# DATA SCIENCE & ANALYTICS

| | | | | | |
|---|---|---|---|---|---|
| **Course Code** | **:** | **CSE3105** | **Credits** | **:** | **02** |
| **Credit Hours** | **:** | **02/week** | **Exam Hours** | **:** | **03** |

**Content of the Course:**

| **No.** | **Topic** | **Lesson Plan** | **Sources of Content** | **Estimated Question Distribution** |
|---|---|---|---|---|
| 1 | **Introduction to data and science** | Types of Data, Scales of measurement, Data sets, Nature of Data Sets, Data Science process | [Rachel Schutt, Cathy O'Neil - Doing Data Science_ Straight Talk from the Frontline-O'Reilly Media (2013).pdf] | 10% |
| 2 | **Data Visualization & Representation** | Graphical methods: histograms, Line graph, Bar chart, Scatter-plot, others<br>Numerical methods: the average, the standard deviation, etc<br>Tabular methods: contingency tables, others<br>Data file format, Vector Space Model, Bag of Words | [Section_2.1_2.2_data_types_and_errors.pdf]<br>[Summarizing and Exploring Data.pdf]<br>[06-VectorSpaceModel.pdf] | |
| 3 | **Exploratory Data Analysis** | Detection of mistakes, Relationships among the explanatory variables, Relationships between explanatory and outcome variables. Types of EDA are univariate non-graphical, multivariate non-graphical, univariate graphical, and multivariate graphical. | [Exploratory Data Analysis.pdf] | 10% |
| 4 | **Data Pre-processing-I** | Data Quality, Data Cleaning: Missing Values, Noisy Data, Data Cleaning Process, Data Integration: The Entity Identification Problem, Redundancy and Correlation Analysis | [illinois_Data_Preprocessing.pdf] | 20% |
| 5 | **Data Pre-processing-II** | Data Reduction: Data Reduction Strategies, Attribute Subset selection, Clustering<br>Data Transformation and Data Discretization: Data Transformation by Normalization, Discretization by Binning | [illinois_Data_Preprocessing.pdf] | |
| 6 | **Knowledge Discovery in Databases** | KDD Process, Database Issues, Databases and Knowledge, Discovered Knowledge, Discovery Algorithms, Application Issues, Introduction to Data Mining | [1992_Frawley_Knowledge discovery in databases An overview.pdf]<br>[1996_Fayyad_From data mining to knowledge discovery in databases.pdf] | 5% |
| 7 | **Statistical Learning** | Why and how to estimate predictive/descriptive models, The Trade-Off Between Prediction Accuracy and Model Interpretability, Supervised Versus | [Introduction to Statistical Learning With Applications in R.pdf] | 50% |

|  |  |  |  |  |
|---|---|---|---|---|
|  |  | Unsupervised Learning, Regression Versus Classification Problems, Assessing Model Accuracy: Measuring the Quality of Fit, The Bias-Variance Trade-Of |  |  |
| 8 | **Predictive Model: Regression** | Linear Regression: Simple Linear Regression: Estimating the Coefficients, Assessing the Accuracy of the Coefficient Estimates, Assessing the Accuracy of the Model, Multiple Linear Regression: Estimating the Regression Coefficients | [Introduction to Statistical Learning With Applications in R.pdf] |  |
| 9 | **Predictive Model: Classification** | Bayes theorem, Naive Bayes, Naive Bayes Classifier, Text Classification, K-nearest-neighbor | [naivebayes.pdf]<br>[knn-1-10.pdf] |  |
| 10 | **Clustering** | Unsupervised learning, Types of clustering, K-Means | [clustering_k-means.pdf] |  |
| 11 | **Metrics for Machine Learning** | Similarity and Dissimilarity measures, evaluation metrics | [http://jcsites.juniata.edu/faculty/rhodes/ml/simdissim.htm]<br>[evaluation_metrics_fall2019-6-22.pdf]<br>[https://drive.google.com/file/d/1Rh5CnVtSrJkSiqhGsdTAffNV_ZMM1bG8/view] |  |
| 12 | **Others** | Inductive Software Engineering, Principles of Inductive Software Engineering, Data Journalism | [2016 Inductive Software Engineering.pdf]<br>[Rachel Schutt, Cathy O'Neil - Doing Data Science_ Straight Talk from the Frontline-O'Reilly Media (2013).pdf] | 5% |

**Text Books:**

(1) Cathy O'Neil, Rachel Schutt. Doing Data Science: Straight Talk from the Frontline. O'Reilly.

(2) Trevor Hastie, Robert Tibshirani, Daniela Witten, Gareth James. “An Introduction to Statistical Learning: With Applications in R”. Springer.

(3) Joseph Adler. “R in a Nutshell”. O’Reilly.

**Reference Books:**

(1) Salvador García, Julián Luengo, Francisco Herrera, “Data Preprocessing in Data Mining”, Springer

**(2)** Russell A. Poldrack, “Statistical Thinking for the 21st Century”.