Prasad V. Potluri Siddhartha Institute of Technology:: Vijayawada.
Department of Computer Science and Engineering

I/II M.Tech. (CSE) - (First Semester)

17CSCS1T3 FUNDAMENTALS OF DATA SCIENCE Credits: 4

Lecture: 4 Periods/week **Internal Assessment: 40 Marks**
Semester end examination: 60 Marks

Course Description:

Data Science is the study of the generalized extraction of knowledge from data. The Data Science keeps tracking integrated skill set spanning strong mathematics, statistics, and computational intelligence, datasets and other branches of computer science along with a good understanding of the problem formulation to develop effective solutions. Students will able to learn the fundamental concepts, techniques and tools that deals with various facets of data science practice, including data collection and integration, exploratory data analysis, predictive modelling, descriptive modelling, data product creation, evaluation, and effective communication. The focus in the treatment of these topics will be on breadth, rather than depth, and emphasis will be placed on integration and synthesis of concepts and their application to solving problems. To make the learning contextual, real datasets from a variety of disciplines will be used.

Course Outcomes:

At the end of the course, the students are able:

CO1: Understand the process of data validation and its role in decision making

CO2: Understand, create, and modify analytic and exploratory algorithms operating over data. Verify and quantify the validity of hypothesis using data analytics.

CO3: Know the privacy and data protection legislation and the data scientist professional code and ethics.

Unit-1

Introduction: What is Data Science? What roles exist in Data Science? Current landscape of perspectives. Define the workflow, tools and approaches data scientists use to analyze data. Define a problem and identify appropriate data sets using the data science workflow. Walk through the data science workflow using a case study.

Unit-2

Statistics Fundamentals: Exploratory Data Analysis and the Data Science Process -analyze datasets using basic summary statistics: mean, median, mode, max, min, quartile, inter-quartile, range, variance, standard deviation and correlation. Data Visualization – scatter plots, scatter matrix, line graph, box blots, and histograms. Identify a normal distribution within a dataset using summary statistics and visualization. Causation Vs Correlation. Test a hypothesis within a sample case study. Validate your findings using statistical analysis.

Unit-3

Foundations of Data Modelling: Introduction Regression – data modelling and linear regression. Categorical variables versus Continuous variables. Build the linear regression/logistic regression model using a dataset. Fit model – regularization, bias and error metrics. Evaluate model fit using loss functions – MSE(Mean Square Error), RMSE (Root MSE), Mean Absolute Error(MAE). Apply different regression models based on fit and complexity. Evaluate model using metrics such as accuracy/error, Confusion matrix, ROC curve and Cross Validation.

Unit-4

Data Science in the real world

Dimensionality Reduction – perform dimensionality reduction using topic models such as PCA and SVD. Refine and extract data/information from sample datasets. Introduction to Classification - define classification model, apply k-NN, Naïve Classifier and Decision trees. Build the classification model using a dataset and evaluate. Working with Time Series Data – Introduction, observations, sub setting data and selecting observations, Time series periodicity and Time Intervals, Plotting Time series,

Text Books:

1. The Art of Data Science: A Guide for Anyone Who Works with Data, Roger D. Peng, Elizabeth Matsui, Lean Pub, 2015.
2. Doing Data Science, Straight Talk from The Frontline, Cathy O'Neil and Rachel Schutt. O'Reilly. 2014.
3. Data Science for Business: What You Need to Know about Data Mining and Data-analytic Thinking, Foster Provost and Tom Fawcett. 2013

4. Elements of Statistical Learning, Trevor Hastie, Robert Tibshirani and Jerome Friedman, Springer, 2009.

Reference Books:

1. Mining of Massive Datasets,Jure Leskovek, Anand Rajaraman and Jeffrey Ullman. , Cambridge University Press. 2014.
2. Machine Learning: A Probabilistic Perspective.Kevin P. Murphy, MIT Press, 2013.
3. Avrim Blum, John Hopcroft and Ravindran Kannan. Foundations of Data Science.
4. Data Mining and Analysis: Fundamental Concepts and Algorithms, Mohammed J. Zaki and Wagner Miera Jr., Cambridge University Press. 2014.
5. R Programming for Data Science, Roger D. Peng, LeanPub, 2015.
6. Python for Data Science for Dummies, Luca Massaron and John Paul Mueller, John Wiley and Sons, 2015.