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

**WORK INTEGRATED LEARNING PROGRAMMES**

**Digital**

Part A: Content Design

|  |  |
| --- | --- |
| **Course Title** | Introduction to Data Science |
| **Course No(s)** | ZG532 |
| **Credit Units** | 5 |
| **Content Authors** | Ms. Seetha Parameswaran |
| **Version** | 2.0 |
| **Date** | August 5h 2022 |

**Course Objectives**

|  |  |
| --- | --- |
| **No** | Course Objective |
| **CO1** | Gain basic understanding of the role of Data Science in various scenarios in the real-world  of business, industry and government. |
| **CO2** | Understand various roles and stages in a Data Science Project and ethical issues to  be considered. |
| **CO3** | Explore the processes, tools and technologies for collection and analysis of structured  and unstructured data. |
| **CO4** | Appreciate the importance of techniques like data visualization, storytelling with data  for the effective presentations of the outcomes with the stakeholders |
| **CO5** | Understand techniques of preparing real-world data for data analytics. |
| **CO6** | Implement data analytic techniques for discovering interesting patterns from data. |

**Text Book(s)**

|  |  |
| --- | --- |
| T1 | Introduction to Data Mining, by Tan, Steinbach and Vipin Kumar |
| T2 | Introducing Data Science by Cielen, Meysman and Ali |
| T3 | Storytelling with Data, A data visualization guide for business professionals, by Cole Nussbaumer Knaflic; Wiley |
| T4 | Data Mining: Concepts and Techniques, Third Edition by Jiawei Han and Micheline Kamber Morgan Kaufmann Publishers, 2006 |

**Reference Book(s) & other resources**

|  |  |
| --- | --- |
| 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 |

**Content Structure**

1. Introduction to Data Science (2 hrs)
   1. Fundamentals of Data Science
   2. Real World applications
   3. Data Science Challenges
   4. Data Science Teams
      1. Defining Data Team
      2. Roles in a Data Science Team
      3. Roles and responsibilities of a Data Scientist, Data Engineers
   5. Software Engineering for Data Science

1.5.1 DataOps

1.5.2 MLOps (intro)

2. Data Analytics ( 2hrs)

2.1. Defining Analytics

2.2. Types of data analytics

2.2.1. Descriptive, Diagnostic, Predictive, Prescriptive

2.3. Data Science Process

2.3.1. CRISP-DM Methodology

2.3.2. SEMMA

2.3.3. BIG DATA LIFE CYCLE

2.3.4. SMAM

3. Data and Data Models (2 hrs)

3.1. Types of Data and Datasets

3.2. Data Quality and Issues: An overview

3.3. Data Models

3.3.1. General Framework of Formal modeling

3.3.2. Association Analyses

3.3.3. Prediction Analyses

3.4. Data Pipelines and patterns;

3.5. Data Pipeline Stages

3.5.1 Data extraction

3.5.2 Data ingestion

3.5.3 Data cleaning

3.5.4 Data exploration

3.5.5 Data wrangling

3.5.6 Data versioning

3.5.7 Data transformation

3.5.8 Feature management

3.6 Modern Data Infrastructure

3.6.1 Diverse data sources

3.6.2 Cloud data warehouses and lakes

4. Data wrangling (4 hrs)

4.1 Data cleaning

4.2 Data Aggregation, Sampling,

4.3 Statistical descriptions of data

4.4 Measuring data similarity & dissimilarity

4.5. Handling Numeric Data

4.5.1 Discretization, Binarization

4.5.2 Normalization

4.5.3 Data Smoothening

4.6 Dealing with textual Data

4.7 Dealing with Images, audio and video data

4.8. Managing Categorical Attributes

4.8.1 Transforming Categorical to Numerical Values

4.8.2 Encoding techniques

4.9 Overview of visualization techniques for Data Exploratory analysis

5. Feature Engineering (4 hrs)

5.1 Feature Extraction

5.2 Feature Construction

5.3 Feature Subset selection

5.3.1 Filter methods

5.3.2 Wrapper methods

5.3.3 Embedded methods

5.4 Feature Learning

5.5 Feature reduction (Dimensionality Reduction)

5.6 Case Study involving FE tasks

5.7 Feature Engineering techniques for text, images, audio, video

6. Classification and Prediction (4 hrs)

6.1. Concepts of classification and prediction

6.2. Decision trees for classification - ID3 algorithm using entropy and Gini Index

6.3. Evaluation of classification algorithms

7. Association Analysis (4 hrs)

7.1. Association analysis concepts

7.2. Apriori Algorithm for frequent itemsets

7.3 FP Growth for frequent itemsets

7.4. Mining association rules

8. Clustering (4 hrs)

8.1. Cluster analysis concepts.

8.2. Partitioning methods – k-Means algorithm

8.3. Hierarchical methods for cluster analysis

8.4. Density based methods for cluster analysis - DBSCAN

8.5. Evaluation of clustering algorithms

9. Anomaly Detection ( 2 hr)

9.1. Concepts of Outliers

9.2. Statistical approaches

9.3. Proximity and Density based outlier detection

10. Storytelling with Data (2 hr)

10.1. The final deliverable

10.2. The Narrative - report / presentation structure

10.3. Building narrative with Data

10.4. Effective storytelling

11. Ethics for Data Science ( 2 hr)

11.1. Bias and Fairness in Data

11.2 Being a data skeptic – examples of misuse of Data

11.3 Five C’s

11.4 Ethical guidelines for Data Scientist

11.5 Ethics of data scraping and storage

11.6 Case Study: IBM AI Fairness 360

**Part B: Learning Plan**

|  |  |
| --- | --- |
| **Academic Term** | S1/22-23 |
| **Course Title** | Introduction to Data Science |
| **Course No** | DSECLZG532 |

|  |  |  |
| --- | --- | --- |
| **Session No.** | **Topic Title** | **Resource Reference** |
| 1 | Introduction to Data Science (2 hrs)   * Fundamentals of Data Science * Real World applications * Data Science Challenges * Data Science Teams   + Defining Data Team   + Roles in a Data Science Team   + Roles and responsibilities of a Data Scientist, Data Engineers * Software Engineering for Data Science   + DataOps   + MLOps (intro) | T3 – Ch 1  T4 – Ch1  T1 – Ch1  Class Room Discussion  Class Notes  Additional Reading (AR) material provided LMS |
| 2 | Data Analytics ( 2hrs)   * Defining Analytics * Types of data analytics   + Descriptive, Diagnostic, Predictive, Prescriptive * Data Science Process   + CRISP-DM Methodology   + SEMMA   + BIG DATA LIFE CYCLE   + SMAM | Beginners Guide to Analytics  <https://www.jigsawacademy.com/blogs/business-analytics/beginners-guide-to-4-types-of-business-analytics/>  Data Analytics Tutorial  https://data-flair.training/blogs/data-analytics-tutorial/  R4 |
| 3 | Data and Data Models (2 hrs)   * Types of Data and Datasets * Data Quality and Issues: An overview * Data Models   + General Framework of Formal modeling   + Association Analyses   + Prediction Analyses * Data Pipelines and patterns * Data Pipeline Stages * Data extraction * Data ingestion * Data cleaning * Data exploration * Data wrangling * Data versioning * Data transformation * Modern Data Infrastructure   + Diverse data sources   + Cloud data warehouses and lakes | T1 – Ch 2.1, 2.2  R1 – Ch 2, Ch 7  Class room discussions |
| 4 | Data wrangling (2 hrs)   * Data cleaning * Data Aggregation, Sampling, * Statistical descriptions of data * Measuring data similarity & dissimilarity * Handling Numeric Data   + Discretization, Binarization   + Normalization   + Data Smoothening | T1 – Ch2.3, 2.4  T4 – Ch4 |
| 5 | Data wrangling (2 hrs)   * Dealing with textual Data * Dealing with Images, audio and video data * Managing Categorical Attributes   + Transforming Categorical to Numerical Values   + Encoding techniques * Overview of visualization techniques for Data Exploratory analysis | Class room discussions  T1 – Ch3.1 |
| 6 | Feature Engineering (2 hrs)   * Feature Extraction * Feature Construction * Feature Subset selection   + Filter methods   + Wrapper methods   + Embedded methods * Feature reduction (Dimensionality Reduction)   (PS: PCA discussed in depth in MFDS course of DSE programme.) | T1 – Ch2  Class room discussions  T1 – Appendix B.1 |
| 7 | Feature Engineering (2 hrs)   * Case Study involving FE tasks * Feature Engineering techniques for text, images, audio, video | Class room discussions  T4 – Ch10.4 |
| 8 | Classification and Prediction (2 hrs)   * Concepts of classification and prediction * Evaluation of classification algorithms | T4 – Ch6.1, 6.2, 6.3  T4 – 6.12, 6.13, 6.15 |
| 9 | Classification and Prediction (2 hrs)   * Decision trees for classification - ID3 algorithm using entropy and Gini Index, Occam’s raxor   (Mutual Information and Gini Index are used as Feature subset selection techniques. ) | T4 – Ch6.1, 6.2, 6.3 |
| 10 | Association Analysis (2 hrs)   * Association analysis concepts * Apriori Algorithm for frequent itemsets | T1 – Ch6  T4 – Ch5 |
| 11 | Association Analysis (2 hrs)   * FP Growth for frequent itemsets * Mining association rules | T1 – Ch6  T4 – Ch5 |
| 12 | Clustering (2 hrs)   * Cluster analysis concepts. * Partitioning methods – k-Means algorithm * Hierarchical methods for cluster analysis | T1 – Ch8  T4 – Ch7 |
| 13 | Clustering (2 hrs)   * Density based methods for cluster analysis – DBSCAN * Evaluation of clustering algorithms | T1 – Ch8  T4 – Ch7  T1 – Ch8.5 |
| 14 | Anomaly Detection ( 2 hrs)   * Concepts of Outliers * Statistical approaches * Proximity and Density based outlier detection | T1 – Ch10  T4 – Ch7.11 |
| 15 | Storytelling with Data (2 hrs)   * The final deliverable * The Narrative - report / presentation structure * Building narrative with Data * Effective storytelling | T3 – Ch10 |
| 16 | Ethics for Data Science ( 2 hrs)   * Bias and Fairness   + Types of Bias   + Identifying Bias   + Evaluating Bias * Being a data skeptic – examples of misuse of Data * Five C’s * Ethical guidelines for Data Scientist * Ethics of data scraping and storage * Case Study: IBM AI Fairness 360   (PS: Ethics for Data is the focus.) | https://hbr.org/2013/04/the-hidden-biases-in-big-data  https://www.oreilly.com/content/on-being-a-data-skeptic/  T4 – Ch11.4  R2 – Ch1, Ch3 |

**Detailed Plan for Lab work**

|  |  |  |
| --- | --- | --- |
| **Lab No.** | **Lab Objective** | **Session Reference** |
| 1 | Introduction to Python, Numpy, Scipy, Python Pandas, | 2 |
| 2 | Data ingestion and extraction, data aggregation techniques | 3 |
| 3 | Exploration and Visualizing using Matplotlib, Seaborn | 4 |
| 4 | Data pre-processing in Python - Discretization, Binarization, Normalization, Data Smoothening, Managing Categorical Attributes | 5 |
| 5 | Feature Engineering using Filter methods, wrapper methods, PCA | 7 |
| 6 | Data pre-processing and Feature Engineering techniques for text, images, audio, video | 8 |
| 7 | Decision trees for classification using Scikit learn | 9 |
| 8 | Association Analysis using Scikit learn | 11 |
| 9 | Clustering analysis by Kmeans, hierarchical methods, DBScan using Scikit learn | 13 |

**Evaluation Scheme**:

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

|  |  |  |  |
| --- | --- | --- | --- |
| No | Name | Type | Weight |
| EC-1(a) | Quizzes | Online | 10% |
| EC-1(b) | Assignments | Take Home | 20% |
| EC-2 | Mid-Semester Test | Closed Book | 30% |
| EC-3 | Comprehensive Exam | Open Book | 40% |

**Note:**

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)

**Important links and information:**

**CANVAS :** https://bits-pilani.instructure.com/courses/1704

Students are expected to visit 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 on the portal.

Evaluation Guidelines:

1. EC-1 consists of two Quizzes. Students will attempt them through the course pages on CANVAS portal. Announcements will be made on the portal, in a timely manner.
2. EC-2 consists of either one or two Assignments. Students will attempt them through the course pages on CANVAS portal. Announcements will be made on the portal, in a timely manner.
3. For Closed Book tests: No books or reference material of any kind will be permitted.
4. 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.
5. 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 informed.

It shall be the responsibility of the individual student to be regular in maintaining the self-study schedule as given in the course hand-out, attend the online 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 hand-out.