ISSN (Online) 2456 -1304

International Journal of Science, Engineering and Management (IJSEM)

Vol 11, Issue 7, July 2024

# Comparative Analysis of Machine Learning Algorithms for Credit Risk Assessment: Identifying the Optimal Model

[1] Dr. Raman Chawla, [2] Kunal Chawla, [3] Tushar Sharma

[1] Professor, Rajiv Academy for Technology & Management, Uttar Pradesh, India
[2] Apex Institute of Technology, Chandigarh University, Punjab, India
[3] Department of Computer Science and Engineering, Lovely Professional University, Punjab, India
Corresponding Author Email: [1] rchawla3@gmail.com, [2] kcchawla85@gmail.com, [3] tushar.rk1990@gmail.com

***Abstract*— The process of determining the possibility that a borrower would miss payments on a loan or meet a contractual commitment is known as credit risk analysis. For lenders, investors, and other financial institutions to make well-informed choices regarding loan extension or investment in a certain company, this study is essential. Analyzing credit risks and loan repayments is one of the biggest challenges that the modern world faces. There are a lot of defaulters in the world in different loan types and variations. According to a recent study conducted by CNBC in 2023 stated that there is an increase in the percentage of defaulters in India to 32.9%. The financial stability report (FSR) of the Reserve Bank of India (RBI) states that the gross non-performing assets (NPA) of public sector lenders in the credit card category was 18%, whilst private sector banks recorded a GNPA of 1.9 percent in FY23. According to a recent report from S&P Global the loan defaults in the U.S.A markets can rise to 3% by September 2024. This creates a demand for tools which can help big banks to grant loans to individuals or companies who have a good credit score. This research paper aims at providing an answer to the question of which machine learning algorithm will be best to perform such kind of predictions and can be used in the future by different credit risk analysis tools.***

***Index Terms: Credit risk, loan repayment default, machine learning algorithms, supervised learning, unsupervised learning.***

## I. INTRODUCTION

The word ‘credit risk analysis’ is used interchangeably with ‘Credit evaluation’, ‘Default risk analysis’, ‘Credit Scoring’ and ‘Credit Profiling’ which means the process of determining the possibility that the borrower would miss payments on a loan or meet a contractual commitment. After the great mortgage crisis of 2007 which lead to the global stock market collapse in 2008 created a high urgency to rely over credit profiling to find out whether a person or an organization should be granted loan or not or whether they will be able to replay it or not. According to a recent report by McKinsey that by 2025 the risk management functions are expected to have a substantial transformation. Changes in client expectations, the growth and modification of legislation, and the evolving nature of risks are the driving forces behind this transition. With the use of cutting-edge analytics and evolving technology, creating new products, services, and risk management strategies is getting easier. Machine learning is widely recognized as a vital technology for risk management, and it plays a significant role in developing more accurate risk models by recognizing complex, non-linear patterns in large data sets. These models can constantly improve their ability to predict outcomes as new data is supplied. In the recent years, Machine learning and Artificial Intelligence based algorithms have played a significant role in transforming industries like technology, software and automobile, etc. AI & ML has been utilized in almost every human activity such as pattern recognition, image classification, autonomous driving, agriculture, etc. This research paper focuses over the scope of machine learning algorithms in the sector of finance. This research paper aims to identify the best machine learning algorithm that can be used credit profiling. Previously, the credit risk assessments were done based on very generic algorithms or by using manual paper work. Some of the algorithms which were used in the past years were based on statistical methods such as Logistic Regression, Linear Discriminant Analysis (LDA) or Linear Regression. The only problem with such kind of algorithms is that it cannot handle large datasets. There are different machine learning based algorithms that are far better that these statistical methods and can provide better results. Some of the algorithms that are considered in this research are Decision Tree Classification, Support Vector Classifiers, K-Nearest Neighbours Classification, Gaussian Naïve Bayes Classification, Linear Discriminant Analysis, Logistic Regression and Random Forest Classification.

This research paper aims to provide a systemic review of the credit risk analysis algorithms. It focuses over statistical and classification based techniques. The final aim is to identify which machine learning algorithm can be a best fit for credit risk assessments. This research tries to bridge a gap between traditional way of banking with the modern machine learning algorithms and market trends which can assist in taking better data-driven decisions so that we can stop a crisis such as the one which happened in the 2008.

The rest of the research paper comprises of the following sections : Section II is the literature survey which will put light of the past work done in the area of machine learning and finance, Section III is the inference from the related work which is done in the past and how this research paper is building over it. Section IV is Methodology where what variables have been considered and the information regarding the dataset used is present along with what steps were taken into consideration while doing the comparison. Section V discusses the results after the comparison and provides the answer about the perfect algorithm for credit risk assessment. Section VI represents the conclusion drawn from the comparison and the future aspects of the research. Section VII represents the references form all the research paper that were considered while performing this comparison.

## II. LITERATURE SURVEY

In the financial field, credit risk and default likelihood have been the subject of much study throughout the years. Understanding and forecasting credit risk is essential for lenders in their decision-making processes. Credit risk is defined as the loss to creditors resulting from debtors' default on credit commitments. In order to accept or further examine cases, researchers Sakprasat and Sinclair (2007) stress the significance of credit evaluation in the early phases. Concerns are raised, nonetheless, by the dependence on credit rating organizations for these evaluations. According to Bolton, Freixas, and Shapiro (2012), credit ratings are artificial and could not truly reflect risk because of possible inflation brought on by market forces. Furthermore, as noted by Bolton et al. (2012), conflicts of interest between agencies and their customers compromise the validity of credit scores for investors. The technique Yu and Zhu (2015) suggested, which relies on synthetic ratings rather than actual application data, was criticized for utilizing credit scores as independent factors in forecasting default risk. This emphasizes the shortcomings of credit ratings, which are biased and intrinsically subjective (Bolton, Freixas, & Shapiro, 2012).

Pan and Singleton (2008) proposed an alternate method that more closely reflects market perceptions: credit risk is inferred using credit default swaps (CDS) spreads. This is corroborated by Longstaff, Mithal, and Neis (2005), who point out the tight relationship between CDS and market assessments. Luo et al. (2017) showed the promise of innovative methods made possible by big data and advancements in computing power by using machine learning techniques on CDS data to categorize credit ratings (Kaastra & Boyd, 1996; Renault, 2017). A departure from conventional statistical techniques is represented by the integration of machine learning into credit risk analysis, which allows computers to examine massive datasets like those from peer-to-peer lending platforms like Lending Club (Kaastra & Boyd, 1996; Renault, 2017). This suggests a viable path forward for financial research, utilizing technology to improve credit risk evaluation and decision-making procedures.

Two popular statistical methods for assessing credit risk are logistic regression and linear discrimination analysis (LDA). One of the first credit scoring techniques was LDA, a parametric model that was criticized for its uneven covariance matrices and categorical credit data (West, 2000). Logistic regression was first used for credit scoring by Henley (1995), providing probability for binary outcomes based on predictor factors. Research indicates that logistic regression performs better than other standard credit rating systems (West, 2000). Although neural networks are receiving more attention, it is still unclear if they are superior than logistic regression (Abellan & Castellano, 2017). Neural network rule extraction strategies were found by Baesens, Setiono, Mues, and Vanthienen (2003) to be competitive with decision tree and logistic regression models. According to Yap, Ong, and Husain (2011), no model performs better than any other. The longevity of logistic regression and LDA in credit risk analysis is attributed to its accuracy and simplicity, as noted by Finlay et al. (2012) and Lessman et al. (2015).

In machine learning, Support Vector Machine (SVM) is a popular classification algorithm with clear advantages over alternative approaches. When SVM was first presented in 2010 by Yu, Wuyi, Shouyang, and Lai, it made less assumptions about the distributions of the input variables and allowed for nonlinear mapping, which made it adaptable to a variety of datasets. By discovering the separating hyperplane during maximizing, it does this. Because SVM may avoid local minima, it outperforms alternatives like fuzzy neural networks in single-agent settings, according to a research by Yu et al. (2010). Furthermore, it was discovered that multi-agent models improved accuracy in comparison to single-agent models. Burgers (1998) confirmed the effectiveness of SVM by pointing out that it performs better or on par with other approaches in a variety of applications. Among the 17 evaluated techniques, Baesens et al. (2003) showed that SVM dominated credit scoring, confirming its high accuracy rate. In their evaluation of SVM against backpropagation neural networks for credit risk assessment, Huang, Chen, Hsu, Chen, and Wu (2004) found that SVM had marginally better results. Later studies (Huang, Chen, & Wang, 2007) compared SVM to decision trees, neural networks, and genetic programming and found that even with fewer input variables, classification accuracy remained comparable. SVM models for credit scoring were improved by Hens and Tiwari (2012), who also achieved competitive accuracy rates and computational efficiency. Yao and Lu (2011) found that SVM outperformed linear discriminant analysis, logistic regression, and neural networks in their credit rating when paired with neighborhood rough sets.

Classification trees, or decision trees, provide a quick and easy-to-understand classification technique. They work especially well with datasets that have low variation since they allow for meaningful model differences (Tsymbal, Pechenizkiy, & Cunninghan, 2005). Based on predetermined guidelines and a particular objective variable, these trees divide large datasets into more manageable, homogeneous groupings (Yap, Ong, & Husain, 2011). Credit card scoring using decision trees and multilayer perceptron neural networks was investigated by Davis, Edelman, and Gammerman (1992), and similar accuracy levels were found. According to Tap & Ong (2011), the classification error rates for credit scorecard, decision tree, and logistic regression models are 27.9%, 28.1%, and 28.8%, respectively. German, Australian, and Japanese credit datasets were investigated by Zhao et al. (2015), who found that decision trees marginally outperformed backpropagation. SAfter comparing decision trees, neural networks, k-nearest neighbor, and probit algorithms, Galindo & Tamayo (2000) concluded that decision trees were better for default prediction. A dual strategy ensemble tree was proposed by Wang, Ma, Huang, and Xu (2012), which improved classification accuracy by reducing noise and redundancy. Decision trees are a prominent paradigm for Bagging ensemble systems in scoring issues, according to Abellan & Castellano (2017).

Pande et al. (2018) used machine learning classifiers to do a credit risk analysis. Using the German credit risk dataset, they investigated techniques such as Artificial Neural Network (ANN), k-Nearest Neighbor, and Naive Bayes (NB) in their study. According to their results, the accuracy of ANN, NB, and KNN was 77.45%, 77.20%, and 72.20%, respectively. They did not, however, use other measures like the F1-Score and the Area Under the Curve (AUC) score to assess their models. An adaptive support vector machine (AdaSVM)-based credit rating technique was introduced by Zhang et al. (2019). Using the Australian credit risk dataset, they evaluated this technique and found an 80% accuracy rate. They did not examine the use of measures like accuracy and recall to assess the quality of the categorization. Gradient Descent and Artificial Neural Networks (ANNs) were used in the development of a consumer credit risk assessment system by Nasser and Maryam (2020). They achieved accuracies of 78.11%, 76.87%, and 68.26%, respectively, using German, Australian, and Japanese credit risk datasets. Using the Taiwan credit risk dataset, Hsu et al. (2021) developed an improved recurrent neural network (RNN) for credit card default prediction. With the addition of Gated Recurrent Units (GRUs), their improved RNN produced a lift index of 0.659 and an AUC of 0.782. Researchers offered a combined approach combining supervised and unsupervised learning for credit risk assessment in a paper that was published in 2022. They reported KNN accuracy of 76.80% and AUC of 0.788, RF accuracy of 72.10% and AUC of 0.811, and ANN accuracy of 78.6% and AUC of 0.811 using the German dataset. Ha et al. (2023) used deep learning (DL) and the feature selection (FS) approach to create an enhanced credit risk prediction model for online peer-to-peer lending systems. They employed a range of machine learning techniques on German and Australian credit risk datasets, such as ANN, KNN, RF, and linear discriminant analysis (LDA). LDA, ANN, KNN, and RF obtained accuracies of 76.50%, 75.8%, 67.10%, and 67.72% for the German dataset, respectively. LDA, ANN, KNN, and RF obtained accuracies of 85.80%, 71.45%, 65.94%, and 67.72%, respectively, for the Australian dataset. They did not, however, take into account other parameters including accuracy, recall, and AUC.

## III. INFERENCE FROM RELATED WORK

The suggested work done by writers in the topic of credit risk analysis during the past ten years is highlighted in the literature review above. The main focus of the research is on combining several algorithms to forecast a person's eligibility for a loan with the maximum accuracy possible. To far, no noteworthy study has been published that compares the many algorithms available to estimate an individual's or company's credit risk rating. Research comparing the accuracy of each and every algorithm and determining which one has the highest potential accuracy to forecast credit risk is desperately needed. The insights gained from this study will help aspiring writers and scholars in the future. Future writers and researchers who are attempting to design an extremely accurate algorithm to choose a suitable algorithm for their research will find valuable insights from this study. The dataset included in this study consists of both individual and corporate loan defaulters as well as those who have never failed on a loan. To provide fair comparison findings, the authors have established the identical settings for all the algorithms utilized in this research. The study paper's results and discussion part includes a discussion of all the findings and conclusions.

## IV. METHODOLOGY & ALGORITHMS USED

The primary objective of this research is to compare several machine learning algorithms in order to determine which algorithm produces predictions for the credit risk analysis with the highest accuracy. First, we wanted to make sure that the algorithms could use the data and provide results without any issues by cleaning and transforming it. Next, we divided the data into distinct training and testing parts. Following that, we ran it via machine learning techniques to obtain results. A table was created by summarizing the findings and presenting them. The following contains more detailed information about the procedure and the algorithms that were employed:

1. ***Preparing the Dataset:*** A variety of resources were utilized to gather the dataset that is the subject of this comparison. The data set had over 900 rows and over 13 columns including various factors that might impact the credit prediction. The information contains a variety of methods by which the loan was granted or denied for the

individual or business. The data set was further separated into two main sections. To build and train a model, the training data, which included over 600 rows and over 13 columns, made up the first portion of the data set. The second is a testing dataset that was constructed in order to test the model, and it had about 300 rows and 13 columns**.** The chart representation of the dataset is showcased in the Fig 1. The variables that were considered in order to predict the credit risk are Gender, Marrital Status, Number of Dependents, Educational Qualification of the applicant, Employment Status, Income earned by the applicant, Co applicant Income, Loan Amount, Loan Amount duration, Credit History and the Property acquired by the applicant.

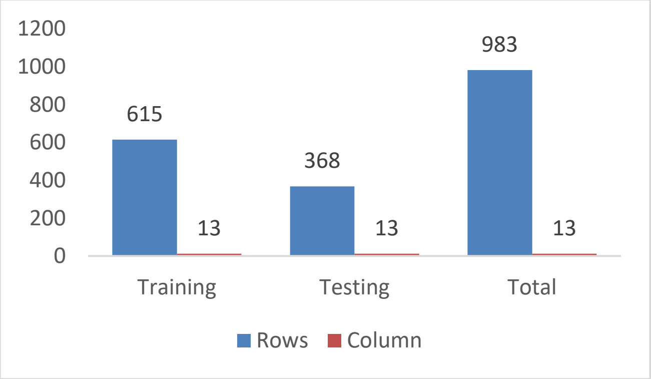


**Fig 1.** Data description of the dataset

2. ***Selecting Algorithms:*** Different Algorithms were taken into consideration to compare them by using the data and those algorithms are as follows:

- *Gaussian Naïve Bayes:*Based on Baye's Theorem, Gaussian NB is a very basic probabilistic classifier and a variation of Naïve Bayes. It is assumed in this approach that every feature has a normal distribution. It is considered that the chance of a characteristic falling into a given class is bell-shaped. It is predicated on the idea that each parameter may independently predict the output and the dependent variable's likelihood of being categorized into each category. The total read under the curve for the variable (X), which has a normal distribution from negative infinity to positive infinity, is 1.
- *Logistic Regression:*When it comes to binary classification problems, the basic objective of the logistic regression model (LR) is to predict the likelihood of a binary result. Instead of being a regression algorithm, it is a classification algorithm. The output of logistic regression is the logistic function, which converts the linear combination of input feature values into a value between 0 and 1. The likelihood of an event occurring given the input attributes is then used to understand the result. In Logistic Regression, maximum likelihood is employed for parameter estimation. Because of its ease of use, effectiveness, and interpretability, this method is frequently employed. frequently employed as a starting point model in binary classification issues and acts as the basis for more intricate algorithms such as neural networks.
- *Decision Tree:* It is a supervised machine learning approach that may be applied to regression and classification problems. This predictive modeling program divides the data into subsets recursively according to the provided feature values. In order to optimize the purity of the subsets, the algorithm chooses the characteristic at each stage that best divides the data into homogenous subsets. In order to optimize the impurity or information gain at each split, the data is divided depending on characteristics in this instance. Entropy and Gini impurity are a couple of the often used impurity metrics. This procedure keeps on till a requirement is satisfied.
- *Random Forest:* It is a popular ensemble learning approach that may be used for both regression and classification applications. During the training phase, many decision trees are created, and their predictions are aggregated to make it. Bootstrapping is the first step in the process; it involves taking random samples from the original dataset and replacing them to create numerous subsets. A decision tree is created for each and every subgroup, and variety is added to the trees by considering a random subset of characteristics at each split. Following construction, each tree makes its own predictions; for classification tasks, the mode of the forecasts is considered the final result.
- *Support Vector Machine:* SVM is a potent supervised machine learning model that may be applied to regression as well as classification. Its main application is in determining which hyperplane best divides the data points into distinct groups. Finding the hyperplane that maximizes the margin—the distance between the hyperplane and the closest data points from each class—is how support vector machine learning (SVM) operates. In addition to 4separating the data, this ideal hyperplane optimizes the margin and aids in enhancing the model's capacity for generalization, which reduces sensitivity. Bioinformatics, text classification, picture classification, and other fields make extensive use of them.
- *LDA:* Linear discriminant analysis (LDA), often referred to as normal discriminant analysis (NDA) or discriminant function analysis (DFA), was first presented by Ronald A. Fisher in 1936. Within a generative model framework, LDA models the distribution of data for every class. In order to categorize new data points, LDA computes conditional probabilities using Thomas Bayes' theorem, which was first presented in 1763. This approach, which makes use of Bayes, forecasts the probability that input data will correspond to particular outputs. LDA simplifies the process of classifying multidimensional data by projecting it into a single dimension and recognizing

linear combinations of characteristics. This method makes it easier to apply in multi-class data classification challenges; it is similar to dimensionality reduction. Because of its versatility, LDA may be used to improve the performance of various classification techniques, including support vector machines, decision trees, and random forests.

- *KNN:* The k-nearest neighbors (KNN) method is a supervised learning classifier that is non-parametric and was first presented by Evelyn Fix and Joseph Hodges in 1951. One of the most straightforward and well-liked classifiers in machine learning, it uses proximity to categorize or predict how data points will be grouped. Although it may be used for regression, its main application is in categorization, presuming that similar points cluster together. It uses a majority vote system for categorization. Literature frequently refers to what is technically known as "plurality voting" as "majority vote." Notably, the latter is appropriate for binary classes and requires more than 50% majority. But in cases when there are several classes, a vote that is more than 25% might be enough to allocate, as explained by Tom Mitchell in 1997. Because of its efficiency and ease of use, KNN is a vital component of machine learning processes.

3. ***Data Preprocessing:*** Several libraries were imported, including matplotlib, pandas, and numpy. Numpy is a mathematical function library that supports massive, multi-dimensional arrays and matrices and is used in numerical calculations. Pandas is used for analysis and data processing. It offers functions and data structures that improve the efficiency of dealing with structured data. For data visualization, Matplotlib is used to create static, interactive, and publication-quality plots and charts. The data was read using pandas, and column names were added. We also counted the total number of rows and columns in each dataset, including the test and train dataset.
4. ***Data Transformation:*** Imported is the Seaborn library, a statistical data visualization package that builds upon Matplotlib to provide more features and eye-catching graphics. To convert category labels into numerical values, a label encoder was loaded. To identify patterns, trends, and correlations between the variables, label encoding was applied to the data frame, followed by the creation and visualization of a correlation matrix and the plotting of a heatmap. The label encoding was done as follows the Gender field was mapped as 1 for Male and 0 for female, the marital status was mapped as 1 for yes and 0 for not married, the number of dependents was mapped as 0 for zero dependents, 1 for one dependent, 2 for two dependents and 3 for three or more dependents, the education status was mapped as 1 for graduated and 0 for not graduated, for property area the mapping was done as 2 for Urban, 0 for Rural and 1 for semi urban.

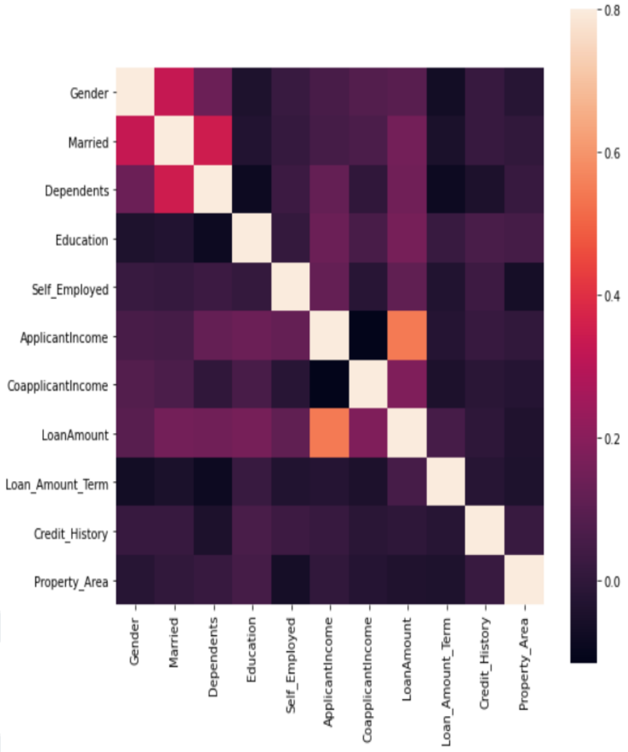


**Fig 2:** Heatmap for the dataset

A numerous strategies for data transformation were used, such as removing unnecessary columns to lower the model's dimensions, computational cost, and interpretability. Binary classification was applied to the data in the data frame to improve the model's assessment, performance, compatibility, and ease of use. Various data transformation methods were employed to eliminate any type of missing values from the dataset, including When it came to the loan amount, a different method was used, which involved filling in the median of the data and adding the gender values. The credit history was filled in with a random number between 0 and 2, and the marital status was filled in with a random integer between 0 and 1. The dependents were determined by taking the median of the remaining cases, and the job status was likewise filled in based on randomization, with the blank fields randomly filled in as 0 or 1.

5. ***Splitting Data & Applying the Models:*** The data was splitted further into 2 major parts in order to test out different models over different range of dataset and the split is termed as X_train, X_test, y_train, y_test After this all the models which were taken into consideration were applied over the data and their accuracy, precision, recall and f1-score was compared.
6. ***Analyzing the Results****:* F1-score, accuracy, precision, recall, were among the important performance indicators used to characterize the algorithms' outcomes. These measures can be interpreted as follows when used in the context of credit risk analysis. The percentage of correctly anticipated default instances, or bad loans, relative to the

total number of loans forecasted as defaults is known as precision (P)(1.1) in credit risk analysis. It assesses how well the algorithm detects hazardous loans without mistakenly classifying non-dangerous loans as risky. The percentage of correctly anticipated default instances (bad loans) in relation to the total number of default cases in the dataset is called recall(R)(1.2). Recall in credit risk analysis shows how well the model can identify all high-risk loans, minimizing the number of defaults that might be overlooked. The harmonic mean of recall and accuracy is known as the F1-score(1.3). It offers a fair assessment of a model's effectiveness by taking both recall and accuracy into account. A model that successfully detects hazardous loans while avoiding false positives and false negatives is indicated by a high F1-score in credit risk analysis. The ratio of accurately anticipated loan outcomes (defaults and non-defaults) to the total number of loans in the dataset is known as accuracy (A)(1.4) in credit risk analysis. Accuracy is a crucial parameter, but in unbalanced datasets where the proportion of defaults to non-defaults is large, accuracy might not be enough on its own. For this reason, in addition to accuracy, precision, recall, and F1-score must be taken into account in order to fully assess the model's performance in credit risk analysis.

**Table I:** Confusion Matrix

| | ***Predicted Positive*** | ***Predicted Negative*** |
|---|---|---|
| ***Positive Instance (P)*** | ***TP (True Positive)*** | ***FN (False Negative)*** |
| ***Negative Instance(N)*** | ***FP (False Positive)*** | ***TN (True Negative)*** |

$$P = \frac{TP}{TP+FP} \quad (1.1)$$

$$R = \frac{TP}{TP+FN} \quad (1.2)$$

$$F1 = 2\,x\,\frac{R\,x\,P}{R+P} \quad (1.3)$$

$$A = \frac{TP+TN}{TP+TN+FP+FN} \quad (1.4)$$

## V. RESULTS & DISCUSSION

Here are the findings from several observations made during the trials conducted on all of these algorithms. Based on these factors, we are summarizing which algorithm is more accurate in determining a person's or a business's credit risk. The applicant's credit history, credit history, gender, marital status, number of dependents, educational background, employment status, income, co-applicant income, loan amount, loan length, and property bought are all taken into consideration. The designed experiment was evaluated on a genuine data set of authorized and defaulted real loans from several institutions. There are 13 columns and around 900 rows in the data collection. Section IV of the considerations states that the data was divided into two primary categories: training and testing. There are 13 columns and around 900 rows in the data collection. As per Section IV of the considerations, the data was divided into two primary training and testing datasets to evaluate the quality of classification and replicate real-world scenarios where a greater amount of data may be utilized for prediction rather than for learning.

Several traditional machine learning techniques were examined, including support vector machines, Gaussian Naïve Bayes, logistic regression, decision trees, random forests, and K closest neighbors. Based on metrics like the Confusion Matrix, accuracy, precision, F1-score, and recall, the algorithms' quality is displayed. The Python programming language was utilized to implement all of the algorithms in this experiment, along with several libraries including numpy, pandas, matplotlib.pyplot, seaborn. Every operation is conducted using a Jupyter Notebook. In the tables that will follow, 0 denotes rejection and 1 approval.

The Gaussian Naïve Bayes has the following average of all training and testing sets:

**Table II:** Gaussian NB Average of all Training and Testing set

| | **Precision** | **Recall** | **F1-Score** | **Accuracy** |
|---|---|---|---|---|
| **0** | 0.63 | 0.51 | 0.56 | 0.78 |
| **1** | 0.82 | 0.88 | 0.85 | |

The Logistic Regression has the following average of all training and testing sets:

**Table III:** Logistic Regression Average of all Training and Testing set

| | **Precision** | **Recall** | **F1-Score** | **Accuracy** |
|---|---|---|---|---|
| **0** | 0.91 | 0.47 | 0.62 | 0.84 |
| **1** | 0.83 | 0.98 | 0.90 | |

The Decision Tree has the following average of all training and testing sets:

**Table IV:** Decision Tree Average of all Training and Testing set

| | **Precision** | **Recall** | **F1-Score** | **Accuracy** |
|---|---|---|---|---|
| **0** | 0.50 | 0.56 | 0.53 | 0.72 |
| **1** | 0.82 | 0.78 | 0.80 | |

The Random Forest has the following average of all training and testing sets:

**Table V:** Random Forest Average of all Training and Testing set

| | **Precision** | **Recall** | **F1-Score** | **Accuracy** |
|---|---|---|---|---|
| **0** | 0.62 | 0.47 | 0.53 | 0.77 |
| **1** | 0.81 | 0.89 | 0.85 | |

The Linear Discriminant Analysis has the following average of all training and testing sets:

**Table VI:** LDA Average of all Training and Testing set

| | Precision | Recall | F1-Score | Accuracy |
|---|---|---|---|---|
| **0** | 0.67 | 0.51 | 0.58 | 0.79 |
| **1** | 0.83 | 0.90 | 0.86 | |

The SVM has the following average of all the training and testing tests:

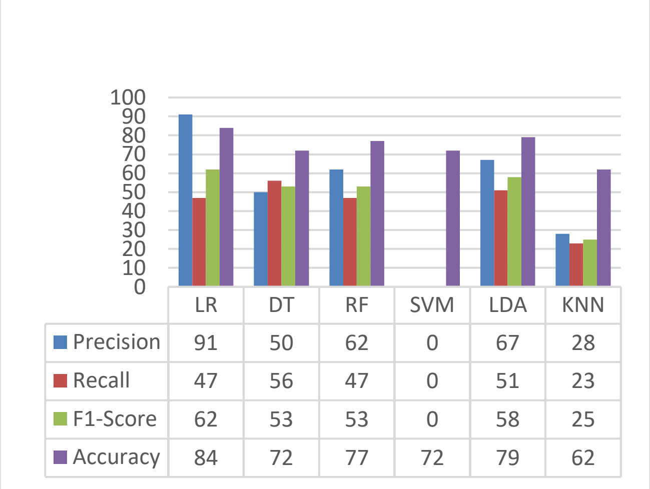
**Table VII:** SVM Average of all Training and Testing set

| | Precision | Recall | F1-Score | Accuracy |
|---|---|---|---|---|
| **0** | 0.00 | 0.00 | 0.00 | 0.72 |
| **1** | 0.72 | 1.00 | 0.84 | |

The K Nearest Neighbors has the following average of all the training and testing tests:

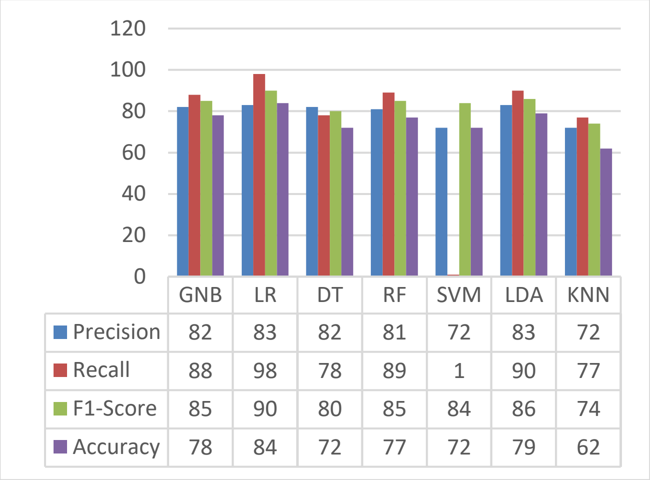
**Table VIII:** KNN Classifier Average of all Training and Testing set

| | Precision | Recall | F1-Score | Accuracy |
|---|---|---|---|---|
| **0** | 0.28 | 0.23 | 0.25 | 0.62 |
| **1** | 0.72 | 0.77 | 0.74 | |



| | LR | DT | RF | SVM | LDA | KNN |
|---|---|---|---|---|---|---|
| 🟦 Precision | 91 | 50 | 62 | 0 | 67 | 28 |
| 🟥 Recall | 47 | 56 | 47 | 0 | 51 | 23 |
| 🟩 F1-Score | 62 | 53 | 53 | 0 | 58 | 25 |
| 🟪 Accuracy | 84 | 72 | 77 | 72 | 79 | 62 |

**Fig 3.** Correctly Rejected Credit Results



| | GNB | LR | DT | RF | SVM | LDA | KNN |
|---|---|---|---|---|---|---|---|
| 🟦 Precision | 82 | 83 | 82 | 81 | 72 | 83 | 72 |
| 🟥 Recall | 88 | 98 | 78 | 89 | 1 | 90 | 77 |
| 🟩 F1-Score | 85 | 90 | 80 | 85 | 84 | 86 | 74 |
| 🟪 Accuracy | 78 | 84 | 72 | 77 | 72 | 79 | 62 |

**Fig 4.** Correctly Approved Credit Results

Following comprehensive evaluation and analysis, it was shown that Gaussian Naïve Bayes, Linear Discriminant Analysis, and Logistic Regression with correctly authorized credit had the highest accuracy (84%, 79%, and 78%, respectively) for identifying correctly allowed credit. The optimal accuracy of Gaussian Naïve Bayes, Decision Tree, Linear Discriminant Analysis, and Logistic Regression is 82%, 83%, and 82%, respectively. The highest recall rates were found in linear regression and linear discriminant analysis, at 98% and 90%, respectively. The results showed that the F1-scores for Linear Discriminant Analysis and Logistic Regression were 90% and 86%, respectively, the highest.

When examining the Correctly Rejected detection, the highest accuracy of 84% and 79%, respectively, was found using Logistic Regression and Linear Discriminant Analysis. The greatest results for precision were obtained with 91%, 67%, and 63% for Gaussian Naïve Bayes, Linear Discriminant Analysis, and Logistic Regression, respectively. The models with the greatest recall rates were Gaussian Naïve Bayes, Linear Discriminant Analysis, and Decision Tree, with 56% and 51%, respectively. For both linear discriminant analysis and logistic regression, the maximum F1-score was 62% and 58%, respectively.

## VI. CONCLUSION & FUTURE WORK

In conclusion, by helping researchers choose the best algorithms for identifying fraudulent activity, the research findings can significantly enhance future efforts in credit risk analysis. This work has the potential to reduce financial fraud risks for both individuals and public entities. The use of many machine learning models in conjunction with the use of multiple methods on preprocessed and modified data was essential to this study.

With the rise in financial fraud in recent years, especially in the increasingly digitalized world, this research is a useful tool for people, financial institutions, and governments fighting fraud. In order to assist future research and development efforts in the industry, this study aimed to determine the best effective algorithm for identifying financial fraud through a comparison analysis.

This comparative research examined seven machine learning algorithms: K-Nearest Neighbors, Support Vector Machine, Decision Tree, Random Forest, Logistic Regression, Gaussian Naïve Bayes, and Linear Discriminant Analysis. Different outcomes were obtained from each method, and these were assessed using both conventional metrics (such as the confusion matrix) and key performance indicators (KPIs) (such as accuracy, precision, F1-score, and recall). The results section included a detailed discussion and presentation of these findings.

It is crucial to carry doing this comparison analysis going ahead by adding any recently developed algorithms in the fields of artificial intelligence and machine learning. Researchers can find the best mix of algorithms that produce the most accurate and effective identification of fraudulent activity in credit risk analysis by extending the scope of comparison. This continuous endeavor seeks to protect financial systems from fraudulent activity and improve the efficacy of credit default detection techniques.

## REFERENCES

[1] Moradi S, Mokhatab RF. A dynamic credit risk assess- ment model with data mining techniques: evidence from Iranian banks. Financ Innov. 2019;5(1):15.

[2] Rehman ZU, Muhammad N, Sarwar B, Raz MA. Impact of risk management strategies on the credit risk faced by commercial banks of Balochistan. Financ Innov. 2019;5(1):44.

[3] Khemakhem S, Boujelbene Y. Predicting credit risk on the basis of fnancial and non-fnancial variables and data mining. Rev Acc Financ. 2018;17(3):316–40.

[4] Dornadula VN, Geetha S. Credit card fraud detection using machine learning algorithms. Procedia Computer Science. 2019;165:631–41.

[5] Garcıa V, Marques AI, S´anchez J.S. Improving Risk Pre- dictions by Preprocessing Imbalanced Credit Data. Neural Information Processing. 2012;67:68–75.

[6] Song Y, Peng Y. A MCDM-Based Evaluation Approach for Imbalanced Classifcation Methods in Financial Risk Prediction. IEEE Access. 2019;7:84897–906.

[7] Guo S, He H, Huang X. A multi-stage self-adaptive classi- fer ensemble model with application in credit scoring. IEEE Access. 2019;7:78549–59.

[8] Liu H, Yu L. Toward integrating feature selection algorithms for classifcation and clustering. IEEE Tran Knowl Data Eng. 2005;17(4):491–502.

[9] Tang PS, Tang XL, Tao ZY, Li JP (2014) Research on feature selection algorithm based on mutual information and genetic algorithm. 11th Int. Comput. Conf. Wavelet Active Media Tech. Inf. Processing (ICCWAMTIP) IEEE, 403–406.

[10] Liu C, Wang Q, Zhao Q, Shen X, Konan M. A new feature selection method based on a validity index of feature subset. Pattern Recogn Lett. 2017;92:1–8.

[11] Pandey TN, Jagadev AK, Mohapatra SK, Dehuri S (2017) Credit risk analysis using machine learning classifers. In: 2017 International Conference on Energy, Communication, Data Analytics and Soft Computing (ICECDS) (pp. 1850–1854). IEEE.

[12] Zhang L, Hui X, Wang L (2009) Application of adaptive support vector machines method in credit scoring. In: International Conference on Management Science and Engineering, 1410–1415.

[13] Mohammadi N, Zangeneh M. Customer credit risk assess- ment using artifcial neural networks. IJ Information Technol Computer Science. 2016;8(3):58–66.

[14] Hsu TC, Liou ST, Wang YP, Huang YS, Che-Lin (2019) Enhanced Recurrent Neural Network for Combining Static and Dynamic Features for Credit Card Default Prediction. In: IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 1572–1576.

[15] Bao W, Lianju N, Yue K. Integration of unsupervised and supervised machine learning algorithms for credit risk assessment. Expert Syst Appl. 2019;128:301–15.

[16] Ha VS, Lu DN, Choi GS, Nguyen HN, Yoon B (2019) Improv- ing credit risk prediction in online peer-to-peer (P2P) lending using feature selection with deep learning. In: 21st International Conference on Advanced Communication Technology, 511–515.

[17] Chen C, Zhang Q, Yu B, Yu Z, Lawrence PJ, Ma Q, Zhang Y. Improving protein-protein interactions prediction accuracy using XGBoost feature selection and stacked ensemble classifer. Comput Biol Med. 2020;123: 103899.

[18] Chakrabarty N, Kundu T, Dandapat S, Sarkar A, Kole DK (2019) Flight arrival delay prediction using gradient boosting classifer. In: Emerging technologies in data mining and information security, 651-659

[19] Weldegebriel HT, Liu H, Haq AU, Bugingo E, Zhang D. A new hybrid convolutional neural network and eXtreme gradient boosting classifer for recognizing handwritten Ethiopian characters. IEEE Access. 2019;8:17804–18.

[20] Liang J, Qin Z, Xiao S, Ou L, Lin X. Efcient & secure decision tree classifcation for cloud-assisted online diagnosis services. IEEE Trans Dependable Secure Comput. 2019;18(4):1632–44

[21] Abdou, H., Pointon, J., & Elmasry, A. (2008). Neural nets versus conventional techniques in credit scoring in Egyptian banking. Expert Systems with Applications, 35(3), 1275-1292.

[22] Abellan, J., & Castellano, J. G. (2017). A comparative study on base classifiers in ensemble methods for credit scoring. Expert Systems With Applications, 1-10.

[23] Altman, E. I. (1968). Financial Ratios, Discriminant Analysis and the Prediction of Corporate Bankrupcy. The journal of Finance, 589-609.

[24] Baesens, B., Gestel, T. V., Viaene, S., Stepanova, M., Suykens, J., & Vanthienen, J. (2003). Benchmarking state-of-the-art classification algorithms for credit scoring. Journal of the Operational Research Society, 627–635.

[25] Baesens, B., Setiono, R., Mues, C., & Vanthienen, J. (2003). Using Neural Network Rule Extraction and Decision Tables for Credit-Risk Evaluation. operations research adn the management science, 312-329.

[26] Beck, J., & Shultz, E. (1986). The use of relative operating characteristic (ROC) curves in test performance evaluation. Archives of Pathology & Laboratory Medicine, 13-20.

[27] Bensic, M., Sarlija, N., & Zekic-Susac, M. (2005). Modelling Small - Business Credit Scoring by Using Logistic

Regression, Neural Networks and Decision Trees. Intelligent Systems in Accounting, Finance and Management, 133–150.

[28] Bolton, P., Freixas, X., & Shapiro, J. (2012). The Credit Ratings Game. The Journal of Finance , 67(1), 85 - 111.

[29] Boser, B. E., Guyon , I. M., & Vapnik, V. N. (1992). A Training Algorithm for Optimal Margin Classiers. Pittsburgh: COLT '92 Proceedings of the fifth annual workshop on Computational learning theory Pages 144-152 .

[30] Brown, I., & Mues, C. (2012). An experimental comparison of classification algorithms for imbalanced credit scoring data sets. Expert Systems with Applications, 3446-3453.

[31] Burgers, C. J. (1998). A Tutorial on Support Vector Machines for Pattern Recognition. Data Mining and Knowledge Discovery, 121–167.

[32] Caldieraro, F., Zhang, J. Z., Cunha Jr, M., & Shulman, J. D. (2018). Strategic Information Transmission in Peer-to-Peer Lending Markets. Journal of Marketing, 82, 42 - 63.

[33] Davis, R., Edelman, D., & Gammerman, A. (1992). Machine learning algorithms for creditcard applications . IMA Journal of Mathematics Applied in Business and Industry , 43-51.

[34] Desai, V., Conway, D., Crook, J., & Overstreet, G. (1997). Credit scoring models in credit union environment using neural network and generic algorithms. . IMA Journal of Mathematics Applied in Business & Industry , 323–346 .

[35] Desai, V., Conway, J., & Overstreet, G. (1996). A comparison of neural networks and linear scoring models in the credit union environment . European Journal of Operational Research , 24-37.

[36] Eggermont, J., Kok, J., & Kosters, W. (2004). Genetic programming for data classification: Partitioning the search space. In Proceedings of the 2004 symposium on applied computing, 1001–1005.

[37] Finlay, S. &. (2012). Instance sampling in credit scoring: An empirical study of sample size and balancing. International Journal of Forecasting, 224-238.

[38] Finlay, S. (2011). multiple classfier architectures adn their application to credit risk assessment. European Journal of Operational Research, 368-378.