Fergusson College (Autonomous) Pune

Learning Outcomes-Based Curriculum

For

M. Sc. I - 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 organizations make sense of their data. As data is collected and analyzed 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 enhances 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.

Programme Outcomes (POs) of M.Sc. Data Science	
PO1	**Disciplinary Knowledge:** Demonstrate comprehensive knowledge of the discipline that form a part of an postgraduate programme. Execute strong theoretical and practical understanding generated from the specific programme in the area of work.
PO2	**Critical Thinking and Problem solving:** Exhibit the skill of critical thinking and understand scientific texts and place scientific statements and themes in contexts and also evaluate them in terms of generic conventions. Identify the problem by observing the situation closely, take actions and apply lateral thinking and analytical skills to design the solutions.
PO3	**Social competence:** Exhibit thoughts and ideas effectively in writing and orally; communicate with others using appropriate media, build effective interactive and presenting skills to meet global competencies. Elicit views of others, present complex information in a clear and concise and help reach conclusion in group settings.
PO4	**Research-related skills and Scientific temper:** Infer scientific literature, build sense of enquiry and able to formulate, test, analyse, interpret and establish hypothesis and research questions; and to identify and consult relevant sources to find answers. Plan and write a research paper/project while emphasizing on academics and research ethics, scientific conduct and creating awareness about intellectual property rights and issues of plagiarism.
PO5	**Trans-disciplinary knowledge**: Create new conceptual, theoretical and methodological understanding that integrates and transcends beyond discipline-specific approaches to address a common problem.
PO6	**Personal and professional competence**: Perform independently and also collaboratively as a part of team to meet defined objectives and carry out work across interdisciplinary fields. Execute interpersonal relationships, self-motivation and adaptability skills and commit to professional ethics.
PO7	**Effective Citizenship and Ethics**: Demonstrate empathetic social concern and equity centred national development, and ability to act with an informed awareness of moral and ethical issues and commit to professional ethics and responsibility.
PO8	**Environment and Sustainability**: Understand the impact of the scientific solutions in societal and environmental contexts and demonstrate the knowledge of and need for sustainable development.
PO9	**Self-directed and Life-long learning:** Acquire the ability to engage in independent and life-long learning in the broadest context of socio-technological changes.

Program Specific Outcomes (PSOs) for M.Sc. Data Science Programme	
PSO1	**Academic competence**: (i) Understand fundamental concepts in statistics, mathematics and computer Science. (ii) Demonstrate an understanding of various analysis tools and software used in data science
PSO2	**Personal and Professional Competence**: (i) Apply laboratory-oriented problem solving and be capable in data visualization and interpretation. (ii) Solve case studies by applying various technologies, comparing results and analysing inferences. (iii) Develop problem solving approach and present output with effective presentation and communication skills
PSO3	**Research Competence:** (i) Design and develop tools and algorithms. (ii) contribute in existing open sources platforms (iii) Construct use case based models for various domains for greater perspective
PSO4	**Entrepreneurial and Social competence:** (i) Cater to/ provide solutions particular domain specific problems by having in depth domain knowledge (ii) Exposure to emerging trends and technologies to prepare students for industry (iii) Develop skills required for social interaction.

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	Emerging Trends in Data Science	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 Stat 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

F.Y. M.Sc. Semester I		
Title of the Course and Course Code	**Probability And Statistics (CSD4101)**	**Number of Credits : 04**
Course Outcome (COs) **On completion of the course, the students will be able to:**		
CO1	Describe basic features of the data.	
CO2	Summarize the sample using different quantitative measures.	
CO3	Apply and compare various counting techniques to analyse a particular problem.	
CO4	Identify different forms of probability distribution for discrete and continuous data.	
CO5	Evaluate and compute the chance of an event.	
CO6	Build predictive models for the sample data.	

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 $P(A \cap B) = P(A).\ P(B|A)$ 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

Title of the Course and Course Code	Applied Linear Algebra (CSD4102)	Number of Credits : 04
Course Outcome (COs) **On completion of the course, the students will be able to:**		
CO1	Describe the concepts of vectors and linear transformations.	
CO2	Explain linearly independent and dependent vectors.	
CO3	Use different concepts of inner products and associated norms.	
CO4	Explain and analyze basic algorithms for massive data problems	
CO5	Determine eigenvalues and eigenvectors of a given matrix and apply the concept for various methods of matrix factorization.	
CO6	Perform different matrix operations and integrate them to solve complex data science problems.	

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. Eigen values, Eigen vectors , orthogonalization
III	**Singular Value Decomposition:** 3.1 L1 norm, L2 norm, regularization of norm, covariance matrix 3.2 Preliminaries, Singular Vectors, Singular Value Decomposition 3.3 Best Rank k Approximations, Left Singular Vectors, Power Method for Singular Decomposition, Singular Vectors and Eigen Vectors 3.4 Applications of Singular Value Decomposition to Centring Data 3.5 Principal Component Analysis, Ranking Documents and Web Pages, Discrete Optimization Problem.
IV	**Algorithms for Massive Data Problems** 4.1 Introduction to algorithms, Characteristics of ideal efficient algorithms, complexity of algorithms 4.2 Big O notation, Scalable algorithms 4.3 Introduction to algorithms for massive data problems (Streaming, Sketching and Sampling), frequency movements of data stream: Number of distinct elements in data stream, Converting the intuition into an algorithm via hashing, two universal hash functions, Analysis of distinct element counting algorithm 4.4 Frequent elements including the majority and frequent algorithm, the second moment, matrix algorithms using sampling 4.5 Sketch of a large matrix, Sketches of documents

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

Title of the Course and Course Code	Data Structures (CSD4103)	Number of Credits : 04
Course Outcome (COs) **On completion of the course, the students will be able to:**		
CO1	Describe the basics of python programming.	
CO2	Explain programming constructs and apply them to build and package python modules for reusability.	
CO3	Use various data structures to gain suitable knowledge about their implementation.	
CO4	Compare various file handling techniques and database interactions.	
CO5	Evaluate patterns , compile expressions and write scripts to extract data.	
CO6	Write an application to solve real life problems by applying Object-Oriented principles.	
Unit No.	**Title of Unit and Contents**	
I	**Introduction To Python** 1.1 Introduction 1.2 Various IDEs	
II	**Data Types** 2.1 Numeric data types: int, float, complex 2.2 String, list and list slicing 2.3 Tuples	
III	**Control Flow, Functions, Modules And Packages** 3.1 Control Flow Conditional blocks using if, else and elif Simple for and while loops in python For loop using ranges, string, list and dictionaries Loop manipulation using pass, continue, break and else 3.2 Functions Arguments, Lambda Expressions, Function Annotations 3.3 Modules Organizing python projects into modules Importing own module as well as external modules	
	3.4 Packages 3.5 Programming using functions, modules and external packages	

IV	**Data Structures** 4.1 Lists as Stacks, Queues, Comprehensions 4.2 Tuples and sequences 4.3 Sets 4.4 Dictionaries
V	**Python File Operation** 5.1 Reading config files in python 5.2 Writing log files in python 5.3 Understanding read functions, read(), readline() and readlines() 5.4 Understanding write functions, write() and writelines() 5.5 Manipulating file pointer using seek 5.6 Programming using file operations
VI	**Object Oriented Programming** 6.1 Concept of class, object and instances 6.2 Constructor, class attributes and destructors, Inheritance, overlapping and overloading operators, 6.3 Adding and retrieving dynamic attributes of classes 6.4 Programming using Oops support
VII	**Regular Expression** 7.1 Powerful pattern matching and searching 7.2 Real time parsing of networking or system data using regex 7.3 Password, email, url validation using regular expression 7.4 Pattern finding programs using regular expression
VIII	**Database Interaction SQL** 8.1 Database connection using python 8.2 Creating and searching tables 8.3 Reading and storing config information on database 8.4 Programming using database connections
IX	**WEB Scraping** 9.1 Basics of scraping, 9.2 Scrape HTML Content From a static / dynamic pages 9.3 Parsing HTML code using packages like request and Beautiful soup

Learning Resources:

1. Learning Python, O'Reilly publication
2. Programming Python, O'Reilly publication
3. https://docs.python.org/3/tutorial/
4. https://realpython.com/beautiful-soup-web-scraper-python

Title of the Course and Course Code	Database Management System (CSD4104)	Number of Credits : 04
Course Outcome (COs) **On completion of the course, the students will be able to:**		
CO1	Describe different concepts of database management systems.	
CO2	Discuss structure of relational databases and apply relational operations on it.	
CO3	Apply the basic and advanced concepts of SQL language to solve the queries in the databases.	
CO4	Analyse database requirements and determine the entities involved in the system and their relationship.	
CO5	Compare traditional relational databases and NoSQL stores and explain types of NoSQL databases.	
CO6	Write the queries to implement different functionalities of SQL language.	

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** 1.1 Overview of NoSQL 1.2 Comparison of relational databases to new NoSQL stores 1.3 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.

Title of the Course and Course Code	**Data Science Practical - I (R Programming) (CSD4105)**	**Number of Credits : 04**
Course Outcome (COs) **On completion of the course, the students will be able to:**		
CO1	Describe concepts of Data Science and its specialised branches. State the use of the R and R-Studio's interactive environment	
CO2	Illustrate fundamentals of R language.	
CO3	Apply the data manipulation and transformation techniques to prepare data for further processing.	
CO4	Analyze the nature of data with help of statistical methods, different tools and visualization techniques.	
CO5	Evaluate various techniques and communicate observations.	
CO6	Write R scripts to solve complex business problems from different domains.	

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

Title of the Course and Course Code	**Data Science Practical - II (Data Structures and RDBMS) (CSD4106)**	**Number of Credits : 04**

Course Outcome (COs) On completion of the course, the students will be able to:	
CO1	Identify the concepts of Data structures and RDBMS to design solutions for different types of problems.
CO2	Explain the use of data structures and stored functions, views.
CO3	Apply different concepts of data structures and write programs.
CO4	Analyse/explain database application scenarios in the form of E-R and transform the ER-model to relational tables.
CO5	Test and validate the outputs of Data structures programs and SQL queries.
CO6	Write SQL queries to implement DDL, DML commands on relational databases to create and manipulate the table data.

Lab Course in Data Structures and RDBMS

Note:- Each Assignment will be based on following concepts

Assignment No.	Topics Covered
1	Strings and Lists
2	OOPS and Packages
3	Stacks, Queues, Tuples, Sets, Dictionaries
4	File Handling
5	Web Scraping / Regular Expression
6	Working with Database
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

F.Y. M.Sc. Semester II

Title of the Course and Course Code	Statistical Inference (CSD4201)	Number of Credits : 04
Course Outcome (COs) **On completion of the course, the students will be able to:**		
CO1	Identify sampling methods from the pattern of the observed data.	
CO2	Predict the future behaviour of the time series data.	
CO3	Predict different models of forecasting of time series data.	
CO4	Analyze sample data and identify the parameters and their probability distributions.	
CO5	Validate the hypothesis to ensure that the entire research process remains scientific and reliable.	
CO6	Hypothesize and test an assumption regarding population parameters using sample data.	
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

Title of the Course and Course Code	Mathematical Foundation(CSD4202)	Number of Credits : 04
Course Outcome (COs) **On completion of the course, the students will be able to:**		
CO1	Describe the basics of mathematical foundations to deal with high-dimensional data.	
CO2	Distinguish stochastic process for continuous or discrete time and state space.	
CO3	Implement and apply machine learning algorithms.	
CO4	Integrate the knowledge of SVD to find the best k-dimensional subspace.	
CO5	Evaluate different ways to implement Markov Chains models in various applications.	
CO6	Construct better and efficient Machine Learning algorithms based on mathematical knowledge.	

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	**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, 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.
III	**Graph Theory ,Random Graphs , Random Walks and Markov Chains:** Simple graphs, directed graphs, Undirected graphs , Digraph, Bipartite

	Graphs, Complete graphs, wheel graphs , regular graph ,edge removal , vertex removal , union of graphs , intersection of graphs ,Adjacency matrix , Incidence matrix , Hand Shaking Lemma.Edge contraction, Degree sequence, connected graphs, weighted Graphs.Generation of random graphs and applications of random graphs, Giant components of random graphs.Stationary Distribution, constructing simple random walks, Random walks on undirected graphs with unit edge weights, Random Walks in Euclidean Space, The Web as a Markov Chain.
IV	**Building Blocks of Machine Learning:** Learning problems , Target functions , Learned functions.The Perceptron Algorithm, Kernel Functions, Generalising to new data, Overfitting, Illustrative Examples and applications, 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.

Learning Resources:

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

Title of the Course and Course Code	**Machine Learning (CSD4203)**	**Number of Credits : 04**
Course Outcome (COs) **On completion of the course, the students will be able to:**		
CO1	Define a problem to find appropriate solutions in the field of data science and other interdisciplinary areas.	
CO2	Classify and apply machine learning techniques to solve real world problems.	
CO3	Apply various classification algorithms and evaluate their performance.	
CO4	Analyze various techniques of machine learning.	
CO5	Evaluate performance of machine learning models by using various performance evaluation parameters.	
CO6	Construct use case based models by analyzing datasets from various domains.	

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)

Title of the Course and Course Code	**Design and Analysis of Algorithms (CSD4204)**	**Number of Credits : 04**
Course Outcome (COs) **On completion of the course, the students will be able to:**		
CO1	Define algorithms and its properties.	
CO2	Differentiate between types of algorithms based on problem solving approach.	
CO3	Demonstrate major algorithms and data structures.	
CO4	Analyze the asymptotic performance of algorithms.	
CO5	Evaluate and select algorithmic design paradigms and methods of analysis.	
CO6	Develop 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

Title of the Course and Course Code	**Soft Computing (CSD4205)**	**Number of Credits : 04**

Course Outcome (COs) **On completion of the course, the students will be able to:**	
CO1	Outline different basics and techniques of soft computing and its importance.
CO2	Interpret the concept of fuzzy logic and its importance.
CO3	Apply ANN or GA techniques in various scenarios to solve different kinds of problems and the fuzzification process to handle the veuness in real world data.
CO4	Discriminate the soft computing techniques on the basis of applications.
CO5	Evaluate the goodness measure of the soft computing techniques by comparing it with other techniques.
CO6	Formulate the combination of one or more soft computing techniques to generate a more optimized solution.

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

Title of the Course and Course Code	**Data Science Practical - III (Machine Learning using R) (CSD4207)**	**Number of Credits : 04**
Course Outcome (CO) **On completion of the course, the students will be able to:**		
CO1	Define real world problem statements by performing data interpretation.	
CO2	Represent large scale data using data visualization techniques in R.	
CO3	Interpret the data using data pre-processing techniques in machine learning	
CO4	Analyze different machine learning models to get better accuracy and results.	
CO5	Evaluate performance of machine learning algorithms using performance metrics.	
CO6	Construct models using machine learning algorithms and compare the results.	

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

Title of the Course and Course Code	**Data Science Practical - IV (Python for Data Science) (CSD4208)**	**Number of Credits : 04**
Course Outcome (COs) **On completion of the course, the students will be able to:**		
CO1	Define various validation curve techniques.	
CO2	Exemplify the numerical computation with “Numpy” library.	
CO3	Apply the data transformation and data manipulation operations using “pandas”.	
CO4	Analyze nature of data with help of different tools and visualization techniques.	
CO5	Assess text data processing techniques.	
CO6	Write scripts, follow data pipelining, build models and measure accuracy to communicate the observations.	

Unit No.	Title of Unit and Contents
I	**Data processing with NumPy** 1.1 NumPy Arrays – indexing, slicing, reshaping etc 1.2 Exploring Universal Functions – ufuncs 1.3 Aggressions 1.4 Computation on Arrays – broadcasting, comparisons, sorting, Fancy indexing etc 1.5 Structured Arrays

II	**Data Manipulation and Pre-processing with Pandas** 2.1 Introducing Pandas Objects – series, data frames, index, 2.2 Processing CSV, JSON, XLS data 2.3 Operations on Pandas Objects – indexing and selection, universal functions, missing data, hierarchical indexing 2.4 Combining Dataset – concat and append, merge and join 2.5 Aggregation and grouping 2.6 Pivot table 2.7 Vectorized string operations 2.8 Working with time series 2.9 High performance Pandas – eval, query
III	**Visualization in Python** 3.1 Introduction to Data Visualization – Matplotlib 3.2 Basic Visualization Tools – area, histogram, bar chart 3.3 Specialized Visualization Tools – pie chart, Box plot, scatter plot, Bobble plot 3.4 Advanced Visualization Tools – Waffle charts, Word cloud, Seaborn 3.5 Creating Maps and Visualizing Geospatial Data

IV	**Working with Text data** 4.1 Splitting and Replacing a Data 4.2 Concatenation of Data 4.3 Removing Whitespaces of Data 4.4 Extracting data (using Regex) 4.5 Bags of words 4.6 Tokenizing 4.7 Counting Frequencies
V	**Data Pipelining** 5.1 Building new feature 5.2 Dimensionality reduction 5.3 Feature selection 5.4 Detection and treatment of outlier
VI	**Learning and validation curves** 6.1 Train Learning Curve 6.2 Validation Learning Curve 6.3 Accuracy vs loss 6.4 AOC and ROC

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