DS 707 DATA ANALYTICS
Term I (2014-15)

GENERAL COURSE INFORMATION

Course Name	DS 707 Data Analytics
Instructors	Prof. Chandrashekar R rc@iiitb.ac.in
Course credits	4
Pre-requisite	a) DS 501 Data Management b) Good knowledge of probability and statistics c) Data Modeling (desirable)

COURSE OVERVIEW

Information explosion pervades all spheres of computing. The computational and regulatory need of day-to-day transactional data ranges from a few days to not more than a few months. However, with decreasing costs of data storage, transactional data is being retained for several years now in order to derive additional insights from the transactional data. The process of deriving this additional insight from vast quantities of data is referred to as Data Analytics. This course builds on the knowledge gained in the Data Modeling course by taking a look at some deeper aspects of data warehouses, online analytical processing (OLAP), and data mining.

Outcomes

At the end of the course, the student should be able to:

- Explain the various terminologies associated with Analytics
- Have a good understanding of OLAP techniques
- Learn basic descriptive analytics using spreadsheet tools
- Understand and apply data mining algorithms on datasets using R
- Obtain hands-on experience of the end-to-end data analytics process

COURSE CONTENTS

- Introduction
  - What is data analytics
  - Different approaches to analytics
  - Related areas of data analytics
- Online Analytic Processing (OLAP)
  - Review the major features and functions of OLAP
  - Dr. Codd's OLAP guidelines
  - hypercubes, drill-down and roll-up, and slice-and-dice
  - Examine the different OLAP models (ROLAP, MOLAP, etc.)
  - OLAP implementation by studying the steps and the tools
- Data Mining
  - Introduction to Data Mining concepts
  - Preliminaries
    - Review of concepts from statistics
    - Data Preparation and Data Reduction techniques
    - Introduction to R
  - Classification
    - Review of Naive Bayes classifier
    - Decision Tree/Random forests, over fitting, performance evaluation of classifiers

        - Maximum likelihood
        - Rule-based classifiers
        - Support vector machines (SVM)
    - Cluster Analysis
        - Review of K-means algorithm
        - Hierarchical clustering techniques (agglomeative, divisive)
        - Iso-data clustering
    - Association analysis
        - Review of Apriori and FP Growth;
        - Recent advances in efficient frequent itemset generation approaches
        - Handling categorical and continuous attributes
        - Sequential and infrequent patterns

Class projects discussions and demos

GRADING

Final grade will be based on weights given below:
30%: Mid-Term Exam
15%: Tests / assignments
20%: Project
30%: End-Term Exam
5%: Instructor Discretion (Class participation, etc.)

REFERENCE MATERIAL

- Data Warehousing Fundamentals by Paularj Ponniah
- Data Mining by Pang-Ning Tan, Michael Steinbach, Vipin Kumar
- Han, J., and Kamber, M., Data Mining Concepts and Techniques, Morgan Kaufman Publisher, 2001.
- Richard J. R., and Michael W. G., Data Mining: A Tutorial-Based Primer, Addison Wisley, 2003.
- Recent Literature.

CHEATING AND PLAGIARISM

This course has zero tolerance for cheating and plagiarism. Any violation may result in an F grade and further disciplinary action may be initiated as per the Institute's policies. Ignorance of what constitutes cheating and plagiarism is not an excuse! If you have any doubts, contact your instructor.

DEADLINES

Unless noted otherwise, all deadlines are due at date and time indicated in LMS

LATE POLICY

- 4 - 24 hours late submission: 25% penalty
- 24 - 48 hours late submissions: 50% penalty
- > 48 hours late submissions: 75% penalty

ANNEXURE

What is Plagiarism

Many people think of plagiarism as copying another’s work, or borrowing someone else’s original ideas. But terms like “copying” and “borrowing” can disguise the seriousness of the offense:

According to the *Merriam-Webster OnLine Dictionary*, to “plagiarize” means

1) to steal and pass off (the ideas or words of another) as one's own
2) to use (another's production) without crediting the source
3) to commit literary theft
4) to present as new and original an idea or product derived from an existing source.

In other words, plagiarism is an act of fraud. It involves both stealing someone else’s work and lying about it afterward. But can words and ideas really be stolen?

According to U.S. law, the answer is yes. In the United States and many other countries, the expression of original ideas is considered intellectual property, and is protected by copyright laws, just like original inventions. Almost all forms of expression fall under copyright protection as long as they are recorded in some media (such as a book or a computer file).

All of the following are considered plagiarism:

- turning in someone else’s work as your own
- copying words or ideas from someone else without giving credit
- failing to put a quotation in quotation marks
- giving incorrect information about the source of a quotation
- changing words but copying the sentence structure of a source without giving credit
- copying so many words or ideas from a source that it makes up the majority of your work, whether you give credit or not (see our section on “fair use” rules)

Attention! Changing the words of an original source is not sufficient to prevent plagiarism. If you have retained the essential idea of an original source, and have not cited it, then no matter how drastically you may have altered its context or presentation, you have still plagiarized

Most cases of plagiarism can be avoided, however, by citing sources. Simply acknowledging that certain material has been borrowed, and providing your audience with the information necessary to find that source, is usually enough to prevent plagiarism.