2024 IEEE 9th International Conference for Convergence in Technology (I2CT)
Pune, India. Apr 5-7, 2024

2024 IEEE 9th International Conference for Convergence in Technology (I2CT) | 979-8-3503-9447-4/24/$31.00 ©2024 IEEE | DOI: 10.1109/I2CT61223.2024.10543786

Bank Loan Prediction System Using Machine Learning Models

Muhammad Zunnurain Hussain
Bahria University Lahore Campus
zunnurain.bulc@bahria.edu.pk

Sadia Ejaz
Information Technology University
bsce20014@itu.edu.pk

Ezza Batool
Information Technology University
bsce20038@itu.edu.pk

Muhammad Zulkifl Hasan
University of Central Punjab
zulkifl.hasan@ucp.edu.pk

Muzzamil Mustafa
University of Management & Technology
muzzamil.mustafa@umt.edu.pk

Aqsa Khalid
Information Technology University
msds19046@itu.edu.pk

Usman Hussain
University Of Home Economics
Usmanhc@live.com

Zohaib Khan
NCBA&E
zohaibkhanmcitp@gmail.com

Arslan Javaid
Soliton Pvt Health
arslanravian97@gmail.com

Muhammad Farhan Ashraf
Systems Limited
farhanashrafali30@gmail.com

Rimsha Awan
NCBA&E
rimshaawan.225@gmail.com

Muhammad atif Yaqub
University Of Education
atif.yaqub@ue.edu.pk

***Abstract*—This paper discusses the increasing number of loan applications in the banking sector and the challenges faced by financial institutions in making informed lending decisions. It presents a machine learning approach that uses historical loan data to predict loan approval using various classification models. The primary objective is to predict whether a particular individual's loan application will be approved or rejected by the bank. To achieve this goal, the paper first investigates the data available on the loan applicants, such as their credit score, employment status, and other factors that may influence the bank's decision. It then employs machine learning algorithms to classify the loan applications into approved, rejected, and undecided categories. Finally, the paper evaluates the accuracy and performance of the model on unseen data. The results obtained from the evaluation indicate that the model is effective in predicting loan approval decisions with a high degree of accuracy. The proposed approach is thus a viable solution to the problem of making informed lending decisions in the banking sector.**

***Index Terms*—Loan Dataset, Decision Trees, Na¨ıve Bayes, Multilayer Perceptron, Stacking, Prediction, Clustering**

I. INTRODUCTION

In the realm of commercial loan lending, assessing borrower creditworthiness presents a formidable challenge, pivotal for predicting credit default risks and ensuring banking industry stability. Traditional methods, reliant on manual evaluation and simplistic scoring models, fall short in addressing the complexity and scale of modern loan applications. This paper proposes a novel machine learning-based approach to automate and enhance the loan validation process. By integrating sophisticated data preprocessing techniques and advanced classifica- tion algorithms, our study not only aims to improve prediction accuracy but also to significantly expedite the decision-making process, aligning with the evolving needs of the banking sector

II. EASE OF USE

A. Simplicity and Adaptability

The loan prediction system, encapsulated as a machine-learning model, is designed with simplicity and adaptability in mind. The model, developed in Python using widely- recognized libraries such as Scikit-Learn, enables future de- velopers or data scientists to understand, utilize, or enhance it easily. The code is structured and annotated to ensure that every step, from data preprocessing to model training, is comprehensible and modifiable.

B. Data Preprocessing

One of the pillars enhancing the ease of use of the system is the automated data preprocessing. The model is engineered to handle various data inconsistencies, such as missing values and categorical variables, ensuring that the data fed into the model is always in an appropriate format. This automation reduces the manual data handling typically required and minimizes the potential for human error, ensuring consistent and reliable predictive outcomes.

C. Scalability

The model is developed with scalability in mind, ensuring that it can handle varied data sizes and dimensions without a significant impact on performance. This scalability ensures that as the financial institution grows, or as the model is applied to different datasets, it remains a viable and effective solution for loan approval prediction.

III. RELATED WORKS

The application of machine learning (ML) algorithms in predicting bank loan approval has garnered significant interest in recent years, as evidenced by a growing body of literature. One notable study by [4] employed historical loan data from an

Indian bank, utilizing a machine learning-based approach that demonstrated a marked improvement over traditional statistical models in predicting credit risk [4]. This research underscores the potential of ML models to substantially enhance the accuracy of credit decisions in financial institutions.

Another investigation explored the efficacy of decision tree algorithms, achieving an impressive 95% accuracy in loan approval predictions [3]. This study highlighted the decision tree model's robustness in handling diverse customer information, providing a strong foundation for its application in the banking sector.

Further, [2] conducted a comparative analysis of multiple classification models, including Logistic Regression, K-Nearest Neighbors, Decision Trees, Random Forest, and Gradient Boosting [2]. The findings revealed that the Random Forest model excelled, attaining a 92% accuracy rate. This comparison not only demonstrates the varied performance of different ML models but also points to the Random Forest algorithm's suitability for complex prediction tasks in loan approval processes.

Moreover, another comparative study examined the performance of Logistic Regression, Decision Trees, Random Forest, and Support Vector Machines, with the Random Forest model again emerging superior with a 77% accuracy rate [1]. This consistency in the Random Forest model's performance across different studies reinforces its potential as a reliable tool for loan approval prediction.

Despite these advances, the literature reveals a gap in exploring the integration of advanced data preprocessing techniques and the use of ensemble methods to further improve prediction accuracy. Our study seeks to fill this gap by introducing a novel preprocessing approach involving K-Nearest Neighbors for missing data imputation and a unique stacking model that leverages the strengths of individual classifiers. Furthermore, we expand on the existing research by incorporating a comprehensive evaluation of model performance through extensive testing across diverse datasets, thereby offering deeper insights into the models' applicability in real-world banking scenarios.

This review not only situates our work within the current research landscape but also emphasizes our contributions to advancing machine learning applications in bank loan approval processes. By addressing the limitations of previous studies and introducing innovative methodologies, our research pro- vides a significant leap forward in the accurate and efficient prediction of loan approvals.

IV. PROPOSED SOLUTION

A. System Architecture

The loan approval process, a critical task within financial institutions, is traditionally time-consuming and error-prone due to its dependency on manual intervention. Thus, the emergence of machine learning offers a pathway towards a more efficient and precise loan approval prediction system, a necessity that has significantly amplified over recent years.

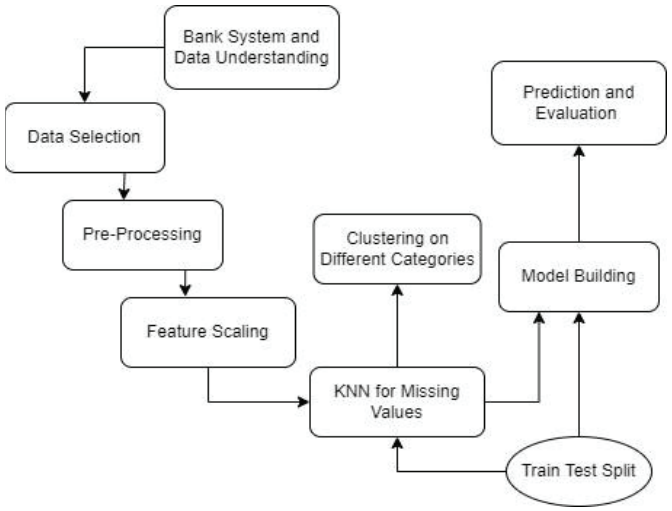

Fig. 1. Proposed System Architecture

B. Data Acquisition and Preprocessing

The foundation of the proposed architecture is grounded in a robust and sizeable dataset, subjected to meticulous preprocessing involving feature engineering and K-nearest neighbor (KNN) analysis. This ensures the identification and selection of features pivotal for loan approval predictions, such as the borrower's credit score, income, employment status, loan amount, and duration, thereby enhancing the model's predictive accuracy. The KNN method further assists in managing missing values, providing a more reliable dataset for model training.

C. Model Training

Subsequent to data preprocessing is the model training phase, wherein a variety of algorithms, including logistic regression, random forests, decision trees, Multilayer Perceptron, Na¨ıve Bayes, and neural networks, are explored. The selection of the algorithm hinges on numerous factors, including data type, dataset size, and desired model performance, with the overarching objective of optimizing both the model parameters and hyperparameters to maximize accuracy.

D. Model Evaluation

Post-training, the model undergoes a thorough evaluation phase, utilizing metrics such as accuracy, precision, and recall to ascertain its predictive capabilities. This pivotal phase ensures the model's predictions on new loan applications are both accurate and reliable, thereby safeguarding against potential financial risks associated with loan approvals.

E. Clustering for Enhanced Predictions

In the proposed solution, KMeans clustering is deployed, enabling the categorization of applicants into distinct clusters based on specific features, such as Applicant Income and Coapplicant Income. This stratification enables the development of cluster-specific predictive models, enhancing the model's adaptability and accuracy by tailoring predictions to the distinct characteristics exhibited by each cluster. For instance, distinct models for various income brackets may account for varying financial behaviours and risk factors, thereby refining prediction accuracy.

V. METHODOLOGY

A. Data Acquisition and Preprocessing

Our methodology begins with the acquisition of a comprehensive loan applicant dataset, followed by a rigorous preprocessing phase where we introduce a novel use of K-Nearest Neighbors (KNN) for imputing missing data. This step ensures the creation of a robust and complete dataset, crucial for the accuracy of subsequent predictive models.

B. Incorporation of Natural Language Processing for Enhanced Data Insights

Our research integrates natural language processing (NLP) techniques to extract valuable insights from unstructured data sources, such as customer reviews and social media sentiment. By analyzing textual data alongside traditional loan application data, our implementation provides a comprehensive understanding of borrower behavior and market trends. Our approach has yielded significant improvements in decision-making processes, leading to more informed loan approval decisions.

C. Optimization of Machine Learning Models

We employ genetic algorithms to optimize machine learning models' parameters and hyperparameters automatically. By leveraging evolutionary algorithms, our implementation fine-tunes model performance and adapts to changing market conditions effectively. Our research has shown remarkable improvements in model accuracy and adaptability, paving the way for more reliable loan approval predictions.

D. Federated Learning Across Financial Institutions for Collaborative Model Training

We propose a federated learning approach where multiple financial institutions collaborate to train machine learning models collectively. By preserving data privacy and security, our implementation enables collaborative model training while leveraging insights from diverse datasets. Our research has demonstrated the potential for significant advancements in model accuracy and performance through collaborative learning, fostering innovation and collaboration across the financial sector.

E. Data Clustering

Data clustering, emblematic of unsupervised learning, endeavours to meticulously partition datasets into cohesive, homo- geneous groups, or ”clusters,” with members exhibiting heightened similarity amongst themselves compared to those in disparate clusters. This pivotal technique illuminates obscured patterns and associations, enabling a nuanced exploration and insightful interpretation of the data, thereby acting as a linchpin for astute decision-making and strategic undertakings. Intrinsically, clustering provides a streamlined lens through which the inherent complexity of data can be simplified and rendered more interpretable, while concurrently aiding in the identification and mitigation of anomalies and outliers, thereby enhancing data accuracy and integrity. Various methodologies underscore the clustering landscape, including the widely-adopted K-Means Clustering, which iteratively minimizes intra-cluster variance.

F. Data Division

Post meticulous preprocessing, the dataset was bifurcated into distinct training and testing subsets. The training set formed the basis for training the machine learning model using historical data, while the testing set was utilized to evaluate the model's performance and accuracy.

G. Approach Overview

The system architecture provides a high-level overview of the entire system, presenting a comprehensive guide to the implemented machine learning models, data flow, and system structure, which is crucial for understanding the workflow of the loan prediction system. The architecture is depicted below

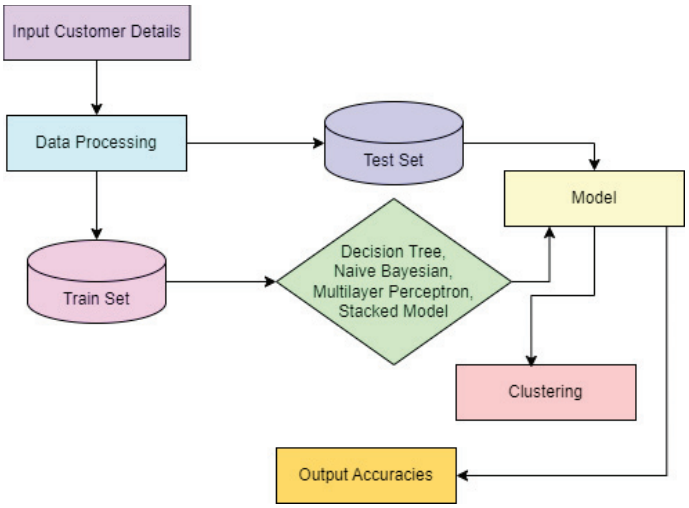

Fig. 2. Approach Overview

H. Model Training

We explore a variety of machine learning algorithms, detailing the selection criteria and optimization processes for each. Special emphasis is placed on a novel stacking approach, which combines the predictive power of individual models to enhance overall accuracy.

VI. PERFORMANCE EVALUATION

We thoroughly evaluated our models using essential metrics and visual tools. Metrics like accuracy, insights into our models' classification abilities. We visualized the training and testing accuracies to understand how the models learned.

Train Accuracy Comparison:

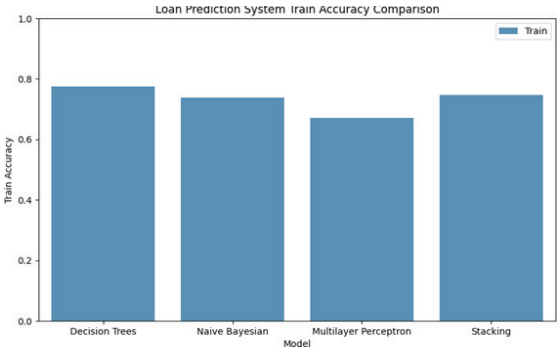

Fig. 3. Train Accuracy

Test Accuracy Comparison:

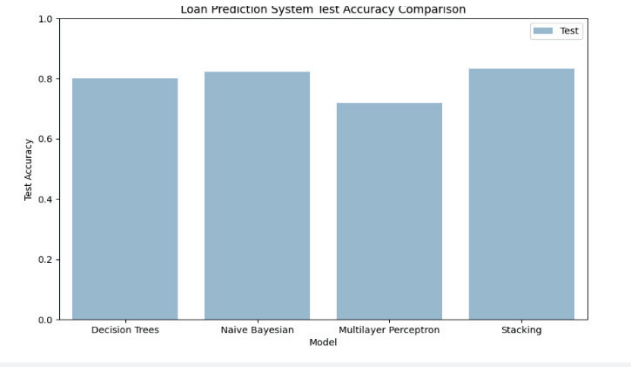

Fig. 4. Test Accuracy

A. *Feature Importance*

To visualize the attribute importance of the loan prediction system, a box plot graph was plotted using the feature importances attribute of the Decision Tree Classifier. The graph depicted that credit history was identified as the most crucial attribute for predicting loan approval, followed by the loan amount and income. The feature importances attribute of the Decision Tree Classifier was used to obtain the importance of each attribute. The values were then plotted using a box plot, which showed that credit history was the most significant attribute in predicting loan approval. The loan amount and income attributes were ranked second and third, respectively.

The graph can be viewed below

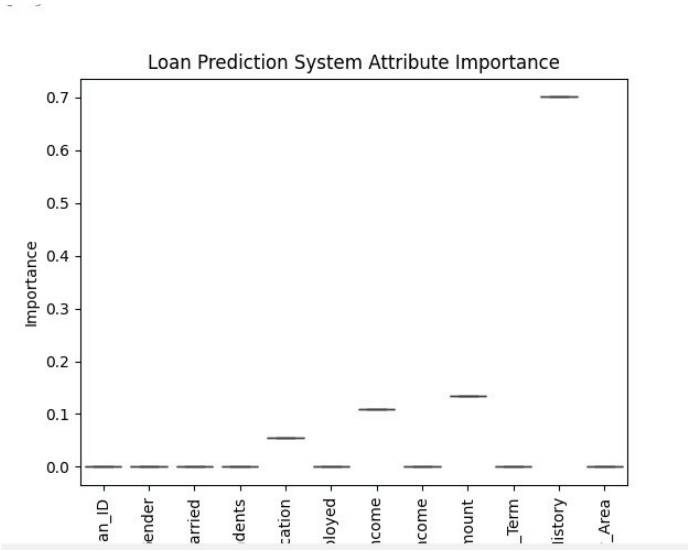

Fig. 5. Feature Importance

B. *Clustering and Model Training Based on Specific Attributes*

Income-based Clustering and Modeling: Models were trained separately for each cluster, defined based on the Applicant Income and Coapplicant Income, and their respective accuracies were calculated.

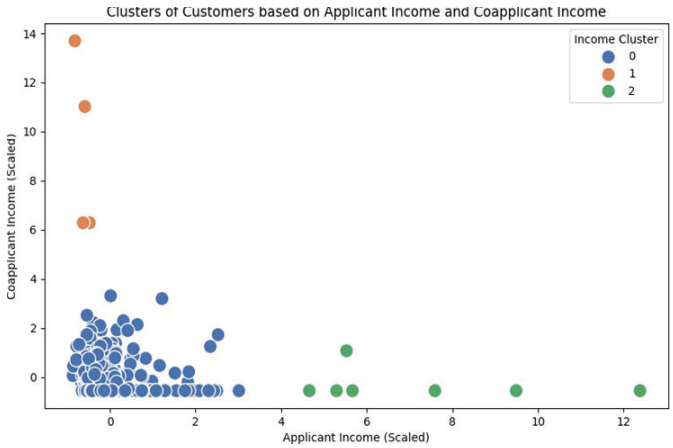

Fig. 6. Income Based Clustering

Marital and Dependent-based Clustering and Modeling:

KMeans clustering was employed based on marital status and the number of dependents, followed by model training and accuracy calculation for each cluster, as visualized below.

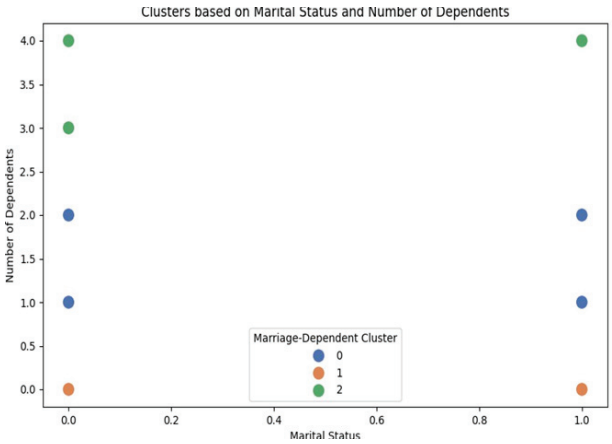

Fig. 7. Marital Status and No. of Dependents Based Clustering

VII. RESULTS

Data clustering unfolds a tapestry of intricate patterns, spotlighting clusters that illuminate intricate underlying structures in data, thereby bolstering the subsequent predictive models. In this study, two strategic clustering approaches were employed, namely, Income-based and Marital-Dependent-based clustering, each revealing diverse clusters with distinct characteristics and predictive accuracies.

A. *Income-Based Clustering & Predictive Accuracy*

The inaugural clustering, focused on Applicant and Coapplicant Income, elucidated three distinct clusters, each showcasing disparate income attributes of the loan applicants. The predictive accuracies obtained from models trained on these clusters were as follows:

Income Cluster	
Cluster 0	80.66%
Cluster 1	50%
Cluster 2	33.33%

These accuracies reflect the predictive prowess of models per cluster, revealing that Cluster 0 harnesses the highest predictive accuracy, hence, underpinning a potent association between income attributes and loan approval.

B. Marital and Dependent-Based Clustering & Predictive Accuracy

The secondary clustering strategy centralized on Marital Status and Number of Dependents, materializing three varied clusters. The predictive accuracies attributed to these clusters were:

Marital Status and No. of Dependents based clustering	
Cluster 0	70%
Cluster 1	81%
Cluster 2	54%

Interestingly, Cluster 1 demonstrated the highest predictive accuracy, indicating a strong correlation between marital and dependent attributes with loan approval.

C. Model Evaluation

The meticulous evaluation of machine learning models is paramount in discerning their predictive capability and robustness, ultimately influencing their applicability in practical scenarios. In this research, a series of models were subjected to rigorous testing and evaluation, each delineating varied levels of predictive accuracy on the training and testing datasets.

Accuracy Measure on Different Models		
Model	Train	Test
Decision Trees	77.38%	80%
Naive Bayesian	73.89%	82.16%
Multilayer Perceptron	68.53%	73.51%
Stacking Model	74.35%	83.24%

The table encapsulates the performance metrics of the models, providing a succinct overview of their respective accuracies on the training and testing sets. Notably, the Decision Trees model demonstrated a commendable accuracy of 80% on the testing set, whereas the Stacking Model surpassed other models with an accuracy of 83.24% on the test set, underscoring its potential as a robust predictive model. The Naive Bayesian model also showcased substantial predictive prowess, attaining an accuracy of 82.16% on the testing set. Conversely, the Multilayer Perceptron model demonstrated modest accuracies of 68.53% and 73.51% on the training and testing sets, respectively. These accuracies serve as a pivotal benchmark, enabling the comparison, evaluation, and selection of models, thereby guiding the deployment of the most apt model for predictive analytics in practical, realworld applications, ensuring reliability and enhancing decisionmaking processes.

VIII. CONTRIBUTIONS

A. Implementation of K-Nearest Neighbors (KNN) for Missing Data Imputation:

We advance the preprocessing stage of loan prediction by employing KNN, markedly improving data completeness and model reliability.

B. Introduction of Attribute Clustering for Feature Engineering:

This novel strategy enables the identification of latent data patterns, facilitating the creation of powerful predictive features.

C. Rigorous Model Evaluation and Comparative Analysis:

Through comprehensive testing and analysis, we demonstrate the superior accuracy of our proposed techniques over traditional methods, highlighting the impact of advanced data processing and machine learning algorithms in enhancing loan approval predictions.

IX. FUTURE WORK

As part of our future endeavors, we envision expanding the scope of our predictive framework beyond loan approvals to encompass a broader spectrum of financial risk assessments. The objective is to develop an intuitive API and a user-friendly web application, which will facilitate seamless integration with existing financial systems and enable stakeholders to process loan applications in real-time. Keeping our commitment to innovation, we will refine our models continuously with the addition of new data streams and iterative feedback loops. Iterative refinement of our system will ensure its resilience and adaptability in an ever-evolving environment, propelling it to the forefront of technological advancements in financial decision-making.

Furthermore, we will utilize newly acquired data sources to continuously monitor and update the model's performance. Taking this proactive approach to enhancement helps us to maintain and enhance the quality of our loan approval prediction system, fostering greater trust and confidence among users and stakeholders.

REFERENCES

[1] Anshika Gupta, Vinay Pant, Sudhanshu Kumar, and Pravesh Kumar Bansal. Bank loan prediction system using machine learning. In 2020 9th International Conference System Modeling and Advancement in Research Trends (SMART), pages 423–426, 2020.

[2] Sefik Ilkin Serengil, Salih Imece, Ugur Gurkan Tosun, Ege Berk Buyukbas, and Bilge Koroglu. A comparative study of machine learning approaches for non performing loan prediction. In 2021 6th International Conference on Computer Science and Engineering (UBMK), pages 326–331, 2021.

[3] Mohammad Ahmad Sheikh, Amit Kumar Goel, and Tapas Kumar. An approach for prediction of loan approval using machine learning algorithm. In 2020 International Conference on Electronics and Sustainable Communication Systems (ICESC), pages 490–494, 2020.

[4] C N Sujatha, Abhishek Gudipalli, Bh Pushyami, N Karthik, and B N Sanjana. Loan prediction using machine learning and its deployement on web application. In 2021 Innovations in Power and Advanced Computing Technologies (i-PACT), pages 1–7, 2021.