# ***Prediction of Salary Class of an Individual***

## Problem:

The problem is to predict the salaried class of an individual is greater than $50,000 or less than $50,000 based on an individual's credentials like education level, age, gender, experience, occupation, etc. For example, the salary of an individual whose experience is above 15 years is most likely to be greater than $50,000. Prediction is not made by considering only one factor but all the factors that affect the income of an individual.

## About Dataset:

The dataset is taken from Kaggle. The US Adult Census dataset is a repository of 32,561 entries with 15 variables. The dataset contains information about age, work class, education, occupation, relationship, country, income. Let's look at dataset details.

| | age | workclass | fnlwgt | education | education.num | marital.status | occupation | relationship | race | sex | capital.gain | capital.loss | hours.per.week | native.country | income |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | 90 | ? | 77053 | HS-grad | 9 | Widowed | ? | Not-in-family | White | Female | 0 | 4356 | 40 | United-States | <=50K |
| 2 | 82 | Private | 132870 | HS-grad | 9 | Widowed | Exec-managerial | Not-in-family | White | Female | 0 | 4356 | 18 | United-States | <=50K |
| 3 | 66 | ? | 186061 | Some-college | 10 | Widowed | ? | Unmarried | Black | Female | 0 | 4356 | 40 | United-States | <=50K |
| 4 | 54 | Private | 140359 | 7th-8th | 4 | Divorced | Machine-op-inspct | Unmarried | White | Female | 0 | 3900 | 40 | United-States | <=50K |
| 5 | 41 | Private | 264663 | Some-college | 10 | Separated | Prof-specialty | Own-child | White | Female | 0 | 3900 | 40 | United-States | <=50K |
| 6 | 34 | Private | 216864 | HS-grad | 9 | Divorced | Other-service | Unmarried | White | Female | 0 | 3770 | 45 | United-States | <=50K |
| 7 | 38 | Private | 150601 | 10th | 6 | Separated | Adm-clerical | Unmarried | White | Male | 0 | 3770 | 40 | United-States | <=50K |
| 8 | 74 | State-gov | 88638 | Doctorate | 16 | Never-married | Prof-specialty | Other-relative | White | Female | 0 | 3683 | 20 | United-States | >50K |
| 9 | 68 | Federal-gov | 422013 | HS-grad | 9 | Divorced | Prof-specialty | Not-in-family | White | Female | 0 | 3683 | 40 | United-States | <=50K |
| 10 | 41 | Private | 70037 | Some-college | 10 | Never-married | Craft-repair | Unmarried | White | Male | 0 | 3004 | 60 | ? | >50K |
| 11 | 45 | Private | 172274 | Doctorate | 16 | Divorced | Prof-specialty | Unmarried | Black | Female | 0 | 3004 | 35 | United-States | >50K |
| 12 | 38 | Self-emp-not-inc | 164526 | Prof-school | 15 | Never-married | Prof-specialty | Not-in-family | White | Male | 0 | 2824 | 45 | United-States | >50K |
| 13 | 52 | Private | 129177 | Bachelors | 13 | Widowed | Other-service | Not-in-family | White | Female | 0 | 2824 | 20 | United-States | >50K |
| 14 | 32 | Private | 136204 | Masters | 14 | Separated | Exec-managerial | Not-in-family | White | Male | 0 | 2824 | 55 | United-States | >50K |
| 15 | 51 | ? | 172175 | Doctorate | 16 | Never-married | ? | Not-in-family | White | Male | 0 | 2824 | 40 | United-States | >50K |
| 16 | 46 | Private | 45363 | Prof-school | 15 | Divorced | Prof-specialty | Not-in-family | White | Male | 0 | 2824 | 40 | United-States | >50K |
| 17 | 45 | Private | 172822 | 11th | 7 | Divorced | Transport-moving | Not-in-family | White | Male | 0 | 2824 | 76 | United-States | >50K |
| 18 | 57 | Private | 317847 | Masters | 14 | Divorced | Exec-managerial | Not-in-family | White | Male | 0 | 2824 | 50 | United-States | >50K |
| 19 | 22 | Private | 119592 | Assoc-acdm | 12 | Never-married | Handlers-cleaners | Not-in-family | Black | Male | 0 | 2824 | 40 | ? | >50K |

Here is the link of Dataset: Adult Census Income | Kaggle

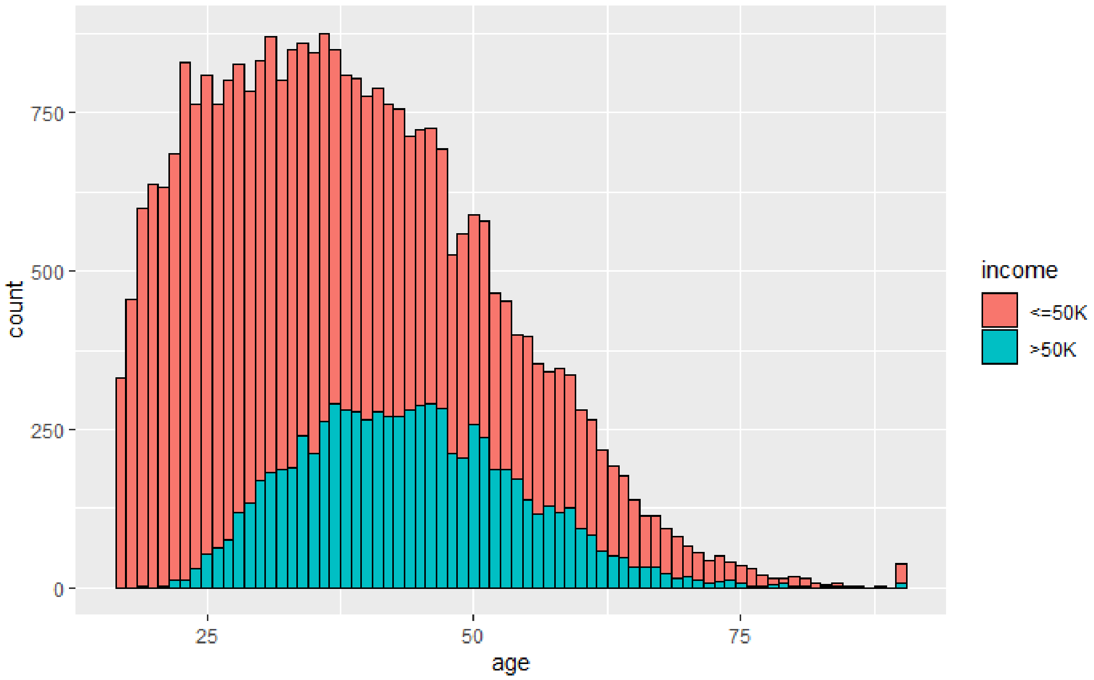
## Approach:

As this is the prediction problem, we can use the Naive Bayes method, Linear Regression and Logistic Regression. The prediction variable (salary class) depends on various variables (both categorical and Numeric), So the Naive Bayes method and Logistic Regression are the best for this problem compared to the Linear Regression. Our dataset has more categorical variables compared to numerical variables, so Logistic Regression is the best suitable solution for this problem.

**<u>Logistic Regression</u>:** Logistic regression is one of the most popular Machine Learning algorithms, which comes under the Supervised Learning technique. It is used for predicting the categorical dependent variable using a given set of independent variables. Therefore, the outcome must be a categorical or discrete value. It can be either Yes or No, 0 or 1, true or False, etc. but instead of giving the exact value as 0 and 1, **it gives the probabilistic values which lie between 0 and 1**.

# Analysis:

From the graph below we can say that the original dataset contains a distribution of 25% entries labelled with >50k and 75% entries labelled with <=50k.




The first step we took was to visualize the distribution of each variable and its effect on the likelihood of earning more than $50,000 per year. From our analysis, we concluded that the most useful variables for prediction were age, education, hours per week, occupation, and sex. That means We can say that age, education, hours per week, occupation, and sex are the variables that are going to impact the salary of an individual. So, to predict we should consider all these variables and develop our model.

## Results:

As this is a prediction problem Accuracy is the measure to find how our model is predicting the salaried class of an individual. The accuracy of the model is shown below.

| | FALSE | TRUE |
|---|---|---|
| <=50K | 5627 | 140 |
| >50K | 1190 | 722 |

Accuracy = (Correct predictions) / (Total predictions)

= (5627+722) / (5627+140+1190+722)

= 6349 / 7679

=0.8268

The accuracy of our model is 82.68% to predict the salary class of an individual by considering age, education, hours per week, occupation, and sex as categorical variables.

## Conclusions:

From the results, we can conclude that the salary of an individual whose income is greater than $50,000 is male, whose education level is higher than a master's degree and who works more than 40 hours per week, and whose occupation is in the private sector.

**GitHub link**: https://github.com/Tarak477/64060_tnunna.git