DISCIPLINE SPECIFIC CORE COURSE – 10 (DSC-19): DATA SCIENCE

CREDIT DISTRIBUTION, ELIGIBILITY AND PRE-REQUISITES OF THE COURSE

Course title & Code	Credits	Credit distribution of the course			Eligibility criteria	Pre-requisite of the course (if any)
		Lecture	Tutorial	Practical/ Practice		
Data Science (DSC 19)	4	2	0	2	Class 12	Maths at Class XII level

Learning Objectives

The course aims to:

- Understand the purpose of Data Science is **to find patterns within data**.
- Use various statistical and mathematical techniques to analyse and draw insights from data.
- Explore Special tools and techniques to analyse the data. Realise that data in real life is usually noisy and messy, therefore special tools and techniques are needed to draw meaningful insights from it.

Learning outcomes

By studying this course, the students will be able to:

- Understand the methods and techniques commonly used in data science.
- Retrieve, organize and explore data
- Demonstrate the ability to clean and prepare data for analysis
- Use the techniques of data analysis, inferential statistics, machine learning, and statistical computing in an integrated capacity.

SYLLABUS OF DSC-19

UNIT–I: Introduction to Data Science [5hours]

Definition of Data, Big Data and Data Science. The current landscape of perspectives - Skill sets needed, Work profile of Data Scientists, Data ethics, valuing different aspects of privacy (eg.GDPR); Data science process overview (Defining Goals – Data acquisition-retrieval-preparation-exploratory analysis-modelling-visualisation); Big Data – problems in handling large data, distributed data storage and processing, Supervised and Unsupervised learning Models, Supervised Learning Models: Classification and regression, bias – variance trade-off.

Basics of Python for Machine Learning.

Unit 2:Data Processing [10hours]

Data pre-processing: (Data Wrangling) Data cleaning - data integration - Data Reduction, Data Transformation and Data Discretization. (Univariate analysis, Handling Missing values, and outliers, imputation of missing values, encoding of nominal and ordinal variables, scaling/standardization of variables).

Exploratory Data Analysis - Basic tools (plots, graphs and summary statistics)

Feature Selection methods – Filter methods (correlation, ANOVA, chi-square, variance threshold, Phi-k correlation), Wrapper methods, Decision Trees; Random Forests.

Unit 3:Clustering and introduction to Data Visualisation [5hours]

Clustering: Choosing distance metrics - Different clustering approaches - hierarchical clustering, K-means, DBSCAN, Relative merits of each method. Data Visualization: Basic principles, ideas and tools for data visualization. (Bars, box plots, heat maps, histograms, normal plots)

Unit 4:Machine Learning [10hours]

Familiarisation with machine learning process (training-testing-validation), Basic Machine,Data Imbalance, Data diversity, Machine Learning Pipeline
Learning Supervised learning algorithms: Linear Regression- Regression diagnostics, checking assumptions of Linear Regression, root mean square error, R2 and adjusted R2
Logistic Regression – Understanding concept and application, Data Imbalance, Evaluation metrics-Classification matrix, Sensitivity, Specificity, ROC curves. Decision trees.

Practical Component (30 practical sessions; total 60 Hours): Practical to be based on Python Programming Language. The student is expected to conduct an end-to-end modelling journey which has Exploratory Data Analysis (EDA), feature engineering, Model development, turning and interpretation of the results. An econometrics-based project to be taken up to constitute the end-term practical examination.

Essential/recommended readings

1. O'Neil Cathy and Schutt Rachel (2014). Doing Data Science, Straight Talk from The Frontline, O'Reilly.
2. McKinney,Wes. (2012). Python for Data Analysis: Data Wrangling with Pandas, NumPy, and IPython, O'Reilly Media.
3. Gareth James, Daniela Witten, Trevor Hastie, Robert Tibshirani, Jonathan Taylor (2023). An Introduction to Statistical Learning: With Applications in Python,Springer Cham.

Suggested readings

1. Harrison,Matt, (2016), Learning the Pandas Library: Python Tools for Data Munging, Analysis, and Visualization, O'Reilly.
2. Grus Joel (2015), Data Science from Scratch: First Principles with Python, O'Reilly Media.