LOAN DEFAULT PREDICTION

GROUP No. - 8

AMAN SARKAR - 27
ANJALI GURJAR - 36
ASHOK GANGWAR - 57
ARNAV GUPTA - 48
ADITYA TYAGI - 18

PROJECT OVERVIEW

1

Project Focus:

The core objective of our project is to predict whether a loan applicant will default or not. This is a binary classification.

2

Why Machine Learning?

Manual credit assessment is:

- *Time-consuming*
- *Subject to human bias*
- *Prone to errors*

Using machine learning (ML), we aim to automate the process and make it more reliable and data-driven.

Problem Statement

Financial institutions face significant risks due to loan defaults, which can lead to heavy losses and affect credit systems. Traditional methods of evaluating a borrower's creditworthiness are often manual, time-consuming, and susceptible to human bias. There is a need for an automated and reliable system to predict whether a loan applicant is likely to default based on historical data.

- *Many financial institutions face high risk due to loan defaults*
- *Manual evaluation of risk is inefficient and error-prone*
- *Need a robust ML model to predict loan defaults accurately*
- *Aim: Build a classification model to flag high-risk applicants*

OUR APPROACH

1. Data Collection

Use a loan dataset containing applicant details
Ensure the dataset includes a target column:
Default (1 = Defaulted, 0 = Not Defaulted).

2. Data Preprocessing

- *Handle missing values (e.g., using mean, median, or dropping rows/columns).*
- *Encode categorical variables (e.g., One-Hot Encoding or Label Encoding).*
- *Normalize/scale numerical features (especially for models like SVM).*
- *Check class imbalance (e.g., if too many '0's vs. '1's – use techniques like SMOTE).*

3. Exploratory Data Analysis:

- *Visualize distributions, correlations, and outliers.*
- *Use bar plots, box plots, and heatmaps to understand relationships.*

4. Model Selection:

Random Forest Classifier:

- *Ensemble of decision trees*
- *Better accuracy and generalization*
- *Provides feature importance*

Decision Tree Classifier:

- *Good interpretability*
- *Handles non-linear relationships*

DATASET

- *Features:*
  - *Age, Income, Credit Score, Employment Type, Loan Amount*
  - *Marital Status, Previous Defaults, Debt-to-Income Ratio, etc.*
- *Target:*
  - *Default (1) or Not Default (0)*
- *Preprocessing:*
  - *Handling missing values, encoding, normalization*

	A	B	C	D	E	F	G	H	I	J	K	L
1	Loan_ID	Gender	Married	Dependen	Education	Self_Empl	ApplicantI	Coapplicar	LoanAmou	Loan_Amc	Credit_His	Property_
2	LP001015	Male	Yes	0	Graduate	No	5720	0	110	360	1	Urban
3	LP001022	Male	Yes	1	Graduate	No	3076	1500	126	360	1	Urban
4	LP001031	Male	Yes	2	Graduate	No	5000	1800	208	360	1	Urban
5	LP001035	Male	Yes	2	Graduate	No	2340	2546	100	360		Urban
6	LP001051	Male	No	0	Not Gradu	No	3276	0	78	360	1	Urban
7	LP001054	Male	Yes	0	Not Gradu	Yes	2165	3422	152	360	1	Urban
8	LP001055	Female	No	1	Not Gradu	No	2226	0	59	360	1	Semiurbar
9	LP001056	Male	Yes	2	Not Gradu	No	3881	0	147	360	0	Rural
10	LP001059	Male	Yes	2	Graduate		13633	0	280	240	1	Urban
11	LP001067	Male	No	0	Not Gradu	No	2400	2400	123	360	1	Semiurbar
12	LP001078	Male	No	0	Not Gradu	No	3091	0	90	360	1	Urban
13	LP001082	Male	Yes	1	Graduate		2185	1516	162	360	1	Semiurbar
14	LP001083	Male	No	3+	Graduate	No	4166	0	40	180		Urban
15	LP001094	Male	Yes	2	Graduate		12173	0	166	360	0	Semiurbar
16	LP001096	Female	No	0	Graduate	No	4666	0	124	360	1	Semiurbar
17	LP001099	Male	No	1	Graduate	No	5667	0	131	360	1	Urban
18	LP001105	Male	Yes	2	Graduate	No	4583	2916	200	360	1	Urban
19	LP001107	Male	Yes	3+	Graduate	No	3786	333	126	360	1	Semiurbar
20	LP001108	Male	Yes	0	Graduate	No	9226	7916	300	360	1	Urban
21	LP001115	Male	No	0	Graduate	No	1300	3470	100	180	1	Semiurbar
22	LP001121	Male	Yes	1	Not Gradu	No	1888	1620	48	360	1	Urban
23	LP001124	Female	No	3+	Not Gradu	No	2083	0	28	180	1	Urban
24	LP001128		No	0	Graduate	No	3909	0	101	360	1	Urban
25	LP001135	Female	No	0	Not Gradu	No	3765	0	125	360	1	Urban
26	LP001149	Male	Yes	0	Graduate	No	5400	4380	290	360	1	Urban
27	LP001153	Male	No	0	Graduate	No	0	24000	148	360	0	Rural

MODEL USED :-

Decision Tree

A Decision Tree is a supervised machine learning algorithm used for both classification and regression tasks. It splits the data into subsets based on feature values and creates a tree-like structure to make predictions.

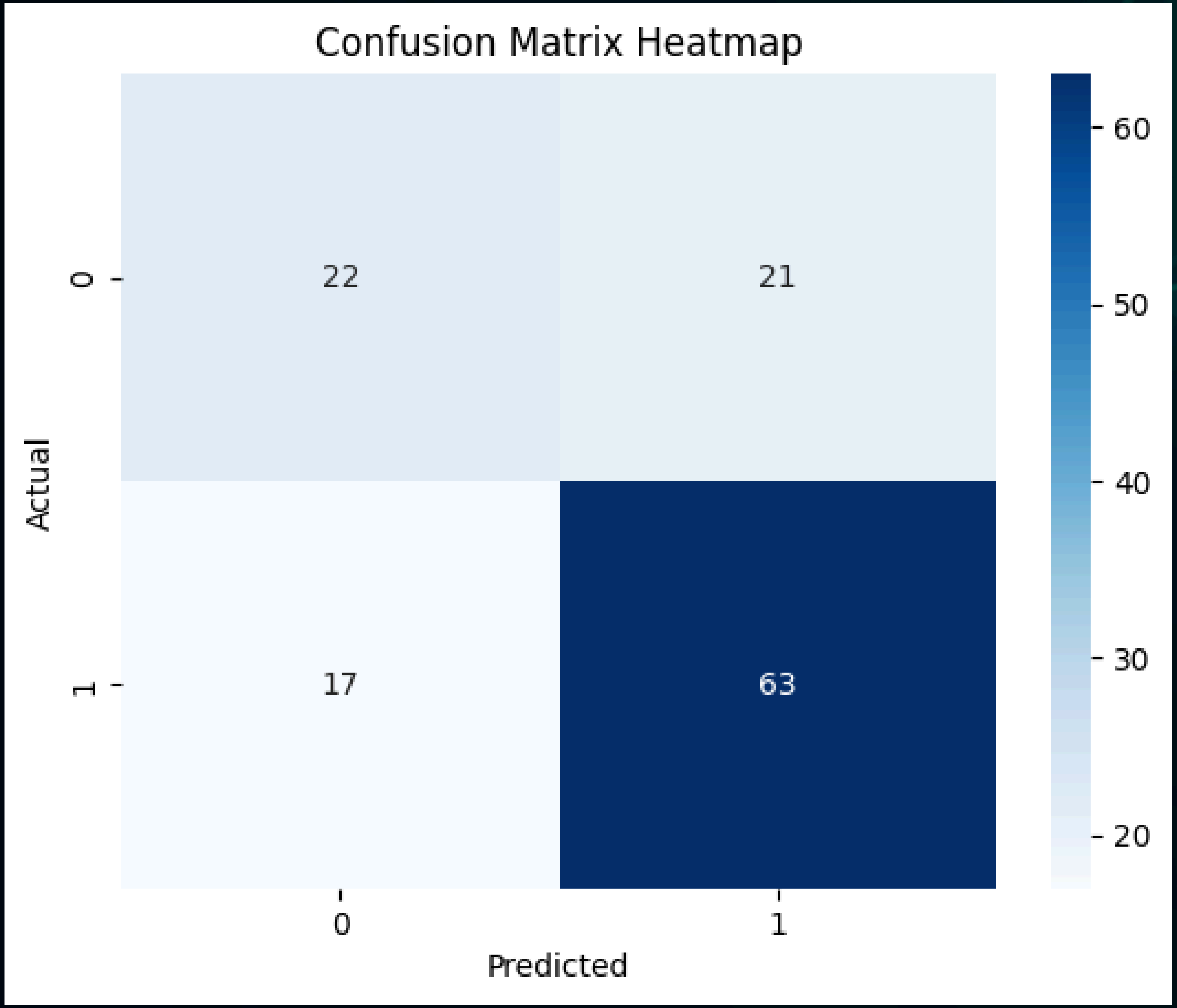
Key Terminologies:

Term	Description
Root Node	The top-most decision node (entire dataset)
Internal Nodes	Feature-based decision points
Leaf Nodes	Final output classes (e.g., Default / Not Default)
Split	Division of data based on feature value
Gini Index / Entropy	Metrics to decide the best split (lower = purer split)

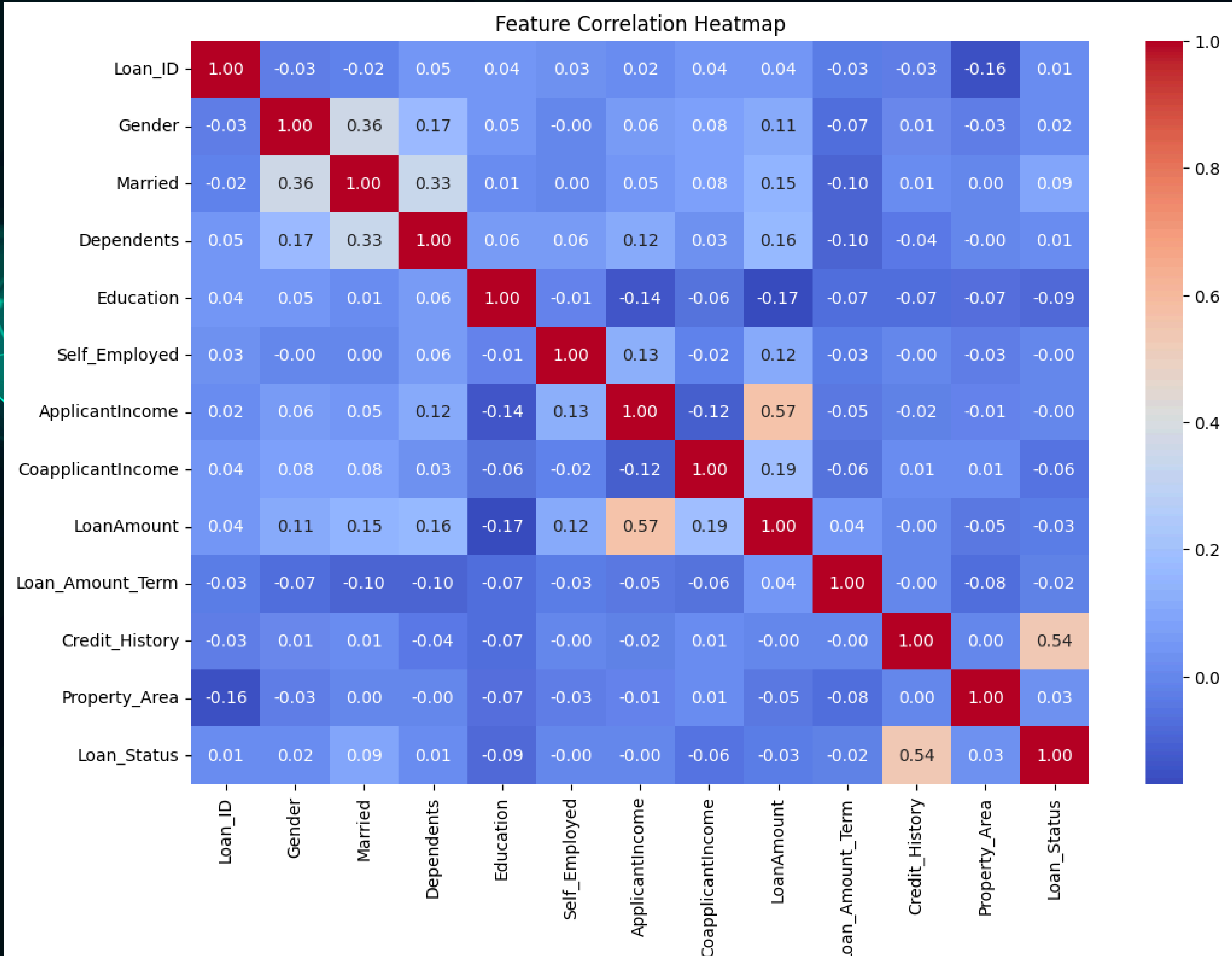
CONFUSION MATRIX

A Confusion Matrix is a table used to evaluate the performance of a classification model. It compares the actual (true) values with the predicted values to show how well the model is performing.

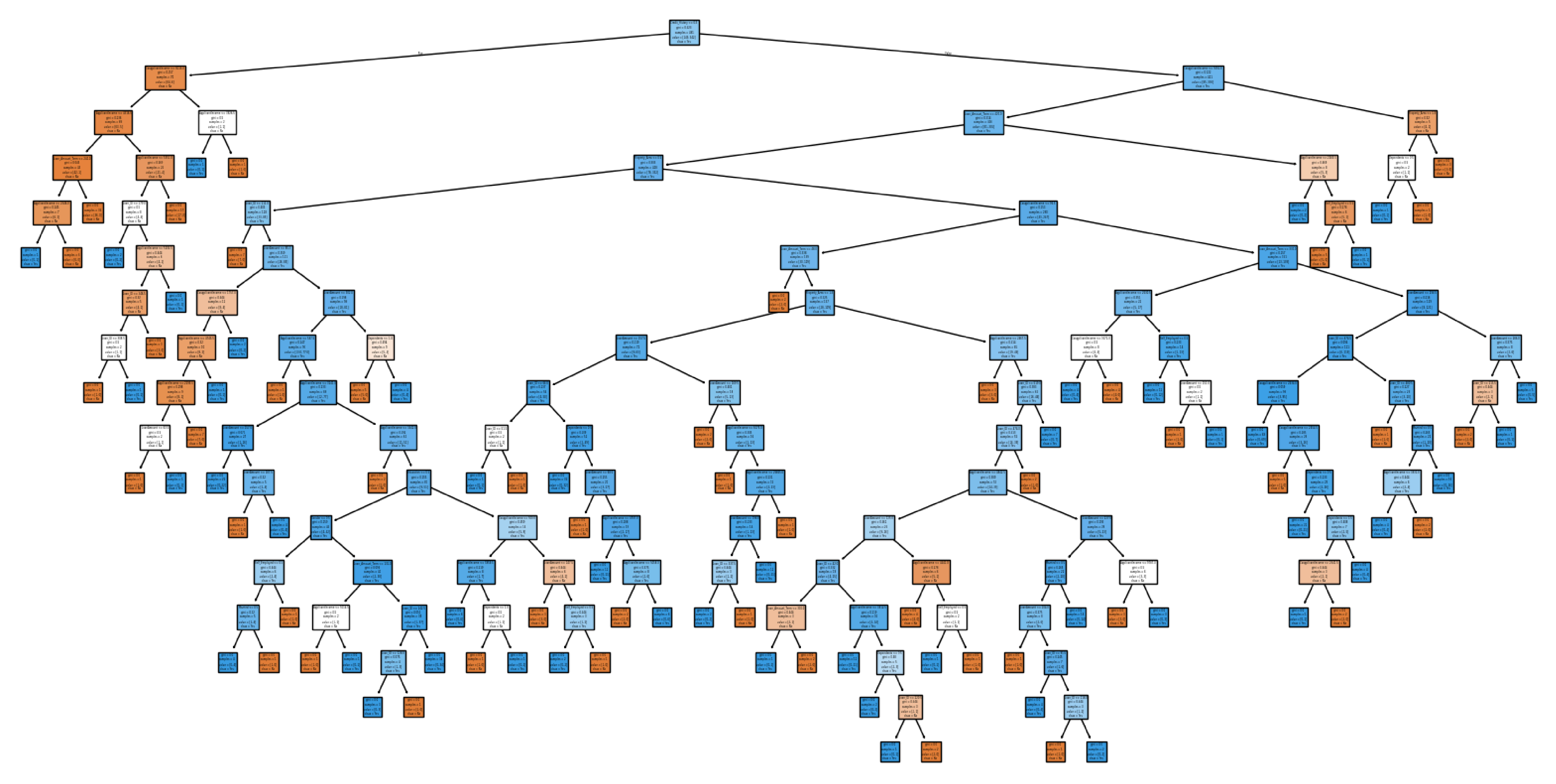
	Predicted: No Default (0)	Predicted: Default (1)
Actual: No Default (0)	True Negative (TN)	False Positive (FP)
Actual: Default (1)	False Negative (FN)	True Positive (TP)



CORELATION HEATMAP MATRIX



DECISION TREE output for dataset:--



THANK YOU!

FOR YOUR ATTENTION