**INTERNSHIP REPORT**

**A Report of Internship**

**at**

**TECHNEX IIT(BHU)**

**Submitted by**

**SAMIR TAMANG**

**Regd.No.: 20781A05H5**

*in partial fulfillment for the award of the degree*

*of*

**BACHELOR OF TECHNOLOGY**

*in*

**COMPUTER SCIENCE ENGINEERING**

****

****

**SRI VENKATESWARA COLLEGE OF ENGINEERING AND TECHNOLOGY(AUTONOMUS)**

(Approved by AICTE, Affiliated to JNTUA)

(Accredited by N.B.A., New Delhi & NAAC. Bangalore)

**R.V.S. Nagar, Chittoor-517127**

**ANDHRA PRADESH**

DECEMBER 2023

**DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING**

**SRI VENKATESWARA COLLEGE OF ENGINEERING &TECHNOLOGY(AUTONOMOUS**)

****

**CERTIFICATE**

This is to certify Certified that this **Report of Internship at**

**“YBI Foundation”**

Being submitted herewith to the

**SRI VENKNATESWARA COLLEGE OF ENGINEERING & TECHNOLOGY(AUTONOMOUS)**

is the bonafide work of

**..SAMIR TAMANG(20781A05H5)..**

who carried out the project work under our guidance and supervision**.**

|  |  |  |
| --- | --- | --- |
| **College internship Coordinator** | **Department Internship Coordinator** | **Head of The Department** |

**INTERNSHIP CERTIFICATION**



**ACKNOWLEDGEMENT**

I would like to express my special thanks of gratitude to my Internship Training guide Mr. Nanda Kumar Agrawal, Head of the Department Dr. P. Jyotheeswari as well as our principal Dr. M. Mohan Babu who gave me the golden opportunity to do this wonderful Internship at “TECHNEX IIT(BHU)”, which provided me an opportunity explore the new horizons.

I sincerely express my gratitude towards the “TECHNEX IIT(BHU)”, for providing this opportunity.

I would like to thank *Mr. Hareram Singh* internship coordinator Department of CSE FOR their support and advices.

Secondly, I would also like to thank my parents and friends who helped me a lot in finalizing this report within the limited time frame.

|  |  |
| --- | --- |
| Date: 08/12/2023 | **SAMIR TAMANG** |
| Place: Chittoor | 20781A05H5 |
|  | IV-CSE’B’ |

**TABLE OF CONTENTS**

|  |  |  |
| --- | --- | --- |
|  | **TITLE** | **Page No.** |
|  | Cover Page | 1 |
|  | Certificate | 2 |
|  | Acknowledgement | 4 |
|  | Table of Content | 5 |
|  |
|  | List of Figures |  |
|  |  |  |
| **1.0** | **INTRODUCTION** | **1** |
|  | 1.1 Sub section | 1 |
|  | 1.2 Sub section | 1 |
| **2.0** | **COMPANY PROFILE** | **2** |
|  | 2.1 Sub section | 2 |
|  | 2.2 Sub section | 2 |
| **3.0** | **DEPARTMENT PROFILE** | **3** |
|  | 3.1 Sub section | 3 |
|  | 3.2 Sub section | 3 |
| **4.0** | **PROJECT WORK** | **4** |
|  | 4.1 Sub section | 4 |
|  | 4.2 Sub section | 4 |
| **5.0** | **CASE STUDY** | **5** |
|  | 5.1 Sub section | 5 |
|  | 5.2 Sub section | 5 |
| **6.0** | **CONCLUSION** | **6** |
|  | **REFERENCES** | **7** |

**LIST OF FIGURES**

|  |  |  |
| --- | --- | --- |
| **Figure No.** | **Title** | **Page No**. |
| Figure 1.1 | Bar chart | 23 |
| Figure 1.1 | Pie chart | 24 |
| Figure 1.1 | Output of code | 37 |

**TABLES OF CONTENT**

INDEX

**S.no CONTENTS Page.no**

1.Introduction to python

1.1 About the Python 10

1.2 Brief overview of data science and machine learning

2.**Data Science**

1. **Data Collection:**
   * Describe the sources of data used in your analysis.
   * Discuss the data collection process, including any challenges or biases. Data cleaning
   * Data transformation
   * Feature scaling
   * Handling missing data
2. **Data Exploration and Preprocessing:**
   * Explore the dataset and provide descriptive statistics.
   * Discuss any missing data, outliers, or data quality issues.
   * Describe the preprocessing steps undertaken to clean and prepare the data

3.**Exploratory Data Analysis (EDA)**:

* This section should include visualizations and statistical summaries of the data to gain insights and identify patterns Histograms
* Box plots
* Scatter plots
* Pair plots
* Heatmaps

4. **Data Visualization:**

* Matplotlib
* Seaborn
* Plotly
* Tableau
* g. gplot2 (R)

5. **Text Analysis**:

* Natural Language Processing (NLP)
* Text mining
* Sentiment analysis
* Topic modelling

3. **Python for machine learning**

3.1 Understanding operator

3.2 Variables and Data Types

3.3 Conditional Statements, For and While Loops

3.4 Functions

3.5 Data Structure

4. **INTRODUCTION**

4.1 Taste of machine learning

4.2 FUTURE OF MACHINE LEARNING

5. **TECHNOLOGY LEARNT**

5.1 Introduction to AI & ML

5.2 Definition of Machine Learning

5.3 Machine Learning Algorithm

6. **TECHINQUES OF MACHINE LEARNING**

6.1 SUPERSIVED LEARNING

6.2 TYPES OF SUPERVISED LEARNING

6.3 FEATURES ENGINEERING

6.4 INTRODUCTION TO DEEP LEARNING

7**.Database and SQL:**

* Relational databases (e.g., SQL Server, MySQL)
* NoSQL databases (e.g., MongoDB, Cassandra)

8. REFERENCES

9.CONCLUSION

**WEEKLY OVERVIEW OF INTERNSHIP ACTIVITES**

|  |  |  |  |
| --- | --- | --- | --- |
| S.NO | Date | Day | Name of the topic |
| 1 | 05/07/23 | Tuesday | Introduction of About the topics (Optional) |
| 2 | 06/07/23 | Wednesday | Background need & importance, introduction & installation, writing first program, I/o in python |
| 3 | 09/07/23 | Sunday | Keywords & Variables Data types - Numbers & Strings Operators in python Hands-on implementation |
| 4 | 10/07/23 | Monday | Indentation & scopes if, else & elif blocks introduction to loops for & While loops Break, continue statements |
| 5 | 13/07/23 | Wednesday | Data Structures |
| 6 | 17/07/23  To  18/07/23 | Sunday& Monday | Functions & Exception Handling |
| 7 | 19/07/23  To  19/07/23 | Thursday | Handling Pattern based problems |
| 8 | 20/07/23  To  20/07/23 | Sunday & Monday | map, filter, reduce, lambda•zip, enumerate, sorted |
| 9 | 21/07/23 | Tuesday | File handling - I/o•Numpy arrays•Case studies |
| 10 | 22/07/23 | Wednesday | About the Data Science and Python on it. |
| 11 | 24/07/23  To  26/07/23 | Friday & Monday | Types of Data science and its process |
| 12 | 28/07/23 | Wednesday | Data Collection like collection, processing, cleaning, transforming etc.  Data Exploration and preprocessing |
| 13 | 30/07/23  To  30/07/23 | Friday | Exploratory Data Analysis (EDA): bar chart, line chart, box, pair chart etc |
| 14 | 01/08/23  To  03/08/23 | Wednesday & Friday | Data Visualization: matplotlib, tableau, seaborn etc  Text Analysis : NLP, text mining etc |
| 15 | 05/08/23  To  06/08/23 | Sunday & Monday | About Machine learning and its types and related with python |
| 16 | 09/08/23 | Wednesday | Types of Machine Learning and its algorithm and how it is used in data science. |
| 17 | 09/08/23 | Wednesday | Logistic RegressionEvaluation metrics |
| 18 | 11/08/23  To  14/08/23 | Friday & Monday | KNN & SVMDecision tree &Ensemble,learning |
| 19 | 17/08/23 | Wednesday | Problem StatementsUnsupervised Learning – I Wine Quality Prediction•Diabetes prediction House price prediction |
| 20 | 19/08/23  To  19/08/23 | Friday | Titanic dataset Need of unsupervised learning K-means clustering Training k-means |
| 21 | 22/08/23  To  23/08/23 | Wednesday & Friday | Mean shift clustering K-means v/s Mean Shift clustering Industrial use cases of unsupervised learning Hyperparameter |
| 22 | 25/08/23  To  26/08/23 | Sunday & Monday | Fundamentals & Tools |
| 23 | 30/08/23 | Friday | Normalization & Transformation Cross Validations |
| 24 | 01/09/23 | Sunday | Tableau Introduction Marks Cards |
| 25 | 02/09/23 | Monday | Basics of charts |
| 26 | 03/09/23 | Tuesday | Different charts in Tableau |
| 27 | 04/09/23 | Wednesday | Database and SQL**:**  Relational databases (e.g., SQL Server, MySQL)  NoSQL databases (e.g., MongoDB, Cassandra |

1. **INTRODUCTION**

**Data Science and Machine Learning**

**Data science** is a multidisciplinary field that involves the use of scientific methods, processes, algorithms, and systems to extract insights and knowledge from structured and unstructured data. It combines expertise from various domains such as statistics, computer science, mathematics, and domain-specific knowledge to analyze and interpret complex data sets. It includes the data collection, data processing, Exploratory data analysis, Data models etc. It used statistical and different algorithm to predict the values.

Machine Learning (ML)  is a sub-category of [artificial intelligence](https://www.hpe.com/in/en/what-is/artificial-intelligence.html), that refers to the process by which computers develop pattern recognition, or the ability to continuously learn from and make predictions based on data, then make adjustments without being specifically programmed to do so.

Machine learning is incredibly complex and how it works varies depending on the task and the algorithm used to accomplish it. However, at its core, a machine learning model is a computer looking at data and identifying patterns, and then using those insights to better complete its assigned task. Any task that relies upon a set of data points or rules can be automated using machine learning, even those more complex tasks such as responding to customer service calls and reviewing resumes.

This are used to perform the different prediction and calculate the values as well as it is used in different AI tools like home appliance, Finnace, Marketing, Healthcare etc.

**Training Objective**: To enhance the knowledge of learner and perform well prediction to get the result.

**Student’s Work Assignment**: They have assigned Minor and Major project during training.

1. **COMPANY PROFILE**

YBI Foundation is a Delhi-based not-for-profit EdTech company that aims to enable the youth to grow in the world of emerging technologies. They offer a mix of online and offline approaches to bring new skills, education, technologies for students, academicians and practitioners. They believe in the learning anywhere and anytime approach to reach out to learners. The platform provides free online instructor-led classes for students to excel in data science, business analytics, machine learning, cloud computing and big data. They aim to focus on innovation, creativity, technology approach and keep themselves in sync with the present industry requirements. They endeavor to support learners to achieve the highest possible goals in their academics and professions.



Fig: India largest Internship platform & certification” YBI Foundation”

Module-1: Introduction to python

Python is a dynamic, interpreted (bytecode-compiled) language. There are no type declarations of variables, parameters, functions, or methods in source code. This makes the code short and flexible, and you lose the compile-time type checking of the source code.

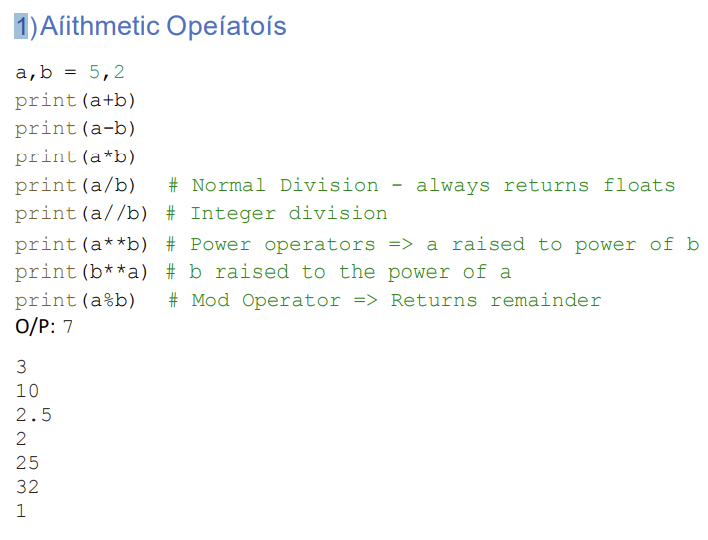
1.1 **Features of Python**

1. **Easy to Learn and Read:**
   * Python has a clear and straightforward syntax that emphasizes readability and reduces the cost of program maintenance. This makes it an excellent language for beginners and experienced developers alike.
2. **Expressive Language:**
   * Python allows developers to express concepts in fewer lines of code than languages like C++ or Java. This leads to more concise and readable programs.
3. **Interpreted Language:**
   * Python is an interpreted language, which means that the source code is executed line by line. This facilitates rapid development and testing.
4. **Cross-Platform Compatibility:**
   * Python is a cross-platform language, meaning that code written in Python can run on various operating systems, including Windows, macOS, and Linux.

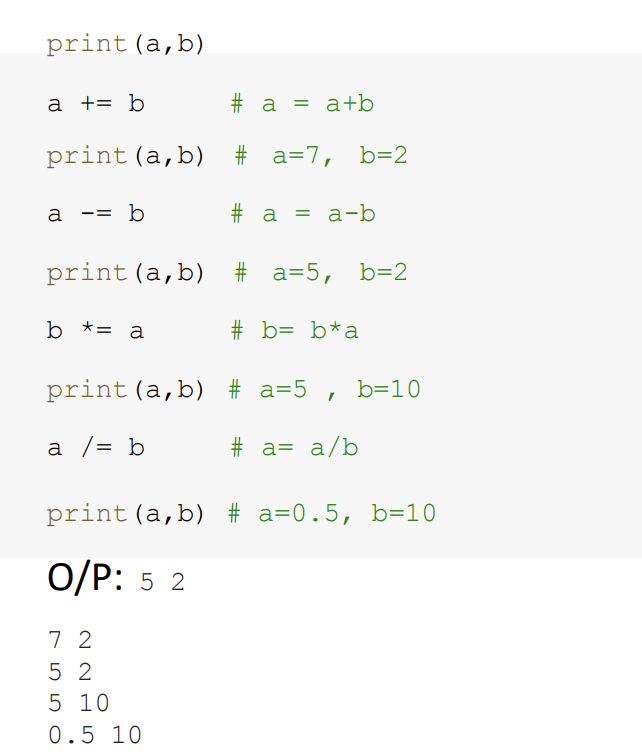
**Module-2: Python operator and operand:**

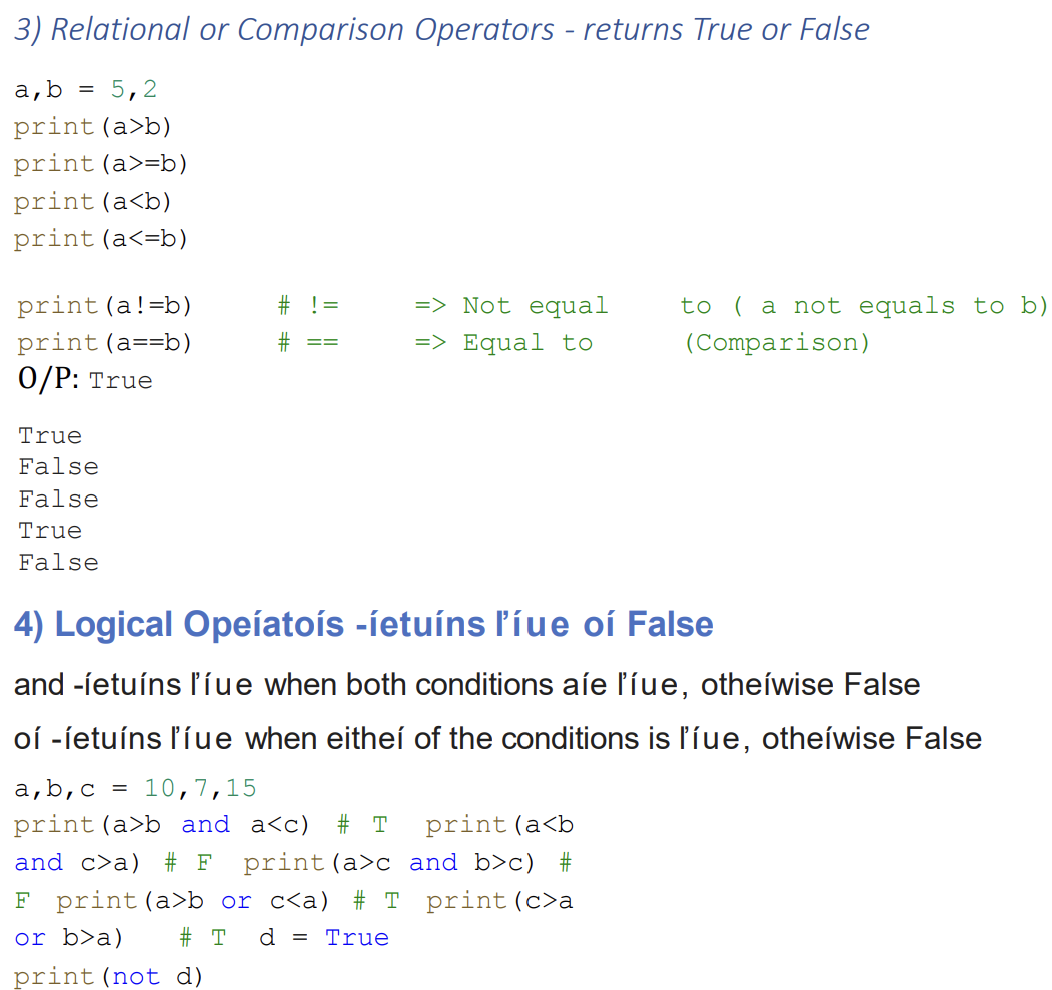
**2.1 Understanding operator**

**Arithmetic operator:** An Arithmetic Operator is a mathematical function that performs a calculation on two operands

****

2)Assignment Operator





O/P: True

False

False

True

True

False



6)Identity Operator:(“is” and “is not”)

X=3

Y=3

Print(x is not y)

O/P:FLASE

7)Bitwise Operator:

Bitwise operators are used to performing bitwise calculations on integers

**2.2 Variables and Data Types**

Variable

1)A variable is a container which is used to store some value. The value can be of any data type 2) Multiple variables can be initialized in the same line by separating them by comma

3) Variables can be reinitialized, reassigned, redeclared

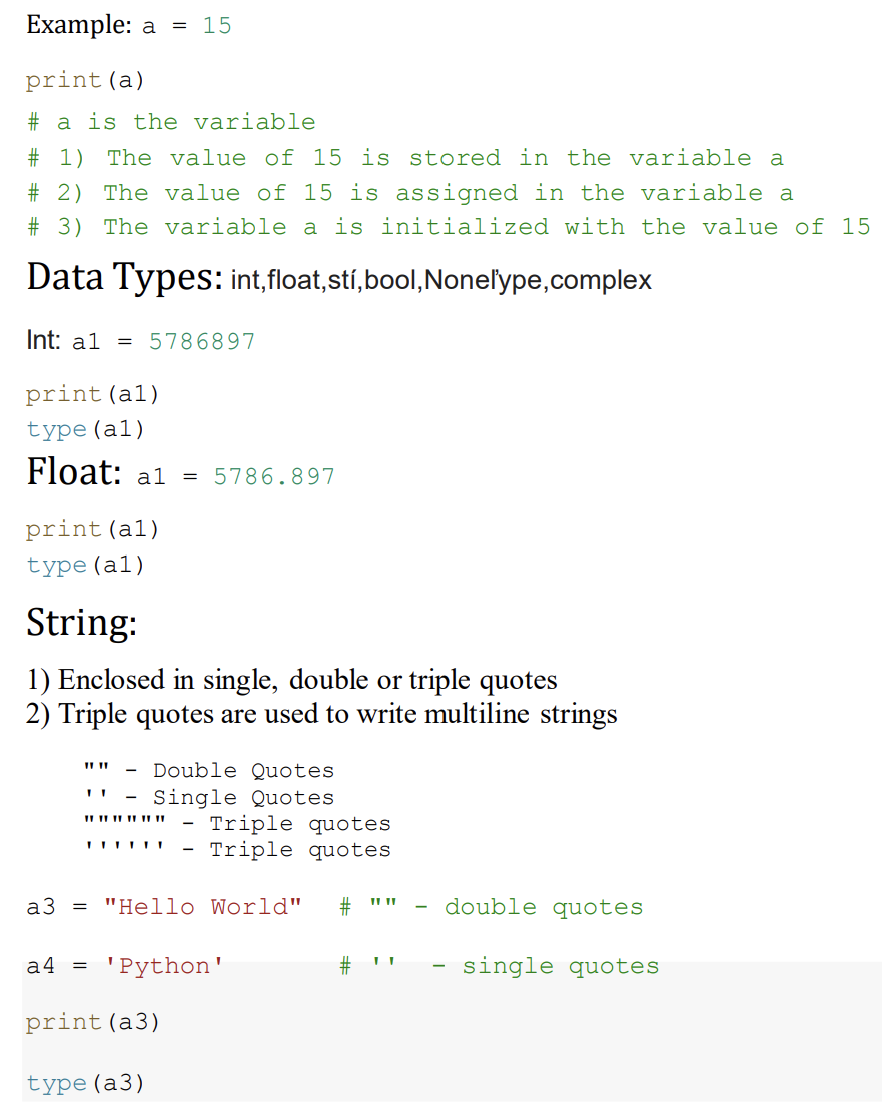
Rules of Naming a variable

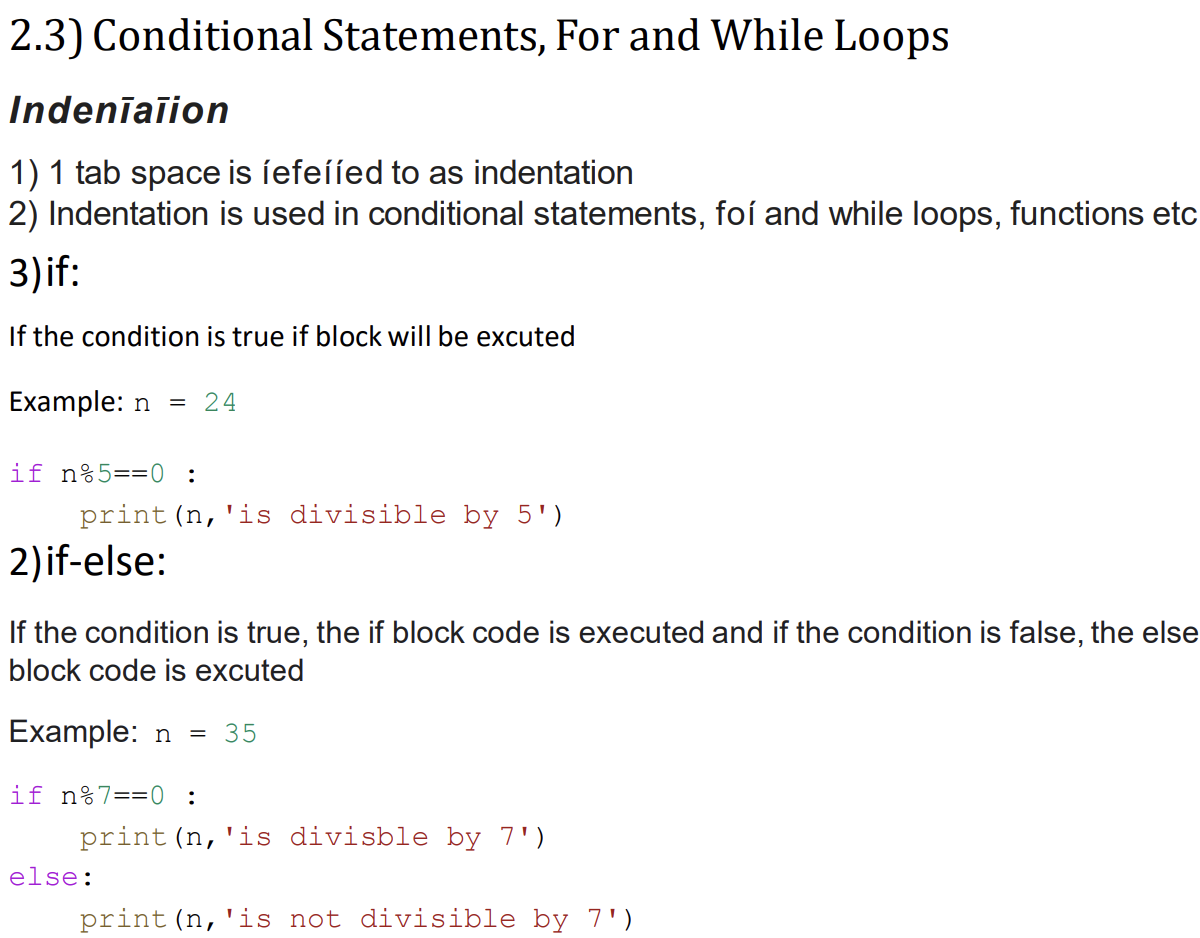
1) Start with A-Z, a-z or underscore (\_)

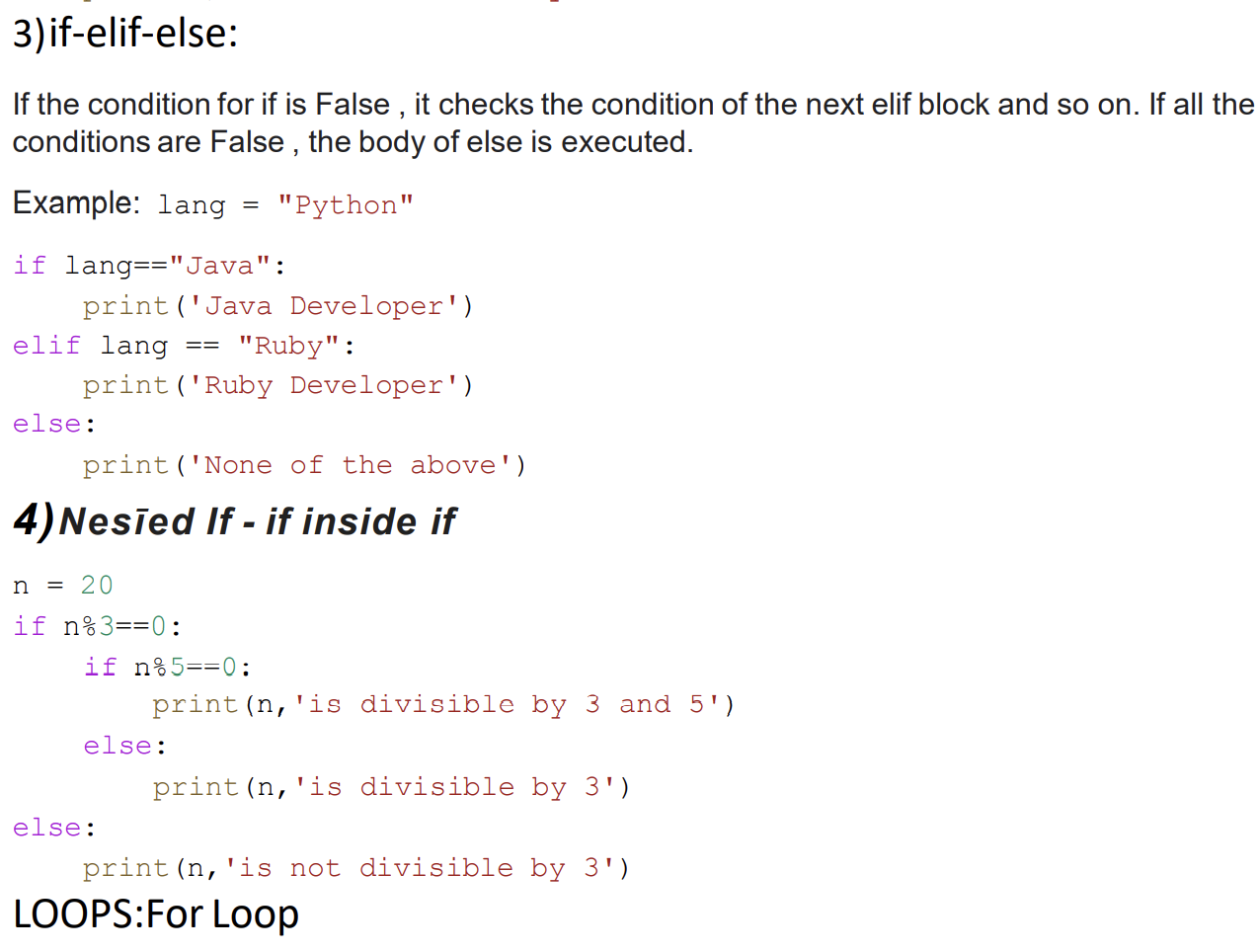
2) Should not start with number

3) Should not contain special characters

4) Should not contain python keywords







For Loops

In range (start, end, step) method 1) In ascending order, iteration proceeds from start till end-1, step size is positive

2) In descending order, iteration proceeds from start till end+1, step size is negative

3) Default step value = +1

4) range () works only on int

5) If only one value is passed in range function, it is considered as end value

6) Default start value is 0, if not mentioned

Example: for i in range (4): # start=0, end=4, step=1 print(i)

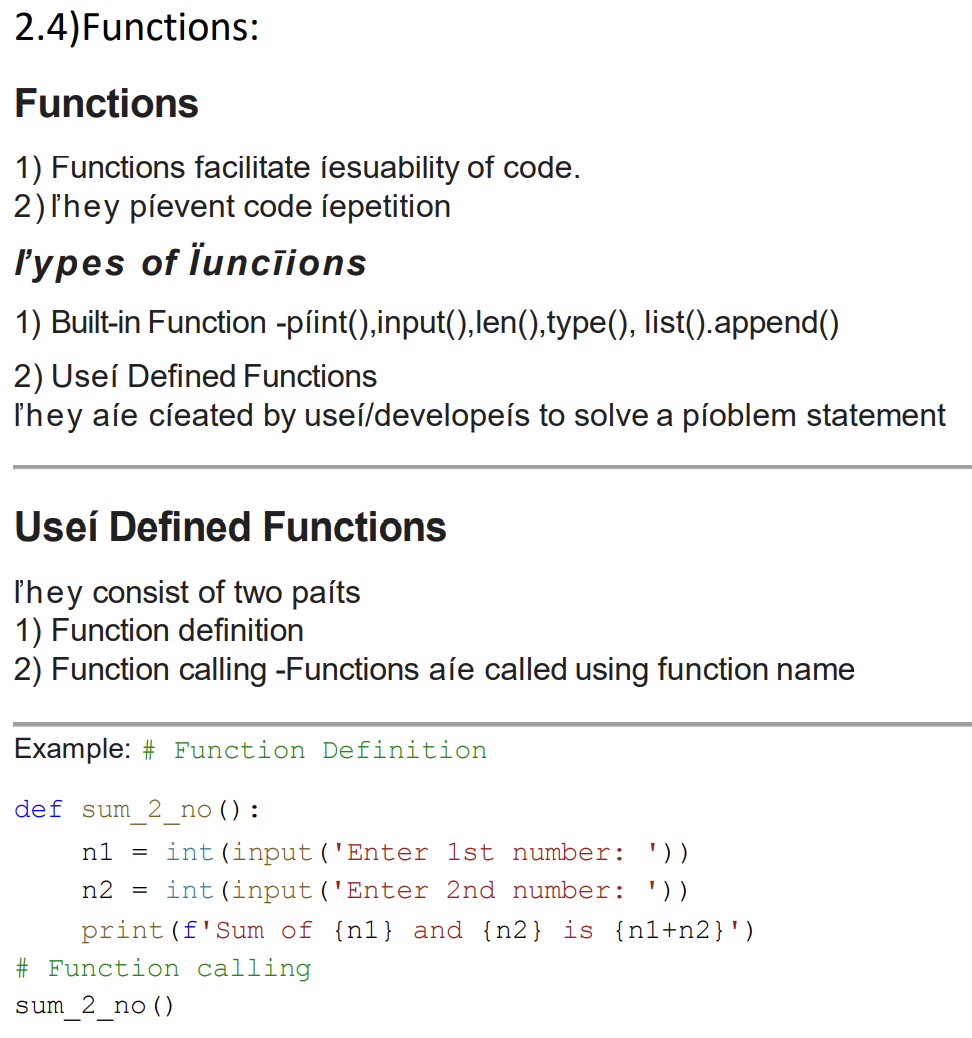
**While Loops**

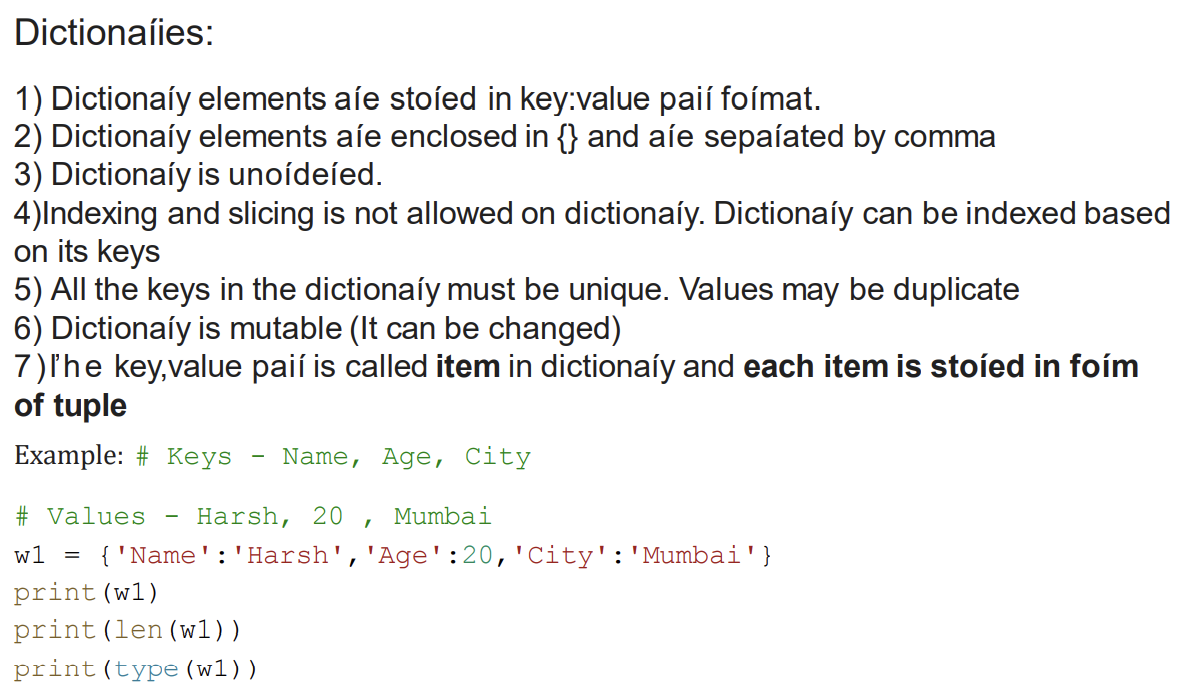
i=0 # start

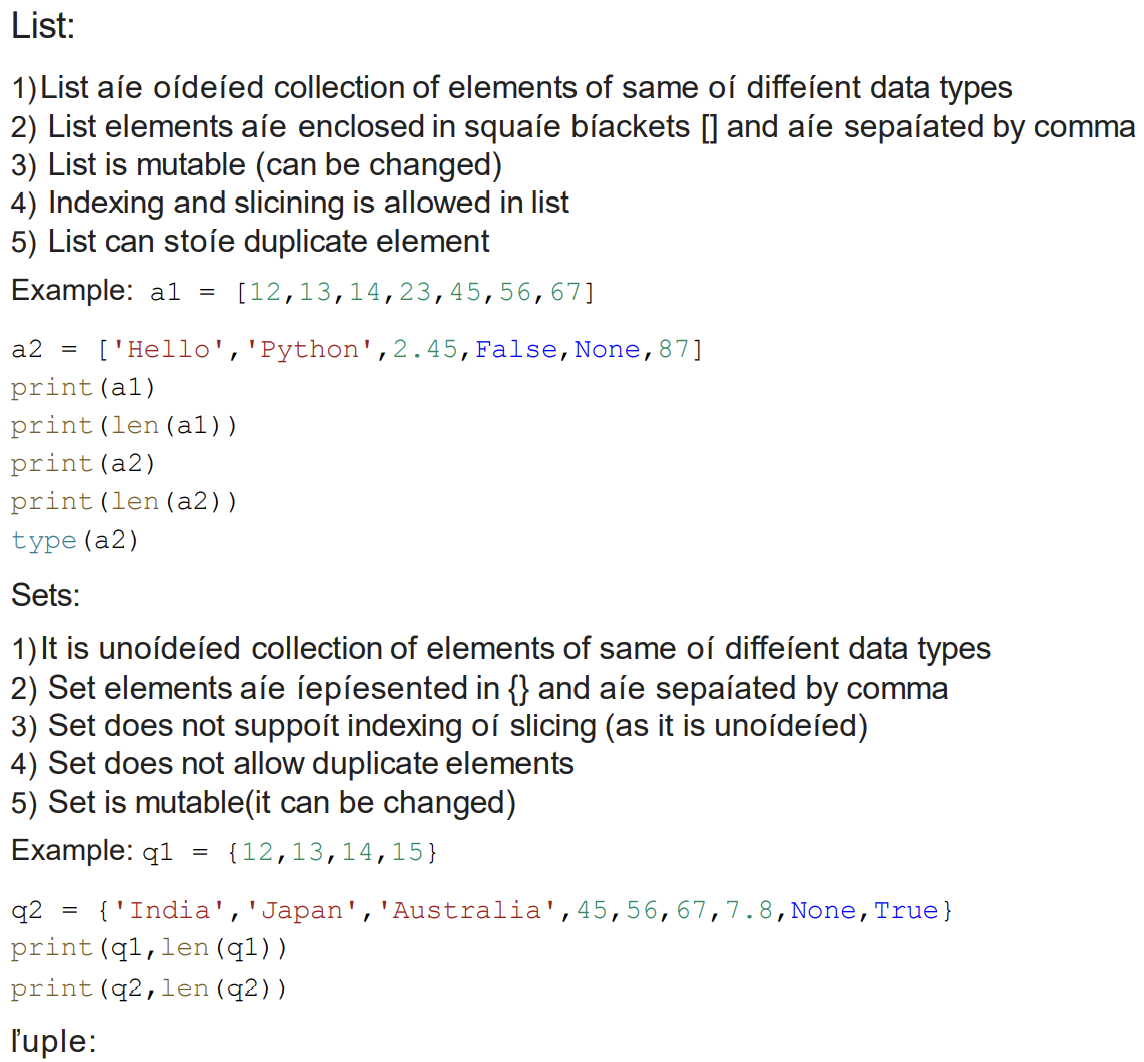
while i<5: #end

print (i,end=” “)

i+=1 #step







1)Tuple elements are represented in () and the elements are separated by comma.

2) Tuple is also an ordered collection of elements of same or different data types

3) Tuple is Immutable (can’t be changed)

4) Indexing and slicing is allowed

5) Duplicate elements are allowed

**Example:** w1 = (12,13,14,15,56,13)

**Data Science**

Data science is a multidisciplinary field that involves the use of scientific methods, processes, algorithms, and systems to extract insights and knowledge from structured and unstructured data. It combines expertise from various domains such as statistics, computer science, mathematics, and domain-specific knowledge to analyze and interpret complex data sets

**Data Collection:**

Data collection is a crucial step in the data science process, as it involves gathering relevant information from various sources to analyse and derive insights. Here are key aspects of data collection in data science:

1. **Define Objectives and Scope:**
   * Clearly define the goals and objectives of the data science project.
   * Determine the scope of the project, including the specific questions you aim to answer.
2. **Identify Data Sources:**
   * Determine where the relevant data is located. Sources may include:
     + Databases (SQL or NoSQL databases)
     + APIs (Application Programming Interfaces)
     + Web scraping
     + Sensor data
     + Log files

3.**Data Extraction:**

* + Extract the relevant data from the identified sources.
  + Transform the data into a format suitable for analysis, addressing issues such as missing values and outliers.

**4.Data Integration:**

* + If the data is spread across multiple sources, integrate it into a single, unified dataset.
  + Ensure consistency in terms of data types, formats, and units.

**Exploratory Data Analysis (EDA**) is a critical phase in the data science process that involves visually and statistically exploring a dataset to gain a deeper understanding of its patterns, characteristics, and relationships. EDA helps in formulating hypotheses, identifying patterns, and guiding subsequent steps in the data analysis process. Here are key aspects of Exploratory Data Analysis:

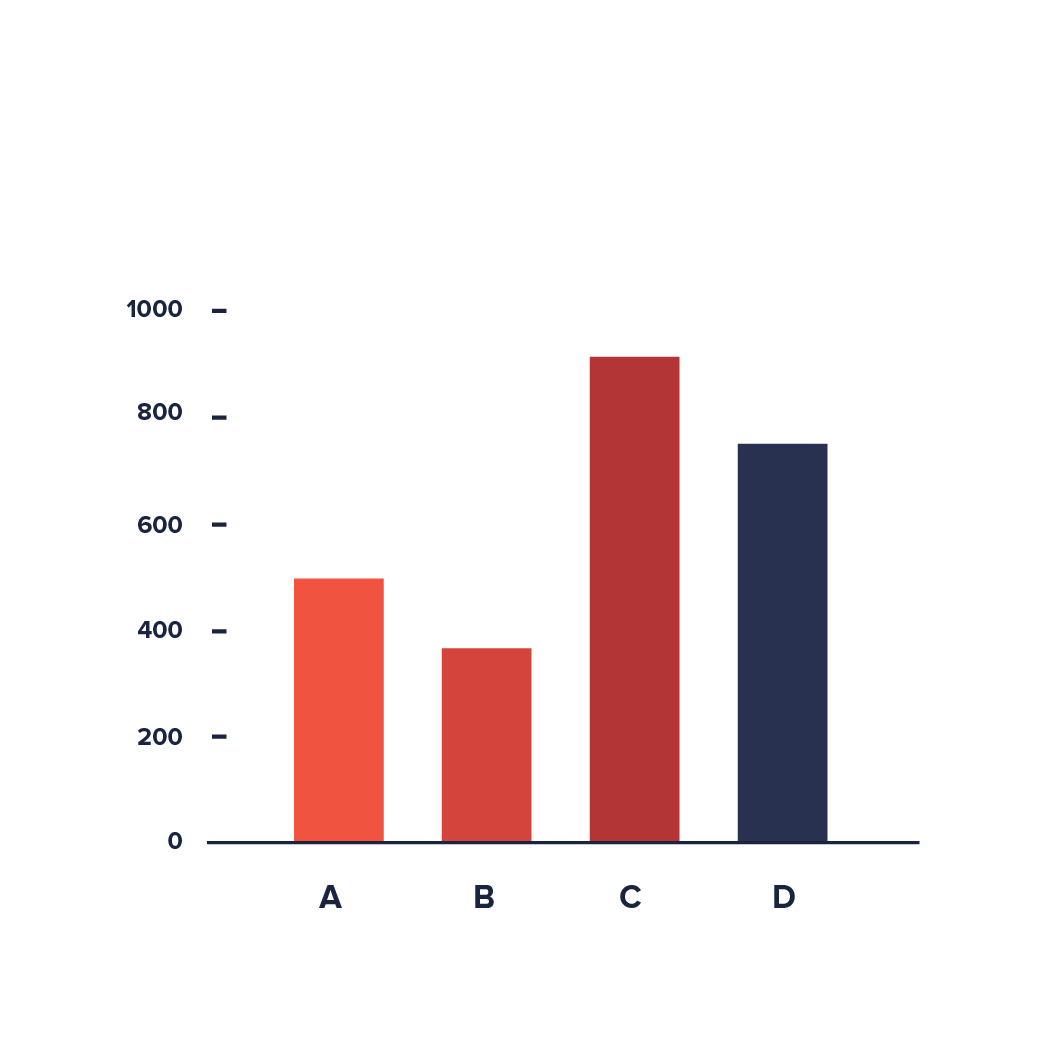
1. **Descriptive Statistics:**
   * Calculate basic summary statistics such as mean, median, mode, range, standard deviation, and quartiles.
   * Understand the distribution of numerical features**.**
2. **Data Visualization:**
   * Create visual representations of the data using charts and graphs.
   * Common visualization tools include histograms, box plots, scatter plots, line plots, and bar charts.
   * Use tools like Matplotlib, Seaborn, and Plotly in Python for visualization.
3. **Univariate Analysis:**
   * Analyse individual variables in isolation.
   * Examine the distribution of each variable to identify outliers and patterns.
4. **Bivariate Analysis:**
   * Explore relationships between pairs of variables.
   * Use scatter plots, line plots, and correlation matrices to understand how variables interact**.**
5. **Multivariate Analysis:**
   * Analyse relationships involving three or more variables simultaneously.
   * Techniques include 3D plots, heatmaps, and pair plots**.**
6. **Handling Missing Data:**
   * Identify and handle missing values appropriately.
   * Understand the impact of missing data on analysis and decision-making.
7. **Outlier Detection:**
   * Identify and examine outliers that may significantly impact the analysis.
   * Evaluate whether outliers are errors or genuine data points**.**

**Data Visualization:**

Data visualization is the representation of data in graphical or visual formats, making it easier to understand patterns, trends, and insights. Effective data visualization is a crucial aspect of data analysis and communication.

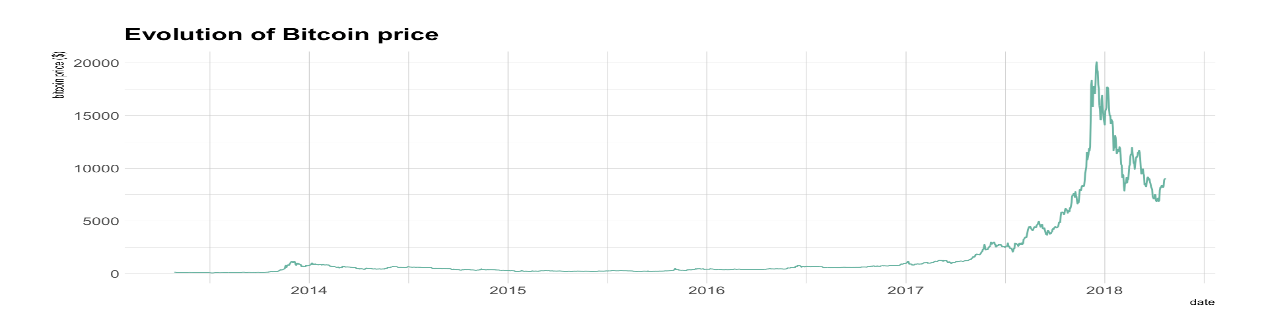
1.Bar Chart:

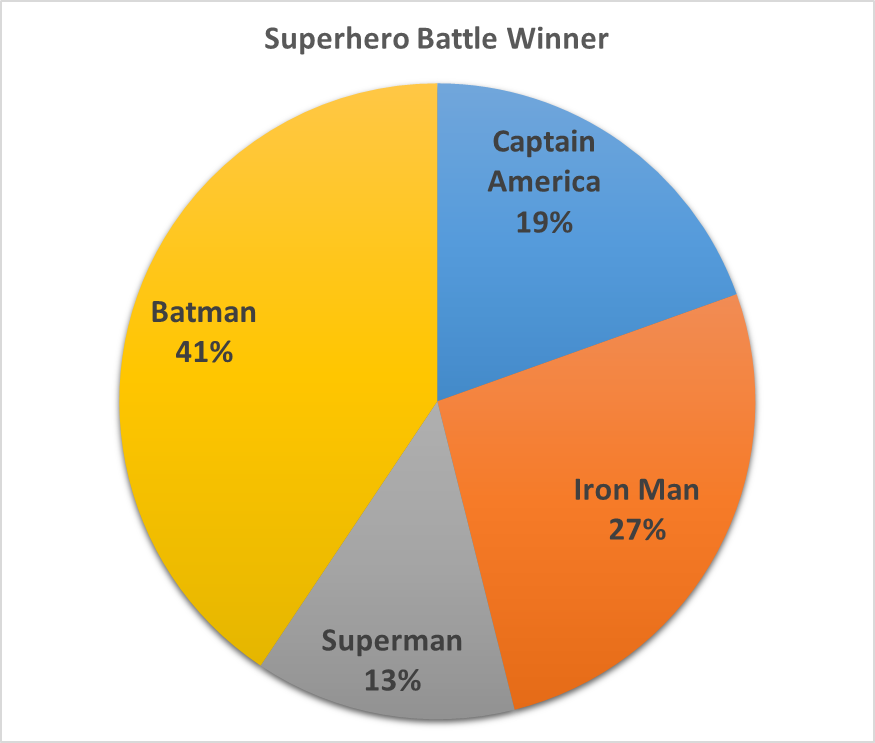
Bar charts are used to compare two or more categories of data. Each category is represented by a bar, and the length of the bar is proportional to the value of the category. Bar charts are a good choice for showing data that is categorical or discrete.

 Fig: bar chart

1. Line Chart:

Line graphs, or line charts, are a simple but effective staple for representing time-series data. They are visually similar to scatterplots but represent data points separated by time intervals with segments joined by a line



1. Pie Chart: pie charts represent a single variable, broken down into percentages or proportions.

**Machine Learning**

Machine learning (ML) is a subfield of artificial intelligence (AI) that focuses on the development of algorithms and statistical models that enable computers to perform tasks without explicit programming. The primary goal of machine learning is to allow computers to learn from data and improve their performance over time. Here are key concepts and aspects of machine learning:

1. **Types of Machine Learning:**
   * **Supervised Learning:** The algorithm is trained on a labeled dataset, where each input is associated with the correct output. The model generalizes patterns from the training data to make predictions on new, unseen data.
   * **Unsupervised Learning:** The algorithm is given data without explicit labels. The model explores patterns and structures in the data, such as clustering similar data points or reducing dimensionality.
   * **Reinforcement Learning:** The algorithm learns by interacting with an environment. It receives feedback in the form of rewards or penalties, allowing it to learn optimal strategies to achieve a goal.

**Supervised learning algorithms - Regression and Classification.**

**Regression:**

1)It is a predictive modelling technique which investigates the relationship between dependent and independent variables(one or more)

2)Dependent variable is continuous in nature eg -Sales, Weight, Profit, Revenue, Price, Distance, Magnitude,Height, Weight etc.

y = dependent variable/output

x = independent variable/input(s)

**LinearRegression:**

1)It is a regression model that estimates the relationship between one independent variable and one dependent variable using a straight line.

2) It has an equation of the form y = ax + b or y = mx + c

Where

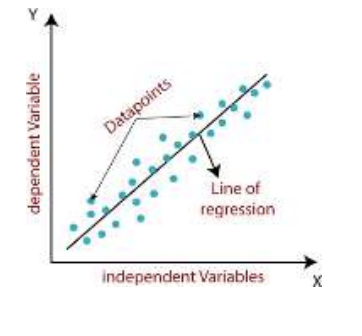
x = independent variable/ input feature/input attribute/input column

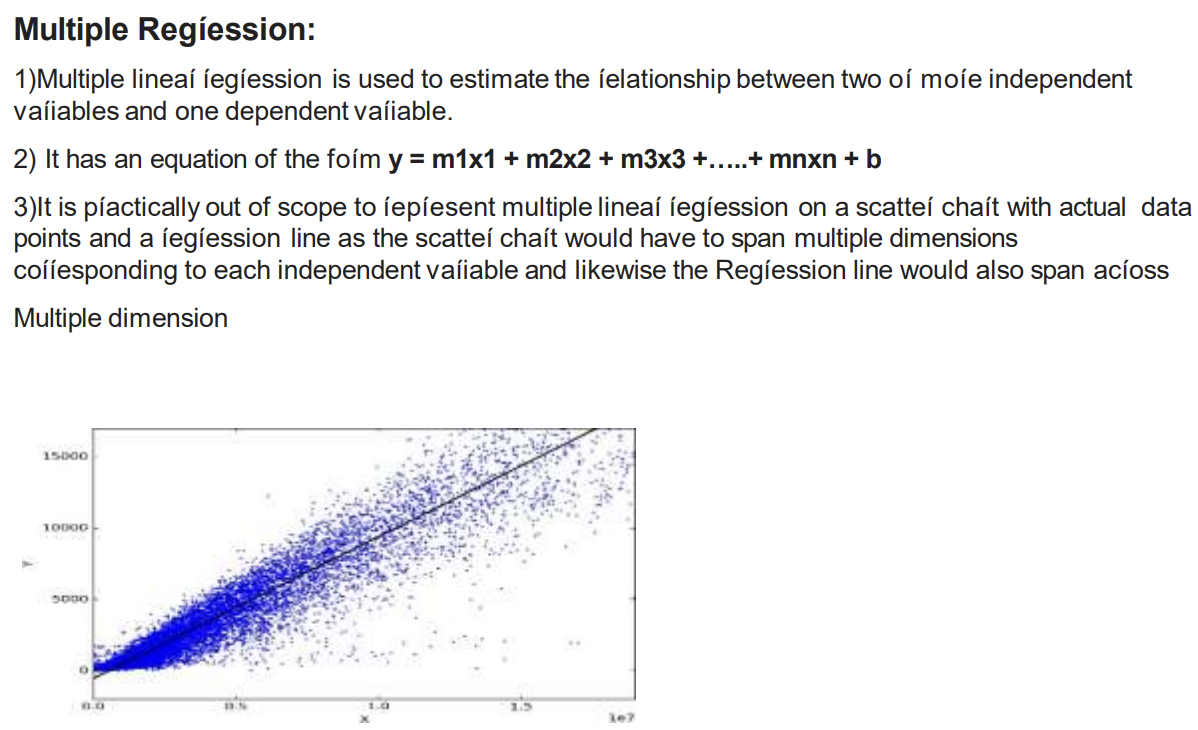
y = dependent variable / output feature/target attribute/ output column

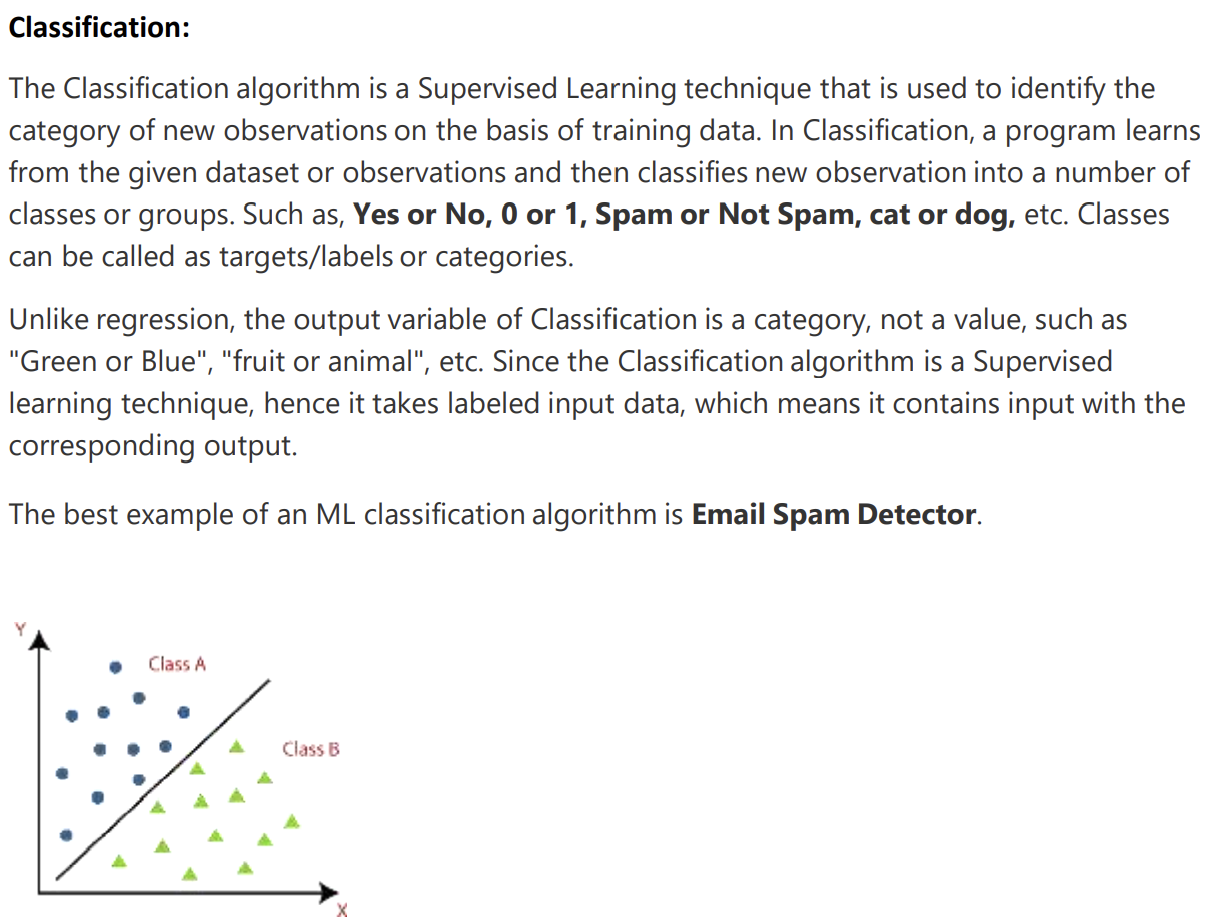
a/m = slope or coefficient or weight or how much we expect y to change as x changes

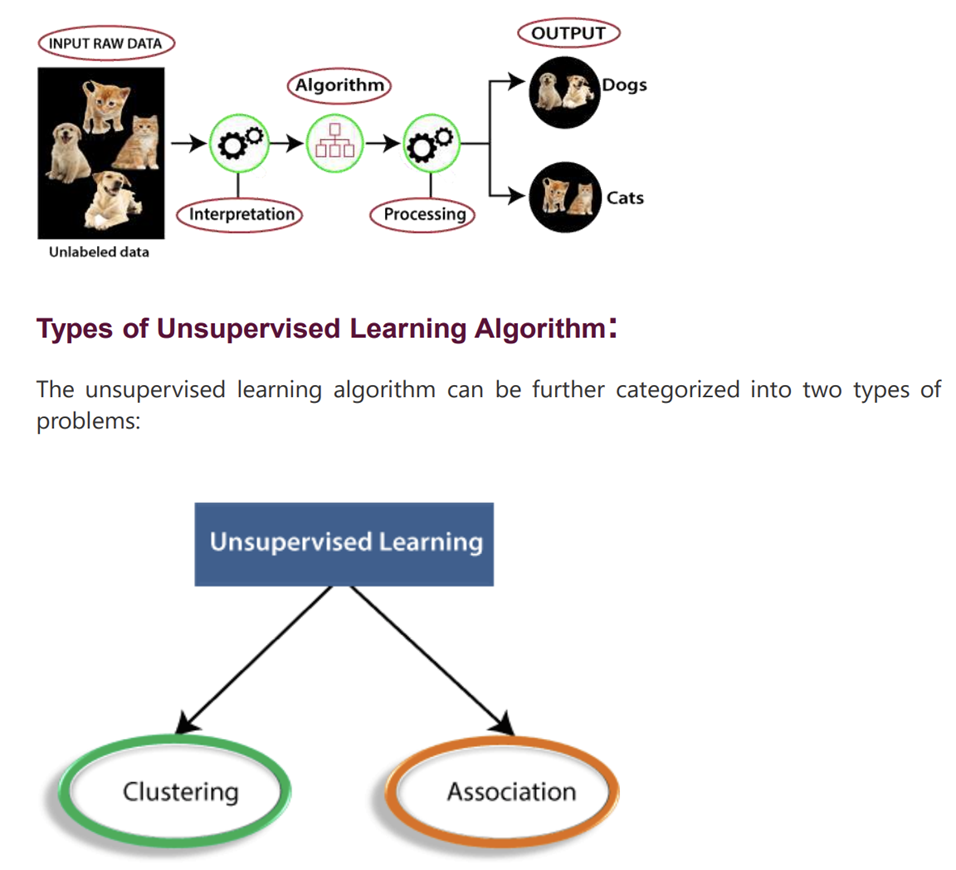
b/c = intercept / constant / bias

In this graph, x = Time spent Studying, y = Marks obtained. The orange dots are the corresponding data points. The blue line is the best fit line for Linear regression (y = mx +c)











**Types of Clustering Algorithms:**

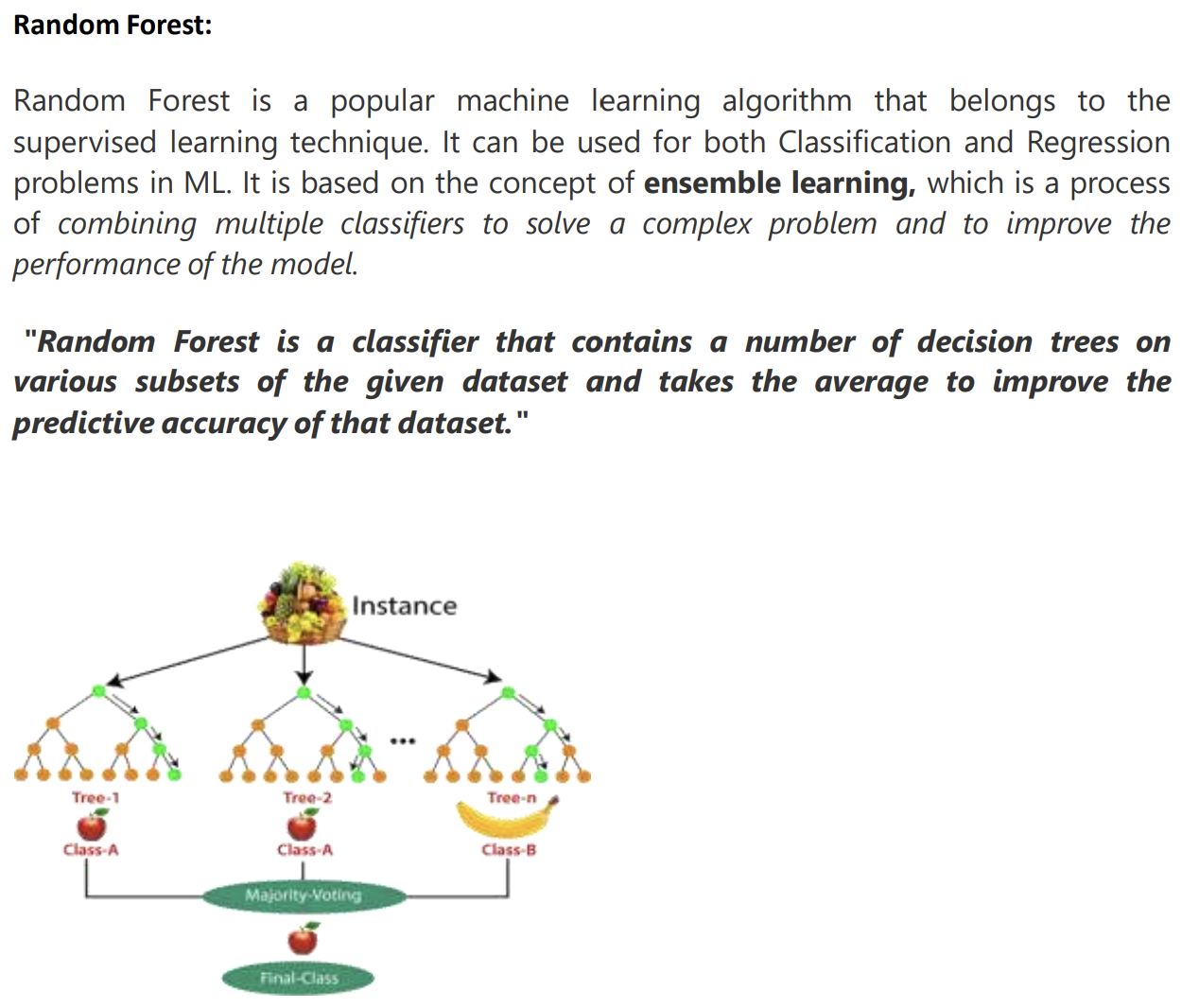
* **K-Means:** Divides data into k clusters based on centroids.
* **Hierarchical Clustering:** Creates a hierarchy of clusters through a tree-like structure.
* **DBSCAN (Density-Based Spatial Clustering of Applications with Noise):** Identifies clusters based on the density of data points.

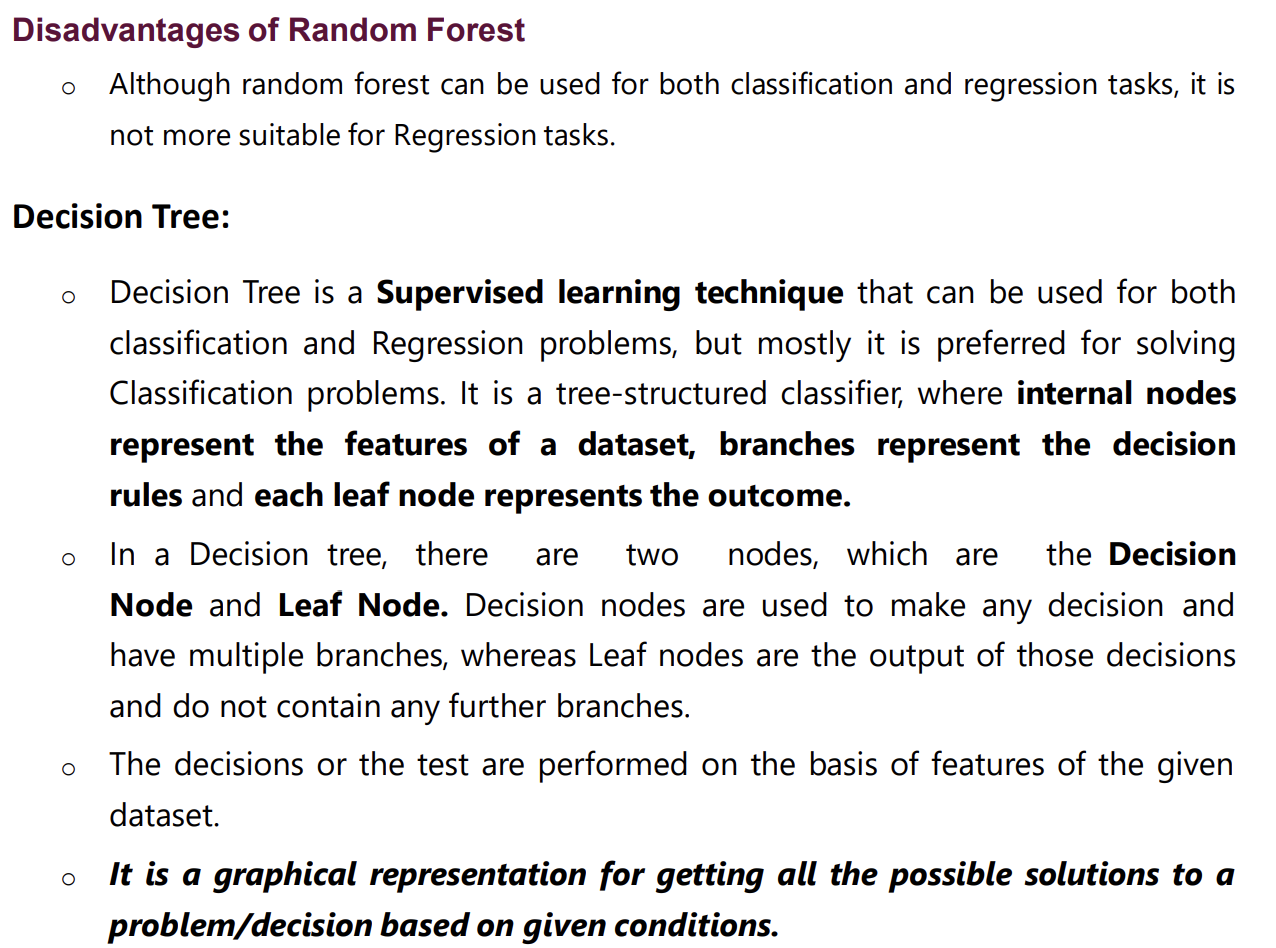
**Association:**

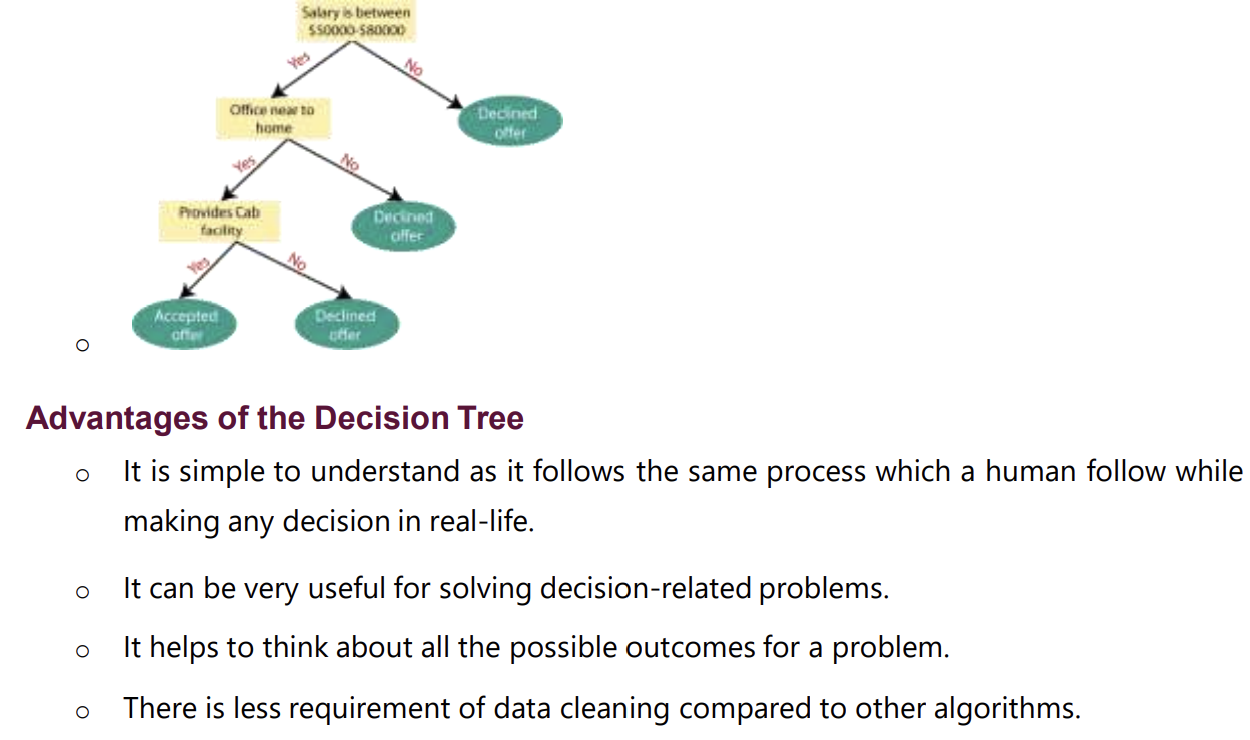
**Definition:** Association in data mining refers to discovering relationships or patterns among variables in large datasets. It focuses on finding associations or connections between items, events, or attributes.

**Key Points:**

1. **Association Rule Mining:**
   * **Support:** Measures the frequency of a set of items in the dataset.
   * **Confidence:** Measures the likelihood that an association rule holds true.
   * **Lift:** Measures the ratio of the observed support to the expected support.
2. **Types of Association Rule Algorithms:**
   * **A-priori Algorithm:** Generates frequent itemset and association rules.
   * **FP-Growth (Frequent Pattern Growth):** Utilizes a tree structure to mine frequent itemset efficiently.







It may have an overfitting issue, which can be resolved using the Random Forest algorithm. For more class labels, the computational complexity of the decision tree may increase.

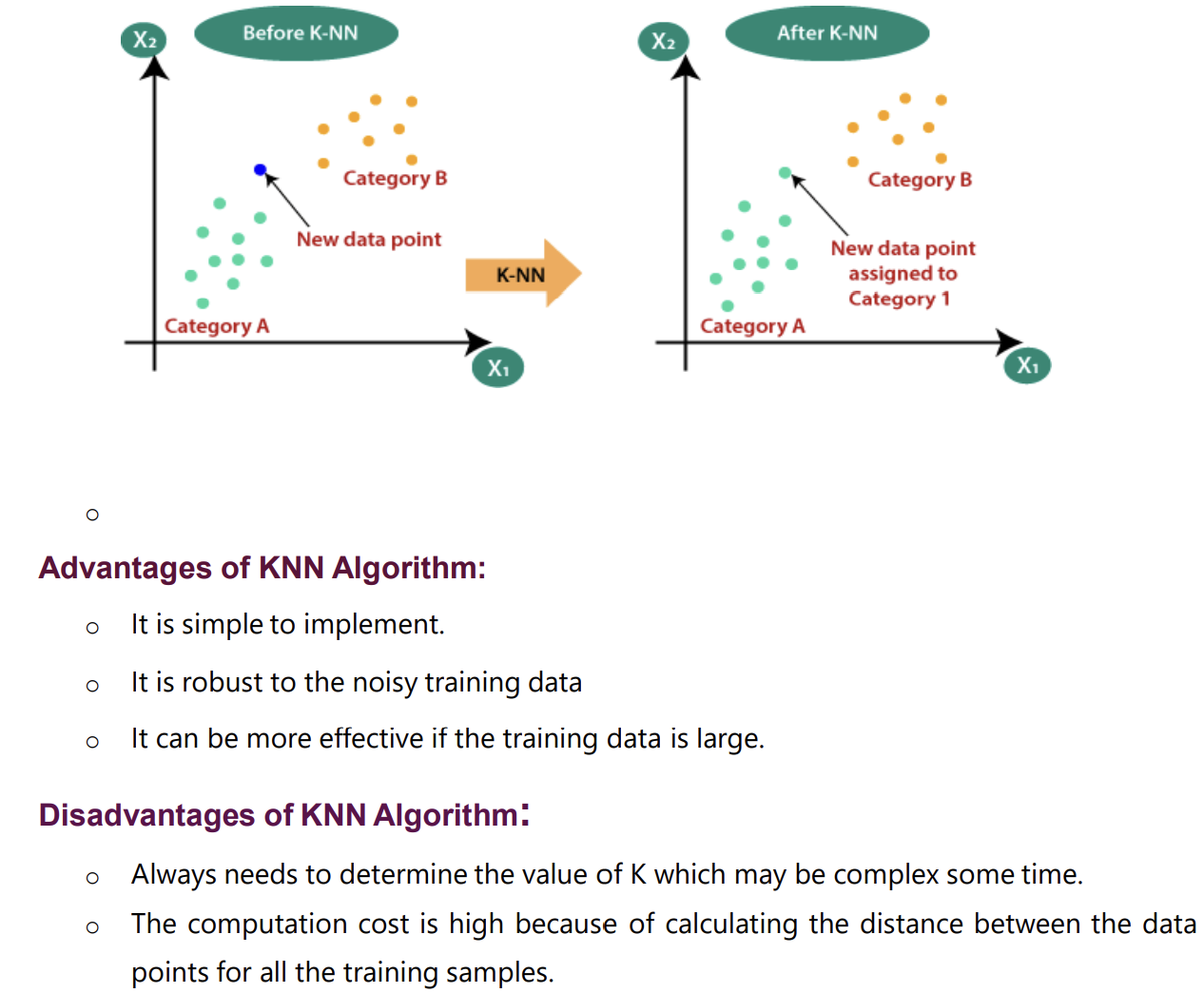
**K-Nearest Neighbor (KNN) Algorithm**

> K-Nearest Neighbour is one of the simplest Machine Learning algorithms based on Supervised Learning technique.

> K-NN algorithm assumes the similarity between the new case/data and available cases and put the new case into the category that is most similar to the available categories.

> K-NN algorithm stores all the available data and classifies a new data point based on the similarity. This means when new data appears then it can be easily classified into a well suite category by using K- NN algorithm.

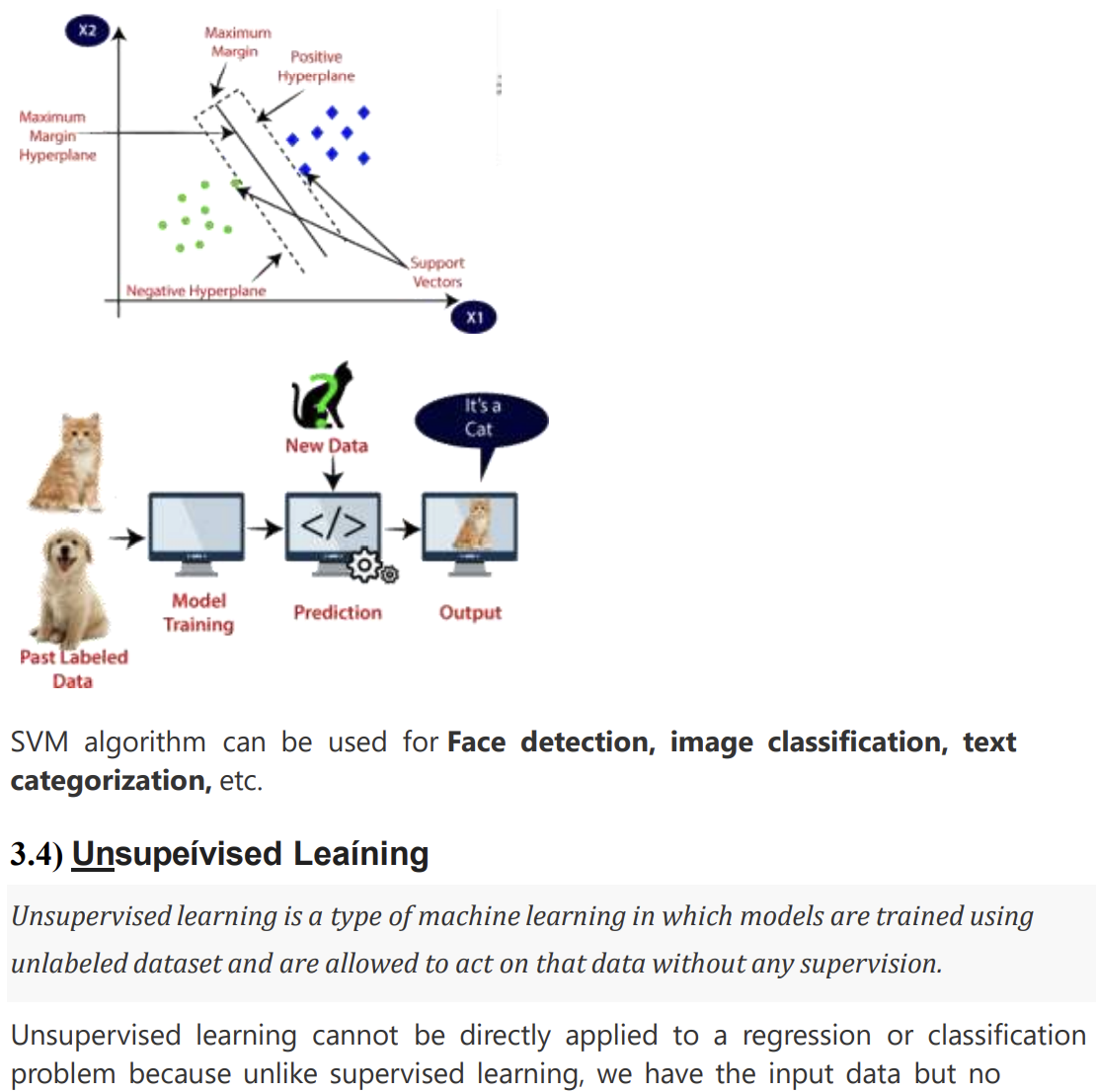
> K-NN algorithm can be used for Regression as well as for Classification but mostly it is used for the Classification problems.



**Support Vector Machine Algorithm**

Support Vector Machine or SVM is one of the most popular Supervised Learning algorithms, which is used for Classification as well as Regression problems. However, primarily, it is used for Classification problems in Machine Learning.

The goal of the SVM algorithm is to create the best line or decision boundary that can segregate n-dimensional space into classes so that we can easily put the new data point in the correct category in the future. This best decision boundary is called a hyperplane.

 SVM chooses the extreme points/vectors that help in creating the hyperplane. These extreme cases are called as support vectors, and hence algorithm is termed as Support Vector Machine. Consider the below diagram in which there are two

**Different Libarires used in Data Science and Machine Learning**

**Matplotlib:**

**Description:**

* **Type:** Python library.
* **Purpose:** Matplotlib is a powerful 2D plotting library for Python. It enables the creation of a wide variety of static, animated, and interactive plots.
* **Features:**
  + Supports line plots, scatter plots, bar plots, histograms, and more.
  + Highly customizable plot elements.
  + Integrates well with Jupiter notebooks.
* **Use Cases:** Matplotlib is widely used in the Python ecosystem for creating static visualizations in various domains, including data science and scientific research.

**Tableau:**

**Description:**

* **Type:** Data visualization software.
* **Purpose:** Tableau is a leading data visualization tool that provides a user-friendly interface for creating interactive and shareable dashboards. It allows users to connect to diverse data sources and explore insights visually.
* **Features:**
  + Drag-and-drop interface for creating dashboards without coding.
  + Offers a variety of pre-built visualizations.
  + Connects to various data sources, facilitating easy data integration.
* **Use Cases:** Tableau is employed for business intelligence, data analysis, and creating interactive dashboards for decision-making.

**Seaborn:**

**Description:**

* **Type:** Python library.
* **Purpose:** Seaborn is a statistical data visualization library built on top of Matplotlib. It provides a high-level interface for creating visually appealing statistical graphics with less code.
* **Features:**
  + Attractive default styles and colour palettes.
  + Specialized functions for visualizing statistical relationships in data.
  + Integration with Pandas data structures.
* **Use Cases:** Seaborn is commonly used in the Python data science community to create aesthetically pleasing and informative statistical visualizations.

**Natural Language Processing (NLP):**

**Definition:**

* **NLP** is a subfield of artificial intelligence that focuses on the interaction between computers and human language. It involves the development of algorithms and models to enable machines to understand, interpret, and generate human-like language.

**Key Concepts and Tasks:**

1. **Tokenization:** Breaking down text into smaller units, such as words or sentences.
2. **Part-of-Speech Tagging:** Assigning grammatical categories (e.g., noun, verb) to words.
3. **Named Entity Recognition (NER):** Identifying and classifying entities (e.g., names, locations) in text.
4. **Sentiment Analysis:** Determining the sentiment or emotion expressed in a piece of text.
5. **Machine Translation:** Automatically translating text from one language to another.
6. **Text Summarization:** Generating concise summaries of longer pieces of text.
7. **Question Answering:** Developing systems that can answer questions posed in natural language.

**Text Mining:**

**Definition:**

* **Text Mining** (also known as Text Analytics) is the process of deriving meaningful information and patterns from large volumes of unstructured text data. It involves extracting valuable insights and knowledge from textual information.

**Key Concepts and Tasks:**

1. **Text Preprocessing:** Cleaning and transforming raw text data for analysis.
2. **Information Retrieval:** Extracting relevant information from a large corpus of documents.
3. **Topic Modelling:** Identifying topics present in a collection of documents.
4. **Clustering:** Grouping similar documents or text snippets together.
5. **Text Classification:** Assigning predefined categories or labels to documents.
6. **Pattern Recognition:** Identifying patterns and trends in textual data.

**Database:**

A database is an organized collection of structured information or data that is stored electronically in a computer system. Databases are designed to efficiently manage, store, and retrieve data. They are a critical component in various applications, enabling the storage and manipulation of vast amounts of information. Key concepts related to databases include:

1. **Database Management System (DBMS):**
   * A software system that provides an interface to interact with the database.
   * Manages the storage, retrieval, and organization of data.
2. **Relational Database:**
   * Organizes data into tables with rows and columns.
   * Uses relationships between tables to establish connections and maintain data integrity.
3. **NoSQL Database:**
   * A type of database that does not strictly adhere to the traditional relational model.
   * Suited for handling unstructured or semi-structured data and provides flexible schema design.
4. **Key-Value Store:**
   * A NoSQL database model where each data item is stored as a key-value pair.
5. **Document Store:**
   * A NoSQL database model that stores data in semi-structured documents (e.g., JSON or XML).
6. **Graph Database:**
   * A database designed for handling data with complex relationships, using a graph structure.
7. **ACID Properties:**
   * A set of properties (Atomicity, Consistency, Isolation, Durability) that guarantee the reliability of database transactions.

**SQL (Structured Query Language):**

SQL is a specialized programming language designed for managing and manipulating relational databases. It is used to perform various operations on databases, such as querying, updating, inserting, and deleting data. Key concepts and statements in SQL include:

1. **SELECT Statement:**
   * Retrieves data from one or more tables.

sqlCopy code

SELECT column1, column2 FROM table WHERE condition;

1. **INSERT Statement:**
   * Adds new records to a table.

sqlCopy code

INSERT INTO table (column1, column2) VALUES (value1, value2);

1. **UPDATE Statement:**
   * Modifies existing records in a table.

sqlCopy code

UPDATE table SET column1 = value1 WHERE condition;

1. **DELETE Statement:**
   * Removes records from a table.

sqlCopy code

DELETE FROM table WHERE condition;

1. **CREATE TABLE Statement:**
   * Defines a new table with its structure.

sqlCopy code

CREATE TABLE table ( column1 datatype1, column2 datatype2, ... );

1. **JOIN Operation:**
   * Combines rows from two or more tables based on a related column.

sqlCopy code

SELECT \* FROM table1 INNER JOIN table2 ON table1.column = table2.column;

1. **INDEX:**
   * Improves the speed of data retrieval operations on a database table.

sqlCopy code

CREATE INDEX index\_name ON table (column);

1. **Normalization:**
   * The process of organizing data in a database to reduce redundancy and improve data integrity.

**PROJECT WORK**

During the period of (6 week) Internship they have assigned two projects to do :

1. Minor Project
2. Major Project

**Minor Project:**

In Minor project, I have done “***Car price prediction*” and “*SMS spam classifier***”

**Car price prediction** models typically involve the use of various modules or components to analyze and predict the prices of cars. The specific modules used can vary depending on the complexity of the model and the data available. Here are some common modules or steps that may be involved in building a car price prediction model:

1. Data Collection and Preprocessing:
   * Data Collection Module: Collecting relevant data on cars, including features such as make, model, year, mileage, fuel type, engine size, etc.
   * Data Cleaning and Preprocessing Module: Handling missing data, removing outliers, and preparing the data for analysis.

2.Exploratory Data Analysis (EDA):

* + Descriptive Statistics Module: Analyzing summary statistics, distributions, and other descriptive measures of the data.
  + Data Visualization Module: Creating visualizations to understand the relationships between different features and the target variable (car prices).

3.Model Building:

* + Model Selection Module: Choosing an appropriate machine learning algorithm for regression tasks (e.g., linear regression, decision trees, random forests, gradient boosting).

**SMS spam prediction** models typically involve several modules to process and analyse text data for classifying messages as spam or not spam. Below are common modules or steps involved in building an SMS spam prediction model:

1.Feature Extraction:

* + Tokenization Module: Breaking down the text into individual words or tokens.

To build an SMS spam classifier, you'll likely use various libraries and frameworks that provide tools for data preprocessing, feature extraction, machine learning, and evaluation. Here are some commonly used libraries in Python for building SMS spam classifiers:

**NLTK (Natural Language Toolkit**):

NLTK is a powerful library for natural language processing. It provides tools for tokenization, stemming, lemmatization, and other text processing tasks.

*import nltk*

*Scikit-learn:*

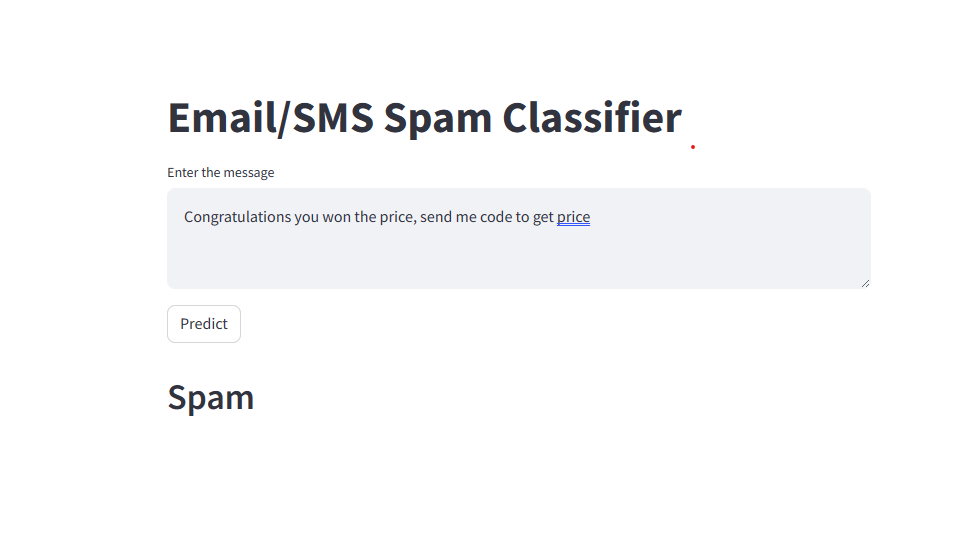
**Scikit-learn** is a popular machine learning library that includes tools for data preprocessing, feature extraction, and implementing machine learning algorithms.

*from sklearn.feature\_extraction.text import CountVectorizer, TfidfVectorizer*

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

*from sklearn.naive\_bayes import MultinomialNB*

*from sklearn.metrics import accuracy\_score, classification\_report, confusion\_matrix*

****

**2.Major Project**

On major project, I have done on **topic “Online Insurance Analysis In Python”**

Online insurance analysis involves the examination and assessment of various aspects related to the digital landscape of the insurance industry. As insurance services increasingly migrate to online platforms, thorough analysis becomes crucial for optimizing operations, enhancing customer experiences, and ensuring the security and efficiency of digital processes.

1. **Pandas:**
   * Pandas is a powerful data manipulation library that provides data structures for efficient data analysis. It is commonly used for tasks such as cleaning, transforming, and exploring datasets.

*import pandas as pd*

1. **NumPy:**
   * NumPy is a fundamental package for scientific computing in Python. It provides support for large, multi-dimensional arrays and matrices, along with mathematical functions to operate on these arrays.

*import numpy as np*

1. **Scikit-learn:**
   * Scikit-learn is a machine learning library that provides simple and efficient tools for data analysis and modeling, including modules for classification, regression, clustering, and more.

*from sklearn.model\_selection import train\_test\_split from sklearn.metrics import accuracy\_score, classification\_report*

1. **Matplotlib and Seaborn:**
   * Matplotlib and Seaborn are popular libraries for data visualization. They are used to create various types of plots and charts to better understand and communicate data insights.

*import matplotlib.pyplot as plt import seaborn as sns*

1. **TensorFlow or PyTorch:**
   * For more advanced machine learning tasks, such as deep learning, libraries like TensorFlow or PyTorch can be used.

*import tensorflow as tf*

1. **Statsmodels:**
   * Statsmodels is a library for estimating and testing statistical models. It is useful for tasks such as regression analysis.
2. Streamlit: Used as webservices.

**Input output dataset/screenshots**

**Home.py**

import streamlit as st

import pandas as pd

import plotly.express as px

from streamlit\_option\_menu import option\_menu

from numerize.numerize import numerize

from query import \*

import time

st.set\_page\_config(page\_title="Dashboard", page\_icon="🌍",layout="wide")

st.subheader("🚮Insurance Details Analysis")

st.markdown("##")

result=view\_all\_data()

df=pd.DataFrame(result,columns=["Policy","Expiry","Location","State","Region","Investment","Construction","BusinessType","Earthquake","Flood","Rating","id"])

#side bar images

st.sidebar.image("logo1.png",caption="Online Analysis")

#switcher

st.sidebar.header("Please filter")

region=st.sidebar.multiselect(

    "Select Region",

    options=df["Region"].unique(),

    default=df["Region"].unique(),

)

location=st.sidebar.multiselect(

    "Select Location",

    options=df["Location"].unique(),

    default=df["Location"].unique(),

)

construction=st.sidebar.multiselect(

    "Select Construction",

    options=df["Construction"].unique(),

    default=df["Construction"].unique(),

)

df\_selection=df.query(

    "Region==@region & Location==@location & Construction==@construction"

)

def Home():

    with st.expander("Tabular"):

        showData=st.multiselect('Filter: ',df\_selection.columns,default=[])

        st.write(df\_selection[showData])

    #compute the top analytical

    total\_investment = float(df\_selection['Investment'].sum())

    investment\_mode = float(df\_selection['Investment'].mode())

    investment\_mean = float(df\_selection['Investment'].mean())

    investment\_median = float(df\_selection['Investment'].median())

    rating = float(df\_selection['Rating'].sum())

    total1,total2,total3,total4,total5=st.columns(5,gap='large')

    with total1:

        st.info('Total Investment',icon="📌")

        st.metric(label="sum TZS",value=f"{total\_investment: ,.0f}")

    with total2:

        st.info('Most Frequent',icon="📌")

        st.metric(label="mode TZS",value=f"{investment\_mode:,.0f}")

    with total3:

        st.info('Average',icon="📌")

        st.metric(label="average TZS",value=f"{investment\_mean:,.0f}")

    with total4:

        st.info('Central Earning',icon="📌")

        st.metric(label="median TZS",value=f"{investment\_median:,.0f}")

    with total5:

        st.info('Rating',icon="📌")

        st.metric(label="Rating",value=numerize(rating),help=f""" Total Rating: {rating} """)

    st.markdown("""\_ \_ \_""")

#graphs

def graphs():

    #total\_investment=int(df\_selection["Investment"]).sum()

    #averageRating=int(round(df\_selection["Rating"]).mean(),2)

    #simple bar graph

    investment\_by\_business\_type=(

        df\_selection.groupby(by=["BusinessType"]).count()[["Investment"]].sort\_values(by="Investment")

    )

    fig\_investment=px.bar(

        investment\_by\_business\_type,

        x="Investment",

        y=investment\_by\_business\_type.index,

        orientation="h",

        title="<b> Investment by  Business type </b>",

        color\_discrete\_sequence=["#0083B8"]\*len(investment\_by\_business\_type),

        template="plotly\_white",

    )

    fig\_investment.update\_layout(

    plot\_bgcolor="rgba(0,0,0,0)",

    xaxis=(dict(showgrid=False))

      )

    #simple line graph

    investment\_state=df\_selection.groupby(by=["State"]).count()[["Investment"]]

    fig\_state=px.line(

        investment\_state,

        x=investment\_state.index,

        y="Investment",

        orientation="v",

        title="<b> Investment by  state </b>",

        color\_discrete\_sequence=["#0083b8"]\*len(investment\_state),

        template="plotly\_white",

    )

    fig\_state.update\_layout(

    xaxis=dict(tickmode="linear"),

    plot\_bgcolor="rgba(0,0,0,0)",

    yaxis=(dict(showgrid=False))

      )

    left,right=st.columns(2)

    left.plotly\_chart(fig\_state,use\_container\_width=True)

    right.plotly\_chart(fig\_investment,use\_container\_width=True)

def Progressbar():

    st.markdown("""<style>.stProgress > div > div > div > div { background(to right, #99ff99, #FFFF00)}</style""",unsafe\_allow\_html=True,)

    target=3000000000

    current=df\_selection["Investment"].sum()

    percent=round((current/target\*100))

    mybar=st.progress(0)

    if percent>100:

        st.subheader("Target done !")

    else:

        st.write("You have (",percent, "% ", "of ", (format(target, 'd')), "TZS")

        for percent\_complete in range(percent):

            time.sleep(0.1)

            mybar.progress(percent\_complete+1,text="Target Percentage")

def sideBar():

    with st.sidebar:

        selected=option\_menu(

            menu\_title="Main Menu",

            options=["Home", "Progress"],

            icons=["house", "eye"],

            menu\_icon="cast",

            default\_index=0

        )

    if selected=="Home":

        st.subheader(f"Page: {selected}")

        Home()

        graphs()

    if selected=="Progress":

        st.subheader(f"Page: {selected}")

        Progressbar()

        graphs()

sideBar()

#theme

hide\_st\_style="""

<style>

#MainMenu{visibility:hidden;}

footer{visibility:hidden;}

header{visibility:hidden;}

</style>

"""

Query.py

import mysql.connector

import streamlit as st

#connecting the sql with python as for the database

conn=mysql.connector.connect(

    host="localhost",

    port="3306",

    user="root",

    passwd="",

    db="mydbs"

)

c=conn.cursor()

#fetch

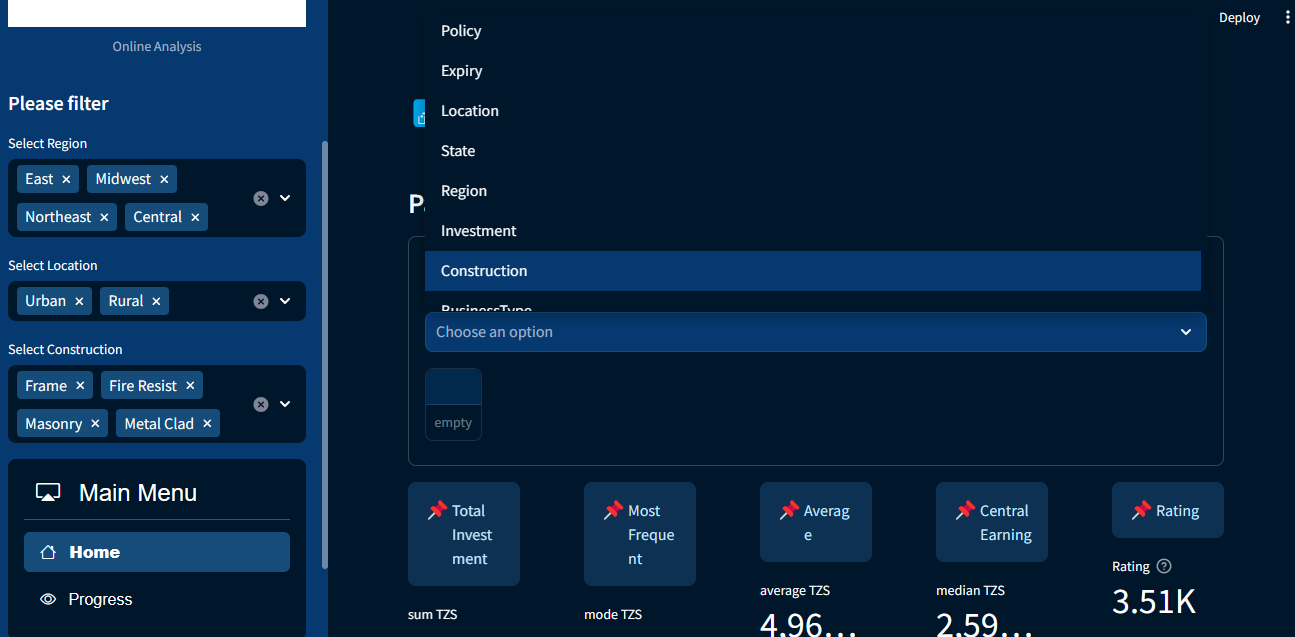
def view\_all\_data():

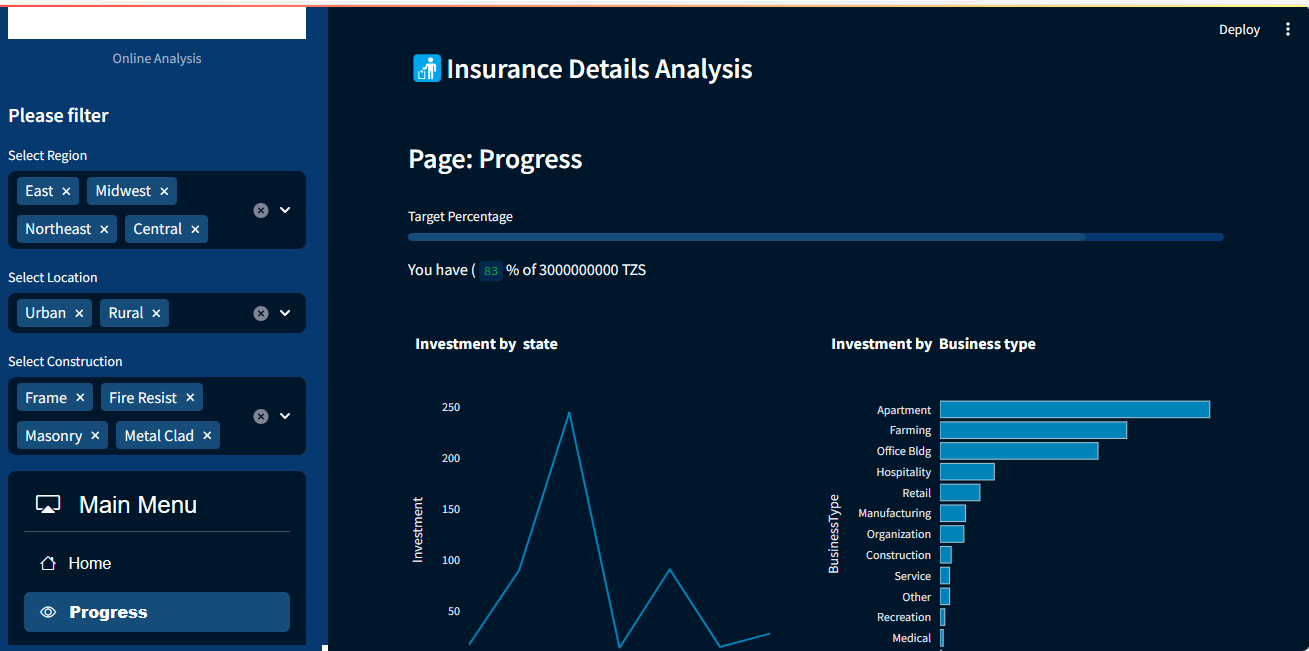
    c.execute('select \* from insurance order by id asc')

    data=c.fetchall()

    return data

**Output:**

****

****

1. **CASE STUDY**

In this 21st century the usage of smartphones has increased exponentially and use of social media too. Concerning future prospects, learning Machine learning and Data Science is strongly recommended for all individuals. And when we come about Machine learning and Data Science, Nodaway’s it become part of all organization. It is part of Artificial Intelligence and whole world is going towards AI and Data Science are used to predict and real-life application. Why machine learning and Dara Science because

* Time effective
* Perform any task automatically
* No need to work yourself etc.
* Easy to predict and classify
* Used in different home appliance and performance is good.
* Used in education, healthcare, security, Autonomous vehicles, Stock prediction etc.

1. **CONCLUSION**

The Machine learning is more useful nowadays as increasing number of people spending time on social media. Data Science used the different Machine learning algorithms to predict and classifier the data set. data science and machine learning represent transformative forces that are reshaping the landscape of industries, decision-making processes, and technological advancements. The fusion of these disciplines has led to unprecedented capabilities in extracting valuable insights from vast and complex datasets, paving the way for innovation and efficiency across diverse sectors.

Looking ahead, the future of data science and machine learning holds promises of even greater advancements. As technology evolves, these fields are poised to tackle increasingly complex challenges, driving innovation, and providing solutions to societal issues.

.

**REFERENCES**

<https://www.ybifoundation.org/#/home>

<https://www.mygreatlearning.com/blog/what-is-data-science/>

<https://cloud.google.com/data-science>