Heart Disease Prediction

Anna Nander, Brian Chang, Celine Habashy, Yeji Sohn

Table of contents

Summary 2

Introduction 2

Data 3
- Data Description . . . 3
- Data Cleaning Process . . . 4
  - Cleaning Data . . . 4
  - Validating Data . . . 5
- Cleaned Dataset . . . 6

Methods 7
- Procedures . . . 7
  - Evaluation Metric: F1 Score . . . 7
  - Model Training and Validation Approach . . . 7
  - Tools and Libraries Used for Analysis . . . 8
- EDA . . . 8
- Model Results . . . 12
  - Decision Tree: Cross-Validation Results . . . 12
  - Decision Tree: Final Results . . . 13
  - Logistic Regression: Cross-Validation Results . . . 13
  - Logistic Regression: Coefficients . . . 15
  - Logistic Regression: Final Results . . . 15

Discussion 16
  - Results . . . 16
- Unexpected Findings: . . . 17
- Future Work: . . . 17

References 18

Summary

In this project, we developed a classification system using Logistic Regression and Decision Tree models to predict a **heart disease diagnosis based on multiple features such as age, blood pressure, cholesterol, and more**. For a full explanation of variables please refer to the Key Features section. The data was sourced from the UCI Heart Disease dataset (Janosi and Detrano 1989), and preprocessing involved cleaning, transforming, and encoding categorical variables for analysis. According to our experiments, the logistic regression model achieved the highest accuracy with 84.44%. Decision Tree provided competitive results, with an accuracy of 74.44%, but lacked the interpretability of logistic regression. The results suggest that machine learning models have the potential to be used to predict heart disease effectively, aiding healthcare providers in early detection and intervention. However, since this report only contains experiments with Decision Trees and Logistic Regression, further experimentation should be conducted to investigate other, potentially better models.

Introduction

Heart disease is one of the leading causes of death worldwide, and early detection is crucial for improving treatment outcomes and patient survival rates. Timely diagnosis can help healthcare providers make more informed decisions, allocate resources more effectively, and ultimately save lives. Traditional diagnostic methods often involve manual interpretation of clinical test results, which can be time-consuming, subjective, and prone to errors. As health data becomes increasingly available, machine learning has emerged as a powerful tool for diagnosing and predicting diseases, including heart disease.

This project explores the application of machine learning models to classify individuals based on their likelihood of having heart disease using clinical data. Specifically, we use the UCI Heart Disease dataset, which contains medical records of patients, including features such as age, chest pain type, blood pressure, cholesterol levels, and other relevant clinical attributes. The dataset also includes a binary diagnosis label indicating the presence or absence of heart disease, which forms the basis for predictive modeling.

For the analysis, we focus on the Heart Disease dataset, which includes 13 features. These features represent key clinical indicators used to assess cardiovascular health, and the target variable categorically indicates the presence or absence of heart disease. For the purpose of this analysis, we focus on a binary classification problem, where we aim to distinguish between individuals with no heart disease and those with some form of heart disease. Additionally, the dataset has been anonymized to protect patient privacy, with identifiers such as names and social security numbers replaced by anonymous values.

The main questions addressed in this analysis are:

1. What is the overall accuracy of a classification model for heart disease prediction?

2. Which features are most predictive of the presence of heart disease?

In this analysis, we aim to determine whether **a machine learning model can predict the likelihood of a patient developing heart disease based on various health indicators and demographic factors**. This prediction can be highly valuable for healthcare providers, enabling them to identify individuals at higher risk and prioritize early interventions or preventive measures. By focusing resources on those most at risk, healthcare institutions can optimize patient care and reduce unnecessary testing or treatments for low-risk individuals. Additionally, this analysis will provide deeper insights into the key factors contributing to heart disease, helping healthcare professionals design more personalized and effective prevention strategies for their patients.

Data

The data was downloaded using (Contributors 2024)

Data Description

The dataset used in this project is UCI Heart Disease dataset consisting of 303 patients records (Janosi and Detrano 1989). The data was collected from 425 patients undergoing angiography at the Hungarian Institute of Cardiology in Budapest, Hungary (Detrano et al. 1989). The dataset is anonymized to protect patient privacy and includes 13 features that provide valuable insights into an individual's health status. Our target/response variable is `diagnosis`.

Key Features:

1. age: The age of the patient in years.
2. sex: The gender of the patient (1 = male, 0 = female).
3. chest_pain_type: Indicates the type of chest pain experienced, categorized as:
    - 0: Typical angina
    - 1: Atypical angina
    - 2: Non-anginal pain
    - 3: Asymptomatic
4. resting_blood_pressure: The patient's resting blood pressure in mmHg.
5. cholesterol: Serum cholesterol levels in mg/dL.
6. fasting_blood_sugar: A binary feature indicating if fasting blood sugar is > 120 mg/dL (1 = true, 0 = false).
7. rest_ecg: Resting electrocardiogram results, coded as:
    - 0: Normal

- 1: Having ST-T wave abnormality
- 2: Showing probable or definite left ventricular hypertrophy.

8. max_heart_rate: Maximum heart rate achieved during exercise.
9. exercise_induced_angina: A binary feature indicating the presence of exercise-induced angina (1 = yes, 0 = no).
10. st_depression: ST depression induced by exercise relative to rest.
11. slope: The slope of the peak exercise ST segment:

    - 0: Upsloping
    - 1: Flat
    - 2: Downsloping.

12. num_of_vessels: The number of major vessels (0–3) colored by fluoroscopy.
13. thalassemia: A categorical feature representing a blood disorder:

    - 0: Normal
    - 1: Fixed defect
    - 2: Reversible defect.

14. diagnosis: The target variable, indicating the presence or absence of heart disease:

    - 0: No heart disease
    - 1: Heart disease (aggregated from severity levels 1–4 in the original dataset).

Data Cleaning Process

In order to prepare the raw heart disease dataset for analysis, several cleaning steps were undertaken to ensure the data is consistent, complete, and properly formatted. These steps include renaming columns, relabeling categorical values, and handling missing data.

Below is a detailed description of each cleaning task performed:

Cleaning Data

1. **Renaming Columns**: The original dataset contained column names that were unclear. To make the dataset more understandable and easier to work with, the column names were renamed to more descriptive, readable terms.
2. **Relabeling Categorical Variables**: Several columns contained categorical data that needed to be relabeled for better interpretability. Table 1 summarizes the newly mapped values.

Table 1: Raw Table

Column Name	Original Values	New Values
Chest Pain Type	1, 2, 3, 4	'Typical angina', 'Atypical angina', 'Non-anginal pain', 'Asymptomatic'
Fasting Blood Sugar	'yes', 'no'	1 (yes), 0 (no)
Resting ECG	0, 1, 2	'Normal', 'ST-T wave abnormality', 'Left ventricular hypertrophy'
Exercise Induced Angina	0, 1	'No', 'Yes'
Slope	1, 2, 3	'Upsloping', 'Flat', 'Downsloping'
Thalassemia	3, 6, 7	'Normal', 'Fixed defect', 'Reversible defect'
Diagnosis	2, 3, 4	1 (presence of heart disease)

3. **Handling Missing Values**: The raw dataset contained missing values that could interfere with analysis. To address this, all rows with missing values were removed from the dataset.

Validating Data

After cleaning, several validation checks were performed to ensure data quality:

- **Empty Observations**: A check was run to ensure there were no rows where all values were missing.
- **Missingness Threshold**: Missing values in numeric and categorical columns were validated to ensure they did not exceed the defined thresholds (e.g., no more than 5% missing data in certain columns).
- **Column Names and Data Types**: The dataset was validated for correct column names and data types (e.g., integer, string, float) to ensure consistency with expectations.
- **Duplicate Records**: A check for duplicate rows was conducted to ensure that no redundant observations were present.
- **Outliers and Anomalous Values**: Numerical columns were validated to ensure values fell within acceptable ranges (e.g., age between 0 and 120, cholesterol between 100 and 600).
- **Correct Categorical Levels**: Categorical variables were checked to ensure they only contained valid values (e.g., 'typical angina', 'normal', etc.).
- **Class Proportions**: The distribution of the target variable, "diagnosis," was checked to ensure balanced proportions of 0 and 1, within a specified tolerance.
- **Feature-Label Correlation**: A deep check was performed to ensure that the correlation between features and the target variable ("diagnosis") did not exceed an acceptable threshold (0.9), which could indicate potential data leakage or multicollinearity.

- **Feature-Feature Correlation**: A check was conducted to ensure no pairs of features exhibited highly correlated relationships that could lead to multicollinearity.

Note: Since we have imported data from the ucimlrepo, we will not be checking for correct data file format.

Cleaned Dataset

This dataset (Table 2) will be used onwards for all analysis.

Table 2: Cleaned Table

Variable Name	Role	Type	Description	Units	Missing Values
age	Feature	Integer	Age of the patient in years	years	no
sex	Feature	Categorical	Male/Female		no
chest_pain_type	Feature	Categorical	Chest pain type: [typical angina, atypical angina, non-anginal, asymptomatic]		no
resting_blood_pressure	Feature	Integer	Resting blood pressure (in mm Hg on admission to the hospital)	mm Hg	no
cholesterol	Feature	Integer	Serum cholesterol in mg/dl	mg/dl	no
fasting_blood_sugar	Feature	Categorical	If fasting blood sugar > 120 mg/dl		no
rest_ecg	Feature	Categorical	Resting electrocardiographic results: [normal, stt abnormality, lv hypertrophy]		no
max_heart_rate	Feature	Integer	Maximum heart rate achieved	bpm	no
exercise_induced_angina	Feature	Categorical	Exercise-induced angina (True/False)		no
st_depression	Feature	Integer	ST depression induced by exercise relative to rest		no
slope	Feature	Categorical	The slope of the peak exercise ST segment		no
num_of_vessels	Feature	Integer	Number of major vessels (0-3) colored by fluoroscopy		no
thalassemia	Feature	Categorical	Thalassemia: [normal, fixed defect, reversible defect]		no
diagnosis	Target	Categorical	Predicted attribute (1 = presence of heart disease, 0 = absence)		no

Methods

Procedures

Evaluation Metric: F1 Score

In this study, we chose the F1 score as our primary evaluation metric for the performance of the predictive model, instead of using recall alone. The decision to use F1 score stems from the importance of balancing both false positives and false negatives in the context of heart disease detection.

While false negatives (missed diagnoses of patients with actual heart disease) are critical, as they may lead to serious consequences such as delayed treatment and even death, false positives are also of significant concern. A false positive in this context refers to a patient being incorrectly diagnosed with heart disease, which may result in unnecessary additional tests, procedures, and treatments. These can be not only expensive but also stressful and potentially harmful to patients.

By using the F1 score, which is the harmonic mean of precision and recall, we aim to achieve a balance between minimizing both false positives and false negatives. This approach ensures that the model not only identifies as many true heart disease cases as possible (high recall) but also avoids the over-diagnosis of patients who do not have heart disease (high precision). In doing so, the F1 score provides a more comprehensive evaluation of the model's performance, considering the trade-offs between these two types of errors in healthcare.

Model Training and Validation Approach

For model evaluation, we employed a 70:30 train-test split, where 70% of the data was used for training the model, and the remaining 30% was reserved for testing. This partition ensures that the model is trained on a sufficient amount of data while allowing for a robust evaluation of its performance on unseen data.

Additionally, to further enhance the reliability of our results, we utilized cross-validation during the training process. Specifically, we performed k-fold cross-validation to compute the average performance across multiple train/validation splits. This approach helps mitigate the risk of overfitting by ensuring that the model is evaluated on different subsets of the training data, leading to a more generalized performance estimate. The average scores across these folds were used to assess the model's effectiveness and consistency in making predictions.

Tools and Libraries Used for Analysis

The Python programming language (Van Rossum and Drake 2009) and the following Python packages were used to perform the analysis: pandas (team 2020), scikit-learn (Pedregosa et al. 2011), pandera (Bantilan 2020), pip (pypa 2024), pandas (team 2024), ipykernel (Ip. D. Team 2024), conda (Anaconda 2024), matplotlib (M. D. Team 2024), requests (Reitz 2024), seaborn (S. Developers 2024), quarto (Q. Developers 2024), click (Ronacher 2024), tabulate (Moore 2024), make (Foundation 2024), deepchecks (Deepchecks 2024), pytest (pytest-dev 2024).

EDA

To better understand the dataset and the relationships between the features and the target variable (diagnosis), we created several visualizations. These allowed us to identify patterns, correlations, and key features that could inform the modelling process.

In the processed data, Classes 1, 2, and 3 were combined into a single category, resulting in two main classes: Class 0 (no or mild disease) and Class 1 (moderate to severe disease). Looking at Figure 1 the group without heart disease is slightly larger. The distribution of these two classes is balanced, with nearly equal representation of patients in each. This balance is beneficial for modelling, as it reduces the risk of bias toward one class and allows the model to learn effectively from both categories. This balance is important as it allows for meaningful analysis and ensures that both groups are adequately represented when building predictive models.

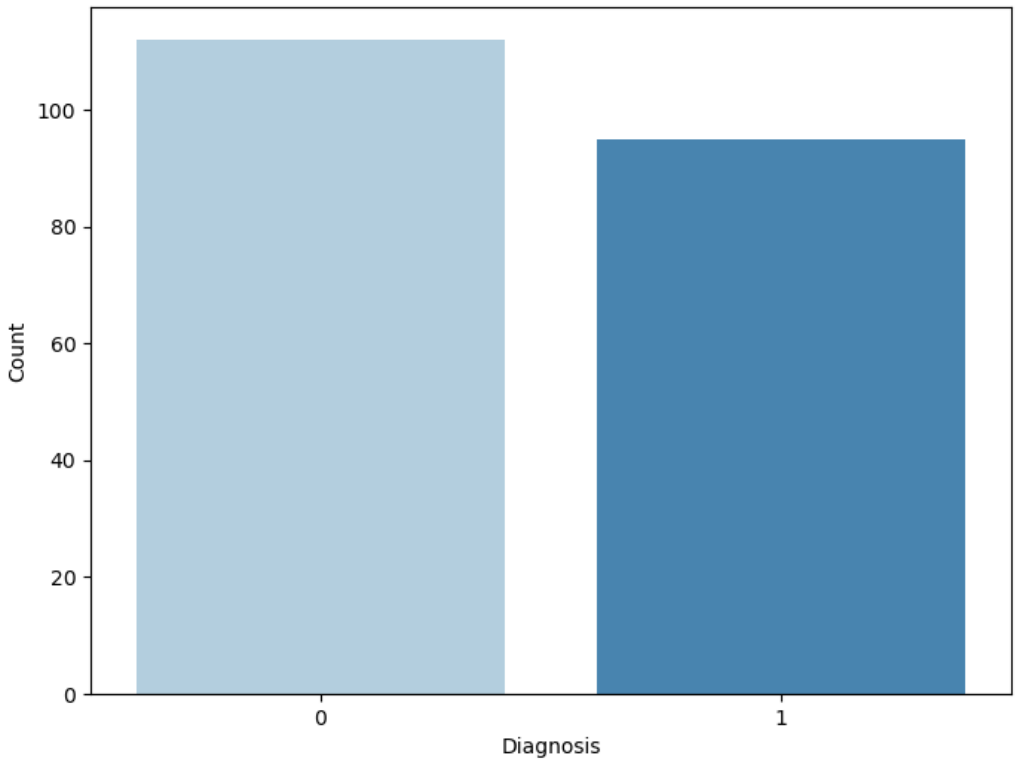

Figure 1: Distribution of Diagnosis

To identify features that might help predict heart disease severity, we examined the distributions and relationships of the continuous predictors. The correlation heatmap Figure 2 showed that st_depression, ca, thal, and max_heart_rate had the strongest relationships with diagnosis, suggesting that these features are likely to be the most valuable. Pairwise plots provided more insights , showing clear trends such as lower max_heart_rate and higher st_depression values being associated with Class 1. In contrast, features like cholesterol and fasting_blood_sugar showed little separation between the two classes, indicating they may be less predictive on their own.

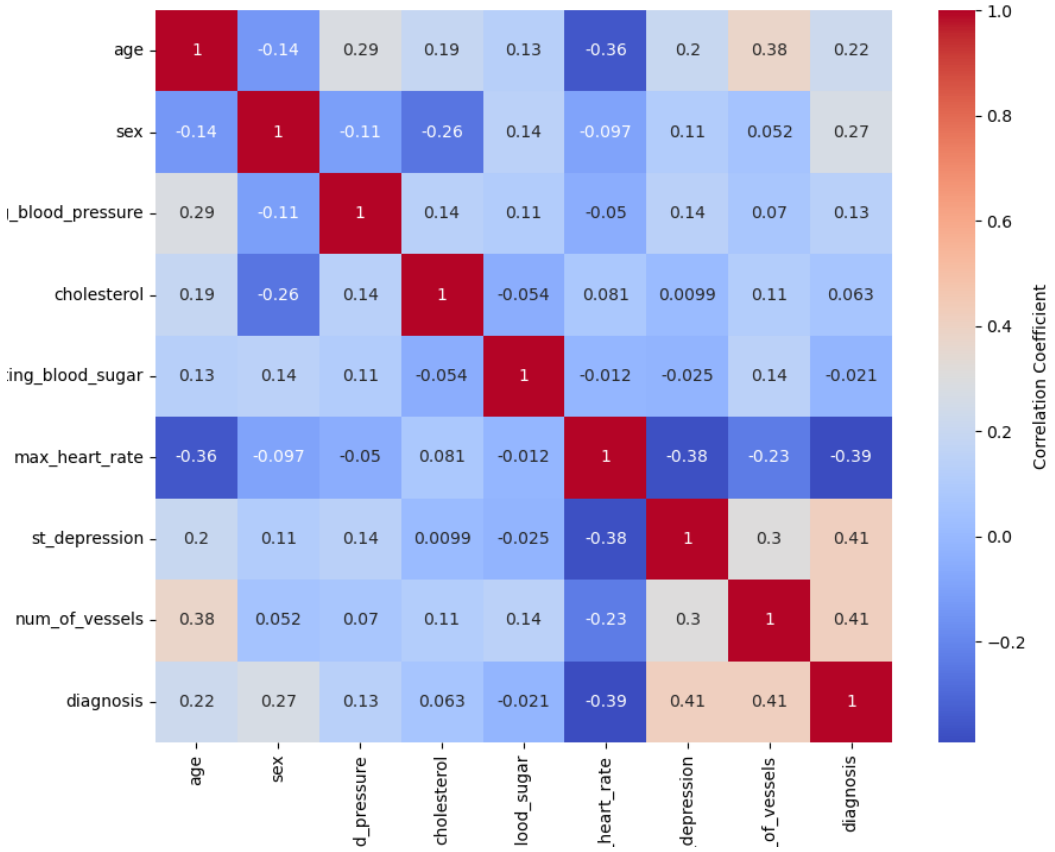

Figure 2: Correlation Heatmap

Figure 3 below visualizes the relationships between key health indicators in the heart disease dataset, grouped by diagnosis. Individuals with heart disease tend to have lower maximum heart rates, as seen in Figure 3. In contrast, cholesterol and resting blood pressure show small separation, indicating lower predictive value. Lastly, Figure 3 shows a mild negative trend between age and max heart rate suggests older individuals have lower max heart rates.

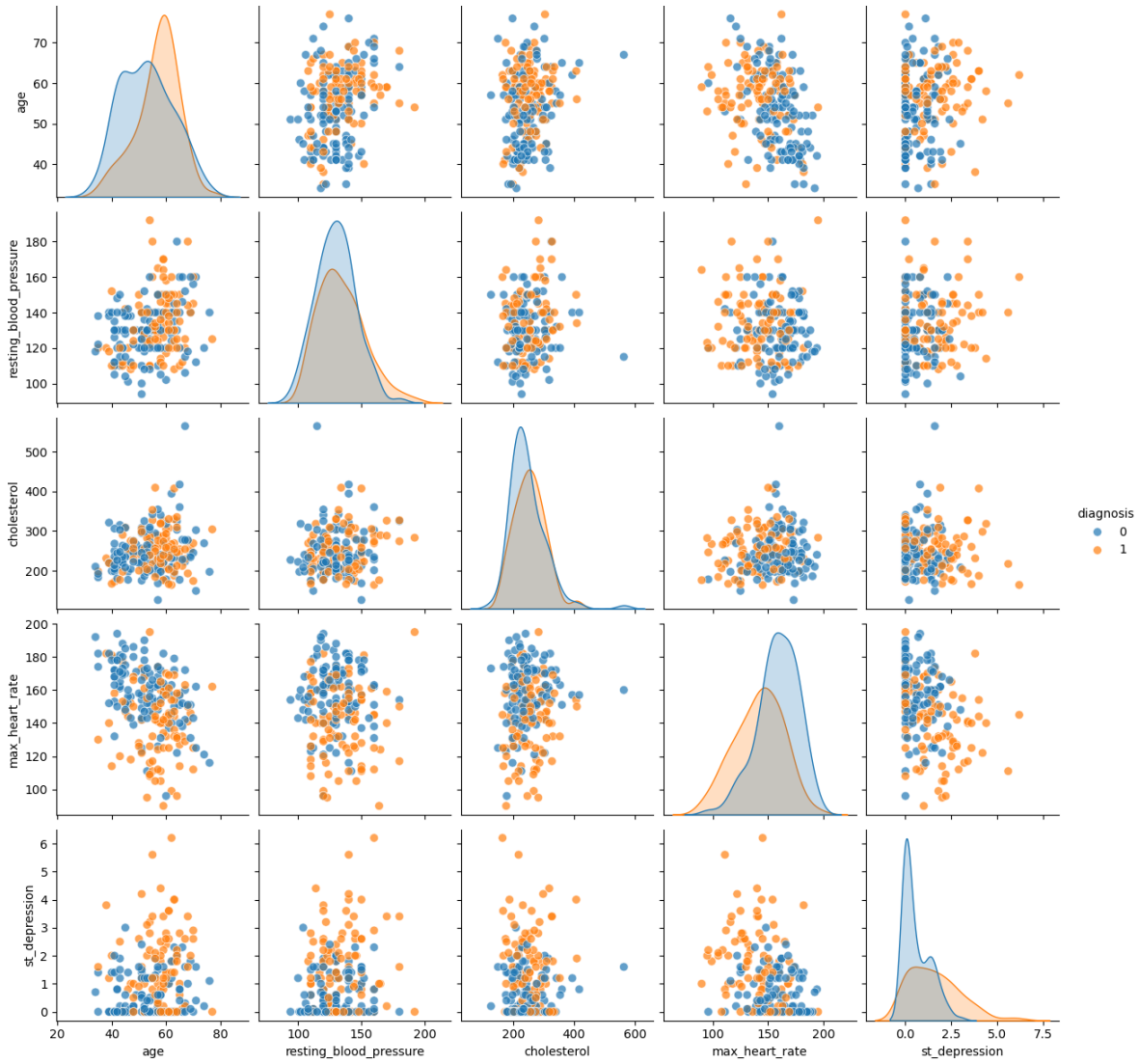

Figure 3: Feature Densities by Diagnosis

Overall, st_depression and max_heart_rate emerge as the most important features for predicting heart disease severity, while features like cholesterol may play a more limited role in the model. The distribution of the target variable shows that the data is well-balanced between the two classes. Class 0 and Class 1 have nearly equal representation in the dataset. The balanced distribution of the two classes ensures the model will have a fair representation of both disease and non-disease cases, helping improve its performance.

Model Results

Decision Tree: Cross-Validation Results

Table 3: Cross-validation results of Decision Tree Model

	mean	std
fit_time	0.003	0
score_time	0.005	0.001
test_accuracy	0.74	0.059
train_accuracy	1	0
test_precision	0.709	0.069
train_precision	1	0
test_recall	0.737	0.129
train_recall	1	0
test_f1	0.718	0.079
train_f1	1	0

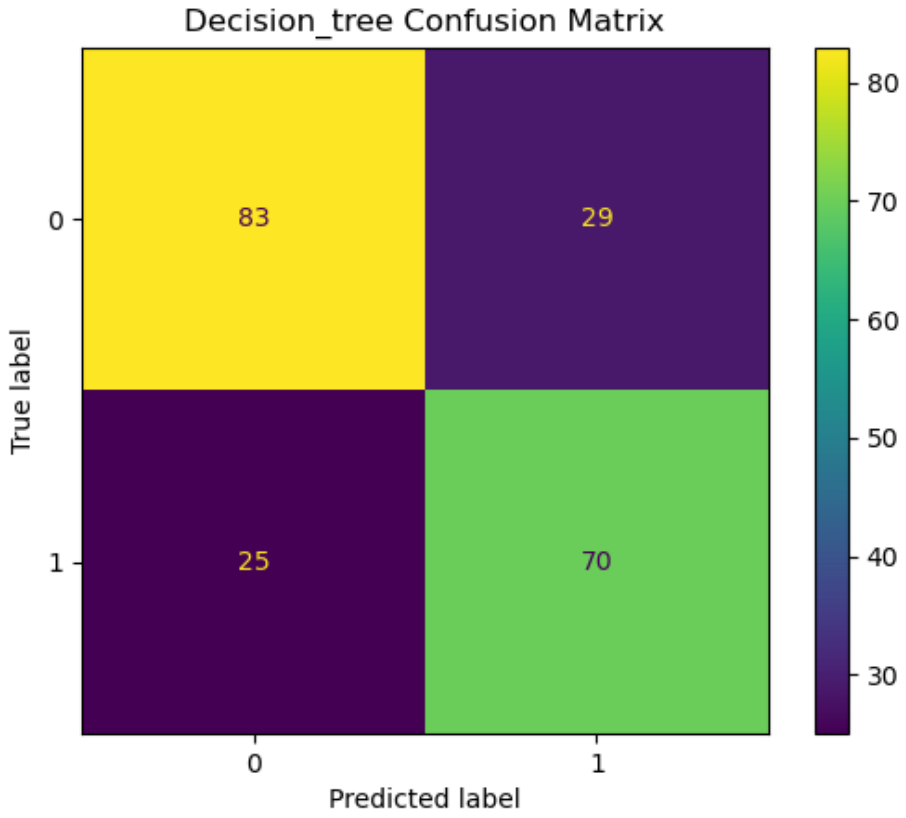

Figure 4: Confusion Matrix of Decision Tree Model

Decision Tree: Final Results

Table 4: Final results of Decision Tree Model

	precision	recall	f1-score
0	0.727273	0.833333	0.776699
1	0.771429	0.642857	0.701299
accuracy	0.744444	0.744444	0.744444

Logistic Regression: Cross-Validation Results

Table 5: Cross-Validation results of Logistic Regression Model

	mean	std
fit_time	0.007	0.001
score_time	0.007	0.001
test_accuracy	0.826	0.021
train_accuracy	0.873	0.009
test_precision	0.822	0.054
train_precision	0.892	0.018
test_recall	0.8	0.044
train_recall	0.824	0.015
test_f1	0.809	0.018
train_f1	0.856	0.009

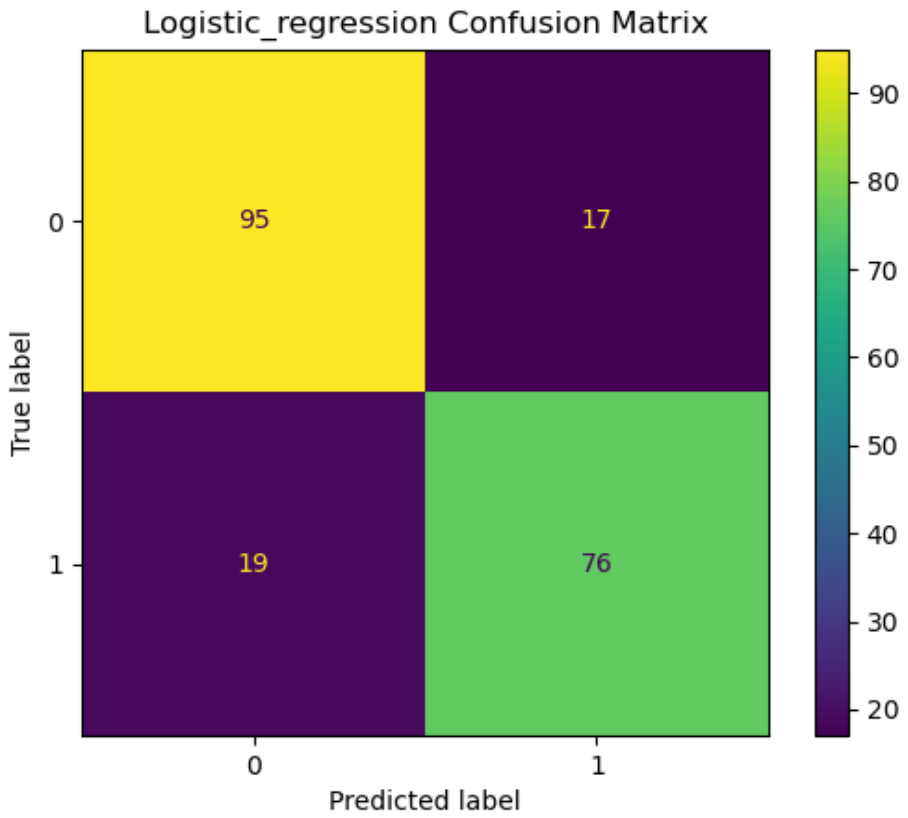

Figure 5: Confusion Matrix of Logistic Regression Model

Logistic Regression: Coefficients

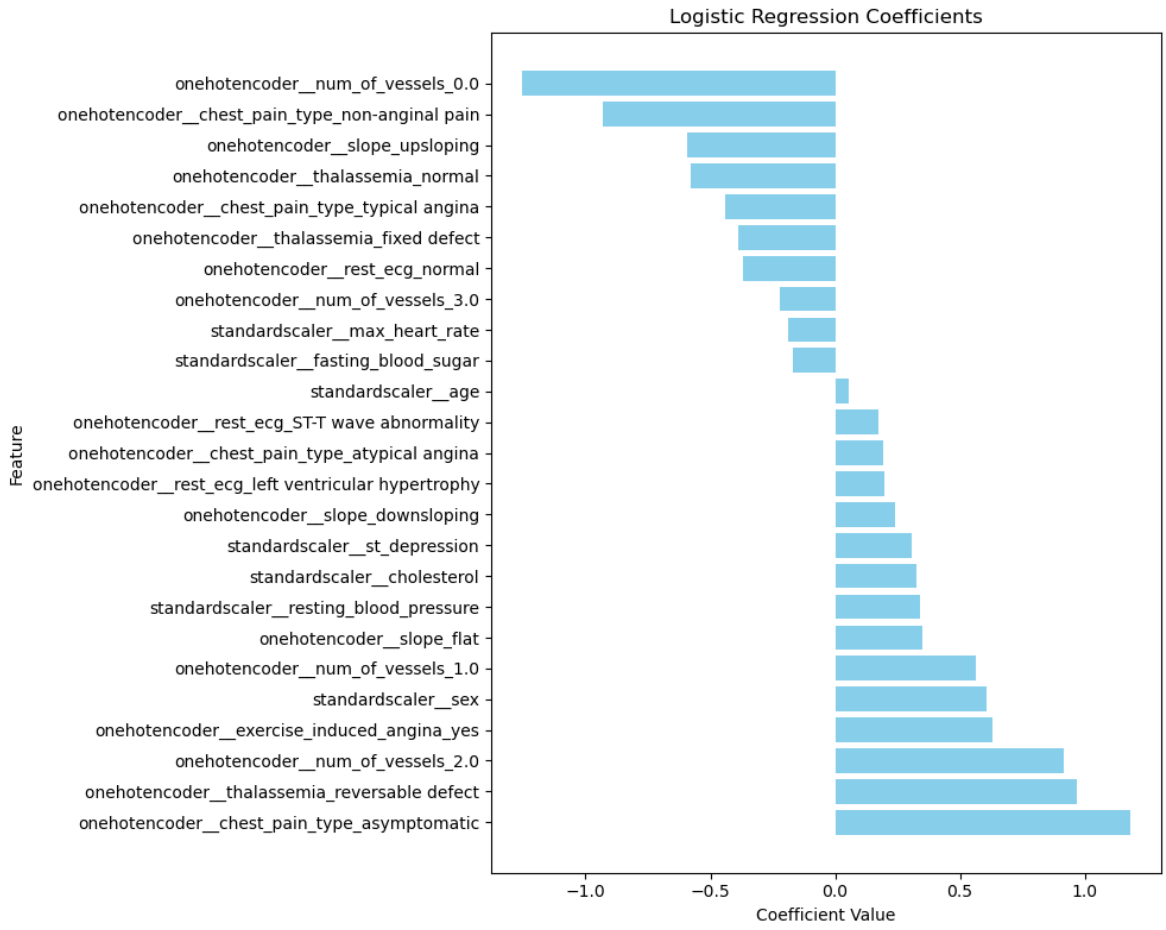

Figure 6: Coefficients of Logistic Regression Model

Logistic Regression: Final Results

Table 6: Final results of Logistic Regression Model

	precision	recall	f1-score
0	0.826923	0.895833	0.86
1	0.868421	0.785714	0.825
accuracy	0.844444	0.844444	0.844444

Discussion

Results

This analysis aimed to evaluate the accuracy and interpretability of machine learning models in predicting heart disease. Specifically, we sought to determine the (1) overall accuracy of classification models, (2) identify key predictive features, and (3) assess the ability of these models to predict the likelihood of a patient developing heart disease based on health indicators and demographic factors.

Overall Accuracy of the Classification Models

Both the logistic regression and decision tree models demonstrated strong predictive capabilities. The decision tree model achieved an accuracy of 73.33%, with a precision of 72.22% and recall of 81.25% for class 0 (absence of heart disease) and a precision of 75.0% and recall of 64.29% for class 1 (presence of heart disease). However, the decision tree model exhibited signs of potential overfitting, as indicated by the 100% training accuracy versus a test accuracy of 72.5%, suggesting that it might not generalize well to unseen data.

In comparison, the logistic regression model outperformed the decision tree model, achieving an overall accuracy of 81.11%. It exhibited balanced precision (78.18% for class 0 and 85.71% for class 1) and recall (89.58% for class 0 and 71.43% for class 1), with a test accuracy of 87.5%, and a training accuracy of 89.6%. This higher performance suggests that logistic regression provides a more reliable model for predicting heart disease, with fewer issues related to overfitting and better generalization to new data.

Key Predictive Features for Heart Disease

An essential objective of this analysis was to identify the most significant features for predicting the presence of heart disease. The logistic regression model, due to its interpretability, proved particularly useful in uncovering key predictors. From the model coefficients, it was clear that the type of chest pain, the slope of the peak exercise ST segment, and the number of vessels colored by fluoroscopy were the most influential features in predicting heart disease. These findings align with clinical knowledge, where symptoms such as chest pain and certain imaging results (such as the number of blocked vessels) are critical indicators of cardiovascular risk.

The decision tree model also provided insights into important features, but its black-box nature made interpretation more challenging. Nevertheless, both models pointed to chest pain type and other cardiovascular factors as essential for clinical decision-making, confirming their relevance in the prediction of heart disease.

Predictive Capability for Patient Diagnosis

The ability to predict whether an individual might develop heart disease based on health indicators and demographic factors was evaluated through both models. Logistic regression demonstrated a strong capability to predict the likelihood of heart disease, with particularly high recall for the absence of heart disease (89.58%), meaning the model was effective in identifying healthy individuals. Conversely, the decision tree model showed a more balanced performance but struggled to consistently predict class 1 (heart disease), as reflected by its lower recall (64.29%) for this group.

Overall, the findings suggest that machine learning models, particularly logistic regression, hold promise for accurately predicting heart disease diagnoses. However, there is still room for improvement in refining these models to reduce misclassifications, especially in predicting the presence of heart disease.

Implications for Clinical Application

The findings of this analysis suggest that machine learning models, particularly logistic regression, can be valuable tools in clinical decision-making for predicting heart disease. Logistic regression demonstrated superior accuracy and interpretability, making it suitable for use in clinical settings where understanding predictor-outcome relationships is crucial. The identified key features, such as chest pain type and the number of blocked vessels, align with existing medical knowledge, reinforcing their importance in diagnosing heart disease. While decision trees offer a visual representation of classification logic, they require further refinement to improve generalization. Overall, these models show promise for enhancing early diagnosis and risk assessment in patients, potentially supporting healthcare providers in making more informed decisions.

Unexpected Findings:

While many features, such as chest pain type and maximum heart rate, had high predictive power, some features demonstrated lower importance than expected. For instance, fasting blood sugar, a commonly discussed indicator in cardiovascular health, showed limited contribution in our models. This finding suggests that some clinical attributes may have less direct influence on heart disease risk than traditionally assumed or that their impact might be context-dependent.

Future Work:

There are several ways to improve upon the findings of this project:

1. Improving the Model: Trying advanced models like Random Forest or Gradient Boosting could help make predictions more accurate and reliable. These models work well with complex data by combining multiple decision-making techniques.
2. Exploring New Features: Adding more details to the data, like lifestyle habits (e.g., smoking, exercise) or family history, could make the model better at predicting heart disease.
3. Making the Model Explainable: Using tools like SHAP or LIME can help us understand why the model makes certain predictions. This is especially important for gaining trust in a healthcare setting.
4. Testing in the Real World: It would be valuable to test the model with real patient data in a clinical environment to see how it performs outside the lab.
5. Dealing with Uneven Data: If the dataset has many more people without heart disease than with it, methods like balancing the data or focusing on the underrepresented group can make the model fairer and more accurate.

References

Anaconda, Inc. 2024. "Nb_conda_kernels." https://github.com/Anaconda-Platform/nb_conda_kernels.

Bantilan, Niels. 2020. "Pandera: Statistical Data Validation of Pandas Dataframes." In *Proceedings of the 19th Python in Science Conference*, edited by Meghann Agarwal, Chris Calloway, Dillon Niederhut, and David Shupe, 116–24. https://doi.org/ 10.25080/Majora-342d178e-010 .

Contributors, UCI Machine Learning Repository. 2024. "Ucimlrepo." https://pypi.org/project/ucimlrepo/.

Deepchecks. 2024. "Deepchecks." https://www.deepchecks.com/.

Detrano, Robert C., András Jánosi, Walter Steinbrunn, Matthias Emil Pfisterer, Johann-Jakob Schmid, Sarbjit Sandhu, Kern Guppy, Stella Lee, and Victor Froelicher. 1989. "International Application of a New Probability Algorithm for the Diagnosis of Coronary Artery Disease." *The American Journal of Cardiology* 64 5: 304–10. https://api.semanticscholar.org/CorpusID:23545303.

Developers, Quarto. 2024. "Quarto." https://quarto.org/.

Developers, Seaborn. 2024. "Seaborn." https://seaborn.pydata.org/.

Foundation, Free Software. 2024. "Make." https://www.gnu.org/software/make/.

Janosi, Steinbrunn, Andras, and Robert Detrano. 1989. "Heart Disease." UCI Machine Learning Repository.

Moore, Matthew D. J. 2024. "Tabulate." https://pypi.org/project/tabulate/.

Pedregosa, F., G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, et al. 2011. "Scikit-learn: Machine Learning in Python." *Journal of Machine Learning Research* 12: 2825–30.

pypa. 2024. "Pip." https://pip.pypa.io/.

pytest-dev. 2024. "Pytest." https://pytest.org/.

Reitz, Kenneth. 2024. "Requests: HTTP for Humans." https://requests.readthedocs.io/.

Ronacher, Armin. 2024. "Click." https://click.palletsprojects.com/.

Team, IPython Development. 2024. "IPython Kernel." https://ipython.org/.

Team, Matplotlib Development. 2024. "Matplotlib." https://matplotlib.org/.

team, The pandas development. 2020. "Pandas-Dev/Pandas: Pandas." Zenodo. https://doi.org/10.5281/zenodo.3509134.

———. 2024. "Pandas." https://pandas.pydata.org/.

Van Rossum, Guido, and Fred L. Drake. 2009. *Python 3 Reference Manual*. Scotts Valley, CA: CreateSpace.