30/5/2022 (E)

SOMAIYA VIDYAVIHAR UNIVERSITY

Semester: January 2022 – May 2022		
Maximum Marks: 100	**Examination: ESE Examination**	**Duration: 03 hrs**
Programme code: 54 **Programme: 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:** COMP	
Course Code: 116h54C401	**Name of the Course: Applied Data Science**	
Instructions: 1)Draw neat diagrams 2)Assume suitable data if necessary		

Question No.		Max. Marks
Q1 (a)	**Attempt any TWO (2) of the following** 1. Draw and explain the data science Process. 2. Explain Chi-squared distribution with suitable example. 3. Explain Significance of variance and bias with respect to machine learning.	10 Marks (05 Marks each)
Q1 (b)	1. Explain K-Fold Cross Validation with suitable diagram. 2. Explain Poisson distribution with example.	10 Marks
Q2 (a)	**Attempt any TWO (2) of the following** 1. Compare Python and R programming. List of packages of R useful for data science. 2. You're appointed as data scientist for second hand material selling company. Company having 1M customer base throughout Maharashtra and Gujrat state. You're asked to provide the data science based solution to increase the revenue. 3. What is X2 (chi-square) test? Perform test on given data and give inference over a result.	10 Marks (05 Marks each)
Q2 (b)	**Attempt any TWO (2) of the following** 1. Explain skewness and kurtosis with a diagram? 2. What is regression? Explain the usability of regression in data science. 3. Explain the use of Brownian motions in the finance model.	10 Marks (05 Marks each)
Q3 (a)	**Attempt any TWO (2) of the following** 1. Why data science is iterative process? Justify your answer with suitable example. 2. Explain how to treat missing values during data cleaning process. 3. What is outlier and error? Explain the treatment to remove it from dataset.	10 Marks (05 Marks each)

	Play chess	Not play chess	Sum (row)
Like science fiction	250(90)	200(360)	450
Not like science fiction	50(210)	1000(840)	1050
Sum(col.)	300	1200	1500

<table>
<tr><td>Q3 (b)</td><td>What is sampling? What is need of sampling? List types of sampling. Explain any of the sampling with suitable example.

Or

What is Data Normalization? Explain need of Data Normalization? Explain Min-Max normalization with suitable example.</td><td>10 Marks</td></tr>
<tr><td>Q4 (a)</td><td>What is K-means clustering? Give Algorithm. Calculate mean and final cluster for Given data: {2, 3, 4, 10, 8, 12, 3, 20, 30, 11, 27, 35, 28, 19}, K=2</td><td>10 Marks</td></tr>
<tr><td>Q4 (b)</td><td>What is house price prediction? Which algorithm is applicable for house price prediction? Explain algorithm in detail.

Or

"Linear regression is not suitable for classification problem", Justify with suitable example. Explain logistic regression</td><td>10 Marks</td></tr>
<tr><td>Q5 (a)</td><td>Consider the problem of comet detection. Suppose we have a dataset of detected and not detected comet and we train a naïve Bayes classifier on the dataset. For ten instances, the figure below shows the predictions of the trained classifier of the probability of a comet being detected. The classifier classifies an instance as comet if and only if the predicted probability is greater than 0.700. Draw the confusion matrix and calculate the accuracy, recall, F-measure (F1), Precision, Error rate, Sensitivity, Specificity.
<table>
<tr><th>Predicted Probability</th><th>Actual label</th></tr>
<tr><td>0.012</td><td>Not Detected</td></tr>
<tr><td>0.201</td><td>Not Detected</td></tr>
<tr><td>0.321</td><td>Not Detected</td></tr>
<tr><td>0.432</td><td>Not Detected</td></tr>
<tr><td>0.699</td><td>Not Detected</td></tr>
<tr><td>0.721</td><td>Detected</td></tr>
<tr><td>0.734</td><td>Detected</td></tr>
<tr><td>0.801</td><td>Detected</td></tr>
<tr><td>0.907</td><td>Detected</td></tr>
<tr><td>0.701</td><td>Detected</td></tr>
</table></td><td>10 Marks</td></tr>
<tr><td>Q5 (b)</td><td>Attempt any TWO (2) of the following
1. Draw and Explain different phases of NLP.
2. What is Word Embedding? What are their types? Explain.
3. Give importance of Feature selection with example.</td><td>10 Marks (05 Marks each)</td></tr>
</table>