UNIVERSITAT POLITÈCNICA DE CATALUNYA

FINAL PROJECT REPORT

Machine Learning

Exploratory Data Analysis and Predictions of Heart Disease with UCI Heart Disease Data Set

Broniewski, Adam

Jahangir, Khushnur Binte

Supervisor

Coma-Puig, Bernat

Universitat Politècnica de Catalunya

June 2022

Table of Contents

1 Introduction

The goal of this project is to explore the UCI - Heart Disease Data Set to discover trends in the data and predict whether a patient would have heart disease based on medical attributes.

The dataset has 303 instances and 14 attributes that are a combination of categorical and real values, which provides room for experimentation with different models and approaches to data pre-processing. There are also 61 other relevant papers that make use of this dataset as identified by UCI website.

2 Related Previous Work

3 Exploratory Data Analysis

3.1 Pre-processing

3.1.1 Reformatting

The dataset column names were renamed from the originally abbreviated version to a full naming to make it more understandable during the exploratory phase.

3.1.2 Encoding

8 of the attributes are categorical and encoded with integer values. All the categorical variables use label encoding. Label encoding was maintained for some of these categories[1] as a review of the dataset indicated an order of severity. The remaining categorical variables[2] were treated with one-hot encoding as there is no relationship or order between each value. The target category of heart disease had increasing severities of sickness and was simplified and encoded as either healthy or sick.

3.1.3 Test and Train Data Set Split

The data set was split at the early stage of pre-processing to eliminate the risk of data leakage. All transformations to the training dataset will be applied to the test dataset independently. To test/train split is stratified to preserve that same proportion of target category examples as observed in the original data set.

It's important to note that the only step of the process that could introduce data leakage is during scaling before the model is trained. This is mitigated with the use of a standard scaler made from the training data set.

3.1.4 Missing Values

The data set is quite clean, without missing values. 6 values in major vessel count and thalassemia were found with a "?". Given the small number of missing values, the rows were dropped.

[1] Label Encoding: chest pain type, resting electrocardiographic, peak exercise ST slope, and thalassemia

[2] One-hot Encoding: sex, fasting blood sugar, exercise induced angina

3.1.5 Outliers

To detect outliers, data that was outside of 150% of the interquartile range (1.5*IQR) was identified. Based on analysis, there were very few values that fell outside of the 1.5*IQR. Given box plot visualizations in Figure 3-1, only the cholesterol value above 450 was deemed to be noise and was removed from the dataset.

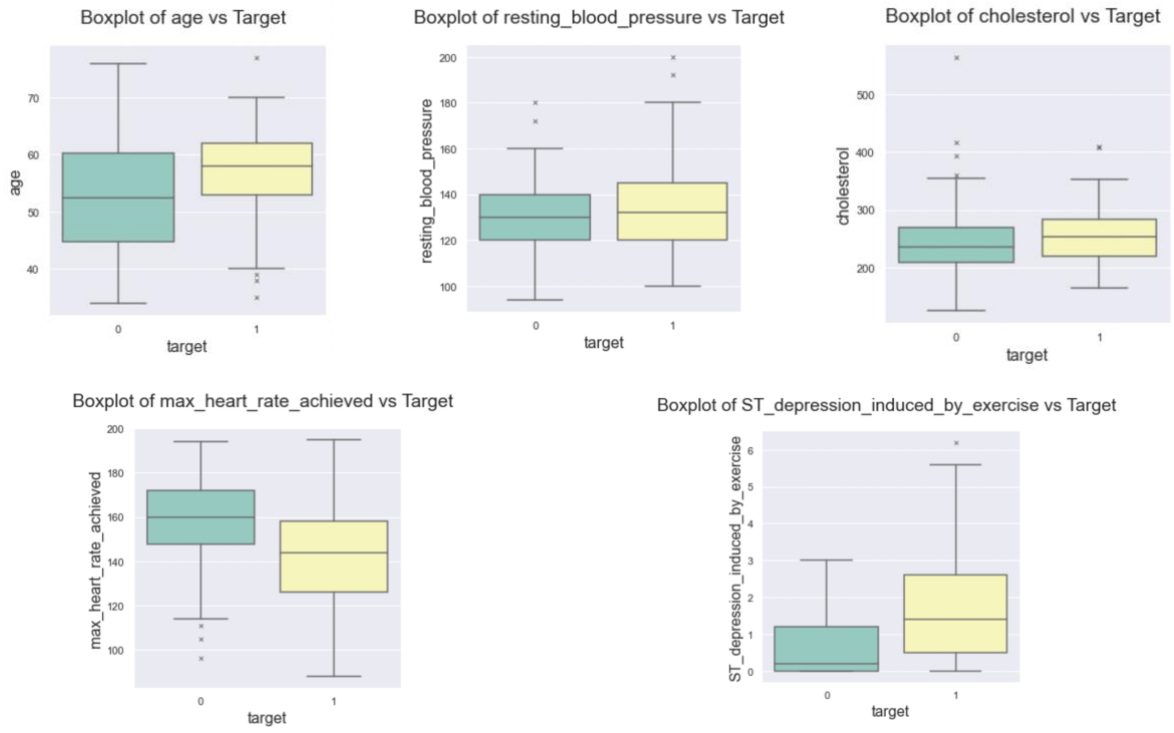

Figure 3-1 Outlier detection and exploration of numerical data

3.2 Visualizations

Visualizations are used to see if there are interesting trends or hypothesis that can be tested. This also provides insight into which models should be tested and which attributes are expected to have the most impact on the prediction.

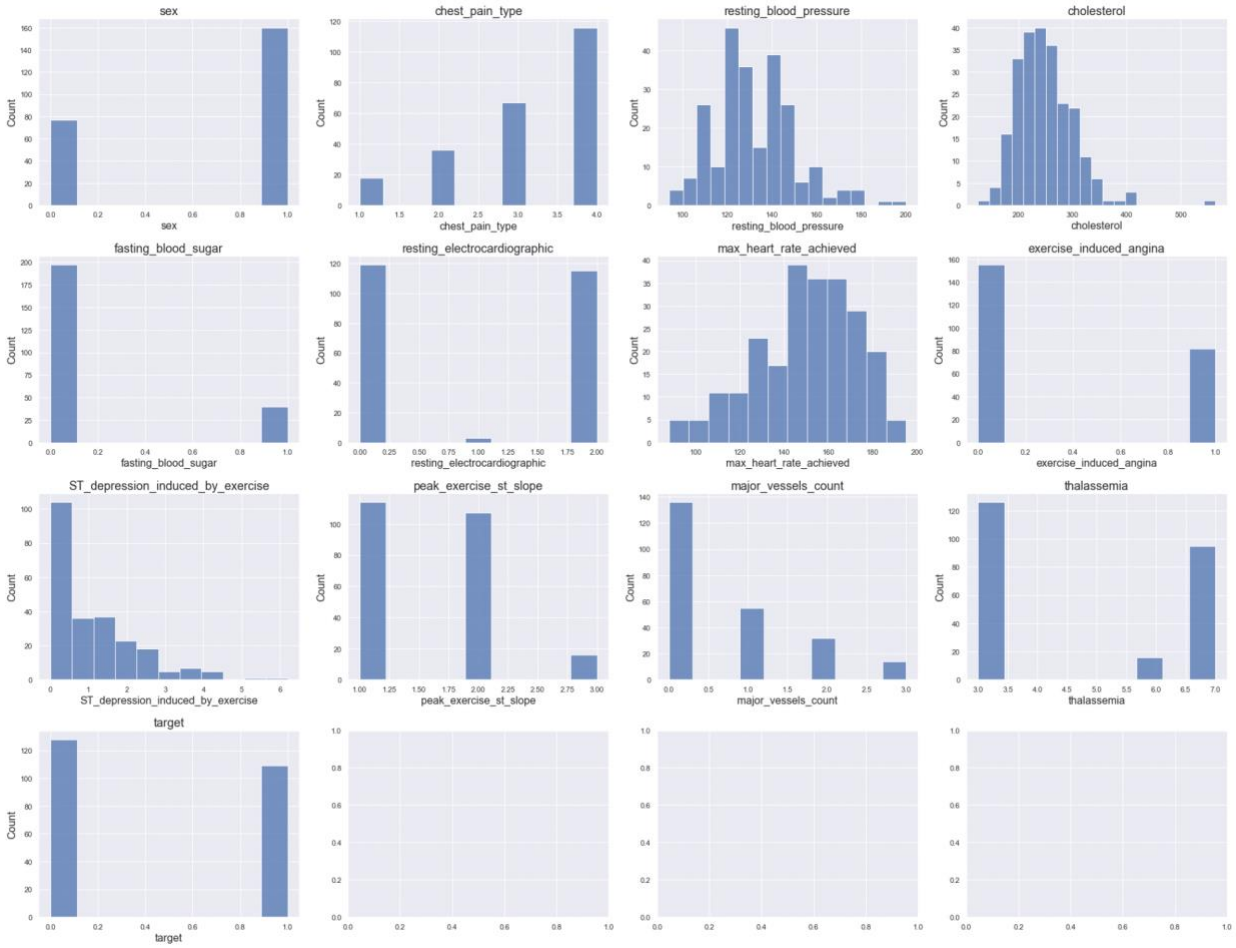

Figure 3-2 Boxplots of each attribute

The boxplots in Figure 3-2, show that the data is in different ranges and will need to be normalized before training models. It also clearly shows the categorical variables that were label encoded. Resting blood pressure appears right skewed, and max heart rate achieved appears left skewed. Although this skew would not impact tree-based models, we will apply a square root and square on the right and left skewed data respectively for our other models.

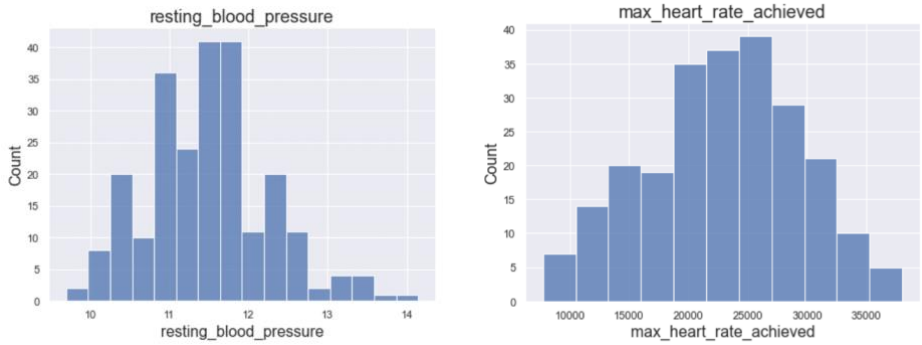

Figure 3-3 Data adjusted for skewness

Figure 3-4 below shows there are lots of individuals that had no ST depression induced by exercise (value of 0). We can explore to see if there is an issue in the data, and see the impact this attribute has on the target. This may be a good attribute to use for clustering individuals for a final predictive model.

Figure 3-4 Pairwise scatterplots and boxplots

3.3 Clustering

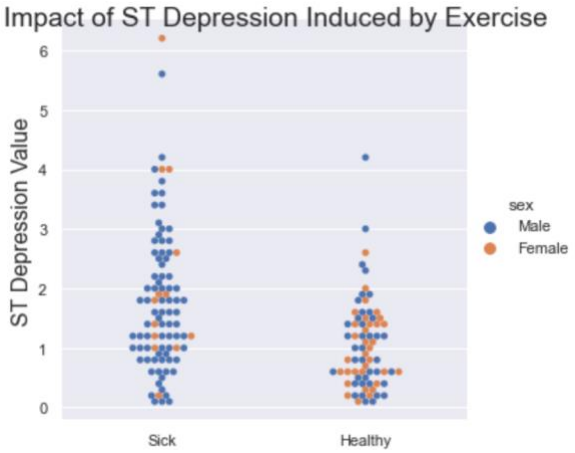

Figure 3-5 ST depression induced by exercise impact on heart disease

ST depression induced by exercise is a measure of change in an ECG reading after exercise [1]. Subjects with a "0" are subjects that did not have any ST depression, where subjects with a non-zero value had some ST depression present. Thus, the "0" values will not be removed and will be considered correct. It could be interesting to come back and bin the data into a "yes/no" categorical for use in model building.

Data also generally shows that males are more likely to have heart disease, however the data set is too small and there are much more instances in the data of males than females.

3.4 Feature Selection

Initial data exploration comparing numerical data against sex and the target category showed potential trends summarized in Figure 3-6 below:

- age seems correlated with heart disease, especially with men
- higher max heart rate is associated with individuals without heart diseases
- cholesterol doesn't have a significant impact on heart disease
- when a participant had ST depression that is seen after exercise, they were more likely to suffer from heart disease

Figure 3-7 below shows a correlation matrix for all attributes. We can see from the matrix that thalassemia looks to be well correlated with the target, and ST depression is correlated with peak exercise at ST slope. The cut-off for correlation in academic literature tends to range between 0.6 - 0.9. Based on this, there is not a strong case to remove any of the features.

For model testing, we will compare datasets where fasting blood sugar, cholesterol and resting blood pressure are removed to compare performance. These attributes do not show strong direct correlation to the target and may be introducing some noise in the predictions.

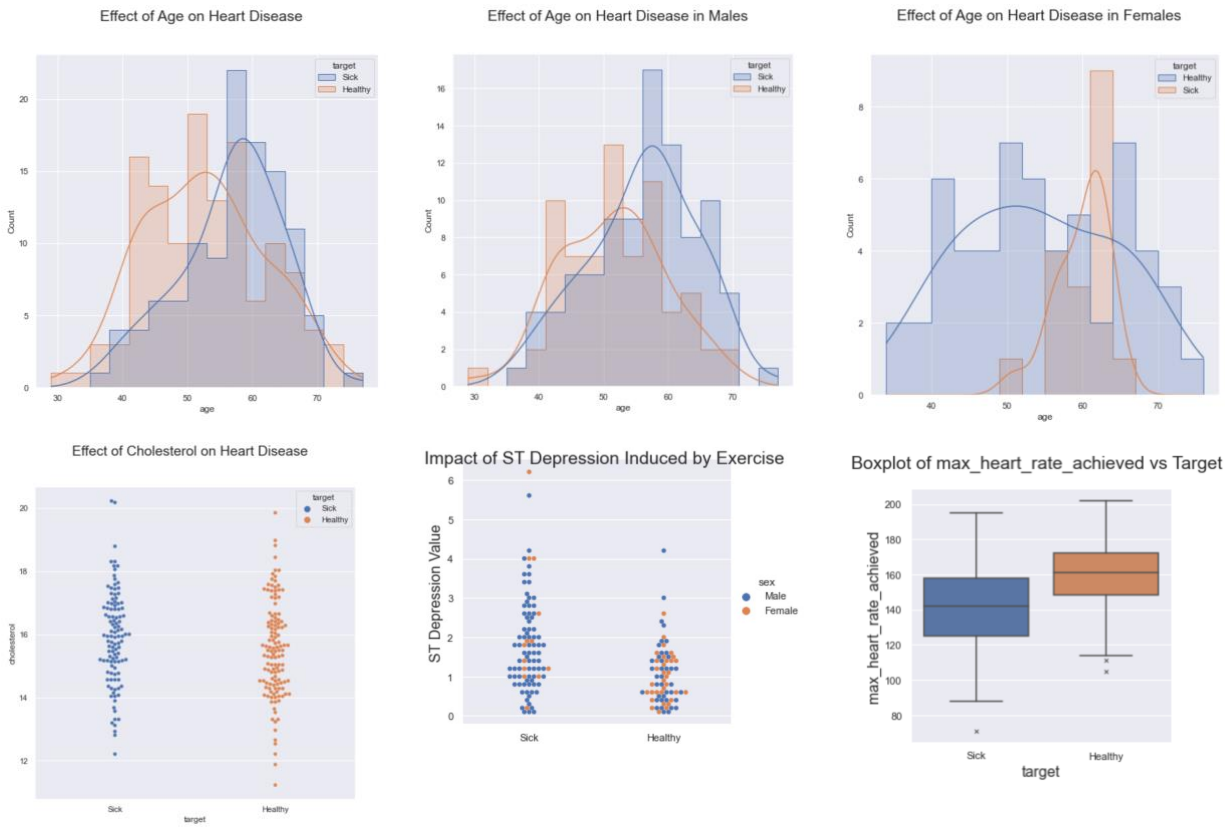

Figure 3-6 Visualizations showing potential feature selection candidates

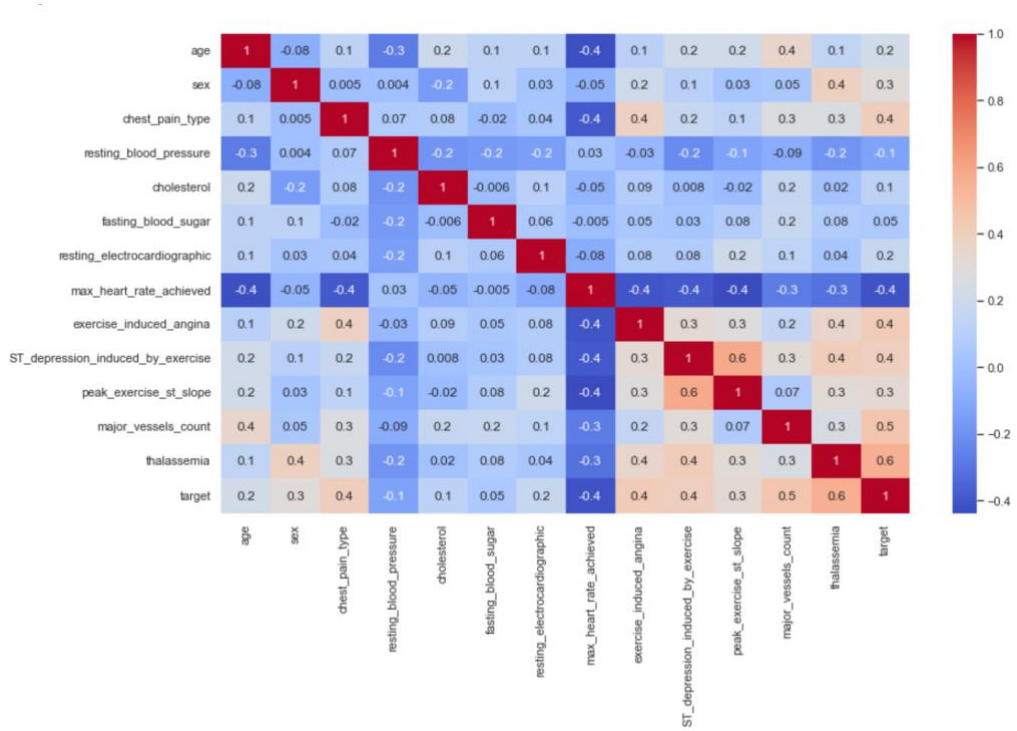

Figure 3-7 Heatmap of attribute correlation

4 Predictive Modeling

Functions are used to implement data cleaning and pre-processing identified during the EDA. Feature scaling will be completed by creating a transformation based on the training dataset for any train/test splits. This will ensure there is no data leakage, and that the way training and test data sets are treated is consistent.

4.1 Modeling Methods Considered

The models tested include:

- Logistic Regression
- Decision Tree
- Random Forest
- Neural Network

4.1.1 Parameter Choice

Each model is tested with various parameters. The models are built first using defaults, then with cross fold validation, and then applying hyper tuning. Models that require more computation time for hyper tuning (like random forest and neural net) are first tuned using a random search to approximate the correct parameters, followed by a full grid search to confirm them.

4.1.2 Feature Selection

Top scoring models that went through hyper parameter tuning were also run against a dataset that had feature selection applied, where three attributes were removed:

- Fasting blood-sugar
- Cholesterol
- Resting blood-pressure

These attributes were removed due to low correlation with the target as seen during the exploratory data analysis.

4.2 Validation Protocol

Validation of each model was completed comparing train/test split and k-Fold cross-validation. Cross-validation was found to result in greater accuracy on all models and was chosen for all validations.

For logistic regression and decision tree models, a 10-fold cross-validation was used as it provided a good balance between computation time and accuracy. Given the relatively small size of the training data set (242 instances), a smaller test instance was not deemed reasonable as it would result in many similar iterations training the model. A 5-fold cross validation was used for random forest and neural network, driven primarily by computation time.

4.3 Results

Figure 4-2 shows the results of the validation of different models. Each of the metrics was evaluated and considered when determining which model was the best choice [2]. Given the potential impact of undiagnosed heart disease, specificity is an important

metric, as we do not want to incorrectly classify an individual that has heart disease. Accuracy was deemed a good balance between true positives and true negatives.

The model chosen for use scored highest in all metrics considered: f-score, precision, recall and accuracy.

4.3.1 Neural Network

The neural network looked to perform well. Most surprisingly, the simple neural network with a single layer and single node operating on the full dataset did extremely well. The hyper tuned and simple neural network using the dataset with feature selection also ranked very well in comparison. Although neural network was performing well, the issue with using it as the final model introduced some issues in this field as the model would not be easily explainable and understandable by subject matter experts.

The tuned neural network model used the following parameters:

- Features: All
- Alpha: 4
- Number of layers: 3
- Hidden layer size: 8
- Activation: logistic

4.3.2 Random Forest

The hyper-tuned random forest model has a depth of 1 with 600 estimators. This is an unexpected result, but it means that each of the estimators is using a different attribute to predict the target.

The best performing random forest used the following parameters:

- Features: All
- Max depth: 3
- Estimators: 600

The feature importance of this model is shown in Figure 4-1. Although not used in model training, these results could be used in feature selection with additional tuning.

	coefficients
age	0.026419
chest_pain_type	0.150468
resting blood pressure	0.010700
cholesterol	0.012176
resting_electrocardiographic	0.004026
max_heart_rate_achieved	0.087889
ST_depression_induced_by_exercise	0.084614
peak_exercise_st_slope	0.040006
major_vessels_count	0.147505
thalassemia	0.241163
male	0.019124
female	0.024350
with fasting blood sugar	0.000223
without fasting blood sugar	0.000271
with_exercise_induced_angina	0.077153
without_exercise_induced_angina	0.073913

Figure 4-1 Feature importance from Random Forest

4.3.3 Logistic Regression

The next best model was the logistic regression model with hyper tuned parameters trained on the dataset with feature selection applied. Since this model can be more easily understood and explained by professionals with semantic knowledge, this was the model that was chosen for final deployment.

The best performing logistic regression model used the following parameters:

- Features: Selected
- C (regularization): 0.1
- Max iterations: 20
- Penalty: l1
- Solver: liblinear

	F-score (cv)	Precision (cv)	Recall (cv)	Accuracy (cv)
MLP (Hyper Parameter Tuning)	0.830799	0.838518	0.829462	0.833673
MLP [1]	0.83049	0.840838	0.829462	0.833673
FS - Logistic Regression (HT)	0.826161	0.836052	0.826011	0.830333
FS - MLP[1]	0.823329	0.831223	0.822339	0.826276
FS - MLP (HT)	0.820369	0.826934	0.820462	0.822279
Logistic Regression (CV)	0.819811	0.828857	0.821744	0.8255
Logistic Regression (Hyper Parameter Tuning)	0.819811	0.828857	0.821744	0.8255
Random Forest (Default)	0.815094	0.818182	0.782609	0.816327
Random Forest (CV)	0.813403	0.828682	0.812177	0.817092
Decision Tree (Hyper Parameter Tuning)	0.796903	0.809892	0.797698	0.801
Random Forest (Hyper Parameter Tuning)	0.795247	0.809506	0.797104	0.805
MLP [2,2,2]-alpha=0.001	0.793905	0.803593	0.794345	0.796259
Decision Tree (Default)	0.775136	0.75	0.782609	0.77551
Random Forest (Randomized Parameters)	0.766414	0.782268	0.773557	0.776
Logistic Regression (Default)	0.754181	0.73913	0.73913	0.755102
Decision Tree (CV)	0.707692	0.717964	0.708021	0.718

Figure 4-2 Model training results

4.4 Final Model Estimation and Performance

The final model was run using Logistic Regression with feature selection applied on the dataset. The F-score and accuracy metrics of the final model were very similar to the training accuracy. The precision of the final test was lower than the training model (78.1% vs 82.6%). The recall of the final model was higher than the training model (89.3% vs 82.6%). This means our model may have been a bit overfit for the data. The F-score does not reflect this change as it is taking an average of the other metrics.

	F-score (cv)	Precision (cv)	Recall (cv)	Accuracy (cv)
Final Model Test (LR)	0.836022	0.78125	0.892857	0.836066
FS - Logistic Regression (HT)	0.826161	0.836052	0.826011	0.830333

Figure 4-3 Final model performance

Looking at the coefficients of the model (Figure 4-4), the major vessels count is the most important indicator of heart disease (coefficient=1.03). This makes sense with the semantic meaning of the data, as the attribute is a measure of the number of unobstructed blood vessels in the heart. Thalassemia (coefficient=0.77), chest pain type (coefficient=0.50), and sex (coefficient=0.26) were the next most important factors in predicting heart disease.

Using the linear regression model makes interpretation and explanation simple and can be used to help diagnosing physicians tune in and look at certain factors in their patients with more weight.

	coefficients
age	-0.001929
chest_pain_type	0.502451
resting_electrocardiographic	0.280766
max_heart_rate_achieved	-0.253334
ST_depression_induced_by_exercise	0.229030
peak_exercise_st_slope	0.265598
major_vessels_count	1.029070
thalassemia	0.771282
male	-0.259844
female	0.259844
with_exercise_induced_angina	-0.182500
without_exercise_induced_angina	0.182500

Figure 4-4 Feature importance used in linear regression model

The feature importance in this linear regression model is generally similar to the random forest feature importance, with some slight differences. It is not surprising that there is some difference as the models and parameters are different.

5 Conclusions

Exploratory data analysis and predictive models were built for a heart disease dataset with 13 attributes and 303 instances. The data was generally clean (6 missing values) and need some adjustment to account for data skewness on continuous numerical values. There were trends seen in the impact of age, ST depression and max heart rate achieved during exercise on and individuals' likelihood to have heart disease.

Decision tree, logistic regression, random forest, and neural networks were trained and tested with a variety of parameters. Additionally, high performing models were tested using a dataset with feature selection performed. The best performing model was a hyper tuned neural network with 8 hidden layers of size 3, however a logistic regression model with feature selection was chosen for the final implementation. The model had a precision of 78.1%, recall of 89.3% and accuracy of 83.6%.

Based on the trained model, the count of unobstructed major blood vessels is the most significant factor in predicting heart disease.

6 Extensions and Limitations

A major limitation of this work was the limited size of the data set. Models were trained and tuned on ~250 instances, which increased the risk of over fitting or bias. Additionally, there was a significant difference in the amount of data from males and females. This could have been a result of how the study was conducted, or potentially a

natural representation of the fraction of the population suffering from heart disease. In either case, the model for females would be improved if there was more data for them.

Given the improvements of models with feature selection, as well as the potential improvements that could be gained in neural networks with fewer parameters, a future focus would be to continue tuning feature selection to improve model performance.

Applying a programmatic manner for feature extraction would also be a good next focus. Feature selection can be done making use of the important features from the random forest model, or using another method to identify the most important attributes and create a subset of data with those selected features.

Another natural extension of this work would be to apply an ensemble method to combine the performance of multiple models. This could allow the use of both logistic regression and neural networks to provide a model that can be generally understood and explained by experts using the coefficient weights, with additional tuning and boosting coming from a neural network to “tweak” the model.

7 Works Cited

[1] P. M. Okin, R. B. Devereux, J. A. Kors, G. v. Herpen, R. S.Crow, R. R. Fabsitz and B. V. Howard, "Computerized ST Depression Analysis Improves Prediction of All-Cause and Cardiovascular Mortality: The Strong Heart Study," *Annals of Noninvasive Electrocardiology,* p. 107–116, 2001.

[2] S. A. Hicks, I. Strümke, V. Thambawita, M. Hammou, M. A. Riegler, P. Halvorsen and S. Parasa, "On evaluation metrics for medical applications of artificial intelligence," *medRxiv,* 2021.