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

Nirmla Sharma, Sameera Iqbal Muhmmad Iqbal

Abstract*: Decision tree study is a predictive modelling tool that is used over many grounds. It is constructed through an algorithmic technique that is divided the dataset in different methods created on varied conditions. Decisions trees are the extreme dominant algorithms that drop under the set of supervised algorithms. However, Decision Trees appearance modest and natural, there is nothing identical modest near how the algorithm drives nearby the procedure determining on splits and how tree snipping 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. Gini index is the name of the level task that has applied to assess the binary changes in the dataset and worked with the definite object variable "Success" or "Failure". Split creation is basically covering the dataset values. Decision trees monitor a top-down, greedy method that has recognized as recursive binary splitting. It has statistics for 15 statistics facts of scholar statistics on pass or fails an online Machine Learning exam. Decision trees are in the class of supervised machine learning. It has been commonly applied as it has informal implement, interpreted certainly, derived to quantitative, qualitative, nonstop, and binary splits, and provided consistent outcomes. The CART tree has regression technique applied to expected standards of nonstop variables. CART regression trees are an actual informal technique of understanding outcomes.

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

I. INTRODUCTION

Decision Trees has supervised machine learning algorithms that have the greatest right 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 works [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 has allocated and leaves, where it developed outcome [2]. The design of a binary tree for supposing whether an Employees has Employed or Not Employed using various statistics like time, work behaviors and movements behaviors [3], has shown under figure 1.

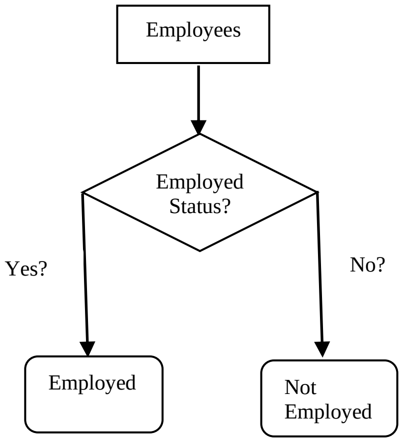

Fig. 1. Decision Tree of Employee [3]

In the above decision tree, the appeal has decision nodes and last outcomes have been 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

Complex the value of Gini index, higher the similarity. A perfect Gini index value is 0 and poorest is 0.5 (for 2 classes difficult). Gini index for a divided has designed with the assistance of following phases −

- First, have analyzed Gini index sub-nodes have get through the formula p^2 + q^2, which has the sum of the square of probability for success and failure [6].
- Next one, analyze Gini index for shared have spent biased Gini score of individually node of that has divided.

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

B. Split Design

It has generated an issued in dataset with the help of next three measures –

- Measure 1: Determining Gini Score

It has required just has deliberated this evaluate 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 index of an attribute and has divided value of that attribute. Afterward receiving the two clusters - right and left, from the dataset, it has analyzed the value of divided by consuming Gini score considered in first measure. Divided value has chosen in which cluster the attribute has exist in.

- Measure 3: Estimating all splits.

Later measure next outcome Gini score and splitting dataset has done the estimation of all splits. For this drive, first, it has done pattern each value related with individually attribute as an applicant split. Then it has 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 has 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

While producing terminal nodes of decision tree, one vital fact has chosen when to end growing tree or generating more terminal nodes. It has ended by applying two standards namely maximum tree depth and minimum node accounts like following steps –

(1) Maximum Tree Depth

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

(2) Minimum Node Records

It has been distinct like the minimum number of preparation arrays that a assumed node is responsible for. It must end addition terminal nodes when tree extended at these minimum node accounts or under this minimum. Terminal node has been applied to create a 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 construct the tree. In this technique, after a node is produced, it has generated the child nodes (nodes extra to an existing node) recursively on individually cluster of data, produced by splitting the dataset, by working the similar purpose again and again. Below figure 2 shows splitting decision tree algorithm [11].

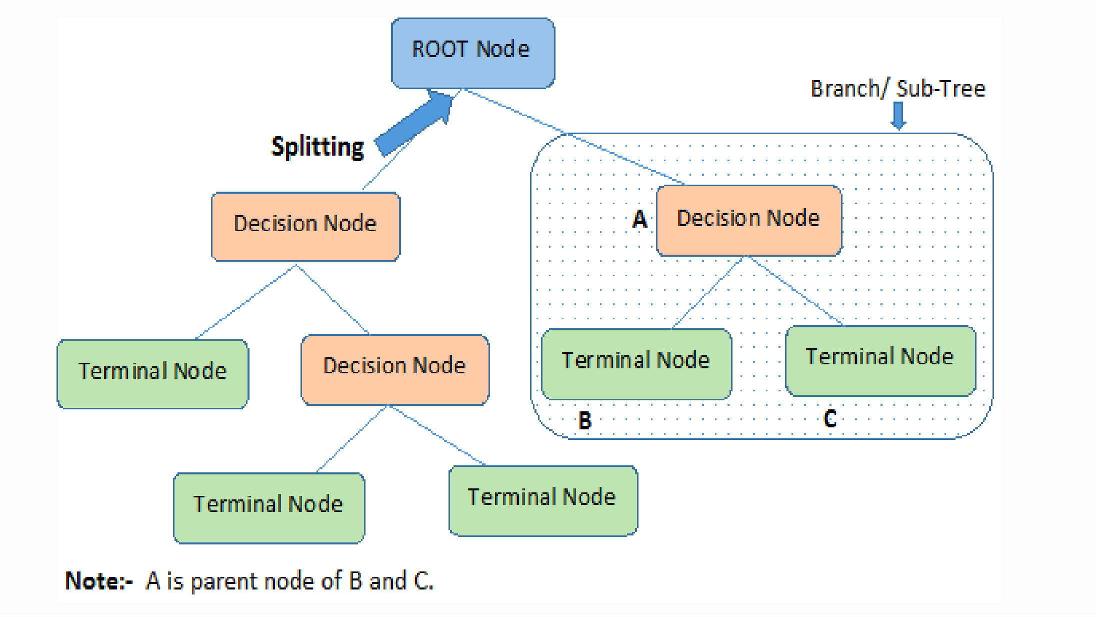

Fig. 2. Splitting Decision Tree Algorithm [11]

II. P PROBLEM STATEMENT

A numeric variable has looked several periods in the data with dissimilar cut offs or thresholds. Also, final classifications have been reiterated. The essential key from data science viewpoint has subsequent difficulties. How to flow of facts through the Decision Tree? This procedure of classification starts with the parent node of the decision tree and develops by relating nearly splitting situations at individually non-leaf node; it splits datasets into a similar subset.

III. RELATED STUDIES

Regression tree is a classification template constructed through relating logistic regression and decision tree. Logistic regression tree is a decision tree with a regression analysis construction.

Retrieval Number:100.1/ijeat.E41950612523
DOI: 10.35940/ijeat.E4195.0612523
Journal Website: www.ijeat.org

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

In this tree structure, logistic regression assessment is completed for individual hierarchy division; formerly, divisions are divided incontrollable the C4.5 decision tree. The last phase is the cut off phase of the tree [6, 12].

This research is an instance of related research effort that we will observe identifying our research better. We will provide a system of similarities with extra way related to ours to improve identify our research paper [7, 13].

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

Decision tree is likely to categorize statistics using a decision tree employed on the statistics. The nodes, leaves, and divisions of a tree are called its functional mechanisms. Interior nodes are the requests that are requested concerning 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 node has a division that tips to a list of likely values for the feature. Unique of the difficulty's class issues is characterized through the nodes at the finish of the diagram, known as child nodes [14]. Machine learning is distinct as identifying designs using well-educated statistics when understanding unnamed input [1]. Machine learning is parted 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 an enhanced ID3 algorithm, which links the information entropy created on unrelated forms with organization point in unfair set model. In ID3, selecting the ideal element is created on the statistics acquire method, but the logarithm in the algorithm starts the computation complex [15]. In this research paper was started on the step that if a simpler method can be recycled, the decision tree structure technique would be prior. The researchers prepared an increased C4.5 decision tree algorithm based on example collection in instruction to progress the categorization precision, decrease the training period of big example, and find the best training set [16]. Their algorithm was initiated on the statistic that decision tree only suited restricted optimal solution and has the better confidence with original standard [17].

IV. RESULT DISCUSSION

It has statistics for fifteen statistics facts of student statistics on Pass or Fail an online Machine Learning exam. It has understood the basic procedure start with a dataset which includes an objective parameter that is binary (Pass/Fail) and different binary or unconditional analyst parameter like:

- Whether registered in New online courses.
- Whether student is from a Game develop, OR and New training.
- Whether Employed or Not Employed.

Table 1: The dataset has been purposed 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 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 like Logistic Regression or Support Virtual Machine that it did not need to move out one warm encrypting to create these into pretend parameters. Let us initial doing the flow of how a decision tree mechanism and then it will joint into the difficulties of how the decisions have really finished.

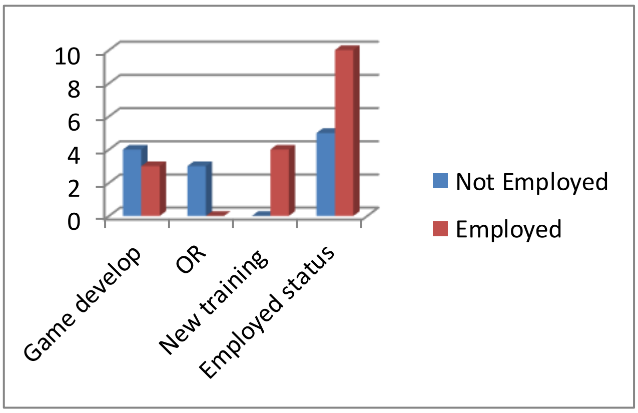

Fig. 3. Dataset for Online Machine Learning Exam

A. Flow of a Decision Tree

A decision tree has started with the Objective parameter. It has frequently named the parent node. The Decision Tree then creates an order of splits based in hierarchical order of influence on this Objective parameter.

After the examination viewpoint the primary node is the parent node, which had the initial parameter that splits the Objective parameter.

To classify the parent node, it has assessed the effect of entirely the parameters that it has presently on the objective parameter to classify the parameters that has divided 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 did it expectation to succeed by this split? Assume it start with Employed status as the parent node. This divided into two sub nodes, separately for Employed and Not Employed. Accordingly, the Pass/Fail position has restructured in individually sub node correspondingly Decision tree figure 4 shown below.

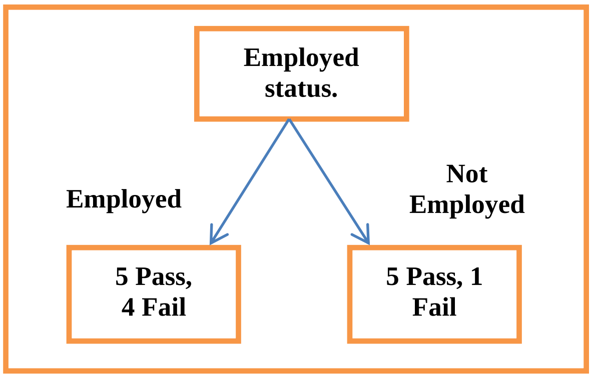

Fig. 4. Decision Tree Flow of Employed Status

Thus, it has done the elementary flow of the Decision Tree. If there is a combination of Pass and Fail in a sub node, here is possibility to divide additional to attempt and acquire it to stand individual group. This is named the clarity of the node. For instance, Not Employed has five Passed and one Failed, later it is cleaner than the Employed node which has five Pass and four Fail. A child node has done be unique which holds either Pass or Fail class individual. A node which is mixed has done be divided additional for refining clarity.

However, it has not done certainly drive down to the fact where individually leaf is 'pure'. It is also significant to recognize that individually node has separated and later the element that finest has divided the "Employed" node has not done the unique that greatest has divided the "Not Employed" node.

V. CONCLUSION

It is frequently detected that decision trees are actual memorable to recognize as of their graphic depiction/clarification. It has controlled the pool of quality statistics that has been authenticated through statistical methods and are cost-effective computationally. It has also handled great dimensional statistics with real decent accurateness. Moreover that, numerous features collection means has applied in constructing the decision tree from parent nodes to child nodes and the decision tree algorithm in Machine Learning. Consequently, that's it for Decision Trees form begins to at least two thirds of the approach. Nearby are numerous difficulties, I have said finish. I hope you enjoyed this study on the inside mechanisms of Decision Trees. Unique article has strong; this is distant from a modest method. I have consequently distant studied individual the difficulties of how parameters hierarchy has selected, and a tree construction has constructed active and how cutting is ended. There have used various kinds of Decision Tree algorithms even in Scikit Study. These contain: ID3, C4.5, C5.0 and CART.

FUTURE WORK

Furthermore, slight study has been completed on the run of evolutionary algorithms for optimal feature assortment, further work requests to be completed in this area as appropriate feature collection in huge datasets can suggestively progress the presentation 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 and consent to participate with evidence.
Availability of Data and Material/ Data Access Statement	Not relevant.
Authors Contributions	All authors have equal participation in this article

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 student's 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]

Retrieval Number:100.1/ijeat.E41950612523
DOI: 10.35940/ijeat.E4195.0612523
Journal Website: www.ijeat.org

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

11. Mhetre, V. and Nagar, M., Classification based data mining algorithms to predict slow, average and fast learners in 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 in King Khalid University Abha, Saudi Arabia as Asst.Prof department of computer science. Initially Graduating from CCS university Meerut U.P. INDIA and then master's in computer science from Rajasthan Vediyapeeth Rajasthan and MCA from IGNOU New Delhi Published 19 Paper in International Journals, 02 in National Journals, 7 National Conferences, attended 14 International Conference 15 National Workshops/Conferences 2 books. Other responsibilities i.e., Head, Dept. 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 different Institutes/Universities time to time i.e., CCSU Meerut/UPTU, Lucknow.

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

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of the Blue Eyes Intelligence Engineering and Sciences Publication (BEIESP)/ journal and/or the editor(s). The Blue Eyes Intelligence Engineering and Sciences Publication (BEIESP) and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.