Adult Income Prediction Project

LaChandra Ash

Topic

The Census Bureau discovered that reported wage and salary income is equivalent to independent estimates of aggregate income, and that wage and salary income is reported far better than income from other sources (United States Census Bureau, n.d.). In many large-scale surveys and initiatives, the Census Bureau collects information on household income (United States Census Bureau, n.d.).

Business Problem

The prediction goal was created to determine if an individual earned less than, equal to, or more than $50K per year.

Background/History

Since 1790, the United States has accumulated population statistics, and it continues to do so every decade (United States Census Income, n.d.). The census records from 1790 through 1940 are now accessible to the public at the National Archives (United States Census Income, n.d.). The federal government has not been the only one to conduct a census; numerous states have also done so (United States Census Income, n.d.). The colonial era is reflected in the construction of several of them. Some of these documents may be discovered at the National Archives.

Data Explanation

Ronny Kohavi and Barry Becker collected these numbers from the U.S. Census Bureau's archives in 1994 (UCI, n.d.). Following the constraints ((AAGE>16) && (AGI>100) && (AFNLWGT>1) && (HRSWK>0), a collection of tidy files was recovered (UCI, n.d.). I

obtained the adult income prediction dataset from the Kaggle website. The dataset characteristics are multivariate. The attribute characteristics are integer and categorical (UCI, n.d.).

The classification is the associated type of task for this project. The dataset has lost values that were replaced with a question mark. There are 32,561 records in the collection, and the 15 columns cover a wide range of individual characteristics (UCI, n.d.). The aim of the prediction is to establish if an individual earns more than $50,000 annually (UCI, n.d.). The dataset has fifteen data attributes:

- **Age**: (17-90)
- **Work class**: Self-employed not incorporated, private, self-employed incorporated, local government, federal government, state government, never worked, and without pay • **Fnlwgt**: Final weight of the number of individuals who participated in the census Survey.
- **Education**: Doctorate, Preschool, Bachelors, Some-college, 11th, HS-grad, Assocacdm, 5 th -6 th, Prof-school, Masters, 12th, Assoc-voc, 9th, 1st -4 th
- **Education-num**: Numerical and continuous
- **Marital Status**: Widowed, Divorced, Never-married, Married-civ-spouse, Married-AFspouse, Separated, Married-spouse-absent
- **Occupation**: Machine-op-inspct, Farming-Fishing, Armed Forces, Priv-house-serv, Admclerical, Transport-moving, Tech-support, Other-service, Craft-repair, Sales, Profspecialty, Exec-managerial, Handlers-cleaners, Protective-serv
- **Relationship**: Own-child, Wife, Husband, Other-relative, Not-in-family, Unmarried

• **Race**: Other, White, Asian-Pac-Islander, Black, Amer-Indian-Eskimo

• **Sex**: Male, Female

• **Capital-Gain**: continuous

• **Capital-loss**: continuous

• **Hours-per-week**: 1-99 hours per week.

• **Native Country**: England, United States, Cambodia, Canada, Puerto Rico, Germany, Japan, India, South, Greece, Cuba, China, Iran, Outlying-US(Guam-USVI-etc.), Philippines, Poland, Vietnam, Portugal, Honduras, Italy, Jamaica, Mexico, Ireland, Dominican-Republic, Taiwan, Columbia, France, Laos, Haiti, Hungary, Nicaragua, Thailand, El-Salvador, Guatemala, Scotland, Yugoslavia, Peru, Holand Netherlands, Trinidad & Tobago, and Hong

• **Income**: <>= 50k per year (UCI, n.d.).

Methods:

The target variable for the project is the income feature of the dataset. I imported the necessary libraries and modules to conduct this project including seaborn, pandas, matplotlib.pyplot, numpy, warnings, and sklearn. I imported the adult.csv into the Jupyter Python Notebook 3. I started the preprocessing of the data by reading the dataset into a pandas dataframe; then I cleansed the data with isna method, fillna method, dropna method, and isnull method. I dropped the education.num column because it was not needed for my analysis.

The categories in the education column are string and were more vital to keep versus the education.num column. The dataset had question marks within it, and I replaced them with NaNs. I explored the data by checking the dataframe's index, value_counts, description, shape, sum, corr, and cov.

View Appendix, Table 1: Countplots

The marital status and income countplot are a comparison of the marital status of each adult and their income. The married-civ-spouse category had the highest number of adults with income

that is lower or equal to $50K per year. The married-civ-spouse category had the highest number of adults with incomes that was greater than $50K. The divorced adults had more adults earning income less than or equal to $50K per year versus divorced adults who earned more income than $50K per year.

The workclass and income countplot is a comparison of the adults within the diverse types of workclasses, and the amount of income that the adult made each year. The private category has the highest number of adults who earned income that was greater than $50k per year. The private category has the highest number of adults who earned income that was less than or equal to $50K per year. The aself-emp-not-inc category has the second highest number of adults who earned income that was less than, greater, or equal to $50K per year. The without-pay and neverworked categories has adults who did not earn income per year. The state-gov, federal-gov, and self-emp-inc had less adults who earned income versus the private category and self-emp-notinc.

The race and income countplot displayed the other and amer-indian-eskimo category had adults who did not earn income greater than $50,000 per year. The white category had the highest number of adults who earned income that was less than or equal to $50K, and greater than $50.000 per year. The black category had the second highest number of adults who earned income that was lower than or equal to $50K per year, and income that was greater than $50,000 income year. The asian-pacific-islander category had the third highest number of adults who earned income that was greater than or equal to $50K per year, and income greater than $50,000 per year.

The education and sex countplot displayed in every education category that there were more males who earned degrees versus the females. The HS-grad category had the highest number of males who graduated, as well as females. Males are most likely to obtain Bachelor's degrees, Associates degrees, Master's degrees, and Doctorates versus females.

The occupation and income countplot revealed the adm-clerical, craft, and other-service occupations had the highest count of adults who received income that was less than or equal to $50K per year. The exec-managerial, prof-speciality, craft-repair, and sales occupations had the highest number of adults who earned income that was greater than $50K per year.

The prof-speciality, other-service, adm-clerical, craft-repair, and sales occupations had the highest number of adults with income that was less than or equal to $50K per year. The farming fish, protective-serv, tech-support, armed-forces, and priv-house-serv occupations had the lowest number of adults who earned income that was less than, equal to, or greater than $50K.

View Appendix, Table 2: Pairplot

The dataframe pairplot displayed various plots of the datafram's features.

View Appendix, Table 3: Scatterplot

The age and income gained by adults in each workclass scatterplot revealed that the adults of all ages in the private, state-gov, federal-gov, self-emp-not-inc, self-emp- inc, and local-gov, earned income each year. The adults of all ages in the without-pay and never-worked categories did not earn income greater than $50K per year. The adults between the ages of thirty and eighty years

old within the self-emp-inc had the highest amount of income that was greater than $50K per year.

The adults between the ages of teenager to approximately thirty-five years old within each workclass category, earned income that was mostly less than or equal to $50K per year. Approximately by the age of sixty, the adults in the local-gov, self-emp-not-inc, state-gov, without-pay, and federal-gov are earning an income that is less than or equal to $50K per year.

View Appendix, Table 4: Boxplot

Females who are between the ages of twenty and fifty earned income that was less than or equal to $50K per year. Females who were between the ages of thirty-five years old and fifty years old earned income that was greater than $50K per year. The males who earned income that was less than or equal to $50K per year, are between the ages of twenty and fifty years old.

The males who earned income that was greater than $50K per year, are in between the ages of thirty-five and fifty-five. The females who are thirty-five years old are the youngest females to earn income that is greater than $50K per year. The males who are approximately forty years old, are the youngest males who earn income that is greater than $50K per year.

View Appendix, Table 5: Heatmap

The heatmap of the dataframe revealed the correlations or lack of correlations between the features of the dataset.

Analysis

I used the label encoder to convert the labels on the columns into numerical form. I dropped the income column from the dataframe to train, test, split the dataframe. The train, test, and split methods were used to create and fit the Gaussian Naïve Bayes model. The confusion matrix, accuracy score, and classification report were used for the prediction test. The accuracy score for the adult income prediction was 80.1%.

The zero represented the income that is less than or equal to $50K. The 0 had a precision score of 0.82%, a recall score of 0.95%, and a f1-score of 0.88%. The 1 represented the income that is greater than $50K per year. The 1 had a precision score of 0.69%, a recall score of 0.33%, and a f1-score of 0.44%.

View Appendix, Table 6: Adult Census Income Prediction Accuracy Score

Conclusion

The Naive Bayes Predictor's classifier method gave strong results. Very minimal adjustment needed to be made to the model's hyperparameters. The minimal adjustments are one of the method's greatest assets. The model can be improved overtime by retesting it for better prediction accuracy. Possibly, rebalancing the dataframe can help improve the accuracy score.

Assumptions and Limitations

Despite the fact that the Naive Bayes model's premise that all predictors are independent of one another is unrealistic in practice, it nonetheless yields acceptable results in the vast majority of cases. There were some errors and missing data within the adult dataset. The dataset has some outliers within it.

Challenges/Issues

The adult income dataset is full of survey data, which could have errors and bias information. There was a slight issue in choosing to keep both of the education attributes within the data frame. I decided to keep the education attribute and drop the education.num. The education.num was the same attribute as the education. The education.num used grade school and post high school numbers to represent the names of each education level category.

Future Uses/Additional Applications

The census income can be used for business payroll comparison for each employee, number of sales completed by each employee, calculate the amount of poverty per country, federal funding, obtain the current income data per country, and predict the average income per year in each country.

Recommendations

The model can be fine-tuned many times for better Naïve Bayes accuracy scores. One classifier can be used for this type of project, and more classifiers can present multiple accuracy results.

Implementation Plan

The model is great for businesses who want to predict their income, profits, for governments, analyze poverty, and payroll for their employees. When the predictor values are constant and are assumed to have a Gaussian distribution, this classifier is used.

Ethical Assessment

The first ethical issue is the data collected from the individuals may be inaccurate. The other ethical issue is some of the data is missing throughout the dataset. The last ethical issue is the Naive Bayes classifier algorithm may not provide accurate results for the adult census income prediction.

References

Kaggle (n.d.) Adult Census Income. Retrieved from Adult Census Income | Kaggle, on October 1, 2022.

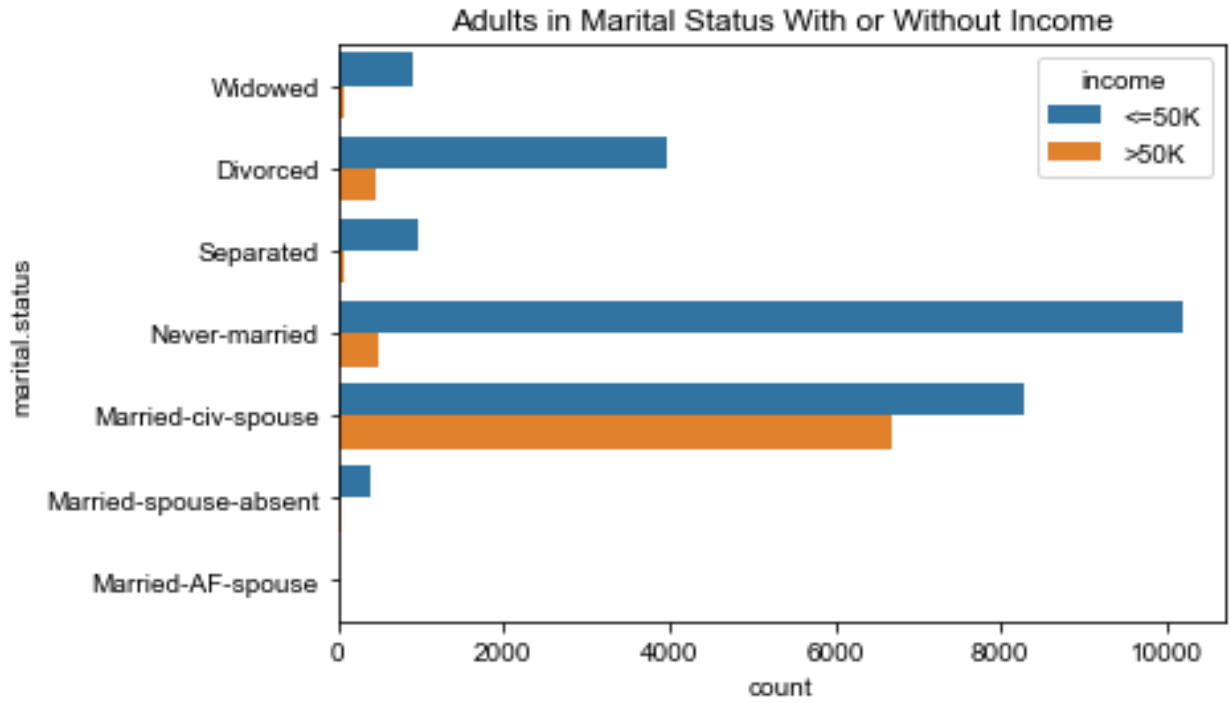
UCI (n.d.). Census Income Data Set. Retrieved from UCI Machine Learning Repository: Census Income Data Set, on October 1, 2022.

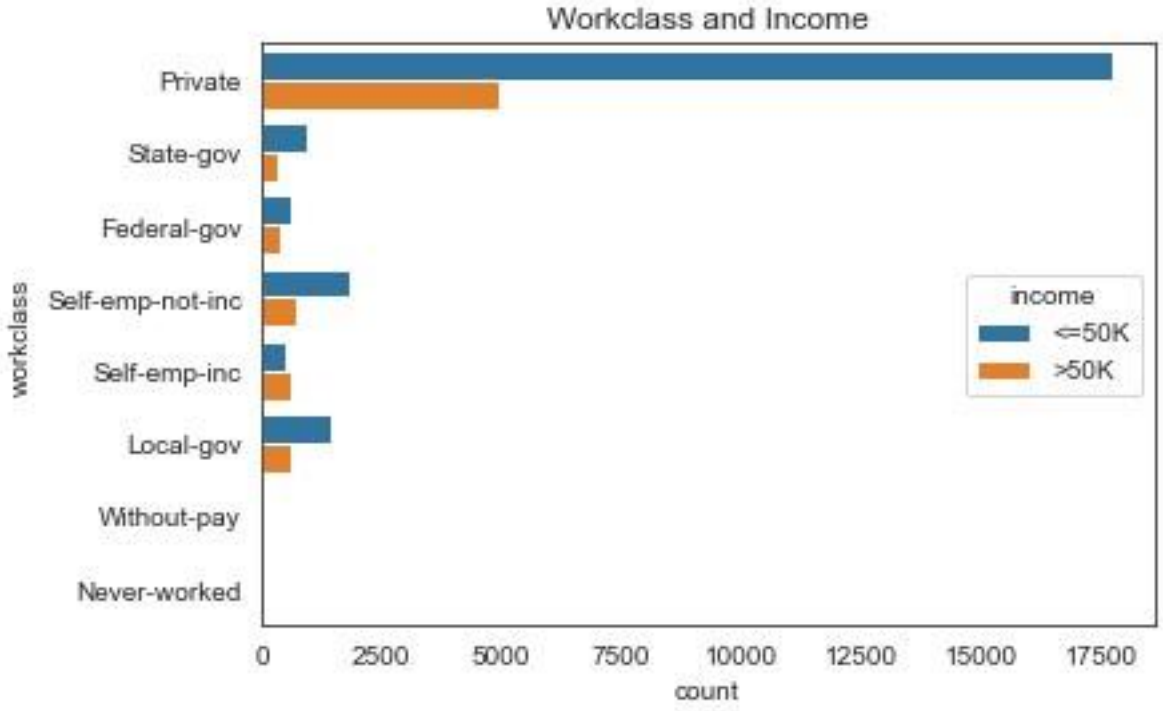
United States Census Bureau (n.d.). About Income. Retrieved from About Income (census.gov), on October 1, 2022.

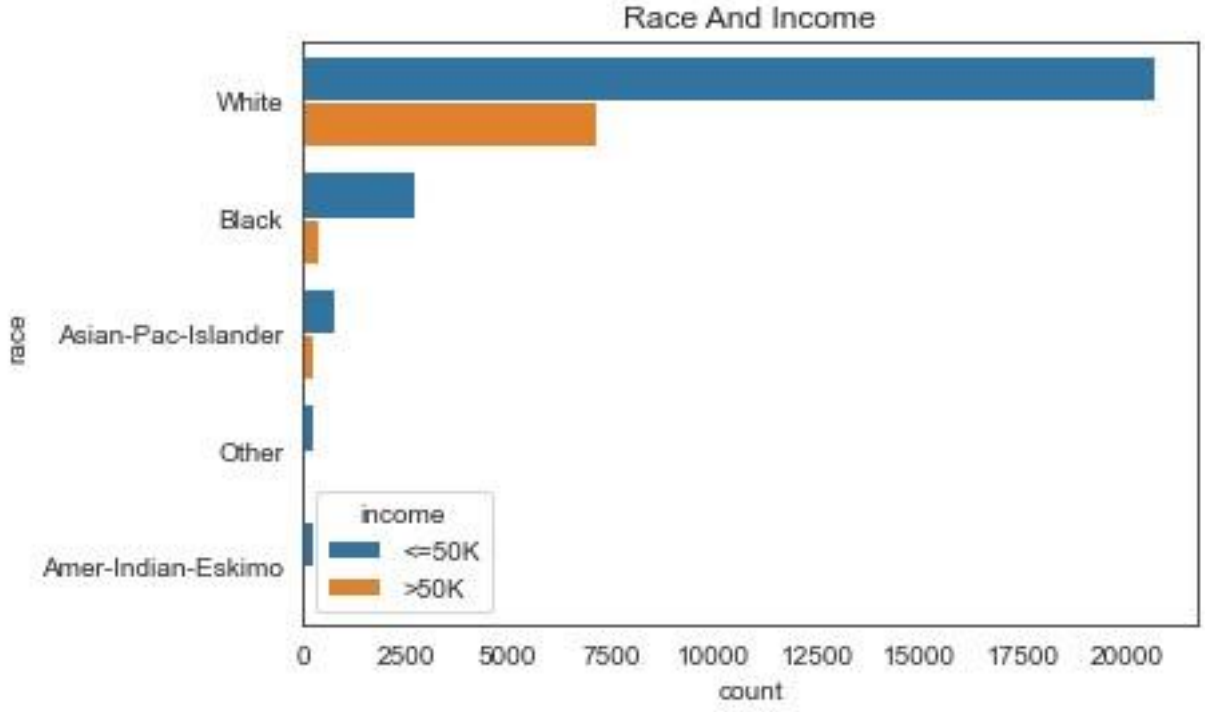
United States Census Bureau (n.d.). Decennial Census Records. Retrieved from Decennial Census Records - History - U.S. Census Bureau, on October 8, 2022.

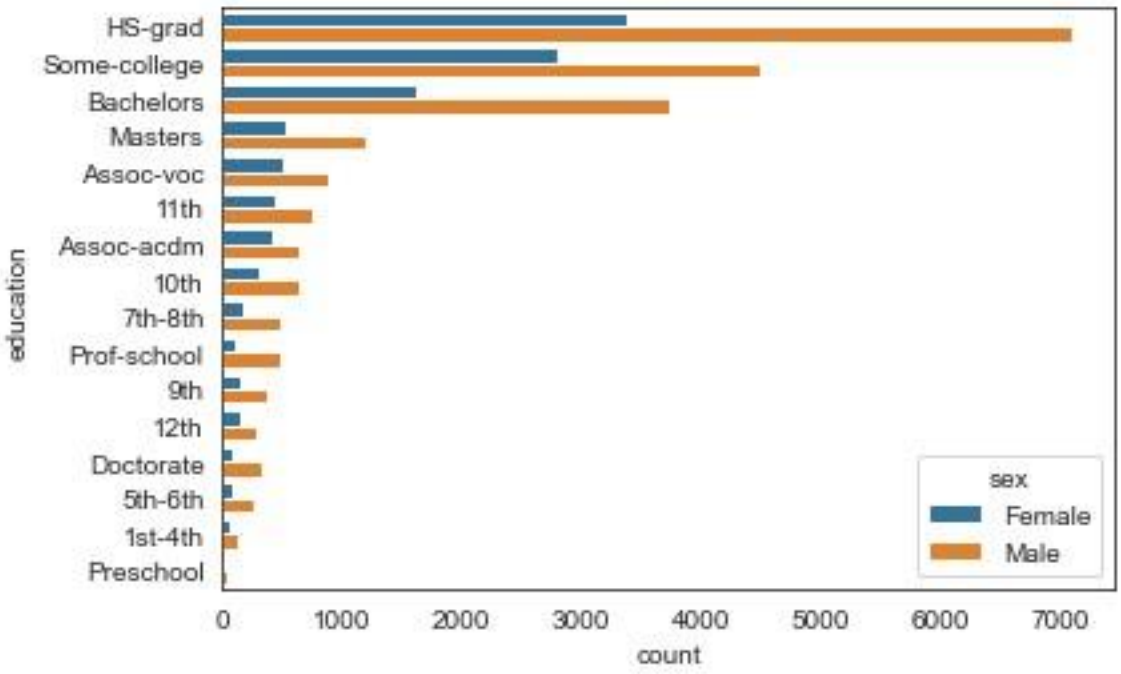
Appendix:

Table 1: Countplots






Race And Income
White
Black
Asian-Pac-Islander
Other
Amer-Indian-Eskimo
race
income
<=50K
>50K
0
2500
5000
7500
10000
12500
15000
17500
20000
count


HS-grad
Some-college
Bachelors
Masters
Assoc-voc
11th
Assoc-acdm
10th
7th-8th
Prof-school
9th
12th
Doctorate
5th-6th
1st-4th
Preschool
education
sex
Female
Male
0
1000
2000
3000
4000
5000
6000
7000
count

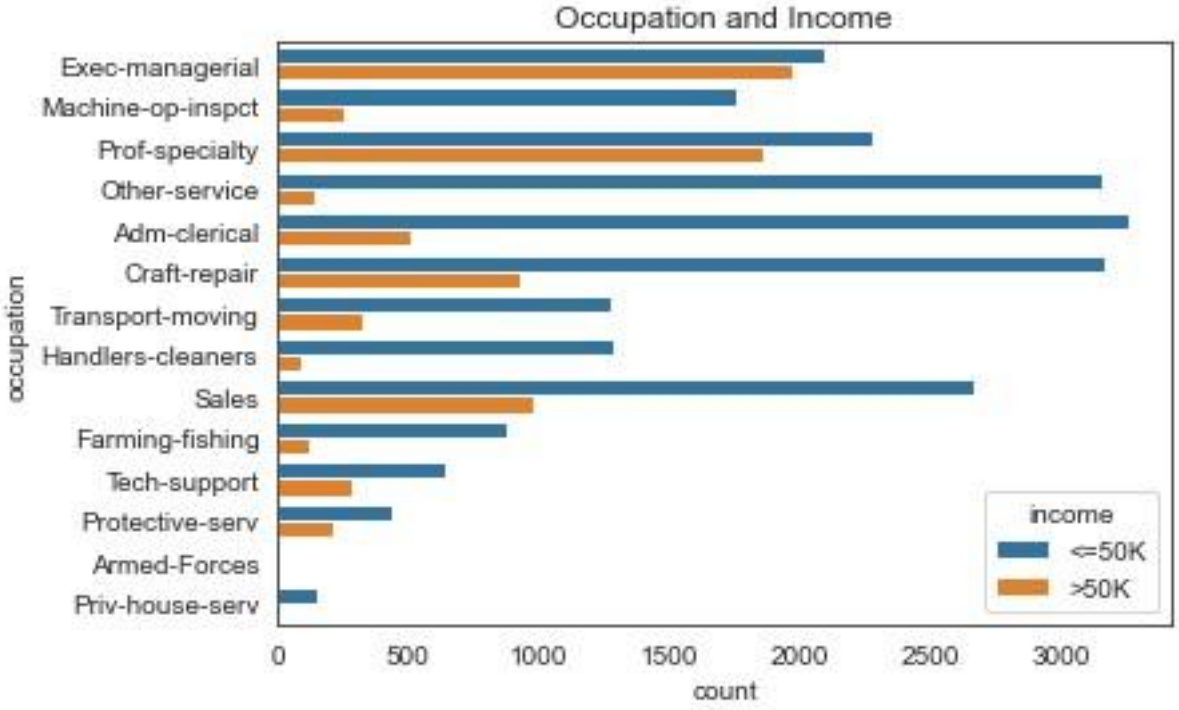

Occupation and Income
Exec-managerial
Machine-op-inspct
Prof-specialty
Other-service
Adm-clerical
Craft-repair
Transport-moving
Handlers-cleaners
Sales
Farming-fishing
Tech-support
Protective-serv
Armed-Forces
Priv-house-serv
occupation
income
<=50K
>50K
0
500
1000
1500
2000
2500
3000
count

Table 2: Pairplot

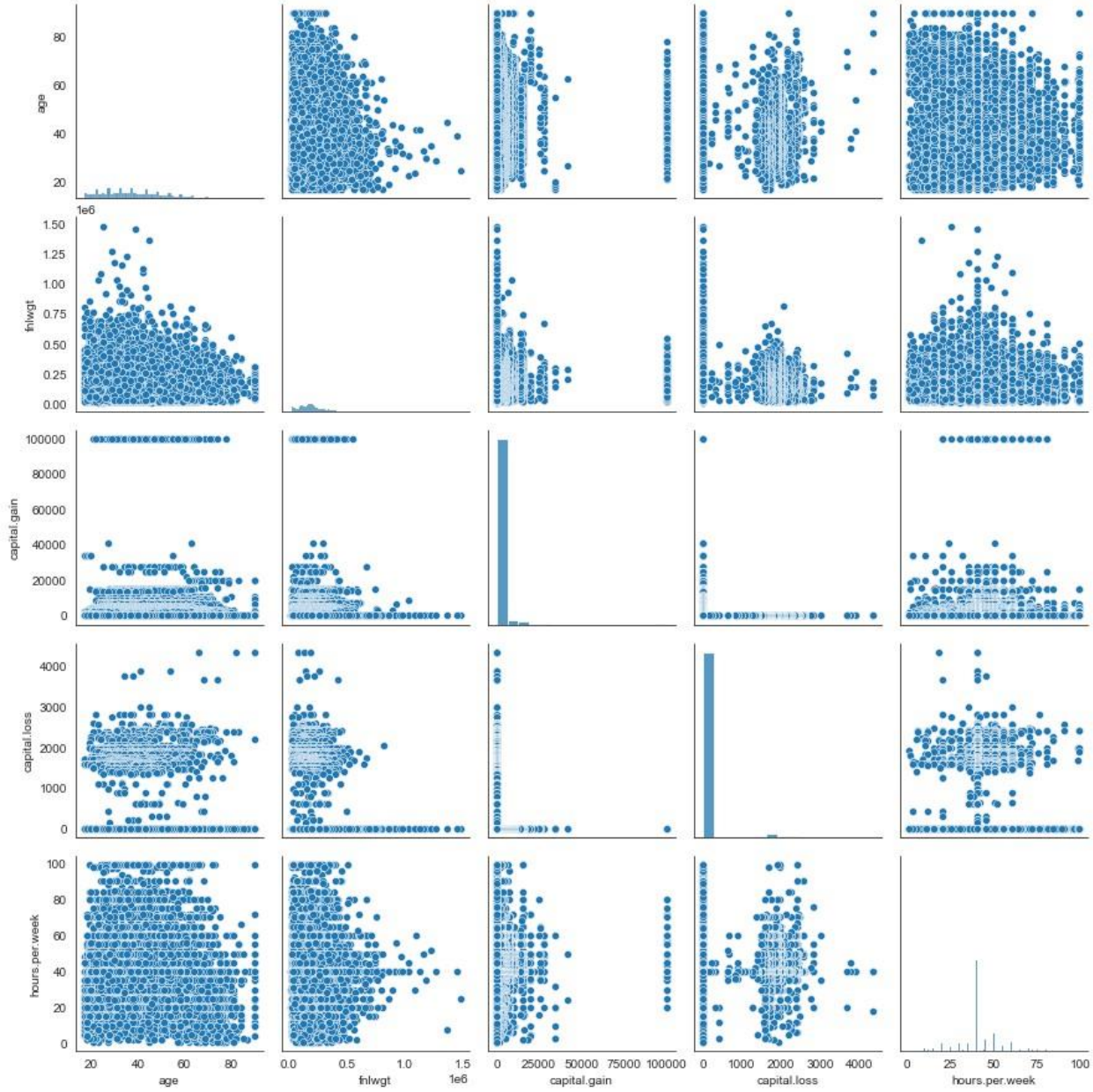


Table 3: Scatterplot

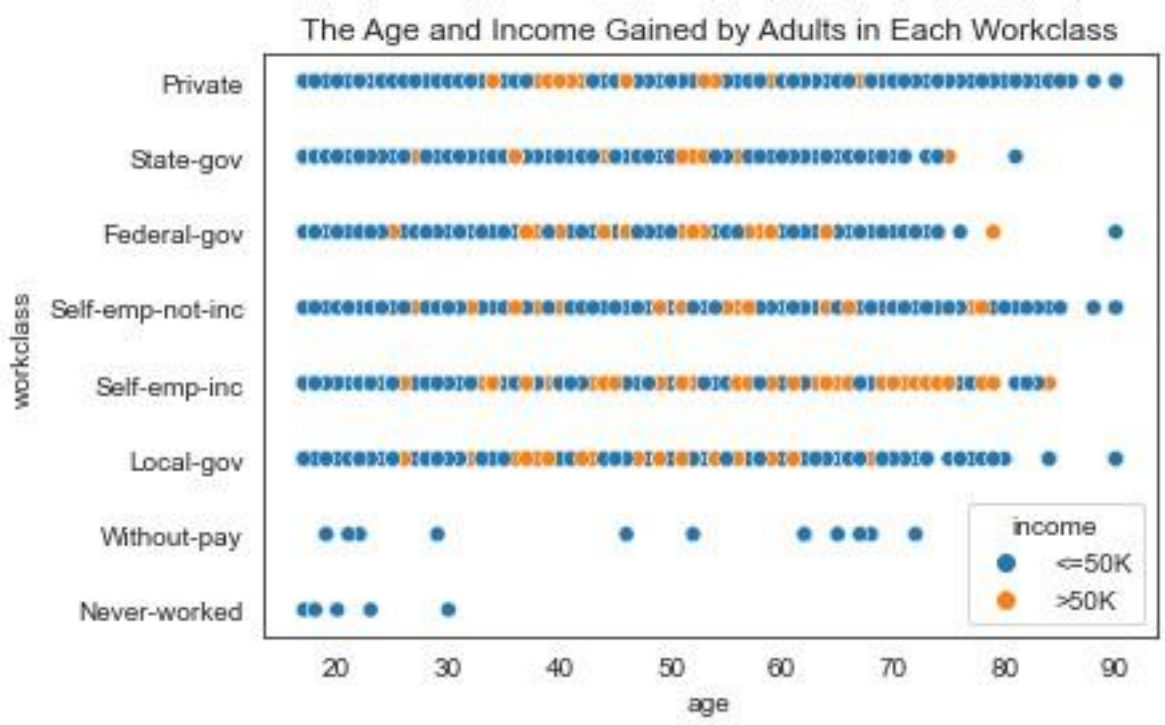


Table 4: Boxplot

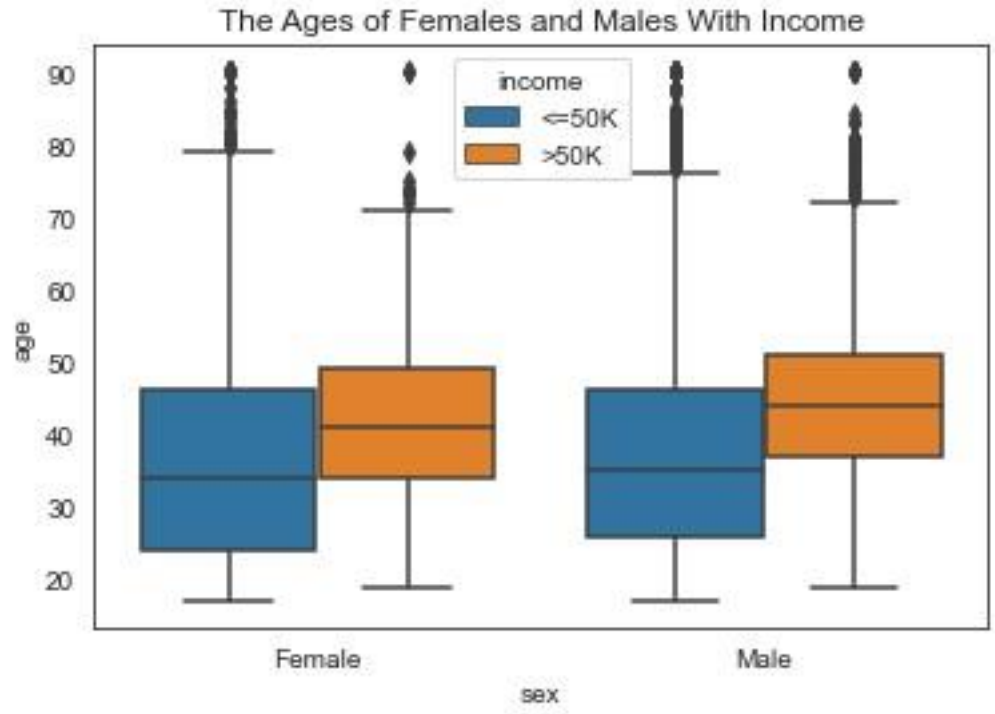


Table 5: Heatmap

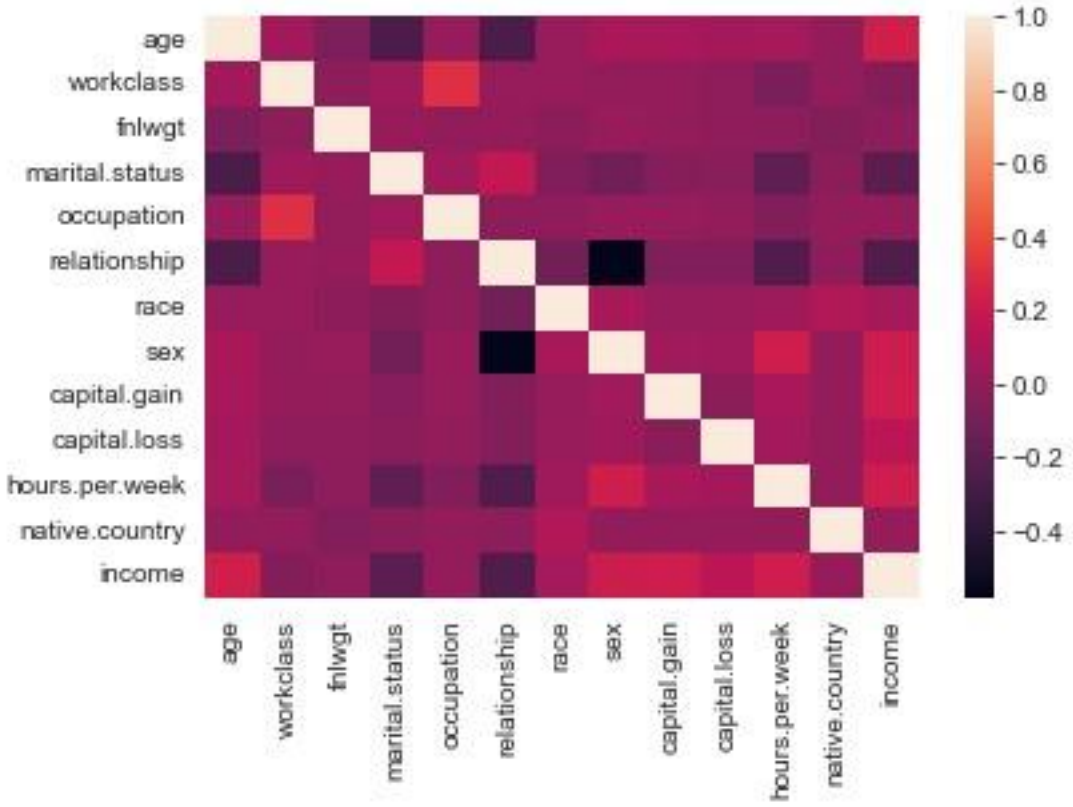


Table 6: Accuracy Metrics

```
[[4704  233]
 [1063  513]]
80.10133578995855
              precision    recall  f1-score   support

           0       0.82      0.95      0.88      4937
           1       0.69      0.33      0.44      1576

    accuracy                           0.80      6513
   macro avg       0.75      0.64      0.66      6513
weighted avg       0.78      0.80      0.77      6513
```