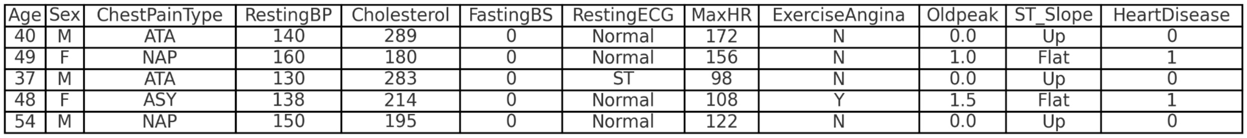
Introduction

The goal of this project is to predict the presence of heart disease using decision tree and neural network models. Accurate predictions can assist healthcare practitioners in making timely diagnoses and provide patients with an opportunity for early screening. We trained models using features such as 'Age', 'RestingBP', 'Cholesterol', and 'MaxHR', with the target column 'HeartDisease' indicating the presence (1) or absence (0) of the condition. To ensure model reliability, we implemented a decision tree with varying depths to evaluate complexity, and a neural network with different hidden layer configurations. Performance metrics like precision, recall, and F1 score were used to evaluate the models. The best results were obtained using a neural network with one hidden layer, achieving a mean F1 score of 0.7702. Increased model complexity did not consistently improve performance, indicating that simpler models can be effective for predicting heart disease. Github repository and presentation links.

Dataset

The dataset embraced in this project comprises of various attributes relevant to the patient health such as age, average blood pressure, cholesterol and heart rate measurements. The salient features include: **Age at patient**: This refers to age of the patient in years. **RestingBP**: The parameter denotes the sitting blood pressure value in mm Hg. **Cholesterol**: Indicates cholesterol level in the range of mg/dL. **MaximumHR**: This represents maximum heart rate achieved during an activity. **FastingBS**: Indicates fasting blood glucose level (1 = OB > 120 mg/dL, 0 otherwise). **Oldpeak**: ST depression induced by exercise in comparison with rest time. The dependent variable, HeartDisease, aims to determine whether a patient has been diagnosed with the disease (1) or not (0). The dataset made available a rich coverage of features suitable for model training and development in which major health parameters were incorporated in predicting heart disease.



Age	Sex	ChestPainType	RestingBP	Cholesterol	FastingBS	RestingECG	MaxHR	ExerciseAngina	Oldpeak	ST_Slope	HeartDisease
40	M	ATA	140	289	0	Normal	172	N	0.0	Up	0
49	F	NAP	160	180	0	Normal	156	N	1.0	Flat	1
37	M	ATA	130	283	0	ST	98	N	0.0	Up	0
48	F	ASY	138	214	0	Normal	108	Y	1.5	Flat	1
54	M	NAP	150	195	0	Normal	122	N	0.0	Up	0

Analysis technique

In this work, we used mainly two machine learning techniques – decision tree and neural network models. **Decision Tree**: In this case, various levels of decision trees were constructed to analyze how the complexity level in the model affects the accuracy. With the use of the max_depth parameter, we defined the scope of bias and variance which in turn increased the accuracy of the model. The ability of the decision tree to classify patients with or without the heart disease was evaluated based on the precision, recall and the F1 score. **Neural Network**: The configuration of the feed forward neural network employed several hidden layers in order to determine how the performance of the neural network is dependent on its depth. To predict heart disease, we used different numbers of hidden layers and units in each layer. Backpropagation was used for training the neural network while precision, recall, F1 score and similar metrics were used for evaluation.

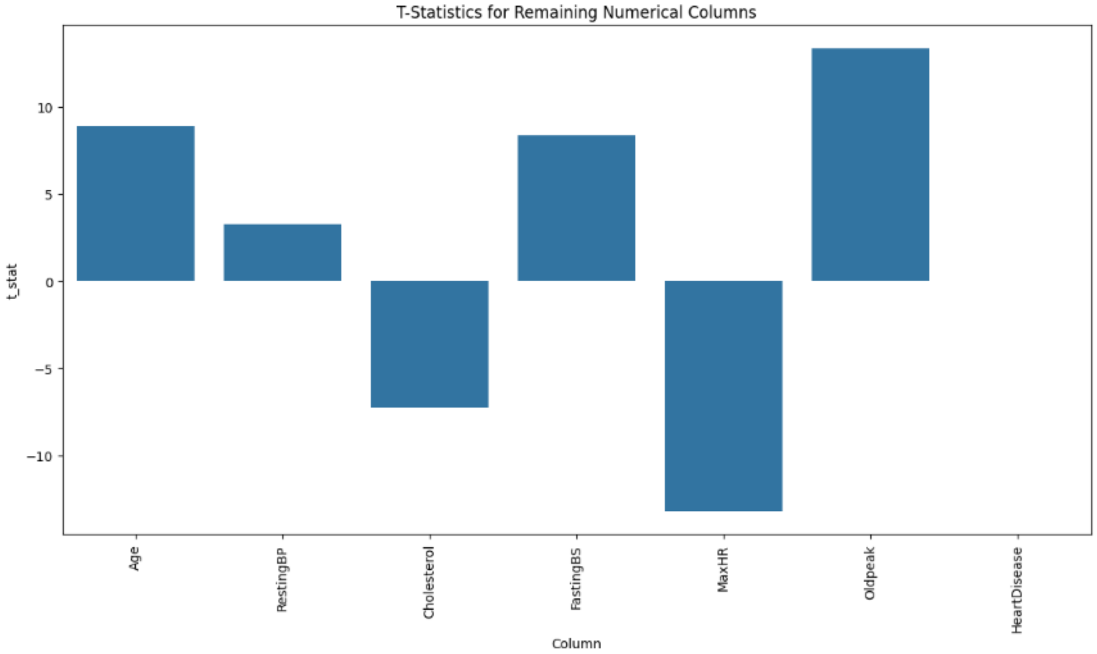
The dataset was divided into training and test sets in an 80-20 ratio. Both models were trained and tested on the training and test set respectively. There was also scaling of features in order to prepare the data for the neural network model to ensure that all features had similar scales and so enhanced convergence during training.

Results

T-Test Analysis of Feature Importance

The first analysis involved plotting T-statistics for each numerical feature in relation to the target variable 'HeartDisease'. The T-statistic values were calculated to assess the significance of each feature in distinguishing between patients with and without heart disease. As shown in the plot, 'Age', 'MaxHR', and 'Oldpeak' had notably high T-statistic values, indicating their strong influence in predicting heart disease. Conversely, features like 'Cholesterol' had lower T-statistic values, suggesting a weaker relationship with the target variable.

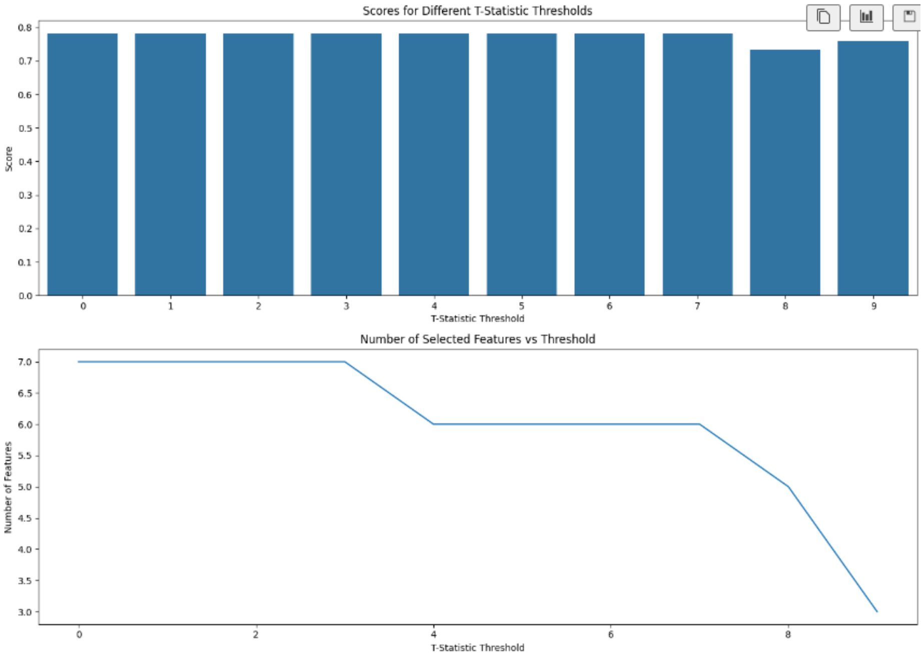
This feature importance analysis helped guide the feature selection process for building the models, emphasizing the importance of including highly correlated features like 'Age', 'MaxHR', and 'Oldpeak' to improve predictive accuracy.



Threshold Analysis foFeature Selection

Also, we looped through different T-statistic thresholds to identify subsets of features that met the specified threshold. For each iteration, we used the selected features to train decision tree models and tested a variety of max_depth values to determine the optimal model parameters. The scores for each T-statistic threshold were recorded to assess model performance.

The results, as depicted in the plot, indicate that using all features (threshold of 0) produced the highest F1 score. This suggesting that including more features generally improved the model's predictive power. However, as the threshold increased and fewer features were used, the model's performance began to decline. This trend highlights the importance of maintaining an appropriate balance between feature selection and model complexity in order to ensure that significant features are retained to maximize predictive accuracy.

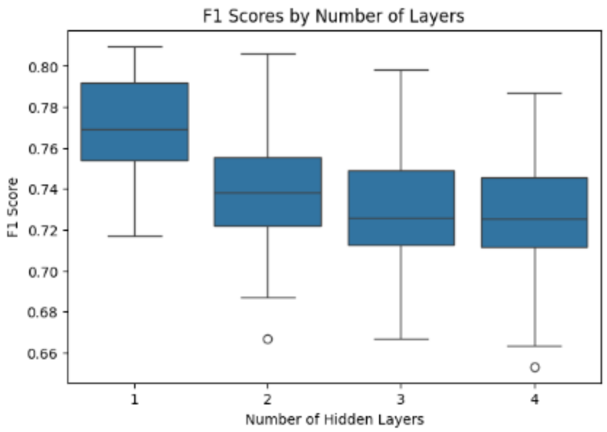


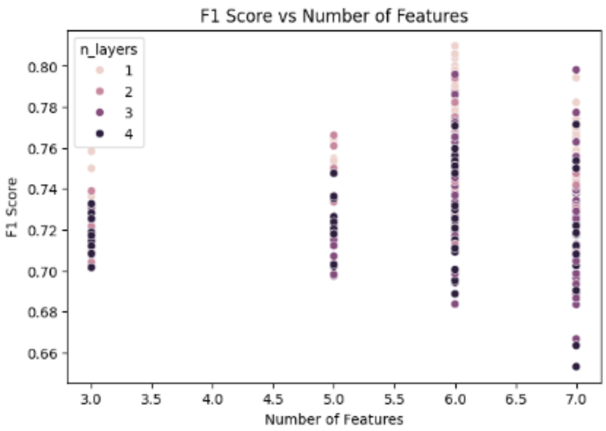
The best results were obtained with a threshold of 0, using all seven features: 'Age', 'RestingBP', 'Cholesterol', 'FastingBS', 'MaxHR', 'Oldpeak', and 'HeartDisease'. The optimal decision tree configuration had a max_depth of 3, resulting in an F1 score of 0.78. These findings demonstrate that retaining more features and an appropriate tree depth can significantly enhance model performance.

Neural Network Performance Analysis

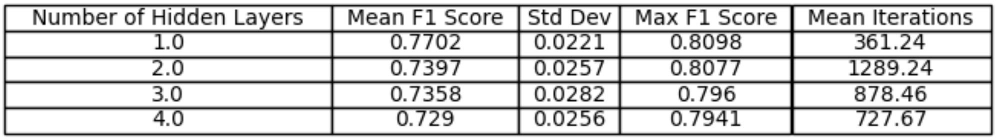
The neural network analysis involved testing different configurations of hidden layers to determine the impact on model performance. As shown in the boxplot, the model with one hidden layer achieved the highest mean F1 score of 0.7702, while increasing the number of layers did not consistently improve performance and sometimes led to lower F1 scores. This suggests that simpler architectures may be more effective for this dataset.

The best configuration for the neural network was obtained using a T-test threshold of 4, resulting in the use of six features: 'Age', 'Cholesterol', 'FastingBS', 'MaxHR', 'Oldpeak', and 'HeartDisease'. The optimal hidden layer configuration consisted of 49 units, and this model achieved an F1 score of 0.81, outperforming the decision tree model. These results indicate that the neural network, with careful selection of features and architecture, can provide superior predictive accuracy for heart disease classification.





Below is a summary table showing the F1 score statistics for different numbers of hidden layers in the neural network model:



Number of Hidden Layers	Mean F1 Score	Std Dev	Max F1 Score	Mean Iterations
1.0	0.7702	0.0221	0.8098	361.24
2.0	0.7397	0.0257	0.8077	1289.24
3.0	0.7358	0.0282	0.796	878.46
4.0	0.729	0.0256	0.7941	727.67

This summary highlights that while a single hidden layer performed the best, increasing the number of layers did not consistently improve the results, reinforcing the notion that simpler architectures were more effective for this dataset.

These results show the application and potential of machine learning models in assisting healthcare practitioners with accurate and early heart disease diagnoses. This will support better decision-making and patient outcomes.

Technical

Data Preparation: Our dataset required formatting and cleaning including handling missing values, normalizing features, and encoding categorical variables. These steps ensured that the data was suitable for analysis and model training.

Analysis: The decision tree and neural network models were chosen as they are well-suited to binary classification problems. The decision tree provided interpretable insights and while the neural network allowed for deeper relationships between features to be captured.

Analysis Process: The analysis involved iterative adjustments to model parameters including testing different tree depths and neural network architectures. Several configurations were tested, and model performance was evaluated using precision, recall, and F1 score to identify the best approach.