DST-Centre for Interdisciplinary Mathematical Sciences

M.Sc. Computational Science and Applications (Semester-III)

Mid- Semester Examination (**CSA 303 - Machine learning)**

Total time: 1 hour

Total Marks: 20

Q1. What is overfitting and underfitting? How can these issues be addressed? 2

Q2. Explain the following steps of data pre-processing with an example: 3

i) Feature discretization ii) Feature encoding iii) Feature scaling

Q3. 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. 3

Q4. Explain Naïve Bayes classifier? Predict the class label of a new instance- Medium, Middle_Aged, Yes using Naïve Bayesian Classifier for the following dataset. 5

Glucose level	Age	Family History	Developing diabetes later in life
High	Youth	Yes	Yes
Medium	Middle_Aged	No	No
Low	Senior	Yes	Yes
low	Middle_Aged	Yes	No
High	Youth	No	No
High	Senior	Yes	Yes
Medium	Senior	Yes	Yes
High	Middle_Aged	No	Yes

Q5. What is information gain? Explain by an example how it is used to find the best feature for splitting? 5

M.Sc. Computational Science and Applications
CSA-304 : Programming with R and Python
Semester: III

Time : 1 Hour **Max. marks: 20**

All questions are compulsory. Question number 1 and 2 carry 5 marks each and rest of the questions carry 1 mark each.

1. What is Data Science? Write the stages of Data Science and explain them in detail.
2. What is Python? What are the advantages of python over other programming languages? Write down the data types available in python.
3. Write a Python code snippet demonstrating the use of a 'for' loop.
4. Write a one-liner in Python to catch and handle a specific exception.
5. Write a line of code to iterate over a Python list using a 'for' loop.
6. Demonstrate how to access a value in a Python dictionary using its key.
7. Write the key difference between list and set data type
8. Create an R code snippet to generate a data frame with three columns: Name, Age, and Gender.
9. Write an R code snippet to append an element to an existing list.
10. What is the significance of visualization in the context of data science?
11. Implement an R code snippet using a 'while' loop to print the numbers from 1 to 3.
12. Create an R code snippet to multiply two matrices, A and B.

M.Sc. Semester III Mid- Sem 2023-24

Computational Science and Applications

CSA- 310: Statistical Data Mining - I

Time: 1hr *Max Marks: 20*

Note: Attempt all the questions is compulsory. Marks of each question is written on the right side.

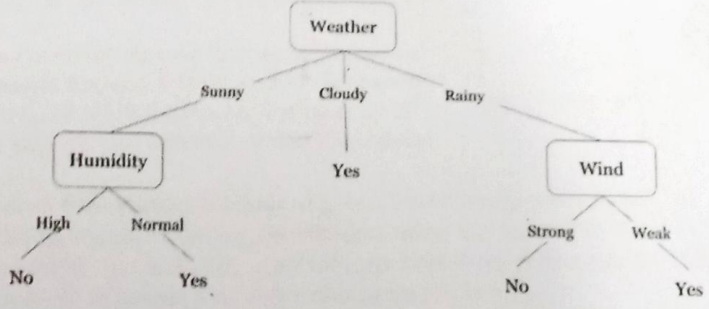
Q1) Explain the terms **(5)**

(a) Data Purging
(b) Data Matrix
(c) Regression
(d) Confusion matrix
(e) KDD (Knowledge Discovery in Database)

Q2) Differentiate Lazy Learning and Eager Learning with an example? **(3)**

Q3) Discuss the term Feature Engineering with its type? **(3)**

Q4) Explain Rule-Based Classifier and how rule-based classifier is different from another classifier. Create Rule by below decision tree? **(5)**



Q5) Using KNN (k=3) find the Class for Manish having Height– 170 and Weight – 57? **(4)**

Student	Height (cm.)	Weight (kg.)	Class
Amit	170	55	Normal
Prabhakar	173	57	Normal
Suraj	174	56	Underweight
Sumit	167	51	Underweight

M.Sc CSA

Practical Examination

Instruction:

Do any one among section A and any one among section B.

Section A

Q1. Using datasets in sklearn, create random data with 50 samples, two features, and two centres with a standard deviation of 1.05. Store them in X and Y. Split the data in the ratio of 80:20 for training and testing. Without using any library, make a class SVM that optimises the values of w and b and predicts the features of testing data.

Q2. Randomly generate a dataset of size 50, where X represents a quantity, and Y defines the corresponding label. The label can either be a class or a numerical value. You are free to generate any random dataset. Using five nearest neighbours, predict the label for the new value.

Section B.

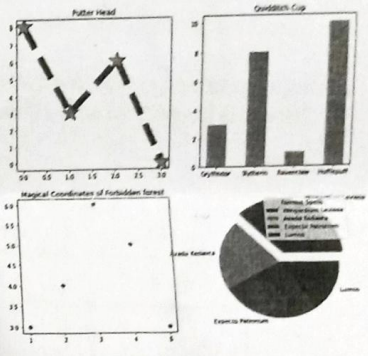
Q3. Make a subplots of graphs having 2 row and 2 columns. Example image has been given for reference.

a) Plot line graph with y values as [8, 3, 6, 0], with dashed line of width 10.5, a star marker with size 30 having yellow face and magenta edge. Give title as 'POTTER HEAD'

b) Plot a bar graph with classes as Gryffindor, Slytherin, Ravenclaw and Hufflepuff with corresponding values as 3,8,1,10. The colour of the bars should be hotpink and width as 0.5. Give title as 'QUIDDITCH CUP'

c) Plot a scatter graph with points (1,3), (3,6), (4,5), (5,3). The colour of the points should be green. Give title as "Magical Coordinates of Forbidden Forest".

d) Make a pie chart with classes Wingardium Leviosa, Avada Kedavra, Expecto Patronum and Lumos with values 35, 25, 20, 15. Explode the first class by 20 %. Include legends too, with title as "Famous spells".



Q4: Import Iris Dataset from sklearn.datasets. Perform the following operations:

a) Use describe command to display the statistics about the dataframe df1.
b) Extract the petal length and petal width for rows from 50 to 100 to another dataframe df2.
c) If any value of df1 is greater than 4.6, it should be set as 4.6.
d) Create a new array of shape (4,) that contains the the mean of corresponding columns of df1.
e) Extract all the values of df1 in a numpy array. Reshape it in 1D array. Name it arr1.
f) Store every even element of arr1 in another array named 'evenarray' and odd element in 'oddarray'.
g) Join evenarray and oddarray. Store it in 'newarray'. And then sort it.

REFNO069795

Roll No. 2241960919

M.Sc. Semester III Examination 2023-24

Computational Science and Applications

CSA-304 : Programming with R and Python

Time : Three hours **Max. Marks : 70**

(Write your roll No. at the top immediately on the receipt of this question paper)

Note: Answer any 5 questions out of the following 8 questions. Each question carries 14 marks each.

Q1. What is Data Science? Write all the stages of Data Science and explain them in detail. Write and explain all the basic data types available in R with examples.

Q2. Explain the differences between **vector**, **matrix**, **data frame**, **list** and **array** in R. Provide an example for each data type mentioned above and demonstrate how to create, access, and manipulate them.

Q3. Describe the use of **for** and **while** loops in R. Provide examples of each loop type and demonstrate their application in solving a specific problem. Discuss the purpose and usage of the next and break statements.
Write the code to make the matrix given below.

	Apple	Banana	Orange	Kiwi
Person_A	1	3	5	3
Person_B	1	5	4	1
Person_C	2	8	8	1

Q4. Write a program to add a row to a given matrix of dimension 4 x 5. After performing the previous operation, add a column to the resulting matrix. Write the output after each operation.
Write a program to create two data frames with some dummy data and then add these two data frames **vertically**.

Q5. What is Python? Write **five** advantages of python over other programming languages? Write down the data types available in python. What is the difference between "**Mutable**" and "**Immutable**" objects in python? Explain with examples.

Q6. Compare and contrast **lists**, **dictionaries**, and **sets** in Python. Provide examples of creating, accessing, and manipulating each type of data structure. Discuss the scenarios where each data structure is most suitable.

P.T.O.

Q7. Explain the concept of **recursion** in Python. Provide an example of a recursive function and discuss its advantages and limitations. Discuss the importance of exception handling in Python and provide examples of **try**, **except**, and **finally** blocks.

Q8. Write a program to read "sample.txt" which contains some text. Count the number of :

a) Words
b) Lines
c) Alphabets
d) Digits
e) Special Characters
f) Spaces
g) Numbers (eg. 1, 10, 100)

After calculating all the values, create a new file named "result.txt" and write these values into this file.

———— o ————

REENO069786

Roll No. 23413065019

M.Sc. Semester III Examination 2023-24

Computational Science and Application

CSA-309: Analysis of Multivariate Data

Time : Three hours **Max. Marks : 70**

(Write your roll No. at the top immediately on the receipt of this question paper)

Note: Attempt any Five questions. All questions carry equal marks.

1. (a) Let $\underline{X}$ be distributed as $N_3(\underline{0}, \mathbf{I_3})$ and $Y_1 = X_2 + X_3$, $Y_2 = 2X_3 - X_1$, $Y_3 = X_2 - X_1$, then find the covariance matrix of $\underline{Y}$.

   (b) Given $(X_1, X_2) \sim N_2(\mu_1, \mu_2, {\sigma_1}^2, {\sigma_2}^2, \rho)$. Find the joint density function of $Y_1 = X_1 + X_2$, and $Y_2 = X_1 - X_2$.

2. (a) Show that if a random vector $\underline{X} \sim N_p(\underline{\mu}, \Sigma)$ and $\underline{Y} = C\underline{X}$, where C is a $q \times p\,(q < p)$ matrix of rank Q, then $Y \sim N_q(C\underline{\mu}, C\Sigma C')$.

   (b) Let $\underline{X} \sim N_p(\underline{\mu}, \Sigma)$. Obtain the moment generating function of $\underline{X}$.

3. (a) Given two random samples from $N_p(\underline{\mu}^{(1)}, \Sigma_1)$ and $N_p(\underline{\mu}^{(2)}, \Sigma_2)$. Provide the test procedure for testing $H_0 : \Sigma_1 = \Sigma_2$.

   (b) Prove that $A/(n-1)$ is an unbiased estimate of Σ, where the symbols have their usual meanings.

4. Let $\underline{X} \sim N_p(\underline{\mu}, \Sigma)$. Find the probability density function of $\underline{X}$.

5. State and prove any two properties of a Hotelling's T^2 statistic known to you.

6. Define the Wishart distribution. Find the characteristic fuction of A if $A \sim W_p(\nu, \Sigma)$.

7. What is classification problem. Discuss the procedure of classification into one of the two known multivariate normal populations. Also obtain the probability of misclassification.

8. Write short notes on

   (a) Canonical correlations and canonical variates,

   (b) Principal components analysis,

   (c) Generalized variance,

   (d) Multiple correlation.

REFNO069796

Roll No. 21515(46 015

M.Sc. Semester III Examination 2023-24

Computational Science and Applications

CSA-310 : Statistical Data Mining-I

Time : Three hours **Max. Marks : 70**

(Write your roll No. at the top immediately on the receipt of this question paper)

Instructions

- Answer any 5 questions. All questions carry equal marks.
- Calculator is allowed.

Q1: (a) Explain the basic types of data. What do you mean by sampling and explain its types in brief with suitable examples.
(b) Explain the term Feature Engineering. Write short notes on data pre-processing steps: (i) Feature Transformation (ii) Feature Aggregation (iii) Feature Discretization (iv) Feature Binarization (v) Feature Sampling

Q2: (a) Differentiate between Parametric and Non-Parametric algorithm? Explain and differentiate between Linear and Logistic Regression in detail with suitable examples.
(b) Define Regularization and shrinkage. Describe method of shrinkage.

Q3: (a) Explain the k-NN Model. What is the need of k in Nearest Neighbour and how does it affect the model? The following dataset contains the length and weight of 15 metal rods along with their cost. Predict the cost for rod with a length of 7.34(M) and a weight of 8.45(KG). Assume (k=5).

Length (M)	Weight (KG)	Cost (Rs.)
10.25	25.78	45.00
11.33	6.54	57.00
12.75	14.26	48.00
7.90	9.74	33.00
9.85	14.27	38.00
15.45	36.23	60.00
12.56	11.00	35.00
15.36	10.21	50.00
14.84	8.75	46.00
7.00	12.12	35.00
10.10	6.85	36.00
13.12	8.12	44.00
19.45	17.14	32.00
5.63	8.96	60.00
5.00	10.33	50.00

(b) Define the term Overfitting and Underfitting. How they can be addressed? What is the process to reduce overfitting in the decision tree?

Q4: (a) What are drawbacks of Rule Based classifier and how it can be overcome? Explain the term Assessment of the Rule in brief and solve the below dataset for the rule.

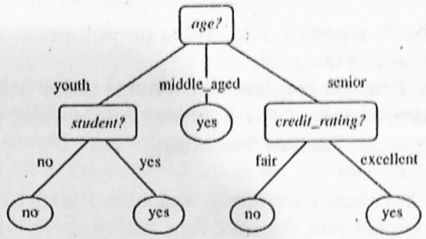
$R1: If\ (Age = Youth)\ AND\ (Student = No)\ Then\ (Buys\ Computer = Yes)$

S.No.	Age	Income	Student	Credit Rating	Buys Computer
1	Youth	High	No	Fair	No
2	Youth	High	No	Excellent	No
3	Mid Age	High	No	Fair	Yes
4	Senior	Medium	No	Fair	Yes

P.T.O.

5	Senior	Low	Yes	Fair	Yes
6	Senior	Low	Yes	Excellent	No
7	Mid Age	Low	Yes	Excellent	Yes
8	Youth	Medium	No	Fair	No
9	Youth	Low	Yes	Fair	Yes
10	Senior	Medium	Yes	Fair	Yes
11	Youth	Medium	Yes	Excellent	Yes
12	Mid Age	Medium	No	Excellent	Yes
13	Mid Age	High	Yes	Fair	Yes
14	Senior	Medium	No	Excellent	No

(b) Explain the Sequential Learning algorithm and briefly explain its mechanism steps. How we create rule from decision tree? Create rule for decision tree given below.



Q5: (a) What do you mean by Proximity measure? Explain the proximity measures for different data type. Obtain suitable measure of proximity for the following dataset.

Name	Gender	Test Type	BP	Fever	Cough	Report
Shiv	M	A	85	Y	N	P
Anjali	F	B	82	Y	Y	P
Ankit	M	A	88	N	Y	N
Sonam	F	B	84	N	N	N

(b) What do you mean by document term matrix? Explain in brief.

Q6: (a) What do you mean by "Curse of Dimensionality"? How it affects the model explain in brief.
(b) Define Principal Component Analysis (PCA) in brief? Obtain the Principal Components (PC's) for the following dataset.

Name	English	Mathematics
Priyanka	4	11
Manish	8	4
Suraj	13	5
Radhe	7	14
Ankita	6	9
Ashwani	3	10

Q7: (a) What do you mean by Subset Selection and why it is important in data mining? Explain any two techniques for subset selection in brief with suitable example.
(b) Explain the Latent Variable Model in brief? How it is different from PCA?

Q8: Write short notes on any three of the following:
(a) Discriminant Analysis **(b)** Data Quality **(c)** Sparse Data Matrix **(d)** Pessimistic error estimate **(e)** Accuracy and Error rate in regression

———— o ————

REFNO069785

Roll No. 22413CAS013

M.Sc. Semester III Examination 2023-24

Computational Science and Applications

CSA-303: Machine Learning

Time : Three hours **Max. Marks : 70**

(Write your roll No. at the top immediately on the receipt of this question paper)

Note: Attempt any five out of eight given questions. Marks of each question are written on the right side.

Q1	a) Write different types of machine learning techniques with suitable examples.	7
	b) Explain at least two techniques for finding optimal number of clusters.?	7
Q2	a) What is the K-fold cross validation method? Write its importance.	7
	b) What is the confusion matrix? Explain any two metrics to evaluate machine learning models.	7
Q3	a) Explain multiple linear regression model. How can we evaluate performance of different regression models ?	7
	b) Explain one hot encoding. What is dummy variable trap? Explain it with an example.	7
Q4	a) When do we use regression splines? Write its advantages over linear/polynomial regression.	7
	b) Explain any one technique for dimensionality reduction.	7
Q5	a) Discuss in detail decision tree splitting based on variables having real values.	7
	b) Explain Chi-square method for decision tree splitting with a suitable example.	7
Q6	a) Explain support vector machine (SVM) for binary classification in case of linear problem.	7
	b) Explain k-means clustering algorithm. What is random initialization trap? What is the solution for that?	7
Q7	a) What do you understand by discriminant function? Define the decision boundary under multivariate normal distribution in the case where all the components of the feature vector X are statistically independent and each component has the same variance and we are assuming this is same for all the classes (w_i).	7
	b) Define Bayes minimum error and Bayes minimum risk classifier. In which case both will be the same?	7
Q8	Write short notes on any of the two a) Data quality and remediations b) Pruning a decision tree c) Gaussian Mixtures as Soft K-means Clustering	7x2

—— o ——

REFNO069781

Roll No. 22515CA5019

M.Sc. Semester III Examination 2023-24

Computational Science and Applications

CSA-301 : Theory of Computation

Time : Three hours **Max. Marks : 70**

(Write your roll No. at the top immediately on the receipt of this question paper)

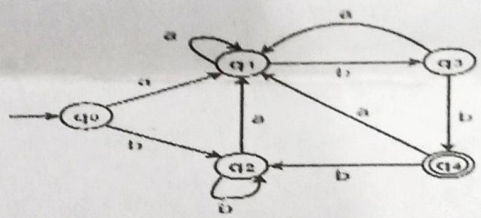
Note: Attempt any FIVE questions. The figures in right margin indicate the marks.

1. **(a).** Define Turing machine and explain how it is different from general purpose computers? 4

   (b). What do you understand by grammar. Discuss the Chomsky classification of grammar by taking appropriate examples. 4

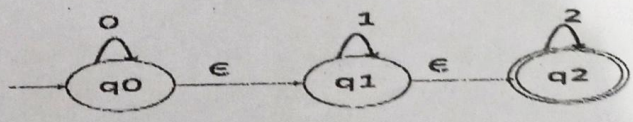
   (c). Differentiate between Deterministic (DFA) and Non-Deterministic (NDFA) Finite Automata. Design a deterministic finite automata machine which accept only even integer numbers. 6

2. **(a)** Explain the Myhill-Nerode theorem for minimization of deterministic finite automata. Find the deterministic finite automata with the minimum number of states (show each step) using the Myhill-Nerode theorem for a given DFA. 9



   (b) State pumping lemma for regular expression? What are the applications of pumping lemma? By using pumping lemma for regular expression prove/disprove whether the given language is regular $L = \{ a^{(n^2)} \mid n \geq 1 \}$. 5

3. **(a).** What do you understand by epsilon (ε) - closure of state of finite automata? Convert the given NFA with epsilon(ε)- moves to NFA without epsilon(ε)-move and find the following 7



   i). Number of Final states and identify their names.

   ii). Number of Initial States and identify their names.

   (b). How many minimum number of states are required to construct the deterministic finite automata for the following: 7

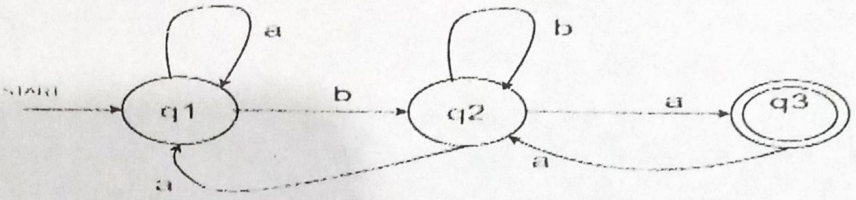
   i) Any language start from some symbols and end by 'n' specific symbols.

   ii) Any language start from 'n' specific symbols and end by some symbols.

P.T.O.

iii) Construct a deterministic finite automata with minimum number states which accept strings of the form $(a+b)^*abba$.

4. (a). Explain and prove Arden's theorem for regular Expression. Find regular expression for the following DFA/NFA using Arden's Theorem. 9



(b). Give a regular expression for representing the set L of string in which every '0' is immediately followed by at least two 1's and prove that the it is also equivalent to the regular expression $\lambda + 1^*(011)^*(1^*(011)^*)^*$ 5

5. What do you understand by simplification of context free grammar? Simplify the given Context free grammar and then convert the simplified grammar in to Chomsky Normal Form (CNF). 14

$S \rightarrow aAa$

$A \rightarrow Sb \mid bCC \mid DaA$

$C \rightarrow abb \mid DdD$

$E \rightarrow aC$

$D \rightarrow aDA$

6. What do you understand by Greibach Normal Form (GNF)? Convert the given context free grammar in to Greibach Normal Form (GNF) and then construct a PDA from the converted GNF grammar. 14

$S \rightarrow AA \mid a$

$A \rightarrow SS \mid b$

7. (a). State Post Correspondence Problem (PCP)? Justify whether the post correspondence problem decidable or undecidable by taking appropriate example. 5

(b). Construct a PDA for language $L = \{WcW^R \mid W \in \{a, b\}^*\}$ and write all Instantaneous Descriptions and make a transition diagram of PDA. 9

8. Differentiate between recursive language and recursive enumerable language. Design a Turing Machine for the following Language 14

$L = \{a^n b^n c^n d^n \mid n \geq 0\}$ and check whether the string 'aabbccdd' is accepted or not by the designed Turing machine (show each moves).

****