**BIRLA INSTITUTE OF TECHNOLOGY & SCIENCE, PILANI**

**WORK INTEGRATED LEARNING PROGRAMMES**

**Digital Learning**

**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 |
|  | NussbaumerKnaflic; 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**

3.1. Types of Data and Datasets

3.2. Data Quality

3.3. Epicycles of Data Analysis

3.4. Data Models

3.4.1. Model as expectation

3.4.2. Comparing models to reality

3.4.3. Reactions to Data

3.4.4. Refining our expectations

3.5 Six Types of the Questions

3.6 Characteristics of Good Question

3.7 Formal modeling

3.7.1 General Framework

3.7.2 Associational Analyses

3.7.3 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

1. **Data visualization**
   1. Need for visualization
   2. Exploratory vs Explanatory Analysis
   3. Tables, Axis based Visualization and Statistical Plots
   4. The Data Visualization Design Process
   5. Lessons in Data Visualization Design
   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

* 1. Ethical guidelines for Data Scientist
  2. Ethics of data scraping and storage
  3. Case Study: IBM AI Fairness 360

**Part B: Course Handout**

|  |  |
| --- | --- |
| **Academic Term** | First Semester, 2020-21 |
| **Course Title** | Introduction to DataScience |
| **Course No** | DSEHLZG553 |
| **Lead Instructor** | Prof. Sumita Narang |

**Contents & Session delivery**

|  |  |  |  |
| --- | --- | --- | --- |
| **Session (2 hrs)** | **Topics to cover** | | **Content Reference** |
|  | **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 |
|  | **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](https://www.jigsawacademy.com/em/Beginners_Guide_to_Analytics.pdf)  [Data Analytics Tutorial](https://data-flair.training/blogs/data-analytics-tutorial/)  R4 and Class Notes  Class Room Discussions |
|  |
|  | **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 |
|  |
|  | **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 |
|  | **Data and Data Models**   * Types of Data and Datasets * Data Quality * Epicycles of Data Analysis * Data Models   + Model as expectation   + Comparing models to reality   + Reactions to Data   + Refining our expectations * Six Types of the Questions * Characteristics of Good Question * Formal modelling   + General Framework   + Associational Analyses   + Prediction Analyses | | T3 – Chapter 2  R1 – Chapter 2  R1 – Chapter 5  R1 – Chapter 3  R1 – Chapter 7 |
|  |
|  | **Data wrangling and Feature Engineering**   * Data cleaning * Data Aggregation, Sampling, * Handling Numeric Data   + Discretization, Binarization   + Normalization   + Data Smoothening * Dealing with textual Data * Managing Categorical Attributes   + Transforming Categorical to Numerical Values   + Encoding techniques * Feature Engineering   + Feature Extraction (Dimensionality Reduction)   + Feature Construction   + Feature Subset selection     - Filter methods     - Wrapper methods     - Embedded methods   + 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 |
|  |
|  |
|  | **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?](https://www.tableau.com/sites/default/files/whitepapers/752750_core_why_visual_analytics_whitepaper_0.pdf)  [Visual Analysis for Everyone](http://cdnlarge.tableausoftware.com/sites/default/files/whitepapers/visual_analysis_for-everyone.pdf)  T2 – Chapter 1  T2 – Chapter 2  T2 – Chapter 8  Class Notes  Class Room Discussion | |
|  |
|  | **Storytelling with Data**   * The final deliverable * The Narrative - report / presentation structure * Building narrative with Data * Effective storytelling | AR  Class Room Discussion  T2 – Chapter 10  R1 – Chapter 10 | |
|  | **Ethics for Data Science**   * Bias and Fairness   + Types of Bias   + Identifying Bias   + Evaluating Bias * Being a data skeptic – examples of misuse of Data * Doing Good Science * Five C’s * Ethical guidelines for Data Scientist * Ethics of data scraping and storage * Case Study: IBM AI Fairness 360 | AR  [Hidden Biases in Big Data](https://hbr.org/2013/04/the-hidden-biases-in-big-data)  [On Being a Data Skeptic](https://www.oreilly.com/data/free/files/being-a-data-skeptic.pdf)  R2 – Chapter 1  R2 – Chapter 3  Class Room Discussion  [Credit Decisions](https://nbviewer.jupyter.org/github/IBM/AIF360/blob/master/examples/tutorial_credit_scoring.ipynb)  [Medical Expenditures](https://nbviewer.jupyter.org/github/IBM/AIF360/blob/master/examples/tutorial_medical_expenditure.ipynb) | |
|  |

**Evaluation Scheme**:

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

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| E | Name | Type | Duration | Wt. | Date/Deadline |
| EC1 | Quiz-I (Post-Mid)  Assignment –Phase 1 (Pre-Mid)  Assignment –Phase 2 (Post-Mid) | Online  Take-home  Take-home |  | 5%  7.5%  7.5% | TBA |
| EC2 | Mid-Semester Exam | Closed Book | 1.5 hours | 30% | As Per Semester Schedule |
| EC3 | Comprehensive Exam | Open Book | 2.5 hours | 50% | As Per Semester Schedule |

\*The above evaluation scheme is tentative and likely to change. In case of any change instructor will keep you posted.

**Notes**

* Quiz and Assignments timelines will be announced on the canvas portal.
* **Deadlines for evaluation components will NOT be extended** and the student is requested not to wait for the deadline to start working on Quiz/Assignment
* 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)
* **Strictly NO MAKEUPS for Quiz and Assignments** and all submissions after the announced deadlines will not be considered for evaluation.
* **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: Students are expected to visit the Canvas portal on a regular basis and stay up to date with the latest announcements and deadlines.

Contact sessions: Students should attend the online lectures as per the schedule provided.

Evaluation Guidelines:

1. EC-1 consists of two quizzes and one assignment. 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. Laptops/Mobiles of any kind are not allowed. Exchange of any material is not allowed.
3. For Open Book exams: Use of prescribed and reference text books, in original (not photocopies) is permitted. Class notes/slides as reference material in filed or bound form 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. The genuineness of the reason for absence in the Regular Exam shall be assessed prior to giving permission to appear for the Make-up Exam. 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 maintaining the self-study schedule as given in the course handout, attend the lectures, 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.