SAVEETHA SCHOOL OF ENGINEERING
SAVEETHA INSTITUTE OF MEDICAL AND TECHNICAL SCIENCES
COMPUTER SCIENCE AND ENGINEERING

Engineer to Excel

CSA16-Data Warehousing and Data Mining

LAB PRACTICAL QUESTIONS

1. Consider the group of 12 sales price records that has been sorted as follows: 5, 10, 11, 13, 15, 35, 50, 55, 72, 92, 204, and 215. Partition them into three bins by each of the following methods.

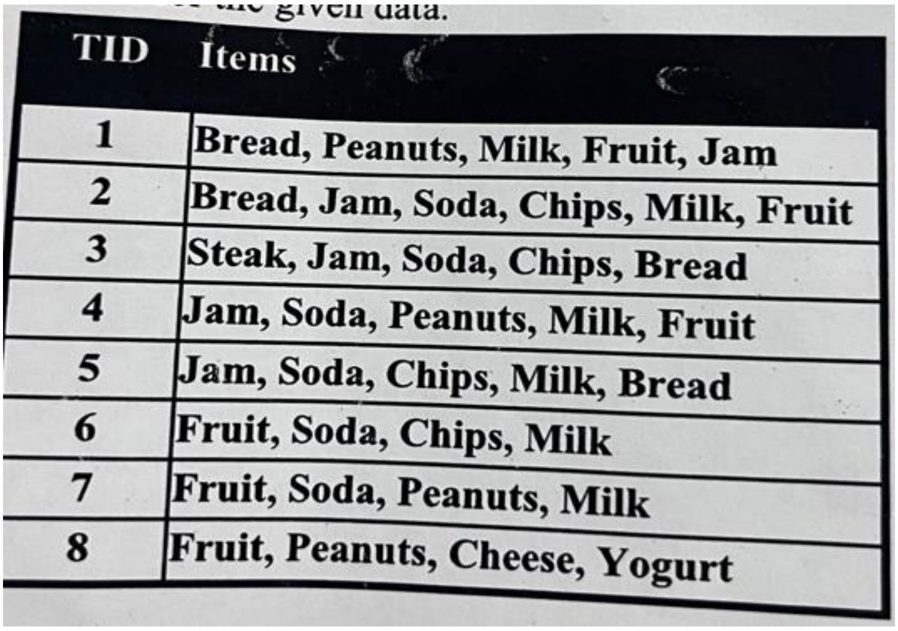
   (a) equal-frequency (equi-depth) partitioning

   (b) equal-width partitioning

   (c) Clustering.

   Implement the same using R.

2. A gadget factory has been quite successful for the past 10 years and Ms.Marry, the manager of the company wondering whether to expand the factory this year or not. The cost to expand factory is $2M. With no expansion, expected revenue is $4M if the economy stays good; while only $1.5M if the economy is bad. If manager expands the factory, expected to receive $7M. if economy is good and $3M if economy is bad. Assume that there is a 45% chance of a good economy and a 55% chance of a bad economy. Draw a Decision Tree showing these choices.
3. Apply Apriori Algorithm for given database below by assuming Minimum support = 2. Implement using WEKA for the given data.



given data.

TID	Items
1	Bread, Peanuts, Milk, Fruit, Jam
2	Bread, Jam, Soda, Chips, Milk, Fruit
3	Steak, Jam, Soda, Chips, Bread
4	Jam, Soda, Peanuts, Milk, Fruit
5	Jam, Soda, Chips, Milk, Bread
6	Fruit, Soda, Chips, Milk
7	Fruit, Soda, Peanuts, Milk
8	Fruit, Peanuts, Cheese, Yogurt

4. Use following group of data: 200, 300, 400, 600, 1000
   (a) min-max normalization by setting min = 0 and max = 1
   (b) z-score normalization using the mean absolute deviation instead of standard deviation
   (c) normalization by decimal scaling

5. Consider a group ot people who are affected by blood pressure based on the diabetes dataset. Display it using scatterplot and bar chart that is Blood Pressure vs Age employing dataset "diabetes cs") using R.

6. Analyze the dataset "diabetes.csv" how the diabetes trend is for different age people, using Linear Regression and Multiple Regression.

7. Suppose a database has five transactions. Let minimum support= 50% (2) and minimum confidence = 80%.

Transactions	Items
T1	(M, O, N, K, E, Y)
T2	(D, O, N, K, E, Y)
T3	(M, A, K, E)
T4	(M, U, C, K, Y)
T5	(C, 0, 0, K, I, E)

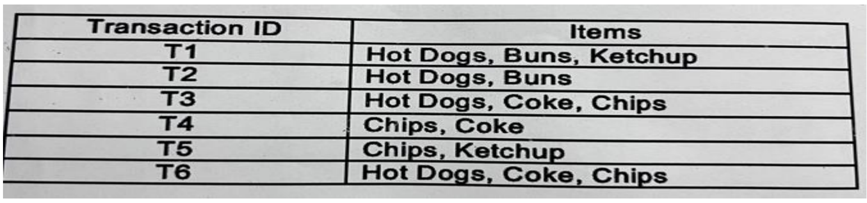
   • Implement using WEKA and find all frequent item sets using Apriori algorithm
   • Also draw FP-Growth Tree

8. Prediction of Categorical Data using Decision Tree Algorithm through WEKA using any datasets. a) Tree b) Preprocess c) Logistic.

9. Suppose that the data for analysis includes the attribute age. The age values for the data tuples are (in increasing order) 13, 15, 16, 16, 19, 20, 20, 21, 22, 22, 25, 25, 25, 25, 30, 33, 33, 35, 35, 35, 35, 36, 40, 45, 46, 52, 70.
   Can you find (roughly) the first quartile (Q1) and the third quartile (Q3) of the data?

10. Download the Dataset "water" From R dataset Link. Find out whether there is a linear relation between attributes "mortality" and "hardness" by plot function. Fit the Data into the Linear Regression model. Predict the mortality for the hardness=88.

11. Create the dataset using ARFF file format:



Transaction ID	Items
T1	Hot Dogs, Buns, Ketchup
T2	Hot Dogs, Buns
T3	Hot Dogs, Coke, Chips
T4	Chips, Coke
T5	Chips, Ketchup
T6	Hot Dogs, Coke, Chips

a. Find the frequent item-sets and generate association rules on this. Assume that minimum support threshold (s = 33.33%) and minimum confident threshold (c = 60%).
b. List the various rule generated by apriori and FP tree algorithm, mention whether it is accepted or rejected.

12. Prediction of Categorical Data using Rule base classification and decision tree classification through WEKA using any datasets. Compare the accuracy using two algorithm and plot the Graph.

13. Imagine that you have selected data from the All Electronics data warehouse for analysis. The data set will be huge! The tollowing data are a list of All Electronics prices for commonly sold items (rounded to the nearest dollar). The numbers have been sorted: 1, 1, 5, 5, 5, 5, 5, 8, 8, 10, 10, 10, 10, 12, 14, 14, 14, 15, 15, 15, 15, 15, 15, 18, 18, 18, 18, 18, 18, 18, 18, 20, 20, 20, 20, 20, 20, 20, 21, 21, 21, 21, 25, 25, 25, 25, 25, 28, 28, 30, 30, 30.

    (i) Partition the dataset using an equal-frequency partitioning method with bin equal to 3
    (i) Apply data smoothing using bin means and bin boundary.
    (i) Plot Histogram for the above frequency division

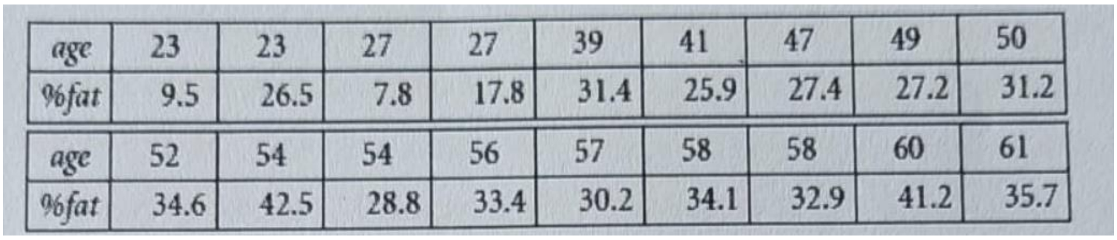
14. Two Maths teachers are comparing how their Year 9 classes performed in the end of year exams. Their results are as follows:
    Class A: 76, 35, 47, 64, 95, 66, 89, 36, 8476,35,47,64,95,66, 89,36,84
    Class B: 51, 56, 84, 60, 59, 70, 63, 66, 5051,56,84,60,59,70,63,66,50

    (i) Find which class had scored higher mean, median and range.
    (ii) Plot above in boxplot and give the inferences

15. Consider a Binary classification model that can be used to predict whether one or more ads on the website will be clicked or not. The models are used to optimize the ad inventory on websites by selecting which ads will have a better chance of being clicked.
16. Consider that Many businesses use cluster analysis to identify consumers who are similar to each other so they can tailor their emails sent to consumers in such a way that maximizes their revenue. Consider a business may collect the following information about consumers:
    - Percentage of emails opened
    - Number of clicks per email
    - Time spent viewing email

    Using these metrics, a business can perform various cluster analyses to identify consumers who use email in similar ways and tailor the types of emails and frequency of emails they send to different clusters of customers. Compare the performance of the applied clustering algorithm.

17. Suppose that a hospital tested the age and body fat data for 18 randomly selected adults with the following results:

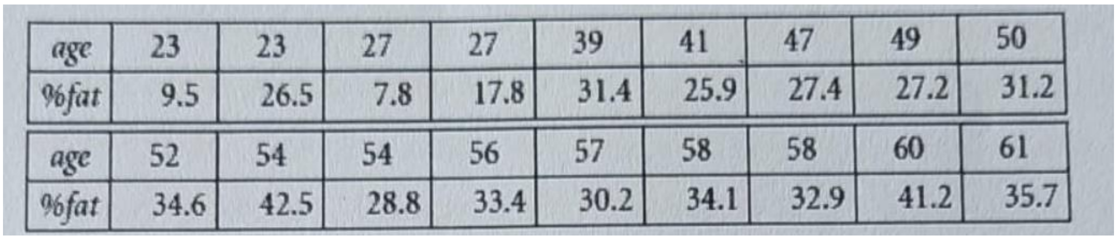


age	23	23	27	27	39	41	47	49	50
%fat	9.5	26.5	7.8	17.8	31.4	25.9	27.4	27.2	31.2
age	52	54	54	56	57	58	58	60	61
%fat	34.6	42.5	28.8	33.4	30.2	34.1	32.9	41.2	35.7

(a) Calculate the mean, median, and standard deviation of age and %fat.
(b) Draw the boxplots for age and /ofat.
(c) Draw a scatter plot and a q-q plot based on these two variables.

Perform the above functions using R – tool

18. Suppose that a hospital tested the age and body fat data for 18 randomly selected adults with the following results:

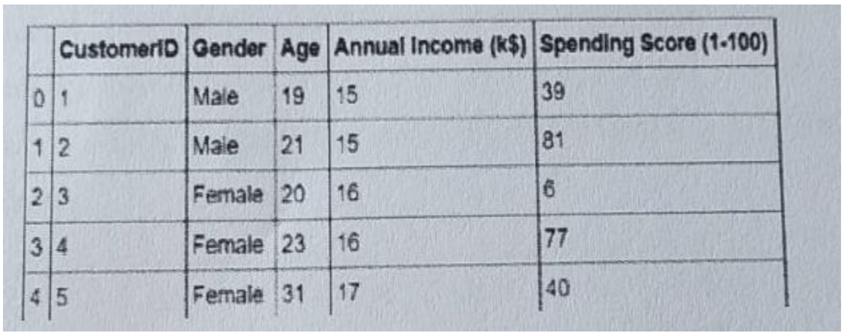


age	23	23	27	27	39	41	47	49	50
%fat	9.5	26.5	7.8	17.8	31.4	25.9	27.4	27.2	31.2
age	52	54	54	56	57	58	58	60	61
%fat	34.6	42.5	28.8	33.4	30.2	34.1	32.9	41.2	35.7

(i) Use min-max normalization to transform the value 35 for age onto the range [0.0, 1.0].
(ii) Use z-score normalization to transform the value 35 for age, where the standard Deviation of age is 12.94 years.
(iii) Use normalization by decimal scaling to transform the value 35 for age.

Perform the above functions using R – tool

19. Consider that you are owning a supermarket mall and through membership cards, you have some basic data about your customers like Customer ID, age, gender, annual income and spending score. For the above scenario, the Problem Statement was You want to understand the customers who can easily converge [Target Customers] so that the data can be given to the marketing team and plan the strategy accordingly. For the above scenario prepare a dataset and perform Clustering Analysis to segment the customers in the Mall. There are clearly Five segments of Customers based on their Annual Income and Spending Score namely Usual Customers, Priority Customers, Senior Citizen Target Customers, and Young Target Customers.

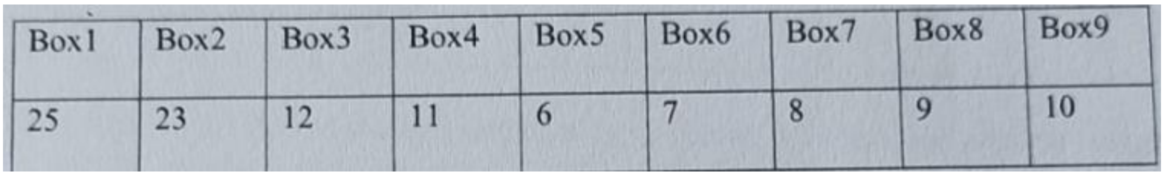


	CustomerID	Gender	Age	Annual Income (k$)	Spending Score (1-100)
0	1	Male	19	15	39
1	2	Male	21	15	81
2	3	Female	20	16	6
3	4	Female	23	16	77
4	5	Female	31	17	40

20. Streaming services often use clustering analysis to identify viewers who have similar behavior. The streaming service may collect the following data about individuals:
    - Minutes watched per day
    - Total viewing sessions per week
    - Number of unique shows viewed per month

    Using these metrics, a streaming service can perform cluster analysis to identify high-usage and low-usage users so that they can know whom they should spend most of their advertising dollars on. Apply the Hierarchical Cluster algorithm and EM clustering algorithm to identify and compare the performance of the clustering technique.

21. The following values are the number of pencils available in the different boxes. Create a vector in R and find out the mean, median and mode values of set of pencils in the given data.



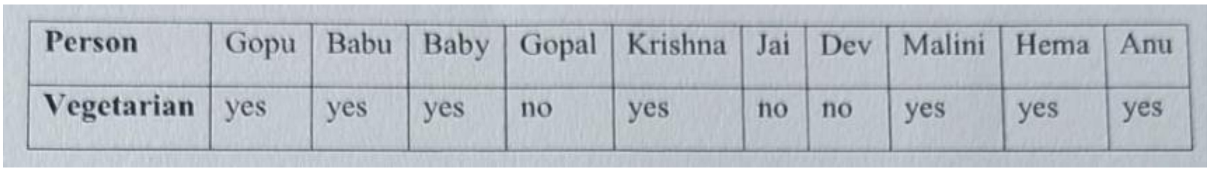
Box1	Box2	Box3	Box4	Box5	Box6	Box7	Box8	Box9
25	23	12	11	6	7	8	9	10

22. Assume the Tennis coach wants to determine if any of his team players are scoring outliers.
    To visualize the distribution of points scored by his players, then how can he decide to develop the box plot? Give suitable example using Boxplot visualization technique using R.

23. Create the following dataset using CSV file format. To perform cluster analysis using K-Means in WEKA. To change the cluster size and plot the graph and illustrate the visualization of cluster.

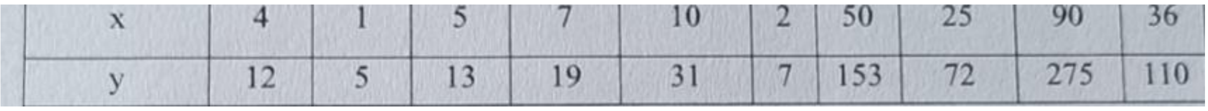
EmployeID	Gender	Age	Salary	Credit
111	Male	28	150000	39
222	Male	25	150000	27
333	Female	26	160000	42
444	Female	25	160000	40
555	Female	30	170000	64
666	Male	29	200000	72

24. Predict the categorical data using Naïve Bayes classification through WEKA using any dataset of your choice. Compare the Naive Bayes algorithm with SVM using the summary of results given by the classifiers and plot the graph.
25. The following list of persons with vegetarian or not details given in the table. How will you tind out how many of them are vegetarian and how many of them are non-vegetarian? Which type of the person total count is greater value?



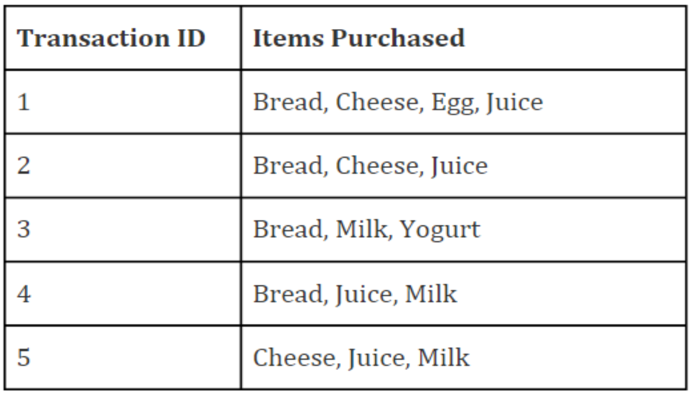
Person	Gopu	Babu	Baby	Gopal	Krishna	Jai	Dev	Malini	Hema	Anu
Vegetarian	yes	yes	yes	no	yes	no	no	yes	yes	yes

26. The following table would be plotted as (x,y) points, with the first column being the x values as number of mobile phones sold and the second column being the y values as money. To use the scatter plot for how many mobile phones sold.



x	4	1	5	7	10	2	50	25	90	36
y	12	5	13	19	31	7	153	72	275	110

27. Generate rules using FP growth algorithm using the given dataset which has the following transactions with items purchased: Consider the values as support=50%
And confidence=75%.



Transaction ID	**Items Purchased**
1	Bread, Cheese, Egg, Juice
2	Bread, Cheese, Juice
3	Bread, Milk, Yogurt
4	Bread, Juice, Milk
5	Cheese, Juice, Milk

28. Predict Diabetes Dataset using Decision tree classifier in WEKA. Compare it with Support Vector Machine classifier. Show the result accuracy and F1 measure calculation. Plot the graph and explain the summary of results.
29. Let us consider marks scored by a student in his model exam that has been sorted as follows: 55, 60, 71, 63, 55, 65, 50, 55,58,59,61,63,65,67,71, 72,75. Partition them into three bins by each of the following methods. Plot the data points using histogram in R.

(a) equal-frequency (equi-depth) partitioning

(b) equal-width partitioning

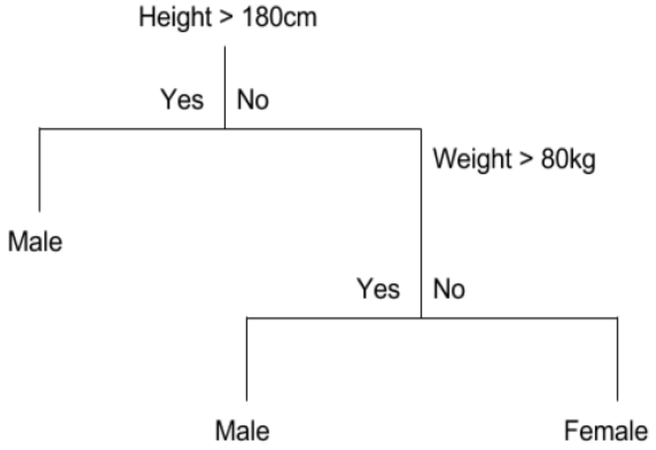
(c) clustering

30. Consider this Decision tree :

a) Create the data set for the below tree using ARFF format and calculate accuracy and decision for the same

b) Using this decision tree generate the rules based on rule-based induction.

c) Compare both the algorithms and plot the confusion matrix.



31. Create an ARFF file for the table below and implement for the Apriori Algorithm and FP growth algonthm and compare the rules generated by both the algorithms. Identify the unique rules generated by the above algorithms.
NOTE: Assume Minimum support-2 and confidence= 50%

T.ID	ITEMS
T1	SONY, BPL, LG
T2	BPL, SAMSUNG
T3	BPL, ONIDA
T4	SONY, BPL, SAMSUNG
T5	SONY, ONIDA
T6	BPL, ONIDA
T7	SONY, ONIDA
T8	SONY, BPL, ONIDA, LG
T9	SONY, BPL, ONIDA

32. The given are the strike-rates scored by a batsman in season 1 in different tournaments. 100, 70, 60, 90, 90. Perform the following in R.

a) Min-Max normalization by setting min = 0 and max = 1
b) z-score normalization
c) z-score normalization using the mean absolute deviation instead of standard deviation
d) normalization by decimal scaling

33. suppose some car is tested for the Avg Speed and Total Time data for 9 randomly selected car with the following result.

Avg Speed (in kph)	78	81	82	74	83	82	77	80	70
Total Time (in Mins)	39	37	36	42	35	36	40	38	46

a) Calculate the standard deviation of Avg. Speed and Total Time.

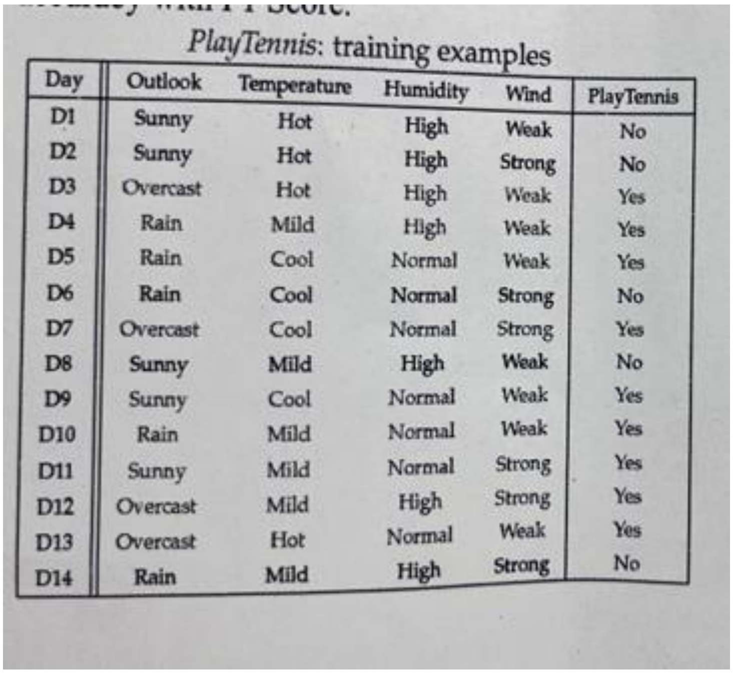
b) Calculate the Variance of Avg. Speed and Total Time for the above dataset.

34. Consider a person want to take a census/plot for the breast-cancer affected people through the years. Create a own dataset with this parameters age, tumor size, in v-nodes [example between age 1-5 = no of count, 6-10-no of count, etc]

Draw the Histogram, scatter plot, box plot.

35. A shepherd boy gets bored tending the town's flock. To have some fun, he cries out, "Wolf!" even though no wolf is in sight. The villagers run to protect the flock, but then get really mad when they realize the boy was playing a joke on them. One night, the shepherd boy sees a real wolf approaching the flock and calls out, "Wolf!" The villagers refuse to be fooled again and stay in their houses. The hungry wolf turns the flock inti lamb chops. The town goves hungry. Panic ensures.

36. Create the ARFF data set for the below mentioned dataset perform the Bayes theorem in addition to that compare the same with decision tree. Identify the efficient classifier with accuracy with F1 Score.

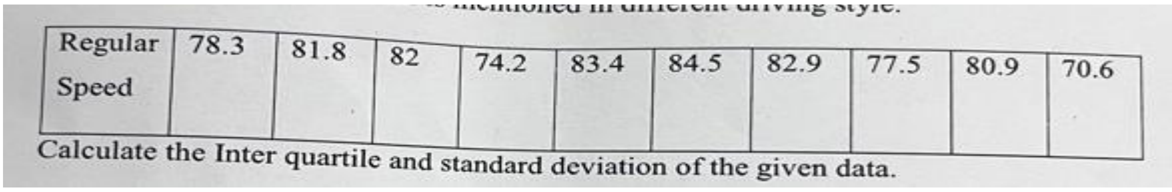


PlayTennis: training examples

Day	Outlook	Temperature	Humidity	Wind	PlayTennis
D1	Sunny	Hot	High	Weak	No
D2	Sunny	Hot	High	Strong	No
D3	Overcast	Hot	High	Weak	Yes
D4	Rain	Mild	High	Weak	Yes
D5	Rain	Cool	Normal	Weak	Yes
D6	Rain	Cool	Normal	Strong	No
D7	Overcast	Cool	Normal	Strong	Yes
D8	Sunny	Mild	High	Weak	No
D9	Sunny	Cool	Normal	Weak	Yes
D10	Rain	Mild	Normal	Weak	Yes
D11	Sunny	Mild	Normal	Strong	Yes
D12	Overcast	Mild	High	Strong	Yes
D13	Overcast	Hot	Normal	Weak	Yes
D14	Rain	Mild	High	Strong	No

37. a).Suppose that the "Diabetes" dataset data for analysis includes the attribute age. The age values for the data are (in increasing order) 30, 57, 68, 96, 39, 40, 20, 19, 42, 12, 25, 25, 65, 35, 30, 23, 23, 35, 45, 85. Use R script to find the mean

    b) Suppose that the speed car is mentioned in different driving style.



mentioned in different driving style.

Regular Speed	78.3	81.8	82	74.2	83.4	84.5	82.9	77.5	80.9	70.6

Calculate the Inter quartile and standard deviation of the given data.

38. (a) Let us consider one example to make the calculation method clear. Assume that the minimum and maximum values for the feature F are \$50,000 and \$100,000 correspondingly. It needs to range F from 0 to 1. In accordance with min-max normalization, v = \$80,

    (b) Use the two methods below to normalize the following group of data: 200, 300, 400,600, 1000

    (a) min-max normalization by setting min = 0 and max = 1

    (b) z-score normalization

39. Consider this table

TID	items bought
T100	{M, O, N, K, E, Y)
T200	{D, O, N, K, E, Y)
T300	{M, A, K, E}
T400	{M, U, C, K, Y)
T500	{C, 0, 0, K, I,E}

(a) Find all frequent item set using Apriori and FP-growth, respectively. Compare the efficiency of the two mining processes.

(b) List all of the strong association rules (with support s and confidence c) matching the following meta rule, where & is a variable representing customers, and itemi denotes variables representing items (e.g., "A", "B", etc.):

Vx E transaction, buys, item1) A buys(X, item2) → buys(X, item3)

40. 4.Suppose we want to classify potential bank customers as good creditors or bad creditors for loan applications. We have a training dataset describing past customers using the following attributes: Marital status {married, single, divorced}, Gender {male, female}, Age {[18.30[[30..50L, [50.65[[65+], Income {[10K..25KL, [25K..5OKL, [5OK..65KL, [65K..100KL, [100K+]}. Using WEKA tool solve this problem.