Microspecialization Project

Optimizing Credit Approval Processes Using Machine Learning Models

Faculty Guide : Prof. Parama Barai

Compiled and Submitted by:
Bhavya Tewari (23BM63042)

Project Overview

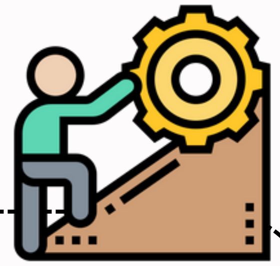
Credit Approval with Machine Learning

Objective: The primary aim of this project is to improve the accuracy of credit risk evaluation, minimizing potential loan defaults and making approval processes more efficient.

Business Benefit: An accurate model reduces the risk of defaults, improves lending decisions, and minimizes losses, enabling institutions to serve creditworthy customers more confidently.

Problem Statement

Current Challenge



Traditional credit approval processes may lack accuracy and efficiency, leading to higher **Type I errors (false positives)**, where applicants with **high risk are approved,** leading to potential loan defaults.

Motivation

The need for precise, reliable models in the financial sector is growing, **especially to ensure sound credit decision-making**. This project focuses on building a **reliable model that helps mitigate risks** associated with approving high-risk applicants.

Data and Methodology

About the Dataset

The dataset consists of **14 Customer Features**

Australian **Credit Approval Dataset**, comprising various categorical and continuous variables related to applicant creditworthiness

Target Variable

Credit Approval Binary Classification:

- 1 => Approved
- 0 => Rejected

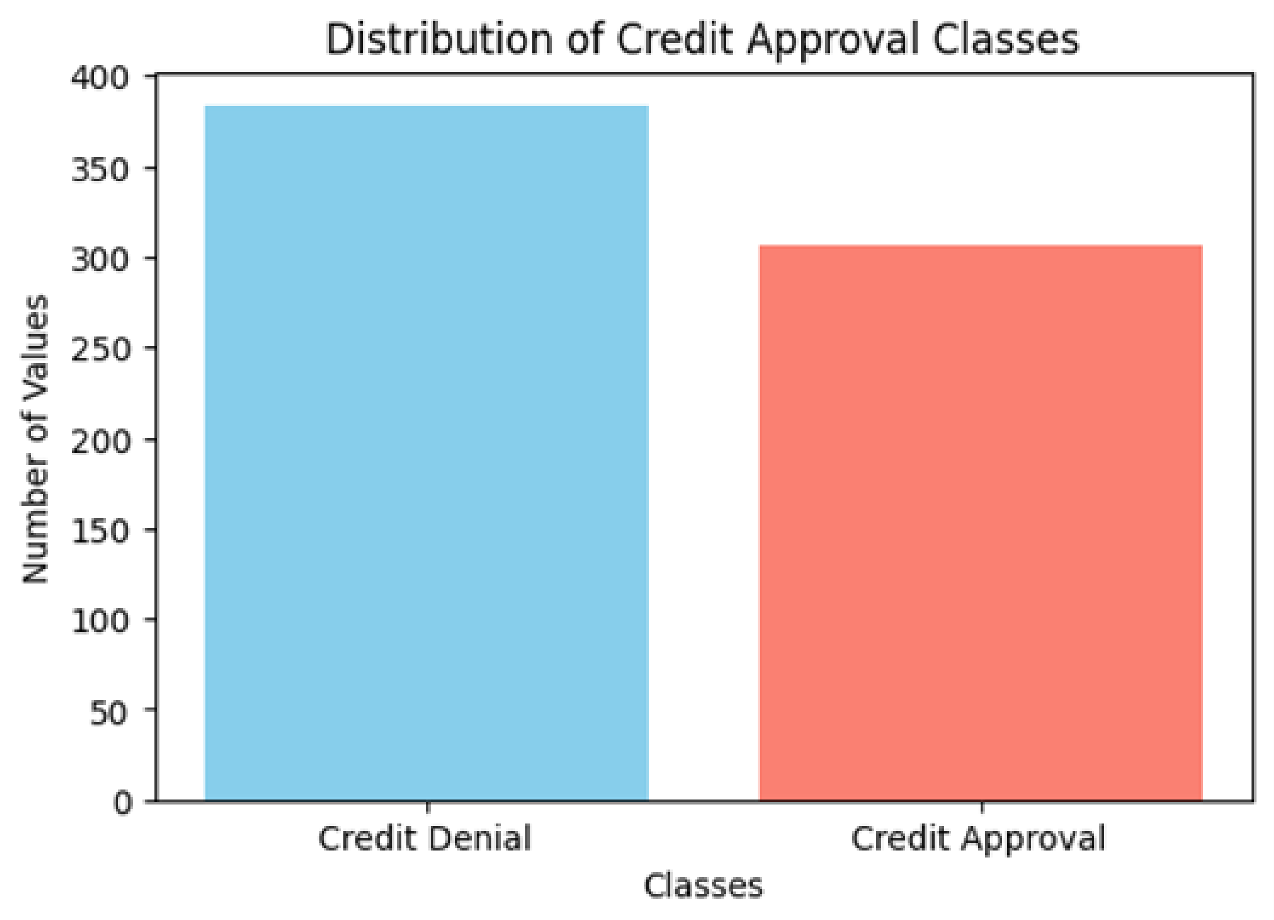
Methodology

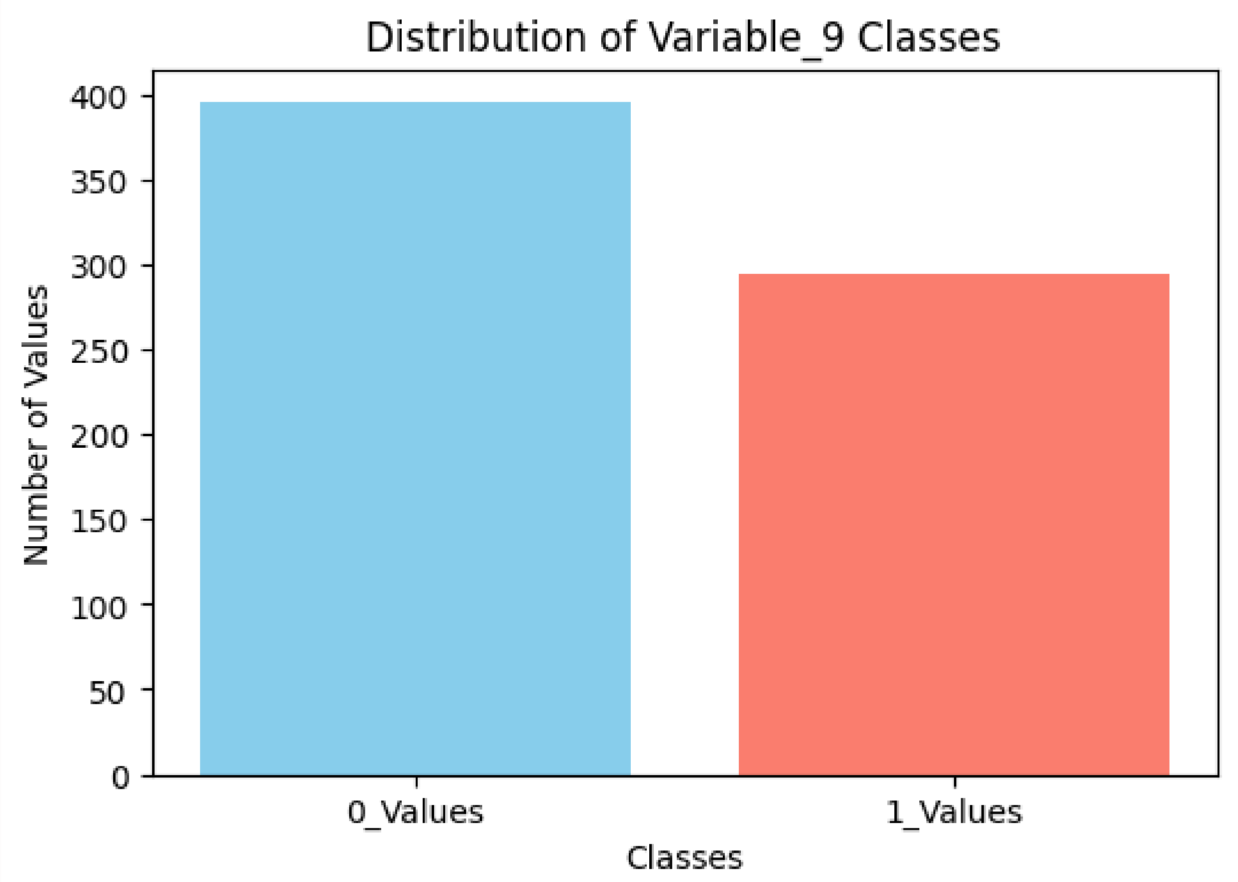
- **Ensemble Models:** Voting and Stacking Classifiers to integrate multiple models for robust results.
- **Deep Learning:** Used a DNN to uncover non-linear patterns in the data.
- **Variable Selection**: Tested subsets to identify the most impactful features.
- **Threshold Tuning:** Adjusted to minimize Type I errors and align with business goals.

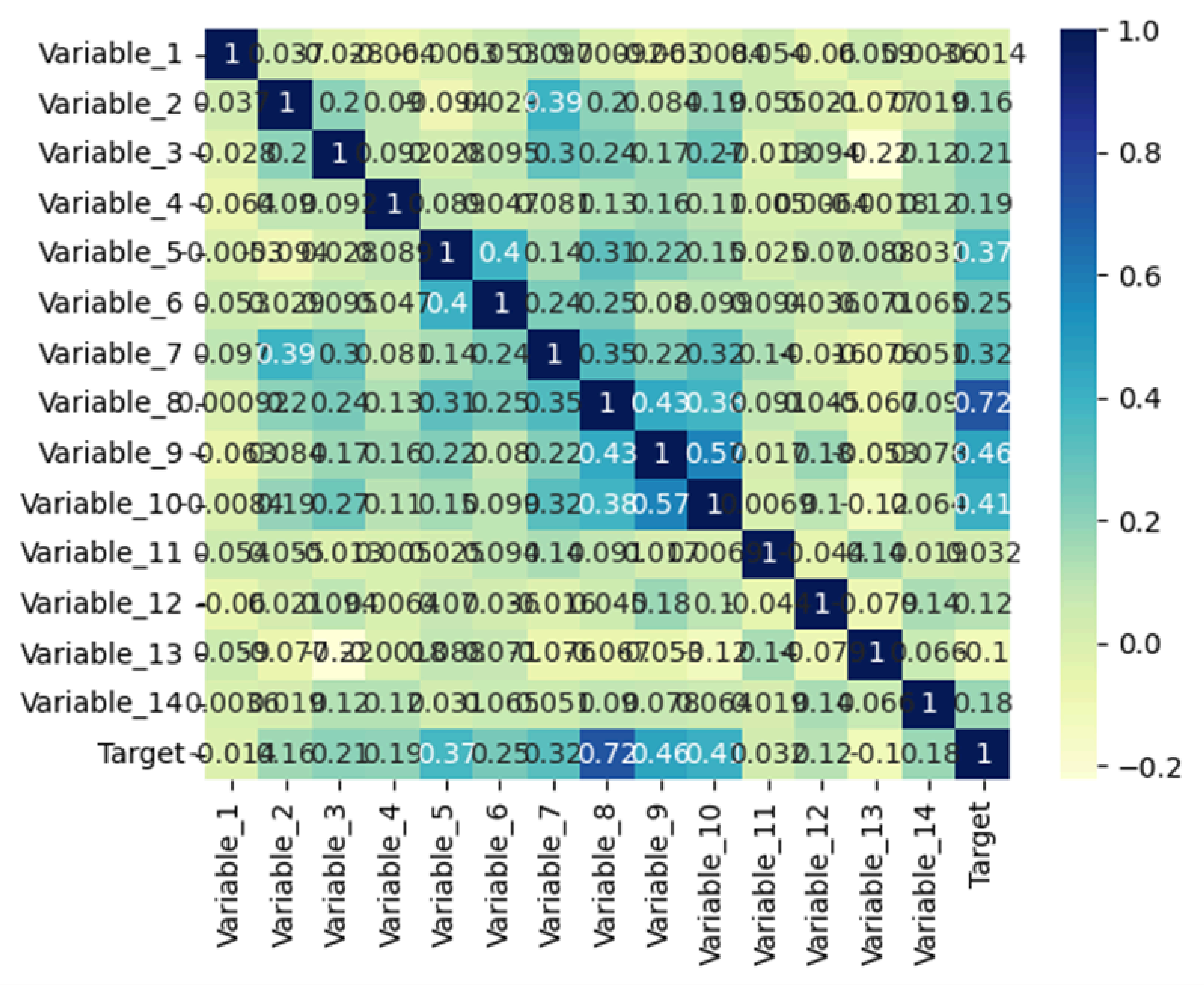
Australian Credit Approval Dataset

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O
1	Variable_1	Variable_2	Variable_3	Variable_4	Variable_5	Variable_6	Variable_7	Variable_8	Variable_9	Variable_10	Variable_11	Variable_12	Variable_13	Variable_14	Target
2	1	22.08	11.46	2	4	4	1.585	0	0	0	1	2	100	1213	0
3	0	22.67	7	2	8	4	0.165	0	0	0	0	2	160	1	0
4	0	29.58	1.75	1	4	4	1.25	0	0	0	1	2	280	1	0
5	0	21.67	11.5	1	5	3	0	1	1	11	1	2	0	1	1
6	1	20.17	8.17	2	6	4	1.96	1	1	14	0	2	60	159	1
7	0	15.83	0.585	2	8	8	1.5	1	1	2	0	2	100	1	1
8	1	17.42	6.5	2	3	4	0.125	0	0	0	0	2	60	101	0
9	0	58.67	4.46	2	11	8	3.04	1	1	6	0	2	43	561	1
10	1	27.83	1	1	2	8	3	0	0	0	0	2	176	538	0
11	0	55.75	7.08	2	4	8	6.75	1	1	3	1	2	100	51	0
12	1	33.5	1.75	2	14	8	4.5	1	1	4	1	2	253	858	1
13	1	41.42	5	2	11	8	5	1	1	6	1	2	470	1	1
14	1	20.67	1.25	1	8	8	1.375	1	1	3	1	2	140	211	0
15	1	34.92	5	2	14	8	7.5	1	1	6	1	2	0	1001	1
16	1	58.58	2.71	2	8	4	2.415	0	0	0	1	2	320	1	0
17	1	48.08	6.04	2	4	4	0.04	0	0	0	0	2	0	2691	1
18	1	29.58	4.5	2	9	4	7.5	1	1	2	1	2	330	1	1
19	0	18.92	9	2	6	4	0.75	1	1	2	0	2	88	592	1
20	1	20	1.25	1	4	4	0.125	0	0	0	0	2	140	5	0
21	0	22.42	5.665	2	11	4	2.585	1	1	7	0	2	129	3258	1
22	0	28.17	0.585	2	6	4	0.04	0	0	0	0	2	260	1005	0
23	0	19.17	0.585	1	6	4	0.585	1	0	0	1	2	160	1	0
24	1	41.17	1.335	2	2	4	0.165	0	0	0	0	2	168	1	0
25	1	41.58	1.75	2	4	4	0.21	1	0	0	0	2	160	1	0
26	1	19.5	9.585	2	6	4	0.79	0	0	0	0	2	80	351	0
27	1	32.75	1.5	2	13	8	5.5	1	1	3	1	2	0	1	1

EDA - Plots and Corrlation Analysis







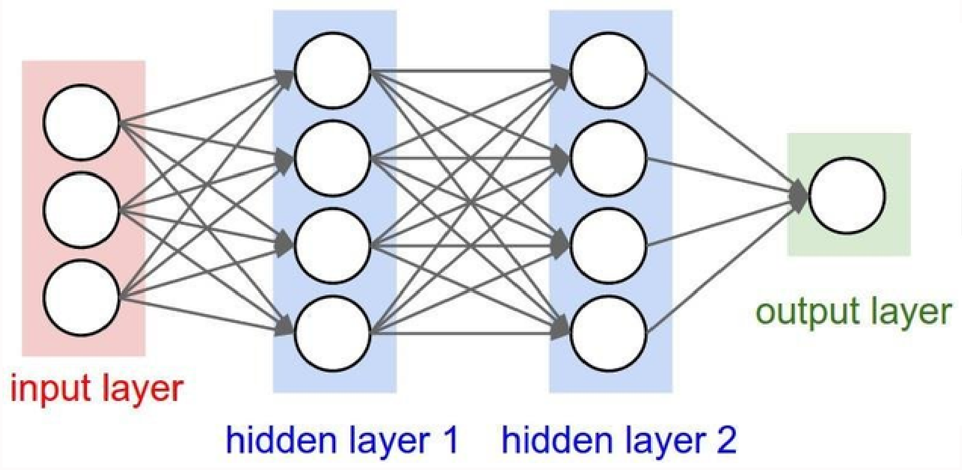
Tools and Technologies Used

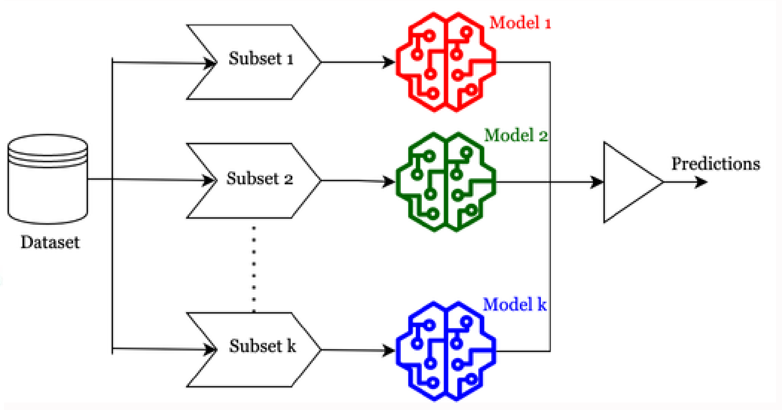
Python Libraries

- **scikit-learn:** Used for model training and ensemble methods.
- **TensorFlow/Keras:** Applied for developing the DNN model.
- **NumPy and Pandas:** For data manipulation and analysis.

Key Algorithms

- **Decision Tree:** To understand feature importance.
- **Random Forest, Logistic Regression, Naive Bayes:** For ensemble models in the Voting and Stacking Classifiers.
- **Deep Neural Network (DNN):** To explore non-linear patterns and complex feature relationships.





Key Findings and Insights

Feature Selection:

- Key variable sets (e.g., 6-10 variables) maintained or improved accuracy.

Threshold Tuning:

- Adjusted for fewer false positives, improving business relevance by reducing risky approvals.

Accuracy:

- Ensemble models outperformed individual classifiers, peaking at 89.85% accuracy with the optimal variable set.

Model	Variables	Accuracy (%)
Decision Tree	14	82.61
Voting Classifier	10	89.37
Voting Classifier	9	89.85
Stacking Classifier	6	89.37
Deep Learning (DNN)	5	86.23

Challenges Faced

- **Data Handling:**
  - Managed missing values and ensured consistent pre-processing.
- **Feature Engineering:**
  - Carefully selected and transformed variables for relevance.
- **Computational Complexity:**
  - Required significant computation for optimal variable and model selection.
- **Threshold Tuning:**
  - Balanced statistical precision with business relevance for Type I error reduction.

Results and Conclusions

- **Accuracy:** Ensemble models, especially the Voting Classifier, reached a peak accuracy of 89.85%.

- **Business Impact:** Reduces loan defaults by minimizing approvals for high-risk applicants, strengthening overall financial stability.

- **Conclusion:** The optimized models achieved high accuracy, demonstrating the value of ensemble and DNN approaches for precise credit risk prediction.

Future Scope

- **Model Enhancements:**
  - Explore advanced models like Gradient Boosting Machines for further accuracy.
- **Data Enrichment:**
  - Incorporate additional features (e.g., spending patterns) to increase model depth.
- **Deployment:**
  - Test model effectiveness in real-world lending scenarios to measure performance.

THANK YOU!