JAYPEE UNIVERSITY OF INFORMATION TECHNOLOGY, WAKNAGHAT

TEST -3 EXAMINATION- 2023

M.Tech-I Semester (CSE/IT)

COURSE CODE (CREDITS): 22M11CI112 (3) MAX. MARKS: 35

COURSE NAME: Introduction to Data Science

COURSE INSTRUCTORS: Dr. Anita MAX. TIME: 2 Hours

Note: *(a) All questions are compulsory.*

(b)Marks are indicated against each question in square brackets.

(c) The candidate is allowed to make Suitable numeric assumptions wherever required for solving problems

Q1(a) Using R, Write a program that reads a CSV file named aas.csv with columns Name, Age and Education and then display the summary statistics of the Age column.(2)

Q1 (b) Write an R program to calculate the factorial of any positive number using recursion concept. (2)

Q1(c) Explain the fundamental differences between data science and real science. Highlight how the interdisciplinary nature of data science bridges the gap between these two sciences. (1)

Q2 (a) You are reviewing four papers submitted to a conference on machine learning for medical expert systems. All the four papers validate their superiority on a standard benchmarking cancer dataset, which has only 5% of positive cancer cases. Which of the experimental setting is acceptable to you? (3)

paper i) We evaluated the performance of our model through a 5-fold cross validation process and report an accuracy of 93%.	paper ii) The area under the ROC curve on a single left out test set of our model is around 0.8, which is the highest among all the different approaches.
paper iii) We computed the average area under the ROC curve through 5-fold cross validation and found it to be around 0.75 – the highest among all the approaches.	paper iv) The accuracy on a single left out test set of our model is 95%, which is the highest among all the different approaches.

Choose among (A) paper i (B) paper i and paper iv (C) paper ii and paper iv (D) paper iii

Q2 (b) Which of these about a dictionary is false? (1)

a) The values of a dictionary can be accessed using keys	b) The keys of a dictionary can be accessed using values
c) Dictionaries aren't ordered	d) Dictionaries are mutable

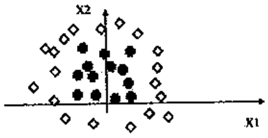
Q2 (c) What is the output of the following codes? (3)

def ChangeList():
L=[]
	numberGames = {}
	numberGames[(1,2,4)] = 8

```
    L1=[]
    L2=[]
    for i in range(1,10):
        L.append(i)
    for i in range(10,1,–2):
        L1.append(i)
    for i in range(len(L1)):
        L2.append(L1[i]+L[i])
    L2.append(len(L)-len(L1))
    print(L2)
ChangeList()
```

```
numberGames[(4,2,1)] = 10
numberGames[(1,2)] = 12
sum = 0
for k in numberGames:
 sum += numberGames[k]
print len(numberGames) + sum
```

Q3 (a) Suppose that we want to build a neural network that classifies two dimensional data (i.e., X = [x1, x2]) into two classes: diamonds and crosses. We have a set of training data that is plotted as follows:



Draw a network that can solve this classification problem. Justify your choice of the number of nodes and the architecture. Draw the decision boundary that your network can find on the diagram. (3)

Q3 (b) What is the false positive rate (FPR) for feature selection? (2)

Q3 (c) How PCA can be implemented in Python (3)

Q3 (d) Explain the advantages of using JSON as a Data format for web applications. Provide examples of real world cases where JSON is mainly used. (2)

Q3 (e) You have a collection of text documents. Implement text preprocessing pipeline that includes tokenization, stop word removal, stemming and lemmatization. Apply this pipeline to the documents and explain how it improves text analysis. (4)

Q4 (a) What is under fitting and how we can manages it using python (2)

Q4 (b) Difference between KNN Regression and Classification with code (2)

Q4(c) What is Bernoulli distribution and in which scenario we can use it with code (2)

Q4 (d) How Geographical data can be visualized in Python (3)