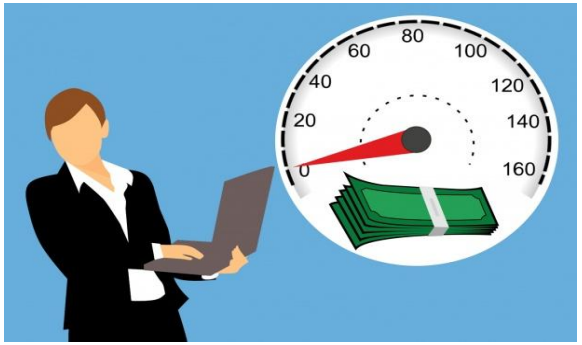
# Adult Income Prediction

Anjali Prakash

# Project Description




The purpose of this project is to use various features (such as age, race, hours worked etc) to predict the income of an individual.

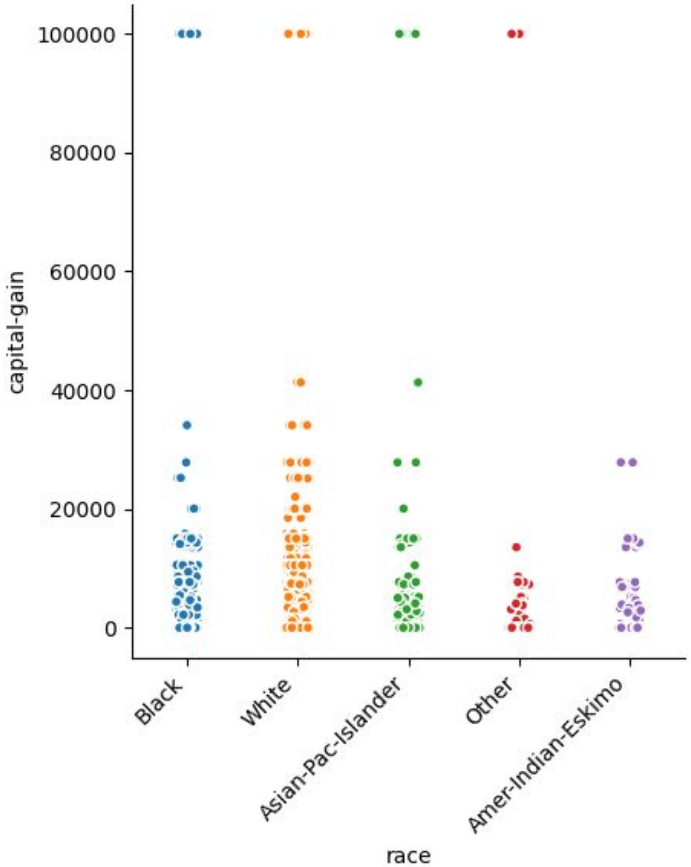
This data set was sourced from Kaggle - Adult Income Data Set. This dataset has 16 columns, utilizing 15 features to predict income of an individual.

# Stakeholders

The potential stakeholders for this project is extensive. Ranging from companies trying to market their product to individuals based on income or government programs that are trying to have an understanding of individuals’ incomes based on various demographics.

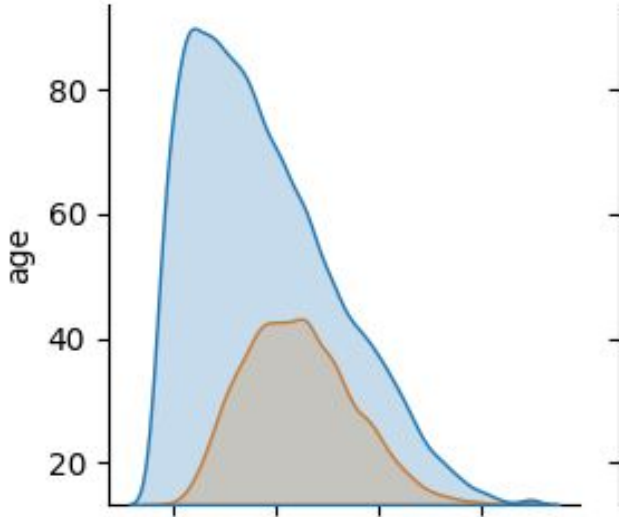
# Key Findings

It appears that race has an interesting correlation to capital gain. White individuals appear to have the highest capital gain and individuals labelled as “other” have the least. This could be due to many factors including opportunity, generational wealth etc.

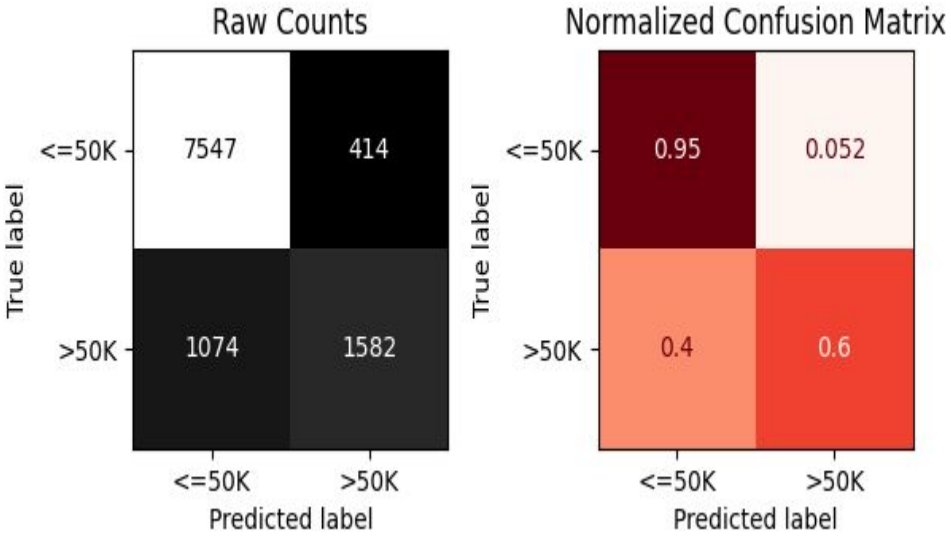


# Key Findings

Looking at the following visualization, it appears that older individuals have a higher income on average. This can be due to individuals who are older having more work experience for example.



# Model Evaluation




It appears that our Random Forest Classification model has a 95% accuracy of accurately identifying individuals who make >50k and a 60% accuracy of identifying individuals who make <=50k. This may be due to our uneven class balance  of 75% <=50k and 24.4% >50k. To help deal with this issue, we can resample the training data to even out the class balance.

# Recommendations

- I would recommend that stakeholders use the Random Forest Classification model as it appears to have the highest recall and accuracy amongst all the classification models used.
- False positives  are still not at the ideal accuracy rate, so ideally they go through a manual review by stakeholders.