**Abstract**

Multiple models of classification and regression were tested on two different datasets within this project to compare to determine which specific models was the most effective. Several different data pre-processing techniques were required to ensure the models would run properly e.g., label encoding or normalization. The findings from the tests suggest that the random forest model of classification was the most successful and the MLR model of regression worked best on their respective datasets.

**Introduction**

The project required us to test the KNN, Dtree and random forest models of classification on a brain stroke dataset and test SLR, MLR, Lasso & Ridge models of regression on airplane crashes and fatalities dataset. The most/least successful models would be determined by specific prediction scores within a classification report for example.

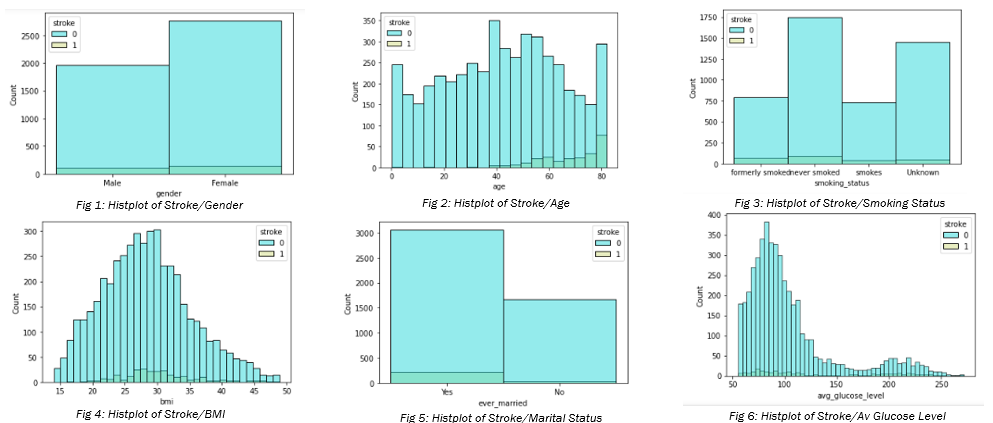
**Data Importing & EDA**

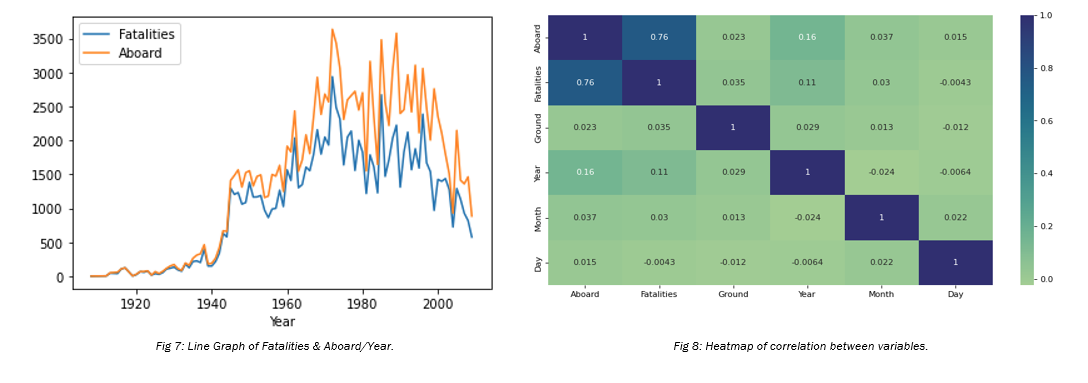
Once each dataset was imported exploratory data analysis (EDA) could begin. EDA includes functions like .head(), .info() to further understand aspects of the data e.g., no. of columns/rows, no. of null values or the data types in each column.The same functions were carried out on both datasets. The EDA process allows us to plan/decide what pre-processes need to be carried out before the data before it can be implemented into models.

**Data Cleansing**

Only the airplane crashes dataset required data cleansing, consisting of multiple processes as there were various issues with the data. Columns with thousands of nulls/deemed not relevant were dropped: ‘Flight #’, ‘Time’, ‘. ‘Date’, 'Route','Registration','Summary','cn/In','Type'column was split into. ‘Year’, ‘Month’, ‘Day’. The remaining nulls were dropped using dropna(inplace = True).

**Visualizations**





**Label/One Hot Encoding**

Label/one hot encoding was used on both datasets, meaning categorical data could be transformed into numerical data so the dataset could be normalized. ‘Gender’, ‘Ever married’, ‘Work type’, ‘Residence Type’ and ‘Smoking type’ from the brain stroke dataset were all encoded. Whereas for the airplane crash dataset only ‘Date’, ‘Location’ and ‘Operator’ were encoded.

**Normalization & Undersampling**

Both datasets were normalised using the MinMaxScaler(), converting all values within the respective dataset to a value between 0 and 1. For the brain stroke data only undersampling was required due to the disparity in class distribution within the dataset, preserving the shape of the original distribution. This improves the prediction/evaluation accuracy of models.

**Splitting Data**

Before creating an instance of any model, the data must be split into X and y variables. ‘y’ variable being the target variable. Data is then trained using train test split. Models are then fitted to the train data to produce prediction scores, some stored within a classification report or a confusion matrix for example.

**Classification Model 1 – KNN**

Once the data was trained and fitted to the KNN model; predictions could be made displayed within a classification report and confusion matrix seen in *fig 9.* The confusion matrix showed a total of 32 false predictions and 69 true predictions. Scores ranged between 0.65 and 0.76, with the majority being the mid to late 60s, with a F1/accuracy score of 68. The K- value used for this model was 10, however the optimal K-value displayed by *fig 10* is 7.

**Classification Model 2 – Dtree**

X and y variables were fitted and trained to the Dtree modelproducing confusion matrix showing a total of 28 false predictions and 72 true predictions. The F1 score for the model in this particular instance was 0.72, suggesting that the predictions for this model were reasonably accurate at 72%.

**Classification Model 3 – Random Forest**

The confusion matrix of the random forest model displayed only 17 false predictions in total and 83 true predictions. This model had pretty high prediction scores all round with an F1 score of 0.83, suggesting predictions for this model are 83% accurate.

**Classification Model Evaluation**

We can compare the confusion matrices and classification reports of the 3 models tested to rank them in terms of accuracy of prediction by focusing on the F1 score and accuracy. The best F1 accuracy score/accuracy score produced was by the random tree model (0.83). Therefore, we can say this was the best performing model. On the other hand, the KNN model was the worst performing with an F1/accuracy score of 0.72.

**Regression Model 1 – SLR**

When comparing the trained data with the test data in we can see that there are some similarities in patterns of the two scatter plots. Regression line is steeper in the test scatter, suggesting a greater rate of change. R2 (R-squared) value is 0.64, shows how well the data fit the regression model from 0-1. Low mean squared error score which suggests the regression line is not far from data points. (0.001).

**Regression Model 2 – MLR**

For the multiple linear regression model, the R2 (R-squared) value produced is 0.65, not optimal but suggests the data fit the model reasonably well. The low mean squared error score which suggests the regression line is not far from data points (0.001).

**Regression Model 3 – Lasso**

The lasso regression model generated a R2 score equal to 0.39 which is quite poor, to add to this the mean square error is 643 meaning the regression line is far from the data points. Essentially, the model did not perform well.

**Regression Model 4 – Ridge**

The low mean squared error score which suggests the regression line is not far from data points (0.001). The R2 value produced was around 0.64, which is a similar value to the other models bar Lasso.

**Regression Model Evaluation**

When evaluating regression models by prediction scores the two most important values are RMSE (root mean squared error) and R2 (R-squared). As we can see from the figures below, the model with both the lowest RMSE and highest R2 value was the MLR model. Therefore, this was the best performing regression model, whilst the worst performing model was Lasso as it produced the lowest RMSE and highest R2 values by a significant margin.

**Conclusion**

For the respective datasets used we can confidently define the best performing models of classification and regression. For Regression the MLR model was the best performing model, producing the best prediction scores whilst Lasso was the worst performing according to R2 and RMSE values. In terms of Classification models the random tree model performed the best by providing the highest number of accurate predictions, displayed In classification reports whilst KNN performed the worst.