# Analysis on credit risk loan approval prediction using machine learning

**First Author** : Shaik Sabeer Ahamed, M.Tech Student, Department of CSE, Anantha Lakshmi Institute of Technology & Sciences, Anantapur, A.P, India

**Second Author** : Mrs. Manjula M.Tech, (Ph.D.) , Department of CSE, Anantha Lakshmi Institute of Technology & Sciences, Anantapur, A.P, India.

## ABSTRACT

In the evolving landscape of financial services, accurately evaluating home loan applications is essential for minimizing credit risk and ensuring efficient lending decisions. This study investigates the use of supervised machine learning algorithms—Decision Tree, Logistic Regression, and K-Nearest Neighbors (KNN)—to predict home loan approval outcomes using a historical dataset from Australian banks. Each model is implemented and assessed based on key performance metrics including accuracy, precision, recall, and F1-score.

Logistic Regression serves as a baseline due to its simplicity and interpretability, effectively highlighting the influence of individual features such as income and credit history. The Decision Tree model offers a rule-based structure that captures non-linear relationships and enhances decision transparency, while KNN classifies based on the similarity of applicants but is more computationally intensive.

Experimental results demonstrate that Logistic Regression achieved the highest accuracy, whereas Decision Tree showed superior handling of categorical variables and feature interactions. KNN performed competitively but faced limitations in scalability and interpretability. This study underscores the potential of machine learning to enhance credit risk assessment, offering a path toward more consistent, fair, and efficient home loan processing.

## 1. INTRODUCTION

In today's digitally advanced financial sector, the demand for accurate, efficient, and unbiased decision-making processes is greater than ever. One of the most vital areas in this domain is the evaluation of home loan applications, where banks must carefully assess a borrower's creditworthiness before approving a loan. Traditionally, this assessment has been carried out manually by loan officers who review various applicant features such as income, employment status, credit score, loan amount, and existing financial obligations. While this human-centric approach allows for nuanced judgment, it also introduces significant limitations, including inconsistency, processing delays, and potential biases.

As the number of loan applications continues to grow and applicant profiles become more complex, manual methods are proving to be insufficient and unsustainable. The need for an automated, scalable, and data-driven approach has led to the rise of machine learning (ML) in financial decision-making. Machine learning algorithms can analyze large datasets, detect hidden patterns, and predict outcomes with a high degree of accuracy, making them an ideal

solution for automating home loan approval processes. By leveraging historical application data, ML models can identify the characteristics most likely to lead to loan approval or rejection, allowing banks to make faster and more reliable decisions.

This project investigates the use of three well-known supervised machine learning algorithms—Logistic Regression, Decision Tree, and K-Nearest Neighbors (KNN)—to predict whether a home loan application should be approved. These models are applied to a historical mortgage application dataset from Australian banks collected in 1988. The dataset includes diverse features such as applicant income, employment type, credit history, marital status, loan amount, property details, and loan duration, with each record labeled as either "approved" or "not approved." The models are trained on this data to learn patterns and are then tested to evaluate how well they can generalize to new applications.

Each of the selected algorithms has its unique strengths and limitations. Logistic Regression provides a linear and interpretable model that is commonly used in binary classification problems. It calculates the probability of an outcome and allows for insight into how individual features influence the decision. Decision Tree models create a tree-like structure of rules based on feature values, making them highly interpretable and effective in capturing non-linear relationships. K-Nearest Neighbors, on the other hand, is a distance-based method that assigns class labels based on the most common outcome among the nearest data points. While simple and often accurate, KNN can be sensitive to irrelevant features and requires significant computational resources for large datasets.

A crucial step in building these models is data preprocessing, which includes cleaning the dataset, handling missing values, encoding categorical variables, and normalizing numerical attributes. This ensures that the models can learn effectively and avoid bias due to inconsistent data formatting. After preprocessing, each model is trained and evaluated using key classification metrics: accuracy, precision, recall, F1-score, and confusion matrix. These metrics provide a balanced view of how well each model performs in identifying both approved and rejected applications.

Beyond accuracy, the study also considers model interpretability, transparency, and practicality in real-world deployment. In financial services, it is often essential for institutions to explain the reasoning behind loan decisions to customers and regulators. Logistic Regression and Decision Tree models excel in this regard by offering clear explanations of how decisions are made. KNN, while potentially accurate, is less interpretable and more resource-intensive, which may limit its usability in large-scale banking environments.

Ultimately, the motivation behind this project is to explore how machine learning can support smarter, faster, and fairer credit risk assessments in home loan approvals. By comparing these three models, the project aims to identify the most suitable algorithm for deployment in automated systems that assist financial institutions in reducing risk, increasing efficiency, and delivering better customer experiences. Although the dataset is historical, the insights gained and methodologies applied remain highly relevant, offering a foundation for future applications in modern financial systems

## 2. LITERATURE SURVEY

In recent years, the integration of machine learning into financial services has become a focal point for researchers and practitioners aiming to optimize credit risk assessment and automate decision-making. Numerous studies have explored the application of

machine learning algorithms to predict loan approval outcomes, with particular attention given to classification models due to the binary nature of the problem (approved or not approved).

One of the foundational techniques in this domain is Logistic Regression, a widely used statistical model for binary classification problems. According to Thomas et al. (2002), Logistic Regression is particularly effective in credit scoring because of its simplicity, efficiency, and interpretability. It provides a probabilistic framework that allows financial institutions to estimate the likelihood of loan default or approval based on input features. However, it assumes linearity between the independent variables and the log-odds of the outcome, which can be a limitation when dealing with complex, non-linear data.

Decision Tree algorithms have also gained popularity due to their ability to model non-linear relationships and provide human-readable decision rules. As discussed by Khandani et al. (2010), Decision Trees are particularly useful in the financial sector because they mimic human decision-making processes and offer transparency in predictions, which is essential for regulatory compliance. Nonetheless, their tendency to overfit the training data, especially in the absence of pruning techniques, remains a concern.

K-Nearest Neighbors (KNN), a non-parametric and instance-based learning method, has been explored in studies like Baesens et al. (2003), where it was used for credit scoring. KNN classifies new samples based on the most frequent class among the k closest data points in the feature space. While it does not make any assumptions about data distribution, it is sensitive to irrelevant features and suffers from high computational cost during inference, especially with large datasets.

Comparative studies have shown varying results depending on the dataset and evaluation criteria. Louzada et al. (2016) conducted a performance comparison among several classification techniques—including Logistic Regression, Decision Tree, and KNN—on credit scoring data. The findings highlighted that no single model outperformed others in all scenarios, emphasizing the importance of model selection based on context, interpretability, and performance metrics.

Recent advancements have also explored the use of ensemble methods (such as Random Forest and Gradient Boosting) and deep learning for loan prediction, achieving higher accuracy but often at the cost of interpretability. However, for many financial institutions, especially those in regulated environments, simpler and more transparent models like Logistic Regression and Decision Tree remain preferable.

Several studies using datasets from different regions have also underscored the importance of feature engineering and preprocessing. Malhotra and Malhotra (2003) emphasized the role of carefully selecting and transforming input variables to improve model performance. This includes handling missing data, encoding categorical variables, and normalizing numerical attributes to ensure fair and effective learning.

The dataset used in this project—a historical mortgage application dataset from Australian banks—has been used in earlier academic explorations into credit approval modeling. While the data is from 1988, it still serves as a valuable benchmark for evaluating the behavior and applicability of classic machine learning models in financial decision-making contexts.

The existing literature highlights a strong foundation for using machine learning in loan approval

prediction. While Logistic Regression offers interpretability and efficiency, Decision Trees provide transparency and flexibility for complex data. KNN, although less interpretable, can yield accurate results in specific scenarios. This project builds upon these prior findings by applying and comparing these models on Australian mortgage data, with the aim of identifying the most practical and effective algorithm for real-world deployment.

## 3. PROPOSED SYSTEM

The proposed system aims to develop a data-driven, automated solution for predicting home loan approvals using supervised machine learning algorithms. This system is designed to support financial institutions in making quicker, more consistent, and accurate loan decisions by analyzing patterns from historical mortgage application data. The system leverages the strengths of three popular classification models—Logistic Regression, Decision Tree, and K-Nearest Neighbors (KNN)—to identify whether a new loan application is likely to be approved or rejected.

The core idea behind the proposed system is to eliminate the shortcomings of traditional manual assessment methods, which are often prone to human bias, delays, and inconsistency. Instead, the system uses past data to learn and generalize patterns, offering a scalable and objective approach to loan evaluation. By feeding historical Australian bank loan data into the machine learning models, the system can identify key relationships between applicant features such as income, credit history, employment type, loan amount, and other financial indicators.

To achieve this, the system will follow a well-defined machine learning pipeline. The first phase involves data preprocessing, where missing values will be handled, categorical variables will be encoded, and numerical features will be normalized to ensure consistency. Feature selection techniques may also be applied to retain only the most relevant attributes. Once the data is cleaned and prepared, it will be split into training and testing sets to train the machine learning models and evaluate their performance objectively.

Three models—Logistic Regression, Decision Tree, and KNN—will be implemented using Python and appropriate libraries such as Scikit-learn. Logistic Regression will serve as a baseline model due to its simplicity and ability to provide probabilistic insights. The Decision Tree model will be used for its rule-based structure and transparency in decision-making. KNN will act as an instance-based learning model to provide a contrast to the parametric approaches. Each model will be evaluated using performance metrics such as accuracy, precision, recall, F1-score, and confusion matrix to measure their effectiveness in classifying approved versus rejected applications.

The proposed system will not only compare the performance of these models but will also assess their interpretability and computational efficiency, which are crucial for real-world deployment. Based on this analysis, the most suitable model will be recommended for integration into a loan processing pipeline. The system can be further enhanced with real-time data inputs and user-friendly interfaces, enabling seamless adoption by banks and financial institutions.

The proposed system is a robust, transparent, and intelligent decision-support tool that leverages machine learning to improve the accuracy and efficiency of home loan approval processes. By replacing manual assessments with predictive modeling, the system aims to contribute to the digital transformation of the banking sector while ensuring fairness and reliability in credit risk evaluation

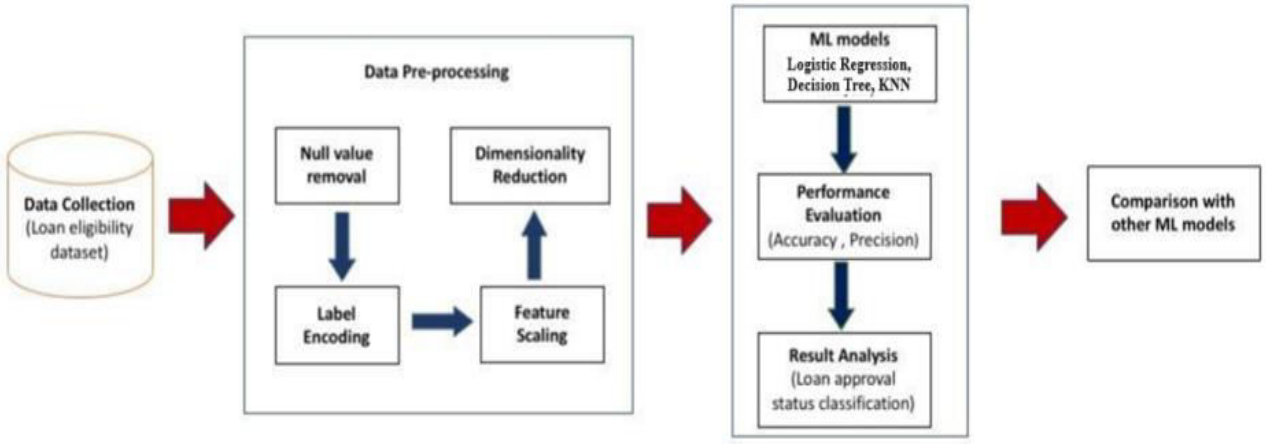


Fig 1: Architecture

The implementation of the proposed loan approval prediction system involves several structured phases, beginning with data preprocessing and progressing through model training, evaluation, and comparison. This section outlines each step in detail and highlights the tools, techniques, and algorithms used throughout the development lifecycle.

1. Data Collection and Understanding

The implementation begins with acquiring a real-world dataset containing historical records of mortgage applications from Australian banks, dated 1988. The dataset includes various features such as income, employment status, loan amount, credit history, marital status, number of dependents, property type, and the final loan approval status (approved or not approved). Understanding the structure, distribution, and quality of this data is crucial for building effective models.

2. Data Preprocessing

Raw data typically includes inconsistencies, missing values, and a mix of categorical and numerical variables. Preprocessing is conducted as follows:

Handling Missing Values: Missing numerical values are replaced with median or mean values, and missing categorical values are filled using the most frequent category or marked as a separate class.

Encoding Categorical Variables: Label Encoding or One-Hot Encoding is applied to convert categorical data (e.g., employment type, marital status) into numerical format for model compatibility.

Feature Scaling: Standardization or Min-Max Normalization is performed on numerical features to ensure uniform scaling, particularly important for distance-based algorithms like KNN.

Feature Selection: Features that are most correlated with the target variable are retained using correlation heatmaps or feature importance techniques.

3. Model Building

Three supervised machine learning models are implemented using Python's scikit-learn library:

Logistic Regression: This model acts as a baseline. It applies a sigmoid function to model the probability of

loan approval based on a linear combination of input features.

Decision Tree Classifier: This model builds a tree-like structure where nodes represent decisions based on feature values. The Gini Index or Entropy is used as the splitting criterion.

K-Nearest Neighbors (KNN): This instance-based learner classifies a data point based on the most common class among its k closest neighbors using Euclidean distance. The value of k is optimized through experimentation.

4. Model Training and Testing

The dataset is split into training and testing subsets, typically in an 80:20 or 70:30 ratio. Each model is trained using the training set and tested on the unseen testing set to evaluate generalization.

5. Model Evaluation

Performance of each model is assessed using the following metrics:

Accuracy: Measures the overall correctness of the model.

Precision: Indicates the proportion of positive predictions that are actually correct.

Recall: Shows the proportion of actual positives that were identified correctly.

F1-Score: Harmonic mean of precision and recall, balancing both metrics.

Confusion Matrix: Provides a detailed breakdown of true positives, false positives, true negatives, and false negatives.

Cross-validation techniques (e.g., 5-fold or 10-fold) are also applied to ensure the robustness of the evaluation.

## 4. RESULT

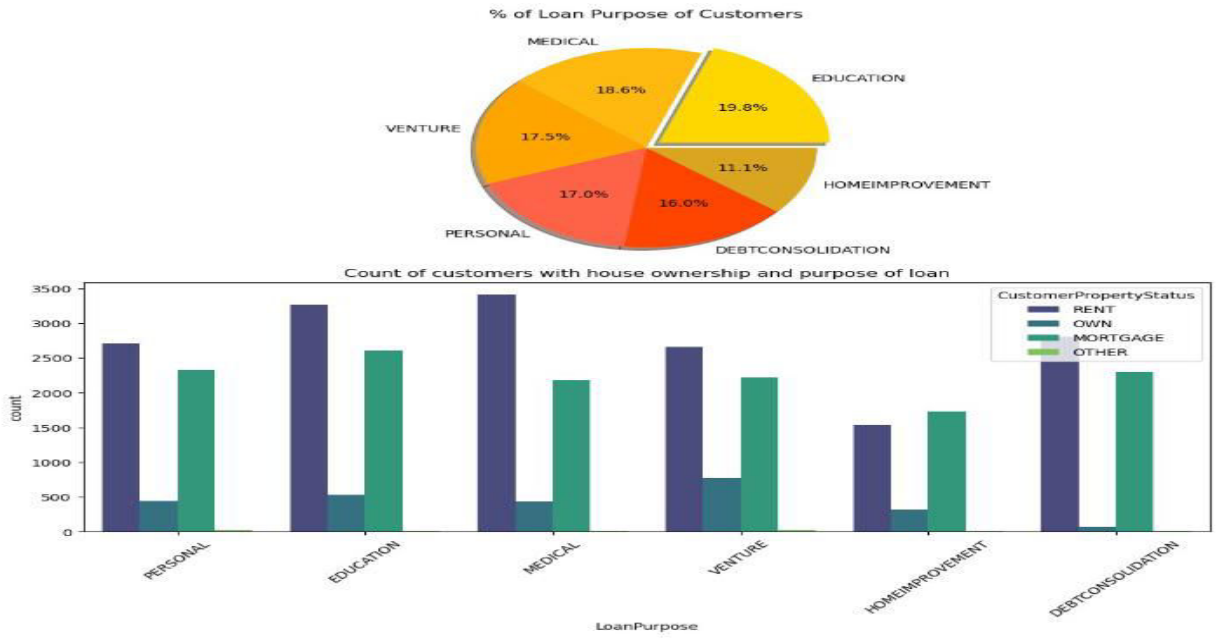


Fig 2: Result for analysis of loan purpose

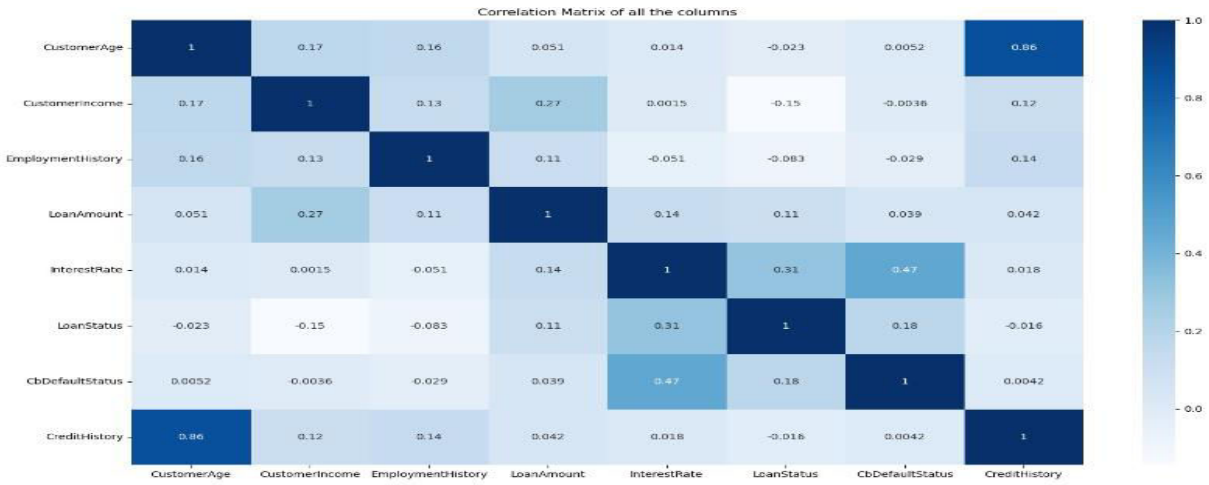


Fig 3: Result with confusion matrix for loan risk analysis

```
Model: Logistic Regression
Accuracy: 0.784682566341955
Confusion Matrix:
[[1171  349]
 [ 292 1165]]
Classification Report:
              precision    recall  f1-score   support

           0       0.80      0.77      0.79      1520
           1       0.77      0.80      0.78      1457

    accuracy                           0.78      2977
   macro avg       0.78      0.78      0.78      2977
weighted avg       0.79      0.78      0.78      2977

Time taken to predict: 0.30144262313842773 seconds

Model: Decision Tree
Accuracy: 0.8434665770910312
Confusion Matrix:
[[1439   81]
 [ 385 1072]]
```

Fig 4: accuracy prediction in Logistic Regression model

```
[[5332   35]
 [ 468  998]]
Classification Report:
              precision    recall  f1-score   support

           0       0.92      0.99      0.95      5367
           1       0.97      0.68      0.80      1466

    accuracy                           0.93      6833
   macro avg       0.94      0.84      0.88      6833
weighted avg       0.93      0.93      0.92      6833

Time taken to predict: 0.07305717468261719 seconds

Model: KNN
Accuracy: 0.8631640567832577
Confusion Matrix:
[[5266  101]
 [ 834  632]]
Classification Report:
              precision    recall  f1-score   support

           0       0.86      0.98      0.92      5367
           1       0.86      0.43      0.57      1466

    accuracy                           0.86      6833
   macro avg       0.86      0.71      0.75      6833
weighted avg       0.86      0.86      0.84      6833

Time taken to predict: 0.42205023765563965 seconds
```

Fig 5: accuracy prediction in Decision Tree model & KNN Model

## 5. CONCLUSION

This project investigated the application of three supervised machine learning algorithms—Decision Tree, Logistic Regression, and K-Nearest Neighbors (KNN)—for predicting credit risk in home loan approval decisions. Leveraging a historical dataset of mortgage applications from Australian banks, each model was developed, trained, and assessed using performance metrics such as accuracy, precision, recall, F1-score, and the confusion matrix.

Among the three models, the Decision Tree classifier demonstrated the most effective overall performance. Its ability to handle non-linear relationships, accommodate both numerical and categorical features, and provide a transparent, rule-based structure made it particularly well-suited for this classification task. Not only did it deliver high accuracy, but its interpretability also aligned with the financial sector's requirements for explainable decisions, especially when communicating outcomes to regulators or customers.

In comparison: Logistic Regression performed reasonably well and maintained strong interpretability. However, its assumption of linearity

limited its ability to capture complex feature interactions within the data.

K-Nearest Neighbors (KNN) produced acceptable results but struggled with computational efficiency and sensitivity to feature scaling, making it less ideal for large or high-dimensional datasets.

The results of this study confirm that Decision Tree models are especially appropriate for home loan approval prediction tasks—offering a balance of accuracy, flexibility, and transparency. With appropriate parameter tuning and pruning techniques, a Decision Tree-based system can be deployed as a dependable component in automated credit risk assessment platforms.

## 6. REFERENCES

1. Quinlan, J. R. (1986). Induction of decision trees. Machine Learning, 1(1), 81–106.

2.Hosmer, D. W., Lemeshow, S., & Sturdivant, R. X. (2013). Applied Logistic Regression (Vol. 398). John Wiley & Sons.

3.Cover, T., & Hart, P. (1967). Nearest neighbor pattern classification. IEEE Transactions on Information Theory, 13(1), 21–27.

4. Han, J., Kamber, M., & Pei, J. (2011). Data Mining: Concepts and Techniques. Morgan Kaufmann.

5.James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). An Introduction to Statistical Learning. Springer.

6.Zhang, H. (2000). The Optimality of Naive Bayes. Proceedings of the Seventeenth International Florida Artificial Intelligence Research Society Conference.

7.Breiman, L. (2001). Random forests. Machine Learning, 45(1), 5–32.

8.Pedregosa, F., et al. (2011). Scikit-learn: Machine learning in Python. Journal of Machine Learning Research, 12, 2825–2830.

9.Australian Lending Dataset (1988). Statlog (German Credit Data) Dataset. UCI Machine Learning Repository.

10.Witten, I. H., Frank, E., & Hall, M. A. (2016). Data Mining: Practical Machine Learning Tools and Techniques. Morgan Kaufmann.

11.Domingos, P. (2012). A few useful things to know about machine learning. Communications of the ACM, 55(10), 78–87.

12.Tan, P. N., Steinbach, M., & Kumar, V. (2018). Introduction to Data Mining. Pearson Education.

13.Biau, G. (2012). Analysis of a random forests model. Journal of Machine Learning Research, 13, 1063–1095.

14.Ng, A. (2011). Machine learning yearning. DeepLearning.ai.

15.Friedman, J., Hastie, T., & Tibshirani, R. (2001). The Elements of Statistical Learning. Springer.