

UNITEDWORLD INSTITUTE OF TECHNOLOGY

(UIT)

Summative Assessment (SA)

Submitted BY

Shaily Antala

(Enrl. No.: 20220701037)

**Course Code and Title: 21BSAI99E32 - Introduction to Machine Learning**

B.Sc. (Hons.) Computer Science / Data Science / AIML

IV Semester – December – April 2024

**UIT**

Dec/Apr 2024

# MACHINE LEARNING PROJECT

## REPORT ON HEART DISEASE PREDICTION

### 

**Table of Contents:**

* **Introduction………………………………………………………………………….3**
* **Problem Statement…………………………………………………………………..3**
* **Objective……………………………………………………………………………..3**
* **Importance…………………………………………………………………………...3**
* **Dataset Description………………………………………………………………….3**
* **Code for the Model:....................................................................................................5**
  + **Importing Libraries…………………………………………………………5**
  + **Data Collection and Data Description……………………………………...6**
  + **Data Cleaning………………………………………………………………..8**
  + **Statistical Analysis…………………………………………………………..9**
  + **Data Processing……………………………………………………………...26**
  + **Logistic Regression………………………………………………………….28**
  + **SVC…………………………………………………………………………..28**
  + **KNeighborsClassifier……………………………………………………….29**
  + **Decision Tree Classifier……………………………………………………..30**
  + **Random Forest Classifier…………………………………………………...30**
  + **Gradient Boosting Classifier………………………………………………..31**
  + **Distribution………………………………………………………………….37**
  + **Correlation Matrix………………………………………………………….38**
  + **Evaluation Metrics and Confusion Metrics……………………………….40**
* **Conclusion…………………………………………………………………………..44**

### INTRODUCTION

Heart disease is a major cause of death worldwide, and there are difficulties in using clinical data analysis to forecast heart disease. Through the analysis of medical data, machine learning (ML) has become a useful tool for the diagnosis and prognosis of cardiac disease. ML approaches have been widely used during research involving medical purposes for heart disease prediction in previous studies. To create the prediction model, a variety of feature combinations and well-known classification techniques were used.

### PROBLEM STATEMENT

Heart disease ranks among the world's top causes of death; successful treatment and prevention depend heavily on early detection. Thus creating a machine learning model that can identify people who have heart related disease on the basis of a variety of clinical and demographic characteristics.

### OBJECTIVE

The promise for saving lives, enhancing health outcomes, and effectively allocating healthcare resources is the driving force behind the indication for the presence of cardiovascular associated disease. Aiming to lessen the cost of CHD to the society as a whole and to individual sufferers at large, the main contributions include early intervention, customized care, technology improvements, the effect on the general population as well as current studies.

### IMPORTANCE

The WHO, or World Health Organization, estimates that heart problems account for 12 million deaths worldwide annually. In the US and other developed countries, 50% of all deaths are caused by cardiovascular sickness. People at high risk of cardiovascular disease can reduce their risk by changing their lifestyle with the support of an early diagnosis.

### DATASET DESCRIPTION

The dataset consists of 13 features to be used for evaluating and prediction. These features are discussed below:

Age: Patients age in years

Sex: Patients gender ( 1=male, 0=female)

Cp: Type of chest pain experienced by the patient ( v1: typical angina, v2: atypical angina, v3: non-anginal pain, v4: asymptomatic)(v= value)

Trestbps: Represents the resting blood pressure of the patient given in mm Hg

Chol: Serum cholesterol level of the patient given in mg/dl

Fbs: It refers to the fasting blood sugar levels of the patient (>120 mg/dl = 1, 0 = otherwise)

Restecg: Resting electrocardiographic results (v0: normal, v1: having ST-T wave abnormality, v2: demonstrating possible or confirmed hypertrophy of the left ventricle by Romhilt-Estes’ criteria)(v=value)

Thalach: Maximum heart rate achieved by the patient

Exang: Exercise-induced angina (1 = yes, 0 = no)

Oldpeak: It represents the ST depression induced by exercise relative to rest

Slope: Slope of the peak exercise ST segment (value 1: upsloping, value 2: flat, value 3: downsloping)

Ca: The number of major vessels (0-3) coloured by fluoroscopy

Thal: Thallium stress test result (value 3: normal, value 6: fixed defect, value 7: reversible defect)

In-depth feature description:

Value 0: <50% narrowing of the diameter (negative for illness) Value 1: > 50% narrowing of the diameter (indicating illness).

Features Related to ECG: slope, old peak, and restecg

Restacg = A non-invasive examination called resting electrocardiography (ECG) can identify problems such as bundle branch blockages, left ventricular hypertrophy, signs in cases of arterial cardiac disease, and arrhythmias. 0: standard,1. exhibiting aberrant ST-T waves; 2. demonstrating likely or certain left ventricular hypertrophy according to Estes' criteria.

Features Affecting Blood: Trestbps, thalach,fbs,chol

Trestbps: Blood Pressure at Rest Hypertensive individuals are more vulnerable to peripheral vascular disease, heart attacks, strokes, renal disease, aneurysms, and heart disease due to blood vessel stress. In line with this, long-term health issues like diabetes, renal disease, sleep apnea, and elevated cholesterol elevate the likelihood of getting hypertension.

Thalach: Heart Rate Maximum : Research indicates that a mere 10 pulses/min rise in heart rate is linked to a minimum 20% higher potential for cardiac death. This risk increase is comparable to that seen with a 10 mm Hg elevation in the systolic blood pressure.

Fasting Blood Sugar (fbs) an examination used to evaluate levels of blood sugar. Higher levels are linked to insulin resistance and diabetes, conditions where the body is unable to process sugar (such as obesity).

Chol: serum cholesterol levels According to traditional wisdom, having high levels of low-density lipoprotein (LDL) cholesterol raises your risk of dying from cardiovascular disorders including heart disease.

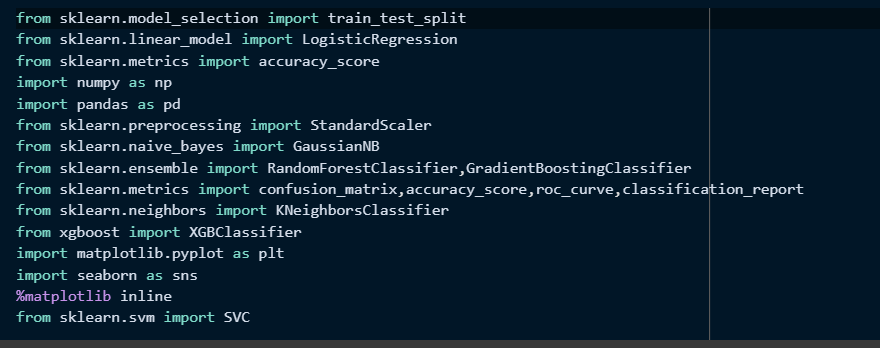
Characteristics of Pain and Defects: cp,exang,thal,ca

Exang: Angina Induced by Activity Chest pain from angina is brought on by a decrease in the flow of blood through the heart muscles. It is a warning indication that you may be at danger of a heart attack or stroke, but it is typically not life-threatening.

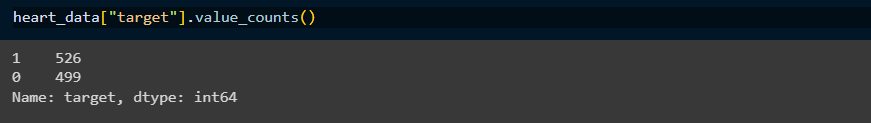
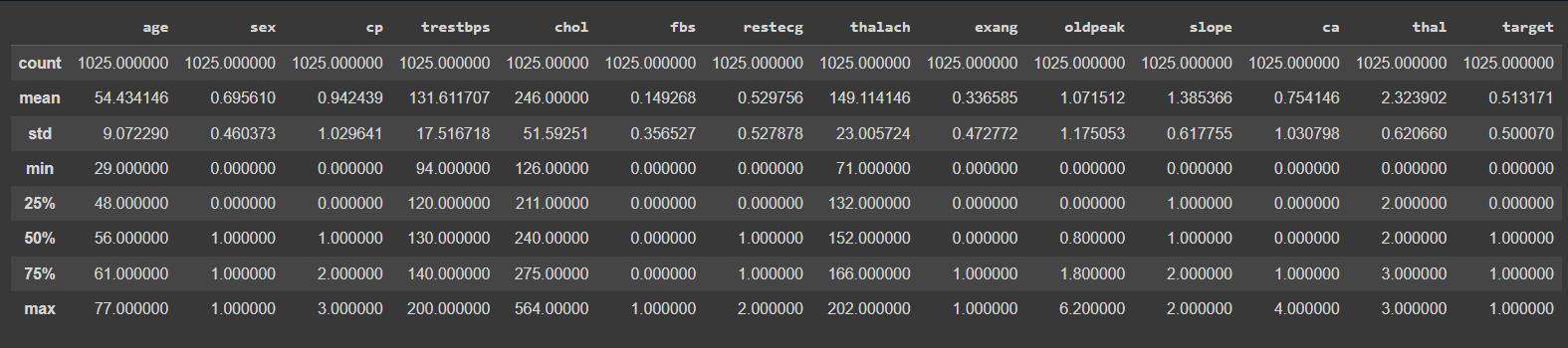
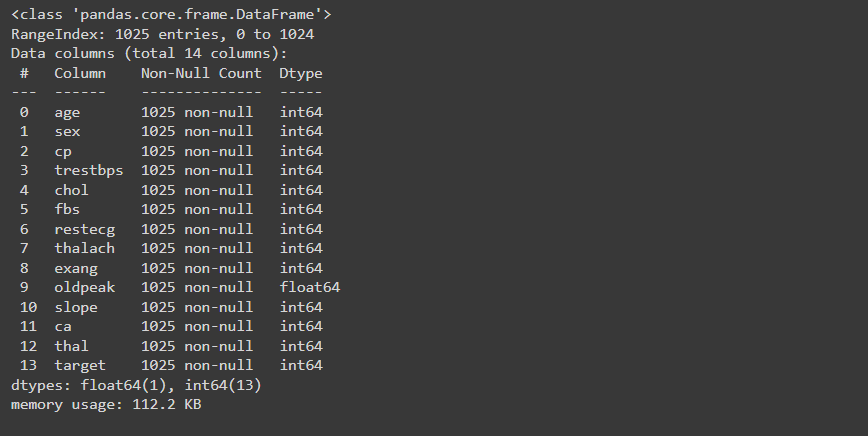
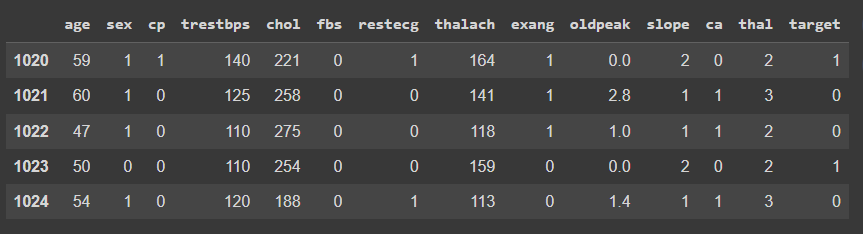
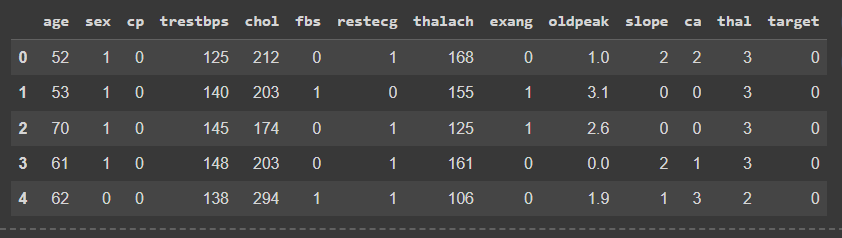
Ca: quantity of large containers (0–3), hued by fluoroscopy The limited sample size of the study population may have contributed to the medium influence of the number of main vessels colored by fluoroscopy on the diagnosis of CAD. It could have to do with the fact that the system learned it from the training set and that the sensitivity of fluoroscopy can be as low as 35% in some situations.

### CODE FOR THE MODEL

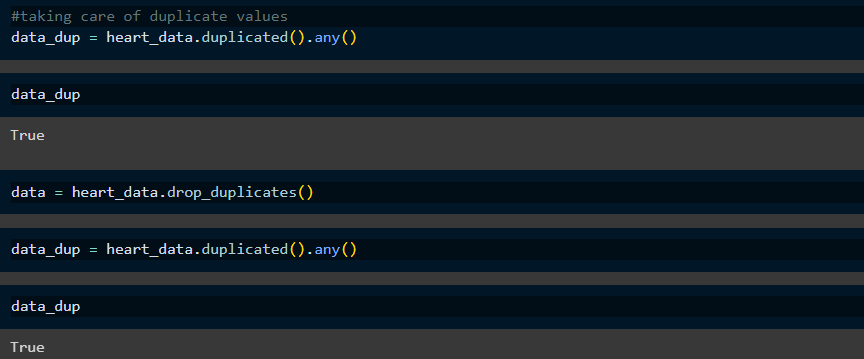
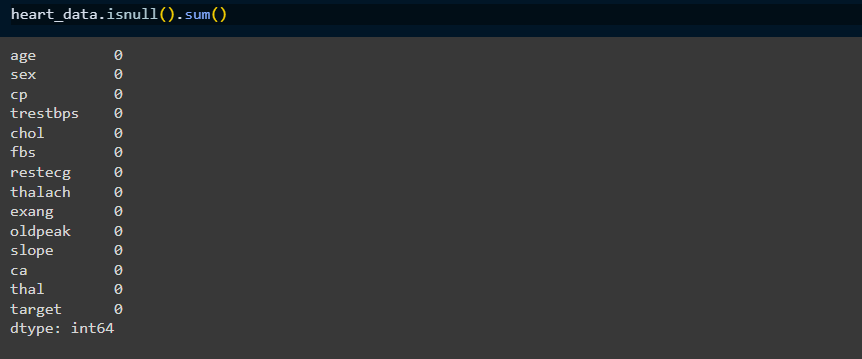
#### IMPORTING LIBRARIES



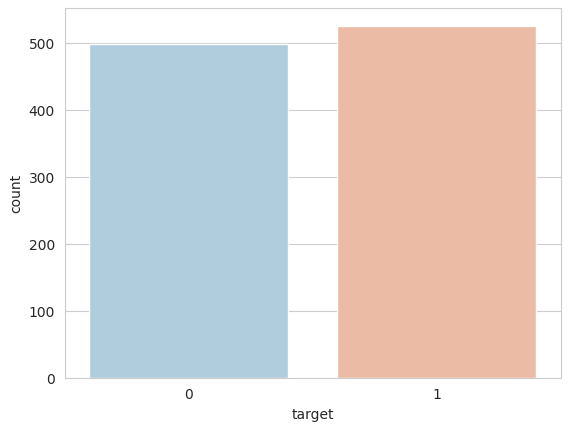
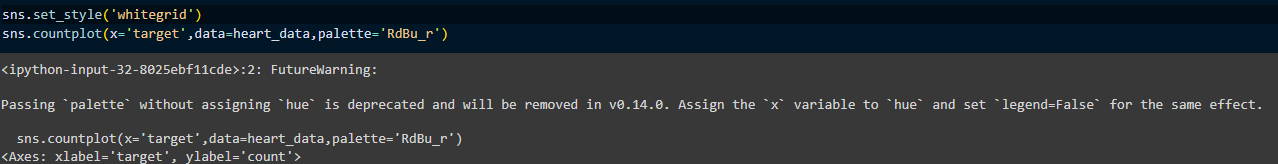
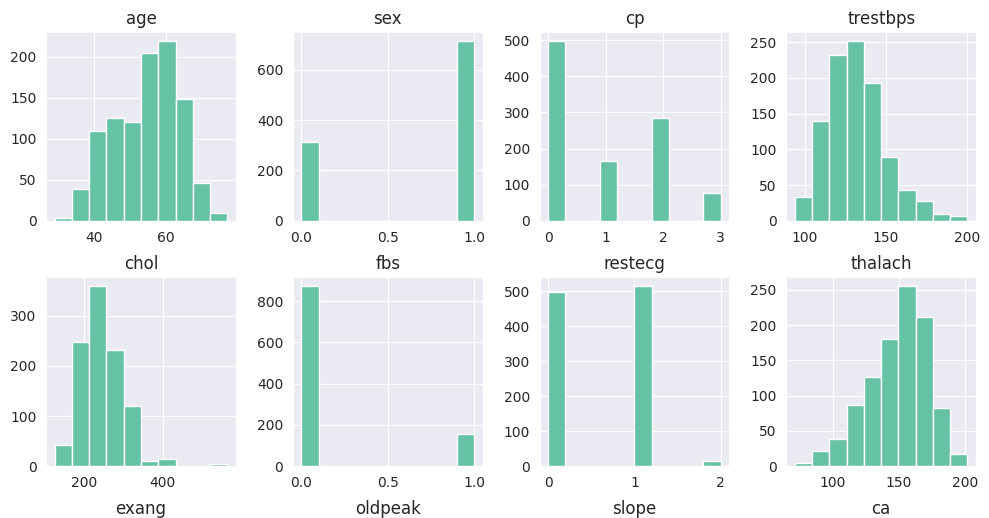
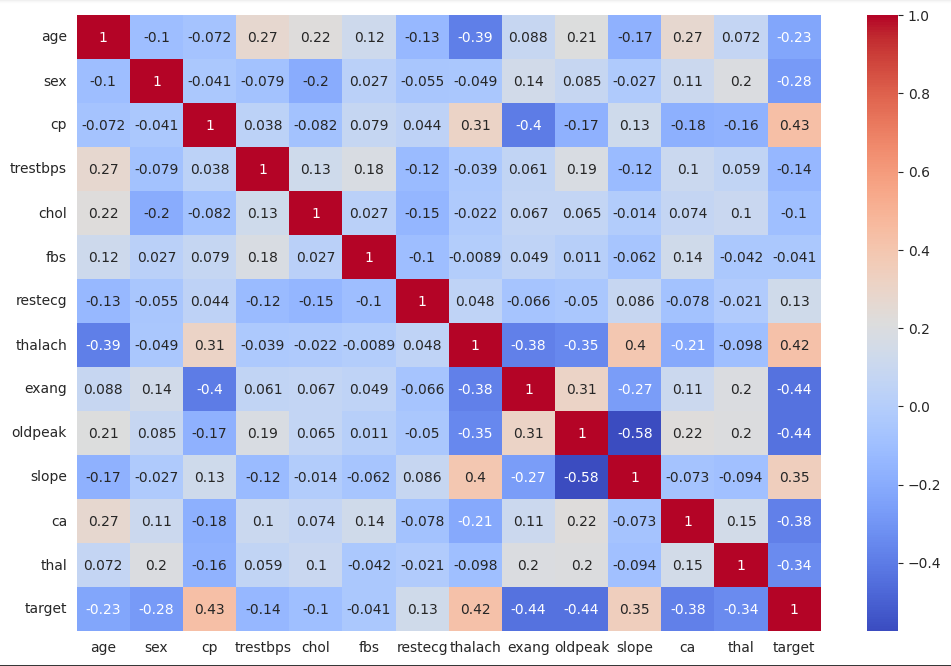
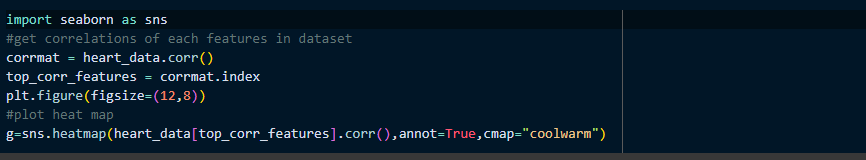
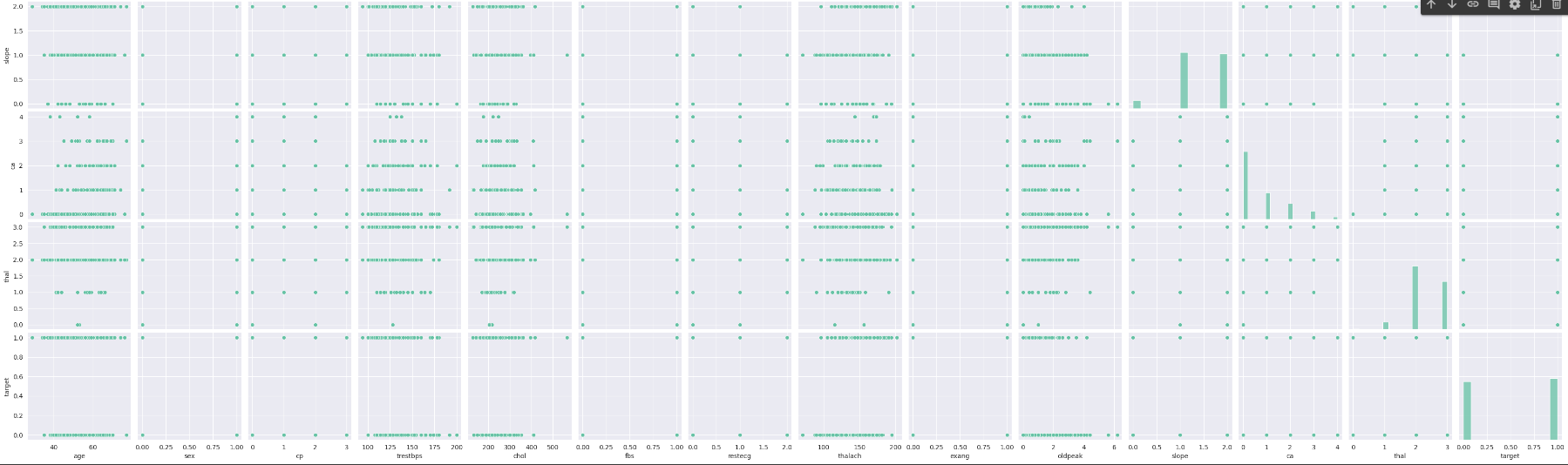
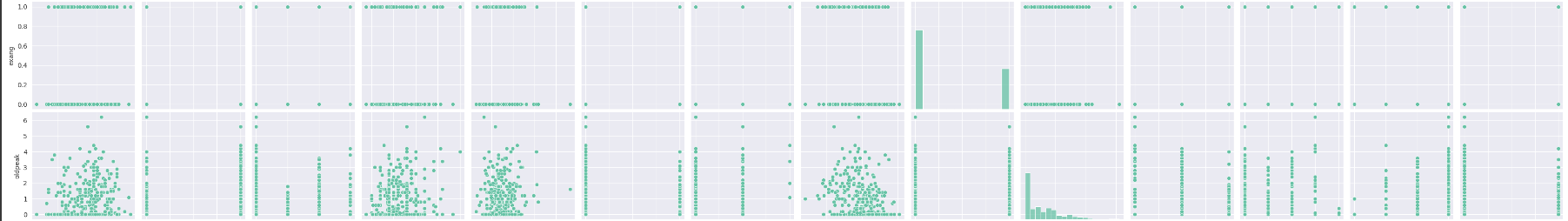
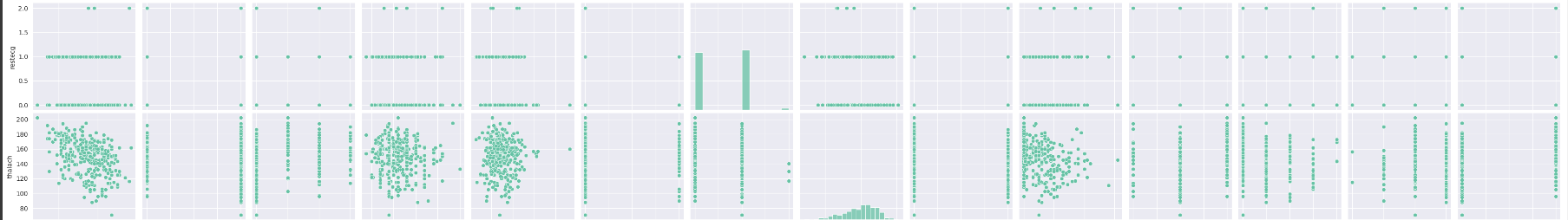
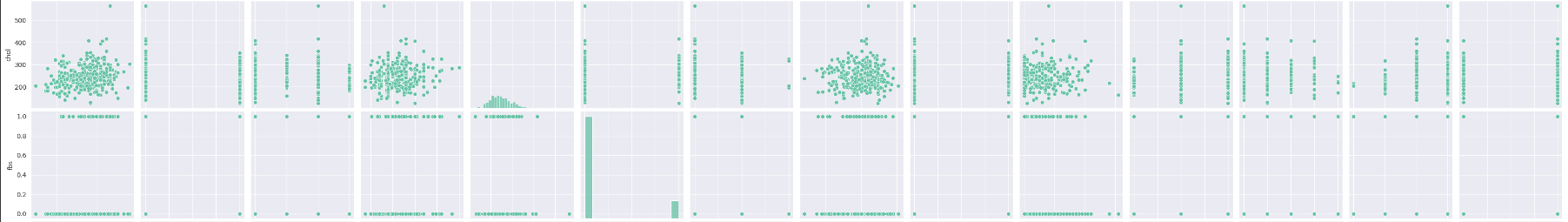
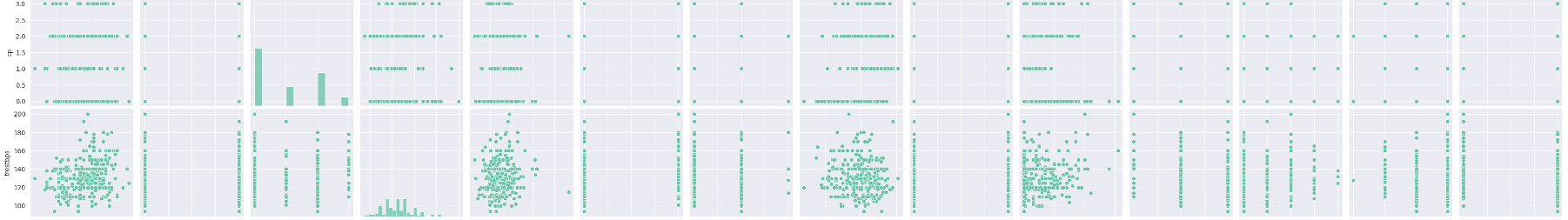
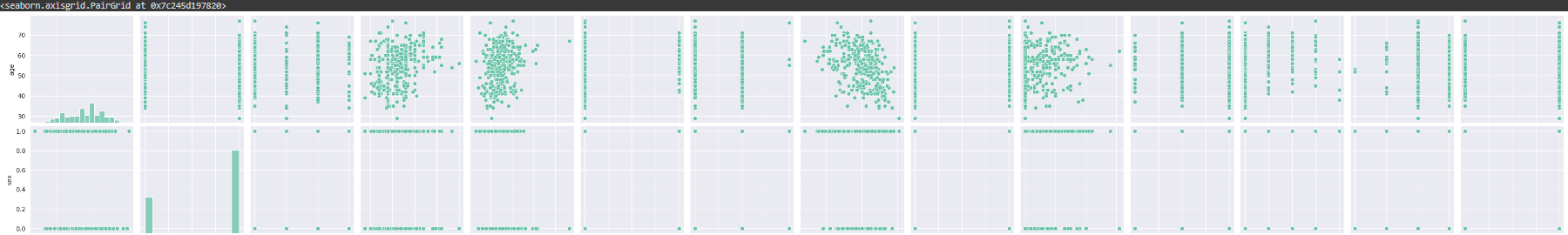
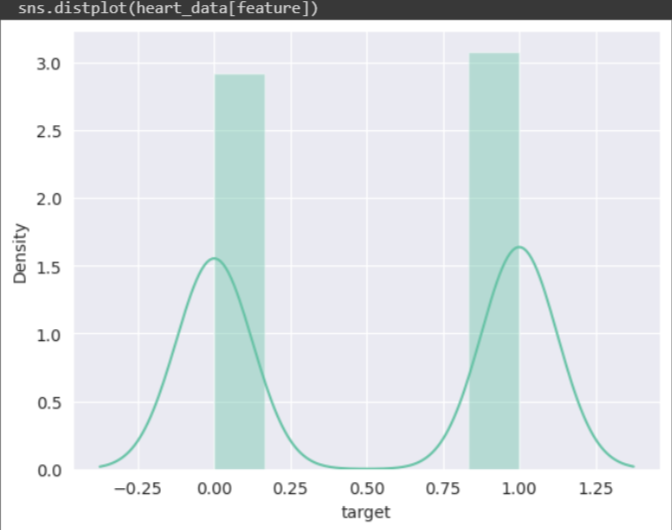
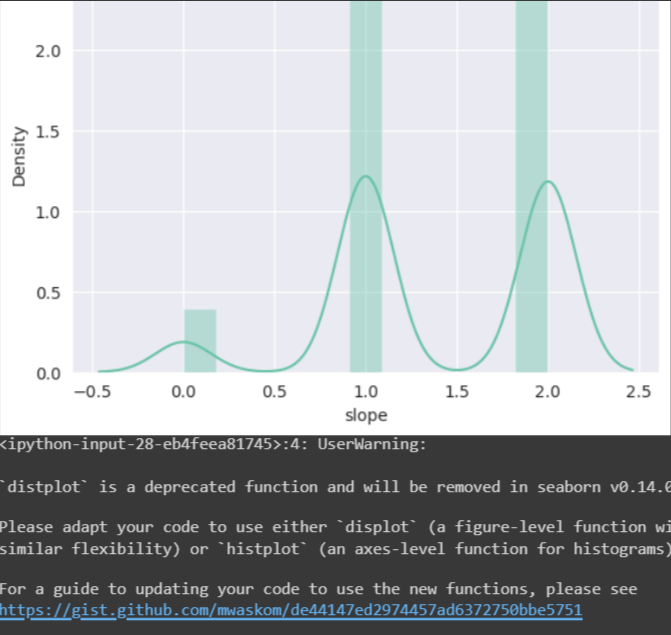
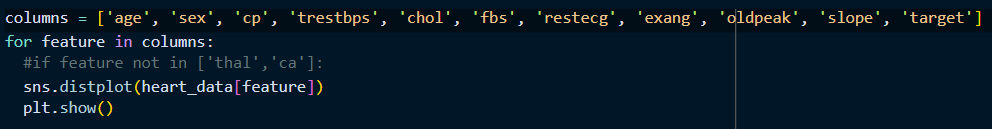
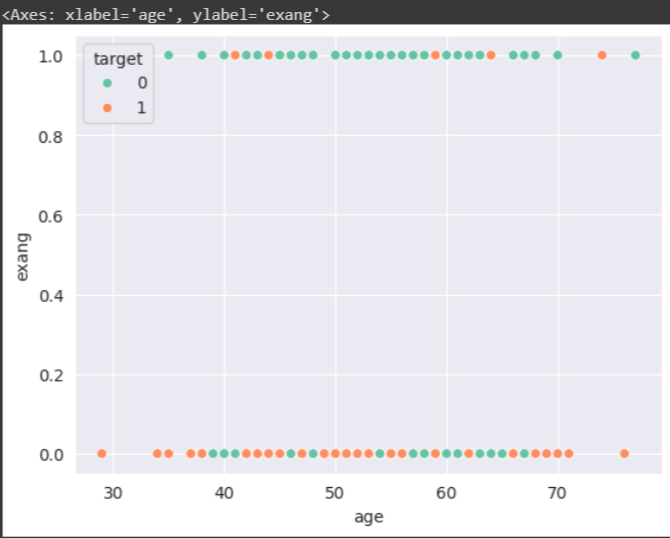
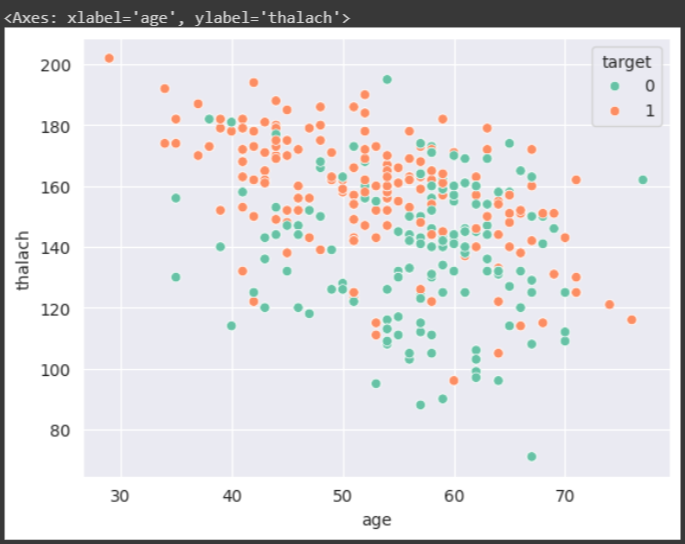
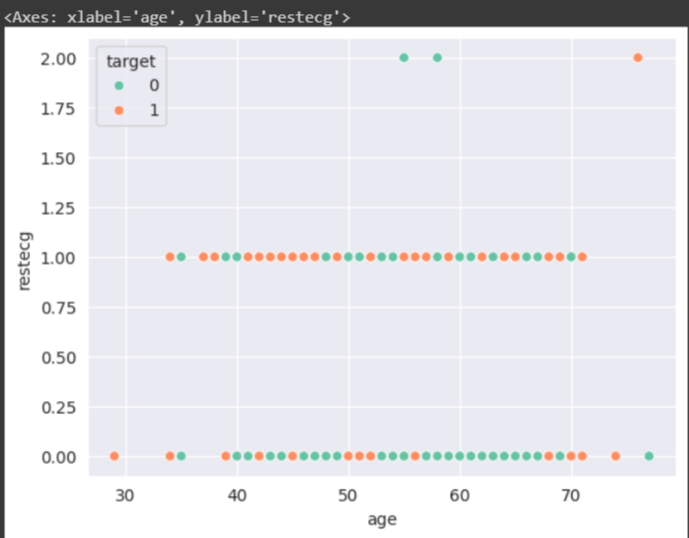
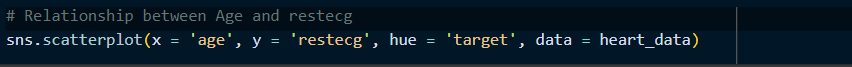
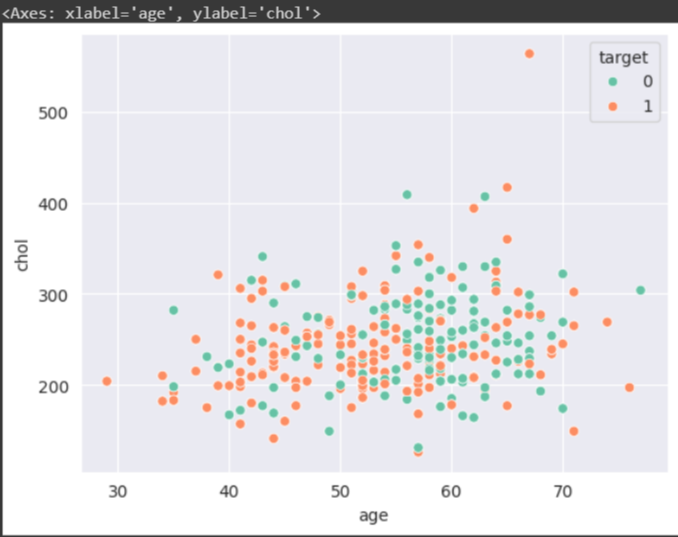
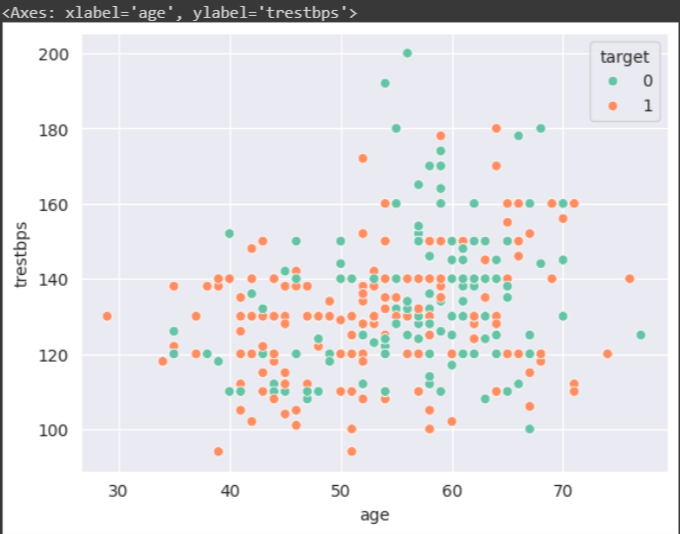
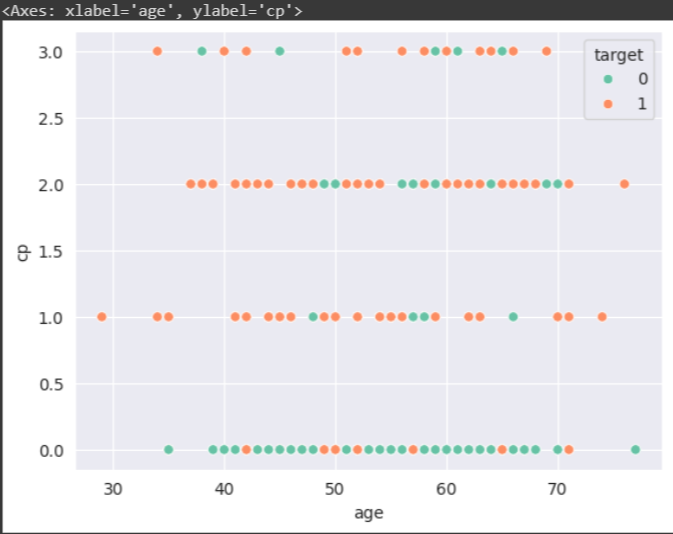
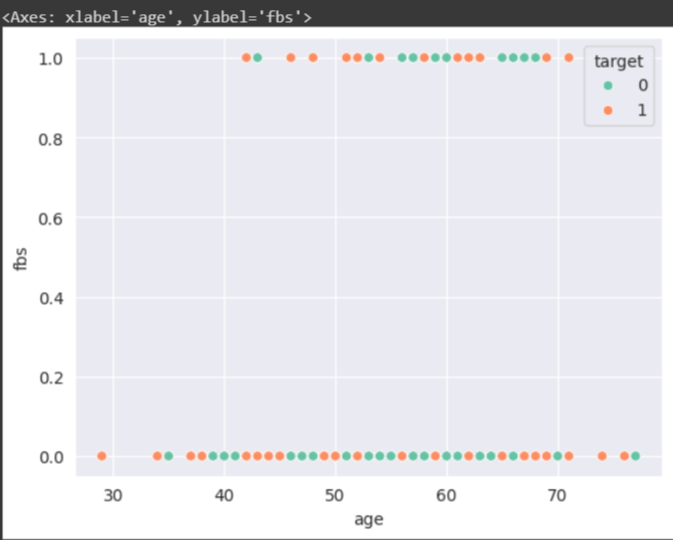
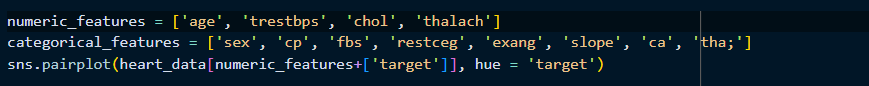
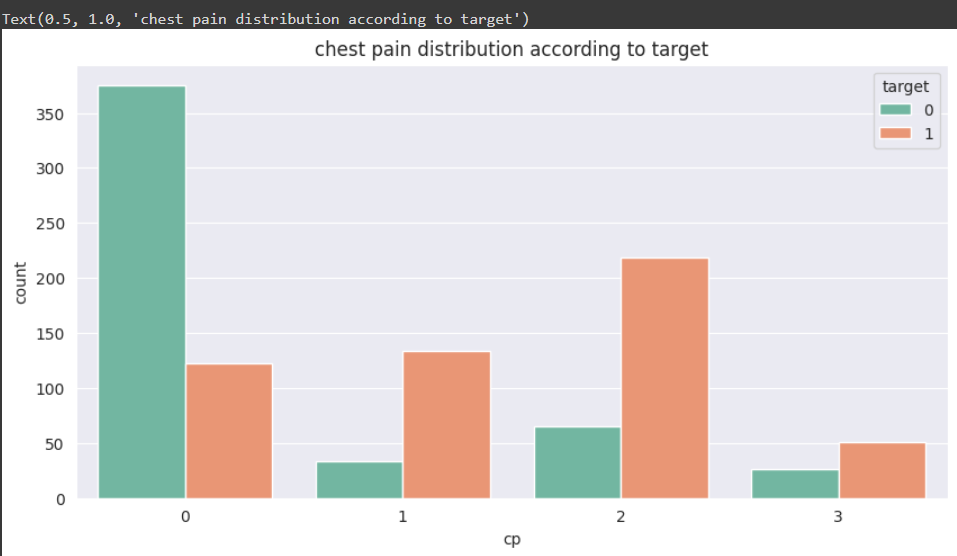
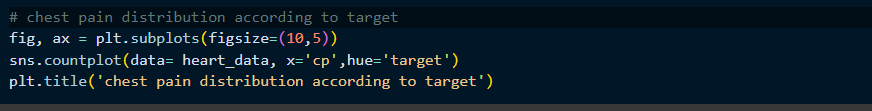
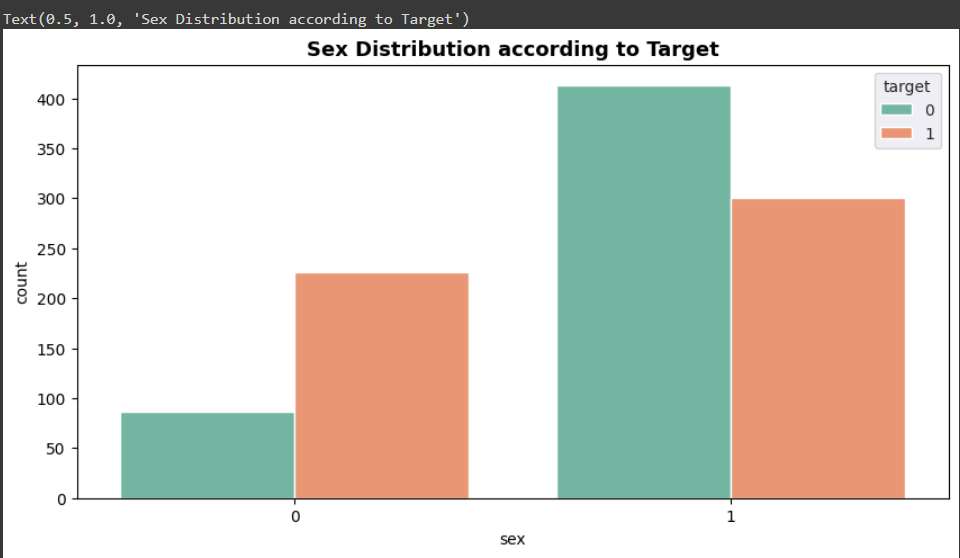
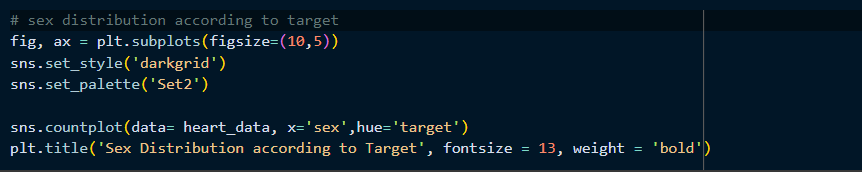
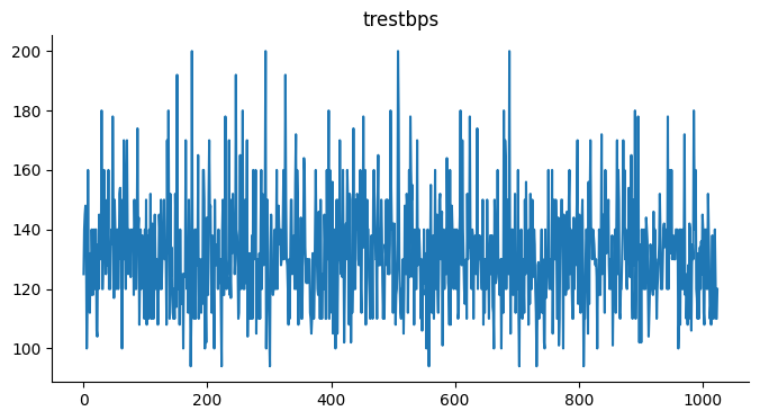
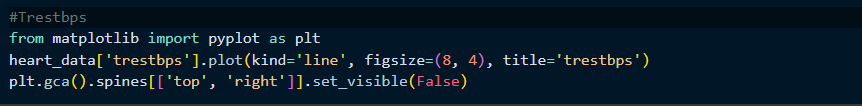
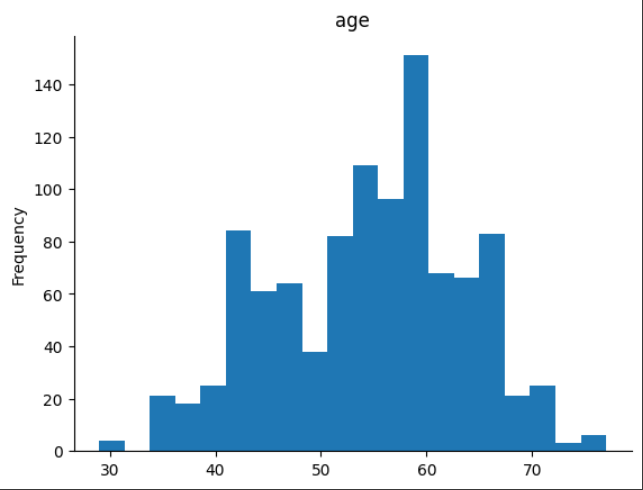
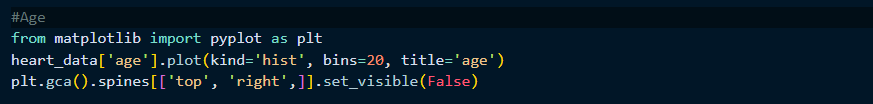
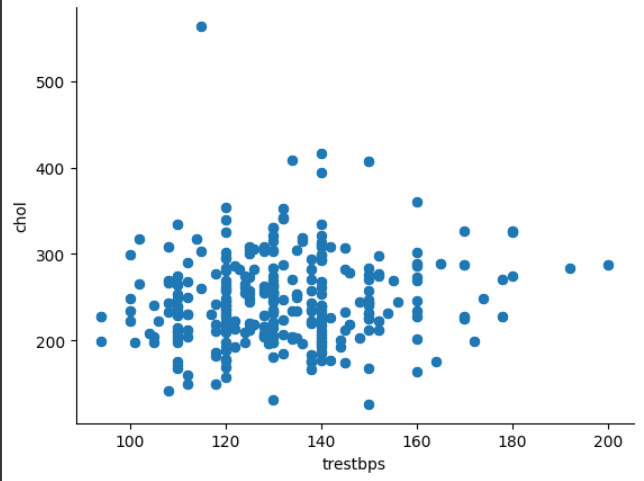
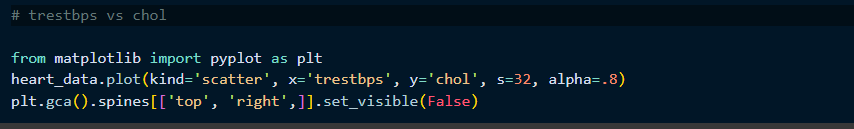
#### DATA COLLECTION AND DATA DESCRIPTION



#### DATA CLEANING

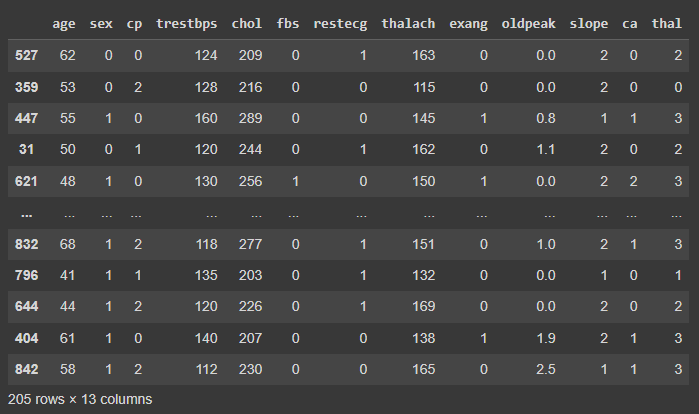
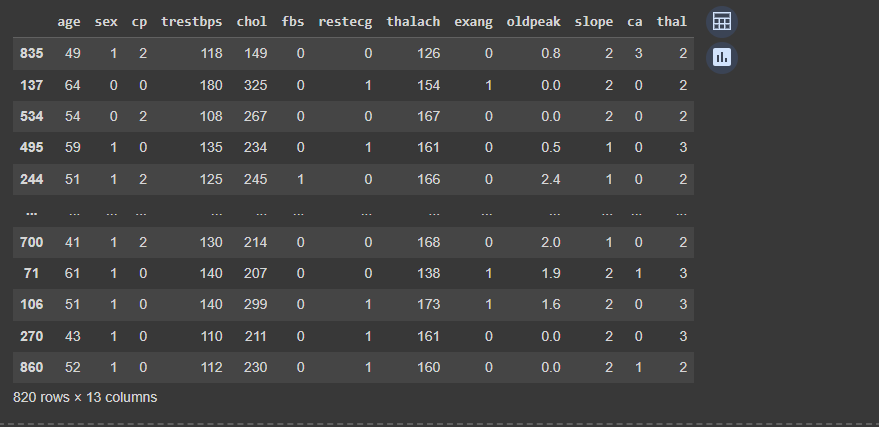


#### STATISTICAL ANALYSIS (EDA)



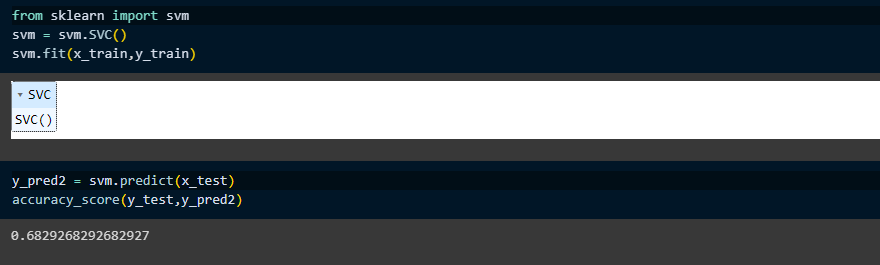
#### **DATA PROCESSING**





#### **LOGISTIC REGRESSION**

#### **SVC**



#### **KNEIGHBORS CLASSIFIER**

#### **DECISION TREE CLASSIFIER**

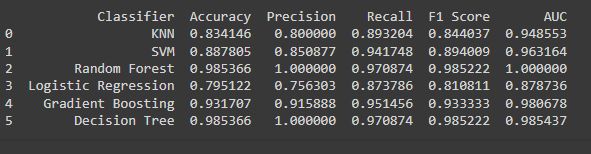
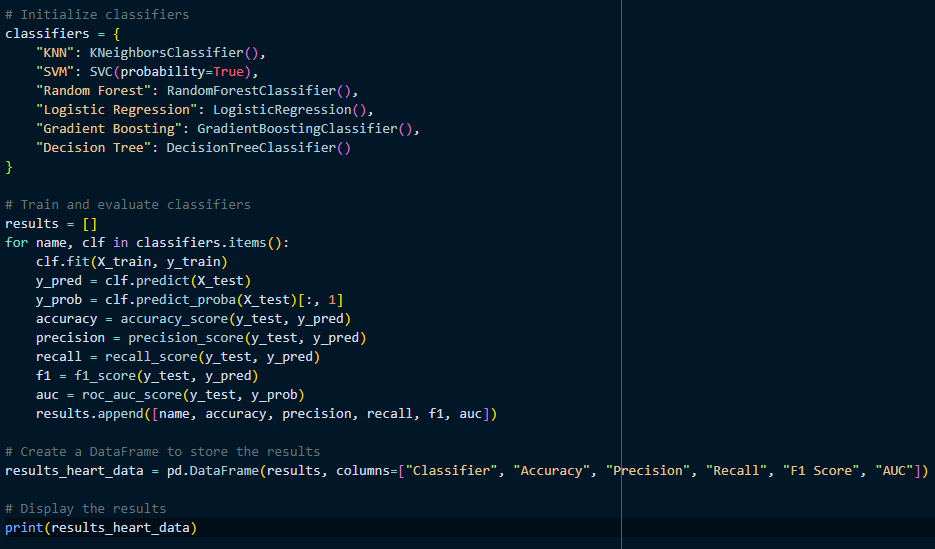
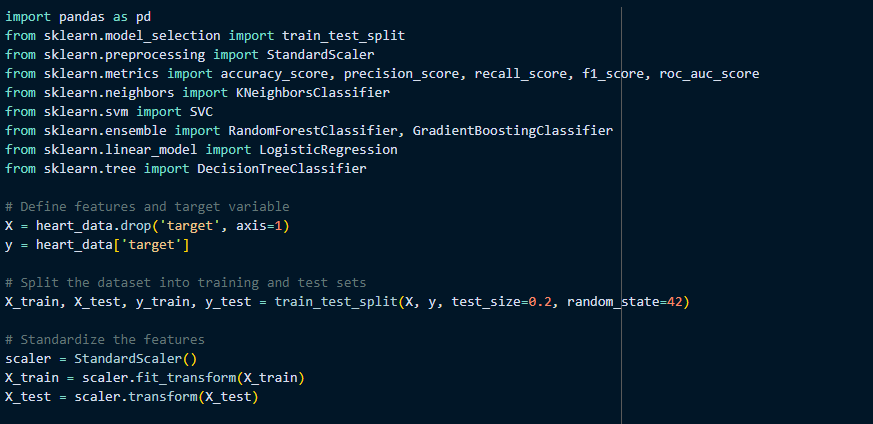
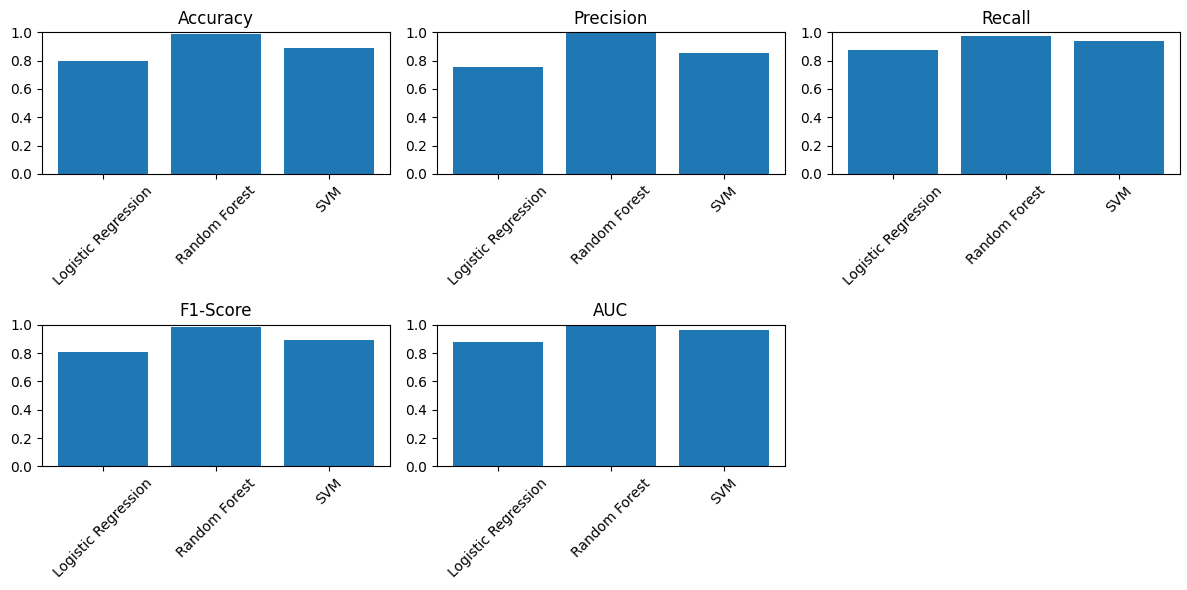
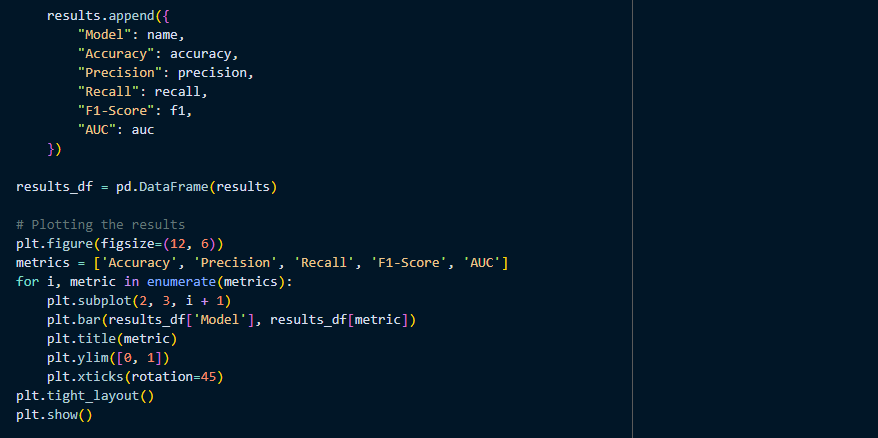
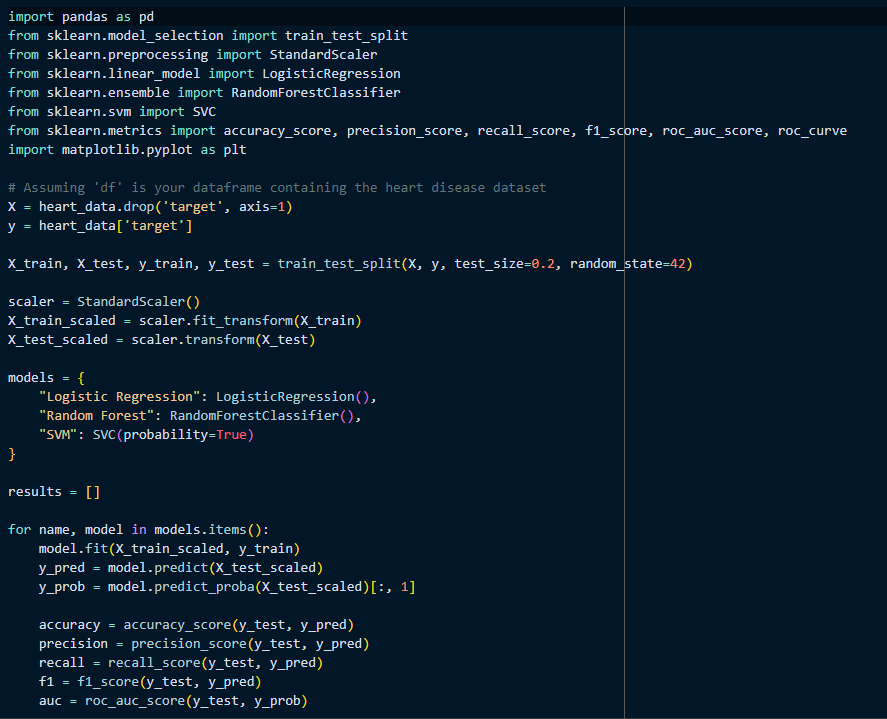
#### **RANDOM FOREST CLASSIFIER**

#### **GRADIENT BOOSTING CLASSIFIER**

#### **DISTRIBUTION**

#### **CORRELATION MATRIX**

#### **EVALUATION METRICS AND CONFUSION MATRIX**



### **CONCLUSION**

**UNDERSTANDING THE DIFFERENT MACHINE LEARNING TECHNIQUES AND ITS APPLICATIONS WITH RESPECT TO THE PROBLEM STATEMENT.**

Through the examination of numerous risk variables and clinical indicators, Machine Learning methodologies are essential in the prognosis of cardiovascular illness. Numerous machine learning algorithms are utilized to evaluate medical data and forecast the likelihood that a person will get heart disease in the context of foreseeing its risk. Here are the description of a few popular machine learning methods and how they're used to predict heart disease:

The Logistic Regression Model

Application: Based on patient characteristics, logistic regression is a popular binary classification approach that works well for estimating the likelihood of cardiac disease.

How to Use It: A binary outcome (the existence or lack of cardiac problems) is modeled using logistic regression using one or more predictor variables (e.g., age, gender, cholesterol level).

Benefits: Implementing logistic regression to assess can help model the chance of heart disease since it is straightforward, easy to understand, and computationally efficient.

Random Forest:

Application: Several decision trees are used in the random forest ensemble learning technique to increase prediction accuracy.

How it Works: During training, random forests construct several decision trees and produce two outputs: the average prediction of an individual (a regression analysis) tree or how the classes are structured (classification).

Benefits: In addition to being strong against overfitting and adept at handling nonlinear interactions, random forest also yields feature relevance ratings that can be used to pinpoint critical heart disease risk factors.

SVM, or support vector machine:

Application: Predicting the presence or lack of the disease is one of the many binary classification tasks that may be handled by SVM, a potent supervised learning algorithm.

How it Works: SVM maximizes the margin between the classes by locating the hyperplane within the feature area that optimally divides the classes.

Benefits: SVM performs well with nonlinear data, is less prone to overfitting, and is efficient in high-dimensional domains.

Gradient Boosting:

Application: Gradient boosting is an ensemble learning method that creates models one after the other, fixing each other's mistakes.

How it Works: To build a powerful prediction model, gradient boosting combines several weak learners, such as decision trees.

Benefits: Gradient boosting frequently produces excellent prediction accuracy, manages intricate relationships with ease, and offers feature importance data.

Decision Tree:

Application: Decision trees can be employed to forecast whether cardiac disease would occur or not based on patient characteristics because they can be utilized for both regression and classification tasks.

How it Works: Decision trees produce a structure like a tree by dividing the information iteratively into subsets according to the values of input features.

Benefits: Decision trees manage data that is both category and numerical, are resistant against outliers, and are easy to comprehend and depict.

K-Nearest Neighbors (KNN):

Application: Based on comparable patient characteristics, KNN is an easy-to-understand classification method that can be used to forecast cardiac disease.

How it Works: A new data point is classified by KNN through the majority vote of its K nearest neighbors in the feature space.

Benefits: KNN can identify complicated patterns in the data, is simple to comprehend and apply, and doesn't require training.

**UNDERSTANDING THE IMPORTANCE OF SIMPLE LINEAR REGRESSION IN PREDICTING NEW OBSERVATIONS WITH RESPECT TO THE PROJECT. CAN SIMPLE LINEAR REGRESSION BE USED, YES/NO, WHY.**

Modeling the link amongst only one independent factor (predictor) with a dependent variable (outcome) is accomplished statistically via simple linear regression. Given how complicated and influenced by many different aspects of cardiology is, simple linear regression might not prove to be the ideal method to anticipate heart disease. It is imperative to acknowledge its significance and constraints within this particular setting, though.

Recognizing the Value of Basic Linear Regression

Basic Analysis: An easy technique to comprehend the correlation between the two parameters is to apply simple linear regression. For example, simple linear regression might assist us in analyzing the way variations in blood pressure may impact the probability of heart disease if we are looking at a dataset with one predictor variable (blood pressure, for example) alongside an outcome indicator (threat of heart illness).

Interpretability: It is easy to understand the values of the coefficients derived from basic linear regression. These correlation coefficients may represent the direction and intensity of the link between an individual predictor (such as cholesterol levels) in addition to the outcome (such instance the risk of heart disease) for the framework of heart disease prediction.

Baseline Model: More sophisticated techniques can be contrasted to a baseline model of simple linear regression. Although it might not fully encapsulate the intricacies of cardiac disease prognosis, it might offer a foundation for examination and contrast.

Assumption Testing: In order to determine whether the model is acceptable for the data, simple linear regression makes it possible to test assumptions like linearity and homoscedasticity. There is optimism regarding the credibility of the link amongst the variables that predict and those that determine if these presumptions are true.

Is Heart Disease Prediction Possible with Simple Linear Regression?

The answer is no in Most Cases: Since heart conditions are complicated illnesses, prognosis usually takes into account a number of factors, including age, gender, blood pressure, cholesterol, and other factors. The capacity of simple linear regression to represent intricate correlations between several predictors and the end variable is constrained.

Yes, under certain conditions: In very particular situations when there is substantial prior proof from a linear correlation involving just one indicator with the risk of heart disease, simple linear regression may be useful. Even in these situations, though, it's important to take into account any confounding factors and the drawbacks of applying a crude model to a complicated illness.

Even though a single predictor can be used with simple linear regression, more complex techniques like logistic regression, multiple line regression, along with algorithms that use machine learning like random forest algorithms or neural networks are frequently added to take into account the multiple contributors of heart disease and increase prediction accuracy.

In conclusion, due to its inherent simplicity, basic linear regression is typically not appropriate for anticipating the progression of heart disease, even though it can provide insightful information about the connection between two variables and act as a jumping off point for further study. Rather, for precise prediction and risk assessment, more intricate techniques taking into account several predictors are usually used.

**UNDERSTANDING THE IMPORTANCE OF MULTIPLE LINEAR REGRESSION IN PREDICTING NEW OBSERVATIONS WITH RESPECT TO THE PROJECT. CAN MULTIPLE LINEAR REGRESSION BE USED, YES/NO, WHY.**

Multiple linear regression is a useful tool when it comes to predicting heart disease. Let's examine the multiple linear regression phenomenon in this context and its significance.

A statistical technique called multivariate linear regression, one may examine the connection between several independent variables, or predictors, along with a dependent variable, or result.It is predicated on the idea that the connection between the indicators and the outcome is linear.. Multiple linear regression is applied to determine which combination of variables (such as age, blood pressure, cholesterol levels, etc.) is linked to an elevated danger of cardiac conditions in the context of heart disease prediction.

The significance of identifying relevant predictors in the estimation of the cardiac disease:

Researchers can identify the factors that are strongly linked to the probability of the ailment by using multiple linear regression. Researchers can determine the most significant predictors by examining a dataset that includes data on a variety of health indicators and outcomes (such as the presence or absence of heart disease). This realization aids in comprehending the fundamental causes of cardiac disease.

Assessing connections: Multiple linear regression is employed to quantify the direction and intensity of the associations between important variables and the chances for the disease after they have been identified. For instance, it can calculate the percentage change in blood pressure, cholesterol, etc. that results in an increase or decrease in the risk of heart disease. When creating preventive measures and prioritizing interventions, this knowledge is essential.

Prediction of New Observations: Using historical data to construct a regression model, one can use it to forecast the probability of cardiac sickness in fresh observations. In clinical settings, this predictive ability is crucial for identifying those who are at high peril of heart illness. With the use of these forecasts, medical professionals can reduce risk by implementing timely interventions, such as medication or lifestyle changes.

Evaluation of Model Performance: The prediction model's performance can be evaluated by researchers using multiple linear regression. The degree to which the model fits the data and forecasts results can be determined using metrics like root mean square error, mean squared error, and R-squared (coefficient of determination). Assessing model performance facilitates improvements and guarantees the model's accuracy in forecasting the risk of the disease.

Can Multiple Linear Regression Be Used?

Indeed, studies aimed at predicting cardiac disease can make use of multiple linear regression. Nevertheless, depending on the complexity of the knowledge along with the nature of the issue, it's critical to recognize its limitations and take into account alternative predictive modeling approaches, such as Logistic Regression, Decision Trees, Random Forests, or SVMs. For accurate results, researchers should also confirm that the multiple the suppositions of linear regression with regards to homoscedasticity, linearity, and independence of errors are satisfied. It may be more acceptable to use alternate modeling methodologies if certain assumptions are not met.

**FOR THE PROJECT DEFINITION, EXPLAIN THE IMPORTANCE OF PARAMETER ESTIMATION. CHECK HOW IT AFFECTS THE OVERALL RESULTS WHEN WE CHANGE THE VALUES OF DIFFERENT PARAMETERS.**

Parameter estimate is essential to assessing the efficacy and precision of the predictive model in a machine learning-based cardiac disease prediction project. This explains its significance and how adjusting the values of various parameters impacts the final outcomes:

Estimating Parameters Is Important for:

Model Tuning: In order to maximize performance, machine learning algorithms frequently contain parameters that need to be adjusted. In order to produce the most accurate forecasts, parameter estimation entails determining the ideal values for these parameters. This can entail modifying settings for algorithm complexity, regularization strength, or feature selection in the context of heart disease prediction.

Broad Generalization Performance: The ability of the model to generalize to previously unobserved data is directly impacted by the standard of parameter estimation. An overfitting phenomenon occurs when a model memorizes the training set of data instead of identifying underlying patterns; this can be avoided with carefully calibrated parameters. An inaccuracy in the assessment of the likelihood of cardiovascular disease can result from overfitting, which can also cause poor performance on new data.

Robust: The robustness of the predictive model is influenced by well-estimated parameters. Reliability and consistency of predictions are increased the moment a model is resilient because it is less susceptible to minute changes or data noise. Robustness is the quality that guarantees the model's performance in predicting heart problems from a variety of datasets and demographics.

Interpretability: The machine learning model's interpretability may be impacted by parameter estimates. Researchers can identify the elements that are most useful in forecasting the risk of heart disease by fine-tuning parameters. This allows them to get important insights into the underlying processes that affect cardiovascular health.

The Effect of Modifying Parameter Values:

Model Complexity: Modifying a few parameters can have a big impact on how well a model performs. Examples of these parameters include the number of features or the algorithm's complexity (e.g., modifying the number of hidden layers in a neural network or the depth of a decision tree). While adding complexity could improve the fit to the training set, it also raises the possibility of overfitting.

Regularization: The degree of penalty imposed to model coefficients is controlled by parameters associated with regularization techniques (e.g., L1 or L2 regularization). The trade-off between model complexity and generalization performance is impacted by changing these parameters. Complex models are penalized more severely by stronger regularization, which helps to avoid overfitting.

Feature Importance: Measures of feature importance are provided by a few machine learning algorithms, including decision trees and ensemble techniques like random forests. The weights given to various features can be altered by modifying the parameters associated with these algorithms, which may have an impact on the model's capacity to forecast cardiac disease and its interpretability.

Algorithmic Effectiveness: The efficiency of various algorithms is influenced by certain parameter configurations. Neural networks, for instance, have architecture-related parameters such as the quantity of nodes and layers, while support vector machine models comprise features such as the kernel type and the regularization value C. Modifying these parameters may have an impact on the model's capacity to represent intricate relationships seen in the data.

In conclusion, creating machine learning frameworks for the heart disease prognosis requires careful consideration of parameter estimation. The best possible model performance, robustness, interpretability, and generalization capacity are all ensured by carefully adjusted parameters. For the purpose of creating precise and trustworthy predictive models for use in healthcare applications, it is crucial to comprehend the effects of altering parameter values.

**COMPARE DIFFERENT MACHINE LEARNING ALGORITHMS USED THROUGH EVALUATION METRICS. SHOW WHICH ML ALGORITHMS WORK BEST FOR THE PROBLEM**

| No. | Classifier | Accuracy | Precision | Recall | F1 Score | AUC |
| --- | --- | --- | --- | --- | --- | --- |
| 0 | KNN | 0.834146 | 0.800000 | 0.893204 | 0.844037 | 0.948553 |
| 1 | SVM | 0.887805 | 0.850877 | 0.941748 | 0.894009 | 0.963164 |
| 2 | Random Forest | 0.985366 | 1.000000 | 0.970874 | 0.985222 | 1.000000 |
| 3 | Logistic Regression | 0.795122 | 0.756303 | 0.873786 | 0.810811 | 0.878736 |
| 4 | Gradient Boosting | 0.931707 | 0.915888 | 0.951456 | 0.933333 | 0.980678 |
| 5 | Decision Tree | 0.985366 | 1.000000 | 0.970874 | 0.985222 | 0.985437 |

For numerous reasons, the Random Forest algorithm frequently produces better results for predicting cardiac disease than other algorithms:

Several Decision Trees get implemented in Random Forest ensemble learning techniques for generation of predictions. Random Forest enhances generalization performance and lowers the possibility of overfitting by combining the predictions of several trees.

In order to determine which features are most important for the prediction task, Random Forest offers a measure of feature importance. This can be especially helpful in the prediction of cardiac disease, as pinpointing the most pertinent risk variables is crucial for a precise diagnosis.

Random Forest is naturally resistant to data noise and anomalies. For reasons such as errors in measurement or individual variation in patient data, medical datasets might contain intrinsic unpredictability and noise, which this resilience aids in managing.

Complex, non-linear correlations amongst features of the input and a target variable can be captured using Random Forest. Random Forest is an excellent modeling technique to forecast cardiac illness where the relationship between risk variables (e.g., age, blood pressure, cholesterol levels) and the presence of cardiac conditions may not be linear.

Imputation is not necessary when using Random Forest for handling values that are missing in the data. This is helpful for medical datasets, because missing data is frequently observed for a variety of reasons, including patient noncompliance or mistakes in data collection.

Random Forest may be effectively applied to big datasets and is parallelizable. Because of its scalability, Random Forest is appropriate for use in real-world scenarios and can handle massive healthcare data sets with hundreds of thousands of different samples and countless features.

Overall, Random Forest is a strong and useful algorithm for heart disease prediction tasks due to its capacity to reduce overfitting, recognize significant features, manage noise, capture non-linear correlations, handle missing values, and be scalable.