St. Xavier's College (Autonomous), Kolkata
Department of Statistics
MSc in Data Science
Syllabus
Semester 1 & 2

Semester	**ONE**
Paper Number	**1**
Paper Code	**MDSC 4111**
Paper Title	**Introduction to Data Science and Data Structures**
No. of Credits	**6**
Course description	Composite Paper Module 1: 2 classes/week. Using Excel, R, Tableau. Module 2**:** 2classes/week. Using Python No. of classes assigned Theory: 4 classes per week Practical: 4 classes per week
Course Objective	At the end of the course, the students should be able to 1. Compute basic descriptive measures 2. Perform exploratory data analysis with descriptive statistics on given data sets. 3. Gain experience in creating a data visualisation for an application domain of their choice. 4. Critically evaluate and interpret a data visualisation. 5. Analyze, evaluate, and implement data structures & algorithms using python.
Syllabus	**Module1: Introduction to Data Science** **Unit 1:** Motivation for data Science: Story telling with data with illustration from different fields. Data abstraction and data wrangling. Storage of data. [4]

	Data: Types of data, scales of measurement. [3] Univariate Data: Descriptive measures related to univariate metric data. [4] Bivariate Data: Descriptive measures related to bivariate metric data: Correlations, linear and polynomial regressions. Descriptive measures related to bivariate categorical data: Measures of associations in a contingency table. [7] **Unit 2:** Exploratory data analysis: Philosophy of EDA, Basic tools of EDA (plots, graphs and summary statistics). [4] Data Visualization: Basic principles, ideas and tools for data visualization. Visualization of qualitative, quantitative, temporal, spatial and panel data. [4] **Module2: Data Structures** Introduction to Data Structures, Arrays, Linked Lists, Stacks, Queues, Binary Trees, Threaded Binary Trees, Binary Search Trees, AVL trees, Sets, Tuples, Dictionaries, Trie Searching and Sorting algorithms Basic ideas on complexity analysis, Big-Oh, Big-Omega, Big-Theta notations. [26]
List of Practical	1. Data visualization techniques for non-frequency data 2. Data visualization techniques for univariate metric data 3. Data visualization techniques for bivariate metric data 4. Descriptive measures for univariate metric data 5. Descriptive measures for bivariate metric data 6. Descriptive measures for bivariate categorical data 7. Introduction to Python. 8. Practical based on Data Structures using Python.
Reading/Reference Lists	1. Jeffrey M. Stanton (2013). Introduction to Data Science.

	2. The Visual Display of Quantitative Information (2nd Edition). E. Tufte. Graphics 3. Trevor Hastie, Robert Tibshirani and Jerome Friedman. Elements of Statistical Learning, Second Edition. ISBN 0387952845. 2009. (free online) 4. Yule G.U. and Kendall M.G (1994) : An Introduction to the theory of Statistics. 14th Edn. Universal Book stall, Delhi. 5. Hogg, R.V., Tanis, E.A. and Rao J.M. (2009): Probability and Statistical Inference, Seventh Ed, Pearson Education, New Delhi. 6. Data Structures and Algorithms in Python, Michael T Goodrich, Roberto Tamassia, Michael H Goldwasser, John Wiley 7. A First Course on Data Structures in Python, Donald R. Sheehy.

Semester	**ONE**
Paper Number	**2**
Paper Code	**MDSC 4112**
Paper Title	**Probability**
No. of Credits	**6**
Course description	Composite Paper One Module. Applications Using R No. of classes assigned Theory: 4 classes per week Practical: 3 classes per week
Course Objective	At the end of the course, the students are expected to have • Knowledge of basic ideas of Probability. • Knowledge of different types of random variables and their probability distributions. • Knowledge of different discrete and continuous standard theoretical distribution and their uses in modelling data through R. • Construction of Mixed distributions and their uses modelling data through R. • Basic knowledge of prior and posterior distributions.

Syllabus	Introduction to Probability: random experiments, sample space, events and algebra of events. Definitions of Probability – classical, statistical and axiomatic. [5] Conditional Probability: Theorem of compound probability, theorem of total probability, Bayes theorem and its applications, independent events. [4] Random variables and their probability distributions: PMF, PDF and CDF, statement of properties of CDF,Empirical distribution functions and their properties, illustrations and properties of random variables. Moments. Joint, marginal and conditional probability distributions, Joint PMF, PDF and CDF, statement of properties of Joint CDF, independence of variables. Markov's and Chebyshev's inequalities. Mixed random variables. Construction of probability distributions of mixed random variables. [12] Standard Univariate Discrete Theoretical Distributions: Binomial, Poisson, geometric, negative binomial, hypergeometric, uniform (Genesis, Statement of properties and applications). [8] Standard Univariate Continuous Theoretical Distributions: Rectangular, normal, exponential, Cauchy, beta, gamma, lognormal, logistic, double exponential and Pareto (Genesis, Statement of properties and applications). [10] Bivariate Normal Distribution (Genesis, Statement of properties and applications). [4] Truncated Distributions. [5]
List of Practical	1. Introduction to R. 2. Generation of random samples from given distributions. 3. Fitting of probability models by introducing the concepts of Method of moments and Likelihood methods 4. Modelling data through fitting of Standard Univariate Discrete Distributions. 5. Modelling data through fitting of Standard Univariate Continuous Distributions.

	6. Modelling data through fitting of truncated distributions. 7. Modelling data through fitting of Mixed Distributions. 8. Simulation based studies: Practical based on Markov's/ Chebyshev's inequalities.
Reading/Reference Lists	1. Ronald E. Walpole; Raymond H. Myers; Sharon L. Myers; Keying E. Ye *Probability and Statistics for Engineers and Scientists*, by Pearson, Ninth Edition (2013). 2. Sheldon Ross *A First Course in Probability*, Pearson, Ninth Edition (2018). 3. Prabhanjan N. Tattar, Suresh Ramaiah, B. G. Manjunath, *A Course in Statistics with R*; Wiley, (2018).

Semester	**ONE**
Paper Number	**3**
Paper Code	**MDSC 4113**
Paper Title	**Linear Algebra and Elements of Statistical Inference**
No. of Credits	**6**
Course description	Composite Paper Module 1: 2 classes/week Module 2**:** 2classes/week No. of classes assigned Theory: 4 classes per week Practical: 4 classes per week
Course Objective	After completion of the course a student is expected to have an idea of • Matrix algebra and determinants. • Vector spaces, subspaces, their dimensions and basis. • Rank of a matrix and systems of linear equations. • Characteristic roots and vectors along with the understanding of classification of quadratic forms. • Applications on least squares and dimension reduction.

	• Understand the concept of an iid sample. • Conceptualise drawing samples from theoretical distributions. • Conceptualise level, size, power of a test and the errors associated with a testing problem. • Applying the results of sampling distributions to build test statistics and critical regions. • Construct confidence intervals for parameters and develop their relation with testing of hypotheses.
Syllabus	**Module1: Linear Algebra** **Unit 1:** Vectors – Concept of a vector, length of a vector, Angle between two vectors, Orthogonal and orthonormal vectors, Linear dependence and independence of vectors, Vector spaces, Spanning set of a vector space, Basis of a vector space, Dimension of a vector space, Projection of a vector on a vector space, Orthogonal Basis, Orthocomplement of a vector space, Gram-Schmidt orthogonalization procedure, Row space & column space of a matrix. Matrices (as a vector of vectors), Square matrices, Matrix operations (Addition, subtraction, multiplication by a scalar and by a matrix, Kronecker Product), Null matrix, Identity matrix, symmetric and skew symmetric matrices, orthogonal matrices, Rank of a matrix, singular and non-singular matrices, A few important results on the rank of a matrix, Inverse of a matrix, idempotent matrices, Elementary Transformations on a matrix, Reduction of a matrix to echelon, and diagonal forms by elementary transformations, QR factorization, Trace of a matrix, Partitioning of matrices and simple properties. [11] **Unit2:** Direct and iterative methods for solving a linear system of equations: Gaussian elimination, LU factorization, Cholesky method, Householder's matrices, Jacobi's method, Gauss-Seidel method. [5] Characteristic roots and Characteristic vector, Properties of characteristic roots, Spectral Decomposition, Singular value decomposition. [5] Quadratic forms: Classification & canonical reduction. [3]

	Module 2 : Elements of Statistical Inference **Unit 3:** Sampling Distributions *Basic Concepts*: Concept of an iid sample, Statistic and its standard error. Drawing of random samples from theoretical distributions. Illustrations with R. (3) *Techniques of Sampling Distributions:* Distribution Function, Moment Generating Function and Transformation of variables technique to obtain sampling distribution of statistics. (2) *Basic Sampling Distributions from Univariate Normal Distributions :* Chi-square, t and F. Degrees of freedom. Sampling Distributions of sample mean, sample variance, their independence, linear combinations of normal variables. Definitions of Non-central chi-square, t and F distributions. Illustrations through simulations. Illustrations through simulations. (5) *Basic Sampling Distributions from Bivariate Normal Distributions :* Sampling distributions of sample correlation coefficient and linear regression coefficients. Illustrations through simulations. (2) **Unit 4:** Elements of Inference Problems *Problems and Paradigms of Inference:* Estimation and Testing of Hypotheses Problems. Parametric and Nonparametric Inference. Classical and Bayesian Inference. (3) *Estimation:* Basic Criteria of a good estimator – Sufficiency, Unbiasedness, Minimum Variance, Consistency and Efficiency. OPEF. (4) *Interval Estimation:* Methods of finding confidence intervals. Shortest confidence intervals. Confidence belts. (3) *Testing of Hypothesis:* Null and Alternative Hypothesis. Simple and Composite Hypothesis. Type-1 and Type-2 Errors. Level and Power of a Test. Power Function. p-value of a test. p-hacking. Application to tests of significance. [4]
List of Practical	1. Orthogonalisation of vectors. 2. Solution of linear system of equations using direct and iterative methods. 3. Characteristic roots and vectors, spectral decomposition and singular value decomposition 4. Classification of Quadratic Forms 5. Drawing random samples from Discrete Distributions

	6. Drawing random samples from Continuous Distributions 7. Studying the basic sampling distributions arising from a normal distribution through simulations 8. Studying consistency and efficiency of estimators through simulations 9. Obtaining empirical level and power of tests through simulations 10. Applying standard tests to real life data 11. Obtaining confidence intervals of parameters based on real life data
Reading/Reference Lists	1. Stephen Boyd and Lieven Vandenberghe, Introduction to Applied Linear Algebra: Vectors, Matrices, and Least Squares (Cambridge University Press, 3rd edition) 2. Lloyd N. Trefethen and David Bau, III: Numerical linear algebra, SIAM (1997) 3. Matrix Algebra: Theory, Computations and Applications in Statistics by J.E. Gentle, Springer, 2007 4. Fundamentals of Matrix Computations by D.S. Watkins, 2nd ed., Wiley, New York, 2002. 5. Gilbert Strang; Linear Algebra and its Applications; Academic Press; Second Edition. 6. Goon A.M., Gupta M.K., Das Gupta.B.: Fundamentals of Statistics, Vol. 1, World Press, 2010. 7. Goon, A.M. Gupta, M.K. and Dasgupta, B. : An outline of Statistical Theory, Vol. 1, World Press, 2010. 8. Ismay, C. and Kim, A.Y., Statistical Inference via Data Science, A Modern Dive into R and the Tidyverse, CRC Press Talor and Francis group, 2020. 9. Moulin, P. and Venugopal, V.V., Statistical Inference for Engineers and Data Scientists, Cambridge University Press. 10. Caffo, B., Statistical Inference for Data Science, Leanpub, 2016.

Semester	**ONE**
Paper Number	**4**
Paper Code	**MDSC 4114**
Paper Title	**Data Base Management System and Data warehousing**
No. of Credits	**6**
Course description	Composite Paper One Module. Practical using SQL. No. of classes assigned Theory: 4 classes per week

	Practical: 3 classes per week
Course Objective	At the end of the course, the students are expected to have • The students will learn the fundamental concepts of database management system • They will learn how to design databases and how to improve the schemas. • The students will also learn about transaction processing, concurrency control and recovery techniques • They would also be familiarized with query processing and optimization techniques • Finally they will be introduced to data warehousing with a discussion on fundamental concepts.
Syllabus	Fundamental concepts of DBMS: Data Abstraction; Data Models (Entity Relationship Model, Relational Model) Database Languages; Structured Query Language. Database design: Constraints; Referential Integrity; Functional Dependencies; Normalization. File Organization; Indexing and Hashing. Transaction Management, Concurrency Control & Recovery: ACID properties; rollback; recoverable schedules; lock-based protocols; deadlock handling; failure; recovery techniques. Query processing and optimization. Data warehousing: Basic Concepts, OLTP, Data Warehouse Architecture; Schemas; Data Marts; Data Warehouse Design; On-line Analytical Processing (OLAP).
Practical	Practical based on the theory component.
Reading/Reference Lists	1. R. Elmasri, S.B. Navathe, Fundamentals of Database Systems 6th Edition, Pearson Education, 2010. 2. A. Silberschatz, H.F. Korth, S. Sudarshan, Database System Concepts 6th Edition, McGraw Hill, 2010. 3. “Data Mining: Concepts and techniques”, J Han and M Kamber, Third Edition, Elsivier. 4. C.J Date, An Introduction to Database systems; Pearson/Addison Wesley; 2003

Semester	**TWO**
Paper Number	**5**
Paper Code	**MDSC 4211**
Paper Title	**Statistical Inference**
No. of Credits	**6**
Course description	Composite Paper Module 1: 2 classes/week Module 2**:** 2classes/week No. of classes assigned Theory: 4 classes per week Practical: 3 classes per week
Course Objective	At the end of the course, the students should be able to 1. Apply likelihood and moment methods to obtain estimates of parameters. 2. Construct tests of hypotheses based on the likelihood function. 3. Apply the chi-square to test goodness of fit and homogeinity on real life data. 4. Apply resampling techniques to obtain estimates, standard errors of estimates and confidence intervals of parameters. 5. Identify the parameters of a Gauss Markov model. 6. Differentiate between ANOVA and regression models. 7. Apply the theorems of least squares to carry out tests in ANOVA and regression models and identify worth of a concomitant variable in an ANOCOVA Model. 8. Apply nonparametric tests to data where the parent distribution is unknown in structure. 9. Differentiate between classical and Bayesian school of thoughts. 10. Select suitable priors of parameters and obtain the posteriors.
Syllabus	**Module-1** **Unit 1 : Parametric Methods (14 L)** *Parametric Methods:* Method of Moments, Maximum Likelihood Estimators. Likelihood Ratio, Rao's Score and Wald Tests. Statements of their large sample properties. Pearsonian Chi-square and its uses. (14) **Unit2: Bayesian Inferential Methods and Resampling Techniques(12 L)**

	Priors and Posteriors: Prior and posterior distributions, conjugate priors. Objective priors. Mixtures of conjugate priors. (6) Concept of Jackniffe and Bootstrap. Resampling methods in estimation. Bootstrap Confidence Intervals. Cross-validation studies. Illustrations with R. (6) **Module-2** **Unit 1 : Nonparametric Methods (10 L)** Basic tests of location and scale. Tests of Goodness of fit, Homogeneity and Associations. (10) **Unit 2 : Linear Models (16 L)** *The Gauss-Markov Model*: Least Square Estimators. Normal Equations and their solutions. Best Linear Unbiased Estimators. The Gauss Markov Theorem. Error and Error Variance. (3) *Linear Models:* ANOVA, Regression and ANOCOVA Models and some related testing problems. (10) *Simultaneous confidence intervals*: Bonferroni, Scheffe, Tukey, HSU and Duncan's Methods. Comparisons. (3)
List of Practical	1. Parametric methods of estimation. 2. Likelihood based tests of hypotheses. 3. Application of the chi-square tests to real life data. 4. Point and interval estimation using resampling techniques. 5. ANOVA, Regression and ANOCOVA Models. 6. Nonparametric tests of hypotheses.
Reading/Reference Lists	1) Goon A.M., Gupta M.K., Das Gupta.B.: Fundamentals of Statistics, Vol. 1, World Press, 2010. 2) Christensen R., Johnson W., Branscum A., Bayesian Ideas and Data Analysis: An Introduction for scientists and statisticians, Chapman & Hall, 2010. 3) Faraway, J., Linear Models with R, CRC Press, Second Edition. 2014. 4) Faraway, J., Extending the Linear Model with R, CRC Press, Second Edition. 2016. 5) Trevor Hastie, Robert Tibshirani and Jerome Friedman. Elements of Statistical Learning,Second Edition. 6) Ismay, C. and Kim, A.Y., Statistical Inference via Data Science, A ModernDive into R and the Tidyverse, CRC Press Talor and Francis group, 2020. 7) Moulin, P. and Venugopal, V.V., Statistical Inference for Engineers and Data Scientists, Cambridge University Press.

	8) Caffo, B., Statistical Inference for Data Science, Leanpub, 2016. 9) Nonparametric Statistical Inference, Gibbons and Chakraborty, CRC Press, First Edition.

Semester	**TWO**
Paper Number	**6**
Paper Code	**MDSC 4212**
Paper Title	**Multivariate Analysis**
No. of Credits	**6**
Course description	Composite Paper Module 1: Unit 1 (2 classes/week) Module 2**:** Unit 2 (2classes/week) No. of classes assigned Theory: 4 classes per week Practical: 3 classes per week
Course objective	At the end of the course, the students should be able to understand • Multivariate Probability Distributions. • Sampling distributions of some statistics drawn from Multivariate Normal distribution. • Basic concepts and applications of Copula. • Multivariate Data Visualisation. • Application of multivariate techniques.
Syllabus	**Module-I: Multivariate Probability Distributions** **Multivariate Data Visualisation:** Mosaic Plots, Scatterplot Matrix, Bivariate qq-plots, Spider Web plots, DD Plots, Parallel coordinate plots, Trellis Displays. [6]

	Multivariate Probability Distributions: Random Vector, Mean vector & Dispersion matrix, Probability mass/density functions, Marginal & Conditional distributions, Multiple and partial correlation coefficient, Multinomial Distribution, Dirichlet Distribution, Multivariate Normal distribution and its properties. [12] **Sampling from Multivariate Normal Distribution:** Sampling distribution for mean vector and variance-covariance matrix, Wishart Distribution, Hotelling T^2 and Mahalanobis D^2. [4] **Copula:** Definition and basic properties, Multivariate Distribution using Copula Functions. [4] **Module-2: Multivariate Techniques** Decomposition of data matrices by factors, Principal Component Analysis, Independent Component Analysis, Factor Analysis, Correspondence Analysis, Canonical Correlation Analysis, Discriminant Analysis, Cluster Analysis, Multidimensional Scaling. [26]
List of Practical	1. Analysis of multivariate frequency type data. 2. Multivariate Probability distributions. 3. Sampling from Multivariate Normal Distributions. 4. Copula 5. Multivariate Data Visualisation 6. Multivariate Techniques.
Reading/Reference Lists	1. Johnson/Wichern; Applied Multivariate Statistical Analysis Sixth Edition, Pearson 2015 2. Hardle Wolfgang, Simar Leopold: Applied Multivariate Statistical Analysis, Second Editon, Springer. 3. Roger B. Nelsen: An introduction to Copulas, Second Editon, Springer.

Semester	**TWO**
Paper Number	**7**

Paper Code	**MDSC 4213**
Paper Title	**Big Data Analytics**
No. of Credits	6
Course Description	Composite Paper One Module Number of classes: Theory – 4 per week Practical – 3 per week
Course Objective	After completion of the course a student is expected to have o Understanding of the challenges of computation related to big data. o Gaining wholesome knowledge about various computational platforms available for big data analytics. o Understanding the advantages and disadvantages of the big data analytics platforms, including the software frameworks. o Gaining wholesome knowledge about parallel computation in various big data analytics platforms. o Gaining hands-on experience in parallel computing with R and Python. o Gaining hands-on experience in cloud computing.
Syllabus	**Introduction:** Examples of big data in natural sciences, engineering, social media, industry, etc. Importance of analysing big data. Limitations of the traditional computational platforms in analysis of big data. **Scaling of big data analytics platforms:** Horizontal and vertical scaling, Peer-to-peer networks, Hadoop, Spark, Berkeley Data Analysis Stack (BDAS), High Performance Computing (HPC) clusters, multicore processors, Graphics Processing Unit (GPU), Field Programmable Gate Arrays (FPGA). **Distributed computing:** Importance of distributed computing for big data, Basic ideas of the communication systems for parallel computing in peer-to-peer networks (Message Passing Interface (MPI)), Hadoop (HDFS, YARN, Map Reduce), Spark, BDAS, (Tachyon+Mesos– improvement over Spark due to more aggressive memory exploitation). Communication systems for vertical scaling – MPI for HPC and multicore processors; CUDA for GPUs, Hardware Descriptive Language (HDL) for FPGA. The concept of cloud computing. **Comparisons of different big data platforms**: communication mechanisms based on scalability, data I/O performance,

	fault tolerance, real-time processing, data size supported, iterative task support. **Pseudocodes:** Illustrative examples of simple pseudocodes of the K-means algorithm in MapReduce, MPI and GPU based platforms.
List of Practical	1. Writing and implementing parallel codes in R and / Python. 2. Implementing the parallel codes in cloud. 3. Implementing Hadoop/Spark's machine learning algorithms in cloud with big data.
Reference List	1. Sourav Mazumder, Robin Singh Bhadoria and Ganesh Chandra Deka (2017), "Distributed Computing in Big Data Analytics", Springer. 2. Martin Van Steen and Andrew S Tanenbaums: *Distributed Systems* 3rd Edition (2017) 3. Singh, D. and G. K. Reddy (2014). A Survey on Platforms for Big Data Analytics, Journal of Big Data 1:8, 1–20

Semester	**TWO**
Paper Number	**8**
Paper Code	**MDSC 4214**
Paper Title	**Predictive Analytics**
No. of Credits	**6**
Course Description	Composite Paper One Module No. of classes assigned Theory: 4 classes per week Practical: 3 classes per week
Course Objective	At the end of the course, the students should be able to o Develop the concept of regression and classification o Understand the different model selection methods

	○ Apply different dimension reduction techniques to real life data ○ Analyse current data and make future predictions
Syllabus	***UNIT 1: Introduction*** Diagnostic versus prognostic models. Regression versus classification problems. Bias-variance trade off. [4] ***UNIT 2: Linear Regression*** Least square method, simple linear regression, multiple linear regression with quantitative & qualitative predictors. Dummy variables, regression diagnostics (Outlier detection, leverage, Influential point, Cook's distance, Model selection via AIC and BIC, adjusted R-Square). K-nearest neighbour regression . [10] ***UNIT 3: Classification*** Logistic regression, multiple logistic regression, multi-category logistic regression (model, parameter estimation and prediction). Multiclass discriminant analysis. Decision trees (CART and CHAID) [15] ***UNIT 4: Model selection and Regularization***

	Subset selection method (forward and backward stepwise selection), Shrinkage methods: Penalized likelihood and Bayesian linear regression; Ridge regression, the LASSO and Elastic NET. Applications of dimension reduction techniques. [15] ***UNIT 5: Generalized linear model*** Components of GLM, link functions (logit, probit, log link). Fitting of GLM (parameter estimation and prediction). Contingency tables, odds ratio and log linear models. Generalized linear mixed models (Inference for model parameters and predictions). [8]
List of Practical	1. Fitting of linear regression and regression diagnostics 2. Logistic regression 3. Multiclass discriminant analysis 4. CART and CHAID 5. Subset selection methods 6. Ridge Regression 7. LASSO 8. Elastic Net 9. Fitting of Generalized linear models Fitting of Generalized linear mixed effects model
Reading/Reference Lists	1. James, Witten, Hastie and Tibshirani: *An Introduction to Statistical Learning.* Second edition, Springer. 2. Hastie, Tibshirani, Friedman: *The Elements of Statistical Learning, Data Mining, Inference and Prediction*. Second Edition, Springer Series in Statistics. 3. McCullagh, P & Nelder, J.A.(1995), Generalized Linear Models. Chapman and Hall. 4. Agresti, A. (2007): An Introduction to Categorical data analysis. Wiley