Instructions: This section consists of **TWO(2) sections**.
Answer **ALL** questions in BOTH SECTIONS.

Course Learning Outcome Coverage	
CLO1	Identify organizational big data problem using analytical techniques (C1, PLO1)
CLO2	Apply data warehousing concepts and data mining models to solve data exploratory and pre-processing issues (C3, PL07)

Section A: Total 30 marks (CLO2)
Answer ALL questions

QUESTION 1: DATA WAREHOUSE AND HADOOP

(a) Explain the concept of a data warehouse. How does it differ from a transactional database in terms of purpose and structure? (6 Marks)

(b) What is Hadoop, and how does it enable processing of big data? Briefly explain the role of MapReduce in the Hadoop ecosystem. (4 Marks)

QUESTION 2 DATA PREPROCESSING AND DATA EXPLORATION

(a) Explain the importance of explorative data analytics (EDA) in data preparation. Provide TWO (2) common techniques used in EDA. (4 Marks)

(b) Define data cleaning and discuss TWO (2) common issues encountered in raw datasets that require cleaning. (3 Marks)

(c) What is data normalization and the necessity of normalization? (2 Marks)

(d) Explain how outliers can impact a dataset (1 Marks)

QUESTION 3 :METHODOLOGY

(a) Explain the six phases of the CRISP-DM process and their significance in structuring a data mining project. (5 Marks)

(b) Compare KDD process with CRISP-DM. Highlight any TWO (2) similarities and TWO (2) differences between them. (5 Marks)

Section B: Total 70 marks (CLO1)
Answer ALL questions

QUESTION 4 : TEXT ANALYSIS

(a) (Explain the Vector Space Model used in text analysis. How does it assist in information retrieval? (5 Marks)

(b) What is sentiment analysis, and how does it enhance decision-making? Provide one real-world example of its application. (5 Marks)

(c) What is tokenization, and how does it help in text analysis? Provide an example of how a sentence is tokenized. (5 Marks)

QUESTION 5 : DECISION TREE

Given dataset with the following attributes (*Study Hours, Sleep Hours, Study Partner, Healthy Diet, Pass/Fail*).

Calculate the entropy for the target variable (Pass/Fail) based on the Study Hours attribute using one iteration of the decision tree construction process. (15 Marks)

Study Hours	Sleep Hours	Study Partner	Healthy Diet	Pass/Fail
High	Normal	Yes	Yes	Pass
Low	Normal	No	No	Fail
High	Low	Yes	Yes	Pass
Low	High	No	No	Fail
High	Normal	No	Yes	Pass
Low	Low	Yes	No	Fail
High	High	Yes	Yes	Pass
Low	Normal	No	No	Fail
High	Low	Yes	Yes	Pass
Low	High	No	No	Fail

QUESTION 6: CLUSTER ANALYSIS

(a) Explain K-Means clustering algorithm. Include key steps involved in the clustering process. (5 Marks)

(b) Given the following data points:

(2,3), (5,4), (9,6), (4,2), (8,7), (3,5)

Assume K = 2, and initial centroids are (2,3) and (8,7).

Assign each point to the closest centroid using Euclidean distance. (1 iteration only)

(6 Marks)

(c) Discuss ONE (1) advantage and ONE (1) limitation of K-Means clustering. (4 Marks)

QUESTION 7 : BIG DATA AND SOCIAL IMPACT

(a) Explain the significance of big data analysis in modern industries. Provide two examples of its applications across different sectors. (6 Marks)

(b) Discuss two key social impacts (positive or negative) of big data analytics. Provide relevant examples to support your argument. (4 Marks)

QUESTION 8 : NAÏVE BAYES

(a) Describe two advantages and two limitations of the Naïve Bayes classifier. (4 Marks)

(b) Given the following dataset for weather-based "Play Tennis" classification: (6 Marks)

Weather	**Play Tennis**
Sunny	No
Overcast	Yes
Rainy	Yes
Sunny	No
Sunny	Yes

Calculate (P({Sunny} | {Yes})), the probability of Tennis is played on a "Sunny" day

QUESTION 9 : ASSOCIATION RULE MINING

Apriori algorithm is significant in finding frequent item sets for association rule mining.

(5 Marks)

Consider the following transactions:

T1: {Bread, Milk, Egg}

T2: {Bread, Milk}

T3: {Milk, Egg}

T4: {Bread, Egg}

T5: {Bread, Milk, Egg}

Find frequent item sets of size 1 and 2 which complies to minimum support = 2.