

**Machine Learning**

Department of Computer Science

Project Report

**Heart Disease Prediction**

Submitted By

Mones Nazih Ksasbeh



**INTRODUCTION**

Machine learning and data mining-based approaches to prediction and detection of heart disease would be of great clinical utility, but are highly challenging to develop. In most countries there is a lack of cardiovascular expertise and a significant rate of incorrectly diagnosed cases which could be addressed by developing accurate and efficient early-stage heart disease prediction by analytical support of clinical decision-making with digital patient records.

**DATA**

One of the impactful processes in every data mining problem is data selection, so we will use the University of California, Irvine (UCI) Cleveland heart Disease dataset, which is clean data and contains 14 attributes the diagnosis (predicted target is included) also it has 303 instances, gathered in 1988 from Cleveland Clinic Foundation.

Cleveland Heart Disease Dataset from the University of California Irvine Repository is real data and has patient records, and there are 13 independent attributes.

Table 1: Attribute Information

|  |  |
| --- | --- |
| Attribute | Description of Attribute |
| Age | Patient Age |
| sex | Patient Sex |
| cp | Chest pain type |
| tbp | The pressure of resting blood in mm Hg |
| chol | Serum of Cholesterol in mg/dl |
| fbs | Sugar rate of fasting blood |
| rstcg | Results of resting electrocardiographic |
| thlch | Max Heart Rate |
| exng | Exercise-induced angina |
| opeak | ST depression induced by exercise relative to rest |
| slope | Peak exercise slope of the ST segment |
| ca | Major vessel count (0–3) colored by fluorescence |
| thal | Thalassemia disease |
| Target | Diagnosis of the Disease |

**APPROACH**:

In order to begin the process of analyzing data, the initial step involves loading the data and conducting exploratory analysis and exploration to gain a better understanding of the data and identify its primary characteristics. Subsequently, the data must be visualized, highlighting the distribution of select attributes and uncovering any insights that may be gleaned from the data. This process allows for a more comprehensive understanding of the data, and aids in the identification of any trends or patterns that may exist within the data set.

We conduct our experiment by implementing the following classification models:

* Naïve Bayes.
* Decision Tree.
* K - Nearest Neighbor.
* Support Vector Machine.

By building an ensemble model from the best models and one from the weak models, we can effectively compare the performance of both models. This approach can help us identify strengths and weaknesses in each model and determine which one works best for specific needs

Looking for an effective way to identify possible groups among patients with heart disease, using a clustering algorithm that can help us accurately classify and group patients based on their health conditions.

**DATA EXPLORATION:**

**Basic Statistics Summary -- Univariate Analysis**

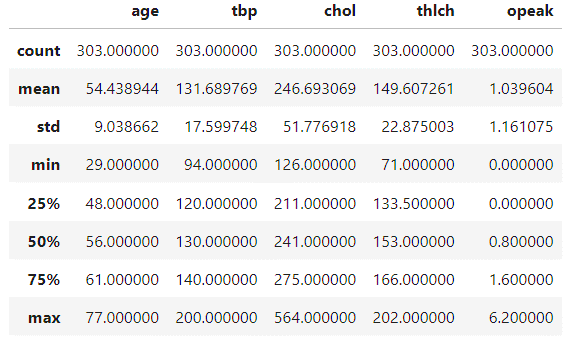
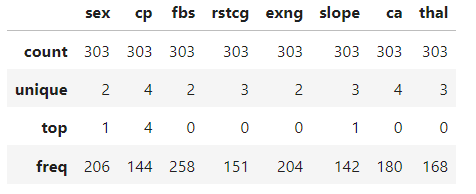
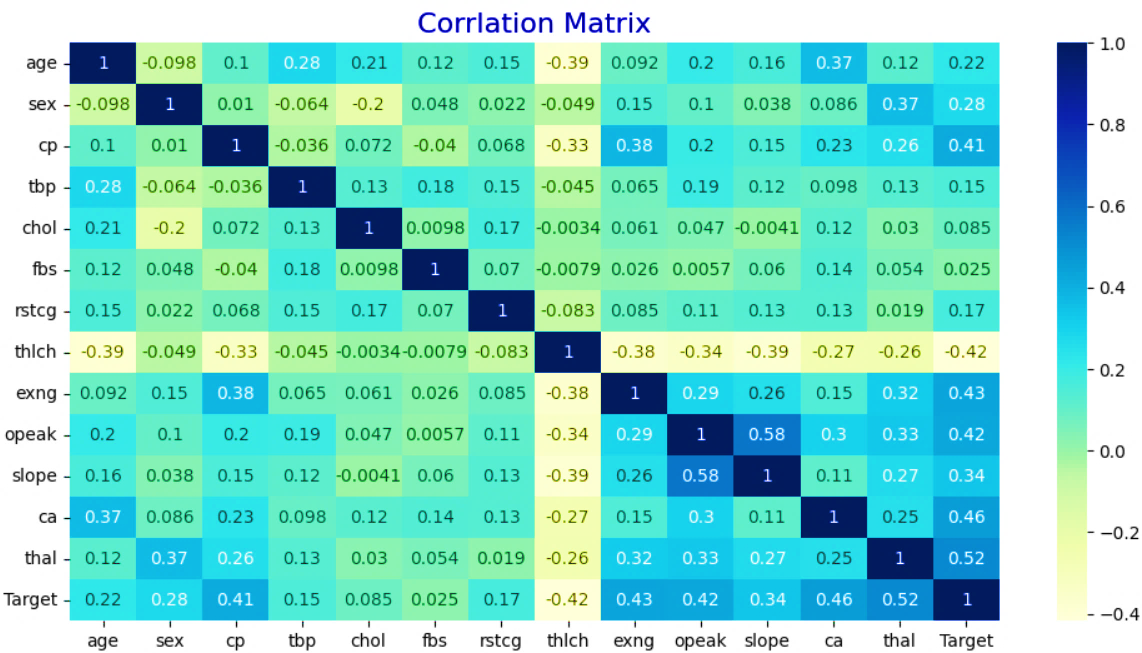
****Table 2: Statistics for numerical attributes

Table 3: Statistics for Categorical attributes

****

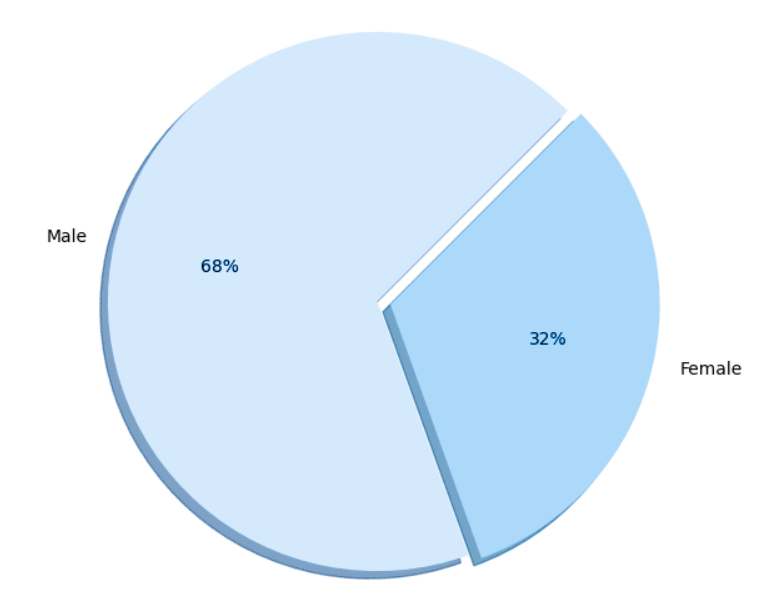
**Multivariate Statistics -- Correlation matrix**

Based in the Correlation Matrix we see that the Top 5 variable that correlated with target variable is **:**

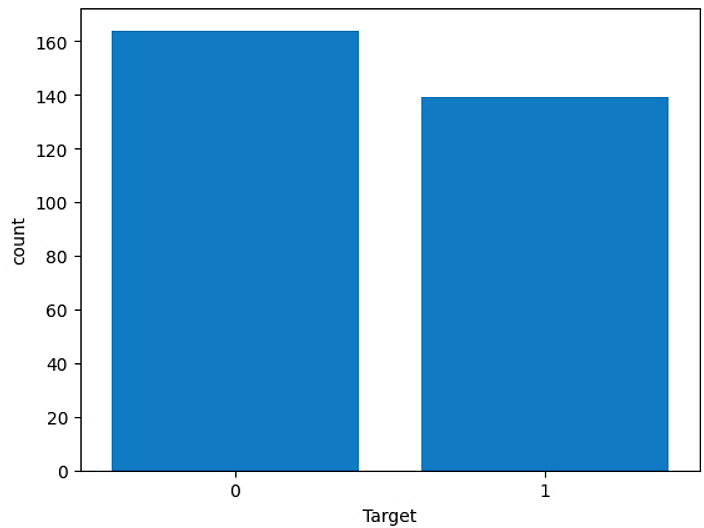
* Max Heart Rate with 0.515.
* Major vessel count colored by fluorescence with 0.460.
* Exercise-induced angina with 0.431.
* ST depression induced by exercise relative to rest with 0.424.
* Chest pain type with 0.414.

The attribute that is most strongly correlated with the target variable is the Maximum Heart Rate having a moderate positive linear relationship.

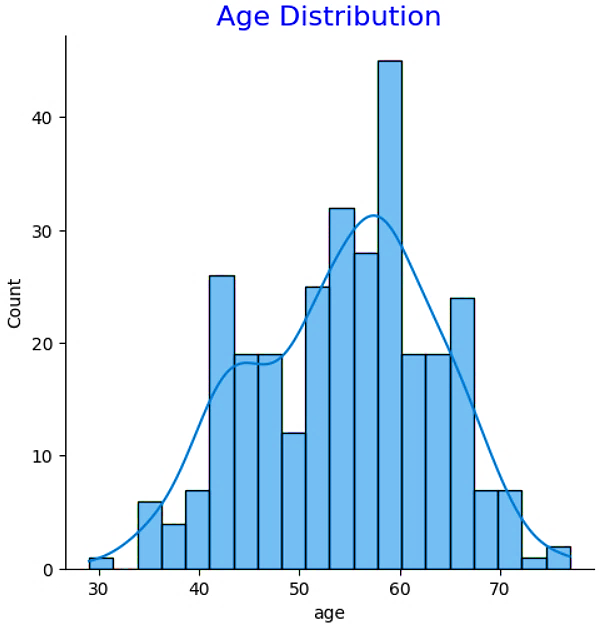
**DATA VISUALIZATION:**

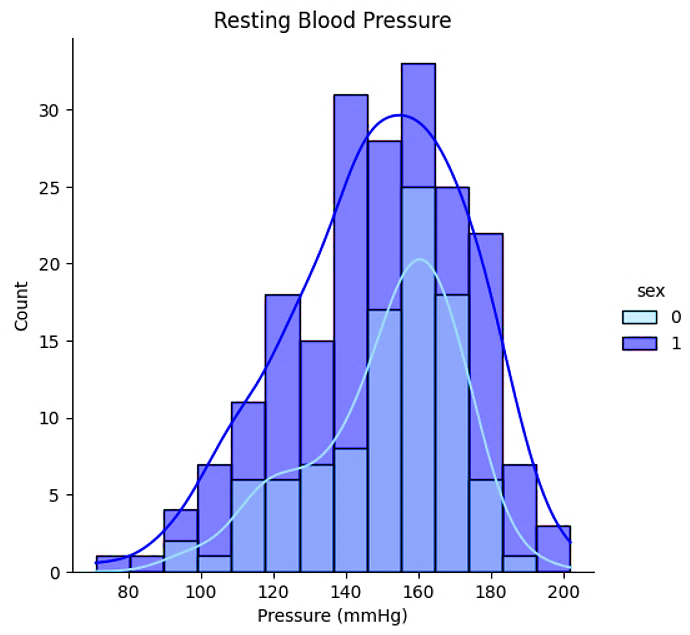
Gender Proportion.

Data distribution based on output class.

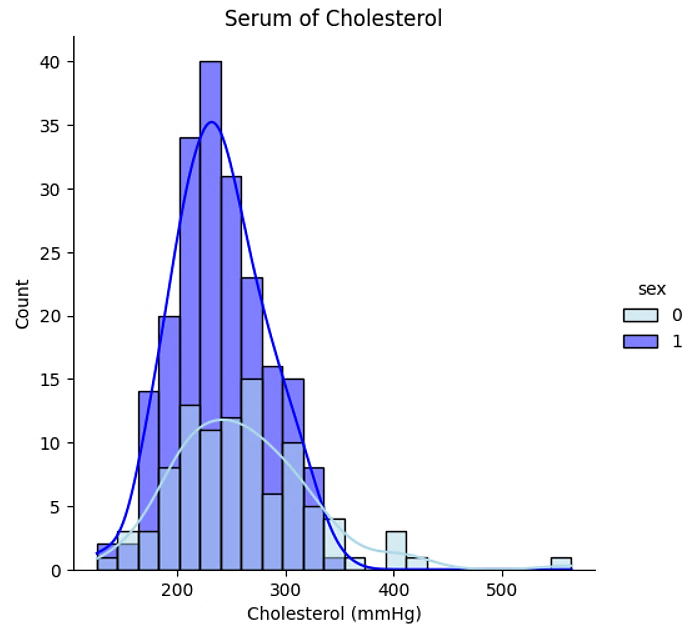
****

Histogram That Show Age Distribution.

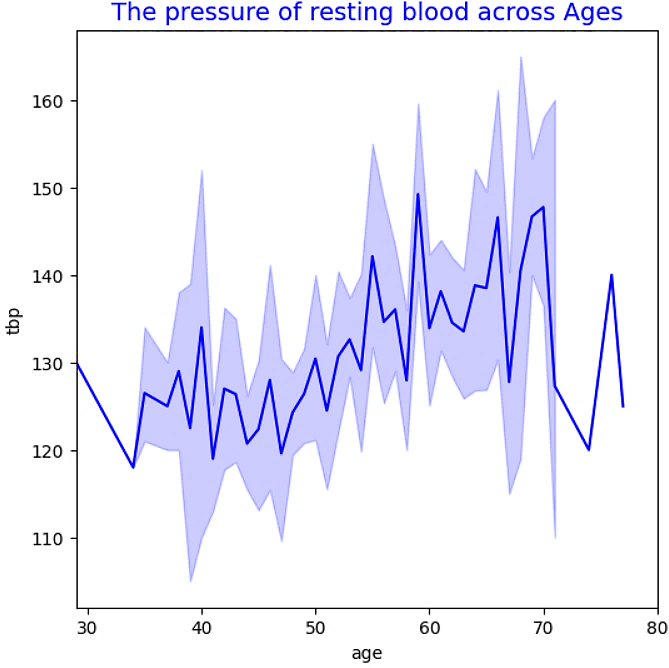
****

The pressure of resting blood Distribution according to Gender.

Serum of Cholesterol blood Distribution according to Gender.

****

Line Chart that show The pressure of resting blood across different age

****

Line Chart that show Max Heart Rate across different age

****

**RESULTS**

We use Confusion Matrix, and ROC curves as the evaluation metrics to analyze the results obtained for our models.

Refer to appendix Figure 1, Figure 2, for detailed description of the results obtained for the classification models.

Table 4: Results Summary

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Model** | **Decision Tree** | **KNN** | **SVM** | **Naïve Bayes** |
| Accuracy | **0.75** | **0.65** | **0.67** | **0.85** |
| Recall | **0.75** | **0.59** | **0.46** | **0.81** |
| Precision | **0.77** | **0.70** | **0.83** | **0.89** |
| F1-Score | **0.76** | **0.64** | **0.60** | **0.85** |

Overall, Naïve Bayes generalizes well when compared to Decision Tree K - Nearest Neighbor , Support Vector Machine.

Table 5: Cross Validation Results Summary

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Model** | **Decision Tree** | **KNN** | **SVM** | **Naïve Bayes** |
| Accuracy | **0.71** | **0.62** | **0.63** | **0.82** |

When evaluating the accuracy of a model, using cross-validation can provide a more reliable measure. While it's true that the accuracy may decrease slightly, cross-validation takes into account the variability of the data and provides a more accurate representation of the model's performance.

Table 6: Ensemble Model Results Summary

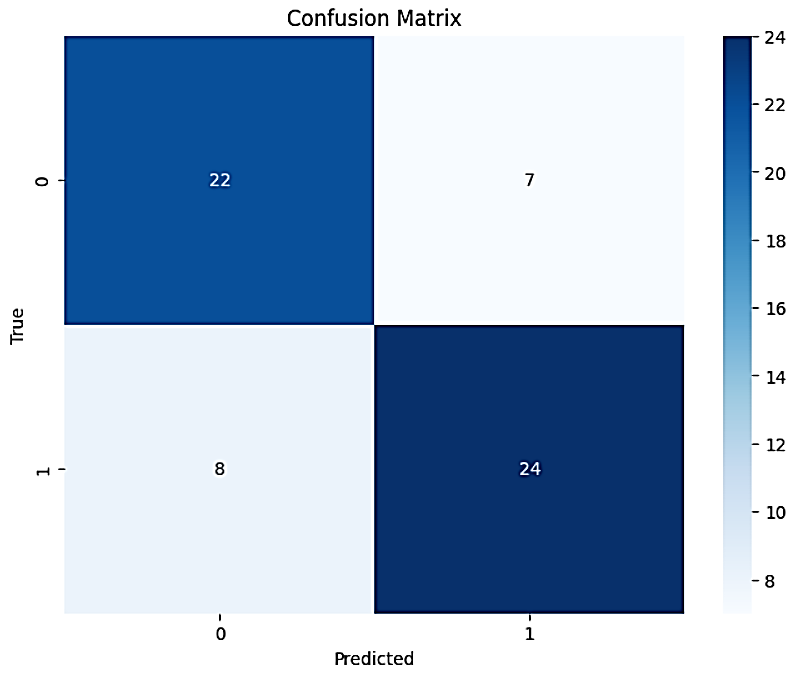
|  |  |  |
| --- | --- | --- |
| **Model** | **Best Models** | **Weak Models** |
| Accuracy | **0.90** | **0.77** |
| Recall | **0.87** | **0.62** |
| Precision | **0.93** | **0.90** |
| F1-Score | **0.90** | **0.74** |

In Ensemble Learning, a technique to build machine learning models, we create two ensemble models: one with the best three models and the other with the weakest three models. By comparing these two models, we can observe that the accuracy of the best three models increases, while the weak three models also show improvement in their results. This indicates that the process of selecting models is critical in machine learning and can have a significant impact on the accuracy of the final output.

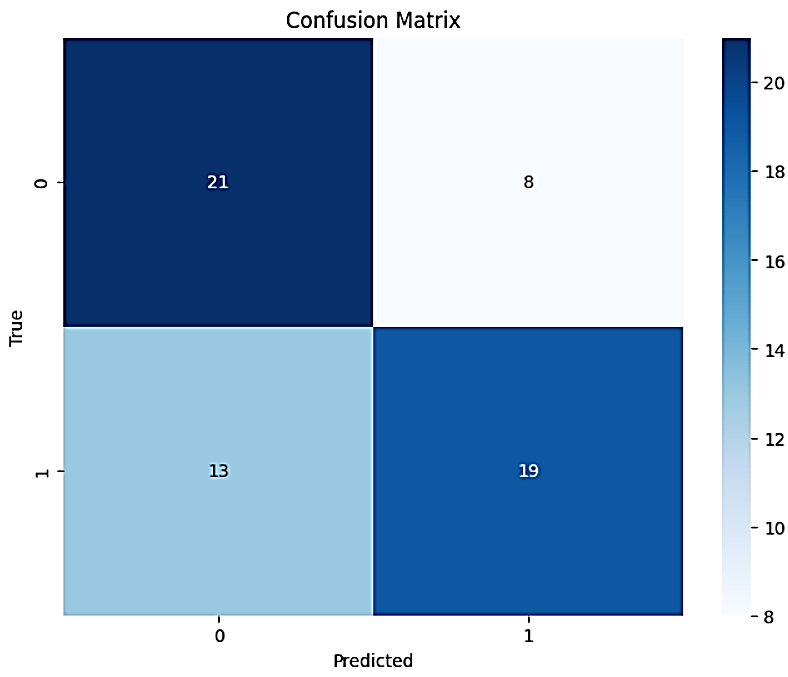
**APPENDIX**

Figure 1: Confusion Matrix

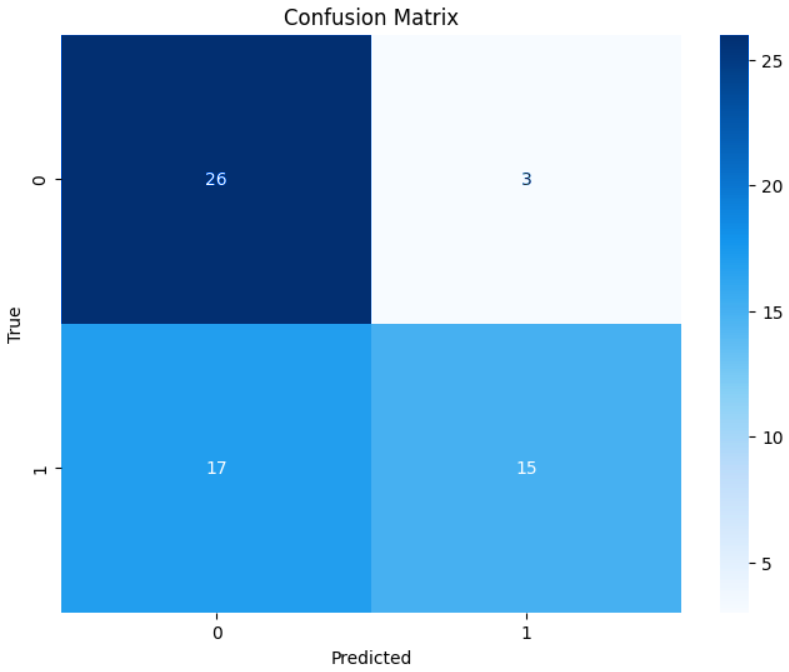
***K - Nearest Neighbor***



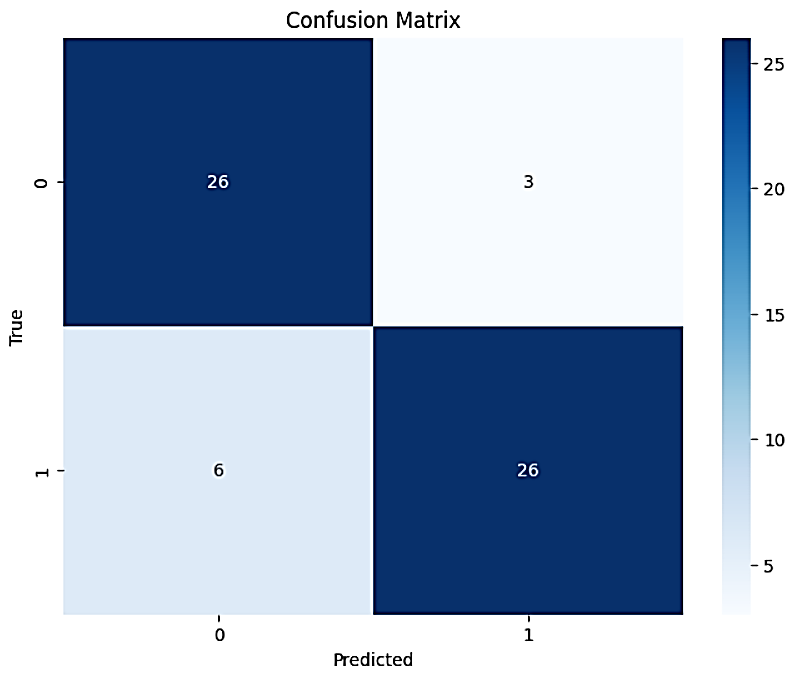
***Decision Tree***

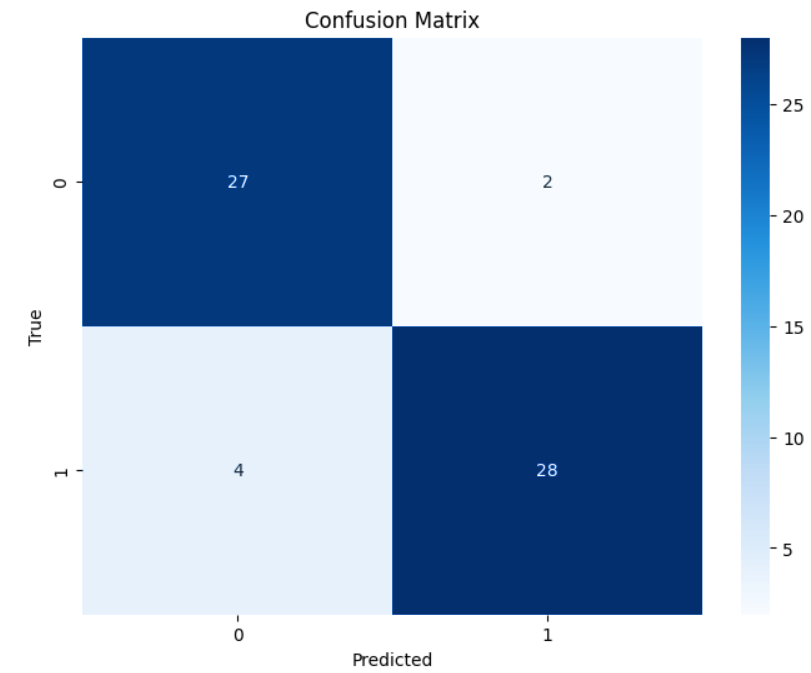


***Support Vector Machine***



***Naïve Bayes***



***The Best Models***

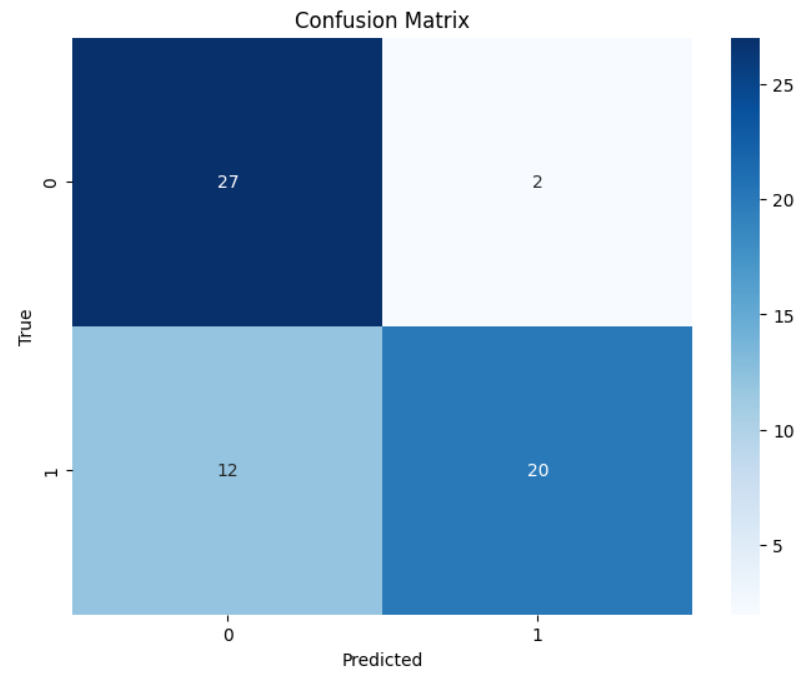
***The Weak Models***

Figure 2: ROC

The Receiver Operating Characteristic (ROC) curve is a graphical representation that illustrates the trade-off between true positive rate (sensitivity) and false positive rate (1-specificity) for a binary classification model across different threshold values. It helps visualize the model's performance and allows for comparison between different models or thresholds.

The AUC (Area Under the Curve) is a metric used to evaluate the performance of binary classification models. In general, an AUC of 0.50 indicates that the model is making random predictions, while an AUC of 1.0 indicates that the model is making perfect predictions.

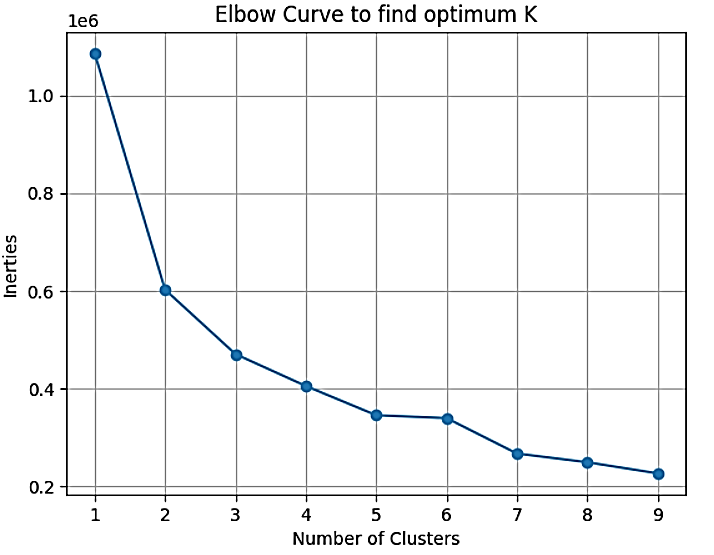
In this case, the AUC for the Decision Tree Classifier is 0.75, which means that it performs better than random prediction. The K-Nearest Neighbor model has an AUC of 0.72, which is also better than random prediction, but not as good as the Decision Tree Classifier. The Support Vector Machine has an AUC of 0.76, which is the highest among the models evaluated. Finally, the Naive Classifier has the highest AUC of 0.91, which indicates that it performs the best among all the models evaluated.

**UNSUPERVISED LEARNING**

[**Clustering**](https://en.wikipedia.org/wiki/Cluster_analysis) is an unsupervised learning method whose job is to separate the population or data points into several groups, such that data points in a group are more similar to each other dissimilar to the data points of other groups. It is nothing but a collection of objects based on similarity and dissimilarity between them.

**KMeans** clustering is an Unsupervised Machine Learning algorithm that does the clustering task. In this method, the ‘n’ observations are grouped into ‘K’ clusters based on the distance. The algorithm tries to minimize the within-cluster variance(so that similar observations fall in the same cluster).

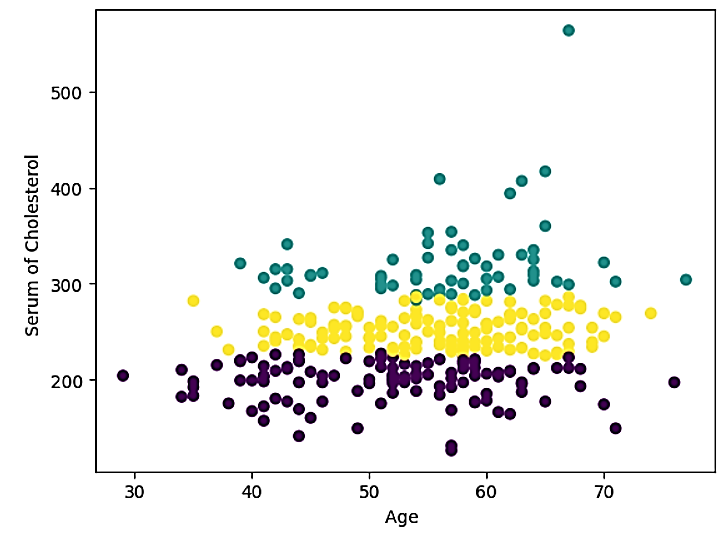
**Elbow** method gives us an idea on what a good k number of clusters would be based on the sum of squared distance (SSE) between data points and their assigned clusters’ centroids.



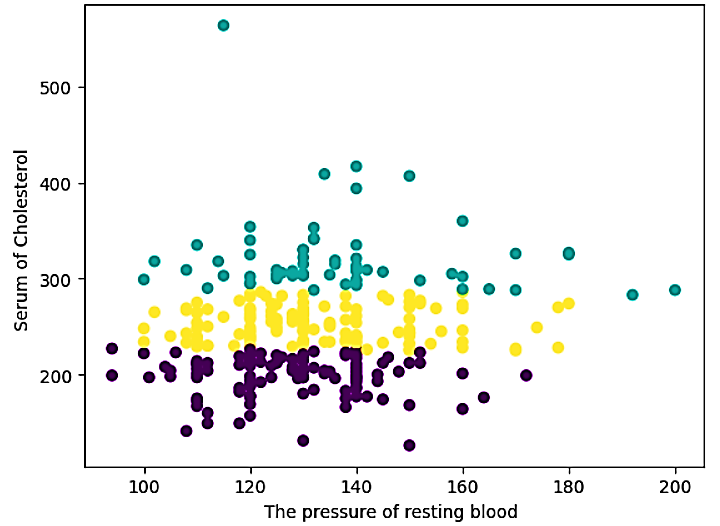
To find the optimal K for a dataset, use the Elbow method; find the point where the decrease in inertia begins to slow. K=3 is the “elbow” of this graph.

when visualizing clusters in a scatter plot, take any two features and plot them to see how the observations are clustered.

Based in Age and Serum of Cholesterol.



Based in The pressure of resting blood and Serum of Cholesterol.



Clustering results help us create customized treatment plans for patient groups with similar characteristics, improving their outcomes.

Silhouette score aka Silhouette Coefficient is an evaluation metric that results in the range of -1 to 1. A score near 1 signifies the best importance that the data point is very compact within the cluster to which it belongs and far away from the other clusters. The score near -1 signifies the least or worst importance of the data point. A score near 0 signifies overlapping clusters.

In our case the Silhouette Score:0.285, So you have overlapping cluster.

And the Davies-Bouldin Index score is defined as the average similarity measure of each cluster with its most similar cluster, where similarity is the ratio of within-cluster distances to between-cluster distances. Thus, clusters that are farther apart and less dispersed will result in a better score. The minimum score is 0, with lower values indicating better clustering.

In our case the Davies-Bouldin Index: 1.134, So you have overlapping cluster.

GitHub Project Link:

https://github.com/Mones-ksasbeh/Heart-Disease-Prediction/tree/main