**REPORT**

**Class**: AI17C

**Subject**: DBM302m

**Instructor**: Nguyen Van Vinh – VinhNV27

**Group**: 1

**Members**:

Ha Khai Hoan - QE170157

Dang Phuc Bao Chau - QE170060

Nguyen Van Thu – QE170147

1. **Which problem are you trying to solve why?**

I am interested in understanding which socio-economic factors most influence an individual's income. Specifically, I would like to explore the relationship between factors such as age, education, occupation, and gender in predicting whether an individual will earn more than $50,000 per year. Additionally, I would like to explore whether there are significant income differences between genders and races.

This would be helpful in areas such as:

* **Advertising and marketing**: Companies can target high or low-income groups to offer suitable products or services.
* **Credit analysis**: Financial institutions and banks can use this information to assess repayment ability, plan loans, and set credit limits.
* **Customer segmentation**: Businesses can divide customers by income to develop tailored business strategies, thereby increasing sales efficiency.
* **Public policy**: Government agencies can use this data to shape social policies, such as welfare support for low-income households.
* **Insurance**: Insurance companies can assess risks or design insurance packages tailored to different income groups.
* **Real estate**: Real estate brokers can use this information to predict housing demand among high or low-income earners.
* **Education**: Educational institutions can offer scholarships or support programs based on income levels.

1. **Where and how do you obtain the data? How big is your data?**

We took the Adult Census Income dataset on Kaggle, which is a popular dataset often used to build machine learning models that predict individual income based on demographic factors.

+ Number of rows of data: 32561

+ Number of columns of data: 15

Note:

|  |  |  |
| --- | --- | --- |
|  | **Feature** | **Description** |
| 1 | Age | Describes the age of individuals. Continuous. |
| 2 | Workclass | Private, Self-emp-not-inc, Self-emp-inc, Federal-gov, Local-gov, State-gov, Without-pay, Never-worked. |
| 3 | fnlwgt | Continuous. A weighting factor created by the US Census Bureau indicating the number of people represented by each data entry. |
| 4 | Education | Bachelors, Some-college, 11th, HS-grad, Prof-school, Assoc-acdm, Assoc-voc, 9th, 7th-8th, 12th, Masters, 1st-4th, 10th, Doctorate, Preschool. |
| 5 | Education-num | Number of years spent in education. Continuous. |
| 6 | Marital-status | Married-civ-spouse, Divorced, Never-married, Separated, Widowed, Married-spouse-absent, Married-AF-spouse. |
| 7 | Occupation | Tech-support, Craft-repair, Other-service, Sales, Exec-managerial, Prof-specialty, Handlers-cleaners, Machine-op-inspct, Adm-clerical, etc. |
| 8 | Relationship | Wife, Own-child, Husband, Not-in-family, Other-relative, Unmarried. |
| 9 | Race | White, Asian-Pac-Islander, Amer-Indian-Eskimo, Other, Black. |
| 10 | Sex | Female, Male. |
| 11 | Capital-gain | Represents the profit from the sale of assets (e.g., stocks or real estate). Continuous. |
| 12 | Capital-loss | Represents the loss from the sale of assets (e.g., stocks or real estate). Continuous. |
| 13 | Hours-per-week | Continuous. |
| 14 | Native-country | List of countries including United-States, Cambodia, England, Puerto-Rico, Canada, Germany, etc. |
| 15 | Salary | >50K, <=50K. |

1. **What are your ideas to solve the problem?**

My approach is to apply various machine learning classification algorithms such as:

+ Logistic Regression for its simplicity and interpretability.

+ Random Forest for handling non-linear relationships and importance weighting of features.

+ KNN (K-Nearest Neighbors) is a supervised learning algorithm.

The pipeline will include:

+ Data preprocessing:

. Missing handle

. Duplicate handle

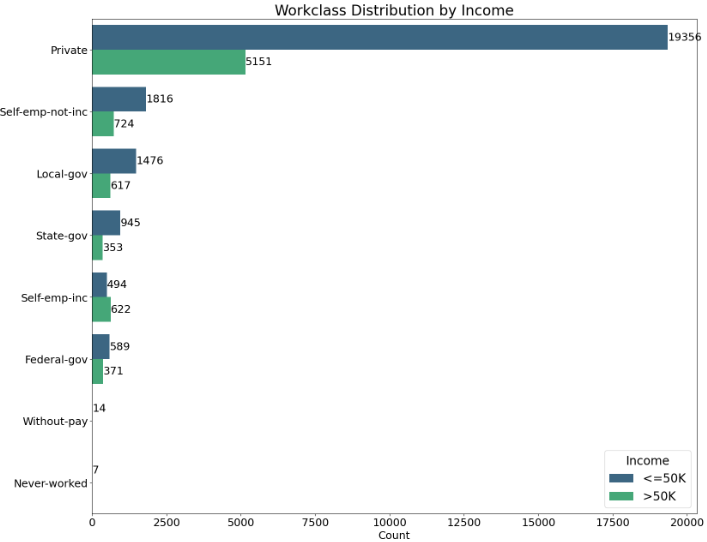
. Outlier handle

+ Feature engineering:

Separate categorical and numerical features for easy management.

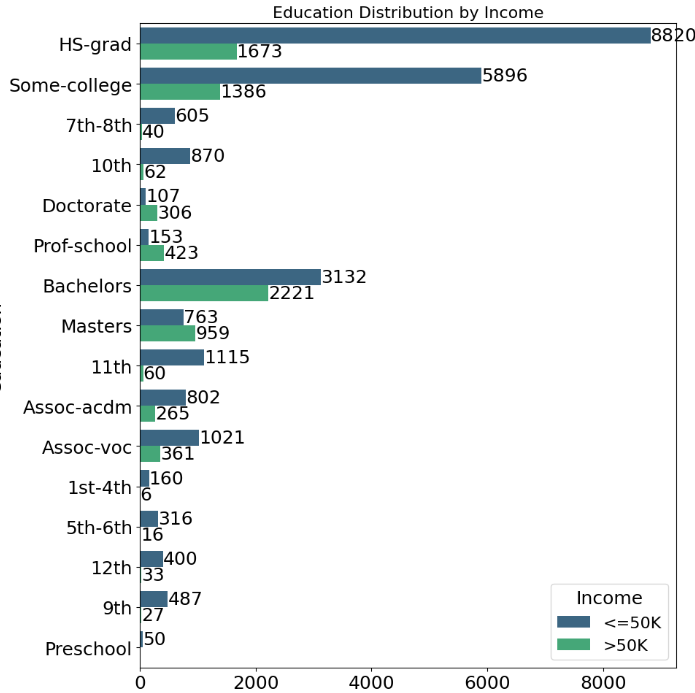
* Categorical features

Example: [“Income”]



* Numerical features

Example: [“education”]



+ Build model

+ Model tuning to optimize performance.

In addition, I also visualized the data to better understand the interactions between features, to identify which groups of factors are important in predicting whether a person is truly high-income or not.

1. **What is your hypothesis for the ideas to work? A more interesting question is how do you verify your hypothesis?**

Hypothesis: Certain features such as **education, age, and occupation** will have the strongest predictive power for determining income. I hypothesize that more educated individuals or those in higher-tier occupations are likely to earn more than 50K USD.

To verify this, I will:

+ Conduct exploratory data analysis (EDA) to check feature distributions.

+ Use feature importance analysis from Random Forest and XGBoost.

+ Compare model performance through accuracy, precision, recall, and F1-score on a test dataset.

+ Validate the models with cross-validation to ensure generalizability.

1. **How does the result look like? Does it confirm your hypothesis?**

*# Pending*

1. **What have you done to make your original ideas better?**

*# Pending*

1. **What is the running time of your algorithm? Is your algorithm scalable?**

*# Pending*

1. **If you are given more time, what can be done to even improve it further?**

*# Pending*

1. **What have you learned from the project?**

*# Pending*