**REPORT ON: DS\_PROJECT INTERNSHIP TEAM 57**

**EDA ANALYSIS FOR MORTALITY RISK RATE DATA SET**

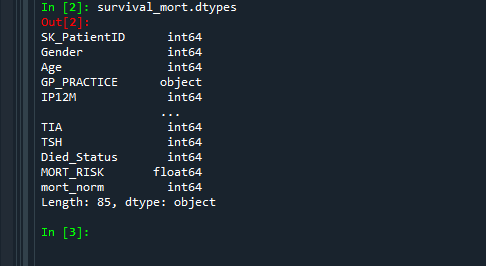
**NAME :SUNKESULA SHAHAN**

**EMAIL:shahan0805@gmail.com**

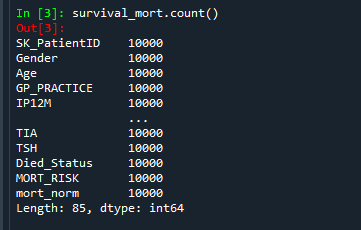
**Mobile:8919009614**

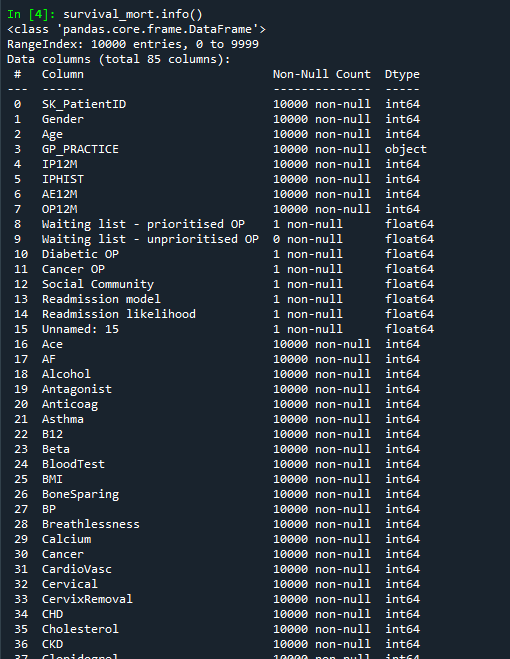
**REPORT:**

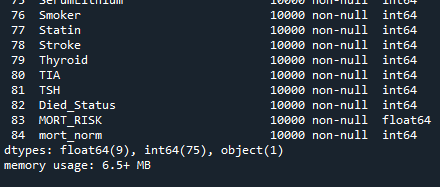
We have loaded the data set,it contains 10000 rows and 85 columns.After reading through the data set we can see majority of the data-set contains binary values as categorical values.if we perform data type analysis of each column we are getting this in python



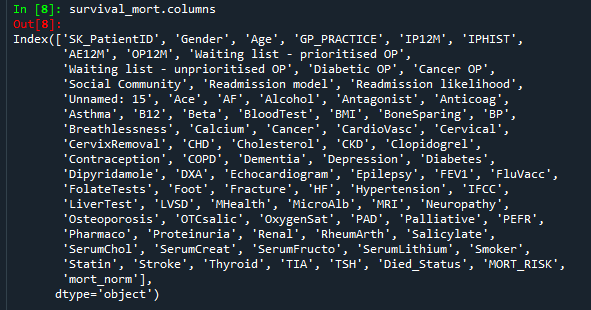
We can see majority of them are int64 datatype columns.some other columns types include float64 and object types.columns contains different characteristics among which for survival analytics only selective columns are utilized for predicting survival analytics.The data set contains mostly binary values, which gives a narrow prediction accuracy than non-binary values.The nature of mortality risk rate in real life depends upon various conditions,diseases,illness,accidents,injuries and other medical conditions.Mortality risk rate depends upon various medically identified conditions and diseases on a human being through t the age or its life span.Each medical condition has its own personal and unique characteristics with different fatality rates.some are vary fatal and some are non fatal.

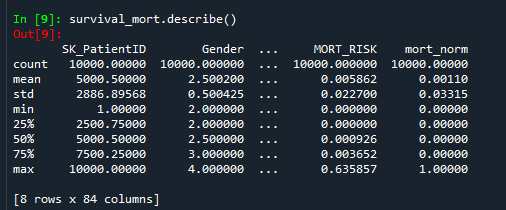


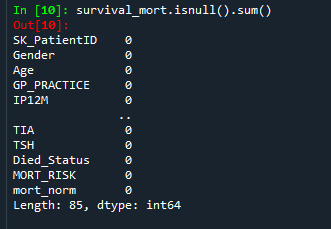












We have drawn meaningful statistics about the dataset given. After carefully analyzing the data set we can understand that due to presence of binary values there is a huge chance of getting low accuracy in prediction.comparison between any two columns merely stays between o and 1 which is not desired.The nature of categorical values is not so detailed and as a result we simply cannot compare the columns with the mortality risk rate which is in decimal form. We need to convert the mortality risk rate column to binary form or we need to convert the binary columns in to numeric or integer values.so due to binary values the comparison graphs,plots are very inaccurate and will give very small inferences about the relationship between different columns and mortality risk rate.Some columns are not necessary or not required for finding the relationship between them.The out put variable will be mortality risk rate where as age will be applied as time frame which will be mostly taken as x-axis on the plots and graphs.mortality risk rate was seen listed in data set in descending order from highest to lowest.so,overall I can see that due to majority of binary values present we cannot apply many models of regression,classification and other algorithms.we need further details and classification among the categorical values.we can still apply survival analytic s but we are limited in accuracy and error rate in prediction will be more due to insufficient data.if we perform binary

comparision there seems to be some non-meaningful inferences that can be derived which may be not accurate. There may be more methods in which we can convert the binary categorical values in to integer or numeric after which we can apply certain operations in python which can make the dataset applicable with different types of regressions and other algorithms required.So we have drawn graphs and plots with binary categorical values and have obtained different scatter-plots,histograms,bar plots,box-plots and other plots required.

Feature importance refers to techniques that assign a score to input features based on how useful they are at predicting a target variable.The role of feature importance in a predictive modeling problem.

The above report is derived from eda analysis and meaningful understanding on the data .so,I have derived inferences as well as insights.I suggest that only binary algorithms can be applied and feature variables can differ and moreover can vary.There may be more complex methods applied but due to the presence of more number of rows it can be stressful.In view of project requirements I submit this report as my own work required for the project work to be completed.These are my views,inferences and insights on the data set and the eda insights derived.