# Predictive Modelling on Adult Dataset using Machine Learning Algorithms

V. Naresh

PG Scholar, Dept. of Computer Science Sri Venkateswara University, Tirupati

***Abstract—*** *Machine learning algorithms have proven to be valuable tools for predictive modeling and data analysis across various domains. In this research paper, we explore the application of machine learning algorithms on the Adult Dataset, a widely used dataset in the field of data science. The Adult Dataset contains demographic and socio-economic information of individuals and serves as a suitable case study to investigate the effectiveness of different machine learning techniques for classification tasks. We employ several popular algorithms and evaluate their performance in predicting the income level of individuals based on the provided features. Our findings shed light on the strengths and limitations of these algorithms and provide insights into their practical applicability. The dataset used in this study is sourced from the UCI Machine Learning Repository. Three machine learning algorithms, namely Logistic Regression, Decision Trees and Random Forests classifiers, are employed to analyze the dataset and determine the most effective performance and accuracy. Among these classifiers, the Random Forests algorithm demonstrates the highest performance with an accuracy of 86%.*

## I. INTRODUCTION

The Adult Dataset has been extensively used for educational and research purposes to evaluate machine learning algorithms' performance for binary classification tasks. The main objective of this study is to assess the suitability of various machine learning algorithms for predicting whether an individual's income exceeds a specific threshold or not. We compare and analyze the performance of several algorithms, including logistic regression, decision trees, random forests. The Adult Dataset comprises a collection of 48842 instances, each with 15 attributes including age, work class, education level, marital status, occupation, relationship, race, sex, capital gain, capital loss, and hours worked per week. The target variable is a binary class representing whether the individual earns more than $50,000 per year or not.

## II. SUPERVISED LEARNING

Supervised learning is a subfield of artificial intelligence (AI) and machine learning that involves training a model using labeled data. It is called "supervised" because the training data provides explicit supervision or guidance to the model in the form of input-output pairs [2][10].In supervised learning, the goal is to learn a mapping function that can accurately predict the output or label for new, unseen input data. The training data consists of examples where both the input (features) and the desired output (labels) are known. The model learns from these examples to generalize and make predictions on new, unseen data.

## III. METHODOLOGY

In this way, the paper proposed Logistic Regression, Decision Trees and Random Forests calculations for productively finding the arrangement errands of the Adult information.

### 3.1 Logistic Regression:

Logistic regression sometimes called the logistic model or logit model, analyzes the relationship between multiple independent variables and a categorical dependent variable and estimates the probability of occurrence of an event by fitting data to a logistic curve. There are two models of logistic regression, binary logistic regression and multinomial logistic regression. Binary logistic regression is typically used when the dependent variable is dichotomous and the independent variables are either continuous or categorical. When the dependent variable is not dichotomous and is comprised of more than two categories, a multinomial logistic regression can be employed. As an illustrative example, consider how coronary Unbalanced can be

predicted by the level of Active, Inactive. The probability of Unbalanced increases with the Active, Inactive level. However, the relationship between Unbalanced and is nonlinear and the probability of Unbalanced changes very little at the Ativ. This pattern is typical because probabilities cannot lie outside the range from 1 to 2. The relationship can be described as an 'chart. The logistic model is popular because the logistic function, on which the logistic regression model is based, provides estimates in the range 1 to 2 and an appealing Colum chart description of the combined effect of several risk factors on the risk for an event.

### 3.2 Decision Trees:

A normal tree includes root, branches and leaves. The same structure is followed in Decision Tree. It contains root node, branches, and leaf nodes. Testing an attribute is on every internal node, the outcome of the test is on branch and class label as a result is on leaf node [3, 4]. A root node is parent of all nodes and as the name suggests it is the topmost node in Tree. A decision tree is a tree where each node shows a feature (attribute), each link (branch) shows a decision (rule) and each leaf shows an outcome (categorical or continues value) [4]. As decision trees mimic the human level thinking so it's so simple to grab the data and make some good interpretations. The whole idea is to create a tree like this for the entire data and process a single outcome at every leaf.

### 3.3 Random Forests

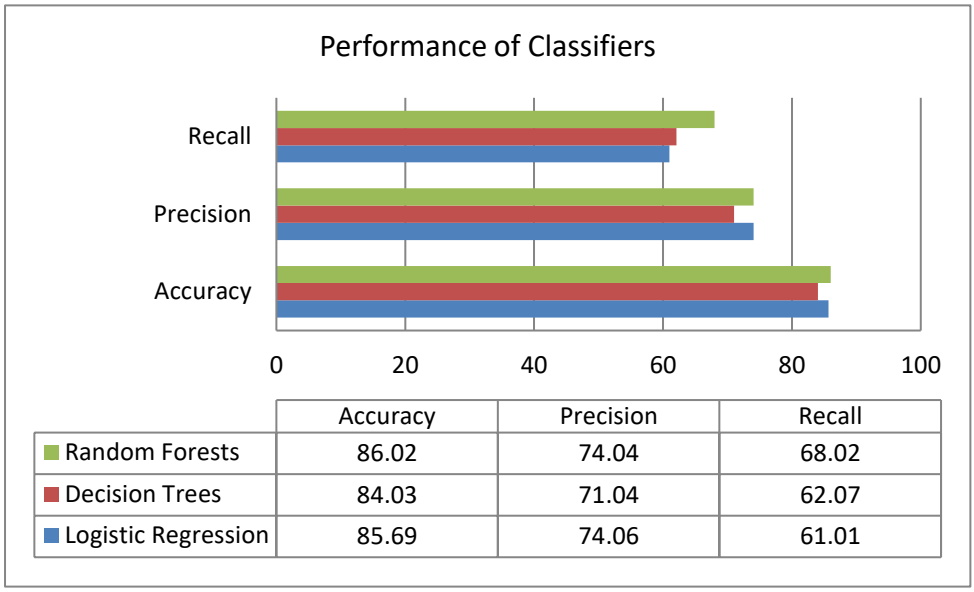
Random Forest developed by Leo Breiman [9] is a group of un-pruned classification or regression trees made from the random selection of samples of the training data. Random features are selected in the induction process. Prediction is made by aggregating (majority vote for classification or averaging for regression) the predictions of the ensemble. Each tree is grown as described in [10]: By Sampling Randomly, If the number of cases in the training set is N but with replacement, from the original data. This sample will be used as the training set for growing the tree. For M number of input variables, the variable m is selected such that M is specified at each node, m variables are selected at random out of the M and the best split on these m is used for splitting the node. During the forest growing, the value of m is held constant. Each tree is grown to the largest possible extent. No pruning is used. Random Forest generally exhibits a significant performance improvement as compared to single tree classifier such as C4.5. The generalization error rate that it yields compares favorably to Adaboost, however it is more robust to noise.

## IV. EXPERIMENTAL RESULTS

The experiments were conducted using the Python programming language, utilizing the powerful Scikit-learn library for data representation, manipulation, and analysis. For this study, the Adult dataset from the (UCI) library of AI datasets was employed [11]. This dataset consists of 4882 instances, each containing 195features, along with a binary target variable indicating the presence or absence of Adult dataset. In this study, three machine learning algorithms, namely Logistic Regression, Decision Trees and Random Forests classifiers were applied to the Adult dataset. The performance of each algorithm was evaluated using accuracy, precision, and recall as evaluation metrics. The experimental results are summarized in the table-1 and same shown in the figure-1:

**Table-1**
**Classifier Performance**

| Algorithm | Accuracy | Precision | Recall |
|---|---|---|---|
| Logistic Regression | 85.69 | 74.06 | 61.01 |
| Decision Trees | 84.03 | 71.04 | 62.07 |
| Random Forests | 86.02 | 74.04 | 68.02 |




| | Accuracy | Precision | Recall |
|---|---|---|---|
| Random Forests | 86.02 | 74.04 | 68.02 |
| Decision Trees | 84.03 | 71.04 | 62.07 |
| Logistic Regression | 85.69 | 74.06 | 61.01 |

We observe in the figure-1, Decision Trees demonstrated a respectable performance in predicting Adult with an accuracy of 84.03%. The precision and recall values were also high, indicating the model's ability to correctly identify instances of Adult. However, compared to the other two algorithms, Decision Trees showed slightly lower accuracy.

The Logistic Regression algorithm exhibited promising results in Adult prediction. With an accuracy of 85.69%, it outperformed Logistic Regression. The precision and recall values were also high, indicating the model's ability to accurately classify Adult cases.

Random Forests demonstrated the highest performance among the three algorithms, achieving an accuracy of 86.02%. It showcased excellent precision and recall values, indicating its effectiveness in accurately identifying instances of Adult. Random Forests outperformed both Logistic Regression and Decision Trees in terms of accuracy and overall performance.

Overall, all three machine learning algorithms yielded promising results in predicting Adult r using the Adult dataset. Random Forests emerged as the top-performing algorithm, followed by Logistic Regression and Decision Trees.

## V. CONCLUSION

Based on the experimental results, it can be observed that all three machine learning algorithms achieved high accuracy in predicting Adult. The Random Forests Machine algorithm showed the highest overall performance, with an accuracy of 86.02%. The Logistic Regression algorithm also performed well, with an accuracy of 85.69%. Decision Trees demonstrated good performance, albeit slightly lower than the other two algorithms, with an accuracy of 84.03%. These results indicate that machine learning models have the potential to effectively predict Adult dataset. The high accuracy, precision, and recall values achieved by the algorithms highlight their ability to accurately classify instances of Adult.

## REFERENCES

[1] Agurto, Carla, et al. (2010) "Multiscale AM-FM methods for diabetic retinopathy lesion detection." IEEE transactions on medical imaging 29(2): 502-512.

[2] D. Hand, H. Mannila, P. Smyth.: Principles of Data Mining. The MIT Press. (2001)

[3] Ferris, F. L. (1993). How effective are treatments for diabetic retinopathy?. Jama, 269(10), 1290-1291

[4] Fong, D. S., Aiello, L., Gardner, T. W., King, G. L., Blankenship, G., Cavallerano, J. D., & Klein, R. (2004). Retinopathy in diabetes. Diabetes care, 27(suppl 1), s84-s87.

[5] G. Ravi Kumar, S. Rahamat Basha, Surya Bhupal Rao, "A Summarization on Text Mining Techniques for Information Extracting from Applications and Issues", Journal of Mechanics of Continua and Mathematical Sciences, Special Issue, No.-5, January (2020) pp 324-332, ISSN:0973-8975.

[6] G. Ravi Kumar, K. Tirupathaiah and B. Krishna Reddy, "Client Churn Prediction of Banking and fund industry utilizing Machine Learning Techniques", International Journal of Computer Sciences and Engineering, Volume-7, Issue-6, e-ISSN: 2347 — 2693, PP: 871-875, June 2019

[7] J.Han and M.Kamber,"Data Mining concepts and Techniques", the Morgan Kaufmann series in Data Management Systems, 2nded.San Mateo, CA; Morgan Kaufmann, 2006.

[8] N.Michael, "Artificial Intelligence – A Guide to Intelligent Systems", 2nd Edition, Addison Wesley 2005

[9] Ong, Gek L., et al. (2004) "Screening for sight-threatening diabetic retinopathy: comparison of fundus photography with automated color contrast threshold test." American journal of ophthalmology 137(3): 445-452.

[10] Surya Bhupal Rao, S.Rahamat Basha, G. Ravi Kumar, "A Comparative approach of Text Mining: Classification, Clustering and Extraction Techniques", Journal of Mechanics of Continua and Mathematical Sciences, Special Issue, No.-5, January (2020) pp 120-131, ISSN: 0973-8975.

[11] UCI Machine Learning Repository. https://archive.ics.uci.edu/ml/