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.

***