A Decision-Making Model for Selecting Criterion in a Decision Tree Algorithm

Ashish P. Joshi[1], Umeshkumar Tank[2], Hasamukhbhai B. Patel[3], Vivek Vyas[4]

[1]Assistant Professor, Computer Science Department, Institute of Science and Technology for Advanced Studies and Research (ISTAR), The CVM University, Vallabh Vidya Nagar, Gujarat, India.
[2]Chief Executive Officer, Vasave Business Solutions, Croydon, London – United Kingdom.
[3]Assistant Professor, Computer Science Department, Natubhai V Patel College of Pure and Applied Sciences, The CVM University, Vallabh Vidyanagar, Gujarat, India
[4]Assistant Professor, School of Management Studies, National Forensic Sciences University, Gandhinagar, Gujarat, India

Received: 16.04.2024 Revised : 17.05.2024 Accepted: 25.05.2024

ABSTRACT
The decision tree algorithm is widely used for classification, which is a part of supervised learning. It classifies the node based on criterions for information gain which are termed as Entropy and Gini index. A comparative study of these techniques is required to determine the most feasible criterion for the specific dataset. As of now, machine learning and healthcare go hand in hand. Diabetes is a common disease that occurs when blood glucose levels exceed the normal range. It is expected to affectmillions of people, with half of the population remaining undiagnosed. The medical field generates a large amount of data, which is required to be analyzed further using machine learning. This paper focuses on machine learning, specifically the Supervised Learning's algorithm decision tree, which is used for classification, to predict whether a patient is diabetic or not. When using a classification technique, the criterion plays an important role. Here, two popular criterions are discussed and compared for the accuracy of both. Ultimately, the prediction is based on a comparative study of both the Gini Index and the entropy to determine the best way to gain more accuracy. The algorithm is used on dataset, which is fed with different parameter values to get the result of the prediction.

Keywords: Machine Learning, Supervised Learning, Decision Tree, Classification, Entropy, Gini Index, Data Preprocessing

1. INTRODUCTION

The machine learning is very well known sub branch of artificial intelligence. It gives computers the power and capability to learn on their own training without being given any specific instructions. The concept of machine learning is very new and stands in contrast to more traditional programming. In traditional programming, user provides the computer with input together with a program that contains the logic in order to obtain an output from the computer. In contrast, the process of machine learning involves the transmission of input and output to a computer for the purpose of training. This gives the computer the ability to construct its own programming logic, which can then be evaluated. Machine learning makes use of computer algorithms to take in and evaluate data in order to make accurate projections about future output. When new data is put into these algorithms, they are able to learn and improve their processes. Over time, this raises both their level of performance and their intelligence.

There are three main approaches to learning: reinforced learning, unsupervised learning, and supervised learning. A training data set is used in supervised learning to instruct and train the model to generate the desired output. Unsupervised learning uses machine learning techniques to analyze and aggregate unlabeled information without the need for human interaction. These algorithms find hidden patterns or data groupings.Reinforced learning is the process of adopting the behavior that maximizes reward in a given situation. Various apps and computers use it to find the optimal potential action or course of action for each event. Reinforced learning is the process of adopting the behavior that maximizes reward in a given situation. Various apps and computers use it to find the optimal potential action or course of action for each event.

Classification and regression problems are the two main types of supervised learning. Regression methods are utilized when there is known to be a correlation between input and output, while classification methods are employed when the output is a categorical variable (e.g., yes/no, male/female, true/false).

A key goal of regression analysis is to deduce the nature of the connection among the independent variables and dependent variables. It's a common instrument for guesstimating things like future revenues for a company. Regression likes logistic regression, polynomial regression and linear regression. Thereare some of the most well-known techniques for doing regression analysis.

Using a classification technique, we can properly sort test data into their respective groups. It finds entities in the dataset and makes an educated guess about what those entities ought to be called or how they ought to be defined. In the following part, we'll dive further into some of the well known classification methods, such as support vector machines (SVM),k-nearest neighbor,linear classifiers,decision trees, and random forest.

This paper is about analyzing the criterion-based method of decision trees for determining node rankings.

1.1 Introduction to Decision Tree

The "nodes" in a decision tree are the properties of interest and the questions we have about them, the "edges" are the possible answers to those questions, and the "leaves" are the final output or class name. In some contexts, a decision tree may be referred to as a decision matrix. They could be helpful in making non-linear decisions when a simple linear decision surface is inadequate. Each node in a decision tree has a specific purpose: the root node provides the framework for the entire structure, the decision node makes the final determination, and the leaf nodes provide the final output.

Classification examples are fed into a decision tree and sorted from the root node to a leaf node, with the latter yielding the classification. For some property, the tree's nodes represent test cases, and the edges emanating from those nodes represent viable solutions. This procedure is recursive and is performed for each new subtree.

Node ordering is crucial, hence measures of randomness and disorder like the entropy and Gini index are often employed. Algorithms utilizing decision trees rely on information gain to partition a node. Information gain is evaluated using the Gini index and entropy. An impure node can be quantified using either the Gini or entropy measure. While a node with a single class is considered pure, a node with many classes is considered impure.

Let’s discuss decision tree with an example. The below table is representing the data about buying cars from the ruler as well as urban area. Also it is representing the income of the buyers from these areas.

Table 1. Dataset about car buyers

Area	Income			Total
	<25K	25K to 50K	>50K	
Rural	1	4	7	12
Urban	3	6	9	18
Total	4	10	16	30

The above table can be classified by two ways because it has two parameters such as income and area. The possibilities of generating the decision tree from above table are as below. Among them which decision tree has more information gain that is decided by one of the criterion such as entropy and Gini index.

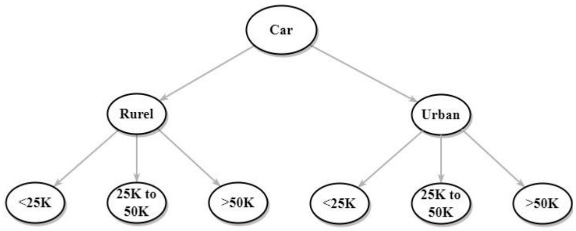


Fig 1. Decision tree sample-1

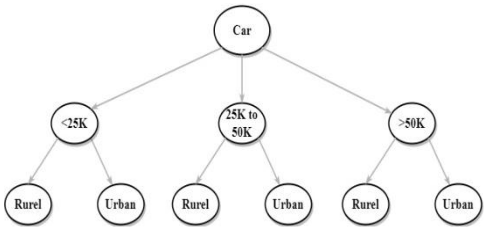


Fig 2. Decision tree sample-2

Entropy and Gini index are both measures used in decision tree algorithms to evaluate the impurity or disorder of a set of data points. They are commonly employed as criteria for splitting nodes in a decision tree. Here's a comparison of entropy and Gini index:

1.2 Introduction to entropy

For both statistics and thermodynamics, entropy is a measure of chaos. If a node belongs to many classes, the node is disorganized. Information gain is the difference between the parent node's entropy and the sum of the offspring nodes' weighted entropies. A child node's weight is calculated by taking its sample count and dividing it by the sum of all of its sample counts[8].

Entropy, in the context of Machine Learning, is a measure of the disorder or unpredictability of the data being analyzed. Entropy, then, is a metric used in machine learning to assess the degree to which a system is chaotic or contaminated.

Note that Pi = Probability of randomly selecting an example in class.

$$Entropy\ (p) = -\sum_{i=1}^{N} p_i \ log_2\, p_i$$

Eq-1: Equation for Entropy Calculation

Properties of Entropy

- Definition: It measures the randomness or disorder in a set of data. The formula for entropy is often expressed as H(S) = -Σ p(i) * log2(p(i)), where p(i) is the proportion of data points in class i.
- Range: The entropy values range from 0 to log2(c), where c is the number of classes. Lower entropy values indicate less disorder.
- Interpretability: The values are more intuitive – a lower Gini index corresponds to a more pure node.
- Computation Complexity: Involves the computation of logarithmic functions, which might be computationally more expensive.
- Sensitivity to class imbalance: Can be sensitive to class imbalance as it considers the distribution of all classes.
- Decision tree splitting: Generally leads to more balanced trees.
- Use Cases: May be preferred when the classes are imbalanced or when you want a more balanced tree.

1.3 Introduction to Gini Index

The efficiency with which a decision tree is partitioned can be quantified with the help of a function called Gini impurity. It helps us zero in on the best splitter to use while building a purely decision tree. This method can be used as a starting point in identifying the best splitter. To calculate the Gini Index, take the sum of the squares of the probabilities for each class and subtract one. It works best with larger divisions and is straightforward to deploy. For each feature, it calculates how likely it is to be incorrectly categorised. The Gini Index is a numeric measure from 0 to 1, where 0 represents perfect categorization and 1 represents a uniform distribution of data across categories. When the Gini Index is 0.5, it means that items are distributed uniformly across all categories. The Gini Index performs simply a binary split on category variables and reports results in terms of success or failure.

Note that Pi = Probability of randomly selecting an example in class.

$$Gini = 1 - \sum_{i=1}^{C} (p_i)^2$$

Eq 2: Equation for Gini Index Calculation

The above both criterion is used for calculating the information gain and based on the information gain the ordering of the node can be decided. The proposed model describes that how to compare accuracy of both criterion and decide the feasible criterion for the specific set of data.

Properties of Gini Index

- Definition: It measures the impurity of a set by calculating the probability of misclassifying an element chosen randomly. The Gini index is defined as Gini(S) = 1 - Σ p(i)^2.
- Range: The Gini index values range from 0 to 1. A Gini index of 0 indicates perfect purity, while 1 indicates maximum impurity.
- Interpretability: The values are more difficult to interpret directly in terms of "purity" or "impurity" compared to the Gini index.
- Computation Complexity: Computationally less expensive as it involves simple squared terms.
- Sensitivity to class imbalance: Also takes into account the class distribution, but it is less affected by imbalances than entropy.
- Decision tree splitting: Tends to favour larger partitions and can create more extensive trees.
- Use Cases: Often used by default and may perform well in practice.

In practice, the choice between entropy and Gini index might not make a significant difference in the resulting decision tree performance. The decision of which impurity measure to use can depend on the specific characteristics of the dataset and the goals of the model. Some machine learning libraries and algorithms allow you to choose the impurity measure as a parameter during the model training phase.

2. Proposed Model

The comparative study of both criterions required a specific model to be developed which helps to decision making for classification. The some steps of this model such as dataset selection, data pre-processing, training data with entropy criterion, train model with Gini index criterion are implementing. Also, measuring the accuracy provided by experiment and compare the accuracy for the decision making process in the final stage.

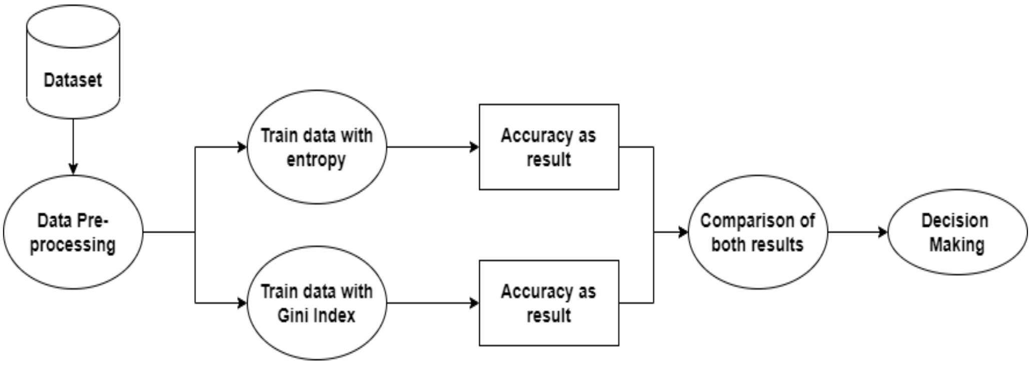


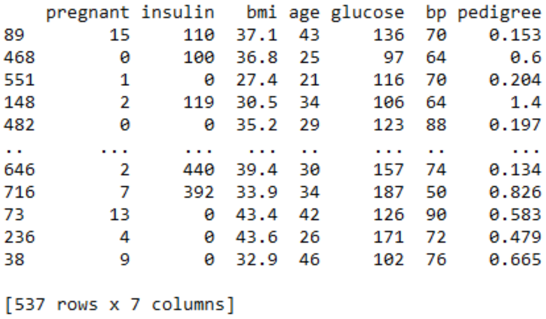
Fig 4. Proposed Model for Comparing Entropy and Gini Index criterion.

3. Dataset Selection and Data Pre-processing

The preliminary stage of the proposed model is the dataset selection. The row dataset is not providing good result. So, it is require pre-processing and generating the meaningful data which can be ready to use for analysis.

3.1 Dataset Selection

This experimental study is using dataset about the diabetic patient[7]which has some important attributes (parameters for diabetes prediction) like pregnant, glucose, bp, skin, insulin, bmi, pedigree, age, and label. Except the label all input data are the independent data and the label is only dependent on the others independent attributes. The label is storing the information about the result of the patient which is in Boolean form as yes or no. This result depends on the value of all independent attributes.



```
     pregnant  insulin   bmi  age  glucose  bp  pedigree
89         15      110  37.1   43      136  70     0.153
468         0      100  36.8   25       97  64       0.6
551         1        0  27.4   21      116  70     0.204
148         2      119  30.5   34      106  64       1.4
482         0        0  35.2   29      123  88     0.197
..        ...      ...   ...  ...      ...  ..       ...
646         2      440  39.4   30      157  74     0.134
716         7      392  33.9   34      187  50     0.826
73         13        0  43.4   42      126  90     0.583
236         4        0  43.6   26      171  72     0.479
38          9        0  32.9   46      102  76     0.665

[537 rows x 7 columns]
```

Fig 5. Dataset of Diabetic Patients[7]

3.2 Data Pre-processing

An essential phase in the pipeline for data analysis and machine learning is data preprocessing . It entails preparing unprocessed data so that robots can comprehend and analyze it with ease. The efficacy and precision of machine learning models can be greatly enhanced by appropriate preprocessing of the data[9]. Here are some common steps involved in data preprocessing such as *Data Cleaning, Data* Transformation, Data Reduction, Data Splitting, Data Merging, dealing with imbalanced Data, handling different kind of Data, Normalization, Noise remove or reduction, and Data Integration.

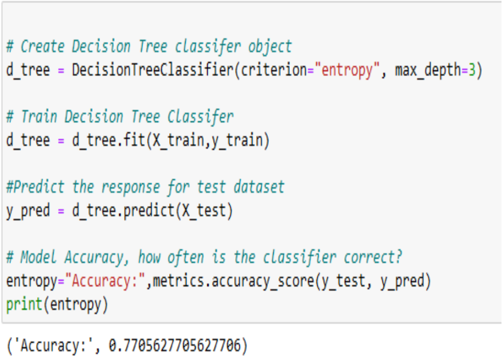
The specific preprocessing steps can vary depending on the nature of the data and the problem you are trying to solve. It's essential to understand the characteristics of your data and choose the preprocessing techniques accordingly.

4. Experiment and Result

The significant of data pre-processing is organizing the cluttered data which makes data analysis efficient and reliable. The experimental study is expedient to deciding whether entropy is suitable for the specific dataset or Gini index.

4.1Train data with entropy

Training data with entropy typically involves introducing a measure of uncertainty or randomness into the dataset. Entropy is a concept from information theory that quantifies the amount of disorder or uncertainty in a system. In the context of machine learning, adding entropy to the training data can be beneficial in certain scenarios, such as promoting model robustness or enhancing generalization. The fig-6 is training the data and showing the result as accuracy by applying the entropy criterion.



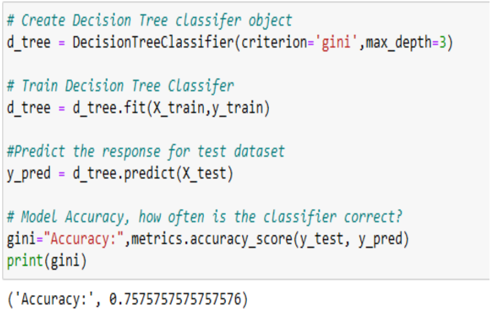
```
# Create Decision Tree classifer object
d_tree = DecisionTreeClassifier(criterion="entropy", max_depth=3)

# Train Decision Tree Classifer
d_tree = d_tree.fit(X_train,y_train)

#Predict the response for test dataset
y_pred = d_tree.predict(X_test)

# Model Accuracy, how often is the classifier correct?
entropy="Accuracy:",metrics.accuracy_score(y_test, y_pred)
print(entropy)
```

```
('Accuracy:', 0.7705627705627706)
```

Fig 6. Train data with entropy and find out the Accuracy

4.2 Train data with Gini Index

The Gini Index is a measure of impurity or inequality often used in decision tree algorithms for binary classification. It quantifies how often a randomly chosen element would be incorrectly classified. The lower the Gini Index, the better the split. The fig-7 is training the data and showing the result as accuracy by applying the gini criterion.



```
# Create Decision Tree classifer object
d_tree = DecisionTreeClassifier(criterion='gini',max_depth=3)

# Train Decision Tree Classifer
d_tree = d_tree.fit(X_train,y_train)

#Predict the response for test dataset
y_pred = d_tree.predict(X_test)

# Model Accuracy, how often is the classifier correct?
gini="Accuracy:",metrics.accuracy_score(y_test, y_pred)
print(gini)
```

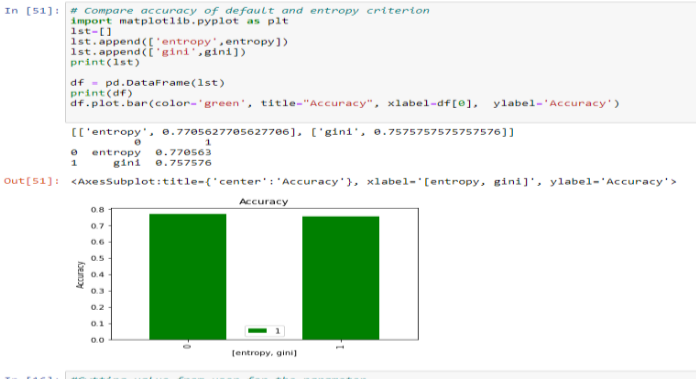
```
('Accuracy:', 0.7575757575757576)
```

Fig 7. Train data with Gini index and find out the Accuracy

4.3 Result Analysis and Visualization

If you're working on a specific problem, it's a good idea to experiment with both criteria and possibly other hyper parameters to determine the optimal configuration for your particular dataset. Many machine learning libraries, provide options to choose between Gini impurity and entropy when building decision trees.

This experimental study is obtaining both results as accuracy of decision tree with different criterion such as entropy and gini index. The data of diabetic patient trained by both criterion and found the accuracy as per the fig-8. It is showing the comparison of accuracy obtained by both criterions. It is found more accuracy obtained by entropy rather thengini index for this specific data. However, it depends on the dataset so, it is more important to decide which criterion will be fit for the desired data analysis.



```
In [51]: # Compare accuracy of default and entropy criterion
         import matplotlib.pyplot as plt
         lst=[]
         lst.append(['entropy',entropy])
         lst.append(['gini',gini])
         print(lst)

         df = pd.DataFrame(lst)
         print(df)
         df.plot.bar(color='green', title="Accuracy", xlabel=df[0],  ylabel='Accuracy')

         [['entropy', 0.7705627705627706], ['gini', 0.7575757575757576]]
                  0         1
         0  entropy  0.770563
         1     gini  0.757576
Out[51]: <AxesSubplot:title={'center':'Accuracy'}, xlabel='[entropy, gini]', ylabel='Accuracy'>
```

Fig 8

To put it briefly, the output of entropy and Gini impurity is a measurement of the disorder or impurity present in a collection of data points. These metrics are used by the decision tree method to determine how to partition the data into a tree structure that can reliably categorize newly discovered and unseen data.

5. CONCLUSION

In conclusion, both entropy and Gini index are effective metrics for measuring impurity, and the choice between them often depends on the specific characteristics of the dataset and the goals of the modeling task. While entropy may be preferred in certain situations, the Gini index has its advantages, and the choice between them might not dramatically impact the overall performance of a decision tree model. It's often a good practice to experiment with both metrics and see which one performs better for a given dataset and problem. The discussion and implementation of both criterions are given in this paper for checking impurity and decision making for which criterion is suitable for the specific dataset.

REFERENCES

[1] Samuel Aning, MałgorzataPrzybyła-Kasperek, “Comparative Study of Twoing and Entropy Criterion for Decision Tree Classification of Dispersed Data”, 26th International Conference on Knowledge-Based and Intelligent Information &EngineeringSystems(KES 2022), Procedia Computer Science 207 (2022) 2434-2443.

[2] Xiaowei L.: Application of decision tree classification method based on information entropy to web marketing. Sixth InternationalConference on Measuring Technology and Mechatronics Automation, 121–127, IEEE. doi: 10.1109/ICMTMA.2014.34. 2014.

[3] PoojaGulati, Amita Sharma, Manish Gupta: Theoretical Study of Decision Tree Algorithms to IdentifyPivotal Factors for Performance Improvement: A Review, International Journal of Computer Applications (0975 – 8887)Volume 141 – No.14, May 2016.

[4] IbomoiyeDomorMienyea , YanxiaSuna, Zenghui Wang: Prediction performance of improved decision tree-based algorithms: a review, 2nd International Conference on Sustainable Materials Processing and Manufacturing (SMPM 2019), Procedia Manufacturing 35 (2019) 698–703

[5] Krzysztof Gr ˛abczewski and Norbert Jankowski: Feature Selection with Decision Tree Criterion, Conference Paper,DOI: 10.1109/ICHIS.2005.43 · Source: IEEE Xplore, https://www.researchgate.net/publication/4219423_Feature_selection_with_decision_tree_criterion (Access date:01-11-23).

[6] DariuszMikulski :Rough Set Based Splitting Criterion for Binary Decision Tree Classifiers, August 2006, Thesis for: U.S. Army Tank Automotive Research, Development and Engineering Center. Source- https://www.researchgate.net/publication/236671728_Rough_Set_Based_Splitting_Criterion_for_Binary_Decision_Tree_Classifiers (access date: 25-10-23).

[7] https://www.kaggle.com/uciml/pima-indians-diabetes-database

[8] Fatma Mohammed Abd El_latif Mousa: Optimal Entropy to enhance the structure of the Wavelet-Packets-Best-Tree for Automatic Speech Recognition, Article 1, Volume 8, Issue 2, September 2021, Page 1-15. Link:https://ejle.journals.ekb.eg/article_195052.html

[9] Chimpiri Sai kiran, S Maragatham: Study on Child Malnutrition using Data Visualization, link: https://www.researchsquare.com/article/rs-3679789/v1