Total No. of Questions: 6 *Total No. of Printed Pages:3*

Enrollment No......................................

Faculty of Engineering

End Sem (Even) Examination May-2022

CS3ED06 / IT3ED07 Data Science

Programme: B.Tech. Branch/Specialisation: CS/IT

Duration: 3 Hrs. **Maximum Marks: 60**

Note: All questions are compulsory. Internal choices, if any, are indicated. Answers of Q.1 (MCQs) should be written in full instead of only a, b, c or d.

Q.1 i. What is the primary difference between a data scientist and a data engineer? **1**

(a) A data engineer collects data, while a data scientist prepares it for analysis.

(b) A data engineer analyses data, while a data scientist prepares it for analysis.

(c) A data engineer prepares data for analysis, while a data scientist does the analysis.

(d) None of these

ii. Which of the following best describes the principal goal of data science? **1**

(a) To collect and archive exhaustive data sets from various source systems for corporate record keeping uses.

(b) To mine and analyse large amounts of data in order to uncover information that can be leveraged for operational improvements and business gains.

(c) To prepare data for analysts to use as part of analytics applications.

(d) None of these

iii. Which of the following is true for a normal distribution? **1**

(a) Mean and Median equal

(b) Mean and Mode equal

(c) Mean, Median and Mode equal

(d) Mode and Median equal

iv. The Pearson correlation coefficient value ranges from **1**

(a) 0 to 1 (b) -1 to 1 (c) -1 to 0 (d) None of these

P.T.O.

v. What is the Exploratory Data Analysis technique? **1**
(a) Analysis of data using quantitative techniques.
(b) Analysis of data using graphical techniques.
(c) Both (a) and (b)
(d) Collect data from various data sources

vi. The visual representation of the statistical five number summary of variable (s) is given by- **1**
(a) Pair-Plot (b) Box and Whisker Plot
(c) Scatter Plot (d) Histogram

vii. Which of the following includes data transformation, merging, aggregation, group by operation, and reshaping? **1**
(a) Data acquisition (b) Data visualization
(c) Data wrangling (d) Machine learning

viii. Which of the following is not a step-in data analysis? **1**
(a) EDA (b) Clean data (c) Obtain data (d) None of these

ix. Which of the following is performed using SciPy? **1**
(a) Website (b) Plot data
(c) Scientific calculations (d) System administration

x. Which of the following libraries is used to extract a web page? **1**
(a) Beautiful soup (b) Sci-Kit learn
(c) Matplotlib (d) Pylearn

Q.2 i. What is data science? Explain need for data science. **2**
ii. Write the difference between business intelligence & data science. **3**
iii. Describe the primary components of data science. **5**
OR iv. Explain approach to solve the data science problem. **5**

Q.3 i. What do you understand by the term normal distribution? Write the properties of normal distribution. **2**
ii. Spam Assassin works by having users train the system. It looks for patterns in the words in emails marked as spam by the user. For example, it may have learned that the word “free” appears in 20% of the mails marked as spam. Assuming 0.1% of non-spam mail includes the word “free” and 50% of all mails received by the user are spam, find the probability that a mail is spam if the word “free” appears in it.? **8**

OR iii. In a TV Game show, a contestant selects one of three doors; behind one of the doors there is a prize, and behind the other two there are no prizes. After the contestant selects a door, the game-show host opens one of the remaining doors and reveals that there is no prize behind it. The host then asks the contestant whether they want to SWITCH their choice to the other unopened door, or STICK to their original choice. Is it probabilistically advantageous for the contestant to SWITCH doors, or is the probability of winning the prize the same whether they STICK or SWITCH? **8**

Q.4 i. Explain the role of data cleaning in data analysis. **3**
ii. Explain Exploratory Data analysis (EDA) and types of EDA. What is the purpose of EDA? **7**
OR iii. Explain Scatter plot and box plot with example? **7**

Q.5 i. What is Data Visualization? Write the advantage and benefits of good data visualization? **4**
ii. Describe the basic principles of data visualization. **6**
OR iii. Name and describe at least five tools of data visualization. **6**

Q.6 Attempt any two:
i. What are the important skills to have in Python with regard to data science? **5**
ii. Explain NoSQL Database? Also explain use of python as a data science tool. **5**
iii. Explain the following python libraries: **5**
(a) Sci-kit Learn (b) Matplotlib

******

Marking Scheme
CS3ED06 / IT3ED07 Data Science

Q.1	i.	What is the primary difference between a data scientist and a data engineer?		**1**
		(c) A data engineer prepares data for analysis, while a data scientist does the analysis.		
	ii.	Which of the following best describes the principal goal of data science?		**1**
		(b) To mine and analyse large amounts of data in order to uncover information that can be leveraged for operational improvements and business gains.		
	iii.	Which of the following is true for a normal distribution?		**1**
		(c) Mean, Median and Mode equal		
	iv.	The Pearson correlation coefficient value ranges from		**1**
		(b) -1 to 1		
	v.	What is the Exploratory Data Analysis technique?		**1**
		(c) Both (a) and (b)		
	vi.	The visual representation of the statistical five number summary of variable (s) is given by-		**1**
		(b) Box and Whisker Plot		
	vii.	Which of the following includes data transformation, merging, aggregation, group by operation, and reshaping?		**1**
		(c) Data wrangling		
	viii.	Which of the following is not a step-in data analysis?		**1**
		(d) None of these		
	ix.	Which of the following is performed using SciPy?		**1**
		(c) Scientific calculations		
	x.	Which of the following libraries is used to extract a web page?		**1**
		(a) Beautiful soup		
Q.2	i.	Definition data science	1 mark	**2**
		Need for data science	1 mark	
	ii.	Difference between business intelligence & data science		**3**
		1 mark for each difference	(1 mark * 3)	
	iii.	Primary components of data science		**5**
		1 mark for each component	(1 mark * 5)	
OR	iv.	Approach to solve the data science problem		**5**
		As per the explanation		
Q.3	i.	Definition normal distribution	1 mark	**2**
		Properties of normal distribution	1 mark	
	ii.	Find the probability that a mail is spam if the word “free” appears in it.		**8**
		As per the solution		
OR	iii.	As per the solution		**8**
Q.4	i.	Role of data cleaning in data analysis.		**3**
	ii.	Exploratory Data analysis (EDA)	2 marks	**7**
		Types of EDA	3 marks	
		Purpose of EDA	2 marks	
OR	iii.	Scatter plot	3.5 marks	**7**
		Box plot	3.5 marks	
Q.5	i.	Data Visualization	2 marks	**4**
		Advantage and benefits	2 marks	
	ii.	Describe the basic principles of data visualization.		**6**
OR	iii.	List of data visualization tool	1 mark	**6**
		Description of each tool		
		1 mark for each (1 mark *5)	5 marks	
Q.6		Attempt any two:		
	i.	Important skills to have in Python with regard to data science		**5**
		As per the explanation		
	ii.	NoSQL Database	3 marks	**5**
		Use of python as a data science tool	2 marks	
	iii.	Explain the following python libraries:		**5**
		(a) Sci-kit Learn	2.5 marks	
		(b) Matplotlib	2.5 marks	

******