**Effect of Various Factors on Patients Heart Disease**

Heart disease refers to all diseases of the heart. It is responsible for around 66,000 deaths in the UK each year (BHF, 2022). This report analyses heart disease presence or absence as a function of the following: Patient\_ID (PID), HeartDisease (HD), BMI, Smoking (SM), AlcoholDrinking (AD), Stroke (ST), PhysicalHealth (PH), MentalHealth (MH), DiffWalking (DW), Sex (S), AgeCategory (AC), Race (R), Diabetic (D), PhysicalActivity (PA), GenHealth (GH), SleepTime (SLT), Asthma (A), KidneyDisease (KD), SkinCancer (SC), Dead/Alive (DA), and 'date'. The analysis investigates:

1. how the features contribute to the presence of heart failure
2. a model that best predicts heart failure from the features
3. how features of patients with heart conditions contributes to death

**Methods used:**

1. Python programming: sklearn, tensorflow, pandas, numpy, dython libraries were used.
2. From the correlation matrix, it is obvious that there exists multicollinearity (**Figure C**) in the data and thus there is need for regularization in logistic regression.
3. Downsampling technique was used to treat the data imbalance.
4. Models trained are random forest, Deep Neural Network (DNN), logistic regression, k-Nearest neighbour (kNN)
5. The analysis can be found at this url: **https://github.com/Enzyme-Muiz/heart\_disease\_analysis**

**Findings:**

1. All features statistically correlate significantly with HeartDisease except Dead/Alive (**Figure E**).
2. Having an history of stroke before is the feature with the highest risk of heart failure while being female reduces the risk (**Figure B**).
3. Random Forest (**Figure D**), KNN, DNN and logistic regression that attained accuracy of average 0.78 were obtained. DNN attained the highest accuracy of 80% on test data for predicting heartdisease from the variables.
4. The following is the logistic regression model for the HeartDisease:
5. Some recommendations of smoking cessation and diffWalking can be made to patients to reduce development of heart disease.
6. Given that a patient has heart disease, kidney disease is the feature that is statistically important in prognosing death.
7. A Random forest model performed better than logistic regression in determining the chance of death of a patient that has had heart disease with 72% accuracy vs 50% accuracy.

**FIGURES**

Graphical user interface

Description automatically generatedChart, bar chart

Description automatically generated

**B**

**A**

Graphical user interface, chart

Description automatically generatedShape

Description automatically generated

**D**

**C**

Graphical user interface

Description automatically generated

**E**

**REFERENCE**

1. https://www.bhf.org.uk/-/media/files/research/heart-statistics/bhf-cvd-statistics---uk-factsheet.pdf