**End of Streak Project for 6 weeks AI\_foundational Course**

**Project Overview**

In this project, you will apply your data science and machine learning skills to develop a predictive model for Easy Visa Dataset. This project will involve data cleaning, exploratory data analysis (EDA), data preprocessing, model training,hyperparameter tuning and evaluation.

**Project Introduction**

Business communities in the United States are facing high demand for human resources, but one of the constant challenges is identifying and attracting the right talent, which is perhaps the most important element in remaining competitive. Companies in the United States look for hard-working, talented, and qualified individuals both locally as well as abroad.

The Immigration and Nationality Act (INA) of the US permits foreign workers to come to the United States to work on either a temporary or permanent basis. The act also protects US workers against adverse impacts on their wages or working conditions by ensuring US employers' compliance with statutory requirements when they hire foreign workers to fill workforce shortages. The immigration programs are administered by the Office of Foreign Labor Certification (OFLC).

OFLC processes job certification applications for employers seeking to bring foreign workers into the United States and grants certifications in those cases where employers can demonstrate that there are not sufficient US workers available to perform the work at wages that meet or exceed the wage paid for the occupation in the area of intended employment.

**Project Objective**

In FY 2016, the OFLC processed 775,979 employer applications for 1,699,957 positions for temporary and permanent labor certifications. This was a nine percent increase in the overall number of processed applications from the previous year. The process of reviewing every case is becoming a tedious task as the number of applicants is increasing every year.

The increasing number of applicants every year calls for a Machine Learning based solution that can help in shortlisting the candidates having higher chances of VISA approval. OFLC has hired the firm EasyVisa for data-driven solutions. You as a data scientist at EasyVisa have to analyze the data provided and, with the help of a classification model:

* *Facilitate the process of visa approvals.*

*Recommend a suitable profile for the applicants for whom the visa should be*

*certified or denied based on the drivers that significantly influence the case*

*Status.*

**Project Phases**

**Phase 1: Data Collection and Preparation**

Task 1.1: Load the dataset from [here](https://raw.githubusercontent.com/ek-chris/Practice_datasets/refs/heads/main/EasyVisa%20(1).csv).

Task 1.2: Load the dataset into a Pandas DataFrame.

Task 1.3: Inspect the dataset for missing values and handle them appropriately.

Task 1.4: Perform data cleaning to ensure the dataset is ready for analysis.

**Phase 2: Exploratory Data Analysis (EDA)**

Task 2.1: Conduct exploratory data analysis to understand the distribution of features and the target variable .

Task 2.2: Visualize the relationships between features and the target variable using scatter plots, histograms, and box plots.

Task 2.3: Identify and handle outliers in the dataset.

**Phase 3: Data Preprocessing**

Task 3.1: Create new features that may help improve the model's performance.

Task 3.2: Encode categorical variables using one-hot or label encoding techniques.

Task 3.3: Normalize or standardize numerical features as needed.

**Phase 4: Model Training and Evaluation**

Task 4.1: Split the dataset into training and testing sets.

Task 4.2: Choose and justify the selection of machine learning algorithms (e.g., Linear Regression, Decision Tree, Random Forest, Gradient Boosting).

Task 4.3: Train multiple models and evaluate their performance using appropriate metrics (e.g.,Confusion matrix(precision, sensitivity, accuracy, F1-score, AUC-ROC).

Task 4.4: Perform hyperparameter tuning to optimize the model's performance.

Task 4.5: Select and evaluate the best-performing model on the testing set.

**Phase 5: Model Interpretation and Reporting**

Task 5.1: Interpret the results of the best-performing model and explain the importance of critical features.

Task 5.2: Create visualizations to support your findings and model interpretations.

Task 5.3: Write a comprehensive report summarizing the project, including the methodology, results, and conclusions.

**Phase 6: Model Interpretation and Reporting**

Task 6.1: Build a machine learning application for your model using gradio.

**Deliverables**

**Code:** Submit the complete codebook used for data preparation, EDA, feature engineering, model training, and evaluation (Jupyter Notebook format)

**Report:** Submit a detailed report (PDF or google doc) documenting your approach, findings, and conclusions. The report should include visualizations and a clear explanation of your steps.

**Dataset:** [The Easy visa dataset](https://raw.githubusercontent.com/ek-chris/Practice_datasets/refs/heads/main/EasyVisa%20(1).csv)

**Data Description**

The data contains the different attributes of the employee and the employer. The detailed data dictionary is given below.

**case\_id**: ID of each visa application

**continent**: Information of continent the employee

**education\_of\_employee**: Information of education of the employee

**has\_job\_experience**: Does the employee have any job experience? Y= Yes; N = No

**requires\_job\_training**: Does the employee require any job training? Y = Yes; N = No

**no\_of\_employees**: Number of employees in the employer's company

**yr\_of\_estab**: Year in which the employer's company was established

**region\_of\_employment**: Information of foreign worker's intended region of employment in the US.

**prevailing\_wage**: Average wage paid to similarly employed workers in a specific occupation in the area of intended employment. The purpose of the prevailing wage is to ensure that the foreign worker is not underpaid compared to other workers offering the same or similar service in the same area of employment.

**unit\_of\_wage**: Unit of prevailing wage. Values include Hourly, Weekly, Monthly, and Yearly.

**full\_time\_position**: Is the position of work full-time? Y = Full Time Position; N = Part Time Position

**case\_status**: Flag indicating if the Visa was certified or denied