Fergusson College (Autonomous) Pune

Learning Outcomes-Based Curriculum

for

M.Sc. Data Science

With effect from July 2019

1. Introduction

Data science combines the knowledge of mathematics, computer science and statistics to solve exciting data-intensive problems in industry and in many fields of science. Data scientists help organisations make sense of their data. As data is collected and analysed in all areas of society, demand for professional data scientists is high and will grow higher.

2. Nature and Extent of M.Sc. Data Science

The M.Sc. Data Science program will provide a unique opportunity to students to obtain skills specially designed for data science stream. It also develops attitude and interest along with necessary skills among the students to encourage them to do research and work in industry. This programme enables students to work on problems specific to various domains with the help of data science techniques.

3. Aims of the Master's programme in Data Science

The objective is to provide technology-oriented students specialized in data science stream with the capability in various areas of data science and business domains too. It helps students to develop skills needed to deal effectively within the areas of data science. The course includes topics in statistical and exploratory analysis, data formats and languages, processing of massive data sets, management of data. The course focuses on overall growth of students and enhance their knowledge in specific domain areas of their interest.

It is a full-time course of Two year and four semesters in which the last semester will be Industrial training.

4. Characteristics attributes of a Post Graduate in Data Science

(a) Disciplinary knowledge and skills – The course provide an opportunity to develop skill-based expertise based on the subject interest of the student.

(b) Skilled Communicator – Assessment and pre-placement training are structured to enhance their communication skills that orients them to communicate information, ideas, problems and solutions to both specialist and non-specialist audiences;

(c) Critical thinker and problem solver - Provides a basis or opportunity for originality in developing and applying ideas, often within a project context; can apply their knowledge and understanding, and problem-solving abilities in new or unfamiliar environments within broader or multidisciplinary perspective related to their field of study.

(d) Sense of inquiry – Assessment is also based on increasing their sense of inquiry by providing a platform to demonstrate their ideas in new upcoming computer science topics through project development and seminars.

(e) Team player/ Skilled Project Manager – Students are encouraged to perform their project in groups to develop their team spirit and handle the peer pressures.

(f) Lifelong learners - have the learning skills to allow them to continue to study in a manner that may be largely self-directed or autonomous.

(g) One of the most demanding programmes in academics and industry.

5. Qualification Descriptors for a Master's programme in Data Science

- To apply their knowledge and understanding to develop applications/algorithms to solve problems in field of data science.
- A wide-ranging knowledge and practical skills in the analysis, design and implementation of software systems in response to application needs and organizational environment.
- A deep and systematic understanding of the academic discipline of Data Science.
- To develop those learning skills which are necessary for students to continue to undertake further study.

6. Programme learning outcomes relating to M.Sc Data Science

- Specialized knowledge of the central concepts, theories, and research methods of data science as well as applied skills.
- Specialized knowledge of computer science theories, methods, practices and strategy.
- Understanding of statistical, mathematical concepts in the context of data science.
- Understanding of various analysis tools and software used in data science.
- Awareness to the rapid technological changes.
- Teamwork and leadership skills through projects.
- Analytical and critical thinking skills.
- Creative thinking skills.
- Time management and organization skills.
- Written and oral communication skills, including presentations and report writing

7. Course Learning Outcomes (Course/ paper wise)

- Understand and analyse the lifecycle of data through application building.
- Apply the major theories in the field of data analysis and data exploration to some characteristic problems.
- Plan a data science project on various application areas using knowledge of the data lifecycle and analysis process.
- Investigate, analyse, document and communicate the core issues and requirements in developing data analysis capability in a global organisation.
- Demonstrate an understanding of data science to a level of depth and sophistication consistent with senior professional practice.
- Review and evaluate data science projects.
- Review, synthesise, apply and evaluate contemporary data science theories through either a significant research thesis component or research-grounded industrial project.

Programme Structure

Year	Course Code	Course Title	Credits
First (Semester - I)	CSD4101	Probability and Statistics	4
	CSD4102	Applied Linear Algebra	4
	CSD4103	Data Structures	4
	CSD4104	Database Management System	4
	CSD4105	Data Science Practical - I (R Programming)	4
	CSD4106	Data Science Practical - II (Data Structures and RDBMS)	4
First (Semester - II)	CSD4201	Statistical Inference	4
	CSD4202	Mathematical Foundation	4
	CSD4203	Machine Learning	4
	CSD4204	Design and Analysis of Algorithms **OR**	4
	CSD4205	Soft Computing **OR**	
	CSD4206	MOOCS-I	
	CSD4207	Data Science Practical - III (Machine Learning using R)	4
	CSD4208	Data Science Practical - IV (Python for Data Science)	4
Second (Semester-III)	CSD5301	Optimization Techniques	4
	CSD5302	Big Data Engineering	4
	CSD5303	Deep Learning	4
	CSD5304	Data Science Case Studies **OR**	4
	CSD5305	Artificial Intelligence **OR**	
	CSD5306	MOOCS-II	
	CSD5307	Data Science Practical - V (Deep Learning)	4
	CSD5308	Data Science Practical – VI (Project)	4
Second (Semester-IV)	CSD5401	Industrial Training (Full-Time Internship with minimum 8 hours per day from Monday to Friday)	8
Total Credits			**80**

Extra Credit Courses

Groups	Particulars	No. of Credits
I	Human Rights Awareness Course (Semester-I)	02
II	Cyber Security Awareness Course (Semester-II)	02
III	Cyber Security Awareness Course (Semester-III)	02
IV	Skill Component Courses (from Semester-I to Semester-IV) • **From any of the following:** (a) Departmental skill component courses: 04 credits (b) Entrepreneurship Development course: 03 credits (c) Participation in Summer/ Winter school / Hands-on-Training programmes (duration not less than 02 weeks): 02 credits (d) Research paper presentation at State / National level: 02 credit (e) Research paper presentation at International (overseas) level: 03 credits (f) Working / undertaking mini project under various schemes at College level: 02 credits (g) Participation in Avishkar research festival: 02 credits Selection in Avishkar at University Level: 03 credits Avishkar Winner at State Level: 04 credits (h) Participation in cultural and cocurricular activities / competitions: At State level: 02 credits Participation in cultural and cocurricular activities / competitions at National level: 02 credits	04

Class: First Year M.Sc. Data Science- I, Semester-I

Course Code: CSD4101

Title of the Course/ Paper: Probability and Statistics

Credits: 4

Learning Outcomes	**Suggested Pedagogical Processes**
Learning to describe basic features of the data in a study.	Black board teaching along with ICT
Provide brief summary about the sample using different quantitative measures.	Provide classroom assignments
To fit predictive models for the sample data	
Develop analytical thinking by using the ability to see a problem or solution from different points of view.	
To find chance of an event based on prior knowledge of conditions that might be related to the event.	
To apply different forms of probability distribution when the values of observed data are discrete and continuous.	Black board teaching along with ICT

Note: Following listed concepts should be explained and executed using large datasets with the help of R

Unit No.	Title of Unit and Contents
I	**Descriptive Statistics:** 1.1 Measures of Central Tendency: Mean, Median, Mode 1.2 Partition Values: Quartiles, Percentiles, Box Plot 1.3 Measures of Dispersion: Variance, Standard Deviation, Coefficient of variation 1.4 Skewness: Concept of skewness, measures of skewness 1.5 Kurtosis: Concept of Kurtosis, Measures of Kurtosis (All topics to be covered for raw data using R software. Manual calculations are not expected.)
II	**Introduction to Probability**: 2.1 Probability - classical definition, probability models, axioms of probability, probability of an event. 2.2 Concepts and definitions of conditional probability, multiplication theorem P(A∩B) =P(A).P(B|A) 2.3 Bayes' theorem (without proof) 2.4 Concept of Posterior probability, problems on posterior probability. 2.5 Definition of sensitivity of a procedure, specificity of a procedure. Application of Bayes' theorem to design a procedure for false positive and false negative. 2.6 Concept and definition of independence of two events. 2.7 Numerical problems related to real life situations.

III	**Introduction to Random Variables** 3.1 Definition of discrete random and continuous random variable. 3.2 Concept of Discrete and Continuous probability distributions. (p.m.f. and p.d.f.) 3.3 Distribution function 3.4 Expectation and variance 3.5 Numerical problems related to real life situations
IV	**Special Distributions** 4.1 Binomial Distribution 4.2 Uniform Distribution 4.3 Poisson Distribution 4.4 Negative Binomial Distribution 4.5 Geometric Distribution 4.6 Continuous Uniform Distribution 4.7 Exponential Distribution 4.8 Normal Distribution 4.9 Log Normal Distribution 4.10 Gamma Distribution 4.11 Weibull Distribution 4.12 Pareto Distribution (For all the probability distributions its pmf/pdf, p-p plot, q-q plot, generation of probabilities and random samples using R software is expected. **)**
V	**Correlation and Linear Regression** 5.1 Bivariate data, Scatter diagram. 5.2 Correlation, Positive Correlation, Negative correlation, Zero Correlation 5.3 Karl Pearson's coefficient of correlation (r), limits of r ($-1 \leq r \leq 1$), interpretation of r, Coefficient of determination (r^2) 5.4 Meaning of regression, difference between correlation and regression. 5.5 Fitting of line Y = a+bX 5.6 Concept of residual plot and mean residual sum of squares. 5.7 Multiple correlation coefficient, concept, definition, computation and interpretation. 5.8 Partial correlation coefficient, concept, definition, computation and interpretation. 5.9 Multiple regression plane. 5.10 Identification and solution to Multicollinearity 5.11 Evaluation of the Model using R square and Adjusted R square All topics to be covered for raw data using R software. Manual calculations are not expected.
VI	**Logistic Regression** 6.1 Introduction to logistic regression 6.2 Difference between linear and logistic regression 6.3 Logistic equation 6.4 How to build logistic regression model in R 6.5 Odds ratio in logistic regression.

Learning Resources:

1. Fundamentals of Applied Statistics (3rd Edition), Gupta and Kapoor, S.Chand and Sons, New Delhi, 1987.
2. An Introductory Statistics, Kennedy and Gentle.
3. Statistical Methods, G.W. Snedecor, W.G. Cochran, John Wiley & sons, 1989.

4. Introduction to Linear Regression Analysis**,** Douglas C. Montgomery, Elizabeth A. Peck, G. Geoffrey Vining, Wiley
5. Modern Elementary Statistics, Freund J.E., Pearson Publication, 2005.
6. Probability, Statistics, Design of Experiments and Queuing theory with applications Computer Science, Trivedi K.S., Prentice Hall of India, New Delhi,2001.
7. A First course in Probability 6th Edition, Ross, Pearson Publication, 2006.
8. Introduction to Discrete Probability and Probability Distributions, Kulkarni M.B., Ghatpande S.B., SIPF Academy, 2007.
9. A Beginners Guide to R, Alain Zuur, Elena Leno, Erik Meesters, Springer, 2009
10. Statistics Using R, Sudha Purohit, S.D.Gore, Shailaja Deshmukh, Narosa, Publishing Company

Class: First Year M.Sc. Data Science- I, Semester-I

Course Code: CSD4102

Title of the Course/ Paper: Applied Linear Algebra

Credits: 4

Learning Outcomes	**Suggested Pedagogical Processes**
Understand the concept of eigenvalues and eigenvectors	Black board teaching along with ICT
Acquire the knowledge of various concepts in Applied Algebra	
Awareness of linear junction models	

Unit No.	Title of Unit and Contents
I	**Vectors** Vector: Vector addition, Scalar Vector multiplication, Inner Product, Complexity of Vector Computations Linear Functions: Linear Functions, Taylor Approximation, Regression Model Norms and Distance: Norm distance, Standard deviation, Angle, Complexity Clustering: Clustering, a clustering Objective, The K means algorithm, Examples and Applications Linear Independence: Linear Dependence, Basis, Orthonormal Vectors, Gram Smith algorithm
II	**Matrices** Matrices: Introduction to Matrices, Zero and identity Matrices, Transpose, addition and norm, Matrix Vector Multiplication, Complexity Matrix Examples: Geometric Transformation, Selectors, Incidence Matrix and Convolution Linear Equations: Linear and affine functions, Linear function models, System of Linear Equations Matrix Multiplication: Matrix Multiplication, Composition of Linear Functions, Matrix Power and QR Factorization Matrix Inverses: Left and right inverses, Inverse, Solving Linear Equations, Examples, Pseudo Inverse

III	**Least Squares** Least Squares: Least Squares Problem, Solution, Solving Least Squares Problems, Examples Least squares data fitting: Least Squares data fitting, Validation, Feature Engineering. Least Squares Classification: Classification, Least Squares Classifier, Multiclassifiers Multi Objective Least Squares: Multi Objective Least Squares, Control, Estimation and Inversion, Regularised data fitting, Complexity Constrained Least Squares: Constrained Least Squares problem, Solution, Solving constrained Least Squares problems Constrained Least Squares Applications: Portfolio Optimization, Linear Quadratic control, Linear Quadratic State Estimation

Learning Resources:

1. Introduction to Applied Linear Algebra Vectors, Matrices and Least Squares by Stephen Boyd (Stanford University) and Lieven Vandenberghe (University of California, Los Angeles) Cambridge University Press

Class: First Year M.Sc. Data Science- I, Semester-I
Course Code: CSD4103

Title of the Course/ Paper: Data Structures

Credits: 4

Learning Outcomes	Suggested Pedagogical Processes
To learn the concept of object oriented programming.	Chalk and talk method is used to discuss the concept, differentiate between OOP and procedure oriented languages.
To understand the concept of ADT	Black board method along with class room discussion to be conducted. With problem solving mechanism, let student realize concept of ADT
To make student aware of the concept of algorithm	Class room discussion to learn the concept of algorithm. Black board method will be used to understand how the goodness of the algorithm is measured. Problem solving method is realize the space and time complexity of any algorithm
To understand the concept of generic class	Using power point presentation, the concept of generic programming has to be explained.
To learn the concept of linked list, its types and applications	Black board method is used to discuss the concepts and programming exercises has to be taken to have deep understanding of the implementation details.

To understand the concept of stacks and queues, their applications and implementations	Classroom discussion along with animated power point presentation is used to discuss the concept. Hands on Programming are considered to realize the applications of the data structure.
To learn the concept of hash table, dictionaries	Different application has to be discussed to let student aware of the concept of key value pair structures. Programming exercises are given to explore the implementation details.
To understand the concept of priority queues, heap and their applications	Power point presentations will be used to discuss the concept. Problem solving session will be considered to understand the application domains.
To learn the concept of search trees, their types, various operations to be performed and make aware with various applications	Power point presentations are used to discuss the basic concept and the operations to be performed. To understand the various applications, programming exercise is used.

Unit No.	Title of Unit and Contents
I	**Introduction to OOP** 1.1 Concept, Benefits and Application of OOP
II	**ADT** 2.1 Abstract Data Types and the C++ Class 2.2 An Introduction to C++ Class- Data Abstraction and Encapsulation in C++ 2.3 The Array as an Abstract Data Type 2.4 The Polynomial Abstract Data type- Polynomial Representation- Polynomial Addition.
III	**Algorithms** 3.1 Performance analysis- time complexity and space complexity 3.2 Templates in C++, Template Functions- Using Templates to Represent Container Classes
IV	**Linked lists** 4.1 Linear lists 4.4.1 Single Linked List and Chains, 4.4.2 Representing Chains in C++ 4.4.3 Designing a Chain Class in C++ 4.4.4 Chain Manipulation Operations, 4.4.5 The Template Class Chain 4.4.6 Implementing Chains with Templates 4.4.7 Chain Iterators 4.4.8 Chain Operations 4.2 Circular List 4.3 Doubly Linked Lists 4.4 Skip list 4.5 Generalized Lists

	4.5.1 Representation of Generalized Lists 4.5.2 Recursive Algorithms for Lists 4.5.3 Reference Counts, Shared and Recursive Lists
V	**Stacks and Queues** 5.1 The Stack Abstract Data Type, 5.2 The Queue Abstract Data Type 5.3 Evaluation of Expressions, Parenthesis Matching 5.4 Implementation of recursion 5.5 Expression - Infix, Postfix, prefix conversions 5.6 Linked Stacks and Queues 5.7 Implementation using template classes in C++.
VI	**Dictionaries and Hash Table** 6.1 Dictionaries 6.2 Hash table 6.2.1. Representation 6.2.2. Hash functions 6.2.3. Collision resolution-separate chaining, open addressing-linear probing, quadratic probing, double hashing, rehashing, extendible hashing 6.2.4. Comparison of hashing and skip lists.
VII	**Priority Queues**: 7.1 Definition 7.2 ADT 7.3 Realizing a Priority Queue using Heaps - Definition, insertion, Deletion 7.4 External Sorting- Model for external sorting, Multiway merge, Polyphase merge.
VIII	**Search Trees** 8.1 Binary Search Trees 8.1.1 Definition 8.1.2 ADT 8.1.3 Implementation 8.1.4 Operations- Searching, Insertion and Deletion 8.2 AVL Trees, 8.2.1 Definition 8.2.2 Height of an AVL Tree 8.2.3 Operations - Insertion, Deletion and Searching 8.3 B-Trees 8.3.1 Definition 8.3.2 B-Tree of order m 8.3.3 Height of a B-Tree 8.3.4 Operations- insertion, deletion and searching 8.3.5 Comparison of Search Trees

Learning Resources:

1. Knuth, D. E. The Art of Computer Programming, Vol. I & III, Addison-Wesley, 1974.
2. Carrano, F. M., Data Abstraction and Problem Solving with C++, Benjamin Cummings, 1995.

3. Horowitz, E., Sahni, S. and Mehta, D., Fundamentals of Data Structures in C++, W.H. Freeman, 1995.
4. Standish, T. A., Data Structures, Algorithms and Software Principles in C, Addison-Wesley, 1995.
5. Tenenbaum, A. M. , Langsam, Augenstein, M. J., Data Structures Using C++, Prentice Hall, 1996.
6. D. Samantha : Classic Data Structures,PHI2002

Class: First Year M.Sc. Data Science- I, Semester-I

Course Code: CSD4104

Subject: Database Management System

Credits: 4

Learning Outcomes	**Suggested Pedagogical Processes**
Understand basics of DBMS	Traditional teaching method using blackboard –chalk method, question answer method
Learn structure of DBMS	Use of ICT tools like presentation
Understand functions of DBMS	Use of ICT tools like presentation
Acquire the knowledge of different data models	Use of ICT tools like presentation
Know basics of E-R Concepts	ICT tools and Case Studies
Understand the fundamentals of Normalization	Presentation
Study SQL : basic structure, Aggregate functions, Simple queries, nested queries	Case studies

Unit No.	Title of Unit and Contents
I	**Introduction** 1.1 Database-system Applications 1.2 Purpose of Database Systems 1.3 View of Data-Data Abstraction, Instance and Schemas 1.4 Relational Databases: Tables, DML, DDL 1.5 Data storage and querying: Storage Manager, The query processor 1.6 Database Architecture 1.7 Speciality Databases
II	**Introduction to Relational Model** 2.1 Structure of Relational Databases 2.2 Database Schema 2.3 Keys 2.4 Relational Operations
III	**Introduction to SQL** 3.1 Overview of SQL query language 3.2 SQL data Definition- Basic Types, Basic schema definition, Date and Time in SQL, Default values, Index creation, Large Object types, user-

	defined types 3.3 Integrity constraint- Constraints on a single relation, Not Null constraint, Unique constraint, The Check clause, referential integrity 3.4 Basic structure of SQL queries- Queries on single relation, queries on multiple relations, The natural join, 3.5 Additional basic operations 3.6 Set operations 3.7 Null Values 3.8 Aggregate Functions-Basic aggregation, Aggregation and grouping, The Having clause, Aggregation with Null and Boolean values 3.9 Nested subqueries- Set membership, Set comparison, Test for Empty Relations, Test for Absence of Duplicate Tuples, Subqueries in the From clause, The **with** clause, Scalar subqueries 3.10 Modification of the Database- Deletion, Insertion, Updates
IV	**Intermediate and advanced SQL** 4.1 Join Expressions- Join conditions, Outer joins, Join types and conditions 4.2 Views- View definition, using views in SQL queries, Materialized views, update a view 4.3 Create table extensions 4.4 Schemas, Catalogs and Environments 4.5 The relational Algebra 4.6 The tuple relational calculus
V	**Database Design and E-R model** 5.1 Overview of the Design process and Entity Relationship Model 5.2 Constraints and Removing Redundant Attributes in Entity Sets 5.3 Entity Relationship Diagrams 5.4 Introduction to UML Relational database model: Logical view of data, keys, integrity rules 5.5 Functional Dependency 5.6 Anomalies in a Databases 5.7 The normalization process: Conversion to first normal form, Conversion to second normal form, Conversion to third normal form, The Boyce-Codd Normal Form (BCNF), Fourth Normal form and fifth normal form 5.8 Normalization and database design 5.9 Denormalization
VI	**Introduction to NoSQL and Graph Database** 6.1 Overview of NoSQL 6.2 Comparison of relational databases to new NoSQL stores 6.3 Types and examples of NoSQL Databases

Learning Resources:

1. Abraham Silberschatz, Henry F. Korth, S. Sudarashan, Database System Concepts, McGraw-Hill International Edition, Sixth Edition
2. Elmasri, Navathe, Fundamentals of Database Systems, Pearson Education, Third Education
3. Ramakrishnan, Gehrke, Database Management Systems, McGrawHill International Edition, Third Edition
4. Peter Rob, Carlos Coronel, Database System Concepts, Cengage Learning, India Edition
5. S.K.Singh, “Database Systems Concepts, Design and Applications”, First Edition, Pearson Education, 2006
6. Redmond,E. & Wilson, Seven Databases in Seven Weeks: A Guide to Modern Databases and the NoSQL Movement Edition:1st Edition.

Class: First Year M.Sc. Data Science- I, Semester-I

Course Code: CSD4105

Subject: Data Science Practical –I (R Programming)

Credits: 4

Learning Outcomes	**Suggested Pedagogical Processes**
Introduce the basics of R Studio	Demonstration along with Hands-on session
Apply different functions for solving statistical problems	Hands – on assignments
Understand visualization in R	Hands – on assignments
Acquire the knowledge of spatial data and graph analysis	Hands – on assignments
Implementation of data manipulation techniques	Hands – on assignments

Lab Course in R Programming

Note: - Each Assignment will be based on following concepts

Assignment No.	Topics Covered
1	Introduction to R-studio, mathematical and logical operators in R, Data types and data structures, simple operations and programs, matrix operations
2	Data frames, string operations, factors, handling categorical data, lists and list
3	Operations Loops and conditional statements, switch and break function
4	Apply functions, Statistical problem solving in R,
5	Visualizations in R – 1
6	Visualizations in R – 2
7	Spatial Data Representation and Graph Analysis.
8	Hands-on data manipulations1: cleaning, sub-setting, sampling, data transformations and allied data operations
9	Hands-on data manipulations2: cleaning, sub-setting, sampling, data transformations and allied data operations
10	Case Study

Class: First Year M.Sc. Data Science- I, Semester-I

Course Code: CSD4106

Subject: Data Science Practical –II (Data Structures and RDBMS)

Credits: 4

Learning Outcomes	Suggested Pedagogical Processes
Acquaint with the concepts of object oriented programming	Hands-on on programming assignments
Implement various concepts of data structures	Hands-on on programming assignments
Designing E-R diagrams	Hands-on on programming assignments
Understand basics of SQL	Hands-on on programming assignments
Study various SQL commands	Hands-on on programming assignments
Using aggregate functions, Simple queries, nested queries and joins	Hands-on on programming assignments

Lab Course in Data Structures and RDBMS

Note:- Each Assignment will be based on following concepts

Assignment No.	Topics Covered
1	Polynomial ADT
2	Concept of linked list
3	Stack and Queues
4	Dictionaries and Hash Table
5	Priority queue and its applications
6	Search tree
7	Introduction to Databases and SQL, DDL and DML Commands
8	Simple queries and Nested queries
9	Joins
10	Views and Stored Functions
11	Case Study

Class: First Year M.Sc. Data Science- I, Semester-II

Course Code: CSD4201

Subject: Statistical Inference

Credits: 4

Learning Outcomes	Suggested Pedagogical Processes
To help make informed judgments based on a pattern of data observed previously.	Provide classroom assignments
To study how hypothesis ensures the entire research process remains scientific and reliable	
To test an assumption regarding population parameter using sample data.	
Study data related to time and predict its future behaviour.	
To study different models of forecasting.	

Unit No.	Title of Unit and Contents
I	**Sampling** 1.1 Introduction to Sampling 1.2 Simple random Sampling 1.3 Stratified Random Sampling 1.4 Cluster Sampling 1.5 Concept of Sampling Error
II	**Sampling Distributions** 2.1 Introduction to Sampling distributions 2.2 Student's t distribution 2.3 Chi square distribution 2.4 Snedecor's F distribution 2.5 Interrelations among t, chi-square and F distributions 2.6 Central Limit Theorem (Various Versions) and its applications.
III	**Testing of hypothesis** 3.1 Definitions: population, statistic, parameter, standard error of estimator. 3.2 Concept of null hypothesis and alternative hypothesis, critical region, level of significance, type I and type II error, one sided and two-sided tests, p-value. 3.3 Large Sample Tests 3.4 Tests based on t, Chi-square and F-distribution **All tests to be taught using R software. Manual calculations are not expected**.
IV	**Analysis of Variance** 4.1 One Way ANOVA 4.2 Two Way ANOVA 4.3 Application of ANNOVA to test the overall significance of Regression. **All topics to be covered using R software. Manual calculations are not expected**.

V	**Time Series** 5.1 Meaning and Utility. 5.2 Components of Time Series. 5.3 Additive and Multiplicative models. 5.4 Methods of estimating trend: moving average method, least squares method and exponential smoothing method. (single, double and triple) 5.5 Elimination of trend using additive and multiplicative models. 5.6 Simple time series models: AR (1), AR (2). **5.7** Introduction to ARIMA Modelling.

Learning Resources:

1. Fundamentals of Applied Statistics (3rd Edition), Gupta and Kapoor, S.Chand and Sons, New Delhi, 1987.
2. Time Series Methods, Brockell and Devis, Springer, 2006.
3. Time Series Analysis,4th Edition, Box and Jenkin, Wiley, 2008.
4. Modern Elementary Statistics, Freund J.E., Pearson Publication, 2005.
5. Probability, Statistics, Design of Experiments and Queuing theory with applications Computer Science, Trivedi K.S. ,Prentice Hall of India, New Delhi,2001.
6. Common Statistical Tests, Kulkarni M.B., Ghatpande S.B., Gore S.D., Satyajeet Prakashan,Pune, 1999.
7. Probability And Statistical Inference, 9th Edition, Robert Hogg, Elliot Tanis, Dale Zimmerman, Pearson education Ltd, 2015
8. A Beginners Guide to R, Alain Zuur, Elena Leno, Erik Meesters, Springer, 2009
9. Statistics Using R, Sudha Purohit, S.D.Gore, Shailaja Deshmukh, Narosa, Publishing Company

Class: First Year M.Sc. Data Science- I, Semester-II

Course Code: CSD4202

Subject: Mathematical Foundation

Credits: 4

Pre-requisite: Set theory

Unit No.	Title of Unit and Contents
I	**High Dimensional Space** Introduction, The Law of Large numbers, The Geometry of High Dimensions, Properties of a Unit Ball, Generating points uniformly from a unit Ball, Gaussians in Higher Dimensions, Random Projection and John Linden Strauss Theorem, Separating Gaussians, Fitting of Spherical Gaussian to Data.
II	**Best Fit Subspaces and Singular Value Decomposition** Introduction, Preliminaries, Singular Vectors, Singular Value Decomposition, Best Rank k Approximations, Left Singular Vectors, Power Method for Singular Decomposition, Singular Vectors and Eigen Vectors, Applications of Singular Value Decomposition to Centering Data, Principal Component Analysis, Clustering a Mixture of Spherical Gaussians, Ranking Documents and Web Pages , Discrete Optimization Problem.
III	**Random Walks and Markov Chains** Stationary Distribution, Markov Chains Monte Carlo Algorithm, Areas and Volumes Convergence of Random Walks in undirected graphs, Electrical Networks and Random Walks, Random walks on undirected graphs with unit edge weights, Random Walks in Euclidean Space, The Web as a Markov Chain
IV	**Machine Learning** Introduction, The Perceptron Algorithm, Kernel Functions, Generalising to new data, Overfitting, Illustrative Examples and Occam’s Razor with Applications to learning decision trees, Regularization: Penalising Complexity, Online Learning, Online to Batch Conversion, Support Vector Machine, VC Dimension, Strong and Weak Learning-Boosting, Stochastic Gradient Descent, Combining Expert Advice, Deep Learning, Semi Supervised Learning, Active Learning and Multi task Learning

Learning Resources:

1. Foundations of Data Science: Alvin Blum, John Hopcroft and Ravindran Kannan

Class: First Year M.Sc. Data Science- I, Semester-II

Course Code: CSD4203

Subject: Machine Learning

Credits: 4

Learning Outcomes	Suggested Pedagogical Processes
Select real-world applications that needs machine learning based solutions	ICT and Hands-on
Implement and apply machine learning algorithms	
Select appropriate algorithms for solving a particular group of real-world problems	
Recognize the characteristics of machine learning techniques that are useful to solve real-world problems	

Unit No.	Title of Unit and Contents
I	**Introduction to Data and Machine Learning** 1.1 Essentials of Data and its analysis 1.2 Framework of Data Analysis
II	**Machine Learning Basics** 2.1 History of Machine Learning 2.2 Machine Learning Vs Statistical Learning 2.3 Types of Machine Learning Algorithms 2.4 Supervised Learning 2.5 Unsupervised Learning 2.6 Reinforcement Learning
III	**Understanding Regression Analysis** 3.1 Linear Regression 3.2 Multiple Regression 3.3 Logistic Regression
IV	**Classification Techniques** 4.1 Decision Tree 4.2 SVM 4.3 Naïve Bayes 4.4 KNN
V	**Clustering** 5.1 K means clustering 5.2 Association Rule Mining 5.3 Apriori Algorithm
VI	**Model Evaluation** 6.1 Introduction 6.2 Performance Measures 6.3 Confusion Matrix
VII	**Ensemble Methods** 7.1 Introduction 7.2 Bagging, Cross Validation

Learning Resources:

1. Jiawei Han, Micheline Kamber, Jian Pei, Data Mining: Concepts and Techniques, 3rd Edition
2. Margaret H. Dunham, S. Sridhar, Data Mining - Introductory and Advanced Topics, Pearson Education 5. Tom Mitchell, Machine Learning‖, McGraw-Hill, 1997
3. R.O. Duda, P.E. Hart, D.G. Stork., Pattern Classification, Second edition. John Wiley and Sons, 2000.
4. Christopher M. Bishop, Pattern Recognition and Machine Learning, Springer 2006 8. Ian H. Witten, Data Mining: Practical Machine Learning Tools and Techniques, Eibe Frank Elsevier / (Morgan Kauffman)
5. Bing Liu: Web Data Mining: Exploring Hyperlinks, Contents and Usage Data, Springer (2006).
6. Soumen Chakrabarti: Mining the Web: Discovering knowledge from hypertext data, Elsevier (2003).
7. Christopher D Manning, Prabhakar Raghavan and Hinrich Schütze: An Introduction to Information Retrieval, Cambridge University Press (2009)

Class: First Year M.Sc. Data Science- I, Semester-II

Course Code: CSD4204

Title of the Course/ Paper: Design and Analysis of Algorithms

Credits: 4

Learning Outcomes	Suggested Pedagogical Processes
Analyse the asymptotic performance of algorithms	Use of ICT and Problem solving
Write rigorous correctness proofs for algorithms	
Demonstrate a familiarity with major algorithms and data structures	
Apply important algorithmic design paradigms and methods of analysis	
Provide analytical and problem-solving skills to design algorithms	

Unit No.	Title of Unit and Contents
I	**Introduction** Definition of Algorithm & its characteristics, Recursive and Non-recursive Algorithms, Time & Space Complexity, Definitions of Asymptotic Notations, Insertion Sort (examples and time complexity), Heaps & Heap Sort (examples and time complexity)
II	**Divide and Conquer** Concept of divide and Conquer, Binary Search (recursive), Quick Sort, Merge sort
III	**Greedy Method** Fractional Knapsack problem, Optimal Storage on Tapes, Huffman codes, Concept of Minimum Cost Spanning Tree, Prim’s and Kruskal's Algorithm

IV	**Dynamic Programming** The General Method, Principle of Optimality, Matrix Chain Multiplication, 0/1 Knapsack Problem, Concept of Shortest Path, Single Source shortest path, Dijkstra’s Algorithm, Bellman Ford Algorithm, Floyd- Warshall Algorithm, Travelling Salesperson Problem
V	**Branch & Bound** Introduction, Definitions of LCBB Search, Bounding Function, Ranking Function, FIFO BB Search, Traveling Salesman problem Using Variable tuple.
VI	**Decrease and conquer** Definition of Graph Representation, BFS, DFS, Topological Sort/Order, Strongly Connected Components, Biconnected Component, Articulation Point and Bridge edge
VII	**Problem Classification** Basic Concepts: Deterministic Algorithm and Non deterministic, Definitions of P, NP, NP-Hard, NP-Complete problems, Cook’s Theorem (Only Statement and Significance)

Learning Resources:

1. Fundamentals of Computer Algorithms, Authors - Ellis Horowitz, Sartaz Sahani, Sanguthevar Rajsekaran Publication: - Galgotia Publications

2. Introduction to Algorithms (second edition) Authors: - Thomas Cormen, Charles E Leiserson, Ronald L.Rivest ,Clifford Stein ,Publication: - PHI Publication

OR

Class: First Year M.Sc. Data Science- I, Semester-II

Course Code: CSD4205

Title of the Course/ Paper: Soft Computing

Credits: 4

Learning Outcomes	Suggested Pedagogical Processes
To make aware of soft computing concept.	Discuss the concept of hard computing and soft computing with student.
To help student understand various techniques used in soft computing and their applications.	Make them aware of various techniques and its applications using discussion and power point presentation.
To introduce the fuzzy logic concepts, Fuzzy principles and relations.	Using black board and practice session the concept is to be discussed.
To learn basics of ANN and Learning Algorithms	What is learning, types of learning, all concepts are put through using discussion and question answer methods.The power point presentations are used to study the ANN and its algorithms.
To understand how ANN can be used as function approximation technique.	Black board and discussion method is proposed to understand what function approximation is. And using power point presentation it is demonstrated how ANN can be used for function approximation.
To make student aware the range of applications ANN can work with	With the help of case studies the various problem domain has to be discussed
To learn the concept of Genetic Algorithm and its applications to soft computing.	Power point presentation to be used to understand the concept of genetic algorithm. Class room discussion about the application of genetic algorithm so that they could handle real world problems.
Hybrid system usage, application and optimization	Classroom discussion and power point methods are used to understand concept of hybrid system. Case studies are discussed to revise the concepts and understand how soft computing can be a tool for optimization

Unit No.	Title of Unit and Contents
I	**Introduction to Soft Computing** 1.1 What is soft computing 1.2 Principle of soft computing (SC Paradigm) 1.3 How is it different from hard computing? 1.4 Constituents of SC (Fuzzy Neural, Machine Learning, Probabilistic reasoning)
II	**Fuzzy Logic - Classical Sets and Fuzzy Sets** 2.1 Operations on Classical sets 2.2 Properties of classical sets 2.3 Fuzzy set operations 2.4 Properties of fuzzy sets: Cardinality, Operations
III	**Classical Relations and Fuzzy Relations** 3.1 Cartesian Product 3.2 Classical Relations-Cardinality, Operations, Properties, Composition 3.3 Fuzzy Relations - Cardinality, Operations, Properties, Composition, Max product
IV	**Membership functions** 4.1 Features of Membership Functions 4.2 Standard Forms and Boundaries 4.3 Fuzzification methods 4.4 Problems on Inference method of Fuzzification
V	**Fuzzy to Crisp conversions** 5.1 Fuzzy Tolerance and equivalence relations 5.2 Lambda (alpha) cuts for fuzzy sets and relations 5.3 Defuzzification methods: Max – Membership, Centroid, Weighted Average method, Mean-Max Membership, Center of Sums, Center of Largest Area, First of Maxima
VI	**Fuzzy Arithmetic and Fuzzy Numbers** 6.1 Fuzzy Arithmetic 6.2 Fuzzy numbers 6.3 Extension Principle
VII	**Logic and fuzzy systems** 7.1 Fuzzy Logic 7.2 Approximate Reasoning 7.3 Fuzzy Implication 7.4 Fuzzy systems
VIII	**Fuzzy Rule based Systems** 8.1 Linguistic Hedges 8.2 Aggregation of Fuzzy Rules
IX	**Artificial Neurons, Neural Networks and Architectures** 9.1 Neuron Abstraction 9.2 Neuron Signal Functions 9.3 Definition of Neural Networks 9.4 Architectures: Feedforward and Feedback 9.5 Salient properties and Application Domains

X	**Binary Threshold neurons** 10.1 Convex Sets 10.2 Hulls and Linear Separability 10.3 Space of Boolean Functions 10.4 Binary Neurons 10.5 Pattern Dicotomizers 10.6 TLN's 10.7 XOR problem
XI	**Perceptrons and LMS** 11.1 Learning and memory 11.2 Learning Algorithms 11.3 Error correction and gradient descent rules 11.4 The learning objectives for TLNs 11.5 Pattern space and weight space 11.6 Perceptron learning algorithm 11.7 Perceptron convergence algorithm 11.8 Perceptron learning and Non-separable sets 11.9 α-Least Mean Square Learning 11.10 Approximate Gradient Descent 11.11 Back Propagation Learning algorithm 11.12 Applications of Neural Networks
XII	**Perceptrons and LMS** 12.1 Learning and memory 12.2 Learning Algorithms 12.3 Error correction and gradient descent rules 12.4 The learning objectives for TLNs 12.5 Pattern space and weight space 12.6 Perceptron learning algorithm 12.7 Perceptron convergence algorithm 12.8 Perceptron learning and Non-separable sets 12.9 α-Least Mean Square Learning 12.10 Approximate Gradient Descent 12.11 Back Propagation Learning algorithm 12.12 Applications of Neural Networks

Learning Resources:

1. S. N. Sivanandam, S. N. Deepa, Principles of Soft Computing (With CD), ISBN:9788126527410, Wiley India
2. Timothy J Ross, Fuzzy Logic: With Engineering Applications, ISBN: 978-81-265- 3126- Wiley India, Third Edition
3. Kumar Satish, Neural Networks: A Classroom Approach, ISBN:9780070482920, 2008 reprint, 1/e TMH
4. David E. Goldberg, Genetic Algorithms in search, Optimization & Machine Learning, ISBN:81-7808-130-X, Pearson Education

Class: First Year M.Sc. Data Science- I, Semester-II

Course Code: CSD4207

Title of the Course/ Paper: Data Science Practical –III (Machine Learning using R)

Credits: 4

Learning Outcomes	Suggested Pedagogical Processes
Perform data preprocessing	Hands – on assignments
Implement various regression methods	
Understand classification algorithms	
Implement clustering k-Means	
Case studies	

Lab Course in Machine Learning Using R

Note: - Each Assignment will be based on Following Concept

Assignment No.	Topics Covered
1	Data Preprocessing – I
2	Data Preprocessing – II
3	Regression Analysis- Linear regression
4	Regression Analysis- Multiple regression
5	Regression Analysis- Logistic Regression
6	Classification Techniques- Decision tree,
7	Classification Techniques- SVM
8	Classification Techniques- Naïve Bayes, KNN
9	Clustering- K- Means clustering
10	Market Basket Analysis

Class: First Year M.Sc. Data Science- I, Semester-II

Course Code: CSD4208

Title of the Course/ Paper: Data Science Practical –IV (Python for Data Science)

Credits: 4

Learning Outcomes	Suggested Pedagogical Processes
To understand importance of Python for data science.	The Lecture method with use of ICT, practical demonstration.
Develop a skill to implement Python Programming for data science.	Brainstorming and solving practical assignments with small application development.
Hands-on Python experience for professional advancement	Developing a Mini project
To use and implement standard programming constructs like data structure, numerical computing, data manipulation and visualization	Practical demonstration and program assignment method

Unit No.	Title of Unit and Contents
I	**Fundamentals of python programming** 1.1 Different IDEs 1.2 Variable of different types 1.3 Control flow, regular expressions, modules, functions and packages 1.4 Data structures 1.5 Working with files and directories
II	**Numerical Computing in Python – NumPy** 2.1 NumPy Arrays – indexing, slicing, reshaping etc 2.2 Exploring Universal Functions – ufuncs 2.3 Aggressions 2.4 Computation on Arrays – broadcasting, comparisons, sorting, Fancy indexing etc 2.5 Structured Arrays
III	**Data Manipulation with Pandas** 3.1 Introducing Pandas Objects – series, data frames, index, 3.2 Processing CSV, JSON, XLS data 3.3 Operations on Pandas Objects – indexing and selection, universal functions, missing data, hierarchical indexing 3.4 Combining Dataset – concat and append, merge and join 3.5 Aggregation and grouping 3.6 Pivot table 3.7 Vectorized string operations 3.8 Working with time series 3.9 High performance Pandas – eval, query

IV	**Visualization in Python** 4.1 Introduction to Data Visualization – Matplotlib 4.2 Basic Visualization Tools – area, histogram, bar chart 4.3 Specialized Visualization Tools – pie chart, Box plot, scatter plot, Bobble plot 4.4 Advanced Visualization Tools – Waffle charts, Word cloud, Seaborn 4.5 Creating Maps and Visualizing Geospatial Data

Learning Resources:

1. Mark Lutz’s, Learning Python, O’Really
2. Mark Lutz’s, Programming Python, O’Really
3. Jake VanderPlas, Python Data Science Handbook, O’ Reilly
4. https://docs.python.org/3/tutorial/
5. https://wiki.python.org
6. https://www.numpy.org/devdocs/user/quickstart.html
7. https://www.learnpython.org/en/Pandas_Basics