Registration No.: | | | | | | | | | | |

Total Number of Pages: 03

Course: M.Tech
Sub_Code: 24PC1005

2nd Semester Regular Examination: 2025-26
SUBJECT: Machine Learning
BRANCH(S): CSEDS, Data Sciences
Time: 3 Hours
Max Marks: 100
Q.Code: V551

Answer Q1 (Part-I) which is compulsory, any eight from Part-II, and any two from Part-III.
The figures in the right-hand margin indicate marks.

Part-I

Q1 **Answer the following questions:** **(2 x 10)**

a) Out of 200 data points, 120 are labelled, and 80 are unlabelled. Identify the appropriate learning paradigm.

b) Define the concept of linear transformation in data representation.

c) A dataset has 4 positive and 4 negative samples. Compute the entropy of the dataset.

d) Define overfitting and its impact on model performance.

e) Explain the roles of training, validation, and test datasets.

f) Define the concept of ensemble learning. A Random Forest has 7 trees giving predictions: A, B, A, A, B, A, B. Determine the final predicted class using majority voting.

g) Given points: (1, 1), (2, 2), (8, 8), (9, 9) and initial centroids C1 = (1, 1), C2 = (9, 9): Compute the new centroids after one iteration.

h) What are generative models? How do they differ from discriminative models?

i) Define the role of loss functions in training ML models.

j) Define PAC learning and its significance.

Part-II

Q2 **Only Focused-Short Answer Type Questions- (Answer Any Eight out of Twelve)** **(6 x 8)**

a) What is feature representation in Machine Learning? Explain the importance of selecting good features for model performance. Illustrate with an example how poor feature representation can affect classification accuracy.

b) What do you mean by problem formulation in Machine Learning? Differentiate between classification and regression problems with suitable examples. A dataset contains information about houses with features such as size (in sq. ft.) and price.

Consider the following data points:

Size (sq. ft.)	Price (in ₹ lakhs)
1000	50
1500	75
2000	100

Identify whether this problem is classification or regression, with justification.

c) State and explain different paradigms of Machine Learning. Describe in detail Supervised Learning and Unsupervised Learning with suitable examples. Also, provide a brief overview of other learning paradigms such as Reinforcement Learning and Semi-supervised Learning.

d) Explain the concept of Principal Component Analysis (PCA) and its role in dimensionality reduction. How does PCA transform high-dimensional data into a lower-dimensional space? Briefly describe the importance of variance in selecting principal components.

e) Write the principles of Nearest Neighbour and K-Nearest Neighbours (KNN) algorithms. How does KNN perform classification? Discuss the role of the parameter *K* and the choice of distance metric. Consider the following training data:

Point	X1	X2	Class
A	1	1	0
B	2	2	0
C	3	3	1
D	6	5	1

A new point P = (2, 3) needs to be classified using KNN with $K = 3$ and Euclidean distance.

(i) Compute the distance of point P from all training points.
(ii) Identify the 3 nearest neighbours.
(iii) Determine the predicted class of point P.

f) Describe the working principles of Linear Regression. How does it model the relationship between dependent and independent variables? Define the terms slope and intercept, and explain their significance. Given the following data points:

x	y
1	2
2	4
3	6

(i) Determine the equation of the best-fit line using the form $y = mx + c$.
(ii) Using the obtained model, predict the value of y when $x = 4$.

g) Describe the concept and underlying principles of Logistic Regression. How is it used for binary classification? Describe the role of the sigmoid function and how probabilities are converted into class labels.

h) What is ensemble learning? Explain how combining multiple models improves performance. Differentiate between Bagging and Boosting. Give one example of each method.

i) What do you mean by overfitting neural networks? Describe how L1/L2 regularisation and early stopping help in improving generalisation.

j) Write the working of Naive Bayes classifier. State the conditional independence assumption and discuss how it simplifies probabilistic classification. Consider a binary classification problem with two classes: C1 and C2.

The following probabilities are given:

$P(C1) = 0.6$, $P(C2) = 0.4$, $P(x1 = 1|C1) = 0.5$, $P(x2 = 1|C1) = 0.4$, $P(x1 = 1|C2) = 0.3$, $P(x2 = 1|C2) = 0.8$

For a new instance with features $x1 = 1$, $x2 = 1$:

(i) Compute the posterior probabilities $P(C1|x)$ and $P(C2|x)$ (ignore normalization constant).
(ii) Determine the predicted class.

k) How does Gradient Descent work, and why is it important for optimizing machine learning models? Describe the working of the Perceptron algorithm for binary classification and how weights are updated.

l) Explain the concept of a Hidden Markov Model (HMM). Define the components: States, Observations, Transition probabilities and Emission probabilities. Briefly describe how HMMs are used for sequence modeling.

Part-III

Only Long Answer Type Questions (Answer Any Two out of Four)

Q3 What are the key factors that have driven the adoption of Machine Learning in modern computer science? Discuss with suitable examples how Machine Learning helps in solving complex real-world problems compared to traditional rule-based programming approaches. **(16)**

Explain the concept of a Linear Support Vector Machine (SVM). Define Hyperplane, Margin, and Support Vectors in SVM. Given a hyperplane defined as: $w.x + b = 0$

Where $w = (2, -1)$, $b = -3$

(i) Determine whether the point $x1 = (2, 3)$ lies above, below, or on the hyperplane.

(ii) Find the distance of the point from the hyperplane.

Q4 Describe the structure and components of a Multilayer Perceptron (MLP). Discuss the role of hidden layers and activation functions. How does an MLP handle classification and regression tasks differently? Describe the backpropagation algorithm used for training an MLP. Explain how errors are propagated backward and how weights are updated using gradient descent. **(16)**

Consider a single neuron with: Input $x = 2$, Weight $w = 0.5$, Bias $b = 0$, Activation function: Linear ($f(x) = x$), Target output $y = 2$, Learning rate $\eta = 0.1$

(i) Compute the predicted output.

(ii) Calculate the error using Mean Squared Error.

(iii) Update the weight w using one step of gradient descent.

Q5 Explain the working of a Decision Tree classifier. Discuss how a tree is constructed using a splitting criterion. Define Entropy and Information Gain. Consider a dataset with 10 samples: 6 belong to Class A and 4 belong to Class B. Calculate the entropy of the dataset. If an attribute splits the dataset into two subsets: **(16)**

- Subset 1: 4 A, 0 B
- Subset 2: 2 A, 4 B

Compute the Information Gain for this split.

Explain the working of the Random Forest algorithm. How does it overcome overfitting in decision trees? Discuss the role of feature randomness.

Q6 What do you mean by Deep Learning? Discuss how deep learning differs from traditional machine learning approaches. Describe the architecture of a Convolutional Neural Network (CNN). Explain the purpose of the following layers: **(16)**

- Convolutional Layer
- Pooling Layer
- Fully Connected Layer

An input image of size $32 \times 32 \times 3$ is passed through a convolutional layer of a CNN with Number of filters = 10, Filter size = 5×5, Stride = 1, Padding = 0. Calculate the output feature map size.

Registration No.: | | | | | | | | | | |

Total Number of Pages: 02 **Course: M.Tech**
Sub_Code: 24PC1006

2nd SEMESTER REGULAR EXAMINATION 2025-26
SUBJECT: RESEARCH METHODOLOGY, ETHICS AND IPR
BRANCH(S): CSEAIML, CSEDS, DATA SCIENCES
Time: 3 Hours
Max Marks: 100
Q.Code: V415
Answer Question No.1 (Part-I) which is compulsory, any eight from Part-II and any two from Part-III.
The figures in the right-hand margin indicate marks.

Part-I

Q1 **Answer the following questions:** **(2 x 10)**

a) What is the full form SPSS?
b) What is copyright piracy?
c) What is trademark?
d) What is Gillette defense?
e) What is licensing?
f) What is hypothesis?
g) What is factor analysis?
h) What is Likert scale?
i) What is stratified sampling?
j) What is research design?

Part-II

Q2 **Only Focused-Short Answer Type Questions- (Answer Any Eight out of Twelve)** **(6 x 8)**

a) Why is research important in business and management?
b) Explain the various steps involved in research process.
c) Research is a systematic search for truth-Discuss this statement.
d) How does research methodology ensure reliable results?
e) What is Type-I and Type-II errors? Explain in detail.
f) Explain the various objectives and steps of factor analysis.
g) Discuss the complete framework of statistical data analysis in research.
h) Discuss data processing steps and its importance.
i) Why is copyright protection necessary for author and artist in modern protection?
j) Why are certain inventions excluded from patent?
k) Discuss the future of IPR in a globalized digital economy.
l) Explain six differences between a copyright and patent.

Part-III

Only Long Answer Type Questions (Answer Any Two out of Four)

Q3	Discuss the role of intellectual property rights in promoting innovation and economic development.	**(16)**
Q4	Define research methodology. Explain the relationship between research problems identification and hypothesis development.	**(16)**
Q5	What do you mean by non-parametric test? Explain the advantages and limitations of non-parametric statistical methods in research.	**(16)**
Q6	Why is copyright protection necessary? Explain its importance in promoting creativity, innovation, and economic development.	**(16)**

Registration No.: | | | | | | | | | | |

Total Number of Pages: 02 **Course: M.Tech**
Sub_Code: 24PC1007

2nd Semester Regular Examination: 2025-26
SUBJECT: Big Data Analytics
BRANCH(S): CSEAIML, CSEDS, Data Sciences
Time: 3 Hours
Max Marks: 100
Q.Code: V468

Answer Q1 (Part-I) which is compulsory, any eight from Part-II, and any two from Part-III.
The figures in the right-hand margin indicate marks.

Part-I

Q1 **Answer the following questions:** **(2 x 10)**

a) Define Big Data and explain its characteristics.
b) What are the challenges in Big Data Analytics?
c) Differentiate between schema-less and schema-based databases.
d) What is a key-value store?
e) Explain Hadoop Distributed File System (HDFS).
f) What is Pig Latin?
g) Define HiveQL and its purpose.
h) What is Spark stream mining?
i) Explain sentiment analysis in Big Data.
j) What is MapReduce?

Part-II

Q2 **Only Focused-Short Answer Type Questions- (Answer Any Eight out of Twelve)** **(6 x 8)**

a) Explain the need for new analytical architecture in Big Data.
b) Discuss document stores and their applications.
c) What are graph databases? Provide an example.
d) Explain HDFS performance tuning.
e) Write a Pig Latin script to filter student records with marks > 60.
f) Explain Hive tables and querying data.
g) What is HBase? Discuss its features.
h) Explain sampling and filtering in data streams.
i) Discuss real-time analytics platforms (RTAP).
j) Explain Big Data applications in e-commerce.
k) What is the role of user-defined functions in Pig/Hive?
l) Explain stock market prediction using Big Data streams.

Part-III

Only Long Answer Type Questions (Answer Any Two out of Four)

Q3	**a)**	Discuss the typical analytical architecture for Big Data with a diagram.	**(8)**
	b)	Explain the challenges in Big Data frameworks with suitable examples.	**(8)**
Q4	**a)**	Compare key-value, document, and graph databases with use cases.	**(8)**
	b)	Design a NoSQL schema for a Twitter analytics application.	**(8)**
Q5	**a)**	Implement MapReduce for word count problem and explain the steps.	**(8)**
	b)	Discuss the Hadoop ecosystem components (Pig, Hive, HBase) with examples.	**(8)**
Q6	**a)**	Design a real-time sentiment analysis system for stock market prediction using Spark streams.	**(8)**
	b)	Explain Big Data applications in healthcare and IoT with case studies.	**(8)**

Registration No.: | | | | | | | | | | |

Total Number of Pages: 02 **Course: M.Tech**
Sub_Code: 24PC1008

2nd Semester Regular Examination: 2025-26
SUBJECT: Advanced Data Visualization
BRANCH(S): CSEDS, Data Sciences
Time: 3 Hours
Max Marks: 100
Q.Code: V609

Answer Question No.1 (Part-I) which is compulsory, any eight from Part-II and any two from Part-III.
The figures in the right-hand margin indicate marks.

Part-I

Q1 **Answer the following questions:** **(2 x 10)**

a) Define data acquisition.
b) What is data annotation?
c) Define data cleaning.
d) State the purpose of data integration.
e) What is data reduction?
f) What is meant by visualizing complex data?
g) State the purpose of correlation visualization.
h) Define residual analysis.
i) What is a prediction interval?
j) What is autocorrelation?

Part-II

Q2 **Only Focused-Short Answer Type Questions- (Answer Any Eight out of Twelve)** **(6 x 8)**

a) Discuss the complete data preprocessing pipeline with suitable illustrations.
b) Explain basic charts and plots and discuss how they are used for effective data communication.
c) Describe multivariate data visualization, its challenges, and commonly used techniques.
d) Explain rank analysis tools and their role in comparative data analysis.
e) Discuss distribution analysis tools, including visualization of statistical distributions.
f) Describe geographical analysis tools and explain how spatial data is visualized.
g) Explain the regression model building framework, including problem definition, data preprocessing, model building, diagnostics, and validation.
h) Discuss various significance tests used in simple linear regression and their interpretations.
i) Explain heteroscedasticity, its detection methods, and corrective measures.

j) Discuss multicollinearity, its causes, effects, and remedies.

k) Explain various transformations of variables and their role in improving regression models.

l) What is meant by transformation of variables? Explain in brief.

Part-III

Only Long Answer Type Questions (Answer Any Two out of Four)

Q3 Explain pixel-oriented visualization techniques in detail, highlighting their working principle, advantages, and limitations. **(16)**

Q4 Describe trend analysis tools with examples showing time-series data visualization. **(16)**

Q5 Explain the coefficient of determination (R^2) and its role in evaluating model performance. **(16)**

Q6 Explain the identification and treatment of outliers in multiple regression. **(16)**

Registration No.: | | | | | | | | | | |

Total Number of Pages: 02 **Course: M.Tech**
Sub_Code: 24PE1005

2nd Semester Regular Examination: 2025-26
SUBJECT: Natural Language Processing
BRANCH(S): CSEDS
Time: 3 Hours
Max Marks: 100
Q.Code: V702

Answer Q1 (Part-I) which is compulsory, any eight from Part-II and any two from Part-III.
The figures in the right-hand margin indicate marks.

Part-I

Q1 **Answer the following questions:** **(2 x 10)**

a) Differentiate between bigram and trigram.
b) Describe why production rules with zero probability are problematic.
c) 3-grams are better than bigrams for part-of-speech tagging. Is it true or false? Explain your answer.
d) Information extraction is harder than text categorization. Is it true or false? Explain your answer
e) What do you understand by cohesion?
f) List the various issues of machine translation system.
g) What are the problems with PCFG?
h) What do you mean by inverse document frequency (IDF)?
i) What do you mean by morphology?
j) What is the difference between wordNet and frameNet?

Part-II

Q2 **Only Focused-Short Answer Type Questions- (Answer Any Eight out of Twelve)** **(6 x 8)**

a) Discuss Viterbi Algorithm in detail.
b) Explain the role of knowledge representation in Question Answering Systems.
c) What do you mean by anaphora resolution? Discuss its types and applications.
d) How to do word sense disambiguation? Explain with example.
e) Write short note on Research corpora.
f) Discuss stochastic part-of-speech tagging.
g) Explain Context-Free Grammar (CFG) in detail.
h) How to deal with spelling error detection and correction? Explain.
i) Discuss Syntax Analysis using Dependency Graph with examples.
j) Explain generative model for parsing.
k) What is text summarization? Explain with an example.
l) What is Binding Theory?

Part-III

Only Long Answer Type Questions (Answer Any Two out of Four)

Q3 Explain the Hidden Markov Model (HMM) approach for POS Tagging. Discuss transition probability, emission probability, Viterbi algorithm, and tagging process with examples. **(16)**

Q4 Explain Earley parser and CYK parser with suitable examples. **(16)**

Q5 Describe different approaches to Word Sense Disambiguation. Explain knowledge-based, supervised, unsupervised, and hybrid approaches with advantages and disadvantages. **(16)**

Q6 Explain the architecture of Statistical Machine Translation (SMT). Discuss the major components such as language model, translation model, decoder, alignment model, and training corpus with a neat diagram. **(16)**