PCAC2002 DATA SCIENCE FOUNDATIONS (3-0-0)

OVERALL COURSE OBJECTIVES: To enhance students' aptitude in implementing scalable data science platforms, and understanding big data landscape with a focus on using statistical measures, data visualization, advanced tools, and specific processes that aid in detecting data trends, minimizing inconsistencies, and improving overall data analysis.

Module 1: Introduction to Data Science in Python [35 Hours]

This course will introduce the learner to the basics of the Python programming environment, including fundamental Python programming techniques such as lambdas, reading and manipulating csv files, and the numpy library. The course will introduce data manipulation and cleaning techniques using the popular Python pandas data science library and introduce the abstraction of the Series and DataFrame as the central data structures for data analysis, along with tutorials on how to use functions such as group by, merge, and pivot tables effectively. By the end of this course, students will be able to take tabular data, clean it, manipulate it, and run basic inferential statistical analyses.

Sub-Topic

Fundamentals of Data Manipulation with Python
Data Processing with Pandas
Answering Questions with Messy Data

Formative Assessments:

4quizzes and 9 coding/lab assignments.

Module 2: Introduction to Big Data [17 Hours]

This course provides an introduction to the Big Data landscape for beginners interested in data science. It includes an overview of key concepts behind big data problems, applications, and systems. The course offers familiarity with the Hadoop framework that simplifies big data analysis, making it more accessible. It covers the characteristics of Big Data, the process of structuring analysis, identification of big data problems, the architectural components, and programming models for scalable big data analysis. It also explores the core Hadoop stack components including the YARN resource and job management system, the HDFS file system, and the MapReduce programming model. Installations and virtual machine operations are required for hands-on assignments. Prior programming experience is not necessary.

Sub-Topic

Big Data: Why and Where
Characteristics of Big Data and Dimensions of Scalability
Data Science: Getting Value out of Big Data
Foundations for Big Data Systems and Programming
Systems: Getting Started with Hadoop

Formative Assessments:

6 quizzes and 1 peer-review assignment.

LEARNING OUTCOMES: On successful completion of the course the students shall be able to:

1. Understand and apply basic statistical measures to identify patterns within large sets of data,

2. Develop proficiency in recognizing various data characteristics, patterns, trends, deviations or inconsistencies, and potential outliers.
3. Employ techniques for dealing with big data like dimension reduction and feature selection methods.
4. Leverage advanced tools and charting libraries to improve the efficiency of big data analysis with partitioning and parallel analysis.
5. Visualize data using 2D and 3D formats achieving a better understanding and interpretation.
6. Get value out of Big Data following a specific 5-step process to structure your analysis.