Heart Disease - Models

Mohsin Mohammed (100744769)

Rhichard Koh (100842848)

Faculty of Science, Engineering & Information Technology, Durham College

Machine Learning - Classification (COSC - 22000)

Professor Noopa Jagadeesh

April 18, 2023

Definition

Project Overview

Any disease that alters the function of the heart in a negative manner can be classified as heart disease (Heart and stroke, n.d.). It is not merely associated with a person having a heart attack but rather a group of conditions that exacerbates the likelihood of a person getting a heart attack.

In this project, the heart dataset was analyzed extensively to derive meaningful patterns that were used to build different machine learning models to help make predictions about wether a person has a heart diseaseor not given a few attributes.

Problem Statement

The goal is to create 5 different machine learning models to predict if a given person has a heart disease or not. One of the five models will be a neural network, while another wil be an ensemble learning model. The tasks involved in accomplishing the problem statement include:

- Downloading and preprocessing the data
- Performing Exploratory Data Analysis
- Performing Data visualizations
- Data Cleaning
- Looking for target class imbalance
- Model building and evaluating algorithms
- Finalizing models

The dataset

The heartdata.csv dataset is a collection of medical data related to heart disease. It contains 303 records with 14 attributes each. The dataset has been widely used for research and analysis of heart disease, and is freely available for use in academic and commercial projects.

The first attribute in the dataset is age, which represents the age of the patient. The second attribute is sex, which represents the gender of the patient. The third attribute is chest pain type, which is a categorical variable that represents the type of chest pain the patient is experiencing. The fourth attribute is resting blood pressure, which is the patient's blood pressure while they are at rest. The fifth attribute is serum cholesterol, which is the patient's cholesterol level in mg/dl. The sixth attribute is fasting blood sugar, which represents whether the patient has fasting blood sugar > 120 mg/dl (1 = true; 0 = false).

The seventh attribute is resting electrocardiographic results, which is a categorical variable that represents the results of the patient's resting electrocardiogram. The eighth attribute is maximum heart rate achieved during exercise, which represents the highest heart rate the patient reached during an exercise test. The ninth attribute is exercise-induced angina, which represents whether the patient experienced angina during exercise (1 = yes; 0 = no).

The tenth attribute is oldpeak, which represents the ST depression induced by exercise relative to rest. The eleventh attribute is the slope of the peak exercise ST segment, which is a categorical variable that represents the slope of the peak exercise ST segment. The twelfth attribute is the number of major vessels (0-3) colored by fluoroscopy, which represents the number of major blood vessels in the heart that are colored during a fluoroscopy test. The thirteenth attribute is thalassemia, which is a categorical variable that represents the type of thalassemia the patient has. The final attribute is the presence of heart disease, which is the target variable and represents whether the patient has heart disease (1 = yes; 0 = no).

The heartdata.csv dataset has been used in a variety of studies related to heart disease. For example, it has been used to develop predictive models to identify patients at high risk of heart disease, to investigate the relationship between chest pain type and heart disease, and to explore the relationship between exercise-induced angina and heart disease.

Overall, the heartdata.csv dataset is a valuable resource for researchers and analysts working in the field of heart disease. Its comprehensive collection of medical data can help inform the development of new diagnostic and treatment strategies for heart disease, as well as support the development of predictive models to identify patients at high risk of heart disease.

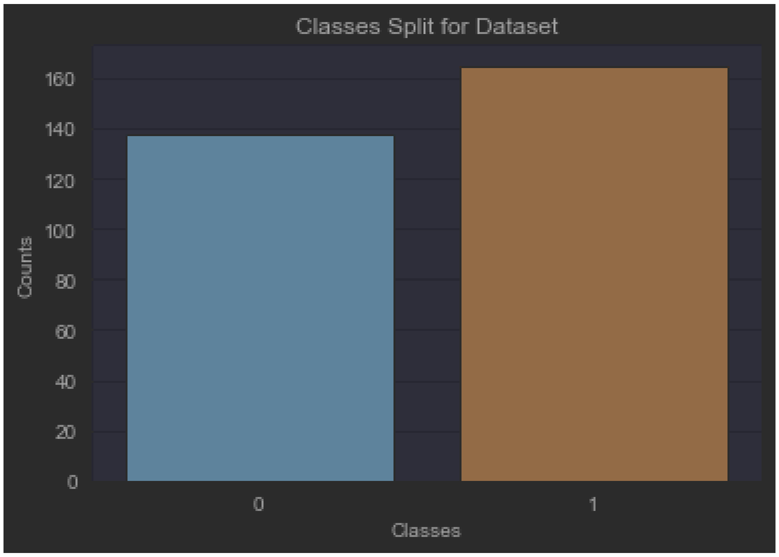
Dataset Attributes

- age - In years.
- sex - Female: 0, Male: 1.
- chest pain (cp) - 0: typical angina, 1: atypical angina, 2: non-anginal pain, and 3: asymptomatic.
- trestbps - resting blood pressure
- chol - serum cholesterol
- fasting blood sugar (fbs) > 120 mg/dl - 0: false, 1: true.
- resting electrocardiographic results (restecg) – 0: Normal, 1: abnormality, 2: LVT.
- thalach - maximum heart rate achieved.
- exercise induced angina (exang) – 0: no, 1: yes.
- oldpeak - ST depression induced by exercise relative to rest.
- the slope of the peak exercise ST segment (slope) - 0:up sloping, 1: flat, 2:down sloping.
- ca - number of major vessels (0-4) colored by fluoroscopy.

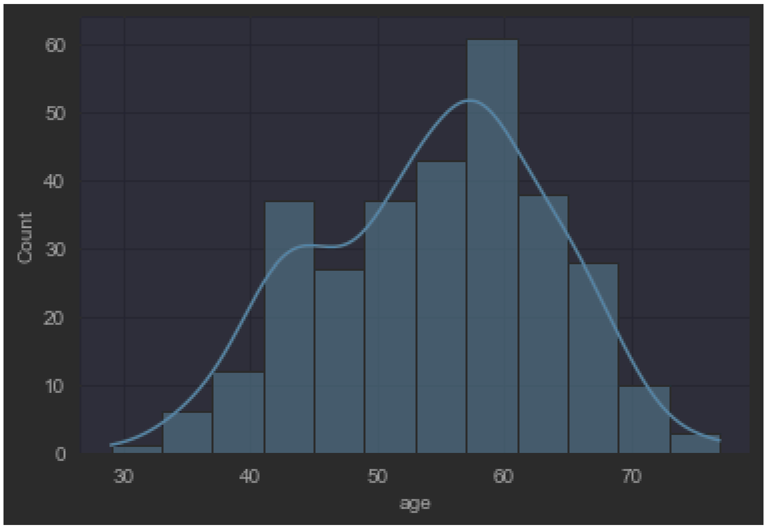
- thal – 0: normal, 1: fixed, 2: reversible, 3: ND.
- target – heart disease 0: no, 1: yes

Exploratory Data Analysis

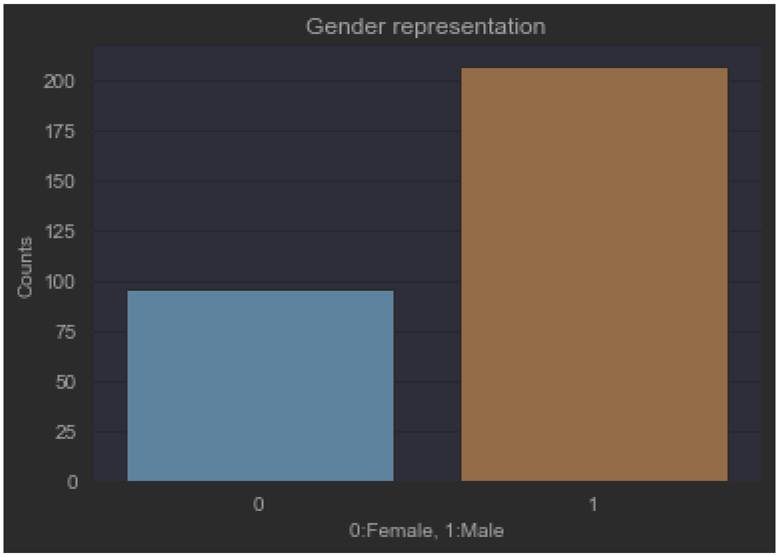
After looking at .info(), it would seem that we have a total of 303 records (rows) belonging to 14 columns. All columns are of integer type except for old peak, which belongs to the float type. We then looked at .describe() for our dataset and the statistical analysis revealed that the maximum age is 77 and 75% of the population in the dataset are 61 or younger, 50% of the population is 55 or younger, and the minimum age is 29. We then checked and plotted for class imbalance:



We found that there was a class imbalance, therefore we could use downsampling methods, upsampling methods, or SMOTE to balance the classes. We then plotted the age distribution to visualize the dataset's age demographic:



We can observe that the dataset has more or less a uniform distribution, however it can also seem like it is bimodal. We then took a look at the sex in the dataset to see if male and female were represented equally:

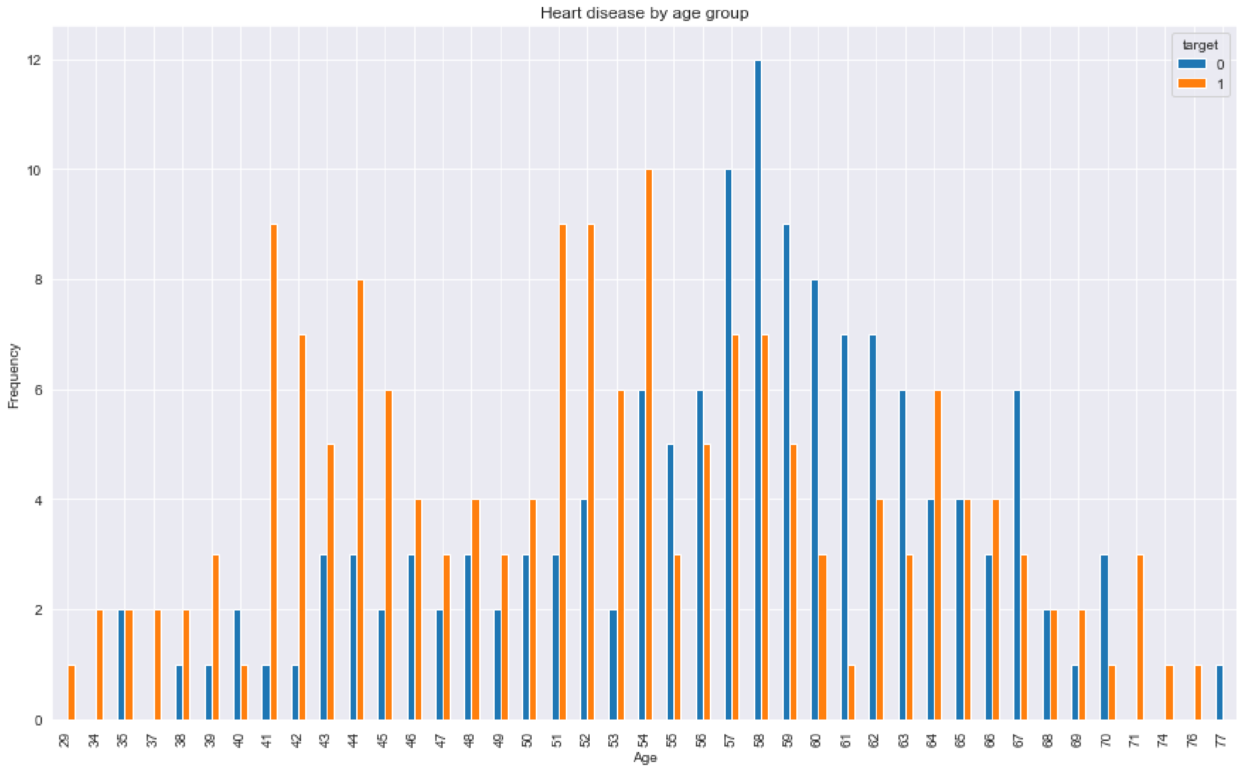


It would seem as there are more males than females in the dataset. Males had a 68.32% representation whereas females only had 31.68% representation.

Preparing the Data

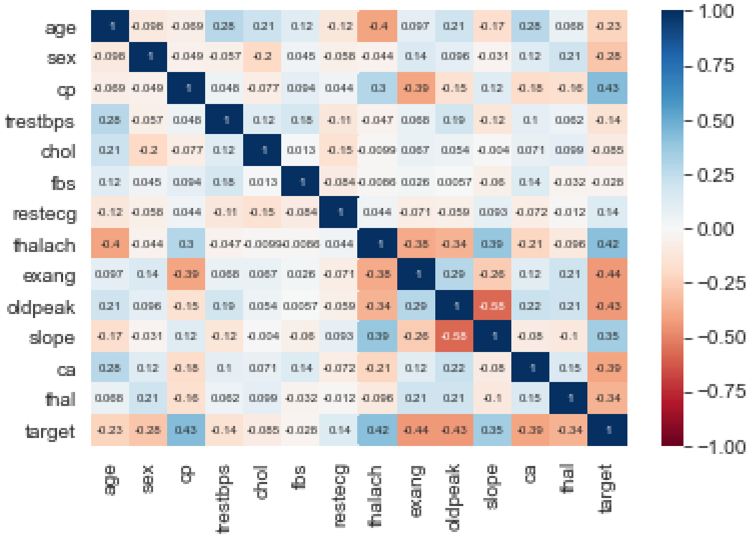
The first thing we checked for the dataset was if there are any missing values for each columns. The dataset has no missing values. Therefore we did not need to touch that problem. We then checked for duplicates, it turns out that there was one duplicate and to deal with that problem we dropped the duplicated rows. We then scaled the columns with continuous variables. Those columns are: [`'age','trestbps', 'chol','thalach', 'oldpeak'`]. We then dropped these same columns when we used pandas' getdummies function to encode the categorical columns of the dataset. Once we had two separate dataframes of the scaled columns and the dummies encoded columns we concatenated both of them to get a full dataframe once more. We checked for class imbalance and there was a significant amount of imbalance between the people with heart disease and without it. Therefore, we had to use SMOTE to balance the classes within the dataset.

Data Visualizations



Data visualization techniques like bar plots help understand how the data was distributed. For example, a bar plot of the gender attribute revealed that man and women were not equally represented. It was found that males made up the majority of the gender distribution accounting for about 68.32%, while women made up 31.68% of the distribution. Ideally our distribution should contain an equal number of classes representing the gender attribute.

A crosstab function was also used to learn the relationship between the ages and the target variable which helped understand the frequency of heart diseases across different age groups. We were able to learn that people aged 54 were the critical group that suffered the most number of heart diseases. It was also learned that people belonging to the age groups of 41, 51 and 52 were equally vulnerable to getting a heart disease. However, people aged 58 were also among those that did not suffer from any heart disease.



A heat map was used to understand how the different attributes were correlated to each other. A color map of blue to red was used in which blue represents positive correlation while red represents a negative correlation. We can see that the target is positively correlated with the features thalach albeit a not so strong correlation because the values are not close to 1. A strong negative correlaton between slope and oldpeak. The target variable is also negatively correlated with the features oldpeak and exang.

The diagnol line just represents the correlation for the variable with itself which is why it shows as 1.

Model Building and Evaluating Algorithms

The first step before we start building any machine learning model is to split our dataset into train, test and validate sets. The dataset was split as train 60%, test 20% and, validate 20%. The idea is to train models using the train set, improve the performance on the test set using hyperparameters and regularization. The validation set is a hold out set which is only used to test the final performance of the models to evaluate the models.

Model 1: RandomForest

The first model built was a random forest which is an ensemble of decision trees.It is a supervised machine learning model. The model is first trained without any hyperparameter tuning so it takes the default values specified in the scikit library. This provides us a baseline to compare to. However, we must be wary about some parameters like the max_depth for which the default value is none. This means that the nodes are expanded until all the nodes are pure. This generally leads to overfitting and should be avoided by either pruning the trees or by setting the hyperparameters.The model was used to predict on test set and accuracy score was measured to be 84%. The model was optimized further by using hyperparameter tuning techniques like RandomizedGridSearchCV which randomly selects hyperparameters from a specified range of values from each hyperparameter. The model was then trained again and used to make predictions on the test set. The models scored 81%. This drop in accuracy score reflects the fact that the first time we trained the model without any hyperparameter tuning, it learnt the data too well and was overfitting it. So the true score of the random forest after tuning hyperparameters is 81%. One thing to note is that since we are using a randomized search, each time we run the algorithm, our accuracy score will differ. The model was finally tested on the hold out validation set and scored 80% on the accuracy which is not too far from the 81% accuracy on the test set.

Model 2: K Nearest Neighbors

K nearest neighbors is a non parametric supervised learning algorithm which does not make any assumptions about the linearity or the distribution of the data. The model works by memorizing the training data and uses it to classify new data points based on their proximity to existing data points. The model was trained on the train set and the test set was used to make predictions. The preliminary training resulted in an accuracy score of 80%. A RandomizedSearchCV was also used to determine the best hyperparameters and the model was trained again on the training data. It was found that the model improved in accuracy with a score of 83%. The model was finally tested on the hold out validation set and it scored 77% in accuracy. This just means that the model may not be the best one for further predictions.

Model 3: Support Vector Classifier

Support vector classifiers are classification algorithms that use support vector machines to find the best boundary between two clases of data. The model was trained using default parameters and scored 85% on the first attempt. The model did not see a change in the accuracy score after hyperparamater tuning using RandomizedGridSearchCV like the other models did. Itstayed consistent at the same accuracy of 85% as before. When the model was tested on the hold out validation set, it dropped to 83% which is still not bad because the model did well on the training and test set and also generalized well to new and unseen data which is the validation set.

Model 4: Voting Ensemble

A Voting Ensemble is a machine learning technique that combines multiple individual classifiers to make a prediction. In a Voting Ensemble, each individual classifier makes a prediction, and the final prediction is determined by combining the predictions of all the classifier. For this model, hard voting was used which means the model picks the majority vote of the predictions made by individual classifiers. Three models were chosen for the ensemble, a decision tree, linear regression and, ada boost classifier and their individual accuracy scores were 74%, 80%, and 79% respectively. The model was tested on the hold out validation set and it scored 80%.

Model 5: Neural Network

A sequential neural network with three layers was trained on the training data. The loss and accuracy of the model as it trained was measured to be 1.39 and 85% respectively. The model was trained further by adding two more layers and the first layer has 128 nodes, uses the ReLU activation function, and includes L2 regularization with a coefficient of 0.01. The next four layers have 64, 32, 16, and 1 nodes, respectively, and use the ReLU and sigmoid activation functions.

Dropout regularization is applied after the second, third, and fourth layers with a rate of 0.5, which randomly drops out some of the nodes during training to prevent overfitting. The model is compiled with the binary cross-entropy loss function, the Adam optimizer, and accuracy as the evaluation metric. The model is trained on the training data for 200 epochs with a batch size of 16 and validated on the test data after each epoch to monitor the model's performance on unseen data. The NN was tested on the hold out validation set and because our outpu tlayer is sigmoid, it gives us a probability which we converted to either being 0 or 1 by applying a threshold of 0.50. The models predictions scored an accuracy score of 81%.

Business Insights

The heartdataset.csv models can be important for an insurance business because heart disease is a major health concern that can increase the risk of mortality and morbidity, which in turn can increase the cost of medical treatments and insurance claims. By using machine learning models to predict the risk of heart disease, insurance companies can better assess the risk of insuring a particular individual and adjust their premiums accordingly. This can help insurance companies to provide more accurate pricing, minimize losses, and ultimately provide better value to their customers. Additionally, by identifying high-risk individuals, insurance companies can offer targeted interventions to help prevent or manage heart disease, which can result in better health outcomes and cost savings for both the individual and the insurer.

Conclusion

After comparing the performance of all the models and testing them two fold on the test set and the validation sets, it was found that the support vector classifier was the best performer with an accuracy score of 83% This model performed well on the training data and the test set and also generalized well to new hold out validation set. This will help us make accurate predictions in predicting wether a person has a heart disease or not given the attributes.