HEART ATTACK PREDICTOR

By: Tharindu Yakkala

SUMMARY

Heart attacks are of major concern for many as you age. Eating healthy and exercise sometimes helps reduce this risk.

The objective is the create a machine learning model for classification that will predict if a person has a higher or lower change of having a heart attack.

After trying multiple models it was found that a custom neural network was able to get 92% training accuracy and 92% testing accuracy as the top model when using all variables to predict without much preprocessing.

VARIABLES

- Age
- Sex – Male/Female
- Cp: Chest Pain type
- Trtbps: resting blood pressure (in mm Hg)
- Chol: cholesterol in mg/dl.
- Resecg: resting electrocardiagraphic result
- Thalach: max heart rate achieved
- Exng: exercise induces angina

VARIABLES

Oldpeak: previous peak, ST Depression induced bu exersice relative to rest.

Slp: Slope of ST Segment

Caa: number of major vessesl colored by flourosopy

Thall: thalium stress test result

output 0 less change of heart attack, 1 greater chance of heart attack

More detailed descriptions and data encoding can be found in the jupyter notebook intro.

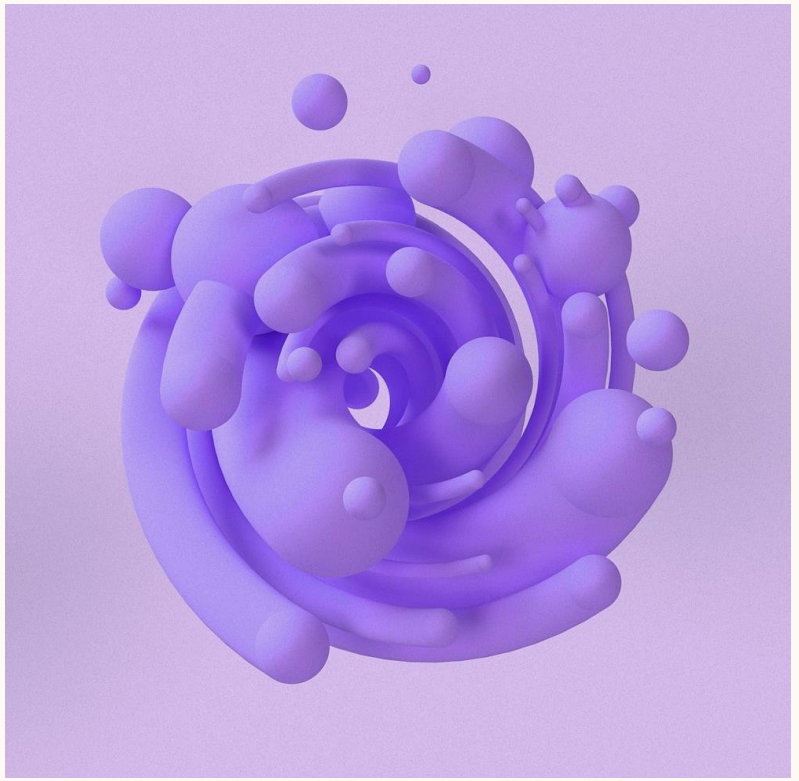
METHODOLOGY: DATA PREP

- No missing data was found, and for some features outliers were either removed or winsorized.
- Winsorize – replacing outliers with nearest min/max boundaring of mean +- 1.5 * IQR
- Other method was using square root to reduce skewness on thalach variable.

METHODOLOGY: MODELS

PYTHON MODELS USED

- Logistic Regression
- Decision Trees
- Support Vector Machines
- Random Forest
- XGBoost
- Neural Net (tensorflow)



METHODOLOGY: MODEL TRAINING

Two versions of the dataset was used.

- (1) Some variables removed, with some winsorizing, transformations, and scaling using RobustScaler. This set will be referred as Training/Test Set 1.
- (2) All variables used with only scaling using Standard Scaler. This set will be referred to Training/Test Set2.

TRAIN/TEST SET 1

Some variables removed, outliers adjusted, and transformed.

Test Set 1

Model	Train accuracy	Test Accuracy
Logistic Regression	88%	84%
Decision Tree	100%	77%
SVC	92%	79%
Random Forest	100%	72%
XGBoost	54%	56%
Neural Net	86%	86%

Test Set 2

Model	Train accuracy	Test Accuracy
Logistic Regression	88%	87%
Decision Tree	85%	84%
SVC	92%	82%
Random Forest	86%	88%
XGBoost	89%	76%
Neural Net	92%	92%

RESULTS SUMMARY

- In both test set 1 and 2, neural networks performed the best with no variance.
- Solution is to use all variables and just use standard scaler.
- After 7 different neural net models, the 7th one able to get 92% train, 92% test accuracy.
- Neural Net Architecture:
  - Dense(128, activation = 'relu') with batchnorm, L2 regularization + Dropout
  - Dense(128, activation='relu') with batchnorm, L2 regularization + Dropout
  - Dense(1, activation='sigmoid')