Machine Learning Assignment

**LOAN PREDICTION**

**Problem Statement**

Dream Housing Finance company deals in all kinds of home loans. They have presence across all urban, semi urban and rural areas. Customer first applies for home loan and after that company validates the customer eligibility for loan.

Company wants to automate the loan eligibility process (real time) based on customer detail provided while filling online application form. These details are Gender, Marital Status, Education, Number of Dependents, Income, Loan Amount, Credit History and others. To automate this process, they have provided a dataset to identify the customers segments that are eligible for loan amount so that they can specifically target these customers.

**Input Attributes for Dataset**

**1.Loan\_id**: A unique identification number assigned to each loan application. It helps in tracking and managing individual loan applications.

**2.Gender:** The gender of the applicant, whether they are male or female.

**3.Marital Status:** The marital status of the applicant, indicating whether they are married, unmarried, divorced, etc.

**4.Dependents:** The number of dependents (such as children, elderly parents, or other family members) that the applicant financially supports.

**5.Education:** The level of education attained by the applicant, typically categorized as Graduate or Not Graduate.

**6.Self\_Employed:** Indicates whether the applicant is self-employed or not.

**7.Applicant Income:** The income earned by the primary applicant. This includes all sources of income before taxes.

**8.Coapplicant Income:** The income earned by the coapplicant (if any) who is applying for the loan with the primary applicant. Coapplicant income contributes to the total household income.

**9.Loan Amount:** The total amount of money requested by the applicant as a loan.

**10.Loan Amount Term:** The period over which the loan amount is to be repaid, usually expressed in months.

**11.Credit History:** A record of the borrower's repayment of debts and their ability to manage credit. It is represented as a binary variable indicating whether the applicant has a credit history meeting the required guidelines.

**12.Property Area:** The area where the property for which the loan is being applied is located. It could be categorized as Rural, Urban, or Semiurban, for example.

**Output for Dataset**

* The output is given in 0’s(not approved) or 1’s (approved).

**Dataset**

**Input Dataset**: <https://drive.google.com/file/d/1LIvIdqdHDFEGnfzIgEh4L6GFirzsE3US/view>

**3-Different Algorithms**

**1.Logistic Regression:**

**Strengths:** Simple and easy to implement. Outputs can be interpreted as probabilities, making it useful for binary classification problems. Less prone to overfitting when the number of features is small.

**Weaknesses:** Assumes a linear relationship between the features and the log-odds of the outcome. Limited expressiveness for capturing complex relationships in data compared to more flexible models like decision trees or neural networks.

**2.Decision Tree:**

**Strengths:** Can capture complex non-linear relationships in the data. No assumptions about the distribution of data. Easy to interpret and visualize.

**Weaknesses:** Prone to overfitting, especially with deep trees. Can be unstable, small variations in the data can result in a completely different tree. Tends to perform poorly on imbalanced datasets.

**3.Artificial Neural Network (Multilayer Perceptron with Error Back Propagation):**

**Strengths:** Can capture complex patterns and relationships in the data. Highly flexible and can learn intricate decision boundaries. Can handle large amounts of data effectively.

**Weaknesses:** Requires a large amount of data for training to avoid overfitting. Computationally intensive, especially with large networks and datasets. Prone to getting stuck in local minima during training.

**Why we chose Logistic Regression?**

**1.Interpretability:** Logistic regression provides easily interpretable results, where the coefficients associated with each feature indicate the strength and direction of their influence on the loan approval decision.

**2.Computational Efficiency:** Logistic regression is computationally efficient compared to neural networks, especially for smaller datasets. Given the relatively small size of many loan application datasets, logistic regression can provide comparable performance without the computational overhead associated with training and tuning more complex models like neural networks.

**3.Robustness to Noise:** Logistic regression tends to be more robust to noise and outliers compared to decision trees and neural networks. In loan prediction tasks, where the data may contain noise or outliers due to various factors, such as incomplete information or errors in reporting, the robustness of logistic regression can lead to more stable and reliable predictions.

**Libraries Used**

import pandas as pd

import numpy as np

import matplotlib.pyplot as plt

import seaborn as sns

**1.sklearn:** Machine learning algorithms and tools for Python.

**2.matplotlib:** Plotting library for Python.

**3.numpy**: Numerical computing library for Python.

**4.pandas:** Data manipulation and analysis library for Python.

**5.scipy:** Scientific computing library for Python.

**Code**

[**https://colab.research.google.com/drive/1T-iLmel0-RGa7uyQXGsGRvxiBIJzxWl9?usp=sharing**](https://colab.research.google.com/drive/1T-iLmel0-RGa7uyQXGsGRvxiBIJzxWl9?usp=sharing)

**Data Preprocessing**:

* **Read Data:** The code reads a CSV file into a pandas DataFrame.
* **Data Transformation:** It transforms the 'Credit\_History' column from numeric to object type, possibly because it represents categorical values.
* **Handling Missing Values:** Missing values are filled for categorical columns with the most frequent value and for numerical columns with the previous value in the same column.
* **Encoding Categorical Variables:** Categorical variables are encoded using label encoding (LabelEncoder), converting categorical labels into numerical format.
* **Separation of Features and Target:** Features and target variable are separated into X (features) and y (target).

**Model Building:**

* **Model Selection:** Logistic Regression is chosen as the classification algorithm.
* **Model Training**: The Logistic Regression model is instantiated and trained on the training data (X\_train and Y\_train).
* **Cross-validation:** Cross-validation is performed using stratified k-fold cross-validation with 10 folds to evaluate the model's performance. Metrics such as precision, recall, F1-score, log loss, and accuracy are computed for each fold and averaged.
* **Model Evaluation:** The model's performance is evaluated on the testing data (X\_test and Y\_test) using metrics like precision, recall, F1-score, log loss, and accuracy.

**Performance Measure:**

The performance of the logistic regression model on the testing data is evaluated using several metrics: precision, recall, F1-score, log loss, and accuracy.

**1.Precision= True Positives/(True Positives+False Positives)**

**2.Recall=True Positives/(True Positives+False Negatives)**

**3.F1= 2 \* ((Precision \*Recall)/(Precision+Recall))**

**4.Log Loss=-1/NΣ [yi In pi + (1-yi) ln(1 – pi)] ,where i=1 to N**

**5.Accuracy=True Positives+True Negatives/Total observations**

where,

1.True Positives (TP): The number of instances correctly predicted as positive.

2.False Positives (FP): The number of instances incorrectly predicted as positive.

3.False Negatives (FN): The number of instances incorrectly predicted as negative.

4.True Negatives (TN): The number of instances correctly predicted as negative.

5.N: Total number of observations or instances.

6.pi: Predicted probability of the positive class for the i-th observation.

7.yi: Actual binary label (0 or 1) for the i-th observation.

Testing Performance:

Precision: 0.85

Recall: 0.4473684210526316

F1-score: 0.5862068965517242

Log Loss: 7.032907978364323

Accuracy: 0.8048780487804879

**Advantages of Logistic Regression:**

**1.Ease of Implementation:** Simple to implement and interpret.

**2.Efficiency:** Fast training and classification, suitable for real-time applications.

**3.Good Performance on Linear Data:** Effective for linearly separable datasets, achieving good accuracy.

**Disadvantages of Logistic Regression:**

**1.Limited Complexity**: Unable to capture complex relationships in data.

**2.Overfitting Risk**: Prone to overfitting with small datasets and many features.

**3.Inability to Handle Non-linear Data:** Unsuitable for non-linear classification problems due to its linear decision surface.