

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	
Version / Date	

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	Appreciate the application of the concepts of Statistics, Linear Algebra and Graph Theory in Data Science/Machine learning etc.
3	Understand various roles and stages in a Data Science Project and ethical issues to be considered.
4	Explore the processes, tools and technologies for collection and analysis of structured and unstructured data

Text Books/References

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

* 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. Definition
- 1.2. Need of Data Science
- 1.3. Motivating Examples
- 1.4. Roles and responsibilities of a Data Scientist
- 1.5. Data Science vs BI
- 1.6. Data Science vs Statistics
- 1.7. Data Science Applications
- 1.8. Data Science Concerns

2. Data Science Process

- 2.1. Roles in a Data Science project
- 2.2. Setting expectations
- 2.3. Data Science methodology
 - 2.3.1. Business understanding,
 - 2.3.2. Data Requirements,
 - 2.3.3. Data Acquisition,
 - 2.3.4. Data Understanding,
 - 2.3.5. Data preparation,
 - 2.3.6. Modelling,
 - 2.3.7. Model Evaluation,
 - 2.3.8. Deployment and feedback.
- 2.4. Case Study
- 2.5. Data Science Proposal Samples
- 2.6. Data Science Proposal Evaluation
- 2.7. Data Science Proposal Review Guide

3. Data

- 3.1. Data quality
- 3.2. Types of Data
- 3.3. Data Formats
- 3.4. High dimensional data
- 3.5. Data representation
 - 3.5.1. Graphs and networks,
 - 3.5.2. Matrices, Vectors,
 - 3.5.3. Data Frames, list
 - 3.5.4. Libraries of Graph, Matrices and vectors
- 3.6. Data Models
 - 3.6.1. Model as expectation
 - 3.6.2. Comparing models to reality
 - 3.6.3. Reactions to Data
 - 3.6.4. Refining our expectations
- 3.7. Data Sampling
 - 3.7.1. Probability sampling
 - 3.7.2. Non-Probability sampling

4. Data Wrangling

- 4.1. Handling Numeric Data
- 4.2. Dealing with textual Data
- 4.3. Managing Categorical Attributes
- 4.4. Transforming Categorical to Numerical Values
- 4.5. Feature Engineering
- 4.6. Feature Selection

- 4.6.1. Curse of dimensionality
- 4.6.2. Dimensionality Reduction
 - 4.6.2.1. Data Correlation,
 - 4.6.2.2. PCA
- 4.6.3. Nonlinear Featurization

5. Data Analytics

- 5.1. Definitions
- 5.2. Types of data analytics
 - 5.2.1. Predictive, Descriptive, Prescriptive, Diagnostic
- 5.3. Analytics terminology
- 5.4. Data analytics - methodologies
 - 5.4.1. CRISP-DM Methodology
 - 5.4.2. SEMMA
 - 5.4.3. BIG DATA LIFE CYCLE
 - 5.4.4. SMAM
 - 5.4.5. ASUM- DM
- 5.5. Applications

6. Data visualization

- 6.1. Need for visualization
- 6.2. Exploratory vs Explanatory Analysis
- 6.3. Tables , Axis based Visualization and Statistical Plots
- 6.4. The Data Visualization Design Process
- 6.5. Lessons in Data Visualization Design
- 6.6. Stories and Dashboards

7. Ethics for Data Science

- 7.1. Why Data science needs Ethics
- 7.2. History, Concept of informed consent
- 7.3. Being a data sceptic
- 7.4. Ethical guidelines for Data Scientist
- 7.5. Data Science concerns
- 7.6. Data Privacy and Legal aspects
- 7.7. Societal consequences
- 7.8. Ethics of data scraping and storage
- 7.9. Rightful use of data science

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. Review

Contents & Session delivery

Session (2 hrs)	Topics to cover	Content Reference
1.	Introduction to Data Science <ul style="list-style-type: none"> • Definition • Need of Data Science • Motivating Examples • Roles and responsibilities of a Data Scientist • Data Science vs BI • Data Science vs Statistics • Data Science Applications • Data Science Concerns 	T1 – Chapter 1
2.	Data Science Process <ul style="list-style-type: none"> • Roles in a Data Science project • Setting expectations • Data Science methodology <ul style="list-style-type: none"> ○ Business understanding, ○ Data Requirements, ○ Data Acquisition, ○ Data Understanding, ○ Data preparation, ○ Modelling, ○ Model Evaluation, ○ Deployment and feedback. • Case Study • Data Science Proposal Samples • Data Science Proposal Evaluation • Data Science Proposal Review Guide 	T1 - Chapter 2
3.		
4.	Data <ul style="list-style-type: none"> • Data quality • Types of Data • Data Formats • High dimensional data • Data representation <ul style="list-style-type: none"> ○ Graphs and networks, ○ Matrices, Vectors, ○ Data Frames, list ○ Libraries of Graph, Matrices and vectors • Data Models <ul style="list-style-type: none"> ○ Model as expectation ○ Comparing models to reality ○ Reactions to Data ○ Refining our expectations • Data Sampling <ul style="list-style-type: none"> ○ Probability sampling ○ Non-Probability sampling 	T1 - Chapter 1 T3- Chapter 5 https://www.researchgate.net/publication/319998246_Sampling_Methods_in_Research_Methodology_How_to_Choose_a_Sampling_Technique_for_Research/link/59c5f8c2a6fdccc719164f0b/download
5.		
6.	Data Wrangling <ul style="list-style-type: none"> • Handling Numeric Data • Dealing with textual Data • Managing Categorical Attributes • Transforming Categorical to Numerical Values • Feature Engineering • Feature Selection <ul style="list-style-type: none"> ○ Curse of dimensionality ○ Dimensionality Reduction <ul style="list-style-type: none"> • Data Correlation, • PCA ○ Nonlinear Featurization 	R4 - Chapter 1, 5 http://www.feat.engineering Class notes
7.		
8.		

Session(2 hrs)	Topics to cover	Content Reference
9.	Data Analytics <ul style="list-style-type: none"> • Definitions • Types of data analytics <ul style="list-style-type: none"> ○ Predictive, Descriptive, Prescriptive, Diagnostic • Analytics terminology • Data analytics - methodologies <ul style="list-style-type: none"> ○ CRISP-DM Methodology ○ SEMMA ○ BIG DATA LIFE CYCLE ○ SMAM ○ ASUM- DM • Applications 	R2
10.		
11.	Data visualization <ul style="list-style-type: none"> • Need for visualization • Exploratory vs Explanatory Analysis • Tables , Axis based Visualization and Statistical Plots • The Data Visualization Design Process • Lessons in Data Visualization Design • Stories and Dashboards 	T2
12.		
13.	Ethics for Data Science <ul style="list-style-type: none"> • Why Data science needs Ethics • History, Concept of informed consent • Being a data sceptic • Ethical guidelines for Data Scientist • Data Science concerns • Data Privacy and Legal aspects • Societal consequences • Ethics of data scraping and storage • Rightful use of data science 	R1 https://hbr.org/2013/04/the-hidden-biases-in-big-data https://www.oreilly.com/data/free/files/being-a-data-skeptic.pdf
14.		
15.	Storytelling with Data <ul style="list-style-type: none"> • The final deliverable • The Narrative - report / presentation structure • Building narrative with Data • Effective storytelling 	T2
16.	Review	

Evaluation Scheme:

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

No	Name	Type	Duration	Wt.	Date/Deadline*
EC-1A	Quiz-I (Pre-Mid / 20MCQ)	Online	3 days open	5%	
EC-1B	Quiz-II (Post-Mid / 20MCQ)	Online	3 days open	5%	
EC-1C	Assignment	Take-home	3 weeks	10%	
EC-2R	Mid-Semester Regular	Closed Book	1.5 hours	30%	
EC-2M	Mid-Sem Makeup	Closed Book	1.5 hours	30%	
EC-3R	Comprehensive Exam	Open Book	2.5 hours	50%	
EC-2M	Compre Makeup	Open Book	2.5 hours	50%	

Notes:

- The release dates of Quiz-1/2 and assignments will be 3 days (for Quiz) and 3 weeks (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 4 / Quiz-II (all Sessions)
- Syllabus for Assignment: Hands-on Python-based Exercise (real-world problem, for individual group of 3 students) / Group formation 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 Quizzes/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 centres.

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