SOMAIYA VIDYAVIHAR UNIVERSITY

26.05.2024 (M)

Semester: January 2024 – April 2024 Maximum Marks: 100 Examination: ESE Examination		Duration:3 Hrs.
Programme code: 54 Programme: Honors in Data Science and Analytics	Class: SY	Semester: IV (SVU 2020)
Name of the Constituent College: K. J. Somaiya College of Engineering	Name of the department: COMPUTER	
Course Code: 116h54C401	Name of the Course: Applied Data Science	
Instructions: 1)Draw neat diagrams 2) All questions are compulsory 3) Assume suitable data wherever necessary		

Que. No.	Question	Max. Marks
Q1	Solve any **Four**	20
i)	Explain the ETL process.	5
ii)	Explain datafication with the help of any five suitable examples	5
iii)	Explain word analysis for text processing with examples.	5
iv)	Explain the 5 V's of big data with the help of suitable examples.	5
v)	With the help of an example explain stratified random sampling.	5
vi)	Differentiate between classification and clustering in data analysis. Discuss the strengths and weaknesses of each technique, and provide real-world examples of where you might use each approach.	5

Que. No.	Question	Max. Marks
Q2 A	Solve the following	10
i)	Suppose that survival drops off rapidly in the year following diagnosis of a certain type of advanced cancer. Suppose that the length of survival (or time-to-death) is a random variable that approximately follows an exponential distribution with parameter 2. What's the probability that a person who is diagnosed with this illness survives a year?	5
ii)	The number of ships to arrive at a harbor on any given day is a random variable represented by x. The probability distribution for x is: **x**: 10, 11, 12, 13, 14 **P(x)**: 0.4, 0.2, 0.2, 0.1, 0.1 a. Find the probability that at most 11 ships arrive. b. Find the variance and standard deviation for the number of ships to arrive at the harbor.	5
	OR	

Q2 What is cross validation ? (2 marks)

A Consider the confusion matrix given below: 10

		Predicted	
		0	1
Actual	0	30	12
	1	8	56

Using the confusion matrix, calculate the following measures and interpret them: (2 marks each)

1. Accuracy
2. Precision
3. Recall
4. F1 score

Q 2 B Solve any **One** 10

i) Consider a dataset with two features X1 and X2 and 15 data points: 10

ID	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
X1	2	2	11	6	6	1	5	4	10	7	9	4	3	3	6
X2	10	6	11	9	4	2	10	9	12	5	11	6	10	8	11

Apply k-means algorithm to this dataset. Assume k=3 and data points with ID 2, 7 and 15 as the initial cluster centroids.

ii) Build the decision tree using the CART algorithm with Gini Index on the following dataset: 10

City Size	Avg. Income	Local Investors	LOHAS Awareness	Decision
Big	High	Yes	High	Yes
Medium	Med	No	Med	No
Small	Low	Yes	Low	No
Big	High	No	High	Yes
Small	Med	Yes	High	No
Med	High	Yes	Med	Yes
Med	Med	Yes	Med	No
Big	Med	No	Med	No
Med	High	Yes	Low	No
Small	High	No	High	Yes
Small	Med	No	High	No
Med	Heigh	No	Med	No

Que. No.	Question	Max. Marks
Q3	Solve any **Two**	20
i)	What is ensemble learning? Explain the random forest algorithm with an example.	10
ii)	Explain Hidden Markov Model (HMM) with the help of a detailed example.	10
iii)	With the help of an example illustrate the process of inference in a Bayesian Belief Network.	10

Que. No.	Question	Max. Marks
Q4	Solve any **Two**	20
i)	Consider the following set of observations:	10

chills	runny nose	headache	fever	flu?
Y	N	Mild	Y	N
Y	Y	No	N	Y
Y	N	Strong	Y	Y
N	Y	Mild	Y	Y
N	N	No	N	N
N	Y	Strong	Y	Y
N	Y	Strong	N	N
Y	Y	Mild	Y	Y

Build a Naive Bayes classifier model with conditional probability tables.

Given a new patient with symptoms: Runny nose=no, Headache=no, Chills=true, Fever=true, use the model to predict whether the patient has flu.

ii)	There are five data points P1,P2, … , P5. Use the distance matrix in the following table to perform single link and complete link hierarchical clustering. Show your results by drawing a dendrogram. The dendrogram should clearly show the order in which the points are merged.	10

	P1	P2	P3	P4	P5
P1	0.00	0.10	0.41	0.55	0.35
P2	0.10	0.00	0.64	0.47	0.98
P3	0.41	0.64	0.00	0.44	0.85
P4	0.55	0.47	0.44	0.00	0.76
P5	0.35	0.98	0.85	0.76	0.00

P.T.O.

Que. No.	Question	Max. Marks
Q4 iii)	Consider the following dataset:	10

Height (in cms)	Weight (in kgs)	T Shirt Size
158	58	M
158	59	M
158	63	M
160	59	M
160	60	M
163	60	M
163	61	M
165	65	L
168	62	L
168	63	L
168	66	L
170	63	L
170	64	L
170	68	L

Apply k-NN algorithm to predict the T-shirt size that would fit a person with a height of 164 cm and a weight of 67 kg.

Que. No.	Question	Max. Marks
Q5	Write short notes on any **four**	**20**
i)	Logistic Regression	5
ii)	Steps in data preprocessing.	5
iii)	Gradient descent algorithm	5
iv)	SVM	5
v)	Curse of dimensionality	5
vi)	Impact of data science on business	5