IncomeForecaster: A **Random Forest** Odyssey

What is the importance of estimating the income of a citizen?

An individual's annual income can be influenced by a multitude of factors, such as education level, age, gender, occupation, and more. By developing a predictive model for income, we gain valuable insights into the dynamics of these contributing factors over time. This allows for a deeper understanding of the patterns and trends that affect income levels annually.

Moreover, employers can use these predictions to make data-driven decisions related to employee compensation, workforce planning, and talent acquisition. Also, policymakers can benefit from insights into income trends to formulate effective economics policies and social programs.

01

Analysis Objectives:

The project aims to address the following **questions:**

01

Exploratory Data Analysis:

1. What is the class distribution in the target variable "Income" (<=50K and >50K)?
2. How are different demographic variables distributed in the dataset?
3. What is the average age of individuals in the dataset?
4. What is the gender ratio?

Correlations:

1. Is there any correlation between age and income?
2. How does education influence income?

02

Data Source

[https://www.kaggle.com/datasets/wenruliu/adult-income-dataset]

02

Attributes:

15

Observations:

48,842

Income - Two classes:

<= 50K

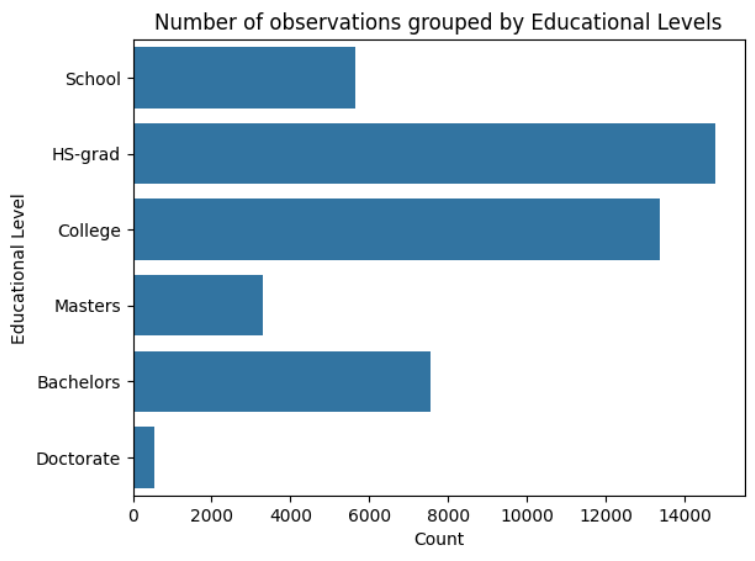
or

> 50K

03

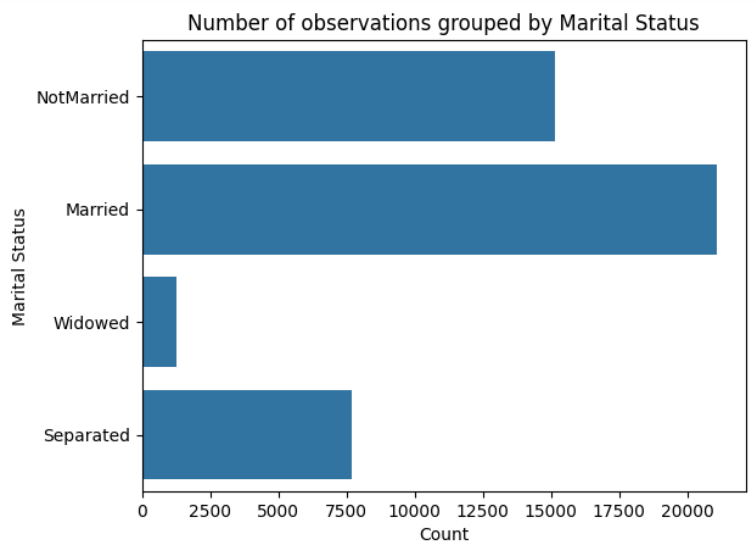
Univariate Exploratory Data Analysis

Categorical Variables

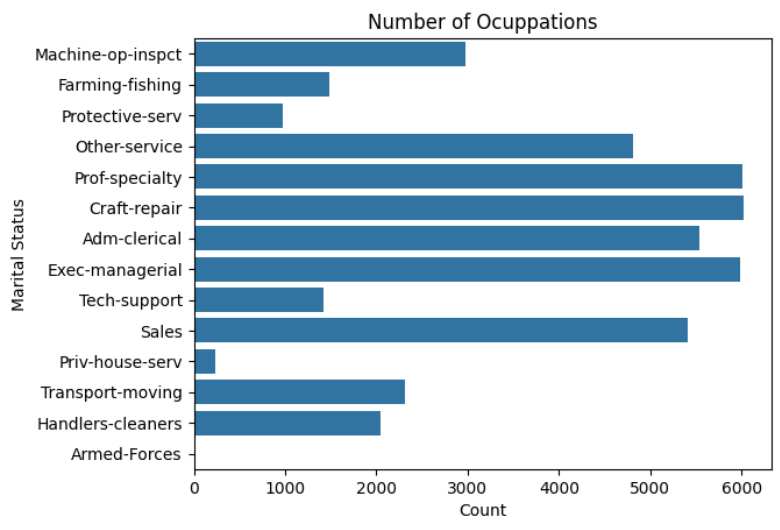


Educational Levels

Marital Status

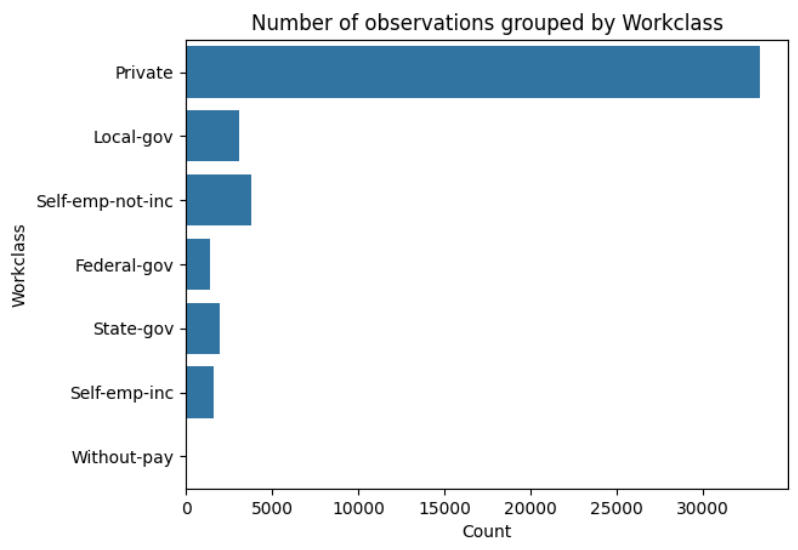


Categorical Variables



Occupations

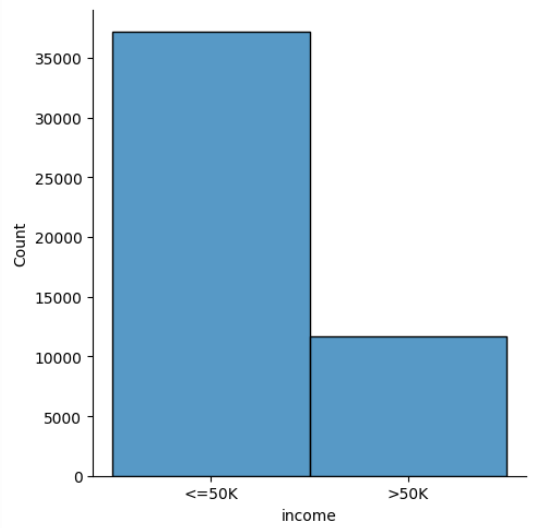
Workclass



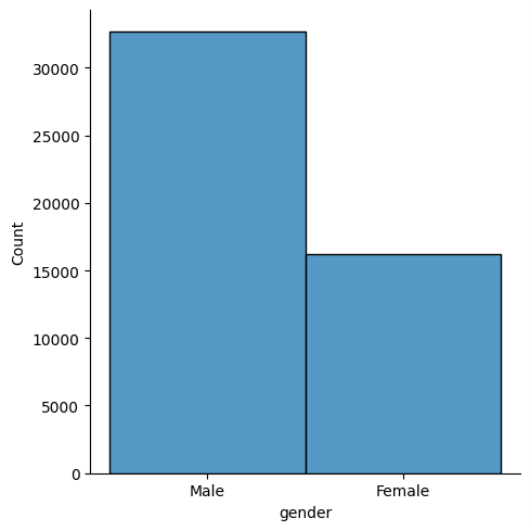
04

Bivariate Exploratory Data Analysis

Income

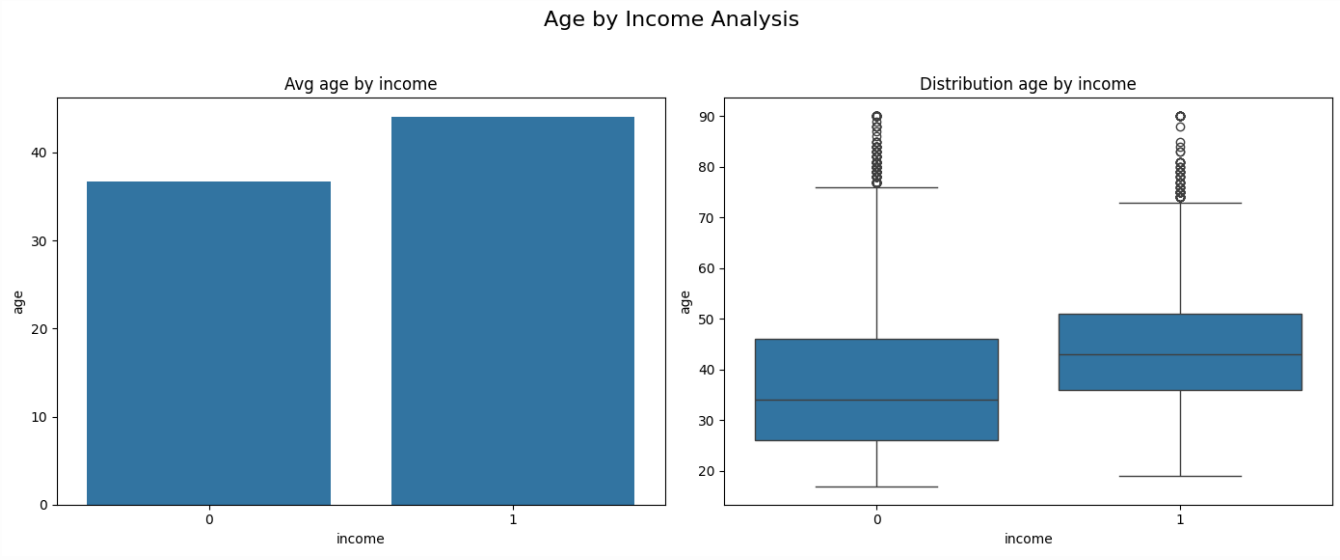


Gender



Income level is less than 50k is **3 times** of those above 50k, indicating that the dataset is skewed. Nevertheless, as there is no available data indicating the upper limit of adults' income beyond 50K, it would be premature to infer that the overall distribution of wealth is skewed toward the high-income group.

Also, the number of males in this dataset is 3 times higher than that of women.



Age by Income (average):

<= 50K: 36.7

> 50K: 44.0

Average age:

38.5

Median age:

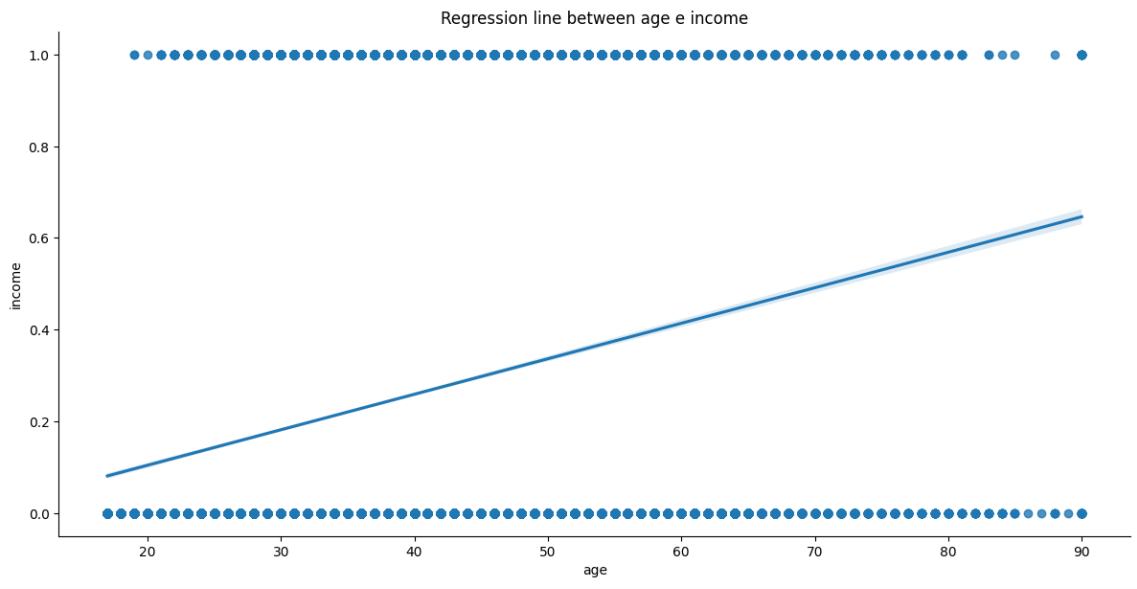
37.0

Max. age:

90.0

Min. age:

17.0



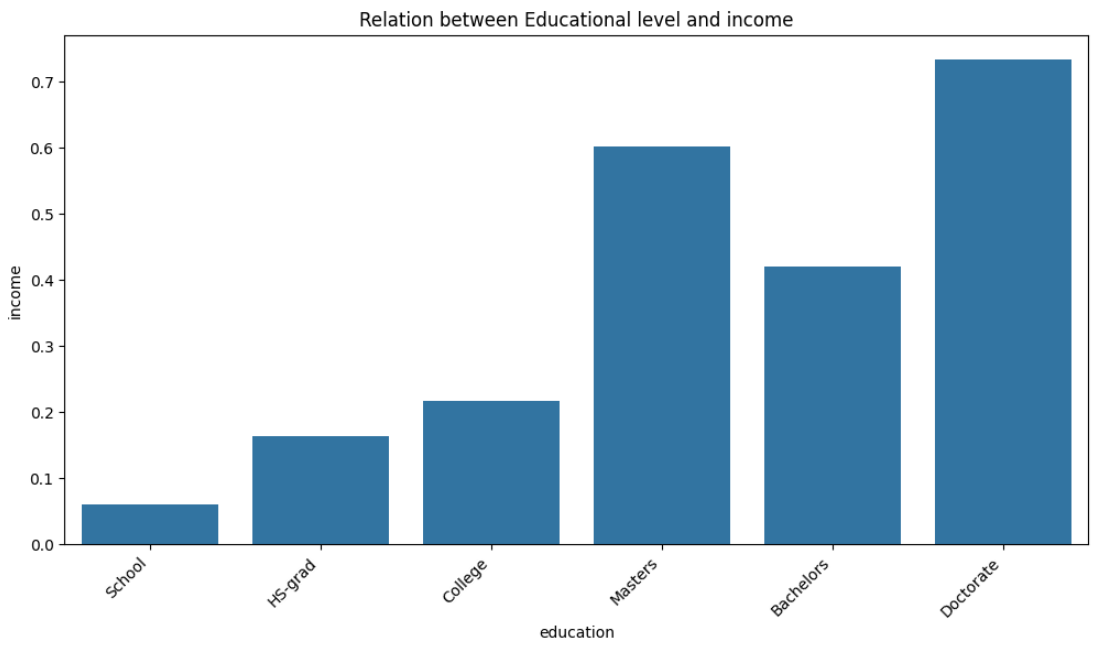
Correlation

0.23

P-Value

0.0

Statistical evidence indicates that there is a **significant association** between the age and income.



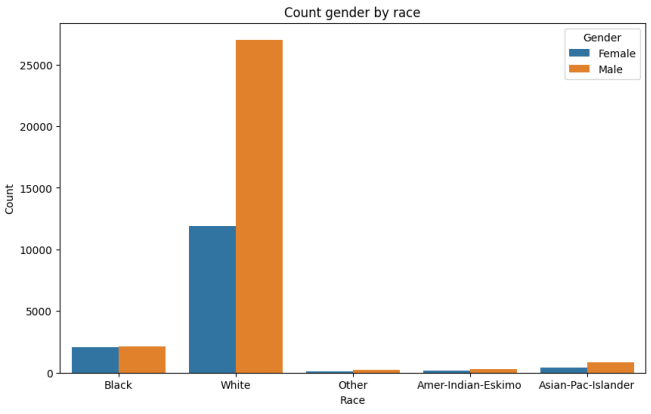
Correlation

0.33

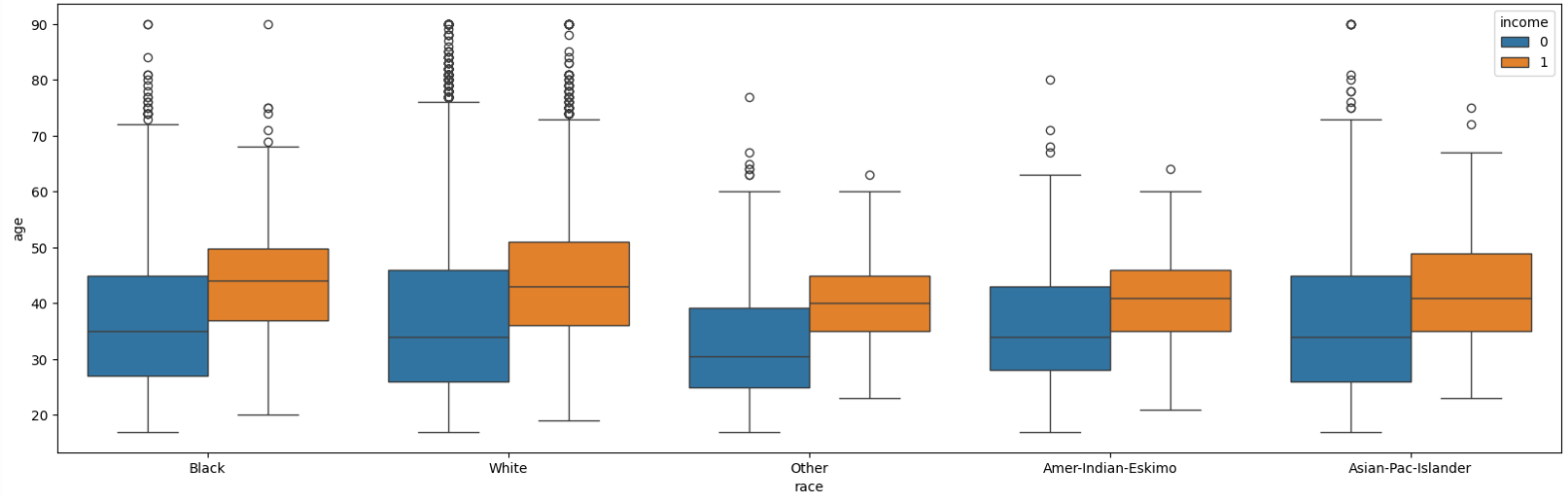
P-Value

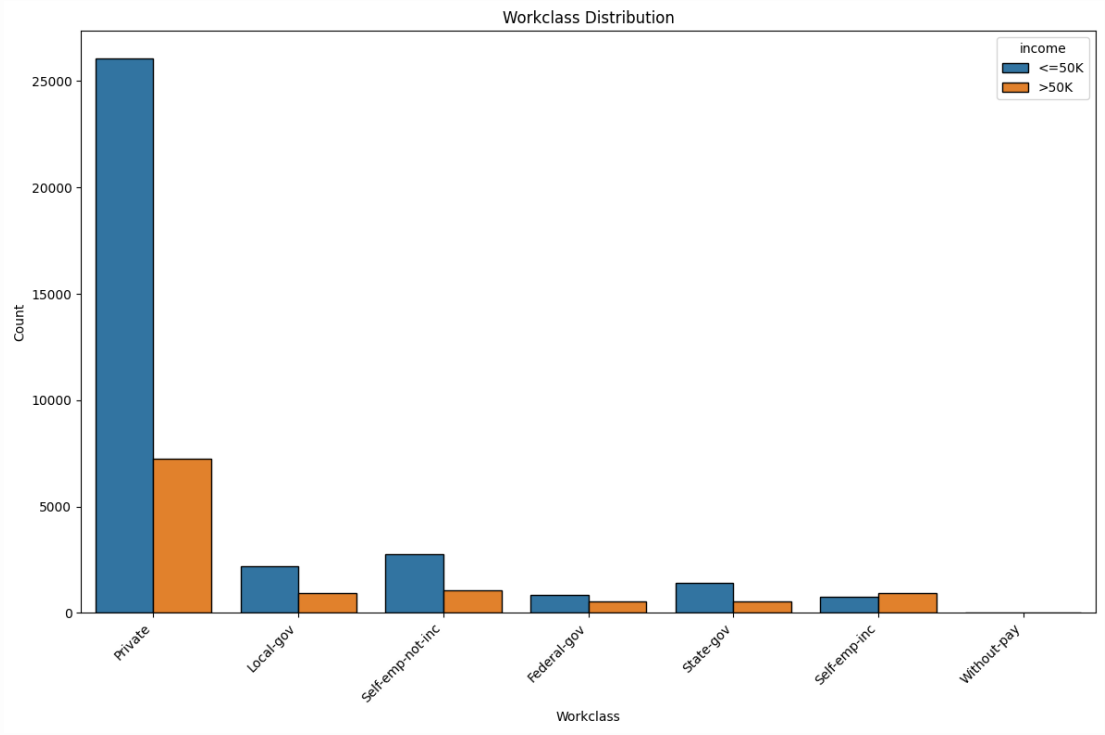
0.0

Statistical evidence indicates that there is a **significant association** between the level of education and income.

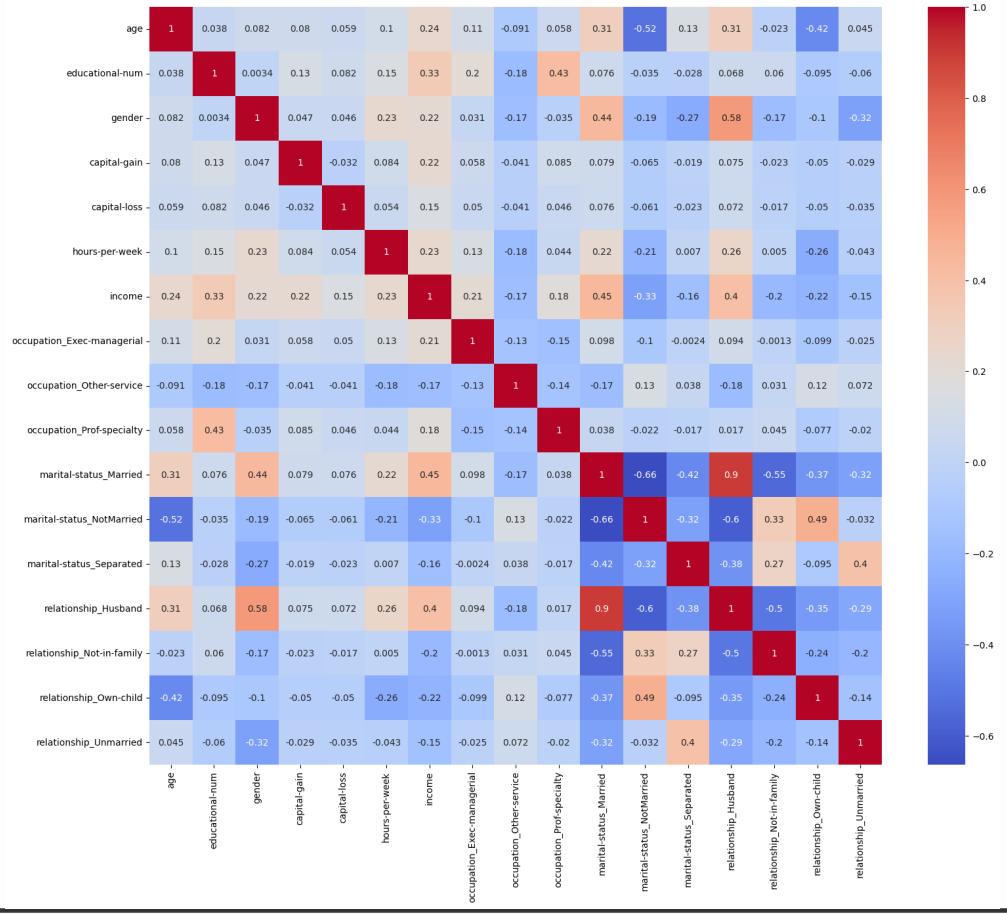


Most of people of our dataset are **white and males**. Secondly, white and females.





The **private sector** is chosen by people, both those whose earnings are less than 50 thousand and those earning more.



A few insights observed through the heatmap:

Positively correlated

1. Marital status is important to the income;
2. Relationship status – husband;
3. Year of experience;
4. High education.

Negatively correlated

1. Never married.

Correlations between features it is not the same as the features importance.

05

Random Forest Model

5.1 Random Forest Algorithm

```
train_df1, test_df1 = train_test_split(df1, test_size=0.2)
```
executed in 12ms, finished 20:49:11 2023-12-13

```
train_X = train_df1.drop('income', axis=1)
train_y = train_df1['income']

test_X = test_df1.drop('income', axis=1)
test_y = test_df1['income']
```
executed in 5ms, finished 20:49:11 2023-12-13

```
forest = RandomForestClassifier()

forest.fit(train_X, train_y)
```

```
# 84% predicting right if this person earn more or less than 50K
forest.score(test_X, test_y)
```

```
0.8465450525152017
```

Using a test size of 20% of the sample**, the Random Forest** predicts **84.6% right,** in our model, if the person earn more or less than 50K.

5.2 Features Importances

Top 5

```
{'age': 0.2304148712102715,
 'educational-num': 0.1328189797978031,
 'hours-per-week': 0.11464508460755216,
 'capital-gain': 0.10607134111652552,
 'marital-status_Married': 0.06683030195880453,
```

Bottom 5

```
 'native-country_Scotland': 8.986745462488411e-05,
 'native-country_Outlying-US(Guam-USVI-etc)': 5.237886881789851e-05,
 'occupation_Armed-Forces': 5.022752656475636e-05,
 'native-country_Honduras': 3.40520874691613e-05,
 'native-country_Holand-Netherlands': 0.0}
```

The random forest has the capability to assess the **importance** of different features. In the **top 5** positions on this list, we observe that as individuals age, attain higher levels of education, work more hours, and are married, their predicted income tends to rise accordingly. Also, check the bottom 5 of the list.

5.3 Hyperparameter Tuning

Hyperparameter tuning iscrcial step to **optimize** the performance of the model. Which involves **fiding the best combination** that maximizes the model's effectiveness. The model has improved and can predict **85.6%** better

```
0.8560530679933664
```

Moreover, **importance** of the features changed. In the **top 5** positions on this list, we observe the same features, but in different values. Also the '*capital-gain*' got more importance than '*hours-per-week*':. However, the bottom 5 did not change order, only values.

```
# they will test in combination
param_grid = {
    'n_estimators': [50, 100, 250],
    'max_depth': [5, 10, 30, None],
    'min_samples_split': [2, 4],
    'max_features': ['sqrt', 'log2']
}
grid_search = GridSearchCV(estimator=RandomForestClassifier(),
                           param_grid=param_grid, verbose=10)
```

Top 5

```
'age': 0.16126817714567646,
'educational-num': 0.13327475749541748,
'capital-gain': 0.1317896594995949,
'hours-per-week': 0.09159195088479438,
'marital-status_Married': 0.08248544292873128,
```

Best estimator

```
RandomForestClassifier(max_depth=30, max_features='sqrt', min_samples_split=4,
                       n_estimators=250)
```

06

Conclusions

Conclusions

After training with the Random Forest model, key features influencing income prediction have been identified.

Age takes the lead, representing approximately 16.1%, suggesting a significant correlation between age and income. Educational level follows closely at about 13.3%, highlighting the relevance of education for higher income levels. Capital gain, contributing around 13.2%, underscores the importance of additional gains. Hours-per-week, with a contribution of 9.2%, indicates the influence of weekly working hours on income determination. Finally, marital status (Married) holds a significance of approximately 8.2%, suggesting a positive impact on income.

Notably, the model achieves an accuracy rate of 85.6% in predictions after training.

07

Thanks!

Do you have any questions?

jonatasv@gmail.com

My portfolio:
https://jonatasv.github.io/portfolio_projetos/#

in/jonatas-vieira/