Course Code	Course name	L	T	P	C
CSDS2001P	Fundamentals of Data Science	4	0	0	4
Total Units to be Covered: 06		Total Contact Hours: 60			
Prerequisite(s):	Basics of mathematics, programming	Syllabus version: 1.0			

Course Objectives

1. To understand the concept of data science.

2. To understand techniques and methods related to the area of data science on real world applications.

Course Outcomes

After the completion of the course the students will be able to

CO1: Understand the fundamentals of data processing.

CO2: Understand and apply mathematical concepts in the field of data science.

CO3: Employ the techniques and methods related to the area of data science in a variety of applications.

CO4: Apply logical thinking to understand and solve the problem in context.

CO5: Apply the entire concept in data analysis tools.

CO-PO Mapping

Program Outcomes / Course Outcomes	PO1	PO2	PO3	PO4	PO5	PO6	PO7	PO8	PO9	PO10	PO11	PO12	PSO1	PSO2	PSO3
CO 1	-	3	2	2	-	-	-	-	-	-	-	-	-	3	3
CO 2	-	3	2	2	-	-	-	-	-	-	-	-	-	3	3
CO 3	-	3	2	3	-	-	-	-	-	-	-	-	-	3	3
CO4	-	3	2	3	-	-	-	-	-	-	-	-	-	3	3
CO5	-	3	2	2	-	-	-	-	-	-	-	-	-	2	3
Average	-	3	2	2.4	-	-	-	-	-	-	-	-	-	2.8	3

1 – Weakly Mapped (Low) 2 – Moderately Mapped (Medium) 3 – Strongly Mapped (High)
"_" means there is no correlation

Syllabus

Unit I: Introduction to Data Science 8 Lecture Hours

Fundamentals of Data Science, Real World Applications, Data Science Challenges, Software Engineering for Data Science (DataOps, MLOps (intro)). Data science process roles, Stages in data science. Defining Analytics, Types of data analytics (Descriptive, Diagnostic, Predictive, Prescriptive) Data Science Process: CRISP-DM Methodology, SEMMA, BIG DATA LIFE CYCLE, SMAM.

Unit II: Probability and statistics for Data Science 12 Lecture Hours

Probability: Introduction, finite sample spaces, conditional probability, independence; Random variables, distribution functions, probability mass and density functions, standard univariate discrete and continuous distributions; Mathematical expectations, moments; Random vectors, joint, marginal, and conditional distributions, independence, covariance, correlation, standard multivariate distributions, functions of random vectors; central limit theorem.

Statistics: Sampling distributions of the sample mean and the sample variance for a normal population; Point and interval estimation; Sampling distributions (Chi-square, t,F,Z), Hypothesis testing; One tailed and two-tailed tests; Analysis of variance, ANOVA, One way and two way classifications

Unit III: Data, Data Sources and Visualization 15 Lecture Hours

Types of Data and Datasets, Data Quality, and Issues, Data Models, General Framework of Formal modeling, Association Analyses, Prediction Analyses, Data Pipelines and patterns, Data from files & working with relational databases, Diverse data sources, data warehouses, data mining, cloud, and Data lake: Characteristics, components, Data Streaming Ingestion, Batch Data Ingestion, Data Cataloging, Data Pipeline Stages (extraction, ingestion, cleaning, exploration, wrangling, versioning, Data transformation, Feature management). Data Visualization: Overview of visualization techniques for Data Exploratory analysis

Unit IV: Feature Engineering and Optimization 10 Lecture Hours

Feature Extraction, Feature Construction, Feature Subset selection, Feature Learning, Feature Reduction (Dimensionality Reduction) Case Study involving FE tasks, and Feature Engineering techniques for text, images, audio, and video.Necessary and sufficiency conditions for optima;

Gradient descent methods; Constrained optimization; Introduction to non-gradient techniques; Introduction to least squares optimization; Optimization view of machine learning.

Unit V: Supervised and unsupervised learning **10 Lecture Hours**

Introduction to Machine Learning, types, Supervised Learning: Overview, workflow, data processing, Linear Regression, Logistic Regression, Decision Trees, Random Forest, Support Vector Machines (SVM), k-Nearest Neighbors (k-NN).

Unsupervised Learning: Overview, clustering algorithms: K-Means Clustering, Hierarchical Clustering, DBSCAN, Gaussian Mixture Models (GMM),

Dimensionality Reduction: Principal Component Analysis (PCA), t-Distributed Stochastic Neighbor Embedding (t-SNE)

Association Rule Mining: Apriori Algorithm, FP-Growth Algorithm, Anomaly Detection, Model Evaluation (Silhouette Score, Inertia, etc.)

Use Cases and Practical Applications

Unit VI: Data Analysis Tool **5 Lecture Hours**

Reading and getting data into R, ordered and unordered factors i.e arrays and matrices – lists and data frames, reading data from files, probability distributions statistical models in R - manipulating objects – data distribution.

Total lecture Hours 60

Textbooks

1. G. Strang, “Introduction to Linear Algebra”, 5th Edition, Wellesley-Cambridge Press, USA, 2016.
2. D. C. Montgomery, and G. C. Runger, “Applied Statistics and Probability for Engineers”, 5th Edition, John Wiley & Sons, Inc., NY, USA, 2011.
3. Nina Zumel, and John Mount, “Practical Data Science with R”, Manning Publications, 2014.
4. Avrim Blum, John Hopcroft, and Ravindran Kannan, “Foundations of Data Science”, 2018. Available online at: https://www.cs.cornell.edu/jeh/book.pdf.

Reference Books

1. Mark Gardener, “Beginning R - The Statistical Programming Language”, John Wiley & Sons, Inc., 2012.
2. W. N. Venables, D. M. Smith and the R Core Team, “An Introduction to R”, 2013. Available online at: https://cran.r-project.org/doc/manuals/R-intro.pdf.
3. S. Abiteboul, R. Hull, V. Vianu, “Foundations of Databases”, Addison Wesley, 1995.
4. J. S. Bendat, and A. G. Piersol, “Random Data: Analysis and Measurement Procedures”, 4th Edition, John Wiley & Sons, Inc., NY, USA, 2010.
5. D. C. Montgomery, and G. C. Runger, “Applied Statistics and Probability for Engineers”, 5th Edition, John Wiley & Sons, Inc., NY, USA, 2011.
6. Cathy O’Neil, and Rachel Schutt, “Doing Data Science”, O’Reilly Media, 2013.

Modes of Evaluation: Quiz/Assignment/ presentation/ extempore/ Written Examination

Examination Scheme

Components	IA	MID SEM	End Sem	Total
Weightage (%)	50	20	30	100

Detailed breakup of Internal Assessment

Internal Assessment Component	Weightage in calculation of Internal Assessment (100 marks)
Quiz 1	15%
Quiz 2	15%
Class Test 1	15%
Class Test 2	15%
Assignment 1/Project	20%
Assignment 2/Project	20%