An Investigation into the Prediction of Annual Income Levels Through the Utilization of Demographic Features Employing the Modified UCI Adult Dataset

Conference Paper · February 2024

DOI: 10.1109/ICCCIS60361.2023.10425394

Md. Aminul Islam
International Islamic University Chittagong
41 PUBLICATIONS **103** CITATIONS
SEE PROFILE

Anindya Nag
Ca' Foscari University of Venice
87 PUBLICATIONS **281** CITATIONS
SEE PROFILE

Nilanjana Roy
Adamas knowledge city
5 PUBLICATIONS **17** CITATIONS
SEE PROFILE

Arpita Dey
Khulna University
2 PUBLICATIONS **82** CITATIONS
SEE PROFILE

2023 International Conference on Computing, Communication, and Intelligent Systems (ICCCIS) | 979-8-3503-0611-8/23/$31.00 ©2023 IEEE | DOI: 10.1109/ICCCIS60361.2023.10425394

An Investigation into the Prediction of Annual Income Levels through the Utilization of Demographic Features Employing the Modified UCI Adult Dataset

Md Aminul Islam
School of Engineering, Computing, and Mathematics
Oxford Brookes University, UK
talukder.rana.13@gmail.com,
19126681@brookes.ac.uk
ORCID ID: 0000-0002-2535-6519

Anindya Nag
Computer Science & Engineering Dscipline
Khulna University
Khulna-9208, Bangladesh
anindyanag58@gmail.com
ORCID: 0000-0001-6518-8233

Nilanjana Roy
Department of Computer Science & Engineering
Adamas University
Kolkata - 700126, India
nilanjanaroy524@gmail.com

Arpita Rani Dey
Computer Science & Engineering Discipline
Khulna University
Khulna-9208, Bangladesh
msc230236@ku.ac.bd

SM Firoz Ahmed Fahim
Department of Computer Science & Engineering
American International University Bangladesh
Dhaka, Bangladesh
firozfahimm@gmail.com

Arjan Ghosh
Computer Science& Engineering Discipline
Khulna University
Khulna-9208, Bangladesh
arjan.ghosh@g.bracu.ac.bd

Abstract- **Predicting an individual's annual income is crucial and has many uses for the planning of government, policymakers, and development partners. This paper uses the UCI Adult Dataset to examine demographic-based annual income predictions. The 16-column dataset classifies "Income" into two categories: <=50K and >50K, where 14 characteristics provide personal information. Appropriate Methods to Handle Categorical Data guidelines govern the study's preprocessing, categorical data management, and missing value strategies. Eleven machine learning models are then studied, including classification, clustering, neural networks, Logistic Regression, Naive Bayes (standard and Gaussian), K-Nearest Neighbors, SVMs, Decision Trees, Random Forest, XGBoost, and Artificial Neural Networks. This study emphasizes optimization, evaluation, and model comparison and how optimization strategies and parameters affect model performance. Comparative analyses compare model performance within a category. The results imply that demographic characteristics help some models predict income levels better. Logistic Regression provides essential knowledge; whereas XGBoost and Neural Networks improve accuracy, Clustering adds knowledge. This study's findings inform revenue-anticipation-related decision-making with an accuracy of 87% from XGBoost and XGBoost-ANN Ensembling. This study emphasizes the need for data pretreatment, algorithm selection, optimization, and evaluation for accurate predictions and valuable insights.**

Keywords— Income Prediction, Demographic Features, XGBoost, ANN Ensembling.

I. Introduction

In the past twenty years, there has been a significant increase in human reliance on data and information within society. Consequently, there has been a corresponding advancement in technology designed to accommodate the storage, analysis, and processing of these vast quantities of data. Data Mining and Machine Learning (ML) disciplines have been utilized to acquire knowledge and make discoveries and investigate concealed patterns and concepts, which have facilitated the prediction of future events that are not readily attainable [1]. The issue of income disparity has garnered significant attention in recent times. The sole criterion for pursuing the eradication of poverty does not appear to be solely focused on improving the well-being of the impoverished. Despite the achievements of machine learning in handling intricate problems related to human prediction, an increasing body of research indicates that "state-of-the-art" models exhibit notably worse accuracy when applied to minority groups compared to the majority population [2-3].

A nation's Gross Domestic Product (GDP) refers to the comprehensive monetary or market worth of all the commodities and services generated within the territorial boundaries of that nation during a designated period [4]. Economists frequently employ the GDP as a key metric for assessing the general state of an economy, its growth trajectory, and the effects of inflation and deflation within a given country. Due to its comprehensive measurement of a nation's production, GDP can be seen as a reliable indicator of its economic prowess and advancement. GDP encompasses the aggregate value of products and services that are created and made available for exchange in the market. GDP also incorporates certain nonmarket output forms, such as government-provided defense or education services [5].

Our objective is to make predictions about income using the Modified UCI Adult Dataset. To develop machine learning models, we have formulated the following issue statements:

a. The examination of income levels by employing demographic features for analysis.

b. Compares models for label and one hot encoding

ISBN: 979-8-3503-0611-8/23/$31.00 ©2023 IEEE

c. Selecting optimal models for performance evaluation

The paper follows a systematic format, where Section 2 includes a comprehensive review of the existing literature, Section 3 outlines the research materials and technique employed, Section 4 gives the analysis and conclusions obtained, and Section 5 summarizes the study.

II. LITERATURE REVIEW

Many supervised ML methods are used to predict the income of a population using various parameters, such as age, gender, marital status, and so on. In this section, various existing procedures are explained briefly.

I. Lopez [1] analyzed the data from the Census 1994 database. Implemented the pre-processed dataset on three machine learning algorithms, i.e., Logistic Regression (LR), Decision Tree (DT), and Naïve Bayes Classifier (NBC), by splitting them into training and testing sets. DT has the highest accuracy of 85.76% in predicting income that is over or equal to 50,000 or less. Sunil Thapa [2] attempted to compare different ML algorithms on a dataset of adult income. In this experiment, they implemented a tuning algorithm called Grid Search to find out the best hyperparameter. The Random Forest Classifier (RF) outperformed all the other models by gaining an accuracy of 86%. Alina Lazar [3] combined the Support Vector Machine (SVM) method with principal component analysis to generate and evaluate income prediction from the Census Bureau database. They claimed that it enhanced the accuracy by 84% and reduced the computational time by 60%. Kim et al. [4] presented the Multiaccuracy Boost algorithm (MBA) on a relatively small dataset, showing its effectiveness in diverse applications. In this experiment, the classification errors are minimized, thus increasing overall accuracy in a data-efficient manner for a black-box predictor. N. Chakraborty et al. [5] proposed a gradient boosting classifier (GBC) method tuned with hyper-parameters. ML techniques can be also used for pair politics, blockchain, breast cancer detection, cyber security etc to make human life better [27], [28], [29]. They modified the model with a grid search algorithm. They compared their output with other existing experiment models, such as Principal Component Analysis (PCA), SVM, GBC, and Extreme Gradient Boosting (XGBOOST). Their income prediction model obtained an accuracy of 88.16%, outperforming the other models they compared. Moe et al. [6] tested over 30000 of data and more than forty countries using three ML models NBC, DT J48, and RF. They tried to determine the impact of GDP growth on a country's financial development. They calculated the correctness and incorrectness of the classification percentages. In this experiment, the overall accuracy of DT J48 is 85.256%, which exceeds 1% and 3% from RF and NBC, respectively. V. Ríos [7] used a supervised learning model collaborating with artificial intelligence (AI) using predictive analytics. They tried to find out the direct relationship between a person's income with his academic education. They used Microsoft Azure ML as a predictive analytic model builder. It showed the graphical output of the income prediction. Their findings suggested that education and occupation have the most impact on predicting income.

TABLE I. SUMMARY OF RELATED WORK

Ref.	Model	Dataset	Types of variables	Accuracy
[1]	LR, DT, NBC	1994 Census dataset	Categoric, Numeric	LR - 84.70%, DT - 85.76%, NBC - 81.99%
[2]	RF, SVC , KNC , LR, NBC	The UCI ML repository	Categoric, Numeric	RF - 86%, SVC - 85%, KNeighbors - 84%, LR - 84%, NBC - 82%
[3]	SVM	CPS data by the US Census Bureau	Categoric, Numeric	84%
[4]	MBA	CelebA data	Binary	--
[5]	Hyper parameter tuned GBM	Adult Census Data	Categoric, Numeric	88.16%
[6]	DTJ48, RF, NBC	Adult income data of US	Categoric, Numeric	DTJ48 - 85.256% RF - 84.2516% NBC - 82.8181%
[7]	AI, ML	1994 Census data	Categoric, Numeric	Graphical analysis

III. MATERIALS AND METHODOLOGY

The work may be categorized into five primary phases, which are as follows: CensusDB Dataset, Data Preprocessing and Transformation, Exploratory Data Analysis (EDA), Feature Engineering and Data splitting, and Optimization outcome and evaluation.

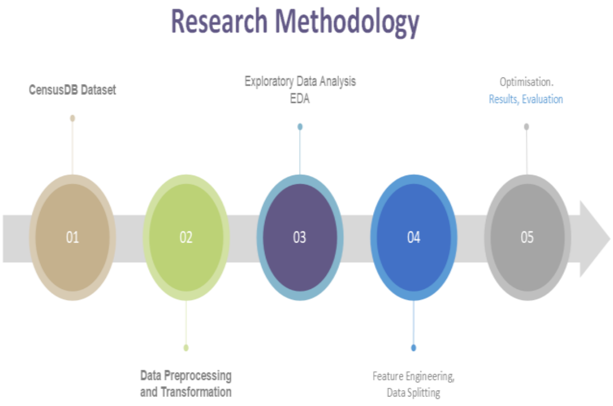

Fig. 1. Research Methodology

A. Dataset Description

In our study, we have used the Census DB dataset for the task of income prediction. It contains various demographic and socio-economic features of individuals, including their age, workclass, fnlwgt, education, education-num, marital-status, occupation, relationship, sex, capital-gain, capital-loss, hours-per-week, native-country, and income. There are total 32561 records which depicts information of 32561 persons taken for preparing the dataset [8].

B. Data Pre-processing

Pre-processing data is an important stage in machine learning research [9]. One hot encodes categorical variables and scale numerical variables to [0,1]. One-hot encoding is a technique used in machine learning and data analysis to convert categorical variables into a numerical format. This method creates binary features for each unique category in the variable. One-hot encoding is a widely employed technique in machine learning models to convert category variables into numerical values [10]. One of the benefits associated with the utilization of one hot encoding is:

i. This feature enables the incorporation of categorical variables into models that necessitate numerical input. Enhancing the model's performance can be achieved by augmenting the amount of information provided to the model regarding the categorical variable.

ii. One potential solution to mitigate the issue of ordinality is to address situations where a categorical variable possesses an inherent order, such as the use of descriptors like "small," "medium," and "large."

C. *Exploratory Data Analysis*

Exploratory Data Analysis (EDA) refers to the systematic examination and analysis of datasets with the objective of identifying key qualities and revealing meaningful patterns that can provide insights and address specific inquiries [11]. This resource facilitates a comprehensive comprehension of many aspects and their interconnections. The following section enumerates the diverse visualization approaches employed in data analysis.

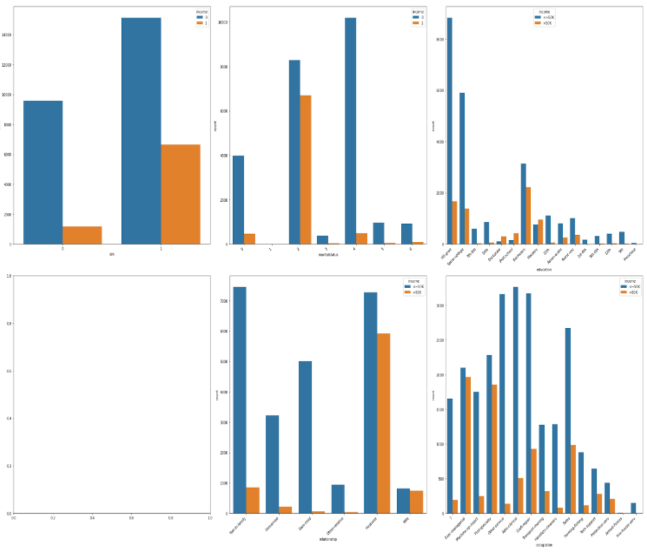

Fig. 2. Displaying the frequency distribution with a countplot

Figure 2 presents a countplot that illustrates the frequency distribution of significant categorical variables, namely "sex," "marital status," "education," "relationship," and "occupation," with respect to income.

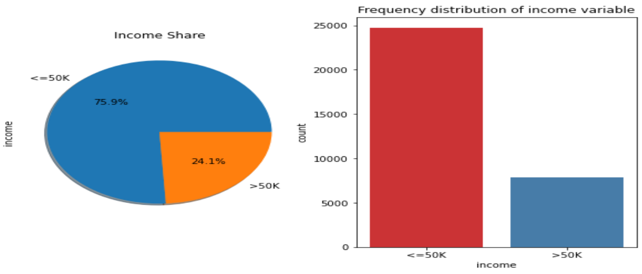

Fig. 3. Distribution plotted as a frequency distribution

Figure 3 presents a visual representation of the frequency distribution of the income variable.

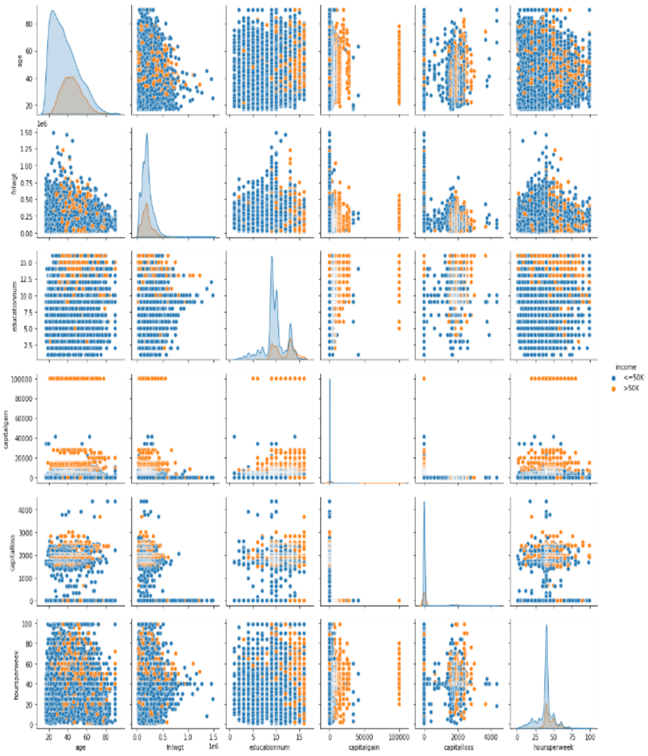

Fig. 4. Pairplot depicting the relationship between each independent feature and Income

Figure 4 depicts a Pairplot that illustrates the association between each independent attribute and Income. Based on the observed data, to control for extraneous variables, it is apparent that the variable fnlwgt has a minimal correlation with the target variable. As a result, this variable is omitted from subsequent analysis. To address the concern of multicollinearity between the variables "education.num" and "education," the variable "education.num" is also omitted from the analysis. Aaddress the issue of missing data, it is necessary to implement appropriate strategies and techniques. Considering the significant amount of data available to us and the relatively small percentage of missing values within the dataset, we have made the decision to exclude rows that include these missing values.

Feature selection is a crucial step in data analysis where the objective is to identify the most pertinent attributes that yield the most ideal outcome. Figure 5 presents the correlation matrix, which showcases the correlations between the chosen qualities and the variable of Income, with a particular emphasis on those correlations that are more robust.

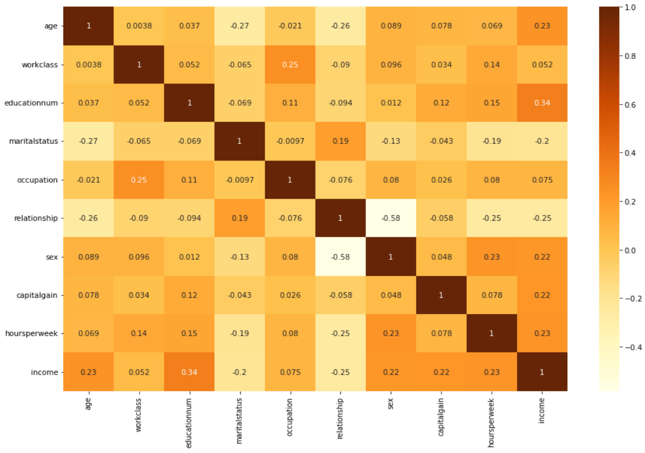

Fig. 5. Correlation matrix illustrates the correlation between selected attributes

D. Model Preparation and Testing

In this section, we will examine the prediction models. The data was divided into distinct strata, and a particular random state was utilized to maintain consistent objective values across all models. It is worth mentioning that the bulk of the implemented complex models have undergone hyperparameter optimization using GridSearchCV. An investigation and training process has been conducted utilizing a total of 11 machine learning models, including various techniques such as classification, clustering, and neural networks. The further proposed models for supervised machine learning are as follows:

1) Logistic regression: Logistic Regression is a supervised machine learning technique as well. This statistical model is characterized by its dependence on binary values. The alternative nomenclature for the Logistic function is the Sigmoid function. The predictor variable is converted into a probability value ranging from 0 to 1. The aforementioned likelihood serves as the fundamental basis for the process of decision making. The objective is to minimize the negative logarithm. Due to its simplicity and significant impact, the statistical toolkit in question has gained widespread recognition and utilization [12].

2) Categorical Naïve Bayes: The Categorical Naïve Bayes algorithm is a classification technique that relies on the principles of the Naïve Bayes theorem. The software is specifically engineered to effectively handle categorical features present in the dataset [13].

3) Gaussian Naïve Bayes: The Gaussian Naïve Bayes method, sometimes referred to as GaussianNB, is a probabilistic classification technique that makes the assumption that the features follow a normal distribution [13].

4) K-Nearest Neighbours (KNN): KNN algorithm is a supervised machine learning technique employed for classification tasks. The algorithm computes the degree of similarity between the dataset and existing cases, afterwards assigning the new data to the category that exhibits the highest level of similarity. The KNN model, which is a pattern identification algorithm, was employed in our method because to the inherent presence of patterns in fraudulent activities [14].

5) Random Forest : The Random Forest (RF) algorithm is a classifier that combines multiple models. RF models are constructed using many independent base classifier techniques, particularly decision trees. The decision tree is constructed on a significant scale. In this experiment, a test sample is fed into a novel classifier, and the outcomes are determined by aggregating the voting results of each individual classifier [15].

6) Decision Tree: It is a graphical representation that exhibits a hierarchical structure reminiscent of a tree. The decision-making process is guided by the utilization of input data by constructing a decision tree entails iteratively dividing the dataset into smaller subgroups, persisting until a predefined level of precision is gained. In this context, the decision is represented by the nodes, whereas the branches represent the different outcomes linked to every decision [16].

7) XGBoost (Extreme Gradient Boosting,) is an ML technique that employs parallel tree boosting to address diverse data science problems in a highly efficient and effective manner to create an XGBoost classification model using a predetermined setup. The training process of the model will involve the utilization of 500 estimators, with a learning rate of 0.05 and the implementation will include an early stopping mechanism, the training process will cease if no improvement is detected for a continuous sequence of 5 iterations to acknowledge that the utilization of a reduced learning rate and a higher number of estimators has the potential to result in improved accuracy inside XGBoost models [17].

8) Support Vector Machines (SVMs) are widely used in the field of machine learning because of their significant effectiveness in addressing categorization tasks. Support Vector Machines (SVMs) are based on the fundamental notion of determining an optimal hyperplane that can efficiently differentiate between different classes of input [18].

9) Artificial neural networks (ANNs)*:* ANNs are computational models that draw inspiration from the structural and functional characteristics of biological neural networks. ANNs are composed of interconnected nodes, referred to as artificial neurons. These neurons receive inputs, which are then subjected to weights and biases. Subsequently, various activation functions are applied to these processed inputs in order to obtain the desired output [19].

10) Ensembling:: The superiority of XGBoost in terms of performance is apparent when compared to the other models. It is important to highlight that the validation score acquired for the ANN applies to a singular instance of the ANN. Hence, it is expected that the validation score will be higher when utilizing a neural network ensemble. Hence, the ensembling strategy can be employed by utilizing both XGBoost and ANN. Considering the exceptional performance demonstrated by XGBoost, it is deemed suitable to allocate it a greater weightage [20-21].

11) Principal Component Analysis (PCA) : Principal Component Analysis (PCA) as a method for reducing dimensionality, optimizing models, and comparing different models. PCA is a commonly employed statistical methodology within the realm of data analysis [21]. PCA is a widely used technique in the field of dimensionality reduction.

Its primary objective is to restructure a given dataset into a lower-dimensional space, while simultaneously preserving the maximum amount of relevant information.The utilization of PCA was implemented in order to efficiently reduce the dimensionality of the dataset [22].

12) k-means clustering method: The k-means clustering technique is a commonly used unsupervised machine learning approach that is applied to discover groupings of data objects within a given dataset. The scikit-learn library, often known as sklearn, provides a variety of hyper-parameters for k-means clustering. These hyper-parameters include n_clusters, init, n_init, max_iter, tol, verbose, random_state, copy_x, and algorithm. The variable n_clusters has been assigned a value of 3. The determination of the optimal number of clusters to be formed [23]. The utilization of the k-means++ algorithm is implemented in order to choose the initial cluster centers for the k-means clustering technique, with the aim of accelerating the convergence process. The argument "max_iter" specifies the number of iterations in which the function will be executed for various centroid seeds. By default, this parameter is set to 10. The option "tol" denotes the upper limit for the number of iterations permitted in each cycle, and it has a default value of 300. The user's text should be changed to reflect a more formal and scholarly tone. The user's text should be rewritten to adhere to academic standards without adding any additional information. The parameter "random_state" is employed to initialize the centroid by utilizing random integers. In this particular instance, the assigned value is null. The function copy_x is frequently utilized to consolidate the data, as is customary in such situations. The algorithm parameter's default value is set to "auto" to ensure compatibility with previous versions. The subject of inquiry is around the diverse approaches utilized to evaluate or gauge the quality, efficacy, or worth of a specific thing, process, or occurrence [24].

13) DBSCAN Clustering: It exhibits a tendency to aggregate data points that are closely grouped together. By utilizing the local density of data points, this method is capable of detecting clusters within extensive spatial datasets. One of the most captivating features of DBSCAN clustering is its ability to effectively handle outliers. Furthermore, unlike the K-means algorithm which necessitates the prior specification of the number of centroids, our method does not impose any such requirement. DBSCAN, a density-based clustering algorithm, necessitates the specification of two parameters: epsilon and minPoints. The value of epsilon represents the radius of the circle that needs to be constructed around individual data points for the purpose of calculating their density [25]. On the other hand, minPoints refers to the minimum number of data points that must be located within this circle in order for the respective data point to be classed as a Core point. In the context of higher dimensions, the geometric shape known as a circle is transformed into a hypersphere. The parameter epsilon, in this case, represents the radius of the hypersphere. Additionally, the term minPoints refers to the minimum threshold of data points that must be present within the confines of the hypersphere [26].

IV. Analysis and Result Discussion

After analyzing we got the results which are depicted in the form of the following visualizations. These graphs provide insights into the distribution and probability of specific features contributing to the Income of an adult person. Each bar correspond to a distinct feature category which includes education level, occupation, marital status, sex, work class, relationship, work time, age and their earning above and below $50k. The extraction of relevant insights from these distributions holds significance in terms of pattern conversion and deriving meaningful conclusions from the investigation.

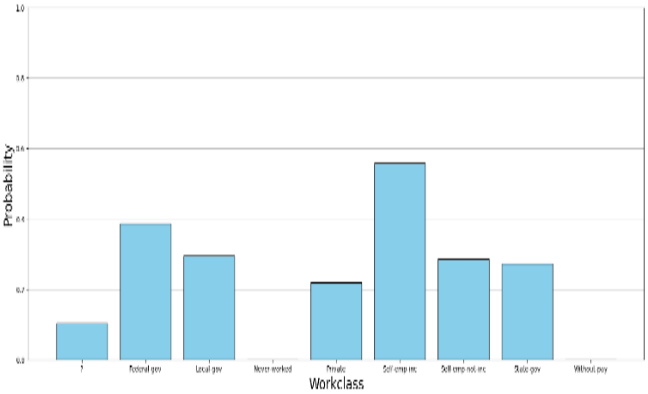

Fig. 6. Probality of earning Vs Workclass

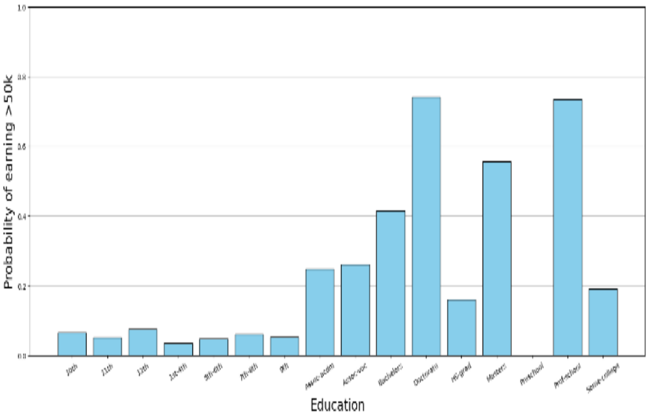

Fig. 7. Probality of earning >50 k Vs Education

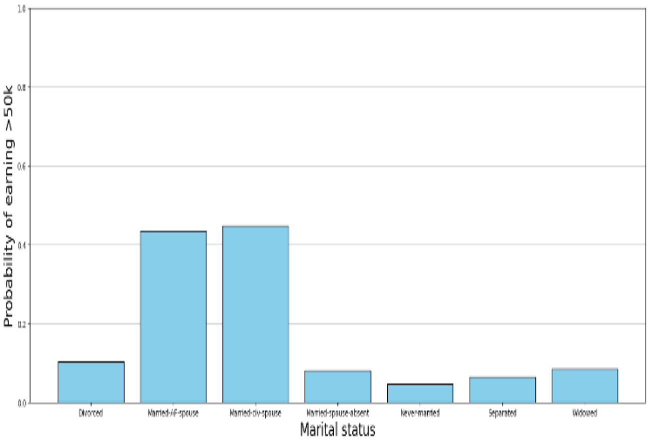

Fig. 8. Probality of earning >50 k Vs Marital status

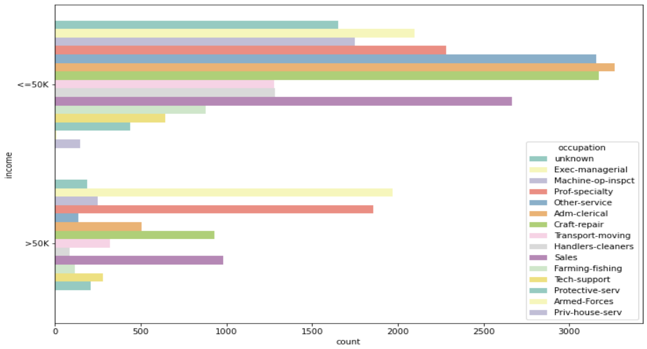

Fig. 9. Income Vs Occcupation

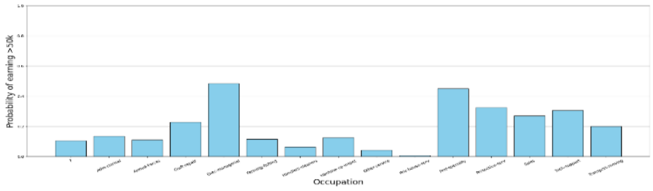

Fig. 10. Probality of earning >50 k Vs Occupation

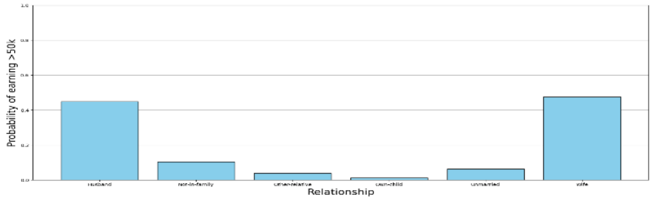

Fig. 11. Probality of earning >50 k Vs Relationship

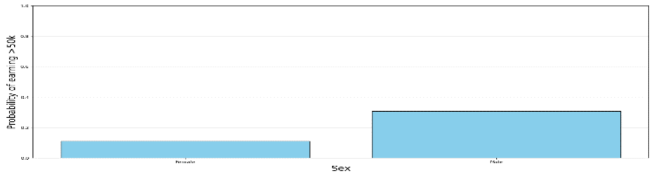

Fig. 12. Probality of earning >50 k Vs Sex

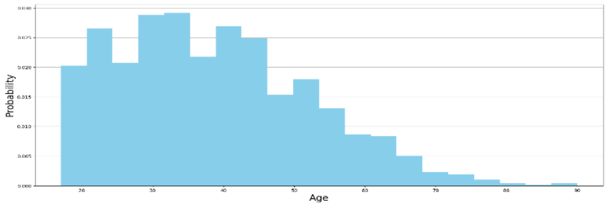

Fig. 13. Probality of earning Vs Age

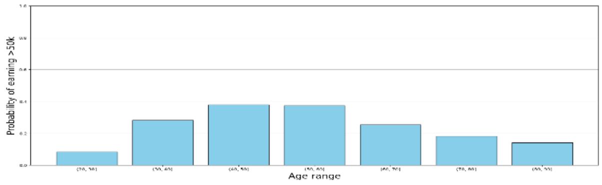

Fig. 14. Probality of earning > 50K Vs Age range

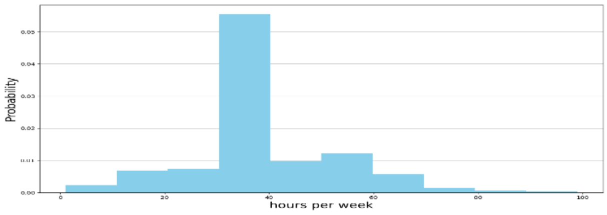

Fig. 15. Probality of earning Vs hours per week

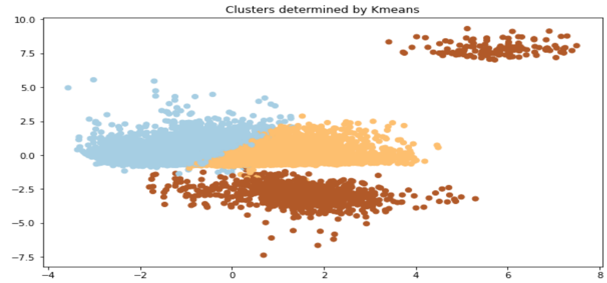

Fig. 16. Visualization of Kmeans Clustering

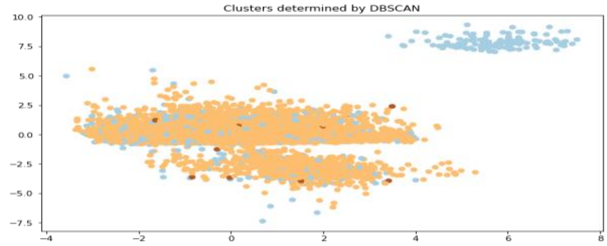

Fig. 17. Clustering Visualization of DBSCAN Clustering

A thorough evaluation is performed to examine the effectiveness of many machine learning models in properly predicting the intended outcome. Following this, a comprehensive evaluation was conducted to compare and evaluate their individual performances. The findings of the research demonstrate that Logistic Regression attained a notable accuracy rate of 84.6%, whereas Categorical Naive Bayes demonstrated a slightly lower accuracy of 75.63%. The Gaussian Naive Bayes algorithm exhibited its predictive capabilities by attaining an accuracy rate of 79.17%. The K-nearest neighbors (KNN) technique exhibited an accuracy rate of 83.18%, whilst the support vector machine (SVM) model yielded a slightly superior accuracy of 84.58%. The performance of Decision Trees was remarkable, as it attained an accuracy rate of 85.12%. The Random Forest model had strong predictive performance, achieving an accuracy rate of 85.56%. The Extended Gradient Boost model demonstrated a higher level of performance in comparison to other alternative models, with an accuracy rate of 86.89%. The aforementioned result highlights the model's adeptness in accurately representing complex associations within the dataset. The efficacy of the artificial neural network (ANN) at capturing intricate patterns is demonstrated by its high accuracy rate of 84.89%. The findings of the research indicate that the integration of Extended Gradient Boost and Artificial Neural Network (ANN) within an ensemble model resulted in the highest level of accuracy, reaching 87%. This finding highlights the potential benefits of integrating these models to improve predictive capabilities. The accuracy values shown offer valuable insights into the respective capabilities of each model, hence aiding in the selection of the best appropriate algorithm for specific prediction tasks.

The performance scores of the clustering algorithms, specifically k-means and DBSCAN, employed in this work were assessed using extrinsic and intrinsic measures. These measures encompassed the Silhouette score, homogeneity score, and other pertinent indicators, as depicted in the table provided below:

TABLE IV. PERFORMANCE TABLE AFTER APPLYING CLUSTERING

Cluster type	Intrinsic Measurers		Extrinsic Measurers	
	Silhouette Score	Cakinski Harabasz Score	Homogeneity Score	V Measure Score
K means Clustering	0.033	518.206	0.1207	0.095
DBSCAN Clustering	0.201	183.242	2.299	2.783

V. . CONCLUTION

This study examines various machine learning techniques using the UCI Adult Dataset to understand income prediction better. The findings demonstrate the need for preprocessing and data management for reliable conclusions. The research follows the "Appropriate Methods to Handle Categorical Data." to ensure analysis integrity and a solid foundation for model evaluation. This research shows how optimization tactics affect prediction powers by doing detailed assessments and comparing models from similar categories where clustering techniques are not suggested for fairness as it go beyond -0.1 and +0.1. The results show that XGBoost and XGBoost-ANN Ensembling can accurately predict income levels based on demographic variables with 87% accuracy. The observation has significant implications for revenue-based decision-making. This study offers guidelines and lessons for future predictive modeling studies for income prediction practitioners and scholars in this complex ecosystem, especially for more complex and recent field data.

REFERENCES

[1] I. Lopez Torres, “Adult UCI Dataset Prediction Analysis,” SSRN Electronic Journal, 2022. doi:10.2139/ssrn.4307371

[2] S. Thapa, “Adult income prediction using various ML algorithms,” SSRN Electronic Journal, 2023. doi:10.2139/ssrn.4325813

[3] A. Lazar, “Income prediction via support Vector Machine,” 2004 International Conference on Machine Learning and Applications, 2004. Proceedings. doi:10.1109/icmla.2004.1383506

[4] M. P. Kim, A. Ghorbani, and J. Zou, “Multiaccuracy,” Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society, 2019. doi:10.1145/3306618.3314287

[5] N. Chakrabarty and S. Biswas, “A statistical approach to adult census income level prediction,” 2018 International Conference on Advances in Computing, Communication Control and Networking (ICACCCN), 2018. doi:10.1109/icacccn.2018.8748528

[6] E. E. Moe, S. S. Win, and K. L. Lai Khine, “Adult income classification using Machine Learning Techniques,” 2023 IEEE Conference on Computer Applications (ICCA), 2023. doi:10.1109/icca51723.2023.10181907

[7] Lekhana_Ganji. (2022, January 11). One Hot Encoding in Machine Learning. GeeksforGeeks. https://www.geeksforgeeks.org/ml-one-hot-encoding-of-datasets-in-python/

[8] Mullin, T. (2022, February 25). DBSCAN Parameter Estimation Using Python. Medium. Retrieved from https://medium.com/@tarammullin/dbscan-parameter-estimation-ff8330e3a3bd

[9] J.-S. Ang, K.-W. Ng, and F.-F. Chua, “Modeling time series data with Deep Learning: A review, analysis, evaluation and future trend,” *2020 8th International Conference on Information Technology and Multimedia (ICIMU)*, 2020. doi:10.1109/icimu49871.2020.9243546

[10] Halford, M. (2022). Prince: Multivariate exploratory data analysis in Python. GitHub. Retrieved from https://github.com/MaxHalford/prince

[11] Chen, T., He, T., Benesty, M., Khotilovich, V., Tang, Y., Cho, H., Chen, K., Mitchell, R., Cano, I. and Zhou, T., 2015. Xgboost: extreme gradient boosting. R package version 0.4-2, 1(4), pp.1-4.

[12] S. Ray, “A Quick Review of Machine Learning Algorithms,” 2019 International Conference on Machine Learning, Big Data, Cloud and Parallel Computing (COMITCon), 2019. doi:10.1109/comitcon.2019.8862451

[13] Kotsiantis, S.B., Zaharakis, I. and Pintelas, P., 2007. Supervised machine learning: A review of classification techniques. Emerging artificial intelligence applications in computer engineering, 160(1), pp.3-24.

[14] Vivek, V., 2022, December. K-Nearest Neighbor (KNN), Soil Evaluation, Classifier and Accuracy. In 2022 5th International Conference on Contemporary Computing and Informatics (IC3I) (pp. 698-705). IEEE.

[15] Osisanwo, F.Y., Akinsola, J.E.T., Awodele, O., Hinmikaiye, J.O., Olakanmi, O. and Akinjobi, J., 2017. Supervised machine learning algorithms: classification and comparison. International Journal of Computer Trends and Technology (IJCTT), 48(3), pp.128-138.

[16] B. Charbuty and A. Abdulazeez, “Classification based on Decision Tree Algorithm for Machine Learning,” Journal of Applied Science and Technology Trends, vol. 2, no. 01, pp. 20–28, 2021. doi:10.38094/jastt20165

[17] C. Janiesch, P. Zschech, and K. Heinrich, “Machine learning and deep learning,” Electronic Markets, vol. 31, no. 3, pp. 685–695, 2021. doi:10.1007/s12525-021-00475-2

[18] D. A. Pisner and D. M. Schnyer, “Support Vector Machine,” Machine Learning, pp. 101–121, 2020. doi:10.1016/b978-0-12-815739-8.00006-7

[19] S. Walczak, “Artificial Neural Networks,” Advanced Methodologies and Technologies in Artificial Intelligence, Computer Simulation, and Human-Computer Interaction, pp. 40–53, 2019. doi:10.4018/978-1-5225-7368-5.ch004

[20] Majumder, S., Chakraborty, J., Bai, G.R., Stolee, K.T. and Menzies, T., 2021. Fair enough: Searching for sufficient measures of fairness. ACM Transactions on Software Engineering and Methodology.

[21] L. Wang, “Research and implementation of machine learning classifier based on Knn,” IOP Conference Series: Materials Science and Engineering, vol. 677, no. 5, p. 052038, 2019. doi:10.1088/1757-899x/677/5/052038

[22] M. Alloghani, D. Al-Jumeily, J. Mustafina, A. Hussain, and A. J. Aljaaf, “A systematic review on supervised and unsupervised machine learning algorithms for Data Science,” Unsupervised and Semi-Supervised Learning, pp. 3–21, 2019. doi:10.1007/978-3-030-22475-2_1

[23] Mullin, T. (2022). Understanding k-Means Clustering in Machine Learning. Towards Data Science, 1(1), 1-10. Retrieved from https://towardsdatascience.com/understanding-k-means-clustering-in-machine-learning-6a6e67336aa1, [Online], (Accessed on 4 August, 2023)

[24] Zhao, W., Yang, W., Wang, H., Zhang, T., Man, D., Liu, T., ... & Guizani, M. (2023). Privacy-Preserving Outsourcing of K-means Clustering for Cloud-Device Collaborative Computing in Space-Air-Ground Integrated IoT. IEEE Internet of Things Journal.

[25] Ruan, Y., Liu, W., Wang, T., Chen, J., Zhou, X., & Sun, Y. (2023). Dominant Partitioning of Discontinuities of Rock Masses Based on DBSCAN Algorithm. Applied Sciences, 13(15), 8917.

[26] Rachwał, A., Popławska, E., Gorgol, I., Cieplak, T., Pliszczuk, D., Skowron, Ł., & Rymarczyk, T. (2023). Determining the Quality of a Dataset in Clustering Terms. Applied Sciences, 13(5), 2942.

[27] Islam, M. A. (2023) ‘Artificial Intelligence for sustainable policy-making and fair politics’. PiP, Westminister Parliament, University of Warwick. doi: 10.13140/RG.2.2.15816.90880

[28] Islam, M. A., Sufian, M. A. and M. H. Sifat (2022) ‘AI, blockchain and self-sovereign identity in higher education’. MRN Annual Conference 2022, UCL, doi: 10.13140/RG.2.2.29117.44008

[29] Sufian, M. A. and Islam, M. A. (2023) ‘Breast cancer detection using machine learning algorithms’. Lets Talk Digital Conference 2023 by NHS. doi: 10.13140/RG.2.2.32320.92168.