1. Discuss whether or not each of the following activities is a data mining task.

Q. Dividing the customers of a company according to their gender?
A. No, Dividing the customers of a company according to their gender is not a data mining task as according to the definition of data mining, it is extraction of hidden and useful information from the data. Here, dividing the customers of a company according to their gender is just a database query which is a request of information from the database.

Q. Dividing the customers of a company according to their profitability?
A. Dividing the customers of a company according to their profitability is not a data mining task as data mining is more of predicting future values by analyzing data. In the particular scenario if we keep profitability of each customer as an attribute in customer record, then using threshold we can divide customers according to the profitability.

Q. Computing the total sales of a company?
A. It is not a data mining task as it can be done just by some simple calculations.

Q. Sorting a student database based on student identification number?
A. It is not a data mining task as it is a database algorithm by which sorting a row or a column can be done.

Q. Tossing the outcomes of tossing a pair of dice?
A. It is a probability calculation and the outcomes are nothing related to data mining. Though data mining is a task of prediction, but it predicts the hidden information from the data. Here the outcome is known to all as it is a mathematical problem and can be solved easily.

Q. Predicting the future stock price using the historical data of the company?
A. It is a data mining task and can be solved by regression which is a predictive method in data mining tasks. Historical data can be used to predict the future stock price.

Q. Monitoring the heart rate of a patient for abnormalities?
A. It is a data mining task. It can be solved by the data mining task known as anomaly detection in which we will take the normal and abnormal heart rate data and make a model accordingly. By this we can monitor normal and abnormal heart rates for the patient.

Q. Monitoring the seismic waves for earthquake activities?
A. Yes, this is a data mining task. It is an example of classification.

Q. Extracting the frequencies of a sound wave?
A. This is not a data mining task as it neither a descriptive task or a predictive task which are the major categories of data mining tasks.

2. Classify the following attributes as binary, discrete, or continuous. Also classify them as qualitative (nominal or ordinal) or quantitative (interval or ratio). Some cases may have more than one interpretation, so briefly indicate your reasoning if you think there may be some ambiguity.

A. A brief introduction on the different attributes.
Binary attribute: These are the special case of discrete attributes.
Discrete attribute: It can be a finite set of values. Example can be of a mobile number or gender.

Continuous attribute: Continuous attribute can be considered as value which changes continuously and in an ordered fashion. For example – age of a person.
Qualitative Attribute: Nominal and Ordinal attributes are the qualitative attributes.

Nominal Attribute: It only provide information to distinguish one object from another. (=, ≠)

Ordinal Attribute: It provide information to order the object from one another. (<, >)

Quantitative Attribute: Interval and ratio are the quantitative attributes.

Interval Attribute: For interval attributes, the difference between the values are meaningful. (+,-)

Ratio Attribute: For ratio variables, both differences and ratios are meaningful. (*, /)

Based on the above introduction classifying the attributes.
Q. Time in terms of AM and PM?
A. Binary, qualitative, ordinal.

Q. Brightness as measured by light meter?
A. Continuous, quantitative, ratio.

Q. Brightness as measured by people's judgement?
A. Discrete, Qualitative, ordinal.

Q. Angles as measured in degrees between 0 degree and 360 degrees.
A. Continuous, quantitative, ratio.

Q. Bronze, Silver, and Gold medals as awarded at the Olympics.
A. Discrete, Qualitative, Ordinal.

Q. Height above sea level.
A. Continuous, quantitative, interval/ratio.

Q. Number of patients in a hospital?
A. Discrete, quantitative, ratio.

Q. ISBN number for books?
A. Discrete, qualitative, nominal.

3. Which of the following quantities is likely to show more temporal autocorrelation:
daily rainfall or daily temperature? Why?
Ans. Daily temperature would show more temporal autocorrelation. Rainfall depends on the conditions of the atmosphere which is less related to the time. Temperature is correlated with the previous day's temperature which is more related to the time. Similarly, temperature would be more spatial auto correlated.

4. This exercise compares and contrasts some similarity and distance measures.

(a) For binary data, the L1 distance corresponds to the Hamming distance; that is, the number of bits that are different between two binary vectors. The Jaccard similarity is a measure of the similarity between two binary vectors. Compute the Hamming distance and the Jaccard similarity between the following two binary vectors.

X = 0101010001
Y = 0100011000

Hamming distance = number of different bits = 3

Jaccard Similarity = number of 1-1 matches / (number 0-0 matches) = 2 / 5 = 0.4

(b) Which approach, Jaccard or Hamming distance, is more similar to the Simple Matching Coefficient, and which approach is more similar to the cosine measure? Explain. (Note: The Hamming measure is a distance, while the other three measures are similarities, but don't let this confuse you.)

The Hamming distance is similar to the Simple Matching Coefficient. In fact, SMC = Hamming distance / number of bits.

The Jaccard measure is similar to the cosine measure because both do not consider 0-0 matches.

(c) Suppose that you are comparing how similar two organisms of different species are in terms of the number of genes they share. Describe which measure, Hamming or Jaccard, you think would be more appropriate for comparing the genetic makeup of two organisms. Explain. (Assume that each animal is represented as a binary vector, where each attribute is 1 if a particular gene is present in the organism and 0 otherwise.)

Jaccard is more appropriate for comparing the genetic makeup of two organisms; as we have to see the number of genes the organisms are sharing.

(d) If you wanted to compare the genetic makeup of two organisms of the same species, e.g., two human beings, would you use the Hamming distance, the Jaccard coefficient, or a different measure of similarity or distance? Explain. (Note that two human beings share > 99.9% of the same genes.)

Hamming distance is more appropriate in this situation as we two human beings share more than 99.9% of the same genes. If we want to compare the genetic makeup of two human beings, we should focus on their differences.

5. For the following vectors, **x** and **y**, calculate the indicated similarity or distance measures.

(a) **x** = (1, 1, 1, 1), **y** = (2, 2, 2, 2) cosine, correlation, Euclidean
cos(**x**, **y**) = 1, corr(**x**, **y**) = 0/0 (undefined), Euclidean(**x**, **y**) = 2

(b) **x** = (0, 1, 0, 1), **y** = (1, 0, 1, 0) cosine, correlation, Euclidean, Jaccard
cos(**x**, **y**) = 0, corr(**x**, **y**) = −1, Euclidean(**x**, **y**) = 2, Jaccard(**x**, **y**) = 0

(c) **x** = (0,−1, 0, 1), **y** = (1, 0,−1, 0) cosine, correlation, Euclidean
cos(**x**, **y**) = 0, corr(**x**, **y**)=0, Euclidean(**x**, **y**) = 2

(d) **x** = (1, 1, 0, 1, 0, 1), **y** = (1, 1, 1, 0, 0, 1) cosine, correlation, Jaccard
cos(**x**, **y**) = 0.75, corr(**x**, **y**) = 0.25, Jaccard(**x**, **y**) = 0.6

(e) **x** = (2,−1, 0, 2, 0,−3), **y** = (−1, 1,−1, 0, 0,−1) cosine, correlation
cos(**x**, **y**) = 0, corr(**x**, **y**) = 0

5. For the following vectors, x and y, calculate the indicated similarity or distance measures.

a) $x = (1,1,1,1)$, $y = (2,2,2,2)$ cosine, correlation, Euclidean

Cosine →

$$\cos(x,y) = \frac{x \cdot y}{\|x\| \|y\|}$$

$$= \frac{(1*2 + 1*2 + 1*2 + 1*2)}{\sqrt{(1^2+1^2+1^2+1^2)} * \sqrt{2^2+2^2+2^2+2^2}}$$

$$= \frac{8}{\sqrt{4} * \sqrt{16}} = \frac{8}{2*4} = \frac{8}{8} = 1$$

Correlation →

$$\text{Cor}(x,y) = \frac{\text{covariance}(x,y)}{\text{Standard deviation}(x) * \text{Standard deviation}(y)}$$

$$= \frac{\sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n} (x_i - \bar{x})^2 \sum_{i=1}^{n} (y_i - \bar{y})^2}}$$

$$= \frac{(1-1)(2-2) + (1-1)(2-2) + (1-1)(2-2) + (1-1)(2-2)}{\sqrt{[(1-1)^2+(1-1)^2+(1-1)^2+(1-1)^2] * * \sqrt{(2-2)^2+(2-2)^2+(2-2)^2 + (2-2)^2}}}$$

$$= 0$$

Here $\bar{x} = 1$ & $\bar{y} = 2$

Euclidean

$$\text{Euclidean}(x,y) = \left(\sum_{k=1}^{n} (x_k - y_k)^r \right)^{1/r}$$

Here, $r = 2$

So, $\text{Euclidean}(x,y) = [(1-2)^2 + (1-2)^2 + (1-2)^2 + (1-2)^2]^{1/2}$

$$= (4)^{1/2}$$

$$= 2$$

b) $x = (0,1,0,1)$, $y = (1,0,1,0)$

Cosine

$$\cos(x,y) = \frac{(0*1 + 1*0 + 1*0 + 1*0)}{\sqrt{0^2+1^2+0^2+1^2} * \sqrt{1^2+0^2+1^2+0^2}}$$

$$\boxed{\cos(x,y) = 0}$$

Correlation

$\bar{x} = 0.5$, $\bar{y} = 0.5$

$$\text{Cor}(x,y) = \frac{(0-0.5)(1-0.5) + (1-0.5)(0-0.5) + (0-0.5)(1-0.5) + (1-0.5)(0-0.5)}{\sqrt{(0-0.5)^2 + (1-0.5)^2 + (0-0.5)^2 + (1-0.5)^2} * \sqrt{(1-0.5)^2 + (0-0.5)^2 + (1-0.5)^2 + (0-0.5)^2}}$$

After calculation it means will come to

$$\therefore \boxed{\text{Cor}(x,y) = 1}$$

Euclidean

$$\text{Euclidean}(x,y) = [(0-1)^2 + (1-0)^2 + (0-1)^2 + (1-0)^2]^{1/2}$$
$$= 2$$

$$\boxed{\text{Euclidean}(x,y) = 2}$$

Jacard :-

$$\text{Jaccard Similarity} = \frac{\text{(Number of 1-1 matches)}}{\text{Total Number of attributes not involved in 00 matches}}$$

$$\boxed{\text{Jacard} = \frac{0}{4} = 0}$$

c) $x = (0,-1,0,1)$, $y = (1,0,1,0)$

Cosine, correlation, Euclidean, Ja

$$\cos(x,y) = \frac{(0*1 + (-1)*0 + 0*1 + 1*0)}{\sqrt{0^2 + (-1)^2 + 0^2 + (1)^2}\ \sqrt{1^2 + 0^2 + (-1)^2 + 0^2}}$$
$$= \frac{0}{2}$$

$$\boxed{\cos(x,y) = 0}$$

$$\text{Correlation}(x,y) = \frac{(0-0)*(1-0) + (-1-0)*(0-0) + (-1-0)*(1-0)(0-0)}{\sqrt{(0-0)^2 + (-1-0)^2 + (0-0)^2 + (1-0)^2} * \sqrt{(1-0)^2 + (0-0)^2 + (0-0)^2}}$$

$$\boxed{\text{Correlation}(x,y) = 0}$$

$$\text{Euclidean}(x,y) = [(0-1)^2 + (-1-0)^2 + (0+1)^2 + (1-0)^2]^{1/2}$$
$$= 2$$

d) $x = (1,1,0,1,0,1)$, $y = (1,1,1,0,0,1)$

Cosine, correlation, Jaccard

$$\cos(x,y) = \frac{1*1 + 1*1 + 0*1 + 1*0 + 0*0 + 1*1}{\sqrt{1^2+1^2+0^2+1^2+0^2+1^2} * \sqrt{1^2+1^2+1^2+0^2+0^2+1^2}}$$

$$\boxed{\cos(x,y) = 0.75}$$

$\bar{x} = 0.67$ & $\bar{y} = 0.67$

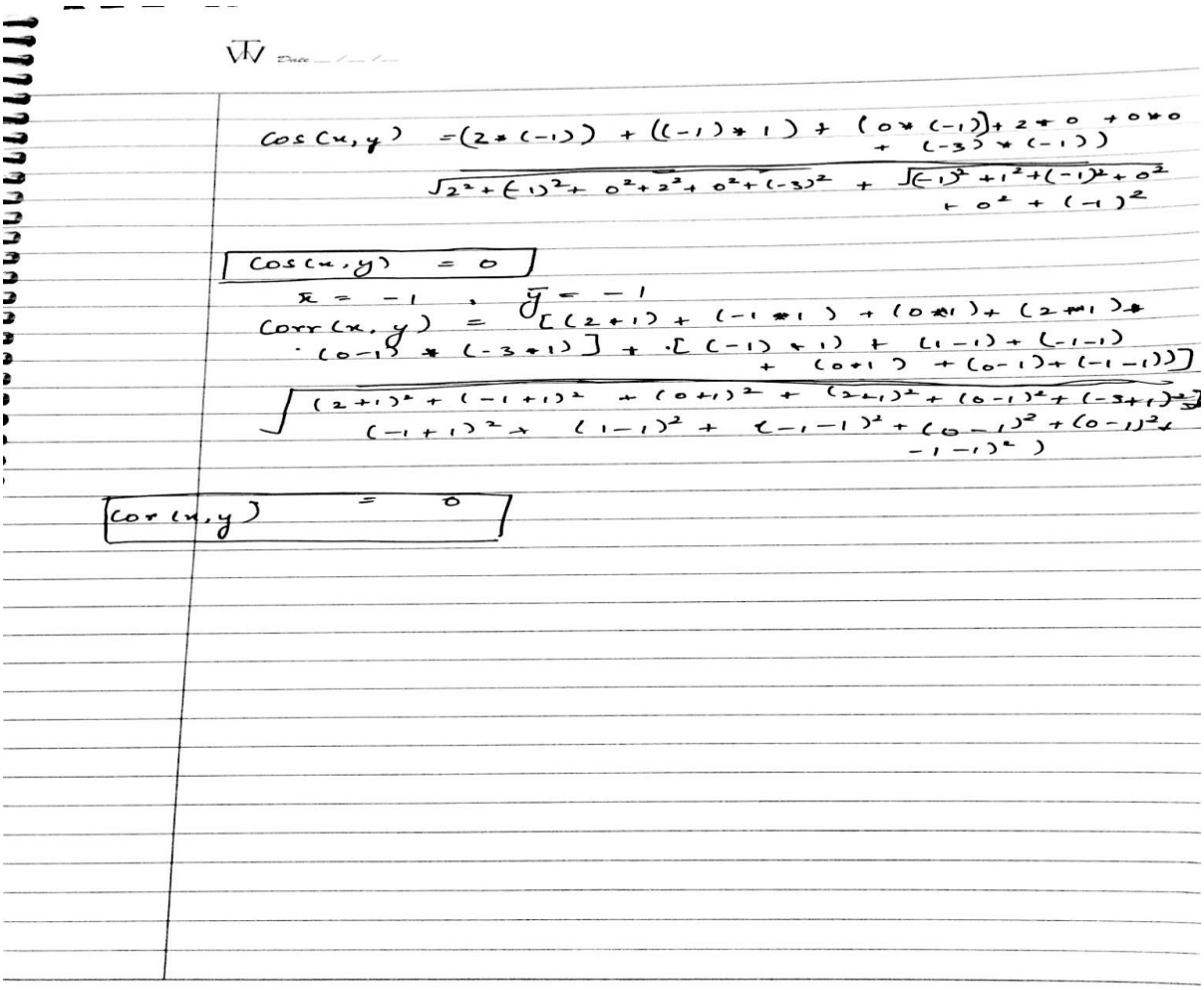
$$\text{cor}(x,y) = \frac{[(1-0.67)*(1-0.67) + (1-0.67)*(1-0.67) + (0-0.67)(1-0.67) + (1-0.67)*(0-0.67) + (1-0.67)(1-0.67)]}{\sqrt{(1-0.67)^2 + (1-0.67)^2 + (0-0.67)^2 + (1-0.67)^2 + (0-0.67)^2 + (1-0.67)^2} * \sqrt{(1-0.67)^2 + (1-0.67)^2 + (1-0.67)^2 + (0-0.67)^2 + (0-0.67)^2 + (1-0.67)^2}}$$

$$\boxed{\text{cor}(x,y) = 0.25}$$

$$\text{Jaccard}(x,y) = \frac{\text{Number of 1-1 matches}}{\text{Total number of attributes not involved in 0-0 matches}}$$
$$= \frac{3}{5}$$
$$= 0.6$$

e) $x = (2,-1,0,2,0,-3)$, $y = (-1,1,-1,0,0,-1)$

Cosine, correlation,



6. Describe how a box plot can give information about whether the value of an attribute is symmetrically distributed. What can you say about the symmetry of the distributions of the attributes shown in Figure 3.11?

Ans. Box plot is a graphically depicting groups of numerical data through their quartiles. Box plot can give whether the value of an attribute is symmetrically distributed if the median of the data is exactly in the middle that is at 50% then the attributes are distributed symmetrically. If it is not at the center then the attributes are distributed, they are skewed. According to the figure 3.11, sepal length's attribute is distributed. For others it is skewed that is Sepal width, petal length, petal width.

7. Consider the training examples shown in Table 4.1 for a binary classification problem.

Ans.

7

a) Compute the Gini index for the overall collection of training examples

Ans.

Nodes	Count
C_0	10
C_1	10

$\therefore$ $Gini(t) = 1 - \sum_{i=0}^{c-1} [p(i|t)]^2$

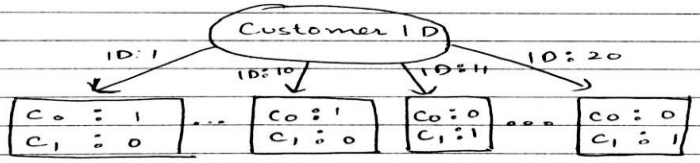
$Gini = 1 - \left(\frac{10}{20}\right)^2 - \left(\frac{10}{20}\right)^2$

$= 1 - (0.5)^2 - (0.5)^2$

$= 1 - 0.5 = 0.5$

b) Compute the Gini index for the Customer ID attribute.

Ans.

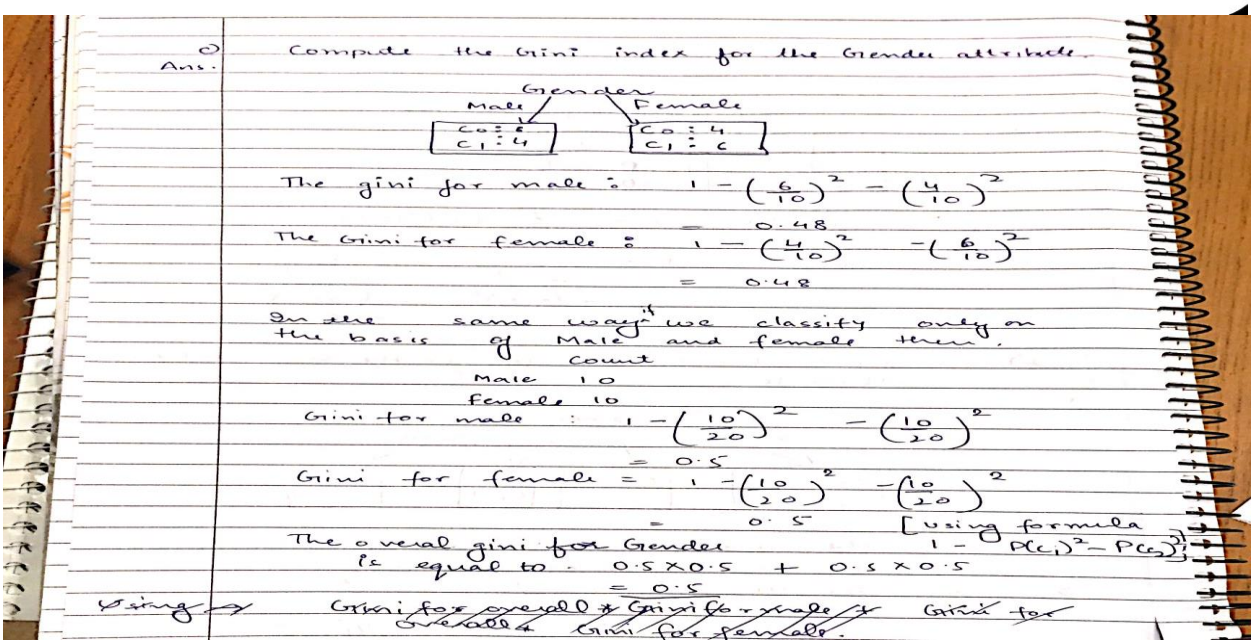


As the Gini for each customer ID is 0.

$\therefore$ the overall gini for customer ID is 0

$\rightarrow$ Gini for ID1 = $1 - (1/1)^2 - (0/1)^2 = 0$

ID 20 = $1 - (0/1)^2 - (1/1)^2 = 0$



c) Compute the Gini index for the Gender attribute.

Ans.

The gini for male : $1 - \left(\frac{6}{10}\right)^2 - \left(\frac{4}{10}\right)^2$

$= 0.48$

The Gini for female : $1 - \left(\frac{4}{10}\right)^2 - \left(\frac{6}{10}\right)^2$

$= 0.48$

In the same way if we classify only on the basis of Male and female then.

	Count
Male	10
female	10

Gini for male : $1 - \left(\frac{10}{20}\right)^2 - \left(\frac{10}{20}\right)^2$

$= 0.5$

Gini for female = $1 - \left(\frac{10}{20}\right)^2 - \left(\frac{10}{20}\right)^2$

$= 0.5$ [using formula $1 - P(C_1)^2 - P(C_0)^2$]

The overal gini for Gender is equal to $0.5 \times 0.5 + 0.5 \times 0.5$

$= 0.5$

Using ~~Gini for overall * Gini for male * Gini for overall * Gini for female.~~

d) Compute the Gini index for the car type attribute using multiway split.

Ans. The multiway split for car type attribute is:-

	Car Type		
	Family	Sports	Luxury
C_0	1	8	1
C_1	3	0	7

The Gini for family car is $1 - \left(\frac{1}{4}\right)^2 - \left(\frac{3}{4}\right)^2$

$= 0.375$

The Gini for Sports car is $1 - \left(\frac{8}{8}\right)^2 - \left(\frac{0}{8}\right)^2$

$= 0$

The Gini for Luxury car is $1 - \left(\frac{1}{8}\right)^2 - \left(\frac{7}{8}\right)^2$

$= 0.2188$

Overall Gini $= \frac{4}{20} \times 0.375 + \frac{8}{20} \times 0 + \frac{8}{20} \times 0.219$

$= 0.163$

f) Which attribute is better, Gender, Car type, or Shirt Size?

Ans. As Car type has lowest Gini among all so it can be treated as better among the three attributes

g) Why customer ID should not be used as the attribute test condition even though it has the lowest Gini.

Ans As it has no predictive power. A new customer will always be assigned a new customer ID's. It is better to make classes by dividing the customers for the attribute test.

c) Compute the Gini index for the Shirt size attribute using multiway split

Ans

	Shirt Size			
	Small	Medium	Large	Extra Large
C_0	3	3	2	2
C_1	2	4	2	2

Fig. Shirt size attribute using multiway split.

Gini for small size is $1 - \left(\frac{3}{5}\right)^2 - \left(\frac{2}{5}\right)^2$

$= 0.48$

Gini for medium size $= 1 - \left(\frac{3}{7}\right)^2 - \left(\frac{4}{7}\right)^2$

$= 0.4898$

Gini for large size $= 1 - \left(\frac{2}{4}\right)^2 - \left(\frac{2}{4}\right)^2$

$= 0.5$

Gini for Extra large size $= 1 - \left(\frac{2}{4}\right)^2 - \left(\frac{2}{4}\right)^2$

$= 0.5$

Overall Gini $= \frac{5}{20} \times 0.48 + \frac{7}{20} \times 0.4898 + \frac{4}{20} \times 0.5 + \frac{4}{20} \times 0.5$

$= 0.4914$

8. Consider the training examples shown in Table 4.2 for a binary classification problem.

Date _/_/_

8.

a) What is the entropy of this collection of training examples with respect to positive class

Ans. According to the data set there are four positive class and five negative class.

$P(+) = 4/9 \qquad P(-) = 5/9$

Entropy for training example $= -\sum_j (j|t) \log p(j|t)$

$= -4/9 \log_2(4/9) - 5/9 \log_2(5/9)$

$= 0.9911$

b) What are the information gains of a_1 and a_2 relative to these training examples.

Ans. For a_1

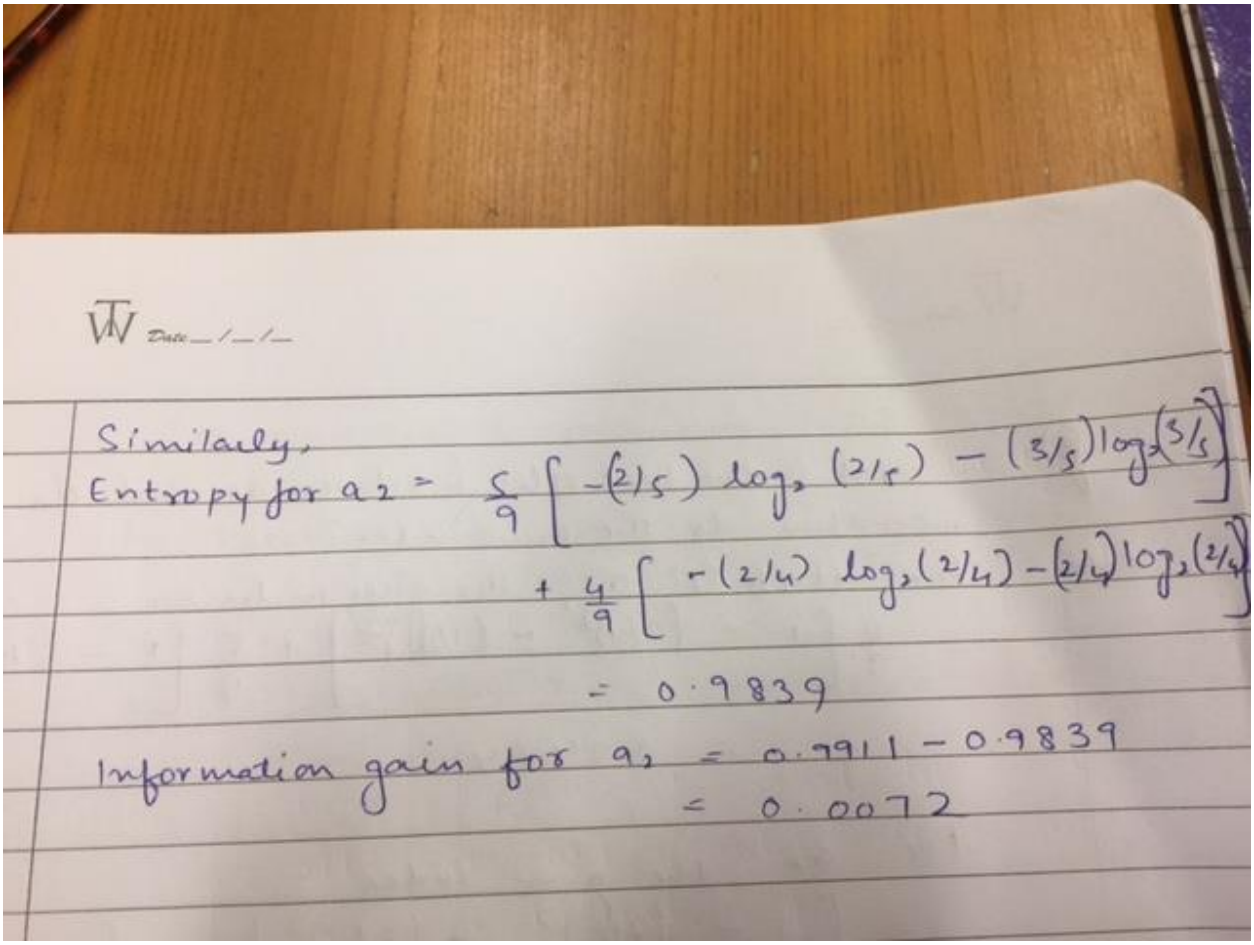
a_1	+	−
T	3	1
F	1	4

Entropy for $a_1 = \frac{4}{9}\left[-(3/4)\log_2(3/4) - (1/4)\log_2(1/4)\right] + \frac{5}{9}\left[-(1/5)\log_2(1/5) - (4/5)\log_2(4/5)\right]$

$= 0.7616$

Information gain for a_1 = Entropy(p) $- \left[\sum_{i=1}^{k} \frac{n_i}{n} \text{Entropy}(i)\right]$

$= 0.9911 - 0.7616 = 0.2294$



d) What is the best split among a1, a2, a3, according to the information gain?
Ans. According to the information a1 produces the best split.

d) What is the best split (between a_1 and a_2) according to the Gini index?

Ans For attribute a_1, the gini index is

$$\frac{4}{9}\left[1 - (3/4)^2 - (1/4)^2\right] + \frac{5}{9}\left[1 - (1/5)^2 - (4/5)^2\right]$$

$$= 0.3444$$

For a_2, the gini index is

$$\frac{5}{9}\left[1 - (2/5)^2 - (3/5)^2\right] + \frac{4}{9}\left[1 - (2/4)^2 - (2/4)^2\right]$$

$$= 0.4889$$

Since Gini index for a_1 is smaller, it produces the better split.

e) What is the best split (between a_1 and a_2) according to the classification error rate?

Ans For a_1 : error rate $= 2/9$

For a_2 : error rate $= 4/9$

$\therefore$ a_1 produces the best split.

2.1 Problem 1

Load the auto-mpg sample dataset into the Orange application, and visualize the dataset. Create a scatterplot between mpg and weight - what is the basic relationship between these variables using just visual inspection? Do the results make sense? Why?

For the problem, auto mpg dataset is loaded in the file in the Orange application. We can visualize the data into a visual format or a tabular format. For the visualizing I am using data tables to visualize and analyze the different attributes and relationship among them.

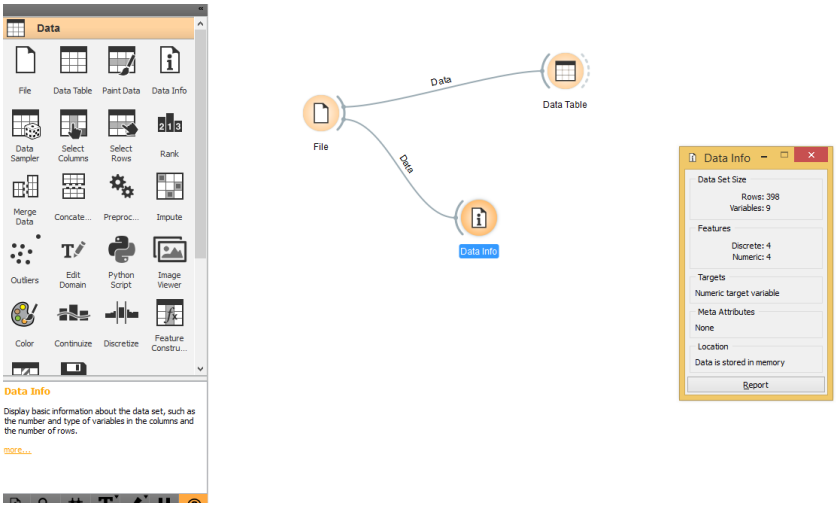
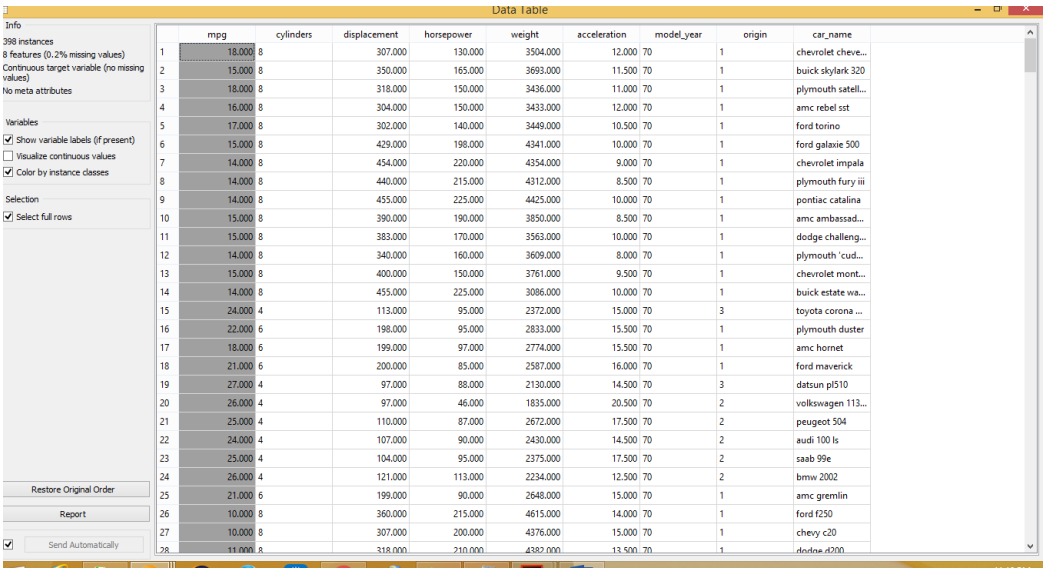


Figure 1. It shows the file loaded with auto-mpg and the data table with the data info.

Once you open the data table, you can visualize the number of attributes, number of instances, also the features. Also you can see the variables.



	mpg	cylinders	displacement	horsepower	weight	acceleration	model_year	origin	car_name
1	18.000	8	307.000	130.000	3504.000	12.000	70	1	chevrolet cheve...
2	15.000	8	350.000	165.000	3693.000	11.500	70	1	buick skylark 320
3	18.000	8	318.000	150.000	3436.000	11.000	70	1	plymouth satell...
4	16.000	8	304.000	150.000	3433.000	12.000	70	1	amc rebel sst
5	17.000	8	302.000	140.000	3449.000	10.500	70	1	ford torino
6	15.000	8	429.000	198.000	4341.000	10.000	70	1	ford galaxie 500
7	14.000	8	454.000	220.000	4354.000	9.000	70	1	chevrolet impala
8	14.000	8	440.000	215.000	4312.000	8.500	70	1	plymouth fury iii
9	14.000	8	455.000	225.000	4425.000	10.000	70	1	pontiac catalina
10	15.000	8	390.000	190.000	3850.000	8.500	70	1	amc ambassad...
11	15.000	8	383.000	170.000	3563.000	10.000	70	1	dodge challeng...
12	14.000	8	340.000	160.000	3609.000	8.000	70	1	plymouth 'cud...
13	15.000	8	400.000	150.000	3761.000	9.500	70	1	chevrolet mont...
14	14.000	8	455.000	225.000	3086.000	10.000	70	1	buick estate wa...
15	24.000	4	113.000	95.000	2372.000	15.000	70	3	toyota corona ...
16	22.000	6	198.000	95.000	2833.000	15.500	70	1	plymouth duster
17	18.000	6	199.000	97.000	2774.000	15.500	70	1	amc hornet
18	21.000	6	200.000	85.000	2587.000	16.000	70	1	ford maverick
19	27.000	4	97.000	88.000	2130.000	14.500	70	3	datsun pl510
20	26.000	4	97.000	46.000	1835.000	20.500	70	2	volkswagen 113...
21	25.000	4	110.000	87.000	2672.000	17.500	70	2	peugeot 504
22	24.000	4	107.000	90.000	2430.000	14.500	70	2	audi 100 ls
23	25.000	4	104.000	95.000	2375.000	17.500	70	2	saab 99e
24	26.000	4	121.000	113.000	2234.000	12.500	70	2	bmw 2002
25	21.000	6	199.000	90.000	2648.000	15.000	70	1	amc gremlin
26	10.000	8	360.000	215.000	4615.000	14.000	70	1	ford f250
27	10.000	8	307.000	200.000	4376.000	15.000	70	1	chevy c20
28	11.000	8	318.000	210.000	4382.000	13.500	70	1	dodge d200

Figure 2. Data table

In the figure 2 we can see that there are 9 attributes, 398 instances, and 8 features for the auto mpg. Different attributes are: mpg, cylinders, displacement, horsepower, weight, acceleration, model year, origin, car name.
We can also check the attribute if it is continuous or not by checking the visualize continues value in the variable section. By checking that you will come to know that mpg, displacement, horsepower, weight, acceleration are continuous variables.

Now plotting the scatterplot between weight on x-axis and mpg on y-axis.

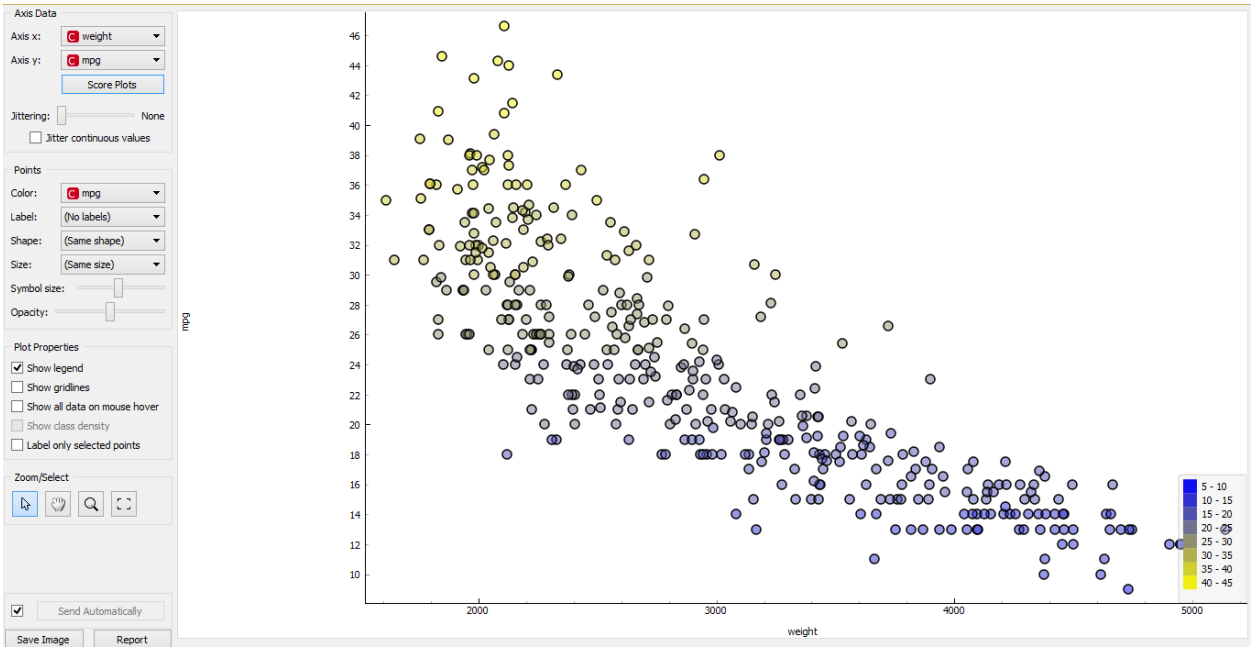


Figure. 3 Scatterplot between weight and mpg.

By visual inspection we can say that weight of the vehicle and the mpg are inversely proportional to each other. We can see that higher the weight, less the mpg and vice versa. Also we can infer from the scatterplot that weight of the vehicle is the explanatory variable and mpg is response variable.

The result make sense as in daily life we can see that heavy weight vehicles have less mpg that is average. For example, a truck will consume more gallons of oil then a car for travelling the same distance so the result make sense with the everyday example.

2. Load the auto-mpg sample dataset into Python using a Pandas dataframe. The horsepower feature has a few missing values with a ? - replace these with a NaN from NumPy, and calculate summary statistics for each numerical column. How do the summary statistics vary when excluding the NaNs, vs. imputing them with the mean (Hint: Use an Imputer from Scikit) - can we do better than just using the overall sample mean?

Ans. Loaded the auto-mpg sample dataset in jupyter using pandas dataframe. After loading the dataset we observed that the mean value for the horse power is not coming. Then according to the problem we replace the ? with a Nan from NumPy. Summary statistics was calculated and was observed that still there was no mean value for the horse power. Now imputer was used and the summary statistics were calculated. It was observed that the mean value was coming now.

It was observed that we cannot do better than just using the overall mean as the mean was coming the same.

	mpg	cylinders	displacement	horsepower	weight	acceleration	year	origin	name
32	25.0	4	98.0	NaN	2046	19.0	71	1	ford pinto
33	19.0	6	232.0	100	2634	13.0	71	1	amc gremlin
34	16.0	6	225.0	105	3439	15.5	71	1	plymouth satellite custom
35	17.0	6	250.0	100	3329	15.5	71	1	chevrolet chevelle malibu
36	19.0	6	250.0	88	3302	15.5	71	1	ford torino 500
37	18.0	6	232.0	100	3288	15.5	71	1	amc matador
38	14.0	8	350.0	165	4209	12.0	71	1	chevrolet impala
39	14.0	8	400.0	175	4464	11.5	71	1	pontiac catalina brougham
40	14.0	8	351.0	153	4154	13.5	71	1	ford galaxie 500

Figure1. After replacing ? with NaN.

	mpg	cylinders	displacement	horsepower	weight	acceleration	year	origin
count	397.000000	397.000000	397.000000	397.000000	397.000000	397.000000	397.000000	397.000000
mean	23.515869	5.458438	193.532746	104.469388	2970.261965	15.555668	75.994962	1.574307
std	7.825804	1.701577	104.379583	38.247388	847.904119	2.749995	3.690005	0.802549
min	9.000000	3.000000	68.000000	46.000000	1613.000000	8.000000	70.000000	1.000000

Figure2. After imputing mean for horsepower.

3. Load the iris sample dataset into Python using a Pandas dataframe. Perform a PCA using the Scikit Decomposition component, and provide the percentage of variance explained by the 1st Principal Component. Use Matplotlib to plot the 1st/2nd Principal Components to recreate the scatterplot shown in class, with colored classes for each ower type.

Ans.

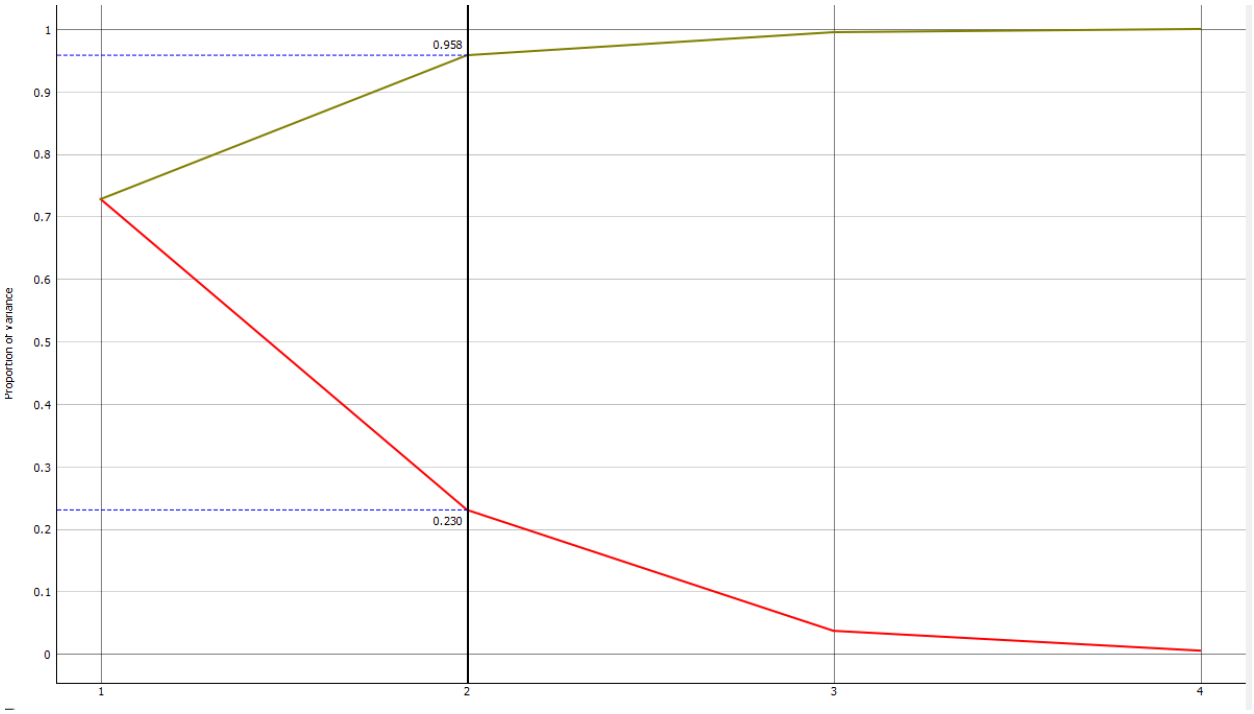


Figure1. PC1 variance

The variance of PC1 is 72% and variance of PC2 is 95%.

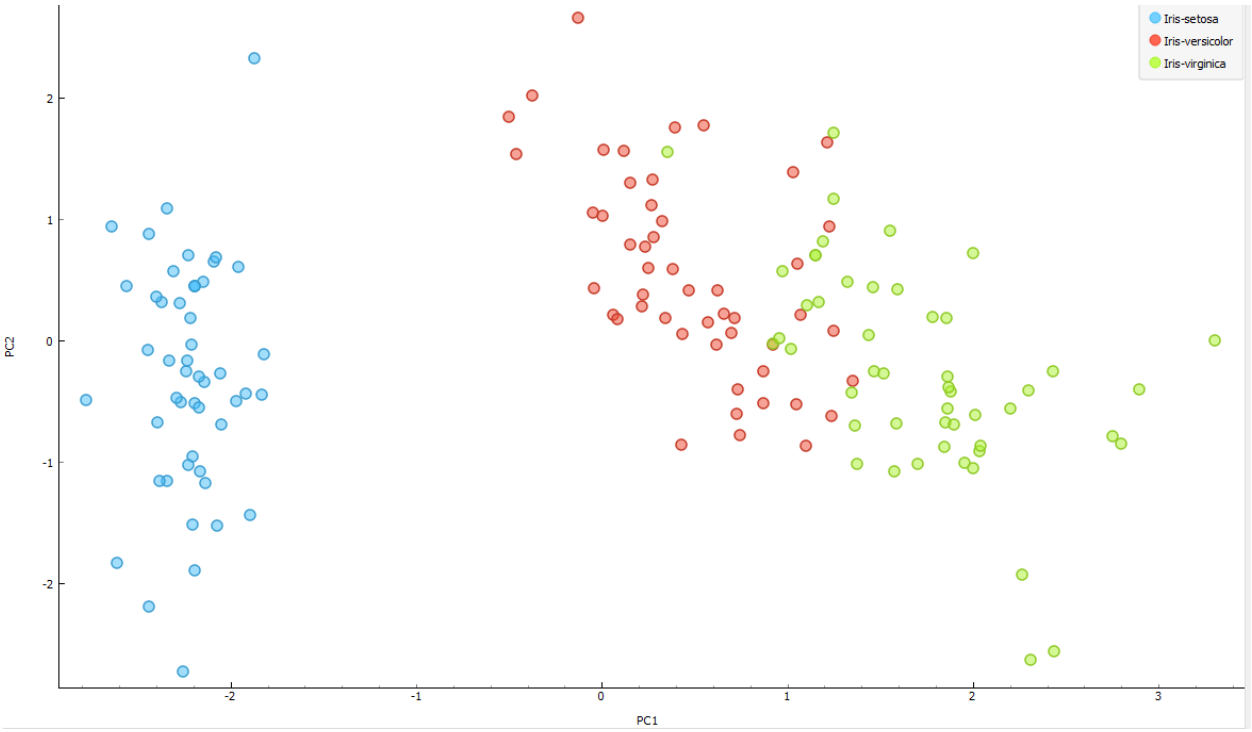


Figure2. Scatterplot of PC1 and PC2.

4. Build two classification trees using the iris sample dataset within the Orange application. Keep all parameters for both classifiers the same (Feature Selection, Pruning), and modify the Limit Depth parameter to a smaller value than the default (e.g., from 10 to 2). How does this affect the Precision and Recall of the classifier? What types of flowers are misclassified? Why? What does Tan refers to as the border where these misclassifications occur?

First iris dataset is loaded and two classification tree with the same parameters are created with one classification tree with the less depth value. For calculating precision and recall we are using test & score which cross validates the set by creating 10 sub sets, out of which 9 are the training model and each one for different which is shown in figure 1. For the Precision and the recall, the value decreases if the depth decreases. As precision and recall depends on the number of true positive value in the attributes. So if the depth of the classification tree is less than the precision and recall will be less as there will be less true positive value. Here is CA in the test and score which is classification accuracy. It reports the proportion of correctly classified instances which is shown in figure 2.

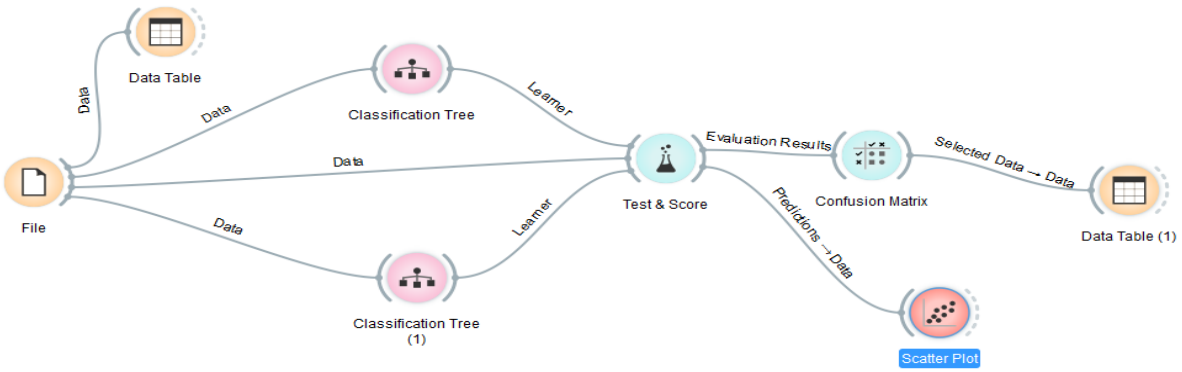


Figure1. Orange Tool for classification.

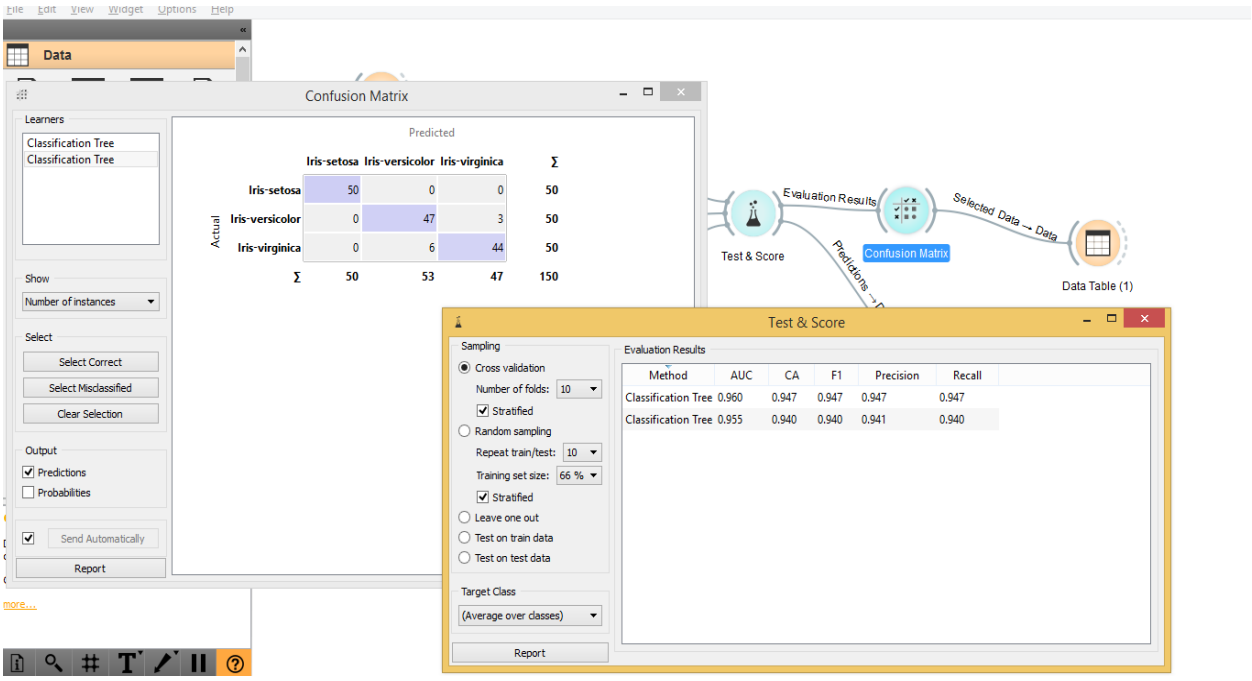


Figure2. Figure shows Score & data.

By using the confusion matrix we can classify the misclassified flowers which are found at the boundary value as the projected values and the actual values are different which can be shown in figure 3. We can also visualize it using a data table. We can visualize that Iris- Sentosa is fine here. Iris- Versicolor and iris-vergonica have misclassified values as they are overlapping each other.

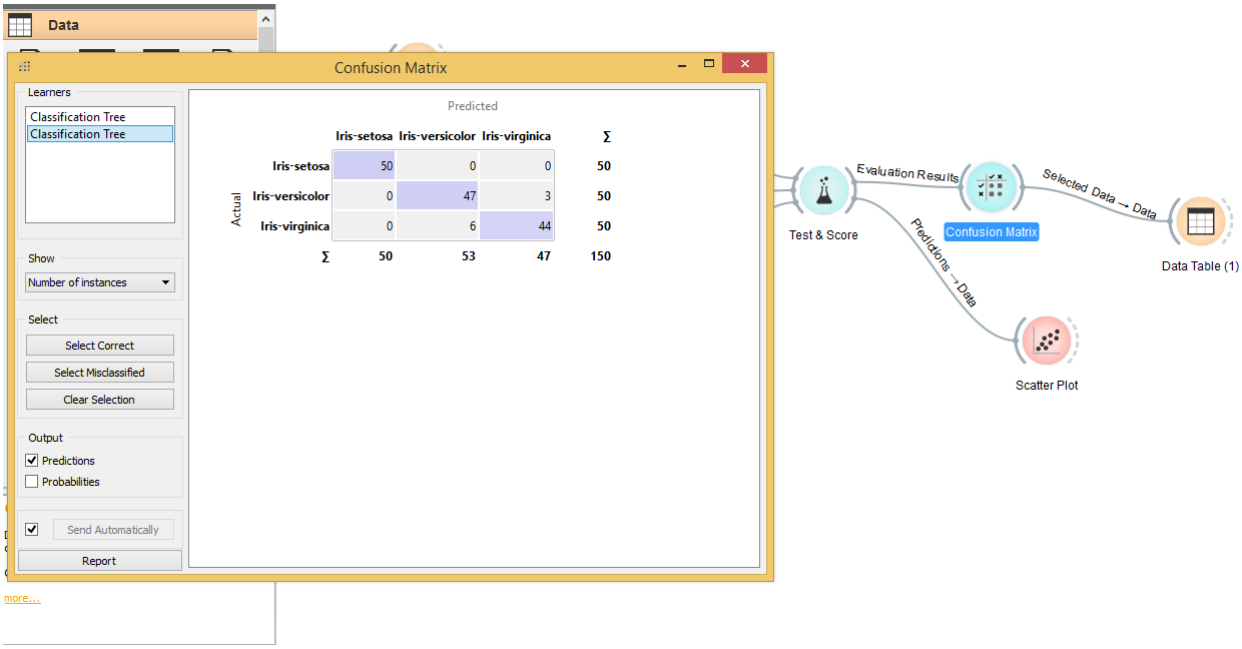


Figure 3. Showing confusion matrix.

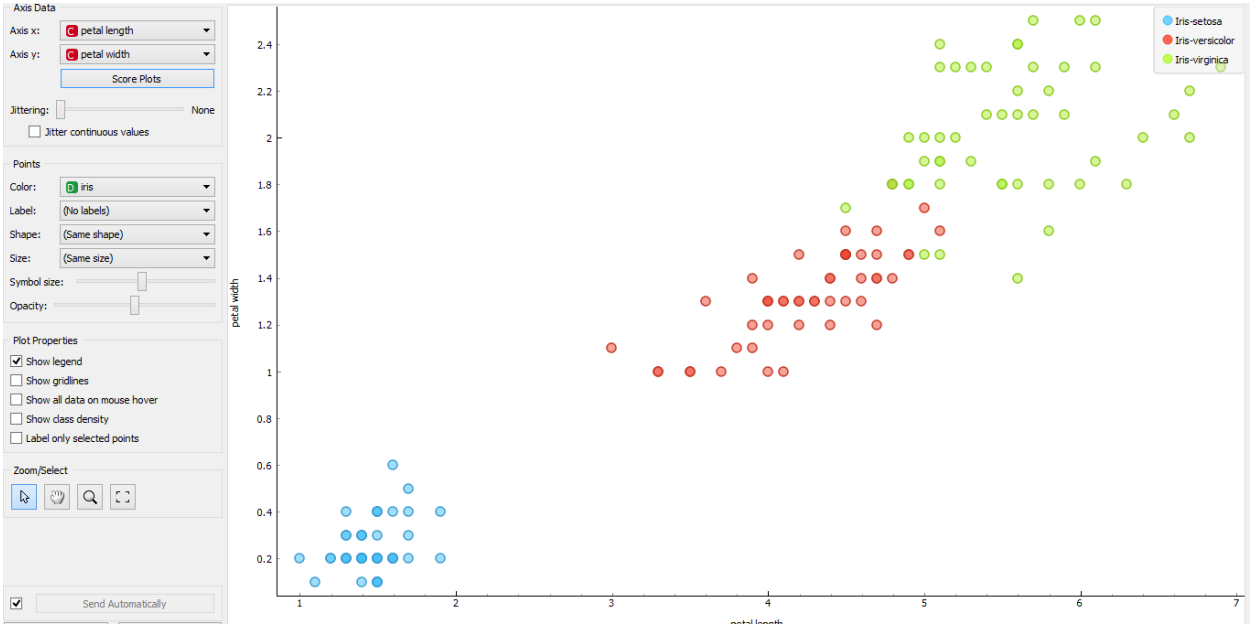


Figure3. Scatter plot of order & data.

Tan refers to decision boundary as the border where the misclassification occurs.