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

**WORK INTEGRATED LEARNING PROGRAMMES**

**COURSE HANDOUT**

**Part A: Content Design**

|  |  |
| --- | --- |
| **Course Title** | Feature Engineering |
| **Course No(s)** |  |
| **Credit Units** |  |
| **Course Author** | Aruna Malapati |
| **Version No** | 3.0 |
| **Date** | 30/08/2018 |
| **Last Updated by** | Pravin Pawar, 8/3/2020 |

**Course Description**

|  |
| --- |
| Feature Engineering is an important step to develop and improve performance of Machine Learning models. This course covers Data wrangling techniques that help transforming the raw data to an appropriate form on which learning algorithms can be applied. The data preprocessing techniques like normalization, discretization, feature subset selection etc. and dimension reduction techniques like PCA will be discussed. The different ways of visualizing the data like Box plots, Contour plots, Heat maps etc. will be illustrated. |

**Course Objectives**

|  |  |
| --- | --- |
| **No** | Objective |
| **CO1** | Apply data wrangling for a given business problem |
| **CO2** | Identify and implement appropriate feature extraction and selection techniques. |
| **CO3** | Apply data preprocessing techniques |
| **CO4** | Implement dimension reduction using PCA |
| **CO5** | Identify and implement appropriate visualization techniques. |

**Text Book(s)**

|  |  |
| --- | --- |
| No | Author(s), Title, Edition, Publishing House |
| T1 | Introduction to Data Mining eBook: Vipin Kumar, Pang-Ning Tan Michael Steinbach |

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

|  |  |
| --- | --- |
| No | Author(s), Title, Edition, Publishing House |
| R1 | Bishop - Pattern Recognition And Machine Learning - Springer 2006 |

**Content Structure**

|  |  |  |
| --- | --- | --- |
| **No** | **Title of the Module** | **References** |
| M1 | Overview of Feature Engineering and Data Preprocessing   1. Introduction to Feature Engineering 2. Types of data and its sources 3. Data quality (Missing values, Noisy data) 4. Aggregation and Sampling 5. Feature Creation 6. Data Transformation | T1 Ch 2 |
| M2 | Proximity measures and Feature Subset Selection   1. Discretization 2. Supervised Discretization using Entropy 3. Binarization 4. Proximity measures for binary attributes 5. Proximity measures for Categorical attributes 6. Proximity measures for Continuous attributes and Nonmetric measures 7. Curse of Dimensionality 8. Feature Subset Selection 9. Feature selection using Filter Methods 10. Feature selection using Chi Squared Test 11. Feature selection using Information Theoretic Measures 12. Feature selection using Fisher Score 13. Feature selection using wrapper methods | T1 Ch 2 |
| M3 | Dimension Reduction   1. Statistics foundations 2. Introduction to Dimension reduction 3. Formulation of Principal Component Analysis (PCA) using Maximum Variance 4. PCA mathematical derivation | R1  Ch 12 Section 12.1 |
| M4 | Visualization   1. Data Visualization and Statistical Analysis 2. Charts and Plots 3. Parallel Coordinates 4. t-SNE Plot 5. Data Visualization Use case and Feature Engineering | T1 Ch 3 |

|  |  |
| --- | --- |
| **Week** | **Content / Assignments / Exercises** |
| Week1 | Video Content: M1  Evaluative Quiz: Nil  Exercises 1:   1. Classify data based on the type of value. 2. Exercises on data quality/Missing Values. 3. Exercises on Data Transformation.   Assignment 1:   1. The numerical / applied problems that assess the students grasping of the data pre-processing concepts. |
| Week 2 | Video Content: M2  Evaluative Quiz: Nil  Exercises 2:   1. Given the description of objects what is the appropriate way of representing the data in the form of features. 2. Identify the appropriate Similarity Measures using Jaccard, Hamming, Manhattan, Euclidean, Mahalanobis, Cosine distances.   Assignment 2:   1. Implementing data pre-processing (aggregation/sampling/ Discretization) using Python off-the-shelf libraries or custom code. |
| Week 3 | Video Content: M3  Evaluative Quiz: Nil  Exercises / Assignment: Nil  Minor Project 1:   1. Python Program to scrape data and get data from various sources like Excel, Flat files, XML, JSON and create a data set. 2. Implement a Python program and apply the Feature Engineering for a problem at hand. |
| Week 4 | Video Content: M4  Minor Project 1: Running from previous week  Evaluative Quiz 1: Q1 on Video Content M1 to M4  Exercises / Assignment: Nil |

Evaluation

|  |  |  |
| --- | --- | --- |
| Evaluation Component | Marks | Type |
| Comprehensive Examination | 40% | Closed |
| Quiz (1) | 10% | Open |
| Minor Projects (1) | 30% | Open |
| Assignments/Exercises (2) | 20% | Open |