Big Data Analytics - COURSE STRUCTURE
wef June 2021

St Xavier's College (Autonomous), Ahmedabad
Department of Big Data Analytics
M.Sc. Big Data Analytics
A.Y (2021-2023)

Pre-requisites for Course Work

1. **Microsoft Excel for Data Analysis**
    a. Excel Tables, Filters, Sorting
    b. Pivot Tables and Charts
    c. Formats, Formulas, Dates
    d. Functions – Mathematical, Statistical, Text, Date

Reference:

On-line courses/Tutorials:

i. Microsoft Virtual Academy:
    a. Analyzing and Visualizing Data with Excel https://mva.microsoft.com/en-US/training-courses/analyzing-and-visualizing-data-with-excel-11157
    b. Data Analysis with Excel https://mva.microsoft.com/en-US/training-courses/data-analysis-with-excel-16654

ii. Edx.Org:
    a. Introduction to Data Analysis using Excelhttps://www.edx.org/course/introduction-to-data-analysis-using-excel-0

iii. Coursera.org:
    a. Introduction to Data Analysis Using Excel https://www.coursera.org/learn/excel-data-analysis

2. **Basic Unix Programming**
    a. Basic Unix Commands
    b. Handling files and folders
    c. Concatenation, find and replace, modify file & texts
    d. Basic summary commands

Reference:

On-line courses/Tutorials:

i. Data Camp:

a. Introduction to Shell for Data Science
https://www.datacamp.com/courses/introduction-to-shell-for-data-science

ii. Linux.Org:
a. Linux Beginner Tutorials
https://www.linux.org/forums/linux-beginner-tutorials.123/
b. Github - Organizing with Unix:
https://rafalab.github.io/dsbook/organizing-with-unix.html

Book:

i. Data Science at the Command Line, Jeroen Janssens,
https://www.datascienceatthecommandline.com/

Program Specific Outcome : Prepare students to understand and apply different tools and techniques of big data analytics through mathematical, statistical and machine learning approaches. Semester 4 of the program is internship based project in which each student will apply their knowledge gained in the program in the real life data which also help them to understand the current trends in industry.

SEMESTER – I

CORE Paper: STATISTICAL METHODS
Course Code: PBD-1801
No. of Credits: 05
Learning Hours: 75 hrs
Practical's to be conducted using R

Course Outcomes

- CO1 :Understand data pre-processing and data cleaning
- CO2 : Identify the suitable descriptive measures to explore the data.
- CO3 :Learn the analysis of attributes and Chi-square tests for categorical data
- CO4: Understand timeseries data and its components and learn to do exploratory analysis
- CO5: Apply basic statistical methods in real data using R

a) Data Collection & Visualization

Concepts of measurement, scales of measurement, design of data collection formats with illustration, data quality and issues with date collection systems with examples from business, cleaning and treatment of missing data.

b) Basic Statistics

Frequency table, histogram, measures of location, measures of spread, skewness, kurtosis, percentiles, box plot, correlation and simple linear regression.

c) Contingency Tables:

Two way contingency tables, measures of association, testing for dependence.

d) **Introduction to Time Series**

Components of time series, decomposition of time series data, Smoothing auto correlation, stationarity

SUGGESTED BOOKS:

1. Statistics: David Freedman, Pobert Pisani & Roger Purves, WW.Norten & Co. 4th Edition 2007.
2. The visual display of Quantitative Information: Edward Tufte, Graphics Press, 2001.
3. Best Practices in Data Cleaning: Jason W. Osborne, Sage Publications 2012
4. Time Series Analysis and Its Applications: Robert H. Shumway and David S. Stoffer, Springer 2010.

CORE Paper: PROBABILITY & STOCHASTIC PROCESS
Course Code: PBD-1802
No. of Credits: 05
Learning Hours: 75 hrs
Practical's to be conducted using R

- CO1 :Apply basic ideas of probability and probability distributions in real life situation
- CO 2 : Apply the concept of stochastic process in different sectors like brand switching in Marketing Analytics .
- CO 3: Understand basic models in time series data
- CO 4: Apply time series data analysis through R

a) Basic Probability

Concepts of experiments, Outcomes, Sample space, Events, Combinatorial probability, Birthday paradox, Principle of inclusion & exclusion, Conditional probability, Independence, Bayes Theorem.

b) Probability Distribution

Random Variables: discrete and continuous probability models, some probability distributions: Binomial, Poisson, Geometric, Hypergeometric, Normal, exponential.

c) Stochastic Process

Markov Chains, Classification of states, Stationery distribution, limit theorems, Poisson process, illustrations and applications.

d) Time series models

Concepts of AR, MA, ARMA & ARIMA models with illustrations.

SUGGESTED BOOKS:

1. A First Course in Probability: Shelden M. Ross, 2014.

2. Introduction to Stochastics Process: Paul G. Hoel, Sydney C. Port & Charles J. Stone, Waveland Press, 1987.

3. Time Series Analysis and Its Applications: Robert H. Shumway and David S. Stoffer, Springer 2010.

CORE Paper: LINEAR ALGEBRA & LINEAR PROGRAMMING
Course Code: PBD-1803
No. of Credits: 05
Learning Hours: 75 hrs
Practical's to be conducted using R

Course Outcomes

- CO 1: Student will be able to perform matrix operations and employ fundamental concepts of matrix theory.
- CO 2: Students will be able to employ linear algebra to solve some scientific problems.
- CO 3: Student will be able to use fundamental concepts like system of simultaneous linear equations, eigenvalues and eigenvectors in some applicable concepts.
  CO 4: Student will be able to formulate and model linear programming problems.
- CO 5: Student will be able to solve real life problems using linear programming problems and interpret solution of linear programming problems.

Course Overview & Course Objectives

Unit A: Linear Algebra
Unit B: Linear Programming

a) Linear Algebra

Linear equations and matrices, matrix operations, solving system of linear equations, Gauss-Jordan method, Concept & Computation of determinant and inverse of matrix, Eigen values and Eigen vectors, Illustrations of the methods, Positive semi definite and position definite matrices, illustrations

b) Linear Programming

Definition of the problem, convex sets, corner points, feasibility, basic feasible solutions, Simplex method

SUGGESTED BOOKS:

i. Linear Algebra and Its Application: Gilbert Strang, 4th Edition, Academic Press.

ii. Hands-On Matrix Algebra Using R (Active and Motivated Learning with Applications), Hrishikesh D Vinod, World Scientific

iii. Linear Programming: G. Hadley, Addison-Wesley.

CORE Paper: COMPUTING FOR DATA SCIENCES
Course Code: PBD-1804
No. of Credits: 05
Learning Hours: 75 hrs
Practical's to be conducted using Java and R

- CO 1: Understand data structures
- CO 2: Learn the concepts of data science using Java
- CO 3: Learn how to do installation of R and application of R in big data.
- CO 4: Learn as programming language for application to compute very large data.
- CO5: Do algorithm from numerical analysis like Newton-Raphson, Steepest ascent method etc.
- CO6: Learn Monte-Carlo method, which is great methodology for computing methodologies.
- CO7: Understand how to handle strategies for big data.

a) Core Java Concepts

Introduction to Java programming, Object-oriented programming concepts, Interface, Exception Handling, Packages, Threads

b) Data Structure & Concepts of Computation Using Java

Algorithms, Convergence, Complexity with illustrations, some sorting & searching algorithms, some numerical methods e.g. Newton-Raphson, Steepest ascent using Java

c) Computing Methodologies Using R

Monte-Carlo simulations of random numbers and various statistical methods, memory handling strategies for big data.

SUGGESTED BOOKS:

1. Introduction to Data Science (Data Analysis and Prediction Algorithms with R), Rafael A. Irizarry, https://rafalab.github.io/dsbook/

2. Hands-On Programming with R - Write Your Own Functions and Simulations, Grolemund Garrett, O'Reilly

3. Data Structures and Algorithm using Java, 6th Ed. Michael T. Goodrich and Roberto Tamassia, John Wiley & Sons, Inc

CORE Paper : DATABASE MANAGEMENT AND DATA MINING
Course Code: PBD-1805
No. of Credits: 05
Learning Hours: 75 hrs

- CO 1: Learn basic data models and Hadoop Ecosystem
- CO 2: Understand few relational and non-relational databases
- CO 3: Explore hands on experience on Oracle/MySql
- CO 4: Implementation of ORACLE SQL/MS SQL/MySQL.

a) Basic Concepts

Different data models, ER and EER diagram, schema, table, Big Data Concepts and Hadoop Ecosystem

b) Relational and Non-Relational Databases

Structure, various operations, normalization, SQL, No-SQL, Graph Database, Parallel and distributed database, Map-Reduce.
Lab using SQL/Oracle/MySql for Relational databases;
Hadoop(any), MangoDB, GraphDB for Big Data

c) Implementation

ORACLE SQL/MS SQL/MySQL, Hadoop Ecosystem, Concept of database security.

d) Introduction to data mining

Knowledge discovery from databases, Data Mining Functionalities

SUGGESTED BOOKS

1. Database system concepts: Abraham Silberschartz, Henry F. Korth and S. Surarshan, McGraw Hill, 2011.

2. Hadoop 2 Quick-Start Guide: Learn the Essentials of Big Data Computing in the Apache Hadoop 2 Ecosystem, Douglas Eadline, Addison-Wesley, Pearson Education India; First edition (1 March 2016)

3. Data Science and Big Data Analytics: Discovering, Analyzing, Visualizing and Presenting Data, EMC Education Services, 2015

4. Data Mining: Concepts and Techniques: Jain Pei, Jiawei Han, Micheline Kamber , 3rd Edition,2012

CORE Paper : Python Programming
Course Code: PBD-1806
No. of Credits: 05
Learning Hours: 75 hrs

- CO 1: Write python functions
- CO 2: Understand packages and importing packages
- CO 3: Learn file handling
- CO 4: Develop OO Programming Concepts and get exposure on Exception Handling along with OO programming

a) **Basic Concepts**

Introduction to Python interpreter, Control statements, Data Types

b) **Writing Functions**

Defining a function, calling a function, passing by value or reference, anonymous function

c) **File Handling**

Opening and Closing Files, Reading and Writing Files, Directories in Python

d) **Packages**

What are Packages? Import package

e) **Exception Handling**

Python errors and Built-in exceptions, user defined exceptions, exception handling

f) **OO Programming Concepts**

OOP, class, Inheritance, overloading

SUGGESTED BOOKS

1. Core Python Programming: Dr. R. Nageswara Rao, DreamTech, Second Edition
2. Python for Everybody: Exploring Data in Python 3: Charles Severance
3. Python Cookbook: Recipes for Mastering Python 3: David Beazley, 3rd Edition

SEMESTER – II

CORE Paper : FOUNDATIONS OF DATA SCIENCE (PROGRAMMING FOR BIG DATA)
Course Code: PBD-2801
No. of Credits: 05
Learning Hours: 75 hrs
Practical's to be conducted using GraphDB, Python, R

- CO 1 : Learn basic concept of graph theory and understand algorithm on connectedness, shortest path algorithm and spanning tree of graph.
- CO 2: Learn High dimensional space and understand geometry of large data set.
- CO 3: Random graph is use in industries, so they learn when Giant component emerge in random graph. When the random graph is connected, cycle? Students will establish these.
- CO 4:Learn above Singular value decomposition, which has application in image processing, principal component analysis etc.
- CO 5: Random walk is use to make prediction, students will learn regarding the same.
- CO 6: Learn few algorithm for massive data problems.

Course Overview & Course Objectives

Unit a: Graph Theory
Unit b: High Dimensional Space
Unit c: Random Graphs
Unit d: Singular Value Decomposition
Unit e: Random Walks
Unit f: Algorithm for Massive Data Problems

a) **Graph Theory**
Basic Concepts, Algorithms for connectedness, shortest path, Minimum Sampling Tree, *Lab: Graph Databases, Python Programming*

b) **High Dimensional Space**
Properties, Law of large numbers, Sphere and cube in high dimension, Generating points on the surface of a sphere, Gaussians in High dimension, Random projection, Applications.
Lab: Graph Databases, Python Programming

c) **Random Graphs**
Large graphs, G(n,p) model, Giant Component, Connectivity, Cycles, Non-Uniform models, Applications.
Lab: Graph Databases, Python Programming

d) **Singular Value Decomposition (SVD)**
Best rank k approximation, Power method for computing the SVD, Applications.
Lab: R and Python Programming (Optional: Matlab/Octave)

e) **Random Walks**
Reflection Principle, Long leads, Changes of Sign, Illustrations.
Lab: R and Python Programming (Optional: Matlab/Octave)

f) **Algorithm for Massive Data Problems**
Frequency Moments of data streams, matrix algorithms.
Lab: R and Python Programming (Optional: Spark, Matlab/Octave)

SUGGESTED BOOK:

1. Foundations of Data Science: John Hopcroft & Ravindran Kannan.

CORE Paper : ADVANCE STATISTICAL METHODS
Course Code: PBD-2802
No. of Credits: 05
Learning Hours: 75 hrs
Practical's to be conducted using R

- CO 1: Identify best estimators by applying knowledge on properties of estimators
- CO 2: Analyze and apply statistical inference by showing how hypothesis testing can be developed for situations involving single population and two populations
- CO 3: Apply concepts of the linear models in real life situation
- CO 4: Analyze data through regression methods as a statistical technique
- CO 5: Estimate best line fit and classify binary outcomes.

Course Overview & Course Objectives

Unit a: Estimation
Unit b: Test of Hypotheses
Unit c: Linear Model
Unit d: Regression

a) Estimation:

Unbiasedness, Consistency, UMVUE, Maximum likelihood estimates; EM algorithm.

b) Test of Hypotheses

Two types of errors, test statistic, parametric tests for equality of means & variances.

c) Linear Model

Gauss Markov Model, least square estimators, Analysis of variance.

d) Regression

Multiple linear regression, forward, backward & stepwise regression (practical's only), Logistic Regression.

SUGGESTED BOOKS:

1. Statistical Inference: P. J. Bickel and K. A. Docksum, 2nd Edition, Prentice Hall.

2. Introduction to Linear Regression Analysis: Douglas C. Montgomery

CORE Paper : Introduction to Machine Learning
Course Code: PBD-2803
No. of Credits: 04
Learning Hours: 60 hrs
Practical's to be conducted using Python

- CO 1: Study the basic concepts and techniques of Machine Learning
- CO 2: Learn supervised and unsupervised algorithms
- CO 3: Evaluate the model performance
- CO 4: Learn importance of Ensemble methods
- CO 5: Understand Association rules in Machine Learning
- CO 6: Clustering algorithms and checking goodness of fit of clusters through Silhouette Analysis

Course Overview & Course Objectives

Unit a: Machine Language Overview
Unit b: Applications of Linear Regression and Logistic Regression
Unit c: Preparing data for classification
Unit d: Resampling Methods
Unit e: Classification using Nearest Neighbors
Unit f: Probabilistic Classifiers
Unit g: Evaluating Model Performance
Unit h: Decision Tree Classification
Unit i : Rule Based Classifier
Unit j: Ensemble methods
Unit k: Clustering
Unit l: Association

a) **Machine Language Overview**: Applications of Machine Learning Algorithms, Steps involved in Machine Learning, Types of machine learning.

b) Practical Applications of Linear Regression and Logistic Regression

c) **Preparing data for classification:** removing outliers, handling missing data, normalizing the data, dimensionality reduction, handling skewed data, SMOTE technique to be implemented practically, using large datasets

d) **Resampling Methods:** cross-validation, the Bootstrap, percentage split

e) **Classification using Nearest Neighbors:** k-NN algorithm

f) **Probabilistic Classifiers:** generative (Naïve Bayes) and conditional(Logistic) classifiers

g) **Evaluating Model Performance:** different performance evaluation metrics

h) **Decision Tree Classification:**

Entropy, Gini index, algorithm, regression tree, tree pruning

i) **Rule Based Classifier**:

Separate and conquer, rules from decision tree

j) **Ensemble Methods**:

Bagging, Random Forest and Boosting

k) **Clustering:**

k-means**,** hard versus soft clustering, Expectation-Maximization, elbow method, silhouette plots,

l) **Association**

Market Basket Analysis and Apriori Algorithm

SUGGESTED BOOKS:

1. Machine Learning: Tom Mitchell
2. Pattern Recognition and Machine Learning: Christopher Bishop, Springer,2006
3. An Introduction to Statistical Learning: Gareth James, Daniela Witten, Trevor Hastie, Robert Tibshirani, Springer,2015
4. Python Machine Learning: Sebastian Raschka,2015

CORE Paper : Enabling Technologies for Data Science I
Course Code: PBD-2804
No. of Credits: 04
Learning Hours: 60 hrs

Course Outcomes

- CO 1: Learn the concept of various big data platforms like Hadoop ecosystem and its major components.
- CO 2 :Learn NoSQL database
- CO3: Learn workflow scheduler tool Oozie in Hadoop environment.

Course overview and Course Objective

Unit a) Big Data and Hadoop
Unit b) Map-Reduce
Unit c) HIVE
Unit d) SQOOP
Unit e) PIG
Unit f) NoSQL database
Unit g) Oozie

a) **Big data and Hadoop**:
Hadoop architecture, Single node & Multi-node Hadoop, Hadoop commands, Hadoop daemon, Task instance, Hadoop ecosystem and its installation, Illustrations.

b) **Map-Reduce**:
Framework, Developing Map-Reduce program, Life cycle method, Serialization, Running Map-Reduce in local and pseudo-distributed mode, Illustrations.

c) **HIVE**:
Data types and commands, Illustrations.

d) **SQOOP**:
Importing data, exporting data, Running, Illustrations

e) **PIG:**
Schema, Commands, Illustrations.

f) **NoSQL database:**
Features, Types, NoSQL vs. SQL, Advantages and Disadvantages

g) **Oozie:**
What is Oozie? Workflow, packaging and deploying an Oozie workflow application, Features.

SUGGESTED BOOKS

1. Hadoop The Definitive Guide : Tom White , 4th Edition, 2017

2. Hadoop in Action : Chuck Lam, 2010

3. Data-intensive Text Processing with Map Reduce : Jimmy Lin and Chris Dyer, Morgan & Claypool Publishers, 2010

CORE Paper : VALUE THINKING
Course Code: PBD-2805
No. of Credits: 02
Learning Hours: 30 hrs

Course Outcomes

- CO 1: Enhance logical thinking, argumentative logic, evidence gathering, and drawing inference from evidences.
- CO 2: Get more awareness of the factors like deep rooted prejudices, pre-conceived ideas, psychological and sociological influences that sub-consciously come into play in decision making and forming impressions.

This course involves watching few movies (list provided below) and reading few books (list provided below) that deals mostly with argumentative logic, evidence, drawing inference from evidences. After watching the movies and reading the books, there will be general discussion amongst the students. Couple of case studies that involve mostly logical thinking will also be presented. Each student will prepare a term paper. Evaluation will be on the basis of this term paper and participation in group discussion.

Movies:

1. Twelve Angry Men
2. Roshoman by Kurosawa
3. Trial of Nuremberg
4. Mahabharata by Peter Brook
5. Jurassic Park
6. Interstellar

Books:

1. The Hound of the Baskervilles by Arthur Conan Doyle
2. Five Little Pigs by Agatha Christie
3. The Purloined Letter by Edger Allan Poe
4. The Case of the Substitute Face
5. Caves of Steel by Issac Assimov
6. Fahrenheit 451 by Ray Bradbury

Case Studies:

Elective course : OPERATIONS RESEARCH

Course Code: PBD-2950
No. of Credits: 05
Learning Hours: 75 hrs
Practical's to be conducted using R/AMPL

Course Outcomes

- CO 1: Student will be able to understand the significance of sensitivity analysis and perform the same on various parameters in an LP model.
- CO 2: Student will be able to formulate and model assignment and transportation problems.
- CO 3: Student will be able to solve real life problems using assignment and transportation problems and interpret solution of assignment and transportation problems
  CO 4: Student will be able to formulate and model non-linear programming problems.
- CO 5: Student will be able to solve real life problems using non-linear programming problems and interpret solution of non-linear programming problems.

- Course Overview & Course Objectives

  Unit a: Sensitivity Analysis in Linear Programming
  Unit b: Assignment Models.
  Unit c: Transportation Models
  Unit d: Non-Linear Programming.

a) Sensitivity Analysis in Linear Programming: Change in Objective function coefficient, availability of resources and Input-output coefficient

b) Assignment Models.: Introduction, Mathematical models of Assignment problems, Solution methods; Hungarian Method , Variations of Assignment Problem; Multiple Optimal Solution, Maximization case, Unbalanced assignment problems and restrictions on Assignment

c) Transportation Models.: Introduction, Mathematical model of Transportation problem, Methods for finding Initial solution (Vogel's approximation method), Test for optimality, Variations in Transportation Problem; Unbalanced supply and demand, Degeneracy and its resolution, Alternate Optimal Solutions, Prohibited Transportation Routes.

d) Non-Linear Programming.: Introduction, General non-linear programming problem, Quadratic Programming, Applications of Quadratic Programming.

SUGGESTED BOOK:

1. Operations Research: Prem Kumar Gupta & D. S. Hira

SEMESTER – III

CORE Paper : ENABLING TECHNOLOGIES FOR DATA SCIENCE-II
Course Code: PBD-3801
No. of Credits: 04
Learning Hours: 60 hrs

- CO 1: Get awareness about big data platforms (Spark, Scala, Mahout)
- CO2 : Explore clustering algorithms in Mahout
- CO3: Train, evaluate and deploy classification models

Course Overview & Course Objectives

Unit a) Spark
Unit b) Scala
Unit c) Mahout

a) **Spark:**
Features, Architecture, Components of Spark, Resilient Distributed Datasets – data structure of Spark, working with Key/Value pairs, Loading and Saving data, Core Programming - RDD Transformations, Actions. Executing a Spark Application.

b) Scala:
Features, Basic Syntax, Data types, Variables, Classes and objects, Access modifiers, operators, if construct, loop statement, functions, OOP concepts, Array, String, Exceptions, Collections, File Handling, Multithreading

c) Mahout:
Features, Installation, **Recommendations**: Introducing recommenders, representing recommender data, making recommendations.
Clustering: exploring distance measure, data representation, clustering algorithms in Mahout, evaluating and improving clustering quality
Classification: training a classifier, evaluating and tuning a classifier, deploying a classifier

SUGGESTED BOOKS:

1. Learning Spark: Holden Karau, Andy Konwinski, Patrick Wendell, and Matei Zaharia, OReilly

2. Programming in Scala: Martin Odersky, Lex Spoon, Bill Venners,2008

3. Mahout in Action: Sean Owen, Robin Anil, Ted Dunning, Ellen Friedman,2012

CORE Paper : Advanced Machine Learning
Course Code: PBD-3802
No. of Credits: 04
Learning Hours: 60 hrs
Practical's to be conducted in Python

Course Outcomes

- CO 1: Learn and apply Support Vector Machine and Neural networks to real life data
- CO2: Learn and apply deep learning algorithms
- CO3 : Understand Recommender system
- CO 4: Learn and apply probabilistic graphical models like Bayesian networks to study conditional independence and inference to estimate probabilities for occurrence of events.
- CO 5: Applications of Computer Vision and Natural Language processing
- CO 6 : To understand the need of MLOps

Course Overview and Course Objectives

Unit a) Support Vector Machines
Unit b) Neural Networks
Unit c) Deep Learning
Unit d) Genetic Algorithm
Unit e) Recommender System
Unit f) Graphical Models
Unit g) Computer Vision and Natural Language processing
Unit h) MLOps

a) **Support Vector Machines**

Model, Large Margin Classification, Kernels, SVMs in practice.

b) **Neural Networks**

Representation Learning, Different Models like single and multi-layer perceptron, back propagation, Application.

c) **Deep Learning:**

Deep Neural Networks, Deep Nets and Shallow Nets, unsupervised Auto encoders probabilistic deep model, Convolutional Deep Neural Networks, Challenges in Deep Learning Algorithms

d) **Genetic algorithm**

e) **Recommender System**

f) **Graphical Models:**

Bayesian Networks, Conditional independence, Markov Random Fields, Inference in Graphical Models

g) Real time practical **applications for Computer Vision and Natural language processing**

h) MLOps

SUGGESTED BOOKS:

1. Pattern Recognition and Machine Learning: Christopher Bishop, Springer,2006
2. An Introduction to Statistical Learning: Gareth James, Daniela Witten, Trevor Hastie, Robert Tibshirani, Springer,2015
3. Python Machine Learning: Sebastian Raschka,2015

CORE Paper : DATA VISUALIZATION AND MODELING IN MANAGEMENT

Course Code: PBD-3803

No. of Credits: 04

Learning Hours: 60 hrs

- CO 1: Develop data visualization skills
- CO2 : Analyse and build statistical models on banking and other financial institution data through R
- CO3: Analyse health care related data and build statistical models to predict package pricing and classify presence and absence of a particular disease.
- CO4: Predict customer churn in Churn Analytics using different data mining techniques

A. DATA VISUALIZATION

a) Introduction to Data Visualization and Visual Perception

b) Fundamentals of Visualization, Data Modeling and Compare and Contrast

c) Data Visualization Best Practice and Not-So-Best Practices

d) The Use of Color in Data Visualization and Dashboard Design

e) Typography and Data Visualization Design

f) Exam and Infographics

g) Interactive Data Visualization

h) Mapping Data

i) Hands-on practice on Tableau.

B. MODELLING IN MANAGEMENT

a) Banking analytics

b) Healthcare analytics

c) Retail analytics

SUGGESTED BOOKS:

1. Information Dashboard Design: Displaying Data for At-a-glance Monitoring by Stephen Few
2. Learning Tableau 10, Joshua N. Milligan, Packt Publishing
3. Practical Tableau : 100 Tips, Tutorials, and Strategies from a Tableau Zen Master, Ryan Sleeper, O'Reilly
4. Tableau Cookbook – Recipes for Data Visualization, Shweta Sankhe-Savale
5. Tableau for Dummies, Molly Monsey, Paul Sochan
6. Tableau Dashboard Cookbook, Jen Stirrup, Packt Publishing

Elective course : TIME SERIES & FORECASTING

Course Code: PBD-3950
No. of Credits: 04
Learning Hours: 60 hrs
Practical's to be conducted using R/Python

- CO 1: Understand Stationary and Non-stationary Time series
- CO 2: Construct different models using time series analysis
- CO 3: Apply forecasting in Time series analysis
- CO4 : Understand and apply modeling in Seasonal Time series data
- CO 5: Do missing data analysis in time series data

a) Stationary and Non-Stationary Time Series.

b) AR, MA, ARMA, ARIMA models, their properties, estimation of parameters.

c) Tests of Non-Stationarity – Unit Root tests.

d) Forecasting, Smoothing,, Minimum MSE Forecast, Forecast Error

e) Modelling Seasonal Time Series.

f) Missing Data Problem in Time Series.

SUGGESTED BOOKS

1. Introduction to Statistical Time Series: W. A. Fuller

2. Introduction to Time Series Analysis: P. J. Brockwell and R. A. Davis

Elective course : INTRODUCTION TO ECONOMETRICS

Course Code: PBD-3951
No. of Credits: 04
Learning Hours: 60 hrs
Practical's to be conducted using R

Course Outcomes

- CO 1: The students will be able to understand how to undertake empirical research and analysis
- CO 2: The students will be able to appreciate the strengths and weaknesses of various econometric techniques
- CO 3: The student will be able to evaluate competing economic theories and alternative policies
- CO4 : The student will be able to understand the intricacies of various economic variables involved in a big data and learn to model them

Course Overview and Course Objectives

Unit a) Analysis of panel data
Unit b) Generalized Method of Moments
Unit c) Simultaneous Equations System
Unit d) Cointegration
Unit e) Models in Econometrics

a) Introduction to econometric data, Analysis of Panel Data.

b) Generalized Method of Moments (GMM).

c) **Simultaneous Equations System**
Least Squares, Bias Problem, Estimation Method.

d) **Cointegration**
Concept, two variable model, Engle-Granger Method, Vector autoregressions (VAR), Vector error correlation model (VECM).

e) ARCH/GARCH/SV models, some important generalizations like EGARCH & GJR models, ARCH –M models.

SUGGESTED BOOKS:

1. The Econometrics of Financial Markets: J. Campbell, A.Lo and C. Mackinlay
2. Econometric Analysis: William H. Greene
3. Introduction to Econometrics : Jeffrey M. Wooldridge.

SEMESTER – IV

Internship based project
Course Code: PBD-4801
No. of Credits: 20
Duration of Project: 20 weeks

A real life project has to be undertaken at an industry for 20 weeks. Each student will have two supervisors: one from academic institution and one from the industry. The project shall involve handling data extensively and use of methodologies learnt during the course work to derive meaningful inferences. A final project report has to be submitted and an "open" presentation has to be made.

Project evaluation may be as follows.

Report from two supervisors: 200 marks (100 each)
Project report: 100 marks
Grand Viva-Voce : 100 marks
Presentation: 100 marks.
Total: 500 marks