

## **Chapter 1:**

### **Problem 1-**

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

- (a) Dividing the customers of a company according to their gender.

Answer: No. Database query.

- (b) Dividing the customers of a company according to their profitability.

Answer: No. Calculation

- (c) Computing the total sales of a company.

Answer: No. Calculation.

- (d) Sorting a student database based on student identification numbers.

Answer: No, Database query.

- (e) Predicting the outcomes of tossing a (fair) pair of dice.

Answer: No, Probability calculation

- (f) Predicting the future stock price of a company using historical records.

Answer: Yes, we can create a model

- (g) Monitoring the heart rate of a patient for abnormalities.

Answer: Yes, we can build a model.

- (h) Monitoring seismic waves for earthquake activities.

Answer: Yes, we can build a model.

- (i) Extracting the frequencies of a sound wave.

Answer: No

## **Chapter 2**

Q. 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.

Example: Age in years. Answer: Discrete, quantitative, ratio

- a) Time in terms of AM or PM.

**Answer:** Binary, qualitative, ordinal.

- b) Brightness as measured by a light meter.

**Answer:** Continuous, quantitative, ratio

c) Brightness as measured by people's judgments.

**Answer:** Discrete, qualitative, ordinal

d) Angles as measured in degrees between 0 and 360.

**Answer:** Continuous, quantitative, ratio

e) Bronze, Silver, and Gold medals as awarded at the Olympics.

**Answer:** Discrete, qualitative, ordinal

f) Height above sea level.

**Answer:** Continuous, quantitative, interval/ratio

g) Number of patients in a hospital.

**Answer:** Discrete, quantitative, ratio

h) ISBN numbers for books. (Look up the format on the Web.)

**Answer:** Discrete, qualitative, nominal

i) Ability to pass light in terms of the following values: opaque, translucent, transparent.

**Answer:** Discrete, qualitative, ordinal

j) Military rank.

**Answer:** Discrete, qualitative, ordinal

k) Distance from the center of campus.

**Answer:** Continuous, quantitative, interval/ratio (depends)

l) Density of a substance in grams per cubic centimeter.

**Answer:** Discrete, quantitative, ratio

m) Coat check number. (When you attend an event, you can often give your coat to someone who, in turn, gives you a number that you can use to claim your coat when you leave.)

**Answer :** Discrete, qualitative, nominal

Q. 7) Which of the following quantities is likely to show more temporal autocorrelation: daily rainfall or daily temperature? Why?

**Answer :** When the locations that are closer to each other are more similar with respect to the values of that feature than the locations that are far away then a feature shows spatial correlation. It is common for physically close locations to have similar temperatures than the similar amount of rainfall, because rainfall can change from one location to other. Hence, daily temperature shows more temporal correlation than daily rainfall.

Q. 18) 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

**Answer:**

Hamming distance = number of different bits = 3

Jaccard Similarity = number of 1-1 matches / (number of bits - number 0-0 matches) =  $\frac{2}{5} = 0.4$

(b) Which approach, Jaccard or Hamming distance, is more similar to the Simple Matching Coefficient(SMC), 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.)

**Answer:**

The Hamming distance is similar to the SMC .

$SMC = \text{Hamming distance} / \text{number of bits}$ .

The Jaccard measure is similar to the cosine measure because both ignore 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.)

**Answer:** We want to see how many genes these two organisms share. Hence, Jaccard is more appropriate for comparing the genetic makeup of two organisms.

(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.)

**Answer:**

The Hamming distance will be more useful in this case because, Two human beings share >99.9% of the same genes. If we want to compare the genetic makeup of two human beings, we should focus on their differences.

19) 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

**Answer:**

$$\cos(x,y) = 1,$$

$$\text{corr}(x,y) = 0/0 \text{ (undefined)},$$

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

(b) x : (0, 1,0, 1), y : (1,0, 1,0) cosine, correlation, Euclidean, Jaccard

**Answer:**

$$\cos(x,y) = 0,$$

$$\text{corr}(x,y) = -1,$$

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

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

(c) x : (0, -1,0,1), y = (1, 0,-1,0) cosine, correlation, Euclidean

**Answer:**

$$\cos(x,y) = 0,$$

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

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

(d) x : (1,1,0,1,0,1), y : (1,1,1,0,0,1) cosine, correlation,Jaccard

**Answer:**

$$\cos(x,y) = 0.75,$$

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

$$\text{Jaccard}(x,y) = 0.6$$

(e) x : (2, -7,0,2,0, -3), y : (-1, 1,- 1,0,0, -1) cosine, relation

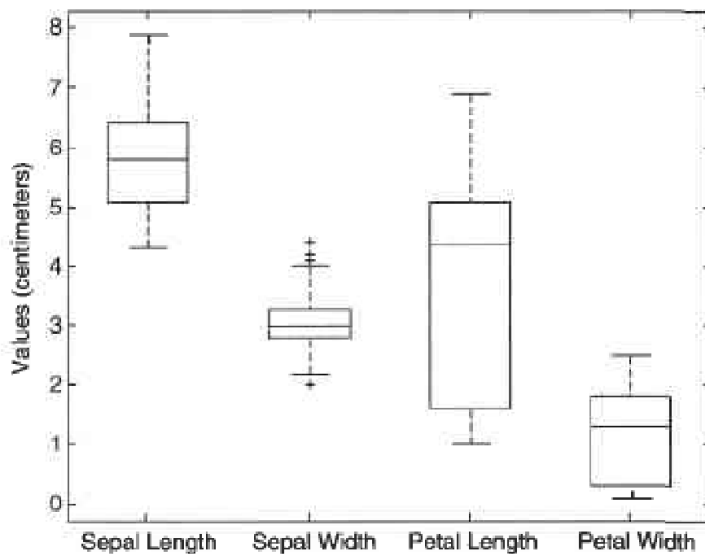
**Answer:**

$$\cos(x,y) = 0,$$

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

### **Chapter 3**

8) 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?



**Figure 3.11.** Box plot for Iris attributes.

- If the line representing the median of the data is in the middle of the box, then the data is symmetrically distributed, at least in terms of the 75% of the data between the first and third quartiles. For the remaining data, the length of the whiskers and outliers is also an indication, although, since these features do not involve as many points, they may be misleading.
- Sepal width and length seem to be relatively symmetrically distributed, petal length seems to be rather skewed, and petal width is somewhat skewed.

#### **Chapter 4:**

Q. 2) Consider the training examples shown in Table 4.7 for a binary classification problem.

**Table 4.7.** Data set for Exercise 2.

Customer ID	Gender	Car Type	Shirt Size	Class
1	M	Family	Small	C0
2	M	Sports	Medium	C0
3	M	Sports	Medium	C0
4	M	Sports	Large	C0
5	M	Sports	Extra Large	C0
6	M	Sports	Extra Large	C0
7	F	Sports	Small	C0
8	F	Sports	Small	C0
9	F	Sports	Medium	C0
10	F	Luxury	Large	C0
11	M	Family	Large	C1
12	M	Family	Extra Large	C1
13	M	Family	Medium	C1
14	M	Luxury	Extra Large	C1
15	F	Luxury	Small	C1
16	F	Luxury	Small	C1
17	F	Luxury	Medium	C1
18	F	Luxury	Medium	C1
19	F	Luxury	Medium	C1
20	F	Luxury	Large	C1

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

**Answer:**

$$\text{Gini} = 1 - 2 * 0.5^2 = 0.5$$

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

**Answer:**

The gini for each Customer ID value is 0. Therefore, the overall gini for Customer ID is 0.

(c) Compute the Gini index for the Gender attribute.

**Answer:**

The gini for Male is  $1 - 2 * 0.5^2 = 0.5$

The gini for Female is also 0.5.

The final gini for gender is  $0.5 * 0.5 + 0.5 * 0.5 = 0.5$

(d) Compute the Gini index for the Car Type attribute using multiway split.

**Answer:**

The gini for Family car is 0.375, Sports car is 0, and Luxury car is 0.2188. The overall gini is 0.1625.

(e) Compute the Gini index for the Shirt Size attribute using multiway split.

**Answer:**

The gini for Small shirt size is 0.48, Medium shirt size is 0.4898, Large shirt size is 0.5, and Extra Large shirt size is 0.5. The overall gini for Shirt Size attribute is 0.4914.

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

**Answer:**

Car Type because it has the lowest gini among the three attributes.

(g) Explain why Customer ID should not be used as the attribute test condition even though it has the lowest Gini.

**Answer:**

The attribute has no predictive power since new customers are assigned to new Customer IDs.

Q. 3) Consider the training examples shown in Table 4.8 for a binary classification problem.

**Table 4.8.** Data set for Exercise 3.

Instance	$a_1$	$a_2$	$a_3$	Target Class
1	T	T	1.0	+
2	T	T	6.0	+
3	T	F	5.0	-
4	F	F	4.0	+
5	F	T	7.0	-
6	F	T	3.0	-
7	F	F	8.0	-
8	T	F	7.0	+
9	F	T	5.0	-

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

**Answer:**

$$P(+) = 4/9$$

$$P(-) = 5/9.$$

The entropy of the training examples is  $-4/9\log_2(4/9) - 5/9\log_2(5/9) = 0.9911$ .

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

**Answer:**

For attribute  $a_1$ , the corresponding counts and probabilities are:

$a_1$	+	-
T	3	1
F	1	4

The entropy for  $a_1$  is:

$$4/9 \left[ -(3/4)\log_2(3/4) - (1/4)\log_2(1/4) \right]$$

+

$$5/9 \left[ -(1/5)\log_2(1/5) - (4/5)\log_2(4/5) \right] = 0.7616.$$

Therefore, the information gain for  $a_1$  is  $0.9911 - 0.7616 = 0.2294$ .

For attribute  $a_2$ , the corresponding counts and probabilities are:

$a_2$	+	-
T	2	3
F	2	2

The entropy for  $a_2$  is:

$$5/9 \left[ -(2/5)\log_2(2/5) - (3/5)\log_2(3/5) \right]$$

+

$$4/9 \left[ -(2/4)\log_2(2/4) - (2/4)\log_2(2/4) \right] = 0.9839$$

Therefore, the information gain for  $a_2$  is  $0.9911 - 0.9839 = 0.0072$ .

(c) For  $a_3$ , which is a continuous attribute, compute the information gain for every possible split.

**Answer:**



a <sub>3</sub>	Class label	Split point	Entropy	Info Gain
1.0	+	2.0	0.8484	0.1427
3.0	-	3.5	0.9885	0.0026
4.0	+	4.5	0.9183	0.0728
5.0 5.0	- -	5.5	0.9839	0.0072
6.0	+	6.5	0.9728	0.0183
7.0 7.0	+ -	7.5	0.8889	0.1022

The best split for a<sub>3</sub> occurs at split point equals to 2.

(d) What is the best split (among a<sub>1</sub> , a<sub>2</sub> , and a<sub>3</sub>) according to the information gain?

**Answers:**

According to information gain, a<sub>1</sub> produces the best split.

(e) What is the best split(between a<sub>1</sub> and a<sub>2</sub>)according to the classification error rate?

**Answers:**

For attribute a<sub>1</sub>: error rate = 2/9.

For attribute a<sub>2</sub>: error rate = 4/9.

Therefore, according to error rate, a<sub>1</sub> produces the best split.

(f) What is the best split (between a<sub>1</sub> and a<sub>2</sub>) according to the Gini index?

**Answers:**

For attribute a<sub>1</sub>, the gini index is

$$4/9*[1-(3/4)^2 - (1/4)^2] + 5/9*[1 - (1/5)^2 - (4/5)^2] = 0.3444$$

For attribute a<sub>2</sub>, the gini index is

$$5/9*[1 - (2/5)^2 - (3/5)^2] + 4/9[1 - (2/4)^2 - (2/4)^2] = 0.4889$$

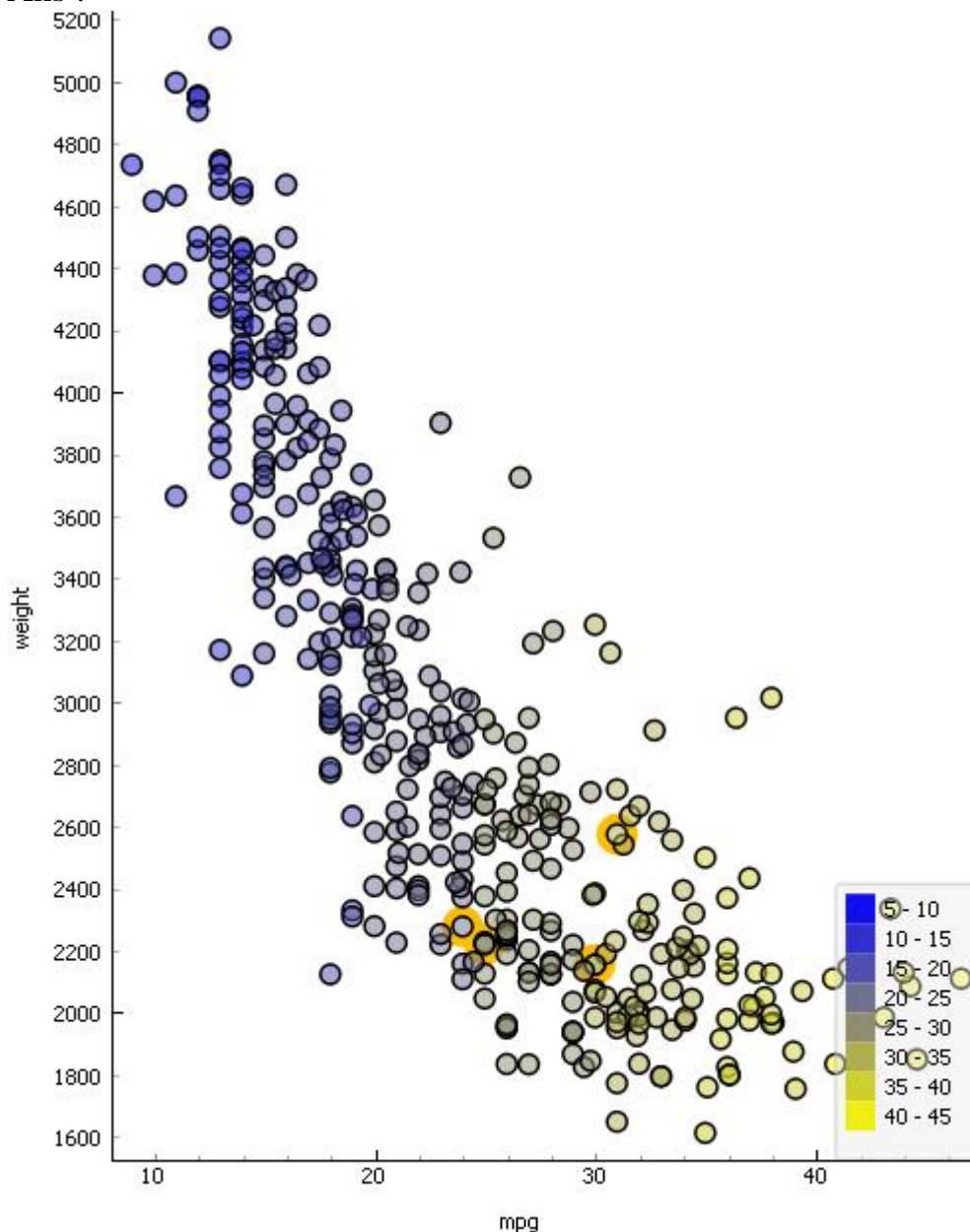
## Practicum Problems

### 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?

Ans :



The scatter plot shows that when the magnitude of mpg is less the value for weight is high. In other words, mpg is inversely proportional to the weight. Hence, the plot indicates that if mpg increases the weight will decrease. This shows that they have negative correlation.

## 2.2 Problem 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?

**Answer:**

When '?' is not replaced

Summary Statistics

	mpg	cylinders	displacement	horsepower	weight	acceleration	model_year	origin	car_name
count	398	398	398	398	398	398	398	398	398
unique	129	5	82	94	351	95	13	3	305
top	13	4	97	150	1985	14.5	73	1	ford pinto
freq	20	204	21	22	4	23	40	249	6

When '?' was replaced by 'NaN' in horsepower:

```
count      392.000000
mean       104.469388
std        38.491160
min        46.000000
25%                NaN
50%                NaN
75%                NaN
max        230.000000
Name: horsepower, dtype: float64
```

When 'NaN' is replaced by Mean:

	horsepower
count	398.000000
mean	104.469388
std	38.199187
min	46.000000
25%	76.000000
50%	95.000000
75%	125.000000
max	230.000000

When Replaced by Median:

	horsepower
count	398.000000
mean	104.304020
std	38.222625
min	46.000000
25%	76.000000
50%	93.500000
75%	125.000000
max	230.000000

When replaced by mode:

	horsepower
count	398.000000
mean	105.155779
std	38.600986
min	46.000000
25%	76.000000
50%	95.000000
75%	130.000000
max	230.000000

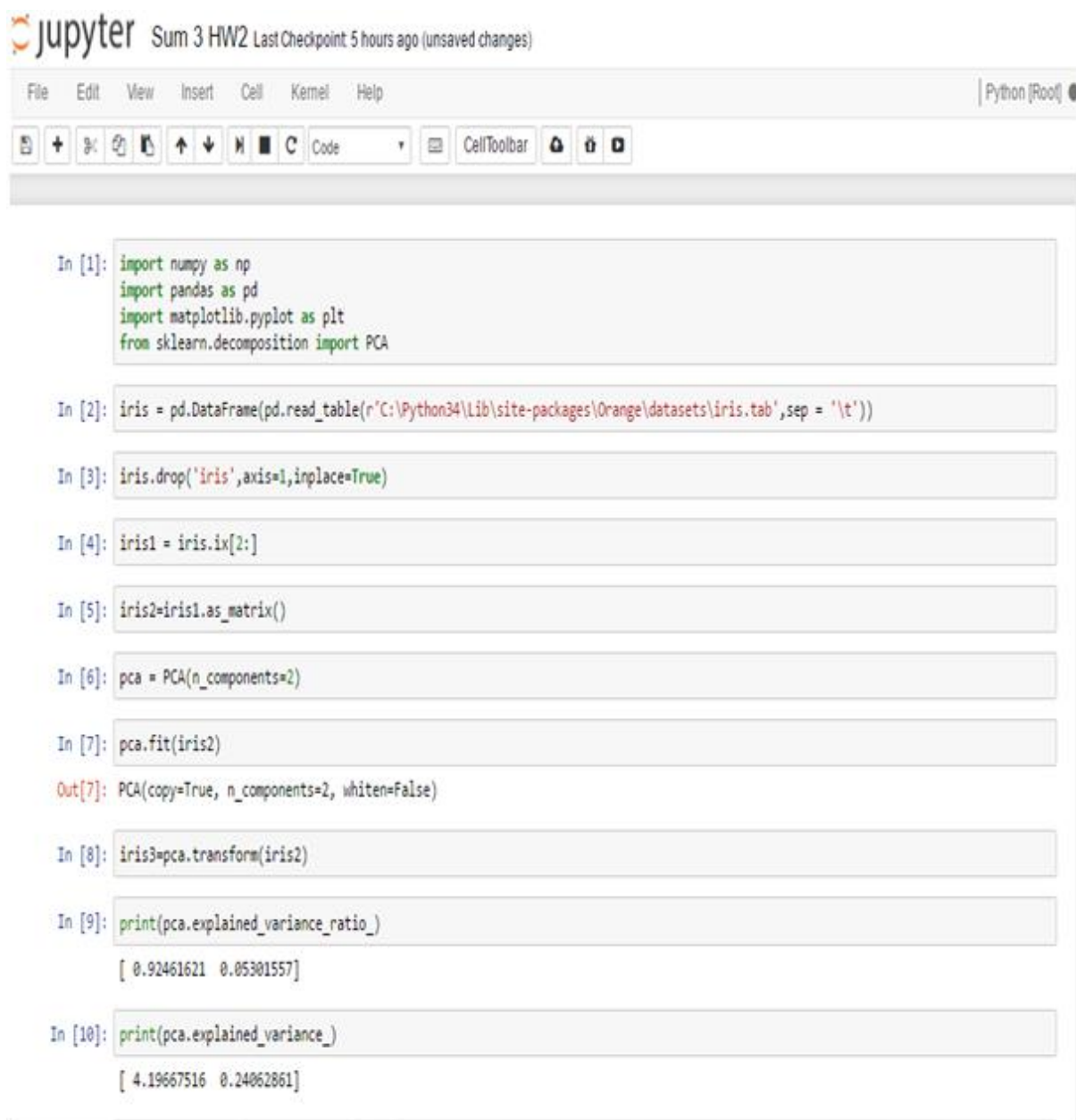
Imputting code :

```
imp = Imputer(missing_values='NaN', strategy='mean', axis=0)
imp.fit(auto2)
auto_median = imp.fit_transform(auto2)
auto['horsepower'] = auto_median
auto.describe()
```

If strategy is mean then it replaces NaN with its mean value.

By using imputer we are making the summary statistics better, as we can see there are no results for summary statistics when there are missing values. When the values are replaced by NaN we get NaN results. But when we impute with mean, we get better results. When we use median and mode instead of mean the standard deviation increases.

3.



The image shows a Jupyter Notebook interface with the title "Sum 3 HW2" and a status "Last Checkpoint 5 hours ago (unsaved changes)". The interface includes a menu bar (File, Edit, View, Insert, Cell, Kernel, Help) and a toolbar with icons for file operations, cell execution, and output viewing. The notebook contains ten input cells and one output cell, all with Python code for PCA analysis.

```
In [1]: import numpy as np
import pandas as pd
import matplotlib.pyplot as plt
from sklearn.decomposition import PCA

In [2]: iris = pd.DataFrame(pd.read_table(r'C:\Python34\Lib\site-packages\Orange\datasets\iris.tab', sep = '\t'))

In [3]: iris.drop('iris',axis=1,inplace=True)

In [4]: iris1 = iris.ix[2:]

In [5]: iris2=iris1.as_matrix()

In [6]: pca = PCA(n_components=2)

In [7]: pca.fit(iris2)

Out[7]: PCA(copy=True, n_components=2, whiten=False)

In [8]: iris3=pca.transform(iris2)

In [9]: print(pca.explained_variance_ratio_)
[ 0.92461621  0.05301557]

In [10]: print(pca.explained_variance_)
[ 4.19667516  0.24062861]
```

```

In [11]: iris4=pd.DataFrame(iris3)

In [12]: iris4 = iris4.ix[2:]

In [14]: irisnew = pd.DataFrame(pd.read_table(r'C:\Python34\Lib\site-packages\Orange\datasets\iris.tab',sep = '\t'))

In [15]: irisnew.drop(['sepal length','sepal width','petal length','petal width'],axis=1,inplace=True)

In [16]: irisnew1 = irisnew.ix[2:]

In [17]: irisnew2=pd.DataFrame(irisnew1)

In [18]: result = pd.concat([iris4, irisnew2], axis=1)

In [19]: result1=result.rename(columns=(0:'PCA1',1:'PCA2','iris':'Iris'))

```

```
In [20]: result1
```

```
Out[20]:
```

	PCA1	PCA2	Iris
2	-2.889820	0.137346	Iris-setosa
3	-2.746437	0.311124	Iris-setosa
4	-2.728593	-0.333925	Iris-setosa
5	-2.279897	-0.747783	Iris-setosa
6	-2.820891	0.082105	Iris-setosa
7	-2.626482	-0.170405	Iris-setosa
8	-2.887959	0.570798	Iris-setosa
9	-2.673045	0.400000	Iris-setosa

150	NaN	NaN	Iris-virginica
151	NaN	NaN	Iris-virginica

150 rows x 3 columns

```

In [21]: gb = result1.groupby(result1['Iris'])

In [22]: versicolor = gb.get_group('Iris-versicolor')

In [23]: virginica = gb.get_group('Iris-virginica')

In [24]: setosa = gb.get_group('Iris-setosa')

In [25]: ver = versicolor.plot.scatter(x='PCA1', y='PCA2', color='Red', label='Group 1');

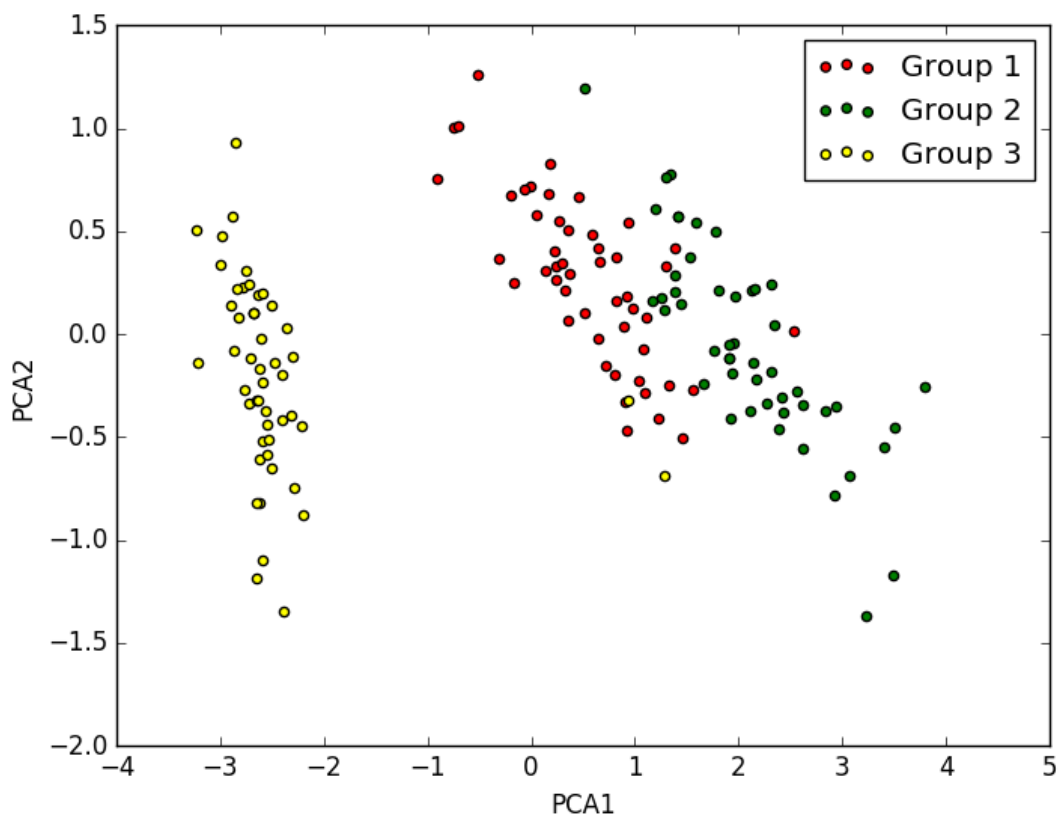
In [26]: vir = virginica.plot.scatter(x='PCA1', y='PCA2', color='Green', label='Group 2', ax=ver);

In [27]: seto = setosa.plot.scatter(x='PCA1', y='PCA2', color='Yellow', label='Group 3', ax=ver);

In [28]: plt.show()

In [ ]:

```



#### 2.4 Problem 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 refer to as the border where these misclassifications occur?

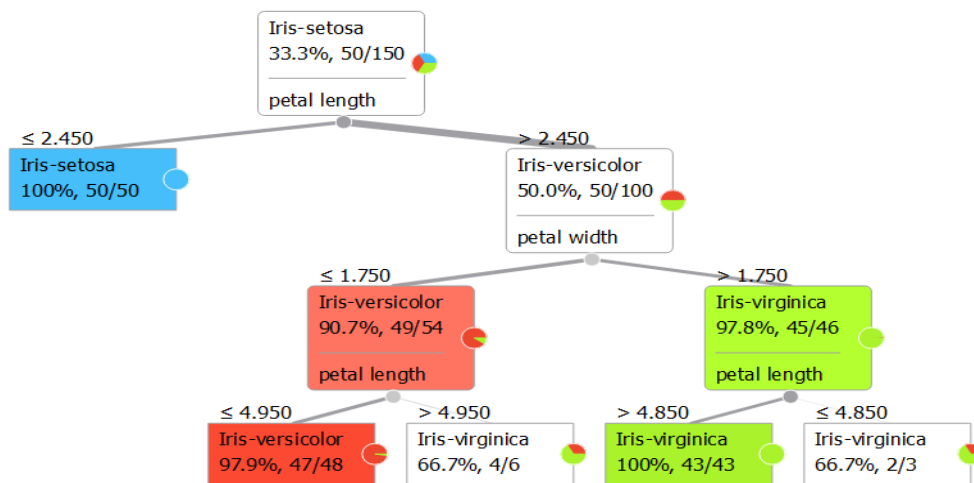
**Answer:**

For the above problem, depth parameters were assumed as 3 and 10. As the depth increases the precision and recall decreases.

Iris-Versicolor and Iris-Verginica are misclassified. The flowers are misclassified as only the petal length and petal width are considered while classifying.

Tan refers to the border where these misclassifications occur as the decision boundary.

Depth=3





**Confusion Matrix**

Wed Sep 21 16, 23:06:42

Confusion matrix for Classification Tree (showing proportion of predicted)

		Predicted			$\Sigma$
		Iris-setosa	Iris-versicolor	Iris-virginica	
Actual	Iris-setosa	100.0 %	0.0 %	0.0 %	50
	Iris-versicolor	0.0 %	93.8 %	9.6 %	50
	Iris-virginica	0.0 %	6.2 %	90.4 %	50
$\Sigma$		50	48	52	150

**Test & Score**

Wed Sep 21 16, 23:06:58

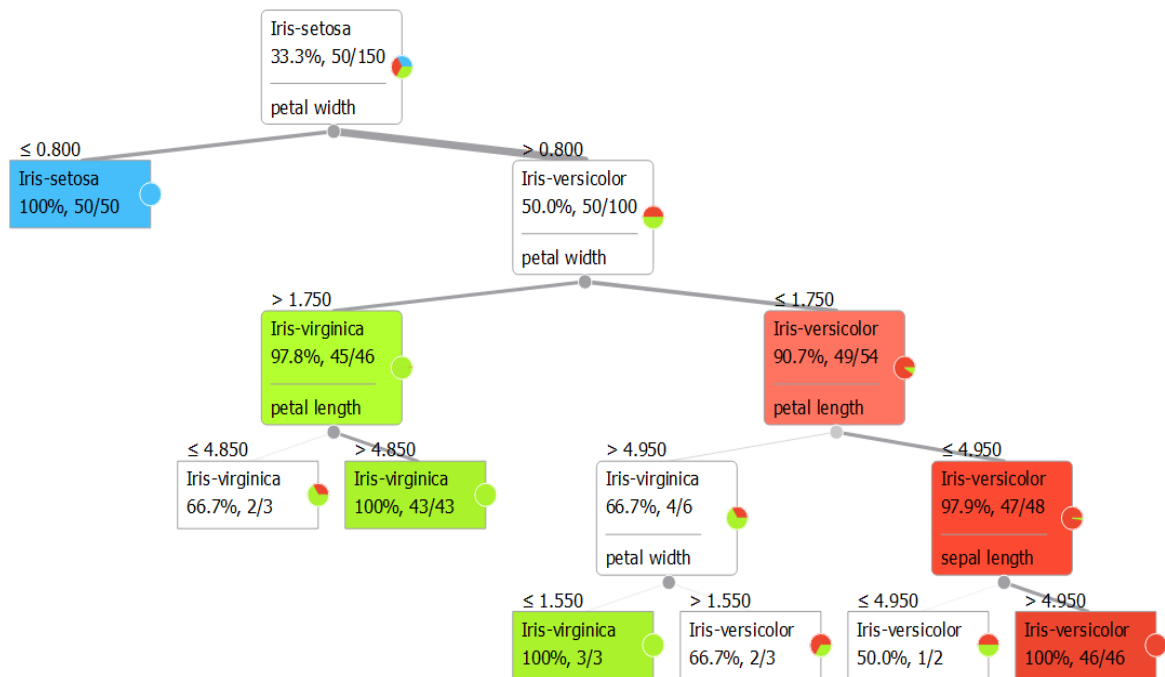
**Settings**

Sampling type: Stratified 3-fold Cross validation  
Target class: Average over classes

**Scores**

Method	AUC	CA	F1	Precision	Recall
Classification Tree	0.959	0.947	0.947	0.947	0.947

Depth=10



## Confusion Matrix

Wed Sep 21 16, 22:57:06

Confusion matrix for Classification Tree (showing proportion of actual)

		Predicted			Σ
		Iris-setosa	Iris-versicolor	Iris-virginica	
Actual	Iris-setosa	100.0 %	0.0 %	0.0 %	50
	Iris-versicolor	0.0 %	92.0 %	8.0 %	50
	Iris-virginica	0.0 %	12.0 %	88.0 %	50
Σ		50	52	48	150

## Test & Score

Wed Sep 21 16, 22:58:29

### Settings

Sampling type: Stratified 3-fold Cross validation  
Target class: Average over classes

### Scores

Method	AUC	CA	F1	Precision	Recall
Classification Tree	0.949	0.933	0.933	0.934	0.933

