DSC 441 Homework 1

Name: Sanchal Sunil Dhurve

PROBLEM 4

a. Describe two ways in which data can be dirty, and for each one, provide a potential solution.

Ans: Ways in Which Data Can Be Dirty and Solutions

a. Missing Data:

- Description: Some values are absent in the dataset, e.g., a customer's age is missing.
- Solution: Use imputation techniques such as filling with the mean/median for numerical data or the mode for categorical data. Alternatively, remove rows or columns with excessive missing values.

b. Inconsistent Data:

- Description: Data values differ in format or scale, e.g., dates in different formats (MM/DD/YYYY vs. YYYY-MM-DD) or inconsistent use of units.
- Solution: Standardize formats and units across the dataset using transformation functions or tools like regular expressions.

b. Explain which data mining functionality you would use to help with each of these data questions.

a. Suppose we have data where each row is a customer and we have columns that describe their purchases. What are five groups of customers who buy similar things?

Ans: Functionality: Clustering. Group customers into segments based on their purchase patterns.

b. For the same data: can I predict if a customer will buy milk based on what else they bought?

Ans: Functionality: Classification. Build a model to classify whether a customer will buy milk based on their purchase history.

c. Suppose we have data listing items in individual purchases. What are different sets of products that are often purchased together?

Ans: Functionality: Association Rule Mining. Discover frequent item sets and association rules (e.g., "Customers who buy bread are likely to buy butter").

c. Explain if each of the following is a data mining task

a. Organizing the customers of a company according to education level.
Ans: No. This is simply data organization, not mining. It doesn’t involve uncovering patterns or insights.

b. Computing the total sales of a company.
Ans: No. This is a simple aggregation task, not data mining.

c. Sorting a student database according to identification numbers.
Ans: No. Sorting is an organizational task, not a data mining process.

d. Predicting the outcomes of tossing a (fair) pair of dice.
Ans: No. Outcomes of fair dice are purely random and lack a discernible pattern to mine.

e. Predicting the future stock price of a company using historical records.
Ans: Yes. This is a predictive modeling task, a key functionality of data mining.