MISSING VALUE ANALYSIS AND MODELLING OF CORONA DATASET

DATASET:

The data relates to characteristics of countries and states that relate to their projected Covid-19 death rate after 10 years if no vaccine is available. There are 14 variables including one Id variable, 2 discrete variable and 11 continuous variables.

PRE-PROCESSING STEPS:

The dataset needs to be cleaned and formatted before modelling the data, as it has missing values and unformatted datatype. The various pre-processing steps done int dataset are as follows

1.The missing values in the dataset are replaced with NA that include values such as -99,--, na, N/A.

2.The continuous variables are changed into double datatype as it is in character datatype.

3.The NA values in health care cost is set to 0 because it is 'Not Applicable' when the health care basis plan is FREE

4. Categorical NA values are not set to none because the given dataset has no 'Not Applicable' values in the categorical variables.

5. The variable "AGEMEDIAN" is removed from the dataset as it is not an important predictor(Obtained from random forest method) and has missing values more than 55%.

6.The observations 28,76,118,168 are removed from the dataset as it has missing values more than 50% in the record.

7. The dataset is divided into two data as train and split for prediction of death rate.

8 The recipe function is used to impute the values for missing data in the dataset. I have imputed all the variables in the dataset using KNN impute. This give 1 Id and 1 Outcome and remaining as predictors.

EDA VISULAISATIONS:

Missing Values:

After data processing the dataset is visualised to see the missing value distribution. The missing values are in the pattern of random. . The given dataset also show that the missing value is **MAR**. The missing values is predicted from variables in the dataset.

Miss Upset:

The miss upset plot shows that there are 2 observations in the dataset which has 6 missing values after pre-processing. Government variable has the minimum missing values in the dataset and Doc10 has the maximum missing values in the dataset after pre-processing.

Correlation:

The correlation plot shows that the GDP of the country is highly correlated with the Vaxrate. If the GDP of the country increases, then the Vaxrate also increases, respectively. The population density is highly correlated with the Deathrate. This shows the death rate increases in the country where the population is higher.

Histogram:

The histogram plot shows the outliers for each variable. Population, GDP, Vaxrate are the variables which has more outliers in the data. Age25prop, Age55prop, Infantmort, DOC10 are the variables which has least outliers in the data.

BoxPlot:

The boxplot shows all the outliers in the variable of the dataset. The boxplot helps to visualise the outlier more clearly than the histogram. Population, GDP, Vaxrate are the variables which has more outliers in the data. Age25prop, Age55prop, Infantmort, DOC10 are the variables which has least outliers in the data.

GG Pairs:

GG pairs plot shows the correlation between the pairs. I used all the continuous variables in the dataset for pairs plot. This pairs plot shows that the Vaxrate and Gdp2019 of the dataset are highly correlated with each other.

Scatter Plot:

The scatter plot shows the relation between the 2 variables. The scatter plot again proves that there is a strong relation between the vaxrate and GDP. Moreover, the scatter plot shows that the infant mortality is dependent with the doctor count.

Rpart Missing:

The rpart helps to predict the number of missing variables in the observation. The rpart plot shows that the country with population greater than or equal to 17 are more in the dataset with missing value of 0.65 and the country with less than 17 has value of 3.7

Random Forest:

The random forest helps to predict the important variables that helps to predict the death rate of the country. Government, Popdensity, Doc10 are the most important variable in the dataset that is required to predict the death rate of the country.

GLM MODEL:

Glmnet is a package that fits a generalized linear model via penalized maximum likelihood. The algorithm is extremely fast and can exploit sparsity in the input matrix . It fits linear, logistic, and multinomial, Poisson, and Cox regression models. A variety of predictions can be made from the fitted models. It can also fit multi-response linear regression.

The glm model is used to predict the death rate of the country. The glm model shows that there are 2 outliers in the prediction than the actual value. The MSE value is 12.58784 for our dataset.

The countries 78 and 61 are the outliers of the model. The given dataset also show that the missing value is **MAR**. The missing values is predicted from variables in the dataset.