Experiment 1

Shashwat Shah
60004220126
TY BTech Comps B

Aim: To perform data preprocessing in terms of handling missing data, removing outliers, eliminating outliers, eliminating duplicate rows & modifying the datatypes, etc.

Theory: Python an easy to learn programming language which makes it the most popular & preferred language for beginners in data science, data analytics and machine learning. It also has a great community of online learners and excellant data centric libraries. With so much data being generated it becomes important that the data we use for data science explications like machine learning & predictive modelling is clean. Data cleaning refers to the process of cleaning the dirty data by identifying the errors in data and rectifying them. Data cleaning is hence a very important step in ML.

Data cleaning consists of :-

- Missing values
we will start by calculating the percentage of values missing in each column and storing information in the dataset.

- Drop Observation - One way
we can drop observations that contain only null values in them for any of the columns. This works when

the percentage of missing values in each column is less.

- Remove columns -

Drop column/features which have significant percentage of missing values.

- Imp missing values -

we can also fill missing value in numerical columns, using mean, median, mode and other such structures,

- Outliers

Its an unusual ~~dat~~ observation that is random & wrong. They can affect the ML model significantly.

- Duplicate records

Data can sometimes contain duplicate values. Its important duplicate records from dataset before we process to ML model.

Fixing a datatype

Often values in dataset went stored in the c correct data type.

Procedure

Firstly we load the dataset in the dataframe, then we list columns & the no. of null values in that column. Its observed that age has 177 null values, cabin has 687 null values. Next we fill the null values. we also observ that the cleaned data includes colums age and encoded values of columns embaarked & p class as dependant values & survived column as largest.

Conclusion - Here we conclude that data cleaning is very important before applying data on ML models.