Heart Disease Prediction Using Decision Trees

GITHUB Repository Link:

https://github.com/Rudransh911/Heart-Disease-Prediction

Table of Contents

1. Introduction

"The World Health Organization estimates that there are 17.9 million deaths worldwide due to heart disease each year". Knowing the timing of when to detect heart disease is important to reduce the burden on public health. Early diagnosis offers the chance to intervene and it is important as it can make a big difference to patient outcome. Unfortunately, however, there is still some way to go in being able to diagnose heart disease based on the complex nature of risk factors and symptoms and the illness can often be diagnosed too late.

In this study, it will try to predict "heart disease using machine learning techniques, especially the Decision Trees. By classifying the patients into two classes: patients having some heart disease and patients not having any heart disease. For this type of prediction, the chosen dataset, the UCI Heart Disease dataset, has effective features such as age, cholesterol levels, blood pressure", etc (kaggle, 2025).

2. Literature Review

2.1 Overview of Heart Disease Prediction:

"As a result, heart disease prediction is an important task in healthcare as it is extremely common and has a considerable negative impact on public health. Cardiovascular diseases (CVDs)" are the leading disease in the world, the cause of over 17 million deaths every year, "as stated by the World Health Organization (WHO)" (Amini *et al.,* 2021). Therefore, "early detection and diagnosis are important for improving patient outcomes, as timely interventions can decrease the risk of heart attacks, and strokes and also prevent other serious complications.

However, in the past heart disease prediction would rely on clinical expertise in which doctors were making a decision based on a patient's risk factors (Romiti *et al.,* 2020) such as cholesterol levels, blood pressure, smoking habits, and family history, manually". In contrast to that, large datasets with high dimensional features, and patterns hidden from the human view are handled by machine learning models that can identify. Thus, these models have proven to be promising for heart disease prediction with more robust, accurate, and automated solutions.

2.2 Machine Learning in Healthcare:

Heart disease prediction is a necessary task in healthcare as it is prevalent throughout a great deal of society and affects the lives of so many. “Cardiovascular diseases (CVDs) are responsible for killing more people all around the world than any other cause, namely more than 17 million every year (Robert and Al Dawish, 2021), as per the World Health Organization (WHO). Getting early detection and diagnosis is important because if it can time the interventions with that, that can reduce a person’s chance of having a heart attack, a stroke, or any of the other really serious complications”.

“Heart disease prediction historically relied on clinicians' expertise which included reviewing risk factors such as cholesterol levels, blood pressure, smoking habits, and family history” (Duval *et al.,* 2020). But these took the form of subjective and error-prone approaches. However, with the combination of medical imaging and statistical models along with machine learning, heart disease prediction has moved from being less accurate, and less efficient with less of a data-driven approach to more accurate, more efficient, and more data-driven.

2.3 Decision Trees:

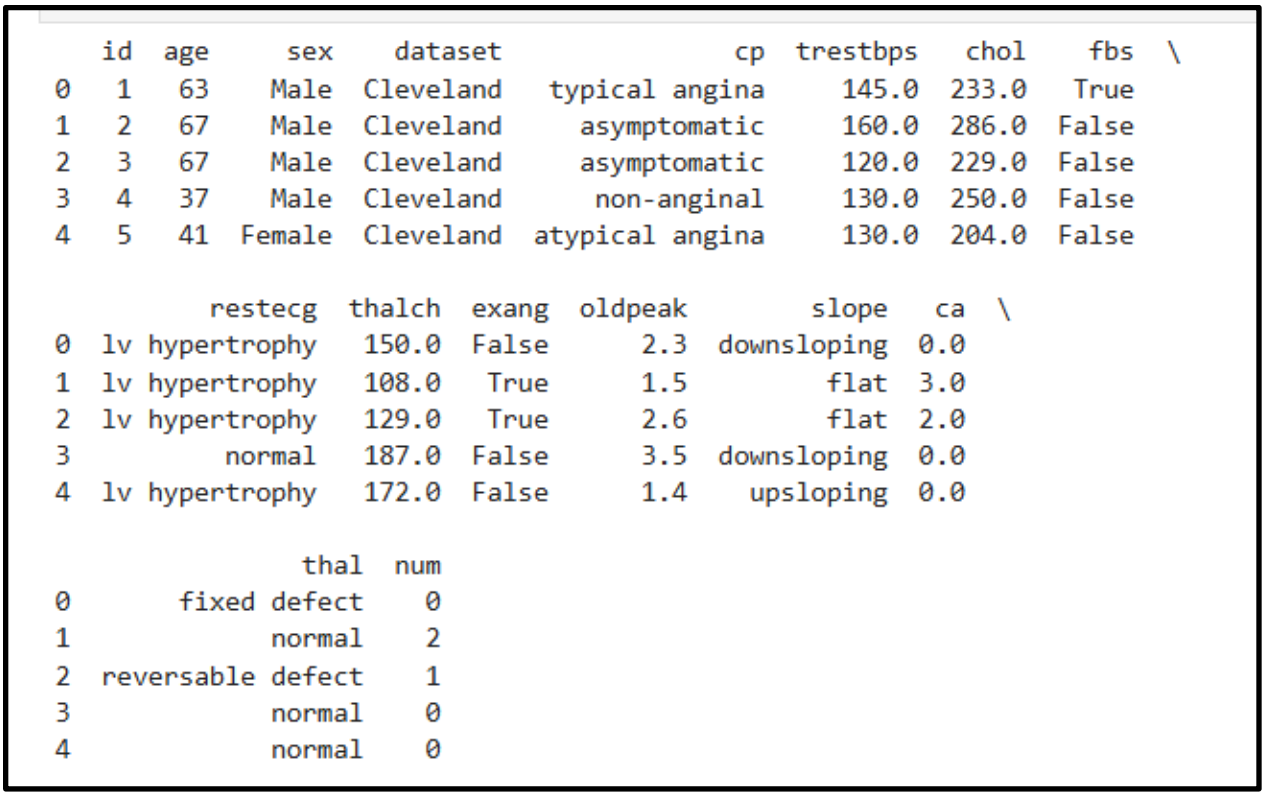
“Decision Trees (DT) are one of the most popular machine learning techniques which are simple and interpretable” (Izza *et al.,* 2020). Given a dataset, a decision tree divides the dataset into subsets on the basis of feature value to form a structure of the tree looking like a tree, where each node represents a decision made using a feature and each branch leads to a leaf node that identifies a class tag: for example, "heart disease" or "no heart disease". This method is recursive and continues until the tree reaches a certain depth or some other condition.

The advantage of a Decision Tree is that it can interpret the model and by visualizing it, even those who are non-experts will be able to. The data it uses can range from numerical to categorical without any need for extensive preprocessing. Decision trees can also capture non-linear relationships of features and hence these are versatile for many different datasets (Zhou *et al.,* 2024). These are however not without their limitations. Their strong drawback is over fitting when allowed to grow deep without pruning, and when it’s very deep, it is more prone to overfit the training data. Such a model leads to a model that generalizes over noise and outliers, and its

generalization ability is hurt. “Additionally, deep trees can become very complicated and computationally high, making decoding and evaluating the trees rather slow.

3. Methodology

3.1 Data Collection:



	id	age	sex	dataset	cp	trestbps	chol	fbs	\
0	1	63	Male	Cleveland	typical angina	145.0	233.0	True	
1	2	67	Male	Cleveland	asymptomatic	160.0	286.0	False	
2	3	67	Male	Cleveland	asymptomatic	120.0	229.0	False	
3	4	37	Male	Cleveland	non-anginal	130.0	250.0	False	
4	5	41	Female	Cleveland	atypical angina	130.0	204.0	False	

	restecg	thalch	exang	oldpeak	slope	ca	\
0	lv hypertrophy	150.0	False	2.3	downsloping	0.0	
1	lv hypertrophy	108.0	True	1.5	flat	3.0	
2	lv hypertrophy	129.0	True	2.6	flat	2.0	
3	normal	187.0	False	3.5	downsloping	0.0	
4	lv hypertrophy	172.0	False	1.4	upsloping	0.0	

	thal	num
0	fixed defect	0
1	normal	2
2	reversable defect	1
3	normal	0
4	normal	0

Figure 1: Display the first few rows of the dataset

The raw dataset is taken from Kaggle, and this figure shows the first few rows of it, the rows contain details like age, sex, blood pressure, cholesterol chest pain, etc. These are important inputs to the model in predicting the chances of having heart disease”.

3.2 Data Preprocessing:

Preprocessed data:

	id	age	sex	dataset	cp	trestbps	chol	fbs	restecg
0	-1.730169	1.007386	1	0	3	0.698041	0.311021	True	1.0
1	-1.726404	1.432034	1	0	0	1.511761	0.797713	False	1.0
2	-1.722639	1.432034	1	0	0	-0.658158	0.274289	False	1.0
3	-1.718873	-1.752828	1	0	2	-0.115679	0.467130	False	0.0
4	-1.715108	-1.328180	0	0	1	-0.115679	0.044717	False	1.0

	thalch	exang	oldpeak	slope	ca	thal	num
0	0.495698	False	1.349421	0	-1.249371	0	-0.871794
1	-1.175955	True	0.589832	1	4.292099	1	0.879408
2	-0.340128	True	1.634267	1	2.444942	2	0.003807
3	1.968345	False	2.488805	0	-1.249371	1	-0.871794
4	1.371326	False	0.494884	2	-1.249371	1	-0.871794

Figure 2: Preprocessed data

This is a picture of the preprocessed dataset after the variables have been encoded and scaled, categorical variance is represented as sex, resting, etc. In order to be fed to machine learning algorithms, data generally needs to go through this transformation.

3.3 Exploratory Data Analysis:

```
Dataset Info:

<class 'pandas.core.frame.DataFrame'>
RangeIndex: 920 entries, 0 to 919
Data columns (total 16 columns):
 #   Column    Non-Null Count  Dtype
---  ------    --------------  -----
 0   id        920 non-null    float64
 1   age       920 non-null    float64
 2   sex       920 non-null    int32
 3   dataset   920 non-null    int32
 4   cp        920 non-null    int32
 5   trestbps  920 non-null    float64
 6   chol      920 non-null    float64
 7   fbs       920 non-null    object
 8   restecg   741 non-null    float64
 9   thalch    920 non-null    float64
 10  exang     920 non-null    object
 11  oldpeak   920 non-null    float64
 12  slope     920 non-null    int32
 13  ca        920 non-null    float64
 14  thal      920 non-null    int32
 15  num       920 non-null    float64
dtypes: float64(9), int32(5), object(2)
memory usage: 97.2+ KB
None
```

Figure 3: Dataset Info

In this figure we've got a brief summary of the dataset for which it has 920 entries and 16 features. It forwards each column's data type and mentions that some of the columns (resting) contain missing values. It enables us to understand better what information is in the dataset, as well as the quality of this dataset for modeling.

Statistical Summary:

	id	age	sex	dataset	cp	\
count	920.000000	9.200000e+02	920.000000	920.000000	920.000000	
mean	0.000000	6.178632e-17	0.789130	1.238043	0.782609	
std	1.000544	1.000544e+00	0.408148	1.130673	0.956350	
min	-1.730169	-2.708286e+00	0.000000	0.000000	0.000000	
25%	-0.865085	-6.912073e-01	1.000000	0.000000	0.000000	
50%	0.000000	5.192709e-02	1.000000	1.000000	0.000000	
75%	0.865085	6.888994e-01	1.000000	2.000000	2.000000	
max	1.730169	2.493654e+00	1.000000	3.000000	3.000000	

	trestbps	chol	restecg	thalch	oldpeak
count	9.200000e+02	9.200000e+02	741.000000	9.200000e+02	9.200000e+02
mean	-6.873729e-16	-1.853590e-16	0.253711	2.162521e-16	-4.633974e-17
std	1.000544e+00	1.000544e+00	0.435428	1.000544e+00	1.000544e+00
min	-7.167915e+00	-1.828588e+00	0.000000	-3.086416e+00	-3.303061e+00
25%	-6.581583e-01	-1.963329e-01	0.000000	-6.983397e-01	-8.343970e-01
50%	-1.156786e-01	2.008263e-01	0.000000	1.808312e-02	-7.480801e-02
75%	4.268011e-01	6.232384e-01	1.000000	7.345059e-01	5.898324e-01
max	3.681680e+00	3.708683e+00	1.000000	2.565364e+00	5.052418e+00

	slope	ca	thal	num
count	920.000000	9.200000e+02	920.000000	9.200000e+02
mean	1.152174	-1.486733e-16	1.158696	-3.089316e-17
std	0.516007	1.000544e+00	0.483493	1.000544e+00
min	0.000000	-1.249371e+00	0.000000	-8.717944e-01
25%	1.000000	-2.050756e-16	1.000000	-8.717944e-01
50%	1.000000	-2.050756e-16	1.000000	3.806963e-03
75%	1.000000	-2.050756e-16	1.000000	8.794083e-01
max	2.000000	4.292099e+00	2.000000	2.630611e+00

Figure 4: Statistical Summary

Key statistics of each feature in the dataset from the statistical summary in this figure: “mean, standard deviation, min, 25th percentile, median, 75th percentile, and max”. In terms of feature distribution and spread, these metrics provide insight into how to preprocess further.

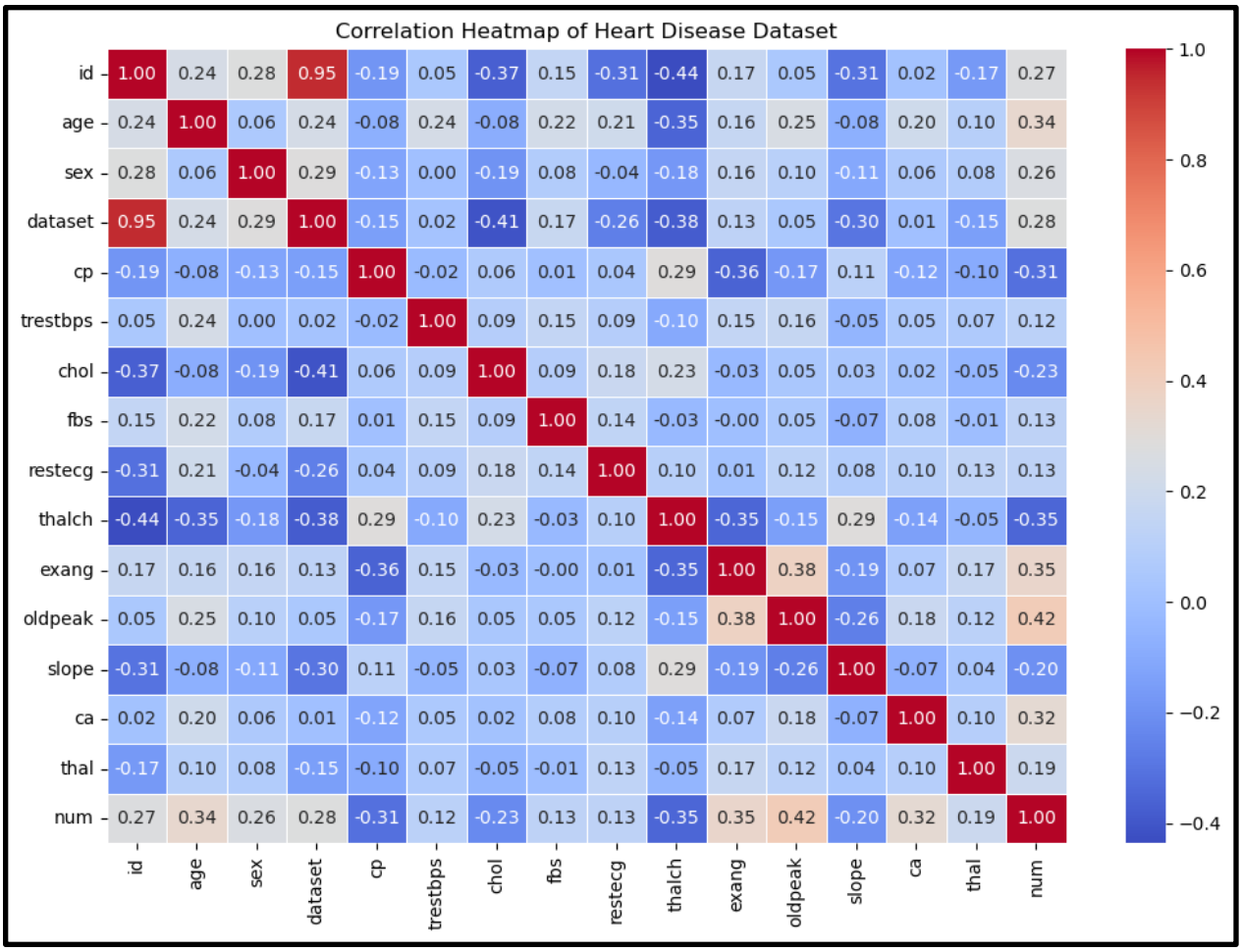


Figure 5: Correlation Heatmap

A correlation heatmap is a visualization of the relationship between the numerical features within the dataset. It also demonstrates the strong correlations corresponding to trestbps, and its weak to negative correlations in chol. These correlations must be identified otherwise one will be negatively selecting features and need to use many features (dividing window) or do feature engineering to reduce multicollinearity during model training.

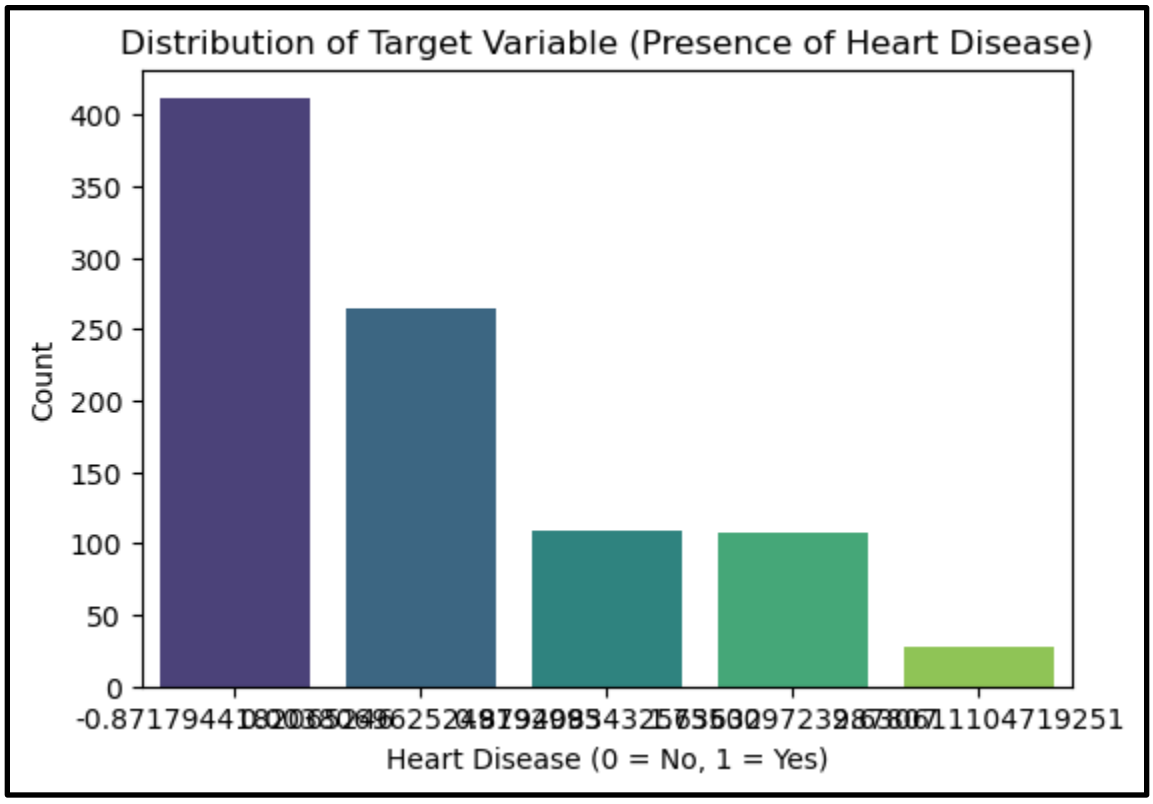


Figure 6: Count Plot of the Target Variable

"This figure shows the target variable distribution (heart disease diagnosis) as represented (or classified as) 0 (no heart disease) and 1 (heart disease). The count plot indicates class imbalance, with a significantly higher number of non-heart disease cases". This type of imbalance may require them to address it in order to produce good model predictions.

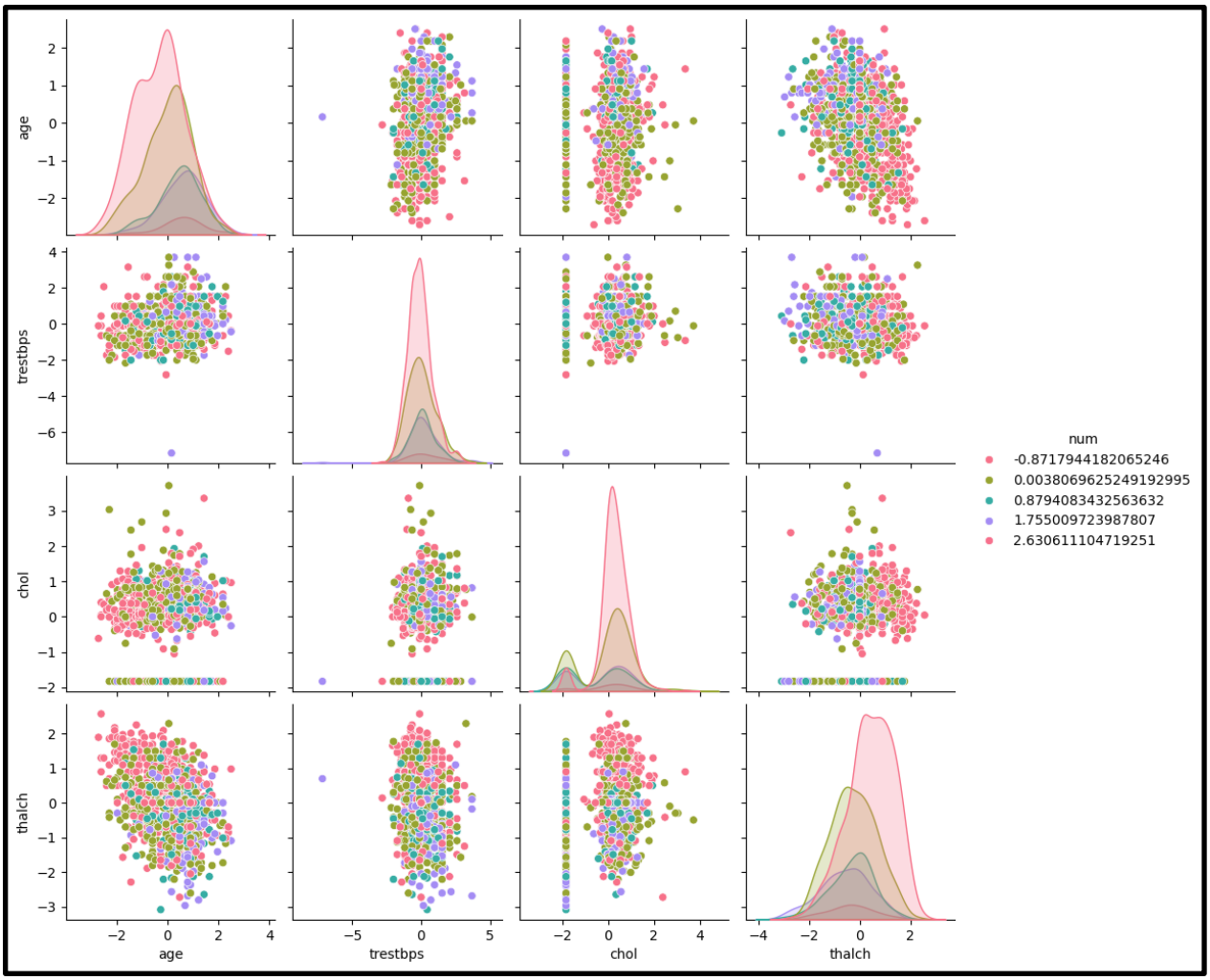


Figure 7: Pair Plot of Selected Features

A pair plot represents the relationships of some of the features, age, treetops, chol, and thatch, coloring the data based on the target variable. This helps to reveal possible patterns or clusters that may be present “between the features and the target to guide the model” and reveal how it will distinguish heart disease from nonheart disease cases.

3.4 Model Selection:

This task was easy to identify due to simplicity, interpretability, and versatility in dealing with numerical and categorical data and so Decision Trees were a natural fit. “Binary classification tasks like predicting the presence or absence of heart disease” are particularly suited to it as it gives clear decision rules that will help to separate the classes (presence/absence of heart

disease). Not only that, but Decision Trees can have non-linear relationships and can model complex decision boundaries thus naturally applying such healthcare apps.

3.5 Model Training:

```
Training data shape: (736, 15)
Testing data shape: (184, 15)
```

Figure 8: Splitting the data

This is the figure of the dataset after dividing it "into training, and testing sets. The size of the training set is 736 instances, and that of the testing set is 184 instances. This split guarantees that the model is trained on a big part of the data and evaluated on data it has not seen yet in order to achieve better generalization.

4. Results and Discussion

4.1 Model Performance:

```
Accuracy: 73.91%

Classification Report:
              precision    recall  f1-score   support

           0       0.78      0.88      0.82       129
           1       0.59      0.42      0.49        55

    accuracy                           0.74       184
   macro avg       0.68      0.65      0.66       184
weighted avg       0.72      0.74      0.72       184
```

Figure 9: Classification Report

This is the classification report of the Decision Tree model with precision, recall, and F1-score for two classes(no heart disease and heart disease). The model achieves an accuracy of 73.91% and a classification report provides some insight into what performance was achieved, particularly with a rather imbalanced precision and recall for class 1 (heart disease).

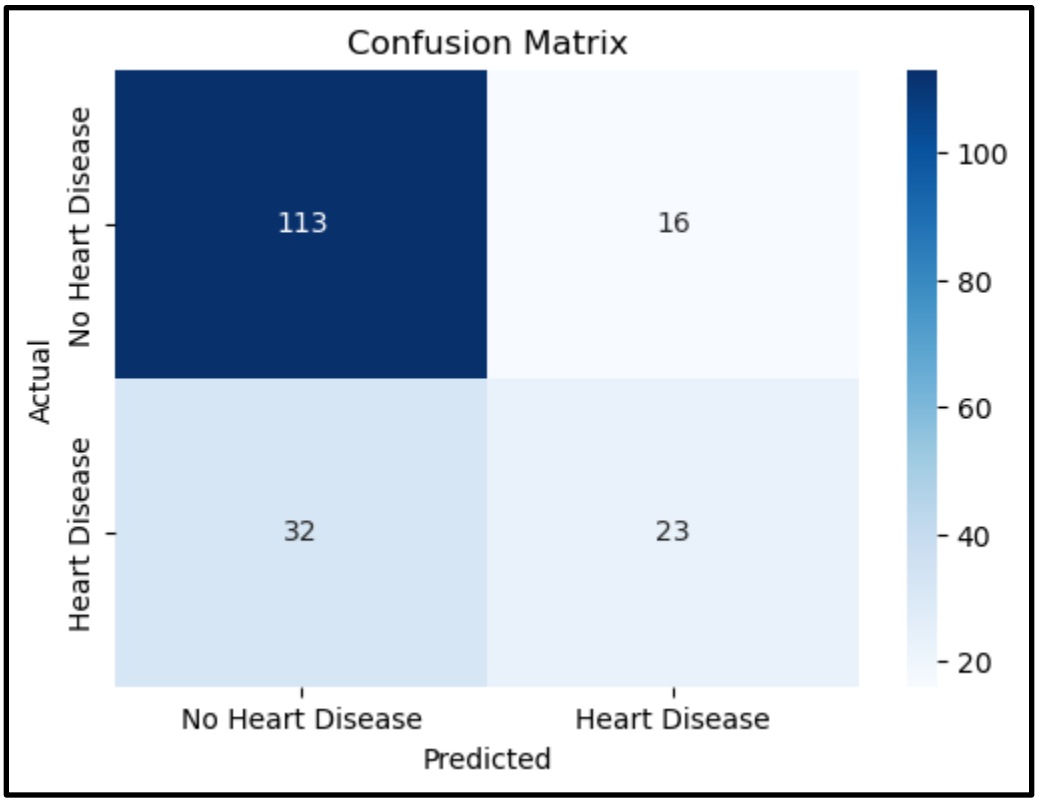


Figure 10: Confusion Matrix

The confusion matrix is a matrix containing true positive, true negative, false positive, and false negative values for the model's predictions. It serves to see how well the model does in separating the two classes: "No Heart Disease" and "Heart Disease". This matrix shows the ability of the model to predict the number of heart disease cases".

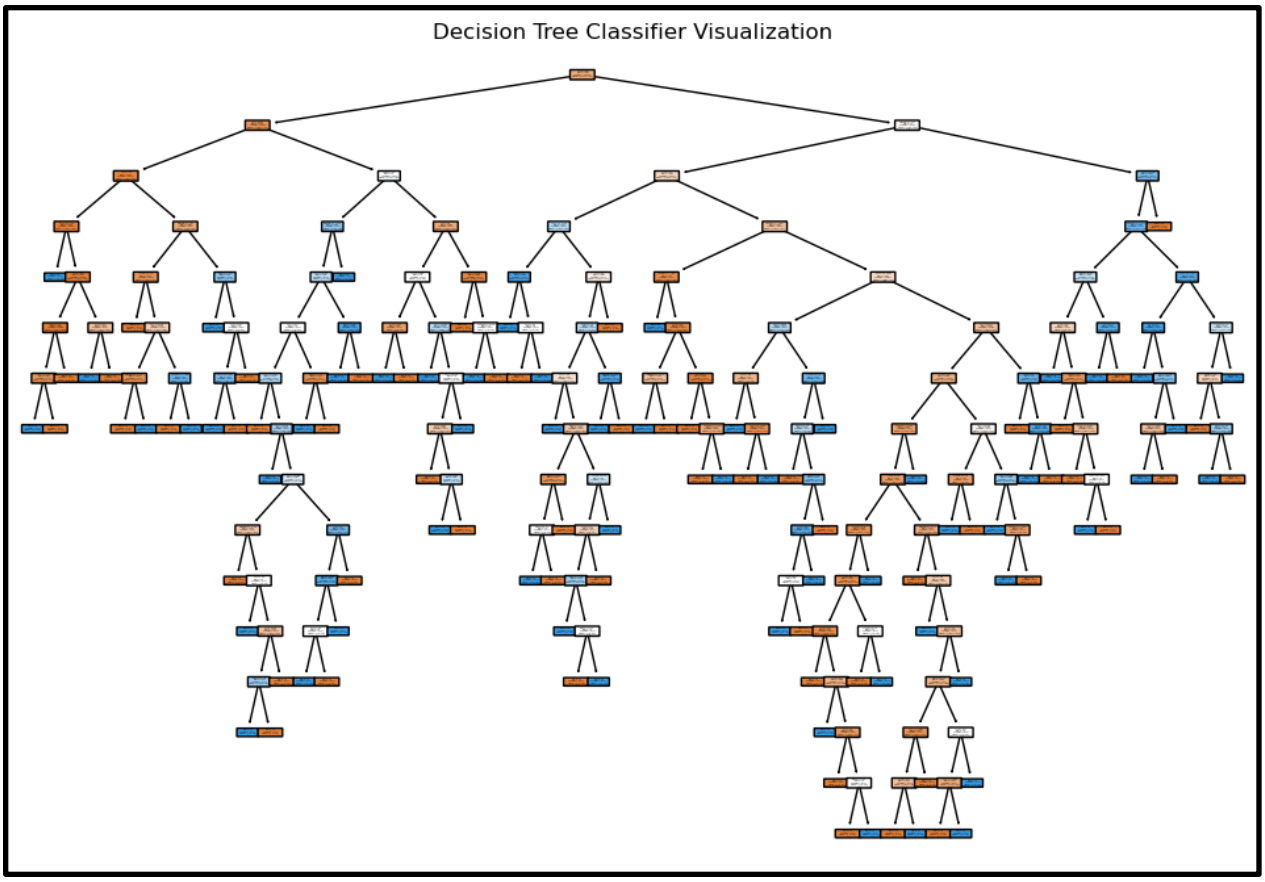


Figure 11: Decision Tree Classifier Visualization

The visualization of the trained Decision Tree Classifier is shown in this figure. The decision-making process is represented by this "tree structure in which each node splits following the feature values". The visual offers a straightforward way of understanding how the model reaches its final predictions and its decision boundary used for classification.

4.2 Model Interpretability:

An important reason Decision Trees are considered advantageous in such domains as healthcare is that are interpretable. Whereas complex models like Neural Networks or SVMs have complex, hard-to-understand rules, Decision Trees are simple, readable decision rules resulting in logic that can easily be understood by nonexperts (Chakraborty, 2024). At each node of the tree, the data is split according to the feature leading to the largest information gain in case of classification or the smallest variance in the case of regression. The decision rule that the model follows for each path from root to leaf is easy to follow and trace and represents each path.

The attributes used by the “Decision Tree which include age, cholesterol levels, blood pressure, and chest pain type are used to create decision paths (Zhou *et al.,* 2020). That leads either to "No Heart Disease" or "Heart Disease" classification in terms of the prediction of heart disease. For instance, the cholesterol threshold that the first node splits on may depend on the fact that patients with higher cholesterol are more likely to develop heart disease”. Across an increasing number of splits, nodes focus on different risk factors and further refine the prediction.

4.3 Discussion:

The Decision Tree model result shows good accuracy in predicting heart disease with an overall accuracy of 73.91% (Dissanayake and Md Johar, 2021). The Precision and Recall values of the classification report for both classes show that the model has a better recall of not having heart disease (88%) but a lower recall of having heart disease (42%). That means that although the model can identify patients who don’t have heart disease, it’s struggling to spot the patients who do have the disease, which is crucial in healthcare.

Overall confusion matrix further highlights the class imbalance difficulty with the model which predicted 113 true negatives and only 23 true positives which predicted correctly that there was no heart disease and predicted correctly that there was heart disease, respectively, and misclassified 32 cases of the disease as negative. This shows that although good model performance for nonheart disease cases, the model can be improved in detecting heart disease.

5. Conclusion

5. Conclusion

Summary of Findings:

“The Decision Tree model demonstrated an accuracy of 73.91% in predicting heart disease”. The classification report indicated that while the model performed well in identifying individuals without heart disease (high recall), it struggled to accurately detect those with heart disease (low recall). The confusion matrix confirmed this, showing a significant number of false negatives.

Implications:

In a real-world healthcare setting, this model could be a valuable tool for aiding healthcare professionals in the early diagnosis of heart disease. By providing a data-driven prediction, the model could assist doctors in identifying at-risk patients and facilitating timely interventions.

Limitations:

One limitation of this approach is the potential for overfitting, especially with complex decision trees. Additionally, the class imbalance in the dataset could have affected the model's performance. Using more advanced models, such as Random Forest or XGBoost, could help mitigate these issues and improve accuracy (Sahin, 2020).

Future Work:

Future research could focus on integrating Decision Trees with ensemble methods to enhance performance. Improving data quality, using larger datasets, and integrating this model into real-time healthcare systems could also lead to better outcomes and more accurate predictions in clinical practice.

6. References

Amini, M., Zayeri, F. and Salehi, M., 2021. Trend analysis of cardiovascular disease mortality, incidence, and mortality-to-incidence ratio: results from global burden of disease study 2017. BMC public health, 21, pp.1-12. https://link.springer.com/content/pdf/10.1186/s12889-021-10429-0.pdf

Chakraborty, M., 2024. Explainable Neural Networks: Achieving Interpretability in Neural Models. Archives of Computational Methods in Engineering, 31(6), pp.3535-3550. https://www.researchgate.net/profile/Manomita-Chakraborty/publication/379163767_Explainable_Neural_Networks_Achieving_Interpretability_in_Neural_Models/links/660052f0a8baf573a1d59145/Explainable-Neural-Networks-Achieving-Interpretability-in-Neural-Models.pdf

Dissanayake, K. and Md Johar, M.G., 2021. Comparative study on heart disease prediction using feature selection techniques on classification algorithms. Applied Computational Intelligence and Soft Computing, 2021(1), p.5581806. https://onlinelibrary.wiley.com/doi/pdf/10.1155/2021/5581806

Duval, S., Van't Hof, J.R., Steffen, L.M. and Luepker, R.V., 2020. Estimation of cardiovascular risk from self-reported knowledge of risk factors: insights from the Minnesota Heart Survey. Clinical Epidemiology, pp.41-49. https://www.tandfonline.com/doi/pdf/10.2147/CLEP.S219708

Izza, Y., Ignatiev, A. and Marques-Silva, J., 2020. On explaining decision trees. arXiv preprint arXiv:2010.11034. https://arxiv.org/pdf/2010.11034

Kaggle, 2025. UCI Heart Disease Dataset. Viewed on 25th MArch 2025. From https://www.kaggle.com/datasets/redwankarimsony/heart-disease-data

Robert, A.A. and Al Dawish, M.A., 2021. Cardiovascular disease among patients with diabetes: The current scenario in Saudi Arabia. Current Diabetes Reviews, 17(2), pp.180-185. https://www.researchgate.net/profile/Asirvatham-Alwin-Robert/publication/341703132_Cardiovascular_Disease_Among_Patients_With_Diabetes_The_

Current_Scenario_in_Saudi_Arabia/links/661f64b4f7d3fc2874665d8f/Cardiovascular-Disease-Among-Patients-With-Diabetes-The-Current-Scenario-in-Saudi-Arabia.pdf

Romiti, S., Vinciguerra, M., Saade, W., Anso Cortajarena, I. and Greco, E., 2020. Artificial intelligence (AI) and cardiovascular diseases: an unexpected alliance. Cardiology Research and Practice, 2020(1), p.4972346. https://onlinelibrary.wiley.com/doi/pdf/10.1155/2020/4972346

Sahin, E.K., 2020. Assessing the predictive capability of ensemble tree methods for landslide susceptibility mapping using XGBoost, gradient boosting machine, and random forest. SN Applied Sciences, 2(7), p.1308. https://link.springer.com/content/pdf/10.1007/s42452-020-3060-1.pdf

Zhou, W., Yan, Z. and Zhang, L., 2024. A comparative study of 11 non-linear regression models highlighting autoencoder, DBN, and SVR, enhanced by SHAP importance analysis in soybean branching prediction. Scientific Reports, 14(1), p.5905. https://www.nature.com/articles/s41598-024-55243-x.pdf

Zhou, Y.Y., Qiu, H.M., Yang, Y. and Han, Y.Y., 2020. Analysis of risk factors for carotid intima-media thickness in patients with type 2 diabetes mellitus in Western China assessed by logistic regression combined with a decision tree model. Diabetology & metabolic syndrome, 12, pp.1-13. https://link.springer.com/content/pdf/10.1186/s13098-020-0517-8.pdf