**Technical Report on Finding donors**

**Introduction**

In this project, the objective is to build machine learning models that classify individuals into two categories: either earning more than $50,000 or less. This classification is based on data from the 1994 U.S. Census. The model aims to help a charity organization called "CharityML" identify individuals who are more likely to donate based on their income.

**1. Algorithms Used:**

The project employs several common classification algorithms. Here’s a brief overview of each:

**A. Decision Tree**

* **Decision Tree classifier was used as one of the potential models and was trained on the data**

**B. Random Forest**

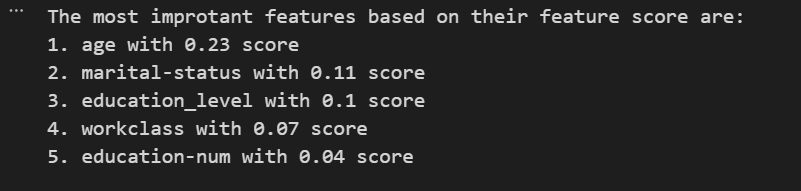
* **Random forest classifier was used in the process of extracting feature importance**

**C. K Neighbors Classifier**

**2. Data Analysis and Feature Selection:**

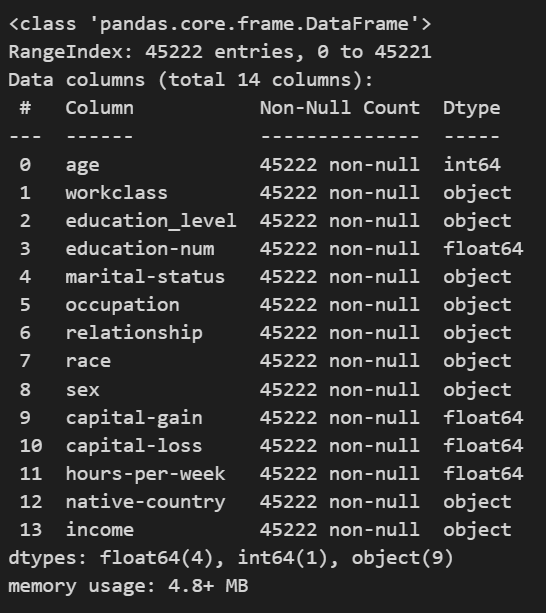
**A. Data Exploration**

* The dataset was imported and analyzed to identify the most impactful features using (feature\_importances\_)\_function

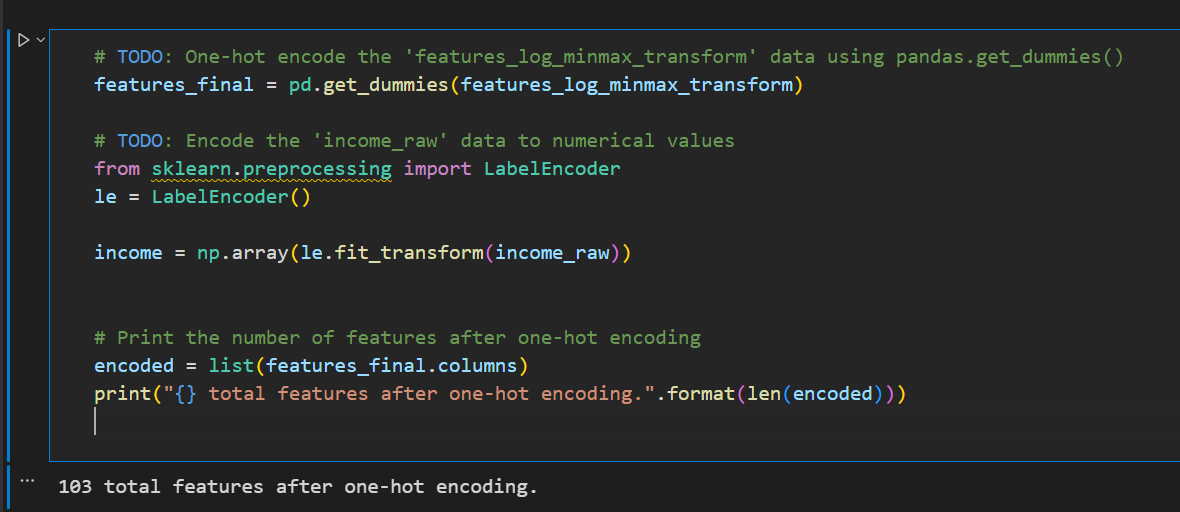


**B. Handling Missing or Dirty Data**

* The data had no missing values.

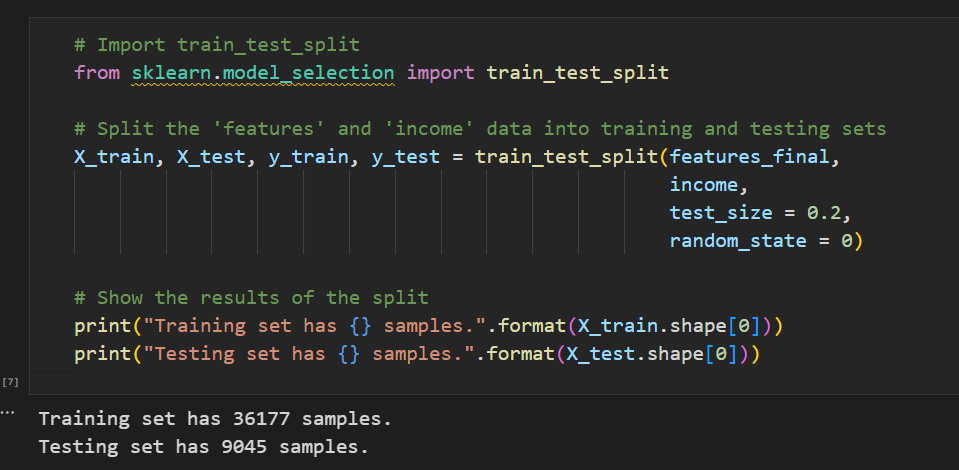


* Categorical variables were converted into numerical values using **one-hot encoding** to transform text-based features like "occupation" into numeric formats that can be used in models.



**C. Data Splitting**

* The data was split into training (80%) and testing sets (20%) to ensure fair evaluation of the models and prevent overfitting.



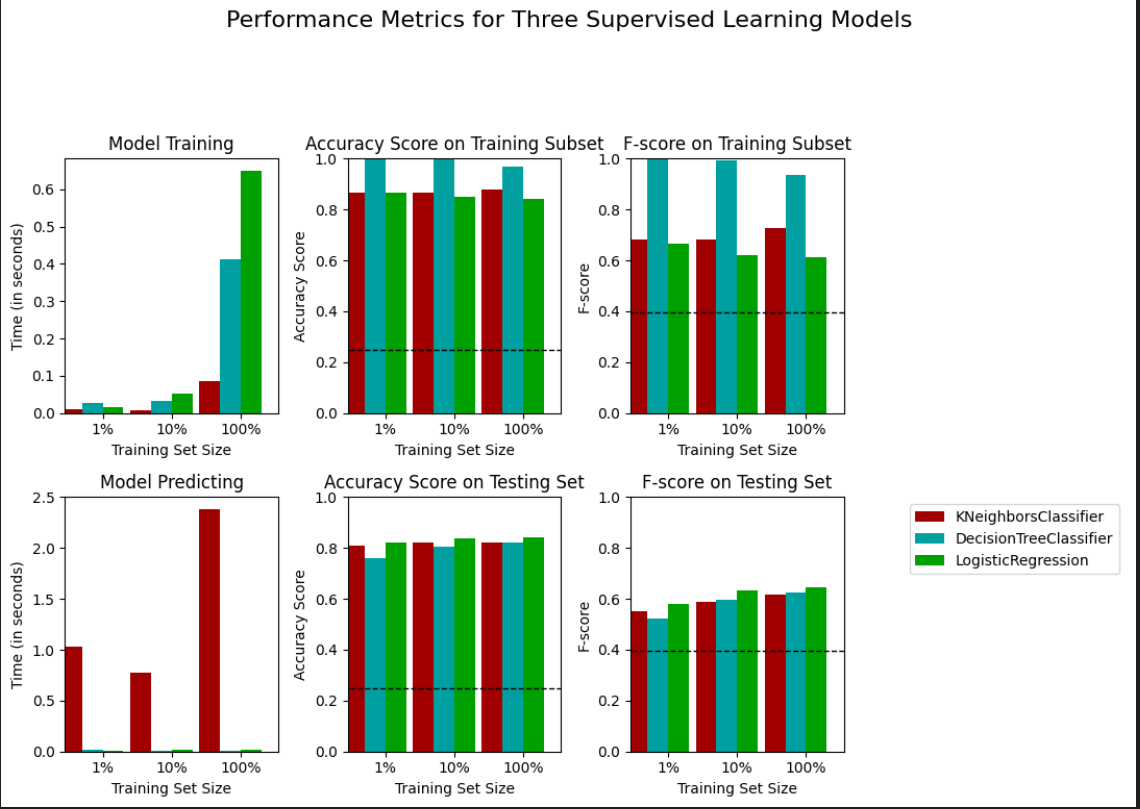
**3. Challenges and Potential Issues:**

**A. Data Imbalance**

* Most individuals in the dataset earn less than $50,000, meaning the data is imbalanced. This could bias the model towards favoring the larger class.
  + Total number of records: 45222
  + Individuals making more than $50,000: 11208
  + Individuals making at most $50,000: 34014
  + Percentage of individuals making more than $50,000: 24.78%
* Potential Solutions: employing the **F1-score** to measure performance instead of relying solely on accuracy.

**B. Choosing the Right Model**

* A major challenge is selecting the most appropriate algorithm. So we compared model performance using to ensure the model works well across different data samples



Based on those results we came to the this conclusion

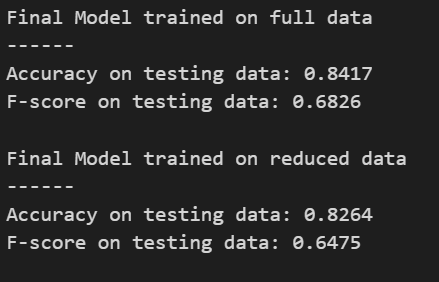
- The model with the best F1-score on 100% of the testing set data is **Logistic regression** model

- The training time of the model is high, but the prediction time is low which is good

**4. Optimization and Evaluation:**

**A. Model Evaluation**

* Metrics such as **accuracy**, **recall**, and **precision** were used to evaluate model performance.



* **Grid Search** was used to tune hyperparameters and improve the final model.

**C. Final Prediction**

* The final model chosen was the most accurate, and it was fine-tuned using techniques like **cross-validation** and **hyperparameter tuning** to ensure it performs well on new data.