ARTIFICIAL INTELLIGENCE

CLASS XII

TEACHER HANDBOOK

Subject Code:843

ARTIFICIAL INTELLIGENCE CURRICULUM

Teacher Handbook for Class XII

Acknowledgments

Patrons

Mr. Rahul Singh, IAS, Chairperson, Central Board of Secondary Education

Strategic Guidance

Dr. Biswajit Saha, Director (Skill Education), Central Board of Secondary Education
Sh. Ravinder Pal Singh, Joint Secretary, Department of Skill Education, Central Board of Secondary Education

Strategic Advisory

Ms. Shipra Sharma, CSR Leader, India/South Asia, IBM
Dr. Mani Madhukar, Program Lead - SkillsBuild, IBM

Curriculum Planning Team

Mr. Manav Subodh, Founder & Chief Mentor, 1M1B
Mr. Saffin Mathew, Programs Director, 1M1B

Lead Curriculum Curator

Ms. Mehreen M Shamim, AI Curriculum Manager, 1M1B

AI Teacher Advisory for Curriculum Revision

Mr. Akhil R, TGT, DAV Public School Pushpanjali Enclave, Delhi
Ms. Anni Kumar, PGT, Vikas Bharati Public School, Delhi
Ms. Harmeet Kaur, PGT, Vasant Valley School, New Delhi
Ms. Jyoti P B, PGT, Jyothis Central School, Thiruvananthapuram, Kerala
Ms. Meena Bisht, TGT, Army Public School Noida, Uttar Pradesh
Mr. Naveen Gupta, PGT, St. Mark's Sr. Sec. Public School, Meera Bagh, Delhi
Ms. Neeru Mittal, PGT, Shaheed Rajpal DAV Public School, Delhi
Ms. Rani Kumari, PGT, DLF Public School, Ghaziabad, Uttar Pradesh
Ms. Roshy Thomas, PGT, Loyola School, Thiruvananthapuram, Kerala
Ms. Shelly Batra, TGT, Mount Carmel School, Dwarka, Delhi
Ms. Smitha R Athreya, PGT, Delhi Public School Rourkela, Odisha
Ms. Soumya Iyer, PGT, Sanskriti School, Pune, Maharashtra
Ms. Swati Sharma, TGT, Heritage Xperiential Learning School, Gurugram, Haryana
Mr. Tushar Upadhyay, TGT, Navrachana Higher Secondary School, Vadodara, Gujarat
Ms. Varsha Vijay K, TGT, Delhi Public School Bangalore North, Bangalore, Karnataka
Ms. Vineeta Garg, PGT, Shaheed Rajpal DAV Public School, Delhi

Foreword

The world around us is undergoing a dramatic transformation, driven by the relentless advancement of Artificial Intelligence (AI). From self-driving cars navigating city streets to virtual assistants understanding complex inquiries, AI is rapidly reshaping industries, societies, and the very way we interact with technology.

This revised textbook, designed for students in Classes XI and XII, dives into the captivating world of AI, offering a comprehensive exploration of its core concepts, applications, and potential impact. As you embark on this journey, you will not only delve into the fascinating algorithms that power AI systems, but also examine its ethical considerations and its profound implications for the future.

This is no longer science fiction. AI is here, and it holds immense potential to improve our lives in countless ways. This textbook equips you, the future generation, with the knowledge and critical thinking skills necessary to navigate this rapidly evolving landscape. Through engaging exercises and thought-provoking questions, you will be challenged to not only understand AI but also to consider its role in your own future.

The Central Board of Secondary Education (CBSE) recognizes the transformative power of Artificial Intelligence (AI) and its impact on the future. Building upon this successful introduction, CBSE extended the AI subject to Classes XI & XII, starting in the 2020-2021 academic session. Thus, allowing students to delve deeper into the world of AI and develop a more comprehensive understanding.

This AI Curriculum has been created with the help of teacher advisors managed by 1M1B and supported by IBM. This curriculum aligns with industry standards as set forth by the National Skills Qualification Framework (NSQF) at Levels 3 & 4.

CBSE acknowledges and appreciates the valuable contribution of IBM India in developing the AI curriculum and conducting training programs. This collaborative effort ensures educators are well-equipped to deliver the AI curriculum effectively.

By working together, CBSE and its partners aim to empower students to embrace the future. By incorporating AI into their learning experience, students gain the knowledge and skills necessary to not only understand AI but also leverage its potential to enhance their learning and future prospects.

The future is full of possibilities, and AI is poised to play a pivotal role. Are you ready to be a part of it?

Embrace the challenge. Explore the potential.

Shape the future with Artificial Intelligence.

INDEX

TOPICS	PAGE NO.
Unit 1: Python Programming-II	1
Unit 2: Data Science Methodology: An Analytic Approach to Capstone Project	21
Unit 3: Making Machines See	56
Unit 4: AI with Orange Data Mining Tool	83
Unit 5: Introduction to Big Data and Data Analytics	107
Unit 6: Understanding Neural Networks	132
Unit 7: Generative AI	159
Unit 8: Data Storytelling	188

UNIT 1: PYTHON PROGRAMMING-II

<table>
<tr><td><b>Title:</b> Python Programming-II</td><td><b>Approach</b>: Hands on, Team Discussion, Web search, Case studies</td></tr>
<tr><td colspan="2"><b>Summary:</b>
This chapter provides a comprehensive review of fundamental python programming and techniques. Students will gain hands-on experience with essential Python libraries, preparing them for more advanced data analysis and machine learning tasks which can be incorporated in their capstone project.</td></tr>
<tr><td colspan="2"><b>Learning Objectives:</b>
1. Review the basics of the NumPy and Pandas library, including arrays, and essential functions.
2. Efficiently import and export data between CSV files and Pandas Data Frames.
3. Implement Linear Regression algorithm, including data preparation, and model training.</td></tr>
<tr><td colspan="2"><b>Key Concepts:</b>
1. Recap of NumPy library
2. Recap of Pandas library
3. Importing and Exporting Data between CSV Files and Data Frames
4. Handling missing values
5. Linear Regression algorithm</td></tr>
<tr><td colspan="2"><b>Learning Outcomes:</b>
Students will be able to:
1. Apply the fundamental concepts of the NumPy and Pandas libraries to perform data manipulation and analysis tasks.
2. Import and export data between CSV files and Pandas Data Frames, ensuring data integrity and consistency.</td></tr>
<tr><td colspan="2"><b>Prerequisites:</b> Foundational understanding of Python from class XI and familiarity with the basic programming.</td></tr>
</table>

Become a Python Powerhouse: A Teacher's Guide to Python Programming-II

This lesson transforms your classroom into a hub of Python programming excellence! Students will master foundational Python libraries and practical techniques, equipping them for advanced data analysis and machine learning.

1. Setting the Stage: Python Libraries

- **Warm-up Activity:** Start with a brief recap of Python libraries like NumPy and Pandas. Engage students in a hands-on activity where they analyze a dataset of their choice, showcasing the power of Python in data manipulation and analysis.

2. Exploring the Detective's Toolkit: NumPy and Pandas Recap

- **NumPy:** Introduce NumPy as a tool for numerical computations and efficient array handling.
  - Activity: Create and manipulate arrays to calculate metrics like mean, median, and standard deviation.
- **Pandas:** Highlight its ability to handle tabular data with ease. Showcase real-world use cases such as analyzing marketing campaigns.
  - Activity: Create a DataFrame from lists or dictionaries and explore basic operations like adding rows and columns.

3. Data Handling and Transformation

- Importing and Exporting CSV Files:
  - Explain the significance of CSV files in data analysis.
  - Activity: Demonstrate reading and writing CSV files in Python, emphasizing best practices for data integrity.
- Handling Missing Values:
  - Discuss strategies like dropping rows or filling missing values using statistical estimates.
  - Activity: Provide a dataset with missing values and guide students in cleaning the data using Pandas.

4. Unleashing the Power of Python: Linear Regression

- Introduction to Linear Regression:
  - Explain the basics of Linear Regression and its use in predicting outcomes.
  - Activity: Use a dataset like "USA_Housing.csv" to implement Linear Regression using Scikit-learn. Guide students through the process of training, testing, and evaluating the model.
- Model Evaluation:
  - Introduce evaluation metrics like Mean Squared Error (MSE) and R-squared values.
  - Activity: Compare the predicted and actual values and calculate the error rates.

5. Practical Applications

- Case Studies:
  - Real-life examples like student performance analysis or marketing campaign insights.
  - Activity: Students choose a small-scale project to apply Python skills, analyse data, and draw actionable insights.

6. Advanced Learning Opportunities

- For Enthusiasts:
  - Introduce advanced concepts like K-Fold Cross-Validation to improve model reliability.
  - Provide online resources and tutorials for deeper exploration of Python programming.

Additional Tips:

- Encourage group work to foster collaboration and problem-solving.
- Provide clear instructions and real-world datasets to make learning relatable.
- Emphasize the iterative nature of data analysis and encourage students to experiment.

By incorporating these elements, educators can equip students with a robust foundation in Python programming, empowering them to tackle real-world challenges with confidence and creativity.

Teachers can ask the following questions to spark curiosity before starting the topics:

- **How do you think businesses analyze the performance of their products or services using data? What tools might they use to make sense of millions of numbers and trends?** (This question introduces the relevance of libraries like Pandas and NumPy in real-world scenarios and connects the lesson to practical applications students can relate to.)
- **If you had a magic tool to predict future events (like house prices or exam scores), what kind of data would you need to make accurate predictions?** (This question sets the stage for discussing Linear Regression and the importance of data handling, preparation, and analysis in predictive modeling.)

1.1. Python Libraries

Python libraries are collections of pre-written codes that we can use to perform common tasks, making our programming life easier. They are like toolkits that provide functions and methods to help us avoid writing code from scratch.

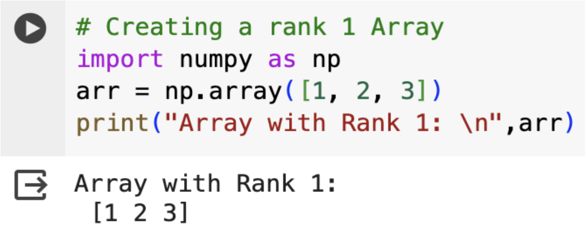
In the realm of data science and analytics, two powerful libraries stand out for their efficiency and versatility: NumPy and Pandas. These libraries form the backbone of data manipulation and analysis in Python, enabling users to handle large datasets with ease and precision.

Let's recap a few of these libraries (covered in Class XI) that are incredibly valuable in the realm of Artificial Intelligence, Data science and analytics.

1.1. 1 NumPy Library

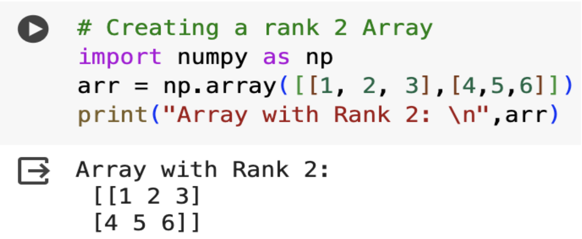
NumPy, short for Numerical Python is a powerful library in Python used for numerical computing. It is a general-purpose array-processing package.

In NumPy, the number of dimensions of the array is called the rank of the array



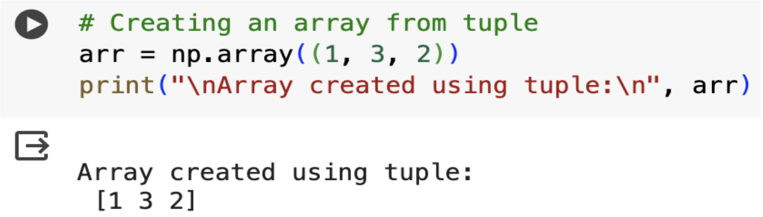
```
# Creating a rank 1 Array
import numpy as np
arr = np.array([1, 2, 3])
print("Array with Rank 1: \n",arr)
```

```
Array with Rank 1:
 [1 2 3]
```



```
# Creating a rank 2 Array
import numpy as np
arr = np.array([[1, 2, 3],[4,5,6]])
print("Array with Rank 2: \n",arr)
```

```
Array with Rank 2:
 [[1 2 3]
 [4 5 6]]
```



```
# Creating an array from tuple
arr = np.array((1, 3, 2))
print("\nArray created using tuple:\n", arr)
```

```

Array created using tuple:
 [1 3 2]
```

1.1.2 Pandas Library

Where and why do we use the Pandas library in Artificial Intelligence?

Suppose we have a dataset containing information about various marketing campaigns conducted by the company, such as campaign type, budget, duration, reach, engagement metrics, and sales performance. **Pandas** is used to load the dataset, display summary statistics, and perform group-wise analysis to understand the performance of different marketing campaigns. We can visualize the sales performance and average engagement metrics for each campaign type using **Matplotlib**, a popular plotting library in Python.

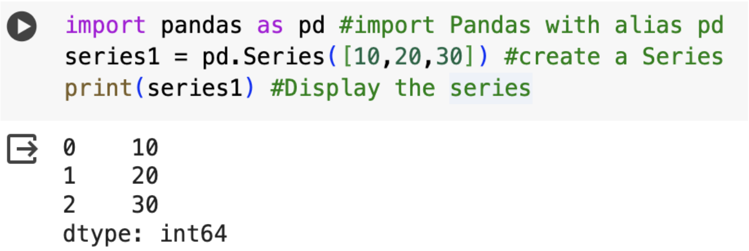
Pandas provides powerful data manipulation and aggregation functionalities, making it easy for us to perform complex analyses and generate insightful visualizations. This capability is invaluable in AI and data-driven decision-making processes, allowing businesses to gain actionable insights from their data.

1.1.2.1 Pandas Data Structures

Pandas generally provides two data structures for manipulating data, They are:

- Series
- Data Frame

i) Creation of a Series from Scalar Values- A Series can be created using scalar values as shown below:



```
import pandas as pd #import Pandas with alias pd
series1 = pd.Series([10,20,30]) #create a Series
print(series1) #Display the series
```

```
0    10
1    20
2    30
dtype: int64
```

ii) Creation of a DataFrame from NumPy arrays

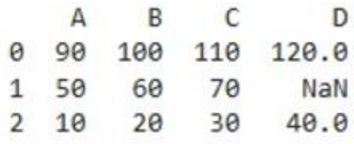
array1=np.array([90,100,110,120])

array2=np.array([50,60,70])

array3=np.array([10,20,30,40])

marksDF = pd.DataFrame([array1, array2, array3], columns=['A', 'B', 'C', 'D'])

print(marksDF)

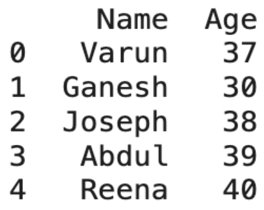


```
    A    B    C      D
0  90  100  110  120.0
1  50   60   70    NaN
2  10   20   30   40.0
```

iii) Creation of a DataFrame from dictionary of array/lists:

```
import pandas as pd
# intialise data of lists.
data = {'Name':['Varun', 'Ganesh', 'Joseph', 'Abdul','Reena'],
       'Age':[37,30,38, 39,40]}
```

```
# Create DataFrame
df = pd.DataFrame(data)
# Print the output.
print(df)
```

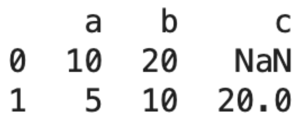


```
     Name  Age
0   Varun   37
1  Ganesh   30
2  Joseph   38
3   Abdul   39
4   Reena   40
```

The dictionary keys become column labels by default in a DataFrame, and the lists become the columns of data.

iv) Creation of DataFrame from List of Dictionaries

```
# Create list of dictionaries
listDict = [{'a':10, 'b':20}, {'a':5,'b':10,'c':20}]
a= pd.DataFrame(listDict)
print(a)
```



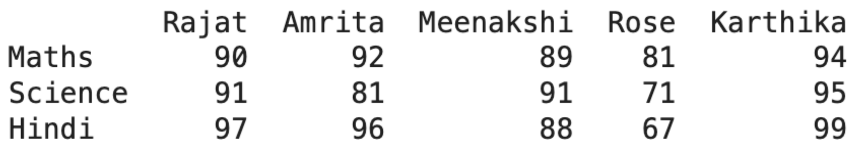
```
    a   b     c
0  10  20   NaN
1   5  10  20.0
```

There will be as many rows as the number of dictionaries present in the list.

1.1.2.2 Dealing with Rows and Columns

i) Adding a New Column to a DataFrame:

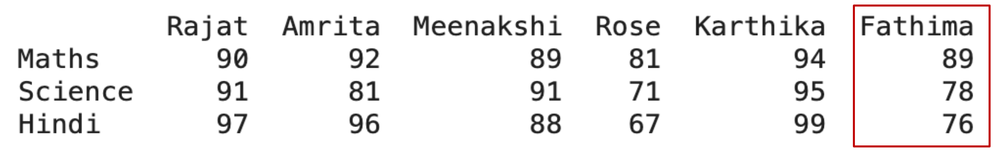
```
ResultSheet={'Rajat': pd.Series([90, 91, 97],index=['Maths','Science','Hindi']),
             'Amrita': pd.Series([92, 81, 96],index=['Maths','Science','Hindi']),
             'Meenakshi': pd.Series([89, 91, 88],index=['Maths','Science','Hindi']),
             'Rose': pd.Series([81, 71, 67],index=['Maths','Science','Hindi']),
             'Karthika': pd.Series([94, 95, 99],index=['Maths','Science','Hindi'])}
Result = pd.DataFrame(ResultSheet)
print(Result)
```



```
         Rajat  Amrita  Meenakshi  Rose  Karthika
Maths       90      92         89    81        94
Science     91      81         91    71        95
Hindi       97      96         88    67        99
```

To add a new column for another student ‘Fathima’, we can write the following statement:

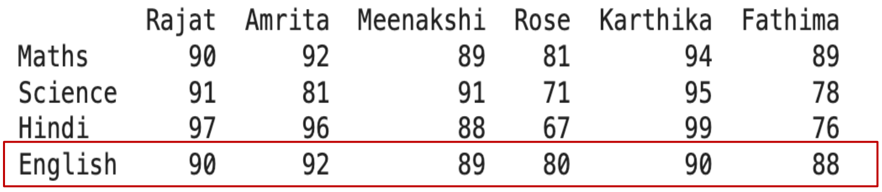
```
Result['Fathima']=[89,78,76]
print(Result)
```



```
         Rajat  Amrita  Meenakshi  Rose  Karthika  Fathima
Maths       90      92         89    81        94       89
Science     91      81         91    71        95       78
Hindi       97      96         88    67        99       76
```

ii) Adding a New Row to a DataFrame:

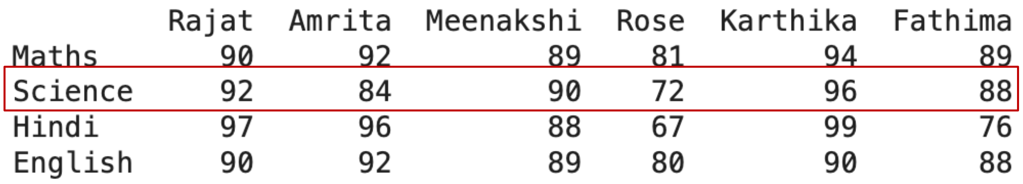
```
Result.loc['English'] = [90, 92, 89, 80, 90, 88]
print(Result)
```



```
         Rajat  Amrita  Meenakshi  Rose  Karthika  Fathima
Maths       90      92         89    81        94       89
Science     91      81         91    71        95       78
Hindi       97      96         88    67        99       76
English     90      92         89    80        90       88
```

DataFRame.loc[] method can also be used to change the data values of a row to a particular value. For example, to change the marks of science.

```
Result.loc['Science'] = [92, 84, 90, 72, 96, 88]
print(Result)
```

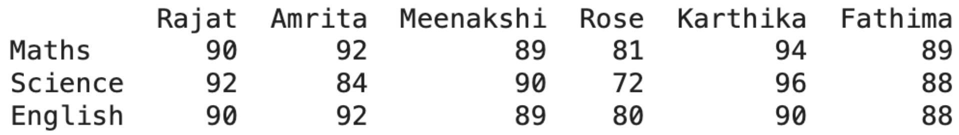


```
         Rajat  Amrita  Meenakshi  Rose  Karthika  Fathima
Maths       90      92         89    81        94       89
Science     92      84         90    72        96       88
Hindi       97      96         88    67        99       76
English     90      92         89    80        90       88
```

1.1.2.3 Deleting Rows or Columns from a DataFrame

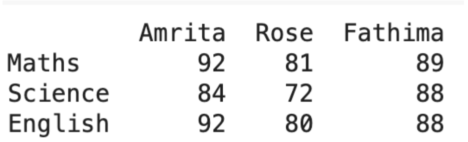
To delete a row, the parameter axis is assigned the value 0 and for deleting a column, the parameter axis is assigned the value 1.

```
Result = Result.drop('Hindi', axis=0) #delete the row "Hindi"
print(Result)
```



```
         Rajat  Amrita  Meenakshi  Rose  Karthika  Fathima
Maths       90      92         89    81        94       89
Science     92      84         90    72        96       88
English     90      92         89    80        90       88
```

```
#delete multiple columns
Result = Result.drop(['Rajat','Meenakshi','Karthika'], axis=1)
print(Result)
```

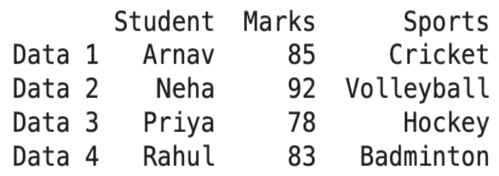


```
         Amrita  Rose  Fathima
Maths        92    81       89
Science      84    72       88
English      92    80       88
```

1.1.2.4 Attributes of DataFrames

We are going to use following data as example to understand the attributes of a DataFrame.

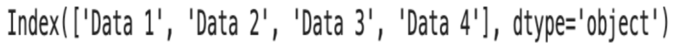
```
import pandas as pd
# creating a 2D dictionary
dict = {"Student": pd.Series(["Arnav","Neha","Priya","Rahul"],
                   index=["Data 1","Data 2","Data 3","Data 4"]),
        "Marks": pd.Series([85, 92, 78, 83],
                 index=["Data 1","Data 2","Data 3","Data 4"]),
        "Sports": pd.Series(["Cricket","Volleyball","Hockey","Badminton"],
                 index=["Data 1","Data 2","Data 3","Data 4"])}
# creating a DataFrame
df = pd.DataFrame(dict)
# printing this DataFrame on the output screen
print(df)
```



```
       Student  Marks      Sports
Data 1   Arnav     85     Cricket
Data 2    Neha     92  Volleyball
Data 3   Priya     78      Hockey
Data 4   Rahul     83   Badminton
```

i) DataFrame.index

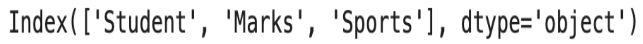
```
>>>df.index
```



```
Index(['Data 1', 'Data 2', 'Data 3', 'Data 4'], dtype='object')
```

ii) DataFrame.columns

```
>>>df.columns
```



```
Index(['Student', 'Marks', 'Sports'], dtype='object')
```

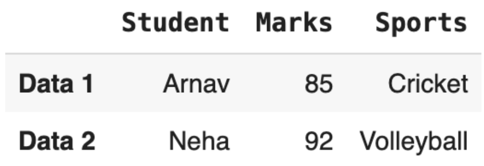
iii) DataFrame.shape

```
>>>df.shape
```

(4,3)

iv) DataFrame.head(n)

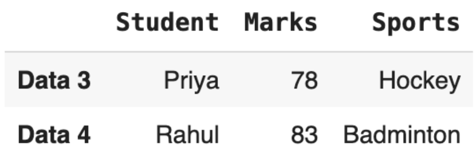
```
>>>df.head(2)
```



	Student	Marks	Sports
Data 1	Arnav	85	Cricket
Data 2	Neha	92	Volleyball

v) DataFrame.tail(n)

```
>>>df.tail(2)
```



	Student	Marks	Sports
Data 3	Priya	78	Hockey
Data 4	Rahul	83	Badminton

1.2. Import and Export Data between CSV Files and DataFrames

CSV files, which stand for Comma-Separated Values, are simple text files used to store tabular data. Each line in a CSV file represents a row in the table, and each value in the row is separated by a comma. This format is widely used because it is easy to read and write, both for humans and computers.

In Python, CSV files are incredibly important for data analysis and manipulation. We often use the Pandas library to load, manipulate, and analyze data stored in CSV files. Pandas provides powerful tools to read CSV files into Data Frames, which are data structures that allow us to perform complex operations on the data with ease. This makes CSV files a go-to format for data scientists and analysts working with Python.

1.2.1 Importing a CSV file to a DataFrame

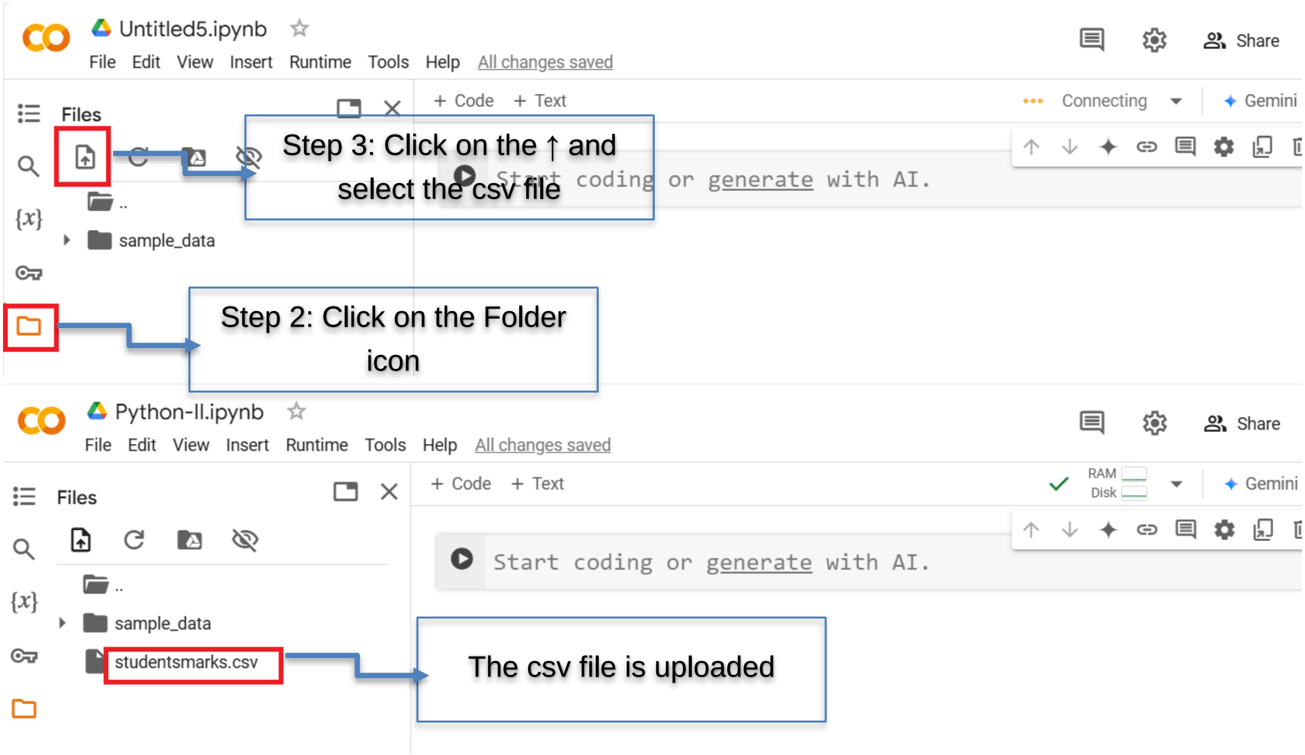
Using the read_csv() function, we can import tabular data from CSV files into pandas DataFrame by specifying a parameter value for the file name **(e.g. pd.read_csv("filename.csv")).**

Let us create a DataFrame from the “studentmarks.csv” file.

Follow the following steps to upload the csv file in the google colab file

Step1: Open the Google colab from the following link https://colab.research.google.com/

And create a new file from the File menu



Step 3: Click on the ↑ and select the csv file

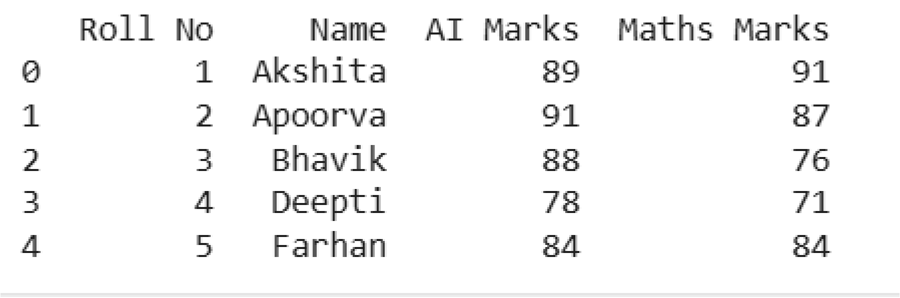
Step 2: Click on the Folder icon

The csv file is uploaded

CO
Untitled5.ipynb
File Edit View Insert Runtime Tools Help All changes saved
Share
Files
+ Code + Text
Connecting
Gemini
Start coding or generate with AI.
..
sample_data
{x}
Python-II.ipynb
File Edit View Insert Runtime Tools Help All changes saved
Share
Files
+ Code + Text
RAM
Disk
Gemini
Start coding or generate with AI.
..
sample_data
studentsmarks.csv

Once the csv file is uploaded then we can execute the following code to convert the csv to DataFrame.

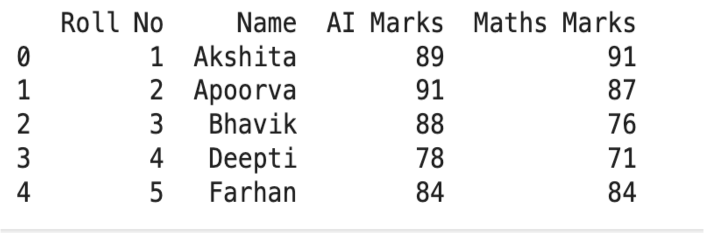
```
import pandas as pd
df=pd.read_csv("studentsmarks.csv")
print(df)
```



```
   Roll No     Name  AI Marks  Maths Marks
0        1  Akshita        89           91
1        2  Apoorva        91           87
2        3   Bhavik        88           76
3        4   Deepti        78           71
4        5   Farhan        84           84
```

On Python IDE we can directly give the complete path name with the csv file in parenthesis.

```
import pandas as pd
import io
df = pd.read_csv('C:/PANDAS/studentsmarks.csv',sep =",", header=0)
print(df)
```

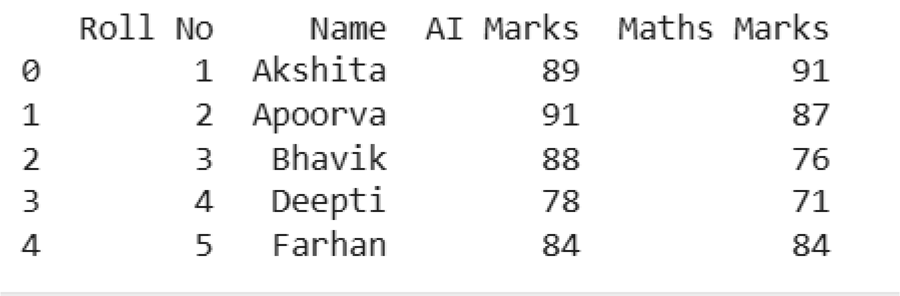


```
   Roll No     Name  AI Marks  Maths Marks
0        1  Akshita        89           91
1        2  Apoorva        91           87
2        3   Bhavik        88           76
3        4   Deepti        78           71
4        5   Farhan        84           84
```

1.2.2 Exporting a DataFrame to a CSV file

We can use the to_csv() function to save a DataFrame to a text or csv file. For example, to save the DataFrame df created in above coding

```
print(df)
```



```
   Roll No     Name  AI Marks  Maths Marks
0        1  Akshita        89           91
1        2  Apoorva        91           87
2        3   Bhavik        88           76
3        4   Deepti        78           71
4        5   Farhan        84           84
```

Program to export this data-

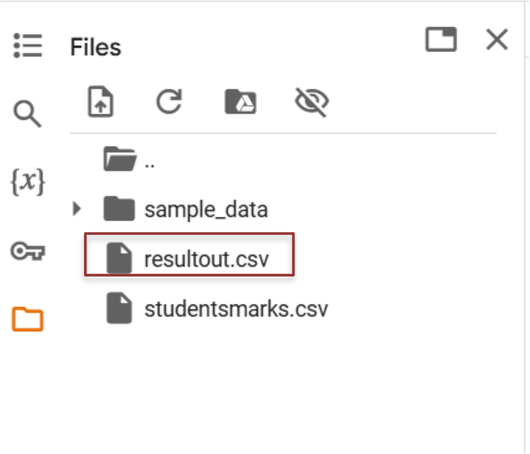
```
df.to_csv(path_or_buf='C:/PANDAS/resultout.csv', sep=',')
```

This creates a file by the name resultout.csv in the hard disk. When we open this file in any text editor or a spreadsheet, we will find the above data along with the row labels and the column headers, separated by comma.

On Google Colab we can write the following code

df.to_csv("resultout.csv",index=False)

The resultout.csv is created



1.3. Handling Missing Values

The two most common strategies for handling missing values explained in this section are:

i) Drop the row having missing values **OR**

ii) Estimate the missing value

Checking Missing Values

Pandas provide a function isnull() to check whether any value is missing or not in the DataFrame. This function checks all attributes and returns True in case that attribute has missing values, otherwise returns False

Drop Missing Values

Dropping will remove the entire row (object) having the missing value(s). This strategy reduces the size of the dataset used in data analysis, hence should be used in case of missing values on few objects.

The dropna() function can be used to drop an entire row from the DataFrame

Estimate the missing value

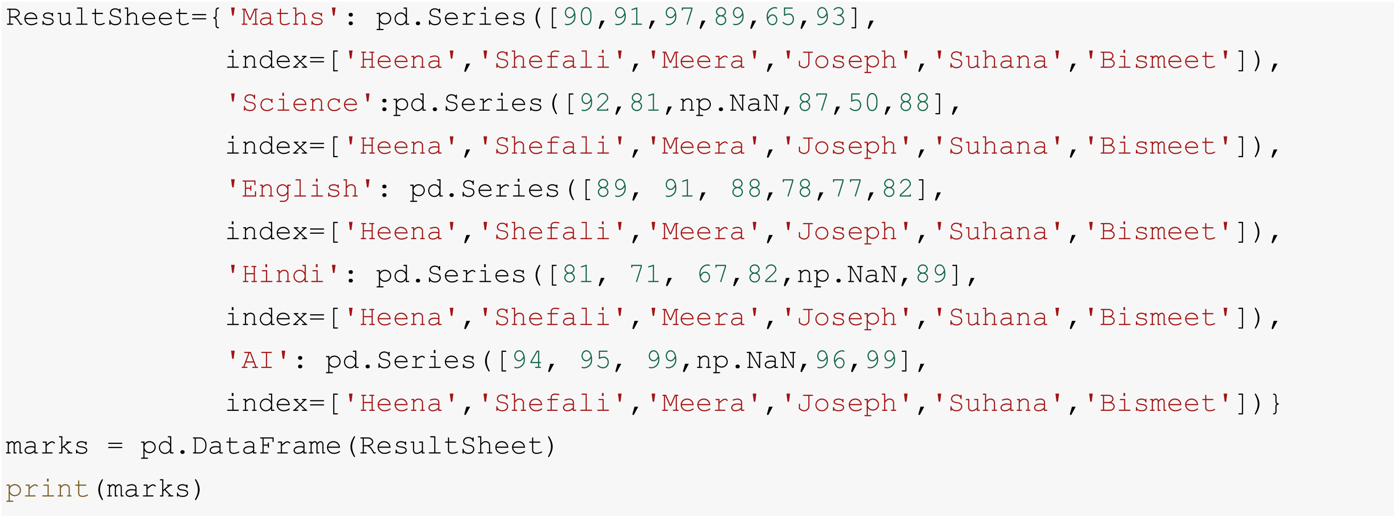
Missing values can be filled by using estimations or approximations e.g a value just before (or after) the missing value, average/minimum/maximum of the values of that attribute, etc. In some cases, missing values are replaced by zeros (or ones).

The fillna(num) function can be used to replace missing value(s) by the value specified in num. For example, fillna(0) replaces missing value by 0. Similarly fillna(1) replaces missing value by 1.

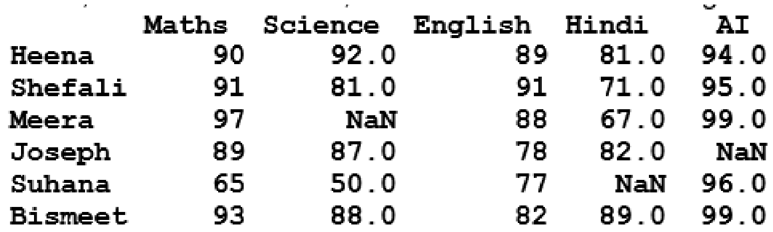
1.4. CASE STUDY

Let's examine a scenario where we have the marks of certain students, but some data is missing in the columns due to specific circumstances. For example, Meera and Suhana couldn't attend the Science and Hindi respectively exam due to fever, Joseph participated in a national-level science exhibition on the day of the AI exam.

Let us feed the data in Python



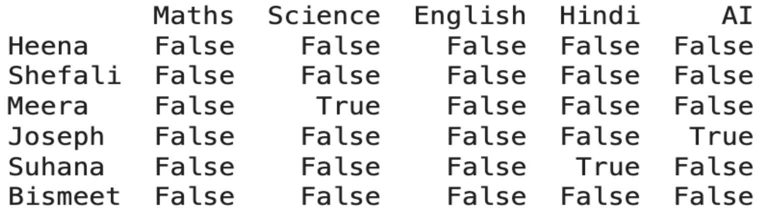
```
ResultSheet={'Maths': pd.Series([90,91,97,89,65,93],
             index=['Heena','Shefali','Meera','Joseph','Suhana','Bismeet']),
             'Science':pd.Series([92,81,np.NaN,87,50,88],
             index=['Heena','Shefali','Meera','Joseph','Suhana','Bismeet']),
             'English': pd.Series([89, 91, 88,78,77,82],
             index=['Heena','Shefali','Meera','Joseph','Suhana','Bismeet']),
             'Hindi': pd.Series([81, 71, 67,82,np.NaN,89],
             index=['Heena','Shefali','Meera','Joseph','Suhana','Bismeet']),
             'AI': pd.Series([94, 95, 99,np.NaN,96,99],
             index=['Heena','Shefali','Meera','Joseph','Suhana','Bismeet'])}
marks = pd.DataFrame(ResultSheet)
print(marks)
```



```
          Maths  Science  English  Hindi    AI
Heena        90     92.0       89   81.0  94.0
Shefali      91     81.0       91   71.0  95.0
Meera        97      NaN       88   67.0  99.0
Joseph       89     87.0       78   82.0   NaN
Suhana       65     50.0       77    NaN  96.0
Bismeet      93     88.0       82   89.0  99.0
```

#check for missing values

>>>print(marks.isnull())



```
         Maths  Science  English  Hindi     AI
Heena    False    False    False  False  False
Shefali  False    False    False  False  False
Meera    False     True    False  False  False
Joseph   False    False    False  False   True
Suhana   False    False    False   True  False
Bismeet  False    False    False  False  False
```

We can see there are three “True” values. So three pieces of data are missing.

>>>print(marks['Science'].isnull().any())

True

print(marks['Maths'].isnull().any())

False

#To find the total number of NaN in the whole dataset

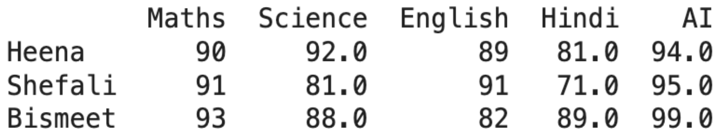
>>>marks.isnull().sum().sum()

3

#apply dropna() for the above case

drop=marks.dropna()

print(drop)

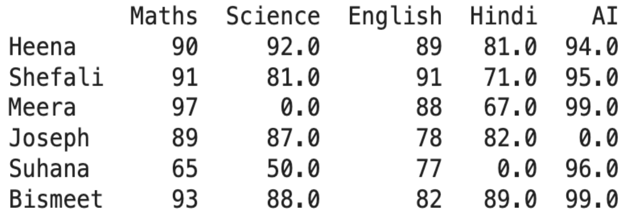


```
         Maths  Science  English  Hindi    AI
Heena       90     92.0       89   81.0  94.0
Shefali     91     81.0       91   71.0  95.0
Bismeet     93     88.0       82   89.0  99.0
```

#Estimate the missing value

FillZero = marks.fillna(0)

print(FillZero)



```
         Maths  Science  English  Hindi    AI
Heena       90     92.0       89   81.0  94.0
Shefali     91     81.0       91   71.0  95.0
Meera       97      0.0       88   67.0  99.0
Joseph      89     87.0       78   82.0   0.0
Suhana      65     50.0       77    0.0  96.0
Bismeet     93     88.0       82   89.0  99.0
```

1.5. PRACTICAL ACTIVITY (**For Advanced Learners)

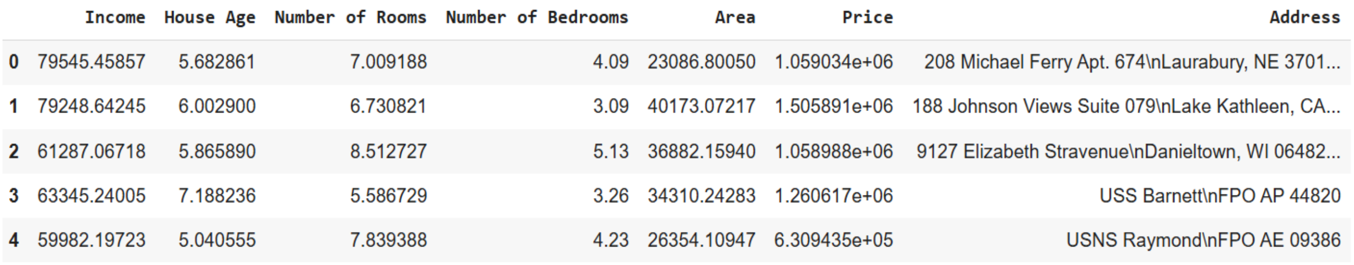
Activity: To implement Linear Regression algorithm

Dataset: https://drive.google.com/drive/folders/1tLbVXWkKzcp6O_-FAvn9usEWuoF3jTp3?usp=sharing

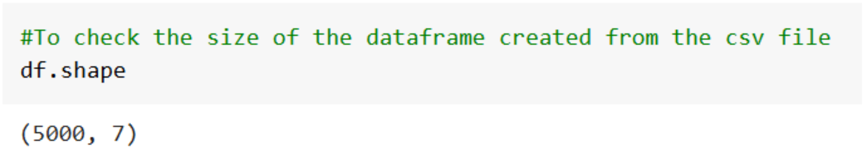
import pandas as pd

df=pd.read_csv('USA_Housing.csv')

df.head()

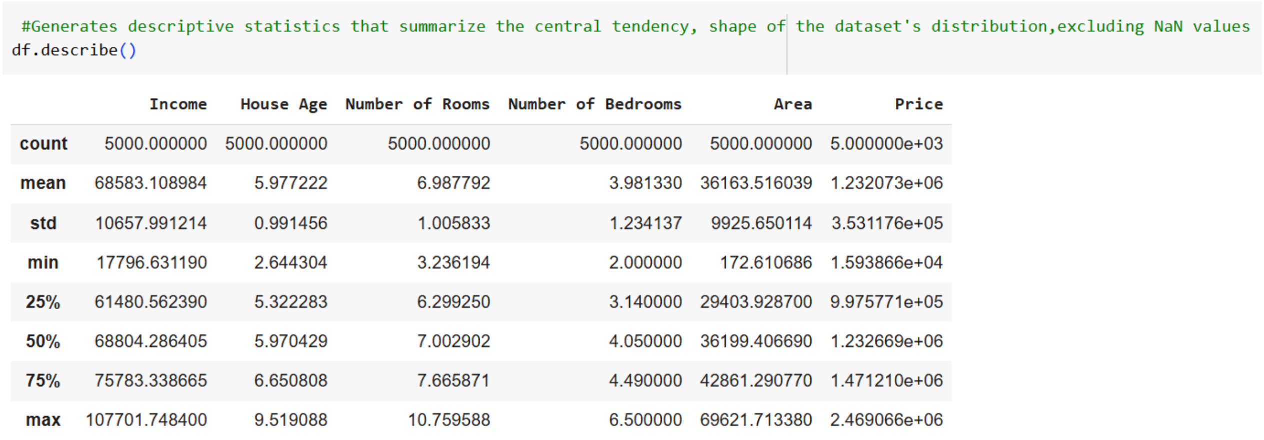


	Income	House Age	Number of Rooms	Number of Bedrooms	Area	Price	Address
0	79545.45857	5.682861	7.009188	4.09	23086.80050	1.059034e+06	208 Michael Ferry Apt. 674\nLaurabury, NE 3701...
1	79248.64245	6.002900	6.730821	3.09	40173.07217	1.505891e+06	188 Johnson Views Suite 079\nLake Kathleen, CA...
2	61287.06718	5.865890	8.512727	5.13	36882.15940	1.058988e+06	9127 Elizabeth Stravenue\nDanieltown, WI 06482...
3	63345.24005	7.188236	5.586729	3.26	34310.24283	1.260617e+06	USS Barnett\nFPO AP 44820
4	59982.19723	5.040555	7.839388	4.23	26354.10947	6.309435e+05	USNS Raymond\nFPO AE 09386



```
#To check the size of the dataframe created from the csv file
df.shape
```

```
(5000, 7)
```

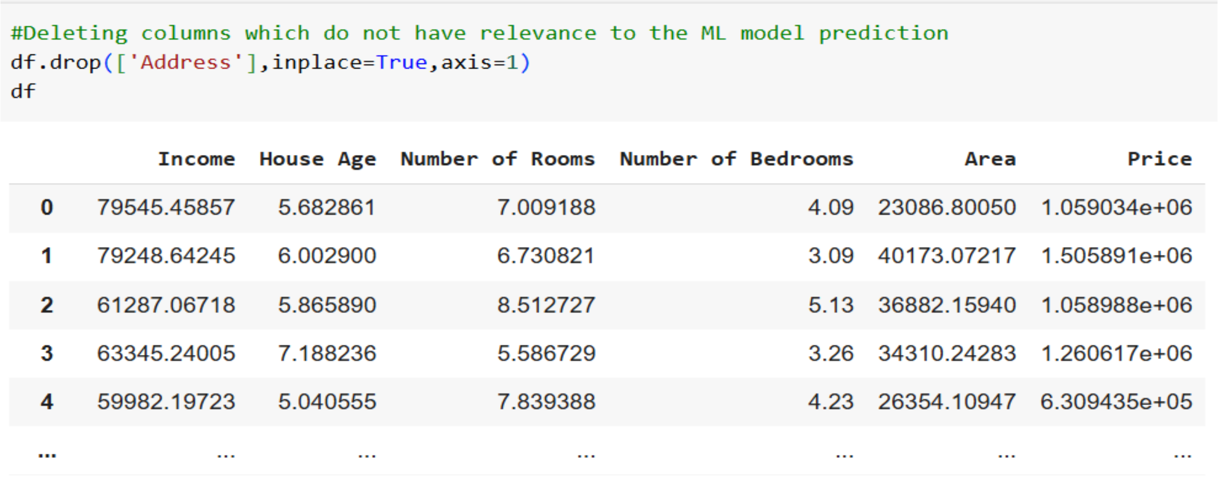


```
#Generates descriptive statistics that summarize the central tendency, shape of the dataset's distribution,excluding NaN values
df.describe()
```

	Income	House Age	Number of Rooms	Number of Bedrooms	Area	Price
count	5000.000000	5000.000000	5000.000000	5000.000000	5000.000000	5.000000e+03
mean	68583.108984	5.977222	6.987792	3.981330	36163.516039	1.232073e+06
std	10657.991214	0.991456	1.005833	1.234137	9925.650114	3.531176e+05
min	17796.631190	2.644304	3.236194	2.000000	172.610686	1.593866e+04
25%	61480.562390	5.322283	6.299250	3.140000	29403.928700	9.975771e+05
50%	68804.286405	5.970429	7.002902	4.050000	36199.406690	1.232669e+06
75%	75783.338665	6.650808	7.665871	4.490000	42861.290770	1.471210e+06
max	107701.748400	9.519088	10.759588	6.500000	69621.713380	2.469066e+06

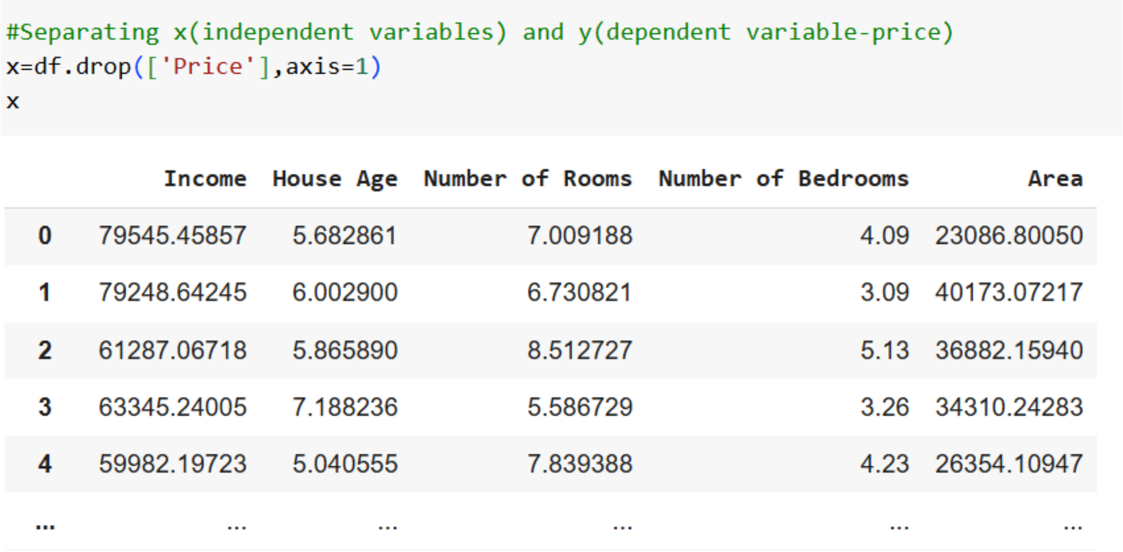
Upon examining the count, it's evident that all columns contain 5000 values. This suggests that there are no missing values in any of the columns.

EXPLORATORY DATA ANALYSIS



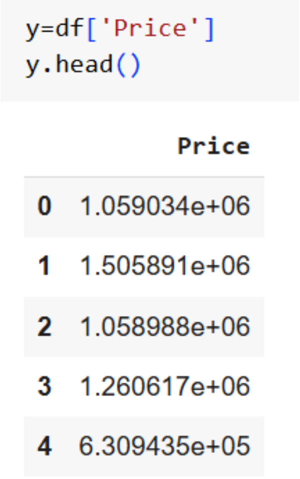
```
#Deleting columns which do not have relevance to the ML model prediction
df.drop(['Address'],inplace=True,axis=1)
df
```

	Income	House Age	Number of Rooms	Number of Bedrooms	Area	Price
0	79545.45857	5.682861	7.009188	4.09	23086.80050	1.059034e+06
1	79248.64245	6.002900	6.730821	3.09	40173.07217	1.505891e+06
2	61287.06718	5.865890	8.512727	5.13	36882.15940	1.058988e+06
3	63345.24005	7.188236	5.586729	3.26	34310.24283	1.260617e+06
4	59982.19723	5.040555	7.839388	4.23	26354.10947	6.309435e+05
...	...	...	...	...	...	...



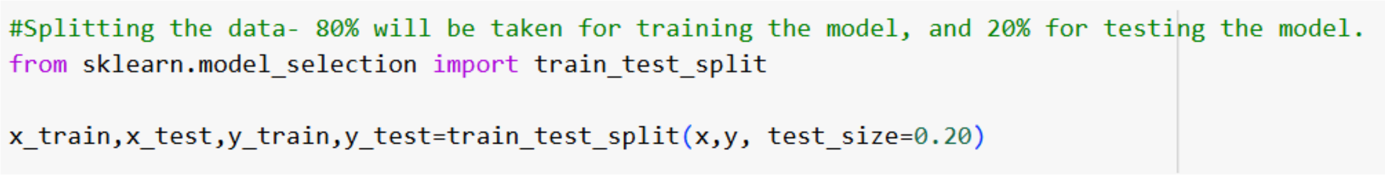
```
#Separating x(independent variables) and y(dependent variable-price)
x=df.drop(['Price'],axis=1)
x
```

	Income	House Age	Number of Rooms	Number of Bedrooms	Area
0	79545.45857	5.682861	7.009188	4.09	23086.80050
1	79248.64245	6.002900	6.730821	3.09	40173.07217
2	61287.06718	5.865890	8.512727	5.13	36882.15940
3	63345.24005	7.188236	5.586729	3.26	34310.24283
4	59982.19723	5.040555	7.839388	4.23	26354.10947
...	...	...	...	...	...

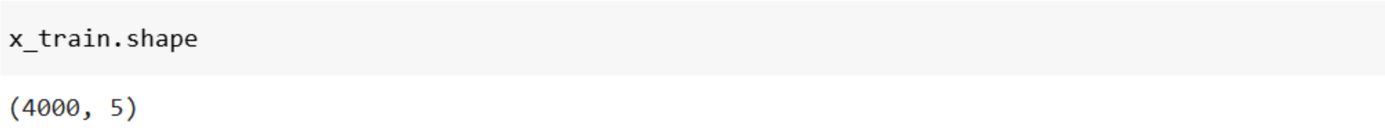


```
y=df['Price']
y.head()
```

	Price
0	1.059034e+06
1	1.505891e+06
2	1.058988e+06
3	1.260617e+06
4	6.309435e+05



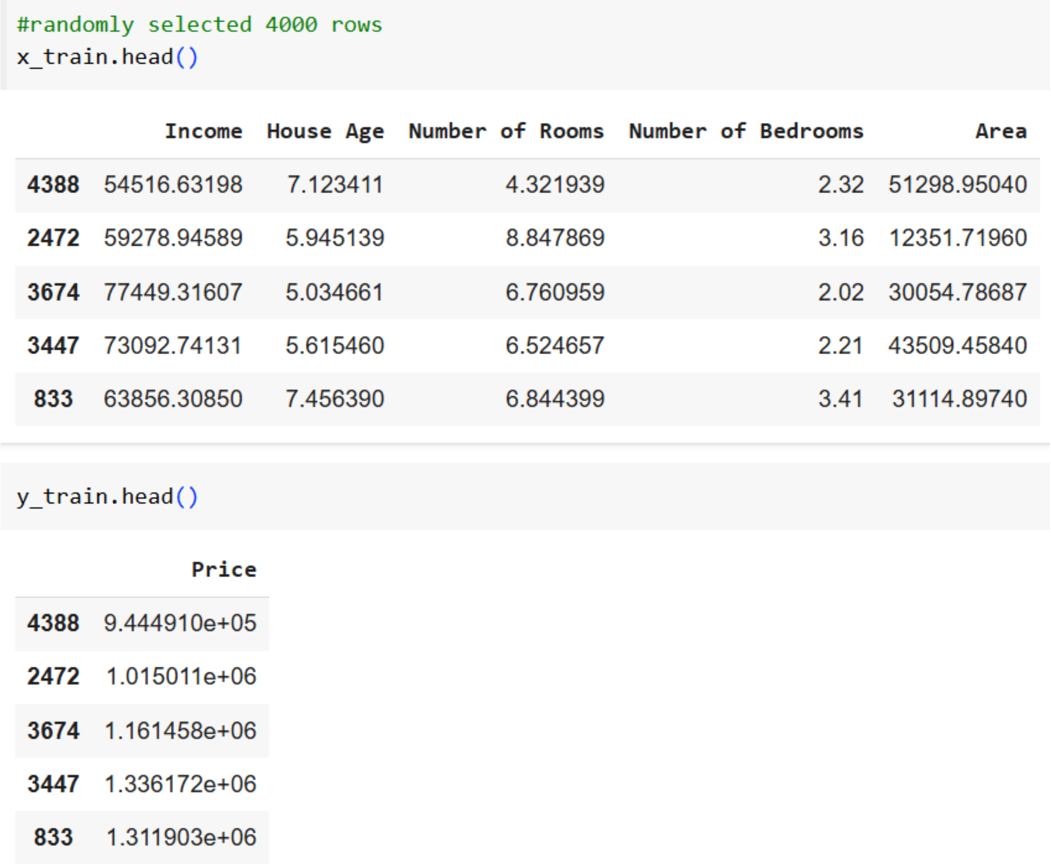
```
#Splitting the data- 80% will be taken for training the model, and 20% for testing the model.
from sklearn.model_selection import train_test_split

x_train,x_test,y_train,y_test=train_test_split(x,y, test_size=0.20)
```



```
x_train.shape
```

```
(4000, 5)
```

From the above output, we understand that 4000 rows (80% of 5000 rows will be used for training the model)

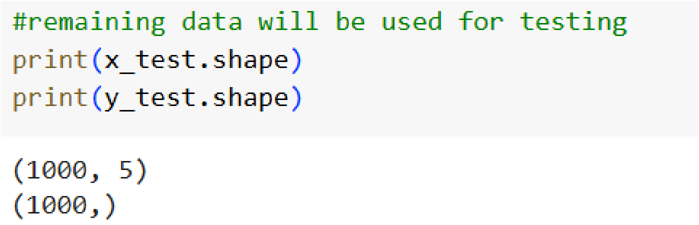


```
#randomly selected 4000 rows
x_train.head()
```

	Income	House Age	Number of Rooms	Number of Bedrooms	Area
4388	54516.63198	7.123411	4.321939	2.32	51298.95040
2472	59278.94589	5.945139	8.847869	3.16	12351.71960
3674	77449.31607	5.034661	6.760959	2.02	30054.78687
3447	73092.74131	5.615460	6.524657	2.21	43509.45840
833	63856.30850	7.456390	6.844399	3.41	31114.89740

```
y_train.head()
```

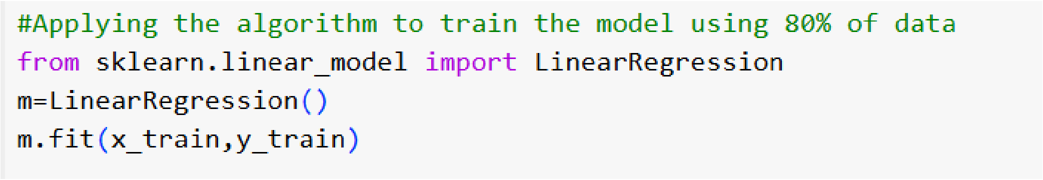
	Price
4388	9.444910e+05
2472	1.015011e+06
3674	1.161458e+06
3447	1.336172e+06
833	1.311903e+06



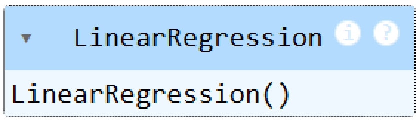
```
#remaining data will be used for testing
print(x_test.shape)
print(y_test.shape)
```

```
(1000, 5)
(1000,)
```

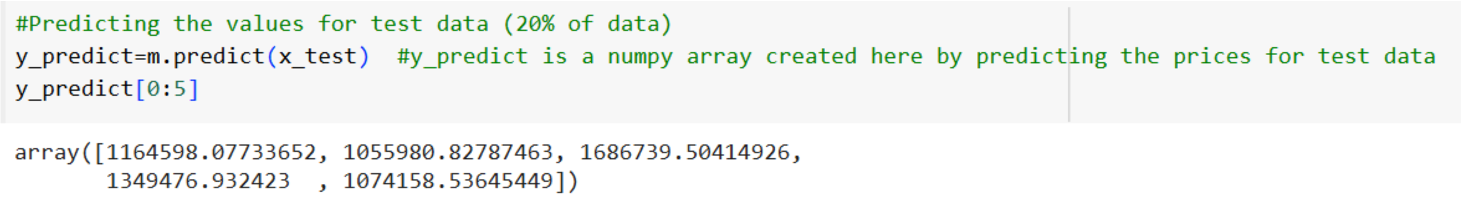
Applying the Linear Regression Algorithm



```
#Applying the algorithm to train the model using 80% of data
from sklearn.linear_model import LinearRegression
m=LinearRegression()
m.fit(x_train,y_train)
```

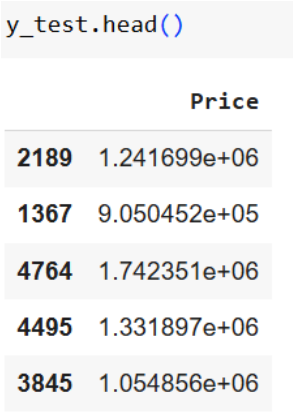


```
▾ LinearRegression
LinearRegression()
```



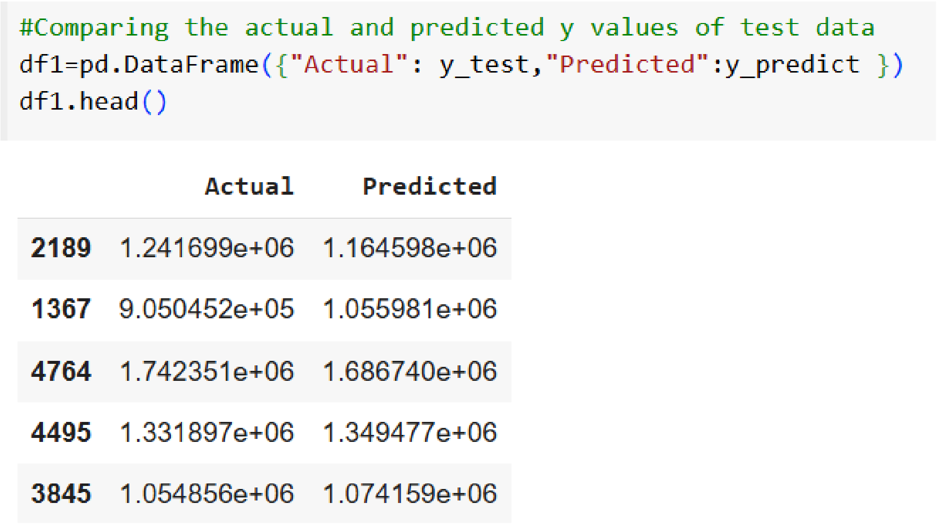
```
#Predicting the values for test data (20% of data)
y_predict=m.predict(x_test)  #y_predict is a numpy array created here by predicting the prices for test data
y_predict[0:5]
```

```
array([1164598.07733652, 1055980.82787463, 1686739.50414926,
       1349476.932423  , 1074158.53645449])
```



```
y_test.head()
```

	Price
2189	1.241699e+06
1367	9.050452e+05
4764	1.742351e+06
4495	1.331897e+06
3845	1.054856e+06



```
#Comparing the actual and predicted y values of test data
df1=pd.DataFrame({"Actual": y_test,"Predicted":y_predict })
df1.head()
```

	Actual	Predicted
2189	1.241699e+06	1.164598e+06
1367	9.050452e+05	1.055981e+06
4764	1.742351e+06	1.686740e+06
4495	1.331897e+06	1.349477e+06
3845	1.054856e+06	1.074159e+06

We observe that there is a difference between the actual and predicted value.

Further, we need to calculate the error, evaluate the model and test the accuracy of the model. This will be covered in the next chapter.

EXERCISES

A. Objective type questions

1. Which of the following is a primary data structure in Pandas?

a) List

b) Tuple

c) Series

d) Matrix

2. What does the fillna(0) function do in Pandas?

a) Removes rows with missing values

b) Fills missing values with zeros

c) Estimates missing values based on averages

d) Converts all data to zero

3. In Linear Regression, which library is typically used for importing and managing data?

a) NumPy

b) Pandas

c) Matplotlib

d) Scikit-learn

4. What is the correct syntax to read a CSV file into a Pandas DataFrame?

a) pd.DataFrame("filename.csv")

b) pd.read_csv("filename.csv")

c) pandas.read_file("filename.csv")

d) pd.file_read("filename.csv")

5. What is the result of the df.shape function?

a) Data type of the DataFrame

b) Number of rows and columns in the DataFrame

c) Memory usage of the DataFrame

d) Column names of the DataFrame

6. Which function can be used to export a DataFrame to a CSV file?

a) export_csv()

b) to_file()

c) to_csv()

d) save_csv()

B. Short Answer Questions

1.What is a DataFrame in Pandas?

Ans: A DataFrame is a 2D data structure in Pandas, similar to a table in a database or Excel sheet. It consists of rows and columns, where each column can hold different types of data.

2.How do you create a Pandas Series from a dictionary?

Ans:

```
import pandas as pd
data = {'a': 1, 'b': 2, 'c': 3}
series = pd.Series(data)
print(series)
```

3.Name two strategies to handle missing values in a DataFrame.

Ans:

- Dropping rows or columns with missing values using dropna().
- Filling missing values using fillna() with mean, median, or a specific value.

4.What does the head(n) function do in a DataFrame?

Ans: It returns the first n rows of the DataFrame.

5.What is the role of NumPy in Python programming?

Ans: NumPy is used for numerical computations. It provides support for arrays, matrices, and mathematical functions like linear algebra and statistical operations.

6.Explain the use of the isnull() function in Pandas.

Ans: The isnull() function checks for missing values in a DataFrame or Series and returns True where data is missing and False otherwise.

C. Long Answer Questions

1.Describe the steps to import and export data using Pandas.

Ans:

- Importing Data: Use pd.read_csv('filename.csv') to read a CSV file into a DataFrame.
- Exporting Data: Use df.to_csv('filename.csv') to save the DataFrame to a CSV file.

Example:

```
import pandas as pd
df = pd.read_csv('data.csv')
df.to_csv('output.csv', index=False)
```

2.Explain the concept of handling missing values in a DataFrame with examples.
Ans:

Missing values can be handled by:

- Dropping rows/columns:

```
df = df.dropna()
```

- Filling with mean/median:

```
df['column'] = df['column'].fillna(df['column'].mean())
```

3.What is Linear Regression, and how is it implemented in Python?
Ans: Linear Regression is a statistical method used to model the relationship between a dependent variable and one or more independent variables.
Example:

```
from sklearn.linear_model import LinearRegression
model = LinearRegression()
model.fit(X_train, y_train)
predictions = model.predict(X_test)
```

4.Compare NumPy arrays and Pandas DataFrames.
Ans:

- NumPy Arrays: Homogeneous, fast for mathematical operations.
- Pandas DataFrames: Heterogeneous, better for data analysis and manipulation.

5.How can we add new rows and columns to an existing DataFrame? Explain with code examples.
Ans:
Add a column:

```
df['new_column'] = [value1, value2, value3]
```

Add a row:

```
df.loc[len(df)] = [value1, value2, value3]
```

6.What are the attributes of a DataFrame? Provide examples.
Ans:

- df.index: Returns the index of the DataFrame.
- df.columns: Returns the column labels.
- df.shape: Returns the dimensions (rows, columns).

D. Case study

1. A dataset of student marks contains missing values for some subjects. Write Python code to handle these missing values by replacing them with the mean of the respective columns.

   Ans:

   ```
   import pandas as pd
   df = pd.DataFrame({'Maths': [90, None, 88], 'Science': [None, 92, 85]})
   df.fillna(df.mean(), inplace=True)
   print(df)
   ```

2. Write Python code to load the file into a Pandas DataFrame, calculate the total sales for each product, and save the results into a new CSV file. Click in the link below to access sales.csv dataset.

   https://drive.google.com/drive/folders/1tLbVXWkKzcp6O_-FAvn9usEWuoF3jTp3?usp=sharing

   Ans:

   ```
   import pandas as pd
   df = pd.read_csv('sales.csv')
   total_sales = df.groupby('Product')['Sales'].sum()
   total_sales.to_csv('total_sales.csv')
   ```

3. In a marketing dataset, analyze the performance of campaigns using Pandas. Describe steps to group data by campaign type and calculate average sales and engagement metrics.

   Ans:

   ```
   grouped = df.groupby('Campaign Type').agg({'Sales': 'mean', 'Engagement': 'mean'})
   print(grouped)
   ```

4. A company has collected data on employee performance. Some values are missing, and certain columns are irrelevant. Explain how to clean and preprocess this data for analysis using Pandas.

   Ans:

   - Drop irrelevant columns using drop().
   - Handle missing values using fillna() or dropna().
   - Normalize or scale data if needed.

     Example:

   ```
   df = df.drop(['Irrelevant Column'], axis=1)
   df.fillna(0, inplace=True)
   ```

References:

https://www.programiz.com/python-programming

https://www.javatpoint.com/python-pandas

https://www.w3schools.com/

UNIT 2: Data Science Methodology: An Analytic Approach to Capstone Project

Title: Capstone Project Using Data Science Methodology

Approach: Hands-on, Team Discussion, Web search, Case studies

Summary:

The Data Science Methodology put forward by John B. Rollins, a Data Scientist in IBM Analytics, is discussed here. The major steps involved in practicing Data Science, from forming a concrete business or research problem, to collecting and analyzing data, to building a model, and understanding the feedback, after model deployment are detailed here. Students can develop their Capstone Project based on this methodology.

Learning Objectives:

1. Understand the major steps involved in tackling a Data Science problem.
2. Define Data Science methodology and articulate its importance.
3. Demonstrate the steps of Data Science Methodology.

Key Concepts:

1. Introduction to Data Science Methodology
2. Steps for Data Science Methodology
3. Model Validation Techniques
4. Model Performance- Evaluation Metrics

Learning Outcomes:

Students will be able to -

1. Integrate Data Science Methodology steps into the Capstone Project.
2. Identify the best way to represent a solution to a problem.
3. Understand the importance of validating machine learning models
4. Use key evaluation metrics for various machine learning tasks

Prerequisites: Foundational understanding of AI concepts from class XI and familiarity with the concept of Capstone Projects and their objectives.

Become a Data Detective: A Teacher's Guide to Data Science Methodology

This lesson equips you to transform your classroom into a data detective agency! Students will learn the Data Science Methodology, a powerful framework to solve problems using data.

1. Case of the Curious Numbers:

- **Warm-up Activity:** Spark curiosity with an interactive game! Begin by saying, "Let's put on our data detective hats!" Present a dataset (e.g., historical sales figures, customer reviews with sentiment analysis scores). Challenge students to ask questions, uncover trends, and identify potential insights hidden within the data. This ignites their interest and highlights the detective-like nature of Data Science.

2. Cracking the Code: Data Science Methodology:

- **Introducing Methodology:** Introduce Data Science Methodology as a structured approach to solving problems using data. Explain how it equips us with the tools and techniques to turn data into actionable insights.

3. The Detective's Toolkit:

- **Key Terminology:** Introduce the key terms within the Data Science Methodology framework:
  - Business Understanding: Defining the business problem and its goals.
  - Problem Approach: Formulating a plan to solve the problem using data.
  - Data Requirements: Identifying the type and amount of data needed.
  - Data Collection: Gathering the required data from various sources.
  - Data Understanding: Exploring and analyzing the data to understand its structure and quality.
  - Data Preparation: Cleaning and transforming the data for analysis.
  - AI Modelling: Building models to make predictions or classifications based on the data (focusing on Linear Regression in this lesson).
  - Evaluation: Assessing the performance of the model (using K-Fold Cross-Validation).
  - Deployment: Integrating the model into a real-world application (briefly discuss for future lessons).
  - Feedback: Monitoring the model's performance and making adjustments as needed (briefly discuss for future lessons).

4. Cracking the Case: Problem Approach Activities:

- **Case Studies:** Present real-world case studies where Data Science Methodology was used to solve a business problem. Guide students through the different stages of the methodology, asking them to consider:
  - What was the business problem?
  - How did they approach the problem with data?
  - What type of data was required?

5. Choosing Your Detective Tools: Introduction to Python:

- **Python Power:** Briefly introduce Python as a popular programming language widely used in Data Science. Highlight its benefits like readability and vast libraries for data analysis (like NumPy, Pandas, Scikit-learn).

6. Building the Model: Train-Test Split and K-Fold Cross-Validation:

- **Train-Test Split:** Explain the concept of train-test split, a technique where data is divided into two sets: training data used to build the model and test data used to evaluate its performance. Demonstrate this with Python code examples using libraries like Scikit-learn.
- **K-Fold Cross-Validation:** Introduce K-Fold Cross-Validation, a robust evaluation technique that splits data into multiple folds, trains the model on different folds, and provides a more reliable estimate of the model's generalizability. Demonstrate this concept with Python code examples using libraries like Scikit-learn.

Additional Tips:

- Encourage students to work in pairs or small groups throughout the lesson to foster collaboration and problem-solving skills.
- Provide online resources and tutorials for students who wish to delve deeper into Python programming.
- Offer opportunities for students to apply the Data Science Methodology to a small-scale project of their own, focusing on a specific problem and utilizing the skills learned in train-test split and K-Fold Cross-Validation.

By incorporating these elements, you can equip students with the skills to approach problems analytically and unlock the power of data to solve real-world challenges.

Teachers can ask the following questions to spark curiosity before starting the topics:

- **Imagine you are at a food festival where dishes are unlabeled, and you need to identify which cuisine each dish belongs to. What strategies would you use to figure it out?** (This question introduces the concept of problem scoping and analytic approaches in an accessible, real-world scenario. It sets the stage for discussing data science methodology by focusing on identifying problems and formulating approaches.)

- **How do you think businesses like Amazon or Netflix decide what products or shows to recommend to you? What kind of questions and data might they consider?** (This question connects students' daily experiences with recommendation systems to concepts like predictive and prescriptive analytics. It encourages them to think about how data can be used to solve practical problems.)

2.1. INTRODUCTION TO DATA SCIENCE METHODOLOGY

A Methodology gives the Data Scientist a **framework** for designing an AI Project. The framework will help the team to decide on the methods, processes and strategies that will be employed to obtain the correct output required from the AI Project. It is the best way to organize the entire project and finish it in a systematic way without losing time and cost.

Data Science Methodology is a process with a prescribed sequence of iterative steps that data scientists follow to approach a problem and find a solution.

Data Science Methodology enables the capacity to handle and comprehend the data.

In this unit, we discuss the steps of Data Science Methodology which was put forward by John Rollins, a Data Scientist at IBM Analytics. It consists of 10 steps. The foundation methodology of Data Science provides a deep insight on how every AI project can be solved from beginning to end. There are five modules, each going through two stages of the methodology, explaining the rationale as to why each stage is required.

1. From Problem to Approach
2. From Requirements to Collection
3. From Understanding to Preparation
4. From Modelling to Evaluation
5. From Deployment to Feedback

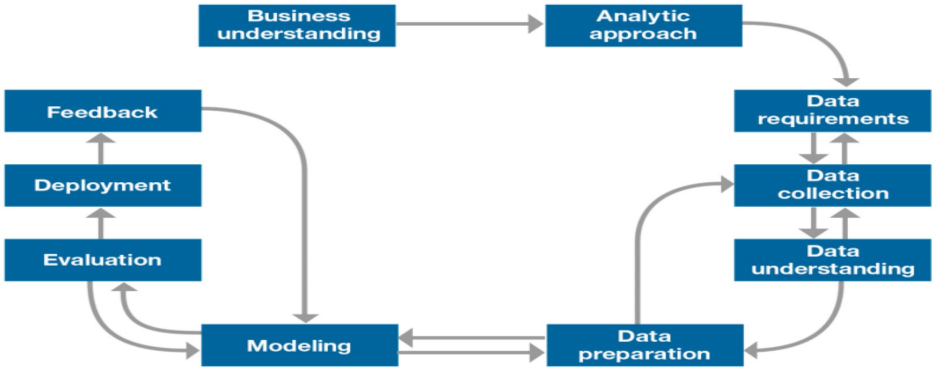


Figure 1. Foundational methodology for data science.

Source: *https://cognitiveclass.ai/courses/data-science-methodology-2*

1. From Problem to Approach

2.1.1 Business understanding

What is the problem that you are trying to solve?

In this stage, first, we understand the problem of the customer by asking questions and try to comprehend what is exactly required for them. With this understanding we can figure out the objectives that support the customer's goal. This is also known as Problem Scoping and defining. The team can use 5W1H Problem Canvas to deeply understand the issue. This stage also involves using DT (Design Thinking) Framework.

To solve a problem, it's crucial to understand the customer's needs. This can be achieved by asking relevant questions and engaging in discussions with all stakeholders. Through this process, we will be able to identify the specific requirements and create a comprehensive list of business needs.

Activity 1:

Mr. Pavan Sankar visited a food festival and wants to sample various cuisines. But due to health concerns, he has to avoid certain dishes. However, since the dishes were not categorized by cuisine, he found it challenging and wished for assistance in identifying the cuisines offered.

Q1. Are You ready to help Pavan Sankar?

Yes

Q2. What do you think is the actual problem of Pavan?

Pavan could not identify the cuisine of dishes kept in the food festival.

Q3. Can we predict the cuisine of a given dish using the name of the dish only?

No, it is difficult as the names of all dishes might be unknown.

Q4. Let's check. Following dish names were taken from the menu of a restaurant in India

- Bubble and Squeak
- Phan Pyat
- Jadoh

Are you able to tell the cuisines of these dishes?

No

Q5. Is it possible to determine the dish name and cuisine of a dish with its images alone?

No, some images will look similar. So, it is difficult to identify the cuisine.

Q6. What about determining the cuisine of a dish based on its ingredients?

To Certain extent we could identify the cuisine with the ingredients of dishes.

Q7. What is the name of the dish prepared from the ingredients given in Fig 2.1?

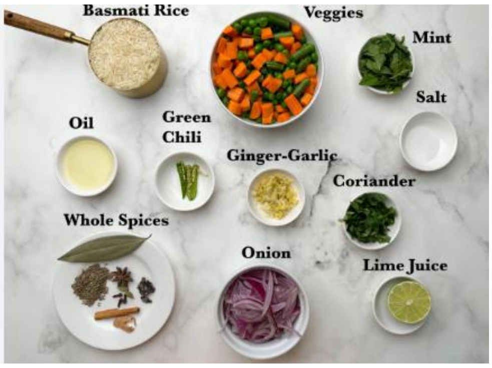


Fig 2.1

Vegetable Pulav

2.1.2 Analytic Approach

How can you use the data to answer the question?

When the business problem has been established clearly, the data scientist will be able to define the analytical approach to solve the problem. This stage involves seeking clarification from the person who is asking the question, so as to be able to pick the most appropriate path or approach. Let us understand this in detail.

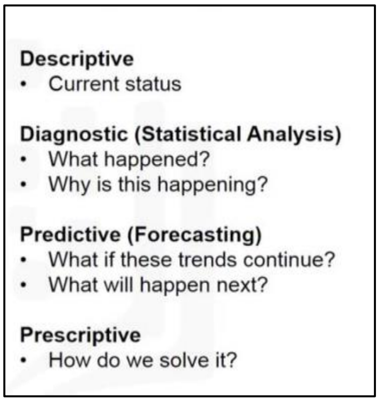
This stage involves asking more questions to the stakeholders so that the AI Project team can decide on the correct approach to solve the problem.

The different questions that can be asked now are:

1. Do I need to find how much or how many? (Regression)
2. Which category does the data belong to? (Classification)
3. Can the data be grouped? (Clustering)
4. Is there any unusual pattern in the data? (Anomaly detection)
5. Which option should be given to the customer? (Recommendation)

To solve a particular problem, there are four main types of data analytics as shown in Fig 2.2. They are:

1. Descriptive Analytics
2. Diagnostic Analytics
3. Predictive Analytics
4. Prescriptive Analytics



Descriptive
- Current status

Diagnostic (Statistical Analysis)
- What happened?
- Why is this happening?

Predictive (Forecasting)
- What if these trends continue?
- What will happen next?

Prescriptive
- How do we solve it?

Fig 2.2

Let's understand each of them.

Descriptive Analytics: This summarizes past data to understand what has happened. It is the first step undertaken in data analytics to describe the trends and patterns using tools like graphs, charts etc. and statistical measures like mean, median, mode to understand the central tendency. This method also examines the spread of data using range, variance and standard deviation.
For example: To calculate the average marks of students in an exam or analyzing sales data from the previous year.

Diagnostic Analytics: It helps to understand the reason behind why some things have happened. This is normally done by analyzing past data using techniques like root cause analysis, hypothesis testing, correlation analysis etc. The main purpose is to identify the causes or factors that led to a certain outcome.
For example: If the sales of a company dropped, diagnostic analysis will help to find the cause for it, by analyzing questions like "Is it due to poor customer service" or "low product quality" etc.

Predictive Analytics: This uses the past data to make predictions about future events or trends, using techniques like regression, classification, clustering etc. Its main purpose is to foresee future outcomes and make informed decisions.
For example: A company can use predictive analytics to forecast its sales, demand, inventory, customer purchase pattern etc., based on previous sales data.

Prescriptive Analytics: This recommends the action to be taken to achieve the desired outcome, using techniques such as optimization, simulation, decision analysis etc. Its purpose is to guide decisions by suggesting the best course of action based on data analysis.
For example: To design the right strategy to increase the sales during festival season by analyzing past data and thus optimize pricing, marketing, production etc.

We can summarize each of these analytics as given in Table 2.1

	Descriptive Analytics	Diagnostic Analytics	Predictive Analytics	Prescriptive Analytics
Focus	Questions on summarizing historical data	Questions on understanding why certain events occurred	Questions on predicting future outcomes based on historical data patterns	Questions on determining the best course of action
Purpose	Identify patterns, trends, and anomalies in past data	Uncover root causes and factors contributing to specific outcomes	Forecast future events or behaviors	Recommend specific actions or interventions based on predictive insights. May indirectly influence classification through recommendations

Table 2.1

[To know more about these analytics, you can go through this Coursera activity]
https://www.coursera.org/learn/data-science-methodology/

Activity 2:

Mr. Pavan Sankar has set his goal to find the dish and its cuisine using its ingredients. He plans to proceed as shown in the flowchart in Fig 2.3.

Observe the flowchart and answer the questions.

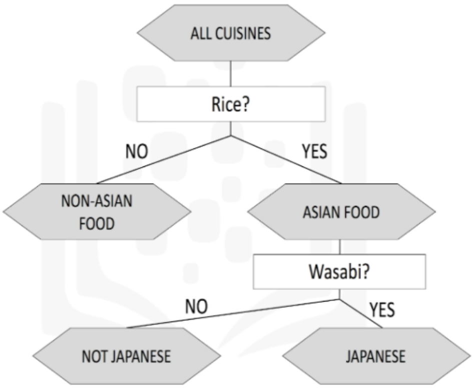


Fig 2.3

Q1. Which type of analytics questioning is being utilized here?

A) Descriptive Analytics

B) Diagnostic Analytics

C) Predictive Analytics

D) Prescriptive Analytics

Q2. What type of approach is chosen here?

Classification Approach

Q3. Which algorithm is depicted in the figure given here?

Decision Tree

By thoroughly understanding the problem at hand and formulating a clear approach to address it, the stage is set for effective decision-making and problem-solving. This initial stage sets the direction for the entire project, ensuring that efforts are focused on solving the right problem in the most efficient and effective manner possible. As such, careful consideration and thoughtful planning during this stage are vital for achieving meaningful results and delivering value to stakeholders.

2. From Requirements to Collection

2.1.3 Data requirements

WHO?
WHAT?
HOW?
WHERE?
WHY?
WHEN?

What are the data requirements?

The requirements of data are determined by the analytic approach chosen in the previous stage. The 5W1H questioning method can be employed in this stage also to determine the data requirements. It is necessary to identify the data content, formats, and sources for initial data collection, in this stage.

Determining the specific information needed for our analysis or project includes:

- identifying the types of data required, such as numbers, words, or images.
- considering the structure in which the data should be organized, whether it is in a table, text file, or database.
- identifying the sources from which we can collect the data, and
- any necessary cleaning or organization steps required before beginning the analysis.

This stage involves defining our data requirements, including the type, format, source, and necessary preprocessing steps to ensure the data is usable and accurate for our needs. Data for a project can be categorized into three types: **structured data** (organized in tables, e.g., customer databases), **unstructured data** (without a predefined structure, e.g., social media posts, images), and **semi-structured data** (having some organization, e.g., emails, XML files). Understanding these data types is essential for effective data collection and management in project development

Activity 3:

Mr. Pavan Sankar is now ready with a classification approach. Now he needs to identify the data requirements.

Q1. Write down the name of two cuisines, five dishes from each cuisine and the ingredients needed for the five dishes separately.

Cuisine: Indian

Dish1 – Aloo gobi-	Ingredients – Potato, Cauliflower, Masalas, Oil, Salt
Dish2 – Naan-	Ingredients – Flour, Yeast, Salt, Milk
Dish3 – Butter Chicken-	Ingredients –Chicken, Butter, Masala, Oil, Salt
Dish4 –Gulab Jamun-	Ingredients – Dough, Oil, Sugar
Dish5 –Poha-	Ingredients – Rice, Potato, Turmeric

Cuisine – Chinese

Dish1—Manchow soup-	Ingredients –Vegetables, Ginger, Garlic, Soy sauce, Chilly
Dish2—Mapo Tofu-	Ingredients –Tofu, Pork,Vegetable, Meat, Soy sauce, Chilly, Garlic etc.
Dish3-Chow Mein-	Ingredients –Noodles, Sesame oil, Chicken, Garlic, Soy Sauce
Dish4-Chicken Fried Rice-	Ingredients –Rice, Vegetables, Chicken, Soy sauce, Oil, Salt
Dish5—Char Siu	Ingredients –Soy Sauce, Garlic, Honey, Spices, Sesame oil

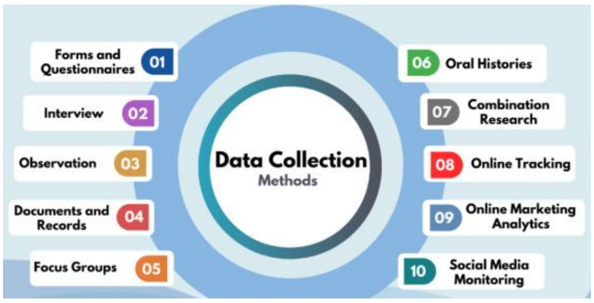
Q2. To collect the data on ingredients, in what format should the data be collected?

Data can be collected in table format.

Text file can also be created.

For available dishes, images can also be collected.

2.1.4 Data collection



What occurs during data collection?

Data collection is a systematic process of gathering observations or measurements. In this phase, the data requirements are revised and decisions are made as to whether the collection requires more or less data. Today's high-performance database analytics enable data scientists to utilize large datasets. There are mainly two sources of data collection:

→ **Primary data Source**

A primary data source refers to the original source of data, where the data is collected firsthand through direct observation, experimentation, surveys, interviews, or other methods. This data is raw, unprocessed, and unbiased, providing the most accurate and reliable information for research, analysis, or decision-making purposes. Examples include marketing campaigns, feedback forms, IoT sensor data etc.

→ **Secondary data Source**

A secondary data source refers to the data which is already stored and ready for use. Data given in books, journals, websites, internal transactional databases, etc. can be reused for data analysis. Some methods of collecting secondary data are social media data tracking, web scraping, and satellite data tracking. Some sources of online data are data.gov, World bank open data, UNICEF, open data network, Kaggle, World Health Organization, Google etc. Smart forms are an easy way to procure data online.

DBA's and programmers often work together to extract data from both primary and secondary sources. Once the data is collected, the data scientist will have a good understanding of what they will be working with. The Data Collection stage may be revisited after the Data Understanding stage, where gaps in the data are identified, and strategies are developed to either collect additional data or make substitutions to ensure data completeness.

Activity 4:

Q1. If you need the names of American cuisine, how will you collect the data?

To get the dish names of American cuisine, we can use Web scraping. Personal Interviews with Americans is also possible in case if Americans are there nearby.

Q2. You want to try out some healthy recipes in the Indian culture. Mention the different ways you could collect the data.

Collect the data directly from the places where the culture is maintained, interview grandparents, refer text book which have the cultural context.

Q3. How can you collect a large amount of data and where can it be stored?

Large amount of data can be collected through Online sources. Many websites provide large data sets free of use. It can be stored in the form of CSV file or relational database in the cloud.

3. From Understanding to Preparation

2.1.5 Data Understanding

Is the data collected representative of the problem to be solved?

Data Understanding encompasses all activities related to constructing the dataset. In this stage, we check whether the data collected represents the problem to be solved or not. The relevance, comprehensiveness, and suitability of the data for addressing the specific problem or question at hand are evaluated. Techniques such as descriptive statistics (univariate analysis, pairwise correlation etc.) and visualization (Histogram) can be applied to the dataset, to assess the content, quality, and initial insights about the data.

Activity 5:

Q1. Semolina which is called *rava or suji* in Indian households is a by-product of durum wheat. Name a few dishes made from semolina. How will you differentiate the data of different dishes?

Upma, Rava Kichadi, Kesari, Suji Pancakes etc.

Main ingredients of all dishes are Suji. Based on salt or sugar, it becomes sweet dish or not.

With different ingredients added to the base ingredient Suji, different type of dishes is made.

Q2. Given below is a sample data collected during the data collection stage. Let us try to understand it.

Table -2.2

Dish	Country	Salt	Chilli	Onion	Oil	Rice	Fish	Chicken	Potato	Milk	Sugar	Vegetable	Wasabi	Soy sauce
Chicken Biriyani	Indian	1	1	1	1	1	0	2	0	0	0	0	0	0
Kheer	Indian	0	0	0	N	1	0	0	0	1	1	0	0	0
Pulao	Indiana	1	1	1	1	One	0	0		0	0	1	0	
Sushi	Japanese	1	0	1	1	1	1	0	0	0	N	0	1	0
Fried Rice	Chinese	Y	1	1	1	1	0	1	0	1	0	1	0	1

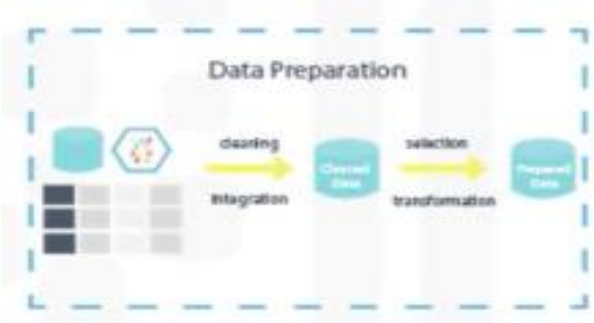
a. Basic ingredients of sushi are rice, soy sauce, wasabi and vegetables. Is the dish listed in the data? Are all ingredients available?
Yes. Vegetables and Soy Sauce is not available in the data.

b. Find out the ingredients for the dish “Pulao”. Check for invalid data or missing data.
Common Ingredients for Fried Rice are Rice, Vegetables, Oil, Garlic, Soy sauce, Salt, Chilli, Onion.
Here in the data Garlic is not found.

c. Inspect all columns for invalid, incorrect or missing data and list them below.
Invalid: Salt (Y), Oil (N), Sugar(N)
Incorrect: Rice (one), Chicken(2)
Missing data: Potato, Soy sauce

d. Which ingredients is common for all dishes? Which ingredient is not used for any dish?
Common- Rice, Not common- Potato

2.1.6 Data preparation



What additional work is required to manipulate and work with the data?

This stage covers all the activities to build the set of data that will be used in the modelling step. Data is transformed into a state where it is easier to work with.

Data preparation includes

- cleaning of data (dealing with invalid or missing values, removal of duplicate values and assigning a suitable format)
- combine data from multiple sources (archives, tables and platforms)
- transform data into meaningful input variables

Feature Engineering is a part of Data Preparation. The preparation of data is the most time-consuming step among the Data Science stages.

Feature engineering is the process of selecting, modifying, or creating new features (variables) from raw data to improve the performance of machine learning models.

For example:

Suppose you're building an ML model to predict price of houses, and you have the following data:

- Raw Data: Area of the house (in sq.ft), number of bedrooms, and the year the house was built.

To improve the model's performance, you might create new features such as age of the house and price per square foot which can be derived from the raw data.

- New Features

1. Age of the house = Current year - Year built.
2. Price per square foot = Price of the house / Area.

These new features can help the model make more accurate predictions.

Activity 6:

Q1. Are there any textual mistakes in the data given in the Table-1? Mention if any.

Yes, For Pulao dish, as value of country “Indiana” is given instead of “Indian”.

Q2. In Table-1, incorrect data was identified in the columns rice and chicken. Write the possible ways to rectify them.

In the column Rice it is written as “one” instead of 1 and in the column Chicken “2” is written.

Q3. Is the first column name appropriate? Can you suggest a better name?

No, as this is the table of dishes and cuisine, column name country can be replaced with” Cuisine”.

Q4. First three values of the first column seem to be similar. Do we need to make any corrections to this data?

First three values are similar as the dishes belongs to the same cuisine. “Indiana’ may be changed to “Indian”.

Q5. Do the dishes with common ingredients come under the same cuisine? Why?

Yes, some ingredients may be common for the dishes under the same cuisine. It is determined by the culture and food habits of people under the cuisine.

Q6. Instead of mentioning whether the ingredient is present or not by using 0's and 1's, can you suggest any alternative ways to display the information?

- Tables of dish names and ingredients list can be taken.
- An image with dish and its ingredients can be collected.

4. From Modelling to Evaluation

2.1.7 AI modelling

In what way can the data be visualized to get to the required answer?

The modelling stage uses the initial version of the dataset prepared and focuses on developing models according to the analytical approach previously defined. The modelling process is usually iterative, leading to the adjustments in the preparation of data. For a determined technique, Data scientists can test multiple algorithms to identify the most suitable model for the Capstone Project.

Data Modelling focuses on developing models that are either descriptive or predictive.

1. **Descriptive Modeling**: It is a concept in data science and statistics that focuses on summarizing and understanding the characteristics of a dataset without making predictions or decisions. The goal of descriptive modeling is to describe the data rather than predict or make decisions based on it. This includes summarizing the main characteristics, patterns, and trends that are present in the data. Descriptive modeling is useful when you want to understand what is happening within your data and how it behaves, but not necessarily why it happens.

   Common Descriptive Techniques:

   - Summary Statistics: This includes measures like:
     - Mean (average), Median, Mode
     - Standard deviation, Variance
     - Range (difference between the highest and lowest values)
     - Percentiles (e.g., quartiles)
   - Visualizations: Graphs and charts to represent the data, such as:
     - Bar charts
     - Histograms
     - Pie charts
     - Box plots
     - Scatter plots

2. **Predictive modeling**: It involves using data and statistical algorithms to identify patterns and trends in order to predict future outcomes or values. It relies on historical data and uses it to create a model that can predict future behavior or trends or forecast what might happen next. It involves techniques like regression, classification, and time-series forecasting, and can be applied in a variety of fields, from predicting exam scores to forecasting weather or stock prices. While it is a powerful tool, students must also understand its limitations and the importance of good data.

   The data scientist will use a training set for predictive modeling. A training set is a set of historical data in which the outcomes are already known. The training set acts like a gauge to determine if the model needs to be calibrated. In this stage, the data scientist will play around with different algorithms to ensure that the variables selected are actually required.

Activity 7:

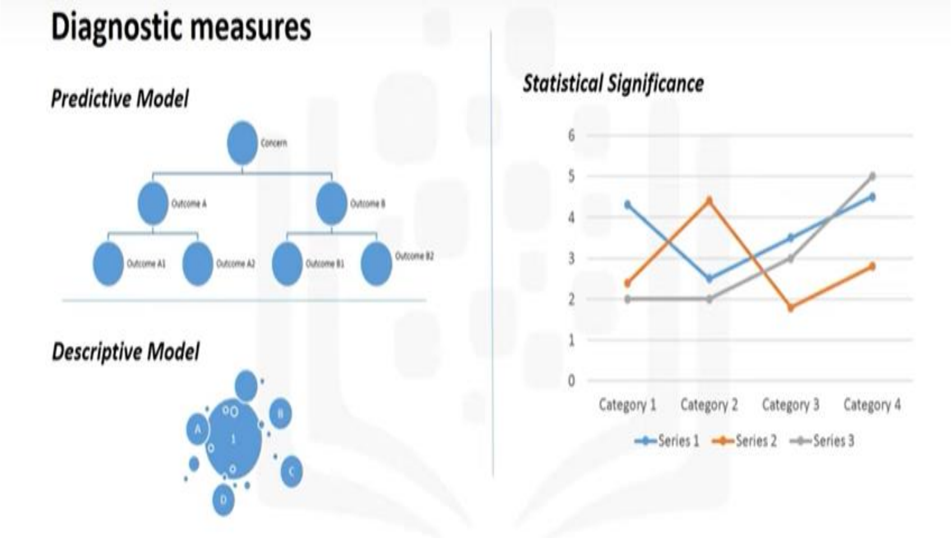
Q1. Name two programming languages which can be used to implement the Decision Tree Algorithm.

Python, R

Q2. In the problem of identifying dish name and cuisine, if we chose the algorithm Decision Tree to solve the problem and choose Python as a tool, name some libraries which will help in the implementation.

numpy, pandas, re, sklearn, matplotlib, itertools, random

2.1.8 Evaluation



Does the model used really answer the initial question or does it need to be adjusted?

https://medium.com/ml-research-lab/part-4-data-science-methodology-from-modelling-to-evaluation-3fb3c0cdf805

Evaluation in an AI project cycle is the process of assessing how well a model performs after training. It involves using test data to measure metrics like accuracy, precision, recall, or F1 score. This helps determine if the model is reliable and effective before deploying it in real-world situations.

Model evaluation can have two main phases.

First phase – Diagnostic measures

It is used to ensure the model is working as intended. If the model is a predictive model, a decision tree can be used to evaluate the output of the model, check whether it is aligned to the initial design or requires any adjustments. If the model is a descriptive model, one in which relationships are being assessed, then a testing set with known outcomes can be applied, and the model can be refined as needed.

Second phase – Statistical significance test

This type of evaluation can be applied to the model to verify that it accurately processes and interprets the data. This is designed to avoid unnecessary second guessing when the answer is revealed.

Activity 8:

Q1. In the cuisine identification problem, on which set will the Decision tree be built: Training or Test?

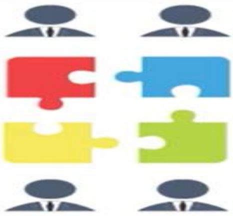
Training

Q2. Name any diagnostic metric which can be used to determine an optimal classification model.

Confusion Matrix, Log loss etc.

5. From Deployment to Feedback

2.1.9 Deployment



How does the solution reach the hands of the user?

Deployment refers to the stage where the trained AI model is made available to the users in real-world applications. Data scientists must make the stakeholders familiar with the tool produced in different scenarios. Once the model is evaluated and the data scientist is confident it will work, it is deployed and put to the ultimate test. Depending on the purpose of the model, it may be rolled out to a limited group of users or in a test environment, to build up confidence in applying the outcome for use across the board.

Deploying a model into a live business process, frequently necessitates the involvement of additional internal teams, skills and technology.

Activity 9:

Q1. Mention some ways to embed the solution into mobiles or websites.

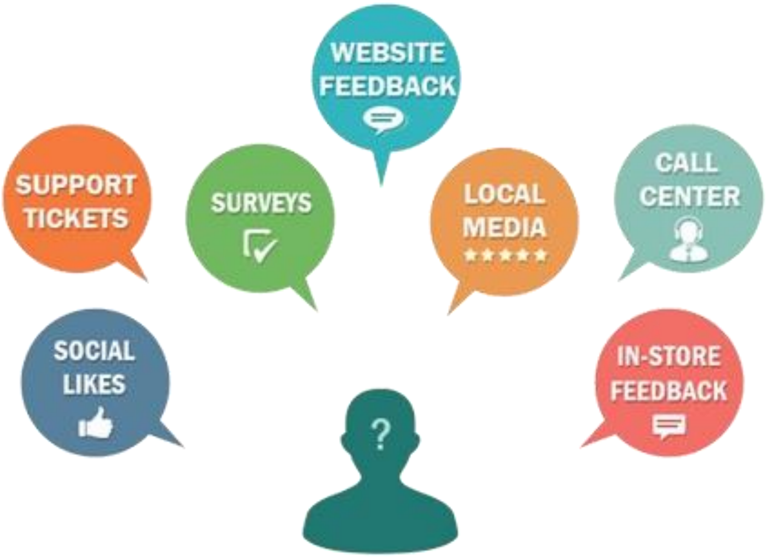
Training model may be downloaded into apk files and this can be integrated into mobile apps (using thunkable) or into websites created by Weebly.

2. 1.10 Feedback

Is the problem solved?
Has the question been satisfactorily answered?

The last stage in the methodology is feedback. This includes results collected from the deployment of the model, feedback on the model's performance from the users and clients, and observations from how the model works in the deployed environment. This process continues till the model provides satisfactory and acceptable results.

Feedback from the users will help to refine the model and assess it for performance and impact. The process from modelling to feedback is highly iterative. Data Scientists may automate any or all of the feedback so that the model refresh process speeds up and can get quick improved results. Feedback from users can be received in many ways.



Throughout the Data Science Methodology, each step sets the stage for the next, making the methodology cyclical and ensuring refinement at each stage.

Teachers can ask the following questions to spark curiosity before starting the topics:

- **Imagine you've trained a self-driving car's AI to recognize stop signs. How would you ensure that it performs well in both clear and foggy conditions? What steps would you take to test its reliability?** (This question introduces the concept of model validation by relating it to a real-world scenario where accuracy and reliability are critical.)

- **If you were developing a system to predict whether an email is spam, how would you measure whether your predictions are accurate? What would you do if your system keeps missing certain spam emails?** (Purpose: This question highlights the importance of evaluation metrics such as precision, recall, and F1-score, encouraging students to think about practical ways to assess model performance.)

2.2. MODEL VALIDATION

Evaluating the performance of a trained machine learning model is essential. Model Validation offers a systematic approach to measure its accuracy and reliability, providing insights into how well it generalizes to new, unseen data.

https://www.geeksforgeeks.org/what-is-model-validation-and-why-is-it-important/

Model validation is the step conducted post Model Training, wherein the effectiveness of the trained model is assessed using a testing dataset. Validating the machine learning model during the training and development stages is crucial for ensuring accurate predictions. The benefits of Model Validation include

- Enhancing the model quality.
- Reduced risk of errors
- Prevents the model from overfitting and underfitting.

Model Validation Techniques

The commonly used Validation techniques are Train-test split, K-Fold Cross Validation, Leave One out Cross Validation, Time Series Cross Validation etc. Let's discuss the Train test split and K-Fold Cross Validation

2.2.1 Train Test Split

The train-test split is a technique for evaluating the performance of a machine learning algorithm. It can be used for classification or regression problems and can be used for any supervised learning algorithm.

The procedure involves taking a dataset and dividing it into two subsets, as shown in Fig 2.4. The first subset is used to fit/train the model and is referred to as the **training dataset**. The second subset is used to test the model. It is with the testing data that predictions are made and compared to the expected values. This second dataset is referred to as the **test dataset**.

Train Dataset: Used to fit the machine learning model.

Test Dataset: Used to evaluate the fit machine learning model.

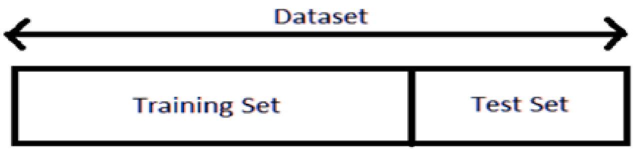


Fig 2.4

The objective is to estimate the performance of the machine learning model on new data, i.e., data not used to train the model. This is how we expect to use the model in practice. The procedure is to fit it on available data with known inputs and outputs, then make predictions on new examples in the future where we do not have the expected output or target values. The train-test procedure is appropriate when there is a sufficiently large dataset available.

How to Configure the Train-Test Split

The procedure has one main configuration parameter, which is the size of the train and test sets. This is most commonly expressed as a percentage between 0 and 1 for either the train or test datasets. For example, a training set with the size of 0.67 (67 percent) means that the remaining percentage 0.33 (33 percent) is assigned to the test set. There is no optimal split percentage. You must choose a split percentage that meets your project’s objectives with considerations that include:

- Computational cost in training the model.
- Computational cost in evaluating the model.
- Representation of the Train set.
- Representation of the Test set.

Nevertheless, common split percentages include:

- Train: 80%, Test: 20%
- Train: 70%, Test: 30%
- Train: 67%, Test: 33%

2.2.2 K-Fold Cross Validation

Cross Validation is a technique used to evaluate a model’s performance. It splits the data into multiple parts or folds. It trains the model on some folds and tests it on other folds and repeats this process for a number fixed by the data scientist. In cross-validation, we run our modeling process on different subsets of the data to get multiple measures of model quality.

k-fold cross validation

In k-fold cross validation we will be working with k subsets of datasets. For example, if we divide the data into 5 folds or 5 pieces, as shown in Fig 2.5, each being 20% of the full dataset, then k=5.

We run an experiment called experiment 1 which uses the first fold as a holdout set (validation), and remaining four folds as training data. This gives us a measure of model quality based on a 20% holdout set. We then run a second experiment, where we hold out data from the second fold. This gives us a second estimate of model quality. We repeat this process, using every fold once as the holdout. Putting this together, 100% of the data is used as a holdout at some point.

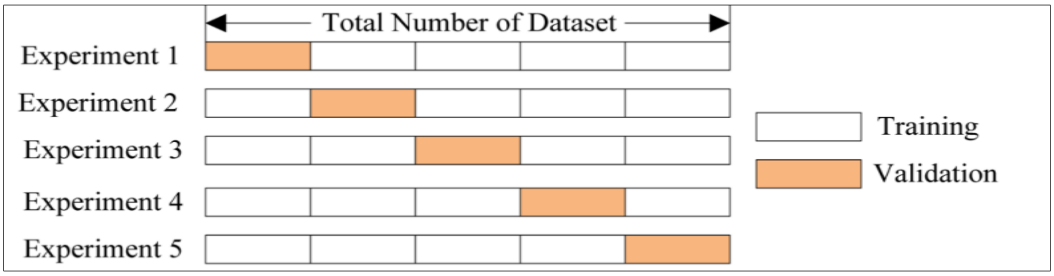


Fig 2.5

Cross-validation gives a more accurate measure of model quality, which is especially important if you are making a lot of modelling decisions. However, it can take more time to run, because it estimates models once for each fold.

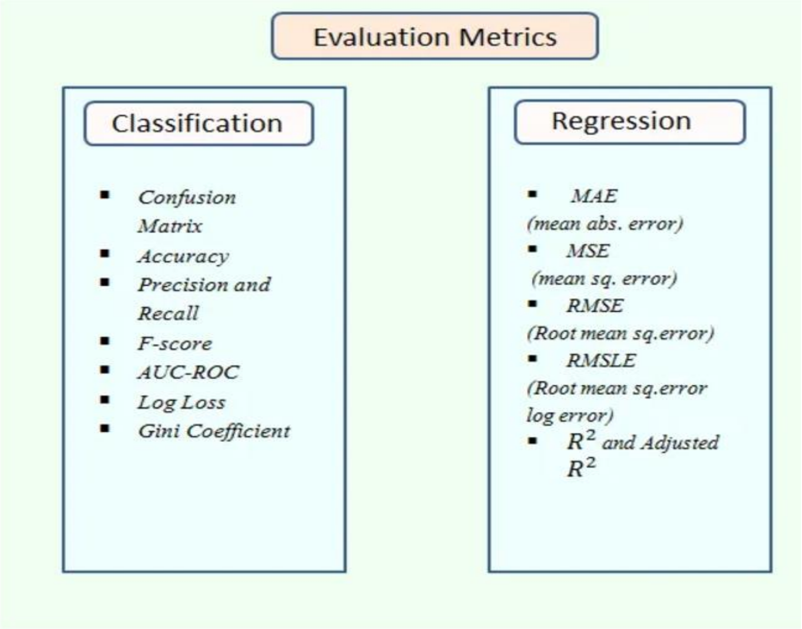
Difference between Train-Test Split and Cross Validation

Train-Test Split	Cross Validation
Normally applied on large datasets	Normally applied on small datasets
Divides the data into training data set and testing dataset.	Divides a dataset into subsets (folds), trains the model on some folds, and evaluates its performance on the remaining data.
Clear demarcation on training data and testing data.	Every data point at some stage could be in either testing or training data set.

2.3. MODEL PERFORMANCE - EVALUATION METRICS

Evaluation metrics help assess the performance of a trained model on a test dataset, providing insights into its strengths and weaknesses. These metrics enable comparison of different models, including variations of the same model, to select the best-performing one for a specific task.

In classification problems, we categorize the target variable into a finite number of classes, while in regression problems, the target variable has continuous values. Hence, we have different evaluation metrics for each type of supervised learning, as depicted in Fig 2.6.



Evaluation metrics for classification and regression models

Fig 2.6
https://medium.com/@ladkarsamisha123/most-popular-machine-learning-performance-metrics-part-1-ab7189dce555

2.3.1 Evaluation Metrics for Classification

1. Confusion Matrix

A Confusion Matrix is a table (Fig 2.7) used to evaluate the performance of a classification model. It summarizes the predictions against the actual outcomes. It creates an N X N matrix, where N is the number of classes or categories that are to be predicted. Suppose there is a problem, which is a binary classification, then N=2 (Yes/No). It will create a 2x2 matrix.

True Positives: It is the case where the model predicted Yes and the real output was also yes.

True Negatives: It is the case where the model predicted No and the real output was also No.

False Positives: It is the case where the model predicted Yes but it was actually No.

False Negatives: It is the case where the model predicted No but it was actually Yes.

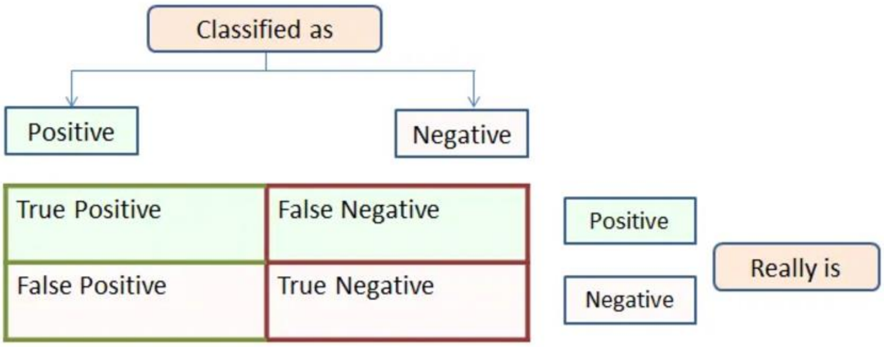


Fig 2.7

2. Precision and Recall

Precision measures “What proportion of predicted Positives is truly Positive?”
Precision = (TP)/(TP+FP).
Precision should be as high as possible.

Recall measures “What proportion of actual Positives is correctly classified?”
Recall = (TP)/(TP+FN)

3. F1-score

A good F1 score means that you have low false positives and low false negatives, so you’re correctly identifying real threats, and you are not disturbed by false alarms. An F1 score is considered perfect when it is 1, while the model is a total failure when it is 0.
F1 = 2* (precision * recall)/(precision + recall)

4. Accuracy

Accuracy = Number of correct predictions / Total number of predictions
Accuracy = (TP+TN)/(TP+FP+FN+TN)

2.3.2 Evaluation Metrics for Regression

1. MAE

Mean Absolute Error is a sum of the absolute differences between predictions and actual values. A value of 0 indicates no error or perfect predictions

2. MSE

Mean Square Error (MSE) is the most commonly used metric to evaluate the performance of a regression model. MSE is the mean(average) of squared distances between our target variable and predicted values.

$$MSE = \frac{\sum_{i=1}^{N} (Predicted_i - Actual_i)^2}{N}$$

3. RMSE

Root Mean Square Error (RMSE) is the standard deviation of the residuals (prediction errors). RMSE is often preferred over MSE because it is easier to interpret since it is in the same units as the target variable.

$$RMSE = \sqrt{\frac{\sum_{i=1}^{N} (Predicted_i - Actual_i)^2}{N}}$$

2.4. PRACTICAL ACTIVITIES

2.4.1. Calculate MSE and RMSE values for the data given below using MS Excel.

Predicted value	14	19	17	13	12	7	24	23	17	18
Actual Value	17	18	18	15	18	11	20	18	13	19

Predicted	Actual	Residual (actual - predicted)	Squared Residuals
14	17	3	9
19	18	-1	1
17	18	1	1
13	15	2	4
12	18	6	36
7	11	4	16
24	20	-4	16
23	18	-5	25
17	13	-4	16
18	19	1	1
		Total	125
		MSE	12.5
		RMSE	3.54

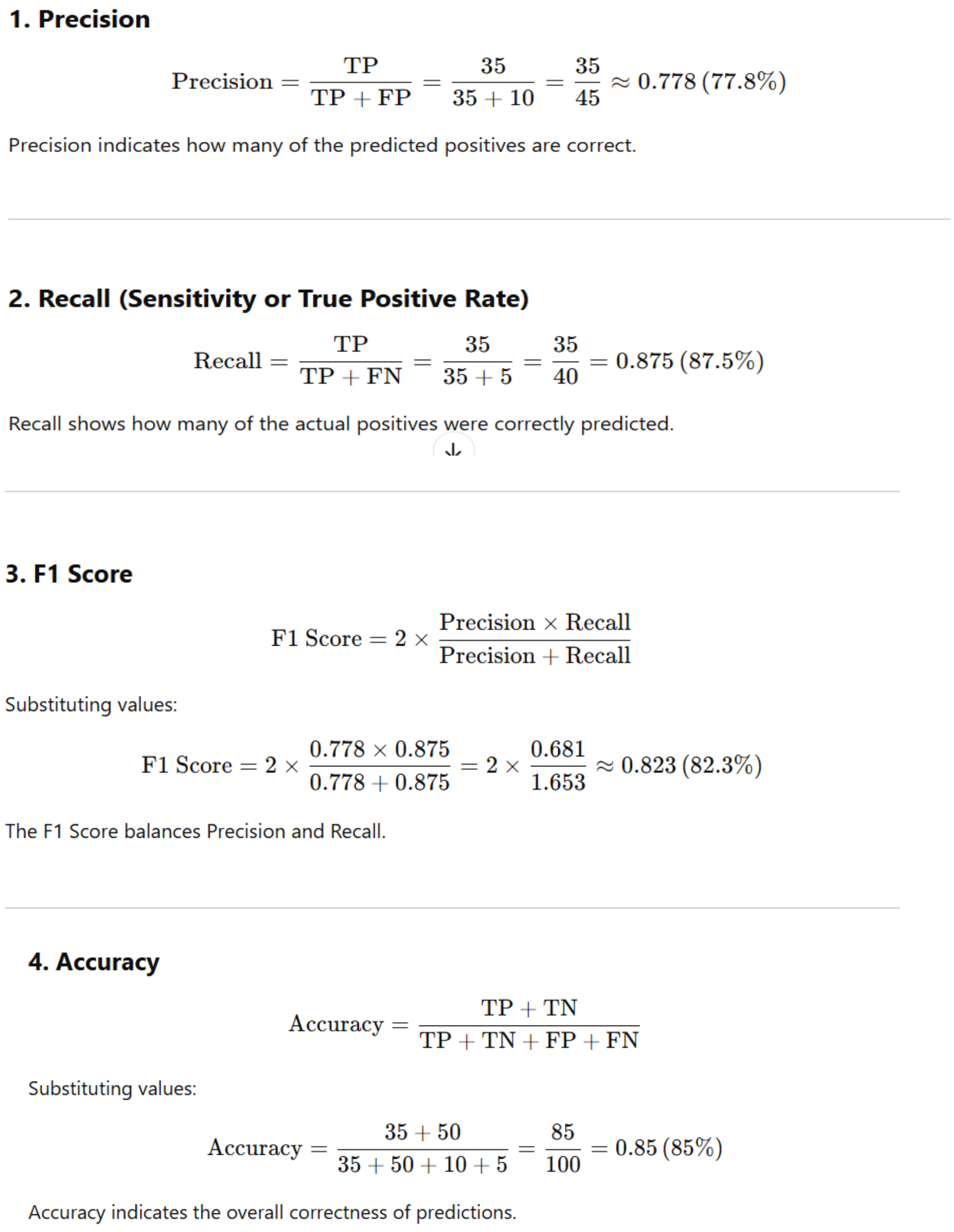
2.4.2. Given a confusion matrix, calculate Precision, Recall, F1 score and Accuracy

Confusion Matrix:

$$\begin{bmatrix} \mathrm{TN} = 50 & \mathrm{FP} = 10 \\ \mathrm{FN} = 5 & \mathrm{TP} = 35 \end{bmatrix}$$

From the confusion matrix:

True Positives (TP) = 35, True Negatives (TN) = 50, False Positives (FP) = 10, False Negatives (FN) = 5



1. Precision

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}} = \frac{35}{35 + 10} = \frac{35}{45} \approx 0.778\,(77.8\%)$$

Precision indicates how many of the predicted positives are correct.

2. Recall (Sensitivity or True Positive Rate)

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}} = \frac{35}{35 + 5} = \frac{35}{40} = 0.875\,(87.5\%)$$

Recall shows how many of the actual positives were correctly predicted.

3. F1 Score

$$\text{F1 Score} = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}$$

Substituting values:

$$\text{F1 Score} = 2 \times \frac{0.778 \times 0.875}{0.778 + 0.875} = 2 \times \frac{0.681}{1.653} \approx 0.823\,(82.3\%)$$

The F1 Score balances Precision and Recall.

4. Accuracy

$$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{TN} + \text{FP} + \text{FN}}$$

Substituting values:

$$\text{Accuracy} = \frac{35 + 50}{35 + 50 + 10 + 5} = \frac{85}{100} = 0.85\,(85\%)$$

Accuracy indicates the overall correctness of predictions.

Summary of Metrics:

Precision: 77.8%

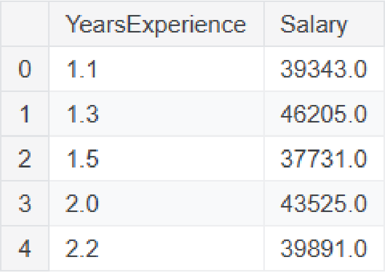
Recall: 87.5%

F1 Score: 82.3%

Accuracy: 85%

2.4.3. Python Code to Evaluate a Model

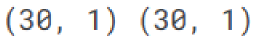
```
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
import seaborn as sns
from sklearn.metrics import mean_squared_error
from sklearn.linear_model import LinearRegression
from sklearn.model_selection import train_test_split

df=pd.read_csv('/kaggle/input/random-salary-data-of-employes-age-
wise/Salary_Data.csv',sep=',')
df.head()
```



	YearsExperience	Salary
0	1.1	39343.0
1	1.3	46205.0
2	1.5	37731.0
3	2.0	43525.0
4	2.2	39891.0

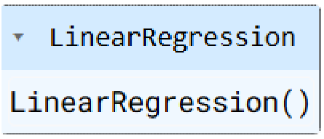
```
df.shape
(30, 2)

df.isnull().sum()
YearsExperience    0
Salary             0
dtype: int64
```

```
#Data Preparation
X=np.array(df['YearsExperience']).reshape(-1,1)
Y=np.array(df['Salary']).reshape(-1,1)
print(X.shape,Y.shape)
```



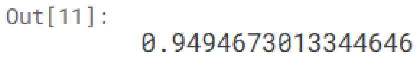
```
(30, 1) (30, 1)
```

```
#Configuring the train-test split
X_train,x_test,Y_train,y_test=train_test_split(X,Y,test_size=0.2,shuffle=True,random_state=10)

#Fitting the model
model=LinearRegression()
model.fit(X_train,Y_train)
```



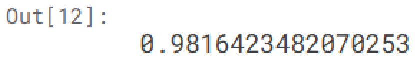
```
▾ LinearRegression
LinearRegression()
```

```
#Evaluation the score of training data
model.score(X_train,Y_train)
```



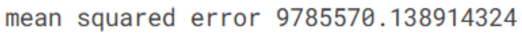
```
Out[11]:
        0.9494673013344646
```

```
#Evaluation the score of testing data
model.score(x_test,y_test)
```



```
Out[12]:
        0.9816423482070253
```

```
#Evaluating Mean Squared error
y_pred=model.predict(x_test)
print("mean squared error",mean_squared_error(y_test,y_pred))
```



```
mean squared error 9785570.138914324
```

Base Reference: https://www.kaggle.com/code/saiteja180/salary-estimation

EXERCISES

A. Objective type questions

1. Which is the hardest stage in the foundational methodology of Data Science?

a. Business Understanding b. Data collection

c. Modelling d. Evaluation

2. Business Sponsors defines the problem and project objectives from a ________ perspective.

a. Economic b. Feedback c. Business d. Data Collection

3. Match the following and choose the correct options:

i. Descriptive approach	A. Statistical Analysis
ii. Diagnostic approach	B. Current Status
iii. Predictive approach	C. How to solve it?
iv. Prescriptive approach	D. Probabilities of action

a. (i)—A , (ii)—B, (iii) – C , (iv)—D b. (i)—B , (ii)—A, (iii) – D , (iv)—C

c. (i)—D , (ii)—B, (iii) – A , (iv)—C d. (i)—A , (ii)—C, (iii) – B , (iv)—D

4. Arrange the following statements in order

i: Gaps in data will be identified and plans to fill/make substitutions will have to be made

ii: Decisions are made whether the collection requires more data or not

iii: Descriptive statistics and visualization is applied to dataset

iv: Identify the necessary data content, formats and sources

a. i,ii,iii,iv b. iv,ii,iii,i c. i,iii,ii,iv d. ii,i,iii,iv

5. Data Modelling focuses on developing models that are either ___________ or ____________

a. Supervised, Unsupervised b. Predictive, Descriptive

c. Classification, Regression d. Train-test split, Cross Validation

6. Statement 1- There is no optimal split percentage

Statement 2- The most common split percentage between training and testing data is 20%-80%

a. Statement 1 is true Statement 2 is false

b. Statement 2 is true Statement 1 is false

c. Both Statement 1 and 2 are true

d. Both Statement 1 and 2 are false

7.Train-test split function is imported from which Python module?

a. sklearn.model_selection b. sklearn.ensemble

c. sklearn.metrics d. sklearn. preprocessing

8. Identify the incorrect statement:

i. cross-validation gives a more reliable measure of your model's quality

ii. cross-validation takes short time to run

iii. cross-validation gets multiple measures of model’s quality

iv. cross-validation is preferred with small data

a. ii and iii b. iii only c. ii only d. ii, iii and iv

9. Identifying the necessary data content, formats and sources for initial data collection is done in which step of Data Science methodology?

a. Data requirements   b. Data Collection   c. Data Understanding   d. Data Preparation

10. Data sets are available online. From the given options, which one does not provide online data?

a. UNICEF   b.WHO   c. Google   d. Edge

11. A ____________ set is a set of historical data in which outcomes are already known.

a. Training set   b. Test set   c. Validation set   d. Evaluation set

12. _____________ data set is used to evaluate the fit machine learning model.

a. Training set   b. Test set   c. Validation set   d. Evaluation set

13. x_train,x_test,y_train,y_test = train_test_split (x, y, test_size=0.2)

From the above line of code, identify the training data set size

a. 0.2   b. 0.8   c. 20   d. 80

14. In k-fold cross validation, what does k represent?

a. number of subsets   b. number of experiments   c. number of folds   d. all of the above

15. Identify the correct points regarding MSE given below:

i. MSE is expanded as Median Squared Error

ii. MSE is standard deviation of the residuals

iii. MSE is preferred with regression

iv. MSE penalize large errors more than small errors

a. i and ii   b. ii and iii   c. iii and iv   d. ii, iii and iv

B. Short Answer Questions

1. How many steps are there in Data Science Methodology? Name them in order.

Ans- There are 10 steps in Data Science Methodology. They are Business Understanding, Analytic Approach, Data Requirements, Data Collection, Data Understanding, Data Preparation, Modelling, Evaluation, Deployment and Feedback

2. What do you mean by Feature Engineering?

Ans- Feature Engineering is the process of using domain knowledge of data to create features(variables) that make the machine learning algorithms work.

3. Data is collected from different sources. Explain the different types of sources with example.

Ans - Data can be collected from two sources—Primary data source and Secondary data source Primary Sources are sources which are created to collect the data for analysis. Examples include Interviews, Surveys, Marketing Campaigns, Feedback Forms, IOT sensor data etc.,

Secondary data is the data which is already stored and ready for use. Data given in Books, journals, Websites, Internal transactional databases, etc. are some examples

4. Which step of Data Science Methodology is related to constructing the data set? Explain.

Ans- Data Understanding stage is related to constructing the data set. Here we check whether the data collected represents the problem to be solved or not. Here we evaluate whether the data is relevant, comprehensive, and suitable for addressing the specific problem or question at hand. Techniques such as descriptive statistics and visualization can be applied to the dataset, to assess the content, quality, and initial insights about the data.

5. Write a short note on the steps done during Data Preparation.

Ans- The most time-consuming stage is Data Preparation. Here data is transformed into a state where it is easier to work with. Feature Engineering is also a part of Data Preparation.

Data preparation includes

- cleaning of data (dealing with invalid or missing values, remove duplicates and give a suitable format)
- combine data from multiple sources (archives, tables and platforms)
- transform data into meaningful input variables

6. Differentiate between descriptive modelling and predictive modelling.

Ans- Descriptive modelling and Predictive modelling are based on the analytic approach that was taken, either statistically driven or machine learning driven.

Descriptive modeling is a mathematical process that describes real-world events and the relationships between factors responsible for them. An example of a descriptive model might examine things like: if a person did this, then they are likely to prefer that.

Predictive modeling is a process that uses data mining and probability to forecast outcomes. For example, A predictive model tries to yield yes/no, or stop/go type outcomes. The data scientist will use a training set for predictive modeling.

7. Explain the different metrics used for evaluating Classification models.

1. Confusion Matrix

A Confusion Matrix is a table used to evaluate the performance of a classification model. It summarizes the predictions against the actual outcomes

2. Precision and Recall

Precision measures “What proportion of predicted Positives is truly Positive?”

Precision = (TP)/(TP+FP).

Precision should be as high as possible.

Recall measures “What proportion of actual Positives is correctly classified?”

Recall = (TP)/(TP+FN)

3. F1-score

A good F1 score means that you have low false positives and low false negatives, so you’re correctly identifying real threats, and you are not disturbed by false alarms. An F1 score is considered perfect when it is 1, while the model is a total failure when it is 0.

F1 = 2* (precision * recall)/(precision + recall)

4. Accuracy
Accuracy = Number of correct predictions / Total number of predictions
Accuracy = (TP+TN)/(TP+FP+FN+TN)

8. Is Feedback a necessary step in Data Science Methodology? Justify your answer.

Ans - Yes, Feedback is necessary. Feedback from the users will help to refine the model and assess it for performance and impact. Data Scientists can automate any or all of the feedback so that the model refresh process speeds up and gets quick improved results. The value of the model will be dependent on successfully incorporating feedback and adjusting for as long as the solution is required

9. Write a comparative study on train-test split and cross validation.

Train-Test Split	**Cross Validation**
Normally applied on large data sets	Normally applied on small data sets
Divides the data into training data set and testing data set.	Divides a dataset into subsets (folds), trains the model on some folds, and evaluates its performance on the remaining data.
Gives the accuracy of the model of the validation data set.	Gives the average accuracy across each validation set.
Clear demarcation on training data and testing data.	Every data point at some stage could be in either testing or training data set.

10. Why is model validation important?

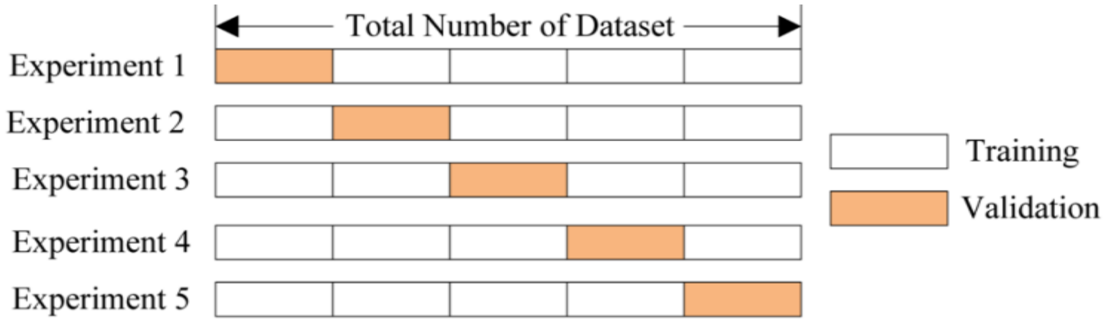
Model Validation offers a systematic approach to measure its accuracy and reliability, providing insights into how well it generalizes to new, unseen data. The benefits of Model Validation include

- Enhancing the model quality.
- Reduced risk of errors
- Prevents the model from overfitting and underfitting.

C. Long Answer Questions

Ans- In k-fold cross validation we will be working with k subsets of datasets. For example, if we could have 5 folds or experiments (here k=5), we divide the data into 5 pieces, each being 20% of the full dataset.

We run an experiment called experiment 1 which uses the first fold as a holdout set, and everything else as training data. This gives us a measure of model quality based on a 20% holdout set. We then run a second experiment, where we hold out data from the second fold (using everything except the 2nd fold for training the model.) This gives us a second estimate of model quality. We repeat this process, using every fold once as the holdout. Putting this together, 100% of the data is used as a holdout at some point.



2. Data is the main part of any project. How will you find the requirements of data, collect it, understand the data and prepare it for modelling?

Ans - For any Model data is made ready with four steps.

1. Data Requirements- In the data requirements stage we should identify the necessary data content, formats, and sources for initial data collection. 5W1H questions may be employed. Here we identifying the types of data required, decides how to store the data considering the structure in which the data should be organized, whether it is in a table, text file, or database. We will be identifying the sources from which we can collect the data and also any necessary cleaning or organization steps required are done.

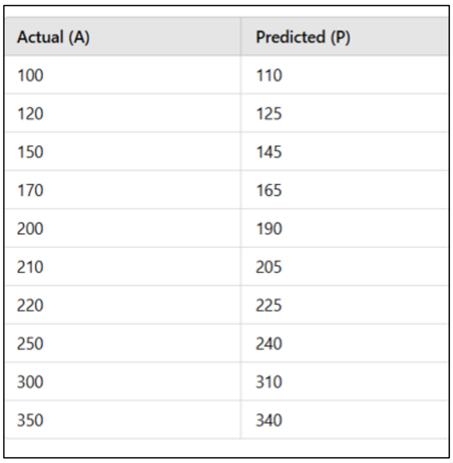
2. Data Collection-Data collection is a systematic process of gathering observations or measurements. In this phase the data requirements are revised and decisions are made as to whether the collection requires more or less data. Today's high performance database analytics enable data scientists to utilize large datasets. Data can be collected from Primary data source (Survey, Interview etc.) or Secondary data source (Social media data tracking, web scraping etc)

3. Data Understanding- Data Understanding stage is related to constructing the data set. Here we check whether the data collected represents the problem to be solved or not. Here we evaluate whether the data is relevant, comprehensive, and suitable for addressing the specific problem or question at hand. Techniques such as descriptive statistics and visualization can be applied to the dataset, to assess the content, quality, and initial insights about the data.

4. Data Preparation- The most time-consuming stage is Data Preparation. Here data is transformed into a state where it is easier to work with. Data preparation includes cleaning of data, combining data from multiple sources and transform data into meaningful input variables. Feature Engineering is also a part of Data Preparation.

D. Case study

1. Calculate MSE and RMSE values for the data given below using MS Excel.



Actual (A)	Predicted (P)
100	110
120	125
150	145
170	165
200	190
210	205
220	225
250	240
300	310
350	340

Steps in Excel:

1. Enter the Data

In Column A, input the actual values.

In Column B, input the predicted values.

Fill in the data under the respective columns.

2. Calculate Squared Errors

Label Column C as Squared Error.

In Cell C2, enter the formula: =$(A2-B2)2$

Drag this formula down to fill all rows (C2:C11). This calculates the squared error for each row.

3. Calculate the Mean Squared Error (MSE)

In an empty cell (e.g., D1), label it MSE.

In D2, use the formula: =$AVERAGE(C2{:}C11)$

This computes the average of all squared errors.

4.Calculate the Root Mean Squared Error (RMSE)

In an empty cell (e.g., E1), label it RMSE.

In E2, use the formula: =$SQRT(D2)$

This takes the square root of the MSE to compute the RMSE.

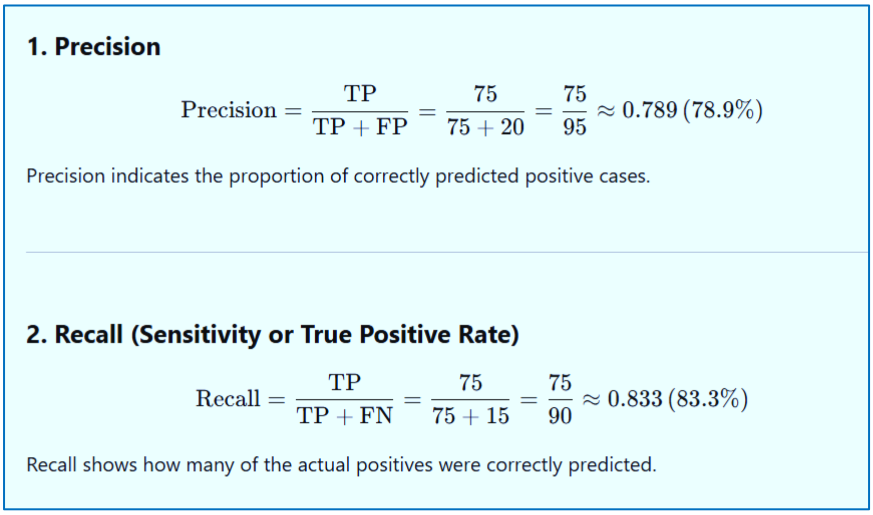
For the given data:

MSE = 58.0

RMSE = 7.62

2. Given a confusion matrix, calculate Precision, Recall, F1 score and Accuracy.

$$\begin{bmatrix} \text{TN} = 90 & \text{FP} = 20 \\ \text{FN} = 15 & \text{TP} = 75 \end{bmatrix}$$



1. Precision

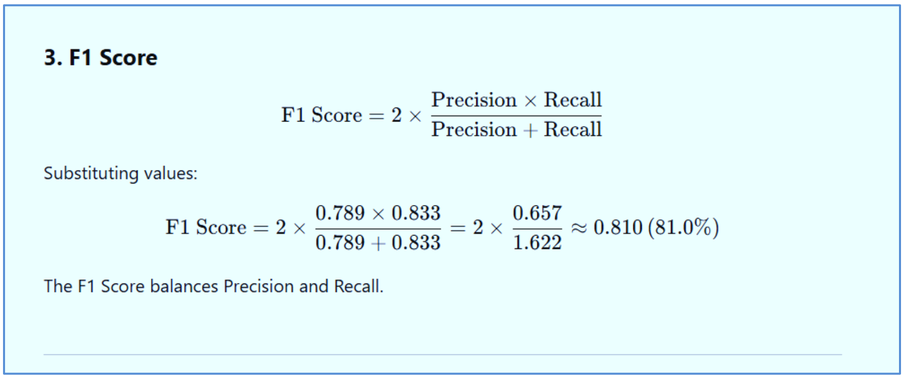
$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}} = \frac{75}{75 + 20} = \frac{75}{95} \approx 0.789\,(78.9\%)$$

Precision indicates the proportion of correctly predicted positive cases.

2. Recall (Sensitivity or True Positive Rate)

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}} = \frac{75}{75 + 15} = \frac{75}{90} \approx 0.833\,(83.3\%)$$

Recall shows how many of the actual positives were correctly predicted.



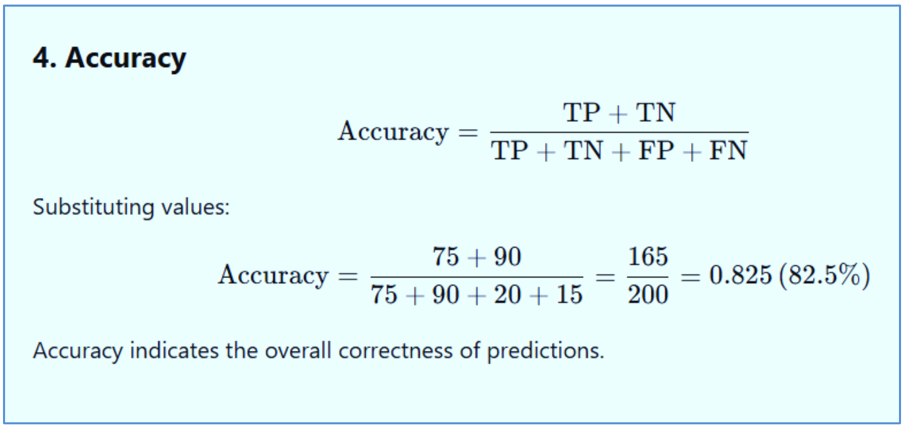
3. F1 Score

$$\text{F1 Score} = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}$$

Substituting values:

$$\text{F1 Score} = 2 \times \frac{0.789 \times 0.833}{0.789 + 0.833} = 2 \times \frac{0.657}{1.622} \approx 0.810\,(81.0\%)$$

The F1 Score balances Precision and Recall.



4. Accuracy

$$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{TN} + \text{FP} + \text{FN}}$$

Substituting values:

$$\text{Accuracy} = \frac{75 + 90}{75 + 90 + 20 + 15} = \frac{165}{200} = 0.825\,(82.5\%)$$

Accuracy indicates the overall correctness of predictions.

Summary of Metrics:

Precision: 78.9%

Recall: 83.3%

F1 Score: 81.0%

Accuracy: 82.5%

E. Competency Based Questions

1.A transportation company aims to optimize its delivery routes and schedules to minimize costs and improve delivery efficiency. The company wants to use Data Science to identify the most optimal routes and delivery time windows based on historical delivery data and external factors such as traffic and weather conditions. Various questions are targeted by data scientist to achieve this business goal. Identify the analytical approach model that can be used for each.

a) determine the most suitable delivery routes for perishable goods, ensuring timely deliveries without explicitly using past data to make predictions.

b) gather insights on the average delivery times for different vehicle types, how they vary based on the complexity of delivery route.

c) group delivery routes into different categories based on the average delivery time and order volume.

Ans. a. Predictive model

b. Descriptive model

c. Classification model

2. A leading investment firm aims to improve their client portfolio management system. They want to know whether Artificial Intelligence could be used to better understand clients' investment preferences and risk tolerance levels. Which stage of Data Science methodology can you relate this to?

Ans. Business Understanding

3. An Online Learning Platform has implemented a recommendation system to suggest personalized courses to users. They need to assess the effectiveness and accuracy of this system. Which stage of Data Science methodology can you relate this to?

Ans. Evaluation

4. A data scientist working to improve public transportation services by analyzing commuter travel patterns. He has encountered a scenario where he needs to understand the impact of major events on commuter behavior. For instance, the city is hosting a large-scale sporting event, and the data scientist needs to assess how this event affects commuting patterns, such as changes in peak travel times, shifts in preferred modes of transportation, and alterations in popular routes. Which stage of Data Science methodology is he in? List the steps he needs to follow.

Ans: The data scientist in this scenario is in the stage of Data Collection within the Data Science methodology. To address the scenario effectively, the data scientist should:

- Identify Relevant Data Sources
- Gather Data
- Clean and Prepare Data
- Analyze Data
- Interpret Results

5.A data scientist is tasked with developing a machine learning model to predict customer churn for a small e-commerce startup. A limited dataset is only available for this task. The dataset contains information about customer demographics, purchase history, website interactions, and whether they churned or not. Considering the challenge posed by the limited dataset size, which approach would you recommend the data scientist to use for training the churn prediction model

- a simple train-test split or cross-validation? Justify your recommendation regarding the dataset's size and generalizability.

Ans: Considering the limited dataset size and the need for robustness and generalizability in the model, I would recommend using cross-validation approach. Cross-validation involves splitting the dataset into multiple subsets (folds), training the model on different combinations of these subsets, and evaluating its performance on the remaining data. This mitigates the risk of overfitting or underfitting the model, which is common with a small dataset. Additionally, cross-validation maximizes the utilization of limited data by using each data point for both training and validation across multiple folds, eliminating the need for additional data.

6.Identify the type of big data analytics (descriptive, predictive) used in the following:

a. A clothing brand monitors social media mentions to understand customer perception. It uses data in the form of social media posts, comments, and reviews containing brand mentions to get a clear picture of overall customer sentiment and areas where they excel or fall short.
b. A factory aims to predict equipment failures before they occur to minimize downtime. It uses Sensor data from machines (temperature, vibration, power consumption) coupled with historical maintenance records to identify patterns in sensor data that indicate an impending equipment failure.

Ans:

a. Descriptive Analytics. This involves summarizing and analyzing historical social media data to understand customer sentiment and perception towards the clothing brand.
b. Predictive Analytics. Specifically, it involves using historical sensor data from machines to predict equipment failures before they occur, minimizing downtime.

7.Identify the type of big data analytics (diagnostic, prescriptive) used in the following:

a. A subscription service experiences a rise in customer cancellations. It uses Customer account information, usage data (frequency of logins, features used), and support ticket logs.to identify potential reasons for churn.
b. A food delivery service wants to improve delivery efficiency and reduce delivery times. It uses Customer location data, order details, historical delivery times, and traffic patterns to calculate the most efficient delivery routes.

Ans:

a. Diagnostic Analytics. This involves analyzing customer account information, usage data, and support ticket logs to diagnose potential reasons for customer cancellations or churn.
b. Prescriptive Analytics. This involves analyzing customer location data, order details, historical delivery times, and traffic patterns to prescribe the most efficient delivery routes and reduce delivery times.

RESOURCES

Courses in Data Science Methodology:

1. https://www.coursera.org/learn/data-science-methodology#modules
2. https://cognitiveclass.ai/courses/data-science-methodology-2

UNIT 3: Making Machines See

Title: Making Machines See	**Approach:** Team Discussion, Web search**,** Hands-on activities

Summary:

Computer vision has become a cornerstone technology in today's digital era, enabling machines to "see" and interpret visual data much like humans. This lesson delves into the fascinating world of computer vision, exploring its fundamental principles, key processes, real-world applications, and future potential.

Learning Objectives:

1. Understand the fundamentals of computer vision and its role in processing and analysing digital images and videos.
2. Explore the various stages involved in the computer vision process.
3. Gain insight into the applications of computer vision across different industries.
4. Identify the challenges and ethical considerations associated with computer vision technology, including privacy concerns, data security, and misinformation.
5. Recognize the future potential of computer vision technology and its impact on society.

Key Concepts:

1. Introduction to Computer Vision
2. Working of Computer Vision
3. Applications of Computer Vision
4. Challenges of Computer Vision
5. The Future of Computer Vision

Learning Outcomes:

Students will be able to -

1. Explain the concept of computer vision and its significance in analysing visual data.
2. Demonstrate an understanding of the key stages involved in computer vision process and their respective roles in interpreting images and videos.
3. Identify real-world applications of computer vision technology in various industries and understand how it enhances efficiency and productivity.
4. Evaluate the ethical implications and challenges associated with computer vision, including privacy concerns and the spread of misinformation.
5. Envision the future possibilities of computer vision technology.

Prerequisites:

Basic understanding of digital imaging concepts, and knowledge of machine learning.

Unveiling the Magic of Computer Vision: A Teacher's Guide to Making Machines See

This lesson equips you to introduce students to the captivating world of Computer Vision (CV), where machines learn to "see" and interpret the visual world around them.

1. Beyond the Screen: How Computers See:

- Captivating Introduction: Begin by sparking curiosity with a thought-provoking question: "Can computers, see?" Introduce the concept of Computer Vision, where machines learn to process and understand visual information from images and videos.

2. The Building Blocks of Machine Vision:

- Demystifying CV Basics: Introduce the fundamental concepts of CV:
  - Image Recognition: Identifying objects or scenes within an image.
  - Neural Networks: Machine learning models inspired by the human brain, playing a crucial role in CV tasks.
  - Feature Extraction: Identifying and extracting key characteristics from images (e.g., shapes, edges, colors).
  - Models: Trained algorithms that can make predictions based on image data.
  - Evaluation: Measuring the performance of CV models and identifying areas for improvement.

3. Real-World Applications: A Glimpse into the Future:

- Bringing CV to Life: Showcase the diverse applications of CV:
  - Object Detection: Identifying and locating objects in images and videos (e.g., self-driving cars detecting pedestrians).
  - Image Classification: Categorizing images into pre-defined classes (e.g., classifying products in e-commerce).
  - Facial Recognition: Identifying individuals based on facial features (e.g., unlocking smartphones).
  - Medical Image Analysis: Assisting doctors in analysing medical scans (e.g., detecting abnormalities in X-rays).
  - Autonomous Vehicles: Enabling vehicles to navigate roads by understanding their surroundings.
  - Augmented Reality: Superimposing digital information on the real world (e.g., overlaying navigation instructions).
  - Quality Control Applications: Automating visual inspection in manufacturing processes.

4. Seeing the Challenges: Obstacles and Considerations:

- Understanding the Roadblocks: Discuss the challenges in CV:
  - Data Quality: Reliance on large amounts of high-quality data for training models.
  - Interpretability: Difficulty in understanding how models reach decisions (often a "black box").
  - Ethical Concerns: Bias in data sets and potential misuse of facial recognition technology.
  - Real-time Processing Constraints: Balancing accuracy with speed for real-world applications.

5. A Glimpse into the Future:

- Emerging Trends: Discuss upcoming advancements in CV:
  - Improved Accuracy: Models with greater ability to recognize objects and understand complex scenes.
  - Real-time Processing: Faster processing times for real-world applications with minimal latency.
  - Integration with AI: CV working seamlessly with other AI functionalities.
  - Enhanced Applications: Greater involvement of CV in robotics, healthcare, and security sectors.

6. Hands-on Learning with Teachable Machine:

- Learning by Doing: Introduce Teachable Machine, a beginner-friendly platform for creating basic machine learning models without extensive coding. This allows students to experiment with image classification and experience model deployment.

Additional Tips:

- Utilize engaging visuals and interactive activities throughout the lesson to enhance student understanding.
- Encourage students to explore real-world examples of CV applications through online resources and demos.
- Provide opportunities for students to discuss the ethical implications and societal impact of CV.

By incorporating these elements, you can ignite student interest in Computer Vision, empower them to explore its potential, and encourage them to contribute responsibly to its future development.

Teachers can ask the following questions to spark curiosity before starting the topics:

- **Imagine you take a picture with a digital camera. How do you think the camera captures the image you see in front of you? What do you think happens to the picture after you take it?** (This question activates students' prior knowledge about how cameras capture visual information digitally and connects it to the concept of digital images being made up of pixels.)
- **Can you think of any examples of how we describe visual information? We might say something is red or shaped like a circle. How do you think computers might understand these visual details in an image?** (This question primes students to consider the challenges of translating visual information into a format computer can understand, which is what computer vision aims to achieve.)

With the rapid expansion of social media platforms such as Facebook, Instagram, and Twitter, smartphones have emerged as pivotal tools, thanks to their integrated cameras facilitating effortless sharing of photos and videos. While the Internet predominantly consists of text-based content, indexing and searching images present a distinct challenge. Indexing and searching images involve organizing image data for quick retrieval based on specific features like colour, texture, shape, or metadata. During indexing, key attributes are extracted and stored in a searchable format. Searching uses this index to match query parameters with stored image features, enabling efficient retrieval. Unlike text, which can be easily processed, algorithms require additional capabilities to interpret image content.

Traditionally, the information conveyed by images and videos has relied heavily on manually provided meta descriptions. To overcome this limitation, there is a growing need for computer systems to visually perceive and comprehend images to extract meaningful information from them. This involves enabling computers to "see" images and decipher their content, thereby bridging the gap in understanding and indexing visual data. This poses a simple challenge for humans, evident in the common practice of teaching children to associate an image, such as an apple, with the letter 'A'.

Humans can easily make this connection. However, enabling computers to comprehend images presents a different dilemma. Similarly to how children learn by repeatedly viewing images to memorize objects or people, we need computers to develop similar capabilities to effectively analyse our images and videos.

3.1. HOW MACHINES SEE?

Computer Vision, commonly referred to as CV, enables systems to see, observe, and understand. Computer Vision is similar to human vision as it trains machines with cameras, data, and algorithms similar to retinas, optic nerves, and a visual cortex as in human vision. CV derives meaningful information from digital images, videos and other visual input and makes recommendations or takes actions accordingly.

Computer Vision systems are trained to inspect products, watch infrastructure, or a production asset to analyse thousands of products or processes in real-time, noticing defects or issues. Due to its speed, objectivity, continuity, accuracy, and scalability, it can quickly surpass human capabilities. The latest deep learning models achieve above human-level accuracy and performance in real-world image recognition tasks such as facial recognition, object detection, and image classification.

Computer Vision is a field of artificial intelligence (AI) that uses Sensing devices and deep learning models to help systems understand and interpret the visual world.

Computer Vision is sometimes called Machine Vision.

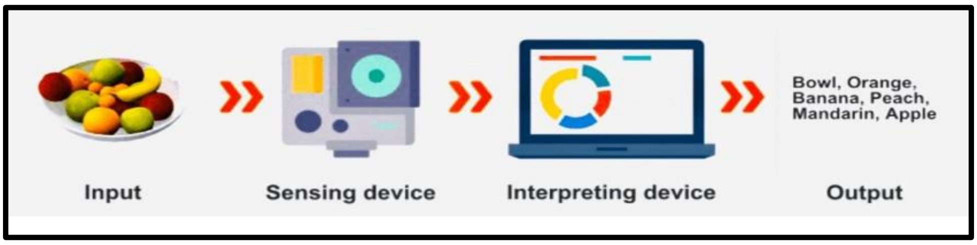


Fig.3.1: Process flow of computer vision

image source: https://www.ciopages.com/wp-content/uploads/2020/07/vision-work.jpg

3.2. WORKING OF COMPUTER VISION

At its core, computer vision is the field of study that focuses on processing and analysing digital images and videos to comprehend their content. A fundamental aspect of computer vision lies in understanding the basics of digital images.

3.2.1. Basics of digital images

A digital image is a picture that is stored on a computer in the form of a sequence of numbers that computers can understand. Digital images can be created in several ways like using design software (like Paint or Photoshop), taking one on a digital camera, or scan one using a scanner.

3.2.2. Interpretation of Image in digital form

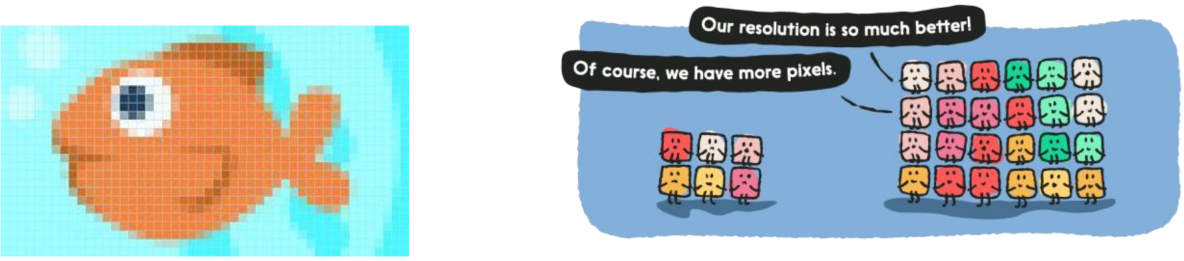
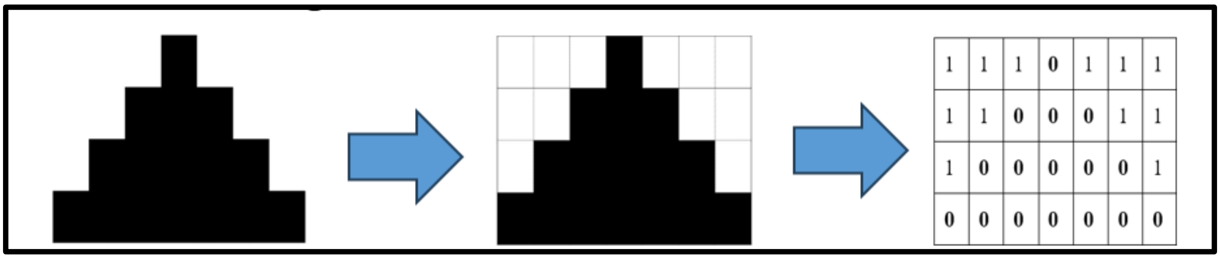

Our resolution is so much better!
Of course, we have more pixels.

Fig.3.2: How pixel affects the image

When a computer processes an image, it perceives it as a collection of tiny squares known as *pixels*. Each pixel, short for "picture element," represents a specific color value. These pixels collectively form the digital image. During the process of digitization, an image is converted into a grid of pixels. The resolution of the image is determined by the number of pixels it contains; the higher the resolution, the more detailed the image appears and the closer it resembles the original scene.



1	1	1	0	1	1	1
1	1	0	0	0	1	1
1	0	0	0	0	0	1
0	0	0	0	0	0	0

Fig. 3.3: Representing an image using 0's and 1's

In representing images digitally, each pixel is assigned a numerical value. For monochrome images, such as black and white photographs, a pixel's value typically ranges from 0 to 255. *A value of 0 corresponds to black, while 255 represents white*.

ACTIVITY 3.1 - Binary Art: Recreating Images with 0s and 1s

Step 1: Choose an Image

- Select any image to work with.
- You can find free images on open-source websites like: Pixabay, Unsplash,pexels, etc.

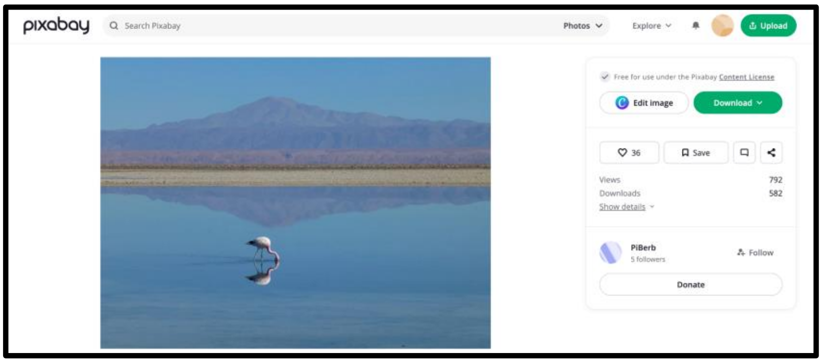
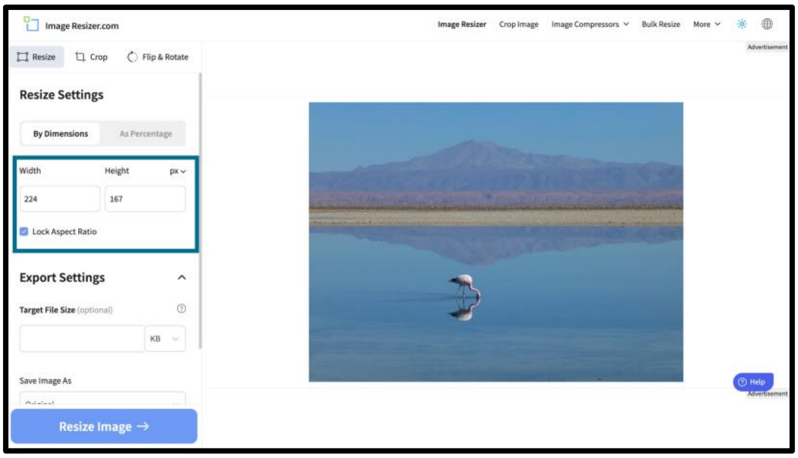


Fig. 3.4: choose an image

Step 2: Resize the Image

- To simplify the activity, resize the image to smaller dimensions (recommended size: width and height between 200 to 300 pixels).
- Use any online resizing tool, such as: Image resizer - https://imageresizer.com/
- Ensure the resized image is saved to your computer.



Resize Settings

By Dimensions

Width 224

Height 167

px

☑ Lock Aspect Ratio

Fig. 3.5: resize of image

Step 3: Convert to Grayscale

- Transform the image into grayscale so it contains only shades of gray (1 channel).
- Use an online grayscale converter, such as Pine tools- https://pinetools.com/grayscale-image
- Upload your resized image, convert it to grayscale, and download the resulting image.

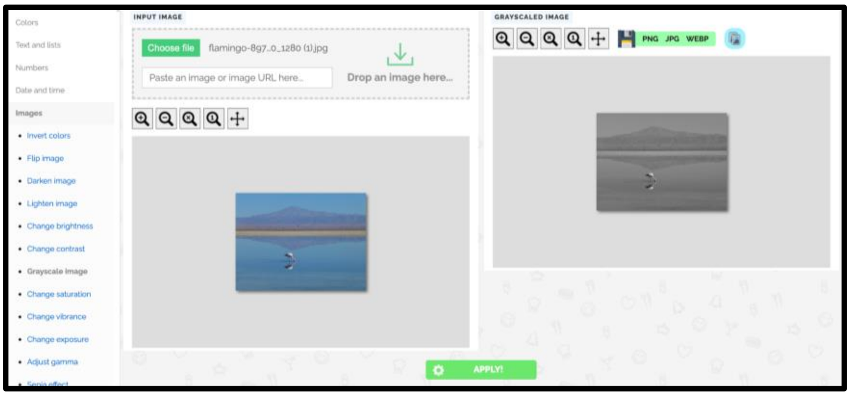
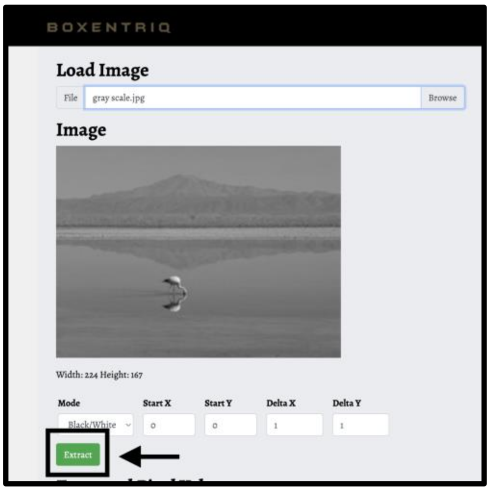


Fig. 3.6: grayscale conversion

Step 4: Extract Pixel Values

- The grayscale image needs to be converted into numerical pixel values (e.g., 0 and 1 for black and white tones).
- Use a pixel value extractor tool, such as: Boxentriq Pixel Value Extractor - https://www.boxentriq.com/code-breaking/pixel-values-extractor
- Upload your grayscale image and press the "Extract" button to generate pixel values.



gray scale.jpg

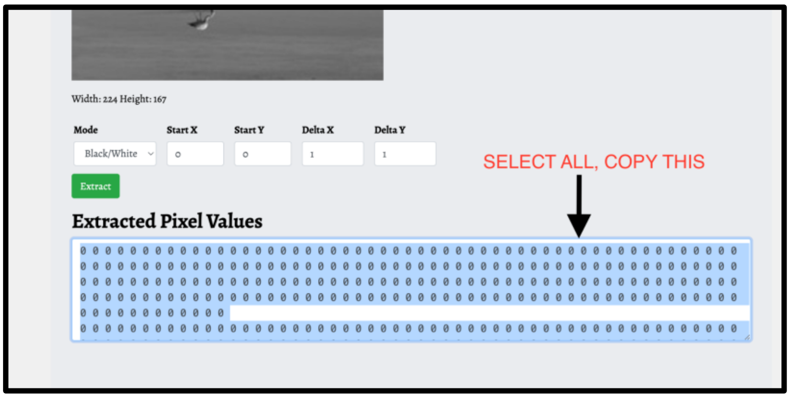
Width: 224 Height: 167

Mode: Black/White
Start X: 0
Start Y: 0
Delta X: 1
Delta Y: 1

Fig. 3.7

Step 5: Copy the Pixel Values

- Once the pixel values are extracted, select all the values from the tool and copy them.



Width: 224 Height: 167

Mode: Black/White, Start X: 0, Start Y: 0, Delta X: 1, Delta Y: 1

Extracted Pixel Values

```
0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
0 0 0 0 0 0 0 0 0 0 0 0
0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
```

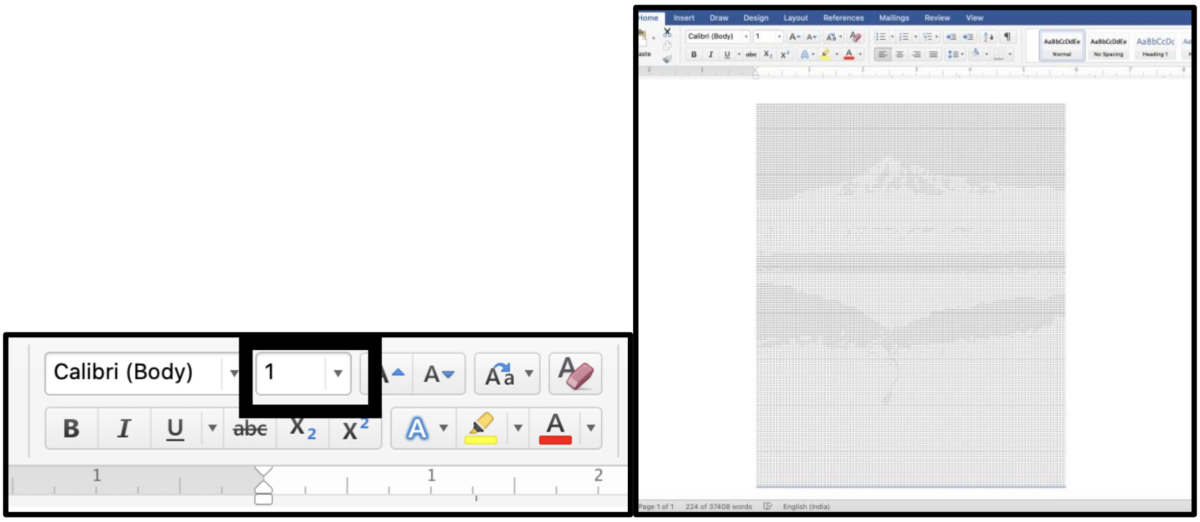
Fig. 3.8: copy this pixel value

Step 6: Paste into a Word Document

- Open a Word document (Google Docs or Microsoft Word).
- Paste the copied pixel values into the document.

Step 7: Adjust the Font Size

- Select all the pasted pixel values in the document.
- Change the font size to **1** for better visualization.
- **Observe the image formation as 0s and 1s recreate the original grayscale image.**



Font: Calibri (Body), size: 1

Fig. 3.9: image formation as 0s and 1s recreate the original grayscale image

In coloured images, each pixel is assigned a specific number based on the RGB colour model, which stands for Red, Green, and Blue.

1 byte= 8 bits so the total number of binary numbers formed will be 2^8=256.

0	0	0	0	0	0	0	0

$= 2^7X0 + 2^6X0 + 2^5X0 + 2^4X0 + 2^3X0 + 2^2X0 + 2^1X0 + 2^0X0 \quad = 0$

1	1	1	1	1	1	1	1

$= 2^7X1 + 2^6X1 + 2^5X1 + 2^4X1 + 2^3X1 + 2^2X1 + 2^1X1 + 2^0X1 \quad = 255$

By combining different intensities of red, green, and blue, a wide range of colours can be represented in an image, each colour channel can have a value from 0 to 255, resulting in over 16 million possible colours.

3.3. COMPUTER VISION – PROCESS:

The Computer Vision process often involves five stages. They are explained below.

3.3.1. Image Acquisition: *Image acquisition is the initial stage in the process of computer vision, involving the capture of digital images or videos*. This step is crucial as it provides the raw data upon which subsequent analysis is based. Digital images can be acquired through various means, including capturing them with digital cameras, scanning physical photographs or documents, or even generating them using design software.

The quality and characteristics of the acquired images greatly influence the effectiveness of subsequent processing and analysis. It is important to understand that the capabilities and resolutions of different imaging devices play a significant role in determining the quality of acquired images. Higher-resolution devices can capture finer details and produce clearer images compared to those with lower resolutions. Moreover, various factors such as lighting conditions and angles can influence the effectiveness of image acquisition techniques. For instance, capturing images in low-light conditions may result in poorer image quality, while adjusting the angle of capture can provide different perspectives of the scene.

In scientific and medical fields, specialized imaging techniques like MRI (Magnetic Resonance Imaging) or CT (Computed Tomography) scans are employed to acquire highly detailed images of biological tissues or structures. These advanced imaging modalities offer insights into the internal composition and functioning of biological entities, aiding in diagnosis, research, and treatment planning.

3.3.2. Preprocessing:

Preprocessing in computer vision aims to enhance the quality of the acquired image. Some of the common techniques are-

a. **Noise Reduction**: Removes unwanted elements like blurriness, random spots, or distortions. This makes the image clearer and reduces distractions for algorithms.
Example: Removing grainy effects in low-light photos.

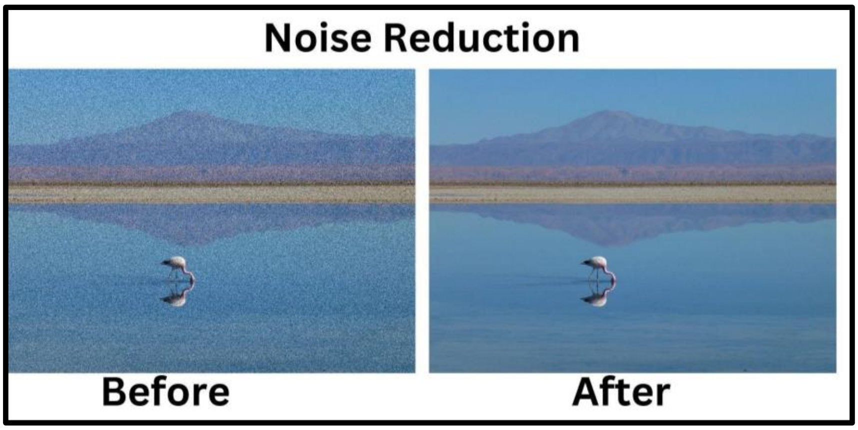


Fig. 3.10: before image is noise image, after image is noise reduced

b. **Image Normalization**: Standardizes pixel values across images for consistency. Adjusts the pixel values of an image so they fall within a consistent range (e.g., 0–1 or -1 to 1).
Ensures all images in a dataset have a similar scale, helping the model learn better.
Example: Scaling down pixel values from 0–255 to 0–1.

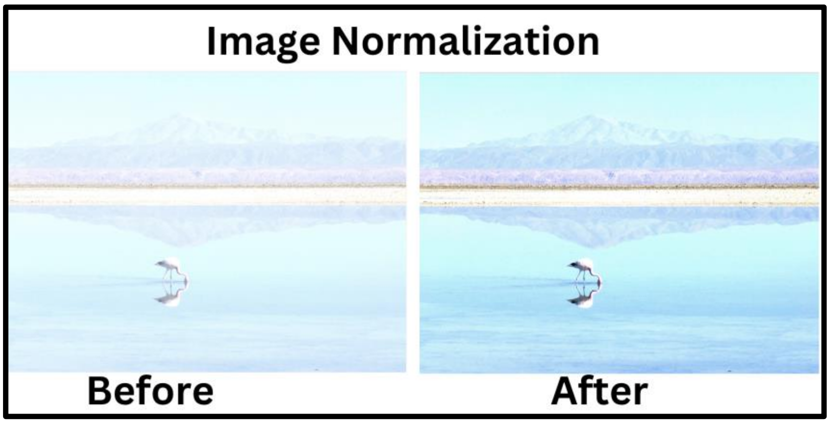


Fig. 3.11: before is distorted with no normalization

c. **Resizing/Cropping**: Changes the size or aspect ratio of the image to make it uniform. Ensures all images have the same dimensions for analysis.
Example: Resizing all images to 224×224 pixels before feeding them into a neural network.

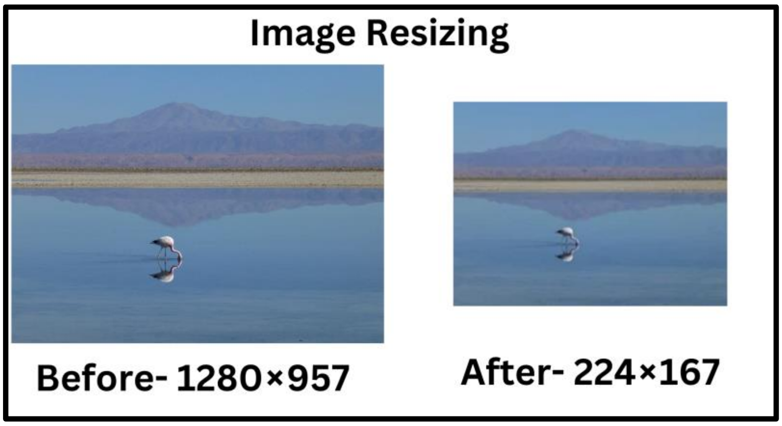


Fig. 3.12: Resizing of image

d. **Histogram Equalization**: Adjusts the brightness and contrast of an image. Spreads out the pixel intensity values evenly, enhancing details in dark or bright areas. Example: Making a low-contrast image look sharper and more detailed.

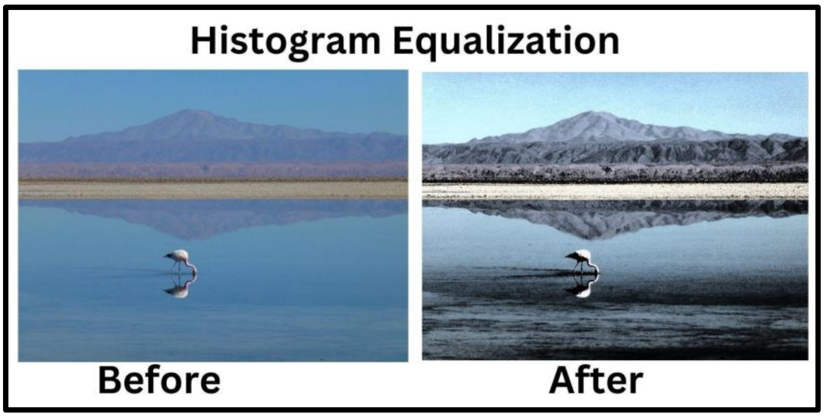


Fig. 3.13

The main goal for preprocessing is to prepare images for computer vision tasks by:

- Removing noise (disturbances).
- Highlighting important features.
- Ensuring consistency and uniformity across the dataset.

3.3.3. Feature Extraction:

Feature extraction involves identifying and extracting relevant visual patterns or attributes from the pre-processed image. Feature extraction algorithms vary depending on the specific application and the types of features relevant to the task. The choice of feature extraction method depends on factors such as the complexity of the image, the computational resources available, and the specific requirements of the application.

- **Edge detection** identifies the boundaries between different regions in an image where there is a significant change in intensity
- **Corner detection** identifies points where two or more edges meet. These points are areas of high curvature in an image, focused on identifying sharp changes in image gradients, which often correspond to corners or junctions in objects.
- **Texture analysis** extracts features like smoothness, roughness, or repetition in an image
- **Colour-based feature extraction** quantifies colour distributions within the image, enabling discrimination between different objects or regions based on their colour characteristics.

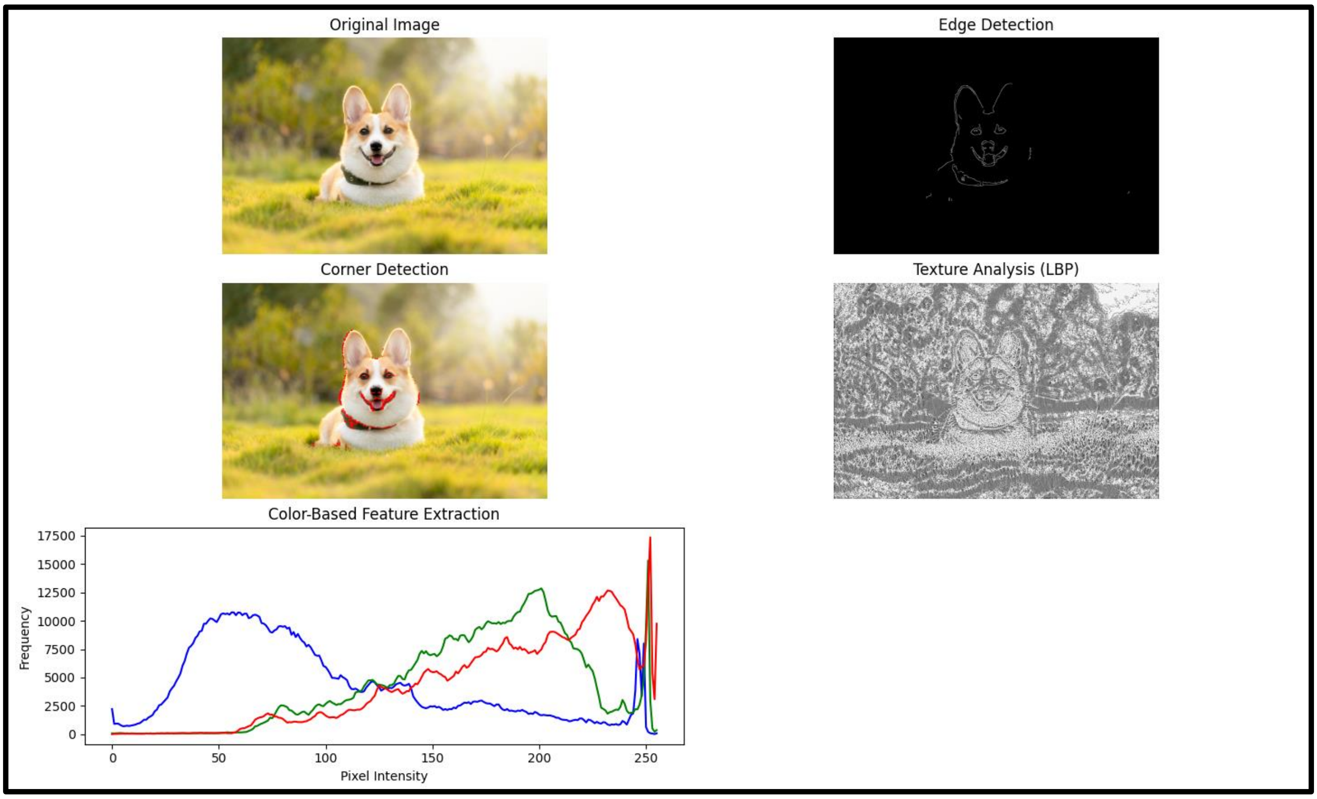


Fig.3.14-Edge detection, corner detection, texture analysis, color-based feature extraction

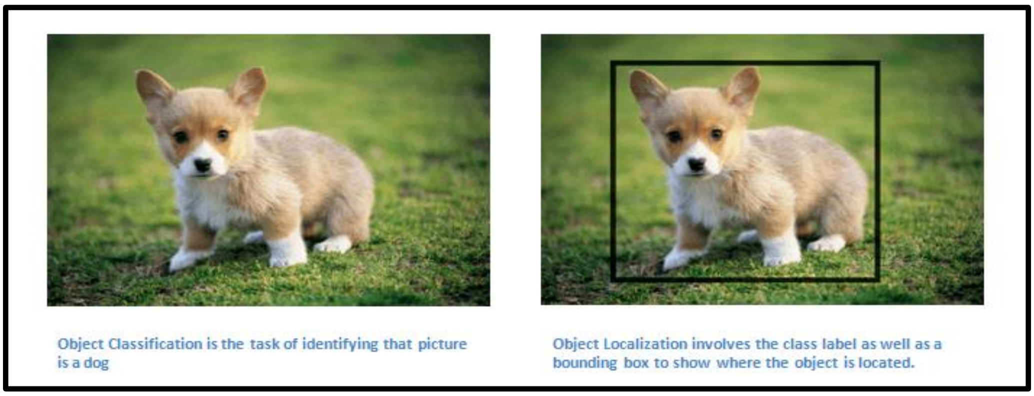
In deep learning-based approaches, feature extraction is often performed automatically by convolutional neural networks (CNNs) during the training process.

3.3.4. Detection/Segmentation:

Detection and segmentation are fundamental tasks in computer vision, focusing on identifying objects or regions of interest within an image. These tasks play a pivotal role in applications like autonomous driving, medical imaging, and object tracking. This crucial stage is categorized into two primary tasks:

1. Single Object Tasks
2. Multiple Object Tasks

Single Object Tasks: Single object tasks focus on analysing/or delineate individual objects within an image, with two main objectives:



Object Classification is the task of identifying that picture is a dog

Object Localization involves the class label as well as a bounding box to show where the object is located.

Fig.3.15: classification, classification+localization

i) **Classification**: This task involves determining the category or class to which a single object belongs, providing insights into its identity or nature. KNN(K-Nearest Neighbour)algorithm may be used for supervised classification while K-means clustering algorithm can be used for unsupervised classification.

ii) **Classification + Localization**: In addition to classifying objects, this task also involves precisely localizing the object within the image by predicting bounding boxes that tightly enclose it.

Multiple Object Tasks: Multiple object tasks deal with scenarios where an image contains multiple instances of objects or different object classes. These tasks aim to identify and distinguish between various objects within the image, and they include:

i) **Object Detection**: Object detection focuses on identifying and locating multiple objects of interest within the image. It involves analysing the entire image and drawing bounding boxes around detected objects, along with assigning class labels to these boxes. The main difference between classification and detection is that classification considers the image as a whole and determines its class whereas detection identifies the different objects in the image and classifies all of them.

In detection, bounding boxes are drawn around multiple objects and these are labelled according to their particular class. Object detection algorithms typically use extracted features and learning algorithms to recognize instances of an object category. Some of the algorithms used for object detection are: R-CNN (Region-

Based Convolutional Neural Network), R-FCN (Region-based Fully Convolutional Network), YOLO (You Only Look Once) and SSD (Single Shot Detector).

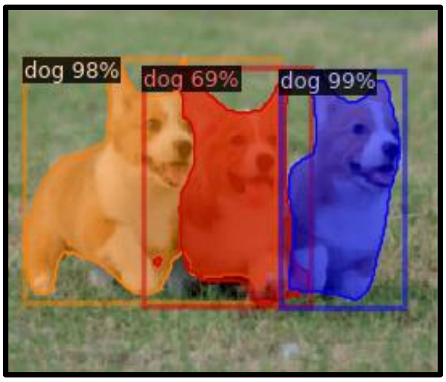


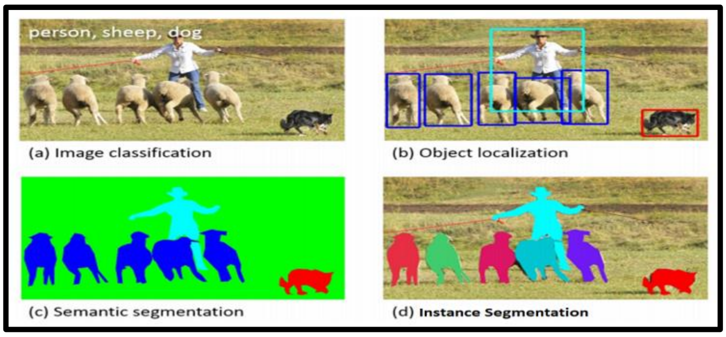
Fig.3.16- object detection

ii) **Image segmentation:** It creates a mask around similar characteristic pixels and identifies their class in the given input image. Image segmentation helps to gain a better understanding of the image at a granular level. Pixels are assigned a class and for each object, a pixel-wise mask is created in the image. This helps to easily identify each object separately from the other. Techniques like Edge detection which works by detecting discontinuities in brightness is used in Image segmentation. There are different types of Image Segmentation available.

Two of the popular segmentation are:

a. Semantic Segmentation: It classifies pixels belonging to a particular class. Objects belonging to the same class are not differentiated. In this image for example the pixels are identified under class animals but do not identify the type of animal.

b. Instance Segmentation: It classifies pixels belonging to a particular instance. All the objects in the image are differentiated even if they belong to the same class. In this image for example the pixels are separately masked even though they belong to the same class.



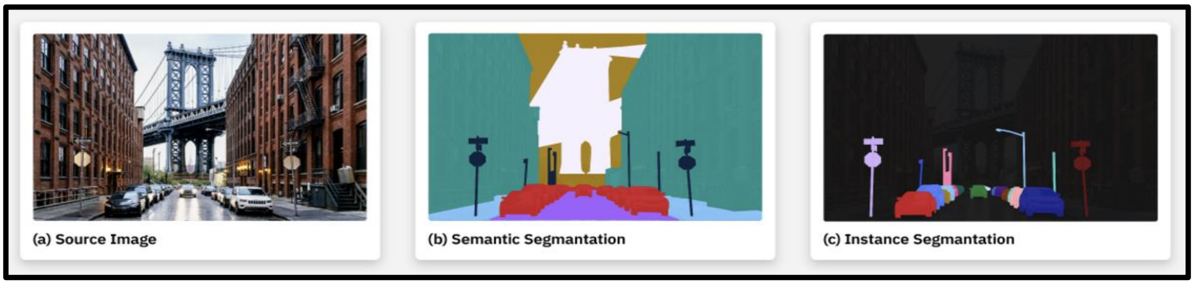


Fig.3.17

3.3.5. High-Level Processing: In the final stage of computer vision, high-level processing plays a crucial role in interpreting and extracting meaningful information from the detected objects or regions within digital images. This advanced processing enables computers to achieve a deeper understanding of visual content and make informed decisions based on the visual data. Tasks involved in high-level processing include recognizing objects, understanding scenes, and analysing the context of the visual content. Through sophisticated algorithms and machine learning techniques, computers can identify and categorize objects, infer relationships between elements in a scene, and derive insights from complex visual data. Ultimately, high-level processing empowers computer vision systems to extract valuable insights and drive intelligent decision-making in various applications, ranging from autonomous driving to medical diagnostics.

3.4. APPLICATIONS OF COMPUTER VISION

Computer vision is one of the areas in Machine Learning whose principle is already integrated into major products that we use every day. Some of the applications are listed below which you might have already learned in lower classes.

- **Facial recognition:** Popular social media platforms like Facebook uses facial recognition to detect and tag users.
- **Healthcare:** Helps in evaluating cancerous tumours, identifying diseases or abnormalities. Object detection & tracking in medical imaging.
- **Self-driving vehicles:** Makes sense of the surroundings by capturing video from different angles around the car. Detect other cars and objects, read traffic signals, pedestrian paths, etc.
- **Optical character recognition (OCR)**: Extract printed or handwritten text from visual data such as images or documents like invoices, bills, articles, etc.
- **Machine inspection:** Detects a machine's defects, features, and functional flaws, determines inspection goals, chooses lighting and material-handling techniques, and other irregularities in manufactured products.
- **3D model building:** Constructing 3D computer models from existing objects which has a variety of applications in various places, such as Robotics, Autonomous driving, 3D tracking, 3D scene reconstruction, and AR/VR.
- **Surveillance:** Live footage from CCTV cameras in public places helps to identify suspicious behaviour, identify dangerous objects, and prevent crimes by maintaining law and order.
- **Fingerprint recognition and biometrics:** Detects fingerprints and biometrics to validate a user's identity.

3.5. CHALLENGES OF COMPUTER VISION

Computer vision, a vital part of artificial intelligence, faces several hurdles as it strives to make sense of the visual world around us. These challenges include:

1. **Reasoning and Analytical Issues:** Computer vision relies on more than just image identification; it requires accurate interpretation. Robust reasoning and analytical skills are essential for defining attributes within visual content. Without such capabilities, extracting meaningful insights from images becomes challenging, limiting the effectiveness of computer vision systems.

2. **Difficulty in Image Acquisition:** Image acquisition in computer vision is hindered by various factors like lighting variations, perspectives, and scales. Understanding complex scenes with multiple objects and handling occlusions adds to the complexity. Obtaining high-quality image data amidst these challenges is crucial for accurate analysis and interpretation.

3. **Privacy and Security Concerns:** Vision-powered surveillance systems raise serious privacy concerns, potentially infringing upon individuals' privacy rights. Technologies like facial recognition and detection prompt ethical dilemmas regarding privacy and security. Regulatory scrutiny and public debate surround the use of such technologies, necessitating careful consideration of privacy implications.

4. **Duplicate and False Content:** Computer vision introduces challenges related to the proliferation of duplicate and false content. Malicious actors can exploit vulnerabilities in image and video processing algorithms to create misleading or fraudulent content. Data breaches pose a significant threat, leading to the dissemination of duplicate images and videos, fostering misinformation and reputational damage.

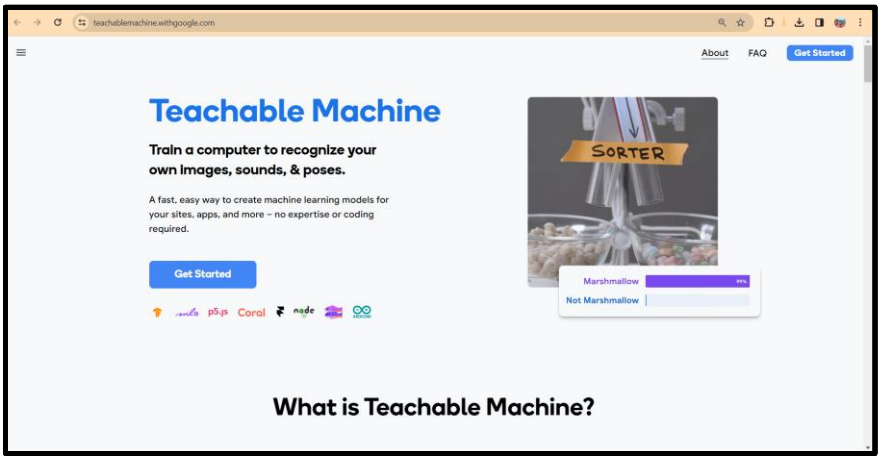
3.6. THE FUTURE OF COMPUTER VISION

Over the years, computer vision has evolved from basic image processing tasks to complex systems capable of understanding and interpreting visual data with human-like precision. Breakthroughs in deep learning algorithms, coupled with the availability of vast amounts of labelled training data, have propelled the field forward, enabling machines to perceive and analyse images and videos in ways previously thought impossible.

As we look to the future, the possibilities by computer vision are awe-inspiring. From personalized healthcare diagnostics to immersive AR experiences, the impact of computer vision on society is set to be profound and far-reaching. By embracing innovation, fostering collaboration, and prioritizing ethical considerations, we can unlock the full potential of computer vision and harness its transformative power for the benefit of humanity.

ACTIVITY 3.2 CREATING A WEBSITE CONTAINING AN ML MODEL

1. Go to the website https://teachablemachine.withgoogle.com/



Teachable Machine

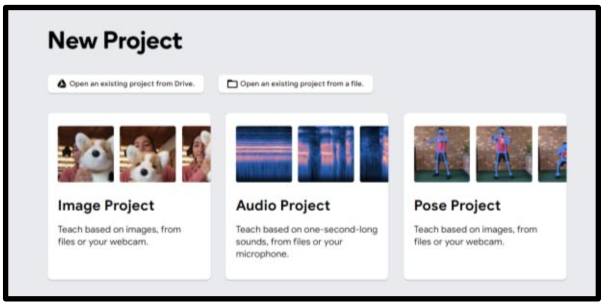
Train a computer to recognize your own images, sounds, & poses.

A fast, easy way to create machine learning models for your sites, apps, and more – no expertise or coding required.

What is Teachable Machine?

Fig.3.18

2. Click on Get Started.
3. Teachable machine offers 3 options as you can see.


Open an existing project from Drive.
Open an existing project from a file.

New Project

Image Project

Teach based on images, from files or your webcam.

Audio Project

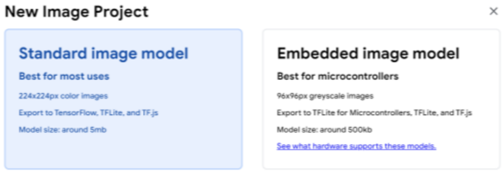
Teach based on one-second-long sounds, from files or your microphone.

Pose Project

Teach based on images, from files or your webcam.

Fig.3.19

4. Choose the 'Image' project.
5. Select 'Standard Image Model'.



New Image Project

Standard image model

Best for most uses

224x224px color images

Export to TensorFlow, TFLite, and TF.js

Model size: around 5mb

Embedded image model

Best for microcontrollers

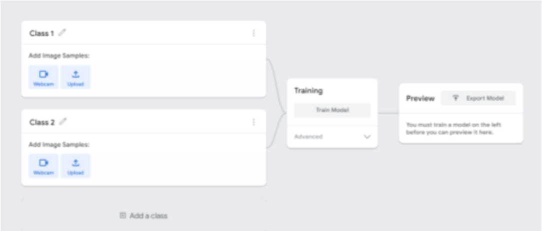
96x96px greyscale images

Export to TFLite for Microcontrollers, TFLite, and TF.js

Model size: around 500kb

Fig.3.20

6. You will get a screen like this.

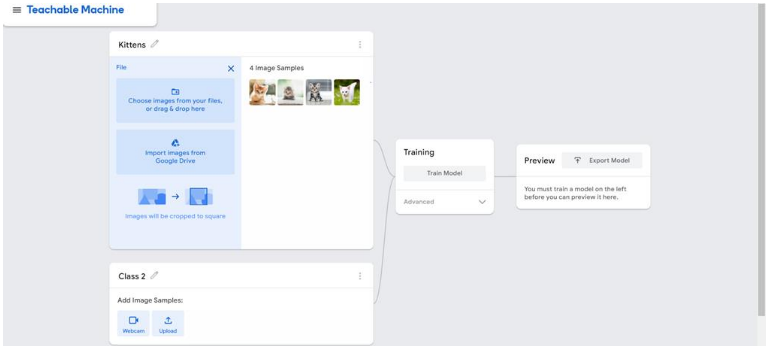


You must train a model on the left before you can preview it here.

Fig.3.21

You have the option to choose between two methods: using your **webcam** to capture images or **uploading** existing images. For the webcam option, you will need to position the image in front of the camera and hold down the record button to capture the image. Alternatively, with the upload option, you have the choice to upload images either from your local computer or directly from Google Drive.

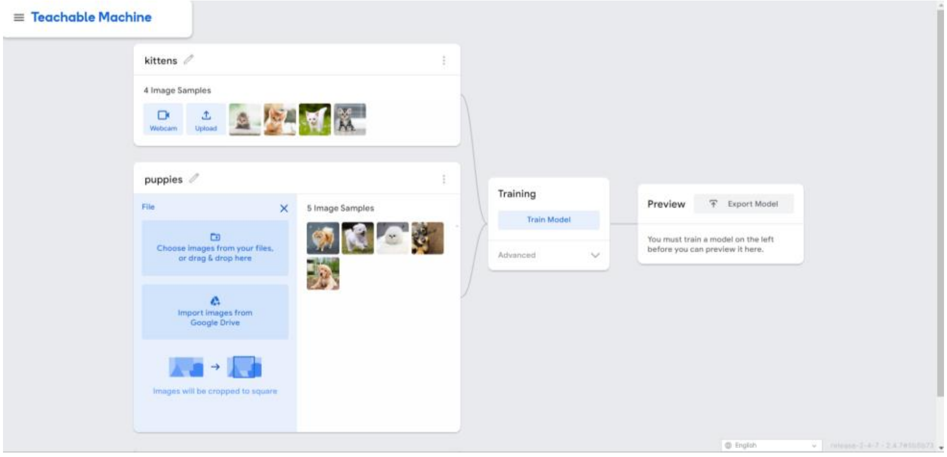
7. Let us name ‘Class 1’ as Kittens and upload pictures already saved on the computer.



You must train a model on the left before you can preview it here.

Fig.3.22

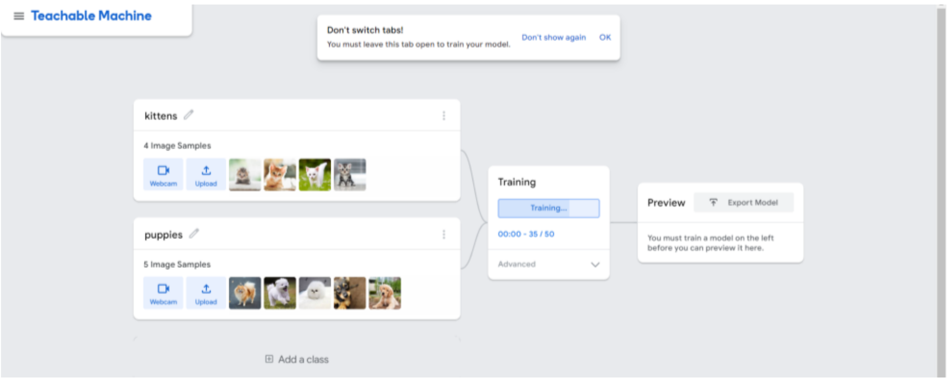
8. Now, let us add another class of images “Puppies” saved on the computer.



You must train a model on the left before you can preview it here.

Fig.3.23

9. Click on Train Model.



Don't switch tabs!

You must leave this tab open to train your model.

You must train a model on the left before you can preview it here.

Fig.3.24

Once the model is trained, you can test the working by showing an image infront of the web camera. Else, you can also upload the image from your local computer / Google drive.

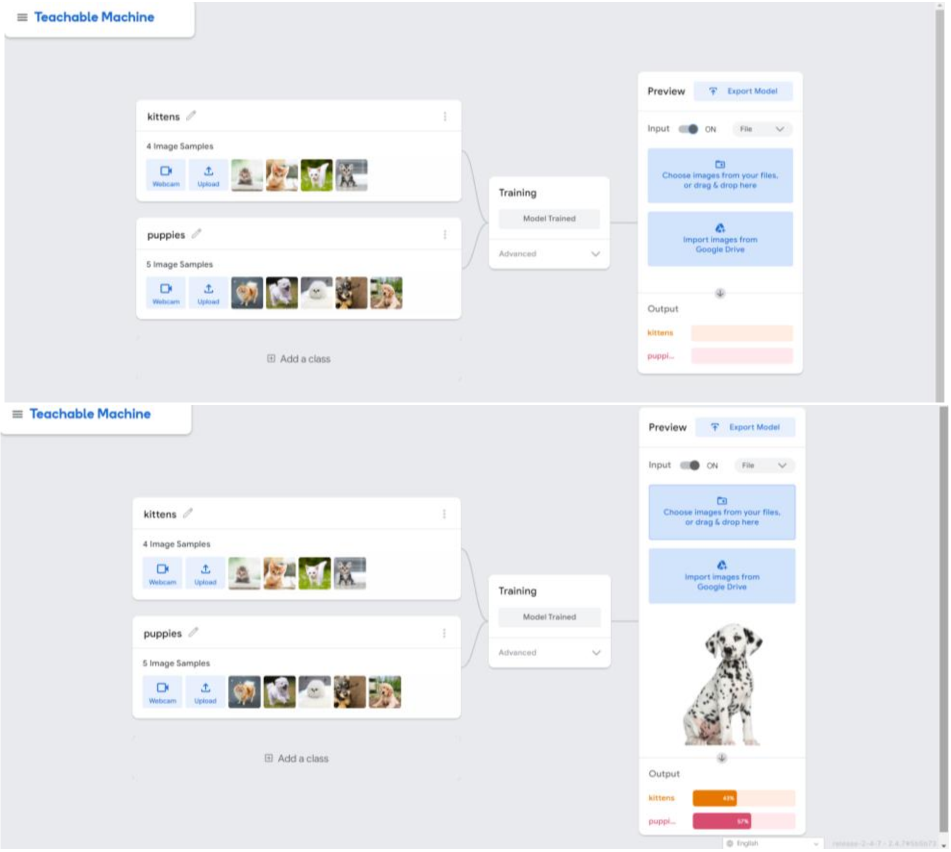
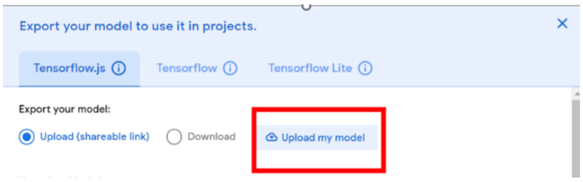


Fig.3.25

10. Now click on ‘Export Model’. A screen will open up as shown. Now, click on Upload my model.



Export your model to use it in projects.

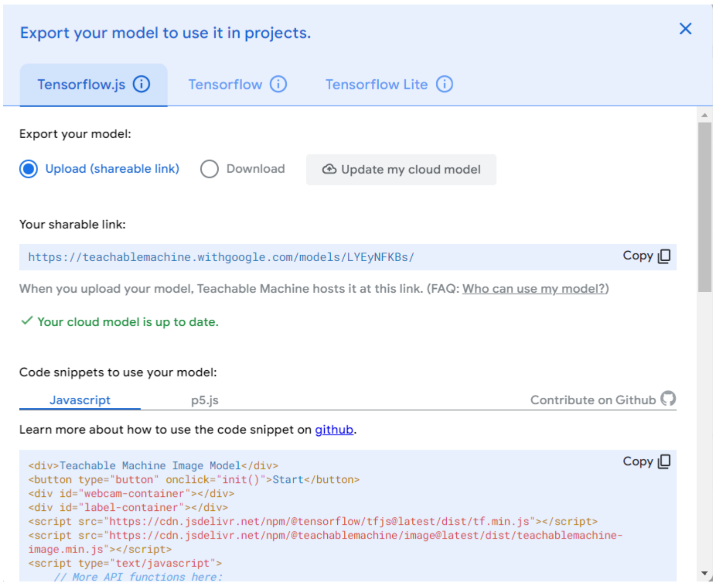
Tensorflow.js ⓘ　Tensorflow ⓘ　Tensorflow Lite ⓘ

Export your model:

● Upload (shareable link)　○ Download　Upload my model

Fig.3.26

11. Once your model is uploaded, Teachable Machine will create a URL which we will use in the Javascript code. Copy the Javascript code by clicking on Copy.



Export your model to use it in projects.

Tensorflow.js ⓘ　Tensorflow ⓘ　Tensorflow Lite ⓘ

Export your model:

● Upload (shareable link)　○ Download　Update my cloud model

Your sharable link:

`https://teachablemachine.withgoogle.com/models/LYEyNFKBs/`　Copy

When you upload your model, Teachable Machine hosts it at this link. (FAQ: Who can use my model?)

✓ Your cloud model is up to date.

Code snippets to use your model:

Javascript　p5.js　Contribute on Github

Learn more about how to use the code snippet on github.

Copy

```
<div>Teachable Machine Image Model</div>
<button type="button" onclick="init()">Start</button>
<div id="webcam-container"></div>
<div id="label-container"></div>
<script src="https://cdn.jsdelivr.net/npm/@tensorflow/tfjs@latest/dist/tf.min.js"></script>
<script src="https://cdn.jsdelivr.net/npm/@teachablemachine/image@latest/dist/teachablemachine-
image.min.js"></script>
<script type="text/javascript">
    // More API functions here:
```

Fig.3.27

12. Open Notepad and paste the JavaScript code and save this file as web.html.
13. Let us now deploy this model in a website.
14. Once you create a free account on Weebly, go to Edit website and create an appealing website using the tools given.

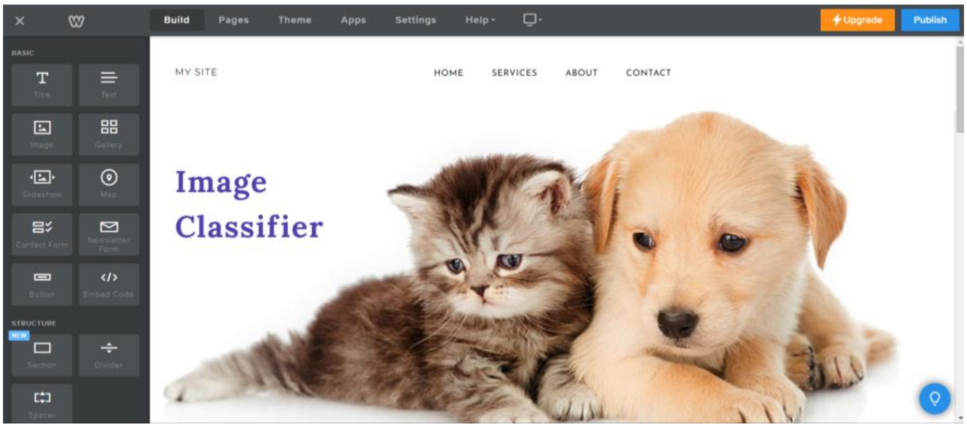


Fig.3.28

15. Click on Embed Code and drag and place it on the webpage.

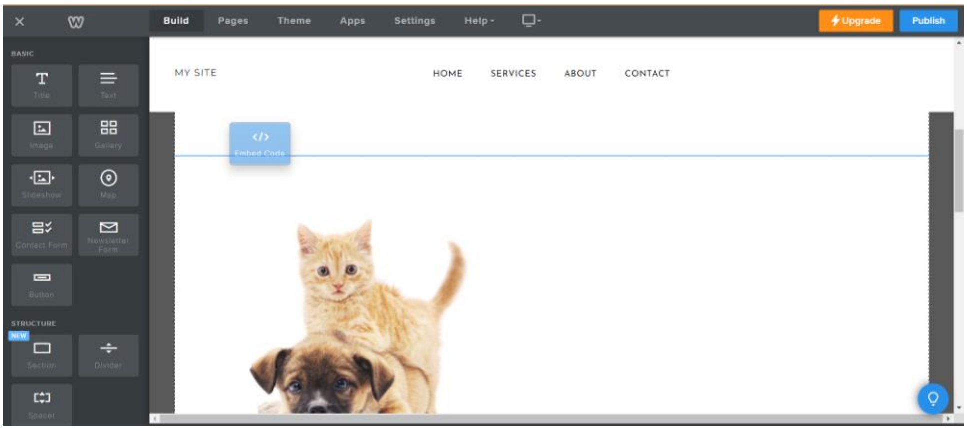
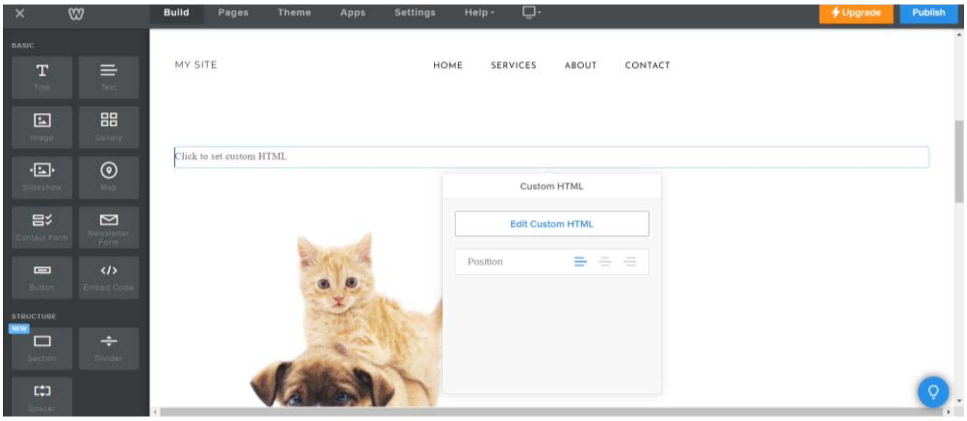


Fig.3.29

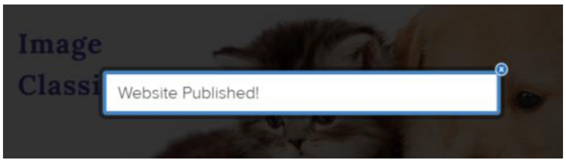
16. Click on 'Edit HTML Code' and copy the Javascript code from the html file (web.html) and paste it here as shown.



Click to set custom HTML

Fig.3.30

17. Click 'Publish' (keep on Weebly subdomain for free version).



Website Published!

Fig.3.31

18. Copy the URL and paste it into a new browser window to check the working of your model.

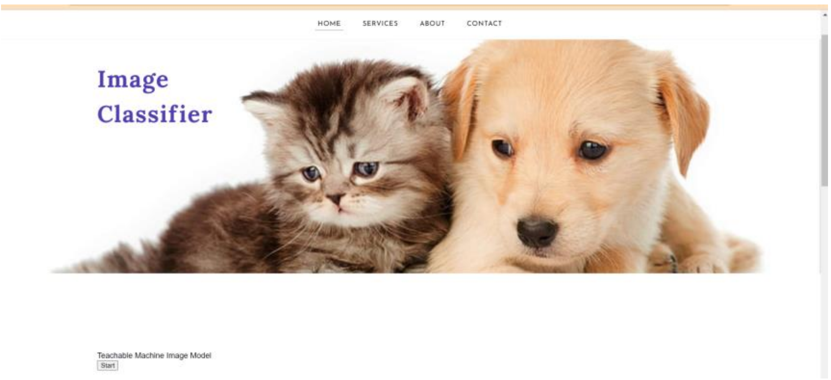
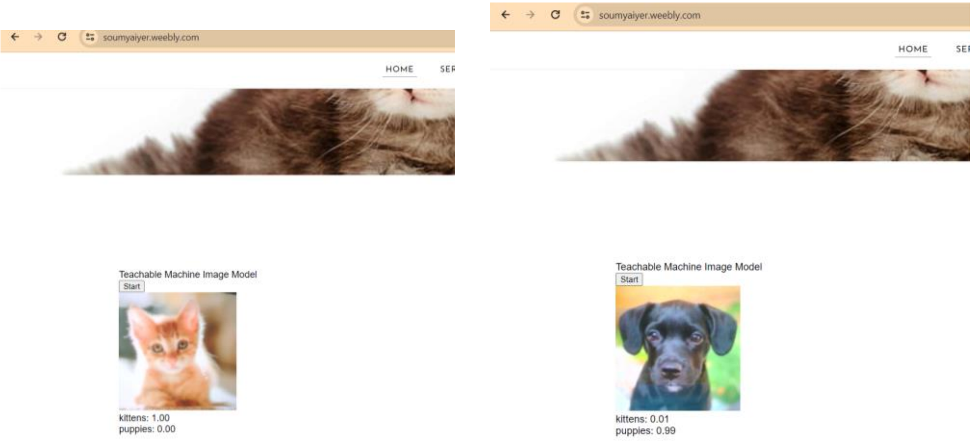


Image Classifier

Teachable Machine Image Model

Fig.3.32

19. Click on start and you can show pictures of kitten and puppy to check the predictions of your model.



Teachable Machine Image Model

kittens: 1.00
puppies: 0.00

Teachable Machine Image Model

kittens: 0.01
puppies: 0.99

Fig.3.33

3.7. Working with OpenCV:(**For Advanced Learners)

3.7.1. Introduction to OpenCV: OpenCV or Open-Source Computer Vision Library is a cross-platform library using which we can develop real-time computer vision applications. It mainly focuses on image processing, video capture, and analysis including features like face detection and object detection. It is also capable of identifying objects, faces, and even handwriting.

To use OpenCV in Python, you need to install the library. Use the following command in your terminal or command prompt:

pip install opencv-python

3.7.2. Loading and Displaying an Image: Let us understand the loading and displaying using a scenario followed by a question.

Scenario- You are working on a computer vision project where you need to load and display an image. You decide to use OpenCV for this purpose.

Question:

What are the necessary steps to load and display an image using OpenCV? Write a Python code snippet to demonstrate this.

sol -

Here's a simple Python script to load and display an image using OpenCV:

```
import cv2
image = cv2.imread('example.jpg')  # Replace 'example.jpg' with the path to your image
cv2.imshow('original image', image)
cv2.waitKey(0)
cv2.destroyAllWindows()
```

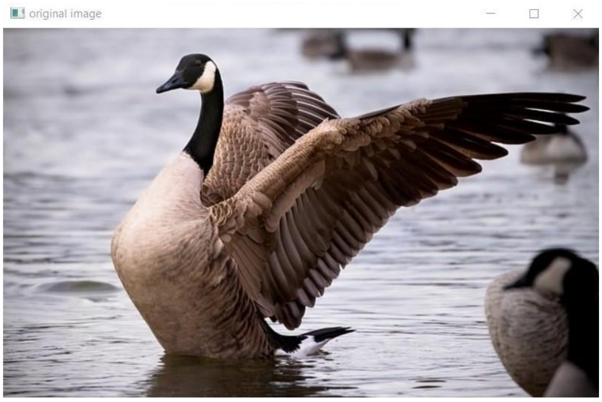


Fig.3.34: loading and displaying image

- `cv2.imread('example.jpg')` loads the image into a variable. Replace `example.jpg` with your image's file name or path.
- `cv2.imshow()` opens a new window to display the image.
- `cv2.waitKey(0)` waits indefinitely for a key press to proceed
- `cv2.destroyAllWindows()` closes any OpenCV windows.

3.7.3. Resizing an image:

Scenario- Imagine we are working on an application where we need to process images of varying sizes. To standardize the input, all images must be resized to 300x300 pixels before further processing. We are using OpenCV to achieve this. Additionally, we want to ensure the aspect ratio is maintained in some cases.

Question:

1. Write a Python script using OpenCV to resize an image named “example.jpg” to a fixed size of 300x300 pixels.

   Sol -

   Here we will use the “cv2.resize()” function to resize the image. We can directly specify the width and height of the new image

```
new_width = 300
new_height = 300
# Resize the image to the new dimensions
resized_image = cv2.resize(image, (new_width, new_height))
```

So, the full code will look like this -

```
import cv2
image = cv2.imread('example.jpg') # Replace 'example.jpg' with the path
to your image
new_width = 300
new_height = 300
# Resize the image to the new dimensions
resized_image = cv2.resize(image, (new_width, new_height))
cv2.imshow('Resized Image', resized_image)
cv2.waitKey(0)
cv2.destroyAllWindows()
```

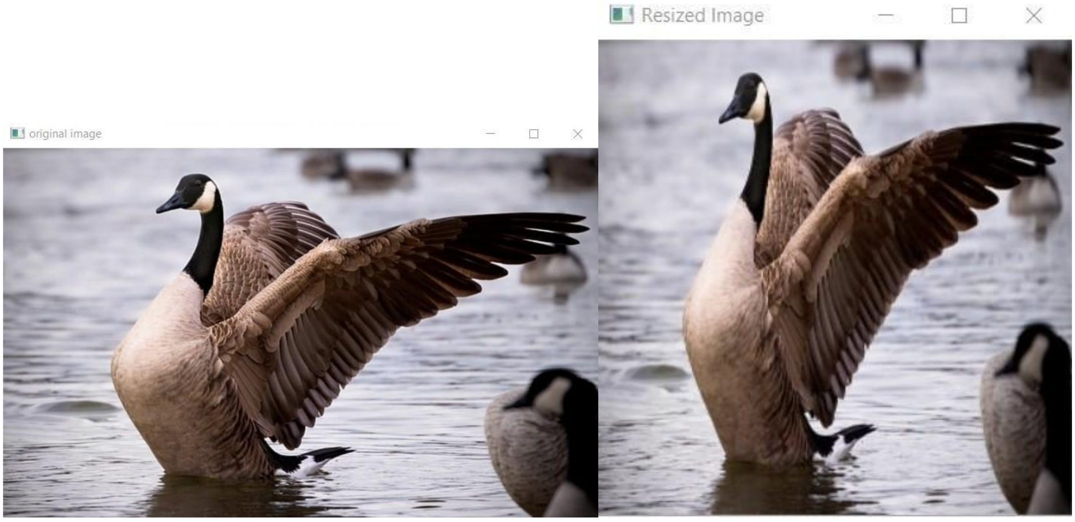


Fig.3.35: Resizing an Image

3.7.4. Converting an Image to Grayscale

Scenario:

Imagine we are working on an application where color images need to be converted to grayscale before further processing, as grayscale images reduce computational complexity. We are using OpenCV to achieve this.

Question: Write a Python script using OpenCV to convert an image named "example.jpg" into a grayscale.

Solution:

Here we will use the `cv2.cvtColor()` function to convert the image into grayscale. The function requires the source image and a color conversion code as inputs. The code for grayscale conversion is `cv2.COLOR_BGR2GRAY`.

```
grayscale_image = cv2.cvtColor(image, cv2.COLOR_BGR2GRAY)
```

So, the full code will look like this:

```
import cv2
image = cv2.imread('example.jpg')# Replace 'example.jpg' with the path to your image
grayscale_image = cv2.cvtColor(image, cv2.COLOR_BGR2GRAY)
cv2.imshow('Grayscale Image', grayscale_image)
cv2.waitKey(0)
cv2.destroyAllWindows()
```

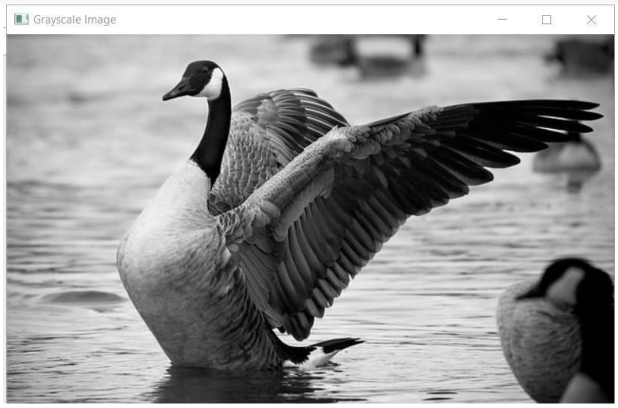


Fig.3.36: converting an image to grayscale

EXERCISES

A. Multiple Choice Questions:

1. The field of study that helps to develop techniques to help computers "see" is_______________.

    a. Python b. Convolution c. Computer Vision d. Data Analysis

2. Task of taking an input image and outputting/assigning a class label that best describes the image is ___________.

    a. Image classification b. Image localization
    c. Image Identification d. Image prioritization

3. Identify the incorrect option

    (i) computer vision involves processing and analysing digital images and videos to understand their content.
    (ii) A digital image is a picture that is stored on a computer in the form of a sequence of numbers that computers can understand.
    (iii) RGB colour code is used only for images taken using cameras.
    (iv) Image is converted into a set of pixels and less pixels will resemble the original image.

    a. ii b. iii c. iii & iv d. ii & iv

4. The process of capturing a digital image or video using a digital camera, a scanner, or other imaging devices is related to ________.

    a. Image Acquisition b. Preprocessing
    c. Feature Extraction d. Detection

5. Which algorithm may be used for supervised learning in computer vision?

    a. KNN b. K-means c. K-fold d. KEAM

6. A computer sees an image as a series of ___________

    a. colours b. pixels c. objects d. all of the above

7. _____________ empowers computer vision systems to extract valuable insights and drive intelligent decision-making in various applications, ranging from autonomous driving to medical diagnostics.

a. Low level processing   b. High insights

c. High-level processing   d. None of the above

8. In Feature Extraction, which technique identifies abrupt changes in pixel intensity and highlights object boundaries?

a. Edge detection   b. Corner detection

c. Texture Analysis   d. boundary detection

9. Choose the incorrect statement related to preprocessing stage of computer vision

a. It enhances the quality of acquired image

b. Noise reduction and Image normalization is often employed with images

c. Techniques like histogram equalization can be applied to adjust the distribution of pixel intensities

d. Edge detection and corner detection are ensured in images.

10. 1 byte = ___________ bits

a. 10   b. 8   c. 2   d. 1

B. Short Answer Questions

1. What is Computer Vision?

Ans - Computer Vision is a field of artificial intelligence (AI) that uses digital images and deep learning models to help systems understand and interpret the visual world.

2. What is the main difference between classification and detection?

Ans - The main difference between classification and detection is that classification considers the image as a whole and determines its class whereas detection identifies the different objects in the image and classifies all of them.

3. Write down any two algorithm which can be used for object detection.

Ans- Algorithms used for object detection are: R-CNN (Region-Based Convolutional Network) and R-FCN (Region-based Fully Convolutional Network).

4. Write down the process of object detection in a single object.

Ans- Single object tasks focus on analyzing individual objects within an image, with two main objectives:

Classification: - This task involves determining the category or class to which a single object belongs, providing insights into its identity or nature.

Classification + Localization: In addition to classifying objects, this task also involves precisely localizing the object within the image by predicting bounding boxes that tightly enclose it.

5. Write any four applications of computer vision.

Ans –

1)Facial recognition: Popular social media platforms like Facebook uses facial recognition to detect and tag users.

2)Healthcare: Helps in evaluating cancerous tumours, identify diseases or abnormalities. Object detection & tracking in medical imaging

3)Surveillance: Live footage from CCTV cameras in public places helps to identify suspicious behaviour, identify dangerous objects, and prevent crimes by maintaining law and order.

4)Fingerprint recognition and biometrics: Detects fingerprints and biometrics to validate a user's identity.

C. Long Answer Questions

1. What do you mean by Image segmentation? Explain the popular segmentations?

Ans - Image segmentation creates a mask around similar characteristic pixels and identifies their class in the given input image. Image segmentation helps to gain a better understanding of the image at a granular level. Pixels are assigned a class and for each object, a pixel-wise mask is created in the image. This helps to easily identify each object separately from the other. Techniques like Edge detection which works by detecting discontinuities in brightness is used in Image segmentation. Two of the popular segmentation are:

Semantic Segmentation: It classifies pixels belonging to a particular class. Objects belonging to the same class are not differentiated. In this image for example the pixels are identified under class animals but do not identify the type of animal.

Instance Segmentation: It classifies pixels belonging to a particular instance. All the objects in the image are differentiated even if they belong to the same class. In this image for example the pixels are separately masked even though they belong to the same class

2. Explain the challenges faced by computer vision.

Ans - The Challenges faced by Computer vision are:

1. **Reasoning and Analytical Issues:** Robust reasoning and analytical skills are essential for defining attributes within visual content. Without such capabilities, extracting meaningful insights from images becomes challenging.
2. **Difficulty in Image Acquisition:** Image acquisition in computer vision is hindered by various factors like lighting variations, perspectives, and scales. Understanding complex scenes with multiple objects and handling occlusions adds to the complexity. Obtaining high-quality image data amidst these challenges is crucial for accurate analysis and interpretation.
3. **Privacy and Security Concerns:** CV raises privacy concerns, potentially infringing upon individuals' privacy rights. Regulatory scrutiny and public debate surround the use of such technologies, necessitating careful consideration of privacy implications.
4. **Duplicate and False Content:** Computer vision introduces challenges related to the proliferation of duplicate and false content. Data breaches pose a significant threat,

leading to the dissemination of duplicate images and videos, fostering misinformation and reputational damage.

COMPETENCY BASED QUESTIONS:

1. A group of students is participating in a photography competition. As part of the competition, they need to submit digitally captured images of various landscapes. However, one of the students, Aryan, is unsure about how to ensure the best quality for his images when digitizing them. Explain Aryan how the resolution of his images can impact their quality and detail when viewed on a computer screen or printed.

   Ans: Resolution determines the quality and level of detail in an image. Higher resolution images contain more pixels, resulting in clearer and more detailed representations of the landscapes.

2. The Red Fort is hosting a grand cultural event, and keeping everyone safe is top priority! A state-of-the-art security system utilizes different "FEATURE EXTRACTION " to analyse live video feeds and identify potential issues. Identify the feature extraction technique that can be used in the following situation.
   a. A large bag is left unattended near a crowded entrance.
   b. A person tries to climb over a wall near a blind spot.
   c. A group of people starts pushing and shoving in a congested area.
   d. A wanted person with a distinctive red scarf enters the venue.

   Ans: a. Texture Analysis
   b. Edge detection
   c. Corner detection
   d. Colour based extraction

3. Which image segmentation technique would be most effective in this scenario: semantic segmentation or instance segmentation?
   a. You are developing a quality control system for a manufacturing plant. The system needs to analyse images captured by cameras above a conveyor belt to identify and isolate defective products.
   b. You are assisting urban planners in developing a comprehensive land-use map for a growing city. They require analysis of aerial imagery to classify large areas into distinct categories like "buildings," "roads," "vegetation," and "parks."

   Ans:
   a. Instance Segmentation
   b. Semantic Segmentation

4. Shreyan joined an architecture firm wherein he was asked to check if AI could be used to create a detailed, representation of a city district for urban planning purposes, utilizing computer vision techniques and data integration from aerial imagery, topographic surveys, and architectural blueprints. Help him identify the application of computer vision that can be used for the same.

   Ans: 3D Model building using computer vision

5. You work for a fact-checking website. Lately, there's been a surge in INCORRECT news articles and videos online, often manipulated using computer vision techniques. These manipulations can make real people appear to be saying things they never did. What is the concern here?

   Ans. False/Fake content

References:

1. https://www.ibm.com/topics/computer-vision
2. https://www.datacamp.com/tutorial/seeing-like-a-machine-a-beginners-guide-to-image-analysis-in-machine-learning
3. https://medium.com/@edgeneural.ai/computer-vision-making-the-machines-see-361a3d0cfc3f
4. https://www.javatpoint.com/computer-vision

UNIT 4: AI with Orange Data Mining Tool

Title: AI with Orange Data Mining Tool

Approach: Practical work, Group activity, Creativity, Data analysis, Group discussion

Summary: Students will learn to use Orange's intuitive interface and component-based visual programming approach in the domains of Data Science, Computer Vision, and Natural Language Processing. They will explore its diverse set of widgets, covering data visualization, preprocessing, feature selection, modeling, and evaluation, gaining practical insights into its applications in these fields.

Learning Objectives:

1. Students will gain a comprehensive understanding of Orange Data Mining tool, empowering them to leverage its capabilities across various domains of Artificial Intelligence.
2. Students will gain practical insights into its applications in Data Science, Computer Vision, and Natural Language Processing (NLP) through detailed exploration of different widgets and functionalities offered by Orange.

Key Concepts:

1. Introduction to Orange Data Mining Tool
2. Components of Orange Data Mining Tool
3. Key domains of AI with Orange data mining tool – Data Science, Computer Vision, NLP

Learning outcomes:

Students will be able to -

1. Develop proficiency in utilizing the Orange Data Mining tool, enabling them to navigate its interface, employ its features, and execute data analysis tasks effectively.
2. Demonstrate the ability to apply Orange in real-world scenarios across diverse domains of Artificial Intelligence, including Data Science, Computer Vision, and Natural Language Processing (NLP), through hands-on projects and case studies.

Prerequisites:

- Awareness and understanding of the terms Data Science, Natural Language Processing, and Computer Vision.
- Basic knowledge of algorithms used in Machine Learning.

Unveiling Data Secrets with Orange Data Mining Tool: A Teacher's Guide

This lesson equips you to introduce students to Orange, a user-friendly data mining tool. Students will learn to manipulate data, create visualizations, and explore machine learning through practical exercises.

1. Welcome to Orange World!

- **Introduction:** Introduce Orange as a powerful data mining tool that helps us explore data, visualize patterns, and build predictive models.

2. Orange Tour:

- **Interface Overview:** Demonstrate the Orange interface, highlighting key areas:
    - Data: Loading and managing datasets.
    - Visualize: Creating various charts and graphs to explore data visually.
    - Preprocess: Cleaning and preparing data for analysis.
    - Models: Building and evaluating machine learning models.
    - Explore: Interactive data exploration tools.
- **Hands-on Exercise:** Guide students through loading a sample dataset (e.g., Iris flower dataset) and exploring its features using the "Data Table" widget.

3. The Power of Visualization:

- **Data Visualization Techniques:** Introduce different data visualization techniques in Orange:
    - Histograms: Exploring the distribution of numerical data.
    - Scatter Plots: Identifying relationships between two variables.
    - Box Plots: Comparing distributions across different groups.
- **Interactive Exploration:** Demonstrate how to create visualizations using Orange widgets (e.g., "Scatter Plot", "Box Plot"). Guide students in visualizing features of the Iris dataset to identify potential patterns.

4. Unveiling Machine Learning:

- **Machine Learning with Orange:** Introduce the concept of Machine Learning and how Orange can be used to build models.
- **Model Evaluation:** Explain how to evaluate the performance of models using metrics like accuracy and confusion matrix.
- **Classification with Orange:** Demonstrate building a classification model to predict Iris flower species based on features. Guide students through:
    - Selecting a classification algorithm (e.g., K-Nearest Neighbors).
    - Splitting data into training and testing sets.
    - Training the model and evaluating its performance.

5. Text Mining with Orange's "Keyphrases" Widget:

- **Unveiling Text Data:** Introduce the "Keyphrases" widget to analyze text data.
  - **Focus:** Highlight Orange's capabilities for text mining tasks like identifying key topics, sentiment analysis, and trends.
- **Practical Demonstration:** Demonstrate using the "Keyphrases" widget on a sample text dataset. Discuss the extracted keywords and their potential insights.

6. A Deeper Look at the Iris Dataset:

- **Exploring Data:** Guide students through a deeper exploration of the Iris dataset:
  - Load the dataset in Orange.
  - Visualize characteristics like sepal length and sepal width using scatter plots.
  - Analyze patterns and discuss potential relationships between features.

7. Classification in Action:

- **Classification Models:** Discuss different classification models available in Orange.
  - **Focus:** Emphasize checking accuracy, visualizing results (e.g., confusion matrix), and tuning hyperparameters to improve model performance.
- **Hands-on Practice:** Guide students in building a classification model to predict Iris flower species. Encourage them to experiment with different algorithms and hyperparameters to optimize the model's accuracy.

8. Beyond Data: Orange for More:

- **Computer Vision with Orange:** Briefly introduce Orange's capabilities for computer vision tasks, including image processing, feature extraction, and object detection.
- **Natural Language Processing (NLP) with Orange:** Briefly discuss how Orange can be used for NLP tasks like text preprocessing, sentiment analysis, topic modeling, and classification.

Additional Tips:

- Utilize real-world examples to showcase the applications of data mining and machine learning.
- Encourage students to ask questions and explore different functionalities within Orange.
- Provide online tutorials and resources for students to delve deeper into specific Orange features.

By incorporating these elements, you can equip students with a valuable toolkit for data exploration, visualization, and machine learning using Orange, empowering them to uncover hidden insights from data.

Teachers can ask the following questions to spark curiosity before starting the topics:

- **Have you ever wondered how businesses analyze vast amounts of data to make decisions, like predicting customer preferences or classifying products? What tools or techniques do you think they use to make sense of such data?** (This question introduces the concept of data mining and piques interest in tools like Orange.)

- **Imagine you're tasked with identifying patterns in social media images or analysing customer reviews for trends. What kind of process or tools might you use to extract meaningful insights from such diverse data sources?** (This question sets the stage for exploring Orange's applications in data science, computer vision, and natural language processing.)

4.1. What is Data Mining?

Data mining is the process of discovering trends, useful information, and patterns from large datasets. It involves analyzing and interpreting data to extract meaningful insights that can aid in decision-making.

4.2. Introduction to Orange Data Mining Tool

In this unit, we embark on an exhilarating journey into the realms of Data Science, Computer Vision, and Natural Language Processing (NLP) using the capabilities of Orange data mining. Orange is a component-based visual programming software package designed for various tasks such as data visualization, machine learning, data mining, and data analysis. Within Orange, components are referred to as widgets, offering a diverse array of functionalities. These widgets range from basic data visualization, subset selection, and preprocessing to the empirical evaluation of learning algorithms and predictive modeling. The essence of visual programming is manifested through an intuitive interface, where workflows are constructed by interconnecting predefined or user-designed widgets. Throughout this unit, we will delve into the applications and challenges of AI techniques and explore how they can be applied to tackle real-world problems and extract valuable insights from data using Orange data mining.

4.3. Beneficiaries of Orange data mining

Data Analysts and Scientists	Orange provides a user-friendly interface for data analysis, making it accessible to professionals who may not have extensive programming skills.
Researchers	Orange provides tools for exploring and analyzing research data, facilitating hypothesis testing, and generating insights from experimental results

Educators and Students	Educators can leverage Orange's intuitive interface and visual programming environment to introduce students to complex topics in a more approachable way.
Business Professionals	Orange's visualizations and predictive modeling capabilities can help businesses identify trends, predict customer behavior, optimize processes, and improve performance.
Open-Source Community	Orange is an open-source software tool, meaning its source code is freely available to the public.

4.4. Getting started with Orange tool

To install the Orange tool on your system, follow these simple steps:

4.4.1. Visit the Orange Website:

Go to the official Orange website using the link: https://orangedatamining.com/download/
Scan the QR code.

Fig.4.1: Scan the QR code

4.4.2. Choose the Correct Version:

- **For Windows Users**: Download the "Standalone installer" tailored for Windows users.
- **For Mac Users**: Select "Orange for Apple silicon" for Mac users.

4.4.3. Download the Installer:

Click on the respective download link to initiate the download process.

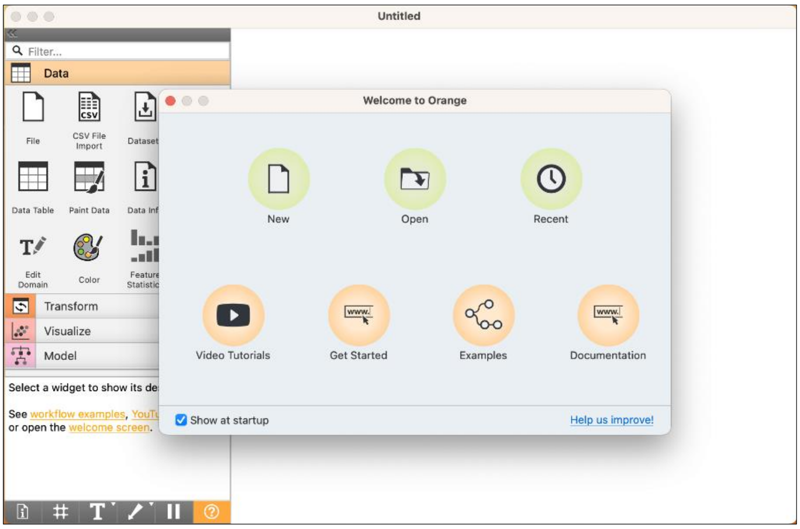
4.4.4. Install the Software:

Once the download is complete, locate the downloaded file and run it.

- For Windows: Double-click the installer file and follow the on-screen instructions to install Orange on your system.
- For Mac: Double-click the downloaded file to mount the disk image. Drag the Orange application to your Applications folder to install it.

4.4.5. Launch Orange:

After installation is complete, launch the Orange tool from your system's applications menu or by double-clicking its icon.



Select a widget to show its de

See workflow examples, YouTu
or open the welcome screen.

Fig.4.2: Orange Data Mining Tool

4.5. COMPONENTS OF ORANGE DATA MINING TOOL:-

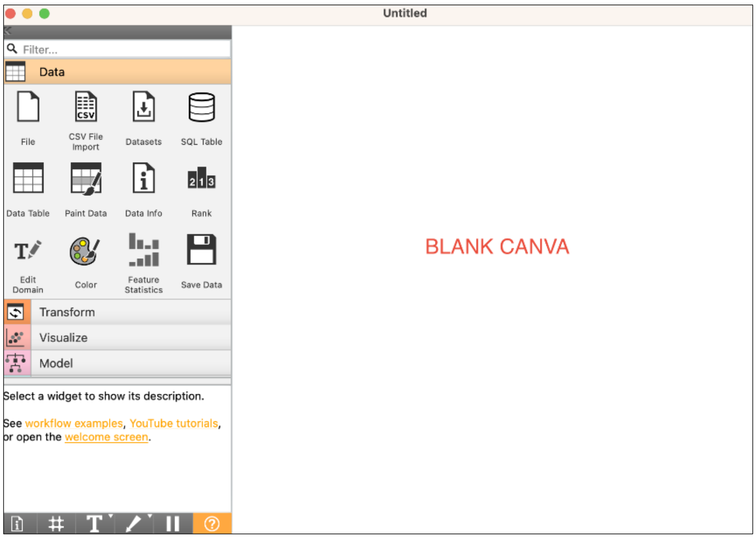
The main components of Orange tools are-

1. Blank Canva 2.Widgets 3. Connectors

1. **Blank Canva:** The Blank Canva is where you build your analysis workflows by dragging and dropping widgets. It serves as the workspace where you connect widgets together to form a data analysis pipeline. You can add widgets to the canvas, rearrange them, and connect them to create a flow of data processing from input to output.

2. **Widgets: -**

Widgets are graphical elements that perform specific tasks or operations on data. When you open Orange, you will typically see a blank canvas where you can drag and drop widgets to create your analysis workflow.



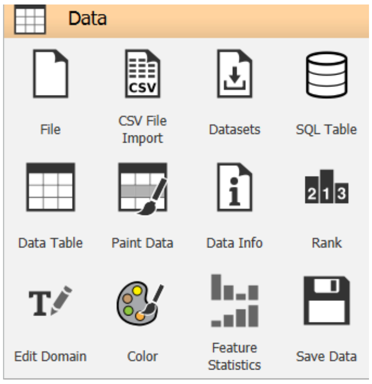
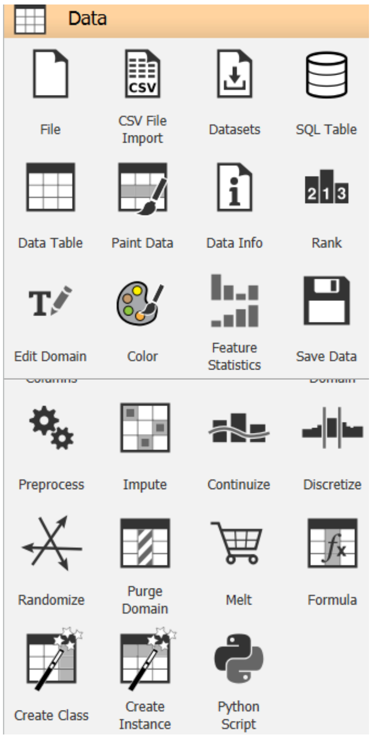
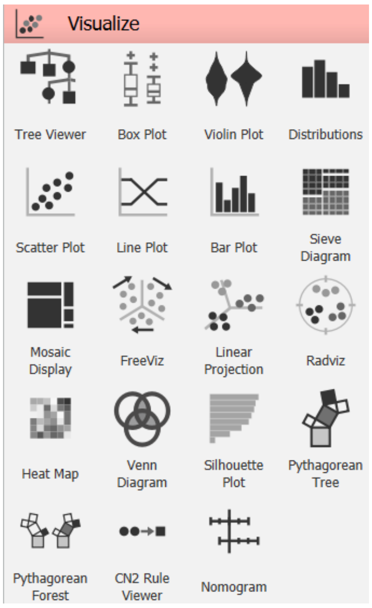
Select a widget to show its description.

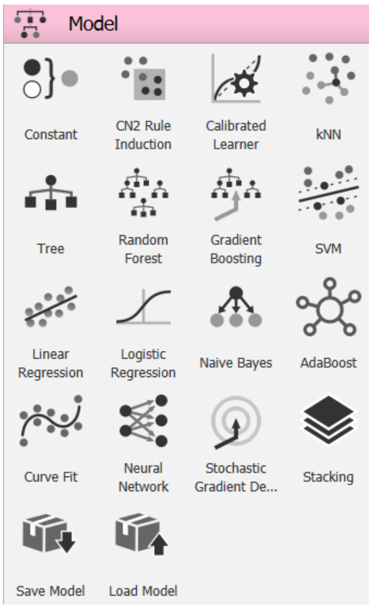
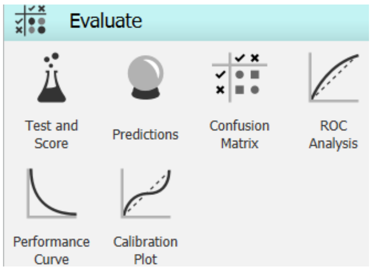
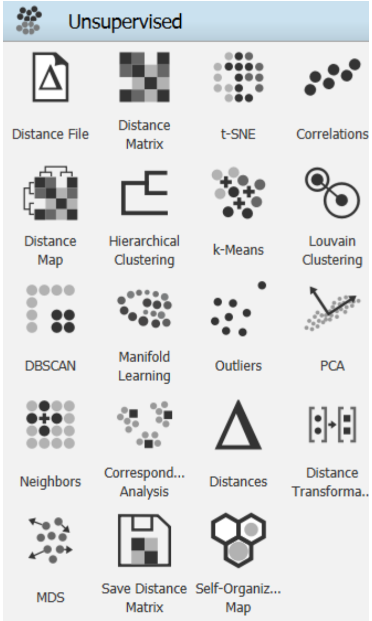
See workflow examples, YouTube tutorials,
or open the welcome screen.

Fig.4.3: Blank Canva

3. **Connectors:-** Connectors are lines that link widgets together on the canvas. They represent the flow of data from one widget to another, indicating how the output of one widget is used as input for another.

4.6. DEFAULT WIDGET CATALOG

Data Widgets	 Data File CSV File Import Datasets SQL Table Data Table Paint Data Data Info Rank Edit Domain Color Feature Statistics Save Data ***Fig.4.4***	These widgets are used for data manipulation, i.e., ● File widget reads the input data file (data table with data instances) and sends the dataset to its output channel. ● Data Table- Displays attribute-value data in a spreadsheet. ● SQL Table - Reads data from an SQL database.
Transform Widgets	 Data File CSV File Import Datasets SQL Table Data Table Paint Data Data Info Rank Edit Domain Color Feature Statistics Save Data Preprocess Impute Continuize Discretize Randomize Purge Domain Melt Formula Create Class Create Instance Python Script ***Fig.4.5***	The Transform Widget in Orange is a powerful tool used for data transformation tasks within your analysis workflow. It allows you to apply various transformations to your dataset
Visualize Widgets	 Visualize Tree Viewer Box Plot Violin Plot Distributions Scatter Plot Line Plot Bar Plot Sieve Diagram Mosaic Display FreeViz Linear Projection Radviz Heat Map Venn Diagram Silhouette Plot Pythagorean Tree Pythagorean Forest CN2 Rule Viewer Nomogram ***Fig. 4.6***	Widgets for visualizing data, including scatter plots, bar charts, and heatmaps.

Model Widgets	 Model Constant CN2 Rule Induction Calibrated Learner kNN Tree Random Forest Gradient Boosting SVM Linear Regression Logistic Regression Naive Bayes AdaBoost Curve Fit Neural Network Stochastic Gradient De... Stacking Save Model Load Model ***Fig. 4.7***	Widgets that enable users to apply machine learning algorithms, including classification, regression, clustering, and anomaly detection, to build predictive models and analyze data patterns.
Evaluate Widgets	 Evaluate Test and Score Predictions Confusion Matrix ROC Analysis Performance Curve Calibration Plot ***Fig. 4.8***	Widgets for evaluating model performance through techniques like cross-validation, confusion matrices etc.
Unsupervised Widgets	 Unsupervised Distance File Distance Matrix t-SNE Correlations Distance Map Hierarchical Clustering k-Means Louvain Clustering DBSCAN Manifold Learning Outliers PCA Neighbors Correspond... Analysis Distances Distance Transforma... MDS Save Distance Matrix Self-Organiz... Map ***Fig. 4.9***	These widgets facilitate exploratory data analysis and pattern recognition without labeled data, enabling tasks such as clustering, dimensionality reduction, and association rule mining.

4.7. KEY DOMAINS OF AI WITH ORANGE DATA MINING TOOL

1. **Data science with Orange Data Mining**
2. **Computer Vision with Orange Data Mining**
3. **Natural Language Process with Orange Data Mining**

4.7.1. Data Science with Orange

Let us begin our journey by diving into the vast ocean of data, where we will learn the fundamental principles of Data Science. Through hands-on exercises with Orange data mining, we will uncover the secrets hidden within datasets, data visualization and how data is classified using classification models.

You want to add a special touch to someone's day with a bouquet, but wouldn't it be even more delightful if you could surprise with their favorite kind of flower? While you can easily visit a flower shop and pick out some iris flowers, there is a catch - iris flowers come in various shapes and sizes. Specifically, when it comes to the violet-colored iris, there are three main types: *Iris Setosa, Iris Versicolor, and Iris Virginica.*

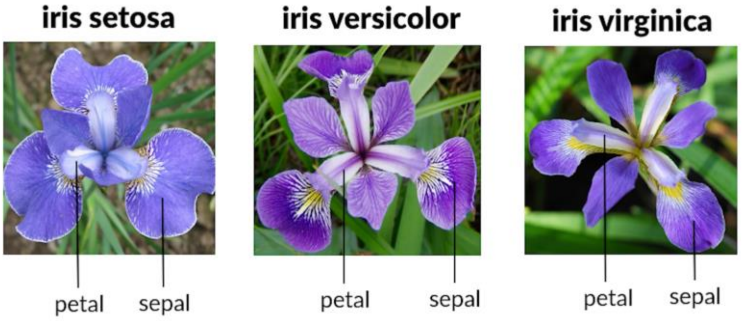


Fig. 4.10: Iris Flower Types

At first glance, distinguishing between these varieties isn't straightforward. However, the key differences lie in the dimensions of the petal and sepal of each flower. To make an informed choice and ensure the surprise is just right, we need data - specifically, measurements of the length and width of the sepal and petal for each iris variety. This data will help us differentiate between the different types of iris flowers accurately. To gather this data efficiently and visualize it for better understanding, we'll turn to Orange data mining tool. With its data visualization capabilities, we can explore the dimensions of iris flowers. Let's dive in and use Orange to enhance our flower surprise with a personal touch!

Data Visualization: - Exploring Iris Flower Dimensions with Orange Data Mining

Step 1- Launch Orange Data Mining Software

Once the software is launched, you will be greeted with a blank canvas where you can construct your analysis workflow.

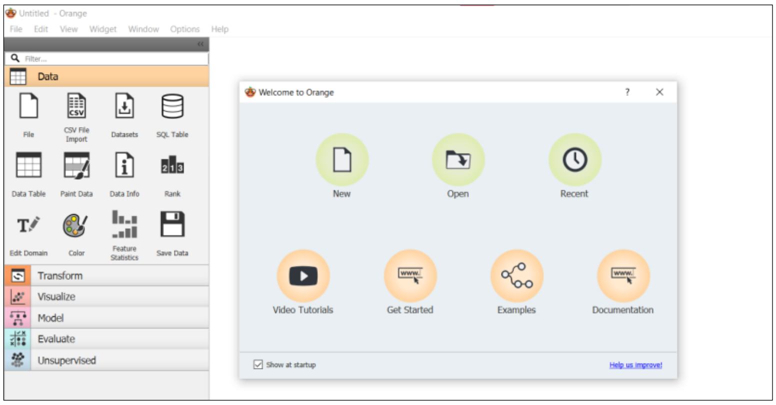
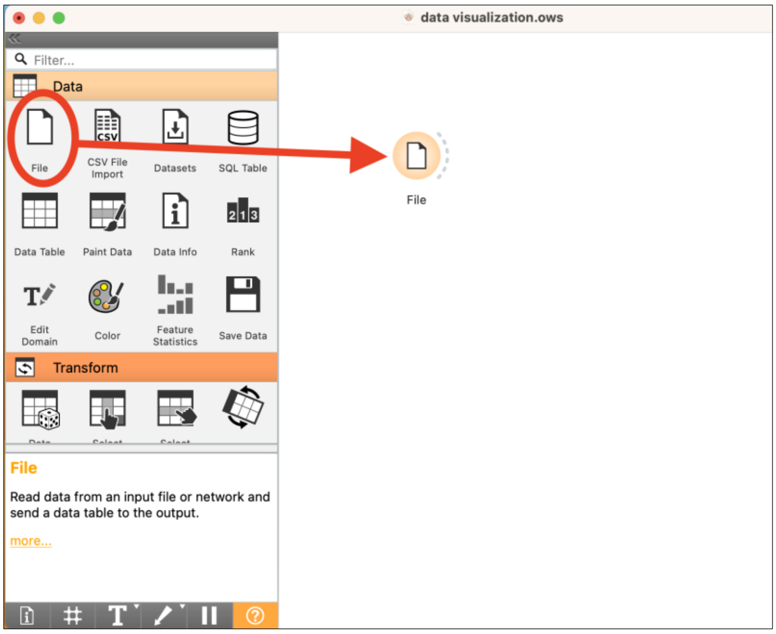


Fig. 4.11: Launch orange data mining

Step 2 – Select the Data Widget

Load the Iris Dataset, Drag the "File" widget from the widget panel onto the blank canvas



File

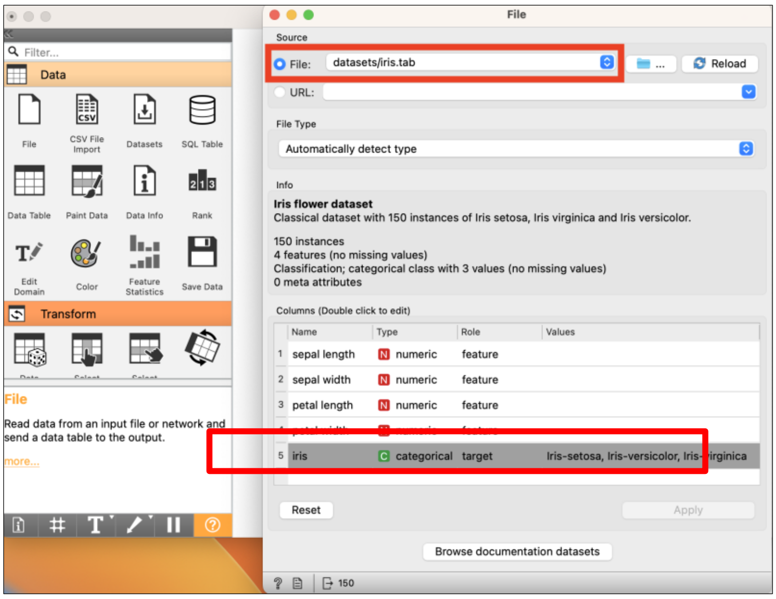
Read data from an input file or network and send a data table to the output.

Fig. 4.12: select the data widget

Step 3 - Load the iris Dataset

Double Click on the "File" widget and select the iris data set.

Change the Role of the column "iris" which has the values of flowers as "**target**"



Read data from an input file or network and send a data table to the output.

● File: datasets/iris.tab

○ URL:

File Type

Automatically detect type

Info

Iris flower dataset
Classical dataset with 150 instances of Iris setosa, Iris virginica and Iris versicolor.

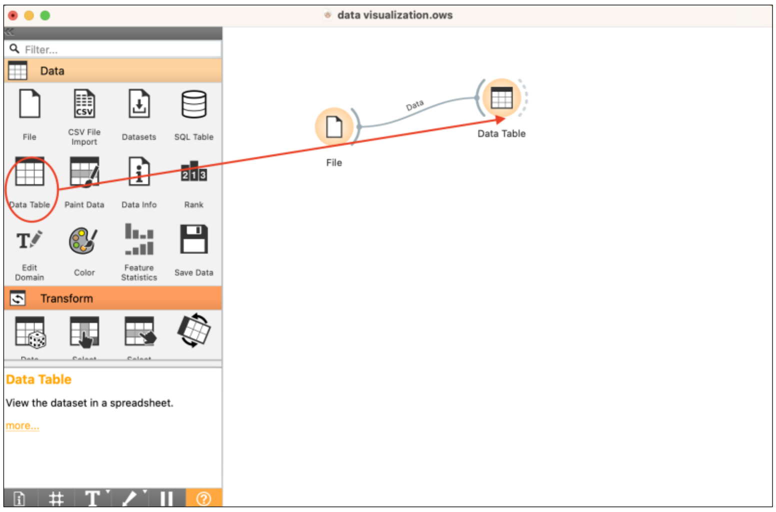
150 instances
4 features (no missing values)
Classification; categorical class with 3 values (no missing values)
0 meta attributes

Columns (Double click to edit)

	Name	Type	Role	Values
1	sepal length	N numeric	feature	
2	sepal width	N numeric	feature	
3	petal length	N numeric	feature	
4	petal width	N numeric	feature	
5	iris	C categorical	target	Iris-setosa, Iris-versicolor, Iris-virginica

Fig. 4.13: Load Iris dataset

Step 4 - Display the Dataset in a Data Table- Drag and drop the "Data Table" widget onto the canvas. Connect the output of the "File" widget to the input of the "Data Table" widget by dragging the connector from one widget to the other.



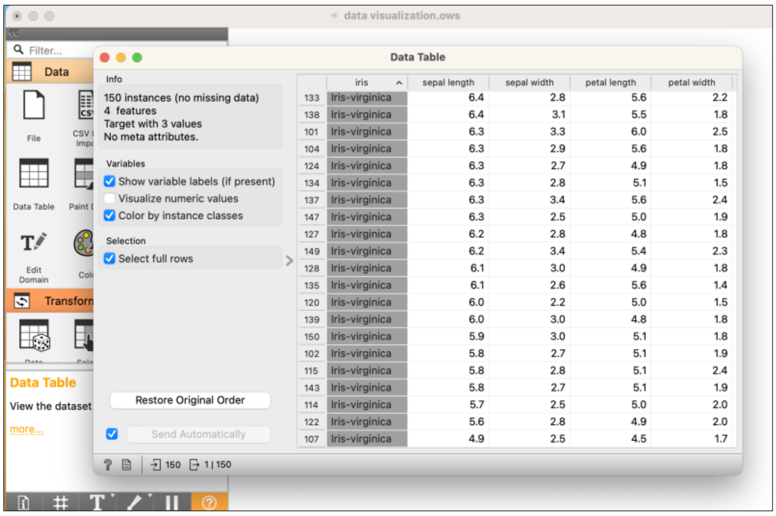
Data Table

View the dataset in a spreadsheet.

more...

Fig. 4.14: Display the dataset

Step 5 – Explore the dataset Double-click on the "Data Table" widget to open it. You will now see a tabular view of the iris dataset, displaying 150 samples of iris flower dimensions, including the length and width of the sepal and petal.



Info

150 instances (no missing data)
4 features
Target with 3 values
No meta attributes.

Variables

☑ Show variable labels (if present)
□ Visualize numeric values
☑ Color by instance classes

Selection

☑ Select full rows

Restore Original Order

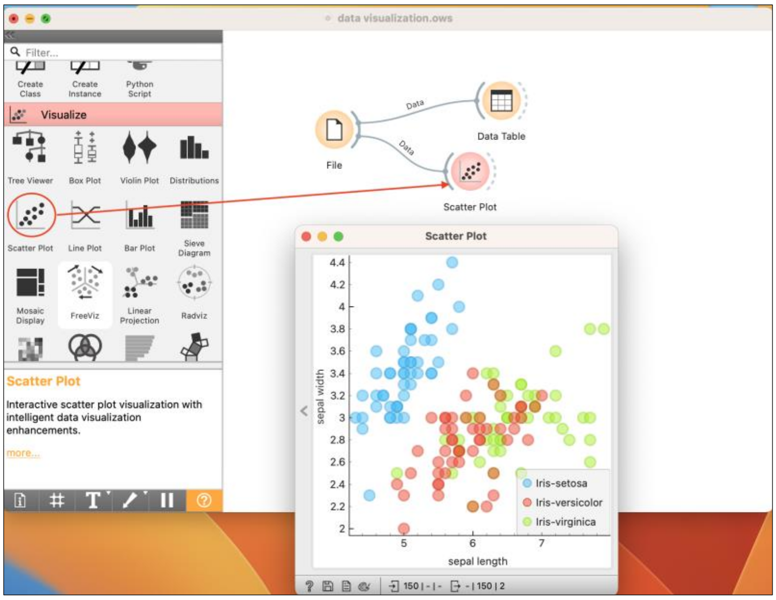
☑ Send Automatically

150 | 1 | 150

	iris	sepal length	sepal width	petal length	petal width
133	Iris-virginica	6.4	2.8	5.6	2.2
138	Iris-virginica	6.4	3.1	5.5	1.8
101	Iris-virginica	6.3	3.3	6.0	2.5
104	Iris-virginica	6.3	2.9	5.6	1.8
124	Iris-virginica	6.3	2.7	4.9	1.8
134	Iris-virginica	6.3	2.8	5.1	1.5
137	Iris-virginica	6.3	3.4	5.6	2.4
147	Iris-virginica	6.3	2.5	5.0	1.9
127	Iris-virginica	6.2	2.8	4.8	1.8
149	Iris-virginica	6.2	3.4	5.4	2.3
128	Iris-virginica	6.1	3.0	4.9	1.8
135	Iris-virginica	6.1	2.6	5.6	1.4
120	Iris-virginica	6.0	2.2	5.0	1.5
139	Iris-virginica	6.0	3.0	4.8	1.8
150	Iris-virginica	5.9	3.0	5.1	1.8
102	Iris-virginica	5.8	2.7	5.1	1.9
115	Iris-virginica	5.8	2.8	5.1	2.4
143	Iris-virginica	5.8	2.7	5.1	1.9
114	Iris-virginica	5.7	2.5	5.0	2.0
122	Iris-virginica	5.6	2.8	4.9	2.0
107	Iris-virginica	4.9	2.5	4.5	1.7

Fig. 4.15: Explore the dataset

Step 6- Visualize the Data with a Scatter Plot
Drag and drop the "Scatter Plot" widget onto the canvas. Connect the output of the "File" widget to the input of the "Scatter Plot" widget. By doing so, you are instructing Orange to use the iris dataset as input for generating a scatter plot.



Scatter Plot

Interactive scatter plot visualization with intelligent data visualization enhancements.

more...

Fig. 4.16: Visualize the data

Step 7- Interpret the Scatter Plot

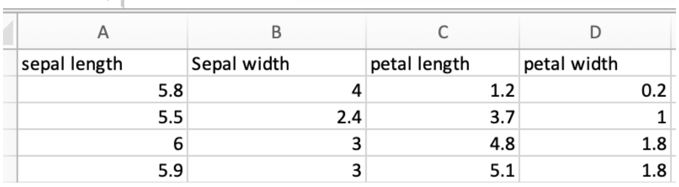
Once the scatter plot is generated, you'll see a graphical representation of the iris Dataset. The scatter plot will display the relationship between two selected variables, such as sepal length and sepal width, with each point representing an individual iris sample. You can further customize the scatter plot and explore other visualizations using additional widgets available in Orange.

You can experiment with other widgets available in Orange for data visualization, such as histograms, box plots, or parallel coordinate plots. Each visualization offers unique insights into the iris dataset, helping you understand the distribution and patterns of flower dimensions.

4.7.1.1. CLASSIFICATION: -

Imagine you have just returned from the flower shop with a bunch of irises, and you have meticulously measured the dimensions of their petals and sepals. Now, armed with these data features, you are ready to determine the type of iris flower - whether it is iris setosa, iris versicolor, or iris virginica. We will collect this testing data and feed it into a spreadsheet for further analysis.

Step 1: Prepare Testing Data in Spreadsheet



A	B	C	D
sepal length	Sepal width	petal length	petal width
5.8	4	1.2	0.2
5.5	2.4	3.7	1
6	3	4.8	1.8
5.9	3	5.1	1.8

Fig. 4.17: creating Testing dataset in excel

- Create columns in a new spreadsheet for sepal length, sepal width, petal length, and petal width, matching the field names in the training data (e.g., "sepal length" in lowercase).
- Then, enter the measured dimensions of the petals and sepals for each iris sample, ensuring consistency with the units used in the training data (e.g., centimeters).

Step 2: Classification with Orange Data Mining

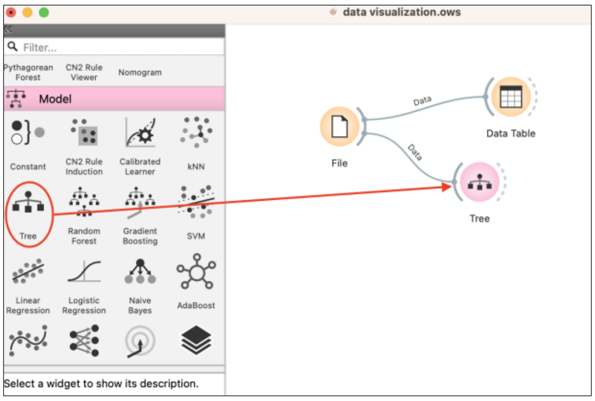


Fig. 4.18: Display the dataset

- Launch the Orange data mining software on your computer.
- Drag the "Tree" widget from the widget panel onto the blank canvas.
- Connect the "Tree" widget to the "File" widget containing the training data(iris inbuilt training data which we used in data visualization) by dragging a connector from the output of the "File" widget to the input of the "Tree" widget.

Note:- "Data table" widget just shows the content of the data in tabular form, it is optional to connect or not.

Step 3: Perform Classification

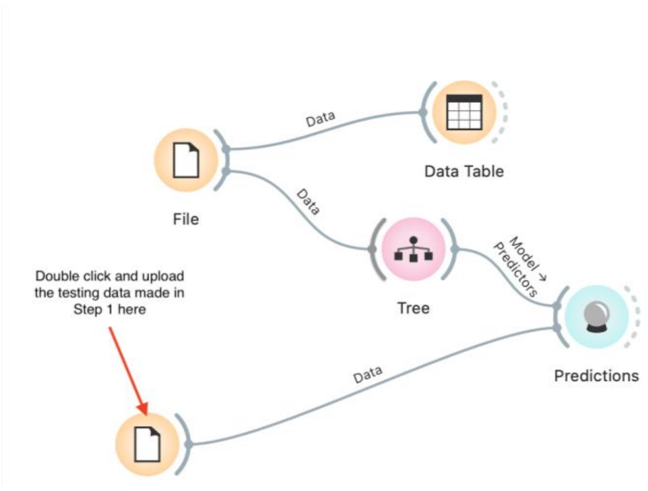
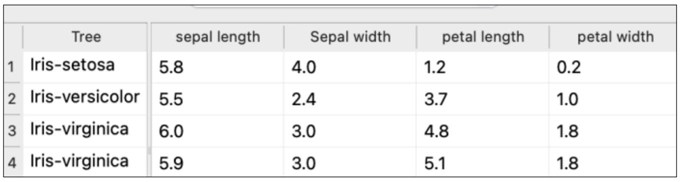

Data
Data Table
File
Data
Tree
Model → Predictors
Double click and upload the testing data made in Step 1 here
Predictions
Data

Fig. 4.19

- Drag and drop the "Predictions" widget onto the canvas. Connect the output of the "File" widget (containing the training data) to the input of the "Predictions" widget.
- Drag another "File" widget onto the canvas and upload the iris testing dataset created in the spreadsheet.
- Connect the output of the second "File" widget (containing the testing data) to the input of the "Predictions" widget. This instructs Orange to use the trained model to classify the samples in testing dataset.

Step 4: Interpret Results



	Tree	sepal length	Sepal width	petal length	petal width
1	Iris-setosa	5.8	4.0	1.2	0.2
2	Iris-versicolor	5.5	2.4	3.7	1.0
3	Iris-virginica	6.0	3.0	4.8	1.8
4	Iris-virginica	5.9	3.0	5.1	1.8

Fig. 4.20: Interpret Results

- Click on the prediction widget
- The "Predictions" widget displays the predicted class labels for each iris sample in the testing dataset, allowing you to assess the accuracy of the classification model.

By following the above steps, you have successfully employed Orange data mining software to classify iris flower types using testing data. This demonstrates the practical application of machine learning techniques in real-world scenarios, offering insights into the classification process and its accuracy.

4.7.1.2. Evaluating the Classification Model with Orange

Now that we have classified iris flower types using our model, it is crucial to assess its performance. Evaluation metrics such as accuracy, precision, recall, and F1 score provide insights into the effectiveness of the classification model. Additionally, we will utilize the confusion matrix with cross-validation to gain a deeper understanding of the model's performance. **Cross-validation**, sometimes called **rotation estimation,** includes resampling and sample splitting methods that use different portions of the data to test and train a model on different iterations.

Step 1: Perform Evaluation: -

- Begin by adding the "File" widget to the canvas and connect it to the output of the classification tree model. This will provide the dataset used in the earlier training phase.
- Connect the same "File" widget to the input of the "Test and Score" widget. This widget evaluates the model's performance and calculates metrics such as accuracy, precision, recall, and F1 score.
- Double-click on the "Test and Score" widget to view the evaluation metrics. Orange utilizes cross-validation, splitting the data into 10 subsets for training and testing. This process is repeated multiple times to ensure robust evaluation.

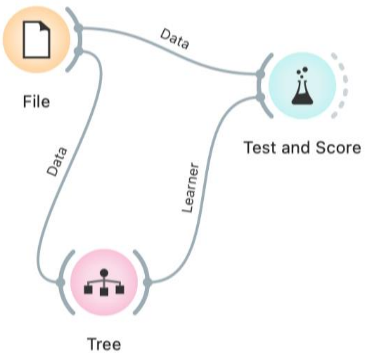
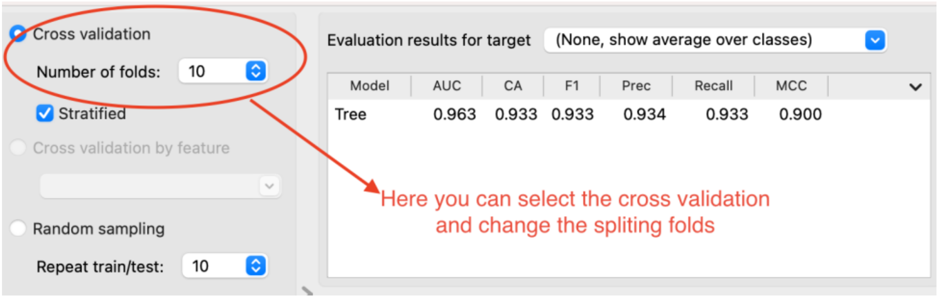


Fig. 4.21



○ Cross validation

Number of folds: 10

☑ Stratified

○ Cross validation by feature

○ Random sampling

Repeat train/test: 10

Evaluation results for target (None, show average over classes)

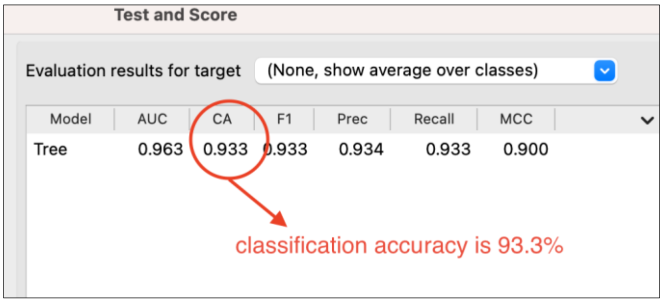
Model	AUC	CA	F1	Prec	Recall	MCC
Tree	0.963	0.933	0.933	0.934	0.933	0.900

Here you can select the cross validation and change the spliting folds

Fig. 4.22

Step 2: Interpret Evaluation Metrics

- Analyze the evaluation metrics displayed by the "Test and Score" widget. The accuracy metric indicates the percentage of correctly classified instances, which in our case is 93%.
- However, to gain deeper insights, we will examine precision, recall, and F1 score. Precision measures the proportion of true positives among all instances classified as positive, while recall measures the proportion of true positives correctly identified. The F1 score is the harmonic mean of precision and recall.
- These metrics collectively provide a comprehensive understanding of the model's performance across different classes.



Test and Score

Evaluation results for target (None, show average over classes)

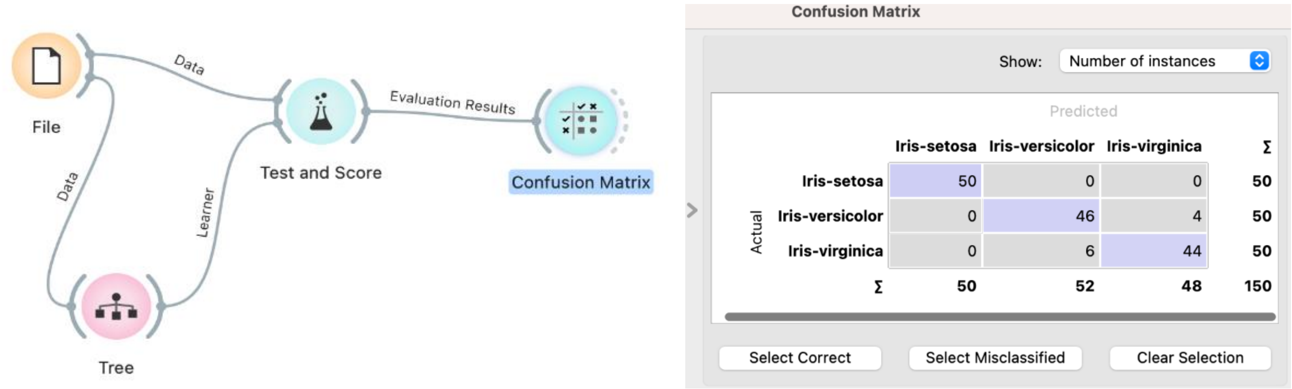
Model	AUC	CA	F1	Prec	Recall	MCC
Tree	0.963	0.933	0.933	0.934	0.933	0.900

classification accuracy is 93.3%

Fig. 4.23: Evaluation Metric

Step 3: Utilize Confusion Matrix

- Add the "Confusion Matrix" widget to the canvas and connect it to the output of the "Test and Score" widget.
- Double-click on the "Confusion Matrix" widget to view the matrix. This matrix provides a detailed breakdown of the model's classifications, showing the number of true positives, true negatives, false positives, and false negatives for each class.



Confusion Matrix

Show: Number of instances

Predicted

Actual	Iris-setosa	Iris-versicolor	Iris-virginica	Σ
Iris-setosa	50	0	0	50
Iris-versicolor	0	46	4	50
Iris-virginica	0	6	44	50
Σ	50	52	48	150

Fig. 4.24: Confusion Matrix

- Interpret the confusion matrix to identify any patterns or areas where the model may be struggling. For instance, while the model accurately identifies iris setosa, it may struggle with distinguishing between versicolor and virginica.

By evaluating our classification model using Orange, we've gained valuable insights into its performance. Through metrics such as accuracy, precision, recall, and F1 score, as well as the confusion matrix with cross-validation, we can assess the model's strengths and weaknesses. This evaluation process empowers us to make informed decisions and further refine our model for improved performance.

Practical Activity

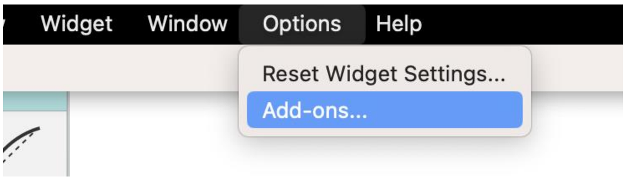
- Your task is to conduct a Data Science project aiming to differentiate between fruits and vegetables based on their nutritional characteristics.
- Begin by gathering a dataset containing nutritional information for a variety of fruits and vegetables, including features such as energy (kcal/kJ), water content (g), protein (g), total fat (g), carbohydrates (g), fiber (g), sugars (g), calcium (mg), iron (mg), magnesium (mg), phosphorus (mg), potassium (mg), and sodium (g).
- You can explore reputable sources like Kaggle for this data.
- Next, split the dataset into training and testing subsets. Utilize classification algorithms available in Orange to train models on the training data, then evaluate their performance using the testing data.
- Utilize evaluation metrics such as accuracy, precision, recall, and F1-score to assess the models' effectiveness and compare their performance to identify the most effective one.

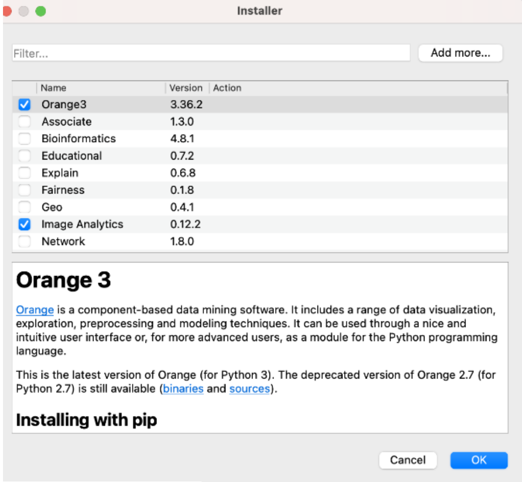
4.7.2. Computer Vision with Orange

In the vast landscape of data, images represent a rich and diverse source of information. With advancements in computer vision and machine learning, we can now extract valuable insights from images using tools like Orange. Let us now explore how Orange transforms images into numerical representations and enables machine learning on image data.

Step 1: Install Image Analytics Add-On

- To get started, we need to install the Image Analytics add-on in Orange. Go to the "Options" menu, click on "Add-ons," and install Image Analytics. Restart Orange to enable the add-on to appear in the widget panel.





	Name	Version	Action
☑	Orange3	3.36.2	
□	Associate	1.3.0	
□	Bioinformatics	4.8.1	
□	Educational	0.7.2	
□	Explain	0.6.8	
□	Fairness	0.1.8	
□	Geo	0.4.1	
☑	Image Analytics	0.12.2	
□	Network	1.8.0	

Orange 3

Orange is a component-based data mining software. It includes a range of data visualization, exploration, preprocessing and modeling techniques. It can be used through a nice and intuitive user interface or, for more advanced users, as a module for the Python programming language.

This is the latest version of Orange (for Python 3). The deprecated version of Orange 2.7 (for Python 2.7) is still available (binaries and sources).

Installing with pip

Fig. 4.25: Install Image Analytics Add on

Step 2: Example Scenario

- Let us dive into an example scenario where we want to build a model to cluster unlabeled images of dogs and cats. We will start by obtaining a dataset containing images of dogs and cats and saving it on our computer. Click on the link below to obtain the dataset:

https://drive.google.com/drive/folders/1tLbVXWkKzcp6O_-FAvn9usEWuoF3jTp3?usp=sharing

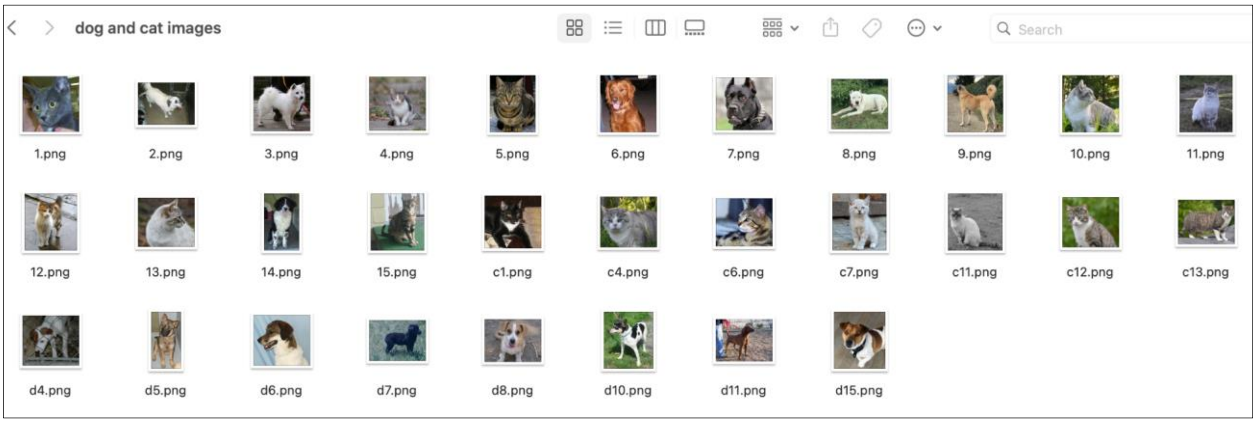
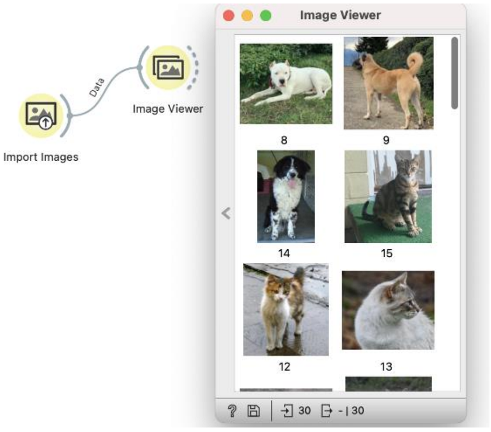


Fig. 4.26: Dog and Cat images

Step 3: Import Images

- Drag and drop the "Import Images" widget onto the canvas and upload the dataset folder containing images of dogs and cats.

Step 4: Image Visualization



- Add the "Image Viewer" widget to the canvas and connect it to the output of the "Import Images" widget. Double-click on the "Image Viewer" to visualize the images in the dataset.

Fig. 4.27: Image Visualization

Step 5: Image Embedding

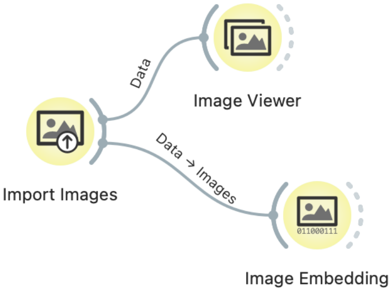
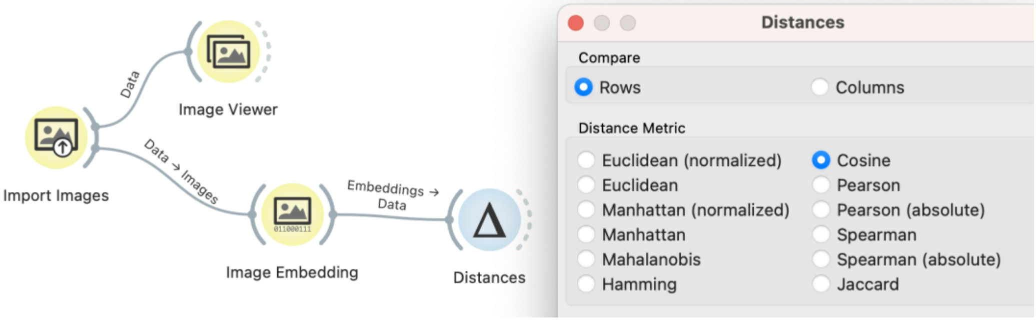


Fig. 4.28: Image Embedding

Now, we need to transform the raw images into numerical representations using image embeddings. Connect the "Image Embedding" widget to the "Import Images" widget. This widget sends information to the server and computes embeddings remotely using a deep neural network trained on millions of real-life images.

Step 6: Computing Similarities

- We will compare the embeddings of images and compute their similarities. Connect the "Distance" widget to the "Image Embedding" output. Double-click on the "Distance" widget and select the cosine option, which typically works best with images.



Distances

Compare

● Rows ○ Columns

Distance Metric

○ Euclidean (normalized) ● Cosine
○ Euclidean ○ Pearson
○ Manhattan (normalized) ○ Pearson (absolute)
○ Manhattan ○ Spearman
○ Mahalanobis ○ Spearman (absolute)
○ Hamming ○ Jaccard

Fig. 4.29: Computing Similarities

Step 7: Hierarchical Clustering

- Pass the distance matrix to the "Hierarchical Clustering" algorithm by dragging and dropping the widget onto the canvas.

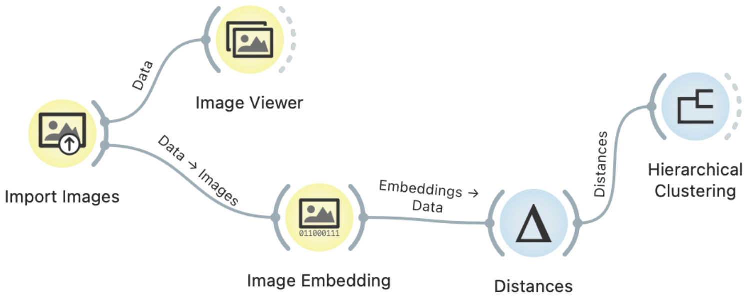


Fig. 4.30: Hierarchical Clustering

- Double-click on the "Hierarchical Clustering" widget to observe the dendrogram, where images are grouped based on their similarities.

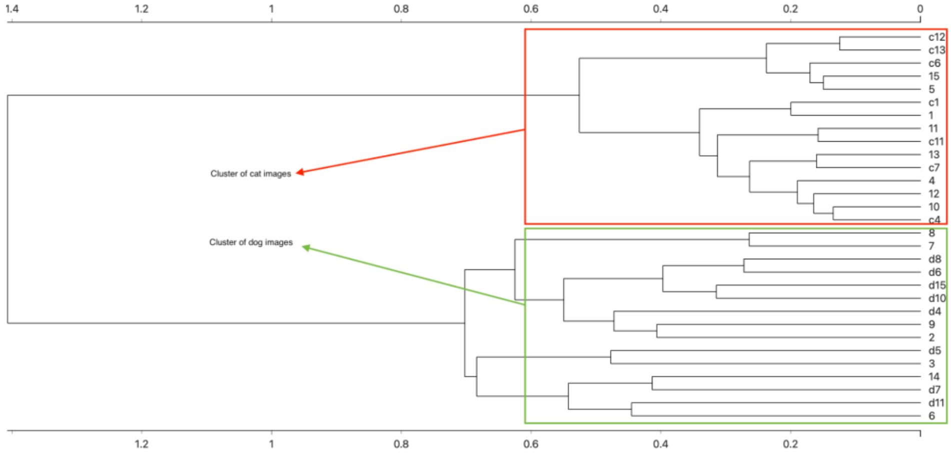
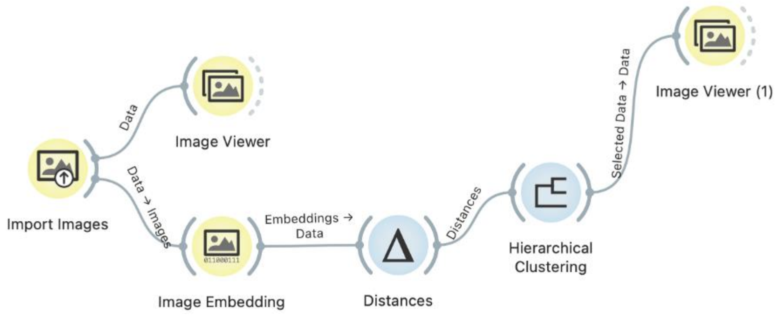


Fig. 4.31: Dendrogram

Step 8: Visualization and Interpretation

- Use the "Image Viewer" widget to visualize each cluster by selecting the corresponding area in the dendrogram.



You will notice that images of dogs and cats are grouped together, demonstrating the effectiveness of image analytics using Orange data mining.

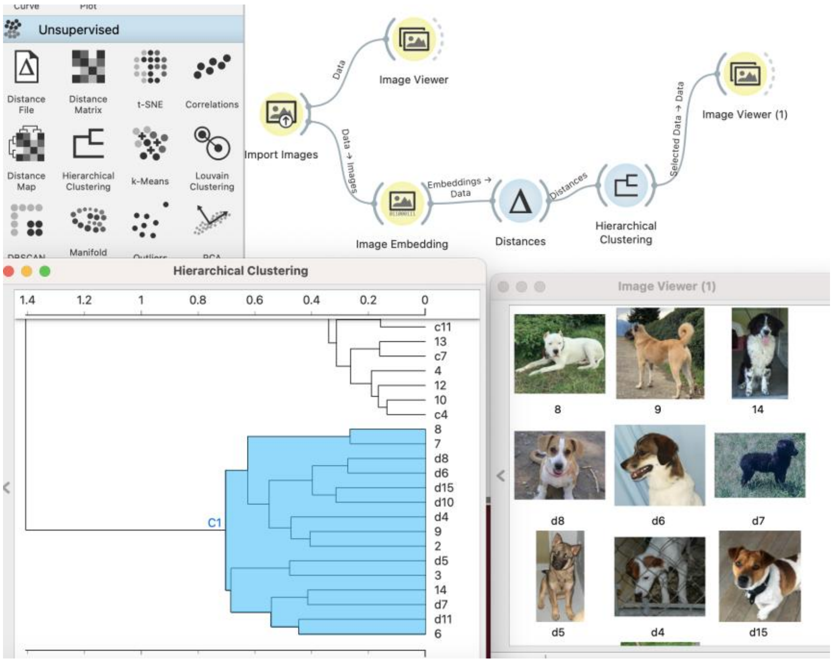


Fig. 4.32: Visualization and Interpretation

With Orange's Image Analytics capabilities, we can unlock the potential of image data and perform sophisticated analysis and machine learning tasks. By leveraging image embeddings and clustering algorithms, we can gain insights and make informed decisions from image datasets, paving the way for innovative applications in various domains.

Practical Activity:

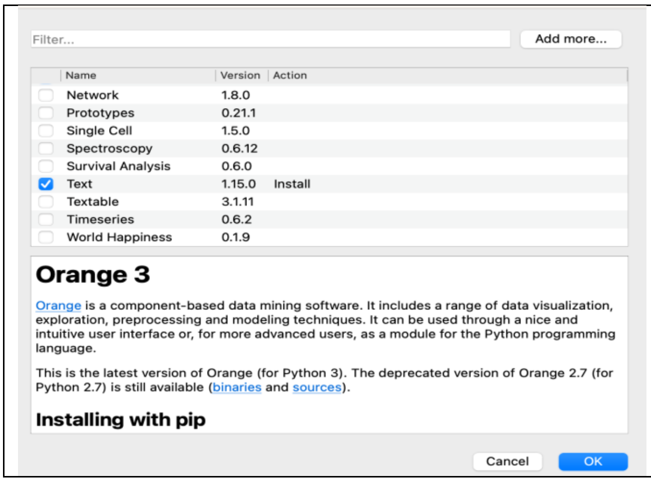
- Cluster images of birds and animals into distinct groups based on their visual characteristics.
- Collect datasets of images containing various species of birds and animals.
- Ensure that each dataset contains a sufficient number of images representing different species within the respective categories.
- Import the collected image datasets into Orange Data Mining.
- Apply clustering algorithms to group similar images together based on their numerical representations.
- Analyze the clustering results and interpret the grouping of images into different clusters.
- Identify any patterns or similarities observed within each cluster and between clusters.

4.7.3. Natural Language Processing with Orange

Now that we have learned about analyzing data in spreadsheets and images, let us move on to working with text using Orange Data Mining. Natural Language Processing (NLP) helps us understand and learn from written words, like finding patterns in documents. With Orange, we will do different tasks in NLP, which can help us in lots of ways, like analyzing and understanding text better.

Step 1: Install Text Add-On

First, we need to install the Text add-on. Navigate to the "Options" menu, select "Add-ons," and choose "Text." Restart Orange Data Mining to activate the add-on.



	Name	Version	Action
□	Network	1.8.0	
□	Prototypes	0.21.1	
□	Single Cell	1.5.0	
□	Spectroscopy	0.6.12	
□	Survival Analysis	0.6.0	
☑	Text	1.15.0	Install
□	Textable	3.1.11	
□	Timeseries	0.6.2	
□	World Happiness	0.1.9	

Orange 3

Orange is a component-based data mining software. It includes a range of data visualization, exploration, preprocessing and modeling techniques. It can be used through a nice and intuitive user interface or, for more advanced users, as a module for the Python programming language.

This is the latest version of Orange (for Python 3). The deprecated version of Orange 2.7 (for Python 2.7) is still available (binaries and sources).

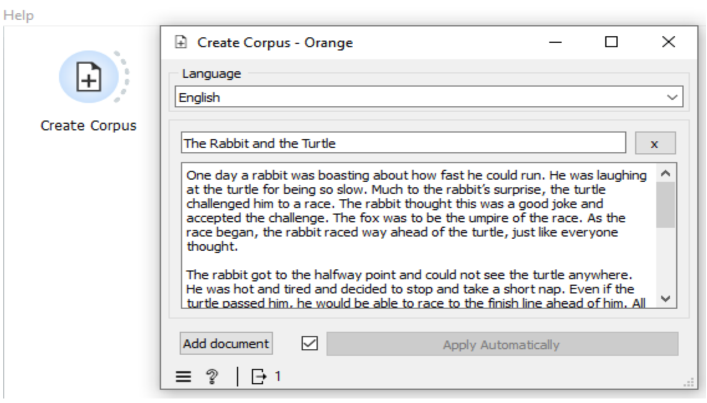
Installing with pip

Fig. 4.33: Install Text Add-on

Step 2: Load or Create Textual Data

Whether we have existing textual data or want to create your own corpus, Orange provides the tools to handle both scenarios. Drag and drop the "Corpus" widget to load data (or) use the "Create Corpus" widget to input your own text.

Double-click on the "Create Corpus" widget to add textual data. You can input any text you wish to analyze, such as articles, reviews, or documents.



Language: English

The Rabbit and the Turtle

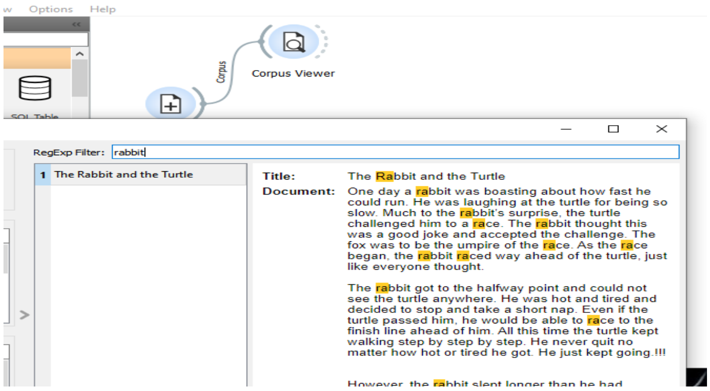
One day a rabbit was boasting about how fast he could run. He was laughing at the turtle for being so slow. Much to the rabbit's surprise, the turtle challenged him to a race. The rabbit thought this was a good joke and accepted the challenge. The fox was to be the umpire of the race. As the race began, the rabbit raced way ahead of the turtle, just like everyone thought.

The rabbit got to the halfway point and could not see the turtle anywhere. He was hot and tired and decided to stop and take a short nap. Even if the turtle passed him, he would be able to race to the finish line ahead of him. All

Fig. 4.34: Load or create Textual Data

Step 3: Visualize Text with Corpus Viewer

Add the "Corpus Viewer" widget to the canvas and connect it with the output of the "Create Corpus" widget. The Corpus Viewer allows us to browse through the text and search for specific words, which it highlights within the corpus



RegExp Filter: rabbit

1 The Rabbit and the Turtle

Title: The Rabbit and the Turtle

Document: One day a rabbit was boasting about how fast he could run. He was laughing at the turtle for being so slow. Much to the rabbit's surprise, the turtle challenged him to a race. The rabbit thought this was a good joke and accepted the challenge. The fox was to be the umpire of the race. As the race began, the rabbit raced way ahead of the turtle, just like everyone thought.

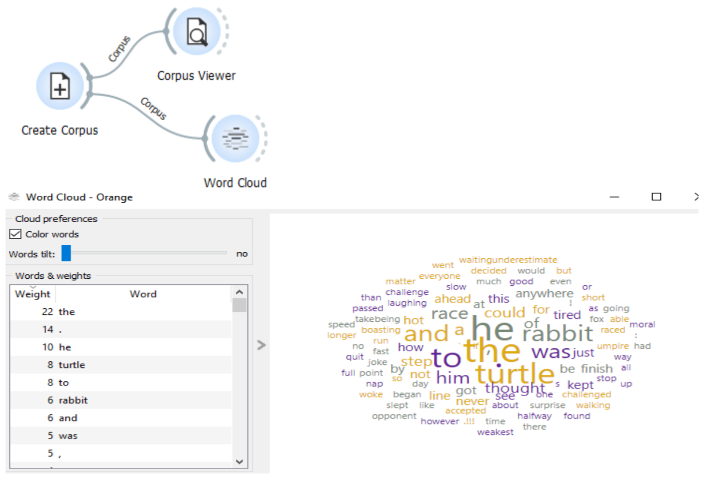
The rabbit got to the halfway point and could not see the turtle anywhere. He was hot and tired and decided to stop and take a short nap. Even if the turtle passed him, he would be able to race to the finish line ahead of him. All this time the turtle kept walking step by step by step. He never quit no matter how hot or tired he got. He just kept going.!!!

However, the rabbit slept longer than he had

Fig. 4.35: Visualize text with corpus viewer

Step 4: Visualize Word Frequencies with Word Cloud

Connect the output of the "Corpus" widget to the "Word Cloud" widget. The Word Cloud visually represents word frequencies in a cloud format, with more frequent words appearing larger. This visualization provides an initial glimpse into the prominent themes or topics within the text.



Word Cloud - Orange

Cloud preferences

☑ Color words

Words tilt: no

Words & weights

Weight	Word
22	the
14	.
10	he
8	turtle
8	to
6	rabbit
6	and
5	was
5	,

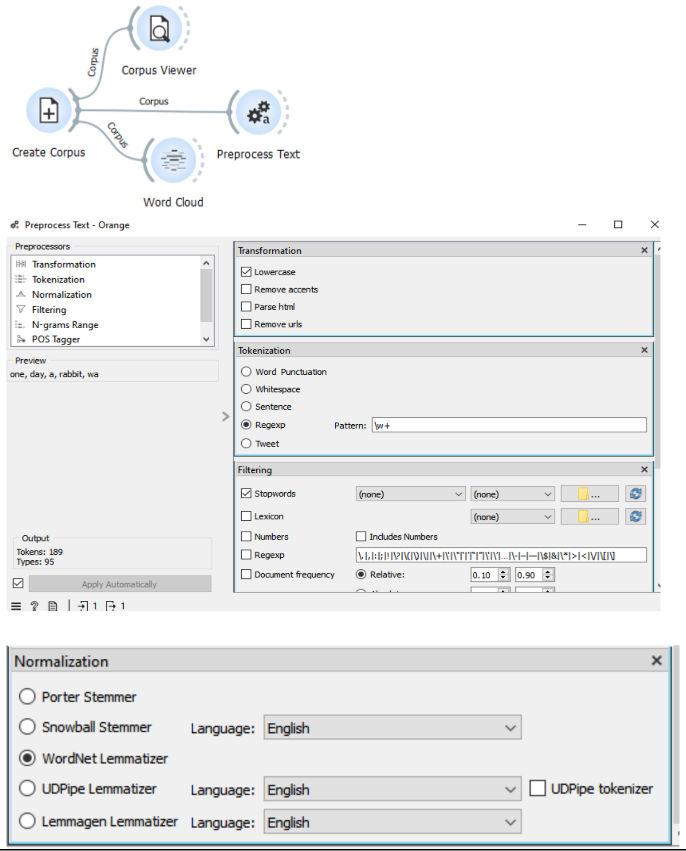
Fig. 4.36: Visualize word frequencies with cloud word

Step 5: Preprocess Text

It is essential to preprocess the text to remove noise and irrelevant information (.,!@#$). Use the "Preprocess Text" widget for this task. Connect this widget to the output of the “Corpus” widget.

“**Preprocess text**” widget performs text normalization by converting text to lowercase, tokenizing it into individual words, removing punctuation, and filtering out stop words.

Optionally, you can choose to perform stemming or lemmatization to further refine the text.



Preprocess Text - Orange

Preprocessors: Transformation, Tokenization, Normalization, Filtering, N-grams Range, POS Tagger

Preview: one, day, a, rabbit, wa

Output: Tokens: 189, Types: 95

☑ Apply Automatically

Transformation
- ☑ Lowercase
- □ Remove accents
- □ Parse html
- □ Remove urls

Tokenization
- ○ Word Punctuation
- ○ Whitespace
- ○ Sentence
- ● Regexp — Pattern: \w+
- ○ Tweet

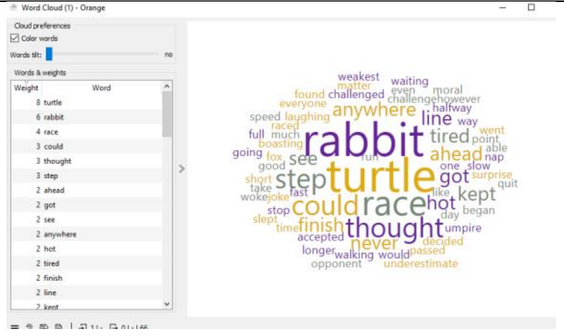
Filtering
- ☑ Stopwords (none) (none)
- □ Lexicon (none)
- □ Numbers □ Includes Numbers
- □ Regexp \.|,|:|;|!|\?|\(|\)|\||\+|\'|\"|‘|’|“|”|\'|\’|…|\-|–|—|\$|&|*|>|<|\/|\[|\]
- □ Document frequency ● Relative: 0.10 0.90

Normalization
- ○ Porter Stemmer
- ○ Snowball Stemmer Language: English
- ● WordNet Lemmatizer
- ○ UDPipe Lemmatizer Language: English □ UDPipe tokenizer
- ○ Lemmagen Lemmatizer Language: English

Fig. 4.37: Preprocess Text

Step 6: Visualize Cleaned Text with Word Cloud

Connect the output of the "Preprocess Text" widget to the "Word Cloud" widget to visualize the cleaned text data.



Word Cloud (1) - Orange

Cloud preferences

☑ Color words

Words tilt: no

Words & weights

Weight	Word
8	turtle
6	rabbit
4	race
3	could
3	thought
3	step
2	ahead
2	got
2	see
2	anywhere
2	hot
2	tired
2	finish
2	line
2	kept

Fig. 4.38: Visualize cleaned text with word cloud

- Now, the Word Cloud displays only meaningful words, allowing us to better understand the main themes or topics within the corpus. Here you can see Turtle and rabbit are larger in size because these words are more frequent in the corpus.

With Orange Data Mining, we can leverage the power of NLP to analyze textual data effectively. Through the steps outlined above, we have seen how to load, visualize, preprocess, and analyze textual data effectively. From exploring word frequencies with the Word Cloud to cleaning and refining text through preprocessing, Orange empowers users to derive meaningful insights from text effortlessly.

Practical Activity
Your task is to create your own corpus by selecting a story or article of your choice. Ensure the text is sufficiently long to provide meaningful insights. Once you have your corpus, use Orange Data Mining to perform: - 1. Text normalization techniques like Converting all text to lowercase to ensure consistency, tokenization to split the text into individual words or tokens, removal of stop words to eliminate common words that do not contribute to the meaning of the text 2. Analyze the word cloud to identify which words appear most frequently and gain insights into the main themes or topics of the text. 3. Compare the results of the word cloud analysis before and after applying lemmatization and stemming to observe any differences in the most commonly used words.

EXERCISES

A. Multiple Choice Questions

1. Which widget helps in knowing the Accuracy, Precision, recall and F1 score values?
   a. Confusion matrix b. Test and score c. Word cloud d. Corpus
2. Which widget provides the detailed breakdown of the model's classifications, showing the number of true positives, true negatives, false positives, and false negatives for each class?
   a. Test and score b. Preprocess text c. Confusion Matrix d. Corpus
3. ______________widget performs text normalization by converting text to lowercase, tokenizing it into individual words, removing punctuation, and filtering out stop words.
   Ans- Preprocess text
4. Cross validation can also be called as _______
   a. Rotation estimation b. Cross hybrid c. Revolution Estimation d. Estimation
5. In word cloud, more the frequency of word in corpus larger the size of word would look.
   a. True b. False
6. Which widget is used to transform the raw images into numerical representations?
   a. Image viewer b. Image analytics c. Image Embedding d. Distance
7. Which widget is used to compare the embeddings of images and compute their similarities?
   a. Distance b. Image Embedding c. Preprocess Text d. Image viewer

8. _____________ are lines that link widgets together on the canvas. They represent the flow of data from one widget to another, indicating how the output of one widget is used as input for another.

   Ans- Connectors

9. Suppose you are working as a data scientist for a social media analytics company. The company is interested in analyzing and categorizing images shared on its platform to better understand user behavior and preferences. Which widget will you use to extract numerical representations of images for analysis?

   Ans- Image embedding

B. Short Answer Questions

1. What are the different components of Orange data mining tool?

   Ans- Main components of Orange tools are-

- Blank Canva- The blank canvas is where you build your analysis workflows by dragging and dropping widgets. It serves as the workspace where you connect widgets together to form a data analysis pipeline. You can add widgets to the canvas, rearrange them, and connect them to create a flow of data processing from input to output.
- Widgets - Widgets are graphical elements that perform specific tasks or operations on data. When you open Orange, you will typically see a blank canvas where you can drag and drop widgets to create your analysis workflow.
- Connectors - Connectors are lines that link widgets together on the canvas. They represent the flow of data from one widget to another, indicating how the output of one widget is used as input for another.

2. Explain cross validation.

   Ans- Cross-Validation, sometimes called rotation estimation, includes resampling and sample splitting methods that use different portions of the data to test and train a model on different iterations.

3. What is the purpose of using the “corpus viewer” widget in Orange?

   Ans- The Corpus Viewer allows us to browse through the text and search for specific words, which it highlights within the corpus.

4. How does the "Preprocess Text" widget contribute to text analysis in Orange Data Mining?

   Ans- “Preprocess text” widget performs text normalization by converting text to lowercase, tokenizing it into individual words, removing punctuation, and filtering out stop words.

C. Competency Based Questions:

1. Manya recently discovered an open-source software tool, a component-based visual programming software package that serves multiple purposes including data visualization, machine learning, data mining, and data analysis. Can you name this tool?
   Ans: Orange data mining tool

2. Arun, a teacher wants to educate his students about distinguishing between animals with two legs and those with four legs. He has a folder containing mixed images of animals and wants to organize them into clusters based on this distinction. Which Add-ons should he install in Orange data mining tool to achieve this image differentiation?
   Ans: Arun should install the add-on called "Image Analytics."

3. Sunita enjoys writing and is currently working on a story about nature. Sunita is curious to know about the most frequently used word in her story and seeks assistance in using the Orange data mining tool to determine this. Which tool can she use to analyze her story's word frequency?
   Ans: Sunita can utilize the Word Cloud tool within Orange Data Mining to analyze her story's word frequency. This tool visually displays the frequency of words in a cloud format, with more frequent words appearing larger and more prominent.

4. Rajat and Misty were having an argument on the effectiveness of natural language processing (NLP) techniques in Orange data mining for text analysis tasks. Can you list the applications of NLP techniques in Orange data mining that contribute to text analysis?
   Ans: NLP techniques in Orange data mining are used for tasks such as text preprocessing, sentiment analysis, topic modeling, and text classification. These techniques enable analysts to extract insights from unstructured text data, automate text processing tasks, and gain valuable information from textual sources in diverse domains.

5. Imagine you are a data scientist working at a recycling plant. The plant recently introduced robots equipped with machine learning models to sort incoming materials. These robots categorize items into different bins like plastic, metal, glass, and paper. To improve the robots' accuracy, you plan to use Orange Data Mining to analyze their performance. Which widget within Orange can provide a detailed breakdown of the model's classifications for each material type? Specifically, you are interested in seeing the number of true positives (correctly identified items), true negatives (correctly rejected items), false positives (incorrectly accepted items), and false negatives (incorrectly rejected items) for aluminum and tin cans. What widget can help you achieve this?
   Ans: Confusion matrix

REFERENCES:

1. https://orangedatamining.com/docs/
2. https://orangedatamining.com/blog/
3. https://www.youtube.com/@OrangeDataMining

UNIT 5: Introduction to Big Data and Data Analytics

Title: Introduction to Big Data and Data Analytics

Approach: Team discussion, Web search

Summary: Students will delve into the world of **Big Data**, a game-changer in today's digital age. Students gain insights into the various types of data and their unique characteristics, equipping them to understand how this vast information is managed and analysed. The journey continues as students discover the real-world applications of Big Data and Data Analytics in diverse fields, witnessing how this revolutionary concept is transforming how we approach data analysis to unlock new possibilities.

Learning Objectives:

1. Students will develop an understanding of the concept of Big Data and its development in the new digital era.
2. Students will appreciate the role of big data in AI and Data Science.
3. Students will learn to understand the features of Big Data and how these features are handled in Big Data Analytics.
4. Students will appreciate its applications in various fields and how this new concept has evolved to bring new dimensions to Data Analysis.
5. Students will understand the term mining data streams.

Key Concepts:

1. Introduction to Big Data
2. Types of Big Data
3. Advantages and Disadvantages of Big Data
4. Characteristics of Big Data
5. Big Data Analytics
6. Working on Big Data Analytics
7. Mining Data Streams
8. Future of Big Data Analytics

Learning Outcomes:

Students will be able to –

1. Define Big Data and identify its various types.
2. Evaluate the advantages and disadvantages of Big Data.
3. Recognize the characteristics of Big Data.
4. Explain the concept of Big Data Analytics and its significance.
5. Describe how Big Data Analytics works.
6. Exploring the future trends and advancements in Big Data Analytics.

Prerequisites: Understanding the concept of data and reasonable fluency in the English language.

Demystifying the Big Data Deluge: A Teacher's Guide to Introduction to Big Data and Data Analytics

This lesson plan equips you to guide students through the vast and exciting world of Big Data.

1. Unveiling the Big Data Universe:

- **Engaging Introduction:** Spark curiosity with a thought-provoking question: "How much data is created every day?" Discuss the explosion of data and introduce the concept of Big Data - massive and complex datasets that require specialized tools and techniques for analysis.

2. The Big Data Journey: From Collection to Insights:

- **Big Data Analytics in Action:** Introduce the various stages of Big Data Analytics:
    - Data Collection: Gathering data from diverse sources (e.g., social media, sensors, transactions).
    - Data Storage: Storing massive datasets using specialized solutions (e.g., distributed file systems).
    - Data Processing: Cleaning, organizing, and preparing data for analysis.
    - Data Analysis: Utilizing various techniques (e.g., machine learning, statistics) to extract insights from data.
    - Data Visualization: Presenting data findings in clear and compelling ways (e.g., charts, graphs).
- **Real-World Applications:** Showcase real-world applications of Big Data Analytics across various sectors:
    - Business: Optimizing marketing campaigns, identifying customer trends, and improving operational efficiency.
    - Healthcare: Predicting disease outbreaks, analyzing medical images, and personalizing patient care.
    - Finance: Detecting fraudulent activities, managing risk, and providing personalized financial recommendations.

3. Demystifying Data Types: Structured, Semi-structured, and Unstructured:

- **Data Diversity:** Define and differentiate between the three main data types:
    - Structured Data: Highly organized data with a defined format (e.g., database tables).
    - Semi-structured Data: Partially organized data with some inherent structure (e.g., emails, logs).
    - Unstructured Data: Data with no predefined format (e.g., text, images, videos).

4. Big Data: A Double-Edged Sword:

- **Advantages and Disadvantages:** Discuss the benefits and limitations of Big Data:

Advantages: * Improved decision-making through data-driven insights. * Enhanced efficiency and innovation across various industries. * Potential for breakthroughs in scientific research and social good.

Disadvantages: * Privacy concerns and potential misuse of personal data. * Challenges in data security and ethical considerations. * Requirement for specialized skills and infrastructure for Big Data analysis.

5. The Big Difference: Big Data vs. Traditional Data:

- **Distinguishing Features:** Highlight the defining characteristics of Big Data that distinguish it from traditional data sources:
  - Volume: The sheer amount of data generated makes traditional methods of storage and analysis impractical.
  - Variety: Big Data encompasses a wide range of data types, structured, semi-structured, and unstructured.
  - Velocity: The speed at which data is generated and needs to be processed requires real-time analysis capabilities.
  - Veracity: Data quality and accuracy become critical considerations when dealing with massive and diverse datasets.

6. Data-Driven Decisions: Empowering Insights with Machine Learning:

- **Leveraging Machine Learning:** Introduce the role of Machine Learning techniques in Big Data analysis. Explain how machines learn from data to identify patterns and make predictions, enhancing the ability to gain valuable insights.

7. A World of Opportunities: Big Data Careers and Growth:

- **The Future of Big Data:** Discuss the booming job opportunities in Big Data Analytics across various industries and the projected growth of the Big Data market. Explore emerging markets where Big Data is revolutionizing various sectors.

Additional Tips:

- Utilize interactive activities, case studies, and real-world data examples to solidify student understanding.
- Encourage students to research specific Big Data applications in their areas of interest.
- Discuss potential solutions and ethical frameworks for addressing Big Data challenges like privacy concerns.

By incorporating these elements, you can equip students with the knowledge and skills to navigate the Big Data landscape and leverage its power to make informed decisions and solve real-world problems.

Teachers can ask the following questions to spark curiosity before starting the topics:

- **Imagine you are running a global online store. Every day, thousands of customers visit, browse, and purchase items. How would you handle and analyse the massive amount of data generated to improve customer experience and boost sales**? (This introduces the concept of Big Data and its applications, making it relatable to real-world scenarios.)
- **Have you ever wondered how streaming platforms like Netflix suggest shows and movies you might like? What kind of data do you think they analyze to make those recommendations?** (This connects Big Data Analytics to familiar examples, sparking curiosity about how data is used for personalization.)
- **In a world where millions of tweets and posts are created every second, how do you think companies or governments use this information to identify trends, predict outcomes, or make decisions?** (This highlights the importance of analysing unstructured data streams and sets the stage for discussing Big Data characteristics like velocity and variety.)

5.1. What is Big Data?

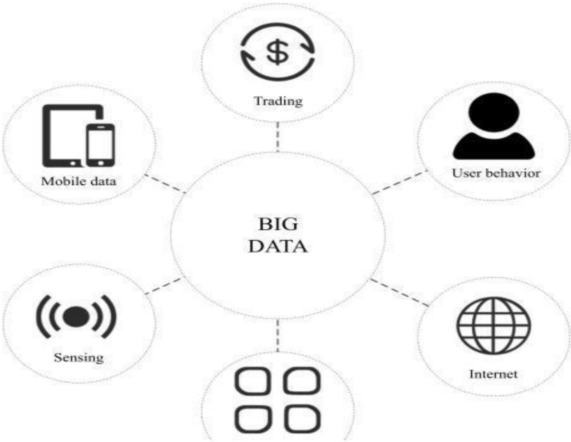


Fig. 5.1 Sources of Big Data

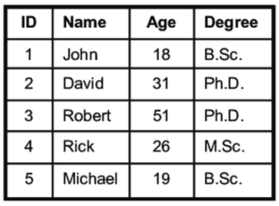
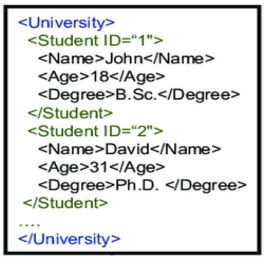
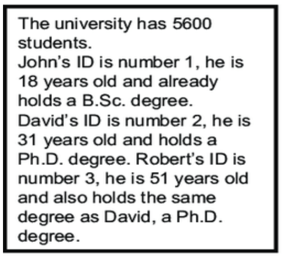
To understand Big Data, let us first understand small data.

Small data refers to datasets that are easily comprehendible by people as they are easily accessible, informative, and actionable, this makes it ideal for individuals and businesses to find useful information and make better choices in everyday tasks. For example, a small store might track daily sales to decide what products to restock.

Big Data refers to extremely large and complex datasets that regular computer programs and databases cannot handle. It comes from three main sources: transactional data (e.g., online purchases), machine data (e.g., sensor readings), and social data (e.g., social media posts). To analyze and use Big Data effectively, special tools and techniques are required. These tools help organizations find valuable insights hidden in the data, which lead to innovations and better decision-making. For example, companies like Amazon and Netflix use Big Data to recommend products or shows based on users' past activities.

5.2. Types of Big Data

Fig. 5.2

Aspect	Structured Data	Semi-Structured Data	Unstructured Data
Definition	Quantitative data with a defined structure	A mix of quantitative and qualitative properties	No inherent structures or formal rules
Data Model	Dedicated data model	May lack a specific data model	Lacks a consistent data model
Organization	Organized in clearly defined columns	Less organized than structured data	No organization exhibits variability over time
Accessibility	Easily accessible and searchable	Accessible but may be harder to analyze	Accessibility depends on the specific data format
Examples	Customer information, transaction records, product directories Structured data  ID | Name | Age | Degree 1 | John | 18 | B.Sc. 2 | David | 31 | Ph.D. 3 | Robert | 51 | Ph.D. 4 | Rick | 26 | M.Sc. 5 | Michael | 19 | B.Sc.	XML files, CSV files, JSON files, HTML files, semi-structured documents Semi-structured data  <University> <Student ID="1"> <Name>John</Name> <Age>18</Age> <Degree>B.Sc.</Degree> </Student> <Student ID="2"> <Name>David</Name> <Age>31</Age> <Degree>Ph.D. </Degree> </Student> …. </University>	Audio files, images, video files, emails, PDFs, social media posts Unstructured data  The university has 5600 students. John's ID is number 1, he is 18 years old and already holds a B.Sc. degree. David's ID is number 2, he is 31 years old and holds a Ph.D. degree. Robert's ID is number 3, he is 51 years old and also holds the same degree as David, a Ph.D. degree.

5.3. Advantages and Disadvantages of Big Data:

Big Data is a key to modern innovation. It has changed how organizations analyze and use information. While it offers great benefits, it also comes with challenges that affect its use in different industries. In this section, we will be discussing a few pros and cons of big data.

Advantages:

- **Enhanced Decision Making**: Big Data analytics empowers organizations to make data-driven decisions based on insights derived from large and diverse datasets.
- **Improved Efficiency and Productivity**: By analyzing vast amounts of data, businesses can identify inefficiencies, streamline processes, and optimize resource allocation, leading to increased efficiency and productivity.
- **Better Customer Insights**: Big Data enables organizations to gain a deeper understanding of customer behavior, preferences, and needs, allowing for personalized marketing strategies and improved customer experiences.
- **Competitive Advantage**: Leveraging Big Data analytics provides organizations with a competitive edge by enabling them to uncover market trends, identify opportunities, and stay ahead of competitors.
- **Innovation and Growth**: Big Data fosters innovation by facilitating the development of new products, services, and business models based on insights derived from data analysis, driving business growth and expansion.

Disadvantages:

- **Privacy and Security Concerns**: The collection, storage, and analysis of large volumes of data raise significant privacy and security risks, including unauthorized access, data breaches, and misuse of personal information.
- **Data Quality Issues**: Ensuring the accuracy, reliability, and completeness of data can be challenging, as Big Data often consists of unstructured and heterogeneous data sources, leading to potential errors and biases in analysis.
- **Technical Complexity**: Implementing and managing Big Data infrastructure and analytics tools require specialized skills and expertise, leading to technical challenges and resource constraints for organizations.
- **Regulatory Compliance**: Organizations face challenges in meeting data protection laws like GDPR (General Data Protection Regulation) and The Digital Personal Data Protection Act, 2023. These laws require strict handling of personal data, making compliance essential to avoid legal risks and penalties.
- **Cost and Resource Intensiveness**: The cost of acquiring, storing, processing, and analyzing Big Data, along with hiring skilled staff, can be high. This is especially challenging for smaller organizations with limited budgets and resources.

Activity: Find the sources of big data using the link UNSTATS

5.4. Characteristics of Big Data

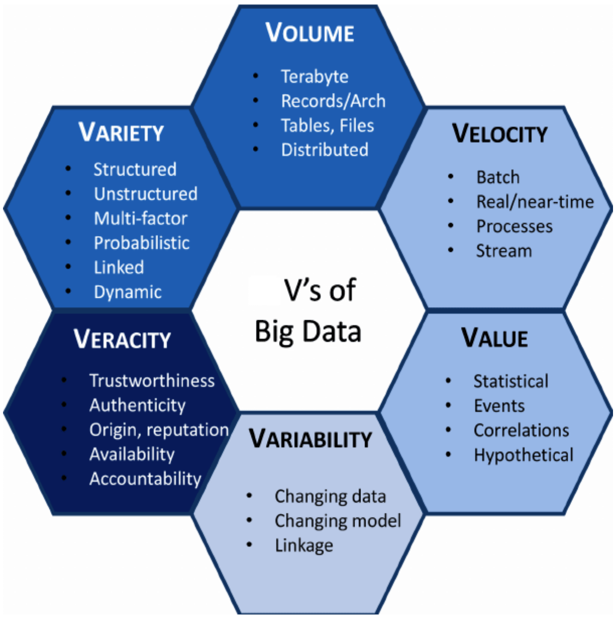


Fig. 5.3 Characteristics of Big Data

The "**characteristics of Big Data**" refer to the defining attributes that distinguish large and complex datasets from traditional data sources. These characteristics are commonly described using the **"3Vs" framework: Volume, Velocity, and Variety.** The **6Vs framework** provides a holistic view of Big Data, emphasizing not only its volume, velocity, and variety but also its veracity, variability, and value. Understanding and addressing these six dimensions are essential for effectively managing, analyzing, and deriving value from Big Data in various domains.

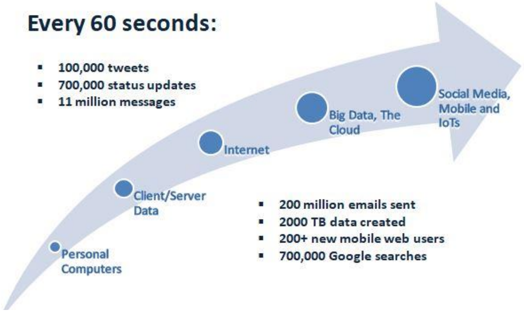


Fig. 5.4 Speed of data generation from various sources

5.4.1. Velocity: Velocity refers to the speed at which data is generated, delivered, and analyzed. In the present world, where millions of people are accessing and storing information online, the speed at which the data gets stored or generated is huge. For example: Google alone generates more than 40,000 search queries per second. See the statistics in the picture provided. Isn't it huge!

5.4.2. Volume: Every day a huge volume of data is generated as the number of people using online platforms has increased exponentially. Such a huge volume of data is considered Big Data. Typically, if the data volume exceeds gigabytes, it falls into the realm of big data. This volume can range from petabytes to terabytes or even exabytes, based on surveys conducted by various organizations. According to the latest estimates, 328.77 million terabytes of data are created each day.

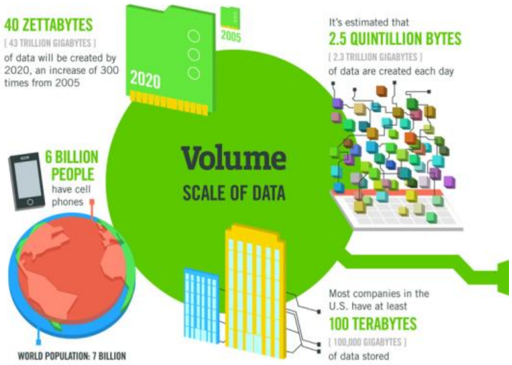

40 ZETTABYTES
[43 TRILLION GIGABYTES]
of data will be created by 2020, an increase of 300 times from 2005
2020
2005
It's estimated that
2.5 QUINTILLION BYTES
[2.3 TRILLION GIGABYTES]
of data are created each day
6 BILLION PEOPLE
have cell phones
Volume
SCALE OF DATA
WORLD POPULATION: 7 BILLION
Most companies in the U.S. have at least
100 TERABYTES
[100,000 GIGABYTES]
of data stored

Fig.5.5 Volume of data

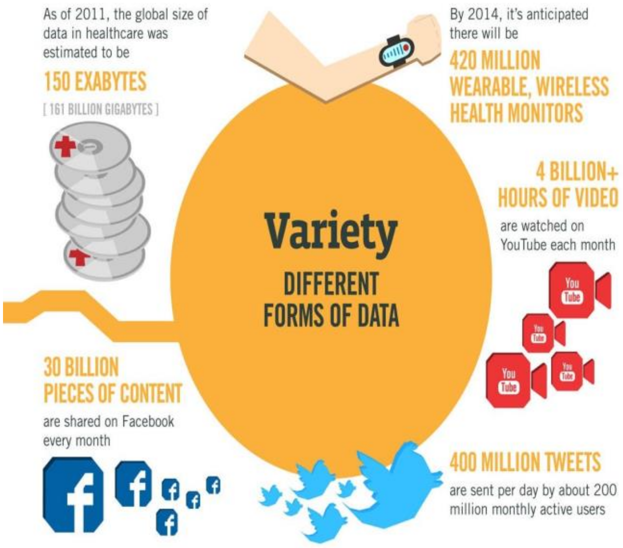

As of 2011, the global size of data in healthcare was estimated to be
150 EXABYTES
[161 BILLION GIGABYTES]
By 2014, it's anticipated there will be
420 MILLION WEARABLE, WIRELESS HEALTH MONITORS
Variety
DIFFERENT FORMS OF DATA
4 BILLION+ HOURS OF VIDEO
are watched on YouTube each month
30 BILLION PIECES OF CONTENT
are shared on Facebook every month
400 MILLION TWEETS
are sent per day by about 200 million monthly active users

Fig.5.6 Varieties in Big data

5.4.3. Variety: Big data encompasses data in various formats, including structured, unstructured, semi-structured, or highly complex structured data. These can range from simple numerical data to complex and diverse forms such as text, images, audio, videos, and so on. Storing and processing unstructured data through RDBMS is challenging. However, unstructured data often provides valuable insights that structured data cannot offer. Additionally, the variety of data sources within big data provides information on the diversity of data.

5.4.4. Veracity: Veracity is a characteristic in Big Data related to consistency, accuracy, quality, and trustworthiness. Not all data that undergoes processing holds value. Therefore, it is essential to clean data effectively before storing or processing it, especially when dealing with massive volumes. Veracity addresses this aspect of big data, focusing on the accuracy and reliability of the data source and its suitability for analytical models.

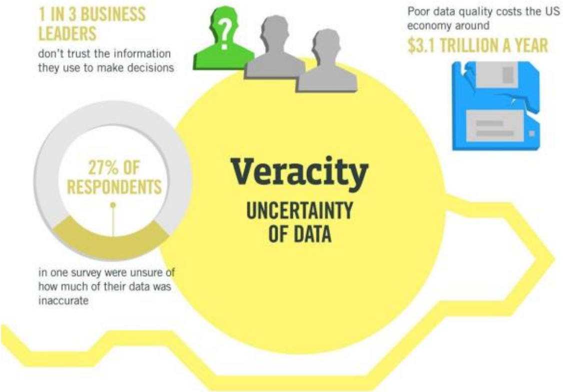

1 IN 3 BUSINESS LEADERS
don't trust the information they use to make decisions
Poor data quality costs the US economy around
$3.1 TRILLION A YEAR
27% OF RESPONDENTS
in one survey were unsure of how much of their data was inaccurate
Veracity
UNCERTAINTY OF DATA

Fig. 5.7

Fig. 5.8 The value of Big Data

5.4.5. Value: The goal of big data analysis lies in extracting business value from the data. Hence, the business value derived from big data is perhaps its most critical characteristic. Without obtaining valuable insights, the other characteristics of big data hold little significance. So, in simple terms Value of Big Data refers to the benefits the big data can provide.

5.4.6. Variability: This refers to establishing if the contextualizing structure of the data stream is regular and dependable even in conditions of extreme unpredictability. It defines the need to get meaningful data considering all possible circumstances.

Fig. 5.9

Case Study: How a Company Uses 3V and 6V Frameworks for Big Data
Company: An OTT Platform 'OnDemandDrama'

3V Framework:

Volume: OnDemandDrama processes huge amounts of data from millions of users, including watch history, ratings, searches, and preferences to offer personalized content recommendations.

Velocity: Data is processed in real-time, allowing OnDemandDrama to immediately adjust recommendations, track the patterns of the users, and offer trending content based on their activity.

Variety: The platform handles diverse data such as user profiles, watch lists, video content, and user reviews which are categorized as structured, semi-structured, and unstructured data.

6V Framework:

Along with the above 3 V of big data, the 6V Framework involves 3 more features of big data named Veracity, Value, and Variability.

Veracity: OnDemandDrama filters out irrelevant or low-quality data (such as incomplete profiles) to ensure accurate content recommendations.

Value: OnDemandDrama uses the data to personalize user experiences, driving engagement and retention by recommending shows and movies that match individual tastes.

Variability: OnDemandDrama handles changes or inconsistencies in data streams caused by factors like user behavior, trends, or any other external events. For example, user preferences can vary based on region, time, or trends.

By using the 3V and 6V frameworks, OnDemandDrama can manage, process, and derive valuable insights from its Big Data, which enhances customer satisfaction and drives business decisions.

5.5. Big Data Analytics

Data Analytics

Data analytics involves analyzing datasets to uncover insights, trends, and patterns. It can be applied to datasets of any size, from small to moderate volumes. Technologies commonly used in data analytics include statistical analysis software, data visualization tools, and relational database management systems (RDBMS).

Big data analytics uses advanced analytic techniques against huge, diverse datasets that include structured, semi-structured, and unstructured data, from different sources, and in various sizes from terabytes to zettabytes.

Big-Data Analytics encompasses the methodologies, tools, and practices involved in analyzing and managing data, covering tasks such as data collection, organization, and storage. The primary objective of data analytics is to utilize statistical analysis and technological methods to uncover patterns and address challenges. In the business realm, big data analytics has gained significance as a means to assess and refine business processes, as well as enhance decision-making and overall business performance. It provides valuable insights and forecasts that help businesses make informed decisions to improve their operations and outcomes. Different types of Big Data Analytics can help businesses and organizations find insights from large and complex datasets. Some of the common types are: Descriptive analytics, Diagnostic analytics, Predictive analytics, and Prescriptive analytics, which we have discussed in Unit 2 of Data Science Methodology.

Big Data Analytics emerges as a consequence of four significant global trends:

1. **Moore's Law**: The exponential growth of computing power as per Moore's Law has enabled the handling and analysis of massive datasets, driving the evolution of Big Data Analytics.
2. **Mobile Computing**: With the widespread adoption of smartphones and mobile devices, access to vast amounts of data is now at our fingertips, enabling real-time connectivity and data collection from anywhere.
3. **Social Networking**: Platforms such as Facebook, Foursquare, and Pinterest facilitate extensive networks of user-generated content, interactions, and data sharing, leading to the generation of massive datasets ripe for analysis.
4. **Cloud Computing**: This paradigm shift in technology infrastructure allows organizations to access hardware and software resources remotely via the Internet on a pay-as-you-go basis, eliminating the need for extensive on-premises hardware and software investments.

5.6. Working on Big Data Analytics

Big data analytics involves collecting, processing, cleaning, and analyzing enormous datasets to improve organizational operations. The working process of big data analytics includes the following steps –

Step 1. Gather data

Each company has a unique approach to data collection. Organizations can now collect structured and unstructured data from various sources, including cloud storage, mobile apps, and IoT sensors.

Step 2. Process Data

Once data is collected and stored, it must be processed properly to get accurate results on analytical queries, especially when it's large and unstructured. Various processing options are available:

- **Batch processing** which looks at large data blocks over time.
- **Stream processing** looks at small batches of data at once, shortening the delay time between collection and analysis for quicker decision-making.

Step 3. Clean Data

Scrubbing all data, regardless of size, improves quality and yields better results. Correct formatting and elimination of duplicate or irrelevant data are essential. Erroneous and missing data can lead to inaccurate insights.

Step 4. Analyze Data

Getting big data into a usable state takes time. Once it's ready, advanced analytics processes can turn big data into big insights.

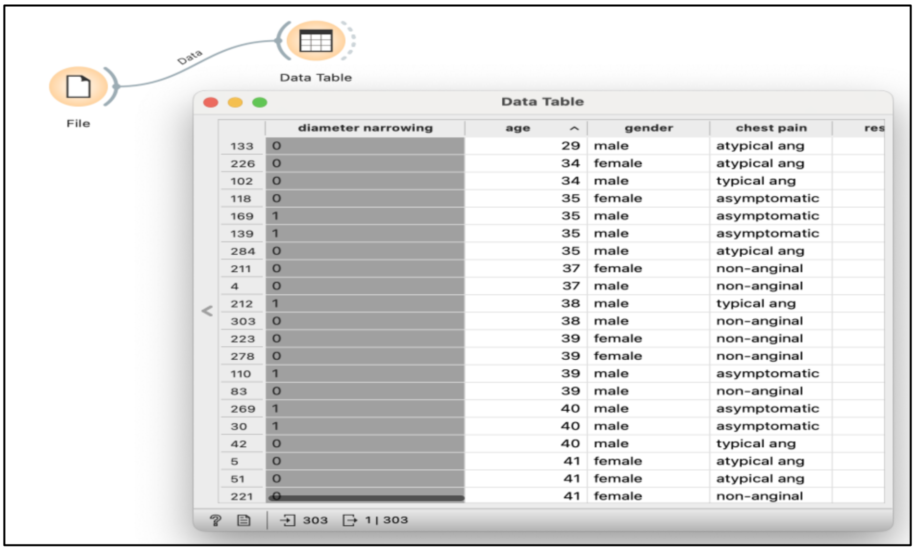
Example: **Data Analytics Tools – Tableau, APACHE Hadoop, Cassandra, MongoDB, SaS**

Using Orange Data Mining for Big Data Analytics

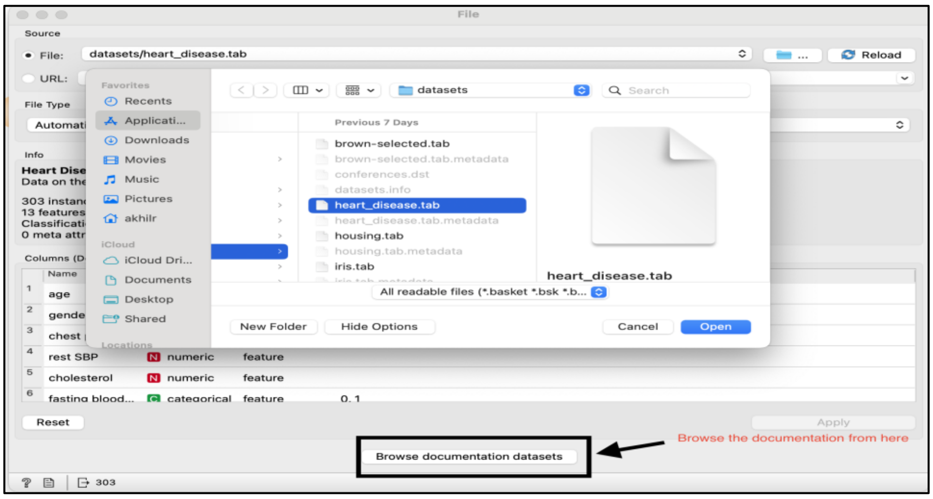
We will explore how big data analysis can be performed using Orange Data Mining.

Step 1: Gather Data

1. Use the **File** widget to load data into Orange.
2. Load the desired dataset. For demonstration, we will use the built-in **Heart Disease** dataset.



	diameter narrowing	age	gender	chest pain	res
133	0	29	male	atypical ang	
226	0	34	female	atypical ang	
102	0	34	male	typical ang	
118	0	35	female	asymptomatic	
169	1	35	male	asymptomatic	
139	1	35	male	asymptomatic	
284	0	35	male	atypical ang	
211	0	37	female	non-anginal	
4	0	37	male	non-anginal	
212	1	38	male	typical ang	
303	0	38	male	non-anginal	
223	0	39	female	non-anginal	
278	0	39	female	non-anginal	
110	1	39	male	asymptomatic	
83	0	39	male	non-anginal	
269	1	40	male	asymptomatic	
30	1	40	male	asymptomatic	
42	0	40	male	typical ang	
5	0	41	female	atypical ang	
51	0	41	female	atypical ang	
221	0	41	female	non-anginal	



File: datasets/heart_disease.tab

	Name			
1	age			
2	gende			
3	chest			
4	rest SBP	N numeric	feature	
5	cholesterol	N numeric	feature	
6	fasting blood...	C categorical	feature	0, 1

Browse the documentation from here

It is important to carefully study the dataset and understand the **features** and **target** variable.

- **Features**: age, gender, chest pain, resting blood pressure (rest_spb), cholesterol, resting ECG (rest_ecg), maximum heart rate (max_hr), etc.
- **Target**: **diameter narrowing**.

> If the value for **diameter narrowing** is **1**, it signifies significant narrowing of the arteries, which is a risk factor for heart disease. If the value is **0**, it indicates healthier arteries with minimal or no narrowing.

Step 2: Process Data

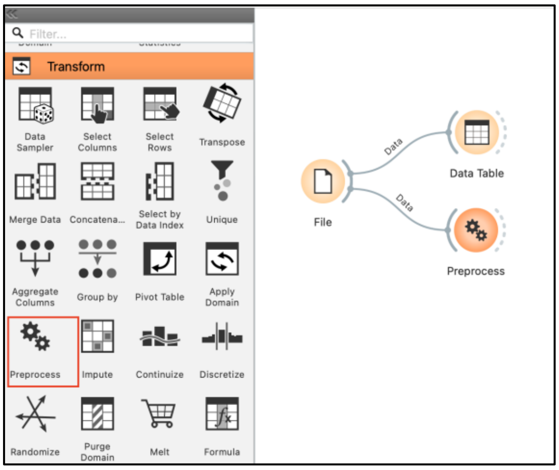
Data processing involves preparing the data for accurate analysis. There are two methods:

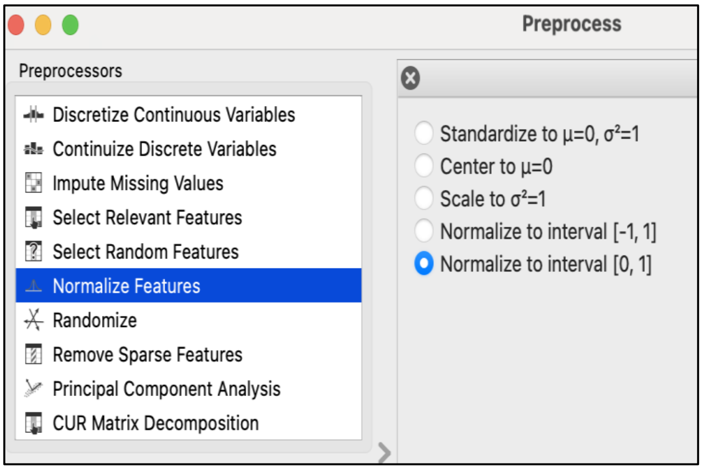
1. **Batch Processing**: Use the **Preprocess** widget to normalize large chunks of structured data at once.
2. **Stream Processing (near-real-time)**: While Orange does not natively support live stream data, you can incrementally process smaller subsets of the data in parallel workflows.

Here, we will focus on the **Normalization** technique.
Normalization in data preprocessing refers to scaling numerical values to a specific range (e.g., 0–1 or -1–1), making them comparable and improving the performance of machine learning algorithms.

Step 2.1: Normalize Data

1. Connect the **Preprocess** widget to the **File** or **Data Table** widget.
2. Double-click on the **Preprocess** widget and select **"Normalize Features"**.
3. Choose an interval, such as **0–1** or **-1–1**.

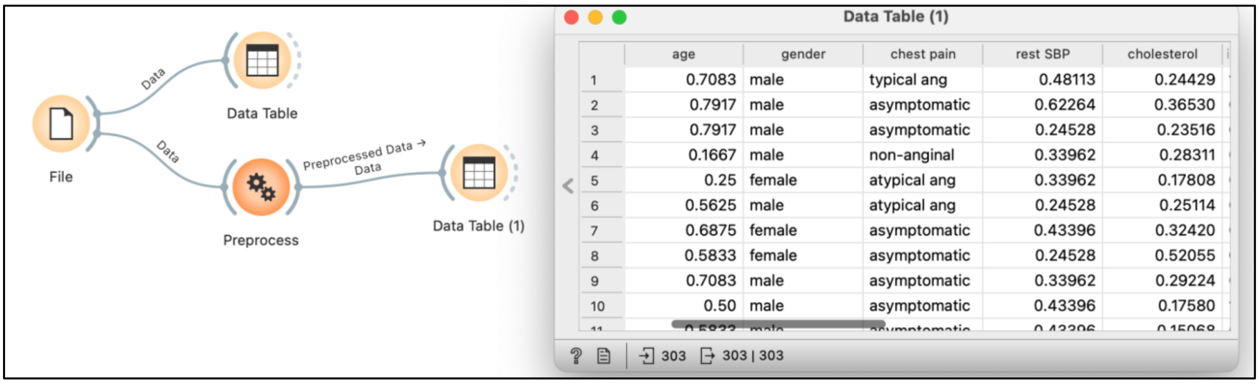




○ Standardize to μ=0, σ²=1
○ Center to μ=0
○ Scale to σ²=1
○ Normalize to interval [-1, 1]
● Normalize to interval [0, 1]

Step 2.2: Verify Normalized Data

1. Connect the **Data Table** widget to the **Preprocess** widget.
2. Open the Data Table to observe the differences in values.



Data Table (1)

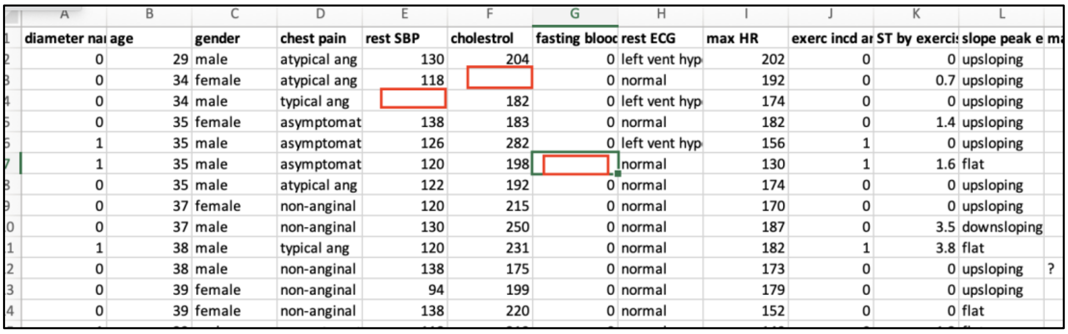
	age	gender	chest pain	rest SBP	cholesterol
1	0.7083	male	typical ang	0.48113	0.24429
2	0.7917	male	asymptomatic	0.62264	0.36530
3	0.7917	male	asymptomatic	0.24528	0.23516
4	0.1667	male	non-anginal	0.33962	0.28311
5	0.25	female	atypical ang	0.33962	0.17808
6	0.5625	male	atypical ang	0.24528	0.25114
7	0.6875	female	asymptomatic	0.43396	0.32420
8	0.5833	female	asymptomatic	0.24528	0.52055
9	0.7083	male	asymptomatic	0.33962	0.29224
10	0.50	male	asymptomatic	0.43396	0.17580
11	0.5833	male	asymptomatic	0.43396	0.15068

303 303 | 303

You will see that all numerical values are now scaled between **0 and 1**.

Step 3: Clean Data

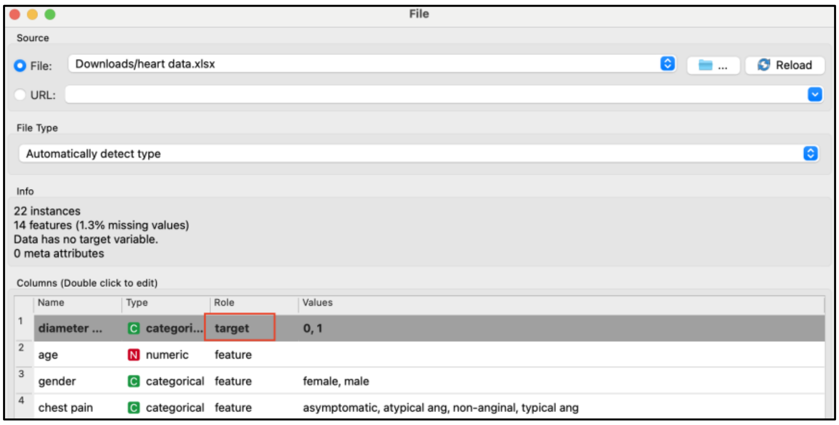
Data cleaning is essential to ensure quality results. We will use the **Impute** widget to handle missing values by replacing them with the **mean**, **median**, **mode**, or a custom value. In this data we all can see that some values are missing in the figure below. This missing value data set is being saved as heart data.xlsx in the computer folder.



	A	B	C	D	E	F	G	H	I	J	K	L	
1	diameter nai	age	gender	chest pain	rest SBP	cholestrol	fasting blooc	rest ECG	max HR	exerc incd ar	ST by exerci:	slope peak e	ma
2	0	29	male	atypical ang	130	204	0	left vent hyp	202	0	0	upsloping	
3	0	34	female	atypical ang	118		0	normal	192	0	0.7	upsloping	
4	0	34	male	typical ang		182	0	left vent hyp	174	0	0	upsloping	
5	0	35	female	asymptomat	138	183	0	normal	182	0	1.4	upsloping	
6	1	35	male	asymptomat	126	282	0	left vent hyp	156	1	0	upsloping	
7	1	35	male	asymptomat	120	198		normal	130	1	1.6	flat	
8	0	35	male	atypical ang	122	192	0	normal	174	0	0	upsloping	
9	0	37	female	non-anginal	120	215	0	normal	170	0	0	upsloping	
10	0	37	male	non-anginal	130	250	0	normal	187	0	3.5	downsloping	
11	1	38	male	typical ang	120	231	0	normal	182	1	3.8	flat	
12	0	38	male	non-anginal	138	175	0	normal	173	0	0	upsloping	?
13	0	39	female	non-anginal	94	199	0	normal	179	0	0	upsloping	
14	0	39	female	non-anginal	138	220	0	normal	152	0	0	flat	

Step 3.1: Upload Data

1. Use the **File** widget to upload a dataset with missing values.
2. Assign the role of **"Target"** to the feature you want to predict.



File

Source

● File: Downloads/heart data.xlsx

○ URL:

File Type

Automatically detect type

Info

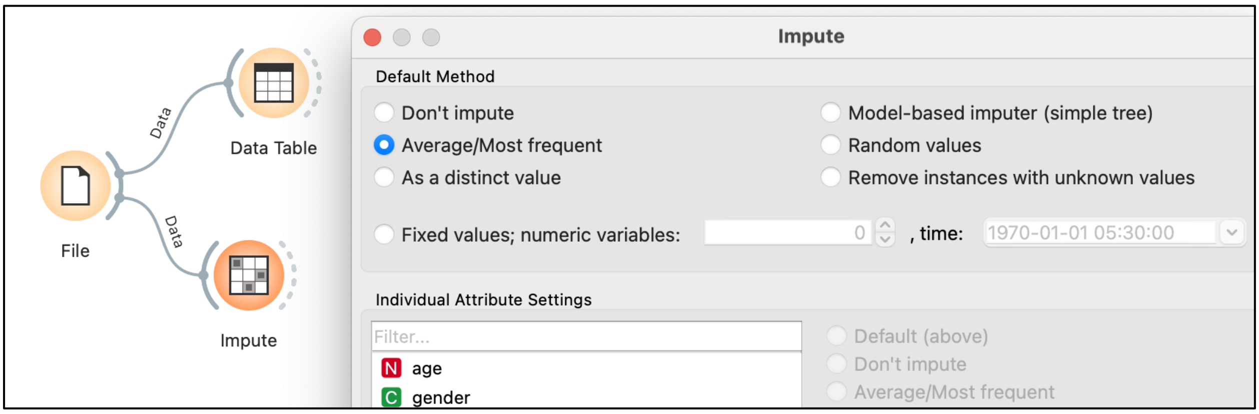
22 instances
14 features (1.3% missing values)
Data has no target variable.
0 meta attributes

Columns (Double click to edit)

	Name	Type	Role	Values
1	diameter ...	C categori...	target	0, 1
2	age	N numeric	feature	
3	gender	C categorical	feature	female, male
4	chest pain	C categorical	feature	asymptomatic, atypical ang, non-anginal, typical ang

Step 3.2: Handle Missing Values

1. Connect the **Impute** widget to the **File** widget.
2. Double-click the **Impute** widget and select an imputation strategy: Average (mean), Most frequent (mode), Fixed value, Random value



Impute

Default Method

○ Don't impute
● Average/Most frequent
○ As a distinct value
○ Model-based imputer (simple tree)
○ Random values
○ Remove instances with unknown values
○ Fixed values; numeric variables: 0 , time: 1970-01-01 05:30:00

Individual Attribute Settings

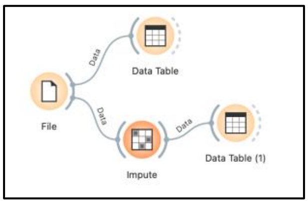
Filter...

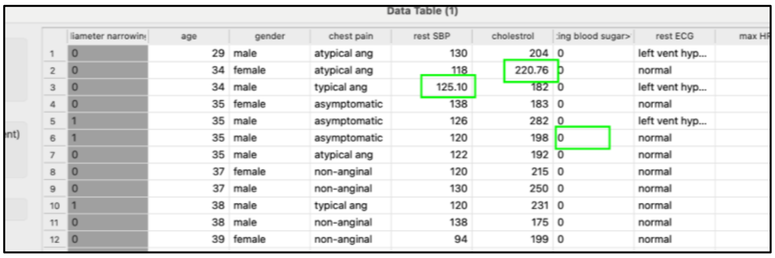
N age
C gender

○ Default (above)
○ Don't impute
○ Average/Most frequent

Step 3.3: Verify Cleaned Data

1. Connect the **Data Table** widget to the **Impute** widget.
2. Open the Data Table to confirm the missing values have been replaced.




Data Table (1)

	liameter narrowin	age	gender	chest pain	rest SBP	cholestrol	:ing blood sugar>	rest ECG	max HF
1	0	29	male	atypical ang	130	204	0	left vent hyp...	
2	0	34	female	atypical ang	118	220.76	0	normal	
3	0	34	male	typical ang	125.10	182	0	left vent hyp...	
4	0	35	female	asymptomatic	138	183	0	normal	
5	1	35	male	asymptomatic	126	282	0	left vent hyp...	
6	1	35	male	asymptomatic	120	198	0	normal	
7	0	35	male	atypical ang	122	192	0	normal	
8	0	37	female	non-anginal	120	215	0	normal	
9	0	37	male	non-anginal	130	250	0	normal	
10	1	38	male	typical ang	120	231	0	normal	
11	0	38	male	non-anginal	138	175	0	normal	
12	0	39	female	non-anginal	94	199	0	normal	

Missing values are now filled with the chosen method (e.g., average values).

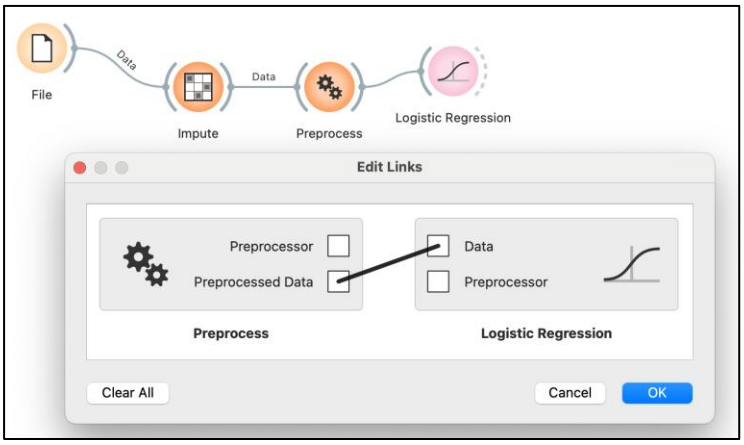
Step 4: Analyze Data

After cleaning, Orange provides various advanced analytics tools to extract insights:

- **K-Means**: For segmenting data into clusters.
- **Logistic Regression / Decision Tree**: For predicting outcomes using labeled data.
- **Scatter Plot / Box Plot / Heat Map**: For visualizing data patterns and relationships.

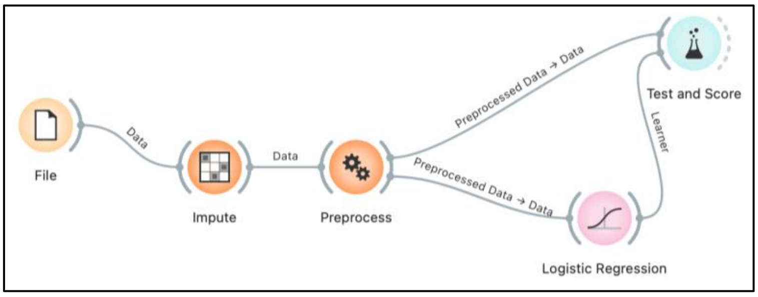
Step 4.1: Build a Logistic Regression Model

1. Drag and drop the **Logistic Regression** widget.
2. Connect it to the cleaned and normalized data.



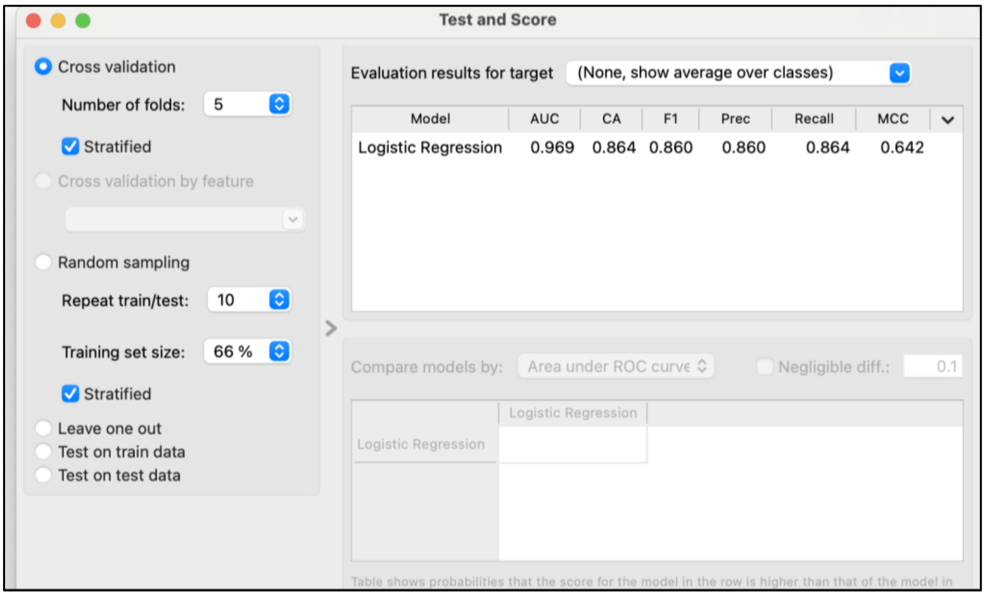
Step 4.2: Test the Model

1. Add the **Test and Score** widget.
2. Connect the **Test and Score** widget to:
   a. The **Logistic Regression** widget (learner data)
   b. The processed data.



Step 4.3: Choose a Validation Method

1. Double-click the **Test and Score** widget. Select a validation method (e.g., **Cross-Validation**).



Test and Score

● Cross validation

Number of folds: 5

☑ Stratified

○ Cross validation by feature

○ Random sampling

Repeat train/test: 10

Training set size: 66 %

☑ Stratified

○ Leave one out

○ Test on train data

○ Test on test data

Evaluation results for target (None, show average over classes)

Model	AUC	CA	F1	Prec	Recall	MCC
Logistic Regression	0.969	0.864	0.860	0.860	0.864	0.642

Compare models by: Area under ROC curve □ Negligible diff.: 0.1

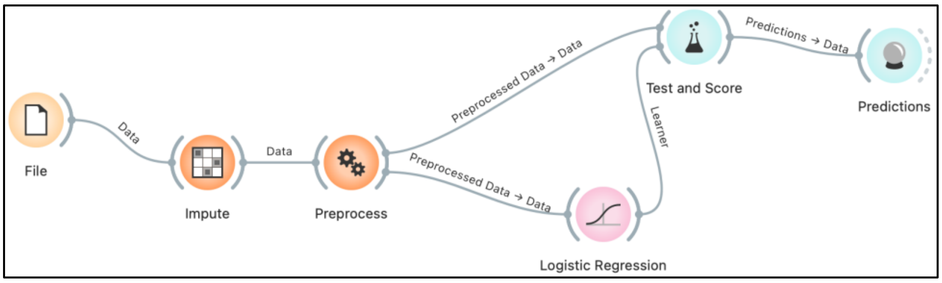
	Logistic Regression
Logistic Regression	

Table shows probabilities that the score for the model in the row is higher than that of the model in

Step 4.4: Generate Predictions

Connect the **Predict** widget to the **Test and Score** widget.

Check the predictions generated using the Logistic Regression model.



5.7. Mining Data Streams

To understand mining data streams, we first understand what data stream is. A data stream is a continuous, real-time flow of data generated by various sources. These sources can include sensors, satellite image data, Internet and web traffic, etc.

Mining data streams refers to the process of extracting meaningful patterns, trends, and knowledge from a continuous flow of real-time data. Unlike traditional data mining, it processes data as it arrives, without storing it completely. An example of an area where data stream mining can be applied is website data. Websites typically receive continuous streams of data daily. For instance, a sudden spike in searches for "election results" on a particular day might indicate that elections were recently held in a region or highlight the level of public interest in the results.

5.8. Future of Big Data Analytics

The future of Big Data Analytics is highly influenced by several key technological advancements that will shape the way data is processed and analyzed. A few of them are:

Real-Time Analytics: It will allow businesses to process data instantaneously, providing immediate insights for decision-making and enabling actions based on live data, such as monitoring customer behavior or tracking supply chain activities.

Development of Advanced Models in Predictive Analytics: Predictive analytics will evolve with the integration of more sophisticated machine learning and AI algorithms, enabling organizations to forecast trends and behaviors with greater precision.

Quantum Computing: Quantum computing promises to revolutionize Big Data analytics by offering unprecedented processing power. Quantum computers will be able to solve complex problems much faster than classical computers.

Activity 1: Note – <u>This is a research-based group activity</u>

i) Watch this video using the link https://www.youtube.com/watch?v=37x5dKW-X5U

ii) Form a group, explore the applications of Big Data & Data Analytics in the following fields, and fill in the table given below:

Field	**Video resource**	**Insights are drawn about this field and its futuristic development**
Education		
Environmental Science		
Media and Entertainment		

Solution: In this activity, the students are to be encouraged to search for any one video source which is related to the above fields given. Their answers may vary, this can be encouraged. In the next column, students can write the insights drawn about the concepts seen in the video. For example: In the field of education, students can write points like- Adaptive Learning, assisting management decisions, adaptive content etc. Encourage the students to write creative answers for each field. Maximum 4 insights are enough for them to write.

Activity-2

List the steps involved in the working process of Big Data analytics.

Step 1:

Step 2:

Step 3:

Step 4:

Solution:

Step 1- Gather data
Step 2- Process data
Step 3-Clean data
Step 4-Analyse data

EXERCISES

A. Multiple Choice questions

1. What does "Volume" refer to in the context of big data?

   a) The variety of data types b) The speed at which data is generated
   c) The amount of data generated d) The veracity of the data

2. Which of the following is a key characteristic of big data?

   a) Structured format b) Easily manageable size
   c) Predictable patterns d) Variety

3. Which of the following is NOT one of the V's of big data?

   a) Velocity b) Volume c) Verification d) Variety

4. What is the primary purpose of data preprocessing in big data analytics?

   a) To increase data volume b) To reduce data variety
   c) To improve data quality d) To speed up data processing

5. Which technique is commonly used for analyzing large datasets to discover patterns and relationships?

   a) Linear regression b) Data mining c) Decision trees d) Naive Bayes

6. Which term describes the process of extracting useful information from large datasets?

   a) Data analytics b) Data warehousing c) Data integration d) Data virtualization

7. Which of the following is a potential benefit of big data analytics?
   a) Decreased data security b) Reduced operational efficiency
   c) Improved decision-making d) Reduced data privacy

8. What role does Hadoop play in big data processing?
   a) Hadoop is a programming language used for big data analytics.
   b) Hadoop is a distributed file system for storing and processing big data.
   c) Hadoop is a data visualization tool.
   d) Hadoop is a NoSQL database management system.

9. What is the primary challenge associated with the veracity aspect of big data?
   a) Handling large volumes of data
   b) Ensuring data quality and reliability
   c) Dealing with diverse data types
   d) Managing data processing speed

B. True or False

1. Big data refers to datasets that are too large to be processed by traditional database systems. **(True)**
2. Structured data is the primary type of data processed in big data analytics, making up the majority of datasets. **(False)**
3. Veracity refers to the trustworthiness and reliability of data in big data analytics. **(True)**
4. Real-time analytics involves processing and analyzing data as it is generated, without any delay. **(True)**
5. Cloud computing is the only concept used in Big Data Analytics. **(False)**
6. A CSV file is an example of structured data. **(False)**
7. "Positive, Negative, and Neutral" are terms related to Sentiment Analysis. **(True)**
8. Data preprocessing is a critical step in big data analytics, involving cleaning, transforming, and aggregating data to prepare it for analysis. **(True)**
9. To analyze vast collections of textual materials to capture key concepts, trends, and hidden relationships, the concept of Text mining is used. **(True)**

C. Short answer questions

1. Define the term Big Data.

Ans - Big Data refers to a vast collection of data that is characterized by its immense volume, which continues to expand rapidly over time.

2. What does the term Volume refer to in Big Data?

Ans - Volume refers to the quantity of data to be stored. In case of big data, huge quantity of data is generated in a very short period of time. For example, Walmart deals with big data. They handle more than 1 million customer transactions every hour, importing more than 2.5 petabytes of data into their database.

3. Mention some important benefits of big data in the health sector.

Ans –

- predictive analysis for predicting disease outbreak, patient analysis and other health risks.
- Personalized medicine
- Clinical decision support
- healthcare resource management

4. Enlist the four types of Big Data Analytics.

Ans - The four types of Big Data Analytics are:

1. Descriptive Analytics: Summarizes historical data to identify patterns and trends.
2. Diagnostic Analytics: Analyses past data to understand the reasons behind specific outcomes.
3. Predictive Analytics: Uses historical data to forecast future events or trends.
4. Prescriptive Analytics: Recommends actions to achieve desired outcomes based on data insights.

These types are designed to provide insights at different levels of decision-making and problem-solving

D. Long answer questions

1. Explain the 6 V's related to Big data.

The **6 V's of Big Data** are:

I. **Volume:** Refers to the massive amount of data generated daily, ranging from terabytes to exabytes. For example, 328.77 million terabytes of data are created every day.

II. **Velocity:** Describes the speed at which data is generated, delivered, and analyzed. For instance, Google processes over 40,000 search queries per second.

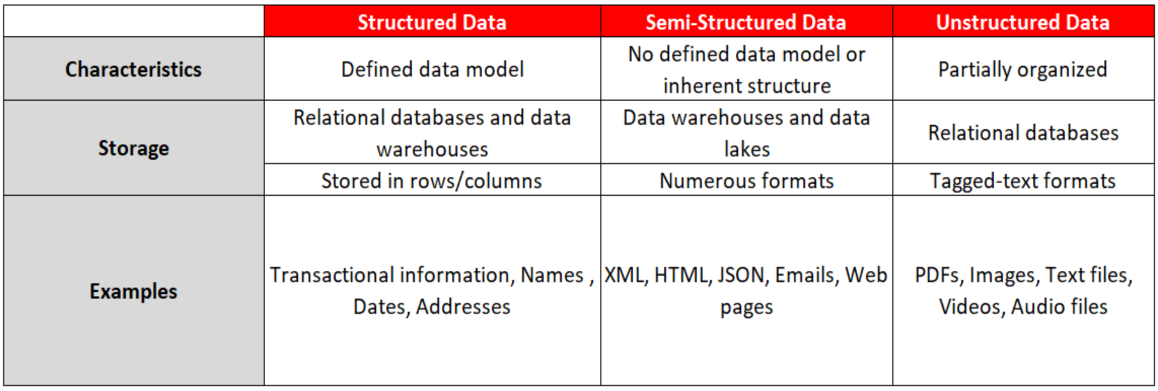
III. **Variety:** Indicates the different forms of data, including structured (e.g., databases), semi-structured (e.g., XML files), and unstructured data (e.g., videos, images, and social media posts).

IV. **Veracity:** Focuses on the accuracy, quality, and trustworthiness of data. It involves ensuring that data is reliable and suitable for analytical models by addressing inconsistencies or inaccuracies.

V. **Value:** Refers to the insights and business benefits that can be extracted from Big Data. Without deriving value, the other characteristics hold little significance.

VI. **Variability:** Highlights the inconsistencies or unpredictability in data flow, requiring systems to adapt and extract meaningful insights even in dynamic conditions

2. Explain the differences between structured, semi-structured, and unstructured data.



	Structured Data	Semi-Structured Data	Unstructured Data
Characteristics	Defined data model	No defined data model or inherent structure	Partially organized
Storage	Relational databases and data warehouses	Data warehouses and data lakes	Relational databases
	Stored in rows/columns	Numerous formats	Tagged-text formats
Examples	Transactional information, Names , Dates, Addresses	XML, HTML, JSON, Emails, Web pages	PDFs, Images, Text files, Videos, Audio files

(Teachers can add a few more points if required.)

3.Explain the process of Big Data Analytics.

Ans- The process of Big Data Analytics can be divided broadly into four major steps. They are as follows:

Step 1. Gather data

Each company has a unique approach to data collection. Organizations can now collect structured and unstructured data from various sources, including cloud storage, mobile apps, and IoT sensors.

Step 2. Process Data

Once data is collected and stored, it must be processed properly to get accurate results on analytical queries, especially when it's large and unstructured. The analysis can be done either batch wise or stream wise.

Step 3. Clean Data
Scrubbing all data, regardless of size, improves quality and yields better results. Correct formatting and elimination of duplicate or irrelevant data are essential. Dirty data can lead to inaccurate insights.
Step 4. Analyze Data

Getting big data into a usable state takes time. Once it’s ready, advanced analytics processes can turn big data into big insights.

4.Why is Big Data Analytics important in modern industries and decision-making processes?

Big Data Analytics is important in modern industries and decision-making processes because it:

1. **Enables Data-**Driven Decisions: By analyzing vast and diverse datasets, organizations can make informed decisions based on insights and trends.
2. **Improves Efficiency and Productivity:** Identifying inefficiencies and optimizing resource allocation helps streamline processes.
3. **Enhances Customer Insights:** Understanding customer behavior and preferences enables personalized marketing and improved customer experiences.
4. **Provides Competitive Advantage:** Leveraging analytics helps organizations uncover market trends, identify opportunities, and stay ahead of competitors.
5. **Fosters Innovation and Growth:** Insights derived from data analysis drive the development of new products, services, and business models.

5.A healthcare company is using Big Data analytics to manage patient records, predict disease outbreaks, and personalize treatments. However, the company is facing challenges regarding data privacy, as patient information is highly sensitive. What are the potential risks to patient privacy when using Big Data in healthcare, and how can these be mitigated?

Potential Risks to Patient Privacy:

1. **Unauthorized Access:** Sensitive patient information could be accessed by unauthorized individuals, leading to breaches of confidentiality.
2. **Data Breaches:** Cyberattacks could expose patient data to malicious actors.
3. **Misuse of Personal Information:** Patient data might be used for purposes beyond its intended scope, such as marketing or profiling.

4. **Regulatory Non-Compliance:** Failing to comply with data protection laws like GDPR or the Digital Personal Data Protection Act, 2023, could lead to legal and financial penalties.

Mitigation Strategies:

1. **Data Encryption:** Encrypt data during storage and transmission to protect against unauthorized access.
2. **Access Controls:** Implement strict access controls to ensure that only authorized personnel can access sensitive data.
3. **Anonymization:** Remove personally identifiable information (PII) from datasets to safeguard patient identity during analysis.
4. **Regular Audits:** Conduct regular security audits to identify and address vulnerabilities.
5. **Compliance with Regulations:** Adhere to data protection laws to ensure ethical handling of sensitive information.
6. **Employee Training:** Educate staff about data privacy practices and the importance of protecting patient information

6.Given the following list of data types, categorize each as Structured, Unstructured, or Semi-Structured:

a) A customer database with fields such as Name, Address, Phone Number, and Email. **Structured**

b) A JSON file containing product information with attributes like name, price, and specifications. **Semi-Structured**

c) Audio recordings of customer service calls. **Unstructured**

d) A sales report in Excel format with rows and columns. **Structured**

e) A collection of social media posts, including text, images, and hashtags. **Unstructured**

f) A CSV file with daily temperature readings for the past year. **Structured**

E. Competency Based Questions:

1. A retail clothing store is experiencing a decline in sales despite strong marketing campaigns. You are tasked with using big data analytics to identify the root cause.
    a. What types of customer data can be analyzed?
    b. How can big data analytics be used to identify buying trends and customer preferences?
    c. Can you recommend specific data visualization techniques to present insights to stakeholders?
    d. How might these insights be used to personalize customer experiences and improve sales?

    Ans:
    a. Analyze purchase history (items bought together, frequency, time of purchase), demographics (age, location, income), and browsing behavior (clicks, time spent on product pages) of the customer.
    b. Big data analytics can help
        i. identify items that are frequently purchased together to optimize product placement and promotions.
        ii. group customers based on demographics and buying habits
        iii. track customer journeys on the website, identify areas of improvement (e.g., checkout process)
    c. understand the key metrics (sales by category, customer demographics) for easy stakeholder comprehension, and to understand the customer browsing behavior on the website (hotspots which indicate the items of interest).
    d. These insights will help the application to
        i. recommend relevant products based on a customer's purchase history and browsing behavior.
        ii. tailor promotions and advertisements to specific customer segments.
        iii. adjust prices based on demand and customer demographics.

2. A research institute is conducting a study on public sentiment towards environmental conservation efforts. They aim to gather insights from various data sources to understand public opinions and perceptions. They collect data from diverse sources such as news articles, online forums, blog posts, and social media comments. Which type of data does this description represent?
    Ans: Unstructured data

3. A global e-commerce platform is experiencing rapid growth in its user base, with millions of transactions occurring daily across various product categories. As part of their data analytics efforts, they are focused on improving the speed and efficiency of processing incoming data to provide real-time recommendations to users during their browsing and purchasing journeys. Identify the specific characteristic of big data (6V's of Big Data) that is most relevant in the above scenario and justify your answer.

Ans:

In the scenario described, the most relevant characteristic of big data from the 6V's perspective is Velocity. The reason being it highlights the need for the e-commerce platform to handle the high speed at which data is generated from millions of transactions daily. The platform needs to process this data quickly to provide real-time recommendations during a user's browsing and purchasing journey. Delays in processing could lead to missed opportunities to influence customer decisions.

Reference links:

- https://www.researchgate.net/publication/259647558_Data_Stream_Mining
- https://www.ibm.com/topics/big-data-analytics
- https://www.researchgate.net/figure/olume-scale-of-Data-from-different-data-sources-26_fig1_324015815Fig

UNIT 6: Understanding Neural Networks

Title: Understanding Neural Networks

Approach: Team Discussion, Web search**,** Case studies

Summary: This unit will introduce students to the fundamentals of neural networks, including their structure, components, and types. Students will learn how neural networks are used in various applications and its impact on society.

Learning Objectives:

1. To understand the basic structure of a neural network and its components.
2. To explore different types of neural networks and their applications.
3. To analyse case studies of neural networks in real-world scenarios.

Key Concepts:

1. Parts of a neural network.
2. Components of a neural network.
3. Working of a neural network.
4. Types of neural networks, such as feedforward, convolutional, and recurrent.
5. Impact of neural network on society.

Learning Outcomes:

Students will be able to –

1. Explain the basic structure and components of a neural network.
2. Identify different types of neural networks and their respective applications.
3. Understand machine learning and neural networks through hands-on projects, interactive visualization tools, and practical Python programming.

Prerequisites:

Basic understanding of machine learning concepts.

Unveiling the Neural Network: A Teacher's Guide to Machine Learning

This lesson empowers you to guide students through the fascinating world of Neural Networks, a cornerstone of Machine Learning.

1. The Brain in a Box: Unveiling Neural Networks:

- **Captivating Introduction:** Spark curiosity with an analogy: "Imagine a computer program inspired by the human brain!" Introduce the concept of Neural Networks (NNs) – machine learning models loosely mimicking how the brain processes information.

2. Demystifying the Structure:

- **NN Breakdown:** Introduce the key components of a Neural Network:
  - Structure: Layers of interconnected nodes (artificial neurons) that process information.
  - Connections: Weighted connections between neurons that determine signal flow.
  - Activation Functions: Functions that determine a neuron's output based on its weighted inputs.
  - Learning Rules: Algorithms that adjust connection weights based on training data, allowing the network to learn.
  - Signal Propagation: The process by which information flows through the network, layer by layer.
- **Real-World Applications:** Showcase the diverse applications of NNs:
  - Image Recognition: Identifying objects and scenes in images (e.g., facial recognition).
  - Natural Language Processing: Enabling machines to understand and generate human language (e.g., chatbots, machine translation).
  - Recommendation Systems: Recommending products or content based on user preferences (e.g., e-commerce platforms).
  - Predictive Analytics: Making predictions based on historical data (e.g., stock market trends).

3. A Network of Networks: Exploring Different Architectures:

- **NN Diversity:** Introduce various types of Neural Networks:
  - Feedforward Networks: Information flows in one direction, from input to output layers.
  - Convolutional Neural Networks (CNNs): Highly effective for image recognition tasks.
  - Recurrent Neural Networks (RNNs): Suitable for sequential data like text or speech.
  - Deep Neural Networks (DNNs): Multiple layers of interconnected neurons, capable of complex learning.

- **Structure, Applications, and Implications:** Discuss the specific structure, applications, and implications of each network type for diverse machine learning domains.

4. Learning by Doing: Hands-on Projects for Interactive Learning:

- **Engaging Activities:** Introduce age-appropriate hands-on projects or simulations to solidify understanding of basic machine learning concepts.
    - Create simple neural networks using online learning platforms or unplugged activities like sorting objects based on features.

5. Building with Code: Introduction to TensorFlow and Keras:

- **Practical Learning:** For advanced students, introduce TensorFlow and Keras – popular libraries used for building and training Neural Networks. Guide them through creating and training a simple Neural Network on a beginner-friendly dataset using these tools.

6. The Power and Responsibility of NNs: Societal Impact and Ethical Considerations:

- **Revolutionizing Society:** Discuss the positive societal impact of NNs:
    - Increased efficiency in various sectors.
    - Personalized experiences for users.
    - Economic growth and innovation.
- **Ethical Scrutiny:** Emphasize the importance of ethical considerations:
    - Data privacy concerns – ensuring responsible data collection and usage.
    - Bias in algorithms – mitigating against biased datasets leading to unfair outcomes.

Additional Tips:

- Utilize visuals like diagrams and animations to illustrate Neural Network structures and processes.
- Encourage students to research specific NN applications in their areas of interest.
- Discuss potential solutions for mitigating ethical concerns related to NNs.
- Explore online resources and interactive tools for further student exploration of Neural Networks.

By incorporating these elements, you can equip students with a solid foundation in Neural Networks and empower them to contribute responsibly to the future of this transformative technology.

Teachers can ask the following questions to spark curiosity before starting the topics:

- **Can you think of any examples in everyday life where we encounter situations that involve making choices based on patterns?** (This will help connect the concept of neural networks to real-world applications. Students might provide examples like filtering spam emails, recommending products on shopping websites, or recognizing faces in photos.)
- **Have you ever learned a new skill by practicing and improving over time? How do you think this process of learning happens in the brain?** (This will help bridge the gap between biological neurons and artificial neurons. By understanding how our brains learn and adapt, students can better grasp the concept of neural networks that mimic this process.)

6.1. WHAT IS A NEURAL NETWORK?

A network of neurons is called "Neural Network". "Neural" comes from the word neuron of the human nervous system. The neuron is a cell within the nervous system which is the basic unit of the brain used to process and transmit the information to all other nerve cells and muscles.

A similar behaviour of neurons is embedded by Artificial Intelligence into a Network called Artificial Neural Network. It can adapt to changing inputs and hence the network generates the best possible outcome without making any changes anywhere as the human brain does.

A neural network is a machine learning program, or model, that makes decisions in a manner similar to the human brain, by using processes that mimic the way biological neurons work together to identify phenomena, weigh options and arrive at conclusions.

The main advantage of neural networks is that they can extract data features automatically without needing any input from the programmer. Artificial Neural Networks (ANN) are very famous these days and also considered to be a very interesting topic due to their applications in chat-bots. It also makes our work easy by auto replying the emails, suggesting email replies, spam filtering, Facebook image tagging, showing items of our interest in the e-shopping web portals, and many more. One of the best-known examples of a neural network is Google's search algorithm.

6.1.1 PARTS OF A NEURAL NETWORK: -

Every neural network comprises layers of interconnected nodes — an input layer, hidden layer(s), and an output layer, as shown in Fig 6.1.

1. **Input Layer**: This layer consists of units representing the input fields. Each unit corresponds to a specific feature or attribute of the problem being solved.
2. **Hidden Layers**: These layers, which may include one or more, are located between the input and output layers. Each hidden layer contains nodes or artificial neurons, which process the input data. These nodes are interconnected, and each connection has an associated weight.

3. **Output Layer**: This layer consists of one or more units representing the target field(s). The output units generate the final predictions or outputs of the neural network.

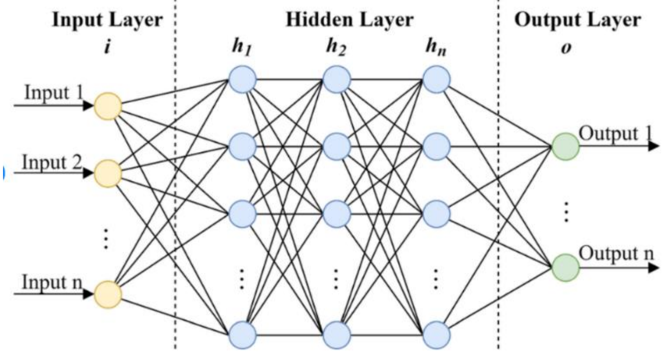


Fig 6.1 Layers of Neural Network

Figure Source : https://www.researchgate.net/figure/General-Neural-Network-Architecture-45_fig3_355944019

Each node is connected to others, and each connection is assigned a weight. If the output of a node exceeds a specified threshold value, the node is activated, and its output is passed to the next layer of the network. Otherwise, no data is transmitted to the subsequent layer.

An Artificial Neural Network (ANN) with two or more hidden layers is known as a **Deep Neural Network**. The process of training deep neural networks is called ***Deep Learning***. The term "deep" in deep learning refers to the number of hidden layers (also called *depth*) of a neural network. A neural network that consists of more than three layers—which would be inclusive of the inputs and the output layers—can be considered a deep learning algorithm. A neural network that only has three layers is just a basic neural network.

6.2. COMPONENTS OF A NEURAL NETWORK

The key components of a neural network are as follows:

1. Neurons:

- Neurons (also known as nodes) are the fundamental building blocks of a neural network.
- They receive inputs from other neurons or external sources.
- Each neuron computes a weighted sum of its inputs, applies an activation function, and produces an output.

2. Weights:

- Weights represent the strength of connections between neurons.
- Each connection (synapse) has an associated weight.
- Weights convey the importance of that feature in predicting the final output.
- During training, neural networks learn optimal weights to minimize error.

3. Activation Functions

- Activation functions in neural networks are like decision-makers for each neuron. They decide whether a neuron should be activated (send a signal) or not based on the input it receives.

- Different types of Activation Functions are Sigmoid Function, Tanh Function, ReLU (Rectified Linear Unit), etc. (A detailed explanation of activation functions is beyond the scope of the syllabus)
- These functions help neural networks learn and make decisions by adding non-linearities to the model, allowing them to understand complex patterns in data.

4. Bias:

- Bias terms are constants added to the weighted sum before applying the activation function.
- They allow the network to shift the activation function horizontally.
- Bias helps account for any inherent bias in the data.

5. Connections:

- Connections represent the synapses between neurons.
- Each connection has an associated weight, which determines its influence on the output of the connected neurons.
- Biases (constants) are also associated with each neuron, affecting its activation threshold.

6. Learning Rule:

- Neural networks learn by adjusting their weights and biases.
- The learning rule specifies how these adjustments occur during training.
- Backpropagation, a common learning algorithm, computes gradients and updates weights to minimize the network's error.

7. Propagation Functions:

- These functions define how signals propagate through the network during both forward and backward passes. Forward pass is known as Forward Propagation and backward pass is known as Back Propagation.
- **Forward Propagation**

  In the forward propagation, input data flows through the layers, and activations are computed. Here, input data flows through the network layers, and activations are computed. The predicted output is compared to the actual target (ground truth), resulting in an error (loss).
- **Back Propagation**

  Backpropagation is the essence of neural network training. It is the practice of fine-tuning the weights of a neural network based on the error rate (i.e. loss) obtained in the previous epoch (i.e. iteration.) Proper tuning of the weights ensures lower error rates, making the model reliable by increasing its generalization and helping the network to improve its prediction over time.

  Backpropagation, (short for "backward propagation of errors") is an optimization algorithm used during neural network training. It adjusts the weights of the network based on the error (loss) obtained in the previous iteration (epoch).

  In Back propagation, gradients are propagated to update weights using optimization algorithms (e.g., gradient descent).

6.3. WORKING OF A NEURAL NETWORK

Teachers can ask the following question:

When you practice a new skill, like riding a bike, what happens with your physical movements and how do you eventually improve? How might similar adjustments happen within a neural network during training?

Imagine each node as a simple calculator. It takes input numbers, multiplies them by certain values (weights), adds them together, adds an extra number (bias), and then gives an output. This output is used as input for the next node in the network. The formula would look something like this:

$$\boxed{\sum wixi + bias = w1x1 + w2x2 + w3x3 + bias}$$

$$output \Rightarrow f(x) = 1, \quad \text{if } \sum w1x1 + b >= 0;$$

$$f(x) = 0, \quad \text{if } \sum w1x1 + b < 0$$

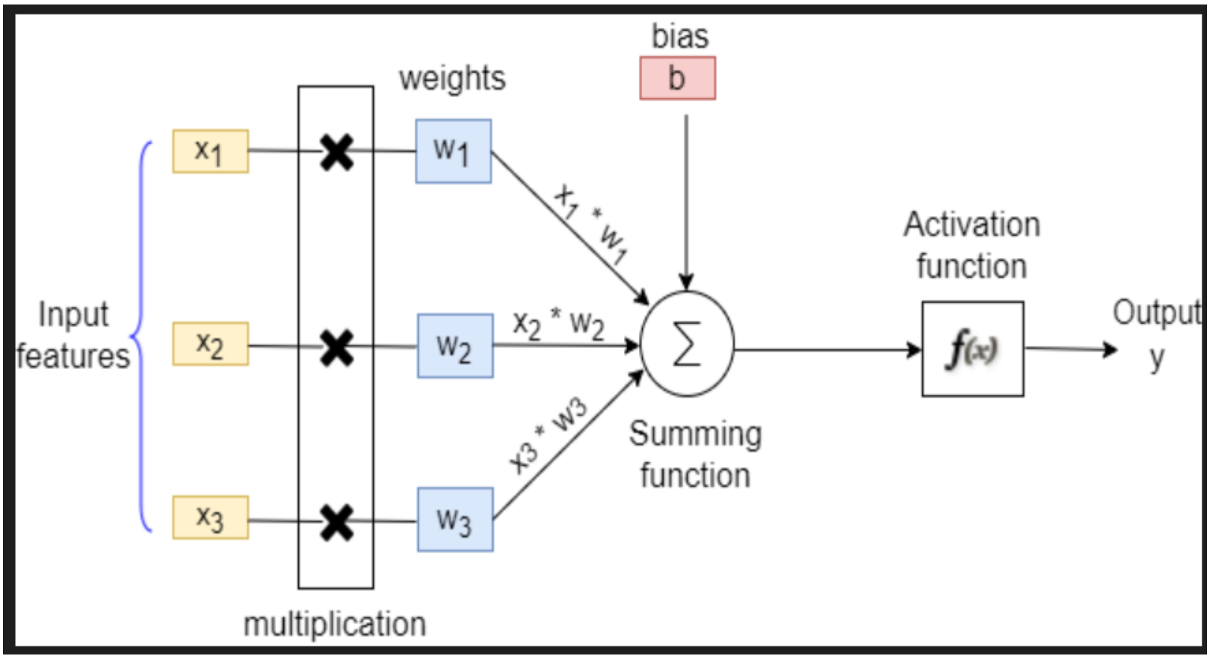


Fig 6.2 Image source : https://www.bombaysoftwares.com/blog/feed-forward-propagation

Fig 6.2 illustrates that the output of a neuron can be expressed as a linear combination of weight ‘w’ and bias ‘b’, expressed mathematically as w * x + b.

Each input is assigned a weight to show its importance. These weights are used to calculate the total input by multiplying each input with its weight and adding them together. This total input then goes through an activation function, which decides if the node should "fire" or activate based on the result. If it fires, the output is passed to the next layer. This process continues, with each node passing its output to the next layer, defining the network as a feedforward network. *One single node might look like using binary values.*

Let us see a simple problem.
CASE I: Let the features be represented as x1,x2 and x3.
Input Layer:
Feature 1, x1 = 2
Feature 2, x2 = 3
Feature 3, x3 = 1

Hidden Layer:
Weight 1, w1 = 0.4
Weight 2, w2 = 0.2
Weight 3, w3 = 0.6
bias = 0.1
threshold = 3.0

Output: Using the formula:
∑wixi + bias = w1x1 + w2x2 + w3x3 + bias
= (0.4*2) + (0.2*3) + (0.6*1) + 0.1

= 0.8 + 0.6 + 0.6 + 0.1

= 2.1

Now, we apply the threshold value:
If output > threshold, then output = 1 (active)
If output < threshold, then output = 0 (inactive)
In this case:
Output (2.1) < threshold (3.0)
So, the output of the hidden layer is:
Output = 0
This means that the neuron in the hidden layer is inactive.

CASE II

Let's say we have another neuron in the output layer with the following weights and bias:

w1 = 0.7
w2 = 0.3
bias = 0.2

The output of the hidden layer (0) is passed as input to the output layer:

Output = w1_x1 + w2_x2 + bias
= 0.7_0 + 0.3_0 + 0.2
= 0.2
Let's assume the threshold value for the output layer is 0.1:
Output (0.2) > threshold (0.1)
So, the final output of the neural network is:
Output = 1

We can now apply this concept to a more tangible example: -

Imagine you are standing on the shore, contemplating whether to go for surfing or not. You glance out at the ocean, assessing the conditions. Are the waves big and inviting? Is the line-up crowded with fellow surfers? And perhaps most importantly, has there been any recent shark activity? Using Neural Network concept decide ***"Whether you should go for surfing or not"***

Note: The decision to go or not to go is our predicted outcome, $\hat{y}$ or y-hat. (The estimated or predicted values in a regression or other predictive model are termed the $\hat{y}$.)

If Output is 1, then Yes (go for surfing), else 0 implies No (do not go for surfing)

Solution:

1. Let us assume that there are three factors influencing your decision-making:
   - Are the waves good? (Yes: 1, No: 0)
   - Is the line-up empty? (Yes: 1, No: 0)
   - Has there been a recent shark attack? (Yes: 0, No: 1)
2. Then, let us assume the following, giving us the following inputs:
   - X1 = 1, since the waves are pumping
   - X2 = 0, since the crowds are out
   - X3 = 1, since there has not been a recent shark attack
3. Now, we need to assign some weights to determine importance. Larger weights signify that particular variables are of greater importance to the decision or outcome.
   - W1 = 5, since large swells do not come often
   - W2 = 2, since you are used to the crowds
   - W3 = 4, since you have a fear of sharks

Factor	Input	Weight
Wave Quality	1	5
Lineup Congestion	0	2
Any Shark Activity	1	4

4. Finally, we will assume a threshold value of 3, which would translate to a bias value of –3.

With all the various inputs, we can start to plug in values into the formula to get the desired output.

$$\hat{y} = (1*5) + (0*2) + (1*4) - 3 = 6$$

If $\hat{y}$ > threshold, then output = 1

If $\hat{y}$ < threshold, then output = 0

Since 6 is greater than 3, we can determine that the output of this node would be 1. **In this instance, you would go surfing.**

If we adjust the weights or the threshold, we can achieve different outcomes from the model.

6.4. TYPES OF NEURAL NETWORKS

Neural networks can be classified into different types, which are used for different purposes. While this is not a comprehensive list of types, the below would be representative of the most common types of neural networks that you will come across for its common use cases:

6.4.1 Standard Neural Network (Perceptron):

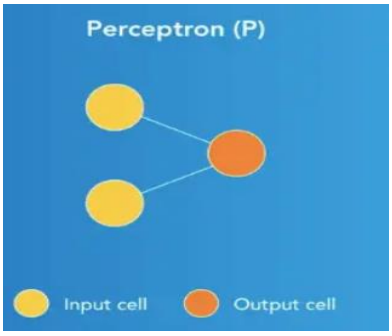


Fig 6.3

The perceptron, created by Frank Rosenblatt in 1958, is a simple neural network with a single layer of input nodes fully connected to a layer of output nodes (Fig 6.3). It uses Threshold Logic Units (TLUs) as artificial neurons.

Application: Useful for binary classification tasks like spam detection and basic decision-making.

6.4.2 Feed Forward Neural Network (FFNN):

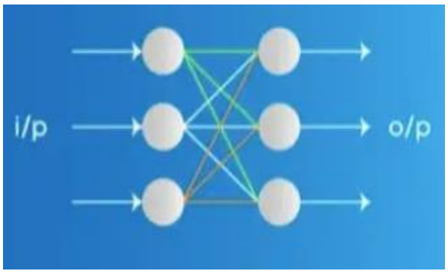


Fig 6.4

FFNN, also known as multi-layer perceptrons (MLPs) have an input layer, one or more hidden layers, and an output layer. Data flows only in one direction, from input to output. They use activation functions and weights to process information in a forward manner (Fig 6.4). FFNNs are efficient for handling noisy data and are relatively straightforward to implement, making them versatile tools in various AI applications.

Application: used in tasks like image recognition, natural language processing (NLP), and regression.

6.4.3 Convolutional Neural Network (CNN):

CNNs utilize filters to extract features from images, enabling robust recognition of patterns and objects (Fig 6.5). CNNs differ from standard neural networks by incorporating a three-dimensional arrangement of neurons, which is particularly effective for processing visual data.

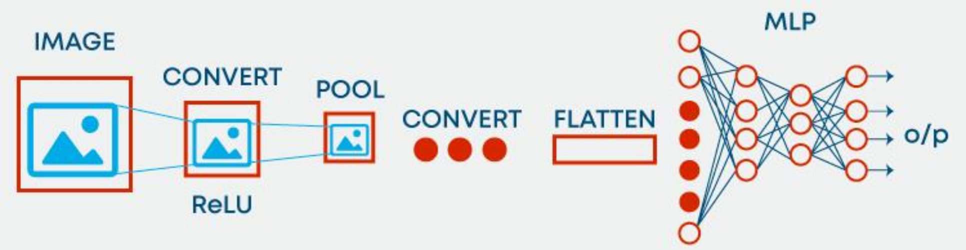


Fig 6.5 Image Source: https://www.mygreatlearning.com/blog/types-of-neural-networks/

Application: Dominant in computer vision for tasks such as object detection, image recognition, style transfer, and medical imaging.

[To see the working of a CNN, you may watch the video on https://www.youtube.com/watch?v=K_BHmztRTpA&t=1s]

6.4.4 Recurrent Neural Network (RNN):

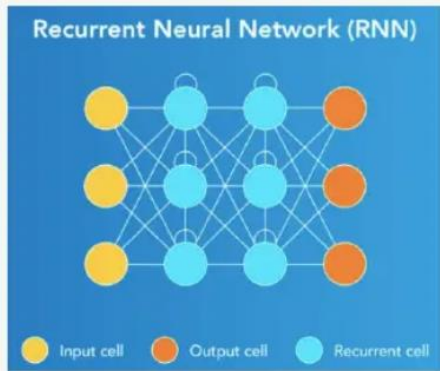


Fig 6.6

RNNs are designed for sequential data and feature feedback loops to allow information persistence across time steps. They have feedback connections that allow data to flow in a loop (Fig 6.6). If the prediction is wrong, the learning rate is employed to make small changes. Hence, making it gradually increase towards making the right prediction during the backpropagation.

Application: Used in NLP for language modeling, machine translation, chatbots, as well as in speech recognition, time series prediction, and sentiment analysis.

6.4.5 Generative Adversarial Network (GAN):

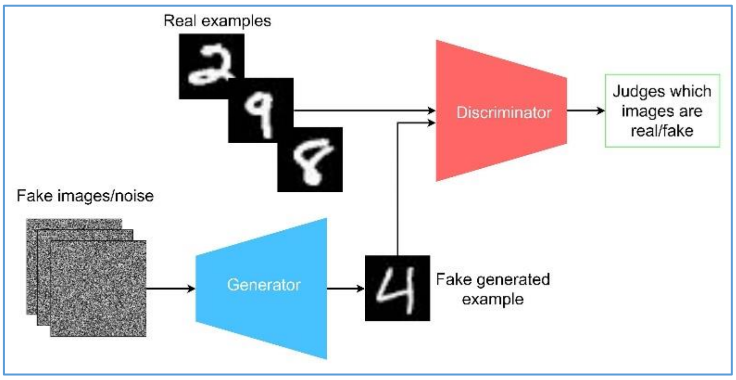


Fig 6.7

GANs consist of two neural networks – a generator and a discriminator – trained simultaneously. The generator creates new data instances, while the discriminator evaluates them for authenticity. They are used for unsupervised learning and can generate new data samples. GANs are used for generating realistic data, such as images and videos.

Application: Widely employed in generating synthetic data for various tasks like image generation, style transfer, and data augmentation.

6.5. FUTURE OF NN AND ITS IMPACT ON SOCIETY:

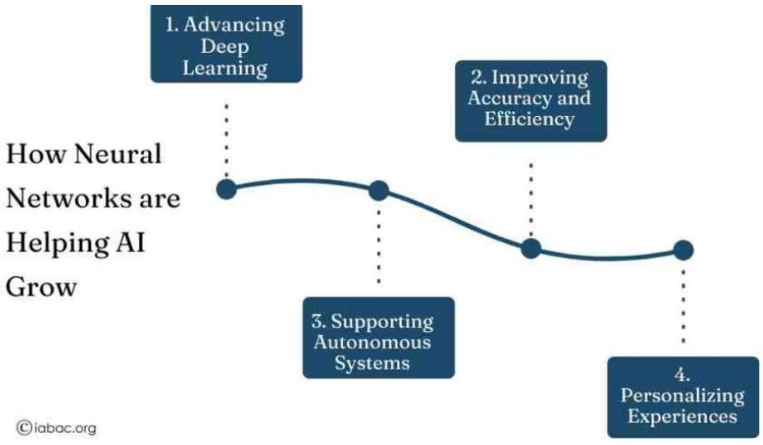

1. Advancing Deep Learning
2. Improving Accuracy and Efficiency
How Neural Networks are Helping AI Grow
3. Supporting Autonomous Systems
4. Personalizing Experiences
©iabac.org

Fig 6.8 https://iabac.org/blog/exploring-the-role-of-neural-networks-in-the-future-of-artificial-intelligence

The widespread integration of neural networks has revolutionized society across various domains. These sophisticated algorithms have significantly enhanced efficiency and productivity by automating tasks, streamlining processes, and optimizing resource allocation in industries ranging from manufacturing to finance. Fig 6.8 illustrates how Neural network are helping A.I. grow tremendously.

Moreover, neural networks have personalized products and services by analysing vast datasets, leading to tailored recommendations and experiences that cater to individual preferences and needs. Economically, the adoption of neural networks has spurred innovation, driven economic growth, and created new job opportunities in burgeoning fields like data science and artificial intelligence.

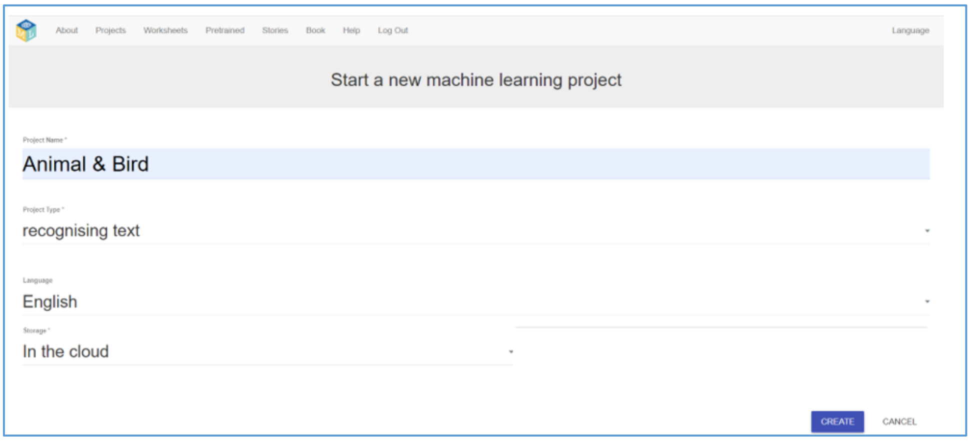
However, alongside these benefits, concerns over ethical and societal implications such as data privacy, algorithmic bias, and job displacement highlight the need for careful consideration and regulation to ensure that neural networks serve the collective well-being of society.

6.6. PRACTICAL ACTIVITIES

6.6.1 Activity 1: (for Practical Concepts Only)

Machine Learning for Kids url: https://machinelearningforkids.co.uk/

Project: Identifying Animals & Birds



Start a new machine learning project

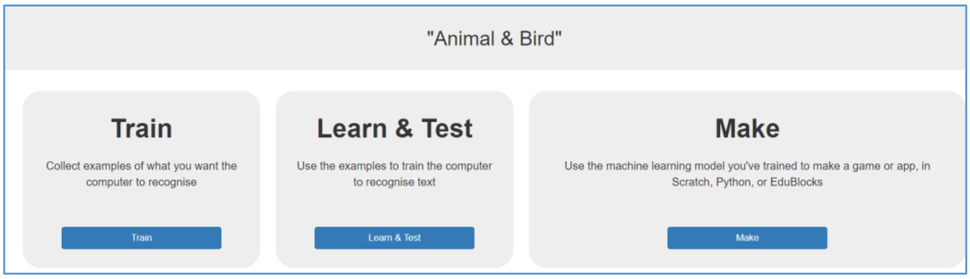
Project Name *
Animal & Bird

Project Type *
recognising text

Language
English

Storage *
In the cloud

Steps in Machine Learning for Kids



"Animal & Bird"

Train

Collect examples of what you want the computer to recognise

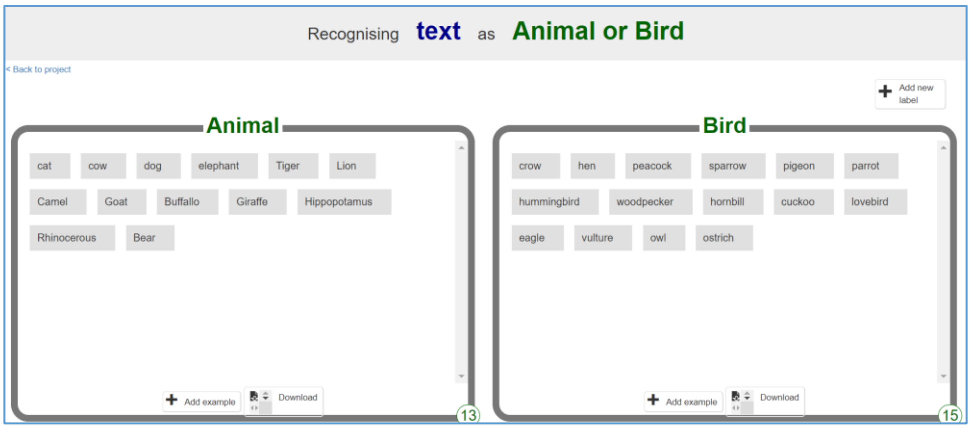
Learn & Test

Use the examples to train the computer to recognise text

Make

Use the machine learning model you've trained to make a game or app, in Scratch, Python, or EduBlocks

After Adding Labels & Contents



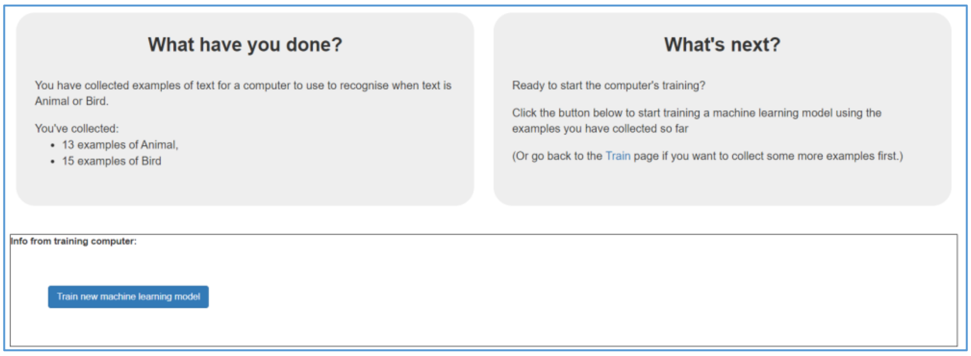
Recognising **text** as **Animal or Bird**

< Back to project

Animal: cat, cow, dog, elephant, Tiger, Lion, Camel, Goat, Buffalo, Giraffe, Hippopotamus, Rhinocerous, Bear

Bird: crow, hen, peacock, sparrow, pigeon, parrot, hummingbird, woodpecker, hornbill, cuckoo, lovebird, eagle, vulture, owl, ostrich

Train the content and test it



What have you done?

You have collected examples of text for a computer to use to recognise when text is Animal or Bird.

You've collected:

- 13 examples of Animal,
- 15 examples of Bird

What's next?

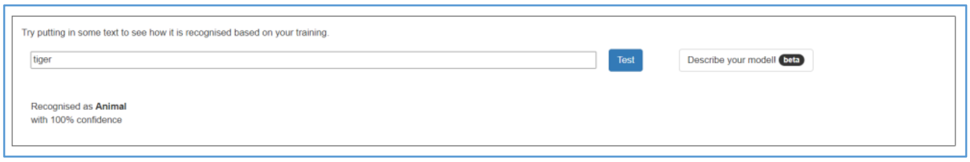
Ready to start the computer's training?

Click the button below to start training a machine learning model using the examples you have collected so far

(Or go back to the Train page if you want to collect some more examples first.)

Info from training computer:

Click on Describe your model to know about neural network

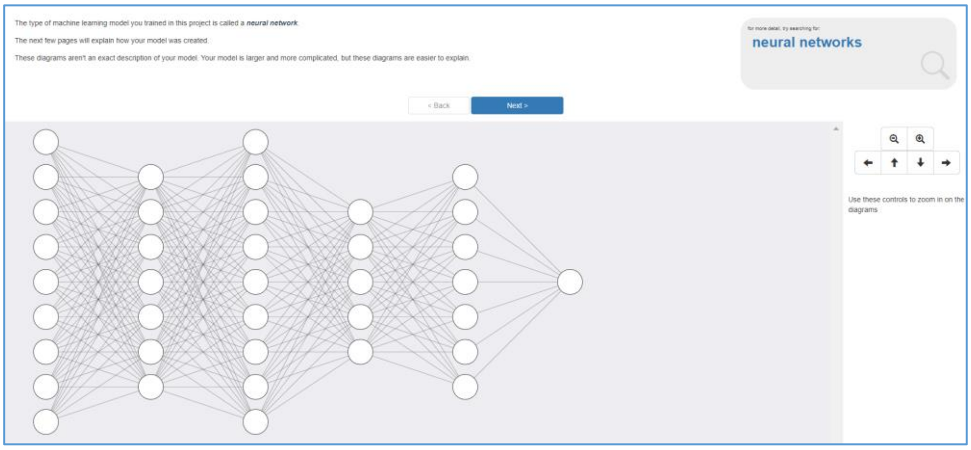


Try putting in some text to see how it is recognised based on your training.

tiger

Recognised as **Animal**
with 100% confidence

Neural Network related to the model will be displayed



The type of machine learning model you trained in this project is called a ***neural network***.

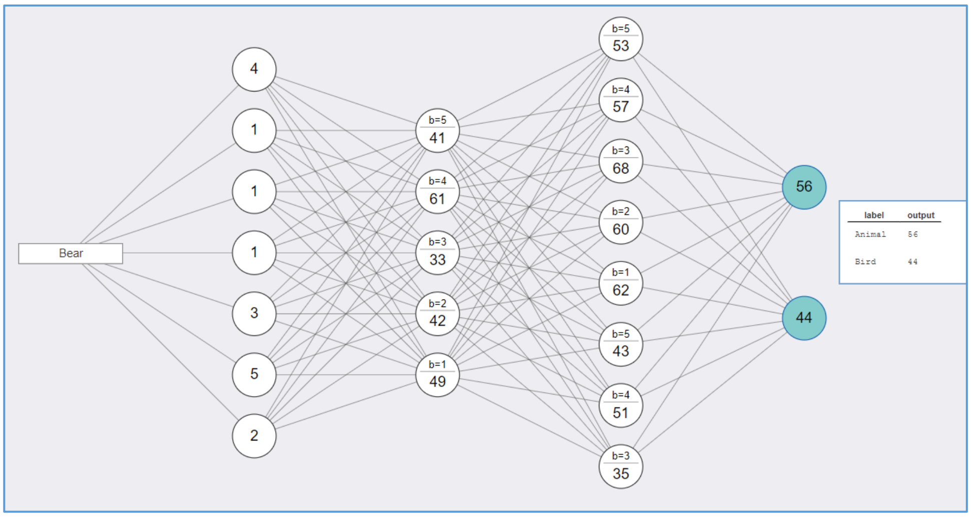
The next few pages will explain how your model was created.

These diagrams aren't an exact description of your model. Your model is larger and more complicated, but these diagrams are easier to explain.

for more detail, try searching for: neural networks

Use these controls to zoom in on the diagrams

Click on next and see the steps of Deep Learning working



6.6.2 Activity 2:

Problem: Convert from Celsius to Fahrenheit where the formula is: f=c×1.8+32

It would be simple enough to create a conventional Python function, but that wouldn't be machine learning. Instead, we will give TensorFlow some sample Celsius values and their corresponding Fahrenheit values. Then, we will train a model that figures out the above formula through the training process.

```
#Importing Libraries
import tensorflow as tf
import numpy as np
import matplotlib.pyplot as plt
#Training Data
c = np.array([-40, -10,  0,  8, 15, 22,  38],  dtype=float)
```

```
f = np.array([-40,  14, 32, 46, 59, 72, 100],  dtype=float)
```

#Creating a model

#Since the problem is straightforward, this network will require only a single layer, with a single neuron.

```
model =
tf.keras.Sequential([tf.keras.layers.Dense(units=1,
input_shape=[1])])
```

The model definition takes a list of layers as argument, specifying the calculation order from the input to the output.

- input_shape=[1] → This specifies that the input to this layer is a single value, i.e., the shape is a one-dimensional array with one member.
- units=1 → This specifies the number of neurons in the layer. The number of neurons defines how many internal variables the layer has to try to learn to solve the problem.
- Loss function — A way of measuring how far off the predictions are from the desired outcome. The measured difference is called the "loss".
- Optimizer function — A

#Compile, loss, optimizer

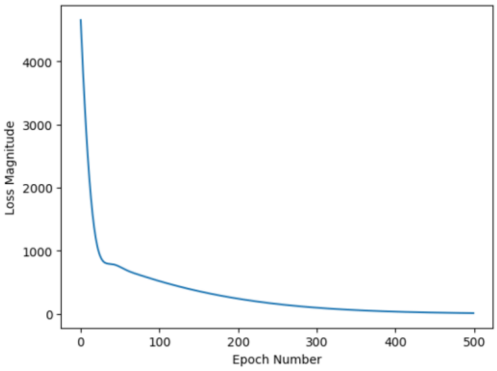
```
model.compile (loss='mean_squared_error',
optimizer=tf.keras.optimizers.Adam(0.1),
metrics=['mean_squared_error'])
```

#Train the model

```
history = model.fit(c, f, epochs=500, verbose=False)
print("Finished training the model")
```

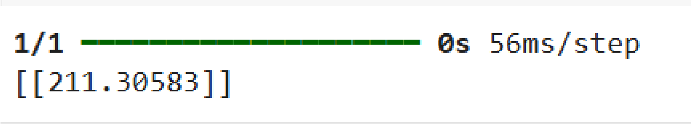
#Training Statistics

```
plt.xlabel('Epoch Number')
plt.ylabel("Loss Magnitude")
plt.plot(history.history['loss'])
plt.show()
```



#Predict Values

```
print(model.predict(np.array([100.0])))
```



```
1/1 ━━━━━━━━━━━━━━━━━━━━ 0s 56ms/step
[[211.30583]]
```

#Comparison with traditional method of using formula to convert Celsius to Fahrenheit

```
f_by_formula = (100*1.8)+32
print(f_by_formula)
212.0
```

https://www.kaggle.com/code/imdevskp/learning-celsius-to-fahrenheit-equation-with-keras/notebook

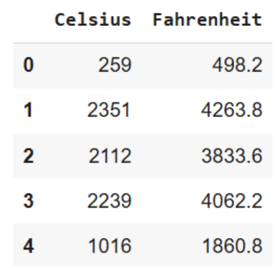
6.6.3 Activity 3: (**For advanced learners)

Creating Neural Network with Python

The following python program creates and trains a simple artificial neural network (ANN) using tensorflow and keras to predict Fahrenheit temperatures based on Celsius temperatures.

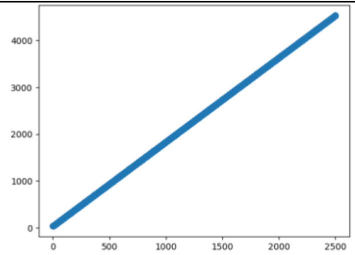
```
In [ ]:  #Import the necessary files

import numpy as np
import pandas as pd
import matplotlib.pyplot as plt
import tensorflow as tf
```

```
In [ ]: #Read the dataset from csv file

temp_df = pd.read_csv(‘cel_fah.csv’)
temp_df.head()
```



	Celsius	Fahrenheit
0	259	498.2
1	2351	4263.8
2	2112	3833.6
3	2239	4062.2
4	1016	1860.8

```
In [ ]: #Visualise relationship between celsius & Fahrenheit using a scatterplot

plt.scatter(temp_df['Celsius'], temp_df['Fahrenheit'])
```



```
In [ ]: #Assign the Celsius and Fahrenheit to variables X_train and Y_train

X_train = temp_df['Celsius']
y_train = temp_df['Fahrenheit']
```

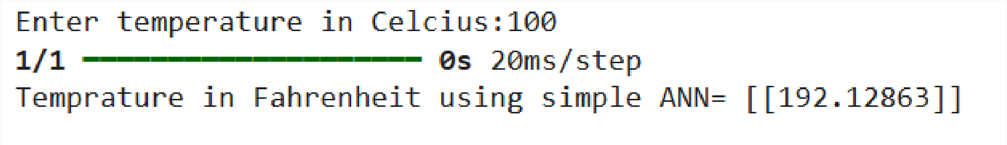
```
In [ ]: #Define a sequential neural network.

# In this neural network, layers are added sequentially, one on top of other.

model = tf.keras.Sequential()
#Define the network with 1 node in input layer,32 nodes in the first hidden layer
model.add(tf.keras.layers.Dense(units= 32 , input_shape = (1,)))
# now we are adding one more hidden layer to the network with 32 nodes
model.add(tf.keras.layers.Dense(units = 32))
# now adding the output layer
model.add(tf.keras.layers.Dense(units = 1))
```

```
In [ ]: model.summary()

'''shows The architecture of the model, showing the layers and their order.
         The output shape of each layer.
         The number of parameters (weights) in each layer.
         The total number of trainable parameters in the model.'''
```

```
In [ ]: #configure the model for training.

model.compile(optimizer= tf.keras.optimizers.Adam(.1), loss =
'mean_squared_error')
```

```
In [ ]: # train the network model by iterating 30 times with 20% of data used for
validation

epochs_hist = model.fit(X_train, y_train, epochs=30, validation_split = 0.2)
```

```
In [ ]: # DEPLOYMENT:Use the same model to perform predictions

Temp_C = float(input("Enter temperature in Celcius:"))
Temp_F = model.predict(np.array([Temp_C]))
print('Temperature in Fahrenheit using simple ANN=', Temp_F)
```



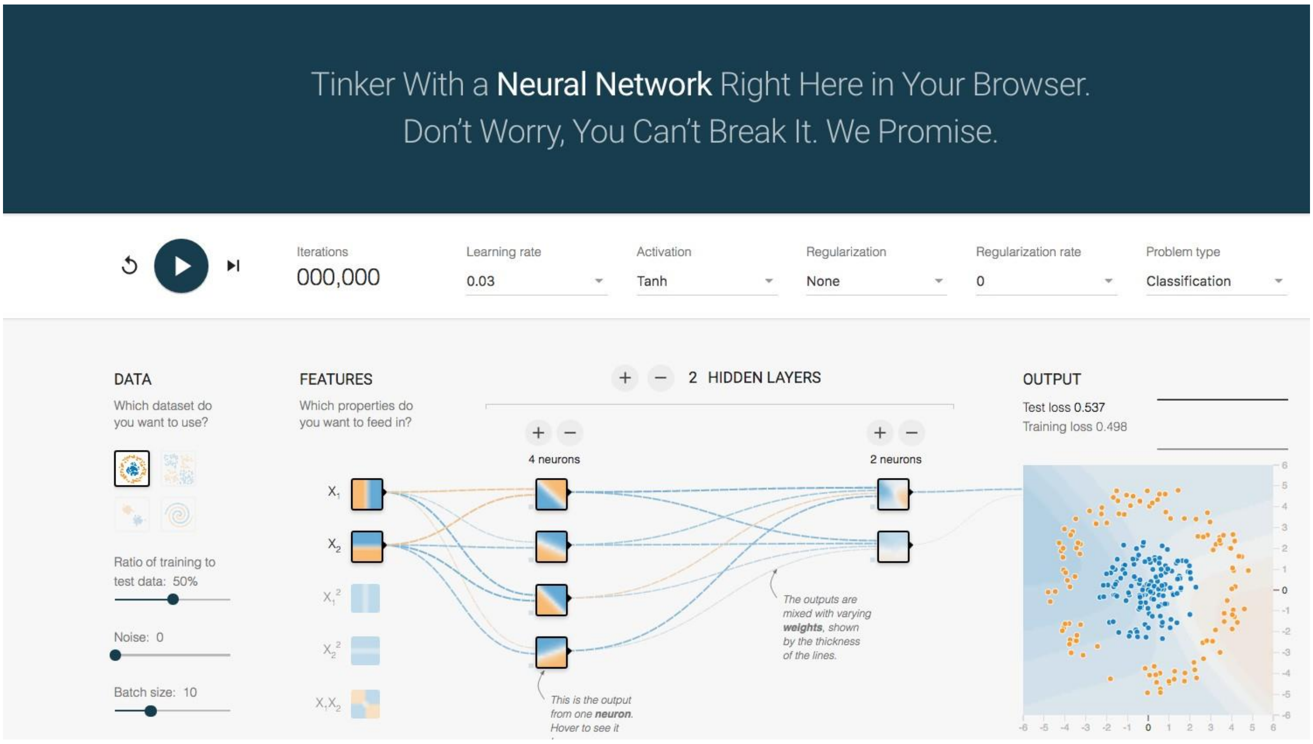
```
Enter temperature in Celcius:100
1/1 ━━━━━━━━━━━━━━━━━━━━ 0s 20ms/step
Temprature in Fahrenheit using simple ANN= [[192.12863]]
```

6.6.4 Activity 4: TensorFlow Playground

We have heard a lot about Artificial Neural Networks (ANN), Deep Learning (DL) and Machine Learning (ML). We have also heard about the different training algorithms like clustering, classification etc. But when we learn about the technology from a textbook, we may feel overwhelmed by mathematical models and formulae.

To make this easy and interesting, there's an awesome tool to grasp the idea of neural networks and different training algorithms like classification and clustering. This tool is called **TensorFlow Playground**, (https://playground.tensorflow.org) a web app written in JavaScript where a real neural network will run in the browser. By tweaking the parameters, it is possible to see how the different algorithms work.

TensorFlow Playground home screen



Tinker With a Neural Network Right Here in Your Browser.
Don't Worry, You Can't Break It. We Promise.

Iterations 000,000
Learning rate 0.03
Activation Tanh
Regularization None
Regularization rate 0
Problem type Classification

DATA
Which dataset do you want to use?
Ratio of training to test data: 50%
Noise: 0
Batch size: 10

FEATURES
Which properties do you want to feed in?

2 HIDDEN LAYERS

OUTPUT
Test loss 0.537
Training loss 0.498

The outputs are mixed with varying **weights**, shown by the thickness of the lines.

This is the output from one **neuron**. Hover to see it

First, let's understand some of the terms in the above picture.

1. **Data**

   There are six pre-loaded data sets Circle, Exclusive OR (XOR), Gaussian, Spiral, Plane and Multi-Gaussian. The first four are for classification problems and last two are for regression problems. Small circles are the data points which correspond to positive one and negative one. In general, positive values are shown in blue and negative in orange.

2. **Features**

   There are seven features or inputs (X1, X2, squares, product and sine). We can turn on and off different features to see which features are more important.

3. **Weights**

   In the hidden layers, the lines are coloured by the weights of the connections between neurons. Blue shows a positive weight, which means the network is using that output of the neuron as given. An orange line shows that the network is assigning a negative weight.

   In the output layer, the dots are coloured orange or blue depending on their original values. The background colour shows what the network is predicting for a particular area. The intensity of the colour shows how confident that prediction is.

4. **Epoch**

   Epoch is one complete iteration through the data set.

5. **Learning Rate**

   Learning rate (alpha) is responsible for the speed at which the model learns.

6. **Activation Function**

   We may skip this term for now but for the purpose of the activity, you may choose any one of the given 4 activation functions (Tanh, ReLu, Sigmoid and Linear).

7. **Regularization**

   The purpose of regularization L1 and L2 is to remove or reduce overfitting.

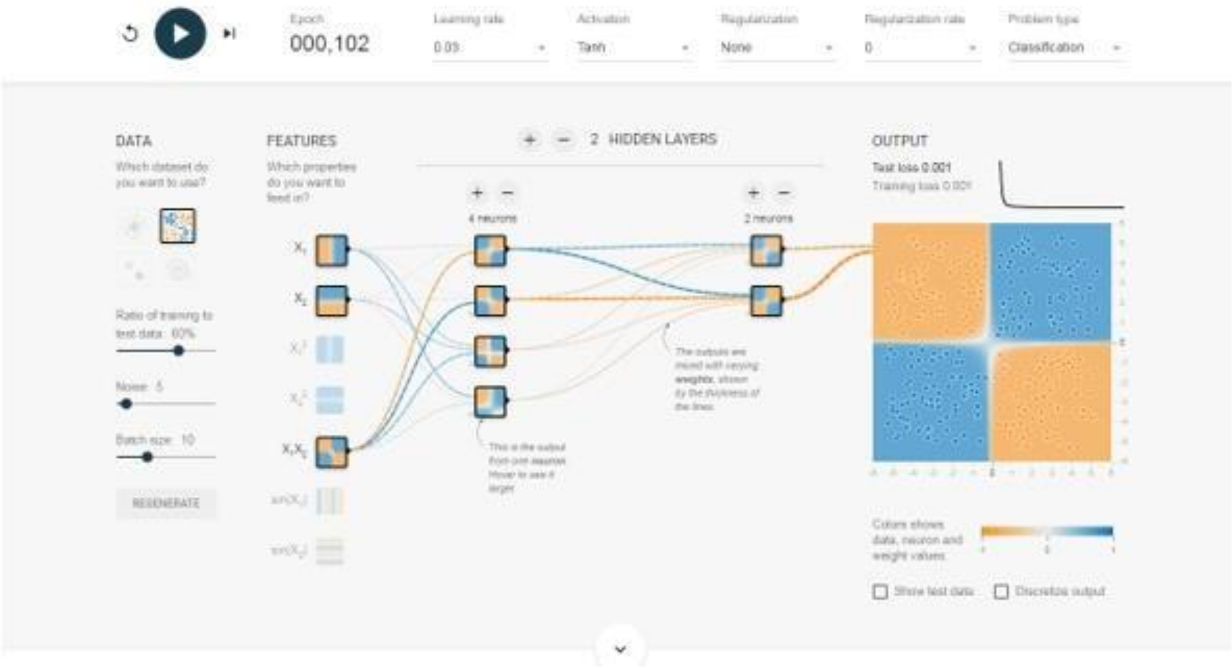
8. **Output**

   Check the model performance after training the neural network. Observe the Test loss and Training loss of the model.

Classification problem using TensorFlow playground.

Below are the steps on how to play in this neural network playground:

- Select the Exclusive OR Data Set Classification problem.
- Set Ratio of training and test data to 60% – which means we have 60% train data and 40% testing data.
- Noise is added to 5 and increase it and do some experiment with it, check how the output losses are changing and select the batch size to 10.
- First Select simple features like X1 and X2 then note down the output losses (Training loss: -0.004, Test loss: – 0.002, Steps: -255)
- Now add the third feature product of (X1X2) then observe the Losses (Training loss: -0.001, Test loss: – 0.001, Steps: -102)
- This is how you can understand the value of features, and how to get good results in minimum steps
- Set the learning rate to 0.03, also check how the learning rate plays an important role in training a neural network
- Since you have already learnt about regression, you may also play with regression, so you have a clear idea about regression.
- Select 2 hidden layers, set 4 neurons for the first hidden layer and 2 neurons for the second hidden layer then followed by the output
- Starting from the first layer the weights are passed on to the first hidden layer which contains output from one neuron, second hidden layer output is mixed with different weights. Weights are represented by the thickness of the lines
- Then the final output will contain the Train and Test loss of the neural network



Epoch 000,102

Learning rate 0.03

Activation Tanh

Regularization None

Regularization rate 0

Problem type Classification

DATA

Which dataset do you want to use?

Ratio of training to test data: 60%

Noise: 5

Batch size: 10

FEATURES

Which properties do you want to feed in?

2 HIDDEN LAYERS

4 neurons

2 neurons

This is the output from one neuron. Hover to see it larger.

The outputs are mixed with varying **weights**, shown by the thickness of the lines.

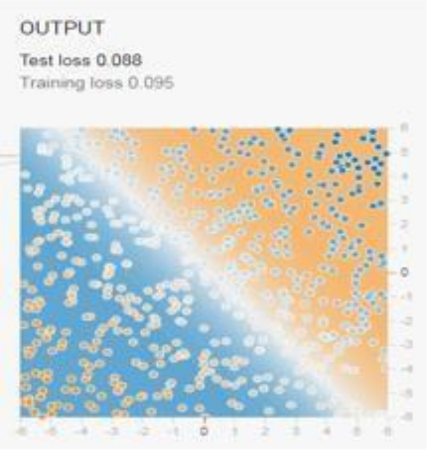
OUTPUT

Test loss 0.001

Training loss 0.001

Colors shows data, neuron and weight values.

□ Show test data □ Discretize output



OUTPUT

Test loss 0.088

Training loss 0.095

EXERCISES

A. Multiple Choice Questions:

1. What is a neural network?

A. A biological network of neurons in the brain.
B. A machine learning model inspired by the human brain.
C. A type of computer hardware used for complex calculations.
D. A mathematical equation for linear regression.

2. What are neurons in a neural network?

A. Cells in the human brain.
B. Mathematical functions that process inputs and produce outputs.
C. Nodes that make up the layers of a neural network.
D. None of the above.

3. What is the role of activation functions in neural networks?

A. They determine the learning rate of the network.
B. They introduce non-linearity, allowing the network to learn complex patterns.
C. They control the size of the neural network.
D. They define the input features of the network.

4. What is backpropagation in neural networks?

A. The process of adjusting weights and biases to minimize the error in the network.
B. The process of propagating signals from the output layer to the input layer.
C. The process of initializing the weights and biases of the network.
D. The process of defining the architecture of the neural network.

5. Which type of neural network is commonly used for image recognition?

A. Feedforward neural network.
B. Convolutional neural network.
C. Recurrent neural network.
D. Perceptron.

6. How do neural networks learn from data?

A. By adjusting their weights and biases based on the error in their predictions.
B. By memorizing the training data.
C. By using a fixed set of rules.
D. By ignoring the input data.

B. Short Answer Questions:

1. What is the purpose of an activation function in a neural network?

Ans - An activation function introduces non-linearity to the output of a neuron, allowing the neural network to model complex patterns in the data.

2. How does backpropagation work in a neural network?

Ans- Backpropagation is a training algorithm for neural networks. It calculates the error between the predicted output and the actual output, then adjusts the weights and biases of the network to minimize this error.

3. What are the key components of a neural network?

Ans- The key components of a neural network include neurons, connections (weights), activation functions, and biases.

4. Describe the structure of a feedforward neural network.

Ans- A feedforward neural network consists of an input layer, one or more hidden layers, and an output layer. The input data flows from the input layer through the hidden layers to the output layer.

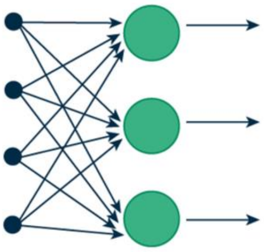
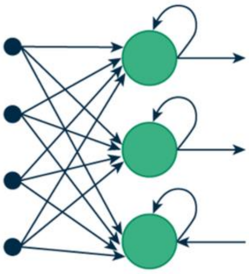
5. What are some common types of neural networks and their applications?

Ans- Common types of neural networks include feedforward neural networks (used for general-purpose learning tasks), convolutional neural networks (used for image recognition and computer vision), and recurrent neural networks (used for sequential data processing, such as natural language processing).

6. How does a neural network learn from data?

Ans- A neural network learns from data by adjusting its weights and biases based on the error in its predictions. Through repeated training on a dataset, the network improves its ability to make accurate predictions.

7. Differentiate between FNN and RNN.

FNN	RNN
Data flows only in one direction, from input to output. They use activation functions and weights to process information in a forward manner.	They have feedback connections that allow data to flow in a loop. If the prediction is wrong, the learning rate is employed to make small changes. Hence, making it gradually increase towards making the right prediction during the backpropagation.
This makes them suitable for tasks with independent inputs, like image classification	This feedback enables RNNs to remember prior inputs, making them ideal for tasks where context is important.
	

8. What are the potential future developments and trends that will shape the evolution of neural networks?

i) The sophisticated NN algorithms have significantly enhanced efficiency and productivity by automating tasks, streamlining processes, and optimizing resource allocation in industries ranging from manufacturing to finance.

ii) Neural networks have personalized products and services by analysing vast datasets, leading to tailored recommendations and experiences that cater to individual preferences and needs.

iii) Economically, the adoption of neural networks has spurred innovation, driven economic growth, and created new job opportunities in burgeoning fields like data science and artificial intelligence.

9. What are the four ways in which neural networks help A.I. grow?

i) Advancing Deep learning

ii) Improving Accuracy and Efficiency

iii) Supporting Autonomous Systems

iv) Personalizing Experience

C. Long Answer Questions:

1. Explain the concept of a neural network and its significance in machine learning. Provide examples of real-world applications where neural networks are used.

Ans- A neural network is a computer system inspired by the human brain's biological neural networks. It consists of interconnected nodes (neurons) that process information and learn patterns from data. Neural networks are significant in machine learning because they can learn from data without being explicitly programmed. They are used in various applications, such as image recognition (e.g., facial recognition in smartphones), natural language processing (e.g., voice assistants like Siri and Alexa), and autonomous driving (e.g., Tesla's self-driving cars).

2. Describe the structure of a neural network and the role of each component, including neurons, connections, activation functions, and biases.

Ans- A neural network consists of layers of neurons organized in a specific structure. The input layer receives data, which is then passed through one or more hidden layers before reaching the output layer. Neurons in each layer are connected to neurons in the adjacent layers through connections (weights). Each neuron computes a weighted sum of its inputs, applies an activation function to produce an output, and passes this output to the next layer. Activation functions introduce non-linearity to the network, allowing it to model complex patterns. Biases are constants added to the weighted sum before applying the activation function, helping the network account for any inherent bias in the data.

3. Explain how backpropagation works in neural networks and its importance in training the network.

Ans- Backpropagation is an algorithm used to train neural networks by adjusting the weights and biases based on the error in the network's predictions. It works by first making a forward pass through the network to make predictions, then calculating the error between the predicted output and the actual output. The algorithm then makes a backward pass through the network, computing the gradient of the error with respect to each weight and bias. These gradients are used to update the weights and biases using an optimization algorithm such as gradient descent. Backpropagation is crucial for training neural networks as it allows the network to learn from the error and improve its predictions over time.

4. Discuss the different types of neural networks, including feedforward neural networks, convolutional neural networks, and recurrent neural networks. Provide examples of real-world applications where each type is used.

Ans- Feedforward neural networks are the simplest type of neural network and are used for general-purpose learning tasks such as classification and regression. Convolutional neural networks (CNNs) are specialized for image recognition tasks and are used in applications like facial recognition and object detection. Recurrent neural networks (RNNs) are designed for sequential data processing and are used in applications such as speech recognition and language translation.

D. Competency Based Questions:

1.You are a data scientist working for a healthcare company, tasked with developing a neural network model to predict the likelihood of a patient having a heart attack based on their medical history. The dataset you have been provided with contains various features such as age, gender, family medical history, blood pressure, cholesterol levels, and previous medical conditions. Your task is to determine which type of neural network model would you build so that you can accurately predict whether a patient is at risk of having a heart attack.

Ans- This problem is that of binary classification so, a feedforward neural network with multiple hidden layers can be effective.

2.You are planning to go out for dinner and are deciding between two restaurants. You consider three factors which are food quality, ambience, and distance. Using a neural network concept, determine the inputs and weights.

Ans-

1. Factors Influencing Decision:

- Food Quality: High (Yes: 1, No: 0)
- Ambience: Cozy (Yes: 1, No: 0)
- Distance: Nearby (Yes: 1, No: 0)

2. Inputs and Weights:

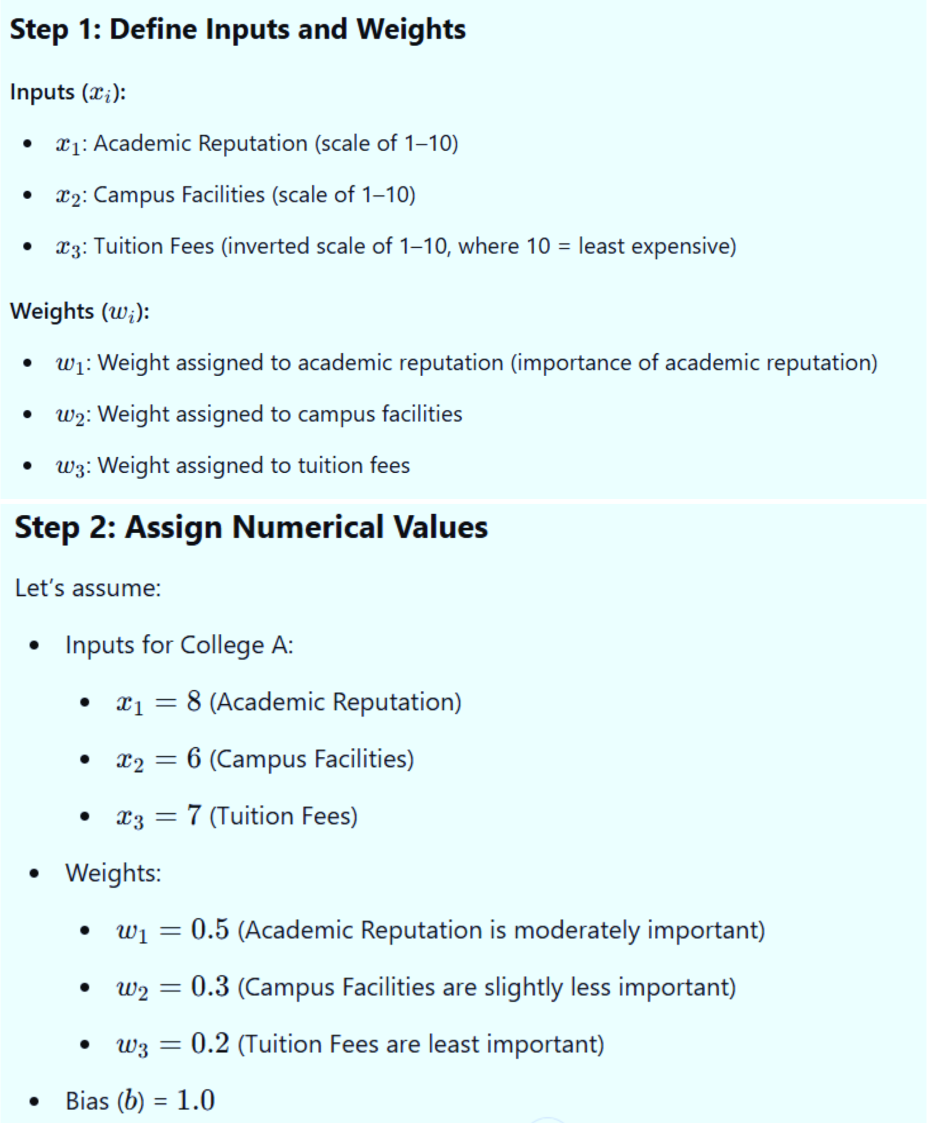
- (X1 = 1) (Food Quality is high)
- (X2 = 1) (Ambience is cozy)
- (X3 = 1) (Distance is nearby)

Weights:

- (W1 = 3) (Food Quality importance)
- (W2 = 2) (Ambience importance)
- (W3 = 1) (Distance importance)

3.You are deciding between two colleges and are considering three factors: academic reputation, campus facilities, and tuition fees. Using a neural network concept: What would be the inputs and weights? Apply the formula and calculate the outcome.

3. You are deciding between two colleges and are considering three factors: academic reputation, campus facilities, and tuition fees. Using a neural network concept: What would be the inputs and weights? Apply the formula and calculate the outcome.



Step 1: Define Inputs and Weights

Inputs (x_i):

- x_1: Academic Reputation (scale of 1–10)
- x_2: Campus Facilities (scale of 1–10)
- x_3: Tuition Fees (inverted scale of 1–10, where 10 = least expensive)

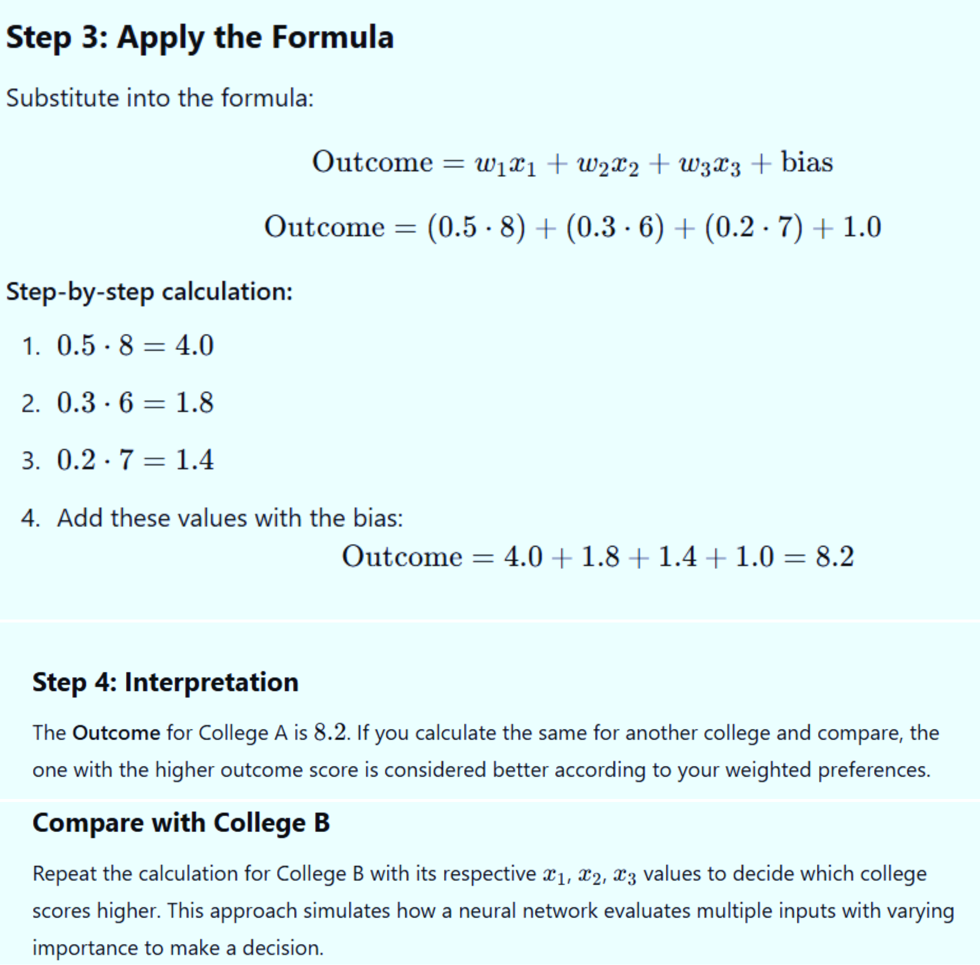
Weights (w_i):

- w_1: Weight assigned to academic reputation (importance of academic reputation)
- w_2: Weight assigned to campus facilities
- w_3: Weight assigned to tuition fees

Step 2: Assign Numerical Values

Let's assume:

- Inputs for College A:
  - $x_1 = 8$ (Academic Reputation)
  - $x_2 = 6$ (Campus Facilities)
  - $x_3 = 7$ (Tuition Fees)
- Weights:
  - $w_1 = 0.5$ (Academic Reputation is moderately important)
  - $w_2 = 0.3$ (Campus Facilities are slightly less important)
  - $w_3 = 0.2$ (Tuition Fees are least important)
- Bias (b) = 1.0



Step 3: Apply the Formula

Substitute into the formula:

$$\text{Outcome} = w_1x_1 + w_2x_2 + w_3x_3 + \text{bias}$$

$$\text{Outcome} = (0.5 \cdot 8) + (0.3 \cdot 6) + (0.2 \cdot 7) + 1.0$$

Step-by-step calculation:

1. $0.5 \cdot 8 = 4.0$
2. $0.3 \cdot 6 = 1.8$
3. $0.2 \cdot 7 = 1.4$
4. Add these values with the bias:
$$\text{Outcome} = 4.0 + 1.8 + 1.4 + 1.0 = 8.2$$

Step 4: Interpretation

The **Outcome** for College A is 8.2. If you calculate the same for another college and compare, the one with the higher outcome score is considered better according to your weighted preferences.

Compare with College B

Repeat the calculation for College B with its respective x_1, x_2, x_3 values to decide which college scores higher. This approach simulates how a neural network evaluates multiple inputs with varying importance to make a decision.

4.You work for a company that develops software for recognizing handwritten digits. Your task is to create a neural network model to accurately classify handwritten digits from the dataset. Which neural network would you use for this?

Ans- Convolutional Neural Network

5.You have built a CNN model for the task in Case study – 2. List the training process that you will use to train the neural network.

Ans-Training the Model:

- Compile the model with suitable loss function and optimizer.
- Train the model on the training dataset for a certain number of epochs.
- Monitor training progress using validation data and adjust hyperparameters accordingly.

6.You want to do time series forecasting by feeding historical data into a neural network and training it to predict future values based on past observations. Which neural network should you use?

Ans- Recurrent Neural Network

REFERENCES:

1. https://machinelearningforkids.co.uk/#!/projects
2. Types of Neural Networks and Definition of Neural Network (mygreatlearning.com)
3. https://web.pdx.edu/~nauna/week7b-neuralnetwork.pdf
4. https://realpython.com/courses/build-neural-network-python-ai/
5. https://www.geeksforgeeks.org/neural-networks-a-beginners-guide/
6. https://medium.com/data-science-365/overview-of-a-neural-networks-learning-process-61690a502fa

UNIT 7: Generative AI

Title: Generative AI

Approach: Hands-on, Team Discussion, Web search, Case studies

Summary:

This unit delves into Generative AI, focusing on its ability to create new content resembling training samples across text, images, audio, and video. It covers how these systems learn from data, the difference between generative and discriminative models, and their diverse practical applications.

Learning Objectives:

1. Understand the foundational principles of Generative AI and its operational mechanisms.
2. Differentiate between generative and discriminative models in machine learning.
3. Explore the diverse applications of Generative AI, including image and text generation, audio production, and video creation.
4. Examine the ethical and social implications of Generative AI technologies.

Key Concepts:

1. Introduction to Generative AI
2. Working of Generative AI
3. Generative and Discriminative models
4. Applications of Generative AI
5. LLM- Large Language Model
6. Future of Generative AI
7. Ethical and Social Implications of Generative AI

Learning Outcomes:

Students will be able to -

1. Articulate the principles behind Generative AI, including how these models are trained and how they generate new content.
2. Demonstrate the ability to use Generative AI tools for creative and analytical purposes, applying their knowledge to generate images, text, audio, and video.
3. Evaluate the use cases of Generative AI, distinguishing between its capabilities and those of discriminative models.
4. Examine the ethical, social, and legal implications of Generative AI, including issues related to deep fakes, copyright, and data privacy.
5. Use Generative AI technologies to create novel content, applying ethical guidelines and critical thinking to their work.
6. Create a chatbot using Gemini API

Prerequisites:

1. Foundational understanding of AI concepts from class XI.

2. Understanding of basic Python, installation and importing of packages.

Unleashing Creativity with Machines: A Teacher's Guide to Generative AI

This lesson empowers you to introduce students to the fascinating world of Generative AI, where machines learn to create entirely new content!

1. The Magic of Machine-Made Content:

- **Captivating Introduction:** Spark curiosity with an engaging question: "Can machines be creative?" Introduce Generative AI, a branch of Artificial Intelligence that allows computers to create new and original content, resembling the training data they are exposed to.

2. From Inspiration to Creation: Understanding Generative AI:

- **Generative AI Basics:** Explain the core concept of Generative AI:
  - Content Creation Resemblance: These models analyse existing data and learn to generate new content that resembles the training data.
- **Applications Across Domains:** Showcase the diverse applications of Generative AI:
  - Image Generation: Creating new images based on existing styles or concepts.
  - Audio Generation: Composing music or generating realistic sound effects.
  - Text Generation: Writing different kinds of creative text formats like poems, scripts, or code.
  - Video Generation: Creating realistic videos from scratch or modifying existing ones.
- **Ethical Considerations:** Introduce the importance of ethical considerations surrounding Generative AI from the outset. Briefly discuss potential challenges like deepfakes and bias.

3. Discriminative vs. Generative: Understanding the Model Types:

- **Model Differentiation:** Distinguish between Generative and Discriminative models:
  - Generative Models: Learn to create new samples resembling the training data.
  - Discriminative Models: Focus on defining class boundaries within data for classification tasks (e.g., identifying spam emails).

4. The Power of Large Language Models (LLMs):

- **Text Powerhouses:** Introduce Large Language Models (LLMs) – a powerful type of Generative AI excelling at text generation. Discuss their capabilities in:
  - Text Generation: Creating different creative text formats, translating languages, and writing different kinds of content.
  - Audio and Video Generation (Advanced): LLMs can be used in conjunction with other AI techniques for audio and video generation.

- **LLM Applications:** Highlight the applications of LLMs:
    - Content creation assistance (e.g., writing summaries or marketing copy).
    - Personalized education and learning experiences.
    - Improved human-computer interaction through chatbots and virtual assistants.

5. Hands-on Learning with Generative AI Tools:

- **Interactive Learning:** Introduce user-friendly Generative AI tools for students to explore:
    - Canva for Education: Offers AI-powered features for image and video creation with design templates and prompts.
    - ChatGPT: A large language model for text generation, allowing students to experiment with creative writing prompts or code generation (with adult supervision and focus on responsible use).
- **Creative Learning Experiences:** Facilitate interactive activities using these tools:
    - Students can generate images based on different artistic styles using Canva.
    - Students can use prompts in ChatGPT to create short stories or poems (focusing on factual and unbiased prompts).

6. The Responsibility of Creativity: Ethical Considerations:

- **Exploring Challenges:** Delve deeper into the ethical challenges of Generative AI:
    - Deepfakes: Creating realistic but fabricated videos that can be used for malicious purposes.
    - Bias: Generative models can perpetuate biases present in the training data.
    - Copyright: Questions arise regarding ownership of content created by AI.
    - Transparency: Understanding how Generative AI models work and their limitations.
- **Fostering Responsible Development:** Discuss the importance of responsible development and deployment of Generative AI technologies. Encourage students to think critically about the potential impact of this technology.

Additional Tips:

- Utilize visuals like examples of Generative AI creations to illustrate the concepts.
- Encourage students to discuss potential future applications of Generative AI across various industries.
- Promote responsible use of AI tools and emphasize the importance of factual prompts and avoiding the spread of misinformation.

By incorporating these elements, you can equip students to understand the potential of Generative AI for creative expression while fostering critical thinking about its ethical implications and responsible use.

Teachers can ask the following questions to spark curiosity before starting the topics:

- **Have you ever seen an image or video online that seemed too good to be true? How could you tell if it might be fake?** (This question prompts students to think critically about the media they consume and the challenges of distinguishing real content from artificial content. It sparks curiosity about how Generative AI can be used to create realistic-looking fakes.)
- **Imagine you could create new and interesting things using a computer program. What kind of content would you generate (e.g., images, stories, music)?** (This question taps into students' creativity and gets them thinking about the potential applications of Generative AI. It helps bridge the gap between understanding the technology and its potential uses in various domains.)

7.1 Introduction to Generative AI

In recent times, we all have come across these pieces of information widely being circulated in news or online media. Celebrities and famous people are being targeted with fake images, damaging their reputations and self-esteem. Additionally, some individuals are using AI tools to generate content and claiming it as their own, thereby misusing the power of AI technology. But do you know what is behind all of this? This chapter takes you on an interesting journey where you will explore a new dimension of AI, known as Generative AI.

INDIA'S CRACKDOWN ON DEEPFAKES
RISING THREAT OF DEEPFAKES IN INDIA
INDIA'S I.T. MINISTRY TO ESTABLISH DEEPFAKE GUIDELINES
WION
ChatGPT
Bard AI
LLM
LARGE LANGUAGE MODEL

Fig. 7.1

Generative AI, a facet of artificial intelligence, creates diverse content like audio, text, images, and more, aiming to generate new data resembling its training samples. It utilises machine learning algorithms to achieve this, learning from existing datasets. Examples include ChatGPT, Gemini, Claude, and DALL-E.

7.2 Working of Generative AI

Generative AI learns patterns from data and autonomously generates similar samples. It operates within the realm of deep learning, employing neural networks to understand intricate patterns. Models such as Generative Adversarial Networks (GANs) and Variational Autoencoders (VAEs) facilitate tasks like image and text generation.

1. **Generative Adversarial Networks (GANs):** A Generative Adversarial Network is a type of neural network architecture. It consists of two networks, a generator, and a discriminator, that compete against each other. The generator generates new data samples, such as images or text (which are fake), while the discriminator evaluates these samples to distinguish between real and fake data. The generator aims to produce samples that are indistinguishable from real data, while the discriminator aims to differentiate between real and generated data. Through adversarial training, where these networks challenge one another, GANs learn to generate increasingly realistic samples. GANs have been successfully applied in various domains, including image generation, style transfer, and data augmentation.

2. **Variational Autoencoders (VAEs):** VAEs, or Variational Autoencoders, are computer programs designed to learn from data uniquely. They consist of two parts: an encoder, which understands the data and converts it into a hidden space called a latent space (latent space is the compressed representation of the actual data), and a decoder, which translates the information back from this hidden space into its original form. Unlike some other similar programs, like GANs, VAEs focus on capturing the underlying patterns of the data to generate new samples. They find applications in various tasks such as Data generation, detecting anomalies in data, and filling in missing information.

Both GANs and VAEs are powerful generative models but have distinct strengths. GANs are excellent for creating visually realistic outputs, while VAEs are better suited for structured data generation and tasks requiring interpretable latent spaces.

7.3 Generative and Discriminative Models

With the vast amount of data generated globally every day, acquiring new data is straightforward. However, addressing this immense volume of data requires either the development of new algorithms or the scaling up of existing ones. These algorithms, rooted in mathematics—particularly calculus, probability, and statistics—can be broadly categorized into two types:

1. Discriminative models 2. Generative models

Discriminative models focus on defining class boundaries within the data, making them suitable for tasks like classification. On the other hand, generative models seek to comprehend the underlying data distribution and generate new samples. In simple words, discriminative models are used to distinguish between different categories or classes. These models learn the boundary or difference between classes based on the features of the data. For example, we want to identify an email as either "spam" or "not spam. A discriminative model would learn the features like certain words or phrases that differentiate spam emails from non-spam emails.

By efficiently tackling complex tasks, generative models play a crucial role in managing large datasets.

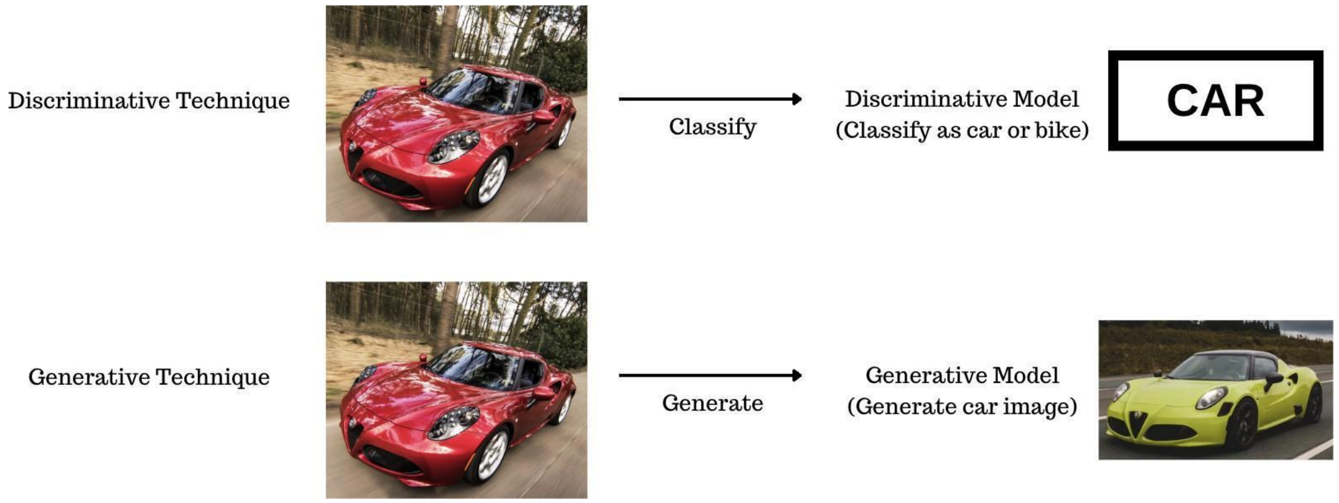


Fig. 7.2

Differences between Generative AI and Discriminative AI:

	Generative AI	**Discriminative AI**
Purpose (What is it for?)	Helps create things like images and stories and finds unusual things. It learns from data without needing to be told precisely what to do.	Helps determine what something is or belongs to by looking at its features. It is good at telling different things apart and making decisions based on that.
Models (What are they like?)	Uses tricks like making things compete against each other or making guesses based on patterns to create new things.	Learn by finding rules to separate things and recognise patterns, like understanding whether something is a dog or a cat.
Training Focus (What did they learn during training?)	Tries to understand what makes data unique and how to create new data that are similar but different.	Focuses on learning how to draw lines or make rules to tell other things apart based on their features.
Application (How are they	Helps artists create new artworks, generate new ideas for stories, and	Powers things like facial and speech recognition and helps

used in the real world?)	find unusual patterns in data.	make decisions like whether an email is spam or not.
Examples of Algorithms used	Naïve Bayes, Gaussian discriminant analysis, GAN, VAEs, LLM, DBMs, Autoregressive models	Logistic Regression, Decision Trees, SVM, Random Forest

7.4 Applications of Generative AI

Image Generation:

It involves creating new images based on patterns learned from existing datasets. These models analyse the characteristics of input images and generate new ones that exhibit similar features. It is like computers producing new pictures resembling ones they have seen before. For example, imagine a computer generating fresh cat images based on the ones it has encountered in its training data. Examples include Canva, DALL-E, Stability AI, and Stable Diffusion.

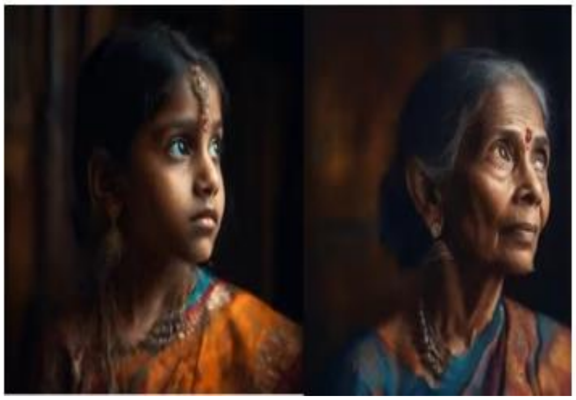
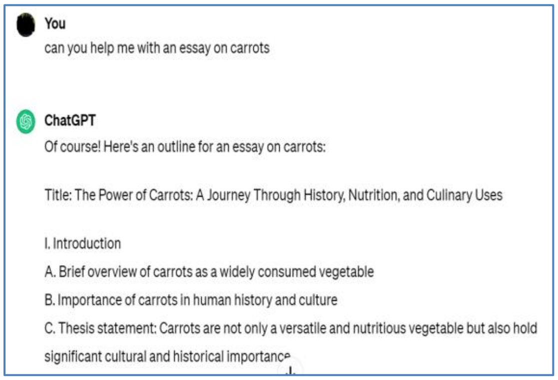


Fig 7.3



You

can you help me with an essay on carrots

ChatGPT

Of course! Here's an outline for an essay on carrots:

Title: The Power of Carrots: A Journey Through History, Nutrition, and Culinary Uses

I. Introduction

A. Brief overview of carrots as a widely consumed vegetable

B. Importance of carrots in human history and culture

C. Thesis statement: Carrots are not only a versatile and nutritious vegetable but also hold significant cultural and historical importance

Source: Software Snapshot

Fig. 7.4

Text Generation:

Text generation is when computers write sentences that sound like people wrote them. It involves creating written content that mimics human language patterns. These models analyse text data to produce coherent and contextually relevant text. They learn from many written words to create new sentences that make sense. For example, an AI tool/application that might compose a story that reads like a human authored it. Examples include OpenAI's ChatGPT, Perplexity and Google's Bard (Gemini)

Video Generation:

It involves creating new videos by learning from existing ones, including animations and visual effects. These models learn from videos to create realistic and unique visuals, producing new videos that look authentic. For instance, an AI tool/application might generate a movie scene that resembles professional filmmaking. Examples include Google's Lumiere and Deepfake algorithms for modifying video content.

A sample video generated by Meta's new AI text-to-video model, Make-A-Video. The text prompt used to create the video was "a teddy bear painting a portrait."
Image: Meta

Audio Generation: Audio generation involves computers producing new sounds, such as music or voices, based on sounds they have heard. It involves generating fresh audio content, including music, sound effects, and speech, using AI models. These models derive inspiration from existing audio recordings to generate new audio samples. They learn from existing sounds to create new ones. For instance, an AI tool/application might compose a song that sounds like a real band performed it. Examples include Meta AI's Voicebox and Google's Music LM.

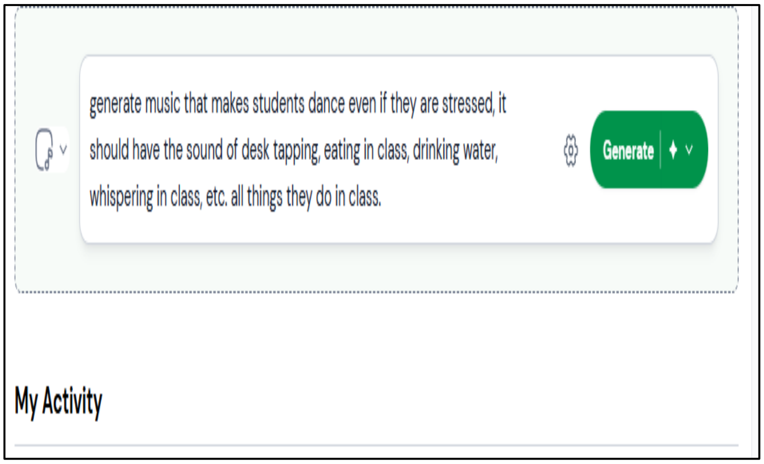

generate music that makes students dance even if they are stressed, it should have the sound of desk tapping, eating in class, drinking water, whispering in class, etc. all things they do in class.

Figure 7.6 Google's Music LM to generate music from text.

7.5 LLM- Large Language Model

Teachers can ask the following questions to spark curiosity before starting the topics:

1. **Can machines be creative? Why or why not? (**This question primes students to consider the capabilities of LLMs in tasks traditionally associated with human creativity, sparking curiosity about how these models can be used for creative writing, music generation, or other applications.)
2. **Imagine you have a magic writing assistant that can help you with different creative tasks. What kind of help would you find most beneficial? (e.g., brainstorming ideas, checking grammar, generating different writing styles)** (This question taps into students' creativity and gets them thinking about the potential benefits of LLMs for tasks like writing and content creation. It helps bridge the gap between understanding the technology and its potential usefulness in various domains.)

A Large Language Model (LLM) is a deep learning algorithm that can perform a variety of Natural Language Processing (NLP) tasks, such as generating and classifying text, answering questions in a conversational manner, and translating text from one language to another.

Fig. 7.7

Large Language Models (LLMs) are called large because they are trained on massive datasets of text and code. These datasets can contain trillions of words, and the quality of the dataset will affect the language model's performance.

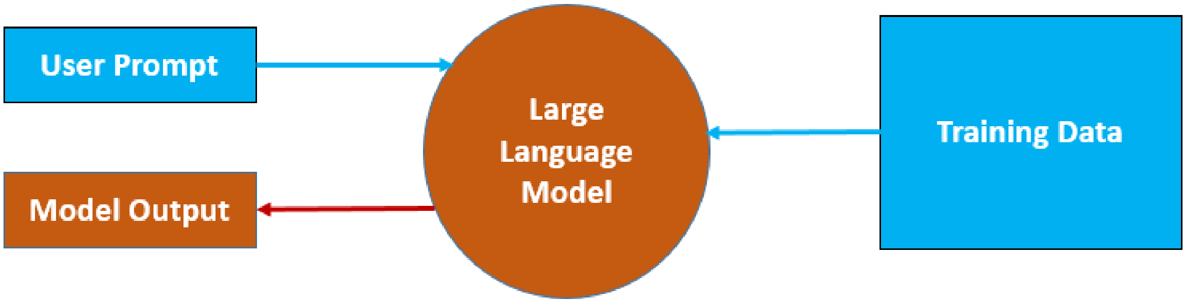


Fig. 7.8

Image source: https://levelup.gitconnected.com/understanding-the-magic-of-large-language-model-architecture-transformer-models-exposed-f5164b5db174

Transformers in LLMs:

Transformers are a type of neural network architecture that has revolutionized the field of Natural Language Processing (NLP), particularly in the context of Large Language Models (LLMs). The primary use of transformers in LLMs is to enable efficient and effective learning of complex language patterns and relationships within vast amounts of text data.

Some leading Large Language Models (LLMs) are:

- OpenAI's GPT-4o: GPT-4 is multimodal and excels in processing and generating both text and images.
- Google's Gemini 1.5 Pro: Integrates advanced multimodal capabilities for seamless text, image, and speech understanding.
- Meta's LLaMA 3.1: Open-source and optimized for high efficiency in diverse AI tasks.
- Anthropic's Claude 3.5: Prioritizes safety and interpretability in language model interactions.
- Mistral AI's Mixtral 8x7B: Implements a sparse mixture of experts for superior performance with smaller model sizes.

Applications of LLMs:

- **Text Generation**: LLMs are primarily employed for text generation tasks, including content creation, dialogue generation, story writing, and poetry generation. They can produce coherent and contextually relevant text based on given prompts or input. Other examples include
  - translating natural language descriptions into working code and streamlining development processes.
  - autocompleting text and generating continuations for sentences or paragraphs, enhancing writing tools, and email auto-completion.

- **Audio Generation**: While LLMs themselves do not directly generate audio signals, they can indirectly influence audio generation tasks through their text-to-speech (TTS) capabilities. By generating textual descriptions or scripts, LLMs enable TTS systems to synthesize natural-sounding speech from text inputs.
- **Image Generation**: LLMs have been adapted for image captioning tasks, where they generate textual descriptions or captions for images. While they do not directly generate images, their understanding of visual content can be leveraged to produce relevant textual descriptions, enhancing image accessibility and understanding.
- **Video Generation**: Similarly, LLMs can contribute to video generation tasks by generating textual descriptions or scripts for video content. These descriptions can be used to create subtitles, captions, or scene summaries for videos, improving accessibility and searchability.

Limitations of LLM:

- Processing text requires significant computational resources, leading to high response time and costs.
- LLMs prioritise natural language over the accuracy, which may result in generating factually incorrect or misleading information with high confidence.
- LLMs might memorize specific details rather than generalize, leading to poor adaptability.

Risks associated with LLM:

- Trained on Internet text, LLMs may exhibit biases, and concerns arise regarding data privacy when personal information is processed.
- Using sensitive data in training can inadvertently reveal confidential information.
- Inputs intentionally crafted to confuse the model may lead to harmful or illogical outputs.

Case Study: Demystifying LLaMA - A Publicly Trained Powerhouse in the LLM Landscape

Large Language Models (LLMs) are a type of deep learning algorithm revolutionizing Natural Language Processing (NLP). These advanced AI models excel at various tasks, including generating human-quality text, translating languages, and answering questions in a conversational way. Their effectiveness hinges on the quality and size of the training data. Traditionally, LLMs are trained on massive, proprietary datasets, limiting transparency and accessibility. This case study explores LLaMA, a unique LLM developed by Meta AI that stands out for its approach to training data and architecture.

LLaMA's Disruptive Training Approach: LLaMA leverages publicly available text and code scraped from the internet for training. This *focus on open data, fosters transparency* within the research community and allows for wider accessibility. Additionally, LLaMA's developers implemented efficient training techniques, *requiring less computational power* compared to some LLMs. This translates to better scalability, making LLaMA a more feasible solution for deployment on a wider range of devices with varying processing capabilities.

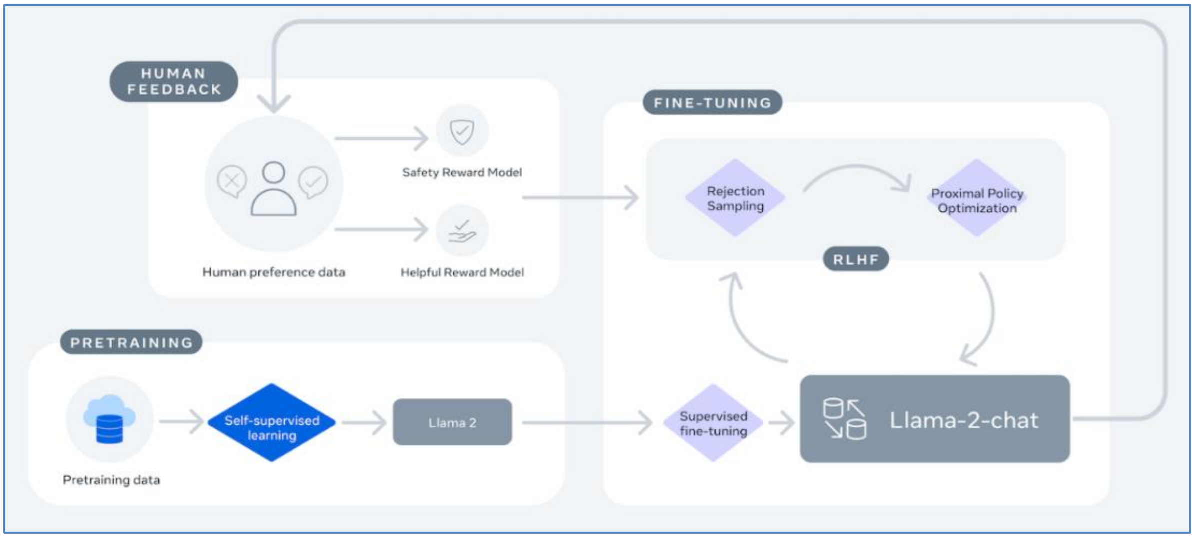


Fig. 7.8

Flexibility Through Multi-Model Design: LLaMA offers flexibility through its multi-model design. Meta releases LLaMA in various sizes, ranging from 7 billion to 65 billion parameters. This allows users to choose the model that best suits their specific needs and computational resources. Smaller models can be deployed on devices with lower processing power, ideal for everyday tasks. Conversely, larger models offer superior performance for complex tasks requiring more intensive processing power.

Impressive Results Despite Publicly Trained Data: Despite its focus on efficient training with publicly available data, LLaMA delivers impressive results. It benchmarks competitively or even surpasses some larger LLMs on various NLP tasks. Areas of particular strength include *text summarization and question answering*, showcasing LLaMA's ability to grasp complex information and generate concise, informative outputs. This positions LLaMA as a valuable tool with promising applications in education (*personalized learning materials*), content creation (*brainstorming ideas, generating drafts, translation*), and research assistance (*analysing vast amounts of data*).

7.6 Future of Generative AI

The future of AI focuses on evolving architectures to surpass current capabilities while prioritizing ethical development to minimize biases and ensure responsible use. Generative AI will address complex challenges in fields like healthcare and education, enhance NLP tasks like multilingual translation, and expand in multimedia content creation. Collaboration between humans and AI will deepen, emphasizing AI's role as a supportive partner across domains.

7.7 Ethical and Social Implications of Generative AI

Teachers can ask the following questions to spark curiosity before starting the topics:

- **Imagine you see a funny video online of a celebrity doing something strange. How can you tell if the video might be fake? What are some reasons why someone might create a fake video?** (This question primes students to think critically about the authenticity of online content and the potential motives behind creating deepfakes. It sparks a discussion about the challenges deepfakes pose to trust in media and the spread of misinformation.)

- **Have you ever seen an advertisement online that seemed to target you specifically? How might companies use AI to personalize their marketing strategies?** (This question gets students thinking about the applications of AI in everyday life and how it can be used to influence them. It can then be transitioned into a discussion about potential biases in AI algorithms and the importance of fairness and transparency in AI development.)

Generative AI, with its ability to create realistic content such as images, videos, and text, brings about a multitude of ethical and social considerations. While the technology offers promising applications, its potential for misuse and unintended consequences raises significant concerns. In this context, understanding the ethical and social implications of generative AI becomes crucial for ensuring responsible development and deployment.

1. **Deepfake Technology:**

The emergence of deepfake AI technology, such as DeepFaceLab and FaceSwap, raises concerns about the authenticity of digital content. Deepfake algorithms can generate compelling fake images, audio, and videos, jeopardising trust in media integrity and exacerbating the spread of misinformation.

Examples: Deepfake AI tools, such as DeepArt's style transfer algorithms, can seamlessly manipulate visual content, creating deceptive and misleading media. For instance, deepfake

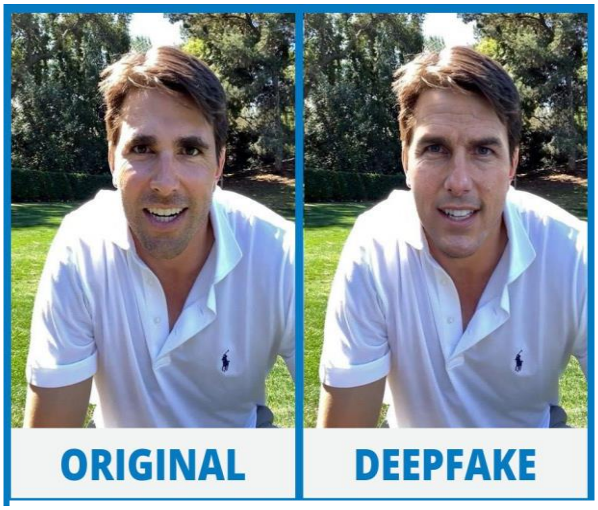


Fig. 7.8

videos have been used to superimpose individuals' faces onto adult content without their consent, leading to privacy violations and reputational damage.

2. Bias and Discrimination:

Generative AI models, exemplified by Clearview AI's facial recognition algorithms, have demonstrated biases that disproportionately affect certain demographic groups. These biases perpetuate social inequalities and reinforce stereotypes, posing ethical dilemmas regarding the fairness and impartiality of AI-generated outputs. **Examples: The AI-powered hiring platform developed by HireVue has faced criticism for perpetuating recruitment bias**. The platform may inadvertently discriminate against underrepresented groups by relying on historical hiring data, hindering diversity and inclusivity in employment opportunities.

3. Plagiarism:

Presenting AI-generated content as one's work, whether intentionally or unintentionally, raises ethical questions regarding intellectual property rights and academic integrity. Moreover, if AI output significantly resembles copyrighted material, it could potentially infringe upon copyright laws, leading to legal ramifications.

4. Transparency:

Transparency in the use of generative AI is paramount to maintaining trust and accountability. Disclosing the use of AI-generated content, particularly in academic and professional settings, is essential to uphold ethical standards and prevent instances of academic dishonesty. Failure to disclose AI use can erode trust and undermine the credibility of research and scholarly work.

Points to Remember:

- Be cautious and transparent when using generative AI.
- Respect copyright and avoid presenting AI output as your own.
- Consult your teacher/institution for specific guidelines.

Citing Sources with Generative AI:

- **Intellectual Property**: Ensure proper attribution for AI-generated content to respect original creators and comply with copyright laws.
- **Accuracy**: Verify the reliability of AI-generated information and cite primary data sources whenever possible to maintain credibility.
- **Ethical Use**: Acknowledge AI tools and provide context for generated content to promote transparency and ethical use.

Citation Example:

1. **Treat the AI as author**: Cite the tool name (e.g., Bard) & "Generative AI tool" in the author spot.
2. **Date it right**: Use the date you received the AI-generated content, not any tool release date.
3. **Show your prompt**: Briefly mention the prompt you gave the AI for reference (optional).

APA reference for text generated using Google Gemini on February 20, 2024:
(Optional in parentheses): "Prompt: Explain how to cite generative AI in APA style."
Bard (Generative AI tool). (2024, February 20). How to cite generative AI in APA style. [Retrieved from [invalid URL removed]]
Note: Generative AI is a weak AI

Hands-on Exploration Activity: Generative AI Tools

Activity 1. Signing Up for Canva for Education

Site URL: Canva for Education

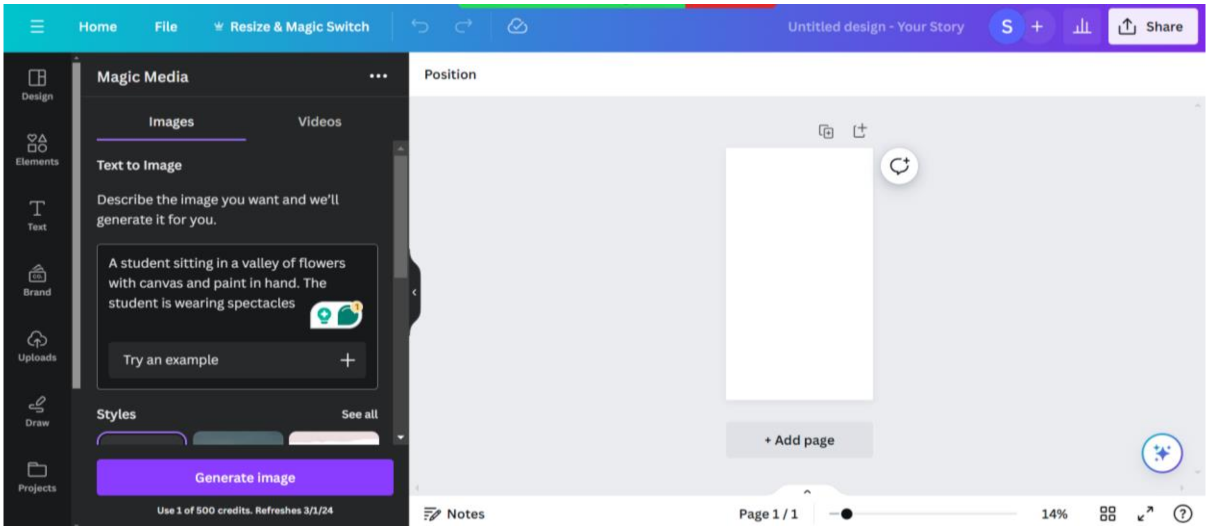
- *Visit* Canva for Education Website: Access the Canva for Education platform through your web browser.
- *Sign Up*: Click the "Sign Up" or "Create Account" option to begin registration.
- *Provide Details*: Create an account and enter your educational institution's information, including your school email address.
- *Verification*: Verify your email address by following the instructions sent to your school email inbox.
- *Access Canva for Education*: Once verified, log in to your Canva for Education account to start using the platform.

Instructions for Using AI Text to Image in Canva

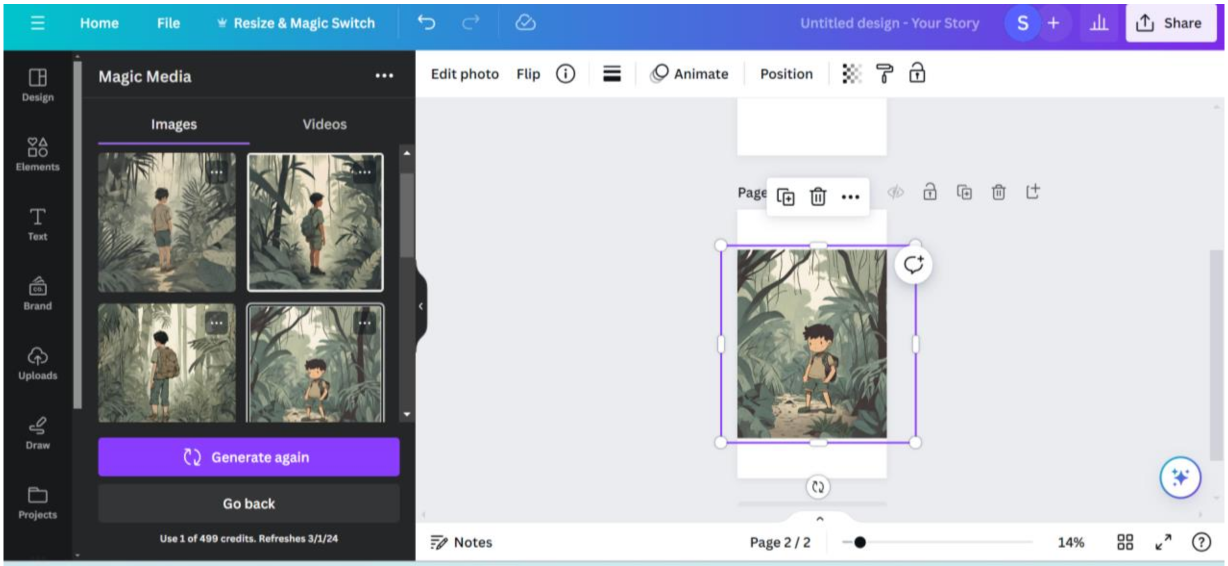
- *Access Text to Image Feature*: Open Canva and navigate to the "Text to Image" tool in the design tools menu.
- *Enter Text Prompt*: Type your text prompt or description into the provided text box. For example, "A red balloon floating in a clear blue sky."
- *Generate Image*: Click the "Generate Image" button to instruct Canva's AI to create an image based on your text prompt.

***Explore Results*: Canva will generate multiple images based on your input. Browse through the generated images to find the one that best matches your vision.**

- *Customize (Optional)*: If desired, you can customize the generated image by adjusting colours, shapes, and backgrounds using Canvas editing tools.
- *Download or Use in Design*: Once satisfied with the image, you can download it to your computer or directly incorporate it into your design project within Canva.



A student sitting in a valley of flowers with canvas and paint in hand. The student is wearing spectacles



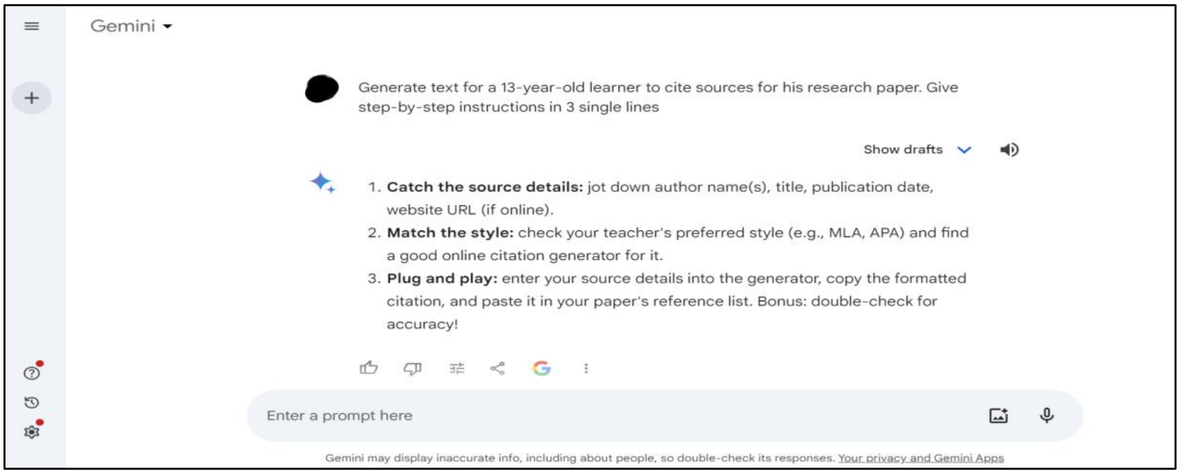
Activity 2. Signing Up for Google Gemini **Site URL: [Google Gemini]**

- Access Google Account: Ensure you have a Google account. If not, create one by visiting the Google account creation page and following the prompts to set up your account.
- Navigate to Google Gemini: Once logged in to your Google account, navigate to the Google Gemini platform through your web browser or Google Workspace.

- Agree to Terms: Review and agree to the terms of service and privacy policy for Google Gemini. This step may vary depending on your region and Google's current policies.

Instructions for Using Google Gemini's AI Text Generation

- Access Gemini Platform: Log in to your Google account and navigate to the Gemini platform.
- Craft Effective Prompt: Create a descriptive and specific text prompt that conveys the content you want to generate.
- Generate Text: Click the text generation option to input your prompt. Wait for Gemini to process the prompt and generate the text output. Review and use the generated text as needed.



[redacted] Generate text for a 13-year-old learner to cite sources for his research paper. Give step-by-step instructions in 3 single lines

1. **Catch the source details:** jot down author name(s), title, publication date, website URL (if online).
2. **Match the style:** check your teacher's preferred style (e.g., MLA, APA) and find a good online citation generator for it.
3. **Plug and play:** enter your source details into the generator, copy the formatted citation, and paste it in your paper's reference list. Bonus: double-check for accuracy!

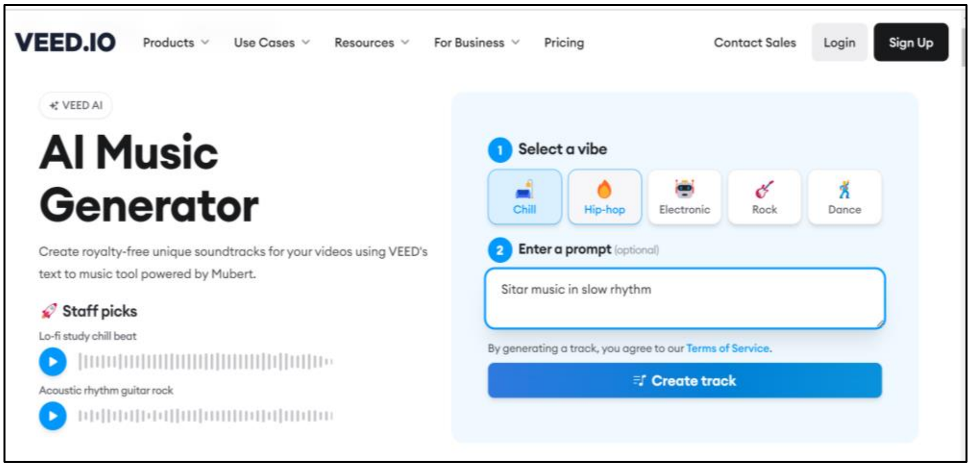
Gemini may display inaccurate info, including about people, so double-check its responses. Your privacy and Gemini Apps

Activity 3. Signing Up for Veed AI **Site URL: Veed AI Music**

- Visit Veed Website: Open your web browser and visit the Veed website.
- Click on Sign Up: Look for the "Sign Up" or "Get Started" button on the homepage and click on it.
- Provide Details: Fill out the sign-up form with your email address, password, and other required information. Once completed, click "Sign Up" to create your Veed AI account.

Instructions for Using Veed AI Text Generation

- Access Veed AI Text Generator: Log in to your Veed AI account and navigate the Text Generation tool.
- Enter Prompt: Input a concise prompt describing the text content you want to generate. For example, "Write a short story about a lost astronaut searching for their way home."
- Generate Text: Click the "Generate Text" or similar button to initiate the text generation process. Veed AI will analyse your prompt and generate a corresponding text output. Review and use the generated text for your projects or creative endeavours.

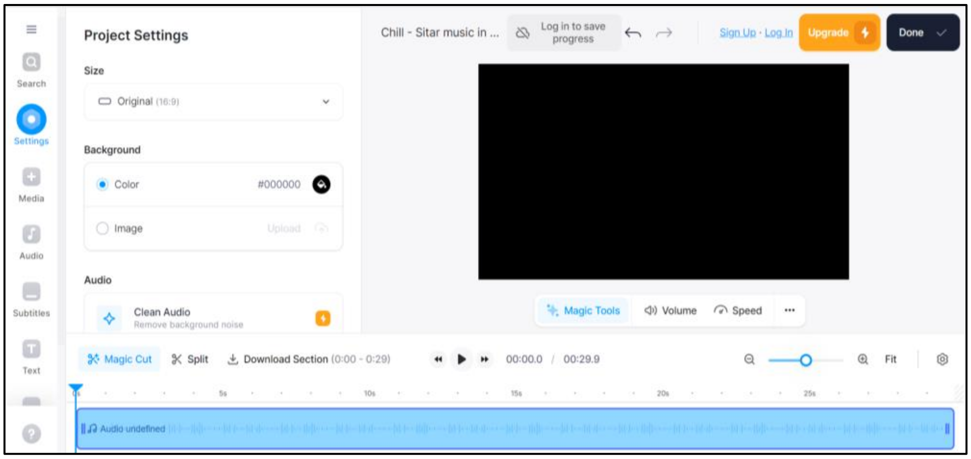


AI Music Generator

Create royalty-free unique soundtracks for your videos using VEED's text to music tool powered by Mubert.

Sitar music in slow rhythm

By generating a track, you agree to our Terms of Service.

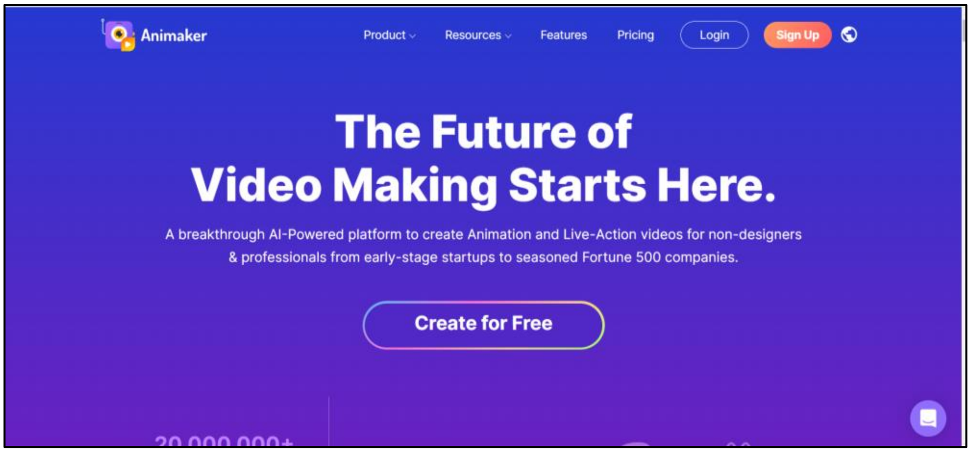


Activity 4. Signing Up for Animaker **Site URL: https://www.animaker.com/**

- Visit Animaker Website: Open your web browser and navigate to the Animaker website.
- Click on Sign Up: Look for the "Sign Up" or "Get Started" button on the homepage and click on it.
- Provide Details: Fill out the sign-up form with your email address, password, and other required information. Once completed, click "Sign Up" to create your Animaker account.

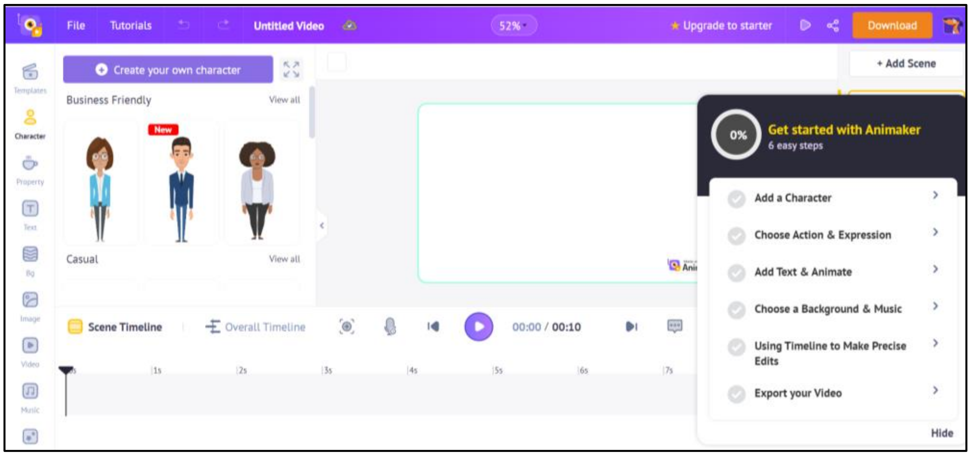
Instructions for Using Animaker AI Video Generation

- Access Animaker AI Video Tool: Log in to your Animaker account and navigate to the AI Video Generation tool.
- Enter Prompt: Input a detailed and descriptive prompt that outlines the video content you want to generate. For example, "Create a promotional video for a new product launch featuring animated characters and dynamic visuals."
- Generate Video: Click the "Generate Video" or similar button to initiate the video generation process. Animaker AI will analyse your prompt and generate a corresponding video output. Review and customise the generated video for your marketing campaigns, presentations, or storytelling projects.

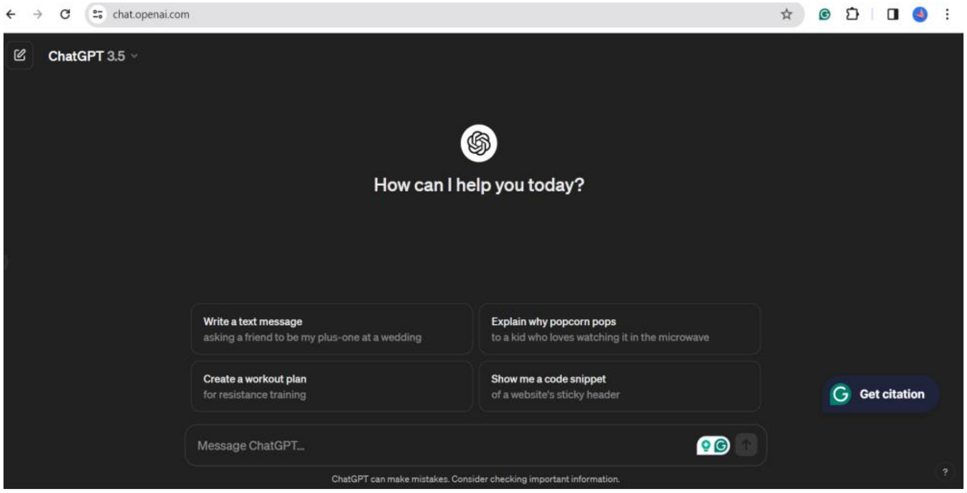


The Future of Video Making Starts Here.

A breakthrough AI-Powered platform to create Animation and Live-Action videos for non-designers & professionals from early-stage startups to seasoned Fortune 500 companies.



Activity 5. ChatGPT **Site URL: https://chat.openai.com/**



How can I help you today?

Write a text message
asking a friend to be my plus-one at a wedding

Explain why popcorn pops
to a kid who loves watching it in the microwave

Create a workout plan
for resistance training

Show me a code snippet
of a website's sticky header

ChatGPT can make mistakes. Consider checking important information.

Activity 6. Creating a Customized Chatbot with Gemini API

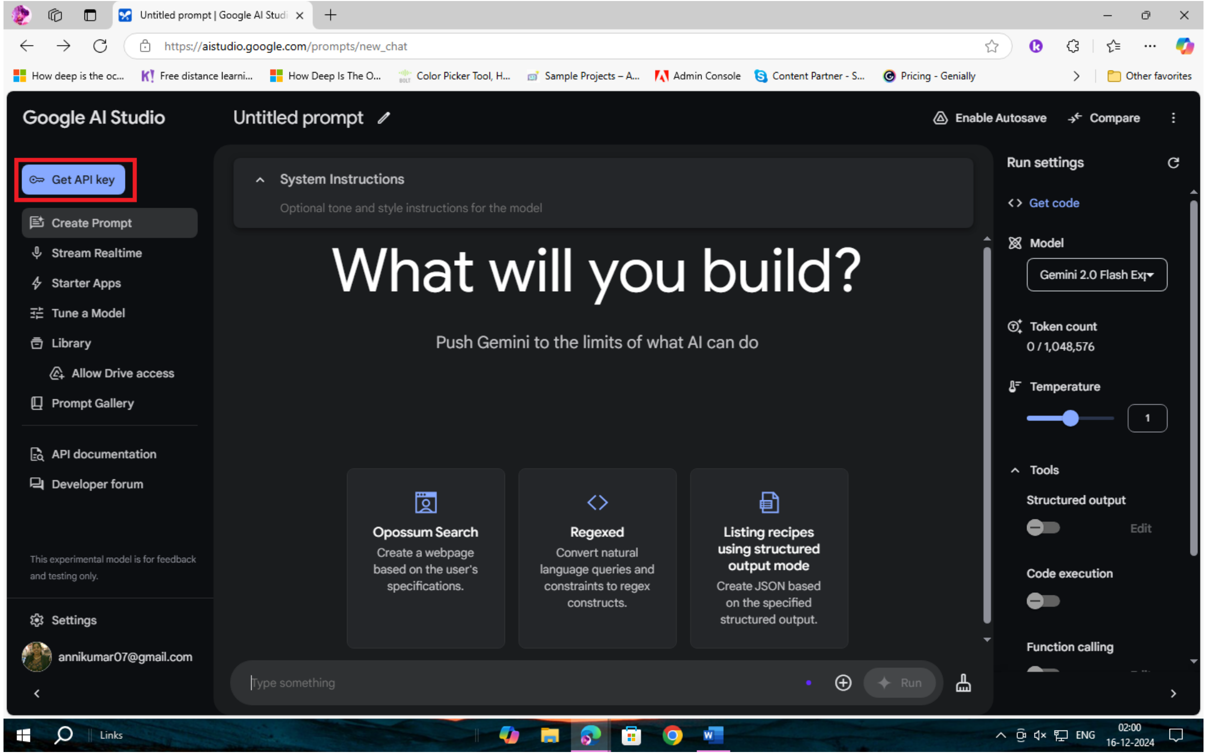
Large e-commerce companies like Amazon have deployed AI-powered chatbots to handle the massive influx of customer queries. These chatbots can instantly answer common questions like order status, return policies, and product information. For more complex issues, they can escalate the query to a human agent, providing initial support and gathering relevant information. This not only improves customer satisfaction by offering quick responses but also reduces the burden on human agents, allowing them to focus on more critical tasks.

Beyond customer service, customized chatbots find applications across various sectors. In education, they can provide personalized tutoring, answer questions, and handle administrative tasks. Researchers can leverage chatbots to analyse data, summarize research papers, and identify relevant articles. In healthcare, chatbots can offer initial medical advice, schedule appointments, and provide mental health support. For the general public, they can be used for language learning, financial advice, and cultural insights. By tailoring chatbots to specific needs, individuals and organizations can unlock their full potential and enhance productivity, efficiency, and user experience.

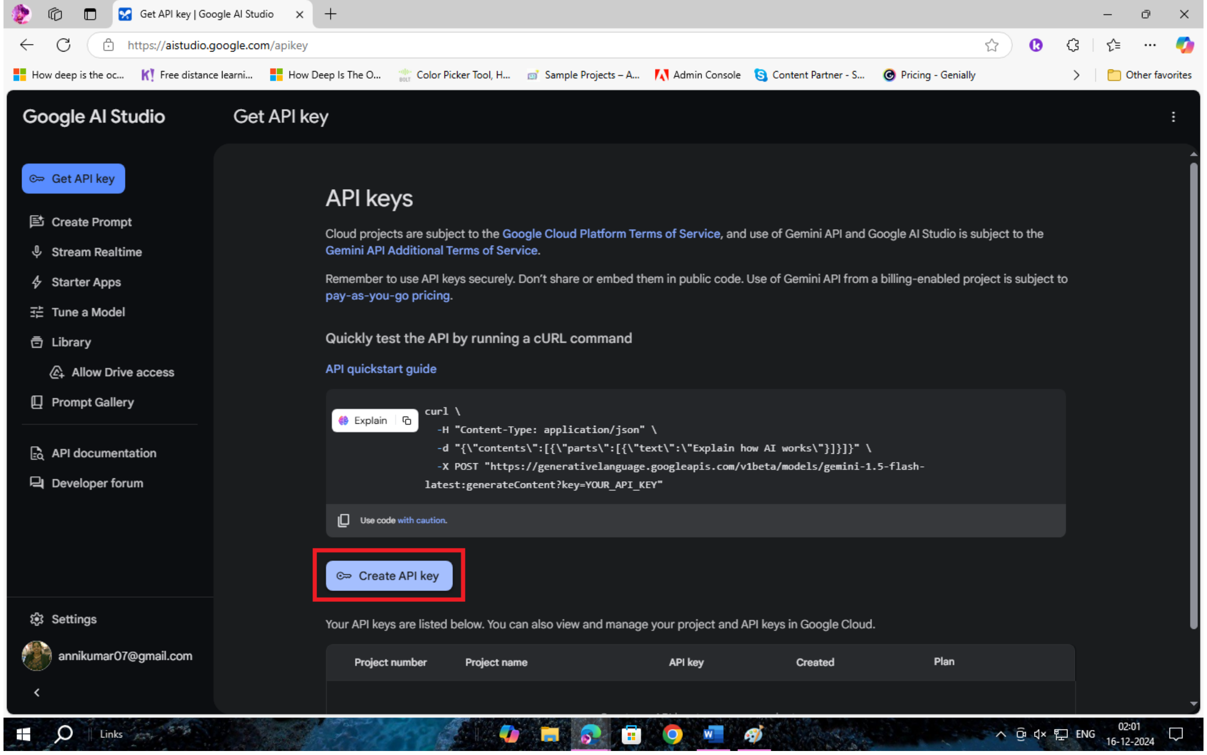
Let us learn to create a customized Chatbot with the help of Gemini API.
We can follow the following steps to create a chatbot:

Step 1. Obtaining an API Key:

- Visit the Google AI Studio platform: https://aistudio.google.com/ and click on the Get API key button



- Click on the Create API key button



API keys

Cloud projects are subject to the Google Cloud Platform Terms of Service, and use of Gemini API and Google AI Studio is subject to the Gemini API Additional Terms of Service.

Remember to use API keys securely. Don't share or embed them in public code. Use of Gemini API from a billing-enabled project is subject to pay-as-you-go pricing.

Quickly test the API by running a cURL command

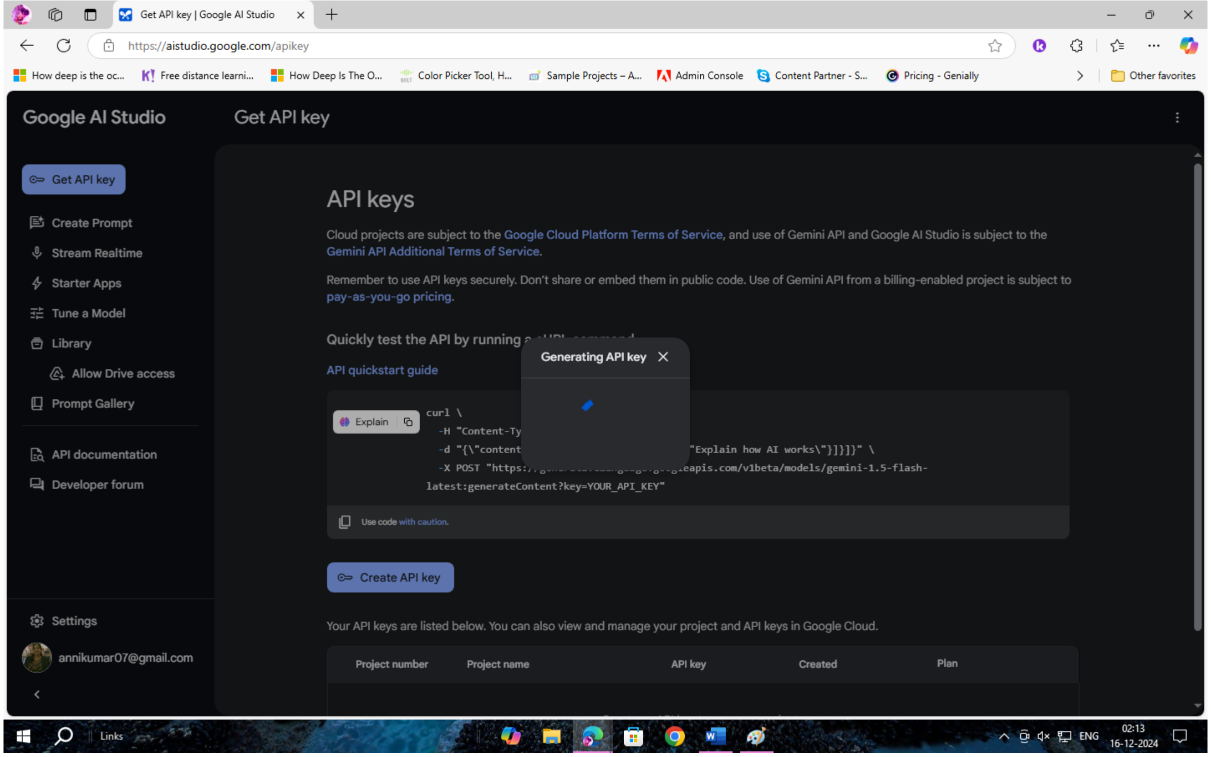
API quickstart guide

```
curl \
  -H "Content-Type: application/json" \
  -d "{\"contents\":[{\"parts\":[{\"text\":\"Explain how AI works\"}]}]}" \
  -X POST "https://generativelanguage.googleapis.com/v1beta/models/gemini-1.5-flash-
latest:generateContent?key=YOUR_API_KEY"
```

Use code with caution.

Your API keys are listed below. You can also view and manage your project and API keys in Google Cloud.

Project number	Project name	API key	Created	Plan



API keys

Cloud projects are subject to the Google Cloud Platform Terms of Service, and use of Gemini API and Google AI Studio is subject to the Gemini API Additional Terms of Service.

Remember to use API keys securely. Don't share or embed them in public code. Use of Gemini API from a billing-enabled project is subject to pay-as-you-go pricing.

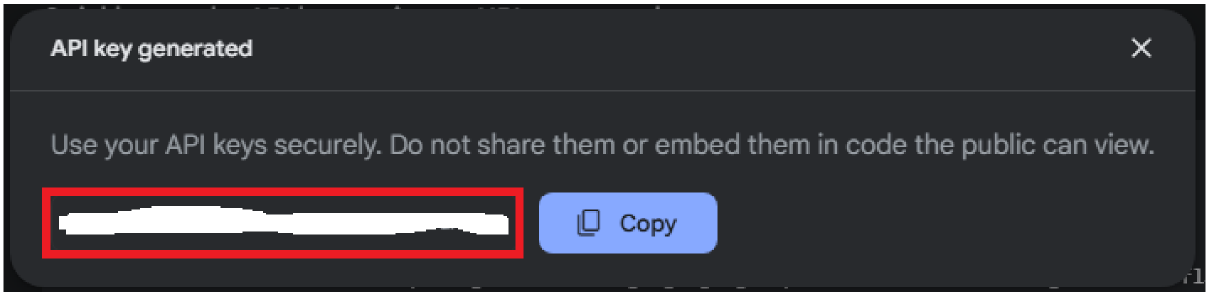
Quickly test the API by running a cURL command

API quickstart guide

Use code with caution.

Your API keys are listed below. You can also view and manage your project and API keys in Google Cloud.

Project number	Project name	API key	Created	Plan


API key generated

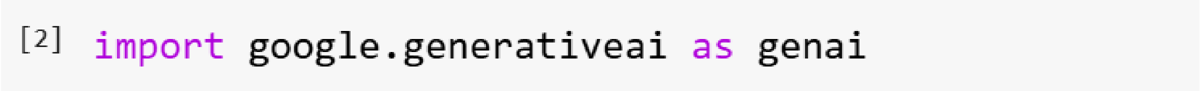
Use your API keys securely. Do not share them or embed them in code the public can view.

[redacted]

Click on Copy and keep it safe.

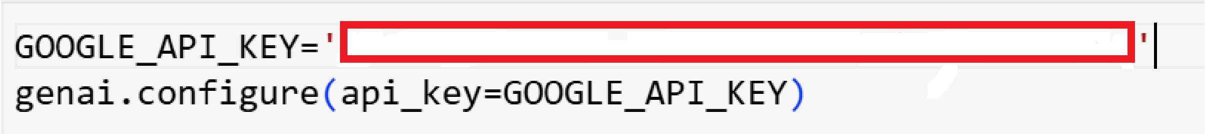
***Step 2.* Setting Up the Python Environment:**
Install the Python SDK for the Gemini API (contained in the google-generativeai package)
pip install -q -U google-generativeai

***Step 3.* Import libraries like generative-ai to interact with the API.**



```
[2] import google.generativeai as genai
```

***Step 4.* Initialize the Gemini API with the API key.**
Pass the API_KEY as value of the variable GOOGLE_API_KEY

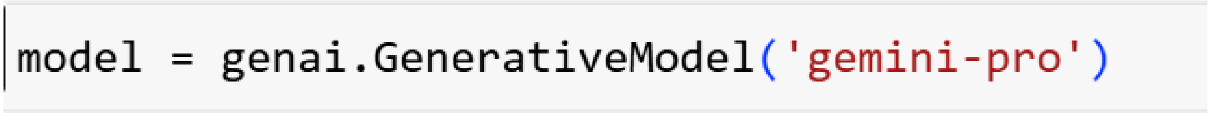


```
GOOGLE_API_KEY='[redacted]'
genai.configure(api_key=GOOGLE_API_KEY)
```

Step 5. Interact with the Gemini model to generate responses.
Gemini offers various models, such as gemini-pro or gemini-pro-vision.

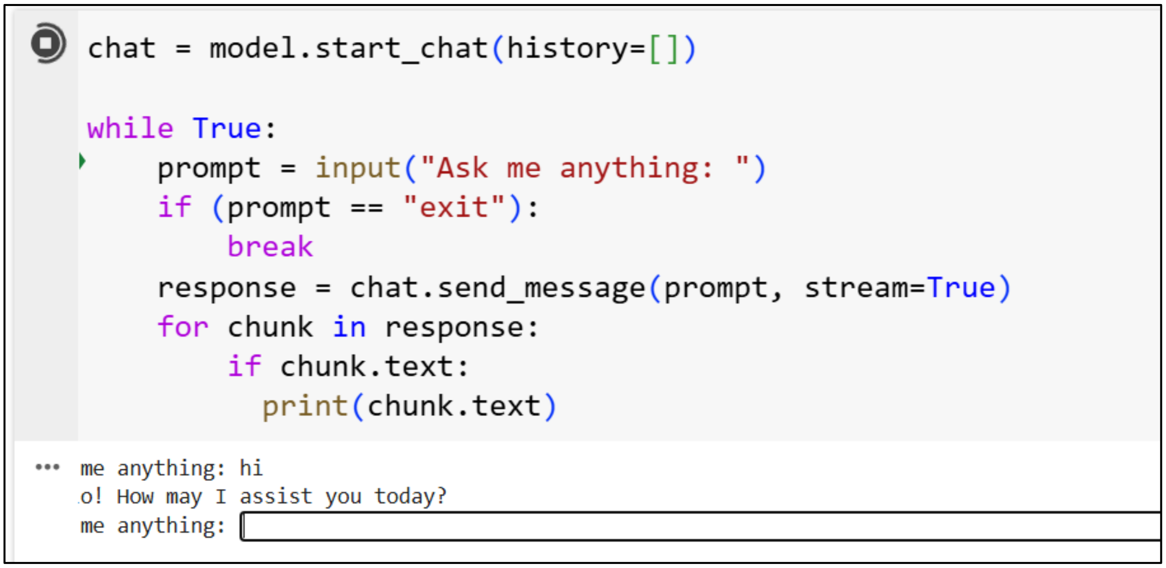
- gemini-pro: optimized for text-only prompts.
- gemini-pro-vision: optimized for text-and-images prompts.

In this example, we will choose gemini-pro



```
model = genai.GenerativeModel('gemini-pro')
```

***Step 6.* Create a chat session**
Now let’s start a chat and see the response:



```
chat = model.start_chat(history=[])

while True:
    prompt = input("Ask me anything: ")
    if (prompt == "exit"):
        break
    response = chat.send_message(prompt, stream=True)
    for chunk in response:
        if chunk.text:
          print(chunk.text)
```

```
me anything: hi
o! How may I assist you today?
me anything:
```

EXERCISES

A. Multiple Choice Questions

1. What is the primary objective of generative AI?
   a) To classify data into different categories
   b) **To generate new data resembling its training samples**
   c) To learn from labelled data for decision-making
   d) To predict outcomes based on input features

2. Which machine learning algorithms are commonly used in generative AI?
   a) Support Vector Machines (SVM)
   b) Decision Trees
   c) **Generative Adversarial Networks (GANs) and Variational Autoencoders (VAEs)**
   d) Random Forests

3. What is the purpose of generative AI in text generation?
   a) To create coherent and contextually relevant text based on given prompts
   b) **To identify patterns in text data for classification tasks**
   c) To analyse sentiment in social media posts
   d) To generate code from natural language inputs

4. Which generative AI model is known for its application in text generation?
   a) Meta AI's Voicebox
   b) Google's Lumiere
   c) **OpenAI's ChatGPT**
   d) Stability AI

5. In image generation, what analogy describes how generative AI models work?
   a) Computers analyse sounds to produce new images.
   b) **Computers make new pictures based on existing ones, resembling patterns they have learned.**
   c) Computers generate images by understanding class boundaries in data.
   d) Computers utilise labelled data to predict outcomes.

6. Which of the following is NOT an application of generative AI?
   a) Creating original music compositions
   b) Crafting promotional videos for marketing campaigns
   c) **Analysing sentiment in social media posts**
   d) Identifying features in image datasets

7. What distinguishes generative models from discriminative models in terms of their training focus?
   a) **Generative models learn to generate new data, while discriminative models focus on identifying class boundaries.**
   b) Generative models focus on learning to draw lines to separate data, while discriminative models analyse the sentiment in text data.
   c) Generative models learn to understand the underlying data distribution, while discriminative models analyse patterns in data.
   d) Generative models aim to identify different classes in data, while discriminative models focus on generating new data samples.

8. Which technology raises concerns about the authenticity of digital content and the spread of misinformation?
   a) **Deepfake AI**
   b) Generative Adversarial Networks (GANs)
   c) Variational Autoencoders (VAEs)
   d) Convolutional Neural Networks (CNNs)

B. True/False

1. Generative AI can only generate text and images, not audio or video. **(False)**
2. Generative Adversarial Networks (GANs) involve competition between the Generator and the Discriminator networks. **(True)**
3. Discriminative models generate new data samples similar to the training data. **(False)**
4. Variational Autoencoders (VAEs) are AI generative models that can be used for tasks such as image generation. **(True)**

5. Deepfake technology is an application of generative AI that can create realistic fake videos. **(False)**

6. Large Language Models (LLMs) like GPT and BERT cannot generate human-like text. **(False)**

C. Fill in the Blanks

1. Generative AI utilises **training** learning algorithms to learn from existing datasets.
2. Examples of Generative AI include ChatGPT, DALL-E, Gemini, and **Claude**
3. The full form of GAN is **Generative Adversarial Networks.**
4. **Image** Generation is an application of Generative AI where computers create new images that resemble the ones they have seen before.
5. OpenAI's ChatGPT and Google's Bard (Gemini) are examples of Generative AI used for **Text** Generation.
6. **Large Language Models (LLMs),** like GPT models developed by OpenAI, are trained on vast amounts of text data to understand and generate human-like text.

D. Short Answer Type Questions

1. What is Generative AI?

Ans- Generative AI is a facet of artificial intelligence that creates new content resembling its training samples, including text, images, audio, and video.

2. How does Generative AI work?

Ans- It learns patterns from existing datasets using machine learning algorithms and intense learning to generate similar new samples autonomously.

3. What distinguishes generative models from discriminative models?

Ans- Generative models aim to understand and replicate the underlying data distribution to generate new samples, while discriminative models focus on distinguishing between different data classes.

4. Give examples of Generative AI applications.

Ans - Examples include ChatGPT for text generation, DALL-E for image generation, and Google's Music LM for audio generation

5. What are Generative Adversarial Networks (GANs)?

Ans- GANs are a class of AI algorithms used in Generative AI for tasks like image generation, where two networks, a generator and a discriminator, work against each other to improve the quality of generated images.

6. What is a Large Language Model (LLM)?

Ans- LLMs are sophisticated AI models trained on vast text data to understand and generate human-like text, excelling in natural language processing tasks.

7. How is Generative AI used in image generation?

Ans - Generative models learn to reproduce the data distribution, while discriminative models learn to separate classes within the data based on features.

8. What ethical considerations surround Generative AI?

Ans - Analysing patterns from input images and generating new images with similar features, like creating new cat images from observed ones.

9. What are the limitations and risks involved with Large Language Models?

Ans- Limitations include high computational costs, the potential for generating incorrect information, and data privacy concerns.

E. Case Study Analysis

Scenario:

A marketing agency, "Creative Horizons", leverage Generative AI technologies to enhance its campaign strategies. The agency uses various AI models for creating unique advertising content, including AI-generated images, personalised text for email campaigns, dynamic video ads, and innovative audio jingles. One of their key projects involves launching a new line of eco-friendly products for a client. The campaign's success hinges on the uniqueness and engagement of the generated content, aiming to highlight the product's sustainability features innovatively.

1. What are the primary types of AI models used by "Creative Horizons" for their campaign?

Ans- Generative AI models, including image and text generation models, video generation algorithms, and audio synthesis tools.

2. How does Generative AI contribute to creating personalised email campaign content?

Ans- Generative AI analyses customer data and previous engagement metrics to produce personalised text that resonates with each segment of the agency's target audience, improving engagement and response rates.

3. Identify one potential ethical consideration the agency must address when using Generative AI in advertising.

Ans- The agency must ensure that the AI-generated content does not inadvertently propagate biases or stereotypes, maintaining ethical standards in representation and messaging.

4.What is a significant advantage of using Generative AI for dynamic video ad creation?
Ans- Generative AI can rapidly produce diverse and innovative video content tailored to different platforms and audience preferences, significantly reducing production time and costs while enhancing creativity.

5.How can "Creative Horizons" ensure their AI-generated content's originality and copyright compliance?
Ans- The agency should implement checks to verify the uniqueness of the content, use AI models trained on copyright-compliant datasets, and provide proper attribution or licenses for AI-generated outputs, ensuring compliance with copyright laws and ethical guidelines.

F. Ethical Dilemma

Read the following ethical dilemma and provide your response:

Ethical Dilemma Scenario:

An AI development company, "InnovateAI," has created a new Generative AI model that can produce original music tracks by learning from a vast database of songs across various genres. The AI's capability to generate music that rivals compositions by human artists has attracted significant attention in the music industry. However, "InnovateAI" faces an ethical dilemma: the AI model inadvertently replicates distinctive styles and melodies of existing copyrighted works, raising concerns about copyright infringement and the originality of AI-generated music. Furthermore, the company discovered that some AI-generated tracks contain elements remarkably similar to unreleased songs by living artists, likely due to including these tracks in the training data without consent.

Discussion Question:

Should "InnovateAI" release this music-generating AI to the public, considering the potential for copyright infringement and ethical concerns regarding the originality of AI-generated content? What measures should be taken to address these ethical and legal issues while advancing technological innovation in the music industry?

Response:
"InnovateAI" faces a complex situation that balances the fine line between innovation and ethical responsibility. Before releasing the AI model to the public, the company should take several steps to mitigate potential legal and ethical issues:

1. Transparency: "InnovateAI" should be transparent about the AI's capabilities and limitations, including the potential for generating content that may resemble existing copyrighted works. This transparency should extend to how the model was trained, including the sources of its training data.
2. Consent and Copyright Compliance: The company must ensure that all data used to train the AI model is either in the public domain, copyrighted with permission, or used under

fair use provisions. This may involve auditing the training dataset to remove any copyrighted material included without consent.

3. Creative Attribution: "InnovateAI" should consider implementing a system for attributing the creative influences of AI-generated music. This could involve tagging AI-generated tracks with metadata referencing the styles or artists that influenced the AI's composition, indirectly acknowledging human artists' contributions.
4. User Guidelines: Providing clear guidelines regarding the ethical use of AI-generated music, including advising against passing off AI-generated compositions as entirely original works without acknowledging the AI's role.
5. Technological Solutions: Developing and integrating technology that can detect and flag potential copyright issues in AI-generated compositions before making them public. This tool could help identify elements too closely resembling existing copyrighted works, allowing for revision or alteration.
6. Engaging with Stakeholders: "InnovateAI" should engage in dialogue with copyright holders, artists, and legal experts to explore collaborative solutions that respect copyright while fostering innovation. This could include licensing agreements or partnerships with music publishers.

By taking these measures, "InnovateAI" can address ethical concerns head-on while contributing positively to the music industry. It acknowledges the importance of balancing innovation with respect for existing creative works and copyright laws, ensuring that advancements in AI benefit all stakeholders in the music ecosystem.

Note to Teachers: Teachers to discuss any other ethical concern.

G. Competency Based Questions:

1. Anita works for a social media company that is exploring the use of Generative Adversarial Networks (GANs) to generate personalized content for its users. However, there are concerns about the ethical implications of using AI to manipulate users' perceptions and behaviours. What are some potential ethical concerns associated with using GANs to generate personalized content on social media platforms?

Ans- Some potential ethical concerns include:

- Manipulation of user perceptions potentially leading to misinformation or exploitation.
- Invasion of privacy
- Amplification of biases leading to unfair treatment or discrimination.

2.A research team is using Generative AI to generate synthetic data for training medical imaging algorithms to detect rare diseases. However, there are concerns about the potential biases and inaccuracies in the generated data. What steps can the research team take to mitigate biases and ensure the accuracy and reliability of the synthetic data generated by Generative AI for medical imaging applications?

Ans- The research team can take the following steps:

- Ensure that the training data used to train the Generative AI model represent a diverse range of demographics, including underrepresented groups, to mitigate biases.
- Validate the synthetic data generated by the model against real medical imaging data to ensure accuracy and reliability.
- Involve domain experts, such as medical professionals and imaging specialists
- Continuously refine and improve the Generative AI model based on feedback from domain experts and the performance of the medical imaging algorithms trained on the synthetic data.

3.A music streaming platform is considering using Generative AI to create personalized playlists for users based on their listening habits. They want to ensure that the generated playlists accurately reflect users' preferences and introduce them to new music they might enjoy. How can the music streaming platform ensure that the playlists generated by Generative AI effectively cater to users' preferences while also introducing them to new music?

Ans- The music streaming platform can:

- Analyze user listening habits to inform playlist generation.
- Introduce randomness to expose users to new music aligned with preferences.

4.An interior design company wants to use Generative AI to create room renderings for client presentations. They aim to ensure that the generated designs are both attractive and practical. How can the interior design company ensure that the room renderings generated by Generative AI are both visually appealing and functional?

Ans- The interior design company can:

- Curate diverse room design data for training.
- Incorporate design principles into the Generative AI model.
- Seek feedback from experienced interior designers.

5.A marketing agency needs visually appealing social media posts for an advertising campaign. How can the marketing agency use Generative AI to create diverse and engaging visuals for their campaign?

Ans- The marketing agency can:

- Collect relevant images for training the Generative AI model.
- Generate visuals aligned with the campaign's branding.
- Review and refine the generated visuals for maximum impact.

References/Links of images used in the lesson

- https://www.hindustantimes.com/ht-img/img/2023/04/25/550x309/Anand_mahindra_shares_video_of_a_girl_aging_1682405476306_1682405479636.png

- https://fdczvxmwwjwpwbeeqcth.supabase.co/storage/v1/object/public/images/d6546a37-4ba1-4a5d-87b1-a698c3960c20/fa7f7d26-0ba6-4275-b469-f5a005828775.png

- https://uploads-ssl.webflow.com/61dfc899a471632619dca9dd/62f2dd5573c01a4523b4ace6_Deepfake-Optional-Body-of-Article.jpeg

- https://arxiv.org/html/2405.11029v1#S6

UNIT 8: Data Storytelling

Title: Data Storytelling

Approach: Team discussion, Web search, Case studies

Summary: Students will learn about the importance of storytelling, which has been used for ages to share knowledge, experiences, and information. They will also understand how to connect storytelling with data storytelling, a key part of Data Analysis. This lesson will teach them to combine the three elements of data storytelling—data, visuals, and narrative—to present complex information engagingly and effectively. This helps the audience make informed decisions at the right time.

Learning Objectives:

1. Students will understand the benefits and importance of powerful storytelling.
2. Students will appreciate the concept of data storytelling in data analysis, which is a key part of data science and AI.
3. Students will learn how to combine the elements of data storytelling—data, visuals, and narrative—to present complex information.
4. Students will learn how to draw insights from a data story.

Key Concepts:

1. Introduction to Storytelling
2. Elements of a Story
3. Introduction to Data Storytelling
4. Why is Data Storytelling Powerful?
5. Essential Elements of Data Storytelling
6. Narrative Structure of a Data Story (Freytag's Pyramid)
7. Types of Data and Visualizations for Different Data
8. Steps to Create a Story Through Data
9. Ethics in Data Storytelling

Learning Outcomes:

Students will be able to -

1. Identify the difference between storytelling and data storytelling.
2. Understand the key elements of data storytelling.
3. Recognize the importance of data storytelling today.
4. Use the appropriate type of visual for the data.
5. Draw insights from data stories and write simple narratives based on the visuals.

Prerequisites: Understanding the concept of data and reasonable fluency in the English language. Ability to understand visual data.

Unveiling the Power of Data Storytelling: A Teacher's Guide

This lesson empowers you to equip students with the art of Data Storytelling – transforming data into captivating narratives that inform, persuade, and inspire.

1. The Magic of Stories: Building Blocks of Narrative:

- **Captivating Introduction:** Begin by engaging students with a powerful story. Discuss the elements that make a story compelling (e.g., characters, plot, setting).
- **Narrative Structure:** Introduce the core elements of narrative structure:
  - Characters: The individuals driving the story.
  - Plot: The sequence of events that unfold.
  - Setting: The time and place where the story take place.
  - Conflict: The central challenge the characters face.
  - Resolution: The outcome of the conflict.
  - Theme: The underlying message or idea conveyed by the story.
- **Hands-on Storytelling Exercise:** Divide students into pairs and have them brainstorm a short story using these elements. Encourage them to share their stories with the class.

2. The Elements of Every Story:

- **Digging Deeper:** Focus on the core elements of a story:
  - Plot: The sequence of events, including exposition, rising action, climax, falling action, and resolution.
  - Characters: Their traits, motivations, and how they develop throughout the story.
  - Setting: The physical and social environment where the story takes place.
  - Conflict: The central problem or challenge that drives the plot forward.
  - Resolution: How the conflict is addressed or resolved.
  - Theme: The underlying message or idea conveyed by the story (e.g., courage, perseverance).
- **Interactive Activities:** Use games and activities to solidify understanding:
  - Match character traits to character descriptions.
  - Sequence events from a story to understand the plot structure.
  - Brainstorm different resolutions for a given conflict.

3. Weaving Data into Narratives: The Power of Data Storytelling:

- **A New Storytelling Frontier:** Introduce Data Storytelling – the art of combining data insights with narrative elements to create compelling stories.
- **Why Data Storytelling Matters:** Discuss the power of Data Storytelling:
  - Makes complex information accessible and understandable.

    - Persuades audiences to take action based on evidence.
    - Creates a deeper emotional connection with the data.

4. The Essential Ingredients of a Data Story:

- **Recipe for Success:** Outline the key ingredients of a Data Story:
    - Data-Driven Insights: The core findings and takeaways from the data analysis.
    - Coherent Narrative Structure: A clear narrative arc with a beginning, middle, and end.
    - Effective Visualizations: Charts, graphs, and maps that complement the narrative and make data understandable.
    - Audience-Focused Communication: Tailoring the story to the interests and needs of the audience.
- **Real-World Examples:** Showcase compelling Data Stories from various fields (e.g., healthcare, business, social impact) to illustrate these elements in action.

5. Data Types and Visualizations: Telling the Visual Story:

- **Understanding Data Types:** Introduce different data types and their appropriate visualizations:
    - Categorical Data (e.g., customer categories): Bar charts, pie charts.
    - Numerical Data (e.g., sales figures): Line charts, scatter plots, histograms.
    - Temporal Data (e.g., time series): Line charts, timelines.
- **Interactive Activity:** Provide students with sample data sets and have them choose the most appropriate visualization for each. Discuss the rationale behind their choices.

6. From Data to Impact: Crafting Your Data Story:

- **The Storytelling Process:** Guide students through the process of creating a Data Story:
    - Analyze Data: Identify key insights and trends.
    - Craft a Narrative: Develop a clear narrative structure.
    - Select Visualizations: Choose visualizations that enhance the story.
    - Communicate Effectively: Tailor the story to the audience.
- **Hands-on Project:** Divide students into groups and assign them datasets related to their interests (e.g., educational data, environmental data). Guide them through the process of creating their own Data Stories using storytelling tools and presentation software.

Additional Tips:

- Encourage students to find creative ways to present their Data Stories (e.g., infographics, videos).
- Integrate technology tools like data visualization platforms to enhance storytelling.
- Discuss ethical considerations when presenting data (e.g., avoiding bias, data source transparency).

Teachers can ask the following questions to spark curiosity before starting the topics:

- **Do you like listening to stories? Why or why not?** (This taps into students' prior knowledge about stories and their enjoyment of them)
- **Can you think of any examples of stories that you've heard that taught you something?** (This encourages students to make the connection between stories and learning)
- **Do you think facts and figures can be presented in a story format? Why or why not?** (This gets students thinking about the possibility of data being presented in a story)

8.1 Introduction to Storytelling

"Once upon a time, there lived a King...". "On a dark night, when it was raining...". "Long long ago...". These familiar sentences spark interest, enthusiasm, and curiosity in all of us. Aren't they familiar?

Yes, they are usually used when telling stories. Stories have been a part of our lives since ancient times. They have existed from the time of cavemen to the present day. Stories are a way to share our imaginations, experiences, and thoughts with others.

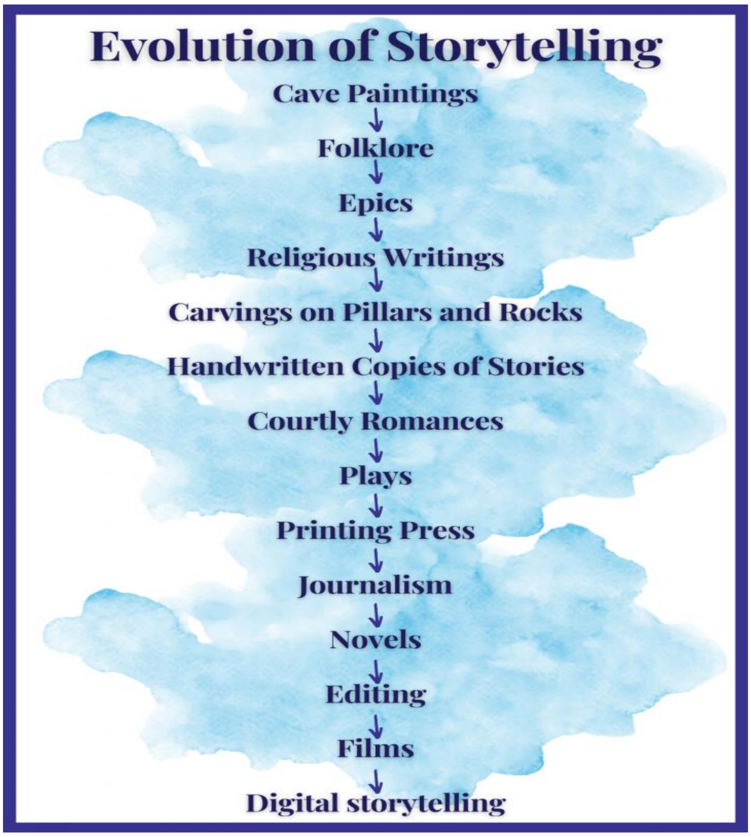


Fig. 8.1 Evolution of Storytelling

source: https://www.trueeditors.com/blog/the-evolution-of-storytelling/

So, what are Stories?

Stories are a valuable form of human expression. They connect us closely with one another and transport us to different places and times. Stories can be of various types, like folk tales, fairy tales, fables, and real-life stories. Each type of story creates a sense of connection, and folk tales, in particular, strengthen our sense of belonging to our community and help establish our identity.

Every story has a theme or topic. There is always a storyteller and a listener, and sometimes the listeners can be a group of people. According to the dictionary, a 'story' is a 'factual or fictional narrative,' meaning it tells about an event that can be true or made up, in a way that the listener experiences or learns something. Stories can be used to share information, experiences, or viewpoints.

What are the benefits of stories?

Stories have a strong impact on human minds as they motivate, ignite, and change our perspectives. **The process of telling a story narratively is known as Storytelling.** It encourages people to make use of their imagination and creativity to express themselves, which improves their communication skills. Storytelling can be in different forms- oral, digital, and written.

Why is storytelling a powerful tool?

Storytelling is a potent tool for several reasons, such as –

- It generates interest, captivates audiences, and draws them to the narrative.
- It captures our attention, keeping us engaged and focused.
- It communicates meaning, making complex ideas more accessible and understandable.
- It inspires us, evokes emotions, and motivates us to take action.

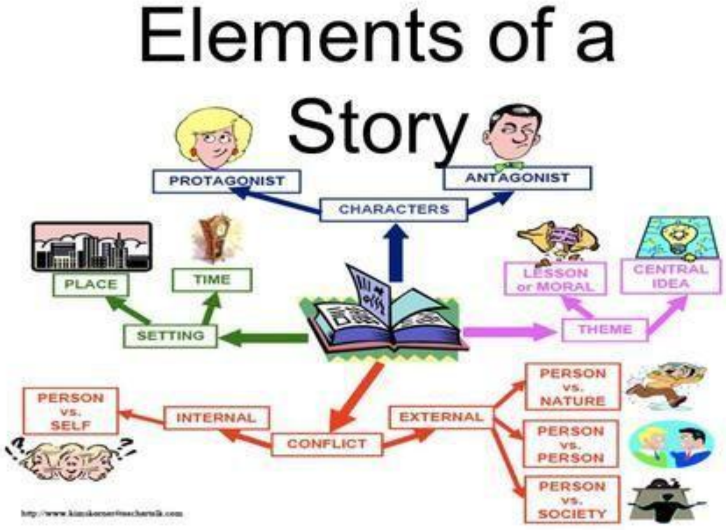


Fig. 8.2 Elements of a story

Source: *https://in.pinterest.com/pin/778419116828184060/*

1. **Characters:** The characters are the people or animals or some things or objects which are featured in a story. They perform the actions and drive the story.
2. **Plot/setting**: Setting refers to the time or location in which the story takes place. Plot refers to the sequence of the events of the story.
3. **Conflict:** It is the problem or the situation the characters are dealing with. It drives the story forward which makes the story engaging and a key element for the characters.
4. **Resolution:** It is the end of the story where the characters arrive at a particular situation to resolve the conflict. It is the stage after climax which is the peak or height of any story.
5. **Insights**: The ability to have a clear, deep, and sometimes sudden understanding of a complicated problem or situation.

Activity 1:

Think of different types of stories, a real story, a mythological story, a fiction story, folk tale, and then complete the table according to the given headings.

Name of the story	**Type of the story**	**Characters**	**Insight Gathered/Moral**
The Diary of Anne Frank	Real Story	Anne Frank, Otto Frank (father), Edith Frank (mother), Margot Frank (sister)	Even in difficult circumstances, hope and the human spirit can endure.
The Lion King	Mythological Story (resembles coming-of-age stories)	Simba (lion cub), Mufasa (father), Scar (uncle), Timon & Pumbaa (meerkat & warthog), Rafiki (mandrill)	Circle of life, overcoming adversity, importance of responsibility.
The Adventures of Pinocchio	Fiction Story	Pinocchio (wooden puppet), Geppetto (carver), Jiminy Cricket (conscience), The Fox & The Cat (tricksters), The Blue Fairy	Honesty, hard work, and good choices lead to a happy life.
The Three Little Pigs	Folk Tale	Three little pigs, Big Bad Wolf	Preparation and hard work are rewarded, while laziness has consequences.

Introduction to Data Storytelling

According to Wikipedia, **data are individual facts, statistics, or items of information, often numeric**. In a more technical sense, data is a set of qualitative or quantitative variables about one or more persons or objects. So, keeping these definitions in mind, when we connect logically related data together, they tell us something. The collection of data when represented in a better way has a greater impact which can be engaging, entertaining, thought provoking, helps in better decision making and can make way for big changes.

So, when we interpret this data in a systematic way, then this concept is known as Data Storytelling. It is a practice that is used lately by analysts and data scientists to communicate their findings and observations from data to technical and non-technical business stakeholders who are in general called audience.

Data storytelling is the art and practice of translating complex data and analytics into a compelling narrative that is easily understandable and relatable to various audiences.

Data can be in the simple form of numbers and digits. This data when it is pictorially represented is known as **Data Visualization**. It can be in the form of different types of charts or graphs. Depending on the requirements, data can be interpreted in the form of narratives known as Data Stories which can reduce ambiguity. It can be clear with respect to context and convey the right meaning which can be used for an effective decision-making process.

Need for Data Storytelling

“Sometimes reality is too complex. Stories give it form.” — *Jean-Luc Godard, film director, screenwriter*

Activity 2:

Read this passage aloud in class:

Rahul eagerly wanted to share the exciting news with his sister, Priya, and his wife, Smita. Priya's friend, Anil, had promised to arrange an interview for Rahul at the software company where he worked. This job was particularly significant to Rahul as it aligned perfectly with his area of expertise and offered a competitive salary. However, just as Rahul was preparing for the interview, he received a call from another company, where Smita had applied. They offered her a position as well, but it required them to relocate to a different city. This sudden twist left Rahul torn between pursuing his dream job and supporting his wife's career aspirations. Unable to reach Priya, Rahul left her a voicemail, brimming with anticipation. A few hours later, when Priya returned his call, she sounded even more thrilled than Rahul. "Guess what," Priya exclaimed, "Anil just helped me secure a position at the software company."

Now, ask the following questions:

Who is Anil's friend? Priya

Who is Rahul's sister? Priya

Who arranged an interview for whom? Anil arranged an interview for Priya

Activity 3:

Show the following content to the students on screen for approximately 2 minutes.

A police officer pulls over a car for driving too slowly on the highway. “This is a highway, you must drive at least **80 km/hr.”** “But the sign says **20**!” said the driver. “This is highway route **20**, not speed limit **20**.” The officer then sees the passenger is unconscious. “Is everything OK?” “She’s been like that since we turned off highway **180**.”	An established finding in traffic safety is that vehicles should maintain speeds consistent with those of neighboring vehicles. For this reason, federal agencies recommend a minimum speed of eighty kilometers per hour to be adopted by local state, and provincial governments. Nearly all local governments have conformed to this recommendation as it promotes safety and minimizes confusion when passing from one jurisdiction to another or when merging from one highway onto another.

Now, ask them “What is the minimum speed limit on a highway?

With these two activities, we understand that data, when presented in a narrative format, is better absorbed, retained, and understood compared to a collection of disjointed facts or figures. Just like the effectiveness of storytelling in memory retention, data storytelling enhances the comprehension and impact of data insights by providing context and structure.

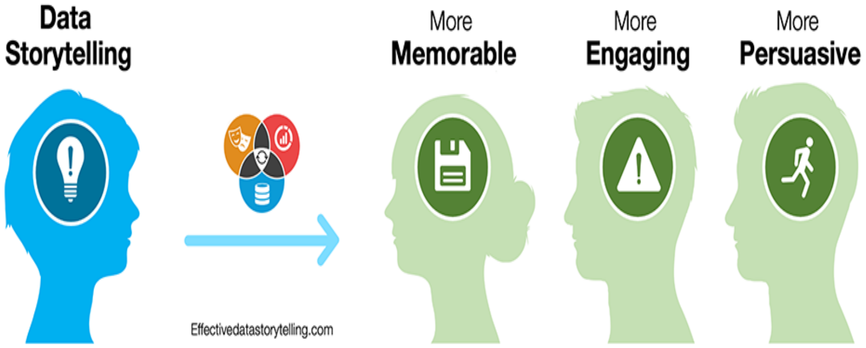
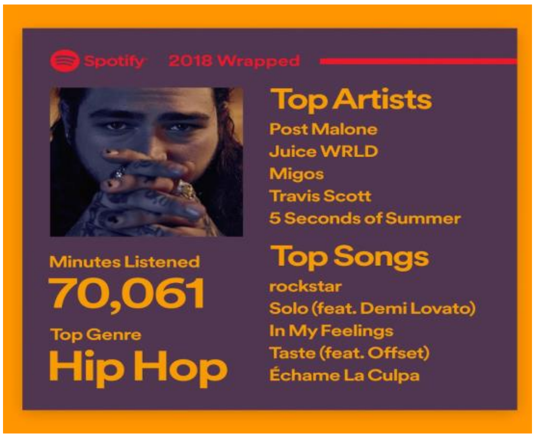


Fig. 8.3 Data Storytelling: Makes information more memorable, engaging, and persuasive

The need for data storytelling is gaining importance in all fields. Many companies and brands are using data storytelling as an effective method of conveying their message and gaining client loyalty. Data storytelling makes complex data more accessible and understandable, allowing audiences or stakeholders to grasp insights easily. Engaging narratives and compelling visuals keep audiences engaged, increasing retention and attention. Storytelling with data empowers better decision-making by presenting evidence-based insights.

Examples of some famous brand Data Stories



Spotify 2018 Wrapped

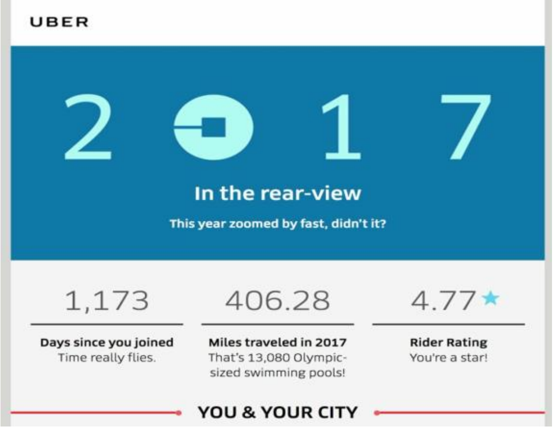
Top Artists
Post Malone
Juice WRLD
Migos
Travis Scott
5 Seconds of Summer

Minutes Listened
70,061

Top Genre
Hip Hop

Top Songs
rockstar
Solo (feat. Demi Lovato)
In My Feelings
Taste (feat. Offset)
Échame La Culpa

Spotify



UBER

2017

In the rear-view

This year zoomed by fast, didn't it?

1,173	406.28	4.77★
Days since you joined Time really flies.	Miles traveled in 2017 That's 13,080 Olympic-sized swimming pools!	Rider Rating You're a star!

YOU & YOUR CITY

Uber

Why has Data Storytelling become very powerful now?

In today's business context, data storytelling becomes so critical due to these characteristics:

1. It makes the insights and key findings memorable to the audience.
2. It is a persuasive way of communicating key insights and findings to both business stakeholders and technical stakeholders.
3. It is also important that the story is engaging to the audience.

Today, data storytelling is filling a vital gap in business as well as the technical analytics process between technology and people. By combining an increasingly essential resource (data) with a familiar and time-tested form of communication (storytelling), this skill can help more people translate their insights into action. From a data literacy perspective, once an organization's people are comfortable with reading and working with data, they should also learn how to communicate insights effectively.

8.2 Essential elements of Data Storytelling

The three key elements of Data Storytelling are:

1. Data 2. Narrative 3. Visuals

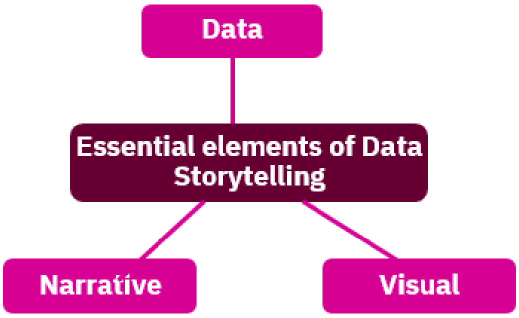


Fig. 8.4 Essential elements of Data Storytelling

Data: Basic facts or raw facts about any entity is known as Data. Data is the primary building block of every data story. It serves as the foundation for the narrative and visual elements of your story.

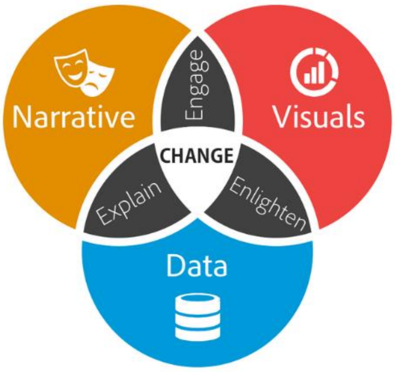
Narrative: It refers to the structure or storyline that is crafted to present insights derived from data in a clear, engaging, and informative manner to the audience. It involves identifying and organizing the key information in a linear and coherent fashion, ensuring that the storyline is crisp, and contextual. A well-defined narrative allows audience to understand the significance of insights and how they relate to the broader context of the data analysis.

Visuals: Visuals refer to the pictorial representations of data using various graphs, charts, and diagrams. They serve as a scene of the data presented in a graphical format, helping to convey complex information more clearly and effectively. They enable the audience to visualize trends, patterns, and anomalies in the data, facilitating better understanding and interpretation of the insights being communicated.

How are these three interlinked with each other?

- When we explain data observations, it helps the audience understand how the data behaves in different situations and why certain insights are generated.
- Visualizing data through charts and graphs allows the audience or stakeholders to see the data from a different perspective, helping them analyse and make the right decisions.
- Combining narrative and visuals can engage or even entertain an audience.

Each element—data, narrative, or visuals—can be powerful on its own. You can achieve some success with a compelling story, an interesting statistic, or impressive data visualization. However, the real impact comes from skillfully blending data, narrative, and visuals in a data story. When you combine the right visuals and narrative with the right data, you create a data story that can influence and drive change.



Narrative structure of a data story

Most stories follow a common arc – a protagonist who faces a complication goes on a journey of resolving a difficulty before returning to their normal lives. Building on Aristotle's simple model, Freytag developed a more robust narrative framework to better understand the arc or progression of a story. The "pyramid-based" dramatic structure with five key stages:

1. **Introduction**: The beginning of the story when the setting is established, and main characters are introduced. It provides the audience with ample background information to understand what is going to happen.
2. **Rising action**: The series of events that build up to the climax of the story.
3. **Climax**: The most intense or important point within the story. It is often an event in which the fortune of the protagonist turns for the better or worse in the story.
4. **Falling action**: The rest of the events that unravel after the main conflict has occurred, but before the final outcome is decided.
5. **Conclusion**: The conclusion of the story where all of the conflicts are resolved and outstanding details are explained.

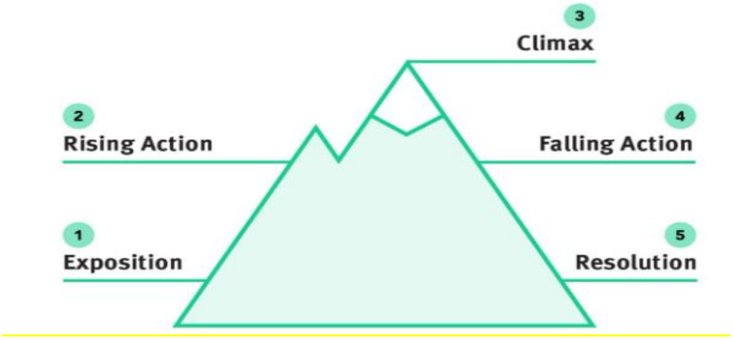


Fig.8.5 Freytag's Pyramid

Image Source: https://www.linkedin.com/pulse/freytags-pyramid-data-storytelling-unlocking-power-ahsan-khurram/

In the context of data storytelling, Freytag's Pyramid can be used as a framework to structure the presentation of data and insights in a way that captivates the audience's attention and guides them through the narrative journey.

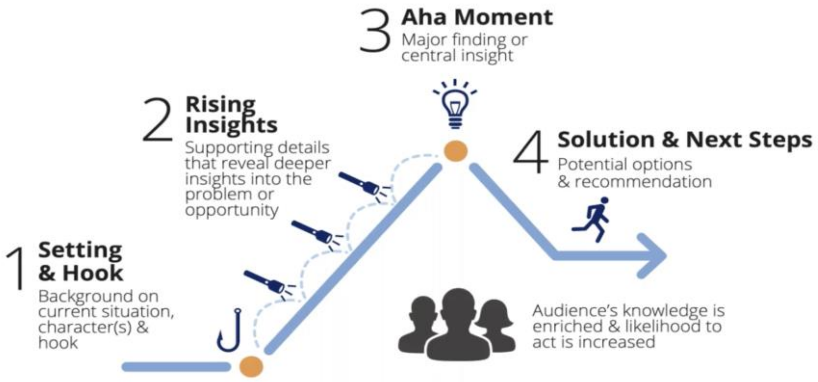
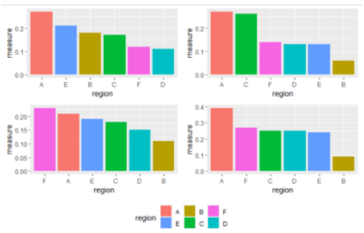
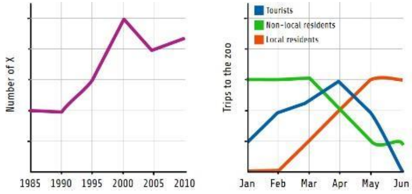
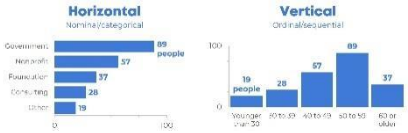
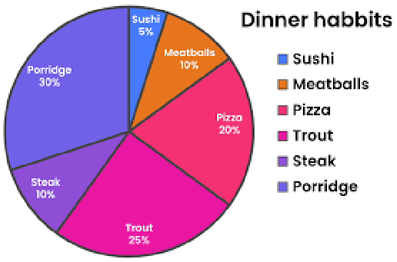
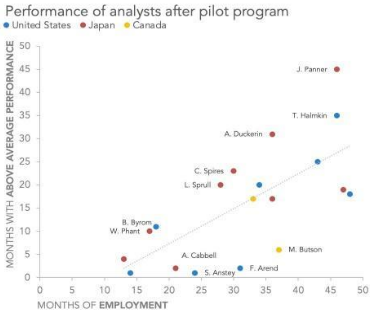
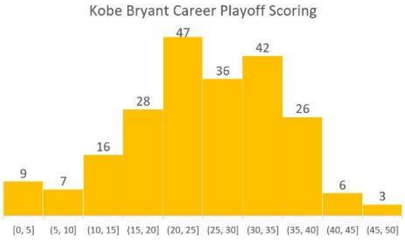
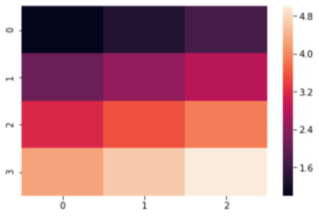
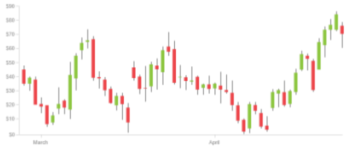

3 Aha Moment
Major finding or central insight
2 Rising Insights
Supporting details that reveal deeper insights into the problem or opportunity
1 Setting & Hook
Background on current situation, character(s) & hook
4 Solution & Next Steps
Potential options & recommendation
Audience's knowledge is enriched & likelihood to act is increased

Fig. 8.6 The Data Storytelling Arc: Crafting Insightful Narratives

Visualizations for different data

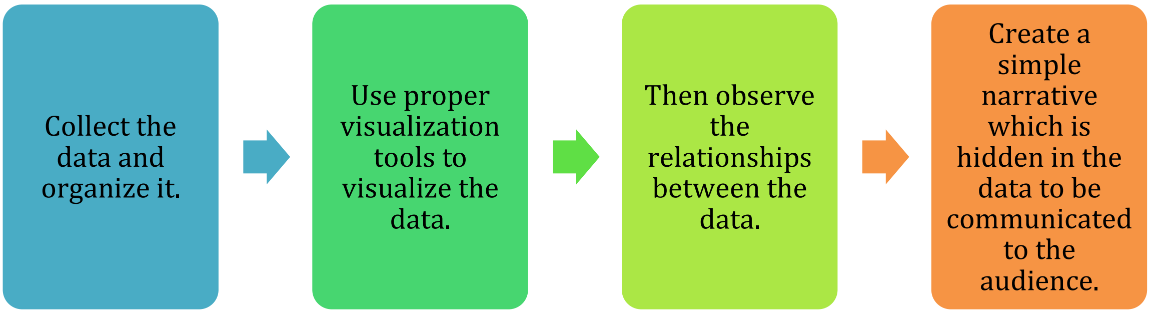
Data visualization is a powerful way to show context. Data charts can reveal crucial deviations or affinities in the data that can lead to insights.

Type	Visualization Type	Description
Text Data	Word Cloud knowledge education learning development wisdom school student value proficiency comprehension motivation	Visual representation of word data where word size indicates frequency and importance.
Mixed Data	FacetGrid 	Multi-axes grid with subplots to visualize distribution and relationships between variables.
Numeric Data	Line Graph 	Illustrates data changing over time, useful for trends.
	Bar Chart 	Compares data between categories using bars.
	Pie Chart 	Circular chart illustrating numerical proportions.

	Scatter Plot  	Visualizes relationships and trends between two variables.
	Histogram  	Represents distribution of continuous data through bars.
	Heat Map  	Compares data across categories using color to identify strong and weak categories.
Stocks	Candlestick Chart  	Visual aid for decision-making in stock, forex, commodity, and option trading.
Geographic	Map Chart Countries by Category A B C D E F Map chart by Category Powered by Bing © GeoNames, Microsoft, Navteq, Wikipedia	Utilizes geographical maps to display data points or statistical information associated with specific locations.

Steps to create a story through data

If the data collected is represented in just a series of graphs and charts, it will not serve the purpose to any organization. It should be communicated well with proper narrative, with proper context and meaning, relevance and clarity. The narrative should be able to take the focus of the audience to the correct spot and not miss out on important facts. To find compelling stories in data sets, the following steps are to be followed:



Collect the data and organize it.

Use proper visualization tools to visualize the data.

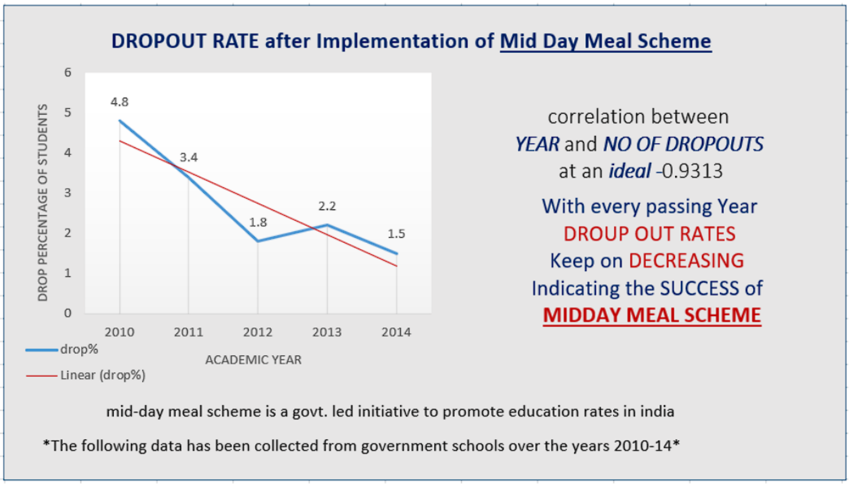
Then observe the relationships between the data.

Create a simple narrative which is hidden in the data to be communicated to the audience.

Example 1:

Using available data on student enrollment, attendance, and dropout rates, create a compelling data story that explores the impact of the Mid-Day Meal Scheme (MDMS) since its launch in 1995. Uncover trends, patterns, and correlations in the data to tell a story about how the implementation of the MDMS may have influenced dropout rates in the state over the years. Consider incorporating visualizations, charts, and graphs to effectively communicate your findings. Additionally, analyze any external factors or events that might have played a role in shaping these trends. Your goal is to provide a comprehensive narrative that highlights the relationship between the MDMS and student dropout rates in the state.

Year	Enrolled	No_of_Dropout	Dropout_percentage
2010	3720068	178490	4.8
2011	3753087	128185	3.4
2012	3728540	68860	1.8
2013	3531426	76204	2.2
2014	3403395	50373	1.5



DROPOUT RATE after Implementation of Mid Day Meal Scheme

correlation between
YEAR and *NO OF DROPOUTS*
at an ***ideal*** -0.9313

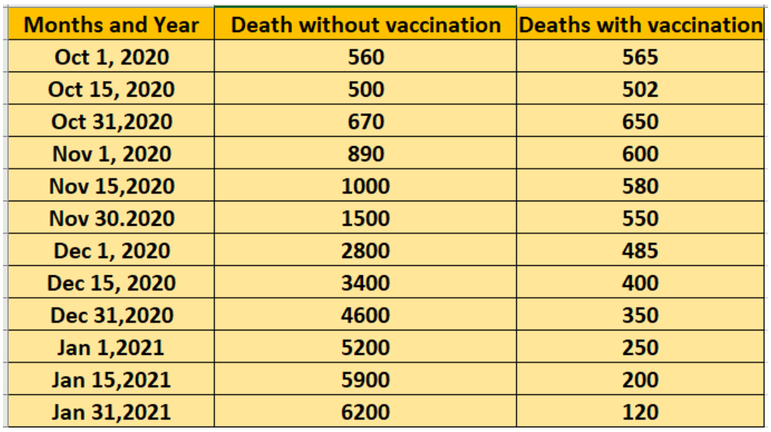
With every passing Year
DROUP OUT RATES
Keep on DECREASING
Indicating the SUCCESS of
MIDDAY MEAL SCHEME

mid-day meal scheme is a govt. led initiative to promote education rates in india

The following data has been collected from government schools over the years 2010-14

Example 2:

Let us do an activity now to create a data story with the information given below. We have collected the data. Use the above steps to create an effective Data Story.

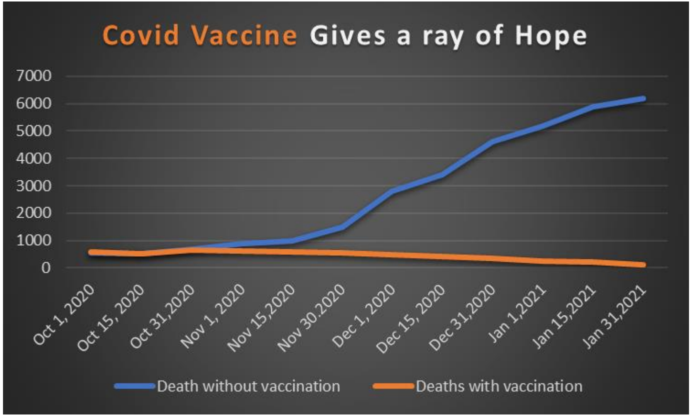


Months and Year	Death without vaccination	Deaths with vaccination
Oct 1, 2020	560	565
Oct 15, 2020	500	502
Oct 31,2020	670	650
Nov 1, 2020	890	600
Nov 15,2020	1000	580
Nov 30.2020	1500	550
Dec 1, 2020	2800	485
Dec 15, 2020	3400	400
Dec 31,2020	4600	350
Jan 1,2021	5200	250
Jan 15,2021	5900	200
Jan 31,2021	6200	120

Solution:

Step 1: Prepare this data sheet in MS-Excel.

Step 2: Visualize the data using Line chart as follows in Ms-Excel.



Step 3: Narrative:

Covid Vaccine- Gives a ray of Hope

Ethics in Data Storytelling

Data storytelling is a powerful tool, but it also requires careful consideration of ethics. Each of the three key elements of data storytelling - Data, Narrative, and Visuals - presents unique ethical challenges that must be addressed to ensure responsible and trustworthy storytelling.

1. Accuracy: Ensure that the data is accurate, reliable, and truthful. Avoid manipulating data to support a predetermined narrative.
2. Transparency: Clearly cite the sources of the data, methods used for analysis, and any limitations or biases. Be transparent about the story's purpose and potential conflicts of interest.

3. Respect for Privacy: Protect the privacy of individuals and groups represented in the data. Avoid sharing personal or sensitive information without consent.

Conclusion

A data story does not just happen on its own—it must be curated and prepared by someone for the benefit of other people. When we effectively combine the right insights with the right narratives and visuals, we can communicate the data in a manner that can inspire change. The data stories can help other people to understand a problem, risk, or opportunity in a meaningful way that compels them to act on it. So, we can define Data storytelling as a persuasive, structured approach for communicating insights using narrative elements and explanatory visuals to inform decisions and drive changes.

Exercises

A. Multiple choice questions

1. Which of the following best describes data storytelling?
   a) Presenting raw data without any analysis
   b) Communicating insights through data in a compelling narrative format
   c) Creating colorful charts and graphs
   d) Analyzing data without any visual aids
2. What is the primary goal of data storytelling?
   a) To confuse the audience with complex charts
   b) To entertain the audience with anecdotes
   c) To communicate insights and findings effectively using data
   d) To hide information from the audience
3. Which of the following is NOT a component of effective data storytelling?
   a) Compelling visuals
   b) Clear narrative structure
   c) Overcomplicating the message
   d) Insightful analysis
4. What role do visuals play in data storytelling?
   a) They make the presentation look fancy
   b) They distract the audience from the data
   c) They help convey complex information quickly and effectively
   d) They are not necessary in data storytelling

5. Why is it important to know your audience when creating a data story?
   a) To impress them with your knowledge
   b) To tailor the message and visuals to their level of understanding and interests
   c) To exclude certain data points that might confuse them
   d) To ignore their feedback and preferences

B. True or False

1. The main purpose of a data visualization is to hide the data from the audience - False
2. Sonnet charts are not a common type of data visualization. - True
3. Data storytelling involves presenting insights and findings in a compelling narrative format. - True
4. Data storytelling involves presenting raw data without any analysis or interpretation. - False
5. Data storytelling is only effective if it excludes certain data points that might confuse the audience. – False

C. Short answer questions:

1. Define Data Storytelling.
   Ans: Data Storytelling is a means of delivering facts with a compelling narrative to a specific audience. It uses a structured approach to deliver data insights that includes the three main elements - data, visuals and narrative.

2. What is the difference between conflict and resolution?
   Ans- Conflict is the struggle between two or more opposing forces in the story whereas resolution is the end of the story where the characters arrive at a particular situation to resolve the conflict.

3. Name the three elements of Data Storytelling.
   Ans- The three elements are - Data, visuals and narrative.

4. Name some graphs which can be used for the following type of data.
   a. Text data- Word cloud
   b. Data which is changing constantly over a period of time- Line chart
   c. Stocks variation- Candle stick graph
   d. Mixed data- facet grid

5. Name some important ethical concerns related to Data Storytelling.

   Ans- The ethical concerns related to each element of Data Storytelling are as follows:

   1. Data-accuracy concerns
   2. Narrative- transparency concerns
   3. Visualizations- fallacy concerns

6. Define insight in storytelling and briefly explain its significance.

   Ans - An insight in data storytelling refers to a valuable and meaningful observation or understanding derived from analyzing data. It involves uncovering patterns, trends, correlations, or anomalies within the data that provide actionable information or shed light on a particular issue or problem. Insights in data storytelling help stakeholders gain a deeper understanding of the underlying data and empower them to make informed decisions or take appropriate actions based on the findings.

7. Describe Freytag's Pyramid and its application in data storytelling.

   Ans - Freytag's Pyramid is a narrative structure with five stages: exposition, rising action, climax, falling action, and resolution. In data storytelling, it serves as a framework to structure the presentation of data and insights, guiding the audience through a compelling narrative journey from introduction to resolution.

8. Discuss the ethical considerations in data storytelling.

   Ans - Ethical considerations in data storytelling include privacy and consent, bias and fairness, accuracy and integrity, accountability and respect, and continuous learning. It is important to be transparent about data sources, mitigate biases, ensure data accuracy, use power responsibly, and stay informed about evolving ethical standards.

D.Long answer questions:

1. Why has data storytelling become very powerful nowadays?

Ans- In today’s data-rich world, whoever is using data is expected to understand and use data to make decisions everyday, regardless of position, function or skill set. For digital business, data can serve both as input as well as output. Today, data storytelling is filling a vital gap in the business as well as the technical analytics process between technology and people. By combining an increasingly essential resource—data—with a familiar and time-tested form of communication—storytelling, this skill can help more people translate their insights into action. From a data literacy perspective, once an organization’s people are comfortable with reading and working with data, they should also learn how to communicate insights effectively.

2. Explain the steps to create a Data story.

   Ans-: If the data collected is represented in just a series of graphs and charts, it will not serve the purpose to any organization. It should be communicated well with proper narrative, with proper context and meaning, relevance and clarity. The narrative should be able to take the focus of the audience to the correct spot and not miss out on important facts. To find compelling stories in data sets, the following steps are to be followed:

   - Collect the data and organize it.
   - Use proper visualization tools to visualize the data.
   - Then observe the relationships between the data.
   - Finally create a simple narrative which is hidden in the data to be communicated to the audience.

3. What are the different types of data and which type of visualization should we use for which data?

Ans- Data can be broadly classified as Qualitative data and Quantitative data. Under the qualitative data we can nominal data and ordinal data. Under quantitative data , we have discrete data and continuous data. Further this data can come under any of these categories, which are: Text data, Mixed data, Numeric data, Stocks data, Geographic data. Based on the data we are handling we can go for different type of visualizations which are: Word cloud, Facet grid, Line chart, Bar chart, Pie chart, Histogram, Bubble chart, Heat map, Scatter plot, Candle stick, Map chart etc.

E. Case Study Based Questions

Consider each of the given scenarios and answer the MCQs:

1. Case Study:

The marketing team of a retail company conducted a survey to understand customer preferences for product packaging. They collected data on packaging design preferences from customers across different age groups and demographics. Based on the survey results, they created a data storytelling presentation to inform product packaging decisions.

What is the primary objective of the marketing team's data storytelling presentation?

A) To analyze sales trends of different products.
B) To understand customer preferences for product packaging.
C) To evaluate the effectiveness of marketing campaigns.
D) To track inventory levels of various products.

2. Case Study:

A city government collected data on traffic accidents at intersections to identify high-risk areas and prioritize safety improvements. They analyzed the data to identify patterns and trends in accident occurrence and severity. Subsequently, they developed a data storytelling report to present their findings to city officials and propose targeted interventions.

What is the primary purpose of the city government's data storytelling report?

A) To analyze public transportation usage.

B) To identify high-risk areas for traffic accidents.

C) To assess air quality levels in the city.

D) To evaluate the performance of road maintenance crews.

3. Case Study:

A healthcare organization conducted a study to analyze patient satisfaction levels at various hospitals within the network. They collected data on patient experiences, wait times, staff responsiveness, and overall satisfaction ratings. Using this data, they created a data storytelling presentation to share insights with hospital administrators and identify areas for improvement.

What type of data did the healthcare organization primarily analyze in their study?

A) Sales data for medical supplies.

B) Patient satisfaction levels at hospitals.

C) Staffing levels at healthcare facilities.

D) Insurance claims data.

4. Case Study:

A technology company analyzed user engagement data to understand the effectiveness of its mobile app features. They collected data on user interactions, session durations, and feature usage patterns. Based on the analysis, they developed a data storytelling presentation to guide future app development efforts and enhance user experience.

What was the main objective of the technology company's data analysis?

A) To measure employee productivity.

B) To understand user engagement with the mobile app.

C) To track inventory levels of hardware components.

D) To evaluate customer satisfaction with tech support services.

5. Case Study:

An educational institution conducted a survey to gather feedback from students on online learning experiences during the COVID-19 pandemic. They collected data on internet connectivity issues, course content satisfaction, and overall learning effectiveness. Using this data, they created a data storytelling presentation to inform future decisions on online course delivery methods.

What motivated the educational institution to conduct the survey and create the data storytelling presentation?

A) To analyze student enrollment trends.

B) To assess campus infrastructure needs.

C) To gather feedback on online learning experiences.

D) To evaluate faculty performance in virtual classrooms.

F.Competency Based Questions:

1.A sales team at a software company conducted a customer feedback survey to gather insights on product satisfaction, user experience, and feature requests. They collected data on customer ratings, feedback comments, and usage patterns. Based on the survey results, they created a data storytelling presentation to improve product development and customer satisfaction. What did the sales team do to gather insights for product improvement and customer satisfaction?
Ans- The sales team conducted a customer feedback survey, collected data on customer ratings, feedback comments, and usage patterns, and then used the survey results to create a data storytelling presentation.

2.The human resources department at a technology company conducted an employee satisfaction survey to understand workplace culture and job satisfaction. They collected data on aspects like work-life balance, career development, and communication effectiveness. Using the survey results, they created a data storytelling presentation to address employee concerns and enhance overall satisfaction.

1. What is the primary goal of the human resources department's data storytelling presentation?
Ans- The primary goal is to address employee concerns and improve job satisfaction.

2. What specific areas of employee feedback did the human resources department collect data on?
Ans- They collected data on workplace culture and job satisfaction.

3. How does the human resources department plan to utilize the data storytelling presentation based on the employee satisfaction survey results?
Ans- The plan is to use it to address employee concerns and enhance employee satisfaction.

3.A retail company conducted a customer behavior analysis to understand shopping patterns and preferences. They collected data on purchase history, browsing behavior, and customer demographics. Based on the analysis, they created a data storytelling presentation to improve marketing strategies. How did the retail company use data storytelling to enhance their marketing strategies?
Ans- The retail company used data storytelling to analyze customer behaviour, understand shopping patterns and preferences, and improve marketing strategies based on data-driven insights.

4.A healthcare organization analyzed patient satisfaction survey data to identify areas for improvement in patient care and services. They collected data on patient feedback, treatment outcomes, and facility experiences. How can data storytelling be used for improvement in patient care?

Ans-

- Data storytelling helps healthcare organizations improve patient care by analysing patient feedback and treatment outcomes.
- It provides insights into areas that need enhancement, allowing for data-driven decisions that enhance patient experience and quality of care.
- This approach enables clear communication of findings and actionable recommendations, leading to targeted improvements in patient care and services.

5.A technology company conducted a market analysis to identify emerging trends and customer preferences in the tech industry. They collected data on market share, competitor strategies, and consumer behavior. Through data storytelling, they created a presentation to guide product development and strategic decision-making. How did the technology company leverage data storytelling to inform strategic decision-making?

Ans- The technology company used data storytelling to analyze market trends, understand customer preferences, and guide product development and strategic decision-making processes.

References:

1. https://www.unicef.org/india/media/8746/file/THE%20STATE%20OF%20THE%20WORLD%20%E2%80%99S%20CHILDREN%202023.pdf

2. https://www.linkedin.com/pulse/what-data-storytelling-ram-narayan/

3. Link for storytelling: https://www.khanacademy.org/humanities/hass-storytelling/storytellingpixar-in-a-box/ah-piab-we-are-all-storytellers/v/storytelling-introb

4. Link for storytelling: https://www.youtube.com/watch?v=uAG8c-sapUE

5. Link for storytelling: http://storywards.com/en/what-is-storytelling/

Class XII| **Artificial Intelligence** |AI Teacher Handbook