Hanoi University of Science and Technology
School of Information Communication and Technology

——————— * * * ———————

Final Report

GRADUATION RESEARCH 2

HEART DISEASE PREDICTION

Class : 143704
Supervisor : PhD. Nguyen Hong Quang

Full name	Student ID
Pham Anh Minh	20194802

Hanoi, January 31, 2024

ABSTRACTION

This paper presents a comprehensive investigation into the development and application of predictive models for heart disease risk assessment. Drawing upon a diverse array of medical data sources, including patient demographics, clinical records, and diagnostic tests, the study explores the efficacy of various machine learning algorithms in predicting the likelihood of cardiovascular complications. Key components of the abstraction include:

1. ***Data Collection and Preprocessing**: A rigorous data collection process is undertaken to assemble a representative dataset encompassing a wide spectrum of cardiovascular health indicators. Preprocessing steps such as data cleaning, normalization, and feature engineering are applied to ensure the quality and compatibility of the input data.*
2. ***Model Selection and Evaluation**: Multiple machine learning algorithms, ranging from traditional statistical methods to more advanced techniques like neural networks, are evaluated for their performance in predicting heart disease risk. Evaluation metrics such as accuracy, sensitivity, specificity, and area under the receiver operating characteristic curve (AUC-ROC) are employed to assess the robustness and generalization capability of the models.*
3. ***Feature Importance Analysis**: The paper investigates the relative importance of different features or risk factors in influencing the prediction of heart disease. Techniques such as feature selection, permutation importance, or SHAP (SHapley Additive exPlanations) values may be utilized to identify the most informative predictors contributing to the model's decision-making process.*
4. ***Model Interpretability**: Efforts are made to enhance the interpretability of the predictive models, enabling clinicians and stakeholders to understand the underlying mechanisms driving the risk predictions. Interpretability techniques such as feature importance plots, decision tree visualization, or model-agnostic explanations are employed to elucidate the relationship between input variables and predicted outcomes.*
5. ***Validation and External Generalization**: The paper discusses the validation process to ensure the reliability and reproducibility of the developed models.*

External validation on independent datasets or cross-validation techniques may be employed to assess the generalizability of the predictive models across diverse populations and healthcare settings.

In summary, the Heart Disease Prediction Paper contributes to the growing body of literature on predictive analytics in cardiovascular medicine, offering insights into the development, evaluation, and clinical application of machine learning models for heart disease risk assessment.

CHAPTER1. INTRODUCTION

1.Dataset Problem:

The dataset utilized in this study is sourced from the UCI machine learning website, comprising medical records of patients alongside their diagnosis outcomes regarding the presence of heart disease. Employing machine learning models becomes imperative to ascertain the presence of heart disease and expedite the diagnostic procedure, leveraging the medical information available for each patient. Furthermore, this analysis will delve deeper into identifying the variables exerting the most significant influence on the likelihood of heart disease occurrence in patients.

2.Notebook Objectives:

The objectives of this notebook include:

1. Conducting comprehensive dataset exploration utilizing diverse types of data visualization techniques.
2. Developing a machine learning model capable of accurately predicting the status of patients.
3. Exporting the prediction results obtained from testing data into files for further analysis.
4. Saving or dumping the entire machine learning pipeline for future usage, ensuring reproducibility and scalability.
5. Executing predictions on new example data provided and exporting the resultant predictions for evaluation and application purposes.

3.Machine Learning Model:

The models used in this notebook:

1. Logistic Regression
2. K-Nearest Neighbour (KNN)
3. Support Vector Machine (SVM)
4. Gaussian Naive Bayes
5. Decision Tree
6. Random Forest
7. Gradient Boosting

CHAPTER 2. READING DATASET

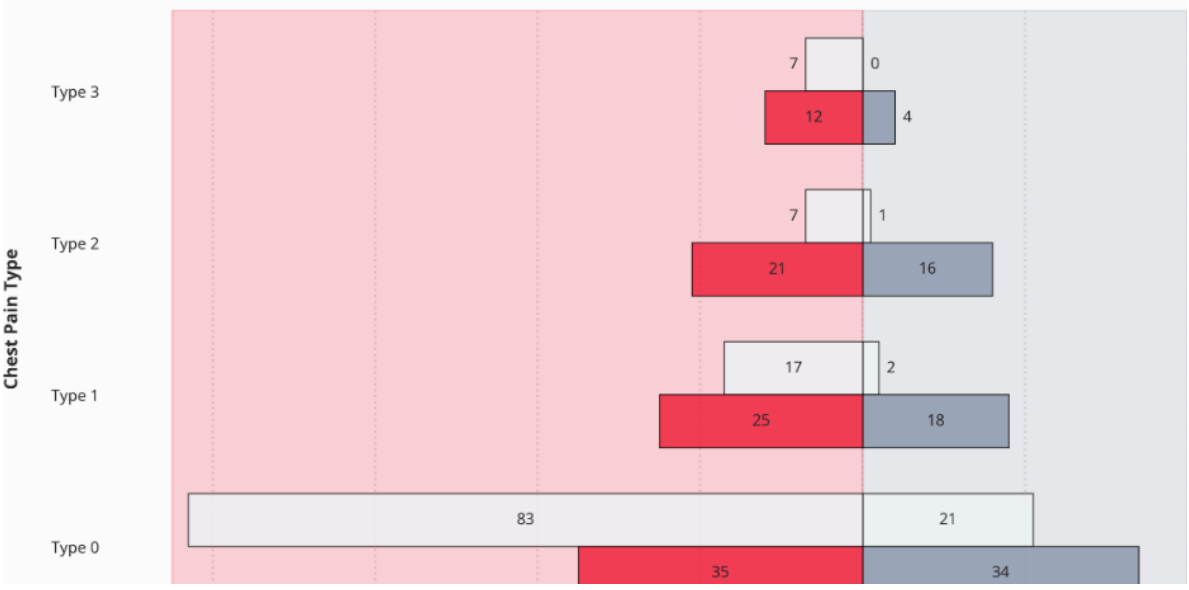
The following is the **structure of the dataset**.

Variable Name	Description	Sample Data
Age	Patient age (in years)	63; 37; ...
Sex	Gender of patient 0 = male 1 = female	1; 0; ...
cp	Chest pain type 0 = typical angina 1 = atypical angina 2 = non-anginal pain 3 = asymptomatic	3; 1; 2; ...
trestbps	Resting blood pressure (in mm Hg)	145; 130; ...
chol	Serum cholestoral (in mg/dl)	233; 250; ...
fbs	Fasting blood sugar > 120 mg/dl 0 = false 1 = true	1; 0; ...
restecg	Resting electrocardiographic results 0 = normal 1 = having ST-T wave abnormality 2 = showing probable or definite left ventricular	0; 1; ...

	hypertrophy by Estes' criteria	
thalach	Maximum heart rate achieved	150; 187; ...
exang	Exercise induced angina 0 = no 1 = yes	1; 0; ...
oldpeak	ST depression induced by exercise relative to rest	2.3; 3.5; ...
slope	The slope of the peak exercise ST segment 0 = upsloping 1 = flat 2 = downsloping	0; 2; ...
ca	Number of major vessels (0-4) colored by flourosopy	0; 3; ...
thal	Thalassemia 3 = normal 6 = fixed defect 7 = reversable defect	1; 3; ...
Target	Target column 0 = not have heart disease 1 = have heart disease	1; 0; ...

CHAPTER 3. EDA

1.Disease Distribution based on Chest Pain Type in Each Gender



Based on the butterfly chart presented earlier and as previously noted, it's evident that the dataset contains a higher number of cases associated with typical angina chest pain and female patients. However, upon closer examination, it becomes apparent that atypical angina, non-anginal pain, and asymptomatic chest pain are associated with a higher proportion of patients diagnosed with heart disease compared to both healthy male and female patients. Notably, within the category of patients experiencing typical angina chest pain, the ratio of male to female patients diagnosed with heart disease is nearly equivalent.

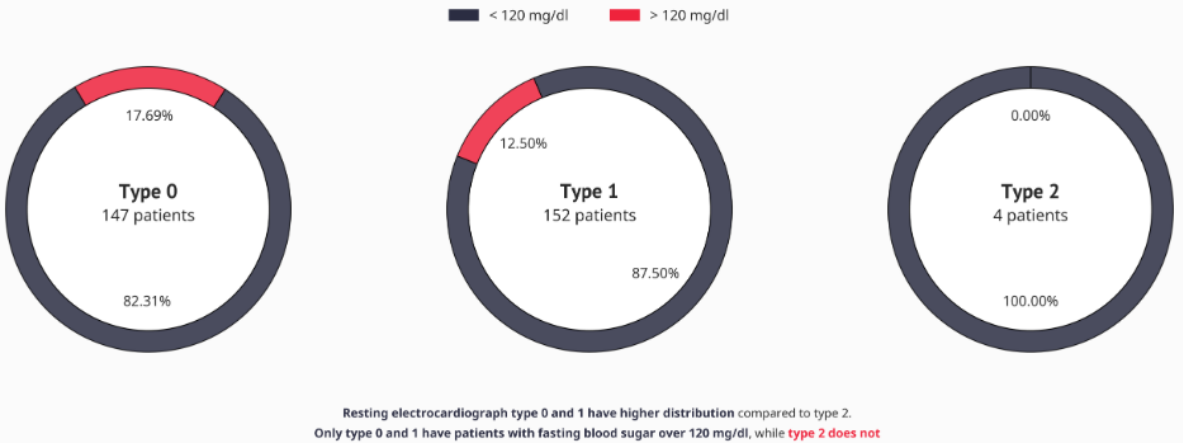
Nevertheless, the number of healthy female patients in this chest pain category surpasses that of healthy male patients.

2. Maximum Heart Rate vs. Age based on Patients Sickness

The scatter plot depicted above reveals that individuals both with and without heart disease fall within the age range of 40 to 70 years old. Furthermore, the dataset illustrates that the maximum heart rate among patients spans from 140 to 180 beats per minute. Upon closer examination, it becomes apparent that individuals prone to heart disease typically exhibit a maximum heart rate exceeding 149 bpm and are under 54 years of age. Additionally, the scatter

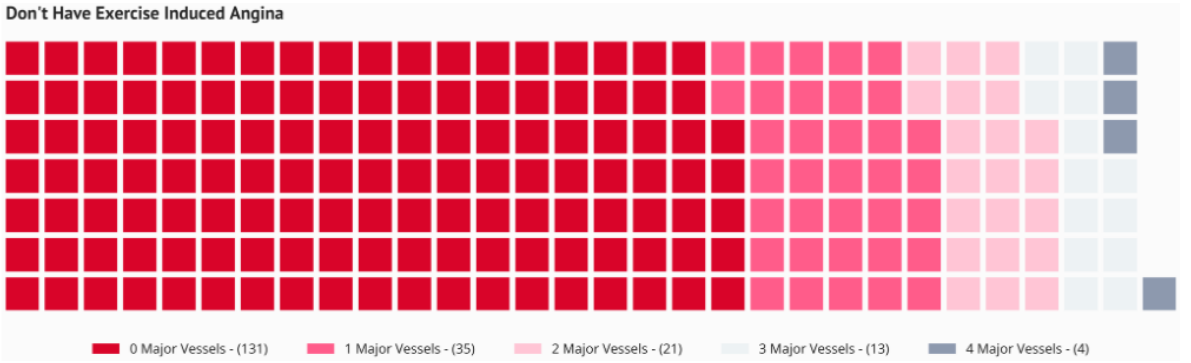
plot highlights a negative correlation between age and maximum heart rate, particularly prominent among patients diagnosed with heart disease. Furthermore, the plot underscores that there are more instances of heart disease patients compared to healthy individuals within the dataset.

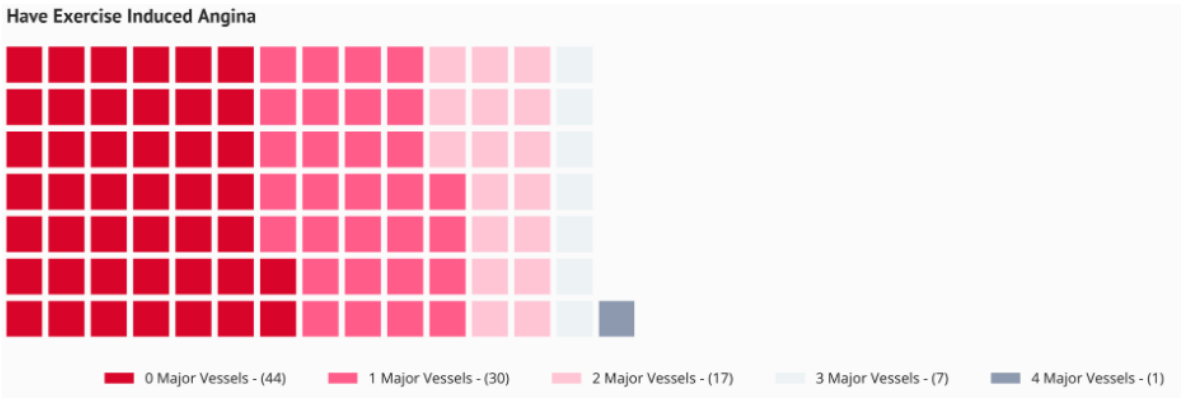
3. Fasting Blood Sugar Distribution by Resting Electrocardiographic Results



As illustrated in the donut chart above, resting electrocardiograph types 0 and 1 exhibit a similar number of patients, while resting electrocardiograph type 2 is notably underrepresented with only four patients. Furthermore, it's noteworthy that patients with resting electrocardiograph types 0 and 1 tend to have fasting blood sugar levels exceeding 120 mg/dl. Conversely, resting electrocardiograph type 2 displays a distinct pattern, as none of its patients have fasting blood sugar levels surpassing 120 mg/dl.

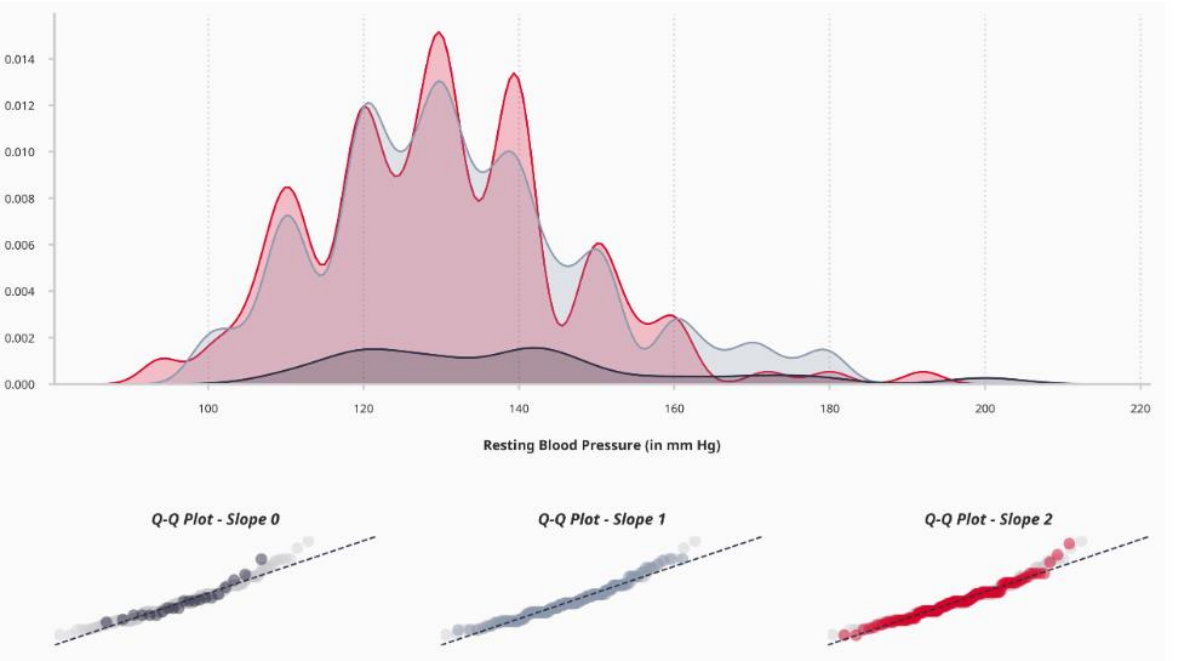
4. Number of Major Vessles Distribution based on Exercise Induced Angina





The waffle charts presented above indicate that the proportion of patients experiencing exercise-induced angina compared to those who do not is nearly equivalent. This observation is evident when comparing the total number of patients across major vessels within each exercise category.

5.Resting Blood Pressure Distribution based on Slope



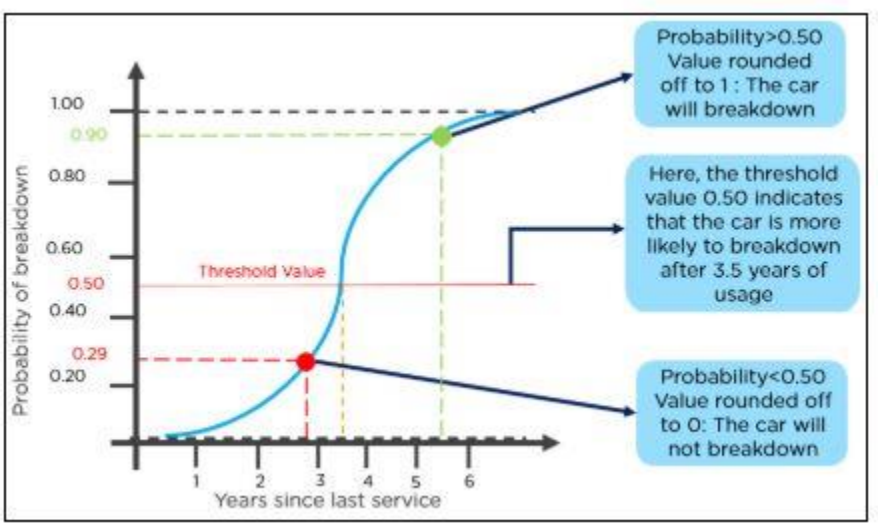
The distribution plot and Q-Q plots displayed above reveal that the distribution of each slope type exhibits moderate right-skewness, attributable to outliers primarily located in the distribution tail on the right side of the plot. Furthermore, the skewness value and the noticeable gap above the 45-degree line in the Q-Q plots indicate that the distribution within this column is non-normal.

CHAPTER 4. MODEL IMPLEMENTATION

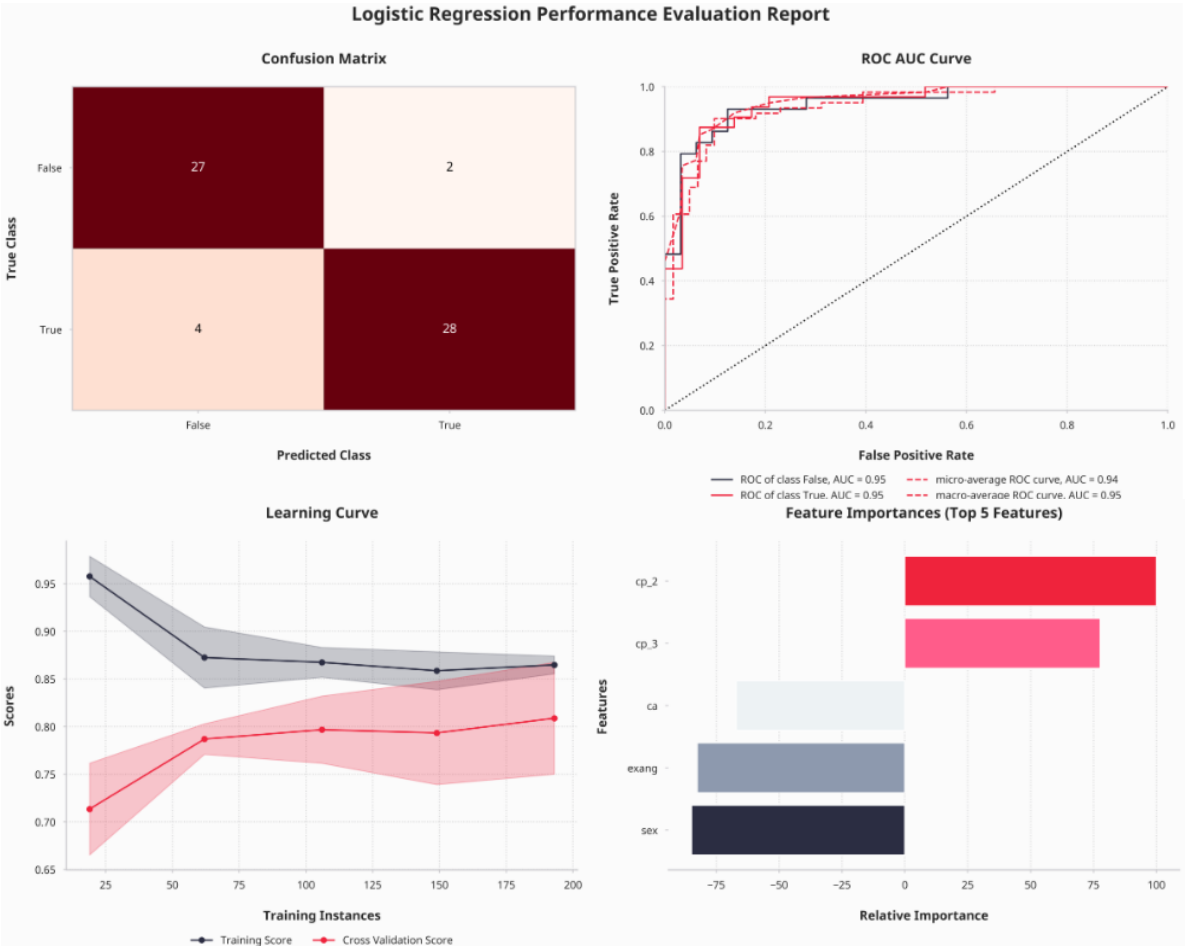
1.Logistic Regression

Logistic regression is a statistical method that is used for building machine learning models where the dependent variable is dichotomous: i.e. binary. Logistic regression is used to describe data and the relationship between one dependent variable and one or more independent variables. The independent variables can be nominal, ordinal, or of interval type.

The name "logistic regression" is derived from the concept of the logistic function that it uses. The logistic function is also known as the sigmoid function. The value of this logistic function lies between zero and one.



	precision	recall	f1-score	support
0	0.87	0.93	0.90	29
1	0.93	0.88	0.90	32
accuracy			0.90	61
macro avg	0.90	0.90	0.90	61
weighted avg	0.90	0.90	0.90	61

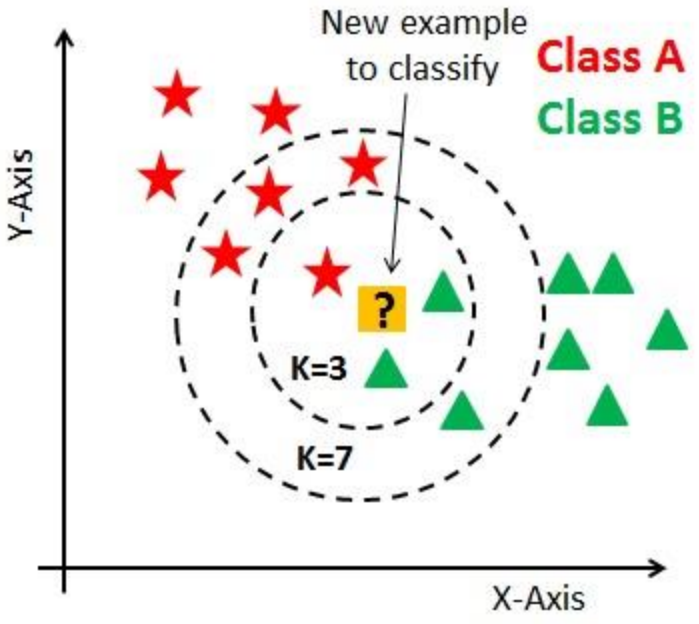


2. K-Nearest Neighbour (KNN)

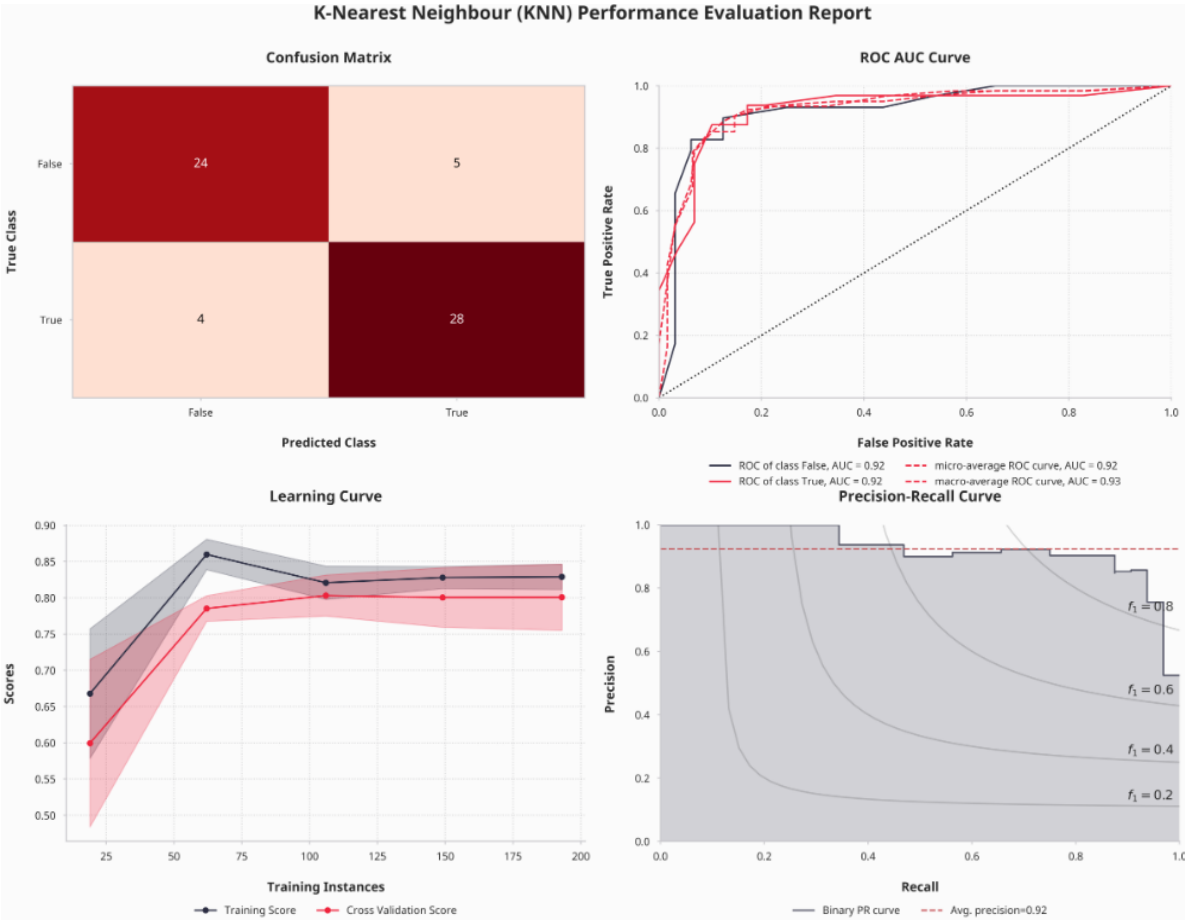
The k-nearest neighbors (KNN) algorithm is a method for data classification, aiming to predict the likelihood of a data point belonging to a particular group based on the groups of its nearest neighbors. This algorithm, falling under the

category of supervised machine learning techniques, is employed for both classification and regression tasks.

Termed as a "lazy learning" algorithm, KNN does not engage in training upon receiving the training data. Instead, it merely stores the data and abstains from executing any computations. It defers the construction of a model until a query is initiated on the dataset. This characteristic renders KNN suitable for data mining applications.



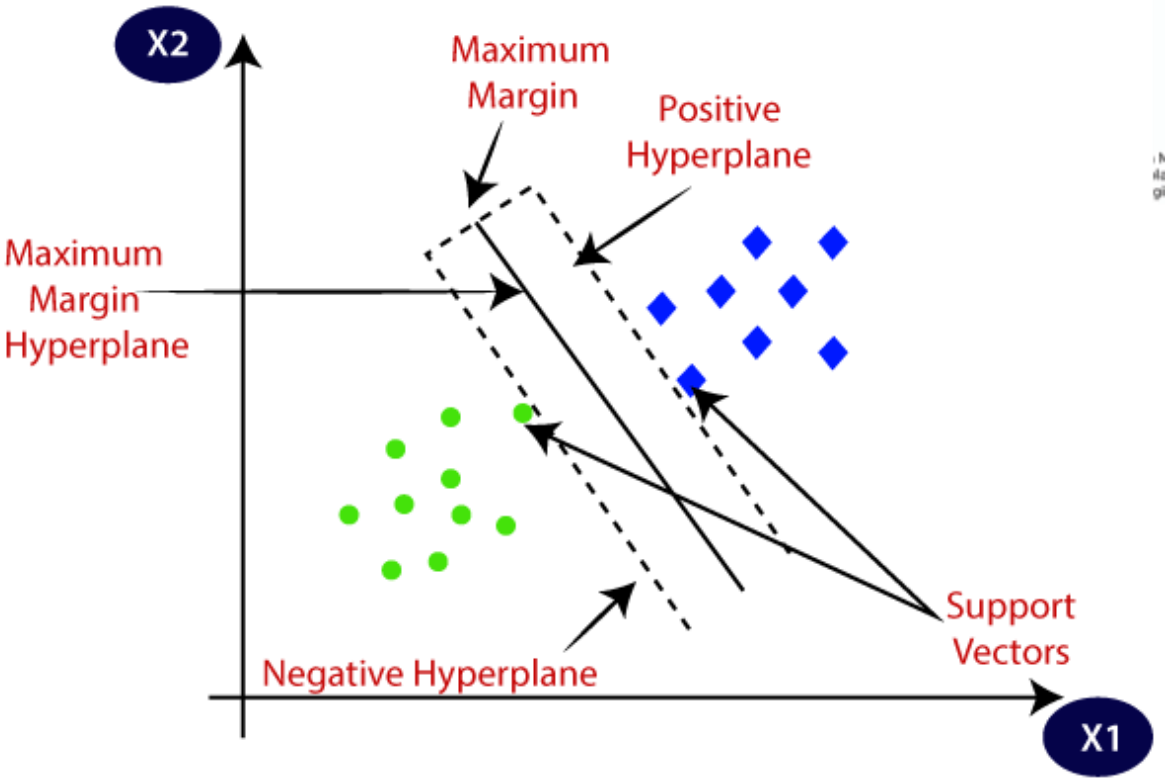
	precision	recall	f1-score	support
0	0.86	0.83	0.84	29
1	0.85	0.88	0.86	32
accuracy			0.85	61
macro avg	0.85	0.85	0.85	61
weighted avg	0.85	0.85	0.85	61



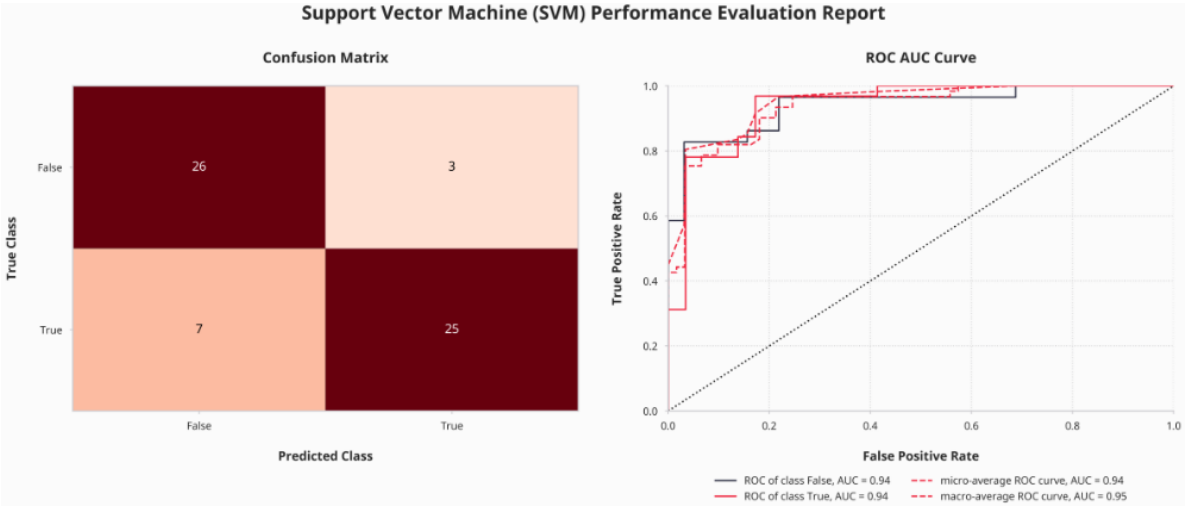
3. Support Vector Machine (SVM)

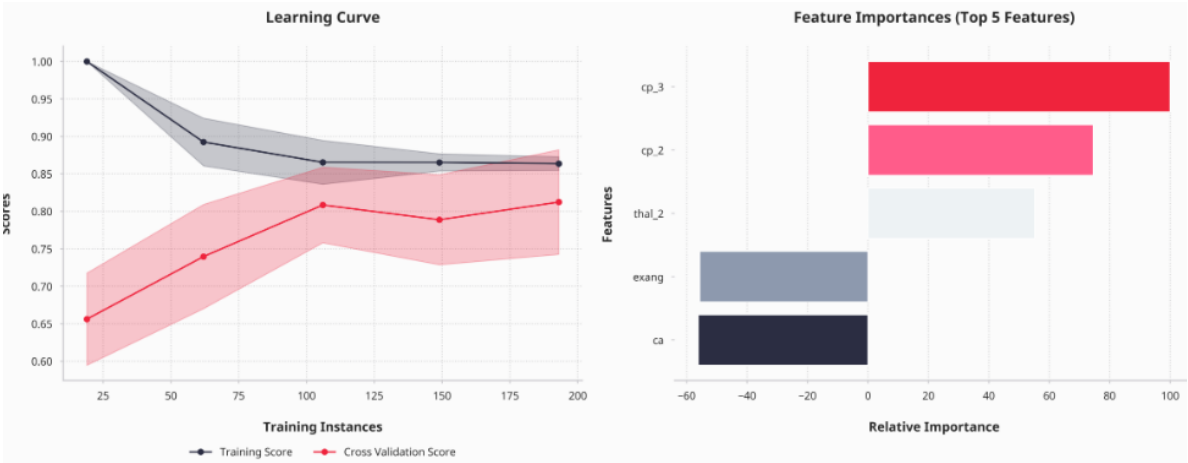
The Support Vector Machine (SVM) stands out as one of the most widely utilized Supervised Learning algorithms, applied to tasks encompassing both Classification and Regression. Its primary objective is to construct an optimal line or decision boundary, known as a hyperplane, effectively partitioning n-dimensional space into distinct classes. This delineation facilitates the accurate categorization of new data points into the appropriate class in subsequent instances.

SVM achieves this by identifying the extreme points or vectors crucial for defining the hyperplane. These critical instances are termed "support vectors," hence lending the algorithm its name, Support Vector Machine.



	precision	recall	f1-score	support
0	0.79	0.90	0.84	29
1	0.89	0.78	0.83	32
accuracy			0.84	61
macro avg	0.84	0.84	0.84	61
weighted avg	0.84	0.84	0.84	61

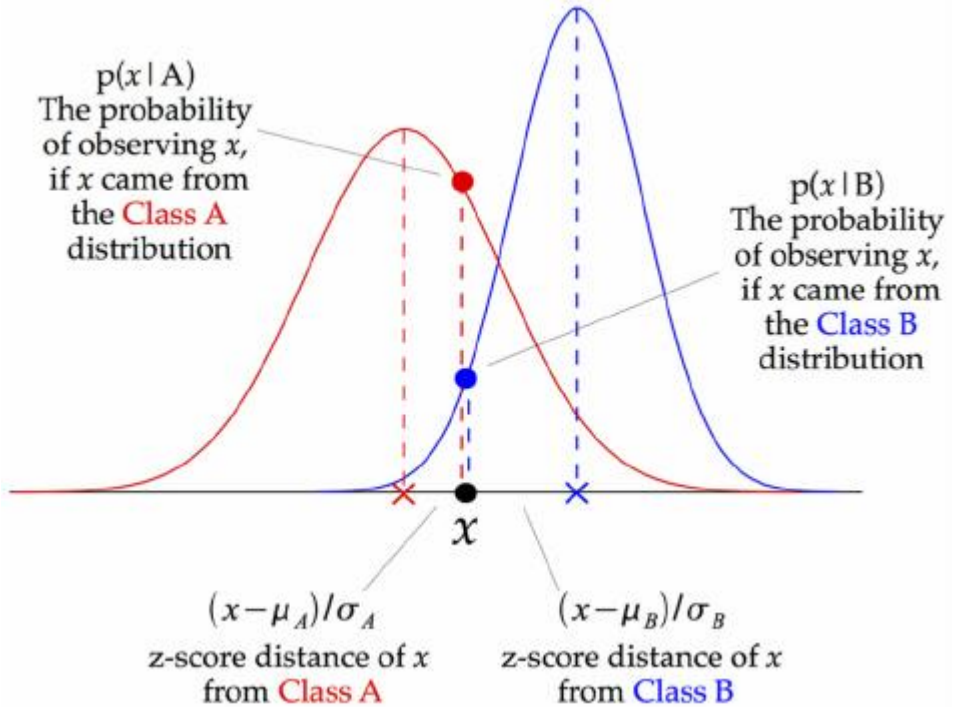




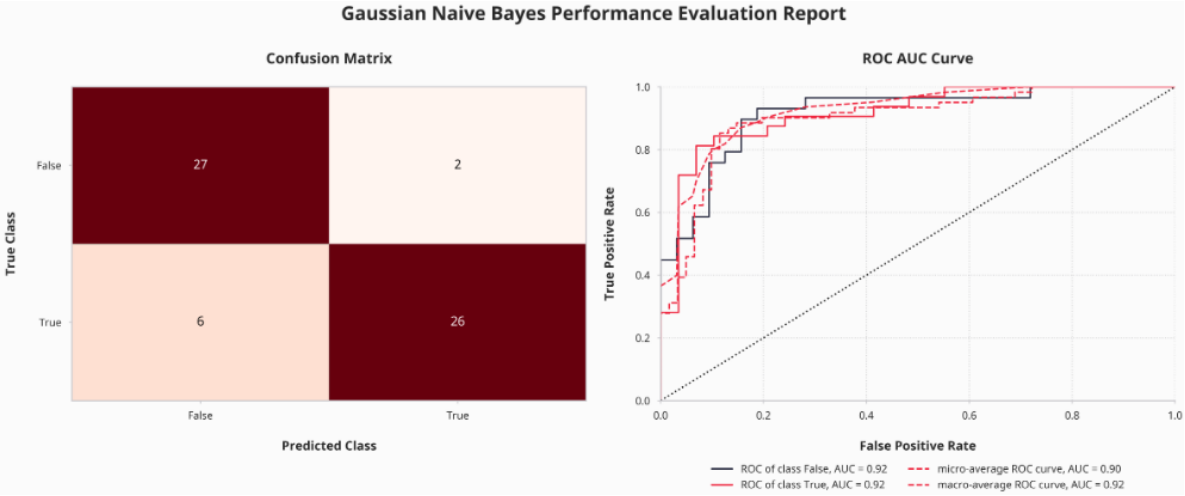
4. Gaussian Naive Bayes

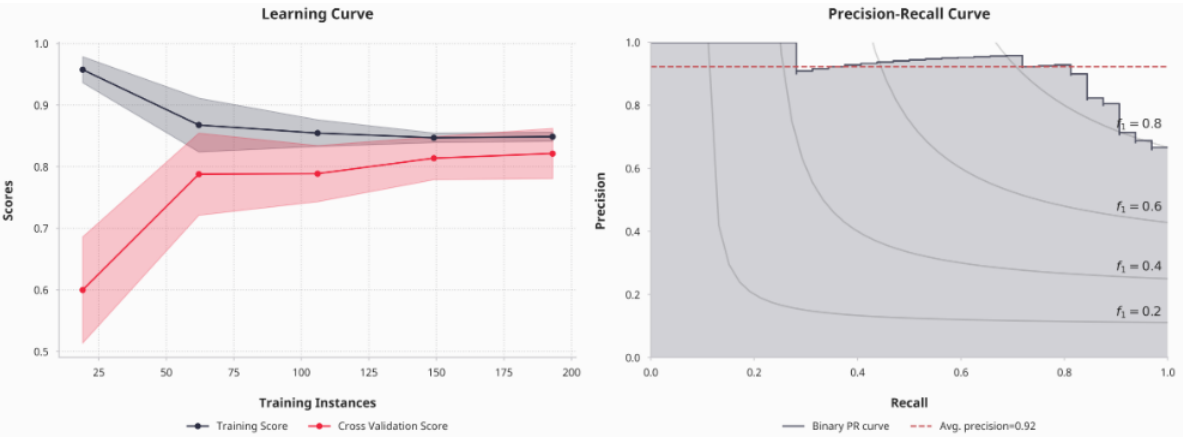
Naive Bayes Classifiers, grounded in the Bayes Theorem, operate under the strong assumption of feature independence, implying that the value of one feature is unrelated to the value of any other feature. In supervised learning scenarios, these classifiers efficiently train on data, requiring only a small dataset to estimate the parameters essential for classification. Their straightforward design and implementation make them applicable to a wide range of real-life situations.

Gaussian Naive Bayes, a variant of Naive Bayes, specifically caters to continuous data by adhering to a Gaussian normal distribution. This model assumes that continuous values associated with each class follow a normal (Gaussian) distribution, facilitating effective handling of continuous data.



	precision	recall	f1-score	support
0	0.82	0.93	0.87	29
1	0.93	0.81	0.87	32
accuracy			0.87	61
macro avg	0.87	0.87	0.87	61
weighted avg	0.88	0.87	0.87	61

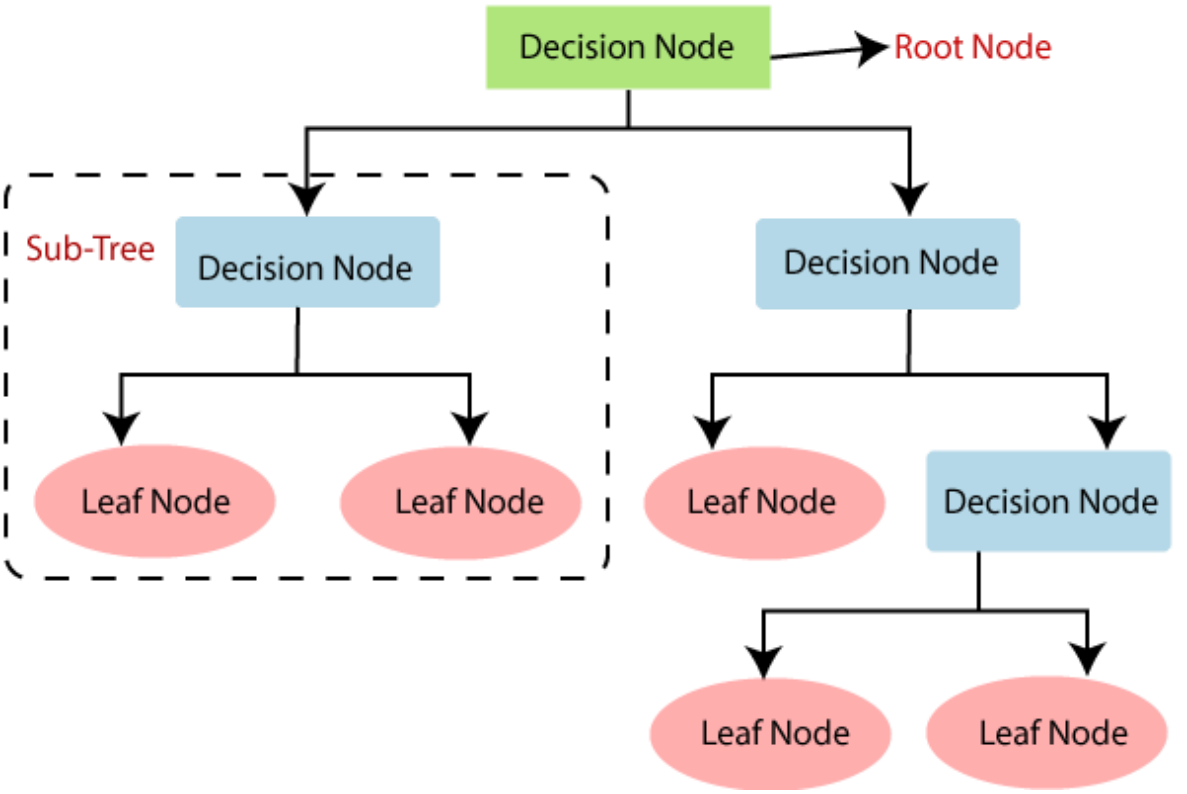




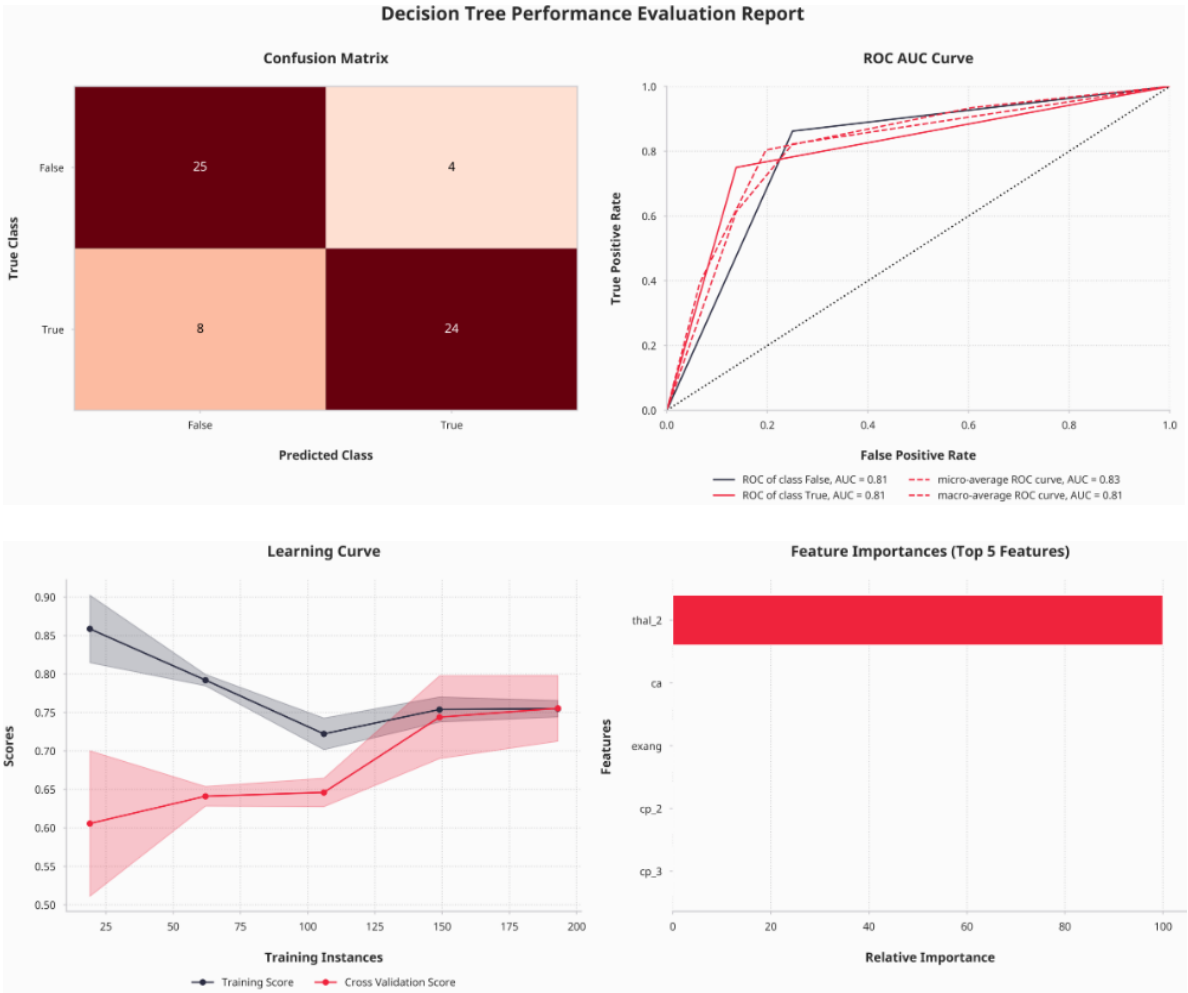
5.Decision Tree

Decision Tree is a versatile Supervised learning method applicable to both classification and regression tasks, although it is predominantly favored for solving classification problems. It operates as a tree-structured classifier, wherein internal nodes correspond to dataset features, branches signify decision rules, and each leaf node denotes an outcome.

Within the Decision Tree framework, two types of nodes are prominent: Decision Nodes and Leaf Nodes. Decision nodes serve to make decisions and possess multiple branches, while leaf nodes represent the final outcomes resulting from these decisions and do not lead to further branching.



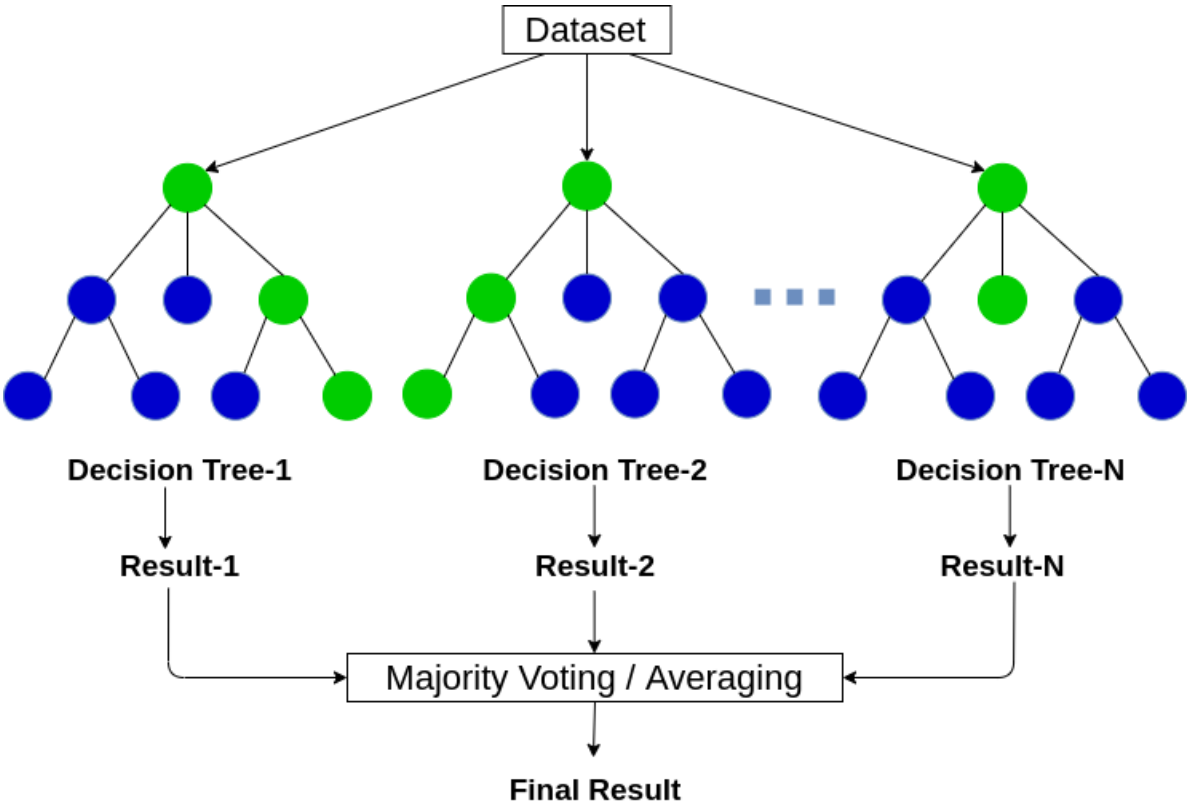
	precision	recall	f1-score	support
0	0.76	0.86	0.81	29
1	0.86	0.75	0.80	32
accuracy			0.80	61
macro avg	0.81	0.81	0.80	61
weighted avg	0.81	0.80	0.80	61



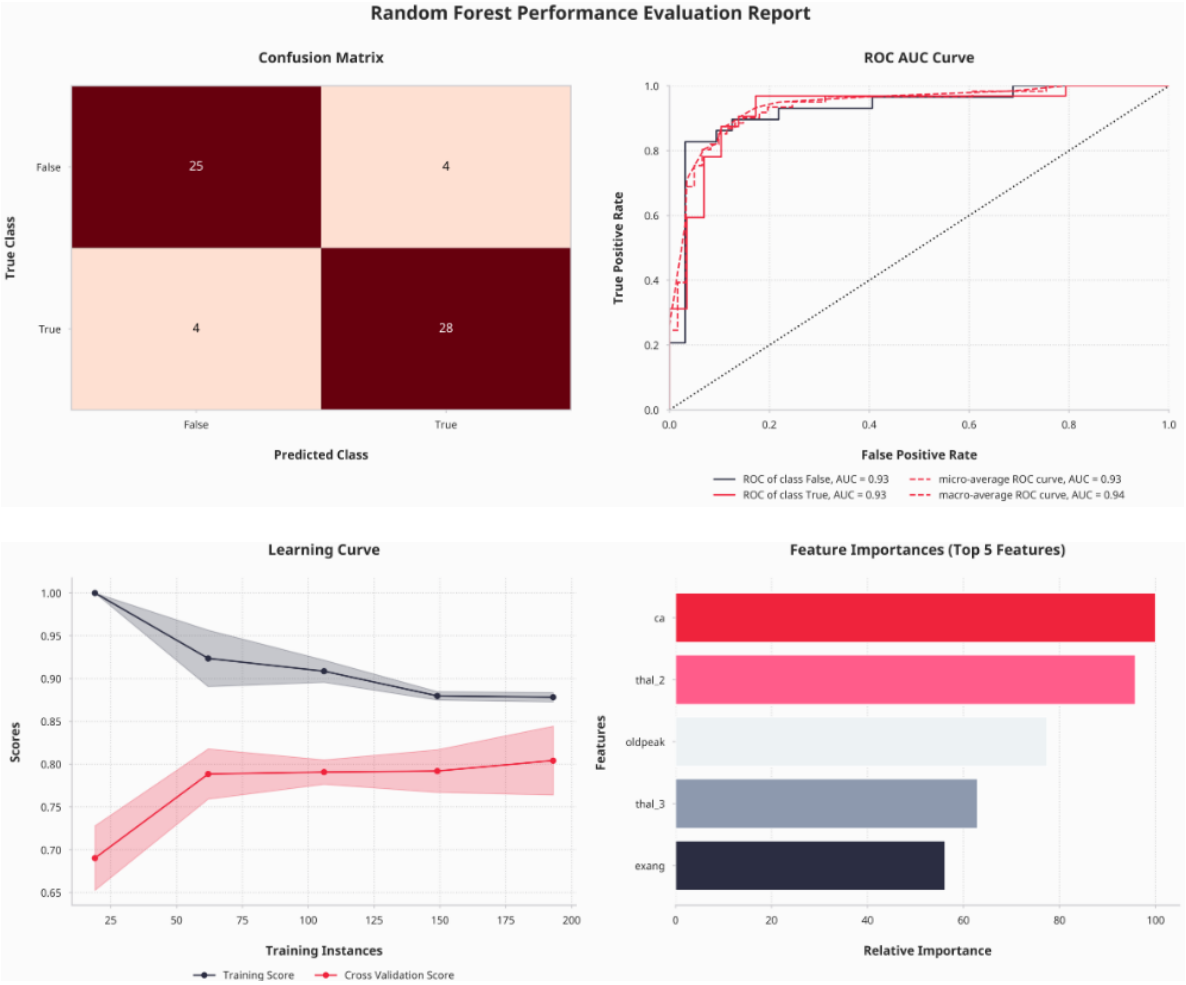
6. Random Forest

Random Forest is a robust tree-based machine learning algorithm that harnesses the collective strength of multiple decision trees to make predictions. Within a Random Forest ensemble, each individual tree generates its own class prediction, and the final model prediction is determined by the class with the most "votes" from all the trees. By aggregating the predictions from numerous trees, each trained on different subsets of the data, Random Forest can achieve

superior performance compared to any single constituent model. This ensemble approach ensures that the models are relatively uncorrelated, leading to improved generalization and robustness.



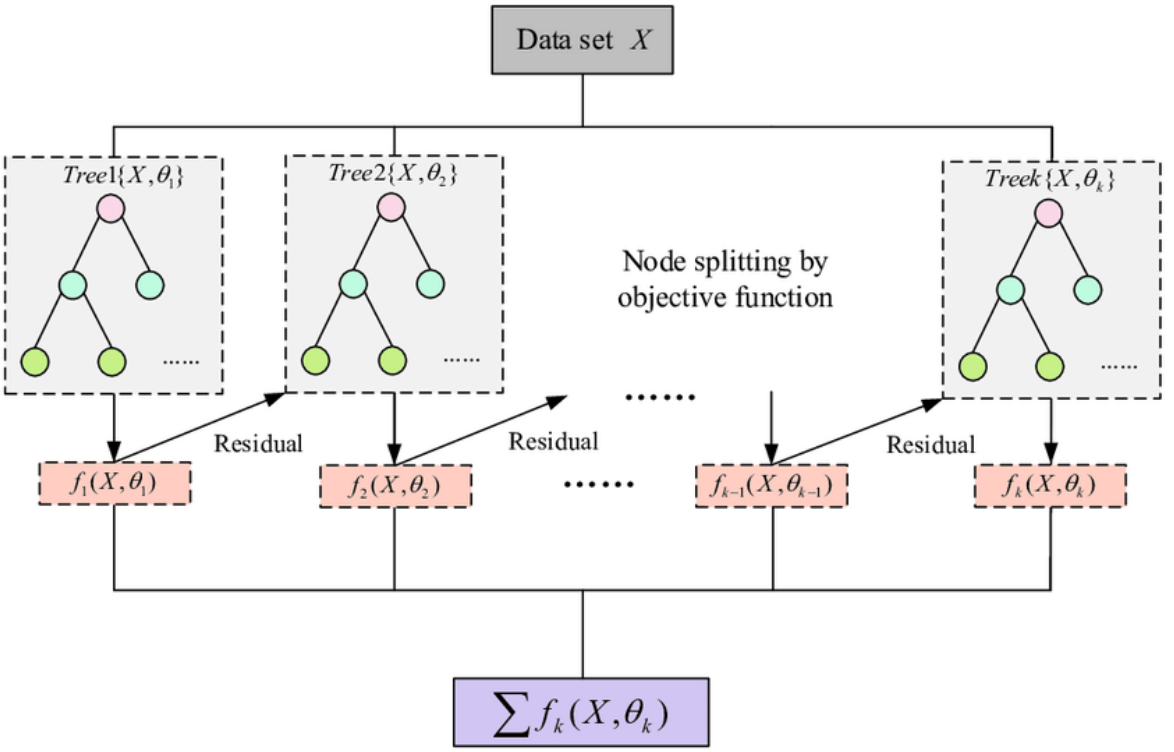
	precision	recall	f1-score	support
0	0.86	0.86	0.86	29
1	0.88	0.88	0.88	32
accuracy			0.87	61
macro avg	0.87	0.87	0.87	61
weighted avg	0.87	0.87	0.87	61



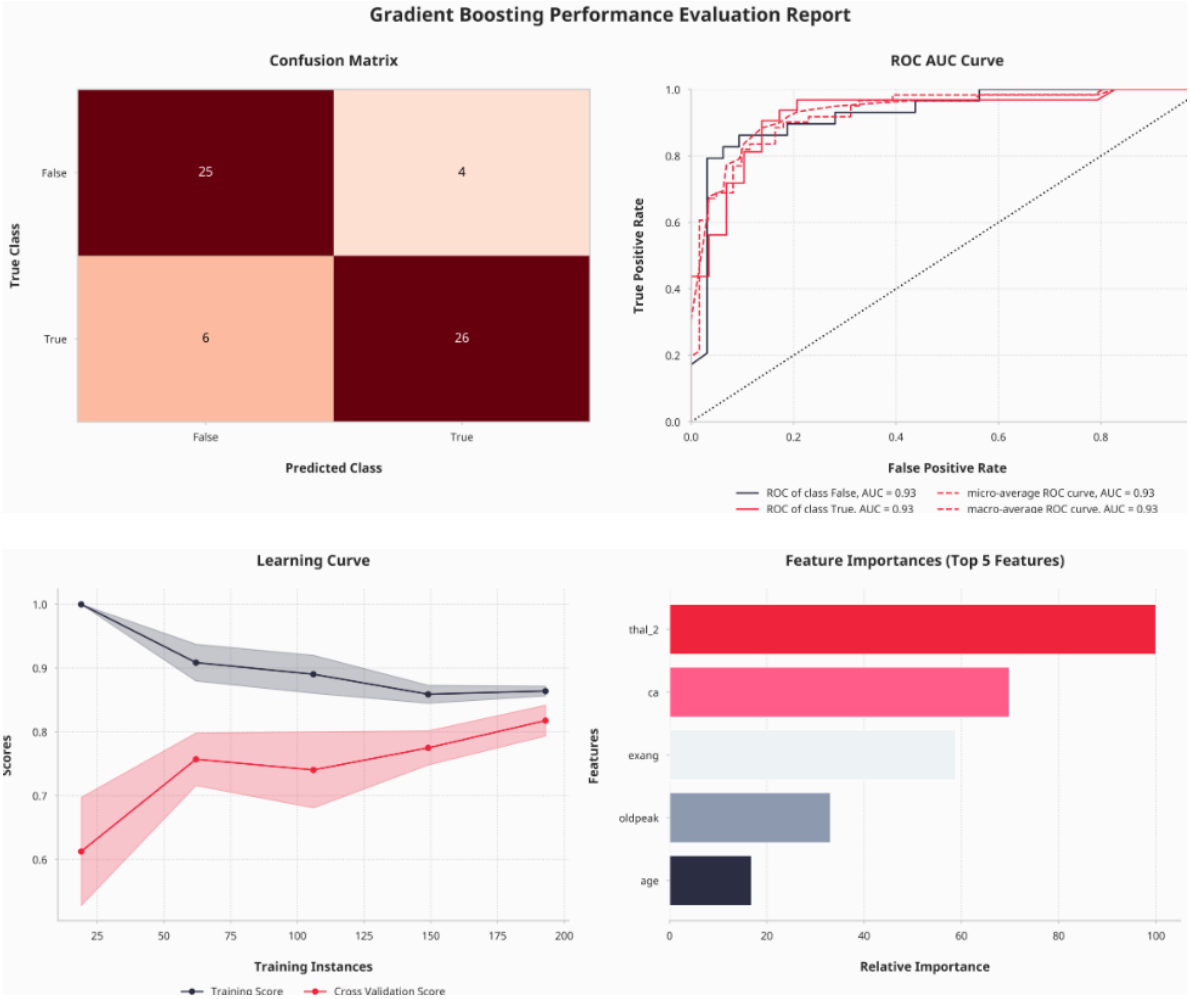
7. Gradient Boosting

Boosting is a technique aimed at enhancing weak learners to become formidable ones. In boosting, each subsequent model is trained on a modified version of the original dataset, with a focus on improving prediction accuracy. This method heavily relies on the expectation that the next model will effectively reduce prediction errors when combined with the previous ones. The fundamental concept revolves around setting target outcomes for the forthcoming model to minimize errors.

Gradient Boosting, a specific approach within the boosting framework, trains numerous models in a gradual, additive, and sequential manner. The term "gradient boosting" arises from the fact that the target outcomes for each instance are determined based on the gradient of the error concerning the predictions. Each model in the sequence works towards reducing prediction errors by iteratively adjusting predictions in the direction that minimizes the error.



	precision	recall	f1-score	support
0	0.81	0.86	0.83	29
1	0.87	0.81	0.84	32
accuracy			0.84	61
macro avg	0.84	0.84	0.84	61
weighted avg	0.84	0.84	0.84	61


Gradient Boosting Performance Evaluation Report
Confusion Matrix
True Class
False
True
25
4
6
26
False
True
Predicted Class
ROC AUC Curve
True Positive Rate
False Positive Rate
ROC of class False, AUC = 0.93
ROC of class True, AUC = 0.93
micro-average ROC curve, AUC = 0.93
macro-average ROC curve, AUC = 0.93
Learning Curve
Scores
Training Instances
Training Score
Cross Validation Score
Feature Importances (Top 5 Features)
Features
thal_2
ca
exang
oldpeak
age
Relative Importance

CHAPTER 5. MODEL COMPARISON

Model	Acccuracy Train	Accuracy Test	Best Score
SVM	85.124	83.607	0.8352
GB	86.364	83.607	0.8352
Random Forest	87.603	86.885	0.8307
Gausian NB	85.950	86.885	0.8303
Logistic Regression	85.950	90.164	0.8230
KNN	83.471	85.246	0.8185
Decision Tree	75.620	80.328	0.7563

Based on the results obtained from the accuracy assessment of the training and test data, it is apparent that several models encountered issues of overfitting or underfitting. However, there are notable exceptions where certain models exhibit a good fit, with minimal discrepancies between the train and test accuracies. Specifically, the Random Forest, Gaussian Naive Bayes models demonstrate promising performance.

Among these models, Random Forest and Gaussian Naive Bayes stand out with the highest accuracy compared to the others. This observation is corroborated by the ROC AUC curve analysis, which indicates that both models possess strong predictive capabilities, as evidenced by the AUC values nearing 1. Additionally, the confusion matrix reveals that the Random Forest

and Gaussian Naive Bayes models outperform others in accurately predicting the target classes in the test data.

Analyzing the F1 scores of both models further underscores their efficacy in differentiating between sick and healthy patients, with scores exceeding 0.85. Notably, Gaussian Naive Bayes demonstrates a precision value of 93%, indicating a high proportion of accurately predicted heart disease cases. Conversely, while the Random Forest model exhibits a slightly lower precision value at 88%, it compensates with a superior recall value of 88%, indicating its ability to effectively identify patients with heart disease.

In terms of the learning curves, the Random Forest model demonstrates a more favorable trend compared to Gaussian Naive Bayes. The Random Forest's learning curve indicates robust generalization ability, as evidenced by the convergence of training and validation scores to similar values with increasing training set sizes. Conversely, the Gaussian Naive Bayes model's learning curve suggests potential issues with both variance and bias, with training and validation scores remaining close together.

Moreover, the feature importance plot of the Random Forest model highlights major vessel number (ca), fixed defect thalassemia (thal_2), ST depression induced by exercise relative to rest (oldpeak), reversible defect thalassemia (thal_3), and exercise-induced angina (exang) as the most influential features. These findings underscore the importance of factors such as the number of major vessels, thalassemia types, ST depression, and exercise-induced angina in predicting heart disease, emphasizing the significance of early detection and appropriate management strategies to improve cardiovascular health.

In conclusion, based on the comprehensive analysis conducted, the Random Forest model emerges as the most effective predictor of heart disease, exhibiting robust performance and highlighting key features contributing to accurate predictions.

CHAPTER 6. CONCLUSIONS

From the analysis and implementation of machine learning models in the previous sections, the following conclusions can be drawn:

Random Forest emerges as the most effective model among the nine machine learning models examined in this notebook. It demonstrates strong performance with both train and test data, outperforming other models in predicting test data outcomes, as evidenced by the performance evaluation graphs and classification reports.

Medical professionals can prioritize examination of the five variables identified as having the greatest influence on whether a patient has heart disease. Focusing on these variables may aid in early detection and targeted intervention strategies.

The prediction results for test data, dummy data, and the complete machine learning pipeline have been successfully exported for further analysis. Additionally, data exploration has been effectively conducted using the ydata-profiling, seaborn, and matplotlib libraries.

Several potential areas for improvement exist for future research or iterations of this notebook. For example, conducting A/B Testing on patients with the same major vessel number to explore variations in outcomes. Additionally, advanced hyperparameter tuning experiments could be performed to enhance model accuracy, aiming for higher levels of performance, potentially reaching around 90%.

By addressing these areas of improvement and leveraging the strengths of the Random Forest model and influential variables, future iterations of this research can contribute to more accurate and effective predictions in the diagnosis and management of heart disease.