GANDHI INSTITUTE OF TECHNOLOGY AND MANAGEMENT (GITAM)

(Deemed to be University)

VISAKHAPATNAM * HYDERABAD * BENGALURU

Accredited by NAAC with A^{++} Grade

GITAM School of Science

CURRICULUM AND SYLLABUS

2 Year Postgraduate Programme
PCSCI02: M.Sc. Data Science

w.e.f. 2023-24 admitted batch

(Updated on 31st July 2023)

Master of Science in Data Science (M.Sc Data Science)
REGULATIONS
(w.e.f. 2023-2024 batch)

1. ADMISSION

1.1 Admission into M.Sc. Data Science program of GITAM University is governed by GITAM University admission regulations.

2. ELIGIBILITY CRITERIA

2.1. A pass in BCA or B.Sc. degree with a minimum aggregate of 50% marks / a pass in anydegree with a minimum aggregate of 50% marks along with Mathematics or Statistics or Computer science as one of the subjects.

2.2. Admission into M.Sc. Data Science (Master of Science in Data Science) will be based on All India GITAM Science Admission Test (GSAT) conducted by GITAM University and the rule of reservation, wherever applicable.

3. CHOICE BASED CREDIT SYSTEM

Choice Based Credit System (CBCS) is introduced with effect from the admitted Batch of 2015-16 based on UGC guidelines in order to promote:

- Student Centered Learning
- Cafeteria approach
- Inter-disciplinary learning

Learning goals/objectives and outcomes are specified leading to what a student should be able to do at the end of the program.

4. STRUCTURE OF THE PROGRAM

4.1 The Program Consists of

i) Foundation Courses (compulsory) which gives general exposure to a Student in communication and subject related areas.

ii) Core Courses(compulsory).

iii) Discipline centric electives which

a) are supportive to the discipline

b) give an expanded scope of the subject

c) give their disciplinary exposure

d) nurture the student skills

iv) Open electives are of general nature either related or unrelated to the discipline.

v) Practical Proficiency Courses, Laboratory and Project work.

4.2 Each course is assigned a certain number of credits depending upon the number of contact hours (lectures/tutorials/practical) per week.

4.3 In general, credits are assigned to the courses based on the following contact hours per week per semester.

- One credit for each Lecture / Tutorial hour per week.
- One credit for two hours of Practical per week.
- Eight credits for project.

4.4 The curriculum of the Four semesters M.Sc. Data Science program is designed to have a total of 80 credits for the award of M.Sc. Data Science degree.

5. MEDIUM OF INSTRUCTION

The medium of instruction (including examinations and project reports) shall be in English.

6. REGISTRATION

Every student has to register himself / herself for each semester individually at the time specified by the Institute / University.

7. ATTENDANCE REQUIREMENTS

7.1. A student whose attendance is less than 75% in all the courses put together in any semester will not be permitted to attend that end - semester examination and he/she will not be allowed to register for the subsequent semester of study. He/she has to repeat the semester along with his / her juniors.

7.2. However, Vice-Chancellor on the recommendation of the Principal / Director of the Institute/School may condone the shortage of attendance to the students whose attendance is between 66% and 74% on genuine grounds and on payment of prescribed fee.

8. EVALUATION

8.1. The assessment of the student's performance in a Theory course shall be based on two components: Continuous Evaluation (40 marks) and Semester-end examination (60 marks).

8.2. A student has to secure an aggregate of 40% in the course in continuous and semester-end examinations the two components put together to be declared to have passed the course, subject to the condition that the candidate must have secured a minimum of 24 marks (i.e. 40%) in the theory component at the semester-end examination.

8.3. Practical / Viva voce etc. courses are completely assessed under Continuous Evaluation for a maximum of 100 marks and a student has to obtain a minimum of 40% to secure Pass Grade. Details of the Assessment Procedure are furnished below in Table1.

Table 1: Assessment Procedure

S. No.	Component of assessment	Marks allotted	Type of Assessment	Scheme of Examination
1	Theory	40	Continuous evaluation	(i) Three mid semester examinations shall be conducted for 15 marks each. The performance in best two shall be taken into consideration. (ii) 5 marks are allocated for quiz. (iii) 5 marks are allocated for assignments.
		60	Semester-end examination	The semester-end examination shall be for a maximum of 60 marks.
	Total	100		
2	Practical	100	Continuous evaluation	60 marks for performance, regularity, record and case study. Weightage for each component shall be announced at the beginning of the semester. 40 marks (30 marks for experiment(s) and 10 marks for practical Viva-voce.) for the test conducted at the end of the semester conducted by the concerned lab Teacher.
	Total	100		
3	Project work	200	Project evaluation	150 marks for evaluation of the project work dissertation submitted by the candidate. 50 marks are allocated for the project Viva-Voce. The project work evaluation and the Viva-Voce shall be conducted by one external examiner outside the University and the internal examiner appointed by the Head of the Department.

9. RE-TOTALING & REVALUATION

9.1 Re-totaling of the theory answer script of the semester-end examination is permitted on request by the student by paying the prescribed fee within one week after the announcement of the results.

9.2 Revaluation of the theory answer scripts of the semester-end examination is permitted on request by the student by paying the prescribed fee within one week after the announcementoftheresult.

10. PROVISION FOR ANSWER BOOK VERIFICATION & CHALLENGE EVALUATION:

10.1 If a student is not satisfied with his/her grade after revaluation, the student can apply for, answer book verification on payment of prescribed fee for each course within one week after announcement of revaluation results.

10.2 After verification, if a student is not satisfied with revaluation marks/grade awarded, he/she can apply for challenge valuation within one week after announcement of answer book verification result/ two weeks after the announcement of revaluation results, which will be valued by the two examiners i.e., one Internal and one External examiner in the presence of the student on payment of prescribed fee. The challenge valuation fee will be returned, if the student is succeeded in the appeal with a change for a better grade.

11. SUPPLEMENTARY EXAMINATIONS & SPECIAL EXAMINATIONS:

11.1 The odd semester supplementary examinations will be conducted on daily basis after conducting regular even semester examinations in April/May.

11.2 The even semester supplementary examinations will be conducted on daily basis after conducting regular odd semester examinations during November/December

11.3 A student who has completed his/her period of study and still has “F” grade in final semester courses is eligible to appear for Special Examination normally held during summer vacation.

12. PROMOTION TO THE NEXT YEAR OF STUDY

12.1 A student shall be promoted to the next academic year only if he/she completes the academic requirements of 60% of the credits till the previous academic year.

12.2 Whenever there is a change in syllabus or curriculum he/she has to continue the course with new regulations after detention as per the equivalency established by the Board of Studies (BoS) to continue his/her further studies.

13. BETTERMENT OF GRADES

13.1 A student who has secured only a pass or second class and desires to improve his/her class can appear for betterment examinations only in ‘n’ (where ‘n’ is no.of semesters of the program) theory courses of any semester of his/her choice, conducted in summer vacation along with the Special Examinations.

13.2 Betterment of Grades is permitted ‘only once’, immediately after completion of the program of study.

14. REPEAT CONTINUOUS EVALUATION

14.1 A student who has secured ‘F’ grade in a theory course shall have to reappear at the subsequent examination held in that course. A student who has secured ‘F’ grade can improve continuous evaluation marks upto a maximum of 50% by attending special instruction classes held during summer.

14.2 A student who has secured ‘F’ grade in a practical course shall have to attend special Instruction classes held during summer.

14.3 A student who has secured ‘F’ grade in a combined (theory and practical) course shall have to reappear for theory component at the subsequent examination held in that course. A student who has secured ‘F’ grade can improve continuous evaluation marks up to a maximum of 50% by attending special instruction classes held during summer.

14.4 The RCE will be conducted during summer vacation for both odd and even semester students. Student can register a maximum of 4 courses. Biometric attendance of these RCE classes has to be maintained. The maximum marks in RCE be limited to 50% of Continuous Evaluation marks. The RCE marks are considered for the examination held after RCE except for final semester students.

14.5 RCE for the students who completed course work can be conducted during the academic semester. The student can register a maximum of 4 courses at a time in slot of 4 weeks. Additional 4 courses can be registered in the next slot.

14.6 A student is allowed to Special Instruction Classes (RCE) ‘only once’ per course.

15. GRADING SYSTEM

15.1 Based on the student performance during a given semester, a final letter grade will be awarded at the end of the semester in each course. The letter grades and the corresponding grade points are as given in Table 2.

Table 2: Grades & Grade Points

Sl.No.	Grade	Grade Points	Absolute Marks
1	O (outstanding)	10	90 and above
2	A+ (Excellent)	9	80 to 89
3	A (Very Good)	8	70 to 79
4	B+ (Good)	7	60 to 69
5	B (Above Average)	6	50 to 59
6	C (Average)	5	45 to 49
7	P (Pass)	4	40 to 44
8	F (Fail)	0	Less than 40
9	Ab. (Absent)	0	-

15.2 A student who earns a minimum of 4 grade points (P grade) in a course is declared to have successfully completed the course, subject to securing an average GPA (average of all GPAs in all the semesters) of 5 at the end of the Program to declare pass in the program.

Candidates who could not secure an average GPA of 5 at the end of the program shall be permitted to reappear for a course(s) of their choice to secure the same.

16. GRADE POINT AVERAGE

16.1 A Grade Point Average (GPA) for the semester will be calculated according to the below formula :

.

Where C = number of credits for the course,

G = grade points obtained by the student in the course.

To arrive at Cumulative Grade Point Average (CGPA), a similar formula is used considering the student’s performance in all the courses taken, in all the semesters up to the particular point of time.

16.2 CGPA required for classification of class after the successful completion of the program is shown in Table 3.

Table 3: CGPA required for award of Class

Class	CGPA Required
First Class with Distinction	≥ 8.0*
First Class	≥ 6.5
Second Class	≥ 5.5
Pass Class	≥ 5.0

* In addition to the required CGPA of 8.0 or more the student must have necessarily passed all the courses of every semester in first attempt.

17. ELIGIBILITY FOR AWARD OF THE MSc Data Science DEGREE

17.1 Duration of the program: A student is ordinarily expected to complete M.Sc. program in four semesters of two years. However a student may complete the program in not more than four years including study period.

17.2 However the above regulation may be relaxed by the Vice-Chancellor in individual cases for cogent and sufficient reasons.

17.3 A student shall be eligible for award of the M.Sc. Degree if he / she fulfills all the following conditions.

a) Registered and successfully completed all the courses and projects.

b) Successfully acquired the minimum required credits as specified in the curriculum corresponding to the branch of his/her study within the stipulated time.

c) Has no dues to the Institute, hostels, Libraries, NCC / NSS etc, and

d) No disciplinary action is pending against him / her.

17.4 The degree shall be awarded after approval by the Academic Council.

18. DISCRETIONARY POWER

Notwithstanding anything contained in the above sections, the Vice-Chancellor may review all exceptional cases, and give his decision, which will be final and binding.

PROGRAMME EDUCATIONAL OBJECTIVES (PEOs)

PEO1: Graduate shall have successful professional career in data science and allied fields with in-depth knowledge and practical/interpersonal skills.

PEO2: Graduates will have the ability to apply knowledge across the disciplines likecomputerscience, optimization, and statistics to handle the realistic problems.

PEO3: Graduates will demonstrate skill in Data management and will demonstrate proficiency in statistical analysis of data.

PROGRAM OBJECTIVES(POS)

1. Ability to apply knowledge of mathematics, probability and statistics, computer science and solve problems.
2. An ability to analyze very large data sets in the context of real-world problems and interpret results using data analytics
3. The ability to apply the knowledge and generate actionable insights from data to solve problems that analysts and business users can translate into tangible business value.
4. Ability to model, analyze, design, visualize and realize physical systems or processes of increasing size and complexity.
5. Ability to develop strategies for analyzing data and visualizing the data by using complex machine learning algorithms to build predictive models for a wide range of application domains

PROGRAM OUTCOMES (POS)

1. **Engineering knowledge**: Apply the knowledge of mathematics, science, engineering fundamentals, and an engineering specialization to the solution of complex engineering problems.
2. **Problem analysis**: To Identify, formulate, research literature, and analyze complex engineering problems reaching substantiated conclusions using first principles of mathematics, natural sciences, and engineering sciences.
3. **Design/development of solutions**: Design solutions for complex engineering problems and design system components or processes that meet the specified needs with appropriate consideration for public health and safety, and the cultural, societal, andenvironmental considerations.
4. **Conduct investigations of complex problems**: Use research-based knowledge and research methods including design of experiments, analysis, and interpretation of data, and synthesis of the information to provide valid conclusions.
5. **Modern tool usage**: Create, select, and apply appropriate techniques, resources, and modern engineering and IT tools including prediction and modeling to complex engineering activities with an understanding of the limitations.
6. **The engineer and society**: Apply reasoning of the contextual knowledge to assess societal, health, safety, legal and cultural issues and the consequent responsibilities relevantto the professional engineering practice.
7. **Environment and sustainability**: Understand the impact of the professional engineering solutions in societal and environmental contexts, and demonstrate the knowledge of, and need for sustainable development.

8. **Ethics**: Apply ethical principles and commit to professional ethics and responsibilities and norms of the engineering practice.
9. **Individual and teamwork**: Function effectively as an individual, and as a member or leaderin diverse teams, and in multidisciplinary settings.
10. **Communication**: Communicate effectively on complex engineering activities with the engineering community and with society at large, such as, being able to comprehend andwrite effective reports and design documentation, make effective presentations, and give and receive clear instructions.
11. **Project management and finance**: Demonstrate knowledge and understanding of the engineering and management principles and apply these to one's own work, as a memberand leader in a team, to manage projects and in multidisciplinary environments.
12. **Life-long learning:** Recognize the need for, and have the preparation and ability to engagein independent and life-long learning in the broadest context of technological change.

PROGRAM SPECIFIC OUTCOMES (PSOS)

PSO1: To equip students with a sound knowledge base of programming language skills and an overall understanding of the domain in order to meet the requirements of the industry.

PSO2: To provide a concrete foundation in Mathematics, Artificial Intelligence, and Data Analytic techniques that can lead to research in the specialized fields as well as Employability.

PSO3: To acquire proficiency in statistical analysis of data, build and assess data models so as to be competent enough to meet global requirements as software professionals.

PSO4: To prepare graduates so as to make them lifelong learners through continuous Professional development.

M. Sc Data Science - Scheme of Instruction

I SEMESTER

S. No	Course Code	Course Title	Category	L	T	P	C	Remarks
1	CSCI6001	Introduction to Modern Databases	PC	4	0	0	4	
2	CSCI6011	Python Programming and Data Visualization	PC	4	0	0	4	
3	CSCI6021	Introduction to Data Structures	PC	3	0	0	3	
4	CSCI6031	Data Mining	PC	4	0	0	4	
5	MATH6001	Statistics for Data Science		4	0	0	4	
6	VDC111	Venture Development		2	0	0	2	
7	CSCI6041	Modern Databases Lab	PP	0	0	2	2	
8	CSCI6051	Python Programming Lab	PP	0	0	2	2	
				21	0	4	25	

II SEMESTER

S. No	Course Code	Course Title	Category	L	T	P	C	Remarks
1	LANG6181	Professional Communication Skills		2	0	0	2	
2	CSCI6061	Machine Learning	PC	4	0	0	4	
3	MATH6011	Mathematics for Data Science		4	0	0	4	
4	MATH6021	Inferential Statistics	PC	4	0	0	4	
5	CSCI6071	Artificial Intelligence	PC	4	0	0	4	
6	CSCI6081	Introduction to Object-Oriented Software Engineering		3	0	0	3	
7	MATH6031	Inferential Statistics Lab	PP	0	0	2	2	
8	CSCI6091	Machine Learning Lab using Python	PP	0	0	2	2	
				22	0	4	25	

III SEMESTER

S. No	Course Code	Course Title	Category	L	T	P	C	Remarks
1	CSCI7001	Deep Learning	PC	4	0	0	4	
2	CSCI7011	Big Data Analytics	PC	4	0	0	4	
3		Generic Elective – I	GE	4	0	0	4	
	CSCI7021	Cloud Computing						
	MATH7001	Applied Multivariate Statistical Analysis						
	CSCI7031	Computational Biology						
	CSCI7041	Web Programming						
	CSCI7051	Data Security and Privacy						
4		Generic Elective – II	GE	4	0	0	4	
	MATH7011	Time Series Analysis and Forecasting						
	CSCI7061	Data Storage Technologies and Networking						
	CSCI7071	Natural Language Processing						
	CSCI7081	Fundamentals of Blockchain Technologies						
	CSCI7091	Web Analytics						
5	CSCI7101	Deep Learning Lab	PP	0	0	2	2	
6	CSCI7111	Big Data Analytics Lab	PP	0	0	2	2	
7	CSCI7121	Industrial Training & Seminar	PP	0	0	2	2	
				16	0	6	22	

IV SEMESTER

S. No	Course Code	Course Title	Category	L	T	P	C	Remarks
1	CSCI7131	Project Work	PP	0	0	3	8	
TOTAL CREDITS= 25+25+22+8=80								

M.Sc DATA SCIENCE

SEMESTER- I

CSCI6001 INTRODUCTION TO MODERN DATABASES

Hours per week: 4End Examination: 60Marks

Credits:4Sessionals: 40 Marks

Preamble:

This course provides fundamental and practical knowledge on database concepts by means of organizing the information, storing and retrieving the information in an efficient and flexible way when data is stored in a well-structured model. This course ensures that every student will gain experience in creating data models and database design.

Course Objectives:

- Demonstration of basic database concept and construction of simple and moderately advanced database queries using Structured Query Language
- Explain and successfully apply logical database design principles, including E-R diagrams and database normalization.
- Demonstrate the concept of a database transaction, indexing, hashing.
- Explain different types of database architecture, the concept of parallel and distributed databases.
- Learn NoSQL features and Compare types of NOSQL Databases.

UNIT-I

Introduction to Relational Database Model: Structure of Relational Database**,** Database Schema, Keys, Schema Diagrams, Relational Query Languages, Relational Operations.

Introduction to SQL: Overview of the SQL query Language**,** SQL data definition, Basic structure of SQL queries**,** Additional basic operations**,** Set operations**,** Null values**,** Aggregate functions, Nested sub queries**,** Modification of the database.

Intermediate SQL: Join expressions, Views, Transactions, Integrity Constraints, SQL data types and schemas, Authorization. (8hours)

Learning Outcomes:

After completion of this unit, student will be able to

- Interpret the basic terminology of DBMS like data, database, database management systems. (L2)
- Compare DBMS over File Systems. (L2)
- Create and modify database using SQL query and apply integrity constraints. (L5)
- Illustrate different types of query forms (simple queries, nested queries, and aggregated queries) in SQL. (L2)
- Compare the difference between views and physical tables and working with views. (L2)

UNIT-II

Database design and the E-R Model: Overview of the Design Process**,** the Entity-Relationship model**,** Constraints**,** Removing Redundant Attributes in Entity sets**,** Entity-Relationship Diagrams, Reduction to relational schemas**,** Entity-Relationship design issues**,** Extended E-R features - Specialization and generalization.

Relational Database Design: Features of Good Relational Designs**,** Atomic Domains and First Normal Form, Decomposition using Functional Dependencies**,** Functional-Dependency Theory**,** Decomposition using Multi-Valued Dependencies**,** More Normal Forms**,** Database-Design Processand Modeling Temporal Data. (10hours)

Learning Outcomes:

After completion of this unit, student will be able to

- Model a given application using ER diagram. (L3)
- Match the integrity constraints from ER model to relational model. (L1)
- Translate an ER Model to a Relational Model and vice versa. (L2)
- Make use of the schema refinement process. (L3)
- Extend the concept of functional dependencies (FDs) and know about anomalies. (L2)
- Illustrates knowledge about different types of normal forms and the importance of normalization. (L2)

UNIT-III

Indexing and Hashing: Basic concepts**,** Ordered Indices, B+ tree index files**,** B+ tree extensions, Multiple-key access, Static Hashing**,** Dynamic Hashing**,** Comparison of ordered indexing and Hashing**,** Bitmap indices**,** Index definition in SQL.

Transactions: Transaction Concept, A Simple Transaction Model, Storage Structure, Transaction Atomicity and Durability**,** Transaction Isolation**,** Serializability**,** Transaction Isolation Levels, Implementation of Isolation Levels**,** Transactions as SQL Statements. (8 hours) **Learning**

Learning Outcomes:

After completion of this unit, student will be able to

- Illustrate the most important high-level file structure tools that include indexing, co-sequential processing, B trees, Hashing. (L2)
- Recognize the difference between various indexing techniques. (L3)
- Recognize the difference between various hashing techniques. (L3)
- Build indexing mechanisms for efficient retrieval of information from databases. (L5)
- Interpret the transaction management in DBMS. (L2)

UNIT-IV

Database System Architecture: Centralized and Client –Server Architectures**,** Server System Architectures**,** Parallel Systems**,** Distributed Systems**,** Network Types.

Parallel Databases: Introduction, I/O Parallelism**,** Inter Query Parallelism**,** Intra query Parallelism**,** Intra Operation Parallelism**,** Interoperation Parallelism**,** **Query** Optimization**,** Design of Parallel Systems**,** Parallelism on Multi Core Processor.

Distributed Databases: Homogeneous and Heterogeneous Databases, Distributed Data Storage, Distributed Transactions. (10 hours)

Learning Outcomes:

After completion of this unit, student will be able to

- Understand the techniques of parallel DBMSs, distributed DBMS architectures, distributed database design, query processing, multi database systems. (L1)
- Interpret the design principles in a distributed database for better resource management. (L2)
- Understand the types of distributed database systems. (L1)

UNIT-V

NoSQL- Need of NoSQL, ValueofRelationalDatabases,impedancemismatch,ApplicationandIntegrationDatabases,Attackofthеclusters,EmergenceofNoSQL.

NoSQL Data Architecture Patterns: NoSQL Data model: Aggregate Models- Document

Data Model-
Key-Value Data Model- Columnar Data Model, Graph Based Data Model Graph Data Model.
Distributed Models- Single Server, Sharding, Master-Slave Replication, Peer to Peer Replication.
Key-Value Databases: Comparison of Relational and Key Value Store, Features.
Document Databases: Comparison of Relational and Document Database, Advantages and Disadvantages, Features.
Column Family Stores- Comparison of Relational and Column Family, Advantages and Disadvantages, feature

Learning Outcomes:

After completion of this unit, students will be able to

- Learn various NoSQL systems and their features: (L1)
- Compare and use four types of NoSQL Databases (Document-oriented, Key-Value, Column-oriented and Graph based data models). (L2)
- Demonstrate and understand the detailed architecture, define objects, load data, query data and performance tune using all the four types of databases. (L2)

Course Outcomes:

Upon completion of the course student will be able to:

- DesignadatabaseforasystemusingE-R and relational datamodeland query usingRelationalDatamodel (L3)
- Design logical databases with all integrity constraints over relations(L3)
- Apply normalization steps in database design and removal of data anomalies(L3)
- Understand the variety of data access techniques and Transaction Processing, protocols used to assure ACID Properties. (L2)
- Understand a variety of parallelization techniques (L2)
- Distinguish different types of NoSQL databases. (L3)

Text book:

1. Database System Concepts by Abraham Silberschatz, Henry F.korth, S.Sudarshan, McGrawHill, Sixth Edition, 2011.(Unit – I–IV).
2. NoSQL Distilled by Pramod J Sadalage, Martin Fowler, Addison-Wesley Professional; 1st edition, 2012. (Unit –V)

Reference Book:

1. Fundamentals of Database Systems by Ramez Elmasri, Shamkant B.Navathe, Addison Wesley, Sixth Edition, 2011.

M.Sc DATA SCIENCE

SEMESTER- I

CSCI6011 PYTHON PROGRAMMING AND DATA VISUALIZATION

Hours per week: 4	End Examination: 60 Marks
Credits:4	Sessionals: 40 Marks

Preamble:

Python is an interpreter oriented, high-level, general-purpose programming language. Created by Guido van Rossum and first released in 1991. Python has a design philosophy that emphasizes code readability, notably using significant white space. It provides constructs that enable clear programming on both small and large scales.

Course Objectives:

- To learn the basic concepts and usage of variables, expressions and practice the use of functions in Python programming language.
- To identify and practice different conditionals and implement recursive functions.
- To understand the concepts of strings, lists and dictionaries, practice the use of classes methods, overloading and polymorphism.
- To learn the basic concepts of raw data and use different statistical methods on the data.
- To implement line properties, use different setter methods and practice different kinds of plots.

UNIT- I

The way of the program: Running Python, Arithmetic Operators, Values and types, Formal and Natural Languages, Debugging.

Variables, expressions and statements: Assignment statements, variable names, expressions and statements, script mode, order of operations, string operations.

Functions: Function calls, math functions, composition, adding new functions, definitions and uses, flow of execution, parameters and arguments. (8 hours)

Learning Outcomes:

By the end of the unit the student will be able to

- Explain different types of operators.(L2)
- Develop and run simple Python program.(L3)
- Describe the concepts of variables, expressions and statements.(L2)
- Use functions and develop programs using functions.(L3)
- Extend the concept of functions using parameters.(L2)

UNIT - II

Conditionals and Recursion: Floor division and modulus, Boolean expressions, logical operators, conditional execution, alternative execution, chained conditionals, nested conditionals, recursion, stack diagrams for recursive functions, infinite recursion.

Fruitful Functions: Return values, incremental development, composition, Boolean functions.

Iteration: Reassignment, updating variables, while statement, break, square roots. (9 hours)

Learning Outcomes:

By the end of the unit the student will be able to

- Use the logical operators in programming.(L3)
- Identify the need of recursive functions.(L1)
- Construct programs using while statements.(L3)
- Explain the use of break statement.(L2)

UNIT – III

Strings: String length, traversal with for loop, string slices, searching, looping and counting, string methods, in operator, string comparison.

Lists:traversing a list, list operations, list slices, list methods, map, filter and reduce, deleting elements, lists and strings, objects and values, aliasing, list arguments.

Dictionaries: looping and dictionaries, reverse lookup, dictionaries and lists, memos, global variables.

Tuples: Tuple assignment, Tuples as return values, Variable-length argument tuples, Lists and tuples, Dictionaries and tuples

Learning Outcomes:

By the end of the unit the student will be able to

- Construct programs to perform operations on strings.(L3)
- Explain basic concepts related to lists.(L2)
- Outline the concepts in dictionaries.(L2)
- Explain the basic concepts of tuples.(L2)

UNIT - IV

Files: Reading and writing, Format operator, Filenames and paths, Catching exceptions

Classes and objects: programmer defined types, attributes, rectangles, instances as return values.

Classes and methods: object oriented features, init method, str method, operator overloading, polymorphism. **Inheritance.**

Learning Outcomes:

By the end of the unit the student will be able to

- Develop simple programs using class.(L3)
- Apply operator overloading and polymorphism.(L3)
- Explain the basic analytics that can be applied on data.(L2)

UNIT - V

Getting Started with Raw Data: The world of arrays with NumPy, Empowering data analysis with pandas, Data cleansing, Data operations.

Plotting Data Using Matplotlib: Plotting using Matplotlib, Customization of Plots: Marker, Colour, Linewidth and Line Style, The Pandas Plot function (Pandas Visualization), Plotting a Line chart, Plotting Bar Chart, Plotting Histogram, Plotting Scatter Chart, Plotting Quartiles and Box plot, Plotting Pie Chart

Making Sense of Data through Advanced Visualization: Area plots, Bubble charts, Hexagon bin plots, Trellis plots, 3D plot of a surface.

Learning Outcomes:

By the end of the unit the student will be able to

- Use keyword arguments and setter methods.(L3)
- Identify the need of plots.(L1)• Explain basic concepts of charts.(L2)
- Illustrate the usage of visualization ToolEcosystem.(L3)

Course Outcomes:

Upon completion of this course, student will be able to:

- List the different types of operators. (L2)
- Understand the concept of variables, expressions, and statements. (L2)
- Understand the concept of functions and recursive functions. (L2)
- Identify the use of iteration. (L3)
- List the operations that can be performed on strings. (L2)

- Identify the differences between lists, dictionaries,tuples. (L3)
- Demonstrate the concepts of class, inheritance and polymorphism. (L2)
- Visualize different types of plots. (L3)

Text Books:

1. Mastering Python for Data Science by Samir Madhavan, PACKT Publishing,2015.
2. Think Python by Allen Downey O'Reilly Publications, 2nd Edition,2016.

Reference Books:

1. Programming Python by Mark Lutz, O'Reilly Publications, 4thEdition,2011.
2. Python in a nutshell by Alex Martelli, Anna Ravenscroft, Steve Holden, O'Reilly Publications, 3rd Edition, 2017.

M.Sc DATA SCIENCE
SEMESTER – I
CSCI6021 INTRODUCTION TO DATA STRUCTURES

Hours per week: 3Continuous Evaluation: 100 Marks
Credits:3

Preamble:

Data Structure is a way of collecting and organizing data in such a way that canperform operations on these data in an effective way. It is about rendering data elements interms of some relationship, for better organization and storage in different ways. This course will help in understanding various strategies require to solve a problem effectively and efficiently.

Course Objectives:

- To teach efficient storage mechanisms of data for an easy access.
- To design and implementation of various basic data structures.
- To introduce various techniques for the representation of the data in the real world.
- To develop applications using data structures.
- To improve the logical ability

UNIT I

Introduction to Data Structures- Data Structure and Algorithms- Introduction, Data Structures, Fundamentals of DS, Operations on Data Structure.

Arrays – Introduction, Memory/Storage Representation of One- and Two-Dimensional Array, Sorting- Definition of Sorting, Comparison of Sorting Method, Bubble Sort, Insertion Sort, Selection Sort

Searching – An Introduction, Linear or Sequential Search, Binary Search, Indexed SequentialSearch

Learning Outcomes:

By the end of this Unit, the student will be able to

- Illustrate the different operations performed on Data Structures (L3)
- Understand the concept of storage representation in arrays (L2)
- Explain the concepts of sorting (L2)
- Describe types of searching techniques (L2)

UNIT II

Stacks- Introduction & Definition, Applications of Stack, Various Representation of Stack,Operation on stack (Push and Pop), Hierarchy of Operation, Representation of ArithmeticExpression (Infix, Postfix, Prefix)

Learning Outcomes:

By the end of this Unit, the student will be able to

- Explain the concept of stack (L2)
- Understand the representation of stack (L2)
- Identify the different applications of stack (L1)
- Understand the concept of arithmetic expression (L2)

UNIT-III

Queues- Introduction, Applications of Queue, Various Representations of Queue, Operations on queue, Concept - Dequeue, Priority Queues, Circular Queue.

Learning Outcomes:

By the end of this Unit, the student will be able to

- Explain the concept of queue (L2)

- Identify the different applications of queue (L1)
- Illustrate the operations on queue (L3)
- Describe different types of queues (L2)

UNIT IV: Recursion and Linked List

Recursion- Introduction, Recursion Properties, Applications of Recursion (Factorial, Addition of TwoNumber, Power of A Number, Fibonacci Series, Quick Sort, Advantages and Disadvantages of Recursion.

Linked List- Introduction, Application of Linked List, and Representation of Linked List, operations on linked list.

Learning Outcomes:

By the end of this Unit, the student will be able to

- Explain the concept of recursion (L2)
- Identify the different applications of recursion (L1)
- Understand the concept of linked lists (L2)
- Illustrate different applications of linked lists (L3)

UNIT V: Trees

Trees- Introduction, Definition of Trees, Binary Tree, Types of Binary Tree, , Binary Search Tree (BST), Operations on Binary Search Tree,Traversal of Binary Search Tree, Expression Trees.

Learning Outcomes:

By the end of this Unit, the student will be able to

- Explain the concept of trees (L2)
- Understand different types of trees (L2)
- Explain the operations on trees (L2)
- Illustrate tree traversals (L3)

Course Outcomes:

Upon completion of the course, the students are expected to:

- Be able to choose appropriate data structure as applied to specified problem definition.
- Be able to handle operations like searching, insertion, deletion,traversing mechanism etc. on various data structures
- Be able to use linear and non-linear data structures like stacks, queuesand linked list,trees.

Text Books:

1. “Data Structures using C”, ISRD group, Second Edition, TMH, 2017
2. “Data Structures through C”, Third Edition, YashavantKanetkar, BPB Publications,2019.
3. “Data Structures Using C” , First Edition, Balagurusamy E. TMH,2017

Reference

1) LipschutzSchaum’s ”Data Structure” Outline Series [TMH],2014.ISBN-0-07-060168-2
2) D. Samanta and Debasis “Classic Data Structure”, Prentice Hall India, 2009. ISBN: 8120318749
3)Deshpande and Kakade, “C and Data Structure”, Dreamtech Press, 1st edition, 2003.

M.Sc DATA SCIENCE
SEMESTER – I
CSCI6031DATA MINING

Hours per week: 4	End Examination: 60Marks
Credits:4	Sessionals: 40 Marks

Preamble:

Due to the advent of technology, internet, and advanced applications like social media, a huge amount of digital data has been accumulated in data centers/cloud databases, which has led to a situation "we are drowning in data but starving from knowledge". To find golden nuggets which are useful for decision making processes, various data mining functionalities like association analysis, classification, clustering, outlier analysis and web mining are used. Data warehousing (DW) is an integral part of the knowledge discovery process, where DW is an integration of multiple heterogeneous data repositories under a unified schema at a single site. The students will acquire knowledge in data modeling, design, architecture, data warehouse implementation and further development of data cube technology.

Course Objectives:

- Understand the importance of Data Mining and its applications.
- Introduce various types of data and pre-processing techniques.
- Learn various multi-dimensional data models and OLAP Processing.
- Study concepts of Association Analysis.
- Learn various Classification methods.
- Learn the basics of cluster analysis.

UNIT – I

Introduction: Need for Data Mining, Definition of Data Mining, Kinds of data, Kinds of patterns to be mined, Technologies used, applications, Major issues in Data Mining.

Data Preprocessing: Need for Preprocessing the Data, Data Cleaning, Data Integration, Data Reduction, Data Transformation and Data Discretization. (10 hours)

Learning outcomes

After completion of this unit, student will be able to

- Understand basic concepts of data mining. (L2)
- Learn the KDD process.(L2)
- Learn different data mining tasks.(L2)
- Learn major challenges in the field of data mining.(L2)
- Understand various types of data sets and attributes.(L2)
- Apply different statistical techniques on different types of attributes to measure the similarities and dissimilarities.(L3)
- Learn different data preprocessing techniques and apply them on data sets.(L2)

UNIT – II

Data Warehouse and OLAP Technology: Data Warehouse – basic concepts, Data Cube and OLAP Technology, Design and Usage, implementation, Data Generalization by Attribute-Oriented Induction. (8hours)

- Learn the basics of data warehousing and different OLAP operations.(L2)
- Understand the relationship between data warehousing and other data generalization methods.(L2)
- Study the methods of data cube computation.(L2)

- Explorations of data cube and OLAP technologies.(L4)

UNIT – III

Mining Frequent Patterns, Associations and Correlations: Basic Concepts and Methods- Basic Concepts, Frequent item set Mining methods, Pattern Evaluation methods.

Advance Pattern Mining: Pattern mining in multilevel, multidimensional space, Constraint based Frequent Pattern Mining. (8hours)

Learning outcomes

After completion of this unit, student will be able to

- Understand the use of frequent patterns in business analysis.(L2)
- Implement Apriori algorithm and FP-growth algorithm.(L3)
- Learn different types of association rules.(L2)
- Identify the importance of each pattern evaluation method.(L3)
- Understand measures for mining correlated patterns.(L2)
- Learn Advanced Pattern Mining Methods.(L2)

UNIT –IV

Classification: Basic Concepts, Decision Tree induction, Bayes' Classification methods, Rule based Classification, Model Evaluation and selection, Techniques to improve classification accuracy, Support Vector Machines, Classification using Frequent patterns, Lazy Learners.

(10 hours)

Learning outcomes

After completion of this unit, student will be able to

- Understand basic concepts of classification.(L2)
- Implement the classification algorithms. (L3)
- Compare the performance of various classification algorithms.(L2)
- Understand model evaluation and selection methods.(L2)
- Identify the method that improves classification accuracy.(L3)

UNIT –V

Cluster Analysis: Definition, Requirements, Basic Clustering methods, Partitioning methods, Hierarchical methods, Density based methods, grid based methods, Evaluation of Clustering.

Outlier Detection: Outliers and Outlier Analysis, Detection methods, Statistical approaches, Proximity Based Approaches. (8 hours)

Learning outcomes:

After completion of this unit, student will be able to

- Understand the basic concepts of clustering. (L2)
- Implement the clustering algorithms. (L3)
- Compare the performance of various clustering algorithms. (L2)
- Learn various outlier detection methods. (L2)

Course Outcomes:

Upon completion of the course student will be able to:

- Understand the functionality of data mining components and data preprocessing techniques. (L2)
- Understand Data warehouse architecture and various OLAP operations. (L2)
- Identify patterns using various pattern mining algorithms on different datasets. (L3)

- Identify and apply appropriate data classification technique. (L3)
- Identify and apply appropriate data clustering technique. (L3)

Text Book:

1. Data Mining Concepts and Techniques by Jiawei Han, Michel Kamber, Elsevier,3rd Edition, 2012.

Reference Books

1. Introduction to Data Mining by Pang-Ning Tan & Michael Steinbach, VipinKumar, Pearson Publications, 1st edition, 2016.
2. Data Mining Techniques and Applications: An Introduction by Hongbo Du, CengageLearning EMEA, 1st edition,2010.
3. Data Mining : Introductory and Advanced topic by Dunham, Pearson Publications, 1st edition, 2006.

M.Sc DATA SCIENCE
SEMESTER- I
MATH6001 STATISTICS FOR DATA SCIENCE

Hours per week: 4	End Examination: 60 Marks
Credits:4	Sessionals: 40 Marks

Preamble: Statistical knowledge is essential to make sense of vast data and extract insights. Probability theory is important when it comes to evaluating statistics. This course treats the most common discrete and continuous distributions, showing how they find use in decision and estimation problems, and constructs computer algorithms for generating observations from the various distributions.

Course Objectives:

- To explain the diagrammatic and graphical representation of data. (L2)
- To interpret basic concepts necessary to derive descriptive statistics using of measures of central tendency and Measures of dispersion. (L2)
- To explain properties of probability and evaluate problems based on addition theorem, multiplication theorem and Bayes' theorem. (L2)
- To summarize different types of correlation for quantitative and qualitative data .(L2)
- To understand regression analysis and to distinguish between correlation and regression analysis. (L2)

UNIT – I

Descriptive Statistics - Data Classification; Tabulation – Frequency and graphic Representation – Measures of Central Tendency; Measures of dispersion; Skewness and Kurtosis; Coefficient of variation.

8 hours

Learning Outcomes :

By the end of this Unit, the student will be able to

- Understand the data representation (L2)
- Describe the aggregate measures of data based using different measures of central tendency. And dispersion (L3)
- Describe the measures of shape of data (L3)

- Evaluate consistency and reliability of data (L5)

UNIT – II

Probability Theory and Random variables - Definition and concepts of probability; Laws of probability – addition and multiplication theorem Bayes Theorem and its applications; Discrete and Continuous Random Variable – probability function and distribution function of random variables, Properties of random variables, Central Limit Theorem and its applications; Law of large numbers (LLN) and problems based on LLN.

8 hours

Learning Outcomes :

By the end of this Unit, the student will be able to

- Define probability. (L1)
- Evaluate problems on addition theorem, multiplication theorem and Bayes' theorem. (L5)
- Understand Random Variable. (L2)
- Infer Central Limit Theorem . (L2)

UNIT – III

Discrete probability distributions : Properties and application of Binomial distribution, Poisson distribution, Negative binomial distribution to solve real world business problems. Use of R program and MS Excel to carry out data analysis.10 hours

Learning Outcomes :

By the end of this Unit, the student will be able to

- Explain Binomial and Poisson distributions. (L2)
- Utilize Binomial and Poisson distributions in data analysis. (L3)

UNIT - IV

Probability distributions and their properties - Continuous Probability Distributions - Rectangular Distribution, Uniform Distribution - Normal Distribution, and Exponential Distribution to solve real world business problems. Use of R program and MS Excel to carry out data analysis.8 hours

Learning Outcomes :

By the end of this Unit, the student will be able to

- Explain the need of Normal distribution. (L2)
- Explain area under the bell shaped curve using Normal distribution. (L5)
- Evaluate Normal curve with given data. (L5)

UNIT – V

Correlation and Regression -Scatter Diagram; Karl Pearson's Correlation; Correlation Coefficient for Bivariate Frequency Distribution; Spearman's Rank Correlation; Fitting of Regression Lines, multiple Regression Coefficients, Applications of regression analysis; Correlation and regression Analysis.

10 hours

Learning Outcomes:

By the end of this Unit, the student will be able to

- Show the Correlation Coefficient for Bivariate Frequency Distribution.(L2)
- To model Linear and Multiple Regression Lines.(L3)
- To mark Correlation and Regression Analysis. (L4)

Course Outcomes:

Upon completion of this course, student will be able to

- Perform exploratory data analysis with ease to present results analytically for business problems. (L3)

- Distinguish application of right probability distributions to solve business problems. (L4)
- Understand the difference between correlation and regression analysis. (L3)
- Comprehend the meaning and interpretation of regression coefficients. (L6)
- Assess performance of regression models using appropriate metrics. (L5)
- Predicting outcome /response variable for given values of independent variables.(L6)

TEXT BOOKS:

1. Fundamentals of Mathematical Statistics by Gupta, S.C. and Kapoor, V.K., Sultan & Chand & Sons, New Delhi, 11th Ed, 2002.
2. The elements of Statistical Learning by Hastie, Trevor, Springer, 2009.
3. Introduction to Probability and Statistics by Ross, S.M., Academic Foundation, 2011.
4. Probability, Random Variables and Stochastic Processes by Papoulis, A. and Pillai, S.U. ,TMH, 2010.

M.Sc Data Science

VENTURE DEVELOPMENT

Hours/Week:2 **Continuous Evaluation:100**

Credits:2

M.Sc DATA SCIENCE
SEMESTER- I
CSCI6041 MODERN DATABASES LAB

Hours per week: 2 Continuous Evaluation:100 Marks
Credits: 2

Relational Databases

1. Implement DDL Statements.
2. Implement DML Statements.
3. Write the queries using Built-In Functions of SQL.
4. Design a Database and create required tables. Apply the constraints like Primary Key, Foreign Key, Not Null, and other constraints to the tables. Perform SQL Queries.
5. Write the queries to implement the Joins.
6. Write the queries to implement subqueries, group by and correlated queries.

Course Outcomes:

Upon completion of this course, the student should be able to:

- Design and implement a database schema for a given problem domain.
- Populate and query a database using SQL DML/DDL commands.
- Declare and enforce integrity constraints on a database
- Retrieve data using different SQL joins, subqueries, and correlated queries.

(a) NoSQL Databases:

1. Demonstrate to create Columnar, Big Table, Document Databases, Graph databases
2. Demonstrate to insert, update and delete data in different types of databases.
3. Demonstrate various techniques used to query the database.
4. Explain techniques to optimize querying using indexing.
    - Use different **FIND** techniques to query the document in MongoDB.
    - Use the **sort()** to sort the records.
    - Use the logical operators to query the document.
    - Use the conditional operators in MongoDB to query the document.

5. Demonstrate the methods to analyze data using aggregation techniques.
6. Explain the techniques of splitting data across machines.
7. Demonstrate **LIMIT()** and **SKIP()** methods in MongoDB.
8. Demonstrate an example for **Text Searching** in MongoDB.
9. Demonstrate the usage of different removalmethodologies to remove documents from collections.

Course Outcomes:

Upon completion of this course, the student is able to:

- Utilize the techniques used to create, insert, update and delete data/documents.(L3)
- Utilize various techniques used to query the database.(L3)
- Utilize techniques to optimize querying using indexing.(L3)

- Apply methods to analyze data using aggregation techniques.(L3)
- Adopt knowledge about the role of NoSQL in business.(L6)

- Select variousdesign aspects and operations of MongoDB.(L1)
- Define objects, load data, query data and performance tune Key-Value Pair NoSQL databases.(L3)

Text Books:

1. SQL, PL/SQL the Programming Language of Oracle by Ivan Bayross, BPB Publications, 4th Edition,2010.
2. MongoDB: The Definitive Guide by Kristina Chodorow, Shroff publisher, 2nd edition, 2013.
3. NoSQL with Mongo DB in 24 hours by Brad Dayley, Pearson Education, 1st edition, 2015.

M.Sc DATA SCIENCE

SEMESTER- I

CSCI6051 PYTHON PROGRAMMING LAB

Hours per week: 2

Credits:2 Continuous Evaluation: 100 Marks

1. Find all numbers which are multiples of 17, but not the multiples of 5, between 2000 and2500?
2. Swap two integer numbers using a temporary variable. Repeat the exercise using the code format: a, b = b, a. Verify your results in both the cases.
3. Given two pairs of Cartesian points such as (x1, y1) and (x2, y2). Find the Euclidean distance between them.
   Hint: Use math module to find the square root.
4. Print the first 2 and last 3 characters in a given string. Use the string slicing concept. Do not use loops. If the length of the string is less than 5, print a suitable message.
5. Implement bubble sort. Do not use the default sort() method.
   Hint: So as to familiarize with the concept of sorting, and nested looping structures.
6. Implement shallow copy and deep copy of a list. You may use the copy module.
   Hint: While we copy a list, just a reference is copied. Hence if we make any changes to one of the lists, the same will reflect in the other as well. This is called shallow copying. Hence, in some cases we might need to deep copy, where a completely independent copy is created. This can be achieved through the deepcopy() method of the copy module.
7. Write a temperature converter program, which is menu driven. Each such conversion logic should be defined in separate functions. The program should call the respective function based on the user’s requirement. The program should run as long as the user wishes so.
8. Find the largest of n numbers, using a user defined function largest().
9. Write a function that capitalizes all vowels in a string.
   Hint: Do not use the ASCII concept. Use the upper() method.
10. Write a function leapYear() which receives a four digit year and returns a Boolean value: True if the year is leap, False if the year is not leap.
11. Read a line containing digits and letters. Write a program to give the count of digits and letters. Hint: Instead of checking ASCII, use the in-built methods like isdigit(), isalpha()etc.
12. Write a function myReverse() which receives a string as an input and returns the reverse of the string.
13. Use the list comprehension methodology in python, to generate the squares of all odd numbers in a given list.
    Hint: List comprehension is one of the powerful techniques in python;
14. Check if a given string is palindrome or not.

Hint: do not use the C philosophy where we compare indices. Instead, copy the string as a new list, reverse the list using reverse(), join the list so that the reversed string is formed, using join(). Compare the new string and the old one.

15. Write a function to see if a given number is prime or not. Do not use any flag variables. Use a math module to find the square root, and its roof which will be fed in to range().

    Hint: Just the return statements are enough. No need for flag variables. The loop has to run up to the roof of the square root of the number.
16. Write a function to find the factorial of a number using recursion.
17. Extend the above problem to find the nCr of given values of n and r. Verify the result with the help of the filter tools module, which helps to find the combinations.
18. Write a program that eliminates duplicates in a list.Donotusetheconceptofsets.Now,convert the original list into a set. Verify the result in both cases.
19. The user will enter five integers separated by commas. Write a program to read these values, and make a list. Print the list.

    Hint: They will need to read the input using raw_input(), and then split the one and only line of input using split(). Then each of the values will need to be appended to a list, which will be empty at first.
20. Generate a dictionary and print the same. The keys of the dictionary should be integers between 1 and 10 (both inclusive). The values should be the cubes of the corresponding keys.
21. Create a nested dictionary. The roll number of a student maps to a dictionary. This inner dictionary will have name, age, and place as keys. Read details of at least three students.

    Hint: A sample output should look like the one given below:

    {11: {‘name’: ‘Sachin’, ‘age’: 18, ‘place’: ‘Kochi’},
    12:{‘name’:’Ammu’, ‘age’: 19, ‘place’: ‘Kannur’},
    13: {‘name’:’jishad’, ‘age’:20, ‘place’:’Calicut’}}
22. Enter a word. Create a dictionary with the letters of this word as keys, and the corresponding ASCII values as values.

    Hint: Students may use the ord() function. Further, this is a simple problem, if list comprehension is used.
23. Write a python program to find the tuples which all the elements are divided by a ‘k’ element from a list of tuples
24. Write a Python program to print all pair combinations of elements from 2 tuples. Hint: Students may use list comprehension.
25. Write a Python program to concatenate consecutive elements in tuple. Hint: use map() and tuple methods for concatenation .
26. Write a pandas program to select a name and score columns in rows 1,3,5,6 from the data frame consists of column values as index, name, score, attempts and qualify
27. Write a pandas program to implement line and box plots on the data frame student having the values as roll, name, exam1, exam2 and exam3.

28. Implement Multiple plots.
29. Implement Scatter plots with histogram.
30. Implement Bubble charts.

Course Outcomes:

Upon completion of this course, student will be able to:

- Able to develop programs in Python.(L4)
- Able to implement functions using parameters.(L3)
- Understand the concept of expressions.(L2)
- Construct programs using while statement.(L6)
- List the operations that can be performed on strings.(L4)
- Identify the differences between lists and dictionaries.(L4)
- List the concepts of operator overloading and polymorphism.(L4)
- Understand various forms of distribution.(L2)
- Able to implement plots and charts.(L4)

Text Books:

1. Mastering Python for Data Science by Samir Madhavan, PACKTPublishing,2015.
2. Think Python by Allen Downey O’Reilly Publications, 2ndEdition,2016.

M.Sc DATA SCIENCE
SEMESTER – II
LANG6181 PROFESSIONAL COMMUNICATION SKILLS

Hours per week: 3 Continuous Evaluation: 100 Marks
Credits: 2

Preamble:

This course is designed to expose students to the basics of academic and professional communication in order to develop professionals who can effectively apply communication skills, theories and best practices to meet their academic, professional and career communication needs.

Course Objectives:

To enable students to

- acquaint themselves with basic English grammar.
- acquire presentation skills.
- develop formal writing skills.
- develop creative writing skills.
- keep themselves abreast with employment-readiness skills.

UNIT - I

BACK TO BASICS: Tenses, Concord – Subject Verb Agreement, Correction of Sentences-Error Analysis, Vocabulary building. (10hours)

Learning Outcomes:

- Utilize structures and tenses accurately.(L3)
- Apply the right verb to the right subject in a sentence.(L3)
- Identify incorrect sentences in English and write their correct form.(L3)
- Choose new vocabulary and use in speaking and writing.(L1)

UNIT - II

ORAL PRESENTATION: What is a Presentation? Types of Presentations, Technical Presentation – Paper Presentation, Effective Public Speaking, Video Conferencing. (8 hours)

- Recall how to overcome speaking anxiety prior to presentation.(L1)
- Plan and structure effective presentations that deliver persuasive messages.(L3)
- Prioritize slides that can catch the attention of the audience. (L5)
- Show the skills in organizing, phrasing, and expressing the ideas, opinions and knowledge.(L1)
- Formulate and participate in a video conference effectively.(L6)

DOCUMENTATION: Letter Writing, E-mail Writing & Business Correspondence, Project Proposal, Report Writing, Memos, Agenda, Minutes, Circulars, Notices, Note Making.(10 hours)

At the end of the unit, the student will be able to:

- Outline a business letter, which includes appropriate greetings, heading, closing and body and use of professional tone.(L2)
- Develop crisp and compelling emails.(L6)
- Develop project proposals, reports and memos.(L6)
- Prepare agenda and draft minutes.(L6)

Prepare circulars, notices and make notes. (L6)

UNIT IV

CREATIVE WRITING: Paragraph Writing, Essay writing, Dialogue Writing, Précis Writing, Expansion of Hints, Story Writing. (6hours)

Learning Outcomes:

At the end of the unit, the student will be able to:

- Outline paragraphs on familiar and academic topics using a topic sentence, supporting detail sentences and a conclusion sentence. (L3)
- Select the structure of a five-paragraph essay and write essays that demonstrate unity, coherence and completeness. (L3)
- Structure natural, lucid and spontaneous dialogues. (L3)
- Examine the elements of a short story and develop their functional writing skills.(L3)

UNIT V

PLACEMENT ORIENTATION: Resume preparation, group discussion – leadership skills, analytical skills, interviews –Types of Interviews, Preparation for the Interview, Interview Process. (8hours)

Learning Outcomes:

At the end of the unit, the student will be able to:

- Formulate professional resume that highlights skills, specific to the student’s career field. (L6)
- Adapt the personality traits and skills required to effectively participate in a G.D.(L6)
- Infer the purpose of interviews.(L2)
- Aware of the processes involved in different types of interviews.(L3)
- Plan on how to prepare for an interview.(L4)
- Conclude on how to answer common interview questions.(L4**)**

Course Outcomes:

Upon completion of this course, student will be able to:

- Develop formal writing skills. (L3)
- Aware of the processes involved in interviews. (L3)
- Keep themselves with employment-readiness skills. (L3)

Text Books :

1. Essentials of Business Communication by Rajendra Pal and J S KorlahaHi, Sultan Chand & Sons.
2. Advanced Communication Skills by V. Prasad, Atma RamPublications.
3. Effective Communication byAshraf Rizvi, McGraw Hill Education; 1st Edition ,2005.
4. Interviews and Group Discussions How to face them by T.S.Jain, Gupta,1st Edition, Upkar Prakashan,2010.
5. High School English Grammar and Composition by P.C.Wren & Martin, N.D.V.PrasadaRao S.Chand.

M.Sc DATA SCIENCE

SEMESTER - II

CSCI6061 MACHINE LEARNING

Hours per week: 4 End Examination: 60Marks
Credits: 4 Sessionals: 40Marks

Preamble : Machine Learning is an application of artificial intelligence (AI) that provides systems the ability to automatically learn and improve from experience without being explicitly programmed. Machine learning focuses on the development of computer programs that can access data and use it to learn for themselves which is used in decision-making processes based on data inputs.

Course Objectives:

- To understand the basic theory underlying machine learning and its applications.
- To learn the Steps involved in designing and interpreting Model Performance.
- To Understand and apply different Learning models.
- To Understand the concept of Artificial Neural Network and its learning process.
- To Learn Different Learning models.

UNIT - I

Introduction: Human Learning definition, Types of Human Learning, Problems not to be solved using Machine Learning, Applications of Machine Learning, Tools in Machine Learning, Issues in Machine Learning.

Preparing to Model: Machine Learning Activities, Basis types of Data in Machine Learning, Exploring structured data, Data Quality and Remediation, Data Preprocessing. (10 hours)

Learning Outcomes:

By the end of the unit, the student is able to

- Identify basic data types in Machine learning. (L3)
- Develop methods for exploring structured data. (L3)
- Relate data quality and remediation. (L1)

UNIT-II

Modelling & Evaluation: Introduction, Selecting a model, Training a model, Model Representation and Interpretability, Evaluating Model Performance, Improving model performance.

Feature Engineering: Introduction, Feature Transformation, Subset selection. (8 hours)

By the end of the unit, the student is able to

- Learn how to select a model. (L1)
- Interpret Model Performance. (L2)
- Select the subset. (L3)

UNIT-III

Introduction, Importance of Bayes Theorem, Bayes theorem and concept learning, Bayesian Belief Network.

Supervised Learning: KNN, Decision Tree, Random forest model, Support vector Machines, Regression. (10 hours)

Learning Outcomes:

By the end of the unit, the student is able to

- Demonstrate the importance of Bayes Theorem. (L2)
- Make use of Bayesian Belief Networks(L3)
- Analyze KNN, Random Forest (L4)

UNIT-IV

Unsupervised Learning: Supervised Vs Unsupervised Learning, Applications of Unsupervised Learning, Clustering, Association Rule.

Basic Neural Networks: Neural Network, Understanding Biological Neuron, Exploring Artificial Neuron, Types of activation function, Early implementation of ANN, Architecture of Neural Networks, Learning process in Artificial Neural Networks, Back Propagation, Deep
Learning. (12 hours)

Learning Outcomes:

By the end of the unit, the student is able to

- Compare Supervised Learning and Unsupervised Learning. (L2)
- Construct Artificial Neural Networks. (L3)
- Make use of Deep Learning. (L3)

UNIT-V

Other Types of Learning: Introduction, Representation Learning, Active Learning, Instance Based learning, Association Rule Learning, Ensemble Learning Algorithm, Regularization algorithm. (8 hours)

Learning Outcomes:

By the end of the unit, the student is able to

- Outline Learning. (L2)
- Develop Instance based Learning. (L3)
- Utilize regularization algorithm. (L3)

Course Outcomes:

By the end of the Course, the student is able to

- Understand the basic concepts of machine learning algorithms. (L2)
- Understand the steps involved in developing a model and feature engineering techniques (L2)
- Learn various supervised and unsupervised learning algorithm (L2)
- Develop the ability to formulate machine learning techniques to respective problems. (L3)
- Understand the basic framework of neural networks.(L2)
- Understand various learning algorithms. (L2)

Text Book:

- Machine Learning by Subramanian, Chandra Mouli, Amit Kumar Das , Saikant Dutt, Pearson Publications, I edition, 2018.
- Machine Learning by Tom Mitchell, McGraw Hill, 2007

M.Sc DATA SCIENCE
SEMESTER – II
MATH6011 MATHEMATICS FOR DATA SCIENCE

Hours per week: 4	End Examination: 60Marks
Credits:4	Sessionals: 40 Marks

Preamble :

This course provides the basic knowledge on mathematics required for a data scientist. This course covers the concepts on Matrices, Normal forms, Rules of inference, Boolean Algebra and Boolean functions, Graph and Tree Concepts.

- To understand the difference between various types of matrices.
- To learn the basic concept and applications of matrices in real life problems.
- To identify and practice the mathematical logic problems with the help of truth tables or without using truth tables.
- Ability to implement features of inference rules in inference calculus.
- To understand the concept of Boolean algebra and Boolean functions.
- To understand the concepts of graphs, directed graphs, and trees.

UNIT - I

Matrices: Definition, addition and multiplication of matrices, various types of matrices, Determinant of a square matrix, Inverse of a matrix, Solution of system of non-homogenous linear equations by Crammer's rule , matrix inversion method, Gauss elimination method, Gauss-Jordan method, rank of a matrix, Normal form of a matrix, Echelon form of a matrix Consistency of linear system of equations, solution of system of linear homogenous equations, Eigen values and
Eigenvectors, norm, condition number. (10 hours)

By the end of this Unit, the student will be able to

- Explain various types of matrices.(L3)
- Evaluate system of equations by Crammer's rule, matrix inverse method, gauss elimination method.(L3)
- Explain various methods to find rank of a matrix.(L3)
- Evaluate Eigen values and Eigen vectors of a matrix.(L3)

UNIT - II

Mathematical Logic: Connectives, Negation, Conjunction, Disjunction, Conditional &Bi-Conditional, Well Formed Formulae, Tautologies, Equivalence of formulae, Duality, Tautological Implications, Functionally Complete Set of Connectives, Principal Disjunctive & Conjunctive Normal Forms, Inference Calculus, Rules of Inference, Indirect method of proof.

(8 hours)

Learning Outcomes:

By the end of this Unit, the student will be able to

- Demonstrate basic concepts of mathematical logic including connectives, tautology, equivalence, and normal forms.(L2)
- Evaluate problems on principal disjunctive normal form.(L5)
- Evaluate problems on principal conjunctive normal form.(L5)
- Describe methods to solve inference calculus problems.(L3)
- Describe indirect method of proof of an argument.(L3)

UNIT - III

Boolean Algebra: Definition and Examples, sub algebra, Direct product and Homomorphism, Boolean Functions, Boolean forms and free Boolean Algebras, Values of Boolean expressions and Boolean functions, Representation of Boolean functions, Minimization of Boolean functions, Karnaugh maps. (8hours)

Learning Outcomes:

By the end of this Unit, the student will be able to

- Define Boolean algebra, Sub Boolean algebra with examples. (L3)
- Explain the need of Boolean functions.(L3)
- Evaluate Boolean expressions and Boolean functions.(L5)
- Explain representation of Boolean functions.(L3)
- Explain minimization of Boolean functions using Karnaugh Maps.(L3)

UNIT - IV

Graph Theory: Definitions, Finite and Infinite graphs, Incidence and Degree, Isolated pendant vertices, Isomorphism, sub graphs, Walk, Path and Circuit, Connected and Disconnected graphs, components, Euler graphs, Euler graph theorem, Operations on graphs, Decomposition of Euler graphs into circuits, Hamiltonian paths and circuits. (10 hours)

Learning Outcomes:

By the end of this Unit, the student will be able to

- Define various types of graphs.(L3)
- Define Euler graphs and prove Euler graph theorem.(L3)
- Evaluate operations on graphs.(L5)
- Evaluate Hamiltonian paths and circuits.(L5)
- Evaluate isomorphism of undirected and directed graphs.(L5)

UNIT-V

Trees: Properties of trees, pendant vertices, distance & centers, rooted & binary trees, spanning trees, fundamental circuit, shortest spanning trees, Kruskal's algorithm, Binary Tree Traversals.

(8 hours)

Learning Outcomes:

By the end of this Unit, the student will be able to

- Define various types of trees and their properties.(L3)
- Explain rooted and binary trees.(L3)
- Construction of spanning trees from a connected graph.(L5)
- Explain Krushkal's algorithm to find minimum spanning tree of a connected graph.(L3)
- Explain Pre-order, Post- order, and In-order traversals of a binary tree.(L3)

Course outcomes:

At the end of the course student will be able to

- Determine inverse of matrix, perform matrix operations, solving systems of simultaneous linear equations (L3)
- demonstrate concepts of mathematical logic for analyzing propositions and proving theorems. (L2)
- Analyze logical propositions via truth tables. (L3)
- Understand the basic properties of Boolean algebra and simplify simple Boolean functions using the basic Boolean properties. (L2)
- Understand and apply the fundamental concepts in graph theory in solving practical problems. (L3)
- Model problems in Computer Science using graphs and trees. (L3)

Text Books :

1. Higher Engineering Mathematics by B.S.Grewal, Khanna Publishers, 43rd edition,2015
2. Numerical methods for scientific and engineering computation by M.K.Jain, S.R.K. Iyengar, R.K. Jain, New Age International publishers,6th edition,2012.

3. Discrete Mathematical Structures with Applications to Computer Science by J.P. Tremblay and R. Manohar, Tata McGraw Hill,1997.
4. Graph Theory with Applications to Engineering and Computer Science by Narsingh Deo, Prentice Hall of India,2006.

Reference Books:

1. Discrete Mathematics and its Applications by Keneth. H. Rosen, Tata McGraw-Hill, 6th Edition, 2009.
2. Discrete Mathematics by Richard Johnsonbaug, Pearson Education, 7th Edition,2008.
3. Discrete Mathematics for Computer Scientists and Mathematicians by J.L. Mott, A.Kandel, T.P. Baker, PrenticeHall.

M.Sc DATA SCIENCE
SEMESTER- II
MATH6021INFERENTIAL STATISTICS

Hours per week: 4 **End Examination: 60 Marks**
Credits:4 **Sessionals: 40 Marks**

Preamble :

The inferential statistics aids the process of drawing inferences on population parameters based on scientific study of sample data. Statistics facilitates the decision making process by quantifying the element of chance or uncertainties and make valid inferences. The inferential statistics formulate the basis of the growth of almost all the disciplines of the contemporary world. The statistical tests finds the relationship between a categorical and a numeric variables and determine whether the relationship is significant or not.

Course Objectives:

- To Understand point estimation and properties of best estimators
- To Construct and interpret confidence intervals for means when the population standard deviation is known
- To Construct and interpret confidence intervals for means when the population standard deviation is unknown
- To Carry out hypothesis tests for population parameters when the population standard deviation is known
- To Carry out hypothesis tests for population parameters when the population standard deviation is unknown
- To acquire the ability to implement features of ANOVA and experimental design.
- To understand the concepts and application of non- parametric tests.

UNIT – I

Estimation and sampling distribution - Types of estimation – point estimation and interval estimation; Properties of good estimators, Standard error of statistics; Sampling distribution of sample mean, sample standard deviation, sample proportion, sample correlation coefficient, application of central limit theorem. 10 hours

Learning Outcomes:

By the end of this Unit, the student will be able to

- Understand Point Estimation and Interval Estimation. (L2)
- Infer the standard error of Statistic.(L2)
- Analyze Central Limit Theorem. (L4)

UNIT-II

Concepts of Testing of Hypothesis - Statistical Hypothesis - Simple and composite hypothesis, Null and Alternative hypothesis - two kinds of errors, level of significance, size and power of a test, most powerful test, Neyman-Pearson lemma with proof. Simple examples using Neyman-Pearson lemma. Uniformly most powerful tests and unbiased tests based on normal Likelihood ratio test (without proof) and its properties.

8 hours

Learning Outcomes:

By the end of this Unit, the student will be able to

- Demonstrate Null and Alternate Hypothesis. (L2)
- Choose the Level of Significance.(L3)

- Apply the tests on sample statistics to derive best estimators. (L3)

UNIT – III

Test of significance for large samples - Test of significance for mean(s), variance(s), proportion(s), correlation coefficient(s) based on Normal distribution.

Test of significance for small samples- Test of significance for mean(s), variance(s), correlation coefficient(s), regression coefficient, based on t, Chi-square and F-distributions. Applications of Chi-square in test of significance (independence of attributes, goodness of fit). 12 hours

Learning Outcomes:

By the end of this Unit, the student will be able to

- Understand the test of Significance. (L2)
- Explain and evaluate the test of significance for large sample. (L5)
- Explain and Evaluate the test of significance for small sample.(L5)

UNIT – IV

Analysis of variance (ANOVA) - Assumptions of ANOVA, Techniques of ANOVA, One-way ANOVA, Analysis of variance of Two-Way Classification Model, Random Block Design, Latin Squares, Randomized Blocks Vs Latin Square, Factorial designs. 10 hours

Learning Outcomes:

By the end of this Unit, the student will be able to

- Model features of ANOVA and experimental design.(L3)
- Explain the need of Analysis of Variance.(L2)
- Outline the different types of experimental designs.(L3)

UNIT – V

Non-parametric tests- Advantages of Non-Parametric Tests, Kolmogorov -Smirnov test, Sign test, Wald-Wolfowitz run test, run test for randomness, median test, Wilcoxon test and Wilcoxon – Mann-Whitney U test.

8 hours

Learning Outcomes:

By the end of this Unit, the student will be able to

- Infer the advantage of Non-Parametric Tests. (L2)
- Able to I
- llustrate Non-Parametric Tests.(L2)

Course Outcomes:

Upon completion of this course, student will be able to

- Gain thorough understanding of concepts of estimation to derive estimators for population parameters.
- Explain the procedure of testing of hypothesis by utilizing the sampling distribution of sample statistics.
- Demonstrate different tests of significance for large samples.
- Understand when to use small sample tests for business problems and conduct tests of significance for small samples.
- Understand the advantages and limitations of non-parametric tests and Demonstrate applications of non-parametric tests

TEXT BOOKS:

1. Probability and Statistics for Engineering and Sciences by Jay L.Devore, Cengage Learning, 2015.

2. Probability and Statistics for Engineers and Scientists by Ronals E.Walpole, Raymond Mayers, Sharon L.Myers, Keying E. Ye, Pearson Publication, Ninth Edition, 2014.
3. Probability and Statistics for Science and Engineering by Shankar Rao, University Press, 2015.
4. Larry Wasserman, All of Statistics: A Concise Course in Statistical Inference, Springer Texts in Statistics, Springer-Verlag, New York, 2004.
5. Larry Wasserman, All of Nonparametric Statistics, Springer Texts in Statistics, Springer Verlag, New York, 2005.

M.Sc DATA SCIENCE
SEMESTER- II
CSCI6071 ARTIFICIAL INTELLIGENCE

Hours per week: 4	**End Examination: 60 Marks**
Credits:4	**Sessionals: 40 Marks**

Preamble:

This course enables the students to think critically about what makes humans intelligent, and how computer scientists are designing computers to act more like us. AI plays an important role in the design and development of systems with intelligent behavior. The primary objective of this course is to provide an introduction to the basic principles and applications of Artificial Intelligence.

Course Objectives:

- To teach fundamentals of Artificial Intelligence, the concept of Intelligent Agents and problem-solving process through uninformed and informed searches.
- To gain an insight into competitive environments which gives rise to adversarial search problems, often known as games.
- To view many problems in AI as problems of constraint satisfaction.
- To gain complete idea of knowledge representation techniques Propositional and First-order logics.
- To learn how to trace the inference mechanism in First-order logics.

UNIT – I

Introduction: AI definition, Foundations, History, State of the Art.

Intelligent Agents: Agents and Environments, Concept of Rationality, Nature of Environments, Structure of Agents.

Solving Problems by Searching: Problem Solving Agents, Example problems, Searching for solutions, Uninformed Search Strategies, Informed search strategies, Heuristic Functions, Local Search Algorithms and Optimization Problems. (10 hours)

Learning Outcomes

After completion of this unit, student will be able to

- Define Artificial Intelligence. (L1)
- How agents work in environments.(L1)
- Recall uninformed search techniques.(L1)
- Illustrate the working of informed search techniques.(L2)

UNIT-II

Adversarial Search: Games, Optimal Decisions in Games, Alpha- Beta Pruning, Imperfect real-time decisions, Stochastic games, Partially observable games, State of art game program. (8 hrs)

Learning outcomes

After completion of this unit, student will be able to

- Understand how games improve intellectual abilities of humans. (L1)
- Choose optimal decisions in games. (L1)
- Illustrate alpha-beta pruning. (L2)
- Compare stochastic and partially observable games.(L2)

UNIT-III

Constraint Satisfaction Problems: Defining Constraint Satisfaction problems, Constraint Propagation, Backtracking search for CSPs, Local Searches for CSPs.

Logical Agents: Knowledge-Based Agents, Wumpus World, Logic, Propositional Logic and Propositional Theorem proving. (8hours)

Learning outcomes

After completion of this unit, student will be able to

- Define constraint satisfaction problems.(L1)
- Illustrate inference in constraint satisfaction problems.(L2)
- Contrast backtracking search and local search for constraint satisfaction problems.(L2)
- Define knowledge-based agents.(L1)
- How to represent real-world facts in propositional logic.(L1)

UNIT-IV

First Order Logic: Syntax and Semantics of First Order Logic, Using First Order Logic, Knowledge Engineering in First Order Logic.

Inference in First Order Logic: Propositional Vs First Order Inference, Unification and Lifting, Forward Chaining, Backward Chaining, Resolution. Introduction to Fuzzy Logic. (8 hours)

Learning outcomes

After completion of this unit, student will be able to

- Infer proofs in propositional and first-order logic. (L2)
- Define propositional and first-order inference.(L1)
- Outline unification and lifting. (L2)
- Experiment with forward chaining and backward chaining. (L3)
- Make use of resolution.(L3)

UNIT-V

Classical Planning: Definition, Algorithms for planning as state space search, Planning Graphs.

Knowledge Representation: Ontological Engineering, Categories and Objects, Events, Mental Events and Mental Objects, Reasoning Systems for Categories, Reasoning with Default information, Internet Shopping World. (8hours)

Learning outcomes

After completion of this unit, student will be able to

- Overview on basic knowledge representation aspects and on ontologies.(L2)
- Design and generate path planning using knowledge representation. (L5)
- Accumulate sophisticated knowledge about the environment for processing tasks ormethods. (L5)

Course Outcomes:

Upon completion of this course, student will be able to:

- Illustrate artificial intelligence, the role of intelligent agents, uninformed and informed search techniques. (L2)
- Examine competitive environments like game problems. (L2)
- Interpret many real-world problems as constraint satisfaction problems. (L3)
- Illustrate what knowledge representation is and able to distinguish propositional and first-order logics. (L3)
- Infer proofs using resolution in first-order logic. (L3)
- A machine-interpretable representation of the world, similar to human reasoning can be designed (L4)

Text Books:

1. Artificial Intelligence - A Modern Approach by Stuart J. Russell, Peter Norvig, Pearson Education, 3rd Edition. 2015.

Reference Books:

1. Artificial Intelligence by Kevin Knight, Elaine Rich, 3rd Edition,TMH,2017.
2. Artificial Intelligence by Saroj Kaushik, Cengage Learning,2011.

M.Sc DATA SCIENCE
SEMESTER – II
CSCI6081 INTRODUCTION TO OBJECT ORIENTED SOFTWARE ENGINEERING

Hours per week: 3 Continuous Evaluation: 100 Marks
Credits: 2

Preamble:

Object-Oriented Software Development is an approach/paradigm of developing software by identifying and implementing a set of objects and their interactions to meet the desired objectives. The first step towards this kind of software development is to learn and master the various concepts, tools and techniques that are to be used design and implementation of such systems.

Course Objectives:

- To learn and understand various O-O concepts along with their applicability context.
- Given a problem, identify domain objects, their properties, and relationships amongthem.
- How to identify and model/represent domain constraints on the objects and (or) on their relationships.
- Develop design solutions for problems on various O-O concept
- To learn various modeling techniques to model different perspectives of object-orientedsoftware design (UML).
- Analysing and designing Object-Oriented solutions for Real-World Problems.
- To learn design solutions for recurring problems.

Unit –I

Introduction to Software Engineering: Introduction, What is Software Engineering, Software Engineering Concepts ,Software Engineering Concepts, Software Engineering Activities , Managing.

Software Development Modeling with UML : An Introduction, An overview of UML, Modeling Concepts, A deeper View into UML

Learning Outcomes

- To explain about software engineering concepts and its related concepts (L2)
- To explain about various UML Concepts (L2)

Unit –II

Requirement Elicitation : Introduction, An Overview of Requirements Elicitation, Requirements Elicitation Concepts, Requirements Elicitation Activities, Managing Requirements Elicitation, ARENA Case Study

Learning Outcomes

- To build various requirement Elicitation concepts (L2)

Unit –III

Analysis: Introduction, An Overview of Analysis, Analysis Concepts, Analysis Activities from use cases to objects, Managing Analysis, ARENA Case Study,

System Design: Decomposing the System, An Introduction, An overview of System Design,System Design Concepts, System Design Activities From objects to Subsystems

Learning Outcomes

- To develop Analysis concepts and its related concepts (L3)
- To explain about various System concepts (L2)

Unit –IV

System Design Addressing Design Goals: Introduction, An overview of system Design Activities, UML Deployment Diagrams, System Design Activities Addressing design goals, Managing System Design, ARENA Case Study

Learning Outcomes

- To develop System Design concepts and its related concepts (L3)

Unit –V

Object Design: Specifying Interfaces ,Introduction, An overview of Testing, Testing Concepts, Testing Activities, Managing Testing.

Learning Outcomes

- To develop Object Design document and its related concepts (L3)
- To explain about various Testing concepts (L2)

TextBook:

Object-Oriented Software Engineering, Using UML, Patterns, Design and Java, Bernd Bruegge & Allen H.Dutoit,Prentice Hall

Reference Textbooks:

Blaha and Rumbaugh "Object-oriented modeling & Design with UML, 2nd Ed, PHI

Martin Fowler, "UML Distilled", 3rd Ed, Pearson Education.

Grady Booch, "Object-oriented analysis and design", 2nd Ed, Pearson Education

Course Outcomes

- Discuss software engineering concepts and UML concepts (L6)
- Discuss Requirement Elicitation Concepts (L6)
- Develop Analysis and System Design Concepts (L3)
- Discuss system design concepts (L6)
- Develop Object design document and Testing Concepts (L3)

M.Sc DATA SCIENCE
SEMESTER- II
MATH6031Inferential Statistics Lab

Hours per week: 2
Credits: 2 **Continuous Assessment: 100 Marks**

1. Confidence limits for population parameters
2. Large sample tests for: (i) single population mean, (ii) difference between two population means
3. Large sample tests for: (i) single population proportion, (ii) difference between two population proportions
4. Large sample tests for Fisher's Z- transformation.
5. Small sample tests for one-sample and two-sample t-test.
6. Paired comparisons test (Paired t-test)
7. F-test for testing equality of population variances
8. Chi square tests for (i) independence of attributes, (ii) goodness of fit
9. One-factor analysis of variance (ANOVA)
10. Two-factor analysis of variance (ANOVA) with and without replication
11. Two-factor analysis of variance (ANOVA) with and without replication
12. Non-parametric tests (run test, median test, sign test)
13. Using R program and Data analysis tool pack of MS-Excel conducting hypothesis tests

M.Sc DATA SCIENCE

SEMESTER - II

CSCI6091MACHINE LEARNING LAB USING PYTHON

Hours per week: 2 Continuous Evaluation: 100 Marks Credits:2

1. Creating a Data Frame in Pandas from csv files.
2. Importing Data with Pandas – adding columns to the data frame.
3. Handling Missing Data- drop, fill, aggregate functions.
4. Indexing Data Frames with Pandas, Indexing Using Labels in Pandas.
5. Exploratory Data Analysis with Pandas- for both one dimensional and two dimensional data (series or data frames) - describe, group data, ANOVA, correlation and correlation methods, rank.
6. Calculating Mean, Trimmed Mean, Weighted Mean, Median,
7. Plotting using pandas- Exploratory analysis based on the plots.
8. Data Visualization with different charts in python.
9. Apply PCA function. Find Eigen Values and EigenVectors.
10. Working with JSON Data with python.
11. Use OpenCV to find a face in an image.
12. A weather prediction model that predicts if there'll be rain or not in a particular day with decision tree regression concept.
13. A Python script to create a confusion matrix on a predicted model.
14. Consider a dataset where we have a value of response y for every features.
    a) Find a line which fits best and predict the response for any new feature values using simple linear regression.
    b) Find the errors using Least Squares technique to fine tune the model.

X								7		
Y								9		

15. Consider a dataset with p features (or independent variables) and one response (or dependent variable). Also the dataset contains n rows/observations.
    a. Find the regression line using multiple linear regression.
    b. Find the residual error of i^{th} observation.
16. A researcher has collected data on three psychological variables, four academic variables (standardized test scores), and the type of educational program the student is in for 600 high school students. She is interested in how the set of psychological variables is related to the academic variables and the type of program the student is in using Multivariate Regression.
17. Demonstrate to find the values of the parameters of a function that minimizes the cost function using Stochastic Gradient Descent.
18. A python program to explore your data with matplotlib and PCA, preprocess your data with normalization. Split the data into training and test sets. construct an unsupervised model (K-means algorithm) to fit the model to the data, predict values, and validate the model that is built.
19. Multidimensional data analysis in Python- import, Clustering, Exploratory Data Analysis.
20. Demonstrate to perform support vector classifier on a non linear dataset using a linear kernel.

Course Outcomes:

Upon completion of the course student will be able to

- Learn to Load data sets.(L1)
- Learn about the various libraries offered by Python to manipulate, preprocess and visualize data.(L3)
- Learn the technique to reduce the number of variables using Feature Selection and Feature Extraction.(L3)
- Learn in building models and model persistence using regression, classification.(L3)
- Learn various machine learning algorithms like KNN, Decision Trees, SVM, Clustering in detail.(L3)
- Learn to use optimization techniques to find the minimum error in your machine learning model.(L3)

Text Books:

1. Python Machine Learning by Sebastian Raschka and Vahid Mirialili, Packt publishing, 2nd Edition, 2017.
2. Introduction toMachine Learning with Python by Andreas C. Müller and Sarah Guido, Orielly, 1st Edition, 2016.
3. Machine Learning in Python by Michael Bowles, Wiley Publishers,2018.

M.Sc DATA SCIENCE
SEMESTER - III
CSCI7001 DEEP LEARNING

Hours per week: 4	End Examination: 60 Marks
Credits:4	Sessionals: 40 Marks

Preamble:

Deep learning is a form of machine learning that enables computers to learn from experience and understand the world in terms of a hierarchy of concepts. Deep learning methods have dramatically improved the state-of-the-art in speech recognition, visual object recognition, object detection and many other domains such as drug discovery and genomics. Deep convolutional nets have brought about breakthroughs in processing images, video, speech and audio, whereas recurrent nets have thrown light on sequential data such as text and speech.

- To learn the fundamental principles, theory and approaches for learning with deep neural networks.
- To demonstrate the key concepts, issues and practices when training and modeling with deep architectures.
- To learn the main variants of deep learning (such as convolutional and recurrent architectures) and their typical applications.

UNIT - I

Review of Machine Learning: The Learning Machines, The Math Behind Machine Learning: Linear Algebra, The Math Behind Machine Learning: Statistics, Work of Machine Learning, Logistic Regression, Evaluating Models, Building an understanding of machine learning.
(10 hours)

Learning Outcomes:

After completion of this unit, student will be able to

- Learn the basics of deep learning. (L2)
- Recollect the basic machine learning models. (L1)
- Learn the measures of evaluating a model. (L2)

UNIT – II

Foundations of Neural Networks and Deep Learning: Neural Networks, Training Neural Networks, Activation Functions, Loss Functions , Hyper parameters. (8 hours)

After completion of this unit, student will be able to

- Understand the foundations of neural networks. (L2)
- Understand the techniques of training neural networks.(L2)
- Learn different activation functions, loss functions, hyper parameters. (L3)

UNIT - III

Fundamentals of Deep Networks: Defining Deep Learning, Common Architectural Principles of Deep Networks, Building Blocks of Deep Networks. (8 hours)

After completion of this unit, student will be able to

- Learn the evolution of deep neural networks. (L2)
- Understand the building blocks and architectural principles of Deep Learning. (L3)

UNIT – IV

Major Architectures of Deep Networks: Unsupervised Pre-trained Networks, Convolutional Neural Networks (CNNs), Recurrent Neural Networks, Recursive Neural Networks. (10 hours)

Learning Outcomes:

After completion of this unit, student will be able to

- Learn the major architectures of deep neural networks. (L2)
- Identify the difference of different architectures. (L3)

UNIT - V

Building Deep Networks: Matching Deep Networks to the Right Problem, The DL4J Suite of Tools, Basic Concepts of the DL4J API, Modelling CSV Data with Multilayer Perceptron Networks, Modelling Handwritten Images Using CNNs, Modelling Sequence Data by Using Recurrent Neural Networks, Applications of Deep Learning in Natural Language Processing.

(10 hours)

Learning Outcomes:

After completion of this unit, student will be able to

- Build a Deep Neural network using APIs. (L4)
- Understand how to apply a variety of learning algorithms to data. (L3)
- Understand how to perform evaluation of learning algorithms and model selection.(L4)

Course Outcomes:

At the end of the course, student will be able to

- Understand the concepts of Machine learning (L2).
- Understand the framework of neural networks (L2)
- Understand the principles and blocks of deep neural network architecture (L2)
- Identify the difference between various major deep neural network architectures (L3)
- Use DL4J API to design deep neural network model (L3)

Text Book:

1. Deep Learning: A practitioners approach by Josh Patterson & Adam Gibson, Oreilly publications, 1st edition, 2017.

Reference Books:

1. Deep Learning by Ian Goodfellow, Yoshua Bengio. Aaron Courville. The MIT Press,2016.
2. Deep Learning, A Practical Approach by Rajiv Chopra, Khanna Book Publishing**,**2018.

M.Sc DATA SCIENCE
SEMESTER - III
CSCI7011 BIG DATA ANALYTICS

Hours per week: 4 End Examination: 60Marks
Credits:4 Sessionals: 40Marks

Preamble:

The internet, Big Data, vastly improved computational power*, and a wide variety of variables are involved in complex, real-world problems that led to a new set of analytic techniques and technologies. The concept of Big Data includes massive volumes of data and huge benefits that can accrue from the analysis of it.*

Course Objectives:

- To introduce an in depth understanding of all the concepts related to Big Data and its uses.
- To provide an insight on the underlying technologies to handle Big Data and the Ecosystem of Hadoop.
- To explore the layers of Big Data Stack and YARN Functionality.
- To Understand the Architecture, benefits and Properties of Hive and Pig.
- To provide learners with a deep and systematic knowledge on Spark.

UNIT – I

Getting an overview of Big Data: Big Data definition, History of Data Management, Structuring Big Data, Elements of Big-data, Big Data Analytics.

Exploring use of Big Data in Business Context: Use of Big Data in Social Networking, Use of Big Data in preventing Fraudulent Activities in Insurance Sector & in Retail Industry. **(**8 hours) **Learning Outcomes:**

After completion of this unit, student will be able to

- Learn various sources of data and forms of data generation. (L2)
- Understand the evolution and elements of Big Data. (L2)
- Explore different opportunities available in the career path. (L3)
- Understand the role and importance of Big Data in various domains. (L2)

UNIT – II

Introducing Technologies for Handling Big Data: Distributed and parallel computing for Big Data, Introducing Hadoop, Cloud computing and Big Data, In-memory Computing Technology for Big Data.

Understanding Hadoop Ecosystem: Hadoop Ecosystem**,** Hadoop Distributed File System, MapReduce, Hadoop YARN, Introducing HBase, Combing HBase and HDFS, Hive, Pig and Pig Latin, Sqoop, ZooKeeper, Flume, Oozie.

Understanding MapReduce Fundamentals and HBase: The MapReduce Framework, Techniques to Optimize Map Reduce Jobs, Uses of Map Reduce, Role of HBase in Big Data Processing.

(10hours)

Learning Outcomes:

After completion of this unit, student will be able to

- Identify the difference between distributed and parallel computing. (L3)
- Learn the importance of Virtualization in Big Data. (L2)
- Learn the details of Hadoop and Cloud Computing. (L2)
- Learn the architecture and features of HDFS. (L2)
- Understand Hadoop Ecosystem, MapReduce and HBase. (L2)
- Apply the technique in optimizing MapReduce jobs. (L3)

UNIT- III

Understanding Big Data Technology Foundations: Exploring the Big Data Stack, Virtualization and Big Data, Virtualization approaches.

Understanding Hadoop YARN Architecture: Background of YARN, Advantages of YARN, YARN Architecture, Working of YARN, YARN Schedulers, YARN Configurations, YARN commands. (10 hours)

Learning Outcomes:

After completion of this unit, student will be able to

- Explore the layers of Big Data Stack. (L2)
- Learn virtualization approaches in handling Big Data operations. (L2)
- Learn the importance of YARN. (L2)
- Understand the use and importance of schedulers and backward compatibility in YARN. (L3)
- Learn the commands, log management and configuration for handling Big Data. (L3)

UNIT – IV

Exploring Hive: Introducing Hive, Getting Started with Hive, Hive Services, Data Types, Built- in Functions, Hive-DDL, Data Manipulation, Data Retrieval Queries, Using Joins.

Analyzing Data with Pig: Introducing Pig, Running Pig, Getting started with Pig Latin, working with operators in Pig, Debugging Pig, Working with Functions in pig, Error Handling in Pig.

Understanding Analytics and Big Data: Comparing Reporting and analysis, Types of Analytics, Developing an Analytic Team, Understanding Text Analytics. (10 hours)

Learning Outcomes:

After completion of this unit, student will be able to

- Learn the working of Hive and query execution. (L2)
- Learn the importance of Pig. (L2)
- Choose the operators in Pig. (L2)
- Understand various types of analytical approaches. (L3)

UNIT – V

Spark: Introduction, Spark Jobs and API, Spark 2.0 Architecture, Resilient Distributed Datasets: Internal Working, Creating RDDs, Transformations, Actions. DataFrames: Python to RDD Communications, speeding up PySpark with DataFrames, Creating Data Frames and Simple DataFrame Queries, Interoperating with RDDs, Querying with DataFrame API, Querying with SQL, DataFrame Scenario, Spark Dataset API

After completion of this unit, student will be able to

- Get an overview of Spark technology and Jobs Organization concept (L2)
- Understand the schema less data structure available in PySpark (L3)
- Get an overview of data frames that bridges the gap between Scala and Python in terms of efficiency. (L2)

Course Outcomes:

Upon completion of this course, student will be able to

- Understand the terminology and the need of Big Data in Real time scenarios (L2)
- Understand the need of different tools in Hadoop ecosystem (L2)
- Understand the concept of virtualization and YARN architecture (L2)
- execute queries in Hive and develop scripts using pig Latin.(L3)
- Learn Apache Spark fundamentals, RDD, DataFrame. (L3)

Text book:

1. Big Data Black Book by Dt Editorial Services, Dreamtech Publications, 2016. (Unit- I-IV)
2. Learning PySpark by Tomasz Drabas, Denny Lee, Packt publishing, 2017. (Unit – V)

Reference Book:

1. Hadoop The Definitive Guide by Tom White, O'reilly ,4th Edition,2016.

M.Sc DATA SCIENCE

SEMESTER - III

Generic Elective – I

CSCI7021 CLOUD COMPUTING

Hours per week: 4 End Examination: 60Marks
Credits:4 Sessionals: 40 Marks

Preamble:

Cloud computing is the on-demand availability of computer system resources, especially data storage and computing power, without direct active management by the user. The term is generally used to describe data centers available to many users over the Internet. Large clouds, predominant today, often have functions distributed over multiple locations from central servers. If the connection to the user is relatively close, it may be designated an edge server.

Course Objectives:

- To learn the basic concepts and services provided by cloud computing.
- To understand the concepts of virtualization in cloud computing.
- To understand the architecture of cloud computing.
- To learn the basic concepts of AWS.
- To understand the working of AWS.

UNIT – I

Introduction: Cloud Computing at a Glance, The Vision of Cloud Computing, Definition of a Cloud, A Closer Look, Cloud Computing Reference Model, Characteristics and Benefits, Challenges Ahead, Historical Developments, Distributed Systems, Virtualization, Web 2.0, Service-Oriented Computing, Utility-Oriented Computing, Building Cloud Computing Environments, Application Development, Infrastructure and System Development, Computing Platforms and Technologies, Amazon Web Services (AWS), Google AppEngine, Microsoft Azure, Hadoop, Force.com and Salesforce.com, ManjrasoftAneka. (8 hours)

Learning`Outcomes:

By the end of the unit the student will be able to

- Explain basic features of cloud computing.(L2)
- Describe the cloud computing reference model.(L2)
- Illustrate the characteristics and benefits of cloud computing.(L3)
- Outline the computing platforms and technologies.(L2)
- Describe the basic features of AWS and other cloud providers.(L2)

UNIT- II

Virtualization: Introduction, Characteristics of Virtualized Environments, Taxonomy of Virtualization Techniques, Execution Virtualization, Other Types of Virtualization, Virtualization and Cloud Computing, Pros and Cons of Virtualization, Technology Examples, Xen: Para-virtualization, VMware: Full Virtualization, MicrosoftHyper-V. (5 hours)

By the end of the unit the student will be able to

- Explain the basic features of virtualization.(L2)
- Identify the different types of virtualization.(L1)
- Describe the concept of VMware.(L2)
- Explain the relation between virtualization and cloud computing.(L2)

UNIT - III

Cloud Computing Architecture: Introduction Cloud Reference Model, Architecture Infrastructure / Hardware as a Service, Platform as a Service, Software as a Service, Types of Clouds, Public Clouds, Private Clouds, Hybrid Clouds, Community Clouds, Economics of the

Cloud, Open Challenges, Cloud Dentition, Cloud Interoperability and Standards, Scalability and Fault Tolerance, Security, Trust, and Privacy, Organizational Aspects. (8 hours)

Learning Outcomes:

By the end of the unit the student will be able to

- Explain the architecture of cloud computing. (L2)
- Describe the different services provided by cloud computing.(L2)
- Identify the different types of clouds.(L1)
- Outline the challenges in cloud computing.(L2)
- Explain the basic concepts of security in cloud computing.(L2)

UNIT - IV

Discovering the AWS Development Environment : Starting AWS Adventure, **Defining** the AWSCloud, Discovering IaaS, Determining Use of AWS, Considering the AWS-Supported Platforms.**Obtaining Development Access to Amazon Web Services**: **Discovering** the Limits of FreeServices, Considering the Hardware Requirements, Getting Signed Up, Testing the Setup,Choosing the Right Services, Getting a Quick Overview of Free-Tier Services, Matching AWS Services to the Application, Considering AWS Security Issues. (10 hours)

By the end of the unit the student will be able to

- Explain basic concepts of AWS.(L2)
- Identify the hardware requirements for AWS.(L1)
- Illustrate basic steps corresponding to AWS.(L3)
- Identify the AWS services required for an application.(L1)
- Describe the AWS security issues.(L2)

UNIT- V

Starting the Development Process: **Considering** AWS Communication Strategies, Defining theMajor Communication Standards, Understanding how REST Works, Creating a DevelopmentEnvironment, Choosing a Platform, Obtaining and Installing Python, Working with the Identityand Access Management Console, Installing the Command Line Interface Software,ConfiguringS3 Using CLI, Configuring S3 Using Node.js, Configuring S3 Using a DesktopApplication, Creating a Virtual Server Using EC2,Getting to Know the Elastic Compute Cloud(EC2), Working with Elastic Block Store (EBS) Volumes, Discovering Images and Instances.

Performing Basic Development Tasks : Understanding AWS Input/Output, **Considering** theInput/Output Options, Working with JSON, Working with XML, Working with Amazon APIGateway, Developing Web Apps Using Elastic Beanstalk, Considering Elastic Beanstalk(EB)Features, Deploying an EB Application, Updating an EB Application, Removing Unneeded
Applications, Monitoring Your Application Using Amazon Cloud Watch. (12 hours)

By the end of the unit the student will be able to

- Explain the major communication standards.(L2)
- Summarize creating a development environment.(L2)
- Illustrate installation of command line interface.(L1)
- Describe the basic concepts of development tasks.(L2)
- Explain the monitoring of an application using Amazon Cloud watch.(L2)

Course Outcomes:

- Understand the need of cloud computing.(L2)
- Identify the differences between different types of ciphers.(L3)

- List different cloud providers. (L2)
- Understand the various features of virtualization. (L2)
- Understand the architecture of cloud computing. (L2)
- List the basic features of AWS.(L2)
- Understand the working of AWS.(L2)
- Understand basic development tasks.(L2)

Text Book:

1. Mastering Cloud Computing by Rajkumar Buyya, Christian Vecchiola, S Thamarai Selvi, Morgan Kaufmann ,2013.
2. AWS for Developers- Dummies by John Paul Mueller, John Wiley & Sons Inc. publications, 2017.

Reference Books:

1. Cloud Computing Concepts Technology Architecture by Thomas Erl,Pearson Education,2014.
2. Cloud Computing Explained by John Rhoton , Recursive Press 2nd edition, 2009.

M.Sc DATA SCIENCE
SEMESTER- III
GENERIC ELECTIVE - I
MATH7001 APPLIED MULTIVARIATE STATISTICAL ANALYSIS

Hours per week: 4 **End Examination: 60 Marks**
Credits: 4 **Sessionals: 40 Marks**

Preamble : This course emphasizes both theory and applications of statistics and is structured to provide knowledge and skills necessary for the employability of students in data analytics. The pre-requisite of this course is system and extensive computer training of statistical computations including standard software packages such as R and Python.

Course objectives:

- Understand and derive the important properties of the multivariate normal distribution
- Decide the distribution of linear combinations of multivariate normal distributions
- Analyze multivariate data sets with the methods included in the course
- Evaluate the results from multivariate statistical analysis and interpretation of results
- Summarize the most important results from a scientific report on some area in multivariate analysis
- Evaluate the applicability of different models for real world business problems and identifying relevant multivariate statistical methods to find solutions

UNIT - I

Multivariate Linear Model and Analysis of Variance and Covariance, Maximum likelihood estimation of parameters, tests of linear hypothesis, distribution of partial and multiple correlation coefficients and regression coefficients. Multivariate linear regression, multivariate analysis of variance of one-way and two-way classification data. Multivariate analysis of covariance, Hoteling T^2 and Mahalanobis D^2 applications in testing and confidence set construction.

10 hours

Learning outcomes:

At the end of this unit, the student will be able to

- use multiple regression techniques to build empirical models for business data (L3)
- understand how the method of least squares extends to fitting multiple regression models. (L2)
- use the regression model to estimate the mean response, and to make predictions and to construct confidence intervals and prediction intervals. (L3)
- build regression models with polynomial terms. (L3)
- Apply Hoteling T^2 and Mahalanobis D^2 statistic to aid in decision making. (L3)

UNIT - II

Multivariate Normal Distribution (MND) : Introduction to multivariate normal distribution, probability density function and moment generating function, singular and non-singular normal distributions, distribution of linear and quadratic form of normal variables, marginal and conditional distributions. Random sampling from multivariate normal distributions. Goodness of fit of multivariate normal distribution. Wishart matrix-its distribution and properties.

10 hours

Learning outcomes:

At the end of this unit, the student will be able to

- gain experience of how the various methods are applied, and results interpreted, in practice. (L3)

- evaluate the suitability of, and compare, different methods in practice. (L5)
- determine the shape of the multivariate normal distribution from the eigenvalues and eigen vectors. (L5)
- perform statistical tests of the mean value vector of a multivariate normal distribution. (L3)

UNIT - III

Multiple Discriminant Analysis and Logistic Regression: Discriminant model and analysis: a two group discriminant analysis, a three group discriminant analysis, the decision process of discriminant analysis, assumptions and estimation of the model, assessing overall fit of a model, interpretation and validation of the results; Logistic Regression model and analysis: regression with a binary dependent variable, representation of the binary dependent variable, estimating the logistic regression model, assessing the goodness of fit of the estimation model, testing for significance of the coefficients, and interpreting the coefficients. 10 hours

Learning outcomes :

At the end of this unit, the student will be able to

- apply discriminant analysis and logistic regression for evaluation of associations between various covariates and a categorical outcome. (L5)
- Design logistic regression models to predict probability of outcome variables in relation to several independent variables. (L6)

UNIT - IV

Principal Component Analysis (PCA): Population and sample principal components, their uses and applications, large sample inferences, graphical representation of principal components, Biplots, the orthogonal factor model, dimension reduction, estimation of factor loading and factor scores, interpretation of factor analysis. 12 hours

Learning outcomes :

At the end of this unit, the student will be able to

- Understand use of PCA for data reduction without losing its properties.(L2)
- Apply PCA to overcome the dimensionality of the real world problems. (L3).
- interpret associations among variables and recognizing variables responsible for distribution. (L5)

UNIT – V

Cluster Analysis and Multidimensional Scaling : Concepts of cluster analysis and multidimensional scaling, similarity measures, hierarchical clustering methods, Ward’s hierarchical clustering method’s, non-hierarchical clustering methods, K-means methods. Clustering based on statistical models. 10 hours

Learning outcomes

After completion of the unit, the student will be able to

- Understand the concept and application of cluster analysis. (L2)
- Interpret the cluster analysis output obtained from the statistical software. (L5)
- Apply multidimensional scaling to quantify similarity judgements (L3)
- Interpret the multidimensional scaling output obtained from the statistical software (L5)

Course Outcomes:

Upon completion of this course, student will be able to

- gain knowledge of the basic concepts underlying the most important multivariate techniques
- apply different techniques of multivariate statistics to get solutions for business problems in diverse disciplines
- describe properties of multivariate normal distribution.
- use principal component analysis effectively for data exploration and data dimension reduction.
- find groupings and associations using cluster analysis.
- Use of statistical software packages such as R for solving problems

Text Books:

1. Multivariate Statistics Made Simple: A Practical Approach (1st ed. BY)Sarma, K.V.S., & Vardhan, R.V. (2018).. Chapman and Hall/CRC .
2. An Introduction to Statistical Learning: With Applications in R by James, Gareth, Daniela Witten, Trevor Hastie, and Robert Tibshirani, 2013.
3. Applied Multivariate Statistical Analysis by Hardly W.K. and Simor L., Richard A. Johnson and Dean W. Wichern, , Prentice hall India, 7th Edition, 2019.

M.Sc DATA SCIENCE

SEMESTER - III

Generic Elective – I

CSCI7031 COMPUTATIONAL BIOLOGY

Hours per week: 4	End Examination: 60Marks
Credits:4	Sessionals: 40 Marks

Preamble:

A large number of prokaryotic and eukaryotic genomes completely sequenced. Mining the genomic information requires the use of sophisticated computational tools. It covers major databases and software programs for genomic data analysis, with an emphasis on the theoretical basis and practical applications of these computational tools.

- Gene and protein sequence acquisition, storage, retrieval and analysis
- Protein structure and function relationship using computational tools
- Development of computational applications for processing of biological data
- Modeling and simulation of biological systems.

UNIT – I

Bioinformatics: Introduction, Goal, Scope, Applications, Limitations, New Themes.
Introduction to Biological Databases: Database and Types of Databases, Biological Databases, Pitfalls of Biological Databases, Information Retrieval from Biological Databases - GENBANK, National Centre for Biotechnology Information, European Bioinformatics Institute. (7 hours)

After completion of this unit, student will be able to

- know what bioinformatics is and why it is important.(L1)
- how bioinformatics data is stored and organized.(L1)
- learn different types of data found at the NCBI and EBI resources.(L2)
- understand how to locate and extract data from key bioinformatics databases and resources.(L2)
- Know the difference between databases, tools, repositories and be able to use each one to extract specific information.(L3)

UNIT – II

Pairwise Sequence Alignment: Evolutionary Basis, Sequence Homology versus Sequence Similarity, Sequence Similarity versus Sequence Identity, Methods, Scoring Matrices, Statistical Significance of Sequence Alignment. **Database Similarity Searching:** Unique Requirements of Database Searching, Heuristic Database Searching, Basic Local Alignment Search Tool (BLAST), FASTA, Comparison of FASTA and BLAST, Database Searching with the Smith– Waterman Method. **Multiple Sequence Alignment:** Scoring Function, Exhaustive Algorithms, Heuristic Algorithms. (10hours)

Learning Outcomes:

After completion of this unit, student will be able to

- Extract and generate pairwise sequence alignments for a protein sequence of interest.L3).
- interpret the metrics used to assess the quality of a pairwise sequence alignment, identity versus similarity.(L2)
- identify mutations between two sequences using pairwise sequence approach.(L3)
- Outline the principles of the BLAST algorithm.(L2)

- Assess the relationships between the protein sequences from the different organisms based on the multiple sequence alignment.(L5)

UNIT – III

Profiles and Hidden Markov Models: Position-Specific Scoring Matrices, Profiles, Markov Model and Hidden Marko Model. **Protein Motifs and Domain Prediction**: Identification of Motifs and Domains in Multiple Sequence Alignment, Motif and Domain Databases Using Regular Expressions, Motif and Domain Databases Using Statistical Models, Protein Family Databases, Motif Discovery in Unaligned Sequences, Sequence Logos. (8 hours)

Learning Outcomes:

After completion of this unit, student will be able to

- Identify possible conserved protein domains and amino acids.(L2)
- Learn various statistical representations of motifs.(L2)

UNIT – IV

Gene Prediction: Categories of Gene Prediction Programs, Gene Prediction in Prokaryotes and Eukaryotes, Promoter and Regulatory Elements in Prokaryotes and Eukaryotes, Prediction Algorithms. **Phylogenetics Basics:** Molecular Evolution and Molecular Phylogenetics, Terminology, Gene Phylogeny versus Species Phylogeny, Forms of Tree Representation, Why Finding a True Tree is Difficult. **Phylogenetic Tree Construction Methods and Programs:** Distance-Based Methods, Character-Based Methods, Phylogenetic Tree Evaluation, Phylogenetic Programs. (10hours)

Learning Outcomes:

After completion of this unit, student will be able to

- Understand concepts of phylogenetic distance & how models of evolution are used. (L2)
- To build phylogenies from pairwise distance matrices. (L3)
- Learn the properties of phylogenetic trees, methods to optimize the topology and branch lengths of a tree. (L3)

UNIT – V

Protein Structure Basics: Amino Acids, Peptide Formation, Dihedral Angles, Hierarchy, Secondary Structures, Tertiary Structures, Determination of Protein Three-Dimensional Structure, Protein Structure Database, Protein Structure Visualization, Comparison and Classification. **Protein Secondary Structure Prediction:** Secondary Structure Prediction for Globular Proteins, Transmembrane Proteins, Coiled Coil Prediction. (8 hours)

Learning Outcomes:

After completion of this unit, student will be able to

- Outline the different levels and organization of protein structures.(L2)
- Describe how protein structures are determined.(L2)
- Predict the structure for a protein sequence based on an identified template.(L6)
- Elaborate the criteria for assessing and refining a predicted protein structure. (L6)

Course Outcomes:

Upon completion of this course, student will be able to:

- Acquire knowledge about biological data bases.(L1)
- Learn Gene prediction methods.(L4)
- Understands Sequence Homology versus Sequence Similarity.(L3)
- Learn about Multiple Sequence Alignment.(L4)
- Understand different structures of protein. (L2)

Text Books:

1. Essential Bioinformatics by Jin Xiong, Cambridge University Press,2006.
2. Bioinformatics and Functional Genomics by Jonathan Pevsner, Wiley Publications, 2nd edition, 2009.

Reference Books:

1. Introduction to Bioinformatics by Lesk, A.M., Oxford UniversityPress,4th Edition, 2014.

2. Bioinformatics: A practical guide to the Analysis of Genes and Proteins by Andreas Baxevanis, B.F. Francis Ouellette, Wiley Publications, 2nd Edition,2001.

M.Sc DATA SCIENCE

SEMESTER – III

GENERIC ELECTIVE -I

CSCI7041 WEB PROGRAMMING

Hours per week: 4	End Examination: 60Marks
Credits:4	Sessionals: 40 Marks

Preamble:

This course enables the students to associate with developing websites for hosting via intranet or internet. The web development process includes web design, web content development, client- side scripting, server-side scripting. Web development is the coding or programming that enables website functionality as per the owner's requirements.

- Design static web pages using Markup languages.
- Design and implement web applications using stylesheets.
- Use of java script for designing web applications with dynamic effects.
- Validations on form input entry and adding dynamic content to web applications.
- The notions of Web servers and Design Methodologies with MVC Architecture.
- Creation of adaptive web pages and implementing cookies.
- Design and implementation of complete applications over the web.

UNIT - I

Overview of HTML5 and Other Web Technologies: Introduction to Internet, Web and Web technologies, HTML5 and its Essentials, New Features of HTML5**,** Structuring an HTML Document - Elements and Attributes , Tags, The DOCTYPE Element, Exploring Editors and Browsers Supported by HTML5, Creating, Saving, Validating ,Viewing a HTML Document, Hosting Web Pages. **Fundamentals of HTML**: Understanding Elements, Describing Data Types, Horizontal Rules, Line Breaks, Paragraphs, Citations, Quotations, Definitions, Comments, Working with Text, Organizing Text in HTML, Exploring Hyperlinks, URL, Understanding and Describing the Table Elements, Inserting Images, Exploring Colors. **Working with Forms**: Exploring the FORM element, Types of INPUT Element, Exploring Button, Multiple Choice, Text Area, Label, Fieldset, Legend, Datalist, Keygen, Output elements, submitting a Form. (10 hours)

After completion of this unit, student will be able to

- Understand various steps to design static websites.(L2)
- Identify the importance of HTML tags for designing webpage.(L3)
- Able to develop a static web page along with user interactive elements.(L5)

UNIT – II

Working with Multimedia: Exploring Audio and Video File Formats, Describing the Multimedia Elements, defining a Multimedia File Using the EMBED, OBJECT Element, Exploring the FIGURE and FIGCAPTION Elements. **Overview of CSS**: Evolution, Syntax of CSS, Exploring CSS selectors, Inserting CSS in a HTML Document, Exploring Background, Color, Font Properties of a Webpage, Properties Table: Using the style Attribute, Creating Classes and IDs, Generating External Style Sheets, Typography, Consistency, Types of styles, Specifying class within HTML document, Style placement: Inline style, Span & div tags, header styles, Text and font attributes: Font Vs CSS, changing fonts, text attributes, Advance CSS properties:
Backgrounds, Box properties and Positioning. (8 hours)

After completion of this unit, student will be able to

- Separate design from content using various levels of StyleSheets.(L2)
- Learn different types of style sheets.(L2)

UNIT – III

Java Script: Features, Using Java Script in a HTML Document, Exploring Programming Fundamentals of JavaScript, Strings, Exploring Functions, Events, Image Maps, Animations.
(10 hours)

Learning Outcomes:

After completion of this unit, student will be able to

- Use Java script to validate user input and perform dynamic documents.(L3)
- Design dynamic and interactive web pages by embedding Java script code in HTML.(L4)

UNIT – IV

PHP: Introducing PHP, History, Unique Features, Basic Development Concepts, Creating First PHP Script, Mixing PHP with HTML, Escaping Special Characters, Using Variables and Operators, Controlling Program Flow, Working with Arrays. **File Handling in PHP**: File operations like opening, closing, reading, writing, appending, deleting etc. on text and binary files,
listing directories. (10hours)

After completion of this unit, student will be able to

- Understands the components of PHP.(L2)
- Learn the basic constructs of PHP, built in functions.(L2)
- Learn the importance of PHP for web application development. (L3)

UNIT – V

Introducing JSON: History of JSON, JSON vs XML, Typical Uses of JSON, JSON DATA Structures, JSON Syntax, Data Types, Creating JSON Objects, Parsing JSON, JSON Data Persistence, Data Interchange, Cross Origin Resources, Posting JSON, Working with templates, JSON with PHP. (8hours)

After completion of this unit, student will be able to

- Learn the features of JSON. (L2)
- Understand the difference between JSON and XML, JSON and RDBMS.(L3)
- Learn JSON Format to serialize and transmit structured data over the internet.(L3)

Course Outcomes:

Upon completion of this course, student will be able to:

- Demonstrate the importance of HTML & DHTML tags for designing web pages and separate design from content using Cascading Style Sheet.(L2)
- Understand various steps to design dynamic websites. (L2)
- Design interactive web pages with client and sserver-sidescripting. (L3)
- Apply validations on user input using javascript(L3)
- Understands the PHP framework and develops reusable component.(L2)
- Apply JSON for storing information in an organized manner.(L3)

Text Books:

1. HTML 5 Black Book , CSS 3, Java Script, XML, XHTML, AJAX, PHP and JQuery by DT Editorial Services, , Dream Tech Press, 2nd Edition,2016.
2. PHP: A Beginner’s Guide by Vikram Vaswani, Tata McGraw Hill,2017.
3. JSON for Beginners by Icode Academy,2107.

Reference Books:

1. HTML5 and CSS3 by Elizabeth Castro & Bruce Hyslop, , Pearson, 7th Edition, 2012
2. HTML 5 in simple steps by Kogent Learning Solutions, Dream Tech,2010.
3. Programming PHP by Rasmus Lerdorf and Kevin Tatroe, Oreilly Publication, 1st edition, 2002
4. Beginning JSON by Ben Smith, Apress publisher, 1st Edition,2015.

M.Sc DATA SCIENCE
SEMESTER – III
GENERIC ELECTIVE -I
CSCI7051 DATA SECURITY AND PRIVACY

Hours per week: 4 End Examination: 60Marks

Credits:4 Sessionals: 40 Marks

Preamble:

Data security refers to protecting digital privacy measures that are applied to prevent unauthorized access to computers, databases and websites. Data security also protects data from corruption. Data security is an essential aspect of IT for organizations of every size and type. Data security is also known as Information Security (IS) or Computer Security.

- To learn the basic concepts related to data security and understand the different types of symmetric key ciphers.
- To understand the concepts of encryption standards.
- To understand the concepts of asymmetric key cryptography and hash functions.
- To learn the basic concepts of hiding data in text and images.
- To understand the concepts of privacy, authentication, web and email security.

Introduction: Security goals, Cryptographic Attacks, Services and Mechanism, Techniques. **Traditional Symmetric Key Ciphers**: Introduction, Substitution Ciphers, Transposition Ciphers, Stream and Block Ciphers.

Introduction to Modern Symmetric-Key Ciphers: Modern Block Ciphers, Modern Stream

By the end of the unit the student will be able to

- Explain different security goals.(L2)
- Develop substitution and transposition ciphers.(L3)
- Describe concepts of symmetric key ciphers.(L2)
- Explain concepts of modern block ciphers.(L2)
- Extend the concept of modern stream ciphers.(L2)

Data Encryption Standard (DES): Introduction, DES Structure, DES Analysis, Security of DES, Multiple DES-Conventional Encryption Algorithms.

Advanced Encryption Standard (AES): Introduction, Transformations, Key Expansion, AES Ciphers, Analysis of AES. (9hours)

By the end of the unit the student will be able to

- Outline the structure of DES.(L2)
- Illustrate the analysis of DES. (L3)
- Explain the concept of AES. (L2)
- Identify the need of key expansion.(L1)
- Illustrate the analysis of AES. (L3)

UNIT - III

Asymmetric-Key Cryptography: Introduction, RSA Cryptosystem, Rabin Cryptosystem, Elgamal Cryptosystem, Elliptic Curve Crypto systems.

Cryptographic Hash Functions: Introduction, Iterated Hash function, SHA-512, WHIRLPOOL.

Digital Signature: Comparison, Process, Services, Attacks on Digital Signature, Digital Signature Standard. (10hours)

Learning Outcomes:

By the end of the unit the student will be able to

- Explain different types of cryptosystems.(L2)
- Identify necessity of a HASH function.(L1)
- Illustrate the use of cryptographic hash functions.(L3)
- Identify different types of attacks on digital signature.(L1)
- Extend the concept of digital signature standard.(L3)

Data Hiding in Text: Basic Features, Applications of Data Hiding, Watermarking, Intuitive Methods, Simple Digital Methods, Data Hiding in Text, Innocuous Text, Mimic Functions.

Data Hiding in Images: LSB Encoding , BPCS Steganography, Lossless Data Hiding, Spread Spectrum Steganography, Data Hiding by Quantization, Patchwork , Signature Casting in Images, Transform Domain Methods, Robust Data Hiding in JPEG Images, Robust Frequency Domain Watermarking, Detecting Malicious Tampering. (10 hours)

By the end of the unit the student will be able to

- Identify the need for data hiding. (L1)
- Illustrate different types of data hiding techniques. (L3)
- Describe the concepts of steganography. (L2)
- Explain the concepts of data hiding in images. (L2)

UNIT - V

Privacy: Privacy Concepts, Privacy Principles and Policies, Authentication and Privacy, Data Mining, Privacy on the Web, E-Mail Security, Impacts on Emerging Technologies.

Legal and Ethical Issues in Computer Security: Protecting Programs and Data, Information and the Law, Rights of Employees and employers, Redress for Software Failures, Computer crime, Ethical Issues in Computer Security. (8hours)

Learning Outcomes:

By the end of the unit, the student will be able to

- Understand the basic concepts of privacy. (L2)
- Identify the need of email security. (L1)
- Illustrate software failures. (L3)
- Describe the concepts of computer crime. (L2)
- Explain the concepts of ethical issues in computer security. (L2)

Course Outcomes:

Upon completion of this course, student will be able to:

- Understand the need of computer security. (L2)
- Identify the differences between different types of ciphers.(L4)
- List the concepts of block ciphers and stream ciphers. (L2)
- Able to differentiate between DES and AES algorithms. (L3)
- Learn Asymmetric cryptography, cryptographic hash functions , various features of digital signature (L2)
- List the concepts of data hiding. (L2)
- Understand different Data Hiding Techniques on Text, Images. (L3)
- Demonstrate various Privacy, Legal and Ethical issues in computer security. (L2)

Text Books:

1. Cryptography and Network Security by Behrouz A. Forouzan, Dedeep Mukhopadhyay, TMH, 2nd edition, 2013. (Unit I , II,III)
2. Data Privacy and Security by Salomon, David, Springer, 2003. (Unit IV only)
3. Security in Computing by Charles Pfleeger, Shari Lawrence Pfleeger, 5th Edition, PHI, 2015. (Unit V only)

1. Information Security: Principles and Practice by Mark Stamp, Wiley Inter Science,2011.
2. Computer Security: Art and Science by Matt Bishop, First Edition, Addison Wesley,2002.
3. Cryptography and Network Security by William Stallings, Pearson Education,7th edition,2017.

M.Sc DATA SCIENCE
SEMESTER- III
GENERIC ELECTIVE – II
MATH7011 TIME SERIES AND FORECASTING ANALYSIS

Hours per week: 4 **End Examination: 60 Marks**
Credits: 4 **Sessionals: 40 Marks**

Course objectives:

At the end of the course, the student should be able to

- Compute and interpret a correlogram and a sample spectrum
- Derive the properties of autoregressive integrated moving average (ARIMA) and state-space models
- Choose an appropriate ARIMA model for a given set of data, fit the model and interpretation of results
- Compute forecasts for a variety of linear methods and models.
- Use R studio to perform time series analysis

UNIT – I

Introduction to time series: Stationery stochastic processes. The autocovariance and Auto correlation functions and their estimation. Standard errors of autocorrelation estimates. Bartlett's approximation (without proof). The periodogram, the power spectrum and spectral density functions. Link between the sample spectrum and autocorrelation function (ACF) Partial autocorrelation function (PACF).

10 hours

Learning outcomes:

At the end of the unit, the student should be able to

- Describe and verify mathematical considerations for analysing time series (L1)
- Understand the concepts of white noise, stationarity, autocovariance, autocorrelation etc.(L2)
- Apply correlogram and periodogram concepts to analyse time series sets and interpret results (L5)
- Use R programming concepts to perform time series analysis (L4)

UNIT – II

Linear Stationary Models: Two equivalent forms for the general linear process. Auto-covariance generating function and spectrum, stationarity and invertibility conditions for a linear process. Autoregressive and moving average processes, Spectrum for AR processes up to two Moving average (MA) process, stationarity and Invertibility conditions. ACF and PACF for M.A. (q), spectrum for MA processes up to order; Duality between autoregressive and moving average processes, Mixed AR and MA (ARMA) process. Stationarity and invertibility properties. The ARMA(1,1) process and its properties.
10 hours

Learning outcomes:

At the end of the unit, the student should be able to

- Understand the Auto-covariance generating function and spectrum, stationarity and invertibility conditions for a linear process. (L1)
- Apply various techniques of time series models, including the seasonal autoregressive moving etc. (L3)
- Identify relevant average (SARIMA) models, regression with ARMA models to apply in timeseries data analysis (L3)

- Use R programming concepts to perform time series analysis (L5)

UNIT – III

Linear Non-Stationary Models: Autoregressive integrated and moving average (ARIMA) processes. The three explicit forms the ARIMA models (viz) Difference equation, random shock and inverted forms. Model Identification–Stages in the identification procedures. Use of autocorrelation and partial autocorrelation, functions in identification. Standard errors for estimated autocorrelation and partial autocorrelations. Initial estimates MA, AR and ARMA processes and residual variance. 10 hours

Learning outcomes:

At the end of the unit, the student should be able to

- Understand the Linear Non-Stationary Models and ARIMA three explicit forms in detail. (L2)
- apply various techniques of time series models, including the ARIMA models for the real time data sets. (L5)
- Estimate MA, AR and ARMA processes and residual variance for timeseries data(L4)

UNIT – IV

Model Estimation: Least squares and Maximum likelihood estimation and interval estimation of parameters; Model Diagnostic checking – checking the stochastic model diagnostic checks applied to residuals. 8 hours

Learning outcomes:

At the end of the unit, the student should be able to

- Find Maximum likelihood and interval estimation of parameters for the for fitted models (L4)
- check model diagnostics stochastic model diagnostic checks applied to residuals. (L4)
- Use R programming concepts to perform time series analysis(L5)

UNIT - V

Time series Forecasting: Principles of forecasting, Minimum mean square error forecasts and their properties, derivation of the minimum mean square error forecasts, calculating and updating forecasts at any lead time.

10 hours

Learning outcomes:

- analyze any time series data using various forecasting approaches (L4)
- produce reasonable forecast values (L5)
- make concise and valid decisions based on forecasts obtained (L6)

Course Outcomes:

After completing this course, the student should be able to

- Get an idea about Time series Forecasting with principles to solve the real time problems
- Compute forecasts for a variety of linear methods and models Data science related problems
- Analyze any time series data using various statistical approaches to generate robust forecast values, and to make concise decisions based on forecasts obtained
- Use R programming concepts to perform time series analysis

Text Books :

1. Time Series Analysis by Box and Jenkins.

2. Forecasting Methods and Applications by Markidakis, S., Wheelwright, S.C. and Hydman, R.J. (1998). 3rd Edition, John Wiley &amp; Sons. Inc., Hoboken.

Reference books:

1. Time Series Analysis by Anderson, T.W.

2. Time Series : Theory and Methods (Second Edition) by Brockwell,P.J., and Davis,R.A.:.Springer–Verlag.

3.) Time Series Analysis and its applications, with examples in R by Shumway &amp; Stoffer (2011 3rd edition, Springer.

M.Sc DATA SCIENCE

SEMESTER - III

Generic Elective - II

CSCI7061 DATA STORAGE TECHNOLOGIES AND NETWORKING

Hours per week:4 End Examination: 60Marks

Credits:4 Sessionals: 40 Marks

Preamble:

This course fill the knowledge gap in understanding varied components of modern information storage infrastructure, including virtual environments. It provides comprehensive learning of storage technology, which will enable one to make more informed decisions in an increasingly complex IT environment. This course builds a strong understanding of underlying storage technologies and prepares one to learn advanced concepts, technologies, and products.

- To learn various storage infrastructure components
- To provide a strong understanding of storage related technologies.
- To learn the architectures, features and benefits of intelligent storage systems.
- To understand various storage networking technologies.

UNIT - I

Introduction to Information Storage and Management: Information Storage, Evolution of Storage Architecture, Data Center Infrastructure, Virtualization and Cloud Computing.

Data Center Environment: Application, DBMS, Host, Connectivity, Storage, Disk Drive Components, Disk Drive Performance, Host Access to Data, Direct Attached Storage, Storage Design based on Application, Introduction to Flash Drives. (8 hours)

Learning Outcomes:

After completion of this unit, student will be able to

- Learn the key elements of a data center environment. (L2)
- Understands the importance of Virtualization. (L2)
- Learn different types of data storage Systems. (L3)

UNIT - II

DataProtection RAID: RAID Implementation methods, RAID Array Components, RAID Techniques, RAID Levels, RAID Comparison, RAID Impact on Disk Performance, Hot Spares.

Intelligent Storage Systems: Components of an Intelligent Storage Provisioning, Types of Intelligent Storage Systems. (8hours)

Learning Outcomes:

After completion of this unit, student will be able to

- Identify the importance of RAID technology. (L3)
- Understand fundamental constructs and various RAID Levels. (L2)
- Learn the Components and types of Intelligent Storage Systems. (L2)
- Understand the benefits of Intelligent Storage Systems.(L3)

UNIT - III

Storage Networking Technologies: Fiber Channel Storage Area Network: Overview, the SAN and its Evolution, Components of FC SAN, FC Connectivity, Fiber Channel Architecture, FC SAN Topologies, Virtualization in SAN. **Network Attached Storage:** General Purpose Servers vs NAS Devices, Benefits of NAS, File Systems and Network File Sharing, Components of NAS, NAS I/O operation, NAS Implementations, NAS File-Sharing Protocols. (10 hours)

Learning Outcomes:

After completion of this unit, student will be able to

- Learn how FC SAN reduces overall operational cost and downtime.(L1)

- Understand how Virtualization minimizes resource management complexity and cost.(L1)
- Identify appropriate storage infrastructure.(L3)

UNIT - IV

Object Based and Unified Storage: Object Based Storage Devices, Content Addressed Storage, CAS Use Cases, Unified Storage. **Introduction to Business Continuity:** Information Availability, BC Terminology, BC Planning Lifecycle, Failure Analysis, Business Impact Analysis, BC Technology Solutions. (8hours)

Learning Outcomes:

After completion of this unit, student will be able to

- Learn how Object based storage manage storing unstructured data.(L1)
- Learn the components of unified storage and the processes of accessing data.(L2)
- Understand the goals of business continuity plan.(L3)
- Learn BC framework.(L2)

UNIT-V

Backup And Recovery: Backup - purpose, Considerations, Granularity, methods, Architecture, Backup and Restore Operations, Backup Topologies, Backup in NAS Environments, Backup Technologies. **Local Replication:** Replication Terminology, Uses of Local Replicas, Replica Consistency, Local Replication Technologies. (10 hours)

Learning Outcomes:

After completion of this unit, student will be able to

- Learn the need of backup, backup methods, technologies and implementations.(L2)
- Understand different backup topologies and backup in virtualized environment.(L3)
- Understand local replication process and uses of local replica.(L2)
- Learn various local replication technologies. (L2)

Course Outcomes:

Upon completion of this course student will be able to

- Understand how to manage the capacity, performance, and reliability of large numbers of disks. (L3)
- Learn how Intelligent Storage Systems provide highly optimized I/O processing capabilities. (L2)
- Understands importance of NAS and identify how NAS improves the performance. (L3)
- Apply to organizations for an effective and cost-efficient disaster recovery and restart procedures in both physical and virtual environments. (L4)

Text Book:

1. Information Storage and Management by EMC Education Services, 2nd Edition 2012.

Reference Books:

1. Storage Area Network Essentials by Richard Barker, Paul Massiglia, Wiley 1st Edition, 2008.
2. Storage Networks – Complete Reference by Robert Spalding,TMH,2003.
3. Building Storage Networks by Marc Farley, TMH,2001.

M.Sc DATA SCIENCE

SEMESTER - III

Generic Elective – II

CSCI7071 NATURAL LANGUAGE PROCESSING

Hours per week: 4 End Examination: 60Marks
Credits:4 Sessionals: 40 Marks

Preamble:

This course provides an introduction to the computational modelling of natural language.

Course Objectives:

- Acquaintance with natural language processing and learn how to apply basic algorithms in this field.
- To recognize the significance of pragmatics for natural language understanding.
- Capable of describing the application based on natural language processing and to show the points of syntactic, semantic and pragmatic processing.

UNIT – I

Regular Expressions, Text Normalization, Edit Distance: Regular Expressions, Words, Corpora, Text Normalization, Minimum Edit Distance. **N-Gram Language Models:** N-grams, Evaluating Language Models, Generalization and Zeros, Smoothing, Kneser-Ney Smoothing, The web and stupid Backoff, Advanced Perplexity's Relation to Entropy. (6 hours)

Learning Outcomes:

After completion of this unit, student will be able to

- Learn features of NLP.(L1)
- Learn pattern matching methods using different types of regular expressions.(L2)
- Understand the concepts of morphology, syntax, semantics and pragmatics of the language.(L3)
- Understand the applications of NLP. (L3)

UNIT – II

Parts of Speech Tagging: English Word Classes, The Penn Tree bank part of speech Tagset, Part of Speech tagging, HMM part of speech tagging, Maximum Entropy Markov Models, Bi-directionality, Part of Speech tagging for other languages. (6 hours)

Learning Outcomes:

After completion of this unit, student will be able to

- Learn basics of English language.(L2)
- Learn basic structure of English sentence and its syntax. (L2)
- Understand the complexity of English language and hence techniques of English language processing.(L3)
- Understand the elements and applications of Part-of-speech tagging. (L3)

UNIT – III

Formal Grammars of English: Constituency, Context Free Grammars, Some Grammar Rules for English, Tree banks, Grammar Equivalence and Normal Form, Lexicalized Grammars. **Syntactic Parsing:** Ambiguity, CYK Parsing, Partial parsing. (8 hours)

Learning Outcomes:

After completion of this unit, student will be able to

- Understand approaches to syntax and semantics in NLP.(L3)
- Learn different types of grammars. (L2)
- Learn different types of parsing techniques.(L2)

UNIT – IV

Dependency Parsing: Dependency Relations, Formalisms, Treebank, Transition Based Dependency Parsing, Graph based dependency parsing, Evaluation.

Representation of Sentence Meaning: Computational Desiderata for Representations, Model – Theoretic Semantics, First Order Logic, Event and State Representations, Description Logics.

(8 hours)

Learning Outcomes:

After completion of this unit, student will be able to

- Understand the way to parse a sentence, recognize its syntactic structure, and construct representation of meaning.(L2)
- Provide the student with knowledge of various levels of analysis involved in NLP.(L4)

UNIT – V

Semantic Parsing : Information Extraction: Named Entity Recognition, Relation Extraction, Extracting Times, Events and their times, Template Filling.

Lexicons for Sentiment, Affect and Connotation: Defining Emotion, Available Sentiment and Affect Lexicons, Creating affect lexicons by human labeling, semi supervised induction of affect lexicons, supervised learning of word sentiment, Using lexicons for Sentiment Recognition.

(10 hours)

Learning Outcomes:

After completion of this unit, student will be able to

- Understand approaches to syntax and semantics in NLP.(L3)
- Presents an introduction to the computational modelling of natural language and identifying the sentiment of the text.(L3)
- Building robust systems to perform linguistic tasks with technological applications.(L4)

Course Outcomes:

Upon completion of this course, student will be able to:

- Understand the features and basic concepts of NLP (L2)
- Demonstrate pattern matching methods using Regular expressions. (L2)
- Understand Sequence labeling Algorithms. (L2)
- Identify the difference between types of parsing. (L3)
- Understand the representation of sentence and perform analysis. (L3)
- Recognize and categorize the sentiment. (L3)

Text Book:

1. Speech and Language Processing- Daniel Jurafsky, James H Martin, 2nd edition, PHI, 2008.

   1. Natural Language Processing using Python by Steven Bird, Ewan Klien, EdwardLoper, 1st edition, Oreilly Publications,2009.

M.Sc DATA SCIENCE
SEMESTER - III
Generic Elective - II
CSCI7081 FUNDAMENTALS OF BLOCK CHAIN TECHNOLOGIES

Hours per week: 4	End Examination: 60Marks
Credits: 4	Sessionals: 40Marks

Preamble:

This is new technology of digital currency. Block chains are to achieve decentralization. T*he system needs to validate transactions without anyone being able to veto transactions or control the network.*

Course Objectives:

- Learn the basic concept of Cryptographic Hash Functions, Hash Pointers and Elliptic Curve Digital Signature Algorithm.
- A technical overview of decentralized digital currencies like Bitcoin, as well as their broader economic, legal and financial context.
- To get an insight into the working of the Bitcoin network, Wallet, Bitcoin mining and distributed consensus for reliability.

UNIT – I

Introduction to Cryptography: Cryptographic Hash Functions, SHA-256, Hash Pointers and Data Structures, Merkle tree. (8hours)

Learning Outcomes:

After completion of this unit, student will be able to

- Learn the basics of hash functions.(L1)
- Identify the importance of each hash function and understand the underlying data structures used.(L3)

UNIT – II

Digital Signatures: Elliptic Curve Digital Signature Algorithm (ECDSA), Public Keys as identities, A Simple Cryptocurrency. (8hours)

Learning Outcomes:

After completion of this unit, student will be able to

- Learn the importance of digital signature.(L2)
- Understand digital signature algorithms.(L2)
- Learn the mechanism of simple crypto currency.(L2)

UNIT – III

Centralization vs Decentralization, Distributed consensus, Consensus without identity using a block chain, Incentives and proof of work.

Mechanics of Bitcoin: Bitcoin Transactions, Bitcoin Scripts, Applications of Bitcoin Scripts, Bitcoin Blocks, The Bitcoin Network. (10hours)

Learning Outcomes:

After completion of this unit, student will be able to

- Understand the structure of a blockchain.(L1)
- Learn why it is better than a simple distributed database.(L2)
- Learn the underlying principles and techniques associated with blockchain technologies. (L3)
- Familiar with the cryptographic building blocks.(L3)
- Understand typical Cryptocurrency such as Bitcoin.(L2)

UNIT – IV

Storage and Usage of Bitcoins: Simple Local Storage, Hot and Cold Storage, Splitting and Sharing Keys, Online Wallets and Exchanges, Payment Services, Transaction Fees, Currency Exchange Markets. (10hours)

Learning Outcomes:

After completion of this unit, student will be able to

- Learn different ways of storing Bitcoin keys, security measures.(L2)
- Understand various types of services that allow you to trade and transact with bitcoins. (L3)

UNIT – V

Bitcoin Mining: The Task of Bitcoin miners, Mining Hardware, Mining pools, Mining incentives and strategies.

Bitcoin and Anonymity: Anonymity Basics, Mixing, Zerocoin and Zerocash.

Applications of Block Chain Technologies. (8 hours)

Learning Outcomes:

After completion of this unit, student will be able to

- Learn how Bitcoin relies on mining. (L1)
- Who are the miners?.(L1)
- How they get into this and operate.(L1)
- Learn Business model for miners. (L2)
- Impact of business model on the environment.(L3)
- The various ways to improve Bitcoin's anonymity and privacy.(L3)

Course Outcomes:

Upon completion of this course, student will be able to

- Learn individual components of the Bitcoin protocol make the whole system tick.(L2)
- Learn the methods of security from a combination of technical methods and clever incentive engineering.(L2)
- Analyze the incentive structure in a blockchain-based system and critically assess its functions, benefits and vulnerabilities. (L4)

Text Books:

1. Bitcoin and Cryptocurrency Technologies: A Comprehensive Introduction by Arvind Narayanan, Joseph Bonneau, Edward Felten, Andrew Miller and Steven Goldfeder, Princeton Press, 2016.

1. Mastering Bitcoin: Programming the Open Blockchain by Andreas M. Antonopoulos Shroff, O'Reilly; 2nd Edition, 2017.

M.Sc DATA SCIENCE
SEMESTER - III
Generic Elective - II
CSCI7091 WEB ANALYTICS

Hours per week: 4 End Examination: 60 Marks
Credits:4 Sessionals: 40 Marks

Preamble:

Web analytics is a way of learning how users interact with websites by automatically recording aspects of the user's behavior and transforming the behavior into data that can be analyzed. It is the measurement and analysis of data to inform an understanding of user behavior across web pages.

- To learn web analytics from a strategic and practical perspective.
- Explore different types of analytics and why they are important for business.
- Learn various web analytics processes and metrics used to measure.
- Learn techniques using Google Web analytics, traffic analysis, click path analysis and segmentation.

UNIT – I

Web Analytics Present and Future: A brief history of web analytics, Current Landscape and Challenges, Traditional Web analytics, Web analytics Activities, Measuring, Trinity. **Datacollection:** Understanding the Data Landscape, Click stream Data, Outcomes Data, Research Data, Competitive Data. (6hours)

Learning Outcomes:

After completion of this unit, student will be able to

- Lay the foundational ground work on the approaches to web analytics.(L2)
- Understand the critical importance of various data collection mechanisms.(L2)
- Focus on qualitative data- why and what the available options.(L2)
- Significantly elevate the ability to listen to the customers.(L3)

Introduction to Web Analytics: Definition, User Experience and Web Analytics Questions.
Web Analytics Approach: Introduction, A model of Analysis, Showcasing the work, Context Matters, Contradicting the data.
Working of Web Analytics: Introduction, Log File Analysis, Page Tagging, Metrics and Dimensions, Interacting with data in Google Analytics. (8 hours)

After completion of this unit, student will be able to

- Understand foundational concepts in web analytics.(L1)
- Learn the analysis process.(L2)
- Learn the importance of viewing data and balancing the desire for perfection.(L2)
- Understand how analytic tools work.(L2)
- Understand how analytic tools organize and segment data.(L2)

UNIT – III

Goals: Introduction, Definition of Goals and Conversions, Conversion Rate, Goal Reports in Google Analytics, Finding the right things to measure as key, Performance Indicators, Measure on a website that can constitute a goal. **Learning about users:** Introduction, Visitor Analysis. (6 hours)

Learning Outcomes:

After completion of this unit, student will be able to

- Identify analytics goals and conversion rates.(L3)
- Learn the key performance indicators to calibrate the web site goals.(L4)
- Identify user's geographical location, technology used and visiting frequency.(L3)

UNIT –IV

Traffic Analysis: Introduction, Source and medium, Organic Search, Search Query Analysis, Referral Traffic, Direct Traffic, Paid Search Keyword.

Analyzing usage of content: introduction, Website content Reports. (8 hours)

Learning Outcomes:

After completion of this unit, student will be able to

- Study the way users actually get to the websites.(L2)
- Analyze the key words user types in search engines.(L3)
- Articulate the information needs and categorize users.(L4)
- Learn analysis metrics used to analyze usage content.(L2)

UNIT – V

Click-Path Analysis: Introduction, Focus on Relationships between pages, Navigation Summary, Visitors Flow Report, Analyzing how users move from one page type to another. **Segmentation:** Introduction, Necessity, Procedure to segment, Ways to Segment, Useful ways to segmentUX questions. (8hours)

After completion of this unit, student will be able to

- Learn Click-Path analysis. (L1)
- Examine the relationships between pairs of pages.(L3)
- Learn methods of filtering data.(L2)
- Analyze the behavior of users.(L3)

Course Outcomes:

Upon completion of this course student will be able to

- Understand the Challenges, activities of web analytics along with types of data used for web analytics (L2)
- Understand the process of analysis and working of tools
- Learn the goals of web analytics and identify the role of web site in organizing and structuring the use of web analytics (L3)
- Analyze and categorize the users (L3)
- Understand the methods used for learning user behavior and filtering techniques.(L2)

Text Books:

1. Web Analytics an hour a Day by Avinash Kaushik, Sybex, 1st Edition, 2007. (Unit –I)
2. Practical Web Analytics for User Experience by Michael Beasley, Morgan Kaufmann, 1st Edition, 2013. (Unit II to UnitV)

Reference Books:

1. Web Analytics 2.0: The Art of Online Accountability and Science of Customer Centricity by Avinash Kaushik, 1st edition, Sybex,2009.
2. Web Data Mining: Exploring Hyperlinks, Content and Usage Data by Bing Liu, 2nd Edition, Springer,2011.
3. Google Analytics by Justin Cutroni, O'Reilly, 2010.

M.Sc DATA SCIENCE
SEMESTER - III
CSCI7101 DEEP LEARNING LAB

Hours per week: 2 Continuous Evaluation: 100 Marks
Credits:2

Preamble : The **Deep Learning** is a foundational program that will help you understand the capabilities, challenges, and consequences of deep learning and prepare you to participate in the development of leading-edge AI technology. The practical knowledge in Deep Learning Lab provides a pathway for students to take the definitive step in the world of AI by helping them gain the knowledge and skills to level up their career.

Course Educational Objectives:

- ToUnderstandthe conceptof Sequential Deep Neural Networkanditslearningprocess
- To understand how to improve Deep Neural Networks: Hyperparameter Tuning, Regularization and Optimization
- To build and train neural network architectures such as Convolutional Neural Networks,
- Understand and implement Recurrent Neural Networks, LSTMs, Transformers,
- To learn how to make them better with strategies such as Dropout, BatchNorm

1. Introduction to Tensorflow
   - install Tensorflow
   - understand Tensorflow library.

   Check if the following libraries are installed
   - scipy • numpy • matplotlib • pandas • statsmodels • scikit-learn
2. Compute the function using Tensorflow library. $f(x,y) = x^2 + y^2 + 2x + y$

   Find Expected Results, Check Tensorflow was installed correctly, Evaluate a function using Tensorflow, View the Tensor graph.
3. Implementation of AND gate using Tensorflow
4. Deep Neural Network : implement a deep neural network using Tensorflow and Keras. • train the DNN with image 2D or 3D dataset.
5. Deep Neural Network with Regularization
   Implement a deep neural network using Tensorflow and Keras.
   • train the DNN with image 2D or 3D dataset.• add regularization
6. Deep Neural Network with Dropout
   implement a deep neural network using Tensorflow and Keras.
   • train the DNN with different dataset.. • add dropout neurons.
7. Deep Neural Network with Early stopping
   implement a deep neural network using Tensorflow and Keras.
   • train the DNN with image 2D or 3D dataset.
   • implement early stopping
8. Build convolutional neural network model (CNN) (Basic model) forimage and other 2D or 3D data
9. Evaluate the model using five-fold cross-validation and Develop an Improved Model
10. Develop a model for Text classification with an RNN

Course Outcomes:

Upon completion of this course student will be able to

- The learners will gain a comprehensive understanding of TensorFlow, its installation, and its library. They will be able to verify the successful installation of TensorFlow and compute functions using the TensorFlow library.

- The students will understand how to model and implement logical operations, like the AND gate, using TensorFlow. This will demonstrate how basic logical functions can be modeled and tested using neural networks.

- Participants will learn how to implement, train, and test a deep neural network using TensorFlow and Keras. They will gain experience working with 2D or 3D image datasets, improving their understanding of neural networks in image processing applications.

- Learners will further enhance their DNNs by implementing advanced strategies like regularization to reduce overfitting, dropout to improve generalization, and early stopping to optimize training time. This will help students understand and utilize these strategies to improve their neural network models.

- The students will learn to construct CNN models for handling 2D or 3D data such as images, implementing text classification with RNNs, and using methods like five-fold cross-validation for model evaluation. This will expose them to a broad range of neural network architectures and their applications.

SEMESTER - III

CSCI7111 BIG DATA ANALYTICS LAB

Hours per week: 2 Continuous Evaluation: 100 Marks

Credits:2

Preamble

The Big Data Analytics Lab is a practical ,hands on course designed to immerse students in the tools and technologies central to big data analysis .Students engage with technologies such as Hadoop and Distributed File System ,Mapreduce, Hbase, Sqoop ,Mahout, Hive ,pig AND Spark .The course involves exploring HDFS,writing MapReduce programs ,managing tables in HBase,importing data with Sqoop,getting recommendations with Mahout,managing data in Hive,Scripting wth PigLatin and Understanding Spark Shell .

Course Educational Objectives

1. Students should gain a deep understanding of the Hadoop ecosystem and HDFS, the backbone of most big data projects. This includes understanding the file system's architecture, commands, and operations.
2. Gain practical experience in writing and running MapReduce jobs, a critical component of processing data in the Hadoop ecosystem. This includes understanding how to create Combiners and Partitioners to optimize data processing tasks.
3. Learn to perform fundamental operations in HBase and Hive, two essential tools for working with structured data in the Hadoop ecosystem. This includes creating tables, inserting and scanning data in HBase, as well as loading data into Hive, executing queries, and partitioning data.
4. Learn to import data from relational databases using Sqoop, write PigLatin scripts for data manipulation, and use Mahout for generating recommendations. These tools play a significant role in data ingestion, processing, and analysis in a big data environment.
5. Get hands-on experience with the Spark Shell, understand the use of RDDs, and perform operations using the PySpark framework. This will equip students with knowledge about in-memory data processing, which offers faster and more flexible alternatives to MapReduce.

List of Experiments

1. Exploring Hadoop Distributed File System (HDFS). Implementation of file system commands in HDFS.
2. Understanding Map Reduce Jobs:
   a. Writing a MapReduce Program
   b. Running a MapReduce Job.
   c. Writing and Implementing a Combiner
   d. Writing a Partitioner
3. Using the HBase Shell perform basic table functions view the results of each operation:
   a. Creating a table,
   b. Putting rows into table
   c. Scanning a table
   d. Manually flush and compact a table
4. Write an HBase program that creates a table, puts several rows, scans the table and

outputs the column values.

5. Importing Data with Sqoop: import data from a relational database using Sqoop.
6. Creating a Mahout Recommender: Use of Mahout to generate recommendations for users based on the data imported using Sqoop.
7. Loading Data into Hive:
    a. Using the dump file present in the HDFS file system.
    b. Using Sqoop to import the table into Hive from MySQL.
8. Executing Hive Queries.
9. Partitioning and Bucketing Data in Hive.
10. Managing an HBase Table with Hive: Create a table in HBase and then create an External Table in Hive which points to the HBase table.
11. Reading and Writing Data using Pig using Grunt Shell and perform the following operations:
    a. DESCRIBE the table.
    b. Use DUMP to print the records you loaded.
    c. Use STORE to save the records in a file. The resulting file should be comma-delimited.
    d. Using a terminal or the Grunt shell, inspect the results.
12. Writing PigLatin Scripts for the following tasks:
    a. Write a Pig Latin program to read u.user and find the users who are female and scientists. Store the results in a file.
    b. What is the average age of the users in u.user? Hint: use the AVG function. Dump the result to the screen.
    c. Create a script using a text editor. It should load the u.user file, find the 3 most common occupations and store the results in HDFS. Execute the script using "pig -f script"
13. Write your Pig script. It should work on all the files in a directory and return a list of the 20 most common words in the directory. Hint: the TOKENIZE function will be useful
14. Understanding Spark Shell-
    a. Create a Spark Session,
    b. Inspect Spark Session using following Commands-
        i. Retrieve Spark Session Version,
        ii. Return Spark Application name, Retrieve Spark Application Id,
        iii. Check and Return Minimum Number of Partitions,
15. Create an RDD which contains key value pairs using parallelize method and Perform the following operations on RDD using pyspark:
    a. List number of partitions RDD consists,
    b. Count the instances of RDD,
    c. Count the instances of RDD by key,
    d. Count the instances of RDD by value,
    e. Calculate Sum of RDD elements,
    f. Check whether RDD is empty or not.
16. Create an RDD which contains range of 100 numbers and perform the following operations on RDD using pyspark-
    a. Find the minimum value in the RDD elements
    b. Find the maximum value in the RDD elements

c. Find the Mean, Median & standard deviation of RDD elements

17. Word Count using RDD- Write a pyspark program for the following

    d. Create a Spark Session
    e. Read the text file using RDD API
    f. Perform transformations (Map & flat Map) on the text file to count the number of words are there in the file.
    g. Save the outcome as a file.

18. Dataframe API Operations

Textbook:

1. Big Data Black Book by Dt Editorial Services, Dreamtech Publications, 2016.
2. Hadoop The Definitive Guide by Tom White, O'reilly, 4thEdition, 2016.
3. Programming Hive- Jason Rutherglen, Dean Wampler, Edward Capriolo, O'reilly Publisher, 1st edition, 2012.

Course Outcomes

Upon completion of the course, the student is able to

1. Students will develop a comprehensive understanding of the Hadoop ecosystem, including the usage and operation of the Hadoop Distributed File System (HDFS).

2. Students will gain practical experience in writing, running, and debugging MapReduce jobs. This will include knowledge of how to optimize MapReduce computations using combiners and partitioners.

3. Students will be able to create and manage tables in HBase, perform data manipulations in Hive, and integrate these tools effectively in the context of a big data project.

4. Students will develop the ability to import data from relational databases using Sqoop, write PigLatin scripts for data analysis, and use Mahout for generating user recommendations. This set of skills will enable students to handle various data processing and analysis tasks in a big data environment.

5. Students will gain a robust understanding of the Spark framework, including its operations, capabilities, and application in big data analytics. This includes the ability to create and manipulate Resilient Distributed Datasets (RDDs) using PySpark, which is critical for efficient data processing in Spark.

M.Sc DATA SCIENCE

SEMESTER - III

CSCI7121 NDUSTRIAL TRAINING AND SEMINAR

Hours per week: 2

Continuous Evaluation: 100 Marks

Credits:2

SEMESTER -IV

CSCI7131 Project Work

Hours per week: 3

Credits: 8

End Examination: 50 Marks

Sessionals: 150 Marks

GITAM School of Science

GITAM (Deemed to be Universtiy)

Visakhapatnam | Hyderabad | Bengaluru