Dr. Homi Bhabha State University
Mumbai

Proposed M. Sc. (Data Science) Syllabus

Semester I

With Effect from Academic Year 2023 - 2024

M.Sc. (Data Science) Semester I

Course	Course Code	Course Name	Credits	Teaching Scheme				Examination Scheme						
								Theory				Practical		
				Lecture Per Week	Practical per Week	Total Periods	Duration in Hours	Max. Marks (Theory)	Max. Marks (Internals)	Total Marks	Minimum Passing Marks	Max Marks	Total Marks	Minimum Passing Marks
DSC1	MSDSDC101T	Probability Theory	4	4	--	4	4	60	40	100	40	--	--	--
DSC2	MSDSDC102T	Programming with Python	4	4	--	4	4	60	40	100	40	--	--	--
DSE1	MSDSDE101T	DBMS and Data Warehousing	4	4	--	4	4	60	40	100	40	--	--	--
RM	MSDSRM101T	Research Methodology	4	4	--	4	4	60	40	100	40	--	--	--
DSC1	MSDSLB101P	Probability Theory LAB	2	--	8	8	4	--	--	--	--	50	50	20
DSC2	MSDSLB102P	Programming with Python LAB	2	--	8	8	4	--	--	--	--	50	50	20
DSE1	MSDSDE101P	DBMS and Warehousing LAB	2	--	8	8	4	--	--	--	--	50	50	20

MSDSDC101T Probability Theory

Course Credit: 06 (Credit: Th(4) + Pr(2)) **Total Contact Hours: 60 Hrs.**

Unit I: Random variables and their distributions (15 Hours)

Review of probability theory, Conditional probability and independence, Bayes theorem.

Discrete and Continuous probability distributions, Cumulative Distribution functions, Expectation and moments, Moment generating functions and characteristic functions and their applications, Probability and moment inequalities

Common statistical distributions: Binomial Distribution, Uniform Distribution, Poisson Distribution, Negative Binomial Distribution, Geometric Distribution, Continuous Uniform Distribution, Exponential Distribution, Normal Distribution, Log Normal Distribution, Gamma Distribution, Weibull Distribution, Pareto Distribution.

Unit II: Multiple random variables and distribution of Functions of random variables (15 Hours)

Multiple random variables and joint distributions, conditional distributions and independence, Covariance and correlation of random variables; Conditional moments

Transformations of discrete and continuous random variables and their distributions, Exact methods (using characteristic functions, transformation formula) and approximate methods (Taylor series method) for computing means and variance of functions of random variables, Cumulative distribution function techniques, moment generating function techniques.

Unit III: Sampling distributions (15 Hours)

Concept of population and sample, Theory of Simple random sampling with and without replacement

Introduction to Sampling distributions, distribution of sample mean, Sampling from normal distributions, Student's t distribution, Chi-square distribution, F distribution, Interrelations among t, chi-square and F distribution, Order statistics and their distributions.

Large sample properties of sample characteristics: Convergence in probability and convergence in distributions, weak law of large numbers and Central limit theorem

Unit IV: Elements of probabilistic simulation (15 Hours)

Probability integral transform, direct simulation of discrete and continuous univariate random variables

Computation of probability by simulation, Monte Carlo method of integration and computation of moments.

Simulations from bivariate distributions, Gibbs sampler and accept reject algorithm

References:

1. L. Wasserman, All of Statistics, Springer.
2. P.G. Hoel, S.C. Port and C.J. Stone, Introduction to Probability, Universal Book Stall, New Delhi.
3. Vijay K. Rohatgi and A. K. Md. Ehsanes Saleh, An Introduction to Probability and Statistics, John Wiley & Sons, Inc
4. A. M. Mood, F. A. Graybill and D. C. Boes, Introduction to The Theory of Statistics, Third Edition, Mc Graw Hill Education.
5. Sheldon M. Ross, Introduction to Probability Models
6. Anirban DasGupta, Probability and Statistics for Machine Learning: Fundamentals and Advanced Topics.
7. A. M. Gun, M. K. Gupta, B. Dasgupta, An Outline of Statistical Theory, Volume Two, World Press.

MSDSDC102T Programming with Python

Course Credit: 06 (Credit: Th(4) + Pr(2)) **Total Contact Hours: 60 Hrs.**

Unit I: Introduction to Python (15 Hours)

Introduction to Python programming, key words and identifiers, basic data types, operators, control flow, arrays, built in functions, user-defined functions,

Sorting and Searching.

Unit II: Object Oriented Programming with Python (15 Hours)

Classes and objects, methods, constructors, inheritance, polymorphism.

Unit III: Fundamental Data Structure (15 Hours)

Introduction- Algorithm Analysis, Finding Complexity. Fundamental data structures - List, Sorted Lists, Double Linked Lists, Stack & Queue application.

Unit III: Binary Trees (15 Hours)

Insertion and Deletion of nodes, Tree Traversals, Polish Notations, Red Black Trees, *B*-Trees, Heaps, Priority Queues.

References

(1) Clifford A Shaffer, Data Structures and Algorithm Analysis, Edition 3.2 (Java Version), 2011.

(2) Michael T. Goodrich, Roberto Tamassia, Michael H. Goldwasser. Data Structures And Algorithms In Java^{TM} Sixth Edition, Wiley Publishers, 2014.

(3) Ellis Horowitz, Fundamentals of Data Structures in C++, University Press, 2015.

(4) Ajay Agarwal, Data Structure through C, A Complete Reference Guide, Cyber Tech Publications, 2005.

(5) Python Data Science Handbook: Essential Tools for Working with Data by Jake VandarPlas, O'Rielly/Shroff Publication, 2016 Edition.

(6) Python for Data Science for Dummies by John Paul Mueller and Luca Massaron, Wiley Publication, Second Edition

(7) Python: The Complete Reference by Martin C. Brown, McGraw Hill Education, 2018 Edition

MSDSDE101T DBMS and Data Warehousing

Course Credit: 06 (Credit: Th(4) + Pr(2)) **Total Contact Hours: 60 Hrs.**

Unit I: Database Systems and Design (15 Hours)

Introduction - Introduction and applications of DBMS, Purpose of data base, Data, Independence, Database System architecture- Levels, Mappings, Database, users and DBA

Database Design - Database Design Process, ER Diagrams - Entities, Attributes, Relationships, Constraints, keys, extended ER features, Generalization, Specialization, Aggregation, Conceptual design with the E-R model.

Unit II: Relational Modelling and Storage Indexing (15 Hours)

The Relational Model - Introduction to the relational model, Integrity constraints over relations, Enforcing integrity constraints, Querying relational data, Logical database design : E-R to relational, Introduction to views, Destroying/altering tables and views.

Relational Algebra and Calculus - Preliminaries, relational algebra operators, relational calculus - Tuple and domain relational calculus, expressive power of algebra and calculus.

Overview of Storage and Indexing : Tree structured indexing - intuition for tree indexes, indexed sequential access method (ISAM), B+ Trees - a dynamic tree structure.

Unit III: Data Preprocessing (15 Hours)

Data Preprocessing, Data Cleaning, Data Integration, Data Reduction, Data Transformation and Data Discretization.

Unit IV: Data Warehousing and Online Analytical Processing (15 Hours)

Basic Concepts, Data Cube and OLAP, Design and Usage, Implementation, Data Generalization by Attribute-Oriented Induction

References

1. Raghurama Krishnan, Johannes Gehrke , Database Management Systems, 3rd edition, Tata McGraw Hill, New Delhi,India.
2. Elmasri Navathe, Fundamentals of Database Systems, Pearson Education, India.

3. Abraham Silberschatz, Henry F. Korth, S. Sudarshan (2005), Database System Concepts, 5th edition, McGraw-Hill, New Delhi,India.

4. Peter Rob, Carlos Coronel (2009), Database Systems Design, Implementation and Management, 7th edition.

5. Data Mining - Concepts and Techniques by Jiawei Han, Micheline Kamber, Jian Pei, Morgan Kaufmann Publication, Third Edition

6. Data Mining and Data Warehousing : Principles and Practical Techniques by Parteek Bhatia, Cambridge University Press, 2019 Edition

MSDSRM101T Research Methodology

Course Credit: 04 (Credit: Th(4)) **Total Contact Hours: 60 Hrs.**

Unit I: Descriptive Statistics (15 Hours)

Measures of Central Tendency: Mean, Median, Mode Partition Values: Quartiles, Percentiles, Box Plot.

Measures of Dispersion: Variance, Standard Deviation, Coefficient of variation.

Skewness: Concept of skewness, measures of skewness Kurtosis: Concept of Kurtosis, Measures of Kurtosis

Unit II: Introduction to R and RStudio (15 Hours)

Overview of R and its applications, Installing R and RStudio, RStudio interface and basic operations Introduction to R packages, Basic data types and variables in R, Arithmetic operations and basic functions

Data Structures and Data Manipulation: Working with vectors and matrices, Indexing and subsetting data, Data frames and tibbles, Importing and exporting data, Data manipulation using dplyr package, Handling missing data, Introduction to data visualization with ggplot2.

Unit III: Programming Concept in R (15 Hours)

Control structures: if-else, for loops, while loops, Functions in R, Writing your own functions, Scoping rules and environments, Debugging and error handling, Vectorization and efficient coding techniques, Applying functions to data using apply family of functions,

Unit IV: Data Visualization and Graphics (15 Hours)

Introduction to data visualization principles, Basic plotting functions in R, Advanced data visualization with ggplot2, Customizing plots and adding aesthetics, Creating interactive visualizations with plotly, Working with geographic data and maps.

Working on a small-scale project using R.

References

1. R for Data Science by Hadley Wickham and Garrett Grolemund.

2. The Art of R Programming by Norman Matloff

3. Data Manipulation with R by Phil Spector.

Dr. Homi Bhabha State University
Mumbai

Proposed M. Sc. (Data Science) Syllabus

Semester II

With Effect from Academic Year 2023 - 2024

M.Sc. (Data Science) Semester II

Course	Course Code	Course Name	Credits	Teaching Scheme				Examination Scheme						
								Theory				Practical		
				Lecture Per Week	Practical per Week	Total Periods	Duration in Hours	Max. Marks (Theory)	Max. Marks (Internals)	Total Marks	Minimum Passing Marks	Max Marks	Total Marks	Minimum Passing Marks
DSC3	MSDSDC201T	Statistical Methods	4	4	--	4	4	60	40	100	40	--	--	--
DSC4	MSDSDC202T	Machine Learning	4	4	--	4	4	60	40	100	40	--	--	--
DSE2	MSDSDE201T	Mathematics for Data Science	4	4	--	4	4	60	40	100	40	--	--	--
DSC3	MSDSLB201P	Data Analysis Using Excel LAB	2	--	8	8	4	--	--	--	--	50	50	20
DSC4	MSDSLB202P	Machine Learning LAB	2	--	8	8	4	--	--	--	--	50	50	20
DSE2	MSDSDE201P	Data Visualization Using Power BI LAB	2	--	8	8	4	--	--	--	--	50	50	20
FP/OJT	MSDSFP201P/ MSDSOJ201P	Field Project/OJT	4	--	4	4	4	60	40	100	40	--	--	--

M.Sc. (Data Science) Syllabus

Semester II

MSDSDC201T Statistical Methods

Course Credit: 04 **Total Contact Hours: 60 Hrs.**

Unit I: Elements of Statistical estimation theory (15 Hours)

Parametric point estimation: Statistic and estimator, sufficient statistics and factorization theorem, unbiasedness and consistency.

Computation of Mean Squared Error (MSE) of estimators and bias variance decomposition, Comparing estimators based on MSE, Rao Blackwell theorem and uniformly minimum variance unbiased estimator.

Unit II: Computational methods for finding estimators (15 Hours)

Method of maximum likelihood (ML), Optimal properties of ML estimators, Numerical computation of maximum likelihood estimator and Fisher information (both univariate and multivariate).

Unit III: Elements of Hypothesis testing (15 Hours)

Basics notions of testing of hypothesis: Concept of null hypothesis and alternative hypothesis, critical region, level of significance, type I and type II error, one sided and two-sided tests.

Likelihood ratio tests and Wald test, Computation of power functions, tests related to normal distribution, tests for mean, tests for equality of variance, tests for equality of means etc.

Unit IV: Elements of Bayesian estimation (15 Hours)

Concepts of prior and posterior distributions, Computation of posterior estimates of parameters, posterior distribution for functions of parameters, Jeffrey's prior.

Posterior inference for multiparameter models, Gibbs sampler, data analysis using software.

References

1. G. Casella and R. L. Berger. Statistical Inference. Duxbury Press.
2. A. M. Mood, F. A. Graybill and D. C. Boes, Introduction to The Theory of Statistics, Third Edition, Mc Graw Hill Education.
3. Laura Chihara and Tim Hesterberg, Mathematical Statistics and Resampling and R.John Wiley & Sons.
4. L. Wasserman, All of Statistics, Springer.
5. Daniel Sabanés Bové and Leonhard Held, Applied Statistical Inference: Likelihood and Bayes, Springer
6. Andrew Gelman, John B. Carlin, Hal S. Stern, David B. Dunson, Aki Vehtari, and Donald B. Rubin, Bayesian Data Analysis, CRC Press
7. Richard McElreath, Statistical Rethinking: A Bayesian course with Examples in R and Stan, Chapman & Hall/ CRC Texts in Statistical Inference.

MSDSDC202T Machine Learning

Course Credit: 06 (Th(4) + Pr(2)) **Total Contact Hours: 60 Hrs.**

Unit I: Regression Problem (15 Hours)

Simple and Multiple Linear Regression: Relationship between attributes using Covariance and Correlation Relationship between multiple variables: Regression (Linear, Multivariate, polynomial) in prediction. Residual Analysis Identifying significant features, feature reduction using AIC, multi-collinearity Non-normality and Heteroscedasticity Hypothesis testing of Regression Model Confidence intervals of Slope R-square and goodness of fit Influential Observations – Leverage.

Unit II: Classification Problems (15 Hours)

Logistic regression, Naive Bayes Classifier, classification tree, random forest, K-nearest neighbour algorithm.

Support Vector Machines: Linear learning machines and Kernel space, Making Kernels and working in feature space SVM for classification and regression problems.

Unit III: Model Selection and Regularization (15 Hours)

Regularization Methods: Regularization methods, Lasso, Ridge and Elastic nets, Categorical Variables in Regression.

Non-Linear Regression: Logit function and interpretation, Types of error measures (ROCR).

Validation Techniques (Cross-Validations).

Decision Trees & Ensembles methods: ID4, C4.5, CART, Bagging & boosting and its impact on bias and variance, C5.0 boosting Random forest Gradient Boosting Machines and XGBoost.

Unit IV: Multivariate Methods (15 Hours)

LDA, QDA, Principal component analysis, Factor analysis, principal component regression, multidimensional scaling.

References

1. Ethem Alpaydin, Introduction to Machine Learning, Second Edition
2. Stephen Marshland, Machine Learning: An Algorithmic Perspective.
3. Christopher M. Bishop, Pattern Recognition and Machine Learning.
4. Tom Mitchell, Machine Learning
5. Venkata Reddy Konasai, Shailendra Kadre, Machine Learning and Deep Learning Using Python and Tensor Flow, Mc Graw Hill Publication.
6. Prateek Gupta, Practical Data Science with Jupyter, BPB Publication.
7. An Introduction to Statistical Learning: with Applications in R (Springer Texts in Statistics) by Gareth James, Daniela Witten, et al
8. The Elements of Statistical Learning: Data Mining, Inference, and Prediction, Second Edition (Springer Series in Statistics)
9. Hands-On Machine Learning with Scikit-Learn and Tensor Flow: Concepts, Tools, and Techniques to Build Intelligent Systems
10. Python Machine Learning: Machine Learning and Deep Learning with Python, scikit-learn, and Tensor Flow, 2nd Edition

MSDSDE201T Mathematics for Data Science

Course Credit: 04 **Total Contact Hours: 60 Hrs.**

Unit I: Eigenvalues, eigenvectors and orthogonality (15 Hours)

Eigenvalues and eigenvectors, diagonalization, similar matrices, application to linear recurrences. (Chapter 3, Sections 3.3, 3.4)

Orthogonality of vectors in $\mathbb{R}^n$, Orthogonal set, Gram-Schmidt orthogonalization process, Orthogonal complement and projections. (Chapter 8, Section 8.1)

Orthogonal diagonalization, positive definite matrix, QR factorization. (Chapter 8, Section 8.2, 8.3, 8.4)

Unit II: Application (15 Hours)

Application to best approximation and Least Square, Singular Value Decomposition, Correlation and Variance, Principal Component Analysis, Constrained optimization, Linear Codes over Finite Fields, Fourier approximation. (Chapter 8, Sections 8.7, 8.9, 8.10, Chapter 10, Section 10.5)

Unit III: Differentiation (15 Hours)

Differentiation of Univariate Functions, maxima and minima, Taylor series, Differentiation Rules.

Partial Differentiation and Gradients, Basic Rules of Partial Differentiation, Chain rule, Jacobian matrix, Gradient of a vector-valued functions, Gradients of Matrices, Higher-Order Derivatives, Hessian Matrix, multivariate Taylor series, multivariate Taylor polynomial, local and global maxima & minima.

Unit IV: Optimization (15 Hours)

Optimization using Gradient Descent, Constrained Optimization and Lagrange Multipliers, Convex Optimization. Riemann integration, improper integrals. Applications of integral.

References

1. Linear Algebra with Application, W. Keith Nicholson, Mc Graw Hill publication

(7th edition).

2. Linear Algebra and its application, Gilbert Strang, Cengage Publication.

3. Linear Algebra and its application, David C. Lay, Steven R. Lay and Judi J. McDonald, Pearson Publication.

4. Linear Algebra Done Right, Sheldon Axler, Springer.

5. Practical Linear Algebra, Gerald Farin, Dianne Hansford, CRC Press.

6. Mathematics for Machine Learning, Marc Peter Deisenroth, A. Aldo Faisal and Cheng Soon Ong, Cambridge University Press.

7. Calculus, James Stewart, Books Cole Publication.

8. Calculus and Analytical Geometry, Thomas and Finney, Addison Wesley Publishing Company.

MSDSLB201P Data Analysis Using Excel LAB

Course Credit: 02 (Pr(2)) **Total Contact Hours: 30 Hrs.**

Unit I: Excel Fundamentals & Excel For Data Analytics (15 Hours)

Excel Fundamentals: Reading the Data, Referencing in formulas , Name Range, Logical Functions, Conditional Formatting, Advanced Validation, Dynamic Tables in Excel, Sorting and Filtering.

Working with Charts in Excel, Pivot Table, Dashboards, Data And File Security, VBA Macros, Ranges and Worksheet in VBA, IF conditions, loops, Debugging.

Excel For Data Analytics: Handling Text Data, Splitting, combining, data imputation on text data, Working with Dates in Excel, Data Conversion, Handling Missing Values, Data Cleaning, Working with Tables in Excel.

Unit II: Data Visualization with Excel (15 Hours)

Charts, Pie charts, Scatter and bubble charts, Bar charts, Column charts, Line charts, Maps

Multiples: A set of charts with the same axes, Matrices, Cards, Tiles.

References

1. Data Analysis Using Microsoft Excel by Michael R. Middleton.
2. Excel Data Analysis: Your Visual Blueprint for Analyzing Data, Charts, and PivotTables by Denise Etheridge
3. Data Analysis with Microsoft Excel: Updated for Office 365 by Kenneth N. Berk and Patrick Carey

---

MSDSDE201P Data Visualization Using Power BI LAB

Course Credit: 02 (Pr(2)) **Total Contact Hours: 30 Hrs.**

Unit I: Introduction to Power BI (15 Hours)

What is BI, Tools supporting BI, Why Power BI: Basics, Architecture, Fundamentals of Power BI, Power BI Building blocks, Installation of Power BI, and supported Data Sources, what is data modelling? Using Data Modelling and Navigation, Managing Time based Data, Create impactful charts and comprehensive reports on Power BI. Power BI Advantages, Excel Integration in Power BI, Power BI DAX.

Unit II: Power BI functionalities (15 Hours)

Visualisation Options: KPI indicators and importance of KPI visualization, Dashboards Options, Power BI and other Tools (Tableau, SSRS) comparison. Difference between Dashboard and Report.

Project using Power BI.

References

1. Fundamentals of Data Visualization, By Claus O. Wilke, April 2019.
2. Visual Analytics with Tableau, By Alexander Loth, May 2019.
3. Mastering Power BI by Chandraish Sinha September 2021.
4. Developing Analytic Talent: Becoming a Data Scientist by Vincent Granville.

MSDSFP201P/MSDSOJ201P Field Project/On Job Training

Course Credit: 04 **Total Contact Hours: 60 Hrs.**

Evaluation will be based on Dissertation submitted and presentation in front of the internal examiner (other than the Guide) within the department and external examiner.

Each students should submit the monthly report of their project to the department.

Internal Marks - 40 (to be given by the Guide)

External Examination - Project content & Project Report (20 marks) + Project Presentation (20 marks) + Viva (20 marks)
