#SYSTEM ARCHITECTURE

Data Ingestion LAYER 1

DATA ACCUMULATION

- Kaggle public datasets
- Transaction & credit behavior data
- MSME loan data

DATA Loading

- Read csv files .
- Input From users.
- Give Me Some Credit (individual borrowers)
- German Credit Dataset

DATA Cleaning

- Handle Missing Values
- Remove Duplicates
- Fix Invalid & Impossible Values
- Outlier Treatment

Feature Engineering LAYER 2

Feature Discovery

- **Raw Data itself in not enough .**
- **You derive new features that describe risk behavior.**
- **for Eg-Credit Utilization Ratio, Payment Stability.**

Feature Validation

- Check missing value percentage
- Remove constant / low-variance features
- Validate logical bounds and skewed features. e.g. Utilization $\in [0,1]$
- EMI $\leqslant$ Income

Feature Selection & Bias Screening

- Select predictive, stable, and fair features

Machine Training LAYER 3

Foundation

SMOTE

Train 80%

Test 20%

Models

Logistic regression = baseline model

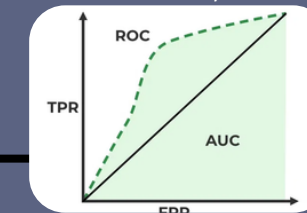**Random Forest** = "Which features actually drive risk?"

KNN = Sanity Check Model

XGBoost = High accuracy + controllable overfitting

scikit learn

Model evaluation LAYER 4

ROC-AUC

- Ability to distinguish from odds, >0.79.

SHAP

- Why did model give this prediction.

Shap

Explainability

- F1 Score
- Accuracy
- Confusion matrix
- Recall
- Precision
-

Predicted Values

	Positive	Negative
Positive	TP	FN
Negative	FP	TN

Actual Values

Feedback LAYER 5

Dashbaord

"We collect loan outcomes, repayment behavior, financial changes, and user feedback to continuously improve model accuracy and fairness."

Testing and Feedback

Taking user input and testing rest 20%

Future scope

- Addng features of AI recomendations.
- API calling.
- Govt.implementation