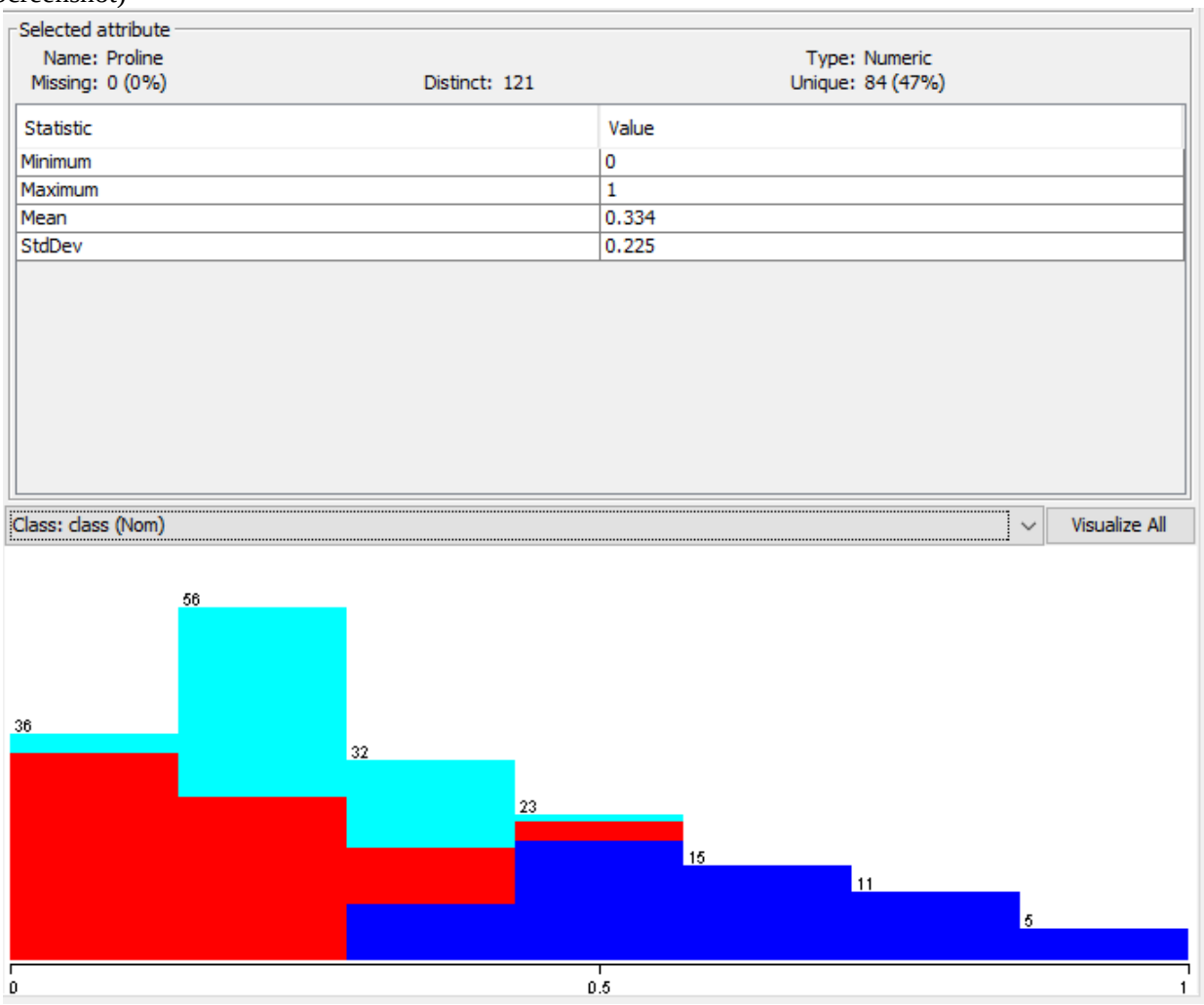
CST8502 - Lab 2
k-Nearest Neighbor (kNN)

Name: Alvin Litani Liauw

Student Number: 41118874

7. (Screenshot)



9.

a. What is the **percentage** of correctly classified items? 94.9438 %
b. What are the True Positive (TP) rates of **each** class?
Class 1 is 1.0. Class 2 is 0.873. Class 3 is 1.0.
c. Look at the confusion matrix, which class is incorrectly classified?
Class 2

10. Fill in the table.

K	Percentage of correctly classified instances	Number of instances misclassified in each class
3	94.9438 %	1: 0 2: 9 3: 0
5	95.5056 %	1: 0 2: 7 3: 1
7	94.9438 %	1: 0 2: 8 3: 1
9	96.0674 %	1: 0 2: 6 3: 1

11. Repeat step 10 with "Percentage Split" of 70. Fill in the following table.

K	Percentage of correctly classified instances	Number of instances misclassified in **each** class
3	100	1: 0 2: 0 3: 0
5	98.1132 %	1: 0 2: 1 3: 0
7	100	1: 0 2: 0 3: 0
9	100	1: 0 2: 0 3: 0

12. Explanation of the process and the screenshot.

Student Number: 41118874
Instance number: 74
Row number: 75
Steps of the Test Process:

1. Normalize each column
2. Calculate the Euclidean distances between other instances with the 74th instance using the normalized data.
3. Since it is 5NN, we look at the 5 instances with the smallest Euclidean distances as they are the nearest neighbours to the 74th distance in the multidimensional space.
4. Check the classes of the 5 nearest neighbours and note the majority class. The 74th instance should be classified similarly as the majority class of the neighbours. In this case, the majority of the neighbours' classes are 1. The 74th instance is wrongly classified in the training data as class 2.

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	AB
1	class	Alcoho	Malic a	Ash	Alcalin	Magnes	Total phen	Flavan	Nonfla	Proanth	Color ir	Hue	OD280/	Proline	normalized	distanc
6	1	13.24	2.59	2.87	21	118	2.8	2.69	0.39	1.82	4.32	1.04	2.93	735	0.581578947	0.7777
27	1	13.05	2.05	3.22	25	124	2.63	2.68	0.47	1.92	3.58	1.13	3.2	830	0.531578947	0.7423
32	1	13.73	1.5	2.7	22.5	101	3	3.25	0.29	2.38	5.7	1.19	2.71	1285	0.710526316	0.7733
37	1	13.48	1.81	2.41	20.5	100	2.7	2.98	0.26	1.86	5.1	1.04	3.47	920	0.644736842	0.7570
57	1	13.56	1.73	2.46	20.5	116	2.96	2.78	0.2	2.45	6.25	0.98	3.03	1120	0.665789474	0.7357
75	2	12.99	1.67	2.6	30	139	3.3	2.89	0.21	1.96	3.35	1.31	3.5	985	0.515789474	0.0000

The AB column is the Euclidean distance from the 74th instance. I truncated the P-AA columns in the excel sheet containing the normalized values for each attribute. The nearest neighbours are class 1 while 74th instance is class 2 therefore it is misclassified.

An example of the formula for the distance for 5th instance is = SQRT((O6-O$75)^2+(P6-P$75)^2+(Q6-Q$75)^2+(R6-R$75)^2+(S6-S$75)^2+(T6-T$75)^2+(U6-U$75)^2+(V6-V$75)^2+(W6-W$75)^2+(X6-X$75)^2+(Y6-Y$75)^2+(Z6-Z$75)^2+(AA6-AA$75)^2)

13. Definition of mean: the average of a dataset or the sum of the data divided by the number of instances

Definition of median: the middle value of the dataset after it has been sorted from smallest to largest

Definition of mode: the most often occurring value in the dataset

Student number: 41118874

List of numbers: 4,1,1,1,8,8,7,4

Mean (show calculation): (4+1+1+1+8+8+7+4) / 8 = 34/8 = 4.25

Median (Show how you find it): ranking the list = 1,1,1,4,4,7,8,8 = (4+4)/2 = 4

Mode (Show how you find it): ranking the list = 1,1,1,4,4,7,8,8 = 1 as most often occurring value

Weighted average (show calculation):

(4*1) + (1*2) + (1*3) + (1*4) + (8*5) +(8*6) + (7*7) + (4*8) / (1+2+3+4+5+6+7+8)

= 4+2+3+4+40+48+49+32/36 = 182/36 = 5.06