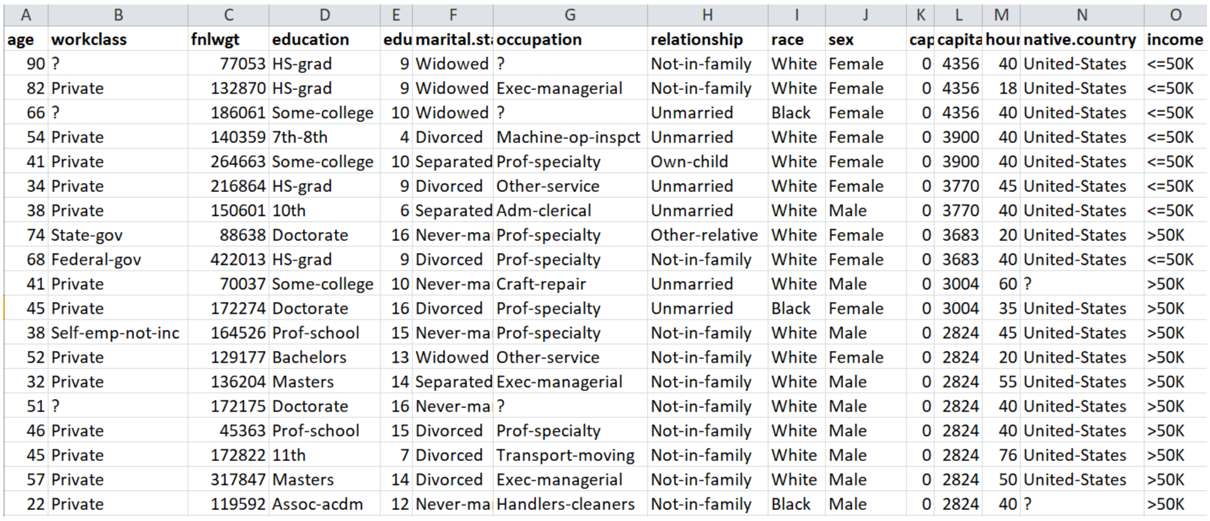
Prediction of Salary Class of an Individual

Problem Statement: Using an individual's credentials like education level, age, gender, experience, occupation, etc., we need to predict whether an individual's salary class is greater than $50,000 or less than $50,000. A salary of $50,000 is most likely to be earned by an employee with over 15 years of experience. An individual's income cannot be predicted based on just one factor, but rather on all factors that influence it.

Dataset: The dataset was taken from Kaggle. There are 32,561 entries in the US Adult Census dataset with 15 variables. Age, work class, education, occupation, relationship, country, and income are all included in the dataset. Here are some details about the dataset.



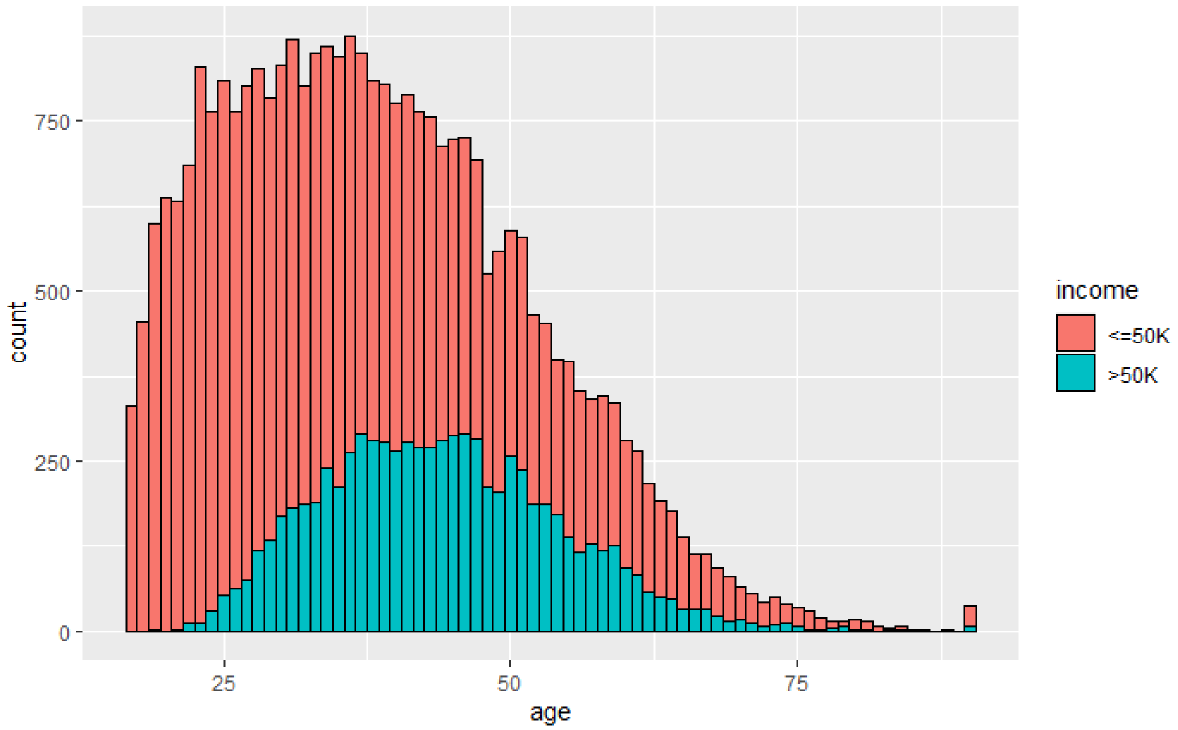
A	B	C	D	E	F	G	H	I	J	K	L	M	N	O
age	**workclass**	**fnlwgt**	**education**	**edu**	**marital.st**	**occupation**	**relationship**	**race**	**sex**	**cap**	**capita**	**hour**	**native.country**	**income**
90	?	77053	HS-grad	9	Widowed	?	Not-in-family	White	Female	0	4356	40	United-States	<=50K
82	Private	132870	HS-grad	9	Widowed	Exec-managerial	Not-in-family	White	Female	0	4356	18	United-States	<=50K
66	?	186061	Some-college	10	Widowed	?	Unmarried	Black	Female	0	4356	40	United-States	<=50K
54	Private	140359	7th-8th	4	Divorced	Machine-op-inspct	Unmarried	White	Female	0	3900	40	United-States	<=50K
41	Private	264663	Some-college	10	Separated	Prof-specialty	Own-child	White	Female	0	3900	40	United-States	<=50K
34	Private	216864	HS-grad	9	Divorced	Other-service	Unmarried	White	Female	0	3770	45	United-States	<=50K
38	Private	150601	10th	6	Separated	Adm-clerical	Unmarried	White	Male	0	3770	40	United-States	<=50K
74	State-gov	88638	Doctorate	16	Never-ma	Prof-specialty	Other-relative	White	Female	0	3683	20	United-States	>50K
68	Federal-gov	422013	HS-grad	9	Divorced	Prof-specialty	Not-in-family	White	Female	0	3683	40	United-States	<=50K
41	Private	70037	Some-college	10	Never-ma	Craft-repair	Unmarried	White	Male	0	3004	60	?	>50K
45	Private	172274	Doctorate	16	Divorced	Prof-specialty	Unmarried	Black	Female	0	3004	35	United-States	>50K
38	Self-emp-not-inc	164526	Prof-school	15	Never-ma	Prof-specialty	Not-in-family	White	Male	0	2824	45	United-States	>50K
52	Private	129177	Bachelors	13	Widowed	Other-service	Not-in-family	White	Female	0	2824	20	United-States	>50K
32	Private	136204	Masters	14	Separated	Exec-managerial	Not-in-family	White	Male	0	2824	55	United-States	>50K
51	?	172175	Doctorate	16	Never-ma	?	Not-in-family	White	Male	0	2824	40	United-States	>50K
46	Private	45363	Prof-school	15	Divorced	Prof-specialty	Not-in-family	White	Male	0	2824	40	United-States	>50K
45	Private	172822	11th	7	Divorced	Transport-moving	Not-in-family	White	Male	0	2824	76	United-States	>50K
57	Private	317847	Masters	14	Divorced	Exec-managerial	Not-in-family	White	Male	0	2824	50	United-States	>50K
22	Private	119592	Assoc-acdm	12	Never-ma	Handlers-cleaners	Not-in-family	Black	Male	0	2824	40	?	>50K

Data Set Link: UCI Machine Learning. (2016). Adult Census Income. Retrieved June 15, 2022, from Kaggle.com website: https://www.kaggle.com/datasets/uciml/adult-censusincome?resource=download

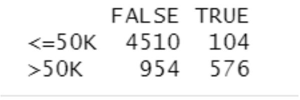
Solution Approach: Naive Bayes, Linear Regression, and Logistic Regression can be used as approaches to this prediction problem. In this problem, the Naive Bayes method and Logistic Regression are better than Linear Regression because the prediction variable (salary class) depends on various variables (both categorical and numerical). Logistic Regression is the most suitable solution for our dataset since there are more categorical variables than numerical variables.

Logistic Regression: Logistic regression is one of the most popular Machine Learning algorithms under the Supervised Learning category. Using a set of independent variables, it predicts the categorical dependent variable. As a result, the outcome must have a discrete or categorical value. The answer can be Yes or No, 0 or 1, true or false, etc., but rather than giving an exact number, it gives a probability ranging from 0 to 1.

Analysis: According to the graph below, the original dataset contains 25% entries labelled with >50k and 75% entries labelled with <=50k. As a first step, we visualized the distribution of each variable and its effect on earning more than $50,000 a year. As a result of our analysis, we concluded that age, education, hours per week, occupation, and sex were the most useful variables for predicting outcomes. An individual's salary will be influenced by their age, education, hours per week, occupation, and sex. Our model should take into account all these variables in order to predict.



Results: Since this is a prediction problem, accuracy measures how well our model predicts an individual's salary class. Model’s accuracy is shown below.



```
        FALSE  TRUE
  <=50K  4510   104
  >50K    954   576
```

Accuracy = (Correct predictions) / (Total predictions)

= (4510+576) / (4510+104+954+576)

= 5086 / 6144

=0.8277

The accuracy of our model is 82.77%

Conclusion: The study suggests that an individual with a salary exceeding $50,000 is male, has higher education than a master's degree, works more than 40 hours per week, and is employed by a private company.