CSE422 Lab Project Report

Spring 25

Student-1

Name: Digonta Das

ID: 22201633

Section : 14

Student-2

Name: Atif Islam Arian

ID: 22201230

Section: 14

Introduction

We have executed a Loan selecting Prediction utilizing diverse Machine Learning Calculations. This project will predict if someone can get a Loan or not. This is a Classification problem and There are 2247 data-points in our dataset .

Dataset description

- **Feature**: There are 16 features -

  Id - (Numerical)

  Loan_amnt - (Numerical)

  Term- (Categorical)

  Int_rate - (Numerical)

  Installment - (Numerical)

  Emp_length- (Categorical)

  Homw_ownership- (Categorical)

  Annual_inc - (Numerical)

  Verification_status- (Categorical)

  Purpose- (Categorical)

  Title- (Categorical)

  Zip_code- (Categorical)

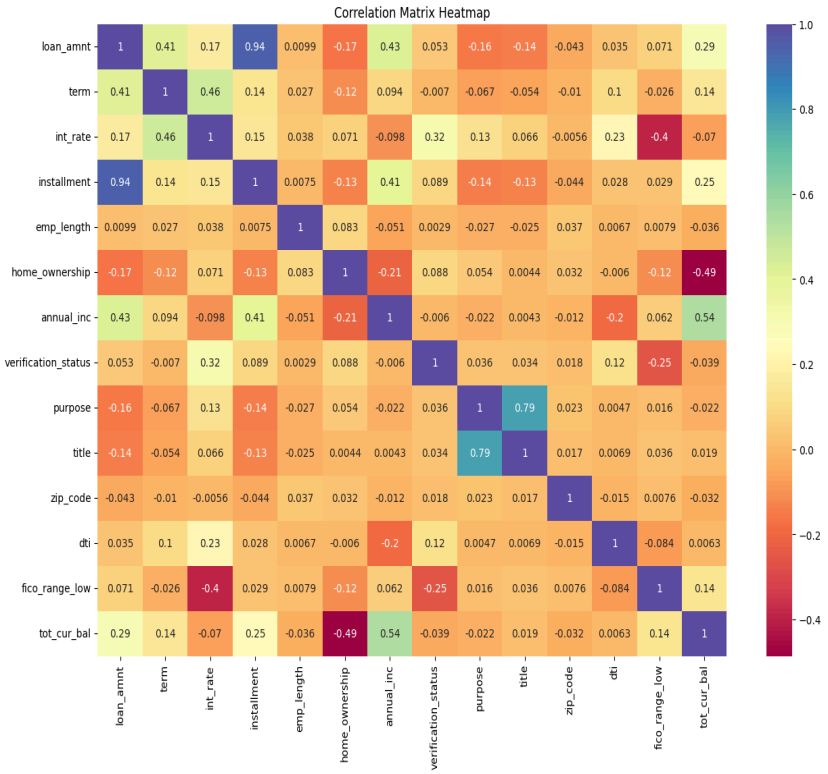
  Dti - (Numerical)

  Fico_range_low - (Numerical)

  Tot_cur_bal- total current balance - (Numerical)

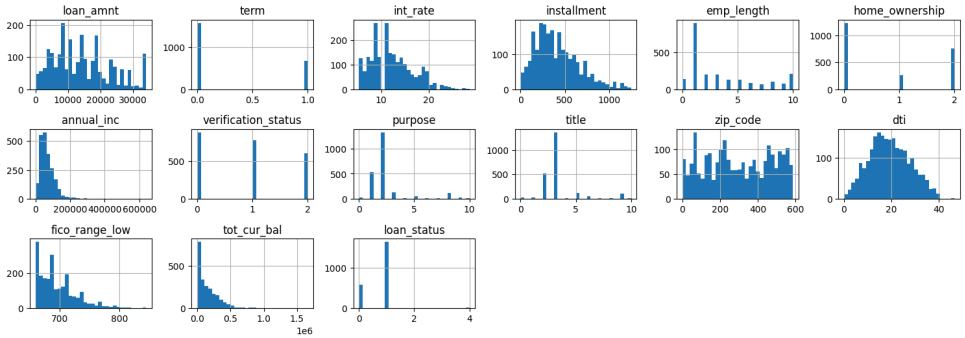
- **Label :** The label is the loan_status- (Categorical), which is a multiclass variable with Fully Paid, Charged Off, In Grace Period,Late (31-120 days), Late (16-30 days)

Correlation test:



The correlation heatmap shows that loan amount has a very strong positive correlation (.94) with installment, suggesting these two variables are highly interconnected. Annual income moderately correlates with loan amount(.43) indicating higher income individuals tend to receive larger loans.

Imbalanced Dataset :



The dataset exhibits noticeable class imbalance across credit categories or loan approval statuses. A frequency distribution analysis revealed that certain loan types (potentially term lengths or approval categories) are significantly overrepresented compared to others. For example, fully verified loans or short-term loans appear much more frequently than other categories. This imbalance could potentially skew model predictions and was carefully considered during model development through techniques like stratified sampling and class weighting.

.

Exploratory Data Analysis (EDA):

The dataset contains several categorical variables that demonstrate clear trends in loan outcomes. For example:

- Lower FICO scores correlate strongly with higher interest rates and smaller loan amounts.
- Home ownership status shows meaningful relationships with total credit balance and loan approval odds.
- Numerical distributions revealed several outliers in variables like annual income and loan amount that required normalization or capping.

- The correlation heatmap illustrated how verification status interacts moderately with interest rates, suggesting verified borrowers may receive different lending terms.
- Purpose of loans shows strong correlation with loan titles, indicating potential feature redundancy that could be addressed through dimension reduction.

Dataset pre-processing : For the data preprocessing,

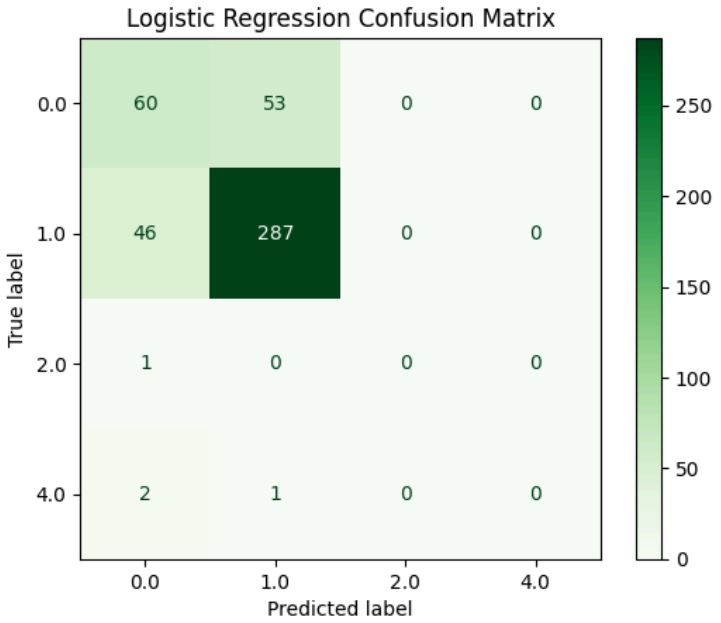
- **Faults:** we firstly checked all the null values and filled them then dropped the duplicate values using drop_duplicates() where we didn't find any duplicate values. Then we converted all the categorical values into numerical values and did our Imputation. For scaling we used StandardScaler which was applied to numerical features and also Minmaxscaler.

- **Solutions:** For the solutions we deleted the "Id" feature because the id was not affecting our loan_status which drastically improved our accuracy and lastly, used normalization after fixing it.

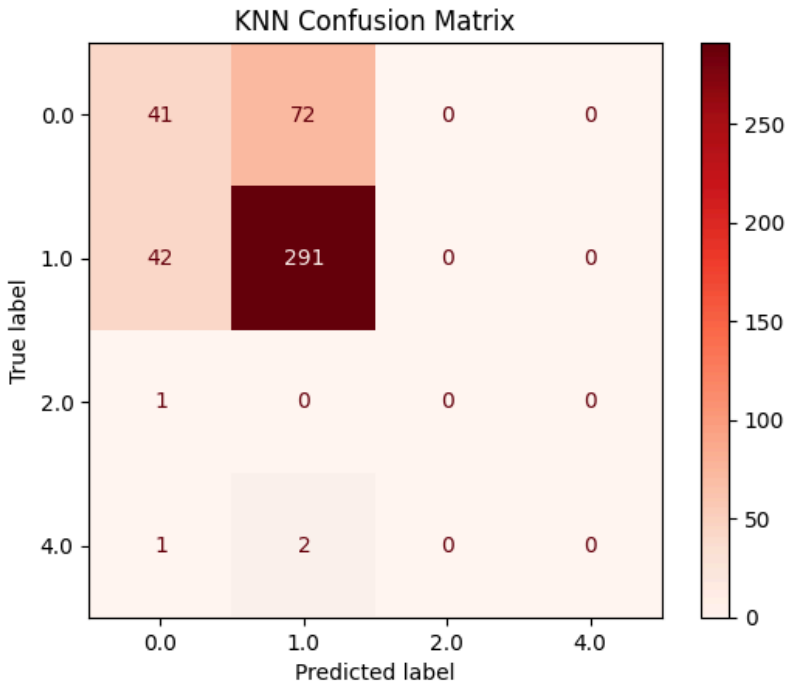
Dataset splitting: Dataset Split: Used a 70-30 train-test split with train_test_split()

Model selection

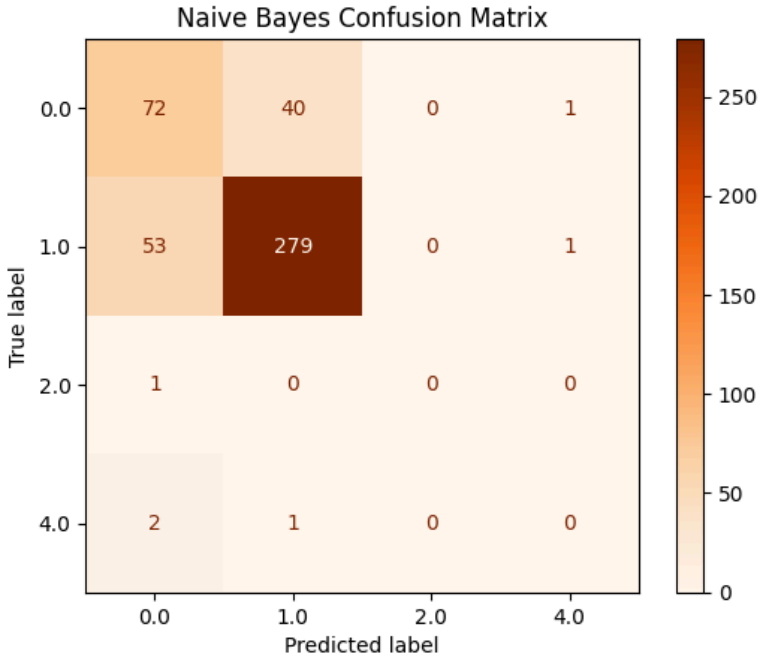
- **Logistic Regression**



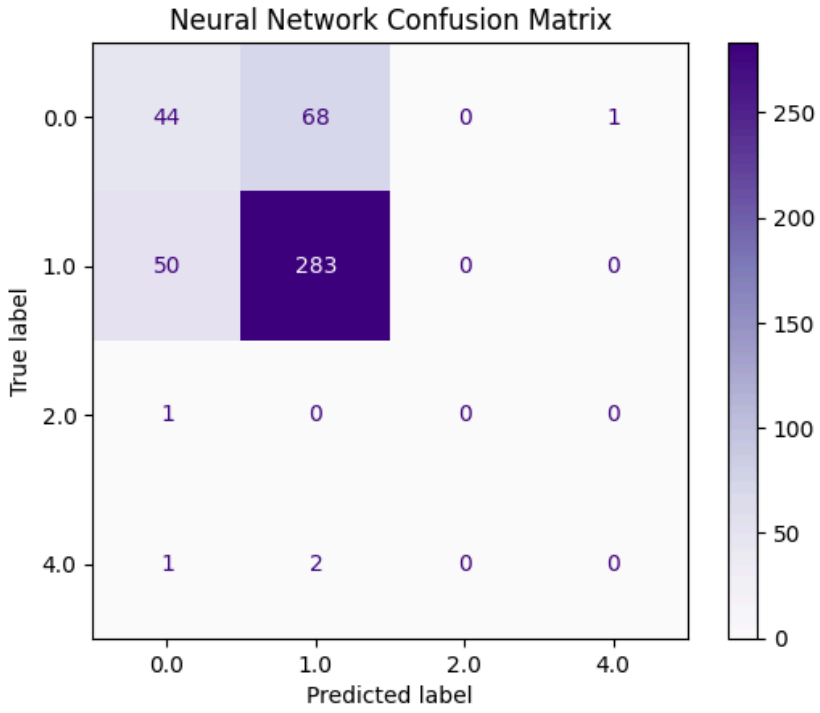
- **KNN**


KNN Confusion Matrix
41
72
0
0
42
291
0
0
1
0
0
0
1
2
0
0
0.0
1.0
2.0
4.0
True label
Predicted label
250
200
150
100
50
0

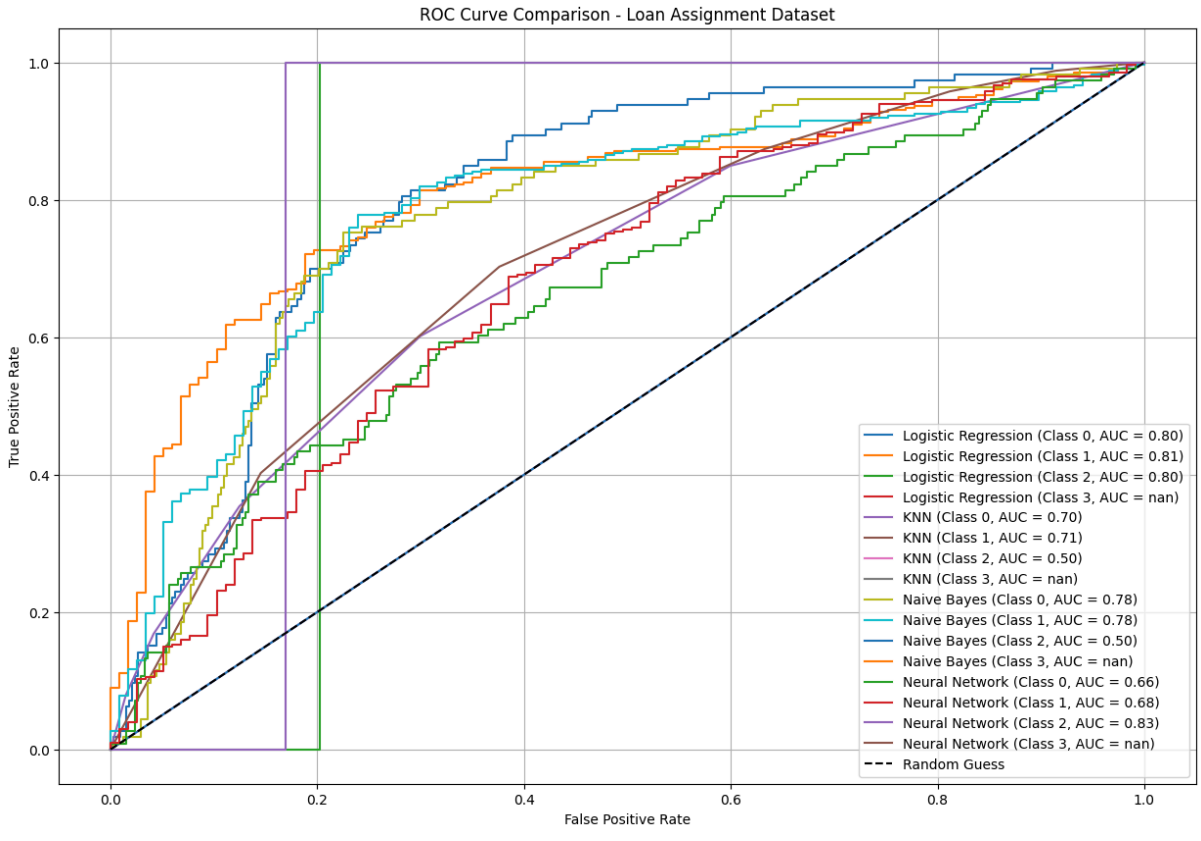
- **Naive Bayes**


Naive Bayes Confusion Matrix
72
40
0
1
53
279
0
1
1
0
0
0
2
1
0
0
0.0
1.0
2.0
4.0
True label
Predicted label
250
200
150
100
50
0

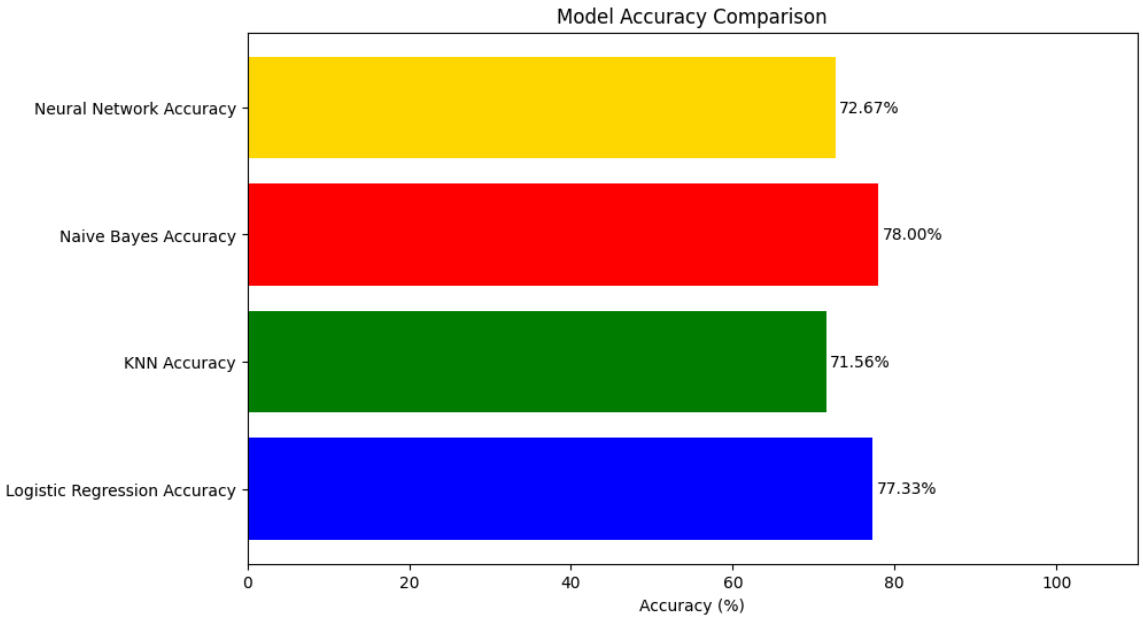
- **Neural Network (MLP Classifier)**



Comparison analysis:



1. **Logistic Regression :** Used a multiclass classification, accuracy was best as we can see from the ROC curve.
2. **KNN:** Moderate accuracy , got bad accuracy for the class 2 which is only .50
3. **Naive Bayes:** Better accuracy than KNN , got bad accuracy for the class 2 which is only .50 like the KNN model
4. **Neural Network:** Here we got the best accuracy for class 2 but bad accuracy for class 1 and 0



Conclusion: This project successfully implemented various classification models to predict loan outcomes. The analysis reveals that Logistic Regression consistently performed well across most classes achieving AUC scores of (0.80-0.81) that indicated strong predictive capability. Naive Bayes also demonstrated comparable performance

with AUC scores of 0.78 for multiple classes. But the Neural Network showed varied performance with its strongest result in Class 2 (AUC = 0.83) though it underperformed for Class 0 (AUC = 0.66). KNN models exhibited inconsistent performance across classes with AUC scores ranging from (0.50 - 0.71) suggesting this algorithm may not be optimal for this particular dataset. The Data preprocessing steps, feature engineering based on correlation analysis and handling of class imbalance contributed to the model's effectiveness. For future improvements the ensemble methods combining the strengths of Logistic Regression and Neural Networks could potentially enhance predictive power.