|  |
| --- |
| Build Data Warehouse/Data Mart for a given problem statement |
| Design dimensional data model i.e. Star schema, ii), Snowflake schema and Fact Constellation schema (if applicable) |
| Identifying the source tables and populating sample data |
| iii) Create tables for schema and insert 5 records in dimensional table and multiple records in fact tables. |

Q2.



Create tables for star schema and insert records in it.

Q3. 



Create tables for star schema and insert records in it.

Q.4 To perform data exploration and data cleaning in python

1) Load the libraries and give list of libraries required for data exploration for following task

2) Use the given file

3) Read the file –select appropriate file read function according to data type of file( refer link 1)

4) Display attributes in the data set. Display first 10 rows of data set.

5) Describe the attributes name, count no of values, and find min, max, data type, range, quartile, and percentile. Give box plot and find outliers. (link 1 and link2-descibe function)

6) Give visualization of statistical description of data – in form of histogram, scatter plot, pie chart.

7) Give correlation matrix

8) Identify missing values and outlier and fill them with average.

Q.5 WAP to implementation of Apriori algorithm and write the disadvantages of apriori algorithm.

Q.6 WAP to implement K-means algorithms for following points for k=2

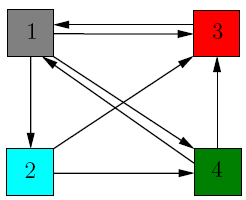
p1(2,3), p2(2,4), p3(8,9) p4(7,8), p5(1,2), p6(8,8), p7(9,7), p8(2,2)

Q.8. WAP to implement of Linear Regression for following data

| **Air Velocity (cm/sec)** | 30,70,110,150,180,220,260,300,350,390 |
| --- | --- |
| **Evaporation Coefficient**  **(mm2 /sec)** | 0.18, 0.37, 0.35, 0.78, 0.56, 0.75, 1.18, 1.36, 1.17, 1.65 |

* Find b1 = Σ [ (xi - x)(yi - y) ] / Σ [ (xi - x)2]
* Find b0 = y - b1 \* x
* Put the values in formula ŷ = b0 + b1x
* Calculate R-squared value.
* Find the value evaporation coefficient for air velocity =250.

Q.9. WAP to implement page rank algorithm for following Graph using Vk= M\*V k-1. Print final K, V, M.



Q.10. Use weka for classification and give decision tree Press the Explorer button on the main panel and load the weather dataset and answer the following questions

* 1. How many instances are there in the dataset?
  2. State the names of the attributes along with their types and values.
  3. What is the class attribute?
  4. In the histogram on the bottom-right, which attributes are plotted on the X,Y-axes? How do you change the attributes plotted on the X,Y-axes?
  5. How will you determine how many instances of each class are present in the data
  6. What happens with the Visualize All button is pressed?
  7. How will you view the instances in the dataset? How will you save the changes?
  8. What is the difference between and attribute filter and an instance filter?

Q.11 Use weka for classification and give decision tree. Press the Explorer button on the main panel and load the iris dataset and answer the following questions

* 1. How many instances are there in the dataset?
  2. State the names of the attributes along with their types and values.
  3. What is the class attribute?
  4. In the histogram on the bottom-right, which attributes are plotted on the X,Y-axes? How do you change the attributes plotted on the X,Y-axes?
  5. How will you determine how many instances of each class are present in the data
  6. What happens with the Visualize All button is pressed?
  7. How will you view the instances in the dataset? How will you save the changes?
  8. What is the difference between and attribute filter and an instance filter?

Q.11 Load the ‘weather.arff’ dataset into Weka and run classification algorithm**.** Answer the following questions

* + List the attributes of the given relation along with the type details
  + observe the classifier output and answer the following questions
    1. Draw the decision tree generated by the classifier
    2. Compute the entropy values for each of the attributes
    3. What is the relationship between the attribute entropy values and the nodes of the decision tree?
  1. Draw the confusion matrix? What information does the confusion matrix provide?
  2. Describe the Kappa statistic?
  3. Describe the following quantities:
     + TP Rate
     + FP Rate
     + Precision, Recall

Q.12 Use weka for clustering **using bank-data.csv**

**Q.13 Create a data cube using union all from following table structure and insert (25-50 records)**

**CREATE TABLE sales (**

**brand VARCHAR NOT NULL,**

**segment VARCHAR NOT NULL,**

**country VARCHAR NOT NULL**

**quantity INT NOT NULL,**

**PRIMARY KEY (brand, segment, country)**

**);**

**Insert records for 5 different brands, 5 different segments, 5 different countries.**