2022 IEEE International Conference on Data Science and Information System (ICDSIS)

Optimized Features Based Machine Learning Model for Adult Salary Prediction

1st Lokesh Pawar
Computer Science and Engineering
Chandigarh University
Mohali, India
lokesh.pawar@gmail.com

2nd Ajay Kumar Saw
Computer Science and Engineering
Chandigarh University
Mohali, India
sawajay9572@gmail.com

3rd Abhay Tomar
Computer Science and Engineering
Chandigarh University
Mohali, India
abhaysinghtomar56@gmail.com

4th Navneet Kaur
Computer Science and Engineering
Chandigarh University
Mohali, India
navneetsehgal5@gmail.com

2022 IEEE International Conference on Data Science and Information System (ICDSIS) | 978-1-6654-9801-2/22/$31.00 ©2022 IEEE | DOI: 10.1109/ICDSIS55133.2022.9915981

***Abstract*—A nation's economic stability is bolstered and long-term progress is ensured by the idea of universal moral equality. Many governments are putting a lot of effort into addressing this problem and coming up with a workable answer. In order to determine what decisions an adult can make in the future and whether or not he is financially independent and secure, it is predicted in this article whether his wage will be larger than $50,000 per year or not. Data mining technologies and machine learning algorithms both play a big part in this. This paper focuses on eliminating useless features using various machine learning approaches and algorithms and there is room for improvement. So, using the Gini Index, prominent features are identified and prioritized which applied on machine learning algorithm boosted the performance upto accuracy of 87.82 percent.**

Keywords—Artificial Intelligence, Machine Learning Algorithm, Bayes Net (BN), Logistic, Decision Tree (DT), Random Forest (RF), Salary, Prediction.

I. INTRODUCTION

An adult [1] lies into the category of people whose income is greater than or equal to $50k/year or less than that, we are predicting so to go one step further and predict his lifestyle and living standard [2] and also, situations in different stages and possibilities of life, one of the most important could be him being financially independent and secure [3]. We will be talking about his situation when an emergency falls on him like a medical emergency if he or his family have had an accident, will he be able to handle it? On the basis of his salary [4] will he be eligible for getting a loan, the choices he will be making for himself and his family either could be daily basic needs like clothes, type of food, house etc., or affording travel mediums, travel destinations etc.

Fig. 1 presents the prominent features of income dataset of census. In the process of income analysis and prediction, there are several important steps. First of all, identification of data selection [5] in improving the chosen dataset by feature selection, visualization of the dataset using different machine learning algorithms, analyzing it over different test options and applying data mining techniques for evaluating the results. Since a classifier provides a function that classifies instances (data item) into several predefined classes which in many cases gives ability to understand the output from the algorithms which is very crucial in designing and analyzing any model. Decision Tree is one among the foremost usually used, reasonable approaches for supervised learning. It can be used to tackle problems of both Regression and Classification jobs with the latter being placed greater into sensible application decision [6]. The Gini Index is calculated by subtracting the sum of the squared chances of every elegance from one. It favors frequently the bigger partitions and is quite simple to implement [7]. In other words, it calculates the opportunity of a randomly selected feature that was classified incorrectly. The Gini Index in equation (1) works on specific variables and offers the effect in "success" or "failure" phrases.

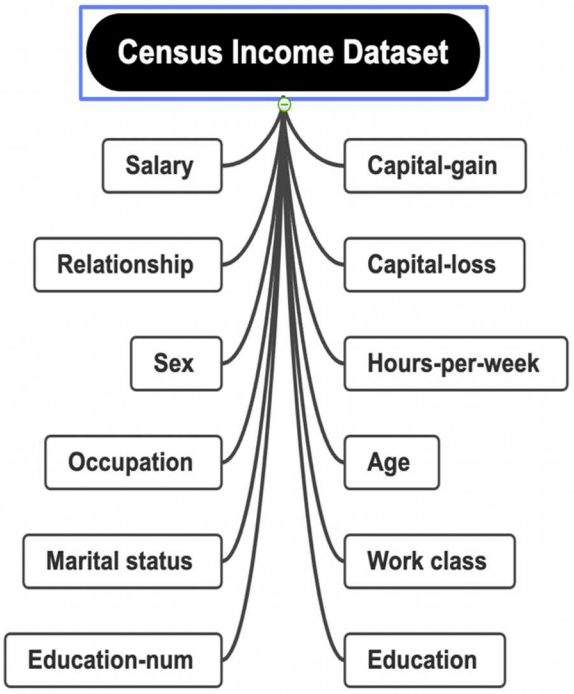

Fig. 1. Prominent features of Census Income Dataset

$$\text{Gini Index} = 1 - \sum_{i=1}^{n} (p_i)^2 \quad (1)$$

Where p_i is the probability of an object being classified to a particular class. We did resampling of the dataset to achieve greater accuracy. Re-sampling [8] is a chain of techniques for reassembling your sample datasets, including training sets and validation sets. It can offer more "useful" different pattern sets for the learning process in some way. The major types of resampling are Cross-validation and bootstrap techniques are majorly used in resampling. We are ranking the features as shown in Table 1 on the basis of their weightage [9-10] so that we could make proper feature selection and improve the

accuracy by removing purposeless and inefficient features [11].

TABLE. I. VARIANCE AND RANKING OF FEATURES

S. No.	Weightage	Rank
1	0.33518	5
2	0.33119	6
3	0.26971	8
4	0.23404	1
5	0.22969	13
6	0.22333	11
7	0.216	10
8	0.15052	12
9	0.10887	4
10	0.09743	7
11	0.08224	9
12	0.07077	2
13	0.03301	14
14	0.00948	3

In this paper, the dataset is procured from UCI machine learning repositories and is used to evaluate the performance of learning models. The characteristics of the dataset are multivariate. Its task is classification. The number of instances in this dataset is 48842, it was given on 1996-05-01 and the no. of attributes is 14.

II. LITERATURE REVIEW

This section focuses on the research of several researchers in this topic and the following are the views and perspectives given:

Sidra et al. focused on predicting whether the income of employees exceeds $50k/year. The data and instance are inconsistent and in a dirty form that we can't even apply the ETL process to make it clean so the author also works on the ETL process to make the data consistent. The Analysis is being done numerically by taking census income dataset which collects basic information about the person.

Navoneel Chakrabarty et al. aim on the role of ensemble learning algorithm, and to show the importance of learning algorithm techniques in providing a best solution to the wealth disparity problem. The author is predicting whether the person falls in the category of people whose income is greater than or equal to $50k or in the category less than that based on certain attributes using the classification process.

Chet et al. studied the census income dataset to predict the income of an individual by visualizing each feature and what impact it has on the possibility to earn more than $ 50k per year. They used feature selection for better prediction and optimization.

Bramesh et al. focused on finding the solution for the problem that the details of employees are not in general to predict the salary of an adult. So to overcome this problem, the author has examined the machine learning classifiers deeply by using a UCI census dataset or public database. The paper also focuses on the best classifier to be used for this purpose.

Sumit et al. aim at predicting the salary by using classification algorithms and exploratory data analysis to get information about the dataset, preprocessing to make data ready for testing and then testing the data. The Author also compared the performances of various classification models with various performance metrics.

William H. Hampton et al. aim at novel analytic approach using various machine learning algorithms to draw the relationship between income and various variables (demographic and oblique) and behaviors like delay discounting. The Author has used the holdout data set to test the validity of their findings.

Sabitha et al. focused on various data mining methods like clustering for the analysis of data and various feature selection techniques that is to be applied to the clustering algorithm on the adult dataset for the comparative study in terms of accuracy and execution time of clustering for predicting the salary of an adult.

III. PROPOSED METHODOLOGY

This section is focused on the prediction of the salary of an adult we are using different traditional machine learning algorithms [12]. We are first identifying the dataset, and then while pre-processing the imported dataset, we do balancing i.e. basically a method that can balance the imbalanced class [13] and find out the prominent features in the dataset using Gini Index.

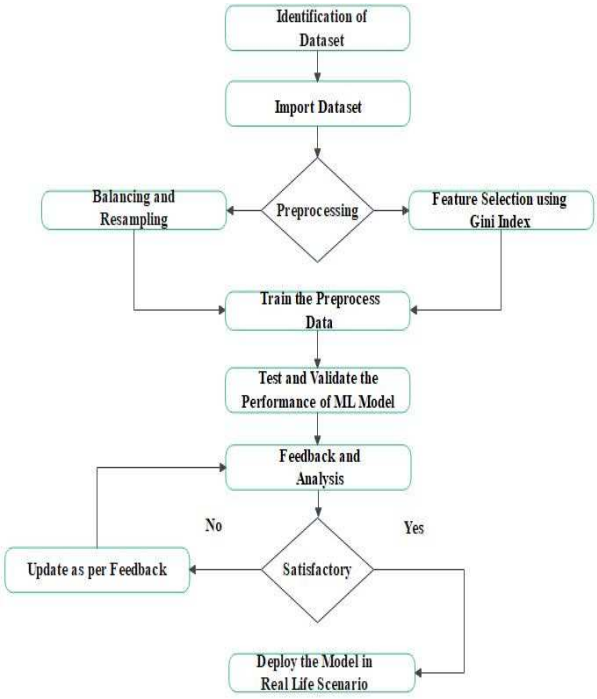

Fig. 2. Proposed Methodology

We are now testing and validating the ML Models and if we get satisfied result, we deploy it to a real-world application otherwise we give feedback [14-15] and again improve it and then after getting a satisfying result, we deploy it as shown in Fig. 2.

Different classifiers [16] are used are as follows:

1) Bayes Net: This is a Probabilistic Graphical Model used to build models and in a huge variety of tasks together with prediction, anomaly detection [17-18], reasoning and so on.

2) Logistic: It is to evaluate and predict the probability of the concerned variable.

3) IBK(Instance-Based learner): It generates a prediction for a test instance just-in-time [19] instead of building a model.

4) AdaBoostM1: It is an iterative ensemble method which ensembles different classifiers to increase the accuracy of classifiers.

5) Randomizable Filtered Classifier: An arbitrary classifier can be applied to data that has been passed through an arbitrary filter using this class.

6) Decision Stump: It is used for obtaining a decision tree with only one single split. It makes a prediction based on the value of just a single input feature.

7) J48: It is utilized to classify many applications [20] and produce accurate classification results.. It is based on a top-down strategy, a recursive divide and a conquer strategy.

8) Random Forest: It's a supervised machine learning approach that's commonly used to solve classification and regression problems. It creates decision trees [21] from various samples, using the majority vote for classification and the average for regression.

TABLE. II. PERFORMANCE ANALYSIS OF THE ORIGINAL DATASET

Classifier	Accuracy	Error	TP Rate	FP Rate	F- Measure	MCC	Precision	Recall	ROC Area
Baye Net(bayes)	81.89%	16.1087	0.83	0.189	0.845	0.604	0.858	0.839	0.917
Logistic (function)	83.23%	14.77%	0.852	0.318	0.847	0.574	0.846	0.852	0.906
IBK(lazy)	78.27%	20.73%	0.793	0.368	0.792	0.428	0.791	0.793	0.712
AdaBoostM1(meta)	82.02%	15.98%	0.84	0.404	0.826	0.52	0.833	0.84	0.87
Randomizable Filtered Classifer	73.97%	25.03%	0.75	0.433	0.75	0.316	0.75	0.75	0.658
Decision Stump(trees)	74.92%	24.08%	0.759	0.759				0.759	0.754
J48	84.11%	13.89%	0.861	0.294	0.857	0.602	0.856	0.861	0.888
Random Forest	82.71%	15.29%	0.847	0.307	0.844	0.565	0.842	0.847	0.896

IV. RESULTS AND DISCUSSIONS

This section is focused on the performance and comparison of different machine learning algorithms between the original dataset and the dataset where the prominent features are selected. Making use of the confusion matrix, we are executing our problem statement. Performance is evaluated and compared for different state of art parameters i.e. accuracy, precision, recall and F1 score. Table 2 shows Performance analysis of the original dataset and Table 3 depicts the performance analysis of the prominent features. Fig. 3 represents the comparison of accuracy measure for original dataset and of the prominent feature dataset. Fig. 4 represents the comparison of error measure for original dataset and of the prominent feature dataset. Fig. 5 represents the comparison of TP rate measure for original dataset and of the prominent feature dataset and Fig. 6 represents the comparison of FP rate measure for original dataset and of the prominent feature dataset. Following graphs represent the performance evaluation over different classifiers:

TABLE. III. PERFORMANCE ANALYSIS OF THE PROMINENT FEATURES

Classifier	Accuracy	Error	TP Rate	FP Rate	F-Measure	MCC	Precision	Recall	ROC Area
Baye net(bayes)	83.82%	15.1794	0.838	0.191	0.844	0.602	0.857	0.838	0.917
Logistic(function)	86.12%	13.83%	0.851	0.32	0.846	0.571	0.845	0.851	0.906
IBK(lazy)	82.25%	17.25%	0.802	0.378	0.799	0.441	0.796	0.802	0.752
AdaBoostM1(meta)	84.02%	15.78%	0.84	0.404	0.826	0.52	0.833	0.84	0.87
Randomizable Filtered Classifer	80.10%	19.85%	0.791	0.379	0.789	0.419	0.788	0.791	0.738
Decision Stump(trees)	76.53%	23.18%	0.759	0.759	-	-	-	0.759	0.754
J48	87.82%	11.78%	0.862	0.294	0.858	0.605	0.857	0.862	0.89
Random Forest	85.11%	14.49%	0.841	0.308	0.838	0.552	0.837	0.841	0.885

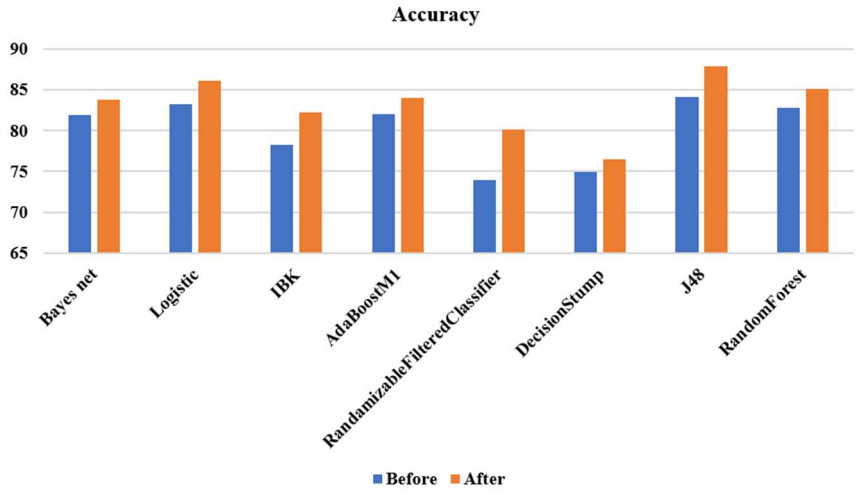

Fig. 3. The graph between the accuracy of dataset between original and prominent features

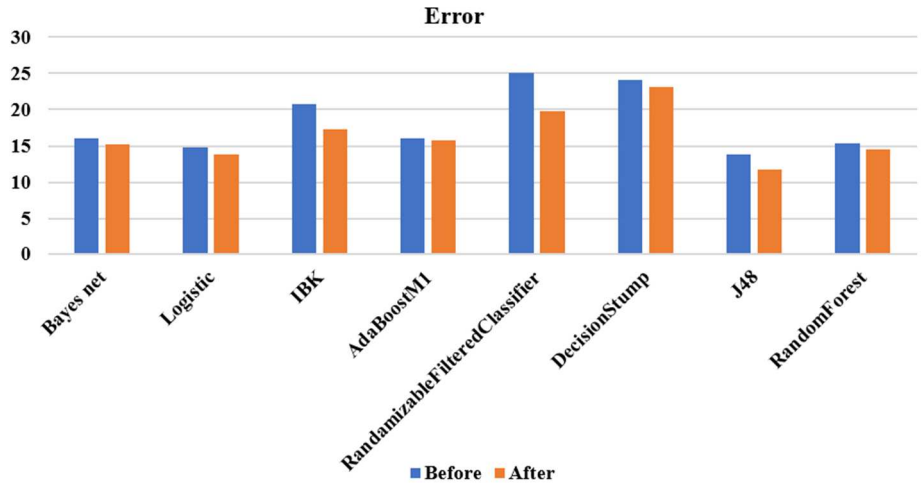

Fig. 4. The graph between the error of dataset between original and prominent feature

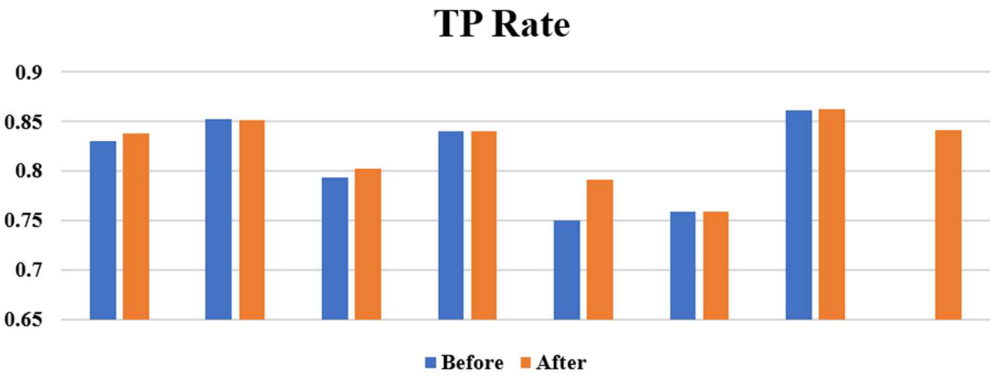

Fig. 5. The graph between the TP rate of dataset between original and prominent features

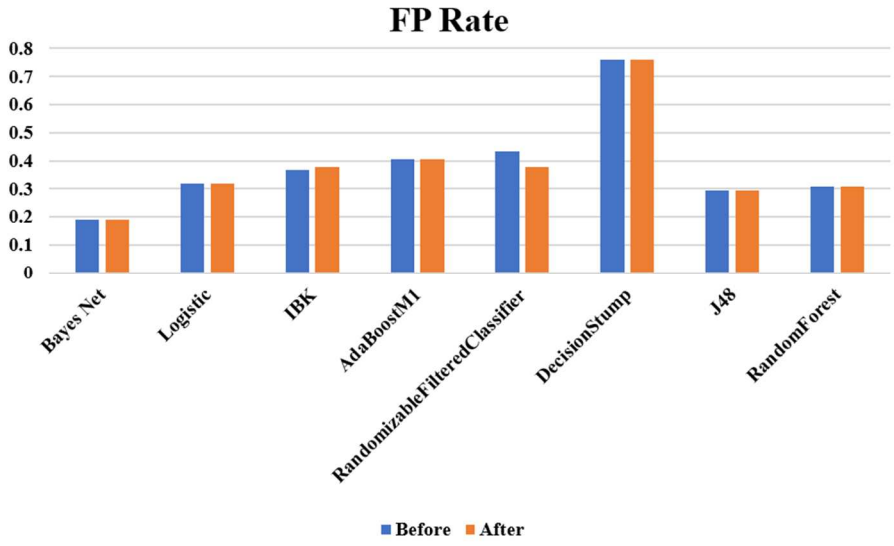

Fig. 6. The graph between the FP rate of the dataset between original and prominent features

V. Conclusion And Future Scope

The paper is focused on the prediction of the salary of an adult. It is necessary because wealth and income are a huge concern these days to bring financial stability in the individual's life and also for the betterment of the nation. To accomplish this aim, we have applied feature selection, machine learning algorithms and data mining tools to identify highly prioritized features from which we achieved more accurate performance and more precise results over the census income dataset. However, no matter how good is the performance, there is always a scope for improvement in the result. In future, one can optimize the result through the ensembling of more than two machine learning algorithms. Feature Selection can also be done by using other methods to enhance the performance of the model. Other methods for class balancing and resampling can be adopted to improve the performance of the model. Performance can be evaluated by other different parameters to check the performance of the model.

REFERENCES

[1] VidyaChockalingam,SejalShahandRonitShaw:"IncomeClassification using Adult Census Data", https://cseweb.ucsd.edu/classes/wi17/cse258- a/reports/a120.pdf.

[2] Sisay Menji Bekena:"Using decision tree classifier to predict income levels", Munich Personal RePEc Archive 30th July, 2017

[3] MohammedTopiwalla:"MachineLearningonUCIAdultdataSetUsing Various Classifier Algorithms And Scaling Up The Accuracy Using Extreme Gradient Boosting", University of SP Jain School of Global Management.

[4] Alina Lazar: "Income Prediction via Support Vector Machine", International Conference on Machine Learning and Applications - ICMLA 2004, 16-18 December 2004, Louisville, KY, USA.

[5] S.Deepajothi and Dr. S.Selvarajan: "A Comparative Study of Classifica- tion Techniques On Adult Data Set", International Journal of Engineering Research Technology (IJERT), ISSN: 2278-0181 Vol. 1 Issue 8, October- 2012.

[6] Chet Lemon, Chris Zelazo and Kesav Mulakaluri: "Predicting if income exceeds $50,000 per year based on 1994 US Census Data with Simple Classification Techniques", https://cseweb.ucsd.edu/ jmcauley/cse190/reports/sp15/048.pdf.

[7] Haojun Zhu: "Predicting Earning Potential us- ing the Adult Dataset", https://rstudio-pubs- static.s3.amazonaws.com/235617 51e06fa6c43b47d1b6daca2523b2f9e4.html

[8] IBM® SPSS® Modeler Professional, "Make Better Decisions Through Predictive Intelligence", IBM Software, Business Analytics.

[9] Jerzy Stefanowski, "Data Mining - Clustering", Institute of Comput ing Sciences Poznan University of Technology, Poznan, Poland, Lecture 7, SE Master Course, 2008/2009.

[10] J. Ross Quinlan, "C4.5: programs for machine learning", Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, 1993.

[11] Chid Apte, "Data Mining: Concepts and Techniques", Second Editionx, University of Illinois at Urbana- Champaign..

[12] Pawar, L., Agrawal, P., Kaur, G., & Bajaj, R. (2021). Elevate Primary Tumor Detection Using Machine Learning. Cognitive Behavior and Human Computer Interaction Based on Machine Learning Algorithm, 301-313.

[13] Kumar, D., Sharma, A. K., Bajaj, R., & Pawar, L. (2021). Feature Optimized Machine Learning Framework for Unbalanced Bioassays. Cognitive Behavior and Human Computer Interaction Based on Machine Learning Algorithm, 167-178.

[14]] Rahi, P., Sood, S. P., Bajaj, R., & Kumar, Y. (2021). Air quality monitoring for Smart eHealth system using firefly optimization and support vector machine. International Journal of Information Technology, 13(5), 1847-1859.

[15] Pawar, L., Sharma, A. K., Kumar, D., & Bajaj, R. (2020). Advanced Ensemble Machine Learning Model for Balanced BioAssays. In Artificial Intelligence and Machine Learning in 2D/3D Medical Image Processing (pp. 171-178). CRC Press.

[16] Sharma, D., Singh Aujla, G., & Bajaj, R. (2021). Deep neuro-fuzzy approach for risk and severity prediction using recommendation systems in connected health care. Transactions on Emerging Telecommunications Technologies, 32(7).

[17] Sharma, D., Singh Aujla, G., & Bajaj, R. (2019). Evolution from ancient medication to human-centered Healthcare 4.0: A review on health care recommender systems. International Journal of Communication Systems.

[18] Singh, Jaspreet, Rohit Bajaj, and Anuj Kumar. "Scaling Down Power Utilization with Optimal Virtual Machine Placement Scheme for Cloud Data Center Resources: A Performance Evaluation." In 2021 2nd Global Conference for Advancement in Technology (GCAT), pp. 1-6. IEEE, 2021.

[19] Singh, Jaspreet. "Genetic Approach based Optimized Load Balancing in Cloud Computing: A Performance Perspective." In 2022 9th International Conference on Computing for Sustainable Global Development (INDIACom), pp. 814-819. IEEE, 2022

[20] Bathla, Gaurav, and Rajneesh Randhawa. "ODA: optimal deployment algorithm for wireless sensor network for coverage enhancement." In Proceedings of 3rd international conference on internet of things and connected technologies (ICIoTCT), pp. 26-27. 2018.

[21] Bathla, Gaurav, and Rajneesh Randhawa. "Virtual Tier structured Grid-based Dynamic Route Adjustment scheme for mobile sink-based Wireless Sensor Networks (VTGDRA)." International Journal of Applied Engineering Research 13, no. 7 (2018): 4702-4707.