**ALGORITHMIC ANALYSIS OF STUDENT PERFORMANCE PREDICTION MODELS**

A Capstone Phase-II project report submitted in partial fulfilment of requirement for the award of degree

BACHELOR OF TECHNOLOGY

in

COMPUTER SCIENCE & ENGINEERING

By

M.SRINITHA (2203A52201)

N.SRI VIDYA (2203A52229)

K. ABHINAYA SRI (2203A52241)

CH. PRANATHI (2203A52242)

Under the guidance of

Mr. Dr. E.L.N. Kiran Kumar

Associate Professor, CS & AI.

CERTIFICATE

This is to certify that this project entitled “ ALGORITHMIC ANALYSIS OF STUDENT PERFORMANCE PREDICTION MODELS ” is the bonafied work carried out by M.SRINITHA , N.SRI VIDYA , K. ABHINAYA SRI , CH. PRANATHI as a Capstone Phase-II project for the partial fulfilment to award the degree BACHELOR OF TECHNOLOGY in COMPUTER SCIENCE & ENGINEERING during the academic year 2023-2024 under our guidance and Supervision.

#### **Mr.E.L.N. Kiran Kumar** **Dr.M.Sheshikala**

Assoc. Prof. CS & AI, Assoc. Prof.HOD(CSE), SR UNIVERSITY, SR UNIVERSITY,

Ananthasagar, Warangal. Ananthasagar, Warangal

ACKNOWLEDGEMENT

We owe an enormous debt of gratitude to our project guide **Mr. Dr. E.L.N. KiranKumar**, **Assoc. Prof. CS and AI** as well as Head of the CSE Department **Dr.M.Sheshikala**, **Associate Professor** for guiding us from the beginning through the end of the Capstone Phase-II project with their intellectual advices and insightful suggestions. We truly value their consistent feedback on our progress, which was always constructive and encouraging and ultimately drove us to the right direction.

We express our thanks to project co-ordinators **Mr.Sallauddin** **Md, Asst. Prof, Y.Chanti** **Asst. Prof**. for their encouragement and support.

Finally, we express our thanks to all the teaching and non-teaching staff of the department for their suggestions and timely support.

CONTENTS ACKNOWLEDGEMENT I LIST OF FIGURES III LIST OF TABLES IV LIST OF ACRONYMS V Chapter No. Title Page No. 1 ABSTRACT 01 1.1 KEYWORDS 01 2 INTRODUCTION 2.1 OVERVIEW 02 3 PROBLEM STATEMENT 03 4 MOTIVATION AND SCOPE OF WORK 03 5 LITERATURE REVIEW 5.1 RELATED WORK 04 6 DATASET 08 7 PROPOSED METHODOLOGIES 7.1 DATA PRE-PROCESSING 12 7.2 SENTENCE BERT APPROACH 12 7.3 LONG SHORT-TERM MEMORY 13 7.4 COMPARED ALGORITHMS 7.4.1 KNN 15 7.4.2 SVM 15 7.4.2 DECISION TREE 15 7.5 HARDWARE AND SOFTWARE TOOLS 16 8 RESULTS & DISCUSSION 17 9 CONCLUSION 30 10 FUTURE WORK 30 REFERENCES 31

**LIST OF ACRONYMS**

• MLP - Multi-Layer Perceptron• LR - Logistic Regression• SVM - Support Vector Machine• LGBM - Light Gradient Boosting Machine• DT - Decision Tree• NB - Naive Bayes

•XGB – XGBoost

**1.ABSTRACT**

The dataset called Higher Education Students Performance Evaluation is designed for students in the Faculty of Engineering and Faculty of Educational Sciences. This data was collected in 2019 with the aim to forecast end-of-term performances among students by means of machine learning techniques. The dataset contains 31 variables that cover a range of demographic, socio-economic, educational and behavioral characteristics such as student age, gender, high-school type, scholarship status, parental education level attained by parents or guardians with whom they live most often during the school year , study habits (e.g., hours per day spent studying), and academic achievements so far (e.g., GPA). The response variable is OUTPUT Grade which represents final grades achieved by students falling into one of seven categories: F – Fail; D -Poor; C -Satisfactory; B – Good; A – Very Good; AA – Excellent. Therefore this data set can be used by teachers together with various machine learning algorithms not only to identify factors influencing success in higher education but also discover how best they can assist those learners who may be struggling most based on their unique needs.

**KEYWORDS**

Index Terms—Role, collaborative leadership, fine of higher education

**2. INTRODUCTION:**

Decision-making, policy formulation, and intervention strategies to improve educational outcomes for all students. In order to gain knowledge about what affects students’ academic performance in modern schools — teachers, administrators, and other interested parties must know the factors that influence it. The Higher Education Students Performance Evaluation dataset contains a variety of attributes relating to socio-economic status as well as personal development; this includes information like age groupings, gender identification etc. These datasets were collected among individuals enrolled at two faculties during 2019 – faculty for engineering sciences and faculty of educational science. It can be used effectively when applying machine learning methods which are good tools for predicting end-of-term performance levels among learners.This data set contains 32 different variables which cover many different aspects of students’ lives such as their age group , whether they went through High school or not , scholarship awarded , parents’ education level achieved by them , how often do they study , what is their routine like . In addition to these there are also other variables that indicate behavior problems at school like drug usage and bullying incidents with others indicating exceptional performance- what grade did the student score in their best subject ever taken ? How many credits does he have? Etc...With its wide range of variables reflecting various components influencing academic outcomes attained by learners, this record provides a lot more than just basic insights into relationship between different factors and student achievement. For instance “OUTPUT Grade” variable classifies final grades into seven classes starting from Fail up until AA which gives detailed understanding about learners’ performances. By studying such data sets coupled with artificial intelligence techniques researchers will be able to identify hidden patterns within learner’s records helpful for teachers/principals when designing curriculum models aimed at enhancing success rates among them within schools

**3. PROBLEM STATEMENT**

Design an AI-driven system to evaluate higher education students' performance utilizing machine learning algorithms. The system should analyze academic data to provide insights into student progress, predict future performance, and recommend personalized interventions for improvement.

**4. MOTIVATION AND SCOPE OF WORK**

The motivation behind this project lies in the need for an efficient and objective method to assess higher education students' performance, leveraging the capabilities of artificial intelligence and machine learning. By employing advanced algorithms, we aim to provide valuable insights into student progress, identify areas of improvement, and personalize educational interventions. The scope of work involves developing a comprehensive system that can analyze diverse academic data, predict student outcomes, and offer actionable recommendations to enhance learning outcomes and academic success.

**5. LITERATIVE REVIEW**

[1] A Abu Saa, M Al-Emran, K Shaalan focused on the literature review examines the application of qualitative research methods to the study of entrepreneurship.

[2] K Struyven, F Dochy, S Janssens considered The literature survey examines the role of qualitative studies technique in entrepreneurship studies.

[3]LD Parker proposed The literature review examines the impact of corporate governance on firm performance in emerging markets.

[4]HU Rahiman, R Kodikal developed The literature survey investigates the relationship among company social obligation (CSR) and financial performance inside the context of rising markets.

[5]Daniel Sullivan,Richard Lakeman,Debbie Massey,Dima Nasrawi,Marion Tower,Megan Lee proposed about The literature survey investigates the impact of formative evaluation practices on student getting to know results in better training.

[6] A Romhild, A Hollederer considered about he literature¨ survey investigates the effectiveness of mindfulness-based interventions in decreasing signs and symptoms of tension and melancholy amongst college students.

[7] W Karunarathne, A Calma proposed about The literature survey examines the influence of cooperative gaining knowledge of on student motivation and educational achievement in secondary education.

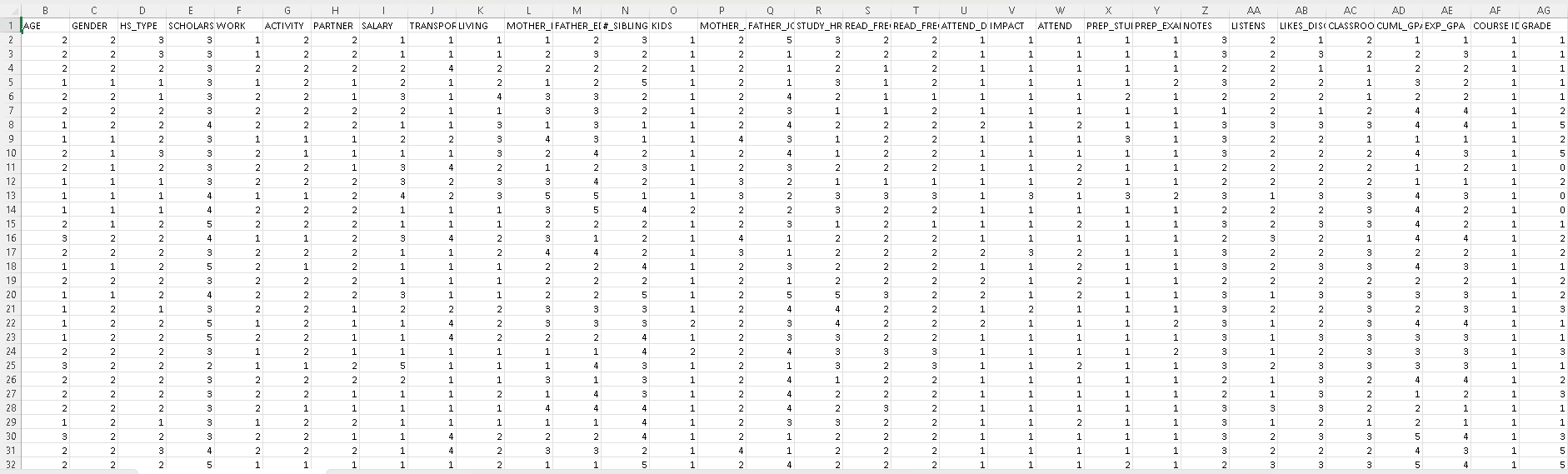
[8]Marie Hutchinson,Rosanne Coutts,Debbie Massey,Dima Nasrawi,Jann Fielden,Megan Lee,Richard Lakeman considered about The literature survey explores the effectiveness of project-primarily based learning in selling vital wondering capabilities among university college students.

[9]X Jin, Q Jiang, W Xiong, Y Feng,W Zhao considered totally about The literature survey examines the effectiveness of peer tutoring programs in enhancing educational performance and mastering effects amongst students. [10]JH Nieminen focused on the literature survey investigates the effectiveness of mindfulness-based interventions in reducing stress and enhancing well-being amongst university students. [11]Mohamad Sudi,Ivon Arisanti,Siti Aisyah Hanim,Ridwan Sya’rani,Kusuma Agdhi Rahwana did reaserch to prove that The literature survey investigates the effectiveness of educational interventions in selling sustainable development recognition among university students.

[12] D Boud, M Bearman proposed about the literature survey investigates the effectiveness of peer tutoring in improving instructional achievement and scholar engagement in secondary training.

[13] Joseph Crawford,Kelly-Ann Allen,Taren Sanders,Roy Baumeister,Philip Parker,Cassandra Saunders focused on The literature survey examines the effect of flipped gaining knowledge of on pupil engagement and educational performance in better schooling.

**6. DATASET**

****

**7. PROPOSED METHODOLOGY**

A.Overview

Machine learning methods are being used to estimate the success rate of higher education students by collecting data from educational institutions and engineering colleges Logistic regression models,lightgbm,XGB boost,Naive Bayes,Multi layer perceptron,decision tree and Support Vector Machines (SVM) are utilized to predict student performance, with logistic regression estimating the probability of outcomes and SVM being a classification technique Predictive models in education focus on evaluating high school student performance using various materials and machine learning techniques, providing valuable insights for educators and policymakers

**7.1 DATA PRE-PROCESSING**

In the dataset the target or the analysing variable is Grade in that there are 0-7 values repeatedly so i came to known that the dataset is a classification.it is also known as multi classification dataset.With original dataset the accuracy is not proper so i changes to a classification dataset by changing 0-3 are 0 and 3-7 are 1.In dataset there is one string with the column name student Id i dropped it.later i filled the null values using fillna().

C. Implementation

logistic regression:

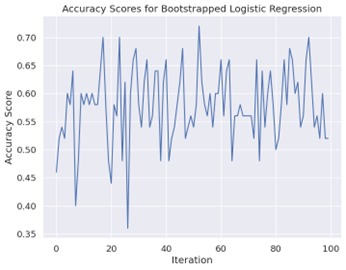
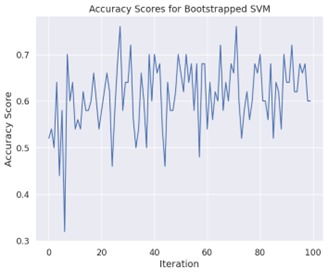
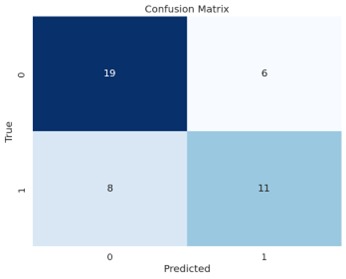
Logistic regression is a mathematical version for binary type duties, wherein the intention is to expect certainly one of two feasible consequences (e.G., 0 or 1, yes or no, real or false) primarily based on one or greater enter traits on. Unlike linear regression, which predicts a continuous outcome, logistic regression predicts the possibility of two consequences. A logistic characteristic (also referred to as a sigmoid characteristic) is used to version this probability. Here is an evidence of logistic regression with its formula. Measurement: A logistic regression model uses the logistic characteristic to estimate the chance (P) of a couple of events (e.G., the probability of a tremendous final results) the usage of the logistic function:

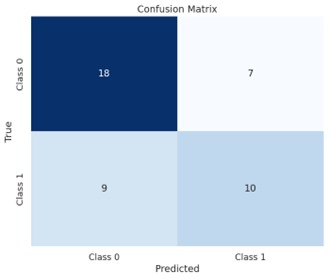
z = w0w1x1w2x2−−−−−−−−−−−−−−−−wnxn

(1) yp = 1/1e−z (2)

z = −Y log(Y P)−(1−Y )INTERPRETATION(1−Y P)

(3)



**7.2 LONG SHORT-TERM MEMORY:**

The proposed model for Higher Education Students Performance Evaluation utilizes clustering techniques followed by classification and regression methods to analyze academic data. Initially, the input data, comprising student performance metrics, is subjected to clustering using diverse algorithms including K-Means, Hierarchical, DBSCAN, Agglomerative, K-Medoids, and Gaussian Mixture Models (GMM). These clustering methods partition the data into meaningful clusters based on similarity. Subsequently, each cluster is analyzed separately using classification techniques such as Logistic Regression, Support Vector Machines, Decision Trees, Naive Bayes, multi-layer perceptrons, convolutional neural networks, XGBoost, and LightGBM, to predict students' academic outcomes. Additionally, regression models including Decision Tree Regression, Support Vector Regression (SVR), Random Forest Regression, K-Nearest Neighbor’s (KNN) Regression, ElasticNet Regression, and Neural Networks (Multi-layer Perceptron) are employed to forecast future performance trends and provide insights into academic progress. This comprehensive approach leverages the power of artificial intelligence and machine learning to enhance student performance evaluation in higher education settings.

**7.3 COMPARED ALGORITHMS**

.•Support Vector Machine:

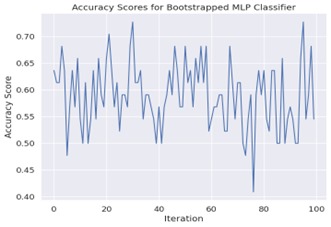
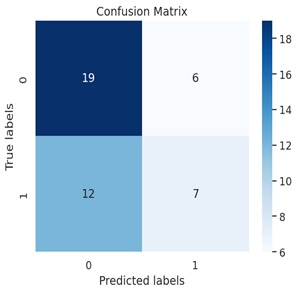
-Support Vector Machine or SVM is one of the most popular Supervised Learning algorithms, it’s far used for Classification in addition to for Regression issues. However, it’s miles specially used for classification issues in machine learning.

-The intention of the SVM set of rules is to create an most useful line or decision boundary that can separate the n-dimensional area into training so that within the destiny we can easily assign a brand new facts point to the suitable range. This choicest choice constraint is called a hyperplane.

Multi-layer Perceptron:

-A Multi-Layer Perceptron (MLP) is a fundamental form of artificial neural network such as more than one layers of interconnected nodes or neurons. It generally accommodates an enter layer, one or greater hidden layers, and an output layer. Each neuron within the community gets enter alerts, strategies them thru a non-linear activation function, and transmits an output to neurons within the next layer.

-MLPs are able to learning complex styles and relationships within statistics, making them widely used in various fields consisting of pattern recognition, class, and regression tasks. Through a manner called backpropagation, MLPs regulate the weights of connections among neurons all through schooling, aiming to decrease the difference between expected and actual outputs. -Despite their effectiveness, MLPs have barriers, together with vulnerability to overfitting and problem in handling excessive-dimensional data. Nevertheless, they remain a cornerstone within the subject of deep studying, forming the idea for more state-of-the-art neural network architectures.

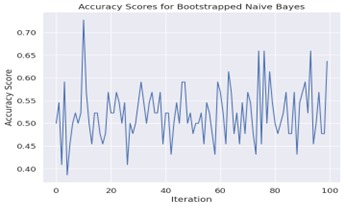
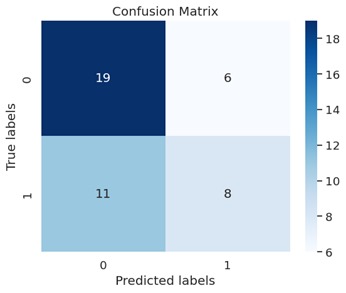
 

• Naive Bayes:

-Naive Bayes is a famous set of rules in machine gaining knowledge of for type duties. It’s based totally on Bayes’ theorem, which calculates the possibility of a hypothesis given the proof. Despite its simplicity, Naive Bayes often performs remarkably nicely, mainly in text type and spam filtering.

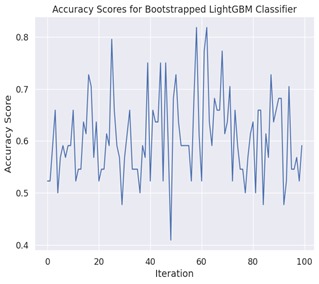
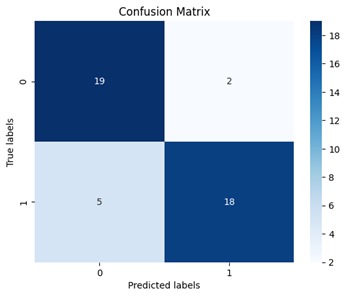
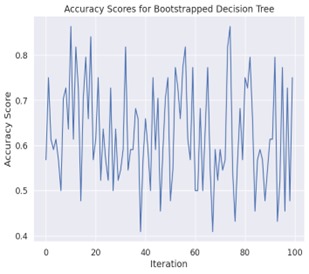
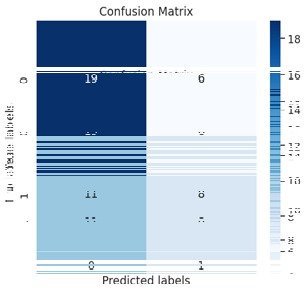
-The ”naive” issue of Naive Bayes comes from its assumption of independence among functions. It assumes that each function contributes independently to the chance of a selected final results, which may not hold actual in real-international scenarios. However, this simplifying assumption allows for immediate schooling and prediction, making Naive Bayes specially beneficial for massive datasets.

-Naive Bayes works with the aid of calculating the possibility of each elegance given the enter features and choosing the class with the very best possibility as the prediction. Despite its ”naive” assumptions, Naive Bayes may be extraordinarily effective and is broadly utilized in numerous programs because of its simplicity, efficiency, and frequently aggressive overall performance

• Lightgbm:

LightGBM, brief for Light Gradient Boosting Machine, is a high-performance gradient boosting framework developed via Microsoft. It’s designed for performance and speed in handling large datasets and is widely used in device studying competitions and enterprise programs. LightGBM makes use of a tree-based mastering set of rules, using a singular technique referred to as Gradientbased totally One-Side Sampling (GOSS) to reduce memory usage and enhance schooling velocity. It additionally utilizes Histogram-primarily based algorithms for quicker computation. Its flexibility, scalability, and capacity to handle specific features make it a famous desire for diverse tasks, along with category, regression, and ranking problems, continuously handing over competitive effects with minimal tuning.

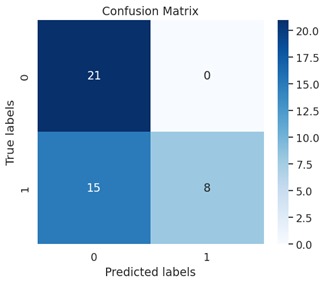
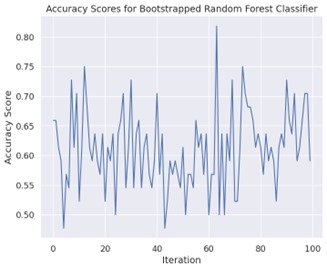
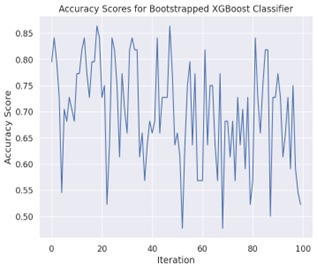
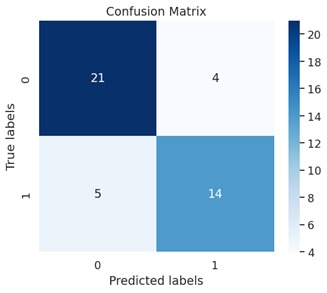
   

• Decision Tree:

A decision tree is a predictive modeling tool utilized in facts mining and machine getting to know. It organizes records right into a tree-like structure of choices and their viable outcomes. Each internal node represents a ”take a look at” on an characteristic, every branch represents the final results of the test, and every leaf node represents a class label or a choice. Decision timber are famous due to their simplicity and interpretability, making them treasured for know-how and explaining the decisionmaking technique. They’re employed in various fields, from finance to healthcare, aiding in class, regression, and solving complicated choice-making troubles easily and transparency.

• XG Boost:

XGBoost, short for Extreme Gradient Boosting, is a powerful machine learning set of rules famend for its efficiency and performance in classification and regression responsibilities. It employs a gradient boosting framework, iteratively combining vulnerable beginners, normally decision trees, to shape a robust predictive model. XGBoost contains regularization techniques to save you overfitting and makes use of parallel computing to beautify velocity. Its versatility, scalability, and effectiveness in handling huge datasets make it a popular desire in numerous domain names, from finance to healthcare and past. With its capability to deliver excessive accuracy and interpretability, XGBoost stands as a cornerstone in modern-day device studying workflows.



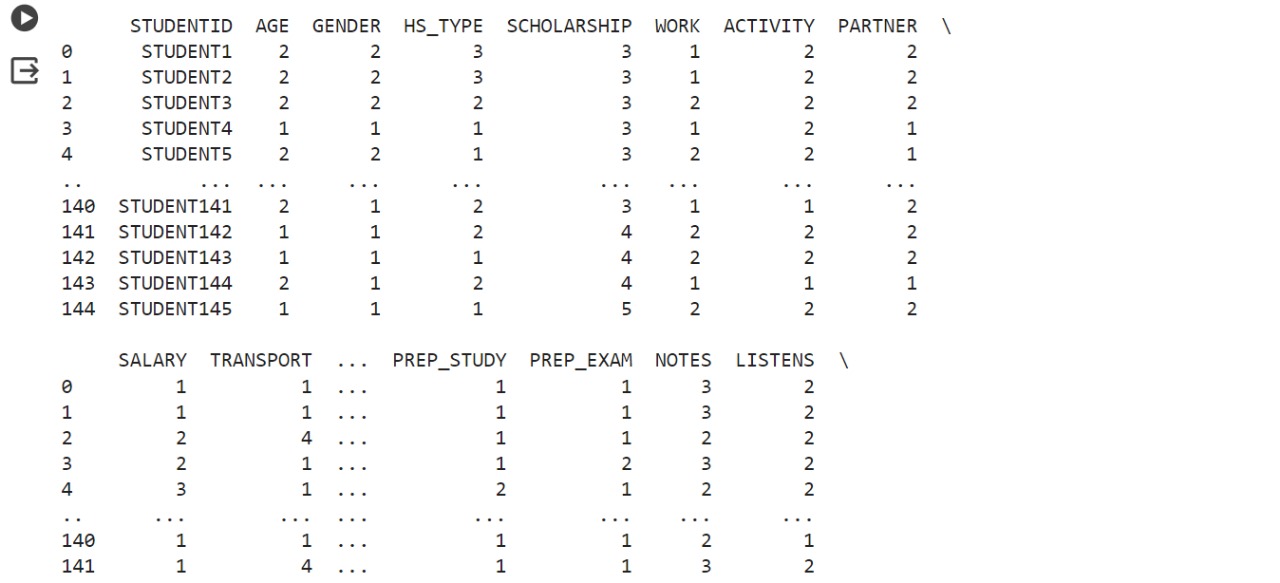
• Random Forest:

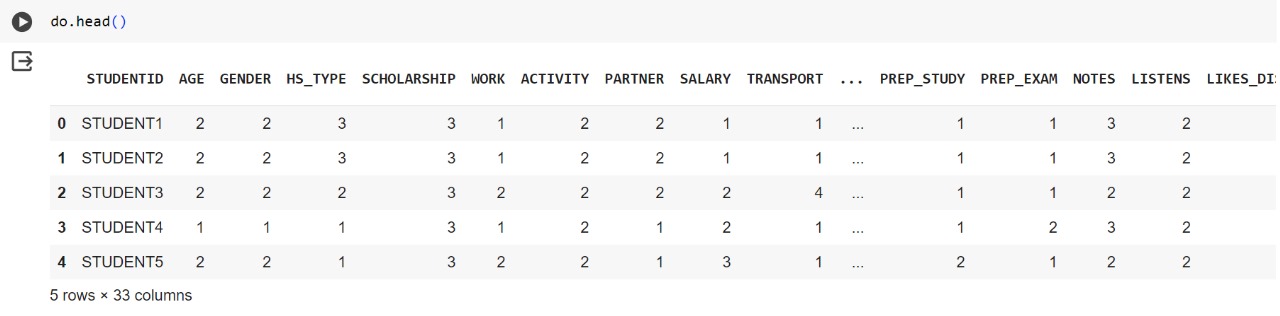
Random Forest is a versatile ensemble studying set of rules widely used for each category and regression duties. It operates by building a mess of decision timber in the course of training and outputs the mode of the training (class) or the imply prediction (regression) of man or woman bushes. Each tree inside the wooded area is trained on a random subset of the schooling information and makes impartial predictions. This approach mitigates overfitting and complements generalization overall performance. Random Forest is famend for its robustness, scalability to big datasets, and capability to address high-dimensional feature areas. It's a popular choice across diverse domain names due to its simplicity, flexibility, and high predictive accuracy.

**8. RESULTS & DISCUSSION**

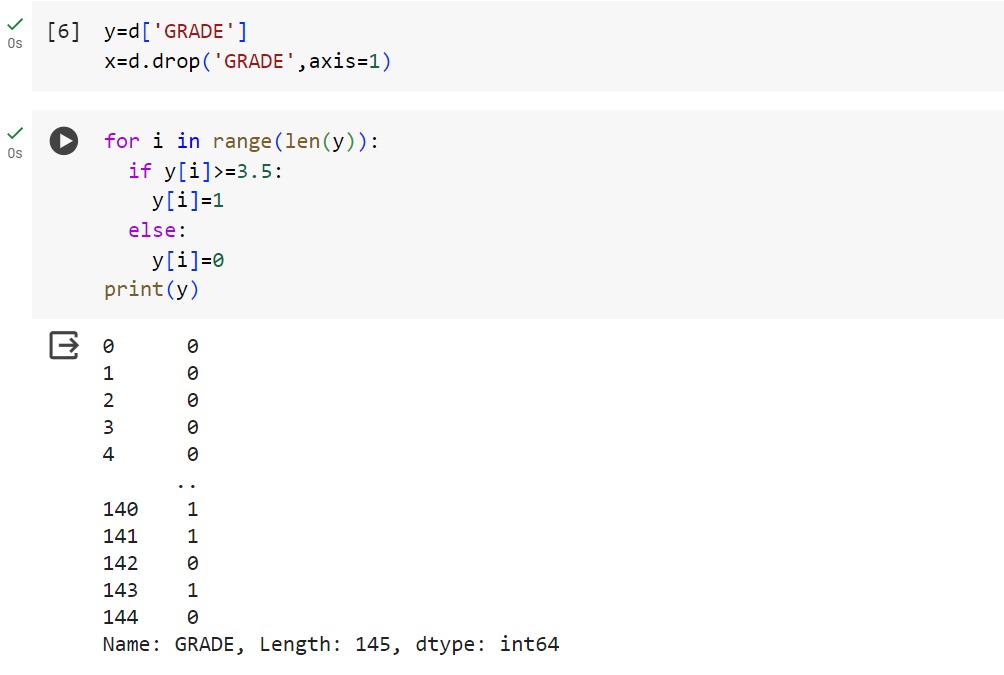
**Code:**

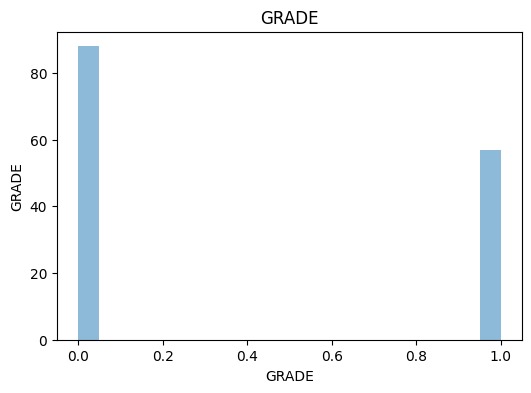
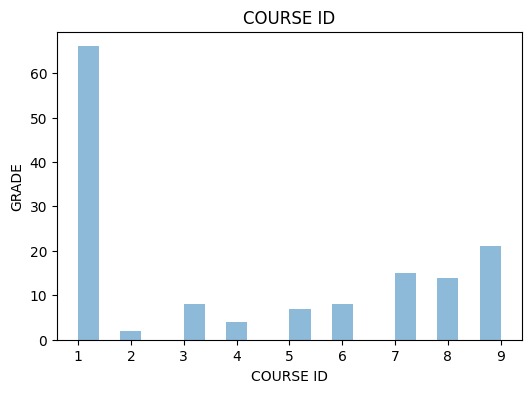
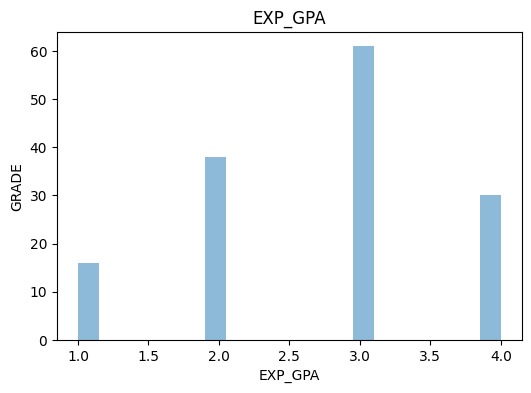
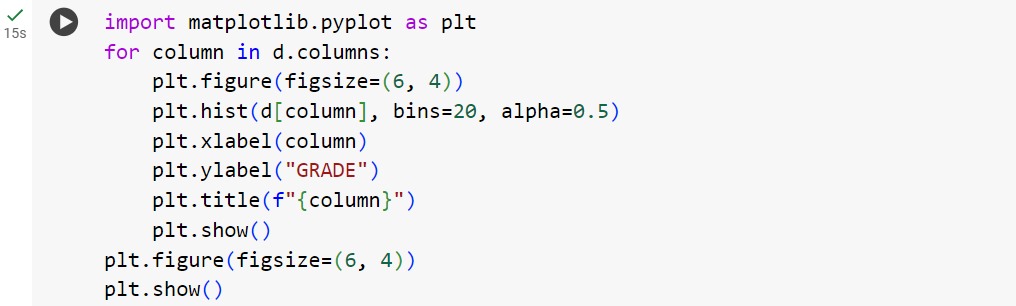


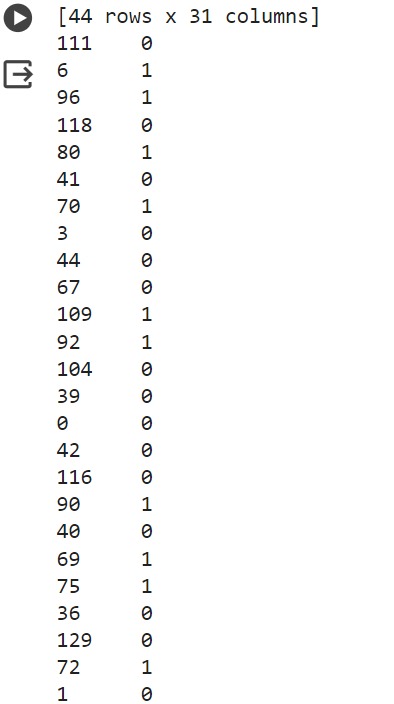
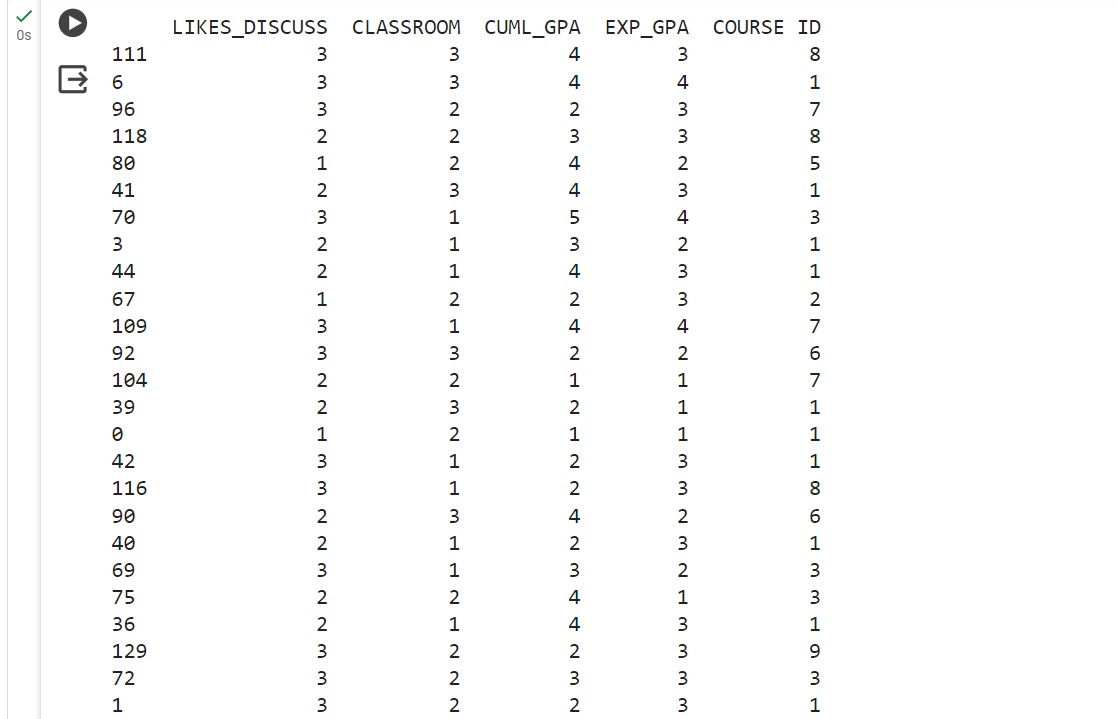
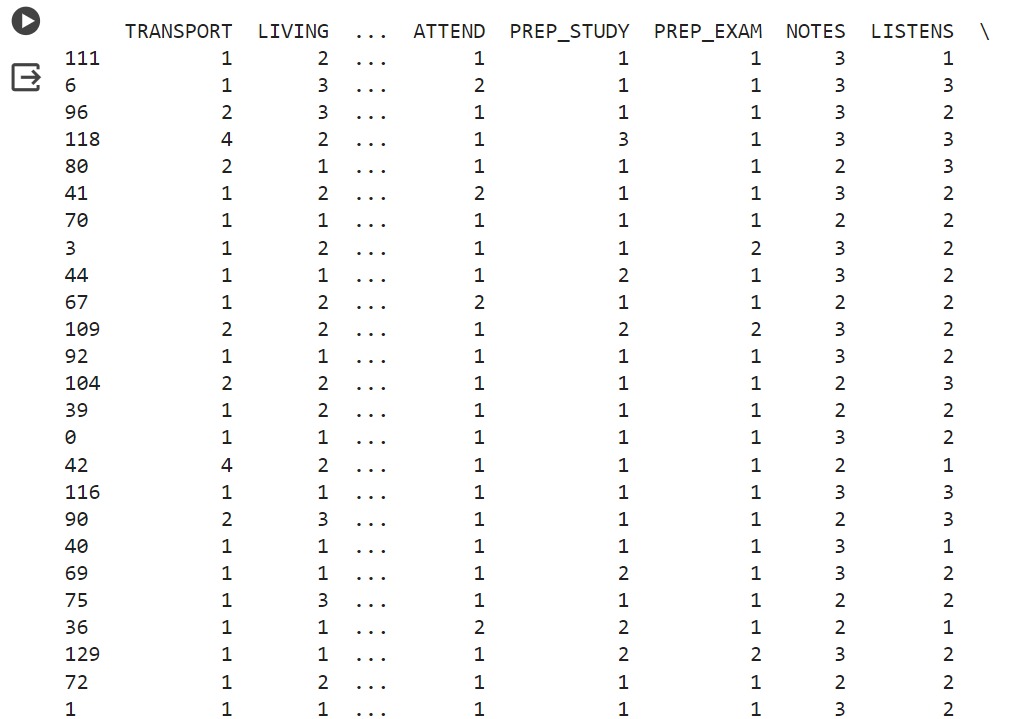
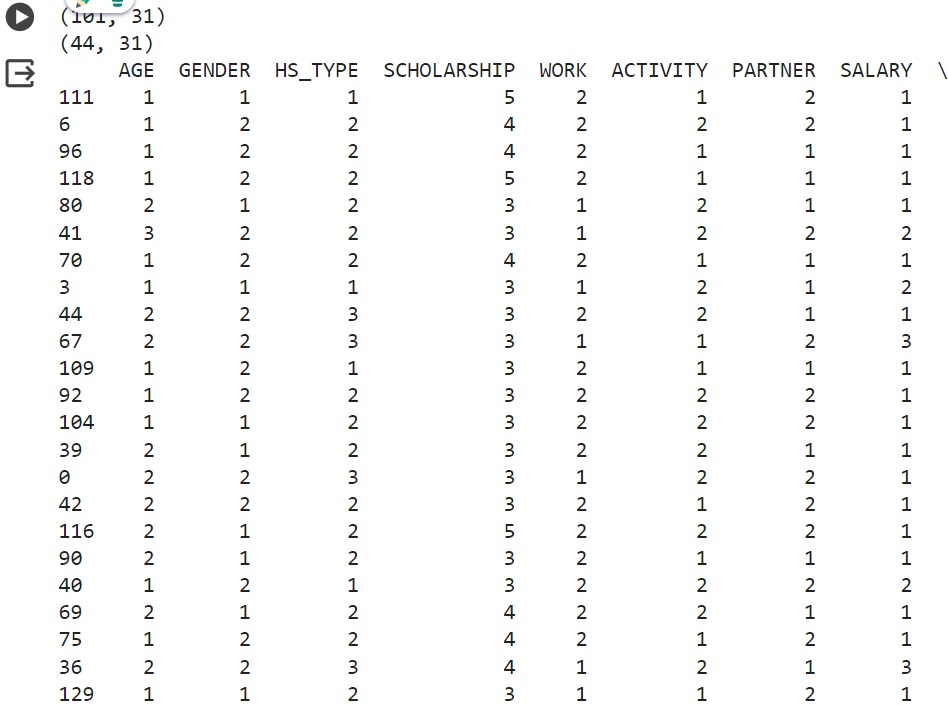
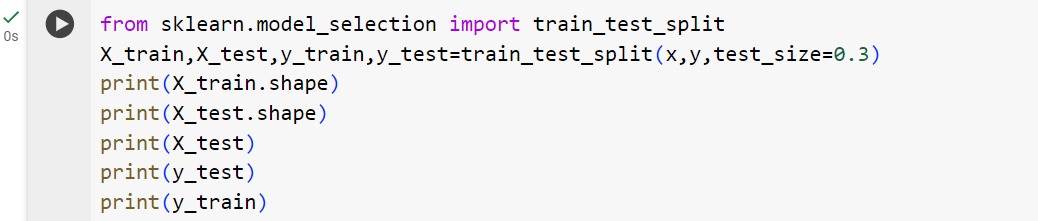


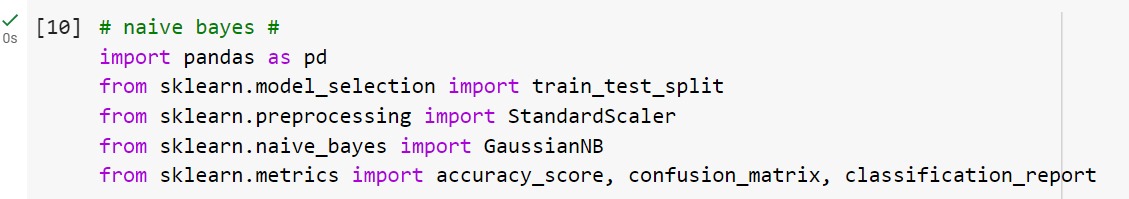


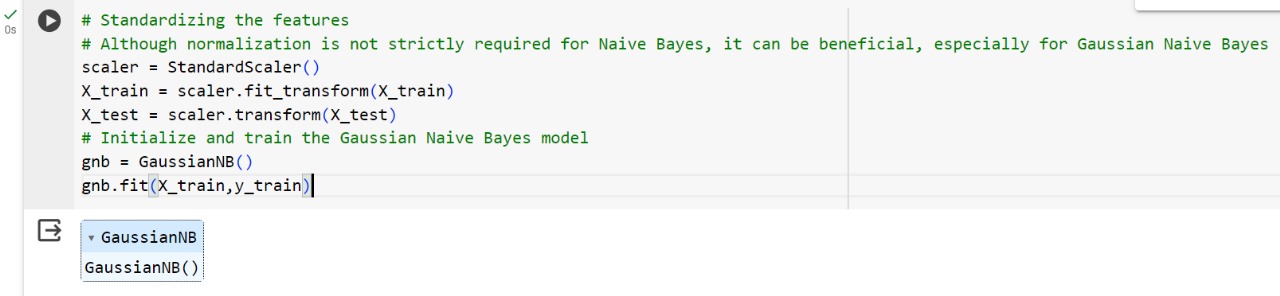


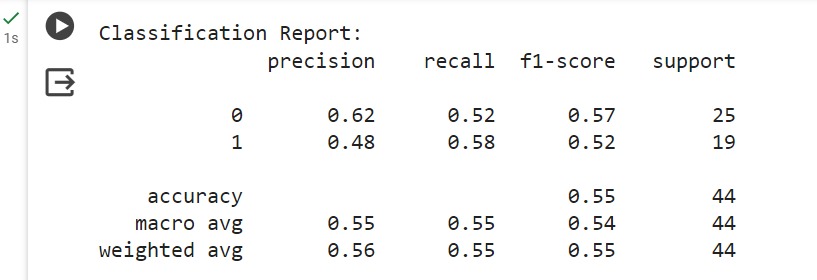
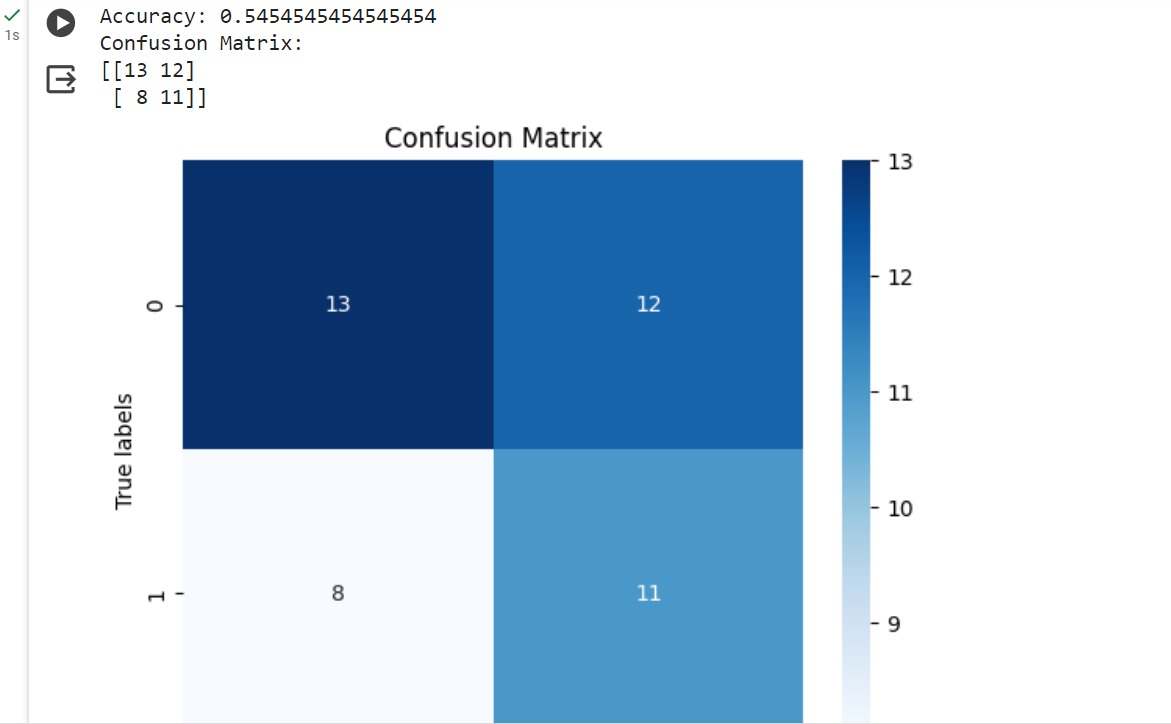
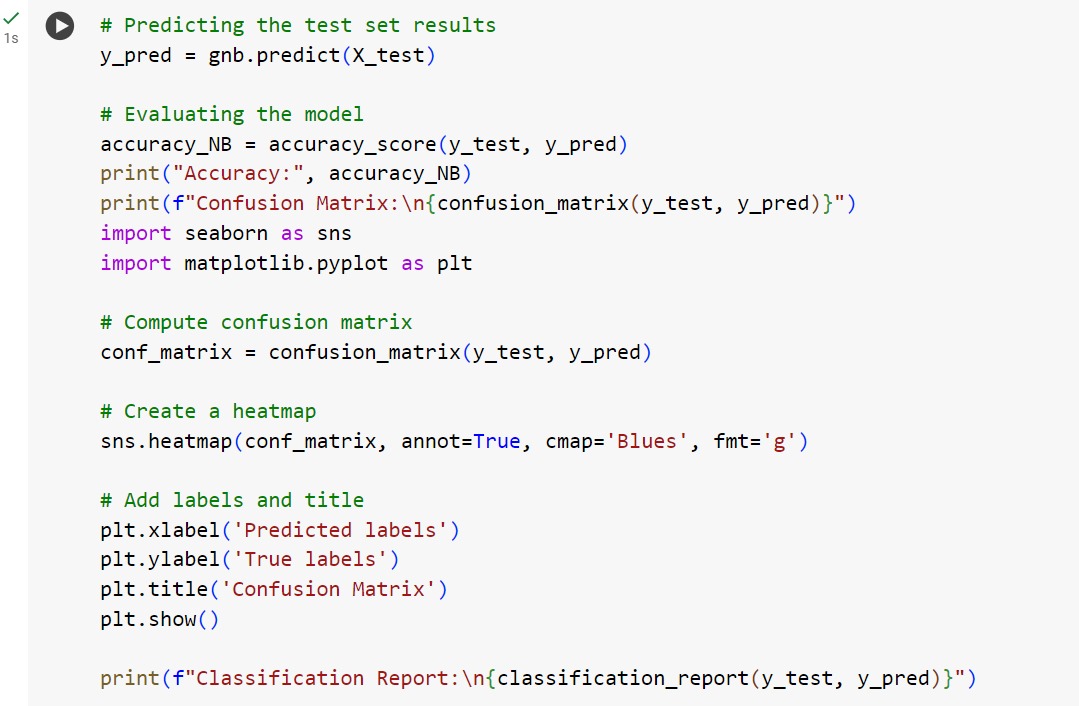


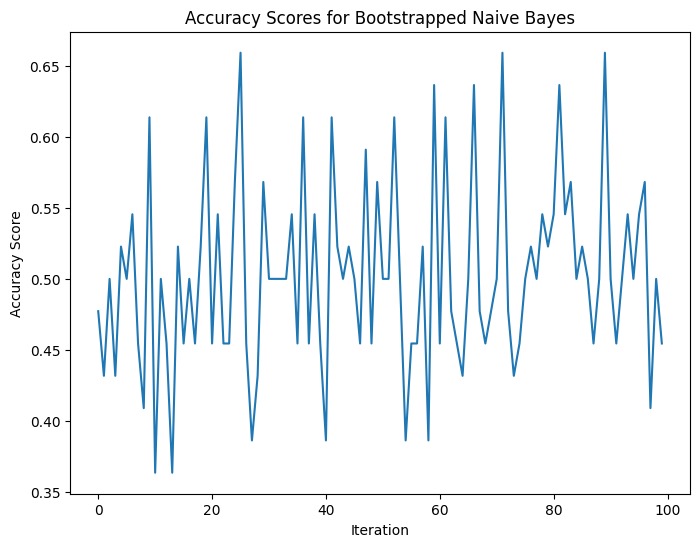
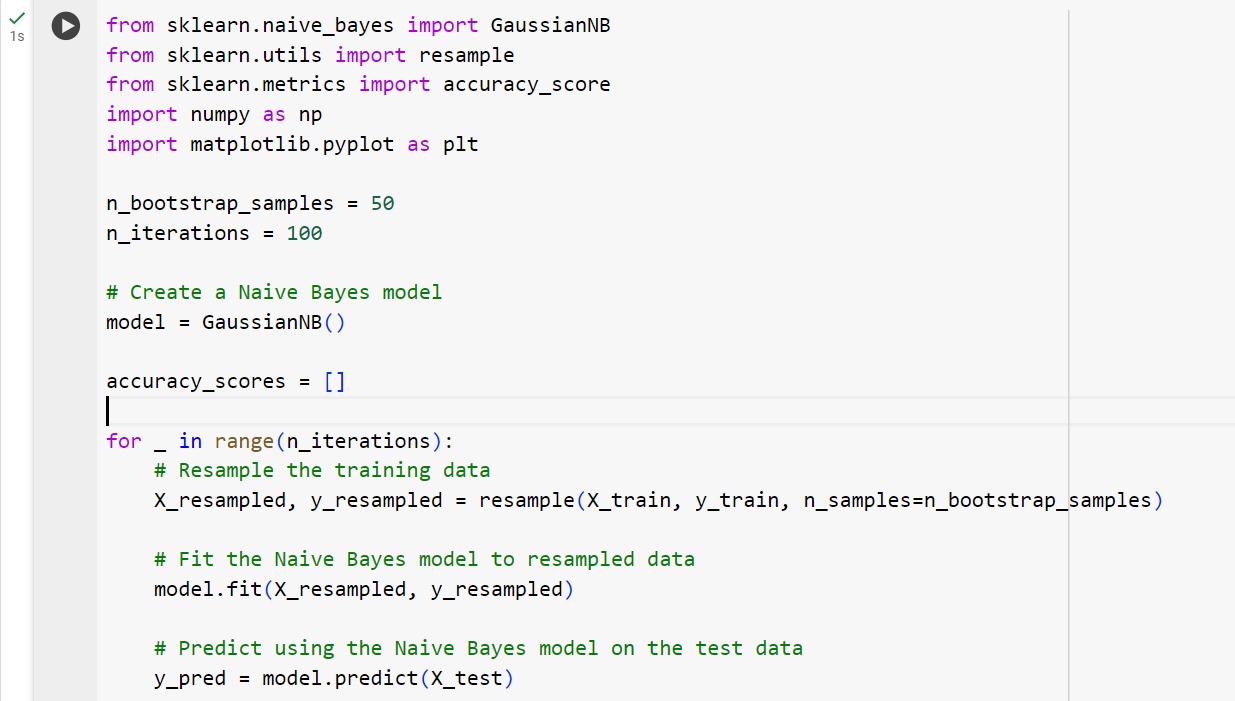


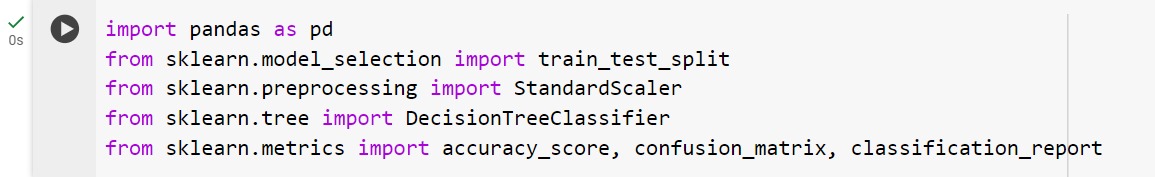


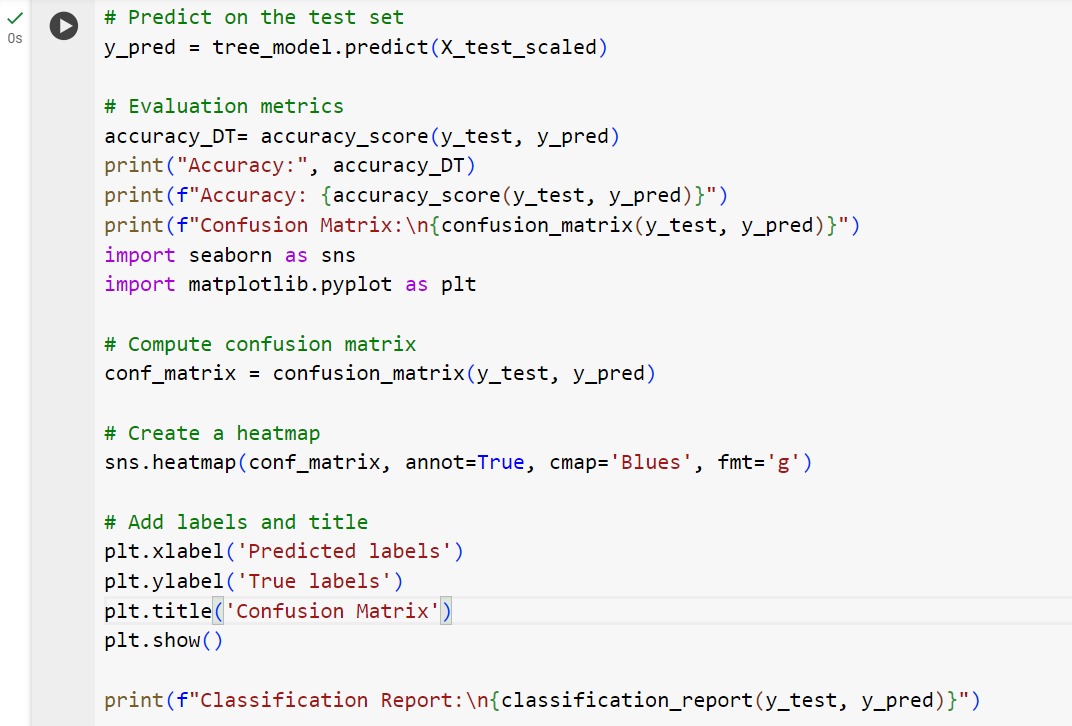


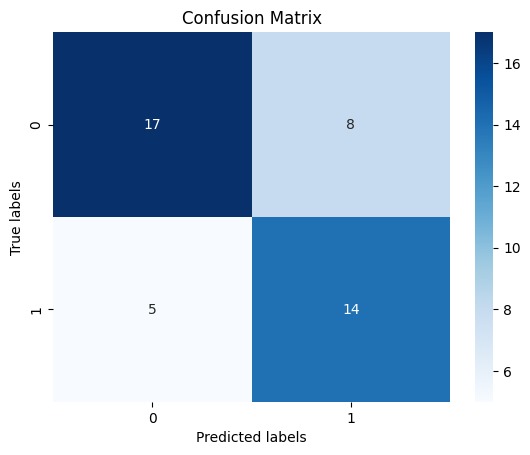


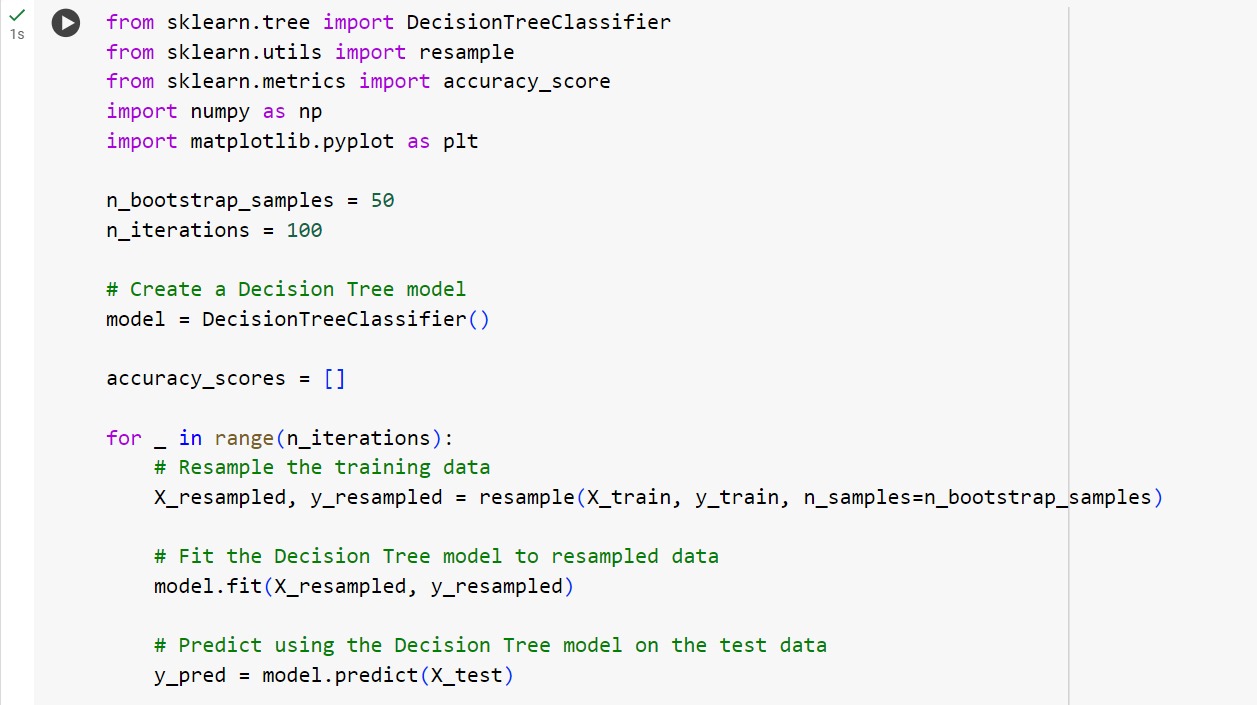


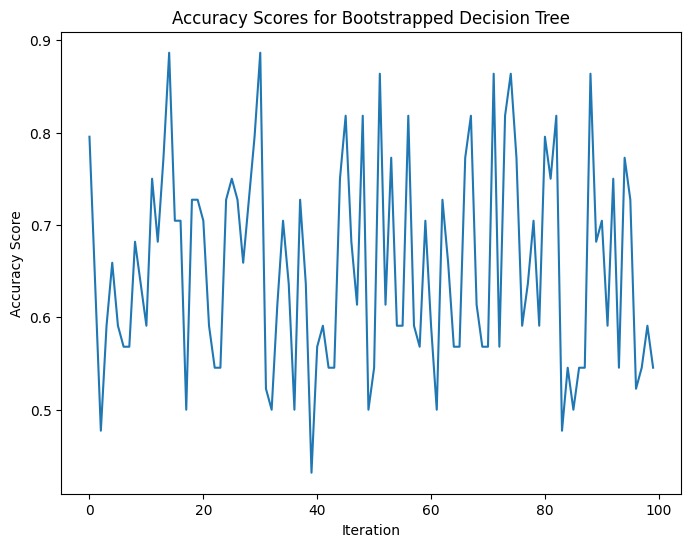
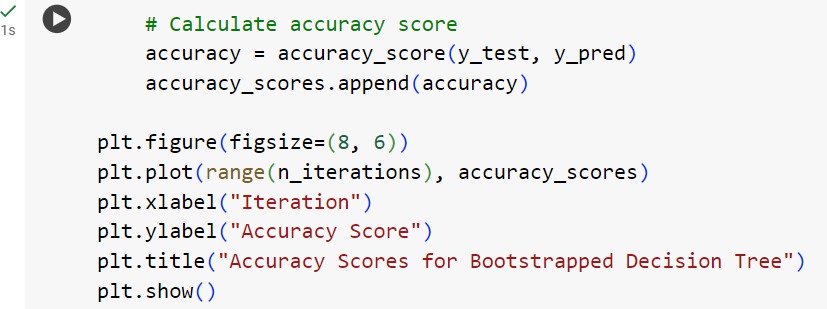


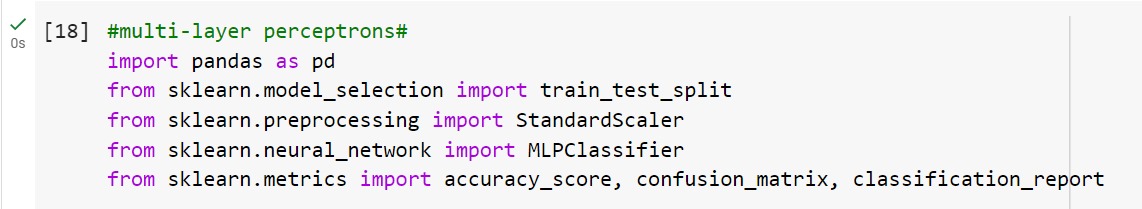


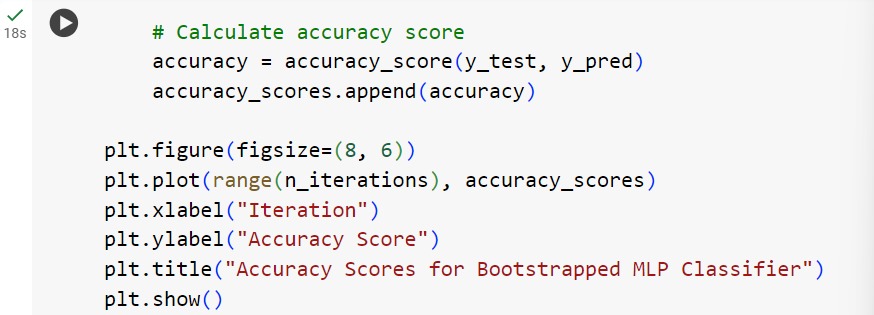
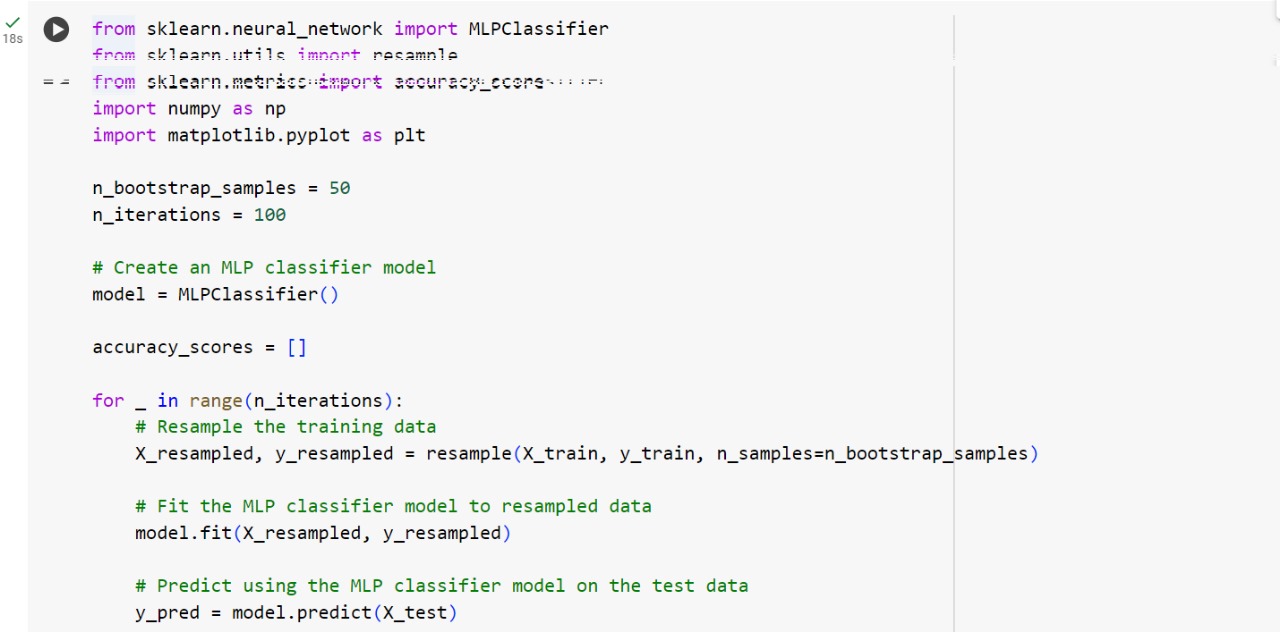


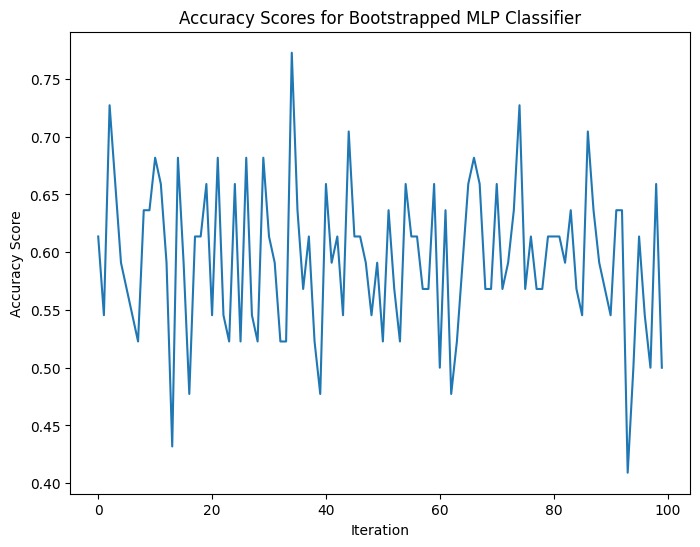




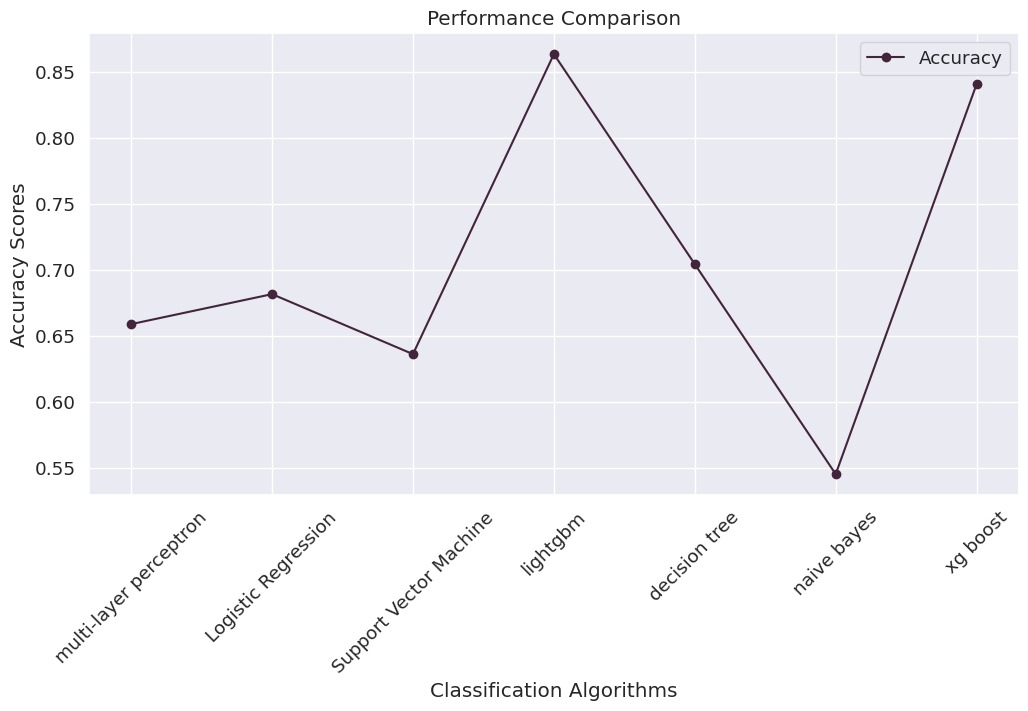
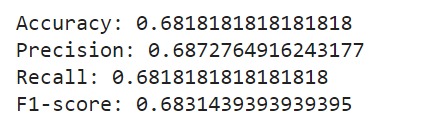
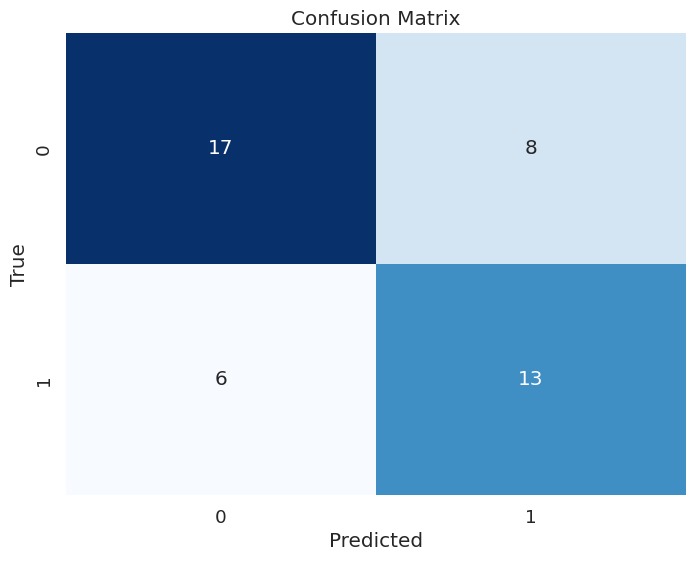
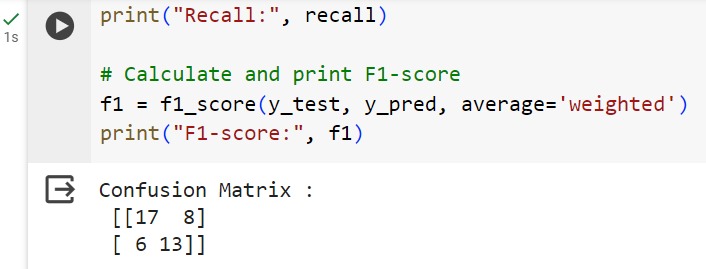
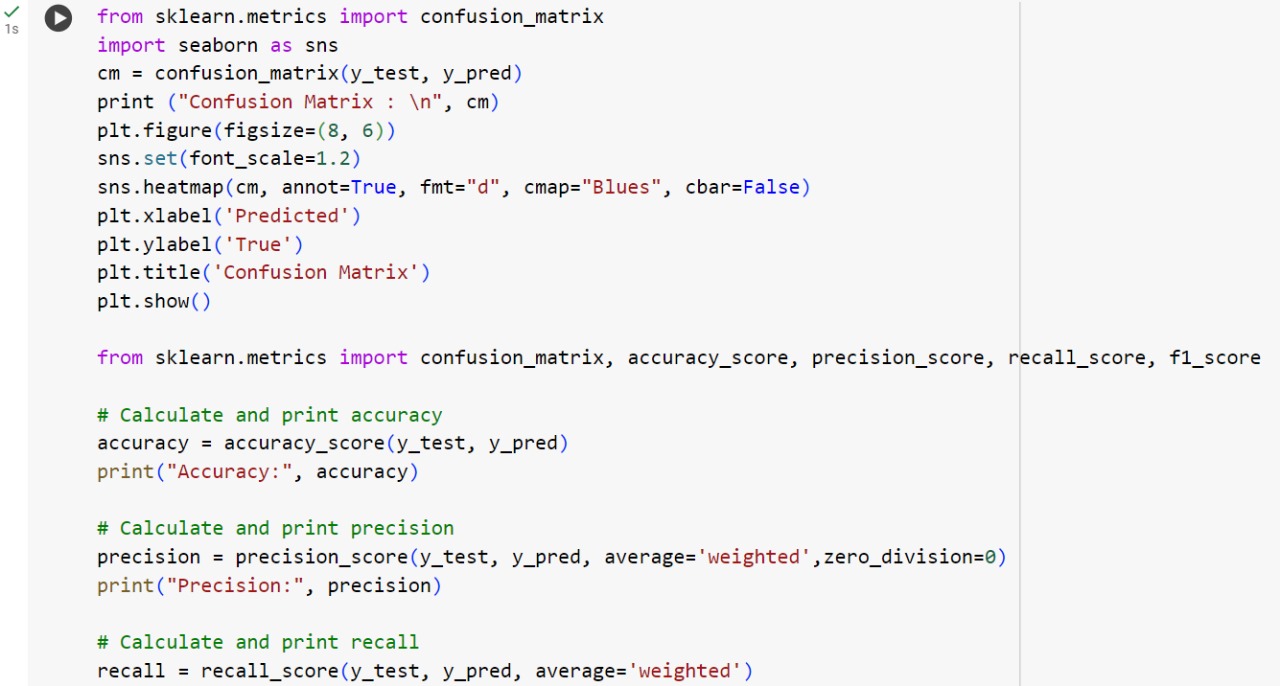
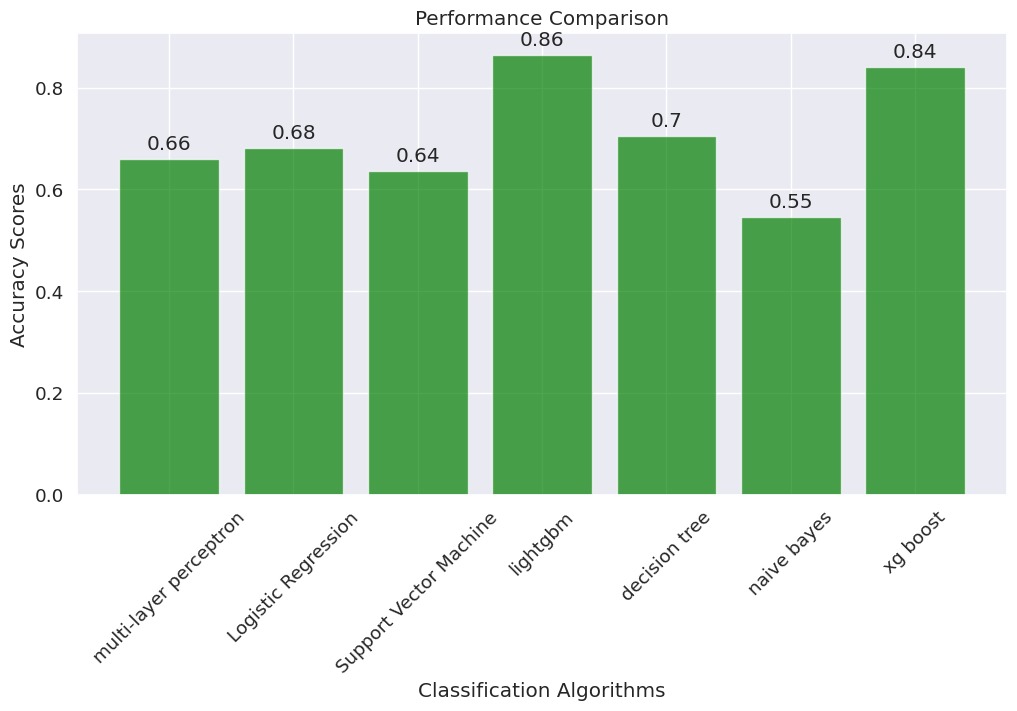
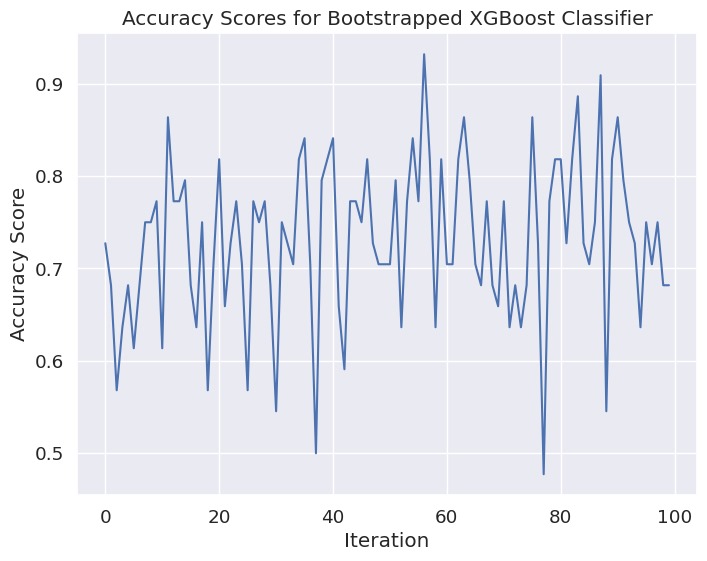
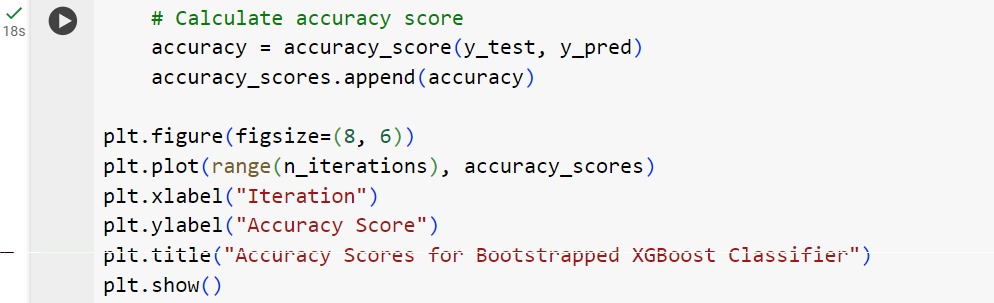
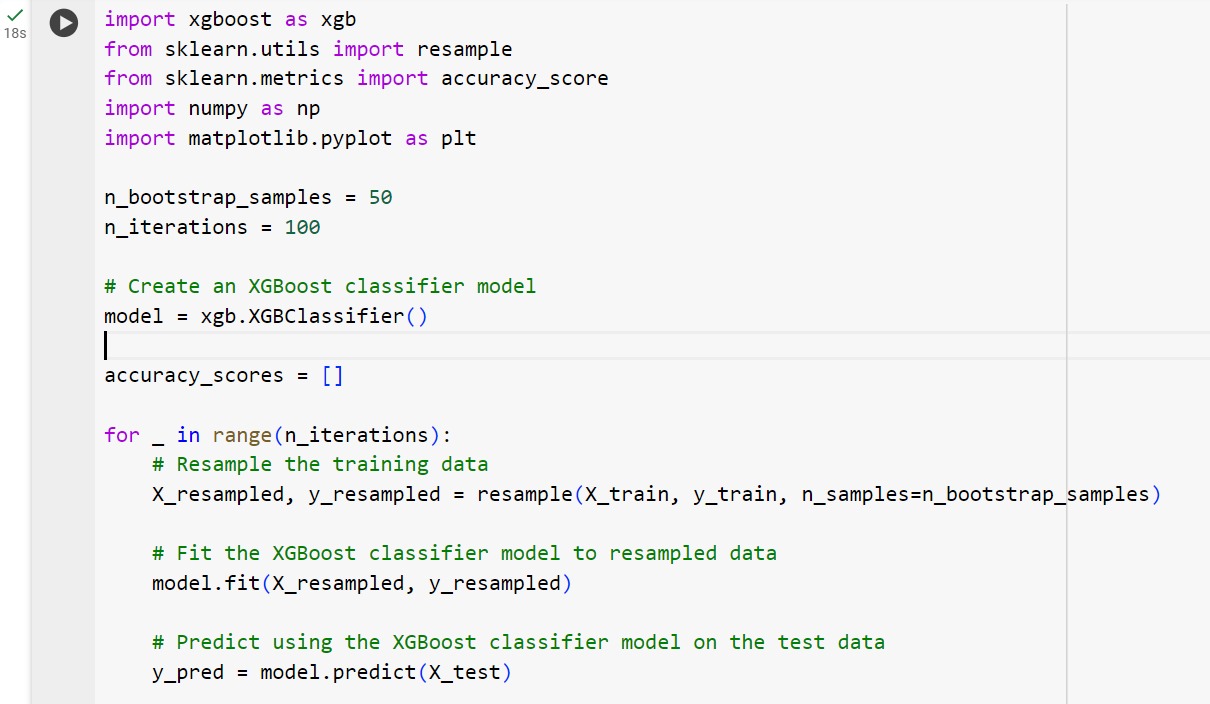
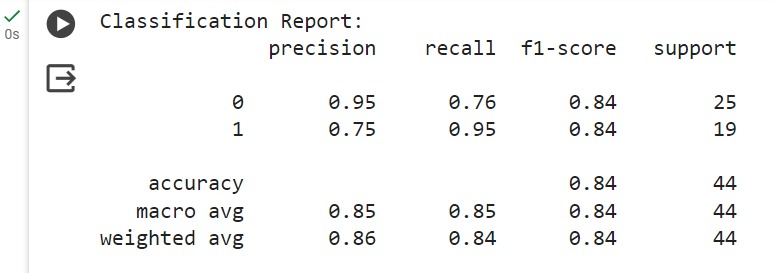
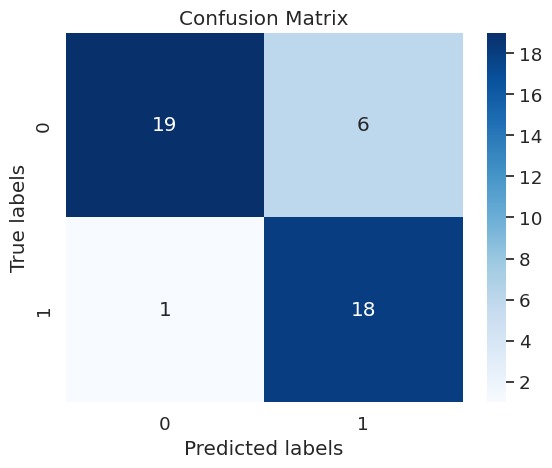
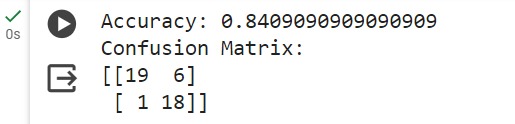
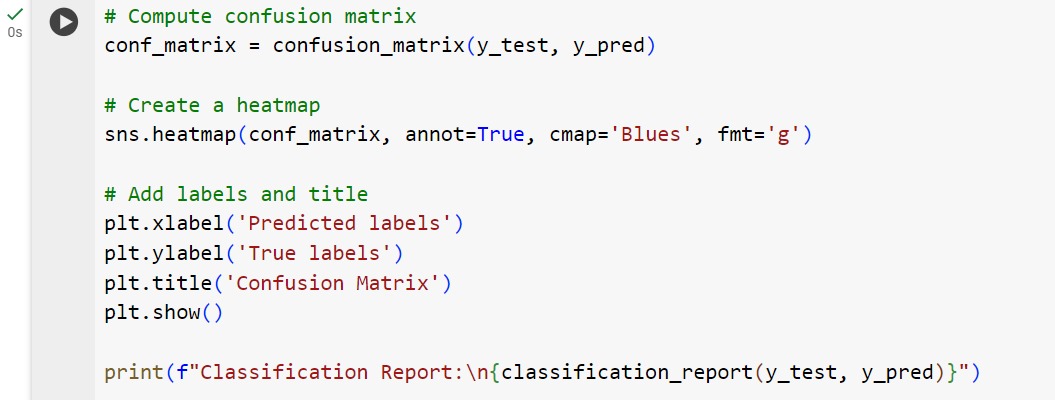
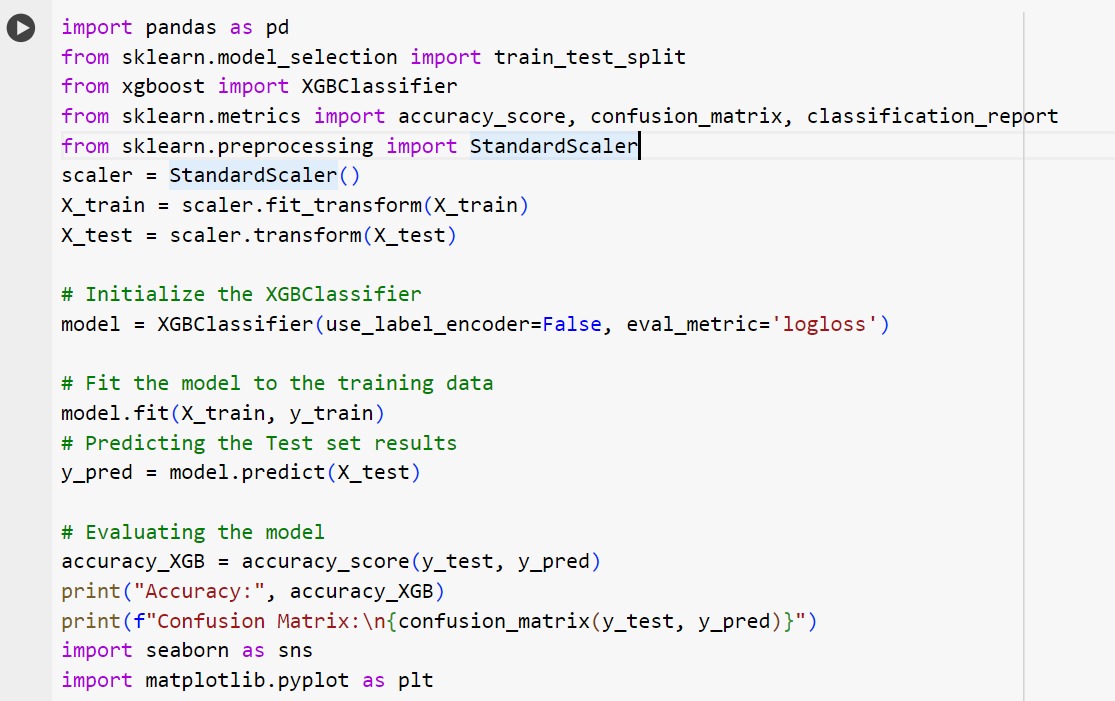
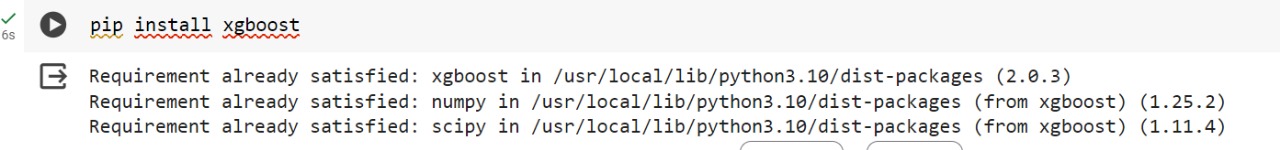
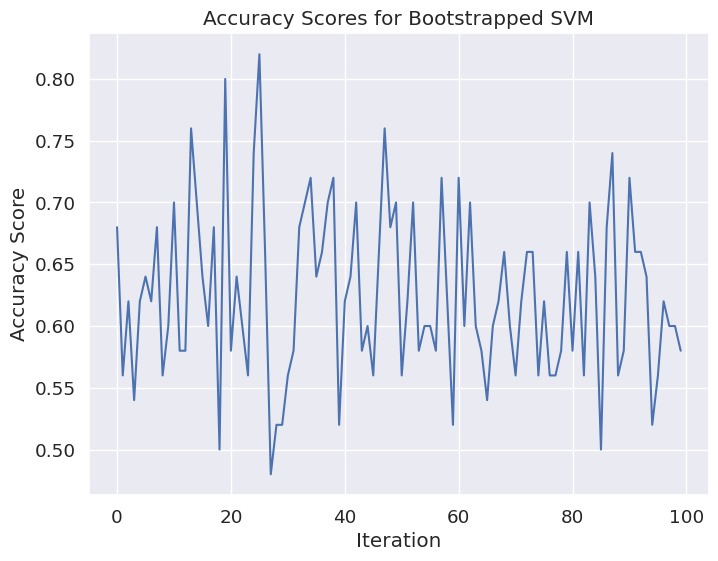
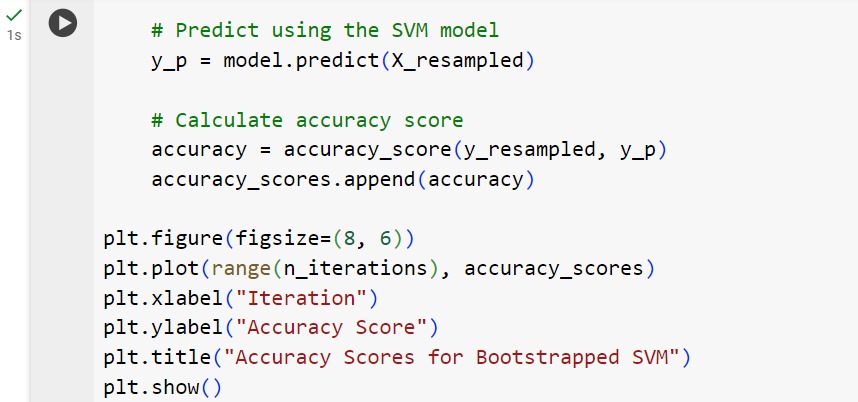
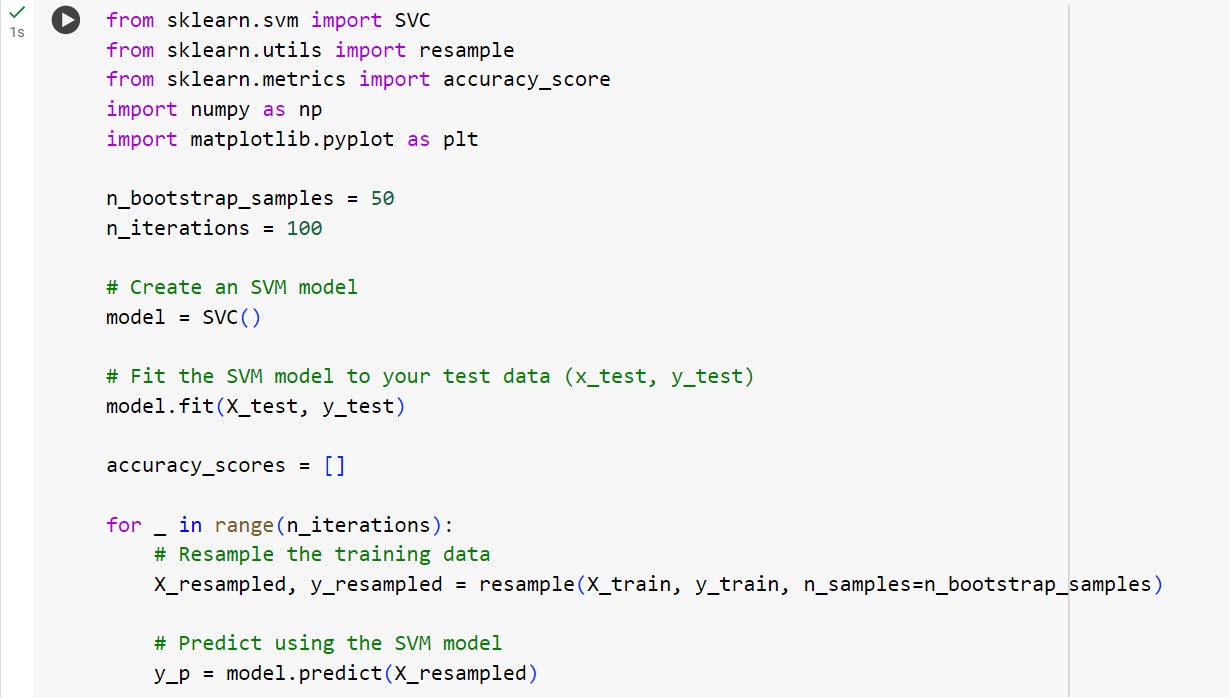
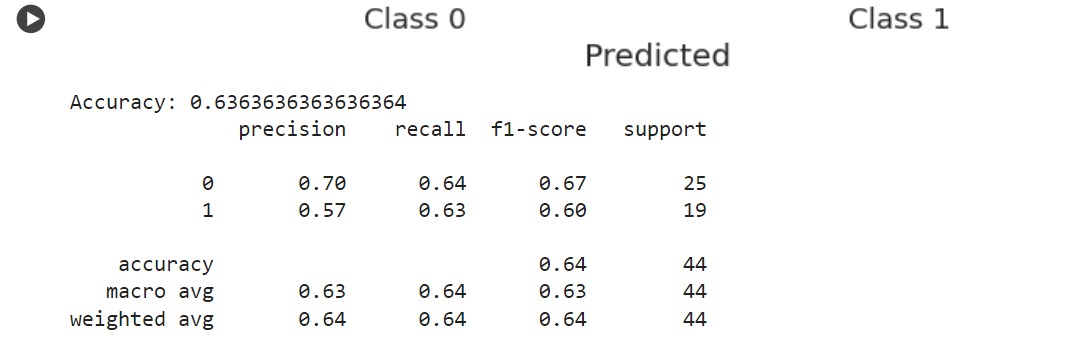
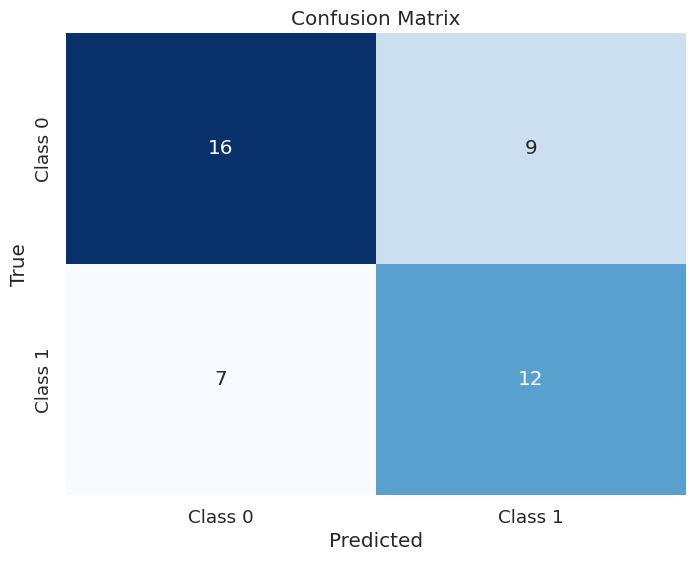
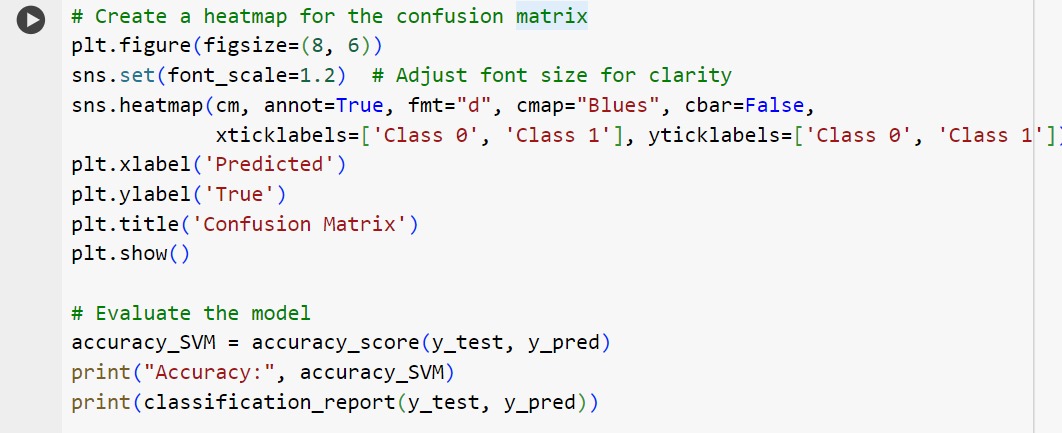
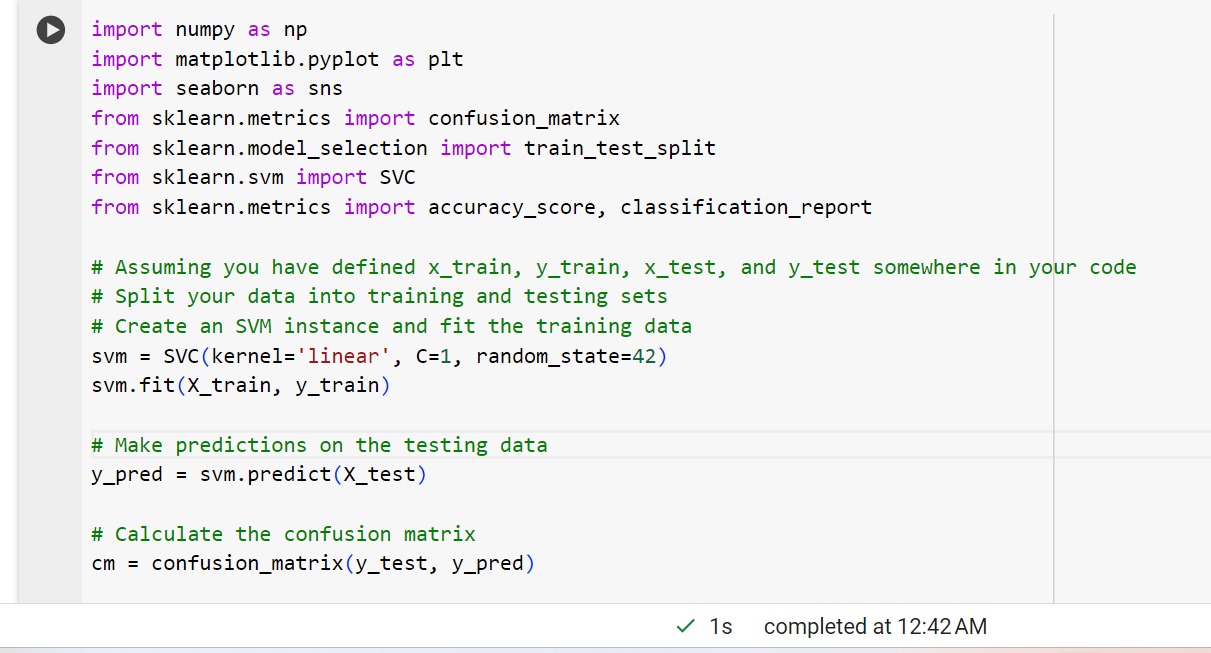
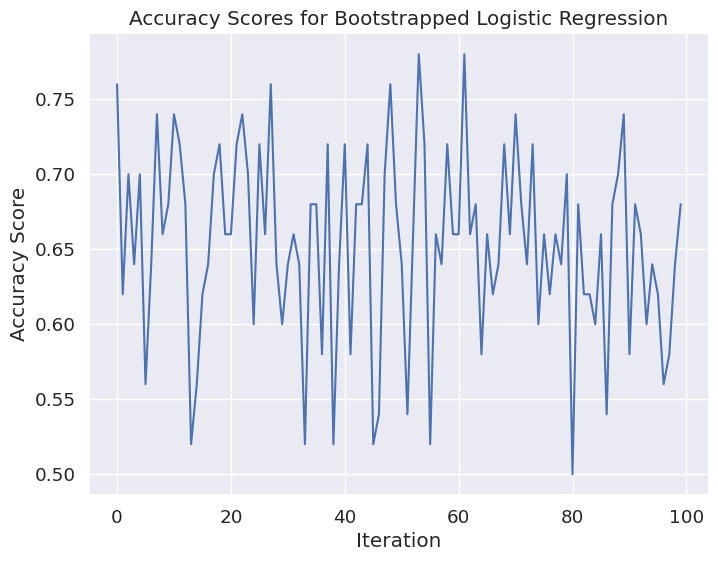
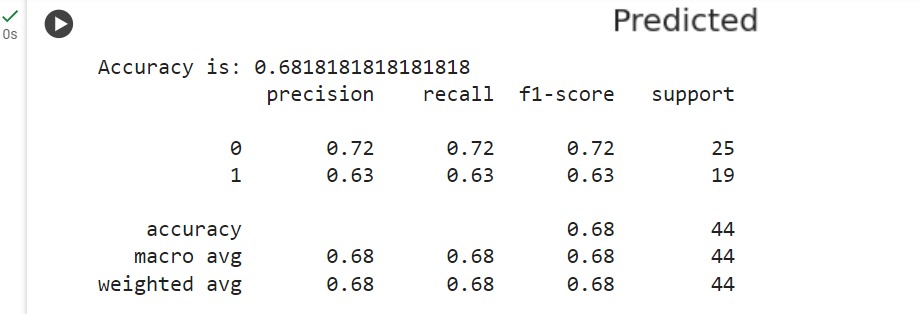
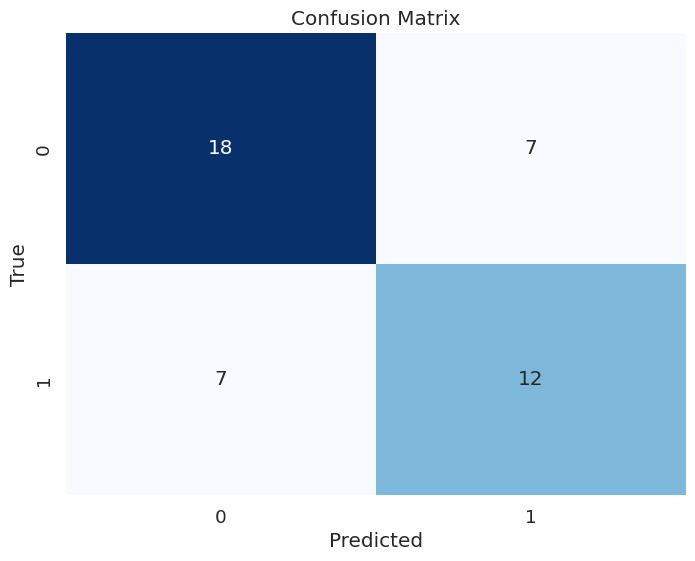
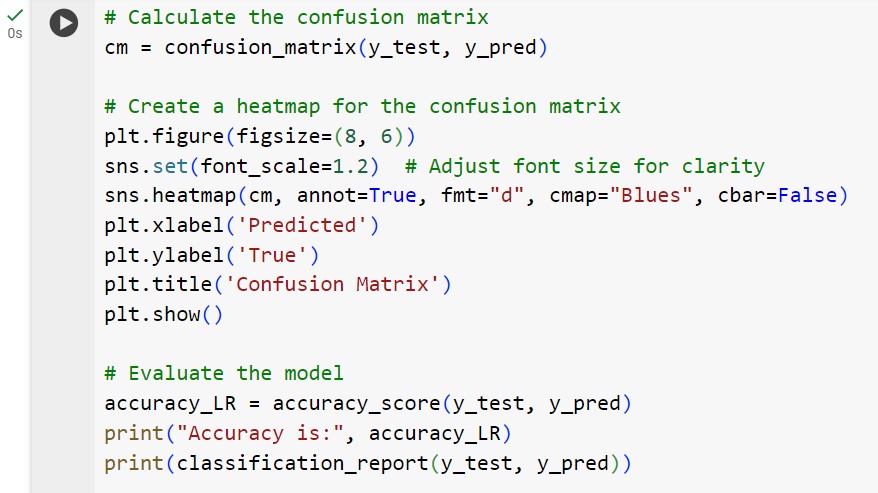
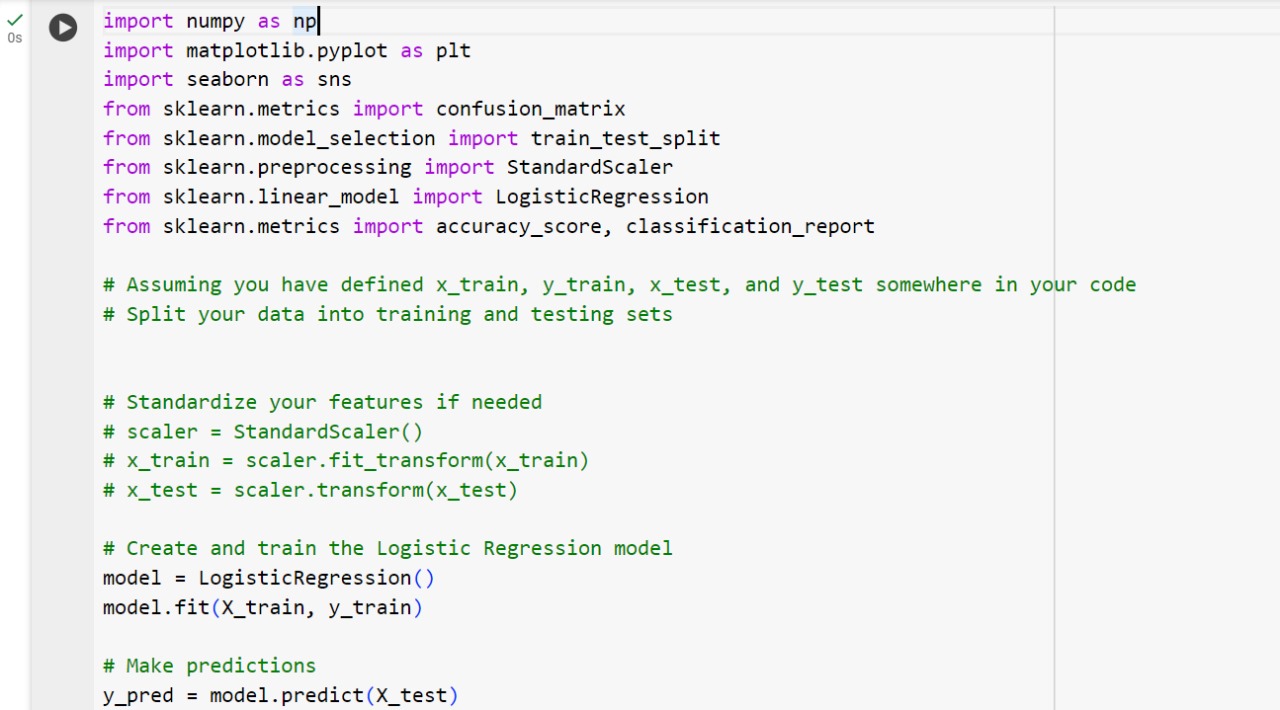
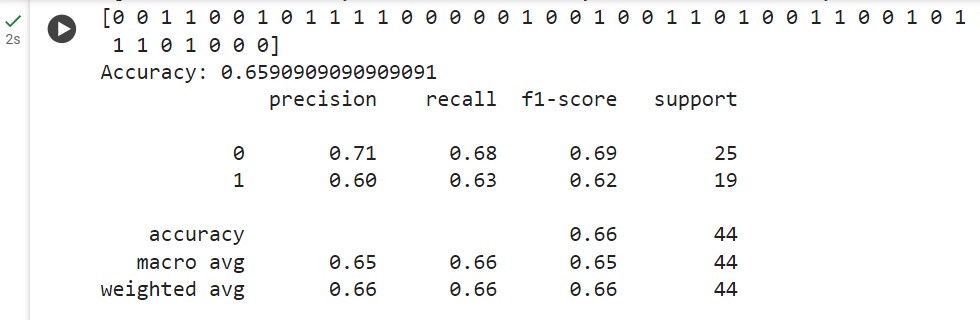
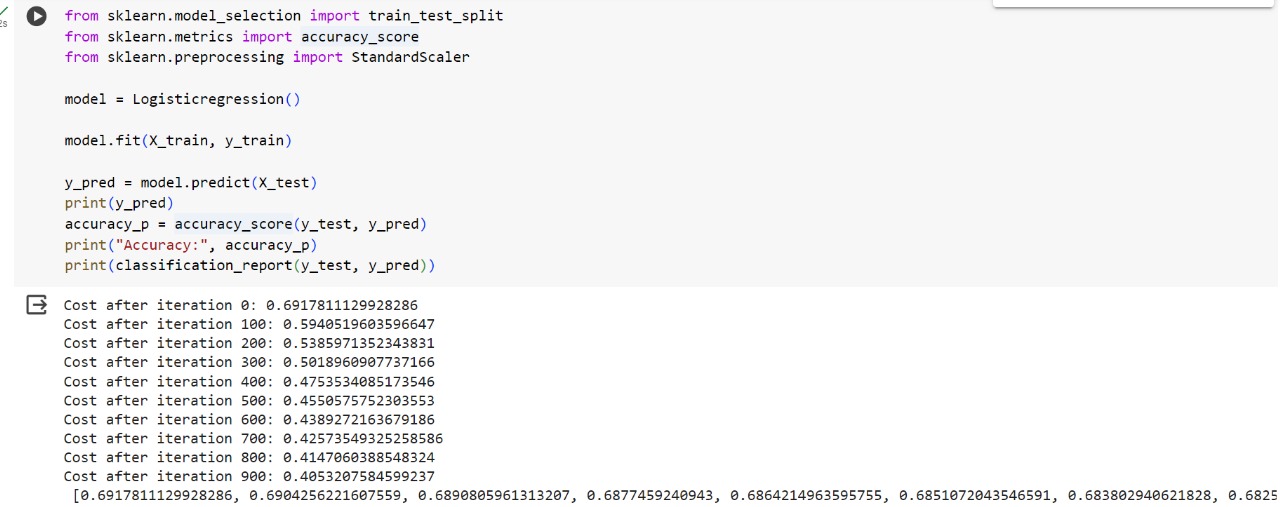
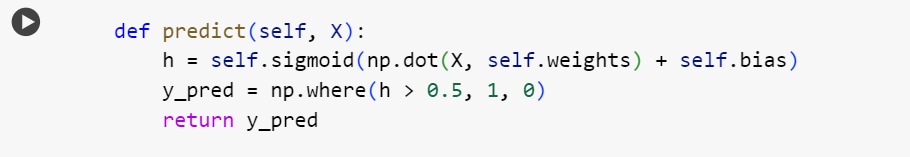
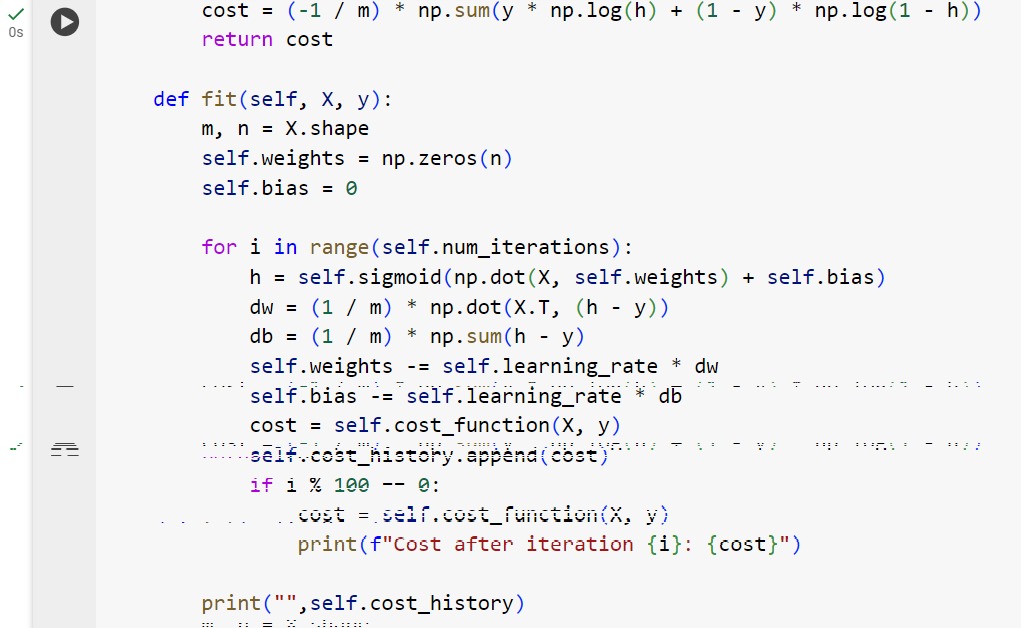
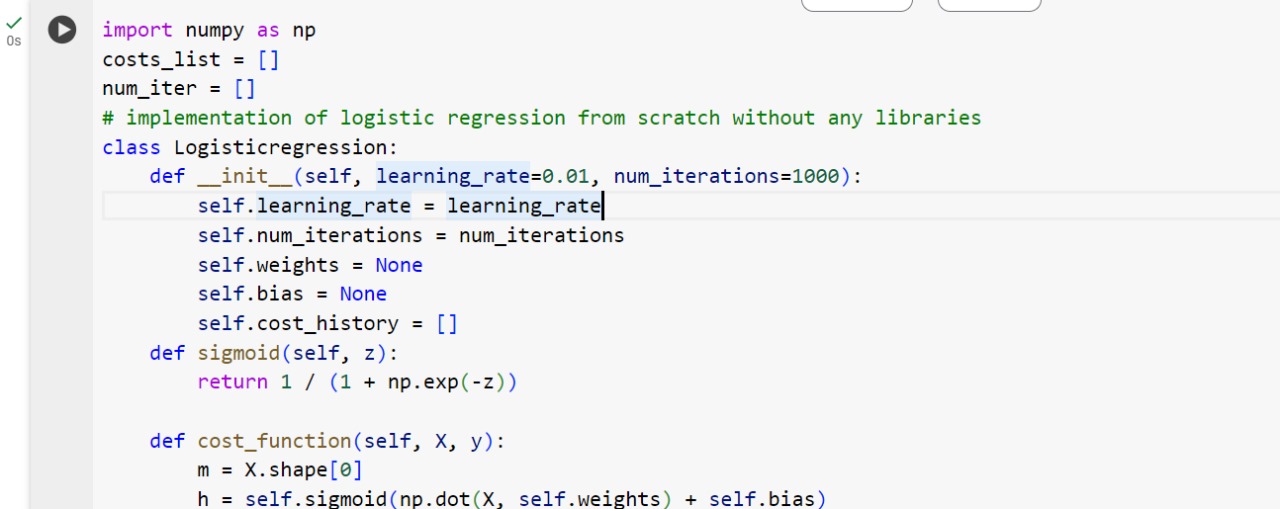
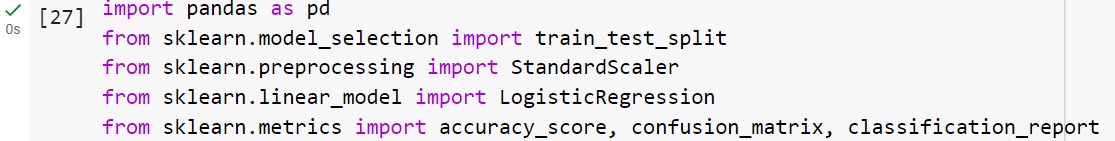
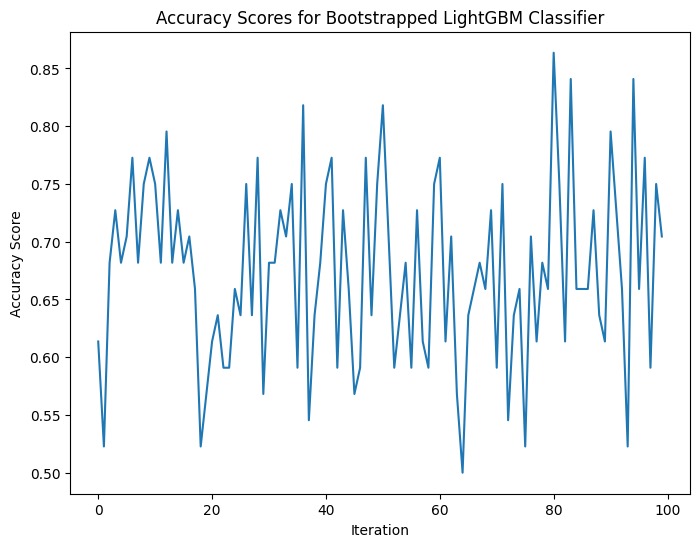
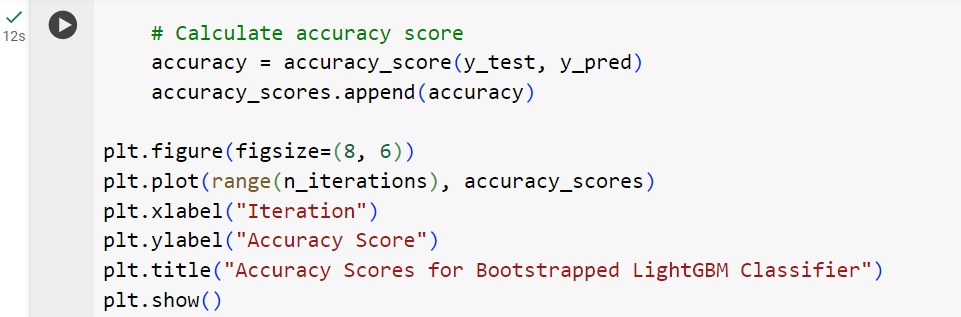
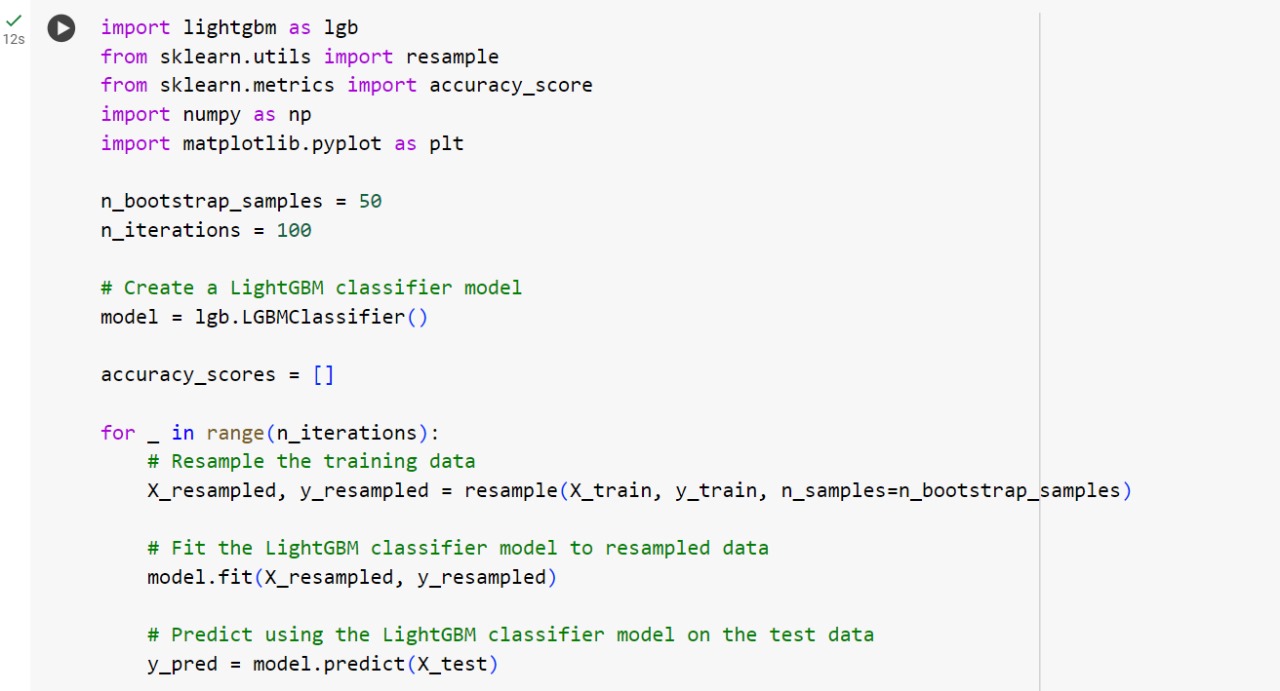
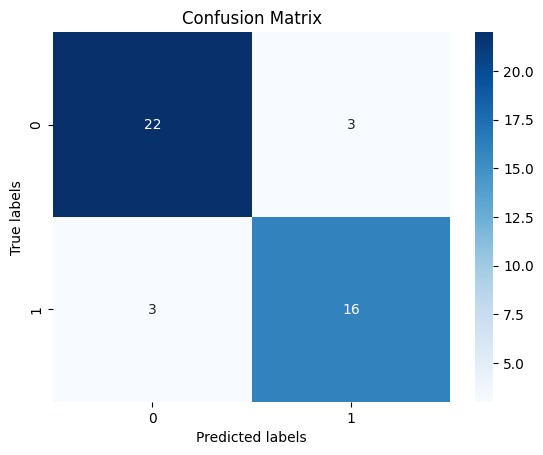
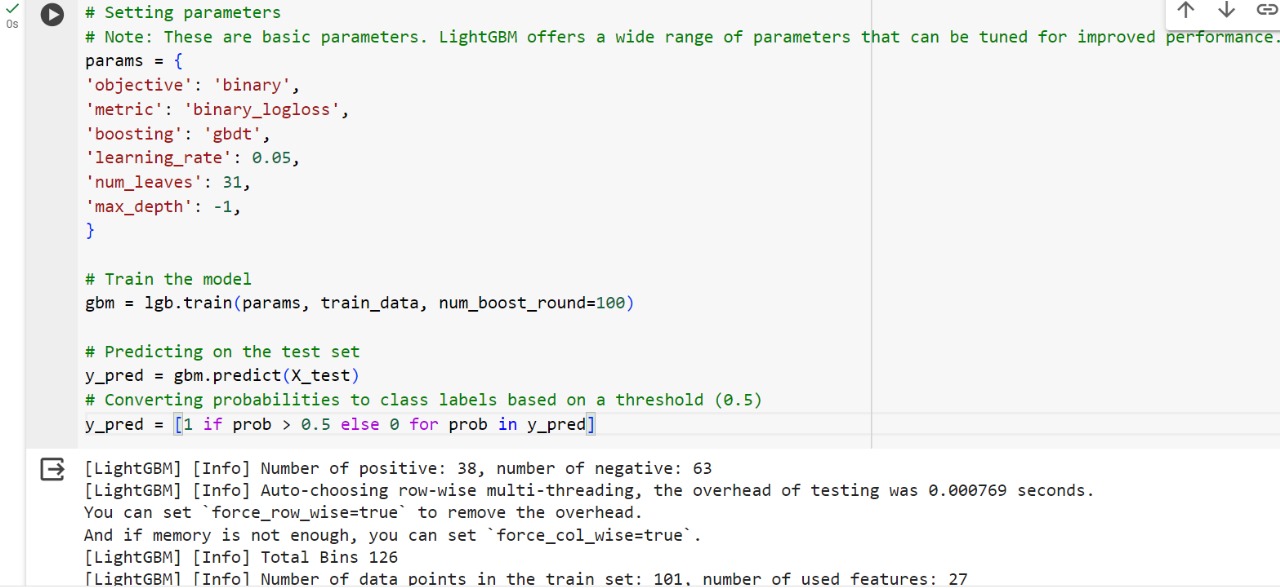












**9. CONCLUSION:**

Experts aimed to estimate student success charges in higher training the use of machine gaining knowledge of strategies Data amassed from academic establishments and engineering schools turned into classified into grades for analysis Factors including gender, high college type, scholarship kind, extra work, dating popularity, general revenue, and transportation manner have been considered in predicting scholar grades.i used the above mentioned classification methods comparing the all methods. I got the best accuracy and the best confusion matrix and also the f1 score is in light gbm and also in the XG Boost.so i conclude that the these methods are perfect for my dataset.

**REFERENCES**