Adult income dataset

The prediction task is to determine whether a person makes over $50K a year in several countries around the world.

Mohamed Iyadh Tajouri

Table of contents

01

Business Problem, Stakeholders and what problem am I solving for them

Business Problem

The central business problem addressed by this dataset is to empower financial institutions, businesses, and researchers to make data-driven decisions by predicting whether an individual's annual income exceeds $50,000 based on a set of demographic and employment-related features.

This predictive capability can have a transformative impact on various sectors.

Stakeholders and what problem am I solving for them

Financial Services

Financial institutions can use income predictions to enhance responsible lending practices and reduce credit risk.

Marketing and Sales

Businesses can tailor their marketing strategies and product offerings to different income segments, boosting customer engagement and revenue

Socioeconomic Research

Researchers gain valuable insights into income distribution and disparities, aiding in a better understanding of societal dynamics.

Human Resources Departments

HR teams can make more informed hiring decisions and negotiate competitive compensation packages for job candidates.

Government and Policy Makers

Policymakers can leverage income predictions to inform policies that reduce income inequality and promote social welfare.

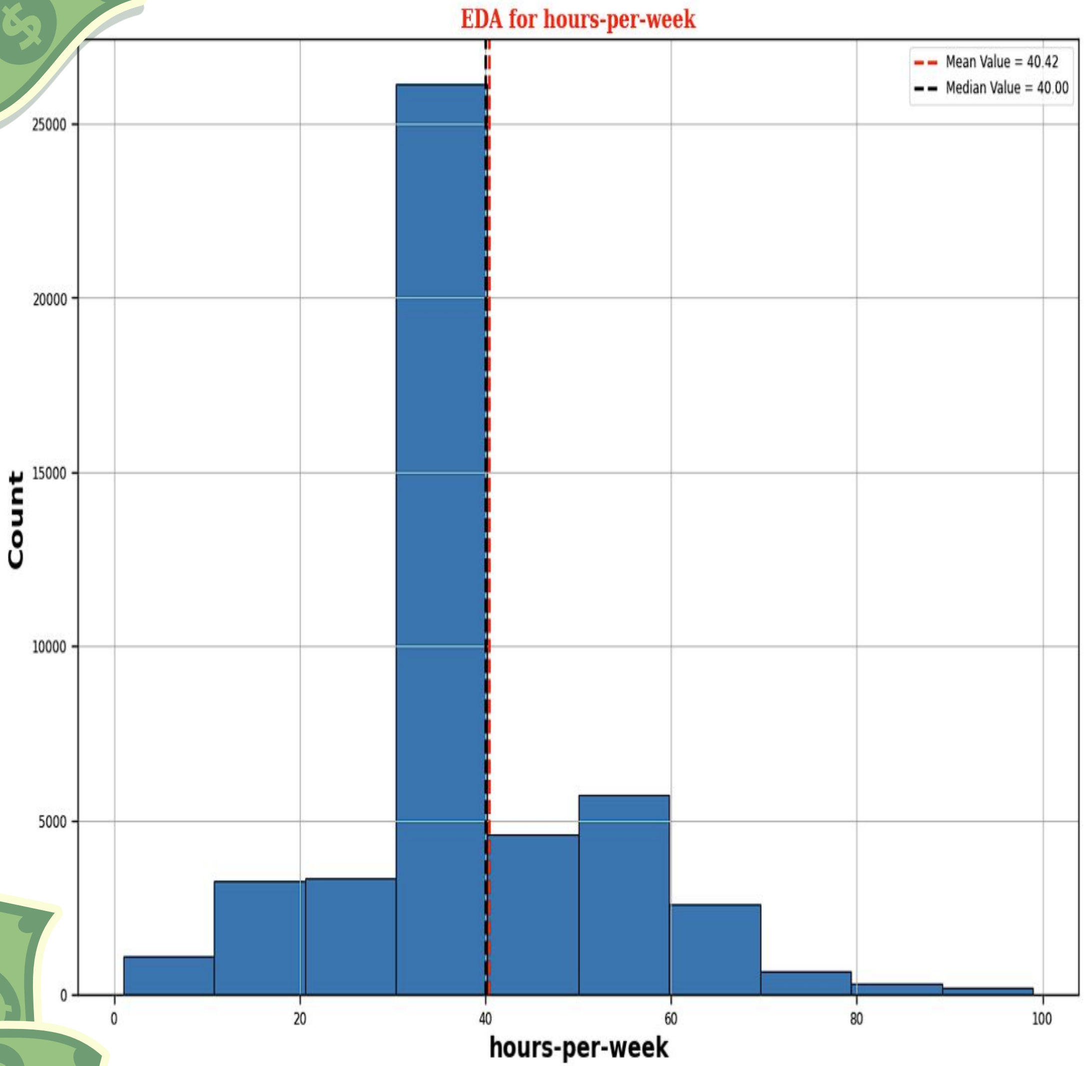
02

All about the data and the visualizations

Introduction to the dataset

- This dataset, a well-known extract from the 1994 Census bureau database by Ronny Kohavi and Barry Becker, holds great significance across multiple domains. Its importance lies in its capacity to shed light on income distribution and disparities, offering valuable insights that can drive progress in various fields.
- This dataset serves as a versatile tool that can catalyze advancements in several critical domains.
- You can download the dataset here
- This shows that we have **48842** observation and **15** attributes including target attribute(**income**).

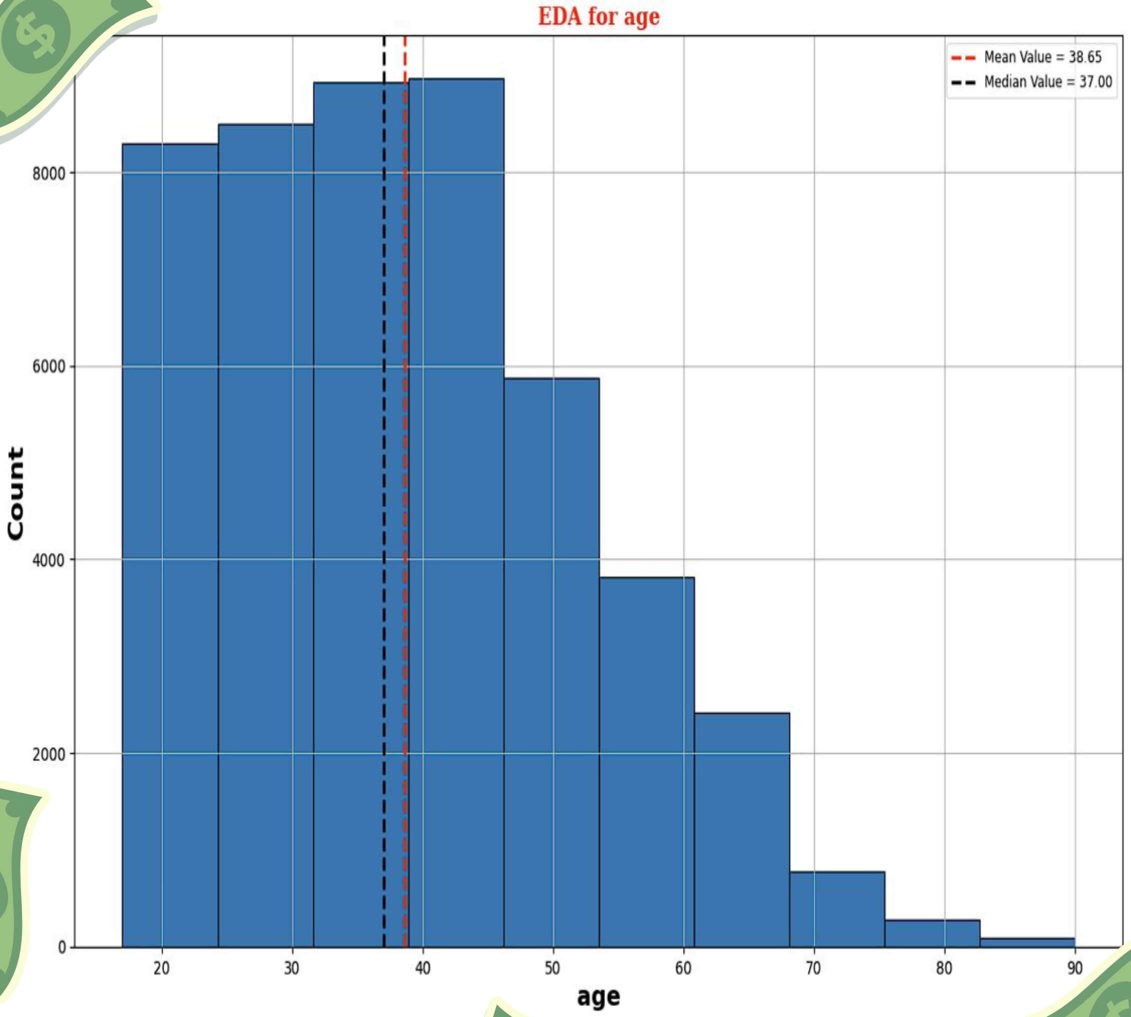
Visualizations



Description about the distribution

In this data the hours per week attribute varies within the range of 1 to 99.

Most people work 30-40 hours per week, they are roughly 27,000 people. There are also few people who works 80-99 hours per week and some less than 20 which is unusual. 75 percentage of the people spend 45 or less working hours per week.

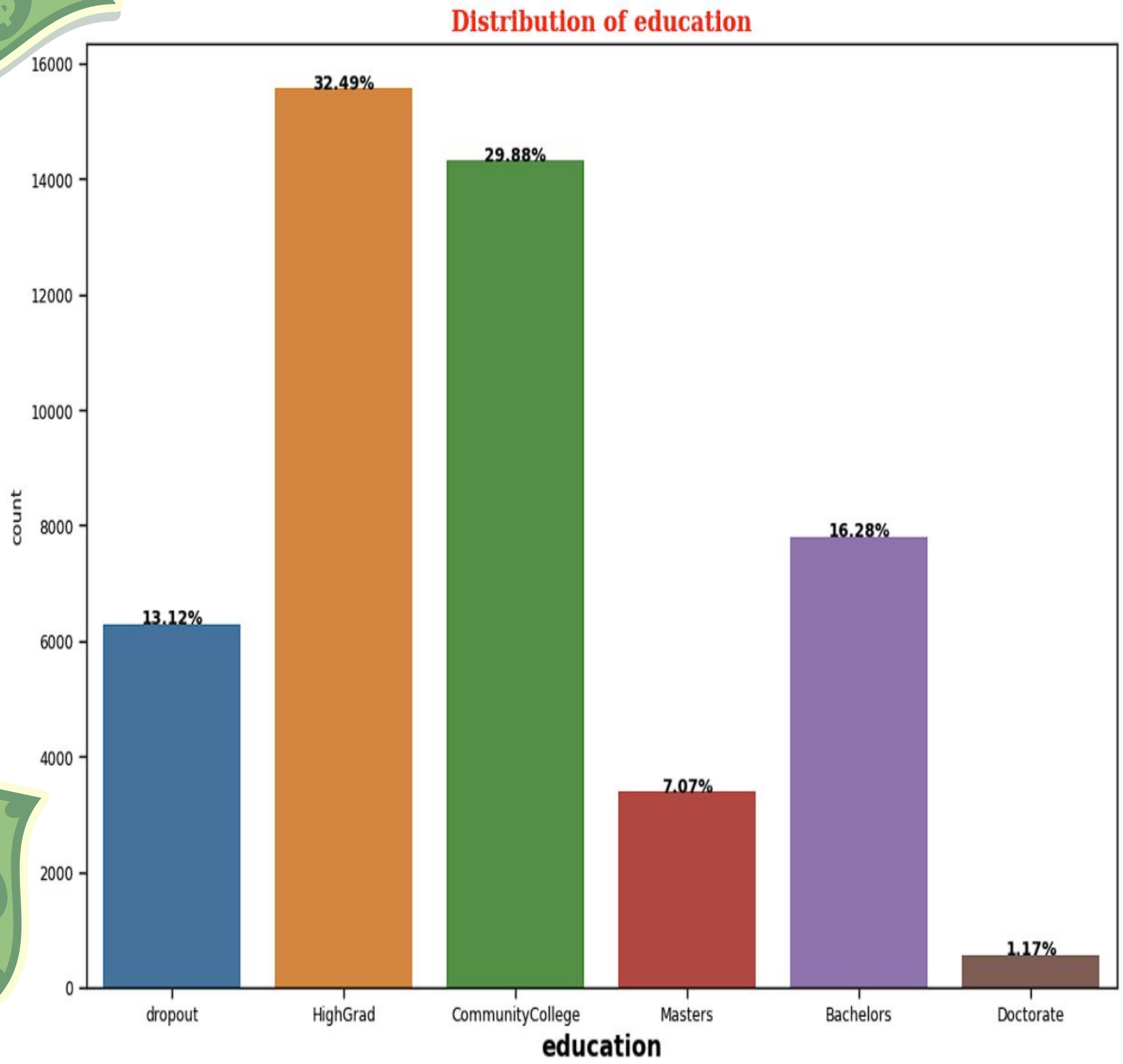


Total Ages above 70yo are: 858

Description about the distribution

"**age**" attribute is not symmetric.

it is right-skewed(But this is totally fine as younger adult earn wages not the older ones)
Minimum and Maximum age of the people is 17 and 90 respectively. The **mean age is around 38 years**
This dataset has fewer observations(868) of people's age after certain age i.e. 70 years.

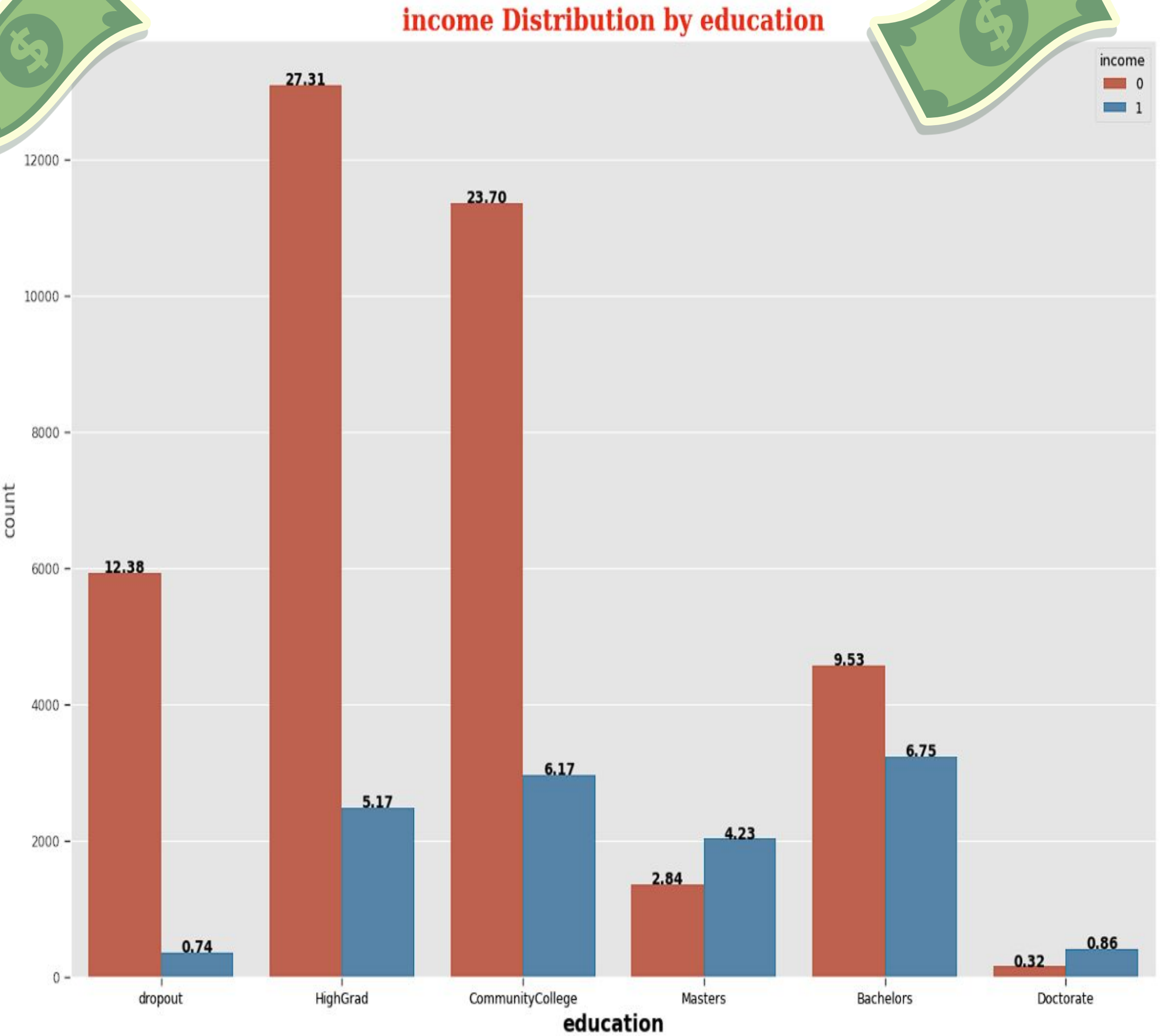


Description about the distribution

There are **6 unique categories** present in the education attribute(after modification).

HighGrad has 32.49% of all the education attribute. HighGrad (15573) has the maximum number of observations followed by **CommunityCollege** (14324) and **Bachelors(**7803).

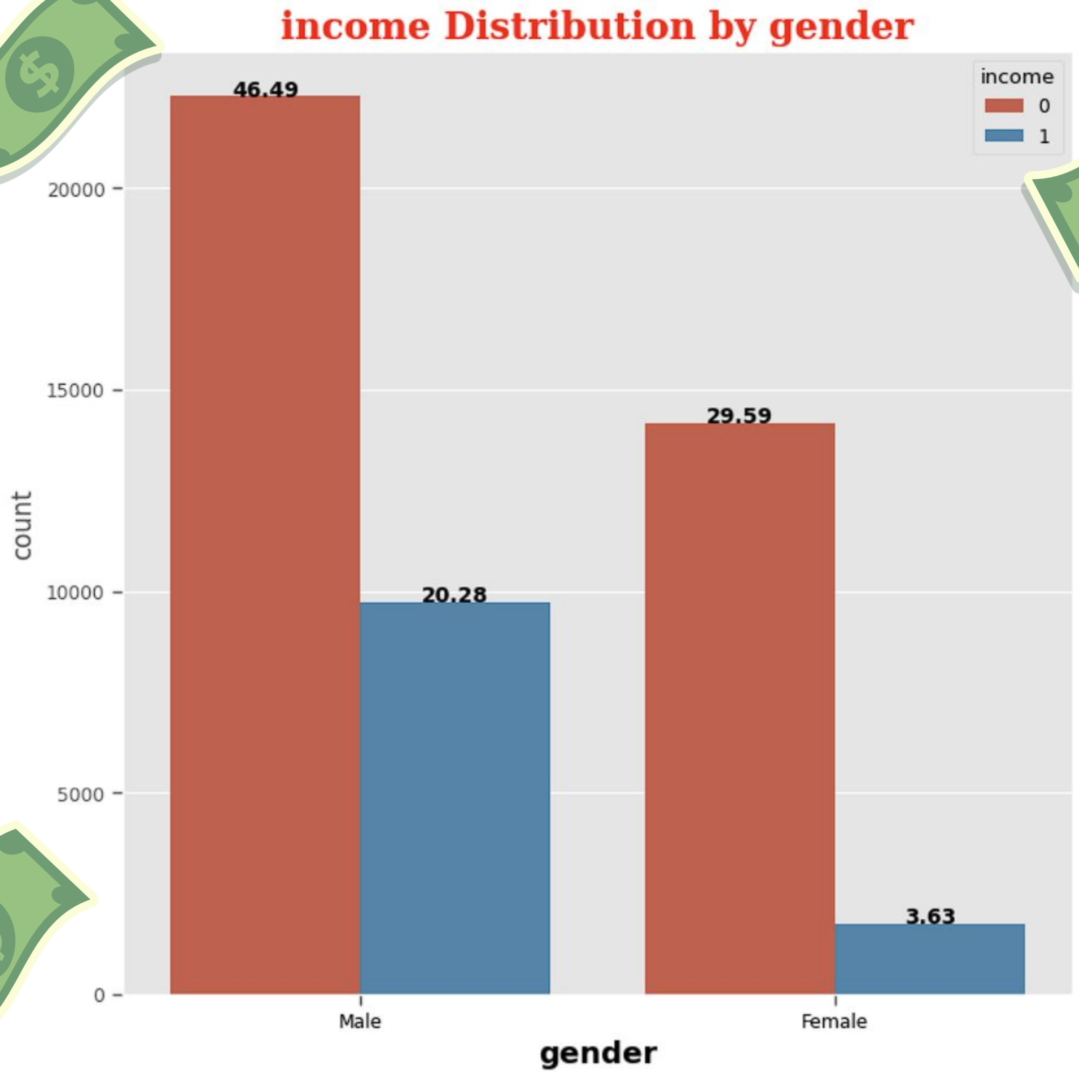
Doctorate has the minimum number with only 562 having a doctorate (1.17%).



Description about the distribution

Despite the fact that **most of the categories fall under the HighGrad** but the interesting thing is only **5.17%** of all people belong to the **income group 1(i.e. earns more than 50k)**, surprisingly less than the Bachelors which is **6.75%**. There are only few categories in "education" attribute whose percentage to fall under income group 1 is greater than the falling under income group 0. These are **masters and doctorate**.

We can also infer that higher education may provide better earnings.



Description about the distribution

This countplot explain following things:

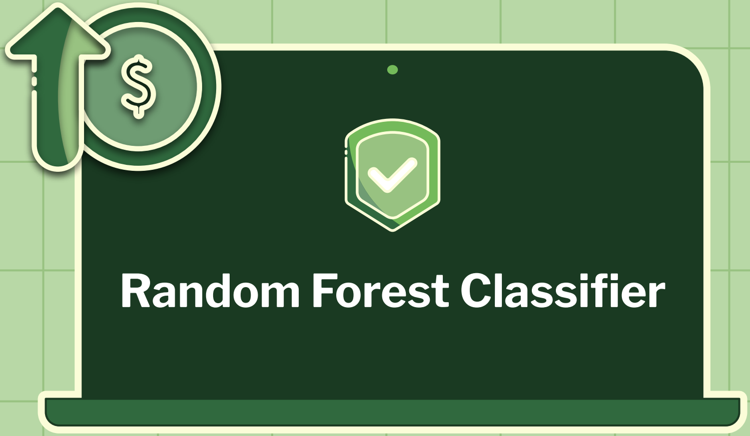
For "**female**" earning more than 50k is rare with only 3.63% of all observations.

But for **male**, 20.28% of all people earn more than 50k.

03

Description about the best model

Description about the best model



Based on the evaluation metrics of the tuned models, it appears that the Random Forest Classifier outperforms both the Logistic Regression and K-Nearest Neighbors (KNN) models. The Tuned Random Forest Classifier achieved the highest accuracy of approximately 87%, coupled with a balanced recall and precision, making it a strong choice for this classification task. It exhibits an impressive ability to correctly predict both income groups (<=50K and >50K) while maintaining good overall precision and recall scores.

- **While all models exhibit room for improvement, the Random Forest Classifier demonstrates the most promising results among the three, offering a robust foundation for further refinement and optimization in subsequent analyses and**

04

Final recommendations based on the results

Final recommendations based on the results

In conclusion, this dataset provides a valuable foundation for addressing income-related challenges across various sectors. Based on our analysis, we recommend the following:

Enhance Data Quality: Continuously monitor and improve data quality to ensure accurate and reliable predictions.

Refine Feature Engineering: Invest in further feature engineering to extract more meaningful insights from the data, potentially incorporating additional external data sources to enrich the dataset.

Fine-Tune Models: Experiment with different machine learning algorithms, hyperparameter tuning, and model ensembles to improve prediction accuracy.

Looking ahead, there are several avenues for enhancing the effectiveness of this project:

Advanced Models: Explore the use of advanced machine learning and deep learning models to further improve prediction accuracy.

Interpretability: Models like Random Forest, which offer interpretability, are preferred in contexts where stakeholders need to understand the reasons behind predictions.applications.

THANK YOU FOR YOUR ATTENTION !

For any additional questions, please contact:

mohamediyadhtajouri@gmail.com

LinkedIn

GitHub