Heart Disease Prediction Using Data Mining Algorithms and Patient Segmentation

Ferial Najiantabriz
Computer Science
University Of Oklahoma
Oklahoma
ferial@ou.edu

Subankar Chowdhury
Computer Science
University Of Oklahoma
Oklahoma
Subankar.Chowdhury-1@ou.edu

Ujwala Vasireddy
Computer Science
University of oklahoma
City State Country
ujwala.vasireddy-1@ou.edu

ABSTRACT

Heart disease is one of the leading causes of death worldwide, making early prediction and patient risk assessment essential for better healthcare outcomes. This project focuses on Heart Disease Prediction Using Data Mining Algorithms and Patient Segmentation. Our objectives are threefold: (1) predict heart disease risk using machine learning models like Logistic Regression, Decision Trees, and k-Nearest Neighbors (k-NN), (2) segment patients based on shared medical characteristics to identify risk factors and trends, and (3) provide a graphical interface for inputting data and visualizing predictions and segmentations.

We implemented and tested machine learning algorithms to evaluate their accuracy in predicting heart disease, achieving promising results. Feature importance analysis highlights the significance of attributes such as cholesterol and age in risk prediction. Additionally, patient segmentation revealed distinct groups based on medical profiles, which correspond to varying risk levels. Our system offers an accessible interface for medical professionals and patients, combining predictive modeling with visualization tools to enhance decision-making.

This project contributes to advancing heart disease research by integrating data mining techniques with user-friendly software, facilitating more effective diagnosis and patient management.

Introduction

Heart disease is a major global health concern, often caused by complex and interrelated risk factors such as age, cholesterol levels, and blood pressure. Despite the availability of advanced medical technologies, predicting heart disease remains challenging due to the diversity and complexity of patient data. To address this challenge, computational techniques like data mining and machine learning are increasingly being applied in healthcare to provide more accurate and actionable insights.

This project focuses on developing a system for predicting heart disease risk and segmenting patients using machine learning algorithms and clustering techniques. The implementation involves analyzing patient data using algorithms such as Logistic Regression, Decision Trees, and k-Nearest Neighbors (k-NN) for prediction tasks. Additionally, clustering methods are applied to group patients based on shared medical characteristics, which can uncover hidden trends and inform better risk management strategies.

The importance of this project lies in its practical application for improving heart disease diagnosis and treatment planning. By using real-world patient datasets, the system identifies key features influencing heart disease risk and provides an interactive graphical interface for users to input data and explore predictions and segmentations. This interface is designed to make the system accessible to healthcare professionals and researchers who may lack technical expertise.

Related Works

Heart disease prediction has been a focal point of research in both the medical and data science communities due to its significant impact on global health. The application of data mining and machine learning techniques has been extensively explored to improve the accuracy of heart disease diagnosis and risk assessment. This section reviews notable studies in this area, highlighting their methodologies, findings, and the gaps that our project aims to address.

Almustafa (2020) conducted a comparative analysis of various machine learning classifiers, including Decision Trees, Support Vector Machines (SVM), k-Nearest Neighbors (k-NN), and Neural Networks, for predicting heart disease. The study also performed sensitivity analysis to determine the impact of different features on model performance. The findings indicated that certain classifiers, particularly Decision Trees and Neural Networks, achieved higher accuracy levels. However, the study primarily focused on model performance metrics without integrating patient segmentation to identify underlying risk patterns.

Ali et al. (2021) provided a comprehensive evaluation of supervised machine learning algorithms such as Logistic Regression, Random Forest, Gradient Boosting, and Deep Learning models using the UCI Heart Disease dataset. Their research emphasized the superiority of ensemble methods like Random Forest and Gradient Boosting in terms of accuracy and robustness. Feature selection techniques were also employed to enhance model performance by eliminating redundant variables. While the study advanced the understanding of algorithm efficacy, it did not extend to patient segmentation or the development of user-friendly tools for clinical application.

Bhatla and Jyoti (2012) analyzed heart disease prediction using different data mining techniques, including Naïve Bayes, Decision Trees, and Neural Networks. They concluded that Decision Trees and Neural Networks outperformed Naïve Bayes classifiers. The study underscored the importance of data preprocessing but lacked emphasis on interactive interfaces or visualization tools to aid healthcare professionals in decision-making processes.

Chandrasekhar and Peddakrishna (2023) enhanced heart disease prediction accuracy by integrating machine learning techniques with optimization algorithms like Genetic Algorithms for feature selection. Their work demonstrated that hybrid models could yield better predictive performance than standalone models. Despite these advancements, the study did not focus on patient segmentation or the practical implementation of these models in clinical settings through accessible software interfaces.

Dey and Rautaray (2014) explored data mining algorithms for healthcare decision support systems, evaluating the efficiency and accuracy of algorithms such as SVM, k-NN, and Bayesian classifiers. The study highlighted the potential of data mining techniques to support medical professionals but did not provide solutions for patient segmentation or the integration of predictive models into user-friendly platforms for practitioners.

Soni et al. (2011) offered an overview of predictive data mining techniques for medical diagnosis, discussing classification, clustering, and association analysis in the context of heart disease prediction. While the study acknowledged the importance of clustering methods to uncover hidden patterns in medical data, it did not implement these techniques to segment patients based on medical characteristics or develop tools for visualizing these clusters.

Despite significant advancements in applying machine learning algorithms for heart disease prediction, existing research predominantly focuses on improving model accuracy through algorithm selection and feature engineering. However, notable gaps remain unaddressed. One major gap is the lack of integration of patient segmentation using clustering techniques, which can uncover hidden trends and enable personalized risk management strategies. This means that while predictions are being made, the opportunity to understand underlying patient groupings based on shared medical characteristics is often missed.

Another gap is the absence of user-friendly interfaces in most studies, limiting the practical adoption of these predictive models in clinical settings. Healthcare professionals without technical expertise may find it challenging to utilize these models without accessible graphical user interfaces (GUIs). Additionally, there is a scarcity of solutions that combine predictive modeling with visualization tools, which are crucial for enhancing the interpretability of models and supporting better decision-making by medical practitioners. This combination could significantly aid clinicians in understanding predictions and patient segments, ultimately leading to improved patient outcomes.

3. The Proposed Work and Results

We worked with the Heart Disease Prediction dataset available from the UCI Machine Learning Repository. The dataset can be accessed through the following link:
https://www.kaggle.com/johnsmith88/heartdisease-dataset
In this project, we will explore various machine learning techniques for heart disease prediction, including Logistic Regression, Decision Trees, and k-Nearest Neighbors (k-NN). We aim to gain a comprehensive understanding of these algorithms by evaluating them using important performance metrics such as accuracy, precision, and others that will be discussed in detail.

The dataset contains several features that we will consider for our machine learning models. These features are:

- Age: The age of the patient in years.
- Sex: The gender of the patient (1 = male, 0 = female).
- Chest Pain Type: The type of chest pain experienced, categorized into four types.1:Typical angina 2: Atypical angina 3: Non-anginal pain 4: Asymptomatic.
- Resting Blood Pressure: The resting blood pressure in millimeters of mercury (mm Hg).
- Serum Cholesterol: The serum cholesterol level in milligrams per deciliter (mg/dl).
- Fasting Blood Sugar: Indicates if fasting blood sugar is greater than 120 mg/dl (1 = true, 0 = false).
- Resting Electrocardiographic Results: Results of the resting electrocardiogram (ECG), represented by values: 0: Normal 1: Having ST-T wave abnormality 2: Showing probable or definite left ventricular hypertrophy
- Maximum Heart Rate Achieved: The highest heart rate achieved during exercise.
- Exercise-Induced Angina: Indicates if exercise-induced angina is present (1 = yes, 0 = no).
- Oldpeak: ST depression induced by exercise relative to rest, which is a measure of abnormal heart activity.
- Slope of the Peak Exercise ST Segment: Describes the slope of the peak exercise ST segment:1: Upsloping 2: Flat 3: Downsloping
- Number of Major Vessels Colored by Fluoroscopy: The number of major blood vessels (0–3) visualized by fluoroscopy.
- Thalassemia (Thal): A blood disorder, with values indicating: 3: Normal 6: Fixed defect 7: Reversible defect

By analyzing these features, we aim to build predictive models that can accurately assess the risk of heart disease in patients. The project will not only focus on prediction but also on understanding which features contribute most significantly to the risk, thereby providing valuable insights for medical professionals.

	age	sex	cp	trestbps	chol	fbs	restecg	thalach	exang	oldpeak	slope	ca	thal	target
0	52	1	0	125	212	0	1	168	0	1.0	2	2	3	0
1	53	1	0	140	203	1	0	155	1	3.1	0	0	3	0
2	70	1	0	145	174	0	1	125	1	2.6	0	0	3	0
3	61	1	0	148	203	0	1	161	0	0.0	2	1	3	0
4	62	0	0	138	294	1	1	106	0	1.9	1	3	2	0

Table1

3.1 Application Description

3.3 System Architecture

The proposed system is a comprehensive framework aimed at predicting heart disease risk and grouping patients into meaningful clusters. It is designed to assist medical decision-making with a focus on being interpretable and easy to understand. The system has several layers and components, each performing a specific function:

3.3.1 Data Handling

The system offers two methods for entering data:

Batch Input: Users can upload a CSV file with patient data, allowing the system to process multiple records at once for predictions and segmentation.

Manual Input: Users can enter details for a single patient to get an immediate heart disease risk prediction and segmentation result.

This dual-mode feature provides flexibility, making it suitable for healthcare providers managing both individual patient assessments and large datasets.

3.3.2 Preprocessing Layer

The preprocessing layer prepares the input data for prediction and segmentation. It includes:

- Categorical Variable Encoding: One-hot encoding is applied to variables like gender, chest pain type, and exercise-induced angina to make them machine-readable.
- Continuous Variable Normalization: Features such as age, cholesterol levels, resting blood pressure, and heart rate are standardized for consistent analysis.
- Outlier Removal: The Interquartile Range (IQR) method is used to remove outliers, improving the accuracy and robustness of the segmentation process.

3.3.3 Prediction module

The prediction module is based on a Decision Tree algorithm, implemented from scratch to ensure simplicity and interpretability. It focuses on the most important features, which we identified and implemented from scratch. This reduces the amount of input data required, making it easier and faster for users to obtain predictions. The module provides:

- Real-Time Prediction: Allows users to enter fewer features manually and receive an immediate prediction of heart disease risk for individual patients.
- Batch Prediction: Processes uploaded datasets to generate predictions for all records, while highlighting the important features that influence the results.

This feature-based approach improves user experience by simplifying the data input process and maintaining accurate predictions.

3.3.4 Patient Segmentation module

The system uses a patient segmentation feature based on a custom K-means clustering algorithm, created from scratch to allow deeper analysis and better use of resources. The best number of clusters (K) is chosen using the Elbow Method, and the algorithm groups patients into four clear categories.

By dividing patients into these meaningful clusters, the system makes it easier for healthcare providers to find patterns, make better decisions, and improve care quality while using resources wisely.

The cluster analysis divides patients into four distinct groups, each with unique health characteristics and specific recommendations.

Cluster 0 consists of older patients with high cholesterol levels and elevated blood pressure. For these patients, interventions such as low-cholesterol diets, blood pressure management, and regular checkups are advised. Cluster 1 includes younger patients with normal cholesterol but moderate exercise capacity, for whom maintaining physical activity and adopting a healthy lifestyle are recommended. Cluster 2 represents patients with lower heart rates and normal blood pressure, who would benefit from routine health monitoring. Finally, Cluster 3 represents patients with higher oldpeak values and multiple risk factors, requiring personalized care plans based on their specific health needs.

3.3.5. Visualization and Insights

The system provides extensive visualizations to support interpretability:

- **Correlation Heatmaps**: Illustrate relationships among features, aiding in understanding dependencies.
- **Confusion Matrices**: Summarize model performance across different classification tasks.
- **ROC Curves**: Evaluate and compare the discriminative abilities of the models through AUC metrics.
- **Feature Importance Plots**: Visualize the ranked importance of features identified by the Decision Tree.

These visualizations assist users in comprehending the data and model performance, fostering a transparent and informed decision-making process.

3.3.6. User Interface

The user interface is built using Streamlit to create a simple and interactive experience for users. It offers both manual and batch input options, allowing users to either enter patient data one at a time or upload a CSV file for batch processing. The system provides detailed clustering results, including group assignments and personalized health recommendations, making the data easy to understand and actionable.

The interface also features dynamic visualizations, such as pair plots for clusters and feature distributions, which give users instant feedback and a better understanding of the data. Additionally, users can download the segmented data along with recommendations as a CSV file for further use. With its clean layout and real-time responsiveness, the Streamlit interface is user-friendly and suitable for healthcare providers with limited technical skills.

3.3.7. Deployment

The system is deployed on a local **Streamlit server** for prototyping and testing. Future enhancements could include deploying the system on cloud platforms for scalability and accessibility in real-world healthcare settings.

3.4. Data Preprocessing:

Data preprocessing is an essential step in preparing the dataset for machine learning to ensure high-quality input data and improve model performance. The preprocessing steps applied to the dataset focused on handling missing values, normalization, and outlier detection to create a clean and reliable dataset for training.

Handling Missing Values: The dataset was thoroughly examined for missing values, and it was confirmed that no data was missing in any of the columns. As a result, no imputation or removal of records was

required, ensuring that the dataset was complete and ready for further preprocessing.

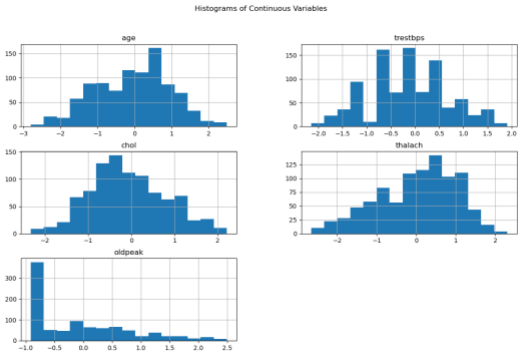
Normalization: Continuous variables, including age, trestbps (resting blood pressure), chol (cholesterol), thalach (maximum heart rate achieved), and oldpeak (ST depression), were standardized. Standardization scaled these features to have a mean of 0 and a standard deviation of 1, ensuring that variables with different scales contributed equally during model training. This step is especially important for machine learning algorithms like k-Nearest Neighbors, which are sensitive to the scale of features.

Outlier Detection and Removal: Outliers in continuous variables were identified based on values exceeding three standard deviations from the mean. Outliers were detected in trestbps, chol, thalach, and oldpeak. After removing these extreme values, the dataset size reduced from 1025 to 946 records. This cleaning step helped minimize noise and ensure that the dataset contained fewer extreme values that could negatively impact the accuracy of machine learning models.

3.5 Exploratory Data Analysis (EDA)

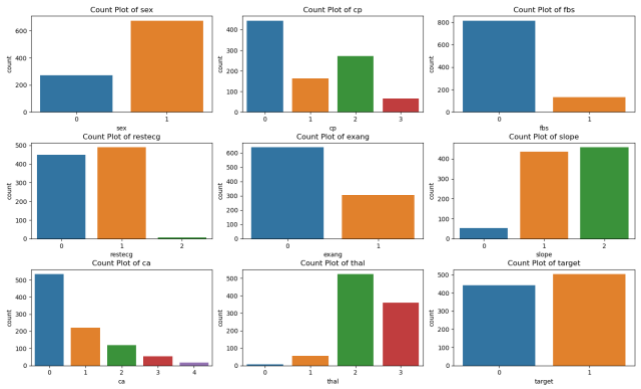
Continues Data Analysis

Figure 1 illustrates the histograms of standardized continuous variables, including age, resting blood pressure(trestbps), cholesterol(chol), maximum heart rate achieved (thalach), and ST depression induced by exercise(oldpeak). Most variables exhibit approximately normal distributions centered around zero, except for oldpeak, which shows a right-skewed distribution. These visualizations provide insights into the data's distribution and variability, highlighting potential trends and outliers for further analysis.



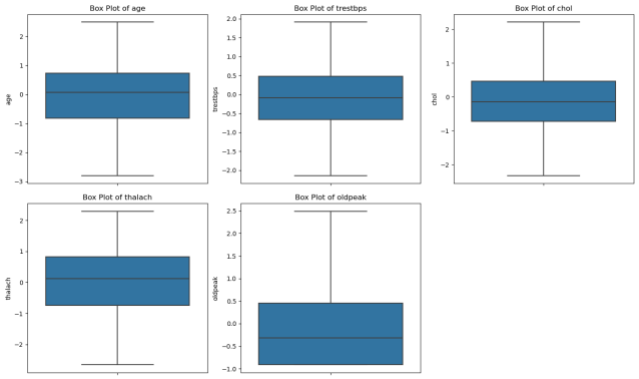
Categorical Data Analysis

Figure 2 presents count plots of categorical variables, including sex, chest pain type (cp), fasting blood sugar (fbs), restecg, exercise-induced angina (exang), slope, ca, thal, and the target variable. The dataset shows a higher proportion of male patients (sex=1) and varied distributions across other categories, such as chest pain types and thalassemia. These plots highlight the diversity in categorical variables, emphasizing the need for proper encoding during preprocessing to ensure accurate modeling and interpretation of heart disease risk.



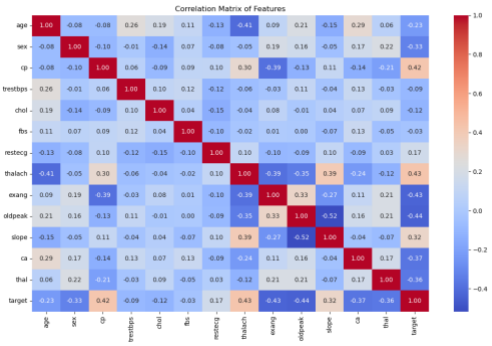
Outlier Detection

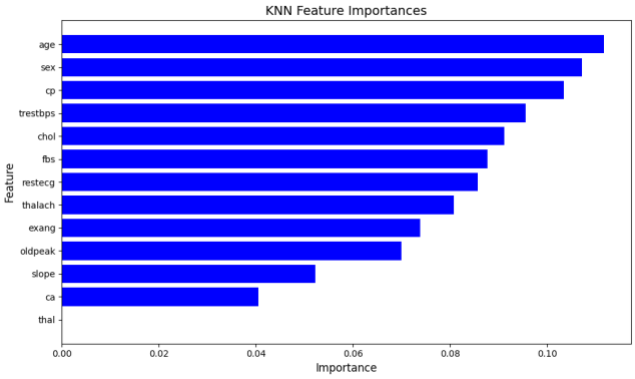
Box plots in Figure 3 highlight the presence of outliers in continuous features like trestbps and chol. These outliers were addressed during preprocessing to ensure model robustness. Detecting and managing outliers is critical to maintaining model reliability, especially in medical data where extreme values may skew predictions.



Correlation Analysis

Figure 4 illustrates the correlation matrix of features, showcasing the relationships between variables in the dataset. Notable correlations include a strong positive relationship between thalach (maximum heart rate achieved) and target (presence of heart disease), as well as a negative correlation between oldpeak (ST depression) and target. Other features, such as age and chol, exhibit weaker correlations with the target variable. These insights guide feature selection and highlight the significance of understanding variable interactions in predicting heart disease.



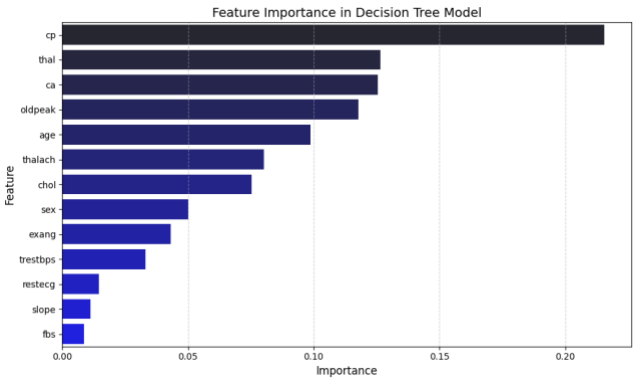


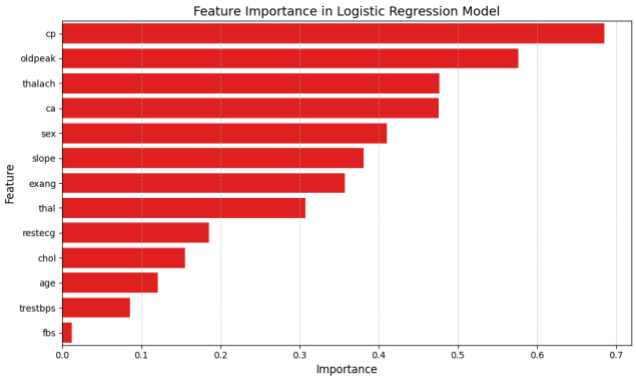
Feature Importance

The above plots show the importance of features across three models: Decision Tree (DT), Logistic Regression (LR), and K-Nearest Neighbors (KNN). Each model ranks features based on how much they contribute to predicting the target variable. This analysis helps us understand which features are most important and supports selecting the best model. After comparing these results, we chose the Decision Tree model for our study.

In the Decision Tree plot, key features like chest pain type (cp), thalassemia status (thal), and the number of major vessels (ca) are identified as the most important. These features are medically significant for diagnosing heart conditions, making the results meaningful. Logistic Regression also shows similar important features, such as cp and oldpeak, but ranks them slightly differently. The KNN model shows more evenly distributed feature importance, which makes it less effective for identifying the most critical features.

We selected the Decision Tree model because it captures non-linear relationships and interactions between features, which are important in datasets like this. Decision Trees are also easy to interpret, as they show clear splits based on features. This makes the model useful not only for achieving good predictions but also for explaining the results to non-technical audiences. Overall, the Decision Tree is a practical and reliable choice for our analysis.



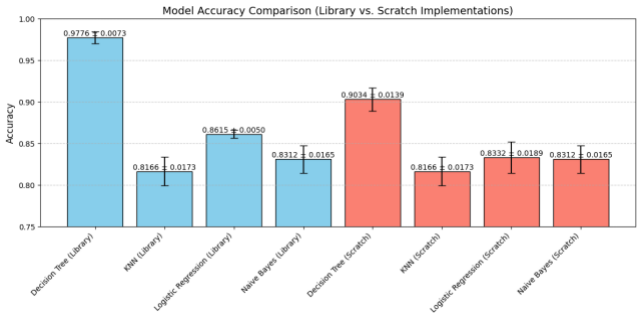


3.5 Performance Evaluation of Scratch-Implemented Models

We implemented Decision Tree, k-Nearest Neighbors (k-NN), Logistic Regression, and Naive Bayes classifiers from scratch to explore their predictive capabilities compared to their library-based counterparts. While library implementations often utilize advanced optimization techniques and pre-built frameworks, our focus was on understanding the internal mechanics of these algorithms. By manually implementing these models, we were able to study how the components and parameters directly influence classification tasks in real-world applications.

Among the scratch-implemented algorithms, Decision Tree and Logistic Regression stood out with the highest predictive accuracy. The Decision Tree classifier achieved a mean accuracy of 90.34% (std = 1.39%), showcasing its ability to manage non-linear decision boundaries effectively. Similarly, Logistic Regression obtained a mean accuracy of 83.32% (std = 1.89%), highlighting its effectiveness in capturing linear relationships between the input features and the target variable. Although these results are slightly less accurate than their library-based counterparts, they validate the correctness of our custom implementations and demonstrate their capability in achieving accurate predictions.

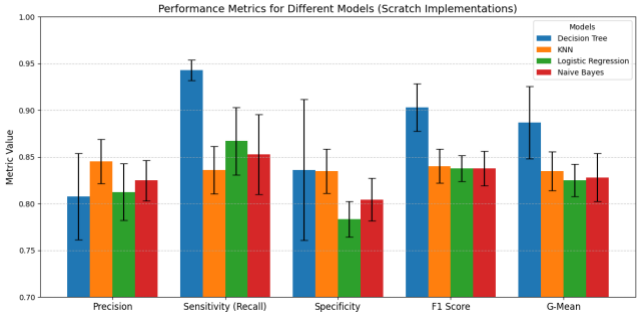
The performance metrics for these models are summarized in bar charts, showing Precision, Sensitivity (Recall), Specificity, F1 Score, and G-Mean for all four algorithms. The Decision Tree performed the best across most metrics, excelling in Precision, F1 Score, and G-Mean due to its ability to focus on key features and balance true positives and negatives effectively. Logistic Regression and k-NN also performed competitively, although k-NN exhibited slightly lower consistency when handling imbalanced data. Naive Bayes, while performing well in sensitivity and specificity, showed lower precision and F1 Score, likely due to its assumption of feature independence, which may not align perfectly with the dataset.



Description of Performance Metrics

The bar plot shows the performance metrics—Precision, Sensitivity (Recall), Specificity, F1 Score, and G-Mean—for four machine learning models implemented from scratch: Decision Tree, KNN, Logistic Regression, and Naive Bayes. Each bar represents the average value of the metric across multiple runs, and the error bars show the standard deviation, which indicates the variability in the results. The Decision Tree model performs the best in most metrics, especially in Sensitivity (Recall) and F1 Score, showing its strength in identifying positive cases while minimizing false negatives. KNN and Logistic Regression have similar performance, but their precision and sensitivity are slightly lower. Naive Bayes, on the other hand, performs worse than the other models in Specificity and F1 Score.

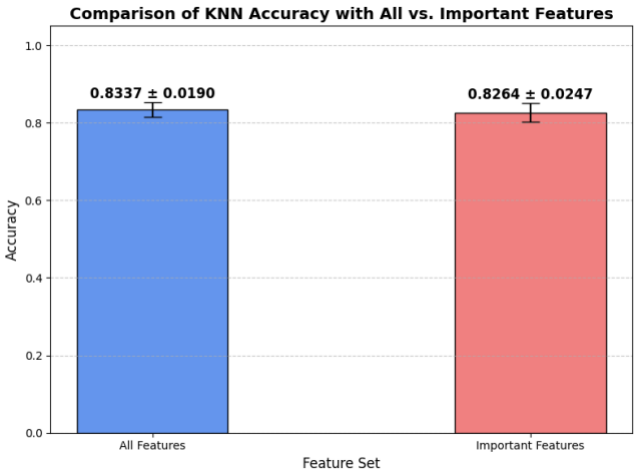
The G-Mean metric, which measures the balance between Sensitivity and Specificity, shows that the Decision Tree achieves the highest balance, making it the best overall model for this dataset. KNN and Logistic Regression have competitive G-Mean values, but they are slightly lower than the Decision Tree. Naive Bayes has the lowest G-Mean, which suggests it struggles more with this dataset, especially with handling imbalanced data. This plot provides a clear comparison of the models' strengths and weaknesses, helping to understand which model is more suitable for tasks that require high recall and balanced performance. The Decision Tree is the most reliable choice for this dataset.

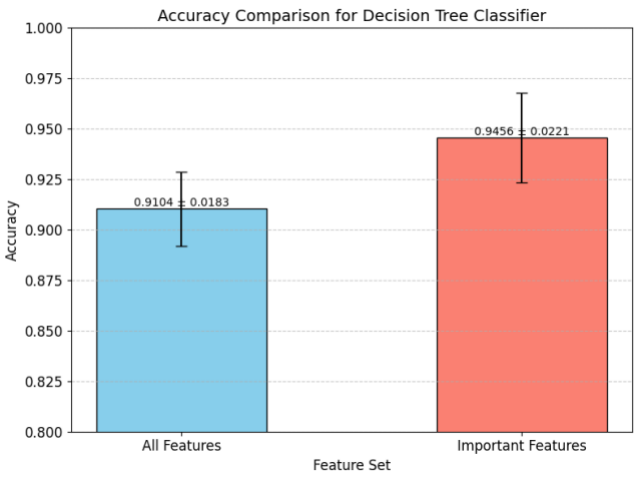


Impact of Feature Selection on Classifier Accuracy

The bar plots compare the accuracy of two classifiers: K-Nearest Neighbors (KNN) and Decision Tree, when trained with all features and when using only the most important features. For the KNN classifier, the accuracy using all features is slightly higher (0.8337 ± 0.0190) compared to the accuracy with selected important features (0.8264 ± 0.0247). This suggests that KNN benefits more from considering all available features, likely due to the nature of distance-based classification where more features provide better differentiation.

In contrast, the Decision Tree classifier shows an improvement in accuracy when using important features (0.9456 ± 0.0221) compared to all features (0.9104 ± 0.0183). This result indicates that Decision Trees are more effective at focusing on the most relevant features, which reduces noise and overfitting. Overall, these plots demonstrate how feature selection impacts the performance of different classifiers, emphasizing that the effect varies depending on the classifier's working mechanism.





3.6 Model Selection

Based on the performance metrics, the Decision Tree model was selected as the preferred model for heart disease risk prediction due to its high accuracy, interpretability, and the ability to visualize feature importance. Decision Trees allow for easy identification of key predictive features,

which can provide clinicians with insights into the factors contributing to heart disease risk.

Figure presents the confusion matrices for the Decision Tree, k-Nearest Neighbors (k-NN), Logistic Regression, and Naive Bayes classifiers, all implemented from scratch. These matrices provide an in-depth analysis of each model's predictive performance by comparing the actual and predicted labels. Among the models, the Decision Tree exhibited the highest accuracy, correctly classifying 84 instances of the negative class and 102 instances of the positive class, with only 19 misclassifications. The k-NN classifier showed competitive performance, albeit with a higher number of misclassified positive instances, indicating potential sensitivity to class distribution.

Logistic Regression and Naive Bayes demonstrated comparable results, effectively predicting the majority of positive instances while misclassifying a notable number of negative cases. Logistic Regression showed a slightly higher rate of false positives (26 instances), which could be attributed to its linear decision boundary. Naive Bayes, on the other hand, effectively captured the patterns in the data, achieving similar results with marginally fewer false positives. These confusion matrices highlight the strengths and weaknesses of each scratch implementation, providing valuable insights into their suitability for heart disease prediction tasks.

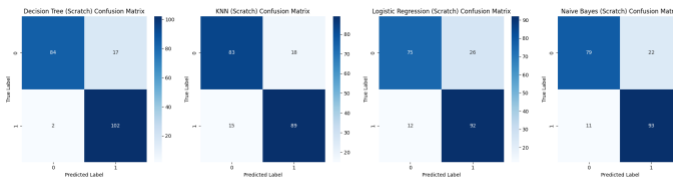
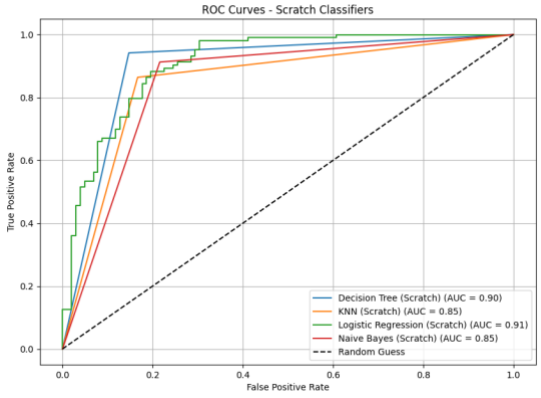
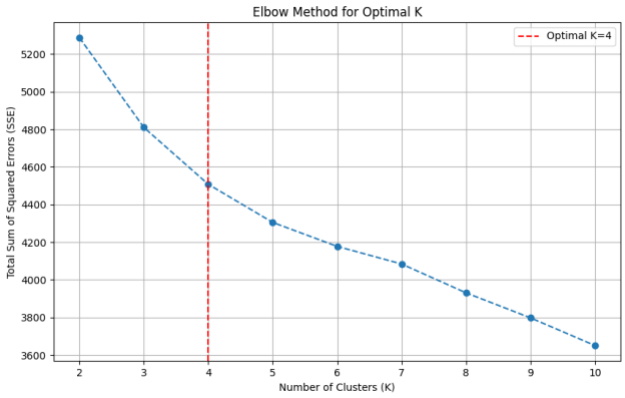


Figure illustrates the Receiver Operating Characteristic (ROC) curves for the scratch implementations of Decision Tree, k-Nearest Neighbors (k-NN), Logistic Regression, and Naive Bayes classifiers. The Area Under the Curve (AUC) values provide a comparative measure of each model's performance in distinguishing between the classes. Logistic Regression achieved the highest AUC of 0.91, closely followed by the Decision Tree with an AUC of 0.90, indicating their superior ability to balance sensitivity and specificity. Both k-NN and Naive Bayes exhibited slightly lower AUCs of 0.85, reflecting comparable but less robust performance. These curves highlight the reliability of the scratch implementations in predicting heart disease risk, with Logistic Regression and Decision Tree demonstrating the most promising results.



3.7 Determining the Optimal Number of Clusters Using the Elbow Method

The graph illustrates the Elbow Method, which is used to determine the optimal number of clusters (K) for the dataset. The x-axis represents the number of clusters, while the y-axis shows the Total Sum of Squared Errors (SSE), a measure of how tightly data points are grouped within their clusters. As the number of clusters increases, the SSE decreases, indicating improved clustering. However, after a certain point, the rate of improvement slows significantly, forming an "elbow" shape. In this case, the elbow occurs at K=4, marked by the red dashed line. This suggests that four clusters provide a good balance between minimizing SSE and avoiding overfitting, making it the optimal choice for segmentation.



3.8 Analysis of Clustered Data Using Pairwise Scatterplots

This scatterplot matrix represents the relationships between five key continuous variables in the dataset: age, trestbps (resting blood pressure), chol (cholesterol level), thalach (maximum heart rate achieved), and oldpeak (ST depression induced by exercise). The dataset has been segmented into four clusters (0, 1, 2, and 3), where each cluster is represented by a unique color. The diagonal plots show the distribution of each feature within the clusters, while the off-diagonal plots highlight pairwise relationships between the variables.

3.8.1 Cluster Characteristics Based on Age and Resting Blood Pressure

The diagonal plot for age shows that Cluster 2 (pink) is primarily composed of younger individuals, as its density is highest in the lower age range. In contrast, Cluster 3 (blue) spans a much broader age range, including older individuals. Similarly, the distribution of trestbps shows that Cluster 0 (orange) and Cluster 3 (blue) are concentrated in the lower blood pressure range, suggesting these clusters may represent healthier individuals in terms of blood pressure. On the other hand, Cluster 1 (green) has a slightly higher density in the mid-range blood pressure values, which could indicate patients with moderate blood pressure levels.

When comparing the relationship between age and trestbps, the clusters show significant overlap. However, Cluster 2 (pink) stands out as it is concentrated around younger ages with relatively low resting blood pressure. Clusters 0 and 1 are more distributed across these two features, indicating a mix of age groups and blood pressure levels in these clusters.

3.8.2 Cholesterol Levels Across Clusters

The cholesterol distribution (chol) reveals some overlap among clusters, but Cluster 2 (pink) has a sharper peak, indicating a more uniform range of cholesterol levels for its members. The scatterplots involving chol, such as chol vs. thalach, demonstrate moderate separation between clusters. Cluster 3 (blue) and Cluster 1 (green) are more spread out across cholesterol levels, suggesting a more diverse population in these clusters regarding cholesterol.

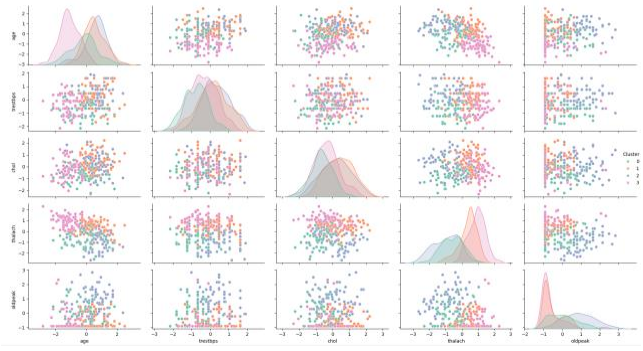
3.8.3 Exercise Capacity and ST Depression

The variable thalach represents maximum heart rate achieved, which can reflect exercise capacity. The diagonal plot for thalach shows that Cluster 0 (orange) and Cluster 1 (green) have higher densities in the mid-range of this feature, while Cluster 2 (pink) shows a concentration in the lower heart rate values, indicating potentially lower exercise capacity. The oldpeak variable, which measures ST depression induced by exercise, shows that Cluster 3 (blue) has a wider range, including individuals with both lower and higher oldpeak values, suggesting varied exercise tolerance levels within this cluster.

The scatterplot for oldpeak vs. thalach highlights that Cluster 0 (orange) is more tightly grouped in this space, while Cluster 3 (blue) has a broader distribution. This suggests that Cluster 0 may represent individuals with more consistent exercise tolerance levels, while Cluster 3 includes a wider variety of cases.

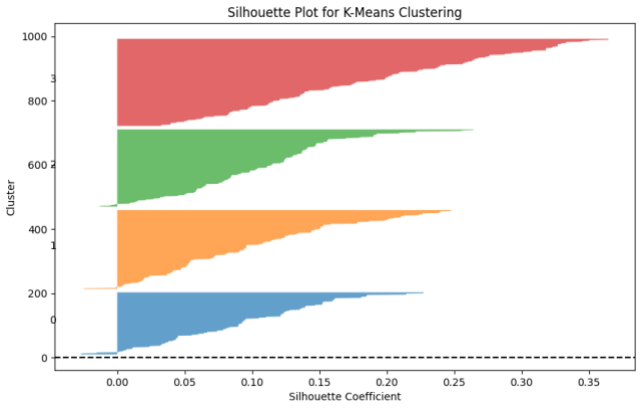
The scatterplot matrix shows that while there is some overlap between clusters in certain feature combinations, distinct patterns emerge in others. For example, Cluster 2 (pink) stands out for its younger members with lower blood pressure and cholesterol levels. Cluster 3 (blue) exhibits the most diversity, spanning a wide range of ages, cholesterol levels, and exercise-induced oldpeak values. Clusters 0 and 1 show more moderate distributions across the variables, suggesting these clusters may represent intermediate-risk groups.

This analysis provides valuable insights into the characteristics of each cluster based on key medical variables. By understanding the patterns revealed in this scatterplot matrix, healthcare professionals can better interpret the segmentation results and identify which clusters might benefit from targeted interventions. For example, younger individuals in Cluster 2 may require less intensive monitoring compared to the diverse and high-risk individuals in Cluster 3. This visualization helps bridge the gap between raw data and actionable knowledge, making it easier to develop patient-specific strategies.



3.9 Silhouette Plot for K-Means Clustering

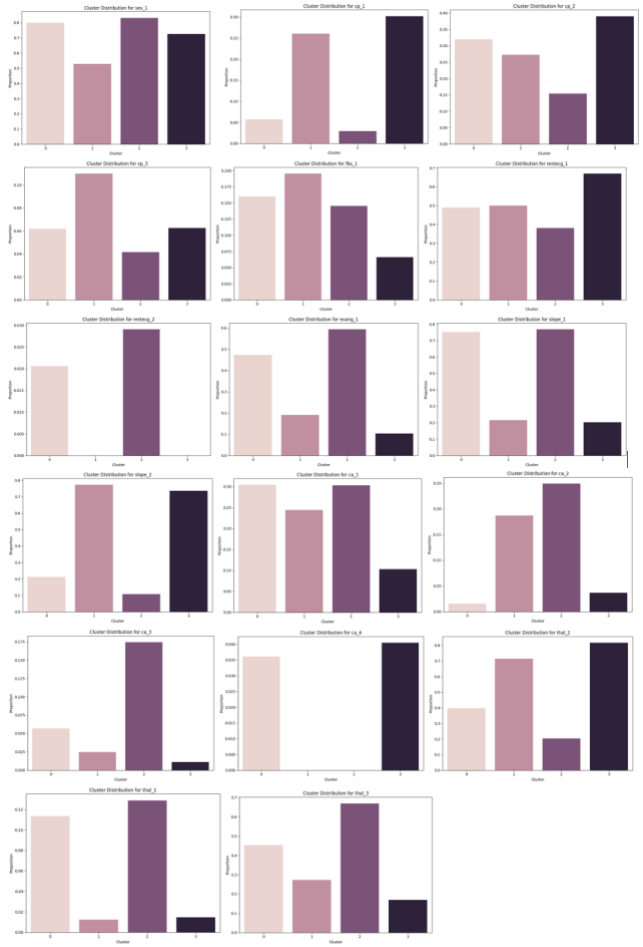
The silhouette plot above evaluates the quality of clustering by measuring how well each data point fits within its assigned cluster. The silhouette coefficient ranges from -1 to 1, where higher values mean better clustering. Each bar represents a cluster, and its thickness shows the number of data points in that cluster. In this plot, all clusters have mostly positive silhouette scores, which means the points are well-grouped. However, some points near zero indicate overlap between clusters, where some data points might not be clearly assigned. Overall, the clustering seems reasonable but could be improved by adjusting features or the number of clusters.

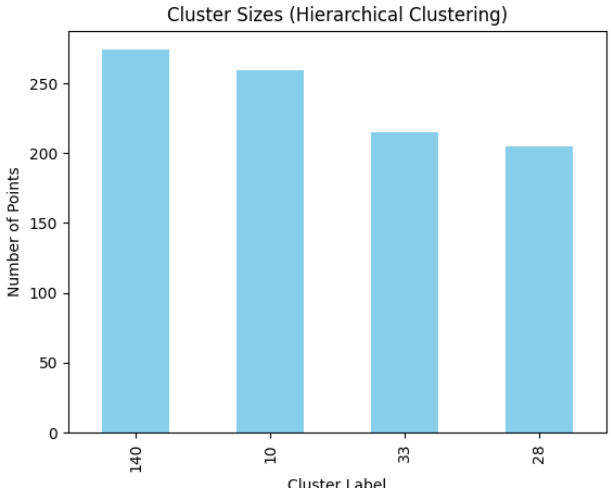


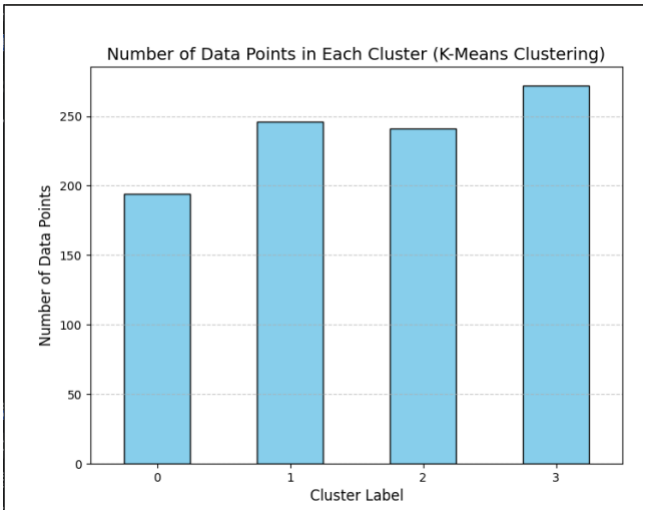
3.10 Cluster Distribution of Categorical Features

The bar charts above show how categorical features are distributed across the four clusters created by the K-means algorithm. Each chart represents a different categorical variable, such as sex, cp (chest pain type), fbs (fasting blood sugar), and others. The x-axis displays the cluster labels (0 to 3), and the y-axis shows the proportion of each feature within each cluster.

The visualizations clearly indicate that the distributions of categorical features differ among the clusters. For example, some clusters, like Cluster 2 or Cluster 3, have higher proportions of specific categories, which shows that each cluster has unique characteristics. This means the clusters are identifying meaningful patterns in the data, such as variations in chest pain types or fasting blood sugar levels. These findings can help better understand the traits of each group and support more personalized medical recommendations. Overall, the bar charts effectively demonstrate the relationship between categorical variables and the identified clusters.





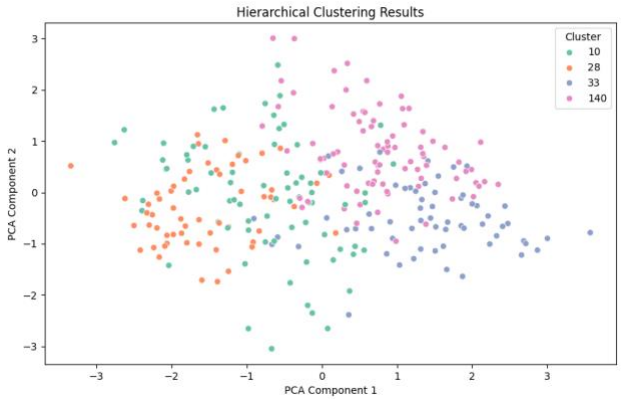


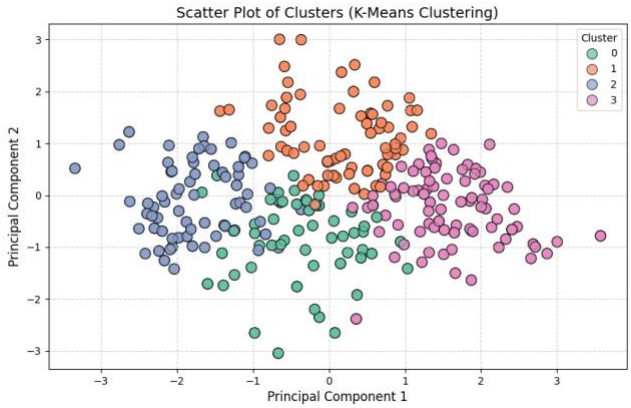
3.11. **Cluster Distribution Analysis and Justification for Choosing K-Means**

The bar plots above illustrate the distribution of data points across clusters generated by K-Means and Hierarchical Clustering. The K-Means plot shows a relatively balanced allocation of data points among the four clusters, indicating its ability to group data into similarly sized, distinct clusters. In contrast, the Hierarchical Clustering plot highlights slight imbalances in cluster sizes, reflecting its focus on hierarchical relationships rather than balanced segmentation. K-Means was chosen for this analysis due to its computational efficiency and suitability for handling large datasets with predefined cluster numbers. Additionally, its iterative optimization ensures compact and well-separated clusters, making it more appropriate for our dataset, where balanced and interpretable segmentation is critical for understanding patient group characteristics.

The scatter plots show the clustering results of K-Means and Hierarchical Clustering using two dimensions created by Principal Component Analysis (PCA). Each point represents a data record, and the colors show the clusters. In the K-Means plot, the clusters are more separated and balanced, which means K-Means worked well in grouping similar data points. In the Hierarchical Clustering plot, there is more overlap between clusters, making it harder to see clear groups.

I chose K-Means because it is faster and works better with larger datasets. K-Means also improves the clusters during its process, which helps find better patterns in the data. The scatter plot for K-Means shows better separation between clusters, proving that it is a better choice for this dataset.


Hierarchical Clustering Results
Cluster
10
28
33
140
PCA Component 2
PCA Component 1


Scatter Plot of Clusters (K-Means Clustering)
Cluster
0
1
2
3
Principal Component 2
Principal Component 1

REFERENCES