Elective 2 (Open Elective)
CSC07E506: Introduction to Data Science

Contact Hours per Week: 3 Lecture +2 Lab + 1 Tutorial

Number of Credits: 4

Course Outcomes:

CO1: To establish a robust foundation in data science and its application areas.
CO2: Understand the overview of data analytics in the context of machine learning and artificial intelligence
CO3: Apply data preparations and analysis to solve real word problems
CO4: Demonstrate an understanding of statistics and machine learning concepts
CO5: Explore the importance of data visualization.
CO6: Familiarize Python Packages for Basic Data Analytics- numpy, pandas, matplotlib.

	PSO1	PSO2	PSO3	PSO4	PSO5	PSO6	PSO7	PSO8
CO1	✔	✔	✔		✔		✔	
CO2		✔		✔		✔		✔
CO3	✔		✔		✔	✔		
CO4	✔	✔	✔	✔	✔	✔	✔	✔
CO5	✔	✔	✔	✔	✔	✔	✔	✔
CO6		✔		✔		✔		✔

Module I

Introduction to Data Science: Definition - Evolution of Data Science - Data Science Process - Data Collection Strategies – Exploring Your Data -Types of Data -Getting Data-Data Preprocessing – Data Cleaning – Data Integration and Transformation – Dimensionality Reduction – Data Ethics-Applications of Data Science in various fields.

Module II

Statistics and Probability for Data Science: Linear Algebra: Vector-Matrices, Statistics: Central Tendencies-Correlation, Statistical Inference: Populations and samples - Statistical modeling - Probability: Conditional probability-Bayes's Theorem-Continuous Distribution -Normal Distribution, Hypothesis and Inferences: - fitting a model - Hypothesis Testing: p-Value-p-Hacking-Bayesian Inference.

Module III

Machine Learning: Introduction - Supervised Learning - Unsupervised Learning, Reinforcement Learning Basic Machine Learning Algorithms: Linear Regression-Logistic Regression-Naive Bayes--Decision Tree-K-Nearest Neighbours-Clustering-Neural Networks-Deep Learning. Confusion matrix - Evaluation Metrics - Overfitting - Underfitting.

Module IV

Python - Basic syntax and data types - operators - String - Data Structures in Python - List - Tuple - Set - Dictionary - Conditional Statements - Python Loops - Functions - Modules - Packages - Numpy - Pandas - Matplotlib.

Module V

Demonstration of Data Analytics implementations using matplotlib-BarChart-LineChart-ScatterPlot.

References

1. Cathy O'Neil and Rachel Schutt, "Doing Data Science, Straight Talk from The Frontline", O'Reilly, 2014.
2. G James, D Witten, T Hastie and R Tibshirani," An Introduction to Statistical Learning with Applications in R", Springer Texts in Statistics, Springer, 2013.
3. Joel Grus, "Data Science from Scratch: First Principles with Python", O'Reilly Media, 2015.
4. Robert J Woz, Data Analytics for Beginners, 2017.
5. Mohammed J. Zaki and Wagner Miera Jr, "Data Mining and Analysis: Fundamental Concepts and Algorithms", Cambridge University Press, 2014.
6. Matt Harrison, "Learning the Pandas Library: Python Tools for Data Munging, Analysis, and Visualization, O'Reilly, 2016.
7. V.K. Jain, Data Science and Analytics, Khanna Publishing, 2019.
8. Anil Maheshwari, Data Analytics, McGraw Hill Education, 2018