INTERNATIONAL JOURNAL OF PROGRESSIVE RESEARCH IN ENGINEERING MANAGEMENT AND SCIENCE (IJPREMS)

e-ISSN : 2583-1062

www.ijprems.com
editor@ijprems.com

Vol. 04, Issue 02, February 2024, pp : 518-521

Impact Factor : 5.725

# STUDY PAPER: DECISION TREE ALGORITHMS IN MACHINE LEARNING

**A. Rexlin Freeda**[1]

[1]M. Sc, Department of Computer Science, Fatima College, Madurai, India.

## ABSTRACT

A Decision tree is a flowchart-like tree structure where each internal node denotes the feature, branches denote the rules and the leaf node denotes the result of algorithm. The ultimate outcomes or predictions are represented by the leaf nodes of the tree, conveying the results generated by the algorithm. A decision tree is one of the most powerful tools of supervised learning algorithms used for both classification and regression tasks. This paper provides the detailed information about the application of decision tree in Machine Learning.

**Keywords:** Machine Learning, Supervised, Classification, Decision Tree.

## 1. INTRODUCTION

In the contemporary era, technological advancements, particularly in the realm of Machine Learning (ML), have reached unprecedented levels, revolutionizing various facets of human work and interaction. Machine Learning, situated within the broader field of artificial intelligence, amalgamates principles from statistics and computer science to construct algorithms capable of enhancing their efficiency through exposure to relevant data, rather than relying on explicit programming instructions. For Machine Learning, there are many uses, the most prominent of which is predictive data mining. Two major mechanisms can be broken into Machine Learning classification fulfilments, model development and model evaluation.

In the realm of supervised machine learning, each input data object is accompanied by a preassigned class label, forming a labeled dataset. The fundamental objective of supervised algorithms is to learn a model that accurately assigns labels to new, unseen data instances based on the patterns discerned from the provided labeled dataset. This process is particularly emphasized in the context of classification algorithms.

The classification process is utilized to categorize datasets, employing an investigative approach to systematically sort objects or events into distinct groups or categories. The significance of classification and identification lies in their ability to enhance our comprehension of relationships and connections between various entities. Moreover, they serve as essential tools for facilitating clear communication among scientists and researchers, providing a standardized language for conveying findings. In the realm of data science and mining, classification algorithms play a crucial role by automating this process.

A decision tree is a method based on a tree structure, where each path starting from the root is defined by a sequence of data partitions until reaching a Boolean outcome at the leaf node. This hierarchical representation encapsulates knowledge relationships through nodes and connections. In the context of classification, nodes serve as representations of specific decision points or criteria. This paper undertakes a thorough examination of recent and highly effective approaches employed by researchers in the last three years within various domains of machine learning, specifically focusing on decision trees. The details of these methods, encompassing the utilization of algorithms/approaches, datasets, and the resulting findings, are comprehensively summarized. Furthermore, the study sheds light on the predominant approaches frequently employed and accentuates the methods that have yielded the highest accuracy.

## 2. DECISION TREE ALGORITHM

One prevalent technique in data mining involves the creation of classifiers . Within data mining, classification algorithms are adept at managing vast amounts of information. They can be utilized to make assumptions about categorical class names, categorize knowledge based on training sets and class labels, and classify newly acquired data. The realm of classification algorithms in machine learning encompasses various methods, and in this study, the paper concentrates on the decision tree algorithm in a general context. Figure 1 illustrates the structure of a decision tree (DT).

INTERNATIONAL JOURNAL OF PROGRESSIVE RESEARCH IN ENGINEERING MANAGEMENT AND SCIENCE (IJPREMS)

e-ISSN : 2583-1062

www.ijprems.com
editor@ijprems.com

Vol. 04, Issue 02, February 2024, pp : 518-521

Impact Factor : 5.725

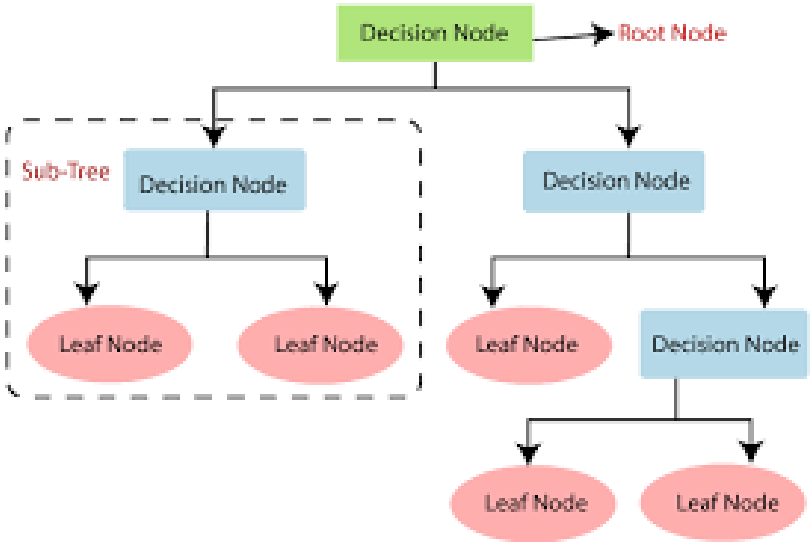



**Figure 1.** Decision Tree

(A)Types of Decision Tree Algorithms

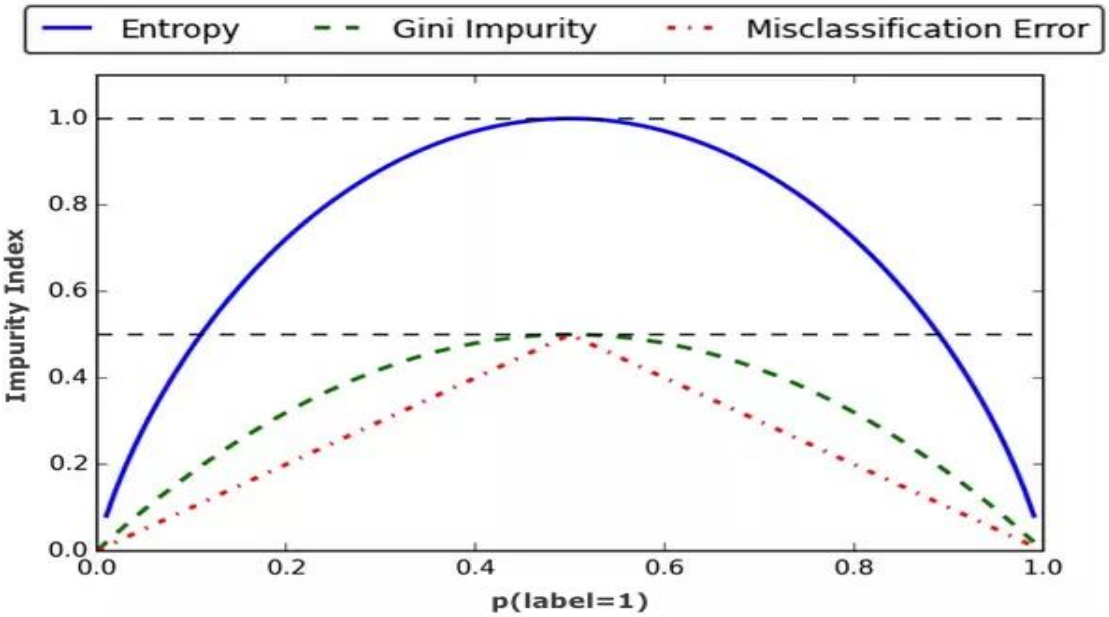
- Iterative Dichotomiser or ID3
- CHAID (Chi-Squared automatic interaction detection
- C4.5 **(**successor of ID3)
- CART (Classification and Regression Tree)

(B)Entropy and Information Gain

Entropy serves as a metric for assessing the impurity or randomness within a dataset [52], [53]. Its value typically ranges from 0 to 1, with lower values indicating better results, as they signify less randomness or impurity. When the target is $G$ with various attribute values, the entropy of the classification set $S$ with respect to $c$ is described by equation (1). It helps determine the best split for building an informative decision tree model.

$$\text{Entropy}(S) = \sum_{i=1}^{C} p_i \log 2\, p_i \qquad (1)$$

Where $p_i$ is the ratio of the sample number of the subset and the $i$-th attribute value.




Information gain, also referred to as mutual information, serves as a metric for segmentation. It provides insight into how much knowledge of a random variable's value is conveyed. Unlike entropy, higher values of information gain are considered better. The data gain, denoted as $Gain(S,A)$, is defined based on the concept of entropy, as depicted in "equation (2)" .

$$\text{Gain}(S, A) = \text{Entropy}(S) - \sum_{v \in V(A)} |s_v|/|s|\ \text{Entropy}(s_v) \qquad (2)$$

Where the range of attribute A is $V(A)$, and $Sv$ is a subset of set $S$ equal to the attribute value of attribute $v$.

INTERNATIONAL JOURNAL OF PROGRESSIVE RESEARCH IN ENGINEERING MANAGEMENT AND SCIENCE (IJPREMS)

e-ISSN : 2583-1062

Impact Factor : 5.725

www.ijprems.com
editor@ijprems.com

Vol. 04, Issue 02, February 2024, pp : 518-521

**Table1.**Comparison between the most used algorithms in decision tree

| Methods | CART | C4.5 | CHAID |
|---|---|---|---|
| **Pruning** | Pre-pruning using a single-pass algorithm | Pre-pruning using a single-pass algorithm | Pre-pruning using Chi-square test for independence |
| **Dependent variable** | Categorical/ Continuous | Categorical/ Continuous | Categorical |
| **Input variables** | Categorical/ Continuous | Categorical/ Continuous | Categorical/ Continuous |
| **Split at each node** | Binary; Split on linear combinations | Multiple | Multiple |

## 3. CHARACTERISTICS OF DECISION TREE ALGORITHMS

When discussing the traits of Decision Trees, the ID3 algorithm is exclusively simulated using the WEKA tool, and the dataset is strictly categorical.

ID3 is unable to handle continuous datasets for simulation. Similarly, CART and C4.5 have same the characteristics of ID3 are shared by both C4.5 and CART.

The sole distinction lies in the fact that C4.5 and CART can process continuous datasets for simulation purposes. Table2

**Table2:** Characteristics of Decision Tree.

| Decision Tree Algorithm | Data Types | Numerical Data Splitting Method | Possible Tool |
|---|---|---|---|
| CHAID | Categorical | N/A | SPSS answer tree |
| ID3 | Categorical | No Restriction | WEKA |
| C4.5 | Categorical, Numerical | No Restriction | WEKA |
| CART | Categorical, Numerical | Binary splits | CART 5.0 |

The decision tree presents all potential options and follows each path to its conclusion within a single view, facilitating easy comparison among alternatives [12].

Its transparency is one of its greatest advantages. Additionally, it excels in selecting the most influential feature and possesses a comprehensible nature.

Decision trees are also straightforward to classify and interpret, suitable for both continuous and discrete datasets. Variable screening and feature selection are notably robust in decision trees [19]. When considering its performance, decision trees remain unaffected by non-linearity across their parameters.

## 4. DATASET DESCRIPTION

The dataset which is used in this experiment is car dataset. By applying decision tree algorithms which are ID3,C4.5 and CART which is described as follows.

The car dataset is in two parts. One is Car Acceptability and other is Technical Characteristic. Overall price (buying) and Price of Maintenance (maint) are two attributes of Car Acceptability. The Number of doors (doors), Capacity in terms of persons carries (persons), the size of luggage boot (lug_boot) and estimated the safety of the car (safety).

**Number of Instances:** 1728

**Number of Attributes:** 6

**Missing Attributes value:** None

**Table.3**

| Attribute | Attribute value |
|---|---|
| Buying | v-high, high, med, low |
| Maint | v-high, high, med, low |
| Doors | 2,3,4,5-more |
| Persons | 2,4-more |
| lug_boot | Small, medium, big |
| Safety | low,med,high |

INTERNATIONAL JOURNAL OF PROGRESSIVE RESEARCH IN ENGINEERING MANAGEMENT AND SCIENCE (IJPREMS)

e-ISSN : 2583-1062

Impact Factor : 5.725

www.ijprems.com
editor@ijprems.com

Vol. 04, Issue 02, February 2024, pp : 518-521

**Class Distribution**

**Table.4**

| class | N | N[%] |
|---|---|---|
| Unacc | 1210 | 70.0233% |
| Acc | 384 | 22.222% |
| good | 69 | 3.993% |
| v-good | 65 | 3.762% |

**Experiment**

The experiment is conducted using the WEKA tool, which serves as a comprehensive resource for data mining tasks. WEKA comprises a range of machine learning algorithms and tools for data pre-processing, classification, regression, clustering, association rules, and visualization. Released under the GNU General Public License, WEKA is open-source software. It is particularly adept at facilitating the development of new machine learning schemes. Users have the flexibility to apply algorithms directly to datasets or invoke them from custom Java code.

**Table 5:** Theoretical Results

| Algorithm | Attribute Type | Missing Value | Pruning Strategy | Outlier Detection |
|---|---|---|---|---|
| ID3 | Only categorical values | No | No | Susceptible to outlier |
| CART | Categorical and Numerical both | Yes | Cost complexity pruning is used | Can handle |
| C4.5 | Categorical and Numerical both | Yes | Error based pruning is used | Susceptible to outlier |

**Usage of Decision Tree Algorithms**

1. **Variable Selection:** Decision trees aid in choosing the most relevant input variables from a pool of potentially numerous variables. Similar to stepwise variable selection in regression analysis, decision tree methods streamline the process by identifying key variables, thereby facilitating hypothesis formulation and subsequent research.
2. **Assessing variable importance:** Once relevant variables are identified, understanding their relative importance becomes crucial. Variable importance is typically determined by assessing the impact on model accuracy or node purity when a variable is removed. Variables that affect a larger number of records generally hold greater importance.
3. **Handling missing values:** Traditional approaches to handling missing data, such as excluding cases with missing values, are both inefficient and prone to bias. Decision tree analysis offers two strategies for managing missing data: treating missing values as a separate category for analysis or utilizing a decision tree model to predict missing values based on other variables.
4. **Prediction:** Prediction is a fundamental application of decision tree models. By leveraging historical data to build a tree model, it becomes straightforward to predict outcomes for future records.

## 5. CONCLUSION

In this paper decision tree algorithms which are ID3, C4.5 and CART discussed in this paper. This paper demonstrates the comparative analysis, usage and datasets are evaluated using these algorithms in Weka tool. The comprehensive about decision tree algorithms is given here.

## 6. REFERENCES

[1] Classification based on Decision tree Algorithm Research Paper
[2] Study and Analysis of Decision tree Based Classification Algorithm Journal
[3] Decision tree methods: applications for classification and prediction Research paper
[4] Application of decision tree in business field Research paper.