Cardiovascular diseases (CVD’s) are a major health concern contributing to millions of lives lost each year. Risk factors include alcohol misuse, smoking cigarettes, obesity, lack of physical activity, diabetes, and hypertension. The aims of this research are to investigate whether death from CVD’s can be predicted by a set of features known to be linked to CVD such as diabetes, tobacco use and high blood pressure. It is hypothesised that smoking, diabetes, high blood pressure and anaemia (decrease of red blood cells or haemoglobin) will predict whether the patient survives.

**Method**

Participants

There was 194 males and 105 females while the average age was 60 years and ranged from 40 to 95 years.

Procedure

The dataset was retrieved from Kaggle and Python 3 using Jupyter notebook was used to explore the data. Pandas, Numpy, matplotlib and sklearn were the main packages used.

Exploratory data analysis

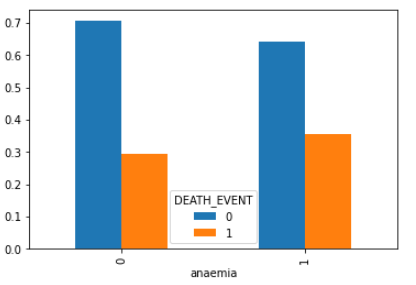
All features considered to be important for predicting death from CVD were first visualised with bar charts. For example, a bar chart depicting high blood pressure and death between groups who had high blood pressure and did not while also exploring the percentage of people who died between the groups.

Machine learning model

A logistic regression was used to classify whether someone survives based on four Boolean features including high blood pressure, diabetes, smoking and anaemia. Considering the data was imbalanced in favour of individuals surviving (203) compared to death (96), metrics other than accuracy were also explored such as precision and recall. Different thresholds for the logistic regression were also explored to investigate the best trade-off between specificity and sensitivity depicted by a ROC curve. A second model only including high blood pressure and anaemia features was also analysed to compare the AUC score to the first model.

5 k-fold cross validation was then used to further evaluate multiple models including one model with all features excluding time, a second model with the feature’s anaemia, smoking, high blood pressure and diabetes. The last model only consisted of two features, anaemia, and high blood pressure.

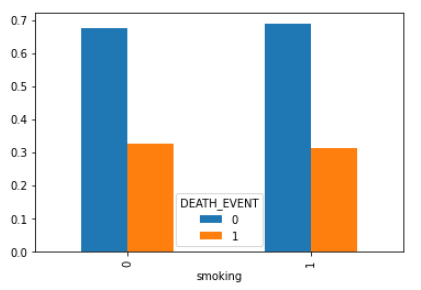
Results



*Figure 1.* A side by side comparison of the proportion of deaths between

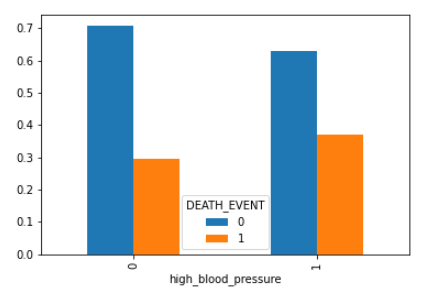
group 1 who did not have anaemia and group 2 who had anaemia.

As can be seen from figure 1, there was a slight increase of deaths when people had anaemia.



*Figure 2.* A side by side comparison of the proportion of deaths between

group 1 who did not smoke tobacco and group 2 who smoked tobacco.

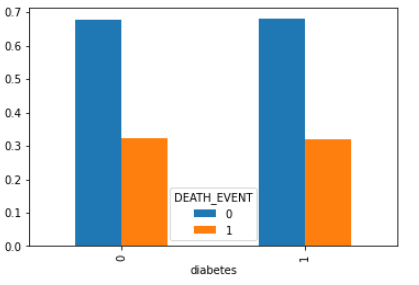
As can be seen from figure 2, contrary to intuition, the smoking group had slightly less deaths.

*Figure 3*. Side by side comparison of the proportion of deaths between

group 1 who did not have high blood pressure and group 2 who had high

blood pressure.

As can be seen from figure 3, people who had high blood pressure resulted in slightly more deaths.

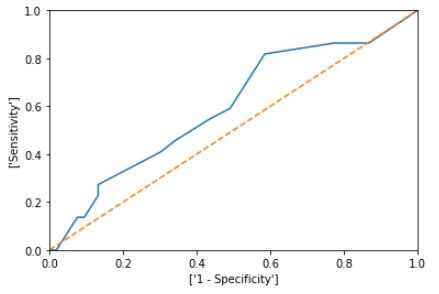


*Figure 4.* Side by side comparison of the proportion of deaths between

group 1 who did not have diabetes and group 2 who had diabetes.

As can be seen from figure 4, having diabetes didn’t seem to be a factor responsible for death since similar amount of people died between both groups.

Classification accuracy was 73% but considering the classes (survive, died) were imbalanced in favour of surviving, this accuracy score can be misleading and should be interpreted with caution. Precision was 0.73 for the survival class and 0 for the death class suggesting the logistic regression algorithm did not predict any correct for the death class (true negatives). In terms of recall, survival class was 1 indicating all positive cases were identified correctly while the other class was 0 suggesting none were recalled. No precision and recall for the minority class is explained by the algorithm only ever predicting person survived and never that the person died.



*Figure 5.* ROC curve depicting different thresholds for the logistic

regression algorithm and showing specificity on x axis and sensitivity on

the y axis.

As can be seen from figure 5, sensitivity (true positive rate) and specificity (false positive rate) are depicted with various thresholds from the logistic regression classifier. In terms of the graph, when sensitivity is approximately 0.80 and specificity is 0.40 seems to be the best trade-off to manage the true/false positive rate. To get an empirical measure of the ROC curve, area under the curve (AUC) was calculated. The AUC score was 0.60 for the model and a second model was created but with only anaemia and high blood pressure features. the second model also scored 0.60 indicating the addition of diabetes and smoking tobacco does not explain any more information in model 1.

*Table 1.* 5 K-fold cross validation between 3 models and average taken over the 5 folds for accuracy, precision, recall and f1 score

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Model | accuracy | precision | recall | F1 score |
| All features | 0.73 | 0.60 | 0.27 | 0.37 |
| 4 features | 0.68 | 0 | 0 | 0 |
| 2 features | 0.68 | 0 | 0 | 0 |

As can be seen from table 1, using all features resulted in the highest classification

accuracy and the highest precision, recall and f1

score. The other models were not far from the classification accuracy

compared to all features model but had 0 precision, recall and f1 score suggesting none of the positive predictions were true and all positive values were never predicted.

**Discussion**

The hypothesis that death from CVD can be predicted by diabetes, high blood pressure, anaemia and smoking was not supported. Even though classification accuracy was 73%, the data was imbalanced, and this metric was not valid. Further inspection confirmed that the algorithm just predicted survive for every data point suggesting the logistic regression classifier did not predict any of the actual deaths correctly. Moreover, it was evident other metrics and thresholds were required to be evaluated from the logistic regression classifier to detect and predict the minority class. An AUC score was generated from a ROC curve and the model scored 0.60 suggesting the model was not much better than flipping a coin. A second model was also generated with two less features and score 0.60 as well. After considering 5 fold cross validation with three different models, the model with features had 0.74 accuracy, precision score of 0.68 and recall of 0.32 and f1 score of 0. indicating this model did an okay job considering the other models had 0 for precision, recall and f1 score. Further work should attempt to use different machine learning classifiers to test whether they can improve classification performance on predicting CVD. Perhaps more features need to be included in a machine learning model to be able to better predict CVD.

In conclusion, this research has shown diabetes, smoking tabaco, high blood pressure and anaemia are not strong predictors of death from CVD. Using all features did help classification accuracy slightly and moderately helped with precision, recall and f1 score.