KNN Based Big Data Analytics Framework for Real-Time Business Decision Support

Twinkle[1], Dr. Parveen Kumar[2]
[1]Research Scholar (Computer Science), School of Engineering and Technology, Shri Venkateshwara University, Gajraula, UP, INDIA.
[2]Research Guide (Computer Science), School of Engineering and Technology, Shri Venkateshwara University, Gajraula, UP, INDIA.

Email: goel.twinkle1996@gmail.com, pk223475@gmail.com

ABSTRACT
Credit risk assessment is a critical aspect of financial decision-making, requiring accurate models to predict the likelihood of loan defaults. Traditional statistical methods often struggle with complex, high-dimensional datasets, making machine learning (ML) techniques a preferred alternative. This study evaluates multiple ML algorithms, including Decision Tree, Support Vector Machine, Adaboost, Random Forest, XGB Classifier, and a proposed K-Neighbors Classifier, to determine their effectiveness in predicting credit risk. The results indicate that the Proposed K-Neighbors Classifier achieved the highest test accuracy (85.25%), but with a higher Mean Squared Error, suggesting potential overfitting. Among other models, Random Forest and XGB Classifier demonstrated strong generalization with balanced accuracy and lower error rates, making them suitable for practical credit risk assessment. The comparative analysis with existing systems highlights the advantages of ML models over traditional approaches in handling large credit datasets. The study emphasizes the need for optimized algorithms that balance accuracy and robustness while minimizing overfitting. Future research should focus on hybrid models, deep learning techniques, and alternative data sources, such as behavioral and transaction-based data, to enhance predictive performance. By leveraging advanced machine learning techniques, financial institutions can improve credit risk evaluation, leading to more informed and efficient lending decisions.

KEYWORDS: KNN, Decision Support Systems, Big Data Analytics.

1. INTRODUCTION

In today's fast-evolving business environment, the exponential growth of data has reshaped traditional decision-making processes, making them more complex yet potentially more insightful. The advent of big data has introduced both significant opportunities and formidable challenges, requiring innovative approaches to harness its full potential for informed, agile, and strategic decision support. In response, the convergence of machine learning techniques and big data analytics has emerged as a transformative paradigm, promising to enhance the effectiveness and intelligence of business decision support systems (DSS). This integration not only addresses the challenges associated with handling large-scale, heterogeneous data but also empowers organizations to extract actionable insights, fostering competitiveness and adaptability in an increasingly dynamic market landscape [1].

The fusion of machine learning and big data analytics represents a pivotal shift in the field of decision support systems, leveraging computational power and advanced algorithms to analyze vast datasets efficiently. By incorporating sophisticated data preprocessing, integration, and analytical techniques, this approach ensures the extraction of meaningful patterns, predictive models, and valuable insights. Machine learning algorithms, including neural networks, decision trees, and ensemble methods, facilitate the identification of hidden trends and

correlations within business data, ultimately improving decision-making accuracy and efficiency [2-3].

Beyond its analytical capabilities, the synergy between machine learning and big data analytics enables the development of adaptive decision-making frameworks. These systems continuously evolve by learning from real-time data streams, ensuring their relevance and accuracy in rapidly changing business environments. The ability to dynamically adjust to new information enhances the resilience of business strategies, allowing organizations to anticipate market trends, mitigate risks, and seize emerging opportunities. This introduction sets the foundation for exploring the intricacies of a machine learning-driven big data analytics framework and its role in revolutionizing modern decision support systems. The subsequent sections delve into the architecture, implementation, and real-world applicability of this approach, demonstrating its potential as a cornerstone for robust, data-driven business decision-making [4].

2. REVIEW OF LITERATURE

A thorough understanding of credit risk assessment is crucial in financial services, where institutions face two primary risks: credit risk exposure and market-related uncertainties. Credit risk exposure involves the potential non-repayment of loans, either partially or entirely, posing significant challenges for lenders [1]. Market-related risks stem from fluctuations in the value of securities, commodities, and services, which are often unpredictable [2]. Additionally, operational risks further complicate credit finance, mortgages, and venture capital investments [3]. To address these challenges, financial institutions increasingly rely on sophisticated credit risk evaluation models, with growing adoption of software robots to automate processes, minimize human effort, and enhance efficiency [5].

Research on credit risk assessment models, particularly those leveraging machine learning, has gained traction in recent years. Chen et al. emphasize the heightened focus on risk assessment post-2008 financial crisis, with a push for predictive algorithms to mitigate business failures [1]. Crook et al. explore consumer credit risk, highlighting key factors like repayment ability and timeliness [2]. Galindo and Tamayo stress the importance of predictor selection in financial risk models, proposing an approach based on error curve research [3]. Twala demonstrates the efficacy of ensemble classifiers in handling noisy data for credit risk assessment [5]. Doumpos et al. examine default probability estimation and its impact on financial stability [7], while Saha et al. propose a data-driven loan approval strategy that integrates expert opinions with data mining [8]. Emerging technologies also play a role—Cai et al. leverage blockchain for credit risk assessment [11], Zhou et al. utilize distributed computing for improved efficiency [12], and Wang et al. develop price forecasting models [14]. Deng et al. introduce k-means clustering for dataset segmentation, showcasing diverse approaches to improving credit risk models [15]. As research progresses, there remains a need to refine machine learning applications, particularly in peer-to-peer lending, to enhance borrower classification accuracy and financial stability.

3. THE PROPOSED METHODOLOGY

Assessing credit risk using traditional statistical models presents challenges due to the complexity and volume of numerical and categorical attributes that require analysis. In contrast, Machine Learning (ML)-based models have gained prominence for their ability to process large and intricate datasets efficiently, making them well-suited for time-sensitive financial applications. While ML models are broadly categorized into different types, various subgroups and specialized algorithms enhance their effectiveness in credit risk evaluation. The proposed research methodology follows a structured approach to credit risk assessment using ML, ensuring accuracy and efficiency in forecasting borrower defaults.

The methodology consists of three key stages. The first stage involves data collection, where relevant financial and demographic information is gathered from credit applicants. In the second stage, the dataset undergoes preprocessing, followed by training various ML models to identify significant patterns and risk indicators. Finally, in the third stage, the trained models are applied iteratively for credit risk prediction, continuously improving accuracy through performance evaluation and refinement. This systematic approach enables robust and scalable credit risk assessment, facilitating informed decision-making in financial institutions.

The K-Nearest Neighbour (KNN) is a method to perform non-parametric classification, whose output is typically a class membership. The KNN approach tries to classify a given sample $'\boldsymbol{XX}'$ into a class $'\boldsymbol{CC}'$ based on the Euclidean Distance $'\boldsymbol{dd}'$ given by:

$dd = \sqrt{f_1^2 + \cdots f_n^2}$ ……(1)

Here, $\boldsymbol{dd}$ denotes the Euclidean Distance

$\boldsymbol{ff_{11}} \ldots \boldsymbol{ff_{nn}}$ denotes the data attributes or features.

A typical illustration of a two-class KNN is depicted in figure 1.

The KNN computes the Euclidean Distance ($\boldsymbol{dd}$) of the data sample $'\boldsymbol{XX}'$ from all the classes. The minim Euclidean distance governs the decision regarding a new data sample being classified into any class $'\boldsymbol{CC}'$.

$yy = \min(dd_1, dd_2 \ldots \ldots dd_{nn})$ …(2)

Thus for $'\boldsymbol{nn}'$ distinct classes, the nearest neighbour based on weighted Euclidean distance can be computed as:

$dd_{ww} = \sqrt{\sum_{i=1}^{n} w_i(C_i - \sigma_i)}$ ….(3)

$\boldsymbol{dd_{ww}}$ denotes the weighted Euclidean Distance

$\boldsymbol{ww_{ii}}$ denotes the weight for element $'\boldsymbol{ii}'$

$\boldsymbol{CC_{ii}}$ denotes the component of $\boldsymbol{ii^{tttt}}$ feature vector.

$\boldsymbol{\sigma\sigma_{ii}}$ denotes the standard deviation of the $\boldsymbol{ii^{tttt}}$ feature vector.

Here, $\boldsymbol{yy}$ denotes the minimum Euclidean distance for the given data class. $\boldsymbol{dd_{11}}, \boldsymbol{dd_{22}} \ldots \ldots \boldsymbol{dd_{nn}}$ denote the individual Euclidean Distances. The weighted version of the KNN is obtained by assigning a weight to the k-nearest neighbour members. Mathematically, the weight assigned to an $\boldsymbol{ii^{tttt}}$ the nearest neighbour is given $\boldsymbol{ww_{nn,ii}}$ with the property of:

$\sum^{nn}_{ii=1} ww_{nn,1} = 1$ …(4)

The objective of the multi-class classification is to attain convergence of error rate for the classifier for $'\boldsymbol{nn}'$ distinct classes.

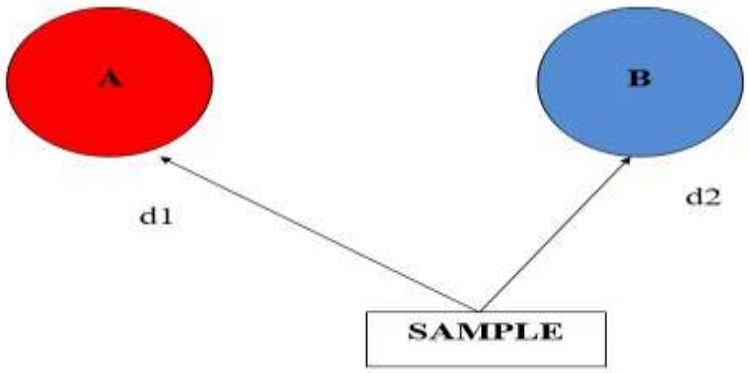

Figure 1: Working of proposed K-Nearest Neighbour

Steps followed to use KNN for credit assessment:

Step 1: Extract data set for training iterations.

Step 2: Split data set into training and testing vectors in the ratio of 75:25.

Step 3: Initialize weights ($\boldsymbol{ww}$) randomly.

Step 4: Update weights using gradient descent to minimize the objective function J given by:

$$= \frac{1}{mm} \sum_{i=1}^{m} (v_i - v'_i)^2 \quad JJ \qquad \ldots(5)$$

Step.5: Compute the error matrix (cost function).

Step.6: Iterate steps (1-4) till the cost function ***JJ*** stabilizes.

4. DATASET

To conduct their research, the authors consulted the credit risk dataset [11] available at UCI's Machine Learning Repository (UCIMLR). A dataset containing 30,000 approved and declined credit applications based on 24 attributes or features is used in the proposed work. This dataset encompasses details related to default payments, incorporating demographic factors, credit data, payment history, and credit card bill statements of clients in Taiwan during the period from April 2005 to September 2005. The dataset comprises 25 variables, including client IDs, credit limits in NT dollars, gender, education level, marital status, age, and repayment statuses across six months. The repayment statuses are denoted on a scale, ranging from timely payments to delayed payments.

Table 1: Description of Dataset Variables and Their Attributes

Variable	Description	Values
ID	Unique identifier for each client	Numeric
LIMIT_BAL	Amount of given credit in NT dollars (individual and family/supplementary credit)	Numeric
SEX	Gender of the client	1 = Male, 2 = Female
EDUCATION	Education level	1 = Graduate school, 2 = University, 3 = High school, 4 = Others, 5 & 6 = Unknown
MARRIAGE	Marital status	1 = Married, 2 = Single, 3 = Others
AGE	Age of the client (in years)	Numeric
PAY_0 to PAY_6	Repayment status from September 2005 to April 2005	-1 = Pay duly, 1 = 1-month delay, ..., 9 = 9+ months delay
BILL_AMT1 to BILL_AMT6	Amount of bill statement from September 2005 to April 2005 (NT dollars)	Numeric
PAY_AMT1 to PAY_AMT6	Amount of previous payment from September 2005 to April 2005 (NT dollars)	Numeric
default.payment.next.month	Default payment status for the next month	1 = Yes, 0 = No

5. PERFORMANCE EVALUATION

The table 2 presents the performance metrics of various machine learning algorithms on a given dataset. It lists the names of the machine learning algorithms used in the analysis. Accuracy (Train) represents the accuracy of each algorithm on the training dataset. This indicates how well the algorithm predicts the target variable on the data it was trained on. Accuracy (Test) indicates the accuracy of each algorithm on a separate test dataset. This dataset is not used during the training phase, and the accuracy on this set gauges how well the algorithm generalizes to

new, unseen data. Test Accuracy (MSE) provides the Mean Squared Error (MSE) for the test dataset. MSE is a measure of the average squared difference between the predicted and actual values. A lower MSE indicates better performance (Figure 2 and Table 2).

Table 2: Performance evaluation of proposed algorithm

Algorithm	Accuracy (Train)	Accuracy (Test)	Test Accuracy (MSE)
Support Vector Machine	61.94%	61.23%	0.3877
Decision Tree Classifier	79.94%	74.13%	0.2587
Adaboost Classifier	75.47%	75.35%	0.2465
Random Forest Classifier	82.94%	81.29%	0.1871
Proposed K-Neighbors Classifier	99.10%	85.25%	0.4575
Logistic Regression	56.08%	55.40%	0.4460
XGB Classifier	85.82%	80.42%	0.1958

The performance comparison of various machine learning algorithms for credit risk assessment is summarized in the table. The Proposed K-Neighbors Classifier achieved the highest training accuracy (99.10%) and a competitive test accuracy (85.25%), though its Mean Squared Error (MSE) was relatively high (0.4575), suggesting possible overfitting. Among traditional classifiers, the Random Forest Classifier exhibited strong generalization capability with a training accuracy of 82.94% and a test accuracy of 81.29%, alongside the lowest MSE (0.1871), indicating reliable predictive performance. The XGB Classifier also performed well, achieving an 85.82% training accuracy and 80.42% test accuracy, with an MSE of 0.1958. In contrast, Support Vector Machine (SVM) and Logistic Regression showed the weakest performance, with test accuracies of 61.23% and 55.40%, respectively, and relatively high MSE values (0.3877 and 0.4460). The Decision Tree Classifier and Adaboost Classifier demonstrated moderate performance, with test accuracies of 74.13% and 75.35%, respectively. Overall, while the Proposed K-Neighbors Classifier had the highest test accuracy, its high MSE suggests further optimization is needed. The Random Forest Classifier and XGB Classifier stand out as robust choices due to their balanced accuracy and low error rates.

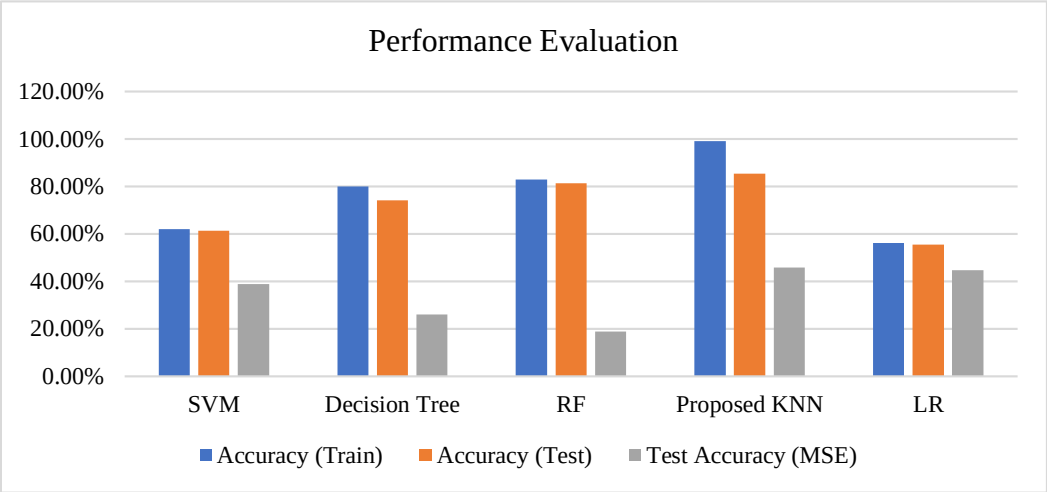

Figure 2: Performance evaluation of proposed system

Table 3: Confusion Matrix

	Predicted 0	Predicted 1
Actual: 0 (Negative Class)	5596 (TN)	277 (FP)

Actual: 1 (Positive Class)	1076 (FN)	551 (TP)

- True Negative (TN): 5596 instances were correctly predicted as class 0.
- False Positive (FP): 277 instances were wrongly predicted as class 1.
- False Negative (FN): 1076 instances were wrongly predicted as class 0.
- True Positive (TP): 551 instances were correctly predicted as class 1.

The model exhibits higher accuracy and precision for predicting class 0 (no default), with a high number of true negatives and a relatively low number of false positives. However, the model's performance is weaker in predicting class 1 (default), with lower recall, precision, and F1-score. It correctly identifies fewer instances of class 1 (low recall) and also misclassifies some class 0 instances as class 1 (moderate false positives). The overall assessment highlights the model's strengths in predicting class 0 but indicates a need for improvement in identifying and predicting class 1 instances more accurately (Figure 3 and Table 3).

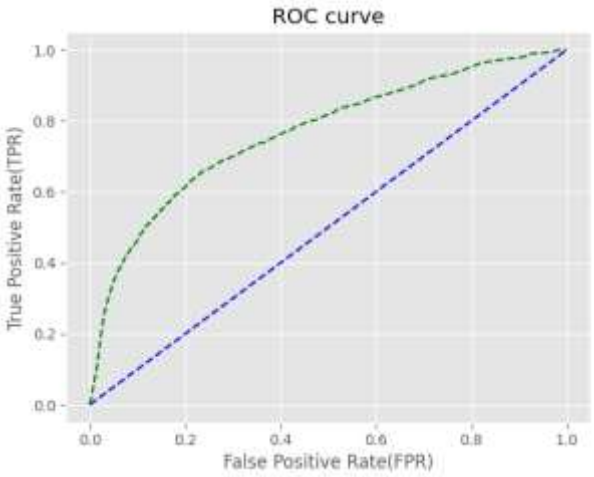

Figure 3: ROC curve for the best model

Table 4: Comparative analysis with existing systems

Research Approach	**Algorithm used**	**Dataset used**	**Classification Accuracy**
Existing work that used K-NN for analysing Tunisian Bank customer's behaviour for checking risk of loan repay [14]	K-Nearest Neighbour Classifier	Tunisian Banks records used	Classification rate 88%
Corporate financial records with features liquidity, solvability, capital, profit margins, return on investment analysed [15]	Ensemble Learning Method by combining K-NN	Corporate financial records	Improved performance by combining methods and got result in order 83%
Presented our work used to decide whether a credit application is accepted or rejected	Decision Tree and the K-Nearest	Credit dataset which has 30,000 samples with	Proposed models attain accuracy of 82.2% and 81.2%

	Neighbour models separately	24 features each	

Table 4 presents a comparative analysis of different research approaches in credit risk assessment. A study on Tunisian bank customers utilized the K-Nearest Neighbor (K-NN) Classifier on Tunisian bank records, achieving a classification accuracy of 88%. Another study analyzing corporate financial records using an Ensemble Learning method that combined K-NN improved classification performance, achieving 83% accuracy. In contrast, the proposed research applied Decision Tree and K-NN models separately on a credit dataset containing 30,000 samples with 24 features, obtaining classification accuracies of 82.2% and 81.2%, respectively. While the existing studies demonstrated high accuracy, the proposed approach provides a robust credit evaluation model using a large and diverse dataset, offering a more scalable solution for real-world credit assessment.

6. CONCLUSION

The evaluation of various machine learning algorithms for credit risk assessment reveals that the Proposed K-Neighbors Classifier achieved the highest test accuracy (85.25%), but its high Mean Squared Error (0.4575) indicates potential overfitting. Among other models, the Random Forest Classifier and XGB Classifier demonstrated strong generalization capabilities, with test accuracies of 81.29% and 80.42%, respectively, and lower MSE values, making them reliable choices for credit risk prediction. In contrast, Logistic Regression and Support Vector Machine showed the lowest test accuracies, suggesting their limitations in handling complex credit data. Overall, the findings emphasize the importance of selecting models that balance accuracy and generalization to minimize financial risk. Future research can explore hybrid models, deep learning techniques, and alternative data sources to further enhance credit risk prediction.

7. REFERENCES

[1] Verma Shruti, Maan Vinod, "Comparative Analysis of Pig and Hive," International Journal of Research in Advent Technology, Vol.6, No.5, 2018. E-ISSN: 2321-9637. Available online at www.ijrat.org. 585 http://www.ijrat.org › paper ID-65201829

[2] N. Chen, B. Ribeiro, and A. Chen, "Financial credit risk assessment: a recent review," Artif Intell Rev, vol. 45, no. 1, pp. 1–23, Jan. 2016, DOI: 10.1007/s10462-015-9434-x.

[3] J. N. Crook, D. B. Edelman, and L. C. Thomas, "Recent developments in consumer credit risk assessment," European Journal of Operational Research, vol. 183, no. 3, pp. 1447–1465, Dec. 2007, DOI: 10.1016/j.ejor.2006.09.100.

[4] J. Galindo and P. Tamayo, "Credit Risk Assessment Using Statistical and Machine Learning: Basic Methodology and Risk Modeling Applications," Computational Economics, vol. 15, no. 1, pp. 107–143, Apr. 2000, DOI: 10.1023/A:1008699112516.

[5] S. Oreski and G. Oreski, "Genetic algorithm-based heuristic for feature selection in credit risk assessment," Expert Systems with Applications, vol. 41, no. 4, Part 2, pp. 2052–2064, Mar. 2014, DOI: 10.1016/j.eswa.2013.09.004.

[6] B. Twala, "Multiple classifier application to credit risk assessment," Expert Systems with Applications, vol. 37, no. 4, pp. 3326–3336, Apr. 2010, DOI: 10.1016/j.eswa.2010.10.018.

[7] L. Yu, S. Wang, and K. K. Lai, “Credit risk assessment with a multistage neural network ensemble learning approach,” Expert Systems with Applications, vol. 34, no. 2, pp. 1434–1444, Feb. 2008, DOI: 10.1016/j.eswa.2007.01.009.

[8] M. Doumpos, K. Kosmidou, G. Baourakis, and C. Zopounidis, “Credit risk assessment using a multicriteria hierarchical discrimination approach: A comparative analysis,” European Journal of Operational Research, vol. 138, no. 2, pp. 392–412, Apr. 2002, DOI: 10.1016/S03772217(01)00254-5.

[9] P. Saha, I. Bose, and A. Mahanti, “A knowledge-based scheme for risk assessment in loan processing by banks,” Decision Support Systems, vol. 84, pp. 78–88, Apr. 2016, DOI: 10.1016/j.dss.2016.02.002.

[10] M. R. Sousa, J. Gama, and E. Brandão, “A new dynamic modeling framework for credit risk assessment,” Expert Systems with Applications, vol. 45, pp. 341–351, Mar. 2016, DOI:10.1016/j.eswa.2015.09.055.

[11] Yeh, I. C., & Lien, C. H., “The comparisons of data mining techniques for the predictive accuracy of probability of default of credit card clients,” Expert Systems with Applications, 36(2), 2473-2480, 2009.

[12] Shousong Cai, Jing Zhang, “Exploration of the credit risk of P2P platform based on data mining technology”, Journal of Computational and Applied Mathematics, Volume 372, 2020.

[13] H. Zhou, G. Sun, S. Fu, J. Liu, X. Zhou and J. Zhou, "A Big Data Mining Approach of PSO-Based BP Neural Network for Financial Risk Management With IoT," in IEEE Access, vol. 7, pp. 154035-154043, 2019, DOI: 10.1109/ACCESS.2019.2948949.

[14] T. Qiu, H. Wang, K. Li, H. Ning, A. K. Sangaiah and B. Chen, "SIGMM: A novel machine learning algorithm for spammer identification in industrial mobile cloud computing", IEEE Trans. Ind. Informat., vol. 15, no. 4, pp. 2349-2359, Apr. 2019. DOI: 10.1109/TII.2018.2799907

[15] Wang Bao, Ning Lianju, Kong Yue, “Integration of unsupervised and supervised machine learning algorithms for credit risk assessment”, Expert Systems with Applications, Volume 128, 2019, Pages 301-315. https://doi.org/10.1016/j.eswa.2019.02.033.