Aim: Creation of a Data warehouse.

→ Build Data Ware house / Dart Mart (using open source tools like pentaho Data Integration Tool, pentaho Business Analytics, or other data ware house tools like Microsoft - SSIS, Information, Business objects etc...

→ Design multi-dimensional data models namely star, snowflake and fact constellation schemas for any one enterprise.

(Eg:- Banking, Insurance, Finance, Healthcare, manufacturing, Automobile, sales etc).

→ write ETL 'scripts and implement using data warehouse tools.

→ perform various OLAP operations such slice, dice, rollup, drillup and pivot.

Objectivies:

Data warehousing is a technique of gathering and analyzing data from many sources to get valuable, business, insights. Typically a data ware house integrates and analyzes business data from many sources. Data ware housing is a vital component of business intelligence.

Preprocess :-

The data that is collected from the field contains many unwanted things that leads to wrong analysis. Those the data must be preprocessed to meet the requirements of the type of analysis you are setting. This is the done in the preprocessing module.

Classifiers :-

Classifiers in WEKA are the models for Predicting nominal or numeric quantities. The learning schemes available in WEKA include decision trees and lists, instance - based classifiers, classificers include bagging, boosting, Stacking, error - correcting output codes and locally weighted learning.

WEKA :-

WEKA(waikota Environment for Knowledge Analysis) is a popular suite of machine learning software written in Java, developed at the university of waikota, NewZealand. weka is free Software available under the GNO General Public license.

weka - an open Source Software provides tools for data pre processing implementation of several machine learning algorithms and visualization tools so that you can develop machine learning techniques and apply them to real - world data mining problems.

Explorer: It is an environment for exploring data Explorer consists of several tabs. They are:

→ Preprocess: It is the first step in machine learning is to preprocess the data. It is used to select the data file, process it and make it fit for applying the various, machine learning algorithms.

→ classify: The classify tab provides you several machine learning algorithms for the classification of your data Such as Linear Regression, logistic Regression

→ cluster: under the cluster tab there are several clustering algorithm provided - such as simple k Means, Filtered Clusterer, Hierarchical cluster.

→ Associate: under the Associate tab you would find Apriori filtered Associator and Fp Growth.

- Select Attributes Tab:

Select Attributes allows you feature selection based on several algorithms such as classifier, subset Eval, Principal components.

- visualize Tab:

The visualize option allows you to visualize your processed data for analysis.

→ simple CLI: It provides a simple command - line interface and allows direct execution of weka commands.

→ Experimenter: It is an environment for performing experiment and conducting statistical tests between learning scheme. ~~knowledge~~ +

→ Knowledge Flow: It is a java - Beans based interface for setting up running machine learning experiments.

Trees J48 classifier:

It is an algorithm to generate a decision tree that is generated by C4.5. It is also known as a statistical classifier. For decision tree classification, we need a database. of

weather nominal:

In weka, attributes can be nominal or numeric. The value of a nominal attribute is represented by a word: sonny, overcast and rainy for the outlook attribute; yes and no for the play attribute.

Steps Required:-

1. open WEKA you can see 5 tabs on the right side of the application. these are: Explorer, Experimenter, knowledge flow, work bench, simple CLI.

2. Click on " Explorer"

3. on preprocess. click on "open file".

4. Go to " c:\program files\ weka - 3-8-6\ data", select " weather.nominal.arff" and click on open.

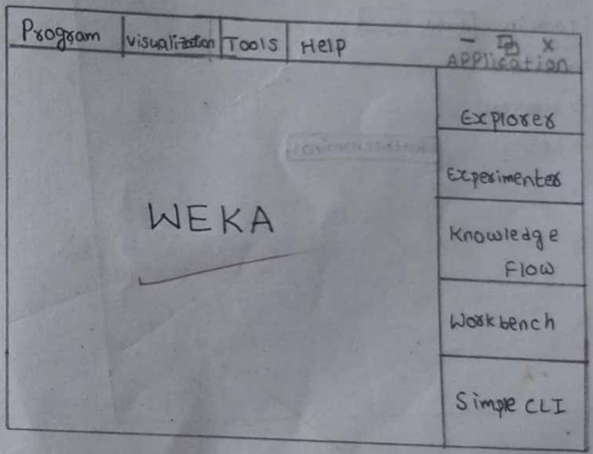
5. click on " classify and then click on choose.

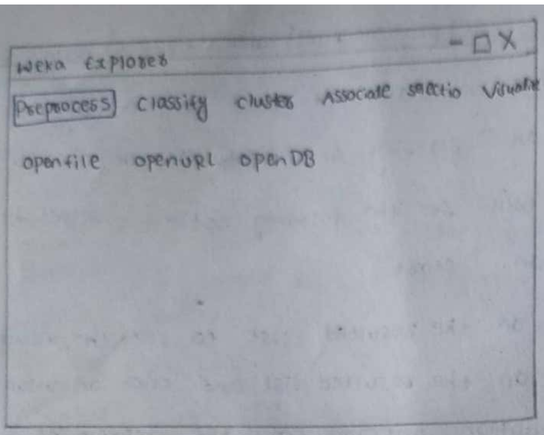
6. you will see the following options. select j48 & click on " start".

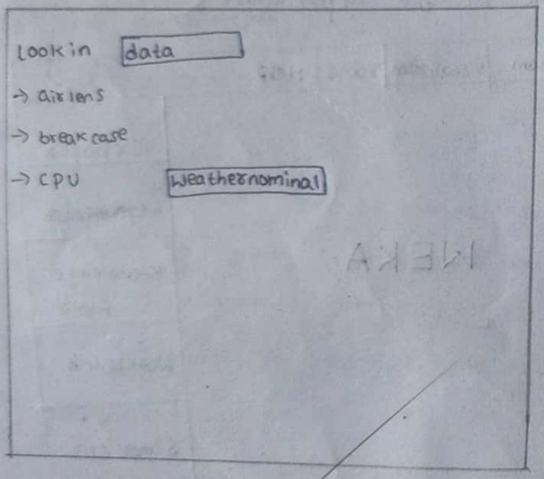
7. click on the resalted list to see the visual.

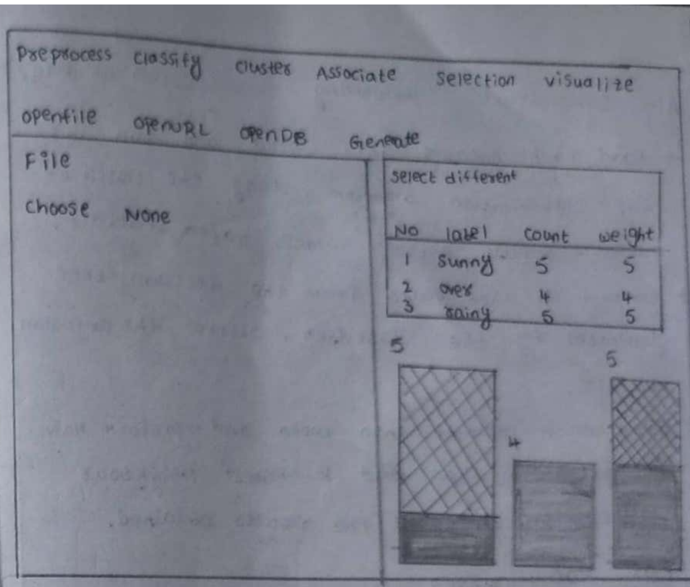
8. click on the resulted list and click on visualize tree option. out come of the experiment:-

⇒ Out come of the Experiment:-









No	label	count	weight
1	Sunny	5	5
2	over	4	4
3	Rainy	5	5

13/10/22

Aim: Demonstrate performing classification on data.

→ Load each dataset into weka and run id3, J48 classification algorithm. study the classifier output. compute entropy values, kappa statistic.

→ Extract if then rules from the decision tree generated by the classifier, observe the confusion matrix.

→ Load each dataset into weka and perform Naive-bays. classification and K-Nearest Neighbour classification. Interpret the results obtained.

→ Plot ROC curves

→ Compare classification result of ID3, J48, Naive-Bayes and K-NN classifiers for each dataset and reduce which classifier is Performing best and poor for each dataset and justify.

Objectives:

The Ultimate objective of classification is to relate a variable of interest with observed variables. The actual variable of interest is meant to be of "Qualitative" type. The algorithm required for performing the classification is known as the classifier.

Zero R:-

→ Zero R is the simplest classification method which relies on the targest and ignores all predictors.

→ Zero R classifier simply predicts the majority category.

→ Although there is no predictability Power in zero R it is useful for determining a baseline performance as a benchmark for other classification methods.

One R:-

→ This method is used in the sequential learning algorithm for leaving the rules.

→ It returns a single rule that covers atleast some examples.

→ However, what makes it really powerful is its ability to create relations among the attributes given. Hence covering a larger hypothesis space.

Explorer: It is an environment for exploring data.

Simple CLI: It provides a simple command-line-interface and allows direct execution of weka commands.

Experimenter: Knowledge Flow:

It is a Java-Beans based interface to setting up and running machine learning experiments.

Experimenter:

It is an environment for performing experiment and conducting statistical tests between learning scheme.

Preprocess:

It is the first step in machine learning to preprocess the data. It is used to select the data file preprocessing and make it fit for applying the various machine learning algorithms.

Classify:

The classify tab provides you several machine learning algorithms for, the classification of your data such as linear Regression, logistic Regression.

Test options:

Before you run the classification algorithm, you need to set test options. set test options in the 'Test options' box. The test options that available are.

1. Use training set:-

Evaluates the classifier on how well it predicts the class of the instances it was trained on.

2. supplied test set:

Evaluates the classifier on how well it predicts the class of a set of instances loaded from a file. clicking on the 'set..' button brings up a dialog allowing you to choose the file to test on.

3. cross validation:

Evaluates the classifier by cross-validation using the number of folds that are entered in the 'Folds' text field.

4. percentage spilt:

Evaluates the classifier on how well it predicts a certain percentage of the data, which is held out for testing. The amount of data held out depends on the value entered in the '%' field.

Steps Required:

1. Open weka you can see 5 tabs on the right side of the application. These are explorer, experimentor, knowledge flow, workbench, Simple CLI.
2. Click on 'explorer'.
3. You can see classify tab click on the classify botton.

4. You can observe choose, test options etc..

5. In test option you can see cross - validation folds. set it as 10.

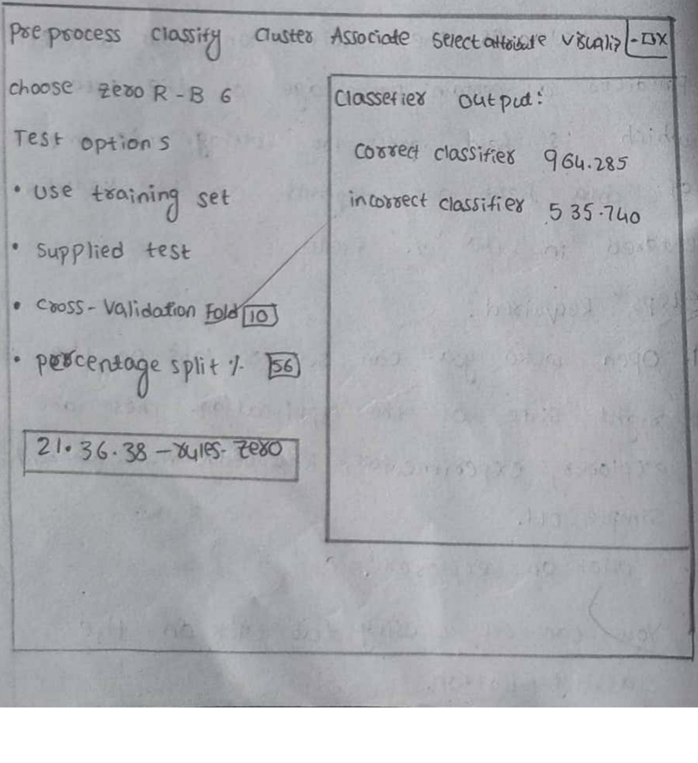
6. Right click on choose option, then select the zero R algorithm or one R algorithm.

7. Then click start button.

8. Zero R algorithm or one R algorithm will execute and it gives the output.

Output:-

Zero R



Preprocess classify Cluster Associate Select attribute visualiz -□X

choose zero R - B 6

Test option s

- use training set
- supplied test
- cross - Validation Fold 10
- percentage split % 56

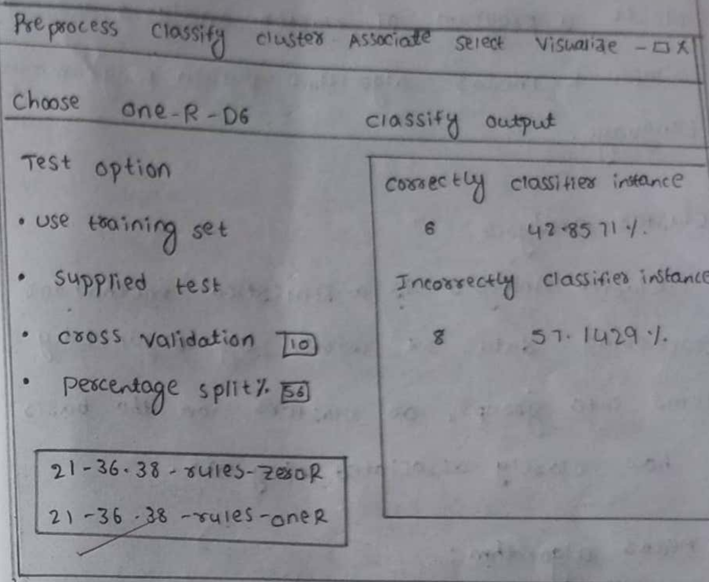
21. 36. 38 - rules- zero

Classefier output:

correct classifier 9 64.285

incorrect classifier 5 35.740

One R :-



Preprocess Classify cluster Associate select Visualize -□X

Choose one-R-D6 classify output

Test option

- use training set
- supplied test
- cross validation 10
- percentage split% 56

21-36.38 - rules-zeroR

21 -36 .38 -rules-oneR

correctly classifier instance

6 42.8571 %.

Incorrectly classifier instance

8 57. 1429 %.

13/10/22

Write a program of cluster analysis using simple k-means algorithm python programming language.

Cluster Analysis:

Cluster Analysis is a statistical method for Processing data. It works by organizing items into groups, or clusters on the basis of how closely associated they are.

K-means algorithm:

K-means algorithm is a simple two steps clustering process. The first step is cluster assignment and the second one is the move centroid step. ~~However~~, this unsupervised algorithm can easily create, implement and handle massive datasets.

Steps involved in k-means Algorithm:

Step 1: Select the number k to decide the number of clusters.

Step 2: Select random k points or centroids.

Step 3: Assign each data point to their closest centroid, which will form the predefined k clusters.

Step 4: Calculate the variance and place a new centroid of each cluster.

Step 5: Repeat the third steps, which means reassign each datapoint to the new closest centroid of each cluster.

Step 6: If any reassignment occurs, then go to step-4, else go to FINISH.

Step 7: The model is ready.

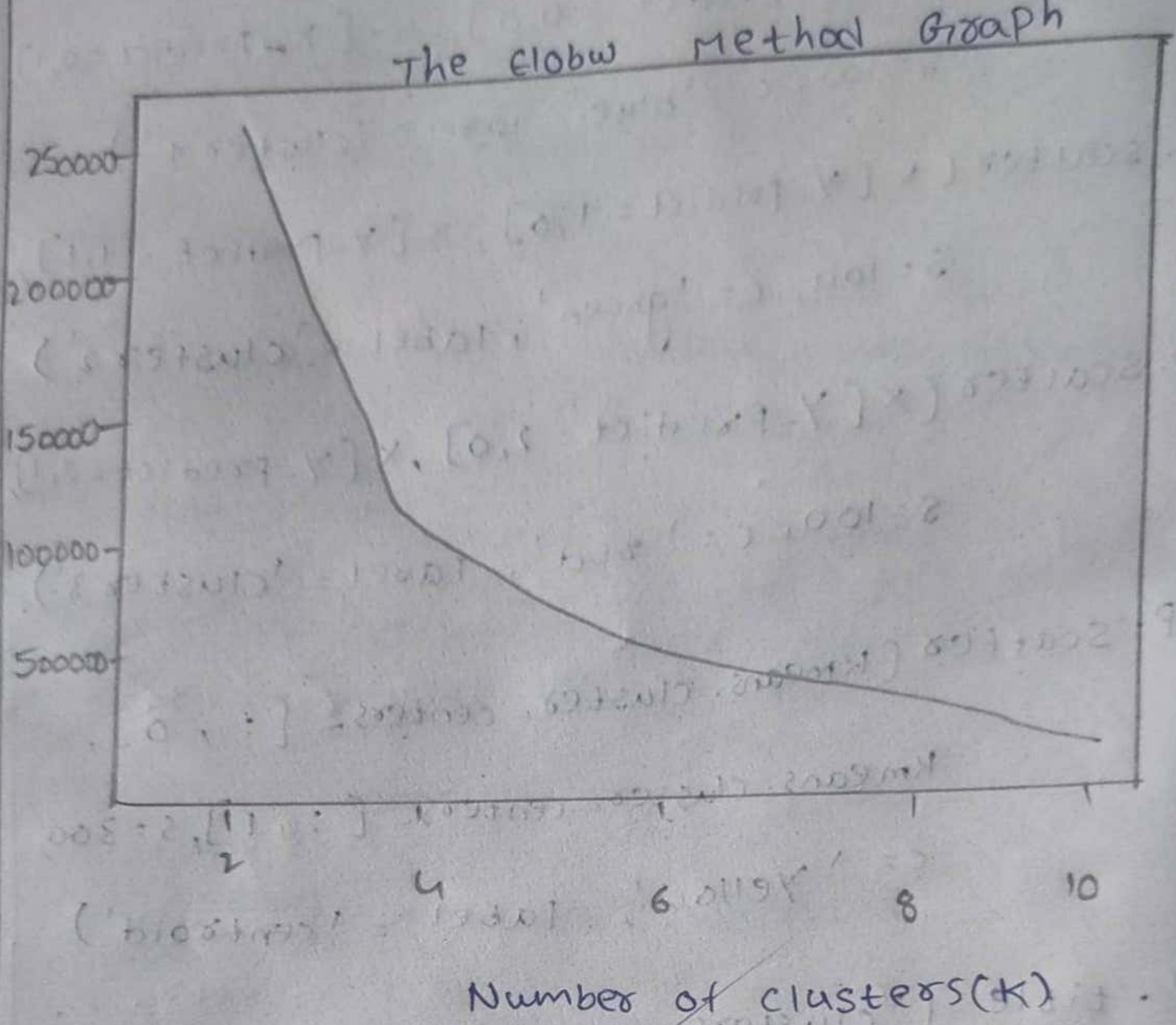
K-means Algorithm asing python programming:

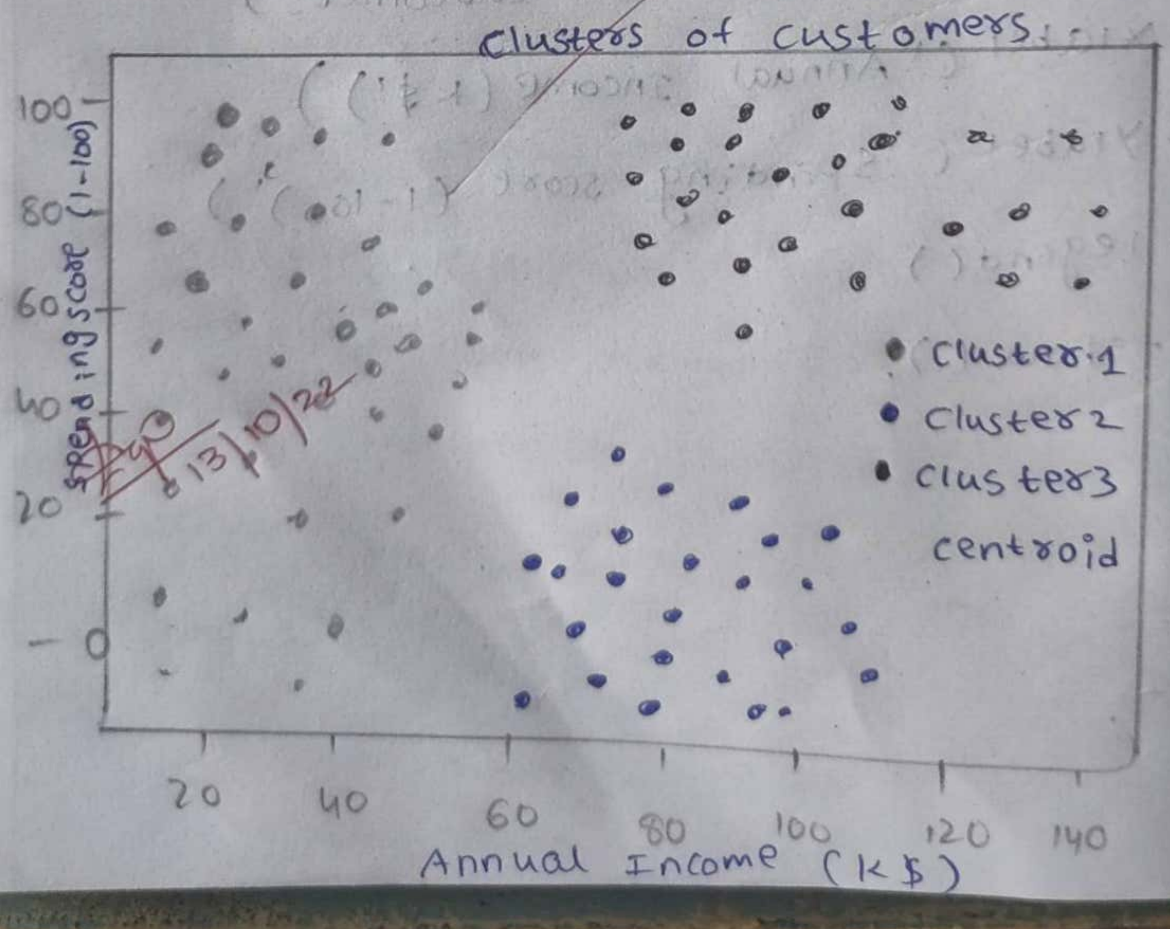
```
import numpy as nm
import matplotlib.pyplot as mtp
import pandas as pd
dataset = pd.read_csv('/content/sample_data/Mall_
                          customers.csv')
x = dataset.iloc[:, [3,4]].values
from sklearn.cluster import KMeans
wcss.list = []
for i in range (1,11):
        kmeans = kmeans(n_clusters = i, init =
                       'k-means++', random_state=42)
        kmeans.fit(x)
        wcss_list.append(kmeans.inertia_)
mtp.plot(range(1,11), wcss_list)
mtp.title('The Elobw Method Graph')
mtp.xlabel('Number of clusters(k)')
mtp.ylabel('wcss_list')
mtp.show()

kmeans = KMeans ~~mode~~ (n_clusters=5, init='k-means++',
          random_state=42)
```

```
y_predict = kmeans.fit_predict(x)
mtp.scatter(x[y_predict = 0,0], x[y_predict=0,1],
            s=100, c='blue', label='cluster1')
mtp.scatter(x[y_predict = 1,0], x[y_predict = 1,1],
            s=100, c='green', label='cluster2')
mtp.scatter(x[y_predict = 2,0], x[y_predict=2,1],
            s=100, c='red', label='cluster3').
mtp.scatter(kmeans.cluster_centers_[:, 0],
            kmeans.cluster_centers_[:, 1], s=300,
            c='yellow', label ='centroid')
mtp.title('clusters of customers')
mtp.xlabel('Annual Income (k$'))
mtp.ylabel('Spending score (1-100)')
mtp.legend()
mtp.show().
```

Out put


The Elobw Method Graph


Clusters of customers

13/10/22