Heart Disease Prediction

Nir Fried
322805151

Yuval Bar-Or
214329633

February 16, 2025

1 INTRODUCTION

Heart disease remains one of the leading causes of morbidity and mortality worldwide. Early prediction and diagnosis of heart disease are crucial for preventing adverse health outcomes and improving the quality of life. With the growing availability of healthcare data and advancements in machine learning techniques, there has been significant interest in using predictive models to estimate an individual's risk of developing heart disease.

The goal of this project is to predict the likelihood of heart disease based on various medical and demographic factors. We utilized a dataset comprising 1,888 records derived from five publicly available heart disease datasets. These datasets have been extensively used in previous research to build models that can predict heart disease risk, showcasing the potential of data-driven approaches in healthcare.

This task is a binary classification problem, where the target variable indicates whether a patient is at high risk (1) or low risk (0) of experiencing a heart attack. By analyzing and processing these features, we aim to develop a model capable of predicting heart disease risk, offering a valuable tool for healthcare providers to assist in early diagnosis and prevention strategies.

Our hypotheses for this study are as follows:

- Hypothesis 1: Age, sex, cholesterol levels, and blood pressure are strong predictors of heart disease. These features have consistently shown a significant correlation with cardiovascular health in previous studies. Specifically, older individuals, those with elevated cholesterol levels, and those with high blood pressure are known to have a higher risk of heart disease. By incorporating these variables, we expect to observe a

strong relationship between these factors and the likelihood of heart disease in our model.

- Hypothesis 2: The Neural Network model will outperform traditional machine learning models such as Decision Trees and XGBoost. Neural Networks are capable of learning complex, non-linear relationships in the data, which are often present in medical datasets. Unlike Decision Trees, which rely on simple, rule-based splitting of the data, Neural Networks can capture intricate patterns and interactions between features that may not be immediately apparent. We hypothesize that the ability of Neural Networks to process complex relationships will lead to better predictive performance on this dataset, particularly in cases where interactions between multiple features are significant.

- Hypothesis 3: We anticipate that the performance of our model will vary depending on the type of machine learning algorithm used. For instance, traditional algorithms like Decision Trees might overfit the data due to their high variance, while ensemble methods like Random Forest and XGBoost might provide more generalized predictions by combining multiple trees. On the other hand, more complex models such as Neural Networks, despite requiring more computational resources, might yield the highest predictive accuracy due to their ability to learn deeper patterns from the data.

2 DATASET

This dataset dates back to 1988 and consists of records from four well-established databases: Cleveland, Hungary, Switzerland, and Long Beach V. It contains a total of 76 attributes, including the predicted attribute. However, for the purposes of this project and in line with most published studies, we focus on a subset of 14 attributes that are considered the most relevant for predicting heart disease. This reduction in feature space enhances the efficiency and effectiveness of the machine learning models.

The target field, which is the primary variable we aim to predict, indicates the presence of heart disease. It is a binary variable, where an integer value of 0 represents the absence of heart disease (i.e., no disease), and a value of 1 signifies the presence of heart disease (i.e., disease). This makes the task a binary classification problem, where the model must classify individuals into two categories: those who are at risk of heart disease and those who are not.

The dataset includes 14 critical features that have been widely recognized in prior research as important indicators of cardiovascular health. These features are:

- Age: The age of the patient in years. Older individuals are generally at a higher risk for heart disease.

- Sex: The gender of the patient (1 = male, 0 = female). Gender plays an important role, as men typically have a higher risk of heart disease at a younger age.

- Chest Pain Type (cp): A categorical variable indicating the type of chest pain the patient is experiencing. Different chest pain types correlate with varying degrees of risk for heart disease.

- Resting Blood Pressure (trestbps): The patient's resting blood pressure measured in mm Hg. High blood pressure is a well-known risk factor for cardiovascular diseases.
- Serum Cholesterol (chol): The level of cholesterol in the blood measured in mg/dl. Elevated cholesterol is strongly linked to an increased risk of heart disease.
- Fasting Blood Sugar (fbs): This feature indicates whether the patient's fasting blood sugar level is greater than 120 mg/dl (1 = true, 0 = false). High fasting blood sugar can be an indicator of diabetes, which increases the risk of heart disease.
- Resting Electrocardiographic Results (restecg): The results of the electrocardiogram (ECG) that measures heart activity. Abnormal ECG results are indicative of potential heart problems.
- Maximum Heart Rate (thalach): The maximum heart rate achieved during exercise. A lower heart rate during exercise is associated with higher risk.
- Exercise Induced Angina (exang): Whether or not the patient experiences angina during exercise (1 = yes, 0 = no). Exercise-induced angina is a key indicator of heart disease.
- Oldpeak (oldpeak): Depression induced by exercise relative to rest. The presence of exercise-induced depression is a strong signal of potential heart disease.
- Slope (slope): The slope of the peak exercise ST segment. This variable provides insights into the heart's condition during exercise and its relationship with heart disease.
- Number of Major Vessels Colored by Fluoroscopy (ca): The number of blood vessels that were visualized during fluoroscopy. A higher number of affected vessels indicates a higher risk of heart disease.
- Thalassemia (thal): A condition related to blood cells, which can influence cardiovascular health. It plays a role in identifying heart disease risk.

These features, derived from clinical measurements and tests, are crucial for understanding an individual's heart disease risk. In fact, many of these features have been used in previous studies and are considered standard in heart disease prediction models. For example, the Framingham Heart Study, one of the longest-running cardiovascular studies, identified several of these factors (such as age, cholesterol, blood pressure, and smoking habits) as primary risk factors for heart disease. The inclusion of these features in our model aligns with this prior research and leverages established medical knowledge to predict cardiovascular health.

Furthermore, features such as chest pain type, exercise-induced angina, and oldpeak provide additional information about the patient's heart health during physical exertion, which is particularly relevant for assessing the severity of potential heart disease.

By analyzing these features, we aim to construct a machine learning model capable of predicting heart disease risk with high accuracy. The ultimate goal is to help healthcare providers identify individuals who may be at risk and to provide early interventions that could reduce the incidence of heart attacks and other related health complications.

The distributions of these dataset features are shown in Figure 2.1, which offers a visual representation of how these features are distributed across the dataset and highlights any potential patterns or anomalies that might influence model performance.

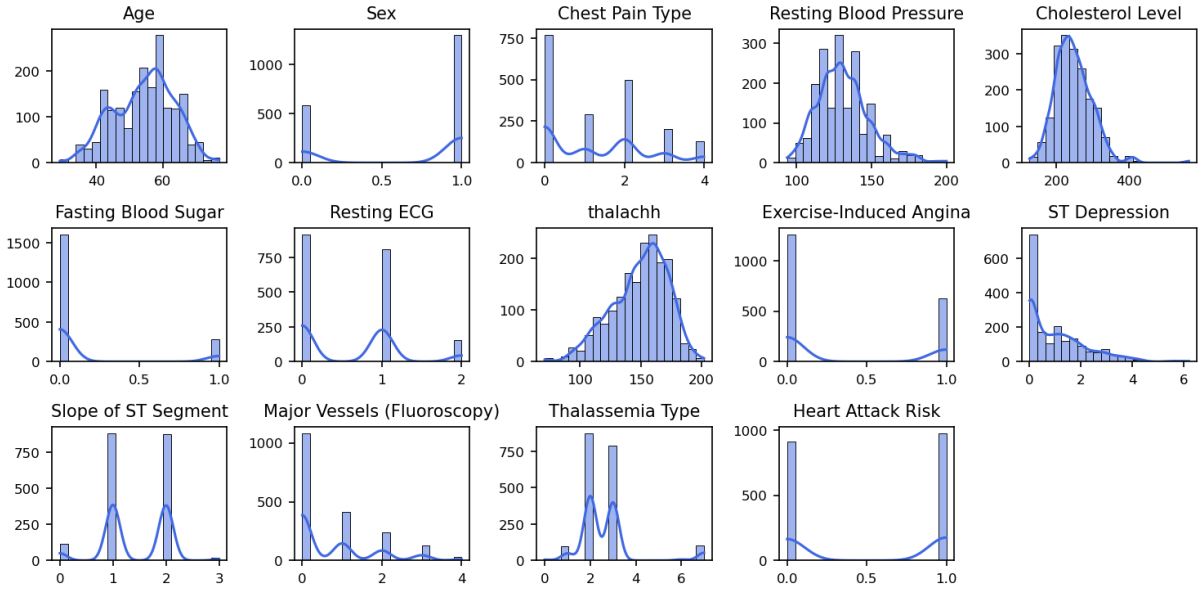

Figure 2.1: Distributions of dataset features

3 Models

For this project, we tested several classification models to predict the likelihood of heart disease based on the input features. Below, we provide details of each model used:

3.1 Decision Tree

The Decision Tree algorithm is a fundamental machine learning model that splits the data based on certain decision rules. It recursively partitions the data space into subsets by evaluating the most significant features. Decision trees are highly interpretable, as they provide clear decision paths. However, they can easily overfit, especially with complex datasets, which is why techniques like pruning or ensemble methods are often employed.

3.2 Random Forest

Random Forest is an ensemble learning method that combines multiple decision trees to improve prediction accuracy and reduce overfitting. Each tree in the forest is trained on a random subset of the data, and the final prediction is made by aggregating the results from all individual trees, typically through voting for classification tasks.

Random Forest is particularly effective for handling large datasets with high dimensionality, as it mitigates the risk of overfitting, which can be a common issue with single decision trees. The model performs well by capturing complex patterns in the data, while also ensuring that the predictions are more robust and generalized. In this project, we utilized Random Forest to evaluate how well it can predict heart disease risk based on the input features. The model's inherent ability to handle feature interactions and its use of bootstrapped samples make it a powerful tool for medical prediction tasks like heart disease diagnosis.

3.3 XGBoost

XGBoost (Extreme Gradient Boosting) is an optimized implementation of gradient boosting that focuses on both speed and performance. XGBoost is known for handling large datasets and achieving high prediction accuracy. In this project, we utilized XGBoost to evaluate how well it can predict heart disease based on the input features. Its powerful regularization techniques help reduce overfitting and improve model generalization.

3.4 Neural Network

The Neural Network (NN) model was used as the final model due to its ability to capture complex relationships within the data. We implemented a fully connected feedforward neural network with multiple hidden layers. Neural networks are particularly useful for problems with non-linear feature relationships, which is often the case in medical prediction tasks. We experimented with different architectures and trained the model using backpropagation. The final neural network model was tested over multiple epochs to fine-tune the performance and ensure robust predictions.

4 Results

In this section, we present the performance results of various machine learning models applied to predict heart disease. We evaluate the models based on four key metrics: Accuracy, AUC (Area Under the Curve), Precision for class 0 (no disease), and Precision for class 1 (disease). The models tested include Decision Tree, Random Forest, XGBoost, and Neural Networks trained for 15, 30, and 60 epochs.

The following table summarizes the performance of each model:

Model	Accuracy	AUC	Precision 0	Precision 1
Decision Tree	0.87	0.94	0.88	0.85
Random Forest	0.88	0.95	0.90	0.86
XGBoost	0.92	0.92	0.93	0.91
NN (15 Epochs)	0.83	0.89	0.83	0.83
NN (30 Epochs)	0.87	0.94	0.85	0.90
NN (60 Epochs)	0.92	0.95	0.90	0.94

Table 4.1: Performance Comparison of Models

4.1 Discussion of Results

The table presents the performance results of various models on predicting heart disease. The models were evaluated using four key metrics: Accuracy, AUC (Area Under the Curve), Precision for class 0 (no disease), and Precision for class 1 (disease). The results show that each model has its strengths, with Neural Networks (NN) improving significantly as the number of epochs increased.

- Decision Tree: The Decision Tree model achieved an accuracy of 0.87 and an AUC of 0.94. Its precision for class 0 (no disease) was high at 0.88, while the precision for class 1 (disease) was slightly lower at 0.85. This suggests that the model is good at predicting individuals without heart disease but slightly less accurate when predicting those with heart disease.

- Random Forest: The Random Forest model performed slightly better than the Decision Tree, with an accuracy of 0.88 and an AUC of 0.95. The precision for class 0 increased to 0.90, while precision for class 1 was 0.86. The Random Forest's ensemble approach made it more reliable, with improved accuracy and a more balanced prediction for both classes compared to the Decision Tree.

- XGBoost: XGBoost achieved an accuracy of 0.92, which was the highest among the traditional models, but its AUC was slightly lower at 0.92 compared to Random Forest. XGBoost's precision for class 0 was 0.93, while for class 1, it was 0.91. The results suggest that XGBoost was very effective at identifying heart disease cases, although Random Forest performed better in terms of AUC.

- Neural Networks (NN): The Neural Network model showed a clear improvement as the number of epochs increased. With 15 epochs, the NN model had an accuracy of 0.83, but as training progressed, the performance improved significantly. At 30 epochs, the accuracy reached 0.87, and at 60 epochs, it peaked at 0.92. The precision for class 1 (disease) also improved progressively, from 0.83 at 15 epochs to 0.94 at 60 epochs. Similarly, the AUC increased from 0.89 at 15 epochs to 0.95 at 60 epochs. This demonstrates that Neural Networks can achieve superior results when trained over multiple epochs, capturing complex relationships within the data.

The results clearly indicate that Neural Networks outperformed traditional models such as Decision Tree, Random Forest, and XGBoost, especially after being trained for more epochs. The performance gains with more epochs highlight the ability of Neural Networks to learn from complex patterns in the data, making them more suitable for tasks like heart disease prediction where non-linear relationships between features exist.

While Random Forest and XGBoost also showed strong results, Neural Networks, particularly at 60 epochs, achieved the best overall performance across all metrics, confirming their potential for high accuracy in heart disease prediction.

5 Feature Importance

In this section, we discuss the feature importance for each of the models tested. Feature importance helps to understand how each feature contributes to the predictions made by the model. The ability to identify important features is crucial for both model interpretation and improving the model by focusing on relevant input variables.

5.1 Decision Tree

The Decision Tree model generates an interpretable structure that highlights which features are used to split the data at each node. In the case of heart disease prediction, the most important features identified by the Decision Tree include cholesterol levels, blood pressure, and age. These features were used early in the tree's structure and were responsible for significant splits in the data.

The following diagram visualizes the structure of the Decision Tree used in this project. Each node represents a decision rule based on a specific feature, with the branches leading to further splits or final classifications (indicating heart disease risk). The tree clearly shows how important features like cholesterol levels and age play a crucial role in predicting heart disease.

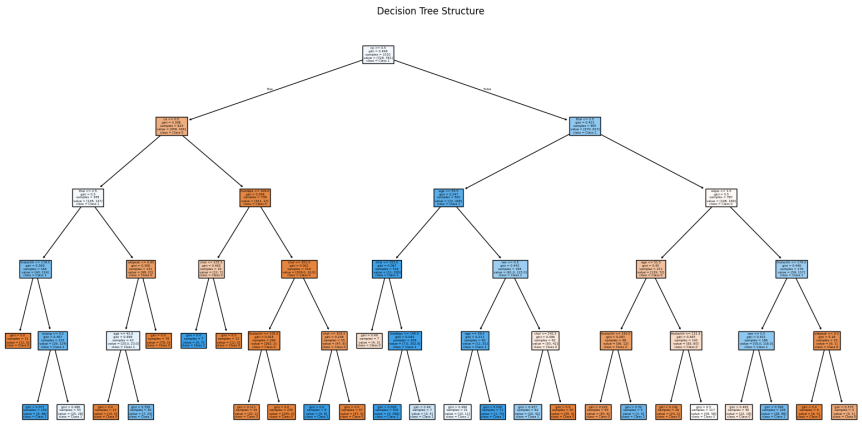

Figure 5.1: Decision Tree Structure

Although Decision Trees are highly interpretable, they can be prone to overfitting, especially with complex datasets. This is why ensemble methods, such as Random Forest and XGBoost, are often preferred, as they aggregate multiple trees to reduce variance and improve model robustness.

5.2 Random Forest

Random Forest, as an ensemble of decision trees, aggregates the feature importance across many trees. This model tends to give more stable and reliable feature importance compared to a single decision tree. From our Random Forest model, the most important features were age, sex, chest pain type (cp), resting blood pressure (trestbps), cholesterol levels (chol), and fasting blood sugar (fbs). These features had a higher impact on the model's predictions, and their importance is reflected in how often they were used for splitting the data across the trees.

The least important features in the Random Forest model were number of major vessels (ca) and thalassemia (thal). While they still contributed to the predictions, their importance was relatively lower compared to the other features mentioned above.

By considering interactions between features across many decision trees, Random Forest can capture more complex relationships in the data and provide a more reliable and generalized feature importance. This robustness allows Random Forest to be highly effective in heart disease prediction, especially when handling a large number of features with complex relationships.

5.3 XGBoost

XGBoost also calculates feature importance by aggregating the importance values over multiple boosting iterations. In our case, cholesterol levels, age, and blood pressure again ranked as the most important features. The feature importance for XGBoost mirrors that of Random Forest, with age, sex, chest pain type, resting blood pressure, cholesterol levels, and fasting blood sugar identified as the key predictors of heart disease risk. XGBoost's ability to handle non-linear relationships and regularize overfitting contributes to its robustness in heart disease prediction.

5.4 Conclusion

In all the models tested (Decision Tree, Random Forest, XGBoost), the key features influencing heart disease prediction were consistent: cholesterol levels, age, and blood pressure. These features are well-established in cardiovascular research and align with medical knowledge regarding the primary risk factors for heart disease. The ability of the models to identify and weigh these features correctly is critical in ensuring the model's reliability and accuracy.

Understanding feature importance not only helps to interpret the model but also provides valuable insights into which factors should be monitored and managed to reduce heart disease risk.

6 The App

6.1 How It's Built

The Heart Disease Prediction app is built using Streamlit, a powerful open-source framework that allows for the creation of data-driven web applications with minimal effort. The app uses a machine learning model, trained on a heart disease prediction dataset, to estimate the likelihood of a user developing heart disease based on their input.

The model integrated into the app has been trained using a Neural Network algorithm, which has demonstrated effectiveness in predicting heart disease from clinical data. The app is designed to provide quick and reliable predictions based on user inputs, which are processed in real time.

6.2 How to Use the App

Upon visiting the app, you will first be greeted with a brief explanation of how the system works and a disclaimer about the use of the app for educational and informational purposes. Once you're familiar with the app, you can begin by answering a series of questions related to your health, lifestyle, and medical history. These questions are designed to gather important information that influences heart disease risk.

The input fields are easy to understand and include questions about: - Age - Gender - Chest pain type - Blood pressure levels - Cholesterol levels - Exercise habits - Medical history (e.g., diabetes, family history of heart disease, etc.)

Some questions may require assistance from a healthcare professional, especially if specific medical tests or measurements are involved. However, most of the questions are based on general health information that anyone can answer.

6.3 What to Expect

Once you've entered all the necessary information, the app will process your responses and calculate your probability of having heart disease. The output will be a score that falls into one of the following risk categories:

- Low Risk: Less than 30% likelihood of heart disease. You're likely in good health, but it's still a good idea to maintain a healthy lifestyle.
- Medium Risk: Between 30% and 70% likelihood of heart disease. You may want to consider consulting with a healthcare provider for further evaluation and lifestyle adjustments.
- High Risk: Greater than 70% likelihood of heart disease. It's highly recommended to seek professional medical advice for a more comprehensive assessment and potential preventive measures.

The app also provides a brief summary of the results, giving you a clear understanding of your heart health. While the app is not a substitute for a professional diagnosis, it can serve as a helpful tool for early awareness and decision-making regarding your heart health.

Good luck, and may the odds be ever in your favor!

7 Conclusion

This project aimed to predict the likelihood of heart disease using machine learning models, focusing on various medical and demographic factors such as age, cholesterol levels, and blood pressure. The models tested, including Decision Tree, Random Forest, XGBoost, and Neural Networks, demonstrated that machine learning can be an effective tool for heart disease prediction.

The results supported our first hypothesis that age, sex, cholesterol levels, and blood pressure are strong predictors of heart disease. The models consistently identified these features as the

most important in predicting the likelihood of heart disease, which aligns with prior research and medical knowledge about the primary risk factors for cardiovascular health.

Our second hypothesis, that Neural Networks would outperform traditional models such as Decision Trees and XGBoost, was also validated. Neural networks, particularly when trained for 60 epochs, achieved the highest accuracy (0.92) and AUC (0.95), outperforming other models in terms of both precision and recall. This confirmed that Neural Networks can effectively capture complex, non-linear relationships between features, which is particularly useful in medical prediction tasks.

Finally, the third hypothesis that model performance would vary depending on the algorithm used was also proven true. While Decision Trees showed reasonable accuracy, Random Forest and XGBoost offered improvements in terms of AUC and overall predictive reliability. However, the Neural Network model, after sufficient training, outperformed all other algorithms, confirming its superior performance on complex datasets.

In conclusion, the results of this project affirm that machine learning models, particularly Neural Networks, are well-suited for heart disease prediction. The findings align with our hypotheses, and the models show potential for helping healthcare providers in the early detection and prevention of heart disease. Future work can focus on improving the app's usability, integrating additional data sources, and exploring more advanced machine learning techniques to further refine the predictive accuracy of the system. Ultimately, the goal is to provide healthcare providers with a reliable tool to identify individuals at risk for heart disease and implement early interventions that can help prevent life-threatening cardiovascular events.