Bansilal Ramnath Agarwal Charitable Trust's

**Vishwakarma Institute of Information Technology, Pune-48**

(An Autonomous Institute affiliated to Savitribai Phule Pune University)

# DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING (ARTIFICIAL INTELLIGENCE)

# Lab Writeup

## CAUA22201: Data Science and Machine Learning

**Name : Komal Suresh Jadhav**
**Prn no. 22320089**
**Roll no.: 73**
**Batch: S3**

Submitted to,

Dr. Anuradha Yenkikar

**SY Semester II Academic Year 2023-24**

# Table of Contents

## ASSIGNMENT NO. 1

**Title:** Perform the following operations using R/Python on suitable data sets:

a) read data from different formats (like csv, xls)

b) indexing and selecting data, sort data,

c) describe attributes of data, checking data types of each column,

d) counting unique values of data, format of each column, converting variable data type (e.g. from long to short, vice versa),

e) identifying missing values and fill in the missing values

**S/W Packages and H/W apparatus used:** Linux OS: Ubuntu/Windows , Jupyter notebook.

**Theory:**

**Data Preparation**

Data preparation also referred as “data preprocessing” is the process of cleaning and transforming raw data prior to processing and analysis. It is an important step prior to processing and often involves reformatting data, making corrections to data and the combining of data sets to enrich data.

**Importance of Data preparation**

● Because most machine learning algorithms require data to be structured in a specific way, datasets must be prepared before they can offer useful insights. a number of databases having missing, invalid, or otherwise difficult to process values for an algorithm. If you're looking for information, The algorithm will be unable to use it if it is missing. If the data is incorrect, the algorithm will produce inaccurate or even incorrect results. the outcomes are deceiving

● Some datasets are relatively clean but need to be shaped (e.g., aggregated or pivoted) and many datasets are just lacking useful business context (e.g., poorly defined ID values), hence the need for feature enrichment. Good data preparation produces clean and well curated data which leads to more practical, accurate model outcomes.

● Before entering the data into the machine learning model, this is the most important step. The reason for this is that the data set must be unique and specific to the model, thus we must identify the data's required characteristics. The data preparation process provides a mechanism for preparing data for project definition as well as project evaluation of machine learning algorithms.

● There are a variety of predicting machine learning models available, each with its own method. However, some processes are common to all models, and they allow us to identify the underlying business problem and its solutions. The following are some of the data preparation procedures:

● 1. Determine the problems

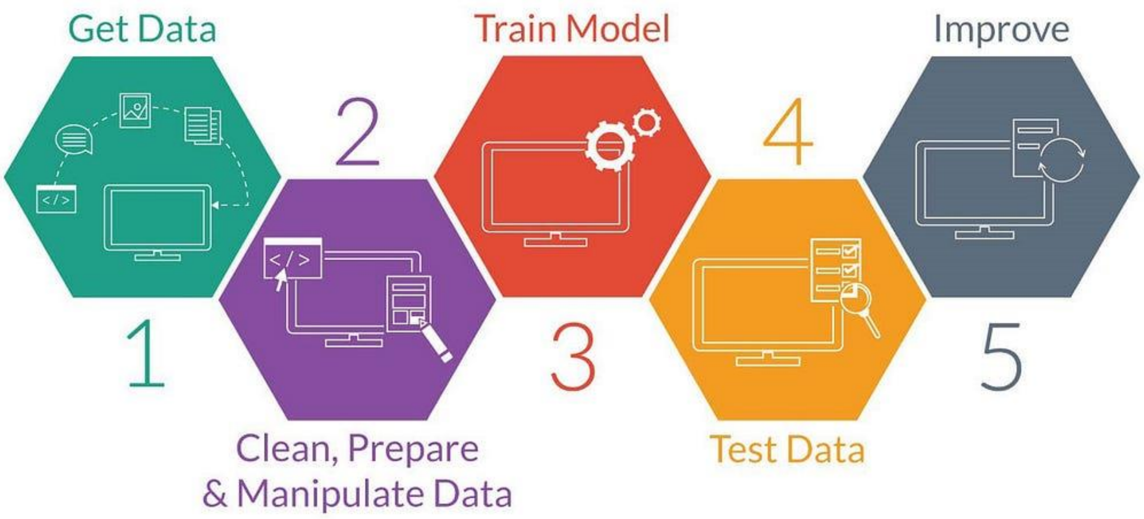
● 2. Data cleaning

● 3. Feature selection

● 4. Data transformation

● 5. feature engineering

● 6. Dimensionality reduction



a) Reading Data from Different Formats:

- CSV: Using `pd.read_csv()`

- Excel: Using `pd.read_excel()`

b) Indexing and Selecting Data, Sorting Data:

- Indexing: Using `.loc[]` and `.iloc[]`
- Selection: Using column names or numeric indices
- Sorting: Using `.sort_values()`

c) Describing Attributes of Data, Checking Data Types:

- Describing: Using `.describe()`
- Data Types: Using `.dtypes`

d) Counting Unique Values, Converting Variable Data Types:

- Unique Values: Using `.value_counts()`
- Converting Data Types: Using `.astype()`

e) Identifying and Handling Missing Values:

- Identifying: Using `.isna()` or `.isnull()`
- Handling: Using `.fillna()` or `.dropna()`

**Working:**

1. Read data from various formats using appropriate Pandas functions.
2. Perform indexing, selection, and sorting operations to extract relevant subsets of data.
3. Use descriptive statistics methods to understand the attributes of the data.
4. Count unique values and convert data types as needed.
5. Identify missing values and handle them by filling or removing them.
6. Utilize Pandas' functionalities to perform any additional data manipulation tasks required.

**Conclusion:**

Data preparation is recognized for helping businesses and analytics to get ready and prepare the data for operations. Pandas is a versatile tool for data manipulation and exploration in Python, offering numerous functionalities to handle various data preprocessing tasks efficiently. By leveraging its capabilities, users can streamline the process of preparing data for analysis and gain valuable insights from their datasets.

## ASSIGNMENT NO. 2

**Title:** Perform the following operations using R/Python on the data sets:

a) Compute and display summary statistics for each feature available in the dataset. (e.g. minimum value, maximum value, mean, range, standard deviation, variance and percentiles

b) Data Visualization-Create a histogram for each feature in the dataset to illustrate the feature distributions.

c) Data cleaning, Data integration, Data transformation, Data model building (e.g. Classification)

**S/W Packages and H/W apparatus used:** Linux OS: Ubuntu/Windows ,

Jupyter notebook.

**Theory/Methodology:**

a) **Summary Statistics:**

Summary statistics provide a concise overview of the characteristics of a dataset. They help in understanding the central tendency, dispersion, and shape of the data distribution. Minimum and Maximum Value: These represent the lowest and highest values observed in a feature.

- Mean: The arithmetic average of the feature values.
- Range: The difference between the maximum and minimum values.
- Standard Deviation: A measure of the dispersion or spread of the values from the mean.
- Variance: The average of the squared differences from the mean.
- Percentiles: Values below which a given percentage of observations fall.

These statistics provide insights into the distribution and variability of data, aiding in understanding the dataset's characteristics and potential issues like outliers.

**b) Data Visualization:**

Data visualization involves representing data visually to better understand patterns, trends, and distributions. Histograms are a common visualization tool for exploring the distribution of individual features. They display the frequency of values within intervals (bins) along the feature's range.Histograms provide a visual summary of the distribution's shape, central tendency, and spread, helping to identify patterns such as normality, skewness, or multimodality.

**c) Data Operations:**

Data operations involve various tasks aimed at preparing data for analysis and modeling:

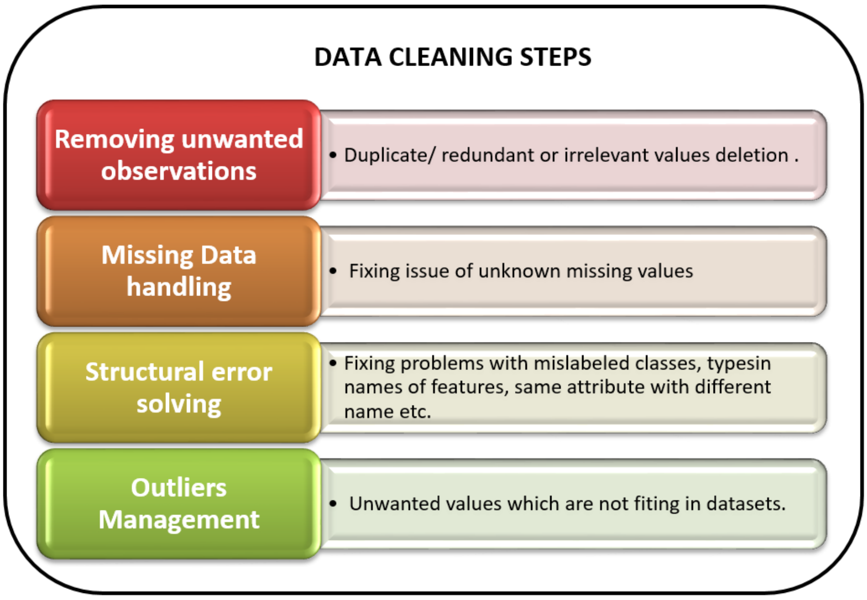
**Data Cleaning**: Removing or correcting errors, dealing with missing values, and handling outliers to ensure data quality.

**Data Integration:** Combining data from multiple sources into a single dataset for analysis.

**Data Transformation**: Converting data into a suitable format for analysis, including normalization, encoding categorical variables, or transforming variables to meet assumptions of statistical models.

**Data Model Building**: Developing predictive or descriptive models using machine learning or statistical techniques. This involves selecting appropriate models, training them on data, and evaluating their performance.

Each of these steps is crucial in the data analysis process to ensure that the resulting models are accurate, interpretable, and generalizable.


DATA CLEANING STEPS
Removing unwanted observations
• Duplicate/ redundant or irrelevant values deletion .
Missing Data handling
• Fixing issue of unknown missing values
Structural error solving
• Fixing problems with mislabeled classes, typesin names of features, same attribute with different name etc.
Outliers Management
• Unwanted values which are not fiting in datasets.

**Advantages/Application:**

a) **Business Analytics and Intelligence:**

1. **Market Analysis:** Analyzing sales data to understand customer preferences, market trends, and competitor performance.
2. **Customer Segmentation:** Using demographic and behavioral data to segment customers for targeted marketing campaigns.
3. **Financial Analysis:** Analyzing financial data to assess business performance, identify risks, and make investment decisions.

b) **Healthcare and Biomedical Research:**

1. **Clinical Trials:** Analyzing patient data to evaluate the effectiveness and safety of medical treatments and interventions.
2. **Epidemiological Studies**: Studying disease patterns and risk factors using population health data.
3. **Genomics and Bioinformatics**: Analyzing genetic data to identify disease markers, understand genetic predispositions, and develop personalized medicine approaches.

**c) Manufacturing and Operations:**

1. **Quality Control:** Analyzing production data to detect defects, optimize processes, and ensure product quality.
2. **Supply Chain Management**: Analyzing inventory data to optimize inventory levels, reduce costs, and improve supply chain efficiency.
3. **Predictive Maintenance**: Analyzing equipment sensor data to predict and prevent equipment failures, minimizing downtime and maintenance costs

**Limitations:**

- Depending on the size of the dataset, certain operations like computing summary statistics or building complex models may require significant computational resources.

- Data cleaning and preprocessing steps can be time-consuming, especially for large and messy datasets.

- The choice of classification model and its performance may vary depending on the specific characteristics of the dataset and the problem being addressed.

**Working :**

1. Load the dataset into a pandas DataFrame.

2. Compute summary statistics using the `describe()` function.

3. Create histograms for each feature using matplotlib.

4. Perform data cleaning, integration, and transformation as needed.

5. Split the dataset into training and testing sets.

6. Choose a classification algorithm and train the model using scikit-learn.

7. Evaluate the model's performance using appropriate metrics.

8. Iterate on the model by tuning hyperparameters or trying different algorithms if necessary.

**Conclusion:**

the practical operations provide a structured approach to extract insights from data. By computing summary statistics, creating visualizations, and performing data operations, practitioners gain understanding, identify patterns, and prepare data for analysis. These practices are crucial across industries, enabling informed decision-making and addressing complex challenges in a data-driven world.

# ASSIGNMENT NO. 3

**Title**: Apply appropriate ML algorithm on a dataset collected in a cosmetics shop showing details of customers to predict customer response for special offers.

**S/W Packages and H/W apparatus used:** Linux OS: Ubuntu/Windows ,

Jupyter notebook.

**Theory :**

**Logistic regression:**

Logistic regression is used for binary classification where we use sigmoid function, that takes input as independent variables and produces a probability value between 0 and 1.

For example, we have two classes Class 0 and Class 1 if the value of the logistic function for an input is greater than 0.5 (threshold value) then it belongs to Class 1 it belongs to Class 0. It's referred to as regression because it is the extension of linear regression but is mainly used for classification problems.

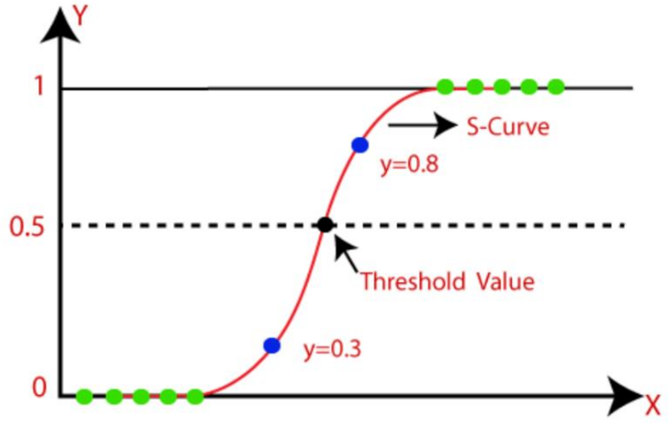
**Logistic Function – Sigmoid Function**

The sigmoid function is a mathematical function used to map the predicted values to probabilities.

It maps any real value into another value within a range of 0 and 1. The value of the logistic regression must be between 0 and 1, which cannot go beyond this limit, so it forms a curve like the "S" form.

The S-form curve is called the Sigmoid function or the logistic function.

In logistic regression, we use the concept of the threshold value, which defines the probability of either 0 or 1. Such as values above the threshold value tends to 1, and a value below the threshold values tends to 0.



**Advantages:**

- Logistic Regression is a simple and interpretable algorithm.
- It works well for binary classification problems like this.
- Easy to implement and understand.

**Limitations:**

- Logistic Regression may not capture complex relationships between features.
- It assumes linear decision boundaries, which may not always be appropriate for the data.

**Working**:

1. Load the dataset and preprocess it.
2. Split the data into features (X) and target variable (y).
3. Split the data into training and testing sets.
4. Initialize and train a Logistic Regression model using the training data.
5. Evaluate the model's performance using metrics such as accuracy, precision, recall, or F1-score.
6. Use the trained model to predict customer responses for new data.

**Conclusion**:

By applying machine learning algorithms like Logistic Regression to the dataset collected in the cosmetics shop, we can predict customer responses for special offers. While Logistic Regression offers simplicity and interpretability, it's essential to consider other algorithms for potentially better performance, especially for more complex datasets. Additionally, careful evaluation of the model's performance and tuning of parameters can lead to improved predictions.

# ASSIGNMENT NO. 4

**Title**: Write a program to do following:

We have given a collection of 8 points. P1=[0.1,0.6] P2=[0.15,0.71] P3=[0.08,0.9] P4=[0.16,

0.85] P5=[0.2,0.3] P6=[0.25,0.5] P7=[0.24,0.1] P8=[0.3,0.2]. Perform the k-mean clustering

with initial centroids as m1=P1=Cluster#1=C1 and m2=P8=cluster#2=C2.

Answer the following:

a) Which cluster does P6 belong to?

b) What is the population of a cluster around m2?

c) What is the updated value of m1 and m2?

**Theory/Methodology:**

K-Means Clustering is an Unsupervised Learning algorithm, which groups the unlabeled dataset into different clusters. Here K defines the number of pre-defined clusters that need to be created in the process, as if K=2, there will be two clusters, and for K=3, there will be three clusters, and so on.

It is an iterative algorithm that divides the unlabeled dataset into k different clusters in such a way that each dataset belongs only one group that has similar properties.

It allows us to cluster the data into different groups and a convenient way to discover the categories of groups in the unlabeled dataset on its own without the need for any training.

It is a centroid-based algorithm, where each cluster is associated with a centroid. The main aim of this algorithm is to minimize the sum of distances between the data point and their corresponding clusters.

The algorithm takes the unlabeled dataset as input, divides the dataset into k-number of clusters, and repeats the process until it does not find the best clusters. The value of k should be predetermined in this algorithm.

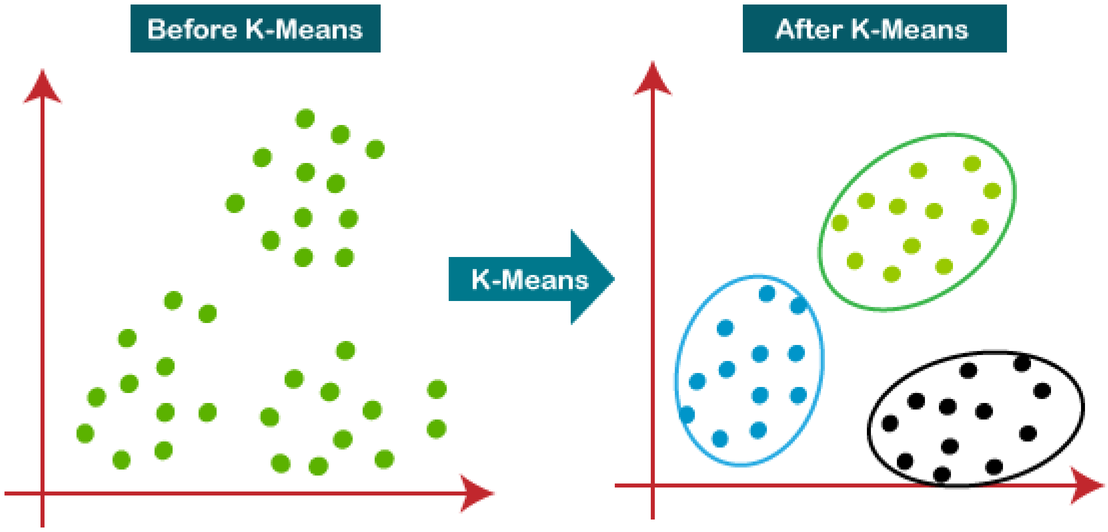
The k-means clustering algorithm mainly performs two tasks:

- Determines the best value for K center points or centroids by an iterative process.

- Assigns each data point to its closest k-center. Those data points which are near to the particular k-center, create a cluster.

Hence each cluster has datapoints with some commonalities, and it is away from other clusters.

The below diagram explains the working of the K-means Clustering Algorithm:



## How does the K-Means Algorithm Work?

The working of the K-Means algorithm is explained in the below steps:

**Step-1:** Select the number K to decide the number of clusters.

**Step-2:** Select random K points or centroids. (It can be other from the input dataset).

**Step-3:** Assign each data point to their closest centroid, which will form the predefined K clusters.

**Step-4:** Calculate the variance and place a new centroid of each cluster.

**Step-5:** Repeat the third steps, which means reassign each datapoint to the new closest centroid of each cluster.

**Step-6:** If any reassignment occurs, then go to step-4 else go to FINISH.

**Step-7**: The model is ready.

**Advantages:**

- Simple and easy to implement.
- Efficient for large datasets.
- Works well with data that is well-separated into clusters.

**Limitations/Examples:**

- Requires the number of clusters (k) to be specified in advance.
- Sensitive to the initial placement of centroids, which can affect the final clustering result.
- May not perform well with clusters of different sizes or densities.

**Working steps:**

1. Initialize centroids (m1 and m2) based on given points P1 and P8.
2. Assign each point to the nearest centroid (cluster) based on Euclidean distance.
3. Calculate the mean of the points in each cluster to update the centroids.
4. Repeat steps 2 and 3 until convergence (centroids do not change significantly).
5. After convergence, determine:
   - Which cluster P6 belongs to.
   - Population of the cluster around m2.
   - Updated values of m1 and m2.

**Conclusion**:

K-Means clustering is a powerful algorithm for partitioning data into clusters based on similarity. In this practical, we implemented K-Means clustering on a collection of points and answered specific questions regarding cluster assignments, cluster populations, and centroid updates. This algorithm provides a straightforward approach to cluster analysis, but it's essential to consider its limitations and understand the impact of parameter choices on the clustering results.

# ASSIGNMENT NO. 5

**Title**: Visualize the data using R/Python by plotting the graphs for assignment no. 1 and 2. Consider a suitable data set.

a) Use Scatter plot, bar plot, Box plot and Histogram

OR

b) Perform the data visualization operations using Tableau for the given dataset.

**S/W Packages and H/W apparatus used:** Linux OS: Ubuntu/Windows ,

Jupyter notebook.

**Theory/Methodology:**

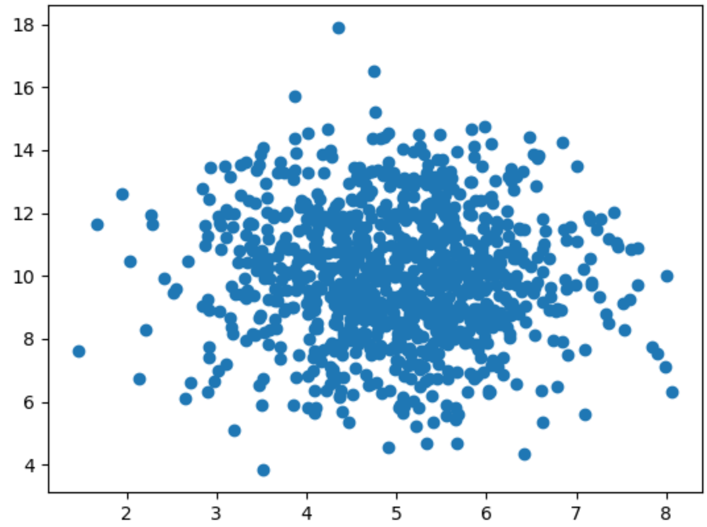
Data visualization is a critical aspect of data analysis, allowing us to explore relationships, patterns, and distributions within the data. In this practical, we'll utilize Python libraries like Matplotlib and Seaborn to create visualizations such as scatter plots, bar plots, box plots, and histograms.

**Scatter Plot:**

Purpose: A scatter plot displays the relationship between two continuous variables. Each point on the plot represents a single observation.

Usage: Scatter plots are useful for visualizing patterns, trends, and correlations between variables. They help identify relationships such as linear, non-linear, positive, negative, or no correlation.

Insights: Scatter plots can reveal clusters of data points, outliers, and the overall distribution of the data.

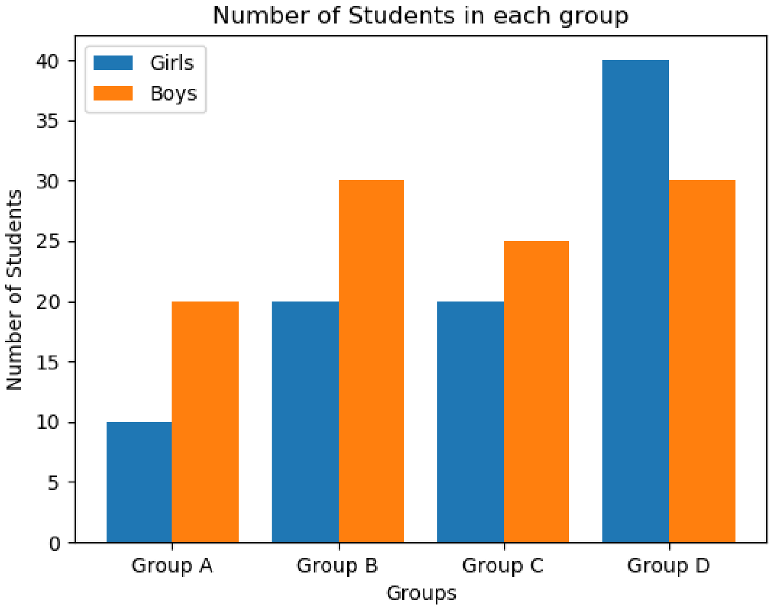


**Bar Plot:**

Purpose: A bar plot displays the distribution of categorical variables or the relationship between a categorical variable and a continuous variable.

Usage: Bar plots are commonly used to compare the frequency or proportion of different categories. They can also show the mean or median value of a continuous variable across different categories.

Insights: Bar plots make it easy to compare values across categories, identify dominant categories, and visualize trends or patterns in categorical data.

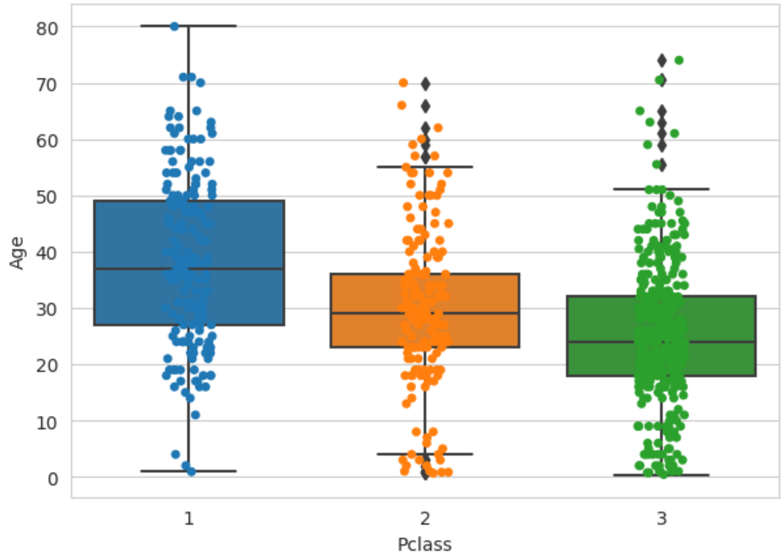


**Box Plot (Box-and-Whisker Plot):**

Purpose: A box plot displays the distribution of a continuous variable and highlights important summary statistics such as the median, quartiles, and outliers.

Usage: Box plots are useful for visualizing the spread and variability of data, detecting outliers, and comparing distributions between different groups or categories.

Insights: Box plots provide a visual summary of the central tendency, spread, and skewness of the data distribution. They help identify variability, symmetry, and the presence of extreme values.

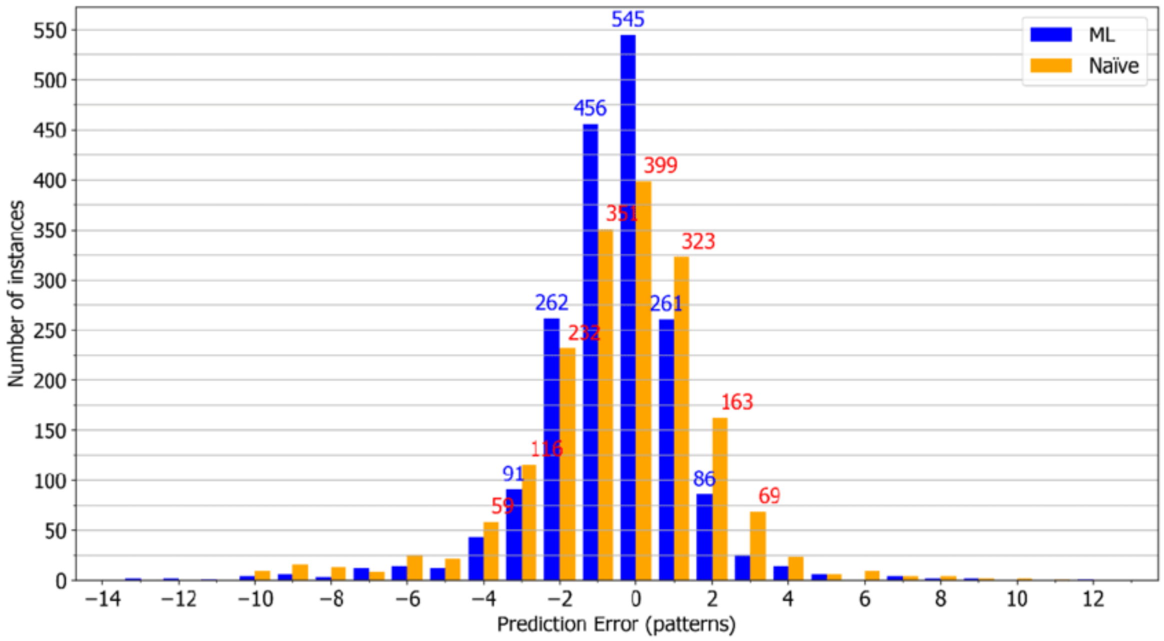


## Histogram:

Purpose: A histogram displays the distribution of a single continuous variable by dividing the data into bins and counting the frequency of observations in each bin.

Usage: Histograms are used to visualize the shape, central tendency, and spread of a data distribution. They help identify patterns such as normality, skewness, multimodality, or outliers.

Insights: Histograms provide insights into the frequency and density of data values across the range of the variable. They are particularly useful for exploring the overall distribution and detecting deviations from expected patterns.



**Advantages:**

**Feature Analysis:**

Visualizations help in understanding the relationships between input features and the target variable. Scatter plots can reveal linear or nonlinear relationships, while box plots can highlight differences in distributions between classes or categories.

**Model Evaluation:**

Visualizations aid in evaluating model performance. For example, histograms of model predictions can reveal if the model is biased towards certain predictions or if it is well-calibrated.

**Model Diagnostics:**

Box plots and scatter plots can help diagnose issues with models, such as heteroscedasticity (unequal variance) or outliers in residuals.

**Working/Algorithm:**:

1. Scatter Plot: Display the relationship between two numerical variables using points on a 2D plane.

2. Bar Plot: Represent categorical data with rectangular bars, where the length of each bar corresponds to the value of a variable.

3. Box Plot: Visualize the distribution of a numerical variable through quartiles, median, and outliers.

4. Histogram: Display the distribution of a numerical variable by dividing the data into intervals (bins) and counting the number of observations in each bin.

**Conclusion**:

By utilizing Python libraries like Matplotlib and Seaborn, we can create informative visualizations to explore and communicate insights from the data effectively. Each type of plot offers unique advantages in representing different aspects of the data, allowing for a comprehensive analysis. However, it's essential to choose the appropriate visualization techniques based on the nature of the data and the insights we aim to convey.

# ASSIGNMENT NO. 6

**Title**: Assignment on Regression technique.

Download temperature data from the link below.

https://www.kaggle.com/venky73/temperaturesof-india?select=temperatures.csv

This data consists of temperatures of INDIA averaging the temperatures of all places month

wise. Temperatures values are recorded in CELSIUS

a) Apply Linear Regression using a suitable library function and predict the Month-wise

temperature.

b) Assess the performance of regression models using MSE, MAE and R-Square metrics

c) Visualize a simple regression model.

**S/W Packages and H/W apparatus used:** Linux OS: Ubuntu/Windows ,

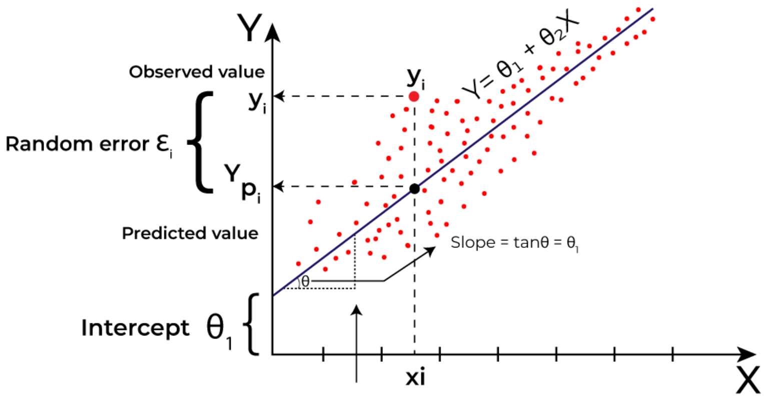
Jupyter notebook.

**Theory:**

**Linear Regression**

It is a statistical method that is used for predictive analysis. Linear regression makes predictions for continuous/real or numeric variables such as sales, salary, age, product price, etc.

Linear regression algorithm shows a linear relationship between a dependent (y) and one or more independent (y) variables, hence called linear regression. Since linear regression shows the linear relationship, which means it finds how the value of the dependent variable is changing according to the value ofthe independent variable.


Y
Observed value
$y_i$
$y_i$
$Y = \theta_1 + \theta_2 X$
Random error $\varepsilon_i$
$Y_{p_i}$
Predicted value
Slope = tanθ = $\theta_1$
θ
Intercept $\theta_1$
xi
X

## Types of Linear Regression

● Simple Linear Regression:

If a single independent variable is used to predict the value of a numerical dependent variable, then such a Linear Regression algorithm is called Simple Linear Regression.

● Multiple Linear regression:

If more than one independent variable is used to predict the value of a numerical dependent variable, then such a Linear Regression algorithm is called Multiple Linear Regression.

Assumptions of Linear Regression

To conduct a simple linear regression, one has to make certain assumptions about the data. This is because it is a parametric test. The assumptions used while performing a simple linear regression are as follows:

● Homogeneity of variance (homoscedasticity)- One of the main predictions in a simple linear regression method is that the size of the error stays constant. This simply means that in the value of the independent variable, the error size never changes significantly.

● Independence of observations- All the relationships between the observations are transparent, which means that nothing is hidden, and only valid sampling methods are used during the collection of data.

● Normality- There is a normal rate of flow in the data. These three are the assumptions of regression methods.

However, there is one additional assumption that has to be taken into consideration while specifically conducting a linear regression.

● The line is always a straight line- There is no curve or grouping factor during the conduction of a linear regression. There is a linear relationship between the variables (dependent variable and independent variable). If the data fails the assumptions of homoscedasticity or normality, a nonparametric test might be used. (For example, the Spearman rank test)

**Applications of Simple Linear Regression**

● 1. Marks scored by students based on number of hours studied (ideally)- Here marks scored in exams are dependent and the number of hours studied is independent.

● 2. Predicting crop yields based on the amount of rainfall- Yield is a dependent variable while the measure of precipitation is an independent variable.

● 3. Predicting the Salary of a person based on years of experience- Therefore, Experience becomes the independent variable while Salary turns into the dependent variable.

**Limitations of Simple Linear Regression**

Indeed, even the best information doesn't recount a total story. Regression investigation is ordinarily utilized in examinations to establish that a relationship exists between variables. However, correlation isn't equivalent to causation: a connection between two variables doesn't mean one causes the other to occur. Indeed, even a line in a simple linear regression that fits the information focuses well may not ensure a circumstances and logical results relationship.Utilizing a linear regression model will permit you to find whether a connection between variables exists by any means. To see precisely what that relationship is and whether one variable causes another, you will require extra examination and statistical analysis.

**Conclusion**:

Simple linear regression is a regression model that figures out the relationship between one independent variable and one dependent variable using a straight line. By applying linear regression to historical temperature data, we can predict month-wise temperatures in India. Evaluation metrics such as MSE, MAE, and R-Squared provide insights into the performance of the regression model. Additionally, visualizing the regression model enhances our understanding of the relationship between variables and the accuracy of predictions.

# ASSIGNMENT NO. 7

**Title**: Assignment on Classification technique

Every year many students give the GRE exam to get admission in foreign Universities. The data set contains GRE Scores (out of 340), TOEFL Scores (out of 120), University Rating (out of 5), Statement of Purpose strength (out of 5), Letter of Recommendation strength (outof 5), Undergraduate GPA (out of 10), Research Experience (0=no, 1=yes), Admitted (0=no, 1=yes). Admitted is the target variable.

Data Set: https://www.kaggle.com/mohansacharya/graduate-admissions

The counselor of the firm is supposed to check whether the student will get an admission or,not based on his/her GRE score and Academic Score. So to help the counselor to take appropriate decisions, build a machine learning model classifier using a Decision tree to predict whether a student will get admission or not.

a) Apply Data pre-processing (Label Encoding, Data Transformation....) techniques if necessary.

b) Perform data-preparation (Train-Test Split)

c) Apply Machine Learning Algorithm

d) Evaluate Model.

**S/W Packages and H/W apparatus used:** Linux OS: Ubuntu/Windows ,

Jupyter notebook.

## Theory/Methodology:

### Classification:

Classification is a process of categorizing a given set of data into classes, It can be performed on both structured or unstructured data. The process starts with predicting the class of given data points. The classes are often referred to as target, label or categories.

What is a Decision Tree?

It uses a flowchart like a tree structure to show the predictions that result

from a series of feature-based splits. It starts with a root node and ends

with a decision made by leaves.

Root Nodes – It is the node present at the beginning of a decision tree.

From this node the population starts dividing according to various features.

Decision Nodes – the nodes we get after splitting the root nodes are called Decision Node

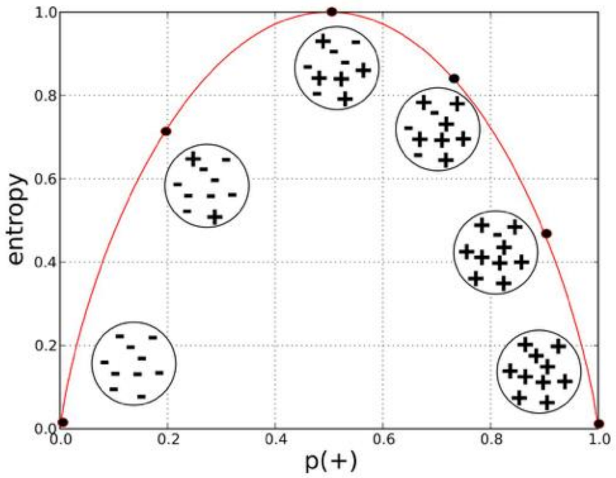
Leaf Nodes – the nodes where further splitting is not possible are called leaf nodes or terminal nodes

Sub-tree – just like a small portion of a graph is called sub-graph similarly a subsection of this the decision tree is called a sub-tree.

Pruning – It is cutting down some nodes to stop overfitting

Entropy:

Entropy is used to calculate the homogeneity of a sample. If the sample is completely homogeneous the entropy is zero and if the sample is equally divided it has entropy of one.

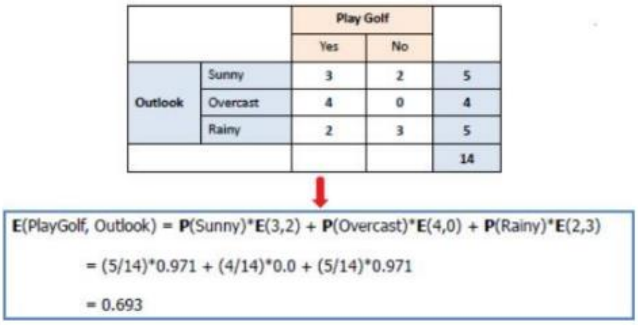


a) Entropy using the frequency table of one attribute:

$$E(S) = \sum_{i=1}^{c} - p_i \log_2 p_i$$

b) Entropy using the frequency table of two attributes:

$$E(T,X) = \sum_{c \in X} P(c)E(c)$$



| | | Play Golf | | |
|---|---|---|---|---|
| | | Yes | No | |
| Outlook | Sunny | 3 | 2 | 5 |
| | Overcast | 4 | 0 | 4 |
| | Rainy | 2 | 3 | 5 |
| | | | | 14 |

**E**(PlayGolf, Outlook) = **P**(Sunny)***E**(3,2) + **P**(Overcast)***E**(4,0) + **P**(Rainy)***E**(2,3)

= (5/14)*0.971 + (4/14)*0.0 + (5/14)*0.971

= 0.693

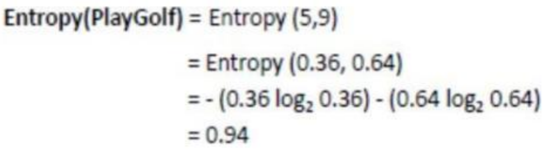
## Information Gain

The information gain is based on the decrease in entropy after a dataset is split on an attribute.

## Constructing a decision tree

IS all about finding attributes that return the highest information gain (i.e., the most homogeneous branches)

Step 1: Calculate entropy of the target.



**Entropy(PlayGolf)** = Entropy (5,9)

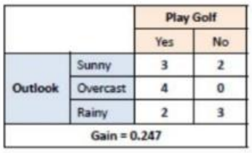
= Entropy (0.36, 0.64)

= - (0.36 $\log_2$ 0.36) - (0.64 $\log_2$ 0.64)

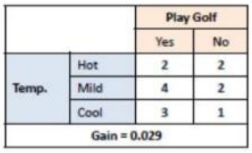
= 0.94

Step 2: The dataset is then split on the different attributes. The entropy for each branch is calculated.

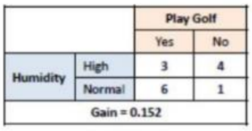
Then it is added proportionally, to get total entropy for the split. The resulting entropy is subtracted from the entropy before the split. The result is the Information Gain, or decrease in entropy.



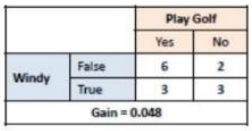
| | | Play Golf | |
|---|---|---|---|
| | | Yes | No |
| Outlook | Sunny | 3 | 2 |
| | Overcast | 4 | 0 |
| | Rainy | 2 | 3 |
| Gain = 0.247 | | | |



| | | Play Golf | |
|---|---|---|---|
| | | Yes | No |
| Temp. | Hot | 2 | 2 |
| | Mild | 4 | 2 |
| | Cool | 3 | 1 |
| Gain = 0.029 | | | |

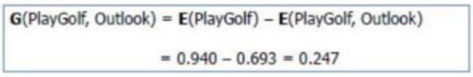


| | | Play Golf | |
|---|---|---|---|
| | | Yes | No |
| Humidity | High | 3 | 4 |
| | Normal | 6 | 1 |
| Gain = 0.152 | | | |



| | | Play Golf | |
|---|---|---|---|
| | | Yes | No |
| Windy | False | 6 | 2 |
| | True | 3 | 3 |
| Gain = 0.048 | | | |

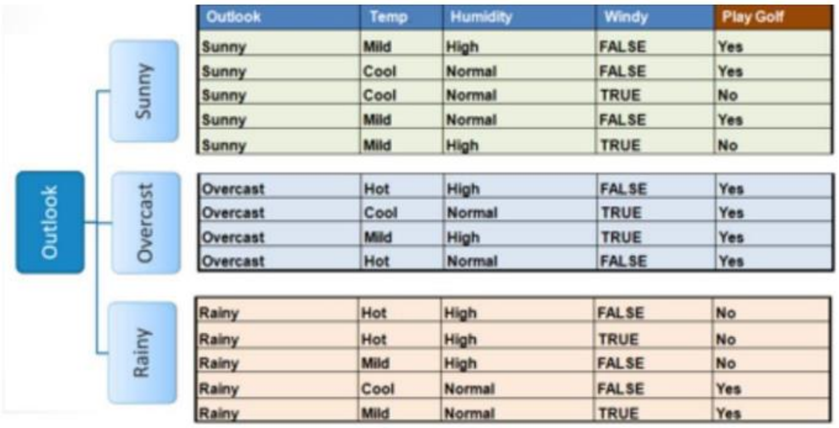
$$Gain(T,X) = Entropy(T) - Entropy(T,X)$$



**G**(PlayGolf, Outlook) = **E**(PlayGolf) – **E**(PlayGolf, Outlook)

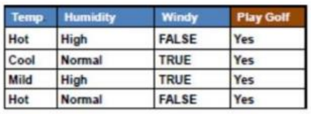
= 0.940 – 0.693 = 0.247

Step 3: Choose the attribute with the largest information gain as the decision node, divide the dataset by its branches and repeat the same process on every branch.

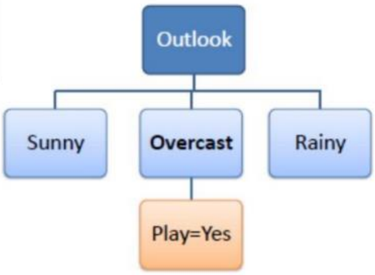


| Outlook | Temp | Humidity | Windy | Play Golf |
|---|---|---|---|---|
| Sunny | Mild | High | FALSE | Yes |
| Sunny | Cool | Normal | FALSE | Yes |
| Sunny | Cool | Normal | TRUE | No |
| Sunny | Mild | Normal | FALSE | Yes |
| Sunny | Mild | High | TRUE | No |
| Overcast | Hot | High | FALSE | Yes |
| Overcast | Cool | Normal | TRUE | Yes |
| Overcast | Mild | High | TRUE | Yes |
| Overcast | Hot | Normal | FALSE | Yes |
| Rainy | Hot | High | FALSE | No |
| Rainy | Hot | High | TRUE | No |
| Rainy | Mild | High | FALSE | No |
| Rainy | Cool | Normal | FALSE | Yes |
| Rainy | Mild | Normal | TRUE | Yes |

Step 4a: A branch with entropy of 0 is a leaf node.



| Temp | Humidity | Windy | Play Golf |
|---|---|---|---|
| Hot | High | FALSE | Yes |
| Cool | Normal | TRUE | Yes |
| Mild | High | TRUE | Yes |
| Hot | Normal | FALSE | Yes |



Step 4b: A branch with entropy more than 0 needs further splitting.

Step 5: The ID3 algorithm is run recursively on the non-leaf branches, until all data is classified.

**Advantages:**

- Decision Trees are easy to understand and interpret.
- They can handle both numerical and categorical data.
- Decision Trees implicitly perform feature selection.

**Limitations:**

- Decision Trees are prone to overfitting, especially with complex datasets.
- They may not perform well with imbalanced datasets.

**Conclusion:**

Classification techniques help in classifying problems and helps figure out relationship between the various variables. By applying a Decision Tree classifier to the admission data, we can predict whether a student will get admission based on their academic scores. Decision Trees offer a simple and interpretable solution for classification tasks. However, it's essential to evaluate the model's performance and consider potential limitations such as overfitting. Additionally, feature selection and data preprocessing play crucial roles in improving the accuracy and generalization of the model.