LPW

Subject : 2ce423-DATA MINING AND WAREHOUSING

1. Implement following normalization method in C/C++ on the given data (age):
   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.
   a) Use Min-Max Normalization to transform the value 25 and 52 for age onto the range[0.0, 1.0].

   b) Use z-score normalization to transform the value 35 for *age*, where the standard deviation of *age* is 12.94 years.

   c) Use normalization by decimal scaling to transform the value 35 for *age*.

2. Implement any three methods in C/C++ to fill the missing values indicated by ? in the given data set

Name	Value
A	45
B	37
C	59
D	?
E	47
F	39
G	?
H	43
I	52
J	?

3. Implement suitable method (using concept of Quartile) in C/C++ for detection of outliers present in the following data set : also take steps of remove these identified outliers from the given data set.

Name	Value
A	45
B	37
C	59
D	150
E	47
F	39
G	5
H	43
I	52
J	100

4. Consider the following Customer database of a Car sales shop :

ID	**Age**	**Income**	**Student**	**Credit Rating**	**Buy Car**
1	Young	High	No	Fair	No
2	Young	High	No	Good	No

3	Middle	High	No	Fair	Yes
4	Old	Medium	No	Fair	Yes
5	Old	Low	Yes	Fair	Yes
6	Old	Low	Yes	Good	No
7	Middle	Low	Yes	Good	Yes
8	Young	Medium	No	Fair	No
9	Young	Low	Yes	Fair	Yes
10	Old	Medium	Yes	Fair	Yes
11	Young	Medium	Yes	Good	Yes
12	Middle	Medium	No	Good	Yes
13	Middle	High	Yes	Fair	Yes
14	Old	Medium	No	Good	No

Assume the attribute 'Buy Car' as decision variable. Suppose we want to construct decision tree, find using entropy or information gain which attribute can be used as 'root' of the decision tree. Implement the same in C/C++ .

5. Implement the Natural partitioning (apply 3-4-5 rule) algorithm in C/C++ for generating concept hierarchy (upto two levels) for the following data (given the attribute 'marks' of some students) : 32, 38, 48, 91, 46, 37, 22, 69, 78, 82, 33, 49, 55, 66, 84, 86, 67, 80, 79, 44.
6. Implement K-Means Algorithm cluster the following eight points (with (*x; y*) representing location) into three clusters.

$A_1(2; 10); A_2(2; 5); A_3(8; 4); B_1(5; 8); B_2(7; 5); B_3(6; 4); C_1(1; 2); C_2(4; 9)$:

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 the three cluster centers and the all the points of clusters after the 2nd round of execution.

7. Getting acquainted with WEKA, R-Package and Anaconda (with Jupyter – IDE) for Python.
8. Performing Pre-processing tasks before Classification, Clustering.
9. Classification using J48/C4.5 Algorithm for the given data (Iris.arff).
10. Classification using Naïve Bayes Algorithm for the given data (Iris.arff) and Comparing result of these two algoritm.
11.