SREENIVASAINSTITUTEOFTECHNOLOGYANDMANAGEMENTSTUDIES
(Autonomous)

Minor - DATA SCIENCE AND ANALYTICS - Offered by CSD Department

S.No	Course Code	CourseTitle	Scheme of Instructions HoursperWeek				Scheme ofExaminationMaximumMarks		
			L	T	P	C	I	E	Total
1	23MRDSA1	Introduction to Data Science	3	-	-	3	30	70	100
2	23MRDSA2	Statistical Learning for Data Science	3	-	-	3	30	70	100
3	23MRDSA3	Data Visualization & Exploratory Data Analysis	3	-	-	3	30	70	100
4	23MRDSA4	Big Data Analytics	3	-	-	3	30	70	100
5	23MRDSA5	Recommender Systems	3	-	-	3	30	70	100
6	23MRDSA6	Data Visualization & EDA Lab	-	-	3	1.5	30	70	100
7	23MRDSA7	Recommender System Lab	-	-	3	1.5	30	70	100

SREENIVASAINSTITUTEOFTECHNOLOGYANDMANAGEMENTSTUDIES
(Autonomous)

CSD Department

DATA SCIENCE

23MRDSA1	INTRODUCTION TO DATA SCIENCE	L	T	P	C
		3	0	0	3

PRE-REQUISITES:

COURSE EDUCATIONAL OBJECTIVES:

- Provide a foundational understanding of data science processes and applications.
- Introduce key tools and techniques such as Python, statistics, data cleaning, visualization, and machine learning.
- Develop practical skills in data analysis, interpretation, and data storytelling.
- Enable students to work on real-world datasets using data science techniques.
- Prepare students for advanced studies or industry roles in data science and analytics.

UNIT –1: Introduction to Data Science **(9)**

What is Data Science?,Role of Data Scientist, Data Science Process (Problem definition, data collection, preprocessing, modeling, evaluation), Applications of Data Science in different domains, Tools: Jupyter, Anaconda, Python/R Overview

UNIT –2: Data Handling and Preprocessing: **(9)**

Introduction to NumPy and Pandas, Reading data from CSV, Excel, SQL, Data Wrangling: Missing values, duplicates, outliers, Data transformation: Scaling, encoding, Feature engineering basics

UNIT –3: Data Visualization: **(9)**

Importance of visualization, Visualization libraries (Matplotlib, Seaborn), Histograms, Boxplots, Pairplots, Heatmaps, Dashboards and Storytelling with Data, Real-time data dashboards (Optional)

UNIT –4: Statistical Foundations for Data Science: **(9)**

Descriptive Statistics, Probability and Probability Distributions, Inferential Statistics: Hypothesis Testing, Confidence Intervals, Correlation and Causation, Use of Scipy/Statsmodels for statistical analysis

UNIT –5: Introduction to Machine Learning: **(9)**

Supervised vs Unsupervised Learning, Classification and Regression problems, Basic ML Algorithms: Linear Regression, Logistic Regression, KNN, Decision Trees, Model Evaluation Metrics: Accuracy, Precision, Recall, F1-Score, Overfitting and Underfitting

Total Hours: 45

COURSE OUTCOMES:

On successful completion of the course- students will be able to		**Bloom's Level**
CO1	Explain the data science lifecycle and its importance in business and research.	Understand (L2)
CO2	Use Python and libraries like Pandas, NumPy, and Matplotlib for data handling.	Apply (L3)
CO3	Perform data cleaning, transformation, and visualization effectively.	Apply (L3)
CO4	Apply basic machine learning models for classification and regression.	Apply (L3)
CO5	Interpret data analysis results and communicate findings clearly.	Analyze (L4)

TEXT BOOKS:

1. **Joel Grus** – *Data Science from Scratch: First Principles with Python*, O'Reilly.
2. **Cathy O'Neil and Rachel Schutt** – *Doing Data Science*, O'Reilly.
3. **Wes McKinney** – *Python for Data Analysis*, O'Reilly.

REFERENCE BOOKS:

1. **Jake VanderPlas** – *Python Data Science Handbook*, O'Reilly.
2. **Andreas Müller & Sarah Guido** – *Introduction to Machine Learning with Python*.
3. **Han, Kamber, & Pei** – *Data Mining: Concepts and Techniques*, Morgan Kaufmann.

REFERENCE WEBSITE:

NPTEL / SWAYAM:

- NPTEL: Introduction to Data Science
  - Instructor: Prof. RaghunathanRengasamy, IIT Madras

Coursera:

- **IBM Data Science Professional Certificate**
  Link: coursera.org
- **Introduction to Data Science in Python (University of Michigan)**
  Link: coursera.org

SREENIVASAINSTITUTEOFTECHNOLOGYANDMANAGEMENTSTUDIES
(Autonomous)

23MRDSA2	STATISTICAL LEARNING FOR DATA SCIENCE	L	T	P	C
		3	0	0	3

PRE-REQUISITES:

COURSE EDUCATIONAL OBJECTIVES:

- Introduce fundamental concepts of statistical learning and its importance in data science.
- Develop understanding of regression, classification, and resampling methods.
- Build skills in model assessment, selection, and regularization techniques.
- Apply statistical learning methods to real-world datasets.
- Interpret, communicate, and evaluate data-driven models.

Unit I: Introduction to Statistical Learning: (9)

What is Statistical Learning?, Supervised vs Unsupervised Learning, Model Accuracy vs Interpretability, Bias-Variance Trade-off, Curse of Dimensionality, Applications of Statistical Learning

Unit II: Linear and Polynomial Regression: (9)

Simple Linear Regression. Multiple Linear Regression, Polynomial Regression, Assumptions of Linear Models, Model Diagnostics and Performance Metrics (R^2, RMSE), Variable Selection Techniques

Unit III: Classification Methods: (9)

Logistic Regression, Discriminant Analysis (LDA, QDA), Naïve Bayes Classifier, K-Nearest Neighbors, Confusion Matrix, Precision, Recall, F1-Score

Unit IV: Resampling and Model Assessment: (9)

Train-Test Split, Cross Validation (k-Fold, LOOCV), Bootstrap Methods, Model Selection and Hyperparameter Tuning

Unit V: Regularization & Advanced Topics: (9)

Ridge Regression, Lasso Regression, Shrinkage Methods, Principal Component Regression (PCR), Partial Least Squares (PLS),Introduction to Support Vector Machines.

Total Hours: 45

COURSE OUTCOMES:

On successful completion of the course- students will be able to		**Bloom's Level**
CO1	Understand the basics of statistical learning and data representation.	Understand (L2)
CO2	Apply linear regression, logistic regression, and classification techniques.	Apply (L3)
CO3	Evaluate models using resampling methods and cross-validation.	Evaluate (L5)
CO4	Analyze regularization methods like Ridge and Lasso to avoid overfitting.	Analyze (L4)
CO5	Create and interpret statistical learning models in practical scenarios.	Create (L6)

TEXT BOOKS:

1. **Gareth James, Daniela Witten, Trevor Hastie, and Robert Tibshirani**
   An Introduction to Statistical Learning with Applications in R – Springer
   (Free PDF: https://www.statlearning.com/)
2. **Trevor Hastie, Robert Tibshirani, Jerome Friedman**
   The Elements of Statistical Learning – Springer

REFERENCE BOOKS:

1. **Norman Matloff** – *Statistical Regression and Classification: from R to Data Science*
2. **Chris Bishop** – *Pattern Recognition and Machine Learning*
3. **T. Ryan** – *Modern Regression Methods*

REFERENCE WEBSITE:

NPTEL / SWAYAM:

- NPTEL: Introduction to Statistical Learning
  - Instructor: Prof. BalaramanRavindran (IIT Madras)

SREENIVASAINSTITUTEOFTECHNOLOGYANDMANAGEMENTSTUDIES
(Autonomous)

23MRDSA3	DATA VISUALIZATION & EXPLORATORY DATA ANALYSIS	L	T	P	C
		3	0	0	3

PRE-REQUISITES:

COURSE EDUCATIONAL OBJECTIVES:

- To understand the role of data visualization and EDA in the data science lifecycle.
- To explore data types, distributions, missing values, and outliers using statistical and visual methods.
- To gain expertise in using tools such as Python (Matplotlib, Seaborn, Plotly) or R for EDA.
- To develop meaningful, interactive visualizations for both univariate and multivariate data.
- To interpret trends and communicate data-driven insights effectively through dashboards and storytelling.

Unit I: Introduction to EDA and Data Types (9)

What is Exploratory Data Analysis?,Types of Data (Categorical, Numerical, Ordinal), Central tendency & dispersion: Mean, Median, Mode, Variance, StdDev, Importance of visualization in EDA

Unit II: Data Cleaning & Preprocessing: (9)

Handling missing values and imputation. Detecting and handling outliers, Feature engineering basics, Data transformations (scaling, normalization, encoding)

Unit III: Data Visualization Tools & Techniques: (9)

Univariate visualizations: Histogram, Boxplot, Violin plot, Pie chart, Bivariate and Multivariate visualizations: Scatter plot, Heatmaps, Pairplot, Time series plots, Density plots, Introduction to Matplotlib, Seaborn, Plotly, and Tableau

Unit IV: Advanced Data Visualization Concepts: (9)

Interactive visualizations: Hover, Zoom, Filters, Geospatial visualization using Folium or Mapbox, Visual perception principles (Gestalt, color theory), Visual storytelling and design best practices

Unit V: Dashboards & Real-World Case Studies: (9)

Creating dashboards using Power BI/Tableau/Plotly Dash, Building EDA pipelines with Python/R, Case studies: Finance, Healthcare, Social media, IoT, Interpreting visualizations for decision-making

Total Hours: 45

COURSE OUTCOMES:

On successful completion of the course- students will be able to		Bloom's Level
CO1	Understand the significance of data exploration and visualization in the data analysis process.	Understand (L2)
CO2	Identify and preprocess anomalies, missing data, and data types through exploratory methods.	Analyze (L4)
CO3	Apply visual and statistical techniques to univariate, bivariate, and multivariate data.	Apply (L3)
CO4	Evaluate different chart types and visualization libraries/tools for effective communication.	Evaluate (L5)
CO5	Create storytelling dashboards and dynamic visual reports for real-world datasets.	Create (L6)

TEXT BOOKS:

1. **Alberto Cairo**, *The Truthful Art: Data, Charts, and Maps for Communication*, New Riders.
2. **Nathan Yau**, *Visualize This: The FlowingData Guide to Design, Visualization, and Statistics*, Wiley.
3. **Hadley Wickham**, *R for Data Science*, O'Reilly (for R users).

REFERENCE BOOKS:

1. **Ben Fry**, *Visualizing Data: Exploring and Explaining Data with the Processing Environment*, O'Reilly.
2. **Claus O. Wilke**, *Fundamentals of Data Visualization*, O'Reilly.
3. **Scott Murray**, *Interactive Data Visualization for the Web*, O'Reilly.

REFERENCE WEBSITE:

Platform	Course Title & Link
Coursera	Data Visualization with Python – IBM

SREENIVASAINSTITUTEOFTECHNOLOGYANDMANAGEMENTSTUDIES
(Autonomous)

23MRDSA4	BIG DATA ANALYTICS	L	T	P	C
		3	0	0	3

PRE-REQUISITES:

COURSE EDUCATIONAL OBJECTIVES:

- Understand the fundamentals of Big Data, its characteristics, and challenges.
- Learn key tools and technologies for storing, processing, and analyzing Big Data (Had oop, Spark, etc.).
- Explore advanced analytical techniques including Map Reduce, machine learning on Big Data, and real-time analytics.
- Develop skills to design and implement Big Data solutions for real-world problems.
- Gain hands-on experience with Big Data frameworks and platforms.

UNIT I: INTRODUCTION TO BIG DATA (9)

Definition and Characteristics (Volume, Velocity, Variety, Veracity, Value), Challenges of Big Data, Traditional Data Processing vs Big Data Analytics, Overview of Big Data applications.

UNIT II: BIG DATA TECHNOLOGIES AND ECOSYSTEM: (9)

Hadoop Framework: HDFS, YARN, MapReduce, Apache Spark: Architecture and Components, NoSQL Databases (HBase, Cassandra, MongoDB), Data ingestion tools (Sqoop, Flume)

UNIT III: MAP REDUCE PROGRAMMING MODEL: (9)

MapReduce fundamentals, Writing Map and Reduce functions, Job configuration and execution, Hands-on example: Word count, log analysis.

UNIT IV: ADVANCED BIG DATA ANALYTICS: (9)

Machine Learning on Big Data using MLlib (Spark), Real-time analytics and streaming data (Spark Streaming, Kafka), Graph processing (GraphX), Data visualization of Big Data

UNIT V: BIG DATA PROJECT AND CASE STUDIES: (9)

Designing Big Data solutions for healthcare, finance, social media analytics, Performance optimization and security considerations, Cloud-based Big Data platforms (AWS EMR, Google BigQuery), Case studies and project presentations.

Total Hours: 45

COURSE OUTCOMES:

On successful completion of the course- students will be able to		Bloom's Level
CO1	Understand the core concepts and architecture of Big Data and related technologies.	Understand (L2)
CO2	Analyze Big Data processing frameworks like Hadoop and Spark for scalability and efficiency.	Analyze (L4)
CO3	Apply MapReduce programming model to solve Big Data problems.	Apply (L3)
CO4	Evaluate different storage options (HDFS, NoSQL) and processing techniques for Big Data.	Evaluate (L5)
CO5	Create end-to-end Big Data analytics pipelines using relevant tools and platforms.	Create (L6)

TEXT BOOKS:

1. **Tom M. Mitchell**, *Machine Learning*, McGraw-Hill, 1997.
   Viktor Mayer-Schönberger, Kenneth Cukier, *Big Data: A Revolution That Will Transform How We Live, Work, and Think*, Houghton Mifflin Harcourt.
2. **Tom White**, *Hadoop: The Definitive Guide*, O'Reilly Media.
3. **Jules S. Damji, Brooke Wenig, Tathagata Das, Denny Lee**, *Learning Spark: Lightning-Fast Big Data Analysis*, O'Reilly Media.

REFERENCE BOOKS:

1. **Raj Kamal**, *Big Data Analytics*, McGraw-Hill Education.
2. **Chris Eaton, Dirk DeRoos, Tom Deutsch, George Lapis, Paul Zikopoulos**, *Understanding Big Data: Analytics for Enterprise Class Hadoop and Streaming Data*, McGraw-Hill.
3. **AnandRajaraman, Jeffrey David Ullman**, *Mining of Massive Datasets*, Cambridge University Press.

REFERENCE WEBSITE:

Platform	Course Title & Link
Coursera	Big Data Specialization by University of California San Diego

SREENIVASAINSTITUTEOFTECHNOLOGYANDMANAGEMENTSTUDIES
(Autonomous)

23MRDSA5	RECOMMENDER SYSTEMS	L	T	P	C
		3	0	0	3

PRE-REQUISITES:

COURSE EDUCATIONAL OBJECTIVES:

- Understand the fundamental concepts and techniques of recommender systems.
- Explore different types of recommendation approaches: content-based, collaborative filtering, and hybrid methods.
- Learn about evaluation metrics and methods to assess recommender system performance.
- Understand challenges such as scalability, sparsity, and cold-start problems.
- Gain hands-on experience implementing and tuning recommender algorithms using real-world datasets.

UNIT –1: Introduction to Recommender Systems (9)

Definition and applications (e-commerce, media, social networks), Types of recommender systems: content-based, collaborative filtering, hybrid, Overview of user-item interactions

UNIT –2: Content-Based Recommendation: (9)

Item profiles and user profiles, Similarity measures (cosine similarity, Pearson correlation), Building content-based recommenders, Advantages and limitations

UNIT –3: Collaborative Filtering Techniques (9)

User-based collaborative filtering, Item-based collaborative filtering, Matrix factorization techniques (SVD, PCA), Addressing sparsity and scalability issues.

UNIT –4: Evaluation of Recommender Systems (9)

Metrics: Precision, Recall, F1-score, RMSE, MAE, Cross-validation and train-test split, Handling cold-start and diversity, Case studies on real-world evaluation.

UNIT –5: Advanced Topics and Hybrid Systems (9)

Hybrid recommender systems (weighted, switching, feature combination), Context-aware recommendations, Deep learning in recommender systems, Privacy, ethics, and future trends

Total Hours: 45

COURSE OUTCOMES:

On successful completion of the course- students will be able to		Bloom's Level
CO1	Explain the basic principles and types of recommender systems.	Understand (L2)
CO2	Analyze various recommendation algorithms including collaborative and content-based filtering.	Analyze (L4)
CO3	Apply different algorithms to build practical recommender system models.	Apply (L3)
CO4	Evaluate the performance of recommender systems using appropriate metrics.	Evaluate (L5)
CO5	Design hybrid recommender systems that address common challenges.	Create (L6)

TEXT BOOKS:

1. **Charu C. Aggarwal**, *Recommender Systems: The Textbook*, Springer, 2016.
2. **Francesco Ricci, LiorRokach, BrachaShapira, Paul B. Kantor (Editors)**, *Recommender Systems Handbook*, Springer, 2015.

REFERENCE BOOKS:

1. **DietmarJannach, Markus Zanker, Alexander Felfernig, Gerhard Friedrich**, *Recommender Systems: An Introduction*, Cambridge University Press, 2010.
2. **Joseph A. Konstan, John Riedl**, *Collaborative Filtering Recommender Systems*, Now Publishers, 2012.
3. **GuibingGuo, Jie Zhang, Neil Yorke-Smith**, *Deep Learning Based Recommender System*, Springer, 2020.

REFERENCE WEBSITE:

Coursera Recommender Systems Specialization by University of Minnesota

SREENIVASAINSTITUTEOFTECHNOLOGYANDMANAGEMENTSTUDIES
(Autonomous)

23MRDSA6	DATA VISUALIZATION & EDA LAB	L	T	P	C
		-	-	3	1.5

PRE-REQUISITES: Nil.

COURSE EDUCATIONAL OBJECTIVES:

- To explore and visualize datasets.
- To derive insights through analysis.

Experiments:

1. Importing data from different sources
2. Handling missing values and outliers
3. Univariate and bivariate analysis
4. Boxplots, Histograms, Heatmaps
5. Correlation matrix
6. Using pandas profiling
7. Data visualization using Plotly
8. Pairplots and violin plots
9. Feature engineering basics
10. Time series visualization
11. Dashboards with Streamlit

Course Outcomes:

- Conduct EDA on datasets.
- Create meaningful visualizations.

SREENIVASAINSTITUTEOFTECHNOLOGYANDMANAGEMENTSTUDIES
(Autonomous)

23MRDSA7	RECOMMENDER SYSTEM LAB	L	T	P	C
		-	-	3	1.5

PRE-REQUISITES: Nil.

COURSE EDUCATIONAL OBJECTIVES:

- To understand collaborative and content filtering.
- To build and test recommender models.

Experiments:

1. Introduction to recommender systems
2. User-item interaction matrix
3. Cosine similarity for recommendations
4. Collaborative filtering – user-based
5. Collaborative filtering – item-based
6. Content-based filtering
7. Surprise library intro
8. Building SVD model
9. Hybrid recommender
10. Evaluation metrics (RMSE, MAE)
11. Recommendation visualization

Course Outcomes:

- Build simple recommender systems.
- Evaluate recommender performance.