The Eurasia Proceedings of Science, Technology, Engineering & Mathematics (EPSTEM)
ISSN: 2602-3199

The Eurasia Proceedings of Science, Technology, Engineering and Mathematics (EPSTEM), 2025

Volume 37, Pages 430-443

ICEAT 2025: International Conference on Engineering and Advanced Technology

Predicting Adult Income Utilizing Various Artificial Intelligence Models

Zahraa Ch. Oleiwi
University of Al-Qadisiyah

Zena H. Khalil
University of Al-Qadisiyah

Salwa Shakir Baawi
University of Al-Qadisiyah

Elaf Hussein Mohammed
University of Al-Qadisiyah

Karrar Khudhair Obayes
University of Al-Qadisiyah

Rahmah Q. Yaseen
University of Al-Qadisiyah

Abstract: This study examines the growth of artificial intelligence (AI) models to forecast adult yearly income, with an objective to use new computational tools to gain a better scientific perspective of economic dynamics. To this end, the study makes a contribution to the development of more effective labour market strategies. Using AI algorithms, the study handles economic information from the UCI Adult Dataset. The performance of four machine learning (ML) models was studied which included support vector machines (SVM), K nearest neighbors (KNN), random forests (RF) and Logistic regression (LR). Top eight predictive features were found using the Recursive Feature Elimination (RFE) approach before the implementation of these models. Among the tested models, SVM was the best performing producing an accuracy of 82.6%, recall of 88.9% and hence the most effective at predicting income level. The results stress the possible role of artificial intelligence in doing financial data analysis and predicting revenue, focusing on its use in employment policies, financial planning, and economic research work.

Keywords: Artificial intelligence, Predictive, Data analysis, UCI adult dataset

Introduction

The world is currently experiencing rapid advancements by using artificial intelligence (AI) techniques through various fronts, including economic and social data analytics. The issues surrounding the inequality of wealth distribution is critical to the societies so that they can enhance economic equality, end the differences in classes and develop sound economic and social policies. Calculations of per capita income is not just a mathematical exercise; it is a kind of tool of greater insight into the economics of societies. Proper forecasting of income will be able to determine employment policies as well as enhance career planning and offer assistance to the disadvantaged groups. Moreover, this analysis also provides useful insights on how to encourage social and economic equality that is a significant objective of sustainable development. The algorithms of artificial

intelligence give the researchers a chance to examine high volueconomicmation regarding the economic, occupational, educational, and demographic characteristics. This way they will be in a position to develop explanatory model, which will provide accurate answers to the complicated questions concerning income distribution. This journal becomes topical in this regard because researchers introduce new solutions to these issues and the path to follow in the direction of AI as the means of seeking answers to social and economic issues.

This paper analyzes the demographics and annual income prediction using the UCI Adult Dataset (Islam et al., 2024). The data pretreatment, management of categorical data and handling missing value are painstakingly controlled to guarantee analytic integrity based on Appropriate Methods to Handle Categorical Data. Various machine learning models are tested (eleven) and their strengths are provided. They consist of neural networks, Naive Bayes (standard and Gaussian), logistic regression, SVMs and decision trees, K-Nearest Neighbors, XGBoost, random forest, and also classification and clustering**.** Concentrating on comparison analysis with optimization evaluation, the research demonstrates that the concentration of optimization strategies and parameter adjustments can enhance the performance of the model. The study reveals that some algorithms can predict income levels correctly with the help of some demographic variables by making comparisons. Although logistic regression provides simple information, more advanced algorithms, such as XGBoost and neural networks, can be used to have a higher quality of predicted results, whereas clustering methods can be used to understand income dynamics better. This paper demonstrates that XGBoost and XGBoost-ANN ensembling can achieve the impressive accuracy level of 87 percent and offer beneficial revenue prediction data. This study highlights the importance of taking good care when pretreating data, choosing an algorithm, optimizing and evaluating predictions, and making relevant conclusions using complex data.

In Chun (2024), the authors addressed the limitations of the current income dynamics prediction tools in Malaysia, noting how they usually do not embrace the state-of-the-art AI tools. These instruments are aimed at improving revenue estimates through machine learning and data mining disciplines. The study examined 423 valid answers of the respondents in Ipoh that do not have extensive income statistics by CRISP-DM. The distribution of income is now better understood in light of findings that depict important economic trends and findings. The initiative also includes an interactive dashboard, which helps to reduce the income divide by visualizing the data in the form of a dashboard. This is a tool that sheds light on the process of decision-making and helps solve problems of economic inequity and income inequality benefiting both society and politicians.

The study by Shuvo (1994) had the task of predicting the annual incomes of people using categorization methods. The data set used is the adult one which comprises of racial and ethnic, gender, marital status, income, education, employment and labour class. The classification methods that were used include the neural networks, logistic regression, random forests and decision trees without extensive use of Bazel. As an example, support vector machines (SVM), and random forests and decision trees, as well as ensemble methods were covered. In the study, hyperparameter tuning and feature engineering were employed to improve the performance of models. This research is one of the most important ones to be mentioned, as it helps to define the most important aspects that affect an annual earnings level, as well as to compare the effectiveness of various methods of categorization. The most accurate model was the Random Forest model (85.73% accuracy, a F1 score of 67.84% versus the 86.1% accuracy and F1 score of 66.64%). It was also found that Artificial Neural Network (ANN) model resulted in the best accuracy. Conversely, the techniques of support vectors and logistic regression were not well. The obtained findings highlight the significance of an unbalanced data consideration and the importance of choosing the right metrics (e.g., F1-score) to evaluate classification algorithms.

The study by Elrahman et al. (2024) applied the Naive Bayes classifier as a probabilistic machine learning model that is well-known because of its effective classification features to construct an income prediction model to determine adult income. The researcher carefully preprocessed the dataset before analysis. This included filling in missing values which may distort the findings and numeric formatting of categorical variables as to fit the model. An annotated training data was used to train the Naive Bayes classifier after preprocessing where the income of the participants was the target variable. The model was created to show the impact of age, education, occupation and matrimonial status on income. The researchers assessed the performance of Naive Bayes on the basis of F1-score, accuracy, recall and precision measures. These indicators showed the weaknesses and strengths of the income forecasts of the model. It was experimentally established that an accuracy of 80.83 percent was gained by Naive Bayes in its prediction.

In this paper, the researcher seeks to discover an application of the method of AI algorithms and feature selection procedures to estimate adult income. It will implement a measure of performance appraisal, which will concentrate on the determinants which affect the distribution of income out there in society. The results of this study contribute to the further realization of the interaction between these components and give grounds to create educational and

economic policies, which will help to improve the quality of life and minimize social tension. This finding may encourage other avenues of AI application in other cardinal societal domains.

In comparison to the previous literature, where the conventional classification models were used without any relevant dataset enhancement directly on the imbalanced data, we propose two-step adder to our work: creation of the balanced dataset by undersampling the majority class; the data-based feature selection using the exhaustive, score-ranked feature selection approach of RFE so that the best features to apply in the classification work are selected.

Research Problem Statement

By examining the factors behind the way income exists, which as we have seen is very critical, we can be able to explain a grave issue of economic inequality, and this can be able to provide some information that perhaps may inform policymaking process successfully. In the era of abundant statistics, one cannot do without AI models in order to make meaningful predictions related to levels of income and the decisive factors that have a dramatic impact on the disparity in income. Nonetheless, among the major technical challenges is the fact that so many attributes are available and identifying the most relevant features is of great challenge. The feature selection procedure that plays a central role in the performance of a model as well as model interpretability also poses a serious question of how to come up with the best subset of features that actually represent the driving forces behind the income variability. This study aims to overcome not only the socioeconomic issue but the technical one as well: by employing smart tools, there should be an improved understanding of the economic processes, the social policy should become more equal, and the strategies of the labor markets should be more effective and competitive.

Research Aim and Objective

Here, we aimed to identify the Adult Income Dataset components are most valuable. The goal of this research is to identify an improved classifier model that uses the most relevant information and is high on accuracy and low on complexity. This ambitious aspiration can be achieved by our research based on five critical aspirations: -

1. Most common and important preprocessing methods were used first and implemented on the Adult Income Dataset to be dataset with high quality and ensure implementing our proposed solutions efficiently. One of these preprocessing steps was focused on and handeled imbalance problem bu undersampling method.
2. Later on, the optimal subset of attributes was selected by appling the method of Recursive Feature Elimination (RFE) on the original dataset, then the system involving four strong techniques of machine learning of Support Vector Machine, Logistic Regression, random Forest as well as K Nearest Neighbors**.** Different benefits are provided by each classifier; each contributes a different methodology for the investigation.
3. Subsequently, a new dataset with the top k features selected from our selection process was built. In addition, a comprehensive function for model assessment using k-fold cross-validation was developed. This sophisticated function is going to easily combine the new dataset with several classifier models, generating useful assessment metrics for the whole folds of each model, and ultimately calculating their average performance.
4. Finally, a comprehensive comparative analysis study was deployed to determine and keep hold of the most efficient classifier model for our data set. This candedated model will work as a reliable instrument, to accurately predict adult income level.

Using this well thought out methodology, what we want to do is work some magic on the Adult Income Dataset and make it into an effective tool by which accurate income predictions are made without compromising its effectiveness or accessibility.

Material and Methods

Data Description

The annual income of individuals depends on several factors. Usually, these factors are influenced by age and gender, education level, occupation, and so on. The annual income dataset, which named “adult”, that used in this study contains 15 variables including target variable which represents the 'Income'. This 'Income' is classified into two classes containing the class of <=50K and the class of >50K. The remaining 14 variables represent the

demographic attributes and other features that describe individuals and enable in prediction of the income level. This dataset is freely found with its description on the UCI machine learning repository at http://www.cs.toronto.edu/~delve/data/adult/desc.html

Logistic Regression (LR)

This classification approach is supervised; logistic regression can be used to predict a target variable. There are just two categories of dependent variables because it is binary. The variable's target class label, as dependent variable, is a simple binary variable, so that it can only take the values 1 or 0, representing success and failure. P(Y=1|X) is mathematically modeled by the logistic regression that consumes X. LR is a probability based statistical technique for the simplest ML algorithms. Such classification issues can be addressed with it- spam detection, diabetes prediction, cancer diagnosis among many others. The probability of a label is determined using sigmoid function when applying logistic regression (LaValley, 2008).

The sigmoid function is a mathematical formulation which turns predicted values to probabilities. It can convert any real number (-∞ to +∞) to (0,1) range number. The sigmoid function can be represented by the below formula (Kyurkchiev & Markov, 2015).

$$\sigma^{z} = \frac{1}{1+e^{-z}} \tag{1}$$

In logistic regression, a cost function quantifies the error, reflecting the divergence between the projected value and the actual value. It measures the model's precision in assessing the correlation between x and y. The resultant value from the cost function is typically referred to as the cost, loss, or mistake. In logistic regression, we utilize a cost function known as cross-entropy or log loss. The formula is as follows (Kim, 2017).

$$\mathcal{L}(p, y) = -(y \log(p) + (1-y) \log(1-p)) \tag{2}$$

Where: y denotes the accurate actual class label, and p signifies the anticipated class label.

Logistic regression class label method mathematical steps can be illustrated as follows:

A. The following formula is used to determine a decision boundary (z) with a dataset with n attributes (Kim, 2017) denoted by x_i , $for\ i = 1,2, \dots n.$

$$z = w_1 x_1 + w_2 x_2 + \cdots + w_n x_n + b \tag{3}$$

B. Use the sigmoid function as described in Eq. (3) to determine the probability of the output z.

C. The error and cost functions are computed in the following manner:

$$J(w, b) = \sum_{i=1}^{M} -(y_i \log(p_i) + (1y_i) \log(1-p_i)) \tag{4}$$

Classification error is measured by the number of classes M. Equations (2) and (3) indicate how to determine the optimal value of w that achieves convergence in terms of least classification error. They are derived from (2) to (5) (Kim, 2017).

Random Forest (RF)

Random Forest is a supervised machine learning approach that is commonly employed and uses ensemble learning techniques for classification and regression applications. The ensemble approach addresses intricate issues by developing and utilizing several classifiers on numerous data subsets. The output class is determined by the

maximum voting of the derived class labels from all classifiers (Elbeltagi et al., 2023). This ensemble approach, Random Forest (RF), during the training phase, independent decision trees are built up on several portions from the original dataset. During testing phase, the output class labels from all decision trees participate in a voting procedure, resulting in the projected output class having the most votes. The advantages of employing Random Forest include its higher performance relative to other machine learning techniques in mitigating overfitting through the utilization of many trees and the aggregation of predictions from these trees, resulting in accurate and exact outcomes (Khajavi & Rastgoo, 2023). Figure 1 illustrates the procedure of RF.

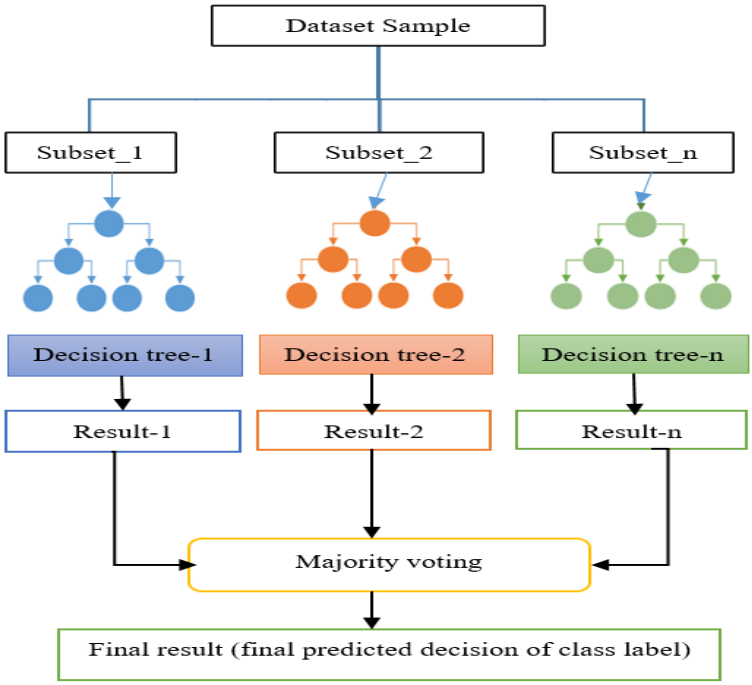

Figure 1. General RF framework

Let d represent the features (dimensions) that exist in the dataset, n denotes the training recordings, and Ntrees signify the number of trees. The training computational cost of Random Forest is O(d*Ntrees*n log n), whereas the testing computational complexity is O(d*Ntrees).

Support Vector Machine (SVM)

It is a supervised machine learning approach that is commonly applied in binary classification. Its fundamental architecture is hyperplanes that, with a broad margin and significant classification accuracy, divides the two classes. Where the margin is the angle at which the two lines joining the nearest class points form a right angle; the points themselves are called support vectors. Using a nonlinear function to transport data to a high-dimensional space where it may be linearly separated, SVM achieves strong classification results for both linear and nonlinear separable.

SVM uses the training set to do linear classification, and it looks like $\{(x_i,y_i)\}$ for each i=1,2,3,…N, in where x_i is an n-dimensional feature vector, n is the original data space's feature count, N is the dataset's sample size, and y_i is a label of target, which might be +1 or -1, signifying the class that includes x_i. Finding the best possible hyperplane that divides x_i points into class +1 and class -1 while leaving the largest possible space between the hyperplane and the closest points x_i to either class is the goal of linear SVM. The group of points with the following requirements is known as a hyperplane:

$$w^T x + b = 0 \tag{5}$$

The bias is represented by b, and the weight normal vector that may be changed to the hyperplane is denoted by w.

$$w^T x + b = +1 \tag{6}$$

The points are classified as -1 for any location on or below the line as follows:

$$w^T x + b = -1 \tag{7}$$

This is because the margin's geometric definition represents the distance between above two lines, which is ($\|w\|$ In order to increase the margin, the term ($\|w\|$) must be minimized, which is defined like: Minimize $\frac{\|w\|^2}{2}$ (function of objective).

$$\textit{Subject to constrain} \rightarrow y_i(w.x + b) \geq 1, \quad \forall\, i = 1,2,3,\ldots..N$$

The following is the formulation of the optimization function in the case of nonlinear separable patterns in the data:

$$\textit{Reduce the objective function} \rightarrow \frac{\|w\|^2}{2} + C\left(\sum_{i=1}^{N} \xi_i\right)$$

$$\textit{Subject to constrain} \rightarrow y_i(w.x + b) \geq 1 - \xi_i, \xi_i \geq 0, \quad \forall\, i = 1,2,3,\ldots..N$$

The slack variable ξ_iis utilized to normalize the overfitting problem, and the regularization parameter C is employed to produce a soft margin by penalizing misclassification cases. It includes two terms in the objective function. Therefore, a big value of C will result in a hard margin and a severely penalized misclassification case (small gap). In order to overcome the overfitting and get superior generalization performance to achieve better prediction accuracy with new example, a misclassification example is penalized lightly when C is small and the margin is soft (big gap).

After extending the nonlinear SVM with inner product kernel notation and moving the original data points. The nonlinear decision function may be effectively treated as the Euclidean space H grows high-dimensional, allowing for the separation of the data by a linear decision boundary.

∅: $R^n \rightarrow H$, where R^n is the feature vectors' initial space. During training, the kernel function K is defined concerning $\emptyset(x_i)$ as:-

$$K(x_i, x_j) = \emptyset(x_i).\emptyset(x_j) \tag{8}$$

Therefore, the w in the space of transfer will be:

$$w = \sum_{i=1}^{N} \alpha_i y_i \emptyset(x_i) \tag{9}$$

The dot product of w and the hyperplane will generate a function of the decision boundary, as:

$$f(x) = w.\emptyset(x) = sign\left(\sum_{i=1}^{N} \alpha_i y_i K(x, x_i) + b\right) \tag{10}$$

The kernel has a large number of functions. The RBF (Gaussian Radial Basis Kernel) is the most popular one in this study because it is described as follows:

$$K(x_i, x_j) = e^{-\left(\frac{\|x_i - x_j\|^2}{2\sigma^2}\right)} \tag{11}$$

The accuracy of SVM classification is significantly affected by the kernel width, σ. Classification outcomes from both intra- and inter-patient schemes show that the regularization parameter C is similarly critical in regulating classification accuracy (Abdullah & Abdulazeez, 2021).

Known to be $O(N^3)$ for training time, $O(N^2)$ for space complexity, and O(kd) for testing (prediction) time, the variables N, d, and k denote the dataset size, dataset dimensions, and support vectors, respectively, in the SVM technique (Wu et al., 2021).

K-Nearest Neighbor (KNN)

It is a straightforward and nonparametric technique used in supervised learning to solve classification and regression problems. It categorizes a new test sample based on its similarity to the training dataset. This method does not assume any characteristics of the underlying data, categorizing it as nonparametric. KNN is classified as "lazy learning" because it just retains the information during the training phase without any learning through the training data. During assessment or forecasting, it uses Euclidean distance to measure the similarity of an unseen sample with retained training samples. Distance in Equation (12).

$$E_{distance} = \sqrt{(x_1 - x_2)^2 + (y_1 - y_2)^2} \tag{12}$$

The KNN algorithm's phases are simple to build and don't require any training time; nonetheless, it requires a lot of memory and prediction time. To improve KNN's efficacy while capitalizing on its simplicity, this project is designed to tackle the problem at hand (Dhivya & Bazilabanu, 2023), the KNN approach has a testing (prediction) time complexity of $O(k_1 dN)$, a space complexity of O(dN), where N is the dataset size, d is a dataset dimension, and k_1 is the Nearest Neighbors number (This, 2022).

Performance Measures

Classification and regression algorithms can be evaluated using various measures. Comparing the predicted labels against the actual labels is a common way to assess supervised learning. The most popular metrics include F1-score, accuracy, recall, precision, and also confusion matrix (Ihsanto et al., 2020).

A. The confusion matrix evaluates the classification model performance. The confusion matrix juxtaposes the projected estimates produced from the model of machine learning against the actual target values (Liang, 2022), as in Figure 2. To make an appropriate assessment, one must comprehend the results of the predictions. By definition, when the actual positive value matches the prediction, it represents the true positive (TP), while the matching of the predicted and actual negative values represents the true negative (TN).The prediction is considered as false positive (FP) when the predicted value is positive and the actual value is negative, and considered as false negative (FN) when model's prediction is negative and the actual value is positive.

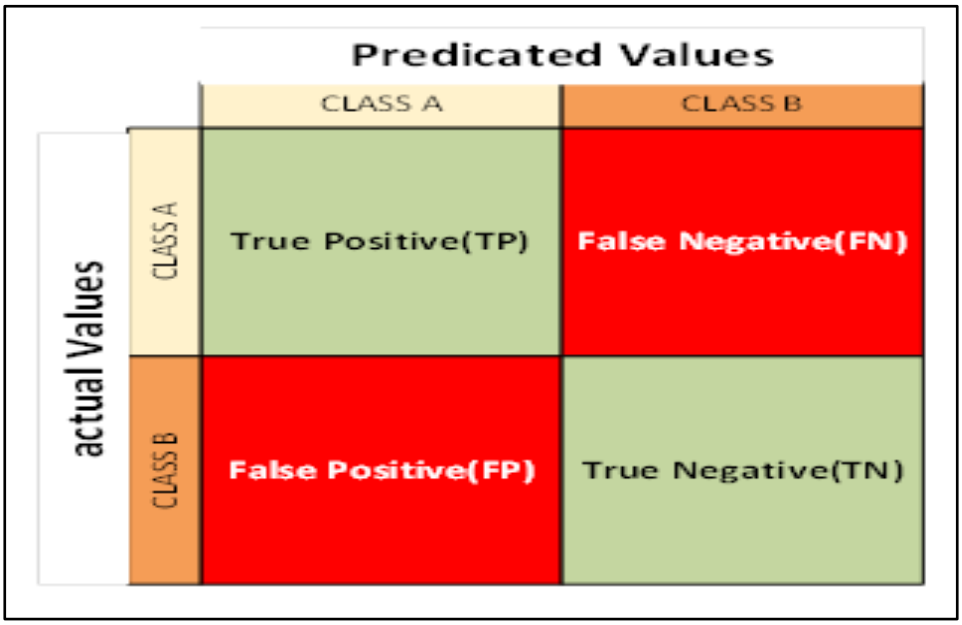

Figure 2. Confusion matrix

B. The percentage of correct predictions of any classifier is measured using accuracy. It computes the ratio of accurate forecasts to total predictions. Nevertheless, accuracy is suboptimal for unbalanced classes. If a model forecasts all instances as the majority class, accuracy may be elevated, although it lacks precision. Accuracy denotes the precision of real predictions and is appropriate for well-balanced, impartial classification tasks devoid of class imbalance. The accuracy is determined as follows (Dalianis, 2018).

$$ACC = \frac{TP + TN}{TP + FN + FP + TN} \tag{13}$$

C. Recall denotes the precision of accurately identified positive observations. It is sometimes referred to as sensitivity. A recall is a dependable assessment indicator when the objective is to identify the maximum number of positives. Recalling metric is the value of the number of accurately identified positive samples divided by the total number of actual positive samples. The value of Recall should be maximizedclosed to 1, which is considered as ideal value. When false negatives (FN) are more consequential than false positives (FP), The recall is especially significant in this case. The recall is computed as follows (Dalianis, 2018)

$$R = \frac{TP}{TP + FN} \tag{14}$$

D. Precision refers to the degree of correctness achieved, which influences the accuracy. It is calculated as the correct positive prediction to the overall sum of positive predictions. Precision is calculated as the number of correctly categorized positive classes to the overall sum of projected positive classes. Ideally, this accuracy should be high (close to 1). When 'false positives (FP)' outweigh 'false negatives (FN)', precision becomes invaluable. Accuracy is determined as follows (Dalianis, 2018)

$$P = \frac{TP}{TP + FP} \tag{15}$$

E. F1-score is bounded of 0 and 1, which represents the harmonic meaning between recall and precision i.e. penalizes outlier values. Because higher values do not affect it, the harmonic meaning is recommended over simple averages. Low precision results in a low F1-score; poor recall values have the same effect. Sometimes, the necessity of precision and recall is unclear. Therefore, we integrate them to get a comprehensive measure. Increasing precision could cause recall to drop and vice versa. The F1-score captures both trends in one number. Ideally, a high F1-score of 1 is desired. The F1-score is derived as follows (Dalianis, 2018)

$$F_{Score} = 2 * \frac{Precision * Recall}{Precision + Recal} \tag{16}$$

Analyzing meticulous/ multilevel targets calls for the average parameter. Metrics can be calculated in two ways:

- Macro" computes the unweighted average for every label by first calculating measures for them. It ignores label imbalance, though.
- "Weighted" measures each label's metrics and averages them weighted by support (the number of actual cases for each label). It compensates for label imbalance, which can cause an F-score beyond the accuracy and recall range.

In unbalanced data, conventional assessment measures like accuracy could be deceptive; thus, special measures such as F1-score, precision and recall should be applied.

Proposed Methodology

The proposed methodology will start with the very thorough preprocessing stage where particular care will have to be taken to the problem of class imbalance in the Adult Income Dataset. In the process, we did not randomize the number of samples like in some of the earlier studies, which employed oversampling methods, which could result in the addition of synthetic data and lead to the distortion of data distribution. This will not deform the genuine nature of the dataset since we will only use real, empirically observed cases, and we feel will result in

more stable and understandable findings. After data balancing, we used ranked feature selection procedure or performance driven method. In particular, we ordered the features by their importance scores, and in each case tested all the possible combinations of features, either ordering by importance and excluding all but one feature or two features, etc, until the order ceased to improve the model, in order to come up with the best sum of features that would lead to the best model performance. This thorough step helped us to make the list of selected features both as short as possible and as informative as possible, which enhanced accuracy and the possibility of understanding the model. The beauty of this two-step approach is that it not only improves the quality of data but also improves the predictive capability of the model thus making our approach unlike others.

An automated system was implemented to figure out the optimal number of features. Instead of human entry of feature counts, the system will test every possible feature count.Where in each test phase, the processes of running feature selection train the classification model using the reliable k-fold validation, and check the performance were done. The feature count that gives the best score is chosen as "optimal number," and the original data will be filtered down to only those best features.

The cross-validation assessment process was used to improve the performance of predictive model, where the overfiiting is addresed and controlled efficiently. The k-fold in this study is used with k=10 to partition the dataset into 10 subsets (folds) equally.Two principals are required as inputs to Algorithm 1: the freshly generated dataset and the model evaluation. It methodically cycles through each fold, employing nine folds for training and one-fold for testing in each iteration. This enhances an in-depth review of the model's efficacy. The outputs of this algorithm are average evaluation measures, accuracy, recall, precision, and F1-score. These outputs provide an overall view of effectiveness of the models. Additionally, it outlines the metrics of evaluation that were collected from each fold, along with the specific datasets for training and testing associated with the fold that best generated the metrics for performance. This thorough view on evaluation ensures the dependability and robustness of the model are seriously tested, therefore minimizing the risk of overfitting the training data.

Algorithm 1: Model Cross Validation Evaluation (MCVE)
Input: Adult dataset (features as x, target as y), Classification model
Output: Each k evaluating metrics, Evaluation metrics' average
Model-Eva ← ({F1[ ], Acc[ ], P [ ], R [ ] }
K-Fold ← Call Model-Selection , Call Stratified-Shuffle-Split
Max ← 0
for each (train-indices, test-indices in K-Fold.Split (new-x, y)
begin
    x-train, x-test, y-train, y-test ← K-Fold-train-test-Split(new-x, y, train-indices, test-indices)
    Model.fit(x-train, y-train), pred ← Model. Predict (x-test)
    Model-Eva Acc[ ]. Append (metrics. Accuracy-score (y-test, pred))
    Model-Eva F1 [ ] . Append (metrics. F1-score (y-test, pred))
    Model-Eva P [ ] . Append (metrics. Precision-score (y-test, pred))
    Model-Eva R [ ]. Append (metrics. Recall-score (y-test, pred))
Endfor
Model-Eva ← pd-DataFram (Model-Eva)
Return Model-Eva

The overall structure of the suggested system are illustrated in Figure (3).

A. Data Preprocessing: initial phase is cleansing the datasets by removing extraneous attributes that do not enhance the predictive task. Subsequently, we methodically transform all nominal and categorical qualities into numerical values by a factorization technique. This transformation is uniformly applied to both datasets to guarantee uniformity in the data structure, which is essential for good modeling.
B. Data Splitting: Following data cleansing, we divide it into two primary components: the input characteristics, which serve as the predictive / target variables, that represents an outcome we intend to forecast.
C. Data Normalization: During training the data should be normalized to provide an equal contribution of all the variables. Normalization is the next phase. This process uses StandardScaler, whereby the features are standardized, that is, brought to mean zero and then the result is divided by the standard deviation. Therefore, the entire attributes are normalised or uniformised, a significant attribute towards improving model behaviour. Both training and test data sets require the scale so as to avoid data leakage.
D. Data Resampling: Due to the unbalanced characteristics of the datasets, marked by a surplus of instances in the majority class, different random undersampling techniques were employed. This strategy aims to match the sample size of both majority and minority classes by diminishing the majority sample size. Subsequent to

the use of these procedures, the resulting balanced dataset is primed for input into the feature selection process to determine the optimal number of feature imports for decision-making to minimize overfitting, reduce model complexity and concentrate on the almost relevant and informative features. Then a newly balanced dataset, including optimum features, is utilized in classification models.

E. By studying the underlying causes of the development of income distribution, we can be in a good place to describe an acute issue of economic inequality and, perhaps, can find valuable insights that may inform policymaking successfully. The need to utilize AI models in such a period of availability of high statistics is that it is only significant enough to predict the various income levels as well as how those significant factors influence the income variations question drastically. Yet it is also a major technical issue to find the most relevant features among all available ones which is a huge number. It is not only that feature selection is a key to both model performance and interpretability, but also the relevant issue of finding the optimal set of features that provides the most accurate representation of the underlying forces behind income between individuals was uncovered in our study through comprehensive ranked assessment strategy. Specifically, we applied a feature-ranking technique to rank features once they were scored in terms of importance and then repeatedly performed all subsets of features in that order (one feature, then two features, etc) and checked the model performance at each step. The subset that is associated with the evaluation metrics was selected as the optimal set of features. This approach to feature selection is data-based and performance-focused due to the addition of this method to the feature selection process, which enhances the credibility and interpretability of the obtained model.

F. Model Development: Following the preprocessing and balancing of the data, we proceed to the model development phase, during which we construct four machine-learning models utilizing the refined dataset.

G. Model Evaluation: After training the models, it is imperative that they are rigorously tested and assessed for their performance. This assessment employs the same procedure as Algorithm 1.

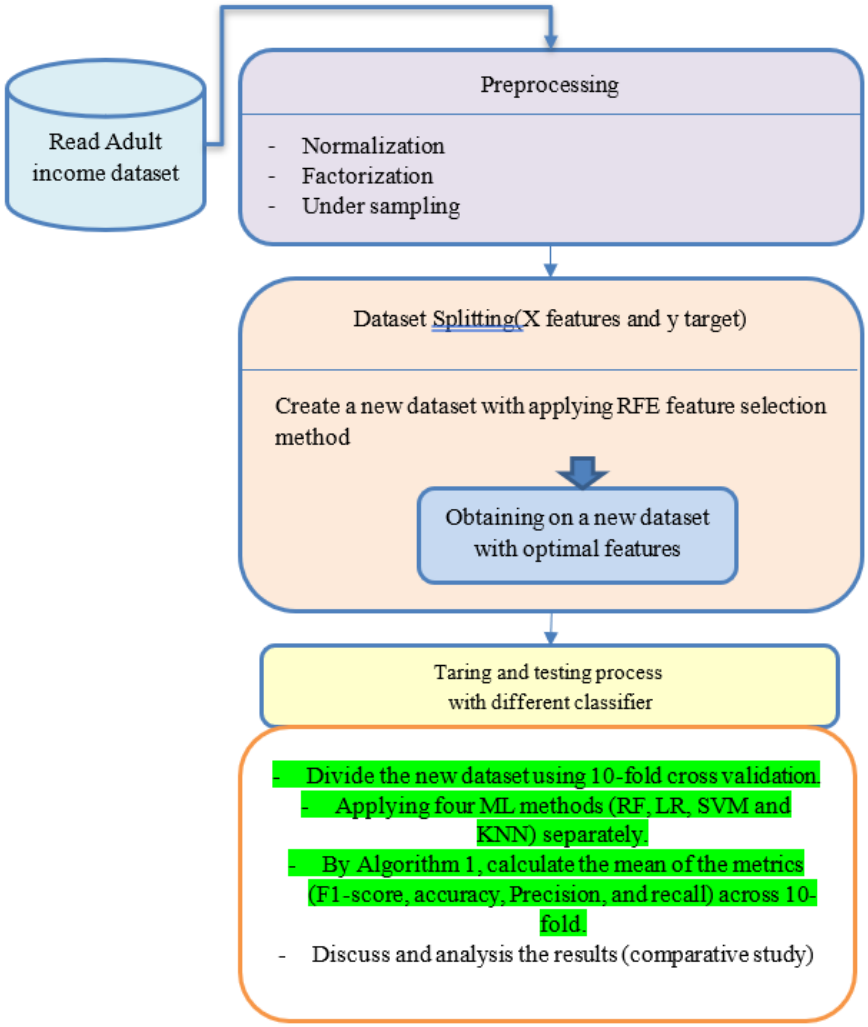

Figure 3. General framework of the proposed system

Results and Discussion

In this section, the results of our implementation in Python were examined, in aim of demonstrating the efficacy of the proposed approach in handling imbalanced problems and enhancing the classification accuracy. The effectivness of the proposed model is evaluated using many metrics: accuracy (ACC), recall (R), precision (P), and F1-score. After the distribution of each dataset to its appropriate classes, we produced an extensive study of the data before and after the use of the under-sampling technique, as outlined in Table 1.

Table 1. Dataset distribution

Dataset size before under-sampling			Dataset size after under sampling		
Total size	Majority class	Minority class	Total size	Majority class	Minority class
48842	37155	11687	20000	10000	10000

The suggested classification scheme comprised two significant phases. The initial phase was the feature selection procedure, which evaluated all 14 prospective characteristics. Figure 4 displays the outcomes of applying Recursive Feature Elimination (RFE), depicting the correlation between the F1 score and no. of utilised features. The ideal of identified characteristics number was 8, which attained the maximum F1 score.

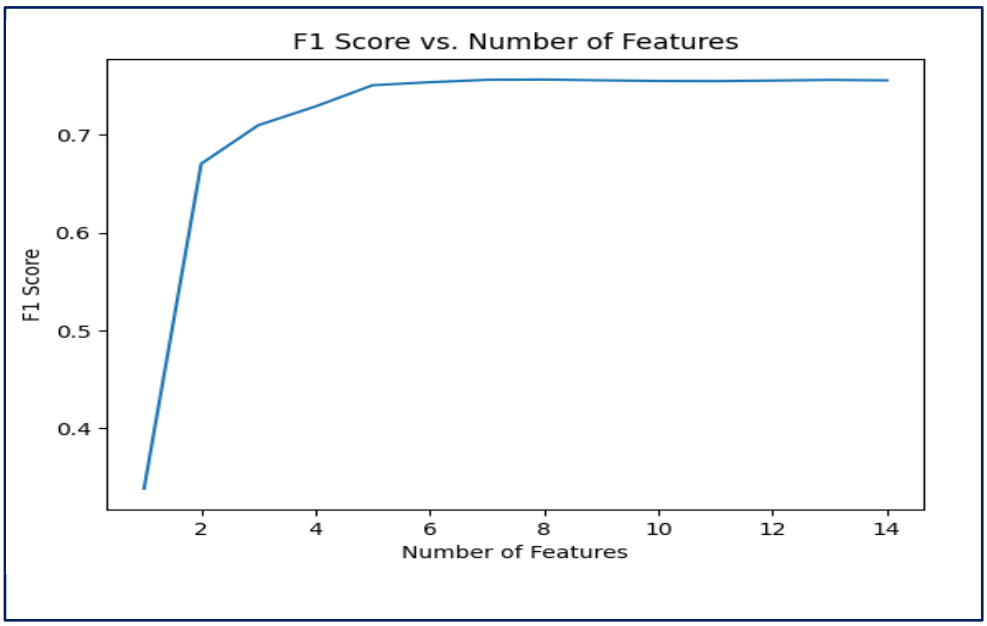

Figure 4. Number of features against F1-score using RFE

This section presents the findings of four classifiers applied to a new dataset, which was constructed by choosing the eight most effective features found by the Recursive Feature Elimination (RFE) approach. Tables 2 and 3 present the evaluation of each classifier's performance based on F1-score, recall, precision, and accuracy.

Table 2. The evaluation measures average of k-fold using different classifiers

The ML algorithm	ACC%	R%	P%	F1 score %
LR	76	74.3	76.8	75.6
KNN	79.7	81	78.9	80
RF	80.7	80.7	80.7	80.7
SVM	82	87.2	79.1	82.9

From Table 2, it can be viewed that the SVM recorded better performance than the other considered models. A bit later, the Random Forest (RF) took the two, K-Nearest Neighbors (KNN) and Logistic Regression (LR), and ran them over. The remarkable success of SVM and RF is attributable to their high capability to accommodate the complexities within the data. Such complications include the maintenance of class disparities, crossing high dimensional zones, and reconciling overlaps of separate classes. Support Vector Machines (SVM) are adeptly able to create perfect hyperplanes for class segregation while Random Forest makes use of ensemble learning to increase accuracy and stability, making both classifiers very ideal for complicated datasets.

Table 3. Maximum evaluation measures of k-fold using different classifiers

ML algorithm	K	ACC%	R%	P%	F1_score%
LR	10	77.1	75.7	77.8	76.7
KNN	7	80.8	82.7	79.7	81.1
RF	8	81.5	81	81.9	81.4
SVM	9	82.6	88.9	78.9	83.6

Recall is a vital statistic used in detection tasks because it evaluates how relevant the model is in recognizing relevant events. The Support Vector Machine (SVM) was the highest recall-based model in the study with recall of 88.9. The higher recall suggests that SVM is a good detector of true positive and can thus be heavily competitive in predicting adult income.

To successfully evaluate the functionality of the model, one would want to look at the classification that is shown in the course of the research that would help to draw attention to the performance of the model in various classes within the dataset. This study provides a critical discussion on the strengths and weaknesses of the model that can be inferred using accuracy, recall and F1-score results of each class. The application of SVM in predicting income in adults is described and detailed in table 4 which offers extensive results on the application of SVM.

Table 4. The classification report using SVM on the dataset of adult income

	Class label	P%	R%	F1_score%
	0	85	77	80
	1	80	87	83
macro avg		82	82	82
Weighted avg		82	82	82

Strong correlation in performance measures across all classes is clear in as Table 4, highlighting the effectiveness of the Support Vector Machine (SVM) method in addressing issues related to minority classes. The ability is essential in the case where some classes are slightly underrepresented in the training data due to the tendency of the model to form a bias against the most prevalent one. This evaluation can be justified by the fact that the given confusion matrix (as in Figure 5) includes detailed information on the true positive and false negative values of each category, so it is possible to focus on the fact that SVM effectively detects the cases that belong to the minority group.

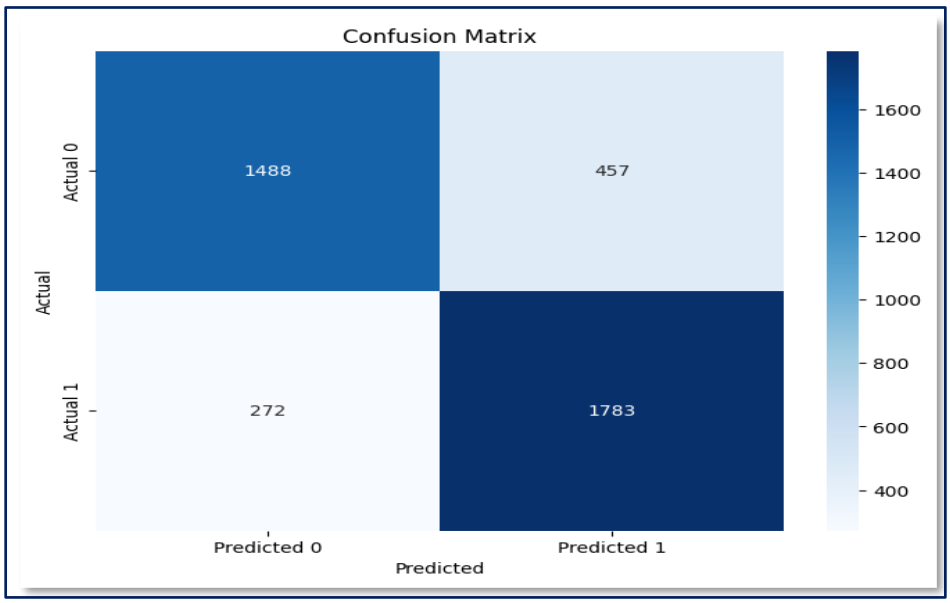

Figure 5. Confusion matrix of SVM

The efficacy of SVM classifier in classification is explained by figure 5. The figures indicate that the SVM falsely denies 457 positive samples as negative (false negatives, FN). Conversely, it identifies accurately 1,488 positive samples (true positives, TP), which means that it has a strong ability to identify positive cases. The SVM performance is good with the negative class in that it has correctly classified the negative samples numbering

1,783 as negative samples (true negatives, TN). It is crucial to admit that the classifier makes errors in this category and misclassifies 272 negative samples as positive ones (false positives, FP).

Along with demonstrating where the SVM classifier has been successful, these figures demonstrate where it has been unsuccessful, and the most apparent is the inability to draw the line between the negative samples. Gaining insight into these misclassifications can help with model tuning in the future, leading to better classification accuracy overall.

Conclusion

The main goals of the Annual Income Forecasting Project were achieved when a reliable forecasting model was developed by analyzing economic and demographic data using AI and ML methods. In a recent study of UCI Adult Dataset through Recursive Feature Elimination (RFE) technique, eight most significant features influencing the annual income of a person were determined. The features that were choosen testified to the highest percent of relevance; hence their predictive importance became evident. To strictly assess the chosen features effectiveness, diverse machine learning classifiers were utilized, including K-Nearest Neighbors, Random Forest, Logistic Regression, and Support Vector Machines. The Support Vector Machine model has repeatedly topped its counterparts in terms of different performance measures and is more accurate predictive and insusceptible to errors. The findings indicate the ability of SVM to classify income and support the need to perform extreme feature selection to performance improvement.

Data imbalances and missing values are two essiential problems this study had limitations due to them although it achived promisie results. Several improvment phases related to data processing methods were applied to handle these problems such as undersampling, normalization, and standardization. The Recursive Feature Elimination (RFE) approach to identify the optimal features for enhancing the model's accuracy was employed by overcoming the overfitting caused by too many features.Trait preferences in decision-making should be emphasized. This initiative represents one of the checkpoints in our efforts to determine the factors influencing income distribution. The lessons learned in social and economic development can be utilized by policymakers to promote sustainable development and economic equity.

Scientific Ethics Declaration

* The authors declares that the scientific ethical and legal responsibility of this article published in EPSTEM journal belongs to the authors.

Conflict of Interest

* The authors declare that they have no conflicts of interest

Funding

* There is no funding.

Acknowledgements or Notes

* This article was presented as an oral presentation at the International Conference on Engineering and Advanced Technology (ICEAT) held in Selangor, Malaysia on July 23-24, 2025.

References

Abdullah, D. M., & Abdulazeez, A. M. (2021). Machine learning applications based on SVM classification: A review. *Qubahan Academic Journal*, *1*(2), 81-90.

Chun, A. S. (2024). *A new perspective on income earning using AI*. http://eprints.utar.edu.my/id/eprint/6617

Dalianis, H. (2018). Evaluation metrics and evaluation. *Clinical Text Mining*, 45-53.

Dhivya, P., & Bazilabanu, A. (2023). Deep hyper optimization approach for disease classification using artificial intelligence. *Data and Knowledge Engineering*, *145*, 1-10.

Elbeltagi, A., Pande, C. B., Kumar, M., Tolche, A. D., Singh, S. K., Kumar, A., & Vishwakarma, D. K. (2023). Prediction of meteorological drought and standardized precipitation index based on the random forest (RF), random tree (RT), and Gaussian process regression (GPR) models. *Environmental Science and Pollution Research*, *30*(15), 43183-43202.

Elrahman, A. A., Elrahman, A. A., Riad, M. R., & Abdelgwad, M. M. (2024). Predicting adults' income using Naive Bayes classifier. 0-7. *ResearchSquare.*

Ihsanto, E., Ramli, K., Sudiana, D., & Gunawan, T. S. (2020). An efficient algorithm for cardiac arrhythmia classification using ensemble of depthwise separable convolutional neural networks. *Applied Sciences*, *10*(2).

Islam, A., Nag, A., Roy, N., Dey, A. R., Firoz, S. M., Fahim, A., & Ghosh, A. (2024, March). An investigation into the prediction of annual income levels through the utilization of demographic features employing the modified UCI adult dataset. *2023 International Conference on Computing, Communication, and Intelligent Systems (ICCCIS)* (pp. 1080-1086).

Khajavi, H., & Rastgoo, A. (2023). Predicting the carbon dioxide emission caused by road transport using a Random Forest (RF) model combined by Meta-Heuristic Algorithms. *Sustainable Cities and Society*, *93*, 1-9.

Kim, P. (2017). MATLAB deep learning:. *With machine learning, neural networks and artificial intelligence*. 130(21), 151.

Kyurkchiev, N., & Markov, S. (2015). Sigmoidal functions: Some computational and modelling aspects. *Biomath Communications*, *1*(2), 1-19.

LaValley, M. P. (2008). Logistic regression. *Circulation*, *117*(18), 2395-2399.

Liang, J. (2022). Confusion matrix: Machine learning. *POGIL Activity Clearinghouse*, *3*(4). Special Issue.

Wu, G., Huang, Y., Zhu, C., Song, L., & Zhang, W. (2021, May). SVM based fast CU partitioning algorithm for VVC intra coding. *Proceedings - IEEE International Symposium on Circuits and Systems* (pp. 1-2).

Author(s) Information

Zahraa Ch. Oleiwi
University of Al-Qadisiyah
Diwaniyah, Iraq
Contact e-mail: *zahraa.chaffat@qu.edu.iq*

Zena H. Khalil
University of Al-Qadisiyah
Diwaniyah, Iraq

Salwa Sh. Baawi
University of Al-Qadisiyah
Diwaniyah, Iraq

Elaf H. Mohammed
University of Al-Qadisiyah
Diwaniyah, Iraq

Karrar K. Obayes
University of Al-Qadisiyah
Diwaniyah, Iraq

Rahmah Q. Yaseen
University of Al-Qadisiyah
Diwaniyah, Iraq

To cite this article:

Oleiwi, Z. C., Khalil, Z. H., Baawi, S. S., Mohammed, E. H., Obayes, K. K., & Yaseen, R. Q. (2025). Predicting adult income utilizing various artificial intelligence models. *The Eurasia Proceedings of Science, Technology, Engineering and Mathematics (EPSTEM), 37,* 430-443.