

CIND 820: CAPSTONE PROJECT

Toronto Metropolitan University

(Formerly Ryerson University)

Anil Nath

500769023

Supervisor: TBA

Table Of Contents:

[**ABSTRACT:**](#_m0enqpslqdi) **3**

[**PROJECT APPROACH:**](#_peayeq62c76p) **5**

[**LITERARY REVIEW:**](#_ou7ddpnhseo5) **6**

[**DATA DESCRIPTION AND EXPLORATORY DATA ANALYSIS (EDA):**](#_jl79yh5g5udk) **12**

[Data Description:](#_yef7zixdmz7f) 12

[About the Dataset:](#_47izlhcvl30u) 12

[Observation:](#_fiqg3jzdb0gj) 13

[Observation:](#_ziaifbgkhxl8) 14

[EDA:](#_gn9adf7z9cho) 15

[Observations:](#_bucfewheku44) 15-24

[**Data Preparation:**](#_8nnt4vlwrzx7) **25**

[**Model Development and Performance Validation:**](#_k06oz8qqv2tt) **25**

[Logistic Regression:](#_6f4dmo7z8feb) 26

[Cross-validation:](#_imkyowkguw8o) 26

[K-fold cross-validation:](#_rfnsktqx8nn8) 26

[Support vector classifier:](#_34ptag7ezbca) 27

[K-nearest neighbors classifier:](#_pjzt273829vo) 27

[Decision tree classifier:](#_3ze2a73nsce6) 27

[Random forest classifier:](#_hl2wr0zg2hv2) 28

[Naive bayes classifier:](#_ip7e4lxfzzkk) 28

[**Models Performance and Comparison:**](#_fjkpsl22oz8h) **29**

[**Final Model:**](#_i2xpo6p6yv9) **29**

[Factor influencing prediction:](#_cax806c34cz6) 31

[**Conclusion:**](#_im4sudk5uqoj) **32**

# **ABSTRACT:**

As inflation and the cost of living continues to rise at a sharp rate, a person's income has a strong influence on their socioeconomic status and overall livelihood. Though causation cannot be proven, this study will explore the correlation between income and various socioeconomic factors including but not limited to a person's age, gender, race, employment status, and the highest level of education completed. Using the Adult data set from the UCI Machine learning Repository, this study will explore census data income. The Adult Data is a multivariate data set consisting of fourteen dependent and independent variables; all data was retrieved by Barry Becker in 1994 from the Census database. It includes 48842 instances which include missing values.

This study will adopt the statistical techniques of predictive analytics via statistical methodologies such as Decision Trees, K-Nearest Neighbor Algorithm (KNN), Support Vector Machine (SVM), Naïve Bayes Classifier, and Logistic Regression, Random Forest. All Data will be normalized, formatted, and filtered prior to analysis. Furthermore, exploratory analysis will be performed. Initially, the exploratory analysis will be conducted between income and the aforementioned socioeconomic parameters, to determine if there exists some relationship between the two. Specifically, this study will focus on the link between income, gender, education level, and hours worked per week. Through the aforementioned analysis, the factors for which income is most dependent will be highlighted. Further studies would be needed to examine if true correlations appear consistent and the influence of covariates.

Specifically, we will investigate if a person's income is above or below 50,000$ and what socioeconomic factors have the most influence on it. The three main factors that we can explore in more detail are: gender, highest education level, and hours worked per week. The aforementioned statistical techniques will be used to evaluate the magnitude and degree to which each of these three factors plays a role in a person's income, and whether or not it exceeds 50,000$ annually. The specific research questions explored will be:

-What role/weight does gender have in determining an individual's income?

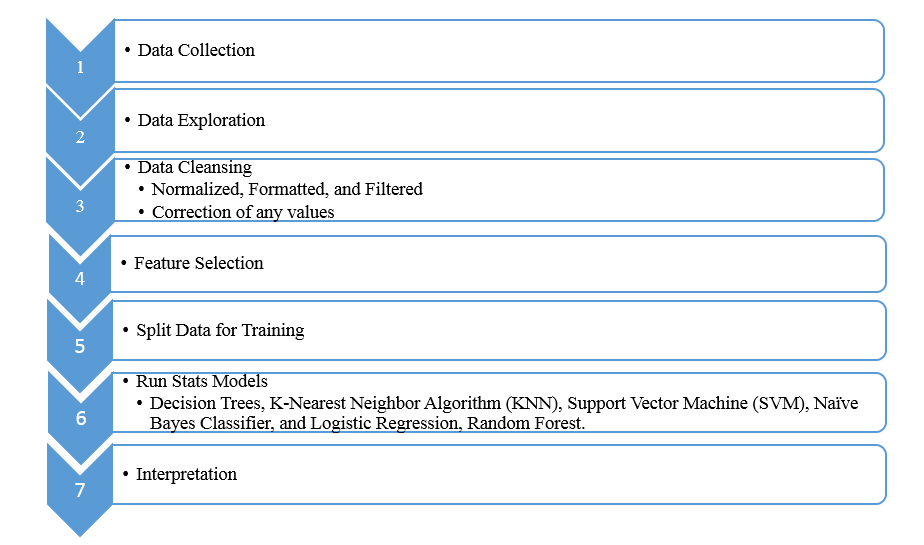
-How does an individual's education level affect their income?

-How does the hour's work per week play on an individual's income?

These questions will relate back to the main theme of looking into socioeconomic factors and determining the correlation to income.

By deploying the previously stated statistical methodologies, through a python based platform; this study will provide graphical as well as descriptive stats to explore the relationship between socioeconomic variables and an individual's income.

# **PROJECT APPROACH:**



The approach for this topic is straightforward. Steps 4,5, and 6 should take up the majority of time and effort. The data cleansing step should be tricky as the data has a multitude of missing values. Step 5 and 6 will prove to be very interesting as we are running a multitude of statistical methods; comparing and diagnosing the results will most likely be cumbersome. Overall, the steps taken will help guide this project and ultimately allow me to interpret the data.

# **LITERARY REVIEW:**

Looking at the role that socioeconomic factors play in an individual's income, is an examination that has been heavily researched, especially within the last few years. Research on how socioeconomic factors affect an individual's livelihood span, health, education, careers, and also psychological health etc. It is well-documented that social factors are closely intertwined with economic factors. For example, “your employment will dictate your income. Your income level often correlates to your level of education and your level of education helps to dictate your employment.”(Dan,2022). It is also important to understand how socioeconomic factors play a role in one's income because as stated by the County Health Rankings site; “These factors affect our ability to make healthy choices, afford medical care and housing, manage stress, and more.” Within the confines of this study, it will be looking into how Gender, level of education, and the number of hours worked per week affect an individual's income. By analyzing these factors I aim to better understand how these socioeconomic factors weigh into whether or not a person is more likely to be making more than $50,000 annually.

In terms of a similar study in regard to this topic, we can look at Mersa B 2018 study on

Factors that Influencing Households Income. Although this study primarily focuses on Women rather than all genders it still looks at the array of socioeconomic factors that this study is focusing on. In this particular study Mersa B, is trying to determine how socioeconomic factors affect female workers' income/ contribution to total family income. In Mersa B study it looks at the following socioeconomic factors: age, level of education, work experience, and many family dependents. Through the use of mostly multiple linear regression, Mersa B was able to determine that age and experience factors played little to no role on income; however, education level and the dependencies factor played a quite significant role. Overall, Mersa B was able to come to the conclusion that “Simultaneously all socioeconomic factors have a significant influence on the income of housewives in Hamparan Perak Sub-District. The large contribution of housewives' income to family income is 32.72%, income that is still below Deli Serdang MSEs in 2018.”(Mersa,2018).

Furthermore, studies on how education affects one's income have been well explored. For instance, according to Pathways to Education, young people who dropout of high school or do not further their education, affect the income they receive long-term when seeking employment. For instance, “Additionally, those with more education have more stable positions and are less susceptible to economic downturns and changes associated with the 21st-century workplace (OECD, 2012).” This illustrates that the importance of obtaining a higher education opens doors to profiting from a higher income than those with lower education. In addition, According to the OECD (2012), Canadians with less than upper secondary education have earnings that are 22% lower than those of high school graduates.” This further examines the importance of education level in regard to income in Canada. When people have little to no education, it creates a barrier when trying to gain employment. As the cost of living arises, people must obtain a high education level to sustain the cost of basic living. This article also highlights the fact that the costs of post-secondary after high school is a barrier itself to those trying to gain a higher education level.

Moreover, the study of how gender affects one's income is a topic that has been heavily explored, especially in the past few years with the push for women's rights and equality. According to the Canadian Women’s Foundation, “ for full-time employees, there is a 16.1% difference between annual median earnings of women and men relative to the annual median earnings of men.” (Howard,2022). One can further look at the case study done by MIT OpenCourseware; instructors Dr. Richard Fletcher, Prof Daniel Frey, Dr. Mike Teodorescu, Amit Gandhi, and Audace Nakeshimana. This case study uses the exact same data set that this study intends to use, the Adult Data set from UCI. In this particular case study, they look into gender bias within the dataset. By using techniques like support vector, random forest, KNN, logistic regression, and MLP classifiers the case study was able to find significant results. The case study was able to find that there is a 1 to 2 ratio of women to men and that 1 in 3 men was making over $50,000 while 1 in 5 was reported to make the same amount. It was also able to determine that “For high salaries, the number of data points in the male population is significantly higher than the number of data points in the female category.” (Fletcher et al, 2020).

In terms of exact identical studies, there is very limited application of studies that pursue to look at gender, education level and age factors and their effect on income all at once. However, studies that include the Adult data set from UCI are bountiful, but the data set is mostly used to explore differences and the effectiveness of different statistical techniques/models. For instance, we can look at Mohammed Topiwalla study *Machine Learning on UCI Adult data Set Using Various Classifier Algorithms And Scaling Up The Accuracy Using Extreme Gradient Boosting,* where he explores how using extreme gradient boosting to further improve the accuracy of the machine learning model. In this study, he aims to compare different statistical models like Decision Tree, Naïve Bayes, KNN, and SVM and then compare them to more complex models in order to improve the accuracy from further previous papers' results. In this study, Topiwalla also uses a method known as stacking, “Stacking is a concept used to boost the accuracy of models by helping one model to learn what another model has already learnt” (Topiwalla, 2017). Based on the investigation, Topiwalla was able to determine that random forest and xgboost models performed the best. Moreover, models that included the country variable were less efficient as the variable is biased.

Moreover in terms of improving statistical models, one can definitely look at Ron Kohavi's work. In his work, Kohavi uses multiple datasets including the Adult Data set. He aims to compare and scale up previous results with statistical techniques. The study is able to improve on past results by deploying the NBTree method. This method is a hybrid that pulls benefits from both the Decision Tree and Naive Bayes models. The NBTree method looks similar to a model described as “a decision tree that is built with univariate splits at each node, but with Naive-Bayes classifiers at the leaves.” (Kohavi, 1996). Kohavi was able to conclude that NBTree techniques work tremendously well in situations where data sets contain a multitude of factors that play a role in predicting.

We can also look at the study by Chakravarty and Biswas 2018. In this study, they try to improve upon existing statistical accuracy previously found in the Adult data set. To do this, the duo deployed a statistical method known as Gradient Boosting Classifier with extensive Parameter Tuning. This technique helped remove biases by boosting which takes a multitude of weak learners in a method, which improves overall observations. These model results were “the gradient boosting classifier model was deployed which clocked the highest accuracy of 88.16%” (Biswas et al, 2018). This technique was able to improve upon the aforementioned work of Mohammed Topiwalla.

In addition, one can also look at the study by Zhang, Lemonie, and Mitchell; *Mitigating Unwanted Biases with Adversarial Learning.* This study aims to help alleviate biases that may occur in machine learning algorithms. “Machine learning is a tool for building models that accurately represent input training data. When undesired biases concerning demographic groups are in the training data, well-trained models will reflect those biases” (Zhang,Lemonie, Mitchell, 2018). By using the Adult UCI dataset, this study utilized a method that in fact does improve previous results, by helping to lower the underlying biases from the dataset itself. It is extremely important to understand how we can mitigate these biases; “For example, systems designed to predict creditworthiness and systems designed to perform analogy completion have been demonstrated to be biased against racial minorities and women respectively.”(Zhang et al, 2018).

Another example that uses census data is a study that aims to evaluate income prediction conducted by Alina Lazar 2004. Unlike previously mentioned studies, Lazar used the Support Vectoring Machine (SVM) method. The goal of this study was to improve prediction time and accuracy across the board. By deploying the PCA method Lazar was able to find that “good classification accuracy was obtained in faster time” (Lazar, 2004). It is to be noted that this was done with a vertically reduced data set. This method, in theory, should allow for “a speed up of the learning algorithm should result since most of the calculations are done at the kernel level.” (Lazar, 2004). This proved to make the SVM method more feasible and practical.

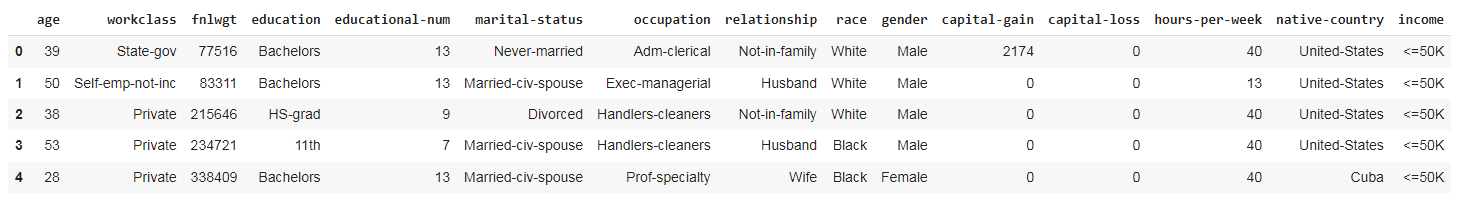
In the works of Deepjatohi and Selvarajan, they aim to compare what is the best and most efficient method of classification on the adult data set. By comparing an array of methods including Naive Bayes, Random Forrest, Zero R, and K star; this study finds that Naive Bayes was the fastest and proved to have higher accuracy than the other models. This study is different from the ones done in the past as the team was able to do Data Preprocessing. This technique allows for improved results. “When applied before mining, can substantially improve the overall quality of patterns mined and the time required for actual mining.” (Deepjatohi, Selvarajan, 2012).

As seen, the study of how socioeconomic factors correlate to one's income has been heavily studied. It is becoming ever more important to understand how these factors are playing into one's income, as it has been proven that it has an effect on health, education, career path etc; which in turn has an effect on income. Studies consisting of UCI Adult data set span from investigating the gender wage gap to analyzing how education affects one's income and finding the predicting factors of one's income. In addition, the data set has been heavily explored in terms of statistical techniques each improving upon the last. Studies were also aimed to help reduce the biases in census data, especially from the Adult Data set. In light of the never-ending work done in this field, this study aims to investigate specifically how gender, level of education, and hours spent working affect/weigh into a person's income. By using a multitude of statistical models we will be able to compare results and see how these specific factors affect income. Taking from previous studies, this study will also take into consideration the biases that are present within the data set and in the statistical models. All in all, this study will help see how socioeconomic factors, specifically, gender, education level and hours spent working, affects whether or not an individual makes more or less than $50,000.

# **DATA DESCRIPTION AND EXPLORATORY DATA ANALYSIS (EDA):**

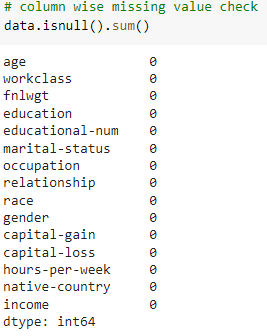
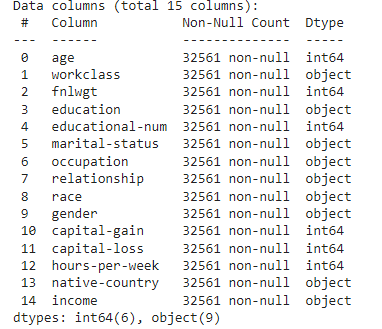
The Jupyter file can be found on <https://github.com/Anil-nath/CIND820-Data-capstone-analysis> or attached in submissions.

## Data Description:



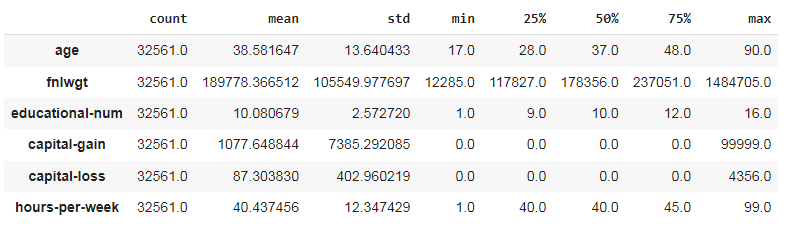
### About the Dataset:

* **Age**: Describes the age of individuals. Continuous.
* **Workclass**: Private, Self-emp-not-inc, Self-emp-inc, Federal-gov, Local-gov, State-gov, Without-pay, Never-worked.
* **fnlwgt**: Continuous.
* **education**: Bachelors, Some-college, 11th, HS-grad, Prof-school, Assoc-acdm, Assoc-voc, 9th, 7th-8th, 12th, Masters, 1st-4th, 10th, Doctorate, 5th-6th, Preschool.
* **education-num**: Number of years spent in education. Continuous.
* **marital-status**: Married-civ-spouse, Divorced, Never-married, Separated, Widowed, Married-spouse-absent, Married-AF-spouse.
* **occupation**: Tech-support, Craft-repair, Other-service, Sales, Exec-managerial, Prof-specialty, Handlers-cleaners, Machine-op-inspect, Adm-clerical, Farming-fishing, Transport-moving, Priv-house-serv, Protective-serv, Armed-Forces.
* **relationship**: Wife, Own-child, Husband, Not-in-family, Other-relative, Unmarried.
* **race**: White, Asian-Pac-Islander, Amer-Indian-Eskimo, Other, Black.
* **sex**: Female, Male.
* **capital-gain**: Continuous.
* **capital-loss**: Continuous.
* **hours-per-week**: Continuous.
* **native-country**: United-States, Cambodia, England, Puerto-Rico, Canada, Germany, Outlying-US(Guam-USVI-etc), India, Japan, Greece, South, China, Cuba, Iran, Honduras, Philippines, Italy, Poland, Jamaica, Vietnam, Mexico, Portugal, Ireland, France, Dominican-Republic, Laos, Ecuador, Taiwan, Haiti, Columbia, Hungary, Guatemala, Nicaragua, Scotland, Thailand, Yugoslavia, El-Salvador, Trinidad & Tobago, Peru, Hong, Holland-Netherlands.
* **salary**: >50K,<=50K



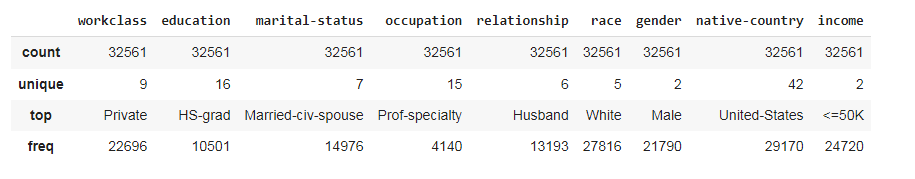
### Observation:

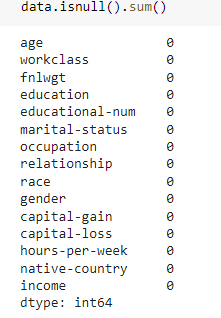
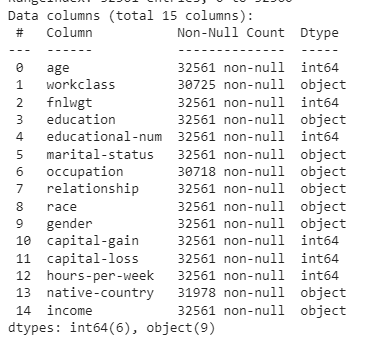
The dataset has 32561 rows and 15 columns. Although the set doesn't have any null values it does contain missing values; we will have to replace these values or mitigate them. Integer columns include: Age, Final Weight, Education Number, Capital Gain, Capital Loss and Hours Per Week. Object data types: Workclass, Education, Marital Status, Occupation, Relationship, Race, Gender, Native Country and Income. In addition, there are float types in the dataset.



### Observation:

When exploring the data we can see the max and min ages are 19 and 90, with the average age of 37. For years spent on education; we min and max years of 1 to 16, where the mean is 10 years. In terms of average capital gain we have a range of 0 to 99999, this looks like it is a calculation error that will have to be fixed before creating the models. In terms of hours spent per week we have a range of 1 to 99 with an average of 40.

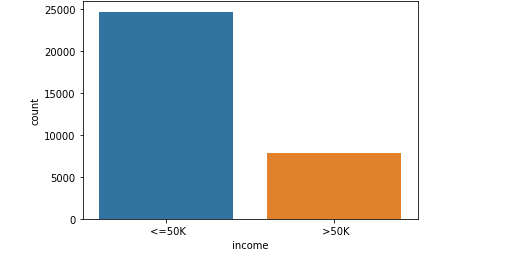




## 

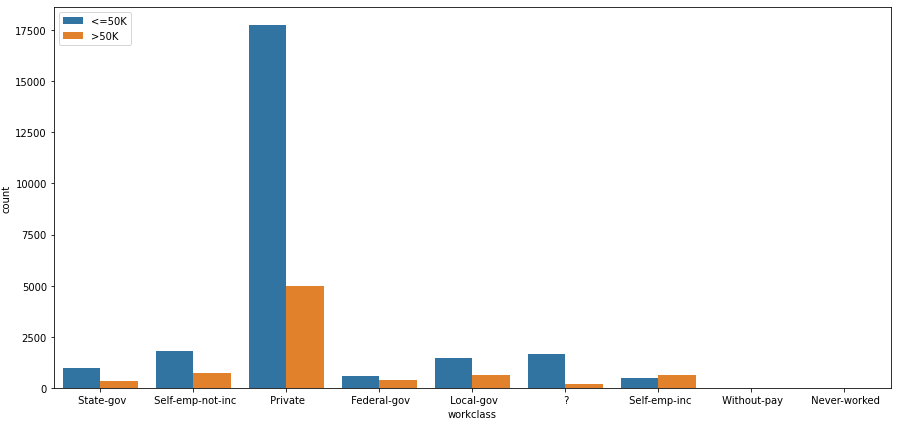
## 

## **EDA:**

\

### Observation:

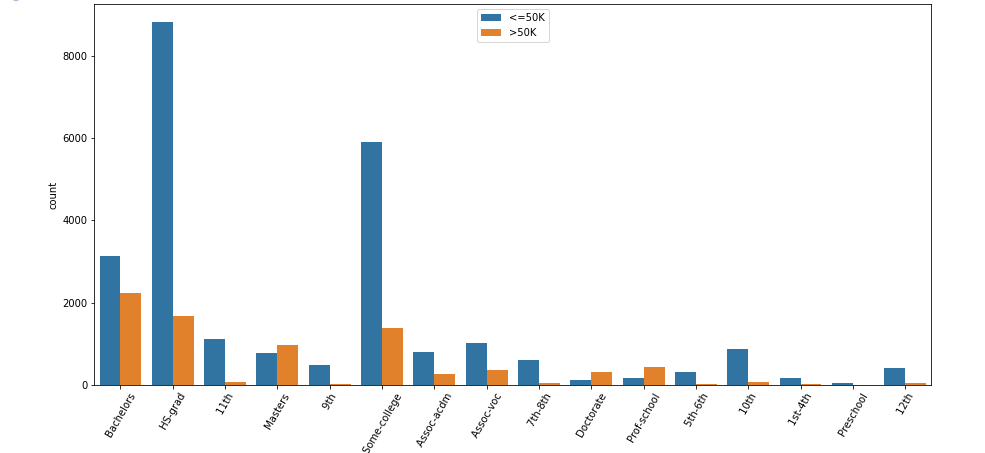
We can clearly see that only about one-third of the people are actually making over 50k. However, we should keep in mind that this data is quite dated. But should be significant in exploring factors that affect income.



### Observation:

In the above graph, we have segregated the incomes of adults on the basis of their different working classes.

Some quick insight indicates that people that are Self-employed have more people earning more than 50k than people earning less than 50k. You can also look at the private sector where there is a gap of around 75% or people making more than 50k. This gap is almost eliminated when you look at the Federal Gov field. We can see there is very little difference.

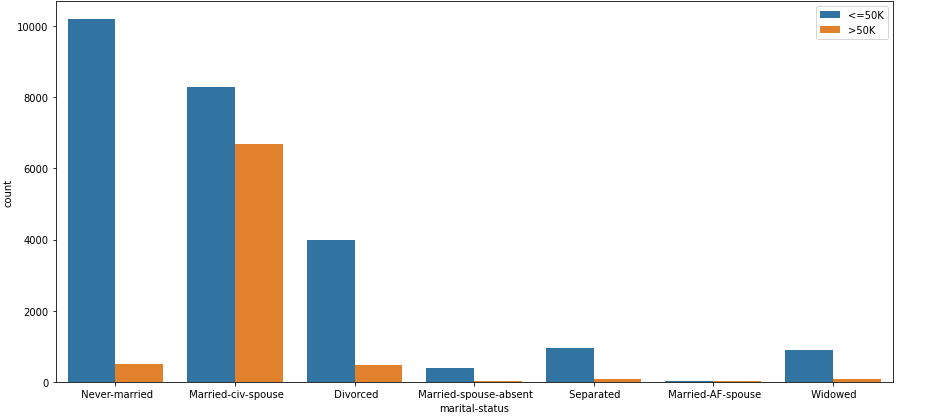


### 

### Observation:

We can draw a few insightful conclusions from this graph.

By looking at the above graph one can clearly see some basic stats. For people who have only a grade 12 education there are only a few people making over 50k. Whereas, people with post-education, where making over 50k seems to be more occurring.

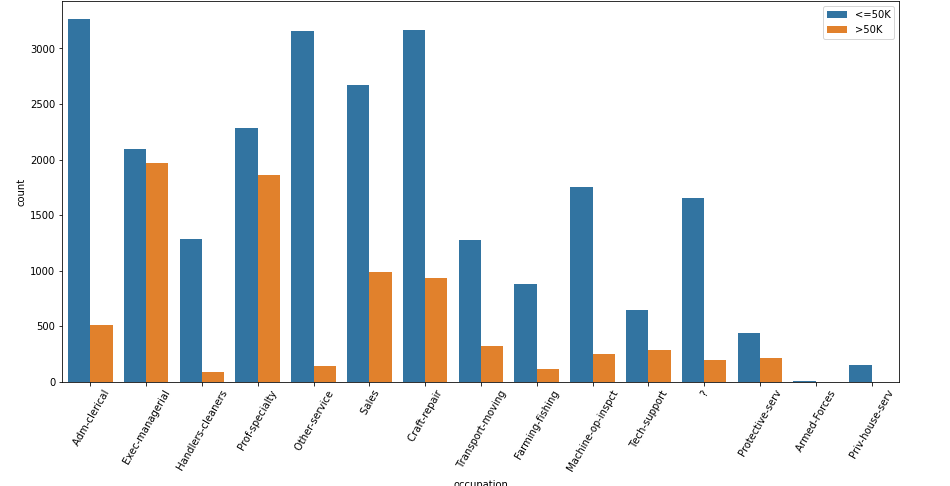


### 

### 

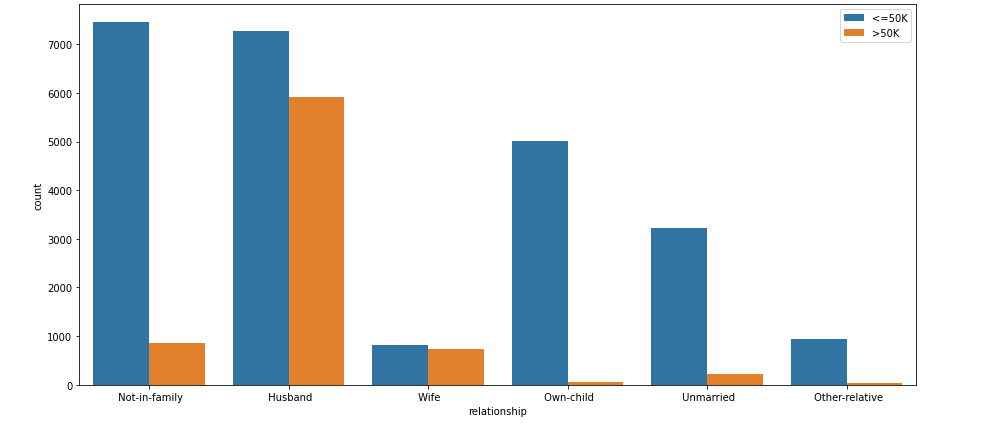
### Observation:

In terms of equal populations within the categories, Married-Civ-Spouse seems to be the only category with comparable populations. It looks like for the other categories there are less than 25% of people with an income of 50k.



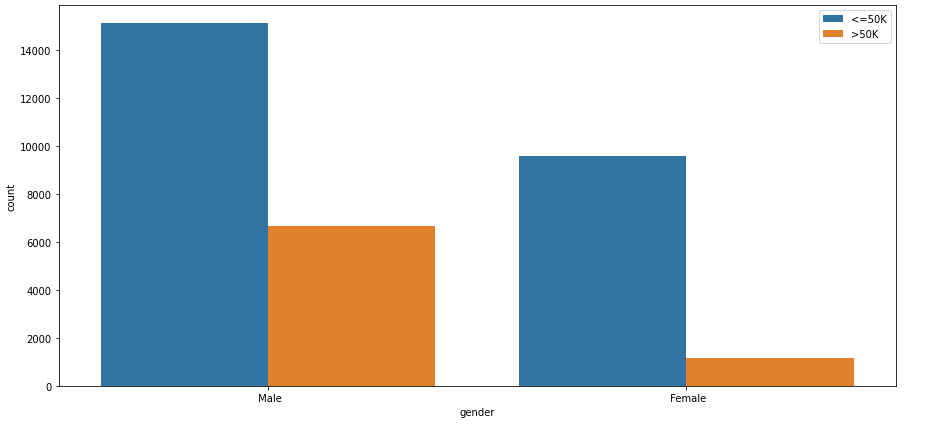
### Observation:

People that fall into the Exec-Managerial role are almost equally likely to be making 50k a year. In addition, the population of people that work in sales, only about 25% make more than 50K. People working in Machine-Op-inspect, Farming-fishing, Adm-Clerical, Transport-moving, and Other-Services, are less likely to be making 50K.



### Observation:

For the category of Wives, people are almost equally likely to be making 50k a year. Whereas people that fall into the Husband section have less probability of them making 50k. We can also look at the unmarried category where we can see only a handful of people make over 50k.



### 

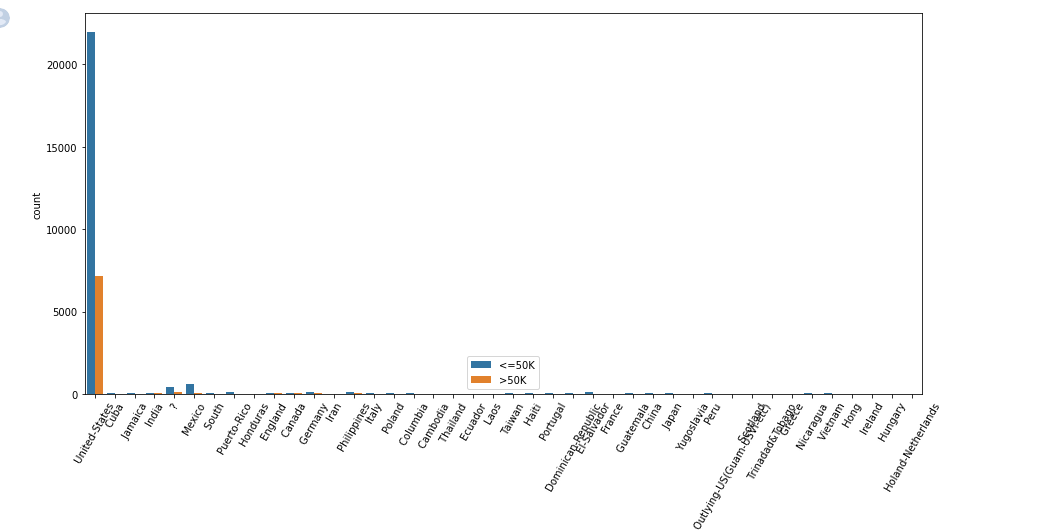
### 

### 

### 

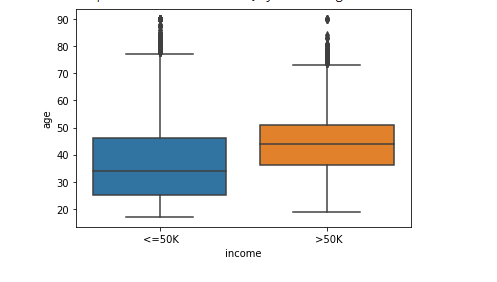
### Observation:

### We can see a significant gap in the earnings between males and females over here. Males have a 33% chance of making 50k. On the other hand, females have less than 10% making 50k or more.



### Observation:

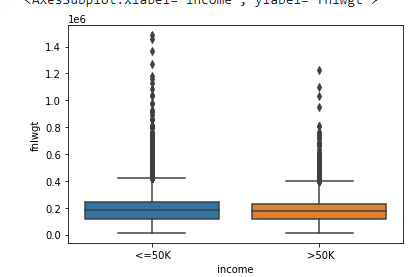
As we can see the data is very skewed. We have a majority of data hailing from the United States, this will cause an over-representation of a certain demographic and an under-representation of another. We will have to find a way to deal with this bias.



### 

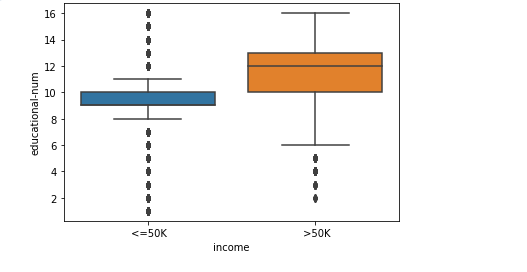
### Observation:

One can quickly see that individuals that are earning 50k or more are older than the level of people earning less than 50k a year.



### Observation:

Individuals with similar ‘fnlwgt’ are quite similar in comparison of who is making >50k and <=50k.

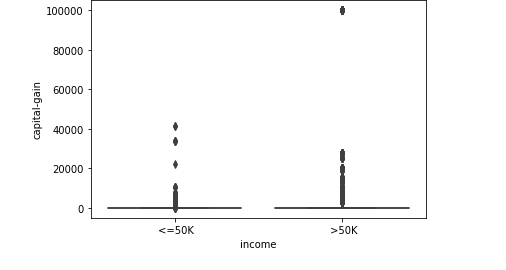


### 

### 

### Observation:

Individuals that are earning more than 50k have more formal education at an aggregate level, than those making less than 50k.



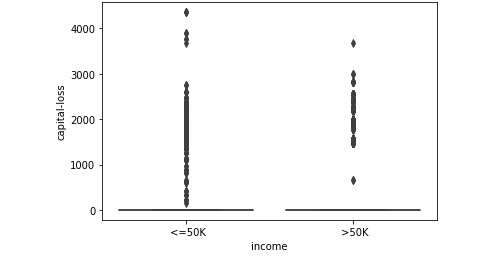
### 

### 

### 

### Observation:

In terms of capital gains, it is quite similar for the two groups. However, those that make more than 50k have more outliers that are high.

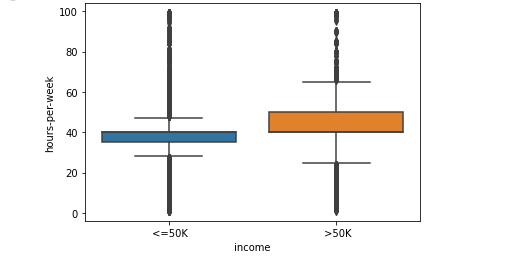


### 

### 

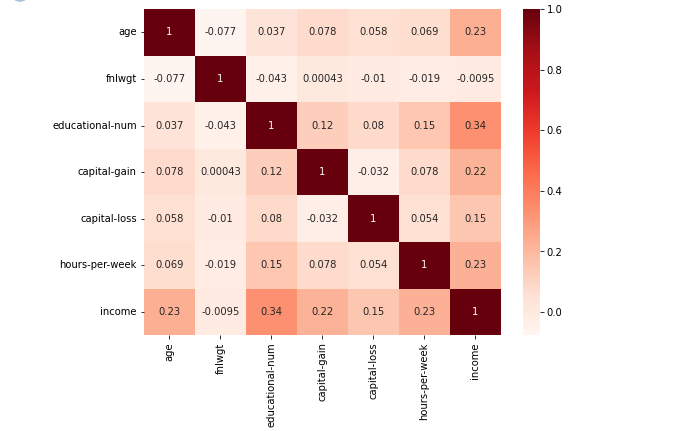
### Observation:

### Similar to capital gains, Capital loss looks similar for people making more than 50k and less than 50k. However, people making less than 50k have more outliers.



### Observation:

At an aggregate level one can see that people who make more than 50k are working more hours than those making less than 50k.



### Observation:

This is a correlation map that I aim to see how well independent and dependent features are related. Most of the categories positively correlate to Income. This map doesn't take into account the object data types.

# **Data Preparation:**

Before running statistical analysis I must first clean and prep the data. To do this we first start off by transforming categorical columns. We do this because the model can only take numerical features. To do this, I deployed all non-numeric features into numerical ones by using the technique of *One hot Encoding (Dummy encoding).* This process creates so-called “dummy variables”. It takes categorical variables where the order is insignificant. In this process for every categorical feature, a new variable is created. For the numerical columns, I decided to standardize/normalize the numerical features in order to get better performance.

# **Model Development and Performance Validation:**

As stated previously this study will be running a multitude of statistical tests, which include: Decision Trees, K-Nearest Neighbor Algorithm (KNN), Support Vector Machine (SVM), Naïve Bayes Classifier, Logistic Regression, Random Forest. This test will then be compared using the Accuracy mean and Accuracy Standard Deviation.

## Logistic Regression:

Logistic Regression is a supervised learning technique, and it is one of the most popular. It is used to predict categorical dependent variables using independent variables. Logistic regression must have a categorical or discrete value because it predicts the categorical dependent value output. It usually gives probabilistic values ​​between 0 and 1. Logistic regression is often described as similar to linear regression and is mainly used to solve classification problems, whereas the latter is used to solve regression problems. One of the biggest positives of Logistic Regression is that it can easily determine the most effective variables used for the classification.

### Cross-validation:

Cross-validation is a method used to estimate the performance or accuracy of a machine-learning model. It helps curb the problem of overfitting, especially in cases where the data is limited. By running a fixed number of partitions on the data, analyzing it gives us the average of the overall error estimate. The training data cannot represent the entirety of real-world data, therefore we must aim for the model to work well based on however small the training data may be.

### 

### K-fold cross-validation:

Similar to Cross-validation, K-fold splits the dataset into a K number of folds, then it is subjected to measure the model's ability to the new data. Because it allows for the data to be used in either training or test sets, this method often has less bias. Although sometimes slower, because the K-fold has to rerun itself K a number of times, it has to involve more time for computation, K-fold cross-valdiation proves highly effective.

## Support vector classifier:

Support Vector Machine or SVM is largely used for both classification and regression problems. However, it is mainly used for classification problems in machine learning. By creating a decision boundary that segregates the nth spaces so that new points can be placed in the correct categories in future instances. The hyperplane is often regarded as the best decision boundary. Taking the extreme point/ vectors (support vectors) the model creates the hyperplane.

## 

## K-nearest neighbors classifier:

K-NN works by putting data into classified categories and then reading the new data and putting them into the already-made category. The training phase stores the dataset, then it takes into the categories that are much similar to the new data. In other words, by creating already made categories K-NN can efficiently sort through new data points and classify them into the categories the share similarities to,

## Decision tree classifier:

Although mostly used for classification problems, Decision Tree models can be used for regression problems as well. Using a tree-like structure to classify the data points, the model uses internal nodes which are features of the dataset, the branches are the decision rules whiles the node represents the outcome. Two nodes are used in this model the decision node and the leaf node. The decision nodes contain multiple branches, and the leaf nodes are the output of those decisions that do not contain more branches.

## 

## 

## Random forest classifier:

By using a process of combining multiple classifiers otherwise known as Ensemble learning. Random Forest models are able to solve complex problems making Random Forest one of the most used supervised learning techniques. Growing from Decision trees, Random forest uses a multitude of Decision trees to classify the dataset, then takes the average to improve the overall accuracy. Moreover, the more trees that are used in the model, will increase accuracy as well as prevent problems of overfitting.

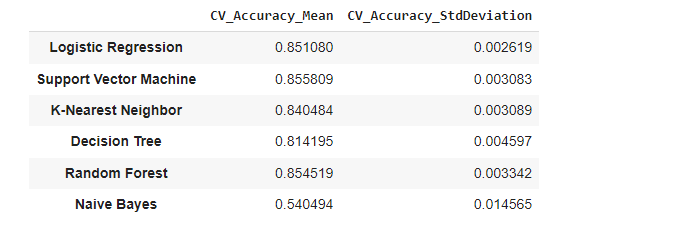
## 

## Naive bayes classifier:

Based on the Bayes theorem, the Naive Bayes Algorithm is one of the fastest models that can be made to make fast predictions. Simply put it the Navies Bayes works like P(A|B) is Posterior probability, the Probability of hypothesis A on the observed event B, P(B|A) is Likelihood probability: Probability of the evidence given that the probability of a hypothesis is true. , P(A) is Prior Probability: Probability of the hypothesis before observing the evidence. P(B) is Marginal Probability: Probability of Evidence.



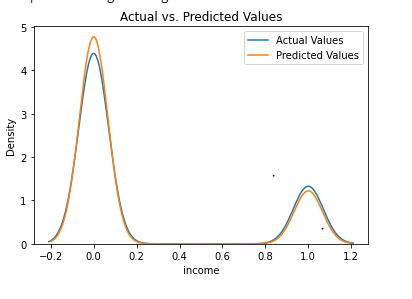
# **Models Performance and Comparison:**



In terms of overall average accuracy all the statistical techniques had quite similar scores. However, Random Forest seems to be the best-performing model when looking at mean accuracy and the standard deviation score. We will build the final model on the Random Forest Classifier algorithm.

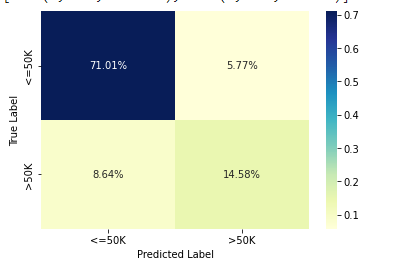
# 

# **Final Model:**



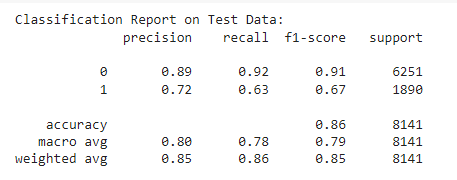
As expected the model was able to predict the income values based on the other characteristics quite well. This can be seen in the above figure. Seeing the distribution of the Actual and Predicted Values, the distributions are almost aligned this indicates, the model is a good fit.

We can also look at the confusion matrix:



The model was able to correctly classify that class <=50k around 71% of the time and >50k 14.5% of the time.

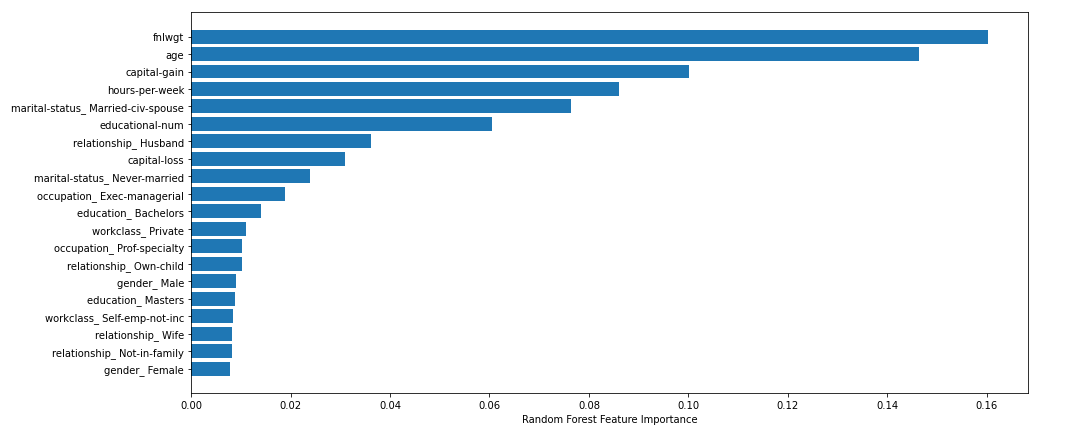
Furthermore, we can also interpret the classification report:



As stated we assume a positive class is =>50k. In terms of accuracy, the model was able to correctly identify the income class of 86% of the overall people. Precision-wise, the model predicted people having an income >50k, 72% of them actually did. In addition, the recall score was 63% meaning the model was able to correctly predict 63% of the people that had an income of >50k. The model was able to produce an F1 score of 67% and a weighted avg of 85%.

## Factor influencing prediction:

As stated the purpose of this study was to see how gender, education level, and hours worked per week affect one's income and the factor they play in helping an individual's income be greater than 50k. By using the results of the Random forest model and running feature importance, we were able to plot this graph.



This graph indicates the features that played the most significant role in the model. Listed in order: fnlwgt, age, capital gain, hours per week, marital status, educational-num, relationship, capital loss, occupation, education, workclass, gender. Whiles looking at the graph we can clearly see, hours worked per week and education level, played a significant role in predicting an individual's income. Whereas gender didn't play such a big role as one may have initially thought.

# **Conclusion:**

As discussed early in this paper, many studies have used the UCI adult database to progressively run different statistical models. In most cases, it saw Random Forest and SVM models producing highly accurate models similar to this study. However, we can looking at studies like *Ron Kohavi's,Chakravarty and Biswas* , *Mohammed Topiwalla and Zhang, Lemonie, and Mitchel.* Where so-called advanced statistical techniques were used to produce highly accurate models. This included gradient boosting, NBT Tree, xgboost etc. This study ultimately differs from this study due to the fact the underlying principle was to produce more highly intensive models essentially to improve upon the last model's prediction. Whereas this model is being used to help see what socioeconomic factors play a role in individuals' income. With that scope in mind, running these basic models like the one in this study is sufficient. Moreover, in the future when these aforementioned advanced techniques are more accessible and easier to process this study should be updated with the models.

Another significant difference between this model and others is the fact that this case does not take into account the biases that are prominent within the dataset. For example, the data is from 1994 and has been heavily explored, one of the biases is the country of origin field. Obviously, the majority of the entries within the data set come from the USA; this causes a misrepresentation of other countries' data. Because this study doesn't aim to classify its results under any country field; we are just looking into how gender, worked hours, and education level factor into one's income, it did not seem valuable to address this basis. However, moving forward in studying this field it would be nice to use an updated dataset with proper country representation. This way we could make other inferences.

All in all, this study aimed to see how different socioeconomic factors affect one's income, in terms of making more than or less than 50k. It was able to reasonably predict the income levels of individuals based on hours worked per week, and education level; whereas gender did not play such a significant role.

REFERENCES:

*Case Study with Data: Mitigating Gender Bias on the UCI Adult Database | Exploring Fairness in Machine Learning for International Development | Supplemental Resources*. (2022, October 20). MIT OpenCourseWare.<https://ocw.mit.edu/courses/res-ec-001-exploring-fairness-in-machine-learning-for-international-development-spring-2020/pages/module-four-case-studies/case-study-mitigating-gender-bias/>

Chakrabarty, N., & Biswas, S. (2018). *A Statistical Approach to Adult Census Income Level Prediction*.

Howard, J. (2022, October 20). *The Gender Pay Gap | Pay Gap in Canada | The Facts*. Canadian Women’s Foundation.<https://canadianwomen.org/the-facts/the-gender-pay-gap/>

Jackson, D. (2022, October 20). Socioeconomic factors. *HSC PDHPE*.<https://pdhpe.net/better-health-for-individuals/what-influences-the-health-of-individuals/the-determinants-of-health/socioeconomic-factors/>

*K-Nearest Neighbor(KNN) Algorithm for Machine Learning—Javatpoint*. (2022, December 6). Www.Javatpoint.Com.<https://www.javatpoint.com/k-nearest-neighbor-algorithm-for-machine-learning>

Kohavi, R. (n.d.). *Scaling Up theaADceccuisriaocny-TorfeNeaHivyeb-Bridayes Classi ers:* 6.

Lazar, A. (2004). *Income prediction via support vector machine.* 143–149.<https://doi.org/10.1109/ICMLA.2004.1383506>

*Logistic Regression in Machine Learning—Javatpoint*. (2022, December 6). Www.Javatpoint.Com.<https://www.javatpoint.com/logistic-regression-in-machine-learning>

*Machine Learning Decision Tree Classification Algorithm—Javatpoint*. (2022, December 6). Www.Javatpoint.Com.<https://www.javatpoint.com/machine-learning-decision-tree-classification-algorithm>

Mesra, M. (2018). Factors that influencing households income and its contribution on family income in hamparan perak subdistrict, deli serdang regency, North Sumatera-Indonesia. *International Journal of Civil Engineering and Technology*, *9*, 461–469.

*Naive Bayes Classifier in Machine Learning—Javatpoint*. (2022, December 6).<https://www.javatpoint.com/machine-learning-naive-bayes-classifier>

S.deepajothi, D. s selvarajan. (2012). A Comparative Study of Classification Techniques On Adult Data Set. *International Journal of Engineering Research*, *1*(8), Article 8.

*Social & Economic Factors*. (2022, October 20). County Health Rankings & Roadmaps.<https://www.countyhealthrankings.org/explore-health-rankings/measures-data-sources/county-health-rankings-model/health-factors/social-and-economic-factors>

*Support Vector Machine (SVM) Algorithm—Javatpoint*. (2022a, December 6). Www.Javatpoint.Com.<https://www.javatpoint.com/machine-learning-support-vector-machine-algorithm>

*Support Vector Machine (SVM) Algorithm—Javatpoint*. (2022b, December 6).<https://www.javatpoint.com/machine-learning-support-vector-machine-algorithm>

Team, G. L. (2020, September 24). *What is Cross Validation in Machine learning? Types of Cross Validation*. Great Learning Blog: Free Resources What Matters to Shape Your Career!<https://www.mygreatlearning.com/blog/cross-validation/>

The Impacts of Socioeconomic Status and Educational Attainment on Youth Success. (2022, October 20). *Pathways to Education*.<https://www.pathwaystoeducation.ca/the-impacts-of-socioeconomic-status-and-educational-attainment-on-youth-success/>

Topiwalla,Mohammed. (n.d.). Machine Learning on UCI Adult data Set Using Various Classifier Algorithms And Scaling Up The Accuracy Using Extreme Gradient Boosting. *University of SP Jain School of Global Management*.

*UCI Machine Learning Repository: Adult Data Set*. (2022, September 23).<https://archive.ics.uci.edu/ml/datasets/Adult>

Zhang, B. H., Lemoine, B., & Mitchell, M. (2018). Mitigating Unwanted Biases with Adversarial Learning. *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society*, 335–340.<https://doi.org/10.1145/3278721.3278779>