## Group 9 Capstone Project Objective & Dataset Summary

### ****Project Title: Predicting Income Level Using U.S. Census Data****

### ****Objective:****

Group 9’s capstone project aims to build a predictive machine learning model to classify whether an individual earns more than $50,000 annually using demographic and socio-economic attributes from U.S. Census data. The project’s goal is to analyze how factors such as education, age, occupation, capital gains, and hours worked contribute to income levels while developing an explainable, high-performing model suitable for practical insights in workforce analytics and policy planning.

### ****Dataset:****

We will use the **UCI Adult (Census Income) Dataset**, which contains approximately 48,000 rows with 14 predictive features plus 1 target variable (income). The features include age, education level, occupation, hours worked per week, marital status, and financial attributes such as capital gain and capital loss. The target variable indicates whether an individual’s income is <=50K or >50K per year, allowing us to frame the task as a binary classification problem.

### ****Methodology:****

Our methodology involves thorough data cleaning and exploratory data analysis to understand distributions and relationships within the dataset. To address class imbalance, we will apply SMOTE, followed by model training using Logistic Regression, Random Forest, and XGBoost, leveraging GridSearchCV for hyperparameter tuning to optimize model performance. We will evaluate models using metrics such as ROC-AUC and F1-score and utilize SHAP for model interpretability to identify and communicate the key drivers influencing income levels transparently.

This project aligns with the course’s objective of applying end-to-end data science processes to real-world data, demonstrating the practical application of classification, evaluation, and interpretability techniques.

### ****Group 9 Members:****

### ThankGod Israel – [tgisrael@osiriuniversity.org](mailto:tgisrael@osiriuniversity.org)

* Oni Akintunde Julius [–akinoni@osiriuniversity.org](mailto:–akinoni@osiriuniversity.org)
* Victory Madu [–vicmadu@osiriuniversity.org](mailto:–vicmadu@osiriuniversity.org)
* Chikezie Amarachi Stephanie [–amachikezie@osiriuniversity.org](mailto:–amachikezie@osiriuniversity.org)
* Doris Akachukwu – [doakachukwu@osiriuniversity.org](mailto:doakachukwu@osiriuniversity.org)
* Mary Paschal Iwundu [–miwundu@osiriuniversity.org](mailto:–miwundu@osiriuniversity.org)

### 🔧 Possible Improvements and Future Work

To further improve the model and enhance this project in future iterations, the following steps could be considered:

✅ **Advanced Feature Encoding**: Implementing techniques like **target encoding** or **embedding layers** for categorical variables instead of one-hot encoding could reduce dimensionality and improve model performance.

✅ **Hyperparameter Optimization**: Using **Bayesian optimization** or **Optuna** instead of GridSearchCV could yield better results faster and with less computational cost.

✅ **Model Stacking or Ensembling**: Combine multiple models (e.g., Logistic Regression + Random Forest + XGBoost) to form a **stacked ensemble**, which can often outperform individual models.

✅ **Cross-Validation with Stratified K-Folds**: While we used stratified train-test split, using **StratifiedKFold** for cross-validation across all training stages can give a more reliable estimate of model performance.

✅ **Model Deployment**: Packaging the final model using **Flask**, **Streamlit**, or **FastAPI** into a web app would make it accessible to end-users or stakeholders.

✅ **Addressing Feature Bias**: Perform **fairness analysis** to ensure the model does not unfairly discriminate across groups like gender or race.