CUSTOMER LOAN ELIGIBILITY PREDICTION AND VISUALIZATION USING MACHINE LEARNING AND DATA ANALYTICS

A PROJECT REPORT

Submitted by

SRINIVASAN S	**513419104043**
VASANTHAVASAN G	**513419104053**
KARTHI KEYAN KALATHI S	**513419104701**

in partial fulfillment for the award of the degree

of

BACHELOR OF ENGINEERING

IN

COMPUTER SCIENCE & ENGINEERING

UNIVERSITY COLLEGE OF ENGINEERING KANCHEEPURAM

(A Constituent College of Anna University, Chennai)

ANNA UNIVERSITY : CHENNAI – 600 025

MAY 2023

ANNA UNIVERSITY:CHENNAI – 600 025

BONAFIDE CERTIFICATE

Certified that this project report titled "**CUSTOMER LOAN ELIGIBILITY PREDICTION AND VISUALIZATION USING MACHINE LEARNING AND DATA ANALYTICS**" is the bonafide work of "**SRINIVASAN S(513419104043), VASANTHAVASAN G(513419104053) AND KARTHI KEYAN KALATHI S(513419104701)**" of Computer Science & Engineering whose carried out this project work under my supervision.

SIGNATURE
Dr. SELVABHUVANESHWARI,
M.E., Ph.D.,
HEAD OF THE DEPARTMENT,
Assistant Professor,
Department of CSE,
University College of Engineering,
Kancheepuram – 631 552.

SIGNATURE
Dr. SELVABHUVANESHWARI,
M.E., Ph.D.,
SUPERVISOR,
Assistant Professor,
Department of CSE,
University College of Engineering,
Kancheepuram – 631 552.

Submitted for the Project Viva Voce held on : ……………

INTERNAL EXAMINER

EXTERNAL EXAMINER

ACKNOWLEDGEMENT

We are always thankful to our college Dean **Dr. V. KAVITHA, M.E, Ph.D., PROFESSOR,** who endorsed us throughout this project.

Our heartfelt thanks to HOD **Dr. K. SELVABHUVANESHWARI, M.E, Ph.D., ASSISTANT PROFESSOR,** Department of Computer Science and Engineering, University College of Engineering, Kancheepuram, for the prompt and limitless help in providing the excellent infrastructure to do the project and to prepare the thesis.

We express our deep sense of gratitude to our guide **Dr.K.SELVABHUVANESHWARI, M.E,Ph.D. ASSISTANT PROFESSOR, HEAD OF THE DEPARTMENT, DEPARTMENT OF COMPUTER SCIENCE & ENGINEERING,** for her invaluable support and guidance and encouragement for successful completion of this project. Her vision and spirit will always inspire and enlighten us.

We express our sincere thanks to the project committee members **Mrs.T.KALA, M.E, ASSISTANT PROFESSOR,** and **Mr.G.MANI M.E, ASSISTANT PROFESSOR** Department of Computer Science and Engineering, Kancheepuram, for their invaluable guidance and technical Support.

We thank all the faculty members of the Department of Computer Science and Engineering for their valuable help and guidance. We are grateful to our family and friends for their constant support and encouragement. We thank the almighty whose showers of blessings made this project a reality.

SRINIVASAN S	**VASANTHAVASAN G**	**KARTHI KEYAN KALATHI S**
(513419104043)	**(513419104053)**	**(513419104701)**

ABSTRACT

Banks are making major part of profits through loan, even if many people are looking for loans. It's challenging to choose a genuine applicant who will return the loan. Many errors may occur when selecting the real applicant when the process is done manually so machine learning model is used. The technology that is being suggested in this project leverages data analytics to create a spectacular dashboard and machine learning to anticipate the loan.

A loan prediction system and dashboard visualization based on machine learning will be created to select qualified loan applicants automatically. The manual selection process is prone to errors, making it challenging for banks to choose genuine applicants who will repay the loan. system will benefit both the bankstaff and the applicant, with a significant reduction in loan sanctioning time. To build the model, a massive amount of data will be collected and trained using machine learning algorithms. The data set consists of 12 columns, each detailing a person who applies for a loan. The data will be split into training and testing portions, with the model trained on the former and itsperformance evaluated on the latter.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLE

LIST OF ABBREVATIONS

ACRONYM	ABBREVATIONS
IEEE	INSTITUTE OF ELECTRICAL AND ELECTRONICS ENGINEERS
CNN	CONVOLUTIONAL NEURAL NETWORK
GUI	GRAPHICAL USER INTERFACE
IDE	INTEGRATED DEVELOPMENT ENVIRONMENT
XGBOOST	EXTREME GRADIENT BOOSTING
IBM	INTERNATIONAL BUSINESS MACHINES CORPORATION
EDA	EXPLORATORY DATA ANALYSIS
ML	MACHINE LEARNING
AI	ARTIFICIAL INTELLIGENCE
P2P	PEER-TO-PEER
ROC	RECEIVER OPERATING CHARACTERISTIC

CHAPTER 1

INTRODUCTION

1.1 OBJECTIVE OF THE PROJECT

Loans are one of the most important sources of revenue in the banking industry. Banks provide several forms of loans to different groups of people based on their income or the type of application. Customers repay the loan amount in installments, together with interest. Credit interest will be included in the bank's profit. Banks also defecate on loan defaulters. Banks occasionally reach loss zones as a result of defaults. According to statistics, a large number of banks have been closed due to large losses caused by loan recovery failure. One of the most important aspects in lowering the default rate is a thorough examination of the applicant's situation while processing the loan. Prediction of loan defaulters can assist the bank in reducing nonperforming assets.

The loan prediction system will also incorporate a data analytics visual dashboard, enabling bank employees to monitor the performance of the model and track important metrics in real-time. The dashboard will offer an intuitive and user-friendly interface with a range of interactive data visualizations, including pie charts, bar graphs, and scatterplots, providing valuable insights into the loan approval process. By leveraging the power of data analytics, bank employees will be able to gain a deeper understanding of the loan approval process, identify trends and patterns, and make more informed decisions. The dashboard will also provide alerts and notifications, enabling bank employees to respond promptly to critical situations and take corrective action. With the integration of the data analytics visual dashboard, the loan prediction system promises to be a game-changer for the banking industry, revolutionizing the way loans are approved and managed.

1.2 MOTIVATION

Due to the multi-billion dollar loan sanctioning and credit scoring industries, processing loan applications is a difficult and time-consuming task for any banking institution. As a result, many people are afraid to approach a bank to request a loan, even though many qualified customers may be turned down due to inadequate facilities. Enhance the customer experience by expediting the loan approval process and lowering the risk of default, the bank may increase the client experience and satisfaction. The motivation behind bank loan prediction is to automate the loan approval process, eliminate the high dependency on manual intervention, extensive domain knowledge, and human bias in loan approval prediction

1.3 SCOPE OF THE PROJECT

The scope of a bank loan prediction and visualization project is to automate the loan approval process, eliminate the high dependency on manual intervention, extensive domain knowledge, and human bias in loan approval prediction. The project aims to provide a quick, immediate, and easy way to choose deserving applicants and provide special advantages to the bank. The project involves exploring loan prediction as a data science and machine learning problem and building a system/application for loan prediction using machine learning and data analytics.

The Projects scope is to build predictive models to automate the process of targeting the right applicants. The primary objective is to find a classification solution that will be as accurate as possible at predicting whether a borrower is likely to repay the loan or not. Data analytics is the process of examining large and complex data sets to extract valuable insights and knowledge that can be used to inform decision-making. This involves using a range of statistical and computational techniques to identify patterns, trends, and relationships within the data.

1.4 LIMITATIONS OF THE PROJECT

Data quality The accuracy and reliability of the predictions generated by the model are highly dependent on the quality of the input data. If the data is incomplete or contains errors, the predictions may be less accurate.

Changing market conditions Market conditions can change quickly, and prior data may not be a fair predictor of future trends. To guarantee that the predictive model remains accurate and relevant, it must be re-evaluated and updated on a regular basis.

Limited scope A predictive model's scope is limited because it can only generate predictions based on the data that is available. If important loan default factors are not captured in the data, the model may be unable to make reliable forecasts.

IBM Account To view an interactive dashboard visualization, an IBM account must be running in the background.

1.5 ORGANIZATION OF THE REPORT

This report helps to understand the problem of Loan approval status among the people and also the solution which is provided through in this system and the technologies that are used for the development of the project and the work which are yet to perform in future for the further enhancement of the project along with the impact which is going to create by the proposed system called CUSTOMER LOAN ELIGIBILITY PREDICTION AND VISUALIZATION USING MACHINE LEARNING AND DATA ANALYTICS

CHAPTER 2

PRELIMINARIES

2.1 SECURITY

Security is an important aspect to consider in a bank loan prediction project, as sensitive data related to loan applications and outcomes will be involved. It is critical to preserve and keep confidential personal and financial information about loan applicants. Personal information is kept as safe as possible and is not shared with third parties outside of the prediction handling system. The project will involve transferring data between different systems and applications. It is important to ensure that the network used for this purpose is secure, with appropriate firewalls and other security measures in place.

Regular security audits: It is critical to examine the security of the loan prediction system on a regular basis in order to discover and resolve any weaknesses. This can be accomplished by frequent penetration testing and other security evaluations.

Security is an important aspect to consider when designing and deploying an IBM Cognos dashboard, as it may contain sensitive business information. Authentication and Authorization: It ensure that only authorized users have access to the dashboard. This can be achieved through authentication mechanisms such as username/password, single sign-on (SSO), or third-party authentication providers. Authorization rules can be used to control what users can see and do within the dashboard. Regular auditing and monitoring of the dashboard can help detect and prevent security breaches.

2.1.1 FEATURES OF LOAN ELIGIBILITY PREDICTION AND VISUALIZATION

- The dashboard should have visually appealing charts, graphs, and other visualizations that present complex data in an easy-to-understand format.

- The dashboard should allow users to interact with the data by using filters and slicers to view different subsets of the data.

- The dashboard should have the ability to display real-time data updates, so users can see the latest information at a glance.

- system objective is to develop a machine learning model to classify with the highest degree of accuracy possible. Work with all algorithm choose highest accuracy algorithm for prediction.

- The project involves exploratory data analysis to identify patterns and trends in the data that can be used to predict loan eligibility

- Responsive web application is developed with prediction and data analytics technique.

- The individual can verify whether he/she is eligible for a loan by filling out the following attributes in a fraction of a second without visiting a bank.

- IBM Cognos Analytics integrates reporting, modeling, analysis, dashboards, stories, and event management to help users and employee of bank to understand their organization’s data

2.2 DOMAIN

2.2.1 MACHINE LEARNING TECHNIQUE

2.2.1.1 WHAT IS MACHINE LEARNING?

Machine learning (ML) is a field of artificial intelligence (AI) that provides machines the ability to automatically learn and improve from experience without being explicitly programmed. Machine learning algorithms use historical data as input to predict new output values. The goal of machine learning is to develop algorithms that can learn from data and make predictions or decisions based on that data.

2.2.1.2 HOW IT WORKS

Machine learning algorithms build a model based on sample data, known as training data, to make predictions or decisions without being explicitly programmed to do so. The machine learning model is trained on a dataset that contains input features and output labels. The model learns the relationship between the input features and output labels and uses this relationship to make predictions on new data. In supervised machine learning, the model is trained on a known set of input data and known responses to the data (output) and trains a model to generate reasonable predictions for the response to new data. The machine learning models require a high quantity of reliable data to perform accurate predictions. The accuracy of the prediction model depends on the quality of the data used to train the model. The machine learning algorithms are used in a wide variety of applications, such as in medicine, email filtering, speech recognition, and predictive analytics

2.2.1.3 DATA ANALYTICS TECHNIQUE

2.2.1.3.1 WHAT IS DATA ANALYTICS?

Data analytics is the process of examining large and complex data sets to extract valuable insights and knowledge that can be used to inform decision-making. This involves using a range of statistical and computational techniques to identify patterns, trends, and relationships within the data. There are several stages in the data analytics process, including data collection, data cleaning and preprocessing, data analysis, and data visualization. Each of these stages requires specialized skills and tools, such as programming languages, statistical software, and visualization platforms.

2.2.1.3.2 HOW IT WORKS

IBM Analytics provides a suite of tools and solutions for collecting, organizing, analyzing, and visualizing data to help organizations make data-driven decisions. IBM Analytics includes IBM Business Analytics, which provides a single point of entry for business planning and analysis for faster data-driven decisions. IBM Planning Analytics with Watson is an AI-driven self-service analytics tool that empowers business users to predict and shape future outcomes IBM Cognos Analytics is a powerful business intelligence and data analytics platform designed for businesses of all sizes. It allows users to easily access and analyze data from a variety of sources, including databases, spreadsheets, and cloud-based applications. One of the key features of IBM Cognos Analytics is its ability to provide a unified view of an organization's data, allowing users to quickly identify trends and patterns across multiple data sources. The platform also includes advanced data visualization capabilities, which allow users to create interactive dashboards, story and report

2.2.2 ALGORITHMS OF MACHINE LEARNING

2.2.2.1 Logistic Regression

1. **Data Preparation** Logistic regression requires a labeled dataset, where each observation has a set of features and a binary label indicating the class to which it belongs.

2. **Model Creation** The logistic regression model estimates the probability of the positive class as a function of the input features. The model creates a decision boundary by optimizing a cost function.

3. **Cost Function** The cost function used in logistic regression is the cross-entropy loss function, which compares the predicted probability to the true label and adjusts the weights accordingly.

4. **Optimization** The model is optimized by minimizing the cost function using an iterative algorithm like gradient descent or stochastic gradient descent.

5. **Evaluation** The model's performance is evaluated using metrics like accuracy, precision, recall, and F1 score.

6. **Prediction** The logistic regression model predicts the probability of the positive class for a new observation, and then applies a threshold to convert this probability to a binary label.

7. **Regularization** Regularization techniques like L1 or L2 can be used to prevent overfitting in the logistic regression model.

8. **Hyperparameter Tuning** The performance of the logistic regression model can be improved by tuning hyperparameters like the learning rate, regularization strength, or number of iterations.

2.2.2.2 Decision Tree

1. **Data Preparation** Decision tree requires a labeled dataset, where each observation has a set of features and a binary label indicating the class to which it belongs.

2. **Model Creation** The decision tree model creates a tree-like structure to classify observations by recursively splitting them based on their features. The model selects the feature that best separates the data at each node and creates a split based on a threshold value.

3. **Splitting Criteria** The decision tree model selects the best feature to split the data by evaluating different criteria such as Gini impurity or information gain. These criteria evaluate the homogeneity of the classes in each split.

4. **Stopping Criteria** The decision tree model stops splitting the data when a stopping criterion is met. These criteria can be a maximum depth of the tree, a minimum number of observations at each leaf node, or a maximum impurity reduction.

5. **Pruning** Pruning techniques like reduced error pruning or cost complexity pruning can be used to avoid overfitting in the decision tree model.

6. **Evaluation** The model's performance is evaluated using metrics like accuracy, precision, recall, and F1 score.

7. **Prediction** The decision tree model predicts the class of a new observation by traversing the tree from the root to the appropriate leaf node.

8. **Ensemble Methods** Ensemble methods like random forests or gradient boosting can be used to improve the performance of decision trees by combining multiple trees.

2.2.2.3 Random Forest

1. **Data Preparation** Random forest requires a labeled dataset, where each observation has a set of features and a binary label indicating the class to which it belongs.

2. **Model Creation** The random forest model creates an ensemble of decision trees to classify observations. Each tree is trained on a random subset of the data and a random subset of the features.

3. **Bootstrap Aggregation (Bagging)** The random forest model uses a technique called bagging to create multiple subsets of the data by sampling with replacement. Each subset is used to train a decision tree model.

4. **Feature Randomness** The random forest model also uses feature randomness by selecting a random subset of features at each node of the decision tree.

5. **Decision Tree Creation** Each decision tree in the random forest is created using the decision tree algorithm with a splitting criterion like Gini impurity or information gain.

6. **Ensemble Voting** The random forest model predicts the class of a new observation by taking the majority vote of the decision trees.

7. **Out-of-Bag Evaluation** The random forest model evaluates each decision tree on data that was not included in its training subset to estimate the generalization error of the ensemble.

8. **Hyperparameter Tuning** The performance of the random forest model can be improved by tuning hyperparameters like the number of trees, the maximum depth of the trees, or the number of features used at each node.Top of Form

2.2.2.4 XGBOOST

1. **Data Preparation** XGBoost requires a labeled dataset, where each observation has a set of features and a binary label indicating the class to which it belongs.

2. **Model Creation** The XGBoost model creates an ensemble of decision trees to classify observations. Each tree is trained on a weighted subset of the data and the errors from previous trees.

3. **Gradient Boosting** The XGBoost model uses a technique called gradient boosting to iteratively train decision trees. At each iteration, a new decision tree is trained to fit the errors of the previous trees.

4. **Regularization** XGBoost uses regularization techniques like L1 or L2 regularization to prevent overfitting in the model.

5. **Objective Function** The XGBoost model uses a custom objective function to evaluate the quality of the splits in the decision trees. The objective function includes a term for the loss and a term for the regularization penalty.

6. **Feature Importance** XGBoost calculates the importance of each feature by measuring how much each feature contributes to the reduction of the objective function.

7. **Cross-Validation** The XGBoost model can use k-fold cross-validation to estimate the generalization error of the model.

8. **Hyperparameter Tuning** The performance of the XGBoost model can be improved by tuning hyperparameters like the learning rate, the maximum depth of the trees, or the regularization strength.

2.3 ARCHITECTURE

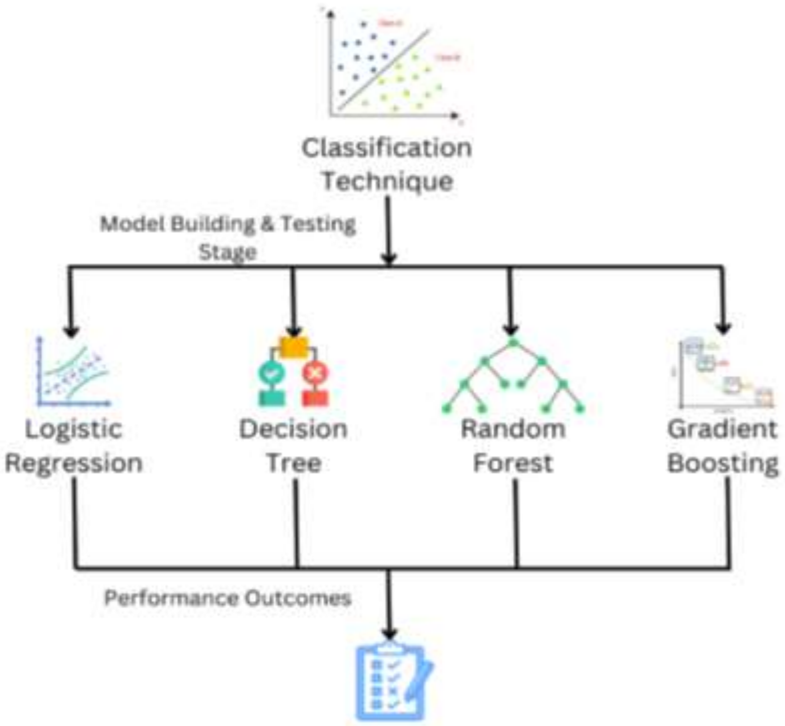


Fig.2.1 Architecture

2.4 APPLICATIONS

- **Credit risk assessment** Bank loan prediction can be used to evaluate potential borrowers’ creditworthiness. Banks can use loan prediction models to assess the chance of a borrower defaulting on a loan by analyzing the borrower’s financial history, credit score, and other pertinent information.

- **Loan approval process** Loan prediction models can help banks speed the loan approval process. Banks can minimize the time and resources required to evaluate loan applications by automating the loan approval process, while simultaneously enhancing the quality and consistency of loan decisions.

- **Fraud detection** Loan prediction models can also be used to identify potential cases of fraud. By analyzing borrower information and comparing it against historical data, banks can identify unusual patterns or inconsistencies that may indicate fraudulent activity.

- **Loan portfolio management** Loan prediction models can also be used to manage loan portfolios. Banks can discover trends and patterns in loan performance over time, which can be utilized to optimize resource allocation and risk management.

CHAPTER 3

LITERATURE SURVEY

3.1 MODEL-AGNOSTIC COUNTERFACTUAL EXPLANATIONS IN CREDIT SCORING

YEAR **:** 2022

AUTHORS **:** XOLANI DASTILE, TURGAY CELIK, HANS VANDIERENDONCK

PUBLISHER: IEEE

CONCEPT

The non-transparent nature of machine learning and deep learning models hampers the application of these models in credit scoring. Address this challenge of non-transparency by generating counterfactuals via a custom genetic algorithm to explain model predictions. Select predictive features and determine Spearman's rank correlations between the target flag and the predictors to enforce sparseness. normalize all continuous features to expedite the generation of counterfactuals. This show the efficacy of the proposed approach on German and HMEQ credit scoring datasets. The experimental results indicate that the proposed approach efficiently generates sparse counterfactuals compared to similar methods**.**

DRAWBACKS

Although the proposed method produces satisfactory results with the default parameter settings of the genetic algorithm, optimal parameter settings may improve the performance of the counterfactual generation. Furthermore, an optimal fitness function capturing different properties of counterfactuals using a genetic programming approach can enhance the performance of the proposed method. In addition, explaining the overall working mechanism of the black-box models instead of explaining individual instances can further improve the transparency and explainability of the credit scoring models

PERFORMANCE MEASURE TABLE

Algorithms	**Accuracy**
Neural network	**80%**
XGBoost	**77%**
Logistic regression	**76.6%**

Table 3.1 Performance Value

3.2 INTERNET FINANCIAL CREDIT SCORING MODELS BASED ON DEEP FOREST AND RESAMPLING METHODS

YEAR **:** 2023

AUTHORS **:** YU ZHONG, HUILING WANG AND ET

PUBLISHER: IEEE

CONCEPT

With the rapid development of Internet finance, preventing and controlling credit risks has become an important problem for the Internet finance industry, and credit scoring provides an effective tool to solve this problem. However, most of the existing Internet financial credit scoring models are based on DNNs, whose structure is difficult to design. Moreover, most of these models do not consider the class imbalance problem on classification performance. This work introduces a new deep learning method DF to credit scoring modeling and employs resampling methods to solve the imbalanced class problem. The experiment was carried out on four Internet financial credit datasets. Four evolution measures, including Recall, AUC, F–measure and G–mean where selected to evaluate the classification performance of the model.

DRAWBACKS

The limitation of this study is that it only focuses on the classification performance of credit scoring models. How to determine the loan amount of the good credit customs is the direction of future research. Besides, loan applicants' social network information has value for the Internet financial credit-scoring model [48], so building a deep learning Internet credit evaluation model based on social network information data is another future research direction

3.3 AN ONLINE TRANSFER LEARNING FRAMEWORK WITH EXTREME LEARNING MACHINE FOR AUTOMATED CREDIT SCORING

YEAR : 2022

AUTHORS : RANA ALASBAHI , XIAOLIN ZHENG and Et al

PUBLISHER: IEEE

CONCEPT

Automated Credit Scoring (ACS) is the process of predicting user credit based on historical data. It involves analyzing and predicting the association between the data and particular credit values based on similar data. Recently, ACS has been handled as a machine learning problem, and numerous models were developed to address it.

DRAWBACKS

The work is to extend the developed algorithm to include optimization of the random weights between the input-hidden layer algorithm and to incorporate dynamic feature selection.

3.4 PREDICTING DEFAULT RISK ON PEER-TO-PEER LENDING IMBALANCED DATASETS

YEAR **:**2022

AUTHORS **:** YEN-RU CHEN1 , JENQ-SHIOU LEU , SHENG-AN HUANG , JUI-TANG WANG AND JUN-ICHI TAKADA

PUBLISHER: IEEE

CONCEPT

In the past few years, Peer-to-Peer lending (P2P lending) has grown rapidly in the world. The main idea of P2P lending is disintermediation and removing the intermediaries like banks. For a small business and some individuals without enough credit or credit history, P2P lending is a good way to apply for a loan. However, the fundamental problem of P2P lending is information asymmetry in this model, which may not correctly estimate the default risk of lending. Lenders only determine whether or not to fund the loan by the information provided by borrowers, causing P2P lending data to be imbalanced datasets which contain unequal fully paid and default loans. Imbalanced datasets are quite common in the real worlds, such as credit card fraud in transactions, bad products in the plant and so on. Unfortunately, the imbalanced data are unfriendly to the normal machine learning schemes.

DRAWBACKS

Peer-to-peer (P2P) lending is a solution to lend money without involving financial institutions and allows borrowers to connect to lenders directly. However, P2P lending has a fundamental problem because its dataset is imbalanced. Therefore, it makes the classifiers are prone to majority class rather than minority class. In this study, we employ a various machine learning algorithm to predict the default risk of P2P lending, use re-sampling and cost-sensitive mechanisms to processing imbalanced datasets. Get the dataset from Lending Club to validate proposed scheme. In the experiment results, random under-sampling shows the best performance in different classifiers. Then after doing preprocessing and feature selection, the proposed scheme can effectively raise the prediction accuracy for default risk.

PERFORMANCE MEASURES TABLE

RANDOM UNDER SAMPLING	**ACCURACY**	**RECALL**	**F1**	**G-MEAN**
Random forest	63.931	60.881	42.923	62.812
Neural Networks	63.559	66.463	44.830	64.567
Logistic regression	63.236	66.146	44.494	64.247

Table 3.2 Performance Value

CHAPTER-4

PROPOSED ARCHITECTURE

4.1 PROPOSED ARCHITECTURE

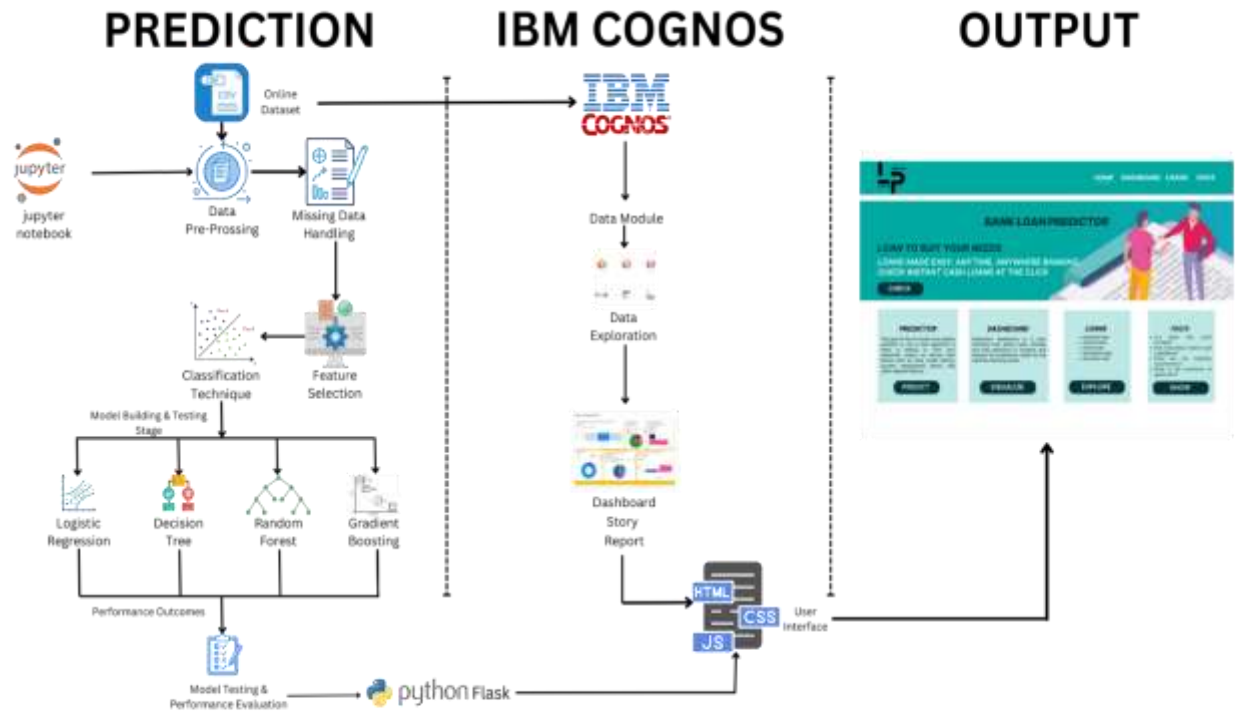


Fig.4.1 Proposed Architecture

4.1.1 PREDICTION MODEL PHASE

The Proposed Architecture shows the flow of different modules of phases involved. Firstly, the required dataset is collected from the standard website called Kaggle. The Dataset is then given as input to Jupyter notebook for data preprocessing which is well known as Exploratory Data Analysis (EDA). After handling the missing data's in the dataset by assigning the mean and median values to the missing data in the dataset. The dataset is then splitted into Categorical and Numerical data to extract the feature variables. The Feature variable is then splitted into test and train data for Classification Techniques (Supervised Learning).

Based on the highest Accuracy the prediction model algorithm is selected for further process. The Trained model is them created and deployed using Python Flask for creating Prediction model using HTML, CSS and JAVASCRIPT.

4.1.2 IBM COGNOS DATA ANALYTICS PHASE

In this phase, the collected dataset is given as input to cognos analytics after cleaning the data(missing values). This Business Intelligence tool is capable of producing stunning Dashboards, Report and Story based on the User requirement. From the dataset can provide varieties of graph together to produce Stunning Dashboard. The Dashboard is an interactive one by showing all relevant details on other graph if click on one graph. The story is an type which will create the scenes of graphs along the transitions and animations. The report is an detailed document in which can use graphs and tables to display information

4.1.3 WEBSITE FINAL PHASE

In this Phase, the Final expected Website is created using joining both the Prediction page as well as Cognos Analytics using the Bootstrap Framework to the required final expected output.

CHAPTER-5
IMPLEMENTATION
MODULES

5.1 MODULES DESCRIPTION AND DIAGRAM

The project's proposed system has divided primarily into 3 modules

5.1.1 PREDICTION MODEL TRAINING MODULE

5.1.2 IBM COGNOS MODULE

5.1.3 USER INTERFACE MODULE

5.1.1 PREDICTION MODEL MODULE

The Proposed Architecture shows the flow of different modules of phases involved. Firstly, the required dataset is collected from the standard website called Kaggle. The Dataset is then given as input to Jupyter notebook for data preprocessing which is well known as Exploratory Data Analysis (EDA). After handling the missing data's in the dataset by assigning the mean and median values to the missing data in the dataset. The dataset is then splitted into Categorical and Numerical data to extract the feature variables. The Feature variable is then splitted into test and train data for Classification Techniques (Supervised Learning). Based on the highest Accuracy the prediction model algorithm is selected for further process. The Trained model is them created and deployed using Python Flask for creating Prediction model using HTML, CSS and JAVASCRIPT.

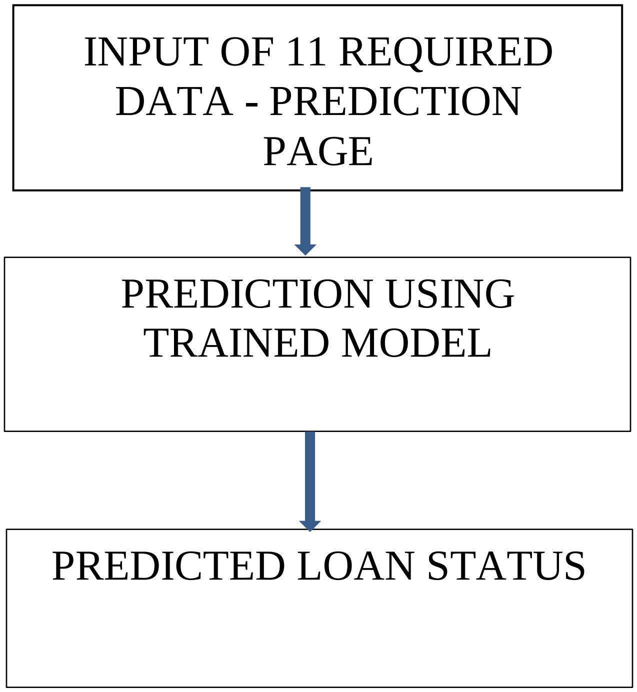

INPUT OF 11 REQUIRED DATA - PREDICTION PAGE
PREDICTION USING TRAINED MODEL
PREDICTED LOAN STATUS

Fig.5.1 Prediction model steps

5.1.2 IBM COGNOS MODULE

In this phase, the collected dataset is given as input to cognos analytics after cleaning the data(missing values). This Business Intelligence tool is capable of producing stunning Dashboards, Report and Story based on the User requirement. From the dataset can provide varieties of graph together to produce Stunning Dashboard. The Dashboard is an interactive one by showing all relevant details on other graph if click on one graph. The story is an type which will create the scenes of graphs along the transitions and animations. The report is an detailed document in which can use graphs and tables to display information

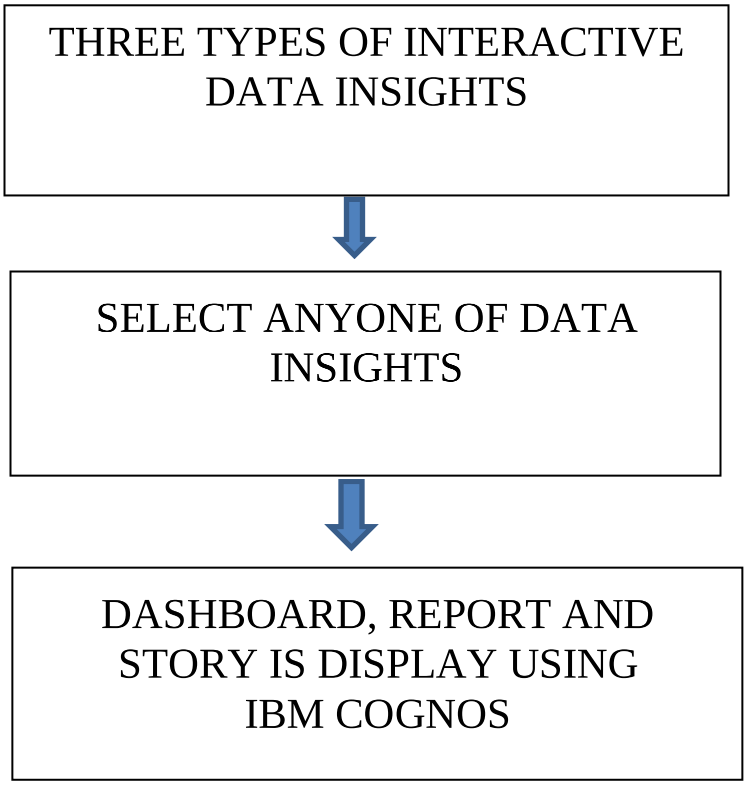

THREE TYPES OF INTERACTIVE DATA INSIGHTS
SELECT ANYONE OF DATA INSIGHTS
DASHBOARD, REPORT AND STORY IS DISPLAY USING IBM COGNOS

Fig.5.2 IBM Cognos module

5.1.3 USER INTERFACE MODULE

User interface module that interacts the user and Prediction model. The user interface design done by Bootstrap Framework. Bootstrap is a free, open source front-end Development framework for the creation of websites and web apps.The website will be beneficial for both the user as well as the banking sector. The main aim of project is to cut short the direct contact of customer with the banking representative for their loan eligibility. In banking POV the website will be useful by providing the beautiful insights to their superior as well as the customer to know about their datasets

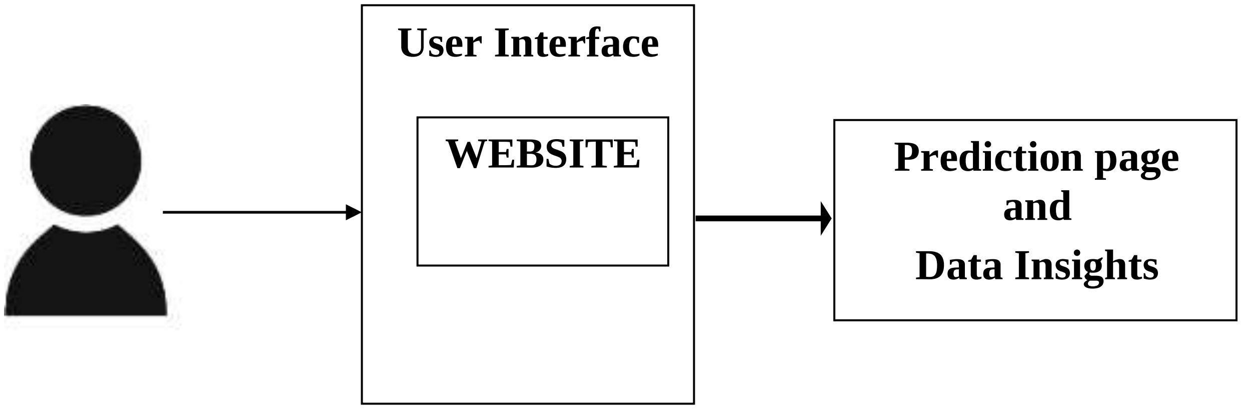


Fig.5.3 User Interface Module

5.2 ALGORITHM

5.2.1 DECISION TREE ALGORITHM

Decision tree is a type of supervised learning algorithm that is mostly used in classification problems. Decision Trees are a non-parametric algorithm that can handle both numerical and categorical variables. They can capture complex interactions between variables and are easy to interpret. They can be used for both binary and multi-class classification problems. Decision Tree is a Supervised learning technique that can be used for both classification and Regression problems, but mostly it is preferred for solving Classification problems. It is a tree-structured classifier, where internal nodes represent the features of a dataset, branches represent the decision rules and each leaf node represents the outcome.

The decisions or the test are performed on thesis of features of the given dataset. It is a graphical representation for getting all the possible solutions to a problem/decision based on given conditions.

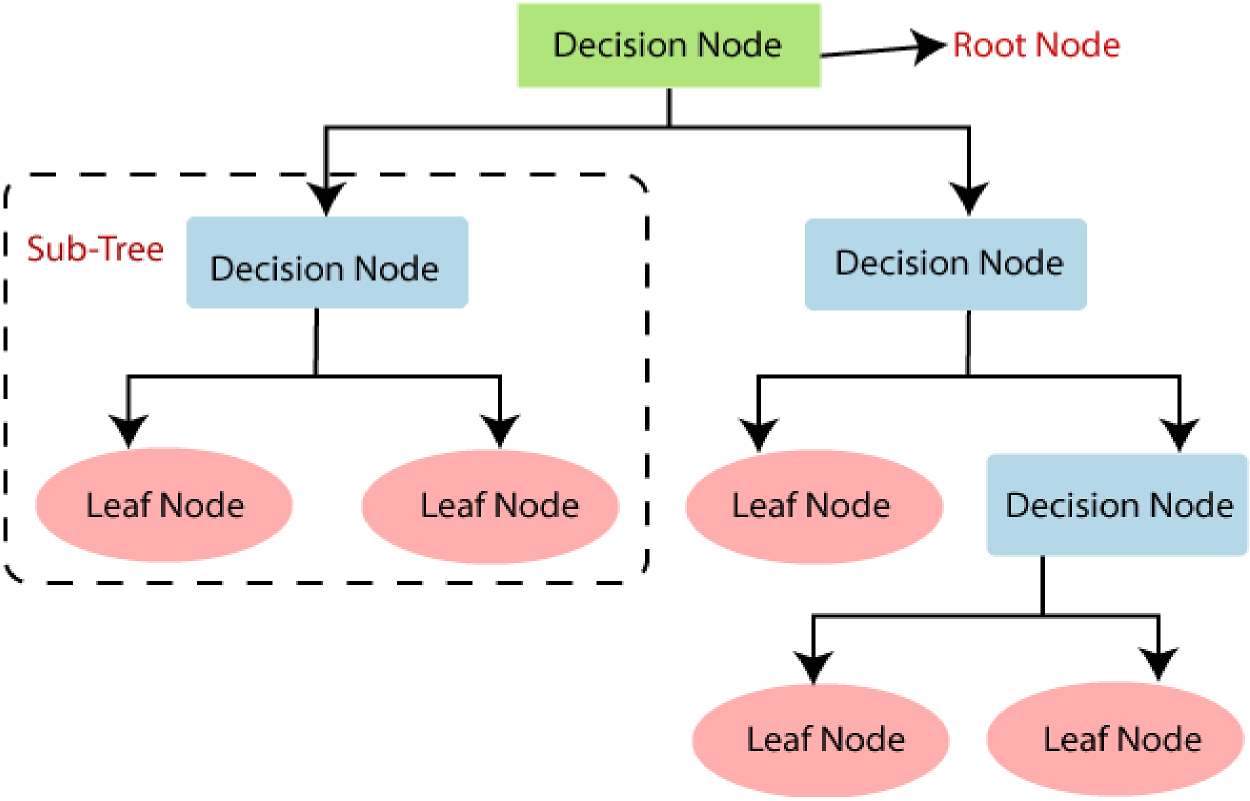


Fig.5.4 Algorithm diagram

5.2.1.1METHODOLOGY

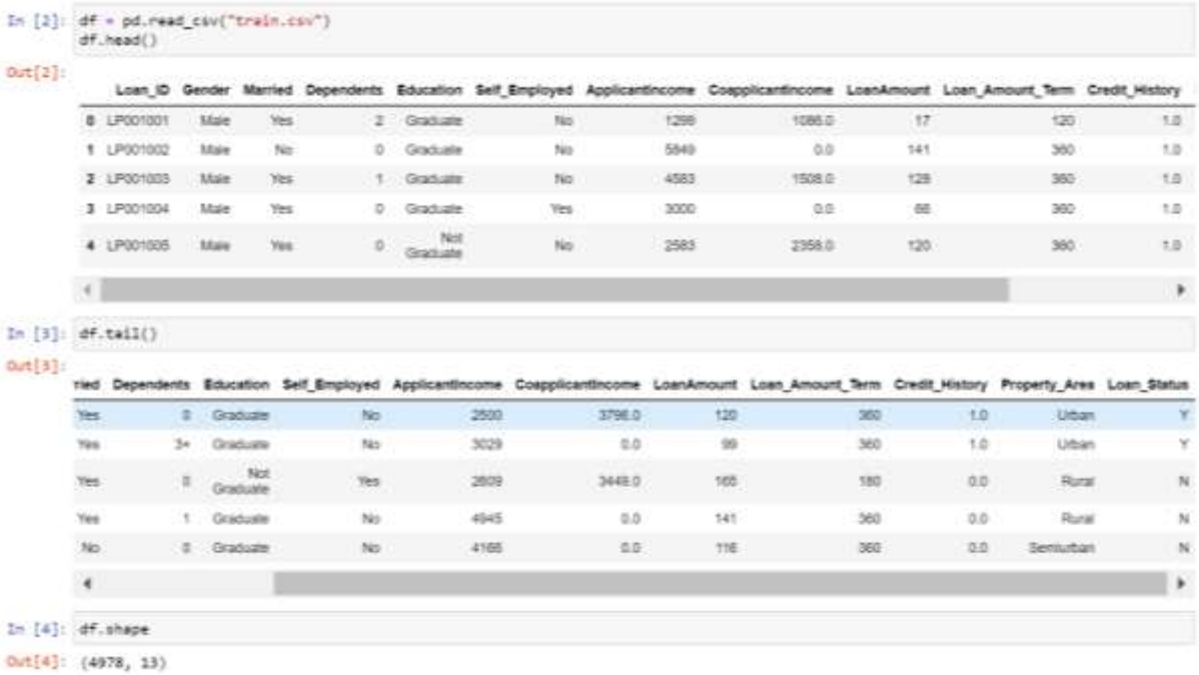
The methodology of the proposed project has three major steps that involves dataset collection, data preprocessing, features extraction, classification, model training, and data analytics

5.2.2 DATASET COLLECTION

Data collection is the process of gathering and measuring information on variables of interest, in an established systematic fashion that enables one to answer stated research questions, test hypotheses, and evaluate outcomes. The data collection component of research is common to all fields of study including physical and social sciences, humanities, business, etc. While methods vary by discipline, the emphasis on ensuring accurate and honest collection remains the same. Using the platform called KAGGLE for dataset collection

5.3 Dataset

Bank customers data set is collected from public repository for implementation. The dataset having 614 records with 13 attributes. Loan_ID, gender, marital_status, dependants, education, self_employed, income of applicant and co_applicant, loan_amount, number of terms, credit_history, asset information and loan_status. It is labeled data, customer loan eligibility outcome Yes/No. Yes, means customer eligible to get the loan and No means customer not eligible for loan sanction.



```
In [2]: df = pd.read_csv("train.csv")
        df.head()
Out[2]:
```

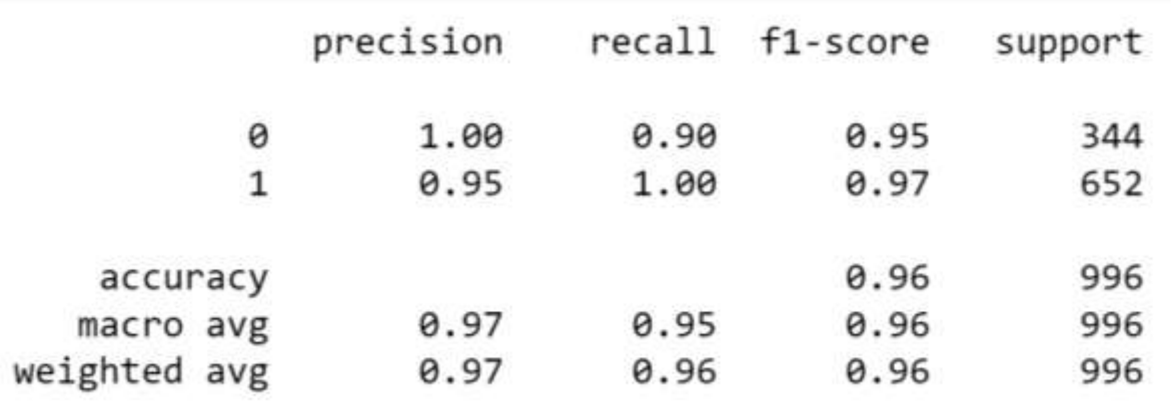
	Loan_ID	Gender	Married	Dependents	Education	Self_Employed	ApplicantIncome	CoapplicantIncome	LoanAmount	Loan_Amount_Term	Credit_History
0	LP001001	Male	Yes	2	Graduate	No	1299	1086.0	17	120	1.0
1	LP001002	Male	No	0	Graduate	No	5849	0.0	141	360	1.0
2	LP001003	Male	Yes	1	Graduate	No	4583	1508.0	128	360	1.0
3	LP001004	Male	Yes	0	Graduate	Yes	3000	0.0	66	360	1.0
4	LP001005	Male	Yes	0	Not Graduate	No	2583	2358.0	120	360	1.0

```
In [3]: df.tail()
Out[3]:
```

ried	Dependents	Education	Self_Employed	ApplicantIncome	CoapplicantIncome	LoanAmount	Loan_Amount_Term	Credit_History	Property_Area	Loan_Status
Yes	0	Graduate	No	2500	3796.0	120	360	1.0	Urban	Y
Yes	3+	Graduate	No	3029	0.0	99	360	1.0	Urban	Y
Yes	0	Not Graduate	Yes	2609	3449.0	165	180	0.0	Rural	N
Yes	1	Graduate	No	4945	0.0	141	360	0.0	Rural	N
No	0	Graduate	No	4166	0.0	116	360	0.0	Semiurban	N

```
In [4]: df.shape
Out[4]: (4978, 13)
```

Fig.5.5 Data set



```
              precision    recall  f1-score   support

           0       1.00      0.90      0.95       344
           1       0.95      1.00      0.97       652

    accuracy                           0.96       996
   macro avg       0.97      0.95      0.96       996
weighted avg       0.97      0.96      0.96       996
```

Fig.5.6 Dataset performance

5.4 DATASET PREPROCESSING

Data preprocessing is a process of preparing the raw data and making it suitable for a machine learning model. It is the first and crucial step while creating a machine learning model. When creating a machine learning project, it is not always a case that come across the clean and formatted data. And while doing any operation with data, it is mandatory to clean it and put in a formatted way. So for this, use data preprocessing task. The dataset is cleaned by assigning the mean and median values to the missing null values or undefined values.

5.4.1 FEATURES EXTRACTION

Feature extraction is a process of dimensionality reduction by which an initial set of raw data is reduced to more manageable groups for processing. A characteristic of these large data sets is a large number of variables that require a lot of computing resources to process. Feature extraction is the name for methods that select and /or combine variables into features, effectively reducing the amount of data that must be processed, while still accurately and completely describing the original data set. In proposed system, the dataset is splitted into Categorical and Numerical variables. The target variable is separated from the given dataset for classification purpose using Dummy Method. The dataset is then further splitted into train and test data by for prediction

5.4.2 CLASSIFICATION

The Processed and extracted variable is the further given as input to the classification . In proposed system using the combination of supervised algorithm . In which the DECISION TREE ALGORITHM has achieved upto 96% accuracy.

5.5 PLATFORM OF IMPLEMENTATION

5.5.1 SOFTWARE REQUIREMENTS

The Software is a set of instructions that are used to command any systems to perform any operations .The software has the advantage to make decisions and to hardware sensible results and is useful in handling complex situations.

OS : Windows 10, Linux

LANGUAGE : Python

BUSINESS TOOL USED: IBM COGNOS ANALYTICS

LIBRARIES : Pandas, Numpy, Matplotlib, Seaborn, Pickle, Python Flask

5.5.1.1 VISUAL STUDIO CODE IDE

Visual Studio Code is a freeware source-code editor made by Microsoft for Windows, Linux and MacOS. Features include support for debugging, syntax highlighting, intelligent code completion, snippets, code refactoring, and embedded Git. Users can change the theme, keyboard shortcuts, preferences, and install extensions that add additional functionality. In the Stack Overflow 2019 Developer Survey, Visual Studio Code was ranked the most popular developer environment tool, with 50.7% of 87,317 respondents reporting that they use it.VS Code can be extended via extensions, available through a central repository. This includes additions to the editor and language support. A notable feature is the ability to create extensions that add support for new languages, themes, and debuggers, perform static code analysis, and add code linters using the Language Server Protocol. Visual Studio Code includes multiple extensions for FTP, allowing the software to be used as a free alternative for web development. Code can be synced between the editor and the server, without downloading any extra software. Visual Studio Code allows users to set the code page in which the active document is saved, the newline character, and the programming language of the active document. This allows it to be used on any platform, in any local, and for any given programming language.

5.5.1.2 ANACONDA

Anaconda distribution comes with 1,500 packages selected from PyPI as well as the conda package and virtual environment manager. It also includes a GUI, Anaconda Navigator, as a graphical alternative to the command line interface(CLI).The big difference between conda and the pip package manager is in how package dependencies are managed, which is a significant challenge for Python data science and the reason conda exists.

When pip installs a package, it automatically installs any dependent Python packages without checking if these conflict with previously installed packages. It will install a package and any of its dependencies regardless of the state of the existing installation. Because of this, a user with a working installation of, for example, Google Tensorflow, can find that it stops working having used pipto install a different package that requires a different version of the dependent numpy library than the one used by Tensorflow. In some cases, the package may appear to work but produce different results in detail.

5.5.1.3 PYTHON (PROGRAMMING LANGUAGE)

Python is an interpreted, high-level and general-purpose programming language. Python's design philosophy emphasizes code readability with its notable use of significant indentation. Its language constructs and object-oriented approach aim to help programmers write clear, logical code for small and large-scale projects. Python is dynamically-typed and garbage-collected.

It supports multiple programming paradigms, including structured (particularly, procedural), object- oriented and functional programming. Python is often described as a "batteries included" language due to its comprehensive standard library. Python was designedto be highly extensible (with modules). This compact modularity has made it particularly popular as a means of adding programmable interfaces to existing applications.Libraries such as NumPy, SciPy and Matplotlib allow the effective use of Python in scientific computing, with specialized libraries such as Biopython and Astropy providing domain-specific functionality. SageMath is a mathematical software with a notebook interface programmable in Python: its library covers many aspects of mathematics, including algebra, combinatorics, numerical mathematics, number theory, and calculus.OpenCV has python bindings with a rich set of features for computer vision and image processing.

Python is commonly used in artificial intelligence projects and machine learning projects with the help of libraries like TensorFlow, Keras, Pytorch and Scikit-learn. As a scripting language with modular architecture, simple syntax and rich text processing tools, Python is often used for natural language processing.

5.5.1.4 PANDAS

Pandas is an open-source library that is made mainly for working with relational or labeled data both easily and intuitively. It provides various data structures and operations for manipulating numerical data and time series. This library is built on top of the NumPy library. Pandas is fast and it has high performance & productivity for users.

5.5.1.5 PICKLE

Python pickle module is used for serializing and de-serializing a Python object structure. Any object in Python can be pickled so that it can be saved on disk. What pickle does is that it “serializes” the object first before writing it to file. Pickling is a way to convert a python object (list, dict, etc.) into a character stream. The idea is that this character stream contains all the information necessary to reconstruct the object in another python script.

5.5.1.6 PYTHON FLASK

Flask Tutorial provides the basic and advanced concepts of the Python Flask framework. Flask tutorial is designed for beginners and professionals.Flask is a web framework that provides libraries to build lightweight web applications in python. It is developed by Armin Ronacher who leads an international group of python enthusiasts (POCCO).Flask is a web framework that provides libraries to build lightweight web applications in python. It is based on WSGI toolkit and jinja2 template engine. Flask is considered as a micro framework.

5.5.1.7 NUMPY

NumPy a library for the Python programming language, adding support for large, multi-dimensional arrays and matrices, along with a large collection of high-level mathematical functions to operate on these arrays. The ancestor of NumPy, Numeric, was originally created by Jim Hugunin with contributions from several other developers. In 2005, Travis Oliphant created NumPy by incorporatingfeatures of the competing Numarray into Numeric, with extensive modifications. NumPy is open-source software and has many contributors. Using NumPy in Python gives functionality comparable to MATLAB since they are bothinterpreted, and they both allow the user to write fast programs as long as most operations work on arrays or matrices instead of scalars.

5.5.1.8 MATPLOTLIB

Matplotlib is a plotting library for the Python programming language and its numerical mathematics extension NumPy. It provides an object-oriented APIfor embedding plots into applications using general-purpose GUI toolkits like Tkinter, wxPython, Qt, or GTK+. There is also a procedural "pylab" interface based on a state machine (like OpenGL), designed to closely resemble that of MATLAB, though its use is discouraged. SciPy makes use of Matplotlib.

5.5.1.9 SEABORN

Seaborn is an amazing visualization library for statistical graphics plotting in Python. It provides beautiful default styles and color palettes to make statistical plots more attractive. It is built on the top of matplotlib library and also closely integrated to the data structures from pandas. Seaborn aims to make visualization the central part of exploring and understanding data. It provides dataset-oriented APIs, so that we can switch between different visual representations for same variables for better understanding of dataset.

5.6 HARDWARE REQUIREMENTS

5.6.1 PERSONAL COMPUTER / LAPTOP

Fig.5.7 Personal computer

The minimum requirement of personal computer or laptop that have Intel i3 or AMD or better processor and virtualization is supported. That have either windowsor linux operating system. Minimum 4GB of RAM required. And storage space needed for this face mask detector is minimum 2GB required.

CHAPTER 6

RESULTS AND PERFORMANCE

6.1 RESULT

In the proposed system developed a responsive website that enables customers and staff to access loan eligibility information and to interact with the loan application process in a seamless and user-friendly manner. This website includes various features and functionalities that allow users to view their loan eligibility status, apply for loans, and track the progress of their applications. This website also includes various security and privacy features to protect customer information and to comply with regulatory requirements. These features include data encryption, secure login, and customer data management tools.

Developed a predictive model for customer loan eligibility using machine learning techniques. Trained and tested model on a dataset consisting of various features related to customers, such as age, income, employment type, credit score, and loan amount. The results show that model achieved an accuracy of 96%. This indicates that model is able to effectively predict whether a customer is eligible for a loan or not.

In addition to the dashboard, In proposed system also created a report and story that provide more detailed information about methodology, dataset, evaluation metrics, and key findings. These resources are designed to provide a comprehensive understanding of analysis and to support further research and decision-making in the field of finance.

Overall, the effectiveness of machine learning techniques in predicting customer loan eligibility, and highlights the importance of specific customer features in this prediction task. These results can be used by financial institutions to develop more accurate and efficient loan approval processes.

6.2 OUTPUT OF AN APPLICATION

6.2.1 LOAN PREDICTION WEBSITE FRONT PAGE

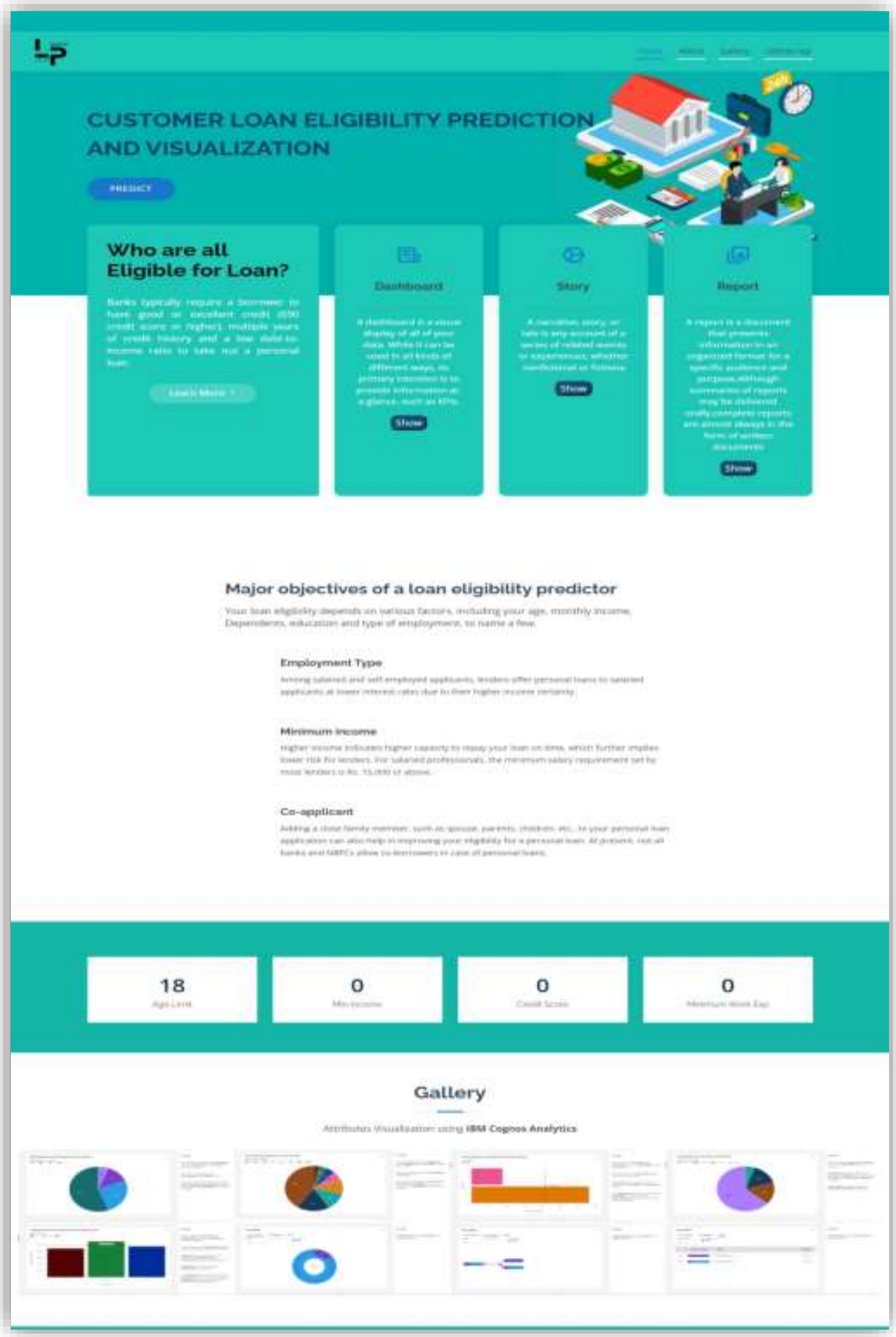
The loan prediction project website is designed to provide information about loan eligibility and display various graphs related to the loan dataset. The website consists of several sections, including a homepage, graphs gallery, and a footer with details about the team members.

The homepage of the website presents an overview of the loan prediction project. It highlights the key concept of loan eligibility and provides a brief explanation of the loan dataset attributes. The homepage also includes a section that lists the individuals who are eligible for a loan based on the predictions made by the model.

The graphs gallery section of the website showcases visual representations of the loan dataset attributes. Each graph represents a specific attribute, such as income, credit score, loan amount, or interest rate. These graphs provide a visual understanding of the dataset and help users analyze the relationships and patterns within the data.

Finally, the footer section of the website displays details about the team members involved in the loan prediction project. This could include information such as their names, roles, areas of expertise, and contact information. The footer serves as a way to give credit to the team members and provide users with additional information about the people behind the project.

Overall, the loan prediction project website offers a user-friendly interface that presents loan eligibility information, visualizes loan dataset attributes through graphs, and acknowledges the team members who contributed to the project.



CUSTOMER LOAN ELIGIBILITY PREDICTION AND VISUALIZATION

Who are all Eligible for Loan?

Banks typically require a borrower to have good or excellent credit (690 credit score or higher), multiple years of credit history and a low debt-to-income ratio to take out a personal loan.

Dashboard

A dashboard is a visual display of all of your data. While it can be used in all kinds of different ways, its primary intention is to provide information at a glance, such as KPIs.

Story

A narrative, story, or tale is any account of a series of related events or experiences, whether nonfictional or fictional.

Report

A report is a document that presents information in an organized format for a specific audience and purpose. Although summaries of reports may be delivered orally, complete reports are almost always in the form of written documents.

Major objectives of a loan eligibility predictor

Your loan eligibility depends on various factors, including your age, monthly income, Dependents, education and type of employment, to name a few.

Employment Type

Among salaried and self employed applicants, lenders offer personal loans to salaried applicants at lower interest rates due to their higher income certainty.

Minimum Income

Higher income indicates higher capacity to repay your loan on time, which further implies lower risk for lenders. For salaried professionals, the minimum salary requirement set by most lenders is Rs. 15,000 or above.

Co-applicant

Adding a close family member, such as spouse, parents, children, etc., to your personal loan application can also help in improving your eligibility for a personal loan. At present, not all banks and NBFCs allow co-borrowers in case of personal loans.

18	0	0	0
Age Limit	Min Income	Credit Score	Minimum Work Exp

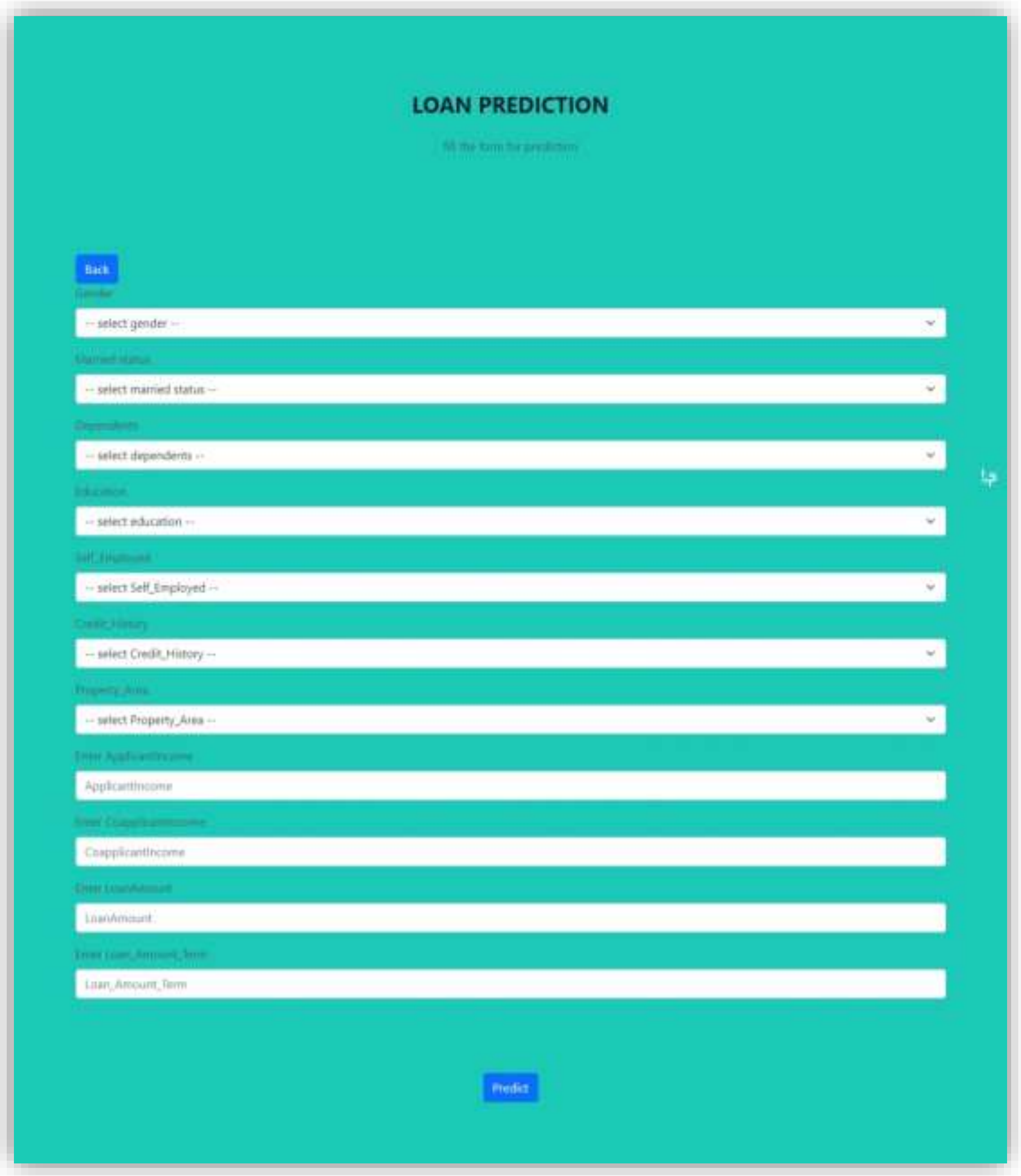
Gallery

Attributes Visualization using **IBM Cognos Analytics**

Fig.6.1 Loan Prediction Website Front Page

6.2.2 PREDICTION PAGE

On this page, users can input their relevant details, such as income, credit score, loan amount, and other necessary information. Once the user submits the form, the loan prediction model is applied to the input data, and a prediction is generated. The prediction indicates whether the user is likely to be approved for the loan or not.



LOAN PREDICTION

fill the form for prediction

Gender

-- select gender --

Married status

-- select married status --

Dependents

-- select dependents --

Education

-- select education --

Self_Employed

-- select Self_Employed --

Credit_History

-- select Credit_History --

Property_Area

-- select Property_Area --

Enter ApplicantIncome

ApplicantIncome

Enter CoapplicantIncome

CoapplicantIncome

Enter LoanAmount

LoanAmount

Enter Loan_Amount_Term

Loan_Amount_Term

Fig 6.2 Prediction Page

6.2.3 INTERACTIVE DASHBOARD

The interactive dashboard created using IBM Cognos provides a dynamic and visual representation of the loan prediction project. It offers users a consolidated view of key metrics, trends, and insights derived from the loan dataset. The dashboard allows users to interact with the data, explore different visualizations, and gain a comprehensive understanding of loan-related patterns and factors. With its intuitive interface and interactive features, the dashboard facilitates data-driven decision-making and supports in-depth analysis of loan prediction

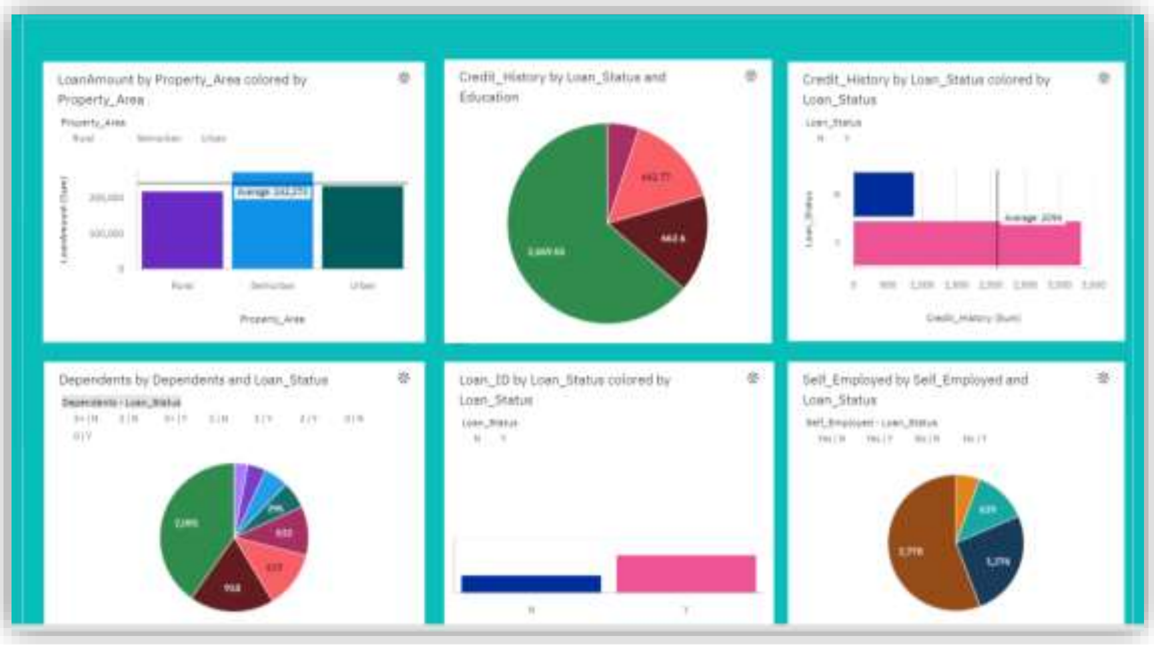

LoanAmount by Property_Area colored by Property_Area
Property_Area
Rural
Semiurban
Urban
LoanAmount (Sum)
200,000
100,000
0
Average
Property_Area
Credit_History by Loan_Status and Education
Credit_History by Loan_Status colored by Loan_Status
Loan_Status
N
Y
Average
Credit_History (Sum)
Dependents by Dependents and Loan_Status
Dependents · Loan_Status
Loan_ID by Loan_Status colored by Loan_Status
Loan_Status
N
Y
Self_Employed by Self_Employed and Loan_Status
Self_Employed · Loan_Status

Fig.6.3 Dashboard

6.2.4 STORY

The story created using IBM Cognos is a narrative-driven presentation of the loan prediction project. It combines data visualizations, text, and multimedia elements to tell a cohesive and engaging story about loan eligibility and prediction. The story takes the audience through the project's background, data exploration, model development, and validation. It highlights notable discoveries, challenges encountered, and the overall impact of the loan prediction project. The story is designed to captivate and educate the audience, conveying the project's objectives, process, and outcomes in a compelling manner.

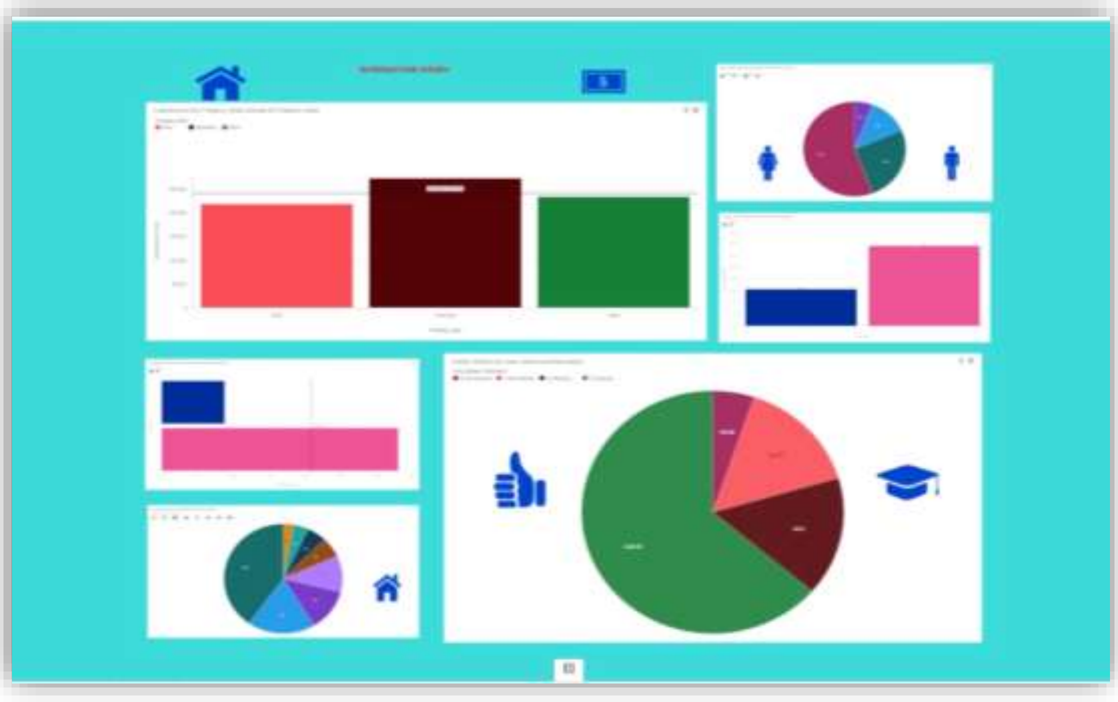


Fig 6.4 Story

6.2.5 REPORT

The report generated using IBM Cognos presents a comprehensive analysis of the loan prediction project. It includes detailed information on the loan dataset, model performance metrics, and key findings. The report highlights important insights such as the significant predictors for loan approval, the accuracy of the loan prediction model, and any observed trends or patterns in the data. Through well-organized sections and visual representations, the report provides stakeholders with a clear overview of the loan prediction project, its methodology, and the outcomes achieved.

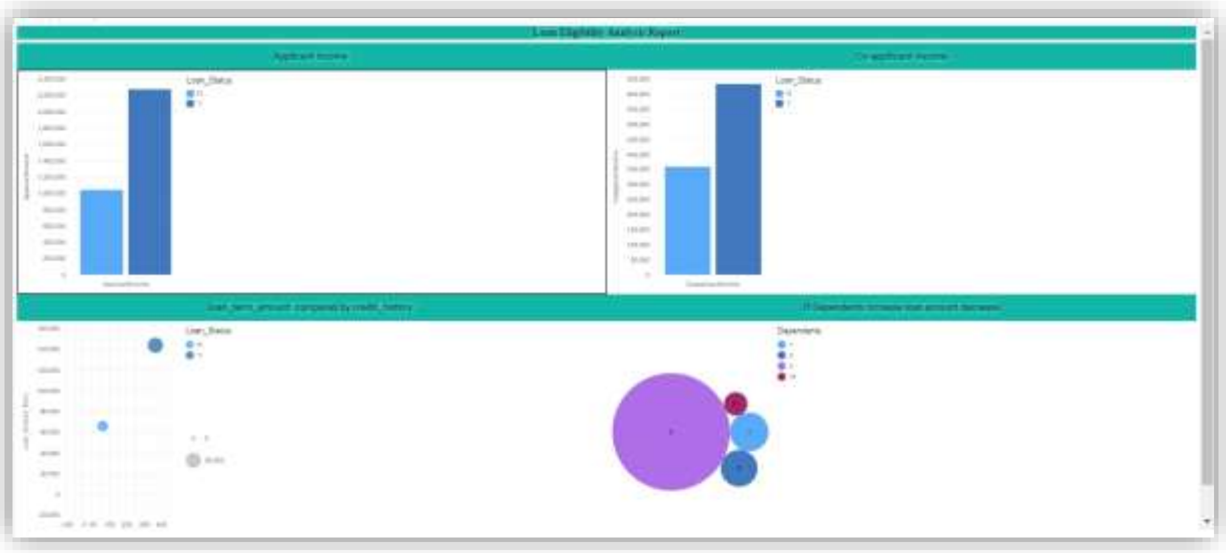


Fig.6.5 Report

6.3 GRAPHS

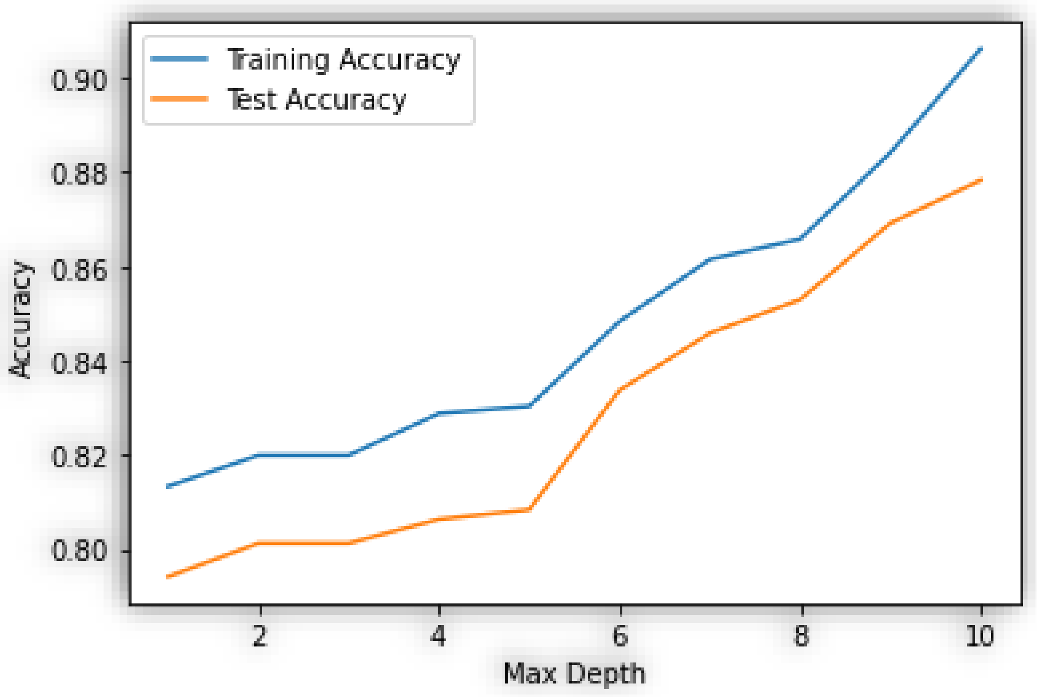


Fig.6.6 accuracy vs. max_depth graph for a decision tree

The graph shows how the model's accuracy changes as we vary the maximum depth of the decision tree. The x-axis represents the maximum depth of the decision tree, which is a hyperparameter that can tune to improve the model's performance. The y-axis represents the accuracy of the model on the training and testing sets, which is a measure of how well the model can predict the loan status of new customers. The blue line represents the model's accuracy on the training set, and the orange line represents the accuracy on the testing set. As increase the maximum depth of the decision tree, the model becomes more complex and can fit the training data better, which results in a higher training accuracy. However, this may lead to overfitting, where the model performs well on the training set but poorly on the testing set. In the graph, can observe that the training accuracy continues to increase as increase the maximum depth, while the testing accuracy reaches a maximum value at a certain depth (in this case, around 4). Beyond this depth, the testing accuracy starts to decrease, indicating that the model is overfitting to the training data. Therefore, can use this graph to select an optimal value for the maximum depth, where the testing accuracy is highest and the model is not overfitting. In this case, might select a maximum depth of 4 to achieve the best trade-off between training and testing accuracy.

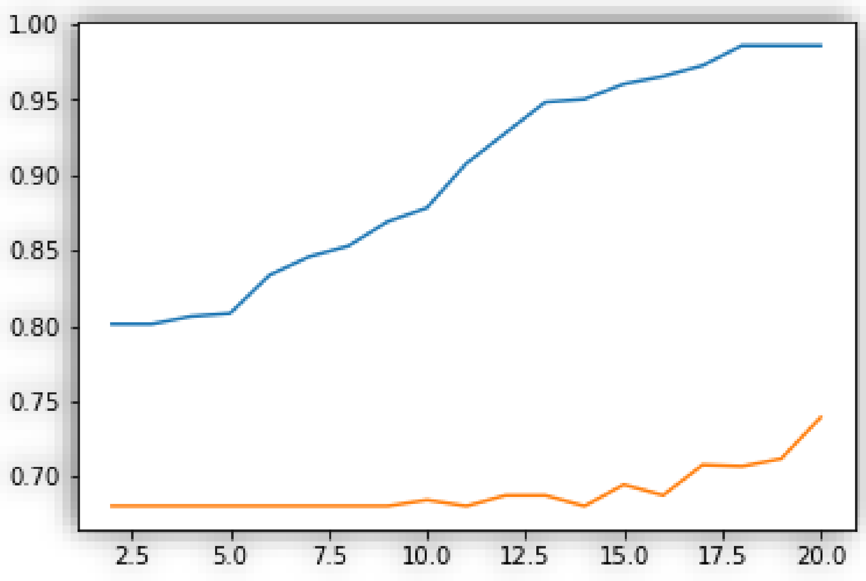


Fig.6.7 accuracy vs. max depth graph for different classifiers

Decision Tree, Random Forest, Logistic Regression, and Gradient Boosting as the maximum depth of the tree increases. The x-axis of the graph represents the maximum depth of the tree, ranging from 2 to 20. The y-axis represents the accuracy score of the classifier. For each classifier, the code trains the classifier with different maximum depths and calculates the accuracy score for each depth. The accuracy scores are then plotted against the maximum depth on the graph. The graph allows us to compare the accuracy of different classifiers and see how their performance changes with different maximum depths. It can help us choose the best classifier and determine the optimal maximum depth for the model.

6.4 PERFORMANCE ANALYSIS

6.4.1 CONFUSION MATRIX

True positives (TP): Predicted positive and are actually positive

False positives (FP): Predicted positive and are actually negative

True negatives (TN): Predicted negative and are actually negative

False negatives (FN): Predicted negative and are actually positive

Confusion Matrix is a representation of the above parameters in a matrix format.

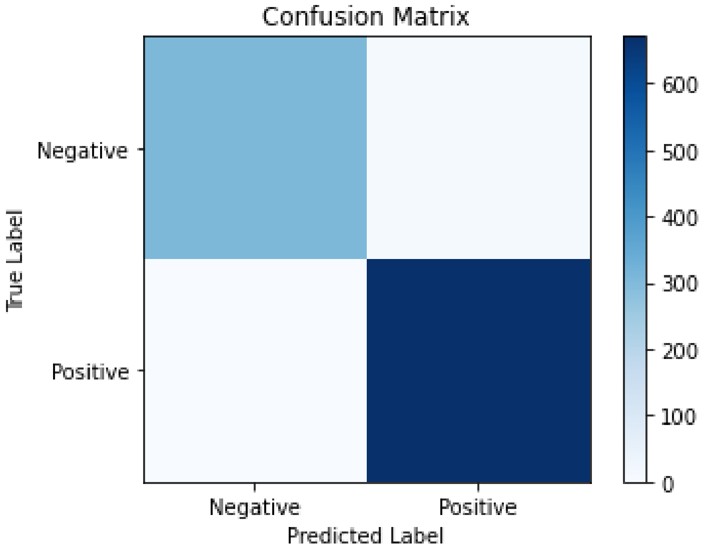


Fig.6.8 Confusion Matrix

6.4.2 ACCURACY

The most commonly used metric to judge a model and is actually not a clear indicator of the performance. The worse happens when classes are imbalanced.

$$\frac{TP + TN}{TP + FP + TN + FN}$$

6.4.3 PRECISION

Percentage of positive instances out of the total predicted positive instances. Here denominator is the model prediction done as positive from the whole given dataset. Take it as to find out ‘how much the model is right when it says it is right’.

$$\frac{TP}{TP + FP}$$

6.4.4 RECALL

Percentage of positive instances out of the total actual positive instances. Therefore denominator (*TP + FN)* here is the *actual* number of positive instances present in the dataset. Take it as to find out ‘how much extra right ones, the model missed when it showed the right one*’*.

$$\frac{TP}{TP + FN}$$

6.4.5 F1 SCORE

It is the harmonic mean of precision and recall. This takes the contribution of both, so higher the F1 score, the better. See that due to the product in the numerator if one goes low, the final F1 score goes down significantly. So, a model does well in F1 score if the positive predicted are actually positives (precision) and doesn't miss out on positives and predicts them negative (recall).

$$\frac{2}{\frac{1}{precision} + \frac{1}{recall}} = \frac{2 * precision * recall}{precision + recall}$$

6.4.6 ROC CURVE

ROC stands for **Receiver Operating Characteristic** and the graph is plotted against TPR and FPR for various threshold values. As TPR increases FPR also increases. As you can see in the first figure, Have four categories and want the threshold value that leads us closer to the top left corner. Comparing different predictors (here 3) on a given dataset also becomes easy as you can see in figure 2, one can choose the threshold according to the application at hand. ROC AUC is just the area under the curve, the higher its numerical value the better.

$$\textit{True Positive Rate (TPR)} = \textit{RECALL} = \frac{TP}{TP+FN}$$

$$\textit{False Positive Rate (FPR)} = 1 - Specificity = \frac{FP}{TN+FP}$$

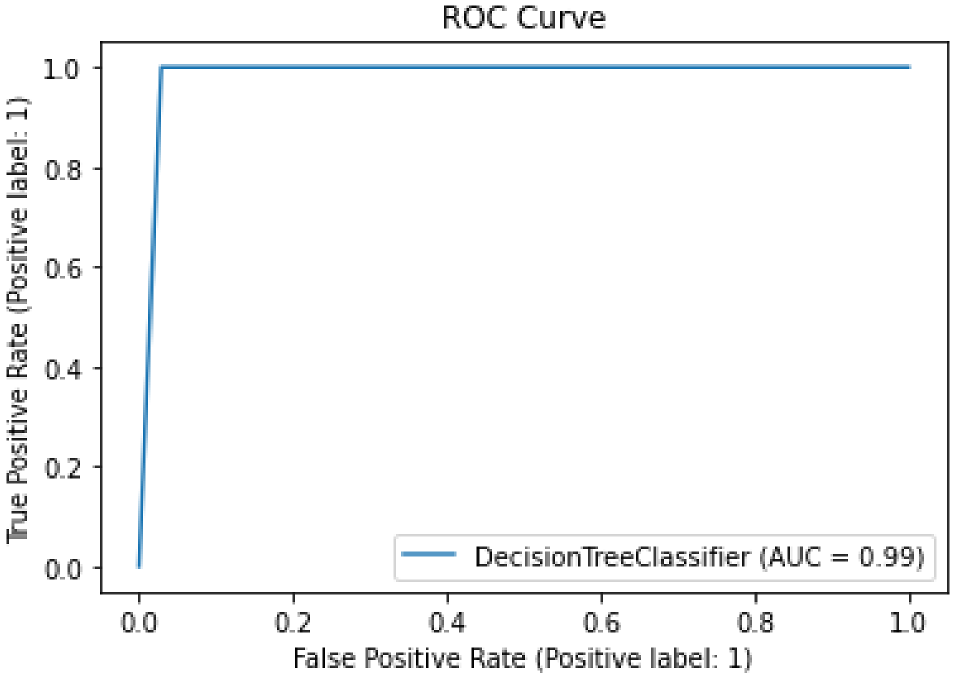
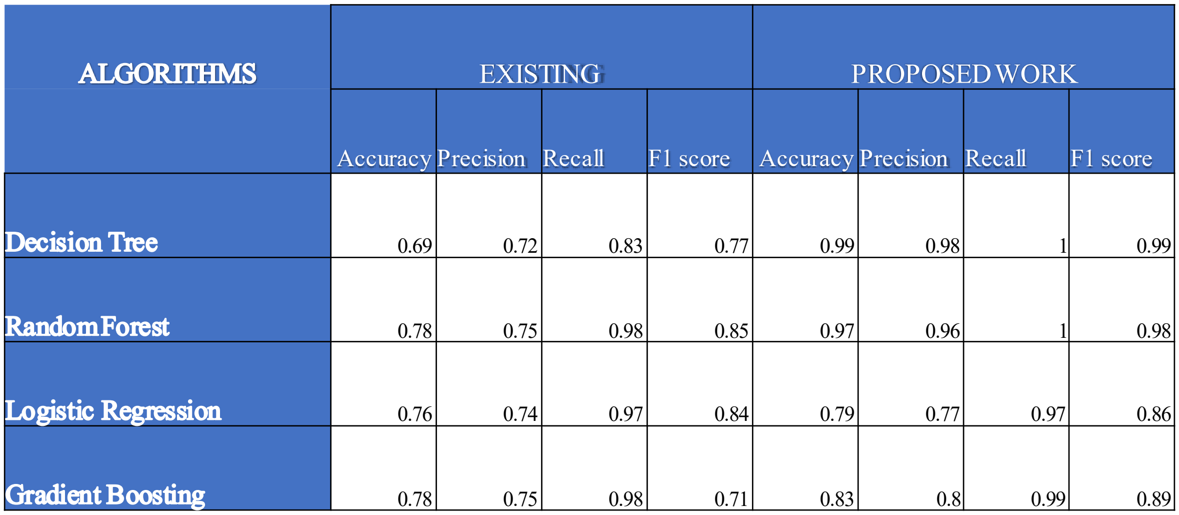


Fig.6.9 ROC Curve

6.5 COMPARATIVE ANALYSIS

The Comparative Analysis Facilitates An Informed Comparison Between Existing And Proposed Algorithms, Enabling Stakeholders To Make Data-driven Decisions Regarding The Loan Prediction Project. It Empowers Project Teams To Identify And Implement Algorithms That Deliver Improved Accuracy And Performance, Ultimately Enhancing The Effectiveness And Reliability Of Loan Approval Predictions.



ALGORITHMS	EXISTING				PROPOSED WORK			
	Accuracy	Precision	Recall	F1 score	Accuracy	Precision	Recall	F1 score
Decision Tree	0.69	0.72	0.83	0.77	0.99	0.98	1	0.99
Random Forest	0.78	0.75	0.98	0.85	0.97	0.96	1	0.98
Logistic Regression	0.76	0.74	0.97	0.84	0.79	0.77	0.97	0.86
Gradient Boosting	0.78	0.75	0.98	0.71	0.83	0.8	0.99	0.89

Table 6.1 Comparative Analysis

6.5.1 ACCURACY COMPARISON

The Accuracy Comparison Summary Showcases The Algorithm With The Highest Accuracy, Highlighting Its Effectiveness In Making Accurate Loan Approval Predictions. It Serves As A Key Performance Indicator For The Loan Prediction Project, Demonstrating The Project's Success In Developing An Accurate Predictive Model.

The Accuracy Comparison Offers a Clear and Concise Assessment of The Algorithms' Performance in Predicting Loan Approvals. It Facilitates Informed Decision-making, Enabling Stakeholders to Select the Most Accurate Algorithm for Loan Prediction and Ensuring The Project's Reliability And Efficacy In Real-world Scenarios

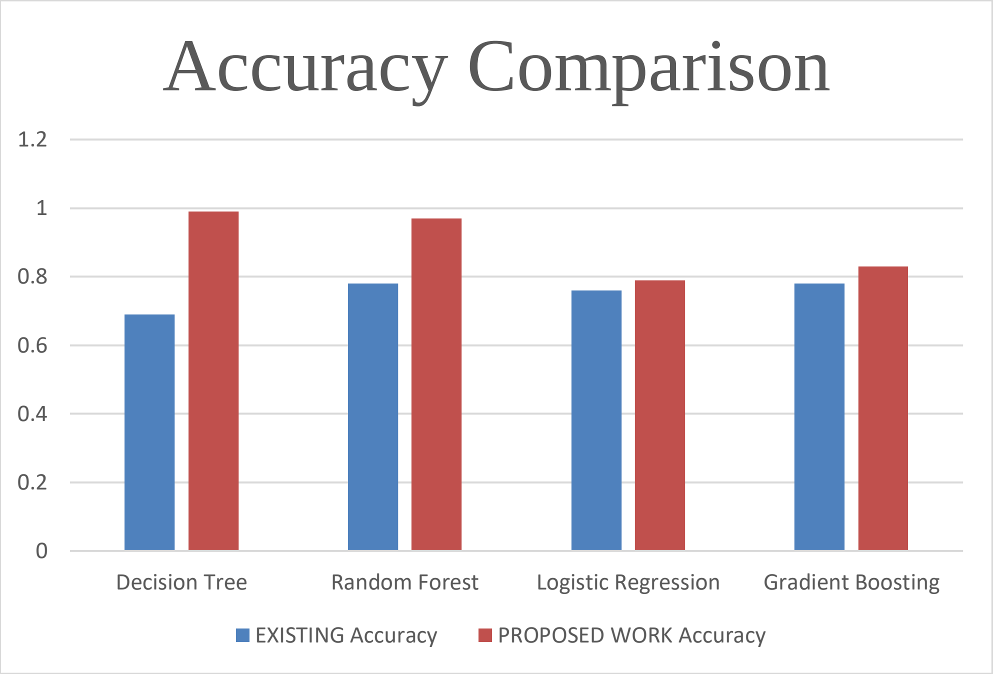


Fig.6.10 Accuracy Comparison

6.5.2 PRECISION COMPARISON

Precision is a metric that measures the proportion of true positive predictions among all positive predictions made by an algorithm. By comparing the precision scores of various algorithms, the analysis aims to identify the algorithm that exhibits the highest precision in predicting loan approvals. A higher precision score indicates that the algorithm is more precise in correctly classifying loan applications as approved or not approved.

The precision comparison provides valuable insights into the algorithms' performance in predicting loan approvals with high precision. It facilitates informed decision-making, allowing stakeholders to choose the algorithm that offers the highest precision and ensures accurate loan approval predictions in the project.

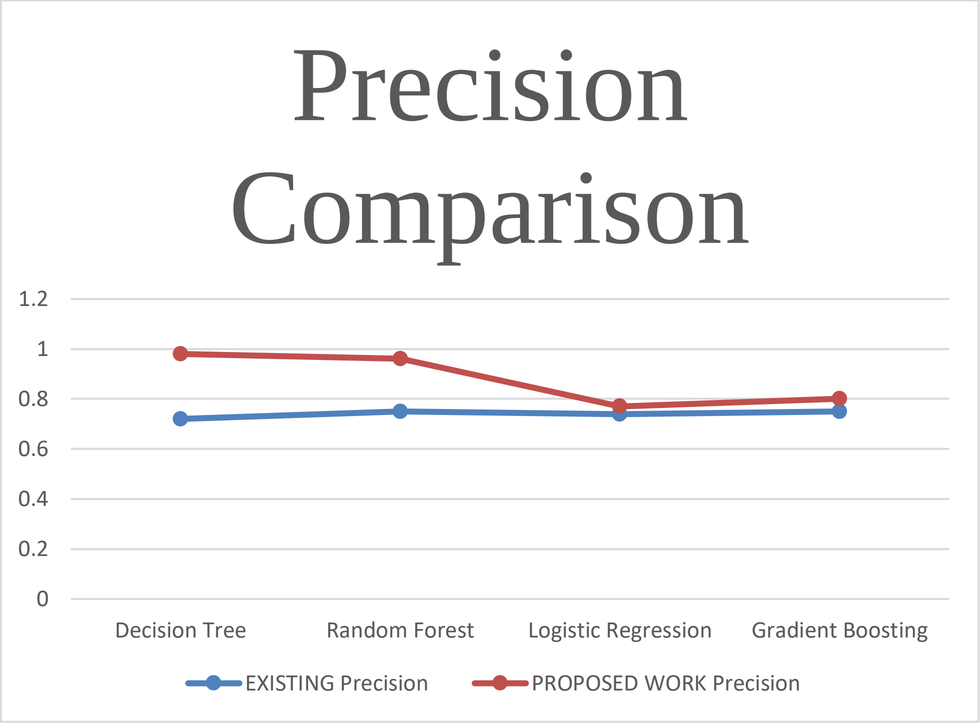


Fig.6.11 Precision Comparison

6.5.3 RECALL COMPARISON

Recall, also known as sensitivity or true positive rate, measures the proportion of actual positive cases that are correctly identified by an algorithm. The comparison examines the recall scores of various algorithms to identify the algorithm that exhibits the highest recall in predicting loan approvals. A higher recall score indicates that the algorithm is more successful in correctly identifying and capturing positive loan approvals.

The recall comparison provides valuable insights into the algorithms' performance in capturing positive loan approvals. It assists stakeholders in selecting the algorithm with the highest recall, ensuring accurate loan prediction outcomes and minimizing missed opportunities for potential borrowers.

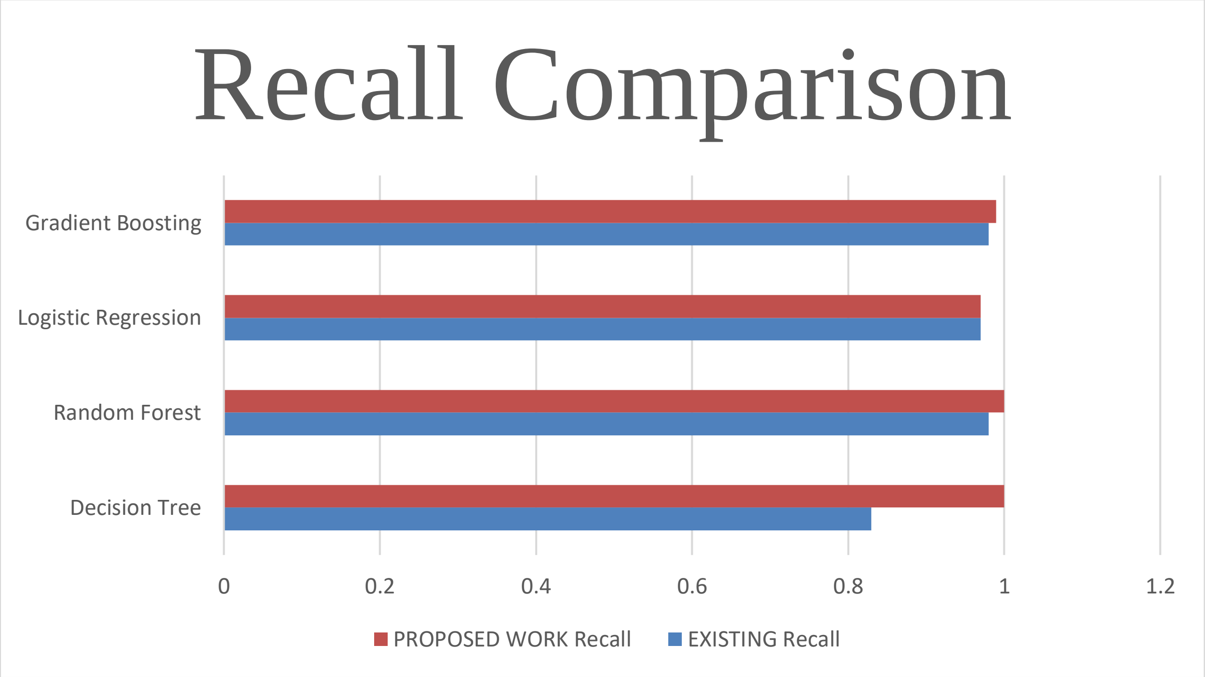


Fig.6.12 Recall Comparison

6.5.4 F1 SCORE COMPARISON

F1 score is a measure that combines both precision and recall, providing a balanced evaluation of an algorithm's performance. By comparing the F1 scores of various algorithms, the analysis aims to identify the algorithm that achieves the highest F1 score in predicting loan approvals. The F1 score takes into account both the algorithm's ability to minimize false positives (precision) and its ability to minimize false negatives (recall).

The F1 score comparison provides valuable insights into the algorithms' performance in predicting loan approvals. It facilitates informed decision-making by considering both precision and recall, ensuring that the selected algorithm achieves a balance between minimizing false positives and false negatives. This ultimately leads to more accurate loan prediction outcomes and better decision-making in the loan approval process.

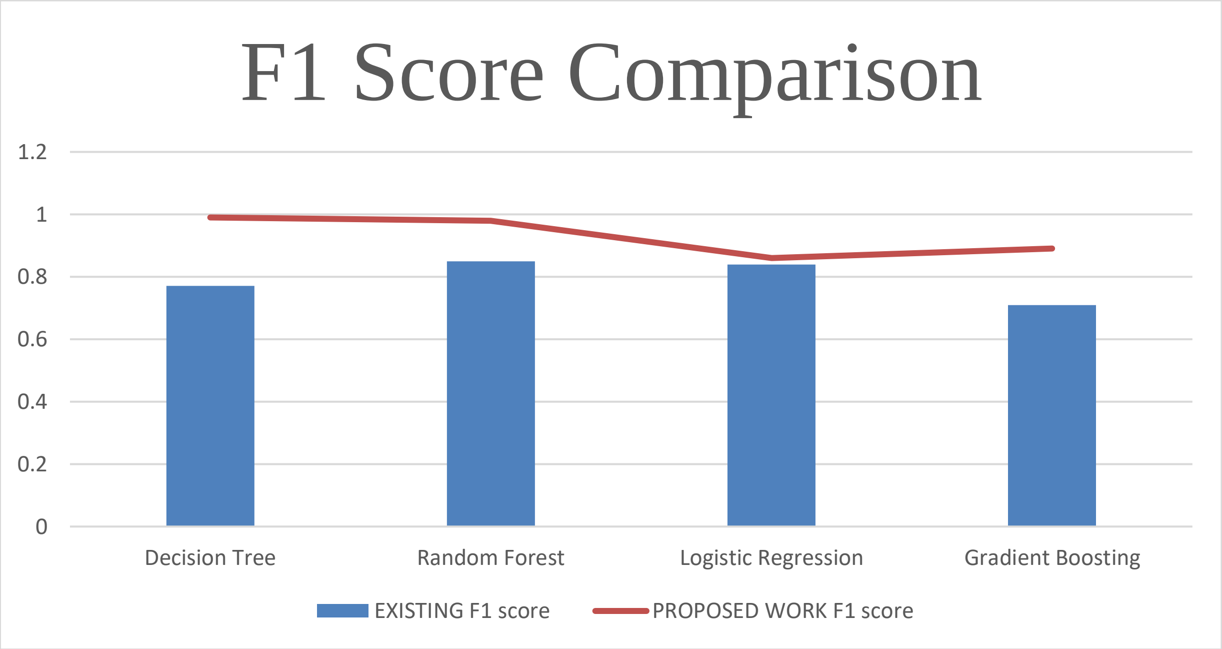


Fig.6.13 F1 Score Comparison

CHAPTER 7

CONCLUSION AND FUTURE WORK

7.1 CONCLUSION

In conclusion, the proposed system of Bank Loan Prediction using Machine Learning and Data Analytics with an Interactive Dashboard has the potential to revolutionize the loan approval process in banks. By automating the loan eligibility prediction using machine learning algorithms and presenting the results in a user-friendly interactive dashboard, this system can significantly reduce the time and effort required by bank staff to process loan applications. The predictive model trained on historical loan data can accurately classify loan applications into eligible or ineligible categories, and the data analytics dashboard provides valuable insights into the loan approval process. This system is a prime example of the power of machine learning and data analytics in transforming traditional industries such as banking.

7.2 FUTURE WORK

The predictive model could be further improved by incorporating additional data sources, such as social media activity, employment history, or educational background, to better assess the creditworthiness of customers. The accuracy of the predictive model could be improved by fine-tuning the model's parameters, such as the learning rate or regularization term, using techniques such as grid search or random search. While the model used in the project achieved good accuracy, The model could be enhanced by incorporating more complex features such as loan payment history or customer behavior patterns, which can help to identify potential risks and opportunities for more accurate loan eligibility predictions. The interactive dashboard and reports could be further improved to make them even more user-friendly and intuitive for bank staff, allowing them to quickly and easily access the information they need to make lending decisions

REFERENCES

[1] xolani dastile , turgay celik, and hans vandierendonck ,"Model-Agnostic Counterfactual Explanations in Credit Scoring",*IEEE ACCESS,2022,VOLUME 10, 2022.*

[2] Richa Manglani and Anuja Bokhare, "Logistic Regression Model for Loan Prediction:A Machine Learning Approach ", *IEEE XPLORE, 2021, ETI 4.0*

[3] Ch. Naveen kumar, D.keerthana, M.Kavitha, M.Kalyani, "Customer Loan Eligibility Prediction using Machine Learning Algorithms in Banking Sector", *IEEE XPLORE, 2022, (ICCES 2022).*

[4] Ugochukwu .E. Orji, Chikodili.H.Ugwuishiwu, Joseph. C. N. Nguemaleu, Peace. N. Ugwuanyi, "Machine Learning Models for Predicting Bank Loan Eligibility", *IEEE XPLORE, 2022, 4th (NIGERCON).*

[5] Vishal Singh, Vishal Singh, Rajat Awasthi, "Prediction of Modernized Loan Approval System Based on Machine Learning Approach"*, IEEE ACCESS, CONIT, Karnataka, India. June 25-27, 2021.*

[6] L. Udaya Bhanu1 , Dr. S. Narayana, "Customer Loan Prediction Using Supervised Learning Technique ", *International Journal of Scientific and Research Publications, 2021, Volume 11, Issue 6, June 2021.*

[7] Yu Zhong And Huiling Wang, "Internet Financial Credit Scoring Models Based on Deep Forest and Resampling Methods", *IEEE ACCESS,2023,Volume 11 2023.*

[8] Rana Alasbahi And Xiaolin Zheng, "An Online Transfer Learning Framework With Extreme Learning Machine for Automated Credit Scoring" , *IEEE ACCESS,2022, Volume 10 2022.*

[9] Min Sue Park , Hwijae Son, Chongseok Hyun, And Hyung Ju Hwang , "Explainability of Machine Learning Models for Bankruptcy Prediction" , *IEEE ACCESS, 2021, Volume 09 2021.*

[10] Yen-Ru Chen , Jenq-Shiou Leu , Sheng-An Huang , Jui-Tang Wang , "Predicting Default Risk on Peer-to-Peer Lending Imbalanced Datasets" , *IEEE ACCESS, 2021.*

[11] Haibo Wang , Wendy Wang , Yi Liu , And Bahram Alidaee ,“Integrating Machine Learning Algorithms With Quantum Annealing Solvers for Online Fraud Detection” , *IEEE ACCESS, 2022, Volume 10 2022.*

[12] Anshika Gupta , Vinay Pant , Sudhanshu Kumar and Pravesh Kumar Bansal, “Bank Loan Prediction System using Machine Learning”, *IEEE 2020.*

[13] Amruta S. Aphale, Dr. Sandeep R. Shinde,” Predict Loan Approval in Banking System Machine Learning Approach for Cooperative Banks Loan Approval”, *IJERT, Volume 09, Issue 08 (August 2020), October 2020*.

[14] Nikhil Madane, Siddharth Nanda, “Loan Prediction using Decision tree”, *Journal of the Gujrat Research History, Volume 21 Issue 14s, December 2019*.

[15] Amruta S. Aphale and R. Prof. Dr. Sandeep. R Shinde, “Predict Loan Approval in Banking System Machine Learning Approach for Cooperative Banks Loan Approval”, *International Journal of Engineering Trends and Applications (IJETA), vol. 9, issue 8, 2020)*

APPENDIX

SAMPLE CODE

app.py

```
# save this as app.py
from flask import flask, escape, request, render_template
import pickle
import numpy as np

app = flask(__name__)
model = pickle.load(open('model.pkl', 'rb'))

@app.route('/')
def home():
    return render_template("prediction.html")


@app.route('/predict', methods=['get', 'post'])
def predict():
    if request.method ==  'post':
        gender = request.form['gender']
        married = request.form['married']
        dependents = request.form['dependents']
        education = request.form['education']
        employed = request.form['employed']
        credit = float(request.form['credit'])
        area = request.form['area']
        applicantincome = float(request.form['applicantincome'])
        coapplicantincome = float(request.form['coapplicantincome'])
        loanamount = float(request.form['loanamount'])
        loan_amount_term = float(request.form['loan_amount_term'])

        # gender
        if (gender == "male"):
            male=1
        else:
            male=0

        # married
        if(married=="yes"):
            married_yes = 1
```

```
else:
    married_yes=0

# dependents
if(dependents=='1'):
    dependents_1 = 1
    dependents_2 = 0
    dependents_3 = 0
elif(dependents == '2'):
    dependents_1 = 0
    dependents_2 = 1
    dependents_3 = 0
elif(dependents=="3+"):
    dependents_1 = 0
    dependents_2 = 0
    dependents_3 = 1
else:
    dependents_1 = 0
    dependents_2 = 0
    dependents_3 = 0

# education
if (education=="not graduate"):
    not_graduate=1
else:
    not_graduate=0

# employed
if (employed == "yes"):
    employed_yes=1
else:
    employed_yes=0

# property area

if(area=="semiurban"):
    semiurban=1
    urban=0
elif(area=="urban"):
    semiurban=0
    urban=1
else:
    semiurban=0
```

```
            urban=0


        applicantincomelog = np.log(applicantincome)
        totalincomelog = np.log(applicantincome+coapplicantincome)
        loanamountlog = np.log(loanamount)
        loan_amount_termlog = np.log(loan_amount_term)

        prediction = model.predict([[credit, applicantincomelog,loanamountlog, 
loan_amount_termlog, totalincomelog, male, married_yes, dependents_1, 
dependents_2, dependents_3, not_graduate, employed_yes,semiurban, urban ]])

        # print(prediction)

        if(prediction=="n"):
            prediction="no"
        else:
            prediction="yes"


        return render_template("prediction.html", prediction_text="loan status is 
{}".format(prediction))




    else:
        return render_template("prediction.html")




if __name__ == "__main__":
    app.run(debug=true)
```