BITS Pilani
Pilani | Dubai | Goa | Hyderabad

AIML FEATURE ENGINEERING ASSIGNMENT 2

Submission Date: 29 December 2019 11.59 PM

Weightage: 8%

Q1. Consider that following employee data set made available to you for carrying out some data mining activity. What are the four potential issues with this dataset? [2 %]

Name	Age	DateOfJoining	Designation	DateOfBirth
A	34	15-Jan-2015	Sr Engineer	Feb 24, 1981
B	33	27-Jan-2015		Mar 27, 1982
A	34	15-Jan-2015	Sr Engineer	Feb 24, 1981
C	32	30-Jan-2015	Staff Engineer	Nov 25,1982

Q2. Consider the following list of observations of a variable. Answer the following: 25, 35, 33, 25, 42, 35, 30, 35, 36, 41, 40, 33, 42, 35 and 99
[4 %]

(a) What is the five-number summary for the given data?
(b) Draw boxplot.
(c) Identify the outliers if any.
(d) Explain, how do the outliers affect the measures of central tendency (Mean, and Median) of data? Comment using the given data set.

Q3. Given the following two objects with four binary attributes. [1%]

	Attribute 1	Attribute 2	Attribute 3	Attribute 4
Object 1	1	1	0	0
Object 2	1	0	1	0

(a) What is the distance between the objects if all variables are symmetric?
(b) What is the distance between the objects if all variables are asymmetric?

Q4. As many data mining algorithms cannot handle missing values, analyst sometimes remove all observations (rows) that contain missing values before the analysis. Give two potential disadvantages of this procedure using an example. [1%]

Submission Details
1. Final document - FE_Assignment2_<Student_ID>.doc