MAHATMA GANDHI UNIVERSITY KOTTAYAM

MASTER OF SCIENCE
IN
DATA ANALYTICS

PROGRAMME STRUCTURE AND SYLLABUS

(2020-21 ADMISSION ONWARDS)

(UNDER MAHATMA GANDHI UNIVERSITY CSS REGULATIONS 2019)

MAHATMA GANDHI UNIVERSITY

2020

Contents

EXPERT COMMITTEE FOR M.Sc. DATA ANALYTICS

Convener : Dr. Annie Cherian, Associate Professor in Statistics, Department of Mathematics & Statistics, Baselius College, Kottayam, email id : annshinto@yahoo.co.in, Ph.No. 9446125298

Members

1. Dr. Gijo E.V, Associate Professor, ISI Bangalore Centre, email : gijoev@gmail.com, Ph.No. 9448324220
2. Dr. K.K Jose , Hon. Director, School of Mathematics & Statistics, M.G University , Kottayam, email id : kkjstc@gmail.com, Ph. No. 9446560608
3. Dr. Varghese Mathew, Principal, Mar Thoma College, Thiruvalla, email id :Varghese_m1@yahoo.co.in, Ph. No. 9447358620
4. Dr. Mendus Jacob , CEO & MD IPSR Group, Email id : ceo@ipsrsolutions.com, Ph. No. 9447053716
5. Dr. Sebastian George, Professor & Head, Department of Statistics, Kannur University, Email id : sthottom@gmail.com, Ph.No. 9447804027
6. Dr. Jayamol K.V, Associate Professor & Head, Department of Statistics, Maharajas College, Ernakulam, email id: jkvalavil@gmail.com, Ph.No. 9447036746
7. Prof. Preetha Rachel George, Associate Professor & Head, Department of Statistics, Mar Thoma College, Thiruvalla, email id : rachelpgeorge@gmail.com, Ph. No: 9745402829
8. Prof. Manesh Jacob , Assistant Professor,Mar Thoma College, Thiruvalla, email id : kochumon13@gmail.com, Phone No. 9895949365
9. Prof. Tessymol Abraham, Assistant Professor , Department of Mathematics & StatisticsBaselius College, Kottayam, email id : tessy02@gmail.com, Ph.No. 9495622403

M. Sc. DATA ANALYTICS PROGRAMME

(Mahatma Gandhi University Regulations CSS 2020 w. e. from 2020-21 Academic Year)

Philosophy

Education should, make people to solve problems and meet the challenges of changing world, suggest better alternatives to cater the needs of the contemporary world and efficient and effective means to predict the phenomena that may happen in future at various fields especially in environmental, economic, health and educational sectors from the existing data.

Need and Significance

Every second, world is driven by millions of decisions at various horizons. Decision making is increasingly becoming data driven within the medical, commercial and the government sectors. There is a large quantity of data readily available for this at present. Smart devices connected to the internet bring more input sources, apart from those generated by humans as they interact with services and one another. Data driven decision making involves the analysis of large volumes of data to identify patterns and build predictive models. This requires a combination of skills ranging from computing, statistics and mathematics, and is broadly labelled as data science. As formal academic programs in data analytics are just emerging, there is a huge gap between the demand for data scientists and the supply of suitable qualified applicants in the job market. The throughput of traditional classroom programs is limited and will not be able to provide the numbers needed to meet the upcoming requirements.

Mathematics and Statistics play a pivotal role in Data Analysis, Machine learning and Artificial intelligence. The National Educational Policy 2020 point out that “With various dramatic scientific and technological advances, such as the rise of big data, machine learning, and artificial intelligence many unskilled jobs worldwide may be taken over by machines while the need for skilled workforce, particularly involving mathematics, computer science and data science in conjunction with multi-disciplinary abilities across the sciences, social sciences and humanities will be increasingly in greater demand.”

Many students from our state wish to join the data science courses, but no such programme is available in government aided sector and the fee for the programme offered by institutes are not affordable to those who are financially backward and so students from weaker and marginalized communities cannot access these programmes.

1. **Eligibility for admissions:**

   B.Sc. Degree with Mathematics / Statistics / Data Science as a core subject or B. Tech./B.E in Computer Science / IT or Bachelor of Computer Applications, provided the candidate has studied at least 2 courses in Probability / Statistics at degree level.

2. **Examination :** Credit and Semester System (CSS)

3. **Medium of instruction and assessment** : English

4. **Duration of the Course : 4 Semesters (2 years)**

5. **Faculty under which the Degree is awarded**: Science

6. **Specializations offered if any**: List of Electives enclosed

7. **Objectives and Outcomes of the Program:**

Objectives:

.

- To Introduce Data Analytics as a multi-disciplinary branch of science for solving everyday problems by analyzing relevant data
- To provide advanced level teaching and training in theory and applications of Data Analytics as well as skills in computer programming and data interpretation.
- To provide a platform for talented students to become leaders in this discipline by undergoing higher studies in the subject as well as to train them to suit for the needs of the society as job providers by establishing startups and business.
- To allow more flexibility to branch out into other emerging areas of Statistics, Computer Science, and Data Analytics.
- To draw together a variety of subject areas to enable students to model real-world data from various contexts by exploring a blend of Applied Mathematics and Statistics with appropriate computing tools including free softwares like R, Python etc.

- To provide special attention to interdisciplinary areas of research and applications in describing, exploring, analyzing and comparing data with an innovative research mind in a data driven world.
- To make them familiar with emerging developments in Big Data Analytics and their applications in various areas.
- To provide the students on the job training in industrial applications and professional development with a view to enable them to get opportunities for teaching, research and employment in India and abroad through industry academia collaborations and linkages with reputed research institutes and industries in India and abroad.

Outcomes:

- After undergoing this program, students will get advanced knowledge in theory and applications in all areas of Data Analytics, Statistical Learning, Machine Learning, Data Base Management, Artificial Intelligence, etc.
- Students have secured practical skills in statistical methods and computer programming to plan and execute projects and decision making using Data Analytics, Machine Learning etc
- Students are well equipped to undertake any work involving exploratory data analysis, fraud analytics, data learning, text mining etc. as future entrepreneurs.
- Students have developed skills in advanced computing softwares like R and Python for big data analytics, computing and data analysis.
- Students are well trained to take up jobs in reputed firms and MNCs etc as Data Analysts, Data Engineers, Risk Analysts, Business Analysts, Financial Analysts, Decision Makers, Entrepreneurs etc.
- Students are motivated to pursue teaching and research in all emerging areas of research in theoretical and applied branches of Data Analytics and related areas.

8. The Programme Structure :M.Sc Data Analytics

Table of Courses and Credits

COURSE CODE	COURSE TITLE	TEACHING (LECTURE+ PRACTICAL)	TYPE	CREDIT S
	SEMESTER I **TOTAL CREDITS 19**			
ST 050101	STATISTICAL FOUNDATION FOR DATA ANALYTICS	**3L+2P**	**THEORY**	**3**
ST 050102	MATHEMATICAL FOUNDATION FOR DATA ANALYTICS 1	**4L+1P**	**THEORY**	**3**
ST 050103	REGRESSION ANALYSIS	**3L+2P**	**THEORY**	**3**
ST 050104	DATA BASE TECHNOLOGY	**3L+2P**	**THEORY**	**2**
ST 050105	PROGRAMMING AND DATA STRUCTURES WITH PYTHON	**2 L+ 3P**	**THEORY**	**2**
ST 050106	PRACTICAL 1		**PRACTICAL ***	**3**
ST 050107	PRACTICAL 2		**PRACTICAL ***	**3**
	SEMESTER II **TOTAL CREDITS 19**			
ST 050201	MATHEMATICAL FOUNDATION FOR DATA ANALYTICS 2	**4L+1P**	**THEORY**	**3**
ST 050202	MULTIVARIATE ANALYSIS	**3L+2P**	**THEORY**	**3**
ST 050203	STOCHASTIC PROCESS AND TIME SERIES ANALYSIS	**3L+2P**	**THEORY**	**3**
ST 050204	DATA VISUALIZATION	**3L+2P**	**THEORY**	**2**
ST 050205	PROGRAMMING USING R	**2L + 3P**	**THEORY**	**2**
ST 050206	PRACTICAL 3		**PRACTICAL ***	**3**
ST 050207	PRACTICAL 4		**PRACTICAL ***	**3**
	SEMESTER III **TOTAL CREDITS 21**			
ST 050301	SAMPLING & DESIGN OF EXPERIMENTS	**4L+2P**	**THEORY**	**3**
ST 050302	OPTIMIZATION TECHNIQUES	**4L+2P**	**THEORY**	**3**
ST 050303	MACHINE LEARNING	**5L+ 2P**	**THEORY**	**3**
ST 050304	BIG DATA ANALYTICS AND HADOOP	**4L+2P**	**THEORY**	**3**
ST 050305	INTERNSHIP			**3**

ST 050306	PRACTICAL 5		**PRACTICAL ***	**3**
ST 050307	PRACTICAL 6		**PRACTICAL ***	**3**
	SEMESTER IV TOTAL CREDITS 21			
	ELECTIVE 1	**3L + 2 P**	**THEORY**	**3**
	ELECTIVE 2	**3L + 2 P**	**THEORY**	**3**
	ELECTIVE 3 - PRACTICAL BASED ON ELECTIVES 1 & 2		**PRACTICAL ***	**3**
ST 050401	INDUSTRIAL VISIT	5		5
ST 050402	PROJECT	10		5
ST 050403	COMPEHENCIVE VIVA VOCE			2
	GRAND TOTAL OF CREDITS			**80**

COURSE CODE	**LIST OF ELECTIVE COURSES**	**TEACHING (LECTURE+ PRACTICAL)**	**TYPE**	**CREDITS**
ELECTIVES – BUNCH 1				
ST 900401	ARTIFICIAL INTELLIGENCE	**3L+2P**	**THEORY**	**3**
ST 900402	EPIDEMIOLOGY AND CLINICAL TRIALS	**3L+2P**	**THEORY**	**3**
ST 900403	DATA SCIENCE PRACTICAL		**PRACTICAL ***	**3**
ELECTIVES – BUNCH 2				
ST 910401	CLOUD COMPUTING	**3L+2P**	**THEORY**	**3**
ST 910402	RELIABILITY MODELING AND STATISTICAL QUALITY CONTROL	**3L+2P**	**THEORY**	**3**
ST 910403	DATA ANALYTICS PRACTICAL		**PRACTICAL ***	**3**
ELECTIVES – BUNCH 3				
ST 920401	WEB ANALYTICS	**3L+2P**	**THEORY**	**3**
ST 920402	ECONOMETRICS	**3L+2P**	**THEORY**	**3**
ST 920403	DATA MANAGEMENT PRACTICAL		**PRACTICAL ***	**3**

* All Practical question papers shall be generated from university

SEMESTER I COURSES

ST 050101– STATISTICAL FOUNDATION FOR DATA ANALYTICS

Objectives: This course is designed to introduce the concepts of theory of probability, random variables, probability distributions, estimation and testing of hypothesis. This paper also deals with the concept of parametric tests for large and small samples. It provides knowledge about non-parametric tests and its applications. An introduction to the Bayesian concepts in Statistics is also provided. It is also expected to give lab illustration of the concepts through original data sets.

Outcomes (i) Demonstrate the concepts of probability theory, random number generation, distribution theory, sampling distributions, point and interval estimation of unknown parameters and their significance using large and small samples. (ii) Apply the idea of sampling distributions of different statistics in testing of hypotheses. (iii) To understand and apply nonparametric tests for single sample and two samples. (iv) To familiarize the students with Bayesian philosophy.

Module 1

Basic elements of probability. Introduction to random variables and probability distributions. Univariate distributions- Binomial, Poisson, Geometric, Exponential, Gamma, Beta, Normal and Lognormal distributions, Sampling Distributions and their properties, Random number generation-Basic principles of Random number generation, inversion method, accept-reject method, Random number generation from common distributions.

Module 2

Concepts of Estimation, Estimators and Estimates. Point and interval estimation. Properties of good estimators- unbiasedness, efficiency, consistency and sufficiency. Methods of moments, maximum likelihood. Invariance property of ML Estimators (without proof). minimum variance. Interval Estimation, $100(1-\alpha)\%$ confidence intervals for mean, variance, proportion, difference of means, proportions and variances.

Module 3

Basic ideas of testing of hypotheses, significance level, power, p-value, Neyman-Pearson fundamental Lemma (statement only), Distributions with monotone likelihood ratio – Problems, Generalization of the fundamental lemma to randomized tests, uniformly most powerful tests, two sided hypotheses – testing the mean and variance of a normal distribution, testing equality of means and variances of two normal distributions. Likelihood ratio tests – locally most powerful tests, Large and small sample tests.

Module 4

Sequential Probability Ratio Tests. Introduction to Non parametric tests – Ordinary Run test, Mann Whitney U test, Wilcoxon Signed Rank test, Kruskal Wallis test, Median test and Sign test. Bayesian Statistics – concept.

References

1. Goon A. M., Gupta M. K., & Dasgupta B. (2005). *Fundamentals of Statistics*, Vol.I, 8th edition, World Press, Kolkatta.
2. Rohatgi, V. K., & Saleh, A. M. E. (2015). *An introduction to probability and statistics*. John Wiley & Sons.
3. Srivastava M. K, Khan A.H & Srivastava N (2014) *Statistical Inference : Theory of Estimation*, PHI Learning pvt ltd.
4. Srivastava M. K, & Srivastava N (2009) *Statistical Inference : Testing of Hypothesis*, PHI Learning pvt ltd.
5. Mood , Graybill & Boes (1974) Introduction to the theory of statistics
6. Mitzenmacher, M., &Upfal, E. (2017). *Probability and computing: Randomization and probabilistic techniques in algorithms and data analysis*. Cambridge university press.

ST 050102 - MATHEMATICAL FOUNDATION FOR DATA ANALYTICS 1

Objectives: To make students familiar with various concepts in linear algebra and matrices and their applications in data analytics.

Outcomes: On successful completion of this course, a student will be able (i) to understand basics concepts of Linear Algebra (ii)To understand concepts of vector spaces and matrices (iii) use the properties of Linear Maps in solving problems in Linear Algebra(iv) Demonstrate proficiency on the topics Eigen values, Eigen vectors and can apply linear algebra for applications in Data Analytics

Module 1

The Geometry of Linear Equations- An Example of Gaussian Elimination- Matrix Notation and Matrix Multiplication - Triangular Factors and Row Exchanges- Inverses and Transposes **Module 2**

Vector Spaces and Subspaces – Solving Ax=0 and Ax=b - Linear Independence, Basis and Dimension- The Four Fundamental Subspaces- Graphs and Networks- Linear Transformations **Module 3**

Introduction- Properties of the Determinant- Formulas for the Determinant - Applications of Determinants, Eigenvalues and Eigenvectors**:** Introduction- Diagonalization of a Matrix - Complex Matrices- Similarity Transformations .

Module 4

Minima, Maxima, and Saddle Points - Tests for Positive Definiteness- Singular Value Decomposition –Minimum principles.

Lab Exercises:

Module 1

1. Matrix multiplication

2. PA=LDU factorization

3. Calculation of A^T

4. Calculation of A^{-1}

- Gauss Jordan Method

Module 2

5. Solving system of linear equations.

6. Find the basis and rank of a matrix.

7. Find the matrix representation of a linear transformation.

Module 3

8. Find the determinant of a matrix.

9. Find A^{-1} using determinant.

10. Find the eigen value and eigen vector of a matrix.

11. Diagonalization of a matrix.

12. Check whether a matrix is

    a) Hermitian
    b) Skew-Hermitian
    c) Unitary

13. Find the Jordan form of a matrix.

Module 4

14. Find the nature of a matrix / Check whether the given matrix is

    a) Positive definite
    b) Negative definite
    c) Positive semi-definite
    d) Negative semi-definite
    e) Indefinite

15. Find the point of maximum or minimum or saddle point of a function.

16. Find the singular value decomposition of a matrix.

References

1. Gilbert Strang(2006). Linear Algebra and Its Application, Fourth Edition, Academic Press.
2. Sheldon Axler, Linear Algebra Done Right, Springer, 2017.
3. E. Davis, Linear Algebra and Probability for Computer Science Applications, CRC Press, 2012.
4. J. V. Kepner and J. R. Gilbert, Graph Algorithms in the Language of Linear Algebra, Society for Industrial and Applied Mathematics, 2011.

ST 050103 - REGRESSION ANALYSIS

Objectives: To introduce the various concepts and techniques in regression analysis for modelling data and forecasting future values.

Outcomes: The students have studied simple linear regression, multiple regression, residual analysis for fitting a suitable model to a given data and to check the suitability. They have studied necessary transformations and modifications to be made when model assumptions are violated. They are capable of fitting logistic and Poisson models, non-linear and polynomial models.

Module 1

Introduction to regression analysis: overview and applications of regression analysis, major steps in regression analysis. Simple linear regression (Two variables): assumptions, estimation and properties of regression coefficients, significance and confidence intervals of regression coefficients, measuring the quality of the fit. Effect of outliers.

Module 2

Multiple linear regression model: assumptions, ordinary least square estimation of regression coefficients, interpretation and properties of regression coefficient, significance and confidence intervals of regression coefficients. Mean Square error criteria, coefficient of determination, Residual analysis, various types of residuals, Departures from underlying assumptions, Departures from normality. Multi-collinearity, sources, effects, diagnostics.

Module 3

Need for transformation of variables; Box-Cox transformation, removal of heteroscedasticity and serial correlation, Leverage and influence (concept only). Generalized least squares and weighted least squares (without derivation). Polynomial regression models, subset regression, Forward, Backward and Stepwise procedures, indicator variables**,** stepwise regression.

Module 4

Introduction to nonlinear regression, linearity transformations, Least squares in the nonlinear case and estimation of parameters, Logistic regression, estimation and interpretation of parameters in a logistic regression model, Poisson regression, Generalized linear models (GLM), Prediction and estimation with the GLM, residual analysis in the GLM.

Lab Exercises using Python:

1. Descriptive statistics
2. Fitting of distribution - binomial, poisson, normal, exponential.
3. Drawing samples from different distributions - binomial, poisson, exponential, gamma, normal.

4. Parametric tests – One sample t-test, two sample independent t-test for equal and unequal variances, paired t test, ANOVA.
5. Non Parametric tests – Mann Whitney U test, Wilcoxon signed rank test, Kruskal Wallis test.
6. Correlation – Pearson correlation coefficient, Spearmans correlation coefficient.
7. Regression Analysis – Simple linear regression, Multiple linear regression, Logistic regression.
8. Checking for multicollinearity.
9. Outlier detection.
10. Normality test.

References

1. D. C Montgomery, E.A Peck and G.G Vining (2003). *Introduction to Linear Regression Analysis*, John Wiley and Sons, Inc.NY,

2. S. Chatterjee and A. Hadi (2013) *Regression Analysis by Example*, 5th Ed., John Wiley and Sons.

3. Seber, A.F. and Lee, A.J. (2003) *Linear Regression Analysis*, John Wiley, Relevant sections from

4. Iain Pardoe (2012) *Applied Regression Modelling*, John Wiley and Sons, Inc,.

5. P. McCullagh, J.A. Nelder,(1989) *Generalized Linear Models*, Chapman & Hall,. John O. Rawlings,

6. Sastry G. Pantula, David A. Dickey (1998) *Applied Regression Analysis*, Second Edition, Springer.

7. Draper, N. and Smith, H. (2012) *Applied Regression Analysis* – John Wiley & Sons

ST 050104 -DATA BASE TECHNOLOGY

Objectives: To understand the basic concepts and the applications of database systems.

Outcomes: (i) Students understood the basics of SQL and can construct queries using SQL. (ii) Understood the relational database design principles and the basic issues of transaction processing and concurrency control. (iv) Understood database storage structures and access techniques. (v) Understood object oriented databases, data warehousing and OLAP tools. (vi)Understood MongoDB and can evaluate the NoSQL databases.

Module 1

Introduction to File and Database systems- History- Advantages, disadvantages- Data views – Database Languages – DBA – Database Architecture – Data Models- Keys – Mapping Cardinalities, Relational Algebra and calculus – Query languages – SQL – Data definition – Queries in SQL – Updates– Views – Integrity and Security –Embedded SQL

Suggested Lab Exercise:

1. Data Definition, Table Creation, Constraints
2. Insert, Select, Update & Delete Commands
3. Nested Queries, Join Queries, Views

Module 2

Design Phases – Pitfalls in Design – Attribute types –ER diagram – Database Design for Banking Enterprise – Functional Dependence – Normalization (1NF, 2NF, 3NF, BCNF, 4NF, 5NF). File Organization – Organization of Records in files – Indexing and Hashing.

Transaction concept – state serializability – Recoverability- Concurrency Control – Locks- Two Phase locking – Deadlock handling – Transaction Management in Multi Databases.

Module 3

Object-Oriented Databases- OODBMS- rules – ORDBMS- Complex Data types – Distributed databases –characteristics, advantages, disadvantages, rules- Homogenous and Heterogeneous Distributed data Storage – XML – Structure of XML Data – XML Document, Introduction to Mongo DB, Overview of NoSQL.

Suggested Lab Exercise:

1. Create a database and collection using MongoDB environment.

Module 4

Introduction to data warehousing, evolution of decision support systems –Modeling a data warehouse, granularity in the data warehouse – Data warehouse life cycle, building a data warehouse, Data Warehousing Components, Data Warehousing Architecture – On Line Analytical Processing, Categorization of OLAP Tools

References

1. Silberschatz, A., Korth, H. F., & Sudarshan, S. (1997). Database System Concepts (Vol. 4). New York: McGraw-Hill.
2. Pratt, P. J. & Adamski, J. J. (2011). Database Systems: Management and Design. Boyd & Fraser Pub. Co.
3. James R Groff and Paul N Weinberg (2003) The Complete Reference SQL –, Second Edition, Tata McGraw Hill,
4. Elmasri, R.& Navathe, S. (2010). Fundamentals of Database Systems. Addison-Wesley Publishing Company.)

ST 050105 PROGRAMMING AND DATA STRUCTURES WITH PYTHON

Objectives: To develop a thorough understanding in data analytics using Python.

Outcomes:(i) Demonstrate the usage of built-in objects in Python (ii) Analyze the significance of python program development environment by working on real world examples(iii) Implement numerical programming, data handling and visualization through NumPy, Pandas and MatplotLib modules.

Module 1

Structure of Python Program- Module Execution-Branching and Looping: Problem Solving Using Branches and Loops-Functions – Lists and Mutability- Problem Solving Using Lists and Functions. Sequences, Mapping and Sets- Dictionaries- -Classes: Classes and Instances-Inheritance-Exceptional Handling-Introduction to Regular Expressions using ―'re' module.

Suggested Lab Exercise:

1. Demonstrate usage of branching and looping statements
2. Demonstrate Recursive functions
3. Demonstrate Lists, Tuples, Sets and Dictionaries
4. Demonstrate inheritance and exceptional handling
5. Demonstrate use of "re", OOP

Module 2

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

Suggested Lab Exercise:

1. Aggregation in NumPy
2. Indexing and Sorting in NumPy

Module 3

Introduction to Pandas Objects- Data indexing and Selection-Operating on Data in Pandas-Handling Missing Data-Hierarchical Indexing – Combining Data Sets. Aggregation and Grouping-Pivot Tables-Vectorized String Operations –High Performance Pandas - eval () and query ()

Suggested Lab Exercise:

1. Demonstrate handling of missing data in Pandas
2. Demonstrate hierarchical indexing in Pandas

Module 4

Basic functions of matplotlib –Simple Line Plot, Scatter Plot-Density and Contour Plots-Histograms, Binnings and Density-Customizing Plot Legends, Colour Bars- Three- Dimensional Plotting in Matplotlib.

Suggested Lab Exercise:

1. Demonstrate various graphs like Scatter Plot
2. Demonstrate 3D plotting using Matplotlib

References

1. Joel Grus (2016) Data Science from Scratch First Principles with Python, O‘Reilly Media.
2. T.R.Padmanabhan(2016) Programming with Python, Springer Publications.
3. Majumdar, A. K., &Bhattacharyya, P. (1996). Database Management Systems. McGraw-Hill.
4. Jake Vander Plas ,Python Data Science Handbook – Essential Tools for Working with Data, O‘Reilly Media,Inc, 2016
5. Zhang.Y. , An Introduction to Python and Computer Programming, Springer Publications, 2016
6. Wes McKinney, (2017) Python for Data Analysis: Data Wrangling with Pandas, NumPy, and Ipython, 2nd Edition, O‘Reilly Media.
7. Haslwanter, T.(2015) An Introduction to Statistics with Python, Springer

SEMESTER II COURSES

ST 050201 MATHEMATICAL FOUNDATION FOR DATA ANALYTICS 2

Objectives: This course aims at introducing data analytics related essential mathematics concepts such as fundamentals of topics on Calculus of several variables, Graph, Finite Difference Method, Numerical methods

Outcomes:(i)Demonstrate the properties of multivariate calculus , (ii) Know the basic terminologies and properties in Graph Theory (iii) Apply various interpolation methods and finite difference concepts (iv) Apply numerical methods to find solution of algebraic equations .

Module 1

Limits: Limits of Functions, Limit Theorems, Some Extensions of the Limit Concept. Continuous Functions: Continuous Functions, Combinations of Continuous Functions, Continuous Functions on Intervals, Uniform Continuity, Continuity and Gauges Monotone and Inverse Functions, Differentiation: The Derivative, The Mean Value Theorem, L'Hospital's Rules, Taylor's Theorem, Partial Derivatives: Functions of Several Variables, Limits and Continuity in Higher Dimensions, Partial Derivatives, The Chain Rule, Directional Derivatives and Gradient Vectors, Extreme Values and Saddle Points. (Without proof)

Module 2

Graph Classes: Definition of a Graph and Graph terminology, isomorphism of graphs, Complete graphs, bipartite graphs, complete bipartite graphs - Vertex degree: adjacency and incidence, regular graphs - subgraphs, spanning subgraphs, induced subgraphs, removing or adding edges of a graph, removing vertices from graphs - Graph Operations: Graph Union, intersection, complement, self-complement, Paths and Cycles, Connected graphs, Matrix Representation of Graphs, Adjacency matrices, Incidence Matrices, Trees and its properties, Bridges (cut-edges), spanning trees, weighted Graphs, minimal spanning tree problems, Shortest path problems, cut vertices, cuts, vertex and edge connectivity, Eulerian and Hamiltonian Graphs. (Without proof)

Module 3

Interpolation: Introduction, Errors in polynomial interpolation, Finite differences, Newton's formula for interpolation. Gauss Center difference interpolation formula, Lagrange's interpolation formula, Newton's general interpolation formula.

Module 4

Solutions of Algebraic and transcendental Equations: Introduction, The bisection method, The iteration method, The method of False position, Newton-Raphson method, Numerical Differentiation and Integration: Introduction, Numerical Differentiation, Numerical integration, Trapezoidal rule, Simpson's 1/3 rd rule, Simpson's 3/8 rule.

References

1. Robert G. Bartle and Donald R. Sherbert - Introduction to Real Analysis,Fourth Edition, John Wiley & Sons, Inc.
2. Thomas , G.B Jr, Weir, M.D and Hass J - Thomas Calculus, 12th edition, Pearson.
3. J Clark, D A Holton, A first look at Graph Theory, Allied Publishers India, 1995. (Unit 4)
4. S. S. Sastry - Introductory Methods of Numerical Analysis, Fourth Edition, Prentice Hall of India.
5. Tom W. Apostol: Calculus: One-Variable Calculus with An Introduction to Linear Algebra, Vol 1, Wiley, Second edition (2007).

ST 050202 MULTIVARIATE ANALYSIS

Objectives: To introduce multivariate data and associated concepts, distributions, testing and estimation, and the theory and applications of discriminant function and classification rules, principal components, canonical correlations, factor analysis etc.

Outcomes: After undergoing this course students can apply multivariate techniques such as discriminant function and classification rules, principal components, canonical correlations, factor analysis, MANOVA etc. They are enabled to apply Hotelling's T^2 and Mahalanobis D^2 etc for testing hypotheses in the case of multivariate data.

Module 1

Basic concepts on multivariate variable. Concept of random vector: Its expectation and dispersion (Variance-Covariance) matrix. Marginal and joint distributions. Conditional distributions and Independence of random vectors. Multivariate normal distribution, Marginal and conditional distribution, additive property, mle.s of mean and dispersion matrix.

Module 2

Sample mean vector and its distribution, Hotelling's T2 and Mahalanobis' D2 statistics and applications. Tests of hypotheses about the mean vectors and covariance matrices for multivariate normal populations. Wishart distribution, Independence of sub vectors and sphericity test.

Module 3

Fisher's criteria for discrimination and classification for two populations, Sample discriminant function. Expected cost of misclassification, classification with two multivariate normal population, classification with several multivariate normal populations. Multivariate analysis of variance (MANOVA) of one and two- way classified data. Multivariate analysis of covariance (concept only).

Module 4

Principal components, sample principal components asymptotic properties. Canonical variables and canonical correlations: definition and estimation. Factor analysis: Orthogonal factor model, factor loadings, estimation of factor loadings, factor scores, cluster analysis-agglomerative and divisive techniques.

Lab Exercises:

1. Computation of Means, Variance, Covariances and Correlations from a multivariate dataset
2. Application of T^2 Statistics to different situations – Test for mean of a single MV normal population, test of equality of mean vectors of two MV normal populations with equal var-cov matrices and unequal var-cov matrices.
3. Manova – One-way & Two-way models.
4. Principal Component Analysis.

5. Factor Analysis.
6. Canonical Correlation Analysis.
7. Fisher’s Discriminant Analysis – Two populations, several populations, classification with prior probabilities.
8. Cluster Analysis.

References

1. Chatfield, C. (2018). *Introduction to multivariate analysis*. Routledge.

2. Rencher, A. C. (2012) *Methods of Multivariate Analysis*.(3rd ed.) John Wiley.

3. Johnson R.A. and Wichern D.W. (2008) *Applied Multivariate Statistical Analysis*. 6th Edition, Pearson Education.

4. Anderson, T.W. (2009). *An Introduction to Multivariate Statistical Analysis*, 3rd Edition, John Wiley.

5. Everitt B, Hothorn T,(2011). *An Introduction to Applied Multivariate Analysis with R*, Springer.

6. Barry J. Babin, Hair, Rolph E Anderson, & William C. Blac, (2013), *Multivariate Data Analysis*, Pearson New International Edition,

ST 050203 STOCHASTIC PROCESSES & TIME SERIES ANALYSIS

Objectives: To introduce the basics of stochastic processes and modelling as well as enable the students to analyse time series data and apply suitable techniques to model them and forecast future values.

Outcomes: Students are aware of various stochastic models and time series models and can apply these to model data for predicting future values to make appropriate planning and decision making.

Module 1

Introduction to stochastic processes:- classification of stochastic processes, wide sense and strict sense stationary processes, processes with stationary independent increments, Markov process, Markov chains- transition probability matrices, Chapman-Kolmogorov equation, first passage probabilities, recurrent and transient states, mean recurrence time, stationary distributions, limiting probabilities, Random walk.

Module 2

Continuous time Markov chains, Poisson processes, properties, inter-arrival time distribution pure birth processes and the Yule processes, birth and death processes, linear growth process with immigration, steady-state solutions of Markovian queues - M/M/1, M/M/s, M/MM/∞ models, Renewal processes – concepts only, examples, Poisson process viewed as a renewal process.

Module 3

Time series data, examples, Time series as stochastic process, Additive and multiplicative models, stationary time series- covariance stationarity, Modelling Time Series Data, Exponential Smoothing Methods - First-Order Exponential Smoothing, Second Order Exponential Smoothing, Forecasting, Exponential Smoothing for Seasonal Data, Exponential Smoothers.

Module 4

Time series modelling, Autocorrelation function (ACF), partial auto correlation function (PACF), correlogram, AR, MA, ARMA, ARIMA Models, Yule- Walker equations, Forecasting future values.

Lab Exercises:

1. Transition Probability Matrix
2. Classification of different states of Markov chain – aperiodic, ergodic,
3. Estimating probability after n stages from initial state probability.
4. Plotting time series data.
5. Decomposing time series data - seasonal and non seasonal.
6. Simple Exponential Smoothing
7. Holt-Winter's Exponential Smoothing.

8. Fitting of ARIMA model.
9. Forecasting for future time point.

References

1. Medhi J. (2017) *Stochastic Processes*, Second Edition, Wiley Eastern, New Delhi

2. Ross S.M. (2007) *Stochastic Processes*. Second Edition, Wiley Eastern, New Delhi

3. Montgomery D. C., Cheryl L. J., and Murat K. (2015) *Introduction to Time Series Analysis and Forecasting*. John Wiley & Sons.

4. Brockwell P.J and Davis R.A. (2002) Introduction to Time Series and Forecasting Second edition, Springer-Verlag.

5. Abraham, B., &Ledolter, J. (2009). *Statistical methods for forecasting* (Vol. 234). John Wiley & Sons.

6. Chatfield, C.(2004).*The Analysis of Time Series - An Introduction (Sixth edition*), Chapman and Hall.

ST 050204 DATA VISUALIZATION

Objectives: To make students familiar with data visualization techniques for displaying data.

Outcomes: Students are now able to use visualization techniques for multidimensional visualization, information visualization applications and systems, visualization packages, grammar of graphics using R etc.

Module 1

Purpose of visualization, visual perception, cognitive issues –other theory and design principles behind information visualization

Module 2

Multidimensional visualization, tree visualization, graph visualization, Time series data visualization techniques.

Module 3

Basic interaction techniques such as selection and distortion, evaluation, Examples of information visualization applications and systems, user tasks and analysis- visualization packages

Module 4

Grammar of graphics using R-Construct/Deconstruct a graphic into a data- order of accuracy of perceptual tasks and its impact and Case study presentations and lab based on R package of Data Visualizations.

Suggested Lab Exercise:

1. Practicals involving visualizing data in different format, using ggplot.

References

1. Wickham, H. (2016). Ggplot2: Elegant Graphics for Data Analysis. Springer.2nd Edition
2. Ward,M., Grinstein, G. , Keim, D. Interactive Data Visualization - Foundations, Techniques, and Applications - Second Edition, CRC press
3. Keen, K. J. (2010). Graphics for Statistics and Data Analysis with R. CRC Press.
4. Buja, A., Swayne, D. F. & Cook, D., (2007). Interactive and Dynamic Graphics for Data Analysis: with R and Ggobi. Springer Science & Business Media.
5. Murrell, P. (2016). R graphics. CRC Press.
6. Cleveland, W. S. (1993). Visualizing data. Hobart Press.
7. Tufte, E. R., Goeler, N. H., & Benson, R. (1990). Envisioning information (Vol. 126). Cheshire, C.T: Graphics press.
8. Tufte, E., & Graves-Morris, P. (2014). The visual display of quantitative information.; 1983.

ST 050205 PROGRAMMING USING R

Objectives: To make students aware of R commands and programming and to impart training in R for Data Analytics using various techniques.

Outcomes: Students have understood the various commands in R and can write programs in R. They have experienced the importance of R in Data Analytics and can apply this for Data Analytics.

Module 1

Why R – Getting started with R – Vectors and Data Frames – Loading Data Frames – Data analysis with summary statistics and scatter plots – Summary tables - Working with Script Files

Linear Regression – Introduction – Regression model for one variable regression – Selecting best model – Error measures SSE, SST, RMSE, R2 – Interpreting R2 – Multiple linear regression – Loess and ridge regression – Correlation.

Logistic Regression – The Logit – Confusion matrix – sensitivity, specificity – ROC curve – Threshold selection with ROC curve – Making predictions – Area under the ROC curve (AUC)

Suggested Lab Exercise:

1. Basic R programs, reading files
2. Linear regression and interpretation
3. Logistic regression and interpretation with ROC curve.

Module 2

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

Suggested Lab Exercise:

1. CART with and without cross validation
2. Random Forest

Module 3

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

Suggested Lab Exercise:

1. Text data classification
2. Time series analytics
3. k-Means clustering

Module 4

Support Vector Machines – Gradient Boosting – Naïve Bayes – Bayesian GLM – GLMNET – Ensemble modeling – Experimenting with all of the above approaches with and without data imputation and assessing predictive accuracy.

Suggested Lab Exercise:

1. Demonstration of SVM, Naïve Bayes

References

1. Michael J. Crawley, Statistics : An Introduction Using R, WILEY, Second Edition, 2015.
2. Garrett Grolemund ,Hands-on programming with R, O'Reilley, 1st Edition, 2014
3. Jared Lander , R for everyone, Pearson, 1st Edition, 2014
4. Seema Acharya, Data Analytics using R , McGraw Hill Publications
5. Nina Zumel, John Mount, Practical Data Science with R, Manning Publications

SEMESTER III COURSES

ST 050301 SAMPLING AND DESIGN OF EXPERIMENTS

Objectives: To make the students familiar with different sampling schemes and their advantages as well as their applications in estimating population mean, total, proportion etc. It is also expected to impart knowledge on basic principles of experimentation, different designs like CRD,RBD, LSD, BIBD, Factorial experiments etc.

Outcomes: (i) After undergoing this course, students are aware of different sample survey methods and are capable of planning and implementing sample surveys, consumer satisfaction surveys, public opinion surveys etc.(ii) they are aware of different designs in experimentation like CRD, RBD, LSD, BIBD, Factorial Designs, etc. and can apply ANOVA technique to analyse the data using Python or R.

Module 1

Census and sampling methods, probability sampling and non-probability sampling, principal steps in sample surveys, sampling errors and non-sampling errors, bias, variance and mean square error of an estimator, simple random sampling with and without replacement, estimation of the population mean, total and proportions, properties of the estimators, variance and standard error of the estimators, confidence intervals, determination of the sample size. Stratified random sampling, estimation of the population mean, total and proportion, properties of estimators, various methods of allocation of a sample, comparison of the precisions of estimators under proportional allocation, optimum allocation and SRS. Systematic sampling.

Module 2

Ratio method of estimation, estimation of population ratio, mean and total, Bias and relative bias of ratio estimator, comparison with SRS estimation. Regression method of estimation. Comparison of ratio and regression estimators with mean per unit method, Cluster sampling, single stage cluster sampling with equal and unequal cluster sizes, estimation of the population mean and its standard error. Multistage and Multiphase sampling (Basic Concepts),

Module 3

Standard Gauss Markoff set up, estimability of parameters, method of least squares, best linear unbiased Estimators, Gauss – Mark off Theorem, tests of linear hypotheses. Planning of experiments, Basic principles of experimental design, uniformity trails, analysis of variance, one-way, two-way and three-way classification models, completely randomized design (CRD), randomized block design (RBD) Latin square design (LSD). Analysis of covariance (ANCOVA)

Module 4

Balanced incomplete block design (BIBD); incidence Matrix, parametric relation; intrablock analysis of BIBD, basic ideas of partially balanced incomplete block design (PBIBD). Factorial experiments, 2n and 3n factorial experiments, Yates procedure, confounding in factorial experiments, basic ideas of response surface designs.

Lab Exercises:

1. Simple Random Sampling With and Without Replacement.
2. Probability proportion to size sampling.
3. Stratified Random Sampling.
4. Sample size determination.
5. Complete Randomized Design
6. Randomized Block Design
7. Latin Square Design.
8. Balanced Incomplete Block Design.
9. Factorial Design - $2^2, 2^3, 3^2$
10. ANOCOVA.

References

1. Cochran W. G. (1999) *Sampling Techniques*, 3rd edition, John Wiley and Sons.

2. Mukhopadyay P. (2009) *Theory and Methods of Survey Sampling*, 2nd edition, PHL, New Delhi.

3. Aloke Dey (1986) *Theory of Block Designs*, Wiley Eastern, New Delhi.

4. Das M.N. and Giri N.C. (1994) *Design and analysis of experiments*, Wiley Eastern Ltd.

5. Arnab, R. (2017). *Survey Sampling: Theory and Applications*. Academic Press.

6. Montgomery, C.D. (2012) *Design and Analysis of Experiments*, John Wiley, New York.

7. Sampath S. C. (2001*) Sampling Theory and Methods*, Alpha Science International Ltd.,

8. Thomas Lumley (1996) *Complex Surveys. A Guide to Analysis Using R*, Wiley eastern Ltd.

9. Des Raj (1967*) Sampling Theory*. Tata McGraw Hill ,NewDelhi

10. Dean, A. and Voss, D. (1999) *Design and Analysis of Experiments*, Springer Texts in Statistics

ST 050302 OPTIMIZATION TECHNIQUES

Objectives: To make the students familiar with modern optimization techniques.

Outcomes: (i) Apply the notions of linear programming in solving transportation problems (2)Understand the theory of games for solving simple games (3)Use linear programming in the formulation of shortest route problem. (4)Apply algorithmic approach in solving various types of network problems (5)Create applications using dynamic programming. .

Module 1

Linear programming: Mathematical Model, assumptions of linear programming, Solutions of linear programming problems – Graphical Method, Simplex method, Artificial Variable Method, Two phase Method, Big M Method, Applications, Duality, Dual simplex method,

Module 2

Special types of Linear programming problems- Transportation Problem (TP) – Mathematical formulation of Transportation Problem, Basic feasible solution in TP, Degeneracy in TP, Initial basic feasible solutions to TP, Matrix Minima Method, Row Minima Method, Column Minima Method, Vogel‘s Approximation Method, Optimal Solution to TP, MODI Method, Stepping Stone Method, Assignment problems – Definition, Hungarian Method

Module 3

Integer Programming: Pure Integer Programming, Mixed Integer Programming, Solution Methods – Cutting plane method, branch and bound method. Binary Integer Linear programming- Travelling salesman problems – Iterative method, Branch and bound method;

Module 4

Dynamic programming: Deterministic and Probabilistic Dynamic programming. Linear programming by dynamic programming approach.

References

1. Ravindran, A., Philips, D.T Solberg, J. J. (2007) Operations Research: Principles and Practice, Wiley
2. J. K. Sharma (2009), Operations Research – Theory and Applications, 4th Ed, Mc Millan Publishing.
3. Taha, H.A. (2007), Operations Research, 8th Ed., Mc Millan Publishing Company
4. Kantiswaroop, P. K. Guptha, & Manmohan (2007) Operations Research, 13th Ed, Sulthan Chand & Sons.
5. Beightler C. S, & Philips D T (2009), _Foundations of Optimization‘, 2nd Ed., Prentice Hall.
6. Mc Millan Claude Jr (1979), _Mathematical Programming‘, 2nd Ed. Wiley Series.

ST 050303 MACHINE LEARNING

Objectives: To familiarize the students with various aspects of machine learning techniques and their applications.

Outcomes: The students have understood different techniques such as unsupervised learning, dimensionality reduction, PCA, SVM, Discriminant function, multilayer preceptors, cluster analysis etc

Module 1

Machine Learning-Examples of Machine Applications – Learning Associations – Classification-Regression- Unsupervised Learning- Reinforcement Learning. Supervised Learning: Learning class from examples- Probably Approximately Correct (PAC) Learning- Noise-Learning Multiple classes. Regression – Model Selection and Generalization.

Introduction to Parametric methods- Maximum Likelihood Estimation: Bernoulli Density-Multinomial Density-Gaussian Density, Nonparametric Density Estimation: Histogram Estimator-Kernel Estimator-K-Nearest Neighbor Estimator.

Module 2

Dimensionality Reduction: Introduction- Subset Selection- Principal Component Analysis, Feature Embedding-Factor Analysis- Singular Value Decomposition- Multidimensional Scaling- Linear Discriminant Analysis- Bayesian Decision Theory. Linear Discrimination: Introduction- Generalizing the Linear Model- Geometry of the Linear Discriminant- Pairwise Separation-Gradient Descent-Logistic Discrimination. Optical separating hyper plane – v-SVM, kernel tricks – vectorian kernel-defining kernel- multiclass kernel machines- one-class kernel machines.

Module 3

Multilayer perceptron, Introduction, training a perceptron- learning Boolean functions- multilayer perceptron- back propagation algorithm- training procedures. Combining Multiple Learners, Rationale-Generating diverse learners- Model combination schemes- voting, Bagging- Boosting- fine tuning an Ensemble.

Module 4

Cluster Analysis, Introduction-Mixture Densities, K-Means Clustering- Expectation-Maximization algorithm- Mixtures of Latent Variable Models-Supervised Learning after Clustering-Spectral Clustering- Hierarchical Clustering- Divisive Clustering- Choosing the number of Clusters.

References

1. E. Alpaydin (2014) Introduction to Machine Learning, 3rd Edition, MIT Press.
2. Frank Kane (2012) Data Science and Machine Learning. Manning Publications.
3. C.M.Bishop, Pattern Recognition and Machine Learning, Springer.
6. T. Hastie, R. Tibshirani and J. Friedman (2016) The Elements of Statistical Learning: Data Mining, Inference and Prediction, Springer, 2nd Edition,2009.

7. Alex Berson & Stephen J. Smith (1997) Data Warehousing, Data Mining & OLAP Computing, Mc Graw Hill
8. Data Mining Techniques: A.K. Pujari, Universities Press, 2001
9. Mastering Data Mining: M. Berry and G. Linoff, John Wiley & Sons., 2000

ST 050304 BIG DATA ANALYTICS AND HADOOP

Objectives: To make the students aware about the different techniques for big data analytics.

Outcomes: After undergoing this course students are enabled to use Hadoop, RDBMS, Mapreduce, HDFS, HIVE & PIG etc for big data analytics.

Module 1

Distributed file system – Big Data and its importance, Four Vs, Drivers for Big data, Big data analytics, Big data applications, Algorithms using map reduce, Matrix-Vector Multiplication by Map Reduce.

Apache Hadoop– Moving Data in and out of Hadoop – Understanding inputs and outputs ofMapReduce – Data Serialization, Problems with traditional large-scale systems-Requirements for a new approach- Hadoop – Scaling-Distributed Framework- Hadoop v/s RDBMS-Brief history of Hadoop.

Lab Exercise

1. Word count application in Hadoop.

2. Sorting the data using MapReduce.

3. Finding max and min value in Hadoop.

Module 2

Understanding MapReduce: Key/value pairs, Hadoop Python API for MapReduce, Writing MapReduce programs, Hadoop-specific data types, Input/output. Developing MapReduce Programs: Analysing a large dataset. Advanced Mapreduce Techniques: Simple, advanced, and in-between Joins, Graph algorithms, using language-independent data structures. Hadoop configuration properties – Setting up a cluster, Cluster access control, managing the NameNode, Managing HDFS, MapReduce management, Scaling.

Lab Exercise:

1. Implementation of decision tree algorithms using MapReduce.

2. Implementation of K-means Clustering using MapReduce.

3. Generation of Frequent Itemset using MapReduce.

Module 3

Hadoop Streaming - Streaming Command Options – Specifying a Python Class as the Mapper/Reducer – Packaging Files With Job Submissions – Specifying Other Plug-ins for Jobs.

HIVE & PIG ; Architecture, Installation, Configuration, Hive vs RDBMS, Tables, DDL & DML, Partitioning & Bucketing, Hive Web Interface, Pig, Use case of Pig, Pig Components, Data Model, Pig Latin.

Lab Exercise:

1. Count the number of missing and invalid values through joining two large given datasets.

2. Using hadoop‘s map-reduce, Evaluating Number of Products Sold in Each Country in the online shopping portal. Dataset is given.

3. Analyze the sentiment for product reviews, this work proposes a MapReduce technique provided by Apache Hadoop.

4. Trend Analysis based on Access Pattern over Web Logs using Hadoop.

5. Service Rating Prediction by Exploring Social Mobile Users Geographical Locations.

Module 4

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

Lab Exercise:

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

References:

1. Boris lublinsky, Kevin t. Smith, Alexey Yakubovich, Professional Hadoop Solutions, Wiley, 2015.

2. Tom White, Hadoop: The Definitive Guide, O‘Reilly Media Inc., 2015.

3. Garry Turkington, Hadoop Beginner‘s Guide, Packt Publishing, 2013.

4. Pethuru Raj, Anupama Raman, DhivyaNagaraj and Siddhartha Duggirala, High-Performance Big-Data Analytics: Computing Systems and Approaches, Springer, 2015.

5. Jonathan R. Owens, Jon Lentz and Brian Femiano, Hadoop Real-World Solutions Cookbook, Packt Publishing, 2013.

SEMESTER IV COURSES

ELECTIVES - BUNCH 1

ST 900401 ARTFICIAL INTELLIGENCE

Objectives: To understand thinking and intelligence in ways that enable the construction of computer systems that are able to reason in uncertain environments.

Outcomes: able to articulate and exemplify the basic knowledge artificial intelligence, Understand the basics of knowledge representation, can use AI programming languages and the methods of AI implementation and can recommend AI strategies based on applications.

Module 1

Artificial Intelligence - Introduction, AI Problems, AI Techniques, The Level of the Model, Criteria For Success. Defining the Problem as a State Space Search, Problem Characteristics, Production Systems, Search: Issues in The Design of Search Programs, Un-Informed Search, BFS, DFS; Heuristic Search Techniques: Generate-And-Test, Hill Climbing, Best-First Search, A*Algorithm, Problem Reduction, AO*Algorithm, Constraint Satisfaction, Means-Ends Analysis.

Knowledge Representation: Procedural Vs Declarative Knowledge, Representations & Approaches to Knowledge Representation, Forward Vs Backward Reasoning, Matching Techniques, Partial Match - ing, Fuzzy Matching Algorithms and RETE Matching Algorithms;

Module 2

Logic Based Programming-AI Programming languages: Overview of LISP, Search Strategies in LISP, Pattern matching in LISP , An Expert system Shell in LISP, Over view of Prolog, Production System using Prolog; Symbolic Logic: Propositional Logic, First Order Predicate Logic: Representing Instance and Relationships, Computable Functions and Predicates, Syntax & Semantics of FOPL, Normal Forms, Unification &Resolution, Representation Using Rules, Natural Deduction; Structured Representations of Knowledge: Semantic Nets, Partitioned Semantic Nets, Frames, Conceptual Dependency, Conceptual Graphs, Scripts, CYC;.

Module 3

Reasoning under Uncertainty: Introduction to Non-Monotonic Reasoning, Truth Maintenance Systems, Logics for non-monotonic Reasoning, Model and Temporal Logics; Statistical Reasoning: Bayes‘ Theorem, Certainty Factors and Rule-Based Systems, Bayesian Probabilistic Inference, Bayesian Networks, Dempster -Shafer Theory, Fuzzy Logic: Crisp Sets ,Fuzzy Sets, Fuzzy Logic Control, Fuzzy Inferences & Fuzzy Systems.

Module 4

Experts Systems: Overview of an Expert System, Structure of an Expert Systems, Different Types of Expert Systems-Rule Based, Model Based, Case Based and Hybrid Expert Systems, Knowledge Ac - quisition and Validation Techniques, Black Board Architecture, Knowledge Building System Tools, Expert System Shells, Fuzzy Expert systems.

References

1. George F Luger (2016) Artificial Intelligence, Pearson Education Publications

2. Elaine Rich and Knight (2017) Artificial Intelligence, Mcgraw-Hill Publications

3. Patterson, D.W.(2005) Introduction to Artificial Intelligence & Expert Systems, PHI

4. Weiss.G, (2000) Multi Agent Systems- A Modern Approach to Distributed Artificial Intelligence, MIT Press.

5. Russell S. and Norvig, P.(2010) Artificial Intelligence : A modern Approach, Printice Hall

ST 900402 EPIDEMIOLOGY AND CLINICAL TRIALS

Objectives: To impart basic knowledge & skills in Controlled Clinical Trials & & their applications

Outcomes: On completion of the course, students should be able to understand basic concepts of clinical trials

Module 1

Introduction to epidemiology, Exposure and outcome, Measures of Disease occurrence - Prevalence and Incidence, Study designs- cross-sectional studies, Case control study-designing a case control study, matching, cohort studies, design of a cohort study. Relative and absolute measures of effect.

Module 2

Censoring-right and left, Functions of survival time- Kaplan Meier (K-M) estimator, Nonparametric Methods for Comparing Survival Distributions - log rank test, Parametric distribution: exponential and Weibull, Cox's proportional hazards model- time dependent covariates.

Module 3

Introduction to clinical trials, ethical issues, protocols, comparative and controlled trials. Objectives and End-points of a clinical trial, Single center and Multi-center trials, ICH and GCP, FDA and EMEA guidelines, Drug Development Process, Overview of phase I-IV trials (Design and analysis), Concept of Blinding in clinical trials, bias and random error in clinical studies. Clinical trial study designs. parallel vs. cross-over designs, Longitudinal study, Bioequivalence trials Adaptive trials, Sample size determination. Randomization methods, Handling of missing data, Handling multiplicity

Module 4

Clinical data Management (CDM),Understanding protocol, Clinical study report, Statistical analysis plan(SAP), Data visualization methods. Data Comprehension, Data Interpretation, Clinical Data Analysis: Analysis methods/models for Continuous data, Categorical data Binary data, Survival data, Parametric and Nonparametric methods, Sub-group Analysis, Sensitivity analysis, Interim analysis, Quality of life data analysis, Meta Analysis

Reference

1. Friedman L.M.,Furberg C.D. &Demets D.L.(1998). Fundamentals of clinical trials, Springer
2. Scott Evans , Naitee Ting, Fundamental Concepts for New Clinical Triallists, Chapman & Hall Book

3. Shein-Chung Chow and Jen-Pei Liu(2004). Design and Analysis of Clinical Trials: Concepts and Methodologies (2nd edition) Wiley-Interscience
4. Stuart J. Pocock (2010)Clinical Trials – A practical approach (Reprint), John Wiley
5. Stephen Senn (2009) Statistical Issues in Drug Development (2nd edition), John Wiley
6. David Collett (2003) Modeling Binary Data (2nd edition), Chapman & Hall/CRC
7. Alan Agresti (2002) Categorical Data Analysis (2nd edition), Wiley-Interscience
8. Atkinson AC and Biswas A - Randomised Response-Adaptive Designs in Clinical Trials
9. Gordis, Leon (2014) – Epidemiology, Fifth edition, Elsevier Saunders
10. Lee E.T., Wang, John Wenyu Hoboken (2003) Statistical Methods for Survival Data Analysis, 3rd Edition, Wiley Inter science

ELECTIVES - BUNCH 2

ST 910401 CLOUD COMPUTING

Objectives: To introduce recent trends in cloud computing, architecture and service models like XaaS, IaaS, SaaS, PaaS and illustrate through case studies.

Outcomes: Students have gained knowledge in recent trends in cloud computing, architecture, service models, platform and storage and web services. Now they are enabled to apply these techniques to real problems.

Module 1

Overview of Computing Paradigm, Recent trends in Computing Grid Computing, Cluster Computing, Distributed Computing, Utility Computing, Cloud Computing Evolution of cloud computing Business driver for adopting cloud computing. Introduction to Cloud Computing. Cloud Computing (NIST Model) Introduction to Cloud Computing, History of Cloud Computing, Cloud service providers. Properties, Characteristics & Disadvantages Pros and Cons of Cloud Computing.

Module 2

Cloud Computing Architecture, Comparison with traditional computing architecture (client/server), Services provided at various levels, How Cloud Computing Works, Role of Networks in Cloud computing, protocols used, Role of Web services.

Module 3

Service Models (XaaS), Infrastructure as a Service(IaaS),Platform as a Service(PaaS), Software as a Service(SaaS), Deployment Models, Public cloud, Private cloud, Hybrid cloud, Community cloud. Introduction to virtualization, Different approaches to virtualization, Hypervisors, Machine Image Virtual Machine (VM). Examples, Amazon EC2, Renting, EC2 Compute Unit,

Module 4

Platform and Storage, pricing, customers Eucalyptus, Platform as a Service(PaaS),Introduction to PaaS, What is PaaS, Service Oriented Architecture (SOA). Examples: Google App Engine, Microsoft Azure, Sales Force.com‘s Force.com platform; Software as a Service (SaaS): Introduction to SaaS, Web services, Web 2.0, Web OS, Case; Study on SaaS Cloud Security. Case Study on Open Source & Commercial Clouds

References

1. Barrie Sosinsky (2010) Cloud Computing Bible*, Wiley-India,*

2. Kirsh D. and Hurwitz J.(2018) Cloud Computing for Dummies 2nd Edition

3. Ruparelia, N.B.(2016) Cloud Computin, MIT Press

4. Orban S.(2019) Ahead in the Cloud: Best Practises for Navgating, AWS.

ST 910402 RELIABILITY MODELING AND STATISTICAL QUALITY CONTROL

Objectives : To understand reliability , to ensure the validity and precision of statistical analysis. Understand the concept of quality control statistical process control.

Outcomes: On completion of the course, students should be able to: Define reliability including the different types and how they assessed. Ensure the validity and precision of statistical analysis. Understand the concept of quality control statistical process control.

Module 1

Basic reliability concepts: Reliability concepts and measures, Components and systems, coherent systems, reliability of coherent systems, cuts and paths, series and parallel system, k-out-of-n systems, Bounds on System Reliability. Failure rate, mean residual life, Mean time to failure in the univariate cases, Exponential, Weibull, Pareto, Inverse Gaussian and Gamma as life distribution models, Characterization of life distribution based on failure rate and mean residual life function.

Module 2

Reliability concepts in discrete set up, Notion of ageing based on failure rate and mean residual life, NBU, NBUE, HNBUE classes and their duals, Interrelationships. Inference in reliability models: Estimation of parameters based on complete and censored samples in exponential, Weibull and Gamma models. Non-parametric estimation of failure rate and reliability function.

Module 3

Statistical process control, Theory of control charts – Shewart control charts for variables- X , R, S charts, Attribute control charts - np, p, c and u charts – OC, ARL & process capability of control charts, CUSUM charts, Acceptance sampling for attributes and variables.

Module 4

Sampling inspection techniques: Single, double and multistage sampling plans and their properties, Chain sampling, Continuous sampling, Taguchi method, Total quality management, ISO standardization, ISO 9001, six sigma concepts.

References

1. Barlow, R.E. and Proschan, F. (1985): Statistical Theory of Reliability and Life Testing, Holt, Rinehart and Winston.
2. Montgomery, D.C. (2009): Introduction to Statistical Quality Control, 8th edition, John Wiley.
3. Cox, D.R. and Oakes, D. (1984): Analysis of Survival Data, Chappman Hall.
4. Duncan, A. J. (1959): Quality Control and Industrial Statistics (5th edition), Irwin, Homewood I.

5. Galambos, J. and Kotz, S. (1978) Characterization of Probability Distributions.
6. Klefjo, B. (1982) The HNBUE and HNWUE Classes of Life distributions, Naval Research Logistic Quarterly, 29, 331-344.
7. Lawless, J. F. (2003): Statistical Models and Methods for Lifetime Data, John Wiley.
8. Nelson, W. (1982): Applied life data analysis, Wiley.
9. Sinha, S. K. (1986) Reliability and Life Testing, Wiley.

ELECTIVES - BUNCH 3

ST 920401 WEB ANALYTICS

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

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

Module 1

Introduction to Web Analytics: Web Analytics Approach – A Model of Analysis – Context matters – Data Contradiction – Working of Web Analytics: Log file analysis – Page tagging – Metrics and Dimensions – Interacting with data in Google Analytics . Goals: Introduction – Goals and Conversions – Conversion Rate – Goal reports in Google Analytics – Performance Indicators – Analyzing Web Users: Learning about users – Traffic Analysis – Analyzing user content – Click-Path analysis – Segmentation

Module 2

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

Module 3

Lab Usability Testing- Heuristic Evaluations- Site Visits- Surveys (Questionnaires) - Testing and Experimentation: A/B Testing and Multivariate Testing-Competitive Intelligence - Analysis Search Analytics: Performing Internal Site Search Analytics, Search Engine Optimization (SEO) and Pay per Click (PPC)-Website Optimization against KPIs- Content optimization- Funnel/Goal optimization - Text Analytics: Natural Language Processing (NLP)- Supervised Machine Learning (ML) Algorithms-API and Web data scarping using R and Python

Module 4

Drill down and hierarchies-Sorting-Grouping- Additional Ways to Group- Creating Sets- Analysis with Cubes and MDX- Filtering for Top and Top N- Using the Filter Shelf- The Formatting Pane- Trend Lines- Forecasting- Formatting- Parameters - Social Network Analysis: Types of social network-Graph Visualization-Network Relationships-Network structures: equivalence-Network

Evolution-Diffusion in networks- Descriptive Modeling-Predictive Modeling-Customer Profiling-Network targeting

Lab Exercises

1. Working concept of web analytics , 2. Evaluation with Intermediate metrics, custom metrics, calculated metrics. 3. Collection of web data and other internet data with the help of web analytics 4. Delivering reports based on collected data 5. Implement the concept of web analytics ecosystem 6. Creation of segmentation in web analytics 7. Visualization, acquisition and conversions of web analytics data 8. Performing site search analytics 9. Analyse the web analytic reports and visualizations 10. Performing visual web analytics 11. Assignments and final discussions 12. Web Analytics case studies

References

1. Beasley M, (2013*), Practical web analytics for user experience: How analytics can help you understand your users*. Newnes, 1st edition, Morgan Kaufmann.
2. Sponder M, (2013), *Social media analytics: Effective tools for building, interpreting, and using metrics*, 1st edition, McGraw Hill Professional.
3. Clifton B, (2012), *Advanced Web Metrics with Google Analytics*, 3rd edition, John Wiley & Sons.
4. Peterson E. T, (2004), *Web Analytics Demystified: AMarketer's Guide to Understanding How Your Web Site Affects Your Business*. Ingram.
5. Sostre P, LeClaire J, (2007), *Web Analytics for dummies*, John Wiley & Sons.
6. Burby J, Atchison S, (2007*), Actionable web analytics: using data to make smart business decisions*, John Wiley & Sons.
7. Dykes B, (2011), *Web analytics action hero: Using analysis to gain insight and optimize your business*, Adobe Press.

ST 920402 ECONOMETRICS

Objectives: To impart the learning of principles of econometric methods and tools. to improve student's ability to understand economics and finance. to apply econometric methods to the investigation of economic relationships and processes. To introduce the students to the traditional econometric methods developed mostly for the work with cross-sections data.

Outcomes: Demonstrate Simple and multiple Econometric models Interpret the models adequacy through various methods , Demonstrate simultaneous Linear Equations model.

Module 1

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

Module 2

Estimator, Heteroscedasticity, Auto-correlation, Multicollinearity, Auto-Correlation, Test of Auto-correlation, Multicollinearity, Tools for Handling Multicollinearity.

Module 3

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

Module 4

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

References

1. *Johnston, J. (1997). Econometric Methods, Fourth Edition, McGraw Hill*
2. *Gujarathi, D., and Porter, D. (2008). Basic Econometrics, Fifth Edition, McGraw-Hill*
3. *Intriligator, M. D. (1980). Econometric Models-Techniques and Applications, Prentice Hall.*
4. *Theil, H. (1971). Principles of Econometrics, John Wiley.*
5. *Walters, A. (1970). An Introduction to Econometrics, McMillan and Co.*

PRACTICALS

- ST 050106 - Practical 1 is Mathematical and Statistical computing using Python based on the courses ST 050101 , ST 050102 & ST 050103
- ST 050107 - Practical 2 is Computer Science practical based on the courses ST 050104 & ST 050105
- ST 050206 - Practical 3 is Mathematical and Statistical computing using R based on the courses ST 050201 , ST 050202 & ST 050203
- ST 050207 - Practical 4 is Computer Science practical based on the courses ST 050204 & ST 050205
- ST 050306 - Practical 5 is Mathematical and Statistical computing using R/Python based on the courses ST 050301 & ST 050302
- ST 050307 - Practical 6 is Computer Science practical based on the courses ST 050303 & ST 050304
- ST 900403 - Data Science Practical is Practical based on the courses ST 900401 & ST 900402
- ST 910403 Data Analytics Practical is Practical based on the courses ST 910401 & ST 910402
- ST 920403 Data Management Practical is Practical based on the courses ST 920401 & ST 920402

PROJECT

1. The Project work shall be carried out in an Industrial Organization where the student go for Industrial visit and shall be under the supervision of a teacher of the department concerned.
2. The Project work shall be evaluated based on the presentation of the project work done by the student, the dissertation submitted and the viva voce on the project.