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

```
In [2]: survival_mort.dtypes
Out[2]:
SK_PatientID      int64
Gender            int64
Age               int64
GP_PRACTICE      object
IP12M             int64
                 ...
TIA               int64
TSH               int64
Died_Status       int64
MORT_RISK       float64
mort_norm         int64
Length: 85, dtype: object

In [3]:
```

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.

```
In [3]: survival_mort.count()
Out[3]:
SK_PatientID    10000
Gender          10000
Age             10000
GP_PRACTICE     10000
IP12M           10000
                ...
TIA             10000
TSH             10000
Died_Status     10000
MORT_RISK       10000
mort_norm       10000
Length: 85, dtype: int64
```

```
In [4]: survival_mort.info()
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 10000 entries, 0 to 9999
Data columns (total 85 columns):
 #   Column                           Non-Null Count  Dtype
---  ------                           --------------  -----
 0   SK_PatientID                     10000 non-null  int64
 1   Gender                           10000 non-null  int64
 2   Age                              10000 non-null  int64
 3   GP_PRACTICE                      10000 non-null  object
 4   IP12M                            10000 non-null  int64
 5   IPHIST                           10000 non-null  int64
 6   AE12M                            10000 non-null  int64
 7   OP12M                            10000 non-null  int64
 8   Waiting list - prioritised OP    1 non-null      float64
 9   Waiting list - unprioritised OP  0 non-null      float64
 10  Diabetic OP                      1 non-null      float64
 11  Cancer OP                        1 non-null      float64
 12  Social Community                 1 non-null      float64
 13  Readmission model                1 non-null      float64
 14  Readmission likelihood           1 non-null      float64
 15  Unnamed: 15                      1 non-null      float64
 16  Ace                              10000 non-null  int64
 17  AF                               10000 non-null  int64
 18  Alcohol                          10000 non-null  int64
 19  Antagonist                       10000 non-null  int64
 20  Anticoag                         10000 non-null  int64
 21  Asthma                           10000 non-null  int64
 22  B12                              10000 non-null  int64
 23  Beta                             10000 non-null  int64
 24  BloodTest                        10000 non-null  int64
 25  BMI                              10000 non-null  int64
 26  BoneSparing                      10000 non-null  int64
 27  BP                               10000 non-null  int64
 28  Breathlessness                   10000 non-null  int64
 29  Calcium                          10000 non-null  int64
 30  Cancer                           10000 non-null  int64
 31  CardioVasc                       10000 non-null  int64
 32  Cervical                         10000 non-null  int64
 33  CervixRemoval                    10000 non-null  int64
 34  CHD                              10000 non-null  int64
 35  Cholesterol                      10000 non-null  int64
 36  CKD                              10000 non-null  int64
```

```
 75  SerumLithium                   10000 non-null  int64
 76  Smoker                         10000 non-null  int64
 77  Statin                         10000 non-null  int64
 78  Stroke                         10000 non-null  int64
 79  Thyroid                        10000 non-null  int64
 80  TIA                            10000 non-null  int64
 81  TSH                            10000 non-null  int64
 82  Died_Status                    10000 non-null  int64
 83  MORT_RISK                      10000 non-null  float64
 84  mort_norm                      10000 non-null  int64
dtypes: float64(9), int64(75), object(1)
memory usage: 6.5+ MB
```

```
In [5]: survival_mort.value_counts()
Out[5]: Series([], dtype: int64)
```

```
In [8]: survival_mort.columns
Out[8]:
Index(['SK_PatientID', 'Gender', 'Age', 'GP_PRACTICE', 'IP12M', 'IPHIST',
       'AE12M', 'OP12M', 'Waiting list - prioritised OP',
       'Waiting list - unprioritised OP', 'Diabetic OP', 'Cancer OP',
       'Social Community', 'Readmission model', 'Readmission likelihood',
       'Unnamed: 15', 'Ace', 'AF', 'Alcohol', 'Antagonist', 'Anticoag',
       'Asthma', 'B12', 'Beta', 'BloodTest', 'BMI', 'BoneSparing', 'BP',
       'Breathlessness', 'Calcium', 'Cancer', 'CardioVasc', 'Cervical',
       'CervixRemoval', 'CHD', 'Cholesterol', 'CKD', 'Clopidogrel',
       'Contraception', 'COPD', 'Dementia', 'Depression', 'Diabetes',
       'Dipyridamole', 'DXA', 'Echocardiogram', 'Epilepsy', 'FEV1', 'FluVacc',
       'FolateTests', 'Foot', 'Fracture', 'HF', 'Hypertension', 'IFCC',
       'LiverTest', 'LVSD', 'MHealth', 'MicroAlb', 'MRI', 'Neuropathy',
       'Osteoporosis', 'OTCsalic', 'OxygenSat', 'PAD', 'Palliative', 'PEFR',
       'Pharmaco', 'Proteinuria', 'Renal', 'RheumArth', 'Salicylate',
       'SerumChol', 'SerumCreat', 'SerumFructo', 'SerumLithium', 'Smoker',
       'Statin', 'Stroke', 'Thyroid', 'TIA', 'TSH', 'Died_Status', 'MORT_RISK',
       'mort_norm'],
      dtype='object')
```

```
In [9]: survival_mort.describe()
Out[9]:
       SK_PatientID        Gender  ...      MORT_RISK    mort_norm
count   10000.00000  10000.000000  ...   10000.000000  10000.00000
mean     5000.50000      2.500200  ...       0.005862      0.00110
std      2886.89568      0.500425  ...       0.022700      0.03315
min         1.00000      2.000000  ...       0.000000      0.00000
25%      2500.75000      2.000000  ...       0.000000      0.00000
50%      5000.50000      2.500000  ...       0.000926      0.00000
75%      7500.25000      3.000000  ...       0.003652      0.00000
max     10000.00000      4.000000  ...       0.635857      1.00000

[8 rows x 84 columns]
```

```
In [10]: survival_mort.isnull().sum()
Out[10]:
SK_PatientID    0
Gender          0
Age             0
GP_PRACTICE     0
IP12M           0
               ..
TIA             0
TSH             0
Died_Status     0
MORT_RISK       0
mort_norm       0
Length: 85, dtype: int64
```

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.