DATA

PREPROCESSING (Replace Missing value).

Aim: To implement data preprocessing for missing numerical value and finding the error using point estimation method.

Definition:

point estimation:

- estimate a population parameter.
- May be made by calculating the parameter for a simple.
- May be used to predict value for missing data.

Estimation Error:

Bias difference between expected value and actual value.

bias = $E(\theta^\wedge) - \theta$.

Mean squared Error:

expected value of the squared difference between the estimate and the actual value MSE = $E(\theta^\wedge - \theta)2$.

Algorithm:

Step 1: open the Excel sheet

step2: create 10 records of student mark list.

step3: Delete some data in that list and save as missing data.csv

step4: open the weka tool and open the csv file.

step5: choose filter-unsupervised in that attribute select replacmissing Values then click apply step6: Then it will replace the missing values

step 7: Then using the formula for find the Prediction accuracy and error accuracy

* prediction accuracy = $\frac{\text{obseved value}}{\text{prediced value} *100}$

* Error Accuracy = error / predicted value * 100

Calculation:

⇒ prediction accuracy = observed value / Predicted value * 100

= 419.4 / 396 *100

= 94.4

⇒ error Accuracy = error / predicted value * 100

= -2 3.4 / 419.4 *100

= 5.6

Good

Result: The program executed successfully.

Experiment - 2

Data PREPROCESSING FILTERS APPILED FOR unsupervised ATTRIBUTE

Aim: To convert nominal to Binary values.

Algorithm:

step1: open weka explorer window

step2: click on preprocess tab

step3: open the file or dataset (eg. weather . arff) which is avaiable in your computer after installing weka tool.

step4: observe the datatype of the features or an attributes on your selected dataset. click edit and see the file content

step5: click the data preprocessing and choose the filter unspervised folder, in that choose the attributes. then choose te nominal to binary option.

step6: Apply the converted work and save the file.

step7: click ~~the~~ edit and check the changed attribute with nominal to binary.

Result: Thus the conversion of data type from nomial to binary values is implemented successfully.

Experiment-3

Data Processing - Add Expression.

Aim: To Add the Enpressions x and y using the weka tool.

Algorithm:

step1: open excel and create table with x & y variables and save in csv format.

step2: Open weka tool Enplorer and open the csv file.

step3: And click choose filter and click on the unsupervised then attribute and open the select the Add Enpression.

step4: Then apply filter then go to show properties then give name z and $(a^2+b^2)*z$.

steps5: Then you get the z value.

step6: To see the z value open the edit option.

Result:

The program enecuted and added the enpression successfully.

9/5/2023

Experiment - 4

Attribute selection

Aim: To select the particular attribute best first

procedure:

step1: open weka explorer window

step2: Go to open file and goto this pc then click on program files select the weka tool select the iris.

step3: Now choose filter and supervised and click on the Attribute selection.

step4: Then show properties select the best first

step5: The best are shown on the scree (Attribute)

Result:

The program executed successfully.

9/5/2023

Experiment-5

Linear Regression

Aim: To find a linear regression equation and predict the salary college graduates whose experience is 10 years.

calculation:

x	y	$x_i - \bar{x}$	$y_i - \bar{y}$	$(x_i - \bar{x})^2$	$(y_i - \bar{y})^2$	$(x_i - \bar{x}) * (y_i - \bar{y})$
3	80	-53	-560	2809	313600	29680
4	45	-52	-595	2704	354025	30940
5	60	-51	-580	2601	336400	29580
7	75	-49	-565	2401	·319225	27685
6	55	-50	-585	2500	342225	29250
2	20	-54	-620	2916	384400	33650
1	10	-55	-630	3025	396900	34650
8	85	-48	-555	2304	308025	26640
9	95	-47	-545	2209	297025	25615
11	115	-45	-525	2205	275625	23625

Algorithm:

Step1: open excel sheet and assign x & y

Step2: Next give some values of table in excel sheets

Step3: Do manual calculation in excel sheet for x mean, y mean xi-xi, yi-yi, xi-x^2, yi-ym= and (xi-x) x (yi-y) step4:- check the Answers

step5:- open a New file in csv format and the Enter values in encel sheet

step6:- save as salary in Desktop

step7: open wekatool and open file and ·choose classify and choose linear Regression.

step8:- By click use training set and start

step9:. check the Answers stop the program

$$b_1 = \Sigma \left[(x_i - x) \cdot (y_i - y) \right] / \left[\Sigma (x_i - x)^2 \right]$$

$$= 97 / 1924$$

$$= 10.5087$$

$$b_2 = y - b_1 * y$$

$$= 59 - 10.5087 * 5.6$$

$$= 0.1515$$

$$y = 0.1515 + 10.5087 \times 10$$

$$= 105.24$$

Result:

Thus the program Executed successfully.

10/5/2023

Experiment - 6

Classification using Naive Bayessian Classifer Algorithm -(multi class classification Problem).

Aim: To find the classification prediction using naive bayesian classifier algorithm the new person will buy a comper or not.

Algorithm:

step1: create a table in the excel sheet and save as a csv file.

step2: Then later open the weka tool and open the computer.csv file

step3: we need save as compuler arff

step4: later Goto edit, see the viewer table, selected all the rows and delete it.

step5: And then Now click add instance. again open the computer.arff file then open classify in the menu bar.

step6: click bayes and click naive bayes click cross validation and click start.

step7: we will get Answer of correctly classified Intances and incorrectly classified instances.

Age	Income	student	credit-rating	Buys-computer
<=30	High	No	Fair	No
<=30	High	No	excellent	yes
31....40	High	NO	Fair	yes
>40	Medinum	NO	Fair	yes
>40	Low	yes	Fair	yes
>40	Low	yes	exclelent	~~yes~~ NO
31..40	Low	yes	Exclent	Yes
<=30	Medimun	NO	Fair	NO
<=30	Low	Yes	Fair	NO
>40	medium	yes	Fair	Yes
<=30	Medium	yes	excellent	Yes
31..40	Medium	NO	excellient	yes
31...40	High	yes	Fair	yes
>40	Medinum	NO	Excelent	NO

E = age <=30, income =medium, student = yes , credit rating = fair

E1 = age <=30

E2 = Income = Medium

E3 = student = yes

E4 = credit rating = faid

$$P(yes|E) = \frac{P(E1|yes)\,P(E2|yes)\,P(E3|yes)\,P(E4|Yes)}{P(E)}$$

P(yes) = 9/14 = 0.643	P(no) = 5/14 = 0.357
P(E1/yes) = 2/9 = 0.222	P(E1|No) = 3/5 = 0.6
P(E2/yes) = 4/9 = 0.444	P(E2|No) = 2/5 = 0.4
P(E3|yes) = 6/9 = 0.667	P(E3|No) = 1/5 = 0.2
P(E4|yes) = 6/9 = 0.667	P(E4|No) = 2/5 = 0.4

$$P(yes|e) = \frac{0.222*0.444*0.6667*0.667*0.643}{P(E)} \qquad \frac{0.028}{P(E)}$$

$$P(no|E) = \frac{0.6*0.4*0.2*0.4*0.357}{P(E)} \qquad \frac{0.006}{P(E)}$$

Experiment - 7

Naive Bayes Classifier Algorithm

working of Naive Bayer's classifer (single class classification)

Aim: To find the classification prediction using naive bayesian classifer Algorithm on weather conditions whether player should play or not.

Algorithm:

step1: create a table in xshell sheet and save file as csv file.

step2: Then open weka tool, select weather nominal data set

step3: keep only outlook and play attribute, remove other attributes save the file as player.

step4: Goto edit, in the viewer delete all the records.

step5: And click add instance, then add only sunny, after save the new file as playertest.

step6: Again the open player file, click classify, again click bayes then naivebayes click the cross folder then click start.

step7: Note the corrected and incorrected values.

step8: Now click the supply test set, open the playertest file, then click more options setup output predictions as on system

step9: Then click start. find out the prediction result with use of right click.

S.NO	outlook	play
1	Sunny	no
2	sunny	no
3	overcast	yes
4	rainy	Yes
5	rainy	yes
6	rainy	no
7	overcast	yes
8	sunny	no
9	sunny	Yes
10	rainy	yes
11	sunny	yes
12	overcast	yes
13	overcast	yes
14	rainy	no

weather conditions

weather	Yes	NO
overcast	4	0
Rainy	3	2
sunny	2	3
Total .	9	5

Applying Bayes theorem:

P(Yes| sunny) = P(sunny|Yes) * P(Yes/ P(sunny)

P(sunny) = 2|9 = 0.22

P(sunny) = 0.36

P(Yes) = 0.64

P(Yes|sunny)

= 0.22 * 0.64 / 0.36

= 0.39

$$P(A/B) = \frac{P(A \cap B)}{P(B)}$$

P(NO|sunny) = P(sunny|NO) * P(NO) | P(sunny)

P(sunny|NO) = 3|5 = 0.60

P(NO) = 0.36

P(sunny) = 0.36

So P(NO| sunny) = 0.60 * 0.36 / 0.36

= 0.60

11/5/2022

Result: Thus the program executed succesfully, hence on sunny day, player cannot play the game.

Experiment-8

DESISION TREE ALGORITHM

Aim: Given the training data, buiid a decision tree and predict the class of the following new. ex: age<=30, income=medium, student=yes

Algorithm:

step1: Credte csv file with name of compulesuy with above mentioned.

step2: Save as arff File in weka tool.

step3: open classify and choose rules in that choose decision tree with name of J48.

step4: start the program.

step5: print the visualize tree.

step6: Apply supply set (open compulesbuy, go to edit, delete all instance, add new instance age<=30, income=medium, student=yes, buys-computer empty) save it another name.

step7: run with supply set and find the answer.

Age	Income	student	Buys-computer
<=30	High	No	NO
<=30	High	NO	NO
31...40	High	NO	yes
>40	medium	NO	Yes
>40	Low	yes	yes
>40	Low	yes	NO
31...40	Low	yes	yas
<=30	Medimum	NO	NO
<=30	~~medium~~ Low	yes	NO
>40	medium	Yes	y.S
<=30	medium	yes	yes

31....40	medium	NO	yes
31..40	High	yes	yes
>40	medium	NO	NO

* $$\text{Gain}(D, A) = \text{Entropy}(D) - \sum_{j=1}^{v} \frac{|D_j|}{|D|} \text{Entropy}(D_j)$$

$I(s_{yes}, S_{NO}) = I(9, 5) = -9/14 \log 2(9/14) - 5/14 \log 2(5/14) = 0.94$

Entropy (age) = 8/14

= 0.6935

Gain (age) = income low (3 yes 1 no)

entropy (income) =

= 0.285714 + 0.393428 + 0.231714

= 0.9108

Gain (income) =

0.94 − 0.9108

= 0.0292

age

<=30

income	student	class
high	no	NO
high	no	NO
medium	no	yes
low	yes	yes
medium	Yes	yes

income	student	class
high	no	Yes
low	yes	Yes
medium	no	Yes
high	yes	yes

>40

Income	student	class
medium	no	Yes
low	Yes	yes
low	yes	No
medium	yes	yes
medium	no	NO

Again the same process is needed for the other branch of age :

Entropy (income)

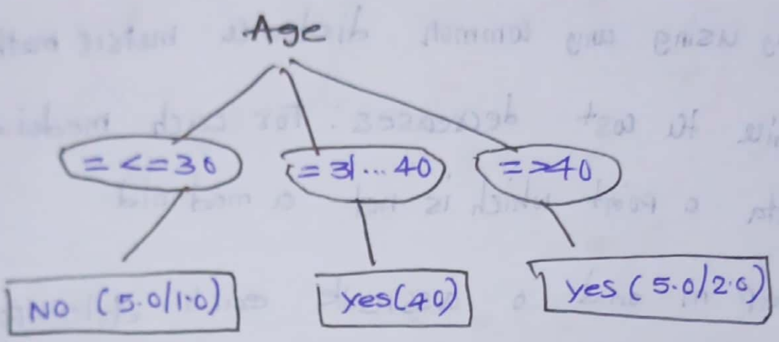
$= 3/5\,(0.9182) + 2/5\,(1) = 0.55 + 0.4 = 0.95$

Gain (income) $= 0.97 - 0.95 = 0.02$

Entropy (student) $= 0.95$

Gain (student) $= 0.97 - 0.95$

$= 0.02$



Result: Thus the program executed succesfully, Hence person will not buy a computer :

13/5/2023

Experiment-9

K-Medoids Algorithm.

Aim: To prove the k-medoids Algorithm using weka tool.

Algorithm:

step1: initialize select k random points out of the n data points as the medoids.

step2: Associate each data point to the closet medid by using any common distance metric methods.

step3: while the cost decreases: For each medoid m, for each data o point which is not a med oid.

step4: swap m and o assicate each data point to the closet medoid, and recompute the cost.

step5: save the csv in enecel sheet and enecute in the wek tool.

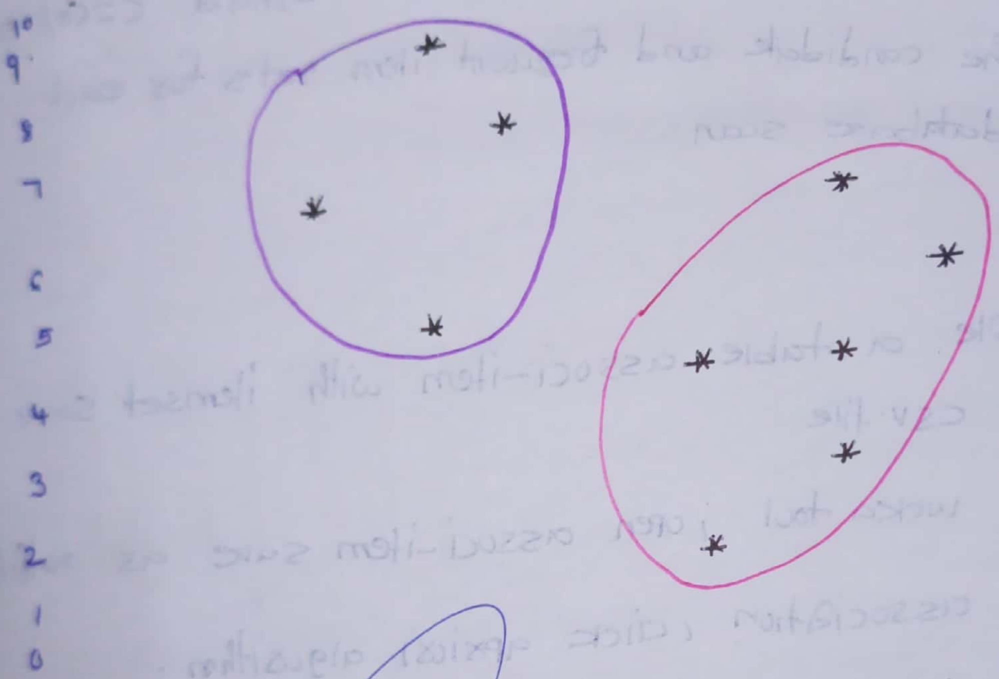
X	Y	Dissimilarity C1	Dissimilarity C2
8	7	6	2
3	7	3	7
4	9	4	8
9	6	6	2
8	5	-	-
5	8	4	6
7	3	5	3
8	4	5	1
7	5	3	1
4	5	-	-

The cost in k-medoids algorithm is given as

$$C = \sum_{ci} \sum_{pi \in ci} |pi - ci|$$

That formula tell that Distance = $|x_1 - x_2| + |y_1 - y_2|$.

* $O(k*(n-k)^2)$.



Result: Thus the program was enecuted succesfully using wek tool.

15/5/2023

Experiment -10

ASSOCIATION RULE MINING

APRIORI ALGORITHM.

Aim: Trace the result of using the apriori algorithm on the grocery store example with support threshold s= 33.34% and confidence threshold c=60% show the candidate and frequent item sets for each database scan.

Algorithm:

step1: create a table associ-item with itemset save as csv file

step2: open weka tool, open associ-item save as arff file

step3: open association, click apriori algorithm.

step4: click show properities, change support 0.33 and confidence 0.60

step5: Start the program

step6: result will display.

Transaction ID	Items
T_1	HotDogs, Buns, ketchup
T_2	HotDogs, Buns
T_3	HotDogs, coke, chips
T_4	chips, coke
T_5	chips, ketchup
T_6	HotDogs, coke, chips

confidence: The confidence of a rule is

$$\text{conf}(X \rightarrow Y) = \text{supp}(X \cup Y) / \text{supp}(X)$$

Transcation ID	hotdogs	buns	ketchup	coke	chips
T1	T	T	T		
T2	T	+			
T3	+			T	T
T4				+	T
T5			T		T
T6	T			T	T

calculation:

support threshold = 33.34%.

=> threshold is at least 2transcations

confidence = 0.66

Association rules:

Itemset	support	confidence
Hot Dogs, Buns	2/6 = 33.33	2/4 = 50
Buns, HotDogs	2/6 = 33.33	2/2 = 1000
Hotdogs, Coke	2/6 = 33.33	2/4 = 50
Coke, HotDogs	2/6 = 33.33	2/3 = 66.66
Hot Dogs, chips	2/6 = 33.33	2/4 = 50
chips, Hot Dogs	2/6 = 33.33	2/4 = 50
Coke, chips	3/6 = 50	3/3 = 100
chips, Coke	3/6 = 50	3/4 = 75
HotDogs -> Coke n chips	2/6 = 33.33	2/4 = 50
coke => chips n Hotdogs	2/6 = 33.33	2/3 = 66.66

chips->Hotdogs n coke	2/6 =33.33	2/4 = 50
Hotdogs n coke->chips	2/6 = 33.33	2/2 =1
chips n Hotdogs->coke	2/6 = 33.33	2/2 = 1
cokenchips >Hot dogs	2/6 = 33.37	2/3 = 66.66

Result: Thus the program have been executed succesfuly through Assocation rule mining.

Generated sets of large itemsets:

-> size of set of large items L(1): 11

size of set of large itemsets L(2): 22

size of set of large itensets L(3): 14

size of set of large itemsets L(4): 3

Best rules found:

coke =T 3 ==> chips=T 3 con

buns=T 2==> hotdogs=T 2 conf:(1)

hotdogs=T chips=T 2 ==> coke=T 2 conf: (1)

Minimum support: 0.25 (1 instances)

Minimum metric < confidence >: 0.9

Number of cycles performed 15

Experiment-11

Classic agglomerative , hierachical clustering methods using with linkage criteria.

Aim: To find the cluster using hieararchical Agglomerative cluster with linkage criteria and find dendrogram.

Algorithm:

step1: open contact uns file in the weka tool.

step2: click cluster and choose hierichal algorithm .change the property link type with single

step3: start the program ,find the results and dendrogram-

step4: change link type completed and start te program, find the results and dendrogram.

step5: change link type average and start the program find the results and dendrogram .

single linkage =

$$L(r,s) = \min (D(x_{ri}, x_{sj})$$

Average linkage =

$$L(r,s) = \frac{1}{n_r n_s} \sum_{i=1}^{n_r} \sum_{j=1}^{n_s} D(x_{ri}, x_{sj})$$

compute linkage :

$$L(r,s) = \max (D(x_{ri}, x_{sj})$$

Euclidean Distance.

$X = (a,b)$ and $Y=(c,d)$

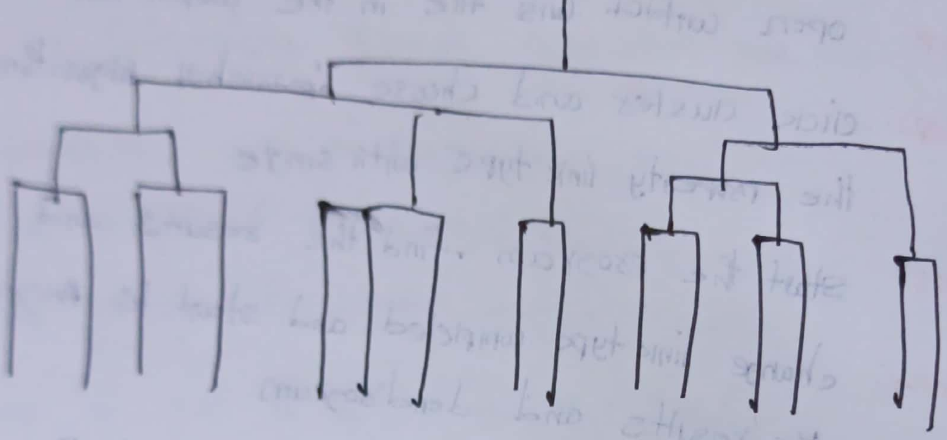
The euclidean distance between x and y

$$\sqrt{(a-c)^2+(b-d)^2}$$

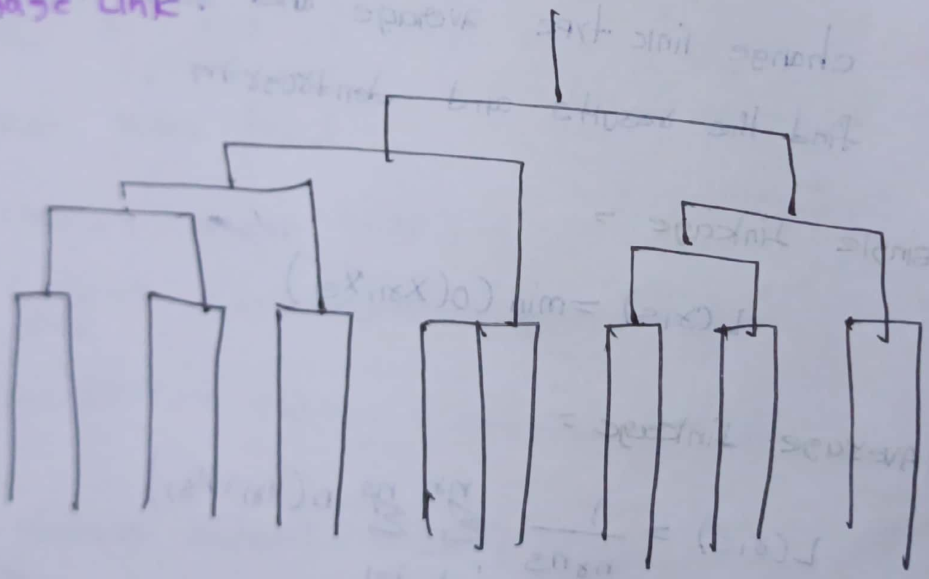
Comparision of link.

Link	Tot record 24	cluster 0	cluster 1
single	20-4	83	17
Completed	15-9	63	37
Average	15-9	63	37

Complete Link:



Average Link:



Result: Thus the program executed succesfuly through classic Agglomerative, hierarchical clustering,

16/5/2025

Experiment -12

comparison various algorithms in classification

Aim: To briefly described three algorithms in terms of how it works, key algorith palameters will be highitened and the algorithm will be demostrated in the weka explorer interfac.

→ Navie Bayes

→ Decision Tree

→ k-Nearest Neibours.

Algorithm:

step1: open the weka GUI chooser

step2: click the explorer button to open the weka explorer.

step3: Load the Ionosphere dataset from data/ionosphere.arff file.

step4: click "classify to open the classify tab.

Decision tree algorithm

step1: click the "choose " button and select "REPTree" under the 'tree" group

step2: click on the name of the algorithm to riview the algorithm configuration.

step3: click "ok" to close the algorithm configuration.

step4: click the "start" button to run the algorithm on the Ionospha dataset.

k-nearest Neighbors algorithm.

step1: click the "choose" button and slect "IBK" under the "lazy"L group

step2: click on the name of the algorithm to review the agorithm configuration.

step3: click "ok" to close the algorithm configuration.

step4: click the "start" button to run the algorithm on the Ionosphera dataset.

Algorithm	Accuracy
Navie bayes	82
decision tree	89
KNN	86

Result: Thus the program was enecuted succesfuly using weka tool.

16/5/2023

Experiment - 13

FP GROWTH ALGORITHM USING WEKA.

Aim: To briefly describe about the FP growth Algorithm using weka and explorer in the weka tool.

Algorithm:

step1: open the data file in weka explorer.

step2: It is presaumed that the resuired data fields have been discretized. in this example it is age attribute

step3: clicking on the associate. tab will bring up the interface for association rule algorithm

step4: we will use FP - growth algorithm. This is the default algorithm.

step5: Inorder to change the parameters for the run (example, support, confidence etc).

step6: we click on the text box immediately to the right of the choose button.

Data set:

shopping.arff
@ relation shopping
@ attribute milk {yes,no}
@ attribute bread {yes,no}
@ attribute honey {yes,no}
@ attribute ghee {yes,no}
@ attribute jam {yes,no}
@ data
yes, yes, no, no, yes
yes, no, yes, no

Yes , yes ,no, yes ,no

yes, no, yes , no, no

no, yes, Yes, no, no

Yes, no, Yes, no, no

Yes, Yes, Yes, no, yes

Yes, Yes, Yes, no, no

Result: Thus the program was executed succesfuly using weka tool.

18/8/2023