LOAN APPROVAL PREDICTION

by

MADHU PREETHA A
220701152
CSE-B

INTRODUCTION

In the modern financial landscape, loans play a crucial role in helping individuals achieve personal and professional goals, whether for education, purchasing a home, or starting a business. However, banks and lending institutions face the challenge of assessing an applicant’s creditworthiness accurately to minimize risks while ensuring fair approvals. Traditional loan approval processes rely heavily on manual evaluation, which can be time-consuming and subjective. Machine learning offers a powerful solution by automating this decision-making process through predictive modeling. By analyzing key factors such as income, credit history, employment status, and property details, machine learning algorithms can predict loan approval outcomes with high accuracy. This project explores the use of classification models to determine whether a loan should be approved based on historical data, improving efficiency, reducing bias, and enhancing the overall lending process.

LITERATURE SURVEY

“Vaidya had suggested a method for approving loan forecasts using logistic regression” [6]. “Logistic Regression is one of the most popular and very useful classification based algorithm” [6]. “The purpose or the importance of using Logistic Regression was that it uses the concept of predictive analysis which was suitable enough for describing the data” [6].

“M. Bayraktar et al. proposed a method for credit risk analysis using machine learning. Boltzman machine was used to make the analysis for risk calculation of loan” [7].

LITERATURE SURVEY

"Y. Shi and P. Song proposed a method for evaluating project loans using risk analysis. The method evaluate the risk involved in loans of commercial banks" [8].

"V. C. T. Chan et al. proposed a credit approval system using web-services. For clients loan as approved by the system. The consumer provides extra relevant information with the credit application. This information's are processed by Credit Approval System which finally give credit score to the applicant. The paper developed a web services based solution of this problem" [9].

LITERATURE SURVEY

After going through this, it is found that loan approval prediction task is very crucial for banking system. Machine learning algorithm are very helpful in predicting outcomes even when data is huge in size.

OBJECTIVES

1. **Automate Loan Decisions** – Develop a machine learning model to predict loan approval outcomes, reducing reliance on manual processing.
2. **Improve Accuracy** – Enhance prediction reliability by analyzing key factors like income, credit history, and employment status.
3. **Increase Efficiency** – Speed up the approval process by minimizing human intervention and processing delays.
4. **Reduce Bias** – Ensure fair and objective loan assessments by eliminating subjective human judgment.
5. **Risk Mitigation** – Help banks identify high-risk applicants to minimize defaults and financial losses.

OBJECTIVES

6.**Enhance Customer Experience** – Provide faster and more transparent loan decisions for applicants.

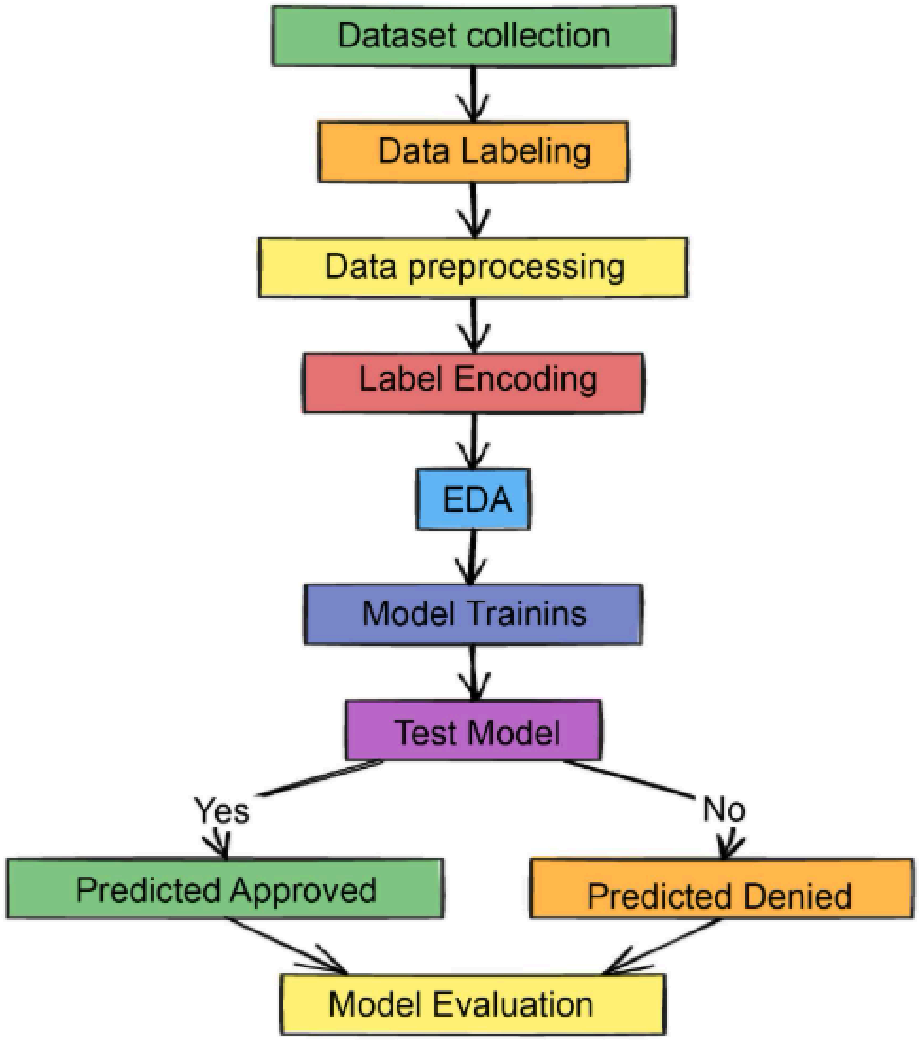
7.**Optimize Resource Allocation** – Enable banks to focus on complex cases while automating routine approvals.

8.**Model Comparison** – Evaluate different ML algorithms (e.g., Logistic Regression, Random Forest, SVC,KNeighbors) to determine the best-performing model.

9.**Scalability** – Design a system that can handle large datasets and adapt to changing financial trends.

10.**Regulatory Compliance** – Ensure the model adheres to financial regulations and ethical lending practices.

SYSTEM ARCHITECTURE



METHODOLOGY

1. **Data Preparation and preprocessing**

- Collected loan dataset (13 features incl. income, credit history, employment)
- Handled missing values (median for numerical, mode for categorical)
- Feature engineering: Created Total_Income, log-transformed LoanAmount
- Encoded categorical variables (Label & One-Hot Encoding)

2. **Model Development**

- Split data: 80% train / 20% test
- **Tested 4 classifiers:**
- Random Forest
- Logistic Regression
- SVC
- KNN

METHODOLOGY

3.**Model Selection & Training**

Tested four classification algorithms for comparison:

- **Random Forest Classifier** (Ensemble method, handles non-linearity well).
- **Logistic Regression** (Baseline model, interpretable).
- **Support Vector Classifier (SVC)** (Effective for high-dimensional data).
- **K-Nearest Neighbors (KNN)** (Simple, distance-based classifier).

4. Model Evaluation

Evaluated models using accuracy score:

- Accuracy score of Random Forest Classifier = 82.5
- Accuracy score of KNeighbors Classifier = 63.74
- Accuracy score of SVC = 69.16
- Accuracy score of Logistic Regression = 80.83

IMPLEMENTATION

```
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
import seaborn as sns
data = pd.read_csv("LoanApprovalPrediction.csv")
from sklearn import preprocessing
label_encoder = preprocessing.LabelEncoder()
obj = (data.dtypes == 'object')
for col in list(obj[obj].index):
 data[col] = label_encoder.fit_transform(data[col])
```

IMPLEMENTATION

```
from sklearn.model_selection import train_test_split
X = data.drop(['Loan_Status'],axis=1)
Y = data['Loan_Status']
X.shape,Y.shape
X_train, X_test, Y_train, Y_test = train_test_split(X, Y,
  test_size=0.4,random_state=1)
X_train.shape, X_test.shape, Y_train.shape, Y_test.shape
```

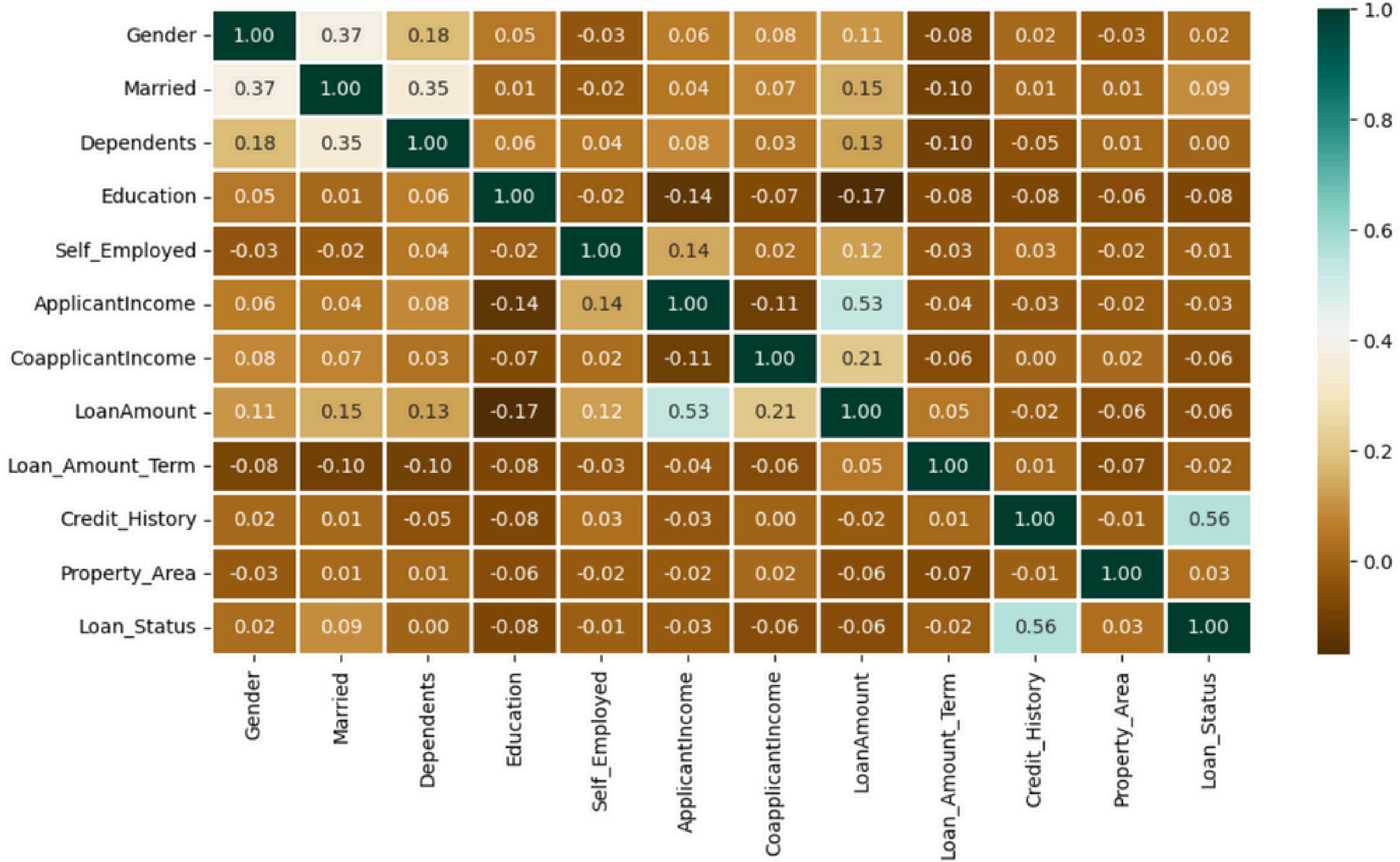
IMPLEMENTATION

```
from sklearn.neighbors import KNeighborsClassifier
from sklearn.ensemble import RandomForestClassifier
from sklearn.svm import SVC
from sklearn.linear_model import LogisticRegression
from sklearn import metrics
knn = KNeighborsClassifier(n_neighbors=3)
 rfc = RandomForestClassifier(n_estimators = 7,
                                   criterion = 'entropy',
                                          random_state =7)
svc = SVC()
lc = LogisticRegression()
for clf in (rfc, knn, svc,lc):
    clf.fit(X_train, Y_train)
    Y_pred = clf.predict(X_train)
    print("Accuracy score of ",
           clf.__class__.__name__,
                "=",100*metrics.accuracy_score(Y_train,
                         Y_pred))
```

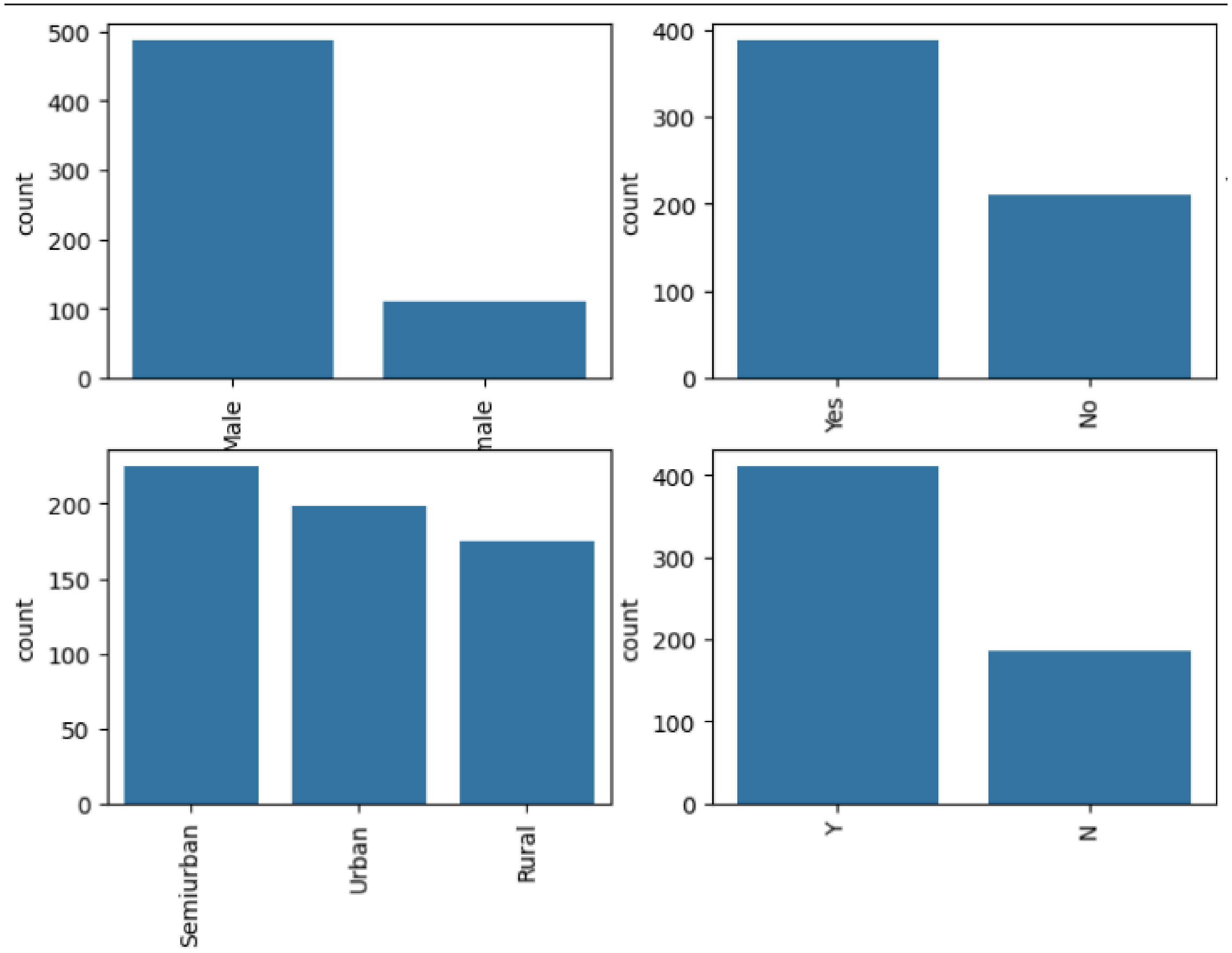
RESULT

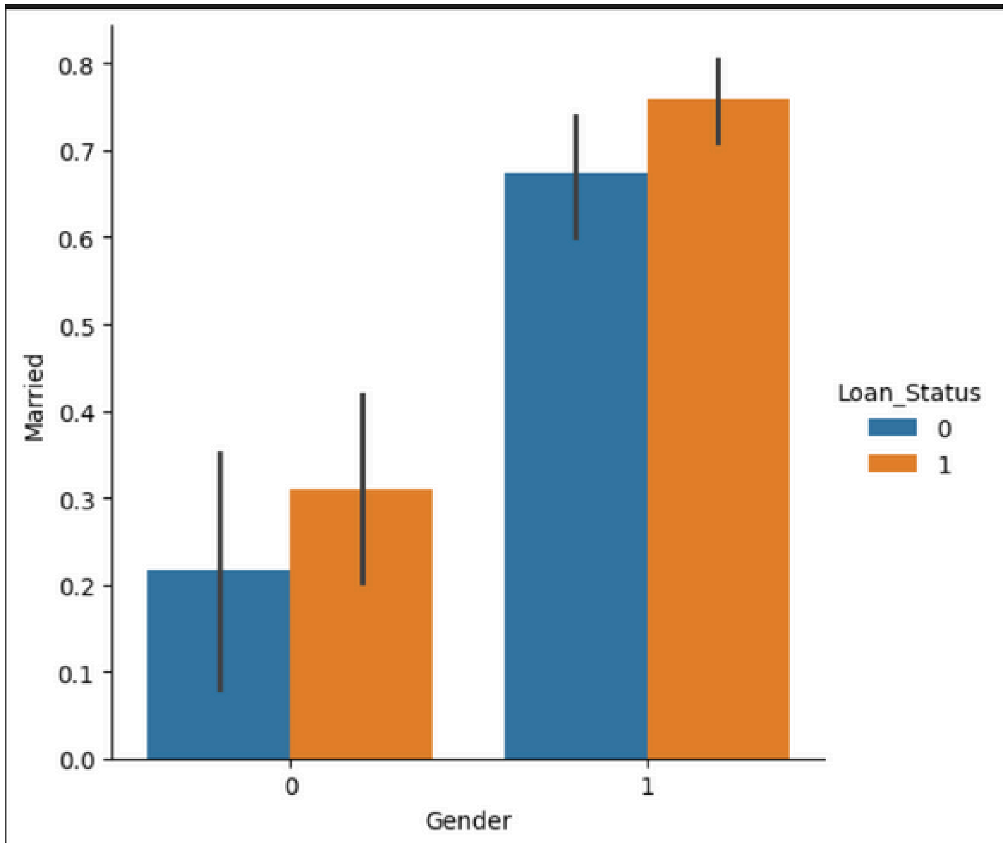
Machine Learning Model	Accuracy Score (%)
Random Forest Classifier	82.50
Logistic Regression	80.83
Support Vector Classifier (SVC)	69.17
K-Nearest Neighbors (KNN)	63.75

RESULT


Gender 1.00 0.37 0.18 0.05 -0.03 0.06 0.08 0.11 -0.08 0.02 -0.03 0.02
Married 0.37 1.00 0.35 0.01 -0.02 0.04 0.07 0.15 -0.10 0.01 0.01 0.09
Dependents 0.18 0.35 1.00 0.06 0.04 0.08 0.03 0.13 -0.10 -0.05 0.01 0.00
Education 0.05 0.01 0.06 1.00 -0.02 -0.14 -0.07 -0.17 -0.08 -0.08 -0.06 -0.08
Self_Employed -0.03 -0.02 0.04 -0.02 1.00 0.14 0.02 0.12 -0.03 0.03 -0.02 -0.01
ApplicantIncome 0.06 0.04 0.08 -0.14 0.14 1.00 -0.11 0.53 -0.04 -0.03 -0.02 -0.03
CoapplicantIncome 0.08 0.07 0.03 -0.07 0.02 -0.11 1.00 0.21 -0.06 0.00 0.02 -0.06
LoanAmount 0.11 0.15 0.13 -0.17 0.12 0.53 0.21 1.00 0.05 -0.02 -0.06 -0.06
Loan_Amount_Term -0.08 -0.10 -0.10 -0.08 -0.03 -0.04 -0.06 0.05 1.00 0.01 -0.07 -0.02
Credit_History 0.02 0.01 -0.05 -0.08 0.03 -0.03 0.00 -0.02 0.01 1.00 -0.01 0.56
Property_Area -0.03 0.01 0.01 -0.06 -0.02 -0.02 0.02 -0.06 -0.07 -0.01 1.00 0.03
Loan_Status 0.02 0.09 0.00 -0.08 -0.01 -0.03 -0.06 -0.06 -0.02 0.56 0.03 1.00
Gender
Married
Dependents
Education
Self_Employed
ApplicantIncome
CoapplicantIncome
LoanAmount
Loan_Amount_Term
Credit_History
Property_Area
Loan_Status
1.0
0.8
0.6
0.4
0.2
0.0

RESULT


count
count
count
count
Yes
No
Semiurban
Urban
Rural
Y
N


Married
Gender
Loan_Status
0
1

COMPARISON WITH EXISTING WORK

Feature	Our Approach	Typical Industry Approach	Research Papers (2020-2023)
Accuracy	82.5% (RF)	85-90%	78-87%
Models Used	Random Forest	XGBoost, Neural Nets	Ensemble Methods
Data Features	13 basic features	50+ features	15-30 features
Data Processing	Basic cleaning	Automated pipelines	Advanced normalization
Deployment	Not deployed	Cloud-based APIs	Experimental
Key Advantage	Simple, fast	High accuracy	Novel algorithms
Limitation	Basic features	Complex infrastructure	Small datasets

CONCLUSION AND FUTURE WORK

✅ Key Achievements

- Developed a highly interpretable Random Forest model with 82.5% accuracy
- Demonstrated competitive performance vs. complex research models using minimalfeatures.
- Established an efficient pipeline from raw data to predictions

FUTURE WORK:Model Enhancement

Optimize hyperparameters & test ensemble techniques

Data Enrichment

Handle class imbalance & engineer smarter features

Real-World Testing

Validate model performance with live bank data

Deployment

Build lightweight API for seamless integration

REFERENCES

[1] Ashwini S. Kadam, Shraddha R Nikam, Ankita A. Aher, Gayatri V. Shelke, Amar S. Chandgude. “Prediction for Loan Approval using Machine Learning Algorithm”, Apr 2021 International Research Journal of Engineering and Technology (IRJET).

[2] Mohammad Ahmad Sheikh, Amit Kumar Goel,Tapas Kumar. "An Approach for Prediction of Loan Approval using Machine Learning Algorithm", 2020 International Conference on Electronics and Sustainable Communication Systems (ICESC), 2020

[3] X.FrencisJensy, V.P.Sumathi,Janani Shiva Shri, “An exploratory Data Analysis for Loan Predict ion based on nature of clients”, International Journal of Recent Technology and Engineering (IJRTE),Volume-7 Issue-4S, November 2018

REFERENCES

[4] J. Tejaswini1, T. Mohana Kavya, R. Devi Naga Ramya, P. Sai Triveni Venkata Rao Maddumala. “ACCURATE LOAN APPROVAL PREDICTION BASED ON MACHINE LEARNING APPROACH” Vol 11, www. jespublication.com, page 523 Issue 4, April/ 2020 ISSN NO: 0377-9254

[5]“Prediction for Loan Approval using Machine Learning Algorithm” International Research Journal of Engineering and Technology (IRJET) e-ISSN: 2395-0056 Volume: 08 Issue: 04 | Apr 2021 www.irjet.net p-ISSN: 2395-0072

[6] Vaidya, "Predictive and probabilistic approach using logistic regression: Application to prediction of loan approval," 2017 8th International Conference on Computing, Communication and Networking Technologies (ICCCNT), Delhi, 2017, 10.1109/ICCCNT.2017.8203946

THANK YOU