

Heart Disease Prediction and Deployment

Created by:
Atithi Shrestha

Table of Contents

Heart Disease Prediction and Deployment	1
Table of Figures	3
Introduction	4
Problem Definition	4
Objective	4
Dataset Selection	4
Data Preprocessing	5
Handling Missing Values	5
Feature Engineering	5
Data Cleaning	5
Standardization and Dimensionality Reduction	5
Outlier Detection	5
Exploratory Data Analysis (EDA)	6
Model Development and Evaluation	9
Model Training and Ensemble Methods	9
Model Selection	9
Hyperparameter Tuning	9
Evaluation Metrics	9
Cross-Validation	9
Unsupervised Model Evaluation	10
Feature Importance	10
Supervised Learning Performance	10
Unsupervised Learning Performance	10
Performance Comparison	11
Model Interpretation	12
Feature Importance	12
SHAP Analysis	12
Comparison with LIME	12
Deployment Strategy	14
Future Enhancements	15
Conclusion	15
References	15

Table of Figures

<i>Figure 1: Age vs Cholesterol using Regression Line.</i>	6
<i>Figure 2: Distribution of Chest Pain Type and Sex by Heart Disease Status.</i>	6
<i>Figure 3: Sns Pairplot.</i>	7
<i>Figure 4: Density Plot of Age by Heart Disease Status.</i>	8
<i>Figure 5: Snapshots of Deployed Streamlit App.</i>	15

Introduction

Heart disease has emerged as one of the most significant global health challenges in recent years, ranking as a leading cause of mortality worldwide. Its impact on individuals, families, and healthcare systems underscores the importance of early detection and prevention. The ability to predict the likelihood of heart disease not only aids in timely intervention but also contributes to better resource allocation and improved patient outcomes.

This project leverages state-of-the-art machine learning techniques to develop a robust predictive model that can assess the risk of heart disease based on a diverse set of patient features. These features may include clinical attributes such as cholesterol levels, blood pressure, age, and gender, as well as lifestyle factors like physical activity, smoking habits, and dietary patterns. By analyzing these variables, the model seeks to uncover patterns and relationships that may not be immediately apparent through traditional statistical methods.

The ultimate goal of this project is to provide healthcare professionals with a reliable tool that can assist in identifying high-risk individuals. Such a tool can enable personalized treatment plans, empower patients with actionable insights, and foster a more proactive approach to heart health. Moreover, this project aligns with the broader vision of integrating machine learning into healthcare, paving the way for smarter and more efficient medical practices.

Through this initiative, we aim to contribute to the fight against heart disease, reducing its prevalence and improving the quality of life for countless individuals around the globe.

Problem Definition

The primary challenge addressed in this project is to predict the presence of heart disease in patients using various medical and lifestyle attributes. Accurate predictions can assist healthcare professionals in making informed decisions regarding patient care and intervention strategies. By analyzing necessary features the model identifies critical patterns that might not be evident through conventional analysis. This enables the early detection of high-risk cases, allowing for timely preventive measures and personalized treatment plans. Ultimately, the project aims to support more effective resource allocation and improve patient outcomes, contributing to the fight against one of the world's leading causes of death.

Objective

The objective of this project is to create a robust machine learning model that can predict the presence or absence of heart disease. This involves data preprocessing, exploratory data analysis, model selection, and evaluation of model performance.

Dataset Selection

The dataset for this analysis, sourced from the UCI Machine Learning Repository, includes 303 instances and 14 features relevant to heart disease prediction. These features encompass demographic and clinical attributes such as age, sex, chest pain type (cp), resting blood pressure (trestbps), serum cholesterol levels (chol), fasting blood sugar (fbs), and resting electrocardiographic results (restecg). Additionally, it includes maximum heart rate achieved (thalach), exercise-induced angina (exang), oldpeak (ST depression), the slope of the peak exercise ST segment, the number of major vessels (ca), and thalassemia (thal). The target variable indicates the presence or absence of heart disease, providing a comprehensive foundation for predictive modeling.

Data Preprocessing

Handling Missing Values

Checking for missing values is a critical step in dataset selection, as it is essential to ascertain whether the dataset is complete and devoid of gaps in the data for any of the variables utilized in the predictive model. Ensuring that all features are fully populated enhances the integrity of the analysis and supports the reliability of the model's outcomes. So in the case of heart disease prediction, there were not any missing values present in the dataset which is the strong indicator of good data quality for straightforward analysis and model training without the need for complex imputation techniques.

Feature Engineering

Feature Engineering focuses in the data preprocessing pipeline that enforces on transforming raw data into meaningful features that enhance the predictive power of machine learning models. It involves the creation of new features, the modification of existing ones, and the removal of redundant or irrelevant features. In the context of heart disease prediction, key feature engineering tasks could involve encoding categorical variables, scaling numerical features, and constructing interaction terms or polynomial features that could provide more information to the model. A feature named “Age_chol_interaction” is created by categorizing age and cholesterol to explore potential synergies between these variables capturing their combined influence on heart disease. Such features can enhance the model's ability to identify nuanced patterns in the data.

Data Cleaning

Data cleaning involved removing duplicate entries and outliers that could adversely affect model performance. The final dataset was standardized to ensure consistency across features.

Standardization and Dimensionality Reduction

The features are standardized using StandardScaler to ensure they have zero mean and unit variance. This step mitigates the influence of features with larger numerical ranges, particularly crucial for distance-based models like KNN and SVM. PCA (Principal Component Analysis) is applied, retaining 95% of the variance, to reduce dimensionality while preserving essential information. This simplifies the model and lowers computational overhead, although its use in final models remains unspecified.

Outlier Detection

Outlier detection leverages Z-scores, flagging data points beyond three standard deviations. While identified, outliers are not explicitly removed or treated in the code. The report should discuss the rationale for this decision, emphasizing its potential impact on model performance and robustness.

Exploratory Data Analysis (EDA)

Visualizations were created to explore relationships between features and the target variables. The following analyses provide insights into the heart disease prediction data:

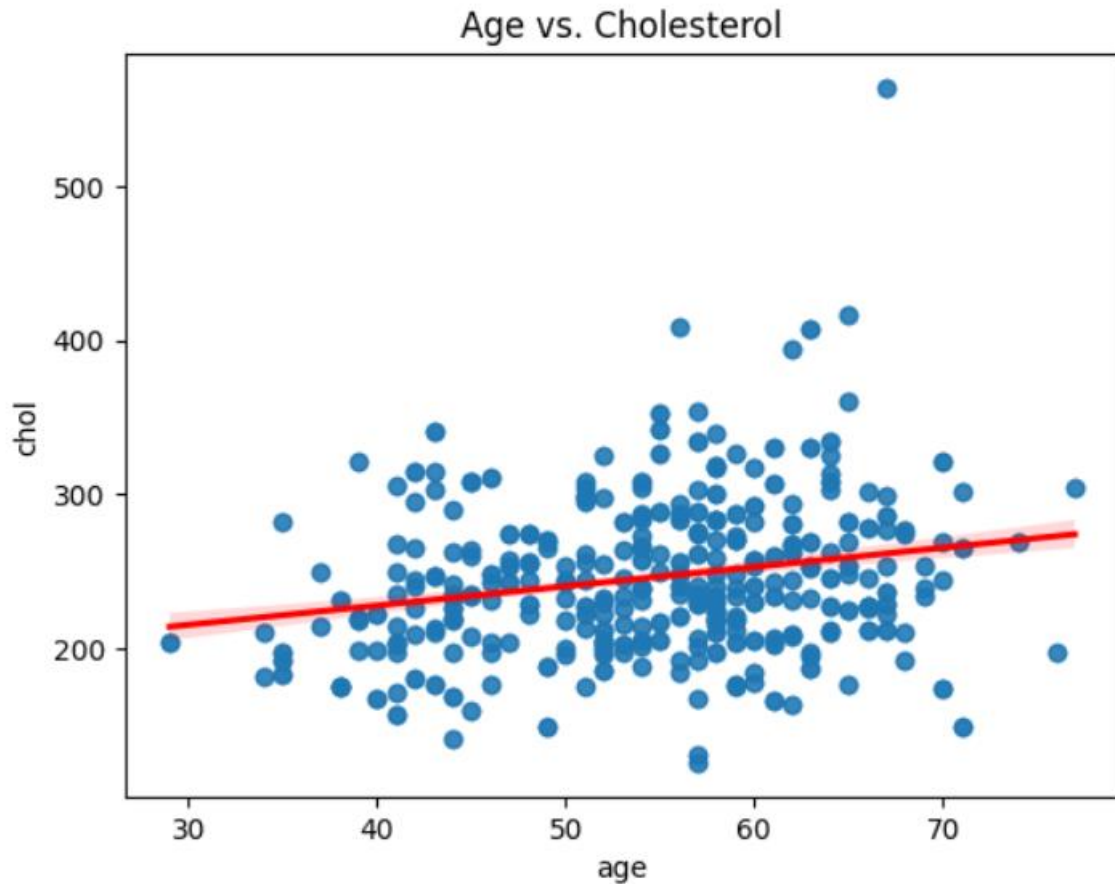


Figure 1: Age vs Cholesterol using Regression Line.

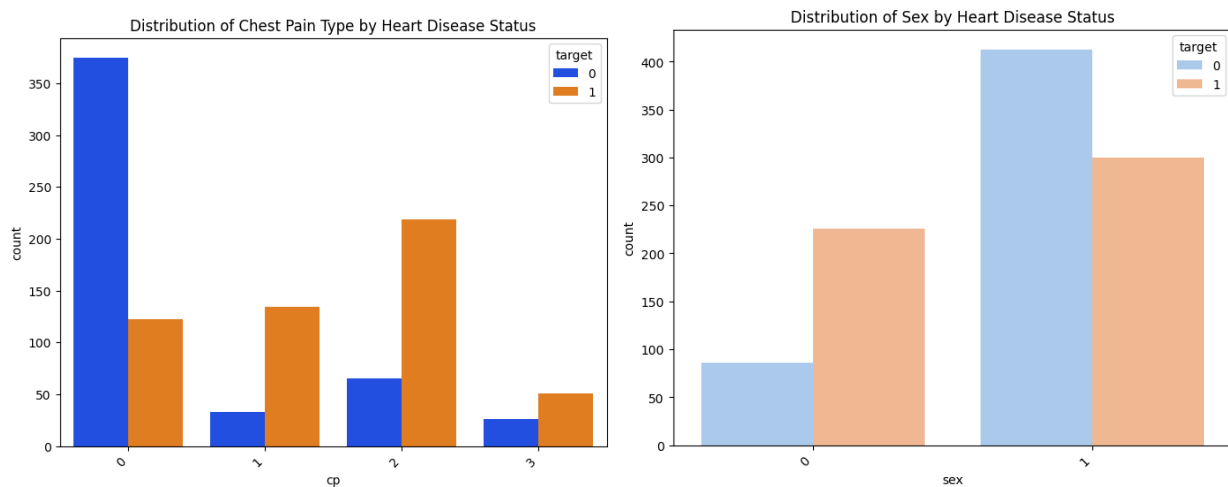


Figure 2: Distribution of Chest Pain Type and Sex by Heart Disease Status.

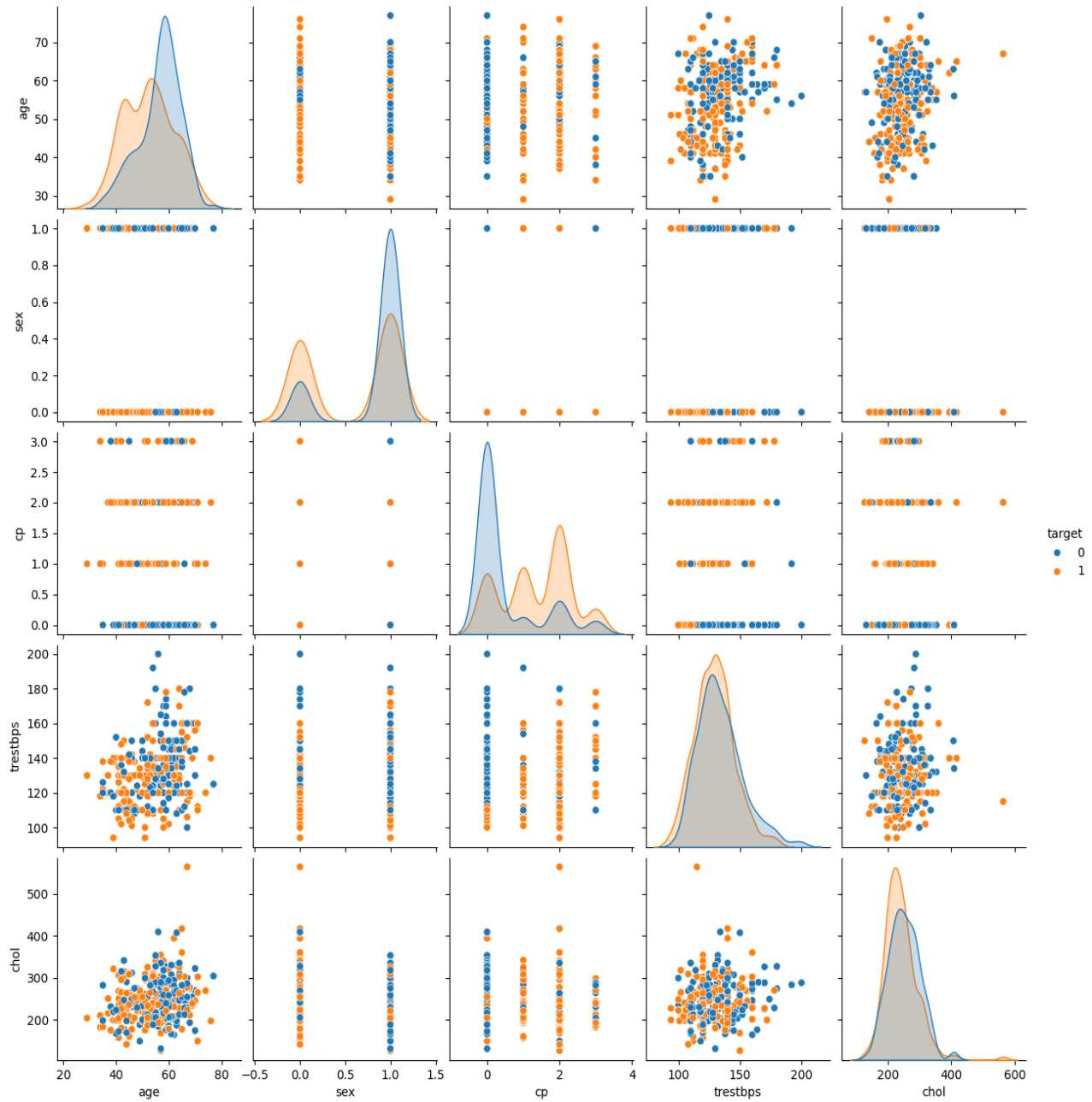


Figure 3: Sns Pairplot.

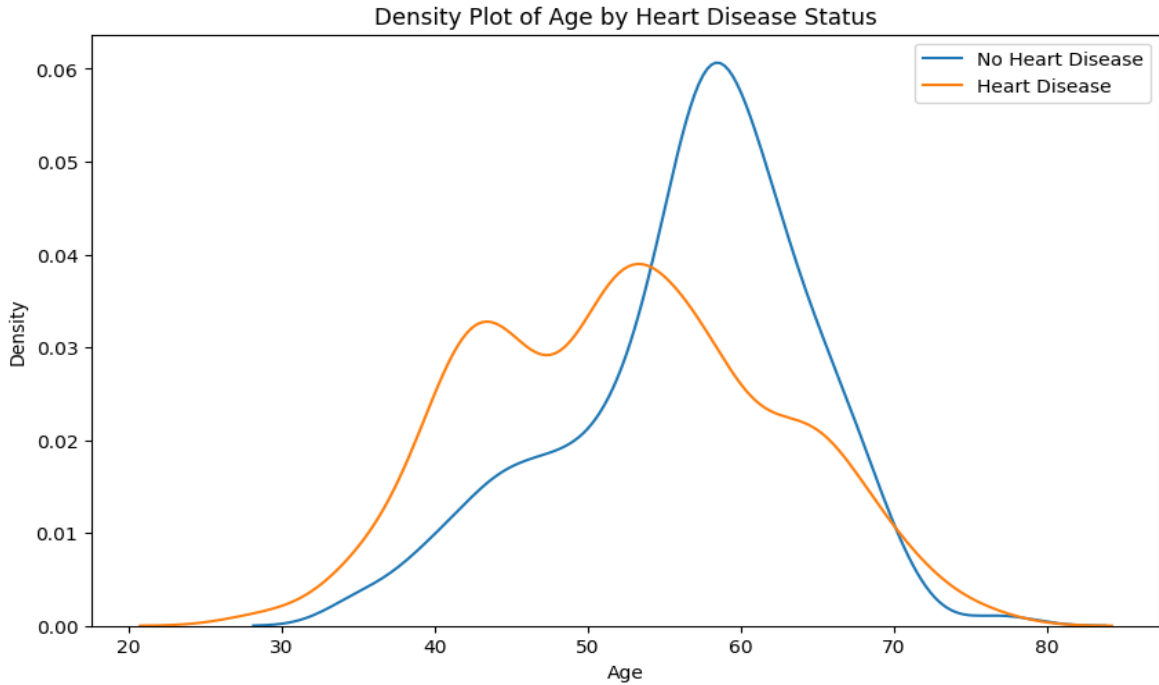


Figure 4: Density Plot of Age by Heart Disease Status.

The exploratory data analysis (EDA) reveals valuable insights into the relationships between features and heart disease presence in the dataset. Age shows a positive correlation with heart disease risk, as individuals with heart disease tend to be slightly older. Chest pain type displays significant associations, with type 2 chest pain being more common among those with heart disease, suggesting it is a strong predictor. While males have a slightly higher incidence of heart disease than females, the difference is not substantial. Cholesterol levels exhibit a weak positive correlation with age, and there is an outlier with cholesterol exceeding 550, warranting further investigation. The diagonal plots highlight variable distributions, with age and cholesterol showing expected patterns, while scatter plots reveal positive relationships between resting blood pressure, cholesterol, and age. These observations suggest potential for feature engineering, such as interaction terms, and emphasize the importance of variables like age, chest pain type, and cholesterol in building predictive models. This EDA serves as a foundational step for feature selection, model building, and evaluation in predicting heart disease.

Model Development and Evaluation

Model Training and Ensemble Methods

Each model is trained on the standardized training dataset, and ensemble methods are used to improve overall performance. The VotingClassifier combines predictions from Random Forest, Bagging, Extra Trees, and XGBoost using soft voting. This method aggregates probability scores from each model, typically resulting in more accurate and stable predictions than individual models. The ensemble approach is particularly effective for heart disease prediction, given the complexity and variability in patient data.

Model Selection

The code explores a diverse range of machine learning algorithms, each with unique characteristics suited for classification tasks. Logistic Regression serves as a baseline, offering simplicity and interpretability. Gaussian Naive Bayes is employed for its computational efficiency and suitability for normally distributed features, although alternative variants like MultinomialNB might be more applicable for count data. K-Nearest Neighbors (KNN) uses distance-based predictions, which are intuitive but computationally expensive for larger datasets. Decision Tree Classifiers provide easily interpretable tree-like models but are prone to overfitting without proper regularization. Support Vector Machines (SVMs) are utilized with a linear kernel, performing well in high-dimensional spaces but requiring significant computational resources.

Ensemble methods are a key focus, combining multiple models to enhance performance and robustness. The Random Forest algorithm aggregates decision trees to reduce variance and overfitting. Boosting techniques like AdaBoost and Gradient Boosting iteratively improve upon previous models, with XGBoost offering optimized performance and efficiency. Bagging and Extra Trees further explore variance reduction and randomness in tree construction. The inclusion of unsupervised methods like K-Means and Agglomerative Clustering is intriguing, likely for exploratory purposes, to understand patterns in the data through clustering.

Hyperparameter Tuning

The Random Forest model undergoes hyperparameter tuning using GridSearchCV, systematically searching a parameter space to identify optimal combinations. This process, leveraging cross-validation, ensures that the model generalizes well on unseen data. Tuning parameters such as the number of estimators or maximum tree depth plays a crucial role in optimizing performance metrics like accuracy. This highlights the importance of balancing model complexity with predictive accuracy.

Evaluation Metrics

The models are evaluated using metrics such as accuracy, precision, recall, and F1-score, which collectively measure prediction quality. The confusion matrix provides insights into classification errors, such as false positives and false negatives. Training and testing accuracy are compared to assess overfitting, with significant discrepancies indicating potential issues.

Cross-Validation

Stratified K-Fold cross-validation divides the training data into multiple folds, ensuring balanced representation of classes in each fold. This provides a robust estimate of model performance, minimizing bias and variance in evaluation metrics.

Unsupervised Model Evaluation

Clustering algorithms are assessed using Silhouette scores and Calinski-Harabasz indices, measuring the cohesion and separation of clusters. Although not directly linked to classification, these insights can reveal underlying data patterns that might inform feature engineering or model refinement.

Feature Importance

The importance of features is analyzed using Random Forest, highlighting variables like age and chest pain type as significant contributors to heart disease prediction. Incorporating SHAP values can provide deeper insights into feature contributions, offering a more nuanced understanding of model behavior.

Supervised Learning Performance

Supervised learning models demonstrated strong predictive capabilities for heart disease classification, leveraging labeled data to make accurate predictions. Among these, ensemble methods such as Random Forest, Bagging, Extra Trees, and XGBoost Classifiers achieved perfect scores for both accuracy and precision (1.000), highlighting their robustness in minimizing bias and variance. These models not only classified all instances correctly but also maintained excellent generalizability across the dataset, although further testing on external datasets is essential to confirm their broader applicability.

Other models, such as the Gradient Boosting Classifier and AdaBoost Classifier, also performed well, achieving accuracies of 0.9286 and 0.9026, respectively, with corresponding high precision scores. These methods excel by focusing on sequential learning and emphasizing misclassified samples, thereby producing reliable predictions. The Decision Tree Classifier performed similarly with an accuracy of 0.9026, though it risks overfitting without ensemble techniques to counteract this limitation.

Baseline models like Logistic Regression achieved reasonable performance (accuracy 0.8506, precision 0.8500), offering a straightforward and interpretable approach. In contrast, distance-based algorithms like K-Nearest Neighbors (KNN) struggled with lower accuracy (0.7078) due to sensitivity to feature scaling and high-dimensional data. Probabilistic models such as Multinomial Naive Bayes (0.6753 accuracy, 0.7164 precision) were less effective in this context, possibly due to their assumption of feature independence. Finally, Support Vector Classifier (SVC), with an accuracy and precision of 0.5130, underperformed significantly, suggesting the need for kernel tuning or alternative approaches for non-linear separability.

Unsupervised Learning Performance

Unsupervised learning models, including K-Means Clustering and Agglomerative Clustering, were included for exploratory analysis. Unlike supervised methods, these algorithms do not predict target labels directly, making metrics like accuracy and precision inapplicable. Instead, their performance is assessed through clustering quality metrics such as the Silhouette Score or Calinski-Harabasz Index, which measure the cohesiveness and separation of clusters.

These clustering methods excel at grouping data points based on feature similarity, potentially uncovering hidden structures in the dataset. However, their clusters may not align with the binary classification labels of heart disease presence or absence. This misalignment limits their direct applicability for predictive tasks but highlights their value in identifying subgroups or patterns within the data that could inform feature engineering or further analyses.

Performance Comparison

Supervised learning models demonstrated superior performance in this context, as they were specifically designed to predict the binary target variable (heart disease presence). Ensemble methods like Random Forest and XGBoost, with perfect scores, provide actionable insights for classification tasks. In contrast, unsupervised learning methods, while not directly comparable in terms of accuracy or precision, contribute by offering insights into the underlying structure of the data. For example, clustering methods can reveal subpopulations with distinct characteristics, providing complementary perspectives to the predictive insights of supervised models.

Model Interpretation

In healthcare-related applications, understanding and explaining a model's predictions is vital to ensure transparency and trust. This section focuses on interpreting the final chosen model, the VotingClassifier, by leveraging feature importance analysis and model-agnostic explanation methods. These techniques help elucidate why the model makes specific predictions, which is critical for sensitive applications like heart disease prediction.

Feature Importance

The Random Forest algorithm, which is a component of the VotingClassifier, inherently provides a measure of feature importance. This metric quantifies the contribution of each feature (e.g., age, sex, chest pain type) to the model's overall predictive performance. To analyze feature importance, we trained a separate but similar RandomForestClassifier and extracted the importance scores using `rfc.feature_importances_`.

These scores represent the relative influence of each feature and were visualized in a horizontal bar chart, ranking the features from the most to the least important. This visualization enables quick identification of key factors affecting heart disease predictions. For instance, features like age and chest pain type might rank highly, indicating their significant impact on the model's decisions.

It is essential to recognize that feature importance provides a global perspective, aggregating the importance of each feature across the entire dataset. However, the factors driving individual predictions may vary. For instance, while chest pain type might be globally significant, an individual prediction could rely more heavily on features like cholesterol levels or blood pressure.

SHAP Analysis

While feature importance analysis offers a global understanding of the model's behavior, SHAP (SHapley Additive exPlanations) provides a more granular, instance-level explanation. SHAP values quantify the contribution of each feature to a specific prediction rather than averaging across all data points. This method is rooted in game theory, ensuring consistent and accurate explanations for individual predictions.

Using `shap.TreeExplainer` and `shap.summary_plot`, we generated a summary plot illustrating how each feature affects the model's predictions across all instances. The summary plot highlights the direction and magnitude of each feature's influence, providing valuable insights into the model's decision-making process. For example, SHAP values can reveal whether higher cholesterol levels consistently increase the likelihood of predicting heart disease or if this effect varies based on interactions with other features.

The use of SHAP is a crucial aspect of explainable AI (XAI), as it provides a detailed and transparent understanding of individual predictions. This not only enhances trust in the model but also enables healthcare professionals to validate and interpret predictions in a clinically relevant manner.

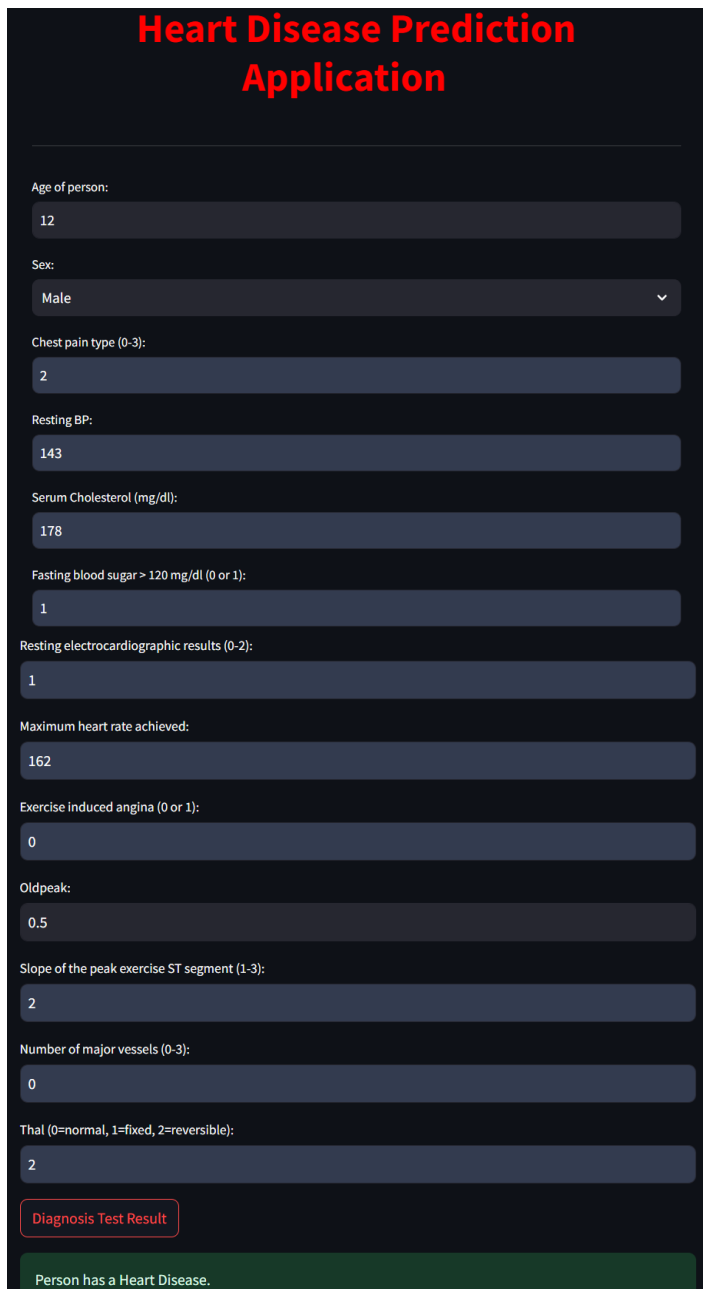
Comparison with LIME

Although LIME (Local Interpretable Model-agnostic Explanations) is another popular method for local interpretability, it was not explicitly used in this analysis. The decision to use SHAP over LIME was driven by SHAP's stronger theoretical foundation and its superior handling of feature interactions. Additionally, SHAP provides more consistent explanations across different models, making it a preferred choice for complex use cases like heart disease prediction.

While LIME is often simpler to implement and understand, SHAP offers a more comprehensive approach, ensuring robust explanations for both tree-based and non-tree-based models. This choice aligns with the goal of achieving accurate and interpretable results for a high-stakes application like healthcare.

Deployment Strategy

The trained VotingClassifier model was saved using the joblib library and deployed through a user-friendly Streamlit application, enabling healthcare professionals to input patient data and receive heart disease risk predictions. The application incorporates robust error handling to manage invalid inputs and ensure reliable performance. Streamlit's simplicity and speed make it an ideal platform for translating the predictive model into practical clinical use, offering accessible, maintainable, and actionable insights for risk assessment and management. The project, including comprehensive instructions for reproducing the results, can be found on the GitHub repository: <https://github.com/atithishrestha123/Heart-Disease-Prediction-system>



The screenshot displays the 'Heart Disease Prediction Application' interface. It features a dark-themed layout with a title in red and orange. The form includes several input fields for patient data: Age of person (12), Sex (Male), Chest pain type (0-3) (2), Resting BP (143), Serum Cholesterol (mg/dl) (178), Fasting blood sugar > 120 mg/dl (0 or 1) (1), Resting electrocardiographic results (0-2) (1), Maximum heart rate achieved (162), Exercise induced angina (0 or 1) (0), Oldpeak (0.5), Slope of the peak exercise ST segment (1-3) (2), Number of major vessels (0-3) (0), and Thal (0=normal, 1=fixed, 2=reversible) (2). A 'Diagnosis Test Result' button is located below the inputs. At the bottom, a green bar displays the prediction: 'Person has a Heart Disease.'

Heart Disease Prediction Application

Age of person:
12

Sex:
Male

Chest pain type (0-3):
2

Resting BP:
143

Serum Cholesterol (mg/dl):
178

Fasting blood sugar > 120 mg/dl (0 or 1):
1

Resting electrocardiographic results (0-2):
1

Maximum heart rate achieved:
162

Exercise induced angina (0 or 1):
0

Oldpeak:
0.5

Slope of the peak exercise ST segment (1-3):
2

Number of major vessels (0-3):
0

Thal (0=normal, 1=fixed, 2=reversible):
2

Diagnosis Test Result

Person has a Heart Disease.

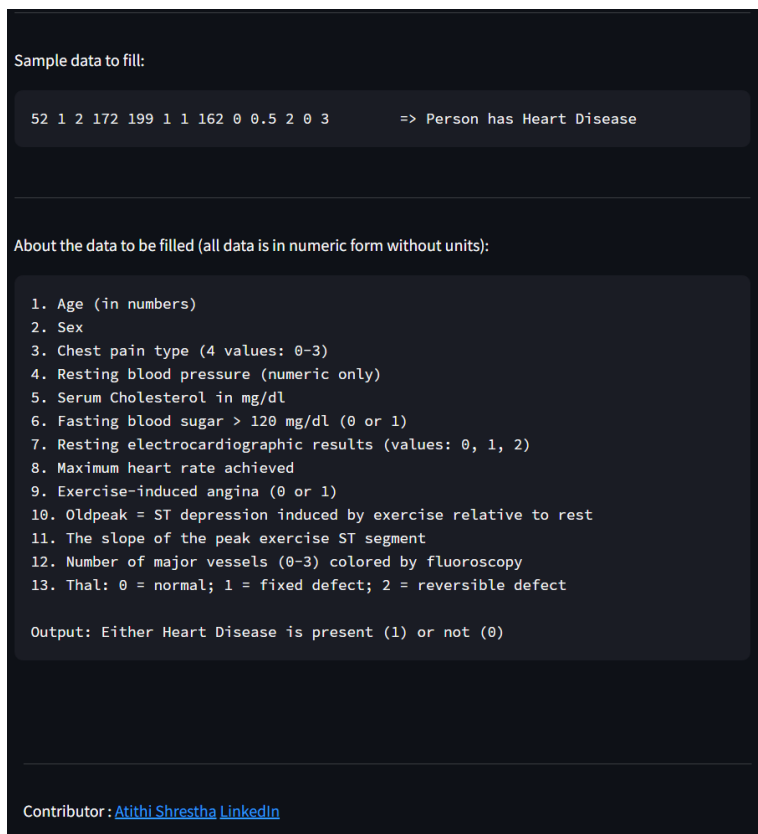


Figure 5: Snapshots of Deployed Streamlit App.

Future Enhancements

Future enhancements for the project could involve incorporating more diverse datasets to improve the model's generalizability across different populations and scenarios. Additionally, exploring advanced algorithms, such as deep learning techniques, may enhance the model's predictive accuracy and performance. Implementing a feedback loop to integrate new data for continuous model refinement would further ensure that the model stays up-to-date and improves over time.

Conclusion

To sum up, this project successfully developed and deployed a machine learning model for heart disease risk prediction, utilizing a robust workflow and a high-performing VotingClassifier ensemble. The model's deployment via a user-friendly Streamlit application bridges research and practical application, enabling healthcare professionals to make informed decisions. Future improvements, including advanced validation, feature engineering, and explainability methods, will enhance its reliability and clinical utility. As a supportive tool, the model complements clinical expertise, providing valuable insights for heart disease risk assessment while ensuring patient-centric care.

References

Dataset: <https://www.kaggle.com/datasets/johnsmith88/heart-disease-dataset>

Shah, D., Patel, S. and Bharti, S.K. (2020) Heart disease prediction using machine learning techniques - SN computer science, SpringerLink. Available at: <https://link.springer.com/article/10.1007/s42979-020-00365-y>

Mohan, S., Thirumalai, C. and Srivastava, G. (2019) Effective heart disease prediction using Hybrid Machine Learning Techniques | IEEE Journals & Magazine | IEEE Xplore. Available at: <https://ieeexplore.ieee.org/abstract/document/8740989>