International Journal of Engineering and Advanced Technology (IJEAT)
ISSN: 2249-8958 (Online), Volume-12 Issue-5, June 2023

Applying Decision Tree Algorithm Classification and Regression Tree (CART) Algorithm to Gini Techniques Binary Splits

Nirmla Sharma, Sameera Iqbal Muhmmad Iqbal

Abstract*: Decision tree analysis is a predictive modelling tool used in various applications. It is constructed through an algorithmic technique that divides the dataset into different methods created under varied conditions. Decision trees are the most dominant algorithms that fall under the set of supervised algorithms. However, the Decision Trees' appearance is modest and natural; there is nothing modest about how the algorithm drives the procedure by determining splits and how tree pruning happens. The initial object to appreciate in Decision Trees is that it splits the analyst field, i.e., the objective parameter, into diverse subsets which are comparatively more similar from the viewpoint of the objective parameter. The Gini index is a level task that has been applied to assess the binary changes in the dataset, working with the definite object variable "Success" or "Failure". Split creation essentially covers the dataset values. Decision trees employ a top-down, greedy method that has been recognised as recursive binary splitting. It provides statistics for 15 key facts about scholar statistics, including pass or fail rates on an online Machine Learning exam. Decision trees are a type of supervised machine learning. It has been commonly applied, with an informal implementation, and has been interpreted as deriving quantitative, qualitative, non-stop, and binary splits, providing consistent outcomes. The CART tree applies a regression technique to expected standards of non-stop variables. CART regression trees are a formal technique for understanding outcomes.

***Keywords*: Decision Trees, Gini index, Objective Parameter and Statistics.**

I. INTRODUCTION

Decision Trees are supervised machine learning algorithms that are particularly well-suited for classification and regression problems. These algorithms have been created by executing the actual splitting situations at individual nodes, breaking down the drill statistics into subsets of yield parameters of the identical class. It has run for composed classification and regression tasks [1].

Manuscript received on 26 May 2023 | Revised Manuscript received on 04 June 2023 | Manuscript Accepted on 15 June 2023 | Manuscript published on 30 June 2023.
*Correspondence Author(s)
Dr. Nirmla Sharma*, Asst. Professor, Department of Computer Science, King Khalid University, Abha, Kingdom of Saudi Arabia. E-mail: nprasad@kku.edu.sa, ORCID ID: 0009-0007-0746-1001
Sameera Iqbal Muhmmad Iqbal, Department of Computer Science, King Khalid University, Abha, Kingdom of Saudi Arabia. Email: eqbal@kku.edu.sa, ORCID ID: 0009-0005-7812-4593

The dual key items of a tree are decision nodes, where the data is allocated, and leaves, where it develops an outcome [2]. The design of a binary tree for supposing whether an employee is employed or Not Employed using various statistics like time, work behaviours and movement behaviours [3], has shown under figure 1.

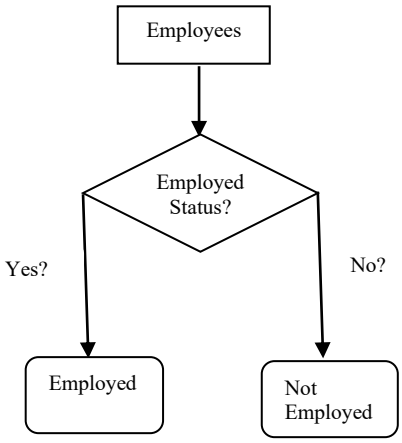

Fig. 1. Decision Tree of Employee [3]

In the above decision tree, the appeal has decision nodes, and the last outcomes are leaves. It has needed the following two categories of decision trees [4].

- Classification decision trees –the decision variable is definite. The above decision tree is an order of classification decision tree.
- Regression decision trees –the decision variable is nonstop [5].

A. Applying Decision Tree Algorithm

Gini Index

The higher the value of the Gini index, the greater the similarity. A perfect Gini index value is 0, and the poorest is 0.5 (for two classes, which is considered a challenging scenario). Gini index for a division has been designed with the assistance of the following phases −

- First, have analyzed Gini index sub-nodes, have got through the formula p^2 + q^2, which has the sum of the squares of the probabilities for success and failure [6].
- Next, analyse the Gini index for the shared have spent biased Gini score of each node that has been divided.

Classification and Regression Tree (CART) algorithm relates the Gini technique to create binary splits [7].

B. Split Design

It has generated an issue in the dataset with the help of the following three measures −

- Measure 1: Determining Gini Score

It has required just as much deliberation as this evaluation in the previous section (Gini Index).

- Measure 2: Splitting a dataset.

It has been distinct as splitting a dataset into two lists of rows, requiring an index of an attribute and dividing the value of that attribute. After receiving the two clusters - right and left - from the dataset, it analysed the value divided by the Gini score considered in the first measure. A divided value has been chosen, indicating which cluster the attribute exists in.

- Measure 3: Estimating all splits.

Later, measure the next outcome, the Gini score, and splitting the dataset has been used to estimate all splits. For this drive, first, it has created a pattern for each value related to an individual attribute as an applicant split. It is then desired to test the top feasible split by estimating the value of the split. The upper split has been applied like a point in the Decision tree [8].

C. Developing a Tree

In this tree, there is a root node and terminal nodes. After generating the root node, [9] it has constructed the tree by following two processes –

- Measure 1: Terminal node creation

When producing terminal nodes of a decision tree, one vital fact is chosen: when to end the growth of the tree or generate more terminal nodes. It has ended by applying two standards, namely maximum tree depth and minimum node accounts, as follows: −

(1) Maximum Tree Depth

This is done by finding the maximum number of nodes in a tree's next root node. It is done to end the count of terminal nodes after a tree has been extended to its maximum depth, i.e., when a tree has grown to its maximum number of terminal nodes.

(2) Minimum Node Records

It has been distinct, like the minimum number of preparation arrays that an assumed node is responsible for. It must end at terminal nodes when the tree is extended to these minimum node accounts or below this minimum node account. The terminal node has been applied to create the last prediction [10].

- Measure 2: Recursive Splitting

Equally, it assumed approximately when to generate terminal nodes; today, it has started constructing this tree. Recursive splitting is a technique to build the tree. In this technique, after a node is produced, it generates child nodes (nodes added to an existing node) recursively on an individual cluster of data, created by splitting the dataset. It repeats this process to achieve the same purpose. Figure 2 below shows the splitting decision tree algorithm [11].

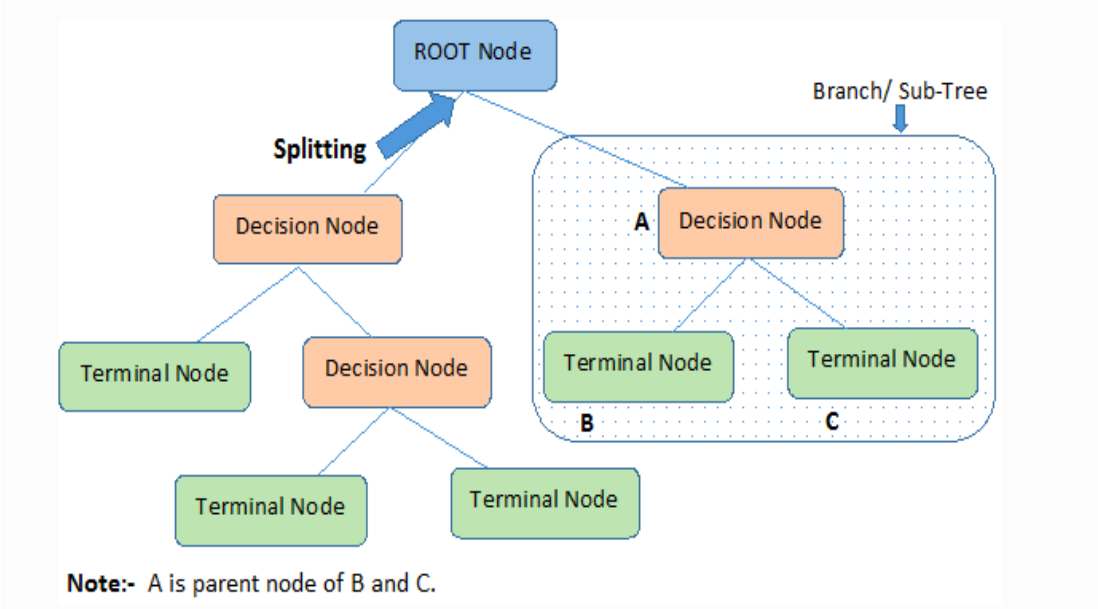

Fig. 2. Splitting Decision Tree Algorithm [11]

II. P PROBLEM STATEMENT

A numeric variable has been examined over several periods in the data with different cut-offs or thresholds. Also, final classifications have been reiterated. The essential key from a data science viewpoint has several subsequent difficulties. How does the flow of facts through the Decision Tree? This classification procedure begins with the parent node of the decision tree and develops by relating nearly splitting situations at individual non-leaf nodes, dividing the datasets into similar subsets.

III. RELATED STUDIES

A regression tree is a classification template constructed by relating logistic regression and decision trees. Logistic regression tree is a decision tree with a regression analysis construction.

In this tree structure, logistic regression assessment is completed for individual hierarchy divisions; formerly, divisions were divided uncontrollably by the C4.5 decision tree. The

Published By:
Blue Eyes Intelligence Engineering
and Sciences Publication (BEIESP)

last phase is the cut-off phase of the tree [6, 12].

This research is an example of related research efforts that will help us better identify our research. We will provide a system of similarities with an extra way related to ours to improve the identification of our research paper [7, 13].

Additional work that we measured was one called "determining the capability of the manufacturer to adopt. Finished this work, the perfect controls the masses of the invisible neurons to enhance the yield [11, 13].

A decision tree is likely to categorise statistics using a decision tree applied to the statistics. The nodes, leaves, and divisions of a tree are referred to as its functional mechanisms. Interior nodes are the requests that concern an explicit feature of the Biomed Research International problem, referred to as "root" or "primary" nodes. There is a node for individual reaction to the desires. Individually, a node has a division that points to a list of likely values for the feature. Unique to the difficulty class issues is characterized through the nodes at the end of the diagram, known as child nodes [14]. Machine learning is distinct in identifying designs using well-educated statistics when understanding unnamed input [1]. Machine learning is divided into supervised and unsupervised learning [2, 13]. Supervised learning weights at decision or forecasting models in a dataset, and the algorithms are respected, for example, either classification or regression [6]. Unsupervised learning focuses on grouping objects in a dataset, removed of known association or models [9]. Familiar supervised learning algorithms are Artificial Neural Network, Decision Tree, Linear Regression, Logistic Regression [1, 14].

The future holds an enhanced ID3 algorithm, which links the information entropy created on unrelated forms with the organisation point in an unfair set model. In ID3, selecting the ideal element is made on the statistics acquisition method, but the logarithm in the algorithm starts the computation complex [15]. This research paper began by considering whether a more straightforward method could be recycled, in which case the decision tree structure technique would be prioritised. The researchers prepared an increased C4.5 decision tree algorithm based on example collection in instructions to improve the categorization precision, decrease the training period of prominent examples, and find the best training set [16]. Their algorithm was initiated on the statistic that a decision tree only suited a restricted optimal solution and has better confidence with the original standard [17].

IV. RESULT DISCUSSION

It provides statistics for fifteen key facts about student performance on Pass/Fail online Machine Learning exams. It has been understood that the basic procedure starts with a dataset which includes an objective parameter that is binary (Pass/Fail) and different binary or unconditional analyst parameters like:

- Whether registered in the New online courses.
- Whether the student is from a game development program or a new training.
- Whether Employed or Not Employed.

Table 1: The dataset has been proposed under

S. No.	Objective parameter	Analyst parameter	Analyst parameter	Analyst parameter
	Exam outcome	New online courses	Student training	Employed status
1	Pass	Y	Game develops	Not Employed
2	Fail	N	Game develops	Employed
3	Fail	Y	Game develops	Employed
4	Pass	Y	OR	Not Employed
5	Fail	N	New training	Employed
6	Fail	Y	New training	Employed
7	Pass	Y	Game develops	Not Employed
8	Pass	Y	OR	Not Employed
9	Pass	N	Game develops	Employed
10	Pass	N	OR	Employed
11	Pass	Y	OR	Employed
12	Pass	N	Game develops	Not Employed
13	Fail	Y	New training	Employed
14	Fail	N	New training	Not Employed
15	Fail	N	Game develops	Employed

Notice that, as shown in Figure 3 below, only one parameter, Student training, has more than 2 levels or groups — Game develops, OR, and New training. The main benefits of Decision Trees compared to other classification models, such as Logistic Regression or Support Vector Machine, are that it does not require one-hot encoding to create these into pseudo parameters. Let us initially discuss the flow of how a decision tree mechanism works, and then we will examine the difficulties involved in making actual decisions.

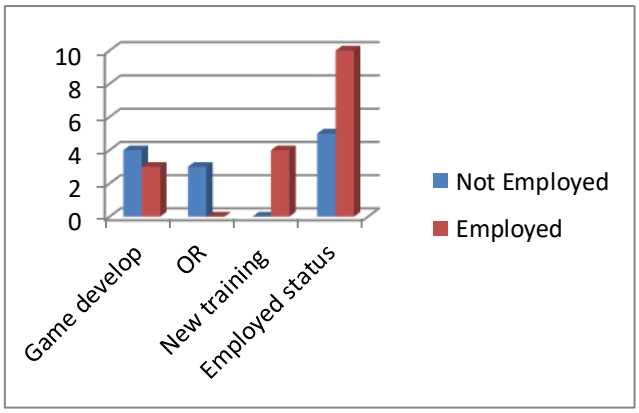

Fig. 3. Dataset for Online Machine Learning Exam

A. Flow of a Decision Tree

A decision tree starts with the Objective parameter. It has frequently been referred to as the parent node. The Decision Tree then creates an order of splits based on the hierarchical order of influence on this Objective parameter.

Published By:
Blue Eyes Intelligence Engineering and Sciences Publication (BEIESP)

From the examination viewpoint, the primary node is the parent node, which has the initial parameter that splits the Objective parameter.

To classify the parent node, it has assessed the effect of all the parameters it has presently on the objective parameter, thereby classifying the parameters that divide the exam Pass/Fail classes into the most similar sets. Our applicants for excruciating this are: Student training, Employed status and New online courses.

What was the expectation to succeed by this split? Assume it starts with the 'Employed' status as the parent node. This is divided into two sub-nodes, one for Employed and the other for Not Employed. Accordingly, the Pass/Fail position has been restructured within an individual sub-node, as shown in Figure 4 below.

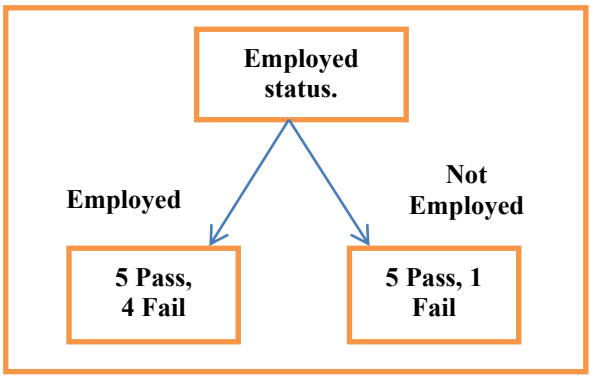

Fig. 4. Decision Tree Flow of Employed Status

Thus, it has done the elementary flow of the Decision Tree. If a sub-node has a combination of Pass and Fail, it is possible to divide the additional attempt and treat it as an individual group. This refers to the clarity of the node. For instance, Not Employed has five Passed and one Failed, whereas the Employed node has five Passed and four Fail. A child node must be unique and hold either a Pass or a Fail class instance. A node that is mixed has been divided further for refining clarity.

However, it has not certainly driven down to the fact that an individual leaf is 'pure'. It is also significant to recognise that the individual nodes have separated, and later, the element that best divided the "Employed" node has not done so that which best divided the "Not Employed" node.

V. CONCLUSION

It is frequently observed that decision trees are particularly memorable due to their graphical depiction and clarity. It has controlled the pool of quality statistics that have been authenticated through statistical methods and are computationally cost-effective. It has also handled great dimensional statistics with real decent accuracy. Moreover, numerous feature selection methods have been applied in constructing the decision tree from parent nodes to child nodes, as well as the decision tree algorithm in Machine Learning. Consequently, that's it for Decision Trees, which form at least two-thirds of the approach. Nearby are numerous difficulties; I have said I will finish. I hope you enjoyed this study on the inner mechanisms of Decision Trees. This article is unique; it is distinct from a modest approach. I have consequently studied the difficulties involved in selecting a parameter hierarchy, constructing a tree, and determining when the cutting process should be terminated. Various types of Decision Tree algorithms have been used, including those in Scikit-Learn. These contain: ID3, C4.5, C5.0 and CART.

FUTURE WORK

Furthermore, a small study has been conducted on the run of evolutionary algorithms for optimal feature selection. Further work is required in this area, as appropriate feature collection in large datasets can significantly improve the performance of the algorithms.

DECLARATION

Funding/ Grants/ Financial Support	No, I did not receive.
Conflicts of Interest/ Competing Interests	No conflicts of interest to the best of our knowledge.
Ethical Approval and Consent to Participate	No, the article does not require ethical approval or consent to participate, as it presents evidence that is not subject to interpretation.
Availability of Data and Material/ Data Access Statement	Not relevant.
Authors Contributions	All authors have equal participation in this arti.cle

REFERENCES

1. Navada, A., Ansari, A., Patil, P., and B. Sonkamble, "Overview of use of decision tree algorithms in machine learning," in 2011 IEEE control and system graduate research colloquium, pp. 37–42, Malaysia, June 2011. [CrossRef]
2. Sekeroglu, B., Hasan, S. S., Abdullah, S. M., Adv. Comput. Vis. 491, 2020 [CrossRef]
3. Lakshmi, T., Aruldoss, M., Begum, R. M., and Venkatesan, V., "An analysis on performance of decision tree algorithms using students' qualitative data," International Journal of Modern Education and Computer Science, vol. 5, no. 5, pp. 18–27, 2013. [CrossRef]
4. Singh, K., "The comparison of various decision tree algorithms for data analysis," International Journal of Engineering and Computer Science, vol. 6, no. 6, pp. 21557–21562, 2017. [CrossRef]
5. Chary, S. N. and Rama, B., "A survey on comparative analysis of decision tree algorithms in data mining," International Journal of Mathematical, Engineering and Management Sciences, vol. 3, pp. 91–95, 2017.
6. Pathak, S., Mishra, I., and Swetapadma A., "An Assessment of Decision Tree Based Classification and Regression Algorithms," in 2018 3rd International Conference on Inventive Computation Technologies (ICICT), pp. 92–95, Coimbatore, India, November 2018. [CrossRef]
7. Moghimipour, I. and Ebrahimpour, M., "Comparing decision tree method over three data mining software," International Journal of Statistics and Probability, vol. 3, no. 3, 2014. [CrossRef]
8. Almasoud, A. M., Al-Khalifa, H. S., and Al-Salman, A., "Recent developments in data mining applications and techniques," in 2015 Tenth International Conference on Digital Information Management (ICDIM), 2015, pp. 36–42. [CrossRef]
9. Anuradha, C. and Velmurugan, T., A data mining-based survey on student performance evaluation system, 2014 IEEE International Conference on Computational Intelligence and Computing Research, 2014, pp. 1–4. [CrossRef]
10. Cherfi, A., Nouira, K., and Ferchichi, A. (2018). Very Fast C4.5 Decision Tree Algorithm, Journal of Applied Artificial Intelligence, 2018, 32(2), pp. 119-139 [CrossRef]

11. Mhetre, V. and Nagar, M., Classification-based data mining algorithms to predict slow, average and fast learners in the educational system using WEKA, in 2017 International Conference on Computing Methodologies and Communication (ICCMC), 2017, pp. 475–479. [CrossRef]
12. Li, M., Application of CART decision tree combined with PCA algorithm in intrusion detection, Presented at the 2017 8th IEEE International Conference on Software Engineering and Service Science (ICSESS), 2017, pp. 38–41. [CrossRef]
13. Rehman, T. U., Mahmud, M., S., Chang, J. K., Jin, Shin, J. Comp. Electron. Agric. 156, 585 (2019). [CrossRef]
14. Chandrasekar, P., Qian, K., Shahriar, H. and Bhattacharya, P., Improving the Prediction Accuracy of Decision Tree Mining with Data Preprocessing, 2017 IEEE 41st Annual Computer Software and Applications Conference (COMPSAC), 2017, pp. 481– 484. [CrossRef]
15. Yi-bin, L., Ying-ying, W. and Xue-wen, R., Improvement of ID3 algorithm based on simplified information entropy and coordination degree, in 2017 Chinese Automation Congress (CAC), 2017, pp. 1526–1530. [CrossRef]
16. Chen, F., Li, X. and Liu, L., Improved C4.5 decision tree algorithm based on sample selection, in 2013 IEEE 4th International Conference on Software Engineering and Service Science, 2013, pp. 779–782.
17. M. A. Muslim, M. A., Nurzahputra, A. and Prasetiyo, B., Improving accuracy of C4.5 algorithm using split feature reduction model and bagging ensemble for credit card risk prediction, in 2018 IEEE International Conference on ICT (ICOIACT), 2018, pp. 141–145. [CrossRef]

AUTHORS PROFILE

Dr. Nirmla Sharma, PhD, from Teerthanker Mahaveer University, Muradabad, U.P., India. Currently working at King Khalid University, Abha, Saudi Arabia, as an Assistant. The Prof department of computer science. Initially, graduated from CCS University, Meerut, U.P., India, and then a master's in computer science from Rajasthan Vidyapeeth, Rajasthan, and MCA from IGNOU, New Delhi. Published 19 papers in International Journals, 02 in National Journals, 7 National Conferences, attended 14 International Conferences, 15 National Workshops/Conferences, and two books. Other responsibilities, e.g., Head, Department. Of CSE and Timetable Convener at AIT, Ghaziabad, INDIA, Head Examiner, for different subjects of C.S. and I.T. in Central Evaluation of M.T.U. NOIDA / U.P.T.U., Lucknow, U.P.- Paper Setter/Practical Examiner in various institutes/universities temporarily, i.e., CCSU Meerut/UPTU, Lucknow.

Sameera Iqbal Muhmmad Iqbal MCS from The Islamia University of Bahawalpur, Pakistan. Currently working at King Khalid University, Abha, Saudi Arabia, as a Lecturer in the Department of Computer Science. Initially graduated from the Islamia University of Bahawalpur, Pakistan. Published two papers in International Journals, and attended 2 International Conferences. Teaching Computer Science courses.