BIN2023R01 – INTRODUCTION TO DATAMINING & MACHINE LEARNING FOR BIOINFORMATICS

Lab Exercise 3- Data quality assessment: Missing value imputation, outliers, and noise

Aim: To evaluate the quality of the datasets, address any missing values, and eliminate outliers from the provided datasets

Using the given dataset,

1. Identify the statistics summary of the dataset.

2. Check and visualize the missing values.

3. Perform missing value imputation.

4. Detect the outliers present in the dataset.

5. Visualize the outliers using a box plot and scatter plot.

6. i) Employ univariate outlier detection techniques, including maximum likelihood, Z-score, and Interquartile Range (IQR) methods and multivariate outlier detection methods such as Mahalanobis distance and $\chi 2$-statistics for outlier identification.

6. ii) Eliminate the detected outliers using the aforementioned methods and evaluate the effectiveness of each approach.

7. Visualize the dataset post-outlier removal using the above methods and interpret their effectiveness in addressing outliers for the given dataset.

Questions:

1. What methods can be used to assess the overall quality of a dataset? How do you identify potential issues or anomalies in the data?
2. Why is it important to understand the quality of the data before analysis?
3. What are common techniques for handling missing values in a dataset?
4. How does imputing missing values impact the statistical properties of the data? Can you name a machine learning algorithm suitable for imputing missing values? Compare and contrast different imputation strategies for handling missing values.
5. Explain the concept of multiple imputation and its advantages.
6. How do outliers affect statistical analyses and machine learning models? What methods can be employed to detect outliers in a dataset? Provide examples of techniques to treat or mitigate the impact of outliers.
7. What distinguishes multivariate outlier detection from univariate methods? In what scenarios is multivariate outlier detection more advantageous than univariate methods? Can univariate outlier detection methods handle situations where outliers are present in multiple variables simultaneously?
8. How might the handling of missing values and outliers introduce bias in data-driven decision-making?

Soft copy deadline: 12th February 11:59PM
Hard copy deadline: 13th February 3:15PM