Detecting Adult Income

Prepared By: Erum Akmal

1. **Problem Statement for Project:**
   In this project, I have predicted the adult income using KNN classification. Adult income was predicted using multiple independent variables.

2. **Data Sources and Description:**
   For data sources I used Kaggle.

   For this project, I worked on a adult income project using KNN classification. My data set included adult income (dependent variable) and age, work class, education, marital status, occupation, race, gender, capital gain, capital loss, hours per week and native-country as independent variable.

3. **Data Mining Operations:**
   -Data wrangling:
   After loading and describing the data set, I proceeded with preprocessing of data. Categorical columns were converted into numbers. Data normalization was done next using min max scaler. The correlation matrix was calculated after that. In the next step, splitting of data was done. Data was partitioned into 80% training and 20% testing.

   Model was trained using KNeighbors Classifiers. Confusion matrix and accuracy were calculated. Model evaluation was done, and classification report was printed. Relevant graphs were generated at the end.

   -Libraries Used: I have used multiple libraries in this project including pandas, NumPy, matplotlib, seaborn, scikit-learn.

   -Data Modeling: I have performed the analysis below in my project:

   - Preprocessing of data
   - Conversion of categorical columns
   - Data normalization using Min Max Scaler
   - Correlation Matrix
   - Splitting dataset into 80-20
   - Modelling with KNeighbors Classifiers.
   - Confusion Matrix
   - K Values and error rates
   - ROC Curve
   - Heatmap
   - Histogram of Differences
   - Scatter Plot

4. **Model Outputs:**
   Below is my evaluation of the model performance for logarithmic:

```
Correlation Coefficient Matrix
age                                0.230369
fnlwgt                            -0.006339
educational-num                    0.332613
capital-gain                       0.223013
capital-loss                       0.147554
                                       ...
native-country_Thailand           -0.004219
native-country_Trinadad&Tobago    -0.009107
native-country_United-States       0.037978
native-country_Vietnam            -0.015542
native-country_Yugoslavia          0.005522
Name: income, Length: 105, dtype: float64
```

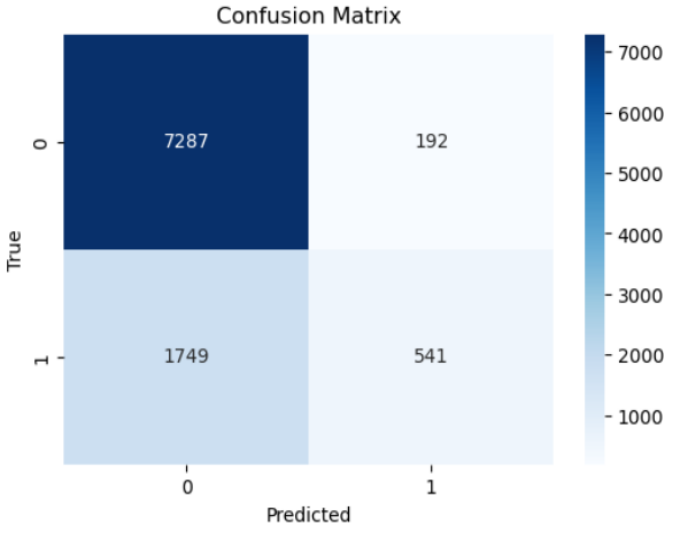
Based on correlation coefficient above, it is observed that educational-num (0.33) and age (0.23) have the strongest positive correlation with the adult income. Capital-gain also has a moderate positive correlation of 0.22.

In the above model, the weakest correlation is observed for native-**country** from -0.015 to 0.037. This means that country of origin has the least impact on adult income.

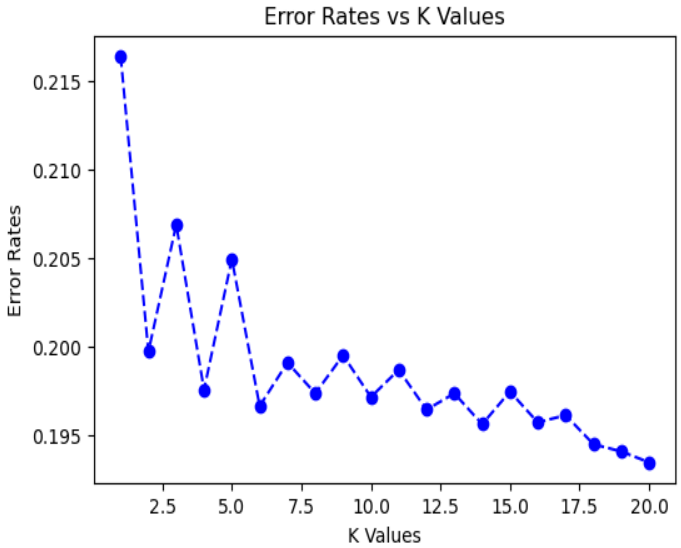
After the analysis, I can observe that model accuracy is good here, Maximum correlation is observed for

marital-status_Married-civ-spouse. Minimum correlation is observed for marital-status_Never-married.

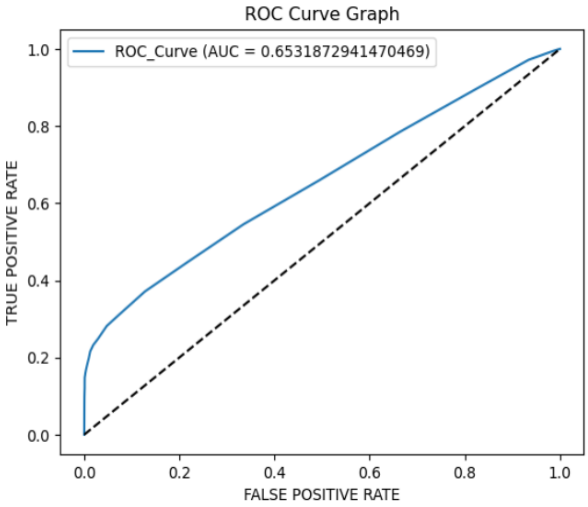
```
Confusion Matrix:
 [[7287  192]
 [1749  541]]
Accuracy:  0.8013102671716654
```



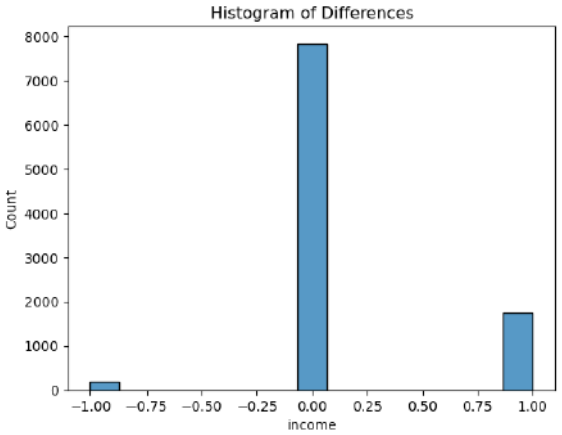
Graph for error rates and K Values was generated after that.



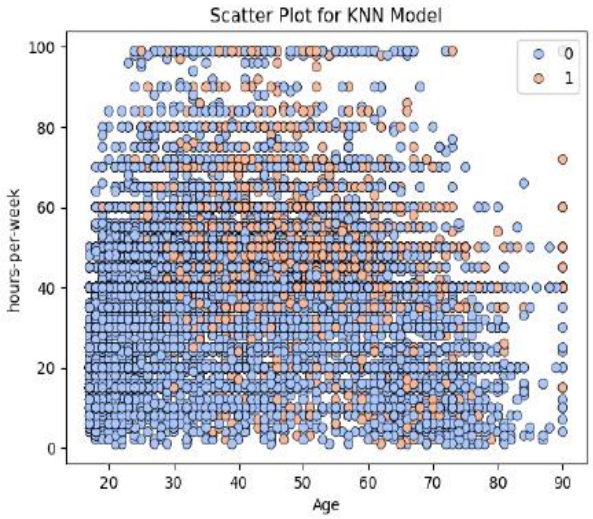
ROC Curve was plotted after that.



After that, a graph for histogram of differences was generated.



Scatter plot for KNN model was plotted at the end.



5. Limitations and Challenges:

Some of the limitations of this study are that I have only used 13 variables in this study for predicting the adult income. There could be multiple other factors that can potentially affect adult income. This model worked well but there might be multiple other reasons that can affect adult income which were not covered in the data set.

Interaction among other variables could have been additionally done here for better model prediction.

Model performance is good over here which can be observed in ROC Curve blue line in graph shown above. Overall this model is performing good since accuracy is 0.80.

6. Conclusion:

I was able to solve the problem partially. By adding more data and variables in this study, a better model could have been predicted for adult income. By adding a greater number of features and rows, we can test multiple factors that can play a role in improving adult income.