**BUCKMAN Hackathon**

Name : Manish M

Github link : <https://github.com/Manish210103/Buckman>

**Problem :**

The project aims to develop a robust recommendation system for investment decisions, utilising machine learning methodologies and thorough data analysis. By delving into demographic distributions, employment particulars, and investment behaviours, the objective is to identify key factors influencing successful investment outcomes.

**Data Exploration :**

* Utilised PowerBI to create an infographic
* Detailed analysis of investment patterns and their correlation with demographic and socioeconomic factors
* Explored relationships between household income, education, and investment returns across different cities
* Investigated variations in risk levels based on investor experience and knowledge
* Provided insights into reasons for investing across different knowledge levels
* Presented a distribution of investors across various education levels
* Offers a comprehensive overview to aid in identifying factors influencing optimal investment decisions

**Preprocessing :**

* Handled missing values in the "Return Earned" column and replaced them accordingly.
* Encoded categorical variables using one-hot encoding to convert them into numerical format.
* Removed instances with non-numeric values such as "Don't Want to Reveal" in the "Percentage of Investment" column.
* Applied dummy encoding for categorical features and performed feature engineering to calculate the average knowledge level.
* Extracted numerical values from the "Household Income" column and preprocessed textual values in the "Return Earned" column. Finally, exported the processed dataset to a new Excel file named "dataset1.xlsx" for further analysis and modelling.

**Model :**

* **RandomForestClassifier** was chosen for its ability to handle **high-dimensional** data and complex feature relationships, making it suitable for predicting risk levels based on demographic and behavioural data.
* This ensemble learning method aggregates **multiple decision trees**, providing robust and accurate predictions for both binary and multiclass classification problems.
* Default parameters were used for simplicity, but they can be fine-tuned for optimal performance using techniques like grid search or random search.
* Feature selection was conducted with **SelectFromModel** to retain the most important features, **reducing dimensionality** and computational complexity while improving model performance.
* Evaluation metrics such as **accuracy**, **precision**, **recall**, and **F1 score** demonstrated the effectiveness of feature selection in enhancing model performance.
* Users interactively input data via dropdown menus, selecting options based on unique values from the training dataset.
* Processed data is saved to "dataset.xlsx" and passed to a Python script for additional preprocessing before model prediction.
* The trained machine learning model **("model.pkl")** is loaded to **predict** the risk level of the new data.
* The **predicted risk level (High, Low, or Medium)** is displayed to the user based on the model's prediction.

**UI / PowerBI Report :**

* **PowerBI** was utilised for creating an **interactive** and **visually appealing infographic**, offering a comprehensive overview of investment patterns and demographic correlations.
* **Streamlit** was employed to develop a **user-friendly interface** for interactive data input, allowing users to seamlessly input data and obtain real-time predictions of risk levels based on the machine learning model.

**Conclusion :**

The implementation of a recommendation system for predicting investment risk levels using machine learning techniques has provided valuable insights into the complex interplay between demographic, socioeconomic, and behavioural factors. Through preprocessing, model selection, and feature engineering, we have successfully built a robust predictive model capable of accurately classifying risk levels for new investment data.

**Future Work:**

* **Enhanced Feature Engineering:** Further exploration of feature engineering techniques, and dimensionality reduction methods like PCA could help uncover additional patterns and improve model performance.
* **Model Tuning:** Conducting a thorough hyperparameter tuning process using techniques like grid search or Bayesian optimization can optimise model parameters for better predictive accuracy and generalisation.
* **Real-Time Data Integration:** Implementing mechanisms to continuously update the model with real-time data streams can ensure its relevance and accuracy over time, especially in dynamic investment environments.