BIRLA INSTITUTE OF TECHNOLOGY & SCIENCE, PILANI

WORK INTEGRATED LEARNING PROGRAMMES

Part A: Content Design

Course Title	Introduction to Data Science
Course No	DSE ZG523
Credit Units	3
Last Revised by	Pravin Y Pawar
Version / Date	1.2 , 15/04/2020

Course Objectives

#	**Course Objectives**
1	Gain basic understanding of the role of Data Science in various scenarios in the real-world of business, industry and government
2	Understand various roles and stages in a Data Science Project and ethical issues to be considered.
3	Explore the processes, tools and technologies for collection and analysis of structured and unstructured data
4	Appreciate the importance of techniques like data visualization, storytelling with data for the effective presentations of the outcomes with the stakeholders

Text Books/References

ID	**Text Book**
T1	Introducing Data Science by Cielen, Meysman and Ali
T2	Storytelling with Data, A data visualization guide for business professionals, by Cole Nussbaumer Knaflic; Wiley
T3	Introduction to Data Mining, by Tan, Steinbach and Vipin Kumar
	References
R1	The Art of Data Science by Roger D Peng and Elizabeth Matsui
R2	Ethics and Data Science by DJ Patil, Hilary Mason, Mike Loukides
R3	Python Data Science Handbook: Essential tools for working with data by Jake VanderPlas
R4	KDD, SEMMA and CRISP-DM: A Parallel Overview , Ana Azevedo and M.F. Santos , IADS-DM, 2008

* The above materials are reference only and are neither conclusive nor exhaustive. However, the student is advised to refer latest content from online sources or instructor supplied materials for more thorough understanding of the topics

Content Structure

1. **Introduction to Data Science**
    1.1. Fundamentals of Data Science
    1.2. Real World applications
    1.3. Data Science vs BI
    1.4. Data Science vs Statistics
    1.5. Roles and responsibilities of a Data Scientist
    1.6. Software Engineering for Data Science
    1.7. Data Scientists Toolbox
    1.8. Data Science Challenges

2. **Data Analytics**
    2.1. Defining Analytics
    2.2. Types of data analytics
        2.2.1. Descriptive, Diagnostic
        2.2.2. Predictive, Prescriptive
    2.3. Data Analytics - methodologies
        2.3.1. CRISP-DM Methodology
        2.3.2. SEMMA
        2.3.3. BIG DATA LIFE CYCLE
        2.3.4. SMAM
    2.4. Analytics Capacity Building
    2.5. Challenges in Data-driven decision making

3. **Data Science Process**
    3.1. Data Science methodology
        3.1.1. Business understanding
        3.1.2. Data Requirements
        3.1.3. Data Acquisition
        3.1.4. Data Understanding
        3.1.5. Data preparation
        3.1.6. Modelling
        3.1.7. Model Evaluation
        3.1.8. Deployment and feedback
    3.2. Case Study
    3.3. Data Science Proposal
        3.3.1. Samples
        3.3.2. Evaluation
        3.3.3. Review Guide

4. **Data Science Teams**
    4.1. Defining Data Team
    4.2. Roles in a Data Science Team
    4.3. Data Scientists
    4.4. Data Engineers
    4.5. Managing Data Team
        4.5.1. Onboarding and evaluating the success of team
        4.5.2. Working with other teams
        4.5.3. Common difficulties

5. Data and Data Models
    5.1. Types of Data and Datasets
    5.2. Data Quality
    5.3. Epicycles of Data Analysis
    5.4. Data Models

    - 5.4.1. Model as expectation
    - 5.4.2. Comparing models to reality
    - 5.4.3. Reactions to Data
    - 5.4.4. Refining our expectations
- 5.5 Six Types of the Questions
- 5.6 Characteristics of Good Question
- 5.7 Formal modeling
    - 5.7.1. General Framework
    - 5.7.2. Associational Analyses
    - 5.7.3. Prediction Analyses

6.Data wrangling and Feature Engineering

- 6.1 Data cleaning
- 6.2 Data Aggregation, Sampling,
- 6.2. Handling Numeric Data
    - 6.2.1 Discretization, Binarization
    - 6.2.2 Normalization
    - 6.2.3 Data Smoothening
- 6.3. Dealing with textual Data
- 6.4. Managing Categorical Attributes
    - 6.4.1 Transforming Categorical to Numerical Values
    - 6.4.2 Encoding techniques
- 6.5 Feature Engineering
    - 6.5.1 Feature Extraction (Dimensionality Reduction)
    - 6.5.2 Feature Construction
    - 6.5.3 Feature Subset selection
        - 6.5.3.1 Filter methods
        - 6.5.3.2 Wrapper methods
        - 6.5.3.3 Embedded methods
    - 6.5.4 Feature Learning
- 6.6 Case Study involving FE tasks

7. Data visualization

- 7.1. Need for visualization
- **7.2.** Exploratory vs Explanatory Analysis
- **7.3.** Tables, Axis based Visualization and Statistical Plots
- **7.4.** The Data Visualization Design Process
- **7.5.** Lessons in Data Visualization Design
- **7.6.** Stories and Dashboards

8. Storytelling with Data

- 8.1. The final deliverable
- 8.2. The Narrative - report / presentation structure
- 8.3. Building narrative with Data
- 8.4. Effective storytelling

9. Ethics for Data Science

- 9.1. Bias and Fairness
    - 9.1.1 Types of Bias
    - 9.1.2 Identifying Bias
    - 9.1.3 Evaluating Bias
- 9.2 Being a data skeptic – examples of misuse of Data
- 9.3 Doing Good Science
- 9.4. Five C’s
- 9.5 Ethical guidelines for Data Scientist
- 9.6 Ethics of data scraping and storage
- 9.7 Case Study: IBM AI Fairness 360

Contents & Session delivery

Session (2 hrs)	Topics to cover	Content Reference
1.	**Introduction to Data Science** • Fundamentals of Data Science • Real World applications • Data Science vs BI • Data Science vs Statistics • Roles and responsibilities of a Data Scientist • Software Engineering for Data Science • Data Scientists Toolbox • Data Science Challenges	T1 – Chapter 1 Class Room Discussion Class Notes Additional Reading (AR) material provided on Canvas LMS Class Room Discussion
2.	**Data Analytics** • Defining Analytics • Types of data analytics ○ Descriptive, Diagnostic ○ Predictive, Prescriptive • Data Analytics – methodologies ○ CRISP-DM Methodology ○ SEMMA ○ BIG DATA LIFE CYCLE ○ SMAM • Analytics Capacity Building • Challenges in Data-driven decision making	Class Notes Beginners Guide to Analytics Data Analytics Tutorial R4 and Class Notes Class Room Discussions
3.		
4.	**Data Science Process** • Data Science methodology ○ Business understanding ○ Data Requirements ○ Data Acquisition ○ Data Understanding ○ Data preparation ○ Modelling ○ Model Evaluation ○ Deployment and feedback • Case Study • Data Science Proposal ○ Samples ○ Evaluation ○ Review Guide	T1 - Chapter 2 R1 – Chapter 2 Class Room Discussion Class Notes and AR
5.		
6.	**Data Science Teams** • Defining Data Team • Roles in a Data Science Team ○ Data Scientists ○ Data Engineers • Managing Data Team ○ On boarding and evaluating the success of team ○ Working with other teams ○ Common difficulties	Class Room Discussion Class Notes and AR

7.	**Data and Data Models** • Types of Data and Datasets • Data Quality • Epicycles of Data Analysis • Data Models o Model as expectation o Comparing models to reality o Reactions to Data o Refining our expectations • Six Types of the Questions • Characteristics of Good Question • Formal modelling o General Framework o Associational Analyses o Prediction Analyses	T3 – Chapter 2 R1 – Chapter 2 R1 – Chapter 5 R1 – Chapter 3 R1 – Chapter 7
8.		
9.	**Data wrangling and Feature Engineering** • Data cleaning • Data Aggregation, Sampling, • Handling Numeric Data o Discretization, Binarization o Normalization o Data Smoothening • Dealing with textual Data • Managing Categorical Attributes o Transforming Categorical to Numerical Values o Encoding techniques • Feature Engineering o Feature Extraction (Dimensionality Reduction) o Feature Construction o Feature Subset selection ▪ Filter methods ▪ Wrapper methods ▪ Embedded methods o Feature Learning • Case Study involving FE tasks	T3 – Chapter 2 T3 – Chapter 2 Class Notes and AR R3 – Chapter 5.4 Class Notes and AR T3 – Appendix B AR Class Room Discussions Class Room Discussions
10.		
11.		
12.	**Data Visualizations** • Data Need for visualization • Exploratory vs Explanatory Analysis • Tables , Axis based Visualization and Statistical Plots • Lessons in Data Visualization Design • The Data Visualization Design Process • Stories and Dashboards	Why Visual Analytics? Visual Analysis for Everyone T2 – Chapter 1 T2 – Chapter 2 T2 – Chapter 8 Class Notes Class Room Discussion
13.		
14.	**Storytelling with Data** • The final deliverable • The Narrative - report / presentation structure • Building narrative with Data	AR Class Room Discussion T2 – Chapter 10

	- Effective storytelling	R1 – Chapter 10
15.	**Ethics for Data Science** - Bias and Fairness - Types of Bias - Identifying Bias - Evaluating Bias - Being a data skeptic – examples of misuse of Data	AR Hidden Biases in Big Data On Being a Data Skeptic
16.	- Doing Good Science - Five C's - Ethical guidelines for Data Scientist - Ethics of data scraping and storage - Case Study: IBM AI Fairness 360	R2 – Chapter 1 R2 – Chapter 3 Class Room Discussion Credit Decisions Medical Expenditures

Evaluation Scheme:
Legend: EC = Evaluation Component; AN = After Noon Session; FN = Fore Noon Session

	Name	Type	Duration	Wt.	Date/Deadline
EC1	Quiz-I (Pre-Mid) Quiz-II (Post-Mid) Assignment	Online (30 mins) Online (30 mins) Take-home	5 days open 5 days open 10 days open	5% 5% 20%	TBA
EC2	Mid-Semester Exam	Closed Book	1.5 hours	30%	Per Schedule
EC3	Comprehensive Exam	Open Book	2.5 hours	40%	Per Schedule

Notes:

- ➔ The release dates of Quiz-1/2 and assignments will be 5 days (for Quiz) and 10 days (for assignments) before the completion/submission deadline.
- ➔ **Deadlines will NOT be extended for whatever reason** and the student is requested not to wait for the deadline to start working on Quiz/Assignment
- ➔ Syllabus for Quiz-I: Sessions: 1 to 3 / Quiz-II : Session 9 to 12
- ➔ Syllabus for Assignment: Hands-on Python-based Exercise (real-world problem, for individual group of 3 / 4 students). Group formation procedure will be announced before Assignment release
- ➔ All Quiz/Assignments will be released and to be answered/submitted in Canvas LMS
- ➔ Syllabus for Mid-Semester Test (Closed Book): Topics in Session Nos. 1 to 8
- ➔ Syllabus for Comprehensive Exam (Open Book): All topics (Session Nos. 1 to 16)
- ➔ The student is strictly advised to stick to regular schedule of Mid-Sem and Compre examinations, and Makeup examinations will be only for those students with business-related absence/health related issues.
- ➔ Strictly NO MAKEUPS for Quiz and Assignments and all submissions after the above stated deadlines will not be considered/evaluated.
- ➔ All students should conform to BITS students' ethical code-of-conduct and all assignments will be subjected to plagiarism check, and if violated will be subject to disciplinary action apart from nullifying all the marks/grades assigned.

Important links and information:

Canvas LMS: All materials/announcements/discussions forums/Online Quizs/Assignment submissions will be via Canvas LMS portal. Students are expected to monitor this portal regularly for any content or announcements.

Contact sessions: Students should attend the online lectures as per the schedule provided in the Course Handout (posted on Canvas LMS)

Evaluation Guidelines:

1. EC-1 consists of 2 Quizzes and 1 Assignments. Students will attempt them through the course pages on the Canvas portal. Announcements will be made on the portal, in a timely manner.
2. For Closed Book tests: No books or reference material of any kind will be permitted.
3. For Open Book exams: Use of books and any printed / written reference material (filed or bound) is permitted. However, loose sheets of paper will not be allowed. Use of calculators is permitted in all exams. Laptops/Mobiles of any kind are not allowed. Exchange of any material is not allowed.
4. If a student is unable to appear for the Regular Test/Exam due to genuine exigencies, the student should follow the procedure to apply for the Make-Up Test/Exam which will be made available on the Elearn portal. The Make-Up Test/Exam will be conducted only at selected exam centers.

It shall be the responsibility of the individual student to be regular in attending the contact-session schedule as given in the course handout, and take all the prescribed evaluation components such as Assignment/Quiz, Mid-Semester Test and Comprehensive Exam according to the evaluation scheme provided in the handout