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

Enrollment No.....

Faculty of Engineering

End Sem (Odd) Examination Dec-2019

OE00055 Data Analytics

Programme: B.Tech. Branch/Specialisation: All

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. Which of the following is not a type of analytics? **1**

(a) Descriptive (b) Deceptive

(c) Prescriptive (d) Predictive

ii. Predictive analytics focuses on: **1**

(a) Report building (b) Model Building

(c) Presentation (d) None of these

iii. Calculate the median from the data set: **1**

55, 19, 24, 86, 36, 41, 66, 11, 15, 32, 71

(a) 41 (b) 36 (c) 24 (d) 32

iv. Calculate the standard deviation for the following data: **1**

55, 19, 24, 86, 36, 41, 66, 11, 15, 32, 71

(a) 23.73 (b) 25.73 (c) 22.73 (d) 21.73

v. Variable cleaning comes under which stage of data analytics: **1**

(a) Data collection (b) Data Preparation

(c) Data Analysis (d) Data Modelling

vi. Data Imputation means: **1**

(a) Deleting data (b) Inserting data

(c) Reducing data (d) None of these

vii. Dimensionality reduction algorithms are one of the possible ways to reduce the computation time required to build a model: **1**

(a) True (b) False

(c) Depends on the scenario (d) Depends on the model

P.T.O.

viii. Which of the following is/are true about PCA? **1**

I. PCA is an unsupervised method

II. It searches for the directions that data have the largest variance

III. Maximum number of principal components <= number of features

IV. All principal components are orthogonal to each other

(a) I and II (b) I and III

(c) I, II and III (d) I, II, III and IV

ix. Decision value to reject null hypothesis in case of a right tail test is said to be **1**

(a) Calculated t must be greater than critical value

(b) Calculated t is less than negative of critical t-value

(c) Calculated t must be less than critical value

(d) Calculated t must be less than critical value in absolute form

x. Test to be applied when number of observations are less than 30 and variance is not known, is said to be **1**

(a) T-test (b) Z-test (c) F-test (d) Chi-square test

Q.2 i. What is the driving force behind data analytics? Justify its need with an example from the current world scenario? **4**

ii. Business intelligence and predictive analytics both have statistics and analytics at their base. State all the differences and similarities between the two? **6**

OR iii. What are basic tasks of data mining? Explain the concept of knowledge discovery in data mining process. **6**

Q.3 i. What is normal distribution? What are the characteristics of normal distribution? **3**

ii. What are the different measures of dispersion? What is the effect of high degree of dispersion on data analysis? **7**

OR iii. What is data visualization? Explain different techniques of data visualization? Why is visualization necessary when summary statistic is already available? **7**

Q.4 i. Describe the possible negative effects of proceeding directly to mine data that has not been processed? **2**

ii. What are the different scaling and normalization methods used for numeric variable transformation? **3**

iii. What is an outlier? Explain the different techniques used to detect an outlier? **5**

OR iv. What is the effect of missing values in a data set? How do we handle missing values? **5**

Q.5 i. How important is transforming the data before applying PCA? What are the issues in assumption about data can be true or can be false? Every such hypothesis needs to data preparation? **4**

ii. Justify the need of PCA in Analytics? Explain the PCA algorithm in detail? **6**

OR iii. Explain the difference between PCA and factor Analysis? What is the drawback of factor analysis? **6**

Q.6 Attempt any two:

i. What is the meaning of margin of error? What are the two ways to reduce margin of error? **5**

ii. An assumption about data can be true or can be false? Every such hypothesis needs to be tested well? Explain the procedure to test a hypothesis for the mean? **5**

iii. Write short notes on: **5**

a) T-test for difference in Mean

b) Z-test for difference in proportion

****

Marking Scheme
OE00055 Data Analytics

Q.1	i.	Which of the following is not a type of analytics? (b) Deceptive		**1**
	ii.	Predictive analytics focuses on: (d) None of these		**1**
	iii.	Calculate the median from the data set: 55, 19, 24, 86, 36, 41, 66, 11, 15, 32, 71 (a) 41		**1**
	iv.	Calculate the standard deviation for the following data: 55, 19, 24, 86, 36, 41, 66, 11, 15, 32, 71 (a) 23.73		**1**
	v.	Variable cleaning comes under which stage of data analytics: (b) Data Preparation		**1**
	vi.	Data Imputation means: (b) Inserting data		**1**
	vii.	Dimensionality reduction algorithms are one of the possible ways to reduce the computation time required to build a model: (a) True		**1**
	viii.	Which of the following is/are true about PCA? (d) I, II, III and IV		**1**
	ix.	Decision value to reject null hypothesis in case of a right tail test is said to be (c) Calculated t must be less than critical value		**1**
	x.	Test to be applied when number of observations are less than 30 and variance is not known, is said to be (b) Z-test		**1**
Q.2	i.	Driving force behind data analytics	1 mark	**4**
		Need	2 marks	
		Example	1 mark	
	ii.	Differences	4 marks	**6**
		Similarities	2 marks	
OR	iii.	Basic tasks of data mining	3 marks	**6**
		Concept of knowledge discovery	3 marks	
Q.3	i.	Definition of normal distribution	1 mark	**3**
		Characteristics of normal distribution	2 marks	
	ii.	Different measures of dispersion	3 marks	**7**
		Effect of high degree of dispersion	4 marks	
OR	iii.	Definition of data visualization	1 mark	**7**
		Different techniques	4 marks	
		Necessity	2 marks	
Q.4	i.	Possible negative effects		**2**
		1 mark for each effect	(1 mark * 2)	
	ii.	At least three transformation methods		**3**
		1 mark for each method	(1 mark * 3)	
	iii.	Definition of outlier	1 marks	**5**
		Different techniques	4 marks	
OR	iv.	Effect of missing values in a data set	2 marks	**5**
		Handling of missing values	3 marks	
Q.5	i.	Importance 2 marks for each	(2 marks * 2)	**4**
	ii.	Need of PCA	2 marks	**6**
		PCA algorithm	4 marks	
OR	iii.	Difference between PCA and factor Analysis	4 marks	**6**
		Drawback of factor analysis	2 marks	
Q.6		Attempt any two:		
	i.	Meaning of margin of error	1 mark	**5**
		Two ways to reduce margin of error 2 marks for each way (2 marks * 2)	4 marks	
	ii.	Procedure to test a hypothesis		**5**
	iii.	Write short notes on:		**5**
		a) T-test for difference in Mean	2.5 marks	
		b) Z-test for difference in proportion	2.5 marks	

****