# Predictive Analytics Tool for Investment Behavior

## Introduction

The goal of this project is to build a predictive analytics tool to understand investment behavior based on demographic and financial information. The tool will predict the risk level and return earned by individuals based on their characteristics.

## Data Collection

The dataset used for this project contains information about individuals' demographic details, financial status, investment behavior, and more. It includes the following columns:

- S. No.
- City
- Gender
- Marital Status
- Age
- Education
- Role
- Number of investors in family
- Household Income
- Percentage of Investment
- Source of Awareness about Investment
- Knowledge level about different investment product

- Knowledge level about sharemarket
- Knowledge about Govt. Schemes
- Investment Influencer
- Investment Experience
- Risk Level
- Return Earned
- Reason for Investment

# Visualizations

1. Demographic Distribution Analysis:
   - Bar plots for Gender Distribution, Marital Status Distribution, and Age Distribution.

2. Employment Details Analysis:
   - Pie charts for Distribution of Roles, Distribution of Career Stages, and Distribution of Household Income Brackets.

3. Relationships Between Variables:
   - Scatter plot for Age vs. Household Income.
   - Box plot for Role vs. Household Income.

4. Investment Analysis: Bar plots for Sources of Awareness about Investments, Knowledge Levels about Different Investment Products, Influencers for Investments, Risk Levels, and Reasons for Investment.

5. Average Percentage of Household Income Invested by Age and Gender:

Bar plot showing the average percentage of household income invested by age and gender.

6.Correlation Matrix: Heatmap showing the correlation matrix between different variables.

7. Feature Importance Plot: Bar plot showing the feature importance based on Random Forest Classifier for feature selection.

## Data Preprocessing

1. Handling Missing Values: Missing values in numerical columns are filled with the mean, while categorical columns are filled with the mode.

2. Encoding Categorical Variables: One-hot encoding is used for categorical variables to convert them into numerical format.

3. Scaling Numerical Variables: Numerical variables are scaled using StandardScaler to bring them to the same scale.

## Feature Selection

Random Forest Classifier is used for feature selection to identify the most important features for predicting risk level and return earned.

## Model Selection

Random Forest Classifier is chosen as the predictive model due to its ability to handle both numerical and categorical variables, handle missing values, and provide feature importances.

## Model Evaluation

The model is evaluated using the following metrics:

- Overall Accuracy
- Classification Report for Risk Level
- Classification Report for Return Earned

## User Interface

A user interface is created using Tkinter to allow users to input their demographic and financial details. The model then predicts their risk level and return earned based on the input.

## Conclusion

The predictive analytics tool provides valuable insights into investment behavior based on demographic and financial information. It can help individuals make informed decisions about their investments.