

A Project Report for course

**CS3142 Predictive Analytics**

On

DECISION TREES

Submitted to

**Manipal University, Jaipur**

Towards the partial fulfilment for the Award of the Degree of

**BACHELOR OF TECHNOLOGY**

In

**Computer Science and Engineering**

2020-24

By

**Siris Gupta 209301062**

Under the guidance of

**Mr Harish Sharma**

Department of Computer Science and Engineering

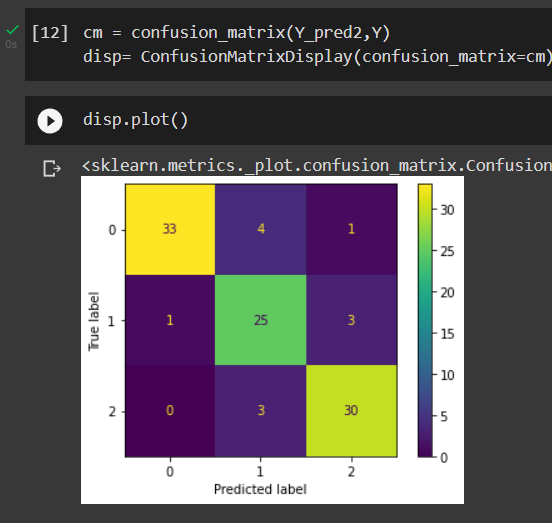
School of Computing and Information Technology

Manipal University Jaipur

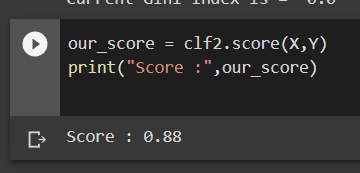
Jaipur, Rajasthan

**Problem Statement: Creating a Decision tree program that can be trained on various datasets to predict target attribute using Gini index or Gain ratio.**

**Confusion Matrix:**

****

**Score:**

****

**About**

A decision tree is a non-parametric supervised learning algorithm, which is utilized for both classification and regression tasks. It has a hierarchical, tree structure, which consists of a root node, branches, internal nodes, and leaf nodes.

**Abstract**

Decision trees are powerful and popular tools for classification and prediction. Decision trees represent rules, which can be understood by humans and used in knowledge system such as database.

A decision tree is a hierarchical model for supervised learning whereby the local region is identified in a sequence of recursive splits in a smaller number of steps. A decision tree is composed of internal decision nodes decision node and terminal leaves. Each decision node m implements a test function fm(x) with discrete outcomes labelling the branches.

Given an input, at each node, a test is applied and one of the branches is taken depending on the outcome. This process starts at the root and is repeated recursively until a leaf node is hit, at which point the value written in the leaf constitutes the output.

A decision tree is also a nonparametric model in the sense that we do not assume any parametric form for the class densities and the tree structure is not fixed a priori but the tree grows, branches and leaves are added, during learning depending on the complexity of the problem inherent in the data.

Decision tree is a classifier in the form of a tree structure which consists of:  Decision node: specifies a test on a single attribute.

 Leaf node: indicates the value of the target attribute.

 Edge: split of one attribute

 Path: a disjunction of test to make the final decision.

**Methodology**

1. Algorithmic Framework for Decision Tree:

Hunt’s algorithm, which was developed in the 1960s to model human learning in Psychology, forms the foundation of many popular decision tree algorithms, such as the following:

- ID3: Ross Quinlan is credited within the development of ID3, which is shorthand for “Iterative Dichotomiser 3.” This algorithm leverages entropy and information gain as metrics to evaluate candidate splits. Some of Quinlan’s research on this algorithm from 1986 can be found here (PDF, 1.4 MB) (link resides outside of ibm.com).

- C4.5: This algorithm is considered a later iteration of ID3, which was also developed by Quinlan. It can use information gain or gain ratios to evaluate split points within the decision trees.

- CART: The term, CART, is an abbreviation for “classification and regression trees” and was introduced by Leo Breiman. This algorithm typically utilizes Gini impurity to identify the ideal attribute to split on. Gini impurity measures how often a randomly chosen attribute is misclassified. When evaluating using Gini impurity, a lower value is more ideal.

All ID3, C4.5 and CART have been used in our project.

1. Univariate Splitting Criteria:

In most of the cases, the discrete splitting functions are univariate. Uni-

variate means that an internal node is split according to the value of a single

attribute. Consequently, the inducer searches for the best attribute upon which

to split. There are various univariate criteria. These criteria can be character-

ized in different ways, such as:

According to the origin of the measure: information theory, dependence,

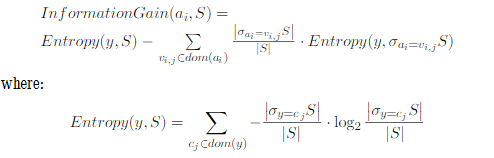
and distance.

According to the measure structure: impurity based criteria, normalized

impurity based criteria and Binary criteria.

Information gain is an impurity-based criterion that uses the entropy mea-

sure (origin from information theory) as the impurity measure (Quinlan, 1987).



Gini index is an impurity-based criterion that measures the divergences be-

tween the probability distributions of the target attribute’s values. The Gini in-

dex has been used in various works such as (Breiman et al., 1984) and (Gelfand

et al., 1991) and it is deﬁned as

A picture containing graphical user interface

Description automatically generated

Consequently the evaluation criterion for selecting the attribute ai is deﬁned as:

Shape, arrow

Description automatically generated

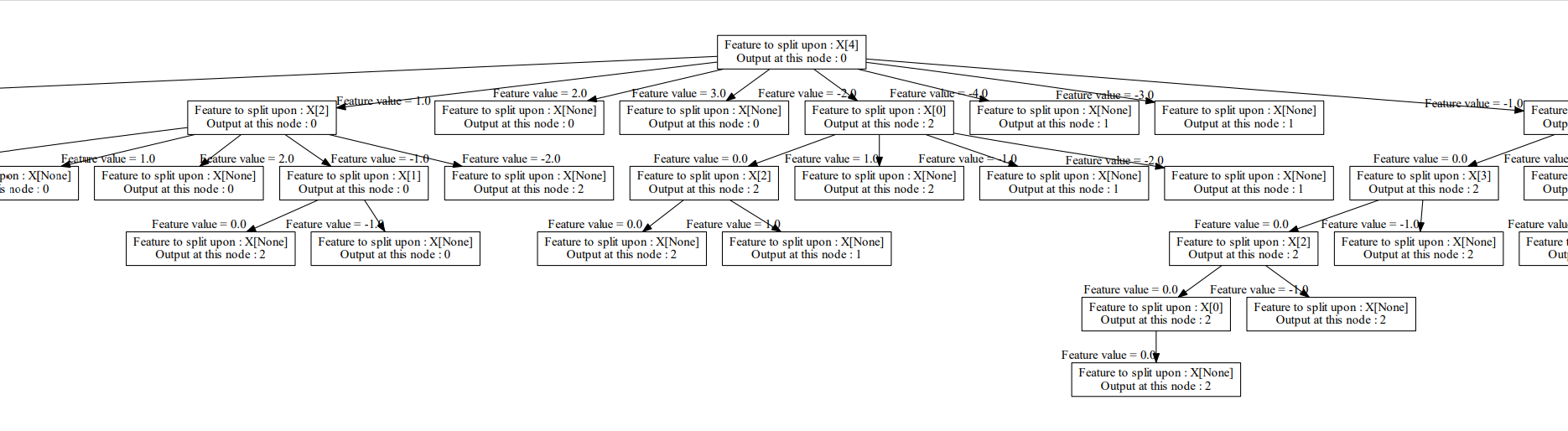
The gain ratio “normalizes” the information gainas follows (Quinlan, 1993):

Text

Description automatically generated

Note that this ratio is not deﬁned when the denominator is zero. Also the ratio

may tend to favor attributes for which the denominator is very small. Consequently, it is suggested in two stages. First the information gain is calculated for all attributes. As a consequence, taking into consideration only attributes that have performed at least as good as the average information gain, the attribute that has obtained the best ratio gain is selected. It has been shown that the gain ratio tends to outperform simple information gain criteria, both from the accuracy aspect, as well as from classiﬁer complexity aspects (Quinlan, 1988)



Graphical user interface, diagram, application, Word

Description automatically generated

Graphical user interface, application

Description automatically generated

Graphical user interface, diagram, application, Word

Description automatically generated

**Advantages**

1. Easy to interpret the decision rules

2. Nonparametric so it is easy to incorporate a range of numeric or categorical data layers and there is no need to select unimodal training data.

3. Robust with regard to outliers in training data.

**Code: Refer to Colab file submitted along with the report.**

**Disadvantages**

1. Decision trees tend to over fit training data which can give poor results when applied to the full data set.

2. Not possible to predict beyond the minimum and maximum limits of the response variable in the training data.

**Application**

1. It is used in filtering of spam emails.

2. Decision tree is used in field of medicine. Ex: To predict the type of people prone to specific type of Virus.

**Conclusion**

As we can see, we have successfully created a decision tree algorithm. We have demonstrated and verified the working of the project.

It has been a great learning experience for me to learn how to implement various aspects of Artificial Intelligence that I had only learnt theoretically and was able to capture how to setup some of them in real life. Overall, it was a great learning experience.