STELLA MARIS COLLEGE (AUTONOMOUS) CHENNAI 600 086
(For candidates admitted during the academic year 2023– 2024 and thereafter)

M.Sc. DEGREE EXAMINATION, NOVEMBER 2024
INFORMATION TECHNOLOGY
THIRD SEMESTER

COURSE : MAJOR CORE
PAPER : DATA ANALYTICS
SUBJECT CODE : 23CS/PC/DA34
TIME : 1 ½ HOURS **MAX. MARKS: 50**

	SECTION – A **Answer all (Internal Choice – Each question carries five marks)** **(6 x 5 = 30)**		
Q.No		**CO**	**CL**
1.	What is a time series data? Provide two examples of time series data and explain. **(OR)** What are the different types of string data and state their uses?	CO1	K1
2.	Explain why handling missing data is important and describe the methods used to handle missing data **(OR)** Explain the term split-apply-combine in group operations with an appropriate example.	CO2	K2
3.	Assume you have a small comma separated text file ex1.csv in examples folder as below. *a,b,c,d,message* *1,2,3,4,hello* *5,6,7,8,world* *9,10,11,12,foo* Explain the different functions that can be used to read the above given csv file into a DataFrame. Print the DataFrame. **(OR)** Determine whether you should use a Pandas Series or a DataFrame for the below given scenarios. Explain your choice and create the same using pandas. **Scenario I** **Scenario II** 0 4 1 7 2 -5 3 3 dtype: int64 pop state year 0 1.5 Ohio 2000 1 1.7 Ohio 2001 2 3.6 Ohio 2002 3 2.4 Nevada 2001 4 2.9 Nevada 2002 5 3.2 Nevada 2003	CO3	K3

…..2

4	Construct a confusion matrix for an imaginary example and calculate the accuracy based on the confusion matrix. **(OR)** With an example, explain Generalization, Overfitting and Underfitting.	CO3	K3
5.	Distinguish Ticks, Tick Labels and Legends with example code. **(OR)** Analyse in detail how a 5-fold cross validation can be used in any given dataset to evaluate the model with an example. Compare the benefits of 5-fold cross validation over a single split of training and testing set.	CO4	K4
6.	Compare and contrast supervised and unsupervised machine learning techniques with examples and list their strengths and limitations. **(OR)** Analyze the decision tree algorithm's process and pre-pruning with an example.	CO4	K4
	SECTION- B **Answer all (Internal Choice – Each question carries ten marks)** **(2 x 10 = 20)**		
Q.No		**CO**	**CL**
7.	Choose the appropriate python built-in string methods to clean a collection of email addresses separated by semicolon(;) where some addresses are invalid due to extra spaces or missing @ symbol or having more than one @ symbol or not ending with .com and explain the chosen string methods. Select other three built-in string methods that are not used for the above cleaning process and explain them with examples. **(OR)** Identify the different types of analytics that can be used on a dataset of a company's sales records that includes columns for OrderID, Product, Quantity, Price, and Date and explain them in detail.	CO3	K3
8	Compare line plot, bar plot, histogram and scatter plot in detail. **(OR)** Distinguish between k-Means Clustering, Agglomerative Clustering and DBSCAN clustering in detail.	CO4	K4

****