Loan Approval Prediction

1.INTRODUCTION

Loan Distribution is the main business part of many

banks. The main portion of banks income comes from the

loan distributed to customers. These banks apply interest

on loan which are distributed to customers.

The main objective of banks is to invest their assets

in safe customers. Up to now many banks are processing

loans after regress process of verification and validation.

But till now no bank can give surety that the customer

who is chosen for loan application is safe or not. So to

avoid this situation we introduced a system for the

approval of bank loans known as Loan Prediction System

Using Python.

Loan Prediction System is a software which checks

the eligibility of a particular customer who is capable of

paying loan or not. This system checks various parameters

such as customer’s martial status, income, expenditure

and various factors. This process is applied for many

customers of trained data set. By considering these factors

a required model is built. This model is applied on the test

data set for getting required output. The output generated

will be in the form of yes or no. Yes indicates that a

particular customer is capable of paying loan and no

indicates that the particular customer is not capable of

paying loan. Based on these factors we can approve loans

for customers

2. LITERATURE SURVEY

Data Analysis for prediction of loan based nature of

clients

The report main intention is to classify the nature of

clients for loans. Depending upon the certain factors the

report classifies the customers. Classification is done

through exploratory data analyses [1].

Exploratory data analysis is a technique that analyzes

and summaries the main features from training dataset.

Prediction of Loan Approval using Machine Learning

Approach

Machine learning [2] is a phenomenon in which

analytical model is build from the trained model.

This model is applied on test data for providing of the

accurate results.

Here the author used three algorithms for prediction of

loan. They are

1. K Nearest Neighbor

2. Decision Tree

3. Random Forests

The main purpose of this report is to provide

immediate and accurate results for the approval of loan to

the eligible customers. In banking sector there will be n

number of people who apply loans. It is difficult to check

customer’s eligibility through paper work. The system can

provide accurate results for the n number of people.Building the model using Random Forest Approach In this report we have discussed about credit risk and

credit analysis. Banking sectors success mainly depends of

credit risk analysis. In this report we have used Random

Forest [3] approach to build the model. The use of Random

Forest is because Random Forest Approach provides

accurate results than the K Nearest Neighbor and Decision

Tree.

Ensemble model survey for loan prediction

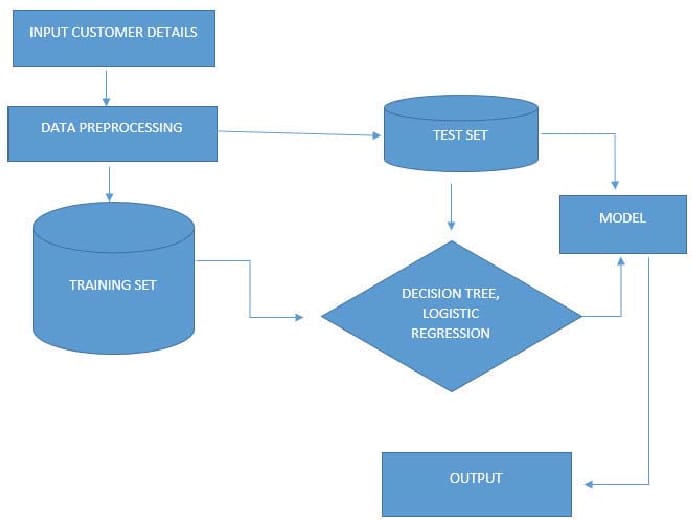
In this report author has used Random Forest approach

for building a model. In this report two or more classifiers

are combined together and identify a perfect model for

loan prediction.

3.THEORITICAL ANALYSIS



HARDWARE / SOFTWARE DESIGNING :

HARDWARE : PC

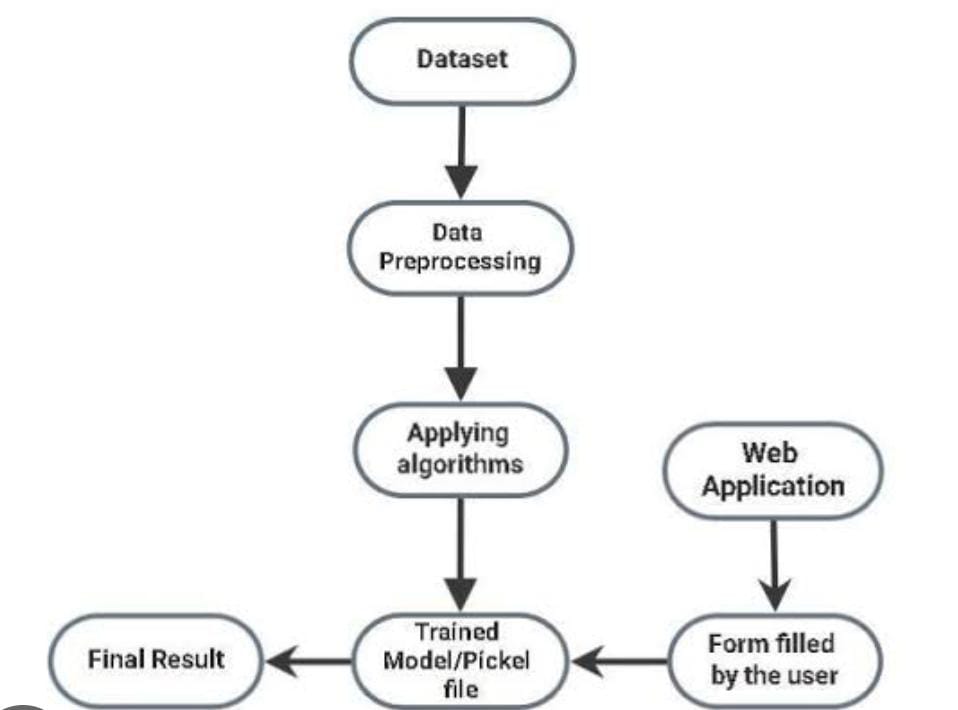
SOFTWARE : Anaconda (jupyter notebook)

Anaconda Navigator is a desktop graphical user interface (GUI) included in Anaconda distribution that allows users to launch applications and manage conda packages, environments and channels without using command-line commands. The Jupyter Notebook application allows you to create and edit documents that display the input and output of a Python or R language script. Once saved, you can share these files with others. NOTE: Python and R language are included by default, but with customization, Notebook can run several other kernel environments.

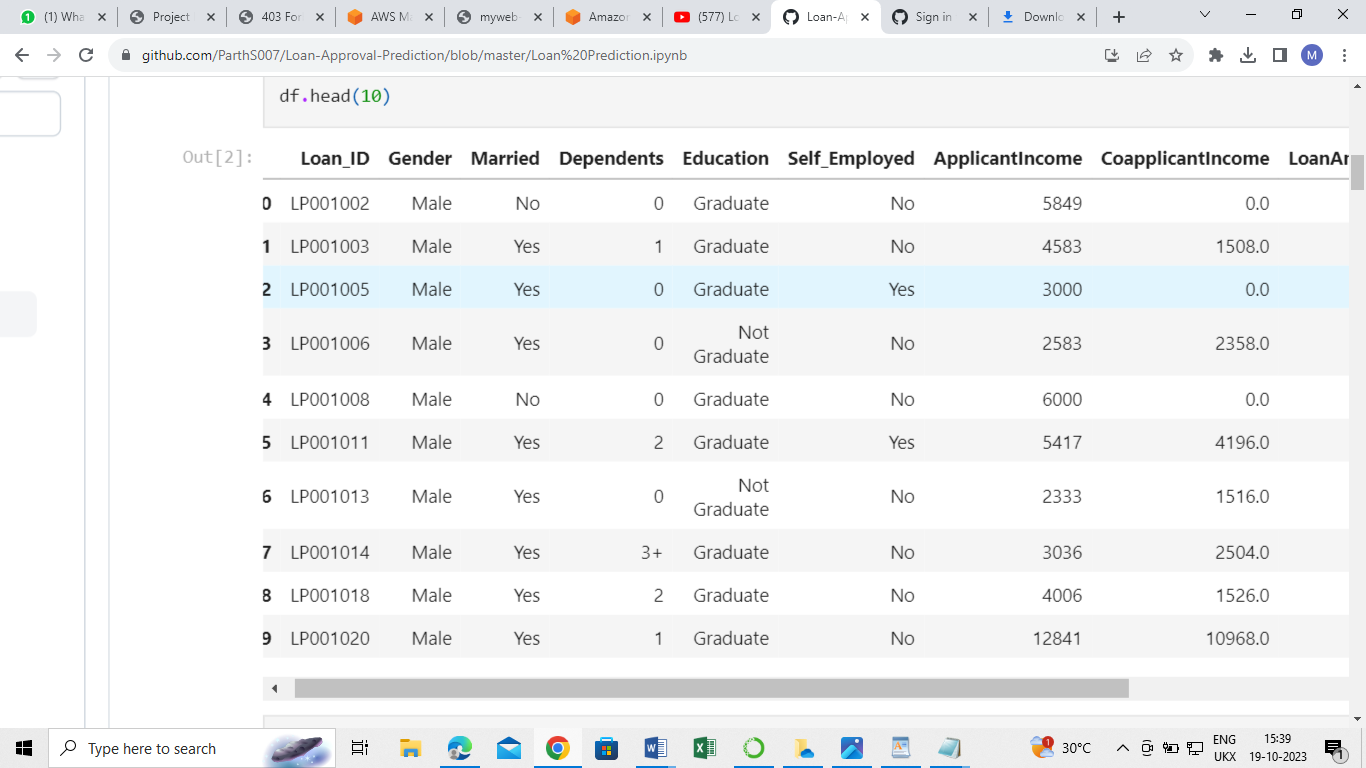
4.EXPERIMENTAL INVESTIGATIONS

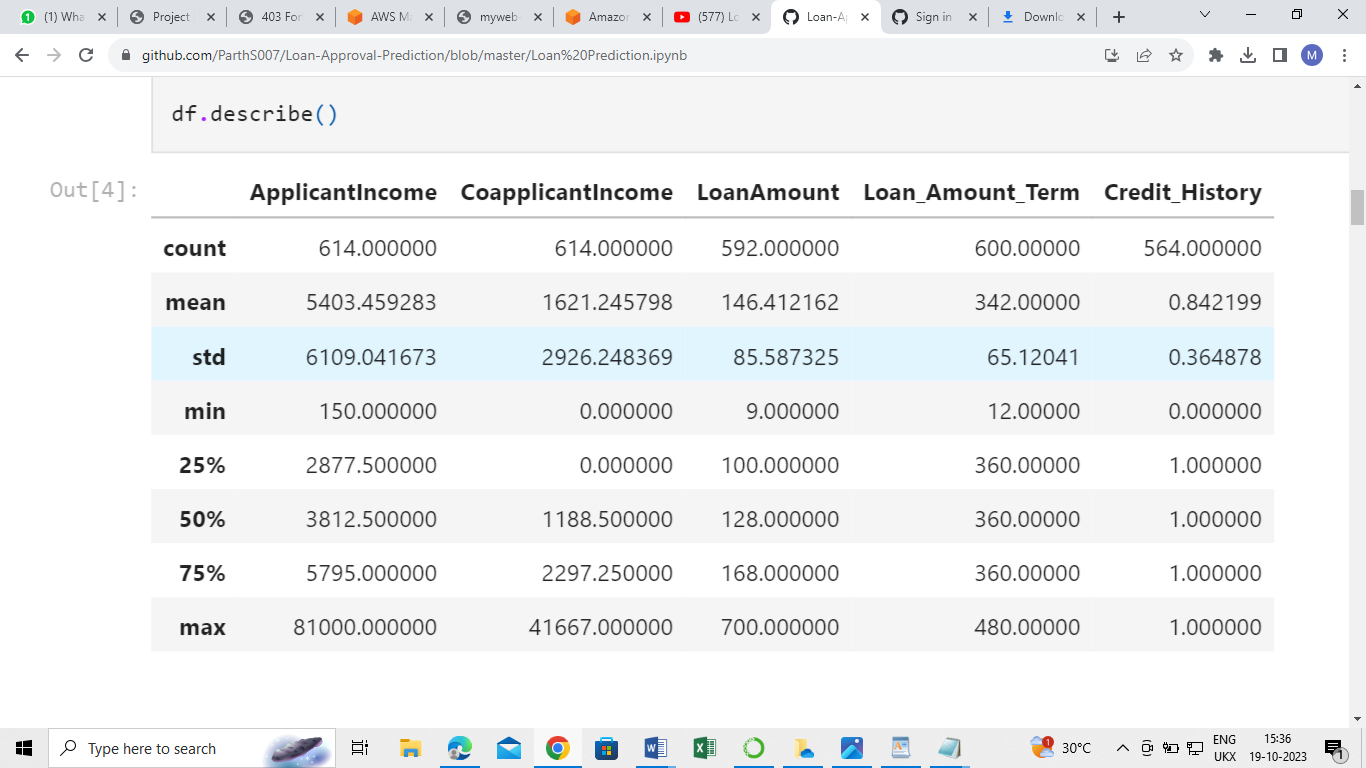
Data Preprocessing:Cleaning and handling missing data in the dataset.Feature engineering to create relevant features from the existing data.Exploratory Data Analysis (EDA) to understand the data distribution, correlations, and outliers.Feature Selection:Identifying the most important features that influence loan approval decisions.Utilizing techniques like feature importance scores and correlation matrices.Model Selection:Evaluating various machine learning algorithms (such as Decision Trees, Random Forest, Logistic Regression, etc.) to determine the best fit for the problem.Comparing model performance using metrics like accuracy, precision, recall, and F1-score.Model Training and Tuning:Splitting the data into training and testing sets to train the model and assess its performance.Hyperparameter tuning to optimize the model for better accuracy and generalization.Handling Imbalanced Data:Addressing class imbalance issues by employing techniques like oversampling, undersampling, or using algorithms designed for imbalanced data.Risk Assessment:Assessing the potential risks associated with false positives (approving a bad loan) and false negatives (rejecting a good loan).Adjusting the model thresholds to balance between these risks based on business requirements.Interpretability and Explainability:Ensuring the model's decisions are interpretable, especially in industries where regulatory compliance and ethical considerations are crucial.Using techniques like SHAP (SHapley Additive exPlanations) values or LIME (Local Interpretable Model-agnostic Explanations) to explain individual predictions.Monitoring and Updating:Setting up mechanisms to monitor the model's performance in real-time.Establishing protocols for model updates and retraining to adapt to changing patterns and behaviors in loan applications.

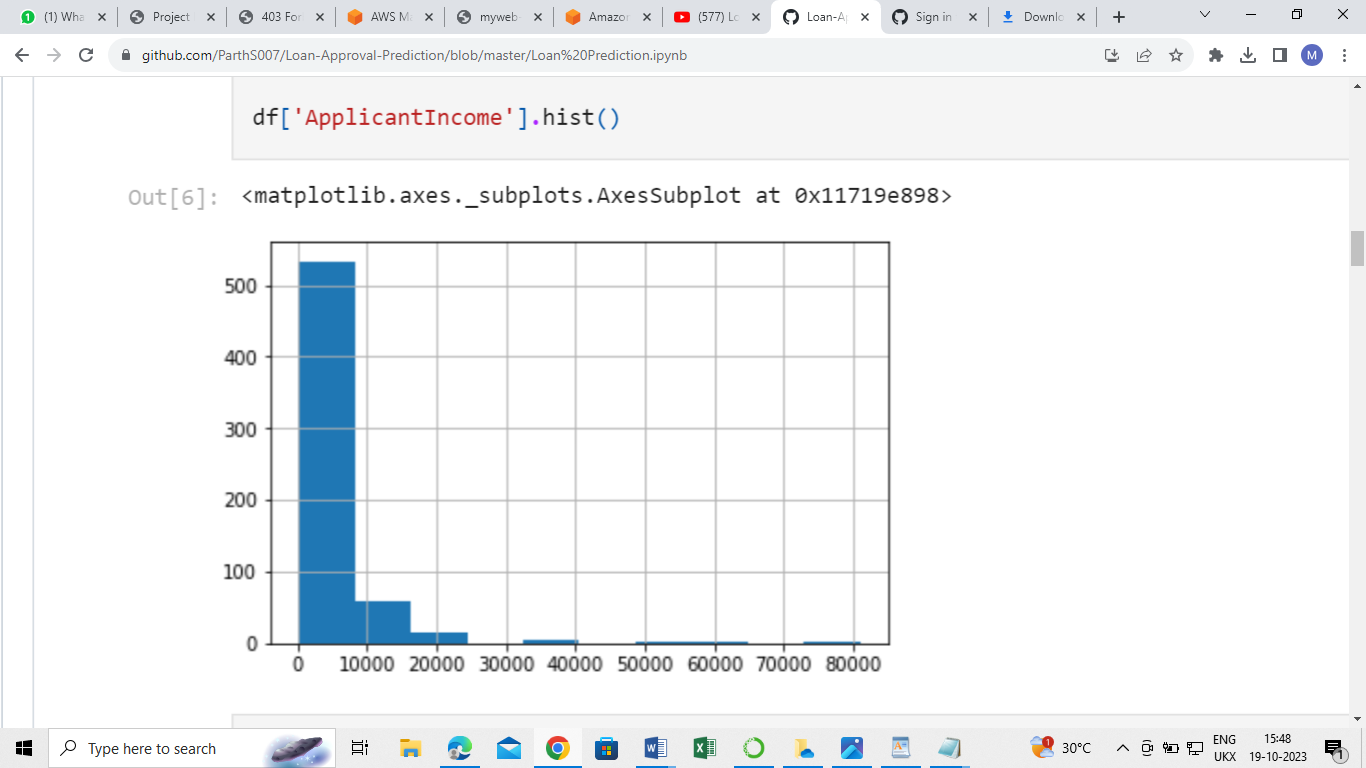
5.FLOWCHART

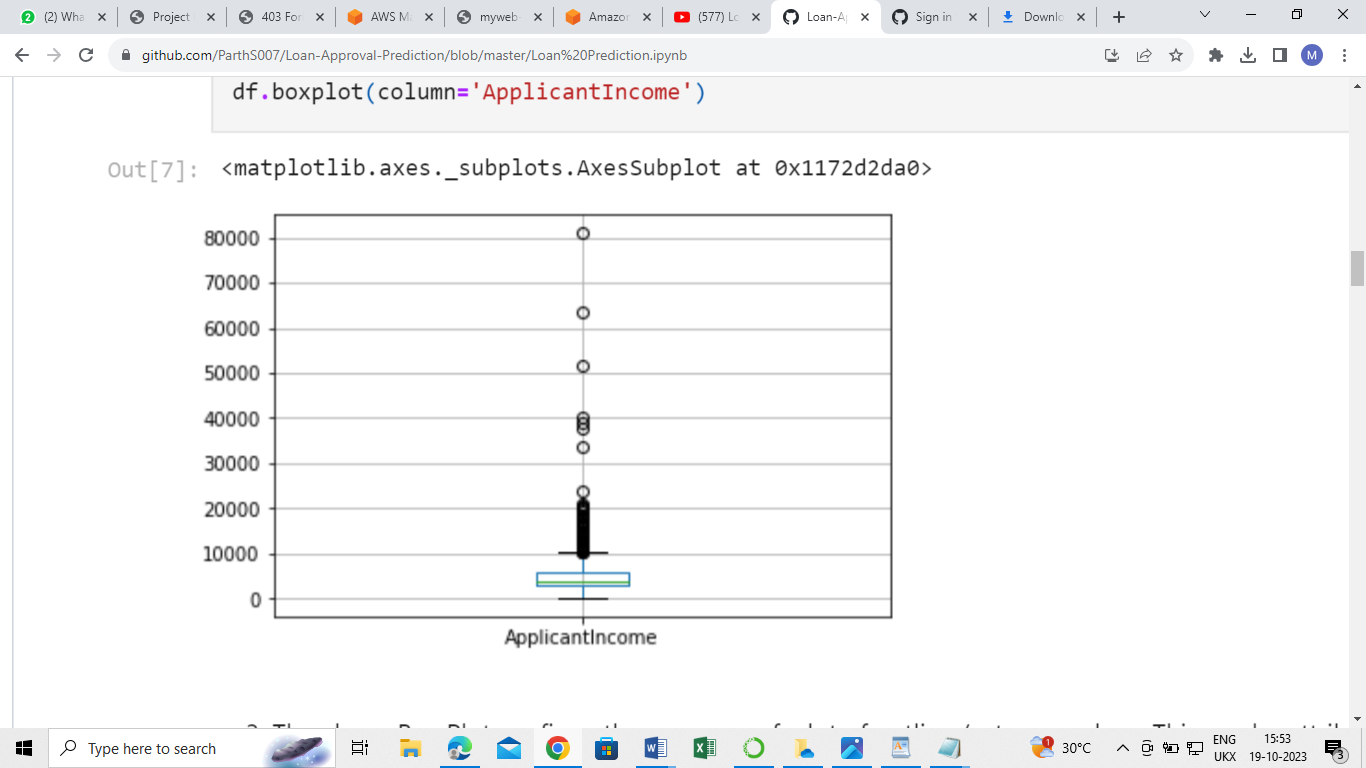


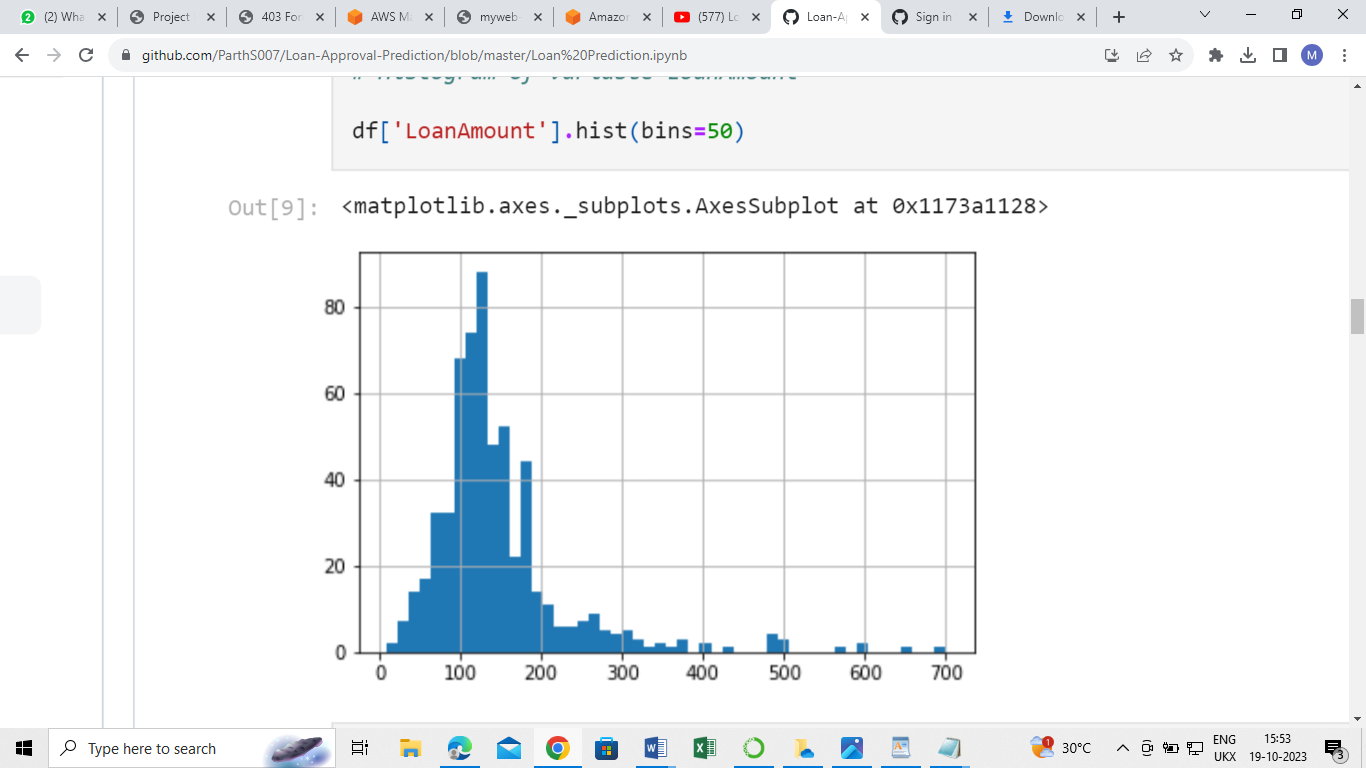
6.RESULT

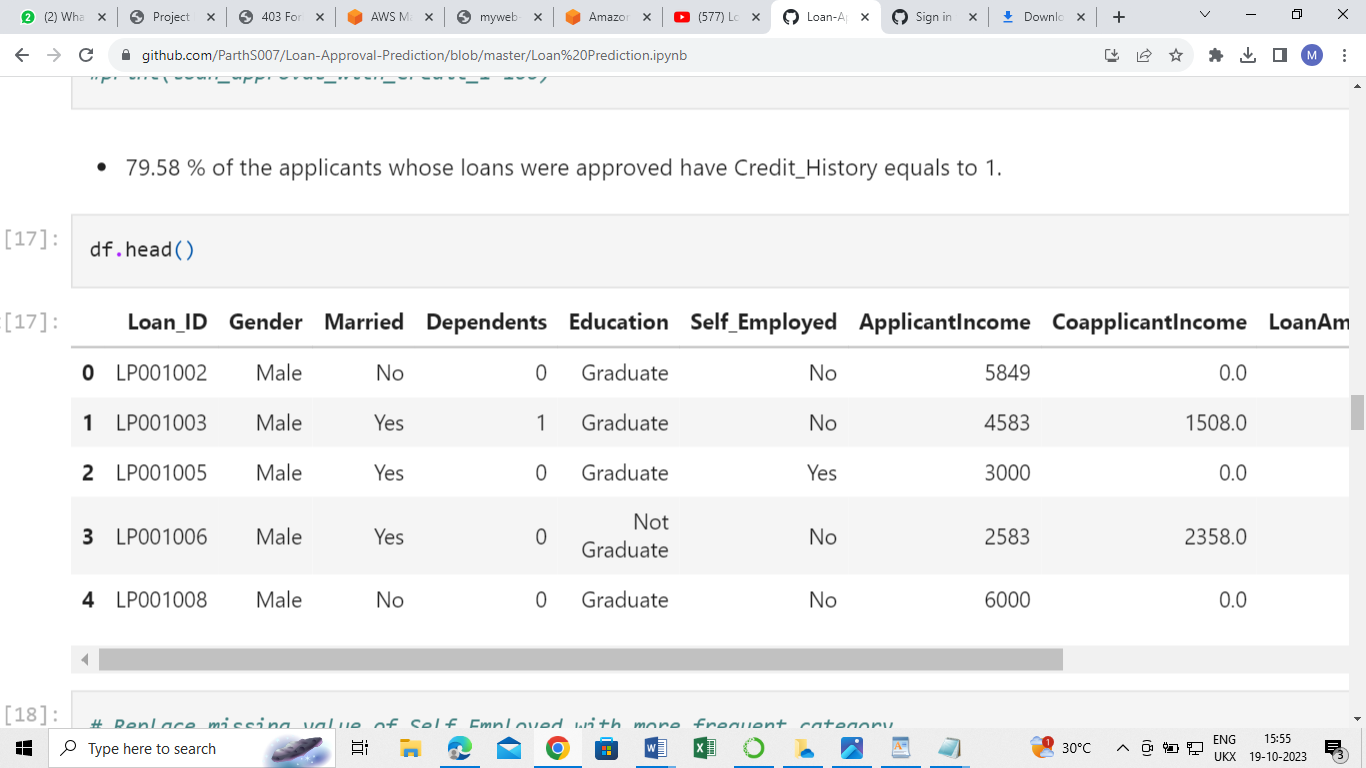


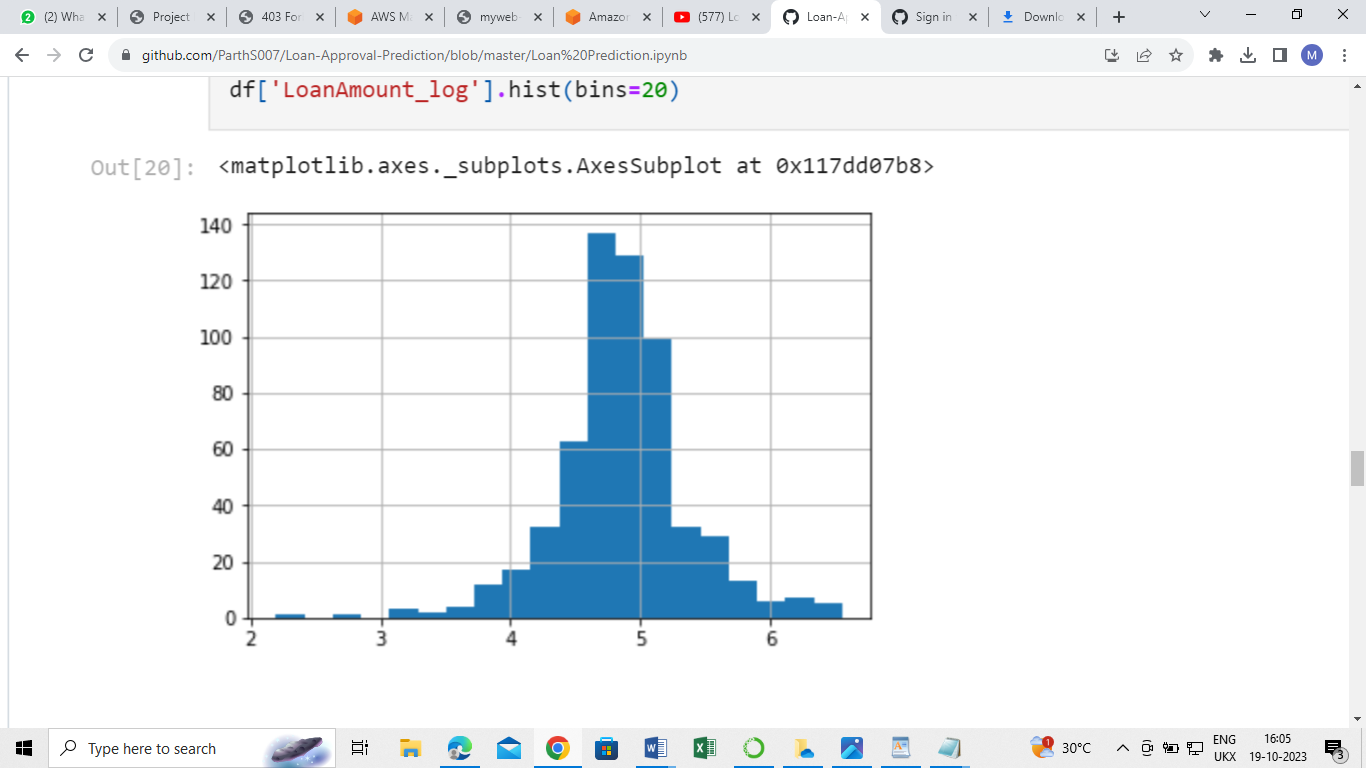


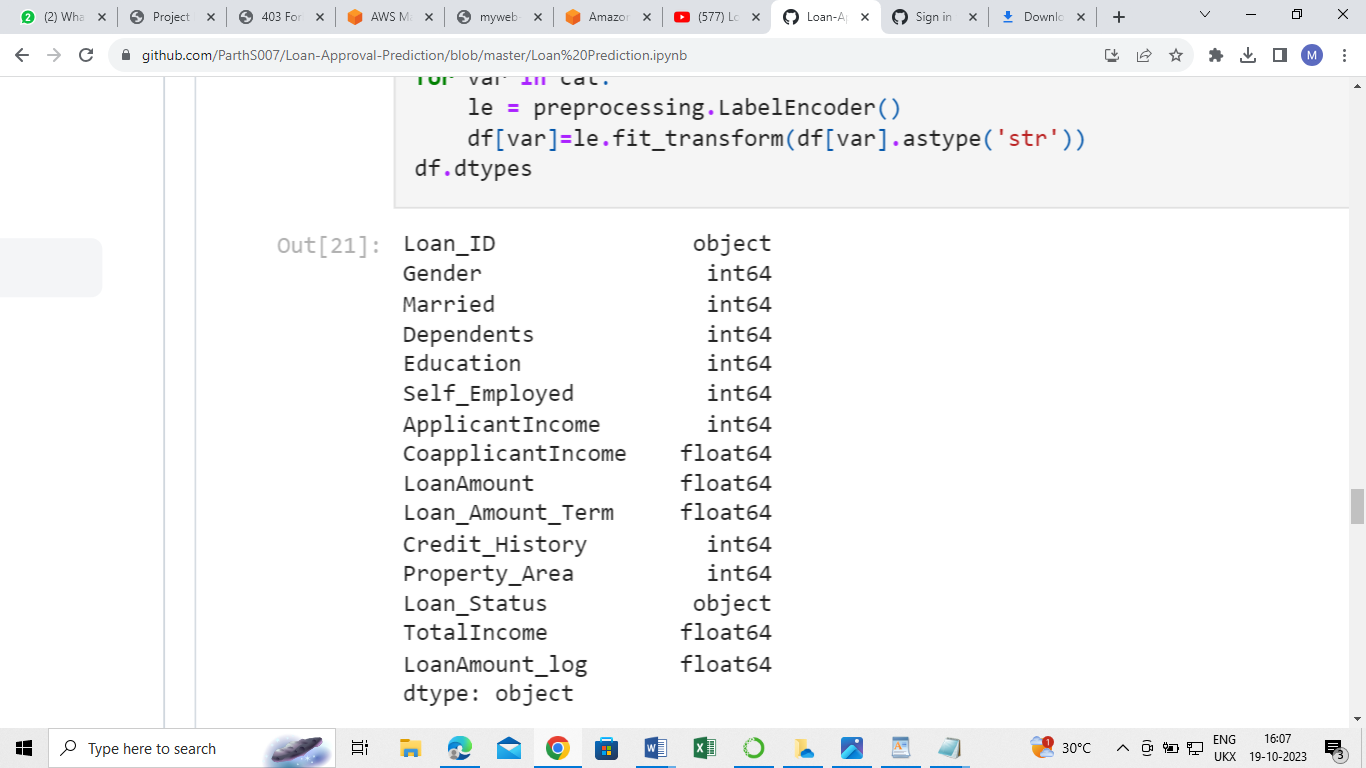


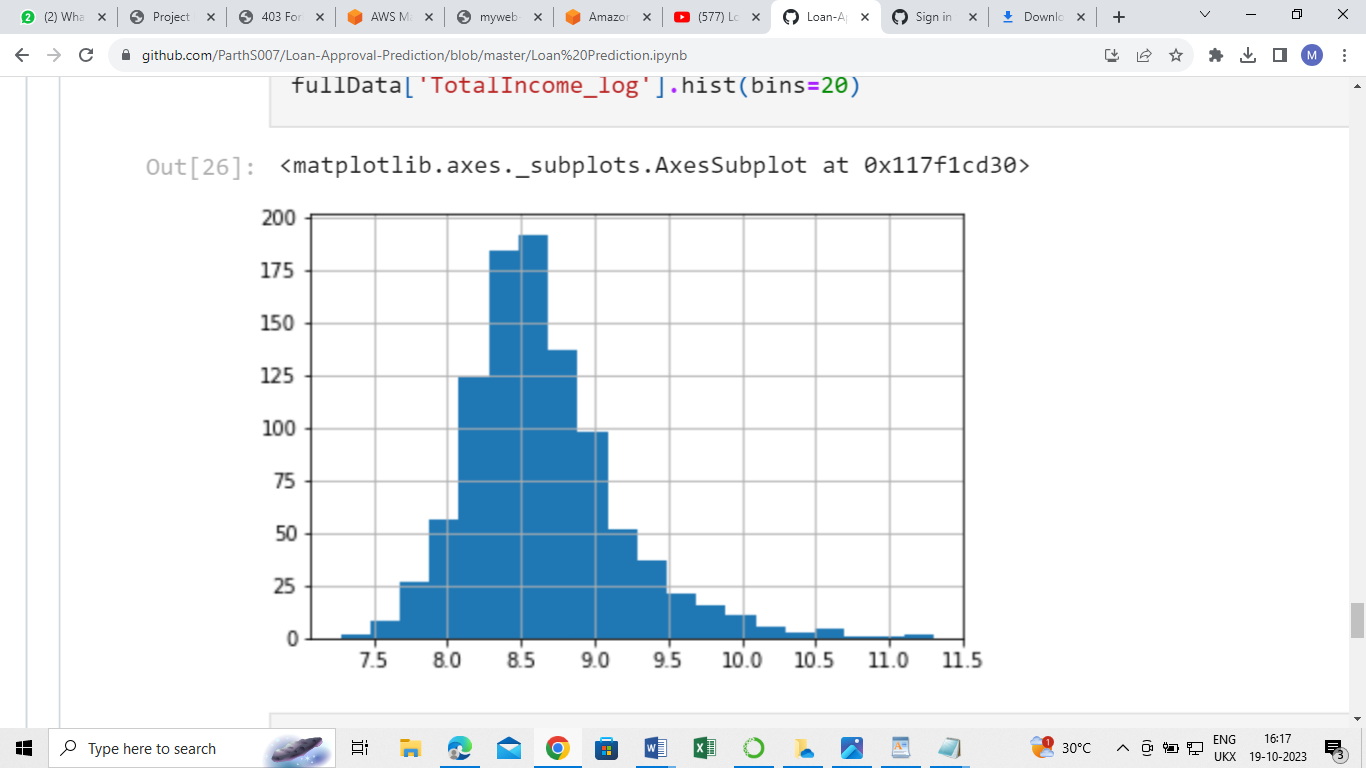












Advantages:

**Improved Efficiency** **:** Automation of the loan approval process leads to faster decision-making, reducing the time applicants have to wait for a response.

**Cost Savings:** By automating the approval process, financial institutions can save costs associated with manual underwriting and processing.

**Consistency:** Automated systems apply consistent criteria to all loan applications, ensuring fairness and reducing the risk of biased decision-making.

**Data-Driven Decisions:** Utilizing historical data and advanced algorithms enable more accurate and data-driven loan approval decisions.

**Risk Mitigation:** Advanced models can assess credit risk more accurately, reducing the chances of approving high-risk loans and minimizing financial losses.Scalability: Automated systems can handle a large volume of loan applications, scaling to meet the demands of growing businesses without significant increases in operational costs.

**Customer Experience**: Faster approval processes enhance customer satisfaction, leading to a positive perception of the financial institution.

**Compliance:** Ensures that loan approval decisions align with regulatory requirements, reducing the risk of legal issues related to discriminatory or unfair lending practices

Disadvantages:

**Data Quality:** The accuracy of predictions heavily depends on the quality and completeness of the historical data. Inaccurate or biased data can lead to flawed predictions.

**Over-reliance on Algorithms:** Relying solely on algorithms may overlook unique or exceptional cases that a human underwriter might consider, leading to potential misapprovals or rejections.

**Algorithmic Bias:** If historical data used for training the model contains biases, the algorithm can perpetuate those biases, leading to discriminatory lending practices.

**Lack of Human Judgment:** Automated systems lack the nuanced judgment and empathy that human underwriters might apply to special cases, such as applicants facing unexpected financial challenges.

**Regulatory Challenges:** Keeping up with evolving regulations and ensuring that the automated system complies with legal requirements can be challenging.

**Security Concerns:** Storing and processing sensitive financial data in automated systems requires robust security measures to prevent data breaches.

**Complexity:** Implementing and maintaining advanced machine learning models can be complex and may require specialized knowledge and skills.

**Customer Trust:** Some customers might be uncomfortable with the idea of automated systems making decisions about their financial future, leading to trust issues.

8.APPLICATIONS

**Traditional Banking**: Banks can use loan approval prediction to automate the loan application process, making it faster and more efficient for personal, business, or mortgage loans.

**Peer-to-Peer (P2P) Lending Platforms:** Online lending platforms can employ these solutions to assess borrower risk and automate the lending process, facilitating peer-to-peer loans.

**Microfinance Institutions**: Microfinance organizations can use loan approval prediction to evaluate small loan applications quickly, enabling financial inclusion for low-income individuals and small businesses.

**Online Lenders and FinTech Companies:** Online lenders and FinTech startups can streamline the loan approval process, making it hassle-free and convenient for borrowers.

**Credit Unions:** Credit unions can leverage these solutions to assess member loan applications efficiently and ensure fair lending practices within their communities.

9.CONCLUSION

For the purpose of predicting the loan approval status ofthe applied customer, we have chosen the machinelearning approach to study the bank dataset. We haveapplied various machine learning algorithms to decidewhich one will be the best for applying on the dataset toget the result with the highest accuracy. Following thisapproach, we found that apart from the logistic regression,the rest of the algorithms performed satisfactory in termsof giving out the accuracy. The accuracy range of the rest of the algorithms were from 75% to 85%. Whereas thelogistic regression gave us the best possible accuracy(88.70%) after the comparative study of all the algorithms.We also determined the most important features thatinfluence the loan approval status. These most importantfeatures are then used on some selected algorithms andtheir performance accuracy is compared with the instanceof using all the features. This model can help the banks infiguring out which factors are important for the loanapproval procedure. The comparative study makes usclear about which algorithm will be the best and ignoresthe rest, based on their accuracy

10.FUTURE SCOPE

**Real-time Data Integration**: Incorporate real-time data streams and integrate alternative data sources (such as social media activity, online behavior, or rental payment history) to enhance the accuracy of creditworthiness assessment.

**Advanced AI Algorithms:** Utilize advanced machine learning algorithms like deep learning and ensemble methods for more accurate predictions, especially in handling complex, nonlinear relationships within data.

**Explainable AI:** Enhance the interpretability of machine learning models to provide transparent and understandable explanations for loan approval decisions, ensuring regulatory compliance and building trust with customers.

**Fairness and Bias Mitigation:** Implement techniques to identify and mitigate biases in data and algorithms, ensuring fairness in lending practices and avoiding discriminatory outcomes.

**Dynamic Threshold Adjustment:** Develop algorithms that can dynamically adjust approval thresholds based on changing market conditions, economic factors, and risk tolerance levels of the lending institution.

11.BIBILOGRAPHY

[1] Kumar Arun, Garg Ishan, Kaur Sanmeet, May-Jun. 2016. Loan Approval Prediction based on

Machine Learning Approach, IOSR Journal of Computer Engineering (IOSR-JCE)

[2] Wei Li, Shuai Ding, Yi Chen, and Shanlin Yang, Heterogeneous Ensemble for Default Prediction of

Peer-to-Peer Lending in China, Key Laboratory of Process Optimization and Intelligent Decision-

Making, Ministry of Education, Hefei University of Technology, Hefei 2009, China

[3] Short-term prediction of Mortgage default using ensembled machine learning models, Jesse

C.Sealand on july 20, 2018.

[4] Clustering Loan Applicants based on Risk Percentage using

K-Means Clustering Techniques

APPENDIX

**import** pandas **as** pd

**import** numpy **as** np

**from** sklearn **import** preprocessing

**from** sklearn.preprocessing **import** LabelEncoder

df**.**head(10)

df\_length **=**len(df)

test\_col **=** len(test**.**columns)

df**.**describe()

df['Property\_Area']**.**value\_counts()

df['ApplicantIncome']**.**hist()

df**.**boxplot(column**=**'ApplicantIncome')

df**.**boxplot(column**=**'ApplicantIncome', by **=** 'Education')

df['LoanAmount']**.**hist(bins**=**50)

df**.**boxplot(column**=**'LoanAmount')

df**.**boxplot(column**=**'LoanAmount', by **=** 'Gender')

loan\_approval **=** df['Loan\_Status']**.**value\_counts()['Y']

print(loan\_approval)

pd**.**crosstab(df ['Credit\_History'], df ['Loan\_Status'], margins**=True**)

**def** percentageConvert(ser):

**return** ser**/**float(ser[**-**1])

df**.**head()

df['Self\_Employed']**.**fillna('No',inplace**=True**)

df['TotalIncome'] **=** df['ApplicantIncome'] **+** df['CoapplicantIncome']

df['LoanAmount']**.**hist(bins**=**20)

df['LoanAmount\_log'] **=** np**.**log(df['LoanAmount'])

df['LoanAmount\_log']**.**hist(bins**=**20)

df['Gender']**.**fillna(df['Gender']**.**mode()[0],inplace**=True**)

df['Married']**.**fillna(df['Married']**.**mode()[0],inplace**=True**)

df['Dependents']**.**fillna(df['Dependents']**.**mode()[0],inplace**=True**)

df['Credit\_History']**.**fillna(df['Credit\_History']**.**mode()[0],inplace**=True**)

cat**=**['Gender','Married','Dependents','Education','Self\_Employed','Credit\_History','Property\_Area']

**for** var **in** cat:

le **=** preprocessing**.**LabelEncoder()

df[var]**=**le**.**fit\_transform(df[var]**.**astype('str'))

df**.**dtypes

df['Type']**=**'Train'

test['Type']**=**'Test'

fullData **=** pd**.**concat([df,test], axis**=**0)

fullData**.**isnull()**.**sum()

ID\_col **=** ['Loan\_ID']

target\_col **=** ["Loan\_Status"]

cat\_cols **=** ['Credit\_History','Dependents','Gender','Married','Education','Property\_Area','Self\_Employed']

fullData['LoanAmount']**.**fillna(fullData['LoanAmount']**.**mean(), inplace**=True**)

fullData['LoanAmount\_log']**.**fillna(fullData['LoanAmount\_log']**.**mean(), inplace**=True**)

fullData['Loan\_Amount\_Term']**.**fillna(fullData['Loan\_Amount\_Term']**.**mean(), inplace**=True**)

fullData['ApplicantIncome']**.**fillna(fullData['ApplicantIncome']**.**mean(), inplace**=True**)

fullData['CoapplicantIncome']**.**fillna(fullData['CoapplicantIncome']**.**mean(), inplace**=True**)

fullData['Gender']**.**fillna(fullData['Gender']**.**mode()[0], inplace**=True**)

fullData['Married']**.**fillna(fullData['Married']**.**mode()[0], inplace**=True**)

fullData['Dependents']**.**fillna(fullData['Dependents']**.**mode()[0], inplace**=True**)

fullData['Loan\_Amount\_Term']**.**fillna(fullData['Loan\_Amount\_Term']**.**mode()[0], inplace**=True**)

fullData['Credit\_History']**.**fillna(fullData['Credit\_History']**.**mode()[0], inplace**=True**)

fullData['TotalIncome']**=**fullData['ApplicantIncome'] **+** fullData['CoapplicantIncome']

fullData['TotalIncome\_log'] **=** np**.**log(fullData['TotalIncome'])

fullData['TotalIncome\_log']**.**hist(bins**=**20)