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 :

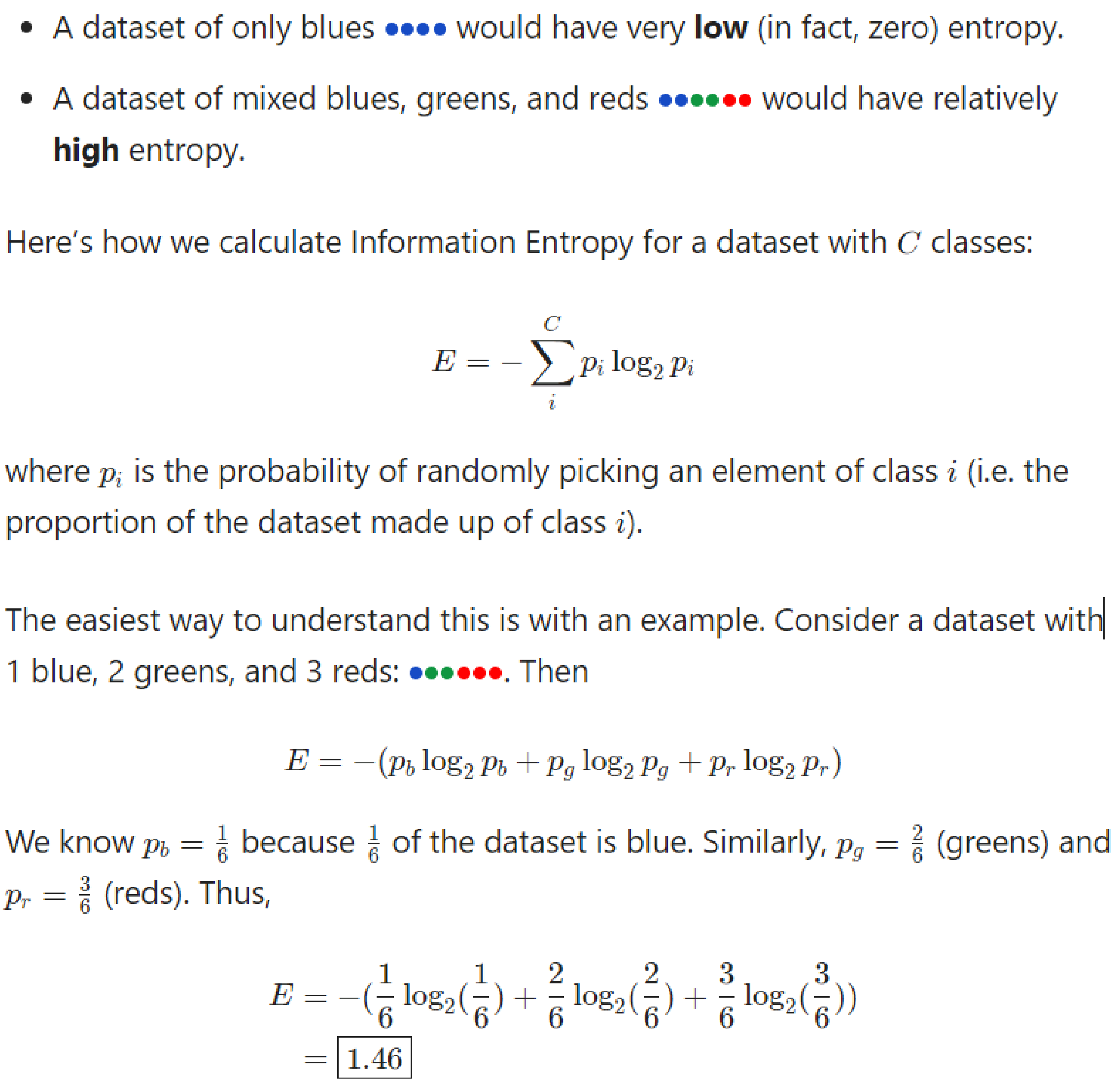
1. Implement the decision tree classifier using the following attribute selection measures and graphically show/visualize the tree:
    a. Information Gain
    b. Gain Ratio
    C. Gini Index
2. Tabulate the results in confusion matrix and evaluate the performance of above classifier using following 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:



- A dataset of only blues ●●●● would have very **low** (in fact, zero) entropy.
- A dataset of mixed blues, greens, and reds ●●●●●● would have relatively **high** entropy.

Here's how we calculate Information Entropy for a dataset with C classes:

$$E = -\sum_{i}^{C} p_i \log_2 p_i$$

where p_i is the probability of randomly picking an element of class i (i.e. the proportion of the dataset made up of class i).

The easiest way to understand this is with an example. Consider a dataset with 1 blue, 2 greens, and 3 reds: ●●●●●●. Then

$$E = -(p_b \log_2 p_b + p_g \log_2 p_g + p_r \log_2 p_r)$$

We know $p_b = \frac{1}{6}$ because $\frac{1}{6}$ of the dataset is blue. Similarly, $p_g = \frac{2}{6}$ (greens) and $p_r = \frac{3}{6}$ (reds). Thus,

$$\begin{aligned} E &= -(\frac{1}{6}\log_2(\frac{1}{6}) + \frac{2}{6}\log_2(\frac{2}{6}) + \frac{3}{6}\log_2(\frac{3}{6})) \\ &= \boxed{1.46} \end{aligned}$$

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 in 1984 as an impurity measure for decision tree learning and is given by the equation/formula;

$$Gini\ (P) = \sum_{i=1}^{n} p_i\,(1 - p_i\ ) = 1 - \sum_{i=1}^{n} (p_i)^2$$

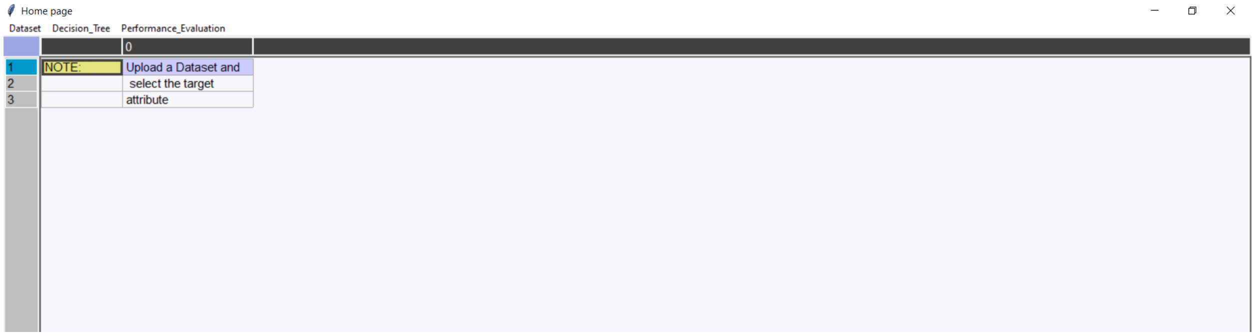
where P=(p1 , p2 ,.......pn) , and pi is the probability of an object that is being classified to a particular class.

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

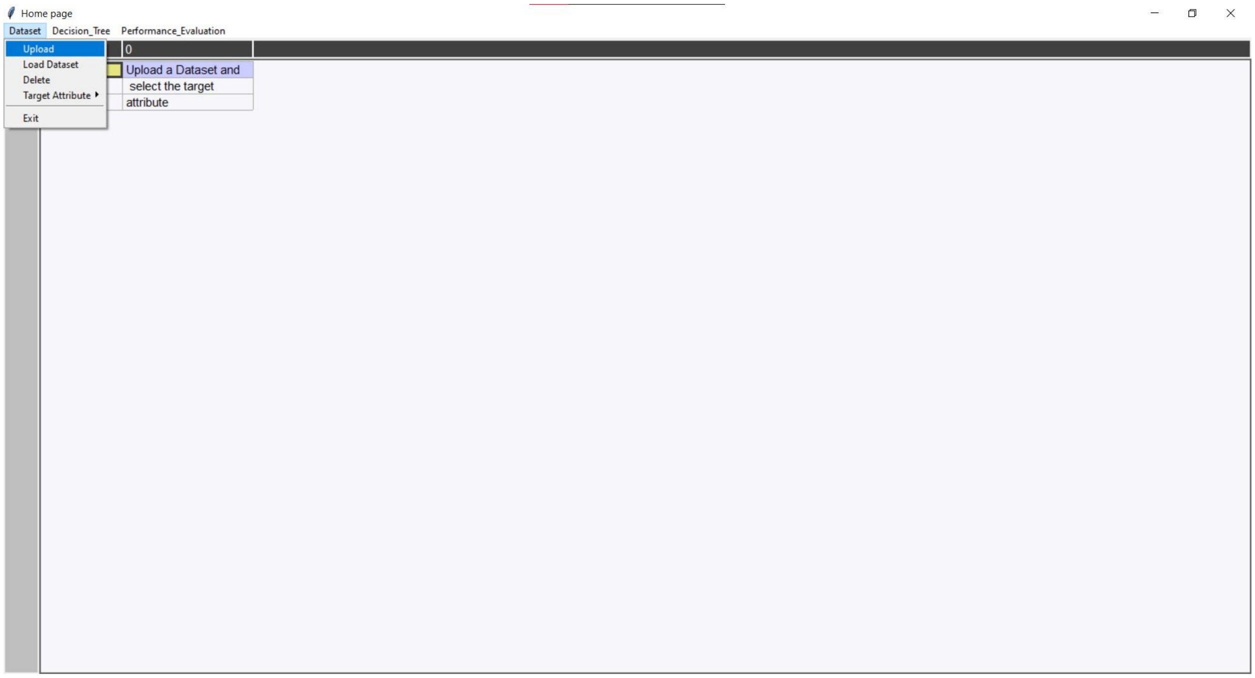
Also, an attribute/feature with least gini index is preferred as root node while making a decision tree.

Measure	*Formula*
accuracy, recognition rate	$\frac{TP+TN}{P+N}$
error rate, misclassification rate	$\frac{FP+FN}{P+N}$
sensitivity, true positive rate, recall	$\frac{TP}{P}$
specificity, true negative rate	$\frac{TN}{N}$
precision	$\frac{TP}{TP+FP}$
F, F_1, F-score, harmonic mean of precision and recall	$\frac{2 \times precision \times recall}{precision + recall}$
F_β, where β is a non-negative real number	$\frac{(1+\beta^2) \times precision \times recall}{\beta^2 \times precision + recall}$

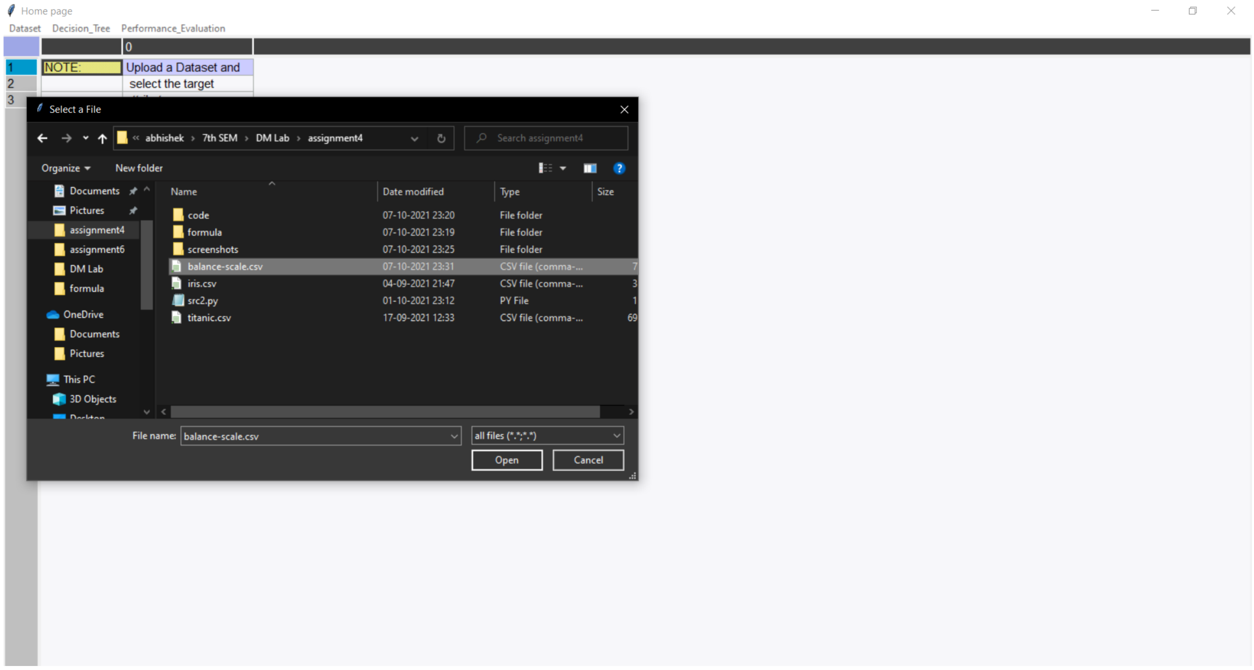
Result/Observations/Screenshots:



		0
1	NOTE:	Upload a Dataset and
2		select the target
3		attribute



Upload a Dataset and select the target attribute

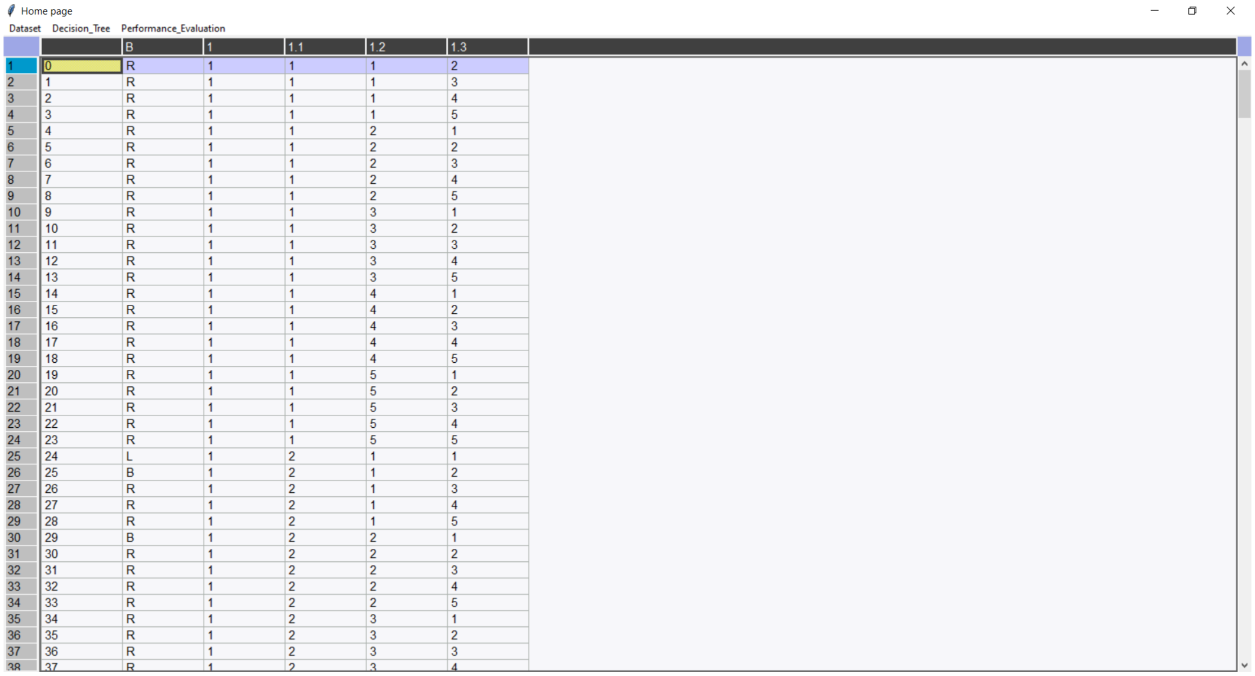


NOTE: Upload a Dataset and select the target

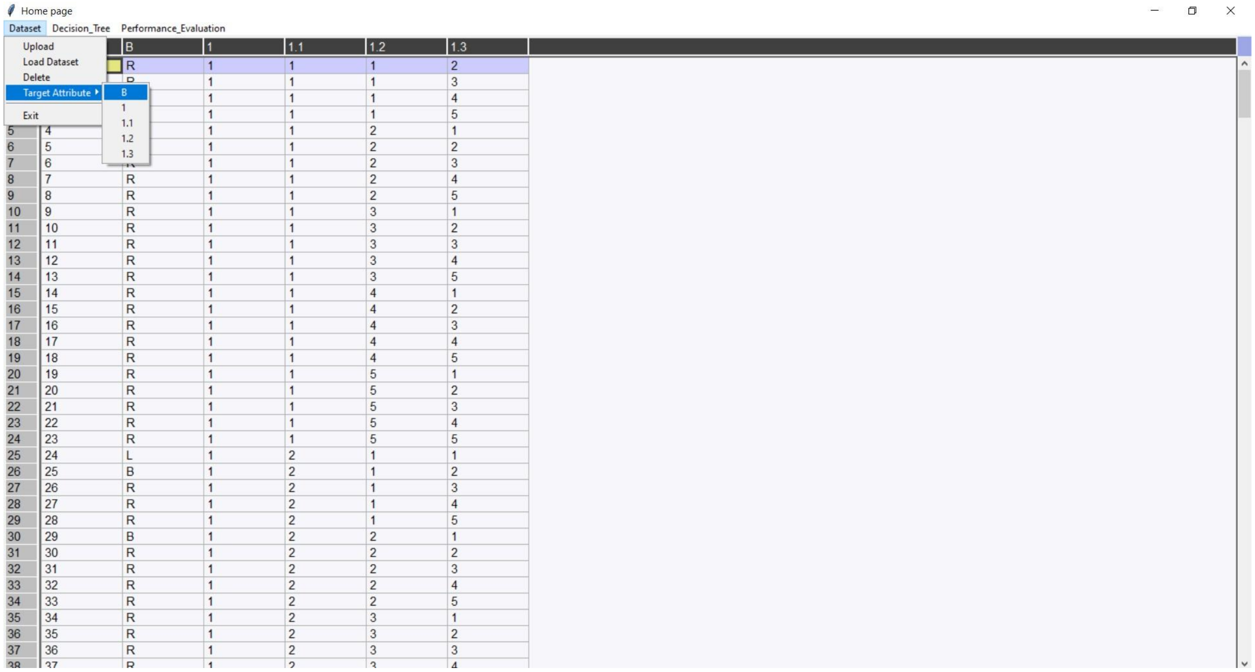
Name	Date modified	Type	Size
code	07-10-2021 23:20	File folder	
formula	07-10-2021 23:19	File folder	
screenshots	07-10-2021 23:25	File folder	
balance-scale.csv	07-10-2021 23:31	CSV file (comma-...	7
iris.csv	04-09-2021 21:47	CSV file (comma-...	3
src2.py	01-10-2021 23:12	PY File	1
titanic.csv	17-09-2021 12:33	CSV file (comma-...	69

File name: balance-scale.csv

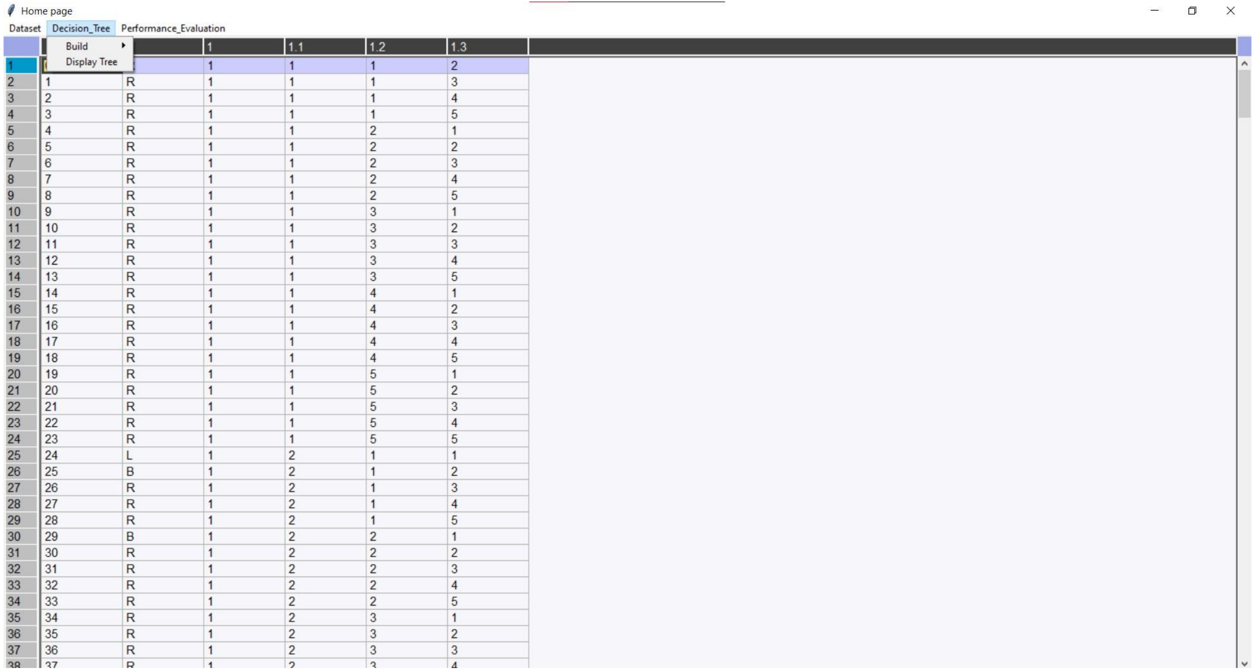
all files (*.*;*.*)



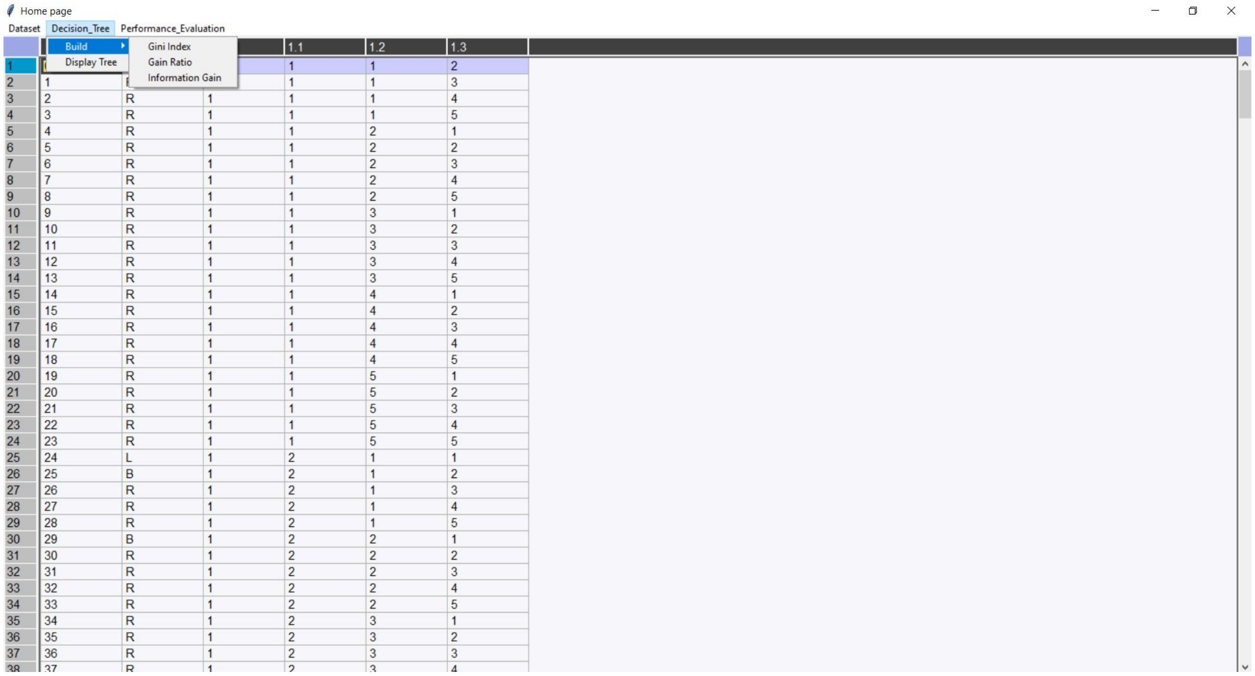
		B	1	1.1	1.2	1.3
1	0	R	1	1	1	2
2	1	R	1	1	1	3
3	2	R	1	1	1	4
4	3	R	1	1	1	5
5	4	R	1	1	2	1
6	5	R	1	1	2	2
7	6	R	1	1	2	3
8	7	R	1	1	2	4
9	8	R	1	1	2	5
10	9	R	1	1	3	1
11	10	R	1	1	3	2
12	11	R	1	1	3	3
13	12	R	1	1	3	4
14	13	R	1	1	3	5
15	14	R	1	1	4	1
16	15	R	1	1	4	2
17	16	R	1	1	4	3
18	17	R	1	1	4	4
19	18	R	1	1	4	5
20	19	R	1	1	5	1
21	20	R	1	1	5	2
22	21	R	1	1	5	3
23	22	R	1	1	5	4
24	23	R	1	1	5	5
25	24	L	1	2	1	1
26	25	B	1	2	1	2
27	26	R	1	2	1	3
28	27	R	1	2	1	4
29	28	R	1	2	1	5
30	29	B	1	2	2	1
31	30	R	1	2	2	2
32	31	R	1	2	2	3
33	32	R	1	2	2	4
34	33	R	1	2	2	5
35	34	R	1	2	3	1
36	35	R	1	2	3	2
37	36	R	1	2	3	3
38	37	R	1	2	3	4



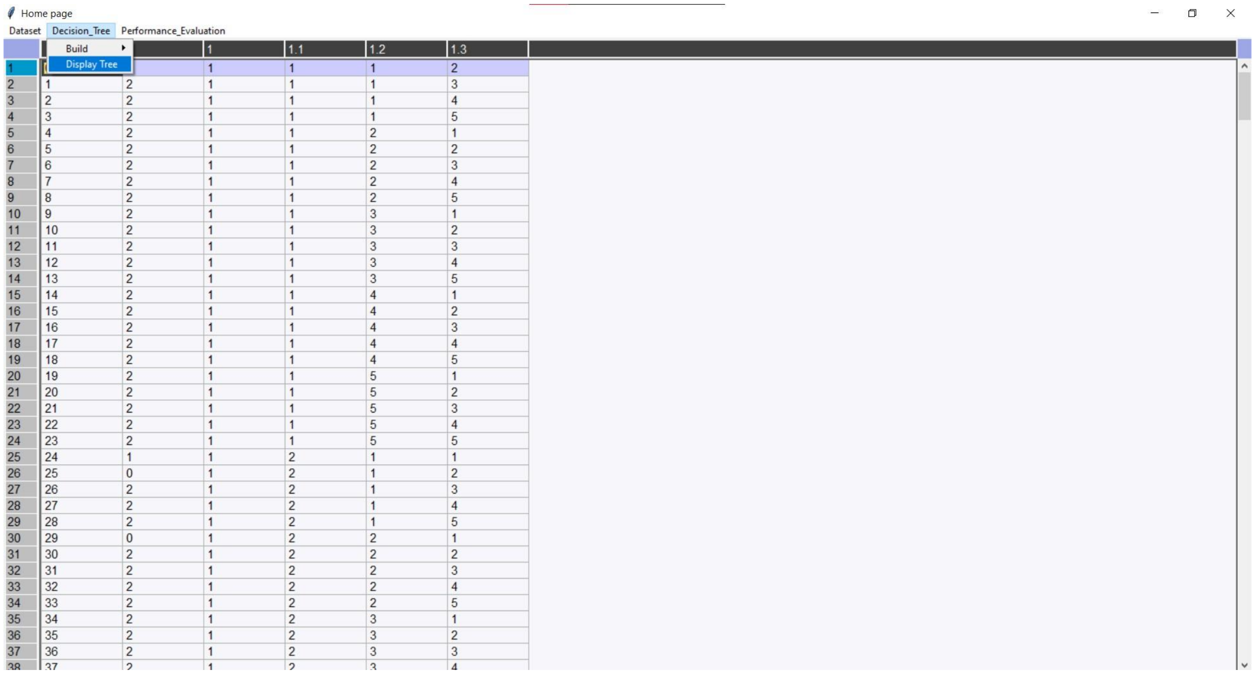
		B	1	1.1	1.2	1.3
		R	1	1	1	2
		R	1	1	1	3
			1	1	1	4
			1	1	1	5
5	4		1	1	2	1
6	5		1	1	2	2
7	6	R	1	1	2	3
8	7	R	1	1	2	4
9	8	R	1	1	2	5
10	9	R	1	1	3	1
11	10	R	1	1	3	2
12	11	R	1	1	3	3
13	12	R	1	1	3	4
14	13	R	1	1	3	5
15	14	R	1	1	4	1
16	15	R	1	1	4	2
17	16	R	1	1	4	3
18	17	R	1	1	4	4
19	18	R	1	1	4	5
20	19	R	1	1	5	1
21	20	R	1	1	5	2
22	21	R	1	1	5	3
23	22	R	1	1	5	4
24	23	R	1	1	5	5
25	24	L	1	2	1	1
26	25	B	1	2	1	2
27	26	R	1	2	1	3
28	27	R	1	2	1	4
29	28	R	1	2	1	5
30	29	B	1	2	2	1
31	30	R	1	2	2	2
32	31	R	1	2	2	3
33	32	R	1	2	2	4
34	33	R	1	2	2	5
35	34	R	1	2	3	1
36	35	R	1	2	3	2
37	36	R	1	2	3	3
38	37	R	1	2	3	4



			1	1.1	1.2	1.3
1			1	1	1	2
2	1	R	1	1	1	3
3	2	R	1	1	1	4
4	3	R	1	1	1	5
5	4	R	1	1	2	1
6	5	R	1	1	2	2
7	6	R	1	1	2	3
8	7	R	1	1	2	4
9	8	R	1	1	2	5
10	9	R	1	1	3	1
11	10	R	1	1	3	2
12	11	R	1	1	3	3
13	12	R	1	1	3	4
14	13	R	1	1	3	5
15	14	R	1	1	4	1
16	15	R	1	1	4	2
17	16	R	1	1	4	3
18	17	R	1	1	4	4
19	18	R	1	1	4	5
20	19	R	1	1	5	1
21	20	R	1	1	5	2
22	21	R	1	1	5	3
23	22	R	1	1	5	4
24	23	R	1	1	5	5
25	24	L	1	2	1	1
26	25	B	1	2	1	2
27	26	R	1	2	1	3
28	27	R	1	2	1	4
29	28	R	1	2	1	5
30	29	B	1	2	2	1
31	30	R	1	2	2	2
32	31	R	1	2	2	3
33	32	R	1	2	2	4
34	33	R	1	2	2	5
35	34	R	1	2	3	1
36	35	R	1	2	3	2
37	36	R	1	2	3	3
38	37	R	1	2	3	4



				1.1	1.2	1.3
1				1	1	2
2	1	R		1	1	3
3	2	R	1	1	1	4
4	3	R	1	1	1	5
5	4	R	1	1	2	1
6	5	R	1	1	2	2
7	6	R	1	1	2	3
8	7	R	1	1	2	4
9	8	R	1	1	2	5
10	9	R	1	1	3	1
11	10	R	1	1	3	2
12	11	R	1	1	3	3
13	12	R	1	1	3	4
14	13	R	1	1	3	5
15	14	R	1	1	4	1
16	15	R	1	1	4	2
17	16	R	1	1	4	3
18	17	R	1	1	4	4
19	18	R	1	1	4	5
20	19	R	1	1	5	1
21	20	R	1	1	5	2
22	21	R	1	1	5	3
23	22	R	1	1	5	4
24	23	R	1	1	5	5
25	24	L	1	2	1	1
26	25	B	1	2	1	2
27	26	R	1	2	1	3
28	27	R	1	2	1	4
29	28	R	1	2	1	5
30	29	B	1	2	2	1
31	30	R	1	2	2	2
32	31	R	1	2	2	3
33	32	R	1	2	2	4
34	33	R	1	2	2	5
35	34	R	1	2	3	1
36	35	R	1	2	3	2
37	36	R	1	2	3	3
38	37	R	1	2	3	4



			1	1.1	1.2	1.3
1			1	1	1	2
2	1	2	1	1	1	3
3	2	2	1	1	1	4
4	3	2	1	1	1	5
5	4	2	1	1	2	1
6	5	2	1	1	2	2
7	6	2	1	1	2	3
8	7	2	1	1	2	4
9	8	2	1	1	2	5
10	9	2	1	1	3	1
11	10	2	1	1	3	2
12	11	2	1	1	3	3
13	12	2	1	1	3	4
14	13	2	1	1	3	5
15	14	2	1	1	4	1
16	15	2	1	1	4	2
17	16	2	1	1	4	3
18	17	2	1	1	4	4
19	18	2	1	1	4	5
20	19	2	1	1	5	1
21	20	2	1	1	5	2
22	21	2	1	1	5	3
23	22	2	1	1	5	4
24	23	2	1	1	5	5
25	24	1	1	2	1	1
26	25	0	1	2	1	2
27	26	2	1	2	1	3
28	27	2	1	2	1	4
29	28	2	1	2	1	5
30	29	0	1	2	2	1
31	30	2	1	2	2	2
32	31	2	1	2	2	3
33	32	2	1	2	2	4
34	33	2	1	2	2	5
35	34	2	1	2	3	1
36	35	2	1	2	3	2
37	36	2	1	2	3	3
38	37	2	1	2	3	4

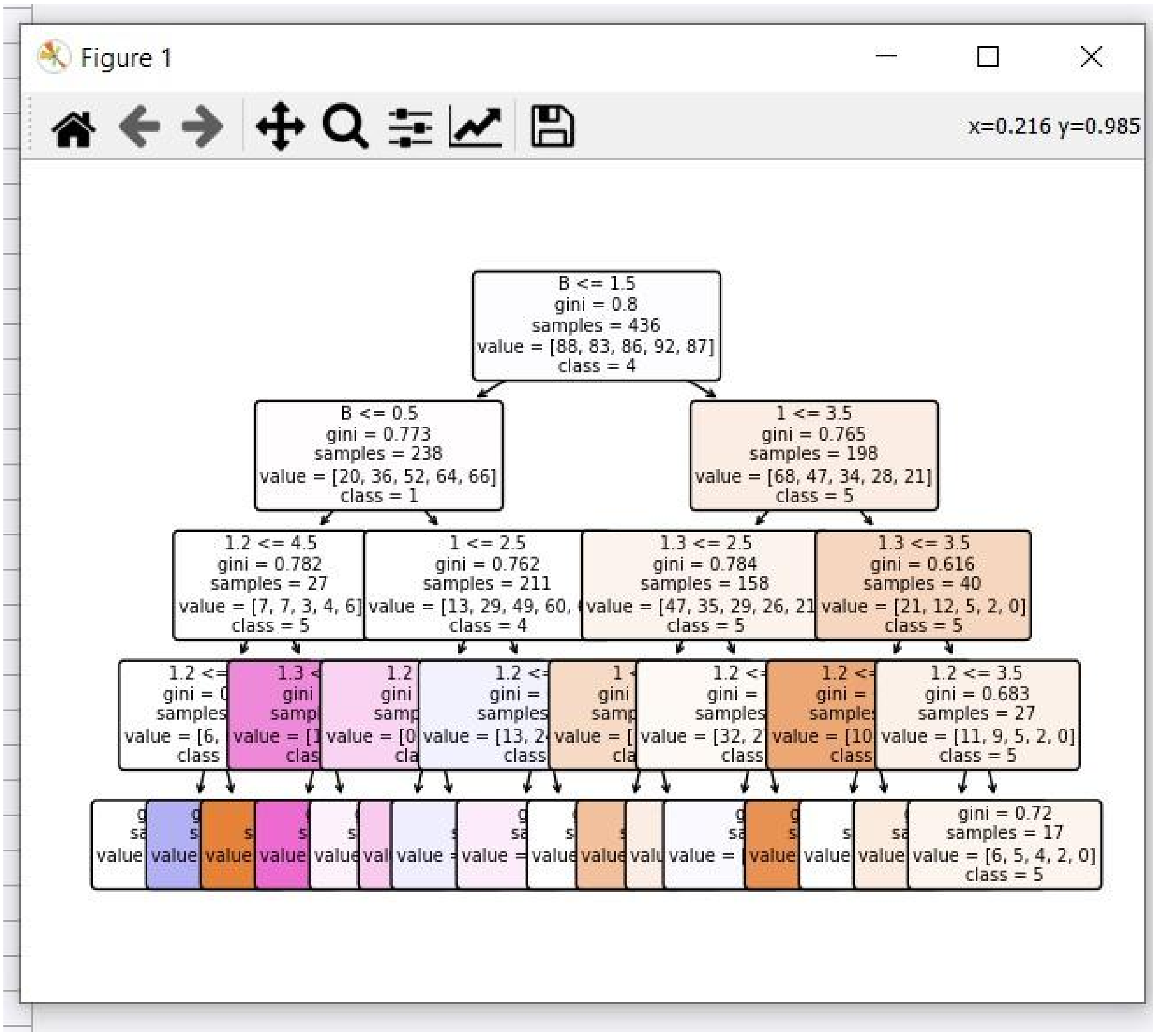
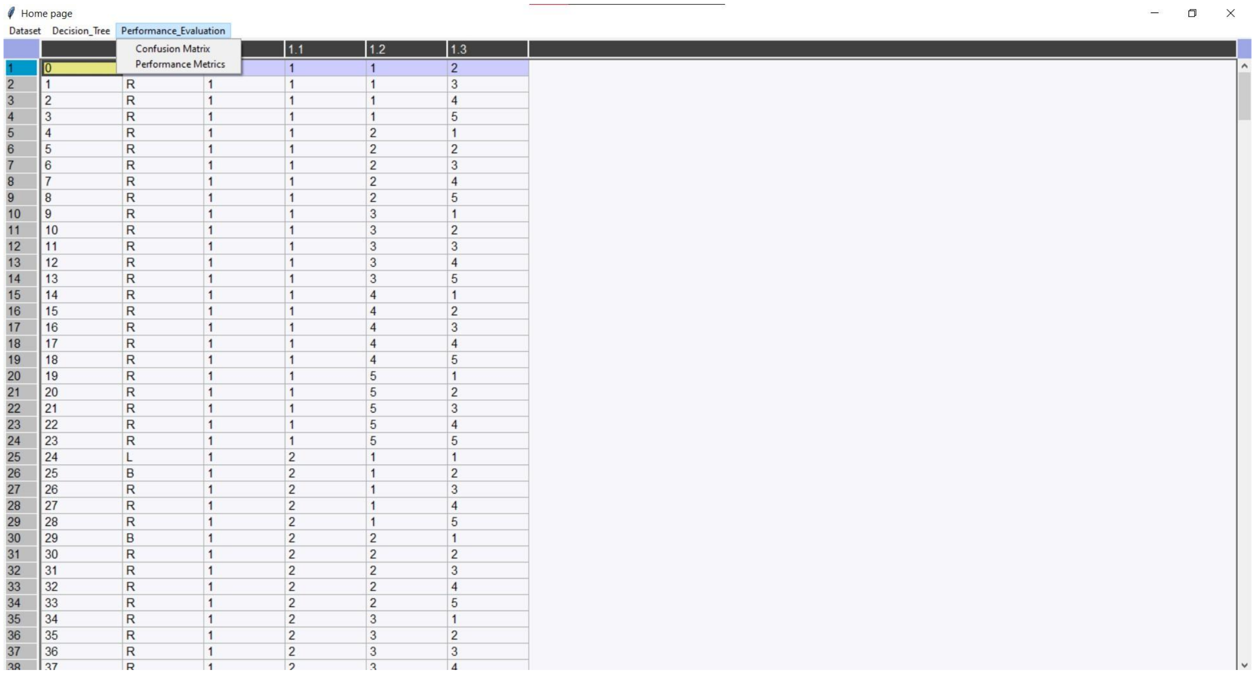
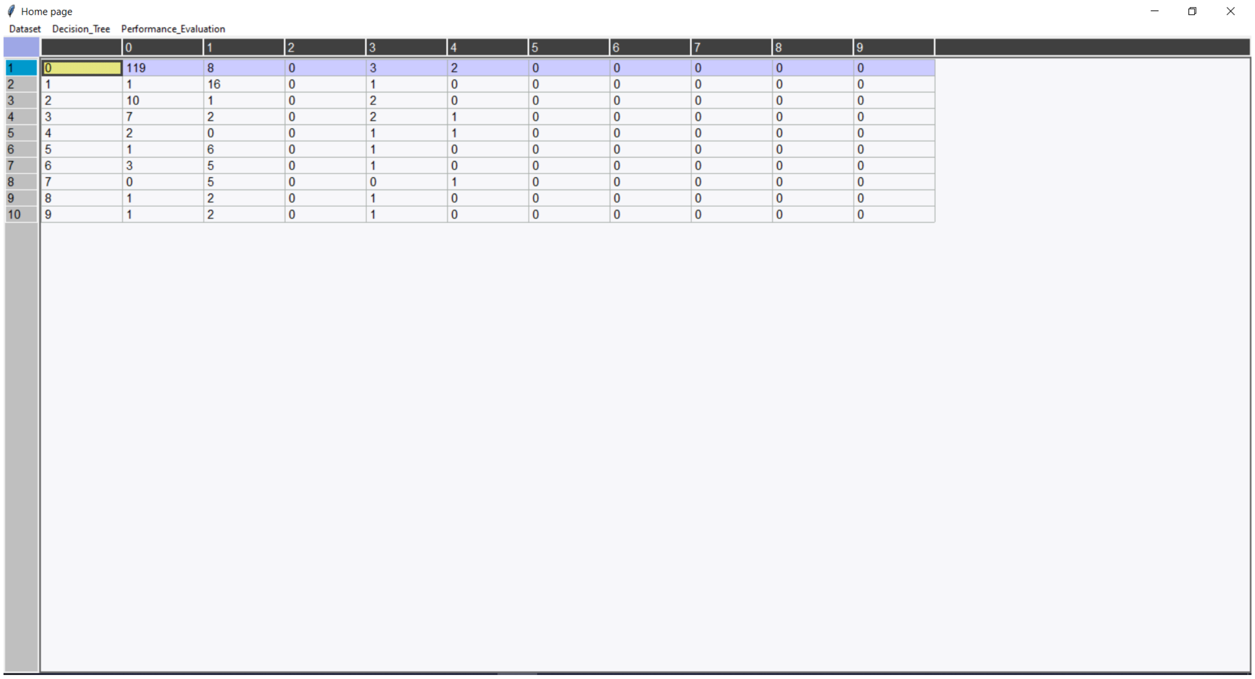


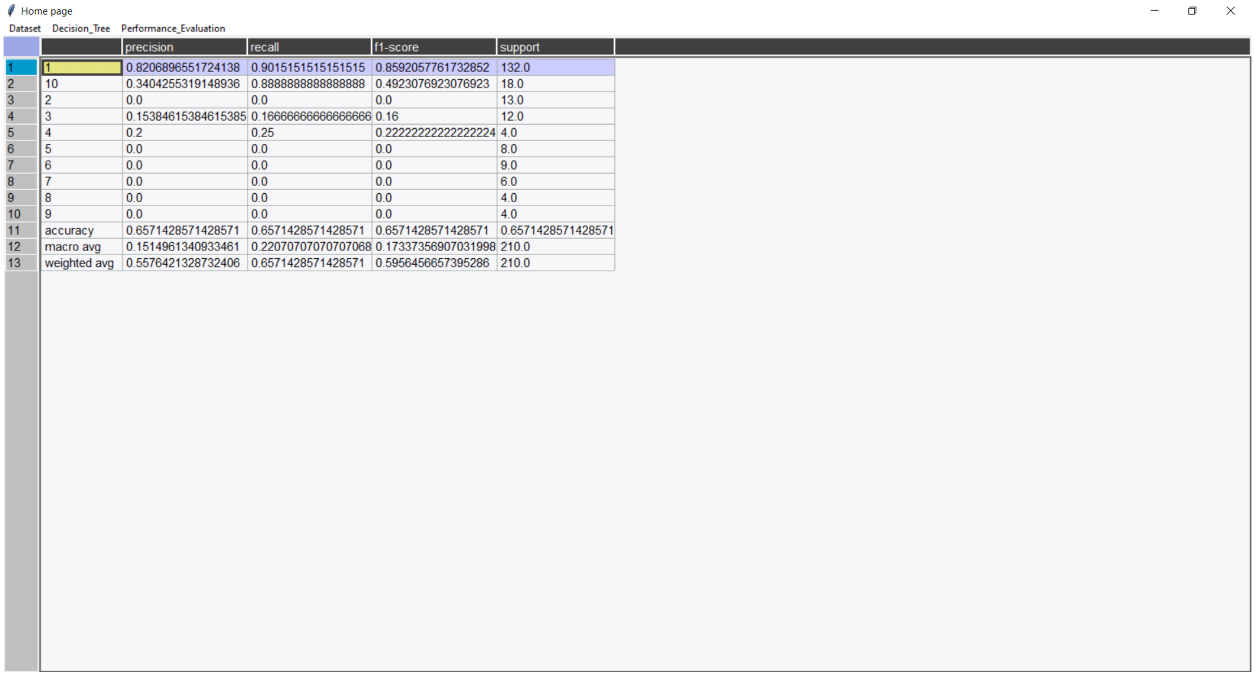
Figure 1
x=0.216 y=0.985
B <= 1.5
gini = 0.8
samples = 436
value = [88, 83, 86, 92, 87]
class = 4
B <= 0.5
gini = 0.773
samples = 238
value = [20, 36, 52, 64, 66]
class = 1
1 <= 3.5
gini = 0.765
samples = 198
value = [68, 47, 34, 28, 21]
class = 5
1.2 <= 4.5
gini = 0.782
samples = 27
value = [7, 7, 3, 4, 6]
class = 5
1 <= 2.5
gini = 0.762
samples = 211
value = [13, 29, 49, 60,
class = 4
1.3 <= 2.5
gini = 0.784
samples = 158
value = [47, 35, 29, 26, 21
class = 5
1.3 <= 3.5
gini = 0.616
samples = 40
value = [21, 12, 5, 2, 0]
class = 5
1.2 <= 3.5
gini = 0.683
samples = 27
value = [11, 9, 5, 2, 0]
class = 5
gini = 0.72
samples = 17
value = [6, 5, 4, 2, 0]
class = 5



				1.1	1.2	1.3
1	0			1	1	2
2	1	R	1	1	1	3
3	2	R	1	1	1	4
4	3	R	1	1	1	5
5	4	R	1	1	2	1
6	5	R	1	1	2	2
7	6	R	1	1	2	3
8	7	R	1	1	2	4
9	8	R	1	1	2	5
10	9	R	1	1	3	1
11	10	R	1	1	3	2
12	11	R	1	1	3	3
13	12	R	1	1	3	4
14	13	R	1	1	3	5
15	14	R	1	1	4	1
16	15	R	1	1	4	2
17	16	R	1	1	4	3
18	17	R	1	1	4	4
19	18	R	1	1	4	5
20	19	R	1	1	5	1
21	20	R	1	1	5	2
22	21	R	1	1	5	3
23	22	R	1	1	5	4
24	23	R	1	1	5	5
25	24	L	1	2	1	1
26	25	B	1	2	1	2
27	26	R	1	2	1	3
28	27	R	1	2	1	4
29	28	R	1	2	1	5
30	29	B	1	2	2	1
31	30	R	1	2	2	2
32	31	R	1	2	2	3
33	32	R	1	2	2	4
34	33	R	1	2	2	5
35	34	R	1	2	3	1
36	35	R	1	2	3	2
37	36	R	1	2	3	3
38	37	R	1	2	3	4



		0	1	2	3	4	5	6	7	8	9
1	0	119	8	0	3	2	0	0	0	0	0
2	1	1	16	0	1	0	0	0	0	0	0
3	2	10	1	0	2	0	0	0	0	0	0
4	3	7	2	0	2	1	0	0	0	0	0
5	4	2	0	0	1	1	0	0	0	0	0
6	5	1	6	0	1	0	0	0	0	0	0
7	6	3	5	0	1	0	0	0	0	0	0
8	7	0	5	0	0	1	0	0	0	0	0
9	8	1	2	0	1	0	0	0	0	0	0
10	9	1	2	0	1	0	0	0	0	0	0



		precision	recall	f1-score	support
1	1	0.8206896551724138	0.9015151515151515	0.8592057761732852	132.0
2	10	0.3404255319148936	0.8888888888888888	0.4923076923076923	18.0
3	2	0.0	0.0	0.0	13.0
4	3	0.15384615384615385	0.16666666666666666	0.16	12.0
5	4	0.2	0.25	0.22222222222222224	4.0
6	5	0.0	0.0	0.0	8.0
7	6	0.0	0.0	0.0	9.0
8	7	0.0	0.0	0.0	6.0
9	8	0.0	0.0	0.0	4.0
10	9	0.0	0.0	0.0	4.0
11	accuracy	0.6571428571428571	0.6571428571428571	0.6571428571428571	0.6571428571428571
12	macro avg	0.1514961340933461	0.22070707070707068	0.17337356907031998	210.0
13	weighted avg	0.5576421328732406	0.6571428571428571	0.5956456657395286	210.0

References:

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