“PREDICTING LOAN ELIGIBILITY : A Machine Learning Approach”

***A project report submitted to***

***MALLA REDDY UNIVERSITY***

***in partial fulfilment of the requirements for the award of degree of***

**BACHELOR OF TECHNOLOGY**

**In**

**COMPUTER SCIENCE AND ENGINEERING (AI & ML)**

**Submitted by**

**Ranjith.K : 2111CS020394**

**Tejaswini.P : 2111CS020395**

**Ravinder.S.R : 2111CS020396**

**Renuka.N : 2111CS020397**

**Revanth Reddy.A : 2111CS020398**

**Rezwan Ali Sheik : 2111CS020399**

***Under the Guidance of***

**Prof.CH. Mohan Babu**

**Asst.Professor**

**DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING(AI&ML)** 

**2023**



**COLLEGE CERTIFICATE**

This is to certify that this is the bonafide record of the application development entitled,”**PREDICTING LOAN ELIGIBILITY : A Machine Learning Project**” Submitted by Ranjith.k (2111cs020294) , Tejaswini.p (2111cs020395) , Ravinder.S.R (2111cs020396) , Renuka.N (2111cs020397) , Revanth Reddy .A (2111cs020398) , Rezwan Ali Sheik (2111cs020399) B. Tech II year II semester, Department of CSE (AI&ML) during the year 2022-23. The results embodied in the report have not been submitted to any other university or institute for the award of any degree or diploma

**PROJECT GUIDE HEAD OF THE DEPARTMENT**

**Prof. Mohan Babu**  **Dr. Thayyaba Khathon**

**CSE(AI&ML)**



**ACKNOWLEDGEMENT**

We would like to express our gratitude to all those who extended their support and suggestions to come up with this application. Special Thanks to our **mentor Prof CH. Mohan Babu** whose help and stimulating suggestions and encouragement helped us all time in the due course of project development.

We sincerely thank our **HOD Dr. Thayyaba Khatoon** for her constant support and motivation all the time. A special acknowledgement goes to a friend who enthused us from the back stage. Last but not the least our sincere appreciation goes to our family who has been tolerant, understanding our moods, and extending timely support.

|  |  |  |
| --- | --- | --- |
| CHAPTER NO. | TITLE | PAGE NO |
| 1. | INTRODUCTION  1.1 Abstract  1.2 Limitations of project | 1-2 |
| 2. | ANALYSIS   * 1. Software requirement specification 2.1.1 Software requirement   2.1.2 Hardware requirement  2.2 Existing System  2.3 Proposed Systems  2.4 Modules  2.5 Architecture | 3-8 |
| 3. | DESIGN    3.1 DFD/ER/UML diagram  (Any other project diagram) | 9-10 |
| 4. | DEPLOYMENT AND RESULTS    4.1 GUI Screens / Output Screens  4.2 Code |  |
| 5. | CONCLUSION:    5.1 Project conclusion  5.2Future enhancement |  |

**CONTENTS**

**1. INTRODUCTION**

**1.1 ABSTRACT**

The goal of this ML project is to build a model that accurately predicts how much loan a user is eligible for based on their personal and employment details, including their marital status, education, number of dependents, and employment status. This project will use machine learning techniques to build a linear regression model that can analyze and identify patterns in data, to make accurate predictions.

Ones the data is cleaned, a linear regression model will be built using the personal and employment details of users as input features and their corresponding loan amounts as output. The model will be trained on a subset of the dateset and then tested on a separate validation set to assess its accuracy. This ML project has the potential to make the loan application process more efficient and accurate. By automating the loan eligibility process, it can help reduce the workload for loan officers and improve the accuracy of loan decisions.

In conclusion, this ML project can help streamline the loan application process and provide more accurate loan decisions. By automating the loan eligibility process, it can reduce the workload on loan officers and improve the accuracy of loan decisions. Additionally, it can help users better understand their loan eligibility and work on areas that can help them improve their chances of being approved for a loan.

**1.2 LIMITATIONS OF THE PROJECT**

1. **Data Quality:** The accuracy of the prediction model heavily depends on the quality of the data. If the data is incomplete, inconsistent, or inaccurate, it can negatively impact the performance of the model.
2. **Bias**: The model may be biased towards certain groups of people or certain types of loans, based on the data used for training the model. This can lead to unfair and discriminatory loan eligibility predictions.
3. **Overfitting:** Overfitting is a common problem in machine learning, where the model performs well on the training data but fails to generalize to new data. Overfitting can occur if the model is too complex or if there is insufficient data to train the model.
4. **Model Explainability**: Machine learning models can be complex, making it difficult to understand how they arrived at a particular loan eligibility prediction This can make it hard for end-users to trust the predictions and may lead to issues with regulatory compliance.
5. **Data Privacy**: Loan application data contains sensitive personal and financial information, which must be protected from unauthorized access and misuse. Ensuring data privacy and security is crucial for the success of the project.
6. **Regulatory Compliance**: The loan eligibility prediction project must comply with various regulatory requirements, such as the Fair Credit Reporting Act (FCRA) and the General Data Protection Regulation (GDPR). Failure to comply with these regulations can result in legal and financial consequences.
7. **Dynamic Market Conditions**: The lending industry is constantly evolving, with changing market conditions and customer behavior. The loan eligibility prediction model must be able to adapt to these changes and remain relevant over time.

**2. ANALYSIS**

**2.1 SOFTWARE REQUIREMENT SPECIFICATION**

**2.1.1 Software Requirement**

* Python language
* Jupyter notebook
* Machine learning libraries
* Data management libraries
* Database management system
* Web framework
* Cloud platform

**2.1.2 Hardware Requirement**

* Processor: Intel Core i3 or higher
* RAM: 4GB or higher
* Hard disk space: 500MB or higher
* Internet connection

**2.2** **EXISTING SYSTEM**

Existing systems for predicting loan eligibility include traditional credit scoring models, which rely on factors such as credit history, debt-to-income ratio, and payment history. These models are often rule-based and have limited flexibility to adapt to changing market conditions or individual borrower characteristics.

**2.3 PROPOSED SYSTEM**

Proposed systems for predicting loan eligibility using a machine learning approach aim to overcome these limitations by incorporating more diverse data sources and leveraging the power of complex algorithms to identify patterns and trends in the data.

Proposed systems for loan eligibility prediction using machine learning can also incorporate additional data sources such as social media activity, spending habits, or demographic information. These data sources can provide a more complete picture of the borrower's financial situation and help identify patterns and trends that may not be captured by traditional credit scoring models.

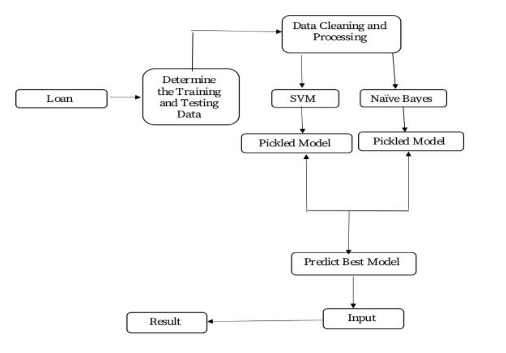
It has advantages like:

* Time period for loan transaction would be reduced.
* Whole process would be automated, so human errors would be reduced.
* Eligible candidate will be sanctioned loan without any delay.

**2.4 MODULES**

1. Data Preprocessing: This module would be responsible for handling missing data, encoding categorical variables, scaling numerical variables, and performing feature engineering.
2. Feature Selection: This module would be responsible for selecting the most relevant features based on correlation analysis, domain knowledge, or other criteria.
3. Model Selection: This module would be responsible for comparing different machine learning models, selecting the most appropriate model for the given problem, and evaluating its performance.
4. Hyperparameter Tuning: This module would be responsible for optimizing the hyperparameters of the selected model using techniques such as grid search or Bayesian optimization.
5. Model Deployment: This module would be responsible for deploying the trained model in a production environment where it can be used to make loan eligibility predictions in real-time.
6. User Interface: This module would be responsible for presenting the loan eligibility predictions to the end-user in an intuitive and user-friendly way, possibly using a web or mobile application
7. Data Visualization: This module would be responsible for visualizing the input data, model performance metrics, and other relevant information using graphs, charts, or other visual aids.
8. Model Explainability: This module would be responsible for providing explanations of how the model arrived at its predictions, possibly using techniques such as feature importance, partial dependence plots, or SHAP values.

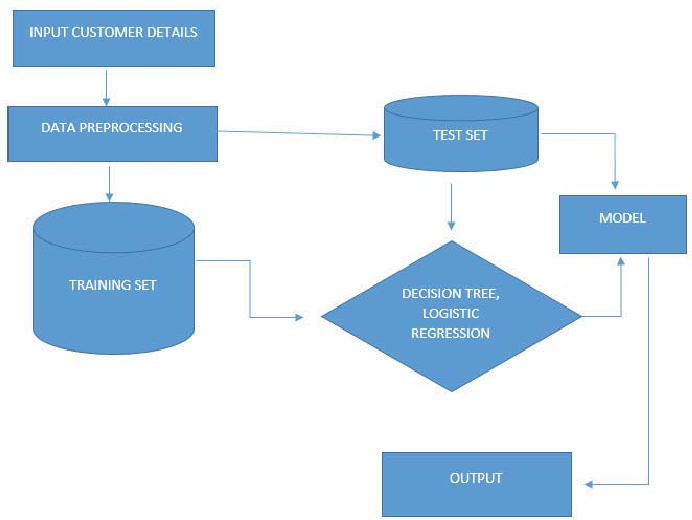
**2.5 ARCHITECTURE**

****

2.5 figure: Architecture of “predicting loan eligibility: a ML approach”

**3.DESIGN**

**3.1 DFD of project**

****

3.1 figure: DFD diagram of “predicting loan eligibility: a ML approach”

**3.2 DATA SET DESCRIPTION**

The Loan Prediction Problem Dateset is an immensely popular resource for machine learning projects related to loan prediction. Created by the esteemed data science community, Analytic Vidhya in India, this dateset contains comprehensive information about loan applicants such as their gender, marital status, education, income, loan amount, loan term, credit history, and much more. Its 614 rows and 13 columns provide ample data for various machine learning algorithms, and the presence of both categorical and numerical variables makes it an excellent learning resource for beginners.

The target variable in this dateset is the Loan\_Status column, which indicates whether a loan has been approved (Y) or not (N). While it's a fantastic dateset, it may require some per-processing and data cleaning before machine learning models can be applied. This is why data scientists and machine learning practitioners use this dateset to experiment with diverse algorithms and techniques that can help improve the accuracy of loan prediction models.

In conclusion, the Loan Prediction Problem dateset is a valuable resource for anyone interested in machine learning and loan prediction. Its comprehensiveness and ample data offer a great opportunity to explore the application of machine learning algorithms in predicting loan approvals.

**3.3 METHODS AND ALGORITHMS**

The loan prediction problem can be tackled using various machine learning algorithms. One of the most commonly used algorithms for binary classification problems, such as loan prediction, is logistic regression. Logistic regression estimates the probability of a binary outcome based on input variables, in this case, whether a loan will be approved or not.

To prepare the data for the logistic regression model, several data per-processing techniques can be used. Missing values can be handled by either imputing the mean or mode of the corresponding variable. Categorical variables can be encoded using one-hot encoding, which creates a binary column for each category of the variable. Numerical variables can be scaled using Mineralogical to ensure that all variables have a similar range.

In addition to data per-processing, cross-validation can be used to estimate the accuracy of the logistic regression model. Cross-validation is a technique used to assess the performance of a model by splitting the data into multiple training and testing sets and evaluating the model's performance on each of them.

Overall, a combination of data per-processing techniques and logistic regression algorithm can be used to predict the loan status (approved or not) based on various input variables such as gender, marital status, education, income, loan amount, loan term, credit history, and more. By using these techniques, the loan prediction model can achieve high accuracy and help banks and other financial institutions to make informed decisions

**3.4 BUILDING A MODEL**

**Data Cleaning and Pte-processing:** This involves removing missing values, handling outliers, and converting categorical variables into numeric form.

**Feature Selection/Engineering:** Selecting or creating the most important features that will be used in the model.

**Splitting Data into Training and Testing Sets:** The dateset is split into training and testing sets so that the model can be trained on a subset of the data and tested on another subset to evaluate its performance.

**Scaling Numerical Features:** Numerical features are often scaled to ensure that they have the same range and are on the same scale.

**Building the Model:** Once the data is per-processed and the features are selected or engineered, a machine learning model is chosen and trained on the training set.

**Evaluating the Model:** The trained model is evaluated on the testing set to measure its performance. The evaluation metrics may include accuracy, precision, recall, and F1 score.

**Tuning the Model:** The hyper-parameters of the model are tuned to improve its performance on the testing set

1. **DEPLOYMENT AND RESULTS**
   1. **SOURCE CODE**

import pandas as pd

from sklearn.model\_selection import train\_test\_split

from sklearn.linear\_model import LogisticRegression

from sklearn.metrics import accuracy\_score

from sklearn.preprocessing import StandardScaler

loan\_data = pd.read\_csv('loan\_data.csv')

loan\_data=loan\_data.dropna()

loan\_data['Married'] = loan\_data['Married'].map({'No': 0, 'Yes': 1})

loan\_data['Education'] = loan\_data['Education'].map({'Not Graduate': 0, 'Graduate': 1})

loan\_data['Self\_Employed'] = loan\_data['Self\_Employed'].map({'No': 0, 'Yes': 1})

loan\_data['LoanAmount'] = loan\_data['LoanAmount']

loan\_data['Loan\_Amount\_Term'] = loan\_data['Loan\_Amount\_Term']

loan\_data['Credit\_History'] = loan\_data['Credit\_History']

X = loan\_data[['Married', 'Education', 'Self\_Employed', 'Credit\_History', 'Property\_Area', 'LoanAmount']]

y = loan\_data['Loan\_Status']

X = pd.get\_dummies(X)

scaler = StandardScaler()

num\_cols = ['LoanAmount']

X[num\_cols] = scaler.fit\_transform(X[num\_cols])

X\_train, X\_test, y\_train, y\_test = train\_test\_split(X, y, test\_size=0.2, random\_state=42)

model = LogisticRegression()

model.fit(X\_train, y\_train)

y\_pred = model.predict(X\_test)

accuracy = accuracy\_score(y\_test, y\_pred)\*100

print("This ML Model Has Genrated the results with accuracy of", accuracy)

* 1. **FINAL RESULT**

This ML Model Has Genrated the results with accuracy of 82.29166666666666

1. **CONCLUSION:**

**5.1 Project conclusion**

In conclusion, predicting loan eligibility is a crucial challenge in the banking and financial industry. Machine learning algorithms provide a promising solution to this problem, with logistic regression being one of the most effective techniques for predicting the likelihood of loan approval based on input variables. However, accurate predictions rely on the quality of data per-processing techniques such as imputation, encoding, and scaling, which play a critical role in preparing data for model training. Additionally, cross-validation can be used to estimate the accuracy of the model, ensuring its effectiveness in real-world scenarios.

By leveraging these techniques, banks and financial institutions can make informed decisions on loan approvals, enabling better risk management and enhancing customer satisfaction. Furthermore, reliable loan prediction models can help prevent fraudulent activities and reduce the likelihood of loan defaults. Therefore, the loan prediction problem is a vital area of research in the machine learning domain, and continued advancements in this field can significantly benefit both financial institutions and customers alike.

**5.2 Future enhancement**

**1.Incorporating alternative data sources:**

Traditional loan prediction models use a limited set of variables such as income, credit score, and employment history. However, the use of alternative data sources such as social media activity, mobile phone usage, and online transactions could provide additional insights into a borrower's creditworthiness, enabling more accurate loan decisions.

**2.Utilizing advanced machine learning algorithms:**

While logistic regression is an effective technique for loan prediction, more advanced machine learning algorithms such as random forests, gradient boosting, and neural networks could potentially yield better results. These algorithms can handle more complex data structures and relationships and can identify non-linear patterns that may be missed by linear models.

**3.Incorporating temporal data**:

Loan prediction models often overlook the temporal aspect of loan repayment. By incorporating time-series analysis techniques, models can analyze how a borrower's financial behavior changes over time, and identify patterns that may indicate future default.

**4.Integrating Explainable AI (XAI) techniques:**

Loan decisions can have a significant impact on an individual's financial well-being. Therefore, it is essential to develop transparent and explainable models that can help borrowers understand why a loan decision was made. Explainable AI techniques such as SHAP (SHapley Additive exPlanations) or LIME (Local Interpretable Model-Agnostic Explanations) can provide insights into the features that contributed most to the loan decision, enabling better transparency and accountability.