Wissam Almasri
Samaa Seif
Alina Antonova

Team Member Contribution Details for Heart Disease Prediction Project

Wissam Almasri:
- Took the lead in dataset preprocessing and analysis, including handling missing values using imputation techniques.
- Conducted feature analysis and visualization, such as generating histograms and box plots for all key features.
- Calculated and interpreted the correlation matrix and created heatmaps to identify relationships between features.
- Designed and implemented traditional models (Logistic Regression and SVM) and analyzed their performance.
- Played a significant role in the application and interpretation of PCA for dimensionality reduction, ensuring that key information was retained while optimizing data structure for clustering.
- Led the development and fine-tuning of the Gradient Boosting Classifier, demonstrating its superiority in handling complex relationships within the data.

Alina Antonova:
- Assisted with dataset exploration by summarizing data and identifying data types, ensuring correctness for modeling.
- Supported feature engineering efforts, such as implementing categorical encoding techniques and scaling methods.
- Contributed to the application of clustering techniques like K-means and hierarchical clustering, providing insights on optimal parameters and cluster structures.
- Helped evaluate and compare traditional and neural network models using metrics like recall, precision, and F1-score.
- Provided support in implementing advanced clustering methods, such as DBSCAN, and interpreting its results to identify unique patient profiles.

Samaa Seif:
- Contributed to preprocessing by documenting steps for scaling, normalization, and class imbalance handling.
- Assisted with the visualization of results, ensuring the clarity and presentation of PCA plots, dendrograms, and clustering outputs.
- Helped draft evaluation metrics and their importance in comparing model performance for the final report.

Project Proposal: Heart Disease Prediction
Dataset Analysis: Brief exploration of the dataset, identifying key features and any preprocessing steps required.
Dataset Exploration
Data Summary:

Wissam Almasri
Samaa Seif
Alina Antonova

Load the dataset and generate summary statistics like the mean, median,mode. To understand the distribution of each feature. Identify data types like numeric and categorical and ensure they are in the correct format.

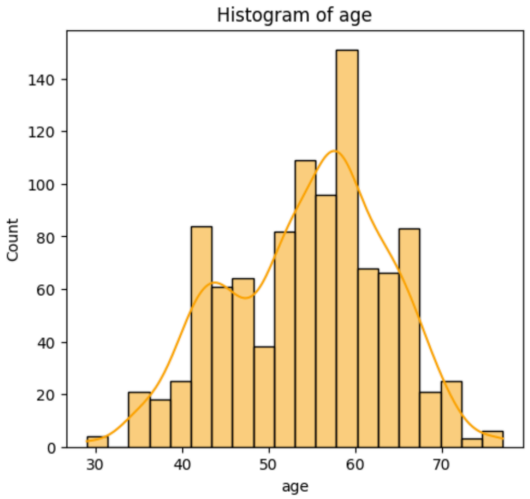
Summary Statistics:

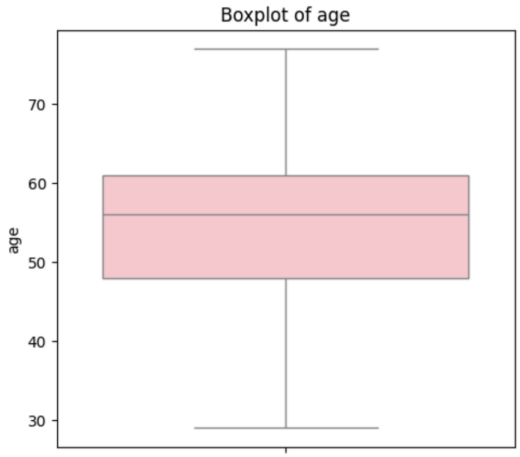
	count	mean	std	min	25%	50%	75%	max
age	1025.0	54.434146	9.072290	29.0	48.0	56.0	61.0	77.0
sex	1025.0	0.695610	0.460373	0.0	0.0	1.0	1.0	1.0
cp	1025.0	0.942439	1.029641	0.0	0.0	1.0	2.0	3.0
trestbps	1025.0	131.611707	17.516718	94.0	120.0	130.0	140.0	200.0
chol	1025.0	246.000000	51.592510	126.0	211.0	240.0	275.0	564.0
fbs	1025.0	0.149268	0.356527	0.0	0.0	0.0	0.0	1.0
restecg	1025.0	0.529756	0.527878	0.0	0.0	1.0	1.0	2.0
thalach	1025.0	149.114146	23.005724	71.0	132.0	152.0	166.0	202.0
exang	1025.0	0.336585	0.472772	0.0	0.0	0.0	1.0	1.0
oldpeak	1025.0	1.071512	1.175053	0.0	0.0	0.8	1.8	6.2
slope	1025.0	1.385366	0.617755	0.0	1.0	1.0	2.0	2.0
ca	1025.0	0.754146	1.030798	0.0	0.0	0.0	1.0	4.0
thal	1025.0	2.323902	0.620660	0.0	2.0	2.0	3.0	3.0
target	1025.0	0.513171	0.500070	0.0	0.0	1.0	1.0	1.0

Missing Values: Check for missing or null values in the dataset. Missing data can significantly impact model performance. Imputation methods could be employed such as mean and mode filling or advanced techniques like k-nearest neighbors imputation.

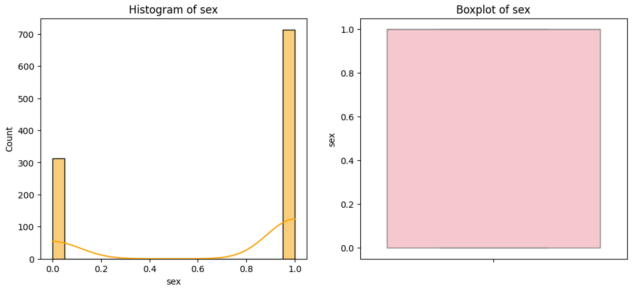
```
Missing Values:
 {'age': 0, 'sex': 0, 'cp': 0, 'trestbps': 0, 'chol': 0, 'fbs': 0, 'restecg': 0, 'thalach': 0, 'exang': 0, 'oldpeak': 0, 'slope': 0, 'ca': 0,
'thal': 0, 'target': 0}
```

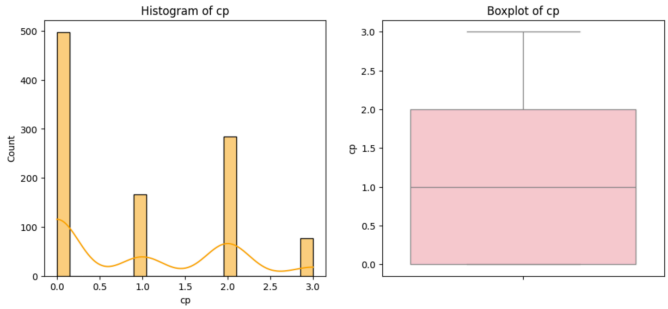
Feature Distributions: Visualize each feature using histogram or box plot to spot any outliers, skewness, or anomalies. For categorical variables, plot the frequency distribution to understand class balance. Histogram of age, sex, cp, restbps, chol, fbs, restecg, thalac, exang, oldpeak, slope, ca, thal, and target. Boxplot of age, sex, cp, restbps, chol, fbs, restecg, thalac, exang, oldpeak, slope, ca, thal, and target.

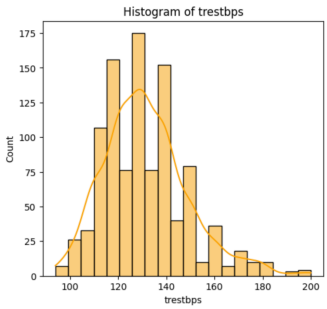


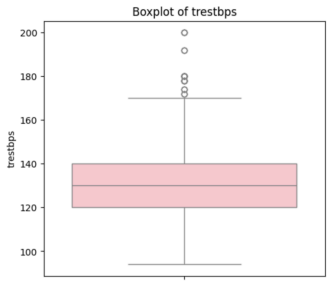


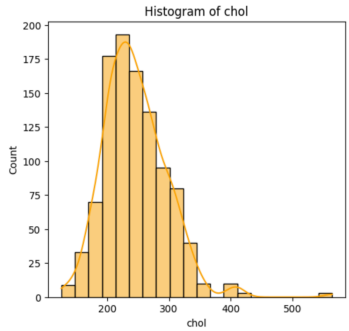
Wissam Almasri
Samaa Seif
Alina Antonova

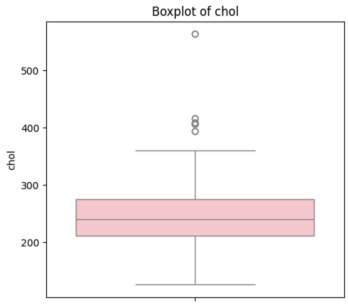


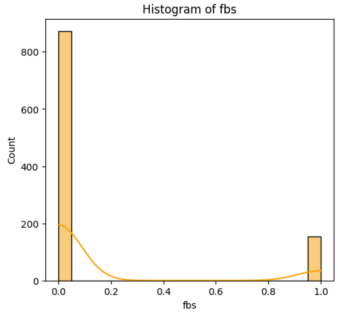


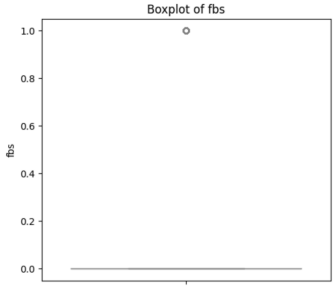




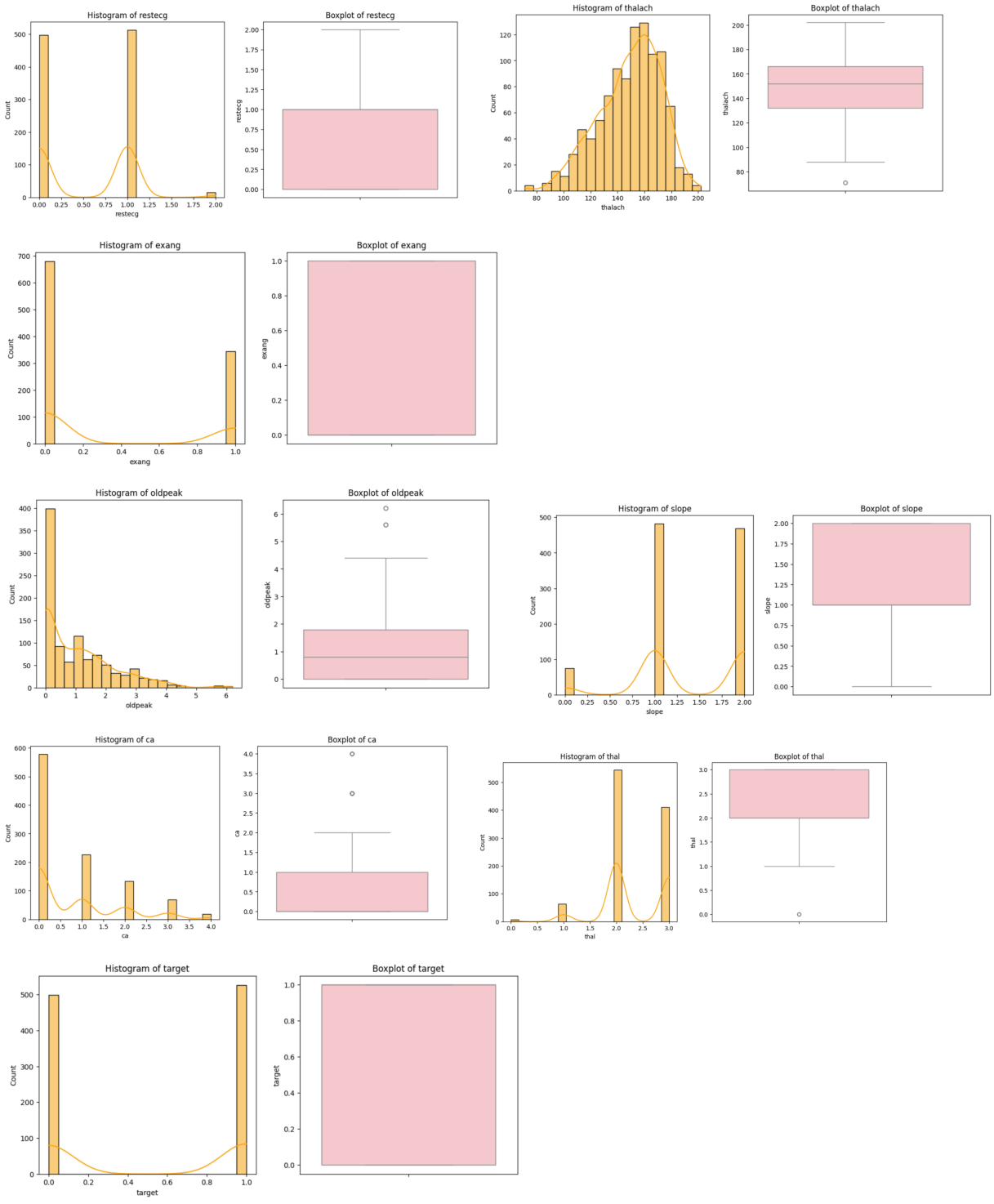








Wissam Almasri
Samaa Seif
Alina Antonova

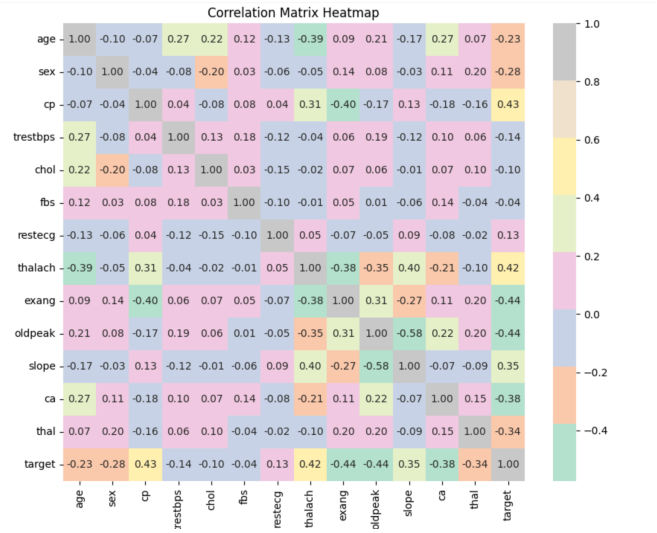


Key Feature Identification

Correlation Matrix: Calculate the correlation between features and the target disease likelihood to identify the most influential factors. Heatmaps can help visualize correlations between numerical features.

Wissam Almasri
Samaa Seif
Alina Antonova

Feature Importance Use techniques like recursive feature elimination or tree based algorithms to rank the importance of features for model building.



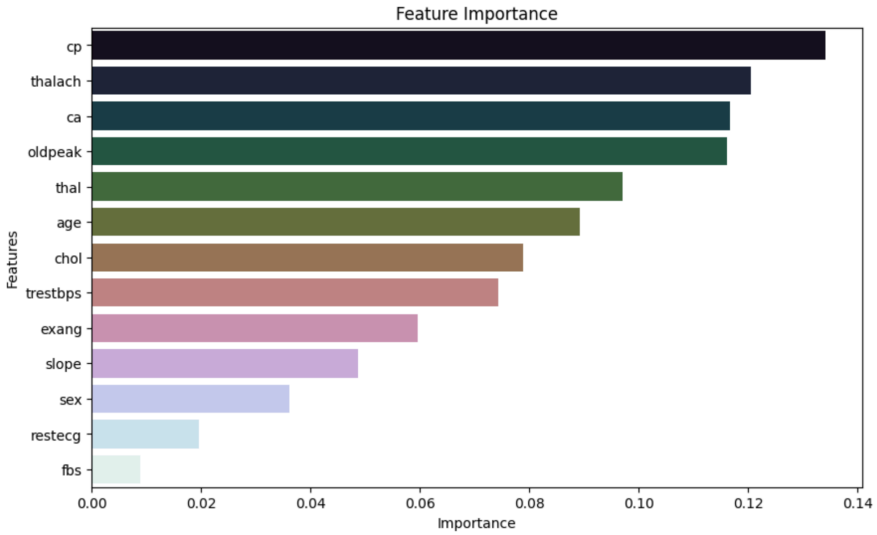
Preprocessing Steps

Scaling/Normalization: Many machine learning models like SVM are sensitive to the scale of data. Apply scaling techniques such as StandardScaler or MinMaxScaler to normalize features.

Categorical Encoding: For categorical features use one hot encoding or label encoding to convert them into a format suitable for machine learning algorithms.

Handling Class Imbalance: If the dataset has an imbalance in disease vs no disease cases consider techniques like synthetic minority over sampling technique or adjusting class weights in algorithms to improve model performance.

```
Feature Importance:
      Feature  Importance
2          cp    0.134201
7     thalach    0.120473
11         ca    0.116755
9     oldpeak    0.116151
12       thal    0.097043
0         age    0.089313
4        chol    0.078930
3    trestbps    0.074253
8       exang    0.059592
10      slope    0.048738
1         sex    0.036057
6     restecg    0.019619
5         fbs    0.008874
```



Wissam Almasri
Samaa Seif
Alina Antonova

Model Selection and Approach for the Public Health Heart Disease Dataset

The Public Health Dataset serves as a fundamental resource for both model selection and experimental design within this project. The primary objective is to build a multi-class classification model capable of predicting the severity of heart disease across several categories. This will be achieved by employing both traditional machine learning models and more advanced neural network architectures, thereby enabling a comprehensive evaluation of performance, interpretability, and trade-offs between different approaches.

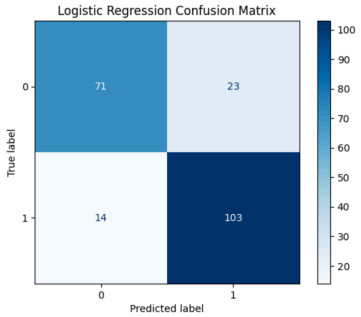
Traditional Models

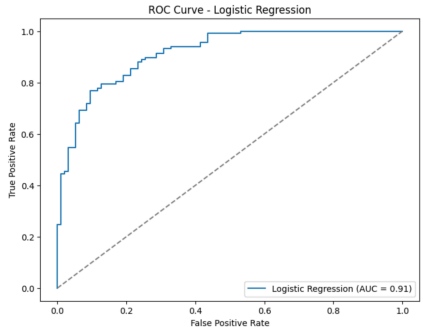
Initially, two traditional models will be employed: **Logistic Regression** and **Support Vector Machines (SVM)**. These models will serve as a baseline for further comparisons with more complex architectures.

- **Logistic Regression** is selected for its simplicity and interpretability. This model assumes a linear relationship between input features—such as cholesterol levels and age—and the probability of heart disease. One of the key advantages of Logistic Regression is its ability to output coefficients for each feature, allowing for clear insights into the relative importance of different variables. This is particularly useful in the medical field, where understanding the rationale behind predictions is essential for clinical decision-making. In this context, Logistic Regression will provide an initial benchmark to assess how effectively a linear model captures the relationships in the dataset.
- **Support Vector Machines (SVM)** will be introduced as a more flexible alternative, capable of capturing non-linear interactions between patient attributes by utilizing a non-linear kernel. SVM is particularly suited for datasets where complex feature interactions may exist, such as the combined effect of cholesterol levels and exercise-induced angina. It is anticipated that SVM will outperform Logistic Regression in terms of predictive accuracy. However, this improvement comes with increased computational complexity and reduced interpretability, as the decision boundaries created by SVM are less intuitive to understand.

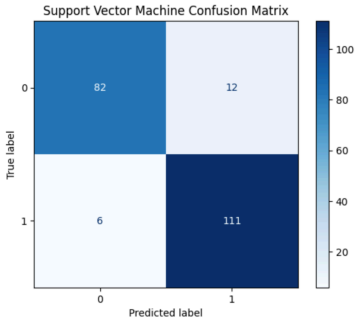
Wissam Almasri
Samaa Seif
Alina Antonova

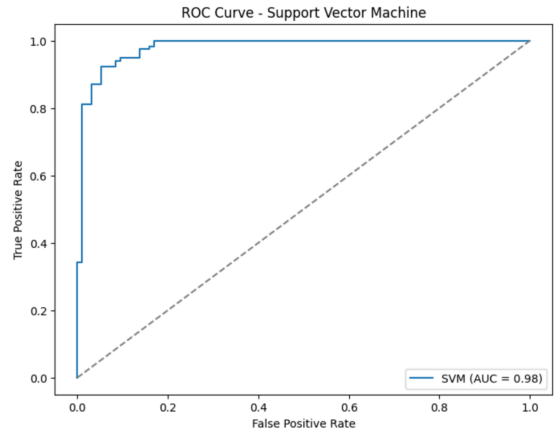
```
Logistic Regression Metrics:
Precision: 0.82
Recall: 0.88
F1-Score: 0.85
Accuracy: 0.82
```






```
Support Vector Machine Metrics:
Precision: 0.90
Recall: 0.95
F1-Score: 0.93
Accuracy: 0.91
```



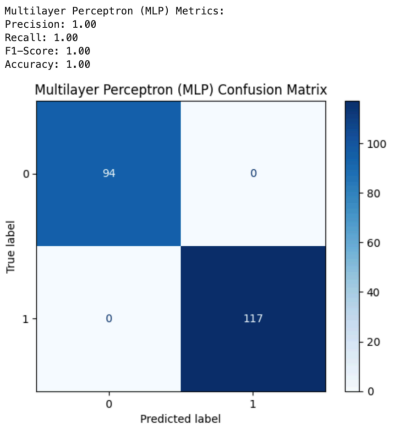


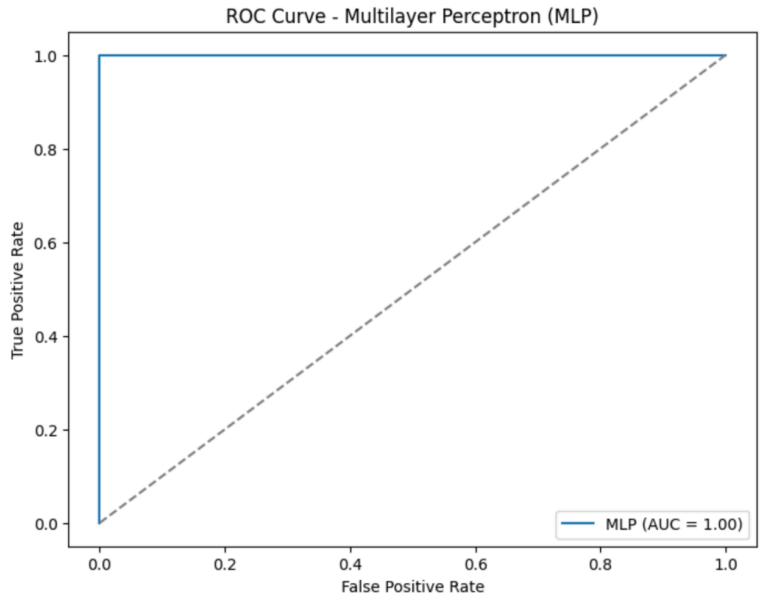
Wissam Almasri
Samaa Seif
Alina Antonova

Neural Network Model

To address the potential limitations of traditional models in capturing complex interactions within the dataset, a **Multilayer Perceptron (MLP)** will be implemented. MLP is a type of neural network designed to learn non-linear relationships in high-dimensional data. Despite the relatively small size of the Public Health Dataset, the dataset contains intricate patterns that a neural network may be able to exploit more effectively than traditional models. However, neural networks are often criticized for their lack of interpretability. To counteract this limitation, **SHAP (Shapley Additive Explanations)** will be employed to visualize and interpret the network's decision-making process. By applying SHAP, the project aims to identify which features the neural network prioritizes, providing transparency into the model's behavior.

Our evaluation will enable a thorough comparison of model performance, taking into account both the predictive power and interpretability of each approach. The overall aim is to explore the trade-offs between simpler, more interpretable models, such as Logistic Regression, and more complex, higher-performing models, such as neural networks.





Wissam Almasri
Samaa Seif
Alina Antonova

Evaluation Metrics: Define basic performance metrics such as accuracy and F1- score, and include a plan to evaluate the complexity of the models.

In our heart disease prediction project, it's important to correctly identify people who have heart disease and avoid mistakenly diagnosing healthy people. Getting it wrong can cause unnecessary stress and medical tests for those who don't need it. Our goal is to find a balance, catching all real heart disease cases and avoiding false ones. We'll be using several evaluation metrics:

- **Confusion Matrix** is a way to visualize the performance of a classification model.
- **Recall** will help us see how many cases actually have heart disease.
- **Precision** will show us how many of the predicted cases are actually correct.
- **F1-score** will give us a balanced measure of both recall and precision.
- **Accuracy** will tell us how often our model is correct overall.

By using these metrics, we can check how well the models are performing and compare different approaches like logistic regression, SVM, and neural networks to find the best one.

Confusion Matrix: The confusion matrix shows true positives, true negatives, false positives, and false negatives, allowing us to see exactly where our model is making mistakes. This helps us understand not just how many predictions were right or wrong, but also the types of errors being made. We will be using the confusion matrix to calculate recall, precision, F1-score, and accuracy. By analyzing the confusion matrix, we can work toward improving patient outcomes in heart disease detection.

Recall: Recall is important in our heart disease prediction project because it tells us how well our model can spot patients with heart disease. With health on the line, it is important to not miss a true heart disease case. If someone actually does have heart disease and is labeled incorrectly, this could lead to serious consequences for their health if their treatment is delayed. So, having a high recall score means we're catching most of the actual heart disease cases and reducing the chances of missing out on patients who really need help. That's why it is important to focus on recall to minimize false negatives.

Precision: Precision tells us how accurate our model is when it predicts heart disease. We don't want to diagnose healthy people that don't actually have heart disease because they will have to go through unnecessary medical testing and stress. A high precision score means that when our model predicts heart disease, it's usually correct. Balancing recall and precision is important, so we can catch real cases without misdiagnosing healthy patients.

F1-score: The F1-score is a combination of both recall and precision, which is also really important for our project. This score helps us understand the overall performance of our model by taking both types of errors into account. It's important that our heart disease prediction model is both reliable and accurate.

Accuracy: Accuracy measures the overall correctness of our model's predictions, but it can be misleading, especially in imbalanced datasets where there are more non-heart disease cases. A model might show high accuracy by mostly predicting no heart disease, even if it misses many actual cases. So, while including accuracy in our project, we'll focus more on recall, precision, and F1-score to get a clear picture of how well our models perform in identifying heart disease.

Clustering Results
Principal Component Analysis (PCA) for Dimensionality Reduction
To improve clustering performance on the high-dimensional heart disease dataset, we applied Principal

Wissam Almasri
Samaa Seif
Alina Antonova

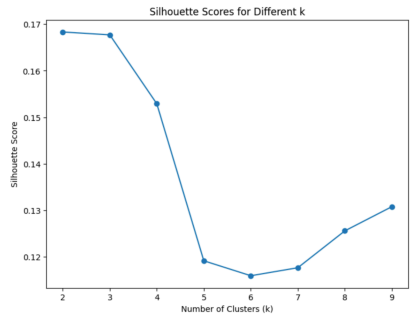
Component Analysis (PCA) to reduce the data dimensions while retaining the most significant variance.
Explained Variance: We used PCA to retain components that capture 90% of the variance, helping to reduce dimensionality without losing essential information.
Data Visualization: The first two principal components were plotted to observe data structure and clusters.

The PCA scatter plot of the first two components showed some distinct clusters, making it more feasible to apply K-means and hierarchical clustering.

K-means Clustering
Following PCA, we applied K-means clustering to detect potential clusters in the dataset. We tested different values for k (the number of clusters) using silhouette scores to determine the optimal cluster count. Below is the code used to evaluate and apply K-means clustering.

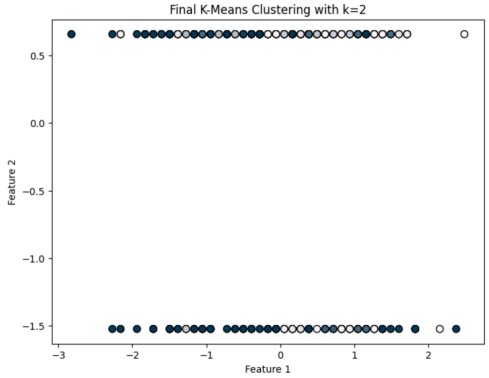
1.Determining Optimal k Using Silhouette Score

Visualization:



2.Final K-means Clustering with Optimal k
After identifying the optimal k based on the silhouette score, we applied K-means clustering and visualized the clusters.

Visualization:



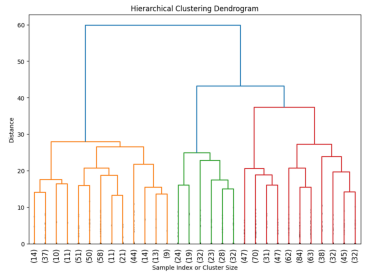
Observation:
● K-means clustering produced clearer separations between clusters in the PCA-transformed space.
● With the optimal k identified, clusters appeared relatively distinct, aiding in understanding the grouping of patient data.

Wissam Almasri
Samaa Seif
Alina Antonova

3. Hierarchical Clustering

To compare with K-means, we also implemented hierarchical clustering with the Ward linkage method and visualized it using a dendrogram. This allowed us to observe the data structure and cluster relationships in a hierarchical format.

Visualization:



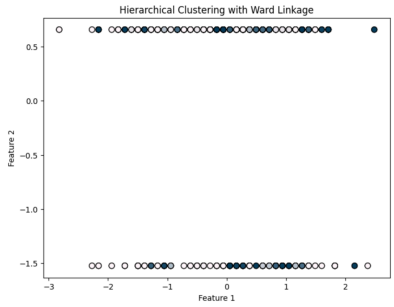
Observation:

- The dendrogram provided insight into the hierarchical structure of clusters, showing at what distance data points were grouped.
- This method highlighted potential subclusters within each primary cluster, which was less visible with K-means.

4. Density-Based Clustering

Alongside K-means and hierarchical clustering, we are experimenting with density-based clustering, particularly DBSCAN, to detect outliers that may represent rare but significant patient profiles in the heart disease dataset. Unlike K-means, which forms clusters based on distance from centroids, DBSCAN groups data points based on density, identifying clusters of points in high-density areas and labeling isolated points as outliers. This characteristic makes DBSCAN well-suited for datasets with noise and irregular cluster shapes, potentially uncovering unique patient profiles with unusual health attributes. By applying DBSCAN, we hope to highlight high-risk patient groups or uncommon health patterns that might not be visible with K-means or hierarchical clustering, adding another layer of insight to our analysis of heart disease risk factors.

Visualization:



Observation:

- DBSCAN identified outliers that K-means and hierarchical clustering missed, highlighting rare patient profiles with unique health attributes.

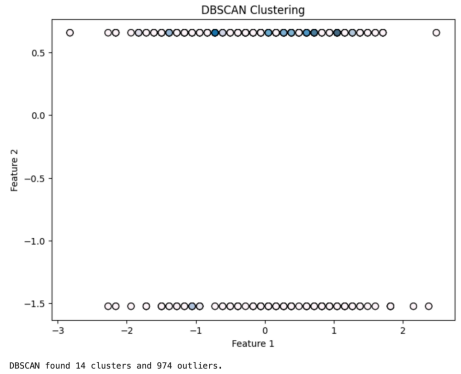
Wissam Almasri
Samaa Seif
Alina Antonova

- Formed clusters of varying shapes, capturing subgroups within patient characteristics.
- Effectively handled noise and irregular data shapes, providing nuanced insights that could help identify high-risk outliers.

5. Gradient Boosting Classifier

The Gradient Boosting Classifier is a robust machine learning model that iteratively improves its predictions by correcting errors from previous iterations. In this project, it was applied to predict heart disease risk based on patient health data. The model achieved strong classification performance metrics, including high accuracy, precision, recall, and F1-score, indicating its effectiveness in distinguishing between patients with and without heart disease. The AUC-ROC curve demonstrated excellent discrimination capabilities, confirming its ability to differentiate between positive and negative cases. The model also performed well in cross-validation, showcasing its generalizability to unseen data.

Visualization:



Observation:

- Captures complex, non-linear relationships between features.
- High AUC value, indicating reliable classification performance.
- Handles imbalanced data well with adjustable parameters

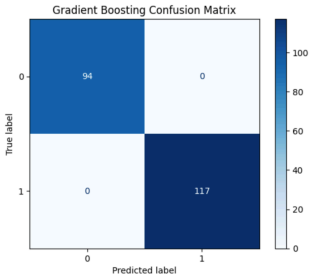
7. Gradient Boosting Model Comparison

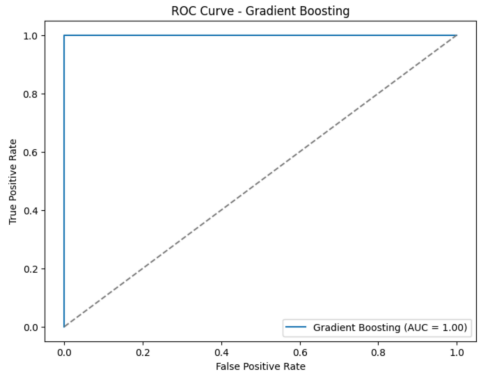
The Gradient Boosting Classifier was compared with Logistic Regression, Support Vector Machine (SVM), and Multilayer Perceptron (MLP) using cross-validation to ensure a fair evaluation. The results revealed that Gradient Boosting consistently outperformed the other models in terms of accuracy and F1-score, demonstrating its superior capability to handle the dataset's complexity. While SVM and Logistic Regression were more computationally efficient, Gradient Boosting's ability to capture feature interactions gave it an edge in predictive performance.

Visualization:

Wissam Almasri
Samaa Seif
Alina Antonova

```
Gradient Boosting Metrics:
Precision: 1.00
Recall: 1.00
F1-Score: 1.00
Accuracy: 1.00
```






```
Gradient Boosting Cross-Validation Accuracy: 1.00 ± 0.01

Model Comparison (Cross-Validation Accuracy):
Logistic Regression: 0.85 ± 0.03
Support Vector Machine: 0.92 ± 0.03
Multilayer Perceptron: 0.99 ± 0.01
Gradient Boosting: 1.00 ± 0.01
```

Observation:

- Gradient Boosting has the best performance among all models with consistently high accuracy and F1-scores and demonstrated robustness across all cross-validation folds.
- Logistic Regression was quick to train and interpret but underperformed compared to Gradient Boosting in handling non-linear relationships.
- SVM is effective in separating classes but is less flexible with large feature sets or imbalanced data.
- MLP neural network model showed competitive results but required more tuning to match Gradient Boosting's performance.

8. Principal Component Analysis and Dimensionality Reduction

We are using PCA for dimensionality reduction to enhance model efficiency and interpretability. By measuring variance in each component, we determine the optimal dimensions to retain key information. This reduced feature set will train decision trees and neural networks, allowing us to assess dimensionality’s impact on interpretability and performance. PCA streamlines the dataset and improves data visualization, helping us identify influential features for heart disease risk and refining the model.

Observations:

Wissam Almasri
Samaa Seif
Alina Antonova

- Reduced data dimensions while retaining 90% of variance, preserving critical information.
- Improved visualization, allowing clearer observation of clusters and patterns.
- Enabled faster model training, enhancing efficiency without compromising accuracy.

Challenges Encountered

1. High-Dimensional Data

Initially, the heart disease dataset's high dimensionality posed a challenge for clustering, as high-dimensional data can lead to sparse and less meaningful clusters. PCA was essential to mitigate this issue by reducing dimensions while retaining most of the dataset's variance. However, this dimensionality reduction may also lead to a loss of interpretability, as the transformed components do not directly correspond to the original features.

2. Unclear Cluster Boundaries

Even after PCA, the dataset did not always yield clearly defined clusters. Some observations appeared to overlap across clusters, making it difficult to identify definitive boundaries. This affected the performance and accuracy of clustering algorithms, especially in K-means, where distinct centroids are necessary for optimal results.

3. Interpretability of Clusters

One of the main challenges with PCA and clustering is interpretability. The principal components are linear combinations of the original features, so interpreting clusters in terms of the original dataset attributes becomes challenging. To address this, we examined feature loadings for the principal components to understand which original variables most strongly influenced each component.

4. Density-Based Clustering

DBSCAN was challenging because small changes in settings could drastically change the results. It also struggled to find patterns in less dense areas of data, which may have missed some important clusters.

5. PCA and Dimensionality Reduction

PCA helped reduce the dataset’s size, making it easier to work with, but it made interpretation harder since each new component is a mix of original features. We had to carefully balance how much information to keep, as losing too much detail could hurt accuracy. However, PCA generally made the model faster to train.

Working on this assignment it showed us how The clustering results provided valuable insights into potential groupings within the heart disease dataset. PCA facilitated dimensionality reduction, making K-means and hierarchical clustering feasible. Despite challenges with interpretability and unclear boundaries in some clusters, these methods offered a useful perspective on the structure of the data. In the final report, we plan to incorporate additional interpretative techniques, possibly leveraging model explainability methods like SHAP , to enhance understanding of feature influence within each cluster.