Date : / /20

Practical No:01

Aim:- Perform Data preprocessing On given Dataset

Apparatus Syder

pandas, numpy, Scipy, matplotlib, python machine learning.

Theory :-

Data preprocessing :-

is process of preparing the raw data and making it suitable for a machine learning model. It is first and crucial step while creating a machine learning model.

When creating a machine learning project it is not always a case that we come across the clean and formatted data. And while doing any operation with data, it is mandatory to clean it and put in a formatted way. So, for this we use data preprocessing task.

Behaviour of data :-

A real-world data generally contains noises, missing values, and maybe in an unusable format which cannot be directly used for machine learning. Dataprocessing helps to increases the accuracy & efficiency of a machine learning model.

Boxplot

It is a type of chart that depicts a group of numerical data through their quartiles. It is a simple way to visualize the shape of the data. Boxplot gives a five-number summary of a set data.

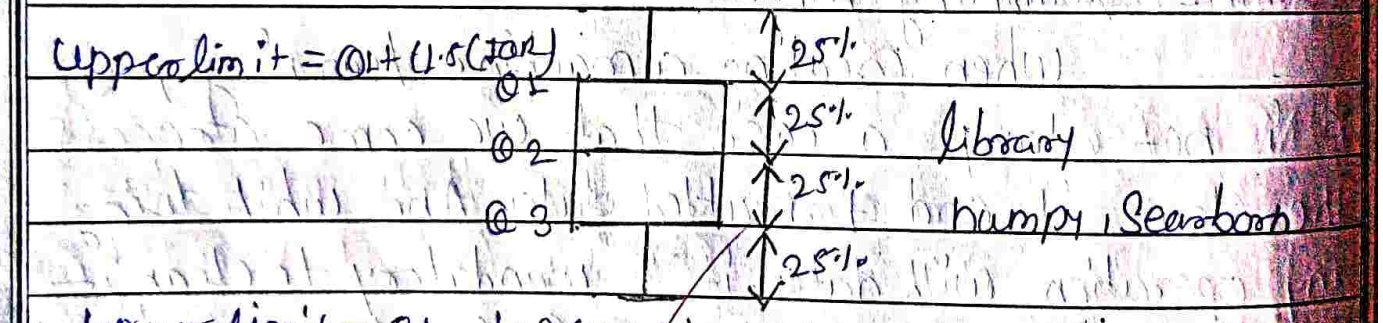
minimum - It is the minimum value in the dataset excluding outliers

First Quartile (Q1) - 25% of data lies below it.

median (Q2) - It is mid point of dataset

Third Quartile - 75% of the data lies below it.

maximum - It is the maximum value in the dataset excluding outliers.



Histogram

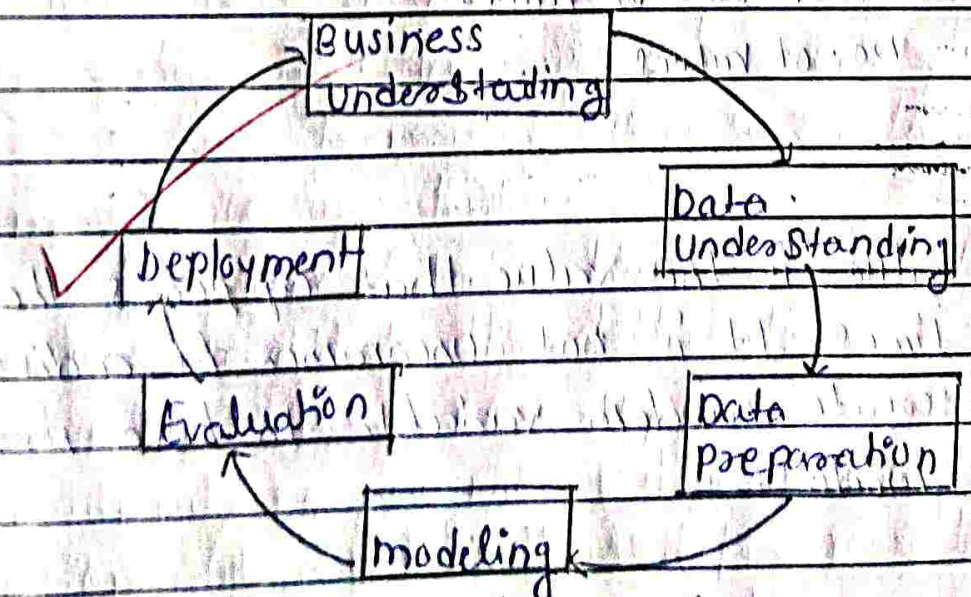
Histogram group the data in bins and is the fastest way to get idea about the distribution of each attribute in dataset.

- It provides us a count of the number of observations in each bin used / created for visualization.

Date : / / 20

- From the shape of the bin, we can easily observe the distribution line, weather it is Gaussian, Skewed or exponential.
- Histogram also helps us to see possible outliers.

CRISP C (Cross-Industry Standard process) for Standard process for Data mining, is an industry proven way to guide your data mining efforts. As methodology, it includes descriptions of the typical phases of a project, the tasks involved with each phase, and an explanation of relation-ship between these tasks. CRISP-DM provides an overview of the data mining life cycle. The CRISP-DM model is flexible and can be customized easily.



Data Mining Life Cycle.

Date : / /20

Mean

The mean value is the average value. To find the mean, find the sum of all values, and divide by the number of values.

formula. $\frac{\sum_{i}^{n} x_i}{N}$

median

The median of a distribution with a random variable depends on whether the number terms in distribution is even or odd. If the number of terms is odd, then the median is the value of the middle. for even or equal number it is difficult.

formula $x\left[\frac{n+1}{2}\right]$ if n is odd

$\frac{x\left[\frac{n}{2}\right] + x\left[\frac{n}{2}+1\right]}{2}$ if n is even.

x - ordered list of values

n - no. of values.

mode

The mode value is the value that appears the most of times. It is not uncommon for a distribution with a discreate random variable to have more than one mode.

Standard Deviation

Standard Deviation is a number that describes how spread out the values are. A low standard deviation means that the most of the numbers are very close to the mean (average) value. A high standard deviation means values are spread out over a wider range.

$$\text{Standard Deviation} = \sqrt{\frac{\Sigma (x-\mu)^2}{N}}$$

- Use the NumPy std() method to find the standard deviation.

Variance

Variance is another number that indicates how spread out values are, in fact, if you take square root of the variance, you get the standard deviation.

$$\text{Variance} = \frac{\Sigma (x-\mu)^2}{N}$$

- Use Numpy var() method to find variance.

Date : / /20

Scatter plot:-

In a dataset, for k set of variables/columns $(X_1, X_2, \ldots X_k)$, the scatter plot all the pairwise scatter between different variables in the form of a matrix.

It gives Answer of following questions like:

i) Are there any outlier in the dataset?

ii) Is there any clustering by groups present in the dataset on the basis of particular variable.

practical No. 02

Aim :- Perform Data cleaning On given Dataset.

Apparatus :- Sydeon library - numpy, pandas, seaborn, winsorize

Theory :-

Data cleaning is a crucial process in Data Mining. It carries an important part in the building of a model. Data cleaning can be regarded as the process needed, but everyone often neglects it. Data quality is main issue in quality information management. Data quality problems occur anywhere in information Systems. These problems are solved by Data cleaning.

Data cleaning is fixing or removing incorrect, corrupted, incorrectly formatted, duplicate, or incomplete data within a dataset. if data is incorrect outcomes and algorithms are unreliable, even though they may look correct.

Generally, data cleaning reduces errors and improves data quality. Correcting errors in data and eliminating bad records can be a time-consuming and tedious process, but it cannot be ignored. Data mining has various techniques that are suitable for data cleaning.

Null Values

The values which is a unknown or mising are called As Null values. This values lead to inaccurate predication. So, the machine learning dev while deal with data cleaning, it ask Remove or replace this Null values.

isNull() → This will return boolean value for every column in data frame.

isNull.sum() → This code will give you total number of Null values in each features in the dataframe.

How to treat Null Values.

i) you can drop the missing values with the method df.fillna() df.drop(). This will drop all the row which contain the missing value.

ii) you can fill the missing values with method df.fillna

Winsorization

is the process of replacing the extreme values of statistical data in order to limit the effect of the outlier on the calculations or the results obtained by using that data. The mean value calculate after the such replacement of the extreme values is called winsorized mean.

Removing duplicate or irrelevant Observations

Remove duplicate values from your dataset including unwanted / irrelevant values. Duplicate value will be happen most often during date Collection. When you combine data sets from multiple places, Scrape data receive data. there are opportunities to create duplicate that data..

dup

duplicated() → function indicate duplicate Series values. The duplicated values are indicated as True Values in the resulting Series.

drop_duplicates() → helps in Removing duplicates from the pandas Dataframe in python.

Outliers

In the Series normal peaks and valleys occur outside of time frame when that Seasonal sequence is normal or as a combination of time series that is in an outlier state as a group. Outliers detecting using Visualization Once we are able to Visualize the Outlier then it becomes quite easy to decide what actions we can take. We can Remove Outlier using the Winsorization method.

Date : / /20

Cause of outliers.

i) A lot of mispredictions.

ii) Bad performance on train and test data.

iii) Accuracy decrease.

One Hot Encoding

One approch to solve the Non-numeric pad data problem can be Lebal Encoding where we will assign a numerical value to these labels for example male and female mapped to 0 and 1. But this can add bias in our model performance and also since the data is string labels it will start high preferna to the female parameter 1>0 ideally both labels are equally important in the dataset. to deal with this issue we will use One Hot Encoding technique.

Ex. pass ATKT fail ~~Column~~ labels in table.

	Pass	ATKT	fail
1	1	0	0
2	1	0	0
3	0	1	0
4	0	0	1

Lebal Encoding

Lebel Encoding refer to Converting the labels into a numeric form so as to convert them into the machine-readable form. machine learning algorithms can then decide in a better way how those labels must be operated. It is an important pre-prossesing step for the structured dataset in Supervised learning.

Example-

Suppose we have Column Height in some dataset. After applying label encoding, the Height column is converted : where 0 is the label for tall, 1 is label formedium and 2 is label for short height.

Normalization

Normalization is no mandate for all datasets available in machine learning. It used whenever the attributes of the dataset have different ranges. It helps to enhance the performance and reliability of machine learning model. It simply contain values in a range of 0 to 1:

formula =

$$\text{Normalization } (X_n) = \frac{(X - X_{min})}{(X_{max} - X_{min})}$$

X_{max} = maximum value of a feature

X_{min} = minimum value of a feature.

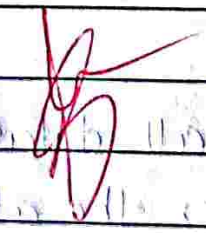
Date : / /20

Standardization.

Standardization is z-score normalization is the transformation of features by subtracting from mean and dividing by standard deviation. This is often called z-score.

$$x\_new = \frac{(x - mean)}{std.\ deviation}$$

Standardization can be helpful in cases where the data follows Gaussian distribution.



Date : / /20

Practical No. 03

Aim :- Apply Hierarchical clustering on given datasets

Apparatus :- Syder, python libraries - numpy, pandas, sklearn, scipy

Theory :-

Hierarchical Clustering

In other clustering algorithm like KMeans Clustering we face some challeges which are a predetermined number of clusters and It always tries to create of the same size. To solve these two challeges, we can use the hierarchical clustering algorithm. It is unsupervised machine learning algorithm.

The hierarchical clustering technique has two Approaches.

1) **Agglomerative**
is a Bottom-up approch, in which the algorithm starts with taking all data points as single clusters and merageing them until one cluster is left.

2) **Divisive**
Divisive algorithm is the reverse of the agglomerative algorithm ~~is the~~ As it is top-Down Approach.

Date : / / 20

Agglomerative Hierarchical Clustering

is a popular example of HCA. To group the data into clusters, it follows the bottom-up approach. It is bottom-up Approach, in which the algorithm starts with taking all the data points as single clusters and merging them until one cluster is left.

The working of AHC

Step-1 → Create each data as single cluster.

Step-2 → Take two closest data points or clusters and merge them to form one cluster.

Step-3 → Again, take two cluster and merge them together to form one cluster.

Step-4 → Repeat 3 until only one cluster left.

Step-5 → Once all clusters combined into one big cluster, develop the dendrogram to divide the cluster as per the problem.

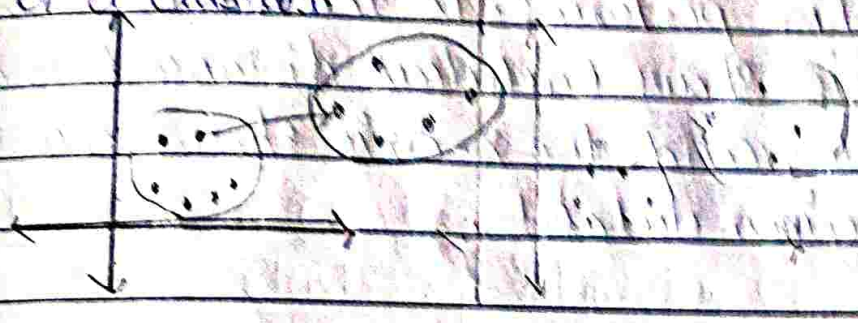
The way of to Cal decide the rule for clustering we use linkage methods.

Date

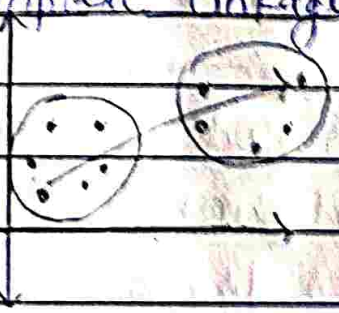
Linkage methods are given below

1) Single Linkage

It is the shortest distance between the closest points of a clusters.

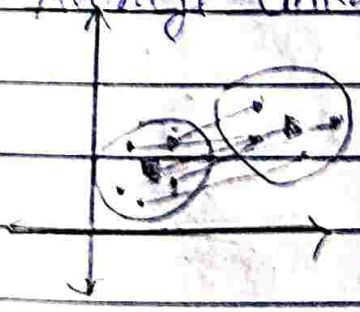


2) Complete Linkage



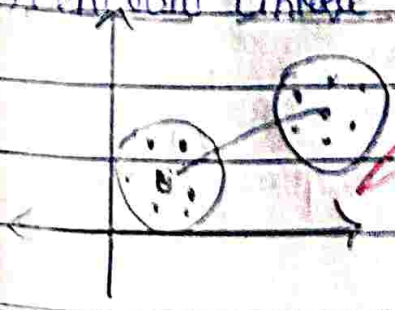
It is the farthest distance between the two points of two different clusters.

3) Average Linkage



It is the linkage method in which the distance between each pair of datasets is added up and then divided by the total number of the dataset to calculate average distance of clust.

4) Centroid Linkage



It is the linkage method in which the distance between the centroid of the cluster is calculated.

Dendrogram

The dendrogram is a tree-like structure that is mainly used to store each step as memory that the HC algorithm performs. In the dendrogram plot, the y-axis shows Euclidean distance between the data points, and the x-axis shows all the points of given dataset.

Practical No. 04

Aim:- Appy k-means clustering on given Dataset

Apparatus:- Syder, numpy, pandas, matplotlib.

Theory:- Sklearn.

K-means Clustering Algorithm

is an ~~se~~ unsupervised learning algorithms that is used to solve the clustering problems in machine learning or data science. It groups the unlabeled dataset into different clusters. Here k define the number of pre-defined clusters that need to be created in the process, as if k=2, 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. 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 distance between the dataset point and their corresponding clusters.

Date : / /20

Working of k-means algorithm.

Step-1 → Select the number of K to decide no. of Cluster

Step-2 → Select random K point Centroids.

Step-3 → Assign each data point to their closet Centroid which will form the predefined K.

Step-4 → Calculate the variance and place a new Centroid of each cluster.

Step-5 → Repeat the third steps, which means data point to the new closest centroid of each other.

Step-6 → if any reassignment occurs, then go to 4 else to FINISH

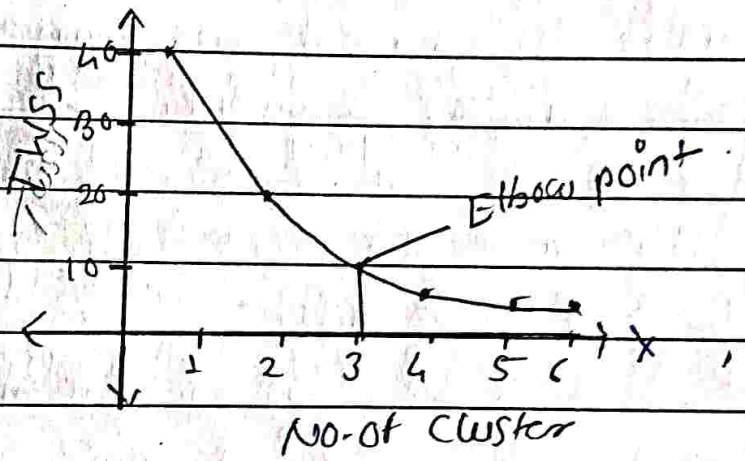
Step-7 → Model is Ready.

Elbow Curve.

The elbow method is a graphical representation of finding the optical 'k' in a k-means clustering. It works by finding TWSS (Total within Sum OF Squares) the sum of the square distance

between points in a cluster and the cluster centroid. The elbow graph shows TWSS values (on the y-axis) corresponding to the different values of k (on the x-axis). When we see an elbow shape in the graph, we pick the k-value where the elbow gets created. We can call this point the Elbow point.

Elbow Curve.



TWSS (Total within Sum of Square)

Twss values were calculated for all the clustering methods used. We compute the distance between each data point in the cluster and the centroid of the cluster for every cluster. This distance is referred to as within-cluster sum of squared errors (wss). We then sum these distances to obtain what we call the total within-cluster sum of squared error (TWSS).

Date : / /20

$$TWSS = (x - \mu)^2$$

Euclidean Distance

It is a distance measure between a pair of sample p and q an n-dimensional features space.

Euclidean distance $= \sqrt{\sum_{i=1}^{n} (q_i - p_i)^2}$

Date : / /20

Practical NO. 05

Aim :- Apply Simple Linear Regression Algorithm of given data set.

Apparatus :- Syder, Python Libries, seaborn, Scipy, Pd lab, Statsmodel.

Theory :-

Simple Linear Regression

is a type of Regression algorithms that model relationship between a dependant variable and a single independant variable. The relationship shown by a simple Linear Regression model is linear sloped straight line, hence it is called Simple Linear Regression. The key point in Simple Linear Regression is that the dependent variable can be measured on continuos or categorial values.

Simple Linear Regression algorithm has mainly two objectives:

1) model the relationship between the two variable. Such as the relationship between income and expenditure ; experiance and Salary etc.

2) Forecasting new observations. Such as weather forecasting according to tempature, Revenue of a Company according investment in a year etc

Equation for represented SLR.

$$y = B_0 + B_1 x + \epsilon$$

where.

B_0 = It is intercept of Regression line.

B_1 = It is the slope of the Regression line, which tells whether the line is increasing

ϵ = The error term.

Confusion matrix

The confusion matrix is a matrix used to determine the performance of the classification models for a given set of test data. It can only determined if the true values for the data are known. The matrix itself can be easily understand but the related terminologies may be confusing. Since it shows the errors in the model performance in the form of matrix, Hence also known as an error matrix.

Date : / /20

		Actual value	
		1	0
Predicted value	1	TP	FP
	0	FN	TN

TP → True negative positive
FP → ~~True~~ false positive
TN → True negative
FN → false negative

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

Regression line

It is used for relationship between dependant & independant values uses.

1) forecasting
2) Trend Analysis
3) To make predication

Type's of Regression line

1) linear
2) logistic
3) Polynomial.

Practical No.06

Aim:- Apply multiple linear Regression algorithm on given dataset.

Apparatus:- Spider, sklearn, ~~k-neighbour~~ multiple ~~regression~~

Theory:-

multiple Linear Regression

is an extension of simple linear regression as it takes more than one predictor variable to predict the response variable. We can define it as: multiple Regression is one of the important regression algorithm which models the linear relationship between a single dependent continues variable and more than one independent variable.

MLR Equation

In multiple linear Regression, the target variable (Y) is a linear combination of multiple predictor variables $x_1, x_2, x_3, \ldots, x_n$. Since it is an enhancement of simple linear Regression. So, the same is applied for the multiple linear regression equation, the equation becomes

Date : / /20

$Y = b_{<sub>} - 0$

$Y = b_0 X_0 +$

$Y = B_0 X_0 + B_1 X_1 + B_2 X_2 + \cdots + B_n X_n$

Where

Y = Output

$b_0, b_1, b_3 \cdots b_n$ = Coefficients of the model

$X_1, X_2 X_3$ = various independent Variable.

Pairplot

Pairplot Visualization Comes handy when you want to go for Exploratory data analysis (EDA) pairplot visualizes given data to find the relationship between them where the Variable Can be Continuous or categorical. plot pairwise relationship in data-set.

pairplot is a module of seaborn library which provides a high-level interface for drawing attractive and informative Statistical graphics.

pairplot is used to understand the best set of features to explain a relationship between two Variable or to form the most separated clusters.

Date : / /20

Residual Leverage plot (Diagnostic plots)

In linear or multiple regression, it is not enough to just fit the model into the dataset. But, it may not give the desired result. To apply the linear or multiple regression efficiently to the dataset.

terms in Regression Diagnostic plots

Outliers

Outliers are the points that are distinct and deviant from the bulk of the dataset.

Leverage points

A leverage point is defined as an observation that has a value x that is far away from mean.

Influential Points

is defined as an observation that has a large influence on the fit of one model.

Residual VS Fitted plot

The residual can be calculated As

$$res = y_{observed} - y_{predicated}$$

Page No. 55

Date : / /20

Practical No. 07

Aim :- Apply K-Nearest Neighbor (KNN) Algorithm on Given Dataset.

Apparatus :- Syder, numpy, pandas, matplotlib, Sklearn, k-neighbor classifer.

Theory :-

K-Nearest Neighbor (KNN) Algorithm.

K-Nearest Neighbor is one of the simplest machine learning algorithms on supervised learning Technique. It assums the similarity between the new case/data and available cases and put the new case into the category that is most similar to the available categories.

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

K-NN algorithm stores all can be used for Regression as well as for classificaton but mostly it is used for the classificaton problems. K-NN is a non-parametric algorithm which means it does not make any assumption on underly data.

Date : / /20

It is also called a lazy learner algorithm because it does not learn from the training set immediately instead it does stores the dataset and at time of classification, it performs an action on the dataset. KNN algorithm at the traing traing phase just stores the dataset and when it gets new data, then it classifies that data into a category that is much similar to new data.

Working of K-NN algorithm :

Step-1 → Select the number k of the neighbors

Step-2 → Calculate the Euclidean distance of k number of neighbors.

Step-3 → Take the k nearest neighbors as per the calculated Euclidean distance.

Step-4) → Among these k neighbor, as per the calculated Euclidean count the number of the data points in each category.

Step-5 → Assign the new data points to that category for which the number of the neighbors is

Date / /20

Advantages of K-NN Algorith.

i) It is simple to implement

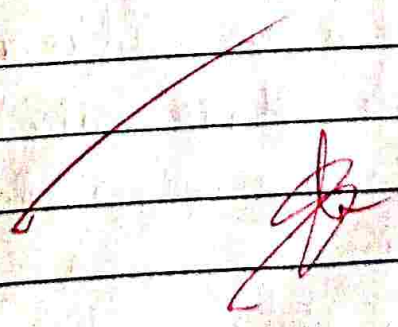
ii) It is robust to the noisy training data.

iii) It can be more effective if the training data is large.

Disadvantages of KNN algorithm.

i) Always needs to determine the value of K which may be complex some time.

ii) The computation cost is high because of calculating the distance between the data points for all the training samples.



Practical No. 08

Aim:- Apply Decision Tree Algorithm on given dataset.

Apparatus:- Sydex, numpy, pandas, matplotlib, sklearn, Decision tree classifier.

Theory:-

Decision Tree Algorithm

Decision Tree is supervised learning technique that can be used for both classification and Regression problems, but mostly it is preferred for solving classification problems. It is a tree-structured classifier, where internal nodes represent the features of a dataset, branches represent the decision rules and each leaf node represents the outcome.

In a Decision tree there are two nodes which are the Decision Node and leaf node. Decision nodes are used to make any decision and have multiple branches, whereas leaf nodes are the output of these decisions and do not contain any further branches. There are various algorithm in machine learning. So choosing the best algorithm for given dataset and problem is the main point to remember while creating a machine learning model.

Date : / /20

Decision Trees usually mimic human thinking ability while making a decision, so it is easy to understand. The logic behind the decision tree can be easily understand because it shows a tree-like structure.

Decision Tree Terminologies

1) Root node → is from where decision tree starts. It represents the entire dataset, which further gets divided into two or more homogeneous sets.

2) Leaf node → are the final output node, and tree cannot be segregated further after getting a leaf node.

3) Splitting → is the process of dividing the tree decision node/root into sub-nodes.

4) Branch / Sub Tree → A tree formed by splitting the tree.

5) Pruning → is the process of removing the unwanted branches from the tree.

6) Parent / child node → The root node of tree is called the parent node, and other nodes are called the child nodes.

Working of Decision Tree Algorithm.

Step-1 → Begin the tree with the root node, Say S. which is complete dataset.

Step-2 → Find the best attribute in the dataset using Attribute Selection Measure (ASM).

Step-3 → Divide the S into subsets that contain possible values for the best attributes.

Step-4 → Generate the decision tree node, which contain best attribute.

Step-5 → Recursively make new decision trees using the subsets of the dataset created.

Step-6 → Continue the decision tree node process until a stage is reached where you cannot further classify the nodes and the final node as a leaf node.

Advantages of Decision Tree

i) It is simple to understand.

ii) It can be very useful for solving decision-related problems.

iii) It helps to think about all the possible outcomes.

Disadvantages of Decision Tree

i) The decision tree contains lots of layers, which makes it complex.

ii) It may be have an overfitting issue.

MEET

Date : / /20

Entropy

Information gain is defined as pattern observed in the dataset and reduction in the entropy

Entropy is defined as the randomness or measuring the disorder of information being in machine learning. entropy is matric that measures the unpredicatability or impurity in the system.

gain = entropy before split - entropy after split

$G(a) = E_B - E_A$.

$$Info_B(D_j) = \sum_{i=1}^{n} p_i \log_i(p_i)$$

$$Info_A(D) = \sum_{i=1}^{n} \frac{|D_j|}{|D|} * info(D_j)$$