Qn. Set Code-1

Semester: 6th
Programme: B.Tech
Branch: All (OE-I/Minor-I)

SPRING END SEMESTER EXAMINATION-2023

6th Semester B.Tech (Open Elective-I/Minor-I)

DATA ANALYTICS

IT 3006

(For 2021 (L.E), 2020 & Previous Admitted Batches)

Time: 3 Hours Full Marks: 50

Answer any SIX questions.
Question paper consists of four SECTIONS i.e. A, B, C and D.
Section A is compulsory.
Attempt minimum one question each from Sections B, C, D.
The figures in the margin indicate full marks.
Candidates are required to give their answers in their own words as far as practicable and all parts of a question should be answered at one place only.

SECTION-A

1. Answer the following questions. [1 × 10]

(a) What is the role of the JobTracker in MapReduce? Explain with a suitable example.

(b) The runs scored in a cricket match by 11 players are: 7, 16, 121, 51, 101, 81, 1, 16, 9, 11, and 16. Find the mean, mode, median of this data.

(c) Find the inter quartile range (IQR), Q0, Q1, Q2, Q3 and Q4 for the data set: 23, 45, 32, 29, 37, 47, 21, 36, and 52.

(d) What log-likelihood function (LLF) in logistic regression?

(e) How is confidence and lift calculated in association rule mining?

(f) What is the purpose of using exponential smoothing in time series forecasting?

(g) How does the ID3 algorithm choose the best attribute for splitting the dataset?

(h) Can SVM be used for multi-class classification? If yes, how?

(i) What is the role of the hyper parameter k in KNN, and how do you choose the optimal value for k?

(j) What is the meaning of support vectors in SVM, and how do they affect the decision boundary?

SECTION-B

2. (a) What are the differences between Hadoop and traditional relational databases, and when should one use Hadoop instead of a database? [4]

(b) In the population, the average IQ is 100 with a standard deviation of 15. A team of scientists want to test a new medication to see if it has either a positive or negative effect on intelligence or not effect at all. A sample of 30 participants who have taken the medication has a mean of 140. Did the medication affect intelligence? The z-value is 1.96. [4]

3. (a) Explain different challenges of Traditional Systems. [4]

(b) A random sample of 30 apples was taken from a large population. On measuring their diameter the mean diameter of the sample was 91 millimeters with a standard deviation of 8 mm. Calculate the 85% confidence limits for the mean diameter of the whole population of apples. [4]

SECTION-C

4. (a) Find frequent item sets and generate association rules for them. Illustrate it with step-by-step process. [4]

Minimum support = 2

Minimum confidence = 50%

Transaction	List of items
T1	I1, I2, I3
T2	I2, I3, I4
T3	I4, I5
T4	I1, I2, I4
T5	I1, I2, I3, I5
T6	I1, I2, I3, I4

(b) Suppose you are the manager of a retail store that sells clothing items. You are trying to forecast the sales of a particular item for the next 4 weeks. You have collected the following data on the weekly sales of this item for the past 7 weeks: [4]

Week	Sales
1	200
2	180
3	220
4	240
5	250
6	280
7	300

You decide to use simple exponential smoothing with a smoothing parameter of alpha=0.3 to forecast the sales of this item for the next 4 weeks. What will be your forecast for the sales of this item for the next 4 weeks using exponential smoothing?

5. (a) Suppose you are a data scientist working for a bank that wants to develop a model to predict whether or not a loan applicant will default on their loan. You have collected a dataset of loan applications with the following features: age, income, credit score, and employment status. You have also labeled each loan application as either "default" or "no default". [4]

You decide to use the ID3 algorithm to build a decision tree for this dataset.

a) Generate a decision tree with following details :

The root node of the tree is the feature "credit score".

The first level of the tree has two branches: one for credit scores >= 700 and one for credit scores < 700.

The second level of the tree has two branches for each of the two branches at the first level: one for income >= 50,000 and one for income < 50,000.

The third level of the tree has two branches for each of the four branches at the second level: one for age >= 35 and one for age < 35.

The leaves of the tree are labeled as either "default" or "no default".

b) You are given the following loan application to predict whether or not they will default on their loan:

Feature	Value
Age	40
Income	60,000
Credit Score	750
Employment	Full-time

What will be the predicted label (default or no default) for this loan application using the decision tree you built with the ID3 algorithm?

(b) What is SVM, and how does it work? [4]

6. (a) Cluster the following eight points (with (x, y) representing locations) into three clusters using K-mean clustering. [4]

A1(2, 10), A2(2, 5), A3(8, 4), A4(5, 8), A5(7, 5), A6(6, 4), A7(1, 2), A8(4, 9)

Initial cluster centers are: A1(2, 10), A4(5, 8) and A7(1, 2).

The distance function between two points a = (x1, y1) and b = (x2, y2) is defined as-

$P(a, b) = |x2 - x1| + |y2 - y1|$

(b) Illustrate K-Nearest Neighbor(KNN) algorithm with a suitable example. [4]

SECTION-D

7. (a) How does the Naïve Bayes classifier algorithm work, and what are the assumptions it makes about the data? [4]

(b) Suppose we have a dataset of **weather conditions** and corresponding target variable "**Play**". So using this below dataset we need to decide that whether we should play or not on a particular day according to the weather conditions. [4]

Problem: If the weather is sunny, then the Player should play or not?

	Outlook	Play
0	Rainy	Yes
1	Sunny	Yes
2	Overcast	Yes
3	Overcast	Yes
4	Sunny	No
5	Rainy	Yes
6	Sunny	Yes
7	Overcast	Yes
8	Rainy	No
9	Sunny	No
10	Sunny	Yes
11	Rainy	No
12	Overcast	Yes
13	Overcast	Yes

8. (a) What are data streams, and how do they differ from traditional batch processing? How do the challenges of data streams, such as high volume, velocity, and variability, affect the mining of data streams, and what are some common techniques for addressing these challenges? [4]

(b) Interpret Basic Model of Stream data [4]

*****