ANNEXURE - II

COCHIN UNIVERSITY OF SCIENCE AND TECHNOLOGY

REGULATIONS and SCHEME

M.Sc. (COMPUTER SCIENCE)

With specialization in

DATA SCIENCE

for

DEPARTMENT OF COMPUTER APPLICATIONS

1. **Aims & Objectives:**
   - To impart proficiency in theoretical as well as practical aspects of Computer Science studies at Post Graduate level
   - To deliver high caliber professionals with expertise in the specialized and relative areas of Soft Computing
   - To cater to the needs of industry, Academia and various Organizations involved in Data Science, Analytics etc.
   - To develop high quality Computer Professionals capable of teaching both graduate and post graduate level
   - Develop computer scientists and professionals having ethical values and social responsibilities
   - To develop Professionals with ability to do research in Computer Science focusing the area of data science and analytics
2. **Conditions for Admission**

   Qualifications

   i) A candidate seeking admission to the M.Sc. Computer Science with specialization in Data Science must have Pass with not less than 60% marks in recognized regular bachelor's degree in Mathematics/Statistics/Computer Science/Information Technology/ Computer Applications/ Computer Science Engineering of minimum 3 years duration.

   Applicable relaxation as per rules followed in the case of SC/ST/SEBC/OBC/OEC candidates.

3. All other regulation pertaining to regular M.Sc. courses of Cochin University of Science and Technology is applicable to this course also.

VISION AND MISSION OF THE DEPARTMENT

Vision

To impart innovation-oriented education, to build globally competent and socially committed professionals.

Mission

M1: To empower students to meet the elementary expectations of Data Science field.

M2: To enhance the technical knowledge with respect to the industries and research scenario.

M3: To mould ethical software professionals and teachers with intensive knowledge in computer science and applications.

M4: To provide globally recognized academic environment through industry – academia collaborations, digital learning and state of the art skill development

M5: To instil research interests among students to solve societal problems

PROGRAM EDUCATION OBJECTIVES

PEO 1: Design and develop innovative ideas to solve data science related problems.

PEO2: Identify lifelong learning for enhancing knowledge base and skills necessary to contribute to the improvement of their profession and community.

PEO3: Analyze the existing solutions to analytics related problems and develop efficient methods.

PEO4 : Communicate effectively for professional enhancement and demonstrate professional ethics with societal responsibilities.

PEO5 : Exhibit appropriate interpersonal skills for working on teams in workplace

Programme Articulation Matrix

	M1	M2	M3	M4
PEO1	3	3		
PEO2			3	3
PEO3			3	3
PEO4			3	
PEO5	3	3	3	

PROGRAMME OUTCOME

PO-1: Apply the knowledge of mathematics, statistics, and computer science to the solution of complex data science problems.

PO-2: Identify, formulate, review research literature, and analyse problems reaching validated conclusions.

PO-3: Design solutions for complex data science problems and design processes that meet the specified needs with appropriate consideration for the society.

PO-4: Use research-based knowledge and research methods to conduct investigations on complex problems and provide valid conclusions

PO-5: Create and apply appropriate techniques for training, prediction, testing and modelling using available tools.

PO-6: Apply reasoning based on contextual knowledge to assess societal, health, safety, legal and cultural issues.

PO-7: Communicate effectively by writing effective reports and design documentation, make effective presentations, and give and receive clear instructions.

Mapping of POs with PEOs

Programme Outcome	**PEO1**	**PEO2**	**PEO3**	**PEO4**	**PEO5**
Engineering Knowledge	**X**				
Problem Analysis	**X**	**X**	**X**		
Design / Development of Solution	**X**		**X**		
Conduct Investigation			**X**		
Modern Tool Usage	**X**			**X**	
Engineer and Society		**X**			
Individual and Team work and Communication				**X**	**X**

M.Sc. COMPUTER SCIENCE WITH SPECIALIZATION IN DATA SCIENCE COURSE STRUCTURE

(2020 Admission onwards)

Semester I

Course Code	Paper				Marks		Credit
		L	T	P	Sessional	Final	
	Statistical Foundations for data Science	4	1	3	50	50	4
	Operating System Concepts*	4	1	3	50	50	4
	Data Structures and Algorithms*	3	1	2	50	50	3
	Python for Data Analytics	3	1	2	50	50	3
	Mathematics for Machine Learning	3	1	2	50	50	3
	Python Programming LAB			2	50	50	1
	Mini Project			2	50		1
TOTAL							19

Semester II

Course Code	Paper				Marks		Credit
		L	T	P	Sessional	Final	
	Networks and Data Communications*	4	1	3	50	50	4
	Database Management Systems*	4	1	3	50	50	4
	R for Data Analytics	3	1	2	50	50	3
	Data Mining & Machine Learning**	3	1	2	50	50	3
	Elective I	3	1	2	50	50	3
	R Programming LAB			2	50	50	1
	Mini Project			2	50		1
		TOTAL					19

Semester III

Course Code	Paper				Marks		Credit
		L	T	P	Sessional	Final	
	Soft Computing Techniques	4	1	3	50	50	4
	Elective II	3	1	2	50	50	3
	Elective III	3	1	2	50	50	3
	Elective IV	3	1	2	50	50	3
	Elective V	3	1	2	50	50	3
	Seminar			2	50		1
	Mini Project			2	50		1
			TOTAL				17

Semester IV

Course Code	Paper	Marks		Credit
		Sessional	Final	
	Internship/Project Work	200	200	18
TOTAL				18

Second Semester Electives

1. Predictive Analytics

2. Text Analytics.

3. Social Network Analysis

Third Semester Electives

1.No SQL Databases

2. Image and Video Analytics

3. Healthcare Data Analytics

4.Fraud Analytics

5. Block Chain Technologies (Industry Oriented course/ MCA)

6.Big Data Analytics **

7. Natural Language Processing **

8. Information Retrieval *

9. Deep Learning **

10.Business Analytics **

11.Data Visualization #

12.Data warehousing #

13.Computational modelling #

14.Time Series Analysis and SEM Modeling #

15.AI & Knowledge representation #

* Subjects from M.Sc. Computer Science with specialization in Soft Computing

** Subjects from MCA

Syllabus to be approved

CAM 2101 STATISTICAL FOUNDATIONS FOR DATA SCIENCE
(December 2019)

Course Outcomes
After completion of this course, students will be able to

CO1	Explain the basics of Probability, Regression and Correlation.	(Cognitive level : Understand)
CO2	Solve the problems using Bayes Theorem, Poisson Distribution and Normal Distribution	(Cognitive level : Apply)
CO3	Apply testing methods relating to Chi square test and t-test.	(Cognitive level : Apply)
CO4	Interpret and models binary data	(Cognitive level : Analyze)
CO5	Explain the concept of Linear Regression and Parameter estimation	(Cognitive level : Understand)
CO6	Interpret parameter in Logistic Regression	(Cognitive level : Analyze)
CO7	Interpret and Analyze data displayed in tables	(Cognitive level : Analyze)

Mapping of course outcomes with programme outcomes - **Low=1, medium=2, High=3**

	PO1	PO2	PO3	PO4	PO5	PO6	PO7
CO1	1						
CO2	2						
CO3	3						
CO4		2					
CO5	1						
CO6		3					
CO7		3					

MODULE I

Probability, basics, Conditional Probability, Bayes Theorem, Distributions - Binomial, Poisson, Normal distributions and related problems. Summary Statistics for Pre-processing and Visualization- Regression and correlation. Hypothesis – generation and testing – Chi square test, t-test, Analysis of variance, correlation

MODULE II

Introduction to Categorical variables, Independent and dependent variables. Distributions for Categorical Data, Statistical Inference for Categorical Data, Statistical Inference for Binomial Parameters, Statistical Inference for Multinomial Parameters, Bayesian Inference for Binomial and Multinomial Parameters.

MODULE III

Models for binary data, The Generalized Linear Model, Generalized Linear Models for Binary Data, Generalized Linear Models for Counts and Rates, Moments and Likelihood for

Generalized Linear Models, Inference and Model Checking for Generalized Linear Models, Fitting generalized Linear Models.

MODULE IV

Linear Regression – Problem formulation, Parameter Estimation, Bayesian Linear Regression, Maximum likelihood as orthogonal projection.

MODULE V

Tabular data- Power and the computation of sample size- Advanced data handling, Logistic Regression - Interpreting Parameters in Logistic Regression, Inference for Logistic Regression, Logistic Models with Categorical Predictors, Multiple Logistic Regression, Fitting Logistic Regression Models, Non Linear models, Multilevel models.

References:

1. Ronald E Walpole , "Probability and Statistics for Engineers and Scientists", 9 ed, Pearson, 2013.
2. Peter Bruce, "Practical Statistics for Data Scientists, 2015
3. Marc Peter Deisenroth, A. Aldo Faisal, Cheng Soon Ong, "Mathematics for Machine Learning**", 2020**

PYTHON FOR DATA ANALYTICS

Course Outcomes
After completion of this course, students will be able to

CO1	Describe the basics of machine learning	(Cognitive level: Understand)
CO2	Apply sci-kit learn for machine learning	(Cognitive level : Apply)
CO3	Explain python data science libraries for data science	(Cognitive level : Understand)
CO4	Create visualizations using python matplotlib	(Cognitive level : Create)
CO5	Analysis of data with Jupyter notebooks, numpy, pandas etc	(Cognitive level : Analyze)
CO6	Apply data analytics in standard datasets and evaluate the results	(Cognitive level : Apply)

Mapping of course outcomes with programme outcomes

	PO1	PO2	PO3	PO4	PO5	PO6	PO7
CO1	1						
CO2			2				
CO3	1						
CO4			3				
CO5			3				
CO6			3				

MODULE I

An Introduction to Data Analysis : Knowledge Domains of the Data Analyst, Understanding the Nature of the Data, The Data Analysis Process, Quantitative and Qualitative Data Analysis, introduction to data analytics using Python- The Interpreter,Python Distributions, IPython, PyPI—The Python Package Index, The IDEs for Python, SciPy

MODULE II

The NumPy Library, Ndarray,Basic Operations ,Indexing, Slicing, and Iterating, Conditions and Boolean Arrays, Shape Manipulation, Array Manipulation,Structured Arrays, Reading and Writing Array Data on Files The pandas Library—An Introduction, Introduction to pandas Data Structures, Other Functionalities on Indexes, Operations between Data Structures, Function Application and Mapping, Sorting and Ranking, Correlation and Covariance, "Not a Number" Data, Hierarchical Indexing and Leveling

MODULE III

pandas: Reading and Writing Data, Reading and Writing HTML Files, Reading Data from XML, Reading and Writing Data on Microsoft Excel Files, JSON Data, The Format HDF5,Pickle—Python Object Serialization, Interacting with Databases, Reading and Writing Data with a NoSQL Database: MongoDB pandas in Depth: Data Manipulation- Data Preparation, Concatenating, Data Transformation, Discretization and Binning, Permutation, String Manipulation, Data Aggregation, Group Iteration, Advanced Data Aggregation

MODULE IV

Data Visualization with matplotlib- The matplotlib Library, matplotlib Architecture, pyplot, Using the kwargs, Adding Further Elements to the Chart and Saving Charts, Handling Date Values, Line Chart, Histogram, Bar Chart, Pie Charts, Advanced Charts, mplot3d, Multi-PanelPlots Machine Learning with scikit-learn: The scikit-learn Library, Machine Learning :Supervised and Unsupervised Learning , Training Set and Testing Set

MODULE V

Supervised Learning with scikit-learn, The Iris Flower Dataset- The PCA Decomposition, K-Nearest Neighbors Classifier, Diabetes Dataset, Linear Regression: The Least Square Regression Support Vector Machines (SVMs), Support Vector Classification (SVC), Nonlinear SVC, Plotting Different SVM Classifiers Using the Iris Dataset, Support Vector Regression (SVR) An Example- Recognizing Handwritten Digits- Handwriting Recognition, Recognizing Handwritten Digits with scikit-learn, The Digits Dataset, Learning and Predicting

TEXT BOOK

Fabio Nelli , "Python Data Analytics Data Analysis and Science Using Pandas, matplotlib, and the Python Programming Language ", Apress, 2015

REFERENCES

1. William McKinney , "Python for Data Analysis: Data Wrangling with Pandas, NumPy, and Ipython",2 nd edition,O'Reilly,2018

2. David Taieb , "Data Analysis with Python: A Modern Approach", 1st EditionPAckt publishers,2018.

Mathematics for Machine Learning

Course Outcomes

After completion of this course, students will be able to

CO1	Solve system of Linear equations using Gauss Elimination method.	(Cognitive level : Apply)
CO2	Explain the basics of Vectors, Spaces and Affine Spaces	(Cognitive level : Understand)
CO3	Evaluate similarity between two vectors.	(Cognitive level : Evaluate)
CO4	Apply different methods to find the Inverse and Rank of a Matrix	(Cognitive level : Apply)
CO5	Calculate Eigen values and Eigen vectors using Linear transformation and power methods.	(Cognitive level : Analyze)
CO6	Solve Derivatives and Partial Derivatives using rules of differentiation	(Cognitive level : Apply)
CO7	Apply optimization using gradient function	(Cognitive level : Apply)

Mapping of course outcomes with programme outcomes

	PO1	PO2	PO3	PO4	PO5	PO6	PO7
CO1	2						
CO2	1						
CO3	2						
CO4	2						
CO5	3						
CO6	2						
CO7				3			

MODULE I

Linear Algebra – System of Linear equations, Solving System of Linear equations, Linear Independence, Vectors , Scalars, Addition, Scalar multiplication, dot product, vector projection, cosine similarity.

MODULE II

Orthogonal vectors, normal and orthonormal vectors , vector norm, vector space, linear combination, basis of vectors, Affine spaces

MODULE III

Matrices – Determinant, Identity matrix, Inverse of a matrix, Rank of a matrix, Nullity , trace of a matrix, Eigen values, Eigen vectors, Matrix decompositions.

MODULE IV

Differentiation, rules of differentiation, Derivatives, Scalar derivatives, Partial derivatives, Principle Component analysis – Concepts and properties. Dimensionality reduction with PCA

MODULE V

Differentiation of univariate functions, Partial differentiation and gradients, Gradient of vector valued function. Gradient of matrices. Optimization using gradient functions, Constrained optimization and Lagrange multipliers. Convex optimization.

Text Books:

1. Marc Peter Deisenroth, A. Aldo Faisal, Cheng Soon Ong, "Mathematics for Machine Learning**", Cambridge University Press, 2020.**

References:

1. Erwin Kreyszig, Advanced Engineering Mathematics, 10thEdition., John Wiley & Sons, (2014).
2. B. S.Grewal, Higher Engineering Mathematics, 38th Edition. Khanna Publications,(2005).

R FOR DATA ANALYTICS

Course Outcomes

After completion of this course, students will be able to

CO1	Explain the basics of R language	(Cognitive level : Understand)
CO2	Solve problems related to classification and clustering using R tool.	(Cognitive level : Apply)
CO3	Apply association rule mining methods using R language.	(Cognitive level : Apply)
CO4	Interpret data using visualization techniques in R	(Cognitive level : Apply)
CO5	Explain how R language can be used for data mining methods.	(Cognitive level : Understand)
CO6	Employ commercially available dataset for clustering, association rule mining, text mining and classification using R language.	(Cognitive level : Apply)
CO7	Design and Develop an application for predicting cancer	(Cognitive level : Create)

Mapping of course outcomes with programme outcomes

	PO1	PO2	PO3	PO4	PO5	PO6	PO7
CO1	1						
CO2					2		
CO3					2		
CO4					2		
CO5	1						
CO6					3		
CO7					3		

MODULE I

R Introduction: Advantages of R over other programming languages, R studio, Expressions, variables, functions in R, Control structures in R, Data structures in R: Vectors, List, Factors, Matrix, Arrays, Data frames. Data: Different types of data, Quality of data, Methods for reading data into Rstudio: CSV and spreadsheets, reading data from packages, Reading data from Web/API,s, Reading from a JSON document, Reading a XML file, Importing data from SQL databases,. Preprocessing: basic preprocessing methods, Data preprocessing using R: Finding the missing values, Invalid values and outliers.

MODULE II

Data visualization: Data visualization tools in market, box plots, qqplot, histogram, Scatter Plot, Bar & Stack Bar Chart, Box Plot, Area Chart, Heat Map, Correlogram. Plotting with base graphics, Plotting with Lattice graphics, Plotting and coloring in R. Data summary statistics using R: Data Range, Frequencies and Mode, Mean and median, Measuring spread, Variance and Standard deviation. Time series data: Reading time series data, plotting time series data, decomposing time series data, Forecasts using exponential smoothing.

MODULE III

Classification in R: Nearest Neighbors classifier, Naïve Bayes classifier, Decision tree: terminologies associated with decision tree, decision tree representation in R, appropriate problems for decision tree learning, basic decision tree algorithm, Measuring features, Issues in Decision tree learning. Forecasting Numeric data- Regression methods, neural networks classifiers. Evaluating Model performance: Measuring performance for classification, Estimating future performance. Improving model performance with meta learning.

MODULE IV

Association rules mining in R: Frequent itemset, data structure overview,Apriori algorithm for association rule learning,measuring rule interest- support and confidence,building a set of rules with the Apriori principle, Mining algorithm interfaces in R, Auxiliary functions, Sampling from transactions. **Text mining in R:** Definition of text mining, Challenges in text mining, Text mining v/s data mining, Text mining in R, General architecture of text mining systems, Pre-processing of documents in R, Core Text mining operations, Mining frequent

patterns, associations and correlations, frequent item sets, closed item sets and association rules, frequent item sets mining methods, Pattern evaluation methods.

MODULE V

Clustering in R: Basic concepts in clustering, Types of clustering, types of clusters, Hierarchical clustering, K-means clustering, CURE algorithm.

Text Books:

1. Seema Acharya ,"Data Analytics using R " , McGrawHill, 2018
2. Brett lantz ,"Machine learning with R", 3rd Ed, , Packt, 2019

References :

3. Bharti Motwani, "Data Analytics with R", Wiley 2019.
4. Roger D. Peng , R programming for Data science, Springer 2008

Soft Computing Techniques

Course Outcomes
After completion of this course, students will be able to

CO1	Explain the applications of soft computing.	(Cognitive level : Understand)
CO2	Design back propagation neural network for real time problems	(Cognitive level: Create)
CO3	Analyze fuzzy classification.	(Cognitive level : Analyze)
CO4	Describe terminologies of Genetic algorithm.	(Cognitive level : Understand)
CO5	Apply GA based back propagation neural networks.	(Cognitive level: Apply)
CO6	Explain hybrid systems	(Cognitive level : Understand)
CO7	Develop applications of soft computing.	(Cognitive level : Create)

Mapping of course outcomes with programme outcomes

	PO1	PO2	PO3	PO4	PO5	PO6	PO7
CO1	2						
CO2		3					
CO3	2						
CO4	2						
CO5	3						
CO6	1						
CO7				3			

MODULE I

Introduction – Soft Computing, Requirements, Areas, Applications.

Neural Networks: Single layer Perceptrons , Back Propagation networks, Architecture of Backpropagation(BP) Networks, Backpropagation Learning,

Associative Memory, Adaptive Resonance theory and Self Organizing Map, Applications.

MODULE II

Fuzzy Systems - Fuzzy Set theory, Fuzzy versus Crisp set, Fuzzy Relation, Fuzzification, Minmax Composition, Defuzzification Method, Fuzzy Logic, Fuzzy Rule based systems, Predicate logic, Fuzzy Decision Making, Fuzzy Control Systems, Fuzzy Classification.

MODULE III

Genetic Algorithm - Working Principle, Various Encoding methods, Fitness function, GA Operators- Reproduction, Crossover, Mutation, Convergence of GA, Bit wise operation in GA, Multi-level Optimization. GA based Backpropagation Networks: GA based Weight Determination, K - factor determination in Columns.

MODULE IV

Hybrid Systems: Sequential Hybrid Systems, Auxiliary Hybrid Systems, Embedded Hybrid Systems, Neuro-Fuzzy Hybrid Systems, Neuro-Genetic Hybrid Systems, Fuzzy-Genetic Hybrid Systems.

MODULE V

Evolutionary Computation - Survival of the Fittest - Fitness Computation – Crossover – Mutation – Reproduction - Rank space Method. Case Studies: Applications of soft computing.

Text books:

1. N.P.Padhy, S.P.Simon, “Soft Computing with MATLAB Programming”, Oxford University Press, 2015.

2. S.N.Sivanandam , S.N.Deepa, “Principles of Soft Computing”, Wiley India Pvt. Ltd., 2nd Edition, 2011.

3. S.Rajasekaran, G.A.Vijayalakshmi Pai, “Neural Networks, Fuzzy Logic and Genetic Algorithm, Synthesis and Applications “, PHI Learning Pvt. Ltd., 2017.

References:

Genetic Algorithms: Search and Optimization, E. Goldberg.

Neuro-Fuzzy Systems, Chin Teng Lin, C. S. George Lee, PHI

PREDICTIVE ANALYTICS

Course Outcomes

After completion of this course, students will be able to

CO1	Describe the basics of predictors	(Cognitive level : understand)
CO2	Apply data reduction and feature extraction	(Cognitive level : Apply)
CO3	Apply regression model to predict.	(Cognitive level : Apply)
CO4	Apply descriptive statistics for prediction	(Cognitive level : Apply)
CO5	Explain various classification analytics techniques.	(Cognitive level : Understand)
CO6	Analyze feature selection	(Cognitive level : Analyze)
CO7	Evaluate different measurement errors in the predictive analytics	(Cognitive level : Evaluate)

Mapping of course outcomes with programme outcomes

	PO1	PO2	PO3	PO4	PO5	PO6	PO7
CO1	1						
CO2		2					
CO3			2				
CO4	2						
CO5	1						
CO6		3					
CO7				2			

MODULE 1

Introduction: - Prediction Versus Interpretation, Key Ingredients of Predictive Models, Predictive Modeling Process . Data Pre-processing:- Data Transformations for Individual Predictors- Centering and Scaling, Transformations to Resolve Skewness, Data Transformations for Multiple Predictors- Transformations to Resolve Outliers, Data Reduction and Feature Extraction, Removing Predictors- Predictor Correlations, Adding Predictors, Binning Predictors.

MODULE II

Over-Fitting and Model Tuning-The Problem of Over-Fitting Model Tuning, Data Splitting, Resampling Techniques. Regression Models- Quantitative Measures of Performance, Linear Regression- Partial Least Squares, Penalized Models, Nonlinear Regression Models - Neural Networks, K-Nearest Neighbors.

MODULE III

Classification Models:- introduction of Classification Models - Discriminant Analysis and Other Linear Classification Models - Nonlinear Classification Model -Na¨ıve Bayes - Support Vector Machines . Classification Trees and Rule-Based Models- Basic Classification Trees , Rule-Based Models -PART, Bagged Trees, Boosting- Ada Boost

MODULE 1V

Introduction to Feature Selection -Consequences of Using Non-informative Predictors-Approaches for Reducing the Number of Predictors-Factors That Can Affect Model Performance- Measurement Error in the Outcome, Measurement Error in the Predictors.

MODULE V

Predicting Cognitive Impairment Predicting Caravan Policy Ownership, The Effect of Class Imbalance- Sampling Methods-Cost-Sensitive Training-Job Scheduling

TEXTBOOK

1. Kuhn, Max, Kjell Johnson. "*Applied predictive modeling*". Springer, 2018 **.**

REFERENCE

1. Siegel, Eric. "*Predictive analytics": The power to predict who will click, buy, lie, or die*. John Wiley & Sons, 2013.
2. Abbott, Dean. "*Applied predictive analytics": Principles and techniques for the professional data analyst*. John Wiley & Sons, 2014.
3. Miner, Gary, "Practical text mining and statistical analysis for non-structured text data applications". Academic Press, 2012.

TEXT ANALYTICS

Course Outcomes

After completion of this course, students will be able to

CO1	Describe the basics of text analytics	(Cognitive level : Understand)
CO2	Interpret different summarization techniques	(Cognitive level : Apply)
CO3	Describe text analysis classification.	(Cognitive level :Understand)
CO4	Apply various text clustering techniques	(Cognitive level : Apply)
CO5	Explain n gram feature extraction and language models	(Cognitive level : Understand)
CO6	Interpret text data using visualization techniques	(Cognitive level : Apply)

Mapping of course outcomes with programme outcomes

	PO1	PO2	PO3	PO4	PO5	PO6	PO7
CO1	1						
CO2			2				
CO3	1						
CO4			2				
CO5	1						
CO6			3				

MODULE 1

An Introduction to Text Mining:- Approaches to Text Analysis, Information Extraction from Text- Named Entity Recognition Rule-based Approach ,Statistical Learning Approach.

Relation Extraction -Feature-based Classification, Kernel Methods, Weakly Supervised Learning Methods - Unsupervised Information Extraction -Relation Discovery and Template Induction, Open Information Extraction.

Basic text processing-tokenization, Stop word removal, Stemming and lemmatization, Language Models and Text Statistics.

MODULE II

Text Summarization Techniques-Topic Representation Approaches -Influence of Context-Indicator Representations and Machine Learning for Summarization- Graph Methods for Sentence Importance Selecting Summary Sentences- Greedy Approaches: Maximal Marginal Relevance, Global Summary Selection

MODULE III

Classification for Text Analysis-Text Classification -Identifying Classification problems , Classifier Models ,Building a Text Classification Application -Cross-Validation , Model Construction ,Model Evaluation ,Model Operationalization

Clustering for Text Similarity-Unsupervised Learning on Text -Clustering by Document Similarity- Distance Metrics , Partitive Clustering , Hierarchical Clustering -Modeling Document -Latent Dirichlet Allocation , Latent Semantic Analysis ,Non-Negative Matrix Factorization.

MODULE IV

Context-Aware Text Analysis-Grammar Based Feature Extraction ,Context-Free Grammars , Syntactic Parsers ,Extracting Key phrases ,Extracting Entities- n-Gram Feature Extraction , An n-Gram-Awar-,Choosing the Right n-Gram Window- n-Gram Language Models - Frequency and Conditional Frequency ,Estimating Maximum Likelihood

Text Visualization.-Visualizing Feature Space - Visual Feature Analysis , Guided Feature Engineering - Model Diagnostics -Visualizing Clusters ,Visualizing Classes ,Diagnosing Classification Error , Visual Steering -Silhouette Scores and Elbow Curves.

MODULE V

Graph Analysis of Text-Graph Computation and Analysis ,Creating a Graph-Based Thesaurus ,Analyzing Graph Structure , Visual Analysis of Graphs , Extracting Graphs from Text - Creating a Social Graph.

Entity Resolution -Entity Resolution on a Graph, Blocking with Structure ,Fuzzy Blocking-Cluster Computing with Spark- Sentiment Analysis- Deep Structure Analysis.

TEXT BOOK

1. Bilbro, Rebecca, Tony Ojeda, and Benjamin Bengfort. "Applied text analysis with Python." ,O'Reilly ,2019
2. Aggarwal, Charu C., and ChengXiang Zhai. "Mining Text Data." Springer Publications, 2012.

REFERENCE

1. Aggarwal, Charu C. "*Machine learning for text"*. New York: Springer, 2012.
2. Ignatow, Gabe, and Rada Mihalcea. "*An introduction to text mining": Research design, data collection, and analysis*. Sage Publications, 2017.

3. Struhl, Steven. "*Practical text analytics": Interpreting text and unstructured data for business intelligence*. Kogan Page Publishers, 2015.
4. Berry, Michael W., and Jacob Kogan, eds. "*Text mining: applications and theory*. John Wiley & Sons, 2010.

Social Network Analysis

Course Outcomes

After completion of this course, students will be able to

CO1	Explain different terminologies of graph and representation of graphs.	(Cognitive level : Understand)
CO2	Interpret social networks.	(Cognitive level : Apply)
CO3	Calculate centrality, betweeness centrality and directional relations.	(Cognitive level : Analyze)
CO4	Analyze structural relations	(Cognitive level : Analyze)
CO5	Evaluate social network for useful applications.	(Cognitive level : Evaluate)
CO6	Analyze social network using UCINET, PAJEK,ETDRAW,StOCNET, Splus-R,NodeXL,SIENA and RSIENA.	(Cognitive level : Analyze)

Mapping of course outcomes with programme outcomes

	PO1	PO2	PO3	PO4	PO5	PO6	PO7
CO1	1						
CO2	2						
CO3	3						
CO4		2					
CO5	2						
CO6					3		

Module I:

Introduction to Social Network Analysis, Mathematical representations of Social Networks: Notations for Social Network data – Graph theoretic, sociometric. Graphs – Subgraphs, Dyads, Triads, Nodal degree, Density, Walks, trails and paths, Connected graphs and components, Geodesics, distance and diameter, Connectivity, Isomorphic graphs and subgraphs.

Module II:

Directed graphs – Dyads, Nodal indegree and outdegree, Density, directed walks, paths and semi paths, Reachability and connectivity, Geodesics, distance and diameter. Signed graphs and signed

directed graphs Matrices – for graphs, digraphs, valued graphs, two-mode networks, Basic matrix operations, Computing simple network properties.

Module III:

Centrality: Actor centrality, Nondirectional relationships – degree, closeness, betweenness centrality, Directional relations – centrality.

Module IV:

Structural relationships – strong and weak ties, homophily, positive and negative relationships, Link analysis.

Module V:

Network dynamics – cascading behavior, small-world phenomenon, epidemics. Tools for Social Network Analysis - UCINET-PAJEK-ETDRAW-StOCNET- Splus-R-NodeXL-SIENA and RSIENA-Real world Social Networks (Facebook-Twitter etc.)

Text Books:

1. Social Network Analysis: Methods and Applications, Book by Katherine Faust and Stanley Wasserman
2. Networks, Crowds, and Markets: Reasoning about a Highly Connected World Book by David Easley and Jon Kleinberg
3. Social and Economic Networks Book by Matthew O. Jackson

NoSQL DATABASES

Course Outcomes

After completion of this course, students will be able to

CO1	Describe the relationship between Big Data and NoSQL databases.	(Cognitive level : Understand)
CO2	Explain the applications of NoSQL databases	(Cognitive level : Understand)
CO3	Design database application to perform CRUD operations.	(Cognitive level : Create)
CO4	Create columnar NoSQL databases using Apache HBASE and Apache Cassandra at enterprise level	(Cognitive level : Create)
CO5	Discuss detailed architecture to use distributed database patterns.	(Cognitive level : Understand)
CO6	Identify NoSQLdatabase issues including data integrity, security, and recovery.	(Cognitive level : Analyze)

Mapping of course outcomes with programme outcomes

	PO1	PO2	PO3	PO4	PO5	PO6	PO7
CO1	1						
CO2		1					
CO3			2				
CO4					2		
CO5	2						
CO6		3					

MODULE I

Database Revolutions- System Architecture- Relational Database- Database Design, Data Storage, Transaction Management- Data warehouse and Data Mining, Information Retrieval.

MODULE II

Overview and History of NoSQL Databases Four Types of NoSQL Database, The Value of Relational Databases, Getting at Persistent Data, Concurrency, Integration, Impedance Mismatch, Application and Integration Databases, Attack of the Clusters, The Emergence of NoSQL, Comparison of relational databases to new NoSQL stores.

MODULE III

Replication and sharding, MapReduce on databases. Distribution Models, Single Server, Sharding, Master-Slave Replication, Peer-to-Peer Replication, Combining Sharding and Replication, NoSQL Key/Value databases using MongoDB, Document Databases. MongoDB – CRUD operations, creating applications.

MODULE IV

Column- oriented NoSQL databases using Apache HBASE, Column-oriented NoSQL databases using Apache Cassandra, Column-Family Data Store- Features, Consistency, Transactions, Availability, Query Features, Scaling, Suitable Use Cases, Event Logging, Content Management Systems, Blogging Platforms, Counters, Expiring Usage.

MODULE V

Graph Databases , Graph NoSQL databases using Neo4,NoSQL in Big Data , Cloud - XML Databases- JSON Document Databases, Graph Databases. SSD and In-Memory Databases, Distributed Database Patterns— Distributed Relational Databases- Non-relational Distributed Databases- MongoDB.

Text Books:

I. NoSQL Distilled: A Brief Guide to the Emerging World of Polyglot Persistence, Sadalage, P. & Fowler, Pearson Education

References:

1. Abraham Silberschatz, Henry F. Korth, S. Sudarshan, "Database System Concepts", Sixth Edition, McGrawHill, 2010

2. Guy Harrison, "Next Generation Databases", Apress, 2015.

3. Eric Redmond, Jim R Wilson, "Seven Databases in Seven Weeks", LLC. 2012.

4. Dan Sullivan, "NoSQL for Mere Mortals", Addison-Wesley, 2015.

5. Adam Fowler, "NoSQL for Dummies ", John Wiley & Sons, 2015.

IMAGE AND VIDEO ANALYTICS

Course Outcomes
After completion of this course, students will be able to

CO1	Describe image representations.	(Cognitive level : understand)
CO2	Apply DFT, DCT, DWT and Hadamard for image transformations	(Cognitive level : Apply)
CO3	Explain filtering, sharpening and edge detection.	(Cognitive level : Understand)
CO4	Employ statistics for modeling.	(Cognitive level : Apply)
CO5	Explain video and image segmentation.	(Cognitive level : Understand)
CO6	Design a system for analytics in WSNs	(Cognitive level : create)
CO7	Design analytical model for real time sensing.	(Cognitive level : Create)

Mapping of course outcomes with programme outcomes

	PO1	PO2	PO3	PO4	PO5	PO6	PO7
CO1	1						
CO2	3						
CO3	2						
CO4	2						
CO5	2						
CO6					3		
CO7					3		

MODULE I

Digital image representation-Visual Perception-Sampling and Quantization-Basic Relations between Pixels.

MODULE II

Mathematical Tools Used in Digital Image Processing: Vector and Matrix Operations-Image Transforms (DFT, DCT, DWT, Hadamard).

MODULE III

Fundamentals of spatial filtering: spatial correlation and convolution-smoothing blurring-sharpening- edge detection-Basics of filtering in the frequency domain: smoothing-blurring-sharpening

MODULE IV

Histograms and basic statistical models of image, Color models and Transformations-Image and Video Segmentation-Image and video denoising-Image and Video Enhancement-Image and Video compression.

MODULE V

Object detection and recognition in image and video-Texture models Image and Video 25 classification models-Object tracking in Video

Applications

Case studies- Industrial-Retail- Transportation & Travel-Remote sensing- Video Analytics in WSN: IoT Video Analytics Architectures.

TEXT BOOKS

1. Ed. Al Bovik ,"Handbook of Image and Video Processing", 2nd Edition, Academic Press, 2000.

2. J. W. Woods, "Multidimensional Signal, Image and Video Processing and Coding",2nd Edition, Academic Press, 2011.

REFERENCE BOOKS

1. Rafael C. Gonzalez and Richard E. Woods," Digital Image Processing", 3rd Edition, Prentice Hall, 2008.

2. A. M. Tekalp, "Digital Video Processing", 2nd Edition, Prentice Hall, 2015.S. Shridhar, "Digital Image Processing", 2nd Edition,

Oxford University Press, 2016.

HEALTHCARE DATA ANALYTICS

Course Outcomes

After completion of this course, students will be able to

CO1	Explain the basics concepts of Electronic healthcare data	(Cognitive level : Understand)
CO2	Apply Strategies for optimizing healthcare data quality	(Cognitive level : Apply)
CO3	Apply Phenotyping Algorithms	(Cognitive level : Apply)
CO4	Apply Sensor Data in Healthcare	(Cognitive level : Apply)
CO5	Explain techniques for analysing health care data.	(Cognitive level : Understand)
CO6	Interpret Temporal Data Mining methods for Healthcare Data	(Cognitive level : Apply)
CO7	Develop efficient clinical decision support systems.	(Cognitive level : Create)

Mapping of course outcomes with programme outcomes

	PO1	PO2	PO3	PO4	PO5	PO6	PO7
CO1	1						

CO2			2				
CO3			2				
CO4			2				
CO5	1						
CO6			2				
CO7					3		

MODULE I

Introduction: Electronic Health care data-Components, Coding Systems, Benefits, Challenges, Sources of data-Surveys, Clinical trials, Medical records, Claims, Other sources, Uses and limitations. Role of analytics in supporting a data-driven learning healthcare system.

MODULE II

Examine epidemiological concepts in healthcare analytics, application, Health statistics-mortality, morbidity, risk adjustment. Methods for selecting, preparing, querying and transforming healthcare data. Strategies for optimizing data quality; querying tools and methods including data preparation and transformation techniques.

MODULE III

Quality control in healthcare systems-tools for identifying quality problems, Simulation methods. Analyse and discover patterns, Medical genomics, Phenotyping Algorithms, Biomedical Image Analysis-Mining of Sensor Data in Healthcare.

MODULE IV

Biomedical Signal Analysis-Genomic Data Analysis for Personalized, Review of Clinical Prediction Models-Temporal Data Mining for Healthcare Data, Visual Analytics for Healthcare. Predictive Models for Integrating Clinical and Genomic Data-Information Retrieval for Healthcare-Privacy-Preserving Data Publishing Methods in Healthcare.

MODULE V

Applications: Data Analytics for Pervasive Health-Fraud Detection in Healthcare-Data Analytics for Pharmaceutical Discoveries-Clinical Decision Support Systems- Computer-Assisted Medical Image Analysis Systems- Mobile Imaging and Analytics for Biomedical Data.

TEXT BOOK

1. Chandan K. Reddy and Charu C Aggarwal, “Healthcare data analytics”, CRC press, 2015

2. Trevor L. Strome."Healthcare Analytics for Quality and Performance Improvement", Wiley,2103

REFERENCES

1. Clinical and Translational Science: Principles of Human Research" ,Elseiver,2nd Edition.

Fraud Analytics

Course Outcomes

After completion of this course, students will be able to

CO1	Explain fraud triangle and defense mechanisms.	(Cognitive level : understand)
CO2	Analyze and detect fraud using data analysis tools and data set	(Cognitive level : Analyze)
CO3	Apply Benford's law to detect a fraud.	(Cognitive level : Apply)
CO4	Apply descriptive statistics for preprocessing	(Cognitive level : Apply)
CO5	Describe various descriptive analytics techniques.	(Cognitive level : Understand)
CO6	Perform fraud detection using Spark, Kaffka and Cassandra	(Cognitive level : Apply)
CO7	Design predictive analytical model for fraud detection using machine learning algorithms	(Cognitive level : Create)

Mapping of course outcomes with programme outcomes

	PO1	PO2	PO3	PO4	PO5	PO6	PO7
CO1	1						
CO2					1		
CO3	2						
CO4	2						
CO5	1						
CO6					2		
CO7				3			

Module I

Introduction – Fraud triangle, Different types of fraud, Importance of Fraud detection, Difference between fraud detection and Audit – goals, outcomes, Fraud prevention – Control measures for prevention, Enterprise risk management, Risk, Risk assessment. Data Analysis techniques – Preventing fraud, Detecting fraud.

Module II

Collection of Sampling and Preprocessing Data – Types of data sources, Merging data sources, sampling, types of data elements, Benford's law. Descriptive statistics – missing values, outlier detection and treatment, dimensionality reduction.

Module III

Predictive analytics for fraud detection - linear regression, logistic regression, decision trees, neural networks, SVM, ensemble models, random forests

Module IV

Descriptive analytics - peer group analysis, break point analysis. Clustering - hierarchical clustering, non-hierarchical clustering. Self-organizing maps.

Module V

Fraud detection models using social network analytics - homophily, featurization, egonets, PageRank. Case Study : Real time fraud detection –using Spark ,Kaffka, Cassandra.

Text Book

1.Fraud Analytics Using Descriptive, Predictive, and Social Network Techniques: A Guide to Data Science for Fraud Detection (Wiley and SAS Business Series) 1st Edition by Bart Baesens

References

1. Computer-Aided Fraud Prevention and Detection 1st Edition by David Corderre's

2.Data Mining –Concepts and techniques by Jiawei Han

Blockchain Technologies

Course Outcomes

After completion of this course, students will be able to

CO1	Describe the basic concepts – public ledger, block chain and block chain data structure	(Cognitive level : understand)
CO2	Explain asymmetric key cryptography and hashing algorithms.	(Cognitive level : Understand)
CO3	Describe consensus and proof of work.	(Cognitive level : Understand)
CO4	Develop smart contract using tools like remix	(Cognitive level : Create)
CO5	Create hyper ledger using Ethereum.	(Cognitive level : Create)
CO6	Describe public block chain	(Cognitive level : Understand)

Mapping of course outcomes with programme outcomes

	PO1	PO2	PO3	PO4	PO5	PO6	PO7
CO1	1						
CO2	1						
CO3	2						
CO4					2		
CO5					3		
CO6	3						

MODULE I Introduction to Blockchain, Blockchain Data structure, Hash chain, distributed database, Index structure, Transactions, Asymmetric-Key Cryptography, Addresses and Address Derivation, Private Key Storage, Ledgers, Blocks, Chaining Blocks.

MODULE II Consensus and multiparty agreements: protocols, Proof of Work, Proof of Stake, Delegated Proof of Stake, Proof of Elapsed Time, Deposit based consensus, Proof of importance.
Federated consensus or federated Byzantine consensus, Reputation-based mechanisms, Practical Byzantine Fault Tolerance.

MODULE III Blockchain implementation. Forking-Soft Fork, Hard Forks. Smart contract programing. Blockchain Platforms – Cryptocurrencies - Bitcoin, Litecoin, Ethereum

MODULE IV Hyperledger, Ethereum. Decomposing the consensus process , Hyperledger
fabric components , Chaincode Design and Implementation.

MODULE V Blockchain-Decentralized identity and public blockchain. IPFS protocol and Blockchain. Blockchain Concurrency and scalability. Network models and timing assumptions, security aspects of decentralized ledger systems, block chain networks.

TEXT BOOK

1. Andreas M Antonopoulos , "Mastering Bitcoin: Unlocking digital crypto currencies", ORELLY,2015.
2. Melanie, "Blockchain: Blue print for new economy", ORELLY,2015.
3. https://www.ibm.com/blockchain/in-en/hyperledger.html.

REFERENCES

1. Don Tapscott, "Block chain and Crypto currency", 2016. Draft NISTIR 8202, Blockchain Technology Overview - NIST CSRC, 2018.
2. Imran Bashir, Mastering Blockchain, 2017.
3. Andreas M. Antonopoulos, Mastering Bitcoin - Programming the Open Blockchain, O'Reilly Media, Inc., 2017
4. Ethereum Programming, Alex Leverington, Packt Publishing Limited, 2017
5. Bitcoin and Cryptocurrency Technologies: A Comprehensive Introduction. A. Narayanan, J. Bonneau, E. Felten, A. Miller, S. Goldfeder, Princeton University Press, 2016
6. The Science of the Blockchain, Roger Wattenhofer, CreateSpace Independent Publishing Platform, 2016
7. Melanie Swan, Blockchain - Blueprint for a new economy, O'Reilly Media, Inc., 2015.
8. Abhijit Das and VeniMadhavan C. E., Public-Key Cryptography: Theory and Practice: Theory and Practice, Pearson Education India, 2009.

****