MAJOR

B.Sc - Data Science Syllabus

Index (Semester II to VIII)

Year	Semester	Paper	Subject	IA	EA	Practical	Total
1	II	II A	Principles of Data Science	25	75	-	100
		II B	Fundamental of Statistics	25	75	-	100
2	III	III A	Introduction to Data science With R Introduction to Data science With R Practical Course	25	75	50	150
		III B	Probability Theory and Distribution	25	75	-	100
		III C	Statistical Methods	25	75	-	100
		III D	Big Data Technology Big Data Technology Practical Course	25	75	50	150
	IV	IV A	Data Mining and Data Analysis Data Mining and Data Analysis Practical Course	25	75	50	150
		IV B	Regression Analysis	25	75	-	100
		IV C	Sampling Techniques and Designs of Experiments	25	75	-	100
3	V	V A	Optimization Technique	25	75	-	100
		V B	Operations Research	25	75	-	100
		V C	Statistical Process and Quality Control	25	75	-	100
		V D	Big data Acquisition and Analysis Big data Acquisition and Analysis Practical Course	25	75	50	150
4	VII	VII A	Techniques and Tools for Data Science Techniques and Tools for Data Science Practical Course	25	75	50	150
		VII B	Data Analysis & Visualization Data Analysis & Visualization Practical Course	25	75	50	150
		VII C	Data Analytical with Python Data Analytical with Python Practical Course	25	75	50	150
5	VIII	VIII A	Time Series Analysis and Forecasting	25	75	-	100
		VIII B	Data Analytics : Descriptive , Predictive and Prescriptive	25	75	-	100
		VIII C	Numerical Methods	25	75		100

MAJOR
B.Sc -Data Science

Syllabus for Semester II

MAJOR
B.Sc Data Science – I Year
II Semester Paper: II A

PRINCIPLES OF DATA SCIENCE

COURSE OBJECTIVES:

To provide strong foundation for data science and application area related to information technology and understand the underlying core concepts and emerging technologies in data science

COURSE OUTCOMES:

Upon completion of this course, the students should be able to:

1. Explore the fundamental concepts of data science
2. Understand data analysis techniques for applications handling large data
3. Understand various machine learning algorithms used in data science process
4. Visualize and present the inference using various tools
5. Learn to think through the ethics surrounding privacy, data sharing and algorithmic decision-making

UNIT-1-INTRODUCTION TO DATA SCIENCE (9HRS)

Definition – Big Data and Data Science Hype – Why data science – Getting Past the Hype – The Current Landscape – Who is Data Scientist? - Data Science Process Overview – Defining goals – Retrieving data – Data preparation – Data exploration – Data modeling – Presentation.

UNIT-2 -BIG DATA (9HRS)

Problems when handling large data – General techniques for handling large data – Case study – Steps in big data – Distributing data storage and processing with Frameworks – Case study.

UNIT-3-MACHINE LEARNING (9HRS)

Machine learning – Modeling Process – Training model – Validating model – Predicting new observations –Supervised learning algorithms – Unsupervised learning algorithms.

UNIT-4-DEEP LEARNING (9HRS)

Introduction – Deep Feedforward Networks – Regularization – Optimization of Deep Learning – Convolutional Networks – Recurrent and Recursive Nets – Applications of Deep Learning.

UNIT-5 - DATA VISUALIZATION (9HRS)

Introduction to data visualization – Data visualization options – Filters – MapReduce – Dashboard development tools – Creating an interactive dashboard with dc.js-summary.

TEXT BOOKS:

1. Introducing Data Science, Davy Cielen, Arno D. B. Meysman, Mohamed Ali, Manning Publications Co., 1st edition, 2016
2. An Introduction to Statistical Learning: with Applications in R, Gareth James, Daniela Witten, Trevor Hastie, Robert Tibshirani, Springer, 1st edition, 2013
3. Deep Learning, Ian Goodfellow, Yoshua Bengio, Aaron Courville, MIT Press, 1st edition, 2016
4. Ethics and Data Science, D J Patil, Hilary Mason, Mike Loukides, O' Reilly, 1st edition, 2018

REFERRENCE BOOKS:

1. Data Science from Scratch: First Principles with Python, Joel Grus, O'Reilly, 1st edition, 2015
2. Doing Data Science, Straight Talk from the Frontline, Cathy O'Neil, Rachel Schutt, O' Reilly, 1st edition, 2013.
3. Mining of Massive Datasets, Jure Leskovec, Anand Rajaraman, Jeffrey David Ullman, Cambridge University Press, 2nd edition, 2014

MAJOR
B.Sc Data Science – I Year
II Semester Paper: II B

FUNDAMENTALS OF STATISTICS

COURSE OBJECTIVES:
To enable the students to understand the fundamentals of statistics to apply descriptive measures and probability for data analysis.

COURSE OUTCOMES:
Upon completion of this course, the students should be able to:

1. Understand the science of studying & analyzing numbers.
2. Identify and use various visualization tools for representing data.
3. Describe various statistical formulas.
4. Compute various statistical measures.

UNIT I - Statistics: (12HRS)
Introduction to Statistics – Origin of Statistics, Features of Statistics, Scope of Statistics, Functions of Statics, Uses and importance of Statistics, Limitation of Statistics, Distrust of Statistics.

UNIT II - Collection of Data: (12HRS)
Introduction to Collection of Data, Primary and Secondary Data, Methods of Collecting Primary Data, Methods of Secondary Data, Statistical Errors, Rounding off Data (Approximation).

UNIT III - Classification of Data Frequency Distribution :(12HRS)
Introduction Classification of Data, Objectives of Classification, Methods of Classification, Ways to Classify Numerical Data or Raw Data. Tabular, Diagrammatic and Graphic Presentation of Data: Introduction to Tabular Presentation of Data, Objectives of Tabulation, Components of a Statistical Table, General Rules for the Construction of a Table, Types of Tables, Introduction to Diagrammatic Presentation of Data, Advantage and Disadvantage of Diagrammatic Presentation, Types of Diagrams, Introduction to Graphic Presentation of Data, Advantage and Disadvantage of Graphic Presentation, Types of Graphs.

UNIT IV - Measures of Central tendency: (12HRS)
Introduction to Central Tendency, Purpose and Functions of Average, Characteristics of a Good Average, Types of Averages, Meaning of Arithmetic Mean, Calculation of Arithmetic Mean, Merit and Demerits of Arithmetic Mean, Meaning of Median, Calculation of Median, Merit and Demerits of Median, Meaning of Mode, Calculation of Mode, Merit and Demerits of Mode, Harmonic Mean-Properties Merit and Demerits.

UNIT V - Measures of Dispersion: (12HRS)

Meaning of Dispersion, Objectives of Dispersion, Properties of a good Measure of Dispersion, Methods of Measuring Dispersion, Range Introduction, Calculation of Range , Merit and Demerits of Range, Mean Deviation, Calculation of Mean Deviation , Merit and Demerits of Mean Deviation, Standard Deviation Meaning, Calculation of Standard Deviation , Merit and Demerits of Standard Deviation, Coefficient of Variation, Calculation of Coefficient Variance, Merit and Demerits of Coefficient of Variation.

TEXT BOOKS:

1. Statistics and Data Analysis, A.Abebe, J. Daniels, J.W.Mckean, December 2000.
2. Statistics, Tmt. S. EzhilarasiThiru, 2005, Government of Tamilnadu.
3. Introduction to Statistics, David M. Lane.
4. Weiss, N.A., Introductory Statistics. Addison Wesley, 1999.
5. Clarke, G.M. & Cooke, D., A Basic course in Statistics. Arnold, 1998.

REFERENCE BOOKS:

1. Banfield J.(1999), Rweb: Web-based Statistical Analysis, Journal of Statistical Software.
2. Bhattacharya,G.K. and Johnson, R.A.(19977), Statistical Concepts and Methods, New York, John Wiley & Sons.

MAJOR
B.Sc -Data Science

Syllabus for Semester III

MAJOR
B.Sc Data Science – II Year
III Semester Paper: III A

Introduction to Data science With R

Objective

Data Science is a fast-growing interdisciplinary field, focusing on the analysis of data to extract knowledge and insight. This course will introduce students to the collection. Preparation, analysis, modeling and visualization of data, covering both conceptual and practical issues. Examples and case studies from diverse fields will be presented, and hands-on use of statistical and data manipulation software will be included.

Outcomes

1. Recognize various disciplines that contribute to a successful data science effort.
2. Understand the processes of data science - identifying the problem to be solved, data collection, preparation, modeling, evaluation and visualization.
3. Be aware of the challenges that arise in data sciences.
4. Develop and appreciate various techniques for data modeling and mining.
6. Be cognizant of ethical issues in many data science tasks.
7. Be comfortable using commercial and open source tools such as the R language and its associated libraries for data analytics and visualization.
8. Learn skills to analyze real time problems using R
9. Able to use basic R data structures in loading, cleaning the data and preprocessing the data.
10. Able to do the exploratory data analysis on real time datasets
11. Able to understand and implement Linear Regression
12. Able to understand and use - lists, vectors, matrices, data frames, etc.

Syllabus:

Unit-1:

Introduction to Data Science- Introduction- Definition - Data Science in various fields - Examples - Impact of Data Science - Data Analytics Life Cycle - Data Science Toolkit - Data Scientist - Data Science Team

Understanding data: Introduction – Types of Data: Numeric – Categorical – Graphical – High Dimensional Data – Classification of digital Data: Structured, Semi-Structured and Un-Structured - Example Applications. Sources of Data: Time Series – Transactional Data – Biological Data – Spatial Data – Social Network Data – Data Evolution.

Unit-2:

Introduction to R- Features of R - Environment - R Studio. Basics of R-Assignment - Modes - Operators - special numbers - Logical values - Basic Functions - R help functions - R Data Structures - Control Structures. Vectors: Definition- Declaration - Generating - Indexing - Naming - Adding & Removing elements - Operations on Vectors - Recycling - Special Operators - Vectorized if- then else-Vector Equality – Functions for vectors - Missing values - NULL values - Filtering & Subsetting.

Unit-3:

Matrices - Creating Matrices - Adding or removing rows/columns - Reshaping - Operations - Special functions on Matrices. Lists - Creating List – General List Operations - Special Functions - Recursive Lists. Data Frames - Creating Data Frames - Naming - Accessing - Adding - Removing - Applying Special functions to Data Frames - Merging Data Frames- Factors and Tables.

Unit- 4:

Input / Output – Reading and Writing datasets in various formats - Functions - Creating User-defined functions - Functions on Function Object - Scope of Variables - Accessing Global, Environment - Closures - Recursion. Exploratory Data Analysis - Data Preprocessing - Descriptive Statistics - Central Tendency - Variability - Mean - Median - Range - Variance - Summary - Handling Missing values and Outliers - Normalization

Data Visualization in R : Types of visualizations - packages for visualizations - Basic Visualizations, Advanced Visualizations and Creating 3D plots.

Unit- 5:

Inferential Statistics with R - Types of Learning - Linear Regression- Simple Linear Regression - Implementation in R - functions on lm() - predict() - plotting and fitting regression line. Multiple Linear Regression - Introduction -comparison with simple linear regression - Correlation Matrix - F-Statistic - Target variables Vs Predictors - Identification of significant features - Implementation of Multiple Linear Regression in R.

References

1. Nina Zumel, John Mount, “Practical Data Science with R”, Manning Publications, 2014.
2. .Jure Leskovec, Anand Rajaraman, Jeffrey D.Ullman, “Mining of Massive Datasets”, Cambridge University Press, 2014.
3. 3.Mark Gardener, “Beginning R - The Statistical Programming Language”, John Wiley & Sons, Inc., 2012.
4. W. N. Venables, D. M. Smith and the R Core Team, “An Introduction to R”, 2013. 5.Tony Ojeda, Sean Patrick Murphy, Benjamin Bengfort, Abhijit Dasgupta, “Practical Data Science Cookbook”, Packt Publishing Ltd., 2014.
5. Nathan Yau, “Visualize This: The FlowingData Guide to Design, Visualization, and Statistics”, Wiley, 2011.
6. Boris lublinsky, Kevin t. Smith, Alexey Yakubovich, “Professional Hadoop Solutions”.

Introduction to Data science With R – Practices

R Programming LAB

1) Installing R and R studio

2) Create a folder DS_R and make it a working directory. Display the current working directory

3) Installing the "ggplot2", "caTools", "CART" packages

	C1	C2	C3	C4	C5
C1	0	12	13	8	20
C2	12	0	15	28	88
C3	13	15	0	6	9
C4	8	28	6	0	33
C5	20	88	9	33	0

- Find the pairs of cities with shortest distance.

4) Load the packages "ggplot2", "caTools".

5) Basic operations in r

6) Working with Vectors:

- Create a vector v1 with elements 1 to 20.
- Add 2 to every element of the vector v1.
- Divide every element in v1 by 5
- Create a vector v2 with elements from 21 to 30. Now add v1 to v2.

7) Getting data into R, Basic data manipulation

8) Using the data present in the table given below, create a Matrix "M"

Find the pairs of cities with shortest distance.

9) Consider the following marks scored by the 6 students

Section	Student no	M1	M2	M3
A	1	45	54	45
A	2	34	55	55
A	3	56	66	64
B	1	43	44	45
B	2	67	76	78
B	3	76	68	37

- Create a data structure for the above data and store in proper positions with proper names
- Display the marks and totals for all students
- Display the highest total marks in each section.
- Add a new subject and fill it with marks for 2 sections.

Three people denoted by P1, P2, P3 intend to buy some rolls, buns, cakes and bread. Each of them needs these commodities in differing amounts and can buy them in two shops S1, S2. The individual prices and desired quantities of the commodities are given in the following table "demand.

	price	
	S1	S2
Roll	1.5	1
Bun	2	2.5
Cake	5	4.5
Bread	16	17

	demand.quantity			
	Roll	Bun	Cake	Bread
P1	6	5	3	1
P2	3	6	2	2
P3	3	4	3	1

- Create matrices for above information with row names and col names.
- Display the demand. Quantity and price matrices
- Find the total amount to be spent by each person for their requirements in each shop
- Suggest a shop for each person to buy the products which is minimal.

10) Consider the following employee details:

employee details as follows		
	emp_no:1	
	name: Ram	
	salary	
		basic: 10000
		hra: 2500
		da: 4000
	deductions	
		pf: 1100
		tax: 200
	total salary	
		gs(Gross Salary):
		ns(Net Salary)

- Create a list for the employee data and fill gross and net salary.
- Add the address to the above list
- display the employee name and address
- remove street from address
- remove address from the List.

11) Loops and functions - Find the factorial of a given number

12) Implementation of Data Frame and its corresponding operators and functions

13) Implementation of Reading data from the files and writing output back to the specified file

14) Treatment of NAs, outliers, Scaling the data, etc

15) Applying summary () to find the mean, median, standard deviation, etc

16) Implementation of Visualizations - Bar, Histogram, Box, Line, scatter plot, etc.

17) Implementation of Linear and multiple Linear Regression

18) Fitting regression line.

MAJOR
B.Sc Data Science – II Year
III Semester Paper: III B

Probability Theory and Distribution

COURSE OBJECTIVES:

At the end of the course, the students will be able to:

- Calculate probabilities by applying probability laws and theoretical results.
- Identify an appropriate probability distribution for a given discrete or continuous random variable and use its properties to calculate probabilities.
- Calculate statistics such as the mean and variance of common probability distributions.
- Calculate probabilities for joint distributions including marginal and conditional probabilities.
- Determine whether random variables are independent and find their covariance and correlation.
- Explain the role of probability in hypothesis testing and describe issues related to interpreting statistical significance.

Syllabus:

Unit-1: Probability:

Introduction, random experiments, sample space, events and algebra of events. Definitions of Probability – classical, statistical, and axiomatic. Conditional Probability, laws of addition and multiplication, independent events, Theorem of total probability, Bayes' theorem and its applications.

Unit -2: Random variables (RV):

Introduction to Random variables, - One dimensional Random Variables, Discrete and Continuous RV- Density and Distribution function of RV, Expectation, Variance, and its properties, Covariance, and Moments.

Unit – 3: Mathematical Expectation

Random variable and of a function of a random variable. Moments and covariance using mathematical expectation with examples. Addition and Multiplication theorems on expectation. Definitions of M.G.F., C.G.F, P.G.F, C.F., Statements of Properties. Chebyshev and Cauchy – Schwartz inequalities.

Unit – 4: Discrete distribution

Binomial, Poisson, Negative Binomial, Hypergeometric distribution (mean and variance only) and properties.

Unit-5 : Continuous Distribution

Rectangular, exponential, gamma, beta of two kinds (mean and variance only) and properties. Normal distribution (mean and variance only) and its properties.

Text book and Reference books:

1. Probability and Statistics for engineers and scientists by R.E.Walpole, R.H.Mayers, S.L.Mayers and K.Ye, 9th Edition, Pearson Education (2012).
2. Probability, Statistics and Reliability for Engineers and Scientists by Bilal M. Ayub and Richard H. McCuen, 3rd edition, CRC press (2011).
3. Mathematical Statistics by M. Ray, H S Sharma, and S Chaudhary, RP & Sons Education.
4. Fundamental of Mathematical Statistics by T Veerarajan, Yes Dee Publishing Pvt Ltd.
5. Probability and Statistics for Engineers by R.A.Johnson, Miller & Freund‘s, 8th edition, Prentice Hall India (2010)

MAJOR
B.Sc Data Science – II Year
III Semester Paper: III C

STATISTICAL METHODS

COURSE OBJECTIVES:

At the end of the course, the students will be able to:

- Knowledge of Statistics and its implementation through practical understanding for various domains related to data science.
- Knowledge of various types of data, their organization and evaluation of summary measures such as measures of central tendency and dispersion etc.
- Knowledge of other types of data reflecting quality characteristics including concepts of independence and association between two attributes, insights into preliminary exploration of different types of data.
- Knowledge of correlation, regression analysis, regression diagnostics, partial and multiple correlations.

Syllabus:

UNIT –I : Curve fitting:

Bi- variate data, Principle of least squares, fitting of degree polynomial. Fitting of straight line, Fitting of Second degree polynomial or parabola, Fitting of power-curve and exponential curves.

UNIT-II: Correlation:

Meaning, Types of Correlation, Measures of Correlation: Scatter diagram, Karl Pearson's Coefficient of Correlation, Rank Correlation Coefficient (with and without ties), Bi- variate frequency distribution, correlation coefficient for bi-variate data and simple problems. Concept of multiple and partial correlation coefficients (three variables only) and properties.

UNIT III: Regression:

Concept of Regression, Linear Regression: Regression lines, Regression coefficients and it's properties, Regressions lines for bi-variate data and simple problems. Correlation vs regression, sigmoid curve, derivation from linear regression to logistic regression.

UNIT-IV: Attributes:

Notations, Class, Order of class frequencies, Ultimate class frequencies, Consistency of data, Conditions for consistency of data for 2 and 3 attributes only , Independence of attributes.

UNIT-V: Attributes:

Association of attributes and its measures, Relationship between association and colligation of attributes, Contingency table: Square contingency, Mean square contingency, Coefficient of mean square contingency**.**

Text book and Reference books:

1. V.K. Kapoor and S.C.Gupta: Fundamentals of Mathematical Statistics, Sultan Chand &Sons, New Delhi.
2. BA/B.Sc I year statistics-descriptive statistics, probability distribution-Telugu Academy-Dr M.Jaganmohan Rao, Dr N.Srinivasa Rao, Dr P.Tirupathi Rao, Smt.D.Vijayalakshmi.
3. K.V.S.Sarma: Statistics

REFERENCEBOOKS:

1. WillamFeller: Introduction to Probability theory and its applications. Volume–I,Wiley
2. Goon AM, GuptaMK, Das GuptaB: Fundamentals of Statistics,Vol-I,the World Press Pvt.Ltd., Kolakota.
3. HoelP.G: Introduction to mathematical statistics, Asia Publishing house.
4. M.Jagan Mohan Rao and PapaRao: ATextbook of Statistics Paper-I.
5. Sanjay Arora and Bansi Lal: New Mathematical Statistics: Satya Prakashan, NewDelhi.

MAJOR
B.Sc Data Science – II Year
III Semester Paper: III D

BIG DATA TECHNOLOGY

COURSE OBJECTIVES:

This course provides practical foundation level training that enables immediate and effective participation in big data projects. The course provides grounding in basic and advanced methods to big data technology and tools, including MapReduce and Hadoop and its ecosystem.

COURSE OUTCOME:

- Learn tips and tricks for Big Data use cases and solutions.
- Acquire knowledge of HDFS components, Name node, Data node, etc.
- Acquire knowledge of storing and maintaining data in cluster, reading data from and writing data to Hadoop cluster.
- Able to maintain files in HDFS
- Able to write MapReduce applications to access data present on HDFS
- Able to read different formats of files into map-reduce application.
- Able to develop MapReduce applications to analyze Big Data related to the real world use cases.
- Able to write MapReduce applications that can take data from multiple datasets and join them
- Able to optimize the performance of Map-Reduce application

Syllabus:

UNIT – I: Introduction to Big Data

Introduction –Distributed File System – Big Data and its importance, Characteristics of Big Data, Limitation of Conventional Data Processing Approaches, Need of big data frameworks, Big data analytics, Limitations of Big Data and Challenges, Big data applications.

UNIT – II: Hadoop:

Basic Concepts of Hadoop and its features -The Hadoop Distributed File System (HDFS)- Anatomy of a Hadoop Cluster - Hadoop cluster modes - Hadoop Architecture, Hadoop Storage - Hadoop daemons (Name node-Secondary name node-Job tracker-Task tracker-Data node,etc) - Anatomy of Read & Write operations – Interacting HDFS using command-line (HDFS Shell and FS shell commands) -Interacting HDFS using Java APIs – Dataflow – Blocks –Replica - YARN.

UNIT – III: Hadoop Ecosystem Components

Schedulers- Fair and Capacity, Hadoop 2.0 Vs Hadoop 3.0 and its new features.

Hadoop Cluster Setup – SSH & Hadoop Configuration –HDFS Administering – Monitoring & Maintenance.

UNIT – IV : Hadoop MapReduce

Introduction - Phases in MapReduce Framework - Anatomy of MapReduce Job run - Failures, Job Scheduling, Shuffle and Sort, Task Execution, Map Reduce

Types and Formats, Map Reduce Features. Understanding Basic MapReduce Program (Word Count program): The Driver Code - The Mapper class - The Reducer class.

UNIT-V

Writing first MapReduce Program - Hadoop's Streaming API - Using Eclipse for Rapid Development – YARN Vs MapReduce Advanced MapReduce Concepts: Partitioner – Combiner – Joins – Map-side Join – Reduce-side Join - Case Study: Weblog Analysis done using Mapper, Reducer, Combiner, Partitioner, etc.

Text Books :

References

1. Boris lublinsky, Kevin t. Smith Alexey Yakubovich, "Professional Hadoop Solutions". Wiley, ISBN : 9788126551071, 2015.
2. Chris Eaton, Dirk Deroos et al., "Understanding Big Data", McGraw Hill , 2010.
3. Tom White, "HADOOP" : The definitive Guide", O Reilly 2012.
4. Srinath Perera, Thilina Gunarathne, "Hadoop MapReduce Cookbook", PACKT publishing, 2013

Student Activity: - BIG DATA LAB

1. **Case Study I**: Centers for Medicare & Medicaid Services: The Integrity of Healthcare Data and Secure Payment Processing.
2. **Case Study II**: Movie Lens Data set Analysis
3. **Case Study III**: Web Server Log Analysis using MapReduce.

RECOMMENDED CO-CURRICULAR ACTIVITIES:

(Co-curricular activities shall not promote copying from textbook or from others work and shall encourage self/independent and group learning)

A. Measurable

1. Assignments (in writing and doing forms on the aspects of syllabus content and outside the syllabus content. Shall be individual and challenging)
2. Student seminars (on topics of the syllabus and related aspects (individual activity))
3. Quiz (on topics where the content can be compiled by smaller aspects and data (Individuals or groups as teams))

4. Study projects (by very small groups of students on selected local real-time problems pertaining to syllabus or related areas. The individual participation and contribution of students shall be ensured (team activity

B. General

1. Group Discussion
2. Try to solve MCQ's available online.
3. Others

RECOMMENDED CONTINUOUS ASSESSMENT METHODS:

Some of the following suggested assessment methodologies could be adopted;

1. The oral and written examinations (Scheduled and surprise tests)
2. Closed-book and open-book tests
3. Problem-solving exercises
5. Practical assignments and laboratory reports
6. Observation of practical skills
7. Individual and group project reports like "Movie Lens Data Analysis", "Youtube Click stream Data Analysis", etc.
8. Efficient delivery using seminar presentations,
9. Viva voce interviews.
10. Computerized adaptive testing, literature surveys and evaluations,
11. Peers and self-assessment, outputs form individual and collaborative work.

MAJOR
B.Sc -Data Science

Syllabus for Semester IV

MAJOR
B.Sc Data Science – II Year
IV Semester Paper: IV A
DATA MINING AND DATA ANALYSIS

Objective

- To learn data analysis techniques.
- To understand Data mining techniques and algorithms.
- Comprehend the data mining environments and application.

Outcomes

Students who complete this course will be able to

- To understand and demonstrate data mining
- Compare various conceptions of data mining as evidenced in both research and application.
- Characterize various kinds of patterns that can be discovered by association rule mining.
- Evaluate mathematical methods underlying the effective application of data mining.
- To Analyze the data using statistical methods
- Gain hands-on skills and experience on data mining tools.

Syllabus:

Unit-1: Data mining

Data mining - KDD Vs Data Mining, Stages of the Data Mining Process-Task Primitives, Data Mining Techniques – Data Mining Knowledge Representation. Major Issues in Data Mining – Measurement and Data – Data Preprocessing – Data Cleaning - Data transformation- Feature Selection - Dimensionality reduction

Unit-2: Classification and Descriptive Analytics

Rule Based Classification – Classification by Back propagation – Support Vector Machines – Associative Classification – Lazy Learners – Other Classification Methods – Prediction.
Descriptive Analytics - Mining Frequent Item sets - Market based model – Association and Sequential Rule Mining.

Unit-3: Predictive Analytics

Classification and Prediction **-** Basic Concepts of Classification and Prediction, General Approach to solving a classification problem- Logistic Regression - LDA - Decision Trees: Tree Construction Principle – Feature Selection measure – Tree Pruning - Decision Tree construction Algorithm, Random Forest, Bayesian Classification-Accuracy and Error Measures- Evaluating the Accuracy of the classifier / predictor- Ensemble methods and Model selection.

Unit- 4: Factor Analysis

Factor Analysis: Meaning, objectives and Assumptions, Designing a factor analysis, Deriving factors and assessing overall factors, Interpreting the factors and validation of factor analysis.

Unit- 5: Cluster Analysis

Cluster Analysis: Basic concepts and Methods – Cluster Analysis – Partitioning methods – Hierarchical methods – Density Based Methods – Grid Based Methods – Evaluation of Clustering – Advanced Cluster Analysis: Probabilistic model based clustering – Clustering High – Dimensional Data – Clustering Graph and Network Data – Clustering with Constraints- Outlier Analysis.

References

1. Adelchi Azzalini, Bruno Scapa, "Data Analysis and Data mining" , 2nd Ediiton, Oxford Univeristy Press Inc., 2012.
2. Jiawei Han and Micheline Kamber, "Data Mining: Concepts and Techniques", 3rd Edition, Morgan Kaufmann Publishers, 2011.
3. Alex Berson and Stephen J. Smith, "Data Warehousing, Data Mining & OLAP", 10th Edition, TataMc Graw Hill Edition , 2007.
4. G.K. Gupta, "Introduction to Data Mining with Case Studies", 1st Edition, Easter Economy Edition, PHI, 2006.
5. Joseph F Hair, William C Black etal, "Multivariate Data Analysis", Pearson Education, 7th edition, 2013.

DATA MINING AND DATA ANALYSIS – Practices

II YEAR IV SEMESTER PAPER– IV

DATA MINIG AND DATA ANALYSIS LAB

1. Data Analysis – Getting to know the Data (Using ORANGE WEKA or R Programming)

- Parametric – Means, T-Test, Correlation
- Prediction for numerical outcomes – Linear regression, Multiple Linear Regression
- Correlation analysis
- Preparing data for analysis
- Pre-Processing techniques

2. Data Mining (Using ORANGE WEKA or R Programming)

- Implement clustering algorithm
- Implement Association Rule mining
- Implement classification using
    - Decision tree
    - Back Propagation
    - Logistic Regression
    - Decision Tree
    - Random Forest
    - Naive Bayes
    - Support Vector Machines
- Visualization methods

MAJOR
B.Sc Data Science – II Year
IV Semester Paper: IV B

Regression Analysis

Objective

After learning this course the student will be able

1. To know about regression techniques, which are powerful tools in statistics,
2. To get an idea of Linear and Multiple Linear regression,
3. To learn about regression diagnostics, residual plots for visualization
4. To perform statistical tests of hypotheses on regression coefficients.
5. To study the structural stability of a regression model.
6. To learn the regression with qualitative independent and dependent variables by dummy variable technique.
7. To learn the selection of the best regression model.

Unit I: Simple Linear Regression:

Simple Linear Regression Model. Least-Squares Estimation of the Parameters - Estimation of $\beta 0$ and $\beta 1$, Properties of the Least-Squares Estimators and the Fitted Regression Model. Hypothesis Testing on the Slope and Intercept -Use of t Tests, Testing Significance of Regression and Analysis of Variance

Unit II: Multiple Linear Regression:

Multiple linear regression: Multiple Linear Regression Model. Estimation of model parameters: Least-Squares Estimation of the Regression Coefficients, Properties of the Least-Squares Estimators. Concept of residual, Residual plots. Test for Significance of Individual Regression Coefficients, and subsets of coefficients. Concept of coefficient of determination.

Unit III: Regressions with Qualitative Independent Variables:

Use of dummy variables to handle categorical independent variables in regression. Estimation of model parameters with dummy variables - Testing the structural stability of regression models, comparing the slopes of two regression models. Multiple linear regression with interaction effects.

Unit IV: Regressions with Qualitative Dependent Variables:

Binary logistic regression with several independent variables, estimation of coefficients, evaluating the Odds Ratio (OR) and its interpretation. The concept of

Unit – V Best Model:

Selecting 'Best' regression model. All possible regressions – R^2, Adjusted R^2, MSRes, Mallow's statistic. Sequential selection of variables – criteria for including and eliminating a variable – forward selection, backward elimination and stepwise regression.

References:

1. Douglas C. Montgomery, Elizabeth A. Peck, G. Geoffrey Vining (2012), Introduction To Linear Regression Analysis, Fifth Edition, John Wiley & Sons
2. Draper, N. R. and Smith, H. (1998). Applied Regression Analysis. 3rd Edition. John Wiley.
3. Hosmer, D. W., Lemeshow, S. and Sturdivant R.X. (2013). Applied Logistic Regression, Wiley Blackwell.
4. Montgomery, D. C., Peck, E. A. and Vining, G. G. (2013). Introduction to Linear Regression Analysis. 5th Edition. Wiley.
5. Neter, J., Kutner, M. H., Nachtsheim, C.J. and Wasserman, W. (1996). Applied Linear Statistical Models, 4th Edition, Irwin USA.
6. Gujarati, D. and Sangeetha, S. (2007). Basic Econometrics, 4th Edition

MAJOR
B.Sc Data Science – II Year
IV Semester Paper: IV C

Sampling Techniques and Designs of Experiments

Objective

- The sampling techniques deals with the ways and methods that should be used to draw samples to obtain the optimum results.
- This paper throw light on understanding the variability between group and within group through Analysis of Variance
- This gives an idea of logical construction of Experimental Design and applications of these designs now days in various research areas.
- Factorial designs allow researchers to look at how multiple factors affect a dependent variable, both independently and together.

Course Learning Outcomes

- The students shall get various statistical sampling schemes such as simple, stratified and systematic sampling.
- Knowledge about comparing various sampling techniques.
- carry out one way and two way Analysis of Variance,
- understand the basic terms used in design of experiments,
- use appropriate experimental designs to analyze the experimental data.

Syllabus:

UNIT I

Simple Random Sampling (with and without replacement): Notations and terminology, various probabilities of selection. Random numbers tables and its uses. Methods of selecting simple random sample, lottery method, method based on random numbers. Estimates of population total, mean and their variances and standard errors, determination of sample size, simple random sampling of attributes.

UNIT II

Stratified random sampling: Stratified random sampling, Advantages and Disadvantages of Stratified Random sampling, Estimation of population mean, and its variance. Stratified random sampling with proportional and optimum allocations. Comparison between proportional and optimum allocations with SRSWOR.

Systematic sampling: Systematic sampling definition when N = nk and merits and demerits of systematic sampling - estimate of mean and its variance. Comparison of systematic sampling with Stratified and SRSWOR.

UNIT III

Analysis of variance : Analysis of variance(ANOVA) –Definition and assumptions. One-way with equal and unequal classification, Two way classification.

Design of Experiments: Definition, Principles of design of experiments, CRD: Layout, advantages and disadvantage and Statistical analysis of Completely Randomized Design(C.R.D).

UNIT IV

Randomized Block Design (R.B.D) and Latin Square Design (L.S.D) with their layouts and Analysis, MissingplottechniqueinRBDandLSD.EfficiencyRBDoverCRD,EfficiencyofLSDoverRBDand CRD.

UNIT V

Factorial experiments – Main effects and interaction effects of 2^2 and 2^3 factorial experiments and their Statistical analysis. Yates procedure to find factorial effecttotals.

Text Books:

1. Telugu AcademyBA/BSc III year paper - III Statistics - applied statistics - Telugu academy by Prof.K.Srinivasa Rao, Dr D.Giri. Dr A.Anand, Dr V.PapaiahSastry.
2. K.V.S. Sarma: Statistics Made Simple: Do it yourself on PC.PHI.

Reference Books:

1. Fundamentals of applied statistics : VK Kapoor and SCGupta.
2. Indian Official statistics - MR Saluja.
3. Anuvarthita SankyakaSastram - TeluguAcademy.

MAJOR

B.Sc -Data Science

Syllabus for Semester V

MAJOR
B.Sc Data Science – III Year
V Semester Paper: V A

Optimization Technique

Course Outcomes (COs):

Upon completion of the course, the students will be able to:

- **Analyse** the real-life systems with limited constraints.
- **Depict** the systems in a mathematical model form.
- **Solve** the mathematical model manually as well as using soft resources/software under the given constraints.
- **Describe** the Concept of optimization and classification of optimization problems.
- **Understand** variety of real industrial problems such as resource allocation, production planning, assignment, transportation, travelling salesman etc. and solve these problems using linear programming approach using software.

Syllabus:

Unit I:
Introduction of OR – Origin and development of OR – Nature and features of OR –Scientific Method in OR – Modeling in OR – Advantages and limitations of Models-General Solution methods of OR models – Applications of Operation Research. Linear programming problem (LPP) -Mathematical formulation of the problem - illustrations on Mathematical formulation of Linear programming of problem. Graphical solution of linear programming problems. Some exceptional cases - Alternative solutions, Unbounded solutions, non-existing feasible solutions by Graphical method.

Unit II:
General linear programming Problem(GLP) – Definition and Matrix form of GLP problem, Slack variable, Surplus variable, unrestricted Variable, Standard form of LPP and Canonical form of LPP. Definitions of Solution, Basic Solution, Degenerate Solution, Basic feasible Solution and Optimum Basic Feasible Solution. Introduction to Simplex method and Computational procedure of simplex algorithm. Solving LPP by Simplex method (Maximization case and Minimization case)

Unit III:
Artificial variable technique - Big-M method and Two-phase simplex method, Degeneracy in LPP and method to resolve degeneracy. Alternative solution, Unbounded solution, Non existing feasible solution and Solution of simultaneous equations by Simplex method.

Unit IV:
Duality in Linear Programming –Concept of duality -Definition of Primal and Dual Problems, General rules for converting any primal into its Dual, Economic interpretation of duality, Relation between the solution of Primal and Dual problem(statements only). Using duality to solve primal problem. Dual Simplex Method.

Unit V:

Post Optimal Analysis- Changes in cost Vector **C**, Changes in the Requirement Vector **b**and changes in the Coefficient Matrix **A**. Structural Changes in a LPP.

Reference Books:

1. S.D. Sharma, Operations Research, Kedar Nath Ram Nath & Co, Meerut.
2. Kanti Swarup, P.K.Gupta, Manmohn, Operations Research, Sultan Chand and sons, New Delhi.
3. J.K. Sharma, Operations Research and Application, Mc.Millan and Company, New Delhi.
4. GassS.I: Linear Programming. Mc Graw Hill.
5. HadlyG :Linear programming. Addison-Wesley.
6. Taha H.M: Operations Research: An Introduction : Mac Millan.

MAJOR
B.Sc Data Science – III Year
V Semester Paper: V B

Operations Research

Outcomes:

After learning this course, the student will be able

1. To solve the problems in logistics
2. To find a solution for the problems having space constraints
3. To minimize the total elapsed time in an industry by efficient allocation of jobs to the suitable persons.
4. To find a solution for an adequate usage of human resources
5. To find the most plausible solutions in industries and agriculture when a random environment exists.

Syllabus:

Unit 1: Transportation Problem:

Introduction, Mathematical formulation of Transportation problem. Definition of Initial Basic feasible solution of Transportation problem- North-West corner rule, Lowest cost entry method, Vogel's approximation method. Method of finding optimal solution-MODI method(U-V method). Degeneracy in transportation problem, Resolution of degeneracy, Unbalanced transportation problem. Maximization TP. Transshipment Problem.

UNIT-II : Assignment Problem

Introduction, Mathematical formulation of Assignment problem, Reduction theorem (statement only), Hungarian Method for solving Assignment problem, Unbalanced Assignment problem. The Traveling salesman problem, Formulation of Traveling salesman problem as an Assignment problem and Solution procedure.

UNIT-III : Sequencing problem:

Introduction and assumptions of sequencing problem, Sequencing of n jobs and one machine problem. Johnson's algorithm for n jobs and two machines problem- problems with n-jobs on two machines, Gantt chart, algorithm for n jobs on three machines problem- problems with n- jobs on three machines, algorithm for n jobs on m machines problem, problems with n-jobs on m-machines. Graphical method for two jobs on m– machines.

UNIT-V Game Theory:

Two-person zero-sum games. Pure and Mixed strategies. Maxmin and Minimax Principles - Saddle point and its existence. Games without Saddle point-Mixed strategies. Solution of 2 x 2 rectangular games.

Graphical method of solving 2 x n and m x 2 games. Dominance Property. Matrix oddment method for n x n games. Only formulation of Linear Programming Problem for m x n games.

UNIT-IV: Network Scheduling

Basic Components of a network, nodes and arcs, events and activities– Rules of Network construction – Time calculations in networks - Critical Path method (CPM) and PERT.

Reference Books:

1. S.D. Sharma, Operations Research, Kedar Nath Ram Nath & Co, Meerut.
2. Kanti Swarup, P.K.Gupta, Manmohn, Operations Research, Sultan Chand and sons, New Delhi.
3. J.K. Sharma, Operations Research and Application, Mc. Millan and Company, New Delhi.
4. Gass: Linear Programming. Mc Graw Hill.
5. Hadly :Linrar programming. Addison-Wesley.
6. Taha : Operations Research: An Introduction : Mac Millan.
7. Dr.NVS Raju; Operations Research, SMS education.

MAJOR
B.Sc Data Science – III Year
V Semester Paper: V C

Statistical Process and Quality Control

Course Objectives:

To understand the concept of quality, process control and product control using control chart techniques and sampling inspection plan. To have an idea about quality management, quality circles, quality movement and standardizations for quality.

Learning Outcomes:

After learning this course, the student will be able

1. To define 'quality' in a scientific way
2. To differentiate between process control and product control
3. To speak about quality awareness in industry
4. To pave a path to an industry to meet the standards
5. To effectively implement various plans to control the quality standards at various stages of an industry.

Syllabus:

Unit I

Meaning of quality, concept of total quality management (TQM) and six-sigma, ISO, comparison between TQM and Six Sigma, Meaning and purpose of Statistical Quality Control (SQC), Seven Process Control Tools of Statistical Quality Control (SQC) (i) Histogram (ii) Check Sheet, (iii) Pareto Diagram (iv) Cause and effect diagram (CED), (v) Defect concentration diagram (vi) Scatter Diagram (vii) Control chart. (Only introduction of 7 tools is expected).

Unit II

Statistical basis of Shewhart control charts, use of control charts. Interpretation of control charts, Control limits, Natural tolerance limits and specification limits. Chance causes and assignable causes of variation, justification for the use of 3-sigma limits for normal distribution, Criteria for detecting lack of control situations:
(i) At least one point outside the control limits
(ii) A run of seven or more points above or below central line.

Unit III

Control charts for Variables: Introduction and Construction of $\bar{X}$ and R chart and Standard Deviation Chart when standards are specified and unspecified, corrective action if the process is out of statistical control.
Control charts for Attributes: Introduction and Construction of p chart, np chart, C Chart and U charts when standards are specified and unspecified, corrective action if the process is out of statistical control.

Unit IV

Acceptance Sampling for Attributes: Introduction, Concept of sampling inspection plan, Comparison between 100% inspection and sampling inspection. Procedures of acceptance sampling with rectification, Single sampling plan and double sampling plan.
Producer's risk and Consumer's risk, Operating characteristic (OC) curve, Acceptable Quality Level (AQL), Lot Tolerance Fraction Defective (LTFD) and Lot Tolerance Percent Defective (LTPD), Average Outgoing Quality (AOQ) and Average Outgoing Quality Limit (AOQL), AOQ curve, Average Sample Number (ASN), Average Total Inspection (ATI).

Unit V

Single Sampling Plan: Computation of probability of acceptance using Binomial and Poisson approximation, of AOQ and ATI. Graphical determination of AOQL, Determination of a single sampling plan by: a) lot quality approach b) average quality approach.
Double Sampling Plan: Evaluation of probability of acceptance using Poisson distribution, Structure of OC Curve, Derivation of AOQ, ASN and ATI (with complete inspection of second sample), Graphical determination of AOQL, Comparison of single sampling plan and double sample plan.

Text Books:

1. **Montgomery, D. C. (2008)**: Statistical Quality Control, 6thEdn., John Wiley, New York.
2. **Parimal Mukhopadhyay:** Applied Statistics, New Central Book Agency.
3. **Goon A.M., Gupta M.K. and Das Gupta B. (1986):** Fundamentals of Statistics, Vol. II, World Press, Calcutta.
4. **S.C. Gupta and V.K. Kapoor:** Fundamentals of Applied Statistics – Chand publications.

References:

1. **R.C. Gupta:** Statistical Quality Control.
2. **Duncan A.J. (1974):** Quality Control and Industrial Statistics, fourth edition D.B. Taraporewala Sons and Co. Pvt. Ltd., Mumbai.

MAJOR
B.Sc Data Science – III Year
V Semester Paper: V D
Big data Acquisition and Analysis

Objective

Learn to develop Hadoop applications for storing processing and analyzing data stored in Hadoop cluster. The course is mainly covering Big Data tools for Data Transformation (Apache PIG), Data Analysis (HIVE) and for handling unstructured data HBase. To Understand the complexity and volume of Big Data and their challenges. To analyses the various methods of data collection. To comprehend the necessity for pre-processing Big Data and their issues.

Outcome

1. Identify the various sources of Big Data
2. Able to collect and store Big Data from various sources
3. Able to write Pig Scripts- Extract, Transform and Load the data on HDFS
4. Able to write Hive Scripts- Extract, Transform, Load and Analyze the data present in HDFS
5. Able to write scripts to extract data from structured and un-structured data for analytics
6. Able to extract and process semi and un-structured data using HBase

Syllabus:
Unit- I

Introduction To Big Data Acquisition: Big data framework – fundamental concepts of Big Data Management and analytics – Current challenges and trends in Big Data Acquisition. Map Reduce Algorithm- Hadoop Storage [HDFS], Common Hadoop Shell commands.

Unit-II

Data Collection And Transmission: Big data collection – Strategies – Types of Data Sources – Structured Vs Unstructured data – ELT vs ETL – storage infrastructure requirements – Collection methods – Log files – sensors – Methods for acquiring network data (Libcap-based and zero-copy packet capture technology).

Unit-III

Apache Pig - Introduction - Pig features - Pig Architecture - Pig Execution modes, Pig Grunt shell and Shell commands. Pig Latin Basics: Data model, Data Types, Operators - Pig Latin Commands - Load & Store , Diagnostic Operators, Grouping, Cogroup, Joining, Filtering, Sorting, Splitting - Built-In Functions, User define functions.

Unit-IV

Hive: Introduction - Hive Features - Hive architecture -Hive Meta store - Hive data types - Hive Tables.

Unit-V

HiveQL–Introduction, HiveQL Select, HiveQL – MapReduce using HiveQL OrderBy, Group By Joins, LIMIT, Distribute By , Cluster By - Sorting And Aggregation.

References

1. Bart Baesens, “Analytics in a Big Data World: The Essential Guide to Data Science and its Applications’, John Wiley & Sons, 2014.
2. Tom White “ Hadoop: The Definitive Guide” Third Edit on, O’reily Media, 2012.
3. Seema Acharya, Subhasini Chellappan, "Big Data Analytics" Wiley 2015.
4. Min Chen. Shiwen Mao, Yin Zhang. Victor CM Leung, Big Data: Related Technologies, Challenges and Future Prospects, Springer, 2014.
5. Michael Minelli, Michele Chambers Ambiga Dhiraj, “Big Data, Big Analytics : Emerging Business Intelligence and Analytic Trends”, John Wiley & Sons, 2013.
6. Raj. Pethuru “ Handbook of Research on Cloud Infrastructures for Big Data Analytics”, IGI Global.

Data Acquisition and Analysis Lab

- Hadoop Cluster Setup
- Install and Run Hive and also use Hive Shell commands to display the list of files in HDFS
- Install and Run HBase and also use HBase Shell commands to display the version and user of HBase
- Use Hive to create, alter, and drop databases, tables, views, functions, and indexes
- Write HiveQL command to create Weather table and to find the year-wise maximum temperature
- Write Hive Query to create database, managed table, external table, join, index, view, etc
- Create a table in HBase and insert the data into with Shell
- Display the data present in a HBase table using Shell

MAJOR
B.Sc -Data Science

Syllabus for Semester VII

MAJOR
B.Sc Data Science – IV Year
VII Semester Paper: VII A

Techniques and Tools for Data Science

Objective

This course deals basics of machine learning algorithms. This course helps to learn and design the simple feed forward neural network model. Also, from this course students are able to demonstrate deep learning-based experiments using real-world data. To gain knowledge on preprocessing the data using WEKA and Excel. To understand how to model a system using Scikit and TensorFlow. To find the solutions using the NLTK tool. To create visualization using Matplotlib and Tableau. To solve the real time problems of data science.

Outcome

Upon completion of this course, the students will be able to

1. Cleaning and preprocessing the data using WEKA and Excel.
2. Modeling a system using Scikit and TensorFlow.
3. Find the solutions using NLTK tool.
4. Create visualization using Matplotlib and Tableau.
5. Solve the real time problems of data science.

Syllabus

Unit I: CLEANING AND PREPROCESSING- Introduction- Preprocessing Data -File Conversion - Opening File from A Local File System –Opening File from A Web Site - Reading Data from a Database - Preprocessing Window-Building Classifier, Cluster, Association-Attribute Selection-Data Visualization. Excel: Statistical Capabilities-Average, Mean, Stand Deviation, Median, Graphs Scatter Plot, Bar Graphs.

Unit II: MODELING - Introduction to Scikit learn – Installation basics – fitting and predicting (estimator basics) - Transformers and pre-processors - Pipelines: chaining pre-processors and estimator - Model evaluation - Automatic parameter searches-Tensor Flow Fundamentals- basic computation - Installation of Tensor Flow - Tensors and NumPy - Loading and Preprocessing data - Linear and Logistic regression with Tensor Flow - Training convolutional neural network in Tensor Flow - deploying model.

Unit III: APPLICATION : Overview of NLTK- Tool Installation -Tokenize Words and Sentences-POS Tagging & Chunking- Stemming and Lemmatization-WordNet with NLTK-Introduction about jupyter notebook-Notebook Basics-Running Code Markdown cells-Importing Jupyter Notebook as module connecting to an existing Ipython kernel using Qt Console

Unit IV: VISUALIZATION: Visualization with Matplotlib- Figures and Subplots- Colors, Line Styles, Ticks, Labels, and Legends - Saving Plots to File - Line Plots, Scatter Plots, Density and Contour Plots, Histograms, Three Dimensional Plotting and Geographic Data with Base map.

Unit V: Visualization with Tableau: Introduction – Adding Data Sources in Tabeau – Creating Data Visualizations – Aggregate Functions, Calculated Fields, and Parameters – Table Calculations – Maps – Advanced Analytics: Trends, Forecasts, Clusters and other Statistical Tools

TEXT BOOKS

1. Aurelian Gerona, "Hands-On Machine Learning with Scikit-Learn and Tensor Flow" O'Reilly, 2017.
2. Bharath Ramsundar, Reza Bosagh Zadeh (2018). "TensorFlow for Deep Learning", O'Reilly, 2018.
3. Statistical Analysis with Excel for Dummies, Joseph Schmuller, John Wiley & Sons, Inc, 2013.
4. Alexander Loth, "Visual Analytics with Tableau", Wiley Publisher, First Edition, 2019.

REFERENCE BOOKS

1. Jake VanderPlas, "Python Data Science Handbook: Essential Tools for Working with Data", O'Reilly, 2017.

Techniques and Tools for Data Science Lab

Detailed Content/ List of Experiments:

1. Excel: Statistical Capabilities-Average, Mean, Stand Deviation, Median, Graphs Scatter Plot, Bar Graphs.
2. Linear and Logistic regression with Tensor Flow
3. Visualization with Matplotlib- Figures and Subplots- Colors, Line Styles, Ticks, Labels, and Legends.
4. Types of charts in tableau, Interactive: visualization in tableau, beautiful visualization in tableau, Tips for More Effective and Engaging
5. Design.

MAJOR
B.Sc Data Science – IV Year
VII Semester Paper: VII B

Data Analysis & Visualization

At the end of the course, the students will be able to:

- **Understand** the basics of data visualization
- **Understand** the importance of data visualization and the design and use of many visual components.
- **Analyse** various visualization structures such as tables, spatial data, time-varying data, tree and network, etc.
- **Apply** basic algorithms in data visualization**.**
- **Understand** the types of transformation the data has undergone to improve the effectiveness of the visualization
- **Explain** characteristics and methods that are needed for the visualization of geospatial data

Syllabus

UNIT 1 Importance of analytics

Importance of analytics and visualization in the era of data abundance. Review of probability, statistics and random processes. Brief introduction to estimation theory.

UNIT 2 Introduction to machine learning

Introduction to machine learning, supervised and unsupervised learning, gradient descent, over fitting, regularization.

UNIT 3 Clustering techniques & Regression:

Clustering techniques: K-means, Gaussian mixture models and expectation-maximization, agglomerative clustering, evaluation of clustering - Rand index, mutual information based scores, Fowlkes-Mallows index. Regression: Linear models, ordinary least squares, ridge regression, LASSO, Gaussian Processes regression.

UNIT 4 Supervised classification methods

Supervised classification methods: K-nearest neighbor, naive Bayes, logistic regression, decision tree, support vector machine. Introduction to artificial neural networks (ANNs), deep NNs, convolutional neural network (CNN).

UNIT 5 Data visualization

Data visualization: Basic principles, categorical and continuous variables.
Exploratory graphical analysis - Creating static graphs, animated visualizations - loops, GIFs and Videos. Data visualization in Python and R, examples.

Text Books References:

1. Hastie, T., Tibshirani, R., Friedman, J. (2009). The elements of statistical learning: data mining, inference and prediction. Springer.
2. Richard O. Duda, Peter E. Hart, and David G. Stork. 2000. Pattern Classification (2nd Edition). Wiley- Interscience, New York, NY, USA.
3. Christopher M. Bishop. 2006. Pattern Recognition and Machine Learning (Information Science and Statistics). Springer-Verlag, Berlin, Heidelberg.

Data Analysis and Visualization Lab:

Course Outcomes (COs):

- At the end of the course, the students will be able to:
- **Use** data analysis tools in the pandas library, Power BI and Tableau.
- **Assess** Load, clean, transform, merge and reshaping of data operation.
- **Apply** pre-processing method to multi-dimensional data, and manipulate time series data.
- **Understand** real world data analysis problems.
- **Design** and Analysis Hierarchical and Topographical Data.
- **Remember** Interactive data plots.

Detailed Content/ List of Experiments:

1. Visualization of Spread sheet Models in Python.
2. Oracle Database Connectivity using Python.
3. Visualization of Semi-Structured Data.
4. Introduction to Tableau/Power BI and Aggregation Methods in Tableau/Power BI.
5. Visual Encodings and Basic Dashboards in Tableau/Power BI.
6. Interactive Plots in Python.
7. Hierarchical and Topographical Data Visualizations in Tableau/Power BI.
8. Calendar Heat maps and Flow Data Visualizations in Python.
9. Time Series Data Visualization in Python.
10. Dashboards, Actions and Story Telling in Tableau/Power BI.

Text book and Reference books:

1. Andy Kirk, Data Visualization A Handbook for Data Driven Design, Sage Publications, 2016

2. Philipp K. Janert, Gnuplot in Action, Understanding Data with Graphs, Manning Publications, 2010.
3. Hastie, T., Tibshirani, R., Friedman, J. (2009). The elements of statistical learning: data mining, inference and prediction. Springer.
4. Richard O. Duda, Peter E. Hart, and David G. Stork. 2000. Pattern Classification (2nd Edition). Wiley- Interscience, New York, NY, USA.
5. Christopher M. Bishop. 2006. Pattern Recognition and Machine Learning (Information Science and Statistics). Springer-Verlag, Berlin, Heidelberg.

MAJOR

B.Sc Data Science – IV Year

VII Semester Paper: VII C

DATA ANALYTICS WITH PYTHON

Objective

This course is designed to introduce learning about the Importance of Data and its importance, Knowing Python fundamentals and Pandas essentials, Learning the Principles of Probability and sampling Methods, Getting knowledge about formulating and testing hypothesis, Learning and analytical comparison's with ANOVA methods, Learning about Performance indicators using ROC methods

Outcome

Upon completion of this course, the students will be able to

- Use of statistical tools and techniques in analyzing the different dimensions of data
- Knowing different functions and packages in Python for data interpretation
- Getting hands on experience in model building using data tools
- Calculating the estimate of variation using ANOVA methods
- Getting out different classifiers with Precision and recall methods

Syllabus

UNIT I

Introduction to Data Analytics: Data and its importance, data analytic and its types, importance of data analytics

Python Fundamentals: Python Language Basics, Jupyter Notebook, Introduction to pandas, Data Structures, Essential Functionality

Central Tendency and Dispersion : Visual Representation of the Data, Measures of Central Tendency, Dispersion

UNIT-II

Introduction to Probability: Classical Probability, Relative Frequency, Sample Space, Events, Types of Probability, conditional Probability, Bayesian Rule, Relative frequency method, Random Variable, Distribution Function, Density Function

Sampling and Sampling Distribution: Random vs Non Random Sampling, Simple random sampling, cluster sampling, concept of sampling distributions, Student's t-test, Chi-square and F-distributions. Central limit theorem and its application, confidence intervals.

UNIT-III

Hypothesis testing: Importance of Hypothesis testing, null and alternative hypotheses, Type-I and Type –II errors, approaches to Hypothesis testing, two sample testing.

UNIT –IV

Analysis of Variance (ANOVA): Introduction to ANOVA, one way ANOVA, two way ANOVA, Post – Hoc test

Regression: Simple Linear Regression, Multiple Linear Regression, Maximum Likelihood Estimation (MLE), Logistic Regression, step-wise methods and algorithms.

UNIT –V

Introduction to ROC Curves: Performance of diagnostic tests, confusion Matrix, true and false positives, precession and recall measures. Roc curves, Area Under the Curve, simple applications and algorithms in machine learning

Reference Books:

1. McKinney, W. (2012). Python for data analysis: Data wrangling with Pandas, NumPy, and IPython. " O'Reilly Media, Inc.".
2. Swaroop, C. H. (2003). A Byte of Python. Python Tutorial.
3. Ken Black, sixth Editing. Business Statistics for Contemporary Decision Making. “John Wiley & Sons, Inc”.
4. Anderson Sweeney Williams (2011). Statistics for Business and Economics. “Cengage Learning”.
5. Douglas C. Montgomery, George C. Runger (2002). Applied Statistics & Probability for Engineering. “John Wiley & Sons, Inc”
6. Jay L. Devore (2011). Probability and Statistics for Engineering and the Sciences. “Cengage Learning”.
7. David W. Hosmer, Stanley Lemeshow (2000). Applied logistic regression (Wiley Series in probability and statistics). “Wiley-Interscience Publication”.
8. Jiawei Han and Micheline Kamber (2006). Data Mining: Concepts and Techniques. “
9. Leonard Kaufman, Peter J. Rousseeuw (1990). Finding Groups in Data: An Introduction to Cluster Analysis. “John Wiley & Sons, Inc”.

DATA ANALYTICS WITH PYTHON Lab

COURSE OBJECTIVES:

- To learn about Exploratory Data Analysis
- To learn about statistics and probability for Data Analytics
- To learn different types of hypothesis testing
- To learn about Linear regression and multiple regression
- To learn different Machine learning Algorithms

List of Experiments:

1. Implement Different types of data and data structures use cases?

2. Perform the following operations using Python on any open-source dataset (e.g., data.csv) i. Import all the required Python Libraries. ii. Locate an open-source data from the web (e.g., https://www.kaggle.com). Provide a clear description of the data and its source (i.e., URL of the web site).
3. Implement Data Cleansing and Data Manipulation Operations using Numpy and pandas?
4. Perform the following operations on any open-source dataset (e.g., data.csv) 1. Provide summary statistics (mean, median, minimum, maximum, standard deviation) for a dataset (age, income etc.) with numeric variables grouped by one of the qualitative (categorical) variable.
5. Build Exploratory Data Analysis on Automobile data?
6. Implement Hypothesis Building using Feature Engineering?
7. Design Different types of plots by using Matplotlib and seaborn in python?

MAJOR
B.Sc -Data Science

Syllabus for Semester VIII

MAJOR
B.Sc Data Science – IV Year
VIII Semester Paper: VIII A

TIME SERIES ANALYSIS AND FORECASTING

Objective

The aim of this course is to introduce Time Series Analysis and forecasting, gaining knowledge on the forecasting models with time series, Regression models with time series and application of time series.

Outcome

Upon completion of this course, the students will be able to

1. Knowledge of basic concepts in time series analysis and forecasting
2. Understanding the use of time series models for forecasting and the limitations of the methods.
3. Ability to criticize and judge time series regression models.
4. Distinguish the ARIMA modelling of stationary and nonstationary time series.
5. Compare with multivariate times series and other methods of applications

Syllabus

UNIT 1 INTRODUCTION OF TIMESERIES ANALYSIS:

Introduction to Time Series and Forecasting -Different types of data-Internal structures of time series Models for time series analysis-Autocorrelation and Partial autocorrelation. Examples of Time series Nature and uses of forecasting-Forecasting Process-Data for forecasting – Resources for forecasting. Practical Component: 1.Time Series Data Cleaning 2.Loading and Handling Times series data 3. Preprocessing Techniques

UNIT 2 STATISTICS BACKGROUND FOR FORECASTING:

Graphical Displays -Time Series Plots - Plotting Smoothed Data - Numerical Description of Time Series Data - Use of Data Transformations and Adjustments- General Approach to Time Series Modeling and Forecasting- Evaluating and Monitoring Forecasting Model Performance.

UNIT 3 TIME SERIES REGRESSION MODEL:

Introduction - Least Squares Estimation in Linear Regression Models - Statistical Inference in Linear Regression- Prediction of New Observations - Model Adequacy Checking -Variable Selection Methods in Regression - Generalized and Weighted Least Squares- Regression Models for General Time Series Data- Exponential Smoothing-First order and Second order.

Practical Component: 1. Moving Average time analysis data. 2. Smoothing the Time analysis Data. 3. Check out the Time series Linear and non-linear trends. 4. Create a modelling .

UNIT 4 AUTOREGRESSIVE INTEGRATED MOVING AVERAGE (ARIMA) MODELS:

Autoregressive Moving Average (ARMA) Models - Stationarity and Inevitability of ARMA Models - Checking for Stationarity using Variogram- Detecting Non stationarity - Autoregressive Integrated Moving Average (ARIMA) Models - Forecasting using ARIMA - Seasonal Data - Seasonal ARIMA Models Forecasting using Seasonal ARIMA Models Introduction - Finding the "BEST" Model -Example: Internet Users Data- Model Selection Criteria - Impulse Response Function to Study the Differences in Models - Comparing Impulse Response Functions for Competing Models .

Practical Component: 1.Modelling time series • moving average • Exponential smoothing • ARIMA 2. Seasonal autoregressive integrated moving average model (SARIMA)

UNIT 5 MULTIVARIATE TIME SERIES MODELS AND FORECASTING:

Multivariate Time Series Models and Forecasting - Multivariate Stationary Process- Vector ARIMA Models - Vector AR (VAR) Models - Neural Networks and Forecasting -Spectral Analysis - Bayesian Methods in Forecasting.

Practical Component: Dependence Techniques • Multivariate Analysis of Variance and Covariance • Canonical Correlation Analysis • Structural Equation Modeling.

Textbooks:

1. Introduction To Time Series Analysis And Forecasting, 2nd Edition, Wiley Series In Probability And Statistics, By Douglas C. Montgomery, Cheryl L. Jen(2015)
2. Master Time Series Data Processing, Visualization, And Modeling Using Python Dr. Avishek Pal Dr. Pks Prakash (2017)
3. Time Series Analysis And Forecasting By Example Søren Bisgaard Murat Kulahci Technical University Of Denmark Copyright © 2011 By John Wiley & Sons, Inc. All Rights Reserved.

References:

1. Peter J. Brockwell Richard A. Davis Introduction To Time Series And Forecasting Third Edition.(2016),
2. Multivariate Time Series Analysis and Applications William W.S. Wei Department of Statistical Science Temple University, Philadelphia, PA, SA This edition first published 2019 John Wiley & Sons Ltd.

MAJOR
B.Sc Data Science – IV Year
VIII Semester Paper: VIII B

DATA ANALYTICS: DESCRIPTIVE, PREDICTIVE AND PRESCRIPTIVE

Objective

This course is designed for undergraduate engineering students to apply computer science knowledge on the raw data in building business model for taking decision more effectively to automate and visualize it. To introduce basic concepts of business analytics and descriptive statistics. Discover best practices of data visualization for different types of data. To determine the similarities in the data and to find existing patterns. To predict trends in data and build business decisions. Explore spread sheet model to analyze the data.

Outcome

1. Learn the basic concepts of business analytics and descriptive statistics.
2. Discover best practices of data visualization for different types of data.
3. Acquire knowledge to determine the similarities in the data and to find existing patterns.
4. Able to predict trends in data and build business decisions.
5. Able to explore spread sheet model to analyze the data.

Syllabus

UNIT 1 Descriptive Statistics

Introduction: Decision Making, Business Analytics Defined, A Categorization of Analytical Methods and Models, Big Data, Business Analytics in Practice, Legal and Ethical Issues in The Use of Data and Analytics. Descriptive Statistics: Overview of Using Data: Definitions and Goals, Types of Data, Modifying Data in Excel, Creating Distributions from Data, Measures of Location, Measures of Variability, Analyzing Distributions, Measures of Association Between Two Variables.

UNIT 2 Data Visualization and Probability

Probability- An Introduction to Modelling Uncertainty: Events and Probabilities, Some Basic Relationships of Probability, Conditional Probability, Random Variables, Discrete Probability Distributions, Continuous Probability Distributions. Data Visualization: Overview of Data Visualization, Tables, Charts, Advanced Data Visualization, Data Dashboards.

UNIT 3 Descriptive Data Mining & Linear Regression

Descriptive Data Mining: Cluster Analysis, Association Rules, Text Mining. Linear Regression: Simple Linear Regression Model, the Multiple Regression Model, Model Fitting, Big Data and Regression, Prediction with Regression.

UNIT 4 Predictive Data Mining & Spreadsheet Models:

Predictive Data Mining: Data Sampling, Preparation, And Partitioning, Performance Measures, Logistic Regression, K-Nearest Neighbors, Classification and Regression Trees. Spreadsheet Models: Building Good Spreadsheet Models, Predictive and Prescriptive Spreadsheet Model.

UNIT 5 Decision Analysis:

Decision Analysis: Problem Formulation, Decision Analysis without Probabilities and With Probabilities, Decision Analysis with Sample Information, Computing Branch Probabilities with Bayes' Theorem.

Textbooks:

1. Business Analytics, Fourth Edition Jeffrey D. Camm, James J. Cochran, Michael J. Fry, Jeffrey W. Ohlmann.

MAJOR
B.Sc Data Science – IV Year
VIII Semester Paper: VIII C
Numerical Methods

Learning Outcomes:
Students after successful completion of the course will be able to

1. understand the subject of various numerical methods that are used to obtain approximate solutions
2. Understand various finite difference concepts and interpolation methods.
3. Work out numerical differentiation and integration whenever and wherever routine methods are not applicable.
4. Find numerical solutions of ordinary differential equations by using various numerical methods.
5. Analyze and evaluate the accuracy of numerical methods.

II. Syllabus

Unit – 1: Finite Differences and Interpolation with Equal intervals (15h)
Introduction, Forward differences, Backward differences, Central Differences, Symbolic relations, nth Differences of Some functions, Advancing Difference formula, Differences of Factorial Polynomial, Summation of Series. Newton's formulae for interpolation. Central Difference Interpolation Formulae.

Unit – 2: Interpolation with Equal and Unequal intervals **(**15h)
Gauss's Forward interpolation formulae, Gauss's backward interpolation formulae, Stirling's formula, Bessel's formula. Interpolation with unevenly spaced points, divided differences and properties, Newton's divided differences formula. Lagrange's interpolation formula, Lagrange's Inverse interpolation formula.

Unit – 3: Numerical Differentiation (15h)
Derivatives using Newton's forward difference formula, Newton's back ward difference formula, Derivatives using central difference formula, Stirling's interpolation formula, Newton's divided difference formula, Maximum and minimum values of a tabulated function.

Unit – 4: Numerical Integration (15h)
General quadrature formula one errors, Trapezoidal rule, Simpson's1/3– rule, Simpson's 3/8 – rule, and Weddle's rules, Euler – McLaurin Formula of summation and quadrature, The Euler transformation.

Unit – 5: Numerical solution of ordinary differential equations (15h)

Introduction, Solution by Taylor's Series, Picard's method of successive approximations, Euler's method, Modified Euler's method, Runge – Kutta methods.

III. References:

1. S.S.Sastry, Introductory Methods of Numerical Analysis, Prentice Hall of India Pvt. Ltd., New Delhi-110001, 2006.
2. P.Kandasamy, K.Thilagavathy, Calculus of Finite Differences and Numerical Analysis. Chand & Company, Pvt. Ltd., Ram Nagar, New Delhi-110055.
3. R.Gupta, Numerical Analysis, Laxmi Publications (P) Ltd., New Delhi.
4. H.C Saxena, Finite Differences and Numerical Analysis, S. Chand & Company Pvt. Ltd., Ram Nagar, New Delhi-110055.
5. S.Ranganatham, Dr.M.V.S.S.N.Prasad, Dr.V.Ramesh Babu, Numerical Analysis, Chand & Company Pvt. Ltd., Ram Nagar, New Delhi-110055.
6. Web resources suggested by the teacher and college librarian including reading material.

Prepared by:

S. Rama Devi (PhD)

HOD, Department of Statistics