**Capstone Project 1**

Name: **Jaswinder Singh Ahuja**

Student ID: **S5037**

Cohort ID: **DS37A**

**Introduction:**

Note:

1. Capstone Project 1 - EDA, ML and MySQL – Word file contains answers to the questions with coding.
2. Capstone\_Project\_1\_Jaswinder\_Singh\_Ahuja\_DS037A\_S5037. .ipynb file contains coding.
3. Credit\_card.csv and Credit\_card\_label.csv files are the datasets given in the LMS.
4. File updated\_merged\_data.csv contains updated and merged files with Credit\_card.csv and Credit\_card\_label.csv
5. Credit\_card\_SQL Queries - SQL Text file contains SQL Queries with answers

**Objective:**

The goal of this project was to leverage machine learning to analyze data and predict credit card approvals. We aimed to understand the key factors influencing approval decisions.

**Approach:**

**Initial Hypotheses:**

Hypothesis 1: Income and credit card approval are correlated.

Hypothesis 2: A random forest classifier will yield the highest accuracy.

**Data Analysis:**

* Collected and merged data from 'Credit\_card' and 'Credit\_card\_label' files.
* Explored data using libraries like pandas, seaborn, and matplotlib.
* Handled missing values through imputation and feature engineering.

**Exploratory Data Analysis (EDA):**

* Visualized target variable distribution.
* Explored relationships between variables using visualizations (e.g., stacked bar charts, boxplots, histograms).
* Calculated correlation matrix and identified strong correlations.

**Machine Learning Models:**

* Employed models such as Random Forest, Logistic Regression, Support Vector Machine, and Gradient Boosting.
* Preprocessed data by encoding categorical variables and standardizing numerical features.
* Trained models, made predictions, and evaluated performance metrics (accuracy, precision, recall, F1-score).

**Hypothesis Testing:**

* Conducted t-tests on variables (e.g., income, age, experience) to assess their significance.
* Utilized chi-square tests for categorical variables (e.g., 'Type\_Occupation').

**Results and Conclusions:**

* Found weak evidence supporting the income-approval correlation hypothesis.
* Random Forest model achieved the highest accuracy (94.19%).

**Hyperparameter Tuning:**

* Applied grid search to optimize Random Forest hyperparameters.

**Final Model Evaluation:**

Assessed the tuned Random Forest model's accuracy, confusion matrix, and classification report.

**Conclusion:**

The project revealed that predicting credit card approvals is complex, with income alone not strongly indicating approval. The Random Forest classifier, after hyperparameter tuning, emerged as the most accurate model, achieving 93.87% accuracy on the test set. This underscores the significance of using advanced machine learning techniques for credit card approval predictions.