Decision Tree Classifier

# **Table Of Contents**

1. Introduction
2. Working of Decision Tree
3. Pros & Cons of Decision Tree
4. Model
5. Conclusion

Introduction

* Decision Tree is a Supervised learning techniquethat can be used for both classification and Regression problems, but mostly it is used for solving classification problems. It is a tree-structured classifier, whereinternal nodes represent the features of a dataset, branches represent the decision rules and each leaf node represents the outcome.
* In a Decision tree, there are two nodes, which are the **Decision Node** and**Leaf Node.** Decision nodes are used to make any decision and have multiple branches, whereas Leaf nodes are the output of those decisions and do not contain any further branches.
* The decisions or the test are performed on the basis of features of the given dataset.
* It is a graphical representation for getting all the possible solutions to a problem/decision based on given conditions.
* A decision tree simply asks a question, and based on the answer (Yes or No), it further split the tree into subtrees.



Working of Decision Tree

Decision Tree follows below mentioned steps:

**Step-1:** The tree starts with the root node, says R, which contains the complete dataset.

**Step-2:** Finding the best attribute in the dataset using Attribute Selection Measure (ASM). There are two methods for ASM:-

1. **Information Gain :**

measurement of changes in entropy after the segmentation of a dataset based on an attribute.

Information Gain =

 Entropy(S)- [(Weighted avg) \*Entropy(each feature)

Entropy(s)= -P(yes)log2 P(yes)- P(no) log2 P(no)

Where, S= Total number of samples

P(yes)= probability of yes

P(no)= probability of no

1. **Gini Index :**

measurement of impurity or purity used while creating a decision tree in the CART (Classification and Regression Tree) algorithm.

Gini Index= 1- ∑jPj2

**Step-3:** Dividing the R into subsets that contains possible values for the best attributes.

**Step-4:** Generating the decision tree node, which contains the best attribute.

**Step-5:** Recursively we make new decision trees using the subsets of the dataset created in step -3. And keep performing this process until a stage is reached where we cannot further classify the nodes and called the final node as a leaf node.



Pros & Cons of Decision Tree

|  |  |
| --- | --- |
| Pros | Cons |
| It follows same concept as human follows while making any decision in real-life. | Due to lots of layers, it is complex. |
| Very helpful for solving decision-making related problem. | Tends to overfitting easily. |
| Helps to think about all possible outcomes of a problem. | For more class labels, it becomes more complex to use this algorithm. |
| Less requirement for data cleaning rather than any other algorithm. | Small changes in data, tends to big changes in results. |

Model Training & Predicting Results

**C4.5 → (successor of ID3) -**C4.5 uses Gain Ratio

The range of the Entropy is from 0 to 1, while the range of the Gini Index is from 0 to 0.5

1. While **Criterion = entropy**
2. **Classification Report:**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | Precision | Recall | F1-score | support |
| 0 [<=50K] | 0.88 | 0.88 | 0.88 | 7452 |
| 1 [>50K] | 0.61 | 0.61 | 0.61 | 2317 |
| Accuracy |  |  | 0.82 | 9769 |
| Macro avg. | 0.75 | 0.74 | 0.74 | 9769 |
| Weighted avg. | 0.82 | 0.82 | 0.82 | 9769 |

1. **Confusion Matrix:**

[[6558 894]

[906 1411]]

True Positives(TP) = 6558

True Negatives(TN) = 1411

False Positives(FP) = 894

False Negatives(FN) = 906

The confusion matrix shows 6558 + 1411 = 7969 correct predictions and 894 + 906 = 1800 incorrect predictions.

In this case, we have

1. **Accuracy:**

Accuracy is: 0.81

1. While **Criterion = gini**
2. **Classification Report:**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | Precision | Recall | F1-score | support |
| 0 [<=50K] | 0.88 | 0.87 | 0.87 | 7452 |
| 1 [>50K] | 0.59 | 0.62 | 0.60 | 2317 |
| Accuracy |  |  | 0.81 | 9769 |
| Macro avg. | 0.73 | 0.74 | 0.74 | 9769 |
| Weighted avg. | 0.81 | 0.81 | 0.81 | 9769 |

1. **Confusion Matrix:**

[[6460 992]

[886 1431]]

True Positives(TP) = 6460

True Negatives(TN) = 1431

False Positives(FP) = 992

False Negatives(FN) = 886

The confusion matrix shows 6460 + 1431 = 7891 correct predictions and 992 + 886 = 1878 incorrect predictions.

In this case, we have

1. **Accuracy:**

Accuracy is: 0.80

Conclusion

While we considering ‘gini’ as criterion, accuracy score is 80% and accuracy score for “ criterion = entropy” is 81%