ASSIGNMENT 1

Title : Data Preparation

Problem Statement : Perform following operation on given dataset.
(a) Find shape of data
(b) Find missing value
(c) Find datatype of each column.
(d) Find out zero's.
(e) Find mean age of patients.
(f) Now extract only age, sex, chest pain, rest BP, cholestrol, randomly divide dataset in Training (75%) and testing (25%
(g) Through the diagnosis test 1 predicted 100 report as COVID positive. Creat confusion matrix 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.

Theory :
Data Preparation :
It is a process of transforming row data so that data scientists and analysts can run it through M.L algorithms.

Why is data preparation ?

We need data preparation, data set needs to be different and specific according to the model so that we have to find out the required features of data.

1. Determine Problem
2. Data cleaning
3. Feature selection
4. Data transformation
5. Feature engineering
6. Dimensionality reduction

1. Determine Problem :

It tells about learning method of the project to find out r.

2. Data cleaning :

After collecting the data, it's very necessary to clean that data and make it proper for ML model.

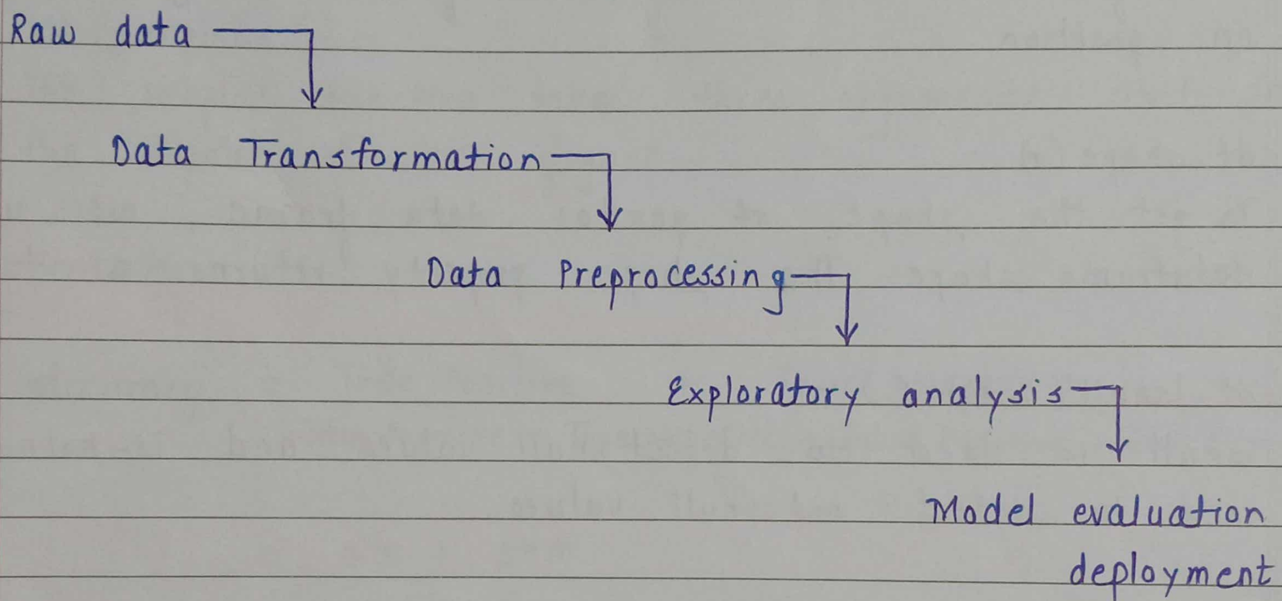
3. Feature selection :

Sometimes we face the problems of identifying related feature from the set of of data and deleting irrelevant and less important data without touching target variables.

4. Data transformation :

It's a process that converts data from one form to another. If ~~is~~ required, we can change type of data.

5. Feature engineering :
Every ML algorithm uses some input for giving required output. and this input requires some features which are in a structured form. To get the proper results the algorithm requires :



Raw data

Data Transformation

Data Preprocessing

Exploratory analysis

Model evaluation deployment

6. Dimensionality reduction :
For ML model, we have to access a large amount of data and that large amount of data can lead in a saturation where we can take possible data that can be available to fed it into a forecasting modes.

* commands :

1. import pandas as pd -
to import the pandas library.

2. pd.read_csv ("filename.csv") -
It is a function used to retrieve data from csv file.

3. df.head (n) -
The head() function is used to get the first n rows. This function returns n rows for the object based on position.

4. df.shape () -
To get the shape of pandas data frame, we use dataframe.shape. The shape property returns a tuple.

5. df.isnull().sum() -
isnull is used to detect null values and is-notnull() used to detect not null values.

6. df.isnull().sum().sum() -
It's used to return total null values of each column.

7. df.dtypes() -
returns a series with the data-types of each column.

8. (df == 0).sum(axis = 0) -
axis = 0 represents rows and axis = 1 represents columns. Therefore to get the sum of values in each row in pandas, this function is used.

9. df.columns -
used to retrieve all columns in the dataset.

10. df 2 = df.filter(['Age'. 'sex', ...])
used to filter out only selective columns from dataset.

11. SNS. countplot (x='AHD', data='df', palette='pastel') -
used to plot a bar graph of specified style.

12. plt.subplots() -
The subplot function takes three arguments that describes the layout of the figure.

* Predictions :

1. Accuracy $= \dfrac{\text{True Positive} + \text{True Negative}}{\text{True Positive} + \text{True Negative} + \text{False Positive} + \text{False Negative}}$

$$= \frac{45 + 395}{500}$$

Accuracy = 88 %

2. Precision - Ratio of correctly ~~clamped~~ classified positive sample to total no. of classified positive sample.

Precision $= \dfrac{\text{True Positive}}{\text{True Positive} + \text{False } \cancel{\text{Negative}} \text{ Positive}}$

$$= \frac{45}{45+55}$$

Precision = ~~45%~~ 45 %

3. Recall : Recall tries / attempts to answer questions - What proportion was identified correctly ?

Ratio between no. of Positive samples correctly classified as positive sample. to total no. of positive samples

$$\text{Recall} = \frac{\text{True Positive}}{\text{True Positive + False Negative}}$$

$$= \frac{45}{45+5}$$

$$= 90\%$$

4. F1 score :

F1 score is weighted average of precision and recall. This score takes both false positive and false negatives into account. It is not as easy to understand as accuracy. Score is more useful than accuracy especially when you have uneven distribution.

$$\text{F1 score} = \frac{2 \times \text{Recall} \times \text{Precision}}{\text{Recall + Precision}}$$

$$= \frac{2 \times 0.45 \times 0.9}{0.45 + 0.9}$$

$$= \frac{81}{135}$$

$$\text{F1 score} = 0.6$$

* Confusion matrix :

It is a table that is used in classification problems to assess where errors in the model were made. It consists of 4 quadrants viz. : True Positive, True Negative, False Positive and False Negative.

TP : Truth is positive and test predicts a positive.
TN : Truth is negative and test predicts a negative.
FP : Truth is ~~positive~~ negative but test predicts a ~~negative~~ positive.
FN : Truth is positive but test predicts a negative.

* Conclusion :

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

C