CD363IA

USN □□□□□□□□□□

R. V. COLLEGE OF ENGINEERING

Autonomous Institution affiliated to VTU

VI Semester B. E. 2025 Examinations

DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING (DATA SCIENCE)

DATA ANALYTICS & VISUALIZATION

(2022 SCHEME)

Model Paper

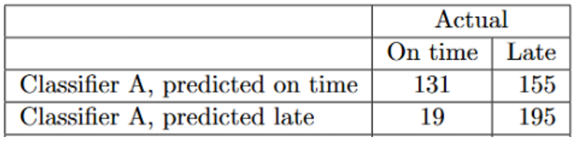
Time: 03 Hours ***Maximum Marks: 100***

Instructions to candidates:

1. Answer all questions from Part A. Part A questions should be answered in first three pages of the answer book only.
2. Answer FIVE full questions from Part B. In Part B question number 2 is compulsory. Answer any one full question from 3 and 4, 5 and 6, 7 and 8 & 9 and 10.

Q. No		**PART-A**	Marks
1	1.1	Define the term Data Wrangling. Give an example.	2
	1.2	Identify the type of data which is found in social media posts, images, or videos.	1
	1.3	Write the steps for Interactive Visualization.	2
	1.4	Compare the features and use cases of Python and R as data science toolkits.	2
	1.5	What is the main goal of data cleaning in a data science project?	1
	1.6	The process of using a large group of people, typically from an online community, to contribute data or solve problems is called ___________.	1
	1.7	List any four general methods used to handle missing data values.	2
	1.8	Define Statistical Data Analysis.	1
	1.9	___ plots are ideal for showing the relationship between two numeric variables.	1
	1.10	Define “Chart junk” in data visualization.	1
	1.11	Identify the type of chart used to represent geographic or spatial data.	1
	1.12	What is Occam’s Razor principle in Modeling?	1

	1.13	What is a Simulation Model used for in data science?	2
	1.14	List the properties of a good Evaluation System.	2
PART-B			
2	a	Illustrate with a real-world example the complete data science workflow with the aid of neat block diagram.	08
	b	Differentiate between structured, semi-structured, and unstructured data with suitable examples.	04
	c	Describe any four conventional methods of data visualization and their appropriate use cases.	04
3	a	Illustrate the concepts of hunting, scraping, and logging in data collection. How do these methods differ in terms of their approach and applications?	10
	b	Discuss how gamification can enhance the effectiveness of crowdsourced data collection.	06
		OR	
4	a	Illustrate how to clean a dataset by handling incompatible data formats between different sources with suitable real-world example.	10
	b	Explain the difference between data errors and artifacts during the data cleaning process.	06
5	a	Illustrate the different types of big data analytics and give real-world examples of each.	10
	b	Describe frequency distribution and its significance in understanding datasets.	06
		OR	
6	a	Discuss the role of Exploratory Data Analysis (EDA) and what visualization brings to the table as a part of the process in the data science workflow. Also, with suitable examples illustrate the basic steps that is encouraged to get acquainted with any new data set or confronting a new data set.	08
	b	Compare and contrast Data Analysis and Data Analytics.	04
	c.	Describe the different types of Statistical data used in analysis with suitable real-world examples.	04
7	a	Discuss the principles underlying design and developing a Visualization Aesthetic.	08
	b	Analyze the strengths and weaknesses of using bar plots and pie charts for data presentation. Considering appropriate real-world examples, provide supporting visuals and reasoning.	08

<table>
<tr><td></td><td></td><td><b>OR</b></td><td></td></tr>
<tr><td>8</td><td>a</td><td>With suitable use cases and mathematical formulations discuss Maximizing Data-Ink Ratio and Minimizing the Lie factor in Visualization. Also support your answer with relevant visualization graphs.</td><td>08</td></tr>
<tr><td></td><td>b</td><td>Illustrate the characteristics of Great Visualizations considering the two case studies i) Marey’s Train Schedule ii) Snow’s Cholera Map</td><td>08</td></tr>
<tr><td></td><td></td><td></td><td></td></tr>
<tr><td>9</td><td>a</td><td>Write the complete Taxonomy of Models. Differentiate between deterministic and stochastic models with examples from real-world data science applications</td><td>08</td></tr>
<tr><td></td><td>b</td><td>Indian Airlines has developed a classifier for the prediction whether a flight originating from Delhi will arrive at its destination on time or not. True or Positive here is ‘On time’ and it refers to the case when the flight is no more than 5 minutes late than the scheduled time. The classifiers were tested on a data-set of 500 flights, and the results are as follows:

<table>
<tr><td></td><td colspan="2">Actual</td></tr>
<tr><td></td><td>On time</td><td>Late</td></tr>
<tr><td>Classifier A, predicted on time</td><td>131</td><td>155</td></tr>
<tr><td>Classifier A, predicted late</td><td>19</td><td>195</td></tr>
</table>
Determine the Evaluation Statistic measures Accuracy, Precision, Recall and F-Score.</td><td>08</td></tr>
<tr><td colspan="4"><b>OR</b></td></tr>
<tr><td>10</td><td>a</td><td>Illustrate how Baseline models for Classification differ from those used in Value Prediction, providing suitable scenarios for each.</td><td>08</td></tr>
<tr><td></td><td>b</td><td>A model predicts a numeric variable with actual values [100, 200, 300] and predicted values [110, 190, 310]. Calculate the Mean Squared Error (MSE) and Root Mean Squared Error (RMSD)</td><td>04</td></tr>
<tr><td></td><td>c.</td><td>If a classifier has an ROC curve with an AUC of 0.95. Interpret this result and its significance.</td><td>04</td></tr>
</table>