2.6.2023 (E)

SOMAIYA VIDYAVIHAR UNIVERSITY

Semester: January 2023 – May 2023		
Maximum Marks:100	**Examination: ESE Examination**	**Duration:3 hrs**
Programme code: 54 **Programme: Computer Engineering** **(Honours 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)Assume suitable data if necessary		

Que. No.	Question	Max. Marks
Q1	Solve any **Four**	**20**
i)	Explain Applied Data Science challenges	5
ii)	What is Skewness w.r.t data explain with diagram	5
iii)	What is supervised and unsupervised methods of data modelling	5
iv)	Explain correlation and types of correlation with diagram	5
v)	Discuss with an example any 2 strategies of data transformation	5
vi)	Explain Statistical data modelling with example	5

Que. No.	Question	Max. Marks
Q2 A	Solve the following	**10**
i)	Explain normal distribution of data with bell curve	5
ii)	What are the Impact of applying Data Science in business -Online Railways Ticket booking	5
	OR	
Q2 A	Explain characteristics of Big data w.r.t Whats App Application	10
Q 2 B	Solve any **One**	**10**
i)	a.In the following table, the third column is the predicted probability(posterior) for the positive class in a binary classification problem. Assume that any test instances whose posterior probability is greater than threshold=0.5 will be classified as positive example. Compute the precision, Recall	10

Instances	True class	P(+/A,.....,Z,M1)
1	+	0.73
2	+	0.69
3	-	0.44
4	-	0.55
5	+	0.67
6	+	0.47
7	-	0.08
8	-	0.15
9	+	0.45
10	-	0.35

b. Give the different measures of central tendency for numerical data. Which is the robust measure? Why?

ii)	The following table shows the data collected by a state highway patrol safety division on stopping distances. **Speed(mph) / Stopping Distance (ft)** 12 / 17.5 15 / 28 20 / 41.5 25 / 56 32 / 77.5 37 / 104 42 / 122 47 / 158 52 / 177.5 32 / 80 Predict the stopping distance for a vehicle travelling at 50mph using regression	10

Que. No.	Question	Max. Marks
Q3	Solve any **Two**	**20**
i)	For which type of data analysis following graphs are use - Histogram - Scatter plot - Box plot Explain with proper example	10
ii)	Why Pre-processing of data is required ? Explain any 3 pre-processing method with example	10
iii)	Explain The Data Science Process on Hospital Management data	10

Que. No.	Question	Max. Marks
Q4	Solve any **Two**	**20**
i)	a. Sales price records has been sorted as follows: (5,10,11,13,15,35,50,55,72,92,204,215) Partition the data into 3 clusters using using k-means (atleast 3 iteration) b. Compare k-Nearest Neighbours(k-NN) and k-means algorithms	10

ii)	Use given table and Bayes theorem to find : If the weather is sunny and wind is weak then the Player should play or not?	10

Sr No.	Wind	Weather	Play
0	weak	Rainy	Yes
1	Weak	Sunny	Yes
2	Strong	Overcast	Yes
3	Weak	Overcast	Yes
4	Strong	Sunny	No
5	Weak	Rainy	Yes
6	Weak	Sunny	Yes
7	Weak	Overcast	Yes
8	Weak	Rainy	No
9	Strong	Sunny	No
10	Weak	Sunny	Yes
11	Strong	Rainy	No
12	Weak	Overcast	Yes
13	Strong	Overcast	Yes

iii)	Using the following data set , predict whether a young person with low income, has own house or rented house by building decision tree.	10

No	Income	Age	Own house
1	Very high	Young	Yes
2	High	Middle	Yes
3	Low	Young	Rented
4	High	Middle	Yes
5	Very high	Middle	Yes
6	Medium	Young	Yes
7	High	Old	Yes
8	Medium	Middle	Rented
9	Low	Middle	Rented
10	Low	Old	Rented
11	High	Young	Yes
12	Medium	Old	Rented
13	Medium	Young	Rented

Que. No.	Question	Max. Marks
Q5	(Write notes / Short question type) on any **four**	**20**
i)	What Skill sets are needed for Applied Data Science	5
ii)	Feature Selection algorithm -Filters	5
iii)	Is Random Forest better compared to Decision tree? Justify your answer	5
iv)	Kurtios	5
v)	Draw confusion Matrix and explain	5