VIT
Vellore Institute of Technology

Continuous Assessment Test – I

Programme Name & Branch: PG

Course Name & Code: Français Fonctionnel/ FRE5001

Faculty : Prof.O.Malathy/ Dr.Vandana Sharma

Class Number: 6474, 6472 **Slot: F1**

Exam Duration: 90 mins **Maximum Marks: 50**

---

1. a) Traduisez en anglais : (5)

1. Est – ce que vous aimez voyager ?
2. Le dimanche, ils jouent au football.
3. Est-ce qu'elles parlent l'allemand ?
4. La famille aime visiter la ville.
5. Il n'aime pas regarder les films français.

b) Traduisez en français: (5)

1. Thank you!
2. Sorry!
3. Good evening!
4. All the best !
5. Please !

2. a) Mettez l'article indéfini : (5)

1. Nous écrivons _____ lettre.
2. Nous achetons _____ livre.
3. Est-ce que vous jouez _____ match de football ?
4. Martha prépare _____ gâteau.
5. Les professeurs rencontrent _____ étudiants.

b) Complétez avec le verbe : **(5)**

1. Tu _____ (travailler) dans un restaurant ?
2. Je _____ (détester) les avions et les aéroports. C'est horrible !
3. Est-ce que tu _____ (inviter) les professeurs?
4. Oui, ils _____ (adorer) les fruits.
5. Elles _____ (danser) dans la pluie.

3. **a) Mettez les phrases au négatif :** **(5)**

1. Les enfants montent l'arbre.
2. Le professeur enseigne le français.
3. Nous aimons les fruits.
4. La femme aide les enfants.
5. L'enfant écoute la musique.

b) Choisissez la meilleure réponse: **(5)**

1. Saint-Placide est ------- ville tranquille (un/une).
2. Monsieur Dumont ------- à Newyork. (habite/ habites)
3. Barbara étudie à VIT. Elle est une ----- (étudiant/ étudiante).
4. Ils achètent ---------- pommes (des/un/une).
5. Je --------- un film (portes/ regarde).

4. **Conjuguez les verbes :** **(10)**

1. *à l'affirmatif* : Avoir, être
2. *au négatif* : Marcher, visiter
3. *à l'interrogatif* : travailler

5. **a) Ecrivez les nombres 21-30 en toutes lettres(5)**
b) Ecrivez les jours de la semaine en français(5)

VIT

Vellore Institute of Technology

(Deemed to be University under section 3 of UGC Act, 1956)

School of Advanced Sciences

Continuous Assessment Test -I (CAT-I) Fall Semester (2023-24)

Course Name & code : **Research Methodology RES5001**

Programme Name & Branch : **M.Sc. Data Science**

Class Number : **VL2023240106607**

Slot : **A1**

Faculty Name : **Dr. Yogiraj Mantri**

Exam Duration : **90 minutes** **Maximum Marks** : **50**

General instruction(s): Answer all questions

Q.No.	Question	Max Marks
1.	What are deductive and inductive approaches to research? Briefly explain the difference between the two approaches.	10
2.	Explain in brief few characteristics of good research	10
3.	Explain the different methods and techniques that can be used to collect data for research	10
4.	a) Explain importance of literature survey while formulating a research problem b) How can you develop your research topic by discussing with colleagues, experts? Write some points you would want to discuss while formulating a research problem	5+5
5.	Give reasons to explain if the following research problem is well formulated. 'Through this research we want to study mental health issues in people' Suggest alternate formulations to improve the formulation of the problem.	10

*****

Continuous Assessment Test - I

Program Name & Branch	: Department of Mathematics
Course Code	: MAT5011
Course Name	: Matrix theory and Linear Algebra
Slot	: D1+TD1
Class Number(s)	: VL2023240106587
Faculty Members	: Dr. M S JAGADEESHKUMAR
Date of the Examination	: 13-09-2023
Duration	**: 90 minutes** **Max. Marks : 50**

General instruction(s): *Please answer all the 5 questions bellow.*

Q. No	Question	Marks
1.	Given a system of linear equations with three unknowns x,y,z. Solve the system using Gauss elimination method. $$10x + y + z = 12$$ $$2x + 10y + z = 13$$ $$2x + 2y + 10z = 14$$	10
2.	Solve the given system of equations with three unknowns x, y, z by LU decomposition method. $$x + y + z = 1$$ $$4x + 3y - z = 6$$ $$3x + 5y + 3z = 4$$	10
3.	Find the rank of the following matrix $$A = \begin{bmatrix} 2 & 3 & -1 & -1 \\ 1 & -1 & -2 & -4 \\ 3 & 1 & 3 & 2 \\ 6 & 3 & 0 & 7 \end{bmatrix}$$	10
4.	Find the inverse of the following matrix by Gauss-Jordan method. $$A = \begin{bmatrix} 0 & 1 & 1 & 2 \\ -2 & 0 & 1 & 1 \\ 1 & 1 & 0 & 3 \\ 2 & 1 & 1 & 0 \end{bmatrix}$$	10
5	Find all the Eigen values and Eigenvectors of the matrix. $$\begin{bmatrix} 3 & -1 & 1 \\ -1 & 5 & -1 \\ 1 & -1 & 3 \end{bmatrix}$$	10

VIT
Vellore Institute of Technology

C1+TC1

SCHOOL OF ADVANCED SCIENCES
CONTINUOUS ASSESSMENT TEST – I
FALL SEMESTER 2023-24

Programme Name & Branch: M.Sc. Data Science
Course Code: MAT 5012
Course Name: Probability theory and distributions
Faculty Name(s): Dr. Venkataramana B
Class Number(s): VL2023240106589 **Exam Duration: 90 minutes** **Maximum Marks: 50**

General instruction(s): **Scientific calculators are permitted.** **Answer all the questions (5 × 10 = 50)**

Q1 An instructor gives her class a set of 10 problems with the information that the final exam will consist of a random selection of 5 of them. If a student has figured out how to do 7 of the problems, what is the probability that he or she will answer correctly:

(i) all 5 problems (ii) at least 4 of the problems ?

Q2 Five percent of patients suffering from a certain disease are selected to undergo a new treatment that is believed to increase the recovery rate from 30 percent to 50 percent. A person is randomly selected from these patients after the completion of the treatment and is found to have recovered. What is the probability that the patient received the new treatment?

Q3 The following distribution function of a discrete random variable X :

X	-3	-1	0	1	2	3	5	8
$F(x)$	0.10	0.30	0.45	0.5	0.75	0.9	0.95	1.00

Find (i) The marginal probability function of X, (ii) $P(X \text{ is even})$,

(ii) $P(X = -3 \mid X < 0)$

Q4 The joint probability density function of the two-dimensional random variables (X, Y) given by

$$f(x,y) = \begin{cases} x^2 + \frac{xy}{3} & ; 0 < x < 1,\ 0 < y < 2 \\ 0 & ; \quad elsewhere \end{cases}$$

Find the correlation coefficient between X and Y.

Q5 The velocity of a particle in a gas is a random variable V with probability distribution

$$f_V(v) = \begin{cases} aV^2 e^{-bv} ; & v > 0 \\ 0, & elsewhere \end{cases}$$

where b is a constant that depends on the temperature of the gas and the mass of the particle.

(a) Find the value of the constant a.

(b) The kinetic energy of the particle is $= mV^2/2$. Find the probability distribution of W.

VIT
Vellore Institute of Technology

DEPARTMENT OF MATHEMATICS
SCHOOL OF ADVANCED SCIENCES
Fall Semester 2023-2024
Continuous Assessment Test - I

Programme Name & Branch: M.Sc. - Data Science
Slot: E1
Course Code & Name: MAT6012 - Programming for Data Analysis
Class Number(s): VL2023240106593
Faculty Name: Dr. B.S.R.V. Prasad (13342)
Exam Duration: 90 Minutes **Date of Exam: 14-Sep-2023** **Max. Marks: 50**

General Instructions: *Answer ALL the questions.*

1. Write the symbolic form of the following statement and then analyse the statement using truth tables:
"If you get more repeats than any other player in betting you will lose, or that if you lose you must be rich". **(10 Marks)**

2. (a) Write a short note on Pölya's four steps of problem solving process. **(5 Marks)**
   (b) A positive integer N is a perfect number if that number N is equal to the sum of its positive divisors, excluding the number N itself. Draw a flowchart to verify whether a given number is a perfect number or not. **(5 Marks)**

3. Consider the sentence `S = "Hello here is some text without meaning. This text will show what a printed text will look like at this place."` of 110 characters. Write Python commands to perform the following operations on the above string: **(10 Marks)**
   (i) Count the number of times the words `lo` and `ex` are appeared in the string `S`.
   (ii) Convert the string `S` into title case and store it into `ST`.
   (iii) Replace the occurrence of all lowercase vowels in the string `S` with uppercase vowels and store the resultant string into `S1`.
   (iv) Reverse the case of letters in string `S1` and store into a string `S2`.
   (v) Split the string `S` into `five` equal parts (i.e., each part containing equal number of characters) and store these parts in `S11, S12, S13, S14, S15`.
   (vi) Form a new string named `S3` by reversing the strings `S11, S13, S15` and joining them by the symbol `@`.
   (vii) Form a new string by concatenating the first 10 characters of string `S1`, last 10 characters of string `S15` and characters 2 to 10 from the string `S`.

4. (a) Write a short note on the various basic data types that are available in Python with providing proper examples. **(4 Marks)**
   (b) Write a short note on Set data type in Python listing its advantages, disadvantages and any five major functions that can be performed on sets in Python with proper examples. **(6 Marks)**

5. A number N is said to be an Abundant number if the sum of the divisors of the number is greater than $2N$.

   Write a Python program (using if conditionals/loops) to accept an integer number from the user and check whether that number is an Abundant number. The Python program should first validate the number to be integer only and then proceed to check if it is Abundant or not. If the number is not an integer, it should print an error message: "Not a valid number. Please enter an integer." **(10 Marks)**

VIT®

Vellore Institute of Technology

(Deemed to be University under section 3 of UGC Act, 1956)

SCHOOL OF ADVANCED SCIENCES

Fall Semester 2023-2024

Continuous Assessment Test – I

Programme Name & Branch : **M.Sc. & Data Science (MDT)**

Slot : **B1+TB1**

Course Name & code : **Foundations of Data Science & MAT5010**

Class Number (s) : **VL2023240106585**

Faculty Name : **Dr. Jisha Francis**

Exam Duration : **90 Min.** **Maximum Marks: 50**

General instructions:

- Answer all the questions. Read all the instructions and questions carefully before you begin answering.
- Write your answers neatly and legibly. Illegible handwriting may affect grading.
- Provide clear and concise answers. Stick to the point and avoid unnecessary elaboration.
- Maintain academic integrity. Do not engage in cheating, plagiarism, or any form of dishonesty.
- Wish you the best of luck with your examination!

<table>
<tr><th>Q.No.</th><th>Questions</th><th>Max Marks</th></tr>
<tr><td>1.</td><td>a) Define Big Data Analytics and explain how it differs from traditional data analytics.
b) Provide two examples of industries or applications where Big Data Analytics has had a significant impact.</td><td>10</td></tr>
<tr><td>2.</td><td>a) Describe the different phases of the Data Analytics Lifecycle. Explain the objectives and activities typically associated with the Discovery phase.
b) How does the Discovery phase lay the foundation for the subsequent phases of the life cycle?</td><td>10</td></tr>
<tr><td>3.</td><td>Find the missing frequency in the following distribution if N=100 and median is 32.
<table>
<tr><td>Marks</td><td>0-10</td><td>10-20</td><td>20-30</td><td>30-40</td><td>40-50</td><td>50-60</td></tr>
<tr><td>No.of Students</td><td>10</td><td>?</td><td>25</td><td>30</td><td>?</td><td>10</td></tr>
</table></td><td>10</td></tr>
</table>

<table>
<tr>
<td>4.</td>
<td>a) Discuss the importance of boxplot in EDA.
b) The math test results for a class of 15 students are given below:

91 95 54 69 80 85 88 73 71 70 66 90 86 84 73.

Draw a boxplot for the test results. Analyze the plot and discuss what insights it provides regarding test results, including outliers and central tendencies.</td>
<td>10</td>
</tr>
<tr>
<td>5.</td>
<td>The performance of 2 groups of students on the final math exam is given the following table.

<table>
<tr><td>Group A</td><td>56</td><td>58</td><td>60</td><td>62</td><td>64</td></tr>
<tr><td>Group B</td><td>40</td><td>50</td><td>60</td><td>70</td><td>80</td></tr>
</table>

Compare the performances of each group of students by using the descriptive statistics including mean,variance; coefficient of variance and range.</td>
<td>10</td>
</tr>
</table>

30

VIT
Vellore Institute of Technology

Continuous Assessment Test – II

Programme Name & Branch: PG

Course Name & Code: Français Fonctionnel/ FRE5001

Faculty : Prof.O.Malathy/Dr. Vandana Sharma

Class Number: 6474, 6472 Slot: F1

Exam Duration:90 mins Maximum Marks: 50

---

1. **Traduisez les phrases en anglais :** **(10)**

Je m'appelle Angélica Summer, j'ai 12 ans et je suis canadienne. Mon père, Frank Summer, il est mécanicien. Il adore les voitures anciennes. Ma mère s'appelle Emilie Summer. Elle est infirmière dans un hôpital. Nous avons une jolie maison avec un grand jardin.

Le week-end, nous visitons ma grand-mère. Elle a quatre-vingt-quatre ans et elle habite à Antibes. J'adore ma grand-mère, elle est très gentille.

Lundi, je retourne à l'école. Je suis contente. J'aime beaucoup l'école.

2. **Traduisez en français :** **(10)**

a). A green bag.
b) A happy family.
c) A charming girl.
d) A big hôtel.
e) The ancient house.
f) A red pen.
g)A naughty boy.
h) A big garden.
i)The brown book.
j)The small boy.

3. **Complétez avec une préposition de la liste « dans, avec, à, de, sur, devant, derrière, pour » :** **(10)**

a) Pierre réserve une chambre ----- Patrick à l'hôtel Taj.
b) Les enfants jouent ---- le jardin ---- un ballon.
c) Elle habite ----- Paris.
d) Le livre ---- Marie est ---- la table.
e) Le magasin est ------le musée.
f) Le garage est ----- la maison.

30

VIT
Vellore Institute of Technology

Continuous Assessment Test – II

Programme Name & Branch: PG

Course Name & Code: Français Fonctionnel/ FRE5001

Faculty : Prof.O.Malathy/Dr. Vandana Sharma

Class Number: 6474, 6472 Slot: F1

Exam Duration:90 mins Maximum Marks: 50

---

1. **Traduisez les phrases en anglais :** **(10)**

Je m'appelle Angélica Summer, j'ai 12 ans et je suis canadienne. Mon père, Frank Summer, il est mécanicien. Il adore les voitures anciennes. Ma mère s'appelle Emilie Summer. Elle est infirmière dans un hôpital. Nous avons une jolie maison avec un grand jardin.

Le week-end, nous visitons ma grand-mère. Elle a quatre-vingt-quatre ans et elle habite à Antibes. J'adore ma grand-mère, elle est très gentille.

Lundi, je retourne à l'école. Je suis contente. J'aime beaucoup l'école.

2. **Traduisez en français :** **(10)**

a) A green bag.
b) A happy family.
c) A charming girl.
d) A big hôtel.
e) The ancient house.
f) A red pen.
g)A naughty boy.
h) A big garden.
i)The brown book.
j)The small boy.

3. **Complétez avec une préposition de la liste « dans, avec, à, de, sur, devant, derrière, pour » :** **(10)**

a) Pierre réserve une chambre ----- Patrick à l'hôtel Taj.
b) Les enfants jouent ---- le jardin ---- un ballon.
c) Elle habite ----- Paris.
d) Le livre ---- Marie est ---- la table.
e) Le magasin est ------le musée.
f) Le garage est ----- la maison.

g) Il habite ---- la famille de Ria.
h) Je viens à l'université ---- 8h00.

4. Mettez les mots en ordre : (10)
(Rearrange the words in correct order)

a) je/pas/les/ai/documents/n'
b) ne/parlons/nous/allemand/pas/l'
c) sœur/est/d'/Amélie/journaliste/la
d) jolies/les/sont/filles
e) joue/enfant/mère/avec/la/ l'
f) une/elle/pomme/mange/rouge
g) actrice/habite/Canada/l' /au
h) Etats-Unis/M. Sharma/va/aux
i) bruns/ et/Kishore/ ronds/les /de/chapeaux/sont
j) une/il/ronde/achète/table

5.a Complétez avec « être, avoir, habiter, aimer ou s'appeler » (Fill up with the correct form of the verb) : (5)

Il ------ John et il est Kenyan. Il ------ vingt-sept ans et il --------- faire du sport. Il -------- marié avec une Française. Ils -------- à Nairobi, au Kenya.

5.b Reliez les phrases : (5)

a) C'	- est italien, ton ami Paolo !
b) Il	- est ton ami de Mexico !
c) Tu	- travaillez à Tours ?
d) Elle	- es mon ami.
e) Vous	- habite à Paris.

VIT

Vellore Institute of Technology

(Deemed to be University under section 3 of UGC Act, 1956)

SCHOOL OF ADVANCED SCIENCES

Fall Semester 2023-2024

Continuous Assessment Test – II

Programme Name	: M.Sc.
Branch	: Data Science (MDT)
Slot	: B1 + TB1
Course Name & code	: Foundations of Data Science & MAT5010
Class Number(s)	: VL2023240106585
Faculty Name	: Dr. Jisha Francis
Exam Duration	: 90 Min.
Max. Marks	: 50

General Instructions:

- Answer all the questions.
- Maintain academic integrity.

No.	Questions	Marks	CO
1.	A random sample of 10 hot drinks from Dispenser A had a mean volume of 203 ml and a standard deviation (divisor $(n-1)$) of 3 ml. A random sample of 15 hot drinks from Dispenser B gave corresponding values of 206 ml and 5 ml. The amount dispensed by each machine may be assumed to be normally distributed. Test, at the 5% significance level, the hypothesis that there is no difference in the variability of the volume dispensed by the two machines.	10	CO2
2.	Marketers believe that 72% of adults in India own a cell phone. A cell phone manufacturer believes that number is actually lower. 200 Indian adults are surveyed, of which, 174 report having cell phones. Use a 5% level of significance. State the null and alternative hypothesis, find the critical-value, state your conclusion, and identify the Type I and Type II errors.	10	CO2
3.	A consumer group, concerned about the mean fat content of a certain grade of chicken burger, submits to an independent laboratory a random sample of 12 chicken burgers for analysis. The percentage of fat in each of the chicken burgers is: 21, 18, 19, 16, 18, 24, 22, 19, 24, 14, 18, 15. The manufacturer claims that the mean fat content of this grade of chicken burger is less than 20%. Assuming percentage fat content to be normally distributed, carry out an appropriate hypothesis test in order to advise the consumer group the validity of the manufacturer's claim.	10	CO2

Please Turn Over

No.	Questions	Marks	CO
4.	a). Assume a null hypothesis, H0, that states the percentage of adults with jobs is at least 88%. Identify the Type I and Type II errors from these four statements. i. Not to reject the null hypothesis that the percentage of adults who have jobs is at least 88% when that percentage is actually less than 88%. ii. Not to reject the null hypothesis that the percentage of adults who have jobs is at least 88% when the percentage is actually at least 88%. iii. Reject the null hypothesis that the percentage of adults who have jobs is at least 88% when the percentage is actually at least 88%. iv. Reject the null hypothesis that the percentage of adults who have jobs is at least 88% when that percentage is actually less than 88%. b). State whether each of the following is true or false. Justify your answer. i. If the null hypothesis is rejected based on sample evidence using the test at the significance level $\alpha = 0.1$, the research has absolutely proven that the null hypothesis is false without any doubts. ii. If a null hypothesis is rejected at the 5% significance level, then using the same data, the null hypothesis will be rejected at the 10% significance level. iii. As the sample size increases to infinity, the variance of the sample mean approaches zero. iv. Let $\{X1, X2, ..., Xn\}$ be n observations, each of which is randomly drawn from a distribution with mean μ and variance σ^2. Let $$S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2 \quad \text{and} \quad \bar{X} = \frac{1}{n}\sum_{i=1}^{n}X_i$$ Then, the distribution of a statistic $$\frac{\bar{X} - \mu}{s/\sqrt{n}}$$ is always given by t-distribution with the degree of freedom $n - 1$.	10	CO2
5.	In a study of pollution in a water stream, the concentration of pollution is measured at 5 different locations. The locations are at different distances to the pollution source. In the table below, these distances (in km) and the average pollution are given: Distance to pollution source: 2, 4, 6, 8, 10 Average concentration: 11.5, 10.2, 10.3, 9.68, 9.32 a). Use the method of least squares to fit a simple linear regression line to the accompanying data points. Give the estimates of the regression coefficients. b.) Plot the points and sketch the fitted least-squares line.	10	CO2

C1+TC1

VIT
Vellore Institute of Technology

FALL SEMESTER 2023-24
CONTINUOUS ASSESSMENT TEST – II
Programme Name: M.Sc. Data Science
Course Code: MAT 5012
Course Name: Probability theory and distributions

Class Number(s): VL2023240106589 Exam Duration: 90 minutes Maximum Marks: 50

General instruction: Scientific calculators are permitted. Answer all the questions $(5 \times 10 = 50)$

Q1 The number of automobiles sold weekly at a certain dealership is a random variable with expected value 16 and the variance 9.

(i) Give an upper bound to the probability that next week's sales exceed 25.

(ii) Give a lower bound to the probability that next week's sales are between 10 and 22, inclusively.

Q2 Fit a binomial distribution for the following data and hence find the expected frequencies.

x	0	1	2	3	4	5	6
f	5	18	28	12	7	6	4

Q3 a) In a clinical study, volunteers are tested for a gene that has been found to increase the risk for a disease. The probability that a person carries the gene is 0.1. What is the probability that 4 or more people will have to be tested before 2 with the gene are detected? (5M)

b) If the probability that a certain test yields a positive reaction equals 0.4, what is the probability that fewer than 5 negative reactions occur before the first positive. (5M)

Q4 The line width of for semiconductor manufacturing is assumed to be normally distributed with a mean of 0.5 micrometer and a standard deviation of 0.05 micrometer.

(a) What is the probability that a line width is greater than 0.62 micrometer?

(b) What is the probability that a line width is between 0.47 and 0.63 micrometer?

(c) The line width of 90% of samples is below what value?

Q5 Find the moment generating function of a random variable X whose probability function is given by

$$f(x) = \begin{cases} \lambda e^{-\lambda x} & ; \quad x \geq 0, \ \lambda > 0 \\ 0 & ; \quad elsewhere \end{cases}$$

And hence obtain the Mean and variance of X using the moment generating function.

VIT

Vellore Institute of Technology

Vellore – 632014, Tamil Nadu, India

SCHOOL OF ADVANCED SCIENCES

Department of Mathematics FALL SEMESTER (2023-2024)

Continuous Assessment Test -II

Program Name & Branch	: **M. Sc (Data Science)**
Course Code	: MAT5011
Course Name	: Matrix theory and Linear Algebra
Slot	: D1+TD1
Class Number(s)	: VL2023240106587
Faculty Members	: Dr. M S JAGADEESHKUMAR (11258)
Date of the Examination	:18th, October 2023
Duration	**: 90 minutes** **Max. Marks : 50**

General instruction(s): *Please answer **all the 5 questions** below.*

Q. No	Question	Marks
1.	Show that the matrix $A=\begin{pmatrix} 2 & 1 & -1 \\ 1 & 1 & -2 \\ -1 & -2 & 1 \end{pmatrix}$ is orthogonally diagonalisable matrix. Find The orthogonal matrix P and diagonal matrix D such that $A = PDP^T$	10
2.	Reduce the quadratic form of $2x^2 + 6y^2 + 2z^2 - 8xz$ to the canonical form by orthogonal transformation. Find the rank, index and signature and nature of the quadratic form.	10
3.	Find the Singular Value Decomposition (SVD) of $A=\begin{bmatrix} 3 & 1 & 1 \\ -1 & 3 & 1 \end{bmatrix}$.	10
4.	Let R be the set of all real numbers. Prove that $P_2[x]=\{a+bx+cx^2 / a,b,c \in R\}$ forms a vector space over R with usual addition and scalar multiplication.	10
5.	Let R be the set of all real numbers and $W=\{(x,y,z)/x+y+z=0 \text{ and } x,y,z \in R\}$, $U=\{(x,y,z)/x,y,z \in R \text{ and } z=1-x-y\}$ (a) Does W forms a subspace of R^3 ?, justify. (b)Does U forms a subspace of R^3 ?, justify.	10

VIT
Vellore Institute of Technology

SCHOOL OF ADVANCED SCIENCES

Fall Semester 2023-2024

Continuous Assessment Test – II

Programme Name & Branch	:	**M.Sc. Data Science**
Slot	:	**A1**
Course Name & code	:	**Research Methodology RES5001**
Class Number	:	**VL2023240106607**
Faculty Name	:	**Dr. Yogiraj Mantri**

Exam Duration: 90 Min. **Maximum Marks: 50**

General instruction(s): Answer all questions

Q.No.	Question	Max Marks
1.	Explain with help of suitable example the four parts of research design: sampling design, observational design, statistical design, and operational design.	10
2.	How is research design for exploratory research different from that of diagnostic research? Give example for both the research objectives.	10
3.	An experiment is conducted using different rates (e.g. ½ X, X, 1.5 X, 2 X) of a pesticide to determine its efficacy to control weeds. Explain the following experimental designs for this study: a) Completely randomized design b) Factorial design	10
4.	What are the different reasons for which we do sampling? Are there any disadvantages to sampling?	10
5.	A university wants to collect data by sampling few students to study their performance in exams. Explain the methods of random sampling and stratified sampling for this study.	10

VIT
Vellore Institute of Technology

DEPARTMENT OF MATHEMATICS
SCHOOL OF ADVANCED SCIENCES
Fall Semester 2023-2024
Continuous Assessment Test - II

Programme Name & Branch: M.Sc. - Data Science
Slot: E1
Course Code & Name: MAT6012 - Programming for Data Analysis
Class Number(s): VL2023240106593
Faculty Name: Dr. B.S.R.V. Prasad (13342)
Exam Duration: 90 Minutes **Date of Exam: 19-Oct-2023** **Max. Marks: 50**

Open Notebook Examination
Answer ALL the questions.

1. Write a Python function named `DecomposeCapsSpecial(string)` which accepts a string with characters, numbers and special characters and outputs a string of capital letters, a string of special characters separately, along with their count.

   For example, consider the string "aBcD@1234#Ef&56!" then the output is `Capital Letter String is: BDE` and `Special Character String is: @#&!` and the message that "There are 3 capital letters and 4 special characters in the given string". (*Remark.* Use Dictionaries in the function code.) **(10 Marks)**

2. A Pythagorean prime is a prime of the form $4n + 1$. For example, the primes $5, 13, 17$ are Pythagorean primes as $5 = 4 \times 1 + 1$, $13 = 4 \times 3 + 1$ and $17 = 4 \times 4 + 1$.

   Write a Python function named `PythagoreanPrimes(n)`, which accepts a positive integer and returns either *True* or *False* depending on whether the given number n is a Pythagorean prime or not? Using this function list all the Pythagorean primes between 100 and 200 and find their sum, product. **(12 Marks)**

3. Write a Python program to count the positive and negative float numbers, separately, in a given mixed list using lambda, map and filter. For example if the list is [1, 'abcd', 3.12, 1.2, 4, 'xyz', 5, 'pqr', 7.2, -5, -12.22], then the output should be 3 positive float numbers and 1 negative float number. Later find the sum of these entries and print them to the user. **(8 Marks)**

4. Write a Python generator function named `divisibleBy_3_n_5`, which takes a positive number as argument and returns the numbers can be divisible by 3 and 5 between 0 and n. Using, the above function, write the Python code snippet to obtain sum of all the numbers which are divisible by 3 and 5 between 0 and 1000. **(5 Marks)**

5. Consider the array

$$A = \begin{bmatrix} 40 & 60 & 80 & 100 & 120 & 140 \\ 42 & 62 & 82 & 102 & 122 & 142 \\ 44 & 64 & 84 & 104 & 124 & 144 \\ 46 & 66 & 86 & 106 & 126 & 146 \\ 48 & 68 & 88 & 108 & 128 & 148 \\ 50 & 70 & 90 & 110 & 130 & 150 \end{bmatrix}$$

Write Python commands to achieve the following tasks. **(15 Marks)**

(a) Write the List comprehension technique, to construct the above array as a `NumPy` array and store it in A.

(b) Create the sub array, $B_1 = \begin{bmatrix} 60 & 100 & 140 \\ 64 & 104 & 144 \\ 68 & 108 & 148 \end{bmatrix}$.

(c) Create the sub array, $B_2 = \begin{bmatrix} 64 & 68 \\ 104 & 108 \\ 144 & 148 \end{bmatrix}$.

(d) Create the sub array, $B_3 = [42 \quad 106 \quad 68 \quad 80 \quad 90]$.

(e) Create the sub array, B_4 containing the elements 60, 82, 104, 126, 148.

(f) Create the sub array, B_5 containing the elements 48, 70, 50, 120, 142, 140.

(g) Find the matrix product $B_6 = B_1 B_2$ and then find $B_{6a} = B_6/B_1$.

(h) Reshape the matrix B_6 into $3 \times 1 \times 2$ and store it in B_7.

(i) Are the matrices B_3 and B_4 are broadcastable? Give reason. If yes, then what's the result of $B_3 - B_4$?

(j) Are the matrices B_7 and B_5 are broadcastable?
—If yes, what's size of the output array for the operation B_7/B_5?
— If not, can you use any reshaping techniques so that you can find the value $B7/B5$?

VIT
Vellore Institute of Technology

Final Assessment Test – November/December 2023

Course: **MAT5010 - Foundations of Data Science**

Time: **Three Hours** Max. Marks: **100**

KEEPING MOBILE PHONE/SMART WATCH, EVEN IN 'OFF' POSITION IS TREATED AS EXAM MALPRACTICE

General Instructions: Statistical tables are permitted

Answer any TEN Questions

(10 X 10 = 100 Marks)

1. a) What is big data and business intelligence (BI)? **[4]**

   b) How does the traditional BI environment differ from the big data environment? **[6]**

2. a) Compute the median for the following frequency distribution **[6]**

Scores	20-29	30-39	40-49	50-59	60-69	70-79	80-89
No. of students	4	6	8	12	9	7	4

   b) Draw a histogram for the above data and describe the key characteristics of the histogram, including the shape, central tendency, and spread of average scores of marks. **[4]**

3. Two judges at a cooking competition placed the ten entries for the 'best fruit cakes' competition in the following order:

Entry	A	B	C	D	E	F	G	H	I	J
Judge 1	2	9	1	3	10	4	6	8	5	7
Judge 2	6	9	2	1	8	4	3	10	7	5

   Is there any linear relationship between the rankings produced by the two judges? Justify your answer.

4. A sample of 400 male students is found to have a mean height 67.47 inches. Can it be reasonably regarded as a sample from a large population with mean height 67.39 inches and standard deviation 1.30 inches? Test at 5% level of significance.

5. A restaurant near the railway station has been having an average sales of 500 tea cups per day. Because of the development of a bus stand nearby, it expects to increase its sales. During the first 12 days after the start of bus stand, the daily sales were as following:

   550, 570, 490, 615, 505, 580, 570, 460, 600, 580, 530, 526.

   On the basis of sample information and assuming that population is normally distributed, can one conclude that restaurant's sales have increased?

6. In practical datasets, it's quite common to encounter instances where certain attributes have missing values. Discuss different approaches for addressing this issue.

7. Describe the process of data integration. Why is this important in data pre-processing?

8. a) What is the difference between supervised and unsupervised learning? [5]

b) What is reinforcement learning? Given a classification task, you can solve it using either a supervised learning approach or a reinforcement learning approach. What is the difference between these two approaches? [5]

9. a) What is multiple and partial correlation? How we can find these correlations for a given data. [4]

b) Use the method of least squares, to fit a simple linear regression line to the following data point. [6]

x	-1	0	2	-2	5	6	8	11	12	-3
y	-5	-4	2	-7	6	9	13	21	20	-9

10. Find the singular value decomposition of a matrix with elements $a_{11} = 2, a_{12} = 2, a_{21} = 1, a_{22} = 1.$

11. Discuss in detail, In what scenarios would principal component analysis be particularly useful for data analysis and dimensional reduction?

12. Compare and contrast the feature selection algorithms: Decision Trees and Random Forests. Highlight their strengths and weaknesses.

⇔⇔⇔

[5]

VIT
Vellore Institute of Technology

Final Assessment Test – November/December 2023

Course: RES5001 - Research Methodology

Class NBR(s): 6607 / 6608 / 6609 Faculty Name: Prof. YOGIRAJ MANTRI

Time: Three Hours Max. Marks: 100

KEEPING MOBILE PHONE/SMART WATCH, EVEN IN 'OFF' POSITION, IS TREATED AS EXAM MALPRACTICE

Answer any TEN Questions

(10 X 10 = 100 Marks)

1. Briefly describe some of the characteristics of scientific method to do research.
2. Explain the different types of research objectives. Give suitable example for each.
3. a) What is the necessity to define a research problem? [5]

   b) What are some points one needs to keep in mind while selecting a research problem. [5]
4. What are some features of a good research design?
5. Industrial psychologists want to conduct workplace observational studies to describe employee behaviours, job performance, and corporate culture. This descriptive research can provide insights into factors that impact productivity, job satisfaction, and employee well-being.

   What are some factors the psychologists should consider while preparing a research design for this study?
6. What are some differences between probability and non-probability sampling? Give two examples for sampling methods within each category.
7. An agricultural researcher wants to study the effect of four new fertilizers that were launched recently on the yield of the crop. He also wants to minimise the effects of extraneous variables such as soil fertility, different varieties of seeds etc, in his study.

   Explain the three basic principles of experimental design for this study with regards to dependent, independent and extraneous variables.
8. The following are marks of 10 students in Statistics and Algebra out of 50 marks each:

Statistics	34	23	45	41	16
	24	33	39	20	28
Algebra	40	28	38	42	21
	19	31	47	14	34

Answer the following questions:

a) What is the mean and standard deviation of marks for each subject.

b) Find Pearson's correlation coefficient between the marks in Statistics and Algebra. Interpret your result.

c) Calculate the equation of regression line between the marks in two subjects and show it in a plot.

9. Briefly explain the layout of a research paper.

10. a) Explain the different types of animal models. [5]

b) What is meant by LD50, ED50 in measuring toxicity of drugs? [5]

11. What are some available resources to find information for your research?

12. What are some effective search strategies to find information using search tools, such as search engines (Google), the library catalogue and online databases?

⇔⇔⇔

VIT
Vellore Institute of Technology

Final Assessment Test – November/December 2023

Course: MAT5011 - Matrix Theory and Linear Algebra

Time: Three Hours

Max. Marks: 100

KEEPING MOBILE PHONE/SMART WATCH, EVEN IN 'OFF' POSITION, IS TREATED AS EXAM MALPRACTICE

General Instructions :

$\mathbb{R}$ is the set of real numbers

Answer any TEN Questions

(10 X 10 = 100 Marks)

1. a) Find the inverse of the following matrix using Gauss-Jordon's elimination method: **[5]**

$$A = \begin{pmatrix} 0 & -3 & -2 \\ 1 & -4 & -2 \\ -3 & 4 & 1 \end{pmatrix}.$$

b) Let $A = \begin{pmatrix} 6 & -2 & 2 \\ -2 & 3 & -1 \\ 2 & -1 & 3 \end{pmatrix}$. If 2 and 8 are two eigenvalues of A, find the third eigenvalue of A. Hence find all the eigenvalues of A^{100}. **[5]**

2. Solve the system of equations by LU decomposition method:

$$x_1 - 5x_2 + x_3 = 2,$$
$$2x_1 + 4x_2 + x_3 = 1,$$
$$x_1 + x_2 + x_3 = 0.$$

3. Let $V = \{(a,b,c,d) \in \mathbb{R}^4 : b - 2c + d = 0\}$; $W = \{(a,b,c,d) \in \mathbb{R}^4 : a = d, b = 2c\}$ be two subspaces of $\mathbb{R}^4(\mathbb{R})$, find the bases and the dimensions of V, W and $V \cap W$.

4. Let T: $\mathbb{R}^3 \to \mathbb{R}^3$ be defined by

$$T(a_1, a_2, a_3) = (2a_2 + a_3,\ -a_1 + 4a_2 + 5a_3,\ a_1 + a_3).$$

Determine the matrix representation of T with respect to the standard basis of $\mathbb{R}^3$.

5. Apply Gram-Schmidt process to the vectors (1, 0, 1), (0, 1, 1), and (1, 3, 3) to obtain an orthogonal basis for $\mathbb{R}^3$ with respect to standard inner product.

6. Find the Singular Value Decomposition (SVD) of A, $U\Sigma V^T$, where

$$A = \begin{pmatrix} 3 & 2 & 2 \\ 2 & 3 & -2 \end{pmatrix}.$$

7. Find the generalized inverse of the following matrix:

$$A = \begin{pmatrix} 1 & 2 & 4 & 3 \\ 3 & 5 & 12 & 9 \\ 2 & 4 & 8 & 6 \end{pmatrix}.$$

8. Determine the Jordan form of the operator represented by the matrix

$$A = \begin{pmatrix} -1 & -1 & 0 \\ 0 & -1 & -2 \\ 0 & 0 & -1 \end{pmatrix}.$$

9. What do you mean by over-determined and under-determined systems? Is the following system over or under-determined?

$$-x_1 + x_2 - x_3 + 3x_4 = 0,$$
$$3x_1 + x_2 - x_3 - x_4 = 0,$$
$$2x_1 - x_2 - 2x_3 - x_4 = 0.$$

Find an approximate solution to the above system of equations.

10. Reduce the quadratic form $5x^2 + 6y^2 + 14z^2 - 14yz - 10zx$ to a canonical form through an orthogonal transformation. Identify the definiteness, rank, index, and signature of the quadratic form.

11. Diagonalize the following matrix if possible

$$A = \begin{pmatrix} 1 & 1 & 4 & 0 \\ 0 & 1 & 1 & 6 \\ 0 & 0 & -2 & 1 \\ 0 & 0 & 0 & -2 \end{pmatrix}.$$

12. a) Let V be an inner product space, and α and β, are orthogonal vectors in V. Prove that $\|\alpha + \beta\|^2 = \|\alpha\|^2 + \|\beta\|^2$. **[5]**

b) Show that the set of vectors $S = \{(4,0,3), (3,0,-4), (0,10,0)\}$ are orthogonal, while the set $T = \left\{\frac{1}{5}(4,0,3), \frac{1}{5}(3,0,-4), (0,1,0)\right\}$ is orthonormal. **[5]**

⇔⇔⇔

VIT
Vellore Institute of Technology

Final Assessment Test – November/December 2023

Course: MAT5012 - Probability Theory and Distributions

Time: Three Hours

Max. Marks: 100

KEEPING MOBILE PHONE/SMART WATCH, EVEN IN 'OFF' POSITION IS TREATED AS EXAM MALPRACTICE

General Instructions: 1. Statistical Tables and Scientific calculator are permitted
2. Programmable calculator is not permitted

Answer any TEN Questions

(10 X 10 = 100 Marks)

1. Urn A contains 3 red and 3 black balls, whereas urn B contains 4 red and 6 black balls. If a ball is randomly selected from each urn, what is the probability that the balls will be the same color?

2. A type C battery is in working condition with probability 0.7, whereas a type D battery is in working condition with probability 0.4. A battery is randomly chosen from a bin consisting of 8 type C and 6 type D batteries

   a) What is the probability that the battery works? [5]

   b) Given that the battery does not work, what is the conditional probability that it was a type C battery? [5]

3. For the joint probability distribution of X and Y given below :

X \ Y	1	2	3	4
1	4/36	3/36	2/36	1/36
2	1/36	3/36	3/36	2/36
3	5/36	1/36	1/36	1/36
4	1/36	2/36	1/36	5/36

Find a) the marginal distributions of X and Y. [5]

b) Conditional distribution of X given the value of $Y = 1$ and that of Y given the value of $X = 2$. [5]

4. The joint probability density function of the two-dimensional variable (X, Y) is of the form :

$$f(x,y) = \begin{cases} 2e^{-x}e^{-2x} & ; \ 0 < x < \infty \ , 0 < y < \infty \\ 0 & ; \ otherwise \end{cases}$$

Compute (i) $P(X > 1, Y < 1)$ [5]

(ii) $\mathrm{Cov}(X, Y)$. [5]

5. The random variable X has the probability distribution

$$f(x) = \begin{cases} \frac{x}{18} & ; \ 0 \le x \le 6 \\ 0 & ; \ otherwise \end{cases}$$

(i) Find the probability distribution of the random variable $Y = 10 + 2X$. [7]

(ii) Also, find the $E(Y)$. [3]

6. If the random variable X is uniformly distributed over $\left(1 - \frac{1}{\sqrt{3}}, \ 1 + \frac{1}{\sqrt{3}}\right)$, compute $P\left\{|X - \mu| \ge \frac{3}{2}\sigma\right\}$ and compare it with the upper bound obtained by Tchebycheff's inequality. Give specific comments on result obtained.

7. Suppose that moment generating function of a random variable X is given by

$$M_X(t) = e^{3(e^t-1)}$$ [4]

Find (i) $P(X = 0)$ [6]

(ii) Mean and Variance of X using it.

8. Fit a Poisson distribution for the following distribution and hence find the expected frequencies.

X	0	1	2	3	4	5	6
f	314	335	204	86	29	9	3

Give specific comments on your findings.

9. a) An excellent free-throw shooter attempts several free throws until she misses. [5]
If $p = 0.9$ is her probability of making a free throw, what is the probability of having the first miss on the 13th attempt or later?

b) If an auditor selects 5 returns from among 15 returns of which 9 contain [5]
illegitimate deductions, what is the probability that the auditor will catch only 2-income tax returns contains illegitimate deductions.

10. The time it takes a cell to divide (called mitosis) is normally distributed with an average time of 60-minutes and a standard deviation of 5 minutes.

a) What is the probability that a cell divides in less than 45 minutes? [3]

b) What is the probability that it takes a cell more than 65 minutes to divide? [3]

c) What is the time that it takes approximately 99% of all cells to complete mitosis? [4]

11. The lifetime of a mechanical assembly in a vibration test is exponentially distributed with a mean of 400 hours.

a) What is the probability that an assembly on test fails in less than 100 hours? [3]

b) What is the probability that an assembly operates for more than 500 hours before failure? [3]

c) If an assembly has been on test for 400 hours without a failure, what is the probability of a failure in the next 100 hours? [4]

12. Suppose the time it takes a data collection operator to fill out an electronic form for a database is uniformly between 1.5 and 2.2 minutes.

a) What is the mean and variance of the time it takes an operator to fill out the form? [4]

b) What is the probability that it will take less than two minutes to fill out the form? [3]

c) Determine the cumulative distribution function of the time it takes to fill out the form. [3]

⇔⇔⇔

VIT
Vellore Institute of Technology

Final Assessment Test – November/December 2023

Course: **MAT6012 - Programming for Data Analysis**

Time: **Three Hours** Max. Marks: **100**

KEEPING MOBILE PHONE/SMART WATCH, EVEN IN 'OFF' POSITION, IS TREATED AS EXAM MALPRACTICE

Answer any TEN Questions

(10 X 10 = 100 Marks)

1. Construct the truth table for the following compound proposition and draw the conclusion regarding the statement $((p \rightarrow q) \rightarrow r) \vee (p \rightarrow \neg r)$.

2. A number $N = d_1 d_2 \ldots d_n$ is said to be Disarium if the sum of each digit powered by their position is equal to the number. For example, $135 = 1^1 + 3^2 + 5^3$ and therefore 135 is a Disarium number.

   Write the Pseudo code and draw the flowchart to verify whether a given number is Disarium or not.

3. a) Define the terms (i) Variables, (ii) Identifiers, (iii) Literals, (iv) Operators, (v) Constant Variables and (vi) Expression in Python and provide examples. **[7]**

   b) Write a short note on Operator Precedence in Python by providing proper examples. **[3]**

4. What is mutability and immutability in Python? What is Set datatype in Python, and what are its uses? Is Set mutable or immutable? Justify your answer. List any six methods that can be operated on Set datatype with the help of proper examples.

5. The digital root of a natural number n is obtained as follows: Add up the digits n to get a new number, then add up the digits of the new number just obtained to get another new number. Repeat this until you are left with a number with only one digit, which is the digital root of the given number.

   For example, if n = 54123, at first, we get = 5 + 4 + 1 + 2 + 3 = 15. Adding the digits of 15, we obtain 1 + 5 = 6. Since 6 has only one digit, 6 is our digital root for n = 54123.

   Draw the flowchart for finding the digital root of a number and write a Python code to determine the digital root of a number (taken input for the user) using while/for/if conditions. The program should check that the number n is a natural number before finding the digital root. If the number is a natural number, the program prints the digital root. Otherwise, the program should print, "Please enter a valid number which is an integer and > 0"

6. Consider a positive integer n. If the sum of all the divisors of the number n is greater than $2n$, then the number n is said to be an Abundant number. If the sum of all divisors of the number n is less than $2n$, then the number n is said to be a Deficient number.

   For example, the number 24 is abundant as the sum of its divisors 1, 2, 3, 4, 6, 8, 12 and 24 is 60, which is greater than 2 x 24 = 48.

   The number 21 is deficient as the sum of its divisors 1, 3, 7 and 21 is 32, which is less than 2 x 21 = 42.

   Write a Python function program AbundantDeficien $t(n)$ that accepts an integer argument n, and then the function checks whether the n is an Abundant number or Deficient number and displays an appropriate message to the user.

7. What are generator functions? Create a generator function named GenPrimes(n). When n is not passed, then the function should consider $n = 2$. The functionality of this function is to generate infinite prime numbers starting from the given number.

Later, use the above function to generate the first 100 hundred primes after 1000 and find their sum.

8. Using the list comprehension technique, create the following matrix as a NumPy array and name it A.

0	1	2	3	4	5
10	11	12	13	14	15
20	21	22	23	24	25
30	31	32	33	34	35
40	41	42	43	44	45
50	51	52	53	54	55

Write the Python code segments to perform the following operations on the NumPy array A.

(i) Extract the subarray $\begin{bmatrix} 30 & 32 & 34 \\ 40 & 42 & 44 \\ 50 & 52 & 54 \end{bmatrix}$ and name it as B.

(ii) Extract the subarray $\begin{bmatrix} 44 & 42 \\ 24 & 22 \\ 4 & 2 \end{bmatrix}$ and name it C.

(iii) Extract the diagonal entries of the matrix A and store them in D.

(iv) Find the matrix product of BC and store it in E.

(v) Find the matrix product of $C^T B$ and store it in F.

(vi) Is the evaluation of EF (element-by-element multiplication of E and F) and B+C (element-by-element addition of B and C) possible? Justify your answer.

(vii) By appropriately reshaping the matrices C and D, generate the matrix of order 6×6 which is obtained by the product of C and D.

9. a) Write a Python program to fit the following data to the function $y = a \sin(bx)$ by identifying the a, b values. **[6]**

x	y
-2.0	0.03204
-1.6	-1.15937
-1.2	-3.13395
-0.8	-3.12632
-0.4	-2.06561
0.0	1.61216
0.4	1.39841
0.8	1.32860
1.2	5.24644
1.6	1.45341
2.0	-1.63063

b) Write the Python commands to evaluate the following integrals [4]

(i) $\int_{x=2}^{-2} \int_{y=x}^{x^2} \int_{z=x+y}^{2y} x + y + z \; dz \; dy \; dx$

(ii) $\int_{y=3}^{y=-3} \int_{x=-1}^{x=1} \frac{\left(e^{x^2+y^2}\right)}{x+y+2} \; dx \; dy$

10. The following data displays the marks obtained by 10 students in Mathematics and Physics subjects.

Student_ID	*1*	*2*	*3*	*4*	*5*	*6*	*7*	*8*	*9*	*10*
Marks_in_Maths	88	92	80	89	100	80	60	100	80	34
Marks_in_Physics	35	79	79	48	100	88	32	45	20	30

Write a Python program to draw a (i) scatter plot, (ii) bar plot and (iii) line plot to compare both marks in a 3 x 1 subplot grid. Label the x-axis as "Student ID" and the y-axis as "Marks in Maths and Physics". Give a legend to the plots and put the title "Comparison of Marks in Maths and Physics Subjects".

11. Illustrate the use of pipe(), apply() and applymap() in Pandas by providing appropriate examples.

12. What are the exceptions in Python? How do you handle exceptions in Python? Discuss the ZeroDivisionError, NameError, IndexError, ValueError in Python and how you handle these errors through examples.

⇔⇔⇔

VIT
Vellore Institute of Technology
(Deemed to be University under section 3 of UGC Act, 1956)

SCHOOL OF ADVANCED SCIENCES

Continuous Assessment Test-I

Winter Semester (2023-24)

Course Name & Code: Regression Analysis and Predictive Models & MAT6002 **SLOT: D1+TD1**

Prog. Name & Branch: M. Sc. Data Science **Exam Duration: 90 min.**

Faculty Name: Dr. MAHAMOOD USMAN **Max. Marks:50**

Class Number: VL2023240501473

Answer All the Questions:

<table>
<tr><th>Q.No.</th><th>Question</th><th>Max Marks</th><th>CO</th><th>BL</th></tr>
<tr><td>1.</td><td>a) Discuss simple regression model with the mean of a real life example.
b) Give the interpretation of slope parameter in a simple linear regression when it is positive and negative.</td><td>7+3</td><td>CO2</td><td>BL2</td></tr>
<tr><td>2.</td><td>Fit a linear trend model for the following time series data
<table><tr><td>Year</td><td>2017</td><td>2018</td><td>2019</td><td>2020</td><td>2021</td><td>2022</td></tr><tr><td>Profit</td><td>45.5</td><td>40.8</td><td>52.7</td><td>50.0</td><td>64.5</td><td>62.2</td></tr></table>Compute all the trend values from 2017-2022 and Estimate the trend projection for 2027.</td><td>10</td><td>CO4</td><td>BL3</td></tr>
<tr><td>3.</td><td>For the data given below
<table><tr><td>Exam score</td><td>94</td><td>88</td><td>71</td><td>75</td><td>72</td></tr><tr><td>Revision time</td><td>80</td><td>80</td><td>56</td><td>47</td><td>43</td></tr></table>The fitted regression line is
Exam score=44.5+0.55*Revision time
Can we assume the slope is 0.7 for this model. Test at 5% level of significance. Also obtain 95% confidence interval for slope parameter.</td><td>10</td><td>CO1</td><td>BL5</td></tr>
</table>

4.	Fit a multiple linear regression model for the following data <table><tr><td>Y</td><td>7</td><td>4</td><td>5</td><td>2</td><td>3</td></tr><tr><td>X_1</td><td>2.5</td><td>1.5</td><td>2.0</td><td>1.0</td><td>3.5</td></tr><tr><td>X2</td><td>8</td><td>9</td><td>11</td><td>10</td><td>8</td></tr></table>Predict the value of Y when $(x_1, x_2) = (4.5,10)$.	10	CO2	BL4
5.	Define the followings in regression analysis *(with mathematical expressions)*: a) Residual b) Sum of squares due to residuals c) Sum of squares of regression d) Total sum of squares. e) R-square f) Adjusted R-square g) Estimate of the variance of error term in multiple regression.	1+(6×1.5)	CO2	BL1

VIT
Vellore Institute of Technology

B1+TB1

SCHOOL OF ADVANCED SCIENCES
CONTINUOUS ASSESSMENT TEST – I
WINTER SEMESTER 2023-24

Programme Name & Branch: M.Sc. (Data Science)
Course Code: MAT5013
Course Name: Statistical Inference
Faculty Name: Dr. Venkataramana B
Class Number: VL2023240501469 **Exam Duration:** 90 minutes **Maximum Marks:** $5 \times 10 = 50$

General instruction(s): Statistical tables are allowed and Answer all the questions.

1 **(a)** Define the sufficient estimator. (2M)

(b) Let $X_1, X_2, \ldots X_n$ be a random sample of size n from a population with the probability density function

$$f(x,\beta) = \frac{x}{\beta} e^{\left(\frac{-x^2}{2\beta}\right)} \quad ; x \geq 0, \ \beta > 0$$

Find a sufficient statistic for the parameter β. (8M)

2 **(a)** What is an efficient estimator? (2M)

(b) Let X_1, X_2, X_3 and X_4 be independent random variables such that $E(X_i) = \mu$ and $Var(X_i) = \sigma^2$ for $i = 1,2,3,4$.

If $Y_1 = \frac{X_1+X_2+X_3+X_4}{4}$, $Y_2 = \frac{2X_1+X_2+X_3+X_4}{5}$ and $Y_3 = \frac{X_1+2X_2+X_3-X_4}{4}$

Examine whether Y_1, Y_2 and Y_3 are unbiased estimators for μ.
Find the best estimator? In what sense is it better? (8M)

3 Let $X_1, X_2, \ldots X_n$ be a random sample of size n from a population with the probability density function

$$f(x,\theta) = \frac{1}{2\theta^3} x^2 e^{-x/\theta} \quad ; X > 0, \ \theta > 0.$$

(i) Obtain the maximum likelihood estimator of the parameter θ.

(ii) Calculate the estimate when $x_1 = 2.00, x_2 = 7.5, x_3 = 4.00, x_4 = 3.0$.

4 Obtain the Minimum Variance Bound estimator for μ in normal population $N(\mu, \sigma^2)$, when σ^2 – is known.

5 Let $Y_1, Y_2, \ldots Y_n$ be a random sample of size n from a population with the probability density function

$$f(y,\alpha) = \frac{1}{\alpha^2} x e^{-x/\alpha} \ ; \ X > 0, \alpha > 0.$$

(i) Obtain the method of moments estimator of the parameter α.

(ii) Calculate the estimate when $x_1 = 0.25, x_2 = 0.75, x_3 = 1.50, x_4 = 2.5, x_4 = 2.0$.

VIT
Vellore Institute of Technology

School of Advanced Sciences (SAS)

Winter Semester 2023-24

Continuous Assessment Test – 1

Programme Name &Branch: MSc Data Science **SLOT – F2**

Course Name & code: Database Systems: Design and Implementation CSE5003

Class Number (s): VL2023240505062

Faculty Name (s): Dr Karthik K.

Exam Duration: 90 Min. **Maximum Marks:** 50

General instruction(s):

Answer all the Questions.

Q.No.	Question	Max Marks	CO	BL
1.	For the **Questions 1 and 2** use the Humane Society database with the following relations: Animals(AnimalID, Name, PrevOwner, DateAdmitted, Type) Adopter(SIN, Name, Address, OtherAnimals) Adoption(AnimalID, SIN, AdoptDate, chipNo, AdopDuration) Design a complete ER diagram for a database for Humane Society. The set of relations with attributes are given, where (a) The primary keys are underlined. (b) Animals stores information about the animals currently at the Humane Society. Type refers to animal like (dog, cat, etc) (c) Adopter is the relation that holds information about animal adopters. SIN is owners ID. OtherAnimals records the number of other animals that the adopter currently has at home. (d) Attribute chipNo stores the number on the microchip that is implanted on the animal for tracking. All other attributes are self descriptive.	10	CO2	BL6
2.	a) Connect the relations given in Q.No.1 of the Humane Society database with a schema diagram. (4 marks) b) Write a create table statement for the **Adoption** table. (Assume you have already created table for Animals and Adopter as per SQL statements). (2 marks) c) Insert a tuples for **Adoption table with and without** using attribute set. (2 marks) d) Write an SQL to Retrieve the total number of dogs that were brought to the Humane Society on 18 April 2000. (2 marks)	10	CO2	BL2

3.	a) Giving a real time example, explain the states of transactions in DBMS (6 Marks) b) Relate your example to ACID properties of transactions. (4 Marks)	10	CO1	BL3
4.	a) Giving an example explain serial and parallel transaction in DBMS. (4marks) b) Explain the different types of attributes available in DBMS by providing appropriate example. (6 Marks)	10	CO1	BL1
5.	Given relation R(A, B, C, D, E, F, G) and the set of functional dependencies F ={BCD→A, BC →E, →F, F→G, C→D, A→G}, decompose R into 3NF. Show your steps. Is this decomposition also BCNF? Justify your answer.	10	CO1	BL5

BL3

VIT
Vellore Institute of Technology

SCHOOL OF ADVANCED SCIENCES

Department of Mathematics

Winter Semester 2023-24

Continuous Assessment Test –I

Programme Name & Branch: (M.Sc. Data Science - 23MDT)

Slot: (C1+TC1)

Course Name & code: MAT 5016 – TIME SERIES ANALYSIS AND FORECASTING

Class Number (s): VL2023240501267

Exam Duration: [90 Minutes] **Maximum Marks: [50]**

General instruction(s):

1. **Answer ALL Questions.**
2. **Scientific calculator and statistical table must be permitted as per need but use of programmable calculator/device is strictly prohibited.**

<table>
<tr><th>Q.No.</th><th>Question</th><th>Max Marks</th></tr>
<tr><td>1.</td><td>a. Explain briefly the additive and multiplicative models of time series. Which of these models is more popular in practice and why? [5]

b. Describe any two methods of trend measurements and examine critically the merits and demerits of these methods. [5]</td><td>10</td></tr>
<tr><td>2.</td><td>Fit a straight line trend by the method of least squares to the following indices and calculate the trend values.
<table>
<tr><td>Yrs.</td><td>2015</td><td>2016</td><td>2017</td><td>2018</td><td>2019</td><td>2020</td><td>2021</td></tr>
<tr><td>Index No.</td><td>127</td><td>101</td><td>130</td><td>132</td><td>126</td><td>142</td><td>137</td></tr>
</table>
Comment on the fitted trend and find the forecast value for the year 2023.</td><td>10</td></tr>
<tr><td>3.</td><td>The following table shows the historical sales demand for a recently released pharmaceutical product.
Table: Historical sales demand (in 000s million RS.)
<table>
<tr><th>Year</th><th>Quarter</th><th>T</th><th>X_t (Sales)</th></tr>
<tr><td rowspan="4">2020</td><td>Q1</td><td>1</td><td>30</td></tr>
<tr><td>Q2</td><td>2</td><td>50</td></tr>
<tr><td>Q3</td><td>3</td><td>80</td></tr>
<tr><td>Q4</td><td>4</td><td>130</td></tr>
<tr><td rowspan="3">2021</td><td>Q1</td><td>5</td><td>180</td></tr>
<tr><td>Q2</td><td>6</td><td>230</td></tr>
<tr><td>Q3</td><td>7</td><td>300</td></tr>
</table></td><td>10</td></tr>
</table>

2022	Q4	8	350
	Q1	9	370
	Q2	10	320
	Q3	11	380
	Q4	12	390

a. Plot the sales demand against time and give your comment on this graph. [3]
b. Using above time series data, fit an **additive model** to the data and find the estimates of the four seasonal factors. [7]

4. a. Conduct the test for **seasonality** using **Kruskal-Wallis** one- way analysis of variance test to the outcomes that were obtained by multiplicative model as shown in following table: 10

T	Quarter	$S_t.e_t$
3	Q3	0.9872
4	Q4	1.074
5	Q1	0.9082
6	Q2	1.0180
7	Q3	1.0109
8	Q4	1.0703
9	Q1	0.9071
10	Q2	1.0041
11	Q3	1.014
12	Q4	1.077
13	Q1	0.8922
14	Q2	1.0197

The chi square critical value at df(degrees of freedom) =3, and α(level of significance)=0.05 is **7.81.** [8]

b. What will be the H_0 and H_1, if you will apply the same test for seasonality using **additive model?** [2]

5. Yearly sales data for a certain product were collected over a 12 month period as shown below in a table form. Apply single exponential smoothing model with smoothing constant α = 0.3 and compute the **MAD** and **RMSE** values. Also, give suitable comments on your findings. 10

t	Sales(X_t) in 000 million
1	14.7
2	16.0
3	15.2
4	19.5
5	19.0
6	22.0
7	25.6
8	21.2
9	21.0
10	20.5
11	21.0
12	24.0

****

A sample of 400 male students is found to h
ably regarded as a sample from

VIT

Vellore Institute of Technology

SCHOOL OF ADVANCED SCIENCES

Winter Semester 2023-2024

Continuous Assessment Test –I

Programme Name & Branch: M.Sc. (Data Science)

Slot: E1+TE1

Course Name & code: MAT6005 & Machine Learning for Data Science

Class Number (s): VL2023240501476

Exam Duration: 90 Min. **Maximum Marks: 50**

General instruction(s): Answer ALL Questions

<table>
<tr><th>Q.No.</th><th>Question</th><th>Max Marks</th></tr>
<tr><td>1.</td><td>Explain machine learning and write its importance with four examples.</td><td>10</td></tr>
<tr><td>2.</td><td>What is an example of a regression problem? Find the quadratic regression model for the following data:
<table>
<tr><td>x</td><td>2</td><td>4</td><td>5</td><td>8</td><td>10</td><td>13</td><td>20</td><td>22</td><td>25</td></tr>
<tr><td>y</td><td>15</td><td>22</td><td>32</td><td>78</td><td>105</td><td>178</td><td>405</td><td>500</td><td>630</td></tr>
</table></td><td>10</td></tr>
<tr><td>3.</td><td>What is Association rule? Trace the results of using the Apriori algorithm on the grocery store example with support threshold S = 40% and confidence threshold C = 60%. Enumerate all the final frequent itemsets. Also indicate the association rules that are generated and highlight the strong ones, sort them by confidence.
<table>
<tr><th>Transaction ID</th><th>Items</th></tr>
<tr><td>T1</td><td>Apple, Chocolate</td></tr>
<tr><td>T2</td><td>Apple, Biscuit, Chocolate</td></tr>
<tr><td>T3</td><td>Biscuit, Black Jam, Cream Soda</td></tr>
<tr><td>T4</td><td>Cream Soda, Apple, Chocolate, Blue Band</td></tr>
<tr><td>T5</td><td>Apple, Cream Soda</td></tr>
</table></td><td>10</td></tr>
<tr><td>4.</td><td>Why SVM is an example of a large margin classifier? Find the hyperplane for the following data:
<table>
<tr><td>x_1</td><td>1</td><td>1</td><td>-1</td><td>-1</td><td>5</td><td>5</td><td>-5</td><td>-5</td></tr>
<tr><td>x_2</td><td>1</td><td>-1</td><td>1</td><td>-1</td><td>4</td><td>-4</td><td>4</td><td>-4</td></tr>
<tr><td>y</td><td>O</td><td>O</td><td>O</td><td>O</td><td>X</td><td>X</td><td>X</td><td>X</td></tr>
</table></td><td>10</td></tr>
</table>

5.	List down the attribute selection measures used by the ID3 algorithm to construct a Decision Tree. Consider the database described in the following table. Identify the Root Node of Decision Tree using the ID3 algorithm.	10

No.	Cloudy	Water	Wind	Play
1	True	Hot	High	No
2.	True	Hot	High	No
3.	False	Hot	High	Yes
4.	False	Cool	Normal	Yes
5.	False	Cool	Normal	Yes
6.	True	Cool	High	No
7.	True	Hot	High	No
8.	True	Hot	Normal	Yes
9.	False	Cool	Normal	Yes
10.	False	Cool	High	No

VIT
Vellore Institute of Technology

School of Advanced Sciences (SAS)

Winter Semester 2023-24

Continuous Assessment Test – 2

SLOT – F2

Programme Name &Branch: MSc Data Science

Course Name & code: Database Systems: Design and Implementation CSE5003

Class Number (s): VL2023240505062

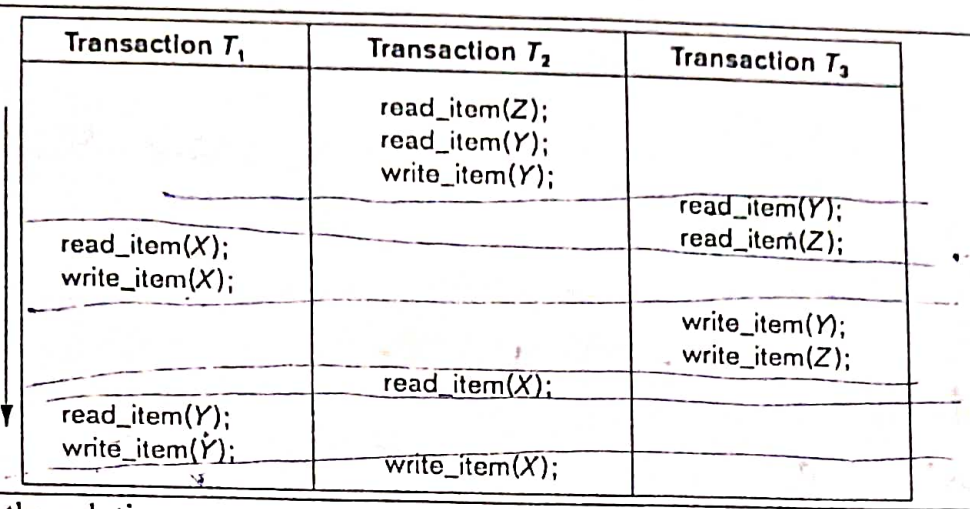
Faculty Name (s): Dr Karthik K.

Exam Duration: 90 Min. **Maximum Marks:** 50

General instruction(s):

Answer all the Questions.

Q.No.	Question	Max Marks
1.	**a)** Consider the following SQL query select t.branch_name from branch t, branch s where t.assets > s.assets and s.branch_city = 'Vellore'; on the relation - **branch(branch_name, branch_city, assets)** (i) What type of Join is used in the above query? Justify your answer. (ii) Write an efficient relational algebra expression that is equivalent to the above query. **b)** Suppose you have the following relations: employee(emp_id, salary, age, dept_id) department(dept_id, budget, status) Write **three** different types of Join Query for the above relations and sample output as per your data. (Note: Sub Join types are not considered as two different Join Queries)	(4+6)
2.	**a)** Consider the following relation Car(Carid, Name, Make, Model, Price, Type) Partition the above table by price. Create 3 fragments of approximately equal size by Horizontal Fragmentation. Describe the three fragments by using appropriate operations. Show the results. **b)** Consider the three transactions T1, T2, and T3, and the schedules S given in next page. Draw the serializability (precedence) graph for S and state whether the schedule is serializable or not. Justify your answer.	(4+6)



3.	Given the relation STUDENT (SID, REG.NO, FIRSTNAME, LASTNAME, COURSE, GRADE) consider the table consists of two tuples (i) How **identity** can be used in SID attribute in generating five tuples for the above relation. Describe how **identity** can be use with syntax and example. Also write necessary SQL statement and sample output. (ii) How do you generate an XML data file for the above relation? Write the query statement and a sample output. (iii) Write a XML Query to generate XML data consisting tags like <student details> <Reg.No.> <FirstName> <LastName> </stdent details> (iv) Write a XML Query to add <Reg.No.>, <FirstName>, <LastName> as attributes to <student details>	10
4.	a) How do industries achieve better performance in Partitioning? With an example illustrate your answer. b) Write a PL/SQL procedure program to assign values of **FirstName, LastName, Course, Grade** from the student table where **Reg.No = '23MDT1234'** and **display it** using appropriate DBMS output statements.	(5+5)
5.	**Mention** and **Illustrate** the different partitioning schemes with necessary diagram and an example with record number=12 considering the scenario – **"which machines will get the data/record"** based on the types of partitioning schemes.	10

VIT
Vellore Institute of Technology

B1+TB1

SCHOOL OF ADVANCED SCIENCES
CONTINUOUS ASSESSMENT TEST – II
WINTER SEMESTER 2023-24

Programme Name & Branch: M.Sc. (Data Science)
Course Code: MAT5013
Course Name: Statistical Inference
Faculty Name: Dr. Venkataramana B
Class Number: VL2023240501469 **Exam Duration:** 90 minutes **Maximum Marks:** $5 \times 10 = 50$

General instruction(s): Statistical tables are allowed and Answer all the questions

1 A survey reveals that of the 1000 randomly selected cases of lung cancer, 823 resulted in death within 10 years of the period.

(i) Construct a 95% two-sided confidence interval on the death rate from lung cancer.

(ii) Construct a 99% two-sided confidence interval on the death rate from lung cancer.

(iii) How large a sample would be required to be at least 95% confident that the error in estimating the 10-year death rate from lung cancer is less than 0.03?

2 Sample weights (in pounds) of newborn babies born in two adjacent counties in western Asia yielded the following data:

	Sample -1	Sample -2
Sample Size	50	45
Sample Mean	7.8	8.2
Sample Variance (S^2)	4.2	3.9

Verify the hypothesis that the mean weight of newborn babies is the same in both counties at a 5% significance level. What is the resulting p-value? Also, estimate the 95% Confidence interval on the difference in mean weights.

3 A company supplies plastic sheets for industrial use. A new type of plastic has been produced and the company would like to claim that the average stress resistance of this new product is at least 30.0, where stress resistance is measured in pounds per square inch (psi) necessary to crack the sheet. The following random sample was drawn off the production line. Based on this sample, would the claim clearly be unjustified?

1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16
30.1, 32.7, 22.5, 27.5, 27.7, 29.8, 28.9, 31.4, 31.2, 24.3, 26.4, 22.8, 29.1, 33.4, 32.5, 21.7

Assume normality and test the claim at 5% and 10% significance level.

4 Two independent samples of 8 and 7 items, respectively had the following values of the variable:

Sample 1	9	11	13	8	15	9	12	14
Sample 2	10	12	11	14	9	8	10	

Can the two samples be regarded as drawn from the same normal population? Use $\alpha = 0.05$.

5 In the following contingency table, 1018 individuals are classified by gender and by whether they favor, oppose, or have no opinion on a complete ban on smoking in public places:

	Smoking in Public Places		
Gender	*Favour*	*Oppose*	*No opinion*
Male	262	231	10
Female	302	205	8

Test the null hypothesis that gender and opinion on smoking in public places are independent at 5% significance level.

VIT

Vellore Institute of Technology

SCHOOL OF ADVANCED SCIENCES

Winter Semester 2023-24

Continuous Assessment Test –II

Programme Name & Branch: (M.Sc. Data Science - 23MDT)

Slot: (C1+TC1)

Course Name & code: MAT 5016 – TIME SERIES ANALYSIS AND FORECASTING

Class Number (s): VL2023240501267

Exam Duration: [90 Minutes] **Maximum Marks: [50]**

General instruction(s):

1. **Answer ALL Questions.**
2. **Student may use non programmable scientific calculator and statistical table is permitted..**

Q1. (Max Marks: 10)

The actual and predicted value of a time series is given in following incomplete table.

t	X_t	$\widehat{X_t}$	CI	PI
1	4	3.07		
2	6	5.57		
3	8	8.07		
4	9	10.57		
5	12	13.07		
6	15	15.57		

If X_t denote actual time series, $\widehat{X_t}$ denote predicted value, CI denote confidence interval and PI denote the predictive interval, and critical value at $\alpha = 0.5$, is $t_{\frac{\alpha}{2},4} = 2.77$; then compute the values of s_e, s_m, and s_p to complete the table. Also, give specific comments on relationship between s_e, s_m, and s_p.

Q2. (Max Marks: 10)

The following table shows the historical sales demand for a recently released pharmaceutical product.

Table: Historical sales demand (in 000s million RS.)

Year	Quarter	T	X_t (Sales)
2020	Q1	1	32
	Q2	2	54
	Q3	3	80
	Q4	4	120
2021	Q1	5	130
	Q2	6	180
	Q3	7	220
	Q4	8	280

2022	Q1	9	300	
	Q2	10	320	
	Q3	11	380	
	Q4	12	400	

a. Calculate a double moving average(DMA) forecast of length 3 **[6]**
b. Calculate the forecast for quarters Q1, Q2, Q3 and Q4 of the year 2023. Write your observations on forecast value obtained. **[4]**

Q3. Yearly sales data for a certain product was collected over a 12-month period is shown in a table given below. A linear regression model was fitted to the data and the following results obtained: 10

$\widehat{\beta_0}$=15.38 ; $\widehat{\beta_1}$= 0.707; R^2 = 0.578; F =13.73; t-statistic for $\widehat{\beta_1}$ =3.705;

Table: Yearly sales(in RS. 000000)

t	1	2	3	4	5	6	7	8	9	10	11	12
X_t	14.7	16.0	15.2	19.5	19	22	25.6	21.2	21	20.5	21	24

a. Using $\alpha = 0.80$ as a parameter for the level and $\beta^* = 0.0001$ as the smoothing parameter for the trend; apply Holt's trend method to forecast the values of the months 13, 14, 15, 16, 17 and 18 and give suitable comments on your findings. [8]
b. Compute the value of $R_{adjusted}{}^2$; and give comment. [2]

Q4. Given the time series 10

t	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
X_t	50	40	60	48	52	42	44	60	45	58	42	55	46	36	64

a. Calculate the sample ACF, $\widehat{\rho_k}$ for k = 1, 2, 3, 4, 5. [6]
b. Compute the associated standard error of ρ_k and corresponding value of the statistic $t_{\rho k}$. Give suitable comments on your findings. [4]

Q5. For the given time series data in Q4 above, calculate and plot the sample PACF, $\widehat{\varphi_{kk}}$ for k = 1, 2, 3, 4, 5. Also, give the suitable comments on your findings. 10

VIT

Vellore Institute of Technology

(Deemed to be University under section 3 of UGC Act, 1956)

SCHOOL OF ADVANCED SCIENCES

Continuous Assessment Test-II

Winter Semester (2023-24)

Course Name & Code: Regression Analysis and Predictive Models & MAT6002 **SLOT: D1+TD1**

Prog. Name & Branch: M. Sc. Data Science **Exam Duration: 90 min.**

Faculty Name: Dr. MAHAMOOD USMAN **Max. Marks:50**

Class Number: VL2023240501473

General Instructions:

(i) Answer all the Questions.

(ii) Statistical Tables and non programmable scientific calculator are allowed in EXAM HALL.

Q.No.	Question	Max Marks	CO	BL
1.	In a residual analysis: e_i : -2.52 2.26 -1.11 2.67 -1.31 0.02 s_i^2 : 3.04 3.78 4.95 3.21 4.69 5.41 R-student: -1.76 1.31 -0.61 1.65 -0.77 0.01 Determine the followings based on above information (i) The number of explanatory variables (ii) Standardized residuals (iii) PRESS residuals (iv) PRESS statistics	10	CO5	BL2
2.	Draw the normal probability plot for residuals from the data given below X: 7 9 10 12 13 14 Y: 24 20 21 26 31 27 $\hat{Y}$: 19.57 22.18 23.48 26.09 27.39 28.70 Comment on the result.	10	CO6	BL3
3.	a) Draw the plot of residual against fitted values	4	CO4	BL1

	(using 12 hypothetical combined points) in which the residual can be contained in an outward opening funnel. Also comment on this. b) Recall the *Introduction (Background) of Box-Cox Method for transformation on y.*	6		
4.	From the data information given below Y: 11, 3, 2, 5, 9, 14 e_i: 3.59, -5.79, -3.57, -1.95, 1.13, 6.59 h_{ii}: 0.167, 0.541, 0.714, 0.19, 0.217, 0.167 $\hat{Y}'$: 2.59, 2.91, 2.16, 2.48, 2.69, 2.59 Check whether the variance of errors are constant. If not then apply the variance stabilizing transformation to the data and comment on the result. (Note: $\hat{Y}'$ here is the fitted y for transformed data)	10	CO4	BL2
5.	a) If for various values of $\lambda = (-3,-2,-1,0,1,2,3)$ the function $SS_{res}(\lambda) = \lambda^2 - 2\lambda + 3$. Then obtain the optimum value of λ. Also determine its approximate 95% confidence interval when $n = 53$.	7	CO6	BL3
	b) Suppose there is only one predictor variable then calculate mean sum of squares of residuals at optimum value of λ. Also write the final response, fit for the model.	3		

VIT
Vellore Institute of Technology
(Deemed to be University under section 3 of UGC Act, 1956)

SCHOOL OF ADVANCED SCIENCES

Winter Semester 2023-2024

Continuous Assessment Test –II

Programme Name & Branch: M.Sc. (Data Science)

Slot: E1+TE1

Course Name & code: MAT6005 & Machine Learning for Data Science

Class Number (s): VL2023240501476

Exam Duration: 90 Min. **Maximum Marks: 50**

General instruction(s): Answer ALL Questions

<table>
<tr><th>Q.No.</th><th>Question</th><th>Max Marks</th></tr>
<tr><td>1.</td><td>Formulate the support vector machine as constrained optimization problem for the following data:
<table>
<tr><td>x_1</td><td>1</td><td>0</td><td>0</td><td>-1</td><td>3</td><td>3</td><td>6</td><td>6</td></tr>
<tr><td>x_2</td><td>0</td><td>1</td><td>-1</td><td>0</td><td>1</td><td>-1</td><td>1</td><td>-1</td></tr>
<tr><td>y</td><td>-</td><td>-</td><td>-</td><td>-</td><td>+</td><td>+</td><td>+</td><td>+</td></tr>
</table></td><td>10</td></tr>
<tr><td>2.</td><td>How does maximum likelihood estimation (MLE) facilitate parameter estimation in statistical modeling? Suppose you have a dataset representing the lifetimes of light bulbs, which follows an exponential distribution. The observed lifetimes are: {100, 150, 200, 250, 300} hours. Determine the maximum likelihood estimate of the parameter λ for the exponential distribution.</td><td>10</td></tr>
<tr><td>3.</td><td>Discuss the factors that influence the construction of histograms, such as the number of bins and their width, and how these choices can impact the interpretation of the resulting histogram. You are given a dataset D = {1, 1, 2, 3, 4, 4, 4, 4, 5, 5, 6, 6, 6}. Using histogram estimation, find a probability density function with a bin-width of 2 and sketch it.</td><td>10</td></tr>
<tr><td>4.</td><td>How does the kernel density function visualize the distribution of data? You are given a dataset $D = \{1,3,5,6,9\}$. Using the Epanechnikov kernel density estimation, find a probability density function and sketch it.</td><td>10</td></tr>
<tr><td>5.</td><td>Explain the concept of k-NN estimation in the context of statistical modeling. Using the following data and the Euclidean distance, what is the result for a student who scored 76 for English, 56 for Physics, 45 for Chemistry, and 88 for Mathematics?
<table>
<tr><th>English</th><th>Physics</th><th>Chemistry</th><th>Mathematics</th><th>Result</th></tr>
<tr><td>25</td><td>76</td><td>23</td><td>78</td><td>Fail</td></tr>
<tr><td>67</td><td>55</td><td>45</td><td>90</td><td>Pass</td></tr>
<tr><td>47</td><td>89</td><td>74</td><td>50</td><td>Pass</td></tr>
<tr><td>23</td><td>44</td><td>54</td><td>60</td><td>Fail</td></tr>
<tr><td>12</td><td>11</td><td>10</td><td>18</td><td>Fail</td></tr>
<tr><td>78</td><td>56</td><td>86</td><td>60</td><td>Pass</td></tr>
<tr><td>34</td><td>40</td><td>30</td><td>25</td><td>Fail</td></tr>
<tr><td>80</td><td>94</td><td>92</td><td>98</td><td>Pass</td></tr>
</table></td><td>10</td></tr>
</table>

Reg. No: 23MDT0005

M/E/TX

VIT® Vellore Institute of Technology

Final Assessment Test – May 2024

Course: MAT5016 - Time Series Analysis and Forecasting

Slot: C1+TC1

Class NBR(s):1267

Max. Marks: 100

Time: Three Hours

- KEEPING MOBILE PHONE/ELECTRONIC DEVICES EVEN IN 'OFF' POSITION IS TREATED AS EXAM MALPRACTICE
- DON'T WRITE ANYTHING ON THE QUESTION PAPER

General Instructions:

1. Non Programmable Scientific calculator is permitted
2. Statistical Tables are permitted

Answer any TEN Questions

(10 X 10 = 100 Marks)

1. a) Explain the forecasting procedure of a given time series and discuss the criteria's to evaluate / judge a good forecast. [6]

   b) Define stationary and non-stationary time series. Suggest a method to achieve stationary from non-stationary behaviour of a time series. [4]

2. a) What are the advantages and disadvantages of a moving average method? [5]

   b) Apply four year (k=4) Centered Moving Average (CMA) method to the following time series and derive suitable conclusion from the result. [5]

   Table: Yearly sales:

Time (t)	Year	Sales (X_t) in RS. Billion.
1	2012	600
2	2013	620
3	2014	650
4	2015	672
5	2016	678
6	2017	700
7	2018	710
8	2019	730
9	2020	710
10	2021	690
11	2022	720
12	2023	760

3. The following table shows the historical sales demand for a recently released pharmaceutical product.

Table: Historical sales demand (in 000s million RS.)

Year	Quarter	T	Xt (Sales)
2020	Q1	1	300
	Q2	2	400
	Q3	3	700
	Q4	4	1000
2021	Q1	5	1800
	Q2	6	2200
	Q3	7	3000
	Q4	8	3500
2022	Q1	9	4000
	Q2	10	3500
	Q3	11	3800
	Q4	12	4200

a) Plot the sales demand against time and give your comment on this graph. [3]

b) Using above time series data, fit a multiplicative model to the data and find the estimates of the four seasonal factors. [7]

4. a) Conduct the Kruskal-Wallis one- way analysis of variance test for seasonality to the outcomes that were obtained by additive model as shown in following table: [8]

T	Quarter	(St + et)
3	Q3	10.56
4	Q4	7.56
5	Q1	-18.39
6	Q2	1.61
7	Q3	14.19
8	Q4	9.89
9	Q1	-26.97
10	Q2	6.64
11	Q3	9.61
12	Q4	-2.07
13	Q1	-16.19
14	Q2	7.66

The chi square critical value at df(degrees of freedom) 3, and α(level of significance) = 0.05 is 7.81.

b) What will be the H_0 and H_1, if you will apply the same test for seasonality using multiplicative model? [2]

5. If X_t, $\widehat{X_t}$, CI, and PI denotes actual time series, predicted values, confidence interval and the predictive interval respectively as shown in below incomplete table: **[10]**

Table: Actual and predicted time series.

t	X_t	$\widehat{X_t}$	*Error* E_t	CI	PI
1	4	4	0		
2	6	5.9	0.1		
3	8	7.8	0.2		
4	9	9.7	-0.7		
5	12	11.6	0.4		

Let the critical value at $\alpha = 0.05$, is $t_{\frac{\alpha}{2},3} = 3.182$; then compute the values of s_e, s_m and s_p and complete the table for CI and PI. Also, give specific comments on

a) the relationship between the CI and PI and
b) The relationship between s_e, s_m and s_p.

6. Write single exponential smoothing model with smoothing constant $\alpha = 0.4$ and apply this model to a given yearly production data for a certain product were collected over a 12 month period as shown below in a table form and compute the MAPE and RMSE values. Also, give suitable comments on your findings. **[10]**

Time(t)	Production(Xt) in million units
1	15
2	18
3	17
4	21
5	20
6	25
7	27
8	23
9	21
10	20
11	26
12	24

7. Yearly sales data for a certain product was collected over a 12-month period is shown in a table given below. A linear regression model was fitted to the data and the following results obtained:

$\widehat{\beta_0}$=15.38 ; $\widehat{\beta_1}$= 0.707; R^2 = 0.578; F =13.73; t-statistic for $\widehat{\beta_1}$ =3.705;

Table: Yearly sales(in RS. 000000)

t	1	2	3	4	5	6	7	8	9	10	11	12
X_t	14.7	16.0	15.2	19.5	19	22	25.6	21.2	21	20.5	21	24

a) Using $\alpha = 0.70$ as a parameter for the level and $\beta^* = 0.0001$ as the smoothing parameter for the trend; apply Holt's trend method to forecast the values of the months 13, 14, 15, 16, 17 and 18 and give suitable comments on your findings. [8]

b) Compute the value of $R_{adjusted}{}^2$; and give comment. [2]

8. A financial time series is shown in below table:

Table: Financial time series

t	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
X_t	60	50	70	48	55	40	45	65	48	58	38	58	44	36	74

a) Calculate the sample ACF, $\widehat{\rho_k}$ for k = 1, 2, 3, 4, 5 and sample PACF $\widehat{\varphi_{kk}}$ for $k = 1, 2, 3$. [7]

b) Compute the associated standard error of ρ_k and corresponding value of the statistic $t_{\rho k}$. Give suitable comments on your findings. [3]

9. Discuss any three of the followings: [10]

a) the basis of Box- Jenkins techniques for examining model adequacy?
b) steps involved in model identification
c) model selection criteria
d) diagnostic checking

10. a) For n = 72, the following are the values of estimated partial autocorrelation. Conduct a statistical test, if an AR (1) model fits the data. [6]

Table: PACF values for given k

k	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
$\widehat{\varphi_{kk}}$	-.40	.19	.01	-.07	-.08	-.15	.05	.00	-.10	.05	.18	-.06	.09	.10	.01

Give specific comment on your findings.

b) Explain the concept and method of variance stabilizing transformation. Write Box cox power transformation model and suggest suitable transformation for λ (known as transformation parameter) = -1.0, -0.5, 0.0, 0.5,1. [4]

11. Write short notes on any three of the followings: [10]

a) ARIMA(1,1,1) models
b) White Noise Model
c) Random Walk Model
d) ARMA model

12. Discuss spectral density function (sdf) and its properties of AR and ARMA models. [10]

⇔⇔⇔ M/E/TX⇔⇔⇔

J/E/TX

Reg. No: 23MD1005

VIT® Vellore Institute of Technology

Final Assessment Test – May 2024

Course: MAT6002 - Regression Analysis and Predictive Models

Class NBR(s):1473

Slot: D1+TD1

Time: Three Hours

Max. Marks: 100

- KEEPING MOBILE PHONE/ELECTRONIC DEVICES EVEN IN 'OFF' POSITION IS TREATED AS EXAM MALPRACTICE
- DON'T WRITE ANYTHING ON THE QUESTION PAPER

Answer any TEN Questions

(10 X 10 = 100 Marks)

1. Fit a simple linear regression model for *Y* from the following data **[10]**

Price in Rs. (X)	10	12	13	12	16	15
Amount demanded (Y)	40	38	43	45	37	43

Predict likely demand when Price is Rs 20.

2. For the data given below **[10]**

Latitude	33.0	34.5	35.0	37.5	39.0
Mortality	219	160	170	182	149

The fitted regression line is: Mortality=389.2-5.96*Latitude

Can we assume that the intercept is 360 for this model. Test at 5% level of significance. Also obtain 95% confidence interval for intercept parameter.

3. Given the following fitted multiple regression model **[10]**

$$y = 14.98 - 0.20 * x_1 + 0.018 * x_2$$

The actual value and predicted values of heart disease for a sample of size 6 are given as follows:

Actual	11.65	3.85	17.17	6.82	5.06	9.56
Pred.	13.81	1.61	14.99	8.82	5.31	10.20

The sample variance of x_1 is 572.0908. The values of *R-Squared* for x_1 is 0.3. Now, can we assume $\beta_1 = -0.25$. Test at 5% level of significance. Also obtain 95% confidence interval for β_1.

4. (a) Give your interpretation of regression coefficients for the following fitted multiple linear regression line $\hat{y} = -10.867 + 5.128x_1 - 1.33x_2$ **[3]**

(b) Given the following data along with the predicted values of response variable (Y) **[7]**

Y	350	460	350	430	350
X_1	5.5	7.5	8.0	8.0	6.8
X_2	3.3	3.3	3.0	4.5	3.0
Pred. Y	413.79	363.84	329.12	440.31	359.08

Now compute the followings:

(i) R^2 and write its interpretation

(ii) Adjusted R^2

(iii) $\hat{\sigma}^2$

5. If $y = \begin{pmatrix} 3 \\ 2 \\ 4 \\ 5 \end{pmatrix}$, Design matrix: $X = \begin{pmatrix} 1 & 2 & 1 \\ 1 & 3 & 5 \\ 1 & 5 & 3 \\ 1 & 7 & 6 \end{pmatrix}$ [10]

and $(XX)^{-1}X' = \begin{pmatrix} 1.02 & 0.39 & 0.17 & -0.57 \\ -0.04 & -0.27 & 0.16 & 0.15 \\ -0.16 & 0.28 & -0.17 & 0.04 \end{pmatrix}$ then

Calculate the followings

(i) HAT matrix

(ii) Fitted values

(iii) Residuals

(iv) Approximated average variance of residuals

6. Discuss Four graphic patterns of Normal Probability Plot along with their interpretations. [10]

7. Examine the linear relationship between y and x in the following data. If it is not linear then apply appropriate transformation to linearize the model. Draw all the required graphs. [10]

y	1038.05	169.20	421.87	634.19	369.29	554.19	934.86	293.15
x	4.66	3.26	3.97	4.28	3.86	4.18	4.58	3.69

8. (a) Write the procedure of Variance Stabilizing Transformations. [5]

(b) Recall the overview of Box-Cox Method of transformation. [5]

9. Discuss the consequences (effects) of multicollinearity. [10]

10. Explain generalized linear model through the meaning of exponential family of distribution. Write the mean and variance of $a(X)$. Convert the probability functions of normal distribution and binomial distribution in the form of exponential family of distribution and determine $a(X)$ and $b(\theta)$ in both distributions. [10]

11. (a) Recall residual analysis in generalized linear model. [7]

(b) Write the canonical links of Normal, Binomial and Poisson distributions. [3]

12. (a) Show that at least in a non-linear regression model, at least one of the derivatives of the expectation function with respect to the parameters depends at least one of the parameters. [6]

(b) Write the normal equations for the following non linear regression models. [4]

(i) $y = \theta_1 e^{\theta_2 x} + \varepsilon$ (ii) $y = \theta_1 + \frac{\theta_2}{x - \theta_3} + \varepsilon$

⇔⇔⇔ J/E/TX⇔⇔⇔

O/E/TX

Reg. No: 23MBT0005

VIT® Vellore Institute of Technology

Final Assessment Test – May 2024

Course: MAT6005 - Machine Learning for Data Science

Class NBR(s): 1476

Time: Three Hours

Slot: E1+TE1

Max. Marks: 100

- KEEPING MOBILE PHONE/ELECTRONIC DEVICES EVEN IN 'OFF' POSITION IS TREATED AS EXAM MALPRACTICE
- DON'T WRITE ANYTHING ON THE QUESTION PAPER

Answer any TEN Questions

(10 X 10 = 100 Marks)

1. What is machine learning? Describe the various types of machine learning using appropriate examples. **[10]**

2. The following data show the number of bedrooms, the number of baths, and the prices at which a random sample of eight one-family houses sold in a certain large housing development: **[10]**

Price (dollars), y	Number of bedrooms, x_1	Number of baths, x_2
292,000	3	2
264,600	2	1
317,500	4	3
265,500	2	1
302,000	3	2
275,500	2	2
333,000	5	3
307,500	4	2

Design a regression network to predict the sales price of a three-bedroom house with two baths in the subject housing development.

3. Define support and confidence in Association rule mining. Trace the results of using the Apriori algorithm on the grocery store example with support $S = 50\%$ and confidence $C = 75\%$. Find all the final frequent itemsets. Also indicate the association rules that are generated and highlight the strong ones, sort them by confidence. **[10]**

Transaction ID	Items
T1	Bread, Cheese, Egg, Juice
T2	Bread, Cheese, Juice
T3	Bread, Milk, Yogurt
T4	Bread, Juice, Milk
T5	Cheese, Juice, Milk

4. A company manufactures light bulbs, and they are interested in estimating the average lifespan (in hours) of their new bulb design. They randomly sample 10 bulbs and record their lifespans (in hours) as follows: [10]

{1000, 950, 1100, 780, 970, 880, 750, 1005, 900, 850}

They believe the lifespans follow an exponential distribution with an unknown parameter λ (lambda), which determines the rate of decay. Find the Maximum Likelihood Estimate of λ for this data set.

5. Consider a hierarchical Bayesian model where the prior distribution of a parameter θ follows a Beta distribution with parameters $\alpha = 2$ and $\beta = 2$, and the likelihood function $P(X|\theta)$ follows a Binomial distribution with $n = 20$ and $p = \theta$. If we observe $X = 12$successes, compute the posterior distribution of θ and discuss how the hierarchical structure influences the estimation. [10]

6. Given a dataset of observations: [3,5,2,6,4,7,3,5,8,2,4,6,3,5,4], estimate the probability density function using kernel density estimation using Epanechnikov kernel function and an appropriate bandwidth. Plot the probability density function. [10]

7. If $X \sim N_3(\mu, \Sigma)$ where $\mu = \begin{bmatrix}5\\8\\7\end{bmatrix}$, $\Sigma = \begin{bmatrix}4 & -1 & 0\\-1 & 4 & 2\\0 & 2 & 9\end{bmatrix}$ and $X = \begin{bmatrix}X_1\\X_2\\X_3\end{bmatrix}$, find [10]

   a) $P(X_1 > 6)$
   b) $P(5X_2 + 4X_3 > 70)$
   c) $P(4X_1 - 3X_2 + 5X_3 < 80)$.

8. Find the missing data for the following dataset: [10]

x:	62	64	65	69	?	71	72	74
y:	126	125	139	145	165	152	180	208

9. Describe the concept of Singular Value Decomposition (SVD) and its significance in machine learning. Given a matrix $A = \begin{bmatrix}3 & 0\\4 & 5\end{bmatrix}$, compute its Singular Value Decomposition $A = U\sum V^T$. [10]

10. Explain the significance of eigenvalues and eigenvectors in PCA. Use Principal Component Analysis to reduce the dimension for the following dataset: [10]

x_1	3	7	11	8
x_2	12	5	5	13

11. What do you mean by Gradient Descent? Determine a minimum of $f(x) = (x-3)^2$, starting from $x_0 = 0$ and applying the method of gradient descent with appropriate values of η. [10]

12. Elaborate on the strengths and weaknesses of k-means clustering, highlighting scenarios where it performs well. Cluster the following eight points into three clusters: [10]

A1(2, 10), A2(2, 5), A3(8, 4), A4(5, 8), A5(7, 5), A6(6, 4), A7(1, 2), A8(4, 9), using k-means algorithm. Initial cluster centers are: A1(2, 10), A4(5, 8) and A7(1, 2).

⇔⇔⇔ O/E/TX⇔⇔⇔

T/E/TX

Reg. No: 23MDT0005

VIT®
Vellore Institute of Technology

Final Assessment Test – May 2024

Course: **CSE5003 - Database Systems: Design and Implementation**
Class NBR(s): **5062** Slot: **F2**
Time: **Three Hours** Max. Marks: **100**

- KEEPING MOBILE PHONE/ELECTRONIC DEVICES EVEN IN 'OFF' POSITION IS TREATED AS EXAM MALPRACTICE
- DON'T WRITE ANYTHING ON THE QUESTION PAPER

Answer any TEN Questions
(10 X 10 = 100 Marks)

1. Design a complete ER diagram by considering the following schema for a Library Database.

   BOOK (Book_id, Title, Publisher_Name, Pub_Year)
   BOOK_AUTHORS (Book_id, Author_Name)
   PUBLISHER (Name, Address, Phone)
   BOOK_COPIES (Book_id, Branch_id, No-of_Copies)
   BOOK_LENDING (Book_id, Branch_id, Card_No, Date_Out, Due_Date)
   LIBRARY_BRANCH (Branch_id, Branch_Name, Address)

2. a) With neat diagram explain the three-schema architecture. **[6+4]**
   b) Explain the concept of data independence in DBMS.

3. Justify your answer for the following questions with appropriate examples.
   i) Is it mandatory for a foreign key to be a primary key in another table?
   ii) Can foreign key values be null?
   iii) After enforcing foreign key constraint using on delete cascade rule, is it possible to delete a record in the child relation whose foreign key value is Not NULL before deleting the corresponding primary key record in the parent relation?
   iv) Can foreign keys have duplicate values?
   v) Can there exist more than one foreign key in a relation?

4. Define Normalization. Assume a relation which is in un-normalized form (UNF) and convert the relation to 1 NF and 2NF with explanation. **[2+4+4]**

5. a) Consider the following relation and functional dependencies. **[2+4+4]**

   R(A,B,C,D,E,F),

   F= { A→BC, AC→DEF, F→AB}

   I. List all of the candidate keys
   II. Give the minimal cover for F.

b) Given a relational schema R(A, B, C, D, E) and a set of functional dependencies P and Q such that: P = {A→B, AB→C, D→AC, D→E}, Q = {A→BC, D→AE}. Check whether P and Q are equivalent?

6. a) Explain shared-nothing architecture with a neat diagram. [4+6]

b) Given that there are 10 disks and 100 tuples that must be assigned to these 10 disks. (i) Using the round-robin partitioning, determine on which disk will the 57th tuple be placed. (ii) For hash partitioning, tuple j is assigned to the disk h (j). Let the hash function h be given as j % 10, determine on which disk will the 57th tuple be placed?

7. a) Mention the conditions to check a schedule is conflict serializable. [4+6]

b) Is the schedule given below conflict serializable? Justify your answer with proper explanation.

T1	T2
R (A)	
W (A)	
	R (A)
	W (A)
R (B)	
W (B)	
	R (B)
	W (B)

8. Consider the Employee schema and the department schema given below: [4+6]

EmpId	FName	LName	DeptId	BDate	Address	Gender	Salary

DeptNo	DeptName	DeptAddress

The following statements are true:

- Employee is stored on site 1 and Department is stored on a different site 2
- There are 10000 records for Employee and each record is 100 bytes long.
- There are 100 records of Department, and each record is 35 bytes long.
- Each employee is associated with exactly one department and the size of the query result is 40 bytes long.

Consider the query:

For each employee, retrieve the FName and LName of the employee and the department name to which the employee is associated.

i. Write the SQL query and relational algebra notation for the above query.

ii. Suppose that the query is submitted to a distinct site called result site which is a different site 3, elucidate the possible strategies for generating the join results at site 3 and their respective costs calculated as the amount of data transfer required, and specify the strategy that minimizes this cost.

9. Write the SQL Query and relational algebric notation to - **retrieve the name and address of every employee who works for the 'Research' department** and Perform the corresponding query tree representation using stepwise transformations. **[3+3+4]**

10. What is an anomaly? Consider the following relational schema. **[1+9]**

EMP_DEPT

Ename	Ssn	Bdate	Address	Dnumber	Dname	Dmgr_ssn

Explain insertion anomaly, deletion anomaly and update anomaly that may be present in a state of the above table with an example.

11. List out the different Database Security Issues and its counter measures to deal with those problems. **[4+6]**

12. a) What is a spatial database? Explain the different types of spatial data types. **[5+5]**

b) Write a note on Spatial Indexing.

⇔⇔⇔ T/E/TX⇔⇔⇔

SAUMYA SINHA
23MDT0005

MAT6012 - Programming for Data Analysis

Python Lab Worksheet on Functions in Python

Q1) Fibonacci sequence is given by $0, 1, 1, 2, 3, 5, 8, \ldots$ The mathematical representation of this sequence is

$$f(n) = \begin{cases} 0, & \text{if } n = 0 \\ 1, & \text{if } n = 1 \\ f(n-1) + f(n-2), & \text{if } n > 1. \end{cases}$$

Write a Python recursive function `fib(n)`, which produces the n^{th} Fibonacci number. Using this function, create another function, `fib_upto_n(n)`, which produces the first n Fibonacci numbers and displays them to the user.

Q2) Collatz conjecture in Mathematics is a famous conjecture which asks whether repeating two simple arithmetic operations will eventually transform every positive integer into 1. It concerns a sequence of integers in which each term is obtained from the previous term as follows: if the previous term is even, the next term is one-half of the previous term. If the previous term is odd, the next term is 3 times the previous term plus 1. Mathematically, this is represented as follows:

$$f(n) = \begin{cases} \frac{n}{2}, & \text{if } n \text{ is even,} \\ 3n + 1, & \text{if } n \text{ is odd.} \end{cases}$$

Recursively,

$$f(n) = \begin{cases} \{n\}, & \text{if } n = 1, \\ \{n, f(n/2)\}, & \text{if } n \text{ is even,} \\ \{n, f(3n+1)\}, & \text{if } n \text{ is odd.} \end{cases}$$

1. Write a Python function script named `collatz_norecur(n)` to print the Collatz conjecture sequence for the given number n, along with the given number of steps taken to reach the end number 1.
2. Write a Python recursive function script named `collatz(n)` to print the Collatz conjecture sequence for the given number n.

Q3) Write a Python function script `multipyMatrices(matrix1, matrix2)`, which takes two matrices and returns the matrix multiplication of these two matrices. Later, extend this function to a function named `multiplyAnyMatrices(*matrices)`, which repeatedly finds the matrix multiplication of any number of matrices.

Remark: For matrix multiplication to work, you need to check the dimensions of the matrices. The first matrix's column dimension should equal the second matrix's row dimension. You can check this with the `arraySize(A)` function that you have worked out previously.

MAT6012 - Programming for Data Analysis

Python Lab Worksheet on Generator Functions

Q1) Prime factors of a composite number. Prime factorisation of a composite number N is defined as the set of prime numbers whose product is equal to the given number N. For example, for the number 1729, the prime factorisation is $7, 13, 19$. Similarly, for the number 1024, the prime factorisation is $2, 2, 2, 2, 2, 2, 2, 2, 2, 2$.

Procedure for finding the prime factors:

1. Divide the given number by the smallest prime number.
2. Again, divide the quotient by the smallest prime number.
3. Repeat the process until the quotient becomes 1.
4. Finally, multiply all the prime factors.

Write a Python generator function, `primeFactors(n)`, which returns prime factors of the given number as an iterable object. You can then use this generator function to display the given number's prime factors and their sum.

Q2) The Generator Version of `range`. The range function creates a sequence. For very large sequences, this consumes a lot of memory. You can write a version of the range which does not create the entire sequence but instead yields the individual values. Using a generator will have the same effect as iterating through a sequence but won't consume as much memory.

Define a generator function, `genrange`, which generates the same sequence of values as `range` without creating a `list` object.

Q3) Generate infinite primes. Create a generator function, `genprimes()`, which generates infinite prime numbers and then, using this function, generate the first 100 hundred primes and then generate the next 100 primes.

VIT

Vellore Institute of Technology

(Deemed to be University under section 3 of UGC Act, 1956)

Winter Semester 2023-2024

School Of Advanced Sciences

Department Of Mathematics

LAB Final Assessment Test (LAB-FAT)

Course Code & Name: MAT6002 & Regression Analysis and Predictive Models Lab Course
Slot: L59+L60 (D1+TD1) **Exam Duration: 60**
Set Code:LF02 **Max. Marks: 40**

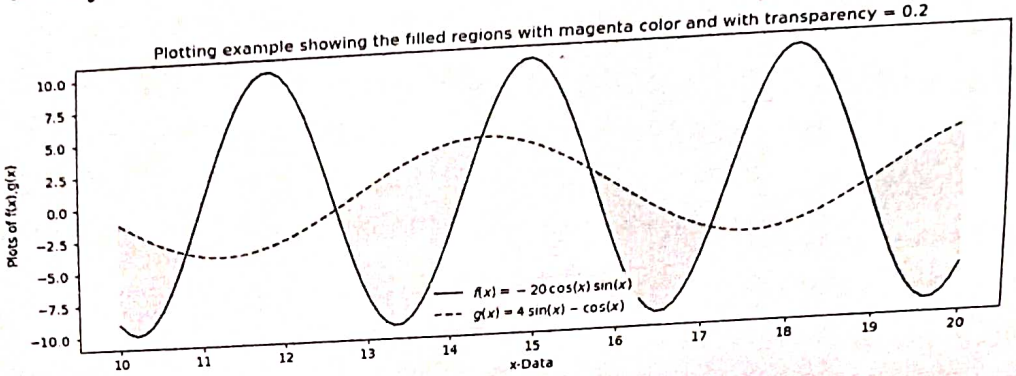
ANSWER ALL QUESTIONS

Sl. No.	Question	Marks
1	Given the following real life data. Firm 1: 199, 236, 167, 263, 254, 210, 225, 189 Firm 2: 108, 104, 153, 218, 210, 96, 117, 115 Firm 3: 162, 86, 160, 135, 207, 201, 90, 122 Now (i) Calculate the Pearson Correlation Coefficient between **Firm 2 and Firm 3.** (ii) Draw correlation matrix using heatmap based on all three variables. **(b)** Fit simple linear regression model for the following data. Give your interpretation for testing of hypothesis for regression parameters. Also obtain 95% confidence interval. Comment on the results in detail. Y: 73 82 72 75 88 73 74 81 86 83 X: 48 57 41 35 45 48 41 48 34 29	**20**
2.	**(a)** In the following sampled data **Detect the outliers** using any one of the methods (prefer Z-Score method) and **apply Mean imputation treatment** for the outliers. Finally **visualize the data after the application of treatment.** **Sample Data**: 10.2, 9.1, 8.5, 8.1, 1.2, 1.8, 80, 3.5, 11.5, 6.9, 99, 12.9, 11.0, 5.8, 8.5, 14.6 **(b)** Fit quadratic and exponential models for the following data. X = np.arange(-8.0, 8.0, 0.3)	**20**

PART-B (Duration 80 Minutes)
Answer ALL Questions (40 Marks)

marks) 1. Write a function block which accepts a number (with at least three digits) and the function will return `True` if the sum of the digits of the number is divisible by 8 and `False` if otherwise. Later use this function to find all the numbers between 2000 and 8000, who satisfy the property that the sum of the digits of the number is divisible by 8 and return the sum and product of these numbers.

(12 marks) 2. Write the Python code to generate the following figure.



(16 marks) 3. From the data file 'WQ_FAT.xlsx' read the 'Aug 04' sheet into the DataFrame named 'WQAugData'. Perform the following operations:

(a) Describe the 'WQAugData' DataFrame statistics and store them in an excel file named 'WQAugStatistics.xlsx'.

(b) Identify the columns in which there are null values, columns in which there are no null values and list both of them separately. Fill the null values by forward fill and then the remaining null values by the value 0.

(c) Append a column named `DIP` containing the values `TP - PO4-P` to the DataFrame.

(d) Draw a scatter plot between `NH4-N` and `DO` plot and interpret the plot.

(e) Add a row with 'index=Minimum' to the DataFrame which contains the minimum of each variable.

(f) Plot the each variable with (i) bar plot (ii) line plot and (ii) area plot (iv) box plot. (Use subplots wherever necessary)