Income Classification Model Report

# Table of Contents

1. **Introduction**

1.1 Importance of Implementing our model and doing statistics

1.2 Objective of the project

1. **Uploading Dataset**

2.1 Importing packages, libraries, and warnings

2.2 Data Dictionary

1. **Data Cleaning**

3.1 Dealing with the missing values; One-Hot Encoding/Duplicates/Outliers/Null Values

1. **Data Pre-Processing, EDA & Feature Engineering**

4.11 Income

4.12 Education (Income Distribution by Education)

4.13 Marital-status

4.14 Occupation

4.15 Gender (Income Distribution by Gender)

4.16 Race (Income Distribution by Race)

4.17 Native Country (Income Distribution by Native-Country)

4.18 Work class (Income Distribution by Work Class)

4.19 Relationship (Income Distribution by Relationship)

4.21 Age (Income Distribution by Age)

4.22 Educational Num (Income Distribution by Educational Num)

4.23 Capital gain and Capital loss; Create a crossing feature Capital Diff (Income Distribution by Capital Diff)

4.24 Hours Per Week (Income Distribution by Hours Per Week)

5. **Feature Selection Encoding**

5.1 Label Encoding

5.2 Standard Scaling

5.3 Dividing Data into training and testing dataset

5.4 Class Balancing

6. **Building Machine Learning Model**

6.1 Logistic Regression

6.2 Confusion Matrix for logistic Regression Model

6.3 Classification Report for logistic Regression Model

6.4 K-Nearest Neighbors

6.5 Confusion Matrix for K-Nearest Neighbors Model

6.6 Classification Report for K-Nearest Neighbors Model

# Introduction

Data preprocessing involved systematic steps to clean and structure the dataset, ensuring optimal model performance

## Importance of Implementing our model and doing statistics

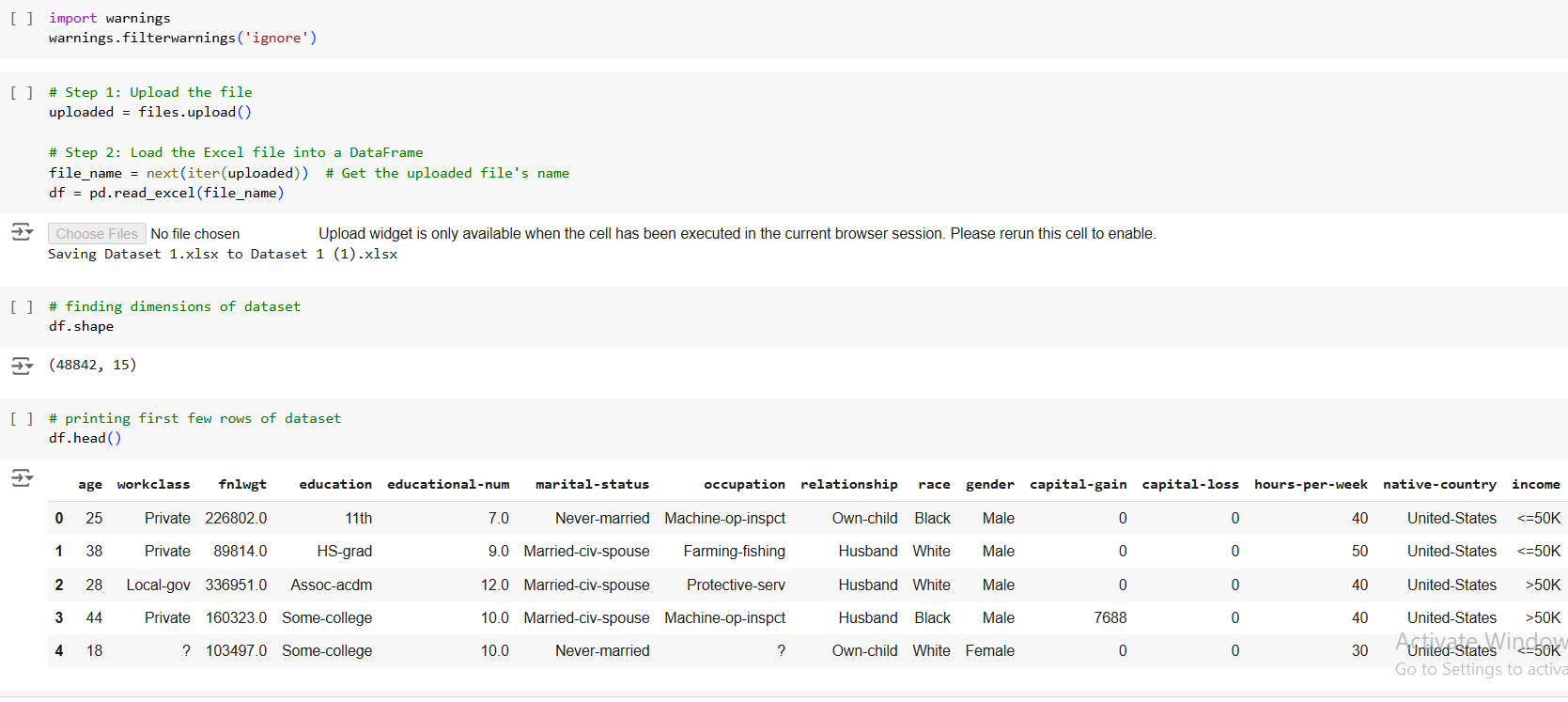
The data as I studied about it is about census which is a special, wide-range activity, which takes place once a decade in the entire country. The purpose is to gather information about the general population, in order to present a full and reliable picture of the population in the country - its housing conditions and demographic, social and economic characteristics. The information collected includes data on age, gender, country of origin, marital status, housing conditions, marriage, education, employment, etc.

This information enables more effective planning of services, enhances the quality of life, and helps address existing challenges.

## Objective of the project

The goal of this machine learning project is to predict whether a person makes over 50K a year or not given their demographic variation. To achieve this, the K-Nearest Neighbors technique is implemented that would yield the best prediction result.

# Uploading Dataset



## 2.2 Data Dictionary

### **Categorical Attributes**

* workclass(Individual work category):Private, Self-emp-not-inc, Self-emp-inc, Federal-gov, Local-gov, State-gov, Without-pay, Never-worked.
* education(Individual's highest education degree): Bachelors, Some-college, 11th, HS-grad, Prof-school, Assoc-acdm, Assoc-voc, 9th, 7th-8th, 12th, Masters, 1st-4th, 10th, Doctorate, 5th-6th, Preschool.
* marital-status(Individual marital status): Married-civ-spouse, Divorced, Never-married, Separated, Widowed, Married-spouse-absent, Married-AF-spouse.
* occupation(Individual's occupation): Tech-support, Craft-repair, Other-service, Sales, Exec-managerial, Prof-specialty, Handlers-cleaners, Machine-op-inspct, Adm-clerical, Farming-fishing, Transport-moving, Priv-house-serv, Protective-serv, Armed-Forces.
* relationship(Individual's relation in a family): Wife, Own-child, Husband, Not-in-family, Other-relative, Unmarried.
* Race(Race of Individual): White, Asian-Pac-Islander, Amer-Indian-Eskimo, Other, Black.
* Gender(Individual's Gender): Female, Male.
* native-country(Individual's native country): United-States, Cambodia, England, Puerto-Rico, Canada, Germany, Outlying-US(Guam-USVI-etc), India, Japan, Greece, South, China, Cuba, Iran, Honduras, Philippines, Italy, Poland, Jamaica, Vietnam, Mexico, Portugal, Ireland, France, Dominican-Republic, Laos, Ecuador, Taiwan, Haiti, Columbia, Hungary, Guatemala, Nicaragua, Scotland, Thailand, Yugoslavia, El-Salvador, Trinadad&Tobago, Peru, Hong, Holand-Netherlands.

### **Continuous Attributes**

* age(Individual's age)
* education-num(Individual's year of receiving education)
* fnlwgt(Individual's final weight)
* capital-gain
* capital-loss
* hours-per-week(Individual's working hour per week)

# Cleaning Data

## 3.1 Dealing with the missing values; One-Hot Encoding/Duplicates/Outliers/Null Values

**Encoding and Dummy Variables:**

Work-class, education, marital-status, occupation, relationship, race, Gender, and native-country are categorical variables encoded to binary values to be used for logistic regression. Moreover, a dummy variable has been created named Capital-diff by subtracting capital-loss from capital-gain to tell the capital difference.

**Duplicates:**

There were 57 duplicate rows found in the dataset, and these rows were dropped.

**Outliers:**

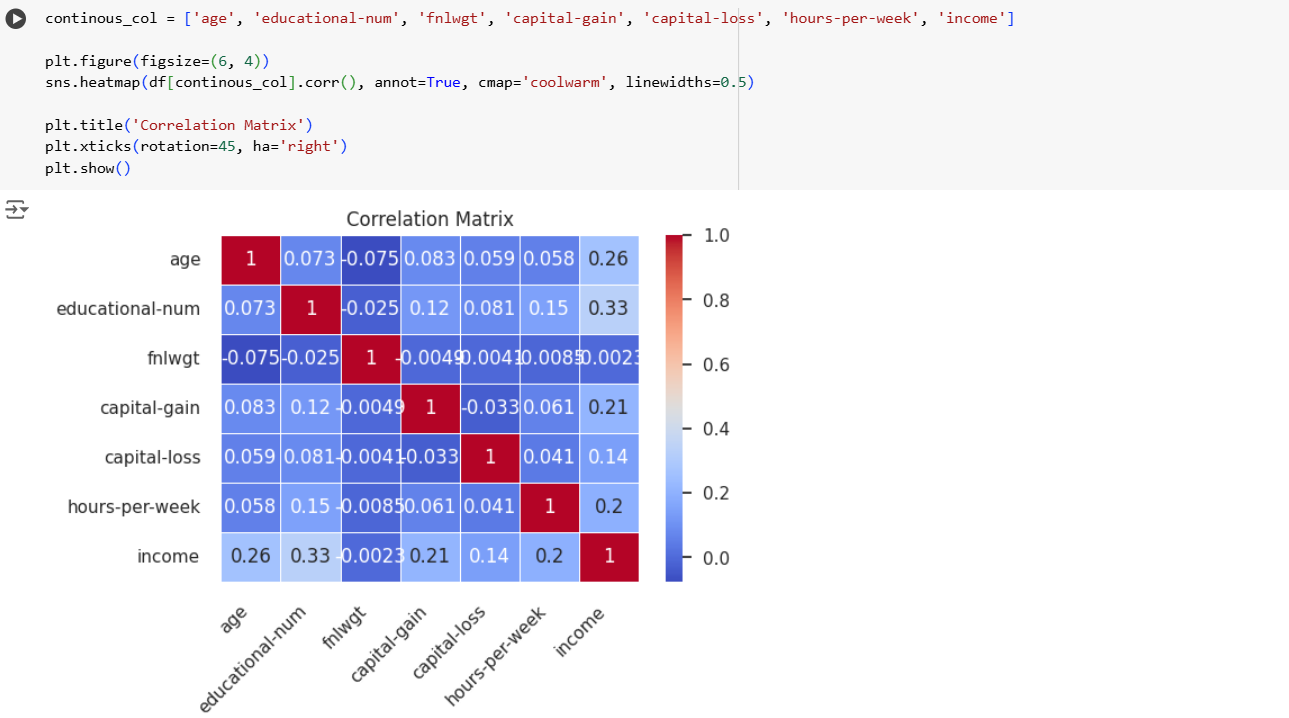
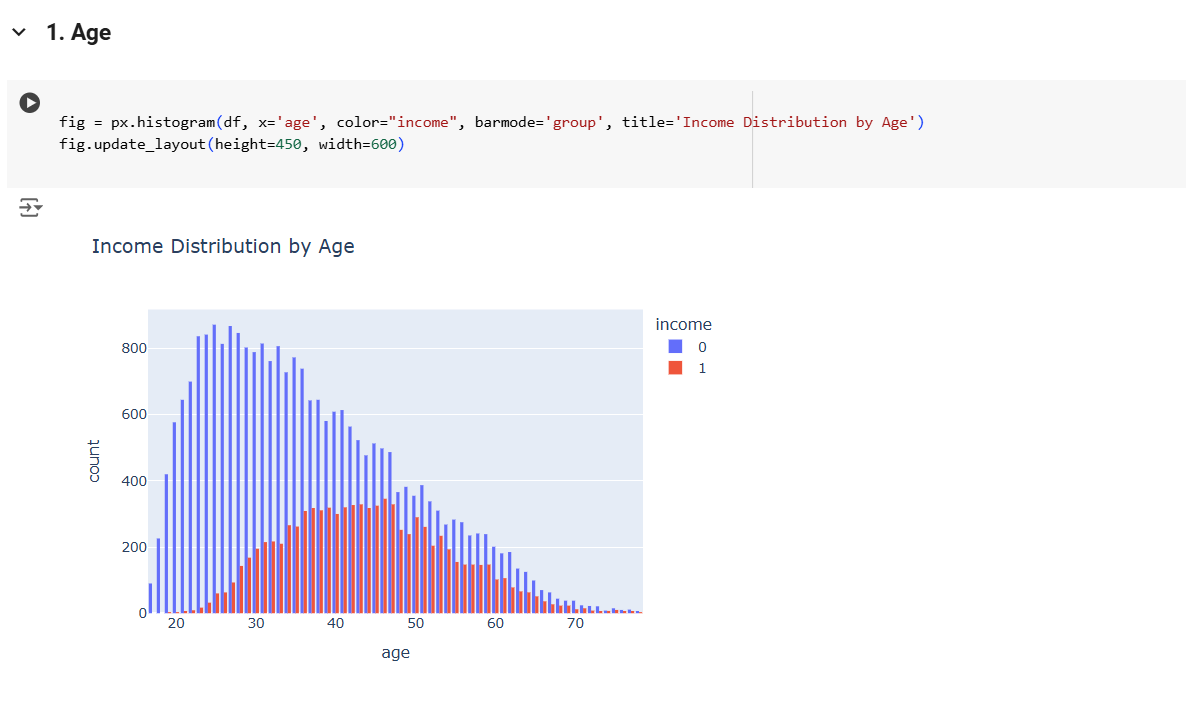
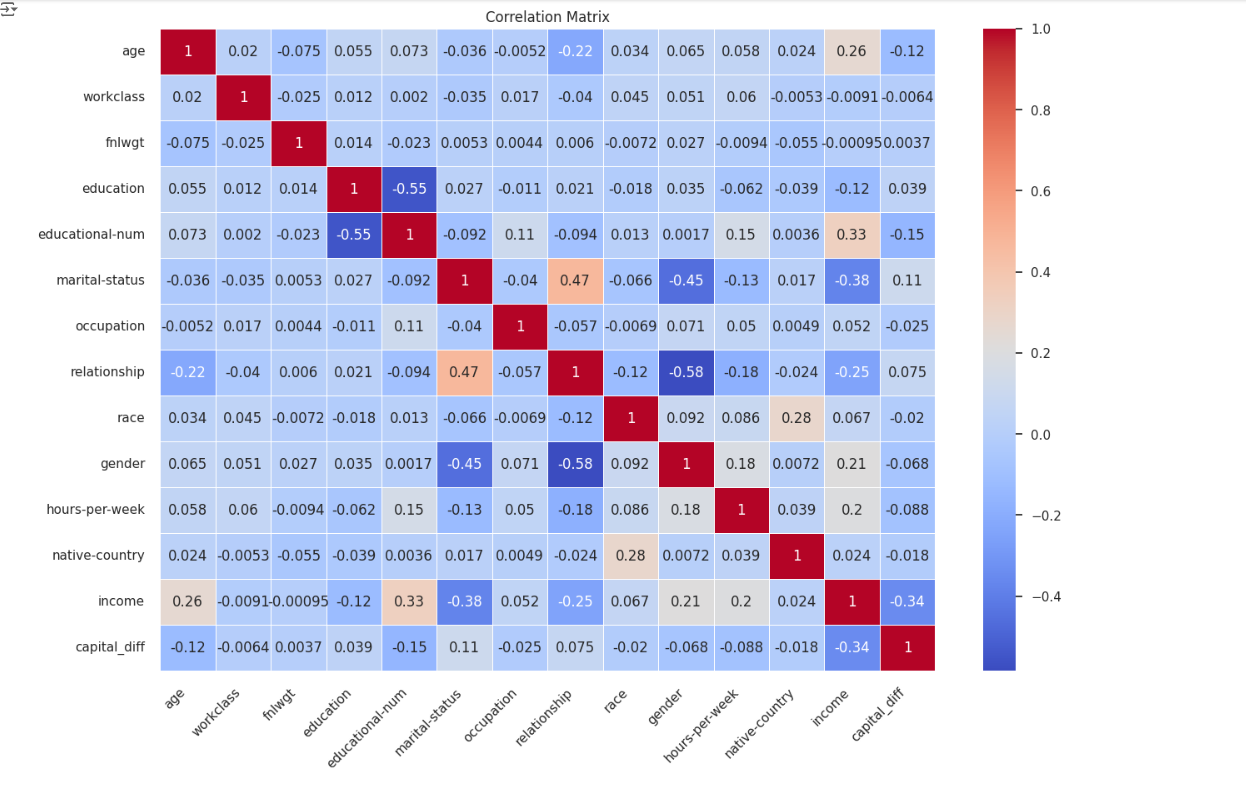
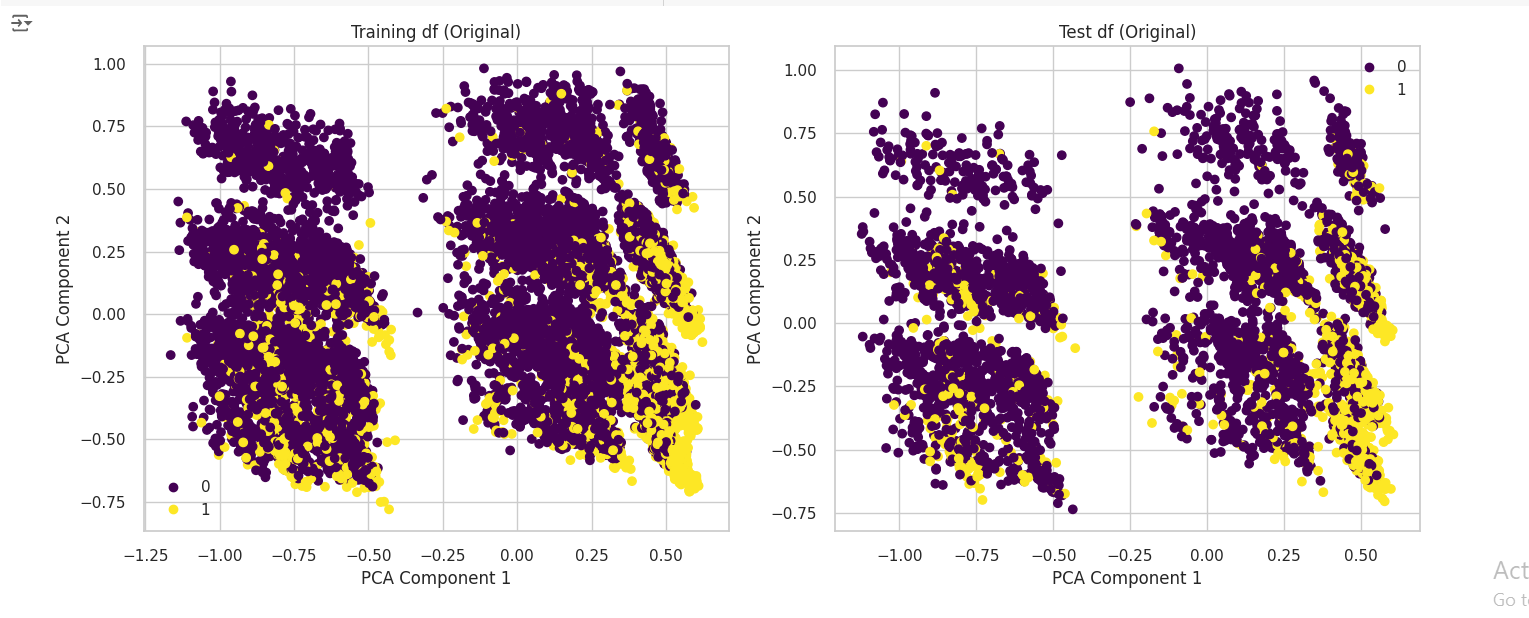
Boxplots show outliers for age, educational-num, fnl-weight, capital-gain and capital-loss. The rows having all these outliers are dropped so that it doesn’t affect our regression (income) variable. Outliers were removed based on IQR and the variables capital-gain and capital loss were dropped because they were having no effect on the model and were reducing the accuracy.

**Null/Missing Values in Dataset:**

Except for age, capital-gain, capital-loss, hours-per-week and income, every feature has null values, which have been treated with appropriate measures of tendency.

# Data Pre-Processing, EDA & Feature Engineering

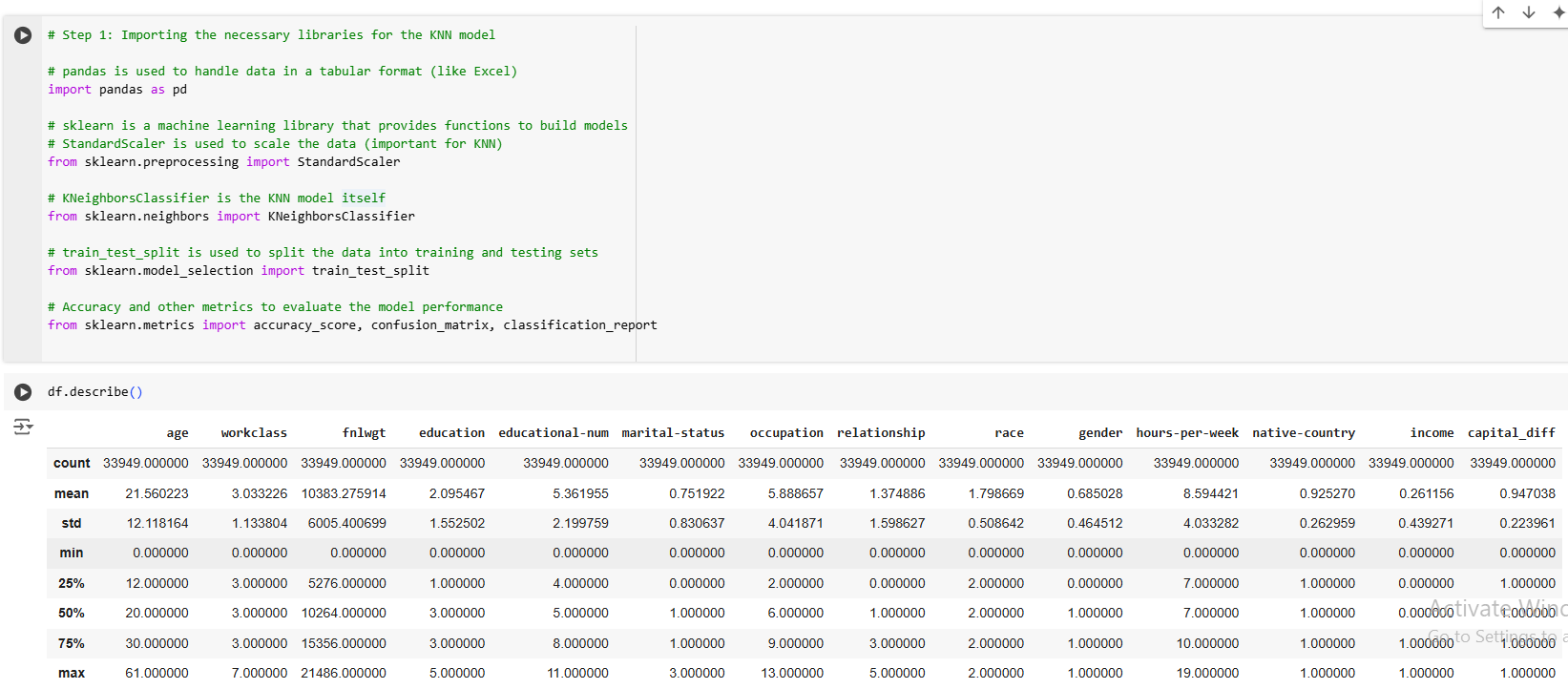
Five most important visualizations are given below.

* For Dependent variable(income):  
    
  In the below chart, we can see that less than 15 thousand people falling in the category of income greater than 50K slot. More than 35 thousand people fall in the category of having income less than a 50K stamp.
* Correlation Matrix for continuous variables:  
    
    
  This correlation matrix shows relationships among continuous variables. Notably, "educational-num" has the highest positive correlation with income (0.33), followed by "age" (0.26) and "hours-per-week" (0.2), suggesting these may influence income. Capital gain and capital loss have almost no correlation. Darker red cells indicate strong positive relationships, while darker blue indicates negative ones. The diagonal cells, all at 1, represent each variable's correlation with itself. Later the variables capital-loss and capital-gain would be dropped.
* Income distribution by age:  
    
  This histogram shows income distribution across different age groups. Blue bars represent lower-income individuals (income = 0), while red bars represent higher-income individuals (income = 1). Younger age groups (under 30) have mostly lower income, with higher income becoming more common in middle age (around 30-50). The proportion of higher-income individuals peaks between ages 40 and 50. Both income groups decrease in count after age 50.
* Correlation Matrix for both categorical and continuous variables:  
  
* Principal component Analysis for trained and tested Datasets:  
    
  These scatter plots visualize the training and test datasets using the first two Principal Component Analysis (PCA) components. Each point represents a data sample, with purple indicating the lower-income group (0) and yellow indicating the higher-income group (1).
* **Training Dataset (left)**: The data points are well-distributed across the PCA space, with some clustering, but there is overlap between income groups.
* **Test Dataset (right)**: The test set mirrors the distribution of the training set, showing a similar pattern with clusters and overlap between income groups.

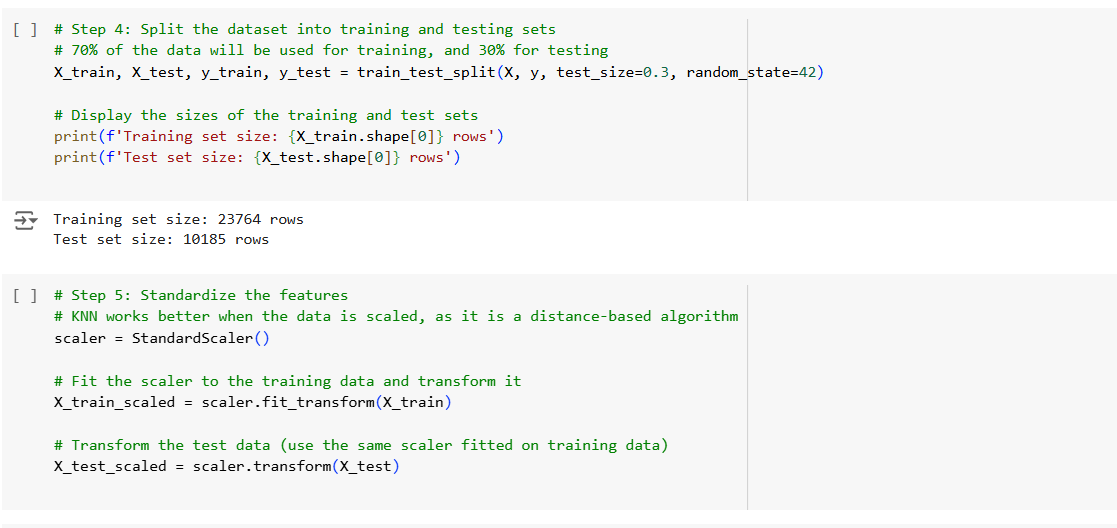
This suggests consistency in the dataset’s structure between training and test sets, with no clear separation between income groups in the PCA-transformed space.

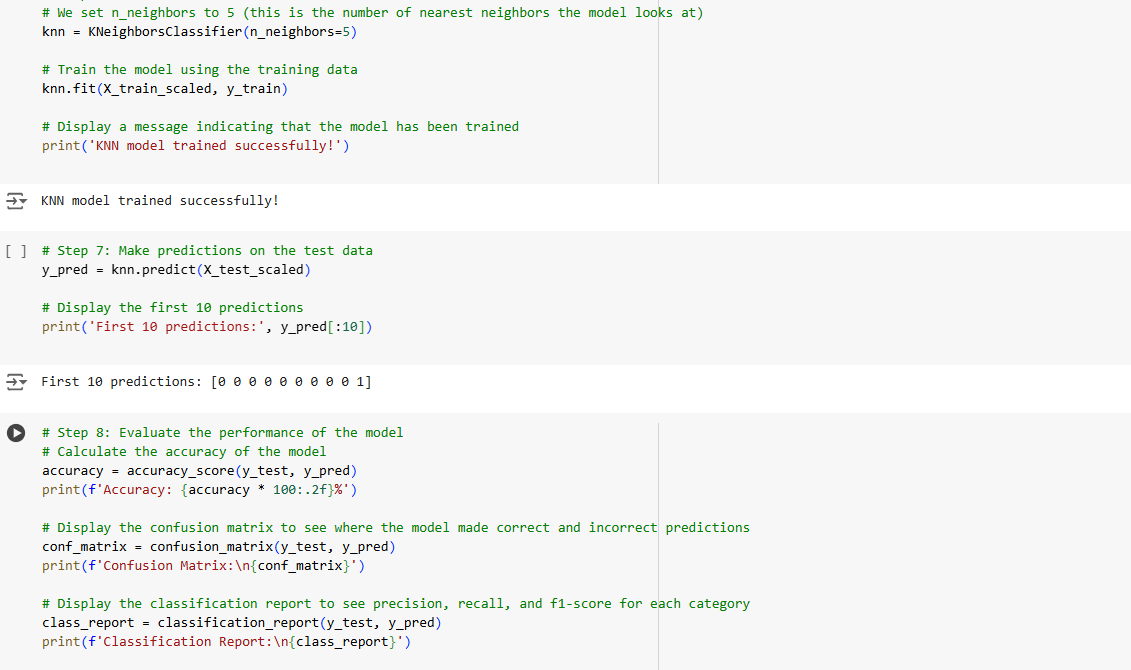
# Building Machine Learning Model

## K Nearest Neighbors

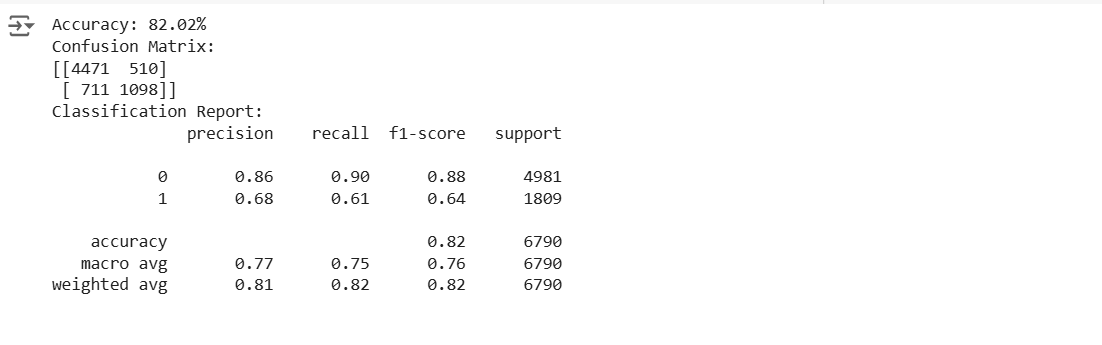


* I imported StandardScalar which is important for KNN and I would use it for scaling the data
* I imported KNeighboursClassifier to classify our dataset, train-test-split for splitting, training and testing the data, and confusion matrix, accuracy score, and classification report to predict better and explain the model further





Following is the **classification report, accuracy, and confusion matrix** that was obtained:



**→ Precision:**

**86%** of the predicted outcome (0→ <=50K) matches exactly the results we actually had in our outcome of test data. Just 14% false positive prediction. **<=50k in actual/<=50k predicted (few were missed)**

Moreover, **68%** precision is recorded for predicting class 1 (1 → >50k), meaning 68% of the predicted result matches the actual outcome. 32% false positive predictions. **>50k in actual/>50k predicted (moderate were missed)**

*The model is best at recognizing class 0 instead of class 1.*

**→ Recall:**

Impressive **90%** of the actual occurrences of class 0 were measured correctly by the model. So, the model only missed **10%** of class 0 instances by predicting them as 1. **<=50k predicted/<=50k actual.**

The model correctly moderately predicted **61%** of the actual occurrences of class 1. It missed 39% of actual >50k cases by predicting them as 0. **>50k predicted/>50k actual.**

**→ F1-Score:**

F1-score is the **harmonic mean** of both precision and recall.

The F1-score of **0.88 or 88% for class 0** shows good performance for class 0, while the f1-score of

**0.64 or 64% for class 1** shows relatively moderately underperforming for class 1

**→ Support:**

It shows the actual number of occurrences of each class in the dataset. As discussed earlier, there does remain a moderate high-class imbalance, with class 0 occurring **approximately 2.5 times more** than class 1.

**→ Accuracy:**

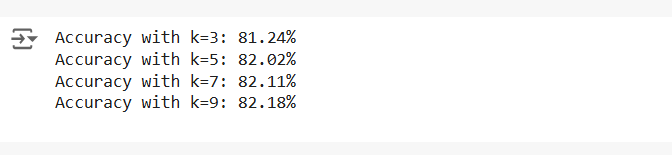
The overall accuracy of the model in predicting the outcome is **82.02%,** which is good but due to a moderate bias towards class 0, it’s difficult to improve the model more as noted by recall and precision score.

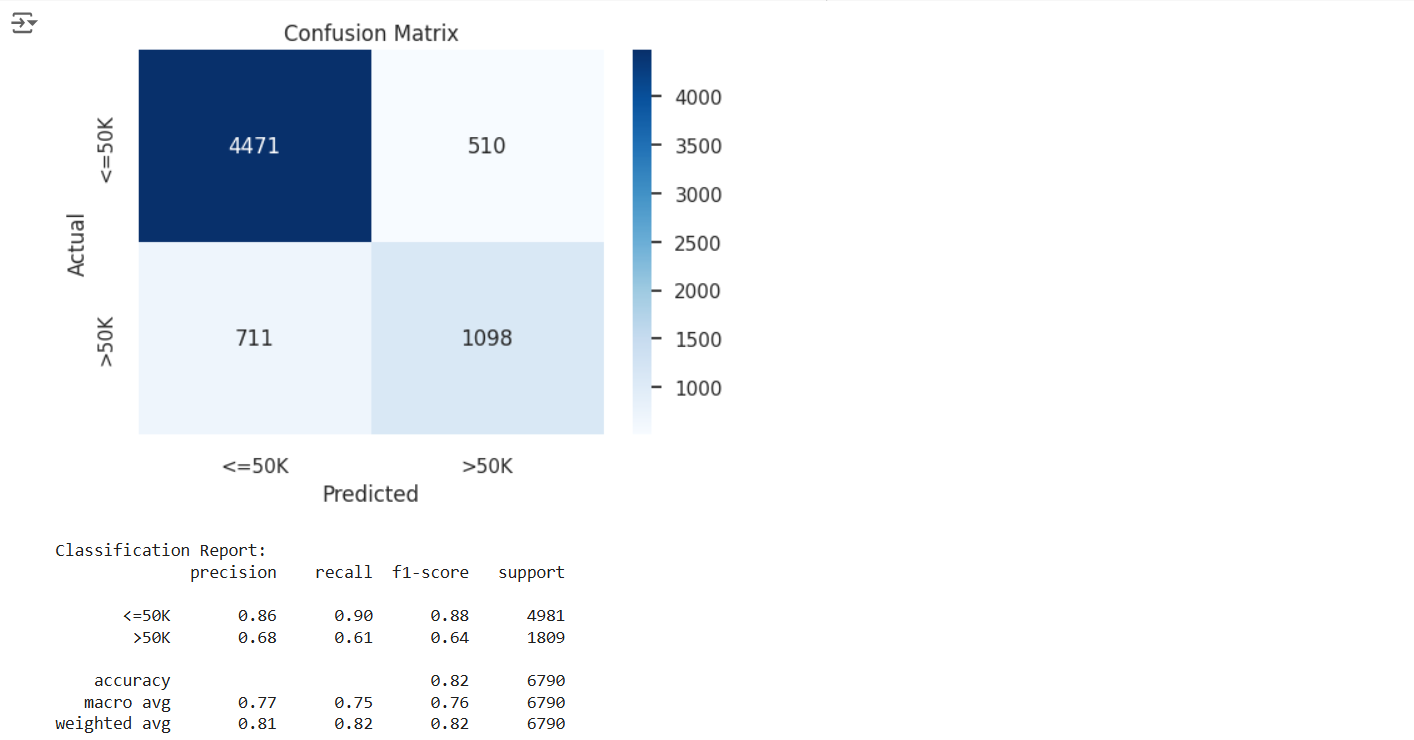
**Overall conclusion:**

Model slightly struggles to predict instances of class 1, but has a reasonable precision and recall for both the classes..

Overall, The model is more effective at predicting **<=50k** than **>50k** cases, with high precision for **<=50k** predictions. However, its lower precision for **>50k** suggests it may produce too slightly more false positives when predicting customer **>50k.**

* **For Different Values of K, accuracy:**





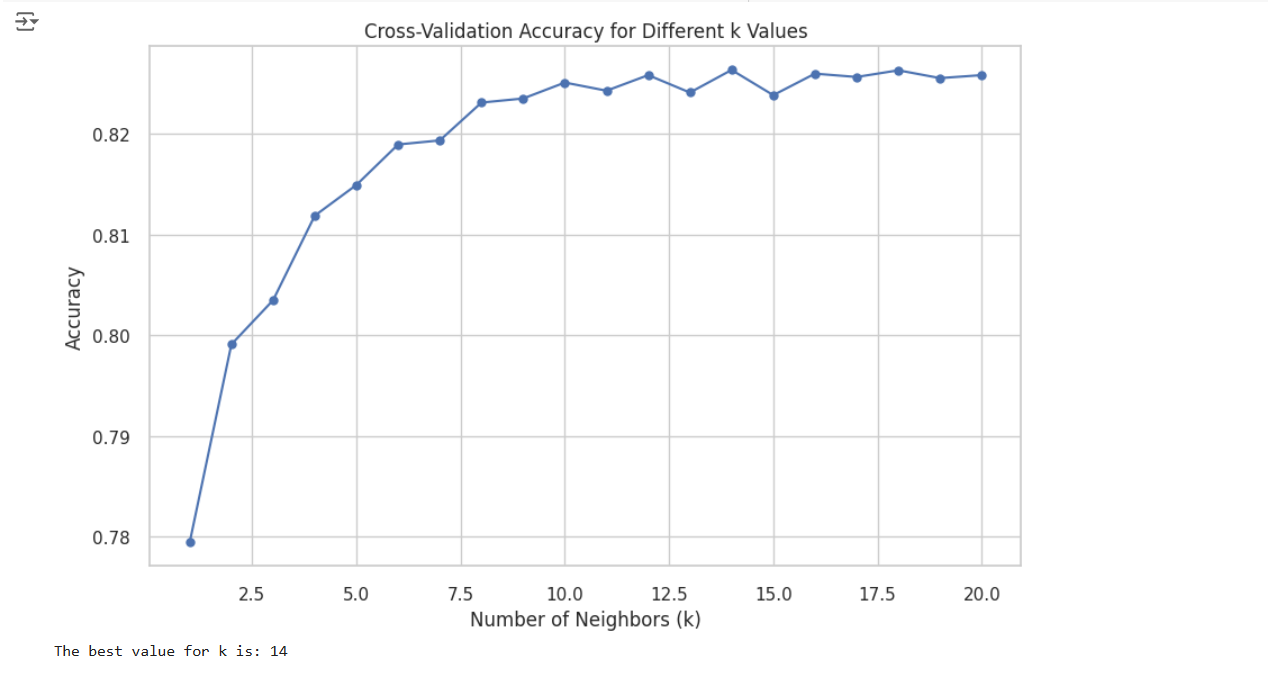
**4471(top left):** True Negatives (TN) → where the actual class is 0, and the model correctly predicted. **510(top right):** False Positives (FP) → where the actual class is 0, but the model incorrectly predicted 1.

**711(bottom left):** False Negatives (FN) → where the actual class is 1, but the model incorrectly predicted 0.

**1098(bottom right):** True Positives (TP) → where the actual class is 1, and the model correctly predicted 1.

True Negatives have the highest value (4471), which shows the model is really good at predicting class 0 that the salary category is <=50k.

* **For best K value:**

****