Deccan Education Society's

Fergusson College (Autonomous)

Pune

NEP 2020-Based Curriculum For

M. Sc. I - Data Science

With effect from Academic Year

2023-2024

Program Outcomes (POs) of M.Sc. Data Science	
PO1	**Disciplinary Knowledge:** Demonstrate comprehensive knowledge of the discipline that forms a part of a 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 a 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 centered 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 Program	
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 modCSD for various domains for greater perspective
PSO4	**Entrepreneurial and Social competence:** (i) Cater to/ provide solutions to 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

Semester	Paper Code	Paper Title	Type	Credits
I	CSD-501	Python Programming	Theory	3
	CSD-502	Probability and Statistics	Theory	3
	CSD-503	Mathematical Foundation - I	Theory	3
	CSD-504 OR	SQL for Data Science (Experiential Learning)	Theory	3
	CSD-505	Design and Analysis of Algorithms		
	CSD-510	Research Methodology	Theory	4
	CSD-520	Practical - I	Practical	2
	CSD-521	Practical - II	Practical	2
		Total Semester Credits		20
II	CSD-551	Machine Learning	Theory	3
	CSD-552	Statistical Inference	Theory	3
	CSD-553	Mathematical Foundation - II	Theory	3
	CSD-554 OR	Soft Computing	Theory	3
	CSD-555	Data Integration		
	CSD-560	On Job Training / Field Project	Project	4
	CSD-570	Practical - III	Practical	2
	CSD-571	Practical - IV	Practical	2
		Total Semester Credits		20
		Total PG-I Credits		40

Teaching and Evaluation (Only for FORMAL education courses)

Course Credits	No. of Hours per Semester Theory/Practical	No. of Hours per Week Theory/Practical	Maximum Marks	CE 40 %	ESE 60%
1	15 / 30	1 / 2	25	10	15
2	30 / 60	2 / 4	50	20	30
3	45 / 90	3 / 6	75	30	45
4	60 / 120	4 / 8	100	40	60

Eligibility: As per the rules and regulations of Savitribai Phule Pune University (SPPU)

F.Y. M.Sc. Semester I

CSD-501	Python Programming	**Credits:** 3 **Hours:** 45

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	Contents	No. of Hours
I	**Introduction To Python** ▪ Introduction ▪ Various IDEs	**02**
II	**Data Types** ▪ Numeric data types: int, float, complex ▪ String, list and list slicing ▪ Tuples	**06**
III	**Control Flow, Functions, Modules And Packages** ▪ Control Flow - Conditional blocks using if, if and elif, Simple for and while loops in python For loop using ranges, string, list and dictionaries Loop manipulation using pass, continue, break and exit ▪ Functions - Arguments, Lambda Expressions, Function Annotations ▪ Modules - Organizing python projects into modules Importing own module as well as external modules ▪ Packages	**10**

	▪ Programming using functions, modules and external packages	
IV	**Data Structures** ▪ Lists as Stacks, Queues, Comprehensions ▪ Tuples and sequences ▪ Sets ▪ Dictionaries	**06**
V	**Python File Operation** ▪ Reading different Types of files (config / log) in python ▪ Understanding Read / Write functions ▪ Manipulating file pointer using seek ▪ Programming using file operations	**06**
VI	**Object Oriented Programming** ▪ Concept of class, object, and instances ▪ Constructor, class attributes and destructors, Inheritance, overlapping and overloading operators, ▪ Adding and retrieving dynamic attributes of classes ▪ Programming using Oops support	**08**
VII	**Regular Expression** ▪ Powerful pattern matching and searching ▪ Password, email, url validation using regular expression ▪ Pattern finding programs using regular expression	**04**
VIII	**Database Interaction SQL** ▪ Database connection using python ▪ Creating and searching tables ▪ Reading and storing config information on database ▪ Programming using database connections	**05**

Learning Resources:

1. Learning Python, O’Reilly publication
2. Programming Python, O’Reilly publication
3. https://docs.python.org/3/tutorial

F.Y. M.Sc. Semester I		
CSD-502	**Probability And Statistics**	**Credits: 3** **Hours:45**

Course Outcome (COs) **On completion of the course, the students will be able to:**	
CO1	Describe basic features of the data.
CO2	Summarise 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.

Unit	Contents	No. of Hours
I	**Descriptive Statistics:** ▪ Measures of Central Tendency: Mean, Median, Mode ▪ Partition Values: Quartiles, Percentiles, Box Plot ▪ Measures of Dispersion: Variance, Standard Deviation, Coefficient of variation ▪ Skewness: Concept of skewness, measures of skewness ▪ Kurtosis: Concept of Kurtosis, Measures of Kurtosis (All topics to be covered for raw data using R software also.)	**07**
II	**Introduction to Probability**: ▪ Probability - classical definition, probability models, axioms of probability, probability of an event. ▪ Concepts and definitions of conditional probability, multiplication theorem P(A∩B) =P(A). P(B|A) ▪ Bayes’ theorem (without proof) ▪ Concept of Posterior probability, problems on posterior probability. ▪ Definition of sensitivity of a procedure, specificity of a procedure. Application of Bayes’ theorem to design a procedure for false positive and false negative. ▪ Concept and definition of independence of two events.	**09**

	▪ Numerical problems related to real life situations.	
III	**Introduction to Random Variables** ▪ Definition of discrete random and continuous random variable. ▪ Concept of Discrete and Continuous probability distributions. (p.m.f. and p.d.f.) ▪ Distribution function ▪ Expectation and variance ▪ Numerical problems related to real life situations	**03**
IV	**Special Distributions** ▪ Binomial Distribution ▪ Uniform Distribution ▪ Poisson Distribution ▪ Negative Binomial Distribution ▪ Geometric Distribution ▪ Continuous Uniform Distribution ▪ Exponential Distribution ▪ Normal Distribution ▪ Log Normal Distribution ▪ Gamma Distribution ▪ Weibull Distribution ▪ 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. **)**	**12**
V	**Correlation and Linear Regression** ▪ Bivariate data, Scatter diagram. ▪ Correlation, Positive Correlation, Negative correlation, Zero Correlation ▪ Karl Pearson's coefficient of correlation (r), limits of r ($-1 \leq r \leq 1$), interpretation of r, Coefficient of determination (r2) ▪ Meaning of regression, difference between correlation and regression. ▪ Fitting of line Y = a+bX ▪ Concept of residual plot and mean residual sum of squares. ▪ Multiple correlation coefficient, concept, definition, computation, and interpretation. ▪ Partial correlation coefficient, concept, definition, computation, and interpretation. ▪ Multiple regression plane. ▪ Identification and solution to Multicollinearity ▪ Evaluation of the Model using R square and Adjusted R square ▪ All topics to be covered for raw data using R software also.	**11**

VI	**Logistic Regression** ▪ Introduction to logistic regression ▪ Difference between linear and logistic regression ▪ Logistic equation ▪ How to build logistic regression model in R ▪ Odds ratio in logistic regression.	**03**

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

F.Y. M.Sc. Semester I

CSD-503	**Mathematical Foundation- I**	**Credits: 03** **Hours:45**

Course Outcome (COs) **On completion of the course, the students will be able to:**	
CO1	Describe the basics of sets and vector algebra.
CO2	Explain the fundamentals of Logic.
CO3	Use various properties of vectors and matrices to store and prepare data for further modeling.
CO4	Analyze different types of relations and functions on sets.
CO5	Evaluate Eigenvalues and vectors of a matrix.
CO6	perform different matrix decompositions and integrate the ideas of linear algebra for dimension reduction of data.

Unit	**Contents**	**No. of Hours**
I	**Sets, Relations & Functions** ▪ Introduction to Sets (Elements, Subsets, Venn diagram to visualize set relationships) ▪ Operations on Sets (Union, Intersection, Complement of Set), Associative, distributive and Identity Laws, Cardinality and Power Sets ▪ Cartesian Product, Types of relations ▪ Equivalence relations and partitions ▪ Introduction to functions and function properties	**08**
II	**Propositional Calculus** ▪ Propositions and logical connectives (AND, OR, NOT) ▪ Truth tables and Logical equivalences ▪ Implication and logical deduction	**06**
III	**Vectors & Matrices** ▪ Introduction to vectors, Vector Operations (addition, Scalar Vector multiplication, Inner Product) ▪ Linear Functions, affine functions ▪ Norm, distance, Standard deviation, Angle between vectors ▪ Clustering K means algorithm), examples and applications ▪ Linear independence, Linear Dependence, Basis, Orthonormal Vectors, Orthogonalisation, Gram Schmith algorithm ▪ Types and operations on matrices, System of Linear Equations, Inverse of matrix (left and right inverse) ▪ Eigenvalues and Eigenvectors, Diagonalisation	**18**

IV	**Matrix Decompositions** ▪ QR factorisation ▪ Singular values, left and right singular vectors, Singular Value decomposition, Best Rank k Approximation ▪ Applications of Singular Value Decomposition to Centring Data ▪ Principal Component Analysis, Ranking Documents and Web Pages	**13**

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.
2. ‘Linear Algebra and its Applications’, David C. Lay; Pearson Education, Third Edition.
3. Introduction to “Elements of Set Theory” by Hebert B. Enderton

F.Y. M.Sc. Semester I		
CSD-504	**SQL for Data Science** **(Experiential Learning)**	**Credits: 3** **Hours:45**

Course Outcome (COs) **On completion of the course, the students will be able to:**	
CO1	Describe basic concepts of SQL
CO2	Distinguish between use of SQL for data science and SQL as a common database management system
CO3	Apply the basic and advanced concepts of SQL language to solve the queries in the databases.
CO4	Analyze the data using SQL statements
CO5	Determine the data elements, their relationships, and inter-dependencies to address a business question by using ER-Diagrams.
CO6	Write the queries to use SQL for data analysis.

Unit	Contents	No. of Hours
I	**Introduction to Relational Database Management System** ▪ Database-system purpose and its applications ▪ View of Data-Data Abstraction, Instance and Schemas ▪ Relational Databases: Tables, DML, DDL ▪ Structure of Relational Databases ▪ Database Schema ▪ Keys ▪ Relational Operations	**03**
II	**Database Design and E-R model** ▪ Overview of the Design process and Entity Relationship Model ▪ Constraints and Removing Redundant Attributes in Entity Sets ▪ Entity Relationship Diagrams ▪ Introduction to UML Relational database model: Logical view of data, keys, integrity rules ▪ Anomalies in a Databases	**10**
III	**Basic SQL** ▪ Overview of SQL query language ▪ SQL data definition- Basic Types, Basic schema definition, Date and Time in SQL, Default values, Index creation, Large Object types, user- defined types ▪ Integrity constraint- Constraints on a single relation, Not Null constraint, Unique constraint, The Check clause, referential integrity ▪ Basic structure of SQL queries ▪ Additional basic operations ▪ Set operations ▪ Null Values ▪ Aggregate Functions-Basic aggregation, Aggregation and grouping, The Having clause, CASE ▪ Modifying and analyzing the data with SQL: working with text strings, date and time ▪ 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 ▪ Modification of the Database- Deletion, Insertion, Updates	**15**
IV	**Intermediate and advanced SQL** ▪ Join Expressions- Join conditions, Outer joins, Join types and conditions ▪ Views- View definition, using views in SQL queries, Materialized views, update a view ▪ Create table extensions Schemas, Catalogs and Environments ▪ Stored procedures ▪ Stored Functions	**15**

	▪ Cursors	
V	**Using SQL for Data Science** ▪ Introduction to Data Cleaning ▪ Use of String functions for data cleaning: Left, Right, Length, Concat, Position, substr, strpos, coalesce ▪ Introduction to window functions: aggregate functions, partition by, over, order by, Row_number, Rank ▪ Comparing row to previous row, percentiles	**02**

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

Web References:

3. www.towardsdatascience.com
4. **https://www.mysql.com/**

F.Y. M.Sc. Semester I		
CSD-505	**Design and Analysis of Algorithms**	**Credits: 03** **Hours: 45**

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	Contents	No. of Hours
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)	**02**
II	**Divide and Conquer** Concept of divide and Conquer, Binary Search (recursive), Quick Sort, Merge sort	**05**
III	**Greedy Method** Fractional Knapsack problem, Optimal Storage on Tapes, Huffman codes, Concept of Minimum Cost Spanning Tree, Prim’s and Kruskal's Algorithm	**08**
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, Traveling Salesperson Problem	**08**

V	**Branch & Bound** Introduction, Definitions of LCBB Search, Bounding Function, Ranking Function, FIFO BB Search, Traveling Salesman problem Using Variable tuple.	**08**
IV	**Decrease and conquer** Definition of Graph Representation, BFS, DFS, Topological Sort/Order, Strongly Connected Components, Biconnected Component, Articulation Point and Bridge edge	**08**
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)	**08**

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

F.Y. M.Sc. Semester I		
CSD-510	**Research Methodology**	**Credits : 04** **Hours : 60**
On completion of the course, the students will be able to:		
CO1	Learn the various aspects of the research process, framing useful research questions, research design, data collection, analysis, writing and presentation	
CO2	Understand the research problem, methods/techniques to be adopted	
CO3	Apply statistical tools for analysing the data while performing their research	
CO4	Develop skills in qualitative and quantitative data analysis and presentation	
CO5	Analyse for fitting, errors in the measurements and able to withdraw conclusions from the analysed data	
CO6	Execute a quality research paper and patents in science and technology	

Unit	Contents	No. of Hours
I	History of research. Indian, Egyptian, Greek ideas methodologies and research in agriculture, chemistry, metallurgy, medical. Ancient Indian research methodology applications.	08
II	Statistical analyses and its significance, Exploratory and confirmatory research, Planned and ad-hoc methods of data collection, Non-response and methods of recovering the missing response, Various software for statistical analysis, case studies of the research performed in various subjects using statistical methods, Error and noise analysis, curve fitting.	10
III	Literature search and Review: selection of research topic (case study based), maintaining laboratory records (case study based). Safety in Laboratories, Ethical considerations, effective verbal and non-verbal communication, field data collection, safety in field, Qualitative and Quantitative Research:	12
IV	Developing Research Plan: Writing research paper and/or thesis, making a presentation, writing a research proposal, and patents in Science, technology, Data Interpretation, paper publications and its ethics	10
V	**Research Tools, Databases and Research Metrics** Experimental measurements, numerical modeling, theoretical derivations & Calculations, Different Databases and metrics	10
VI	**Research Methods in Data Science** Business Requirement Analysis, Data Collection using different methods, Modeling, Methods for evaluating and monitoring performance of models, Handling domain specific case studies in healthcare, finance, social media,	10

	Human Behaviour Analysis etc.	

Learning Resources:

1. ‘History of the Scientific Methods’ by Martin Shuttleworth, https://explorable.com/history-of-the- scientific-method.
2. The Statistical Analysis of Experimental Data’ by, John Mandel, ISBN: 0486646661, ISBN13: 9780486646664

F.Y. M.Sc. Semester I		
CSD-520	**Practical - I** **(Python Programming)**	**Credits: 2** **Hours:60**

Course Outcome (COs) **On completion of the course, the students will be able to:**	
CO1	Identify the concepts of Data structures and design solutions for different types of problems.
CO2	Explain the use of data structures.
CO3	Apply different concepts of data structures and write programs.
CO4	Test and validate the outputs of Data structures programs.
CO5	Problem solving using OOPS concepts.
CO6	Building application using Python programming

Lab Course in Python Note: - Each Assignment will be based on following concepts	
Sr. No.	**Topics Covered**
1	Strings and Lists
2	Functions and Packages
3	Stacks, Queues, Tuples, Sets, Dictionaries
4	File Handling
5	Regular Expression
6	OOPS

7	Working with Databases

F.Y. M.Sc. Semester I		
CSD-521	**Practical - II** **Emerging Tools and Techniques in Data Science (Experiential Learning)**	**Credits: 2** **Hours:60**
Course Outcome (COs) **On completion of the course, the students will be able to:**		
CO1	Outline different terms and concepts in Data science.	
CO2	Explain the importance of different steps in data processing to get the desired result.	
CO3	Implement different models and concepts in Data Science.	
CO4	Analyze different visualizations to display the result.	
CO5	Measure and test the performance of different models in different domains.	
CO6	Develop and deploy Dashboard using different tools.	

Unit.	**Contents**	**No. of Hours**
I	**Introduction** ▪ Data science Concept ▪ Different Methodology ▪ Cloud and Data Science ▪ Data Preparation, Data Transformation ▪ Data visual representation ▪ Machine learning Concept and algorithm	**20**
II	**Technology Used in Data Science** ▪ Technology Implementation method with data ▪ Exploring and Preparing auto data ▪ Validating automotive data ▪ Visualize preliminary data wrangling results ▪ Run summary statistics on the results ▪ Exploring visualization tool for data ▪ Implementation of ML concept	**20**

III	**Various Domain based Case study implementation with Technology.**	**20**

F.Y. M.Sc. Semester II	
Machine Learning (CSD551) **Credits: 3**	
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 by using various performance evaluation parameters.
CO6	Construct a use case based on analyzing datasets from various domains.

Unit	Contents	No. of Hours
I	**Machine Learning Basics** ▪ Frameworks of Data Analysis ▪ Basics of Machine Learning ▪ Types of Machine Learning ▪ Models of Machine Learning	**04**
I I	**Machine Learning Theory** ▪ Features: Feature Construction, Transformation, Feature Selection ▪ Dimensionality Reduction: Subset selection, the Curse of Dimensionality ▪ Principle Components analysis ▪ Linear and Quadratic Discriminant Analysis ▪ Bias Variance Tradeoff ▪ Model Evaluation Metrocs	**06**

III	**Regression Analysis** ▪ Linear Regression ▪ Multiple Regression ▪ Lasso & Ridge Regression	**06**
IV	**Classification Techniques** ▪ Logistic Regression ▪ Decision Tree Algorithms ▪ Support Vectors: Maximal margin classifier, Support Vector Machines as a linear and non-linear classifier ▪ Naïve Bayes ▪ KNN	**10**
V	**Clustering Techniques** ▪ Types of clustering, methods for choosing number of clusters ▪ K means clustering ▪ Association Rule Mining ▪ Apriori Algorithm	**10**
VI	**Ensemble Learning Methods** ▪ Bagging, Boosting ▪ Random Forest Classifier Techniques ▪ Reinforcement Learning	**05**
VII	**Case Studies on Machine Learning Techniques** Diagnosis of human disease, Text Mining, Prediction, and forecasting	**04**

Learning Resources:

1. Trevor Hastle, Robert Tibshirani, Jerome Friedman , The Elements of Statistical Learning Data Mining, Inference and Prediction, second edition, Springer Series in Statistics
2. Hastie, Tibshirani, Friedman, Introduction to statistical machine learning with applications in R,
3. Tom Mitchell, Machine Learning, McGraw Hill, 1997, 0-07-042807-7
4. Jiawei Han, Micheline Kamber, Jian Pei, Data Mining: Concepts and Techniques, 3rd Edition
5. Margaret H. Dunham, S. Sridhar, Data Mining - Introductory and Advanced Topics, Pearson Education 5. Tom Mitchell, Machine Learning‖, McGraw-Hill, 1997
6. R.O. Duda, P.E. Hart, D.G. Stork., Pattern Classification, Second edition. John Wiley and Sons, 2000.
7. Christopher M. Bishop, Pattern Recognition and Machine Learning, Springer 2006 8. Ian H. Witten, Data Mining: Practical Machine Learning Tools and Techniques, Eibe Frank CSDevier / (Morgan Kauffman)

Courses: ● Introduction to Machine Learning, By prof. Balaraman Ravindran

F.Y. M.Sc. Semester II	
Statistical Inference (CSD552) **Credits: 3**	
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	**Contents**	**No. of Hours**
I	**Sampling** ▪ Introduction to Sampling ▪ Simple random Sampling ▪ Stratified Random Sampling ▪ Cluster Sampling ▪ Concept of Sampling Error	**04**
II	**Sampling Distributions** ▪ Introduction to Sampling distributions ▪ Student's t distribution ▪ Chi square distribution ▪ Snedecor's F distribution ▪ Interrelations among t, chi-square and F distributions	**08**

	▪ Central Limit Theorem (Various Versions) and its applications.	
III	**Testing of hypothesis** ▪ Definitions: population, statistic, parameter, standard error of estimator. ▪ 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. ▪ Large Sample Tests ▪ Tests based on t, Chi-square, and F-distribution **All tests to be taught using R software also.**	**15**
IV	**Analysis of Variance** ▪ One Way ANOVA ▪ Two Way ANOVA ▪ Application of ANNOVA to test the overall significance of Regression. **All topics to be covered using R software also.**	**06**
V	**Time Series** ▪ Meaning and Utility. ▪ Components of Time Series. ▪ Additive and Multiplicative models. ▪ Methods of estimating trend: moving average method, least squares method and exponential smoothing method. (single, double and triple) ▪ Elimination of trend using additive and multiplicative models. ▪ Simple time series models: AR (1), AR (2). Introduction to ARIMA Modelling.	**12**

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

F.Y. M.Sc. Semester II	
Mathematical Foundation- II (CSD553) **Credits: 3**	
Course Outcome (COs) **On completion of the course, the students will be able to:**	
CO1	Identifying role of calculus in machine learning algorithms
CO2	Distinguish between discrete time markov chain and continuous time markov chain.
CO3	Apply various tools to implement different optimization methods.
CO4	Analyze and appreciate a variety of performance measures for various optimization problems.
CO5	Validate mathematical minima/maxima problems into optimization framework.
CO6	Evaluate different ways to implement Markov Chains models in various applications.

Unit No.	Title of Unit and Contents	No. of Hours
I	**Graph Theory, Random Graphs, Random Walks and Markov Chains:** ▪ Simple graphs, directed graphs, undirected graphs, Digraphs, Bipartite graphs Adjacency matrix, Incidence matrix . Hand shaking Lemma. ▪ Edge contraction, Degree sequence, connected graphs, weighted graphs, strongly connected components. Edge contraction. ▪ Generation of random graphs and applications of random graphs. Giant components of random graphs. ▪ Introduction to Markov chain, Stationary distributions, constructing simple random walks. Random walks on undirected graphs with unit edge weights. o Random walks in Euclidean Space. ▪ The web as a Markov chain. Applications of Markov Chain.	18
II	**Multivariable calculus** ▪ Maxima, Minima, critical points of single variable functions and multi variable functions. Gradient, Jacobian, Hessian. ▪ Differentials and Chain rule in multiple dimensions. ▪ Single variable optimizations with and without constraints. ▪ Multi-variable optimization with and without constraints. ▪ method of Penalty methods, Lagrange multipliers, Kuhn-Tucker conditions.	15

III	**Hyperparameter optimization** ▪ Gradient Descent Method. ▪ Steepest descent method. ▪ Nelder Mead's Simplex search method, Newton's method.	**12**

Learning Resources:

1. Foundations of Data Science: Alvin Blum, John Hopcroft and Ravindran Kannan.
2. Multivariable Calculus 7th edition by James Stewart, published by Brooks/Cole, Cengage Learning.
3. Optimization Theory and Applications, 2nd Ed., 1984, S.S. Rao, Wiley Eastern Ltd.

Web links: Mathematical Foundation of Data Analysis. J. Philips

Download link: http://www.cs.utah.edu/~jeffp/M4D/M4D-v0.6.pdf

F.Y. M.Sc. Semester II		
CSD-554	**Soft Computing**	**Credits: 3** **Hours:45**

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 vagueness in real world data.
CO4	Discriminate the soft computing techniques based on 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**	**No. of Hours**

I	**Introduction to Soft Computing** ▪ What is soft computing ▪ Principle of soft computing (SC Paradigm) ▪ How is it different from hard computing? ▪ Constituents of SC (Fuzzy Neural, Machine Learning, Probabilistic reasoning)	**02**
II	**Fuzzy Logic** ▪ Difference between fuzzy and classical set (Operations and properties) ▪ Fuzzy Relations - Cardinality, Operations, Properties, Composition, ▪ Membership functions – features, Standard Forms and Boundaries, Fuzzification methods ▪ Fuzzy to crisp conversion - Fuzzy Tolerance and equivalence relations ▪ Lambda (alpha) cuts for fuzzy sets and relations, Defuzzification methods	**09**
III	**Fuzzy Arithmetic and and fuzzy systems** ▪ Fuzzy Arithmetic ▪ Fuzzy numbers ▪ Extension Principle ▪ Fuzzy Logic ▪ Approximate Reasoning ▪ Fuzzy Implication ▪ Fuzzy systems ▪ Linguistic Hedges ▪ Aggregation of Fuzzy Rules	**10**
IV	**Artificial Neurons, Neural Networks and Architectures** ▪ Neuron Abstraction ▪ Neuron Signal Functions ▪ Definition of Neural Networks ▪ Architectures: Feedforward and Feedback ▪ Salient properties and Application Domains	**03**
V	**Binary Threshold neurons** ▪ Convex Sets ▪ Hulls and Linear Separability ▪ Space of Boolean Functions ▪ Binary Neurons ▪ Pattern Dicotomizers ▪ TLN's XOR problem	**05**

VI	**Perceptron and LMS** ▪ Learning and memory ▪ Learning Algorithms ▪ Error correction and gradient descent rules ▪ The learning objectives for TLNs ▪ Pattern space and weight space ▪ Hebbian learning ▪ Perceptron learning algorithms ▪ Perceptron learning and non-separable sets ▪ α-Least Mean Square Learning ▪ Approximate Gradient Descent ▪ Applications of Neural Networks	**10**
VII	**Genetic algorithm** ▪ Biological background ▪ Search space ▪ Basic terminologies in GA ▪ A simple GA – General GA ▪ Operators in GA (Encoding, Selection, Crossover – mutation) ▪ Stopping conditions ▪ Constraints ▪ Problem solving ▪ The schema theorem ▪ Advantages ▪ Applications**.** ▪ Differences and similarities between GA and other traditional methods	**06**

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

F.Y. M.Sc. Semester II		
CSD-555	**Data Integration**	**Credits: 3**

Course Outcome (COs) **On completion of the course, the students will be able to:**	
CO1	Understand the problems related to querying heterogeneous and autonomous data sources
CO2	Explain the differences and similarities between the data integration / exchange, data warehouse, and Big Data analytics approaches
CO3	Apply ETL techniques to the data
CO4	Learn techniques for expressing schema mappings
CO5	Differentiate between data integration vs. data exchange
CO6	Create or build Data pipeline to populate Data Warehouse

Unit No.	**Title of Unit and Contents**	**No. of Hours**
I	**Introduction** ▪ Heterogeneity of data ▪ Uncertainty and incompleteness ▪ Autonomous and distributed data sources – Structured vs. unstructured data.	**04**
II	**Pre-processing and Cleaning** ▪ Entity resolution – Data fusion ▪ Cleaning	**04**
III	**Data Integration** ▪ Mediated schemata and query rewrite – Schema matching ▪ Schema mappings	**08**
IV	**Data Exchange** ▪ Data exchange transformations ▪ Universal solutions	**08**
V	**Data Warehousing** ▪ Extract-transform-load (ETL) ▪ Data cubes ▪ Star- and snowflake schemas ▪ Efficient analytics (OLAP) and relationship to transactional relational systems (OLTP)	**12**

VI	**Big Data Analytics** ▪ Big Data analytics platforms and programming models ▪ Differences between Big Data analytics and traditional warehousing approaches – Big Data integration	**05**
VII	**Data Provenance** ▪ Why- and Where-provenance ▪ Provenance polynomials ▪ Provenance in data integration – Provenance for missing answers	**04**

Learning Resources:

1. Doan, Halevy, and Ives, Principles of Data Integration, 1th Edition, Morgan Kaufmann, 2012
2. Elmasri and Navathe, Fundamentals of Database Systems, 6th Edition, Addison-Wesley, 2003
3. Silberschatz, Korth, and Sudarshan, Database System Concepts, 6th Edition, McGraw Hill, 2010

Web references:

1. CS520 - Index - CS520 - Data Integration, Warehousing, and Provenance - 2022 Spring (iit.edu)

F.Y. M.Sc. Semester II		
CSD-570	**Practical - III** **(Machine Learning)**	**Credits: 2** **Hours:60**

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 visualisation techniques.
CO3	Interpret the data using data pre-processing techniques in machine learning
CO4	Analyse different machine learning models to get better accuracy and results.
CO5	Evaluate performance of machine learning algorithms using performance metrics.
CO6	Construct a model using machine learning algorithms and compare the results.

Lab Course in Machine Learning

Note: - Each Assignment will be based on Following Concept

Assignment No.	Topics Covered
1	Regression Analysis- Linear regression
2	Regression Analysis- Multiple regression
3	Regression Analysis- Logistic Regression
4	Classification Techniques- Decision tree,
5	Classification Techniques- SVM
6	Classification Techniques- Naïve Bayes, KNN
7	Advanced Classification Algorithms
8	Clustering- K- Means clustering, Market Basket Analysis, Apriori
9	Case Study 1
10	Case Study 2

F.Y. M.Sc. Semester II

CSD-571	**Practical - IV** **Python for Data Science (Experiential Learning)**	**Credits: 2** **Hours:60**

Course Outcome (COs)

On completion of the course, the students will be able to:

CO1	Understand usage of functions and packages for respective problem-solving techniques.
CO2	Exemplify the numerical computation with “Numpy” library.
CO3	Apply the data transformation and data manipulation operations using “pandas.”
CO4	Analyse nature of data with help of different tools and visualisation 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	No. of Hours
I	**Data processing with NumPy** ▪ NumPy Arrays - indexing, slicing, reshaping etc ▪ Exploring Universal Functions – ufuncs ▪ Aggressions ▪ Computation on Arrays - broadcasting, comparisons, sorting, Fancy indexing etc. ▪ Structured Arrays	**12**
II	**Data Manipulation and Pre-processing with Pandas** ▪ Introducing Pandas Objects – series, data frames, index, ▪ Processing CSV, JSON, XLS data ▪ Operations on Pandas Objects – indexing and selection, universal functions, missing data, hierarchical indexing ▪ Combining Dataset - concat and append, merge and join ▪ Aggregation and grouping ▪ Pivot table ▪ Vectorized string operations ▪ Working with time series ▪ High performance Pandas - eval, query	**16**
III	**Visualization in Python** ▪ Introduction to Data Visualization – Matplotlib ▪ Basic Visualization Tools – area, histogram, bar chart ▪ Specialized Visualization Tools – pie chart, Box plot, scatter plot, Bobble plot ▪ Advanced Visualization Tools – Waffle charts, Word cloud, Seaborn	**16**
IV	**Web Scraping** ▪ Basics of scraping, ▪ Scrape HTML Content From a static / dynamic pages ▪ Parsing HTML code using packages like request and Beautiful soup	**08**
V	**Data Pipelining** ▪ Building new feature ▪ Dimensionality reduction ▪ Feature selection ▪ Detection and treatment of outlier	**08**