Paper / Subject Code: 37471 / Data Analytics and Visualization

1T01876 - T.E. Computer Science and Engineering (Artificial Intelligence and Machine Learning) (Choice Based)
(R-19-20 'C' Scheme)SEMESTER - VI / 37471 - Data Analytics and Visualization
QP CODE: 10029185 DATE: 08/05/2023

Duration: 3 Hrs **[Max Marks: 80]**

Notes: (1) Question No. 1 is Compulsory.
(2) Attempt any **THREE** questions out of the remaining **FIVE.**
(3) All questions carry equal marks.
(4) Assume suitable data, if required, and state it clearly.
(5) Figures to the right indicate full marks.

Q1	a)	What is an analytic sandbox, and why is it important?	5
	b)	Why use autocorrelation instead of autocovariance when examining stationary time series?	5
	c)	Difference between Pandas and NumPy.	5
	d)	What is regression? What is simple linear regression?	5
Q2	a)	Explain in detail how dirty data can be detected in the data exploration phase with visualizations.	10
	b)	List and explain methods that can be used for sentiment analysis.	10
Q3	a)	List and explain the main phases of the Data Analytics Lifecycle.	10
	b)	Describe how logistic regression can be used as a classifier.	10
Q4	a)	Suppose everyone who visits a retail website gets one promotional offer or no promotion at all. We want to see if making a promotional offer makes a difference. What statistical method would you recommend for this analysis?	10
	b)	List and explain the steps in the Text Analysis.	10
Q5	a)	How does the ARMA model differ from the ARIMA model? In what situation is the ARMA model appropriate?	10
	b)	Explain with suitable example how the Term Frequency and Inverse Document Frequency are used in information retrieval.	10
Q6		**Write short notes on:**	
	a)	Evaluating the Residuals in Linear regression.	5
	b)	Box-Jenkins Methodology	5
	c)	Seaborn Library.	5
	d)	Data import and Export in R	5

****