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

Heart disease is an umbrella for all diseases of the heart. It is responsible for around 66,000 deaths in the UK each year (BHF, 2022). This report involves the analysis of 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 is carried out to investigate:

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:**

1. Python programming: sklearn, tensorflow, pandas, numpy, dython
2. From the correlation matrix, it is obvious that there exists multicollinearity in the data and thus there is need for regularization in logistic regression.
3. Data was cleaned with **'Goo\_d', 'goo\_d'** and **'ExCellent'** from GH column changed to **Good, Good** and **Excellent**
4. Downsampling technique was used to treat the data imbalance.
5. Models trained are random forest,

**Findings:**

1. All features statistically correlate significantly with HeartDisease except Dead/Alive
2. Having had stroke before is the feature with the highest risk of heart failure.
3. Random Forest, KNN, Deep Neural Network and logistic regression that attained Mathew Correlation Coefficient (MCC) of average 0.55 were obtained.
4. The following is the logistic regression model for the HeartDisease:
5. Being a female reduces the chance of having heart failure by almost half.
6. Some recommendations of increased sleeptime, smoking cessation and diffWalking can be made to patients to reduce development of heart disease.
7. Given that a patient has heart disease, kidney disease is the feature that is statistically important in deciding if the patient dies.
8. 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.

**Appendix**

Graphical user interface, chart

Description automatically generatedShape

Description automatically generated

Graphical user interface

Description automatically generated

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