1. Write a high level program for the following:

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, 36, 40, 45, 52, 70.

1. What is the mean of the data? What is median?
2. What is the mode of the data? Comment on the data’s modality (bimodal, trimodal, etc..)
3. What is mid-range of the data?
4. Can you find the first quartile(Q1) and the third quartile (Q3) of the data?
5. Give the five number summary of the data.
6. Write a C program to calculate the *correlation coefficient. Use the following data to check your code.*

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



Are these two variables positively or negatively correlated?

1. 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. Write a C program to implement *smoothing by bin means* to smooth the data, using a bin depth of 3.
2. Write a C program to implement:

(a) min-max normalization

(b) z-score normalization

(c) Normalization by decimal scaling.

1. Write a program to find frequent item sets using apriori algorithm for the following transactional database. Let min sup = 60% and min conf = 80%



1. Write high level language programs to implement Naïve Bayesian classification for the given case study.

The following table consists of training data from an employee database. The data have been generalized. For example, “31 : : : 35” for age represents the age range of 31 to 35. For a given row entry, count represents the number of data tuples having the values for department, status, age, and salary given in that row.



Given a data tuple having the values “systems,” “26 . . . 30,” and “46–50K” for the

attributes department, age, and salary, respectively, what would a naive Bayesian

classification of the status for the tuple be?

1. Write high level language programs to implement KNN classification for the following case study.

Suppose we have height, weight and T-shirt size of some customers and we need to predict the T-shirt size of a new customer given only height and weight information we have. Data including height, weight and T-shirt size information is shown below

|  |  |  |
| --- | --- | --- |
| Height (in cms) | Weight (in kgs) | T Shirt Size |
| 158 | 58 | M |
| 158 | 59 | M |
| 158 | 63 | M |
| 160 | 59 | M |
| 160 | 60 | M |
| 163 | 60 | M |
| 163 | 61 | M |
| 160 | 64 | L |
| 163 | 64 | L |
| 165 | 61 | L |
| 165 | 62 | L |
| 165 | 65 | L |
| 168 | 62 | L |
| 168 | 63 | L |
| 168 | 66 | L |
| 170 | 63 | L |
| 170 | 64 | L |
| 170 | 68 | L |

New customer named 'Monica' has height 161cm and weight 61kg. find out the T-Shirt size using KNN Classification.

1. Write high level language programs to implement k-means clustering for the following case study.

Suppose that the data mining task is to cluster points (with (*x*, *y*) representing location) into three clusters, where the points are

*A*1(2,10), *A*2(2,5), *A*3(8,4), *B*1(5,8), *B*2(7,5), *B*3(6,4), *C*1(1,2), *C*2(4,9).

The distance function is Euclidean distance. Suppose initially we assign *A*1, *B*1, and *C*1 as the center of each cluster, respectively. Use the *k-means* algorithm to show *only*

(a) The three cluster centers after the first round of execution.

(b) The final three clusters.

1. Write high level language programs to implement k-medoids clustering for the following data points. assuming no of clusters=2.

(2,10), (2,5), (8,4), (5,8), (7,5), (6,4),(1,2), (4,9).