EC 9560 DATA MINING
LAB 01

PRIYATHARSINI S.

2020/E/122

SEMESTER 07

02 OCT 2024

Tittle:- Loan Approval Prediction

Objective

The objective of this project is to develop a supervised machine learning model that predicts whether a loan application will be approved or rejected. By tuning model parameters, applying cross-validation techniques, and evaluating model performance, the project aims to provide insights into loan approval prediction based on the provided dataset. The final goal is to accurately predict outcomes for an unlabeled test dataset using the optimized model.

Methodology

1. **Data Collection**
   The dataset selected for this project is sourced from the Kaggle competition "Loan Approval Prediction" created by Walter Reade and Ashley Chow in 2024.

   Dataset link: Loan Approval Prediction

2. **Data Preprocessing**

   - **Handling Missing Values**: Missing values will be identified and handled either by imputation (mean/median for numerical features, mode for categorical features) or by removing rows/columns with excessive missing data.
   - **Categorical Feature Encoding**: Categorical variables (such as Gender, Marital Status, etc.) will be converted into numerical form using techniques like One-Hot Encoding or Label Encoding.
   - **Feature Scaling**: Continuous numerical features will be standardized or normalized to ensure that they contribute equally to the model training.
   - **Feature Selection**: By Using correlation coefficient

3. **Splitting Data**

   - **Training Set**: Used to train the model (80% of the labeled dataset).
   - **Validation Set**: Used to evaluate model performance during hyperparameter tuning (20% of the labeled dataset).

4. **Model Selection**

   - Logistic Regression
   - Decision Trees
   - Random Forests
   - Support Vector Machines (SVM)

- Gradient Boosting Machines (GBM)

5. **Cross-Validation and Hyperparameter Tuning**

    - **K-Fold Cross-Validation**: The data will be split into K equal subsets. Each model will be trained on K-1 subsets and validated on the remaining subset. This process will be repeated K times to ensure robust performance evaluation.
    - **Hyperparameter Tuning**: Techniques such as Grid Search or Random Search will be used to find the best hyperparameters for each model, improving the prediction accuracy.

6. **Model Evaluation**
   After training and tuning the models, the performance will be evaluated using the following metrics:

    - **Accuracy**: Percentage of correctly predicted instances.
    - **Precision**: Ratio of correctly predicted positive observations to the total predicted positives.
    - **Recall**: Ratio of correctly predicted positive observations to all actual positives.
    - **F1 Score**: Harmonic mean of precision and recall, offering a balance between both metrics.
    - **Confusion Matrix**: Provides insight into the number of true positives, true negatives, false positives, and false negatives.

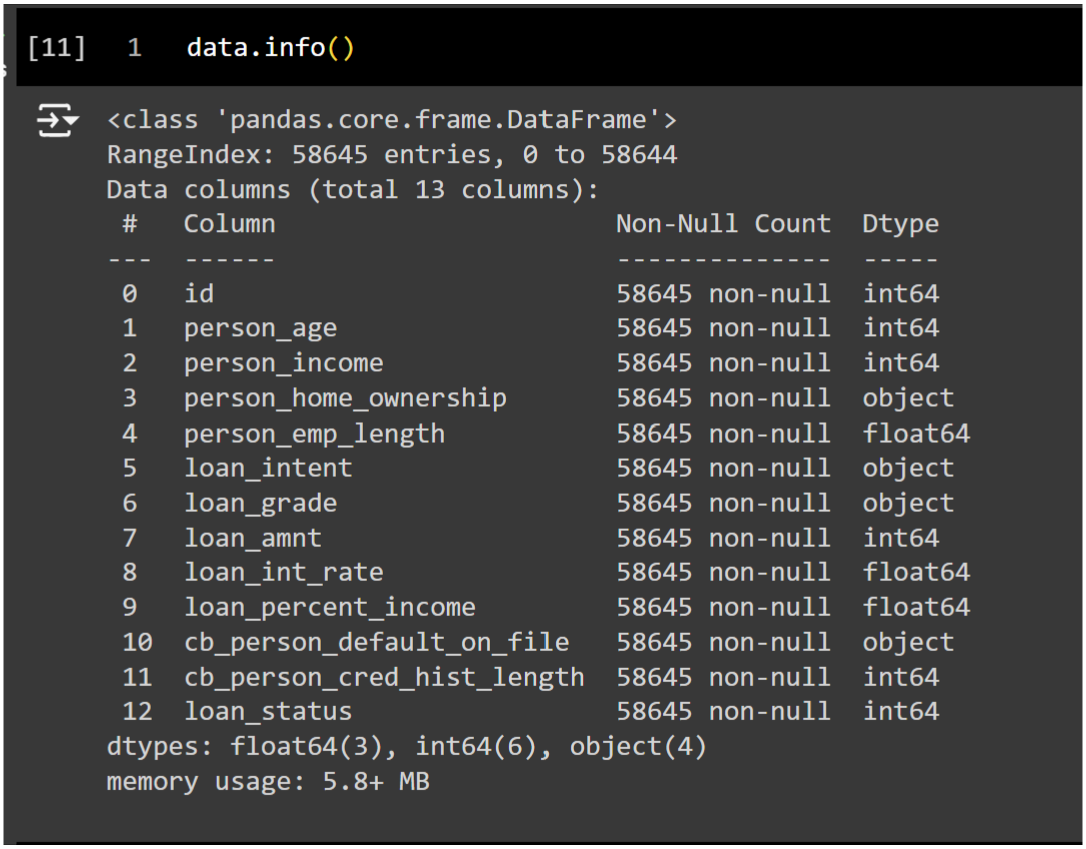
Data Description

The dataset used in this project is the "Loan Approval Prediction" dataset, which contains the following key features

Features:

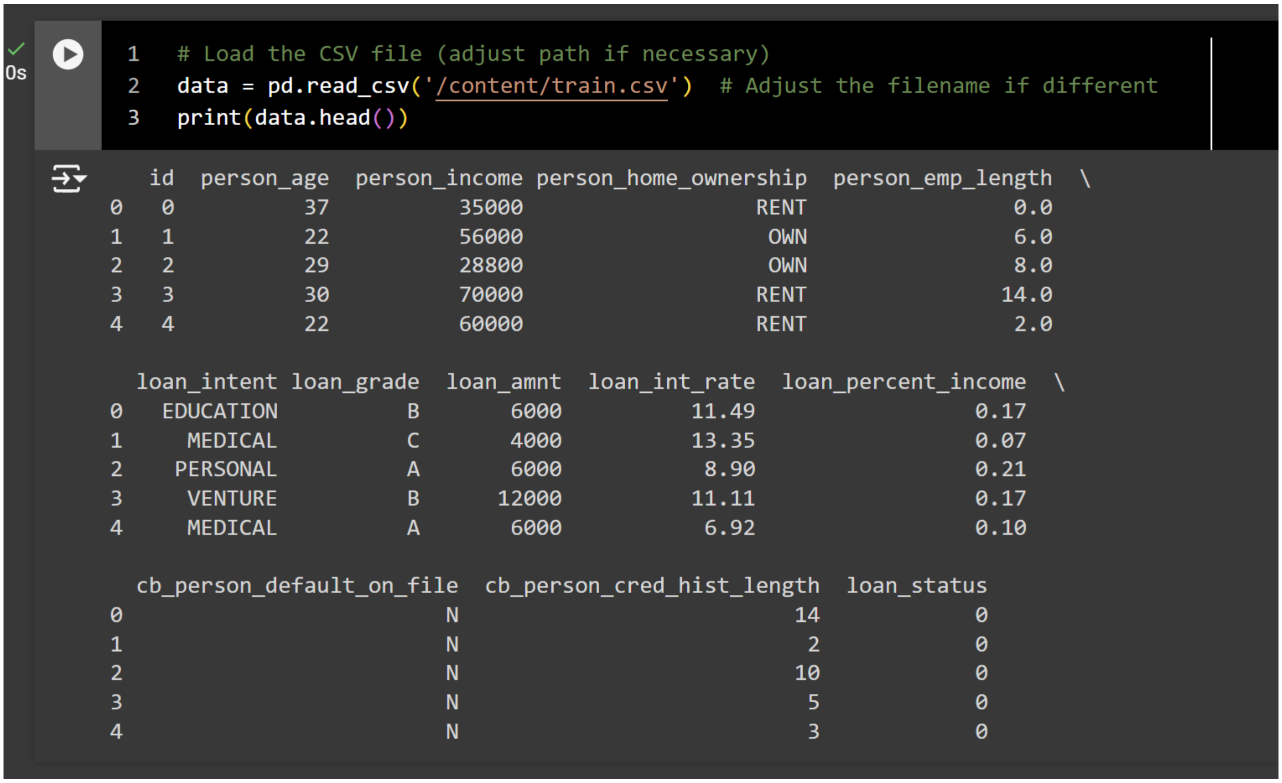
- **id**:Unique identifier for each loan application.
- **person_age**:Age of the applicant.
- **person_income**:Annual income of the applicant.
- **person_home_ownership**:Home ownership status (e.g., rent, own).
- **person_emp_length**:Length of employment in years.
- **loan_intent**:Purpose of the loan (e.g., personal, business, education).
- **loan_grade**:Grade assigned to the loan based on risk.

- **loan_amnt**:Amount of the loan requested.
- **loan_int_rate**:Interest rate on the loan.
- **loan_percent_income**:Percentage of income allocated for loan payments.
- **cb_person_default_on_file**:Indicator of past defaults (Yes/No).
- **cb_person_cred_hist_length**:Length of credit history in months.
- **loan_status**:Target variable indicating loan approval status (approved/rejected).



```
[11]   data.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 58645 entries, 0 to 58644
Data columns (total 13 columns):
 #   Column                      Non-Null Count  Dtype  
---  ------                      --------------  -----  
 0   id                          58645 non-null  int64  
 1   person_age                  58645 non-null  int64  
 2   person_income               58645 non-null  int64  
 3   person_home_ownership       58645 non-null  object 
 4   person_emp_length           58645 non-null  float64
 5   loan_intent                 58645 non-null  object 
 6   loan_grade                  58645 non-null  object 
 7   loan_amnt                   58645 non-null  int64  
 8   loan_int_rate               58645 non-null  float64
 9   loan_percent_income         58645 non-null  float64
 10  cb_person_default_on_file   58645 non-null  object 
 11  cb_person_cred_hist_length  58645 non-null  int64  
 12  loan_status                 58645 non-null  int64  
dtypes: float64(3), int64(6), object(4)
memory usage: 5.8+ MB
```

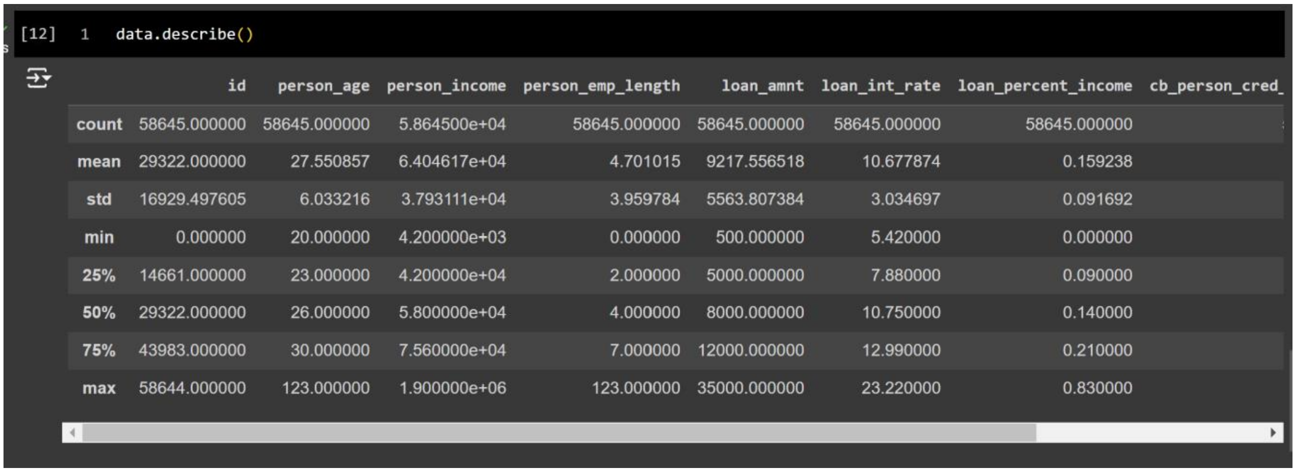


```python
# Load the CSV file (adjust path if necessary)
data = pd.read_csv('/content/train.csv')  # Adjust the filename if different
print(data.head())
```

```
   id  person_age  person_income person_home_ownership  person_emp_length  \
0   0          37          35000                  RENT                0.0
1   1          22          56000                   OWN                6.0
2   2          29          28800                   OWN                8.0
3   3          30          70000                  RENT               14.0
4   4          22          60000                  RENT                2.0

  loan_intent loan_grade  loan_amnt  loan_int_rate  loan_percent_income  \
0   EDUCATION          B       6000          11.49                 0.17
1     MEDICAL          C       4000          13.35                 0.07
2    PERSONAL          A       6000           8.90                 0.21
3     VENTURE          B      12000          11.11                 0.17
4     MEDICAL          A       6000           6.92                 0.10

  cb_person_default_on_file  cb_person_cred_hist_length  loan_status
0                         N                          14            0
1                         N                           2            0
2                         N                          10            0
3                         N                           5            0
4                         N                           3            0
```



```python
data.describe()
```

	id	person_age	person_income	person_emp_length	loan_amnt	loan_int_rate	loan_percent_income	cb_person_cred_
count	58645.000000	58645.000000	5.864500e+04	58645.000000	58645.000000	58645.000000	58645.000000	
mean	29322.000000	27.550857	6.404617e+04	4.701015	9217.556518	10.677874	0.159238	
std	16929.497605	6.033216	3.793111e+04	3.959784	5563.807384	3.034697	0.091692	
min	0.000000	20.000000	4.200000e+03	0.000000	500.000000	5.420000	0.000000	
25%	14661.000000	23.000000	4.200000e+04	2.000000	5000.000000	7.880000	0.090000	
50%	29322.000000	26.000000	5.800000e+04	4.000000	8000.000000	10.750000	0.140000	
75%	43983.000000	30.000000	7.560000e+04	7.000000	12000.000000	12.990000	0.210000	
max	58644.000000	123.000000	1.900000e+06	123.000000	35000.000000	23.220000	0.830000	

The dataset link: Loan Approval Prediction Dataset

References

- Walter Reade, Ashley Chow. (2024). Loan Approval Prediction. Kaggle. Available at: https://kaggle.com/competitions/playground-series-s4e10