P P SAVANI UNIVERSITY

Second Semester of M.Sc. CS Examination

September 2022

SSDM7061 Introduction to Data Science

16.08.2022, Tuesday **Time: 09:00 a.m. To 11:30 a.m.** **Maximum Marks: 60**

Instructions:

1. The question paper comprises of two sections.
2. Section I and II must be attempted in separate answer sheets.
3. Make suitable assumptions and draw neat figures wherever required.
4. Use of scientific calculator is allowed.

SECTION – I

Q - 1	Answer the Following:	**[05]**
(i)	Suppose that the minimum and maximum values for the attribute income are \$12000 and \$98000, respectively. We would like to map income to the range [0.0, 1.0]. By min-max normalization, a value \$73,600 for income is transformed to _____	
(ii)	ANOVA stands for _____	
(iii)	PCA stands for_____	
Q - 2 (a)	Explain Data Preprocessing Techniques.	**[05]**
Q - 2 (b)	Explain Data Transformation Techniques.	**[05]**
	OR	
Q - 2 (a)	Explain Box Plot with example.	**[05]**
Q - 2 (b)	Explain Data Reduction and types of data reduction techniques.	**[05]**
Q - 3 (a)	Smooth the give data [8 16, 9, 15, 21, 21, 24, 30, 26, 27, 30, 34]. Smooth the data by equal frequency, smoothing by bean means, Smoothing by bin boundaries.	**[05]**
Q - 3 (b)	Explain Data Cleaning Process.	**[05]**
Q - 4	Attempt any one.	**[05]**
(i)	Explain challenges of Data Science.	
(ii)	What is normalization? Explain normalization techniques.	

SECTION – II

Q - 1	Answer the Following: (Any Five)	**[05]**
(i)	How do you find RMSE and MSE in a linear regression model?	
(ii)	What is the significance of p-value?	
(iii)	Write the equation to calculate the precision and recall rate.	
(iv)	We want to predict the probability of death from heart disease based on three risk factors: age, gender, and blood cholesterol level. What is the most appropriate algorithm for this case?	
(v)	What do you understand about true positive rate and false-positive rate?	
(vi)	Which of the following is a categorical outcome? a) RMSE b) RSquared c) Accuracy d) All of the mentioned	
Q - 2 (a)	Explain clustering Technique and its three application in real life.	**[05]**
Q - 2 (b)	Explain Regression and types of regression Technique.	**[05]**
	OR	
Q - 2 (a)	Briefly explain the various model involved in Regression Analysis.	**[05]**
Q - 2 (b)	Explain the procedure of doing prediction analysis using Ridge Regression Model.	**[05]**
Q - 3 (a)	What is Residual Plot? Explain Residual plot with suitable example.	**[05]**
Q - 3 (b)	Define a search space as a grid of hyper parameter values and evaluate every position in the grid.	**[05]**

OR

Q - 3 (a)	Differentiate between Overfitting and Under fitting with suitable example.	**[05]**
Q - 3 (b)	Explain K means clustering algorithm.	**[05]**
Q - 4	Attempt anyone.	**[05]**
(i)	Difference between supervised learning and unsupervised learning.	
(ii)	Explain difference between classification and clustering.	

****