EXPERIMENT-7 (PART-I)

* Tasks 6-

Q: (i) 1) 150

2) ~~Sepallength~~
sepalwidth
petallength
petalwidth
class

3) Iris-setosa

4) Class, sepallength, sepalwidth, petallength, petalwidth are the attributes plotted on the X-Y axes.
The attributes can be changed by selecting the any of the listed ~~any~~ attributes in the drop-down menu at bottom right. The bars can be used to change the attributes.

5) When the "visualize all" button is pressed, all the ~~class~~ attributes graphs are shown individually on a different window.

Q: (ii) 1) A simple instance filter that allows no instances to pass through. Methods in weka.filters with parameters of type Filter. static instances. Filter. useFilter (Instances data, Filter filter) Filters an entire set of instances through a filter and returns the new set.

2) Filtering panel is used to filter a dataset.

③ There are two main types of filters in Weka?
a) Supervised b) Unsupervised

④ In supervised learning, the category/labels data is assigned to are known before computation. So, the labels, classes or categories are being used in order to "learn" the parameters that are really significant for those clusters.
In unsupervised, datasets are assigned to segments, without the clusters being known.

Q: ④ 2) Visualize panel lets you look at a dataset and select different attributes - preferably numeric ones - for the x and y-axes. Instances are shown as points, with different colors for different classes. You can sweep out a rectangle and focus the dataset on the points inside it. You can also apply a classifier and visualize the errors it makes by plotting the "class" against the "predicted class".

Q: ⑤ The ~~visual~~ visualize panel can be used. Also the confusion matrix shows the number of instances of each class present in the data.

EXPERIMENT - 7 (PART-a)

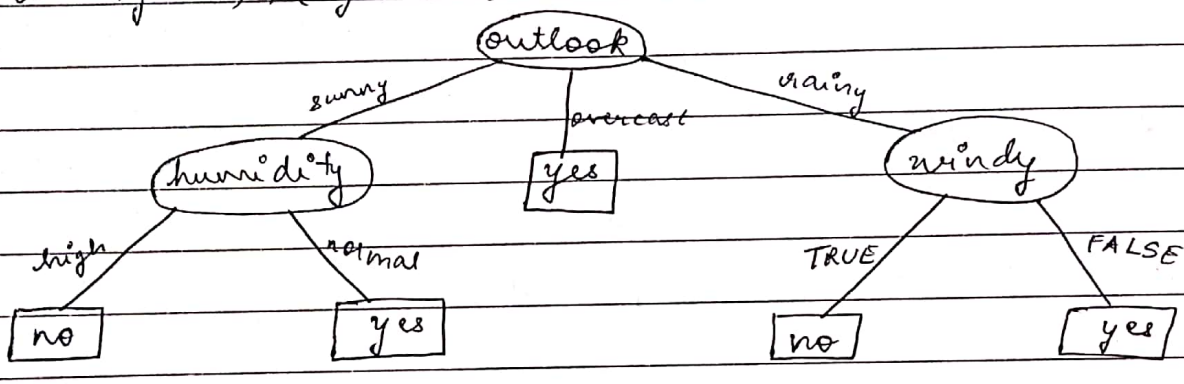
Task:

Q:① 1) There are 5 attributes :- (image Q1 ID3.jpg)

outlook, temperature, humidity, windy, play.

(sunny, overcast, rainy), (hot, mild, cool), (high, normal), (true, false), (yes, no)

2) ①



3)

id	outlook	temperature	humidity	windy	play
1	sunny	hot	high	False	no
2	sunny	hot	high	True	no
3	overcast	hot	high	False	yes
4	rainy	mild	high	False	yes
5	rainy	cool	normal	False	yes
6	rainy	cool	normal	True	no
7	overcast	cool	normal	True	yes
8	sunny	mild	high	False	no
9	sunny	cool	normal	False	yes
10	rainy	mild	normal	False	yes
11	sunny	mild	normal	True	yes
12	overcast	mild	high	True	yes
13	overcast	hot	normal	False	yes
14	rainy	mild	high	True	no

③② Entropy of dataset $H(S) = \sum - p(t) \log_2 p(t)$ $= 0.94$

First attribute : outlook (sunny, rainy, overcast)

$H(\text{outlook} = \text{sunny}) = -\frac{2}{5}\log_2\left(\frac{2}{5}\right) - \frac{3}{5}\log_2\left(\frac{3}{5}\right)$

$= 0.5288 + 0.4422$

$= 0.971$

$H(\text{outlook} = \text{rainy}) = -\frac{3}{5}\log_2\left(\frac{3}{5}\right) - \frac{2}{5}\log_2\left(\frac{2}{5}\right)$

$= 0.971$

$H(\text{outlook} = \text{overcast}) = -\frac{4}{4}\log_2\left(\frac{4}{4}\right) - \frac{0}{4}\log_2\left(\frac{0}{4}\right)$

$= 0$

$\text{Avg. entropy(outlook)} = \frac{5}{14} \times 0.971 + \frac{5}{14} \times 0.971 + \frac{4}{14} \times 0$

$= 0.6936$

$IG(\text{outlook}) = H(S) - \sum p(t) H(t)$

$= 0.94 - 0.6936$

$= 0.246$

Second attribute : temperature (hot, mild, cool)

$H(\text{temp} = \text{hot}) = -\frac{2}{4}\log_2\left(\frac{2}{4}\right) - \frac{2}{4}\log_2\left(\frac{2}{4}\right) = 1$

$H(\text{temp} = \text{mild}) = -\frac{4}{6}\log_2\left(\frac{4}{6}\right) - \frac{2}{6}\log_2\left(\frac{2}{6}\right)$

$= 0.39 + 0.53 = 0.92$

$H(\text{temp} = \text{cool}) = -\frac{3}{4}\log_2\left(\frac{3}{4}\right) - \frac{1}{4}\log_2\left(\frac{1}{4}\right)$

$= 0.3113 + 0.5 = 0.8113$

$IG(\text{temp}) = 0.94 - \left(\frac{4}{15} \times 1 + \frac{6}{15} \times 0.92 + \frac{4}{15} \times 0.8113\right)$

$= 0.0292$

Third attribute : humidity (high, normal)

$$H(\text{humidity} = \text{high}) = -\frac{3}{7}\log_2\left(\frac{3}{7}\right) - \frac{4}{7}\log_2\left(\frac{4}{7}\right)$$

$$= 0.9852$$

$$H(\text{humidity} = \text{normal}) = -\frac{6}{7}\log_2\left(\frac{6}{7}\right) - \frac{1}{7}\log_2\left(\frac{1}{7}\right)$$

$$= 0.5917$$

$$IG(\text{humidity}) = 0.94 - \left(\frac{7}{14} \times 0.9852 + \frac{7}{14} \times 0.59\right)$$

$$= 0.153$$

Fourth attribute : wind (strong, weak)

$$H(\text{wind} = \text{strong}) = -\frac{3}{6}\log_2\left(\frac{3}{6}\right) - \frac{3}{6}\log_2\left(\frac{3}{6}\right)$$

$$= 1$$

$$H(\text{wind} = \text{weak}) = -\frac{6}{8}\log_2\left(\frac{6}{8}\right) - \frac{3}{8}\log\left(\frac{3}{8}\right)$$

$$= 0.84$$

$$IG(\text{wind}) = 0.94 - \left(\frac{6}{14} \times 1 + \frac{9}{14} \times 0.84\right)$$

$$= 0.048$$

3)

② Confusion Matrix

	Yes	No
Yes	5	4
No	3	2

⇒

	Class 1	Class 2
Class 1	TP	FN
Class 2	FP	TN

Confusion matrix gives the accuracy, Recall, Precision, and F-measure.

③ Cohen's Kappa statistic measures interrater reliability. Interrater reliability, or precision, happens when your data raters give the same score to the same data item.

This statistic should only be calculated when:

① Two raters each rate one trial on each sample or

② One rater rates two trials on each sample.

In addition, Cohen's kappa has the assumption that the raters are deliberately chosen. If your raters are chosen at random from a population of raters, use Fleiss' Kappa instead.

The Kappa statistic varies from 0 to 1, where

① 0 = agreement equivalent to chance

② 0.1 – 0.20 = slight agreement

③ 0.21 – 0.40 = fair agreement

④ 0.41 – 0.60 = moderate agreement

⑤ 0.61 – 0.80 = substantial agreement

⑥ 0.81 – 0.99 = near perfect agreement

⑦ 1 = perfect agreement

④ a) TP rate ⇒ You predicted positive and it's true.

b) FP rate ⇒ You predicted positive and it's false.

c) Precision ⇒ $\text{Precision} = \frac{TP}{TP+FP}$

Out of all +ve classes we have predicted correctly, how many are actually positive.

d) Recall => $\text{Recall} = \frac{TP}{TP+FN}$

Out of all the +ve classes, how much we predicted correctly. It should be high as possible

5 ① There are 10 attributes in the dataset- (image 5.png)

② The attributes are :-

RI, Na, Mg, Al, Si, K, Ca, Ba, Fe, Type

③

Classes	instances
build wind float	70
build wind non-float	76
vehic wind float	17
vehic wind non-float	0
containers	13
tableware	9
headlamps	29

④ The "visualize" panel where the "type" class has a clear representation of color assigned to each class.

build wind float -	blue
build wind non-float	red
vehic wind float	light blue
vehic wind non-float	no colour
containers	pink
tableware	green
headlamps	yellow

Q:5) steps :- (image Q3 ~~ONER~~.png)

1. Open Weka Tool
2. Click on Weka Explorer
3. Click on Preprocessing tab button
4. Click on open file button
5. Choose Weka folder in drive
6. ~~Select and click on data ofe~~ Choose weather.arff data set and open file.
7. Click on Classify tab button
8. Click on Choose button and then drop down weka → classifiers → rules → OneR and select it
9. Click on start button

Q:6) a) Present in the image IBK.png

b) Present in the image IBK.png

c) IBK is a lazy classifier.

Q:7) a) Present in the image J48.png

b) Present in the image J48.png

c) J48 is a trees classifier

Q:8) IBK classifier is better.

Q:9) a) Present in the images J48.png and IBK.png

b) Present in images J48 holdout.png, IBK holdout.png

The accuracy level is better in cross-validation strategy than holdout strategy with 3 % levels.

Q:10) a) Present in image Q10(1).png

b) Present in image Q10(2) IBK.png, Q10(2) J48.png

The accuracy is better than the previous one.

Q:12 ① The important attributes are (using J48 classifier results):-

vehicle.arff => Radius Ratio, hr.axis aspect ratio, max. length aspect ratio, scatter ratio, elongatedness, hr.axis rectangularity, scaled variance-major, scaled variance_minor, skewness about_major, skewness about_minor, kurtosis about_major

kr-vs-kp.arff => V3, V10, V15, V16, V21, V32, V33

glass.arff => RI, Na, Mg, Al, K, Ca, Ba, Fe

waveform-5000.arff => x4, x5, x6, x7, x8, x9, x10, x11, x12, x13, x14, x15, x16, x17, x18