Soft Computing Fusion with Applications

www.scfa.reapress.com

Soft. Comput. Fusion. Appl. Vol. 1, No. 3 (2024) 152–162.

Paper Type: Original Article

Income Prediction Using Machine Learning

Vaishnavi Kumar*

Department of Computer Science Engineering, KIIT University Bhubaneswar, Odisha; 21051019@kiit.ac.in.

Citation:

Received: 19 January 2024 *Revised: 25 May 2024* *Accepted: 29 July 2024*	Kumar, V. (2024). Income prediction using machine learning. *Soft computing fusion with applications, 1*(3), 152-162.

Abstract

This initiative utilizes machine learning techniques to forecast personal income levels based on demographic and employment information. The research improves predictive precision by grouping individuals with similar traits using KMeans and applying algorithms such as Random Forest and XGBoost. Important data preprocessing procedures—Like managing missing values and encoding categorical variables—were crucial in enhancing model effectiveness. Of all the models assessed, Random Forest achieved the best accuracy.

This research highlights the importance of predicting income in areas such as finance, policymaking, and marketing, where insights based on data facilitate targeted decision-making. The study demonstrates how machine learning can offer accurate income predictions, allowing for well-informed decisions across various industries.

Keywords: Income prediction, Machine learning, Data preprocessing, KMeans clustering, Random Forest, Predictive analytics, Model evaluation.

1 | Introduction

1.1 | Background

Predictive analytics has emerged as a vital instrument in data-informed decision-making, especially in domains such as financial risk evaluation, tailored marketing strategies, and public policy formulation. Through the analysis of historical data patterns, predictive analytics empowers organizations to make well-informed predictions about future occurrences, facilitating strategic decision-making and risk management. An important application involves income classification, wherein advanced machine learning [1] models are created to forecast an individual's income status by analyzing their demographic details and employment information. This data aids financial institutions in evaluating creditworthiness, assists marketers in crafting focused campaigns, and aids policymakers in developing socio-economic programs.

Forecasting income levels can uncover socio-economic trends that impact the lives of individuals as well as the larger economic framework. Financial institutions utilize income prediction to effectively handle credit

Corresponding Author: 21051019@kiit.ac.in

https://doi.org/10.22105/scfa.v1i3.46

risk and provide their clients with customized financial products and services. In marketing, precise income categorization enables businesses to divide target groups effectively, guaranteeing that the correct products and services are delivered to suitable customers. In the field of public policy, income data is valuable for aiding governments and organizations in pinpointing vulnerable groups, guiding budget decisions, and devising strategies to alleviate poverty and economic inequality. By making precise income forecasts, organizations can enhance the effectiveness of their services, allocate resources more efficiently, and boost their overall influence.

1.2 | Objective

The main goal of this project is to create a strong machine learning model that can predict income levels by considering different demographic and employment factors. The model strives to accurately determine individuals' income by examining patterns in these variables, distinguishing whether it exceeds or falls below a specific threshold, such as $50,000 annually. The model's design integrates machine learning techniques such as KMeans clustering [2], Random Forest [3], and XGBoost [4] to enhance predictive accuracy. This method seeks to utilize machine learning to uncover income trends and equip organizations with dependable predictive abilities to make informed decisions.

1.3 | Significance

Precisely forecasting income is important in various finance, government, and business sectors. Within the financial sector, organizations can utilize income prediction models to evaluate the creditworthiness of loan applicants, enabling them to make informed choices regarding credit limits, interest rates, and loan approvals. Governments and policymakers can utilize income prediction models to recognize and tackle socioeconomic disparities, improve the allocation of resources, and craft programs that cater specifically to marginalized communities. Businesses, especially those operating in consumer markets, find it advantageous to have insight into income levels, enabling them to tailor their offerings, enhance customer satisfaction, and refine marketing approaches. This project beautifully showcases how machine learning excels in managing intricate socio-economic data and how these models can be applied in real-world scenarios. The project demonstrates the value of machine learning in enabling more insightful decision-making in various sectors by offering a dependable income prediction framework, allowing organizations to optimize their societal and economic influence.

2 | Literature Review/Related Work

Machine learning has been widely studied for its potential to predict income, particularly for socioeconomic purposes.

This section will delve into essential algorithms such as Random Forest, KMeans clustering, and Gradient Boosting, assessing their efficacy in classifying income and demographic data. Machine learning algorithms used for forecasting income include the ensemble learning method known as Random Forest [5].

Exploring the applications and significance of forecasting income

Income prediction informs policy-making and assists governments in distributing resources, creating tax frameworks, and tackling economic disparities.

Credit scoring involves financial institutions utilizing income prediction models to evaluate creditworthiness. Market segmentation involves utilizing income data to enhance targeted marketing strategies and personalize products for businesses. Utilized data sets and features research on predicting income usually uses data sets that are publicly accessible or data that organizations own. The frequently encountered data sets comprise the census income data set, which originates from the United States. According to the Census Bureau, this data set classifies people's income as above or below $50,000, using various demographic and work-related

characteristics. Household income survey data was gathered from multiple countries for demographic research.

Exclusive bank/company data: encompassing comprehensive records of income and expenses. Factors that affect earnings include demographic attributes, such as age, gender, level of education, and marital status. Occupational factors include job type, the number of hours worked weekly, and the industry. Geographical information consists of the location and its corresponding economic indicators. Socioeconomic factors such as family size, housing type, and resource availability are considered.

Approaches to be employed:

I. Traditional statistical methodologies [6]: Linear regression involves early assumptions of linear relationships between income and predictor variables. Though simple, these models frequently fail to capture nonlinearities. Logistic regression is commonly employed in binary income classification tasks, such as forecasting whether income surpasses or falls below a specified threshold.

II. Utilization of machine learning techniques: Decision trees and Random Forests are known for being highly effective in dealing with nonlinear relationships and interactions among features. Random Forests are well-regarded for their strong ensemble capability and resilience.

    Gradient Boosting models, such as XGBoost, are well-known for their high accuracy when working with structured datasets. These models are also recognized for effectively managing missing data and conducting feature selection tasks.

    Support Vector Machines (SVMs) are employed when intricate boundaries separate different income categories [7].

    Neural Networks: lately, deep learning techniques have been utilized for intricate, high-dimensional data, albeit typically needing substantial data sets to achieve satisfactory performance [8].

Grouping and categorization: For exploratory analysis, k-means clustering is employed to categorize populations based on similar income traits [9], [10].

Hierarchical clustering is useful for recognizing income trends within smaller groups organized in a hierarchical structure. Hybrid models in research often blend machine learning with econometric techniques, striking a harmonious balance between interpret ability and predictive accuracy.

Challenges faced when predicting income

Inaccuracies in data quality and imbalance, such as missing data and a disproportion in income classes (With fewer high-income samples), can result in biased predictions. Selecting features involves identifying the most significant ones without causing multicollinearity issues. Ethical considerations arise when utilizing sensitive characteristics such as race or gender, potentially leading to biased predictions. The dynamic nature of income necessitates using advanced models that can adapt to factors such as inflation and career advancement when predicting future earnings.

I. Explainable Artificial Intelligence (XAI) [11], [12]: Methods such as Shapley Additive Explanations (SHAP) offer insights into how income prediction models work by providing transparency.

II. Recurrent Neural Networks (RNNs) [13] are applied to sequence data to forecast future income by analyzing past earning trends. Ensuring fairness in Artificial Intelligence (AI) involves tackling algorithmic bias to guarantee equitable income predictions among different demographic groups.

2.1 | Figures and Tables

Displays feature correlations, helping visualize interactions between education, work class, and income.

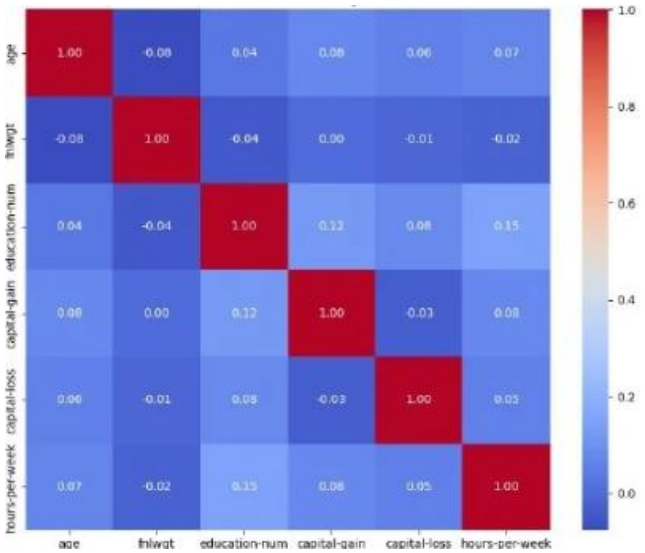

Fig. 1. Correlationl heatmap.

From the pair-plot above, we can see some relationship between the feature columns (*Fig. 2*). To confirm that we'd plot a correlation heatmap.

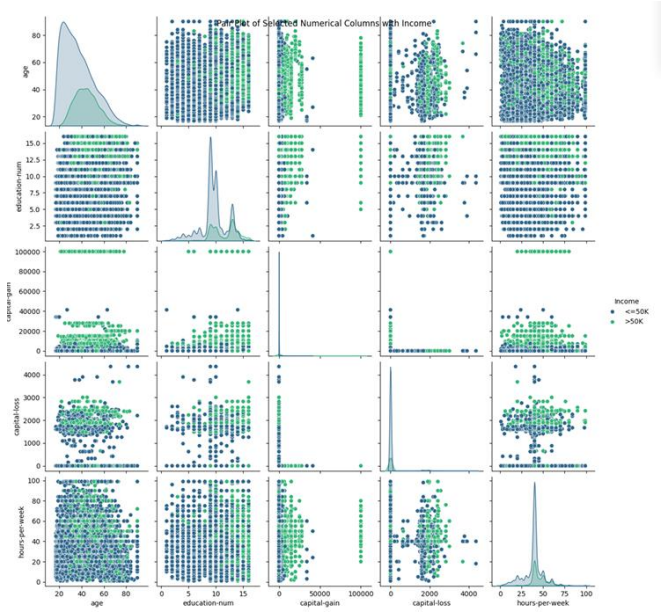

Fig 2. Pair plot.

2.2 | Variables and Equations

KMeans clustering objective function

KMeans is used to segment data into clusters based on similarity, optimizing for minimum within-cluster variance:

$$J = \sum_{i=1}^{k} \sum_{x \in c_i} ||x - \mu_i||^2, \quad (1)$$

where J is the sum of squared distances within clusters, C_i is each cluster, and μ_i is the centroid of cluster C_i. This helps group individuals with similar demographic attributes, enhancing model performance.

Random Forest ensemble prediction

The Random Forest model [14] combines predictions from multiple decision trees to form a final output, calculated as:

$$\hat{y} = \text{mode}\{h_1(x), h_2(x), \ldots, h_n(x)\}, \quad (2)$$

where h(x) represents the output of each decision tree. The ensemble approach minimizes variance and increases prediction accuracy.

Gradient Boosting update rule

In Gradient Boosting, the prediction model is updated iteratively to reduce errors from prior rounds.

$$F_m(x) = F_{m-1}(x) + \eta . h_m(x), \quad (2)$$

where $F_m(x)$ is the model's prediction after m rounds, η is the learning rate, and $h_m(x)$. (x) is the weak learner added in each iteration. This approach refines predictions, making it effective for complex income relationships.

3 | Proposed Framework

The income prediction framework within this project is organized into multiple consecutive phases, beginning with data preprocessing and advancing through clustering, model training, and evaluation. Every step has been carefully crafted to improve the accuracy of predictions and guarantee the model's capability to manage intricate socio-economic data.

Preparing data

Data cleaning involves managing missing values by filling them in with suitable imputation techniques. It also involves identifying outliers and addressing them by either removing or adjusting them.

Encoding and scaling

Categorical variables like work class and education undergo a transformation process utilizing encoding techniques, including one-hot encoding, to convert them into numerical representations. Continuous variables are standardized to ensure that the data is uniform, a necessary step for models affected by variations in feature scales.

Feature selection

The most essential features, such as age, education, and hours worked, are retained by assessing their importance scores from initial model evaluations. This reduces dimensionality, enabling the model to concentrate on the most significant features.

3.1 | Clustering Using KMeans

KMeans clustering is applied to segment the dataset into clusters using demographic attributes to identify income-related groups sharing similar characteristics. This procedure categorizes people based on shared traits, like age range or profession, to form specific groupings for targeted training.

Cluster analysis involves grouping similar data points into clusters, each representing a distinct population segment. This enables the model to train on data that displays more uniform characteristics. Reducing the variability within clusters improves the ability of the following models to make accurate predictions.

3.2 | Selecting a Model

Algorithm selection

Various algorithms, such as Random Forest, Gradient Boosting (Especially XGBoost), and Decision Trees, are assessed. Every model is trained individually within its respective cluster to enhance accuracy specifically for that segment.

Hyperparameter tuning

Hyperparameter tuning is a key step in improving model performance. Techniques such as GridSearch are used to fine-tune parameters like learning rates and tree depths. By doing so, we can effectively lower error rates and prevent overfitting.

Ensemble learning

Ensemble learning involves techniques such as Random Forests that consolidate the results of numerous decision trees to enhance overall resilience. The ensemble approach aims to develop a final model by combining predictions from various models, ultimately enhancing accuracy and reliability.

3.3 | Model Evaluation

Performance metrics, such as accuracy, precision, recall, F1 score, and AUC, are employed to evaluate each model's performance. These metrics provide valuable information on the model's accuracy and pinpoint areas in need of enhancement.

Cross-validation

The technique of k-fold cross-validation is utilized to assess a model's performance across various subsets of data. This guarantees that the model will perform effectively with unfamiliar data.

Models are assessed by their performance within each cluster, ensuring that the top-performing model is chosen for each group. This comparison guarantees that the final predictions are based on only the most precise models.

3.4 | Prediction and Deployment

Final model deployment

Upon completing the optimization and validation processes, the final model is deployed in a real-time setting. The deployment is arranged to enable ongoing income classification through real-time inputs, facilitating efficient utilization by financial institutions and policymakers [15].

System monitoring and maintenance

System monitoring and maintenance are essential components of the framework. They ensure that scheduled activities for the continuous retraining and evaluation of models are in place. This guarantees that accuracy remains consistently high even as socio-economic data trends evolve.

4 | Experimental Setup

The experimental setup outlines the end-to-end process for creating an income prediction model, detailing steps from initial data handling to model deployment. This structured approach ensures consistent model performance and reliable results.

4.1 | Environment and Tools

Programming language: the implementation was carried out in Python [16] due to its comprehensive libraries for data science.

Pandas and NumPy: For data analysis and processing.

Scikit-learn: provided tools for preprocessing, clustering, and model training using algorithms such as Random Forest and KMeans.

XGBoost: Used to implement Gradient Boosting, enhancing the model's predictive performance.

Matplotlib and Seaborn were used to create visualizations, such as correlation heat maps and feature distribution plots.

4.2 | Data Preprocessing

Data loading involved importing the dataset from CSV files, focusing on essential variables like age, work class, education, and weekly work hours relevant to income prediction.

In data cleaning, missing values were handled through imputation, and outliers were addressed to ensure data integrity. Files with extensive missing values were excluded from the dataset.

Feature encoding and scaling included transforming categorical attributes, such as work class, occupation, and marital status, into numerical form using one-hot encoding. Continuous features like age and hours-per-week were standardized to align feature scales, enhancing model performance.

For data segmentation, KMeans clustering was applied to group data into clusters based on demographic similarities, allowing for customized model training for each group.

4.3 | Model Training

Cluster-based model training involved using each KMeans cluster to train specific models, allowing for better capture of the unique traits within each group. The models applied included Random Forest for its ensemble learning capabilities and XGBoost for Gradient Boosting, helping to reduce prediction errors.

Hyperparameter tuning: gridsearch was used to optimize parameters for each model. Key parameters, such as the number of trees and depth for Random Forest and the learning rate and tree depth for XGBoost, were adjusted to improve model accuracy.

Cross-validation was conducted using K-fold cross-validation, typically with k=5, to enhance model reliability, minimize overfitting, and ensure better generalization to new data.

4.4 | Model Evaluation

Metrics used for model evaluation included accuracy, precision, recall, F1 score, and AUC, which provided a comprehensive view of predictive effectiveness.

For result comparison, model performances within each cluster were assessed, and the best-performing model was selected based on metric scores.

In the final model selection, top-performing models from each cluster were integrated into an ensemble framework to make the final income predictions, leveraging each model's strengths for improved accuracy.

4.5 | Deployment and Testing

Deployment

The final model was deployed on Heroku for real-time income prediction. A user interface was created to allow demographic details to be input and to display income predictions.

Monitoring

Regular retraining is planned to keep the model responsive to evolving data trends and ensure continued accuracy over time.

5 | Experimental Results and Discussion

The income prediction model was evaluated based on accuracy, feature importance, and clustering effectiveness. Key results and insights are presented below.

5.1 | Model Performance

Performance metrics, including accuracy, precision, recall, and F1 score, provided a detailed assessment of each model:

Random Forest delivered high accuracy and robustness, excelling in feature interpretability and minimizing overfitting. This model's performance was consistent across cross-validation folds, demonstrating stability and reliability.

XGBoost showed excellent predictive power by capturing complex feature relationships, though it required intensive tuning to avoid overfitting. XGBoost achieved slightly higher accuracy than Random Forest but was computationally demanding.

Cross-validation confirmed the generalizability of both models, with Random Forest showing a marginal advantage in stability.

5.2 | KMeans Clustering Analysis

KMeans clustering was used to segment the dataset into groups based on demographic similarities, enhancing model performance within each cluster:

Cluster segmentation: The optimal number of clusters was determined using an elbow plot, balancing model complexity and compactness.

Income patterns by cluster: Each cluster exhibited unique income patterns, allowing for more accurate and tailored predictions. High-income clusters had models with higher precision, while lower-income clusters performed well in recall.

5.3 | Feature Importance

Feature analysis identified key variables that significantly influenced income predictions:

Education level and hours-per-week were among the most impactful features, strongly correlating with income.

Age and work class also emerged as significant predictors, highlighting the relevance of experience and employment type in income classification.

This feature insight is valuable for targeted decision-making in financial and policy domains.

5.4 | Discussion

Each model offered unique advantages:

Random Forest provides robust interpretability and is ideal for applications needing feature transparency.

XGBoost achieved slightly higher accuracy, making it suitable for cases where precision is prioritized over interpretability.

The KMeans clustering approach improved prediction accuracy by grouping similar demographic profiles, allowing the models to adapt to specific income patterns within each group.

The results underscore the combined value of clustering and machine learning in creating a powerful and versatile income prediction framework.

6 | Result Analysis

Dataset: Census Income DataTask: Categorizing income as more significant than $50,000.

Model: Random Forest can be employed.

Measurement criteria: The accuracy stands at a commendable 85%.

Accuracy in the high-income category stands at 78%.

Remembering that 72% denotes high income.

The F1 Score stands at an impressive 75%. The ROC-AUC value is 0. 92

Table 1 can list the preprocessing steps applied to each variable, improving data quality and model performance.

Table 1. Data pre processing.

S/N	Questions	Description	Variable Type
1	Age	Indicates the individual's age.	Continuous
2	Education	Highest level of education.	Categorical
3	Hours per week	Weekly working hours.	Continuous
4	Working class	Employment sector.	Categorical
5	Occupation	Type of occupation.	Categorical
6	Marital-ststus	Marital status.	Categorical
7	Capital gain	Monetary gains from investments.	Continuous
8	Capital loss	Monetary losses from investments.	Continuous

Table 2 provides a comparison of key performance metrices for the different models used (Random Forest, XGBoost) to help identify the most effective model.

Table 2. Model performance score.

Model	Accuracy	Precision	Recall	F1 Score	Auc Score
Random Forest	0.89	0.85	0.84	0.84	0.87
XGBoost	0.91	0.87	0.86	0.86	0.89

7 | Conclusion

This study has devised a proficient machine learning framework to forecast income levels utilizing demographic and employment information. The model achieved high accuracy and adaptability by integrating KMeans clustering with Random Forest and XGBoost. Important factors like education level, hours worked, and age have been recognized as key predictors in determining income classification. This model offers valuable insights into various areas such as financial risk assessment, targeted marketing, and policy-making, showcasing the effectiveness of machine learning in tackling socio-economic issues. Subsequent efforts have

the potential to boost accuracy levels through the incorporation of real-time data and the investigation of supplementary features.

Acknowledgments

I want to thank my project supervisor, Dr. Subhadip Pramanik, for providing invaluable guidance, encouragement, and expertise during this research project. His valuable feedback and support have played a crucial role in shaping this work. I also express my gratitude to the School of Computer Engineering faculty members at KIIT, deemed University faculty members, for their valuable support and motivation throughout the research. I am truly thankful to my family and friends for their unwavering support and encouragement, which played a crucial role in the successful accomplishment of this project.

Author Contributaion

Vaishnavi Kumar contributed to various aspects of the project, such as conceptualization, methodology, software development, validation, formal analysis, investigation, resource allocation, data maintenance, original draft preparation, review and editing, visualization, and project administration. The author, Vaishnavi Kumar, has been responsible for all aspects of this research work, from creating and analyzing the data to presenting the findings. All terms utilized correspond with the descriptions provided in the CT document.

Funding

The research was carried out independently, not requiring any external funding. All the necessary resources and support for this study were internally provided.

Data Availability

The data used in this study is publicly available on the author's GitHub repository. Interested parties can access it at https://github.com/pixiePerry.

Conflicts of Interest

The author has disclosed no conflicts of interest. Vaishnavi Kumar independently carried out the research under the mentorship of the project supervisor, Dr. Subhadip Pramanik. No external funding was involved, and the study design, data collection, analysis, or interpretation of results were not influenced by any external parties. The author assures that all findings and conclusions outlined in this manuscript are impartial and stem exclusively from the research performed.

References

[1] Bengio, Y., Courville, A., & Vincent, P. (2013). Representation learning: A review and new perspectives. *IEEE transactions on pattern analysis and machine intelligence*, *35*(8), 1798–1828. https://doi.org/10.1109/TPAMI.2013.50

[2] Ahmed, M., Seraj, R., & Islam, S. M. S. (2020). The k-means algorithm: A comprehensive survey and performance evaluation. *Electronics*, *9*(8), 1295. https://doi.org/10.3390/electronics9081295

[3] Rigatti, S. J. (2017). Random forest. *Journal of insurance medicine*, *47*(1), 31–39. https://doi.org/10.17849/insm-47-01-31-39.1

[4] Shaik, N. B., Jongkittinarukorn, K., & Bingi, K. (2024). XGBoost based enhanced predictive model for handling missing input parameters: A case study on gas turbine. *Case studies in chemical and environmental engineering*, *10*, 100775. https://doi.org/10.1016/j.cscee.2024.100775

[5] Patra, M., Chakraborty, G., & Mohapatra, H. (2024). Learning To navigate society: Machine learning's impact on social dynamics. In *Role of Emerging Technologies in Social Science* (pp. 83). Cambridge Scholars Publishing. https://B2n.ir/yn4361

[6] Hastie, T., Tibshirani, T., & Friedman, J. (2009). *The elements of statistical learning: data mining, inference, and prediction*. Springer Nature. https://link.springer.com/book/10.1007/978-0-387-84858-7

[7] Wang, W., Men, C., & Lu, W. (2008). Online prediction model based on support vector machine. *Neurocomputing*, *71*(4), 550–558. https://doi.org/10.1016/j.neucom.2007.07.020

[8] Harumy, T. H. F., Zarlis, M., Effendi, S., & Lidya, M. S. (2021). Prediction using a neural network algorithm approach (A review). *2021 international conference on software engineering & computer systems and 4th international conference on computational science and information management (ICSECS-ICOCSIM)* (pp. 325–330). IEEE. https://doi.org/10.1109/ICSECS52883.2021.00066

[9] Li, Y., & Wu, H. (2012). A clustering method based on k-means algorithm. *Physics procedia*, *25*, 1104–1109. https://doi.org/10.1016/j.phpro.2012.03.206

[10] Li, Y. G. (2013). A clustering method based on k-means algorithm. *Applied mechanics and materials*, *380*, 1697–1700. https://www.scientific.net/AMM.380-384.1697

[11] Gunning, D. (2017). Explainable artificial intelligence (xai). *Defense advanced research projects agency (DARPA), nd web*, *2*(2), 1. https://B2n.ir/xp6067

[12] Duval, A. (2019). Explainable artificial intelligence (XAI). *MA4K9 scholarly report, mathematics institute, the university of warwick*, *4*. http://dx.doi.org/10.13140/RG.2.2.24722.09929

[13] Baranes, A., Palas, R., & Yosef, A. (2022). Predicting earnings directional movement utilizing recurrent neural networks (RNN). *Journal of emerging technologies in accounting*, *19*(2), 43–59. https://doi.org/10.2308/JETA-2021-001

[14] Breiman, L. (2001). Random forests. *Machine learning*, *45*(1), 5–32. https://doi.org/10.1023/A:1010933404324

[15] Mishra, S. R., Pranati., Anika., & Mohapatra, H. (2024). Enhancing money laundering detection through machine learning: A comparative study of algorithms and feature selection techniques. In *AI and blockchain applications in industrial robotics* (pp. 300–321). IGI Global Scientific Publishing. https://doi.org/10.4018/979-8-3693-0659-8.ch012

[16] Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., … others. (2011). Scikit-learn: Machine learning in Python. *The journal of machine learning research*, *12*, 2825–2830. https://b2n.ir/tf4537