**Final Year B. Tech. (CSE) – I : 2021-22**

**4CS462 : PE2 - Data Mining Lab**

**Assignment No. 4**

**Group id: DM21G12**

**Group members:**

**Abhishek More(2018BTECS00037)**

**Sushil Wagh(2018BTECS00031)**

**Title :** Use / extend the data analysis tool (menu driven GUI) developed in Assignment No. 2

to perform the following classification task

**Objective/Aim :**

1. Implement the decision tree classifier using the following attribute selection

measures and graphically show/visualize the tree:

a. Information Gain

b. Gain Ratio

C. Gini Index

2. Tabulate the results in confusion matrix and evaluate the performance of above

classifier using following metrics :

a) Recognition rate

b) Misclassification rate

c) Sensitivity

d) Specificity

e) Precision & Recall

3. Use the following categorical data sets from UCI machine learning repository :

a. Balance Scale data set

b. Car evaluation data set

c. Breast-cancer data set

**Introduction:**

Information Gain

ID3 uses information gain as its attribute selection measure. This measure is based on

pioneering work by Claude Shannon on information theory, which studied the value or

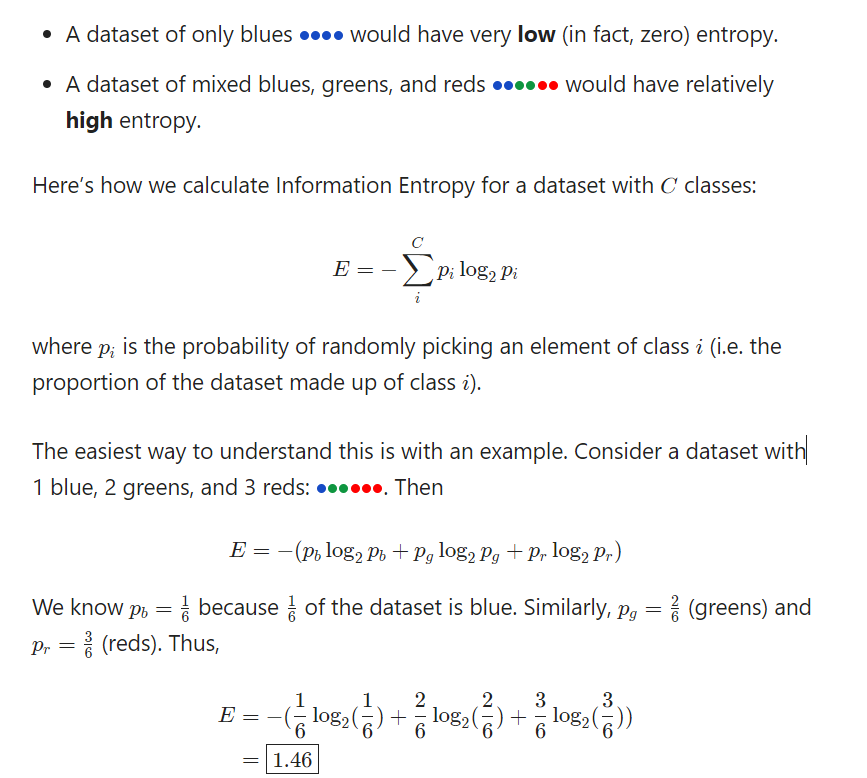
“information content” of messages. Let node N represent or hold the tuples of partition

D. The attribute with the highest information gain is chosen as the splitting attribute for

node N. This attribute minimizes the information needed to classify the tuples in the

Before we get to Information Gain, we have to first talk about Information Entropy. In the context of training Decision Trees, Entropy can be roughly thought of as how much variance the data has.

For example:

****

**Information Gain** = Entropy before splitting - Entropy after splitting

**What is Gain Ratio?**

Proposed by John Ross Quinlan, Gain Ratio or Uncertainty Coefficient is used to normalize the information gain of an attribute against how much entropy that attribute has. Formula of gini ratio is given by

**Gain Ratio=**Information Gain/Entropy

From the above formula, it can be stated that if entropy is very small, then the gain ratio will be high and vice versa.

Be selected as splitting criterion, Quinlan proposed following procedure,

First, determine the information gain of all the attributes, and then compute the average information gain.

Second, calculate the gain ratio of all the attributes whose calculated information gain is larger or equal to the computed average information gain, and then pick the attribute of higher gain ratio to split.

**What is Gini Index?**

The gini index, or gini coefficient, or gini impurity computes the degree of probability of a specific variable that is wrongly being classified when chosen randomly and a variation of gini coefficient. It works on categorical variables, provides outcomes either be “successful” or “failure” and hence conducts binary splitting only.

The degree of gini index varies from 0 to 1,

Where 0 depicts that all the elements be allied to a certain class, or only one class exists there.

The gini index of value as 1 signifies that all the elements are randomly zdistributed across various classes, and

A value of 0.5 denotes the elements are uniformly distributed into some classes.

It was proposed by Leo Breiman in 1984 as an impurity measure for decision tree learning and is given by the equation/formula;Gini index formula

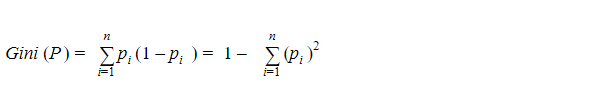
where P=(p1 , p2 ,.......pn ) , and pi is the probability of an object that is being classified to a particular class.

## What is Gini Index?

The gini index, or gini coefficient, or gini impurity computes the degree of probability of a specific variable that is wrongly being classified when chosen randomly and a variation of gini coefficient. It works on categorical variables, provides outcomes either be “successful” or “failure” and hence conducts binary splitting only.

The degree of gini index varies from 0 to 1,

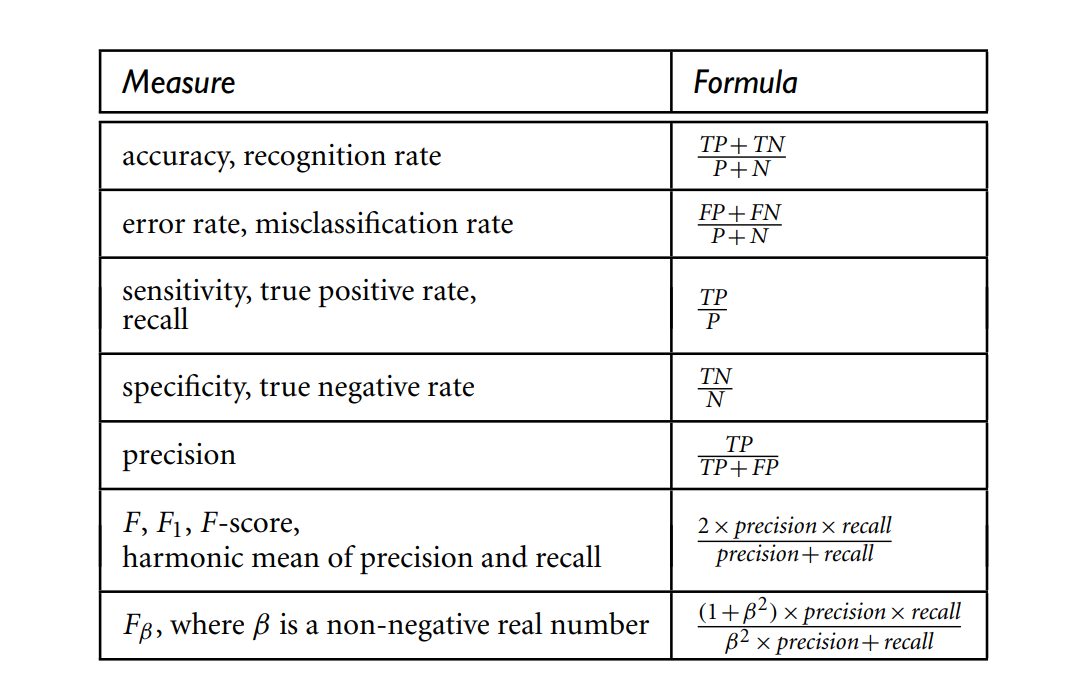
* Where 0 depicts that all the elements be allied to a certain class, or only one class exists there.
* The gini index of value as 1 signifies that all the elements are randomly zdistributed across various classes, and
* A value of 0.5 denotes the elements are uniformly distributed into some classes.

It was proposed by [Leo Breiman](https://scholar.google.com/citations?user=mXSv_1UAAAAJ&hl=en) in 1984 as an impurity measure for decision tree learning and is given by the equation/formula;

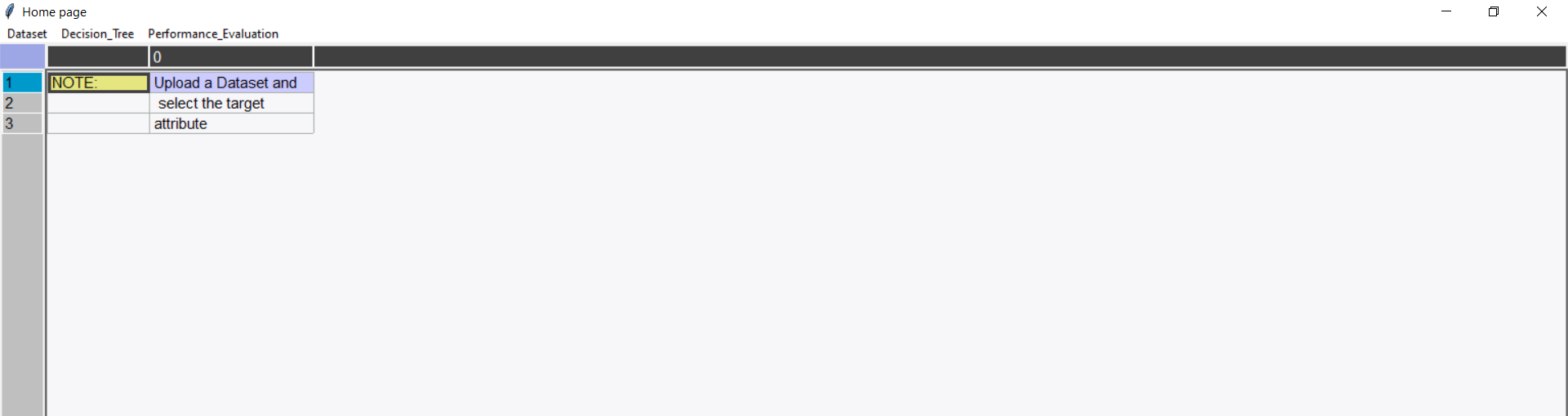
where P=(p1 , p2 ,.......pn ) , and pi is the probability of an object that is being classified to a particular class.

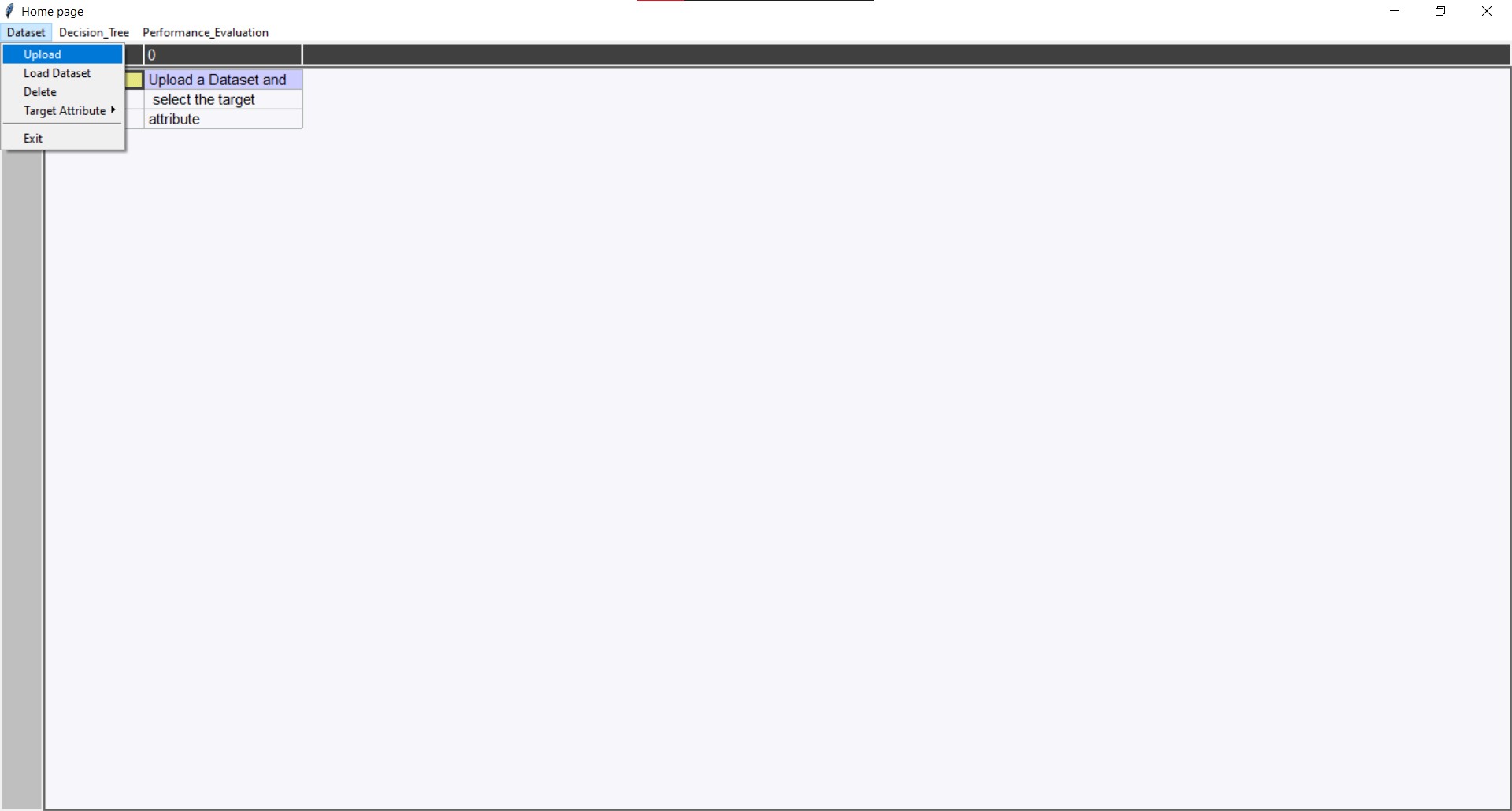
Also, an attribute/feature with least gini index is preferred as root node while making a decision tree.

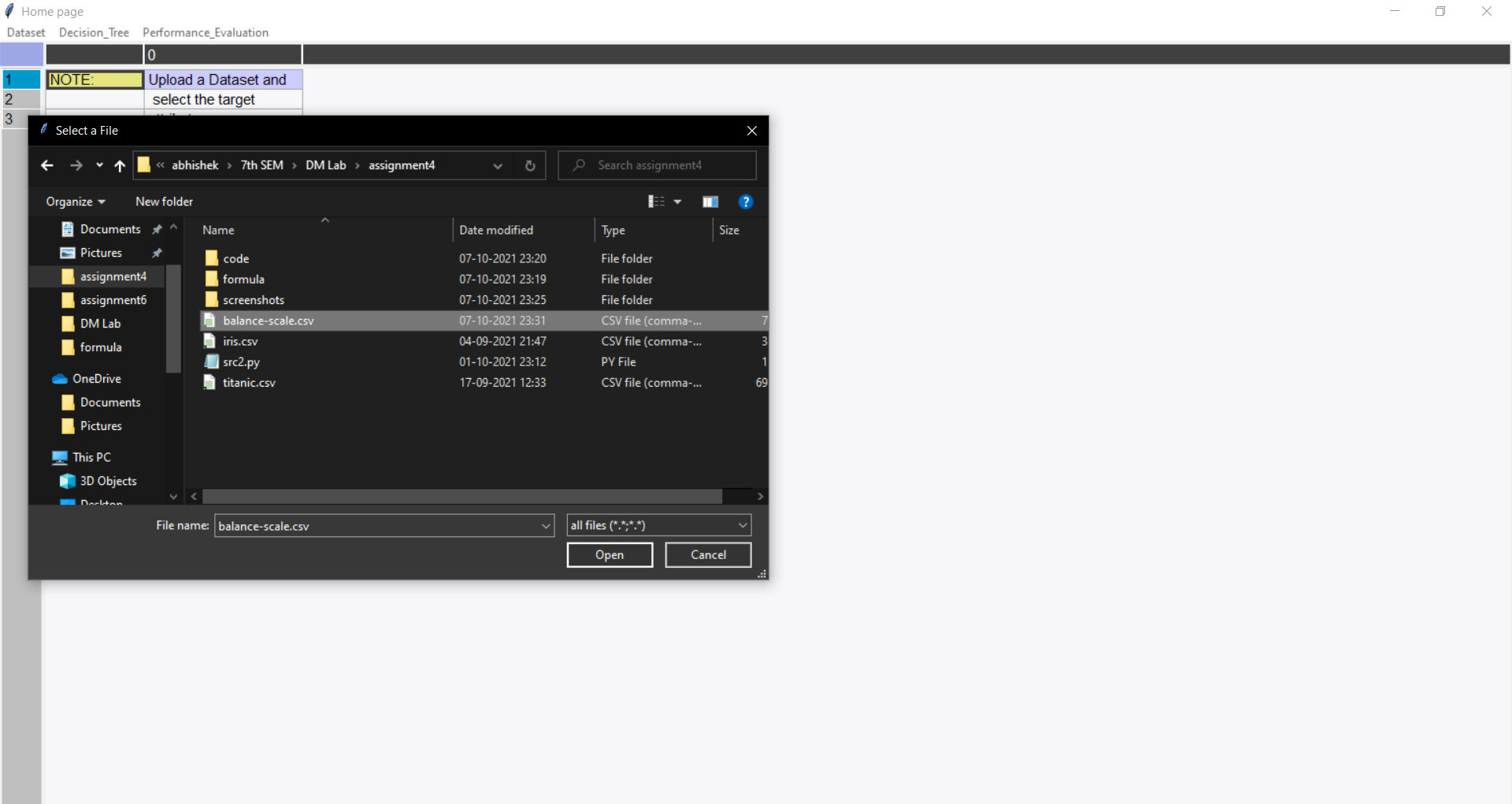
Also, an attribute/feature with least gini index is preferred as root node while making a decision tree.

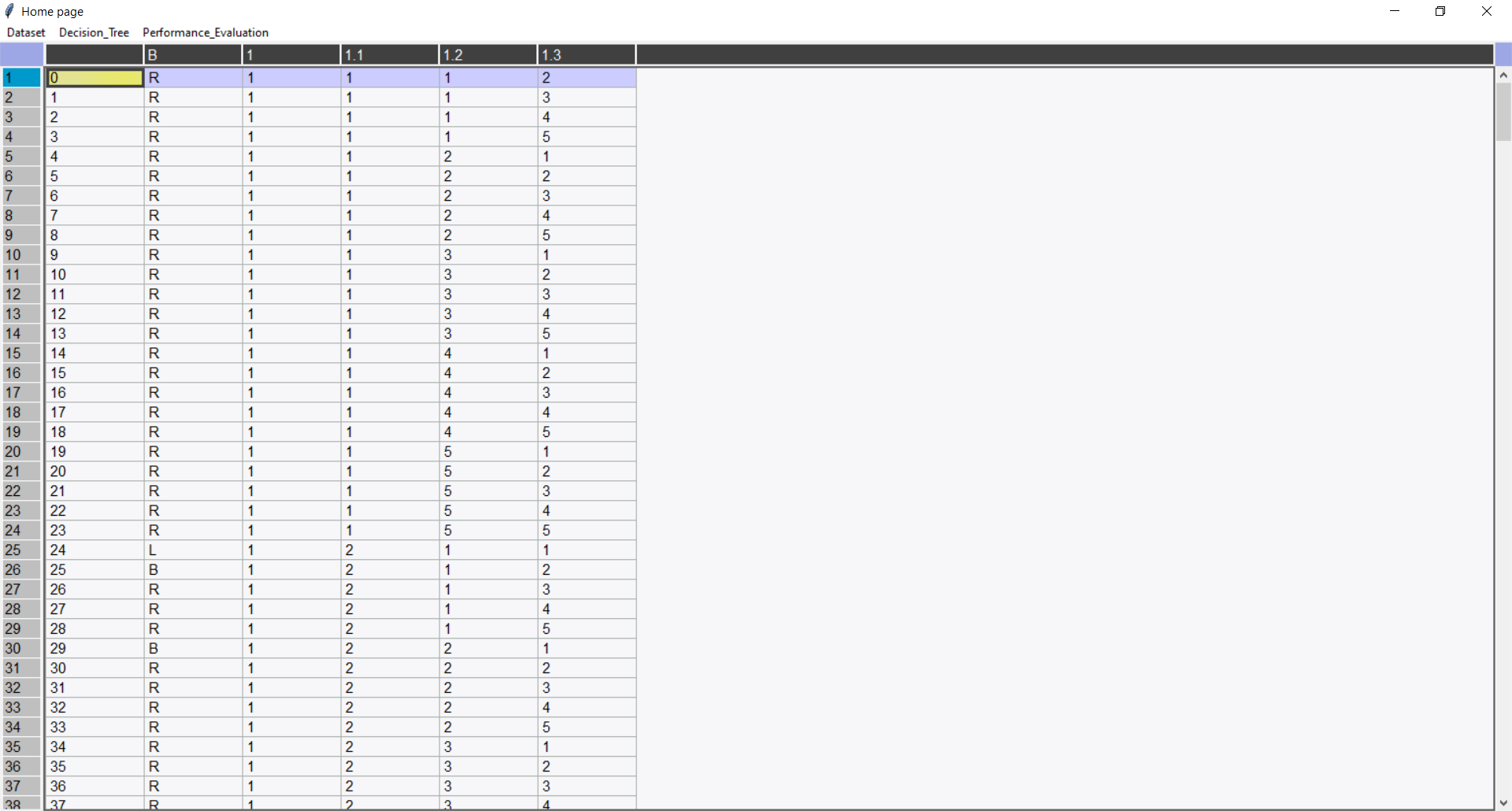


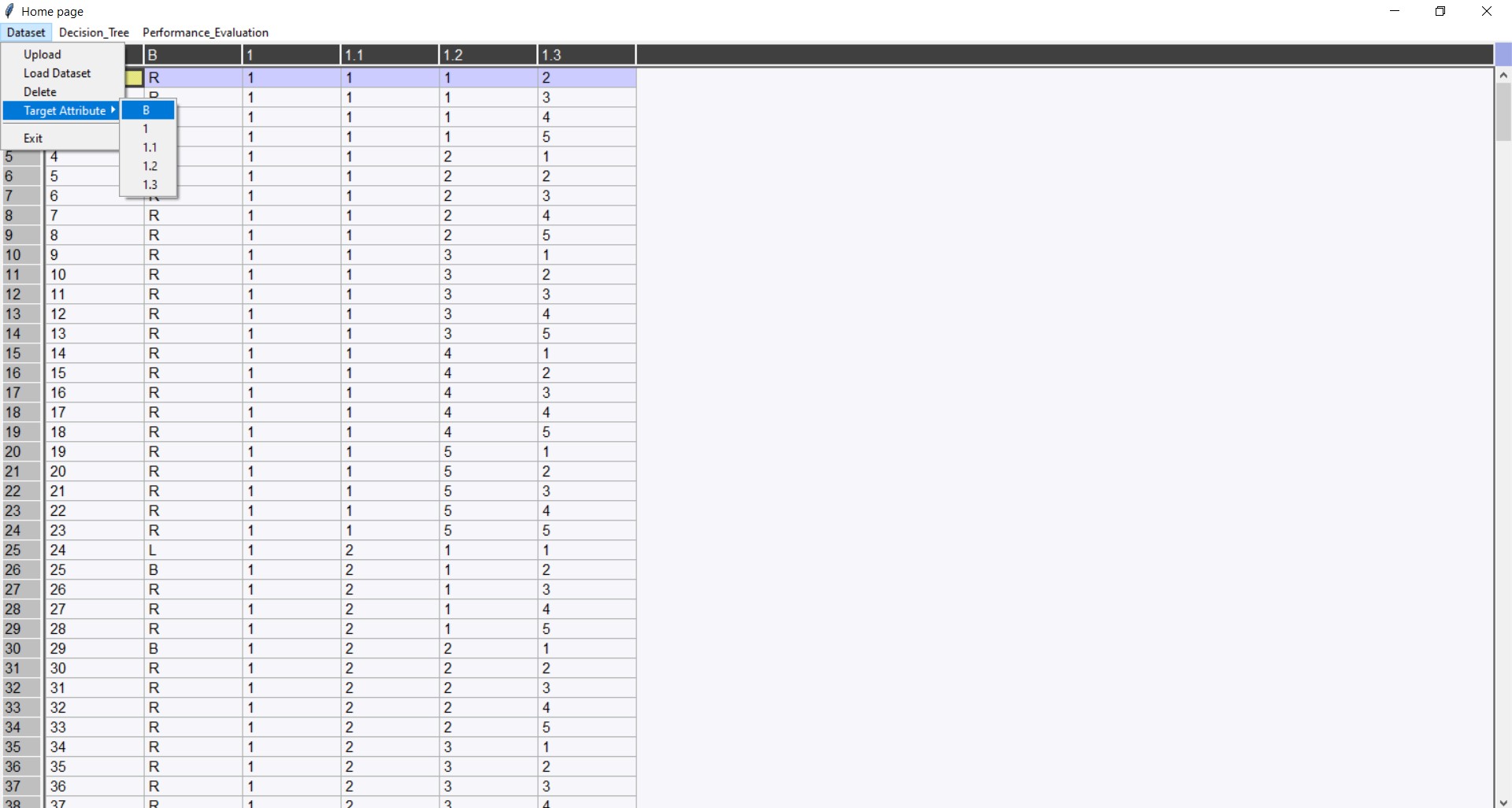
**Result/Observations/Screenshots:**

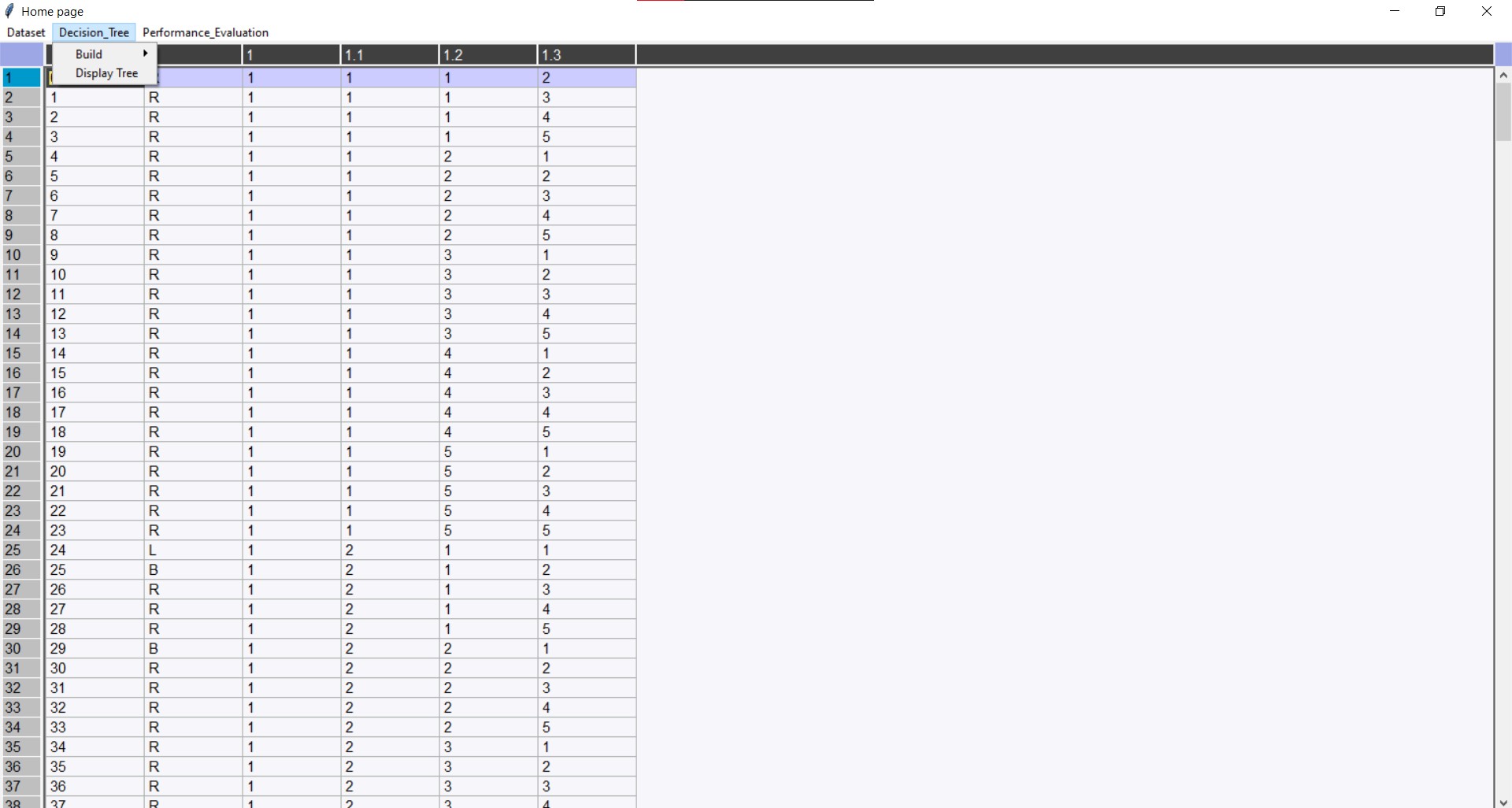
****

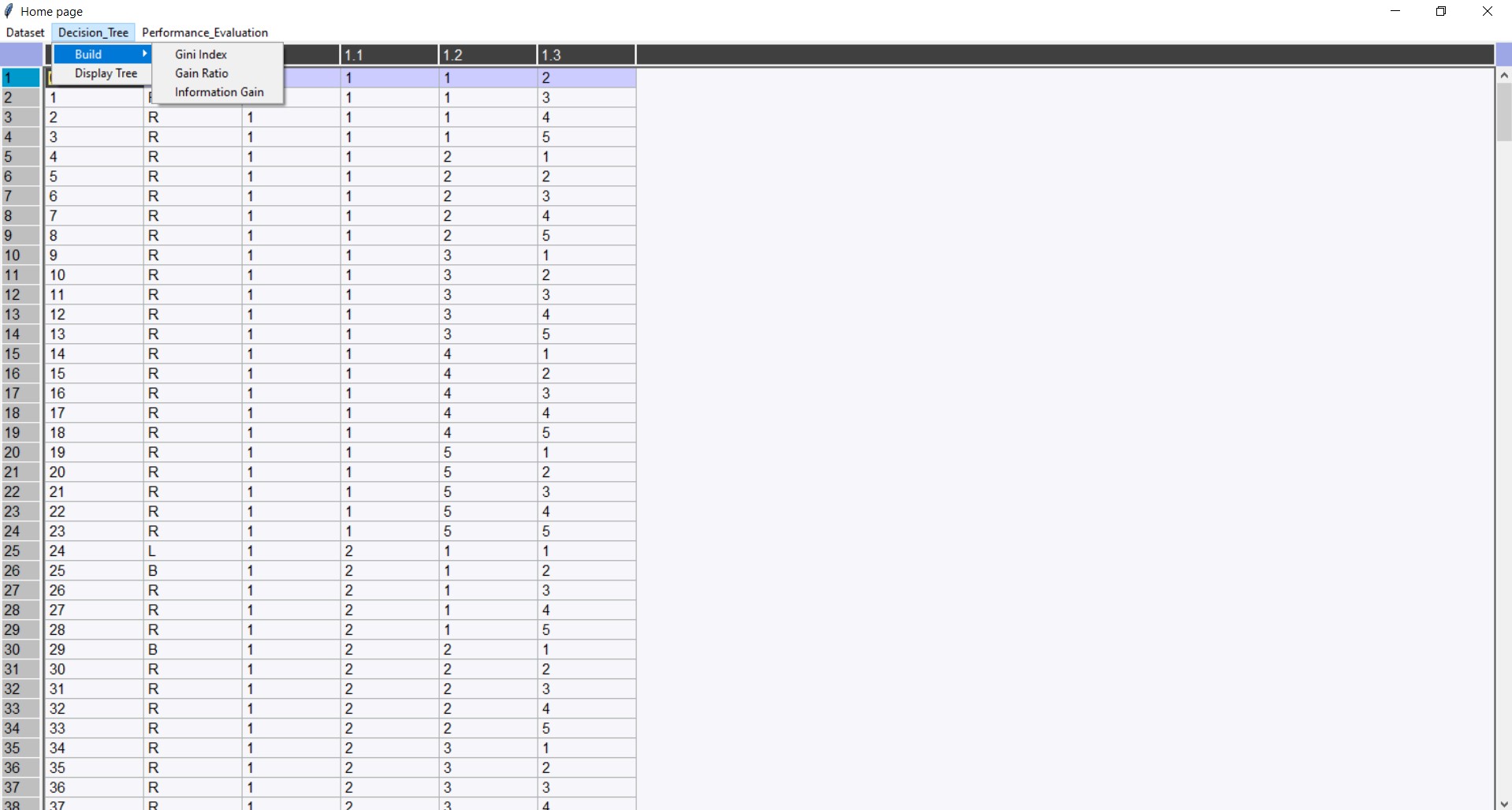
****

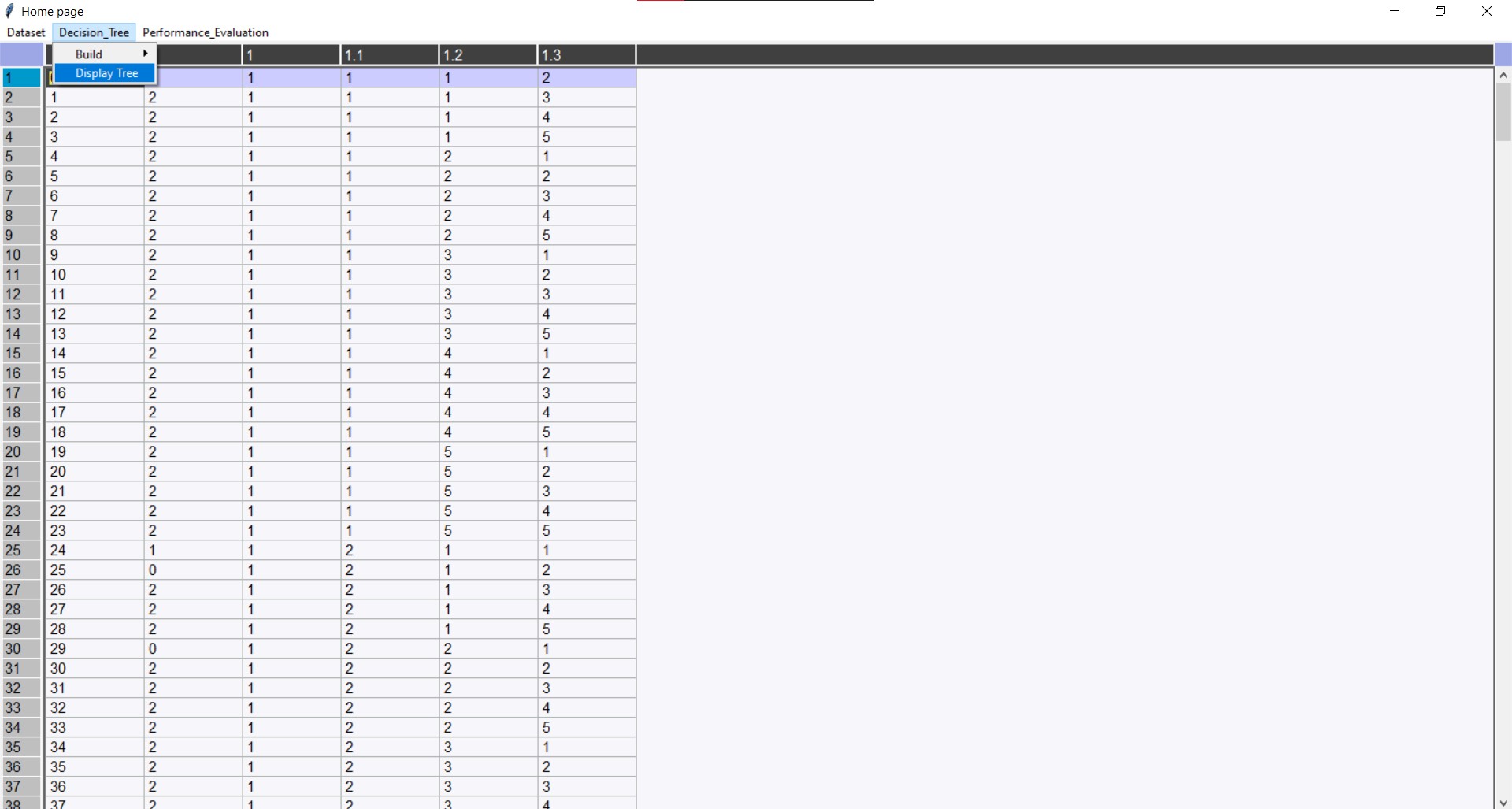
****

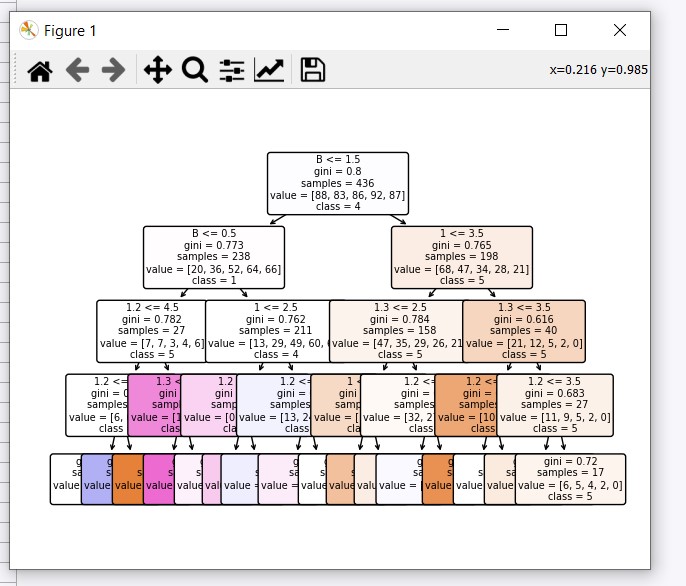
****

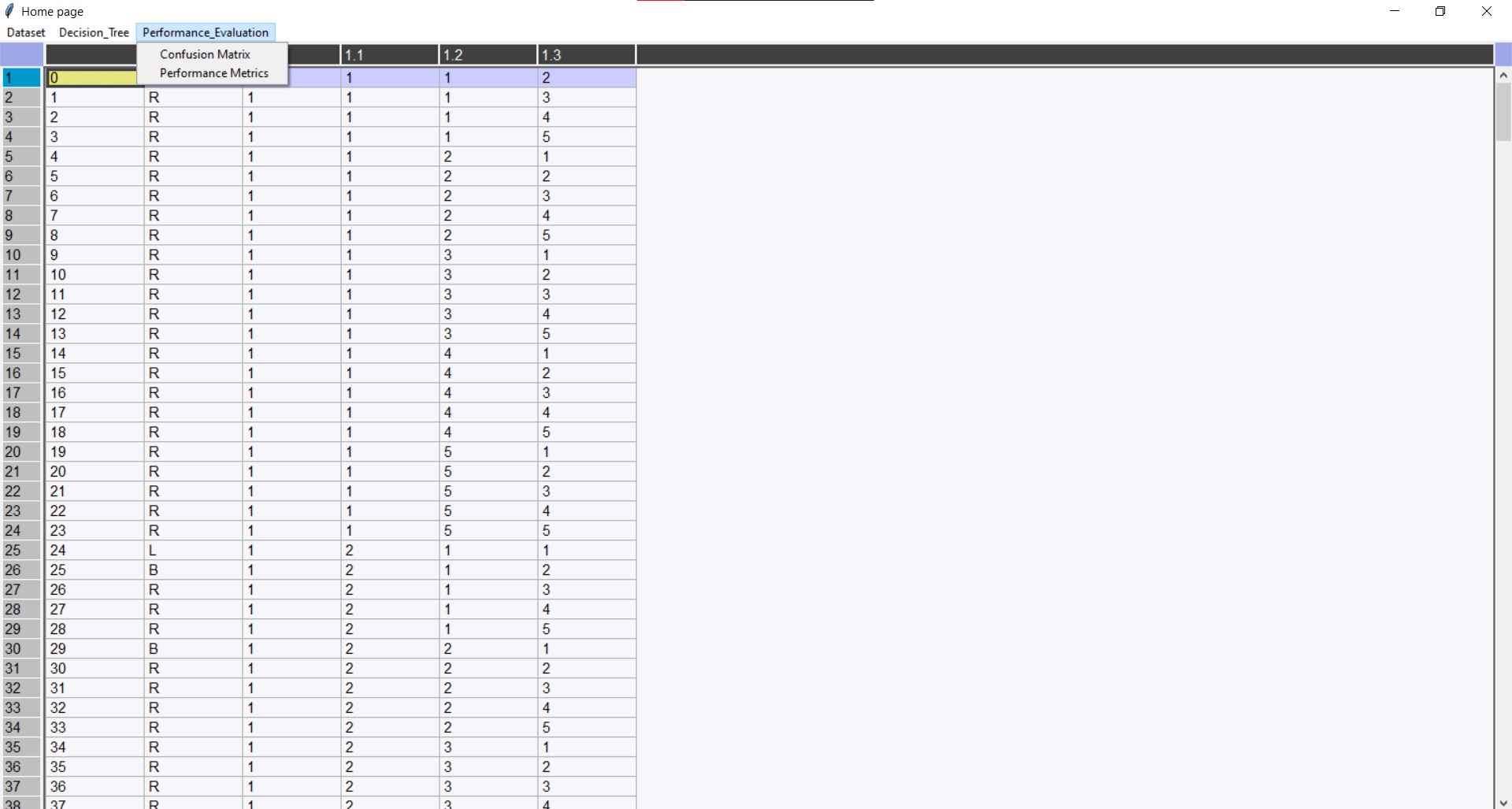
****

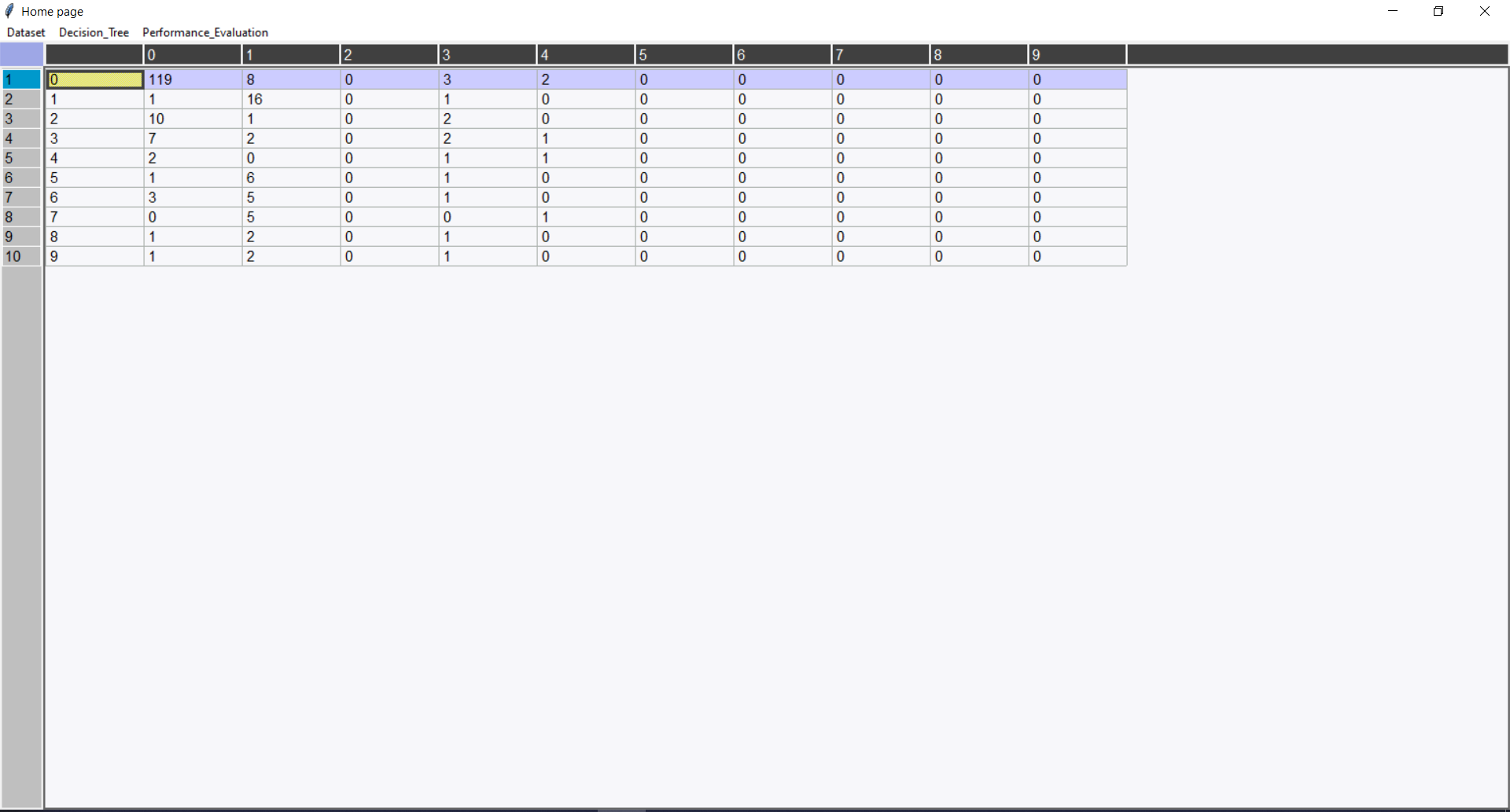
****

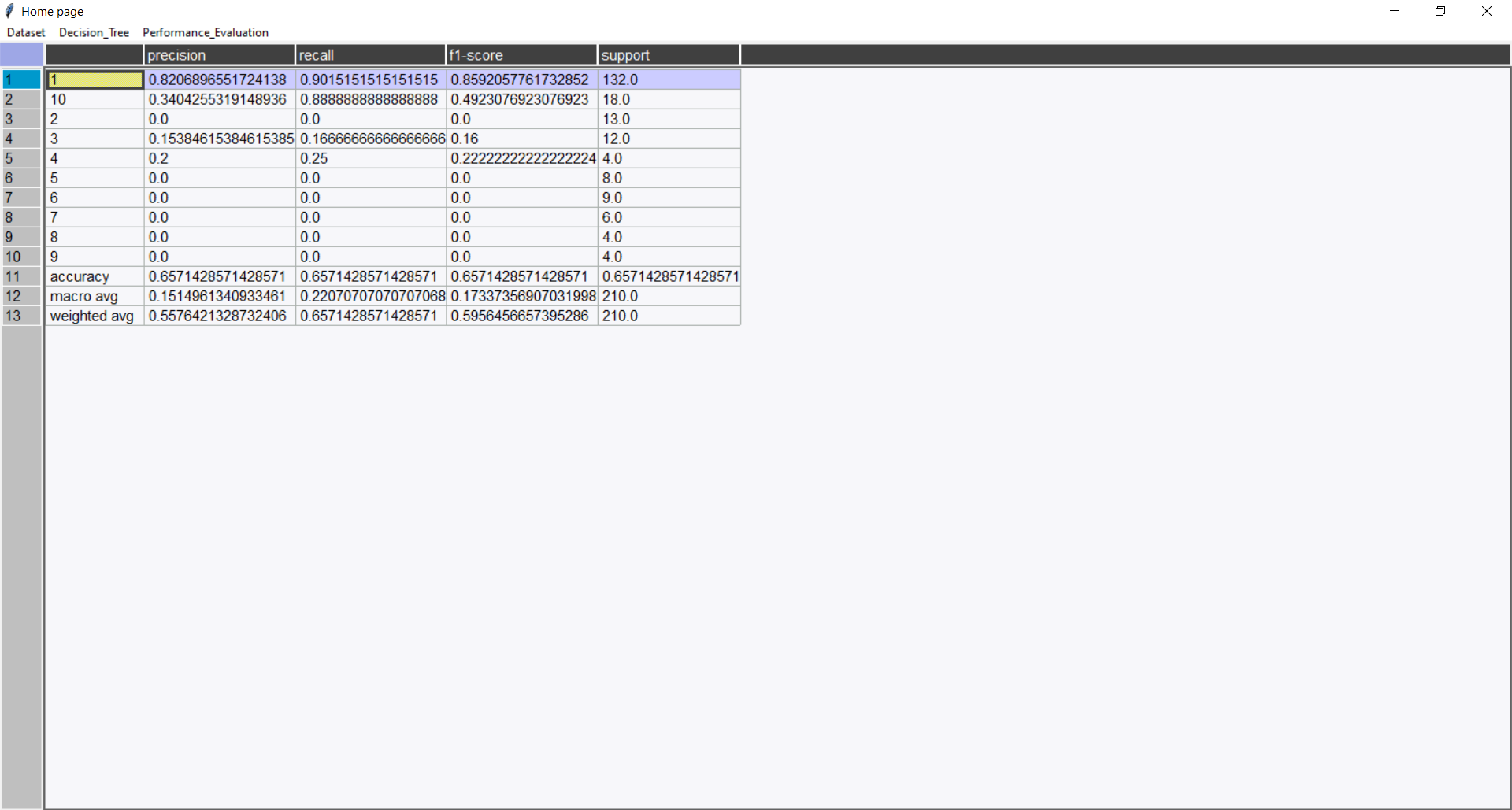
****

****

****

****

****

****

**References:**

* <https://victorzhou.com/blog/information-gain/>
* <https://www.analyticssteps.com/blogs/what-gini-index-and-information-gain-decision-trees>