

A SUMMER TRAINING PROJECT

On PREDICTING

*Meteorites Origin Classification (ML)*

*Popular Brand Logos Detection (DL)*

*Person's Disease and Precautions (AI)*

Submitted In partial fulfilment of the requirement

For the award of the degree of

**Master of Computer Application**

To

Guru Gobind Singh Indraprastha University, Delhi



Mr. Prateek Gupta  
(External Guide)

**Submitted By**

Adarsh Maurya

(02013704424)

MCA - II

## ACKNOWLEDGEMENT

I am profoundly thankful to everyone who contributed to the successful completion of my summer training project.

Firstly, I would like to extend my deepest gratitude to **Mr. Prateek Gupta, Sr. Data Scientist**, whose invaluable guidance and insights were essential throughout the training. His expertise and patience were crucial in helping me navigate the complexities of advanced data science and machine learning.

I also wish to sincerely thank IITM Janakpuri for collaborating with ShapeMySkills by DUCAT. to provide this exceptional learning opportunity. Special appreciation goes to our course coordinator, **Mr. Ashish Kumar Nayyar, Training Coordinator** for his dedicated efforts and support during the training. Furthermore, I am deeply grateful to **Dr. Ganesh Kumar Wadhwani, Head of the Department**, for his encouragement and for providing the necessary resources and environment to facilitate this learning experience.

Finally, I acknowledge the unwavering support of my family and friends, who have been my pillars of strength and encouragement throughout this journey.

Thank you all for making this experience valuable and memorable.

Sincerely,

Adarsh Maurya  
02013704424

## CERTIFICATE

This is to certify that Adarsh Maurya (02013704424) has successfully completed the project titled "*Meteorites Origin Classification (ML), Popular Brand Logos Detection (DL), Person's Disease and Precautions (AI)*" as part of the Artificial intelligence and machine learning summer training program organized by IITM Janakpuri in collaboration with ShapeMySkills by DUCAT.

This project was conducted under the esteemed guidance of Mr. Prateek Gupta (Senior Data Scientist), whose expertise and mentorship were instrumental in its successful completion. The project exemplifies a thorough understanding of machine learning techniques, highlighting the skills acquired during the training program.

We commend Adarsh Maurya (02013704424) for his dedication, hard work, and enthusiasm throughout the project duration.

Mr. Prateek Gupta  
(External Guide)

Mr. Ashish Kumar Nayyar  
(Training Coordinator)

(Internal Guide)

Dr. Ganesh Kumar Wadhwani  
(Head of Department)

## **LIST OF FIGURES**

<b>S No.</b>	<b>Name</b>	<b>Page</b>
1	Figure 5.1: Figure represents the Meteorites Dataset Loading	14
2	Figure 5.2: Figure represents the Shape of dataset	15
3	Figure 5.3: Figure represents the Information of dataset	15
4	Figure 5.4: Figure represents the Head of dataset	15
5	Figure 5.5: Figure represents the Tail of dataset	15
6	Figure 5.6: Figure represents the Count of non – null values	16
7	Figure 5.7: Figure represents the Minimum value in each column	16
8	Figure 5.8: Figure represents the Maximum value in each column	16
9	Figure 5.9: Figure represents the Mean	16
10	Figure 5.10: Figure represents the Median	16
11	Figure 5.11: Figure represents the Standard Deviation	17
12	Figure 5.12: Figure represents the Describing the dataset	17
13	Figure 5.13: Figure represents the Renaming the column	18
14	Figure 5.14: Figure represents the Unwanted Columns	18
15	Figure 5.15: Figure represents the Null Values	19
16	Figure 5.16: Figure represents the Duplicates	19
17	Figure 5.17: Figure represents the Scatter Plot	20
18	Figure 5.18: Figure represents the Regression Plot	20
19	Figure 5.19: Figure represents the Box Plot	21
20	Figure 5.20: Figure represents the Heat Map	22
21	Figure 5.21: Figure represents the IQR Method	24
22	Figure 5.22: Figure represents the Z – Score Method	24
23	Figure 5.23: Figure represents the Removing Outliers using the IQR Method	26

24	Figure 5.24: Figure represents the change Categorical into Numerical data.	26
25	Figure 5.25: Figure represents the Iloc Splitting	27
26	Figure 5.26: Figure represents the Train and Test Split	29
27	Figure 5.27: Figure represents the Logistic Regression	30
28	Figure 5.28: Figure represents the Decision Tree Classifier	31
29	Figure 5.29: Figure represents the Random Forest Classifier	33
30	Figure 5.30: Figure represents the Naive Bayes Classifier	34
31	Figure 5.31: Figure represents the Support Vector Machine Classifier	35
32	Figure 5.32: Figure represents the K - Nearest Neighbour Classifier	36
33	Figure 5.33: Figure represents the Comparison of Classifier Algorithm using Bar Chart	39
34	Figure 5.34: Figure represents the Importing Dependencies	40
35	Figure 5.35: Figure represents the Extracting and Verifying Dataset	41
36	Figure 5.36: Figure represents the Generating CSV	42
37	Figure 5.37: Figure represents the Preprocessing	43
38	Figure 5.38: Figure represents the Model Definition	44
39	Figure 5.39: Figure represents the Model Training	45
40	Figure 5.40: Figure represents the Model Accuracy	46
41	Figure 5.41: Figure represents the Classname into label	46
42	Figure 5.42: Figure represents the Result Comparison	47
43	Figure 5.43: Figure represents the Scikit-Learn Installation	48
44	Figure 5.44: Figure represents the Uploading Data to Colab	49
45	Figure 5.45: Figure represents the Reading and model saving	50
46	Figure 5.46: Figure represents the Precaution Dataset Loading	51
47	Figure 5.46: Figure represents the AI Interface	52
48	Figure 5.46: Figure represents the Result and output	53

## **LIST OF TABLES**

<b>S No.</b>	<b>Name</b>	<b>Page</b>
1	Table 5.1: Shows the Comparison Table	37

## **TABLE OF CONTENT**

<b>S. No.</b>	<b>Title</b>	<b>Page No.</b>
1.	Acknowledgements	ii
2.	Certificate	iii
3.	List of Figures	iv-v
4.	List of Tables	vi
<b>5.</b>	<b>Chapter 1:</b> 1.1 Introduction to Python	<b>1</b>
	1.2 Features of python	1
	1.3 Advantages of Python	1
	1.4 Limitation of python	2
	1.5 Application of python	2
	1.6 Python Libraries	3
	1.6.1 NumPy (Numerical Python)	3
	1.6.2 Pandas (Panel Data)	4
	1.6.3 Scikit-learn	4
	1.6.4 Statsmodels	4
	1.6.5 Matplotlib	5
	1.6.6 Seaborn	5
<b>6.</b>	<b>Chapter 2:</b> 2.1 Introduction to Machine Learning	<b>6</b>
	2.2 Types of Machine Learning	6
	2.3 Machine Learning Workflow	7
	2.4 Machine Learning Algorithms	7
	2.5 Advantages of Machine Learning	8

	2.6 Disadvantages of Machine Learning	8
<b>7.</b>	<b>Chapter 3:</b> 3.1 Introduction to Deep Learning	<b>9</b>
	3.2 How Deep Learning Works	9
	3.3 Key Concepts in Deep Learning	9
	3.4 Applications of Deep Learning	10
	3.5 Deep Learning Frameworks	10
	3.6 Deep Learning Workflow	10
<b>8.</b>	<b>Chapter 4:</b> 4.1 Introduction to Artificial Intelligence	<b>11</b>
	4.2 Types of Artificial Intelligence	11
	4.3 How AI Works	11
	4.4 Key AI Techniques	12
	4.5 Applications of Artificial Intelligence	12
	4.6 Advantages of Artificial Intelligence	12
	4.7 Disadvantages of Artificial Intelligence	12
<b>9.</b>	<b>Chapter 5: Results and Discussions</b>	<b>14</b>
	5.1 Machine Learning Project	14-39
	5.2 Deep Learning Project	40-48
	5.3 Artificial Intelligence Project	49-54
<b>10.</b>	<b>Chapter 6: Conclusion</b>	<b>55</b>
<b>11.</b>	<b>Chapter 7: Future Scope</b>	<b>56-57</b>
<b>12.</b>	<b>Chapter 8: References</b>	<b>58</b>



# **Chapter 1**

## **Python and Its Libraries**

### **1.1 Introduction to Python**

Python is a high-level, interpreted, and general-purpose programming language that emphasizes code readability and developer productivity. It was created by Guido van Rossum and released in 1991. Over the years, Python has gained immense popularity due to its simplicity, clean syntax, and versatility. The language supports multiple programming paradigms, including object-oriented, procedural, and functional programming, making it suitable for a wide range of applications.

Python has become the language of choice for various fields such as web development, data science, artificial intelligence, automation, and more. Its open-source nature and strong community support have contributed significantly to the development of an extensive ecosystem of third-party libraries and frameworks.

### **1.2 Features of Python**

Python's popularity is attributed to a range of powerful features, including:

#### **1.2.1 Easy to Learn and Use**

Python's syntax is simple and readable, making it accessible for beginners and efficient for experienced developers.

#### **1.2.2 Interpreted Language**

Python code is executed line-by-line, which simplifies debugging and testing processes.

#### **1.2.3 Dynamically Typed**

Variable types are determined at runtime, eliminating the need for explicit declarations.

#### **1.2.4 Cross-Platform Compatibility**

Python runs seamlessly on various operating systems such as Windows, Linux, and macOS without requiring changes to the code.

#### **1.2.5 Extensive Standard Library**

Python includes a rich standard library that provides modules for file handling, regular expressions, threading, databases, and more.

#### **1.2.6 Community Support**

A vast and active community ensures continuous development, problem-solving, and resource availability.

### **1.3 Advantages of Python**

Python offers numerous benefits, which have contributed to its widespread adoption:

- **Beginner-Friendly:** Its clean and concise syntax makes it ideal for those new to programming.
- **Rapid Development:** Facilitates fast prototyping and iterative development.
- **Extensive Libraries:** Access to a wide array of libraries for data manipulation, scientific computing, machine learning, and more.
- **Integration Capabilities:** Can integrate with other languages like C, C++, and Java.
- **Strong Community and Documentation:** Comprehensive tutorials, guides, and forums available online.
- **Versatility:** Suitable for web applications, automation, data analysis, artificial intelligence, and more.

## 1.4 Limitations of Python

Despite its advantages, Python has some limitations:

- **Slower Execution Speed:** Being an interpreted language, it is slower compared to compiled languages like C or Java.
- **Memory Consumption:** Python programs can consume more memory, which may be inefficient for certain tasks.
- **Not Ideal for Mobile Development:** It is less commonly used for mobile application development due to limited support.
- **Multithreading Limitations:** Python's Global Interpreter Lock (GIL) restricts concurrent execution of multiple threads.
- **Runtime Errors:** Dynamic typing can lead to type-related runtime errors.

## 1.5 Applications of Python

Python is used across a wide variety of domains. Some of its major application areas include:

### 1.5.1 Web Development

Python supports frameworks such as Django and Flask, which facilitate the development of dynamic and secure web applications.

### 1.5.2 Data Science and Analytics

Libraries like Pandas and NumPy enable data manipulation, statistical analysis, and data visualization.

### 1.5.3 Machine Learning and Artificial Intelligence

Python is the dominant language in the AI/ML community. Libraries such as scikit-learn, TensorFlow, and PyTorch are extensively used for building and training models.

### **1.5.4 Automation and Scripting**

Python is widely used to automate repetitive tasks, file operations, and testing processes.

### **1.5.5 Game Development**

Libraries like Pygame support the development of 2D games and interactive applications.

### **1.5.6 Desktop Applications**

Frameworks such as Tkinter and PyQt help in building user-friendly graphical user interfaces.

### **1.5.7 Cybersecurity and Ethical Hacking**

Python is employed in penetration testing, vulnerability scanning, and malware analysis.

### **1.5.8 IoT and Embedded Systems**

MicroPython and CircuitPython extend Python to microcontrollers and embedded platforms.

## **1.6 Python Libraries**

Python's strength lies in its vast collection of libraries that enhance functionality and reduce development time. Some of the most commonly used libraries are:

### **1.6.1 NumPy (Numerical Python)**

NumPy is the foundational library for numerical computations. It provides support for multi-dimensional arrays and matrices, along with a large collection of mathematical functions.

#### **Key Features:**

- Efficient array operations
- Broadcasting and vectorization
- Linear algebra and statistical functions
- Basis for other libraries like SciPy, Pandas, and scikit-learn

#### **Use Cases:**

- Scientific computing
- Data analysis
- Signal processing

### **1.6.2 Pandas (Panel Data)**

Pandas is built on NumPy and offers powerful tools for data manipulation and analysis. It introduces two primary data structures: Series (one-dimensional) and DataFrame (two-dimensional).

**Key Features:**

- Easy handling of missing data
- Data filtering, grouping, merging, and reshaping
- Rich input/output tools for reading and writing data
- Integration with time-series data

**Use Cases:**

- Exploratory data analysis (EDA)
- Data cleaning and transformation
- Business intelligence

**1.6.3 scikit-learn**

scikit-learn is a machine learning library built on NumPy, SciPy, and Matplotlib. It provides simple and efficient tools for data mining and data analysis.

**Key Features:**

- Support for classification, regression, clustering, and dimensionality reduction
- Model selection and evaluation tools
- Preprocessing utilities for scaling and encoding
- Consistent API and documentation

**Use Cases:**

- Predictive analytics
- Model development and validation
- Feature selection

**1.6.4 Statsmodels**

Statsmodels is used for statistical modeling and testing. It provides classes and functions for estimating statistical models and performing hypothesis tests.

**Key Features:**

- Linear and generalized linear models (GLMs)
- Time series analysis and forecasting
- Hypothesis testing (t-test, chi-square, etc.)
- Compatibility with Pandas

**Use Cases:**

- Econometrics

- Data interpretation and inference
- Academic research

### **1.6.5 Matplotlib**

Matplotlib is a versatile plotting library that enables the creation of static, interactive, and animated visualizations in Python.

#### **Key Features:**

- Line, bar, scatter, histogram, and pie charts
- Full control over plot aesthetics
- Export in various formats (PNG, SVG, PDF)

#### **Use Cases:**

- Scientific visualization
- Data reporting and presentations
- Exploratory analysis

### **1.6.6 Seaborn**

Seaborn is built on top of Matplotlib and simplifies the process of creating attractive and informative statistical graphics.

#### **Key Features:**

- High-level interface for complex visualizations
- Built-in themes for aesthetics
- Integration with Pandas DataFrames
- Automatic aggregation and statistical estimation

#### **Use Cases:**

- Correlation analysis
- Heatmaps, boxplots, and violin plots
- Enhanced visual storytelling in research

## **Chapter 2**

### **Introduction to Machine Learning**

#### **2.1 What is Machine Learning?**

Machine Learning (ML) is a subfield of artificial intelligence that enables machines to automatically learn patterns from data and improve their performance on a specific task without being explicitly programmed. The fundamental goal of machine learning is to create systems that can generalize from past experiences (data) to make accurate predictions or decisions when presented with new, unseen information.

ML algorithms use mathematical models to identify patterns in data, and over time, refine their predictions through a process of training, validation, and testing. It is widely used in domains such as healthcare, finance, e-commerce, and robotics.

#### **2.2 Types of Machine Learning**

Machine learning techniques are generally categorized into three main types:

##### **2.2.1 Supervised Learning**

In supervised learning, the algorithm is trained on a labeled dataset, where both the input and the expected output are provided. The goal is to learn a mapping from inputs to outputs that can be used to predict future data.

**Examples:** Predicting house prices, spam detection, loan approval.

##### **2.2.2 Unsupervised Learning**

In this type, the data does not contain labeled outputs. The algorithm tries to learn patterns or groupings inherent in the input data.

**Examples:** Customer segmentation, anomaly detection, clustering.

##### **2.2.3 Reinforcement Learning**

Here, an agent learns to make decisions by interacting with an environment and receiving feedback in the form of rewards or penalties. It aims to find a strategy that maximizes cumulative reward.

**Examples:** Game playing, robotics, autonomous driving.

## 2.3 Machine Learning Workflow

A standard ML process includes the following stages:

1. **Data Collection:** Gather relevant datasets for the problem at hand.
2. **Data Cleaning:** Handle missing, inconsistent, or noisy data.
3. **Data Visualization:** Plot and understand trends, patterns, and outliers.
4. **Feature Selection and Engineering:** Select and transform variables that influence predictions.
5. **Train-Test Split:** Divide data into training and testing sets.
6. **Model Selection and Training:** Choose an appropriate algorithm and train the model.
7. **Evaluation:** Use metrics such as accuracy, precision, recall, or  $R^2$  score.
8. **Prediction and Deployment:** Apply the trained model to new data and deploy it in real-time applications.

## 2.4 Popular Machine Learning Algorithms

### 2.4.1 Linear Regression

Used for predicting continuous values by modeling the relationship between dependent and independent variables using a straight line.

### 2.4.2 Logistic Regression

Used for binary classification problems (e.g., yes/no, spam/not spam).

### 2.4.3 Decision Trees

A flowchart-like structure where internal nodes represent tests on features, and leaf nodes represent outputs.

### 2.4.4 Random Forest

An ensemble technique that combines multiple decision trees to improve accuracy and avoid overfitting.

### 2.4.5 K-Nearest Neighbors (KNN)

Classifies new data based on the majority class among its k-nearest neighbors in the dataset.

### 2.4.6 Naive Bayes

A probabilistic classifier based on Bayes' theorem. Assumes independence among predictors.

### 2.4.7 Support Vector Machines (SVM)

Finds the optimal hyperplane that separates data points into distinct classes with maximum margin.

## **2.5 Advantages of Machine Learning**

- Learns and adapts from new data.
- Automates complex tasks and reduces manual work.
- Performs well with large and high-dimensional datasets.
- Applicable in diverse fields such as finance, marketing, and medicine.

## **2.6 Disadvantages of Machine Learning**

- Requires large, high-quality datasets for training.
- Computationally intensive and time-consuming.
- Difficult to interpret some models (e.g., deep learning).
- Vulnerable to data bias and overfitting.



## Chapter 3

### Introduction to Deep Learning

#### 3.1 What is Deep Learning?

Deep Learning is a subset of machine learning that uses algorithms known as artificial neural networks, which are inspired by the structure and functioning of the human brain. These models are capable of learning complex patterns from large amounts of data by using multiple layers of abstraction, which is why the term “deep” is used.

Deep learning has revolutionized areas such as image recognition, natural language processing, and autonomous systems.

#### 3.2 How Deep Learning Works

Deep learning models consist of:

- **Input Layer:** Takes the raw data.
- **Hidden Layers:** Multiple intermediate layers that transform the input data into higher-level features.
- **Output Layer:** Produces the final prediction or classification.

Each neuron in a layer is connected to the neurons in the next layer. These connections have weights that are adjusted during training.

The model is trained using **backpropagation**, a technique that minimizes the difference between the predicted output and the actual output by adjusting the weights using gradient descent.

#### 3.3 Key Concepts in Deep Learning

##### 3.3.1 Neural Networks

Artificial neural networks are composed of nodes (neurons) arranged in layers. These networks can model non-linear relationships and are highly flexible.

##### 3.3.2 Activation Functions

They introduce non-linearity into the network. Common activation functions include:

- ReLU (Rectified Linear Unit)
- Sigmoid
- Tanh

##### 3.3.3 Loss Functions

Loss functions measure the model's performance. Examples:

- Mean Squared Error (MSE) for regression
- Cross-Entropy Loss for classification

##### 3.3.4 Regularization Techniques

Methods like **dropout** and **early stopping** help prevent overfitting by limiting the model's capacity to memorize training data.

### 3.4 Applications of Deep Learning

- **Image Recognition:** Identifying objects and faces in images.
- **Speech Recognition:** Converting speech to text in digital assistants.
- **Natural Language Processing (NLP):** Translation, summarization, sentiment analysis.
- **Medical Diagnosis:** Analyzing X-rays, MRIs, and pathology slides.
- **Recommendation Systems:** Personalized content suggestions on streaming platforms.

### 3.5 Deep Learning Frameworks

#### 3.5.1 TensorFlow

An open-source framework developed by Google for building and training deep learning models. It supports GPU acceleration, distributed training, and visualization through TensorBoard.

#### 3.5.2 Keras

A high-level API built on top of TensorFlow that simplifies the creation of neural networks. It offers a user-friendly syntax and is ideal for rapid prototyping.

### 3.6 Deep Learning Workflow

1. Collect and preprocess data.
2. Define the neural network architecture.
3. Compile the model (select optimizer and loss function).
4. Train the model using labeled data.
5. Evaluate model performance on test data.
6. Deploy the model for real-time predictions.

## **Chapter 4**

### **Introduction to Artificial Intelligence**

#### **4.1 What is Artificial Intelligence?**

Artificial Intelligence (AI) is a field of computer science focused on creating machines that exhibit behaviours associated with human intelligence. These behaviours include learning, reasoning, problem-solving, understanding language, and perception.

AI systems can perform tasks that typically require human cognition, such as interpreting data, recognizing speech and images, making decisions, and translating languages.

#### **4.2 Types of Artificial Intelligence**

##### **4.2.1 Narrow AI**

Also known as weak AI, it is designed to perform a specific task or a narrow range of tasks. Most current AI systems fall under this category.

**Examples:** Voice assistants, chatbots, recommendation systems.

##### **4.2.2 General AI**

A theoretical form of AI that can perform any intellectual task a human can. This level of AI does not yet exist.

##### **4.2.3 Reactive Machines**

These systems respond to specific inputs but do not store any past experiences.

##### **4.2.4 Limited Memory**

These systems can use past data to make decisions, such as self-driving cars analyzing recent traffic data.

##### **4.2.5 Theory of Mind**

An advanced AI concept where machines would be able to understand human emotions, beliefs, and intentions.

##### **4.2.6 Self-Aware AI**

The most advanced theoretical form of AI, where machines have consciousness and self-awareness.

#### **4.3 How AI Works**

AI systems function through a combination of data, algorithms, and computation. The general process includes:

- **Data Collection:** Gathering relevant information.
- **Preprocessing:** Cleaning and transforming the data.

- **Training:** Feeding data into machine learning or deep learning models.
- **Inference:** Making predictions on new data.
- **Learning and Adaptation:** Updating the model based on feedback and new inputs.

Tools commonly used in AI include Python, Scikit-learn, TensorFlow, and Pandas.

#### 4.4 Key AI Techniques

- **Machine Learning:** Uses algorithms to learn patterns from data.
- **Deep Learning:** Employs neural networks for complex problem solving.
- **Natural Language Processing (NLP):** Enables machines to understand and generate human language.
- **Computer Vision:** Interprets and processes visual information.
- **Robotics:** Combines AI with hardware to perform physical tasks autonomously.

#### 4.5 Applications of Artificial Intelligence

- **Virtual Assistants:** Such as Siri and Alexa.
- **Autonomous Vehicles:** AI enables self-driving cars to perceive surroundings and make decisions.
- **Healthcare:** AI supports diagnosis, treatment recommendations, and medical imaging analysis.
- **Finance:** Fraud detection, algorithmic trading, and customer service bots.
- **E-Commerce:** Personalized recommendations and dynamic pricing.
- **Smart Homes:** Automated control of lighting, temperature, and appliances.

#### 4.6 Advantages of Artificial Intelligence

- Automates repetitive and complex tasks.
- Enhances decision-making through data-driven insights.
- Learns and improves with experience.
- Efficiently processes large volumes of data.

#### 4.7 Disadvantages of Artificial Intelligence

- High dependency on large, high-quality datasets.
- Expensive computational and infrastructure requirements.
- Raises ethical and privacy concerns.

- Lack of transparency in decision-making (black-box models).
- Potential to displace human jobs in certain sectors.

## CHAPTER - 5

### RESULT AND DISCUSSION

#### 5.1 Machine Learning Project

Prediction of Origin of Meteorites

##### 5.1.1 Project Aim

To predict the origin of meteorites based on their physical properties using classification algorithms.

##### 5.1.2 Loading the Dataset

In this step, the dataset is imported into the Python environment using the pandas library. The data file (usually in .csv format) is loaded into a DataFrame so that it can be easily viewed and used for further steps like cleaning, visualization, and model training.

```
# Importing the library
import pandas as pd
df = pd.read_csv(r"C:\Users\91892\Downloads\meteorite_classification_dataset.csv")
df
```

	Unnamed: 0	Weight (kg)	Velocity (km/s)	Metallic Content (%)	Magnetic Strength (T)	Origin
0	4922	NaN	10.821243	88.274805	0.404645	Asteroidal
1	539	43.106142	24.964433	50.336015	0.687528	Lunar
2	2480	NaN	14.833294	81.801750	0.367695	Asteroidal
3	2469	NaN	21.145692	65.705708	0.552385	Martian
4	1703	51.120673	21.401973	NaN	0.740538	Lunar
...	...	...	...	...	...	...
5770	1472	49.785287	27.214585	133.824575	NaN	Lunar
5771	4176	66.131661	13.743410	180.508899	0.437541	Asteroidal
5772	3432	46.751220	22.188273	115.498628	0.718928	Lunar
5773	167	NaN	12.172203	205.919258	0.492338	Asteroidal
5774	3470	40.802312	22.264605	45.492852	0.730193	Lunar

5775 rows × 6 columns

**FIG 5.1: METEORITES DATASET LOADING**

### 5.1.2 Information Gathering:

In this step, we explore the dataset to understand its structure, features, and basic statistics. It includes checking the number of rows and columns, data types, missing values, and getting an overall idea of what the dataset contains before preprocessing or model building.

1. **Shape** : `df.shape` returns the total number of rows and columns in the dataset. It helps us quickly know the size of the data.

```
# Getting number of rows and columns
df.shape

(5775, 6)
```

FIG 5.2: SHAPE

2. **Information** : `df.info()` provides important information such as column names, data types, and the number of non-null values in each column. It is useful for identifying missing data and understanding the structure of the dataset.

```
# Checking data types and non-null values count in the dataset
df.info()

<class 'pandas.core.frame.DataFrame'>
RangeIndex: 5775 entries, 0 to 5774
Data columns (total 6 columns):
 #   Column              Non-Null Count  Dtype  
---  -
 0   Unnamed: 0          5775 non-null  int64  
 1   Weight (kg)         4902 non-null  float64
 2   Velocity (km/s)     4902 non-null  float64
 3   Metallic Content (%) 4906 non-null  float64
 4   Magnetic Strength (T) 4905 non-null  float64
 5   Origin              5775 non-null  object 
dtypes: float64(4), int64(1), object(1)
memory usage: 270.8+ KB
```

FIG 5.3: INFORMATION

3. **Head** : `df.head()` gives the top 5 rows of the dataset .

```
# Top 5 rows of the dataset
df.head()
```

	Unnamed: 0	Weight (kg)	Velocity (km/s)	Metallic Content (%)	Magnetic Strength (T)	Origin
0	4922	NaN	10.821243	88.274805	0.404645	Asteroidal
1	539	43.106142	24.964433	50.336015	0.687528	Lunar
2	2480	NaN	14.833294	81.801750	0.367695	Asteroidal
3	2469	NaN	21.145692	65.705708	0.552385	Martian
4	1703	51.120673	21.401973	NaN	0.740538	Lunar

FIG 5.4: HEAD

4. **Tail** : `df.tail()` gives the bottom 5 rows of the dataset.

```
# Bottom 5 rows of the dataset
df.tail()
```

	Unnamed: 0	Weight (kg)	Velocity (km/s)	Metallic Content (%)	Magnetic Strength (T)	Origin
5770	1472	49.785287	27.214585	133.824575	NaN	Lunar
5771	4176	66.131661	13.743410	180.508899	0.437541	Asteroidal
5772	3432	46.751220	22.188273	115.498628	0.718928	Lunar
5773	167	NaN	12.172203	205.919258	0.492338	Asteroidal
5774	3470	40.802312	22.264605	45.492852	0.730193	Lunar

FIG 5.5: TAIL

5. **Count** : `df.count()` gives the number of non-null (filled) entries in each column. This helps identify which columns have missing values.

```
# Counting non-null values in each column
df.count()

Unnamed: 0          5775
Weight (kg)         4902
Velocity (km/s)     4902
Metallic Content (%) 4906
Magnetic Strength (T) 4905
Origin              5775
dtype: int64
```

FIG 5.6: COUNT

6. **Minimum** : `df.min()` show the minimum values of each numeric column, respectively.

```
# Minimum value in each column
df.min()

Unnamed: 0          0
Weight (kg)        36.046345
Velocity (km/s)     8.764348
Metallic Content (%) 34.870838
Magnetic Strength (T) 0.244
Origin             Asteroidal
dtype: object
```

FIG 5.7: MINIMUM

7. **Maximum** : `df.max()` show the maximum values of each numeric column, respectively.

```
# Maximum value in each column
df.max()

Unnamed: 0          5499
Weight (kg)        69.632001
Velocity (km/s)     31.392793
Metallic Content (%) 242.888867
Magnetic Strength (T) 0.853262
Origin             Martian
dtype: object
```

FIG 5.8: MAXIMUM

8. **Mean** : `df.mean()` calculates the average value of each numeric column, giving us a central value for comparison.

```
# Mean value of the Metallic Content (%) column
df['Metallic Content (%)'].mean()

89.63750059462132
```

FIG 5.9: MEAN

9. **Median** : `df.median()` returns the middle value of each numeric column, which is less sensitive to extreme values (outliers).

```
# Median value of the Metallic Content (%) column
df['Metallic Content (%)'].median()

72.24633211026891
```

FIG 5.10: MEDIAN



- 10. Standard Deviation :** `df.std()`, or standard deviation, measures how much the data varies from the mean. A higher standard deviation indicates greater spread in the data.

```
# Standard deviation of the Metallic Content (%) column
df['Metallic Content (%)'].std()
```

47.330269691693786

FIG 5.11: STANDARD DEVIATION

- 11. Describe :** `df.describe()` provides a summary of key statistical measures for each numeric column, including count, mean, standard deviation, minimum, 25th percentile, median (50th percentile), 75th percentile, and maximum. It is one of the most commonly used functions for initial data exploration.

```
# Describing the dataset
df.describe()
```

	Unnamed: 0	Weight (kg)	Velocity (km/s)	Metallic Content (%)	Magnetic Strength (T)
count	5775.000000	4902.000000	4902.000000	4906.000000	4905.000000
mean	2746.927100	53.269275	19.949355	89.637501	0.550306
std	1586.749082	7.022404	4.528869	47.330270	0.133237
min	0.000000	36.046345	8.764348	34.870838	0.244000
25%	1373.500000	46.747039	16.113816	57.172800	0.433584
50%	2745.000000	54.636884	19.920738	72.246332	0.550352
75%	4117.500000	58.853084	23.803757	101.660296	0.669880
max	5499.000000	69.632001	31.392793	242.888867	0.853262

FIG 5.12: DESCRIBE

### 5.1.3 Data Cleaning:

Data cleaning is the process of correcting or removing inaccurate, incomplete, or duplicate data to improve the quality and reliability of the dataset.

- 1. Renaming the Column :** Renaming a column means changing the name of one or more columns in a dataset to make them more meaningful, readable, or standardized for analysis and modeling.

```
# Renaming the Column Weight (kg)
df.rename(columns={'Weight (kg)': 'Weight'}, inplace=True)
df
```

	Unnamed: 0	Weight	Velocity (km/s)	Metallic Content (%)	Magnetic Strength (T)	Origin
0	4922	NaN	10.821243	88.274805	0.404645	Asteroidal
1	539	43.106142	24.964433	50.336015	0.687528	Lunar

```
# Renaming the Column Velocity (km/s)
df.rename(columns={'Velocity (km/s)': 'Velocity'}, inplace=True)
df
```

	Unnamed: 0	Weight	Velocity	Metallic Content (%)	Magnetic Strength (T)	Origin
0	4922	NaN	10.821243	88.274805	0.404645	Asteroidal
1	539	43.106142	24.964433	50.336015	0.687528	Lunar

```
# Renaming the Column Metallic Content (%)
df.rename(columns={'Metallic Content (%)': 'Metallic_Content'}, inplace=True)
df
```

	Unnamed: 0	Weight	Velocity	Metallic_Content	Magnetic Strength (T)	Origin
0	4922	NaN	10.821243	88.274805	0.404645	Asteroidal
1	539	43.106142	24.964433	50.336015	0.687528	Lunar
2	2480	NaN	14.833294	81.801750	0.367695	Asteroidal

```
# Renaming the Column Magnetic Strength (T)
df.rename(columns={'Magnetic Strength (T)': 'Magnetic_Strength'}, inplace=True)
df
```

	Unnamed: 0	Weight	Velocity	Metallic_Content	Magnetic_Strength	Origin
0	4922	NaN	10.821243	88.274805	0.404645	Asteroidal
1	539	43.106142	24.964433	50.336015	0.687528	Lunar
2	2480	NaN	14.833294	81.801750	0.367695	Asteroidal

FIG 5.13: RENAMING THE COLUMN

- Unwanted Columns :** Unwanted columns are those that do not contribute to the analysis or model and are removed to simplify the dataset.

```
# Removing an unwanted column
df.drop(columns=['Unnamed: 0'], inplace=True)
df
```

	Weight	Velocity	Metallic_Content	Magnetic_Strength	Origin
0	NaN	10.821243	88.274805	0.404645	Asteroidal
1	43.106142	24.964433	50.336015	0.687528	Lunar
2	NaN	14.833294	81.801750	0.367695	Asteroidal
3	NaN	21.145692	65.705708	0.552385	Martian
4	51.120673	21.401973	NaN	0.740538	Lunar

FIG 5.14: UNWANTED COLUMNS

- Null Values :** Null values represent missing or empty data in the dataset and need to be handled to avoid errors during processing or model training.

```
# Total null values in each column
df.isnull().sum()
```

Weight	873
Velocity	873
Metallic_Content	869
Magnetic_Strength	870
Origin	0
dtype:	int64

```
# Check percentage of null values in each column
null_percentage = (df.isnull().sum() / len(df)) * 100
print(null_percentage)
```

Weight	15.116883
Velocity	15.116883
Metallic_Content	15.047619
Magnetic_Strength	15.064935
Origin	0.000000
dtype:	float64

### Filling the null values :

```
# Filling null values with median of each column
df['Weight'].fillna(df.Weight.median(),inplace=True)
df['Velocity'].fillna(df.Velocity.median(),inplace=True)
df['Metallic_Content'].fillna(df.Metallic_Content.median(),inplace=True)
df['Magnetic_Strength'].fillna(df.Magnetic_Strength.median(),inplace=True)
df
```

	Weight	Velocity	Metallic_Content	Magnetic_Strength	Origin
0	54.636884	10.821243	88.274805	0.404645	Asteroidal
1	43.106142	24.964433	50.336015	0.687528	Lunar
2	54.636884	14.833294	81.801750	0.367695	Asteroidal
3	54.636884	21.145692	65.705708	0.552385	Martian
4	51.120673	21.401973	72.246332	0.740538	Lunar
...	...	...	...	...	...

FIG 5.15: NULL VALUES

- Duplicates :** Duplicate rows are repeated entries in the dataset that can affect the accuracy of the model and are usually removed during data cleaning.

```
# Count duplicate rows
df.duplicated().sum()
```

181

### Removing the duplicates :

```
# Remove duplicate rows
df.drop_duplicates(inplace=True)
df
```

	Weight	Velocity	Metallic_Content	Magnetic_Strength	Origin
0	54.636884	10.821243	88.274805	0.404645	Asteroidal
1	43.106142	24.964433	50.336015	0.687528	Lunar
2	54.636884	14.833294	81.801750	0.367695	Asteroidal
3	54.636884	21.145692	65.705708	0.552385	Martian
4	51.120673	21.401973	72.246332	0.740538	Lunar
...	...	...	...	...	...
5769	57.580559	22.527021	58.888648	0.550352	Martian
5770	49.785287	27.214585	133.824575	0.550352	Lunar
5772	46.751220	22.188273	115.498628	0.718928	Lunar
5773	54.636884	12.172203	205.919258	0.492338	Asteroidal
5774	40.802312	22.264605	45.492852	0.730193	Lunar

### Resetting the index :

```
# Resetting the index after dropping duplicates
df.reset_index(drop=True, inplace=True)
df
```

	Weight	Velocity	Metallic_Content	Magnetic_Strength	Origin
0	54.636884	10.821243	88.274805	0.404645	Asteroidal
1	43.106142	24.964433	50.336015	0.687528	Lunar
2	54.636884	14.833294	81.801750	0.367695	Asteroidal
3	54.636884	21.145692	65.705708	0.552385	Martian
4	51.120673	21.401973	72.246332	0.740538	Lunar
...	...	...	...	...	...
5589	57.580559	22.527021	58.888648	0.550352	Martian
5590	49.785287	27.214585	133.824575	0.550352	Lunar
5591	46.751220	22.188273	115.498628	0.718928	Lunar
5592	54.636884	12.172203	205.919258	0.492338	Asteroidal
5593	40.802312	22.264605	45.492852	0.730193	Lunar

5594 rows × 5 columns

FIG 5.16: DUPLICATES

### 5.1.4 Data Visualization:

Data visualization is the process of representing data in graphical form (like charts, graphs, and plots) to better understand patterns, trends, and relationships within the dataset.

Types of Data Visualization :

1. **Scatter Plot** : A scatter plot displays the relationship between two numerical variables using dots, helping to identify patterns or correlations.

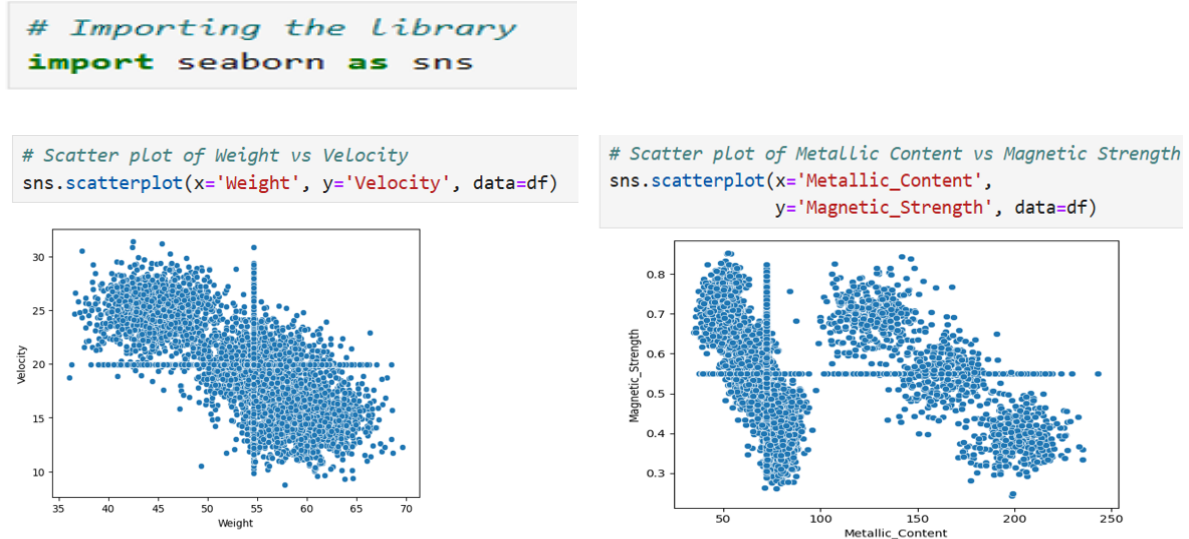


FIG 5.17: SCATTER PLOT

2. **Regression Plot** : A regression plot shows the best-fit line through the data points in a scatter plot, used to visualize the relationship and prediction trend.

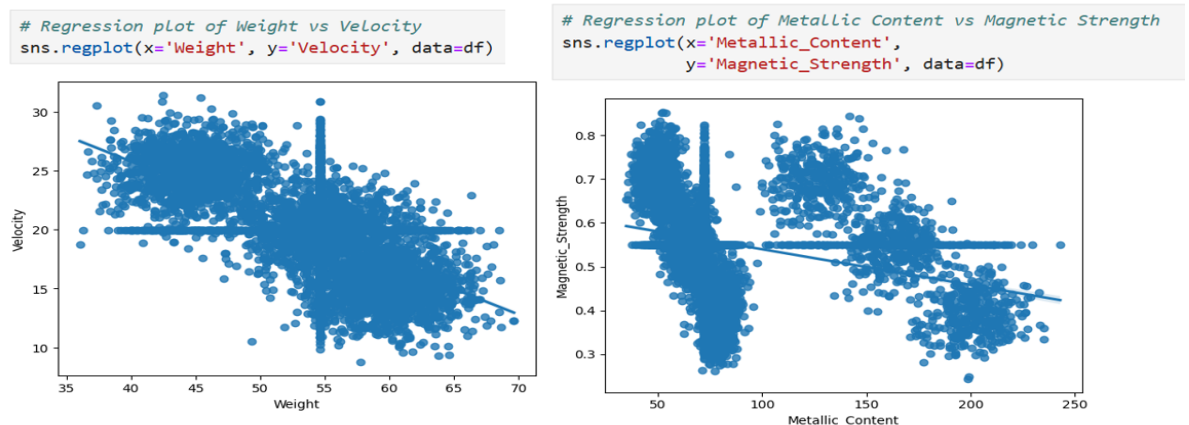


FIG 5.18: REGRESSION PLOT

3. **Box Plot** : A box plot represents the distribution of a dataset based on five summary statistics—minimum, first quartile, median, third quartile, and maximum—while also showing outliers.

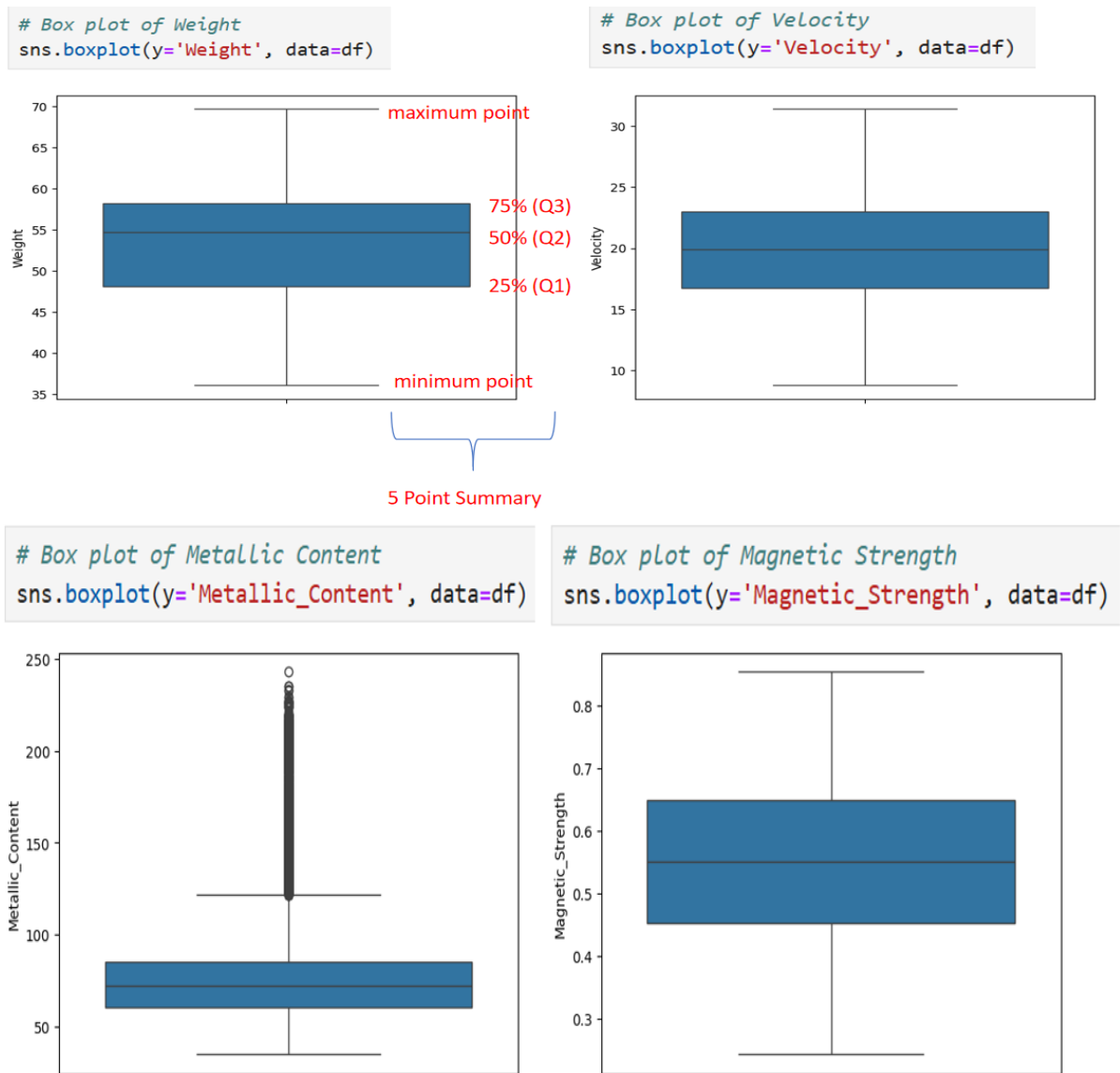


FIG 5.19: BOX PLOT

4. **Heat Map** : A heat map uses color to represent values in a matrix or table, often used to visualize correlations between multiple features.



```
#Heat Map
sns.heatmap(df.corr(numeric_only=True),annot=True)
x=df[['Weight','Velocity','Metallic_Content','Magnetic_Strength']]
y=df['Origin']
```

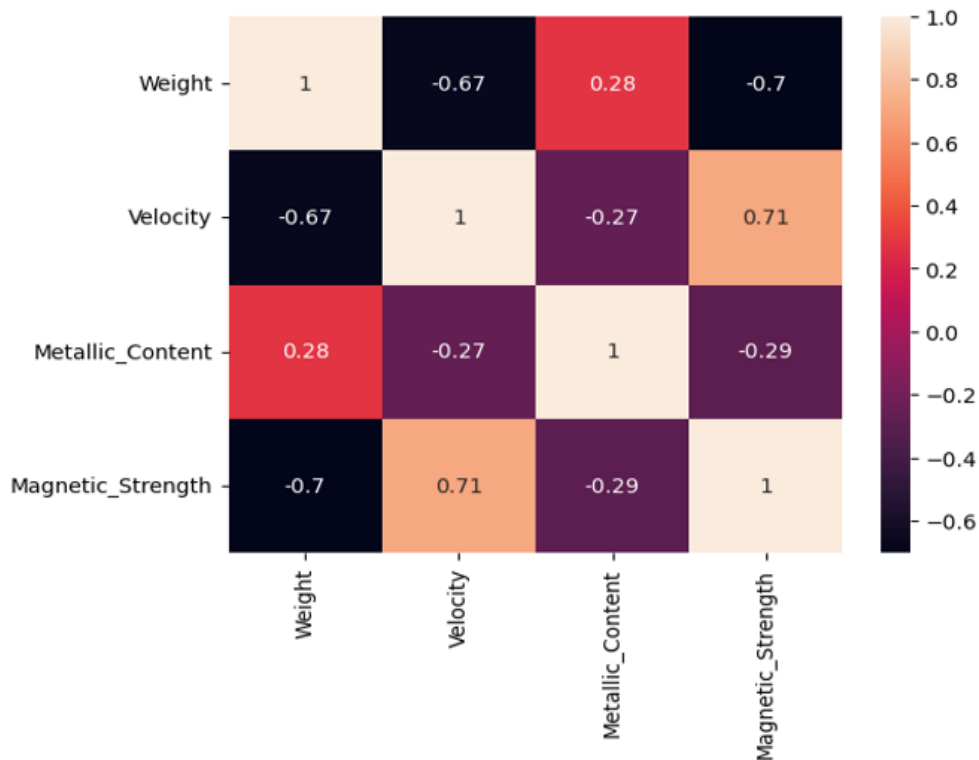


FIG 5.20: HEAT MAP

### 5.1.5 Detecting Outliers in the Dataset:

1. **IQR Method** : The Interquartile Range (IQR) method identifies outliers by measuring the spread of the middle 50% of the data. Any value below  $Q1 - 1.5 \times IQR$  or above  $Q3 + 1.5 \times IQR$  is considered an outlier.

```

# Importing the library
import pandas as pd
import statistics as stats

# Define the function
def find_outliers_iqr(df, columns):

    any_outliers = False # Flag to track if at least one column has outliers

    # Loop through each selected column
    for column in columns:

        # Getting the column data as a sorted list
        values = sorted(df[column].tolist())

        # Calculating the median (Q2)
        q2 = stats.median(values)

        # Splitting data into lower and upper halves
        n = len(values)
        mid = n // 2

        if n % 2 == 0: # even number of elements
            lower_half = values[:mid]
            upper_half = values[mid:]
        else: # odd number of elements
            lower_half = values[:mid]
            upper_half = values[mid+1:]

        # Calculating Q1 and Q3
        q1 = stats.median(lower_half) if lower_half else 0
        q3 = stats.median(upper_half) if upper_half else 0

        # Calculating the Interquartile Range (IQR)
        iqr = q3 - q1

        # Determining bounds for outlier detection
        lower_bound = q1 - 1.5 * iqr
        upper_bound = q3 + 1.5 * iqr

        # Identifying the outliers
        outliers = []
        for value in values:
            if value < lower_bound or value > upper_bound:
                outliers.append(value)

        # Count outliers and total values to calculate percentage
        outlier_count = len(outliers)
        total_count = len(values)

        # Printing the result
        if outliers:
            any_outliers = True
            percent_outliers = (outlier_count / total_count) * 100
            print(f"\nColumn : {column}")
            print(f"Q1 : {q1} , Q3 : {q3} , IQR : {iqr:.2f}")
            print(f"Outlier condition : value < {lower_bound:.2f} or value > {upper_bound:.2f}")
            print(f"\nOutliers detected : {outliers}")
            print(f"\nTotal data points: {total_count}, Outliers: {outlier_count}")
            print(f"Outlier Percentage: {percent_outliers:.2f}%\n")

        if not any_outliers:
            print("No outliers found in any column.")

# Calling the function
find_outliers_iqr(df, ['Weight', 'Velocity', 'Metallic_Content', 'Magnetic_Strength'])

```

```

Column : Metallic_Content
Q1 : 60.43893730669026 , Q3 : 84.98754308433111 , IQR : 24.55
Outlier condition : value < 23.62 or value > 121.81

Total data points: 5594, Outliers: 1074
Outlier Percentage: 19.20%

```

FIG 5.21: IQR METHOD

2. **Z - Score** : The Z-score measures how many standard deviations a data point is from the mean. A value with a Z-score greater than 3 or less than -3 is typically treated as an outlier.

```

# Importing the library
import numpy as np

# Define the function
def find_outliers_zscore(df, columns, threshold=3):

    any_outliers = False # Flag to track if at least one column has outliers

    # Loop through each selected column
    for column in columns:

        # Getting the column data as a list
        values = df[column].tolist()

        # Calculating mean and standard deviation
        mean = round(np.mean(values), 2)
        std_dev = round(np.std(values), 2)

        # Identifying the outliers
        outliers = []
        for value in values:
            z_score = round((value - mean) / std_dev, 2)
            if z_score > threshold or z_score < -threshold:
                outliers.append(value)

        # Count outliers and total values to calculate percentage
        outlier_count = len(outliers)
        total_count = len(values)

        # Printing the result
        if outliers:
            any_outliers = True
            percent_outliers = (outlier_count / total_count) * 100
            print(f"\nColumn : {column}")
            print(f"Mean : {mean} , Standard Deviation : {std_dev}")
            print(f"Outlier condition : Z-score > {threshold} or Z-score < -{threshold}")
            print(f"\nOutliers detected : {outliers}\n")
            print(f"\nTotal data points: {total_count}, Outliers: {outlier_count}")
            print(f"Outlier Percentage: {percent_outliers:.2f}%\n")

        if not any_outliers :
            print("\nNo outliers found in any column.\n")

# Calling the function
find_outliers_zscore(df, ['Weight', 'Velocity', 'Metallic_Content', 'Magnetic_Strength'])

```

```

Column : Metallic_Content
Mean : 87.47 , Standard Deviation : 44.38
Outlier condition : Z-score > 3 or Z-score < -3

Total data points: 5594, Outliers: 21
Outlier Percentage: 0.38%

```

FIG 5.22: Z - SCORE



### 5.1.6 Removing Outliers using the IQR Method:

Removing outliers using the IQR method involves eliminating data points that fall significantly below or above the typical value range defined by the interquartile range.

```
# Importing the Library
import pandas as pd
import statistics as stats

# Define the function
def remove_outliers_iqr(df, columns):

    # Loop through each selected column
    for column in columns:

        # Getting the column data as a sorted list
        values = sorted(df[column].tolist())

        # Calculating the median (Q2)
        q2 = stats.median(values)

        # Splitting data into lower and upper halves
        n = len(values)
        mid = n // 2

        if n % 2 == 0: # even number of elements
            lower_half = values[:mid]
            upper_half = values[mid:]
        else: # odd number of elements
            lower_half = values[:mid]
            upper_half = values[mid+1:]

        # Calculating Q1 and Q3
        q1 = stats.median(lower_half) if lower_half else 0
        q3 = stats.median(upper_half) if upper_half else 0

        # Calculating the Interquartile Range (IQR)
        iqr = q3 - q1

        # Determining bounds for outlier detection
        lower_bound = q1 - 1.5 * iqr
        upper_bound = q3 + 1.5 * iqr

        # Number of rows before removing outliers
        before_rows = df.shape[0]

        # Remove outliers
        df = df[(df[column] >= lower_bound) & (df[column] <= upper_bound)]

        # Number of rows after removing outliers
        after_rows = df.shape[0]
        removed = before_rows - after_rows

    # Print result
    if removed > 0:
        print(f"\nColumn: {column}")
        print(f"Q1: {q1}, Q3: {q3}, IQR: {iqr:.2f}")
        print(f"Outlier range: < {lower_bound:.2f} or > {upper_bound:.2f}")
        print(f"Outliers removed: {removed}")
    else:
        print(f"\nColumn: {column} - No outliers found.")

    return df # Return cleaned DataFrame

# Calling the function
df = remove_outliers_iqr(df, ['Weight', 'Velocity', 'Metallic_Content', 'Magnetic_Strength'])
```

```

Column: Weight - No outliers found.

Column: Velocity - No outliers found.

Column: Metallic_Content
Q1: 60.43893730669026, Q3: 84.98754308433111, IQR: 24.55
Outlier range: < 23.62 or > 121.81
Outliers removed: 1074

Column: Magnetic_Strength - No outliers found.

```

FIG 5.23: REMOVING OUTLIERS USING THE IQR METHOD

### 5.1.7 Change Categorical data into Numerical Data:

Converting categorical data helps algorithms process and learn from the data more effectively, improving model performance and predictive accuracy.

```

# Displaying the unique values in the 'Origin' column
df['Origin'].unique()

array(['Asteroidal', 'Lunar', 'Martian'], dtype=object)

```

```

# Mapping Origin
df['Origin']=df['Origin'].map({'Asteroidal':5,'Lunar':4,'Martian':3})
df

```

	Weight	Velocity	Metallic_Content	Magnetic_Strength	Origin
0	54.636884	10.821243	88.274805	0.404645	5
1	43.106142	24.964433	50.336015	0.687528	4
2	54.636884	14.833294	81.801750	0.367695	5
3	54.636884	21.145692	65.705708	0.552385	3
4	51.120673	21.401973	72.246332	0.740538	4
...	...	...	...	...	...
5587	43.525656	26.472705	53.633524	0.746377	4
5588	59.844237	14.405552	81.176930	0.511826	5
5589	57.580559	22.527021	58.888648	0.550352	3
5591	46.751220	22.188273	115.498628	0.718928	4
5593	40.802312	22.264605	45.492852	0.730193	4

4520 rows × 5 columns

FIG 5.24: CHANGE CATEGORICAL DATA INTO NUMERICAL DATA:

### 5.1.8 Iloc Splitting:

Iloc splitting is the process of selecting specific rows and columns from a DataFrame using integer-based indexing to separate input features (X) and target labels (y) for model training.

```
# Splitting the dataset into features (X) and target variable (y)
x = df.iloc[:, 0:4]
y = df.iloc[:, 4]
x
y
```

```
0      5
1      4
2      5
3      3
4      4
..
5587   4
5588   5
5589   3
5591   4
5593   4
```

```
Name: Origin, Length: 4520, dtype: int64
```

FIG 5.25: ILOC SPLITTING

### 5.1.9 Train And Test Split:

Train and test split is an essential step in machine learning where the dataset is divided into two parts: a training set and a testing set. The training set is used to teach the model, while the testing set is used to evaluate how well the model performs on new, unseen data. This helps in checking the model's accuracy and prevents overfitting. Typically, 80% of the data is used for training and the remaining 20% for testing.

```
# Splitting the data into training and testing sets (80% train, 20% test)
from sklearn.model_selection import train_test_split

x_train,x_test,y_train,y_test=train_test_split(x,y,test_size=0.20,random_state=42)
```

1. **x\_train** : Contains the input features used to train the model.

x_train				
	Weight	Velocity	Metallic_Content	Magnetic_Strength
4667	54.479214	20.772835	60.716737	0.472388
1234	54.636884	15.729924	72.246332	0.550352
311	59.947590	15.406201	79.658199	0.405409
2669	46.294479	27.504921	72.246332	0.752545
5062	59.425141	14.070453	84.705149	0.427230
...	...	...	...	...
5473	41.037865	25.884393	72.246332	0.719157
583	60.633907	15.250617	71.975554	0.393926
3833	58.108927	15.870372	76.941995	0.541533
4656	43.435111	26.113660	55.472585	0.627235
1082	54.636884	16.539892	75.885013	0.468985

3616 rows × 4 columns

2. **y\_train** : Contains the target labels corresponding to the training features.

y_train	
4667	3
1234	3
311	5
2669	4
5062	5
...	..
5473	4
583	5
3833	5
4656	4
1082	5

Name: Origin, Length: 3616, dtype: int64

3. **x\_test** : Holds the input features used to test the model's performance.

x_test				
	Weight	Velocity	Metallic_Content	Magnetic_Strength
3572	57.603228	21.837693	65.122581	0.611392
1013	57.545281	16.289265	80.713693	0.291106
4158	49.548351	18.612640	67.800778	0.581678
2916	42.919324	24.133152	51.083183	0.550352
5283	52.548068	19.920738	60.266779	0.516083
...	...	...	...	...
3640	59.994973	19.514775	59.858705	0.607108
4430	41.795887	27.835004	43.772778	0.738131
3300	61.505479	19.920738	86.358876	0.416102
4037	58.255938	19.920738	61.312854	0.534532
5238	48.425097	21.649294	57.647767	0.526697

904 rows × 4 columns

4. **y\_test** : Holds the actual target labels for the test set to compare with predictions.

```
y_test
3572    3
1013    5
4158    3
2916    4
5283    3
..
3640    3
4430    4
3300    5
4037    3
5238    3
Name: Origin, Length: 904, dtype: int64
```

FIG 5.26: TRAIN AND TEST SPLIT

### 5.1.10 Logistic Regression:

Logistic Regression is a supervised learning algorithm used for binary classification problems (e.g., yes/no, 0/1). It estimates the probability that a data point belongs to a certain class using a sigmoid function. Though it's called "regression," it is mainly used for classification tasks.

```
# Importing the Libraries
from sklearn.tree import DecisionTreeClassifier
from sklearn.metrics import accuracy_score
```

1. **Building the Model** : This step involves selecting and initializing a machine learning algorithm suitable for the problem.

```
DTClassifier=DecisionTreeClassifier()
```

2. **Training the Model** : The model learns patterns from the training data by adjusting internal parameters.

```
LRClassifier.fit(x_train,y_train)
```

3. **Testing the Model** : The trained model is evaluated on unseen test data to check its performance.

```
y_pred = LRClassifier.predict(x_test)
```

4. **Calculating the Accuracy** : Accuracy is measured by comparing the model's predictions with the actual values in the test set.

```
AScore_LR = accuracy_score(y_test,y_pred)
print(AScore_LR)
0.9502212389380531
```

5. **Classification Report :** The classification report provides detailed metrics such as precision, recall, F1-score, and support for each class, helping evaluate the overall performance of a classification model.

```
print("\nClassification Report : \n",classification_report(y_test,y_pred))
```

```
Classification Report :
              precision    recall  f1-score   support

     3         0.91      0.95      0.93        271
     4         0.97      0.94      0.95        324
     5         0.97      0.96      0.97        309

 accuracy          0.95
 macro avg         0.95
 weighted avg      0.95
```

6. **Prediction :** The final model is used to make predictions on new or real-world data.

```
LRC1 = LRClassifier.predict([
    [54.636884, 10.821243, 88.274805, 0.404645]
])
LRC1

array([5], dtype=int64)
```

**FIG 5.27: LOGISTIC REGRESSION**

### 5.1.11 Decision Tree Classifier without Parameters:

A Decision Tree is a flowchart-like structure that makes decisions based on yes/no questions. It splits the dataset into branches using features until it reaches a decision or classification. It is simple, easy to interpret, and works well with both numerical and categorical data.

```
# Importing the libraries
from sklearn.tree import DecisionTreeClassifier
from sklearn.metrics import accuracy_score
```

1. **Building the Model :** This step involves selecting and initializing a machine learning algorithm suitable for the problem.

```
DTClassifier=DecisionTreeClassifier()
```

2. **Training the Model :** The model learns patterns from the training data by adjusting internal parameters.

```
DTClassifier.fit(x_train,y_train)
```

3. **Testing the Model** : The trained model is evaluated on unseen test data to check its performance.

```
y_pred=DTClassifier.predict(x_test)
```

4. **Calculating the Accuracy** : Accuracy is measured by comparing the model's predictions with the actual values in the test set.

```
AScore_DT = accuracy_score(y_test,y_pred)
print(AScore_DT)
```

```
0.9623893805309734
```

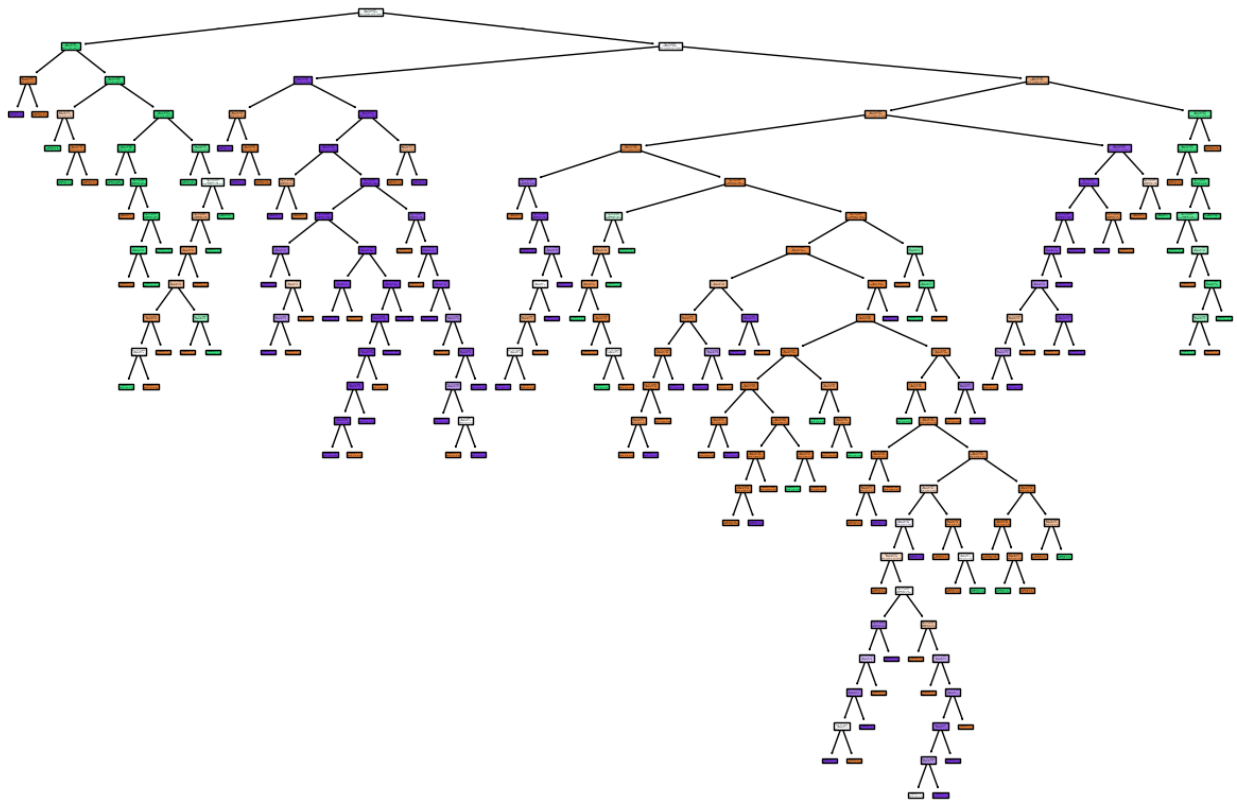
5. **Prediction** : The final model is used to make predictions on new or real-world data.

```
DTC1 = DTClassifier.predict([[54.636884, 10.821243, 88.274805, 0.404645]])
DTC1
```

```
array([5], dtype=int64)
```

6. **Construction** :

```
from sklearn import tree
import matplotlib.pyplot as plt
plt.figure(figsize=(15,10))
tree.plot_tree(DTClassifier,filled=True)
plt.show()
```



**FIG 5.28: DECISION TREE CLASSIFIER WITHOUT PARAMETERS**

### 5.1.12 Random Forest Classifier:

Random Forest is an ensemble learning method that builds multiple decision trees and combines their outputs to improve accuracy. It helps reduce overfitting and increases model stability and performance, making it more reliable than a single decision tree.

```
from sklearn.ensemble import RandomForestClassifier
from sklearn.metrics import accuracy_score, classification_report
```

1. **Building the Model :** This step involves selecting and initializing a machine learning algorithm suitable for the problem.

```
RFClassifier=RandomForestClassifier()
```

2. **Training the Model :** The model learns patterns from the training data by adjusting internal parameters.

```
RFClassifier.fit(x_train,y_train)
```

3. **Testing the Model :** The trained model is evaluated on unseen test data to check its performance.

```
y_pred=RFClassifier.predict(x_test)
```

4. **Calculating the Accuracy :** Accuracy is measured by comparing the model's predictions with the actual values in the test set.

```
AScore_RF=accuracy_score(y_test,y_pred)
print(AScore_RF)
```

0.9767699115044248

5. **Classification Report :** The classification report provides detailed metrics such as precision, recall, F1-score, and support for each class, helping evaluate the overall performance of a classification model.

```
print("\nClassification Report : \n",classification_report(y_test,y_pred))
```

```
Classification Report :
              precision    recall  f1-score   support

     3         0.97        0.95        0.96         271
     4         0.98        0.99        0.99         324
     5         0.97        0.99        0.98         309

 accuracy          0.98          0.98          0.98          904
  macro avg        0.98          0.98          0.98          904
 weighted avg        0.98          0.98          0.98          904
```



6. **Prediction :** The final model is used to make predictions on new or real-world data.

```
RFC1 = RFClassifier.predict([[54.636884, 10.821243, 88.274805, 0.404645]])  
RFC1  
  
array([5], dtype=int64)
```

**FIG 5.29: RANDOM FOREST CLASSIFIER**

### 5.1.13 Naive Bayes Classifier:

Naive Bayes is a probabilistic classifier based on Bayes' Theorem, assuming that all features are independent. It's simple, fast, and highly effective for text classification tasks like spam detection and sentiment analysis.

```
from sklearn.naive_bayes import GaussianNB  
from sklearn.metrics import accuracy_score, classification_report
```

1. **Building the Model :** This step involves selecting and initializing a machine learning algorithm suitable for the problem.

```
NBClassifier=GaussianNB()
```

2. **Training the Model :** The model learns patterns from the training data by adjusting internal parameters.

```
NBClassifier.fit(x_train,y_train)
```

3. **Testing the Model :** The trained model is evaluated on unseen test data to check its performance.

```
y_pred=NBClassifier.predict(x_test)
```

4. **Calculating the Accuracy :** Accuracy is measured by comparing the model's predictions with the actual values in the test set.

```
AScore_NB=accuracy_score(y_test,y_pred)  
print(AScore_NB)  
  
0.9712389380530974
```

- Classification Report :** The classification report provides detailed metrics such as precision, recall, F1-score, and support for each class, helping evaluate the overall performance of a classification model.

```
print("\nClassification Report : \n",classification_report(y_test,y_pred))
```

```

Classification Report :
              precision    recall  f1-score   support

     3         0.94      0.97      0.95         271
     4         0.99      0.97      0.98         324
     5         0.98      0.97      0.98         309

 accuracy          0.97
 macro avg         0.97
 weighted avg      0.97

```

- Prediction :** The final model is used to make predictions on new or real-world data.

```

NBC1 = NBClassifier.predict([[54.636884, 10.821243, 88.274805, 0.404645]])
NBC1

array([5], dtype=int64)

```

**FIG 5.30: NAIVE BAYES CLASSIFIER**

#### **5.1.14 Support Vector Machine Classifier:**

SVM is a classification algorithm that finds the best boundary (hyperplane) to separate data points from different classes. It works well for high-dimensional data and is effective for both linear and non-linear classification using kernels.

```

from sklearn.svm import SVC
from sklearn.metrics import accuracy_score,classification_report,confusion_matrix

```

- Building the Model :** This step involves selecting and initializing a machine learning algorithm suitable for the problem.

```
SVMClassifier=SVC(kernel='linear')
```

- Training the Model :** The model learns patterns from the training data by adjusting internal parameters.

```
SVMClassifier.fit(x_train,y_train)
```

- Testing the Model :** The trained model is evaluated on unseen test data to check its performance.

```
y_pred=SVMClassifier.predict(x_test)
```

- Calculating the Accuracy :** Accuracy is measured by comparing the model's predictions with the actual values in the test set.

```
AScore_SV=accuracy_score(y_test,y_pred)
print(AScore_SV)
```

0.9535398230088495

- Confusion Matrix and Classification Report :** A confusion matrix is a table that shows the actual vs. predicted classifications, helping identify correct predictions and types of errors like false positives and false negatives. The classification report provides detailed metrics such as precision, recall, F1-score, and support for each class, helping evaluate the overall performance of a classification model.

```
print("Confusion Matrix : \n",confusion_matrix(y_test,y_pred))
print("\nClassification Report : \n",classification_report(y_test,y_pred))
```

```
Confusion Matrix :
[[257  7  7]
 [ 19 304  1]
 [ 8  0 301]]

Classification Report :
              precision    recall  f1-score   support

     3             0.90      0.95      0.93         271
     4             0.98      0.94      0.96         324
     5             0.97      0.97      0.97         309

 accuracy              0.95
 macro avg              0.95
 weighted avg           0.95
```

- Prediction :** The final model is used to make predictions on new or real-world data.

```
SVC1 = SVMClassifier.predict([[54.636884, 10.821243, 88.274805, 0.404645]])
SVC1
array([5], dtype=int64)
```

**FIG 5.31: SUPPORT VECTOR MACHINE CLASSIFIER**

### 5.1.15 K - Nearest Neighbour Classifier:

KNN is a non-parametric algorithm that classifies data based on the majority class of its 'K' closest neighbors. It's easy to understand and implement, but can be slow with large datasets. It works well for pattern recognition and recommendation systems.

```
from sklearn.neighbors import KNeighborsClassifier
from sklearn.metrics import accuracy_score,classification_report,confusion_matrix
```

- Building the Model :** This step involves selecting and initializing a machine learning algorithm suitable for the problem.

```
KNNClassifier = KNeighborsClassifier()
```

2. **Training the Model** : The model learns patterns from the training data by adjusting internal parameters.

```
KNNClassifier.fit(x_train,y_train)
```

3. **Testing the Model** : The trained model is evaluated on unseen test data to check its performance.

```
y_pred=KNNClassifier.predict(x_test)
```

4. **Calculating the Accuracy** : Accuracy is measured by comparing the model's predictions with the actual values in the test set.

```
AScore_KNN=accuracy_score(y_test,y_pred)
print(AScore_KNN)
```

```
0.9601769911504425
```

5. **Confusion Matrix and Classification Report** : A confusion matrix is a table that shows the actual vs. predicted classifications, helping identify correct predictions and types of errors like false positives and false negatives. The classification report provides detailed metrics such as precision, recall, F1-score, and support for each class, helping evaluate the overall performance of a classification model.

```
print("Confusion Matrix : \n",confusion_matrix(y_test,y_pred))
print("\nClassification Report : \n",classification_report(y_test,y_pred))
```

```
Confusion Matrix :
[[245  10  16]
 [  4 320   0]
 [  6   0 303]]

Classification Report :
              precision    recall  f1-score   support

     3             0.96       0.90       0.93         271
     4             0.97       0.99       0.98         324
     5             0.95       0.98       0.96         309

 accuracy          0.96
 macro avg         0.96
 weighted avg      0.96
```

6. **Prediction** : The final model is used to make predictions on new or real-world data.

```
KNNC1 = KNNClassifier.predict([[54.636884, 10.821243, 88.274805, 0.404645]])
KNNC1
```

```
array([5], dtype=int64)
```

**FIG 5.32: K - NEAREST NEIGHBOUR CLASSIFIER**

### 5.1.16 Comparison Table:

This comparison table includes columns for the **algorithm name**, its **accuracy score**, and three sets of **input values**. For each input, the corresponding **actual output** and the model's **predicted output** are there. This helps in comparing how accurately each model performs on specific data points.

	ALGORITHM	ACCURACY	INPUT 1	ACTUAL 1	PREDICTION 1	INPUT 2	ACTUAL 2	PREDICTION 2	INPUT 3	ACTUAL 3	PREDICTION 3
0	LOGISTIC REGRESSION	0.950221	[54.636884, 10.821243, 88.274805, 0.404645]	5	[5]	[43.525656, 26.472705, 53.633524, 0.746377]	4	[4]	[57.580559, 22.527021, 58.888648, 0.550352]	3	[3]
1	DECISION TREE CLASSIFIER WITHOUT PARAMETERS	0.962389	[54.636884, 10.821243, 88.274805, 0.404645]	5	[5]	[43.525656, 26.472705, 53.633524, 0.746377]	4	[4]	[57.580559, 22.527021, 58.888648, 0.550352]	3	[3]
2	RANDOM FOREST CLASSIFIER	0.976770	[54.636884, 10.821243, 88.274805, 0.404645]	5	[5]	[43.525656, 26.472705, 53.633524, 0.746377]	4	[4]	[57.580559, 22.527021, 58.888648, 0.550352]	3	[3]
3	NAIVE BAYES CLASSIFIER	0.971239	[54.636884, 10.821243, 88.274805, 0.404645]	5	[5]	[43.525656, 26.472705, 53.633524, 0.746377]	4	[4]	[57.580559, 22.527021, 58.888648, 0.550352]	3	[3]
4	SUPPORT VECTOR MACHINE CLASSIFIER	0.953540	[54.636884, 10.821243, 88.274805, 0.404645]	5	[5]	[43.525656, 26.472705, 53.633524, 0.746377]	4	[4]	[57.580559, 22.527021, 58.888648, 0.550352]	3	[3]
5	K - NEAREST NEIGHBOUR CLASSIFIER	0.960177	[54.636884, 10.821243, 88.274805, 0.404645]	5	[5]	[43.525656, 26.472705, 53.633524, 0.746377]	4	[4]	[57.580559, 22.527021, 58.888648, 0.550352]	3	[3]

**Table 5.1: COMPARISON TABLE**

### 5.1.17 Comparison of Classifier Algorithm using Bar Chart:

```
# Importing the Libraries
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt

# Figure size
plt.figure(figsize=(15,6))

# Subplot
ax1 = plt.subplot(121)
ax2 = plt.subplot(122)

# -----
# Bar Chart 1: Accuracy Comparison
# -----

# Creating the data
algorithm_names = np.array(["Logistic \nRegression","Decision \nTree Classifier \nwithout \nParameters",
                             "Random \nForest \nClassifier","Naive \nBayes \nClassifier",
                             "Support \nVector \nMachine\nClassifier","K - Nearest \nNeighbour \nClassifier"])
accuracy_values = np.array([AScore_LR,AScore_DT,AScore_RF,AScore_NB,AScore_SV,AScore_KNN])

bars1 = ax1.bar(algorithm_names, accuracy_values, color=['c', 'b', 'm', 'g', 'r', 'y'])
ax1.bar_label(bars1, fmt='%.2f',fontSize=6)
ax1.set_title('Accuracy Comparison',fontSize=14)
ax1.set_ylabel('Accuracy',size=12)
ax1.set_xlabel('Algorithm Names',size=12)
ax1.set_ylim(0, 1)

# -----
# Bar Chart 2: Predicted vs Actual Values
# -----

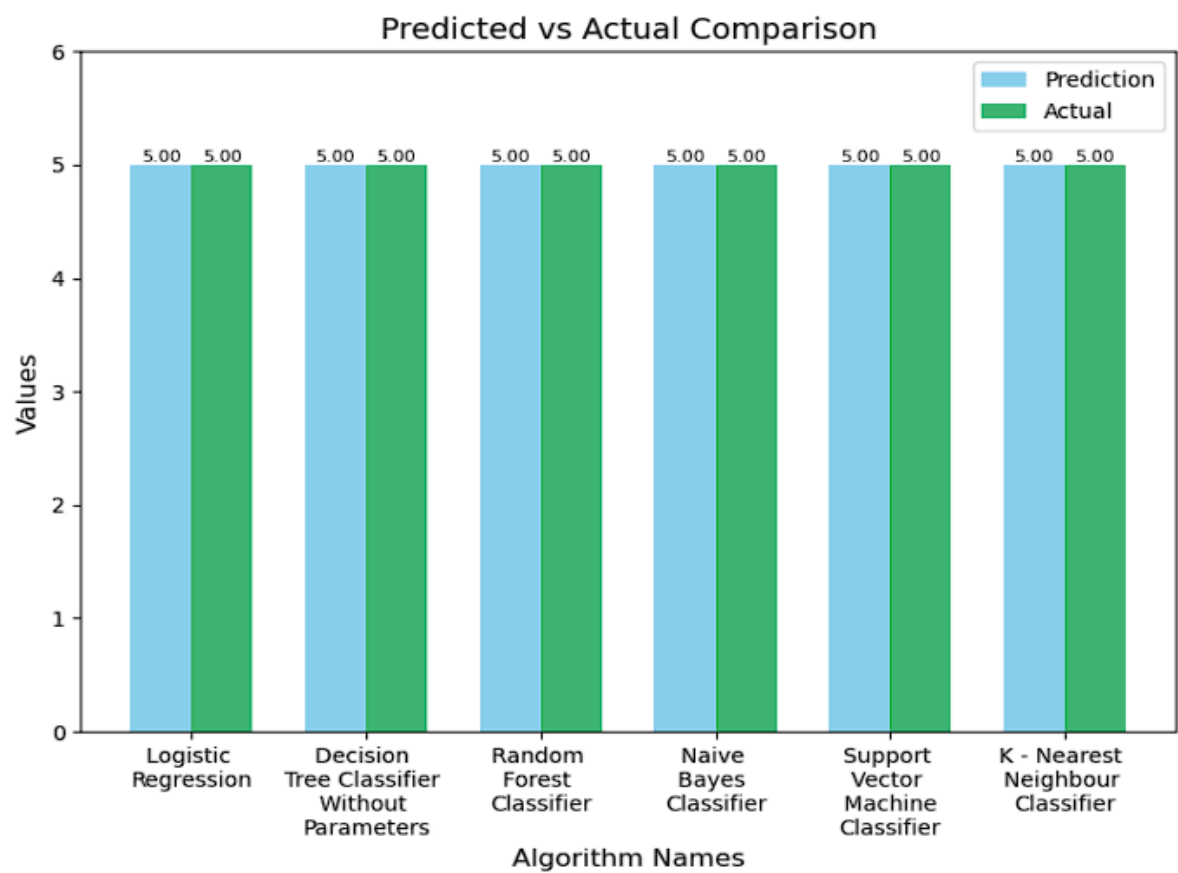
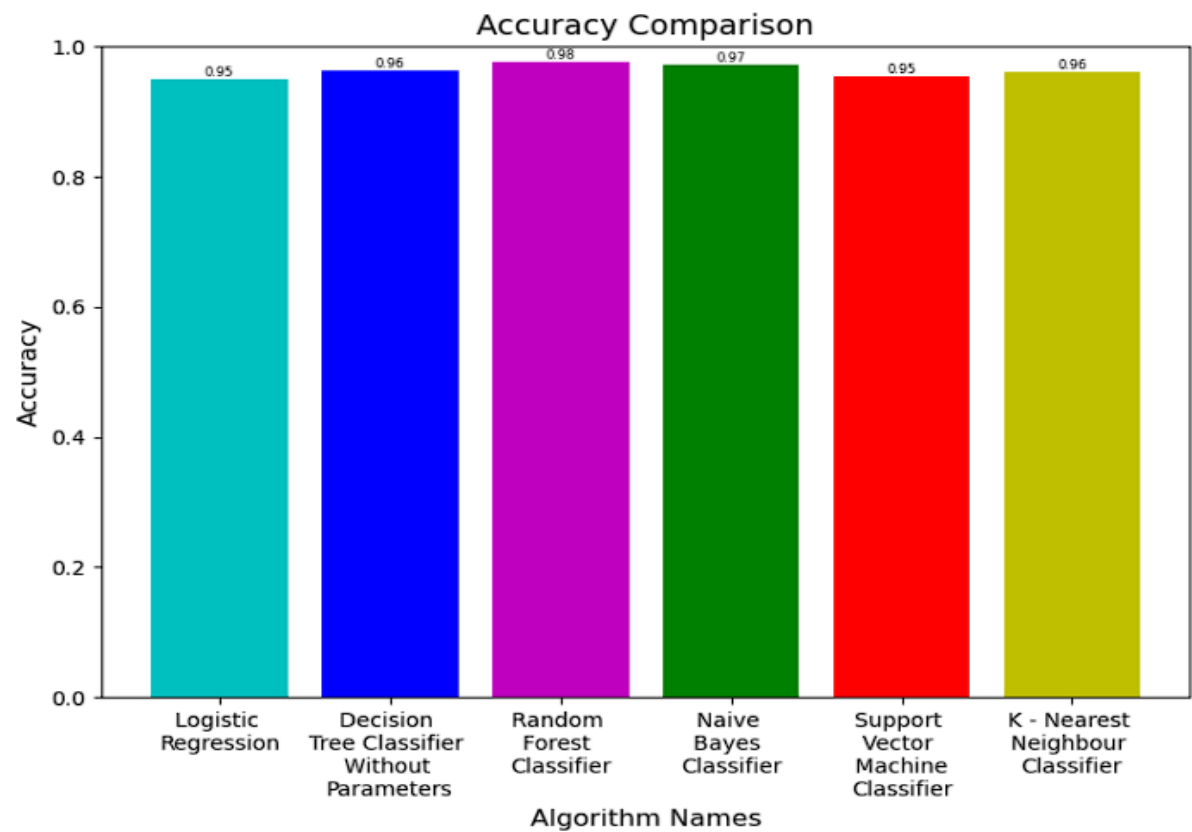
# Creating the data
prediction_values = np.array([x.item() for x in [LRC1, DTC1, RFC1, NBC1, SVC1, KNNC1]])
actual_values = np.array([5,5,5,5,5,5])

x = np.arange(len(algorithm_names)) # Position of bars on x-axis
width = 0.35 # Width of each bar

# Plotting bars
bars_pred = ax2.bar(x - width/2, prediction_values, width, label='Prediction', color='skyblue')
bars_act = ax2.bar(x + width/2, actual_values, width, label='Actual', color='mediumseagreen')

ax2.set_title('Predicted vs Actual Comparison', fontSize=14)
ax2.set_ylabel('Values',size=12)
ax2.set_xlabel('Algorithm Names',size=12)
ax2.bar_label(bars_pred, fmt='%.2f',fontSize=8)
ax2.bar_label(bars_act, fmt='%.2f',fontSize=8)
ax2.set_xticks(x)
ax2.set_xticklabels(algorithm_names)
ax2.set_ylim(0, 6)
ax2.legend()
plt.tight_layout()

plt.show()
```



**FIG 5.33: COMPARISON OF CLASSIFIER ALGORITHM USING BAR CHART**

## 5.2 Deep Learning Project

Brand Logo Detection

### 5.2.1 Project Aim:

Image classification model using convolutional neural networks (CNNs) to accurately identify and classify brand logos from a given dataset of logo images.

### 5.2.2 Importing Dependencies and Uploading Dataset

Upload Datasets and Import Dependencies

First, all of the required libraries and dependencies should be imported before starting the project. These are:

- NumPy and Pandas are libraries for manipulating data and performing numerical operations.
- Matplotlib is used for visualization of the data.
- OS is used for managing files.
- For building and training the deep learning model, TensorFlow and Keras are used.

Next, when importing the libraries and dependencies is complete, the dataset with pictures of different brand logos is uploaded. This dataset will usually be a directory with folders named after the different brands along with image files showing the logos of each respective brand. Later on, when running automatic labels will reference this folder structure.

```
[ ] import tensorflow as tf
    from tensorflow.keras.models import Sequential
    from tensorflow.keras.layers import Conv2D, MaxPooling2D, Flatten, Dense, Dropout
    from tensorflow.keras.preprocessing.image import ImageDataGenerator
    import matplotlib.pyplot as plt
    import numpy as np
```

```
[ ] from google.colab import files
    uploaded = files.upload()
```



Choose Files No file chosen

Upload widget is only available when the cell has been executed

Saving archive.zip to archive.zip

**FIG 5.34: Importing Dependencies**



### 5.2.3 Extracting and Verifying Dataset

A basic Python program executes during this stage for displaying all image files found throughout the dataset directory. The main purpose of this process includes the following three points:

- The dataset has been successfully uploaded,
- The dataset organization becomes clear,
- Files that are either missing or corrupted can be detected.
- The correct listing of image filenames confirms that the dataset is prepared for preprocessing and training.

```
[ ] import zipfile

with zipfile.ZipFile("archive.zip", 'r') as zip_ref:
    zip_ref.extractall("logos_data")
```

```
▶ import os
os.listdir("/content/logos_data/Logos")
```

```
⇒ ['hp-inc-logo-vector-download-400x400.jpg',
   'playboy-tv-eps-vector-logo-400x400.png',
   'plks-pewel-mala-vector-logo-400x400.png',
   'find-us-on-facebook-logo-vector-400x400.png',
   'saint-gobain-logo-vector-download-400x400.jpg',
   'moncler-vector-logo-400x400.png',
   ...]
```

**FIG 5.35: Extracting and Verifying Dataset**

### 5.2.4 Generating CSV with Filenames and Labels

The process of converting folder-based datasets into CSV format serves as an effective method to streamline training tasks. Here:

- The system retrieves filenames through the `os.listdir()` function,
- The program extracts brand names by analyzing the folder labels,
- A CSV file is generated which links every image file to its designated label.
- Tracking and visualization of datasets along with debugging functions become possible through this method which also proves beneficial for custom training pipelines.

```
import os
import pandas as pd

image_folder = '/content/logos_data/Logos'

# List all image files
image_files = os.listdir(image_folder)
image_files = [f for f in image_files if f.endswith((''.jpg', '.jpeg', '.png'))]

# Function to generate label
def extract_label(filename):
    return filename.split('-')[0].split('.')[0].capitalize()

# Create DataFrame
df = pd.DataFrame({
    'filename': image_files,
    'label': [extract_label(f) for f in image_files]
})
df['file_path'] = df['filename'].apply(lambda x: os.path.join(image_folder, x))

# Save to CSV
df.to_csv('/content/logos_data/LogoDatabase.csv', index=False)
df.head()
```

**FIG 5.36: Generating CSV**

### 5.2.5 Preprocessing and Data Splitting

This step uses the ImageDataGenerator class from Keras, which allows:

- **Image rescaling** (to normalize pixel values),
- **Automatic label assignment** based on folder names,
- **Shuffling** of the dataset,
- **Splitting** into training and validation sets.

The generator also allows for **real-time image augmentation** (if needed), but in this project, its primary role is to simplify data loading and ensure the model is trained on uniformly sized and preprocessed data.



**FIG 5.37: Preprocessing**

### 5.2.6 Model Definition Using Keras Sequential

The Keras Sequential model functions as the foundation for defining neural network architectures. The model's structure consists of:

- The model begins with a Flatten layer which transforms two-dimensional images into one-dimensional arrays.
- The model employs Dense (fully connected) layers which learn complex patterns from the input data.
- The Dropout layer functions as a regularization method which disables neurons randomly during training to stop overfitting.
- The basic architecture of this model provides an accessible entry point for new users while maintaining solid results in classifying moderately complex data.

```
model = Sequential([
    Flatten(),
    Dense(128, activation='relu'),
    Dropout(0.5),
    Dense(train_generator.num_classes, activation='softmax')
])
```

```
model.compile(
    optimizer='adam',
    loss='categorical_crossentropy',
    metrics=['accuracy']
)
```

```
history = model.fit(
    train_generator,
    validation_data=val_generator,
    epochs=10
)
```

**FIG 5.38: Model Definition**

## 5.2.7 Model Compilation and Training

The model undergoes compilation with specific settings before beginning its training process:

- The model uses Adam as its optimizer because it delivers adaptive learning rate capabilities together with efficient training performance.
- The model uses categorical\_crossentropy as its loss function because it works well for multi-class classification tasks.
- The training progress is monitored through the accuracy metric.
- The training process runs for 10 epochs which means the model processes the entire training dataset once during each epoch. The training process generates historical data which tracks both loss values and accuracy scores for each epoch.

```
history = model.fit(
    train_generator,
    validation_data=val_generator,
    epochs=10
)
```

/usr/local/lib/python3.11/dist-packages/keras/src/trainers/data\_adapters/py\_dataset\_adapter.py:121: UserWarning: Your `PyDataset` self.\_warn\_if\_super\_not\_called()  
Epoch 1/10  
/usr/local/lib/python3.11/dist-packages/keras/src/ops/nn.py:907: UserWarning: You are using a softmax over axis -1 of a tensor  
warnings.warn(  
/usr/local/lib/python3.11/dist-packages/keras/src/losses/losses.py:33: SyntaxWarning: In loss categorical\_crossentropy, expecte  
return self.fn(y\_true, y\_pred, \*\*self.\_fn\_kwargs)  
36/36 ————— 7s 130ms/step - accuracy: 1.0000 - loss: 0.0000e+00 - val\_accuracy: 1.0000 - val\_loss: 0.0000e+00  
Epoch 2/10  
36/36 ————— 4s 110ms/step - accuracy: 1.0000 - loss: 0.0000e+00 - val\_accuracy: 1.0000 - val\_loss: 0.0000e+00  
Epoch 3/10  
36/36 ————— 3s 92ms/step - accuracy: 1.0000 - loss: 0.0000e+00 - val\_accuracy: 1.0000 - val\_loss: 0.0000e+00  
Epoch 4/10  
36/36 ————— 4s 110ms/step - accuracy: 1.0000 - loss: 0.0000e+00 - val\_accuracy: 1.0000 - val\_loss: 0.0000e+00  
Epoch 5/10  
36/36 ————— 4s 110ms/step - accuracy: 1.0000 - loss: 0.0000e+00 - val\_accuracy: 1.0000 - val\_loss: 0.0000e+00  
Epoch 6/10  
36/36 ————— 3s 92ms/step - accuracy: 1.0000 - loss: 0.0000e+00 - val\_accuracy: 1.0000 - val\_loss: 0.0000e+00  
Epoch 7/10  
36/36 ————— 3s 90ms/step - accuracy: 1.0000 - loss: 0.0000e+00 - val\_accuracy: 1.0000 - val\_loss: 0.0000e+00

**FIG 5.39: Model Training**

### 5.2.8 Model Evaluation

The evaluation of the model takes place through the validation dataset after its training phase. The process evaluates:

- The validation accuracy demonstrates the model's ability to generalize to new data points.
- Loss functions as the prediction error measurement.
- The evaluation process shows whether the model experiences overfitting through high training accuracy combined with low validation accuracy or underfitting by maintaining low accuracy across both training and validation datasets.

```
loss, acc = model.evaluate(val_generator)
print(f"Validation Accuracy: {acc:.2f}")
```

9/9 ————— 1s 72ms/step - accuracy: 1.0000 - loss: 0.0000e+00  
Validation Accuracy: 1.00

```
label_names = list(val_generator.class_indices.keys())
```

FIG 5.40: Model Accuracy

### 5.2.9 Extracting Class Names

The data generation process automatically derives class labels from folder naming conventions. During this step the program extracts class names (e.g., "Nike", "Adidas", "Apple") to enable:

- You can convert model predictions (which are numerical indices) back into human-readable labels.
- You can utilize the extracted class names to display them while visualizing results.
- This step functions as an essential validation method to determine if the model prediction aligns with the actual class label.

```
[ ] # Pick a batch from validation set
    sample_images, sample_labels = next(val_generator)

    # Predict first image
    prediction = model.predict(np.expand_dims(sample_images[2], axis=0))
    predicted_class = np.argmax(prediction[0])
    actual_class = np.argmax(sample_labels[2])

    print("Predicted:", predicted_class)
    print("Actual:", actual_class)

    # Visualize
    plt.imshow(sample_images[2])
    plt.title(f"Predicted: {predicted_class}, Actual: {actual_class}")
```

FIG 5.41: Class name into label

### 5.2.10 Predictions and Result Comparison

The trained model receives several sample images from the validation set to process. The predicted classes are compared with the actual labels while both are shown together with their respective images. Through visual evaluation the following results become apparent:

- The model demonstrates its ability to assign correct classifications to images which were not part of the training set.
- The model predictions show agreement with the actual class labels.
- This validation step functions as both a validation tool and a qualitative evaluation of model accuracy.

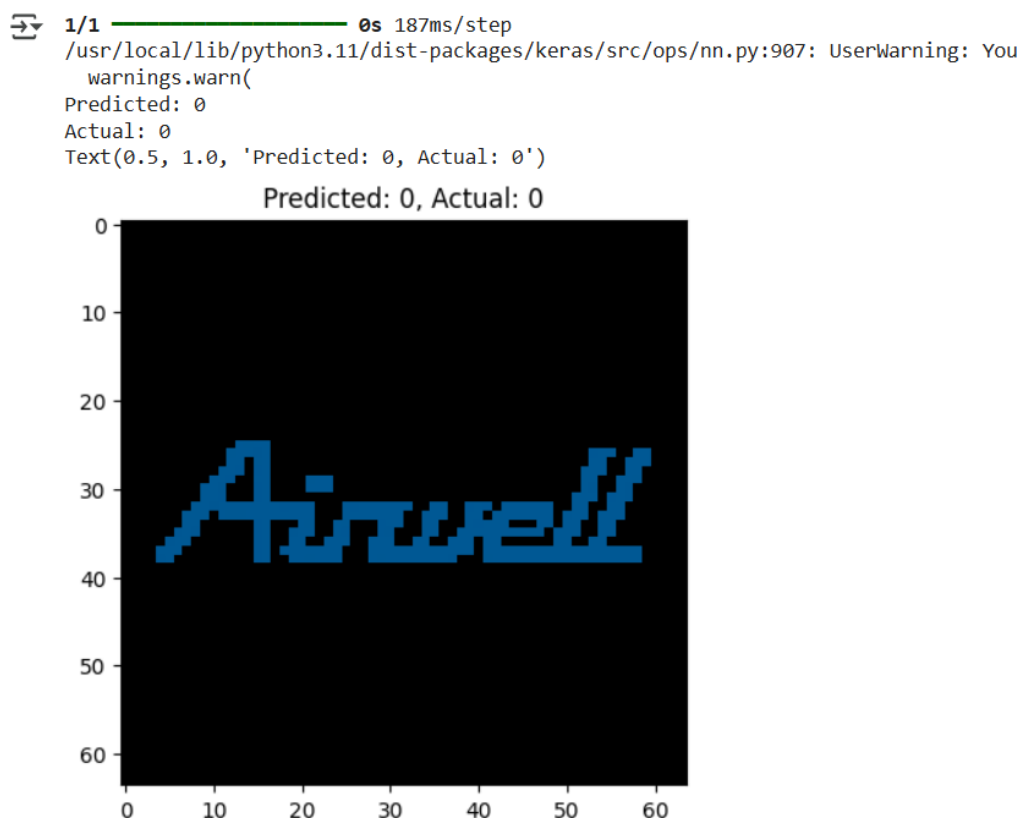


FIG 5.42: Result Comparison

## 5.3 Artificial Intelligence Project

Person's Disease Detection

### 5.3.1 Project Aim

To build AI assistant that predicts a person's disease based on symptoms using a ML data training model and suggests precautions.

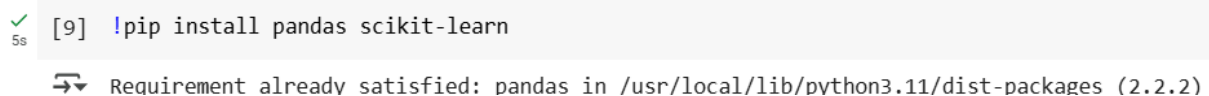
### 5.3.2 Why Scikit-Learn and How Training Works

**Scikit-learn** is a widely used Python library for machine learning. It simplifies building, training, testing, and deploying ML models. The project uses it for multiple steps:

- **Data preprocessing:** Cleaning and formatting the dataset into features (symptoms) and targets (diseases).
- **Model selection:** Using the **Random Forest Classifier**, which is robust, efficient, and provides high accuracy in classification tasks.
- **Training:** The ML model is trained on a dataset where symptoms are input features, and diseases are target outputs. The model learns the relationships between them during this phase.
- **Prediction:** After training, the model can predict diseases based on new symptom data.

This library provides all required tools in one place with clean syntax and effective performance.

---



```
✓ 5s [9] !pip install pandas scikit-learn
↔ Requirement already satisfied: pandas in /usr/local/lib/python3.11/dist-packages (2.2.2)
```

**FIG 5.43: Scikit-Learn installation**



### 5.3.3 Uploading Data to Colab

Google Colab is chosen as the development environment because:

- It provides free GPU and cloud storage.
- No installation is required—everything runs in the browser.
- It's suitable for educational and experimental ML projects.

In this step, the CSV files (datasets) are uploaded directly to Colab. These datasets contain rows of data where each row includes symptoms and a label for the diagnosed disease. Colab allows quick testing and iteration during model development.

```
✓ [11] import pandas as pd
0s

# Load symptom dataset
df = pd.read_csv("symptom_dataset.csv")

# Separate features and target
X = df.drop("Disease", axis=1)
y = df["Disease"]

# Get list of all symptoms (used for later user input)
all_symptoms = X.columns.tolist()
```

**FIG 5.44: Uploading Data to Colab**

### 5.3.4 Reading Dataset and Training Dataset

The uploaded dataset is read using **Pandas**, a data manipulation library. In this step:

- Columns representing various symptoms are separated as **input features**.
- The **target label** (disease) is isolated for training.

This structured approach allows the model to focus on patterns within the symptoms that are indicative of specific diseases. Features are often encoded or cleaned (if necessary), and the data may be split into training and testing sets to evaluate model performance later.

#### **Model Saving using Joblib**

Once the model is trained successfully and performs well on validation data, it is saved using the **Joblib** library. This step is crucial because:

- It eliminates the need to retrain the model every time the assistant runs.
- The model file (.pkl) can be loaded quickly into memory and used for predictions.
- It supports real-time performance in practical applications.


This step makes the assistant efficient, scalable, and ready for deployment.

```
[12] from sklearn.ensemble import RandomForestClassifier
      from sklearn.model_selection import train_test_split
      import joblib

      # Split dataset
      X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.2, random_state=42)

      # Train Random Forest model
      model = RandomForestClassifier(n_estimators=100, random_state=42)
      model.fit(X_train, y_train)

      # Save model
      joblib.dump(model, "disease_model.pkl")
```

 ['disease\_model.pkl']

**FIG 5.45: Reading and model saving**

### 5.3.5 Precautions Dataset

In addition to predicting diseases, the AI assistant also provides **health precautions** tailored to each disease. For this, another dataset is used that maps each disease to a list of precautions (like drink water, take rest, consult doctor, etc.).

When a prediction is made, the assistant looks up this dataset and shows the relevant precautions to the user. This makes the tool not just diagnostic but also **advisory and preventive** in nature.

```
[13] precaution_df = pd.read_csv("precaution_dataset.csv")
```

**FIG 5.46: Precaution Dataset Loading**

### 5.3.6 AI Assistant Interface

The assistant's interface is designed to be user-friendly:

- It accepts **user inputs**: name, age, gender, and symptoms (via a text or dropdown input).
- The trained model then **predicts the disease** based on the symptoms.
- After prediction, **precautions** are displayed using the second dataset.

This interface makes the assistant behave like a digital health bot that can assist users in real-time. It combines input gathering, ML prediction, and result presentation in one coherent workflow.

```
import joblib

# Load trained model
model = joblib.load("disease_model.pkl")

# Ask user for info
name = input("Enter your name: ")
age = input("Enter your age: ")
gender = input("Enter your gender (Male/Female): ")
# print("\nAvailable symptoms:")
# print(", ".join(all_symptoms))

# Ask for symptoms
user_symptoms = input("\nEnter your symptoms (comma separated): ").lower().split(',')

# Create input vector for prediction
input_vector = [1 if symptom.strip() in user_symptoms else 0 for symptom in all_symptoms]


# Predict disease
predicted_disease = model.predict([input_vector])[0]

# Display prediction
print(f"\n{name} (Age {age}, Gender {gender})")
print(f"Based on the symptoms, you may be suffering from: {predicted_disease}")

# Fetch precautions
row = precaution_df[precaution_df["Disease"] == predicted_disease]
if not row.empty:
    print("\nPrecautions to take:")
    for i in range(1, 5):
        print(f"- {row[f'Precaution_{i}'].values[0]}")
else:
    print("Precautions not available.")
```

FIG 5.47: AI Interface

## Final result output:

 Enter your name: sneha  
Enter your age: 118  
Enter your gender (Male/Female): female  
  
Enter your symptoms (comma separated): headache, fatigue, nausea  
  
sneha (Age 118, Gender female)  
Based on the symptoms, you may be suffering from: Malaria  
  
Precautions to take:  
- Use mosquito nets  
- Take antimalarial drugs  
- Avoid stagnant water  
- Rest  
/usr/local/lib/python3.11/dist-packages/sklearn/utils/validation.py:2739: UserWarning: warn(  
warnings.warn(  

**FIG 5.48: Result / Output**

## Chapter 6

### Conclusion

#### 6.1 Conclusion of Machine Learning

- The dataset was tested with multiple machine learning algorithms.
- **Random Forest Classifier** gave the best performance.
- Achieved **98% accuracy**, the highest among all models tested.
- Predictions across all models were consistent, but Random Forest was most reliable.
- Random Forest reduces overfitting by combining outputs from multiple decision trees.
- Provided high stability, generalization, and accuracy.
- Concluded as the **best algorithm for the dataset**.

#### 6.2 Conclusion of Deep Learning

- Developed a **deep learning model for brand logo classification**.
- Used **ImageDataGenerator** for preprocessing and data augmentation.
- Model architecture included **Flatten, Dense, and Dropout layers**.
- Trained with the **Adam optimizer** and **categorical cross-entropy** loss function.
- Achieved **high validation accuracy**.
- Verified predictions by comparing actual labels and displaying test images.
- Demonstrated deep learning's effectiveness in image-based classification tasks.

#### 6.3 Conclusion of Artificial Intelligence

- Built an **AI-based assistant for disease prediction**.
- Trained the model using a **symptom-disease dataset**.
- Employed **Random Forest Classifier** for accurate results.
- Accepted user input: **name, age, gender, and symptoms**.
- Predicted possible diseases and provided related **precautionary suggestions**.
- Developed using **Python, Scikit-learn, and Pandas** on **Google Colab**.
- Found useful for **early disease detection, health awareness, and guidance**.

#### 6.4 Combined Insight

- All three technologies—ML, DL, and AI—demonstrated strong real-world applicability.
- **Machine Learning** provided accurate prediction using structured data.
- **Deep Learning** handled visual tasks effectively through neural networks.
- **Artificial Intelligence** offered a complete intelligent system for healthcare use.

## **Chapter 7**

### **Future Scope**

#### **7.1 Future Scope of Meteorite Type and Origin Classification (Machine Learning)**

##### **1. Global Dataset Expansion**

- Integration of international meteorite datasets from organizations like NASA can improve model accuracy and geographical classification.

##### **2. Geospatial Mapping and Prediction**

- Machine learning models can be integrated with GIS data to predict likely impact zones or meteorite origins.

##### **3. Real-time Event Detection**

- Using satellite data and sensor inputs, future versions can enable real-time meteor detection and classification.

##### **4. Automated Research Tools**

- The project can evolve into a tool for researchers to analyze newly discovered meteorites quickly and with minimal manual effort.

##### **5. Educational Applications**

- The system can be turned into an educational module or interactive platform for schools, museums, and scientific institutions.

#### **7.2 Future Scope of Popular Brand Logos Detection (Deep Learning)**

##### **1. Real-time Logo Recognition in Videos**

- Integration into surveillance and sports broadcasting for detecting and tracking brand visibility.

##### **2. Marketing and Advertising Analytics**

- Detection systems can be used to measure brand presence in public places, media, and advertisements.

##### **3. Augmented Reality Applications**

- Coupling logo detection with AR can trigger dynamic marketing content based on the user's camera input.

##### **4. Scalable Dataset Integration**

- Inclusion of thousands of brand logos across industries and global markets to build a comprehensive detection engine.

##### **5. Robustness Against Noise and Occlusion**

- Advanced architectures like YOLOv5 or EfficientNet can be used to make the model work in real-world, cluttered environments.

##### **6. Mobile Deployment**

- Optimization for edge computing to allow logo detection on smartphones and embedded systems.

#### **7.3 Future Scope of Person's Disease and Precautions Prediction (Artificial Intelligence)**

##### **1. Multilingual and Voice-Based Assistant**

- Expanding the AI to accept voice input in regional languages can make it more inclusive and accessible.

2. **Integration with Wearables and Health Sensors**
  - Real-time symptom monitoring using smartwatches and medical sensors can make predictions more timely and personalized.
3. **Chronic Disease Risk Assessment**
  - Future models can predict risk of chronic diseases like diabetes or heart conditions based on longitudinal data.
4. **Telemedicine Integration**
  - The system can serve as a front-end triage assistant before connecting users with medical professionals.
5. **Dynamic Precaution Updates**
  - Linking to medical databases can allow the model to update its advice based on the latest health guidelines and disease outbreaks.
6. **Privacy-Preserving AI**
  - Techniques like federated learning and differential privacy can be used to secure patient data while allowing model improvement.



## **Chapter 8**

### **References**

#### **8.1 Datasets and Data Sources**

1. NASA Meteorite Landings Dataset – <https://data.nasa.gov>
2. Kaggle Meteorite Dataset – <https://www.kaggle.com/nasa/meteorite-landings>
3. Brand Logos Image Dataset – Kaggle: <https://www.kaggle.com/datasets>
4. Symptom-Disease Mapping Dataset – Public Health Data Repository, Kaggle

#### **8.2 Tools and Frameworks**

1. **Python Programming Language**  
Python Software Foundation, <https://www.python.org>

#### **8.3 Platforms and Development Environment**

1. **Google Colab** – Cloud-based Python notebook environment  
<https://colab.research.google.com>
2. **Kaggle** – Data science competition platform and datasets  
<https://www.kaggle.com>
3. **Jupyter Notebook** – Open-source tool for creating and sharing documents with live code  
<https://jupyter.org>