**Heart Failure Prediction**

Gerard Corrales

Gabriel Duffy

Shiley-Marcos School of Engineering, University of San Diego

# **Abstract**

**Table of Contents**

[**Abstract** 2](#_Toc173946747)

[**Introduction** 4](#_Toc173946748)

[**Methodology** 5](#_Toc173946749)

[CART 6](#_Toc173946750)

[C5.0 7](#_Toc173946751)

[Random Forest 8](#_Toc173946752)

[Naïve Bayes 9](#_Toc173946753)

[Neural Networks 10](#_Toc173946754)

[**Results** 11](#_Toc173946755)

[**Conclusion** 12](#_Toc173946756)

[**References** 13](#_Toc173946757)

[**Appendix** 14](#_Toc173946758)

# **Introduction**

Heart disease is one of the leading causes of mortality worldwide (Roth & Mensah, 2020). Early detection and intervention are crucial for reducing heart disease and improving patient results. This project aims to leverage machine learning techniques to predict the presence of heart disease based on a set of descriptive features collected from patients. The first objective of this project is to conduct feature analysis and exploration to identify and understand which features are most indicative of heart disease. This involves exploratory data analysis (EDA) to uncover the relationships and patterns among the variables. Second, to develop and evaluate various classification models to predict heart disease, including CART, C5.0, Random Forests, Naïve Bayes Classification, and Neural Networks. Each model will be evaluated based on criteria such as accuracy, sensitivity, specificity, and misclassification.

The data set used in this project includes a variety of features such as age, resting blood pressure, cholesterol levels, fasting blood sugar, maximum heart rate, exercise-induced angina, and other medical indicators. These features will be preprocessed and standardized to ensure model performance. The expected outcome of this project is to identify the most accurate machine learning model for predicting heart disease.

# **Methodology**

The dataset for this project was obtained from a public repository on GitHub. The dataset contains a total of 918 records and 12 features. The dataset includes a set of features collected from patients, such as age, sex, chest pain, resting blood pressure, cholesterol levels, fasting blood sugar, resting electrocardiogram results, maximum heart rate, exercise-induced angina, numeric value measured in depression (oldpeak), the slope of the peak exercise (ST), and heart disease.

Initial observations revealed no missing values in the dataset. However, outliers were identified in the MaxHR and Oldpeak columns, particularly for the z-scores of MaxHR and Oldpeak, indicating some extreme values that required attention. The dataset also exhibited varying degrees of skewness across features, with notable skewness in the Cholesterol, FastingBS, and Oldpeak columns, among others.

To ensure the data was suitable for analysis, a series of preprocessing steps were carried out. Initially, wrong values in the dataset were handled using K-Nearest Neighbors (KNN) imputation to estimate and fill in the gaps for numerical variables. Categorical variables such as Sex, ChestPainType, RestingECG, ExerciseAngina, and ST\_Slope were transformed into numerical format using one-hot encoding, creating dummy variables for each category level. The Oldpeak variable, which had negative values, was standardized using Min-Max scaling to bring all features onto a common scale. This helped in ensuring that all features contribute equally during model training. Identified outliers in the MaxHR and Oldpeak columns were treated to ensure they do not skew the results. Figure 1 and Figure 2 show histograms of age distribution for patients with and without heart disease, with Figure 2 normalizing the distribution to proportionally compare the two groups.

**Figure 1**

*Histogram of Age with Response Overlay*

A graph of a graph with blue and orange bars

Description automatically generated

**Figure 2**

*Normalized Histogram of Age with Response Overlay*

A graph with a number of age and response levels

Description automatically generated

## **CART**

The CART algorithm was employed to develop a predictive model for heart disease. CART is a decision tree algorithm that uses the Gini impurity criterion to split the data into homogenous groups. The CART model was trained on the preprocessed dataset, which included both numerical and categorical variables transformed via one-hot encoding. The dataset was split into training and testing sets, with 75% of the data used for training and 25% reserved for testing. The training set was balanced 50% yes and 50% no responses, it was used to build the decision tree, and it was utilized to evaluate the model's performance.

The model identified several key predictors of heart disease. Among the most significant features were ST\_Slope, Sex, MaxHR, and ExerciseAngina (as seen in Figure 3). These variables exhibited the strongest influence on the model’s predictions, highlighting their critical role in assessing heart disease risk.

**Figure 3**

*CART Model Decision Tree*

A diagram of a network

Description automatically generated with medium confidence

The resulting decision tree provided a clear and interpretable set of rules for predicting heart disease. For instance, the root node split on ST\_Slope, indicating that patients with a ST\_Slope value of less than or equal to 0.5 were more likely to not have heart disease. Subsequent splits on features like Sex and MaxHR further refined the classification, demonstrating that male patients with a MaxHR of less than or equal to 150.5 were more likely to have heart disease. The decision tree also indicated that patients with an ExerciseAngina value of less than or equal to 0.5 were more likely to not have heart disease. The accuracy of the model was found to be 80%, indicating a strong ability to correctly classify patients as having heart disease or not. Additionally, the confusion matrix showed a good balance between sensitivity (true positive rate) and specificity (true negative rate), suggesting that the model effectively identifies both positive and negative cases. In Table 1 shows the results of the classification report.

**Table 1**

*Results of Classification Report*

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | Precision | Recall | F1-Score | Support |
| 0 | 0.80 | 0.77 | 0.79 | 111 |
| 1 | 0.80 | 0.82 | 0.81 | 119 |
| Accuracy |  |  | 0.80 | 230 |

## **C5.0**

The C5.0 model was trained on the training dataset. The training and testing splits were identical, with 75% of the data used for training and 25% reserved for testing. The training set was used to construct the decision tree, and the testing set was employed to evaluate the model's performance. The C5.0 model identified several critical predictors of heart disease. The most influential features were ST\_Slope, Sex, MaxHR, and Oldpeak. The resulting decision tree provided an interpretable set of rules for heart disease prediction. As seen in Figure 4, the root node split on ST\_Slope, indicating that patients with an ST\_Slope value of less than or equal to 0.5 were more likely to not have heart disease. For example, male patients (Sex\_M <= 0.5) with a MaxHR of less than or equal to 150.5 were more likely to have heart disease. Additionally, patients with an Oldpeak value of less than or equal to 0.45 were more likely to not have heart disease.

**Figure 4**

*C5.0 Results*

A diagram of a network

Description automatically generated with medium confidence

The C5.0 model demonstrated robust performance on the test dataset, with an accuracy of 83%. The confusion matrix (Table 2) revealed a good balance between sensitivity and specificity, indicating the model's efficacy in identifying both positive and negative cases of heart disease.

**Table 2**

*C5.0 Confusion Matrix*

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | Precision | Recall | F1-Score | Support |
| 0 | 0.78 | 0.90 | 0.83 | 111 |
| 1 | 0.89 | 0.76 | 0.82 | 119 |
| Accuracy |  |  | 0.83 | 230 |

## **Random Forest**

The Random Forest model provided valuable insights into the importance of different features in predicting heart disease. As seen in Figure 5, the most important features identified by the model included ExerciseAngina\_Y, ST\_Slope\_Up, Cholesterol and Sex\_M, MaxHR and Oldpeak, and RestingBP. ExerciseAngina indicates that patients without exercise-induced angina (ExerciseAngina\_Y <= 0.5) had a higher likelihood of not having heart disease. The slope of the ST segment during peak exercise was also a crucial factor. Patients with an upward ST slope (ST\_Slope\_Up <= 0.5) showed a lower likelihood of heart disease. Resting blood pressure appeared multiple times in the tree, with various thresholds influencing the likelihood of heart disease. Lower resting blood pressure values were associated with a higher probability of heart disease. Cholesterol levels and gender (Sex\_M) also played significant roles in the decision tree, with different thresholds impacting the classification. Maximum heart rate achieved and the ST depression induced by exercise relative to rest (Oldpeak) were other important predictors, influencing subsequent splits in the tree.

**Figure 5**

*Random Forest Decision Tree*

A diagram of a computer network

Description automatically generated with medium confidence

## **Naïve Bayes**

When evaluating the Naive Bayes model, the confusion matrix (Table 3) yielded the following results: 88 true negatives (cases where no heart disease was correctly identified), 23 false positives (cases where heart disease was incorrectly predicted), 20 false negatives (cases where heart disease was not predicted but was present), and 99 true positives (cases where heart disease was correctly predicted). This evaluation resulted in an overall model accuracy of 81%. The feature log probabilities for each class (heart disease present or not) were visualized using a heatmap.

**Table 3**

*Naïve Bayes Confusion Matrix*

|  |  |  |  |
| --- | --- | --- | --- |
| Predicted | False | True | Total |
| Actual |  |  |  |
| 0 | 88 | 23 | 111 |
| 1 | 20 | 99 | 119 |
| Total | 108 | 122 | 230 |

## **Neural Networks**

The model was compiled using the Adam optimizer and binary cross-entropy loss function, suitable for binary classification tasks. It was trained over 150 epochs with a batch size determined automatically by Keras. The training dataset was used to fit the model, achieving an accuracy of approximately 87% (see Figure 6), indicating the model's ability to correctly classify the presence or absence of heart disease. After training, the model was evaluated on the test dataset. While neural networks do not provide feature importance in the same way as decision trees or other interpretable models, the architecture and weights learned during training highlight the complex interactions between features.

**Figure 6**

*Neural Networks Accuracy*

**A graph of a line and a line

Description automatically generated with medium confidence**

# **Results**

# **Conclusion**

# **References**

Larose, C., & Larose, D. (2019). *Data Science Using Python and R*. Wiley.

Tan, P.-N., Steinbach, M., Karpatne, A., & Kumar, V. (2020). *Introduction to data mining*

(Second Edition). Pearson.

Roth, G. A., Mensah. (2020). Global Burden of Cardiovascular Diseases and Risk Factors, 1990-2019: Update From the GBD 2019 Study. *Journal of the American College of Cardiology*, *76*(25), 2982–3021. https://doi.org/10.1016/j.jacc.2020.11.010

‌ Jayachandru001. (2021). *GitHub - jayachandru001/Heart-Failure-Prediction-: This project involves training of Machine Learning models to predict the Heart Failure for Heart Disease event. In this KNN gives a high Accuracy of 89%.* GitHub. https://github.com/jayachandru001/Heart-Failure-Prediction-/tree/main

# **Appendix**