Fairness in Machine Learning

1 Team members:

Atharva Barwe
Siddharth Das
Jaswanth Gorthi
Sagar Rudagi
Ahaan Hegde

2 Method attributes and presentation:

For our project we are considering a binary classification problem, i.e, to predict the annual income of an individual, if its greater or smaller than 50k. We can interpret our classifier as a random variable by considering $\hat{Y} = f(X)$.

For initial study we considered the the **adult dataset**, we first checked whether the data is unbalanced or not. An imbalance in data is dangerous because it can lead to bias. A Machine Learning model may learn the wrong lessons simply because data was not collected thoroughly enough.

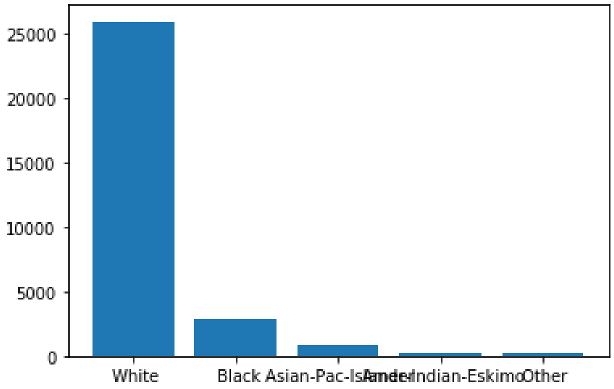
Our dataset contains several protected classes i.e. classes that should not have the outcome but they sometimes do anyway. Race, sex, and native country are the protected classes we have.

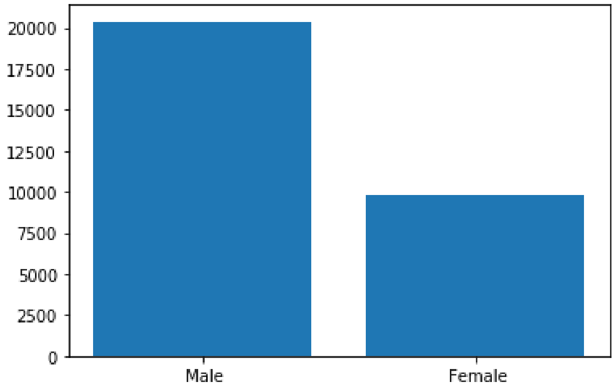
We tested if our dataset for disparate impact. Disparate Impact is a metric to evaluate fairness. It is responsible for comparing the proportion of individuals that have a positive outcome. There are two groups we compare - privileged and unprivileged. Essentially, what we are doing is measuring the success rate of the less favored group against that of the more favored one.

We can see that for all groups except the Asian Pacific Islander, we have calculated the ratio of probabilities to be lesser than 0.8, which is the industry standard. This means that the unprivileged group receives a positive outcome less than 80 percent of their proportion of the privileged group, making it a disparate impact violation.

This is the formula we use to know if our dataset has disparate impact. If the ratio is lesser than 0.8, it confirms that our dataset has disparate impact -

$$P(class => 50K|X = Black)/P(class => 50K|X = White)$$





Race - Black , Probability ratio = 0.49266804540332104
Race - Amer-Indian-Eskimo , Probability ratio = 0.45078871998012227
Race - Asian-Pac-Islander , Probability ratio = 1.0507243618386497
Race - Other , Probability ratio = 0.3447207858671523

These numbers show us that our dataset has disparate impact and we will need to address this issue so that a trained model is not unfairly biased toward or against a certain group.