Predictive Prowess: Unleashing the Power of Machine Learning for Futuristic Loan Approval Precision

Mohamed Elkmeshi[1],Fathy Farag[2],
Wael[3],Ahmed Osama[4],Ahmed Mohamed [5]
Faculty of Computer Science
Misr International University, Cairo, Egypt
mohamed2108926[1], fathy2111262[2],
nada1914465[3], hady1907151[4], Ahmed2110223 [5]{@miuegypt.edu.eg}

Abstract*—The landscape of financial decision-making underscores the paramount importance of accurate loan approval prediction models.The ever-evolving financial ecosystem demands sophisticated tools beyond traditional methods to scrutinize diverse features within loan datasets,making the utilization of machine learning algorithms a pivotal strategy to enhance precision and efficiency in evaluating loan applications.In This paper, we addressed the critical need for robust loan approval prediction models through a comprehensive exploration of machine learning classifiers, including K-Nearest Neighbors (KNN), Random Forest Classifier (RFC), Gaussian Naive Bayes (GNB), Logistic Regression (LC), Decision Tree Classifier (DTC), and Gradient Boosting Classifier (GBC). The classifiers undergo rigorous evaluation, and preprocessing steps, encompassing the encoding of categorical variables and handling missing values, are essential in preparing the dataset for analysis and the datasets used in this study are LoanApprovalPrediction.csv,loan.approval.dataset.csv,and Loan.Approval.csv This study aims to propose advanced loan approval prediction algorithms and evaluate their performance using metrics like accuracy, precision, recall, and F1-score on training and testing sets. The results are as follows: In the first dataset,the best-performing models are Gradient Boosting and Random Forest. For the second dataset, Logistic Regression stands out as the best model, with Random Forest maintaining a good balance.In the third dataset, Logistic Regression and Gradient Boosting are identified as the top-performing models. In conclusion, this research highlights machine learning's essential contribution to enhancing loan approval processes in financial technology. The proposed algorithms provide precision for creating robust systems in the evolving financial landscape. Continuous development of predictive models remains crucial for sustaining effective loan approval systems in the face of advancing financial technologies.

Keywords: Loan Approval Prediction; Machine Learning; Classification; K-Nearest Neighbors; Random Forest Classifier; Gaussian Naive Bayes; Logistic Regression; Decision Tree Classifier; Gradient Boosting Classifier.

I. Introduction

Loans constitute a fundamental offering of banks and various financial institutions. With a substantial number of individuals seeking financial assistance for diverse purposes, the customer base has expanded, and some financial institutions anticipate significant earnings through interest on loans. Nonetheless, the extension of loans comes with the inherent risk of default, where certain borrowers may struggle to repay their debts. Consequently, a high prevalence of non-performing loans poses a potential threat to the stability of the banking sector, potentially leading to financial insolvency. To mitigate these risks, one crucial aspect for banks is to thoroughly assess whether loan applicants possess the capability to repay the loan within the specified terms.

According to the Malaysian central bank's report, loans outstanding to Small and Medium Enterprises (SMEs) by the banking industry in the country was worth RM 104.6 billion in 2006 . This amount, representing 44.5% of the total loans outstanding at that time, not only highlights the strong contribution of the SME sector to loan demand for the banking industry, but also a heavy reliance of the SME sector [1] on the banking sector for financing purposes.

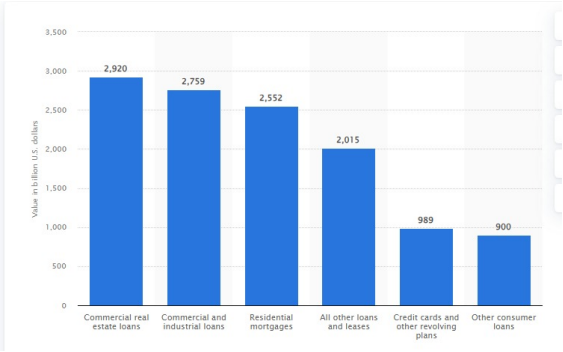


Fig. 1. Value of loans owned by commercial banks in the United States as of May 2023, by category

Machine learning is a subfield of artificial intelligence that is still in its infancy. Its principal vision is to develop systems that can learn and form hypotheses based on their experiences. It builds a model through the training of machine learning algorithms on a training dataset. The model predicts the likelihood of loan approval based on the new input data. It constructs

models through the detection of obscure patterns in the input dataset using machine learning. For novel datasets, it makes accurate predictions. After the dataset has been processed and any null values have been filled. By using the new input data the model is assessed for accuracy , it then predicts the probability of loan approval. Machine learning techniques are categorized into supervised learning, unsupervised learning, and reinforcement learning.

In supervised learning, the model is trained on a labelled dataset. It has input data and output data. The data is categorized and partitioned into training and test datasets. The training dataset is used to train the model, while the testing dataset is used to evaluate the model's accuracy. The dataset contains models and their output. Its application is exemplified through classification and regression.

The data being used to train in unsupervised learning is not labelled or classed in the dataset. The objective is to find hidden patterns in the data. The model is being taught to recognize patterns. It may effortlessly anticipate hidden patterns for each new input dataset, After reviewing data, it draws inferences from datasets in order to define hidden patterns. There are no results in the dataset while using this method. Unsupervised learning is illustrated by the clustering approach.

In reinforcement learning, it will not utilize labelled datasets, nor are the outputs related with data; instead,The model has been honed via experience. The model improves its presentation depending on its relationship with the environment, and determines how to resolve its shortcomings and accomplish the intended result by assessing and evaluating a variety of options. Classification algorithms are prominent supervised learning techniques for determining the likelihood of loan approval circumstance.

This is where Machine Learning takes part. Machine Learning is a technology that can apply to many fields and produce unexcelled results [2], just some statistics need to be collected, and usually they already are, and Machine Learning can start working and picking up on patterns that are too vague for traditional statistics to detect. professionals already collect enormous amounts of data from their work, so a lot of data can be passed into the model to produce more accurate models and producing better results. Therefore, this makes the field of predicting loan approval a prime field for Machine Learning to operate in.

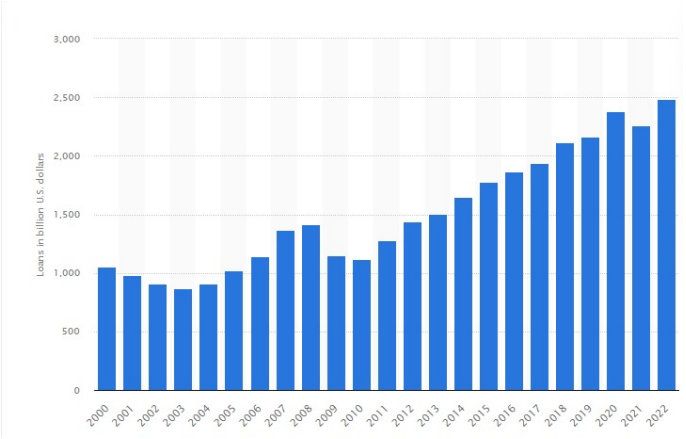


Fig. 2. Value of commercial and industrial loans granted by FDIC-insured commercial banks in the United States from 2000 to 2022

This research paper: Applying machine learning in the loan approval process involves several steps. Firstly, relevant data, encompassing details like financial history and credit scores, is collected. Subsequently, the data undergoes preprocessing to handle missing values and convert categorical variables for model suitability. Key features influencing loan decisions, such as credit scores and income, are identified. Machine learning algorithms are then employed for model training, learning patterns from historical data to predict loan approval outcomes. Model performance is evaluated using metrics like accuracy and precision and Recall and F1-Score. Once validated, the model is deployed into the loan approval system to analyze new applications. Regular monitoring and updates ensure ongoing accuracy, efficiency, and alignment with evolving trends and regulations. This integration of machine learning streamlines decision-making, boosts efficiency, and fosters data-driven and risk-aware loan approval processes.

The contributions made to this topic are:

- The loan approval Prediction with machine learning.
- The testing of 6 machine learning algorithms (K-Nearest Neighbors (KNN) and Random Forest Classifier (RFC) and Naïve Bayes and Logistic Regression (LC) and Decision Tree Classifier (DTC) and Gradient Boosting Classifier (GBC)) and we found that In a comprehensive evaluation across three datasets, machine learning algorithms were assessed for their performance. Notably, in the first dataset, Gradient Boosting and Random Forest outperformed others with accuracies of 0.972 and 0.977, respectively. Logistic Regression, k-NN, and Naïve Bayes exhibited comparatively lower accuracies. In the second dataset, Logistic Regression achieved the highest accuracy (0.817), while Random Forest demonstrated a balanced performance with a high accuracy of 0.800 and a notable F1-Score of 0.869. The third dataset revealed high accuracies for Logistic Regression and Gradient Boosting (0.907), though with perfect precision and low recall, indicating potential class imbalance. Random Forest showcased balanced performance, K-NN displayed room for improvement, Naïve Bayes effectively captured positive

instances, and Decision Tree demonstrated competence in classification tasks.
- The use of 3 datasets, for a total of 574,224 entries with different sets of features spanning between 12 and 34 features.

The remaining sections in this paper are ordered as the following; related work is discussed in the second section. Moreover, The third section clarifies the proposed methodology of the research; it consists of dataset description and used algorithms. The results of the proposed algorithms can be found in the fourth section and their analysis. The conclusion is located in the fifth section. An acknowledgement towards all the supporting figures of this research is present in the sixth section.

II. Related Work

The realm of loan approval prediction is not unexplored numerous researchers have investigated this field and achieved satisfactory results We will reference some of the papers we read and analyzed to aid our research with all mentioned papers appropriately cited in the references section.

The paper.[3] addresses a critical problem faced by banks, namely the challenge of efficiently selecting safe loan applicants to optimize asset allocation. The primary objective is to enhance the precision of predicting whether assigning a loan to a particular individual is a safe decision, ultimately reducing the risk associated with selecting loan applicants and conserving bank resources. The results of the paper involve a thorough comparison of various machine learning models applied to collected data, followed by the training of the system on the most promising model and subsequent testing. The ultimate outcome is the development of a highly efficient loan prediction system. This system can seamlessly integrate with other banking systems, offering a swift, immediate, and streamlined method for banks to identify and choose deserving loan applicants, thereby fostering a more effective and risk-aware approach to loan allocation.

This paper.[4] addresses the prevalent challenge faced by banking organizations in accurately predicting credit defaulters, emphasizing the imperative for a more precise predictive modeling system for loan approval. The objective is to delve into the efficacy of machine learning algorithms, specifically Logistic Regression, Decision Tree, and Random Forest, in predicting loan approval and to conduct a comparative analysis of their accuracy. The study's results reveal that the Decision Tree machine learning algorithm exhibits superior accuracy in predicting loan approval compared to Logistic Regression and Random Forest. These findings underscore the potential of machine learning techniques to enhance the loan approval prediction process, offering significant advantages to the banking industry. Furthermore, the application of these techniques can streamline the validation of features, automate aspects of the approval process, and provide a swift and efficient means for selecting deserving loan applicants, contributing to a more effective and informed decision-making framework within the banking sector.

The paper.[5] addresses challenges faced by financial institutions in determining loan eligibility, proposing automation through machine learning algorithms to predict loan safety. The primary goal is risk reduction in selecting individuals capable of timely repayment, thereby minimizing non-performing assets. Experimental tests highlight Naïve Bayes' superior performance in loan forecasting. The outcome is a comprehensive system for efficient loan approval decisions, encompassing data collection, preprocessing, model selection, evaluation, and classification. The paper's contribution lies in the proposed loan prediction model, offering a swift and user-friendly approach. This model aims to assist both banking employees and applicants by streamlining the loan approval process, ensuring quick and informed decisions on deserving loan applicants.

This paper.[6] addresses the challenge faced by financial organizations in precisely estimating and approving loan applications, as inaccurate estimations or lack of information can impact operational financial processes and increase credit risks. The objective is to investigate the loan prediction process through the application of various machine learning algorithms. The proposed methodology involves data preprocessing to enhance data quality, and three machine-learning algorithms—Logistic Regression, Decision Tree, and Random Forest—are trained and tested to compare their accuracy in predicting loan status. Results reveal that Logistic Regression outperformed the other algorithms in accuracy, precision, recall, F1, and Area under the curve (AUC). The identification of Logistic Regression as the most accurate prediction model suggests its suitability for loan prediction, offering financial organizations a valuable tool for making more precise and timely decisions regarding loan approvals.

This paper.[7] addresses challenges faced by many banks in accurately determining customer eligibility for loan approval highlighting the limitations of existing time-consuming verification processes The objective is to introduce a Loan Approval Prediction System utilizing machine learning specifically the Random Forest algorithm to classify customers accurately based on various parameters and enhance the efficiency of loan approval decisions. Results from the study indicate that the implemented Random Forest algorithm within the Loan Approval Prediction System yields accurate and immediate results for approving loans to eligible customers efficiently handling a large volume of applications. The outcome is the development of a software system capable of accurately predicting customer eligibility for loan approval leveraging machine learning and the Random Forest algorithm to process extensive customer data and provide swift precise loan approval predictions.

The paper "Loan Approval Prediction Using Machine Learning".[8] outlines the challenges faced by banks in manually selecting loan applicants and emphasizing potential inefficiencies and misunderstandings the objective is to address this issue by employing machine learning algorithms including Logistic Regression,Random Forest,Decision Tree,Support Vector Machine, Naive Bayes to automatically identify suitable candidates for bank loans.The result is a loan approval prediction model based on historical loan data with the accuracy of each algorithm evaluated.The outcome highlights the identification of the most suitable machine learning algorithm or combination for accurate predictions offering insights into optimizing the loan approval process.Through a comparative analysis of different algorithms the paper aims to guide practitioners in selecting the most effective approach for loan approval predictions.

The problem statement of this paper.[9] is the need to predict whether a loan applicant is a defaulter or non-defaulter in a given time.The objective of the paper is to use machine learning algorithms to analyze loan applications and determine the credibility of loan applicants.The results found in the paper show that the decision tree,random forest,logistic regression models can be used to predict loan repayment with high accuracy.The outcome of the results is the development of a loan prediction system that can automate the validation of loan application features and provide a quick and easy way to choose deserving applicants.This system can provide special advantages to banks and financial companies and help them make informed decisions about loan approvals.

The paper [10] focuses on building a machine learning model for predicting the creditworthiness of bank customers. The primary objective is to identify key features influencing creditworthiness and evaluate the performance of selected algorithms. The study employs linear regression to create a predictive model and follows a methodology involving data collection, analysis, algorithm training, testing, and performance assessment using metrics like accuracy, sensitivity, specificity, and precision. Results indicate that various machine learning algorithms achieve accuracy rates ranging from 76% to over 80%. The paper highlights significant features affecting creditworthiness and compares algorithm performance based on these features. Throughout, standard mathematical notations and symbols, including Greek letters and abbreviations like SVMs and KNN, are utilized to express mathematical and algorithmic concepts.

The paper [11] addresses challenges faced by Small and Medium Enterprises (SMEs) in obtaining loans from local commercial banks, with a specific focus on the banks' perspective. Its objective is to propose a more systematic research approach by categorizing key factors and offering a holistic understanding of the considerations made by bank officers when evaluating SME loan applications. The methodology involves integrating existing literature on SME financing and introducing a conceptual framework to comprehensively depict the factors considered by bank officers. Results emphasize the importance of moving beyond subjective relationships and incorporating qualitative information into the SME loan application process. The paper, however, does not provide specific result percentages. Key notations include SME (Small and Medium Enterprises), RM (Malaysian Ringgit), and % (Percentage).

The research paper[12] "Loan Approval Prediction based on Machine Learning Approach" addresses the risk associated with loan approval by employing machine learning to predict the safety of granting loans to individuals. The objective is to create a loan prediction system capable of automatically assigning weights to features in the loan processing and determining whether a new applicant is suitable for loan approval. The methodology involves utilizing six machine learning classification models, including Decision Trees, Random Forest, Support Vector Machine, Linear Model, AdaBoost, and Neural Network, to train and test the loan prediction system. Results indicate that the Random Forest model outperforms others with an accuracy of 80.94%. Key notations in the paper include variables like Dependents, EducationQualification, SelfEmployed, ApplicantIncome, CoApplicantIncome, LoanAmount, LoanAmountTerm, CreditHistory, PropertyArea, and LoanStatus.

The problem [13] statement of the study is the difficulty faced by banking organizations in accurately predicting loan approval. The objective of the study is to explore the use of machine learning algorithms to predict loan approval and to compare the accuracy of different algorithms. The methodology involves collecting a dataset of loan applications and using six different machine learning algorithms, including Decision Trees and Random Forest, to predict loan approval. The results show that the Decision Tree algorithm has the highest accuracy in predicting loan approval, with a percentage of 81.54%. The notations used in the study include Loan ID, Gender, Married, Dependents, Education, SelfEmployed, ApplicantIncome, CoapplicantIncome, LoanAmount, LoanAmountTerm, CreditHistory, PropertyArea, and LoanStatus.

The study [14] addresses the challenge of reducing loan approval time and associated risks by comparing various loan prediction models. The objective is to identify the most effective model for expediting approval time and mitigating risk, with a focus on parameters such as credit score, income, age, marital status, and gender. The methodology involves a comprehensive comparison of different prediction models, analyzing their limitations and advantages, and developing a modified prediction model for optimal accuracy and performance. The results indicate that the Random Forest prediction model outperforms others, with a specific result percentage not provided. Notations used in the study include

PLsFi for the filtering function used in attribute analysis, DS2 for the dataset yielding the highest accuracy, and CSV as the file format for importing training set data.

The paper [15] addresses the issue of Jordanian commercial banks' reluctance to adopt artificial intelligence (AI) in credit decision-making, leading to subjective and biased loan decisions. The objective is to develop an AI-based loan decision model using a multi-layer feed-forward neural network with a backpropagation learning algorithm. The model aims to simplify loan officers' tasks, enhance efficiency, and achieve productivity in Jordanian commercial banks. The methodology involves validating the model with representative loan application cases from various banks in Jordan. Results show that the neural network model successfully classified 95% of cases in the testing set, highlighting the effectiveness of AI in improving loan application evaluation in Jordanian commercial banks. The paper utilizes notations such as Business Intelligence (BI), Artificial Intelligence (AI), Artificial Neural Networks (ANN), Backpropagation (BP) algorithm, and Credit Scoring (CS).

The paper[16] "Machine Learning Application for Selecting Efficient Loan Applicants in Private Banks of Bangladesh" addresses the challenge of reducing loan default rates and improving the performance of private banks in Bangladesh by effectively identifying financially capable loan applicants. The objective is to leverage machine learning techniques to categorize loan applicants based on their repayment capabilities, thereby minimizing the risk of default. The methodology involves employing statistical dependency to establish a comprehensive background for the eligibility of loan candidates, incorporating fundamental business development processes and benchmark features from prior research. Results indicate a significant improvement, with approximately 98.3% of selected applicants making regular payments and only 4 individuals showing no payments in a 3-month interval. The paper employs notations such as XGBoost, Random Forest, and financial ratios as performance indicators for banks in Banglades

The paper [17] "Analysis of Loan Availability using Machine Learning Techniques" addresses the crucial challenge of predicting loan defaulters to minimize a bank's Non-Performing Assets. The study aims to maximize earnings by accurately identifying customers suitable for loan granting, given the increased demand for credit products. It employs machine learning models, primarily logistic regression, to analyze key variables influencing credit decisions and predicting loan defaulters. The paper emphasizes the importance of considering diverse customer characteristics beyond checking account information. The objective is to provide insights into using machine learning for improved loan availability, comparing the performance of logistic regression with other models. Results show logistic regression's effectiveness, with an accuracy score of 0.785, highlighting its reliability in predicting loan availability. The study contributes valuable insights into credit decision-making and risk assessment in the banking industry, showcasing the potential of machine learning models, particularly logistic regression, in enhancing loan availability prediction.

The article [18] "Prediction of Loan Behaviour with Machine Learning Models for Secure Banking" tackles the crucial issue of loan default prediction's impact on credit scores and financial stability. Emphasizing the necessity of machine learning models, it aims to predict loan default using various classification algorithms and assess their performance through metrics like accuracy and precision. Key findings include the identification of crucial parameters, such as employment experience and debt income, influencing loan default. The paper successfully applies multiple classification algorithms, providing insights into performance metrics and feature importance. Overall, it contributes valuable knowledge to using machine learning in predicting loan behavior and improving the basis for effective loan credit approval by identifying problematic clients among applicants.

The study[19] addresses the challenge faced by lenders in predicting the risk of loan default by employing machine learning and regression techniques for loan prediction. The objective is to enhance accuracy in the lending industry, enabling informed decisions to mitigate financial risk. The paper evaluates three algorithms—Logistic Regression, Random Forest, and Decision Tree—finding that Logistic Regression outperforms the others with an accuracy of 89.7059%. Key findings include the impact of credit history, income levels, and demographic factors on loan approval, highlighting that applicants with low credit history are less likely to be accepted, while high income increases eligibility. The results underscore the significance of machine learning algorithms, particularly Logistic Regression, in improving loan prediction accuracy, facilitating more effective risk assessment and resource management for lenders and borrowers alike.

The paper[20] addresses the challenge faced by the banking industry in predicting credit defaulters by proposing a machine learning-based credit scoring model. The authors aim to improve the accuracy of credit data analysis and minimize the risk of loss in lending decisions through the development of a logistic regression-based analysis model. The proposed model incorporates Min-Max normalization and linear regression combination to optimize precision for critical details and emphasizes the automation of the loan lending process using machine learning algorithms. Although specific results are not explicitly provided, the paper underscores the importance of credit scoring in the banking sector, discusses logistic regression for loan approval prediction, advocates for data preprocessing techniques, and highlights the significance of accurate credit data analysis to mitigate the risk of loss in lending decisions.

The thesis[21] addresses the application of machine learning to predict loan defaults, with a focus on identifying credit-worthy customers overlooked by traditional credit scores and minimizing the risk of loan default. The objective involves gathering loan data from the Lending Club website and utilizing machine learning techniques to predict whether a customer is likely to repay the loan or become a defaulter. The results reveal the development of a model achieving an accuracy of 0.35 and a precision of 0.93, allowing investors to make informed decisions with a higher likelihood of profit. While adopting a conservative approach, the thesis suggests future enhancements, including the exploration of different machine learning techniques like Random Forests and Neural Networks, to potentially improve the model's performance further. Overall, the outcome suggests that the developed machine learning model provides a valuable tool for investors seeking to confidently identify credit-worthy borrowers and reduce the risk of loan default.

The study[22] addresses the risk of user loan default in the context of P2P online lending platforms and aims to build a loan default prediction model using the random forest algorithm. The objective is to mitigate the risk of default and contribute to the sustainable development of these platforms. The paper evaluates various machine learning algorithms and finds that the random forest algorithm outperforms others, including logistic regression, decision tree, and support vector machine, with an impressive 98% accuracy in predicting loan default. The study highlights the strong generalization ability of the random forest model, and effective handling of imbalanced class issues through the SMOTE method. The outcomes suggest the significant potential of the random forest algorithm in real-world scenarios related to P2P lending platforms, providing valuable insights for the development of robust models to predict loan default and ensuring the sustainable and healthy growth of P2P online lending platforms.

III. Proposed Methodology

Numerous algorithms were used, and a research was done on each algorithm before training the model using them on the datasets. The following diagram represents the steps the datasets went through to get the results.

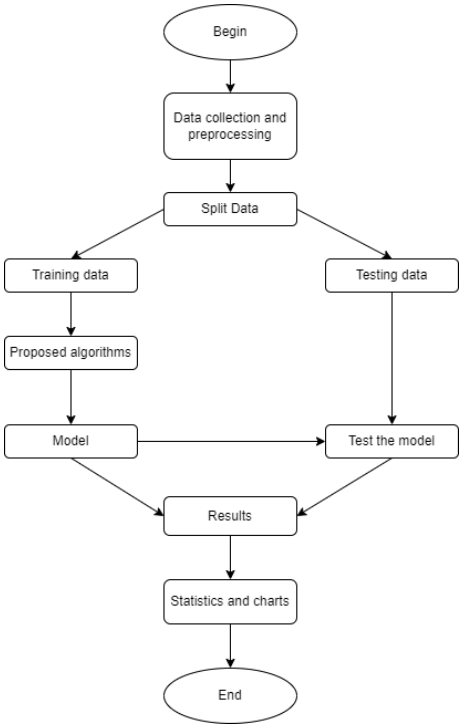


Fig. 3. loan approval prediction process

A. Data sets Descriptions

The first dataset consists of 12 features, and it has 4270 records.The dataset was normalized, then it was split into two partitions: 70% for training, and 30% for testing. A detailed description of the features can be found below.

loan status is the target, and it represents whether the person has or will have take the loan or not. education present if the person graduated or not . self employed if the person work for his own or not . income amount the amount that the person gain . loan amount the amount that person want to loan . loan term the time the person want to pay his loan. cibil score is like a rate for the person . residential asset value . commercial assets value. luxury assets value .luxury assets value. bank assets value .

TABLE I
FEATURES OF DATASET 1

Feature	Type	Values
Loan Status (Target)	Classification	Approved or rejected
Number of Dependents	Numerical	from 0 to 5
Education	Classification	Yes or No
Self-Employment	Classification	Yes or No
Annual Income	Numerical	From 200000 to 9900000
Loan Amount	Numerical	from 300000 to 39500000
Loan Term	Numerical	from 0 to 20
CIBIL Score	Numerical	300 to 900
Residential Assets Value	Numerical	100000 to 29100000
Commercial Assets Value	Numerical	from 0 to 19400000
Luxury Assets Value	Numerical	from 300000 to 39200000
Bank Asset Value	Numerical	from 0 to 14700000

The second dataset consists of 12 features, and it has 599 records.The dataset was normalized, then it was split into two partitions 70% for training, and 30% for testing. A detailed description of the features can be found below.

loan status (target) the person will take the loan or not . gender either male or female . dependents how many dependents he depend . married his miarrige status . education is he a student or not . self employment is he a employee or not .applicant income the applicant salary .co applicant income . loan amount the amount of money he want to loan . loan amount term the duration person will pay the loan . credit history does he pay latest loans or not . property area .

TABLE II
FEATURES OF DATASET 2

Feature	Type	Values
Loan Status	Classification	Y or N
Gender	Classification	Male or Female
Dependents	Numerical	0 to 3
Education	Classification	Graduate or Not Graduate
Self-Employed	Classification	Yes or No
Applicant Income	Numerical	From 150 to 81000
Coapplicant Income	Numerical	From 0 to 41667
Loan Amount	Numerical	From 9 to 650
Credit History	Classification	1 or 0
Property Area	Classification	Urban or Rural

The third and final data set consists of 34 features, and it has 67464 records. The data set was split into two partitions: 70% for training, and 30% for testing. A detailed description of the features can be found below.

loan status (target) the person will take the loan or not. loan amount ,funded amount, founded amount investor , term , batch enrolled , interest rate , grade ,sub grade ,employment duration , home ownership , verification status , payment plane , loan title , debit income , delinquency-two years , inquires-six months , open account , public record , revolving balance , revolving utilities , total accounts , initial list status , total received interest , total received late fee , recoveries collection recovery fee , collection 12 months medical , application type , last week pay , accounts delinquent , total collection amount , total current balance , total revolving credit limit .

TABLE III
FEATURES OF DATASET 3

Feature	Type	Values
loan status (target)	Classification	0 or 1
loan amount	numerical	from 1014 to 35000
funded amount	numerical	from 1014 to 34999
founded amount investor	numerical	from 1127.25 to 34999.7
term	numerical	from 36 to 59
batch enrolled	numerical	
interest rate	Numerical	from 5.32001 to 27.1823
grade	classification	A or B or C or D or E or F
sub grade	numerical	from A1 to F5
employment duration	classification	Mortgage or rent or own
home owner ship	numerical	from 14573.5 to 406562
verification status	Classification	not verified , verified
paymetn plan	Classification	N
loan title	Classification	debt consolidation or other
debit income	numerical	from 0.6753
delinquency - two years	Classification	0 or 1
inquires -six months	Classification	0 or 1 or 2 or 3
open account	numerical	from 2 to 37
public record	Classification	0 or 1 or 2 or 3
revolving balance	numerical	from 0 to 116933
revolving utilities	numerical	from 0.00517 to 100.859
total accounts	numerical	from 4 to 72
initial state	Classification	W or F
total received interest	numerical	from 4.73675 to 14301.4
total received late fee	numerical	from 0.000101 to 42.58806
recoveries	numerical	from 0.000221 to 4353.467
collection recovery fee	numerical	from 0.000144 to 166.833
collection 12 months medical	classification	0 or 1
application type	Classification	individual
last week pay	numerical	from 0 to 161
accounts delinquent	Classification	0 or 1
total collection amount	numerical	from 1 to 16421
total revolving balance	numerical	from 617 to 1177412
total revolving credit limit	numerical	from 1000 to 21169

B. *Used Algorithms*

The mentioned datasets were passed into 6 different Machine Learning algorithms which were Logistic Regression, Gradient Boosting, K Nearest Neighbor (k-NN), Random Forest, Decision Tree, and Naive Bayes. For each of the algorithms there were statistics generated, these statistics were: Accuracy, Recall, Precision, and F1-Score.

1) **Gradient Boosting:**
Gradient Boosting is an ensemble learning method that improves predictive performance by combining the outputs of weak learners, typically decision trees, in a sequential manner. It focuses on minimizing errors from previous models, enhancing overall accuracy through an iterative optimization process.

2) **Decision Tree:**
The supervised learning type includes the decision tree algorithm. Both regression and classification issues may be handled using them.Each node in the tree corresponds to a class label, with attributes expressed on the tree's

inner node. Any Boolean function with discrete characteristics may be described using the decision tree. The entropy varies when a node is employed in a decision tree, it breaks down the training dataset into smaller groupings. the increase in entropy is denoted by the information. Definition: Suppose S is a set of instances, A is an attribute, S_v is the subset of S with $A = v$, and $\text{Values}(A)$ is the set of all possible values of A, then

$$\text{Gain}(S, A) = \text{Entropy}(S) - \sum_{v \in \text{Values}(A)} \mid \frac{|S|}{S \mid} \cdot \text{Entropy}\,(S_v) \tag{1}$$

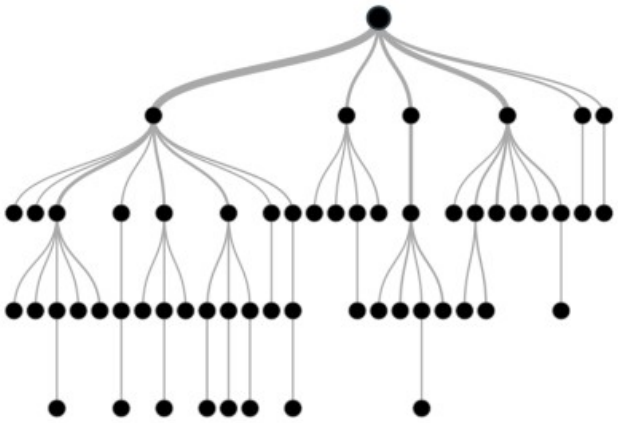


Fig. 4. Illustration of decision tree

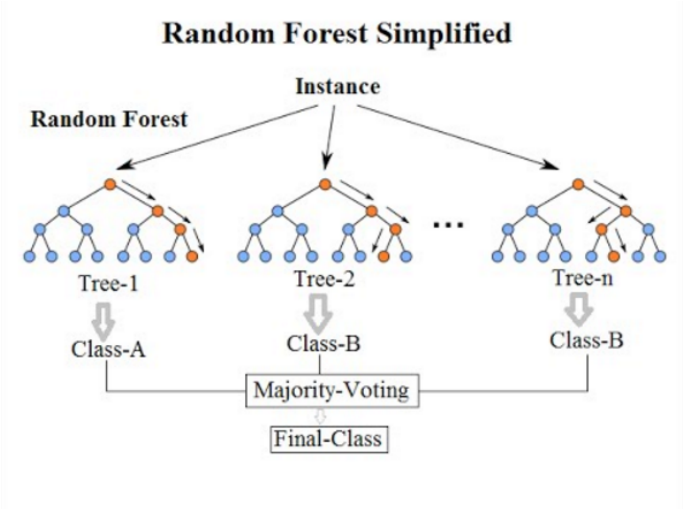
3) **Naïve Bayes:**
Built on the Bayes Theorem, Naive Bayes is a simple yet capable categorization algorithm. It presupposes predictor independence, which means that the traits or characteristics are not related to one another or connected in any way. Even though there is a dependence, each of these qualities or attributes contributes to the probability independantly, which is why it is termed Naive.

$$P(c \mid x) = \frac{P(x \mid c)P(c)}{P(x)} \tag{2}$$

$$P(c \mid X) = P\,(x_1 \mid c) \times P\,(x_2 \mid c) \times \cdots \times P\,(x_n \mid c) \times P(c) \tag{3}$$

4) **K – Nearest Neighbor:**
K-Nearest Neighbors (KNN) is a simple yet effective supervised machine learning algorithm used for classification and regression tasks. In KNN, a data point is classified or predicted based on the majority class or average of its k nearest neighbors in the feature space. The "k" refers to the number of neighbors considered. The algorithm is non-parametric and lazy, meaning it doesn't make assumptions about the underlying data distribution and defers computation until predictions are needed. KNN is intuitive and easy to understand, making it a popular choice for various applications.

5) **Random Forest:**
Random Forest is one of the supervised machine learning algorithms that can be used for both classification and regression tasks but, it works better in classification tasks. This algorithm considers multiple decision trees before giving an output. This technique is founded on the notion that a greater number of trees would ultimately guide to the rectified selection. It employs a voting approach for classification and then determines the class, whereas it uses the mean of all the decision tree outputs for regression. Random Forest Algorithm is extremely efficient with large datasets with high dimensionality.



6) **Logistic Regression:**
Logistic Regression is a widely used statistical method for binary and multiclass classification. Despite its name, it's used for classification rather than regression. The algorithm models the probability that a given instance belongs to a particular class. It employs the logistic function (sigmoid) to squash the output between 0 and 1, interpreting it as a probability.
In simpler terms, Logistic Regression calculates the odds of an instance belonging to a class and transforms these odds into probabilities. If the probability exceeds a certain threshold, the instance is classified into the corresponding class. It's a linear model with a logistic transformation, making it efficient, interpretable, and commonly used in various fields.[23].

C. *Performance Metrics*

Accuracy is the count of legitimately anticipated data from all the data. The count of accurately anticipated positives taken from the anticipated positives is the Precision Recall is the number of correctly anticipated

positives from all the true positives. The number of accurately anticipated negatives out of all the expected negatives is known as specificity.

$$\text{Accuracy} = (\text{TN}+\text{TP})/(\text{TN}+\text{TP}+\text{FN}+\text{FP}) \quad (4)$$

$$\text{Precision} = \text{TP}/\text{TP}+\text{FP} \quad (5)$$

$$\text{Recall} = \text{TP}/\text{TP}+\text{FN} \quad (6)$$

$$\text{F1-Score} = \frac{2 \times \text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \quad (7)$$

IV. Results and Analysis

The results collected from Gradient Boosting, Naïve Bayes, Logistic Regression, Random Forest, k-Nearest Neighbor, Decision Tree are shown below.

The following results are from the first dataset.

TABLE IV
STATISTICS OF ALGORITHMS

Model	Accuracy	Precision	Recall	F1-Score
Logistic Regression	0.628	0.333	0.002	0.003
Gradient Boosting	0.972	0.961	0.965	0.963
k-NN	0.560	0.386	0.312	0.345
Random Forest	0.977	0.971	0.967	0.969
Decision Tree	0.969	0.950	0.967	0.959
Naïve Bayes	0.784	0.884	0.481	0.623

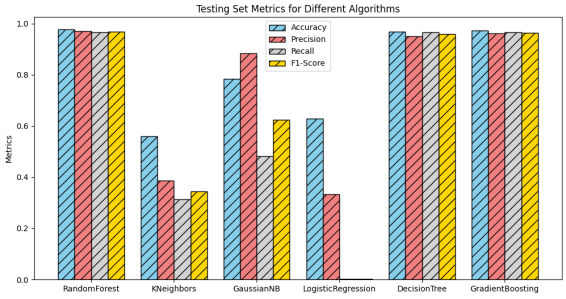


Fig. 5. First dataset performance chart with data split

Logistic Regression achieved an accuracy of 0.628, with precision at 0.333, recall at 0.002, and an F1-Score at 0.003.Gradient Boosting outperformed other models with an accuracy of 0.972. It demonstrated high precision (0.961), recall (0.965), and an F1-Score (0.963).k-NN had a lower accuracy of 0.560, with precision at 0.386, recall at 0.312, and an F1-Score at 0.345.Random Forest excelled with an accuracy of 0.977. It achieved high precision (0.971), recall (0.967), and an F1-Score (0.969).Decision Tree showed competitive performance with an accuracy of 0.969. It had precision at 0.950, recall at 0.967, and an F1-Score at 0.959. Naive Bayes demonstrated moderate performance with an accuracy of 0.784. It showed higher precision (0.884), lower recall (0.481), and an F1-Score (0.623).

The following results are from the second dataset.

TABLE V
STATISTICS OF ALGORITHMS

Model	Accuracy	Precision	Recall	F1-Score
Random Forest	0.800	0.815	0.930	0.869
k-NN	0.628	0.733	0.750	0.741
Naïve Bayes	0.556	0.761	0.547	0.636
Logistic Regression	0.817	0.810	0.969	0.883
Decision Tree	0.628	0.752	0.711	0.731
Gradient Boosting	0.739	0.779	0.883	0.829

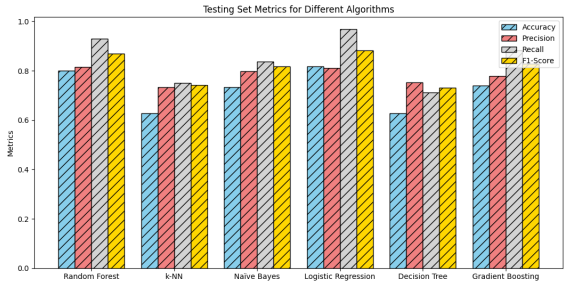


Fig. 6. Second dataset performance chart

The second dataset consistently demonstrated lower accuracy across all conditions compared to the first dataset. Notably, Logistic Regression and Gradient Boosting shared the same accuracy of 0.908 in both datasets, with Logistic Regression exhibiting slightly higher specificity in the second dataset. K-NN experienced a substantial drop in accuracy, registering at 0.628. Although Random Forest exhibited a reduced accuracy of 0.903 in the second dataset, it outperformed other algorithms and ranked relatively higher. This comparison highlights the varying impacts on algorithmic performance when datasets undergo different splits, emphasizing the importance of dataset characteristics in influencing model outcomes.

The following results are from the third dataset.

TABLE VI
STATISTICS OF ALGORITHMS

Model	Accuracy	Precision	Recall	F1-Score
Random Forest	0.904	0.069	0.003	0.862
k-NN	0.887	0.083	0.021	0.855
Naïve Bayes	0.577	0.094	0.41	0.665
Logistic Regression	0.907	1.000	0.000	0.863
Decision Tree	0.819	0.101	0.118	0.826
Gradient Boosting	0.907	1.000	0.000	0.863

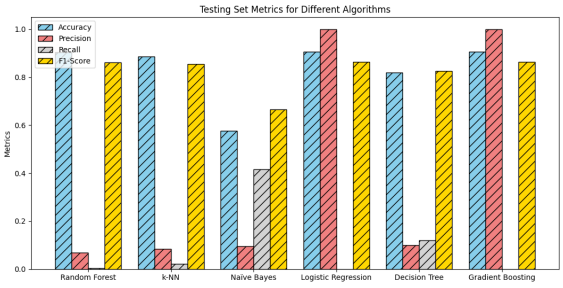


Fig. 7. Thrid dataset performance chart

In the context of the third dataset, the provided table outlines the algorithmic statistics, revealing diverse performance metrics for each model. Notably, Logistic Regression and Gradient Boosting exhibit strikingly high accuracies of 0.907, accompanied by perfect precision, yet with a recall of 0.000, emphasizing a potential imbalance in their predictions. Random Forest achieves a solid accuracy of 0.904 with a considerable F1-Score of 0.862, suggesting a balanced trade-off between precision and recall. K-NN demonstrates a relatively high accuracy of 0.887, but its precision and recall metrics indicate room for improvement. Naïve Bayes, while achieving a lower overall accuracy of 0.577, showcases a noteworthy recall of 0.410, indicating its efficacy in identifying positive instances. Conversely, Decision Tree, with an accuracy of 0.819, presents balanced precision and recall, reflecting its competence in classification tasks. The diverse performance profiles underscore the nuanced strengths and weaknesses of each algorithm in handling the specific characteristics of the third dataset.

V. Conclusion

Finally, in the first dataset, the larger size may have contributed to the success of more complex models like Gradient Boosting and Random Forest. The second dataset showed that simpler models like Logistic Regression can perform well with a smaller dataset, while Random Forest maintained a good balance. The third dataset, being significantly larger, influenced the performance of Logistic Regression and Gradient Boosting but also highlighted potential challenges in handling certain aspects of the data. Machine Learning can be applied to many fields, not just financial, it can be used to predict anything from stock prices to results of sports matches which makes it a very useful tool for humanity. And this tool will only keep improving and producing better results.

VI. Acknowledgment

First and foremost, we express our deep gratitude to the dedicated staff at Misr International University and, in particular, the Faculty of Computer Science, for their unwavering commitment to the success of this institution. Our heartfelt thanks go to Dr. Ayman Nabil, Dean of the Faculty of Computer Science, for providing us with the invaluable opportunity to pursue our education in this esteemed university and for overseeing its seamless operation.

Special appreciation is extended to Dr. Diaa AbdelMoneim, for his consistent guidance and unwavering support, especially in our studies related to the AI course. His expertise and encouragement have played a crucial role in our academic journey at Misr International University.

References

[1] M. A. Sheikh, A. K. Goel, and T. Kumar, "An approach for prediction of loan approval using machine learning algorithm," in *2020 International Conference on Electronics and Sustainable Communication Systems (ICESC)*. IEEE, 2020, pp. 490–494.

[2] B. Mahesh, "Machine learning algorithms-a review," *International Journal of Science and Research (IJSR).[Internet]*, vol. 9, pp. 381–386, 2020.

[3] K. Arun, G. Ishan, and K. Sanmeet, "Loan approval prediction based on machine learning approach," *IOSR J. Comput. Eng*, vol. 18, no. 3, pp. 18–21, 2016.

[4] J. Tejaswini, T. M. Kavya, R. D. N. Ramya, P. S. Triveni, and V. R. Maddumala, "Accurate loan approval prediction based on machine learning approach," *Journal of Engineering Science*, vol. 11, no. 4, pp. 523–532, 2020.

[5] A. S. Kadam, S. R. Nikam, A. A. Aher, G. V. Shelke, and A. S. Chandgude, "Prediction for loan approval using machine learning algorithm," *International Research Journal of Engineering and Technology (IRJET)*, vol. 8, no. 04, 2021.

[6] S. M. Fati, "Machine learning-based prediction model for loan status approval," *Journal of Hunan University Natural Sciences*, vol. 48, no. 10, 2021.

[7] P. S. Murthy, G. S. Shekar, P. Rohith, and G. V. V. Reddy, "Loan approval prediction system using machine learning," *Journal of Innovation in Information Technology*, vol. 4, no. 1, pp. 21–24, 2020.

[8] H. S. Sandhu, V. Sharma, and V. Jassi, "Loan approval prediction using machine learning."

[9] S. Bhattad, S. Bawane, S. Agrawal, U. Ramteke, and P. Ambhore, "Loan prediction using machine learning algorithms." *International Journal of Computer Science Trends and Technology*, vol. 9, no. 3, pp. 143–146, 2021.

[10] A. S. Aphale and S. 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, no. 8, 2020.

[11] Y. T. Sheng, N. S. A. Rani, and J. M. Shaikh, "Impact of smes character in the loan approval stage," *Business and Economics Research*, vol. 1, pp. 229–233, 2011.

[12] K. Arun, G. Ishan, and K. Sanmeet, "Loan approval prediction based on machine learning approach," *IOSR J. Comput. Eng*, vol. 18, no. 3, pp. 18–21, 2016.

[13] J. Tejaswini, T. M. Kavya, R. D. N. Ramya, P. S. Triveni, and V. R. Maddumala, "Accurate loan approval

prediction based on machine learning approach," *Journal of Engineering Science*, vol. 11, no. 4, pp. 523–532, 2020.

[14] A. Khan, E. Bhadola, A. Kumar, and N. Singh, "Loan approval prediction model a comparative analysis," *Advances and Applications in Mathematical Sciences*, vol. 20, no. 3, 2021.

[15] S. F. Eletter, S. G. Yaseen, and G. A. Elrefae, "Neuro-based artificial intelligence model for loan decisions," *American Journal of Economics and Business Administration*, vol. 2, no. 1, p. 27, 2010.

[16] H. M. Sami, M. Rafatuzzaman, and A. Bar, "Machine learning application for selecting efficient loan applicants in private banks of bangladesh," *Int. J. Manag. Account*, vol. 3, no. 5, pp. 114–121, 2021.

[17] S. Dosalwar, K. Kinkar, R. Sannat, and N. Pise, "Analysis of loan availability using machine learning techniques," *International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)*, vol. 9, no. 1, pp. 15–20, 2021.

[18] M. Anand, A. Velu, and P. Whig, "Prediction of loan behaviour with machine learning models for secure banking," *Journal of Computer Science and Engineering (JCSE)*, vol. 3, no. 1, pp. 1–13, 2022.

[19] P. Dutta, "A study on machine learning algorithm for enhancement of loan prediction," *International Research Journal of Modernization in Engineering Technology and Science*, vol. 3, 2021.

[20] S. Sreesouthry, A. Ayubkhan, M. M. Rizwan, D. Lokesh, and K. P. Raj, "Loan prediction using logistic regression in machine learning," *Annals of the Romanian Society for Cell Biology*, pp. 2790–2794, 2021.

[21] A. Bhagat *et al.*, "Predicting loan defaults using machine learning techniques," Ph.D. dissertation, California State University, Northridge, 2018.

[22] L. Zhu, D. Qiu, D. Ergu, C. Ying, and K. Liu, "A study on predicting loan default based on the random forest algorithm," *Procedia Computer Science*, vol. 162, pp. 503–513, 2019.

[23] T. G. Nick and K. M. Campbell, "Logistic regression," *Topics in biostatistics*, pp. 273–301, 2007.