LAB 2: Data mining – Data Preprocessing

Part A.

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)

a) What is the mean of the data?

b) What is the median?

c) What is the mode of the data

d) Can you find (roughly) the first quartile (Q1) and the third quartile (Q3) of the data?

f) Show a boxplot of the data.

2. Based on the numbers in the following table, please describe if these numbers are measured according to (1) nominal scale, (2) ordinal scale, (3) interval scale, and (4) ratio scale, respectively, then among W_1, W_2, W_3, W_4, W_5, W_6, which will denote the same measure as V(.)?

	A	B	C	D	E
$V(.)$	**1**	**2**	**3**	**4**	**5**
$W_1(.)$	100	200	300	400	500
$W_2(.)$	10	11	12	13	14
$W_3(.)$	8	13	45	6	7
$W_4(.)$	1	4	9	16	25
$W_5(.)$	-10	-8	-6	-4	-2
$W_6(.)$	3	6	9	12	15

3. Confusion matrix

Assume that dataset includes:

39 samples, two classes (positive and negative)

the performance of a classification algorithm has been shown as confusion matrix

Classified as→	Positive	Negative
Positive	22	0
Negative	17	0

a) Determine TP, TN, FP, FN
b) Calculate accuracy, recall, precision, sensitivity, specificity.
c) Calcualte f1-score

Part B. Choose a real-world dataset. (Kaggle or other resources)

1. Perform data cleaning tasks such as handling missing values, outliers, inconsistent data, and so on. You can use programming tools such as Python to complete this assignment.
2. Apply data transformation techniques such as normalization, standardization, encoding, and so on.

[1] https://www.kaggle.com/code/abhishekmamidi/titanic-data-preprocessing-and-visualization

[2] https://www.kaggle.com/code/nkitgupta/advance-data-preprocessing

[3] https://www.kaggle.com/code/burhanykiyakoglu/predicting-house-prices

[4] https://www.kaggle.com/code/faressayah/logistic-regression-data-preprocessing

[5] https://www.kaggle.com/code/vikassingh1996/extensive-data-preprocessing-and-modeling

[6] https://www.kaggle.com/code/utkarshm25/data-preprocessing-basics

[7] https://www.kaggle.com/code/anirban7/data-preprocessing-for-beginners