*Advances in Parallel Computing Algorithms, Tools and Paradigms*
*D.J. Hemanth et al. (Eds.)*


*doi:10.3233/APC220079*

# Higher Classification Accuracy of Income Class Using Decision Tree Algorithm over Naive Bayes Algorithm

Mohamed Zaid[a] and RajendranT[b,1]
[a]*Research Scholar, Dept. of CSE, Saveetha School of Engineering, SIMATS, Chennai*
[b]*Asso. Prof., Dept. of CSE, Saveetha School of Engineering, SIMATS, Chennai*

**Abstract:** Developing two machine learning classifiers with higher accuracy for classifying income classes for people earning less and a higher salary scale between 50,000. Decision Tree Algorithm (DTA) and Naive Bayes Algorithm (NBA) are the two classifier mechanisms employed. On a dataset of 32516 records, the methods were implemented and tested. Implemented each algorithm through programs and performed ten rounds on both methods to determine distinct scales of income class for who earns lesser and higher salary scale between 50,000. The G-power test is around 80% accurate. The findings of the programming experiment showed that the Decision Tree Algorithm had a mean accuracy of 84.3790 and the Naive Bayes Algorithm had a mean accuracy of 79.3170 for classifying income categories. The variation in accuracy between the two classifiers is statistically significant (p=0.53), which is insignificant when employing the unpaired samples t-test. The primary purpose of this work is to apply a novel technique to modern Machine Learning Classifiers to forecast income class classification. When the Decision Tree Algorithm is compared to the Naive Bayes Algorithm, the results show that the DTA outperforms the NBA.

**Keywords**: Innovative Income Class Classification, Income Prediction, DTA, NBA, Machine Learning, Statistical Analysis.

## 1. Introduction

Income disparities have become a serious issue across the globe. A person's income is influenced by various things [1]. The majority of the factors influencing income are socioeconomic restrictions [2]. As a result, the significance of analysis and comprehension for income forecasting becomes critical [3]. The use of this study aids in understanding the causes of income inequality, as well as the underlying elements that influence it, such as education and marital status [4].

Nearly 54000 publications relating to Income Class Classification have been published in various indexed journals. In a research article [5], an open-source program that allows practitioners to use minimal coding to investigate, display, and analyze machine learning systems was cited 110 times. In a research article [6], a constant

[1]Rajendran T, Department of Computer Science and Engineering Saveetha School of Engineering Saveetha Institute of Medical and Technical Sciences, Chennai, Tamil Nadu, India, Email:rajendrant.sse@saveetha.com.

explanation of Machine Learning Algorithms using an Adult dataset is referred to 32 times in various research manuscripts. A research article [7] shows a model for which income is predictable from the digital footprints using machine learning, cited 18 times. This research article [8] proposes a methodology for estimating local income data in real-time, cited four times. Among these, the most relevant research article is [9], which provides an open-source program that allows practitioners to use little coding to probe, visualize, and analyze machine learning systems cited 110 times.

The previously utilized approaches do not accurately examine income prediction, are less dependable, and are ineffective in income prediction class classification. It shows the experience in our research of income class classification with knowledge of machine learning algorithms and Jupyter Lab Notebook. The proposed work is under the guidance of our department team, which has helped in this income prediction algorithm to get an accurate result in work. The study's primary goal is to improve classification accuracy by comparing the suggested algorithm Decision Tree algorithm to the current algorithm Naive Bayes algorithm, implementing creative income class classification through machine learning classifiers, and contrasting their results.

## 2. Materials and Methods

The research was conducted at the DWDM lab of the SSE, SIMATS. The two supervised learning algorithms DTA and NBA, have been taken. For these two algorithms, run two complete cycles on both algorithms, the first cycle for factors impacting income and the second for the income class category. Ten iterations have been performed on each group with ten samples to discover unique scales through the programming experiment to determine varying degrees of difficulty for income class classification. The G-power is set at 80%. The alpha error rate, which is 0.95, is a type-I error that distinguishes between the two procedures utilized. Based on the input data, the enrollment ratio in the study is around one.

### *2.1. Dataset Description*

The "adult income dataset" is the dataset that was used in this study. The data was acquired with the help of the open-source Kaggle platform. A person's annual income is governed by various factors such as educational level, age, gender, occupation, and other characteristics. 48842 occurrences (train=32561, test=16281) from this, around 44231 outlier records are removed MLC++ GenCVFiles (2/3, 1/3 random) were used to divide the data into train-test pairs. There are a total of 84 qualities. There are eight nominal qualities and six continuous attributes. Attribute Information: Age: The age of an individual. Workclass: The sector of the economy in which the individual is employed.Fnlwgt: A continuous measure that denotes the socio-economic condition of the individual. Education Number: Education level completed. Marital Status and their Relationship: Wife, Own-child, Husband, Not-in-family, Other-relative, Unmarried. Race: White, Asian-Pac-Islander, Amer-Indian-Eskimo, Other, Black. Sex: Male, Female. Capital gains: Continuous value of capital gain received that year. Capital Loss: Continuous value of capital loss incurred that year. Hours Per Week: Number of hours subject worked per week. Class: greater than or less than 50000. Missing Attribute Values: 7% of attribute values are missing.

### *2.2. Decision Tree Algorithm*

A DTA is a diagram or chart that individuals use to decide what to do next or demonstrate statistical probability. A supervised learning algorithm is a Decision Tree; this can be used for regression and classification. By learning fundamental choice rules from training data, a Choice Tree can be used to develop an innovative model that could be utilized to forecast the target variable's score.

**Pseudocode for DTA**

Input: The Collected Dataset

Output: Classifier accuracy is learned

1. The classifier should be fed the training dataset.
2. Create the dtree class.
   dtree is a kind of class.
3. Get all you need from the preceding inputs.
4. Create a new class that will be used to test the attribute.
   def assessment (test attribute)
   in the event that (end loop ifnon-vertex)
   throw accurate value
   else
   throw child[testattribute]. assessment(testattribute)
   end
5. Obtain final prediction score(test_attribute).

### *2.3. Naïve Bayes Algorithm*

We use Gaussian Naive Bayes, a variant of Naive Bayes that accepts continuous data and follows the Gaussian normal distribution. The Gaussian Nave Bayes model is more straightforward to construct and can handle big datasets.

**Pseudocode for Naive Bayes Algorithm**

Input: The Collected Dataset

Output: Classifier accuracy is learned

1. Read the dataset
2. Divide the dataset into two parts: training and testing
3. To make a prediction, calculate the mean and standard deviation
4. Repeat
   Calculate the gauss density equation on each iteration until you have calculated the probability of all Income
5. Define class
   Define Gaussian NB()
   return the accuracy
6. Accuracy.

### *2.4. Experiment Setup*

The machine learning methods were tested using the python IDE jupyter lab. The testing technique entails picking the data within it and transforming it into the format required by the classifier. The dataset should be divided into two parts: training and testing. Now utilize the training dataset to train the machine learning classification model. The classification model is evaluated using a testing dataset to establish its expected accuracy once it has been trained.

The dataset comes from Kaggle and is called Adult Income. The data is preprocessed before being used. Data cleaning removes non-essential attributes like title, subject, and date from the dataset, as well as concatenating and shuffling. The context of the dataset is revealed through data exploration.

### *2.5. Statistical Analysis*

The SPSS program does statistical calculations such as independent sample T-tests and classifier findings for various test sizes. We use the Independent Sample T-test in the SPSS application for statistical analysis. Factors such as education, age, and marital status in the training dataset are utilized as independent variables. In contrast, income higher than $50,000 and less than $50,000 is used as a dependent variable in the testing dataset. The DTA and the NBA are compared in detail, and accuracy is discovered. For two classifiers, the statistically significant variation in accuracy is 0.553. It is insignificant when employing independent samples t-tests. As a result, the statistics have improved.

## 3. Results

The Table 1, Accuracy values (DTE, NBA), where the DTA has an accuracy of roughly 84 percent and the NBA has an accuracy of approximately 79 percent.The accuracy varies depending on the decimal test size. The algorithm's accuracy varies due to a random fluctuation in the test size.

**Table 1.** Accuracy values for NBA and DTA obtained with different Test Sizes .

| **Test Size** | 0.33 | 0.55 | 0.77 | 0.88 |
|---|---|---|---|---|
| **NBA** | 79.33 | 79.43 | 79.35 | 79.33 |
| **DTA** | 84.36 | 84.40 | 84.30 | 84.39 |

In table 2, the accurateness and SD for DTA are 84.3790 and 0.08850, respectively. The Naive Bayes Algorithm's mean accuracy and standard deviation are 79.3170 and 0.16391, respectively. Decision Tree had an SD of 0.08850 with a standard error of 0.02799 in a statistical analysis of ten samples, while Naive Bayes had an SD of 0.16391 with a standard error of 0.05183. Our hypothesis was found to be negligible, with a significance value of 0.553

**Table 2.** Group Statistics, the mean precision and standard deviation for NBA and DTA.

| | DTE, NBA | N | Mean | Std. Deviation | Std. Mean Error |
|---|---|---|---|---|---|
| Accuracy | DTA | 10 | 84.3790 | 0.08850 | 0.02799 |
| | NBA | 10 | 79.3170 | 0.16391 | 0.05183 |

Table 3, With a significance of 0.553 and a standard error difference of 0.05891, the Independent Samples Test compares the accuracy of the DTA and the NBA for income class categorization. The proposed Decision Tree classifier outperformed the Naive Bayes classifier compared to the current techniques' performance.

**Table 3**. Independent Sample Test, the correlation of precision for NBA and DTA.

| | | Levene's Test for Equality of Variances (1) | Levene's Test for Equality of Variances (2) | T-test for Equality of Means (3) | T-test for Equality of Means (4) | T-test for Equality of Means (5) |
|---|---|---|---|---|---|---|
| | | F | Sig. | Std.Error Difference | 95% Confidence lower | 95% Confidence upper |
| Accuracy | Equal Variances assumed | .366 | 0.553 | 0.05891 | 4.93824 | 5.18576 |
| | Equal Variances not assumed | | | 0.05891 | 4.93824 | 5.18576 |

In Figure 1, the architecture for predicting Income Class Classification Accuracy consists of the steps included in the procedure for the prediction of Income Class Classification Accuracy. It consists of steps: Data Collection, Cleaning, Exploration, Model Classifier, Implementation, and Accuracy Prediction.

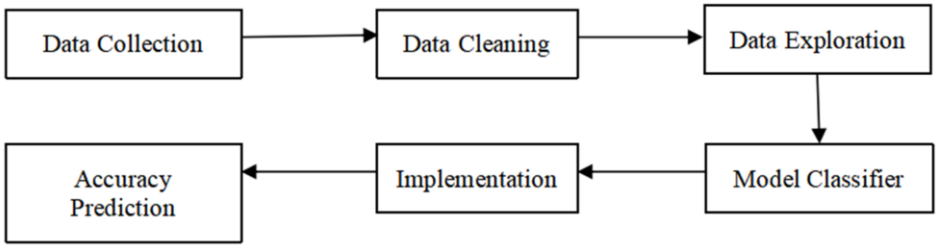


**Figure 1.** Machine learning classifier architecture.

Figure 2, Simple Bar Chart shows the obtained accuracy by DTA, NBA is 84.3790 and 79.3170, respectively, in a bar chart comparing the mean accuracy of DTA and NBA. The DTA has an error rate of 0.02799, and the NBA has an error rate of 0.02799. An independent t-test was used to assess. Two algorithms were compared for accuracy, and a statistically significant difference was found as 0.553, which is insignificant. The Decision Tree model was 84.42 percent accurate. The proposed Decision Tree classifier

outperformed the Naive Bayes classifier compared to the current techniques' performance.

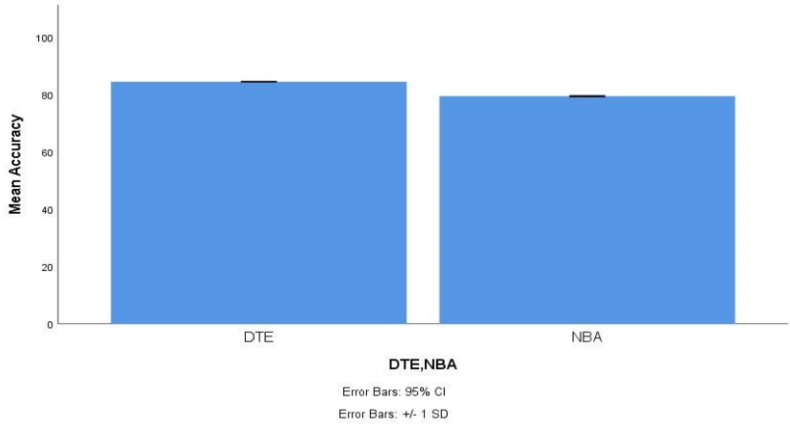


**Figure 2.** Simple Bar Chart depicts the Mean of Accuracy by NBA and DTA.

## 4. Discussion

The Decision Tree Algorithm outperforms the Naive Bayes Algorithm in terms of accuracy. The information is obtained throughout multiple research cycles to define various accuracy rate scales. The data is analyzed using a t-test with independent samples. The Decision Tree has a greater accuracy of about (84 percent) in this investigation of innovative income class classification than the Naive Bayes Algorithm (79 percent ). When compared to the insignificant independent samples t-test, the Decision Tree Algorithm has a higher significance of 0.553 [10].

The Decision Tree Algorithm's mean accuracy and standard deviation are 84.3790 and 0.08850, respectively. The study [11] shows the performance in detection with an accuracy of 90% compared to the Decision Tree, which has 85.87% accuracy. However, according to [12], The SVM is 91% more accurate than the Decision Tree. A literature review has determined that the Decision Tree Algorithm outperforms the Naive Bayes Algorithm in terms of accuracy.

When using the SPSS statistical tool to do independent sample tests, the statistically significant difference in accuracy between the two algorithms is $p>0.05$, which is insignificant. The SPSS statistical program also calculates the mean and standard deviation [13]. The standard error difference defines the error level of the Decision Tree Algorithm, which has an error rate of 0.02799, and the Naive Bayes Algorithm, which has an error rate of around 0.05183. With a recall score of 0.942, classifier Decision Trees had the best mistake rate in previous research. With a recall score of 0.63, the XGBoost classifier finished second [14].

## 5. Conclusion

The findings of this study reveal that an innovative income class categorization utilizing Decision Tree Algorithms has a higher accuracy of 84 percent than Naive Bayes Algorithms, which have a lower accuracy of 79 percent. According to past research studies, the classifiers utilized, the DTA has superior accuracy to the NBA. The dataset provides only a few signs that may be utilized to determine accuracy percentages for unique income class classification, which is the study's principal flaw. The higher the number of independent and dependent variables, the better the accuracy. The dataset will have numerous properties in the future that will allow the classifier to perform successfully and improve the accuracy of income prediction. More attributes can be included to improve the accuracy percentage and get an increased precision score.

## References

[1] Lazar A. Income prediction via support vector machine. 2004 International Conference on Machine Learning and Applications, 2004. Proceedings. DOI: 10.1109/icmla.2004.1383506.
[2] Mothilal RK, Sharma A, Tan C. Explaining machine learning classifiers through diverse counterfactual explanations. Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency. Epub ahead of print 2020. DOI: 10.1145/3351095.3372850.
[3] Farrell D, Greig F, Deadman E. Estimating Family Income from Administrative Banking Data: A Machine Learning Approach. AEA Papers and Proceedings 2020; 110: 36–41.
[4] Chakrabarty N, Biswas S. A Statistical Approach to Adult Census Income Level Prediction. 2018 International Conference on Advances in Computing, Communication Control and Networking (ICACCCN). Epub ahead of print 2018. DOI: 10.1109/icacccn.2018.8748528.
[5] Wexler J, Pushkarna M, Bolukbasi T, et al. The What-If Tool: Interactive Probing of Machine Learning Models. IEEE Trans Vis Comput Graph 2020; 26: 56–65.
[6] Goel N, Yaghini M, Faltings B. Non-Discriminatory Machine Learning through Convex Fairness Criteria. Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society. Epub ahead of print 2018. DOI: 10.1145/3278721.3278722.
[7] Matz SC, Menges JI, Stillwell DJ, et al. Predicting individual-level income from Facebook profiles. PLoS One 2019; 14: e0214369.
[8] Ivan K, Holobâcă I-H, Benedek J, et al. VIIRS Nighttime Light Data for Income Estimation at Local Level. Remote Sensing 2020; 12: 2950.
[9] Wexler J, Pushkarna M, Bolukbasi T, et al. The What-If Tool: Interactive Probing of Machine Learning Models. IEEE Trans Vis Comput Graph 2020; 26: 56–65.
[10] Bekena SM. Using decision tree classifier to predict income levels, https://mpra.ub.uni-muenchen.de/83406/ (2017, accessed 9 August 2021).
[11] Carrillo A, Cantú LF, Tejerina L, et al. Individual Explanations in Machine Learning Models: A Case Study on Poverty Estimation, http://arxiv.org/abs/2104.04148 (2021, accessed 9 August 2021).
[12] A machine learning approach to improving occupational income scores. Explor Econ Hist 2020; 75: 101304.
[13] Liu L, Liu P, Liu J, et al. Unfold income myth: Revolution in income models with advanced machine learning techniques for better accuracy. Model Assisted Statistics and Applications 2018; 13: 319–327.
[14] Narodytska N, Shrotri A, Meel KS, et al. Assessing heuristic machine learning explanations with model counting. In: Lecture Notes in Computer Science. Cham: Springer International Publishing, 2019, pp. 267–278.