

# Predicting Credit Scores Using Machine Learning

## Data Science Group Project


- Bhanu Prakash Akepogu
- Tarun Emmanuel Majhi
- Charan Reddy Devapatla

## Clark University

- Advisor: Dr. Salem Othman
- 



# Agenda

- Introduction
  - Objective
  - Dataset Description
  - Methodology
  - Models Used
  - Results and Discussion
  - Conclusion
  - References
- 

# Introduction:

---



Credit scoring is an essential component of the financial industry, used in evaluating whether an individual is able to repay a loan, affecting loan approvals, interest rates, and financial services.



It's because traditional credit-scoring systems are based on static models, which have shortcomings in catching the complex patterns or relationships of financial data.



Machine learning offers a new paradigm to enable more dynamic, accurate, and fair credit scoring.



This project focuses on developing predictive models to enhance the accuracy of credit scoring, helping financial institutions make better decisions, decrease default risks, and increasing access to credit for meritorious people.

# Purpose:

- The project tries to apply machine learning for credit score prediction so that the lenders can manage the risk in a better way.
- Benefits also include providing access to financial services for people with thin credit files and increasing transparency in the credit evaluation process.

## Objectives include

- Enhancing prediction accuracy with borrower attributes.
- Identifying important factors that influence credit score.
- Developing flexible models for diverging borrower profiles. This project demonstrates the potential of machine learning to support better risk management and decision-making for financial institutions.

# Project Objective

1

Create predictive models to classify credit scores (Good, Standard, Bad).

2

Improve risk assessment for financial institutions.

3

Enable better decision-making and promote financial inclusion.

# Dataset Overview

## Dataset Description:

- Source: <https://www.kaggle.com/datasets/parisrohan/credit-score-classification>
- 100,000 records, 15 features.

## Key Features:

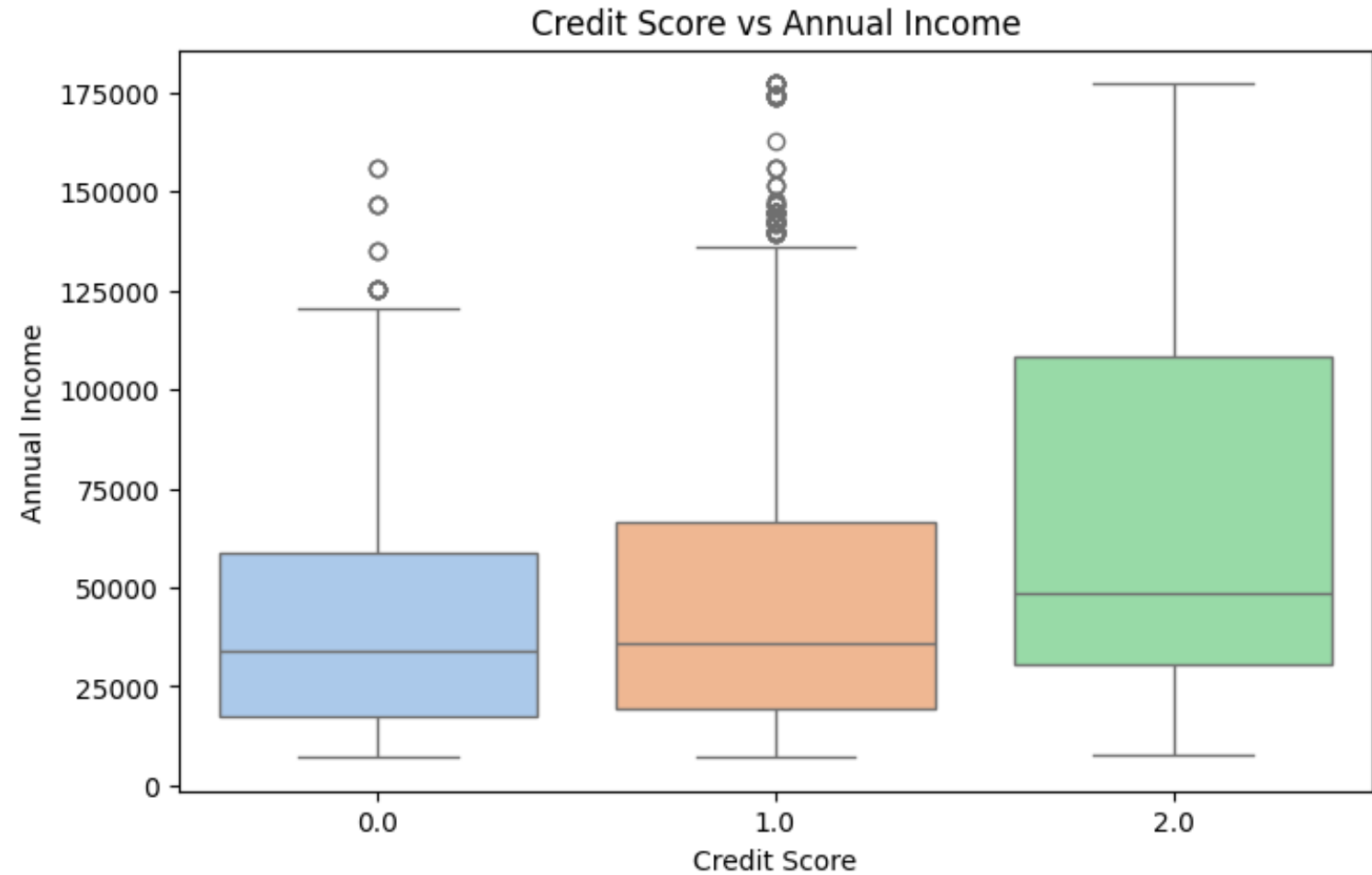
- Age, Income, Credit History, Debt-to-Income Ratio, etc.

## Summary Statistics Table:

- Mean, Median, Min, Max values for critical variables.

Feature	Mean	Std Dev	Min	25%	50%	75%	Max
Age	33.31	10.76	14.00	24.00	33.00	43.00	72.00
Annual Income	50,505.12	38,299.42	7,005.93	19,342.97	38,340.65	74,020.54	810,000.00
Monthly In-Hand Salary	4,198.77	3,187.49	303.64	1,626.76	3,268.23	6,014.88	30,000.00
EMI Per Month	107.04	130.03	0.00	29.20	76.00	139.00	1,200.00
Number of Bank Accounts	5.37	2.59	0.00	3.00	5.00	7.00	20.00
Number of Credit Cards	5.53	2.07	0.00	4.00	5.00	7.00	15.00
Interest Rate	14.53	8.74	1.00	7.00	12.00	20.00	40.00
Number of Loans	3.53	2.45	0.00	2.00	3.00	5.00	15.00
Outstanding Debt	1,426.22	1,155.13	0.23	566.07	1,034.18	1,789.08	10,000.00
Credit UtilizationRatio	32.29	5.12	20.00	28.05	31.00	36.00	60.00
Credit History Age	221.21	99.68	1.00	144.00	210.00	288.00	600.00

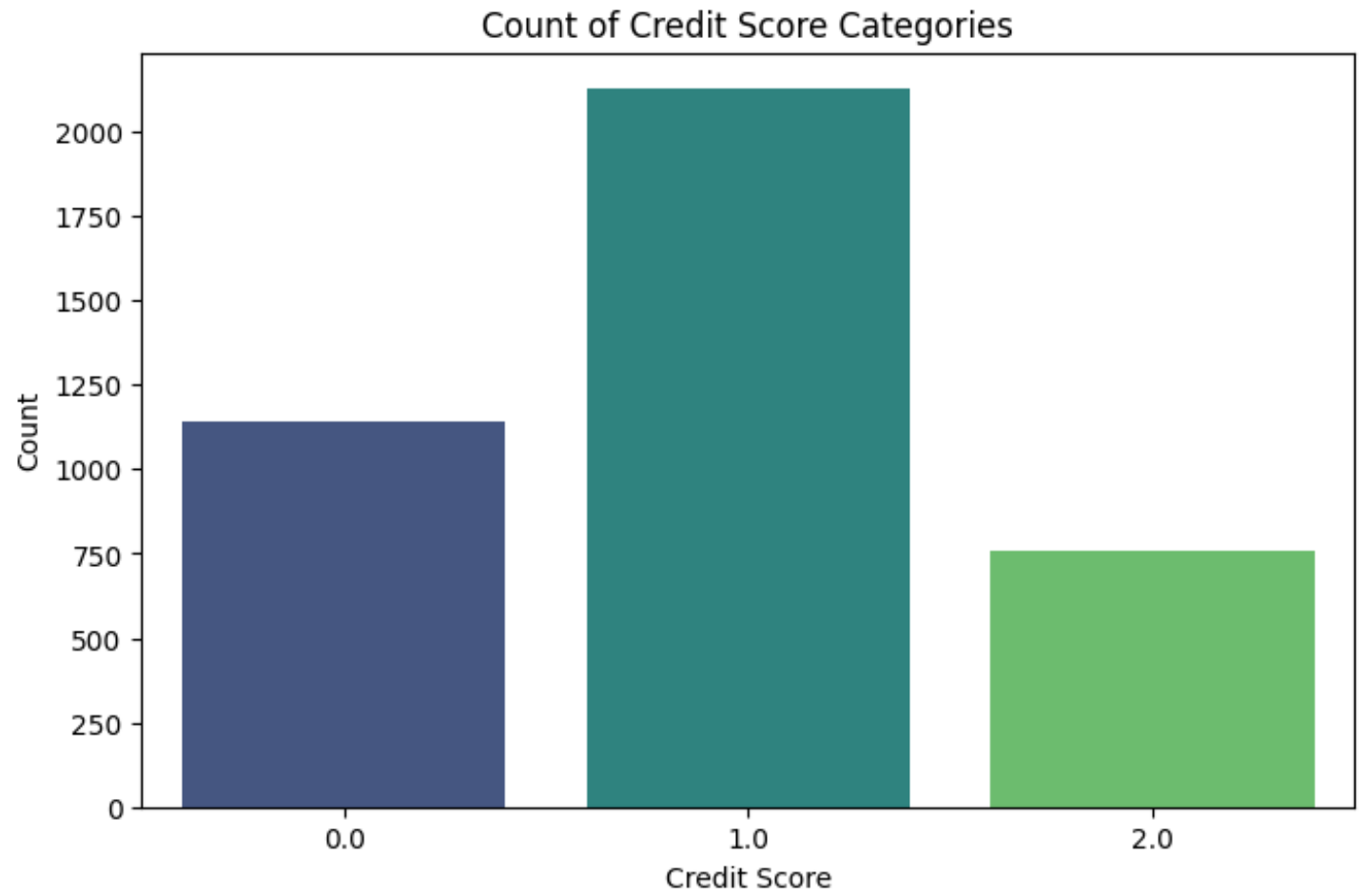
## Exploratory Data Analysis (EDA)



- Credit Score 0.0: Individuals in this category tend to have lower annual incomes, with some outliers at higher income levels.
- Credit Score 1.0: This category shows slightly higher incomes compared to 0.0, with fewer extreme outliers.
- Credit Score 2.0: This group has the highest median and overall income range, indicating that individuals with better creditscores tend to earn more annually.



- 
- Distribution of Target Variable





# Encoding Categorical Variables

- Description:

Categorical features were converted to Numerical using one-hot encoding.

- Reason:

Enables models to process non-numerical data effectively.

- 
- Result:

Avoids misinterpretation of categories as numerical magnitudes.



# Dimensionality Reduction (PCA)

- PCA applied to reduce redundancy.

## Explained Variance Ratios:

- 80% Variance: 11 components, 72.7% Random Forest accuracy.--->
- 90% Variance: 16 components, 75.4% Random Forest accuracy.

PCA reduced redundancy while retaining 90% of variance

PERFORMANCE METRICS WITH 80% VARIANCE PCA DATASET

Model	Accuracy	Precision	Recall	F1-Score
Logistic Regression	0.640133	0.638133	0.640133	0.633754
LDA	0.636100	0.637923	0.636100	0.635295
Decision Tree	0.616767	0.617279	0.616767	0.617014
Random Forest	0.727267	0.726662	0.727267	0.725943
Multilayer Perceptron	0.692533	0.692469	0.692533	0.692103

# 80% vs 90% explained variance:

- The 80% explained variance-managed to achieve an accuracy of 72.7% for Random Forest with 11 features, showing that there is little gain by adding more features.
- The peak of Random Forest accuracy is 75.4% with 16 features in the second table-a variance of 90% explained-showing a significant improvement with the added features until that point.
- In both cases, Logistic Regression, LDA, and Decision Tree performed consistently, with minor variations.
- On the 90% variance , Multilayer Perceptron improves significantly, increasing from 69.3% into 71.4%, showing its sensitivity for more features in the dataset.

PERFORMANCE METRICS WITH 90% VARIANCE

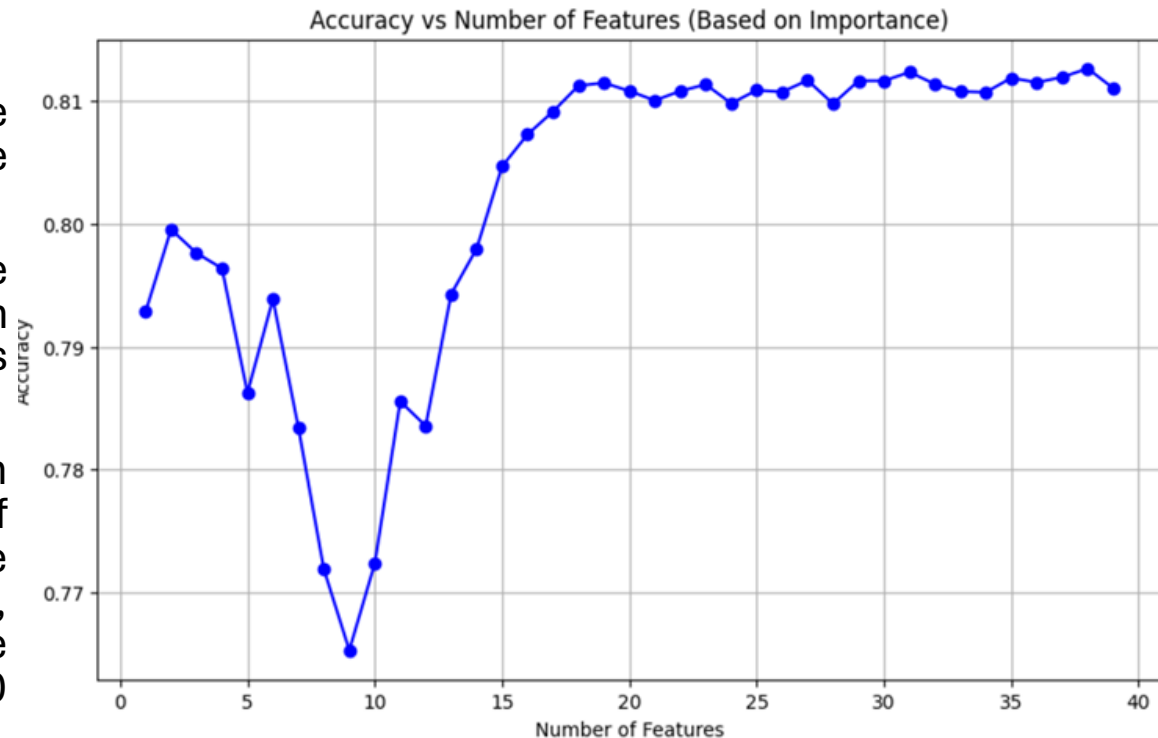
Model	Accuracy	Precision	Recall	F1-Score
Logistic Regression	0.645067	0.644007	0.645067	0.640413
LDA	0.644467	0.647340	0.644467	0.644575
Decision Tree	0.648333	0.648459	0.648333	0.648385
Random Forest	0.754133	0.753856	0.754133	0.753596
Multilayer Perceptron	0.714167	0.713803	0.714167	0.713287

# Feature Importance:

Random Forest inherently provides feature importance based on how much a feature reduces impurity in its decision trees.

It is common to use Random Forest for feature selection since it works well with both categorical and numerical data, and it handles non-linear relationships effectively.

In this diagram accuracy changes when features are added one by one in order of importance. Accuracy fluctuates at the beginning; between 5 and 10 features, accuracy takes a big dip, possibly due to some added features as they are irrelevant. From 10 to 20 features, accuracy improves constantly



# Features and their importance %

	Feature	Importance
12	outstanding_debt	0.101454
6	interest_rate	0.069474
14	credit_history_age	0.062652
8	delay_from_due_date	0.057226
10	changed_credit_limit	0.055896
31	credit_mix_Good	0.052885
16	monthly_balance	0.048123
15	amount_invested_monthly	0.046727
13	credit_utilization_ratio	0.046130
11	num_credit_inquiries	0.042939
3	total_emi_per_month	0.041492
32	credit_mix_Standard	0.040046
2	monthly_inhand_salary	0.039104
1	annual_income	0.039056
5	num_credit_card	0.038293
9	num_of_delayed_payment	0.036981
0	age	0.035954
4	num_bank_accounts	0.029385
7	num_of_loan	0.023847
33	payment_of_min_amount_Yes	0.016813
38	payment_behaviour_Low_spent_Small_value_payments	0.006730
34	payment_behaviour_High_spent_Medium_value_paym...	0.006303
37	payment_behaviour_Low_spent_Medium_value_payments	0.005610
35	payment_behaviour_High_spent_Small_value_payments	0.005227
36	payment_behaviour_Low_spent_Large_value_payments	0.004890
19	occupation_Doctor	0.003526
28	occupation_Scientist	0.003510
23	occupation_Lawyer	0.003500
20	occupation_Engineer	0.003444
26	occupation_Media_Manager	0.003412
21	occupation_Entrepreneur	0.003399
29	occupation_Teacher	0.003337
18	occupation_Developer	0.003315
25	occupation_Mechanic	0.003313
24	occupation_Manager	0.003277
27	occupation_Musician	0.003269
17	occupation_Architect	0.003264
22	occupation_Journalist	0.003124
30	occupation_Writer	0.003073

- Feature selection was applied using RandomForestClassifier to rank the important features. It attributes the largest importance of 0.101 to outstanding\_debt, followed by interest\_rate and credit\_history\_age. Random Forest yielded the best performance of about 81.7%, followed by Decision Tree with 79.7%.

# 14 Features:

- Based on feature importance we have selected 1,2,3 & 10-20 features only they are as follows
- 'outstanding\_debt', 'interest\_rate', 'credit\_history\_age', 'num\_credit\_inquiries', 'total\_emi\_per\_month', 'credit\_mix\_Standard', 'monthly\_inhand\_salary', 'annual\_income', 'num\_credit\_card', 'num\_of\_delayed\_payment', 'age', 'num\_bank\_accounts', 'num\_of\_loan', 'payment\_of\_min\_amount\_Yes'

# Feature selection

As we can observe from above diagram

Dip Between 5 and 10 Features:

- Increasing the number of features to between 5 and 10, you can see a drop in accuracy

Improvement Between 10 and 20 Features:

- From 10 to 20 features, the accuracy increases at a constant phase
- In 90% variance, Random Forest stands out at no.1 among other models

Model	Accuracy	Precision		F1-Score
Logistic Regression	0.637633	0.637307	0.637633	0.635724
LDA	0.639133	0.653649	0.639133	0.641580
Decision Tree	0.797533	0.797576	0.797533	0.797553
Random Forest	0.817933	0.818076	0.817933	0.817965
Multilayer Perceptron	0.699633	0.701599	0.699633	0.698979



# Class balancing:

---



The dataset also required balancing because of class imbalance; this was dealt with by using SMOTE, which oversampled the minority classes to create a balanced training set.



It ensures the models get balanced input data, reducing the bias toward the majority class. It improves the ability of models to learn patterns in minority classes



Hence improve overall classification accuracy. The balancing of the classes significantly reduced misclassification of underrepresented categories.

# Results for Class Balanced data set

- The Random Forest outperformed all others, with an accuracy of 82.2% and an F1-score of 82.3% after balancing; hence, strong in performance on all measures. Decision Tree is competitive with an accuracy of 79.3%, while Logistic Regression and LDA showed slight improvement from their imbalanced classes

Model	Accuracy	Precision	F1-Score
Logistic Regression	0.65	0.7	0.66
LDA	0.66	0.7	0.66
Decision Tree	0.79	0.79	0.79
Random Forest	0.82	0.82	0.82
Multilayer Perceptron	0.67	0.73	0.67

# Random Forest: The Best Model

- Class 1 is the one with the highest false negative rate at 22.7% for the imbalanced classes. The false positive rate for Class 2 was very high, being 21.3%, Class 0 has a false negative rate of 17.2%

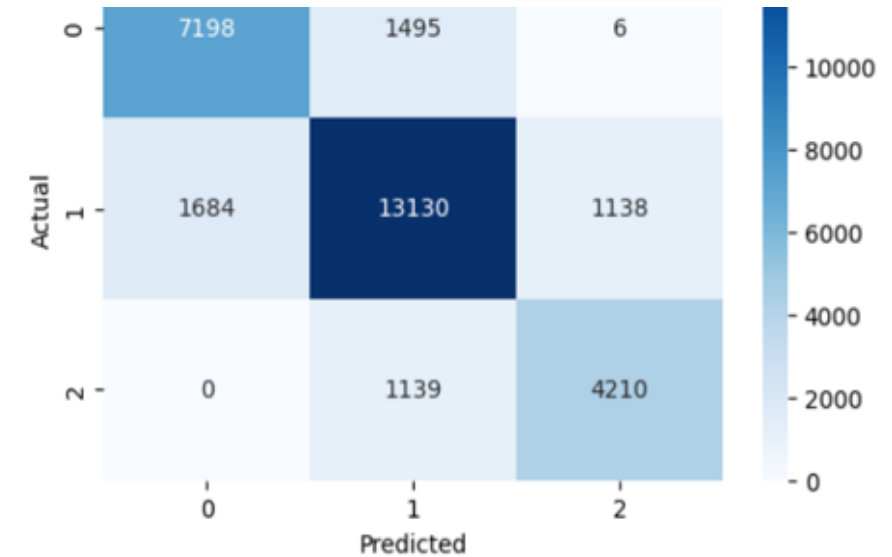
## Performance:

- Before Balancing: Accuracy – 81.7%, F1-Score – 81.79%.
- After Balancing: Accuracy – 82.24%, F1-Score – 82.25%.

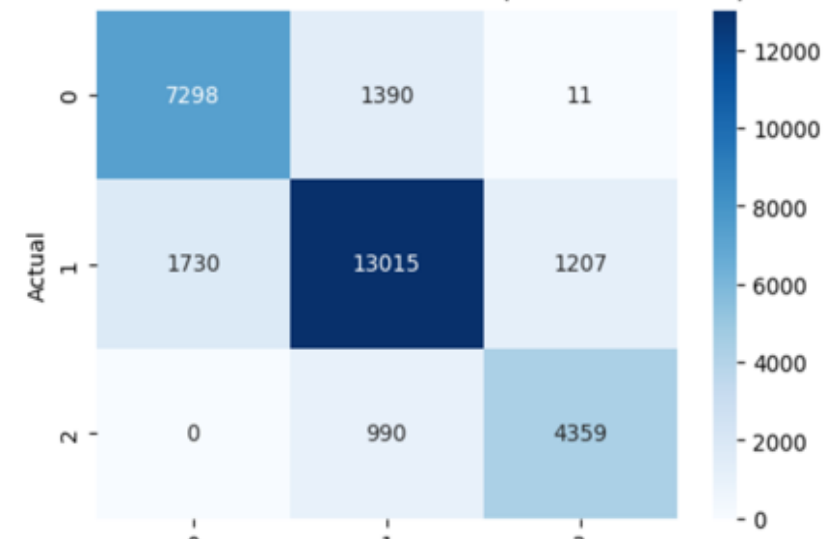
Noted as the best model with strong metrics.

## Strength:

- Best-performing model.
- Handles overfitting better than individual trees.



Confusion Matrix for Random Forest (Balanced Classes)



*A Stacking Classifier is an ensemble learning method that combines predictions from multiple base models (weak learners) to improve overall performance. It uses a meta-model to learn how to best combine the outputs of the base models.*



# Stacking

- Description:
  - Combines Logistic Regression, Decision Tree, Random Forest, and MLP.
  - Here meta-model is a Logistic Regression, which predicts based on the outputs of the base models.
- Performance:
  - Accuracy: 82.4%
  - F1-Score: 82.4%
- Observations:
  - Enhances performance by leveraging strengths of multiple models.

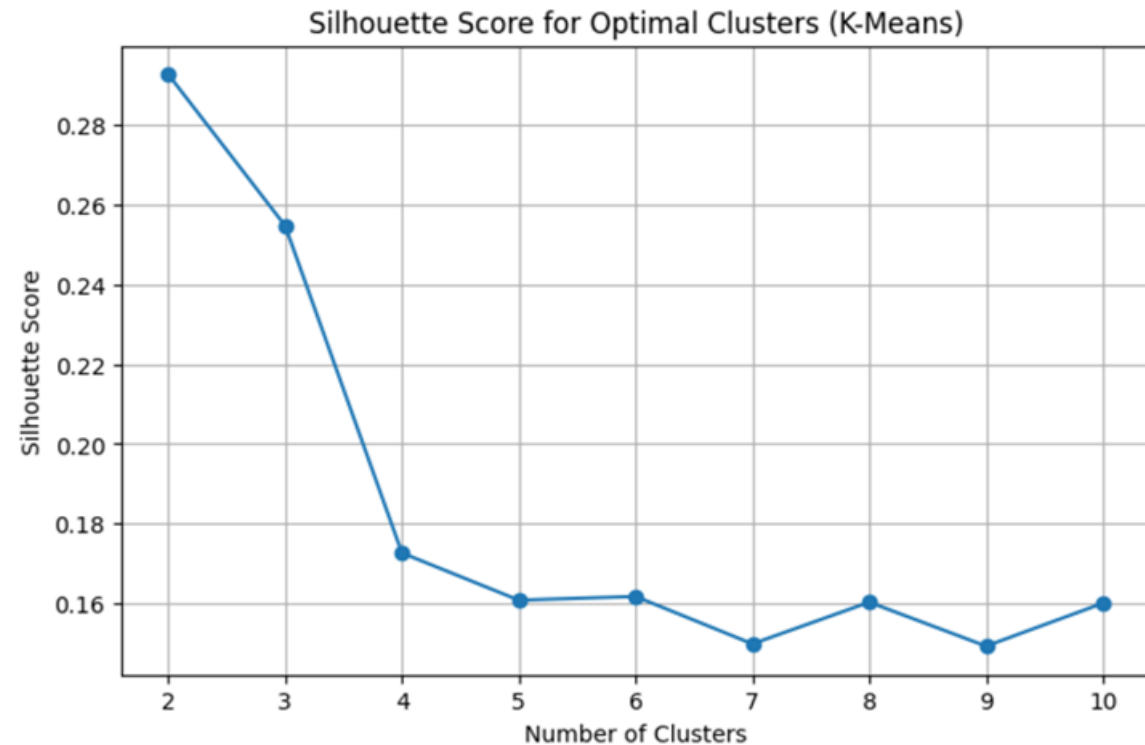
# Clustering Analysis

- The highest score, about 0.28, was achieved with 2 clusters. Beyond 2 clusters, the Silhouette Score drops significantly to about 0.16 for 5 or more clusters

- Findings:

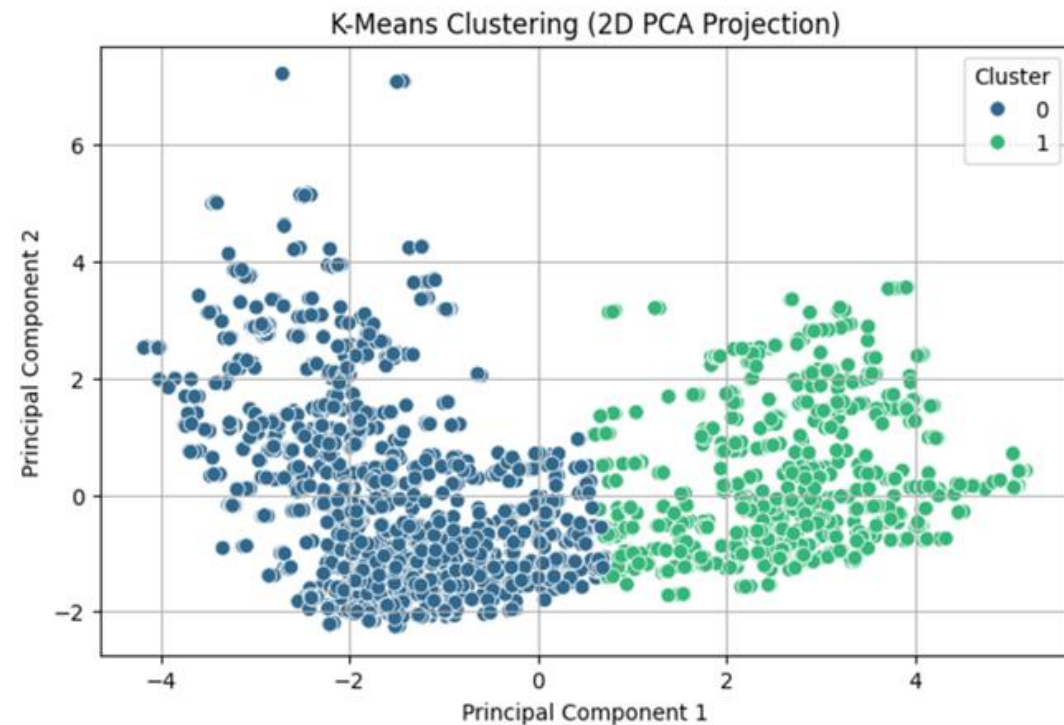
Optimal clusters: 2 (Silhouette Score).

Clusters represent distinct credit patterns.



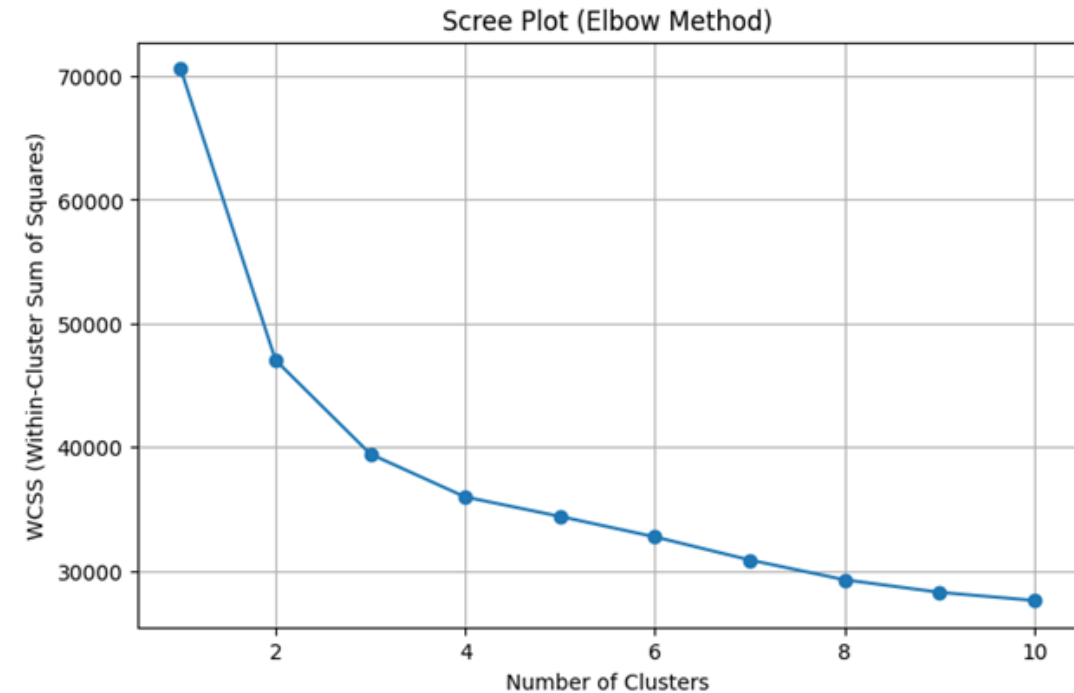
# K-Means Clustering

- Principal Component 1:
- Credit Mix Standard: based on limited credit sources.
- Amount Invested Monthly: financial planning or income levels
- Total EMI Per Month: Reflects the loan
- Principal Component 2:
- Payment of Minimum Amount (Yes):  
Indicates financial behavior
- Number of Bank Accounts: Differentiates individuals
- Number of Credit Cards: Captures credit usage



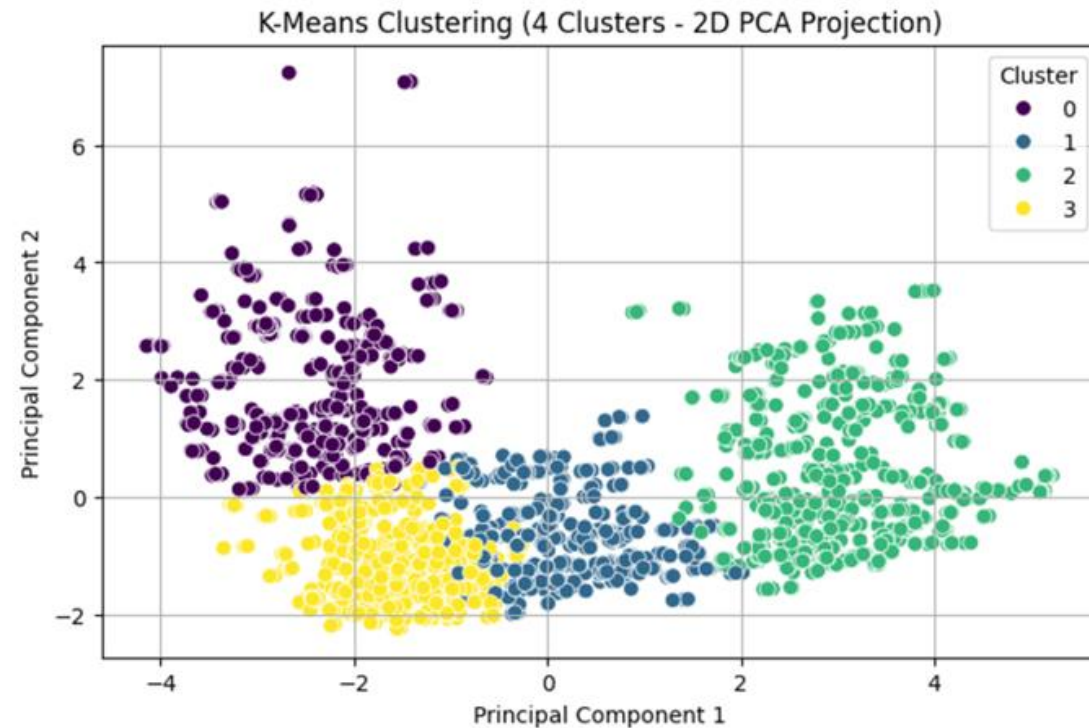
# Scree Plot(elbow method)

- The plot shows that the WCSS drops very steeply up to 1 to 4 clusters. From 4 onwards, the rate of decrease of WCSS really slows down, and the curve is almost flat. This would strongly suggest that the optimal number of clusters is probably 4



# 4 clusters:

- From this PCA plot we observed well-separated clusters that show clear behavioral and financial patterns. While cluster 1 (Blue) characterizes the high financial activities of participants, cluster 2 (Green) and cluster 3 (Yellow) point to some separation like credit dependence versus debt pay-offs.





# Balanced Distribution

- Cluster 0 and Cluster 2 contain the largest number of data points, meaning that these clusters represent the most frequent patterns or groups in the dataset.
- Cluster 1 and Cluster 3 contain fewer data points and so may represent niche or less frequent patterns in the data.



# Conclusion

---

- Analysis of feature importance and dimensionality reduction through PCA significantly improved the models' performance by focusing on the most relevant predictors, such as payment behavior and outstanding debt.
- Random Forest outperformed other models with an accuracy of 82.2% and an F1-score of 82.3%, demonstrating its robustness in credit score prediction and ability to handle class imbalances effectively.
- The use of the SMOTE algorithm reduced class imbalance, improving model generalization and fairness in predictions for underrepresented classes, such as low credit scores.

The integration of machine learning models in credit scoring can transform financial institutions by reducing default risks, promoting fair lending, and improving operational efficiency through automated and precise decision-making processes.



Thank You