INSTITUTE OF TECHNICAL EDUCATION AND RESEARCH

(SOA Deemed to be University)

Theory Assignment
Data Mining and Predictive Analytics
(CSE 4859)

Submitted By

Name: _____

Registration No.:_____

Branch: _____

Semester:_____ Section:_____

INDEX

Name: Section:

Registration No.:

S. No.	Title	Page No.	Date of Submission	Teacher's Sign/Remark

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}

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

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.

INSTITUTE OF TECHNICAL EDUCATION AND RESEARCH

(SOA Deemed to be University)

Assignment
Deep Learning using Python (CSE 4685)

Submitted By

Name: _____

Registration No.:_____

Branch: _____

Semester:_____ Section:_____

ASSIGNMENT
Deep Learning using Python (CSE 4685)

Programme: B. Tech (CSE) **Semester: 7th**
Full Marks: 20 **Date of Submission: 20-12-2025**

Subject/Course Learning Outcome	*Taxonomy Level	Question Nos.	Marks
Able to apply the key fields of linear algebra and probability to build complex neural models.	L2	1	2
Able to analyse, design and implement Artificial Neural Networks (ANNs) and optimization techniques for solving complex learning problems.	L3	2,9	4
Able to analyse, build and implement Convolutional Neural Networks (CNNs) using PyTorch for solving classification problems.	L3	3,4,5,10	8
Able to analyse, design and implement autoencoder using PyTorch for nonlinear data handling.	L3	7	2
Able to design and interpret Recurrent Neural Networks (RNNs) and long short-term memory (LSTM) for sequential data analysis.	L3	6	2
Able to develop Generative Adversarial Networks (GANs) to synthesize new data.	L3	8	2

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

Answer all questions. Each question carries equal mark.

- **Assignment scores/markings depend on neatness, clarity and date of submission.**
- **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.**

1. What is the prerequisite for deep learning? Explain in details, how these concepts are helpful to build Deep learning models.

2. Given perceptron weights $w_1=1$, $w_2=-1$, $b=-0.2$. Write the equation of the decision boundary and classify the points (0.5,0.5) and (0.8,0.2)

3. Explain the role of the following layers in CNN:
    - Convolution layer
    - Pooling layer
    - Fully connected layer

4. Table1: Network parameter details

Layer	Parameters
Conv 2D	Filter size=5×5, stride=1, padding=0, #filters=6
Max Pooling 2D	Stride=2, filter size=2×2
Conv 2D	Filter size=5×5, stride=1, padding=0, #filters=16
Max Pooling 2D	Stride=2, filter size=2×2
Conv 2D	Filter size=5×5, stride=1, padding=0, #filters=32
Max Pooling 2D	Stride=2, filter size=2×2
Flatten	
Dense 1	120
output	10

Draw the CNN architecture with the following information given in **Table.1** for image classification. The input image size is 128×128×3. Find the total no. of learnable parameters in the network.

5. State batch normalization. Discuss the limitations of batch normalization and how layer and group normalization overcomes it with example.

6. Draw a RNN network with input(2) →hidden(2) →output(1), calculate the Y predicted for the following given parameter values in **Table.2** with input x_1=[1, 0], x_2=[0, 1] ,h_0=[0,0].

Table2: Network parameter details

Layer	**Weights**	**Bias**
Input to hidden	[0.5 0.8 , −0.3 0.2]	-
Hidden-to hidden	[0.1 −0.4, 0.2 0.3]	0.0,0.1
Hidden to output	[1.0, −1.0]	0.0

7. What is autoencoder? How does it work? What are the loss functions used in autoencoder, elaborate?

8. Answer the followings
    - What is a Generative Adversarial Network (GAN)?
    - Name the two main components of a GAN and state their roles.
    - What is meant by adversarial training?
    - Define generator loss and discriminator loss.

9. Implement a feedforward neural network with backpropagation using PyTorch to learn the XOR logic function for the following network architecture.

   Input layer: 2 neurons, Hidden layer: 2 neurons, Output layer: 1 neuron, Activation function: Sigmoid, Loss function: Mean Squared Error (MSE), Learning rate: 0.1.

10. Implement a Convolutional Neural Network (CNN) using PyTorch to classify images from the MNIST handwritten digits dataset. Display training and testing outcome in-terms of confusion matrix, accuracy, precision, recall and F1-score.

Dataset:
MNIST (Image size: 28 × 28 pixels, Color channels: 1 (grayscale), Number of classes: 10 (digits 0–9), Training samples: 60,000, Testing samples: 10,000)

Architecture details:
Input Layer: Input shape (28, 28, 1)
Convolution Layer 1: Number of filters: 32 Kernel size: 3 × 3 Stride: 1 Padding: valid
Activation function: ReLU
Max Pooling Layer 1: Pool size: 2 × 2 Stride: 2
Convolution Layer 2: Number of filters: 64 Kernel size: 3 × 3 Activation function: ReLUMax Pooling Layer 2: Pool size: 2 × 2
Fully Connected Layer: Number of neurons: 128 Activation function: ReLU
Output Layer: Number of neurons: 10 Activation function: Softmax

INSTITUTE OF TECHNICAL EDUCATION AND RESEARCH

(SOA Deemed to be University)

Assignment
Large Language Models (CSE 4357)

Submitted By

Name: _____

Registration No.:_____

Branch: _____

Semester:_____ Section:_____

ASSIGNMENT
Large Language Models (CSE 4357)

Programme: B. Tech (CSE) **Semester: 7th**
Full Marks: 20 **Date of Submission: 20-12-2025**

Subject/Course Learning Outcome	*Taxonomy Level	Question Nos.	Marks
Understand the fundamental concepts of language models, including tokenization and the representation of text as vector embedding for language processing	L2	1	2
Understand and explain the core mechanisms of the Transformer architecture used in modern large language models	L2	2	2
Develop skills to categorize and cluster text data using large language models for text classification and clustering tasks.	L2	3, 4	4
Apply dense retrieval, reranking and retrieval augmented generation methods to enhance traditional keyword-based search systems	L3	5, 6, 7	6
Develop the ability to work with multimodal LLMs by understanding image-to-vector transformations and applying them to visual reasoning tasks	L2	8	2
Understand and implement end-to-end adaptation of LLMs—including data preparation, task-specific fine-tuning, and performance assessment	L2	9, 10	4

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

Answer all questions. Each question carries equal mark.

- **Assignment scores/markings depend on neatness, clarity and date of submission.**
- **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.**

1. Consider a small training corpus with word types and frequencies: low (5), lowest (2), newer (6), and wider (3). Using a SentencePiece tokenization procedure, perform 2 merge iterations and generate the sub-words**.** At each iteration:

    a. Compute Log- likelihood loss computation.
    b. Find the candidates for removal.
    c. Update the corpus representations and recompute the frequencies for the next iteration.

2. Explain the Transformer architecture, focusing on the self-attention mechanism. Discuss the role of the Query (Q), Key (K), and Value (V) vectors in computing the attention score for a token.

3. Consider three document embeddings:

D1 = [0.1, 0.5, 0.9]
D2 = [0.2, 0.6, 0.8]
D3 = [0.9, 0.2, 0.4]

a. Calculate pairwise Cosine similarity, Euclidean distance, and Dot product between all documents.
b. Explain which similarity metric is generally preferred for clustering text embeddings in high-dimensional spaces.

4. Explain how prompting affects the performance of large language models and provides examples of different types of prompting.

5. Compare semantic search with keyword-based search in terms of scalability, accuracy, and computational cost. Provide one example where keyword search may still be preferable. What is the role of embeddings in semantic search pipelines? How do they capture meaning beyond keywords?

6. How does the perplexity metric help in evaluating the quality of text generated by a language model? Explain briefly. Additionally, a language model assigns probabilities of 0.25, 0.10, 0.20, and 0.05 to a four-token sequence. Using these probabilities, calculate the perplexity of the model for this sequence and show the formula and final numerical value.

7. The figure below shows the output of an information retrieval system on two queries. Crosses correspond to the relevant documents, dashes to non-relevant documents. Let the two documents contain 3 and 6 relevant documents, respectively, but only those shown in the figure are retrieved by the system, not the others.

Rank	1	2	3	4	5	6	7	8	9	10
Q1	X			X	X					
Q2		X		X		X	X		X	

a. Calculate precision at K (where K=5)
b. Compute the average precision(AP)
c. Compute the mean average precision(MAP)
d. Normalized Discounted Cumulative Gain (nDCG) (where relevance grade is binary)

8. Explain what makes a language model “multimodal.” How does it differ from unimodal text-only LLMs?

9. State and compare sentence-level embeddings with document-level embeddings. Which would be more suitable for a legal search engine, and why?

10. Explain the difference between fine-tuning, instruction-tuning, and RLHF. How do these methods improve LLM performance and safety?

INSTITUTE OF TECHNICAL EDUCATION AND RESEARCH

(SOA Deemed to be University)

Assignment
Introduction to Software Engineering Approach (CSE 4658)

Submitted By

Name: _____

Registration No.:_____

Branch: _____

Semester:_____ Section:_____

INDEX

Name: Section:

Registration No.:

S. No.	Title	Page No.	Date of Submission	Teacher's Sign/Remark

Assignment-1

1. What is the main difference between programming and software engineering?

2. What does the phrase “Hyrum’s Law” state in software engineering?

3. Mention two reasons why long-term projects need continuous upgrades?

4. What is meant by the phrase “Shifting Left” in software development?

5. What is trade-offs in engineering decisions?

6. Mention two challenges faced by software engineers when working in teams.

7. How does knowledge sharing benefit teams?

8. Define the role of a team lead.

9. Why should team members admit mistakes openly?

10. How does knowledge sharing benefit teams?

11. Explain with example why a low Bus Factor is dangerous for software projects.

12. Why is feedback culture important in software teams?

ASSIGNMENT 2

1. What does "equity" mean in the context of software engineering teams and why is it important in software engineering workplaces?
2. Why conducting blind hiring processes promotes equity in software engineering teams?
3. What was the main reason Google's image recognition system made serious classification mistakes?
4. Who discovered the mistakes in Google's image recognition system? Why internal testing at Google failed to detect the bias?
5. Explain haunting graveyard in code and why they are haunted?
6. Write short note on tribal knowledge.
7. Explain Googleness in Software Development.
8. Low bus factor is a red flag for project continuity. Explain with real life example.
9. Draw a use case diagram for a ticket distributor for a train system. The system includes two actors: a traveler, who purchases different types of tickets, and a central computer system, which maintains a reference database for the tariff. Use cases should include: BuyOneWayTicket, BuyWeeklyCard, BuyMonthlyCard, UpdateTariff. Also include the following exceptional cases: Time-Out (i.e., traveler took too long to insert the right amount), TransactionAborted (i.e., traveler selected the cancel button without completing the transaction), DistributorOutOfChange, and DistributorOutOfPaper.
10. Hans and Jacqueline live in Amsterdam; she is a pilot and he is a physician, both have a very busy schedule. To schedule their shopping and household activities, they develop software, Easy Shop, H & J (Hans and Jacqueline) Easy Shop. H&J should be able to enter their presence for meals for each day, to register invitations for guests for each meal to fix a menu for cold meals with given ingredients, to choose whether they will cook in their agenda, and if so, fix a recipe for each warm meal to prepare a weekly shopping list, and fax the shopping list to the supermarket for delivery, to keep track of ingredients in the kitchen; give a report when needed to show recipes (e.g., when cooking a dish).

Draw use case diagram and class diagram for the given problem.