I Semester

PH 901.1 : STATISTICS FOR DATA SCIENCE

Total No. of Lectures : 45	Total Marks : 100	[L – T – P – S]
No. of Lectures / Week : 4	Credits : 4	[3 – 1 - 0 - 2]

Learning Objectives: This course is about learning from data, in order to gain useful predictions and insights. Separating signal from noise presents many computational and inferential challenges, which we approach from a perspective at the interface of computer science and statistics

Learning Outcomes: On successful completion of the course students will be able:

CO1 : To demonstrate proficiency with statistical analysis of data.

CO2 : To develop the ability to build and assess data-based models.

CO3 : To execute statistical analyses with professional statistical software.

CO4 : To apply data science concepts and methods to solve problems in real-world contexts

CO5 : To create and communicate these solutions effectively

Unit - I

Data science in a big data world - Benefits and uses of data science and big data, Facets of data, the data science process, the big data ecosystem and data science. The data science process - Overview of the data science process,

Handling large data on a single computer - The problems you face when handling large data, General techniques for handling large volumes of data, General programming tips for dealing with large data sets

Text mining and text analytics-Text mining in the real world,Text mining techniques Data visualization to the end userData visualization options Crossfilter, the JavaScript MapReduce library,Creating an interactive dashboard with dc.js, Dashboard development tools

(9 Hrs)

Unit II

Introduction to Statistics – Definition, functions of statistics: condensation, comparison, forecasting & estimation, Importance and scope, Limitations of statistics, Types of statistics: Descriptive and inferential, variables and organization of data: quantitative (numerical): discrete, continuous, and qualitative (categorical),

Methods of collecting data: primary and secondary, Advantages and disadvantages. Measurement of

data: Nominal Data, Ordinal, Interval and ratio.

Classification of data: definition, characteristics of classification, objectives, guiding principles. Types of classifications: Geographical, Chronological, Qualitative and Quantitative. Preparation of tables: frequency distribution, ungrouped and grouped, class intervals, types of class intervals: exclusive and inclusive, Relative frequency, Cumulative frequency.

Presentation of data: pie chart, bar chart, stem and leaf display, histograms, scatterplots, frequency polygon , Lorenz curve.

(9 Hrs)

Unit III

Descriptive Statistics- Measures of Central Tendency: Requisites of a good measure, Mean: Arithmetic Mean, Geometric mean, Harmonic Mean, Weighted mean and Combined mean, Median, Mode, properties, merits and demerits, for grouped and ungrouped data Percentiles, Quartiles, Box Plots, Histogram, Bar chart, Relationship between arithmetic mean, geometric mean and harmonic mean, Need and importance.

Measure of Dispersion – Range, Interquartile Division, Quartile Division, Mean Division, Variance, Standard Division, Co-efficient of Variation, and combined variance, Need and Importance

(9 Hrs)

Unit IV

Sampling – Need and importance, Sampling Methods, Probability and non-probability sampling methods, sampling error and Non-sampling errors

Statistical Inference for single population: Estimating population mean using z statistic with population standard deviation known, Estimating population mean using t statistic with population standard deviation unknown, Estimating the population proportion, Estimating the population Variance, Estimating Sample size

Statistical tests-Hypotheses, types of hypothesis, types of errors, test statistics, significance level, power of the test, one sided and two sided testing using Z & t statistic, Constructing a test statistic

(9 Hrs)

Unit V

Statistical Inference: Hypothesis Testing for single Populations: Introduction to Hypothesis Testing, Testing Hypotheses about a population mean using the z statistic known, Testing Hypotheses about a population mean using the t statistic unknown, Testing Hypotheses about a proportion, Testing Hypotheses about a variance.

Statistical Inference about Two Population : Hypotheses Testing and confidence intervals about the

difference in two means using the z statistic population variance known, Hypotheses Testing and confidence intervals about the difference in two means: independent samples and population variance unknown, Statistical inference for two related populations, Statistical inference about two population proportions, Testing Hypotheses about two population variance, ANOVA.

Text Book :

[1] Michael Minelli, Michael Chambers, Ambiga Dhiraj, "Big Data : Big Data Analytics", 1st Edition, 2012, Wiley International

[2] S. Mohanthy, Madhu Jagadish, Harsh Srivatsa, "Big Data Imperatives: Enterprise Big Data Warehouse, BI Implementations and Analytics", 1st Edition, 2015, Wiley Apress

[3]. Thomas , Seemon, "Basic Statistics", 2014, Narosa publishing house

[4]. Rice , John A, "Mathematical Statistics and data Analysis", 2014, Cengage learning India

Reference Books:

[1]. Jey Liebowitz, "Big Data and Business Analytics", 1st Edition, 2014, CRC Press

[2]. Halper, Hurwitz, Nugent, Kaufman, "Big Data", 1st Edition, 2013, Wiley Dreamtech

[3]. Rajendra Akerkar, "Big data Computing", 1st Edition, 2013, CRC Press

[4]. Sheldon M Ross, "Probability and Statistics for engineers and scientists", 4th edition, 2009, Elsevier Publications

[5]. Douglas C. Montgomery & George C. Runger, "Applied Statistics and Probability for Engineers", 3rd edition, 2003, John Wiley & Sons

[6]. Albright, Zappe, Winston, "Data analysis, Optimization and Simulation Modeling", 4th edition, 2011, Cengage learning

[7]. Ken Black, "Applied Business Statistics making better business decision", 7th edition, 2013, Wiley Publishers

[8]. Sarab Boslaugh, Paul Andrew Watters, "Statistics in a nutshell", 2008, Shroff publishers and distributors India

[9]. Gupta , S; Kapoor V.K, "Fundamentals of mathematical Statistics", 2012, Sulthan Chand and Sons. New Delhi, India

[10]. B L Agarwal, "Programmed Statistics", 2nd edition, 2003, New Age International

[11]. David Freedman, Robert Pisani, Roger Purves, "Statistics", 4th edition, 2009,Viva books

II Semester

PH 901.2 : MULTIVARIATE TECHNIQUES FOR DATA SCIENCE

Total No. of Lectures : 45 Total Marks : 100 [L – T – P – S]

No. of Lectures / Week : 4 Credits : 4 [3 – 1 - 0 - 2]

Learning Objectives:

- Demonstrate knowledge and understanding of the basic ideas behind several common statistical techniques for analysing multivariate data (linear regression analysis, logistic regression analysis and principal component analysis)
- Analyse real data by applying these techniques using SPSS and interpret the resulting output
- Write short statistical reports based on these analyses

Learning Outcome: On successful completion of the course the student

CO1 : Will appreciate the range of multivariate techniques available,

CO2 : Will be able to summarize and interpret multivariate data,

CO3 : Will have an understanding of the link between multivariate techniques and corresponding univariate techniques,

CO4 : Will be able to use multivariate techniques appropriately at application level

CO5 : Undertake multivariate hypothesis tests, and draw appropriate conclusion

Unit - I

Preparation for Analysis: Multivariate Analysis; Examples; Characterizing Data for Analysis; Variables: their definition, classification, and use; Defining statistical variables; Stevens's classification of variables; variables are used in data analysis; Examples of classifying variables; other characteristics of data

Preparing for Data Analysis: Processing data so they can be analyzed; Choice of a statistical package Techniques for data entry; organizing the data

Data Screening and Transformations: Transformations, assessing normality and independence; Common transformations; Selecting appropriate transformations; Assessing independence (9 Hrs)

Unit - II

Selecting Appropriate Analyses: analysis to perform; Difficulty in selection; appropriate statistical measures; Selecting appropriate multivariate analyses

Applied Regression Analysis: Simple Regression and Correlation; use of regression and correlation; Data example; Regression methods: fixed-X case; Regression and correlation: variable-X case; Interpretation: fixed-X case; Interpretation: variable-X case; other available computer output;

Robustness and transformations for regression; Other types of regression; Special applications of regression.

Multiple Regression and Correlation Regression methods: fixed-X case; Regression and correlation: variable-X case; Interpretation: fixed-X case; Interpretation: variable-X case; Regression diagnostics and transformations; Other options in computer programs; Discussion of computer programs **(9 Hrs)**

Unit - III

Variable Selection in Regression: variable selection methods used; Data example; Criteria for variable selection; A general F test; Stepwise regression; Subset regression; Discussion of computer programs; Discussion of strategies

Special Regression Topics: Missing values in regression analysis; Dummy variables; Constraints on parameters; Regression analysis with multicollinearity; Ridge regression

Multivariate Analysis: Canonical Correlation Analysis; Use of canonical correlation analysis; Data example; Basic concepts of canonical correlation; Other topics in canonical correlation

Discriminant Analysis: Use of discriminant analysis; Data example; Basic concepts of classification; Theoretical background; Interpretation; Adjusting the dividing point; How good is the discrimination; Testing variable contributions; Variable selection;

(9 Hrs)

Unit - IV

Logistic Regression: Use of logistic regression; Data example; Basic concepts of logistic regression; Interpretation: Categorical variables; Interpretation: Continuous variables; Interpretation: Interactions; Refining and evaluating logistic regression; Nominal and ordinal logistic regression; Applications of logistic regression; Poisson regression; Discussion of computer programs.

Regression Analysis with Survival Data: Use of survival analysis used; Data examples; Survival functions; Common survival distributions; Comparing survival among groups; The log-linear regression model; The Cox regression model; Comparing regression models; Discussion of computer programs

Principal Components Analysis: Use of principal components analysis; Data example; Basic concepts; Interpretation; Other uses; Discussion of computer programs **(9 Hrs)**

Unit - V

Factor Analysis: Use of factor analysis; Data example; Basic concepts; Initial extraction: principal components; Initial extraction: iterated components; Factor rotations; Assigning factor scores; Application of factor analysis; Discussion of computer programs

Cluster Analysis: Use of cluster analysis; Data example; Basic concepts: initial analysis; Analytical clustering techniques; Cluster analysis for financial data set; Discussion of computer programs

Log-Linear Analysis: Use of log-linear analysis; Data example; Notation and sample considerations; Tests and models for two-way tables; Example of a two-way table; Models for multiway tables; Exploratory model building; Assessing specific models; Sample size issues; The logit model;

Correlated Outcomes Regression: Use of correlated outcomes regression; Data example; Basic concepts; Regression of clustered data; Regression of longitudinal data; Other analyses of correlated outcomes. **(9 Hrs)**

Text Books:

[1] Afifi A., May S. and Clark V.A., "Practical Multivariate Analysis", 2nd Edition, 2012, CRC Press, Taylor &Francis,

[2] Joseph F hair, William C Black, Barry J Babin, Ralph E Anderson, "Multivariate Data Analysis A Global Perspective", 6th Edition, 2017, Pearson

[3] Alvin C Renchar, "Methods for Multivariate Analysis", 2nd Edition, 2013, Wiley International

Reference Books:

[1] Sadabnori Konoshi, "Introduction to Multivariate Analysis – Linear and Non Linear Modeling", 2nd Edition, 2016, CRC Press

[2] Richard J Johnson, Dean W Wichern, "Applied Multivariate Statistical Analysis", 3rd Edition, 2016, Pearson Education Asia

[3] Boca Raton. Johnson R.A. and Wichern D.W., "Applied Multivariate Statistical Analysis", 4th Edition, 2014, Prentice Hall of India Pvt Ltd., New Delhi

[4] Anderson, T.W, "An Introduction to Multivariate Statistical Analysis", 2nd Ed, 2003, Wiley Eastern Ltd.

[5] Johnson, R. A and. Wichern D.W, "Applied Multivariate Statistical Analysis", 6 /e, 2007, PHI

[6] Singh, B.M., "Multivariate Statistical Analysis", 2nd Edition, 2012, South Asian Publishers Pvt. Ltd.,

[7] Barbara C Tabachink, Linda S Fidel, "Using Multivariate Statistics", 6th Edition, 2016, Pearson

[8] W Hardle, L Suimar, "Applied Multivariate Statistical Analysis", 2nd Ed, 2012, Springer

[9] Randal E Schumaker, "Using R with Multivariate Statistics", 2013, Elsevier Publishers

III Semester

PH 901.3 : COMPUTATIONAL INTELLIGENCE AND DEEP LEARNING

Total No. of Lectures : 45 Total Marks : 100 [L – T – P – S]

No. of Lectures / Week : 4 Credits : 4 [3 – 1 - 0 - 2]

Learning Objective : The subject aims to introduce students to

- fundamentals of key intelligent systems technologies including knowledge-based systems, neural networks, fuzzy systems, and evolutionary computation, and
- practice in integration of intelligent systems technologies for engineering applications..

Learning Outcome: Upon completion of the subject, students shall be able to

CO1 : Gain a working knowledge of knowledge-based systems, neural networks, fuzzy systems, and evolutionary computation;

CO2 : Apply intelligent systems technologies in a variety of engineering applications;

CO3 : Implement typical computational intelligence algorithms in Python

CO4 : Present ideas and findings effectively; and e. Think critically and learn independently.

CO5 : Application of Fuzzy and Genetic Algorithms in the mainline areas

Unit – I

Computational Intelligence: Adaption – Adaption versus Learning, Types & spaces of adaption, Self-Organizing and Evolution, Self-Organization.

Introduction: Neural Networks overview, Neuro Computing, Artificial Neural Networks; Basic Building Blocks of Artificial Neural Networks; Structure and Function of a Neuron; Neural Net Architectures; Applications. (9 hrs)

Unit – II

Fundamental Models of Neural Networks : McCulloch-Pitts Neuron Model-Architecture; Learning Rules; HebbNet- Architecture & Algorithm;.

Perceptron Networks: Single Layer Perceptron – Architecture & Algorithm, Perceptron Algorithm for several output classes; Multilayer Perceptron Networks. Linear Seperability, Pocket Algorithm

Adaline & Madaline Networks : ADALINE – Architecture & Algorithm; MADALINE – Architecture & MRI, MRII Algorithms.

(9 hrs)

Unit – III

Associative Memory Networks – Algorithm for Pattern Association; Hetero Associative Memory Networks – Architecture & Algorithm; Auto Associatve Memory Network – Architecture & Training Algorithm; Bi-directional Associative Memory – Architecture, Types of BAM, Algorithm, Hamming distance.

Feed Forward Networks : Back Propagation Network (BPN)- Architecture & Algorithm, Applications; Radial Basis Function Network (RBFN)- Architecture & Algorithm..

Feedback Networks: Discrete Hopfield Network – Architecture & Algorithm; Continuous Hopfield Network **(9 hrs)**

Unit – IV

Self Organizing Maps – Methods used for determining the winner, Principal Component Analysis(PCA), Kohonen Self Organizing feature Maps (SOM) – Architecture & Algorithm; Learning Vector Quantization – Architecture & Algorithm; Max Net – Architecture & Application; Maxican Net – Architecture & Algorithm; Hamming Net – Architecture & Algorithm.

Counter Propagation Network : Full Counter Propagation Network (Full CPN)- Architecture & Algorithm; Forward Only Counter Propagation Network- architecture & Algorithm

Adaptive Resonance Theory – ART Fundamentals, Architecture, Operations, Learning in ART, Training Steps; ART 1 – Architecture, Algorithm; ART 2 – Architecture, Algorithm (**9 hrs)**

Unit – V

Fuzzy set theory – Fuzzy vs Crisp, Crisp sets, Fuzzy sets, Crisp Relations, Fuzzy relations, **Fuzzy Systems** – Crisp Logic, Predicate logic, Fuzzy Logic, Fuzzy rule based system, deffuzzification methods, Applications of Fuzzy Logic.

Evolutionary Computing – Concepts, Creation of offsprings, Working principle, Encoding, Fitness Function, Reproduction. **Genetic Modelling** – Inheritance Operators, Cross Over, Inversion and Deletion, Mutation Operator, Bit-wise operator, Bit-wise operators used in GA, Generational cycle, Convergence of GA, Applications, Multi-level optimization.

(9 hrs)

Text Books :

[1]. Russel Eberhart, Yuhui Shi, " Computational Intelligence Concepts to Implementation", 2nd Edition, 2010, Elsevier Publications

[2]. S.N Sivanandam, S.N Deepa, " Principles of Soft Computing", 3rd Edition, 2015, Wiley.

References:

[1]. Haykin S., "Neural Networks-A Comprehensive Foundations", 3rd Ed, 2012, Pearson

[2]. S. Rajasekaran, G.A Vijayalakshmi Pai, "Neral Networks, Fuzzy Logiz, Genetic Algorithms", 3rd Ed, 2011, Prentice Hall of India Ltd.

[3]. Anderson J.A., "An Introduction to Neural Networks", 3rd Ed, 2014, Pearson Asia

[4]. Martin T agan, Howard B Demuth, "Neural Network Design", 1st Ed, 2012, Cengage

[5]. Satish Kumar, "Neural Networks", 2nd Edition, 2011, Tata McGraw Hill Publisher

[6]. Robert J. Schalkoff, "Artificial Neural Networks", 3rd Ed, 2013, McGraw Hill.

[7]. FU, Li-min, “Neural Networks in Computer Intelligence”, 5th Ed, 2015, McGraw Hill.

[8]. Klir, Yuan, “Fuzzy sets and Fuzzy Logic : Theory and Application”, 3rd Ed, 2012, PHI

PH 902.1 : BIG DATA AND DATA MANAGEMNT

Total No. of Lectures : 45	Total Marks : 100	[L – T – P – S]
No. of Lectures / Week : 4	Credits : 4	[3 – 1 - 0 - 2]

Learning Objectives: This course aims to provide an introduction to big data technologies, starting with MapReduce, which is the first of these datacenter-scale computation abstractions and whose Hadoop implementation lies at the core of an application stack that has been gaining widespread adoption in both industry and academia. Because of the success of Hadoop, a large number of big data tools, with specialization ranging from cluster resource management to complex data analytics, were built on and around Hadoop, creating a complete big data application stac

Learning Outcomes: On successful completion of the course students will be able:

CO1: Understand MapReduce as a computation model and an execution framework

CO2: Work with following tools in the big data application stack: Hadoop, YARN, Hive, Pig, Spark, and perhaps others…

CO3: Realize how different tools in the Hadoop stack fit in the big picture of big data analytics

CO4: Design distributed machine learning algorithms

CO5: Use cloud computing services (Amazon Web Services) to build your clusters and run large-scale data processing applications

Unit I

Fundamentals of Big Data: Understanding Big Data; Concepts and Terminology, Datasets, Data Analysis, Data Analytics, Descriptive Analytics, Diagnostic Analytics, Predictive Analytics, Prescriptive Analytics, Business Intelligence (BI), Key Performance Indicators (KPI), Big Data Characteristics –Volume, Velocity, Variety, Veracity, Value; Different Types of Data - Structured Data, Unstructured Data, Semi-structured Data, Metadata, Technical Infrastructure and Automation Environment; Business Goals and Obstacles Case Study Example - Identifying Data Characteristics 26: Volume, Velocity, Variety, Veracity, Value, Identifying Types of Data.

(9 Hrs)

Unit II

Business Motivations and Drivers for Big Data Adoption: Marketplace Dynamics, Business Architecture, Business Process Management, Information and Communications Technology, Data Analytics and Data Science, Digitization, Affordable Technology and Commodity Hardware, Social Media, Hyper-Connected Communities and Devices, Cloud Computing, Internet of Everything (IoE)

Case Study Example

Big Data Adoption and Planning Considerations: Organization Prerequisites, Data Procurement, Privacy, Security, Provenance, Limited Realtime Support, Distinct Performance Challenges, Distinct Governance Requirements, Distinct Methodology, Clouds,

(9 Hrs)

Unit III

Big Data Analytics Lifecycle, Business Case Evaluation; Data Identification, Data Acquisition and Filtering, Data Extraction, Data Validation and Cleansing, Data Aggregation and Representation, Data Analysis; Data Visualization, Utilization of Analysis Results; Case Study Example.

Enterprise Technologies and Big Data Business Intelligence: Online Transaction Processing (OLTP); Online Analytical Processing (OLAP); Extract Transform Load (ETL); Data Warehouses; Data Marts, Traditional BI, Ad-hoc Reports, Dashboards, Big Data BI, Traditional Data Visualization, Data Visualization for Big Data

Case Study Example.- Enterprise Technology; Big Data Business Intelligence **(9 Hrs)**

Unit IV

Storing & Analyzing Big Data - Big Data Storage Concepts: File Systems and Distributed File Systems, NoSQL, Sharding, Replication, Master-Slave, Peer-to-Peer, Sharding and Replication, Combining Sharding and Master-Slave Replication, Combining Sharding and Peer-to-Peer Replication,

CAP Theorem, ACID, BASE, Case Study Example.

Big Data Processing Concepts : Parallel Data Processing, Distributed Data Processing, Hadoop, Processing Workloads, Batch, Transactional, Cluster, Processing in Batch Mode, Batch Processing with MapReduce, Map and Reduce Tasks, Map, Combine, Partition, Shuffle and Sort, Reduce

MapReduce Example - Understanding MapReduce Algorithms, Processing in Realtime Mode, Speed Consistency Volume (SCV), Event Stream Processing, Complex Event Processing, Realtime Big Data Processing and SCV, Realtime Big Data Processing and MapReduce, Case Study Example

(9 Hrs)

Unit V

Big Data Storage Technology : On-Disk Storage Devices, Distributed File Systems, RDBMS Databases, NoSQL Databases, Characteristics, Rationale, Types, Key-Value, Document, Column-Family, Graph,

NewSQL Databases, In-Memory Storage Devices, In-Memory Data Grids, Read-through, Write-through, Write-behind, Refresh-ahead, In-Memory Databases, Case Study Example

Big Data Analysis Techniques: Quantitative Analysis, Qualitative Analysis, Data Mining, Statistical Analysis, A/B Testing, Correlation, Regression, Machine Learning, Classification (Supervised Machine Learning), Clustering (Unsupervised Machine Learning), Filtering, Semantic Analysis, Natural Language Processing, Text Analytics, Sentiment Analysis, Visual Analysis, Heat Maps; Time Series Plots, Network Graphs, Spatial Data Mapping,

Case Study Example -Correlation, Regression, Time Series Plot, Clustering, Classification

(9 Hrs)

Text Book :

[1]. Thomas Erl, Wajid Khattak, Paul Buhler, "Big Data Fundamentals Concepts, Drivers & Techniques", 1st Edition, 2016, CRC Press

[2]. Bill Franks, "Taming the Big Data Tidal Wave: Finding Opportunities in Huge Data Streams with Advanced Analytics", 1st Edition, 2015, John Wiley& sons

Reference Books :

[1]. Chris Eaton, Dirk DeRoos, Tom Deutsch, George Lapis, Paul Zikopoulos, "Understanding Big Data: Analytics for Enterprise Class Hadoop and Streaming Data", 1st Edition, 2014, McGrawHill Publishing International

[2]. Giles , David Corrigan, "Harness the Power of Big Data The IBM Big Data Platform", 2nd Edition, 2015, Tata McGraw Hill Publications Intl

[3]. Bart Baesens "Analytics in a Big Data World: The Essential Guide to DataScience and its Applications (WILEY Big Data Series)", 1st Edition, 2014, John Wiley & Sons

[4]. Arshdeep Bahga, Vijay Madisetti, "Big Data Science & Analytics: A Hands-On Approach ",1st Edition, 2016, Narosa Publishers

[5]. Michael Berthold, David J. Hand, "Intelligent Data Analysis", 1st Edition, 2012, Springer

[6]. Da Ruan, Guoquing Chen, Etienne E.Kerre, Geert Wets, "Intelligent Data Mining", 1st Ed, 2010, Springer Publishers

[7]. Anand Rajaraman and Jeffrey David Ullman, "Mining of Massive Datasets", 2nd Edition, 2015, Meditech Press, New Delhi

[8]. Glenn J. Myatt, "Making Sense of Data", 1st Edition, 2012, John Wiley & Sons, India

PH 902.2 : MACHINE LEARNING ALGORITHMS

Total No. of Lectures : 45 | Total Marks : 100 | [L – T – P – S]

No. of Lectures / Week : 4 | Credits : 4 | [3 – 1 - 0 - 2]

Learning Objectives:

- To understand the concepts of machine learning.
- To appreciate supervised and unsupervised learning and their applications.
- To understand the theoretical and practical aspects of Probabilistic Graphical Models.
- To appreciate the concepts and algorithm of reinforcement learning.

Learning outcomes:

CO1 : To implement a neural network for an application of your choice using an available tool.

CO2 : To implement probabilistic discriminative and generative algorithms for an application of your choice and analyze the results.

CO3 : To use a tool to implement typical clustering algorithms for different types of applications.

CO4 : To design and implement an HMM for a sequence model type of application

CO5 : To identify applications suitable for different types of machine learning with suitable justification.

Unit - I

Introduction: Machine Learning Foundations, Overview, Applications, Types of Machine Learning, Basic Concepts in Machine Learning, Examples of Machine Learning, Applications.

Supervised Learning: Probably Approximately Correct (PAC) Learning, Learning Multiple Classes, Regression, Model Selection and Generalization, Dimensions of a Supervised Machine Learning Algorithm.

Bayesian Decision Theory: Introduction, Classification, Discriminant Functions, Bayesian Networks, Association Rules.

Dimensionality Reduction: Introduction, Subset Selection, Principal Components Analysis, Linear Discriminant Analysis. **(9 hrs)**

Unit - II

Clustering: Introduction, Mixture Densities, k-Means Clustering, Expectation-Maximization Algorithm, Hierarchical Clustering, Choosing the Number of Clusters.

Nonparametric Methods: Introduction, Nonparametric Density Estimation, Nonparametric Classification, Nonparametric Regression: Smoothing Models.

Decision Trees: Introduction, Univariate Trees, Pruning, Rule Extraction from Trees, Multivariate Trees.

Linear Discrimination: Introduction, Generalizing the Linear Model, Geometry of the Linear

Discriminant, Pairwise Separation, Gradient Descent, Logistic Discrimination, Support Vector Machines.

(9 hrs)

Unit - III

Multilayer Perceptrons: Introduction, Training a Perceptron, Backpropagation Algorithm, Training Procedures, Tuning the Network Size.

Hidden Markov Models: Introduction, Discrete Markov Processes, Three Basic Problems of HMMs, Evaluation Problem, Learning Model Parameters, Model Selection in HMM.

Accessing and Comparing Classification Algorithms: Introduction, Cross-Validation and Resampling Methods, Measuring Error, Assessing a Classification Algorithm's Performance, Comparing Two Classification Algorithms, Comparing Multiple Classification Algorithms: Analysis of Variance.

Reinforcement Learning: Introduction, Elements of Reinforcement Learning, Model-Based Learning, Temporal Difference Learning, Generalization. (9 hrs)

Unit - IV

Introduction: Historical trends in Deep learning, Deep learning: Overview of Methods.

Applied Math for Machine Learning: Linear Algebra, Probability and Information Theory, Numerical Computation.

Feedforward Networks – Feed Forward networks, Gradient Based learning, Backpropogation; Regularization- Overview, Parameter Penalties, Data Augmentation, Multi Task learning, Bagging, Dropout; Optimization for Training Deep Models – Optimization vs training, Basic Algorithms, Adaptive learning Rates; Convolution Networks – The Convolution operation and CNNs, Convolution Networks, Pooling; Sequence Modeling: Recurrent and recursive Nets – Sequence Modeling, Unfolding Graphs, Recurrent Neural networks, Bidirectional RNNs, Deep Recurrent Networks.

(9 hrs)

Unit - V

Machine Learning research: Linear Factor Models- PCA and factor Analysis, ICA; Autoencoders – Stochastic Encoders and Decoders, Denoising Autoencoders, Applications; Representation Learning- Greedy Layer wise Unsupervised Pretraining, Transfer Learning and Domain Adaptation, Semi-supervised Distangling of causal factors, Distributed Representation; Structured Probabilistic Models for Deep Learning – Using Graphs to describe model structure, Sampling from Graphical models; Monte Carlo Methods – Markov Chain Monte Carlo Methods, Gibbs Sampling, Deep generative Models – Boltzmann Machines, Deep Belief Networks, Directive Generative nets.

(9 hrs)

Text books:

[1] Ethem Alpaydin, “Introduction to Machine Learning”, 1st 2004, MIT press

[2] Ian Goodfellow, Yoshua Bengio, Aaron Courville, “Deep Learning”, 1st 2016 MIT press

Reference books:

[1] Christopher Bishop, “Pattern Recognition and Machine Learning”, 1st 2006, Springer

[2] Kevin P. Murphy, “ Machine Learning: A Probabilistic Perspective”, 1st 2012, MIT press

[3] Tom Mitchell, “Machine Learning”, 1997, McGraw-Hill

[4] Michael Nielson, “Neural Networks and Deep leraning”, 2nd 2015, Determination Press

PH 902.3 : DATA SCIENCE AND INTERNET OF THINGS

Total No. of Lectures : 45	Total Marks : 100	[L – T – P – S]
No. of Lectures / Week : 4	Credits : 4	[3 – 1 - 0 - 2]

Learning Objectives: As part of this course, students will study,

- Explain in a concise manner how the general Internet as well as Internet of Things work.
- Understand constraints and opportunities of wireless and mobile networks for Internet of Things.
- Use basic measurement tools to determine the real-time performance of packet based networks.
- Analyse trade-offs in interconnected wireless embedded sensor networks

Course Learning Outcomes – upon successful completion of this course, the participant will be able to:

CO1 : Demonstrate knowledge, understanding of the security and ethical issues of the Internet of Things

CO2 : Conceptually identify vulnerabilities, including recent attacks, involving the Internet of Things

CO3 : Conceptually describe countermeasures for Internet of Things devices

CO4 : Analyze the societal impact of IoT security events & Develop critical thinking skills

CO5 : Compare and contrast the threat environment based on industry and/or device type

Unit I

IoT Web Technology: The Internet of Things Today, Time for Convergence, Towards the IoT Universe, Internet of Things Vision, IoT Strategic Research and Innovation Directions, IoT Applications, Future Internet Technologies, Infrastructure, Networks and Communication, Processes, Data Management, Security, Privacy & Trust, Device Level Energy Issues, IoT Related Standardisation, Recommendations on Research Topics.

Internet of Things Privacy, Security and Governance: Introduction, Overview of Governance, Privacy and Security Issues, Contribution from FP7 Projects, Security, Privacy and Trust in IoT-Data-Platforms for Smart Cities, First Steps Towards a Secure Platform, Smartie Approach. Data Aggregation for the IoT in Smart Cities, Security. **(9 hrs)**

Unit II

Architectural Approach for IoT Empowerment : Introduction, Defining a Common Architectural Ground, IoT Standardisation, M2M Service Layer Standardisation, OGC Sensor Web for IoT, IEEE, IETF and ITU-T standardization activities, Interoperability Challenges, Physical vs Virtual, Solve the Basic First, Data Interoperability, Semantic Interoperability, Organizational Interoperability, Eternal Interoperability, Importance of Standardisation, Plan for validation and testing, Important Economic Dimension, Research Roadmap for IoT Testing Methodologies. Semantic as an Interoperability Enabler and related work.

Identity Management Models in IoT : Introduction, Vulnerabilities of IoT, Security requirements,

Challenges for a secure Internet of Things, identity management, Identity portrayal, Different identity management model: Local identity, Network identity, Federated identity, Global web identity, Identity management in Internet of Things, User-centric identity management, Device-centric identity management, Hybrid identity management.

Trust Management in IoT: Introduction, Trust management life cycle, Identity and trust, Third party approach, Public key infrastructure, Attribute certificates, Web of trust models, Web services security, SAML approach, Fuzzy approach for Trust, Access control in IoT, Different access control schemes, Authentication and Access control policies modelling. **(9 hrs)**

Unit - III

Defining IOT Analytics and challenges: The situation; Defining IoT analytics; IoT analytics challenges Business value concerns.

IoT Devices and Networking Protocols: IoT devices, Networking basics, IoT networking connectivity protocols, IoT networking data messaging protocols, Analyzing data to infer protocol and device characteristics.

IoT Analytics for Cloud: Building elastic analytics; Elastic analytics concepts; Designing for scale Cloud security and analytics; The AWS overview; Microsoft Azure overview; The Thing Worx overview;
The AWS Cloud Formation overview; The AWS Virtual Private Cloud (VPC) setup walk-through; terminate and clean up the environment **(9 hrs)**

Unit - IV

Collecting the Data, Strategies and Techniques: Designing data processing for analytics; Applying big data technology to storage; Apache Spark for data processing; To stream or not to stream; Handling change.

Exploring IoT Data: Exploring and visualizing data, attributes that might have predictive value R. Adding internal datasets; Adding external datasets

Visualizing and Dashboarding : Common mistakes when designing visuals; The Hierarchy of Questions method; Designing visual analysis for IoT data; Creating a dashboard with Tableau; Creating and visualizing alerts. **(9 hrs)**

Unit -V

Applying Geospatial Analytics to IoT Data: Geospatial analytics for IoT; Basics of geospatial analysis Vector-based methods; Raster-based methods; Storing geospatial data; Processing geospatial data; Solving the pollution reporting problem

Data Science for IoT Analytics: Machine learning (ML); Anomaly detection using R; Forecasting using ARIMA; Deep learning

Strategies to Organize Data for Analytics: Linked Analytical Datasets; Managing data lakes; The data retention strategy.

Economics of IoT Analytics: The economics of cloud computing and open source, Cost considerations

for IoT analytics, revenue opportunities, predictive maintenance example

(9 hrs)

Text Books :

[1]. Andrew Minteer, "Analytics for the Internet of Things (IoT)", 2nd Ed, 2017, PACKT

[2]. Arsheep Bhaga, Vijay Madasetti, "Internet of Things : A Hands on Approach", 1st Edition, 2016, University Press India Ltd

[3]. Donald Norris, "The Internet of Things: Do-It-Yourself at Home Projects for Arduino, Raspberry Pi and BeagleBone Black", 1st Edition, 2015, McGraw Hill Publishers

References:

[1]. Hwaiyu Geng, "Internet of Things and Data Analytics Handbook", 1st Edition, 2016, Wiley

[2]. Peter Waher, "Mastering Internet of Things", 1st Edition, 2018, PACKT Publishers

[3]. Perry Lea, "Internet of Things for Architects", 2nd Edition, 2017, PACKT Publishers

[4]. Qusay F. Hassan, Atta ur Rehman Khan, Sajjad A. Madani, "Internet of Things: Challenges, Advances, and Applications", 2nd Edition, 2014, CRC Press

[5]. Pethuru Raj, Anupama C. Raman, "The Internet of Things: Enabling Technologies, Platforms, and Use Cases", 2nd Edition, 2014, CRC Press.

[6]. Olivier Hersent, David Boswarthick, Omar Elloumi, "The Internet of Things: Key Applications and Protocols", 2nd Edition, 2015, Wiley International

[7]. Adrian McEwen, Hakim Cassimally, "Designing the Internet of Things", 1st Ed, 2013, Wiley

[8]. Mile Lukodes, John Bruner, "What is the Internat of Things", 1st Ed, 2015, O'Reilly Publishers

[9]. Hakima Chaouchi, "The Internet of Things: Connecting Objects", 1st Edition, 2014, Wiley

[10]. Fawzi Behmann, Kwok Wu, "Collaborative Internet of Things (C-IoT): for Future Smart Connected Life and Business", 1st Edition, 2015, Wiley International

[11]. Fei Hu, "Security and Privacy in Internet of Things (IoTs): Models, Algorithms,", 2016, CRC Press

[12]. Weber, Rolph H, Romana, "Internet of Things : Legal Perspectives", 1st Ed, 2015, Springer

PH 903.1 : MATHEMATICS FOR DATA SCIENCE

Total No. of Lectures : 45	Total Marks : 100	[L – T – P – S]
No. of Lectures / Week : 4	Credits : 3	[3 – 1 - 0 - 2]

Learning Objectives: The Mathematical Course focuses on algebraic and numerical skills in a context of applications and problem solving to prepare students for Statistics or Contemporary Mathematics. Topics will include quantitative relationships, patterning and algebraic reasoning, functional reasoning, probabilistic and statistical reasoning, incorporating quantitative communication skills and technology.

Learning Outcomes: Upon successful completion of this course, a student will be able to:

CO1 : Students will formulate complete, concise, and correct mathematical proofs.

CO2 : Students will frame problems using multiple mathematical and statistical representations of relevant structures and relationships and solve using standard techniques.

CO3 : Students will create quantitative models to solve real world problems in appropriate contexts.

CO4 : Students will effectively use professional level technology tools to support the study of mathematics and statistics.

CO5 : Students will clearly communicate quantitative ideas both orally and in writing to a range of audiences.

Unit - I

Mathematical Foundations: Introduction and Motivation, Finding Words for Intuitions.

Linear Algebra: Systems of Linear Equations, Matrices, Solving Systems of Linear Equations, Vector Spaces, Linear Independence, Basis and Rank, Linear Mappings, Affine Spaces

Analytic Geometry: Norms, Inner Products, Lengths and Distances, Angles and Orthogonality, Orthonormal Basis, Orthogonal Complement, Inner Product of Functions, Orthogonal Projections, Rotations

(9 hrs)

Unit - II

Matrix Decompositions: Determinant and Trace, Eigenvalues and Eigenvectors, Cholesky Decomposition, Eigen decomposition and Diagonalization, Singular Value Decomposition, Matrix Approximation, Matrix Phylogeny

Vector Calculus: Differentiation of Univariate Functions, Partial Differentiation and Gradients, Gradients of Vector-Valued Functions, Gradients of Matrices, Useful Identities for Computing Gradients, Backpropagation and Automatic Differentiation, Higher-Order Derivatives, Linearization and Multivariate Taylor Series

(9 hrs)

Unit - III

Probability and Distributions: Construction of a Probability Space, Discrete and Continuous Probabilities, Sum Rule, Product Rule, and Bayes' Theorem, Summary Statistics and Independence, Gaussian Distribution, Conjugacy and the Exponential Family, Change of Variables/Inverse Transform

Continuous Optimization: Optimization Using Gradient Descent, Constrained Optimization and Lagrange Multipliers, Convex Optimization

(9 hrs)

Unit - IV

Central Machine Learning Problems: When Models Meet Data, Data, Models, and Learning, Empirical Risk Minimization, Parameter Estimation, Probabilistic Modeling and Inference, Directed Graphical Models, Model Selection

Linear Regression: Problem Formulation, Parameter Estimation, Bayesian Linear Regression, Maximum Likelihood as Orthogonal Projection

Dimensionality Reduction with Principal Component Analysis: Problem Setting, Maximum Variance Perspective, Projection Perspective, Eigenvector Computation and Low-Rank Approximations, PCA in High Dimension, Key Steps of PCA in Practice, Latent Variable Perspective.

(9 hrs)

Unit - V

Density Estimation with Gaussian Mixture Models: Gaussian Mixture Model, Parameter Learning via Maximum Likelihood, EM Algorithm, Latent-Variable Perspective,

Classification with Support Vector Machines: Separating Hyperplanes, Primal Support Vector Machine, Dual Support Vector Machine, Kernels, Numerical Solution. **(9 hrs)**

Text Book :

[1]. Marc Peter, A. Aldo Faisal, Cheng Soon Ong, " Mathematics for Machine Learning", 2nd Edition, Cambridge University Press, 2020

Reference Books :

[1]. David C.Lay, Steven R.Lay and J.J.McDonald: Linear Algebra and its Applications, 5 th Edition, Pearson Education Ltd., 2015

[2]. E. Kreyszig, "Advanced Engineering Mathematics", 10 th edition, Wiley, 2015.

[3]. Scott L.Miller,Donald G.Childers: "Probability and Random Process with application to Signal Processing", Elsevier Academic Press, 2 nd Edition,2013.

[4]. Gilbert Strang: Introduction to Linear Algebra, 5th Edition, Wellesley-Cambridge Press., 2016

[5]. Richard Bronson: "Schaum's Outlines of Theory and P roblems of Matrix Operations", McGraw-Hill,

[6]. Elsgolts, L.:"Differential Equations and Calculus o f Variations", MIR Publications, 3 rd Edition, 2012

[7]. T.Veerarajan "Probability, Statistics and Random Pr ocess", 3 rd Edition, Tata Mc-Graw Hill Co.,2016.

PH 904.1 : ALGORITHMS FOR ADVANCED DATA ANALYTICS

Total No. of Lectures : 45	Total Marks : 100	[L – T – P – S]
No. of Lectures / Week : 4	Credits : 4	[3 – 1 - 0 - 2]

Learning Objectives: This course introduces students to a number of highly efficient algorithms and data structures for fundamental computational problems across a variety of areas. Students are also introduced to techniques such as amortized complexity analysis.

Learning Outcomes: On successful completion of the course students will:

CO1 : Analyze the asymptotic performance of algorithms.

CO2 : Write rigorous correctness proofs for algorithms.

CO3 : Demonstrate a familiarity with major algorithms and data structures.

CO4 : Apply important algorithmic design paradigms and methods of analysis.

CO5 : Synthesize efficient algorithms in common engineering design situations.

Unit – I

Data Structures and Algorithms: A Philosophy of Data Structures - The Need for Data Structures, Costs and Benefits; Abstract Data Types and Data Structures; Design Patterns - Flyweight, Visitor, Composite, Strategy, Problems, Algorithms, and Programs.

Mathematical Preliminaries: Sets and Relations, Miscellaneous Notation, Logarithms, Summations and Recurrences, Recursion, Mathematical Proof Techniques - Direct Proof, Proof by Contradiction, Proof by Mathematical Induction; Estimation.

Algorithm Analysis: Best, Worst, and Average Cases; Asymptotic Analysis - Upper Bounds, Lower Bounds, Θ Notation, Simplifying Rules, Classifying Functions; Calculating the Running Time for a Program; Analyzing Problems; Common Misunderstandings; Multiple Parameters; Space Bounds; Speeding Up Your Programs; Empirical Analysis.

(9 hrs)

Unit – II

Lists, Stacks, and Queues: Lists - Array-Based List Implementation, Linked Lists, Comparison of List Implementations, Element Implementations, Doubly Linked Lists, Stacks, Array-Based Stacks - Linked Stacks, Comparison of Array-Based and Linked Stacks, Implementing Recursion; Queues - Array-Based Queues, Linked Queues, Comparison of Array-Based and Linked Queues; Dictionaries

Binary Trees: Definitions and Properties; The Full Binary Tree Theorem; A Binary Tree Node ADT; Binary Tree Traversals; Binary Tree Node Implementations - Pointer-Based Node Implementations, Space Requirements, Array Implementation for Complete Binary Trees; Binary Search Trees; Heaps and

Priority Queues; Huffman Coding Trees - Building Huffman Coding Trees; Assigning and Using Huffman Codes; Search in Huffman Trees.

Non-Binary Trees: General Tree Definitions and Terminology - An ADT for General Tree Nodes, General Tree Traversals; The Parent Pointer Implementation; General Tree Implementations - List of Children, The Left-Child/Right-Sibling Implementation, Dynamic Node Implementations, Dynamic "Left-Child/Right-Sibling" Implementation; K-ary Trees; Sequential Tree Implementations.

(9 hrs)

Unit – III

Sorting and Searching: Internal Sorting - Sorting Terminology and Notation; Three Θ (n2) Sorting Algorithms - Insertion Sort, Bubble Sort, Selection Sort; The Cost of Exchange Sorting; Shellsort; Mergesort; Quicksort; Heapsort; Binsort and Radix Sort; An Empirical Comparison of Sorting Algorithms; Lower Bounds for Sorting.

File Processing and External Sorting: Primary versus Secondary Storage - Disk Drives, Disk Drive Architecture, Disk Access Costs; Buffers and Buffer Pools; The Programmer's View of Files; External Sorting - Simple Approaches to External Sorting, Replacement Selection, Multiway Merging.

Searching: Searching Unsorted and Sorted Arrays; Self-Organizing Lists; Bit Vectors for Representing Sets; Hashing - Hash Functions, Open Hashing, Closed Hashing, Analysis of Closed Hashing, Deletion.

Indexing: Linear Indexing; ISAM; Tree-based Indexing; 2-3 Trees; B-Trees - B+-Trees, B-Tree Analysis.

(9 hrs)

Unit – IV

Graphs: Terminology and Representations; Graph Implementations; Graph Traversals - Depth-First Search, Breadth-First Search, Topological Sort; Shortest-Paths Problems -Single-Source Shortest Paths; Minimum-Cost Spanning Trees; Prim's Algorithm; Kruskal's Algorithm

Lists and Arrays Revisited: Multilists; Matrix Representations; Memory Management; Dynamic Storage Allocation; Failure Policies and Garbage Collection.

Advanced Tree Structures: Tries; Balanced Trees; The AVL Tree; The Splay Tree; Spatial Data Structures - The K-D Tree, The PR quadtree, Other Point & Spatial Data Structures (9 hrs)

Unit – V

Theory of Algorithms: Analysis Techniques - Summation Techniques; Recurrence Relations - Estimating Upper and Lower Bounds, Expanding Recurrences, Divide and Conquer Recurrences, Average-Case Analysis of Quicksort; Amortized Analysis.

Lower Bounds: Introduction to Lower Bounds Proofs; Lower Bounds on Searching Lists - Searching in Unsorted Lists, Searching in Sorted Lists; Finding the Maximum Value; Adversarial Lower Bounds Proofs; State Space Lower Bounds Proofs; Finding the ith Best Element; Optimal Sorting.

Patterns of Algorithms: Dynamic Programming; The Knapsack Problem; All-Pairs Shortest Paths; Randomized Algorithms; Randomized algorithms for finding large values; Skip Lists; Numerical Algorithms; Exponentiation; Largest Common Factor; Matrix Multiplication; Random Numbers; The Fast Fourier Transform.

Limits to Computation: Reductions; Hard Problems; The Theory of N P-Completeness; N P-Completeness Proofs; Coping with N P-Complete Problems; Impossible Problems –Uncountability, The Halting Problem Is Unsolvable.

(9 hrs)

Text Books:

[1]. Michael Goodrich, Roberto Tamssia, "Data Structures and Algorithms", 6th Ed, 2013, Wiley International

[2]. Nell Dale, Daniel T Joyce, Chip Weems, "Object Oriented Data Structures using Java", 3rd Edition, 2012, Jones & Bartlett International

[3]. Jeoffrey McConnel, "Analysis of Algorithms", 2nd Edition, 2014, Jones & Bartlett Publishers.

Reference Books:

[1]. Elliot B Koffman, Paul A T Wolfgang, "Data Structures, Abstraction and Design using Java", 2nd Edition, 2010, Wiley India

[2]. Mark A Johnson, "A concise introduction to Data Structures in Java", 1st Ed, 2012, CRC Publishers.

[3]. D. S Malik, "Data Structures using C++", 2nd Edition, 2014, Cengage Learning India

[4]. Ananda Rao Akepogu, Radhika Raju Palagiri, "Data Structures and Algorithms using C++", 2nd Edition, 2012, Pearson Education Asia

[5]. Langsam Yedidyah, Augustine Moshe J, Tenenbaum Aaron M, "Data Structures Using C and C++", 2nd Edition, 2009, PHI Learning India

[6]. Pradeep Dey, Manas Ghosh, Reema Thareja, "Computer Programming and Data Structures", 2nd Edition, 2012, Oxford University Press.

[7]. Elliot Koffman, Paul A T, "Objects, Abstracts, Data Structures and Design using C++", 1st Edition, 2008, Wiley India Publishers

[8]. Mark A Weiss, "Data Structures and Problem Solving using Java", 4th Edition, 2009, Pearson Education

[9]. Robert Sedgewick, Philippe Flajolet, "Introduction to Analysis of Algorithms", 2nd Edition, 2010, Wiley

[10]. Parag H Dave, H.B Dave, "Design and Analysis of Algorithms", 1st Ed, 2009, Pearson.

PS 904.2 : PYTHON FOR DATA SCIENCE

Total No. of Lectures : 45 | Total Marks : 100 | [L – T – P – S]

No. of Lectures / Week : 4 | Credits : 4 | [3 – 1 - 0 - 2]

Learning Objectives: The course gives you a set of practical skills for handling data that comes in a variety of formats and sizes, such as texts, spatial and time series data. These skills cover the data analysis lifecycle from initial access and acquisition, modeling, transformation, integration, querying, application of statistical learning and data mining methods, and presentation of results. This includes data wrangling, the process of converting raw data into a more useful form that can be subsequently analysed. The course is hands-on, using python, in the iPython interactive computing framework.

Learning Outcomes: At the end of this course students will

CO1 : Be familiar with the entire procedure of conducting a data analysis project

CO2 : Perform exploratory data analysis using Python's Scipy/Numpy libraries , visualization tools

CO3 : Prepare data for complex statistical analysis by working with SQL databases, complex joins/merging using Python's Pandas library

CO4 : Build statistical inference models with regression, bayesian methods, clustering techniques

CO5 : Data visualization and presentations with Python's matplotlib, Bokeh, Plotly libraries

Unit - I

Python: Python program; Python's execution model; Everything is an object –Numbers, Immutable sequences, Mutable sequences, Set types, Mapping types – dictionaries, The collections module

Iterating & Making Decisions: Conditional programming, Looping, Putting this all together, A quick peek at the itertools module

Functions: Scopes and name resolution, Input parameters, Return values, Recursive functions, Anonymous functions, Function attributes, Built-in functions, Documenting code, importing objects.

(9 hrs)

Unit - II

Saving Time & memory: map, zip, and filter; Comprehensions; Generators; performance considerations; Name localization; Generation behavior in built-ins.; Object-oriented programming; custom iterator.

Testing, Profiling & Dealing with Exceptions: Testing your application, Test-driven development, Exceptions, Profiling Python

Edges, GUIs and Web Development: First approach – scripting; Second approach – a GUI application; The Django web framework, A regex website, future of web development. (9 hrs)

Unit – III

Python for Data Analysis: Building NumPy, SciPy, matplotlib, and IPython from source; NumPy arrays; Using IPython as a shell; Reading manual pages; IPython notebooks.

NumPy Arrays: NumPy array object, Creating a multidimensional array, Selecting NumPy array elements, NumPy numerical types, One-dimensional slicing and indexing, Manipulating array shapes, Creating array views and copies, Fancy indexing, Indexing with a list of locations, Indexing NumPy arrays with Booleans, Broadcasting NumPy arrays

Statistics and Linear Algebra: NumPy and SciPy modules; Basic descriptive statistics with NumPy; Linear algebra with NumPy; eigenvalues and eigenvectors with NumPy; NumPy random numbers; Creating a NumPy-masked array **(9 hrs)**

Unit – IV

Pandas Primer: pandas DataFrames; pandas Series; Querying data in pandas; Statistics with pandas DataFrames; Data aggregation with pandas DataFrames; Concatenating and appending DataFrames; Joining DataFrames; Handling missing values; Dealing with dates; Pivot tables;Remote data access

Retrieving, Processing and Storing Data: Writing CSV files with NumPy and pandas, Comparing the NumPy .npy binary format and pickling pandas DataFrames, Storing data with PyTables, Reading and writing pandas DataFrames to HDF5 stores, Reading and writing to Excel with pandas, Using REST web services and JSON, Reading and writing JSON with pandas, Parsing RSS and Atom feeds, Parsing HTML with Beautiful Soup **(9 hrs)**

Unit – V

Data Visualization: matplotlib subpackages; Basic matplotlib plots; Logarithmic plots; Scatter plots; Legends and annotations; Three-dimensional plots; Plotting in pandas; Lag plots; Autocorrelation plots; Plot.ly

Signal Processing and Time Series: statsmodels subpackages; Moving averages; Window functions; Defining cointegration; Autocorrelation; Autoregressive models; ARMA models; Generating periodic signals; Fourier analysis; Spectral analysis; Filtering.

Working with Databases: Lightweight access with sqlite3; Accessing databases from pandas; SQLAlchemy; Pony ORM; Dataset – databases for lazy people; PyMongo and MongoDB; Storing data in Redis; Apache Cassandra **(9 hrs)**

Text Books:

[1]. Febrizo Lomano, “Mastering Python”, 1st Edition, 2018, PACKT Publishers

[2]. Ivan Idrics “ Python For Data Analytics”, 1st Edition, 2017, PACKT Publishers

[3]. Magnus Vilhelm Persson, Luiz Felipe Martins, “Mastering Python Data Analysis”, 1st Edition, 2018, PACKT Publishers

Reference Books:

[1] Davy Cielen, “Introducing Data Science: Big Data, Machine Learning, and More, Using Python Tools”, 3rd Edition, 2016, Wiley International

[2] Nelli, Fabio, “Python Data Analytics : Data Analysis and Science using pandas, matplotlib and the Python Programming Language”, 2nd Edition, 2017, Apress International

[3] Henley, A.J., Wolf, Dave, “Learn Data Analysis with Python Lessons in Coding”, 2nd Edition, 2016, Apress International Edition.

[4] Daniel Y. Chen, “Pandas for Everyone: Python Data Analysis”, 2nd Edition, 2107, Pearson

[5] Jesus Rogel-Salazar, “Data Science and Analytics with Python”, 1st Edition, 2016, CRC Press

[6] David M. Beazley, “ Python Essential Reference”, Fourth Edition, 2014 , Addison Wesley

[7] Leonard Richardson, Dave Aitel, Eric Foster-johnson, Alex Samuel, Aleatha Parker, Michael Roberts, Peter C Norton, Jason Diamond, "Beginning Python", 2nd Ed, 2010, Wiley Wrox

[8] Mark Lutz , "Learning Python", 5th Edition; O'Reilly Publishers.

[9] Jeff Forcier, Paul Bissex, “Python Web Development with Django", , 2nd Edition, 2013, Pearson

[10] Wesley J. Chun, “ Core PYTHON Applications Programming”, Third Edition, 2012, Prentice Hall, 2012 Pearson Education, Inc

[11] Martin C Brown. "Python the Complete Reference", 1st Ed, 2011, Tata McGraw Hill

[12] Hetland, "Beginning Python", 1st Edition, 2008, Wiley-Apress

[13] James Payne, "Beginning Python using Python 2.6 and Python 3.1", 1st Ed, 2013, Wiley

[14] Brian K Jones, "Python Cookbook", 3rd Edition, 2013, Shroff/ O'reilly Publisher

[15] Mark Lutz, Programming Python", 1st Edition, 2011, O'reilly Publishers

PS 904.3 : BIG DATA ANALYTICS WITH SCALA AND SPARK

Total No. of Lectures : 45	Total Marks : 100	[L – T – P – S]
No. of Lectures / Week : 4	Credits : 3	[3 – 1 - 0 - 2]

Learning Objectives: This course introduces students to Big Data and the benefits it can provide to business. Students learn the main concepts in relation to Big Data storage and analytics, and security issues that arise in relation to Big Data. Students also learn how to ethically use Big Data to advance their business while taking care that no harm is done to their employees, customers and the community at large.

Learning Outcomes: Upon successful completion of this course, the student will be able to:

CO1 : Understand what Big Data is and why classical data analysis techniques are no longer adequate

CO2 : Understand the benefits that Big Data can offer to businesses and organisations

CO3 : Understand conceptually how Big Data is stored, retrieved and used

CO4 : Understand how Big Data can be analysed to extract knowledge

CO5 : Communicate with data scientists and apply the methods in real scenarios

Unit- I

Scala: Purposes of Scala, Platforms and editors, Installing and setting up Scala, Scala: the scalable language, Scala for Java programmers, Scala for the beginners

Object Oriented Scala: Variables in Scala, Methods, classes, and objects in Scala, Packages and package objects, Java interoperability, Pattern matching, Implicit in Scala, Generic in Scala, SBT and other build systems.

Functional Programming Concepts: Introduction to functional programming, Functional Scala for the data scientists, FP and Scala for learning Spark; Pure functions and higher-order functions; Using higher-order functions; Error handling in functional Scala; Functional programming and data mutability

Collection of APIs : Scala collection APIs, Types and hierarchies, Performance characteristics, Java interoperability, Using Scala implicits **(9 hrs)**

Unit- II

Spark : Introduction to data analytics, Introduction to big data, Distributed computing using Apache Hadoop, Apache Spark.

Spark – REPL & RDDS: Dig deeper into Apache Spark, Apache Spark installation, Introduction to RDDs, Using the Spark shell, Actions and Transformations, Caching, Loading and saving data.

Special RDD Operations: Types of RDDs, Aggregations, Partitioning and shuffling, Broadcast variables, Accumulators.

Spark SQL: Introducing Spark Session, Understanding Spark SQL concepts, Using Spark SQL in streaming applications; Spark SQL and DataFrames, DataFrame API and SQL API, Aggregations, Joins.

Spark SQL for Processing Structured and Unstructured Data: data sources in Spark applications, Spark

with relational databases, Spark with MongoDB (NoSQL database), Spark with JSON data, Spark with Avro files, Spark with Parquet files, Defining and using custom data sources in Spark **(9 hrs)**

Unit- III

Spark SQL for Data Exploration: Exploratory Data Analysis (EDA), Spark SQL for basic data analysis, Visualizing data with Apache Zeppelin, Sampling data with Spark SQL APIs, Spark SQL for creating pivot tables

Spark SQL for Data Munging: Data munging, Exploring data munging techniques, Munging textual data, Munging time series data, Dealing with variable length records, Preparing data for machine learning

Spark Streaming: Spark Streaming, Discretized streams, Stateful /stateless transformations Check pointing, Interoperability with streaming platforms (Apache Kafka), Structured streaming.

Graphx: A brief introduction to graph theory, GraphX, VertexRDD and EdgeRDD, Graph operators, Pregel API, PageRank; Exploring graphs using GraphFrames, Analyzing JSON input modeled as a graph Processing graphs containing multiple types of relationships, Understanding GraphFrame internals **(9 hrs)**

Unit- IV

Spark MLLIB & ML: Introduction to machine learning, Spark machine learning APIs, Feature extraction and transformation, Creating a simple pipeline, Unsupervised machine learning, Binary and multiclass classification

Bayes and Nave Bayes: Multinomial classification, Bayesian inference, Naive Bayes, The decision trees

Clustering Data with MLLIB: Unsupervised learning, Clustering techniques, Centroid-based clustering (CC), Hierarchical clustering (HC), Distribution-based clustering (DC), Determining number of clusters, A comparative analysis between clustering algorithms, Submitting Spark job for cluster analysis.

Text Analytics using Spark ML : Understanding text analytics, Transformers and Estimators, Tokenization, StopWordsRemover, NGrams, TF-IDF, Word2Vec, CountVectorizer, Topic modeling using LDA, Implementing text classification. **(9 hrs)**

Unit- V

Spark Tuning: Monitoring Spark jobs, Spark configuration, Common mistakes in Spark app development, Optimization techniques;

Deploying Spark on a Cluster – Spark architecture in a cluster, Deploying the Spark application on a cluster

Testing and Debugging in Spark: Testing in a distributed environment, Testing Spark applications, Debugging Spark applications.

PySparrk &Sparkr: Introduction to PySpark, Installation and configuration, Introducing SparkR, the SparkR architecture, SparkR DataFrames, SparkR for EDA and data munging tasks, SparkR for computing summary statistics, SparkR for data visualization, SparkR for machine learning

(9 hrs)

Text Book :

[1] Md. Rezaul Karim, Sridhar Alla, “Scala and Spark for Big Data Analytics - Harness the power of Scala to program Spark and analyze tonnes of data in the blink of an eye”, 1st Ed, 2017, PACKT

[2] Romeo Kienzler, “Mastering Apache Spark 2.x”, 2nd Edition, 2016, PACKT Publishers

Reference Books :

[1] Ilya Ganelin, Ema Orhian, Kai Sasaki, Brennon York, “Spark: Big Data Cluster Computing in Production”, 1st Edition, 2016,Wiley

[2] Rishi Yadav, “Apache Spark 2.x Cookbook”, 2nd Edition, 2016, PACKT Publishers

[3] Dean Wampler, Alex Payne, “Programming Scala, Scalability = Functional Programming + Objects”, 2nd Edition, 2016, O’Reilly Publishers

[4] Jason Swartz, “Learning Scala Practical Functional Programming for the JVM”, 2016, O’Reilly

[5] Guller, Mohammed, “Big Data Analytics with Spark A Practitioner's Guide to Using Spark for Large Scale Data Analysis”, 2nd Edition, 2016, Apress

[6] Luu, Hien, “Beginning Apache Spark 2 With Resilient Distributed Datasets, Spark SQL, Structured Streaming and Spark Machine Learning library”, 2nd Ed, 2016, Apress Edition

[7] Nabi, Zubair, “Pro Spark Streaming The Zen of Real-Time Analytics Using Apache Spark”, 2nd Edition, 2017, Apress Edition.

[8] Matei Zaharia, Holden Karau, Andy Konwinski, Patrick Wendell, “Learning Spark Lightning-Fast Big Data Analysis”, 2nd Edition, 2017, O’Reilly Publishers

[9] Matei Zaharia, Bill Chambers, “Spark: The Definitive Guide Big Data Processing Made Simple”, 3rd Edition, 2106, O’Reilly Publications

[10] Paul Chiusano, “Functional Programming in Scala”, 2nd Edition, 2017, Wiley

[11] Janek Bogucki, Alessandro Lacava, Aliaksandr Bedrytski, Matthew de Detrich, Benjamin Neil, “Professional Scala”, 2nd Edition, 2016, Wiley Wrox

PS 905.2: NoSQL, MapReduce with Hadoop

Total No. of Lectures : 45 | Total Marks : 100 | [L – T – P – S]

No. of Lectures / Week : 4 | Credits : 3 | [3 – 1 - 0 - 2]

Learning Objectives:

- Understand the Big Data Platform and its Use cases
- Provide an overview of Apache Hadoop
- Provide HDFS Concepts and Interfacing with HDFS
- Understand Map Reduce Jobs
- Provide hands on Hodoop Eco System
- Apply analytics on Structured, Unstructured Data.

Learning Outcomes: The students will be able to:

CO1 : Identify Big Data and its Business Implications.

CO2 : List the components of Hadoop and Hadoop Eco-System

CO3 : Access and Process Data on Distributed File System

CO4 : Manage Job Execution in Hadoop Environment

CO5 : Develop Big Data Solutions using Hadoop Eco System

Unit – I

Enterprise Data Architecture Principles: Data architecture principles; metadata; Data governance; Data security; Data as a Service; Evolution data architecture with Hadoop.

Hadoop Life cycle Management: Data wrangling; Data masking; Data security

Hadoop Design Consideration: Understanding data structure principles; Installing Hadoop cluster; Exploring HDFS architecture; Introducing YARN; Configuring HDFS high availability; Configuration of HA NameNodes with QJM; Hadoop cluster composition; Best practices Hadoop deployment; Hadoop file formats

(9 hrs)

Unit – II

Data Movement Techniques: Batch processing versus real-time processing; Apache Sqoop; Flume; Apache NiFi; Kafka Connect.

Data Modeling in Hadoop: Apache Hive, Supported datatypes, How Hive works, Hive architecture, Hive data model management, JSON documents using Hive, Apache HBase

NoSQL: Need of NoSQ; List of NoSQL Databases. Application; RDBMS approach; Challenges; NoSQL approach; NoSQL Storage types – Storage types; Comparing the models.

Advantages and Drwabacks: Transactional application, Computational application, Web-scale

application (9 hrs)

Unit – III

MapReduce: The basic philosophy underlying MapReduce; MapReduce - Visualized And Explained; MapReduce - Digging a little deeper at every step; "Hello World" in MapReduce; The Mapper; The Reducer; The Job; Get comfortable with HDFS; Run your first MapReduce Job

MapReduce - Combiners, Shuffle, Sort & Streaming API: Parallelize the reduce phase - use the Combiner; mappers and reducers with MapReduce; Parallelizing reduce using Shuffle And Sort; MapReduce language independence; Introducing the Streaming API; Python for MapReduce (9 hrs)

Unit – IV

HDFS & YARN: HDFS - Protecting against data loss using replication; Name nodes; Checkpointing to backup name node information; Yarn - Basic components, Submitting a job to Yarn, Plug in scheduling policies, Configure the scheduler.

MapReduce Customization: Setting up your MapReduce to accept command line arguments; The Tool, ToolRunner and GenericOptionsParser; Configuring properties of the Job object; Customizing the Partitioner, Sort Comparator, and Group Comparator.

K-Means Clustering: MapReduce job for K-Means Clustering; Measuring the distance between points; Custom Writables for Input/Output; Configuring the Job; The Mapper and Reducer; The Iterative MapReduce Job (9 hrs)

Unit – V

Inverted Index, Custom Data Types: Search engines - The Inverted Index; Generating the inverted index using MapReduce; Custom data types for keys - The Writable Interface; Represent a Bigram using a Writable Comparable; MapReduce to count the Bigrams in input text; MapReduce job using MRUnit.

Input-Output Formats and Customized Partitioning: File Input Format; Text And Sequence File Formats; Data partitioning using a custom partitioner; custom partitioner real in code; Total Order Partitioning; Input Sampling, Distribution, Partitioning and configuring these Secondary Sort

Hadoop as Database : Structured data in Hadoop; Running an SQL Select with MapReduce; Running an SQL Group By with MapReduce; A MapReduce Join - The Map Side; The Reduce Side;Sorting and

Partitioning,Putting it all together. (9 hrs)

Text Book :

[1]. Naresh Kumar, Prashant Shindgikar, "Modern Big Data Processing with Hadoop", 1st Ed, 2018, PACKT Publishers International

[2]. Gaurav Vaish, "Getting Started with NoSQL", 1st Edition, 2017, PACKT Publishers

[3]. Loonycorn, "Learn By Example: Hadoop, MapReduce for Big Data problems", 2018, PACKT

Reference Books :

[1]. Garry Turkington, "Hadoop Beginner's Guide", 1st Edition, 2015, PACKT Publishers

[2]. Venner, Jason, Wadkar, Sameer, Siddalingaiah, Madhu, "Pro Apache Hadoop", 2015, Apress

[3]. Vohra, Deepak, "Practical Hadoop Ecosystem - A Definitive Guide to Hadoop-Related Frameworks and Tools", 2nd Edition, 2017, Apress Publishers

[4]. Koitzsch, Kerry, "Pro Hadoop Data Analytics, "Designing and Building Big Data Systems using the Hadoop Ecosystem", 2nd Edition, 2016, Apress Publishers

[5]. Srinath Perera, Thilana Gunarathne, "Hadoop MapReduce",1st Ed, 2013, O'Reilly

[6]. Tom White, "Hadoop – Definitive Guide - Storage and Analysis at Internet Scale", 1st Edition, 2015, O'Reilly Publishers

[7]. Eric Sammer, "Hadoop Operations", 1st Edition, 2014, O'Reilly Publishers

[8]. Membrey, Peter, Plugge, Eelco, Hawkins, DUPTim, "The Definitive Guide to MongoDB The NoSQL Database for Cloud and Desktop Computing", 2nd Edition, 2015, Wiley Apress

[9]. Shashank Tiwari, "Professional NoSQL", 1st Edition, 2016, Wiley Wrox Publishers

[10]. Srinath Perera, Thilina Gunarathne, "Hadoop MapReduce Cookbook", 2nd Edition, 2017, PACKT Publishers International

[11]. Donald Miner, Adam Shook, "MapReduce Design Patterns Building Effective Algorithms and Analytics for Hadoop and Other Systems", 1st Edition, 2017, O'Reilly Publishers

[12]. Frank Kane, "Taming Big Data with MapReduce and Hadoop - Hands On", 2nd Edition, 2016, PACKT Publishers

PS 905.3 : DATA VISUALIZATION USING TABLEAU

Total No. of Lectures : 45	Total Marks : 100	[L – T – P – S]
No. of Lectures / Week : 4	Credits : 3	[3 – 1 - 0 - 2]

Learning Objectives: Data visualization techniques allow people to use their perception to better understand this data. The goal of this course is to introduce students to data visualization including both the principles and techniques. Students will learn the value of visualization, specific techniques in information visualization and scientific visualization, and how understand how to best leverage visualization methods.

- Develop skills to both design and critique visualizations
- Understand why visualization is an important part of data analysis
- Understand the components involved in visualization design
- Understand the type of data impacts the type of visualization

Learning Outcomes: Upon successful completion of this course, the student will be able to:

CO1 : Key concepts in data science, including tools, approaches and application scenarios

CO2 : Topics in information design, interaction design and user engagement

CO3 : Understand and apply the fundamental concepts and techniques in data visualization

CO4 : Solve specific real-world problems related to the visualisation and interpretation of data analysis results

CO5 : Use of Data Visualization in Business, Retail, Health, Sports Analtics

Unit – I

Dashboard: Preparing the dashboard, Showing the power of data visualization, Connecting to data sources, Introducing the Tableau interface, Interacting with your first data visualization, Sharing visualization with the world.

Summarizing Data for Dashboards: Dashboards and dates, Grouping your data with calculations, Correlation with calculations, Using cross-tabs flexibly, Simplifying your business rules with customer calculations

Interfacing with Data for Dashboards: grouping data with clarity, Hierarchies for revealing the dashboard message, Classifying data for dashboards, Actions and interactions, Drilling into the details, Working with input controls

Using Dashboards to get Results: Enriching data with mashups, Page trails, Guided analytics with Tableau, Sharing results in a meeting, Notes and annotations, Using external data to enrich your dashboard. **(9 hrs)**

Unit – II

Putting the Dash into Dashboards: Choosing the visualization, Using parameters in dashboards, Using custom geocoding in Tableau, Profiting from Big Data to rev your visualization, Filtering data for focus Creating choices in dashboards using conditional logic

Making Dashboards Relevant: Adding an infographic to your Tableau dashboard, String manipulation in dashboards, Correcting data exports from Tableau to Excel, Blending data, Optimizing tips for efficient, fast visualization

Visual Best Practices: Coloring the numbers, Dueling with dual axes, three dimensional data, pie charts or not? Sizing to make a data story.

Connecting to Data Sources: Connecting to text files- to Excel files - to Access databases - to a SQL Server; Pasting from a clipboard; Connecting to other databases, Connecting to Windows Azure Marketplace, dimensions and measures, Changing data types, Applying filters, Merging multiple data sources. **(9 hrs)**

Unit – III

Creating Univariate Charts: Creating tables, bar graphs, pie charts, sorting the graphs, creating histograms, line charts, Using the Show Me toolbar, stacked bar graphs, box plots, Showing aggregate measures.

Creating Bivariate Charts: Creating tables, Creating scatter plots, Swapping rows and columns, Adding trend lines, Selecting color palettes, Using dates.

Creating Multivariate Charts: Creating facets, area charts, bullet graphs, dual axes charts, Gantt charts, heat maps

Creating Maps: Setting geographic roles, Placing marks on a map, Overlaying demographic data, Creating choropleth maps, Using polygon shapes, Customizing maps. **(9 hrs)**

Unit – IV

Calculating User Defined Fields : Using predefined functions, Calculating percentages, Applying the If-Then logic, Applying logical functions, Showing totals, percentage of totals, Discretizing data, Manipulating text, Aggregating data.

Advanced Features: Viewing data, Changing the mark size, Using the presentation mode, Adding annotations, Excluding data on the fly, Customizing mark shapes, Adding drop-down selectors, Adding search box selectors, Adding slider selectors, Creating dashboards, Creating animated visualizations,

Creating parameters

Tableau Public: Tableau Public overview, Telling story with Tableau Public, Installing Tableau Public Opening files and creating profile,. Discover, Explore; Tableau Public user interface, Using the Marks card, The Show Me tool; Connecting Data - Public data, Tables and databases, data sources that Tableau Public connects to, The databases, tables, dimensions, facts, field formats and conventions, Connecting to the data in Tableau Public. **(9 hrs)**

Unit – V

Calculations: Creating calculated fields, Types of calculations -number functions, date functions, Type conversions, string functions, aggregate functions, logic functions, Blending data sources; Creating quick table calculations, Changing over time, Compute using Manually editing table calculations, Ranking The level of detail calculations

Dashboard Design and Style: Dashboard design process, Best practices for dashboard design, Creating a dashboard, dashboard tab interface, Setting the size of dashboard elements, Adding and using Filters, Filtering across Data sources with parameters, Actions, URL actions. **(9 hrs)**

Text Book :

[1] David Baldwin, "Mastering Tableau - Master the intricacies of Tableau to create effective data visualizations", 1st Ed, 2017, PACKT

[2] Acharya, Seema, Chellappan, Subhashini, "Pro Tableau A Step-by-Step Guide", 2017, Apress

[3] Joshua N. Milligan, "Learning Tableau 10", 2nd Edition, 2016, PACKT Publishers

Reference Books :

[1] Jen Stirrup et al., "Tableau: Creating Interactive Data Visualizations", 2nd Ed, 2016, PACKT

[2] Joshua N. Milligan, Donabel Santos, "Tableau 10 Bootcamp",1st Edition, 2015, PACKT

[3] Ryan Sleeper, "Practical Tableau -100 Tips, Tutorials, and Strategies from a Tableau Zen Master", 2nd Edition, 2016, O'Reilly Publishers

[4] Ben Jones, "Communicating Data with Tableau - Designing, Developing, and Delivering Data Visualizations", 1st Edition, 2015, O'Reilly Publishers

[5] Jen Stirrup, Ruben Oliva Ramos, "Advanced Analytics with R and Tableau", 2nd Edition, 2017, PACKT Publishers

[6] Joshua N. Milligan, "Learning Tableau", 1st Edition, 2016, PACKT

[7] Ashutosh Nandeshwar, "Tableau Data Visualization Cookbook", 2nd Edition, 2016, PACKT

[8] Chandraish Sinha, "Tableau 10 for Beginners : Step by Step Guide to Developing Visualizations in Tableau 10", 1st Edition, 2015, Create Space Publishers

[9] Khan, Arshad, “Jumpstart Tableau A Step-By-Step Guide to Better Data Visualization”, 1st Edition, 2016, Apress International

PS 906.1 : DATA WAREHOUSING AND DATA MINING

Total No. of Lectures : 45 | Total Marks : 100 | [L – T – P – S]

No. of Lectures / Week : 4 | Credits : 3 | [3 – 1 - 0 - 2]

Learning Objectives:

- To introduce the basic concepts and techniques of data mining.
- To develop the skills using recent data mining software for solving practical problems.
- To assess the strengths and weaknesses of various methods and algorithms
- Identify the key processes of data mining, data warehousing and knowledge discovery process.
- Basic principles and algorithms used in practical data mining and understand their strengths, weaknesses
- Apply data mining techniques to solve problems in other disciplines in a mathematical way.

Learning Outcomes: By the end of the module, the student should

CO1 : Display a comprehensive understanding of different data mining tasks and the algorithms most appropriate for addressing them.

CO2 : Evaluate models/algorithms with respect to their accuracy.

CO3 : Perform a self directed piece of practical work that requires the application of data mining techniques.

CO4: Develop hypotheses based on the analysis of the results obtained and test them.

CO5: Conceptualize a data mining solution to a practical problem.

Unit – I

Defining Data Warehouse Concepts and Terminology : Common data warehouse definitions, Data warehouse properties and characteristics, Warehouse development approaches, Components of data warehouse design and implementation, Components of a data warehouse, Data warehouse compared with data mart, Dependent and independent data marts

Planning and Managing the Data Warehouse Project : Managing financial issues, Obtaining business commitment, Gathering business and user requirements, Evaluating the warehouse project, Implementation processes and requirements.

Modeling the Data Warehouse : Data warehouse database design phases, Defining the business model, Choosing the architecture, Creating the dimensional model, Using time in the data warehouse, Using summary data, Query rewrite, Creating the physical model

(9 hrs)

Unit – II

Building the Warehouse - Extracting Data **:** Extracting, transforming, and loading data, Examining data sources, Extracting data, Extraction techniques Transforming Data **:** Transformation, Transforming data: problems and solutions, Resolving quality data issues, Transformation techniques, Transformation tools; Loading Warehouse Data **:** Loading data into the warehouse, Building the loading process, Loading the data, Post-processing of loaded data, Verifying data integrity**;** Refreshing Warehouse Data : Capturing and applying changed data, Batch load requirements, Limitations of methods in applying change, Purging and archiving data ; Leaving a Metadata Trail : Defining warehouse metadata, Developing a metadata strategy, Examining types of metadata, Metadata management tools, Common warehouse metadata; Managing the Data Warehouse : Managing the transition to production, Managing growth Managing backup and recovery, Identifying data warehouse performance issues **(9 hrs)**

Unit – III

Data Mining: Data, Types of Data, Data Mining functionalities – Data Mining Task Primitives; Issues - Integration of Data Mining system with a Data Warehouse.

Data Preprocessing: Descriptive Data Summarization, Data Cleaning, Data Integration and Transformation, Data Reduction, Data Discretization and Concept Hierarchy Generation

Association Rule Mining: Scalable Frequent Itemset Mining Methods, Improving efficiency, Mining various kinds of Association Rules, Association Mining to Correlation Analysis, Constraint based Association Mining.

(9 hrs)

Unit – IV

Classification and Prediction: Issues; Classification by Decision Tree Introduction; Bayesian Classification; Rule Based Classification; Classification by Back propagation; Support Vector Machines; Associative Classification; Lazy Learners; Other Classification Methods; Prediction;

Accuracy and Error Measures; Evaluating the Accuracy of a Classifier or Predictor; Ensemble Methods; Model Section. (9 hrs)

Unit – V

Cluster and Outlier Analysis: Types of Data in Cluster Analysis; Categorization of Major Clustering Methods; Partitioning Methods; Hierarchical methods; Density; Based Methods; Grid-Based Methods; Model-Based Clustering Methods; Constraint-Based Cluster Analysis; Outlier Analysis.

Data Mining Applications: Data mining applications; Social impacts of data mining: Ubiquitous and invisible data mining; data mining privacy and data security; Case Studies: Mining the WWW -Text mining.

(9 hrs)

Text Books:

[1] Jain Pei, Jiawei Han, Micheline Kamber, "Data Mining : Concepts and Techniques", 3rd Ed, 2011, Elsevier

[2] Alex Berson, Stephen J. Smith "Data Warehousing, Data Mining & OLAP", 3rd Ed, 2011, McGraw Hill

[3] Witten, Frank, Hall, "Data Mining : Practical Machine Learning Tools & Techniques", 3rd 2010, Elsevier

Reference Books:

[1] Reema Theraja "Data Warehousing", 1st Edition, 2011, Oxford University Press.

[2] Paulraj Ponniah, "Data Warehousing: Fundamentals for IT Professionals", 2nd Ed, 2012, Wiley India

[3] Prabhu C.S.R., "Data Warehousing Concepts, Techniques, Products and Applications", 3rd Edition, 2011, PHI Learning Private Limited.

[4] K.P. Soman, ShyamDiwakar, V. Ajay "Insight into Data mining Theory and Practice", 2nd Ed, 2010, PHI

[5] Pang-Ning Tan, Michael Steinbach, Vipin Kumar "Introduction to Data Mining", 3rd Ed, 2011, Pearson

[6] Vikram Pudi, Radhakrishna, "Data Mining", 2nd Edition, 2011, Oxford University Press.

[7] Richard Roiger, Michael Getz, "Data Mining : A Practical Based Primer", 1st Ed, 2010, Pearson

[8] Margaret Dunham, "Data Mining : Introductory & Advanced Topics", 1st Edition, 2011, Pearson

[9] Arun K Pujari, "Data Mining Techniques", 2nd Ed, 2013, University Press

[10] Peter Adrians, Rolf Zantinge, "Data Mining", 1st Edition, 2010, Pearson Education