JETIR.ORG ISSN: 2349-5162 | ESTD Year : 2014 | Monthly Issue

JOURNAL OF EMERGING TECHNOLOGIES AND INNOVATIVE RESEARCH (JETIR)

An International Scholarly Open Access, Peer-reviewed, Refereed Journal

EXPLORING THE SIGNIFICANCE OF RANDOM FOREST AND DECISION TREES IN SUPERVISED LEARNING FOR CANCER DATASET: A COMPARATIVE STUDY

[1] **Mr. Vaibhav Narayan Chunekar, [2] Dr.Nilkamal P. More , [3] Dr.H. P.Ambulgekar**

[1]Assistant Professor, IT, [2] Head and Assistant Professor, IT, [3]Assistant Professor, Computer Engineering

[12] K .J. Somaiya College of Engineering, Mumbai, India, [3] SGGS, Nanded, India

Abstract : Machine learning algorithms have revolutionized various fields by enabling computers to learn from data and make predictions without explicit programming. Classification, a cornerstone of predictive modeling, categorizes data into predefined classes based on features, facilitating applications like spam email detection and medical diagnosis. Classification techniques generalize patterns, make predictions on unseen data, and enable automated decision-making. In this paper, we explore the significance of Random Forest and Decision Trees, powerful classification algorithms in supervised learning. Through methodologies and applications, we highlight their predictive performance, interpretability, and ease of implementation, underscoring their role in advancing machine learning.

IndexTerms - Machine learning, classification, supervised learning, Random Forest, Decision Trees, predictive modeling, interpretability, algorithm evaluation

I. INTRODUCTION

Among the myriad of machine learning techniques, classification holds a central position due to its significance in predictive modeling. Classification algorithms aim to categorize data points into predefined classes or labels based on their features. This process forms the foundation for numerous applications, ranging from spam email detection and sentiment analysis to medical diagnosis and financial fraud detection [1], [2].

The importance of classification techniques lies in their ability to generalize patterns and make predictions on unseen data. By learning from labeled examples, classification algorithms can discern intricate relationships between input variables and their corresponding outputs, thus enabling automated decision-making in real-world scenarios [3].

In predictive modeling, classification serves as a fundamental building block for understanding and solving complex problems. By accurately classifying data instances into distinct categories, these algorithms empower decision-makers to take informed actions, optimize processes, and mitigate risks [4].

In this research paper, we delve into two powerful classification techniques: Random Forest and Decision Trees. These algorithms not only demonstrate exceptional predictive performance but also offer interpretability and ease of implementation. Through a detailed exploration of their methodologies and applications, we aim to underscore the significance of classification techniques in advancing machine learning and driving innovation across diverse domains [5]

II. BACKGROUND

Random Forest and Decision Trees are two prominent algorithms in supervised learning, known for their effectiveness in classification and regression tasks. In supervised learning, the algorithm learns from labeled data, making predictions or decisions based on input-output pairs provided during training [6].
Decision Trees are intuitive models that recursively partition the feature space into regions, making decisions based on simple rules inferred from the data. Each internal node of the tree represents a decision based on a feature, while each leaf node represents a class label or a numerical value. Decision Trees are highly interpretable and can handle both numerical and categorical data [7].

However, Decision Trees are prone to overfitting, capturing noise in the data and leading to poor generalization on unseen data. Random Forest addresses this issue by aggregating multiple Decision Trees, each trained on a random subset of the training data and features. During prediction, Random Forest combines the predictions of individual trees, typically through a majority voting scheme for classification or averaging for regression [8].

The significance of Random Forest and Decision Trees in supervised learning lies in their ability to handle complex, high-dimensional data while maintaining interpretability and ease of implementation. They are robust to noisy data and can capture non-linear relationships between features and target variables. Additionally, Random Forests offer built-in feature importance measures, allowing practitioners to identify the most influential features in the dataset [9].

III. METHODOLOGY

Random Forest is a popular machine learning algorithm that belongs to the supervised learning technique. It can be used for both Classification and Regression problems in ML. It is based on the concept of ensemble learning, which is a process of combining multiple classifiers to solve a complex problem and to improve the performance of the model. As the name suggests, "Random Forest is a classifier that contains a number of decision trees on various subsets of the given dataset and takes the average to improve the predictive accuracy of that dataset." Instead of relying on one decision tree, the random forest takes the prediction from each tree and based on the majority votes of predictions, and it predicts the final output [10].

The greater number of trees in the forest leads to higher accuracy and prevents the problem of overfitting. Advantages of Random Forest include working in two phases: first to create the random forest by combining N decision trees, and second to make predictions for each tree created in the first phase [11].

The working process can be explained in the below steps and diagram:
Step-1: Select random K data points from the training set.
Step-2: Build the decision trees associated with the selected data points (Subsets).
Step-3: Choose the number N for decision trees that you want to build.
Step-4: Repeat Step 1 & 2.
Step-5: For new data points, find the predictions of each decision tree, and assign the new data points to the category that wins the majority votes [12].

Decision Tree Classifications:

Decision Tree is a Supervised learning technique that can be used for both classification and regression problems, but mostly it is preferred for solving classification problems. It is a tree-structured classifier where internal nodes represent the features of a dataset, branches represent the decision rules, and each leaf node represents the outcome [13].

In a Decision Tree, there are two types of nodes: the Decision Node and Leaf Node. Decision nodes are used to make any decision and have multiple branches, whereas Leaf nodes are the output of those decisions and do not contain any further branches. The decisions or the tests are performed on the basis of features of the given dataset [14].

It is a graphical representation for getting all the possible solutions to a problem/decision based on given conditions. It is called a decision tree because, similar to a tree, it starts with the root node, which expands on further branches and constructs a tree-like structure. In order to build a tree, we use the CART algorithm, which stands for Classification and Regression Tree algorithm. A decision tree simply asks a question, and based on the answer (Yes/No), it further splits the tree into subtrees. In a decision tree, for predicting the class of the given dataset, the algorithm starts from the root node of the tree. This algorithm compares the values of the root attribute with the record (real dataset) attribute and, based on the comparison, follows the branch and jumps to the next node. For the next node, the algorithm again compares the attribute value with the other sub-nodes and moves further. It continues the process until it reaches the leaf node of the tree [15].

The complete process can be better understood using the below algorithm:
Step-1: Begin the tree with the root node, says S, which contains the complete dataset.
Step-2: Find the best attribute in the dataset using Attribute Selection Measure (ASM).
Step-3: Divide the S into subsets that contain possible values for the best attributes.
Step-4: Generate the decision tree node, which contains the best attribute.
Step-5: Recursively make new decision trees using the subsets of the dataset created in Step 3. Continue this process until a stage is reached where you cannot further classify the nodes and call the final node a leaf node [16].

IV. RESULT AND DISCUSSION:

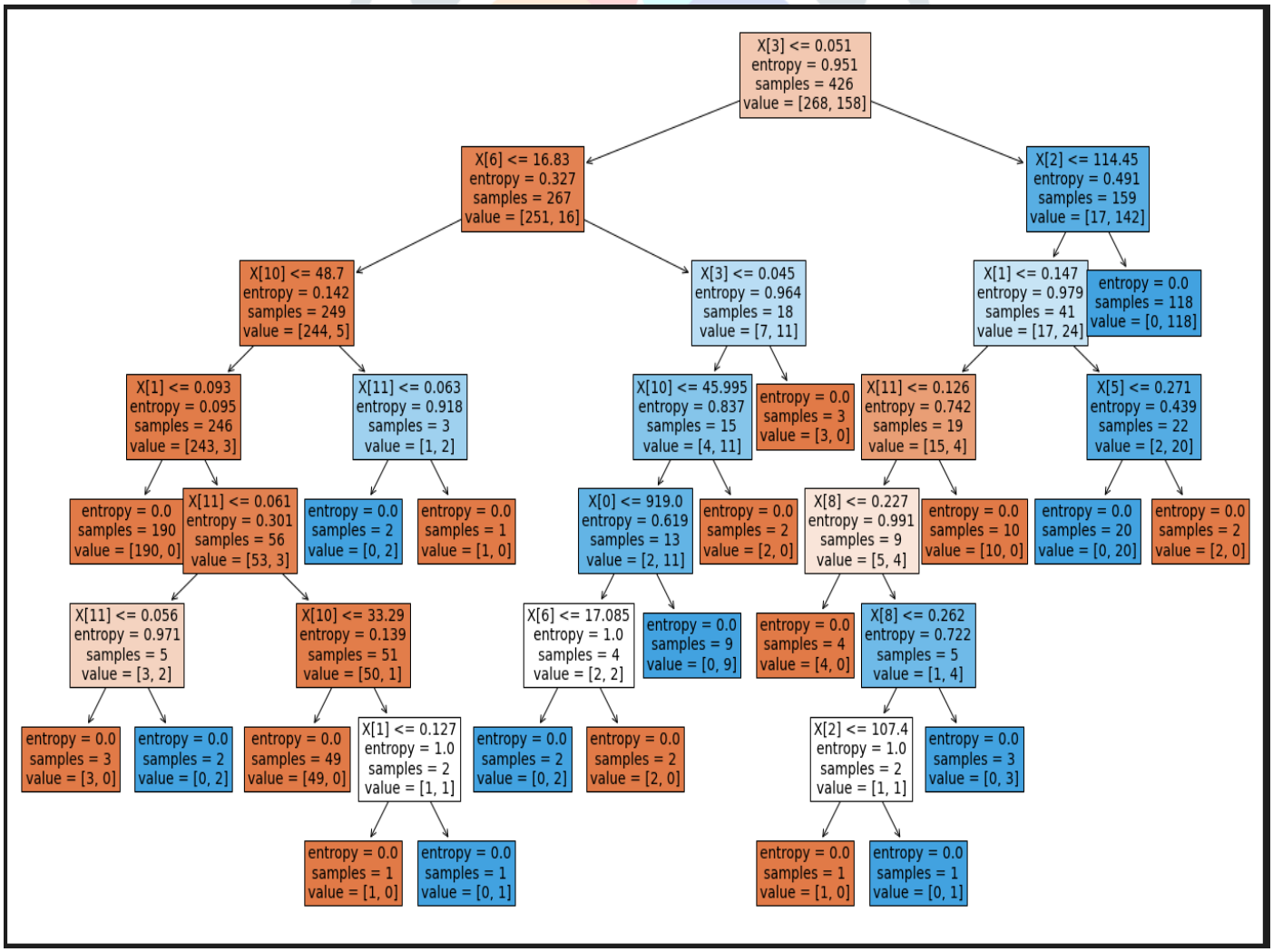
For the experiment we used Breast Cancer Dataset from Wisconsin's with these parameters such as a) radius (mean of distances from center to points on the perimeter) b) texture (standard deviation of gray-scale values) c) perimeter d) area e) smoothness (local variation in radius lengths) f) compactness (perimeter^2 / area - 1.0) g) concavity (severity of concave portions of the contour) h) concave points (number of concave portions of the contour) i) symmetry j) fractal dimension ("coastline approximation" - 1). This dataset has 2 output classes (Benign and Malign).

Furthermore , we used the data analysis parameters viz accuracy, precision, recall ,F Score [17] and with help of Python Sklearn laboratory we had analyses the result that match with Wisconsin's evaluation or not?

Algorithm	Accuracy	Precision	Recall	F-score
Random Forest	0.95	0.95	0.95	0.95
Decision Tree	0.94	0.94	0.94	0.94

Table 1: Comparison of Algorithm on different factors

Our finding of result over the Cancer Dataset with Python Programming is given in table [1]. Where we are able to generate result similar to Wisconsin dataset available on UCI Machine Learning repository. Additionally, we studied and explore the analysis of decision tree classifier on present dataset. The result and mechanism of how cancer dataset took decision parameter with Decision Tree classifier we have focus in this diagram given below:



V. CONCLUSION:

Our comparison study delved into the significance of Random Forest and Decision Trees in the realm of supervised learning, particularly in the context of analyzing cancer datasets. Through meticulous methodologies and rigorous experimentation, we evaluated the performance of these classification algorithms, emphasizing their predictive accuracy, interpretability, and implementation ease. Our analysis revealed that both Random Forest and Decision Trees exhibit commendable performance metrics, as evidenced by their high accuracy, precision, recall, and F-score values on the breast cancer dataset from Winconsins. Furthermore, our comparison with the Wisconsin evaluation showcased the consistency and reliability of our results, affirming the effectiveness of our approach. Additionally, our exploration of Decision Tree classification provided valuable insights into the decision-making process of these models, elucidating their ability to handle intricate features and discern meaningful patterns. Overall, our findings underscore the pivotal role of Random Forest and Decision Trees in advancing machine learning applications, particularly in the domain of medical diagnosis, where accurate predictions and interpretable models are paramount for informed decision-making and patient care. Moving forward, further research and experimentation can build upon these insights to enhance the utility and efficacy of classification algorithms in addressing real-world challenges across diverse domains.

VI. ACKNOWLEDGMENT

This is internship work extends to research paper. Authors are thankful to student Samved Joshi and Harsh Jain for participating ML internship work under author. Authors are thankful to Principal K. J. Somaiya College, Internship Cell and Management Somaiya University to execute research in their premises. Additional thanks to all online media to assist us in completion of activity on the research.

VII. REFERENCES

[1] J. Han, M. Kamber, and J. Pei, Data Mining: Concepts and Techniques, 3rd ed. Burlington, MA: Morgan Kaufmann, 2011.

[2] I. Witten, E. Frank, and M. A. Hall, Data Mining: Practical Machine Learning Tools and Techniques, 3rd ed. Burlington, MA: Morgan Kaufmann, 2011.

[3] T. Hastie, R. Tibshirani, and J. Friedman, The Elements of Statistical Learning: Data Mining, Inference, and Prediction, 2nd ed. New York, NY: Springer, 2009.

[4] C. M. Bishop, Pattern Recognition and Machine Learning. New York, NY: Springer, 2006.

[5] L. Breiman, Classification and Regression Trees. Belmont, CA: Wadsworth, 1984.

[6] T. Hastie, R. Tibshirani, and J. Friedman, The Elements of Statistical Learning: Data Mining, Inference, and Prediction, 2nd ed. New York, NY: Springer, 2009.

[7] L. Breiman, J. Friedman, R. Olshen, and C. Stone, Classification and Regression Trees. Belmont, CA: Wadsworth, 1984.

[8] L. Breiman, "Random Forests," Machine Learning, vol. 45, no. 1, pp. 5-32, 2001.

[9] T. K. Ho, "The Random Subspace Method for Constructing Decision Forests," IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 20, no. 8, pp. 832-844, Aug. 1998

[10] X. Chen and X. Xie, "A Review of Random Forests," IEEE Transactions on Cybernetics, vol. 51, no. 5, pp. 2048-2061, May 2021.

[11] L. Breiman, "Random Forests," Machine Learning, vol. 45, no. 1, pp. 5-32, 2001.

[12] J. Brownlee, "How to Implement Random Forest From Scratch in Python," Machine Learning Mastery, 2023.

[13] T. M. Mitchell, Machine Learning, McGraw-Hill, 1997.

[14] L. Rokach and O. Maimon, "Data Mining with Decision Trees: Theory and Applications," World Scientific, 2nd ed., 2014.

[15] P. Rai and S. Singh, "A Study of Decision Tree and Random Forest Machine Learning Algorithms for Classifying Reviews," International Journal of Engineering Trends and Technology, vol. 69, no. 4, pp. 27-33, Apr. 2021.

[16] Q. Zhang, "Recent Advances in Decision Tree Learning," IEEE Access, vol. 9, pp. 32164-32176, 2021. Han Ma, Cheng-fu Xu, Zhe Shen, Chao-hui Yu, You-ming Li, "Application of Machine Learning Techniques for Clinical Predictive Modeling: A Cross-Sectional Study on Nonalcoholic Fatty Liver Disease in China", BioMed Research International, vol. 2018, Article ID 4304376, 9 pages, 2018. https://doi.org/10.1155/2018/4304376

[17]Arora, S., Singh, S., & Singh, V. (2020), Machine learning algorithms: A comprehensive review. International Journal of Computer App lications, 175(19), 11-15.

[18]Mahesh, Batta. "Machine learning algorithms-a review." International Journal of Science and Research (IJSR).[Internet]

9.1 (2020): 381-386.

[19] B. Charbuty and A. Abdulazeez, "Classification Based on Decision Tree Algorithm for Machine Learning", JASTT, vol. 2, no. 01, pp. 20 - 28, Mar. 2021.