Data Mining - EX7

Deadline: Friday, Azar 16, 1403 - December 06, 2024

Question 1:

A company wants to segment its customers based on purchasing patterns and consumption habits. The data team considers two approaches: one for predicting future purchases based on previous data and another for discovering groups of customers with similar behaviors.

1. Explain the difference between supervised and unsupervised methods. Which of these approaches would be more suitable for segmenting customers without specific labels?
2. Which type of data mining tasks are associated with supervised methods, and which are associated with unsupervised methods?
3. In which of the two methods should the data be labeled?

Question 2:

A data mining team wants to create a model to predict the probability of customer purchases. They have a dataset that can be divided into training, test, and validation sets.

1. Describe the differences between the training set, test set, and validation set.
2. Should we strive for the highest possible accuracy with the training set? Why or why not? Answer the same for the validation set.
3. If we want to prevent overfitting, how should we adjust our model?

Question 3:

A company is building a model to detect fraudulent transactions. They realize that only 4% of their training data consists of fraudulent transactions, so they decide to balance the data.

1. If we want to increase the proportion of fraudulent transactions to 20%, how many new fraudulent samples should we add?
2. Why is it necessary to report a baseline performance before presenting the model's results?

3. Explain the difference between reporting an absolute difference and a relative difference, and indicate which one is more suitable for reporting model results.