**ML Assignment No - 1** Title: . Data preparation:

Name:Bindi Shah Class: TE 09 Batch: K 09

Roll no.: 33110

**Problem Statement:**Download heart dataset from following link. https://www.kaggle.com/zhaoyingzhu/heartcsv

Perform the following operation on a given dataset. a) Find Shape of Data

b) Find Missing Values

c) Find data type of each column d) Finding out Zero's

e) Find Mean age of patients

f) Now extract only Age, Sex, ChestPain, RestBP, Chol.

Randomly divide the dataset in training (75%) and testing (25%). Through the diagnosis test I predicted 100 reports as COVID positive, but only 45 of those were actually positive. Total 50 people in my sample were actually COVID positive. I have a total of 500 samples.

Create a confusion matrix based on the above data and find I. Accuracy II. Precision III. Recall IV. F-1 score

**Objective:** This assignment will help the students to realize what is need of data preparation

**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

**Conclusion:** Data preparation is recognized for helping businesses and analytics to get ready and prepare the data for operations.

**ML Assignment No - 2** Title : Regression technique

Name:Bindi Shah Class: TE 09 Batch: K 09

Roll no.: 33110

**Problem Statement:** This data consists of temperatures of INDIA averaging the temperatures of all places month-wise.

Temperature 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.

**Objective:**This assignment will help the students to realize how Linear Regression can be used and predictions using the same can be performed.

**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 of the independent variable.

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.

**ML Assignment No - 3** Title: Classification using Machine Learning

Name:Bindi Shah Class: TE 09 Batch: K 09

Roll no.: 33110

**Problem Statement:**

Perform following operations on given dataset:

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

b. Perform data-preparation (Train-Test Split) c. Apply Decision tree classification Algorithm d. Evaluate Model.

**THEORY:**

**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:

b) Entropy using the frequency table of two attributes:

**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.

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.

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.

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

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.

**Decision Tree to Decision Rules**

A decision tree can easily be transformed to a set of rules by mapping from the root node to the leaf nodes one by one.

**Pruning:**

It is another method that can help us avoid overfitting. It helps in improving the performance of the tree by cutting the nodes or sub-nodes which are not significant. It removes the branches which have very low importance. There are mainly 2 ways for pruning:

(i) Pre-pruning – we can stop growing the tree earlier, which means we can prune/remove/cut a node if it has low importance while growing the tree. (ii) Post-pruning – once our tree is built to its depth, we can start pruning the nodes based on their significance.

**CONCLUSION:**

Classification techniques help in classifying problems and helps figure out relationship between the various variables.

**ML Assignment No - 5** Title: K Means Clustering

Name:Bindi Shah Class: TE 09 Batch: K 09

Roll no.: 33110

**Problem Statement:**

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.

e) Apply Cross-Validation and Evaluate Model

**Objective:**This assignment will help the students to realize how to do Clustering using the K- Means Clustering algorithm.

**Theory:**

● K-Means Clustering : K-Means Clustering is an unsupervised learning algorithm that is used to solve clustering problems in machine learning or data science

● What is the K-Means Algorithm? :

K-Means Clustering is an Unsupervised Learning algorithm, which groups the unlabeled dataset into different clusters. Here K defines the number of predefined 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 allows us to cluster the data into different groups and is 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 data points with some commonalities, and it is away from other clusters.

Algorithm: 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 reassigning 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

Some examples of use cases are:

1. Behavioral segmentation: Segment by purchase history, Segment by activities on application, website, or platform, Define personas based on interests , Create profiles based on activity monitoring

2. Inventory categorization: Group inventory by sales activity , Group inventory by manufacturing metrics.