Syllabus

Savitribai Phule Pune University
Fourth Year of Computer Engineering (2015 Course)
410243 : Data Analytics

Teaching Scheme	Credit	Examination Scheme :
TH : 3 Hours/Week	03	In-Sem (Paper) : 30 Marks
		End-Sem (Paper) : 70 Marks

Prerequisite Courses : 310242-Database Management Systems

Companion Course : 410246-Laboratory Practice I

Course Objectives

- To develop problem solving abilities using Mathematics
- To apply algorithmic strategies while solving problems
- To develop time and space efficient algorithms
- To study algorithmic examples in distributed, concurrent and parallel environments

Course Outcomes

On completion of the course, student will be able to–

- Write case studies in Business Analytic and Intelligence using mathematical models.
- Present a survey on applications for Business Analytic and Intelligence.
- Provide problem solutions for multi-core or distributed, concurrent/Parallel environments.

Course Contents

UNIT I : Introduction and Life Cycle (08 Hours)

Introduction: Big data overview, state of the practice in Analytics- BI Vs Data Science, Current Analytical Architecture, drivers of Big Data, Emerging Big Data Ecosystem and new approach. Data Analytic Life Cycle: Overview, phase 1- Discovery, Phase 2- Data preparation, Phase 3- Model Planning, Phase 4- Model Building, Phase 5- Communicate Results, Phase 6- Opearationalize. Case Study: GINA **(Refer Chapter 1)**

UNIT II : Basic Data Analytic Methods (08 Hours)

Statistical Methods for Evaluation- Hypothesis testing, difference of means, wilcoxon rank–sum test, type 1 type 2 errors, power and sample size, ANNOVA. Advanced Analytical Theory and Methods: Clustering- Overview, K means- Use cases, Overview of methods, determining number of clusters, diagnostics, reasons to choose and cautions. **(Refer Chapter 2)**

(08 Hours)

UNIT III : Association Rules and Regression

Advanced Analytical Theory and Methods: Association Rules- Overview, a-priori algorithm, evaluation of candidate rules, case study-transactions in grocery store, validation and testing, diagnostics. Regression- linear, logistics, reasons to choose and cautions, additional regression models. **(Refer Chapter 3)**

(08 Hours)

UNIT IV : Classification

Decision trees- Overview, general algorithm, decision tree algorithm, evaluating a decision tree. Naïve Bayes – Bayes' Algorithm, Naïve Bayes' Classifier, smoothing, diagnostics. Diagnostics of classifiers, additional classification methods. **(Refer Chapter 4)**

(08 Hours)

UNIT V : Big Data Visualization

Introduction to Data visualization, Challenges to Big data visualization, Conventional data visualization tools, Techniques for visual data representations, Types of data visualization, Visualizing Big Data, Tools used in data visualization, Analytical techniques used in Big data visualization. **(Refer Chapter 5)**

(08 Hours)

UNIT VI : Advanced Analytics-Technology and Tools

Analytics for unstructured data- Use cases, Map Reduce, Apache Hadoop. The Hadoop Ecosystem- Pig, HIVE, HBase, Mahout, NoSQL. An Analytics Project-Communicating, operationalizing, creating final deliverables.

(Refer Chapter 6)

❑❑❑

UNIT I

Chapter 1 : Introduction and Life Cycle 1-1 to 1-23

Syllabus :

Introduction : Big data overview, state of the practice in Analytics - BI Vs Data Science, Current Analytical Architecture, drivers of Big Data, Emerging Big Data Ecosystem and new approach.

Data Analytic Life Cycle : Overview, phase 1 - Discovery, Phase 2 - Data preparation, Phase 3 - Model Planning, Phase 4 - Model Building, Phase 5 - Communicate Results, Phase 6 - Opearationalize. Case Study : GINA.

UNIT II

Chapter 2 : Basic Data Analytic Methods 2-1 to 1-17

Syllabus :

Statistical Methods for Evaluation- Hypothesis testing, difference of means, wilcoxon rank–sum test, type 1 type 2 errors, power and sample size, ANNOVA. Advanced Analytical Theory and Methods : Clustering- Overview, K means- Use cases, Overview of methods, determining number of clusters, diagnostics, reasons to choose and cautions.

UNIT III

Syllabus :

Advanced Analytical Theory and Methods : Association Rules- Overview, a-priori algorithm, evaluation of candidate rules, case study-transactions in grocery store, validation and testing, diagnostics. Regression- linear, logistics, reasons to choose and cautions, additional regression models.

UNIT IV

Syllabus :

Decision trees - Overview, general algorithm, decision tree algorithm, evaluating a decision tree. Naïve Bayes -Bayes' Algorithm, Naïve Bayes' Classifier, smoothing, diagnostics. Diagnostics of classifiers, additional classification methods.

UNIT V

Syllabus :

Introduction to Data visualization, Challenges to Big data visualization, Conventional data visualization tools, Techniques for visual data representations, Types of data visualization, Visualizing Big Data, Tools used in data visualization, Analytical techniques used in Big data visualization

UNIT VI

Syllabus :

Analytics for unstructured data - Use cases, MapReduce, Apache Hadoop. The Hadoop Ecosystem - Pig, HIVE, HBase, Mahout, NoSQL. An Analytics Project - Communicating, operationalizing, creating final deliverables.

❑❑❑