**LOAN APPROVAL**

In our daily lives, we deal with various factors that tend to impact our money-spending habits, leading to us having to ask for loans from multiple sources, from small money lenders like friends and families to huge firms like banks and financial institutions.

**Project Objective**

This project is intended to help one understand various factors that have influence over their loan status and learn if there are any areas that they can improve on in order to increase their chances of loan approval

**About the Data Set**

This dataset is obtained from Kaggle ([Loan-Approval-Prediction-Dataset (kaggle.com)](https://www.kaggle.com/datasets/architsharma01/loan-approval-prediction-dataset) ). This dataset is a collection of financial records and associated information used to determine the eligibility of individuals or organizations for obtaining loans from a lending institution. I chose this dataset due to its rich features that can be used to find out the factors and accomplish the objective of our project.

**Research Questions**

1. Does credit score have a relationship with the loan approval?
2. Is there a relationship between income level, loan status, and credit score?
3. Do factors like the number of dependents, education level, and self-employment relate to credit score and loan approval?
4. What is the average loan-to-income ratio? Does the loan-to-income ratio influence the loan approval status?
5. Does the loan term and amount (conditional probability) affect the approval status?
6. Is there a relationship between the asset value and a person’s self-employment status, income, and loan amount? Does an individual’s asset value affect their loan status

**PROCEDURES**

Loading the necessary libraries for the time being

Loading the data set

**Data Cleaning**

I started by checking for null values and duplicates. Its outcome showed that no records have null or duplicated

I changed the datatypes of some columns such as loan id, education, self-employed, number of dependents

**Outliers:** I used a function to eliminate the first phase of outliers and for the other phase I removed them directly

I renamed various columns to bring out the ease of understanding

I created some other columns that I would use during my analysis, this is Monthly Income and Annual Income to Loan Ratio. Some other columns are created at later times as the need arises

The summary of the data marked the end of the data cleaning process

Statistical Modeling

Q1.

I used a t-test to analyze the relationship between loan status and credit score. I then used a box plot to explore the relationship between them.

Q2

I started by analyzing the relationship between credit score and income level. I used correlation and ANOVA but the results were still the same, the credit score doesn't have a relationship with income level

Then, while analyzing if someone's income level is related to the loan status, I used a t-test which showed the outcome of rejecting the null hypothesis. Still, the outcome showed no relationship between income level and loan status

Q3

I started by checking if the number of dependents has a pattern and what is the distribution

I then analyzed the relationship between credit score and number of dependents to see if credit score has a relationship with the number of dependents. It doesn't

I analyzed the relationship between the number of dependents and the loan amount, still the relationship wasn't found

The number of dependents has no relationship with anything

The analysis between Education level credit score and loan amount was done using an ANOVA table which showed no relationship.

etc

**Machine Learning Modeling**

Since we are dealing with a Binary prediction problem, I chose to use logistic regression because of its nature