Comparative Analysis of Decision Tree Attribute Selection Measures for Breast Cancer Diagnosis

Meruva Anusha[1], G V Ramesh Babu[2]

[1]PG Scholar, Dept. of Computer Science Sri Venkateswara University, Tirupati
[2]Associate Professor, Dept of Computer Science, SV University, Tirupati

***Abstract*—** *This research paper presents a comprehensive evaluation of the decision tree algorithm using two attribute selection measures, namely Gini (information gain) and Entropy, for the classification of Breast Cancer Wisconsin (Diagnostic) Data. The primary objectives were to assess the number of selected features for the root node and the resulting learning accuracy. The study found that while both attribute selection measures yield promising results, Entropy outperforms Gini in terms of accuracy and precision. This research sheds light on the importance of feature selection in machine learning models for medical diagnosis.*

I. INTRODUCTION

Disease cancers are brought about by the wild development of cells in the bosom. Quite possibly of the most continuous threat in ladies is bosom disease [1]. Bosom disease is delegated harmless and dangerous. Harmless cancer cells just fill in the bosom and don't divide all through different cells. A harmful growth is comprised of carcinogenic cells that can extend wildly, spread to different region of the body, and contaminate different tissues [5[7]].

Breast cancer is a critical health concern worldwide, and early diagnosis is essential for effective treatment [10][11]. Machine learning algorithms, such as decision trees, have shown promise in diagnosing breast cancer based on patient data. However, the choice of attribute selection measure can significantly impact the model's performance. In this study, we compare the performance of decision trees using two popular attribute selection measures, Gini (information gain) and Entropy, on Breast Cancer Wisconsin (Diagnostic) Data.

II. METHODOLOGY

We employed the decision tree algorithm with two different attribute selection measures, Gini and Entropy, to construct classification models for breast cancer diagnosis. Our primary focus was on evaluating the number of selected features for the root node and the resulting learning accuracy. The dataset used in this study comprises various patient characteristics, such as tumor size, shape, and texture, as well as diagnosis outcomes (benign or malignant).

III. DECISION TREE

The decision tree calculation comprises of arranging information through the investigation of its credits to productively address the information acquired through the information set. In the model being referred to, the hubs address the tests performed on the quality qualities; the circular segments demonstrate the conceivable result for a given test, lastly, the leaves show the last order of the tree over the dataset [3]. Three boundaries were utilized to carry out the decision tree calculation: measure, arbitrary state, and most extreme profundity. The model is the capability that actions the nature of a division; two boundaries are upheld: Gini (Gini contamination) and entropy (data gain) [4]. What's more, the tree has the greatest tree profundity variation, which will characterize how far the tree will be fanned, and the boundaries range from 0 to endlessness.

Decision tree is a different evened out data structure that tends to data through a detachment and conquer method. In portrayal, the goal is to acquire capability with a decision tree that tends to the readiness data so much that names for new models can be settled [6]. The essential focuses of decision tree classifiers are: 1) to arrange successfully whatever amount of the readiness test as could be anticipated; 2) summarize past the planning test so disguised models could be portrayed with as high of an accuracy as could truly be anticipated.

3.1 Attribute Choice Measures

For picking the splitting model that "best" secludes the data portion, D, of class-named planning tuples into solitary classes, we used trademark assurance measure which is heuristic for such decision. Assuming we some way or another figured out how to part D into additional unobtrusive sections according to the consequences of the splitting premise, ideally each bundle would

be pure (i.e., all the tuples that fall into a given fragment would have a spot into a comparable class) [2][5]. The result of this present circumstance is actually the "great" proportion of the huge number of norms taken. Characteristic decision measure chooses how to part the tuples at a given center and are thusly in any case called separating rules [6][8]. The partitioning properties can be perpetual regarded or it will in general be restricted to twofold trees. For constant regarded qualities, a split point ought to be settled as a component of the splitting measure while for the twofold trees a splitting subset ought to be settled. The tree center for package is named with the splitting principle, branches are created for each aftereffect of standard and the tuples are separated suitably. The most notable property decision measures are - Entropy (Data Gain), Gain Proportion and Gini List.

Decision tree is a different evened out data structure that tends to data through a division and defeat method. In portrayal, the goal is to acquire capability with a decision tree that tends to the planning data so much that names for new models can be settled [4]. The essential focuses of decision tree classifiers are: 1) to arrange successfully whatever amount of the readiness test as could be anticipated; 2) summarize past the planning test so hid models could be described with as high of an accuracy as could truly be anticipated.

3.2 Entropy

Entropy is a proportion of vulnerability related with an irregular variable. The entropy increments with the expansion in vulnerability or haphazardness and diminishes with a reduction in vulnerability or irregularity. The worth of entropy goes from 0-1.

Entropy(D) = $\sum_{i=0}^{c} -pi\ log2(pi)$

where pi is the non-zero likelihood that an erratic tuple in D has a place with class C and is assessed by |Ci,D|/|D|. A log capability of base 2 is utilized in light of the fact that as expressed over the entropy is encoded in bits 0 and 1.

3.3 Information Gain

ID3 utilizes data gain as its trait determination measure. Claude Shannon concentrated on the worth or "data content" of messages and gave data gain as an action in his Data Hypothesis [5]. Data Gain is the contrast between the first data gain prerequisite (for example in view of only the extent of classes) and the new necessity (for example gotten after the parceling of A).

Gain (D, A) = Entropy(D)-$\frac{|Dj|}{|D|}$ Entropy (Dj)

Where,

D: A given information segment

A: Trait

V: Assume we parcel the tuples in D on some characteristic A having v particular qualities

D is parted into v segment or subsets, {D1, D2, Dj} were Dj contains those tuples in D that have result aj of A.The characteristic that has the most noteworthy data gain is picked.

IV. EXPERIMENTAL RESULTS

The target of this segment is to assess the choice tree calculation with two characteristic determination estimates like entropy and data gain regarding number of chosen highlights for root hub, and learning precision on chose highlights for Bosom Disease Wisconsin (Indicative) Information have been explored different avenues regarding information taken from the UCI AI Vault [9]. The Bosom disease (Indicative) informational index has 569 columns and 32columns. This information contains two class lebels i.e., the 357 harmless class has 357 and dangerous class has 212. We have utilized the Python Language to explore our proposed calculations. The Python Scikit-learn is a bundle for information grouping, relapse, bunching and perception. The information is partitioned in two sets. The preparation set is 70% and the excess 30% are utilized for testing. Our trial results exhibit the accompanying characterization execution measurements for the chose AI calculations as displayed in the table-1 and figure-1.

Table-1
Performance of Decision Tree Method

Decision Tree Attribute Selection Method	Accuracy	Precision	Recall
Gini	96.74	96.7	96.8
Entropy	97.56	97.56	97.6

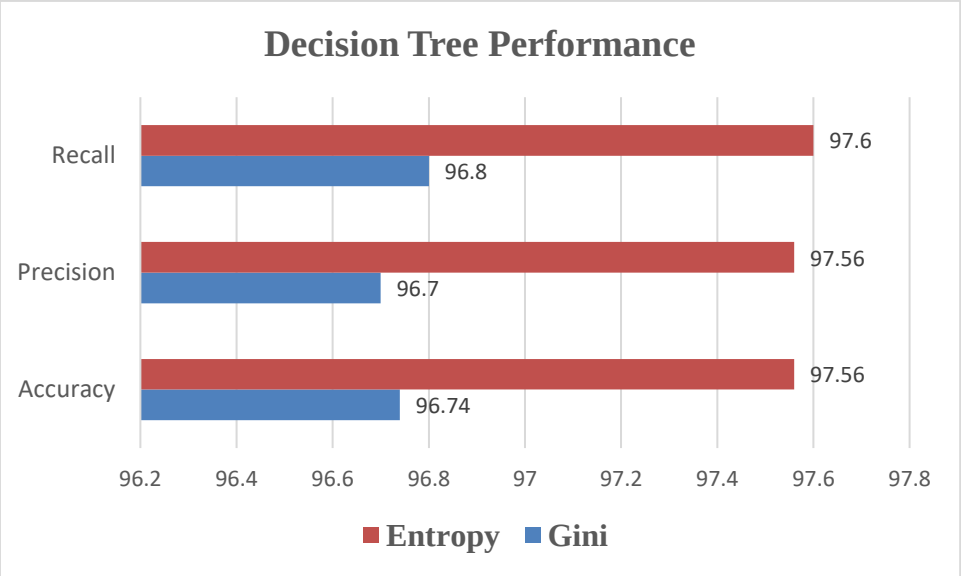

Figure-1: Performance of Decision Tree Method

4.1 Results

Our experimental results from figure-1, indicate the following outcomes:

Gini: The Gini-based decision tree achieved an accuracy of 96.74%, with precision and recall scores of 96.7% and 96.8%, respectively.

Entropy: The decision tree utilizing Entropy as the attribute selection measure outperformed Gini, achieving an accuracy of 97.56% and precision and recall scores of 97.56% and 97.6%, respectively.

These results underscore the importance of selecting the appropriate attribute selection measure when constructing decision tree models for medical diagnosis tasks. In the context of breast cancer diagnosis, Entropy appears to be a more effective choice for identifying relevant features and achieving higher classification accuracy.

V. CONCLUSION

In this study, we conducted a comparative analysis of decision trees using Gini and Entropy as attribute selection measures for breast cancer diagnosis. Our findings indicate that the choice of attribute selection measure significantly impacts the model's performance. Entropy, in particular, demonstrated superior accuracy, precision, and recall compared to Gini.

Ultimately, this research contributes to the ongoing efforts to improve breast cancer diagnosis through machine learning techniques and emphasizes the need for careful consideration of attribute selection methods in model development. Further investigations could explore additional attribute selection measures and their impact on the performance of decision tree models for breast cancer diagnosis.

REFERENCES

[1] Chen SI, Tseng HT, Hsieh CC. Evaluating the impact of soy compounds on breast cancer using the data mining approach. Food & function. 2020;11(5):4561–70

[2] G. Ravi Kumar, K. Venkata Sheshanna, S. Rahamat Basha, and P. Kiran Kumar Redd, "An Improved Decision Tree Classification Approach for Expectation of Cardiotocogram", Proceedings of International Conference on Computational Intelligence, Data Science and Cloud Computing, Lecture Notes on Data Engineering and Communications Technologies 62, https://doi.org/10.1007/978-981-33-4968-1_26

[3] Ian H. Witten and Eibe Frank. Data Mining: Practical machine learning tools and techniques. 2nd ed. San Francisco: Morgan Kaufmann, 2005.

[4] 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.

[5] J. R. Marsilin and G. Wiselin Jiji, "An efficient cbir approach for diagnosing the stages of breast cancer using knn classifier," Bonfring International Journal of Advances in Image Processing, vol. 2, no. 1, 2012.

[6] N. Michael, "Artificial Intelligence - A Guide to Intelligent Systems",2nd edition, Addison Wesley, 2005.

[7] M. A. Mohammed, B. Al-Khateeb, A. N. Rashid, D. A. Ibrahim, M. K. A. Ghani, and S. A. Mostafa, "Neural network and multi-fractal dimension features for breast cancer classification from ultrasound images," Computers & Electrical Engineering, vol. 70, pp. 871–882, 2018

[8] M. V. Lakshmaiah, G. Ravi Kumar and G. Pakardin, "Frame work for Finding Association Rules in Bid Data by using Hadoop Map/Reduce Tool", International Journal of Advance and Innovative Research, Volume 2, Issue1(1), PP:6-9,2015, ISSN: 2394-7780

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

[10] World Health Organization (2021), https://www.who.int/cancer/detection/breastcancer/en/.

[11] Zhang X, Shengli SU, Hongchao WA. Intelligent diagnosis model and method of palpation imaging breast cancer based on data mining. Big Data Research . 2019;5(1):2019005.