**WEEK 1**

1. **Artificial Intelligence (AI) is promising cutting-edge technology providing intelligent solutions in all sectors today. Define AI and describe applications of AI in different domains**

**Ans**: Artificial intelligence: It is the simulation of human intelligence processes by machines, especially computer systems

AI is also defined as,

* It is the science and engineering of making intelligent machines, especially intelligent computer programs
* Artificial Intelligence (AI) is the machine-displayed intelligence that simulates human behavior or thinking and can be trained to solve specific problems
* Artificial intelligence leverages computers and machines to mimic the problem-solving and decision-making capabilities of the human mind.

A**pplications of AI in different domains**

### E-Commerce

#### Personalized Shopping:.

#### AI-powered Assistants

#### Fraud Prevention

#### Administrative Tasks Automated to Aid Educators:

#### Creating Smart Content:

#### Voice Assistants:

### Personalized Learning

#### Autonomous Vehicles

#### Spam Filters

#### Facial Recognition

#### Recommendation System

### Navigation: [Convolutional Neural Network](https://www.simplilearn.com/tutorials/deep-learning-tutorial/convolutional-neural-network) and Graph Neural Network, which makes lives easier for users by automatically detecting the number of lanes and road types behind obstructions on the roads.

### Robotics

* Carrying goods in hospitals, factories, and warehouses
* Cleaning offices and large equipment

### Inventory management

### Human Resource: Using machine learning software, you can examine applications based on specific parameters. AI drive systems can scan job candidates' profiles, and resumes to provide recruiters an understanding of the talent pool they must choose from

### Healthcare: [AI applications are used in healthcare](https://www.simplilearn.com/ai-in-healthcare-article) to build sophisticated machines that can detect diseases and identify cancer cells. Artificial Intelligence can help analyze chronic conditions with lab and other medical data to ensure early diagnosis

### Agriculture: Artificial Intelligence is used to identify defects and nutrient deficiencies in the soil

### Gaming: AI can be used to create smart, human-like NPCs to interact with the players.

### Automobiles: AI can be used along with the vehicle’s camera, radar, cloud services, GPS, and control signals to operate the vehicle

### Social Media

#### Instagram

#### Facebook

#### Twitter

### Marketing: Using AI, marketers can deliver highly targeted and personalized ads , help with content marketing in a way that matches the brand's style and voice and AI can provide users with real-time personalization’s based on their behavior and can be used to edit and optimize marketing campaigns to fit a local market's needs.

### Chatbots: [AI chatbots can comprehend natural language](https://www.concurrency.com/blog/august-2019/role-of-artificial-intelligence-in-chatbot-development) and respond to people online who use the "live chat" feature that many organizations provide for customer service.

### Finance: It has been reported that [80% of banks recognize the benefits that AI can provide](https://www.businessinsider.com/ai-in-finance). Whether it’s personal finance, corporate finance, or consumer finance, the highly evolved technology that is offered through AI can help to significantly improve a wide range of financial services

1. **Groups of developers want to work Collaboratively on big software project. Each developer in team is assigned one module. How git and GitHub help these developers to build project effectively?**

**Ans:**

This project will use two fictional team members – Harish and Riya – with two separate GitHub accounts.

### Step 1: To create two different GitHub accounts

Create two different GitHub accounts.

### Step 2: Set up your local development environment

We are going to use a local development environment and set up Git on it.

We have to make sure Git is installed in our system:

git –version (This command should return the version of Git that is installed in your system.)

## **Get Started Working on the Project**

### Step 1: Create the repository and build the team

* To create the repository, she simply used the GitHub web interface and clicked on the Repositories tab.
* then on the New button.
* She named the repository TeamProject
* she added a short description and a Readme.md file.
* After she invited Harish to work on it.
* To do so, she clicked on the Settings tab in the TeamProject repository
* then in the Manage access option, and finally in the Add people button.
* After she send invitation link for Harish
* A couple of seconds later, Harish received the invitation to his email:

He accepted it, and by doing so, both team members were ready to start working on their project.

### Step 2: Create a file

* Riya started the project by creating the initial file the TeamProject SaaS product will use: index.html.
* she created the file using the GitHub web interface by positioning herself in the repository and clicking on the Add file > Create new file buttons.
* Then she added the name of the file, its content, and a meaningful commit message.
* After clicking on the Commit new file button, the file was created on the repository.

### Step 3: Create an issue and work on it

* she proceeded to create an **issue** in the repository by clicking on the Issues tab
* clicking on the New issue button.

1. **Summarize the challenges associated with Machine Learning**

**Ans:**

### 1. Inadequate Training Data

The major issue that comes while using machine learning algorithms is the lack of quality as well as quantity of data.

* **Noisy Data-** It is responsible for an inaccurate prediction that affects the decision as well as accuracy in classification tasks.
* **Incorrect data-** It is also responsible for faulty programming and results obtained in machine learning models. Hence, incorrect data may affect the accuracy of the results also.
* **Generalizing of output data-** Sometimes, it is also found that generalizing output data becomes complex, which results in comparatively poor future actions.
* **Poor quality of data:** unclean data lead to less accuracy in classification and low-quality results. Hence, data quality can also be considered as a major common problem while processing machine learning algorithms.

### 2. Non-representative training data

To make sure our training model is generalized well or not, we have to ensure that sample training data must be representative of new cases that we need to generalize. The training data must cover all cases that are already occurred as well as occurring.

### 3. Overfitting and Underfitting

**Overfitting:**

Overfitting is one of the most common issues faced by Machine Learning engineers and data scientists. Whenever a machine learning model is trained with a huge amount of data, it starts capturing noise and inaccurate data into the training data set. It negatively affects the performance of the model.

**Underfitting:**

Underfitting occurs when our model is too simple to understand the base structure of the data, just like an undersized pant. This generally happens when we have limited data into the data set, and we try to build a linear model with non-linear data.

### 4. Monitoring and maintenance: As we know that generalized output data is mandatory for any machine learning model; hence, regular monitoring and maintenance become compulsory for the same. Different results for different actions require data change; hence editing of codes as well as resources for monitoring them also become necessary.

### 5. Getting bad recommendations

A machine learning model operates under a specific context which results in bad recommendations and concept drift in the model. This incident is called a Data Drift. It generally occurs when new data is introduced or interpretation of data changes. However, we can overcome this by regularly updating and monitoring data according to the expectations.

### 6. Lack of skilled resources

The absence of skilled resources in the form of manpower is also an issue. Hence, we need manpower having in-depth knowledge of mathematics, science, and technologies for developing and managing scientific substances for machine learning.

### 7. Customer Segmentation

Customer segmentation is also an important issue while developing a machine learning algorithm. an algorithm is necessary to recognize the customer behavior and trigger a relevant recommendation for the user based on past experience.

### 8. Process Complexity of Machine Learning

The machine learning process is very complex, which is also another major issue faced by machine learning engineers and data scientists.

### 9. Data Bias

Data Biasing is also found a big challenge in Machine Learning. Biased data leads to inaccurate results, skewed outcomes, and other analytical errors. However, we can resolve this error by determining where data is actually biased in the dataset

### 10. Lack of Explainability

This basically means the outputs cannot be easily comprehended as it is programmed in specific ways to deliver for certain conditions. Hence, a lack of explainability is also found in machine learning algorithms which reduce the credibility of the algorithms.

### 11. Slow implementation and results

Machine learning models are highly efficient in producing accurate results but are time-consuming. Slow programming, excessive requirements and overloaded data take more time to provide accurate results than expected.

### 12. Irrelevant features

Although machine learning models are intended to give the best possible outcome, if we feed garbage data as input, then the result will also be garbage. Hence, we should use relevant features in our training sample. A machine learning model is said to be good if training data has a good set of features or less to no irrelevant features.

1. **How AI Software Development life cycle differs from traditional software development? Explain**

The AI software development life cycle differs from traditional software development in a few key ways:

1. Data collection and preparation: In AI software development, the first step is typically to collect and prepare the data that will be used to train the AI model. This involves cleaning and formatting the data, as well as potentially using techniques such as feature engineering to extract relevant information from the data.

2. Model training: Once the data is ready, the next step is to train the AI model using the data. This involves using machine learning algorithms to analyze the data and identify patterns and relationships within it.

3. Model evaluation: After the model has been trained, it is important to evaluate its performance to ensure that it is accurate and effective. This may involve testing the model on a separate dataset and comparing the results to a baseline performance.

4. Model deployment: Once the model has been trained and evaluated, it can be deployed to a production environment where it can be used to make predictions or decisions.

Overall, the AI software development life cycle involves more data-focused steps than traditional software development, as well as more iterations and evaluations of the model to ensure its accuracy and effectiveness.

1. **Summarize any two cloud deployment models**

• Private Cloud: Resource managed and used by the organization.

• Public Cloud: Resource available for the general public under the Pay as you go model.

• Community Cloud: Resource shared by several organizations, usually in the same industry.

• Hybrid Cloud: This cloud deployment model is partly managed by the service provided and partly by the organization.

**Public Cloud Model**

As its names suggest, the public cloud is available to the general public, and resources are shared between all users. They are available to anyone, from anywhere, using the Internet. The public cloud deployment model is one of the most popular types of cloud.

**Public Cloud Architecture**

This computing model is hosted at the vendor’s data center. The public cloud model makes the resources, such as storage and applications, available to the public over the WWW. It serves all the requests; therefore, resources are almost infinite

**Characteristics of Public Cloud**

Here are the essential characteristics of the Public Cloud:

• Uniformly designed Infrastructure

• Works on the Pay-as-you-go basis

• Economies of scale

• SLA guarantees that all users have a fair share with no priority

• It is a multitenancy architecture, so data is highly likely to be leaked

**Advantages of Public Cloud Deployments**

Here are the pros/benefits of the Public Cloud Deployment Model:

• Highly available anytime and anywhere, with robust permission and authentication mechanism.

• There is no need to maintain the cloud.

• Does not have any limit on the number of users.

• The cloud service providers fully subsidize the entire Infrastructure. Therefore, you don’t need to set up any hardware.

• Does not cost you any maintenance charges as the service provider does it.

• It works on the Pay as You Go model, so you don’t have to pay for items you don’t use.

• There is no significant upfront fee, making it excellent for enterprises that require immediate access to resources.

**Disadvantages of Public Cloud Deployments**

Here are the cons/drawbacks of the Public Cloud Deployment Model:

• It has lots of issues related to security.

• Privacy and organizational autonomy are not possible.

• You don’t control the systems hosting your business applications.

**Private Cloud Model**

The private cloud deployment model is a dedicated environment for one user or customer. You don’t share the hardware with any other users, as all the hardware is yours.

It is a one-to-one environment for single use, so there is no need to share your hardware with anyone else. The main difference between private and public cloud deployment models is how you handle the hardware. It is also referred to as “internal cloud,” which refers to the ability to access systems and services within an organization or border.

1. **Write steps to Create repository in GitHub and add file.**

**Step 1:**

Create a local directory using the following command:

$ mkdir test

$ cd test

**Step 2:**

The next step is to initialize the directory:

$ git init

**Step 3:**

Go to the folder where "test" is created and create a text document named "demo." Open "demo" and put any content, like "Hello Simplilearn." Save and close the file.

**Step 4:**

Enter the Git bash interface and type in the following command to check the status:

$ git status

**Step 5:**

Add the "demo" to the current directory using the following command:

$ git add demo.txt

**Step 6:**

Next, make a commit using the following command:

$ git commit -m "committing a text file"

**Step 7:**

Link the Git to a Github Account:

$ git config --global user.username

**Step 8:**

Open your Github account and create a new repository with the name "test\_demo" and click on "Create repository." This is the remote repository. Next, copy the link of "test\_demo."

**Step 9:**

Go back to Git bash and link the remote and local repository using the following command:

$ git remote add origin <link>

Here, <link> is the link copied in the previous step.

**Step 10:**

Push the local file onto the remote repository using the following command:

$ git push origin master

**Step 11:**

Move back to Github and click on "test\_demo" and check if the local file "demo.txt" is pushed to this repository.

1. **Is data which is collected by various applications ethical in nature? Justify your answer**

Yes, most of the times, the data collected by various applications is ethical in nature as the users agree to it by clicking on allow when the application asks for various permissions. They ask for our data for various facilities like - to show us personalized recommendations and advertisements and to make their app more accurate and efficient.

**OR**

No, the data collected by various applications is not always ethical in nature. Sometimes, we just share our data to non – trusted third party applications without reading what happens to our data. This may lead to unethical use of our data. If one does not want to share his/her data with anyone, he/she can opt for alternative applications which are of similar usage and keep the data private.

**For example**: an alternative to WhatsApp is the Telegram app which does not collect any data from us.

**WEEK 2**

1. **How Big data is different from the data stored in traditional databases? Elaborate**

**Ans:**

|  |  |
| --- | --- |
| **Traditional Data** | **Big Data** |
| Traditional data is generated in enterprise level. | Big data is generated outside the enterprise level. |
| Its volume ranges from Gigabytes to Terabytes. | Its volume ranges from Petabytes to Zettabytes or Exabytes. |
| Traditional database system deals with structured data. | Big data system deals with structured, semi-structured, database, and unstructured data. |
| Traditional data is generated per hour or per day or more. | But big data is generated more frequently mainly per seconds. |
| Traditional data source is centralized and it is managed in centralized form. | Big data source is distributed and it is managed in distributed form. |
| Data integration is very easy. | Data integration is very difficult. |
| Normal system configuration is capable to process traditional data. | High system configuration is required to process big data. |
| The size of the data is very small. | The size is more than the traditional data size. |
| Traditional data base tools are required to perform any data base operation. | Special kind of data base tools are required to perform any databaseschema-based operation. |
| Normal functions can manipulate data. | Special kind of functions can manipulate data. |
| Its data model is strict schema based and it is static. | Its data model is a flat schema based and it is dynamic. |
| Traditional data is stable and inter relationship. | Big data is not stable and unknown relationship. |
| Traditional data is in manageable volume. | Big data is in huge volume which becomes unmanageable. |
| It is easy to manage and manipulate the data. | It is difficult to manage and manipulate the data. |
| Its data sources include ERP transaction data, CRM transaction data, financial data, organizational data, web transaction data etc. | Its data sources include social media, device data, sensor data, video, images, audio etc. |

**1.**[**Traditional data**](https://www.geeksforgeeks.org/traditional-data-mining-life-cycle-crisp-methodology/)**:** Traditional data is the structured data that is being majorly maintained by all types of businesses starting from very small to big organizations. In a traditional database system, a centralized database architecture used to store and maintain the data in a fixed format or fields in a file. For managing and accessing the data [Structured Query Language (SQL)](https://www.geeksforgeeks.org/sql-tutorial/) is used.

**2.**[**Big data**](https://www.geeksforgeeks.org/what-is-big-data/)**:** We can consider big data an upper version of traditional data. Big data deal with too large or complex data sets which is difficult to manage in traditional data-processing application software. It deals with large volume of both structured, semi structured and unstructured data. Volume, Velocity and Variety, Veracity and Value refer to the[5’V characteristics of big data](https://www.geeksforgeeks.org/5-vs-of-big-data/). Big data not only refers to large amount of data it refers to extracting meaningful data by analysing the huge amount of complex data sets. semi-structured.

1. **Differentiate between supervised machine learning and Unsupervised machine learning**

|  |  |
| --- | --- |
| **Supervised Learning** | **Unsupervised Learning** |
| Supervised Learning can be used for 2 different types of problems i.e. regression and classification | Unsupervised Learning can be used for 2 different types of problems i.e. clustering and association. |
| Input Data is provided to the model along with the output in the Supervised Learning. | Only input data is provided in Unsupervised Learning. |
| Output is predicted by the Supervised Learning. | Hidden patterns in the data can be found using the unsupervised learning model. |
| Labeled data is used to train supervised learning algorithms. | Unlabeled data is used to train unsupervised learning algorithms. |
| Accurate results are produced using a supervised learning model. | The accuracy of results produced are less in unsupervised learning models. |
| Training the model to predict output when a new data is provided is the objective of Supervised Learning. | Finding useful insights, hidden patterns from the unknown dataset is the objective of the unsupervised learning. |
| Supervised Learning includes various algorithms such as Bayesian Logic, Decision Tree, Logistic Regression, Linear Regression, Multi-class Classification, Support Vector Machine etc. | Unsupervised Learning includes various algorithms like KNN, Apriori Algorithm, and Clustering. |
| To assess whether right output is being predicted, direct feedback is accepted by the Supervised Learning Model. | No feedback will be taken by the unsupervised learning model. |
| In Supervised Learning, for right prediction of output, the model has to be trained for each data, hence Supervised Learning does not have close resemblance to Artificial Intelligence. | Unsupervised Learning has more resemblance to Artificial Intelligence, as it keeps learning new things with more experience. |
| Number of classes are known in Supervised Learning. | Number of classes are not known in Unsupervised Learning |
| In scenarios where one is aware of output and input data, supervised learning can be used. | In the scenarios where one is not aware of output data, but is only aware of the input data then Unsupervised Learning could be used. |
| Computational Complexity is very complex in Supervised Learning compared to Unsupervised Learning | There is less computational complexity in Unsupervised Learning when compared to Supervised Learning. |
| Supervised Learning will use off-line analysis | Unsupervised Learning uses Real time analysis of data. |
| Some of the applications of Supervised Learning are Spam detection, handwriting detection, pattern recognition, speech recognition etc. | Some of the applications of Unsupervised Learning are detecting fraudulent transactions, data preprocessing etc. |

1. **Explain different machine learning types**

**There are 3 types of machine learning algorithms:**

* + - 1. Supervised Learning
      2. Unsupervised Learning
      3. Reinforcement Learning

**Supervised Learning**

Supervised learning is one of the most basic types of machine learning. In this type, the machine learning algorithm is trained on labeled data. Even though the data needs to be labeled accurately for this method to work, supervised learning is extremely powerful when used in the right circumstances.

In supervised learning, the ML algorithm is given a small training dataset to work with. This training dataset is a smaller part of the bigger dataset and serves to give the algorithm a basic idea of the problem, solution, and data points to be dealt with. The training dataset is also very similar to the final dataset in its characteristics and provides the algorithm with the labeled parameters required for the problem.

The algorithm then finds relationships between the parameters given, essentially establishing a cause and effect relationship between the variables in the dataset. At the end of the training, the [algorithm](https://it.toolbox.com/blogs/shrutiumathe/mphasis-unveils-its-deep-learning-algorithms-on-amazon-web-services-marketplace-for-machine-learning-073119) has an idea of how the data works and the relationship between the input and the output.

This solution is then deployed for use with the final dataset, which it learns from in the same way as the training dataset. This means that supervised machine learning algorithms will continue to improve even after being deployed, discovering new patterns and relationships as it trains itself on new data. **L**

**Unsupervised Learning**

Unsupervised machine learning holds the advantage of being able to work with unlabeled data. This means that human labor is not required to make the dataset machine-readable, allowing much larger datasets to be worked on by the program.

In supervised learning, the labels allow the algorithm to find the exact nature of the relationship between any two data points. However, unsupervised learning does not have labels to work off of, resulting in the creation of hidden structures. Relationships between data points are perceived by the algorithm in an abstract manner, with no input required from human beings.

The creation of these hidden structures is what makes unsupervised learning algorithms versatile. Instead of a defined and set problem statement, unsupervised learning algorithms can adapt to the data by dynamically changing hidden structures. This offers more post-deployment development than supervised learning algorithms.

## [**Reinforcement learning**](https://it.toolbox.com/article/openais-robot-learns-to-solve-a-rubiks-cube-with-one-hand-peter-welinder-research-lead-openai-shares-insights)

 It directly takes inspiration from how human beings learn from data in their lives. It features an algorithm that improves upon itself and learns from new situations using a trial-and-error method. Favorable outputs are encouraged or ‘reinforced’, and non-favorable outputs are discouraged or ‘punished’.

Based on the psychological concept of conditioning, reinforcement learning works by putting the algorithm in a work environment with an interpreter and a reward system. In every iteration of the algorithm, the output result is given to the interpreter, which decides whether the outcome is favorable or not.

In case of the program finding the correct solution, the interpreter reinforces the solution by providing a reward to the algorithm. If the outcome is not favorable, the algorithm is forced to reiterate until it finds a better result. In most cases, the reward system is directly tied to the effectiveness of the result.

In typical reinforcement learning use-cases, such as finding the shortest route between two points on a map, the solution is not an absolute value. Instead, it takes on a score of effectiveness, expressed in a percentage value. The higher this percentage value is, the more reward is given to the algorithm. Thus, the program is trained to give the best possible solution for the best possible reward.

1. **Discuss various attributes of a high-quality data**

## Metrics of data quality

### 1. Completeness

Completeness relates to whether all required information is present in the dataset. For example, if the customer information in a database is required to include both first and last names, any record in which the first name or last name field is not populated is marked as incomplete. The metric used in assessing this dimension is the percentage of records that are complete.

### 2. Validity

Data is characterized as valid if it matches the rules specified for it. Those rules typically include specifications such as format (number of digits, etc.), allowable types (integer, floating-point, string, etc.), and range (minimum and maximum values). For example, a telephone number field that contains the string ‘1809 Oak Street’ is not valid. The metric for this dimension is the percentage of records in which all values are valid.

### 3. Timeliness

Timeliness relates to whether the information is up to date for the intended use. In other words, is the correct information available when needed?

For example, if a customer has notified the company of an address change, but the new address is not in the database at the time billing statements are processed, that entry fails the timeliness test. The metric used to measure timeliness is the time difference between when data is needed and when it is available.

### 4. Consistency

A data item is consistent if all representations of that item across data stores match.

If, for example, a birth date is entered in one system using the U.S. format (mm/dd/yyyy), but it is imported into another system where the date is entered using the [European standard](https://en.wikipedia.org/wiki/Date_and_time_notation_in_Europe) (dd/mm/yyyy), that data lacks consistency.

And

* 1. Accuracy: High-quality data is accurate and free from errors.
  2. Relevance: High-quality data is relevant to the problem that the AI model is trying to solve.
  3. Diversity: High-quality data is diverse and representative of the real-world scenario that the model will be applied to.
  4. Completeness: High-quality data is complete and includes all necessary information.
  5. Timeliness: High-quality data is up-to-date and current.
  6. Consistency: High-quality data is consistent and follows established standards and conventions.

**WEEK 3**

* + - 1. Here given the bating statistics of Indian great cricketer Anil Kumble. Perform the following operations-

1. Aggregation
2. Grouping
3. Time series
4. Filter
5. Vectorized

**Ans:**

**i.Aggregation**

import numpy as np

import pandas as pd

df=pd.read\_csv("match.csv")

df

df.mean()

df.max()

df.median()

df.min()

**ii. grouping**

**#Grouping and perform count over each group**

import pandas as pd

dept\_match\_num =df.groupby('ipl')['ipl'].count()

print(dept\_match\_num)

**iii.Time series operations**

%%time

import numpy as np

df=pd.read\_csv("match.csv")

df

**iv .Filter**

c=df.loc[:,'test']

print('printing the details of all Branches',c)

b\_arr=np.array(c)

filterarray=b\_arr == 68

print(filterarray)

**v. vectorization operation**

tic = time.process\_time()

dot = 0.0;

for i in range(len(a)):

dot += a[i] \* b[i]

toc = time.process\_time()

print("dot\_product = "+ str(dot));

print("Computation time = " + str(1000\*(toc - tic )) + "ms")

n\_tic = time.process\_time()

n\_dot\_product = numpy.dot(a, b)

n\_toc = time.process\_time()

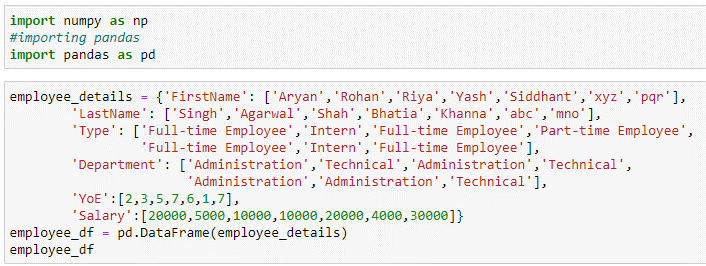
print("\nn\_dot\_product = "+str(n\_dot\_product))

print("Computation time = "+str(1000\*(n\_toc - n\_tic ))+"ms")

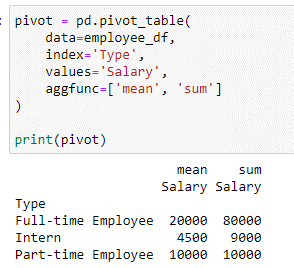
* + - 1. **Create a dataframe with following data.**

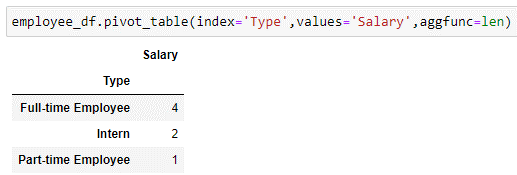


1. Make a pivot table which shows average salary of each type of employee for each department.

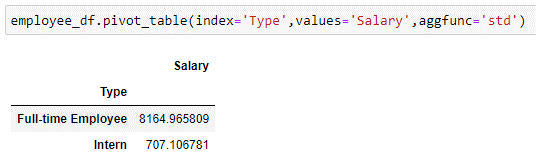


1. Make a pivot table which shows the sum and mean of the salaries of each type of employee and the number of employees of each type.



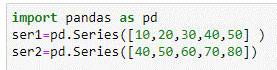


1. Make a pivot table which shows standard deviation for salary column.

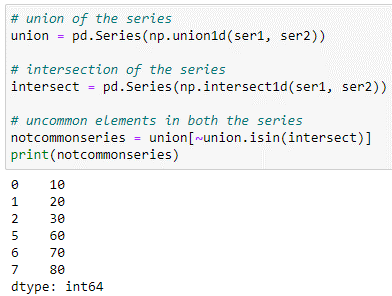


**3. Create two series as shown using pd.series() function.**

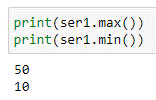
Series\_A = [10,20,30,40,50] Series\_B = [40,50,60,70,80].



i. Get the items not common to both.



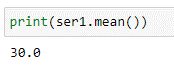
ii. Identify the smallest and largest element in the series A



iii. Find the sum of series B



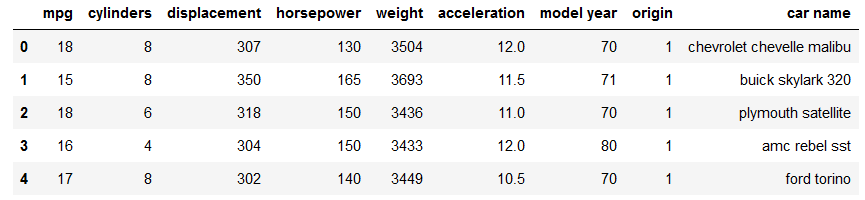
iv. Calculate average in the series A



v. Find median in the given series B



1. **Perform the following operations on Car manufacturing company dataset auto-mpg.csv given below using pandas**

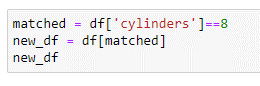


Read data from an existing file

b) statistical details of dataset



c) Get all cars with 8 cylinders



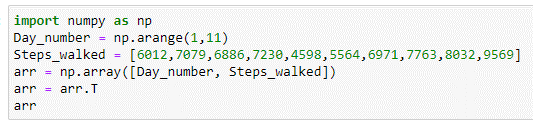
d) Get the number of cars manufactured in each year.



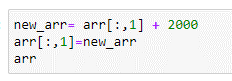
1. **Lee decides to walk 10000 steps every day to combat the effect that lockdown has had on his body’s agility, mobility, flexibility and strength. Consider the following data from fitness tracker over a period of 10days**

|  |  |
| --- | --- |
| Day number | Steps walked |
| 1 | 6012 |
| 2 | 4079 |
| 3 | 6386 |
| 4 | 5230 |
| 5 | 4598 |
| 6 | 5564 |
| 7 | 6971 |
| 8 | 7763 |
| 9 | 8032 |
| 10 | 8569 |

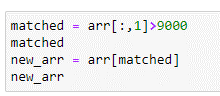
**i) Represent the above data in a 10x2array. In each row, the first element should contain day number and second element should contain steps walked.**



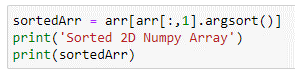
**ii) Lee notices that the tracker’s battery dies every day at 7 pm. Lee discovers that on an average, he walks 2000 steps every day after 7 pm. Perform an appropriate**



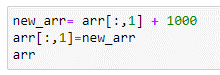
**iii)Write a program that returns the steps walked if the steps walked are more than 9000.**



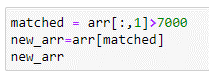
**iv) Print an array containing steps walked in sorted order.**



**v) Perform an appropriate operation to add 1000 steps to all the observations using pandas**



**vi) Find out the days on which he walked more than 7000 steps using pandas**



**6. Write python code to explain map (), filter (), reduce (), lambda()**

**Ans:**

**Lambda**

# Example of lambda function using if-else

Max = lambda a, b : a if(a > b) else b

Max(4,5)

**Map**

Being a higher-order function, the map function takes another function and an iterable (e.g., a list, set, tuple) as input, applies the function to the iterable, and returns an output. It’s syntax is defined as follows:

map(function, iterable)

**Example:**

def add4(x):

return x+4

list1 = [4,6,7,8,9]

list2 = list(map(add4,list1))

print(list2)

**Filter()**

Similar to map(), the filter() higher-order function takes a function and an iterable as inputs. The function in case needs to be of a Boolean nature, returning True/False values corresponding to the filter conditions. As output, it returns a subset of the input data that meets the conditions stipulated by the function.

**Example:**

def oddeven(x):

if x%2 == 0:

return True

else:

return False

list1 = [4,5,6,7,8,9]

evenlist = list(filter(oddeven,list1))

print(evenlist)

**Reduce**

It returns a single value (i.e., it reduces the input to a single element). Commonly, this would be something like the sum of all elements in a list.

**Example:**

from functools import reduce

def sum(x,y):

return x+y

list1 = [6,7,8,9]

s = reduce(sum,list1)

print(s)

7. **Assume Iris dataset and write the code**

a. print first 5 record

b. print the size of the data for given data set

b. Use scatter plot to compare petal length and petal width

c. check for missing values

b. print summarizes of the dataset

e. Count plot for the spices

c. Visualize the distribution of any one column

d. Visualize the relationship between any two variable

e. Print the information of all column in the dataset

f. Visualize the spices column using bar graph

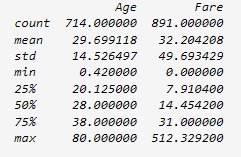
1. df.head(5)
2. print(df.shape)
3. plt.scatter(df.index,df['petal.width'])

plt.show()

1. df.isna().sum()
2. df.describe()
3. sns.countplot(x=df['variety'])
4. sns.displot(df['petal.width'],kde=False,color='blue',bins=10)
5. sns.heatmap(data=df,annot=True) df.info()
6. df['variety'].value\_counts().plot.bar()

**WEEK 4**

1. **The statistical summary of titanic dataset with respect to age and fare is as follows. Analysis and explain statistical matrices from below summary.**



If the Data Frame contains numerical data, the description contains this information for each column:

count - The number of not-empty values.

mean - The average (mean) value.

std - The standard deviation.

min - the minimum value.

25% - The 25% percentile\*.

50% - The 50% percentile\*.

75% - The 75% percentile\*.

max - the maximum value.

1. **Difference between Covariance and Correlation.**

|  |  |  |
| --- | --- | --- |
| **Basis for comparison** | **Covariance** | **Correlation** |
| **Definition** | Covariance is an indicator of the extent to which 2 random variables are dependent on each other. A higher number denotes higher dependency. | Correlation is a statistical measure that indicates how strongly two variables are related. |
| **Values** | The value of covariance lies in the range of -∞ and +∞. | Correlation is limited to values between the range -1 and +1 |
| **Change in scale** | Affects covariance | Does not affect the correlation |
| **Unit-free measure** | No | Yes |

1. **Describe the univariate and multivariate analysis.**

Univariate analysis:- It is the simplest of the three analyses where the data you are analyzing is only one variable. There are many different ways people use univariate analysis. The most common univariate analysis is checking the central tendency (mean, median and mode), the range, the maximum and minimum values, and standard deviation of a variable.

**Types:**

* calculations of frequencies
* central tendency
* dispersion

**main purpose of univariate analysis**

* Univariate analysis **explores each variable in a data set, separately**.
* It looks at the range of values, as well as the central tendency of the values.
* It describes the pattern of response to the variable.
* It describes each variable on its own.

**Working**

Univariate Analysis works by examining the effects of a singular variable on a set of data. For example, a frequency distribution table is a form of univariate analysis as frequency is the only variable being measured. Alternative variables may be age, height, weight, etc., however it is important to note that as soon as a secondary variable is introduced it becomes bivariate analysis. With three or more variables, it becomes multivariate analysis.

**Advantages**

* Univariate analysis refers to the quantitative data exploration we do at the beginning of any analysis.
* These analyses provide us with descriptions of single variables.
* we are interested in using in more advanced tests and help us narrow down exactly

**Multivariate Analysis:-**

Multivariate analysis is similar to Bivariate analysis but you are comparing more than two variables. For three variables, you can create a 3-D model to study the relationship (also known as Trivariate Analysis). However, since we cannot visualize anything above the third dimension, we often rely on other softwares and techniques for us to be able to grasp the relationship in the data.

**Types**

* Dependence techniques, which look at cause-and-effect relationships between variables, interdependence techniques, which explore the structure of a dataset.

**Main purpose of Multivariate analysis**

* Multivariate analysis is used **to describe analyses of data where there are multiple variables or observations for each unit or individual**.
* Often times these data are interrelated and statistical methods are needed to fully answer the objectives of our research.

**Working**

Multivariate is a controlled or supervised Machine Learning algorithm that **analyses multiple data variables**. It is a continuation of multiple regression that involves one dependent variable and many independent variables. The output is predicted based on the number of independent variables.

**Advantage**

* The main advantage of multivariate analysis is that **it considers more than one factor in data analysis**.
* It looks at the various independent variables that influence the dependent variable.
* The conclusions you draw from MVA are also more likely to be accurate.

1. **Justify the Significance of Exploratory Data Analysis?**

Exploratory Data Analysis refers to the critical process of performing initial investigations on data so as to discover patterns, to spot anomalies, to test hypothesis and to check assumptions with the help of summary statistics and graphical representations.

The purpose of Exploratory Data Analysis is essential to tackle specific tasks such as:

* Spotting missing and erroneous data;
* Mapping and understanding the underlying structure of your data;
* Identifying the most important variables in your dataset;
* Testing a hypothesis or checking assumptions related to a specific model;
* Establishing a model that can explain your data using minimum variables
* Estimating parameters and figuring the margins of error.

Importance of using EDA for analyzing data sets is:

* Helps identify errors in data sets.
* Gives a better understanding of the data set.
* Helps detect outliers or anomalous events.
* Helps understand data set variables and the relationship among them.
* Improve understanding of variables by extracting averages, mean, minimum, and maximum values, etc.

1. **A data set is given to you creating a machine learning model. What are the steps followed before using the data for training the model? Elaborate each step.**

**Steps followed before using the data to train the model**

**1. Collecting Data:**

As you know, machines initially learn from the data that you give them. It is of the utmost importance to collect reliable data so that your machine learning model can find the correct patterns. The quality of the data that you feed to the machine will determine how accurate your model is. If you have incorrect or outdated data, you will have wrong outcomes or predictions which are not relevant.

Make sure you use data from a reliable source, as it will directly affect the outcome of your model. Good data is relevant, contains very few missing and repeated values, and has a good representation of the various subcategories/classes present.

**2. Preparing the Data:**

After you have your data, you have to prepare it. You can do this by :

• Putting together all the data you have and randomizing it. This helps make sure that data is evenly distributed, and the ordering does not affect the learning process.

• Cleaning the data to remove unwanted data, missing values, rows, and columns, duplicate values, data type conversion, etc. You might even have to restructure the dataset and change the rows and columns or index of rows and columns.

• Visualize the data to understand how it is structured and understand the relationship between various variables and classes present.

• Splitting the cleaned data into two sets - a training set and a testing set. The training set is the set your model learns from. A testing set is used to check the accuracy of your model after training.

**After the data is prepared the following steps are applied**

**3. Choosing a Model:**

A machine learning model determines the output you get after running a machine learning algorithm on the collected data. It is important to choose a model which is relevant to the task at hand. Over the years, scientists and engineers developed various models suited for different tasks like speech recognition, image recognition, prediction, etc. Apart from this, you also have to see if your model is suited for numerical or categorical data and choose accordingly.

**4. Training the Model:**

Training is the most important step in machine learning. In training, you pass the prepared data to your machine learning model to find patterns and make predictions. It results in the model learning from the data so that it can accomplish the task set. Over time, with training, the model gets better at predicting

**5. Evaluating the Model:**

After training your model, you have to check to see how it’s performing. This is done by testing the performance of the model on previously unseen data. The unseen data used is the testing set that you split our data into earlier. If testing was done on the same data which is used for training, you will not get an accurate measure, as the model is already used to the data, and finds the same patterns in it, as it previously did. This will give you disproportionately high accuracy.

**6. Parameter Tuning:**

Once you have created and evaluated your model, see if its accuracy can be improved in any way. This is done by tuning the parameters present in your model. Parameters are the variables in the model that the programmer generally decides. At a particular value of your parameter, the accuracy will be the maximum. Parameter tuning refers to finding these values

**7. Making Predictions**

In the end, you can use your model on unseen data to make predictions accurately

1. **Explore different types of data in machine learning.**

**Types of Data**

**1. Type 1**

**a) Cross-sectional Data**

Cross-section data is collected in a single time period and is characterized by individual units - people, companies, countries, etc. Some examples include:

* + Student grades at the end of the current semester;
  + Household data of the previous year - expenditure on food, unemployment, income, etc.
  + Car data - average speed, horsepower, colour, etc.
  + With cross-sectional data the ordering of the data does not matter. In other words, we can order the data by ascending, descending or even randomized order and this will not affect out modelling results.
  + Cross-sectional data or cross-section of a population is obtained by taking observations from multiple individuals at the same point in time.

**b) Time Series Data**

Data collected at a number of specific points in time is called time series data. Such examples include stock prices, interest rates, exchange rates as well as product prices, GDP, etc.

* + Time series data can be observed at many different frequencies (hourly, daily, weekly, monthly, quarterly, annually etc)
  + Unlike cross-sectional data, the ordering of the data is important in time-series data.
  + Each point represents the values at specific points in time.
  + As such, time series data are typically presented in chronological order.
  + Changing the order of the data ignores the time-dimensionality of the data.

Example:

* + If we consider only one country, and take a look at its military expenses and central government debt for a span of 10 years from 2001 to 2010, that would get two time series - one about the military expenditure and the other about debt of central government debt.

**2.Type 2**

**a) Univariate data**

* + This type of data consists of only one variable.
  + The analysis of univariate data is thus the simplest form of analysis since the information deals with only one quantity that changes.
  + It does not deal with causes or relationships and the main purpose of the analysis is to describe the data and find patterns that exist within it.
  + The example of a univariate data can be height.

**b) Bivariate data**

* + This type of data involves two different variables.
  + The analysis of this type of data deals with causes and relationships and the analysis is done to find out the relationship among the two variables.
  + Example of bivariate data can be temperature and ice cream sales in summer season.
  + Suppose the temperature and ice cream sales are the two variables of a bivariate data.
  + Here, the relationship is visible from the table that temperature and sales are directly proportional to each other and thus related because as the temperature increases, the sales also increase.
  + Thus, bivariate data analysis involves comparisons, relationships, causes and explanations.
  + These variables are often plotted on X and Y axis on the graph for better understanding of data and one of these variables is independent while the other is dependent.

**c) Multivariate data**

* + When the data involves three or more variables, it is categorized under multivariate.
  + Example of this type of data is suppose an advertiser wants to compare the popularity of four advertisements on a website, then their click rates could be measured for both men and women and relationships between variables can then be examined.
  + It is similar to bivariate but contains more than one dependent variable.
  + The ways to perform analysis on this data depends on the goals to be achieved.
  + Some of the techniques are regression analysis, path analysis, factor analysis and multivariate analysis of variance (MANOVA)

**WEEK 5**

1. **For the given data set Perform the following operations:**
2. **Check statistical info of the data set**
3. **Plot a line plot showing total profit on y axis and number column on x axis**
4. **Find the missing values**
5. **Find the sum of total profit**
6. **Find the max value from Drawing sheets column**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Number | Pencil | textbooks | Drawing sheets | Total units | profit |
| 1 | 300 | 250 | 100 | 700 | 80000 |
| 2 | 350 | 350 | 125 | 1075 | 9500 |
| 3 | 400 | 400 | 190 | 1320 | 10256 |
| 4 | 500 | 420 | 210 | 1510 | 12000 |
| 5 | 520 | 500 | 250 |  | 15000 |

**Ans:**

import numpy as np

import pandas as pd

book=pd.read\_csv('Book.csv')

**i) Check statistical info of the data set**

book.info()

book.describe()

**ii) line plot showing total profit on y axis and number column on x axis**

import matplotlib.pyplot as plt

book.plot(x='number',y='total profit')

plt.show()

**iii) The missing values**

book.isnull().sum()

**iv) sum of total profit**

book['total profit'].sum()

**v) max value from textbook**

book['textbooks'].agg(['max'])

1. **How to handle the missing values in the dataset? Explain.**

Ways to handle missing values in the dataset:

1. Deleting Rows with missing values
2. Deleting attributes having large number of missing values
3. Impute missing values for continuous variables using mean and median
4. Impute missing values for categorical variables using mode
5. Other Imputation Methods such as forward fill and backward fill

## **Delete Rows with Missing Values:**

Missing values can be handled by deleting the rows or columns having null values. If columns have more than half of the rows as null then the entire column can be dropped. The rows which are having one or more columns values as null can also be dropped.

**Deleting attributes having large number of missing values**

If one or two columns in the dataset are missing a large number or values, then instead of deleting all the rows containing missing values, we can delete the attributes where such missing values are present.

## **Impute missing values with Mean/Median:**

Columns in the dataset which are having numeric continuous values can be replaced with the mean, median, or mode of remaining values in the column. This method can prevent the loss of data compared to the earlier method. Replacing the above two approximations (mean, median) is a statistical approach to handle the missing values.

## **Imputation method for categorical columns with mode:**

When missing values is from categorical columns (string or numerical) then the missing values can be replaced with the most frequent category. If the number of missing values is very large then it can be replaced with a new category.

**Other Imputation Methods such as forward fill and backward fill**

Pandas dataframe.ffill() function is used to fill the missing value in the dataframe. ‘ffill’ stands for ‘forward fill’ and will propagate last valid observation forward. Pandas**dataframe.bfill()** is used to backward fill the missing values in the dataset. It will backward fill the NaN values that are present in the pandas dataframe.

**3. Explain different challenges involved in Data Integration.**

**Challenges in Data Integration**

**Challenge 1 – entity identification**

* The entity identification challenge may occur when the data sources are being integrated by adding attributes.
* The challenge is that the data objects in all the data sources are the same real-world entities with the same definitions of data objects, but they are not easy to connect due to the unique identifiers in the data sources.
* For instance, if the sales department and the marketing department did not use a central customer unique identifier for all their customers.

**Challenge 2 – unwise data collection**

* This data integration challenge happens, as its name suggests, due to unwise data collection.
* For instance, instead of using a centralized database, the data of different data objects is stored in multiple files.
* Regardless, in these situations, our goal is to make sure that the data is integrated into one standard data structure. This type of data integration challenge happens when data objects are being added.

**Challenge 3 – index mismatched formatting**

* When we start integrating data sources by adding attributes, we will use the pandas DataFrame .join() function to connect the rows of two DataFrames that have the same indices.
* To use this valuable function, the integrating DataFrames needs to have the same index formatting; otherwise, the function will not connect the rows.

**Challenge 4 – aggregation mismatch**

* This challenge occurs when integrating data sources by adding attributes.
* When integrating time series data sources whose time intervals are not identical, this challenge arises.

**Challenge 5 – duplicate data objects**

* This challenge occurs when we're integrating data sources by adding data objects.
* When the sources contain data objects that are also in the other sources, when the data sources are integrated, there will be duplicates of the same data objects in the integrated dataset.

**Challenge 6 – data redundancy**

* Unlike the previous challenge, this challenge may be faced when you're integrating data sources by adding attributes.
* As the name suggests, after data integration, some of the attributes may be redundant.
* This redundancy could be shallow as there are two attributes with different titles but the same data.

1. **How to handle the outliers in the dataset? Explain**

**1. Deleting the values:** You can delete the outliers if you know that the outliers are wrong or if the reason the outlier was created is never going to happen in the future. For example, there is a data set of peoples ages and the usual ages lie between 0 to 90 but there is data entry off the age 150 which is nearly impossible. So, we can safely drop the value that is 150.

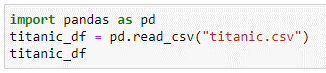
**2. Changing the values:** We can also change the values in the cases when we know the reason for the outliers. We can use the upper cap and lower cap values to replace the outlier values.

**3. Data transformation(log):** Data transformation is useful when we are dealing with highly skewed data sets. By transforming the variables, we can eliminate the outliers for example taking the natural log of a value reduces the variation caused by the extreme values. This can also be done for data sets that do not have negative values.

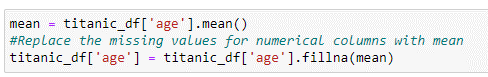
4. **Using different analysis methods**: You could also use different statistical tests that are not as much impacted by the presence of outliers – for example using median to compare data sets as opposed to mean or use of equivalent nonparametric tests etc.

**5. Valuing the outliers:** In case there is a valid reason for the outlier to exist and it is a part of our natural process, we should investigate the cause of the outlier as it can provide valuable clues that can help you better understand your process performance. For example, normal delivery of orders takes 1-2 days, but a few orders took more than a month to complete. Understanding the reason why it took a month and fixing this process can help future customers as they would not be impacted by such large wait times.

1. **Write python code to imputation the missing values in the dataset using, mean and median method**



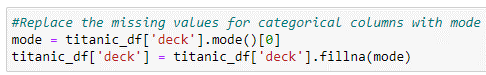
**Replacing with mean**



**Replacing with median**

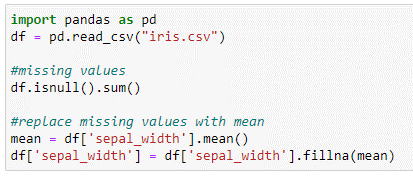


**Replacing with Mode**



6. A company wants to study iris dataset to make predictions. However, the data gathered is not clean for analysis. The company requests you to write a python code to perform the following operations for data driven competitive advantage (Assume dataset with missing values)

* Check for missing values
* Replace missing values with mean value



**7. A company has collected customer comments on its products, rating them as safe or unsafe, using decision trees. The training dataset has the following features: id, date, full review, full review summary, and a binary safe/unsafe tag. During training, any data sample with missing features was dropped. In a few instances, the test set was found to be missing the full review text field. For this use case, which is the most effective course of action to address test data samples with missing features. Justify**

**Ans:** In this use case, the most effective course of action to address test data samples with missing features would likely be to impute the missing values.

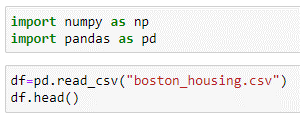
Imputing missing values involves filling in missing values with estimates based on the other values in the dataset. This can help to ensure that all observations in the dataset are included in the analysis and can help to reduce bias.

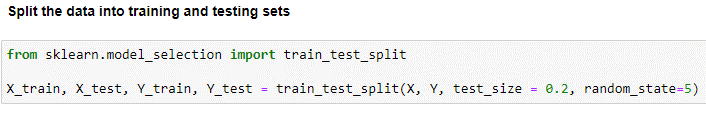
There are several options for imputing missing values, including mean imputation, median imputation, and mode imputation. Mean imputation involves filling in missing values with the mean value of the feature, while median imputation involves filling in missing values with the median value of the feature. Mode imputation involves filling the missing values using frequently occurring values.

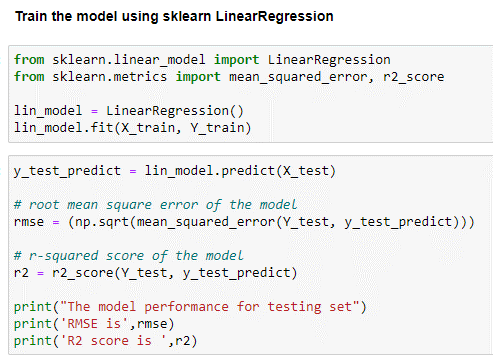
**But in this case, a full review summary usually contains the most descriptive phrases of the entire review and is a valid replacement for the missing full review text field.**

**WEEK 6**

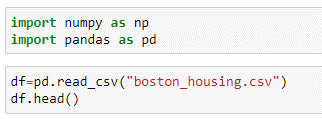
1. **Assume a Boston housing dataset having two columns built\_up\_area (independent variable) and rent (dependent variable). Build linear regression model**

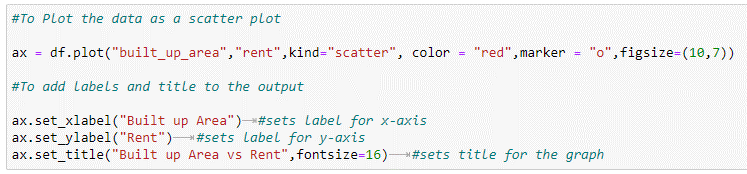






1. **Assume a Boston housing dataset having two columns built\_up\_area (independent variable) and rent (dependent variable).** 
   1. **Import libraries**
   2. **Read data**
   3. **Write scatter plot to compare to show relationship between area and price**





1. **Assume that you are given a train data set having 1000 columns and 1 million rows the data set in based on a classification problem. your manager has asked you to reduce the dimension of this data so that model computations time can be reduced, your machine has memory constraints what would you do? (you are free to make practical assumptions).**

Processing a high dimensional data on a limited memory machine is a strenuous task. Following are the methods you can use to tackle such situation:

1. Since we have lower RAM, we should close all other applications in our machine, including the web browser, so that most of the memory can be put to use.
2. We can randomly sample the data set. This means, we can create a smaller data set, let’s say, having 1000 variables and 300000 rows and do the computations.
3. To reduce dimensionality, we can separate the numerical and categorical variables and remove the correlated variables. For numerical variables, we’ll use correlation. For categorical variables, we’ll use chi-square test.
4. Also, we can use [PCA](https://www.analyticsvidhya.com/blog/2016/03/practical-guide-principal-component-analysis-python/) and pick the components which can explain the maximum variance in the data set.
5. Using online learning algorithms like Vowpal Wabbit (available in Python) is a possible option.
6. Building a linear model using Stochastic Gradient Descent is also helpful.
7. We can also apply our business understanding to estimate which all predictors can impact the response variable. But, this is an intuitive approach, failing to identify useful predictors might result in significant loss of information.

Example: **Using PCA:**

INPUT

fromsklearn.datasets importload\_digits

digits = load\_digits()

digits.data.shape

OUT PUT

(1797, 64)

The data consists of 8×8 pixel images, which means that they are 64-dimensional. To gain some understanding of the relationships between these points, we can use PCA to project them to lower dimensions, like 2-D:

fromsklearn.decomposition import PCA

pca = PCA(2) *# project from 64 to 2 dimensions*

projected = pca.fit\_transform(digits.data)

print(digits.data.shape)

print(projected.shape)

OUT PUT

(1797, 64)

(1797, 2)

1. **For Breast cancer dataset build a machine learning model to predict or identify it and to perform the following operations**

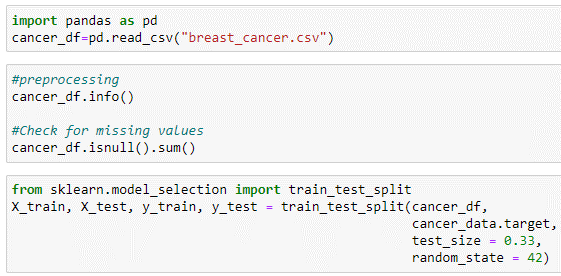
**i) import libraries**

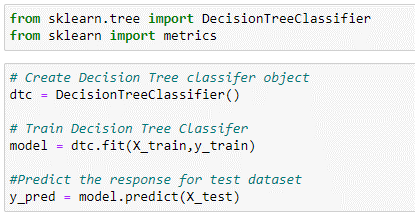
**ii) Perform pre-processing**

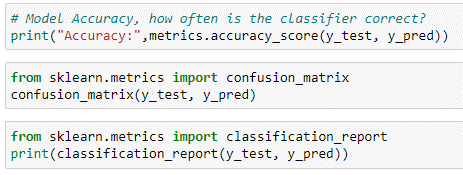
**iii) Split the data set**

**iv) Find the accuracy**

1. **Data prediction**







1. **Assume employee data with two column years of experience (independent variable) and salary (dependent variable).**
2. **import libraries**

**ii) Perform pre-processing**

**iii) Split the data set**

**iv) Find the accuracy**

1. **Data prediction**
2. **Build and test model**
3. import numpy as np

import pandas as pd

import matplotlib.pyplot as plt

import seaborn as sns

1. df=pd.read\_csv(“Salary\_Data.csv")

df

x=df[["Salary"]].values

y=df["YearsExperience"].values

y=y.reshape(len(y),1)

df.isna().sum()

sns.boxplot(data=df['YearsExperience'])

sns.boxplot(data=df['Salary'])

1. from sklearn.model\_selection import train\_test\_split

x\_train,x\_test,y\_train,y\_test=train\_test\_split(x,y, test\_size=0.33,random\_state=42)

1. from sklearn.linear\_model import LinearRegression

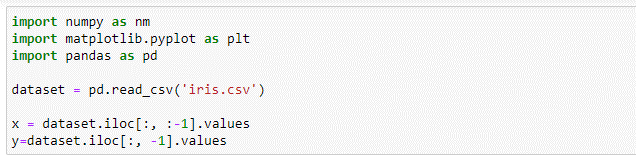
regressor=LinearRegression()

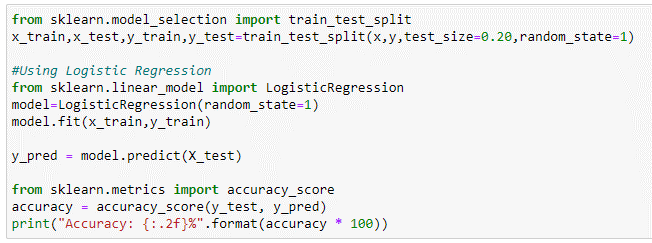
regressor.fit(x\_train,y\_train)

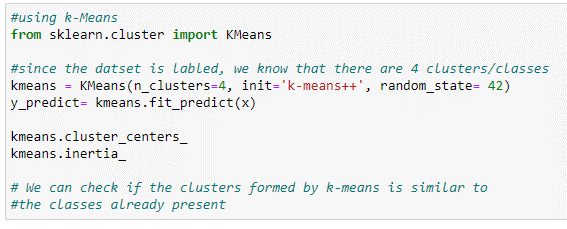
1. y\_pred=regressor.predict(x\_test)
2. train\_accuracy=regressor.score(x\_train,y\_train)

test\_accuracy = regressor.score(x\_test,y\_test)

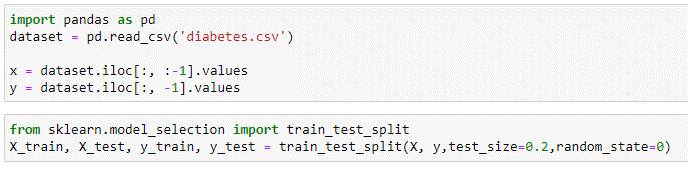
1. Perform comparative analysis and find model accuracy by splitting data for training and testing on iris dataset using K-means clustering and logistic regression.



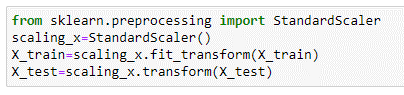




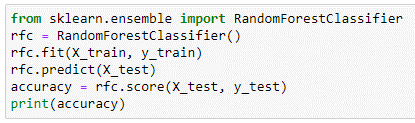
1. **Build a random forest model on diabetic’s dataset and check accuracy.**



Scaling the values



Building the Random Forest classifier and finding accuracy



1. **Illustrate meaning of supervised learning and its types of classification**

In [**SUPERVISED LEARNING**](https://builtin.com/data-science/supervised-learning-python)**,** algorithms learn from labeled data. After understanding the data, the algorithm determines which label should be given to new data by associating patterns to the unlabelled new data.

Learning under supervision directly translates to being under guidance and learning from an entity that is in charge of providing feedback through this process. When training a machine, supervised learning refers to a category of methods in which we teach or train a machine learning algorithm using data, while guiding the algorithm model with labels associated with the data. However, it is essential for [data scientists](https://www.springboard.com/blog/data-science/what-does-a-data-scientist-do/) and machine learning engineers to understand algorithm models and which ones should be applied in particular circumstances.

Supervised learning can be divided into two categories: classification and regression.

Classification is a technique for determining which class the dependent belongs to based on one or more independent variables.

**CLASSIFIER**

A classifier is a type of machine learning algorithm that assigns a label to a data input. Classifier algorithms use labeled data and statistical methods to produce predictions about data input classifications.

### Types of classifications

### 1.logistic regression

Logistic regression is kind of like linear regression, but is used when the dependent variable is not a number but something else (e.g., a "yes/no" response). It's called regression but performs classification based on the regression and it classifies the dependent variable into either of the classes.

Logistic regression is used for prediction of output which is binary.

### 2. K-nearest neighbors (k-nn)

K-NN algorithm is one of the simplest classification algorithms and it is used to identify the data points that are separated into several classes to predict the classification of a new sample point.

K-NN works well with a small number of input variables (p), but struggles when the number of inputs is very large.

### 3. Support vector machine (svm)

### Support vector is used for both regression and classification. It is based on the concept of decision planes that define decision boundaries. A decision plane (hyperplane) is one that separates between a set of objects having different class memberships. 4. Naive bayes

The naive Bayes classifier is based on Bayes’ theorem with the independence assumptions between predictors. Even if these features depend on each other, or upon the existence of the other features, all of these properties independently. Thus, the name naive Bayes.

Based on naive Bayes, Gaussian naive Bayes is used for classification based on the binomial (normal) distribution of data.

### 5. Decision tree classification

Decision tree builds classification or regression models in the form of a tree structure. It breaks down a dataset into smaller and smaller subsets while at the same time an associated decision tree is incrementally developed. Entropy and information gain are used to construct a decision tree.

1. **Compare overfitting with underfitting**

**Reasons for Underfitting:**

1. High bias and low variance

2. The size of the training dataset used is not enough.

3. The model is too simple.

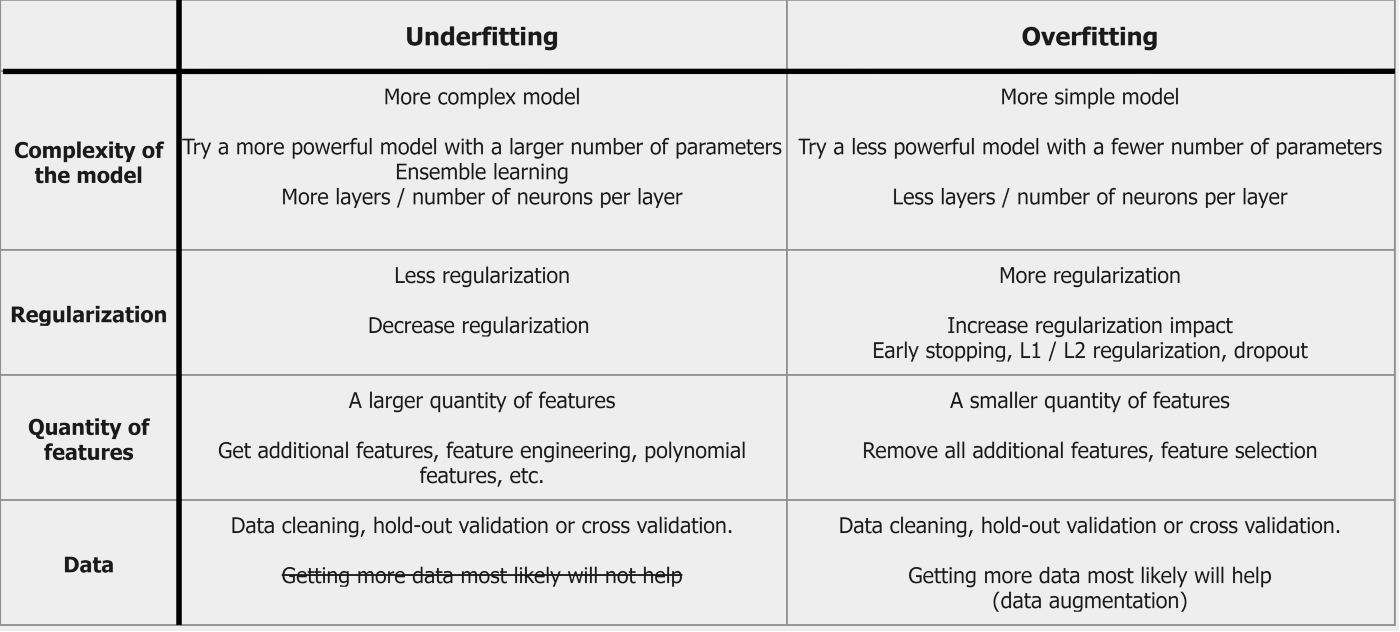
4. Training data is not cleaned and also contains noise in it.

**Reasons for Overfitting are as follows:**

1. High variance and low bias

2. The model is too complex

3. The size of the training data

****

1. **Do comparative analysis by splitting data for training and testing on iris dataset**
2. **Discuss different techniques of cross validation**

## **Types of Cross Validation**

There are different types of cross validation methods, and they could be classified into two broad categories – Non-exhaustive and Exhaustive Methods. We’re going to look at a few examples from both the categories.

## **1. Non-exhaustive Methods**

Non-exhaustive cross validation methods, as the name suggests do not compute all ways of splitting the original data.

## **a. Holdout method**

* This is a quite basic and simple approach in which we divide our entire dataset into two parts viz- training data and testing data.
* As the name, we train the model on training data and then evaluate on the testing set.
* Usually, the size of training data is set more than twice that of testing data, so the data is split in the ratio of 70:30 or 80:20.
* In this approach, the data is first shuffled randomly before splitting.

## **b. K fold cross validation**

* [K-fold cross validation](https://www.mygreatlearning.com/academy/learn-for-free/courses/k-fold-cross-validation/?gl_blog_id=17418) is one way to improve the holdout method.
* This method guarantees that the score of our model does not depend on the way we picked the train and test set.
* The data set is divided into k number of subsets and the holdout method is repeated k number of times. Let us go through this in steps:

1. Randomly split your entire dataset into k number of folds (subsets)
2. For each fold in your dataset, build your model on k – 1 folds of the dataset. Then, test the model to check the effectiveness for kth fold
3. Repeat this until each of the k-folds has served as the test set
4. The average of your k recorded accuracy is called the cross-validation accuracy and will serve as your performance metric for the model.

## **c. Stratified K Fold Cross Validation**

* Using K Fold on a classification problem can be tricky.
* Since we are randomly shuffling the data and then dividing it into folds, chances are we may get highly imbalanced folds which may cause our training to be biased.
* Stratification is the process of rearranging the data so as to ensure that each fold is a good representative of the whole.
* For example, in a binary classification problem where each class comprises of 50% of the data, it is best to arrange the data such that in every fold, each class comprises of about half the instances.

## **2. Exhaustive Methods**

Exhaustive cross validation methods and test on all possible ways to divide the original sample into a training and a validation set.

## **a. Leave-P-Out cross validation**

* When using this exhaustive method, we take p number of points out from the total number of data points in the dataset(say n).
* While training the model we train it on these (n – p) data points and test the model on p data points.
* We repeat this process for all the possible combinations of p from the original dataset. Then to get the final accuracy, we average the accuracies from all these iterations.
* This is an exhaustive method as we train the model on every possible combination of data points.

## **b. Leave-one-out cross validation**

* This is a simple variation of Leave-P-Out cross validation and the value of p is set as one.
* This makes the method much less exhaustive as now for n data points and p = 1, we have n number of combinations.

1. **Demonstrate Simple Linear Regression considering a dataset that has two variables: salary (dependent variable) and experience (Independent variable).**
2. **Build a machine learning model and find the salary if years of experience is 3, test score and interview score is 18.**

|  |  |  |  |
| --- | --- | --- | --- |
| Exp | Test marks (20) | Interview marks (20) | Salary (Rs) |
| 2 | 20 | 15 | 15,000 |
| 3 | 15 | 13 | 20,000 |
| NAN | 12 | 16 | 10,000 |
| 5 | 19 | 10 | 50,000 |

**14. In a hospital doctor found that many patients arrived for Breast cancer test, Help the doctor to build a machine learning model to predict or identify it and to perform the following operations**

**i) import libraries**

**ii) perform pre-processing**

**iii) split the data set**

**iv) Build the accuracy**

**v) Data prediction**

1. **Import the required libraries**

import numpy as np

import pandas as pd

from sklearn.model\_selection import train\_test\_split

import sklearn.metrics as sm

from sklearn.tree import DecisionTreeClassifier

1. **pre-processing**

breast.head()

breast.info()

breast.describe()

breast.isnull().sum()

breast.fillna(1)

1. **splitting the data set**

x=breast.drop([“date”],axis=1)

y=breast.date.values

x\_train,x\_test,y\_train,y\_test=train\_test\_split(x,y,test\_size=0.3,random\_state=42)

**iv ) Data prediction**

from sklearn.linear\_model import LinearRegression

y\_preds=reg.predict(x\_test)

import pandas as pd

compare=pd.DataFrame({"Actual":y\_test,"Predicted":y\_preds})

print(compare)

* 1. **Find the accuracy**

import sklearn.metrics as sm

accuracy=sm.r2\_score(y\_test,y\_preds)

print(“Accuracy of Decision is {:,2f} % ”.format(accuracy\*100))

**WEEK 7**

**1. Explain the evaluation matrix for classification as follows.**

**1) confusion matrix**

**2) Accuracy**

**3) F1 – score**

**4) AUC\_ROC**

**5) Precision and Recall**

**confusion matrix**

A Confusion matrix is an N x N matrix used for evaluating the performance of a classification model, where N is the number of target classes. The matrix compares the actual target values with those predicted by the machine learning model. This gives us a holistic view of how well our classification model is performing and what kinds of errors it is making.



* Rows in confusion matrix refer to the occurrence or non-occurrence of an event
* Columns refer to the model’s predictions
* Confusion matrices make it easy to tell how accurate a model's outcomes are likely to be by exposing when the model is repeatedly confusing two classes.
* They evaluate the performance of a classification model, allowing business users to determine which data their model may be unable to classify correctly.

from sklearn.metrics import confusion\_matrix

confusion\_matrix(Y\_test,y\_preds)

**Accuracy**:-

The accuracy of a machine learning classification algorithm is **one way to measure how often the algorithm classifies a data point correctly**. Accuracy is the number of correctly predicted data points out of all the data points.



Importance of Accuracy :

If the data is not qualitatively rich, it will result in inaccurate insights, leading to unpredictable decisions in human context models. AI model algorithms are classified based on the extent to which they can be interpreted or on the degree of accuracy.

Accuracy measures the model’s ability to correctly classify all instance

from sklearn.metrics import accuracy\_score

accuracy\_score(y\_test,y\_pred)

**F1 – score :**

The F1-score **combines the precision and recall of a classifier into a single metric by taking their harmonic mean**. It is primarily used to compare the performance of two classifiers. Suppose that classifier A has a higher recall, and classifier B has higher precision.

**An F1 score is considered perfect when it's 1** , while the model is a total failure when it's 0 . Remember: All models are wrong, but some are useful. That is, all models will generate some false negatives, some false positives, and possibly both.

from sklearn.metrics import f1\_score

f1\_score(y\_true, y\_pred)

**AUC\_ROC:**

It is used to evaluate the performance of classification model

**AUC = Area Under Curve**

**AUC** stands for "Area under the ROC Curve." That is, AUC measures the entire two-dimensional area underneath the entire ROC curve (think integral calculus) from (0,0) to (1,1).

**ROC = Reciever Operator curve**

ROC is probability curve and AUC represents the degree or measure of separability.It will tell how much the model is capable of distinguishing b/w classes

**ROC curve uses:**

The ROC curve is used to assess the overall diagnostic performance of a test and to compare the performance of two or more diagnostic tests.

**Advantages of ROC:**

ROC curves are frequently used to show in a graphical way the connection/trade-off between clinical sensitivity and specificity for every possible cut-off for a test or a combination of tests. In addition the area under the ROC curve gives an idea about the benefit of using the test(s) in question.

**Precision and Recall:**

Precision tells us how many of the correctly predicted cases actually turned out to be positive.

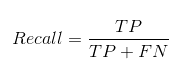
Here’s how to calculate Precision:

[Confusion Matrix Precision](https://cdn.analyticsvidhya.com/wp-content/uploads/2020/04/Confusion-matrix_Precision.png)

This would determine whether our model is reliable or not.

Recall tells us how many of the actual positive cases we were able to predict correctly with our model.

And here’s how we can calculate Recall:

[](https://cdn.analyticsvidhya.com/wp-content/uploads/2020/04/Confusion-matrix_Recall.png)

Precision is a useful metric in cases where False Positive is a higher concern than False Negatives.

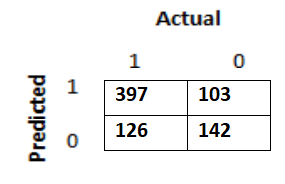
Precision is important in music or video recommendation systems, e-commerce websites, etc. Wrong results could lead to customer churn and be harmful to the business.

Recall is a useful metric in cases where False Negative trumps False Positive.

Recall is important in medical cases where it doesn’t matter whether we raise a false alarm but the actual positive cases should not go undetected

**2. Build a data set and predict the heart disease based on BP, Sugar, Age, Gender and Cholesterol by using relevant operations**

**3. The confusion matrix for a model is as shown below. Evaluate accuracy, precision, recall, Specificity and F1-Score, AUC-ROC**



4. Given a dataset perform comparative analysis using decision tree and SVM algorithm and check accuracy

5. A machine learning model was built to classify people based on whether they speak English or Hindi. The confusion matrix for the model is as shown below. Compute accuracy, precision, recall, F1-Score and Specificity.



**WEEK8**

1. **Define clustering and compare various clustering techniques.**

**Clustering** is basically defined as division of data into groups of similar objects. Each group called a cluster, consists of objects that are similar between themselves and dissimilar compared of other groups. Let’s compare among different type of clusters.

The algorithms under discussion are*:*k-means algorithm, hierarchical clustering algorithm, self-organizing maps algorithm, and expectation maximization clustering algorithm.



1. **Describe the advanced ensemble techniques**

## Advanced Ensemble techniques

### 1. Stacking

* Stacking is an [ensemble learning](https://courses.analyticsvidhya.com/courses/ensemble-learning-and-ensemble-learning-techniques?utm_source=blog&utm_medium=comprehensive-guide-for-ensemble-models) technique that uses predictions from multiple models (for example decision tree, knn or svm) to build a new model.
* This model is used for making predictions on the test set.

1. The train set is split into 10 parts.
2. A base model (suppose a decision tree) is fitted on 9 parts and predictions are made for the 10th part. This is done for each part of the train set.
3. The base model (in this case, decision tree) is then fitted on the whole train dataset.
4. Using this model, predictions are made on the test set.
5. Steps 2 to 4 are repeated for another base model (say knn) resulting in another set of predictions for the train set and test set.
6. The predictions from the train set are used as features to build a new model.
7. This model is used to make final predictions on the test prediction set.

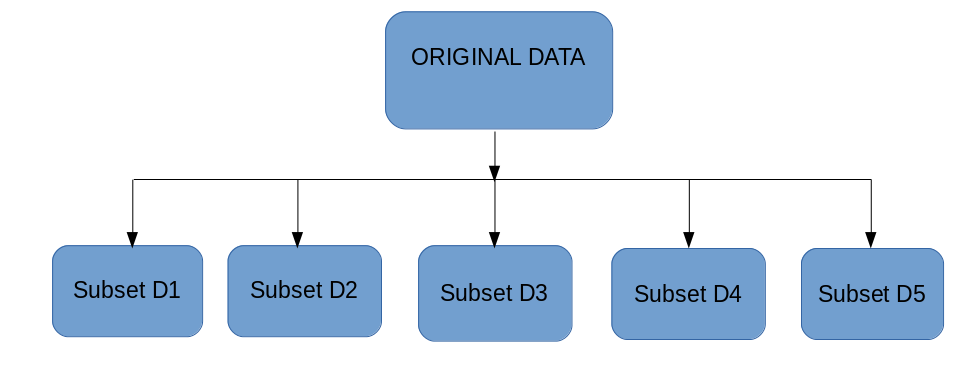
### 2. Blending

* Blending follows the same approach as stacking but uses only a holdout (validation) set from the train set to make predictions.
* In other words, unlike stacking, the predictions are made on the holdout set only.
* The holdout set and the predictions are used to build a model which is run on the test set.
* Here is a detailed explanation of the blending process:

1. The train set is split into training and validation sets.
2. Model(s) are fitted on the training set.
3. The predictions are made on the validation set and the test set.
4. The validation set and its predictions are used as features to build a new model.
5. This model is used to make final predictions on the test and meta-features.

### 3. Bagging

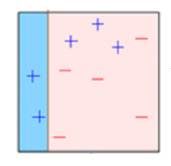
* The idea behind bagging is combining the results of multiple models (for instance, all decision trees) to get a generalized result.
* If you create all the models on the same set of data and combine it, will it be useful? There is a high chance that these models will give the same result since they are getting the same input.
* So how can we solve this problem? One of the techniques is bootstrapping.
* Bootstrapping is a sampling technique in which we create subsets of observations from the original dataset, **with replacement**.
* Bagging (or Bootstrap Aggregating) technique uses these subsets (bags) to get a fair idea of the distribution (complete set). The size of subsets created for bagging may be less than the original set.

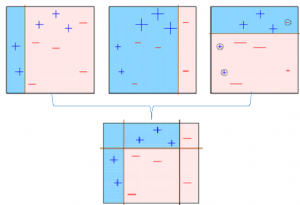


1. Multiple subsets are created from the original dataset, selecting observations with replacement.
2. A base model (weak model) is created on each of these subsets.
3. The models run in parallel and are independent of each other.
4. The final predictions are determined by combining the predictions from all the models.

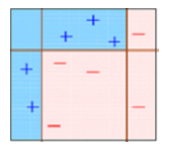
### 4. Boosting

* If a data point is incorrectly predicted by the first model, and then the next (probably all models), will combining the predictions provide better results? Such situations are taken care of by boosting.
* Boosting is a sequential process, where each subsequent model attempts to correct the errors of the previous model. The succeeding models are dependent on the previous model. Let’s understand the way boosting works in the below steps.

1. A subset is created from the original dataset.
2. Initially, all data points are given equal weights.
3. A base model is created on this subset.
4. This model is used to make predictions on the whole dataset.  
   
5. Errors are calculated using the actual values and predicted values.
6. The observations which are incorrectly predicted, are given higher weights. (Here, the three misclassified blue-plus points will be given higher weights)
7. Another model is created and predictions are made on the dataset. (This model tries to correct the errors from the previous model)
8. Similarly, multiple models are created, each correcting the errors of the previous model.
9. The final model (strong learner) is the weighted mean of all the models (weak learners).



1. Thus, the boosting algorithm combines a number of weak learners to form a strong learner. The individual models would not perform well on the entire dataset, but they work well for some part of the dataset. Thus, each model actually boosts the performance of the ensemble.



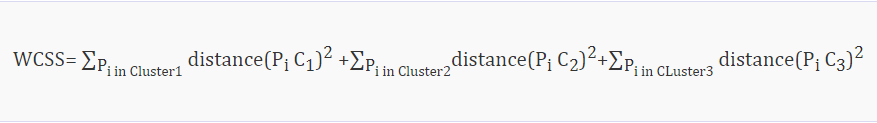
1. **How to Choose the Right Number of Clusters in k-means clustering? Explain any one method**

**Elbow method**

The performance of the K-means clustering algorithm depends upon highly efficient clusters that it forms. But choosing the optimal number of clusters is a big task. The most appropriate method to find the number of clusters or value of K is Elbow method.

**Elbow Method**

The Elbow method is one of the most popular ways to find the optimal number of clusters. This method uses the concept of WCSS value. **WCSS** stands for **Within Cluster Sum of Squares**, which defines the total variations within a cluster. The formula to calculate the value of WCSS (for 3 clusters) is given below:



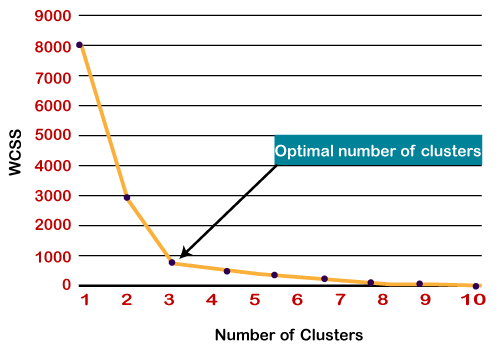
**In the above formula of WCSS,**

* ∑Pi in Cluster1 distance(Pi C1)2: It is the sum of the square of the distances between each data point and its centroid within a cluster1 and the same for the other two terms.
* To measure the distance between data points and centroid, we can use any method such as Euclidean distance or Manhattan distance.

To find the optimal value of clusters, the elbow method follows the below steps:

* It executes the K-means clustering on a given dataset for different K values (ranges from 1-10).
* For each value of K, calculates the WCSS value.
* Plots a curve between calculated WCSS values and the number of clusters K.
* The sharp point of bend or a point of the plot looks like an arm, then that point is considered as the best value of K.

Since the graph shows the sharp bend, which looks like an elbow, hence it is known as the elbow method. The graph for the elbow method looks like the below image:



1. **Cluster the following eight points (with (x, y) representing locations) into three clusters:**

**A1(2, 10), A2(2, 5), A3(8, 4), A4(5, 8), A5(7, 5), A6(6, 4), A7(1, 2), A8(4, 9)**

**Initial cluster centres are: A1(2, 10), A4(5, 8) and A7(1, 2).**

**The distance function between two points a = (x1, y1) and b = (x2, y2) is defined as**

**Ρ(a, b) = |x2 – x1| + |y2 – y1|.**

**Use K-Means Algorithm to find the three cluster centres after the first iteration**

**Solution-**

We follow the above discussed K-Means Clustering Algorithm-

**Iteration-01:**

* We calculate the distance of each point from each of the center of the three clusters.
* The distance is calculated by using the given distance function.

The following illustration shows the calculation of distance between point A1(2, 10) and each of the center of the three clusters-

**Calculating Distance Between A1(2, 10) and C1(2, 10)-**

Ρ(A1, C1)

= |x2 – x1| + |y2 – y1|

= |2 – 2| + |10 – 10|

= 0

In the similar manner, we calculate the distance of other points from each of the center of the three clusters.

Next,

* We draw a table showing all the results.
* Using the table, we decide which point belongs to which cluster.
* The given point belongs to that cluster whose center is nearest to it.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Given Points** | **Distance from center (2, 10) of Cluster-01** | **Distance from center (5, 8) of Cluster-02** | **Distance from center (1, 2) of Cluster-03** | **Point belongs to Cluster** |
| A1(2, 10) | 0 | 5 | 9 | C1 |
| A2(2, 5) | 5 | 6 | 4 | C3 |
| A3(8, 4) | 12 | 7 | 9 | C2 |
| A4(5, 8) | 5 | 0 | 10 | C2 |
| A5(7, 5) | 10 | 5 | 9 | C2 |
| A6(6, 4) | 10 | 5 | 7 | C2 |
| A7(1, 2) | 9 | 10 | 0 | C3 |
| A8(4, 9) | 3 | 2 | 10 | C2 |

From here, New clusters are-

**Cluster-01**

First cluster contains points-

* A1(2, 10)

**Cluster-02:**

Second cluster contains points-

* A3(8, 4)
* A4(5, 8)
* A5(7, 5)
* A6(6, 4)
* A8(4, 9)

**Cluster-03:**

Third cluster contains points-

* A2(2, 5)
* A7(1, 2)

Now,

* We re-compute the new cluster clusters.
* The new cluster center is computed by taking mean of all the points contained in that cluster.

**For Cluster-01:**

* We have only one point A1(2, 10) in Cluster-01.
* So, cluster center remains the same.

**For Cluster-02:**

Center of Cluster-02

= ((8 + 5 + 7 + 6 + 4)/5, (4 + 8 + 5 + 4 + 9)/5)

= (6, 6)

**For Cluster-03:**

Center of Cluster-03

= ((2 + 1)/2, (5 + 2)/2)

= (1.5, 3.5)

1. **Compare classification algorithms with clustering algorithm**

**Classification** and **clustering** are techniques used in data mining to analyze collected data. Classification is used to label data, while clustering is used to group similar data instances together.

Classification vs. Clustering

### Classification

* The number of classes is known.
* Training data (collection of labeled instances) is required.
* Based on the training data, the classification model is used to classify future instances into already defined classes.
* Popular algorithms for classification include Naive Bayes Classifier, Decision Trees, and Random Forests.

### Clustering

* The number of classes is unknown.
* No training data is required.
* Clustering is used to make sense of existing data.
* Popular algorithms used for clustering include K-Means, Mean-Shift Clustering, and Density-Based Spatial Clustering of Applications with Noise.

1. **K-means clustering with Euclidean distance suffer from the curse of dimensionality. Is the statement true and why?**

Yes, the statement that k-means clustering with Euclidean distance suffers from the curse of dimensionality is true.

The curse of dimensionality refers to the phenomenon that as the number of dimensions (features) in a dataset increases, the distance between points in the dataset tends to become increasingly large, making it difficult to identify patterns or clusters in the data. This can be a problem for k-means clustering, which relies on Euclidean distance to measure the similarity between data points.

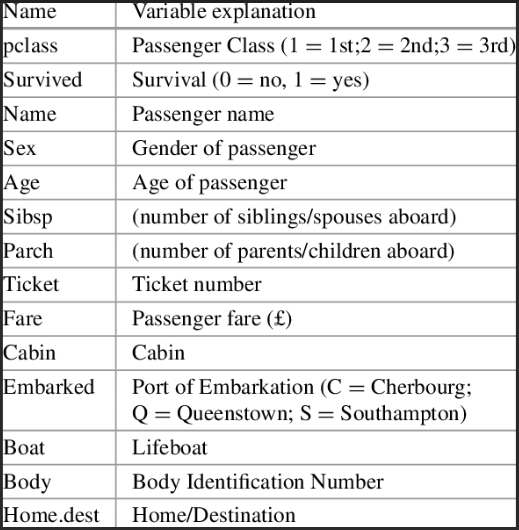
The Euclidean distance between two points is calculated as the square root of the sum of the squares of the differences between the values of each feature. As the number of dimensions increases, the Euclidean distance between points becomes increasingly large, making it more difficult to identify clusters in the data.

To illustrate this, consider a simple example with two dimensions (features). Suppose we have two points in a two-dimensional space, with values (1, 1) and (2, 2). The Euclidean distance between these two points is sqrt((2-1)^2 + (2-1)^2) = sqrt(1 + 1) = sqrt(2) = 1.4.

Now suppose we add a third dimension to the space. The two points now have values (1, 1, 1) and (2, 2, 2). The Euclidean distance between these two points is sqrt((2-1)^2 + (2-1)^2 + (2-1)^2) = sqrt(3) = 1.7.

As we can see, the Euclidean distance between the points increases as the number of dimensions increases, making it more difficult to identify clusters in the data. This is why k-means clustering with Euclidean distance can suffer from the curse of dimensionality, especially when working with high-dimensional datasets.

1. The sinking of the Titanic is one of the most infamous shipwrecks in history. You are asked to build a machine learning model to predict whether a passenger survived or not. Describe each step you will follow to build this model. Description of dataset titanic.csv is as below



1. K-means clustering with Euclidean distance suffer from the curse of dimensionality. Is the statement true and why?
2. Compare Bagging and Boosting.

**WEEK9**

1. Discuss importance of dimensionality reduction in machine learning
2. Explain dimensionality reduction using PCA

**WEEK 10**

1. Discuss activation functions in Neural Network
2. Explain neural network architecture
3. Differentiate between forward propagation and Back propagation

**WEEK 11**

1. **For the given sentence “Machine Learning is best platform. To group over selves”. Visualize the data for the better understanding and perform the following operations.**
2. **Word tokenization**
3. **Sentence tokenization**
4. **Dropping Stop word**
5. **Stemming**
6. **Lemmatization**

**Ans:**

**i) Word tokenization**

INPUT:

import nltk

from nltk import \*

from nltk.tokenize import word\_tokenize

nltk.download('wordnet')

sent = "Machine Learning is best platform. To group overselves"

print(nltk.word\_tokenize(sent))

OUTPUT:

['Machine', 'Learning', 'is', 'best', 'platform', '.', 'To', 'group', 'overselves']

1. **Sentence tokenization**

INPUT

from nltk.tokenize import sent\_tokenize

sent = "Machine Learning is best platform. To group overselves"

sent\_tokenize(sent)

1. **Dropping Stopword**

INPUT:

from nltk.corpus import stopwords

wo\_stop\_words = []

stop\_words\_set = set(stopwords.words("english"))

for i in vocab\_wo\_punct:

if i not in stop\_words\_set:

wo\_stop\_words.append(i)

print(wo\_stop\_words)

OUTPUT:

['Learning', 'Machine', 'To', 'best', 'group', 'overselves', 'platform']

1. **Stemming**

INPUT:

stemmed\_vocab=[]

stemObj = SnowballStemmer("english")

for i in vocab\_wo\_punct:

stemmed\_vocab.append(stemObj.stem(i))

print(stemmed\_vocab)

OUT PUT:

['learn', 'machin', 'to', 'best', 'group', 'is', 'overselv', 'platform']

1. **Lemmatization**

INPUT:

from nltk.stem.wordnet import WordNetLemmatizer

#nltk.download('wordnet')

#nltk.download('omw-1.4')

lemmaObj = WordNetLemmatizer()

lemmaObj.lemmatize("grouping",pos='v')

OUT PUT:

'group'

1. **Companies monitor their call center agents’ live phone interactions or chat sessions with customers in real-time. Call duration with speech recognition automatically detects customer emotions. Companies can better understand how customer satisfaction varies by product and call center services. Using this scenario Explain the applications and working of sentiment analysis in business.**

**Ans:**

**Definition**: Sentiment analysis, also referred to as opinion mining, is an approach to natural language processing (NLP) that identifies the emotional tone behind a body of text

**OR**

Customer sentiment analysis refers to the automated process of discovering and measuring how customers feel about your product, brand, or service.

**AI in call center**

AI-powered customer service allows users to understand and incorporate the tools of customer sentiment detection. **Through deep machine learning, it automatically detects emotions from natural language processing, voice quality, and speech recognition.** **It then trains models to detect those emotions**.

**Applications of AI in call center**

**Predictive Call Routing**

This technology relies on customer behavior profiles to give [AI technology](https://www.youtube.com/playlist?list=PLlw9qxNtFom2JqbUWKTALkvQ8TYl2Z5F0) a comprehensive understanding of the customer journey and customer personas. Meaning customer service (and the customer experience overall) can be hyper-personalized to each customer.

### Interactive Voice Response (IVR)

Interactive voice response (IVR) is the AI that most of us have interacted with during our customer service experiences. This is when you answer recorded questions such as what language you speak, your name, account number, etc. It's true that many of us dislike this type of AI because we've had calls where we had to repeat the information.

**Conversational AI**

Conversational AI is mostly known as [chatbots](https://blog.hubspot.com/service/customer-service-chatbots) nowadays. This is when a call center will have an online chat option that is powered by AI. And it's a necessary form of customer service since [85% of consumers](https://pr.liveperson.com/index.php?s=43&item=635) worldwide would like to message with brands.

### Emotional Intelligence AI

Another form of artificial intelligence in call centers is emotional intelligence AI that can track customer sentiment during a phone call. This type of AI is trained in different languages and cultural contexts, so it can be used in countries with different linguistic and cultural styles. It can analyze the tone of voice and cadence of language to try to detect the caller's mood.

### AI-Powered Recommendations

AI tools can give recommendations to a customer support rep during a call. This technology also uses sentiment analysis to understand what a customer is trying to accomplish. It can then give recommendations for the best solutions to the support rep.

This helps reduce call times and provides a personalized, positive customer experience. The technology can analyze how many times a customer has called or referenced canceling their account, then it can give that customer a customer risk score so agents are aware during the phone call

### Call Analytics

One of the main ways that AI is used in call centers is to provide in-depth analytics on call times, first resolution, and more. These technologies can spot trends and have access to customer data that will provide insight on whether customers are having a positive or negative experience.

Since AI measures customer sentiment, tone, and personality, it can provide more well-rounded analytics than a human customer support manager could.

1. You are working on an NLP model. So, you are dealing with words and sentences, not numbers. Your problem is to categorize these words and make sense of them. Your manager told you that you have to use embeddings. Explain Count Vector and TF-IDF Vector
2. List and explain tools used to deploy the AI model using google cloud platform
3. Demonstrate stemming and Lemmatization concepts with suitable examples
4. Write different algorithms used in sentimental analysis with neat diagram.

**WEEK 12**

1. **What are ethics in AI and Why ethical practices should be followed while developing solutions using AI?**

**Ans:** AI ethics: It is a system of moral principles and techniques intended to inform the development and responsible use of artificial intelligence technology

**Ethics in AI**

* Safety
* Failure transparence
* Judicial
* Responsibility use of AI technology
* Value alignment
* Human values
* Personal Privacy
* Liberty Privacy
* Share benefit
* Human control
* Alarm Race
* Examples of AI in real world
* Manufacturing robots.
* Self-driving cars.
* Smart assistants.
* Healthcare management.
* Automated financial investing.
* Virtual travel booking agent.
* Social media monitoring.
* Marketing chatbots.

**Ethical practices**

* Ethical practices are very important while developing solutions using AI because the solution is made to solve the problem or reduce the complexity of the problem.
* If ethical practices are not followed properly then the solution might create more problems and that leads to more problems.
* Even unethical practices increase risk and damage to the surroundings. Which is not a good cause for the problem solution.
* An AI is considered an ethical AI when it is built on an ethical framework, where the end goal of the software is not just monetary gains, but to make society better.
* Responsible AI is the practice of AI development under the principles of fairness, transparency and explain ability, human-centeredness, and privacy and security.

1. Write Ethical challenges in Artificial Intelligence
2. Create container and build docker image using docker file
3. **Imagine a scenario where the developer’s code is only working on his machine and he has to set it up in the machine of the testing team for every build. How docker containers solve this problem.**

Docker is a tool designed to make it easier for developers to develop, ship, and run applications by using containers. Containers allow devs to package an application with all of its requirements and configurations, such as libraries and other dependencies and deploy it as a single package. By doing so, developers can rest easy knowing that the software will work properly on other systems besides the one they used for writing the code.

Docker saves you a lot of time by doing most of the repetitive tasks automatically, which not only conserves time but also resources. Even deployment is a piece of cake because of light-weight Docker images which are easy to handle. Docker is mostly favored for Continuous Integration and Continuous Delivery (CI/CD) workflows.

Docker is a containerization platform that is free and open source. It allows developers to package programs into containers, which are standardized executable components that combine application source code with the OS libraries and dependencies needed to run that code in any environment

Containerization enables "write once, run anywhere" programs. In terms of the development process and vendor compatibility, portability is critical. Containers are widely referred to as "lightweight," implying that they share the machine's operating system kernel and eliminate the overhead of associating an operating system with each application.

Containers have a lower capacity and require less start-up time than virtual machines, allowing significantly more containers to run on the same computational power as a single VM. As a result, server efficiencies improve, lowering server and licensing costs.

1. **How will you deploy a trained machine learning model as a predictive service in a production environment? Explain.**

**Ans:**

**Definition of Deployment:**

**Deployment** is the process by which a ML model is moved from an offline environment and integrated into an existing production environment, such as a live application. It is a critical step that must be completed in order for a model to serve its intended purpose and solve the challenges it is designed for.

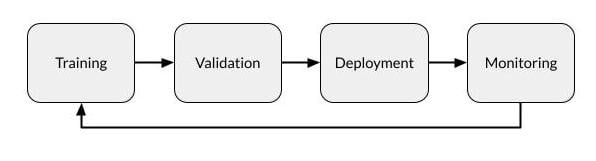
**Deploying the ML models:**

* Develop and create a model in a training environment. To deploy a machine learning application, you first need to build your model. .
* Optimize and test code, then clean and test again. .
* Prepare for container deployment.
* Plan for continuous monitoring and maintenance.

**Working:**

* Most data science projects deploy machine learning models as an on-demand prediction service or in batch prediction mode. Some modern applications deploy embedded models in edge and mobile devices.
* Each model has its own merits. For example, in the batch scenario, optimizations are done to minimize model compute cost. There are fewer dependencies on external data sources and cloud services. The local processing power is sometimes sufficient for computing algorithmically complex models.
* It is also easy to debug an offline model when failures occur or tune hyperparameters since it runs on powerful servers.
* On the other hand, web services can provide cheaper and near real-time predictions. Availability of CPU power is less of an issue if the model runs on a cluster or cloud service. The model can be easily made available to other applications through API calls and so on.
* One of the main benefits of embedded machine learning is that we can customize it to the requirements of a specific device.
* We can easily deploy the model to a device, and its runtime environment cannot be tampered with by an external party. A clear drawback is that the device needs to have enough computing power and storage space.

**Work flow:**



1. Understand the difference between DevOps and MLOps and justify which is best suggested process for developing Machine Learning Models.
2. N-grams are defined as the combination of N keywords together. Consider the given sentence: “AI is simulation of human intelligence by machine. It includes expert system NLP speech recognition and machine vision”

* Generate bi grams for the above sentence
* Generate tri-grams for the above sentence

1. Summarize different strategies of production deployment
2. What are MLOps? brief different stages that are involved in the MLOps lifecycle
3. With a neat diagram explain components of docker.
4. For the following scenarios you are required to build a predictive model. Which machine learning technique/ algorithm can be applied / best suited for stated problems. Justify your recommendation.
   1. Predicting the food delivery time
   2. Predicting whether the transaction is fraudulent
   3. Predicting the credit limit of a credit card applicant
   4. To group similar customers of an online grocery store, based on their purchasing patterns, to offer discounts to its customers.
   5. Predict the probability of a mechanical system breakdown, based on its system vibration and operating temperature
5. For the below given scenarios, suggest best suited cloud deployment model and list the challenges with it.

1. For ,

a. Variable workload

b. Test and Development

2. For,

a. Cloud bursting

b. On demand access

c. Sensitive data

13. How will you deploy a trained machine learning model as a predictive service in a production environment. Explain.