***Overview of the Analysis***

This analysis focuses on **predicting loan risk** using **machine learning models**, specifically **Logistic Regression**. The purpose of this analysis is to develop a model that classifies loans as either **healthy (0)** or **high-risk (1)** based on financial data.

**Data Overview**

The dataset includes key financial factors that influence loan risk:

• **Loan Size (loan\_size)** – Total amount borrowed.

• **Interest Rate (interest\_rate)** – Percentage of the loan charged as interest.

• **Borrower’s Income (borrower\_income)** – Applicant’s annual income.

• **Debt-to-Income Ratio (debt\_to\_income)** – Percentage of income allocated to debt payments.

• **Number of Accounts (num\_of\_accounts)** – Total credit accounts held by the borrower.

• **Derogatory Marks (derogatory marks)** – Negative marks on the borrower’s credit history.

• **Total Debt (total\_debt)** – The borrower’s outstanding debt.

• **Loan Status (loan\_status)** – **Target variable** (0 = Healthy Loan, 1 = High-Risk Loan).

The goal is to predict whether a loan falls into the **healthy** or **high-risk** category.

**Target Variable Overview**

Using value\_counts(), we observed the distribution of loan statuses:

• **Healthy Loans (0)**: **75,036 instances**

• **High-Risk Loans (1)**: **2,500 instances**

This indicates an **imbalanced dataset**, with significantly more healthy loans than high-risk loans.

**Stages of the Machine Learning Process**

1. **Data Loading & Exploration** – Import and review dataset structure.

2. **Feature-Label Separation** – Extract features (X) and target variable (y).

3. **Train-Test Splitting** – Divide data into training (80%) and testing (20%) sets.

4. **Model Training** – Train a **Logistic Regression** model.

5. **Prediction & Evaluation** – Assess model performance using accuracy, precision, and recall.

**Methods Used**

• **Logistic Regression (LogisticRegression)** – Used for binary classification.

• **Confusion Matrix (confusion\_matrix)** – Evaluates model performance by analyzing prediction accuracy.

**Results**

**Machine Learning Model 1: Logistic Regression**

• **Accuracy:** **99.2%** – The model correctly classified **99.2%** of test data.

• **Precision (Healthy Loans 0)**: **99.7%** – When predicting a healthy loan, the model is correct **99.7%** of the time.

• **Recall (Healthy Loans 0)**: **99.5%** – The model correctly identifies **99.5%** of actual healthy loans.

**Summary & Recommendation**

**Which Model Performed Best?**

• **Logistic Regression performed the best** based on overall accuracy and recall.

• **Overall Accuracy:** **99.2%**

• **Precision for Healthy Loans:** **99.7%**

• **Recall for Healthy Loans:** **99.5%**

While the model performs well overall, its ability to predict **high-risk loans (1)** may be impacted by the dataset imbalance. **Future improvements** could include **resampling techniques** (e.g., SMOTE for oversampling) or exploring **alternative models** such as **Random Forest** or **Gradient Boosting** for enhanced prediction accuracy.