

Comparative Review of Machine Learning Methods in Revenue Forecasting

Ali Taghiyev^{1*}, Özkan İnik²

¹ Faculty of Computer Engineering Tokat Gaziosmanpaşa University, Tokat, Turkey ali.taghiyev3424@gop.edu.tr

² Faculty of Computer Engineering Tokat Gaziosmanpaşa University, Tokat, Turkey ozkan.inik@gop.edu.tr

Abstract – In this study, a classification problem is discussed to predict whether individuals' annual income levels are above \$50,000. The UCI Adult Income Dataset, which is derived from the 1994 US Census (Census) data and is widely used in the field of machine learning, was preferred as the data source. This dataset contains 48,842 samples and 14 independent variables; It covers categorical and numerical characteristics such as age, educational status, marital status, profession. In addition, some categorical variables in the dataset have missing values and the classes of the target variable are unbalanced. Individuals with incomes below 50K account for 76%, while those above 24%. Therefore, in order to interpret model performance more accurately, metrics such as the ROC curve and AUC (Area Under Curve) were taken into account as well as the accuracy rate. In the study, after the data analysis and preprocessing process was completed, four different classification algorithms were applied: Logistic Regression, Support Vector Machines (SVM), Random Forest and Naive Bayes. The Logistic Regression model achieved an accuracy rate of 83.2%, while the SVM algorithm showed a moderate performance with an accuracy of 80.3%. While the Naive Bayes algorithm was moderately successful with an accuracy of 78.7%, the highest accuracy rate was obtained in the Random Forest algorithm with 85.1%. These results revealed that despite the unbalanced class distribution, the Random Forest algorithm is an effective method in such classification problems thanks to its high accuracy.

Keywords — Machine Learning, Income Estimation, Income Classification, UCI Adult Dataset, Data Analysis, Logistic Regression, Support Vector Machine, Random Forest, Naive Bayes, Classification Algorithms, Model Performance Comparison, Accuracy, Income Inequality, Data Mining, Prediction Models

1 INTRODUCTION

The massive increase in society's dependence on data and information in recent years has led to significant advances in technologies developed for storing, analyzing, and processing large-scale data. Machine learning (ML) and data mining techniques are widely used to gain information and predict future events that are difficult to observe. [1] One of the social problems that has been emphasized recently is income inequality. The main goal here is not only to reduce poverty, but also to understand the main factors that lead to economic inequality.

Despite the success of machine learning in solving complex predictive problems, recent studies show that the most advanced models exhibit significantly lower accuracy in minority groups compared to majority groups. This problem of bias reveals the need for careful model selection and evaluation, especially in socioeconomic practices where fairness and accuracy are critical.

Gross Domestic Product (GDP) is a key economic indicator that measures the total monetary value of goods and services produced within a country's borders during a given period. Economists use GDP to assess economic health, growth trends, and inflation-deflation effects.

However, GDP does not directly reflect the distribution of individual income, which is a critical component of economic well-being. [2] Therefore, estimating income levels based on demographic and socioeconomic factors offers valuable insights into economic inequalities.

In this study, using various machine learning algorithms, over the Modified UCI Adult Data Set It is estimated whether the annual income of individuals exceeds \$50,000. [3]

The research addresses the following key objectives:

- Analysis of income levels using demographic characteristics and determination of influencing factors.
- Evaluation and selection of the best performing models based on accuracy, precision, sensitivity, F1-score, and AUC-ROC [4] metrics.

Logistic Regression [5], Support Vector Machines (SVM) [6], Random Forest [7] and Naive Bayes [8] algorithms were applied and their performances were compared.

Our findings reveal that the Random Forest algorithm performed best with the highest accuracy (85.1%), followed by Logistic Regression (83.2%) and SVM (80.3% accuracy), and the lowest performance with the Naive Bayes algorithm with 78.7% accuracy.

1.1. Literature Review

Deepika et al. [12] analyze the performance of different machine learning classification models to predict the annual income levels of adults. Individuals are divided into two groups based on their annual income, those who earn more than \$50,000 and those who earn less than or equal to \$50,000. Another aim of the study is to determine the main factors affecting the income level. Decision Tree, Artificial Neural

Networks (ANN), Support Vector Machines (SVM), k-Nearest Neighbors (KNN), Naive Bayes and Ensemble methods were used. Random Forest, AdaBoost and Logistic Regression were applied as ensemble methods. As a result of the analysis, it was observed that the AdaBoost model reached the highest F-measure (0.6834%), precision (0.7708%) and accuracy (0.8637%) values. Artificial Neural Networks also showed high performance but lagged behind AdaBoost. Most models have achieved an accuracy of over 84%. The Support Vector Machine model, on the other hand, exhibited lower accuracy than other models. As a result, it is concluded that Ensemble methods are more successful in this classification problem.

M. A. Islam et al. [13] aimed to estimate the annual income levels of adults based on demographic characteristics and used the modified UCI Adult Dataset. Revenue was divided into two categories, $\leq 50K$ and $> 50K$, and 14 personal characteristics were used for the prediction. Classification algorithms such as Logistic Regression, Naive Bayes, K-Nearest Neighbors (KNN), Support Vector Machines (SVM), Decision Trees, Random Forest, XGBoost, Artificial Neural Networks (ANN) were used in the study. In addition, the ensemble model, which is a combination of XGBoost and ANN, showed the highest performance with 87% accuracy. XGBoost alone achieved an accuracy of 86.89%. Logistic Regression was 84.6%, Decision Trees was 85.12% and Random Forest was 85.56%. K-means and DBSCAN were used as clustering algorithms, but it was stated that these methods were not recommended in terms of fairness. In addition, PCA was used as the Size Reduction Algorithm. The study emphasizes the importance of data preprocessing, choosing the right algorithm, and optimization. As a result, demographic characteristics have proven to be influential in revenue forecasting.

E. E. Moe et al. [14]: Naive Bayes analyzes the performance of classification algorithms such as Decision Tree J48, and Random Forest. Information Gain, Gain Ratio, Pearson Correlation, and ReliefF were used as Feature Selection Algorithms. Experiments have revealed that the Decision Tree J48 classifier outperforms the other two classifiers in terms of accuracy and classification time. J48 achieved 1% better results than Random Forest and 3% better than Naïve Bayes. All three classifiers showed high success in correctly classifying the " $\leq 50K$ " income class, but the accuracy rate for the " $> 50K$ dollars" class was as low as 62% with Random Forest. According to the attribute importance analysis, features such as workclass, native.country, race, and fnlwgt were found to be less important for prediction. Removing these features has often reduced the accuracy of classifiers. Naïve Bayes achieves the lowest Mean Absolute Error (MAE), while the Root Mean Square Error (RMSE) rate is higher than J48 and Random Forest. J48 achieved the lowest RMSE and had a slightly higher MAE ratio than Naïve Bayes. As a result, Decision Tree J48 has emerged as the most appropriate classifier for this type of census data.

L.-P. Chen. [15]: This study aims to estimate the income level (whether or not they are over \$50K) of U.S. adults based on social factors. Logistic Regression, Linear Discriminant Analysis (LDA), Quadratic Discriminant Analysis (QDA), Support Vector Machines (SVM), Random Forest, and Boosting algorithms were used in the study. Logistic Regression presents a linear and interpretable model, while LDA and QDA assume a normal distribution of intraclass

variables. SVM classifies in high-dimensional space using a linear kernel, while Random Forest makes decisions based on the majority vote principle. The boosting algorithm tries to reduce the error rate with successive trees. The results revealed that the Boosting method performed best, with an accuracy of 86.33% and an AUC of 0.9192. Logistic Regression, SVM and Random Forest were successful with an AUC $>$ of 0.9. The performance of LDA and QDA remained poor due to violation of model assumptions. According to the boosting model, the variables that most affected the income level were relationship and capital.gain. As a result, the Boosting method stood out as the most effective model and social factors were successfully identified

This study by Wan, Z. [16] aims to compare machine learning models to predict the income level of adults. In the study, the Adult Income Dataset of 32,561 adults was used, which was taken from Kaggle and included 15 characteristics such as age, education level, and occupation. The dataset is divided into a 7:3 ratio for training and testing. In the study, Decision Trees, Random Forests and Neural Networks models were trained and compared. Data preprocessing steps included quantification of categorical data, correlation analysis, and normalization. The performance of the models was evaluated with accuracy, recall and F1 score metrics. The results showed that the Random Forest model performed best with an accuracy of 81.5%, a recall score of 0.77 and a positive F1 score of 0.63. The Decision Tree model achieved 79.6% accuracy and 0.72 recall, while Neural Networks achieved 79.8% accuracy and 0.72 recall. According to the trait importance analysis, it was determined that the most effective factor on income was the year of education. It has been observed that the proportion of high-income individuals increases with increasing age. The failure of Neural Networks to perform as expected has been associated with weak correlations between the 2D structure of the dataset and features. The success of Random Forest, on the other hand, is due to its ability to combine the predictions of multiple decision trees and prevent overfitting. As a result, it is recommended to use the Random Forest model for revenue forecasting in this dataset.

2 MATERIALS AND METHOD

2.1. Machine Learning Algorithms

Logistic Regression: It is a statistical model used especially for binary classification problems. The model uses the logit function to predict the relationship between the independent variables and the dependent variable. It excels especially well in linear relationships, and its output specifies the probability of a class [5].

SVM (Support Vector Machine): SVM is a classification algorithm that aims to maximize the widest margin (difference) between classes. By classifying data points in high-dimensional space, it can also be effective in nonlinear data. However, it can experience performance issues with very large data sets or unbalanced classes [6].

Random Forest: Random Forest is an ensemble learning method consisting of a combination of decision trees. The final result is obtained by a majority vote of the outputs of many decision trees. This method is used to reduce the problem of overfitting and to increase the ability to generalize [7].

Naive Bayes: Naive Bayes is a classification algorithm based on Bayes' theorem and assumes that properties are independent of each other. It is preferred in large datasets due to its fast and efficient operation, but dependencies between features can be ignored [8].

2.2. Dataset

Census Income Dataset (Adult Dataset) was selected as the data set. It is derived from data from the 1994 U.S. Census (Census) and is a popular dataset for the classification problem with machine learning.

The main features of the dataset are as follows:

- Objective: To estimate whether an individual's annual income is more than 50K
- Transaction Type: Classification problem
- Feature Types: Categorical and numeric variables
- Total Samples: 48,842
- Total Number of Features: 14 (Independent variable) + 1 (Target variable)
- Missing Data: Yes (Some categorical variables have missing values)
- Class imbalance: If the number of instances of the classes in the dataset is very different from each other, this dataset will be unbalanced. [9]

| index | age | workclass | fnbgt | education | education-num | marital-status | occupation | relationship | race | sex | capital-gain | capital-loss | hours-per-week | native-country | income |
|-------|-----|--------------|--------|------------|---------------|-----------------------|-------------------|---------------|-------|--------|--------------|--------------|----------------|----------------|--------|
| 0 | 39 | State-gov | 77536 | Highschool | 9 | Never-married | Adm-clerical | Not-in-family | White | Male | 2174 | 0 | 40 | United-States | <=50K |
| 1 | 50 | Self-emp-inc | 83311 | Highschool | 9 | Married-civ-spouse | Exec-managerial | Married | White | Male | 0 | 0 | 51 | United-States | >50K |
| 2 | 38 | Private | 216048 | HS-grad | 9 | Divorced | Handlers-cleaners | Not-in-family | White | Male | 0 | 0 | 40 | United-States | <=50K |
| 3 | 52 | Private | 236271 | 10th | 7 | Married-civ-spouse | Handlers-cleaners | Married | Black | Male | 0 | 0 | 40 | United-States | <=50K |
| 4 | 29 | Private | 138469 | Highschool | 9 | Married-civ-spouse | Prof-specialty | Wife | Black | Female | 0 | 0 | 40 | Cuba | <=50K |
| 5 | 32 | Private | 296282 | Master | 14 | Married-civ-spouse | Exec-managerial | Wife | White | Female | 0 | 0 | 40 | United-States | >50K |
| 6 | 49 | Private | 102107 | 10th | 7 | Married-spouse-absent | Other-service | Not-in-family | Black | Female | 0 | 0 | 40 | United-States | <=50K |
| 7 | 52 | Self-emp-inc | 309642 | HS-grad | 9 | Married-civ-spouse | Exec-managerial | Married | White | Male | 0 | 0 | 40 | United-States | >50K |
| 8 | 31 | Private | 47091 | Master | 14 | Never-married | Prof-specialty | Not-in-family | White | Female | 18284 | 0 | 50 | United-States | >50K |
| 9 | 42 | Private | 158462 | Highschool | 9 | Married-civ-spouse | Exec-managerial | Married | White | Male | 2124 | 0 | 40 | United-States | >50K |

Fig. 1 Contents of the dataset

Figure 1 shows sample records of the data set used in the study. This dataset includes the demographic and socioeconomic characteristics of individuals. Each line represents an individual; It covers variables such as age, gender, place of birth, education level, marital status, occupation, race, weekly working time, capital gain/loss. The "income" column is the target variable of the classification problem and indicates whether an individual's annual income is above or below \$50,000. The other columns are the inputs used to predict this target variable.

There are 2 classes in this dataset: those with incomes up and down from 50k. Fig 2 shows the percentage distributions of these classes. The number of instances of these classes is 76% and 24%. That is, the data set is unbalanced. In this case, attention will also be paid to the ROC curve and AUC values for a more accurate examination of the results [10]. If the AUC value is higher than 0.5, the model predicts correctly, if it is 0.5 or lower, the model does not predict or produces completely incorrect results.

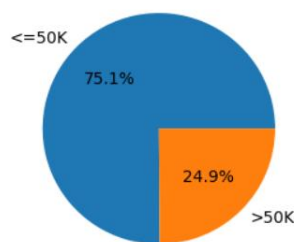


Fig. 2 Income Distribution

2.3. Performance Metrics

A quantitative evaluation using performance metrics is required to compare the results of the models. Metrics such as accuracy, precision, recall, F1-score, and area under the ROC curve (AUC-ROC) are commonly used methods to measure the success of classifiers [17], [18], [19], [20].

Accuracy: refers to the ratio of all correct predictions of the model to the total predictions. This metric is a simple and common method of measuring the overall accuracy of the model. Accuracy can be highly effective, especially in balanced data sets, but it can be misleading in data sets with unbalanced class distributions. For example, even if the model correctly predicts the majority class, not correctly predicting the minority class can keep the accuracy high. Therefore, accuracy, while only reflecting general accuracy, is not sufficient on its own to assess whether each class was correctly predicted [17].

Precision: Measures how much of the positive class predictions are correct. That is, it shows how many of the outcomes that the model predicts as positive are actually positive. Precision is especially critical in situations where false positives are costly, such as cancer diagnosis or fraud detection. High precision means that the model's positive predictions are reliable, meaning that there are very few false positives. However, because precision focuses only on false positives, it can overlook the model's ability to capture all the correct positives.

Recall: Recall shows how accurately the model predicted all the true positives. In other words, it measures how much of the true positives you correctly identify. Recall is used as an important metric, especially in cases where false negatives come at a great cost (for example, a false-negative cancer test result for a patient). A high recall indicates that the model's ability to detect positive class is strong, but false positives can also increase. Recall is especially prominent in applications where the detection of missing positives is critical.

F1-Score: is the harmonic mean of precision and recall and summarizes the overall performance of the model by balancing both metrics. This metric is especially useful when it is necessary to strike a balance between precision and recall. If you only want to use one metric, and both metrics are important, F1-Score is a good choice because it compensates for both false positives and false negatives. F1-Score favors models that perform well in both metrics. A low F1-Score indicates that the model produces too many false positives or false negatives [18].

ROC Curve (Receiver Operating Characteristic Curve): is a technique used to visually evaluate the performance of the classification model. This curve shows how the True Positive Rate (TPR) and False Positive Rate (FPR) values change at different classification threshold values of the model. TPR (recall) and FPR provide important information about the accuracy of the model. Each point on the ROC curve reflects the performance achieved with a different classification threshold. The curve visualizes the ability to correctly detect positive classes and reduce false positives, regardless of which threshold the model uses. The closer the curve is to the left and the wider its top, the more successful the model is. The ideal state of the ROC curve is those where only true positive classes are predicted and false positives are kept to a minimum. The ROC curve provides an effective performance evaluation of the model even in unbalanced data sets [19].

AUC (Area Under the Curve): AUC represents the area under the ROC curve and expresses the overall classification success of the model with a numerical value. AUC has a value between 0 and 1; An AUC close to 1 indicates the presence of a perfect pattern, while an AUC of 0.5 indicates that the model is predicting randomly. AUC measures the model's ability to correctly predict positive classes and prevent false positives. AUC is especially important in the case of imbalance between classes, as it accurately reflects the overall success of the model, even in cases where classes are close to each other or unbalanced. A high AUC value indicates that the model both correctly predicts the positive class and minimizes false positives. AUC values of 0.7 and above are generally considered success, but an AUC value of 0.9 and above is a sign of a very good model [19].

3 EXPERIMENTAL SETUP

3.1. Examining the Data Set

In this section, the data set was analyzed in detail, the necessary data pre-processing steps were applied and the model was prepared for training. Various visualization techniques have been used to better understand the data set. The findings and graphs obtained as a result of the analyzes are presented below.

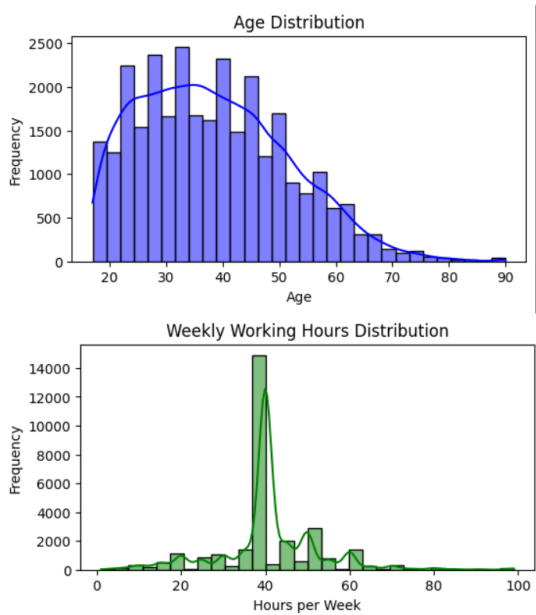


Fig.3 Age and Working Hours Distribution

In Fig 3, the graphs showing the distribution of demographic characteristics in the data set are examined. The age distribution is in the form of a bell curve with a peak in the 20-90 age group and in the 30-40 age band. In the distribution of weekly working hours, a significant concentration is observed in 40 hours, which is the standard full-time employment. Secondary densities were formed in time zones 20 and 60. These two variables have a potentially important place in revenue forecasting modeling.

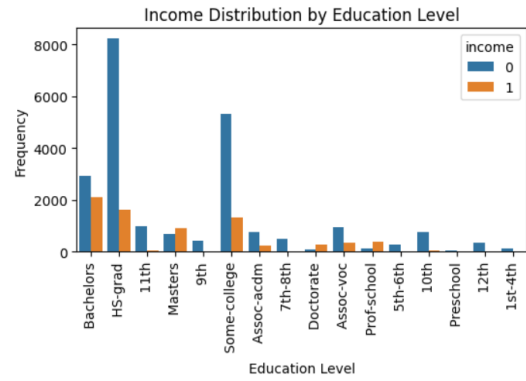


Fig. 4 Distribution by educational attainment

According to Fig 4, the highest number of low-income (Income 0) individuals is seen in "HS-grad" and "Some-college" education levels. The education level with the highest number of high-income (Income 1) individuals draws attention as "Bachelors". In general, it is observed that as the level of education increases, the proportion of high-income (Income 1) individuals also increases. At low levels of education (e.g., "1st-4th", "5th-6th", "Preschool"), the number of high-income individuals is very small. This shows that the level of education is an important factor on the level of income, but it is not a determinant on its own.



Fig. 5 Income and Gender distribution

Figure 5 shows the distribution of income by gender. The graph shows that men are represented in greater numbers than women in both the low (0) and high (1) income groups. Among men, the number of low-income individuals is higher than among high-income individuals. Similarly, the number of low-income individuals is significantly higher among women than among high-income individuals. In general, it is observed that low-income individuals are dominant in both genders, but men are more numerous than women in the high-income group.

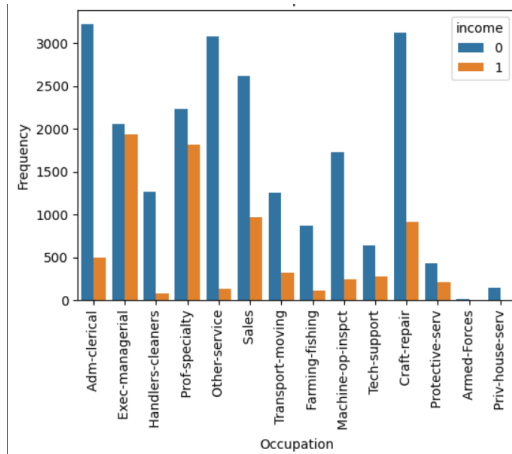


Fig.6 Distribution of Income and Occupation

Figure 6 summarizes the distribution of income in different occupational groups. Each pair of bars on the graph shows the frequency of low-income (blue - 0) and high-income (orange - 1) individuals in a given occupational group. In general, it is seen that the number of low-income employees is significantly higher than the number of high-income employees in many occupational groups. Especially in professions such as "Adm-clerical", "Other-service" and "Craft-repair", this difference is quite striking.

However, it is observed that the number of high-income individuals is closer to that of low-income individuals in more specialized professions such as "Exec-managerial" and "Prof-specialty"

3.2. Preparation of Test and Training data

Afterwards, the data is divided into 2 different sets: 70% Training and 30% Testing. The reason for the 70/30 separation of the data set is that the best results are obtained at this rate during model training in the studies [11]. This distinction allows us to evaluate the performance of our model on data that it has not seen before.

- **X:** Input (features) data are features or arguments used for the model to learn. This data helps the model learn patterns so that it can make accurate predictions.
- **A:** Target data is the target values or dependent variables that the model is trying to predict. That is, the results that the model is trying to learn about.

Fig 7 shows the distribution of training and test sets.

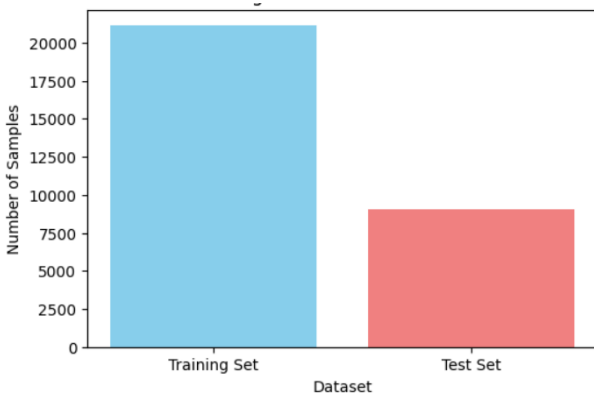


Fig.7 Training and Test data distribution

3.3. Experimental Results

In the analysis, the performance of different machine learning algorithms in predicting the annual income levels of adults was evaluated. The results reveal that each algorithm has different strengths and various factors that should be considered in model selection.

Logistic Regression performed successfully with a value of 0.832 in terms of accuracy. The Precision value has a good precision of 0.720, while the recall value is 0.547, meaning that the rate at which the model correctly predicts the positive class has remained relatively low. The F1-Score is 0.622, indicating a reasonable balance between precision and recall.

The Fig 8 chart shows that the ROC curve of the Logistic Regression model performs quite well thanks to the curve curving towards the upper left corner and having an AUC value of 0.885. This indicates that the model is good at distinguishing between positive and negative classes and is able to achieve a high true positive rate with a low false positive rate.

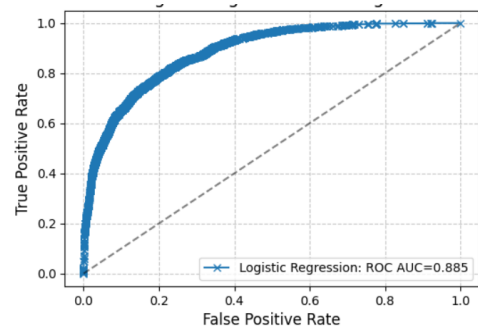


Fig.8 Logistic Regression – ROC Curve

SVM (Support Vector Machines) is particularly notable for its high recall value (0.833). This suggests that the model is able to capture positive classes largely accurately. However, the precision value is relatively low at 0.575; This indicates that the model makes some false positives.

Accuracy: 0.803 and F1-score: 0.681 indicate that the model provides a good balance overall. The ROC AUC score is 0.631, indicating that the model is able to make a reasonable distinction between classes.

As a result, the SVM model can be considered as a generally balanced and feasible model with high sensitivity (recall) for this problem. It can be preferred especially in scenarios where erroneously negative classification is critical.

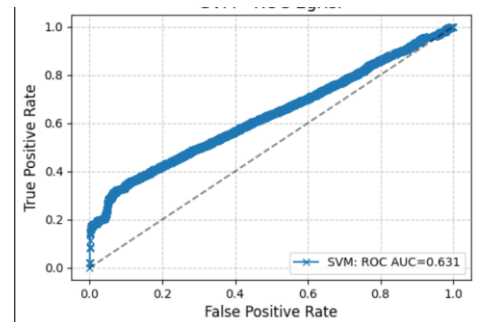


Fig.9 SVM – ROC Curve

Random Forest stands out for its high accuracy value and has achieved a solid result of 0.851. The precision and recall values are fairly stable at 0.737 and 0.637, respectively, indicating the model's ability to accurately predict both classes. The F1-Score value is 0.683. The ROC is the model with the highest AUC value of 0.902, indicating that Random Forest is effective in all classification tasks.

When the ROC curve of the Random Forest model in Fig 10 is examined, it is seen that the curve is significantly closer to the upper left corner. This indicates that the model exhibits a high classification performance. The area under the curve (AUC) value was calculated as 0.902. This high AUC value indicates that the model's ability to distinguish between positive and negative classes is quite good and makes generally successful predictions. As a result, the Random Forest model provides a reliable and effective solution to this particular classification problem.

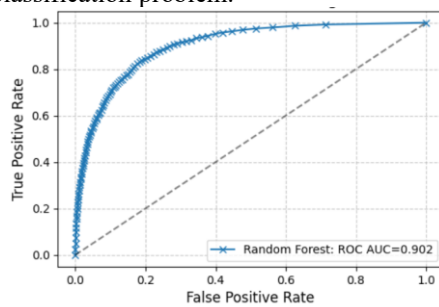


Fig.10 Random Forest – ROC Curve

Naive Bayes is particularly strong in terms of precision (0.666), performing reasonably well with an accuracy of 0.787. However, the recall value is only 0.310, which indicates that the model is not capturing positive classes well enough. The F1-Score value is very low at 0.423, and there is a significant imbalance between precision and recall. According to the graph in Fig 11, the ROC AUC value is 0.824, indicating that Naive Bayes is still a valid model, but further improvements need to be made.

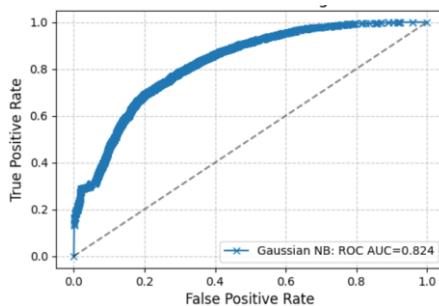


Fig.11 Naive Bayes – ROC Curve

As a result, Random Forest is the best performing model in this data set, with high accuracy, precision, recall and F1-Scores. Logistic Regression is also a good option, providing a solid balance between unbalanced classes. Naive Bayes and SVM are weaker options for this problem, especially with their low precision and ROC AUC values.

Table 1. Results Table

| Classifier | Accuracy | Precision | Recall | F1 | Roc Auc |
|--------------|----------|-----------|--------|-------|---------|
| Logistic_reg | %83.2 | %72.0 | %54.7 | %62.2 | 0.865 |
| SVM | %80.3 | %57.5 | %83.3 | %68.1 | 0.631 |
| R_Forest | %85.1 | %73.7 | %73.7 | %68.3 | 0.902 |
| N_Bayes | %78.7 | %66.6 | %31.0 | %42.3 | 0.824 |

As can be seen from the results in Table 1, the Random Forest algorithm showed the highest performance in all metrics. Despite its high recall value (83.3%), the SVM model performed poorly compared to other models in terms of accuracy (80.3%) and precision (57.5%). Logistic Regression and Naive Bayes algorithms performed moderately. Especially when the ROC AUC values are examined, it is seen that Random Forest has the strongest classification ability with 0.902.

4 DISCUSSION

In this study, the performance of four different machine learning algorithms was compared using the UCI Adult Dataset to predict whether individuals' annual income levels were above \$50,000. The results obtained offer important implications for how different algorithms behave, especially in datasets with unbalanced class distributions.

The fact that the **Random Forest** algorithm achieves the highest accuracy (85.1%) and ROC AUC (0.902) values reveals the advantages of ensemble methods in such classification problems. Behind the success of Random Forest lies in the fact that the combined estimations of multiple decision trees reduce variance and prevent overfitting. In addition, the ability of this algorithm to work effectively with both categorical and numerical features has made it easier to deal with the heterogeneous structure in the dataset. However, Random Forest's high computational cost and relatively low model interpretability may be a disadvantage in some applications.

Logistic Regression has demonstrated a balanced performance with an accuracy rate of 83.2%. The simplicity and high interpretability of the model provide an advantage, especially for researchers who want to understand the impact of socioeconomic factors on income. However, the recall value (54.7%) was relatively low, indicating that the model had difficulty in detecting the positive class (high-income individuals). This may be due to a class imbalance in the data set. The fact that logistic regression is a linear model has caused it to be limited in capturing complex and nonlinear relationships.

The Naive Bayes algorithm has shown reasonable performance with an accuracy rate of 78.7%. However, the recall value (31.0%) was quite low, indicating that the model was insufficient to detect the positive class. This may be due to the fact that Naive Bayes' assumption of independence between traits is not valid in the actual data set. In particular, the presence of correlations between traits in demographic data reduced the performance of this algorithm. However, the fact that Naive Bayes is computationally efficient and can work quickly on large data sets can be an advantage for some applications.

It is noteworthy that the **SVM** algorithm performs at a good level with an accuracy rate of 80.3%. However, high recall (83.3%) and low precision (57.5%) values of SVM indicate that the model predicts positive samples largely correctly, but there are also many false positives. This may be due to the fact that the margin between classes cannot be fully optimized. In addition, SVM's dependency on certain kernel functions and the unstable structure of the dataset are among the factors that limit the performance of the model.

Another important finding of the study is that evaluating the accuracy metric alone in unbalanced data sets can be misleading. In particular, the fact that the SVM model has a high recall but low precision value reveals the necessity of evaluating multiple metrics together in classification problems. The ROC AUC metrics confirmed that the Random Forest algorithm is more balanced and performs superior compared to other models.

In conclusion, this study showed that algorithm selection is critical in socioeconomic classification problems such as income estimation. Random Forest's high performance has proven that ensemble methods are effective in such complex and imbalanced data sets. In future studies, it may be recommended to apply sampling strategies (oversampling/undersampling) or to test different ensemble methods (XGBoost, LightGBM) to eliminate class imbalance. In addition, using methods such as SHAP (SHAP Additive exPlanations) to increase model interpretability can provide a clearer understanding of the factors that lead to income inequality.

5 CONCLUSION

In this study, various machine learning algorithms were compared to predict the annual income levels of individuals using the UCI Adult Data Set. Among the four algorithms implemented, the Random Forest model showed the highest success with an accuracy rate of 85.1%. Logistic Regression and Naive Bayes achieved 83.2% accuracy and 78.7%; The SVM algorithm performed well with an accuracy of 80.3%. Due to the uneven class distribution of the data set, not only the accuracy rate but also additional metrics such as the ROC curve and AUC values were evaluated. The results reveal that the Random Forest algorithm is a strong choice for such data sets in terms of both accuracy and stability in classification problems.

REFERENCES

- [1] Sarker, I.H. Machine Learning: Algorithms, Real-World Applications and Research Directions. *SN COMPUT. SCI.* 2, 160 (2021). <https://doi.org/10.1007/s42979-021-00592-x>
- [2] Stiglitz, J. E., Sen, A., & Fitoussi, J.-P. (2009). *Report by the Commission on the Measurement of Economic Performance and Social Progress*.
- [3] Kohavi, R. (1996). Census Income [Dataset]. UCI Machine Learning Repository. <https://doi.org/10.24432/CSGP7S>.
- [4] Powers, D. M. W. (2011). Evaluation: From precision, recall and F-measure to ROC, informedness, markedness & correlation. *Journal of Machine Learning Technologies*, 2(1), 37-63. <https://doi.org/10.9735/2229-3981>
- [5] Berkson, J. (1944). Application of the Logistic Function to Bio-Assay. *Journal of the American Statistical Association*, 39(227), 357–365. <https://doi.org/10.2307/2280041>
- [6] Cortes, C., Vapnik, V. Support-vector networks. *Mach Learn* 20, 273–297 (1995). <https://doi.org/10.1007/BF00994018>
- [7] Ho, T. K. (1995). Random Decision Forests. *Proceedings of the Third International Conference on Document Analysis and Recognition*, 1, 278-282. <https://doi.org/10.1109/ICDAR.1995.598994>
- [8] Frank, E., Trigg, L., Holmes, G. et al. Technical Note: Naive Bayes for Regression. *Machine Learning* 41, 5–25 (2000). <https://doi.org/10.1023/A:1007670802811>
- [9] X. Guo, Y. Yin, C. Dong, G. Yang and G. Zhou, "On the Class Imbalance Problem," *2008 Fourth International Conference on Natural Computation*, Jinan, China, 2008, pp. 192-201, doi: 10.1109/ICNC.2008.871.
- [10] Soupcioglu ŞK, Aksel G. Receiver operating characteristic curve analysis in diagnostic accuracy studies: A guide to interpreting the area under the curve value. *Turk J Emerg Med.* 2023 Oct 3; 23(4):195-198. doi: 10.4103/tjem.tjem_182_23. PMID: 38024184; PMCID: PMC10664195.
- [11] Q. H. Nguyen et al., "Influence of Data Splitting on Performance of Machine Learning Models in Prediction of Shear Strength of Soil," *Mathematical Problems in Engineering*, vol. 2021, Article ID 4832864, 2021, doi: 10.1155/2021/4832864.
- [12] D. Deepika, A. Vijaya Lakshmi, P. G. Sravya, P. S. P. Abhiram, P. C. Chandu, and P. S. Kiran, "Performance analysis of machine learning classification models for predicting the adult income levels," in *Proc. 2023 Global Conf. Inf. Technol. Commun. (GCITC)*, Bangalore, India, Dec.2023,doi:10.1109/GCITC60406.2023.10425912.
- [13] M. A. Islam, A. Nag, N. Roy, A. R. Dey, S. M. F. A. Fahim, and A. Ghosh, "An investigation into the prediction of annual income levels through the utilization of demographic features employing the modified UCI adult dataset," in *Proc. 2023 Int. Conf. Comput., Commun. Intell. Syst. (ICCCIS)*, Greater Noida, India, Nov. 2023, doi: 10.1109/ICCCIS60361.2023.10425394.
- [14] E. E. Moe, S. S. M. Win, and K. L. L. Khine, "Adult income classification using machine learning techniques," in *Proc. 2023 IEEE Conf. Comput. Appl. (CCA)*, Yangon, Myanmar, Feb. 2023,doi:10.1109/CCA51723.2023.10181907.
- [15] L.-P. Chen, "Supervised learning for binary classification on US adult income," *Journal of Modeling and Optimization*, vol. 13, no. 2, pp. 80–90, Dec. 2021, doi:10.32732/jmo.2021.13.2.80.
- [16] Wan, Z. (2023). Performances evaluation of machine learning models on income forecasting. *Applied and Computational Engineering*, 27(1), 24-29. <https://doi.org/10.54254/2755-2721/27/20230111>
- [17] Labatut, V., & Cherifi, H. (2011, July). *Accuracy measures for the comparison of classifiers*. 5th International Conference on Information Technology (ICIT), Amman, Jordan. arXiv:1207.3790. <https://doi.org/10.48550/arXiv.1207.3790>
- [18] Lipton, Z. C., Elkan, C., & Narayanaswamy, B. (2014). Thresholding Classifiers to Maximize F1 Score. arXiv preprint arXiv:1402.1892. <https://doi.org/10.48550/arXiv.1402.1892>
- [19] Thambawita, V., Jha, D., Hammer, H. L., Johansen, H. D., Johansen, D., Halvorsen, P., & Riegler, M. A. (2020). An Extensive Study on Cross-Dataset Bias and Evaluation Metrics Interpretation for Machine Learning applied to Gastrointestinal Tract Abnormality Classification. arXiv preprint arXiv:2005.03912. <https://doi.org/10.48550/arXiv.2005.03912>
- [20] Obi, J. C. (2023). A Comparative Study of Several Classification Metrics and Their Performances on Data. *World Journal of Advanced Engineering Technology and Sciences*, 8(1), 308-314 <https://doi.org/10.30574/wjaets.2023.8.1.0054>.