BISHOP HEBER COLLEGE (AUTONOMOUS), TIRUCHIRAPPALLI 620017

M. Sc., Data Science

(Applicable to Candidates admitted from Academic Year 2023-2024 onwards)

Sem	Part	Course	Course Title	Course Code	Hours /Week	Credits	Marks		
							CIA	ESE	Total
I	Part A	Core Paper I	Programming and Data Structures with Python	P23DS101	6	4	25	75	100
		Core paper II	Modern Database Systems	P23DS102	5	4	25	75	100
		Core Practical I	Programming and Data Structures with Python Lab	P23DS1P1	5	2	40	60	100
		Core Practical II	Modern Database Systems Lab	P23DS1P2	4	2	40	60	100
		Elective I	Mathematical Foundation for Data Science	P23DS1:A	4	3	25	75	100
			Design and Analysis of Algorithms	P23DS1:B					
		Elective II	Essential Statistics for Data Science	P23DS1:C	4	3	25	75	100
			Basics of Bioinformatics	P23DS1:D					
		SEC I	Data Visualization Lab	P23DSPS1	2	2	40	60	100
II	Part A	Core Paper III	Data Analytics and Storytelling	P23DS203	5	4	25	75	100
		Core Paper IV	Applied Machine Learning	P23DS204	5	4	25	75	100
		Core Practical III	Data Analytics and Storytelling Lab	P23DS2P3	4	2	40	60	100
		Core Practical IV	Applied Machine Learning Lab	P23DS2P4	4	2	40	60	100
		Elective III	Natural Language Processing	P23DS2:A	4	4	25	75	100
			Multivariate Analysis	P23DS2:B					
		Elective IV	Time Series Analysis and Forecasting	P23DS2:C	4	4	25	75	100
			Health Care Data Analytics	P23DS2:D					
		SECII	Natural Language Processing Lab	P23DSPS2	4	2	40	60	100
III	Part A	Core Paper V	Principles of Deep Learning	P23DS305	5	5	25	75	100
		Core Paper VI	Big Data Management and Analytics	P23DS306	6	5	25	75	100
		Core Practical V	Principles of Deep Learning Lab	P23DS3P5	5	4	40	60	100
		Core Practical VI	Big Data Management and Analytics Lab	P23DS3P6	5	4	40	60	100
		Elective V	Computer Vision	P23DS3:A	4	4	25	75	100
			Supply Chain Management	P23DS3:B					
		SEC III	Data Science Project Management Lab	P23DSPS3	5	2	40	60	100
		Internship	Internship/ Industrial Activity	P23DS311	---	2	---	---	100
IV	Part A	Core Paper VII	Digital Innovation Management	P23DS407	5	4	25	75	100
		Core Practical VII	Web Application Development Lab	P23DS4P7	5	4	40	60	100
		Core Project	Project with Viva-Voce	P23DS4PJ	14	8	40	60	100
		Elective VI	Social Media and Graph Analytics	P23DS4:A	4	4	25	75	100
			Customer Relationship Management	P23DS4:B					
		SEC IV	Data Science Soft skills	P23DS4S4	2	2	25	75	100
		Extension	Extension Activity	P23ETA41	---	1	---	---	---
	Part B	VLO	The Big Picture	P23VLO41	2	2	100	---	100
			Flying High	P23VLO42					
				Total Credits		**91+2**			

Core I: Programming and Data Structures with Python

Course Objectives

This course introduces students the language features of Python. Specifically, data structures, regular expressions, data visualization and internet programming features are introduced, besides sorting, searching and classic data structures.

Course Outcomes

On successful completion of this course, students will be able to:
CO1: Develop applications using python features and files
CO2: Develop applications using Python data structures
CO3: Implement sorting, searching and data structures in Python
CO4: Develop OOP in Python, access internet and database data
CO5: Develop GUI and database applications

Unit-1. Python Basics, Functions, Loops, Strings and Files

Variables – Getting Inputs – Conditions – Catching exceptions – Function calls – Built-in functions – Type conversion functions and math functions – Parameters and arguments –While statement – Infinite loops – Continue statement – For loops – Strings – Slice –– The in operator – String comparison – String methods – parsing strings – Format operator - Opening files – Text files – Reading files – Searching through files – Writing files

Unit-2. Python built-in Data Structures

Traversing list – List operations – List slice – List methods – Deleting elements – Built-in list functions – Objects, value and aliasing – List arguments. Dictionaries – Files and dictionaries – Looping and dictionaries – Tuples – Comparing tuples – Tuple assignments – Dictionaries and tuples – Tuples as keys in dictionaries – Map, filter and reduce functions – Regular expressions

Unit-3. Sorting, searching and classic Data Structures

Recursion - Algorithm analysis – Big-O notation – Bubble sort – Selection sort – Insertion Sort – Merge sort – Quick sort – Python implementation of sorting methods - Linked List - Stack – Queue – Deque - Priority queue – Heaps – Search trees – AVL Trees - Python implementation of classic data structures

Unit-4. OOP and Internet Programming

Creating objects – Encapsulation – Classes as types – Object lifecycle – Instances – Inheritance. Regular expressions – Character matching – Extracting data – Escape character – Designing simple web browser using sockets – Retrieving images using HTTP – Retrieving web pages using urllib – Reading binary files using urllib

Unit-5. GUI and Database Programming

GUI using Tkinter – Create text, button, radio button, check button and listbox in tkinter - Create, delete and add data to tables in SQLite – Querying using SQL Select – Updating and deleting rows – Exceptions – CRUD Operations

Text Books

1. Charles R. Severance, Python for Everybody: *Exploring data using Python 3*, Schroff Publishers, 1ed, 2017, ISBN 978-9352136278
2. Paul J. Deitel, Harvey M. Deitel. Intro to Python for Computer Science and Data Science. Pearson Ed. 2022
3. Brad Miller, David Ranum, Problem Solving with Algorithms and Data Structures using Python, Franklin, Beedle and Associates, 2ed, 2013 (Unit II and III)
4. Tony Gaddis, Starting out with Python, 5ed, Pearson Education, 2021 (Unit V)

References

1. Allen B. Downey, —Think Python: How to Think like a Computer Scientist, 2ed, OReilly, 2016
2. M. Goodrich, R. Tamassia and M. Goldwasser. Data structures and algorithms in Python. Wiley. 2016
3. Robert Sedgewick, Kevin Wayne, Robert Dondero, Introduction to Programming in Python: An Inter - disciplinary Approach, Pearson India Education Services Pvt. Ltd., 2016.

Core II: Modern Database Systems

Course Objectives

The widespread emergence of big data storage needs has driven the development and adoption of a new class of non-relational databases commonly referred to as NoSQL databases. This course will explore the origins of NoSQL databases and the characteristics that distinguish them from traditional relational database management systems. Core concepts of NoSQL databases will be presented, followed by an exploration of how different database technologies implement these core concepts.

Course Outcomes

On successful completion of this course, students will be able to:

CO1: Model data using ER diagrams
CO2: Demonstrate competency in designing SQL and NoSQL database management systems.
CO3: Demonstrate competency in describing how NoSQL databases differ from relational databases
CO4: Demonstrate competency in selecting a particular NoSQL database for specific use cases.
CO5: Implement databases using SQL, MongoDB and Neo4J

Unit-1. Structured Query Language-I

ER Model: Entity types, Attribute types, Relationship types – Weak entity types, Ternary relationship types – Examples of ER model. Enhanced ER model: Specialization/Generalization – Categorization - Aggregation – Examples of EER. Relational DB Process and outcome approach - Simple Queries on one table – First look at joins – Sub queries.

Unit-2.Structured Query Language-II

Self Joins: Self relationships, Questions involving Both – Multiple relations between tables – Set operations – Aggregate Operations – Window functions – Efficiency considerations: Indexing and Join Techniques.

Unit-3. MongoDB-I

Introduction: MongoDB document, collection and database – Basic Operations – Datatypes – Creating, deleting, updating documents: insert, batch insert, remove, find, findone, update – arrays – insert – Updating multiple documents

Unit-4. MongoDB-II

Comparison operators – OR and NOT queries – Querying arrays – Querying on embedded documents – WHERE queries – Limits, skips and sort – Compound Index – Unique index – Sparse Index – Pipeline aggregation: MATCH, PROJECT, GROUP and UNWIND clauses.

Unit-5. Neo4J and Cypher

Labeled Property Graph Model – Querying graphs using Cypher: CREATE AND ASSERT, MATCH, WHERE and RETURN clauses– ORDER BY – WITH clause – Case Study: Telent.net Social recommendations application.

Text Books

1. Clare Churcher. *Beginning SQL Queries: From Novice to Professional*, APress, 2ed, 2016. ISBN 9781484219546
2. WilfriedLemahieu, SeppevandenBroucke and Bart Baesens. *Principles of Database Management: The Practical Guide to Storing, Managing and Analyzing Big and Small Data*, Cambridge University Press, 2018. ISBN 978-1-107-18612-5 (Chapter 3 ER diagram only)
3. Kristina Chodorow, *MongoDB: The Definitive Guide*, 2ed, Oreilly Publishers
4. Ian Robinson, Jim Webber and Emil Eifrem. Graph Databases: New Opportunities for connected data. 2ed, Oreilly Publishers. ISBN 978-1491930892.

References

1. Eric Redmond; Jim R. Wilson. *Seven Databases in Seven Weeks: A Guide to Modern Databases and the NoSQL Movement*. Pragmatic Bookshelf. 2012. ISBN: 1934356921Pramod J. Sadalage; Martin Fowler. *NoSQL Distilled: A Brief Guide to the Emerging World of Polyglot Persistence*. Addison-Wesley. 2012 ISBN: 0321826620
2. Adam Fowler. *NoSQL for Dummies*. John Wiley. 2015. ISBN 978-1-118-90574-6

3. Guy Harrison. *Next Generation Databases*. APress. 2016. 978-1-484213-30-8
4. Thomas M. Connolly and Carolyn E. Begg. Database Systems: *"A Practical Approach to Design, Implementation, and Management",* 6th Edition, Pearson, 2015.

Core Practical I: Programming and Data Structures with Python Lab

Course Objectives

This lab course enables students to master the language features of Python programming languages. Specifically, data structures, regular expressions, files and data visualization features are introduced, besides sorting, searching and classic data structures

Course Outcomes

On successful completion of this course, students will be able to:

CO1: Write simple Python programs using Python data structures
CO2: Develop object oriented programs in Python
CO3: Manipulate files using Python
CO4: Access internet and database data
CO5: Implement and apply sorting and searching methods and data structures

Lab Exercises

Develop applications that will demonstrate the following Python and data structures features

- Functions
- String processing
- List processing
- Dictionaries
- Tuples
- File processing
- Regular Expressions
- OOP
- Retrieving webpages from web
- Data visualization in Matplotlib, Seaborn
- Database programming
- Concurrent programming
- Sorting, searching and data structures

Core Practical II: Modern Database Systems Lab

Course Objectives
This lab course will explore the features of SQL and NoSQL languages for solving business applications. Language features of MongoDB and Neo4J are introduced and applied in great depth. This course also enables students to learn and get certifications from MongoDB and Neo4J.

Course Outcomes
On successful completion of this course, students will be able to:

CO1: Model data using ER diagrams
CO2: Create databases and write queries using SQL
CO3: Create databases and write queries using MongoDB
CO4: Create databases and write queries using Neo4J
CO5: Process large relational, document and graph databases for business applications

Lab Exercises

Develop applications that will demonstrate the following concepts using SQL and NoSQL

- Drawing E-R Diagrams for the given business problem
- Designing tables from E-R Diagrams
- Creating tables using SQL and NoSQL features
- Developing queries and subqueries
- Performing Join operations
- Creating Indexes
- Solving real world business use cases using SqLite3, MongoDB and Neo4J
- Processing big datasets of various file formats such as CSV, TXT and others

Elective I: Mathematical Foundation for Data Science

Course Objectives

Every day all over the world, large amounts of data are generated by various businesses, research organizations and social media. Data Science is the study, application, and development of methods to learn from this data. So that many improvements can be made in products, services, advertising campaigns, public health and safety and others. Linear algebra, geometry, calculus and probability play a fundamental role in the theory of Data Science. This course introduces the basic notions of vector spaces, inner product, matrix decomposition, derivatives and probability distributions.

Course Outcomes

On successful completion of this course, students will be able to

CO1: Solve systems of linear equations by use of the matrix
CO2: Determine the Orthogonality and Basis
CO3: Compute and use eigenvectors and eigenvalues
CO4: Analyze the gradients and PDE
CO5: Use the probability concepts in Data science

Unit - I: Linear Algebra

Systems of Linear Equations – Matrices - Solving Systems of Linear Equations - Vector Spaces - Linear Independence - Basis and Rank - Linear Mappings - Affine Spaces.

Unit - II: Analytic Geometry

Norms - Inner Products - Lengths and Distances - Angles and Orthogonality - Orthonormal Basis - Orthogonal Complement - Inner Product of Functions - Orthogonal Projections - Rotations

Unit - III : Matrix Decompositions

Determinant and Trace - Eigenvalues and Eigenvectors - Cholesky Decomposition – Eigen decomposition and Diagonalization - Singular Value Decomposition Matrix Approximation - Matrix Phylogeny -

Unit – IV : 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

Unit – V : Probability and Distributions

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.

Text Book

1. Marc Peter Deisenroth, A. Aldo Faisal, Cheng Soon Ong, **"** *Mathematics for Machine Learning***"**, Cambridge Press, 2019 (Chapters 2, 3, 4,5,6)

References

1. Gilbert Strang,*"Introduction to Linear Algebra***",** 3ed, Cambridge Press, 2003.
2. M. D. Weir, J. Hass, and G. B. Thomas*, "Thomas' calculus*", Pearson, 2016.

Elective-I: Design and Analysis of Algorithms

Course Objectives
This course introduces students the concepts of designing and analysing algorithms, sorting data, greedy algorithms and graph algorithms.

Course Outcomes
On successfulcompletion of this course, students will be able to:
CO1: Analyse the asymptotic performance of algorithms.
CO2: Solve problems using key techniques of algorithm design.
CO3: Develop optimal solution by applying various methods and differentiate polynomial and non-polynomial problems.

Unit-1. Basics of Algorithm Analysis
Basics of Algorithms and Mathematics: Introduction to an algorithm, Mathematics for Algorithmic Sets, Linear Inequalities and Linear Equations. The efficient algorithm, Average, Best and worst-case analysis of Time Complexity and Space Complexity.**Analysis of Algorithm:** Amortized analysis, Asymptotic Notations, analysing control statement, Loop invariant and the correctness of the algorithm, Master's Theorem.**Sorting Algorithms and analysis:** Bubble sort, Selection sort, Insertion sort, Shell sort, Heap sort, sorting in linear time: Bucket sort, Radix sort and Counting sort

Unit-2.Divide and Conquer Algorithm
Introduction, Recurrence and different methods to solve recurrence, multiplying large Integers Problem, Problem Solving using divide and conquer algorithm.Binary Search, Max-Min problem, Sorting (Merge Sort, Quick Sort), Matrix Multiplication.

Unit-3. Dynamic Programming and Greedy Algorithm
Dynamic Programming: Introduction, The Principle of Optimality, Problem Solving using Dynamic Programming–Calculating the Binomial Coefficient, Assembly Line-Scheduling. Matrix chain multiplication, Longest Common Subsequence, All Points Shortest path.**Greedy Algorithm:** General Characteristics, Problem solving using Greedy Algorithm- Activity selection problem. Minimum Spanning trees (Kruskal's algorithm, Prim's algorithm).

Unit-4. Graphs
Graphs: Shortest paths, The Knapsack Problem, Job Scheduling Problem, Huffman code.Exploring Graphs Algorithms: - Applications of DFS- bi-connectivity, topological sort, Articulation point, Connected components. **Backtracking and Branch and Bound:**Introduction, The Eight queen's problem, Minimax principle.

Unit-5. Selected Problems in Algorithms
String Matching: Introduction, The naive string matching algorithm, The Rabin Karp algorithm, String Matching with finite automata, The Knuth Morris-Pratt algorithm.**Approximation algorithms:**Travelling Salesman problem, Hamiltonian problem, Vertex Cover Problem.**Introduction to NP-Completeness:** The class P and NP, Polynomial reduction, NP Completeness Problem, NP-Hard Problems.

Text Books
1. Cormen, Leiserson, Rivest, and Stein, Introduction to Algorithms (3rd Edition), MIT Press, 2009. ISBN- 978-0262033848.
2. Jon Kleinberg and Eva Tardos, Algorithm Design, 1ed, Pearson, 2005. ISBN- 978-0321295354.

References
1. V. Aho, J. E. Hopcroft and J. D. Ullman, Data Structures and Algorithms, (Addison Wesley), Pearson, 1983. ISBN- 978-0201000238.

Elective-II: Essential Statistics for Data Science

Course Objectives

This course covers topics in Statistics from basics to advanced level that every Data Science student should master and apply for the industry applications. Great depth of coverage for the topics from Regression Analysis is also given in this course.

Course Outcomes

On successful completion of this course, students will be able to
CO1: Identify the methods of descriptive statistics and variablity.
CO2: Examine the different tests of the statistical inferences
CO3: Demonstrate the nonparametric statistics methods
CO4: Classify the different types of regression methods for data analytics
CO5: Analyze the different properties of the regression methods.

Unit I: Descriptive Statistics

Introduction to Statistics - Organizing Data Using Tables and Graphs- Measures of Central Tendency: Mode – Median – Mean. Measures of Variability: Variability – Range - Interquartile Range - Standard Deviation.

Unit II: Inferential Statistics – I

Sampling Distribution of Means: Sampling Distribution - Central Limit Theorem. Hypothesis Testing: Hypothesis Testing Steps -Effect Size for a Z-Test - Assumptions – Errors – Power. One-Sample t Test: t-Statistics – t- Distributions - One-Sample t Test – Effect Size – Assumptions. Two-Sample t Test: Independent Samples Design: Calculations – Hypothesis Testing – Effect Size – Assumptions. Two-Sample t Test: Related Samples Design: Calculations – Hypothesis Testing – Effect Size – Assumptions.

Unit III: Inferential Statistics - II

Confidence Interval versus Point Estimation: Introduction- Point Estimates - Confidence Intervals – One Sample t- Test - Two-Sample t Test: Independent Samples Design – Repeated Measure Design - Degree of Confidence Vs. Degree of Specificity One-Way Analysis of Variance: Introduction – Variance – F- statistics – Hypothesis Testing with F- Statistic - F- Distribution Table - Notations for ANOVA - Calculations – Hypothesis Testing – Effect Size – Assumptions. Chi-Square: Chi-Square - Chi-Square Statistic – Assumptions- Goodness of Fit - Goodness of Fit for Known Proportions- Goodness of Fit for No Preference – Test of Independence - Nonparametric Statistics for Ordinal Data: Mann-Whitney U Test - Kruskal-Wallis H Test. Correlation: Introduction – Scatter Plot - Pearson Product Moment Correlation - Hypothesis Testing - Coefficients of Determination and Nondetermination – Interpretation and Uses of The Pearson Correlation.

Unit IV: Regression Analysis - I

Regression Model - Goals of Regression Analysis - Statistical Computing in Regression Analysis - Simple Linear Regression – Multiple Linear Regression – Logistic Regression – Poisson Regression

Unit V: Regression Analysis - II

Detection of Outliers and Influential Observations: Detection of Outliers in Multiple Linear Regression - Detection of Influential Observations in Multiple Linear Regression - Test for Mean-shift Outliers - Graphical Display of Regression Diagnosis. Model Selection: Effect of Underfitting and Overfitting - All Possible Regressions – Stepwise Selection. Model Diagnostics: Test Heteroscedasticity - Detection of Regression Functional Form

Text Books:

1. Cheryl Ann Willard, *"Statistical Methods: An Introduction to Basic Statistical Concepts and Analysis"* ,Routledge, 2020. (Unit – I – III)
2. Xin Yan &Xiaogang Su, *"Linear Regression Analysis : Theory and Computing*"**,** World Scientific Pulishing Ltd, 2009. (Unit – IV: Chapter 1,2,3, 8.5,8.6; Unit – V: Chapters 4.2,4.3, 4.4, 4.5, 5.1-5.3, 6.1, 6.2)

Reference Books:

1. John.E.Freund, Irwin Miller, Marylees Miller *"Mathematical Statistics with Applications*", 8th, Prentice Hall of India, 2014

2. Ross, Sheldon. M, *"Introduction to Probability and Statistics for Engineers and Scientists"*, Academic Press, 2009
3. D.C Montgomery, E.A Peck and G.G Vining, *"Introduction to Linear Regression Analysis"*, John Wiley and Sons, 2003.
4. S. Chatterjee and AHadi, **"***Regression Analysis by Example***"**, 4th Ed., John Wiley and Sons, Inc, 2006

Elective-II : Basics of Bioinformatics

Course Objectives
This course will offer basic understanding of Bioinformatics and its applications to molecular biology, clinical medicine and other disciplines. It enables students to create profound advances in their understanding of life and improvements in the health of humans and other living organisms.

Course Outcomes
On successful completion of this course, students will be able to:
CO1: Design their own databases according to their data of research.
CO2: Acquire knowledge on collecting latest and updated biological information from the browsers.

Unit-1. Introduction to Bioinformatics
Overview- Definition and History. Milestones in Bioinformatics. Methods in Bioinformatics. Role of Bioinformatics in various fields. Useful Bioinformatics web sites. Dogmas: Central and Peripheral. Introduction to single letter code of aminoacids, Symbols used in nucleotides.

Unit-2. Biological Data and Databases
Introduction to Biological Databases- Nucleotide sequence database, Protein sequence & Structure Databases, Organism specific databases, Metabolic pathway databases, Bibliographic databases, Biodiversity databases and Specialized databases.

Unit-3. Sequence formats and Information Retrieval
Sequence Formats in Biological databases- FASTA, Phylip, Clustal, Genbank, EMBL, SWISS PROT. Data retrieval- Entrez, SRS, Protein identification resources (PIR), Expasy, Ensembl.

Unit-4. Database Searches
Similarity, homology, assessing significance of sequence similarity-Z score, P value, E value, Similarity search programs- fast searching methods-BLAST, FASTA, Dynamic programming searching methods, profile based methods-PSI-BLAST, Sensitivity and Specificity.

Unit-5. Applied Bioinformatics
Commercial bioinformatics, Survey of bioinformatics companies in India and abroad –Economics prospects, pharamainformatics, combinatorial chemistry, HT screening – in silico screening - from lead to commercialization.

Text Books
1. Arthur M. Lesk, Introduction to Bioinformatics, Oxford University Press, New Delhi, 2003.

References
1. Attwood, T.K and D.J. Parry-Smith, Introduction to Bioinformatics, Pearson Education Ltd., New Delhi, 2004.
2. Westhead D.R., J.H Parish and R.M. Twyman, Instant notes in Bioinformatics, Viva Books Pvt. Ltd., 2003.
3. Manju Bansal, Basic Bioinformatics, Atlantic Publishers & Distributors, 2009.

SEC-I: Data Visualization Lab

Course Objectives

The world generates huge amount of data every day. Businesses and society would use this huge amount of data in order to solve problems from business and society. In this course, students will be able to generate powerful reports and dashboards using Tableau and PowerBI which will help business and society to make effective decisions.

Course Outcomes

On successful completion of this course, students will be able to:

CO1: Connect and prepare data using Tableau Desktop
CO2: Explore and Analyze data using Data Prep tools
CO3: Share insights through Tableau Public
CO4: Prepare and model data using Power BI and Power Query
CO5: Visualise and analyse data using Power BI tools
CO6: Deploy and manage deliverables using Power BI service

Lab Exercises

Develop applications that will demonstrate the following concepts using Tableau and PowerBI

Tableau

- Connecting to data sources and data preparation: create and modify data connections, manage data properties
- Understanding Tableau concepts: Dimensions and Measures, Discrete and continuous fields, Aggregation, Calculations
- Exploring and analysing data: Create basic charts, Organise data and apply data filters, apply analytics to worksheet, creating maps
- Sharing insights: Formatting view of presentation, modifying dashboards and storytelling

Power BI

- Data Preparation: Getting data from different sources, Data Profiling, Cleaning, Transforming and loading the data
- Modelling the data: Design and develop a data model, create measures using DAX, Optimise model performance
- Visualising the data: Create reports and dashboards, improve reports for usability
- Analysing the data: Enhance reports to delineate insights, perform advanced analysis
- Deploying and Maintaining deliverables: Managing the datasets, create and manage workspaces

Core III: Data Analytics and Storytelling

Course Objectives

Data analytics is the process of analyzing raw data to find trends and answer questions given by any business organization. This course introduces students the concepts of data and visual analytics such as array processing, data wrangling and time series plotting using Python language and its libraries. It also helps students to create effective stories from data

Course Outcomes

On successful completion of this course, students will be able to:

CO1: Apply NumPy functions for array processing
CO2: Perform data wrangling and create various plots using Matplotlib and Seaborn
CO3: Perform data aggregation and group operations
CO4: Use Date and Time classes to create time series plots
CO5: Create and present effective stories from data

Unit-1. NumPy and Pandas Basics

Why Python for Data analysis – Essential Python libraries – ndarray – Universal functions – Data processing using arrays – File I/O with arrays – Random number generation – Series, Data Frames – Indexing, re-indexing, sorting, ranking –Summarizing descriptive statistics – Handling missing data – Hierarchical indexing

Unit-2. Data Loading, Wrangling Plotting and Visualization

Data Loading: reading and storing data in text format, binary format – Data Wrangling: Combining and merging data sets – Reshaping – Pivoting – Data transformation – String manipulation - Figures – Subplots – Colors – Ticks – Label – Legends – Annotation – Saving plots to file – Plots: Line, Bar, Histogram, Density Plots – Scatter Plots

Unit-3. Data Aggregation and Group Operations

Iterating over groups – Selecting columns – Grouping with Series and functions –Data aggregation: Column wise aggregation, returning aggregated data – General-Split-Apply-Combine – Quantile and bucket analysis – Pivot table and cross tabulation

Unit-4. Time Series Analysis

Date and Time – Time Series – Date Range, Frequencies and Shifting – Periods and period arithmetic – Resampling and frequency conversion – Time Series Plotting

Unit-5. Storytelling with Data

Anatomy of data story – Building blocking a story – Exploration to explanation - Storyboarding data story – Visualizing and polishing your data story

Text Books

1. Wes. Mc Kinney, "*Python for Data Analysis*", 3ed, Oreilly Publishers, 2022. ISBN 978-1098104030
2. Brent Dykes, Effective data storytelling, John Wiley and sons, 2020. ISBN9781119615712

References

1. CyrilleRossant. "Learning I*Python for interactive Computing and data visualization*", First edition [Packt]
2. Jake VanderPlas ,Python Data Science Handbook - Essential Tools for Working with Data, O'Reily, 2016
3. Zhang.Y ,An Introduction to Python and Computer Programming, Springer Publications,2016
4. Cole Nussbaumer, Storytelling with data, John Wiley and sons, 2020.

Core IV: Applied Machine Learning

Course Objectives

Machine learning is the kind of programming which gives computers the capability to automatically learn from data without being explicitly programmed. This course introduces students the various ML models for supervised and unsupervised ML problems. Various methods for evaluating and deploying the developed ML model are discussed.

Course Outcomes

On successful completion of this course, students will be able to:

CO1: Learn data gathering from various file formats
CO2: Apply pre-processing methods and reduce dimensions of data
CO3: Design a supervised ML model for the given business problem
CO4: Design a unsupervised ML model for the given business problem
CO5: Perform training, testing, evaluation and deployment of the designed ML model

Unit-1. Supervised Learning classifiers

Machine Learning types – ML workflow for predictive modeling – Types of ML systems - Regression - Gradient descent – Classification types: binary, multiclass and multi-label - Logistic regression – Support vector machines – Kernel SVM – Decision Trees – K-Nearest Neighbour classifier – Random Forest - Developing a web application with Flask

Unit-2. Pre-processing and Dimensionality Reduction

Pre-processing: Missing data, categorical data, feature scaling, feature selection. *Dimensionality reduction:* Principal Component Analysis, Linear Discriminant Analysis, Kernel PCA – Nonlinear Dim. reduction

Unit-3. Model evaluation and tuning

Regression metrics - Pipelines - K-fold cross validation - Grid search - Confusion matrix, Precision, Recall, ROC curves, Scoring metrics –Majority vote classifier – Bagging, Bootstrapping, Adaptive Boosting

Unit-4. Unsupervised Learning classifiers and neural networks

K-Means, K-Means++, Finding optimal no. of classifiers - Agglomerative Hierarchical clustering, Density based clustering - Perceptron – Adaline neural network - Multilayer Neural Network: Feed forward, Back Propagation Training, Multilayer Perceptron

Unit-5. Machine Learning Engineering

Machine Learning Project life cycle – Why ML projects fail – Imbalanced data handling – Data versioning - Feature selection using Boruta algorithm – Before selecting a model for training – Model evaluation – Model deployment – Model serving, monitoring and maintenance.

Text Books

1. Aurelien Geron, *Hands on Machine Learning with Scikit Learn, Keras and Tensorflow*, Oreilly, 3ed, 2023
2. Sebastian Raschka, Yuxi Liu and Vahid Mirjalili, *"Machine Learning with PyTorch and Scikit Learn"*, First Edition, PACKT, 2022.
3. Andriy Burkov, *Machine Learning Engineering*, 2022. Leanpub Publishers

References

1. Andreas C Muller and Sarah Guido, *Introduction to Machine Learning with Python*, Shroff Publishers, ISBN 978935213451
2. Joel Grus, *"Data Science from Scratch"*, First Edition, O'Reilly,2015
3. Gavin Hackeling, "*Mastering machine learning with scikit-learn"*, First Edition, [PACKT] , 2014

Core Practical III: Data Analytics and Storytelling Lab

Course Objectives

Data analytics is the process of analyzing raw data to find trends and answer questions that every business organization gives. This lab course enables students to implement the concepts of data and visual analytics such as array processing, data wrangling and time series plotting using Python language and its libraries.

Course Outcomes

On successful completion of this course, students will be able to:

CO1: Apply NumPy functions for array processing
CO2: Create Series and Data Frames for data wrangling
CO3: Create various plots using Matplotlib
CO4: Perform data aggregation and group operations
CO5: Use Date and Time classes to create time series plots

Lab Exercises

Develop applications that will demonstrate data analysis features of the following technologies.

- Data analytics using NumPy
- Pandas data wrangling features
- Data cleaning
- Data visualization
- Pandas Time series Analysis
- Advanced data wrangling features

Core Practical IV: Applied Machine Learning Lab

Course Objectives

Machine learning is the kind of programming which gives computers the capability to automatically learn from data without being explicitly programmed. This lab course enables students to practice various ML models for supervised and unsupervised Machine Learning problems and develop solutions for business, science and society applications.

Course Outcomes

On successful completion of this course, students will be able to:

CO1: Collect data from business customers
CO2: Apply pre-processing methods and reduce dimensions of data
CO3: Design a ML model for the given business problem
CO4: Perform training, testing and evaluation of the designed ML model
CO5: Provide recommendations to customers

Lab Exercises

Develop applications that will demonstrate the following Machine Learning models for the given business use case.

- Pre-processing methods
- Dimensionality reduction and feature selection using PCA, Kernel PCA, SVD and Entropy, Information Gain and Gini Index.
- Performance evaluation methods
  - Accuracy
  - MSE and SSE
  - Confusion Matrix
  - Precision, Recall and F-Score
- Supervised Machine Learning Models
  - Logistic Regression
  - Support Vector Machine
  - Decision Trees
  - Random Forests
  - Boosting
  - K-Nearest Neighbour
  - Perceptron
- Unsupervised Machine Learning Models
  - KMeans and KMeans++
  - Agglomerative Clustering
  - DBSCAN
- Evaluation of ML Models
  - K-Fold Cross Validation
  - Grid Search
  - Pipeline

Elective-III: Natural Language Processing

Course Objectives

Natural Language Processing is one important technology that has penetrated deeply and widely in the market, irrespective of the industry and domains. This course provides an introduction to the field of NLP, synthesizing research from linguistics and computer science.This coursecovers formal models for representing and analyzing syntax and semantics of words, sentences, and documents.

Course Outcomes

On successful completion of this course, students will be able to
CO1: perform Text Classification
CO2: create language models and POS tagging
CO3: create grammars and parse sentences
CO4: understand meaning of sentences
CO5: design NLP applications

Unit-I Text Pre-processing, Classification and Evaluation

Text tokenization, normalization, Lemmatization and Stemming – Minimum Edit Distance between two strings – Bigram, Trigram and NGrams – Perplexity – Smoothing Methods – Naïve Bayes Classifier – Binary Naïve Bayes for Sentiment Analysis and Spam Detection – Precision, Recall, F1-score, Cross Validation - Bootstrapping.

Unit-II Language Modeling

Term Document Matrix and Word Word Matrix – Cosine Similarity – TFIDF weighting – Positive PMI – Skip Gram with Negative Sampling – Neural Network Unit –Feed Forward Neural Network – Neural Network for Language Model - English word classes for POS Tagging – Penn Treebank POS Tagset – POS Tagging Process – HMM POS Tagger – Viterbi Decoding algorithm for HMM.

Unit-III Context Free Grammars

Context Free Grammars: Definition – Grammar rules for English – Tree Banks as Grammars – Grammar equivalence and normal forms – Lexical Grammars – Syntactic Parsing: Types of Ambiguities - CKY Parsing Algorithm – Chunking – Probabilistic CFG for Disambiguation and Language Modeling – Probabilistic CKY Parsing Algorithm for PCFGs – Probabilistic Lexicalized CFGs: Collins Parser – Parser Evaluation Methods.

Unit-IV Dependency Parsing

Dependency Parsing: Dependency Relations from Universal Dependency Set – Dependency Tree – Shift Reduce Parsing – Arc Eager Transition Based Parsing – Edge Factored Parsing model using MST – First Order Logic for meaning representation: elements, variables, quantifiers, lambda notation, inference – Event Representation – Event Time Representation – Representation and Inference using Description Logic

Unit-V NLP Applications

NLP Applications: Named Entity Recognition – Word Sense Disambiguation - Factiod Question Answering – Rule based and corpus based chatbots – Dialogue State Architecture for chatbots.

Text Books

1. Jurafsky and Martin, Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics and Speech Recognition, 3ed. 2020.

References

1. Steven Bird, Ewan Klein, and Edward Loper , Natural Language Processing with Python– Analyzing Text with the Natural Language Toolkit.
2. Indurkhya, Nitin and Fred Damerau, Handbook of Natural Language Processing, 2ed, 2010, Chapman & Hall/CRC.
3. Christopher Manning and HinrichSchutze, Foundations of Statistical Natural Language Processing, MIT Press.

Elective-III : Multivariate Analysis

Course Objectives
This course introduces the concepts of Multivariate data analysis to understand the research data, its presentation and analysis. Topics such as multivariate data structure, multinomial and multivariate normal distribution, estimation andtesting of parameters and data reduction methods, are covered in this course.

Course Outcomes
On successful completion of this course, students will be able to:
CO1: Understand multivariate data structure, multinomial and multivariate normaldistribution
CO2: Apply Multivariate analysis of variance (MANOVA) of one and two- way classifieddata.

Unit-1. Introduction
Basic concepts on multivariate variable - Multivariate normal distribution, Marginal andconditional distribution - Concept of random vector: Its expectation and Variance-Covariancematrix - Marginal and joint distributions - Conditional distributions and Independence ofrandom vectors - Multinomial distribution

Unit-2. Distribution
Sample mean vector and its distribution - Likelihood ratio tests: Tests of hypotheses about themean vectors and covariance matrices for multivariate normal populations - Independence ofsub vectors and sphericity test.

Unit-3. Multivariate Analysis
Multivariate analysis of variance (MANOVA) of one and two- way classified data - Multivariate analysis of covariance. Wishart distribution - Hotelling's T2 and Mahalanobis'D2 statistics - Null distribution of Hotelling's T2 - Rao's U statistics and its distribution

Unit-4. Classification and Discriminant Procedures
Bayes, minimax, and Fisher's criteria for discrimination between two multivariate normalpopulations - Sample discriminant function - Tests associated with discriminant functions - Probabilities of misclassification and their estimation - Discrimination for several multivariatenormal populations

Unit-5. Principal Component and Factor Analysis
Principal components, sample principal components asymptotic properties –Canonicalvariables and canonical correlations: definition, estimation, computations - Test forsignificance of canonical correlations - Factor analysis: Orthogonal factor model, factor loadings, estimation of factor loadings,factor scores - Applications

Text Books
1. Anderson, T.W. An Introduction to Multivariate Statistical Analysis, 3ed,John Wiley. 2009
2. Everitt B, Hothorn T. An Introduction to Applied Multivariate Analysis with R,Springer. 2011
3. Barry J. Babin, Hair, Rolph E Anderson, and William C. Blac, Multivariate DataAnalysis, Pearson New International Edition, 2013

References
1. Giri, N.C. Multivariate Statistical Inference. Academic Press. 1977.
2. Chatfield, C. and Collins, A.J. Introduction to Multivariate analysis. Prentice Hall.1982.
3. Srivastava, M.S. and Khatri, C.G.An Introduction to Multivariate Statistics. NorthHolland. 1979.

Elective IV: Time Series Analysis and Forecasting

Course Objectives
This course covers not only common statistical approaches and time series models, including ARMA, SARIMA, VAR, GARCH and state space and Markov switching models for (non)stationary, multivariate and financial time series, but also modern machine learning procedures and challenges for time series forecasting to solve real world data science problems.

Course Outcomes
On successful completion of this course, students will be able to:

CO1: Learn basic models for time series data
CO2: Learn auto-regressive and model averaging models
CO3: Learn non stationary and financial models
CO4: Learn multivariate time series models
CO5: Learn state space and markov models

Unit I: Basic Concepts of Time Series
Time Series Concepts: The Concept of Time Series - Time Series Moment Functions and Stationarity -Time Series Data Visualization. *Exploratory Time Series Data Analysis:* Partial Autocorrelation Functions - White Noise Test - Simple Time Series Compositions - Time Series Decomposition and Smoothing.

Unit II: Stationary Time Series Models
Stationary Time Series Models: Backshift Operator, Differencing, and Stationarity Test - Moving Average Models - Autoregressive Models - Autoregressive Moving Average Models. *ARMA and ARIMA Modeling and Forecasting***:** Model Building Problems - Estimation Methods - Order Determination - Diagnosis of Models - Forecasting.

Unit III: Non-stationary Time Series Models
Nonstationary Time Series Models: The Box-Jenkins Method - SARIMA Model Building- REGARMA Models. *Financial Time Series and Related Models:* Stylized Facts of Financial Time Series - GARCH Models: ARCH Models and GARCH Models- Other Extensions: EGARCH Models - TGARCH Models.

Unit IV: Multivariate Time Series Analysis
Multivariate Time Series Analysis : Basic Concepts - Covariance and Correlation Matrix Functions - Stationarity and Vector White Noise - Sample Covariance and Correlation Matrices - Multivariate Portmanteau Test - VARMA Models - Definitions - Properties - VAR Model Building and Analysis - VAR(1) Representation of VARMA Processes - VAR Model Building Steps - Granger Causality - Impulse Response Analysis

Unit V: State Space Models
State Space Models and Markov Switching Models: State Space Models and Representations - State Space Models - State Space Representations of Time Series - Kalman Recursions - Local-Level Model and SARIMAX Models - Local-Level Model - SARIMAX Models - Markov Switching Models.

Text Books
1. Changquan Huang, Alla Petukhina. Applied Time Series Analysis and Forecasting with Python. Springer, 2022.

References
1. Aileen Nielsen, “Practical Time Series Analysis_ Prediction with Statistics and Machine Learning”, O’Reilly Media, 2019
2. Terence C. Mills, “Applied Time Series Analysis_ A Practical Guide to Modeling and Forecasting”, Academic Press, 2019.
3. Shumway and Stoffer. Time Series Analysis and its applications, with examples in R. 4ed, Springer. 2016.
4. Prado & West. Time Series: Modeling, Computation, and Inference Chapman & Hall. 2010

Elective-IV : Health Care Data Analytics

Course Objectives

Healthcare data analytics is a methodology that processes real-time and historical healthcaredata to predict trends, get actionable insights, and drive long-term growth. It allows healthcare professionals to find opportunities for improvement in health system management, patient engagement, spending, and diagnosis. This course will introduce students the various forms of electronic health care information. It helps students to learn the techniques to analyse health care data and to create predictive models for clinical data.

Course Outcomes

On successful completion of this course, students will be able to:

CO1: Understand Electronic Health Records
CO2: Analyse health care data using analytical techniques.
CO3: Apply analytics for decision making in healthcare services.
CO4: Integrate health data from multiple sources
CO5: Developefficient clinical decision support systems.

Unit-1. Introduction

Introduction to Healthcare Data Analytics- Electronic Health Records–Components of EHR- Coding Systems- Benefits of EHR- Barrier to Adopting HER Challenges-Phenotyping Algorithms.

Unit-2. Analysis

Biomedical Image Analysis- Mining of Sensor Data in Healthcare- BiomedicalSignal Analysis- Genomic Data Analysis for Personalized Medicine.

Unit-3. Analytics

Natural Language Processing and Data Mining for Clinical Text- Mining theBiomedical- Social Media Analytics for Healthcare.

Unit-4. Advanced Data Analytics

Advanced Data Analytics for Healthcare– Review of ClinicalPrediction Models- Temporal Data Mining for Healthcare Data- Visual Analytics forHealthcare- Predictive Models for Integrating Clinical and Genomic Data- InformationRetrieval for Healthcare- Privacy-Preserving Data Publishing Methods in Healthcare.

Unit-5. Applications

Applications and Practical Systems for Healthcare– Data Analytics forPervasive Health- Fraud Detection in Healthcare- Data Analytics for PharmaceuticalDiscoveries- Clinical Decision Support Systems- Computer-Assisted Medical ImageAnalysis Systems- Mobile Imaging and Analytics for Biomedical Data.

References

1. Chandan K. Reddy and Charu C Aggarwal, “Healthcare data analytics”, Taylor &Francis, 2015
2. Hui Yang and Eva K. Lee, “Healthcare Analytics: From Data to Knowledge to Healthcare Improvement, Wiley, 2016.

SEC II: Natural Language Processing Lab

Course Objectives

Natural Language Processing is one important technology that has penetrated deeply and widely in the market, irrespective of the industry and domains. It is extensively applied in every business organization. This lab course enables students to represent and analyze syntax and semantics of words, sentences, and documents.

Course Outcomes

On successful completion of this course, students will be able to:

CO1: Design NLP pipeline
CO2: Categorize and tag words
CO3: Classify text using supervised classifiers
CO4: Extract entities and named entities
CO5: Analyze sentences and their meanings

Lab Exercises

Develop NLP applications that will demonstrate the following NLP concepts and features.

- Frequency Distribution using Histograms
- NLP pre-processing such as tokenization, stemming, lemmatization
- Document similarity using VSM
- POS Tagging
- Bigrams and nGram Tagging
- Text Classification
- Named Entity Recognition
- Context Free Grammars
- Dependency parsing
- Word sense disambiguation

Core V: Principles of Deep Learning

Course Objectives

Several Artificial Intelligence tasks demand us to build intelligence applications that are capable of extracting meaningful representations from high-dimensional data. In order to solve these AI tasks, deep learning has emerged as an important subdomain of AI. This course will introduce students to the recent and exciting developments of various deep learning methods and applications.

Course Outcomes

On successful completion of this course, students will be able to
CO1: Understand and apply the foundational concepts of Deep Learning
CO2: Solve real world applications using MLP and improve its performance
CO3: Develop CNN models and optimize the performance for CNN applications
CO4: Develop RNN models and optimize the performance for RNN applications
CO5: Develop Autoencoders and GAN models for real time applications

Unit-I: Deep Learning Foundation

Linear algebra: scalar, vector, matrix, tensor – Multiplying matrix and vectors – L1 and L2 norms – Eigen decomposition – Singular Valued Decomposition – Determinant. Linear Regression: Model, Cost function, Loss function - Model improvement: overfitting vs underfitting; bias vs variance; hyper parameter tuning: random, coarse to fine - Logistic regression.

Unit-II: MLP, Regularization and Optimization

Multilayer perceptron: Layers: Output units: linear, sigmoid, softmax; Hidden units: ReLU, Sigmoid, tangent, RBF – Forward propagation – MLP training with back propagation. Regularization: L1, L2 regularization, Data augmentation, Early stopping, Drop out, Data normalization. Optimization for Training Deep Models: Challenges: Local minima, plateau, saddle points, vanishing and exploding gradients – Algorithms: Minibatch, Stochastic Gradient Descent, RMSProp, Adam.

Unit-III: Convolutional Neural Networks

CNN: Convolution, pooling, striding, padding, 1x1 convolution – Popular CNN models: AlexNet, VGG, ResNet - CNN Applications: Transfer Learning, Image classification, face detection, object detection, face/instance recognition.

Unit-IV: Recurrent Neural Networks

Time series analysis - RNN with no outputs – RNN with outputs - Forward propagation – Backpropagation Through Time (BPTT) algorithm – Bidirectional RNN – Sequence to sequence RNN - Word Embedding - LSTM, GRU. RNN Application: Sentiment analysis, text generation, machine translation, attention model, speech recognition, video classification.

Unit-V: Autoencoders and Generative Models

Auto-encoder, DenoisingAutoencoder; AE Applications: Data compression, retrieval, classification, document clustering, sentiment analysis. Generative Learning: Variational Auto-encoders, Generative Adversarial Neural Networks; GL Applications: animal face/celebrity face generation, video generation.

Textbooks

1. Ian Goodfellow, YoshuaBengio and Aaron Courville, Deep Learning, MIT Press, 2016, ISBN-13: 978-0262035613. http://www.deeplearningbook.org, 2016.
2. Charu C. Aggarwal, Neural Networks and Deep Learning, 1ed, Springer International Publishing AG, part of Springer Nature, 2018, ISBN-13: 978-3319944623.

References

1. Francois Chollet, Deep Learning with Python, 2ed, Manning Publications, 2021, ISBN-13: 978-1617296864.
2. AurélienGéron, Hands-on Machine Learning with Scikit-Learn and TensorFlow, 1ed, O'Reilly Media, 2017, ISBN-13: 978-1491962299.

Core VI: Big Data Management and Analytics

Course Objectives

Big data analytics analyses large volumes of data to find undiscovered patterns, correlations, and other insights. This course will cover concepts, algorithms and standard tools used to analyze Big Data such as MapReduce over modern distributed analysis platforms such as Hadoop, Pig and Spark and it also covers data summarization framework HIVE, sparse data storing tool HBase, configuration and state management Apache project ZooKeeper.

Course Outcomes

On successful completion of this course, students will be able to:
CO1: Store and manipulate data using HDFS
CO2: Develop applications using Hadoop and analyse HDFS stored data
CO3: Explore data using Pig and perform Data Warehousing using Hive
CO4: Develop ML applications using Spark
CO5: Perform sparse data handling by HBase and state management using zookeeper

Unit I. Introduction to Big Data and HDFS

What is Big Data? - Overview of Big Data Industrial examples - Big Data Technology (Hadoop, Cloud, BI, Crowdsourcing Analytics, Business Analytics) – **HDFS:** Block - Namenode & Datanode - Caching – File system operations in command line - Java Interface to Basic Hadoop (Reading data and writing data, Anatomy of File Read and Write).

Unit II. MapReduce

MapReduce model (Weather dataset, Analyzing data with Hadoop, Combiner functions, Hadoop streaming with Python) - Steps of developing MapReduce application - Working of MapReduce (Running Jobs, failure, Shuffle and sort, Task execution) - MapReduce Types (Input formats, Output formats) - MapReduce features (Counters, Sorting, Joins).

Unit III. Pig and Hive

PIG: Exploring large datasets using Pig – Structure – Statements – Expressions – Types – Schemas – Functions - Macros – Data Processing Operators (Loading and Storing Data, Filtering Data, Grouping and Joining Data, Sorting Data, Combining and Splitting Data). **HIVE:** Data Warehousing using Hive - Comparison with Traditional Databases HiveQL (Data Types, Operators and Functions) – Tables (Managed Tables and External Tables, Partitions and Buckets, Storage Formats, Importing Data, Altering Tables, Dropping Tables) - Querying Data(Sorting and Aggregating, MapReduce Scripts, Joins, Subqueries, Views).

Unit IV. Spark

Spark Applications, Jobs, Stages, and Tasks - A Scala Standalone Application - A Java Example - A Python Example - Resilient Distributed Datasets(RDD)(Creation, Transformations and Actions, Persistence, Serialization) - Shared Variables - Anatomy of a Spark Job Run (Job Submission, DAG Construction, Task Scheduling & Task Execution) - Executors and Cluster Managers (Spark on YARN). Case Study: Spark on DataBricks

Unit V. Hbase and ZooKeeper

HBase: Installation – Clients (Java, MapReduce) - Building an Online Query Application (Schema Design, Loading Data, Online Queries) - HBase vs RDBMS. **ZooKeeper:** Installing and Running - Group Membership in ZooKeeper - Creating the Group - Joining a Group - Listing Members in a Group - Deleting a Group - The ZooKeeper Service (Data Model, Operations, Implementation, Consistency, Sessions, States) - Building Applications with ZooKeeper (A Configuration Service, The Resilient ZooKeeper Application, A Lock Service, More Distributed Data Structures and Protocols).

Text Books

1. Tom White Hadoop: The Definitive Guide, Fourth Edition, O'reilly Media, 2015.

2. Jules S. Damji, Brooke Wenig, Tathagata Das & Denny Lee. Learning Spark Lightning-Fast Data Analytics, 2ed, O'Reilly, 2020 (Unit IV)
3. Michael Minelli, Michele Chambers and AmbigaDhiraj. Big Data, Big Analytics: Emerging Business Intelligence and Analytic Trends for Today's Businesses, 1ed, Wiley CIO Series, 2013. ISBN 9781118147603

References

1. Nathan Marz and James Warren, Big Data Principles and Practice of Scalable Real Time Data Systems, Manning Publications. 2015
2. Michael Berthold, David J. Hand, Intelligent Data Analysis, Springer, 2007
3. Bill Franks, Taming the Big Data Tidal Wave: Finding Opportunities in Huge Data Streams with Advanced Analytics, John Wiley& sons, 2012.
4. Glenn J. Myatt, Making Sense of Data, Volume I and II. John Wiley & Sons, 2007.
5. Mark Grover, Ted Malaska, Jonathan Seidman, Gwen Shapira. Hadoop Application Architecture, Shroff Publishers.2015
6. Chris Eaton, Dirk DeRoos, Tom Deutsch, George Lapis, Paul Zikopoulos, Understanding Big Data: Analytics for Enterprise Class Hadoop and Streaming Data, McGrawHill Publishing, 2012.

Core Practical V: Principles of Deep Learning Lab

Course Objectives

Deep learning is a subset of machine learning where artificial neural networks learn from large amounts of data.This lab course will introduce various Deep Learning methods and techniques to solve real world and business problems related to classification, language modelling, machine translation and dialogue systems.

Course Outcomes

On successful completion of this course, students will be able to:

CO1: Practice and apply Deep Learning libraries
CO2: Develop Deep Neural Networks for classification
CO3: Develop Deep Neural Networks for text analysis
CO4: Develop machine translation networks
CO5: Develop dialogue systems for business applications

Lab Exercises

Develop applications that will demonstrate the following applications using Deep Learning features.

- Image classification
- Object detection
- Object classification
- Text classification
- Sentiment analysis
- Text summarization
- Text entailment
- Word and Sentence Embedding
- Question Answering
- Machine Translation
- Dialogue Systems

Core Practical VI: Big Data Management and Analytics Lab

Course Objectives

This lab course will enable students to analyse large amount of datausing modern distributed analysis platforms such as HDFS, Spark, MapReduce, Pig and Hive technologies

Course Outcomes

On successful completion of this course, students will be able to:

CO1: Develop applications using Hadoop
CO2: Store and manipulate data using HDFS
CO3: Explore very large datasets using Pig
CO4: Perform Data Warehousing operations using Hive
CO5: Perform data analytics using Spark

Lab Exercises

Develop applications for the following tasks

- Installation and setup of Hadoop
- File management tasks in Hadoop
- Benchmarking and stress testing on Hadoop cluster
- Map Reduce applications for Word Counting
- Stop word elimination using Map Reduce
- Weather data analytics using Map Reduce
- Perform data analytics using Spark
- Perform sort, group, join, project, and filter operations on Pig
- Design vector space model for text collection using Pig
- Create, alter, and drop databases, tables, views, functions, and indexes on Hive

Elective-V: Computer Vision

Course Objectives
Computer vision aims to compute properties of the 3D world from 2D digital images. There are several applications for computer vision such as creation of virtual characters, autonomous driving, medical imaging, virtual and augmented reality and others. In this course, students will learn computer vision algorithms in order to solve problems including object tracking, object recognition, human recognition and their activities by analysing images and videos.

Course Outcomes
On successful completion of this course, students will be able to:

CO1: Learn image coordinates, transformations and image processing concepts
CO2: Apply feature and edge detection methods
CO3: Apply image and video understanding methods on humans
CO4: Apply image alignment and stitching methods on 2D and 3D images
CO5: Understand and apply motion and pose estimation methods for images
CO6: Understand the applications of computer vision

Unit-I
What is Computer Vision? - Geometric primitives and transformations – Sampling and aliasing – Color spaces. **Image processing:** Point operators - Linear filtering – Nonlinear filtering – Bilateral filtering - Interpolation - Geometric transformations

Unit-II
Feature detection and matching: Feature detectors, descriptors, matching and tracking – Edge detection - Contour detection - Contour tracking - Image segmentation using Normalized cut

Unit-III
Image recognition: Bag of words feature based image classification - Face recognition - Face detection - Pedestrian detection - Instance segmentation - Panoptic segmentation - Pose estimation - Video understanding using neural networks - Image captioning - Text to image generation

Unit-IV
Image alignment and stitching: 2D image alignment – RANSAC - 3D image alignment. *Image stitching*: Parametric motion models - rotational panoramas

Unit-V
Motion and pose estimation: Hierarchical motion estimation - Parametric motion models - Spline based motion - Optical flow: Two frame motion estimation, Multi frame motion estimation. Video object segmentation - Video object tracking - **Pose estimation**: Linear algorithms, iterative algorithms

Text book
1. Richard Szeliski. Computer Vision: Algorithms and Applications. Springer, 2021. ISBN 978-3030343712

Reference books
1. Bishop, C. M. Pattern Recognition and Machine Learning. Springer, NY. 2006
2. Zhang, A., Lipton, Z. C., Li, M., and Smola, A. J. (2019). Dive into deep learning. Corwin Publishers. 2019
3. Brown, M. S. ICCV 2019 tutorial on understanding color and in-camera image processing pipeline for computer vision. 2019. https:// www.eecs.yorku.ca/~mbrown/ICCV2019_Brown.html

Elective-V : Supply Chain Management

Course Objectives

This course teaches students supply chain strategy and concepts. It provides solid understanding on building framework to analyze supply chain, designing supply chain network, managing demand and supply and inventories, besides designing transportation networks.

Course Outcomes

On successful completion of this course, students will be able to:
CO1: Understand key elements of supply chain
CO2: Design supply chain network
CO3: Understand demand forecasting and manage supply and demand
CO4: Understand planning and managing inventories and mode of transportation
CO5: Provide recommendations for the given business problems

Unit-1. Building strategic framework

Supply chain: Definition, 3 streams of knowledge, objectives and importance - Decision phases and process views of a supply chain (SC) - Examples of supply chain - Competitive strategy and SC strategy - 3 steps of achieving strategic fit - Improving SC performance by expanding scope of strategic fit, challenges to achieving strategic fit - Financial measures and drivers of SC performance - Logistical drivers: Roles in SC and decision components - Cross functional drivers: Roles in SC and decision components - Role of infrastructure in SC performance.

Unit-2. Designing SC network

Key factors influencing distribution network design - Design options for a distribution network - Impact of online sales on customer service and cost - Network design decisions: Influencing factors, framework - Capacitated plant location model for network optimization - Gravity location model for network design - Model for demand allocation and locating plants - Global supply chain: Dimensions to evaluate total cost, SC risks, tailored risk mitigation strategies - Discounted cash flow analysis to evaluate network design decision - Decision tree analysis: Basics, Evaluating flexibility at Trip Logistics.

Unit-3. Planning and coordinating demand and supply

Demand forecasting: role, characteristics, components and methods - Static demand forecasting methods - Adaptive demand forecasting methods - Measures of demand forecasting error - Aggregate planning: role, identifying aggregate units, strategies - Aggregate planning using Linear programming - Managing supply and demand to improve synchronization in SC - Lack of SC coordination: Bullwhip effect, effect on performance - Obstacles to coordination in SC - Managerial levers to achieve coordination of demand and supply in SC.

Unit-4. Planning and managing inventories

Cycle inventory terminologies: Lot size, Average flow time, Inventory holding cost, Ordering cost - Computing optimal lot size for single product: Economic order quantity, for Production environment, with Capacity constraint - Lot size based discount schemes: All unit quantity discounts, Marginal unit quantity discount - Trade promotions: Goals, Forward buying, Impact on lot size and cycle inventory - Factors affecting the level of safety inventory - Evaluating required safety inventory: Given a replenishment policy, Desired cycle service level, Desired fill rate - Impact of desired product availability and uncertainty on safety inventory - Impact of supply uncertainty on safety inventory - Factors affecting optimal level of product availability - Managerial levers of inventory to improve SC profitability.

Unit-5. Transportation and cross functional drivers

Modes of transportation in SC - Design options for a transportation network - Transportation and inventory cost trade off - Transportation cost and customer responsiveness trade off - Tailored transportation - Sourcing decisions: In house or Outsource - Sharing risk and reward in SC - Pricing and revenue management for multiple customer segments - Pricing and revenue management for perishable assets - Pricing and revenue management for seasonal demand.

Text Books

1. Sunil Chopra, Peter Meindl and DV Karla. “Supply Chain Management: Strategy, planning and operation”, 6th edition, Pearson, 2016. ISBN 978-9332548237 (Excluding Excel Examples)

References

1. David Simchi-Levi and Philip Kaminsky. “Designing and managing the supply chain: Concepts, strategies and case studies”, 3rd edition, McGraw Hill, 2007.

SEC III-Data Science Project Management Lab

Course Objectives

This lab course will enable students to apply MLOps principles, tools and techniques to develop, deploy and continuously monitor Machine Learning and Deep Learning applications in production environments.

Course Outcomes

On successful completion of this course, students will be able to

CO1: Create or select a dataset that you're interested
CO2: Train a model on that dataset tracking your experiments
CO3: Create a model training pipeline
CO4: Deploy the model in batch, web service or streaming
CO5: Monitor the performance of your model and follow the best practices

Develop applications that will demonstrate the following features

- Cloud: AWS, GCP, Azure or others
- Experiment tracking tools: MLFlow, Weights & Biases
- Workflow orchestration: Prefect, Airflow, Flyte, Kubeflow, Argo
- Monitoring: Evidently, WhyLabs/whylogs
- CI/CD: Github actions, Gitlab CI/CD
- Infrastructure as code (IaC): Terraform, Pulumi, Cloud Formation

Core VII: Digital Innovation Management

Course Objectives

The course will cover the basic concepts of innovation, basic theories and representative practical cases. The course will help students to break through the existing thinking patterns, improve their ability to think creatively, understand the main problems in the innovation process of products, services, organizations and industries, and master the general principles and laws of promoting and managing innovation activities.

Course Outcomes

On successful completion of this course, students will be able to

CO1: Describe different approaches explain digital innovation processes
CO2: Develop innovation strategy and create innovative organization
CO3: Identify innovation networks and create new digital products and services
CO4: Promote open innovation and entrepreneurship
CO5: Describe business and social value of innovation

Unit-1. Basics of Innovation

Innovation - What It Is and Why It Matters - Digital Is Different? - Innovation as a Core Business Process

Unit-2. Strategy Development

Developing an Innovation Strategy - Building the Innovative Organization - Sources of Innovation

Unit-3. Construct innovative products and services

Search Strategies for Innovation - Innovation Networks - Dealing with Uncertainty - Creating New Products and Services

Unit-4. Exploit open innovation and entrepreneurship

Exploiting Open Innovation and Collaboration - Promoting Entrepreneurship and New Ventures

Unit-5. Capturing values from innovation

Capturing the Business Value of Innovation - Creating Social Value - Capturing Learning from Innovation

Text Book

1. Joe Tidd, John Bessant. Managing innovation: integrating technological, market and organizational change, John Wiley, 7ed, 2021

References

1. John M Jordan. Information, technology, and innovation: resources for growth in a connected world, John Wiley, 2012
2. Bernard Burnes. Managing change, Pearson, 7ed, 2017

Core Practical VII: Web Application Development Lab

Course Objectives

Javascript is the standard programming language of the web and the web has been the most used computing platform. Javascript powers billions of websites, mobile devices and IoT devises in the world. This lab course gives students the power to create dynamic content for both frontend and backend web applications.

Course Outcomes

On successful completion of this course, students will be able to

CO1: Develop problems using the function programming features
CO2: Apply OOP features to develop advanced Javascript code
CO3: Demonstrate and apply the Document Object Model (DOM) and asynchronous programs
CO4: Develop interactive websites using HTML, Javascript and AJAX
CO5: Create NoSQL databases and perform CRUD operations using Mongoose and MongoDB

Develop applications that will demonstrate the following features using Javascript language

- Functional features such as arrays and functions
- Object Oriented features
- Regular expressions
- Static webpages using HTML, CSS and Javascript
- DOM tree traversal and item selection
- HTML forms and event handling
- Dynamic webpages using AJAX and event handling
- Database design using Mongoose
- Database design using MongoDB

Core Project-Project with Viva voce

Course Objectives

Students will do a masters project based on application scenario or research based problems. For all application based projects, they will follow software engineering methodologies. Students will follow a two phase project development. In the first phase, students will develop Software Requirements Specifications (SRS) document. The SRS will contain a problem statement, scope & justification, requirements specification, cost estimation, limitations, methodology identified and tools and programming languages to be used in the development of the project. In the second phase, students will implement the proposed software design outlined in SRS. Students will write code, debug source code, test them and deploy at the client site. As part of the deployment, they will develop various documentations such as user manuals.

For research projects, students will define research questions, perform literature survey, propose methodology and validate their research with various experimental results. This way, they are expected to advance the current state of the art in their chosen area of research.

Course Outcomes

On successful completion of this course, students will be able to:

CO1: Identify and define the problem statement
CO2: Define and justify scope of the proposed problem
CO3: Gather and analyzeapplication requirements
CO4: Propose an optimized solution among the existing solutions
CO5: Develop a functional application based on the software design
CO6: Apply coding, debugging and testing tools and construct a new software system
CO7: Prepare documentation and user manuals by following the standard guidelines
CO8: Learn technical report writing and oral presentation skills

Text Books

1. Lynn E. Miner & Jeremy T. Miner, “Proposal Planning and Writing”, Third Edition, Greenwood Publishing Group, 2003.

References

1. William Navidi, “Statistics for Engineers and Scientists”, 2nd Edition, McGraw-Hill, 2007.

Elective VI: Social Media and Graph Analytics

Course Objectives

Social media analytics is the process of representing, analyzing, and extracting actionable patterns from social media data. This course will introduce students the necessary theories and the state-of-the art techniques in Web mining, networks analysis, and predictive modeling to study emerging problems with social media. These problems include information diffusion, information retrieval, recommendations, behaviour analysis and event analytics in social media.

Course Outcomes

On successful completion of this course, students will be able to:

CO1: Explain the essentials of graphs for social networks
CO2: Measure social network nodes and simulate social network models
CO3: Evaluate the community detection and link prediction methods
CO4: Detect cascade behaviour and anomalies in social networks
CO5: Apply and learn graph representations from OSNs

Unit-1. Introduction and Network measures

Fundamentals of social networks-Applications in different domain-Types of Social Networks- Social Networks as Graphs- Network Basics-Graph Representation- Graph Connectivity- Centrality Measures – Assortativity-Transitivity and Reciprocity, Similarity- Degeneracy - Graph Visualization Tools

Unit-2. Social Network Growth Models and Link Analysis

Network Models: Properties – Random Network Models – Ring Lattice – Watts Strogatz – Preferential Attachment Model- Price model – Local world NNetwork growth model - Applications of Link Analysis-Signed Networks-String and Weak Ties-Link Analysis and Algorithms- Page Rank – Div rank – Sim rank – PathSim algorithms

Unit-3. Community Detection and Link Prediction

Types of Communities and their Applications- Disjoint, overlapping and local community detection - Evaluation of Community Detection Methods- Temporal Changes in Networks- Evaluation of Link Prediction-Heuristics and Probabilistic Models- Supervised random Walk- Information-theoretic Model – Latest trends

Unit-4. Cascade Behaviour and Anomaly Detection

Cascade Model- Probabilistic Cascades- Epidemic Models- Independent Cascade Models - Cascade Prediction-Outliers vs Network based Networks- Anomaly Detection in Static and Dynamic Networks

Unit-5. Graph Representation Learning and Applications

Machine Learning Pipeline-Intuition and benefits of Representation Learning – GRL Pipeline- Representation Learning Methods: Matrix Factorization, Graph embedding, Random walk, Graph convolution network, GraphSAGE - Malicious Activities on OSN's- Sock puppets and collision - Modelling the Spread of COVID19-Recommender Systems

Text Books

1. Tanmoy Chakraborty, Social Network Analysis, Wiley India, 2021, ISBN: 9789354247835

References

1. Matthew A. Russell. Mining the Social Web. 3rd Edition. O'Reilly Media. 2019
2. Reza Zafarani, Mohammad Ali Abbasi, and Huan Liu. Social Media Mining: An Introduction, Cambridge University Press, 2014
3. Jennifer Golbeck. Analyzing the Social Web. Morgn Kaufmann. 2013. ISBN 978-0124055315
4. Ricardo Baeza-Yates and BerthierRibeiro-Neto. Modern Information Retrieval: The Concepts and Technology behind Search. 2ed. ACM Press Books, 2011. ISBN 978-0321416919
5. Charu C. Aggarwal. Social Network Data Analytics. Springer. 2011

Elective VI: Customer Relationship Management

Course Objectives

This course will introduce students the concepts of customer relationship management with industry case studies and strategies for implementing them in any organization. It helps students to better understand customer needs and to maintain long-term customer relationships.

Course Outcomes

On successful completion of this course, students will be able to:
CO1: Understand concept of Customer Relationship Management (CRM) and implementation of CRM
CO2: Provide insight into CRM marketing initiatives and customer service
CO3: Design CRM strategy
CO4: Understand CRM implementation strategies
CO5: Understand new trends in CRM, challenges and opportunities for organizations

Unit-1. Introduction

Introduction to Customer Relationship Management: Concept- Evolution of Customer Relationships: Customers as strangers- acquaintances- friends and partners. Objectives- Benefits of CRM to Customers and Organizations- Customer Profitability Segments- Components of CRM: Information- Process- Technology and People-Barriers to CRM. Relationship Marketing and CRM: Relationship Development Strategies: Organizational Pervasive Approach- Managing Customer Emotions- Brand Building through Relationship Marketing- Service Level Agreements- Relationship Challenges.

Unit-2. Marketing and Data Management

CRM Marketing Initiatives- Customer Service and Data Management: CRM Marketing Initiatives: Cross-Selling and Up-Selling- Customer Retention- Behaviour Prediction-Customer Profitability and Value Modeling-Channel Optimization- Personalization and Event-Based Marketing. CRM and Customer Service: Call Center and Customer Care: Call Routing- Contact Center Sales-Support- Web Based Self Service- Customer Satisfaction Measurement- Call-Scripting- Cyber Agents and Workforce Management. CRM and Data Management: Types of Data: Reference Data- Transactional Data-Warehouse Data and Business View Data-Identifying Data Quality Issues- Planning and Getting Information Quality- Using Tools to Manage Data- Types of Data Analysis: Online Analytical Processing (OLAP) - Clickstream Analysis- Personalization and Collaborative Filtering- Data Reporting.

Unit-3. Strategy

CRM Strategy- Planning: Understanding Customers: Customer Value- Customer Care-Company Profit Chain: Satisfaction- Loyalty- Retention and Profits. Objectives of CRMStrategy- The CRM Strategy Cycle: Acquisition-Retention and Win Back- Complexities of CRM Strategy.

Unit-4. Implementation and Evaluation

CRM Implementation and Evaluation: Planning and Implementation of CRM: Business to Business CRM- Sales and CRM- Sales Force Automation- Sales Process/ Activity Management- Sales Territory Management-Contact Management- Lead Management-Configuration Support- Knowledge Management CRM Implementation: Steps- Business Planning- Architecture and Design- Technology Selection- Development-Delivery and Measurement.

Unit-5. Evaluation

CRM Evaluation: Basic Measures: Service Quality- Customer Satisfaction and Loyalty-Company 3E Measures: Efficiency- Effectiveness and Employee Change. CRM New Horizons: e-CRM: Concept- Different Levels of E-CRM- Privacy in E-CRM -Software App for Customer Service:# Activity Management- Agent Management-Case Assignment- Contract Management- Customer Self Service- Email Response Management- Escalation-Inbound Communication Management- Invoicing- Outbound Communication Management- Queuing and Routing- Scheduling - Social Networking and CRM - Mobile-CRM - CRM Trends- Challenges and Opportunities - Ethical Issues in CRM.

References

1. Anderrson Kristin and Carol Kerr. Customer Relationship Management. Tata McGraw-Hill, 2002.
2. Ed Peelen. Customer Relationship Management. Prentice Hall, 2005.

3. BhasinJaspreet Kaur. Customer Relationship Management. Dreamtech Press, 2012
4. Valarie A Zeithmal, Mary Jo Bitner, Dwayne D Gremler and Ajay Pandit. Services Marketing Integrating Customer Focus Across the Firm. Tata McGraw Hill, 2010.
5. UrvashiMakkar and Harinder Kumar Makkar. CRM Customer Relationship Management. McGraw Hill Education, 2013.
6. Baran Roger J. & Robert J. Galka. Customer Relationship Management: The Foundation of Contemporary Marketing Strategy, Routledge Taylor & Francis Group. 2014

SEC IV: Data Science Soft Skills

Course Objective

The overall objective of a data science soft skills program is to enhance the effectiveness and impact of data scientists in their roles by developing their soft skills. This course will provide data scientists with the necessary skills and competencies to effectively communicate their findings, collaborate with others, solve complex problems, and make data-driven decisions in a responsible and ethical manner.

Course Outcome

CO1: Develop ability to communicate complex data concepts, insights, and recommendations to technical and non-technical stakeholders.

CO2: Foster ability to work collaboratively in multidisciplinary teams.

CO3: Strengthen critical thinking abilities to analyse and solve data-related problems.

CO4: Cultivate understanding of the business context in which they operate.

CO5: Promote understanding and practice of ethical considerations in data science.

Unit 1 Effective Communication: Technical Communication, Non-technical Communication- Data Documentation- Storytelling and presentation-Listening Skills

Unit 2 Team Work: Team Dynamics- Collaboration and Cooperation- Conflict Resolution- Manging Diverse Perspectives

Unit 3 Critical Thinking and Problem Solving: Cognitive Bias- Risk Assessment and Management- Root Cause Analysis- Ethical Reasoning- Reflective thinking and learning

Unit 4 Time Management: Setting goals and priorities- Time tracking and assessment- Planning and Scheduling- Procrastination Management- Dealing with time pressures and deadlines- Self care and work-life balance

Unit 5 Creativity and Attention to Detail: Fostering creative mindset- enhancing observational skills- Strategies for overcoming creative blocks-Multitasking

NISI DOMINUS FRUSTRA