Code No.: MDS 501

Course Title: **Fundamentals of Data Science** Full Marks: 75

Nature: Theory(Compulsory) Credit: 3

***Course Description*:**

This is an introductory course to teach the basics of data science, it's applications and commonly used tools and techniques. The course is designed to introduce key ideas and methodologies used in the domain of data science. The goal of this course is to help understand the fundamental building blocks of data science.

Learning Objectives:

Upon the conclusion of the course, students should be able to:

- Describe Data Science, skill sets needed to be a data scientist and be familiar with common tools used for data science. Understand the importance of data quality and familiarize with common data munging techniques.
- Understand and apply commonly used data analysis and machine learning techniques in data science
- Identify the challenges in handling big data, and gain a general understanding of tools to handle big data
- Reason around ethical and privacy issues in data science and understand the common biases affecting data science.

***Course Contents*:**

Unit 1: Introduction to Data Science [10 Hrs.]

Introduction to data science, Applications of data science; Limitations of data science

Commonly used tools in data science, their strengths and common use-cases: R/RStudio, Python/Pandas/Jupyter Notebooks, Excel/Tableau/PowerBI;

Data Science life-cycle/Common methodologies for data science: CRISP-DM, OSEMN Framework, TDSP lifecycle;

Review of statistics and probability: Probability distributions, compound events and independence. Statistics: Centrality measures, variability measures, interpreting variance. Correlation analysis: Correlation coefficients, autocorrelation

Unit 2: Data Munging [8 Hrs.]

Data quality, common issues with real world data: Duplicates, Missing Data, Non-standard data, Unit mismatch;

Ways to clean up and standardize data; Data enrichment: Need for data enrichment; Common ways to enrich data: correction, extrapolation, augmentation;

Data Validation: Common methods of data validation: type check, range & constraint check, consistency check;

Data format conversion: Commonly used formats: JSON, XML, Tabular, Relational - their strengths and weaknesses,

Motivation behind format conversion.General methods of conversion between data formats. Tabular data: Row based vs column based (Parquet, ORC, CSV). Wide vsnarrow(long) table format. Converting between wide vs narrow formats

Unit 3: Data Analysis Technique **[10 Hrs.]**

Feature generation and feature selection algorithms: filters, wrappers, decision trees, random forests;

Common techniques: Linear regression, logistic regression, k-NN, k-means ;

Predictive data analysis: Introduction to predictive data analysis and its common applications.;

Regression based models: linear regression, logistic regression.;

Time series data analytics

Unit 4: Machine Learning **[8 Hrs.]**

Introduction to machine learning, type of machine learning methods.;

Supervised vsUnsupervised learning;

Naive Bayes, Decision Trees, SVMs;

Introduction to deep learning, backpropagation.

Unit 5: Introduction to Big Data **[8 Hrs.]**

Introduction to big data and the challenges of handling big data;

Commonly used tools for big data: The map-reduce programming paradigm.Hadoop, HDFS, (py)Spark, Hive.

Data warehousing and data lake architecture.

Real-time analytics with Apache Kafka

Unit 6: Ethical Issues in Data Science **[4Hrs.]**

Issues with fairness and bias in data science:

Common biases: In group favoritism and out-group negativity, Fundamental attribution error, Negativity bias, Stereotyping, Bandwagon effect, Bias blind spot.

Addressing biases: Group unaware selection, Adjusted group thresholds, Demographic parity, Equal opportunity, Precision parity;

Common issues with privacy and data ethics.

References:

1. O'Neil, Cathy and Schutt, Rachel(2013), *Doing Data Science*, *Straight Talk From TheFrontline*, O'Reilly Media
2. Provost, Foster and Fawcett, Tom (2013). *Data Science for Business*: *What You Need to Know about Data Mining and Data-analytic Thinking,* O'Reilly Media.

***