

ASSIGNMENT-3

Data Mining and Predictive Analytics (CSE 4859)

Programme: B. Tech. (CSE)

Semester: 7th

Full Marks:

Date of Submission: 24/12/2025

| Subject/Course Learning Outcome | *Taxonomy Level | Ques. Nos. | Marks |
|---|-----------------|----------------|-------|
| Comprehend fundamental concepts of Data Mining, Predictive Analytics, CRISP-DM process and explain their applications. | | | |
| Apply appropriate data preprocessing, EDA, and dimension-reduction methods to prepare datasets for effective analysis. | | | |
| 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. | L1, L2, L3, L4 | Q1, Q2, Q3, Q4 | |
| Explain and demonstrate the use of k-Nearest Neighbor (k-NN) algorithm for classification and prediction tasks with their applicability in different problem domains. | L1, L3 | Q5, Q6 | |

*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. Differentiate between simple linear regression and multiple linear regression.
2. Sam found how many hours of sunshine vs how many ice creams were sold at the shop from Monday to Friday. Find equation of regression line for the given data using the “Least Squares Estimates” method.

| Hours of Sunshine | Ice Creams Sold |
|-------------------|-----------------|
| 2 | 4 |
| 3 | 5 |
| 5 | 7 |
| 7 | 10 |
| 9 | 15 |

3. Consider following dataset that shows the number of hours studied by six different students along with their final exam scores. Here “Exam Score” is dependent variable whereas “Hours Studied” is independent variable.

| Hours Studied | Exam Score |
|---------------|------------|
| 1 | 68 |
| 2 | 77 |
| 2 | 81 |
| 3 | 82 |
| 4 | 88 |
| 5 | 90 |

- a. Find equation of regression line that can best fit.
 - b. Calculate Sum of Squares Error (SSE).
 - c. Calculate Sum of Squares Regression (SSR).
 - d. Calculate Sum of Squares Total (SST).
 - e. Calculate Coefficient of Determination (r^2).
 - f. Calculate Mean Square Error (MSE) and standard error s .
4. Given the following data for two variables, X and Y, calculate the Pearson correlation coefficient.

| X | Y |
|---|---|
| 1 | 2 |
| 2 | 3 |
| 3 | 5 |
| 4 | 7 |
| 5 | 8 |

5. Consider following table:

| Record | Age | Income (in \$1000) |
|--------|-----|--------------------|
| 1 | 22 | 4.6 |
| 2 | 33 | 2.4 |
| 3 | 28 | 2.8 |
| 4 | 51 | 2.3 |
| 5 | 25 | 4.7 |
| 6 | 39 | 3.3 |
| 7 | 54 | 2.8 |
| 8 | 55 | 4.9 |
| 9 | 50 | 4.6 |
| 10 | 66 | 3.6 |

- a. Standardize the attributes with min-max normalization.
 - b. Compute the Euclidean distance of record no. 10 with other records.
 - c. Find the nearest neighbours for $k=3$.
6. Discuss the advantages and drawbacks of using a small value versus a large value for k in k -Nearest Neighbor classifier.