CHRIST
(DEEMED TO BE UNIVERSITY)
PUNE LAVASA CAMPUS
The Hub of Analytics

# SYLLABUS

# MSC

## DATA SCIENCE (2022-23)

**CHRIST (Deemed to be University)**
Pune Lavasa Campus - 'The Hub of Analytics'

**1800 123 2009**
lavasa.christuniversity.in

Syllabus for Master of Science (MSc Data Science) 2022-23 approved by the Board of Studies, Department of Computer science and Academic Council, CHRIST(Deemed to be University), Bangalore, India.

Published by the Centre for Publications, CHRIST (Deemed to be University), Hosur Road, Bangalore, 560 029, India. publications@christuniversity.in

2022

# Index

**Department Overview:**

Department of Computer Science of CHRIST (Deemed to be University) strives to shape outstanding computer professionals with ethical and human values to reshape nations destiny. The training imparted aims to prepare young minds for the challenging opportunities in the IT industry with a global awareness rooted in the Indian soil, nourished and supported by experts in the field.

**Vision and Mission:**

Vision The Department of Computer Science endeavours to imbibe the vision of the University "Excellence and Service". The department is committed to this philosophy which pervades every aspect and functioning of the department.

Mission: To develop IT professionals with ethical and human values. To accomplish our mission, the department encourages students to apply their acquired knowledge and skills towards professional achievements in their career. The department also moulds the students to be socially responsible and ethically sound.

**Introduction To The Program:**

Data Science is popular in all academia, business sectors, and research and development to make effective decision in day to day activities. MSc in Data Science is a two year programme with four semesters. This programme aims to provide opportunity to all candidates to master the skill sets specific to data science with research bent. The curriculum supports the students to obtain adequate knowledge in theory of data science with hands on experience in relevant domains and tools. Candidate gains exposure to research models and industry standard applications in data science through guest lectures, seminars, projects, internships, etc.

## Programme Objective

To acquire in-depth understanding of the theoretical concepts in statistics, data analysis, data mining, machine learning and other advanced data science techniques.

To gain practical experience in programming tools for data sciences, database systems, machine learning and big data tools.

To strengthen the analytical and problem solving skill through developing real time applications.

To empower students with tools and techniques for handling, managing, analyzing and interpreting data.

To imbibe quality research and develop solutions to the social issues.

**Programme Outcome**

PO1 Engage in continuous reflective learning in the context of technology and scientific advancement.
PO2 Identify the need and scope of the Interdisciplinary research.
PO3 Enhance research culture and uphold the scientific integrity and objectivity
PO4 Understand the professional, ethical and social responsibilities
PO5 Understand the importance and the judicious use of technology for the sustainability of the environment
PO6 Enhance disciplinary competency, employability and leadership skills

**Programme Specific Outcomes**

PSO1: Abstract thinking: Ability to understand the abstract concepts that lead to various data science theories in Mathematics, Statistics and Computer science.
PSO2: Problem Analysis and Design Ability to identify analyze and design solutions for data science problems using fundamental principles of mathematics, Statistics, computing sciences, and relevant domain disciplines.
PSO3: Modern software tool usage: Acquire the skills in handling data science programming tools towards problem solving and solution analysis for domain specific problems.
PSO4: Innovation And Entrepreneurship: Produce innovative IT solutions and services based on global needs and trends.
PSO5: Societal And Environmental Concern: Utilize the data science theories for societal and environmental concerns.
PSO6: Professional Ethics: Understand and commit to professional ethics and cyber regulations, responsibilities, and norms of professional computing practices.
PSO7: Conduct Investigations of complex computing problems: Use research-based knowledge and research methods including design of experiments, analysis and interpretation of data, and synthesis of the information to provide valid conclusions.
PSO8: Individual and Team work: Function effectively as an individual and as a member or leader in diverse teams and in multidisciplinary environments.
PSO9: Applications in Multi disciplinary domains: Understand the role of statistical approaches and apply the same to solve the real life problems in the fields of data science.
PSO10: Project Management: Apply the research-based knowledge to analyse and solve advanced problems in data science.

## 1 Semester

| Course Code | Course Title | Hours Per Week | Credits | Marks |
|---|---|---|---|---|
| MDS131 | MATHEMATICAL FOUNDATION FOR DATA SCIENCE - I | 4 | 4 | 100 |
| MDS132 | PROBABILITY AND DISTRIBUTION THEORY | 4 | 4 | 100 |
| MDS133 | PRINCIPLES OF DATA SCIENCE | 4 | 4 | 100 |
| MDS134 | RESEARCH METHODOLOGY | 2 | 2 | 50 |
| **Choose Any One (Foundational Elective)** | | | | |
| MDS161A | INTRODUCTION TO STATISTICS | 2 | 2 | 50 |
| MDS161B | INTRODUCTION TO COMPUTERS AND PROGRAMMING | | | |
| MDS161C | LINUX ADMINISTRATION | | | |
| MDS171 | DATA BASE TECHNOLOGIES | 6 | 5 | 150 |
| MDS172 | INFERENTIAL STATISTICS | 6 | 5 | 150 |
| MDS173 | PROGRAMMING FOR DATA SCIENCE IN PYTHON | 6 | 4 | 100 |
| HOLODD | HOLISTIC EDUCATION | - | - | - |
| Total | - | 34 | 30 | 800 |

## 2 Semester

| Course Code | Course Title | Hours Per Week | Credits | Marks |
|---|---|---|---|---|
| MDS231 | MATHEMATICAL FOUNDATION FOR DATA SCIENCE - II | 4 | 4 | 100 |
| MDS232 | REGRESSION ANALYSIS | 4 | 4 | 100 |
| **Choose Any One (Statistics Elective)** | | | | |
| MDS241A | MULTIVARIATE ANALYSIS | 4 | 4 | 100 |
| MDS241B | STOCHASTIC PROCESS | | | |
| MDS241C | CATEGORICAL DATA ANALYSIS | | | |
| MDS271 | MACHINE LEARNING | 6 | 5 | 150 |
| **Choose Any One (Computer Science Elective)** | | | | |
| MDS272A | HADOOP | 6 | 5 | 150 |
| MDS272B | IMAGE AND VIDEO ANALYTICS | | | |
| MDS272C | INTERNET OF THINGS | | | |
| MDS273 | PROGRAMMING FOR DATA SCIENCE IN R | 6 | 4 | 100 |
| HOLEVEN | HOLISTIC EDUCATION | - | - | - |
| Total | - | 30 | 26 | 700 |

## 3 Semester

| Course Code | Course Title | Hours Per Week | Credits | Marks |
|---|---|---|---|---|
| MDS331 | NEURAL NETWORKS AND DEEP LEARNING | 4 | 4 | 100 |
| **Choose Any One (Statistics Elective)** | | | | |
| MDS341A | TIME SERIES ANALYSIS AND FORECASTING TECHNIQUES | 4 | 4 | 100 |
| MDS341B | BAYESIAN INFERENCE | | | |
| MDS341C | ECONOMETRICS | | | |
| MDS341D | BIO-STATISTICS | | | |
| MDS371 | CLOUD ANALYTICS | 6 | 5 | 150 |
| MDS372 | JAVA PROGRAMMING | 5 | 4 | 100 |
| **Choose Any One (Computer Science Elective)** | | | | |
| MDS373A | NATURAL LANGUAGE PROCESSING | 6 | 5 | 150 |
| MDS373B | WEB ANALYTICS | | | |
| MDS373C | BIO INFORMATICS | | | |
| MDS373D | EVOLUTIONARY ALGORITHMS | | | |
| MDS373E | OPTIMIZATION TECHNIQUES | | | |
| MDS381 | SPECIALIZATION PROJECT | 4 | 2 | 100 |
| MDS382 | SEMINAR | 2 | 1 | 50 |
| Total | - | 31 | 25 | 750 |

## 4 Semester

| Course Code | Course Title | Hours Per Week | Credits | Marks |
|---|---|---|---|---|
| MDS481 | INDUSTRY PROJECT | 2 | 12 | 300 |
| Total | - | 2 | 12 | 300 |

## MDS131-MATHEMATICAL FOUNDATION FOR DATA SCIENCE - I

**Total Teaching Hours For Semester:60** **No of Lecture Hours/Week:4**
**Max Marks:100** **Credits:4**

Course Description and Course Objectives

Linear Algebra plays a fundamental role in the theory of Data Science. This course aims at introducing the basic notions of vector spaces, Linear Algebra and the use of Linear Algebra in applications to Data Science.

Course Outcomes

CO1: Understand the properties of Vector spaces
CO2: Use the properties of Linear Maps in solving problems on Linear Algebra
CO3: Demonstrate proficiency on the topics Eigenvalues, Eigenvectors and Inner Product Spaces
CO4: Apply mathematics for some applications in Data Science

**Unit-1** **Teaching Hours:12**

INTRODUCTION TO VECTOR SPACES
Vector Spaces: Rn and Cn, lists, Fn and digression on Fields, Definition of Vector space Subspaces, sums of Subspaces, Direct Sums, Span and Linear Independence, bases, dimension.

**Unit-2** **Teaching Hours:12**

LINEAR MAPS
DefinitionofLinearMaps-AlgebraicOperationson L(V,W) - Null spaces and Injectivity-RangeandSurjectivity-FundamentalTheoremsofLinearMaps-Representing a Linear Mapbya Matrix-InvertibleL inearMaps-IsomorphicVectorspaces-LinearMap as Matrix Multiplication - Operators - Products of Vector Spaces - Product of Direct Sum - Quotients of Vector spaces.

**Unit-3** **Teaching Hours:12**

EIGENVALUES, EIGENVECTORS, AND INNER PRODUCT SPACES
Eigenvalues and Eigenvectors - Eigenvectors and Upper Triangular matrices - Eigenspaces and Diagonal Matrices - Inner Products and Norms - Linear functionals on Inner Product spaces.

**Unit-4** **Teaching Hours:12**

BASIC MATRIX METHODS FOR APPLICATIONS
Matrix Norms – Least square problem - Singular value decomposition-Householder Transformation and QR decomposition- Non Negative Matrix Factorizatio – bidiagonalization.

**Unit-5** **Teaching Hours:12**

MATHEMATICS APPLIED TO DATA SCIENCE
Handwritten digits recognition using simple algorithm - Classification of handwritten digits using SVD bases and Tangent distance - Text Mining using Latent semantic inde; Clustering, Non-negative Matrix Factorization and LGK bidiagonalization.

Essential References

1. S. Axler, Linear algebra done right, Springer, 2017.
2. Eldén Lars, Matrix methods in data mining and pattern recognition, Society for Industrial and Applied Mathematics, 2007.

Recommended References

1. E. Davis, Linear algebra and probability for computer science applications, CRC Press, 2012.
2. J. V. Kepner and J. R. Gilbert, Graph algorithms in the language of linear algebra, Society for Industrial and Applied Mathematics, 2011.
3. D. A. Simovici, Linear algebra tools for data mining, World Scientific Publishing 2012.
4. P. N. Klein, Coding the matrix: linear algebra through applications to computer science, Newtonian Press, 2015.

Evaluation Pattern

CIA - 50%
ESE - 50%

## MDS132-PROBABILITY AND DISTRIBUTION THEORY

**Total Teaching Hours For Semester:60** **No of Lecture Hours/Week:4**
**Max Marks:100** **Credits:4**

Course Description and Course Objectives

Probability and probability distributions play an essential role in modeling data from the real-world phenomenon. This course will equip students with thorough knowledg in probability and various probability distributions and model real-life data sets with an appropriate probability distribution

Course Outcomes

CO1: Describe random event and probability of events
CO2: Identify various discrete and continuous distributions and their usage.
CO3: Evaluate condition probabilities and conditional expectations
CO4: Apply Chebychev's inequality to verify the convergence of sequence in probability

**Unit-1** **Teaching Hours:12**

DESCRIPTIVE STATISTICS AND PROBABILITY
Data – types of variables: numeric vs categorical - measures of central tendency – measures of dispersion - random experiment - sample space and random events – probability - probability axioms - finite sample space with equally likely outcomes - conditional probability - independent events - Baye's theorem

**Unit-2** **Teaching Hours:12**

PROBABILITY DISTRIBUTIONS FOR DISCRETE DATA
Random variable – data as observed values of a random variable - expectation – moments & moment generating function - mean and variance in terms of moments - discrete sample space and discrete random variable – Bernoulli experiment and Binary variable: Bernoulli and binomial distributions – Count data: Poisson distribution – overdispersion in count data: negative binomial distribution – dependent Bernoulli trails: hypergeometric distribution.

**Unit-3** **Teaching Hours:12**

PROBABILITY DISTRIBUTIONS FOR CONTINUOUS DATA
Continuous sample space - Interval data - continuous random variable – uniform distribution - normal distribution (Gaussian distribution) – modeling lifetime data: exponential distribution, gamma distribution, Weibull distribution.

**Unit-4** **Teaching Hours:12**

JOINTLY DISTRIBUTED RANDOM VARIABLES

Joint distribution of vector random variables – joint moments – covariance – correlation - the correlation - independent random variables - conditional distribution – conditional expectation - sampling distributions: chi-square, t, F (central).

**Unit-5** **Teaching Hours:12**

LIMIT THEOREMS

Chebychev's inequality - weak law of large n u mbers (iid): examples - strong law of large numbers (statement only) - central limit theorems (iid case): examples.

Essential References


1. Ross, Sheldon. A first course in probability. 10th Edition. Pearson, 2019.
2. An Introduction to Probability and Statistics, V.K Rohatgi and Saleh, 3rd Edition, 2015


Recommended References


1. Introduction to the theory of statistics, A.M Mood, F.A Graybill and D.C Boes, Tata McGraw-Hill, 3rd Edition (Reprint), 2017.
2. Ross, Sheldon M. Introduction to probability models. 12th Edition, Academic Press, 2019.


Evaluation Pattern

CIA: 50%

ESE: 50%

## MDS133-PRINCIPLES OF DATA SCIENCE

**Total Teaching Hours For Semester:60** **No of Lecture Hours/Week:4**
**Max Marks:100** **Credits:4**

Course Description and Course Objectives

To provide strong foundation for data science and application area related to information technology and understand the underlying core concepts and emerging technologies in data science

Course Outcomes

CO1:Explore the fundamental concepts of data science
CO2:Understand data analysis techniques for applications handling large data
CO3:Understand various machine learning algorithms used in data science process
CO4:Visualize and present the inference using various tools
CO5:Learn to think through the ethics surrounding privacy, data sharing and algorithmic decision-making

**Unit-1** **Teaching Hours:10**

INTRODUCTION TO DATA SCIENCE
Definition – Big Data and Data Science Hype – Why data science – Getting Past the Hy
– The Current Landscape – Who is Data Scientist? - Data Science Process Overview –
Defining goals – Retrieving data – Data preparation – Data exploration – Data modelin
Presentation.

**Unit-2** **Teaching Hours:12**

BIG DATA
Problems when handling large data – General techniques for handling large data – Case study – Steps in big data – Distributing data storage and processing with Frameworks – Case study.

**Unit-3** **Teaching Hours:12**

MACHINE LEARNING
Machine learning – Modeling Process – Training model – Validating model – Predictin
new observations –Supervised learning algorithms – Unsupervised learning algorithms.

**Unit-4** **Teaching Hours:12**

DEEP LEARNING
Introduction – Deep Feedforward Networks – Regularization – Optimization of Deep Learning – Convolutional Networks – Recurrent and Recursive Nets – Applications of Deep Learning.

**Unit-5** **Teaching Hours:14**

DATA VISUALIZATION
Introduction to data visualization – Data visualization options – Filters – MapReduce – Dashboard development tools – Creating an interactive dashboard with dc.js-summary.
ETHICS AND RECENT TRENDS
Data Science Ethics – Doing good data science – Owners of the data - Valuing different aspects of privacy - Getting informed consent - The Five Cs – Diversity – Inclusion – Future Trends.

Essential References

[1]. Introducing Data Science, Davy Cielen, Arno D. B. Meysman, Mohamed Ali, Manning Publications Co., 1st edition, 2016
[2]. An Introduction to Statistical Learning: with Applications in R, Gareth James, Daniela Witten, Trevor Hastie, Robert Tibshirani, Springer, 1st edition, 2013
[3]. Deep Learning, Ian Goodfellow, Yoshua Bengio, Aaron Courville, MIT Press, 1 edition, 2016
[4]. Ethics and Data Science, D J Patil, Hilary Mason, Mike Loukides, O' Reilly, 1st edition, 2018

Recommended References

[1]. Data Science from Scratch: First Principles with Python, Joel Grus, O'Reilly, 1s edition, 2015
[2]. Doing Data Science, Straight Talk from the Frontline, Cathy O'Neil, Rachel Schutt, O'Reilly, 1st edition, 2013
[3]. Mining of Massive Datasets, Jure Leskovec, Anand Rajaraman, Jeffrey David Ullman, Cambridge University Press, 2nd edition, 2014

Evaluation Pattern

CIA : 50 %
ESE : 50 %

**MDS134-RESEARCH METHODOLOGY**

**Total Teaching Hours For Semester:30** **No of Lecture Hours/Week:2**
**Max Marks:50** **Credits:2**

Course Description and Course Objectives

This course is intended to assist students in planning and carrying out research work.The students are exposed to the basic principles, procedures and techniques of implementing a research project. To introduce the research concept and the variou research methodologies is the main objective. It focuses on finding out the research gap from the literature and encourages lateral, strategic and creative thinking. This course also introduces computer technology and basic statistics required for research and reporting the research outcomes scientifically emphasizing on research ethics.

Course Outcomes

CO1: Understand the essense of research and the necessity of defining a research problem.
CO2: Apply research methods and methodology including research design,data collection, data analysis, and interpretation.
CO3: Create scientific reports according to specified standards.

**Unit-1** **Teaching Hours:8**

RESEARCH METHODOLOGY
Defining research problem:Selecting the problem, Necessity of defining the problem ,Techniques involved in defining a problem- Ethics in Research.

**Unit-2** **Teaching Hours:8**

RESEARCH DESIGN
Principles of experimental design,Working with Literature: Importance, finding literatu Using your resources, Managing the literature, Keep track of references,Using the literature, Literature review,On-line Searching: Database ,SCIFinder, Scopus, Science Direct ,Searching research articles , Citation Index ,Impact Factor ,H-index.

**Unit-3** **Teaching Hours:7**

RESEARCH DATA
Measurement of Scaling: Quantitative, Qualitative, Classification of Measure scales, Da Collection, Data Preparation.

**Unit-4** **Teaching Hours:7**

REPORT WRITING

Scientific Writing and Report Writing: Significance, Steps, Layout, Types, Mechanics and Precautions, Latex: Introduction, Text, Tables, Figures, Equations, Citations, Referencing, and Templates (IEEE style), Paper writing for international journals, Writing scientific report.

Essential References

[1] C. R. Kothari, Research Methodology Methods and Techniques, 3rd. ed. New Delhi: New Age International Publishers, Reprint 2014.
[2] Zina O'Leary, The Essential Guide of Doing Research, New Delhi: PHI, 2005.

Recommended References

[1] J. W. Creswell, Research Design: Qualitative, Quantitative, and Mixed Methods Approaches, 4thed. SAGE Publications, 2014.
[2] Kumar, Research Methodology: A Step by Step Guide for Beginners, 3rd. ed. Indian: PE, 2010.

Additional Information

-

Evaluation Pattern

CIA - 50%
ESE - 50%

**MDS161A-INTRODUCTION TO STATISTICS**

**Total Teaching Hours For Semester:30** **No of Lecture Hours/Week:2**
**Max Marks:50** **Credits:2**

Course Description and Course Objectives

To enable the students to understand the fundamentals of statistics to apply descripti measures and probability for data analysis.

Course Outcomes

CO1: Demonstrate the history of statistics and present the data in various forms.
CO2: Infer the concept of correlation and regression for relating two or more related variables.
CO3: Demonstrate the probabilities for various events.

**Unit-1** **Teaching Hours:8**

ORGANIZATION AND PRESENTATION OF DATA
Origin and development of Statistics, Scope, limitation and misuse of statistics. Types o data: primary, secondary, quantitative and qualitative data. Types of Measurements: nominal, ordinal, discrete and continuous data. Presentation of data by tables: construction of frequency distributions for discrete and continuous data, graphical representation of a frequency distribution by histogram and frequency polygon, cumulative frequency distributions

**Unit-2** **Teaching Hours:8**

DESCRIPTIVE STATISTICS
Measures of location or central tendency: Arthimetic mean, Median, Mode, Geometric mean, Harmonic mean. Partition values: Quartiles, Deciles and percentiles. Measures of dispersion: Mean deviation, Quartile deviation, Standard deviation, Coefficient of variation. Moments: measures of skewness, Kurtosis.

**Unit-3** **Teaching Hours:7**

CORRELATION AND REGRESSION
Correlation: Scatter plot, Karl Pearson coefficient of correlation, Spearman's rank correlation coefficient, multiple and partial correlations (for 3 variates only). Regressior Concept of errors, Principles of Least Square, Simple linear regression and its propertie

**Unit-4** **Teaching Hours:7**

BASICS OF PROBABILITY
Random experiment, sample point and sample space, event, algebra of events. Definitio of Probability: classical, empirical and axiomatic approaches to probability, properties o

probability. Theorems on probability, conditional probability and independent events, Laws of total probability, Baye’s theorem and its applications

Essential References

[1]. Rohatgi V.K and Saleh E, An Introduction to Probability and Statistics, 3rd edition, John Wiley & Sons Inc., New Jersey, 2015.
[2]. Gupta S.C and Kapoor V.K, Fundamentals of Mathematical Statistics, 11th edition, Sultan Chand & Sons, New Delhi, 2014.

Recommended References

[1]. Mukhopadhyay P, Mathematical Statistics, Books and Allied (P) Ltd, Kolkata, 2015.
[2]. Walpole R.E, Myers R.H, and Myers S.L, Probability and Statistics for Enginee and Scientists, Pearson, New Delhi, 2017.
[3]. Montgomery D.C and Runger G.C, Applied Statistics and Probability for Engineers, Wiley India, New Delhi, 2013.
[4]. Mood A.M, Graybill F.A and Boes D.C, Introduction to the Theory of Statistics McGraw Hill, New Delhi, 2008.

Additional Information

-

Evaluation Pattern

CIA - 50%
ESE - 50%

**MDS161B-INTRODUCTION TO COMPUTERS AND PROGRAMMING**

**Total Teaching Hours For Semester:30** **No of Lecture Hours/Week:2**
**Max Marks:50** **Credits:2**

Course Description and Course Objectives

To enable the students to understand the fundamental concepts of problem solving a programming structures.

Course Outcomes

CO1: Demonstrate the systematic approach for problem-solving using computers.
CO2: Apply different programming structures with suitable logic for computational problems.

**Unit-1** **Teaching Hours:10**

COMPUTERS AND DIGITAL BASICS
Number Representation – Decimal, Binary, Octal, Hexadecimal and BCD numbers – Binary Arithmetic – Binary addition – Unsigned and Signed numbers – one's and two's complements of Binary numbers – Arithmetic operations with signed numbers - Numbe system conversions – Boolean Algebra – Logic gates – Design of Circuits – K - Map

**Unit-2** **Teaching Hours:5**

GENERAL PROBLEM SOLVING CONCEPT
Types of Problems – Problem solving with Computers – Difficulties with problem solvi – problem solving concepts for the Computer – Constants and Variables – Rules for Naming and using variables – Data types – numeric data – character data – logical data rules for data types – examples of data types – storing the data in computer - Functions Operators – Expressions and Equations

**Unit-3** **Teaching Hours:5**

PLANNING FOR SOLUTION
Communicating with computer – organizing the solution – Analyzing the problem – developing the interactivity chart – developing the IPO chart – Writing the algorithms – drawing the flow charts – pseudocode – internal and external documentation – testing tl solution – coding the solution – software development life cycle.

**Unit-4** **Teaching Hours:10**

PROBLEM SOLVING
Introduction to programming structure – pointers for structuring a solution – modules ar their functions – cohesion and coupling – problem solving with logic structure. Problem solving with decisions – the decision logic structure – straight through logic – positive logic – negative logic – logic conversion – decision tables – case logic structure -

examples.

Essential References

[1] Thomas L.Floyd and R.P.Jain,“Digital Fundamentals”,8th Edition, Pearson Education,2007.
[2] Peter Norton “Introduction to Computers”,6th Edition, Tata Mc Graw Hill, New Delhi,2006.
[3] Maureen Sprankle and Jim Hubbard, Problem-solving and programming concept PHI, 9th Edition, 2012

Recommended References

[1]. E Balagurusamy, Fundamentals of Computers, TMH, 2011

Additional Information

NA

Evaluation Pattern

CIA: 50%
ESE: 50%

## MDS161C-LINUX ADMINISTRATION

**Total Teaching Hours For Semester:30** **No of Lecture Hours/Week:2**
**Max Marks:50** **Credits:2**

Course Description and Course Objectives

To Enable the students to excel in the Linux Platform

Course Outcomes

CO1: Demostrate the systematic approach for configure the Liux environment
CO2: Manage the Linux environment to work with open source data science tools

**Unit-1** **Teaching Hours:10**

Module-1
RHEL7.5,breaking root password, Understand and use essential tools for handling files, directories, command-line environments, and documentation - Configure local storage using partitions and logical volumes

**Unit-2** **Teaching Hours:10**

Module-2
Swapping, Extend LVM Partitions,LVM Snapshot - Manage users and groups, includin use of a centralized directory for authentication

**Unit-3** **Teaching Hours:10**

Module-3
Kernel updations,yum and nmcli configuration, Scheduling jobs,at,crontab - Configure firewall settings using firewall config, firewall-cmd, or iptables , Configure key-based authentication for SSH ,Set enforcing and permissive modes for SELinux , List and identify SELinux file and process context ,Restore default file contexts

Essential References

1. https://access.redhat.com/documentation/en-US/Red_Hat_Enterprise_Linux/7/
2. https://access.redhat.com/documentation/en-US/Red_Hat_Enterprise_Linux/7/

Recommended References

Additional Information

NA

Evaluation Pattern

CIA:50%
ESE:50%

## MDS171-DATABASE TECHNOLOGIES

**Total Teaching Hours For Semester:90** **No of Lecture Hours/Week:6**
**Max Marks:150** **Credits:5**

Course Description and Course Objectives

The main objective of this course is to fundamental knowledge and practical experience with, database concepts. It includes the concepts and terminologies whicl facilitate the construction of relational databases, writing effective queries comprehend data warehouse and NoSQL databases and its types

Course Outcomes

CO1: Demonstrate various databases and Compose effective queries
CO2: Understanding the process of OLAP system construction
CO3: Develop applications using Relational and NoSQL databases.

**Unit-1** **Teaching Hours:18**

INTRODUCTION
Concept & Overview of DBMS, Data Models, Database Languages, Database Administrator, Database Users, Three Schema architecture of DBMS. Basic concepts, Design Issues, Mapping Constraints, Keys, Entity-Relationship Diagram, Weak Entity Sets, Extended E-R features
Lab Exercises
1. Data Definition,
2. Table Creation
3. Constraints

**Unit-2** **Teaching Hours:18**

RELATIONAL MODEL AND DATABASE DESIGN
SQL and Integrity Constraints, Concept of DDL, DML, DCL. Basic Structure, Set operations, Aggregate Functions, Null Values, Domain Constraints, Referential Integrit Constraints, assertions, views, Nested Subqueries, Functional Dependency, Different anomalies in designing a Database, Normalization: using functional dependencies, Boy Codd Normal Form, 4NF
Lab Exercises
1. Insert, Select, Update & Delete Commands
2. Nested Queries & Join Queries
3. Views

**Unit-3** **Teaching Hours:18**

DATA WAREHOUSE: THE BUILDING BLOCKS
Defining Features, Data Warehouses and Data Marts, Architectural Types, Overview of the Components, Metadata in the Data warehouse, Data Design and Data Preparation: Principles of Dimensional Modeling, Dimensional Modeling Advanced Topics From Requirements To Data Design, The Star Schema, Star Schema Keys, Advantages of the

Star Schema, Star Schema: Examples, Dimensional Modeling: Advanced Topics, Updates to the Dimension Tables, Miscellaneous Dimensions, The Snowflake Schema, Aggregate Fact Tables, Families Oo Stars

Lab Exercises:

1. Importing source data structures
2. Design Target Data Structures
3. Create target multidimensional cube

**Unit-4** **Teaching Hours:18**

DATA INTEGRATION and DATA FLOW (ETL)

Requirements, ETL Data Structures, Extracting, Cleaning and Conforming, Delivering Dimension Tables, Delivering Fact Tables, Real-Time ETL Systems

Lab Exercises:

1. Perform the ETL process and transform into data map
2. Create the cube and process it
3. Generating Reports
4. Creating the Pivot table and pivot chart using some existing data

**Unit-5** **Teaching Hours:18**

NOSQL Databases

Introduction to NOSQL Systems, The CAP Theorem, Document-Based NOSQL Systen and MongoDB, NOSQL Key-Value Stores, Column-Based or Wide Column NOSQL Systems, Graph databases, Multimedia databases.

Lab Exercises:

1. MongoDB Exercise - 1
2. MongoDB Exercise - 2

Essential References

[1]. Henry F. Korth and Silberschatz Abraham, “Database System Concepts”, Mc.Graw Hill.

[2]. Thomas Cannolly and Carolyn Begg, “Database Systems, A Practical Approach Design, Implementation and Management”, Third Edition, Pearson Education, 2007

[3]. The Data Warehouse Toolkit: The Complete Guide to Dimensional Modeling, 2 John Wiley & Sons, Inc. New York, USA, 2002

Recommended References

[1] LiorRokach and OdedMaimon, Data Mining and Knowledge Discovery Handbook, Springer, 2nd edition, 2010.

Additional Information

Evaluation Pattern

CIA: 50%

ESE: 50%

## MDS172-INFERENTIAL STATISTICS

**Total Teaching Hours For Semester:90** **No of Lecture Hours/Week:6**
**Max Marks:150** **Credits:5**

Course Description and Course Objectives

Statistical inference plays an important role in modeling data and decision-making from the real-world phenomenon. This course is designed to impart the knowledge c testing of hypothesis and estimation of parameters for real-life data sets.

Course Outcomes

CO1: Demonstrate the concepts of population and samples.
CO2: Apply the idea of sampling distribution of different statistics in testing of hypothesis
CO3: Test the hypothesis using nonparametric tests for real world problems.
CO4: Estimate the unknown population parameters using the concepts of point and interval estimations.

**Unit-1** **Teaching Hours:18**

INTRODUCTION
Population and Statistics – Finite and Infinite population – Parameter and Statistics – Types of sampling - Sampling Distribution – Sampling Error - Standard Error – Test of significance –concept of hypothesis – types of hypothesis – Errors in hypothesis-testing Critical region – level of significance - Power of the test – p-value.
Lab Exercise:
1. Calculation of sampling error and standard error
2. Calculation of probability of critical region using standard distributions
3. Calculation of power of the test using standard distributions.

**Unit-2** **Teaching Hours:18**

TESTING OF HYPOTHESIS I
Concept of large and small samples – Tests concerning a single population mean for known σ – equality of two means for known σ – Test for Single variance - Test for equality of two variance for normal population – Tests for single proportion – Tests of equality of two proportions for the normal population.

Lab Exercise:
4. Test of the single sample mean for known σ.
5. Test of equality of two means when known σ
6. Tests of single variance and equality of variance for large samples
7. Tests for single proportion and equality of two proportion for large samples.

**Unit-3** **Teaching Hours:18**

TESTING OF HYPOTHESIS II
Students t-distribution and its properties (without proofs) – Single sample mean test – Independent sample mean test – Paired sample mean test – Tests of proportion (based o

distribution) – F distribution and its properties (without proofs) – Tests of equality of two variances using F-test – Chi-square distribution and its properties (without proofs) – chisquare test for independence of attributes – Chi-square test for goodness of fit.

Lab Exercise:
8. Single sample mean test
9. Independent and Paired sample mean test
10. Tests of proportion of one and two samples based on t-distribution
11. Test of equality of two variances
12. Chi-square test for independence of attributes and goodness of fit.

**Unit-4** **Teaching Hours:18**

ANALYSIS OF VARIANCE
Meaning and assumptions - Fixed, random and mixed effect models - Analysis of variance of one-way and two-way classified data with and without interaction effects – Multiple comparison tests: Tukey's method - critical difference.

Lab Exercise:
13. Construction of one-way ANOVA
14. Construction of two-way ANOVA with interaction
15. Construction of two-way ANOVA without interaction
16. Multiple comparision test using Tukey's method and critical difference methods

**Unit-5** **Teaching Hours:18**

NONPARAMETRIC TESTS
Concept of Nonparametric tests - Run test for randomness - Sign test and Wilcoxon Signed Rank Test for one and paired samples - Run test - Median test and Mann-Whitney-Wilcoxon tests for two samples.

Lab Exercise:
17. Test of one sample using Run and sign tests
18. Test of paried sample using Wilcoxon signed rank test
19. Test of two samples using Run test and Median test
20. Test for two samples using Mann-Whitney-Wilcoxon tests

Essential References

1. Gupta S.C and Kapoor V.K, Fundamentals of Mathematical Statistics, 12th edition, Sultan Chand & Sons, New Delhi, 2020.
2. Brian Caffo, Statistical Inference for Data Science, Learnpub, 2016.

Recommended References

1. Walpole R.E, Myers R.H and Myers S.L, Probability and Statistics for Engineers and Scientists, 9th edition, Pearson, New Delhi, 2017.
2. John V, Using R for Introductory Statistics, 2nd edition, CRC Press, Boca Raton, 2014.

3. Rajagopalan M and Dhanavanthan P, Statistical Inference, PHI Learning (P) Ltd, New Delhi, 2012.
4. Rohatgi V.K and Saleh E, An Introduction to Probability and Statistics, 3rd edition, JohnWiley & Sons Inc, New Jersey, 2015.

Evaluation Pattern

CIA: 50%
ESE:50%

**MDS173-PROGRAMMING FOR DATA SCIENCE IN PYTHON**

**Total Teaching Hours For Semester:90** **No of Lecture Hours/Week:6**

**Max Marks:100** **Credits:4**

Course Description and Course Objectives

The objective of this course is to provide comprehensive knowledge of python programming paradigms required for Data Science.

Course Outcomes

CO1: Demonstrate the use of built-in objects of Python

CO2:Demonstrate significant experience with python program development environment

CO3:Implement numerical programming, data handling and visualization through NumPy, Pandas and MatplotLibmodules.

**Unit-1** **Teaching Hours:17**

INTRODUCTION TO PYTHON

Structure of Python Program-Underlying mechanism of Module Execution-Branching a Looping-Problem Solving Using Branches and Loops-Functions - Lists and Mutability-Problem Solving Using Lists and Functions

Lab Exercises

1. Demonstrate usage of branching and loopingstatements
2. Demonstrate Recursivefunctions
3. DemonstrateLists

**Unit-2** **Teaching Hours:17**

SEQUENCE DATATYPES AND OBJECT-ORIENTED PROGRAMMING

Sequences, Mapping and Sets- Dictionaries- -Classes: Classes and Instances-Inheritance Exceptional Handling-Introduction to Regular Expressions using “re” module.

Lab Exercises

1. Demonstrate Tuples andSets
2. DemonstrateDictionaries
3. Demonstrate inheritance and exceptionalhandling
4. Demonstrate use of“re”

**Unit-3** **Teaching Hours:13**

USING NUMPY

Basics of NumPy-Computation on NumPy-Aggregations-Computation on Arrays-Comparisons, Masks and Boolean Arrays-Fancy Indexing-Sorting Arrays-Structured Data: NumPy’s Structured Array.

Lab Exercises

1. DemonstrateAggregation

2. Demonstrate Indexing andSorting

**Unit-4** **Teaching Hours:13**

DATA MANIPULATION WITH PANDAS -I

Introduction to Pandas Objects-Data indexing and Selection-Operating on Data in Panda
Handling Missing Data-Hierarchical Indexing - Combining Data Sets
Lab Exercises
1. Demonstrate handling of missingdata
2. Demonstrate hierarchicalindexing

**Unit-5** **Teaching Hours:17**

DATA MANIPULATION WITH PANDAS -II

Aggregation and Grouping-Pivot Tables-Vectorized String Operations -Working with Time Series-High Performance Pandas- and query()
Lab Exercises
1. Demonstrate usage of Pivottable
2. Demonstrate use of andquery()

**Unit-6** **Teaching Hours:13**

VISUALIZATION AND MATPLOTLIB

Basic functions of matplotlib-Simple Line Plot, Scatter Plot-Density and Contour Plots-Histograms, Binnings and Density-Customizing Plot Legends, Colour Bars-Three-Dimensional Plotting in Matplotlib.
Lab Exercises
1. DemonstrateScatterPlot
2. Demonstrate3Dplotting

Essential References

[1]. Jake VanderPlas ,Python Data Science Handbook - Essential Tools for Working with Data, O'Reily Media,Inc, 2016
[2]. Zhang.Y ,An Introduction to Python and Computer Programming, Springer Publications,2016

Recommended References

[1].JoelGrus,DataSciencefromScratchFirstPrincipleswithPython,O'ReillyMedia,201
[2]. T.R.Padmanabhan, Programming with Python,SpringerPublications,2016

Evaluation Pattern

CIA: 50%
ESE: 50%

**MDS231-MATHEMATICAL FOUNDATION FOR DATA SCIENCE - II**

**Total Teaching Hours For Semester:60** **No of Lecture Hours/Week:4**

**Max Marks:100** **Credits:4**

Course Description and Course Objectives

This course aims at introducing data science related essential mathematics concepts such as fundamentals of topics on Calculus of several variables, Orthogonality, Convex optimization and Graph Theory.

Course Outcomes

CO1: Demonstrate the properties of multivariate calculus
CO2: Use the idea of orthogonality and projections effectively
CO3: Have a clear understanding of Convex Optimization
CO4: Know the about the basic terminologies and properties in Graph Theory

**Unit-1** **Teaching Hours:14**

Calculus of Several Variables
Functions of Several Variables: Functions of two, three variables - Limits and continu in HIgher Dimensions: Limits for functions of two variables, Functions of more than tv variables - Partial Derivatives: partial derivative of functions of two variables, part derivatives of functions of more than two variables, partial derivatives and continui second order partial derivatives - The Chain Rule: chain rule on functions of two, thı variables, chain rule on functions defined on surfaces - Directional Derivative a Gradient vectors: Directional derivatives in a plane, Interpretation of directioı derivative, calculation and gradients, Gradients and tangents to level curves.

**Unit-2** **Teaching Hours:10**

Orthogonality
Perpendicular vectors and Orthogonality - Inner Products and Projections onto line: Projections of Rank one - Projections and Least Squares Approximations - Projecti Matrices - Orthogonal Bases, Orthogonal Matrices and Gram-Schmidt orthogonalizatioı

**Unit-3** **Teaching Hours:12**

Introduction to Convex Optimization
Affine and Convex Sets: Lines and Line segments, affine sets, affine dimensi andrelative interior, convexsets, cones - Hyperplanes and half-spaces - Euclidean ba and ellipsoids- Norm balls and Norm cones - polyhedra - simplexes, Convex h description of polyhedra - The positive semidefinitecone.

**Unit-4** **Teaching Hours:12**

Graph Theory - Basics
Graph Classes: Definition of a Graph and Graph terminology, isomorphism of grapl

Completegraphs, bipartite graphs, complete bipartite graphs-Vertex degree: adjacency and incidence, regular graphs - subgraphs, spanning subgraphs, induced subgraphs, removing or adding edges of a graph, removing vertices from graphs - Graph Operations: Graph Union, intersection, complement, self complement, Paths and Cycles, Connected graphs, Eulerian and HamiltonianGraphs.

**Unit-5** **Teaching Hours:12**

Graph Theory - More concepts

Matrix Representation of Graphs, Adjacency matrices, Incidence Matrices, Trees and its properties, Bridges (cut-edges), spanning trees, weighted Graphs, minimal spanning tree problems, Shortest path problems, cut vertices, cuts, vertex and edge connectivity, Grap Algorithms - Applications of Graph Theory

Essential References


1. M. D. Weir, J. Hass, and G. B. Thomas, Thomas' calculus. Pearson, 2016. (Unit 1)
2. G Strang, Linear Algebra and its Applications, 4th ed., Cengage, 2006. (Unit 2)
3. S. P. Boyd and L.Vandenberghe, Convex optimization.Cambridge Univ. Pr., 2011.(Unit 3)
4. J Clark, D A Holton, A first look at Graph Theory, Allied Publishers India, 1995. (Unit 4)


Recommended References


1.J. Patterson and A. Gibson, Deep learning: a practitioner's approach. O'Reilly Med 2017.

2.S. Sra, S. Nowozin, and S. J. Wright, Optimization for machine learning. MIT Pre 2012.

3.D. Jungnickel, Graphs, networks and algorithms. Springer, 2014.

4.D Samovici, Mathematical Analysis for Machine Learning and Data Mining, Worl Scientific Publishing Co. Pte. Ltd, 2018

5.P. N. Klein, Coding the matrix: linear algebra through applications to computer science. Newtonian Press, 2015.

6.K H Rosen, Discrete Mathematics and its applications, 7th ed., McGraw Hill, 201


Evaluation Pattern

CIA:50%
ESE :50%

**MDS232-REGRESSION ANALYSIS**

**Total Teaching Hours For Semester:60** **No of Lecture Hours/Week:4**
**Max Marks:100** **Credits:4**

Course Description and Course Objectives

This course aims to provide the grounding knowledge about the regression model building of simple and multiple regression.

Course Outcomes

CO1: Demonstrate deeper understanding of the linear regression model.
CO2: Evaluate R-square criteria for model selection
CO3: Understand the forward, backward and stepwise methods for selecting the variables
CO4: Understand the importance of multicollinearity in regression modelling
CO5: Ability touse and understand generalizations of the linear model to binary and count data

**Unit-1** **Teaching Hours:13**

SIMPLE LINEAR REGRESSION
Introduction to regression analysis: Modelling a response, overview and applications of regression analysis, major steps in regression analysis. Simple linear regression (Two variables): assumptions, estimation and properties of regression coefficients, significanc and confidence intervals of regression coefficients, measuring the quality of the fit.

**Unit-2** **Teaching Hours:13**

MULTIPLE LINEAR REGRESSION
Multiple linear regression model: assumptions, ordinary least square estimation of regression coefficients, interpretation and properties of regression coefficient, significan and confidence intervals of regression coefficients.

**Unit-3** **Teaching Hours:12**

CRITERIA FOR MODEL SELECTION
Mean Square error criteria, R2 and criteria for model selection; Need of the transformation of variables; Box-Cox transformation; Forward, Backward and Stepwise procedures.

**Unit-4** **Teaching Hours:12**

RESIDUAL ANALYSIS
Residual analysis, Departures from underlying assumptions, Effect of outliers, Collinearity, Non-constant variance and serial correlation, Departures from normality, Diagnostics and remedies.

**Unit-5** **Teaching Hours:10**

NON LINEAR REGRESSION

Introduction to nonlinear regression, Least squares in the nonlinear case and estimation of parameters, Models for binary response variables, estimation and diagnosis methods for logistic and Poisson regressions. Prediction and residual analysis.

Essential References

[1].D.C Montgomery, E.A Peck and G.G Vining, Introduction to Linear Regression Analysis, John Wiley and Sons,Inc.NY, 2003.
[2]. S. Chatterjee and AHadi, Regression Analysis by Example, 4th Ed., John Wiley and Sons, Inc, 2006
[3].Seber, A.F. and Lee, A.J. (2003) Linear Regression Analysis, John Wiley, Relevant sections from chapters 3, 4, 5, 6, 7, 9, 10.

Recommended References

[1]. Iain Pardoe, Applied Regression Modeling, John Wiley and Sons, Inc, 2012.
[2]. P. McCullagh, J.A. Nelder, Generalized Linear Models, Chapman & Hall, 1989.

Additional Information

-

Evaluation Pattern

CIA - 50%
ESE - 50%

**MDS241A-MULTIVARIATE ANALYSIS**

**Total Teaching Hours For Semester:60** **No of Lecture Hours/Week:4**

**Max Marks:100** **Credits:4**

Course Description and Course Objectives

This course lays the foundation of Multivariate data analysis. The exposure provide to multivariate data structure, multinomial and multivariate normal distribution, estimation and testing of parameters, various data reduction methods would help the students in having a better understanding of research data, its presentation and analysis.

Course Outcomes

CO1: Understand multivariate data structure, multinomial and multivariate normal distribution

CO2: Apply Multivariate analysis of variance (MANOVA) of one and two-way classified data.

**Unit-1** **Teaching Hours:12**

INTRODUCTION

Basic concepts on multivariate variable. Multivariate normal distribution, Marginal and conditional distribution, Concept of random vector: Its expectation and Variance-Covariance matrix. Marginal and joint distributions. Conditional distributions and Independence of random vectors. Multinomial distribution. Sample mean vector and its distribution.

**Unit-2** **Teaching Hours:12**

DISTRIBUTION

Sample mean vector and its distribution. Likelihood ratio tests: Tests of hypotheses abo the mean vectors and covariance matrices for multivariate normal populations. Independence of sub vectors and sphericity test.

**Unit-3** **Teaching Hours:12**

MULTIVARIATE ANALYSIS

Multivariate analysis of variance (MANOVA) of one and two- way classified data. Multivariate analysis of covariance. Wishart distribution, Hotelling's T2 and Mahalanobis' D2 statistics, Null distribution of Hotelling's T2. Rao's U statistics and it distribution.

**Unit-4** **Teaching Hours:12**

CLASSIFICATION AND DISCRIMINANT PROCEDURES

Bayes, minimax, and Fisher's criteria for discrimination between two multivariate norm populations. Sample discriminant function. Tests associated with discriminant functions Probabilities of misclassification and their estimation. Discrimination for several multivariate normal populations

**Unit-5** **Teaching Hours:12**

PRINCIPAL COMPONENT and FACTOR ANALYSIS
Principal components, sample principal components asymptotic properties. Canonical variables and canonical correlations: definition, estimation, computations. Test for significance of canonical correlations.
Factor analysis: Orthogonal factor model, factor loadings, estimation of factor loadings, factor scores. Applications

Essential References


[1]. Anderson, T.W. 2009. An Introduction to Multivariate Statistical Analysis, 3rd Edition, John Wiley.
[2]. Everitt B, Hothorn T, 2011. An Introduction to Applied Multivariate Analysis with R, Springer.
[3]. Barry J. Babin, Hair, Rolph E Anderson, and William C. Blac, 2013, Multivaria Data Analysis, Pearson New International Edition,


Recommended References


[1] Giri, N.C. 1977. Multivariate Statistical Inference. Academic Press.
[2] Chatfield, C. and Collins, A.J. 1982. Introduction to Multivariate analysis. Prenti Hall
[3] Srivastava, M.S. and Khatri, C.G. 1979. An Introduction to Multivariate Statistic North Holland


Additional Information

-

Evaluation Pattern

CIA - 50%
ESE - 50%

**MDS241B-STOCHASTIC PROCESS**

**Total Teaching Hours For Semester:60** **No of Lecture Hours/Week:4**
**Max Marks:100** **Credits:4**

Course Description and Course Objectives

This course is designed to introduce the concepts of theory of estimation and testing hypothesis. This paper also deals with the concept of parametric tests for large and small samples. It also provides knowledge about non-parametric tests and its applications.

Course Outcomes

CO1: Demonstrate the concepts of point and interval estimation of unknown parameters and their significance using large and small samples.
CO2: Apply the idea of sampling distributions of difference statistics in testing of hypotheses.
CO3: Infer the concept of nonparametric tests for single sample and two samples.

**Unit-1** **Teaching Hours:12**

INTRODUCTION TO STOCHASTIC PROCESSES
Classification of Stochastic Processes, Markov Processes – Markov Chain - Countable State Markov Chain. Transition Probabilities, Transition Probability Matrix. Chapman - Kolmogorov's Equations, Calculation of n - step Transition Probability and its limit.

**Unit-2** **Teaching Hours:12**

POISSON PROCESS
Classification of States, Recurrent and Transient States - Transient Markov Chain, Random Walk and Gambler's Ruin Problem. Continuous Time Markov Process:, Poisso Processes, Birth and Death Processes, Kolmogorov's Differential Equations, Applications.

**Unit-3** **Teaching Hours:12**

BRANCHING PROCESS
Branching Processes – Galton – Watson Branching Process - Properties of Generating Functions – Extinction Probabilities – Distribution of Total Number of Progeny. Concep of Weiner Process.

**Unit-4** **Teaching Hours:12**

RENEWAL PROCESS
Renewal Processes – Renewal Process in Discrete and Continuous Time – Renewal Interval – Renewal Function and Renewal Density – Renewal Equation – Renewal theorems: Elementary Renewal Theorem. Probability Generating Function of Renewal Processes.

**Unit-5** **Teaching Hours:12**

STATIONARY PROCESS

Stationary Processes: Discrete Parameter Stochastic Process – Application to Time Series. Auto-covariance and Auto-correlation functions and their properties. Moving Average, Autoregressive, Autoregressive Moving Average, Autoregressive Integrated Moving Average Processes. Basic ideas of residual analysis, diagnostic checking, forecasting.

Essential References


[1]. Stochastic Processes, R.G Gallager, Cambridge University Press, 2013.
[2]. Stochastic Processes, S.M Ross, Wiley India Pvt. Ltd, 2008.


Recommended References


[1]. Stochastic Processes from Applications to Theory, P.D Moral and S. Penev, CRC Press, 2016
[2]. Introduction to Probability and Stochastic Processes with Applications, B..C. Liliana, A Viswanathan, S. Dharmaraja, Wiley Pvt. Ltd, 2012.


Additional Information

-

Evaluation Pattern

CIA - 50%
ESE - 50%

**MDS241C-CATEGORICAL DATA ANALYSIS**

**Total Teaching Hours For Semester:60** **No of Lecture Hours/Week:4**

**Max Marks:100** **Credits:4**

Course Description and Course Objectives

Categorical data analysis deals with the study of information captured through expressions or verbal forms. This course equips the students with the theory and methods to analyse and categorical responses.

Course Outcomes

CO1: Describe the categorical response.
CO2: Identify tests for contingency tables.
CO3: Apply regression models for categorical response variables.
CO4: Analyse contingency tables using log-linear models.

**Unit-1** **Teaching Hours:12**

INTRODUCTION
Categorical response data - Probability distributions for categorical data - Statistical inference for discrete data

**Unit-2** **Teaching Hours:12**

CONTINGENCY TABLES
Probability structure for contingency tables - Comparing proportions with 2x2 tables - The odds ratio - Tests for independence - Exact inference - Extension to three-way and larger tables

**Unit-3** **Teaching Hours:12**

GENERALIZED LINEAR MODELS
Components of a generalized linear model - GLM for binary and count data - Statistical inference and model checking - Fitting GLMs

**Unit-4** **Teaching Hours:12**

LOGISTIC REGRESSION
Interpreting the logistic regression model - Inference for logistic regression - Logistic regression with categorical predictors - Multiple logistic regression - Summarising effec - Building and applying logistic regression models - Multicategory logit models

**Unit-5** **Teaching Hours:12**

LOGLINEAR MODELS FOR CONTINGENCY TABLES
Loglinear models for two-way and three-way tables - Inference for Loglinear models - t log-linear-logistic connection - Independence graphs and collapsibility - Models for matched pairs: Comparing dependent proportions, Logistic regression for matched pairs Comparing margins of square contingency tables - symmetry issues

Essential References

1. Agresti, A. (2012). Categorical Data Analysis, 3rd Edition. New York: Wiley

Recommended References

1. Le, C.T. (2009). Applied Categorical Data Analysis and Translational Research, 2nd Ed., John Wiley and Sons.
2. Agresti, A. (2010). Analysis of ordinal categorical. John Wiley & Sons.
3. Stokes, M. E., Davis, C. S., & Koch, G. G. (2012). Categorical data analysis usin SAS. SAS Institute.
4. Agresti, A. (2018). An introduction to categorical data analysis. John Wiley & Sons.
5. Bilder, C. R., & Loughin, T. M. (2014). Analysis of categorical data with R. Chapman and Hall/CRC.

Evaluation Pattern

CIA:50%
ESE:50%

## MDS271-MACHINE LEARNING

**Total Teaching Hours For Semester:90** **No of Lecture Hours/Week:6**
**Max Marks:150** **Credits:5**

Course Description and Course Objectives

Theobjectiveofthiscourseistoprovideintroductiontotheprinciplesanddesignofmachine learning algorithms. The course is aimed at providing foundations for conceptual aspects of machine learning algorithms along with their applications to solve real world problems.

Course Outcomes

CO1: Understand the basic principles of machine learning techniques.
CO2:Understandhowmachinelearningproblemsareformulatedandsolved.
CO3:Applymachinelearningalgorithmstosolverealworldproblems.

**Unit-1** **Teaching Hours:18**

INRTODUCTION
MachineLearning-ExamplesofMachineApplications-LearningAssociations-Classificatio Regression-UnsupervisedLearning-Reinforcement Learning.Supervised Learning: Learning class from examples- Probably Approach Correct(PAC) Learning-Noise-Learning Multiple classes. Regression-Model Selection and Generalization. IntroductiontoParametricmethods-MaximumLikelihood Estimation:Bernoulli Density-Multinomial Density-Gaussian Density, Nonparametric Density Estimation: Histogram Estimator-Kernel Estimator-K-Nearest NeighbourEstimator.
Lab Exercise:

1. Data Exploration using parametric methods
2. Data Exploration using non-parametric methods
3. Regression analysis

**Unit-2** **Teaching Hours:18**

DIMENSIONALITY REDUCTION
Dimensionality Reduction: Introduction- Subset Selection-Principal Component Analys Feature Embedding-Factor Analysis-Singular Value Decomposition-Multidimensional Scaling-Linear Discriminant Analysis- Bayesian Decision Theory.
Lab Exercise:

1. Data reduction using Principal ComponentAnalysis
2. Data reduction using multi-dimensional scaling

**Unit-3** **Teaching Hours:18**

SUPERVISED LEARNING - I
Linear Discrimination: Introduction- Generalizing the Linear Model-Geometry of the Linear Discriminant- Pairwise Separation-Gradient Descent-Logistic Discrimination. Kernel Machines: Introduction- optical separating hyperplane- v-SVM, kernel tricks-vertical kernel- vertical kernel- defining kernel- multiclass kernel machines- one-class kernel machines.

Lab Exercise

1. Lineardiscrimination
2. Logisticdiscrimination
3. Classification using kernel machines

**Unit-4** **Teaching Hours:18**

SUPERVISED LEARNING - II
Multilayer Perceptron:
Introduction, training a perceptron- learning Boolean functions- multilayer perceptron- backpropogation algorithm- training procedures.
Combining Multiple Learners
Rationale-Generating diverse learners- Model combination schemes- voting, Bagging- Boosting- fine tuning an Ensemble.
Lab Exercise

1. Classification using MLP
2. Ensemble Learning

**Unit-5** **Teaching Hours:18**

UNSUPERVISED LEARNING
Clustering
Introduction-Mixture Densities, K-Means Clustering- Expectation-Maximization algorithm- Mixtures of Latent Varaible Models-Supervised Learning after Clustering- Spectral Clustering- Hierachial Clustering-Clustering- Choosing the number of Clusters
Lab Exercise

1. K means clustering
2. Hierarchical clustering

Essential References

[1]. E. Alpaydin, Introduction to Machine Learning, 3rd Edition, MIT Press, 2014.

Recommended References

1. C.M.Bishop,PatternRecognitionandMachineLearning,Springer,2016.
2. T. Hastie, R. Tibshirani and J. Friedman, The Elements of Statistical Learning: Data Mining, Inference and Prediction, Springer, 2nd Edition,2009
3. K.P.Murphy,MachineLearning:AProbabilisticPerspective,MITPress,2012.

Evaluation Pattern

CIA: 50%
ESE: 50%

**MDS272A-HADOOP**

**Total Teaching Hours For Semester:90** **No of Lecture Hours/Week:6**
**Max Marks:150** **Credits:5**

Course Description and Course Objectives

The subject is intended to give the knowledge of Big Data evolving in every real-tin applications and how they are manipulated using the emerging technologies. This course breaks down the walls of complexity in processing Big Data by providing a practical approach to developing Java applications on top of the Hadoop platform. It describes the Hadoop architecture and how to work with the Hadoop Distributed Fil System (HDFS) and HBase in Ubuntu platform.

Course Outcomes

CO1: Understand the Big Data concepts in real time scenario
CO2: Understand the big data systems and identify the main sources of Big Data in t real world.
CO3: Demonstrate an ability to use Hadoop framework for processing Big Data for Analytics.
CO4: Evaluate the Map reduce approach for different domain problems.

**Unit-1** **Teaching Hours:15**

INTRODUCTION
Distributed file system – Big Data and its importance, Four Vs, Drivers for Big data, Bi data analytics, Big data applications, Algorithms using map reduce, Matrix-Vector Multiplication by Map Reduce.
Apache Hadoop– Moving Data in and out of Hadoop – Understanding inputs and outpu ofMapReduce - Data Serialization, Problems with traditional large-scale systems-Requirements for a new approach-Hadoop – Scaling-Distributed Framework-Hadoop v/ RDBMS-Brief history of Hadoop.

Lab Exercise

1. Installing and Configuring Hadoop

**Unit-2** **Teaching Hours:15**

CONFIGURATIONS OF HADOOP

Hadoop Processes (NN, SNN, JT, DN, TT)-Temporary directory – UI-Common err when running Hadoop cluster, solutions.
Setting up Hadoop on a local Ubuntu host: Prerequisites, downloading Hadoop, setting SSH, configuring the pseudo-distributed mode, HDFS directory, NameNode, Examples MapReduce, Using Elastic MapReduce, Comparison of local versus EMR Hadoop.
Understanding MapReduce:Key/value pairs,TheHadoop Java API for MapRedu Writing MapReduce programs, Hadoop-specific data types, Input/output.
Developing MapReduce Programs: Using languages other than Java with Hado Analysing a large dataset.

Lab Exercise
1. 1. Word count application in Hadoop.
2. 2. Sorting the data using MapReduce.
3. 3. Finding max and min value in Hadoop.

**Unit-3** **Teaching Hours:15**

ADVANCED MAPREDUCE TECHNIQUES
Simple, advanced, and in-between Joins, Graph algorithms, using language-independent data structures.
Hadoop configuration properties - Setting up a cluster, Cluster access control, managing the NameNode, Managing HDFS, MapReduce management, Scaling.
Lab Exercise:
1. Implementation of decision tree algorithms using MapReduce.
2. Implementation of K-means Clustering using MapReduce.
3. Generation of Frequent Itemset using MapReduce.

**Unit-4** **Teaching Hours:15**

HADOOP STREAMING
Hadoop Streaming - Streaming Command Options - Specifying a Java Class as th Mapper/Reducer - Packaging Files With Job Submissions - Specifying Other Plug-ins fo Jobs.
Lab Exercise:
1. 1. Count the number of missing and invalid values through joining two large given datasets.
2. 2. Using hadoop's map-reduce, Evaluating Number of Products Sold in Each Country in the online shopping portal. Dataset is given.
3. 3. Analyze the sentiment for product reviews, this work proposes a MapReduce technique provided by Apache Hadoop.

**Unit-5** **Teaching Hours:15**

HIVE & PIG
Architecture, Installation, Configuration, Hive vs RDBMS, Tables, DDL & DML, Partitioning & Bucketing, Hive Web Interface, Pig, Use case of Pig, Pig Components, Data Model, Pig Latin.
Lab Exercise
1. Trend Analysis based on Access Pattern over Web Logs using Hadoop.
2. Service Rating Prediction by Exploring Social Mobile Users Geographical Locations.

**Unit-6** **Teaching Hours:15**

Hbase
RDBMS VsNoSQL, HBasics, Installation, Building an online query application – Schem design, Loading Data, Online Queries, Successful service.
Hands On: Single Node Hadoop Cluster Set up in any cloud service provider- How to create instance.How to connect that Instance Using putty.InstallingHadoop framework o this instance. Run sample programs which come with Hadoop framework.
Lab Exercise:

1. 1. Big Data Analytics Framework Based Simulated Performance and Operational Efficiencies Through Billons of Patient Records in Hospital System.

Essential References

[1] Boris lublinsky, Kevin t. Smith, Alexey Yakubovich, Professional Hadoop Solutions, Wiley, 2015.
[2] Tom White, Hadoop: The Definitive Guide, O’Reilly Media Inc., 2015.
[3] Garry Turkington, Hadoop Beginner's Guide, Packt Publishing, 2013.

Recommended References

[1] Pethuru Raj, Anupama Raman, DhivyaNagaraj and Siddhartha Duggirala, High-Performance Big-Data Analytics: Computing Systems and Approaches, Springer, 2015.
[2] Jonathan R. Owens, Jon Lentz and Brian Femiano, Hadoop Real-World Solution Cookbook, Packt Publishing, 2013.
[3] Tom White, HADOOP: The definitive Guide, O Reilly, 2012.

Additional Information

-

Evaluation Pattern

CIA - 50%
ESE - 50%

**MDS272B-IMAGE AND VIDEO ANALYTICS**

**Total Teaching Hours For Semester:90** **No of Lecture Hours/Week:6**
**Max Marks:150** **Credits:5**

Course Description and Course Objectives

This course will provide a basic foundation towards digital image processing and video analysis. This course will also provide brief introduction about various Object Detection, Recognition, Segmentation and Compression methods which will help th students to demonstrate real-time image and video analytics applications.

Course Outcomes

CO1: Understand the fundamental principles of image and video analysis
CO2: Apply the image and video analysis approaches to solve real world problems

**Unit-1** **Teaching Hours:18**

INTRODUCTION TO DIGITAL IMAGE AND VIDEO PROCESSING
Digital image representation, Sampling and Quantization, Types of Images, Basic Relations between Pixels - Neighbors, Connectivity, Distance Measures between pixels, Linear and Non Linear Operations, Introduction to Digital Video, Sampled Video, Vide Transmission.
Gray-Level Processing: Image Histogram, Linear and Non-linear point operations on Images, Arithmetic Operations between Images, Geometric Image Operations, Image Thresholding, Region labeling, Binary Image Morphology.
Lab Programs:
1. Program to perform Resize, Rotation of binary, Gray-scale and color images using various methods.
2. Program to implement contrast stretching.

**Unit-2** **Teaching Hours:18**

IMAGE AND VIDEO ENHANCEMENT AND RESTORATION
Spatial domain-Linear and Non-linear Filtering, Introduction to Fourier Transform and the frequency Domain– Filtering in Frequency domain, Homomorphic Filtering, Br introduction towards Wavelets, Wavelet based image denoising, A model of The Image Degradation / Restoration, Noise Models and basic methods for image restoration. Blot detection and Removal.
Lab Programs:
3. Program to implement various image enhancement techniques using Built-in and use defined functions.
4. Program to implement Non-linear Spatial Filtering using Built-in and userdefined functions.

**Unit-3** **Teaching Hours:18**

IMAGE AND VIDEO ANALYSIS
Image Compression: Huffman Coding, Run length Coding, LZW Coding, Basics of Wavelets based image compression.
Video Compression: Basic Concepts and Techniques of Video compression, MPEG-1 a

MPEG-2 Video Standards.
Lab Programs:
5. Program to implement homomorphic Filtering
6. Extraction of frames from videos and analyzing frames

**Unit-4** **Teaching Hours:18**

FEATURE DETECTION AND DESCRIPTION
Introduction to feature detectors, descriptors, matching and tracking, Basic edge detecto – canny, sobel, prewitt etc., Image Segmentation - Region Based Segmentation – Regio Growing and Region Splitting and Merging, Thresholding – Basic global thresholding, optimum global thresholding using Otsu's Method.
Lab Programs:
7. Implement multi-resolution image decomposition and reconstruction using wavelet
8. Implement image compression using wavelets.

**Unit-5** **Teaching Hours:18**

OBJECT DETECTION AND RECOGNITION
Descriptors: Boundary descriptors - Fourier descriptors - Regional descriptors - Topological descriptors - moment invariants
Object detection and recognition in image and video: Minimum distance classifier, K-N classifier and Bayes, Applications in image and video analysis, object tracking in videos
Lab Programs:
9. Extracting feature descriptors from the image dataset.
10. Implement image classification using extracted relevant features.

Essential References

[1] Rafael C. Gonzalez and Richard E. Woods, Digital Image Processing, 4th Editio Pearson Education, 2018.
[2] Alan Bovik, Handbook of Image and Video Processing, Second Edition, Academic Press, 2005.

Recommended References

[1] Anil K Jain, Fundamentals of Digital Image Processing, PHI, 2011.
[2] RichardSzeliski,ComputerVision–AlgorithmsandApplications,Springer,2011.
[3] Oge Marques, Practical Image and Video Processing Using MatLab, Wiley, 2011.
[4] John W. Woods, Multidimensional Signal, Image, Video Processing and Coding, Academic Press, 2006.

Additional Information

-

Evaluation Pattern
CIA: 50%
ESE: 50%

## MDS272C-INTERNET OF THINGS

**Total Teaching Hours For Semester:90** **No of Lecture Hours/Week:6**
**Max Marks:150** **Credits:5**

Course Description and Course Objectives

The explosive growth of the "Internet of Things" is changing our world and the rapi growth of IoT components is allowing people to innovate new designs and products home. Wireless Sensor Networks form the basis of the Internet of Things. To latch c to the applications in the field of IoT of the recent times, this course provides a deep understanding of the underlying concepts of IoT and Wireless Sensor Networks.

Course Outcomes

CO1: Understand the concepts of IoT and IoT enabling technologies
CO2: Gain knowledge on IoT programming and able to develop IoT applications
CO3: Identify different issues in wireless ad hoc and sensor networks
CO4: Develop an understanding of sensor network architectures from a design and performance perspective
CO5: Understand the layered approach in sensor networks and WSN protocols

**Unit-1** **Teaching Hours:18**

Introduction to IOT
Introduction to IoT - Definition and Characteristics, Physical Design Things- Protocols, Logical Design- Functional Blocks, Communication Models- Communication APIs- Introductiontomeasurethephysicalquantities,IoTEnablingTechnologies-WirelessSensor Networks, Cloud Computing Big Data Analytics, Communication Protocols- Embeddec System- IoT Levels and DeploymentTemplates.
Lab Exercises
1. 1. Introduction to ICs and Sensors. A basic program can be shown which makes use of logic gates IC s for understanding the basics of sensor nodes. Different sensors which find application in IoT projects can be shown,their working explained.
2. 2.Introduction to Arduino/Raspberry Pi. Sample sketches or code can be selected from theArduinosoftwareandexecuted,making use of different sensors.

**Unit-2** **Teaching Hours:18**

IOT Programming
Introduction to Smart Systems using IoT - IoT Design Methodology- IoT Boards (Raspberry Pi,Arduino)andIDE-CaseStudy:Weather Monitoring- Logical Designusing Python, Data types & Data Structures- Control Flow, Functions- Modules- Packages, Fi Handling - Date/Time Operations, Classes- Python Packages of Interest forIoT.
Lab Exercises
3. Use of sensors to detect the temperature/humidity in a room and having appropriate actions performed such as changing the LED color and turning the speaker ( as an alarm and using serial monitor to see these values.
4. A basic parking system making use of multiple IR sensors, Ultrasonic Sensors, LED bulbs, Speakers etc, to identify if a slot is empty or full and using the LED and speakers to alert the user about the availability.

**Unit-3** **Teaching Hours:18**

IOT Applications
Home Automation – Smart Cities- Environment, Energy- Retail, Logistics- Agriculture, Industry- Health and Lifestyle- IoT and M2M.
Lab Exercises
5. An Agricultural System (Greenhouse System) that makes use of sensors like humidity temperature etc, to identify the current situation of the agricultural area and taking necessary measures such as activating the water spraying motor, the alarm system (to indicate if there is excess heat) etc.
6. Create a basic sound system by making use of knobs, speakers, LED bulbs etc., to mimic the sound produced by a race car, ambulance, siren etc.
7. A basic obstacle avoiding robot by making use of Ultrasonic sensors, dc motors, and the chassis kit for robotic car.

**Unit-4** **Teaching Hours:18**

Network of wireless sensor nodes
SensingandSensors-WirelessSensorNetworks,ChallengesandConstraints-Applications: Structural Health Monitoring, Traffic Control, Health Care - Node Architecture - Operating system.
Lab Exercise
8. Making use of GSM for communication in the obstacle avoiding robot. Using sensors such as flame sensors, PIR human motion sensor, IR sensor, LED bulbs etc for better inputs regarding the environment.
9. A garbage level indicator which makes use of IR proximity sensors, WiFi modules etc to detect the rising amount of garbage and sending data to a server and channelling that data to the owner of the module. Can be introduced as the application IoT. If needed, IoT introduction can be done much earlier and the sharing of data can be shown, for better functionality of later projects.
10. Elderly care: We want to monitor very senior citizens whether they had a sudden fall. If a very senior citizen falls suddenly while walking, due to stroke or slippery ground etc, a notification should be sent out so that he/she can get immediate medical attention. shown, for better functionality of later projects.

**Unit-5** **Teaching Hours:18**

MAC, Routing and Transport Protocols in WSN
Introduction – Fundamentals of MAC Protocols – MAC protocols for WSN – Sensor MAC CaseStudy–RoutingChallengesandDesignIssues–RoutingStrategies–TransportControl Protocols–TransportProtocolDesignIssues–PerformanceofTransportProtocols
Lab Exercise
11. Smart street lights: The street lights should increase or decrease their intensity based on the actual requirements of the amount of light needed at that time of the day. This will save a lot of energy for the municipal corporation.
12. Implement 3-bit Binary Counter using 3 LED Module.
a. Glow RED if the Binary bit is '0'. Glow GREEN if the binary bit is '1'
i. For example:

ii. 000 = 0 (all LED should be RED)
iii. 001 = 1 (Two LEDs Should be RED , and one LED should be GREEN)
iv. If Button is pressed in between, Reset the counter and Re-start from 0.
Theft prevention system for night: When the room is dark and Board is moved or tilted (say around 90 degree), it should alarm.

Essential References

[1] Arshdeep Bahgaand, Vijay Madisetti, Internet of Things: Hands-on Approach, Hyderabad University Press, 2015.
[2] Kazem Sohraby, Daniel Minoli and TaiebZnati, Wireless Sensor Networks: Technology. Protocols and Application, Wiley Publications, 2010.
[3] Waltenegus Dargie and Christian Poellabauer, Fundamentals of Wireless Sens Networks: Theory and Practice, A John Wiley and Sons Ltd., 2010.

Recommended References

[1] Edgar Callaway, Wireless Sensor Networks: Architecture and Protocols, Auerbach Publications, 2003.
[2] Michael Miller, The Internet of Things, Pearson Education, 2015.
[3] Holger Karl and Andreas Willig, Protocols and Architectures for Wireless Sen Networks, John Wiley & Sons Inc., 2005.
[4] Erdal Çayırcı and Chunming Rong, SecurityinWirelessAdHocandSensorNetworks,John Wiley and Sons, 2009.
[5] Carlos De MoraisCordeiro and Dharma Prakash Agrawal, Ad Hoc and Sensor Networks: Theory and Applications, World Scientific Publishing, 2011.

[6] Adrian Perrig and J.D.Tygar, Secure Broadcast Communication: In Wired and Wireless Networks, Springer, 2006.

Evaluation Pattern

CIA - 50%
ESE - 50%

## MDS273-PROGRAMMING FOR DATA SCIENCE IN R

**Total Teaching Hours For Semester:90** **No of Lecture Hours/Week:6**
**Max Marks:100** **Credits:4**

Course Description and Course Objectives

This lab is designed to introduce implementation of practical machine learning algorithms using R programming language. The lab will extensively use datasets fro real life situations.

Course Outcomes

CO1: Demonstrate to use R in any OS (Windows / Mac / Linux).
CO2: Analyse the use of basic functions of R Package.
CO3: Demonstrate exploratory data analysis (EDA) for a given data set.
CO4: Create and edit visualizations with R
CO5: Implement and assess relevance and effectiveness of machine learning algorithms for a given dataset.

**Unit-1** **Teaching Hours:18**

R INSTALLTION, SETUP AND LINEAR REGRESSION

Download and install R – R IDE environments – Why R – Getting started with R – Vectors and Data Frames – Loading Data Frames – Data analysis with summary statistics and scatter plots – Summary tables - Working with Script Files
Linear Regression – Introduction – Regression model for one variable regression – Selecting best model – Error measures SSE, SST, RMSE, R2 – Interpreting R2 – Multiple linear regression – Lasso and ridge regression – Correlation – Recitation – A minimum of 3 data sets for practice

**Unit-2** **Teaching Hours:18**

LOGISTIC REGRESSION

Logistic Regression – The Logit – Confusion matrix – sensitivity, specificity – ROC curve – Threshold selection with ROC curve – Making predictions – Area under the ROC curve (AUC) - Recitation – A minimum of 3 data sets for practice

**Unit-3** **Teaching Hours:18**

DECISION TREES
Approaches to missing data – Data imputation – Multiple imputation – Classification an Regression Tress (CART) – CART with Cross Validation – Predictions from CART – ROC curve for CART – Random Forests – Building many trees – Parameter selection –

K-fold Cross Validation – Recitation – A minimum of 3 data sets for practice

**Unit-4** **Teaching Hours:18**

TEXT ANALYTICS AND NLP

Using text as data – Text analytics – Natural language processing – Bag of words – Stemming – word clouds – Recitation – min 3 data sets for practice – Time series analysis – Clustering – k-mean clustering – Random forest with clustering – Understanding cluster patterns – Impact of clustering – Heatmaps – Recitation – min 3 data sets for practice

**Unit-5** **Teaching Hours:18**

ENSEMBLE MODELLING
Support Vector Machines – Gradient Boosting – Naive Bayes - Bayesian GLM – GLMNET - Ensemble modeling – Experimenting with all of the above approaches (Uni 1-5) with and without data imputation and assessing predictive accuracy – Recitation – min 3 data sets for practice PROJECT – A concluding project work carried out individually for a common data set

Essential References

[1]. Statistics : An Introduction Using R, Michael J. Crawley, WILEY, Second Edition, 2015.

Recommended References

[1].Hands-on programming with R, Garrett Grolemund, O'Reilley, 1st Edition, 2014
[2]. R for everyone, Jared Lander, Pearson, 1st Edition, 2014

Evaluation Pattern

CIA - 50%
ESE - 50%

## MDS331-NEURAL NETWORKS AND DEEP LEARNING

**Total Teaching Hours For Semester:60** **No of Lecture Hours/Week:4**
**Max Marks:100** **Credits:4**

Course Description and Course Objectives

The main aim of this course is to provide fundamental knowledge of neural network and deep learning. On successful completion of the course, students will acquire fundamental knowledge of neural networks and deep learning, such as Basics of neural networks, shallow neural networks, deep neural networks, forward & backwa propagation process and build various research projects

Course Outcomes

CO1: Understand the major technology trends in neural networks and deep learning
CO2: Build, train and apply neural networks and fully connected deep neural netwoı
CO3: Implement efficient (vectorized) neural networks for real time application

**Unit-1** **Teaching Hours:12**

INTRODUCTION TO ARTIFICIAL NEURAL NETWORKS
Neural Networks-Application Scope of Neural Networks- Fundamental Concept of ANI The Artificial Neural Network-Biological Neural Network-Comparison between Biological Neuron and Artificial Neuron-Evolution of Neural Network. Basic models oı ANN-Learning Methods-Activation Functions-Importance Terminologies of ANN.

**Unit-2** **Teaching Hours:12**

SUPERVISED LEARNING NETWORK
Shallow neural networks- Perceptron Networks-Theory-Perceptron Learning RuleArchitecture-Flowchart for training Process-Perceptron Training Algorithm for Single and Multiple Output Classes.
Back Propagation Network- Theory-Architecture-Flowchart for training process-Trainiı Algorithm-Learning Factors for Back-Propagation Network.
Radial Basis Function Network RBFN: Theory, Architecture, Flowchart and Algorithm.

**Unit-3** **Teaching Hours:12**

CONVOLUTIONAL NEURAL NETWORK
Introduction - Components of CNN Architecture - Rectified Linear Unit (ReLU) Layer Exponential Linear Unit (ELU, or SELU) - Unique Properties of CNN -Architectures oı CNN -Applications of CNN.

**Unit-4** **Teaching Hours:12**

RECURRENT NEURAL NETWORK
Introduction- The Architecture of Recurrent Neural Network- The Challenges of Trainiı Recurrent Networks- Echo-State Networks- Long Short-Term Memory (LSTM) - Applications of RNN.

**Unit-5** **Teaching Hours:12**

AUTO ENCODER AND RESTRICTED BOLTZMANN MACHINE
Introduction - Features of Auto encoder Types of Autoencoder Restricted Boltzmann Machine- Boltzmann Machine - RBM Architecture -Example - Types of RBM.

Essential References

1. S.N.Sivanandam, S. N. Deepa, Principles of Soft Computing, Wiley-India, 3rd Edition, 2018.
2. Dr. S Lovelyn Rose, Dr. L Ashok Kumar, Dr. D Karthika Renuka, Deep Learning Using Python, Wiley-India, 1st Edition, 2019.

Recommended References

1. Charu C. Aggarwal, Neural Networks and Deep Learning, Springer, September 2018.
2. Francois Chollet, Deep Learning with Python, Manning Publications; 1st edition, 2017
3. John D. Kelleher, Deep Learning (MIT Press Essential Knowledge series), The M Press, 2019.

Evaluation Pattern

CIA: 50%
ESE: 50%

## MDS341A-TIME SERIES ANALYSIS AND FORECASTING TECHNIQUES

**Total Teaching Hours For Semester:60** **No of Lecture Hours/Week:4**
**Max Marks:100** **Credits:4**

Course Description and Course Objectives

This course covers applied statistical methods pertaining to time series and forecasti techniques. Moving average models like simple, weighted and exponential are dealt with. Stationary time series models and non-stationary time series models like AR, MA, ARMA and ARIMA are introduced to analyse time series data.

Course Outcomes

CO1: Ability to approach and analyze univariate time series
CO2: Able to differentiate between various time series models like AR, MA, ARMA and ARIMA models
CO3: Evaluate stationary and non-stationary time series models
CO4: Able to forecast future observations of the time series.

**Unit-1** **Teaching Hours:12**

INTRODUCTION TO TIME SERIES AND STOCHASTIC PROCESS
Introduction to time series and stochastic process, graphical representation, components and classical decomposition of time series data.Auto-covariance and auto-correlation functions, Exploratory time series analysis, Test for trend and seasonality, Smoothing techniques such as Exponential and moving average smoothing, Holt- Winter smoothing Forecasting based on smoothing.

**Unit-2** **Teaching Hours:12**

STATIONARY TIME SERIES MODELS
Wold representation of linear stationary processes, Study of linear time series models: Autoregressive, Moving Average and Autoregressive Moving average models and their statistical properties like ACF and PACF function.

**Unit-3** **Teaching Hours:12**

ESTIMATION OF ARMA MODELS
Estimation of ARMA models: Yule- Walker estimation of AR Processes, Maximum likelihood and least squares estimation for ARMA Processes, Residual analysis and diagnostic checking.

**Unit-4** **Teaching Hours:12**

NON-STATIONARY TIME SERIES MODELS
Concept of non-stationarity, general unit root tests for testing non-stationarity; basic formulation of the ARIMA Model and their statistical properties-ACF and PACF; forecasting using ARIMA models

**Unit-5** **Teaching Hours:12**

STATE SPACE MODELS
Filtering, smoothing and forecasting using state space models, Kalman smoother, Maximum likelihood estimation, Missing data modifications

Essential References

1. George E. P. Box, G.M. Jenkins, G.C. Reinsel and G. M. Ljung, Time Series analysis Forecasting and Control, 5th Edition, John Wiley & Sons, Inc., New Jersey, 2016.
2. Montgomery D.C, Jennigs C. L and Kulachi M,Introduction to Time Series analysis  and Forecasting, 2nd Edition,John Wiley & Sons, Inc., New Jersey, 2016.

Recommended References

1. Anderson T.W,Statistical Analysis of Time Series, John Wiley& Sons, Inc., N Jersey, 1971.
2. Shumway R.H and Stoffer D.S, Time Series Analysis and its Applications wit R Examples, Springer, 2011.
3. P. J. Brockwell and R. A. Davis, Times series: Theory and Methods, 2nd Edition, Springer-Verlag, 2009.
4. S.C. Gupta and V.K. Kapoor, Fundamentals of Applied Statistics, 4th Edition, Sultan Chand and Sons, 2008.

Additional Information

NA

Evaluation Pattern

CIA: 50%
ESE: 50%

**MDS341B-BAYESIAN INFERENCE**

**Total Teaching Hours For Semester:60** **No of Lecture Hours/Week:4**
**Max Marks:100** **Credits:4**

Course Description and Course Objectives

To equip the students with the knowledge of conceptual, computational, and practica methods of Bayesian data analysis.

Course Outcomes

CO1: Understand Bayesian models and their specific model assumptions.
CO2: Identify suitable informative and non-informative prior distributions to derive posterior distributions
CO3: Apply computer intensive methods like MCMC for approximating the posteric distribution.
CO4: Analyse the results obtained by Bayesian methods.

**Unit-1** **Teaching Hours:12**

INTRODUCTION
Basics on minimaxity: subjective and frequents probability, Bayesian inference, Bayesia estimation , prior distributions, posterior distribution, loss function, principle of minimu expected posterior loss, quadratic and other common loss functions, Advantages of bein a Bayesian HPD confidence intervals, testing, credible intervals, prediction of a future observation.

**Unit-2** **Teaching Hours:12**

BAYESIAN ANALYSIS WITH PRIOR INFORMATION
Robustness and sensitivity, classes of priors, conjugate class, neighbourhood class, density ratio class different methods of objective priors: Jeffrey's prior, probability matching prior, conjugate priors and mixtures, posterior robustness: measures and techniques

**Unit-3** **Teaching Hours:12**

MULTIPARAMETER AND MULTIVARIABLE MODELS
Basics of decision theory, multi-parameter models, Multivariate models, linear regressic asymptotic approximation to posterior distributions

**Unit-4** **Teaching Hours:12**

MODEL SELECTION AND HYPOTHESIS TESTING
Selection criteria and testing of hypothesis based on objective probabilities and Bayes' factors, large sample methods: limit of posterior distribution, consistency of posterior distribution, asymptotic normality of posterior distribution.

**Unit-5** **Teaching Hours:12**

BAYESIAN COMPUTATIONS
Analytic approximation, E- M Algorithm, Monte Carlo sampling, Markov Chain Monte Carlo Methods, Metropolis – Hastings Algorithm, Gibbs sampling, examples, convergence issues

Essential References

1. Albert Jim (2009) Bayesian Computation with R, second edition, Springer, New York
2. Bolstad W. M. and Curran, J.M. (2016) Introduction to Bayesian Statistics 3rd Ed Wiley, New York
3. Christensen R. Johnson, W. Branscum A. and Hanson T.E. (2011) Bayesian Ideas and data analysis : A introduction for scientist and Statisticians, Chapman and Hall, London
4. A. Gelman, J.B. Carlin, H.S. Stern and D.B. Rubin (2004). Bayesian Data Analys 2nd Ed. Chapman & Hall

Recommended References

1. Congdon P. (2006) Bayesian Statistical Modeling, Wiley, New York.
2. Ghosh, J.K. Delampady M. and T. Samantha (2006). An Introduction to Bayesian Analysis: Theory and Methods, Springer, New York.
3. Lee P.M. (2012) Bayesian Statistics: An Introduction-4th Ed. Hodder Arnold, Ne York.
4. Rao C.R. Day D. (2006) Bayesian Thinking, Modeling and Computation, Handbo of Statistics, Vol.25.

Additional Information

NA

Evaluation Pattern

CIA: 50%
ESE: 50%

## MDS341C-ECONOMETRICS

**Total Teaching Hours For Semester:60** **No of Lecture Hours/Week:4**
**Max Marks:100** **Credits:4**

Course Description and Course Objectives

The course is designed to impart the learning of principles of econometric methods and tools. This is expected to improve student's ability to understand of econometric in the study of economics and finance. The learning objective of the course is to provide students to get the basic knowledge and skills of econometric analysis, so th they should be able to apply it to the investigation of economic relationships and processes, and also understand the econometric methods, approaches, ideas, results and conclusions met in the majority of economic books and articles. Introduce the students to the traditional econometric methods developed mostly for the work with cross-sections data.

Course Outcomes

CO1: Demonstrate Simple and multiple Econometric models
CO2: Interpret the models adequacy through various methods
CO3: Demonstrate simultaneous Linear Equations model.

**Unit-1** **Teaching Hours:15**

INTRODUCTION
Introduction to Econometrics- Meaning and Scope – Methodology of Econometrics – Nature and Sources of Data for Econometric analysis – Types of Econometrics

**Unit-2** **Teaching Hours:15**

CORRELATION
Aitken's Generalised Least Squares(GLS) Estimator, Heteroscedasticity, Auto-correlation, Multicollinearity, Auto-Correlation, Test of Auto-correlation, Multicollinearity, Tools for Handling Multicollinearity

**Unit-3** **Teaching Hours:15**

REGRESSION
Linear Regression with Stochastic Regressors, Errors in Variable Models and Instrumental Variable Estimation, Independent Stochastic linear Regression, Auto regression, Linear regression, Lag Models

**Unit-4** **Teaching Hours:15**

LINEAR EQUATIONS MODEL
Simultaneous Linear Equations Model : Structure of Linear Equations Model, Identification Problem, Rank and Order Conditions, Single Equation and Simultaneous Equations, Methods of Estimation- Indirect Least squares, Least Variance Ratio and Tw Stage Least Square

Essential References

1. Johnston, J. (1997). Econometric Methods, Fourth Edition, McGraw Hill
2. Gujarathi, D., and Porter, D. (2008). Basic Econometrics, Fifth Edition, McGraw-Hill

Recommended References

1. Intriligator, M. D. (1980). Econometric Models-Techniques and Applications, Prentice Hall.
2. Theil, H. (1971). Principles of Econometrics, John Wiley.
3. Walters, A. (1970). An Introduction to Econometrics, McMillan and Co.

Additional Information

NA

Evaluation Pattern

CIA : 50%
ESE : 50%

**MDS341D-BIO-STATISTICS**

**Total Teaching Hours For Semester:60** **No of Lecture Hours/Week:4**

**Max Marks:100** **Credits:4**

Course Description and Course Objectives

This course provides an understanding of various statistical methods in describing ar analyzing biological data. Students will be equipped with an idea about the applications of statistical hypothesis testing, related concepts and interpretation in biological data.

Course Outcomes

CO1: Demonstrate the understanding of basic concepts of biostatistics and the proce involved in the scientific method of research.
CO2: Identify how the data can be appropriately organized and displayed.
CO3: Interpret the measures of central tendency and measures of dispersion.
CO4: Interpret the data based on the discrete and continuous probability distribution
CO5: Apply parametric and non-parametric methods of statistical data analysis.

**Unit-1** **Teaching Hours:12**

INTRODUCTION TO BIOSTATISTICS
Presentation of data - graphical and numerical representations of data - Types of variables, measures of location - dispersion and correlation - inferential statistics - probability and distributions - Binomial, Poisson, Negative Binomial, Hyper geometric and normal distribution.

**Unit-2** **Teaching Hours:12**

PARAMETRIC AND NON - PARAMETRIC METHODS
Parametric methods - one sample t-test - independent sample t-test - paired sample t-test one-way analysis of variance - two-way analysis of variance - analysis of covariance - repeated measures of analysis of variance - Pearson correlation coefficient - Non-parametric methods: Chi-square test of independence and goodness of fit - Mann Whitn U test - Wilcoxon signed-rank test - Kruskal Wallis test - Friedman’s test - Spearman’s correlation test.

**Unit-3** **Teaching Hours:12**

GENERALIZED LINEAR MODELS
Review of simple and multiple linear regression - introduction to generalized linear models - parameter estimation of generalized linear models - models with different link functions - binary (logistic) regression - estimation and model fitting - Poisson regressic for count data - mixed effect models and hierarchical models with practical examples.

**Unit-4** **Teaching Hours:12**

EPIDEMIOLOGY
Introduction to epidemiology, measures of epidemiology, observational study designs:

case report, case series correlational studies, cross-sectional studies, retrospective and prospective studies, analytical epidemiological studies-case control study and cohort study, odds ratio, relative risk, the bias in epidemiological studies.

**Unit-5** **Teaching Hours:12**

DEMOGRAPHY

Introduction to demography, mortality and life tables, infant mortality rate, standardized death rates, life tables, fertility, crude and specific rates, migration-definition and concepts population growth, measurement of population growth-arithmetic, geometric and exponential, population projection and estimation, different methods of population projection, logistic curve, urban population growth, components of urban population growth.

Essential References

1. Marcello Pagano and Kimberlee Gauvreau (2018), Principles of Biostatistics, 2nd Edition, Chapman and Hall/CRC press
2. David Moore S. and George McCabe P., (2017) Introduction to practice of statistics, 9th Edition, W. H. Freeman.
3. Sundar Rao and Richard J., (2012) Introduction to Biostatistics and research methods, PHI Learning Private limited, New Delhi

Recommended References

1. Abhaya Indrayan and Rajeev Kumar M., (2018) Medical Biostatistics, 4th Edition Chapman and Hall/CRC Press.
2. Gordis Leon (2018), Epidemiology, 6th Edition, Elsevier, Philadelphia
3. Ram, F. and Pathak K. B., (2016): Techniques of Demographic Analysis, Himalaya Publishing house, Bombay.
4. Park K., (2019), Park's Text Book of Preventive and Social Medicine, Banarsidas Bhanot, Jabalpur.

Additional Information

-

Evaluation Pattern

CIA:50%
ESE:50%

**MDS371-CLOUD ANALYTICS**

**Total Teaching Hours For Semester:90** **No of Lecture Hours/Week:6**
**Max Marks:150** **Credits:5**

Course Description and Course Objectives

The objective of this course is to explore the basics of cloud analytics and the major cloud solutions. Students will learn how to analyze extremely large data sets, and to create visual representations of that data. Also aim to provide students with hands-on experience working with data at scale.

Course Outcomes

CO1: Interpret the deployment and service models of cloud applications.
CO2: Describe big data analytical concepts.
CO3: Ingest, store, and secure data.
CO4: Process and Visualize structured and unstructured data.

**Unit-1** **Teaching Hours:18**

INTRODUCTION
Introduction to cloud computing - Major benefits of cloud computing - Cloud computing deployment models - Private cloud - Public cloud - Hybrid cloud - Types of cloud computing services -Infrastructure as a Service – PaaS – SaaS - Emerging cloud technologies and services - Different ways to secure the cloud - Risks and challenges with the cloud - What is cloud analytics? Parameters before adopting cloud strategy - Technologies utilized by cloud computing
1.Creating Virtual Machines using Hypervisors
2.IaaS: Compute service - Creating and running Virtual Machines

**Unit-2** **Teaching Hours:18**

CLOUD ENABLING TECHNOLOGIES
Virtualization - Load Balancing - Scalability & Elasticity – Deployment –Replication – Monitoring - Software Defined Networking - Network Function Virtualization – MapReduce - Identity and Access Management - Service Level Agreements - Billing
1. Storage as a Service: Ingesting & Querying data into cloud
2. Database as a Service: Building DB Server

**Unit-3** **Teaching Hours:18**

BASIC CLOUD SERVICES & PLATFORMS
Compute Services
Amazon Elastic Compute Cloud - Google Compute Engine - Windows Azure Virtual Machines
Storage Services
Amazon Simple Storage Service - Google Cloud Storage - Windows Azure Storage
Database Services
Amazon Relational Data Store - Amazon DynamoDB - Google Cloud SQL - Google Cloud Datastore - Windows Azure SQL Database - Windows Azure Table Service

1. PaaS: Working with GoogleAppEngine

**Unit-4** **Teaching Hours:18**

DATA INGESTION AND STORING
Cloud Dataflow - The Dataflow programming model - Cloud Pub/Sub - Cloud storage - Cloud SQL - Cloud BigTable - Cloud Spanner - Cloud Datastore - Persistent disks
1. Database as a Service: Building DB Server
2. Transforming data
PROCESSING AND VISUALIZING
Google BigQuery - Cloud Dataproc - Google Cloud Datalab - Google Data Studio
1. Visualize structured data and unstructureddata

**Unit-5** **Teaching Hours:18**

MACHINE LEARNING, DEEP LEARNING AND AI
Services on Artificial intelligence - Machine learning - Cloud Natural Language API – TensorFlow - Cloud Speech API - Cloud Translation API - Cloud Vision API - Cloud Video Intelligence – Dialogflow – AutoML
1. Load and query data in a data warehouse
2. Setting up and executing a data pipeline job to load data into cloud

Essential References

1. Sanket Thodge, Cloud Analytics with Google Cloud Platform, Packt Publishing, 18.
2. Arshdeep Bahga and Vijay Madisetti, Cloud computing - A Hands-On Approach Create Space Independent Publishing Platform, 2014.

Recommended References

1. Deven Shah, Kailash Jayaswal, Donald J. Houde, Jagannath Kallakurchi, Clouc Computing - Black Book, Wiley, 2014.
2. Thomas Erl, Ricardo Puttini, Zaigham Mahmood, Cloud Computing: Concepts Technology & Architecture, Prentice Hall, 2014.

Additional Information

NA

**MDS372-JAVA PROGRAMMING**

**Total Teaching Hours For Semester:75**

**Max Marks:100** **Credits: 4**

Course Description and Course Objectives

This course of study builds on the skills gained by students in Java Fundamentals to help them to apply Java programming skills in Data science applications. Students w design object-oriented applications with Java and will create Java programs using hands-on, engaging activities.This course will help the learner to gain a sound knowledge in object-oriented principles, GUI application design with data base and Servlets.

Course Outcomes

CO1: Understanding and applying the principles and practice of object-oriented programming in the construction of robust maintainable programs

CO2: Competence in the use of Java Programming Language in the development of small to medium sized applications that demonstrate professionally acceptable codin and performance standards

CO3: To prepare the students to address the challenging requirements coming from the enterprise applications.

**Unit-1**

INTRODUCTION

OVERVIEW OF JVM AND JAVA BASICS

Overview of JVM

Introduction to JVM-JVM Architecture-JDK&JRE-Class Loader-Overview of Bootstra Extension and Application Class Loader

Java Basics

Class and Object Concept-Method Overloading and Overriding-Constructor-this and static keyword-finalize () method in java

**Unit-2**

INHERITANCE, INTERFACES & PACKAGES AND EXCEPTION HANDLING IN JAVA

Inheritance in Java

Inheritance Basics - Multilevel Hierarchy- Using super - Dynamic Method Dispatch- Abstract keyword- Using final with inheritance – Aggregation and Composition in Java

Interfaces and Packages

Defining Interfaces - Implementing Interfaces - Extending Interfaces- Creating Packages - Importing Packages - Interfaces in a Package.

Exception Handling in Java

try-catch-finally mechanism - throw statement - throws statement - Built-in-Exceptions – Custom Exceptions.

**Unit-3**

MULTITHREADING, GENERICS AND THE COLLECTIONS FRAMEWORK
Multithreading Java
Thread Model - Life cycle of a Thread - Java Thread Priorities - Runnable interface and Thread Class- Thread Synchronization – Inter Thread Communication.
Generics
Generics Concept - General Form of a Generic Class – Bounded Types – Generic Class Hierarchy - Generic Interfaces – Restrictions in Generics
The Collections Framework
The Collections Overview – Collection Interface – List Interface – Set Interface – SortedSet Interface – Queue Interface - ArrayList Class – LinkedList Class – HashSet Class – Using an Iterator – The For Each Statement

**Unit-4**

INTRODUCING GUI PROGRAMING WITH SWING, EVENT HANDLING
Introducing GUI Programing with Swing
Swing Basics – Components and Containers – JLabel and ImageIcons- JTextField – Swing Buttons – JTabbedPane – JScrollPane – JList – JComboBox – JTable – Swing Menus
Event Handling
Delegation Event Model - Event Classes – Key Event Class – Event Listener Interface - Adapter Classes

**Unit-5**

DATABASE PROGRAMMING AND DATA SCIENCE WITH JAVA
Database Programming
Connecting to and querying a database –Connecting to the database - Creating a Statement for executing query - Executing a query - Processing a Query’s ResultSet – PreparedStatements.
Data Science with Java
Importance of JAVA in Data Science-Creating Simple Plots-Plotting Mixed Chart Types-Saving a Plot to a File

Lab Exercises:
1. Implement the concept of class, data members, member functions and access specifiers.
2. Implement the concept of function overloading & Constructor overloading
3. Implement the static keyword – static variable, static block, static function and static class
4. Implement String and String Buffer classes.
5. Implement this keyword and command line arguments.
6. Implement the concept of inheritance, super, abstract and final keywords
7. Implement package and interface
8. Implement Exception Handing in java
9. Implement multithreading – Thread class, Runnable interface, thread synchronization and thread communication.
10. Implement collection Interfaces and classes
11. Implement basic CRUD operations in JDBC with SWING
12. Visualizing Data with Plots

13. Implement Java Servlets

Essential References

1. Schildt Herbert, Java: The Complete Reference, Tata McGraw-Hill, 12th Edition, 2021.
2. Michael R. Brzustowicz, Data Science with Java: Practical Methods for Scientists and Engineers, Shroff/O'Reilly; 1st edition,2017

Recommended References

1. Paul Deitel, Java How to Program, Pearson Education Asia, 11th Edition, 2017
2. Cay S Horstmann, Core Java Volume 1 Fundamentals, Prentice Hall, 11th Edition 2018.

1. www.w3cschools.com
2. www.javatpoint.com
3. http://stackoverflow.com/

Additional Information

NA

# MDS373A: Natural Language Processing

**Total Teaching Hours for Semester : 90**
**Max Marks: 150** **Credits:05**

**Course Objectives**

The goal is to make familiar with the concepts of the study of human language from a computational perspective. It covers syntactic, semantic and discourse processing models, emphasizing machine learning concepts.

**Course Outcomes**
**CO1:** Understand various approaches on syntax and semantics in NLP
**CO2:** Apply various methods to discourse, generation, dialogue and summarization using NLP.
**CO3:** Analyze various methodologies used in Machine Translation, machine learning techniques used in NLP including unsupervised models and to analyze real time applications

**Unit-1** **Teaching Hours:12**
**INTRODUCTION**
Introduction to NLP- Background and overview- NLP Applications -NLP hard Ambiguity- Algorithms and models, Knowledge Bottlenecks in NLP- Introduction to NLTK, Case study

**Unit-2** **Teaching Hours:12**
**PARSING AND SYNTAX**
Word Level Analysis: Regular Expressions, Text Normalization, Edit Distance, Parsing and Syntax- Spelling, Error Detection and correction-Words and Word classes- Part-of Speech Tagging, Naive Bayes and Sentiment Classification: Case study

**Unit-3** **Teaching Hours:12**
**SMOOTHED ESTIMATION AND LANGUAGE MODELLING**
N-gram Language Models: N-Grams, Evaluating Language Models -The language modelling problem
**SEMANTIC ANALYSIS AND DISCOURSE PROCESSING**
Semantic Analysis: Meaning Representation-Lexical Semantics- Ambiguity-Word Sense Disambiguation. Discourse Processing: cohesion-Reference Resolution- Discourse Coherence and Structure.

**Unit-4** **Teaching Hours:12**
**NATURALLANGUAGE GENERATION AND MACHINE TRANSLATION**
Natural Language Generation: Architecture of NLG Systems, Applications
Machine Translation: Problems in Machine Translation- Machine Translation Approaches- Evaluation of Machine Translation systems.

Case study: Characteristics of Indian Languages

**Unit-5** **Teaching Hours:12**

**INFORMATION RETRIEVAL AND LEXICAL RESOURCES**

Information Retrieval: Design features of Information Retrieval Systems-Classical, Non-classical, Alternative Models of Information Retrieval – valuation Lexical Resources: Word Embeddings - Word2vec- Glove.

**UNSUPERVISED METHODS IN NLP** Graphical Models for Sequence Labelling in NLP

**Lab Exercises:** **Total Hours:30**

1. Write a program to tokenize text
2. Write a program to count word frequency and to remove stop words
3. Write a program to program to tokenize Non-English Languages
4. Write a program to get synonyms from WordNet
5. Write a program to get Antonyms from WordNet
6. Write a program for stemming Non-English words
7. Write a program for lemmatizing words Using WordNet
8. Write a program to differentiate stemming and lemmatizing words
9. Write a program for POS Tagging or Word Embeddings.
10. Case study-based program (IBM) or Sentiment analysis

**Essential Reading**

1. Speech and Language Processing, Daniel Jurafsky and James H., 2nd Edition, Martin Prentice Hall,2013.

2. Foundations of Statistical Natural Language Processing. Cambridge, MA: MIT Press, 1999.

**Recommended Reading**

1. Foundations of Computational Linguistics: Human-computer Communication in Natural Language, Roland R. Hausser, Springer, 2014.

2. Steven Bird, Ewan Klein and Edward Loper Natural Language Processing with Python, O'Reilly Media; 1 edition, 2009.

**Web resources:**

1. https://web.stanford.edu/~jurafsky/slp3/ed3book.pdf
2. https://nptel.ac.in/courses/106101007/
3. NLTK – Natural Language Tool Kit- http://www.nltk.org

## MDS373B-WEB ANALYTICS

**Total Teaching Hours For Semester:90** **No of Lecture Hours/Week:6**
**Max Marks:150** **Credits:5**

Course Description and Course Objectives

The objective of this course is to provide an overview and the importance of Web analytics and helps to understand role of Web analytic. This course also explores the effective of Web analytic strategies and implementation

Course Outcomes

CO1: Understand the concept and importance of Web analytics in an organization ar the role of Web analytic in collecting, analyzing and reporting website traffic.
CO2: Identify key tools and diagnostics associated with Web analytics.
CO3: Explore effective Web analytics strategies and implementation and Understan the importance of web analytic as a tool for e-Commerce, business research, and market research.

**Unit-1** **Teaching Hours:18**

INTRODUCTION TO WEB ANALYTICS
Introduction to Web Analytics: Web Analytics Approach – A Model of Analysis – Context matters – Data Contradiction – Working of Web Analytics: Log file analysis – Page tagging – Metrics and Dimensions – Interacting with data in Google Analytics
Lab Exercise
1. Working concept of web analytics
2. Evaluation with Intermediate metrics, custom metrics, calculated metrics.

**Unit-2** **Teaching Hours:18**

LEARNING ABOUT USERS THROUGH WEB ANALYTICS
Goals: Introduction – Goals and Conversions – Conversion Rate – Goal reports in Goog Analytics – Performance Indicators – Analyzing Web Users: Learning about users – Traffic Analysis – Analyzing user content – Click-Path analysis – Segmentation
Lab Exercise
1. Collection of web data and other internet data with the help of web analytics
2. Delivering reports based on collected data
3. Implement the concept of web analytics ecosystem

**Unit-3** **Teaching Hours:18**

GOOGLE ANALYTICS
Different analytical tools - Key features and capabilities of Google analytics- How Google analytics works - Implementing Google analytics - Getting up and running with Google analytics -Navigating Google analytics – Using Google analytics reports -Googl metrics - Using visitor data to drive website improvement- Focusing on key performanc indicators- Integrating Google analytics with third-Party applications

Lab Exercise
1. Creation of segmentation in web analytics
2. Visualization, acquisition and conversions of web analytics data

**Unit-4** **Teaching Hours:18**

OVERVIEW OF QUALITATIVE ANALYSIS
Lab Usability Testing- Heuristic Evaluations- Site Visits- Surveys (Questionnaires) - Testing and Experimentation: A/B Testing and Multivariate Testing-Competitive Intelligence - Analysis Search Analytics: Performing Internal Site Search Analytics, Search Engine Optimization (SEO) and Pay per Click (PPC)-Website Optimization against KPIs- Content optimization- Funnel/Goal optimization - Text Analytics: Natural Language Processing (NLP)- Supervised Machine Learning (ML) Algorithms-API and Web data scarping using R and Python
Lab Exercise
1. Performing site search analytics
2. Analyse the web analytic reports and visualizations
3. Performing visual web analytics

**Unit-5** **Teaching Hours:18**

VISUAL ANALYTICS
VISUAL ANALYTICS: Drill down and hierarchies-Sorting-Grouping- Additional Way to Group- Creating Sets- Analysis with Cubes and MDX- Filtering for Top and Top N- Using the Filter Shelf- The Formatting Pane- Trend Lines- Forecasting- Formatting- Parameters - SOCIAL NETWORK ANALYSIS: Types of social network-Graph Visualization-Network Relationships-Network structures: equivalence-Network Evolution-Diffusion in networks- Descriptive Modeling-Predictive Modeling-Customer Profiling-Network targeting
Lab Exercise
1. Assignments and final discussions
2. Web Analytics case studies

Essential References

1. Beasley M, (2013), Practical web analytics for user experience: How analytics can help you understand your users. Newnes, 1st edition, Morgan Kaufmann.
2. Sponder M, (2013), Social media analytics: Effective tools for building, interpreting, and using metrics, 1st edition, McGraw Hill Professional.
3. Clifton B, (2012), Advanced Web Metrics with Google Analytics, 3rd edition, John Wiley & Sons..

Recommended References

1. Peterson E. T, (2004), Web Analytics Demystified: AMarketer's Guide to Understanding How Your Web Site Affects Your Business. Ingram.
2. Sostre P, LeClaire J, (2007), Web Analytics for dummies, John Wiley & Sons.
3. Burby J, Atchison S, (2007), Actionable web analytics: using data to make smart

business decisions, John Wiley & Sons.
4. Dykes B, (2011), Web analytics action hero: Using analysis to gain insight and optimize your business, Adobe Press.

Evaluation Pattern

CIA 50%
ESE 50%

## MDS373C-BIO INFORMATICS

**Total Teaching Hours For Semester:90** **No of Lecture Hours/Week:6**
**Max Marks:150** **Credits:5**

Course Description and Course Objectives

To enable the students to learn the information search and retrieval, Genome analysi and Gene mapping, alignment of multiple sequences, and PERL for Bioinformatics.

Course Outcomes

CO1: To understand the molecular Biology and Bioinformatics applications.
CO2: Apply the modeling and simulation technologies in Biology and medicine.
CO3: Evaluate the algorithms to find the similarity between protein and DNA sequences.

**Unit-1** **Teaching Hours:18**

BIOINFORMATICS
Introduction, Historical Overview and Definition, Applications, Major databases in Bioinformatics, Data management and Analysis, Central Dogma of Molecular Biology.
INFORMATION SEARCH AND RETRIEVAL
Introduction, Tools for web search, Data retrieval tools, Data mining of Biological databases.
Lab Exercise
1. Test and verify the basic Linux commands and Filters.
2. Create the file(s) and verify the file handling commands.

**Unit-2** **Teaching Hours:18**

GENOME ANALYSIS AND GENE MAPPING
GENOME ANALYSIS AND GENE MAPPING Introduction, Genome analysis, Genon mapping, Sequence assembly problem, Genetic mapping and linkage analysis, Physical maps, Cloning the entire Genome, Genome sequencing, Applications of Genetic maps, Identification of Genes in Contigs, Human Genome Project. ALIGNMENT OF PAIRS OF SEQUENCES Introduction, Biological motivation of alignment, Methods of sequen alignments, Using score matrices, Measuring sequence detection
Lab Exercise
1. Create directories and verify the directory commands.
2. Perform basic mathematical operations using PERL.
3. Write a PERL script to demonstrate the Array operations and Regular expressions.

**Unit-3** **Teaching Hours:18**

ALIGNMENT OF MULTIPLE SEQUENCES
ALIGNMENT OF MULTIPLE SEQUENCES Methods of multiple sequence alignment Evaluating multiple alignments, Applications of multiple alignments, Phylogenetic analysis, Methods of phylogenetic analysis, Tree evaluation, Problems in Phylogenetic analysis.

TOOLS FOR SIMILARITY SEARCH AND SEQUENCE ALIGNMENT Introduction, Working with FASTA, Working with BLAST, Filtering and Gapped BLAST, FASTA and BLAST algorithm comparison.

Lab Exercise

1. Write a PERL script to concatenate DNA sequences.
2. Write a PERL script to transcribe DNA sequence into RNA sequence
3. Write a PERL script to calculate the reverse complement of a strand of DNA.

**Unit-4** **Teaching Hours:18**

PERL FOR BIOINFORMATICS

Sequences and Strings: Representing sequence data, Program to store a DNA sequence, Concatenating DNA fragments, Transcription DNA to RNA, Proteins, Files and Arrays, Reading Proteins in Files, Arrays, Scalar and List Context.

Motifs and Loops: Flow control, Code layout, Finding motifs, Counting Nucleotides, Exploding strings and arrays, Operating on strings. Subroutine and Bugs: Subroutines, Scoping and Subroutines, Command line arguments and Arrays, Passing data to Subroutines, Modules and Libraries of Subroutines.

Lab Exercise

1. Write a PERL script to read protein sequence data from a file.
2. Write a PERL script to search for a motif in a DNA sequence.

**Unit-5** **Teaching Hours:18**

THE GENETIC CODE

Hashes, Data structure and algorithms for Biology, Translating DNA into Proteins, Reading DNA from the files in FASTA format, Reading Frames. GenBank: GenBank files, GenBank Libraries, Separating Sequence and Annotation, Parsing Annotations, Indexing GenBank with DBM. Protein Data Bank: Files and Folders, PDB Files, Parsin PDB Files.

1. Write a PERL script to append ACGT to DNA using a subroutine.
2. Case Study: a. To retrieve the sequence of the Human keratin protein from UniProt database and to interpret the results. b. To retrieve the sequence of the Human keratin protein from GenBank database and to interpret the results.

Essential References

[1] Bioinformatics: Methods and Applications, S. C. Rastogi, Namita Mendirata and Parag Rastogi, 4th Edition, PHI Learning, 2013.

[2] Beginning Perl for Bioinformatics, Tisdall James, 1st edition, Shroff Publishers (O'Reilly), 2009.

Recommended References

[1] Introduction to Bioinformatics, Arthur M Lesk, 2nd Edition, Oxford University Press,4th edition, 2014.

[2] Bioinformatics Technologies, Yi-Ping Phoebe Chen (Ed), 1st edition, Springer, 2005.

[3] Bioinformatics Computing, Bryan Bergeron, 2nd Edition, Prentice Hall, 1st edition, 2003.
Web resources:
[1] http://cac.annauniv.edu/PhpProject1/aidetails/afug_2013_fu/24.%20BIO%20MED.pdf
[2] https://www.amrita.edu/school/biotechnology/academics/pg/introduction-bioinformaticsbif410
[3] https://canvas.harvard.edu/courses/8084/assignments/syllabus
[4] https://www.coursera.org/specializations/bioinformatics
[5] http://www.dtc.ox.ac.uk/modules/introduction-bioinformatics-bioscientists.html

Evaluation Pattern

CIA 50%
ESE 50%

**MDS373D-EVOLUTIONARY ALGORITHMS**

**Total Teaching Hours For Semester:90** **No of Lecture Hours/Week:6**
**Max Marks:150** **Credits:5**

Course Description and Course Objectives

Able to understand the core concepts of evolutionary computing techniques and popular evolutionary algorithms that are used in solving optimization problems.Students will be able to implement custom solutions for real-time problem applicable with evolutionary computing.

Course Outcomes

CO1:Basic understanding of evolutionary computing concepts and techniques
CO2:Classifyrelevantreal-time problems for the applications of evolutionary algorithms
CO3:Design solutions using evolutionary algorithms

**Unit-1** **Teaching Hours:18**

INTRODUCTION TO EVOLUTIONARY COMPTUTING
Terminologies – Notations – Problems to be solved – Optimization – Modeling – Simulation
– Search problems – Optimization constraints
Lab Program
1. Implementation of single and multi-objectivefunctions
2. Implementation of binaryGA

**Unit-2** **Teaching Hours:18**

EVOLUTIONARY PROGRAMMING
Continuous evolutionary programming – Finite state machine optimization – Discrete evolutionary programming – The Prisoner's dilemma
EVOLUTION STRATEGY
One plus one evolution strategy – The 1/5 Rule – (μ+1) evolution strategy – Self adapti evolution strategy
Lab Program
1. Implementation of continuousGA
2. Implementation of evolutionaryprogramming

**Unit-3** **Teaching Hours:18**

GENETIC PROGRAMMING
Fundamentals of genetic programming – Genetic programming for minimal time contro
EVOLUTIONARY ALGORITHM VARIATION
Initialization – Convergence – Population diversity – Selection option – Recombination Mutation
Lab Program
1. Implementation of geneticprogramming
2. Implementation of Ant ColonyOptimization

**Unit-4** **Teaching Hours:18**

ANT COLONY OPTIMIZATION
Pheromone models – Ant system – Continuous Optimization – Other Ant System
PARTICLE SWARM OPTIMIZATION
Velocity limiting – Inertia weighting – Global Velocity updates – Fully informed Partic
Swarm
Lab Program
1. Implementation of Particle SwarmOptimization
2. Implementation of Multi-ObjectOptimization

**Unit-5** **Teaching Hours:18**

MULT-OBJECTIVE OPTIMIATION
Pareto Optimality – Hyper volume – Relative coverage – Non-pareto based EAs – Paret
based EAs – Multi-objective Biogeography based optimization
Lab Program
1. Simulation of EA in Planning problems (routing, scheduling, packing) and Design problems (Circuit, structure,art)
2. Simulation of EA in classification/predictionmodelling

Essential References


[1] D. Simon, Evolutionary optimization algorithms: biologically inspired and population-based approaches to computer intelligence. New Jersey: John Wiley, 201


Recommended References


1. Eiben and J. Smith, Introduction to evolutionary computing. 2nd ed. Berlin: Springer, 2015.
2. D.Goldberg,Geneticalgorithmsinsearch,optimization,andmachinelearning.Bosto Addison-Wesley,2012.
3. K. Deb, Multi-objective optimization using evolutionary algorithms. Chicheste John Wiley & Sons,2009.
4. R. Poli, W. Langdon, N. McPhee and J. Koza, A field guide to genetic programming. [S.l.]: Lulu Press,2008.
5. T.Bäck,Evolutionaryalgorithmsintheoryandpractice.NewYork:OxfordUniv.Pres 1996.


Web Resources:


1 E.A.EandS.J.E,"IntroductiontoEvolutionaryComputing|Theon-line accompaniment to the book Introduction toEvolutionary Computing",Evolutionarycomputation.org,2015.[Online].Available: http://www.evolutionarycomputation.org/.
2 F.Lobo,"EvolutionaryComputation2018/2019",Fernandolobo.info,2018.[Online Available:http://www.fernandolobo.info/ec1819.
3 "EClabTools",Cs.gmu.edu,2008.[Online].Available: https://cs.gmu.edu/~eclab/tools.html.

4 "Kanpur Genetic Algorithms Laboratory", Iitk.ac.in, 2008. [Online]. Available: https://www.iitk.ac.in/kangal/codes.shtml.
5 "Course webpage Evolutionary Algorithms", Liacs.leidenuniv.nl, 2017. [Online]. Available:http://liacs.leidenuniv.nl/~csnaco/EA/misc/ga_demo.htm.

Evaluation Pattern

CIA: 50%
ESE : 50%

**MDS373E-OPTIMIZATION TECHNIQUES**

**Total Teaching Hours For Semester:90** **No of Lecture Hours/Week:6**

**Max Marks:150** **Credits:5**

Course Description and Course Objectives

This course will help the students to acquire and demonstrate the implementation of the necessary algorithms for solving advanced level Optimization techniques.

Course Outcomes

CO1: Apply the notions of linear programming in solving transportation problems
CO2: Understand the theory of games for solving simple games
CO3: Use linear programming in the formulation of the shortest route problem.
CO4: Apply algorithmic approach in solving various types of network problems
CO5: Create applications using dynamic programming.

**Unit-1** **Teaching Hours:18**

INTRODUCTION
Operations Research Methods - Solving the OR model - Queuing and Simulation model – Art of modelling – phases of OR study.
MODELLING WITH LINEAR PROGRAMMING
Two variable LP model – Graphical LP solution – Applications. Simplex method and sensitivity analysis – Duality and post-optimal Analysis- Formulation of the dual problem.
Lab Exercise

1. Simplex Method
2. Dual Simplex Method

**Unit-2** **Teaching Hours:18**

TRANSPORTATION MODEL
Determination of the Starting Solution – Iterative computations of the transportation algorithm. Assignment Model: The Hungarian Method – Simplex explanation of the Hungarian Method – The trans-shipment Model.
Lab Exercise

1. Balanced Transportation Problem
2. Unbalanced Transportation Problem
3. Assignment Problems

**Unit-3** **Teaching Hours:18**

NETWORK MODELS
Minimal Spanning tree Algorithm – Linear Programming formulation of the shortest-route problem. Maximal Flow Model: Enumeration of cuts – Maximal Flow Diagram – Linear Programming Formulation of Maximal Flow Model.
CPM and PERT
Network Representation – Critical Path Computations – Construction of the time Schedule – Linear Programming formulation of CPM – PERT networks.

Lab Exercise:
1. Shortest path computations in a network
2.Maximum flow problem

**Unit-4** **Teaching Hours:18**

GAME THEORY
Strategic Games and examples - Nash equilibrium and examples - Optimal Solution of two person zero sum games - Solution of Mixed strategy games - Mixed strategy Nash equilibrium - Dominated action with example.
GOAL PROGRAMMING
Formulation – Tax Planning Problem – Goal Programming algorithms – Weights methc – Preemptive method.
Lab Exercise:
1. Critical path Computations
2. Game Programming

**Unit-5** **Teaching Hours:18**

MARKOV CHAINS
Definition – Absolute and n-step Transition Probability – Classification of states.
DYNAMIC PROGRAMMING
Recursive nature of computation in Dynamic Programming – Forward and Backward Recursion – Knapsack / Fly Away / Cargo-Loading Model – Equipment Replacement Model.
Lab Exercise:
1. Goal Programming
2. Dynamic Programming

Essential References

1. Hamdy A Taha, Operations Research, 9th Edition, Pearson Education, 2012.
2. Garrido José M. Introduction to Computational Models with Python. CRC Pre: 2016.

Recommended References

1. Rathindra P Sen, Operations Research – Algorithms and Applications, PHI Learning Pvt. Limited, 2011
2. R. Ravindran, D. T. Philips and J. J. Solberg, Operations Research: Principles and Practice, 2nd ed., John Wiley & Sons, 2007.
3. F. S. Hillier and G. J. Lieberman, Introduction to operations research, 8th ed., McGraw-Hill Higher Education, 2004.
4. K.C. Rao and S. L. Mishra, Operations research, Alpha Science International, 2005.
5. Hart, William E. Pyomo: Optimization Modeling in Python. Springer, 2012.
6. Martin J. Osborne, An introduction to Game theory, Oxford University Press, 2008

Additional Information

NA

Evaluation Pattern

CIA: 50%
ESE: 50%

## MDS381-SPECIALIZATION PROJECT

**Total Teaching Hours For Semester:60** **No of Lecture Hours/Week:4**
**Max Marks:100** **Credits:2**

Course Description and Course Objectives

The course is designed to provide a real-world project development and deployment environment for the students.

Course Outcomes

CO1: Identify the problem and relevant analytics for the selected domain.
CO2: Apply appropriate design/development strategy and tools.

**Unit-1** **Teaching Hours:60**

Specialization Project
Project will be based on the specialization domains which students are opted for during this semester.

Essential References

-

Recommended References

-

Evaluation Pattern

CIA: 50%
ESE: 50%

## MDS382-SEMINAR

**Total Teaching Hours For Semester:30** **No of Lecture Hours/Week:2**
**Max Marks:50** **Credits:1**

Course Description and Course Objectives

The course is designed to provide to enhance the soft skills and technical undetstanding of the students.

Course Outcomes

CO1: Understand new and latest trends in data science
CO2: Demonstrate the professional presentation abilities
CO3: Apply the acquired knowledge in their Research

**Unit-1** **Teaching Hours:30**

Students will be giving presentations on any advanced concepts and technologies in data science and submit the report
-

Essential References

Research Articles / Books / Web resources related to data science domain

Recommended References

Recommended References

Evaluation Pattern

CIA 100%

## MDS481-INDUSTRY PROJECT

**Total Teaching Hours For Semester:30** **No of Lecture Hours/Week:2**
**Max Marks:300** **Credits:12**

Course Description and Course Objectives

This course helps the student to develop students to become globally competent and to inculcate Entrepreneurial skills among students.

Course Outcomes

CO1: Develop Real time Projects.
CO2: Practices different data science principles and strategies in the project.

**Unit-1** **Teaching Hours:30**

Project Work
It is a full time project to be taken up either in the industry or in an R&D organization

Essential References

-

Recommended References

-

Evaluation Pattern

CIA: 50%
ESE: 50%