ASSIGNMENT-1

Data Mining and Predictive Analytics (CSE 4859)

Programme: B. Tech. (CSE) **Semester: 7th**

Full Marks: 10 **Date of Submission: 24/10/2025**

Subject/Course Learning Outcome	***Taxonomy Level**	**Ques. Nos.**	**Marks**
Comprehend fundamental concepts of Data Mining, Predictive Analytics, CRISP-DM process and explain their applications.	L1, L2	Q1, Q2, Q3	
Apply appropriate data preprocessing, EDA, and dimension-reduction methods to prepare datasets for effective analysis.	L1, L2, L3, L4	Q4, Q5, Q6, Q7, Q8, Q9, Q10	
Apply univariate and multivariate statistical analysis on the data in order to assess underlying patterns and relationships.			
Describe and apply key data preparation techniques including Cross-validation, Bias Variance trade-off, Overfitting Control, etc. to enhance model training and validation.			
Analyze predictive modeling techniques such as Simple Linear Regression and Multiple Regression to model relationships between variables.			
Explain and Demonstrate the use of k-Nearest Neighbor (k-NN) algorithm for classification and prediction tasks with their applicability in different problem domains.			

*Bloom's taxonomy levels: Knowledge (L1), Comprehension (L2), Application (L3), Analysis (L4), Evaluation (L5), Creation (L6).

- **Write your answers with enough detail about your approach and concepts used, so that the grader will be able to understand it easily.**
- **You are allowed to use only those concepts which are covered in the lecture class till date.**
- **Assignment scores/markings also depend on neatness, clarity and date of submission.**

1. What is Data Mining? Describe the steps involved in data mining when viewed as a process of knowledge discovery.
2. Define Predictive Analytics and explain its relationship with Data Mining.
3. For each of the following meetings, explain which phase in the CRISP-DM process is represented:
   a. Managers want to know by next week whether deployment will take place. Therefore, analysts meet to discuss how useful and accurate their model is.
   b. The data mining project manager meets with the data warehousing manager to discuss how the data will be collected.
   c. The data mining consultant meets with the Vice President for Marketing, who says that he would like to move forward with customer relationship management.
   d. The data mining project manager meets with the production line supervisor, to discuss implementation of changes and improvements.

e. The analysts meet to discuss whether the neural network or decision tree models should be applied.

4. What is an outlier? Why do we need to treat outliers carefully?

5. Use the following stock price data (in dollars) to compute mean, median, mode and standard deviation.

Stock Price	10	7	20	12	75	15	9	18	4	12	8	14

6. Use the following stock price data (in dollars) to answer the following questions.

Stock Price	10	7	20	12	75	15	9	18	4	12	8	14

a. Find the min-max normalized stock price for the stock worth $20

b. Compute the midrange stock price.

c. Compute the z-score standardized stock price for the stock worth $20

d. Compute the decimal scaling stock price for the stock worth $20

e. Compute the skewness for the stock price data

7. Use the given data set for the following questions: 1 1 1 3 3 7

a. Bin the data into three bins of equal width (width = 3).

b. Bin the data into three bins of two records each.

8. Answer following questions:

a. What is the graphical counterpart of a contingency table?

b. What is the difference between taking row percentages and taking column percentages in a contingency table?

9. For each of the following descriptive methods, state whether it may be applied to categorical data, continuous numerical data, or both.

a. Bar charts

b. Histograms

c. Summary statistics

d. Cross-tabulations

e. Correlation analysis

f. Scatter plots

g. Web graphs

h. Binning

10. Find Q1, Q2, and Q3 for the following data set, and draw a box-and-whisker plot.
{2,6,7,8,8,11,12,13,14,15,22,23}