ASSIGNMENT-2

Data Mining and Predictive Analytics (CSE 4859)

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

Full Marks: **Date of Submission: 13/12/2025**

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.	L1, L2, L3, L4	Q1, Q2, Q3, Q4, Q5, Q6, Q7, Q8, Q9	
Describe and apply key data preparation techniques including Cross-validation, Bias Variance trade-off, Overfitting Control, etc. to enhance model training and validation.	L1, L2, L3, L4	Q10, Q11	
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. Answer following questions:
   a. What do you mean by hypothesis testing for some parameter in a population?
   b. How can we use a confidence interval to conduct hypothesis testing of the population parameters?
2. The duration of customer service calls to an insurance company is normally distributed, with mean 20 minutes, and standard deviation 5 minutes. For the following sample sizes, construct a 95% confidence interval for the population mean duration of customer service calls. (**Refer to the attached table for the critical t-values.**)
   a. n = 25,
   b. n = 100,
   c. n = 200.

3. Of 1000 customers who received promotional materials for a marketing campaign, 100 responded to the promotion. For each of the following confidence levels, construct a confidence interval for the population proportion who would respond to the promotion. (Given the standard normal variate Z satisfies the followings: $P(Z < 1.645) = 0.95$, $P(Z < 1.96) = 0.975$, $P(Z < 2.576) = 0.995$**).**
   a. 90%,
   b. 95%,
   c. 99%.
4. Discuss about Type I and Type II errors in the context of hypothesis testing with an example.
5. In the churn problem discussed in Chapter 3, recall that 483 of 3333 customers in the sample had churned the company. Using level of significance $\alpha = 0.10$, test whether the population proportion π differs from 0.15 by computing the p-value from the Z_{data}. (Given the standard normal variate Z satisfies the followings: $P(Z < 1.645) = 0.95$, $P(Z < 1.96) = 0.975$, $P(Z > 0.8246) = 0.2048$**).**
6. A sample of 100 donors to a charity has a mean donation amount of $55 with a sample standard deviation of $25. Test using α = 0.05 whether the population mean donation amount exceeds $50. (**Refer to the attached table for the critical t-values.**)
   a. Define the null hypothesis and the alternative hypothesis on μ.
   b. What is the rejection rule?
   c. What is the meaning of the test statistic *T*?
   d. What are the values of the test statistic T_{data} and the p-value in this example?
   e. What is our conclusion after comparing the p-value with the level of significance?
   f. Interpret our conclusion so that a non-specialist could understand it.
7. The following table contains information on the mean duration of customer service calls between a training and a test data set. Test whether the partition is valid for this variable, using α = 0.10. (Given the T variate satisfies the followings: $P(T > 0.4322) = 0.3328$**.**)

Data set	Sample Mean	Sample Standard Deviation	Sample Size
Training set	x_1 = 20.5	s_1 = 5.2	n_1 = 2000
Test set	x_2 = 20.4	s_2 = 4.9	n_2 = 600

8. The multinomial variable payment preference takes the values credit card, debit card, and cheque. Now, suppose we know that 50% of the customers in our population prefer to pay by credit card, 20% prefer debit card, and 30% prefer to pay by cheque. We have taken a sample from our population, and would like to determine whether it is representative of the population. The sample of size 200 shows 125 customers preferring to pay by credit card, 25 by debit card, and 50 by cheque. Test whether the sample is representative of the population, using α = 0.05. Given that $P(\chi^2 > 38.82) = 10^{-6}$.
9. Suppose a multinomial variable "**movie choices**" has the following values {**'romantic'**, **'science fiction'**}. We have a set of 1000 males and a set of 250 females with the following frequencies:

	Romantic	Science Fiction	Total
Male	350	650	1000
Female	175	75	250
Total	525	725	1250

Test whether significant differences exist between the multinomial proportions of the two gender groups, given that $P(\chi^2 < 102.17) = 0.9999$.

10. Describe the differences between the training set, test set, and validation set.
11. How is the bias–variance trade-off related to the issue of overfitting and underfitting? Is high bias associated with overfitting and underfitting, and why?

TABLE: Critical values for t-distribution with various degrees of freedom

df	α=.25	.20	.15	.10	.05	.025	.02	.01	.005	.0025	.001	.0005
24	0.685	0.857	1.059	1.318	1.711	2.064	2.172	2.492	2.797	3.091	3.467	3.745
25	0.684	0.856	1.058	1.316	1.708	2.060	2.167	2.485	2.787	3.078	3.450	3.725
99	0.677	0.845	1.042	1.290	1.660	1.984	2.081	2.364	2.626	2.871	3.174	3.391
100	0.677	0.845	1.042	1.290	1.660	1.984	2.081	2.364	2.626	2.871	3.174	3.390
199	0.676	0.843	1.039	1.286	1.653	1.972	2.067	2.345	2.601	2.839	3.132	3.340
200	0.676	0.843	1.039	1.286	1.653	1.972	2.067	2.345	2.600	2.838	3.131	3.339
∞	0.674	0.841	1.036	1.282	1.640	1.960	2.054	2.326	2.576	2.807	3.091	3.291