5KS04 DATA SCIENCE AND STATISTICS

Course Prerequisite: Basic Knowledge of Mathematics &Discrete Structures

Course Objectives: Throughout the course, students will be expected to demonstrate their understanding Data Science and Statistics by being able to do each of the following:

1. To understand the need of data science and Statistics
2. To understand the knowledge of statistics data analysis techniques utilized in business decision making.
3. To understand and apply the different data modeling strategies.
4. To apply the learned concept for the skillful data management.

Course Outcomes (Expected Outcome): On completion of the course, the students will be able to

1. Explain basics and need of data science
2. Demonstrate proficiency with statistical analysis of data.
3. Perform linear and multiple linear regression analysis.
4. Develop the ability to build and assess classification-based models
5. Evaluate outcomes and make decisions based on data.
6. Compare machine learning techniques to solve data science business problems.

Unit I: Introduction to Data Science Hours: 6

What Is Data Science? Where Do We See Data Science? How Does Data Science Relate to Other Fields? The Relationship between Data Science and Information Science, Computational Thinking, Skills for Data Science , Tools for Data Science, Issues of Ethics, Bias, and Privacy in Data Science Data: Data types, Data Collection, Data Pre-processing.

Unit II: Statistical Learning &Inference Hours: 8

Need of Statistics in Data Science, Measures of central tendency: Mean, Median, Mode,. Measures of Dispersion: Range, Variance, Standard Deviation. Techniques: Introduction, Data Analysis and Data Analytics, Descriptive Analysis, Diagnostic Analytics, Predictive Analytics, Prescriptive Analytics, Exploratory Analysis, Mechanistic Analysis, Regression.

Unit III: Regression and its Techniques Hours: 6

Linear Regression, Multiple Linear Regression, Other Considerations in the Regression Model Comparison of Linear Regression with K-Nearest Neighbors

Unit IV: Classification Hours: 8

An Overview of Classification, Why Not Linear Regression? Logistic Regression, Support Vector Machine (SVM): Maximal Margin Classifier, Support Vector Classifier

Unit V: Tree Based Methods Hours: 6

Tree-Based Methods: The Basics of Decision Trees, Regression and Classification Trees, Trees Versus Linear Models, Advantages and Disadvantages, Bagging, Random Forests, Boosting

Unit VI: Unsupervised Learning Hours: 6

Unsupervised Learning: Introduction, Agglomerative Clustering, Divisive Clustering, Expectation Maximization (EM) Introduction to Reinforcement Learning.

Text Book:

1. Chirag Shah," A Hands-on Introduction to Data Science ", Cambridge University Press (2020) ISBN:978-1-108-47244-9.
2. Data Mining: Concepts and Techniques By Jiawei Han, Jian Pei, Micheline Kamber (Chapter 2: point 2.2)
3. Gareth James, Daniela Witten, Trevor Hastie, Robert Tibshirani: An Introduction to Statistical Learning with Applications in Python ,978-3-031-38747-0 Published: 30 June 2023

Reference Books:

1. Cathy O'Neil and Rachel Schutt: Doing Data Science, First Edition, 2014, O'reilly Publications, ISBN:978-1-449-35865-5.
2. DT Editorial Services, "Big Data, Black Book", DT Editorial Services, ISBN: 9789351197577, 2016 Edition