Tribhuvan University

Institute of Science and Technology

SCHOOL OF MATHEMATICAL SCIENCES

Course Structure

Master's in Data Sciences (MDS)

1. Introduction

The Modern Computerized World demands the human resource having all three – analytical ability, data processing capability and fast computing efficiency, i.e., the combined knowledge of **Mathematics, Statistics, and Computer Science and Information Technology**. Tribhuvan University has taken up this as a challenge and has decided to run Bachelor's and Master's Degree Program in Mathematical Sciences that will help produce at least a critical mass of experts with sound knowledge of fundamentals of Mathematics, Statistical and Analytical capability and fluent computational skills. To run these programs, Tribhuvan University established School of Mathematical Sciences under Institute of Science and Technology in 2016 at Kirtipur as its autonomous body.

Computational simulations are everywhere and the amount of data available for many enterprises is increasing exponentially. The internet makes these large quantities of data readily available for many enterprises. Many areas of science, engineering, and industry are now concerned with building and evaluating mathematical models, exploring them computationally, and analyzing enormous amounts of observed and computed data. These activities are all inherently mathematical in nature. Thus, Master's Program in Data Science is an ideal program to start at SMS TU.

2. Objectives

This interdisciplinary program is the first of its kind in the country. After graduation, the students will be able to

- Collect, clean, store and query data from a variety of private and public data sources.
- Assess, evaluate and respond to decision-making needs and requirements.
- Apply appropriate analytic techniques to provide estimates that support decision-making and action.
- Communicate actionable information and findings in easy-to-understand written, oral and visual formats.

3. Duration and Nature of Course

Master in Data Science is full time, of 4 Semesters in 2 years in duration. This program basically comprises of some compulsory foundational courses consisting of fundamentals of Mathematics, Statistics, and Computer Science and Information Technology plus some elective courses from a list of courses which may vary from year to year as a multi-exit model decided by the subject committee.

Total Credit: 60

Nature of course: Theory, Practical, Project, Seminar, Intern, Thesis.

4. Evaluation System

a) 40% internal evaluation and 60% external exam. Internal exams are based on: Attendance/Assignment work / Oral test / Class test / Presentation / Class seminar / Project work/ Term exam etc. End semester exam by School in permission of exam board of TU.

b) Evaluation of project or thesis: research / project monitoring by supervisor; Pre viva by the school after submission; evaluation of thesis by the Research Committee of the School with consent of the supervisor and the external.

c) In each of the semester Exam and Internal Assessment, the student must secure at least 50% in order to complete the course.

5. Eligibility:

Students applying to the program are expected to have a Bachelor's Degree with a strong quantitative and computational background including coursework in calculus, linear algebra and introductory statistics. So students with B Sc CSIT, B Math Sc, B Sc. (Math), B Sc (Stat), B Sc/BA with Math / Stat in the first 2 years, BE, BIT, BCA (with two Math and one Stat).

6. Course Structure

In the **First and Second Semester,** students must take four compulsory courses in each semester and one course from elective courses (the necessary and relevant to them). In the **Third Semester**, student must take three compulsory courses and two courses from elective courses**.** In the **Fourth Semester,** students must take two compulsory courses and two courses from elective course.

FIRST SEMESTER

Compulsory Courses

Course Code	Course Titles	Credits	Nature
MDS 501	Fundamentals of Data Science	3	Th.
MDS 502	Data Structure and Algorithms	3	Th.+ Pr.
MDS 503	Statistical Computing with R	3	Th.+ Pr.
MDS 504	Mathematics for Data Science	3	Th.

Elective Courses (Any One Available on School)

Course Code	Course Titles	Credits	Nature
MDS 505	Data Base Management Systems	3	Th.+ Pr.
MDS 506	Programming Concepts and Techniques	3	Th.+ Pr.
MDS 507	Linear and Integer Programming	3	Th.+ Pr.

SECOND SEMESTER

Compulsory Courses

Course Code	Course Titles	Credits	Nature
MDS 551	Programming with Python	3	Th.+ Pr.
MDS 552	Applied Machine Learning	3	Th.+ Pr.
MDS 553	Statistical Methods for Data Science	3	Th.+ Pr.

MDS 554	Multivariable Calculus for Data Science	3	Th.

Elective Courses (Any One Available on School)

Course Code	Course Titles	Credits	Th.+ Pr.
MDS 555	Natural Language Processing	3	Th.+ Pr.
MDS 556	Artificial Intelligence	3	Th.+ Pr.
MDS 557	Learning Structure and Time Series	3	Th.+ Pr.

THIRD SEMESTER

Compulsory Courses

Course Code	Course Titles	Credits	Nature
MDS 601	Research Methodology	3	Th.
MDS 602	Advanced Data Mining	3	Th.+ Pr.
MDS 603	Techniques for Big Data	3	Th.+ Pr.

Elective Courses (Any Two Available on School)

Course Code	Course Titles	Credits	Nature
MDS 604	Cloud Computing	3	Th.+ Pr.
MDS 605	Regression Analysis	3	Th.+ Pr.
MDS 606	Decision Analysis	3	Th.+ Pr.
MDS 607	Monte Carlo Methods	3	Th.

FOURTH SEMESTER

Compulsory Courses

Course Code	Course Titles	Credits	Nature
MDS 651	Data Visualization	3	Th.
MDS 652	Capstone Project / Thesis	6	Project + Report

Elective Courses (Any Two Available on School)

Course Code	Course Titles	Credits	Nature
MDS 653	Social Network Analysis	3	Th.+ Pr.
MDS 654	Actuarial Data Analysis	3	Th.+ Pr.
MDS 655	Deep Learning	3	Th.+ Pr.
MDS 656	Business Analytics	3	Th.+ Pr.
MDS 657	Bioinformatics	3	Th.+ Pr.
MDS 658	Economic Analysis	3	Th.+ Pr.

●●●●●

Tribhuvan University

Institute of Science and Technology

SCHOOL OF MATHEMATICAL SCIENCES

Syllabus

Master's in Data Sciences (MDS)- FIRST SEMESTER

Compulsory Courses

Course Code	Course Titles	Credits	Nature
MDS 501	Fundamentals of Data Science	3	Th.
MDS 502	Data Structure and Algorithms	3	Th.+ Pr.
MDS 503	Statistical Computing with R	3	Th.+ Pr.
MDS 504	Mathematics for Data Science	3	Th.

Elective Courses (Any One Available on School)

Course Code	Course Titles	Credits	Nature
MDS 505	Data Base Management Systems	3	Th.+ Pr.
MDS 506	Programming Concepts and Techniques	3	Th.+ Pr.
MDS 507	Linear and Integer Programming	3	Th.+ Pr.

Code No.: MDS 501

Course Title: **Fundamentals of Data Science** Full Marks: 75

Nature: Theory(Compulsory) Credit: 3

***Course Description*:**

This is an introductory course to teach the basics of data science, it's applications and commonly used tools and techniques. The course is designed to introduce key ideas and methodologies used in the domain of data science. The goal of this course is to help understand the fundamental building blocks of data science.

Learning Objectives:

Upon the conclusion of the course, students should be able to:

- Describe Data Science, skill sets needed to be a data scientist and be familiar with common tools used for data science. Understand the importance of data quality and familiarize with common data munging techniques.
- Understand and apply commonly used data analysis and machine learning techniques in data science
- Identify the challenges in handling big data, and gain a general understanding of tools to handle big data
- Reason around ethical and privacy issues in data science and understand the common biases affecting data science.

***Course Contents*:**

Unit 1: Introduction to Data Science [10 Hrs.]

Introduction to data science, Applications of data science; Limitations of data science

Commonly used tools in data science, their strengths and common use-cases: R/RStudio, Python/Pandas/Jupyter Notebooks, Excel/Tableau/PowerBI;

Data Science life-cycle/Common methodologies for data science: CRISP-DM, OSEMN Framework, TDSP lifecycle;

Review of statistics and probability: Probability distributions, compound events and independence. Statistics: Centrality measures, variability measures, interpreting variance. Correlation analysis: Correlation coefficients, autocorrelation

Unit 2: Data Munging [8 Hrs.]

Data quality, common issues with real world data: Duplicates, Missing Data, Non-standard data, Unit mismatch;

Ways to clean up and standardize data; Data enrichment: Need for data enrichment; Common ways to enrich data: correction, extrapolation, augmentation;

Data Validation: Common methods of data validation: type check, range & constraint check, consistency check;

Data format conversion: Commonly used formats: JSON, XML, Tabular, Relational - their strengths and weaknesses,

Motivation behind format conversion.General methods of conversion between data formats. Tabular data: Row based vs column based (Parquet, ORC, CSV). Wide vsnarrow(long) table format. Converting between wide vs narrow formats

Unit 3: Data Analysis Technique [10 Hrs.]

Feature generation and feature selection algorithms: filters, wrappers, decision trees, random forests;

Common techniques: Linear regression, logistic regression, k-NN, k-means ;

Predictive data analysis: Introduction to predictive data analysis and its common applications.;

Regression based models: linear regression, logistic regression.;

Time series data analytics

Unit 4: Machine Learning [8 Hrs.]

Introduction to machine learning, type of machine learning methods.;

Supervised vsUnsupervised learning;

Naive Bayes, Decision Trees, SVMs;

Introduction to deep learning, backpropagation.

Unit 5: Introduction to Big Data [8 Hrs.]

Introduction to big data and the challenges of handling big data;

Commonly used tools for big data: The map-reduce programming paradigm.Hadoop, HDFS, (py)Spark, Hive.

Data warehousing and data lake architecture.

Real-time analytics with Apache Kafka

Unit 6: Ethical Issues in Data Science [4Hrs.]

Issues with fairness and bias in data science:

Common biases: In group favoritism and out-group negativity, Fundamental attribution error, Negativity bias, Stereotyping, Bandwagon effect, Bias blind spot.

Addressing biases: Group unaware selection, Adjusted group thresholds, Demographic parity, Equal opportunity, Precision parity;

Common issues with privacy and data ethics.

References:

1. O'Neil, Cathy and Schutt, Rachel(2013), *Doing Data Science***,** *Straight Talk From TheFrontline*, O'Reilly Media
2. Provost, Foster and Fawcett, Tom (2013). *Data Science for Business***:** *What You Need to Know about Data Mining and Data-analytic Thinking*, O'Reilly Media.

***

Code No.: MDS 502

Course Title:**Data Structures and Algorithms** Full Marks: 75

Nature: Theory +Practical (Compulsory) Credit: 3

***Course Description*:**

This course includes the basic foundations in of data structures and algorithms and various data structures like stack, queue, list, tree and graph. Additionally, the course includes idea of sorting and searching.

***Learning Objectives*:**

After successful completion of this course, the student will be able to

- Introduce basic concepts data structures and algorithms, abstract data types, asymptotic notations
- Design and use of data structures such as stack, queue, linked list, tree and graph.

***Course Contents*:**

Unit 1: Introduction to Data Structures & Algorithms [3 Hrs.]

Data types, Data structure and Abstract date type
Dynamic memory allocation
Introduction to Algorithms
Asymptotic notations and common functions

Unit 2: Stack [6 Hrs.]

Basic Concept of Stack, Stack as an ADT, Stack Operations, Stack Applications
Conversion from Infix to Postfix Expressions and Evaluation of Postfix Expressions using Stack

Unit 3: Queue [4 Hrs.]

Basic Concept of Queue, Queue as an ADT, Primitive Operations in Queue
Linear Queue, Circular Queue, Priority Queue, Queue Applications

Unit 4: Recursion [4 Hrs.]

Principle of Recursion, Comparison between Recursion and Iteration, Tail Recursion

Factorial, Fibonacci Sequence, GCD, Tower of Hanoi(TOH)

Applications and Efficiency of Recursion

Unit 5: Lists [8 Hrs.]

Basic Concept, List and ADT, Array Implementation of Lists, Linked List
Types of Linked List: Singly Linked List, Doubly Linked List, Circular Linked List.
Basic operations in Linked List: Node Creation , Insertion and Deletion from Linked List
Stack and Queue as Linked List

Unit 6: Sorting [8Hrs.]

Introduction and Types of sorting: Internal and External sort
Comparison Sorting Algorithms: Bubble, Selection and Insertion Sort, Shell Sort
Divide and Conquer Sorting: Merge, Quick and Heap Sort
Efficiency of Sorting Algorithms.

Unit 7: Searching and Hashing [7Hrs.]

Introduction to Searching, Search Algorithms: Sequential Search, Binary Search
Efficiency of Search Algorithms
Hashing : Hash Function and Hash Tables, Collision Resolution Techniques

Unit 8: Trees and Graphs [8Hrs.]

Concept and Definitions, Basic Operations in Binary Tree, Tree Height, Level and Depth
Binary Search Tree, Insertion, Deletion, Traversals, Search in BST
AVL tree and Balancing algorithm, Applications of Trees
Definition and Representation of Graphs, Graph Traversal, Minimum Spanning Trees: Kruskal and Prims Algorithm
Shortest Path Algorithms: Dijksrtra Algorithm.

Laboratory Works:

The laboratory work consists of implementing different algorithms and data structures studied in the course using C programming.

References:

1. Langsam, Y. ,Augenstein, M.J. &Tanenbaum, A.M. (2015).*Data Structures using C&C++*. , 2nd Edition, Pearson, India.
2. LeenAmmeral.*Programmes and Data Structures in C*, Wiley Professional Computing.
3. Rowe, G.W. (2016).*Introduction to Data Structure and Algroithms with C and C++* , prentice Hall India.
4. Kruse, R.L., Leung, B.P. &Tondo, C.L.(2013).*Data Structure and Program Design in C*, 2nd Edition, Pearson Education , New Delhi, India.

***

Code No.: **MDS 503**
Course Title:**Statistical Computing with R** Full Marks: 75
Nature: Theory +Practical (Compulsory) Credit: 3

Course Description:

This is an outcome based course to introduce basic programming in R software followed by use of R software for Statistical Computing. It focuses on the use of R software for data manipulation, data summary/data visualization, supervised and unsupervised learning sand communicate the findings.

Learning Objectives:

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

- Understand, use and apply R software for basic programming (program)
- Understand, use and apply R software for data manipulation (wrangle)
- Understand, use and apply R software for data summary and visualization (explore)
- Understand, use and apply R software for supervised learning (model)
- Understand, use and apply R software for unsupervised learning (model)
- Understand, use and apply R software to communicate findings (communicate).

Course Contents:

Unit 1: R Software for Basic Programming [8Hrs.]

R software, Statistics, Big Data and Data Science. Downloading and installing R software in Windows, Linux and Unix systems. Variables, Data types, Vectors, Lists and Matrix in R. Factors, Data Frames and Dealing with missing values in R. Logical statements, Loops, Functions and Pipes in R. Coding and naming conventions in R. Reproducible Analysis: Markdown Language, YAML Language; R Markdown/knitr document in R IDE (RStudio). Profiling and optimizing codes/scripts in R.

Unit 2: R Software for Data Manipulation [6 Hrs.]

Using R packages in R. Reading and Reviewing data in R. Manipulating and Tyding data in R. Data Wrangling in R. Data Transformation in R. Data/Text Mining in R. Big Data in R: Subsampling, Hex and 2D Density Plots.

Unit 3: R Software for Data Summary and Visualization [10Hrs.]

Basic graphics/plots in R: Bar chart and histogram, Line chart and Pie chart, Scatterplot and Boxplot, Scatterplot matrix, Social Network Analysis. The Grammar of Graphics: Data, Aesthetic mapping, Geometric objects, Statistical transformation, Scales, Coordinate system, Position adjustment and Faceting using ggplot2 package in R/RStudio. Computing measures of central tendency, dispersion, moments and relative positions in R using packages and functions/scripts.

Unit 4: R Software for Supervised Learning [10 Hrs.]

Probability Distribution Functions: Use of apply(), lapply() and sapply() fuctions in R for Breakdown Analysis. Random Sampling, Covariance and Correlation; Hypothesis Testing using common parametric and non-parametric statistical tests in R. Machine Learning and Supervised Learning. Specifying supervised models: Linear regression, Logistic Regression, Model matrices and formula. Validating models: Evaluating regression models, evaluating classification models, cross-validation, training, testing and holdouts. Supervised learning packages and its use: Decision Trees, Random Forests, Neural Networks, Support Vector Machines and Naïve Bayes.

Unit 5: R Software for Unsupervised Learning [8 Hrs.]

Dimensionality Reduction: Principle component analysis, Principle Axis Factoring, Multidimensional scaling; Clustering: k-Means clustering, Hierarchical clustering; Association rules and Monte-Carlo simulations.

Unit 6: R Software for Communication [6Hrs.]

Markdown Language, R Markdown/knitr document to produce publishable/industry level documents in HTML, PDF and Word formats. Use R Markdown to create reports, websites and dashboards. Use R Markdown to create Shiny apps for effective communication.

Practical Works:
The practical works include of class/computer lab using R/RStudiowith individual project work.

References:

1. Mailund Thomas (2017).*Beginning Data Sciences in R: Data Analysis, Visualization, and Modelling for the Data Scientists*.Apress: Aarhus, Denmark.
2. Goh Eric &Hui Ming (2019).*Learn R for Applied Statistics*. Apress: Singapore.
3. Wichham Hadley &GloremundGarrette (2017).*R for Data Science*. O'Reilly Media Inc: Sebastopol, Canada.

***

Code No.: **MDS 504**
Course Title:**Mathematics for Data Science** Full Mark: 75
Nature: **Theory** (Compulsory) Credit: 3

***Course Description*:**

The course will cover basic topics in linear algebra to understand high-dimensional vector spaces, matrices and graphs as popular mathematical structures with which to model data (e.g., as models for term-document corpora, high-dimensional regression problems, ranking/classification of web data, adjacency properties of social network data, etc.); and geometric approaches to eigendecompositions, least-squares, principal components analysis, etc. The course requires to solve problems using programming R.

***Learning Objectives*:**

After successful completion of this course the student will be able to

- Understand basic linear algebra techniques which are useful in data science.
- Explain \why different methods do and don't work.
- Understand tools that are used to diagnose problems, to develop new methods, etc.
- Understand how some discrete probability and optimization are used with matrices and graphs, two very common ways to model data.
- Understand the connections between the discrete probability ideas and very related linear algebra ideas.
- Acquire a basic understanding and intuition of why various methods work, so that the student can use them in practical applications data science.
- Use programming R to solve problems of this course.

***Course Contents*:**

Unit 1: Introduction, Motivation, and Overview [9 Hrs.]

Linear algebra and machine learning,
Representing data as flat tables versus matrices and graphs;
Different ways probability/randomness/noise interacts with data;
Probability and matrices/graphs in data science versus other areas;
Quantification of the inference step.

Unit 2: Introduction to Matrices and Vectors [15Hrs.]

Vectors, Basic properties of R^n;
Norms and balls;
Vector addition and scalar multiplication.
Vector spaces and subspaces;
Matrices,
Operations on matrices, including matrix multiplication;
Functions, linear functions, and linear transformations; Matrices as transformations.
Dot products, angles, and perpendicularity;
Linear combinations, span, and linear independence; Bases, orthonormal bases, and projections.
Applications in the theory of probabilityand data science

Unit 3: Spectral Theorems **[14Hrs.]**

Eigenvectors and Eigenvalues:
Quadratic forms and matrices
Symmetric bi-linear functions;
Connections with conic sections;
Definiteness, indefiniteness, and quadratic forms as a sum/difference of squares;
EigenValue Decomposition (EVD)
Singular Value Decomposition (SVD)
Properties of the SVD
Orthogonal subspaces;
Uses of the spectral decomposition
Applications in data science

Unit 4: System of Linear Equations **[10 Hrs.]**

Solving system of linear equations:
Geometry of linear equations;
Gaussian elimination;
Row exchanges;
Networks and incidence matrices;
The four fundamental subspaces.
Basis transformations;
Orthogonal bases;
Gram-Schmidt Orthogonalization;
Numerical issues.
Applications in data science

References:

1. Nick Fieller (2015).*Basics of Matrix Algebra for Statistics with R*, Chapman and Hall/CRC.
2. Shayle R. Searle& André I. Khuri (2017).*Matrix Algebra Useful for Statistics*, John Wiley & Sons, Inc..
3. Michael W. Mahoney (2018).*Linear Algebra for Data*, University of California Berkeley.
4. Deisenroth, M. P. ,Faisal, A. A. and Ong, C. S. (2019). *Mathematics for Machine Learning*, Cambridge University Press.
5. Jason Brownlee (2018).*Basics of Linear Algebra for Machine Learning*, https://www.mobt3ath.com/uplode/book/book-33342.pdf .

***

Code No.: **MDS 505**
Course Title:**Data Base Management Systems** Full Marks: 75
Nature: Theory +Practical (Elective) Credit: 3

Course Description:

The course covers on the fundamentals of knowledgebase and relational database management systems, and the current developments in database theory and their practice.

Course Objectives:

After the completion of this course, the students should be able to

- Familiarize the students to the fundamentals of Database Management Systems.
- Understand the relational model, ER diagrams and SQL.
- Understand the fundamentals of Transaction Processing and Query Processing.
- Familiarize the different types of database.
- Understand the Security Issues in Databases.

Course Content:

Unit 1: Fundamental Concept of DBMS [6 Hrs.]

Database and database management system, Data Abstraction and Data Independence,Schema and Instances, Concepts of DDL, DML and DCL, Purpose of Database System, Database System Terminologies, Database characteristics, Data models , Types of data models , Components of DBMS, Relational Algebra. Relational DBMS – Codd's Rule – Entity- Relationship model,

Unit 2: Relational Languages and Relational Model [7Hrs.]

Introduction to SQL, Features of SQL, Queries and Sub-Queries,Set Operations, Relations (Joined, Derived), Queries under DDL and DML Commands, Embedded SQL,Views, Relational Algebra, Database Modification, QBE and domain relational calculus

Unit 3: Database Constraints and Normalization [6 Hrs.]

Integrity Constraints and Domain Constraints, Assertions and Triggering, Functional Dependencies, Different Normal Forms (1st, 2nd, 3rd, BCNF, DKNF)

Unit 4: SQL & Query Optimization [6 Hrs.]

SQL Standards ,Data types , Database Objects- DDL-DML-DCL-TCL, Embedded SQL, Static Vs Dynamic SQL, QUERY OPTIMIZATION: Query Processing and Optimization , Heuristics and Cost Estimates in Query Optimization.

Unit 5: Transaction Processing and Concurrency Control [6 Hrs.]

Properties of Transaction, Serializability, Concurrency Control, Locking Mechanisms, Two Phase Commit Protocol, Deadlock handling and Prevention

Unit 6: Trends in Database Technology [9Hrs.]

Overview of Physical Storage Media ,RAID , Tertiary storage , File Organization, Organization of Records in Files ,Indexing and Hashing ,Ordered Indices , B+ tree Index Files , B tree Index Files , Static Hashing ,Dynamic Hashing , Introduction to Distributed Databases, Client server technology, Multidimensional and Parallel databases, Spatial and multimedia databases, Mobile and web databases, Data Warehouse, data Mining, Data marts.

Unit 7: Advanced Topic **[8Hrs.]**

Concept of Object-Oriented and Distributed Database Model, Properties of Parallel and Distributed Databases, Threats and risks ,Database access Control, Types of Privileges ,Cryptography, Statistical Databases, Distributed Databases Architecture, Transaction Processing, Data Warehousing and Mining, Classification, Association rules-Clustering, Information Retrieval, Relevance ranking, Crawling and Indexing the Web, Object Oriented Databases,XML Databases.

Practical Works:

- Fundamental concept of MS-Access or MySQL or any suitable DBMS
- Database Server Installation and Configuration (MS-SQLServer, Oracle)
- DB Client Installation and Connection to DB Server
- Practice with DDL Commands. (Create Database and Tables)
- Practice of Procedure/Trigger and DB Administration & other DBs (MySQL, PG-SQL, DB2.)
- Group Project Development

References:

1. RamezElmasri&Shamkant B. Navathe (2015).*Fundamentals of Database Systems*, Seventh Edition, Pearson Education.
2. Korth, H. F. &Silberschatz, A. (2010).*Database system concepts*, McGraw Hill.
3. Majumdar, K.&Bhattacharaya, P. (2004).*Database Management Systems*, Tata McGraw Hill, India.
4. Abraham Silberschatz, Henry F. Korth& S. Sudharshan (2011).*Database System Concepts*, Sixth Edition, Tata McGraw Hill.
5. Date, C.J., Kannan, A.&Swamynathan, S. (2006).*An Introduction to Database Systems*, Eighth Edition, Pearson Education.
6. AtulKahate (2006) .*Introduction to Database Management Systems*, Pearson Education, New Delhi.
7. Alexis Leon & Mathews Leon(2003).*Database Management Systems*, Vikas Publishing House Private Limited, New Delhi.
8. Raghu Ramakrishnan (2010).*Database Management Systems*, Fourth Edition, Tata McGraw Hill.
9. Gupta, G.K.(2011).*Database Management Systems*, Tata McGraw Hill.

***

Code No.: MDS 506

Course Title: **Programming Concepts and Techniques** Full Marks: 75

Nature: Theory +Practical (Elective) Credit: 3

***Course Description*:**

This course covers concepts of program and programming language, different program design tools, and different concepts of programming using C programming language.

Learning Objectives:

After the completion of this course, the students should be able to

- Understand concepts of program and programming languages
- Know different program design tools
- Write programs using different concepts of C programming.

Course Contents:

Unit 1: Basic Concepts [4 Hrs.]

Program and Programming Languages; Program Design Tools: Algorithm, Flowchart, and Pseudocode; Coding, Compilation and Execution, History of C, Structure of C program, Debugging, Testing and Documentation.

Unit 2: Elements of C [4 Hrs.]

C Standards(ANSI C and C99), C Character Set, C Tokens, Escape sequence, Delimiters, Variables, Data types (Basic, Derived, and User Defined), Compiling and Executing a C program, Constants and Literals, Expressions, Writing Comments.

Unit 3: Input and Output [2 Hrs.]

Conversion specification, Reading a character, Writing a character, I/O operations, Formatted I/O.

Unit 4: Operators and Expression [4 Hrs.]

Arithmetic operator, Relational operator, Logical or Boolean operator, Assignment Operator, Ternary operator, Bitwise operator, Increment or Decrement operator, Conditional operator, Special Operators(sizeof and comma), Evaluation of Expression, Operator Precedence and Associativity.

Unit 5: Control Statement [6Hrs.]

Conditional Statements, Decision Making and Branching, Decision Making and Looping, Exit function, Break and Continue.

Unit 6: Arrays [6 Hrs.]

Introduction to Array, Types of Array (Single Dimensional and Multidimensional), Declaration and Memory Representation of Array, Initialization of array, Character Array and Strings, Reading and Writing Strings, Null Character, String Library Functions.

Unit 7: Functions [6Hrs.]

Library Functions, User defined functions, Function prototype, Function call, and Function Definition, Nested and Recursive Function, Function Arguments and Return Types, Passing Arrays to Function, Passing Strings to Function, Passing Arguments to Functions, Scope visibility and lifetime of a variable, Local and Global Variable.

Unit 8: Structure and Union [6Hrs.]

Defining and Accessing Structures, Array of structure, Passing structure to function, Passing array of structure to function, Nested Structure, Union, Comparing Structures with Unions.

Unit 9: Pointers [6 Hrs.]

Introduction, The & and * operator, Declaration of pointer, Pointer Arithmetic, Pointers and Arrays, Pointers and Character Strings, Array of Pointers, Pointers as Function Arguments, Function Returning pointers, Pointers and Structures, Dynamic Memory Allocation.

Unit 10: File Handling in C [4 Hrs.]

Concept of File, Opening and closing of File, Input Output Operations in File, Random access in File, Error Handling in Files.

Laboratory Works:

The laboratory work includes writing computer programs that covers all the concepts of C programming language including data types, variables, operators, all control statements, arrays, functions, structures and unions, pointers, and file handling.

References:

1. Byron Gottfried. *Programming with C*, Fourth Edition, McGraw Hill.
2. Herbert Schildt(2000). C *The Complete Reference*, Fourth Edition, Osborne/McGraw-Hill Publication.
3. Paul Deitel, Harvey Deitel, C.(2016).*How to Program*, Eighth Edition, Pearson Publication.
4. Al Kelley, Ira Pohl(2000) .*A Book on C*, Fourth Edition, Pearson Education.
5. Brian W. Keringhan, Dennis M. Ritchiem (1988).*The C programming Language*, Second Edition, Prentice Hall India .
6. Ajay Mittal (2010).*Programming in C.A Practical Approach*, Pearson Publication.
7. Stephen G. Kochan (2001) .*Programming in C*, CBS publishers & distributors.
8. E. Balagurusamy(2008).*Programming in ANSI C*, Third Edition, Tata McGraw- Hill publishing, New Delhi.

***

Code No.: MDS 507
Course Title:**Linear and Integer Programming** Full Marks: 75
Nature: Theory+Practical (Elective) Credit: 3

Course Description:

The course covers basic introduction of linear and integer based optimization problems, their few solution techniques and implementation of the solution techniques to solve real world problems formulated as linear programming problem and integer programming problem.

Course Objectives:

After the completion of this course, the students should be able to

- Know the concept and importance of convexity in optimization.
- Formulate real world problems in the form of linear programming problem(LPP) and integer programming problem (IPP).
- Solve LPP and IPP manually and using software as well.
- Solve the LPP and IPP using graphical and simplex methods.
- Know the total unimodularity, understand and implement cutting plane algorithm and branch and bound technique.

Course Contents:

Unit 1: Convexity and Optimization [6Hrs.]

Affine and Convex Sets, Convex Function, Convex Optimization Problem.

Unit 2: Problem Formulation [10 Hrs.]

Real world problems, Linear programming problem (LPP) formulation, Integer programming problem(IPP) formulation, Non-linear programming problem (NLP) formulation, Matlab and Python tutorial.

Unit 3: Linear Programming Problem [12Hrs.]

Three forms of LPP, Graphical Method, The Simplex Method, The General Problem, Linear Equations and Basic Feasible Solutions, Introduction to the Simplex Method, Theory of the Simplex Method, The Simplex Tableau and Examples, Articial Variables, Redundant Systems, A Convergence Proof, Linear Programming and Convexity, Spreadsheet Solution of a Linear Programming Problem.

Unit 4: Duality and Sensitivity Analysis [10 Hrs.]

Introduction to Duality, Definition of the Dual Problem, Examples and Interpretations, The Duality Theorem, The Complementary Slackness Theorem Examples in Sensitivity Analysis, Matrix Representation of the Simplex Algorithm, Changes in the Objective Function, Addition of a New Variable, Changes in the Constant-Term Column Vector, The Dual Simplex Algorithm, Addition of a Constraint.

Unit 5: Integer Programming Problem [10 Hrs.]

Introduction to Integer Programming, Total Unimodularity, Gomory's Cutting Plane Algorithm, A Branch and Bound Algorithm.

Practical Works:

The practical works includes Python and Matlabsoftwares.

References:

1. Stephen Boyd &LievenVandenberghe (2009) .*Convex Optimization*, Cambridge University Press.
2. Alexander Schrijver (1999).*Theory of Linear and Integer Programming*, John Wiley.
3. Paul R. Thie and G. E. Keough (2008).*An Introduction to Linear Programming and Game Theory*, John Wiley.
4. Laurence A Wolsey (1998).*Integer Programming*, John Wiley.

***

Tribhuvan University

Institute of Science and Technology

SCHOOL OF MATHEMATICAL SCIENCES

Syllabus

Master's in Data Sciences (MDS)- SECOND SEMESTER

Compulsory Courses

Course Code	Course Titles	Credits	Nature
MDS 551	Programming with Python	3	Th.+ Pr.
MDS 552	Applied Machine Learning	3	Th.+ Pr.
MDS 553	Statistical Methods for Data Science	3	Th.+ Pr.
MDS 554	Multivariable Calculus for Data Science	3	Th.

Elective Courses (Any One Available on School)

Course Code	Course Titles	Credits	Th.+ Pr.
MDS 555	Natural Language Processing	3	Th.+ Pr.
MDS 556	Artificial Intelligence	3	Th.+ Pr.
MDS 557	Learning Structure and Time Series	3	Th.+ Pr.

Code No.: **MDS 551**

Course Title:**Programming with Python** Full Marks: 75

Nature: Theory +Practical **(**Compulsory) Credit: 3

Course Description:

Python is a popular language for data science related activities. This course covers the concept of computer programming with python as an implementation language with focus on data processing, visualization and analysis.

Course Objectives:

This course is designed to familiarize students to the techniques of programming in python.

Course Contents:

Unit 1: Introduction to Programming [6 Hrs.]

Problem analysis, Algorithms and Flowchart, Coding, Compilation and Execution modern computer systems: hardware architecture, data representation in computers, software and operating system.

Installing Python; basic syntax, interactive shell, editing, saving, and running a script.

Unit 2:Data Types and Operators [6Hrs.]

Arithmetic Operators, Comparison Operators, Logical Operators, Logical Expressions Involving Boolean Operands, Logical Expressions Involving Non-Boolean Operands, Chained Comparisons, Bitwise Operators, Identity Operators, Operator Precedence, Augmented Assignment Operators.

Data Types: Python numbers, Strings, Lists, Dictionaries, Tuples, Sets, Using data type methods.

Unit 3:Control Statement [5 Hrs.]

Conditions, Boolean logic,ranges; Control statements: Decision Making with branching (if-else), Decision making with loops (for, while); short-circuit (lazy) evaluation.

Unit 4:String and Text files [6Hrs.]

Strings and text files; manipulating files and directories, os and sys modules; text files: reading/writing text and numbers from/to a file; creating and reading a formatted file (csv or tab-separated).

String manipulations: subscript operator, indexing, slicing a string; strings and number system: converting strings to numbers and vice versa.

Unit 5: List and Dictionaries [6Hrs.]

List Literals and Basic Operators: Replacing an Element in a List, List Methods for Inserting and Removing Elements, Searching a List, Sorting a List. Example program with List.

Dictionary Literals Adding Keys and Replacing Values Accessing Values Removing Keys Traversing a Dictionary, Example Program with Dictionary.

Unit 6:Functions **[6 Hrs.]**

Design with functions: hiding redundancy, complexity; arguments and return values; formal vs actual arguments, named arguments, Program structure and design.

Recursive functions: Tracing a Recursive Function, Using Recursive Definitions to Construct Recursive Functions, Infinite Recursion.

Unit 7: Python Libraries for Data Sciences **[11Hrs.]**

Numpy: Introduction, Environment Setup, Data Types, Array Attributes, Array Creation, I/O with Numpy, Array from Existing Data, Array from Numerical Ranges, Indexing & Slicing, Broadcasting, Iterating Over Array, Statistical Functions Sort, Search & Counting Functions.

Scipy: Introduction, Basic Functionality Cluster Constants Integrate Interpolate Input and Output Linalg.

Pandas: Series and DataFrames, Creating DataFrames from scratch (using list, Dictionaries, Numy array and another DataFrame) , Reading data from CSV and JSON, DataFrame Operations: Head and tail, Attributes and underlying data, handling of missing data, slicing, fancy indexing, and subsetting , merging and joining DataFrames.

Unit 8: Data Visualization with Matplotlib **[2Hrs.]**

Matplotlib: Setting up environment, Pyplot API, Simple Plot, Multi-plots, Subplots () Function, Subplot2grid () Function, Grids Formatting Axes. Setting Limits, Bar Plot, Histogram, Pie Chart, Scatter Plot, Contour Plot.

Laboratory Works:

Each programing concept is implemented as a laboratory work. This course should be carried out as practical based course.

References:

1. Kenneth A Lampart:*Fundamental of Python*, Cengage Learning Publishing.
2. Cody Jackson (2018):*Learn Programming in Python with Cody Jackson*, Packt Publishing, Wesley.

***

Code No.: **MDS 552**
Course Title: **Applied Machine Learning** Full Marks: 75
Nature: Theory+Practical **(**Compulsory) Credit: 3

***Course Description*:**
This course covers the concept of machine learning and it application in real world task. It includes Supervised, Unsupervised and reinforcement learning algorithms and evaluation metrics to choose the best algorithm for a particular task.

Course Objectives:
This course is designed to familiarize students to the concept of machine learning and their application.

Course Contents:

Unit 1: Introduction to Machine Learning [6Hrs.]

The Motivation & Applications of Machine Learning, The Definition of Machine Learning, Supervised Learning, Unsupervised Learning and Reinforcement Learning, Overview of Learning theory and Evaluation Metrics.

Unit 2: Supervised Learning [10Hrs.]

Supervised Learning, Linear Regression, Gradient Descent, Batch Gradient Descent, Stochastic Gradient Descent (Incremental Descent), The Concept of Under fitting and Over fitting, Locally Weighted Regression, Logistic Regression, Supervised learning Setup, Least Mean squares, Perceptron Learning Algorithm.

Classification, Linear Classifiers: Support Vector Machines, K-Nearest Neighbors, Multi-Class Classification, Kernelized Support Vector Machines, Naïve Bayes Classifiers, Decision Trees and Random Forest, Cross-Validation, Ensemble Learning, ensemble Size, Bagging, Boosting, Stacking.

Unit 3: Unsupervised Learning [10 Hrs.]

Clustering: Cluster Analysis, Partitioning Method: K-Means, Agglomerative and Divisive Clustering, Density Based Clustering: DBSCAN, Mixture Models and EM Algorithm,.

High Dimensional Data: Principal Component Analysis, Variants of PCA, Low Rank Approximations, Canonical Correlation Analysis, Latent Semantic Analysis.

Outlier Detection: Outlier Analysis, Outlier Detection Method, Clustering based approaches, Classification based Approach.

Unit 4: Model Evaluation and Selections [6 Hrs.]

Model Evaluation & Selection, Confusion Matrices & Basic Evaluation Metrics, Classifier, Decision Functions, Precision-recall and ROC curves, Multi-Class Evaluation, Regression Evaluation, Model Selection: Optimizing Classifiers for Different Evaluation Metrics.

Unit 5: Reinforcement Learning [6Hrs.]

Applications of Reinforcement Learning, Markov Decision Process (MDP), Defining Value &Policy Functions, Value Function, Optimal Value Function, Value Iteration, Policy Iteration,Generalization to Continuous States, Discretization & Curse of Dimensionality.

Unit 6: Neural Network and Deep Learning [10Hrs.]

Neural Network, Activation functions, learning rules, Back-propagation, Multi-layer Neural Networks, Feed Forward Neural Network, Recurrent Neural Network

Deep Neural Network: Convolution Neural Network, Image classification with CNN, Text Processing with RNN, Vanishing gradient and Dropout.

Laboratory Works:

Student are advised to implement supervised, unsupervised machine learning algorithm using any high level programing language (Python and Scikit-Learn preferred). The deep learning algorithms such as CNN and RNN should be implemented from scratch (Using library are not preferred).

References:

1. Forsyth, D.A. (2019) .*Applied Machine Learning*, 1st Edition, SpingerVerlag.
2. Sattari ,H. (2017).*Applied Machine Learning with Python*, Packt Publishing.

***

Code No.: **MDS 553**
Course Title:**Statistical Methods for Data Science** Full Mark: 75
Nature: **Theory and Practical (**Compulsory) Credit: 3

Course Description:

The course explains different probability distributions and their applications, some non-parametric statistical tests and their applications, different aspects of the testing of hypothesis along with Neymann-Pearson's lemma, uniformly most powerful tests, likelihood ratio tests for testing means and variance in exponential families.

Course Objectives:

After completion of this course, students will be able to

- Understand the concept of multinomial probability distributions, their major characteristics and applications
- Be familiar with probability functions of extreme value distributions, their major characteristics and their applications
- Understand the concept of generalized power series distribution with special focus to Binomial, Poisson, Negative Binomial distributions, and examples
- Know meaning and importance of prior and posterior distributions, applications focusing on some particular distributions and examples
- Understand how the distributions are compounding, understand mixed type distributions and their applications
- Know the difference between parametric and non-parametric statistical tests
- Apply non-parametric statistical tests appropriately in real life data analysis
- Understand the different aspects of testing of hypothesis, likelihood ratio tests and their applications.

Course Contents:

Unit 1:Multinomial Distribution [4 Hrs.]

Probability mass function, moment generating and characteristic function, moments, covariance and correlation, distribution fitting and examples.

Unit2: Extreme Value Distributions [4 Hrs.]

Probability density, distribution functions, moments, properties and examples.

Unit3: Generalized Power Series Distribution [6Hrs.]

Unified Probability mass function, it's special cases - Binomial, Poisson, Negative Binomial distributions and examples.

Unit4: Prior and Posterior Distributions [6Hrs.]

Meaning and examples including cases where Binomial, Beta, Exponential, Gamma, Poisson, Negative Binomial distributions and examples.

Unit 5: Compound and Mixed Type Distribution [6 Hrs.]

Compound Negative Exponential Distribution: Compounding of distributions, its moments.

Mixed Type Distribution: Mixed random variable, meaning and examples, computation of moments of mixed random variables, examples.

Unit6: Non-Parametric Tests [11Hrs.]

An overview of parametric tests, need of non-parametric statistical tests, Wilcoxon-Mann-Whitney U test, Median test, Fisher exact test for 2×2 tables, median test, Wilcoxon Sign ranks test, McNemar test, Kruskal-Wallis one-Way Analysis of Variance, Kolmogorov-Smirnov one sample and two sample tests, Friedman two way analysis of variance, relevant examples.

Unit 7: Testing of Hypothesis [11Hrs.]

General concept of simple and composite hypothesis, two types of errors, level of significance, power and size of a test.Most powerful test – Neymann Pearson's lemma and its application.Uniformly most powerful test- application to standard statistical distribution, unbiased test. Likelihood ratio test - Principle and properties, likelihood ratio test for testing means and variance in exponential families (without derivation), relevant examples.

Laboratory Works:

The applications of different probability distributions, testing of hypothesis using different statistical tests in real life data will be performed using appropriate software.

References:

1. Biswas, S. (1991). *Topics in Statistical Methodology*. India : Wiley Eastern
2. Chandra, T.K. and Chatterjee, D. (2003). *A First Course in Probability*. India: Narosa Publishing House.
3. Hoel, P.G., Port, S.C. and Stone, C.J. (1971). *Introduction to Probability Theory*. New Delhi India:Universal Book Stall.
4. Hogg, R.V. and Tanis, E.A. (2001). *Probability and Statistical Inference*. India: Pearson Education.
5. Kale, B.K. (1999). *A First Course on Parametric Inference*. Nindia: Narosa Publishing House.
6. Lehmann E.L. (1986). *Testing Statistical Hypotheses*. John Wiley and Sons.
7. Meyer, P.L. (1970). *Introductory Probability and Statistical Applications*. USA: Addison-Wesley.
8. Rohatgi, V.K. and Saleh, A.K.Md.E. (2005). *An Introduction to Probability and Statistics*. Singapore: John Wiley and Sons.
9. Shrestha, S. L. (2011). *Probability and Probability Distributions*. Kathmandu Nepal: S. Shrestha.
10. Zacks,S. (1971). *Theory of Statistical Inference*. John Wiley and Sons.

***

Code No.: **MDS 554**
Course Title:**Multivariable Calculus for Data Science** Full Mark: 75
Nature: **Theory(**Compulsory) Credit: 3

***Course Description*:**

This course extends single variable calculus to higher dimensions. It will cover the vocabulary for understanding fundamental processes and phenomena and provide important background needed for further study in many diverse fields, particularly in data science. It will build tools to describe geometric objects and apply problem solving methods to answer a variety of questions, mathematical and otherwise.

Course Objectives:

After successful completion of this course, the student will be able to

- Learn vectors and the geometry of space
- Work with Vector functions
- Learn partial derivatives
- Compute multiple Integrals
- Learn vector calculus

Course Contents:

Unit 1: Vectors and the Geometry of Space **[6 Hrs.]**

Three-Dimensional Coordinate Systems
Vectors
The Dot Product
The Cross Product
Equations of Lines and Planes

Unit 2: Vector Functions **[8 Hrs.]**

Vector functions and space curves
Derivatives and integrals of vector functions
Arc length and curvature
Motion in space

Unit 3: Partial Derivatives **[12Hrs.]**

Functions of several variables
Limits and continuity
Partial derivatives
Tangent planes and linear approximation
Chain rule
Directional derivatives and gradient vector
Maximum and minimum values
Lagrange multipliers

Unit 4: Multiple Integrals **[10Hrs.]**

Double integrals
Polar coordinates
Applications of double integrals
Surface area
Triple integrals
Change of variables in multiple integrals

Unit 5: **Vector Calculus** **[12Hrs.]**

Vector fields
Line integrals
Green's theorem
Curl and divergence
Parametric surfaces and their areas
Surface integrals
Stokes' theorem
Divergence theorem

References:

1. Edwards, Henry C., and David E. Penney (2002) .*Multivariable Calculus*. Prentice Hall,

2. Oliver Knill (2018).*Multivariable Calculus*, Harvard University

http://people.math.harvard.edu/~knill/teaching/summer/

3. James Stewart, *Multivariable Calculus*(2009).*Concepts and Contexts*, CengageLearning .
4. Denis Auroux (2010). *Multivariable Calculus.* Massachusetts Institute of Technology: MIT Open Course Ware, https://ocw.mit.edu..

***

Code No.: **MDS 555**
Paper: **Natural Language Processing** Full Marks: 75
Nature: Theory +Practical (Elective) Credit: 3

Course Description:

The course covers the introductions, methods and approaches used in many real-world NLP applications such as Computational Linguistics, Morphology, Syntax, Semantics, Discourse.

Course Objectives:

After successful completion of this course, the student will be able to

- Provide the students a general overview of the basics as well as the advanced concepts of Natural Language Processing (NLP)
- Apply the different concepts of NLP both theoretically and practically.

Course Contents:

Unit 1: Introduction to NLP [4 Hrs.]

Introduction to NLP, Origins and importance of NLP, Challenges in NLP (Difficulties, Ambiguities and Evolution), Language and Knowledge (Syntax, Semantics, Pragmatics and Discourse), A Multi-disciplinary field (Psychology, Information Retrieval), Applications of NLP.

Unit 2 : Words and Morphology [7Hrs.]

Finite State Machines (FSM) and Morphology, Introduction to FSM and FST, Morphological Processes, Principles of Word Construction (Suffix, Prefix, Stem, Affixes), Morphological Representation and FSM, Lexicon, Morphotactic and Orthographic rules, Morphological Parsing and FST, Mealy machines, FST operations.

Unit 3: Part of Speech Tagging [7 Hrs.]

Parts of Speech (PoS) Tagging and Hidden Markov Models (HMM), PoSTagsets, Rule-based PoS Tagging, Stochastic PoS Tagging, Transformation based tagging.

Unit 4: Syntax [9Hrs.]

Syntactic Analysis, Context Free Grammar (CFG) & Probabilistic CFG, Word's Constituency (Phrase level, Sentence level), Parsing (Top-Down and Bottom-Up), CYK Parser, Probabilistic Parsing.

Unit 5: Lexical Semantics [7 Hrs.]

Lexical Semantics, Lexeme, Lexicon, Senses, Lexical relations, WordNet (Lexical Database), Word Sense Disambiguation (WSD), Word Similarity.

Unit 6: Discourse [7Hrs.]

Pragmatic & Discourse Analysis, Monologue and Dialogue, Reference Resolution, Coherence and Cohesion, Discourse Structure.

Unit 7: Applications of NLP **[7Hrs.]**

Applications of NLP, Question Answering, Machine Translation, Sentiment Analysis, Summary Generation.

Lab and Practical Works:

In the lab and practical works, the students will basically get practical concepts of NLP in the Python Programming Language . A lot of these would be hands-on exercises and writing the codes of NLP problem- solving.

References:

1. Daniel Jurafsky and James H. Martin (2009). *Speech and Language Processing* , Second Edition, Pearson Education.
2. Stephen Bird, Ewan Klein& Edward Loper (2009). *Natural Language Processing with Python*. O'Reilly Media, http://www.nltk.org/book/

***

Code No.: **MDS 556**
Paper: **Artificial Intelligence** Full Marks: 75
Nature: Theory + Practical (Elective) Credit: 3

Course Description:

This course covers the underlying principles and theories of artificial intelligence. The course covers the design of intelligent agents, problem solving, searching techniques, knowledge representation systems, concepts of neural networks, machine learning techniques. In covers applications of AI in the field of natural language processing, expert systems, machine vision as well.

Course Objectives:

The main objectives of the course are to

- Understand concepts of artificial intelligence
- Learn about intelligent agents and design the agents,
- Identify AI problems and solve the problems using AI techniques,
- Design knowledge representation systems and expert systems,
- Understand concepts of machine learning
- Understand concepts of artificial neural networks
- Understand application of AI.

Course Contents:

Unit 1: Introduction [4Hrs.]

Introduction of Artificial Intelligence
Defining Artificial Intelligence: acting and thinking humanly:Turing Test, acting and thinking rationally
Foundations of Artificial Intelligence
History of Artificial Intelligence
Applications of Artificial Intelligence

Unit 2:Agents [6Hrs.]

Agent, Intelligent Agent, Rational Agent
Structure of Intelligent Agent: Agent Function, Agent Program
Configuration of Agents: PEAS/PAGE description of Agents
Agent Types
Environment Types

Unit 3: Solving Problems by Searching [10 Hrs.]

Problem, State Space Representation,
Formulating Problems, Solving Problems by Searching, Types of Search
Blind Search: Depth First Search, Breadth First Search, Depth Limited Search, Iterative Deepening Search, Uniform Cost Search, Bidirectional Search
Informed Search: Heuristic, Heuristic Function, Greedy Best first search, A* search, Admissibility and Optimality of A*
Local Search: Hill Climbing, Simulated Annealing
AND-OR Search Trees
Adversarial Search: Mini-max Algorithm, Alpha-Beta Pruning.
Constraint Satisfaction Problems

Unit 4: Knowledge Representation and Reasoning [15Hrs.]

Knowledge, Knowledge Representation in agents

Knowledge Representation Systems

Types of Knowledge Representation Systems: Semantic Network, Frame, Conceptual Dependency, Script, Rule Based System, Propositional Logic, Predicate Logic

Propositional Logic(PL):Syntax and Semantics, Proof by Resolution, Conjuctive Normal Form, Resolution Algorithm

Predicate Logic: FOPL, Syntax and Semantics, Quantifiers, Unification and Lifting, Inference using Resolution Algorithm

Uncertain Knowledge: Uncertainity, Radom Variables, Probability, Prior and Posterior Probability, Probabilistic Reasoning, Bayes' Rule and its use, Bayesian Networks

Fuzzy Logic and Fuzzy Rule Base System

Unit 5: Concepts of Machine Learning [7Hrs.]

Introduction to Machine Learning

Supervised, Unsupervised and Reinforcement Learning

Learning with Neural Networks: Artificial Neural Networks (ANN), Mathematical Model of ANN, Types of ANN, ANN for simulation of gates, Learning by ANN, Perceptron Learning, Back-propagation Algorithm

Deep Learning

Statistical-based Learning: Naive Bayes Model

Learning by Evolutionary Approach: Genetic Algorithm

Unit VI: Applications of AI [6 Hrs.]

Expert System

Natural Language Processing

Robotics

Machine Vision

AI in Data Science

Laboratory Works:

Students should implement intelligent agents, expert systems, various search techniques, knowledge representation systems and machine learning techniques usingappropriate programming language.

References:

1. Russel, S.&Norvig, P..*Artificial Intelligence A Modern Approach*, Pearson.
2. Rich, E., Knight, K. &Nair, S. B. .*Artificial Intelligence*, Tata McGraw Hill.
3. G. F. Luger, Artificial Intelligence: *Structures and Strategies for Complex Problem Solving*, Addison Wesley.
4. Winston, P. H. .*Artificial Intelligence*, Addison Wesley.
5. Jackson, P. C. .*Introduction to Artificial Intelligence*, Dover Publications Inc.
6. Patterson, D. W. ,*Artificial Intelligence and Expert Systems*, Prentice Hall.
7. Konar, A..*Artificial Intelligence and Soft Computing*: *Behavioral and Cognitive Modeling of the Human Brain*, CRC Press.

***

Code No.: **MDS 557**

Paper: **Learning Structure and Time Series** Full Marks: 75

Nature: Theory +Practical (Elective) Credit: 3

Course Description:

In this course students will learn the fundamental concept of Learning Structure along with cluster analysis and dimensional reduction. Similarly, students will study the concept of time series with different stationary and non-stationary models with forecasting. The goal of the course is to prepare the student to formulate and solve learning problems and time series problems in multiple domains. The course will use R extensively for solving all the problems practically.

Learning Objectives:

After completion of the course, students will be able to

- Formulate data-driven learning problems.
- Differentiate between supervised and unsupervised learning tasks
- Use R in Regression, Cluster analysis and Dimension Reduction
- Decompose time series data into its constituent parts and use for policy analysis and forecasting
- Explore graphically and summaries time series data using R.

Course Contents:

Unit 1: Introduction to Learning Structure [5 Hrs.]

Introduction to learning structure: Supervised VsUnsupervised learning, model Assessment, linear regression, Estimating Coefficients and Estimating the accuracy of coefficients focusing on learning structure and itsusagein R.

Unit 2: Cluster Analysis [7Hrs.]

Types of Data in Cluster Analysis, hierarchical clustering, Bayesian clustering: spectral clustering, Partitioning methods: K means clustering, mixture method, Application of R in Cluster Analysis.

Unit 3: Dimension Reduction [8 Hrs.]

Concept of Dimension Reduction, Unsupervised embedding techniques: Principal Component Analysis (PCA), Kernel PCA, Multidimensional Scaling (MDS), Supervised reduction techniques: Feature selection, forward selection and backward selection. Dimension Reduction using R.

Unit 4: Time Series Analysis [7 Hrs.]

Exploratory analysis and graphical display (Time and seasonal plot),Time series decomposition (trend, seasonal, cyclical and irregular), additive and multiplicative models, moving average, exponential smoothing and Usage of R

Unit 5: Time Series Models [15Hrs.]

Auto Regressive (AR), Moving Average (MA), and ARMA Models, Box-Jenkins Correlogram analysis, (Auto-Correlation Function) ACF and (Partial Auto-Correlation Function) PACF, Choice of AR and MA orders, Autoregressive Integrated Moving Average (ARIMA) model, Deterministic and stochastic trends, ARCH (Autoregressive Conditional Heteroscedasticity) and Generalized ARCH (GARCH) model,Vector Error Correction Models and Cointegration, State-Space Models, Granger Causality and their application in R

Unit 6: Forecasting in Time Series [6 Hrs.]

Forecasting Using Exponential Smoothing and Box-Jenkins Methods and Residual Analysis,Artificial Neural Networks, Fuzzy time-series, DBSCAN algorithms and their analysis in R.

Practical Works:

The practical work consist of lab work usingR/RStudio software.

References:

1. Bishop, Christopher M. (2006). *Pattern recognition and machine learning.* New York :Springer,
2. Shumway, R. H., &Stoffer, D. S. (2011).*Time series analysis and its applications: With R examples*. New York: Springer.
3. Friedman, J., Hastie, T., &Tibshirani, R. (2008):*The elements of statistical learning*. New York: Springer series in statistics.
4. Konar, A., Bhattacharya,D. (2017):*Time-Series Prediction and Applications*: *A MachineIntelligence Approach*, Switzerland: Springer.

***

Tribhuvan University

Institute of Science and Technology

SCHOOL OF MATHEMATICAL SCIENCES

Syllabus

Master's in Data Sciences (MDS)- THIRD SEMESTER

Compulsory Courses

Course Code	Course Titles	Credits	Nature
MDS 601	Research Methodology	3	Th.
MDS 602	Advanced Data Mining	3	Th.+ Pr.
MDS 603	Techniques for Big Data	3	Th.+ Pr.

Elective Courses (Any Two Available on School)

Course Code	Course Titles	Credits	Nature
MDS 604	Cloud Computing	3	Th.+ Pr.
MDS 605	Regression Analysis	3	Th.+ Pr.
MDS 606	Decision Analysis	3	Th.+ Pr.
MDS 607	Monte Carlo Methods	3	Th.

Code No. : MDS 601
Course Title: Research Methodology **Full Marks:** 75
Nature: Theory (Compulsory) **Credit: 3**

Course Description

The paper deals with different essential components of research methodology mostly focusing on quantitative research work. It explains the fundamental notions of research methodology, formulation of research problems and hypothesis, different aspects of review of literature and study designs. It also deals with the practical aspects of sampling, choice of appropriate data analysis technique(s) and interpretations of statistical findings with reference to the data problem. Finally, how a research report, journal article is prepared, tools and techniques used in preparing journal articles, ethical issues and the issues of plagiarism are included. The tutor should involve the students practically in preparing the project proposals and thesis proposals, which they will be going to conduct in the next semester.

Learning Objectives

After completion of the course, students will be able to

- Understand the need and importance of research methodology
- Formulate research problem, research hypothesis
- Know the skills of reviewing literature, citation and referencing
- Understand the different types of research design
- Perform online searching and prepare research/ project proposal
- Choose appropriate statistical tools and interpret the results
- Know how to prepare research reports and journal articles

Course Contents:

Unit 1: Introduction to Research Methodology **[5 Hrs.]**

Meaning, objectives, need, utility, research methods and methodology, Deductive and inductive theory, Characteristic of scientific method, Research language, concept, constructs, research process

Unit 2: Research Problem Identification and Formulation **[5 Hrs.]**

Understanding research problem and its formulation, Statement of research problem, Research question, Research hypothesis, Statistical hypothesis, Linkage between research hypothesis and statistical hypothesis

Unit 3: Review of Literature and Research Design **[12 Hrs.]**

Importance of review of literature in research work, Review of research journals, Use of encyclopedia, reference books, research guides, handbooks, academia database in the relevant field, Citations and referencing, American Psychological Association(APA), IEEE style, Bibliography; Need and importance of research design, Exploratory, descriptive, analytical, and experimental research design; Online searching: Different database, SCIFinder, Scopus, Science direct, Searching research articles, Citation index, Impact factor, H-index; Preparation of project proposals and research proposals

Unit 4: Data Analysis and Interpretations **[15 Hrs.]**

Concept of measurement and scaling, Validity and reliability, Levels of measurement- nominal, ordinal, interval and ratio scales; Sampling: Concepts of statistical population, sample survey, sampling frame, target population, sampling and non-sampling errors, Issues of choosing appropriate sampling technique(s) while selecting samples in the study, sample size, issues of practical considerations in sampling and sample size; Use of appropriate statistical tests and models, interpretations of statistical findings in terms of problem specific sense

Unit 5: Preparation of Research Report **[11 Hrs.]**

Preparation of academic and other reports, Formats of research reports (both academic and others), Research paper writing, layout of research paper, Journals in relevant field, Impact factor journals, Communications in the journals for publishing research article, Plagiarism, ethical issues in research and publishing the articles, Techniques and tools used for research: Effective way of searching research materials, Handling reference management software such as Mendeley, research paper formatting software such as LaTeX, MS Office.

References:

1. Kothari C. R. (2014). *Research Methodology, Methods and Techniques*, 3rd Edition, New Age international Publishers, New Delhi, India.
2. Kumar Ranjit(2011). *Research Methodology: A Step by Step Guide for Beginners*, SAGE Publications PVt. Ltd, India.
3. Creswell J.W.(2014). *Research Design: Qualitative, Quantitative and Mixed Methods Approaches*, 4th Edition, SAGE Publications, India.
4. Day Robert A. (1996). *How to Write and Publish Scientific paper*, Cambridge University Press
5. Relevant updated References from the internet

●●●●●

Code No.: MDS 602
Course Title: Advanced Data Mining **Full Marks: 75**
Nature: Theory + Practical (Compulsory) **Credit: 3**

***Course Description*:**

This course covers some basic and advanced concepts of data mining techniques including techniques for association analysis, classification, clustering, and outlier detection.

***Learning Objectives*:**

After completion of this course, students should be able to

- Understand concepts like pre-processing, measures of similarity, summary statistics and visualization of data
- Understand and implement different algorithms for association, classification, clustering, and anomaly detection

***Course Contents*:**

Unit 1: Introduction **[6 Hrs.]**

Introduction of data mining, Origins of data mining, Data mining tasks, Types of data, Data quality, Data pre-processing, Measures of similarity and dissimilarity, Summary statistics and visualization.

Unit 2: Association Analysis **[7 Hrs.]**

Basic concepts, Apriori algorithm, Pattern-growth approach, Handling categorical and continuous attributes, Handling concept hierarchy, Sequential and subgraph patterns

Unit 3: Classification **[14 Hrs.]**

Decision trees induction, Rule-based classifier, Nearest-neighbor classifier, Bayesian classifier, Artificial neural network (ANN), Support vector machines, Ensemble methods, Model evaluation and selection

Unit 4: Clustering **[12 Hrs.]**

Overview; K-Means, Agglomerative hierarchical clustering, DBSCAN, Fuzzy clustering, Graph based clustering, Clustering evaluation

Unit 5: Outlier Detection **[9 Hrs.]**

Preliminaries, Outliers and types, Statistical approaches, Proximity-based approaches, Clustering-based approaches, Classification based approaches

Laboratory Works:

Laboratory work includes writing computer programs using high level programming language like Python to implement all the data mining techniques and algorithms studied in the course.

References:

1. Pang-Ning Tan, Michael Steinbach, and Vipin Kumar (2014). *Introduction to Data Mining*, Pearson.
2. Jiawei Han, Micheline Kamber, and Jian Pei, Data Mining (2012). *Concepts and Techniques*, Third Edition, MK.
3. David Forsyth (2019). *Applied Machine Learning*, Springer.

●●●●●

Code No.: MDS 603

Course Title: Techniques for Big Data **Full Marks: 75**

Nature: Theory + Practical (Compulsory) **Credit: 3**

***Course Description*:**

In this course students will learn fundamental knowledge to handle the challenges of Big Data. This course shall first introduce the overview of Big Data, its scope, applications, challenges and current trends. Then, it will introduce the fundamental platforms, such as Hadoop, MongoDB and HBase as NoSQL databases, Spark, Hive and Pig. Students will then have fundamental knowledge on Big Data Analytics to solve various real-world problem posed by Big Data.

Learning Objectives:

Upon the conclusion of the course, students should be able to:

- Understand the fundamentals of big data, its characteristics, scopes and challenges
- Setup Hadoop cluster and perform basic file operations in HDFS
- Write MapReduce Programs to solve big data problems in real life run the job in Hadoop Environment
- Understand different types of NoSQL database and analyze Structured, Unstructured and Semi Structured Data
- Perform Real Time Analytics with Spark Streaming
- Preform SQL Like queries on top of Hadoop using Spark SQL, Pig and Hive.

***Course Contents*:**

Unit 1: Introduction **[4 Hrs.]**

Introduction to big data, Characteristic of big data, Current trend and real-life applications of big data, Scope and challenges of big data, Tools and technologies used in big data.

Unit 2: Hadoop Ecosystem **[6 Hrs.]**

Introduction to Hadoop, History of Hadoop, Hadoop ecosystem,
Core components of the Hadoop ecosystem: Hadoop common, Hadoop distributed file system (HDFS), Map reduce framework and YARN (Yet another resource negotiator),
Hadoop master/slave architecture, Hadoop daemons, Hadoop configuration modes, Hadoop cluster setup, Hadoop streaming.

Unit 3: Distributed Storage and Processing of Big Data **[12 Hrs.]**

HDFS: The Design of HDFS, HDFS Concepts, Command line interface, Hadoop file system interfaces, Data flow, Basic HDFS commands,
MapReduce: Functional programing, MapReduce fundamentals, Execution overview of MapReduce, Basic MapReduce API Concepts, Setting up the development environment, Writing unit test with MRUnit, Running locally on test data, Running on a cluster, Anatomy of a MapReduce job run, Failures, Shuffle and sort, Task execution, Map reduce types and formats.

Unit 4: NoSQL Databases [10 Hrs.]

Types of data, Introduction to NoSQL, Need of NoSQL, Types of NoSQL Databases, NoSQL v's Relational databases, The CAP theorem,

MongoDB (Collections, Documents, Object Ids, Queries on MongoDB, Aggregation pipeline, Nested documents),

HBase (Overview, HBase vs. RDBMS, HBase vs. HDFS, HBase Architecture, HBase Data Model, Concept of Row Keys and Column Families, HBase Regions, Creating a Table, Writing Queries to insert and retrieve data to and from HBase.

Unit 5: Data Analytics with Spark [6 Hrs.]

Introduction to spark, Need of spark, Evolution of spark, Spark shell, Spark context, Resilient distributed dataset (RDD), Transformations, Programming with RDD, Spark core, Spark SQL, MLib, Spark streaming, GraphX.

Unit 6: Querying Big Data with Pig and Hive [10 Hrs.]

Pig: Introduction to pig, Execution modes of pig, Comparison of pig with databases, Grunt, Pig Latin, User defined functions, Data processing operators.

Hive: Hive shell, Hive services, Hive metastore, Comparison with traditional databases, HiveQL, tables, Querying data and User defined functions

Practical Works:

1. Setup Hadoop cluster in pseudo distributed and fully distributed mode
2. Perform basic file system operation on HDFS
3. Write MapReduce programs in Java/Python and run the job in Hadoop cluster
4. Analyze Big Data (CSV and JSON) using Hadoop MapReduce
5. Install MongoDB and perform CRUD operations
6. Twitter data analysis using MongoDB
7. Install HBase in pseudo distributed and fully distributed mode and perform CRUD operations
8. Use SparkSQL and dataframe to analyze CSV and JSON data
9. Use spark streaming for real time analytics
10. Installing and running pig
11. Installing and running hive
12. Project work to analyze big data (CSV and JSON) using any 3 technologies among Hadoop MapReduce, MongoDB, HBase, Spark, Pig and Hive

References:

1. Tom White (2015). *Hadoop: The Definitive Guide Fourth Edition*, O'Reilly Media.
2. Jeffrey Dean and Sanjay Ghemawat (2004). *MapReduce: Simplified Data Processing on Large Clusters*, Google.
3. Judith Hurwitz (2013). Alan Nugent, Dr. Fern Halper, and Marcia Kaufman, *Big Data for Dummies*, Wiley
4. Chuk Lam (2011). *Hadoop in Action*, Manning
5. Adam Fowler (2015), *NoSQL for Dummies*, Wiley

●●●●●

Code No.: MDS 604
Course Title: Cloud Computing **Full Marks: 75**
Nature: Theory +Practical (Elective) **Credit: 3**

Course Description:

In this course students will learn the fundamental concept of cloud computing, the course contents the cloud computing infrastructure, specification of cloud system structure and implementation of specifications, the evolution of cloud computing methods and tools for development of cloud infrastructure in an economic and timely manner. The goal of the course is to provide the knowledge of as software as service, platform as service and infrastructure as service.

Learning Objectives:

After completion of the course, students will be able to

- Understand the concepts and principles of cloud computing its role
- Describe about Cloud reference model, Cloud deployment models, Cluster and Grid computing: Grid computing Versus Cloud computing.
- Explain the concept of virtualization, MapReduce and benefit of virtualization.
- Describe the concept, principles and practice of security use in cloud computing and Security Architecture Design
- Describe the cloud platform, application and some case studies related to cloud computing

***Course Contents*:**

Unit 1: Introduction to Cloud Computing [6 Hrs.]

History and development of Cloud computing , Characteristics of Cloud computing, Types of cloud, Cloud services: Benefits and challenges of cloud computing, Evolution of Cloud computing , Applications cloud computing, Business models around Cloud, Cloud Architecture, Cloud storage, Cloud services requirements, Cloud and dynamic infrastructure, Cloud adoption.

Unit 2: Architecture of Cloud Computing [6 Hrs.]

Cloud computing Characteristics, Cloud reference model -platform as service, software as a service, infrastructure as service, Cloud deployment models -Public clouds, private clouds, Community cloud, hybrid clouds, Cloud design and implementation using SOA, security, trust and privacy

Unit 3: Data in the Cloud [8 Hrs.]

Map-Reduce and extensions: Parallel computing, The map-Reduce model, Parallel efficiency of Map-Reduce, Relational operations using Map-Reduce, Enterprise batch processing using Map-Reduce, Introduction to cloud development, Map-Reduce model

Unit 4: Cloud Virtualization Technology [10 Hrs.]

Virtualization defined, Types of Virtualization, Implementation Levels of Virtualization Structures, virtualization benefits, server virtualization, hypervisor management software, virtual infrastructure requirements.

Unit 5: Cloud Security **[6 Hrs.]**

Introduction to security, Cloud security challenges and risks, Software as-a-service security, Security monitoring, Security architecture design, Data security, Application security, Virtual machine security, Identity management and access control, Autonomic security**.**

Unit 6: Cloud Platforms, Applications and Case Studies **[12 Hrs.]**

Web services, appEngine, azure Platform, Aneka, open challenges, scientific applications, business and consumer applications.

Practical Works:

The practical work consists of Cloud programming models such as Thread programming, Task programming and Map-reduce programming.

References:

1. Dr. Kumar Saurabh (2012). *Cloud Computing, 4th ed .,* John Wiley and Sons
2. Raj Kumar Buyya, Christian Vecchiola, S. Thamarai Selvi (2013). *Mastering Cloud Computing* , Tata McGraw-Hill Education
3. David S. Linthicum (2010). *Cloud Computing and SOA Convergence in your enterprise,* Pearson Education, Inc., India.
4. Barrie Sosinsky (2011). *Cloud Computing Bible* John Wiley and Sons.
5. Saurabh, K. (2011). *Cloud Computing – Insights into New -Era Infrastructure,* Wiley India.

●●●●●

Code No. : MDS 605

Course Title: Regression Analysis **Full Marks:** 75

Nature: Theory and Practical (Elective) **Credit: 3**

Course Description:

The paper covers important intermediate to advanced level topics on regression analysis useful for statistical data analysis and modeling. It deals with multiple regression analysis, its uses and associated model adequacy tests. Additionally, it covers issues on modeling techniques and methods to deal with quantitative and qualitative variables. Finally, techniques of choosing the best regression is included.

Learning Objectives:

The general objective of the course is to make students understand and enable them to apply regression analysis methods effectively so that they are able to associate and establish relationship between variables statistically. Specifically, the objectives are to make them capable with handling multiple linear regression, regression with qualitative variables, and selection of best regression appropriately during data analysis with focus in data science.

***Course Contents*:**

Unit 1: The Multiple Linear Regression **[10 Hrs.]**

The *k* variable linear regression model, model specification including in matrix form, assumptions, OLS estimation of parameters and their interpretations, the hat matrix, properties of least square estimates, standard error and confidence interval of estimates and mean response, test of significance, analysis of variance for goodness of fit, prediction, unadjusted and adjusted multiple coefficient of determinations, standardized regression coefficients, lack of fit tests, use of computer packages in regression analysis including R package.
Problems and examples focusing data science

Unit 2: Model Adequacy Tests: Normality and Multicollinearity **[10 Hrs.]**

Normality: Kolmogorov-Smirnov and Anderson-Darling tests, Normal probability plots: P-P and Q-Q plots for assessing normality
Multicollinearity: Nature of multicollinearity, detection of multicollinearity, checking for correlations between explanatory variables, wrong regression coefficient signs, variance inflation factors, condition index, consequences and remedial measures

Problems and examples focusing data science

Unit 3: Model Adequacy Tests: Heteroscadasticity and Residual analysis **[10 Hrs.]**

Heteroscadasticity: Nature of heteroscedasticity and its detection through residual plots, Goldfeld-Quandt test, Breusch-Pagan-Godfrey test, Park test, treatment of heteroscedasticity: method of weighted least squares.

Autocorrelation and residual analysis: Definitions of residuals, raw, standardized and studentized residuals, the PRESS residuals, the nature of autocorrelation, detection and consequences of auto-correlation, residual plots for detecting autocorrelation including partial residual plots, Durbin Watson test, remedial measures: changing the functional form, Cochrane-Orcutt iterative procedure, detection and treatments of outliers.

Problems and examples focusing data science

Unit 4: Regression with Categorical Variables [12 Hrs.]

Regression with explanatory categorical or indicator variables: Nature of categorical or dummy variables and their effects and uses in regression models, regression with categorical independent variables: regression with one quantitative and one qualitative variables with two classes and more than two classes, use of multiple quantitative and qualitative independent variables, Use of measuring interaction effects using dummy variables.

Regression with qualitative dependent variables: The nature of categorical dependent variables in regression modeling, The binary logistic model, its importance and uses in data modeling, assumptions, estimation of parameters and their interpretations, odds ratio and its interpretation, model adequacy tests: Hosmer-Lemeshow test, ROC curve.
Problems and examples focusing data science

Unit 5: Methods of Selection of Best Regression [6 Hrs.]

All possible regressions and best subset regression, Stepwise regression: Forward and backward methods, Selecting significance levels in stepwise regression, The use of R^2 Statistic, Residual mean square and the Mellows C_p Statistic, Akaike information criterion
Problems and examples focusing data science.

Practical works:

The practical works include analysis and modeling for multiple regression, models for categorical data and selection of best possible regression including their statistical model adequacy tests using any statistical software.

References:

1. Douglass C. Montgomery, Elizabeth A. Peck and G. Geoffrey Vining (2012). *Introduction to linear regression analysis*, Fifth edition, Wiley.
2. Drapper, N. R. and Smith, H. (1998). *Applied Regression Analysis*, Third edition, Wiley.
3. Maddala, G. S. (2002). *Introduction to Econometrics*, John Wiley and Sons.
4. Ramanathan, B. (2002). *Introductory Econometrics with Applications*, South-Western Thomson Learning, Singapore.
5. Brian Caffo (2015). *Regression models for data science in R*, Lean publishing.

●●●●●

Code No. : MDS 606
Course Title: Decision Analysis **Full Marks: 75**
Nature: Theory + Practical (Elective) **Credit: 3**

***Course Description*:**

This course provides an overview on formal, structured, systematic and visual approach to evaluate problems that leads to decisions and action. It covers both philosophical aspects of decision theory and practical aspects of decision making environments under probabilistic and non-probabilistic situations. Moreover, it covers an interactive decision approach: theory of games which is indispensible while assessing the enterprise risks. It also covers the multi-objective functions related to optimization problem under goal programing.

Learning Objectives:

After the completion of the course, students will be able to

- Utilize a range of methods and tools to aid in the capture, analysis and synthesis of information for the effective decision making
- Identify the values, objectives, attributes, decisions, uncertainties, consequences, and trade-offs in a real decision problem
- Assess the enterprise risk management and adapt the interactive decision approach in real life
- Deal with multiple objectives taking into consideration of their priority
- Use Microsoft Excel / LINGO or latest software in decision analysis

Course Contents:

Unit 1: Introduction to Decision Theory [6 Hrs.]

History and introduction of decision analysis, Decision making philosophy, Elements of decision problems, Framework of decision making, Decision making processes, Problem solving and creative thinking, , Problem or opportunity findings, Types of decision , Decision making conditions, Cognitive biases, Decision making styles, Decision theories: Classical decision theory, Behavioral theory, Normative and descriptive decision theory; Group decision making , Techniques of group decision making

Unit 2: Decision Making Under Uncertainty and Risk [16 Hrs.]

Structuring a decision problem, Decision making under uncertainty (Criterion of optimism, Criterion of pessimism, Laplace criterion, Criterion of realism, Criterion of regret), Conditional probability and use of Bayes theorem, Posterior probabilities and Bayesian analysis ,Value functions, Decision tree analysis, Decision making with utilities: Utility function and loss function, Expected utility, Visualizing uncertainty, Utility curves ; Decision making under risk (EMV, EOL, EVPI), Decision making under ignorance ; Risk analysis and Sensitivity analysis.

Unit 3: Enterprise Risk Management [6 Hrs.]

Introduction, History, Trends in ERM, Risk management standard and guidance, Enterprise risk management integrated framework, Risk profile, Risk Appetite and Risk tolerance

Unit 4: Theory of Games [8 Hrs.]

Introduction, Game models: Two-Person Zero-Sum games, pure strategies: games with saddle point, mixed strategies: games without saddle point, Principles of dominance, solution methods of games without saddle points.

Unit 5: Goal Programming **[12 Hrs.]**

Philosophy of goal programming: Satisficing, Optimizing, Ordering, Balancing; Goal programming variants: Generic goal program, Distance metric based variants, Decision variables, Goal-based variants; Formulation of goal programming: Formulating goals and setting targets, Variant choice, Normalization; solving and analyzing goal programming, application of goal programming in various areas.

Practical Works:

The practical work consist of real-world problems with application of various techniques and methods including multi-attribute utility models, decision trees, and Bayesian models using Microsoft Excel and LINGO/ or latest Software. It also includes Project work with applications to real-world data.

References:

1. Charles Yoe. (2019). *Principle of Risk Analysis Decision Making Under Uncertainty*. New York: Taylor and Francis Group . Retrieved from https://b-ok.asia/book/3711279/e97f5b

2. Anderson D.R., Sweeney D.J., Williams T.A. & Martin K. (2011). *An introduction to Management Science Quantitative Approaches to Decision Making*. Delhi: Cengage Learning India Private Limited.

3. Dylan Jones & Mehrdad Tamiz. (2010). *Practical Goal Programming*. New York: Springer. Retrieved from https://b-ok.asia/book/1056846/f876da

4. Howard, Ronald A. and Abbas, Ali E. (2015). *Foundations of Decision Analysis. England: Pearson Education Ltd.*
5. Sharma J.K. (2004). *Quantitative Techniques for Managerial Decisions*. New Delhi: Rajiv Beri for Macmillan India Ltd.
6. Martin Peterson. (2009). *An introduction to Decision Theory*. New York: Cambridge University Press. Retrieved from https://b-ok.asia/book/2479794/2849b4
7. Vohra N.D. (2004). Quantitative *Techniques in Management*. New Delhi: Tata McGraw-Hill Publishing Company Limited.
8. Sven Ove Hansson, (2015). *Decision Theory: A Brief Introduction.* Stockholm: Royal institute of Technology

•••••

Code No.: MDS 607
Course Title: Monte Carlo Methods **Full Marks: 75**
Nature: Theory +Practical (Elective) **Credit: 3**

Course Description:

The goal of this course is to prepare the student to get an idea of Monte Carlo methods and able to apply to solve problems of data science. In this course, students will learn the fundamental concept of Bayesian statistics and the difference between frequentist and Bayesian inference. The students will study the concept of simple and Markov chain Monte Carlo methods and will be able to apply these methods to solve some common problems in Mathematics, Statistics and to handle big data.

Learning Objectives:

After completion of the course, students will be able to

- Differentiate between Frequentist and Bayes statistics
- Differentiate between simple and Markov Chain Monte Carlo methods
- Write appropriate programs language to apply Monte Carlo methods to solve some problems of data science
- Understand some of algorithms of Marko Chain Monte Carlo methods

Course Contents:

Unit 1: Approaches for Statistical Inference [6 Hrs.]

Introduction, Motivating vignettes, Defining the approaches, Bayes vs frequentist approach , Some basic Bayesian models

Unit 2: The Bayes Approach [6 Hrs.]

Introduction, Prior distributions, Bayesian inference, Hierarchical modeling, Model assessment

Unit 3: Monte Carlo Methods [13 Hrs.]

Introduction, Motivating examples, Random numbers, Pseudorandom number generators, Random walks, Markov processes; Simple, Importance and Rejection sampling

Fundamental concepts of transformation, reweighting; Monte Carlo integration, estimation of pie by using programing code in R /Python/Fortran /C (Note: Students are supposed to design algorithm and code for such simple problems in practical classes)

Unit 4: Markov Chain Monte Carlo [6 Hrs.]

Introduction, Definition and transition probabilities, Decomposition of the state space, Stationary distributions, Limiting theorems, Reversible chains, Continuous state space

Unit 5: Gibbs Sampling (Algorithms) [10 Hrs.]

Introduction, Definition and properties, implementation and optimization, forming the sample, scanning strategies, Using the sample, Reparametrization, convergence diagnostics: Rate of convergence, Informal convergence monitaors, convergence prescription, formal convergence methods, Applications: Hierarchical models, Dynamics models, spatial models, MCMC Code (Students should be able to write simple MCMC code in python/R/Fortran/C)

Unit 6: Metropolis-Hastings Algorithms [7 Hrs.]

Introduction, Definition and properties, Special cases, Hybrid Algorithms, Applications – case studies and examples

Practical Works:

The practical work consist of Applying Monte Carlo and Markov Chain Monte Carlo methods to solve different problems from Text books writing codes in Python/R/Frotran/C (any one).

References:

1. Dani G., & Lopes H.F. (2006). *Markov Chain Monte Carlo Stochastic Simulation for Bayesian Inference*, 2nd edition, CRC Boca Raton FL: Chapman & Hall
2. Carlin B.P., & Louis T.A. (2009). *Bayesian Methods for Data Analysis* 3rd edition, CRC Boca Raton FL: Chapman & Hall
3. Gilks W.R., Richardson S. & Spiegelhalter D.J. (1996). *Markov Chain Monte Carlo in practice*, CRC Boca Raton FL: Chapman & Hall
4. Kendall W.S., Liang F. & Wang J.-S. (2005). *MARHOV CHAIN MONTE CARLO Innovations and Applications*, Singapore World Scientific
5. Anagnostopoulos K. N. (2014). *Computational Physics*, National Technical University of Athens, Greece

●●●●●

Tribhuvan University

Institute of Science and Technology

SCHOOL OF MATHEMATICAL SCIENCES

Syllabus

Master's in Data Sciences (MDS)- FOURTH SEMESTER

Compulsory Courses

Course Code	Course Titles	Credits	Nature
MDS 651	Data Visualization	3	Th.
MDS 652	Capstone Project / Thesis	6	Project + Report

Elective Courses (Any Two)

Course Code	Course Titles	Credits	Nature
MDS 653	Social Network Analysis	3	Th.+ Pr.
MDS 654	Actuarial Data Analysis	3	Th.+ Pr.
MDS 655	Deep Learning	3	Th.+ Pr.
MDS 656	Business Analytics	3	Th.+ Pr.
MDS 657	Bioinformatics	3	Th.+ Pr.
MDS 658	Economic Analysis	3	Th.+ Pr.

Code No.: MDS 651
Course Title: Data Visualization **Full Marks: 75**
Nature: Theory (Compulsory) **Credit: 3**

***Course Description*:**

This course presents comprehensive introduction to several topics on data visualization and its application. It provides the board overview of techniques of the visualization process, detailed view of visual perception, the visualized data and the actual visualization, interaction and distorting techniques.

.

Learning Objectives:

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

- Explain the concept of visualization in the processing and analysis of data.
- Understand and apply visualization models
- Implement attribute and spatial data visualization and applications
- Implement text and document visualization
- . Evaluate visualization techniques and explain its isuues

***Course Contents*:**

Unit 1: Introduction [6 Hrs.]

Introduction of visual perception, Visual representation of data, Data abstraction, Visual encodings, Use of color, Perceptual issues, Information overloads

Unit 2: Visual Representations [6 Hrs.]

Visualization reference model, Visual mapping, Visual analytics, Design of visualization applications.

Unit 3: Attribute Data Visualization [12 Hrs.]

Visualization of one, two and multi-dimensional data, Tabular data, quantitative values (scatter plot), Separate, Order and align (Bar, staked bar, dots and line charts), Tree data, Displaying Hierarchical structures, Graph data, Rules for graph drawing and labeling, Time series data, Characteristics of time data, Visualization time series data, Mapping of time.

Unit 4: Spatial Data Visualization [8 Hrs.]

Scalar fields**,** Isocontours **(**Topographic Terrain maps), scalar volumes**,** Direct volume Rendering (Multidimensional transfer functions), Maps (dot, pixel), vector fields
Defining marks and channels

Unit 5: Text and Document Visualization [8 Hrs.]

Text and document data, Levels of text representation, The vector space model, Visualizations of a single text document, Word cloud, Word tree, Text arc, Themescapes and self organiging maps

Unit 6: Evaluating Visualization Techniques and Issues [8 Hrs.]

User and data characteristics, Visualization characteristics, Structures for evaluating visualizations, Visualization bench marking, Issues of data, Issues of cognition, Perception and reasoning, Issues of hardware and software.

Practical Works:

The practical works include the techniques of the data visualization using software tools like MS spread sheet, Python, Matlab, Java, Tableau etc.

References:

1. Fry (2008). *Visualizing Data*. O'Reilly Media.
2. Ware (2012), *Information Visualization: Perception for Design*, Morgan Kaufmann.
3. Telea (2007). *Data Visualization: Principles and Practice*. A. K. Peters, Ltd.

●●●●●

Code No.: MDS 652

Course Title: Capstone Project / Thesis

Nature: Project + Report (Compulsory) **Credit: 6**

***Course Description*:**

A research project is an important element of Master's Degree Course in Data Science and is written up in the form of a Capstone Project. The project is often seen as the culmination of graduate work, and it is the formal scholarly work. It allows students to reflect and integrate their learning over their earlier semesters of study, and create a descriptive and original work in an area of their interest related to any area of Mathematics, Statistics, Computer Science and Information Technology approved by the Research Committee of SMS TU. While similar in some ways to a thesis, capstone projects may take a wide variety of forms, but most are long-term investigative projects that culminate in a final product, presentation, or performance. For example, students may be asked to select a topic, profession, or problem that interests them, conduct research on the subject, maintain a portfolio of findings or results, create a final product demonstrating their learning acquisition or conclusions and give an oral presentation on the project to a panel of teachers, experts, and practitioners who collectively evaluate its quality.

Learning Objectives:

The aims of the PROJECT are to:

- Put into practice theories and concepts learned on the program;
- Provide an opportunity to study a particular topic in depth through project ;
- Show evidence of independent investigation -data analysis, writing and presentation, and critical analysis;
- Combine relevant theories and suggest alternatives relating the project works ;
- Enable interaction with practitioners (where appropriate to the chosen topic);
- Show evidence of ability to plan and manage a project within deadlines

Learning Outcomes:

After the completion of their a Capstone Project , students should be able to:

- Define, design and deliver an academically rigorous piece of research;
- Understand the relationships between the theoretical concepts taught in class and their application in specific situations;
- Show evidence of a critical and holistic knowledge and have a deeper understanding of their chosen subject area;
- Appreciate practical implications and constraints of the specialist subject;
- Understand the process and decisions to be made in managing a project within strict deadlines.

Capstone Project Activities and Contents

The Capstone Project should focus on a well-formulated research question and show the student's capacity to conduct independent and comprehensive analysis of the subject, taking into account the relevant literature. The following activities and steps will be involved in the project report writing:

- Selecting a relevant topic or issue for the study;
- Getting approval of the Research Committee to pursue the proposed study;
- Locating the relevant literature;
- Locating the sources of data and information;
- Extracting the relevant information from different sources;
- Organizing and analyzing the data;
- Drawing conclusions; and
- Writing a Capstone Project report.

For detail information in writing the Capstone Project Report, see the guidelines

(**THE PROJECT REPORT: BASIC GUIDELINES FOR STUDENTS, 2020)**

●●●●●

School of Mathematical Sciences

Master's Degree in Data Science

TRIBHUVAN UNIVERSITY

THE CAPSTONE PROJECT REPORT/THESIS:
BASIC GUIDELINES FOR STUDENTS

Office of the Director
School of Mathematical Sciences
Institute of Science and Technology
Tribhuvan University
Kirtipur

2020

TABLE OF CONTENTS

Section 1

NATURE OF THE CAPSTONE PROJECT

INTRODUCTION

These guidelines are prepared and designed to help Master's Degree in Data Science students of School of the Mathematical Sciences, Tribhuvan University (SMSTU) in the preparation of their Thesis report. The guidelines address only matters of format and presentation, such as arrangement of content, paper, spacing, headings, data analysis, interpretation and referencing. It is the responsibility of each student to ensure that his or her work conforms to the guidelines set put below. The final Capstone Project is approved by the concerned supervisor and the Research Committee of SMSTU with regards to questions of quality and content.

OBJECTIVES OF THE CAPSTONE PROJECT

The Project is an integral part of the postgraduate studies at SMSTU. Towards the end of your study at SMSTU, you are required to undertake a research assignment and prepare an integrative research report in any areas of Mathematics, Statistics, Computer Science, Information Technology and other related Industries as approved by SMSTU.

The Project explores science, engineering, mathematic, technology and business questions as they relate to data available for many enterprises is increasing exponentially. It focuses on findings in related research plus the methodological alternatives. The project thus involves conceptualizing, planning, implementing, and writing up report, which extends knowledge in the subject area under investigation.

This assignment specifically aims to develop knowledge, skills and abilities necessary for conduct of individual research at a level which will make a distinct contribution to knowledge. You are expected to demonstrate the use of appropriate research, methodology, and written skills through the preparation and presentation of a substantial investigation.

Specifically, the objectives of the PROJECT include:

- To provide an opportunity for the students to integrate classroom knowledge and practice.
- To enable graduate students to do an independent study to reflect a creative endeavor that can make a significant contribution to knowledge in a given field.
- To develop students' ability to read professional literature, reports, and other works critically in their design, treatment of data, and conclusions.
- To strengthen the ability of students in presenting their research in building and evaluating mathematical models, exploring them computationally, and analyzing enormous amounts of observed and computed data.

ACTIVITIES INVOLVED IN THE CAPSTONE PROJECT

The following activities will be involved in the Project:

- Selecting a relevant topic or issue for study.
- Locating the relevant literature.
- Locating the sources of data and information.
- Extracting the relevant information from these sources.
- Identifying the various dimensions of the problem or issues.
- Organizing and analyzing the data effectively.
- Drawing inferences and conclusions
- Writing a report.

STUDY REQUIREMENTS

Your PROJECT report shall comply with the following requirements:

- The proposed field of study or topic of project related research must be approved by the concerned supervisor and the Director of SMSTU.
- The work must comply with any requirements advised by the concerned supervisor.
- The work in a PROJECT must reach a satisfactory standard of expression and presentation.
- You must maintain close and regular contact with your supervisor and the Director of SMSTU.

REPORTING REQUIREMENTS

You shall prepare a PROJECT report embodying the results of your research. The report submitted by you shall:

- Be an accurate account of research.
- Relate to the approved research topic.
- Not include work which has been submitted for any other academic award.
- Be written in English.
- Achieve a satisfactory standard of expression and presentation.
- Acknowledge any substantial assistance provided to you during the conduct of research and writing the report
- Conform to the rules and format of SMSTU for the preparation of the CAPSTONE PROJECT report.

LENGTH OF THE PROJECT REPORT

The length of the PROJECT report shall be around 15,000 – 18,000 words (approximately 70-75 pages). This length is exclusive of the materials included in appendices.

MULTIPLE COPIES REQUIRED

i. You shall submit two loose-bound copies of PROJECT report to SMSTU for evaluation. After satisfactory completing all recommended corrections with final *viva-voce*, you shall submit three hard-bound copies and one electronic copy. Each hard bound copy shall be bound black.

ii. The duplicate copies of the original are to be produced using a method which gives a clear and permanent copy (Laser Copy). The use of spirit duplication or carbon copy is not acceptable.

EVALUATION OF THE CAPSTONE PROJECT REPORT

The PROJECT report shall be evaluated by two examiners, one of whom shall be the faculty member of SMSTU. The internal examiner shall be appointed by the Director of SMSTU and the external examiner by the Examination Section, Office of the Dean, Institute of Science and Technology in consultation with Director of SMSTU.

You will be required to attend the *viva-voce* examination and defend your work. You will also be required to give a seminar presentation of your report as organized by SMSTU. The weightage given for viva and the PROJECT report will be 40% (Viva) and 60% (Report) respectively.

Section 2

CAPSTONE PROJECT PROPOSAL

- Having to prepare the proposal for your PROJECT is part of your graduate training. It is an opportunity to organize your thoughts about your research topic, to decide how you will pursue the work, and to spell out what resources (financial, material, and technical) you need to carry out the research. A PROJECT proposal seeks to convince the supervisor(s) or the Graduate Research Committee (GRC) that the research project is feasible.
- The writing of a project proposal is an exercise that you will repeat many times in your professional career. The ability to "sell" a project convincingly is a crucial part of your toolbox of skills.
- A clearly defined research problem (or question) is central to the success of a research project. It helps you to determine that your project is doable before you begin writing the PROJECT report.
- The proposal should explicitly state the problem being addressed or gap in knowledge to be filled, describe the objectives and research techniques to be employed, and include a review of the principal relevant published literature.
- The proposal needs a thread of logic. It should build from a statement of the research problem or gap in knowledge, and follow an outline of detailed objectives that must be achieved (or questions that must be answered) if the problem is to be solved. The presentation of methodology should be clearly connected to stated objectives.
- A PROJECT proposal usually contains some formulations of the following sections:
    - Background of the PROJECT Study
    - Statement of the research problem
    - Research objectives
    - Research questions or hypotheses
    - Rationale for the study
    - Related literature
    - Theoretical/ conceptual framework
    - Methodological design – a plan outlining how and when each step of the project will be done.
    - References
- Maximum length of the proposal should not exceed 2,500 words (excluding references, figures or tables).
- A CAPSTONE PROJECT proposal shall be reviewed by the GRC of SMSTU. The GRC may either (a) accept the proposal as written, or (b) return the proposal for revision.
- Although you are expected to seek guidance from your supervisor in the choice of topic and the method of solving the problem involved, the responsibility for the proposal writing lies with you. You will, as far as possible, work independently and demonstrate the ability to plan outline an acceptable research project.
- Upon approval of the project proposal by the Research Committee, you will be enrolled in the GRC of your department.

Section 3

PROJECT REPORTING FORMAT

As a student of Master's Degree in Data Science at SMSTU, you are required to write reports for different project assignments in different courses. However, the Project report writing is different from other reports. It is an organized, issues-focused, evidence and data based and creative piece of academic writing. Hence the following structure and guidelines have to be followed while preparing your project report.

STRUCTURE OF THE CAPSTONE PROJECT REPORT

You must carefully read your course information details to ensure that you comply with what instructor stipulates. A PROJECT report is typically made up of three main divisions: (1) preliminary, (2) body, and (3) supplementary. Each of the sections contains different kind of contents.

Preliminary Materials

Title page of the PROJECT
Certification
Deceleration of Authenticity
Table of Contents
List of Tables and Figures
Common Abbreviations Used
Executive Summary

Body of the Report

Chapter I	Introduction
Chapter II	Related Literature and Theoretical Framework
Chapter III	Research Methods
Chapter IV	Analysis and Results
Chapter V	Discussion, Conclusions and Implications

Supplementary Materials

References or Bibliography
Appendices

BODY OF THE PROJECT REPORT

- **Introduction** – Background information on the topic so that you are able to 'place' your research in the context. Details are given of your problem statement, objectives, hypothesis, scope and significance, definition of terms, limitations and an outline of the structure of the PROJECT report

- **Related Literature and Theoretical Framework** – You must carefully structure your findings of the literature survey. It may be useful to do a chronological format where you discuss from the earliest to the latest research, placing your research appropriately in the chronology. Alternately, you could write in a thematic way, outlining the various themes that you discovered in the research regarding the topic. Again, you will need to state where your research fits. Finally, at the end of this chapter, you present your theoretical framework, briefly explaining the measurement of variables.

- **Research Methods –** You have to clearly outline what methodology you used in your research i.e. what you did and how you did it. It must be clearly written so that it would be easy for another researcher to duplicate your research if they wished to. The

contents of this chapter may include research design, population and sample, instrumentation, sources and methods of data collection, and data analysis.

- It is usually written in a 'passive' voice (e.g. the participants were asked to fill in the questionnaire attached in Appendix 1) rather than an 'active' voice (e.g. I asked the participants to fill in the questionnaire attached in Appendix 1).
- Clearly reference any material you have used from other sources. Clearly label and number any diagrams, charts, and graphs. Ensure that they are relevant to the research and add substance to the text rather than just duplicating what you have said. You do not include or discuss the results here.

- **Analysis and Results –** Data are analyzed statistically with mathematical models and results are presented and interpreted. Hypotheses are tested. This is where you indicate what you found in your research. You give the results of your research. Based on these results you give your interpretation. You also discuss the relevance of your results and how your findings fit with other research in the area. It will relate back to your literature review and your introductory project problem statement.

- **Discussion, Conclusions and Implications**

  - **Discussion of the Findings** – This section is the most important section of your project report. Make sure that you allocate enough time and space for a good discussion. This is your opportunity to show that you have understood the significance of your findings and that you are capable of applying theory in an independent manner. The discussion will consist of argumentation. In other words, you investigate a phenomenon from several different perspectives. To discuss means to a question your findings, and to consider different interpretations.

  - **Conclusions and Implications –** This includes key facts from your research findings to help explain your results as needed: you have to summarize, compare and evaluate your research results in context of existing theories, and make comments about its success and effectiveness.

    An implication refers to something which is implied or suggested as naturally being inferred or understood in a certain policy or practice. Your research needs to identify why and how the analyses and interpretations were made and the way key concepts in the analyses evolved. In addition, you need to inform the reader of any unexpected findings or patters that emerged from the data and report a range of evidence to support assertions or interpretations presented.
    You could also indicate some areas where your research has limits and where further research would be useful. Implications of the research for furthering understanding of the research problem need to be explored.

Section 4

LANGUAGE, TYPING, EDITING AND FORMATTING

KEEP THE LANGUAGE CLEAR AND STRAIGHTFORWARD

- Avoid jargon and acronyms, especially those terms which might be common within your field of work but might be unknown to the general public.
- Use clear and concise writing, and include charts and graphs where appropriate.
- Keep sentences and paragraphs short.
- Delete unnecessary words and phrases.
- Use active verbs as much as possible.

MAINTAIN OBJECTIVITY

- Do not use emotionally charged language when describing your findings, like "very" or "extremely". This can make you sound like a program advocate, thus reducing your objectivity and credibility.
- Use disappointing results to guide recommendations for enhancing services or addressing implementation barriers, rather than dismissing or hiding them.
- Discuss limitations in terms of how information was collected, so that audiences can judge the degree of confidence to place in the results. Every evaluation study has limitations, and it is important to know what they are so stakeholders can consider the findings in context.

TYPING AND EDITING

- The PROJECT report shall be a typescript paper document. It shall not only be submitted in an electronic format.
- The PROJECT report is to be typed on ISO A4 size white bond paper. If diagrams, maps, tables and similar presentations do not fit readily on this sheet size, ISO B4 size may be used. B4 size pages are to be folded and bound so as to open out at the top and the right.
- Typing is to be done on one side of each sheet only, with pages numbered consecutively throughout the report. The following minimal margins are to be observed:

    - Left 3.5cm
    - Top 2.5cm
    - Bottom 2cm
    - Right 2cm

- The PROJECT report must be 1.5-spaced. Single spacing may be used only in the table of contents, footnotes and endnotes, charts, graphs, tables, quotations, appendices, and references.
- Text material should be typed on one side of the paper. The manuscript is to be neat in appearance and without error.
- Typing should be done using the Times New Roman and font size of 12 or equivalent, except for text in the tables.
- SMS TU expects a high standard of editing of the work submitted to it for examination.

FORMATTING

- PROJECT Report writing format should follow the Latest version of APA styles of citation and references.
- Except for text in the tables, all other text must always be justified.

APPENDIX –A

TITLE OF THE CAPSTONE PROJECT/THESIS REPORT

BY

Candidate's Full Name
(Roll. No./Registration no.)

A Project Report Submitted to in partial fulfillment of the requirements for the degree of
Master's Degree in Data Science

at the

School of Mathematical Sciences
Institute of Science and Technology
Tribhuvan University

Kirtipur
September , 2020

APPENDIX – B

CERTIFICATION

We, the undersigned certify that we have read and hereby recommend for the acceptance by the School of Mathematical Sciences, Tribhuvan University, a PROJECT /THESIS report submitted by
…… entitled
……, in a partial fulfillment of the requirements for the award of Master's Degree in Data Sciences of Tribhuvan University.

_____
PROJECT/Thesis Supervisor
Signature

_____
External Examiner
Signature

_____
GRC chairman
Signature

_____
Director, SMSTU
Date: _____

APPENDIX – C

DECLARATION OF AUTHENTICITY

I, ……., declare that this PROJECT REPORT /THESIS is my own original work and that it had fully and specifically acknowledged wherever adapted from other sources. I also understand that if at any time it is shown that I have significantly misinterpreted material presented to SMSTU, any credits awarded to me on the basis of that material may be revoked.

Signature: _____
Name:

Date:

Code No.: MDS 653
Course Title: Social Network Analysis **Full Marks: 75**
Nature: Theory +Practical (Elective) **Credit: 3**

***Course Description*:**

This course covers different concepts of social network analysis including basics, data collection, descriptive methods, and inferential methods of social network analysis. This course also covers different applications of social network analysis.

***Learning Objectives*:**

After completion of this course, students should be able to

- Understand different concepts, foundations, data collection related to social network analysis
- Know and use different methods in social network analysis
- Identify and study different applications of social network analysis

Course Contents:

Unit 1: Basics of Social Network Analysis [8 Hrs.]

Introduction , Social network and its representation, Types of networks , Network parts and levels of analysis, Network as social structure and institution, Causality in social network studies, History of social network analysis

Unit 2: Data Collection [5 Hrs.]

Boundary specification, Data collection process, Informal bias and issue of reliability, Archival data

Unit 3: Descriptive Methods in Social Network Analysis [11 Hrs.]

Graph and matrix for social network representation, Density , Centrality, centralization, and prestige, Cliques, Multidimensional scaling (MDS) and dendogram, Structural equivalence, Two-mode networks and bipartite matrix

Unit 4: **Inferential Methods in Social Network Analysis [6 Hrs.]**

Permutation and QAP (Quadratic Assignment Procedure) Correlation; P* or Exponential Random Graph Model (ERGM)

Unit 5: Social Network Analysis of Work and Organizations [6 Hrs.]

Personal connections and labour market processes, Intra-organizational networks , Inter-organizational relations

Unit 6: Social Network Analysis in Crime and Terrorism [6 Hrs.]

Personal networks, delinquency, and crime; Neighbourhood networks, Criminal networks

Unit 7: Social Network Analysis in Emotional and Physical Health [6 Hrs.]

Social network Analysis in physical fitness, Social network analysis and Illicit drug use, Social network analysis and Sexually transmitted disease

Laboratory Work:

Students should implement and realize social network analysis using appropriate tools and programming language.

References:

1. Song Yang, Franziska B Keller, Lu Zheng (2016). *Social Network Analysis: Methods and Examples*, Sage Publications (Verlag),
2. Krishna Raj P.M., Ankith Mohan, Srinivasa K.G. (2018). *Practical Social Network Analysis with Python*, Springer.
3. Mehmet Kaya, Jalal Kawash, Suheil Khoury, Min-Yuh Day (2018). *Social Network Based Big Data Analysis and Applications*, Springer International Publishing.
4. Reda Alhajj, Jon Rokne (2018). *Encyclopedia of Social Network Analysis and Mining*, Springer New York.
5. Xiaoming Fu, Jar-Der Luo, Margarete Boos (2017). *Social Network Analysis: Interdisciplinary Approaches and Case Studies*, CRC Press.

●●●●●

Code No.: MDS 654
Course Title: Actuarial Data Analysis **Full Marks: 75**
Nature: Theory +Practical (Elective) **Credit: 3**

***Course Description*:**

In this course students will learn the fundamental concept of Actuarial Data Analysis and they will able to apply necessary tools and techniques to model the actuarial data by using R. The goal of the course is to prepare the student to examine quality of provided data and fitting appropriate predictive model such as generalized linear models, decision trees, Survival Analysis in insurance. Moreover, in this course students will get idea how insurance company apply predictive modeling techniques to solve the business problems.

***Learning Objectives*:**

After completion of the course, students will be able to

- Explore graphically and summaries time series data using R.
- Understand the different types of predictive modeling problems.
- Create a variety of graphs using the ggplot2 package.
- Select and validate a GLM appropriately, classification problem.
- Select appropriate hyperparameters for regularized regression.
- Construct regression and classification trees.
- Select appropriate hyperparameters for decision trees and related techniques.
- Do survival analysis in insurance
- Simulate of modeles

***Course Contents*:**

Unit 1: Data Visualization **[5 Hrs.]**

Write effective graphs in RStudio, key principles of constructing graphs, create effective graphs in RStudio, graphs using the ggplot2 package.

Unit 2: Data Types and Exploration **[6 Hrs.]**

Work with various data types, principles of data design, construct a variety of common visualizations for exploring data, structured, unstructured, and semi-structured data. Methods of handling missing data, univariate and bivariate data exploration techniques, effective data design with respect to time frame, sampling, and granularity.

Unit 3: Data Issues and Resolutions **[7 Hrs.]**

Evaluate data quality, Resolve data issues, Identify regulatory, Ethical issues, Data sources, Outliers and other data issues, Non-linear relationships via transformations, Regulations, standards, and ethics surrounding predictive modeling and data collection.

Unit 4: Generalized Linear Models **[10 Hrs.]**

Implement ordinary least squares regression in R and understand model assumptions, GLM model assumptions, interpret model coefficients, interaction terms, offsets, and

weights, validate a GLM, bias, variance, model complexity, the bias-variance trade-off, hyperparameters for regularized regression.

Unit 5: Decision Trees **[10 Hrs.]**

The basics of decision trees, motivation behind decision trees, regression trees, classification trees, trees versus linear models, bagging & random forests to improve accuracy, boosting to improve accuracy, fitting classification trees, fitting regression trees, selecting appropriate hyperparameters for decision trees and related techniques.

Unit 6: Survival Analysis in Insurance **[10 Hrs.]**

Kaplan-Meier and Nelson-Aalen estimator, Log-rank test, Parametric regression models, Cox regression models, Simulation: Model comparison, Accelerated failure Time models.

Practical Works:

The practical works consist of lab work using R/RStudio software. Project work on predictive modeling techniques to solve (business) problems must be highlighted.

References:

1. Edward W. Frees (2010). *Regression Modeling with Actuarial and Financial Applications*. New York:Cambridge.
2. James, Witten, Hastie, Tibshirani (2013). *An Introduction to Statistical Learning, with Applications in R*. New York: Springer.
3. Healy E. (2019). *Data Visualization: A Practical Introduction*. Princeton University Press.
4. Lander (2017). *R for Everyone: Advanced Analytics and Graphics*, 2nd ed. Boston: Addison-Wesley.
5. Ambrose L. (2019). *ACTEX Study Manual for SOA Exam Predictive Analytics*. USA.
6. D. C. M. Dickson, M. R. Hardy and H. R. Waters (2009) *Actuarial Mathematics for Life contingtent Risks*, Cambridge University Press, New York.

●●●●●

Code No.: MDS 655
Course Title: Deep Learning **Full Marks: 75**
Nature: Theory + Practical (Elective) **Credit: 3**

***Course Description*:**

The course is designed to provide the students of Graduate level with the fundamental concepts of Deep Learning. The course is divided into two parts: first one is the basic foundation with the introduction to Machine Learning basics, The neural network, Training feed-forward neural networks and beyond gradient descent; and the second one deals with convolutional neural networks and models for sequence analysis.

***Learning Objectives*:**

The major objective of the course is to prepare the students to design Deep Learning models/solutions to solve the real-world problems. Students will familiarize with the Deep Learning libraries and tools like Keras, Tensor Flow, and Pytorch thus learning to understand the different theoretical and implementation aspects of Deep Learning models.

***Course Contents*:**

Unit 1: Machine Learning Basics **[4 Hrs.]**

Learning algorithms, Capacity, overfitting and underfitting, Hyper parameters and validation sets, Estimators, bias and variance , Maximum likelihood estimation, Bayesian statistics, Supervised learning algorithms, Unsupervised learning, Algorithms, Stochastic gradient descent, Building a machine learning algorithm

Unit 2: The Neural Network **[4 Hrs.]**

Building intelligent machines, The limits of traditional computer programs, The mechanics of machine learning, The neuron, Expressing linear perceptrons as neurons; Feed-forward neural networks, Linear neurons and their limitations, Sigmoid, Tanh, and ReLU neurons, Softmax output layers.

Unit 3: Training Feed-Forward Neural Networks **[4 Hrs.]**

The Fast-Food problem, Gradient descent, The delta rule and learning rates, Gradient descent with sigmoidal neurons, The Back propagation algorithm, Stochastic and mini batch gradient descent, Test sets, validation sets, and over fitting; Preventing over fitting in deep neural networks

Unit 4: Beyond Gradient Descent **[8 Hrs.]**

The challenges with gradient descent, Local minima in the error surfaces of deep networks , Model identifiability ,Pesky in deep networks, Flat regions in the error surface, Gradient points, Momentum-based optimizations, A brief view of second-order methods, Learning rate adaptation, AdaGrad – Accumulating historical gradients, RMSProp – Exponentially weighted moving average of gradients, Adam – Combining momentum and RMSProp

Unit 5: Convolutional Neural Networks [8 Hrs.]

Neurons in human vision, The shortcomings of feature selection, Vanilla deep neural networks don't scale, Filters and feature maps, Full description of the convolution layer , Max pooling, Full architectural description of convolution networks, Closing the loop on MNIST with convolutional networks, Image preprocessing pipelines Enable more Robust models, Accelerating training with batch normalization, Building a convolutional network for CIFAR-10,
Visualizing learning in convolutional, Networks, Leveraging convolutional filters to replicate artistic styles, Learning convolutional filters for other problem domains

Unit 6: Embedding and Representation Learning [10 Hrs.]

Learning lower-dimensional representations, Principal component analysis,
Motivating the autoencoder architecture, Implementing an autoencoder in tensor flow , Denoising to Force Robust Representations; Sparsity in Autoencoders
Context and input vector, The Word2Vec framework, Implementing the Skip-Gram architecture

Unit 7: Models for Sequence Analysis [10 Hrs.]

Analyzing variable-length inputs,Tackling seq2seq with neural N-Grams,
Implementing a part-of-speech tagger, Dependency parsing and syntax net,
Beam search and global normalization, A case for stateful deep learning models,
Recurrent neural networks, The challenges with vanishing gradients,
Long Short-Term Memory (LSTM) units, Tensor flow primitives for RNN models,
Implementing a sentiment analysis model, Solving Seq2Seq tasks with recurrent neural networks, Augmenting recurrent networks with attention, Dissecting a neural translation network.

Laboratory Works:

The lab works will introduce the students with the Deep Learning architecture and framework along with the different learning algorithms. Lab work would give students a hands-on knowledge on the different libraries, modules and frameworks of Deep Learning.

References:

1. Ian Goodfellow, Yoshua Bengio, Aaron Courville (2016). *Deep Learning*, MIT Press.
2. Nikhil Buduma and Nicholas Lacascio (2017). *Fundamentals of Deep Learning – Designing Next-Generation Machine Intelligence Algorithms*, O'Reilly.
3. Sudarsan Ravichaniran (2019). *Hands-On Deep Learning Algorithms with Python*, Packt Publishing.
4. Francois Chollet(2018). *Deep Learning with Python*. Manning Publications.
5. Josh Patterson & Adam Gibson(2017). *Deep Learning – A Practitioner's Approach*, O'Reilly.
6. Michael Nielson. *Neural Networks and Deep Learning*. [http://neuralnetworksanddeeplearning.com]
7. Nikhil Ketkar (2017). *Deep Learning with Python – A Hands-on Introduction*, APress.

●●●●●

Code No. : MDS 656
Course Title: Business Analytics **Full Marks: 75**
Nature: Theory + Practical (Elective) **Credit: 3**

***Course Description*:**

This course provides an overview of business analytics from both models and strategy perspectives. It mainly focuses multi-criteria decision making, performance analytics and people analytics. The course comprises particularly the widely used tools in business analytics: AHP and DEA.

Learning Objectives:

After the completion of the course, students will be able to

- Identify the existing business analytics models in real world
- Link between the strategy at functional level and business analytics
- Solve the complex multi-criteria decision problems
- Use optimization tools and techniques in performance analysis
- Link people analytics with overall organizations' mission

Course Contents:

Unit 1: **Introduction to Business Analytics** **[4 Hrs.]**

Meaning of business analytics, Evolution of business analytics, Scope of business analytics, Data for business analytics, Models in business analytics, Problem solving with analytics

Unit 2: Business Analytics at Strategy level **[8 Hrs.]**

Link between strategy and the deployment of business analytics, Strategy and business analytics: No formal link between strategy and business analytics, Business analytics supports strategy at a functional level, Dialogue between the strategy and business analytics functions, Information as a strategic resource ; Information prioritization: The product and innovation perspective, Customer relations perspectives , Analyst's role in the business analytic model

Unit 3: Multi-criteria Decision Making **[12 Hrs.]**

Analytic hierarchy process (AHP), Establishing priorities using AHP: Pairwise comparisons, Pair wise comparison matrix, synthesization, consistency; Overall priority ranking, Application of AHP in various areas

Unit 4: Performance Analytics **[12 Hrs.]**

Basic concept of efficiency measurement, Frontier analysis**,** Mathematical programming aspects of data envelopment analysis (DEA) ; output maximization and Input minimization DEA programs, Charnes, Cooper and Rhodes (CCR) model, Banker, Charnes and Cooper (BCC) model.

Unit 5: People Analytics **[12 Hrs.]**

Introduction, Linking people analytics with overall business plan, Workforce planning analytics, Employee engagement surveys, Big data and people analytics , HR information system

Practical Works:

Use of Microsoft Excel and related software to solve the problem

Research project to carry on the on-hand experience in real world phenomena

References:

1. Anderson D.R. Sweeney D.J., Williams T.A. & Martin K. (2011). *An introduction to Management Science Quantitative Approaches to Decision Making*. Delhi: Cengage Learning India Private Limited.
2. Gert H.N. Laursen & Jesper Thorlund. (2017). *Business Analytics for Managers*. New Jersey: John Wiley & Sons Inc. Retrieved from https://b-ok.asia/book/2871501/a24d0a
3. James R. Evan . (2017). *Business Analytics Methods, Models and Decisions*. England: Pearson Education Limited
4. Jean Paul Isson & Jesse S. Harriott .(2016). *People Analytics in the Era of Big Data* . New Jersey: John Wiley & Sons Inc. Retrieved from https://b-ok.asia/book/2765101/ccbe6e
5. Ramanathan R. (2003). *An Introduction to Data Envelopment Analysis A Tool for Performance Measurement*. New Delhi: Sage Publications India Pvt. Ltd.

●●●●●

Code No.: MDS 657
Course Title: Bioinformatics **Full Marks: 75**
Nature: Theory +Practical (Elective) **Credit: 3**

***Course Description*:**

This course is an application of computational methods used in biological data for data analysis and discovery. It provides basic concepts on molecular biology, explores important biological databases and analysis of those data using various algorithms and available tools. It also encourages students to research advanced computer technology in order to solve complex biological problems.

***Learning Objectives*:**

Upon completion of the course students must be able to:

- Understand basic concepts of molecular biology and biological data.
- Know different biological databases and tools to retrieve and analyze the relevant data.
- Perform pairwise and multiple sequence alignments and construct phylogenetic trees.
- Predict structure and function of protein.
- Use computational tools for genomic data.
- Gain knowledge in drug discovery and emerging fields in bioinformatics.

Course Contents:

Unit 1: Introduction [8 Hrs.]

Overview, history, scope and application areas in bioinformatics, Major databases in bioinformatics, File formats, Molecular biology, Central dogma of molecular biology, Information search and data retrieval tools (Entrez, SRS), Data mining of biological data, Genome analysis, Genome mappings, Genome sequencing, Sequence assembly and annotation tools, Human Genome project.

Unit 2: Pairwise and Multiple Sequence Alignment [10 Hrs.]

Pairwise sequence alignment, Substitution matrices PAM & BLOSUM, Dot matrix method, Dynamic programming:- Local Alignment and Global Alignment, Heuristic Method:- FASTA, BLAST, comparisons, Multiple sequence alignment, Guide tree, Types of MSA, Applications.

Unit 3: Phylogenetic Analysis [7 Hrs.]

Introduction, Distances matrix, Types of trees, Tree construction methods.

Distance based method (UPGMA, Fitch Margolis method, Neighbor-Joining), Character base method (Parsimony and maximum likelihood method), Application of phylogenetic analysis.

Unit 4: Gene Prediction, Expression and Microarrays. [6 Hrs.]

Basics of gene prediction, Pattern recognition- gene prediction methods, Tools- gene expression, Working with DNA microarrays, Clustering gene expression profiles- data sources and tools for microarray analysis, Applications.

Unit 5: Protein Classification and Structure Prediction. [10 Hrs.]

Overview of protein structure, Protein structure visualization - structure based protein classification, Protein structure databases, Tools - protein structure alignment, Domain architecture databases, Tools - Protein classification approaches - Protein identification and characterization - Primary structure analysis and Prediction- Secondary structure analysis and Prediction - Motifs, profiles, patterns, fingerprints search - Methods of sequence based protein prediction, 2D and 3D structure prediction, Protein function prediction.

Unit 6: Drug Discovery and Current Advancement in Bioinformatics. [7 Hrs.]

Areas influencing drug discovery, Pharmacogenetics, Applications , Parameters in drug discovery - drug discovery technologies, Target discovery strategy, Precision medicine, Machine learning, Artificial informatics,Chemoinformatics, Immuno informatics.

Practical Works:

The practical work includes the important aspects of biological databases and their queries. Sequence analysis and alignment using Basic Alignment Search Tool (BLAST). Tools used for multiple sequence alignment and phylogenetic analysis, gene prediction, secondary structure and tertiary structure predictions are also included.

References:

1. Attwood T.K. and Parry-Smith.(1999). *Introduction to Bioinformatics*, Addison Wesley Longman.
2. David W Mount, Bioinformatics.(2004). *Sequence and Genome Analysis*, 2nd edition, CBS publishers.
3. Arun Jagota(2001).*Data Analysis and Classification for Bioinformatics*, Pine Press.
4. Des Higgins and Willie Taylor.(2000). *Bioinformatics Sequence, Structures and Databanks*, Oxford University Press, USA.
5. Durbin R., Eddy S., Krogh A.& Mitchison G.(1999). *Biological Sequence Analysis*
6. Dov Stekel(2003*). Microarray bioinformatics*, Cambridge.
7. Des Higgins, Willie Taylor.(2010). *Bioinformatics Sequence Structure & Data Banks. A practical approach*.

●●●●●

Code No.: MDS 658
Paper: Economic Analysis **Full Marks: 75**
Nature: Theory +Practical (Elective) **Credit: 3**

***Course Description*:**

In this course students will learn the methods and tools of economic analysis for giving advanced knowledge of economic theory and suggesting ways to apply the knowledge in formulating and analyzing economic models and theories.

***Learning Objectives*:**

The objective of this course is to provide advanced knowledge on micro and macroeconomic analysis. Upon the completion of this course, students will be able to understand analytical tools and apply them in formulating and analyzing economic models and theories.

Course Contents:

Unit 1 : Theory of Consumer Behaviour [6 Hrs.]

Total and marginal utility, Consumer equilibrium, Indifference curves, The marginal rate of substitution, Characteristics of indifference curves, The budget constraint line, Consumer equilibrium, The price-consumption curve and the consumer's demand curve, Separation of the substitution and income effects, Theory of revealed preference, Consumer surplus and elasticity of demand, The problem of choice in situations involving risk and uncertainty (attitude towards risk and insurance).

Unit 2: Theory of Production and Costs [9 Hrs.]

Production with one variable input: Total, average, and marginal product. The shapes of the average and marginal product curves, Stages of production,
Production with two variable inputs: Isoquants, Short-Run total cost curves, The long-run average cost curve, The long-run marginal cost curve, The long-run total cost curve, The Cobb-Douglas production function.

Unit 3: Price-Output under Perfect Competition and Monopoly [9 Hrs.]

Perfect competition: Short-run and long-run equilibrium, Supply curves of the firm and industry, Dynamic changes and industry equilibrium.
Monopoly: Short-run and Long-run equilibrium, Predictions in dynamic changes, Regulated monopoly (Taxation, and price regulation), Govt. regulated monopoly, Discriminating monopoly. Comparison competitive and monopoly firms and excess capacity.

Unit 4: Price- Output under Monopolistic Competition and Oligopoly: Monopolistic Competitive Market: [9 Hrs.]

Product differentiation and demand curve, Industry and Group, Chamberlin's model: with entry and price competitions and equilibrium of firms, Comparison competitive and monopolistic competitive markets, Concept of excess capacity; Non-collusive oligopoly: Cournout's model, Bertrand's model, Chamberlin's model, Stackelberg's model and Kinked demand model of oligopoly.

Unit 5: Classical Theory of Output and Employment [6 Hrs.]

The Classical postulates; Say's law of market, Full employment – Demand for and Supply of labour; Labour supply and money wages; Unemployment and wage rigidity; Overall equilibrium in the basic static model (Goods, labour & money markets).

Unit 6: GDP, Growth, and Instability: Measuring Domestic Output and National Income [9 Hrs.]

Gross domestic product: A monetary measure / Avoiding multiple counting / GDP Excludes nonproduction transactions / Two ways of looking at GDP: Spending and Income, The expenditures approach: Personal consumption expenditures (*C*)/Gross private domestic investment (*I g*)/Government purchases (*G*)/Net exports (*X n*)/ Putting it all together: GDP *C _ Ig _ G _ X n,* The income approach: Compensation of employees/ Rents/ Interest *l* Proprietors' Income/ Corporate Profits/ Taxes on production and imports/ From national income to GDP, Other national accounts: Net domestic product/National income/ Personal income/ Disposable income / The circular flow revisited, Nominal GDP versus Real GDP: Adjustment process in a one-product economy/An alternative method / Real-world considerations and data.

Practical Works:

The practical work includes: The theory will be empirical test with the deductive logical approach (microeconomics and macroeconomics theory to real world).

References:

1. Dominick Salvatore (2006). *Managerial Economics in a Global Economy* (4th ed.). Thomson Publication.
2. Mark Hirschey and James L. Pappas (1992) *Fundamentals of Managerial Economics*-4th ed. A Harcourt Brace Javanovich College Publication.
3. Campbell R. McConnell; & Stanley L. Brue Sean M. Flynn (2009). *Economics: Principles, Problems, and Policies* -18th ed. McGraw-Hill/Irwin Publication.
4. Paula A. Samuelson; & William d. Nordhaus, (2010). *Principles of Economics*-19th ed. McGraw-Hill/Irwin Publication.
5. N. Gregory Mankiw (2013). *Principles of Economics* -17th ed. Cengage Learning Publication.

●●●●●