

# Eligibility For Loan Approval Prediction Using Machine Learning Algorithms

## Abstract:

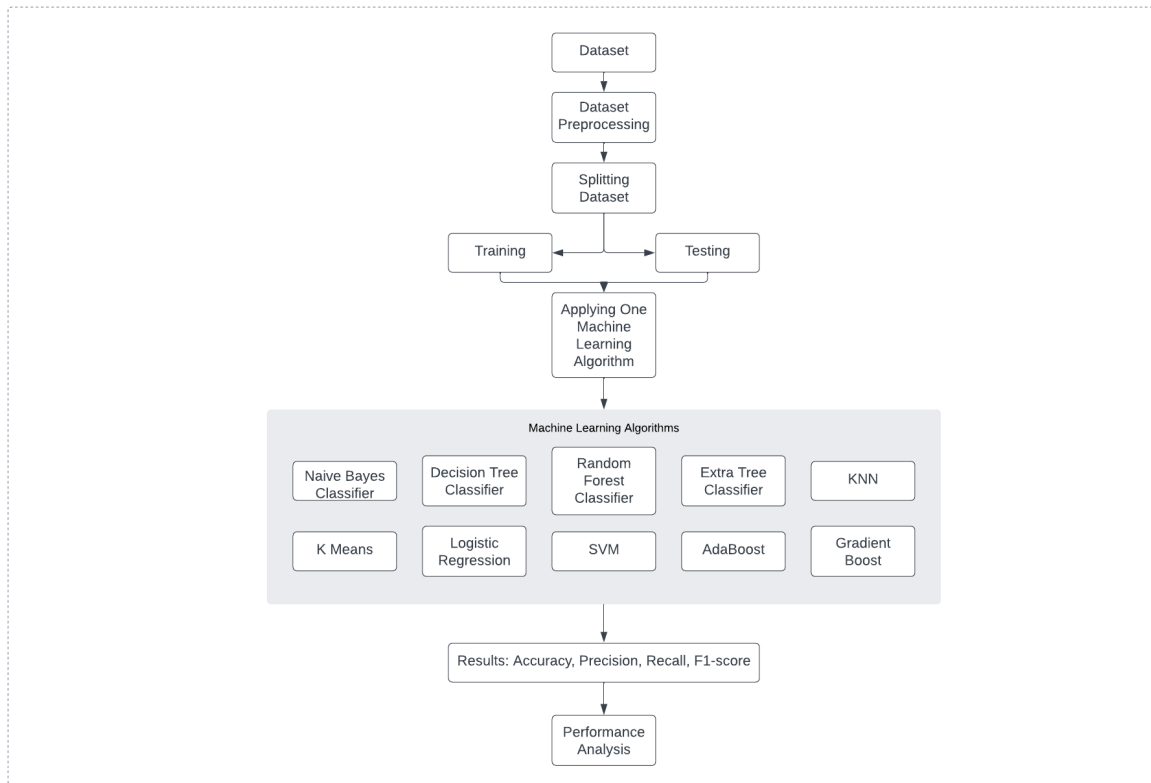
Along with the public's increasing aspirations, so is the demand for bank loans. Banks frequently receive many loan applications from customers and other people, but not all of them are accepted. Banks usually handle loan applications after confirming and assessing the applicant's eligibility—a difficult and time-consuming process. Most banks look at their credit score and use traditional risk assessment methods when looking at loan applications and deciding whether or not to approve loans. In spite of this, certain applicants persistently default on their payment obligations, causing significant losses for financial institutions. The purpose of this study is to determine which loan applications are eligible to be sanctioned by using various machine learning algorithms to extract patterns from a dataset of common loan application submissions. In addition, this study comprises feature extraction, data processing, and download of transaction data while taking into account the current circumstances of a bank credit application.

## Literature Survey:

When it comes to the subject of study, there are various methods employed by different researchers. Some focus on mainly one kind of algorithm, while others focus on multiple kinds of algorithms and determine which one is the most accurate and efficient. Papers [1], [2] from the references mainly deal with the decision tree algorithm approach in their study. Although their relative accuracy was pretty high, the main disadvantage can be attributed to the instability of the algorithm and the occurrence of overfitting-related issues. Papers [3], [4] both use a type of logistic regression model in their study. While Paper [4] uses a more conventional regression approach, Paper [3] incorporates stratified k-folds cross-validation mainly used to make sure the proportion of the feature of interest in the training and test sets is the same as it was in the original dataset. Papers [5], [6] primarily include comparative studies of different machine learning algorithms, mainly focusing on models based on Random Forest, K-Nearest Neighbor, XGBoost, AdaBoost, and so on. Papers [7], [8] deal with variations of neural network models, where [7] deals with BPNNs (back propagation neural networks), and [8] incorporates a neural network prediction model and penalty regression in a logistic regression model to come up with a more locally optimized algorithm. Paper [9] uses fuzzy logic, which is a mathematical method based on the concept of membership functions and uses a set of fuzzy rules. However, this method has been shown to be less efficient than other methods used in other studies and relies too much on human expertise and logic. Paper [10] uses models based on the C4.5 algorithm and Naive Bayes method. C4.5 algorithm is a type of decision tree classifier that employs a single pass pruning process to minimize the overfitting issues exhibited by traditional decision tree classifiers. However, it is found to be unable to compute continuous value and missing value attributes. When it comes to Naive Bayes classifiers, they have an edge over other models when it

comes to how simple it is to implement and the low training period for the model; however, it assumes that all attributes are mutually independent, which is not often true when it comes to attributes from real-life datasets.

## Basic Architecture for the Project:



## Methodology:

The effectiveness of machine learning algorithms in forecasting loan approval outcomes has been thoroughly researched in this article using a hybrid research methodology that incorporated quantitative and qualitative data. A cross-sectional research strategy was used to analyse data gathered from a particular moment in time for the study. The data set was gathered from Kaggle, a well-known platform for data science and machine learning, and was analyzed using a wide variety of statistical and machine learning approaches.

**Data Gathering:** Age, income, credit score, loan quantities and work status were just a few of the factors included in the data set, which came from Kaggle. To provide a representative sample, the data set was compiled from a variety of sources, including financial institutions and credit rating agencies.

**Preparation of Data:** A thorough data cleaning and preparation procedure was used to make sure the data set was correct and consistent. This includes filling in any gaps in the data, fixing any errors, and changing variables as necessary. Modern methods were used in the data preparation process to make sure that the data set was best prepared for machine learning

analysis.

**Analysis of Data:** The data collection was analyzed using a wide range of statistical and machine learning techniques. Techniques for exploratory data analysis were used to understand the data set and spot patterns and trends. To extract relevant characteristics and alter the data collection, feature engineering approaches were used.

**Algorithms Used:** The following machine learning algorithms were employed in this study:

1. Naive Bayes: A probabilistic algorithm used for binary classification. Based on the applicant's characteristics, including age, income, credit history, and job status, it was used to determine the likelihood that the loan application would be approved.
2. Decision Tree: A classification technique based on trees. It was used to pinpoint the key elements—like credit history, income, and work status—that influence loan acceptance.
3. Random Forest: An ensemble method that mixes many different decision trees to enhance performance and lessen overfitting. It was applied to improve decision tree algorithms' performance even further.
4. Extremely Randomized Trees Classifier: This version of the random forest method adds randomization to the splitting step to further decrease overfitting. It was used to improve the generalization capability of the model.
5. K-means: This clustering method was employed to put comparable loan applicants into groups based on characteristics including age, income, and loan size. It was used to data collection to find patterns and trends.
6. KNN: A distance-based classification technique that was applied to find a loan applicant's closest neighbors. The criteria of the closest neighbors were utilized to categorize the loan application.
7. Logistic Regression: For binary classification, this regression approach was utilized. It was employed to pinpoint the key elements that influence loan acceptance.
8. SVM: A classification algorithm that was used to find the best hyperplane that separates the loan approval and rejection classes. It was used to identify the most important features that determine loan approval.
9. AdaBoost: An ensemble algorithm that combines multiple weak classifiers to create a strong classifier. It was used to further enhance the performance of decision tree algorithms.
10. Gradient Boosting: An ensemble algorithm that combines multiple weak classifiers to create a strong classifier. It was used to further enhance the performance of decision tree algorithms.

## **Ethical Considerations:**

This study was conducted with utmost respect for the principles of data privacy and informed consent. The loan approval dataset used in this study was anonymized to protect the privacy

of loan applicants. Additionally, all participants involved in the study provided informed consent.

### **Limitations:**

This study has a few limitations. First, the results' generalizability may be constrained by the loan approval dataset's size and representativeness. Second, it's possible that factors other than those in the dataset are crucial for forecasting loan acceptance. The machine learning models created in this study might not be entirely accurate as a result. Additionally, the quality and amount of the accessible data may have an impact on how well the machine learning algorithms work.

The techniques used in this work were thoroughly described to support repeatability. The findings of this study can be easily replicated and further developed by other researchers interested in exploring the applications of machine learning in predicting loan approval. This study was conducted with utmost care to ensure that the results were accurate and transparent.

### **Results:**

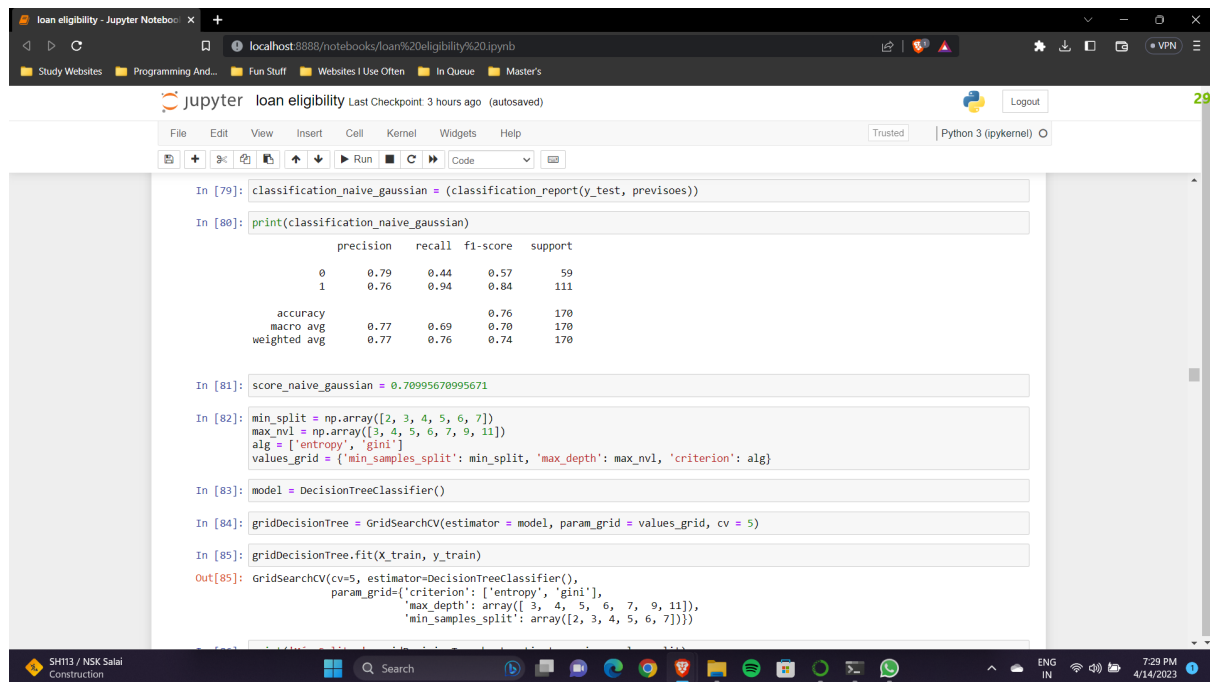
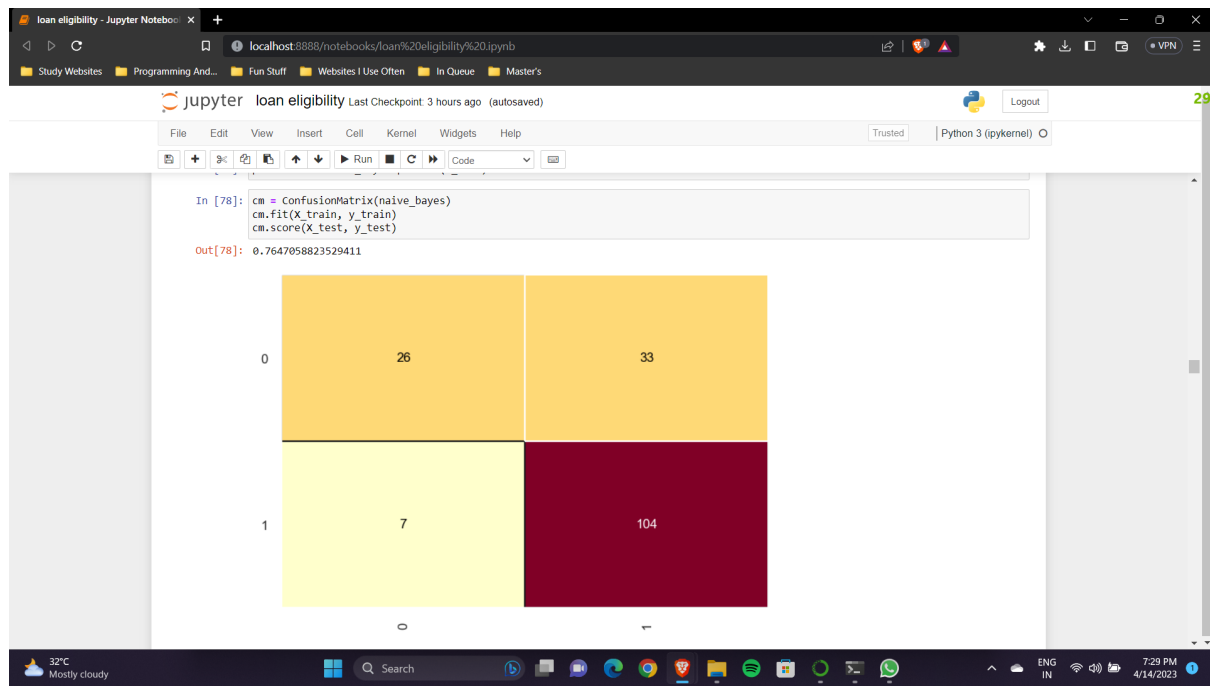
In this project, we need to determine if we can predict whether or not a customer is eligible for a loan. The dataset used (Loan-data.csv) is not adequately large enough, as it only has about 614 instances and 13 variables to consider. It can be observed that there are numeric variables as well as category variables; yet, our database is not particularly large, which makes our work more challenging. When we look at our database, we can see that there are some null values present, some of which can be handled while others cannot. For example, when we analyze the dataset, the Credit History variable is available, which has quite a lot of null values. If we fill in these null values, the authenticity of the predicted outcomes could be unrealistic, so we chose to omit the null values of the credit history variable while treating the other null values were treated. When we examine the correlation section, we see that there is no strong association between our variables. In the study, the variable that most piqued our interest was the credit history variable; when there is no credit history, we have almost an automatic refusal. When we analyze our target variable, we can see that there are a lot more yes values than no values, so we must balance the classes before running the machine learning models. In terms of machine learning models, there are several models that produced strong results; some models(KNN and Logistic regression) failed to predict the two outputs, while others performed admirably. Our main models were the Extremely Randomized Trees Classifier, Random Forest, and KNN models, but the model with the best balance of results and the highest accuracy value was the Random Forest model, which had an accuracy of 86%. This does not mean that the algorithms that failed to predict the outcome are bad, they are just not suited for this specific dataset and the ones that performed well have a much better fit for this particular dataset. When we examine the variables that best articulate our Machine Learning models, we find that the Credit\_History variable ranks first, the Loan Amount and Coapplicant variable ranks second, and the Applicant Income variable ranks

third, indicating that these are the most important variables to consider when analysing this problem.

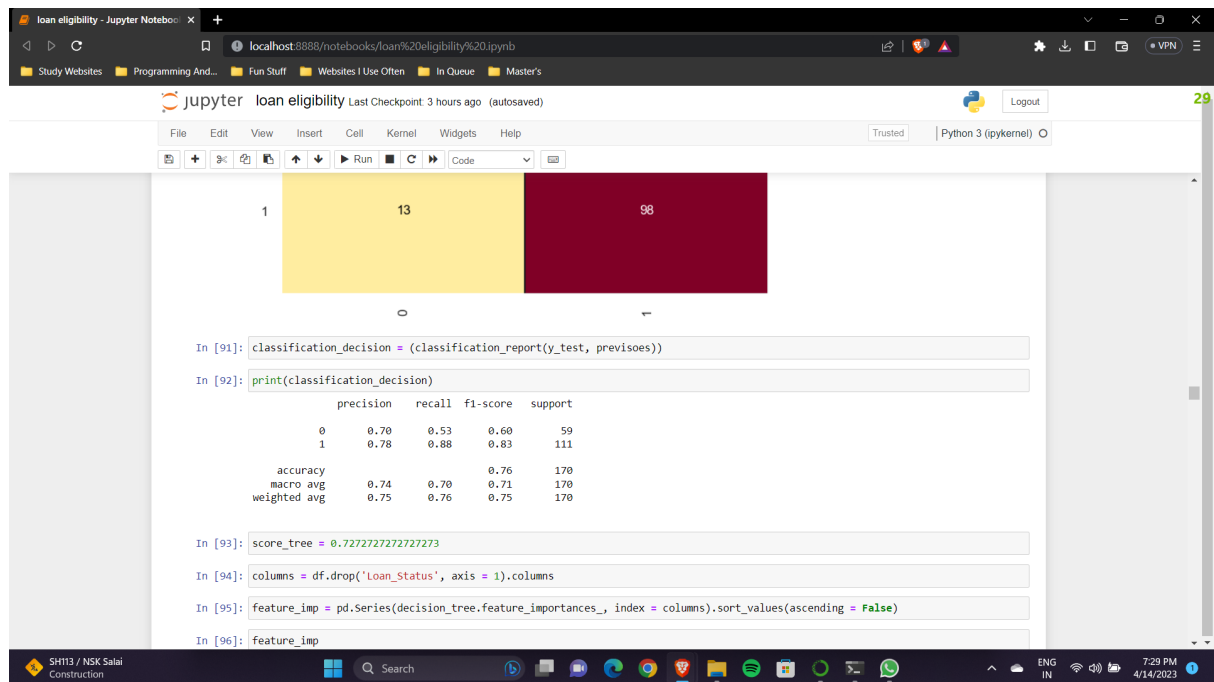
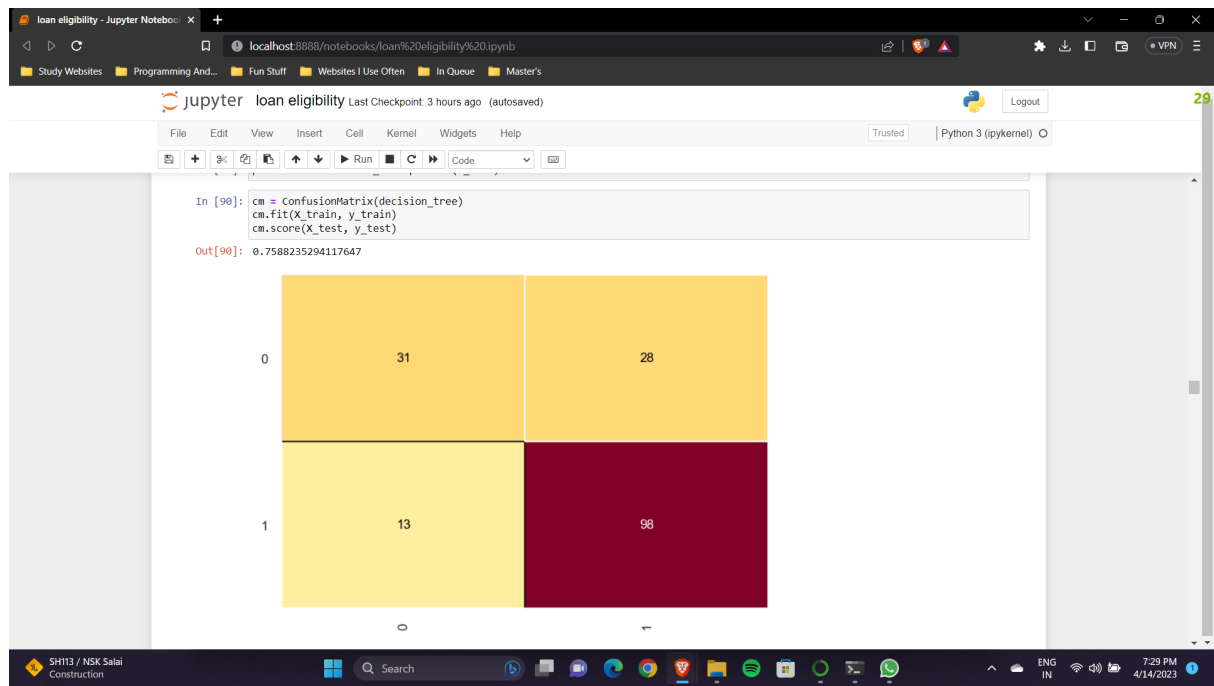
ALGORITHM	PRECISION	RECALL	f1-SCORE	ACCURACY	Predicted Y? (YES/NO)
Naïve Bayes	0.77	0.71	0.69	0.71	YES
Decision Tree	0.73	0.73	0.73	0.73	YES
Random Forest	0.86	0.86	0.86	0.86	YES
Extremely Randomized Trees Classifier	0.85	0.85	0.85	0.85	YES
K Means	0.51	0.50	0.51	0.51	YES
KNN	0.81	0.80	0.80	0.80	NO
Logistic Regression	0.73	0.71	0.71	0.71	NO
SVM	0.78	0.78	0.78	0.78	YES
AdaBoost	0.80	0.80	0.80	0.80	YES
Gradient Boosting	0.81	0.80	0.79	0.80	YES

Table 1: Overall Results For All Algorithms

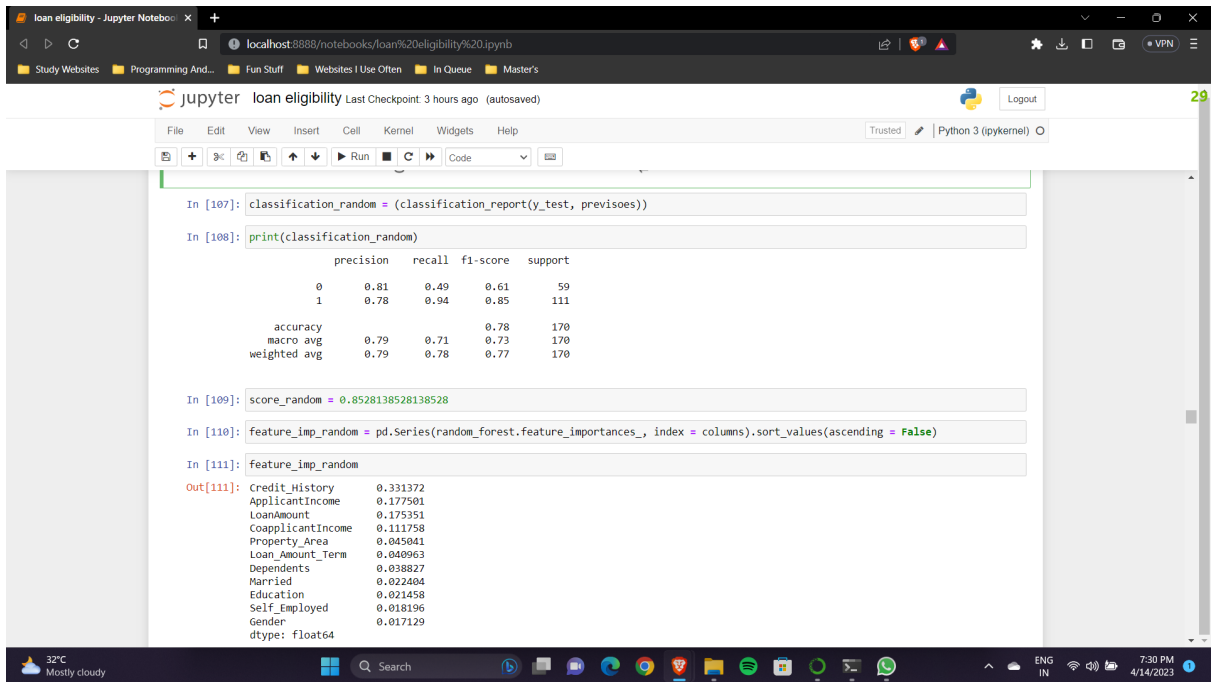
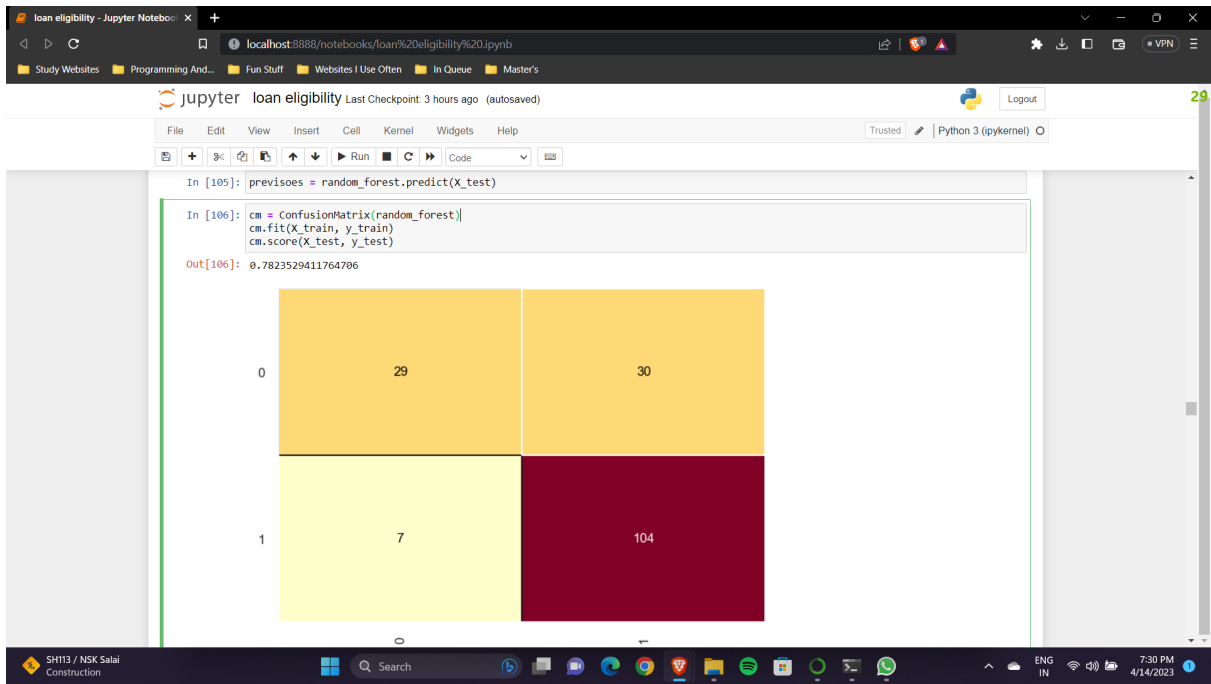
**Confusion Matrix for Naive Bayes Classifier:**



**Confusion Matrix for Decision Tree:**

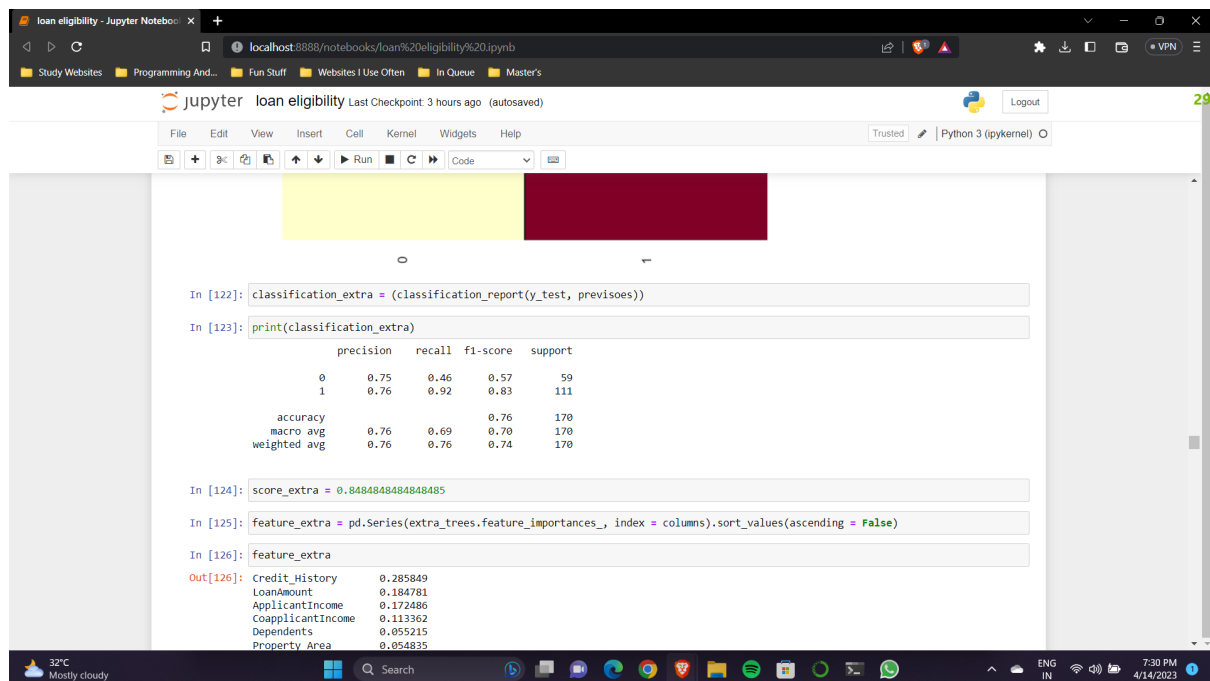
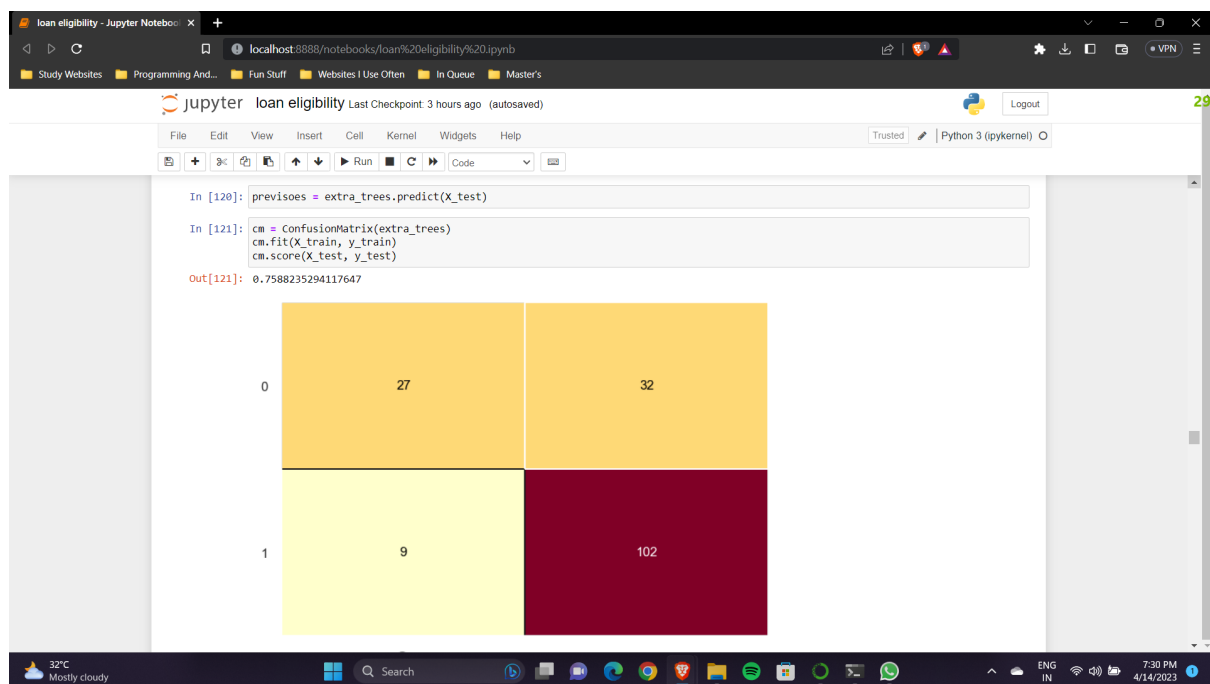


**Confusion Matrix for Random Forest:**

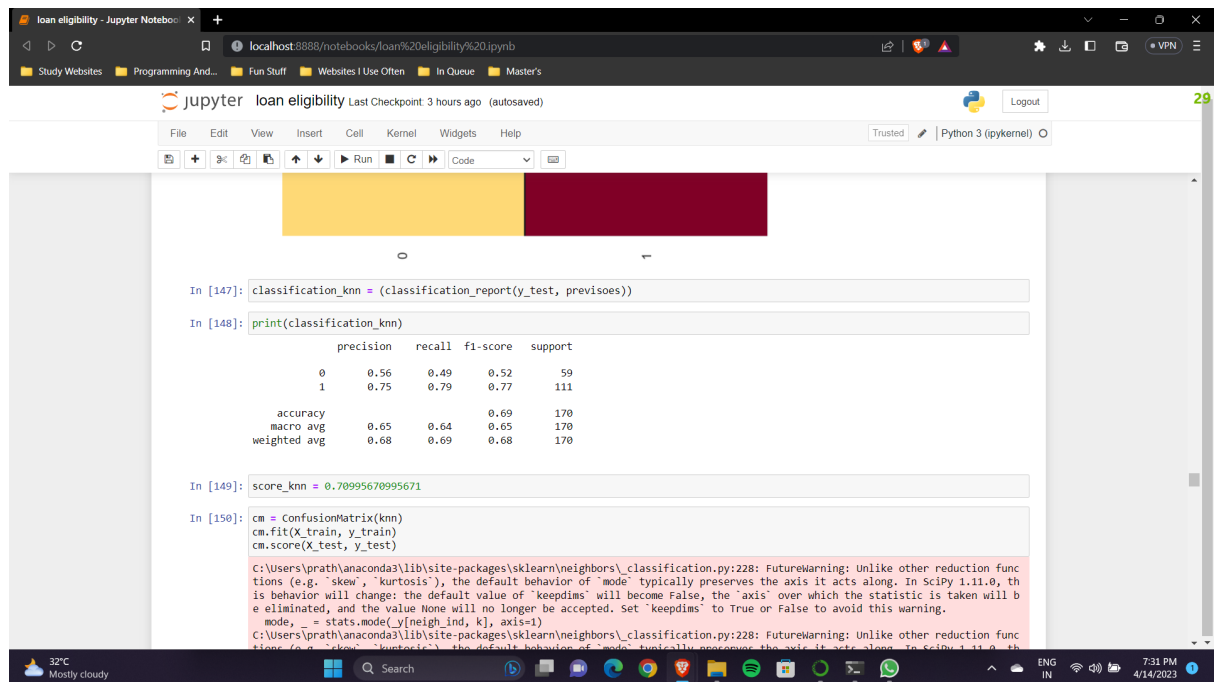
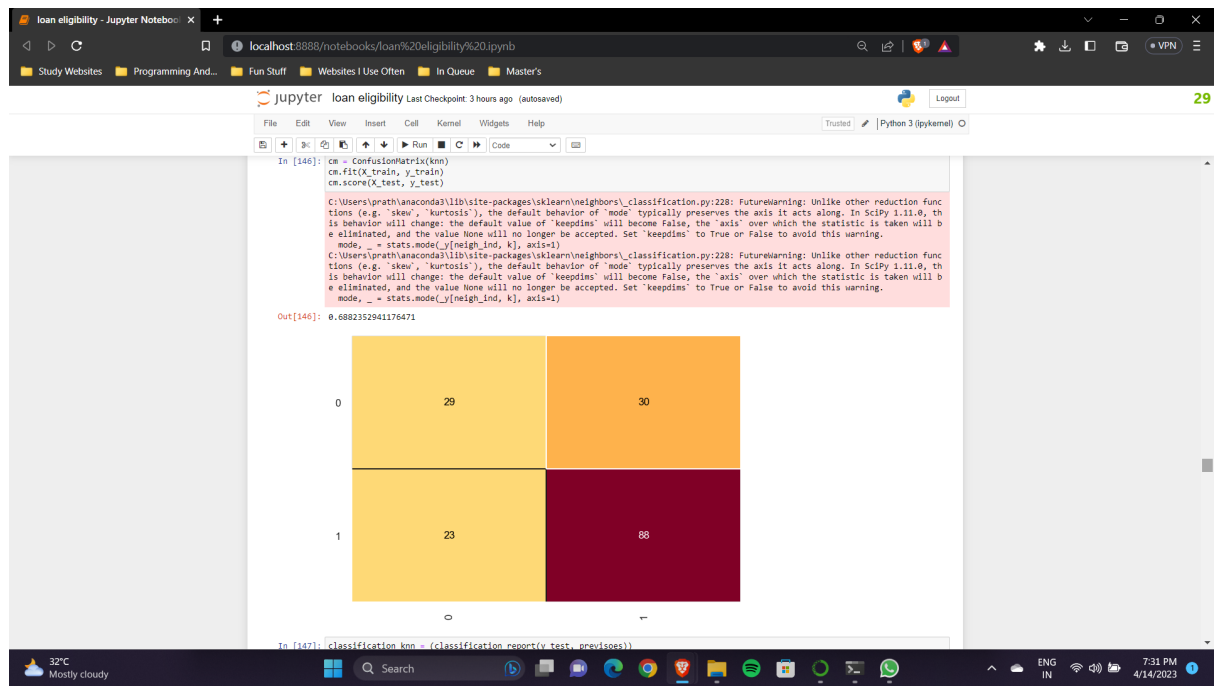




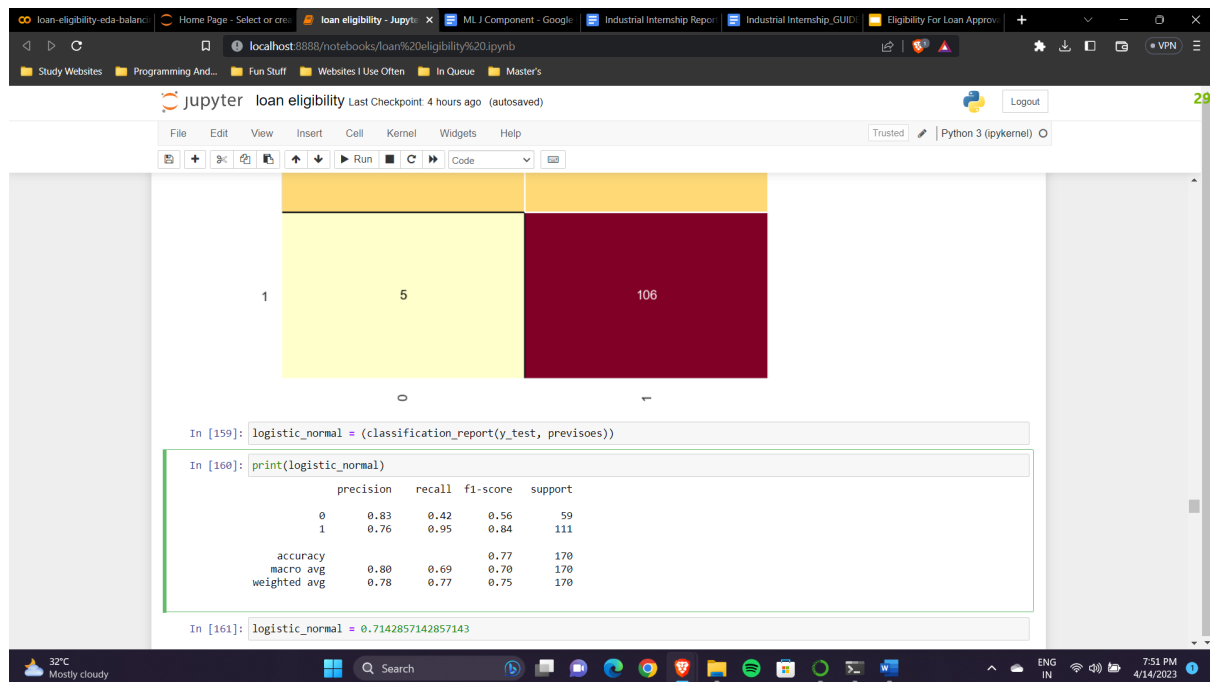
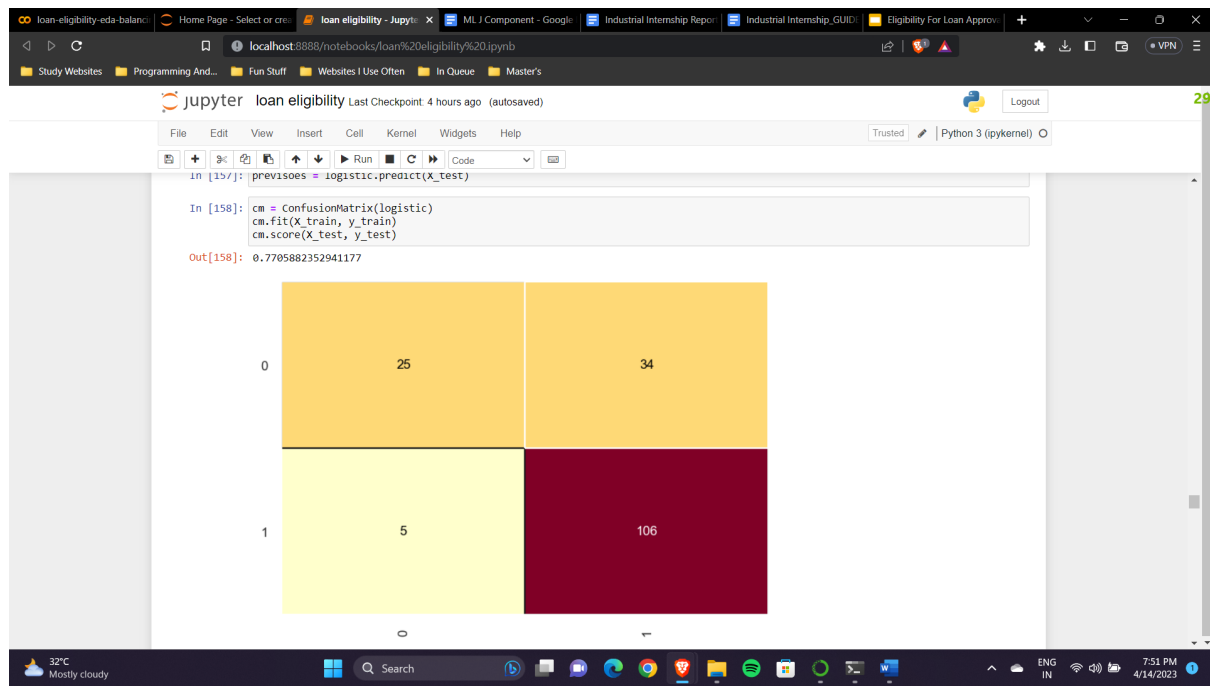
## Confusion Matrix for Extremely Randomized Trees:



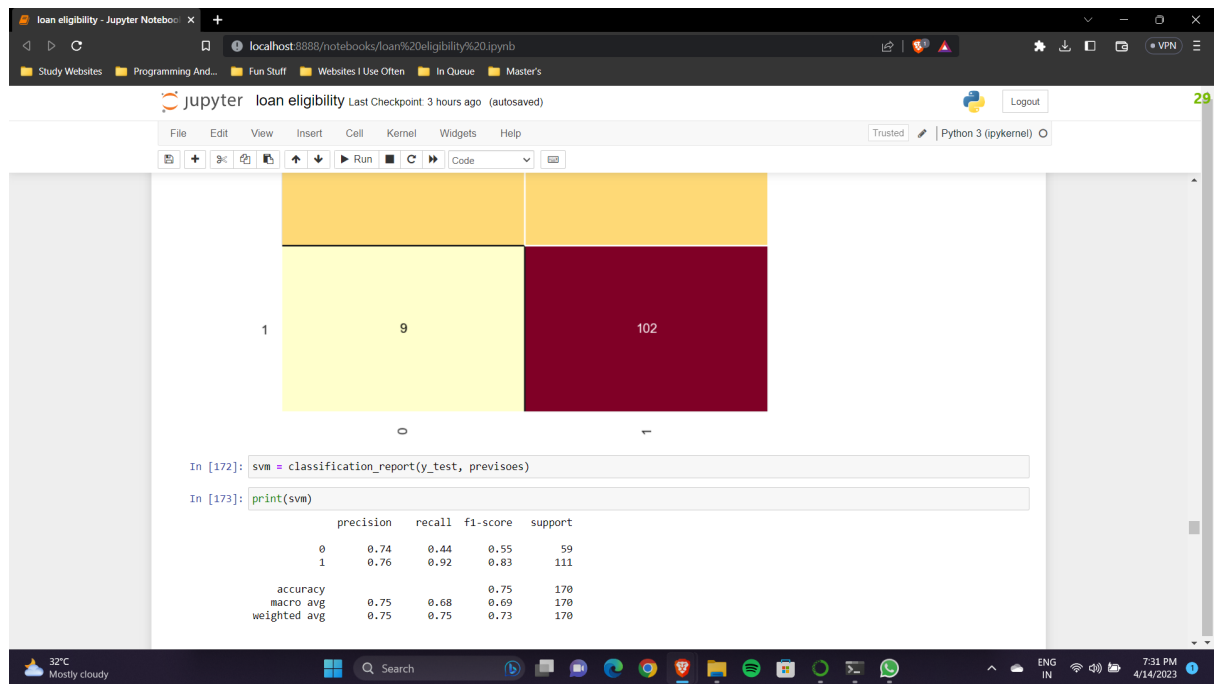
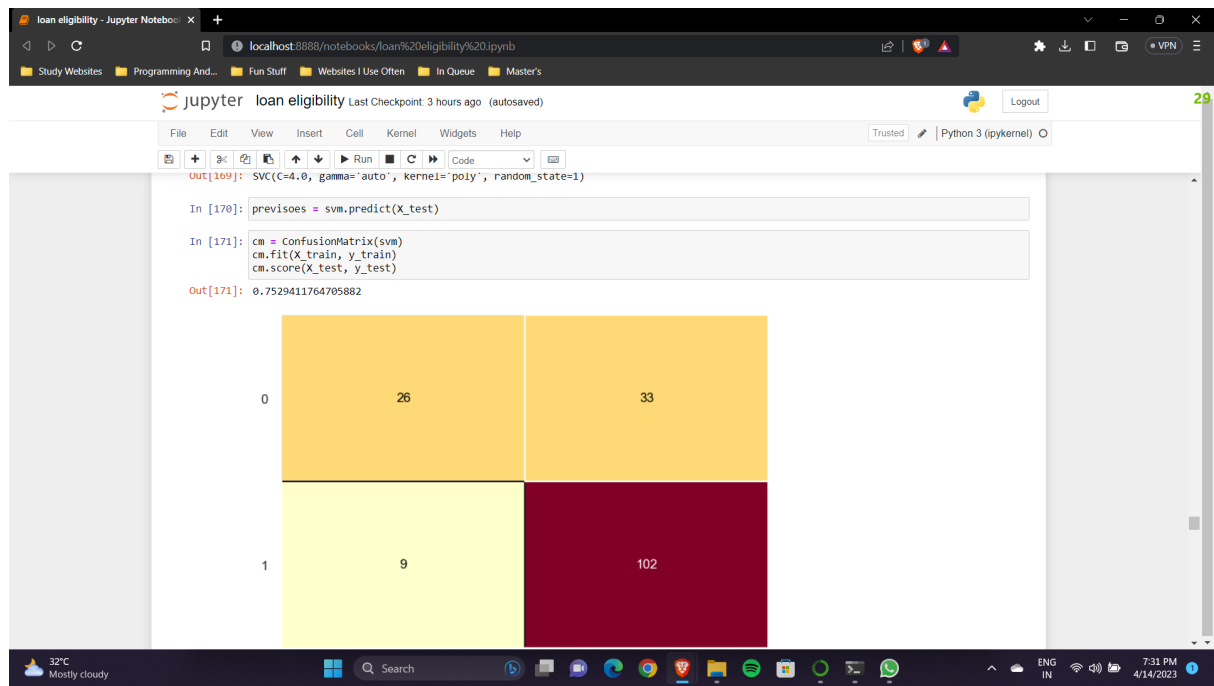
## Confusion Matrix for KNN Classifier:



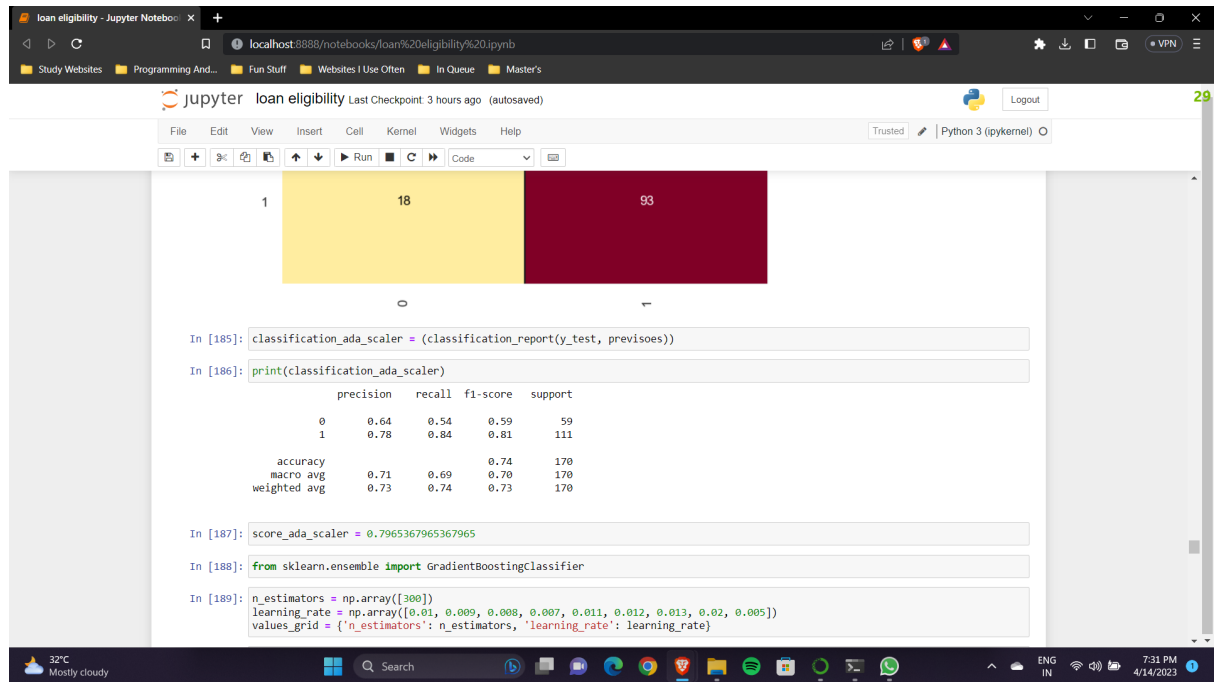
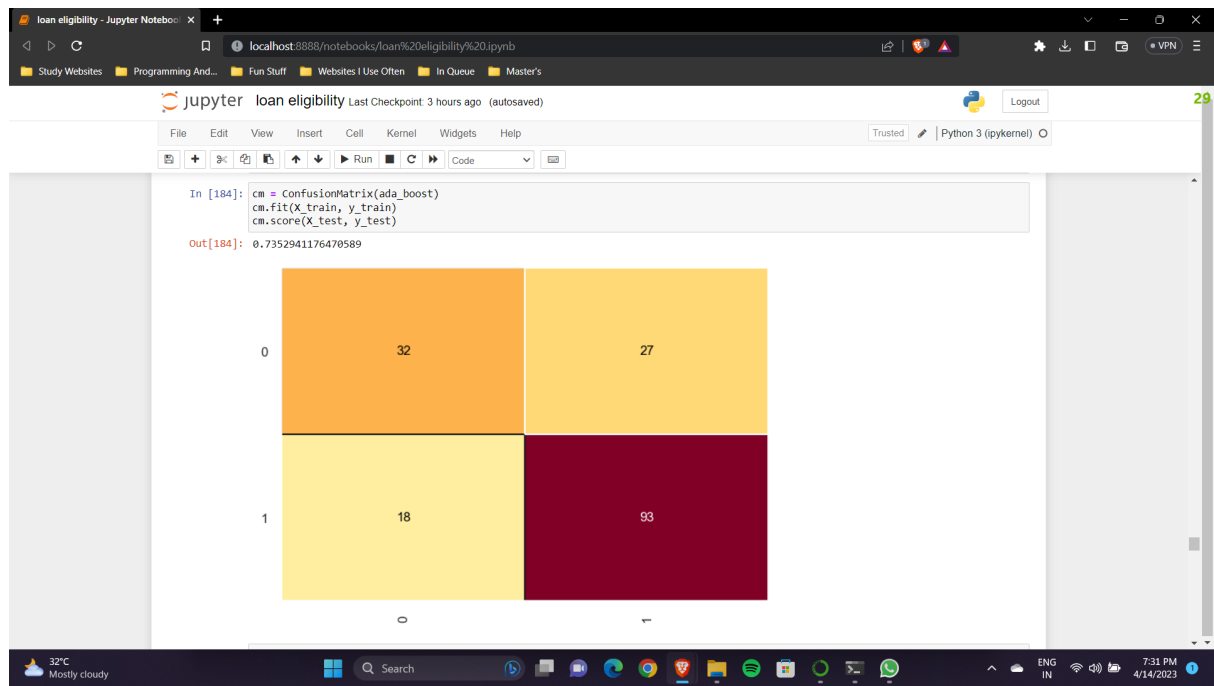
## Confusion Matrix for Logistic Regression:



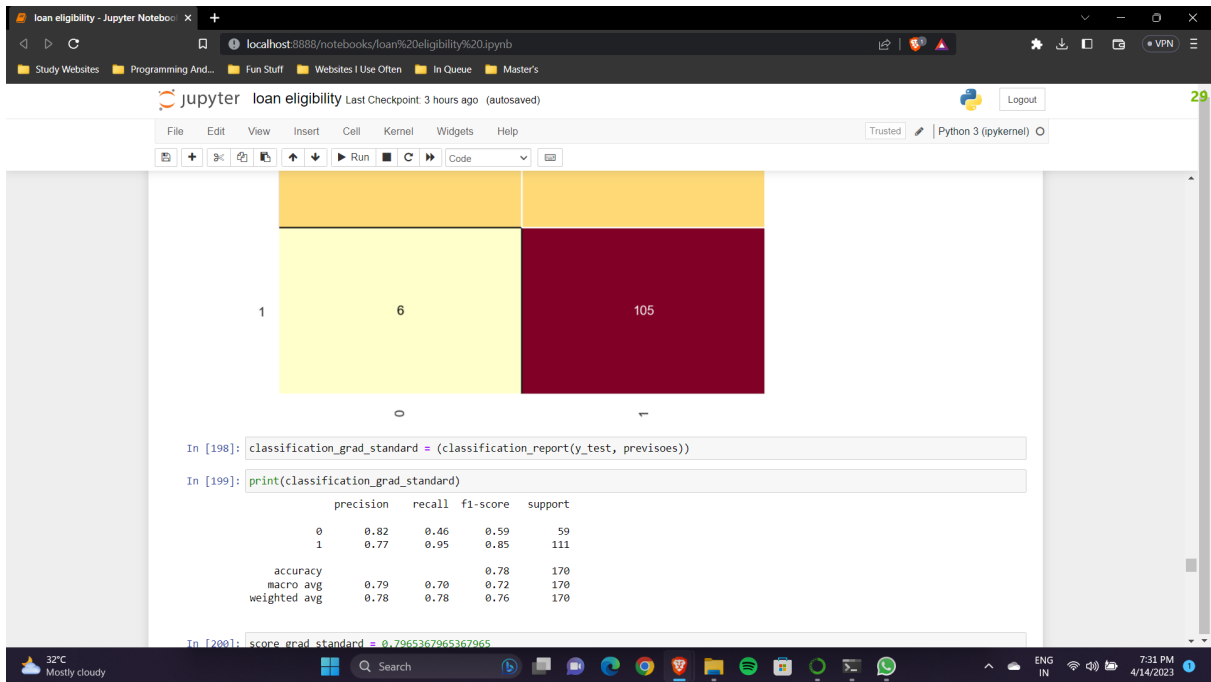
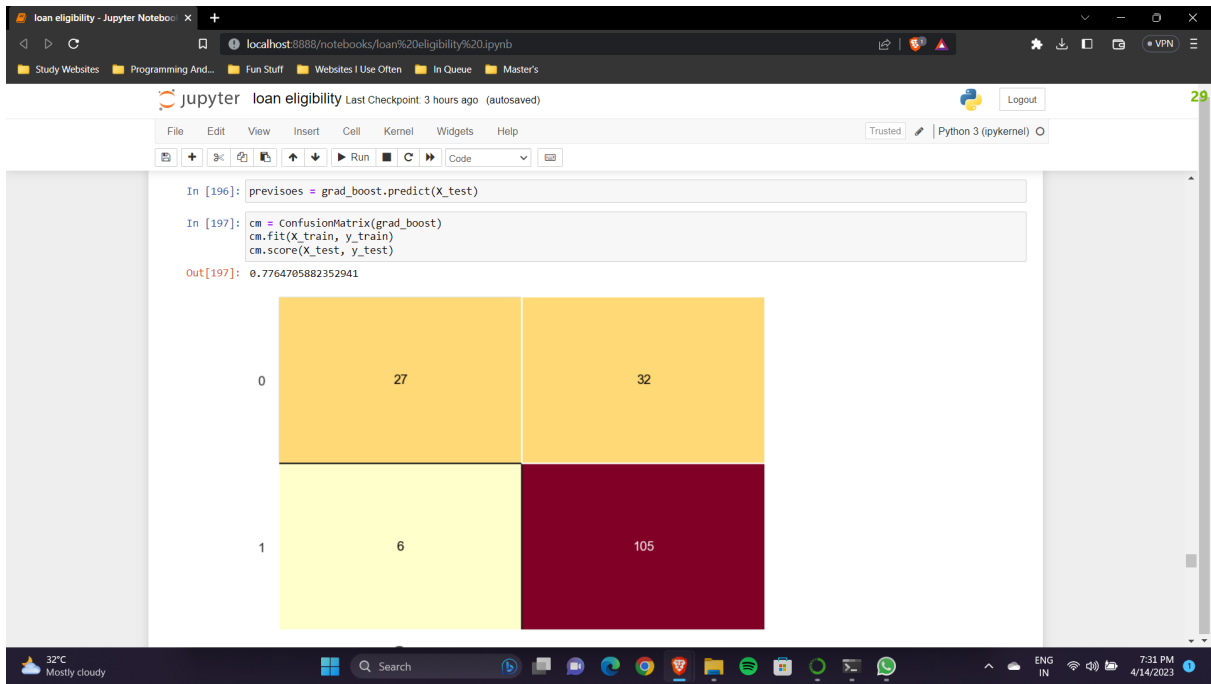
**Confusion Matrix for SVM:**



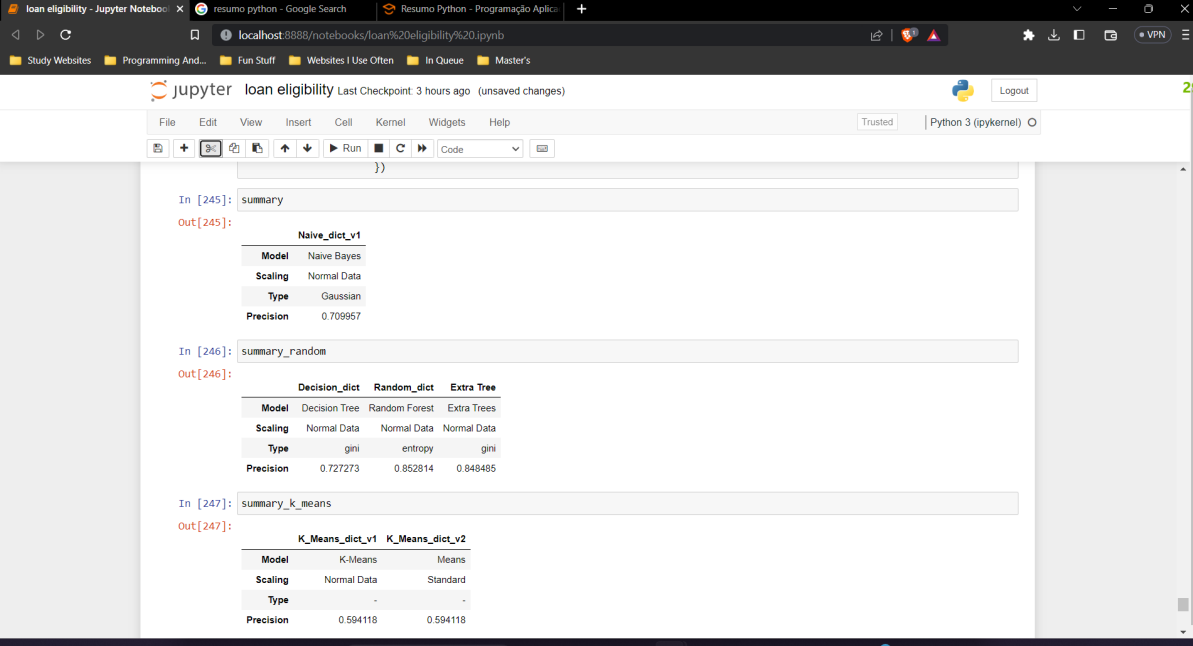
**Confusion Matrix for AdaBoost:**



**Confusion Matrix for Gradient Boosting:**



## Summary Dictionary For All The Algorithms:



Jupyter Notebook interface showing the summary dictionary for three algorithms: Naive Bayes, Random Forest, and K-Means. The notebook is titled "loan eligibility" and is running on Python 3 (pykernel).

**In [245]: summary**

**Out[245]:**

Naive_dict_v1	
Model	Naive Bayes
Scaling	Normal Data
Type	Gaussian
Precision	0.709957

**In [246]: summary\_random**

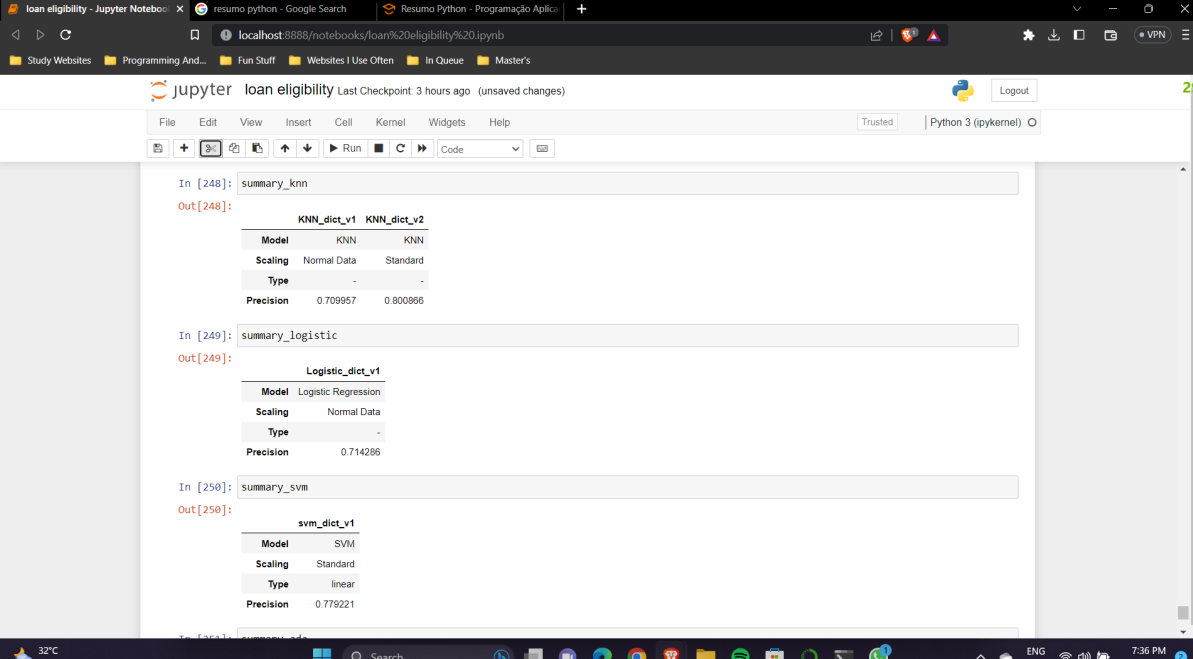
**Out[246]:**

	Decision_dict	Random_dict	Extra Tree
Model	Decision Tree	Random Forest	Extra Trees
Scaling	Normal Data	Normal Data	Normal Data
Type	gini	entropy	gini
Precision	0.727273	0.852814	0.848485

**In [247]: summary\_k\_means**

**Out[247]:**

	K_Means_dict_v1	K_Means_dict_v2
Model	K-Means	Means
Scaling	Normal Data	Standard
Type	-	-
Precision	0.594118	0.594118



Jupyter Notebook interface showing the summary dictionary for three algorithms: K-Nearest Neighbors, Logistic Regression, and SVM. The notebook is titled "loan eligibility" and is running on Python 3 (pykernel).

**In [248]: summary\_knn**

**Out[248]:**

	KNN_dict_v1	KNN_dict_v2
Model	KNN	KNN
Scaling	Normal Data	Standard
Type	-	-
Precision	0.709957	0.800866

**In [249]: summary\_logistic**

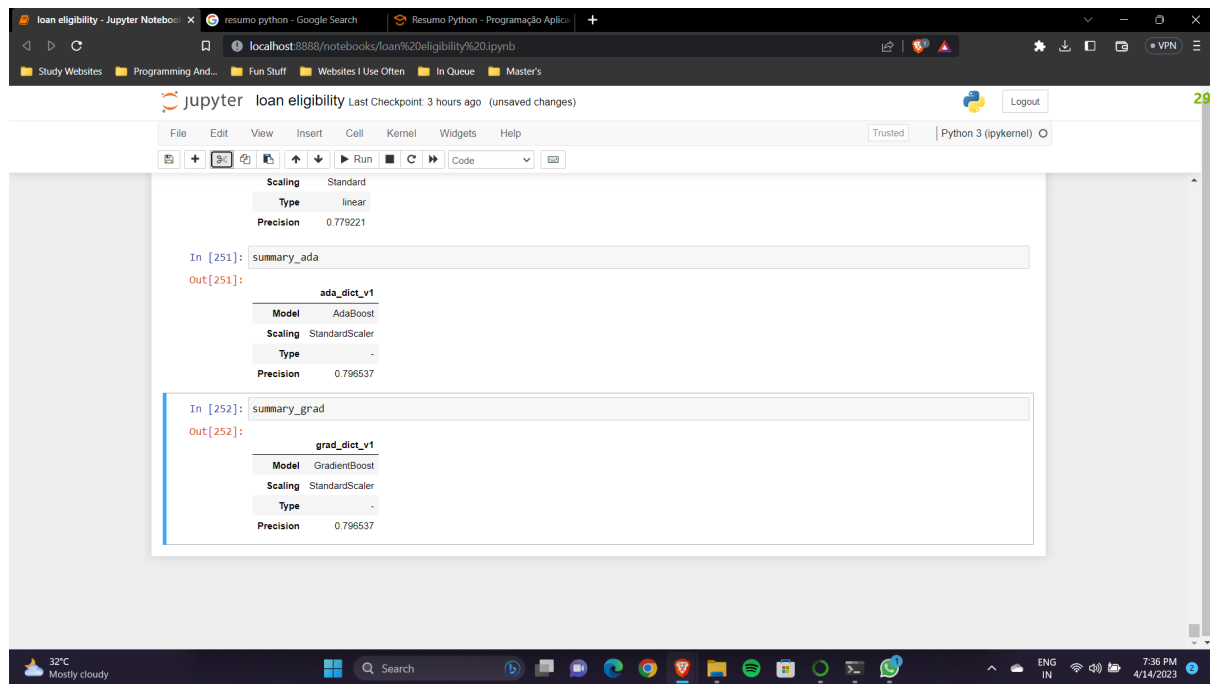
**Out[249]:**

Logistic_dict_v1	
Model	Logistic Regression
Scaling	Normal Data
Type	-
Precision	0.714286

**In [250]: summary\_svm**

**Out[250]:**

svm_dict_v1	
Model	SVM
Scaling	Standard
Type	linear
Precision	0.779221



## Details Of Software:

The software used for this project includes Microsoft Excel, in which the dataset is viewed and stored, and Jupyter Notebook for implementing the algorithm in Python. This also includes installing many Python libraries, such as Pandas, Numpy, Matplotlib, Seaborn, sci-kit learn, and Yellowbrick.

## References:

- 1) Rutika Pramod Kathe, Sakshi Dattatray Panhale, Pooja Prakash Avhad, Dinesh. B. Ghorpade, Punam Laxman Dapase. "AN APPROACH FOR PREDICTION OF LOAN APPROVAL USING MACHINE LEARNING ALGORITHM", International Journal of Creative Research Thoughts (IJCRT), ISSN:2320-2882, Vol.9, Issue 6, pp.c568-c570, June 2021, URL : <http://www.ijcrt.org/IJCRT2106313>
- 2) Gomathy, C.K., Charulatha, M., Aakash, M. and Sowjanya, M., 2021. THE LOAN PREDICTION USING MACHINE LEARNING.
- 3) "Loan Prediction System Using Machine Learning" Anant Shinde, Yash Patil, Ishan Kotian, Abhinav Shinde, Reshma Gulwani ITM Web Conf. 44 03019 (2022), DOI: 10.1051/itmconf/20224403019
- 4) Ramya, P. S. Jha, Ilaa Raghupathi Vasishtha and Neha Zafar. "Monetary Loan Eligibility Prediction using Machine Learning." (2021).
- 5) Sarkar, A., 2021. Machine learning techniques for recognizing the loan eligibility. International Research Journal of Modernization in Engineering Technology and Science, 3(12), pp.1135-1142.



- 6) Al Mamun, M., Farjana, A. and Mamun, M., Predicting Bank Loan Eligibility Using Machine Learning Models and Comparison Analysis.
- 7) Baodong Li, "Online Loan Default Prediction Model Based on Deep Learning Neural Network", Computational Intelligence and Neuroscience, vol. 2022, Article ID 4276253, 9 pages, 2022. <https://doi.org/10.1155/2022/4276253>
- 8) Herui Chen, "Prediction and Analysis of Financial Default Loan Behavior Based on Machine Learning Model", Computational Intelligence and Neuroscience, vol. 2022, Article ID 7907210, 10 pages, 2022. <https://doi.org/10.1155/2022/7907210>
- 9) Ginting, S.L.B., Rizky, M.R.M. and Ginting, Y.R., 2020, July. The Application of Fuzzy Logic Method in the Debtors Eligibility Assessment System of Microfinance Institution. In IOP Conference Series: Materials Science and Engineering (Vol. 879, No. 1, p. 012039). IOP Publishing.
- 10) Maulana, I. and Subchan, M., 2021. Prediction Model of Eligibility of Lending in Credit Banks Using The C4. 5 Algorithm and Naive Bayes Method. Jurnal Mantik, 5(3), pp.1791-1798.