# Roadmap for Loan Approval Prediction using ML

# Introduction

(  Satyendra Prakash B.Tech IIT Roorkee )

LOANS are the major requirement of the modern world. By this only, Banks get a major part of the total profit. It is beneficial for students to manage their education and living expenses, and for people to buy any kind of luxury like houses, cars, etc. But when it comes to deciding whether the applicant's profile is relevant to be granted with loan or not. Banks have to look after many aspects.

## Problem Statement

We are going to develop one such model that can predict whether a person will get his/her loan approved or not by using some of the background information of the applicant like the applicant's gender, marital status, income, etc.

# RoadMap

- **Data Collection :**

    Gather relevant data for model training. This may include historical loan data, customer information, credit scores, employment history, and other relevant features. Ensure that the data is representative and diverse.

Download the dataset from here: https://www.kaggle.com/datasets/altruistdelhite04/loan-prediction-problem-dataset

I have used the above dataset for my model.

This dataset contains information about loan applications, including various attributes related to applicants and whether their loan applications were approved or denied.

The Dataset Contains 13 features.Let's discuss how each feature in the dataset could potentially impact the target feature, which is "Loan_Status"  (indicating whether the loan was approved or not).

## 1. Loan_ID:

A unique identifier for each loan application. It doesn't contribute to the decision-making process but can be useful for record-keeping.

## 2. Gender:

Lending institutions might consider gender as a factor in loan approval, depending on historical data or institutional policies. For instance, if there's evidence of gender-based discrimination, it could affect loan approval.

## 3. Married:

Married individuals may be perceived as more financially stable and responsible.

Lenders might be more inclined to approve loans for married applicants.

## 4. Dependents:

The number of dependents could influence loan approval, as more dependents might mean higher financial responsibilities. Lenders may assess the applicant's ability to repay the loan considering their family size.

## 5. Education:

The level of education might be a proxy for the applicant's earning potential and financial stability. Graduates may be perceived as having better job prospects and, consequently, higher repayment capabilities.

## 6. Self_Employed:

Self-employed individuals may face different income patterns compared to salaried individuals. Lenders might scrutinize the stability of self-employed applicants' income sources.

## 7. ApplicantIncome: Higher income generally indicates a better ability to repay a loan. However, extremely high or low incomes might be red flags. Lenders may set income thresholds for loan approval.

8. **CoapplicantIncome:** The income of the coapplicant can supplement the household income, affecting the overall repayment capacity. A higher coapplicant income may positively influence loan approval.

9. **LoanAmount:** The amount of the loan applied for is crucial. Lenders will assess whether the requested loan amount aligns with the applicant's income and financial situation.

10. **Loan_Amount_Term:** The term of the loan affects monthly repayment amounts. Shorter terms might indicate a quicker repayment ability, while longer terms might be associated with higher overall interest payments.

11. **Credit_History:** This is likely one of the most critical factors. A good credit history (1.0) is generally associated with a higher likelihood of loan approval. Lenders heavily rely on credit history to assess risk.

12. **Property_Area:**

    The location of the property can influence loan approval. Urban areas might have different risk profiles than rural areas, and lenders may have specific criteria for different regions.

# Data Cleaning

Clean the data to handle missing values, outliers, and inconsistencies. This step is crucial for the model's accuracy and generalization. We may need to impute missing values, standardize or normalize features.

**Exploratory Data Analysis (EDA):** Conduct exploratory data analysis to understand the relationships between different variables, identify patterns, and gain insights. Visualization tools can be helpful in this phase.

**Feature Engineering:** Create new features or modify existing ones to improve the model's performance

**Data Splitting:** Split the dataset into training and testing sets. The training set is used to train the model, and the testing set is used to evaluate its performance.

**Model Selection:** Choose an appropriate machine learning algorithm for your problem. For loan approval prediction I used :-

- logistic regression , Decision trees & Random forests.

**Model Training:** Train the selected model using the training dataset. This involves feeding the algorithm the features and corresponding labels and letting it learn the patterns in the data.

**Model Evaluation:** Evaluate the model's performance on the testing dataset using appropriate metrics such as accuracy, precision, recall, and F1 score.
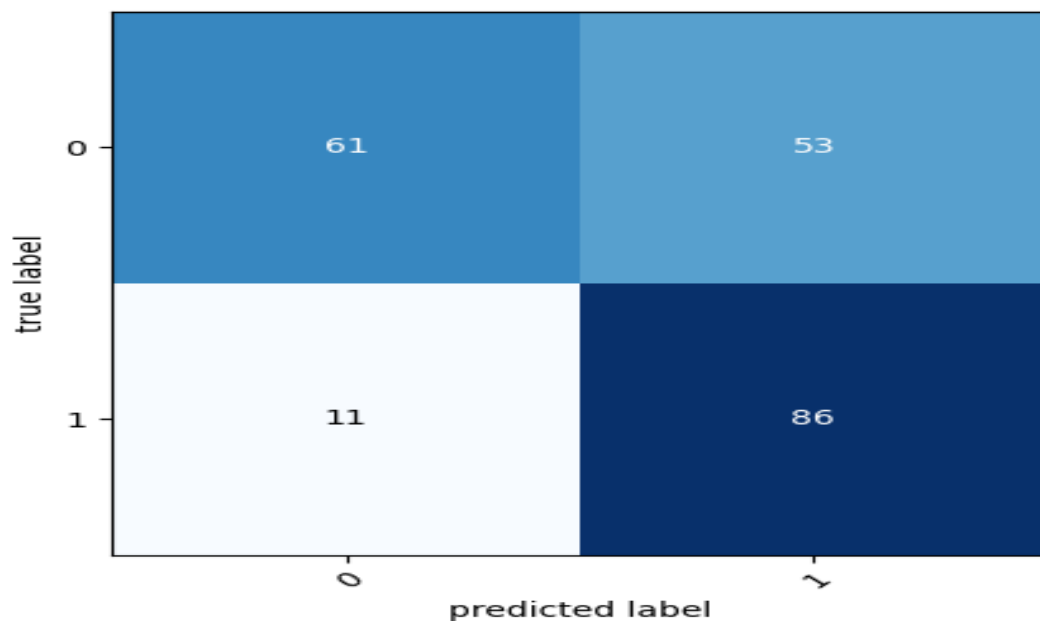
# Results :-

## 1) Logistic Regression Model :-

Accuracy of Logistic Regression model is:  **69.67 %**

Its classification matrix :-

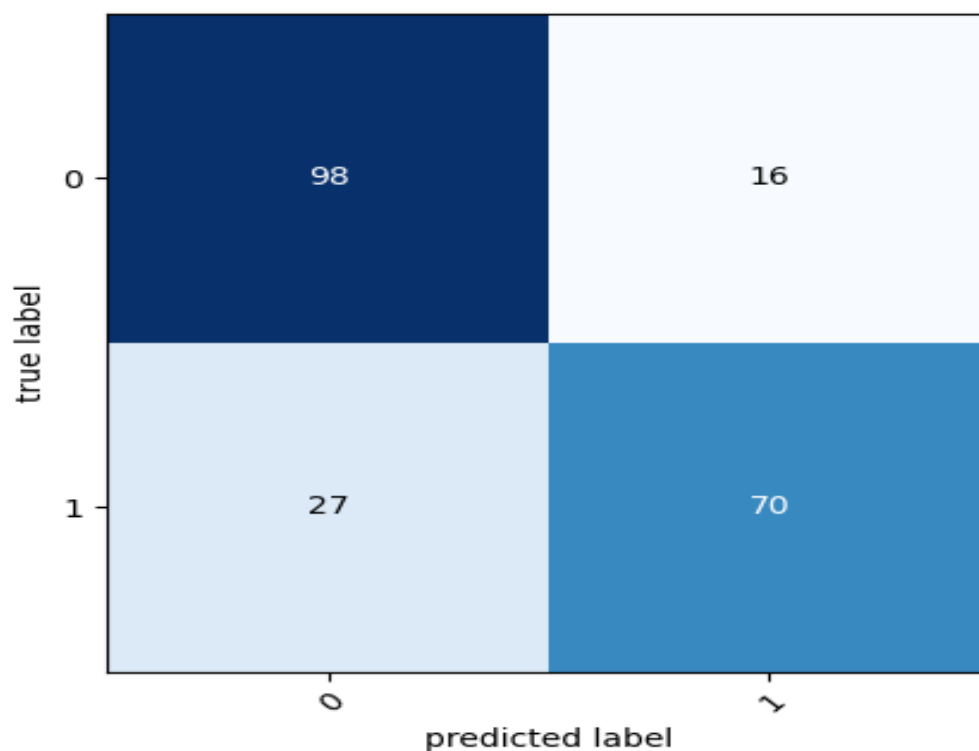|  | precision | recall | f1-score | support |
|---|---|---|---|---|
| 0 | 0.85 | 0.54 | 0.66 | 114 |
| 1 | 0.62 | 0.89 | 0.73 | 97 |
| accuracy |  |  | 0.70 | 211 |
| macro avg | 0.73 | 0.71 | 0.69 | 211 |
| weighted avg | 0.74 | 0.70 | 0.69 | 211 |

Its Confusion matrix :-

## 2) Decision Tree Classifier

Accuracy of Decision Tree Classifier is: **79.62 %**

Its classification matrix :-

|  | precision | recall | f1-score | support |
|---|---|---|---|---|
| 0 | 0.78 | 0.86 | 0.82 | 114 |
| 1 | 0.81 | 0.72 | 0.77 | 97 |
| accuracy |  |  | 0.80 | 211 |
| macro avg | 0.80 | 0.79 | 0.79 | 211 |
| weighted avg | 0.80 | 0.80 | 0.79 | 211 |

Its Confusion matrix :-

## 3) Random Forest Classifier :-

Accuracy of Random Forest Classifier is: **88.15 %**

Its classification matrix :-

|  | precision | recall | f1-score | support |
|---|---|---|---|---|
| 0 | 0.92 | 0.86 | 0.89 | 114 |
| 1 | 0.85 | 0.91 | 0.88 | 97 |
| | | | | |
| accuracy | | | 0.88 | 211 |
| macro avg | 0.88 | 0.88 | 0.88 | 211 |
| weighted avg | 0.88 | 0.88 | 0.88 | 211 |

Its Confusion matrix :-