

Registration No.:

--	--	--	--	--	--	--	--	--	--

Total Number of Pages: 02

Course: M.Tech  
Sub\_Code: 23PC1005

2<sup>nd</sup> Semester Regular Examination: 2023-24  
SUBJECT: Machine Learning  
BRANCH(S): COMPUTER SCIENCE & ENGINEERING(DATA SCIENCE)

Time: 3 Hour

Max Marks: 100

Q.Code: P131

Answer Question No.1 (Part-1) 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) Under what conditions is successful learning possible?
- b) What do you mean by noise in data? How it affects the results?
- c) Define MLP.
- d) What is the representational power of perceptron?
- e) What is confusion matrix and why do we need it?
- f) What do you understand by overfitting of data? Give any two methods to avoid over fitting.
- g) Derive the comparison between Instance based and Bayesian learning.
- h) What are various algorithms for dimensionality reduction?
- i) Define Precision and Recall.
- j) State what is Machine Learning? Also mention when to use it.

**Part-II**

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

- a) Explain the conditions in which Gradient Descent is applied? Distinguish between Gradient Decent and Stochastic Gradient Descent.
- b) Explain in detail Principal Component Analysis for Dimension reduction process.
- c) What do you mean by Training, Validation, and Testing? Explain k-fold cross validation techniques.
- d) Explain the probabilistic generative model in brief.
- e) What do you mean by feature extraction? Explain the PCA for feature extraction.
- f) Explain how Support Vector Machine can be used for classification of linearly separable data.
- g) Consider a set of five 2-dimensional points  $p_1(0, 0)$ ,  $p_2(5, 0)$ ,  $p_3(5, 1)$ ,  $p_4(0, 1)$  and  $p_5(0, 0.5)$ . Complete linkage clustering is used to cluster the points into two clusters. Find the two clusters.

- h) Why the estimate accuracy of Hypothesis is necessary? Explain the process to estimate the error between any two learning methods.
- i) How does perceptron act as a heuristic learning algorithm for linear classifier? Explain.
- j) Explain the terms Gain and Entropy? How it is used to build the Decision tree in algorithm? Illustrate using an example.
- k) What is kernel in SVM? Explain popular kernels used in SVM along with a scenario of their application.
- l) Suppose 10000 patients get tested for flu; out of them 9000 are actually healthy and 1000 are actually sick. For the sick people, a test was positive for 620 and negative for 380. For the healthy people, the same test was positive for 180 and negative for 8820. Construct a confusion matrix for the data and compute the precision and recall for the data.

### Part-III

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

- Q3 Discuss what is Support Vector Machine? Illustrate how to compare the margin. Explain how SVM can be used to find Optimal Hyper plane to classify linearly separable data with suitable example. (16)
- Q4 What are the limitations of back propagation algorithm? Why back propagation algorithm is important in neural network? (16)
- Q5 Discuss how a multilayer network learns using a gradient decent algorithm. (16)
- Q6 Discuss the Bayes Rule with an example and illustrate how it is applied to pattern classification problems. Justify that in multiclass classification task the Bayes decision rule minimizes the error probability. State the significance of Radial basis function. (16)

Registration No.:

--	--	--	--	--	--	--	--	--	--

Total Number of Pages: 02

Course: M.Tech  
Sub\_Code: 23PE1005

2<sup>nd</sup> Semester Regular Examination: 2023-24

SUBJECT: Natural Language Processing

BRANCH(S): CSE (Data Science)

Time: 3 Hours

Max Marks: 100

Q.Code: P577

Answer Question No.1 (Part-1) 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)**

- Differentiate between bigram and trigram.
- Describe why production rules with zero probability are problematic.
- 3-grams are better than bigrams for part-of-speech tagging. Is it true or false? Explain your answer.
- Information extraction is harder than text categorization. Is it true or false? Explain your answer.
- What do you understand by ambiguity?
- List the various issues of machine translation system.
- What are the problems with PCFG?
- What do you mean by lemmatization?
- What do you mean by stemmer?
- What is the difference between phrase-based and feature-based NLP?

**Part-II**

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

- Discuss various relations among the word senses.
- Discuss the relative advantages and disadvantages of partial versus full parsing.
- What do you mean by anaphora resolution? Discuss its types and applications.
- How to do a discourse analysis? Explain with example.
- Write short note on WordNet.
- Discuss stochastic part-of-speech tagging.
- Discuss any two advanced applications of NLP.
- How to deal with spelling error detection and correction? Explain.
- Why is word sense disambiguation important for language technology? Explain with examples.

- j) Explain how unification is implemented.
- k) What is text summarization? Explain with an example.
- l) Explain the application of semantics.

**Part-III**

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

- Q3** Explain the issues in computational morphology with suitable example. **(16)**
- Q4** Explain Earley parser and CYK parser with suitable examples. **(16)**
- Q5** Explain the natural language understanding (NLU) system architecture with neat sketch diagram. **(16)**
- Q6** Elaborate on the impact of Natural language processing in online search. **(16)**

Registration No.:

--	--	--	--	--	--	--	--	--	--

Total Number of Pages: 02

Course: M.TECH  
Sub\_Code: 23PC1006

2<sup>nd</sup> Semester Regular Examination: 2023-24  
SUBJECT: RESEARCH METHODOLOGY ETHICS AND IPR  
BRANCH(S): COMPUTER SCIENCE & ENGINEERING (DATA SCIENCE)

Time: 3 Hour

Max Marks: 100

Q.Code: P224

Answer Question No.1 (Part-1) 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)
- Write about the types of research.
  - What are the applications of Z-test?
  - What are COPE and WAME?
  - Write about the need for intellectual property right.
  - What are the conflicts of interest?
  - What can be protected by a patent?
  - What are non-Patentable subject matters?
  - Write about Gillette defense.
  - What is covered by copyright?
  - Write the difference between the trademarks and trade Secret.

Part-II

- Q2 Only Focused-Short Answer Type Questions- (Answer Any Eight out of Twelve) (6 x 8)
- Write about the importance of scientific research in decision making.
  - Explain about the identification of research problem and formulation of hypothesis.
  - Explain about the Krushall Wallis test.
  - Write the procedure of hypothesis testing.
  - Write about the publication ethics and its importance.
  - Explain the problems that lead to unethical behavior in publication misconduct.
  - Discuss about the nature and significance of intellectual property right (IPR).
  - Write about patent law, patentability requirements with suitable examples.
  - Explain about rights of patentee and working of Patent.
  - Discuss about transfer and termination of transfers and its applications.
  - Explain about the Clickwrap agreements.
  - Explain about the requirements and misappropriation of trade secret.

**Part-III**

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

- Q3** Critically explain measurement and data Collection. Primary data, secondary data, design of questionnaire; with suitable relevant examples. Also explain about factor analysis. **(16)**
- Q4** Explain about the violation of publication ethics, authorship, and contributorship. Identification of publication misconduct, complaints, and appeals. **(16)**
- Q5** Write the procedure for domestic and international filing of applications of patent. Also write about IPR in India Genesis and development. **(16)**
- Q6** Discuss about equivalent doctrine, Pith and Marrow doctrine. Also discuss about Literal and non-literal infringement. **(16)**

Registration No.:

--	--	--	--	--	--	--	--	--	--

Total Number of Pages: 02

Course: M.Tech  
Sub\_Code: 23PC1007

2<sup>nd</sup> Semester Regular Examination: 2023-24

SUBJECT: Big Data Analytics

BRANCH(S): COMPUTER SCIENCE & ENGINEERING (DATA SCIENCE)

Time: 3 Hour

Max Marks: 100

Q.Code: P287

Answer Question No.1 (Part-1) 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)

- What is MongoDB?
- What is MapReduce?
- When and why do we need data normalization?
- Explain CRUD operations.
- Define the various file formats supported by HIVE.
- Mention different ways of data discovery in data analytics.
- What is the command to format the NameNode?
- What is HIVE?
- Which are the application areas of Big Data analytics?
- State the common input formats in Hadoop.

Part-II

Q2 Only Focused-Short Answer Type Questions- (Answer Any Eight out of Twelve)

(6 x 8)

- Discuss Hadoop 2 architecture.
- Explain the map reduce data flow with single reduce and multiple reduce.
- Discuss the NoSQL Data Architectural Patterns in Big-Data System.
- What are the various operational modes of Hadoop cluster configuration? Explain in detail about configuring/installing Hadoop in fully distributed mode.
- What is the difference between the 'set' & a 'map' data structure? Explain about various map implementations in java with suitable examples.
- Give a detail note on HBASE.
- What are the three V's of Big Data system with examples?
- What is benchmarking how it works in Hadoop?

- i) Explain how map reduce jobs run on YARN.
- j) Discuss Data Analytics Lifecycle with examples.
- k) How security does is done in Hadoop? Justify.
- l) Explain about Big Data Computational Limitations.

### Part-III

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

- Q3** Explain the Hadoop distributed file system architecture with a neat sketch. How google file system differs from the Hadoop file system? Explains the google file system architecture with a neat sketch. **(16)**
- Q4** Explain the architecture of HIVE with a neat sketch. What is PIG? Explain its installation process. Explain two execution types or modes in PIG. **(16)**
- Q5** What do you mean by NOSQL? Describe various business drivers, data architectural patterns of NOSQL. Briefly explain How NOSQL can be used to manage BIG DATA. **(16)**
- Q6** How to Build a Real-Time Twitter Analysis Application Using Big Data Tools? **(16)**



Registration No.:

--	--	--	--	--	--	--	--	--	--

Total Number of Pages: 02

Course: M.Tech  
Sub\_Code: 23PC1008

2<sup>nd</sup> Semester Regular Examination: 2023-24  
SUBJECT: Advanced Data Visualization  
BRANCH(S): COMPUTER SCIENCE & ENGINEERING (DATA SCIENCE)

Time: 3 Hour

Max Marks: 100

Q.Code: P417

Answer Question No.1 (Part-1) 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)

- Define Data Acquisition.
- Why data visualization is important?
- What is the geographical analysis tools used for spatial data?
- What are the key components of effective data visualization?
- What are the basic charts and plots are used in data visualizations?
- Mention few data transformation techniques used in data processing?
- What do you mean by data annotation?
- Define data reduction.
- What do you mean by Multivariate Data Visualization?
- What do you mean by outliers?

Part-II

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

- Explain Data Transformation with suitable examples.
- Describe Pixel-Oriented Visualization Techniques.
- Explain Icon-Based Visualization Techniques with examples.
- Describe Rank Analysis Tools in brief.
- What are the different Multivariate Analysis Tools? Write their uses.
- Describe Confidence and Prediction intervals.
- Explain Heteroscedasticity with example.
- What do you mean by Autoregression and transformation of variables? Explain with example.
- How can more than 3D be represented in a single chart?
- Explain in brief how regression model building is done.
- What is the significance of data cleansing in data visualization?
- What are the most important qualities of good data visualization?

**Part-III**

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

- Q3** Explain different types of data used in design process. How do you customize the appearance of a plot in Matplotlib? **(16)**
- Q4** Discuss techniques for visualizing temporal data, such as time series plots, calendar heatmaps, and event timelines. **(16)**
- Q5** Discuss techniques such as parallel coordinates, radar charts, and scatterplot matrices. **(16)**
- Q6** What is a box plot? Explain in detail how outliers are represented in a box plot. **(16)**