**Program: General/Intelligent Systems/Cybersecurity**

'....;;-,. .:-. -'.. -

**Level: Third**

**Course Code: 02-24-01203**

**Total points: 100**

**Course Title: Data Science Tools &Software Professor name: Dr. Mohamed Abd EI-Hafeez**

,, .

**ALEXANDRIA**

I\IER ITY

**..::.U4.JI .J ..::.t,. WI'--:!lS**

**Assignment #3**

Ql) Given the following dissimilarity matrix between three objects

O l*P(X)* 0 2

41

1

r 4 2 0

1. Sketch a dendrogram using single linkage strategy?
2. Sketch a dendrogram using complete linkage strategy?
3. Write down the corresponding Python steps of parts (a) include a proper cut-off points?
4. Using the first and second object as initial two centers for clusters C1 and C2respectively. What is the hard membership of the third object in C1 and C2
5. What is the center of the new formed cluster in part (d) if x1=[l 

Q2) Hierarchical, Medoid-based and Density-based clustering

* 1. Starting with the two cluster medoids: mt= (1 1) and



'.

5 ...... ···:···········'···..····· ! ........,. ...·······-'.····· .....

..

.

'.

.'

'.

.'

.

..

..

.

'

.

4 ........:........---:.... ···•· .............. ·:· ........

' '

'

'

'

.

. . .

'

'

'

. .

. .

'

.

..................(.............................].........

'

. . .

'

.

. .

. .

'

2 .......................... : .........:.........:.........

.

. .

1 · · ................i........ i · ........;...... ··'.·........

' '

.

.

. .

'.

o - -- ---- -- -

m2= (3 3 . What is the cost of replacing ml with (2,2)?

* 1. Cluster above data using complete linkage

strategy and sketch the corresponding dendrogram

* 1. Sketch k-dist graph for *k=3* and estimate suitable value for Epson.

Apply DBSCAN on the given dataset starting from {4,4) using k=l and Epson=v'Z

* 1. Estimate cutting edges for the dendrogram in part b to produce the same clusters produced by DBSCAN in part d

')(,N3

0 2 3

y

4 5 6

Q3) Given the following dissimilarity matrix between five objects

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | **0** | 1 | ***5*** | **6.4** | **7.4** |
| **1** | **0** | **4.2** | **5.7** | **6.7** |
| **P(X)** | **5** | **4.2** | **0** | **1.4** | **2.5** |
|  | **6.4** | **5.7** | **.4** | **0** | **1.1** |
|  | **7.4** | **6.7** | **2.5** | **LI** | **0** |

* + 1. Draw a dendrogram using single linkage strategy?
    2. Select a proper cut off point and output the corresponding clustering?
    3. Draw a dendogram using complete linkage strategy?
    4. Write down the corresponding Python steps of parts a,b,c?
    5. comment on the resulting dendrograms in a and c?
    6. What is the complexity of the algorithm in terms of number of objects?

Q4) Given Dataset X of size nxd

1. The size of similarity matrix between elements of X is-----
2. The size of membership matrix of X ink clusters is-----
3. The size of Linkage matrix output by matlab is------
4. The number of internal nodes of dendrogram is ------
5. in order to determine the pair of clusters that is going to be merged at the *r* + *1 level,* ­

pairs of clusters have to be considered.

1. The complexity of heirarchical clustering is---------

Q5) Select (True/False) to answer the following questions:

* 1. Number of clusters is a required parameter for hierarchical algorithms [
  2. Cancer Diagnosis is a clustering application [
  3. Image segmentation is a classification problem
  4. Noise is a random error or variance in a measured variable

Q6) Compute Accuracy, Sensitivity and Specificity of the following classification results?

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| ID | Income | Marital Stanis | Refund | Actual Class | Predicted las |
| I | 95K | Married | Yes | -1 | I |
| 2 | 120K | Single | Yes | -1 | - I |
| 3 | 140K | Single | 0 | I | I |
| 4 | 80K | Married | No | -1 | - I |
| 5 | 160K | Divorced | Yes | -1 | I |
| 6 | IOOK | Married | Yes | -I | - I |
| 7 | 90K | Single | Yes | -I | I |
| 8 | 75K | Divorced | 0 | -I | - I |
| 9 | 170K | Divorced | 0 | I | I |
| 10 | 125K | Single | 0 | I | - I |

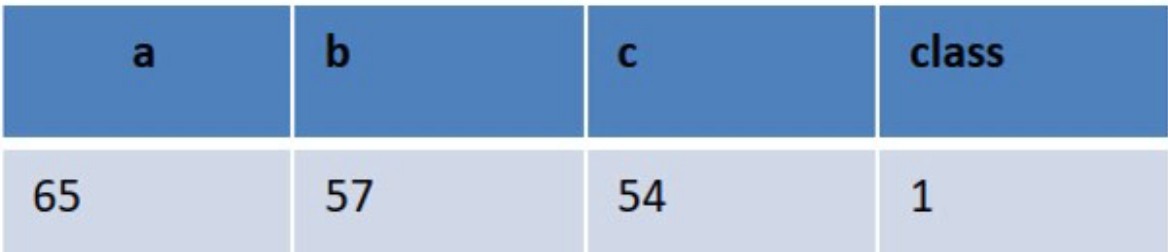
Q7) Consider the following k-nearest table of data points. What are the accuracy, specificity precision, recall, f-measure and sensitivity of k-nearest neighbor on them assuming the actual labels of xl,x3, and x5 is 1 and x2,x4 and x6 is -1 using leave-one-out cross validation method (k=l)?

|  |  |
| --- | --- |
| Object xi  xl x2 x3 | x3\_x .x6  **X** .x6.x2  **X** .x2.x3 |
| x4 x6.x3.xl  **X** x4.x2.xl  x6 x4. ·I. ·2 | |

Q8) Assume we have the following two models and the following test data to be classified. Compute accuracy, sensitivity and specificity of Ml and M2

Ml: if a>60 and b >60 then class= 1 otherwise class=-1

M2: if a>SS and c >65 then class= 1 otherwise class= -1



|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| 45 | 65 |  | 48 | -1 |
| 70 | 46 |  | 62 | -1 |
| 48 | 91 |  | 87 | 1 |
| 61 | 33 |  | 38 | 1 |
| 66 | 59 |  | 76 | -1 |
| 58 | 84 |  | 53 | 1 |

Q9) Predict the house price index if age= 38 and loan $55,000 using kNN regressor with k=3 (lab work)

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Age | Loan | House | Price | Index | Distance |
| 25 | $40,000 | 135 | 102000 | | |
| 35 | $60,000 | 256 | 82000 | | |
| 45 | $80,000 | 231 | 62000 | | |
| 20 | $20,000 | 267 | 122000 | | |
| 35 | $120,000 | 139 | 22000 | | |
| 52 | $18,000 | 150 | 124000 | | |
| 23 | $95,000 | 127 | 47000 | | |
| 40 | $62,000 | 216 | 80000 | | |
| 60 | $100,000 | 139 | 42000 | | |
| 48 | $220,000 | 250 | 78000 | | |
| 33 | $150,000 | 264 | 8000 | | |

## Q10) Consider the data points at the right. Using visual inspection compute Accuracy Sensitivity and Specificity when kNN with *k* =l is applied to this

**14**

**03**

04

**6005**

012

**013**

Os

**07**

**1 2 9**

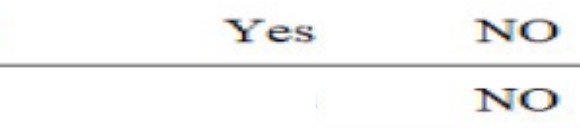
a

**10**

**0** 11

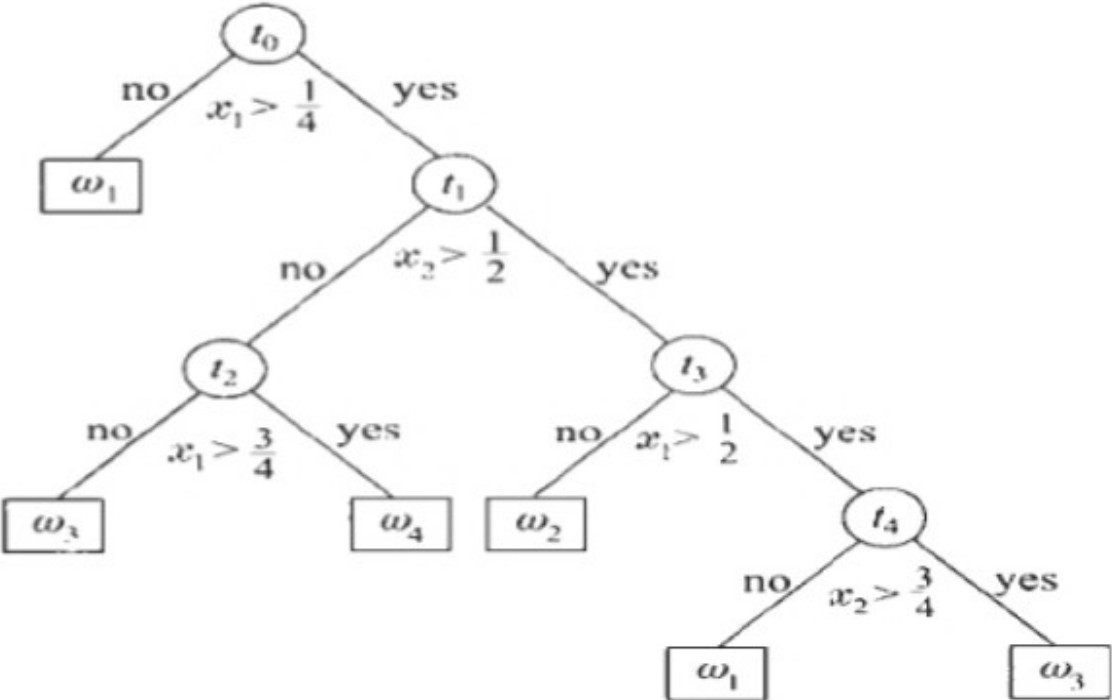
dataset using leave-one-out cross validation. Note: create a table contains id actual and predicted labels where circle represents negative sample and square represents positive sample before computing the metrics.

Q11) Regarding the following Decision Tree generate induction rules for all classes vv1-vv4\_



**Yes**

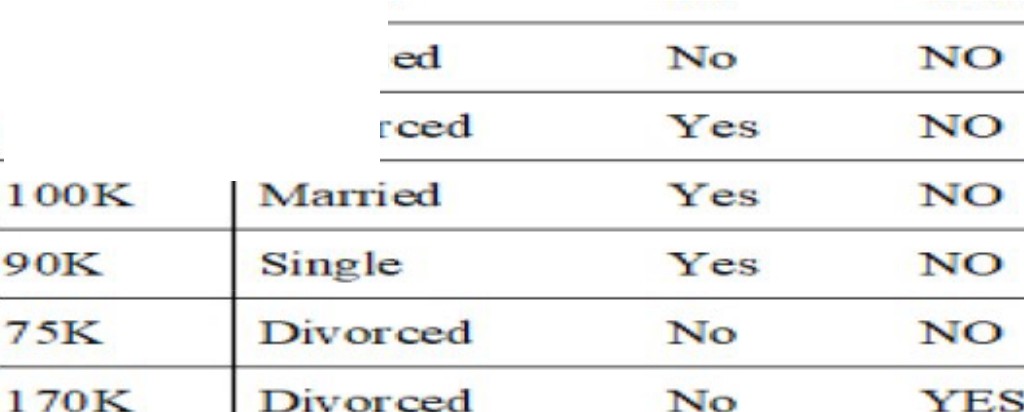
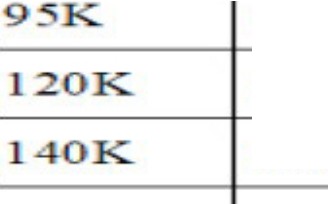
What is the class that will be assigned for[x1x2] equals [0.4 0.6]?



Ql2) Given the following training set where the column with caption Cheat is the class label, compute the information gain of Refund and Marital Status and select the best of them to start with and complete your decision tree by choosing the unselected attribute followed by the income attribute as your next choices to build your model. Compare the perfonnance of decision tree to kNN using python in terms of accuracy, sensitivity and specificity (lab work)



**Income "Marital S atus Cheat**



**"Married**

**Single**

**14-0K**

**80**

**160K**

**100**

**Sing!**

**YES**

**"Married**

**Divorced**

**125K Single 0 YES**

Q13) Construct a box plot for the following data set.

3,5,8,8,9,11,12,12,13,13,163,5,8,8,9,11,12,12,13,13,16

Q13) add label and information to nodes and edges in the following example and show the output (lab work)

>>> g.add\_edges\_from([(1 ,2) ,(1 ,3)])

>>> g.add\_node('a')

>>> g.number\_of\_nodes() # *aiso g.order{)*

*4*

>>> g.number\_of\_edges() # *aiso g.size{)*

2

>>> g.nodes()

[1, 2, 3, 'a']

>>g>.dges()

[(1, 2), (1, 3)]

>>>g.neighbors(1)

[ 2, 31

>>> g.degree(1)

2

Q14) reverse the direction of edges and rerun the following example to show the output

>>> dg = nx.DiGraph()

>>> dg.add\_weighted\_edges\_from([(l,4,0.5), (3,1,0.75)))

>>> dg.out\_degree(l,weighted=True) 0.5

>>> dg.degree(l,weighted=True)

1.25

>>> dg.successors(l)

[4]

>>> dg.predecessors(l)

[3]

Q14) if two edges {l,2,0.75) and (2,1,0.25) is added to after the first step of creating directed graph in the following code fragments, show the new output of each of the following code

>>> dg = nx.DiGraph()

>>> dg.add\_weighted\_edges\_from([(l,4,0.S), (3,1,0.75)])

>>> dg.out\_degree(l,weighted=True) 0.5

>>> dg.degree(l,weighted=True) 1.25

>>> dg.successors(l) [4]

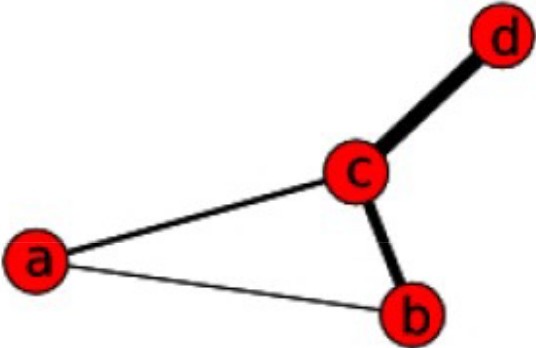
>>> dg.predecessors(l)

[3]

QlS) replace the dataset in the following code by any of your choice and print nodes, edges count and average degree (lab work)

hartford = nx.read\_edgelist('hartford.txt', create\_using=nx.DiGraph() ,nodetype=int)

Q16) add new node e with edge to c and a having weight 0.9 and 1.3 respectively and recompute the shortest path between a and d and the minimum spanning tree.

>>> import networkx as nx

>>> g = nx.Graph{)

>>> g.add\_edge{'a','b' ,weight=0.1)

>>> g.add\_edge{'b','c' ,weight=l.5)

>>> g.add\_edge{'a','c' ,weight=l.0)

>>> g.add\_edge{'c','d' ,weight=2.2)

>>> print nx.shortest\_path(g,'b','d')

['b', 'c', 'd']

>>> print nx.shortest\_path(g,'b','d' ,weighted=True)

['b', 'a', 'c', 'd']

Q17) Given the following dissimilarity matrix between three objects

o 1

41

*P(X)* 1 0 2

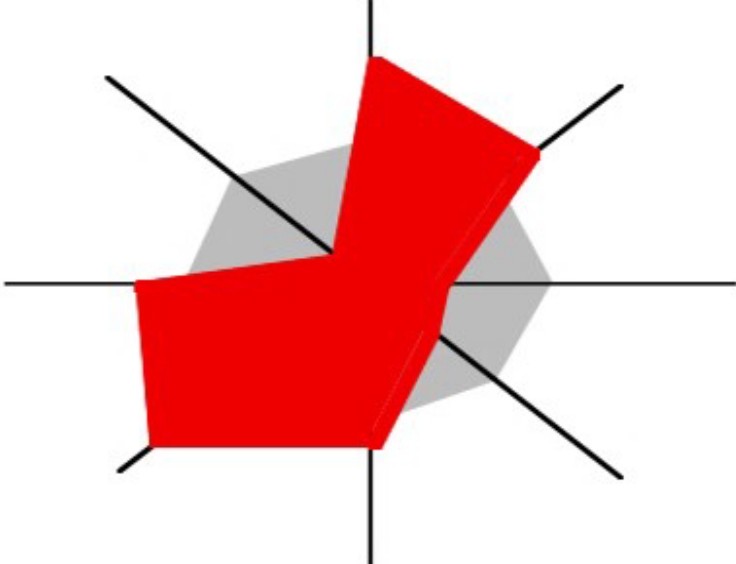
r 4 2 0

The height of the dendrogram produced by any linkage strategy for the above dataset is

1. 2 b)1 c) >2 d) 3
2. Sketch a dendrogram using complete linkage strategy?
3. Write down the corresponding Python steps of parts (a) include a proper cut-off points?
4. Using the first and second object as initial two centers for clusters C1and C2 respectively. What is the hard membership of the third object in C1 and C2

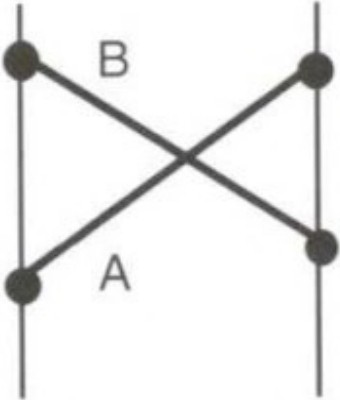
**Q17) select the best answer (use e for all of the above and f for none of the above):­**

1. Import pandas as pd
   1. Df= pd.DataFrame('x':[1,2,3],'y':[40,S0,60],'z'=[70,80,90])
   2. Print df[df.x>2][['x','y']]
   3. Df.to\_csv('filename.csv')
2. **Star Plots is used to examine the relative values for**
   1. **a single data point**
   2. **multi data points**
   3. **statistical indices**
   4. **frequencies**



1. **the process of enlarging the apparent size (not the physical size) of something to show a relative attribute is called**
   1. boxing
   2. Magnification
   3. Zooming
   4. Mapping
2. Parallel coordinate plots can be used to examine the relative values for
   1. a single data point
   2. multi data points
   3. statistical indices
   4. frequencies

b



Price Number of bedrooms

1. the use of pictorial images to make actions, objects, and concepts in a display easier to find, recognize, learn, and remember is called
   1. Iconic representation
   2. Spidergram
   3. Magnification Chernoff Faces
   4. What is the library used in the following line of code
2. the following function is defined in

lmplot(x='Attack', y='Defense', data=dfl, fit\_reg=False)

* 1. Seaborn
  2. matplotlib
  3. sklearn
  4. network

1. What is the type of the chart produced by the following line of code:-

g s, s. faictorplot(x='Ty e 1',

y= A ac ,

data=dfl,

hue='Stage, col= Stage, kind= s ar ')

1. **Box plot**
2. **scatter plot**
3. **histogram**
4. **line chart**
5. What is the center of a cluster containing X1=[l 
   1. [6 15]
   2. [2.5 5]
   3. [2 5]
   4. [1 7]
6. The size of similarity matrix between elements of dataset X is
   1. 2IXI

# IX I 2

* 1. IXl3

# IXI

1. The size of membership matrix of X of size n objects in k clusters is
   1. n2k
   2. k2
   3. nk
   4. nk2
2. The complexity of hierarchical clustering (HC) is (than) kmeans and DBSCAN
   1. Less
   2. greater
   3. equal to
   4. less or equal
3. Number of clusters is a required parameter for
   1. DBSCAN
   2. HC
   3. kmeans
   4. HC and DBSCAN
4. Which of the following is false regarding the kNN classifier

*a}* multi-class

1. efficient
2. simple
3. lazy
4. Root Mean Square Error is a performance metric for
   1. Regression
   2. classification
   3. clustering
   4. association rules mining
5. Dimensionality reduction techniques are primarily used for:
   1. Data visualization b) Data compression

c) oise removal d) b and c

1. Which of the following is an application of dimensionality reduction?
   1. Customer relationship management b) Image retrieval

c) Face recognition d) gene expression analysis

1. Image classification involves:
   1. Classifying unlabeled documents

c) Removing noise from documents

1. Gene expression microarray analysis deals with:
   1. Classifying novel samples into disease types
   2. Analyzing high-dimensional microarray data
   3. Reducing noise in gene expression data
   4. reducing the dimensionality
2. Storing and retrieving documents efficiently

d) Data analysis and visualization

1. Feature selection differs from feature reduction in that it always produces
   1. all original features b) a subset of original features

c) another features d) three features only for visualization

1. The MRMR feature selection algorithm aims to
   1. Maximize redundancy between features b) Minimize relevance of target features

c) Maximize relevance and minimi7e redundancy d) minimize the variance

1. Which library in Python can be used for PCA?
   1. umPy b) Pandas c) Scikit-leam d) Matplotlib