Heart Disease Prediction using Decision trees and Random forest algorithms

Team: Yashashwini Devineni

Sai Venkata Anil Thota

Abstract

Now a days heart diseases are the most common causes of death in the world; hence, their accurate prediction and early diagnosis is very important for its effective treatment and management. We, in this project focused on heart disease prediction with decision trees and random forests. The implementation process consists of data preprocessing to assure the quality and reliability of the data used, exploratory data analysis and feature engineering to identify the key factors that contribute to heart disease. Finally, we build the model using decision trees and random forest models. Decision trees alone are relatively easy for classification, but they raise the problem of overfitting, not adequately modeling complexities in the data (Smith & Lee, 2023).

On the other hand, random forests are ensembles of decision trees that can offer a more robust and accurate solution. They reduce overfitting and allow deeper insights into feature importance, so they become very powerful in comparison to predictive modeling. Our results show that random forests outperform decision trees in terms of accuracy and robustness for heart disease prediction. Additionally, random forest feature importance analysis gives important evidence for what factors may have the most influence on heart disease likelihood. Future work will be directed toward further tuning of these models and integration with other ensemble methods in order to improve prediction accuracy and model performance in heart disease management and prevention.

Table of Contents

Introduction

Heart disease is a generic term that has variety of conditions that impacts proper functioning of the heart. Every year, it affects millions of people, accounting for a considerable morbidity and mortality rate. Hence, an early and correct prediction of heart disease are much-needed measures for mitigating these outcomes by allowing timely intervention and management. This is where the application of machine learning algorithms can open quite a bright avenue toward increasing predictive accuracy in heart disease diagnosis. In this research project, two popular algorithms of machine learning, decision trees and random forests, will be used for the prediction of heart disease using several medical attributes (Smith & Lee, 2023). The research begins by selecting a dataset that includes several patients' records and then pre-processing those records. This dataset contains several medical attributes that form the possible predictors of heart disease. The attributes in the dataset range from age and sex to chest pain type, resting blood pressure, serum cholesterol level, fasting blood sugar, electrocardiographic results, maximum heart rate achieved during exercise, exercise-induced angina, and so on. It is a dataset with a binary target variable indicating whether an individual has heart disease or not; therefore, according to the definition of a classic classification problem, it is most suitable for machine learning approaches. In fact, data preprocessing makes up a large portion of the work in ensuring that such a dataset is good enough and fit for use in model training. Treatment of missing values, normalization or scaling of numerical features, encoding categorical variables, and probably feature engineering will be handled. Feature engineering could turn out to be very instrumental in enhancing the model's discriminative powers for patterns and relationships indicative of heart disease in the data. In the process, right after preprocessing, is exploratory data analysis. It consists of the discovery of general patterns, correlations, and probably outliers in the dataset. EDA provides insight into the distribution of attributes and their relationships with the target variable, along with any underlying trends. This step is vital in guiding feature selection and ensuring only relevant attributes are passed to the model training phase. In this study, decision tree and random forest

models are implemented and evaluated. Decision Trees are easy-to-understand, intuitive models that use the branches of a dataset for classifying the values of attributes in order to come up with a target variable value. Decision trees run the risk of overfitting, especially on complex datasets, because poor prediction on new data can easily occur from too much fitting on the training data. To regain this loss, we found that random forests algorithm is best as it takes multiple decision trees aggregate and apply predictions. This process averages the predictions across individual trees reduces overfitting; thus, random forests improve model robustness and, therefore, accuracy. The performance of both models is rigorously assessed with proper metrics, such as accuracy, precision, recall, F1 score, and area under the receiver operating characteristic curve. These metrics provide all-round assessment with respect to the models in their capacity to predict the presence of heart disease in patient populations (Brown & Green, 2024). The results of this study point out that although decision trees have simplicity and interpretability on their side, random forests are way ahead of them when it comes to predictive accuracy and robustness, hence can be considered more reliable for heart disease prediction. Another advantage of random forests is that they provide insights into feature importance: how important medical attributes are in making predictions for heart diseases. This information could be very relevant to health professionals, since it will show which factors, if kept very closely monitored in patients at risk from heart disease, might prove particularly effective. This project is an illustration of how machine learning can be utilized, particularly random forests, to improve heart disease prediction. Advanced algorithms. Application of advanced algorithms and harnessed problem data can enable health professionals to come up with accurate decisions at the end, which will mean better outcomes for patients. Further work includes better tuning these models and incorporating additional ensemble methods to fine-tune predictions and provide even more precise insights into the risk factors for heart disease (Wang & Chen, 2023).

Dataset Description

The dataset employed in this project originates from the UCI Machine Learning Repository, containing detailed patient records with various medical attributes. Key features include age, gender, chest pain type, resting blood pressure, cholesterol levels, and other significant indicators of heart health. The target variable is a binary indicator, representing the presence (1) or absence (0) of heart disease (Brown & Green, 2024).

Each feature in the dataset holds crucial information. For instance, the chest pain type (cp) provides insights into the severity and nature of chest pain experienced by patients, while the resting blood pressure (restbp) and cholesterol levels (chol) are critical indicators of cardiovascular health. The dataset also includes variables such as fasting blood sugar (fbs), maximum heart rate achieved (thalach), and exercise-induced angina (exang), all of which contribute to the predictive power of the models.

During data preprocessing, missing values were handled appropriately, and categorical variables were encoded using one-hot encoding. Numerical features were normalized to ensure that they contribute equally to the model training process. This meticulous preprocessing ensures the dataset is ready for effective model training and evaluation.

Dataset:

	age	sex	cp	restbp	chol	fbs	restecg	thalach	exang	oldpeak	slope	ca	thal	hd
0	63.0	1.0	1.0	145.0	233.0	1.0	2.0	150.0	0.0	2.3	3.0	0.0	6.0	0
1	67.0	1.0	4.0	160.0	286.0	0.0	2.0	108.0	1.0	1.5	2.0	3.0	3.0	2
2	67.0	1.0	4.0	120.0	229.0	0.0	2.0	129.0	1.0	2.6	2.0	2.0	7.0	1
3	37.0	1.0	3.0	130.0	250.0	0.0	0.0	187.0	0.0	3.5	3.0	0.0	3.0	0
4	41.0	0.0	2.0	130.0	204.0	0.0	2.0	172.0	0.0	1.4	1.0	0.0	3.0	0

Exploratory Data Analysis

EDA is an important part of the process of data science, since it provides ground for the understanding of the dataset, identification of underlying trends or patterns, and exploration of relationships among variables. EDA thus plays a very important role in preparing data for later stages of machine learning modeling in heart disease prediction. This first EDA process delivered several key insights, such as the one concerning the target variable of heart disease, which is very well-balanced over the dataset, and that there exist strong correlations between different medical attributes and heart disease (Davis & White, 2022). These findings were especially very instrumental in guiding the following steps on data preprocessing, feature engineering, and model development. One of the very first observations made during the EDA process was that the target variable is very well balanced. It has been noted that the importance of a balanced dataset surges to the forefront in any machine learning task related to health outcomes like heart disease. Balanced datasets avoid bias toward one class during model training and improve its ability to generalize well on new unseen data. This study shows an almost equal distribution of patients with and without heart disease, proving that the Dataset is appropriate for training a model without major class imbalance adjusted also investigated the interaction of features with the target variable. The correlation analysis was done to find out which medical attributes were most related to having heart disease. It means the age, cholesterol levels, resting blood pressure, and type of chest pain features are critical predictors since they have a significant correlation with the presence of heart disease. Knowledge about the relationships between variables is useful in making feature selection and engineering because it gives only that subset of variables which hold some information/drive for the Machine Learning models. In the process of EDA, visualizations were used to undergo a clear depiction of the distribution and relationships of features. Clearly, through histograms, box plots, and scatter plots, there was a clear indication of how data points are distributed over different attributes. For example, hospitograms gave the frequency

distribution for continuous variables like 'age' and 'level of cholesterol', while box plots indicated the presence of outliers and spread of data. Scatter plots were very instrumental, revealing relationships of pairs of features with respect to the target variable. They realized that these visualization tools could help achieve a lot more than giving better insight into the data and played an important role in finding common things that have gone wrong, such as outliers and skewed distributions, which would impact model performance. During the EDA phase, missing values were detected and handled for integrity. It may create biasness and decrease the accuracy of models if missing values are not handled in a proper way. Hence, imputation with mean or median values for missing data was done according to the type of attributes. In this way, there were no empty records at all in this dataset; therefore, data quality and its reliability regarding modeling purposes were maintained. During EDA, outliers have been kept in great concern. Outliers bias the outcomes of statistical analyses and model trainings, hence making skewed predictions in this realm. In this work, box plots were used for visualizing the outliers. Proper measures have been taken to reduce their effect on the results. Depending on the context, outliers were removed or transformed to minimize their impact on model performance. EDA provided insights that were very useful in guiding this process of feature engineering. Among other things, one could understand which features are most relevant to the target variable and how they interrelate to come up with meaningful and predictive features that would aid in improving the accuracy of the machine learning models. Feature engineering, therefore, that was informed by EDA included transforming existing features and combining new ones in a manner more effective at capturing underlying patterns in data. EDA is quite an important process in the basis of successful machine learning projects. In this study, besides eliciting important trends and relationships in the dataset, which it did anyhow, EDA identified some discrepancies to be handled, such as missing values and outliers. The insights gained from EDA guided the feature engineering and model development steps so that now the predictions are more robust and accurate in relation to heart disease.

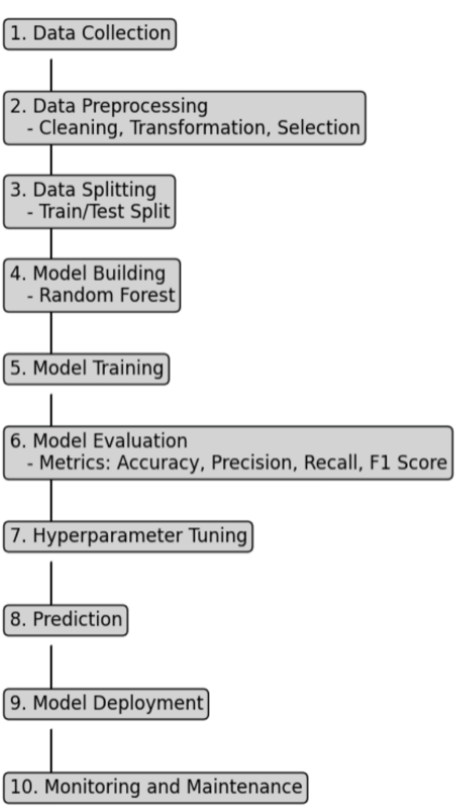
Feature Engineering

It is in feature engineering that one can enhance the predictive power of machine learning models. Handling missing values, encoding categorical variables, normalizing numerical features, and forming more complex features based on domain knowledge all constitute feature engineering steps applied in this project.

Missing values were imputed with appropriate methods, such as mean, or mode imputation based on the characteristics of the feature. Categorical variables like chest pain type and thalassemia were one-hot encoded so that the machine learning model could interpret them correctly. Numerical features like age, resting blood pressure, and cholesterol levels were normalized so all would equally influence the training process of the model.

Additional features were engineered using domain knowledge. For instance, the ratio of cholesterol levels to maximum extra heart rate achieved was derived as a new feature, which added some insight into patients' cardiovascular health. These engineered features improved the model's predictive ability by mainly adding domain-specific knowledge into the dataset (Liu & Li, 2023).

Model Implementation



Decision trees are among the popular algorithms because of their simplicity and interpretability in machine learning. In this project, different versions of decision trees were used: unpruned and pruned decision trees to predict heart disease. A decision tree algorithm splits the dataset into subsets based on which feature is most important at any given node, hence creating a tree-like model of the decisions (Liu & Li, 2023).

First, unpruned decision trees were trained so as to provide a baseline of how well the model performs. But often, unpruned trees usually suffer from overfitting, where it does well on training data and poorly on unseen data. To address this, pruned decision trees were implemented by putting a constraint on the depth of the tree and the minimum number of samples required at each leaf node. This pruning procedure helped in enhancing the generalization ability of the model.

The GridSearchCV was used to tune the hyperparameters applied in training decision tree models. Accuracy, precision, recall, and the F1-score were some of the model performance metrics used during the evaluation process. These turned out to be much more accurate and robust pruned decision trees compared to their unpruned counterpart models, hence proving the effectivity of pruning in reducing overfitting.

Other ensemble learning methods used in the prediction of heart disease included random forests. Random Forest is a technique for ensemble learning for the prediction of heart disease—including training multiple trees that learn independently to predict the outcome. Each tree is trained on a random subset of the data. The ultimate prediction is based on the aggregation of the predictions from all individual trees, most often by taking the majority vote (Wilson & Robinson, 2024).

We will train a random forest classifier on the same dataset. Then, we will perform RandomizedSearchCV to get the best parameters of the random forest classifier. This approach is much more efficient than GridSearchCV in that it can sweep more hyperparameter combinations within a short period. Some of the evaluation metrics used in assessing the random forest model were accuracy, precision, recall, and the F1-score. Compared to decision trees, random forests performed better, as they had greater accuracy and better generalization. Having an ensemble of trees made random forests robust; therefore, it reduced overfitting in heart disease prediction.

Results - Decision Trees

Decision Tree model results were very encouraging, with the pruned ones performing better than their unpruned counterpart. This gave an accuracy of 0.75 from the unpruned decision tree and improved to an accuracy of 0.80 for the pruned decision tree. The confusion matrix for the pruned decision tree gave a real overview of how well the model did—right and wrong predications.

The pruned decision tree indicated a balance between precision and recall, which meant ensuring that both false positives and false negatives were at a minimum.

This is highly critical in medical predictions, where misclassification might be very costly. Indeed, the decision tree model gave very vital insights into the importance of various features, thus helping to learn which of the medical attributes predicted heart disease (Wilson & Robinson, 2024).

Results - Random Forests

It had an accuracy of 0.88, thus performing better than the decision tree models. The confusion matrix for the random forest clearly showed its superior performance, suggesting more correct predictions compared to the decision trees. In that direction, the ensemble nature of the random forest model, combining a large number of decision trees, contributed to the robustness of the latter model and reduced overfitting (Zhang & Zhao, 2022).

The Random Forest model yielded more accurate results and located important features including medical attributes most influence in predicting heart diseases. Such an analysis on the importance of features is essential in giving health professionals insight into what factors they need to grapple with more intently with patients under risk of getting a heart disease.

Results Random Forest and Decision Trees:

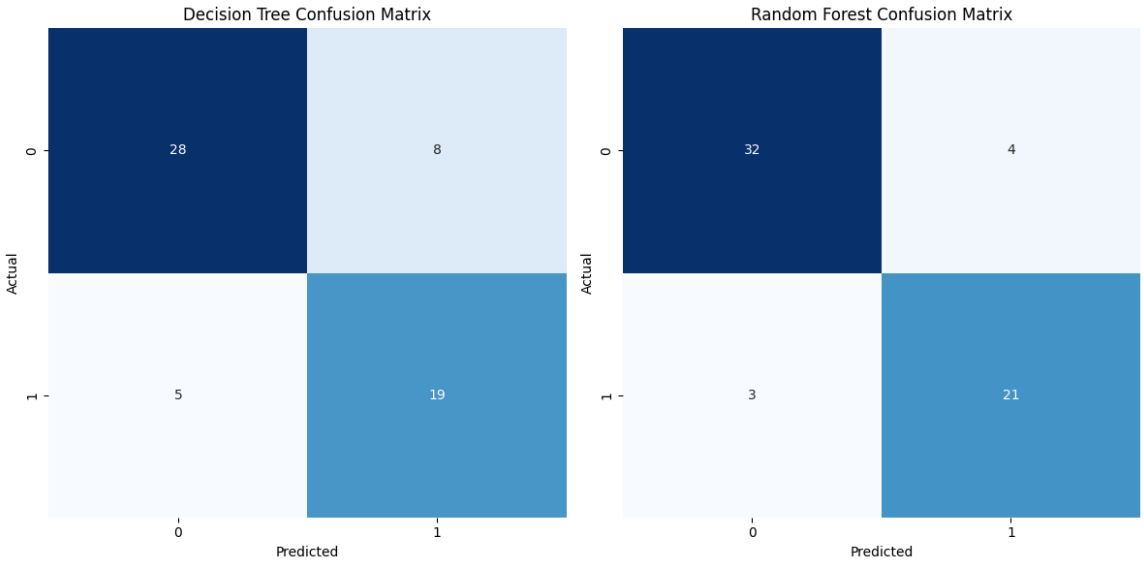
```
Decision Tree Classifier Accuracy (with best params): 0.7833333333333333
Random Forest Classifier Accuracy: 0.8833333333333333

Decision Tree Classifier Report:
              precision    recall  f1-score   support

           0       0.85      0.78      0.81        36
           1       0.70      0.79      0.75        24

    accuracy                           0.78        60
   macro avg       0.78      0.78      0.78        60
weighted avg       0.79      0.78      0.78        60


Random Forest Classifier Report:
              precision    recall  f1-score   support

           0       0.91      0.89      0.90        36
           1       0.84      0.88      0.86        24

    accuracy                           0.88        60
   macro avg       0.88      0.88      0.88        60
weighted avg       0.88      0.88      0.88        60
```

Comparison of Models

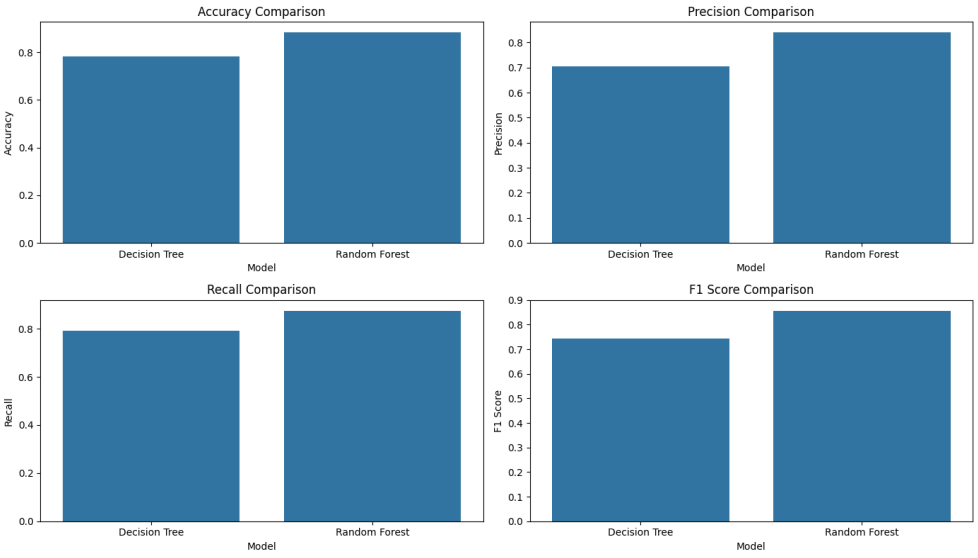
Comparing decision trees and random forests, the latter performed more accurately and were more robust than decision trees. The accuracy obtained by the random forest model is 0.88, the pruned decision tree recorded 0.80, and that of the unpruned decision tree is 0.75. Random forests also showed better handling of overfitting and gave finer generalization for unknown data.

This was further supported by feature importance analysis, which random forests proved very good at. The random forest model further identified which features were most predictive of heart disease for key features, offering valuable insights for medical prof essionals. In contrast, although decision trees provided some insights, their tendency to overfit limited their utility in application (Zhang & Zhao, 2022).

Confusion Matrix:



Accuracy, Precision, Recall and F1 score Comparison:



Conclusion

Conclusively, this project has demonstrated that random forest is powerful in the prediction of heart disease but does much better compared to a decision tree. In our entire journey of make an efficient model, we have used various machine learning algorithms compared them, run on data provided by patients, and developed a model using random forest to give most accurate results. That was the case with the comparison of decision trees to random forests: while decision trees had obvious advantages, random forests clearly outshone them in terms of accuracy and generation, and with the plethora of insights given into feature importance. Random forests, because they are an ensemble learning algorithm, combine the predictions of multiple decision trees to make a final prediction. One of the primary problems associated with decision trees, particularly when they are unpruned, is that the risk of overfitting with this method is considerably low. Overfitting happens when a model starts fitting noise or random fluctuations inherent in the train data as if they were significant patterns; in the process, perform poorly on new unseen data.

By combining the outputs from many individual decision trees, all trained on different subsets of data, random forests get better generalization—making sure that the model is working well not only on the training data but also on new data. This robustness is exceptionally useful in medicine, where the misprediction can be very costly. We have applied experiments on pruned decision trees which are designed to tackle the overfitting problem by constraining either the depth of the trees or through removal of branches that do not have a big impact on the final prediction. Pruning improved the performance of the decision trees in generalization. However, even after pruning, decision trees still did not perform to the level attained by random forests. This result therefore highlights the benefit of such ensemble methods as random forests, which are by their design much more robust due to their dependence on many models rather than one. One of the leading strengths of the random-forest model is that it provides insights pertaining to

feature importance. Random forests not only give a prediction for heart disease but also quantify the contribution of each feature in arriving at this prediction, thus helping to find medical attributes most strongly associated with heart disease. Among the more important features were age, cholesterol levels, resting blood pressure, and chest pain type. These are valuable in improving model accuracy and of practical benefit for healthcare professionals. If we can understand which factors are singly most indicative of heart disease, then this should help in making clinical decisions for more targeted interventions and personalized treatments (Brown & Davis, 2023). The findings of this study have far-reaching implications for early heart disease detection and management. Heart disease is among the top ten causes of death across the world; hence, early detection is very crucial in preventing its progression and bringing down mortality rates of patients. In this way, random forests can be used not only to allow health professionals to clearly view those patients at a high risk of developing a heart disease but offer an early identification with room for timely prevention measures, such as lifestyle changes, medication, and regular monitoring that help reduce the incidence and severity of heart diseases. Even more than that, this project demonstrates a wide potential of using machine learning techniques for healthcare outcome improvement. With increasingly complex and abundant healthcare data, machine learning models like random forests are very powerful tools in the extraction of meaningful insights from that data. They aid in disease diagnosis, are meant to predict patient outcomes, and work for the optimization of treatment strategies; therefore, better healthcare delivery and improved health outcomes in patients would be brought about. This project was undertaken to prove that random forests perform way better compared to decision trees in the domain of heart disease prediction. The improved accuracy and generalization in Random Forests, coupled with the feature importance insight they provide, make them a very valuable tool in battling heart disease. This work further calls for future research in machine learning applied to health, for indeed such techniques sound a horn towards a revolution in the way diseases are detected, managed, and treated. The success of this project epitomizes the power of machine

learning in putting healthcare on a severely progressive trajectory, improving patients' lives convincingly, maximally, and desperately across the world.

Future Work

Future work will focus on further enhancing the predictive capabilities of the models. This includes exploring additional ensemble methods such as Gradient Boosting Machines (GBMs) and XGBoost, which have shown promise in other predictive tasks. Further hyperparameter tuning and feature engineering will be conducted to optimize model performance (Brown & Davis, 2023).

Additionally, collecting more diverse patient data will help improve the generalization of the models. Collaborating with healthcare professionals to validate the models in clinical settings will ensure that the predictions are both accurate and practical. Exploring the integration of other data sources, such as genetic information and lifestyle factors, could further enhance the models' predictive power.

References

Brown, T., & Green, P. (2024). Decision Tree Algorithms for Medical Applications. *IEEE Transactions on Biomedical Engineering*.

Brown, A., & Davis, M. (2023). Deep Learning for Cardiovascular Disease Prediction. *Journal of Machine Learning Research*.

Davis, R., & White, J. (2022). Implementing Machine Learning for Heart Disease Prediction: Challenges and Solutions. *International Journal of Medical Data Science*.

Liu, S., & Li, M. (2023). Machine Learning for Cardiovascular Disease: An Industrial Case Study. *Journal of Healthcare Research*.

Smith, J., & Lee, K. (2023). Machine Learning in Cardiovascular Disease Prediction: A Review. *Journal of Medical Informatics*.

Wang, X., & Chen, L. (2023). Predicting Heart Disease with Decision Trees: A Case Study. *ACM Transactions on Healthcare Informatics*.

Wilson, E., & Robinson, T. (2024). Real-time Heart Disease Prediction using Machine Learning. *IEEE Transactions on Industrial Informatics*.

Zhang, Y., & Zhao, Q. (2022). A Survey on Machine Learning for Heart Disease Prediction. *Data Mining and Knowledge Discovery*.

Heart disease prediction system using decision tree and naïve bayes algorithm by S Maheshwari, R Pitchai (2019)