R6363

Sub. Code
7BD1C1

P.G. DIPLOMA IN BIG DATA ANALYTICS EXAMINATION, NOVEMBER – 2021

First Semester

FUNDAMENTALS OF BIG DATA ANALYTICS

(CBCS – 2018 onwards)

Time : 3 Hours Maximum : 75 Marks

Part A (10 × 2 = 20)

Answer **all** questions.

1. Define data science.
2. List data scientist roles in the data science projects.
3. What is data modeling?
4. What are single variable models?
5. Define factors and its types.
6. What is list? Give an example in R to create a list.
7. Define Map function.
8. Define shuffling.
9. What are the commands for displaying multivariate data?
10. What is the use of R kint package?

Part B (5 × 5 = 25)

Answer **all** questions, choosing either (a) or (b).

11. (a) Write the steps of data science lifecycle.

Or

(b) Write short notes on data exploration.

12. (a) Write about the evaluation strategies of clustering models.

Or

(b) Explain briefly about logistic regression model with commands to build it.

13. (a) Explain arrays with indexing and what is the use of array( ) function.

Or

(b) Explain how to define statistical models in R.

14. (a) Write short notes on distributed file system.

Or

(b) Write how map phase is executed in Hadoop.

15. (a) Write notes on deployment models.

Or

(b) Explain plot ( ) function.

Part C

$(3 \times 10 = 30)$

Answer any **three** questions.

16. Explain cleaning for modeling and validation.

17. Explain k-means algorithm with example.

18. Explain how data is read from files using functions in R.

19. Explain map reduce architecture.

20. Explain how effective presentation can be produced.

———————

R6364	Sub. Code 7BD1C2

P.G. DIPLOMA EXAMINATION, NOVEMBER – 2021

First Semester

Big Data Analytics

ADVANCED COMPUTING FOR BIG DATA ANALYTICS

(CBCS – 2018 onwards)

Time : 3 Hours Maximum : 75 Marks

Part A $(10 \times 2 = 20)$

Answer **all** questions.

1. What is big data?
2. What big data is important?
3. What are the components of hadoop ecosystem?
4. Define data serialization.
5. Define bloom filters.
6. Write any two HDFS sheel commands with its purpose.
7. What are the scheduler options in YARN?
8. How delay scheduling in YARN works?
9. Define Aggregating.
10. What is Pig?

Part B (5 × 5 = 25)

Answer **all** questions, choosing either (a) or (b).

11. (a) What are the characteristics of big data?

Or

(b) Write a note on big data analytics.

12. (a) Discuss on basic of map reduce.

Or

(b) How to move data in to hadoop?

13. (a) What is the usage of HDFS administering?

Or

(b) Write a note on coherency model.

14. (a) What are the main benefits of HDFS federation?

Or

(b) Elaborate lifespan of a YARN application.

15. (a) Discuss on maintaining cluster.

Or

(b) How to install and run pig?

Part C (3 × 10 = 30)

Answer any **three** questions.

16. How big data helping in real time applications? Explain any two applications.

17. Explain the concept of map reduce in detail.

18. Discuss in detail on HDFS architecture.

19. Illustrate how YARN runs an application.

20. Explain different types of Joins in Hive.

R6365

Sub. Code
7BD1G1

P.G. DIPLOMA EXAMINATION, NOVEMBER – 2021

First Semester

Big Data Analytics

MARKETING ANALYSIS

(CBCS – 2018 onwards)

Time : 3 Hours Maximum : 75 Marks

Part A $(10 \times 2 = 20)$

Answer **all** questions.

1. Define Market Analysis.
2. What is cell in Excel?
3. What is Linear Regression?
4. What is Correlation?
5. What is Logistic regression?
6. What are the main components in discrete choice analysis?
7. What is segmentation?
8. What are the different types of collaborative filtering?
9. What is the scope of marketing research?
10. Define Demographics.

Part B (5 × 5 = 25)

Answer **all** questions, choosing either (a) or (b).

11. (a) Explain how to slice marketing data using Pivot table in Excel.

Or

(b) How Excel charts are used to summarize marketing data? Explain.

12. (a) How is simple linear regression calculated? Explain.

Or

(b) Explain Winter's Forecasting technique.

13. (a) Briefly explain Conjoint analysis.

Or

(b) Write a short note on Discrete Choice Analysis.

14. (a) Explain the cluster analysis.

Or

(b) What are the disadvantages of user-based collaborative filtering?

15. (a) Explain the key parameters for study of marketing research in retailing.

Or

(b) Explain the Principal Components analysis.

Part C

$(3 \times 10 = 30)$

Answer any **three** questions.

16. Explain in detail about Excel functions used in marketing data.

17. How to use Neural networks to forecast sales? Explain.

18. Write a detailed note on Customer value.

19. Explain, in detail, collaborative filtering.

20. Discuss Market research tools.

———————

R6366	Sub. Code
	7BD1G2

P.G. DIPLOMA EXAMINATION, NOVEMBER – 2021

First Semester

Big Data Analytics

MATHEMATICAL LOGICS FOR ANALYTICS

(CBCS – 2018 onwards)

Time : 3 Hours Maximum : 75 Marks

Part A (10 × 2 = 20)

Answer **ALL** the questions.

1. Define Data Scientists.

2. List any two roles of analytic project.

3. Write down formula for measures of kurtosis.

4. Define Mode

5. What are Type I and Type II errors?

6. Define F-variance.

7. What is Regression Analysis?

8. Define Multicollinearity.

9. Define Time series.

10. Write the basic design of experiments.

Part B (5 × 5 = 25)

Answer **ALL** the questions, choosing either (a) or (b).

11. (a) What are the practices followed by Data Scientists?

Or

(b) Discuss the main phases of life cycle in Data Analytics.

12. (a) Find the Quartile deviation for the following data:

x	0	1	2	3	4	5	6	7	8
f	1	9	26	59	72	52	29	7	1

Or

(b) Find the Mode for the following data:

x	1	2	3	4	5	6	7	8
f	4	9	16	25	22	15	7	3

13. (a) 15.5 percent of random sample of 1600 undergraduates were smokers, whereas 20% of random sample of 900 post graduates were smokers in a state. Can we conclude that less number of undergraduates are smokers than the postgraduates?

Or

(b) A sample of 100 students is taken from the large population. The mean height of the students in this sample is 160cm. Can it be reasonably regarded that, in the population, the mean height is 165cm and the Standard Deviation is 10cm?

14. (a) The lengths and weights of a sample of six articles manufactured by a factor are given here. Find the Pearson's correlation coefficient.

Length (X)	3	5	6	7	10	11
Weight (Y)	8	12	11	14	16	17

Or

(b) Discuss about curve fitting and goodness of fit in detail.

15. (a) Four doctors each test four treatments for a certain disease and observe the number of days each patient takes to recover. The results are as follows (recover time in days)

Doctor	Treatment			
	1	2	3	4
A	10	14	19	20
B	11	15	17	21
C	9	12	16	19
D	8	13	17	20

Discuss the difference between

(i) doctors

(ii) treatments.

Or

(b) Analyse the variance in the following Latin square of yields (in kgs) of paddy where A,B,C,D denotes the different methods of cultivation.

D 122 A 121 C 123 B 122

B 124 C 123 A 122 D 125

A 120 B 119 D 120 C 121

C 122 D 123 B 121 A 122

Examine whether the different methods of cultivation name given significantly different yields.

Part C (3 × 10 = 30)

Answer any **THREE** questions.

16. Explain the concept of developing core deliverables for stakeholders.

17. Calculate mean, median, mode for the following data.

Marks	20 – 30	30 – 40	40 – 50	50 – 60	60 – 70	70 – 80
Frequency	3	61	132	153	140	51

18. The following table gives the number of aircraft accidents that occurred during the various days of the week. Test whether the accidents are uniformly distributed over the week.

Days	Mon	Tue	Wed	Thu	Fri	Sat	Total
No.of accidents	14	18	12	11	15	14	84

19. Explain the following:

    (a) Multiple Correlation

    (b) Least Square

20. To test the significance of the variation of the retail prices of certain commodities in the four principle status namely A,B,C and D, seven shops where chosen at random in each city and the prices observed were as follows: (prices in paise)

A	82	79	73	69	69	63	61
B	84	82	80	79	76	68	62
C	88	84	80	68	68	66	66
D	79	77	76	74	72	68	64

Do the data indicate that the prices in the four cities are significant different?

R6367	Sub. Code
	7BD1E2

P.G. DIPLOMA EXAMINATION, NOVEMBER 2021.

First Semester

Big Data Analytics

PRINCIPLES OF RDBMS AND NOSQL

(CBCS – 2018 onwards)

Time : 3 Hours Maximum : 75 Marks

Part A $(10 \times 2 = 20)$

Answer **all** questions.

1. What is relational Data Model?

2. What is field?

3. Define SQL * plus.

4. Write SQL quries to select a column from the tables.

5. What is NOSQL? What query language does NOSQL use?

6. List out CAP properties.

7. What is a Document in Mongo DB?

8. How do add data in Mongo DB?

9. How to define arrays in Mongo DB?

10. Write Syntax for Remove ( ).

Part B (5 × 5 = 25)

Answer **all** questions, choosing either (a) or (b).

11. (a) Explain about the commonly used constraints available in SQL.

Or

(b) Consider student table

Reg No.	SNAME	DEPT

perform the following.

(i) Rename the table student as student_info.

(ii) Add a new column semester with note to the existing table student.

(iii) Rename the column SNAME to Student name in the student table.

(iv) Change the datatype of column Reg No. as Char with size 10.

(v) Delete table.

12. (a) What is keys in SQL? Write notes on Primary key, superkey, candidate key and foreign key with example.

Or

(b) Explain about Transactions Management in SQL.

13. (a) Explain in detail about key value pairs based and column based NOSQL databases.

Or

(b) Discuss about uses of NOSQL in Industries.

14. (a) Describe the various datatypes in Mongo DB with example.

Or

(b) Write notes on Sharding.

15. (a) Write short notes on

(i) Find ( ) Field selection

(ii) ordering

(iii) count?

Or

(b) Differentiate Mongo DB (Vs) MySQL.

Part C $(3 \times 10 = 30)$

Answer any **three** questions.

16. State and explain the command DDL, DML DCL with suitable example.

17. What is SQL functions? List out the various Aggregate functions in SQL.

18. (a) Differentiate SQL and NOSQL.

(b) Discuss about the Advantage and uses of NOSQL.

19. Explain in details

(a) Storing binary data in Mongo DB and

(b) Replication in Mongo DB.

20. State and Discuss the Map reduce command in Mongo DB in detail.

———————