SREENIVASAINSTITUTEOFTECHNOLOGYANDMANAGEMENTSTUDIES
(Autonomous)

Minor - DATA SCIENCE - Offered by CSD Department

S.No	Course Code	CourseTitle	Scheme of Instructions HoursperWeek				Scheme ofExaminationMaximumMarks		
			L	T	P	C	I	E	Total
1	23MRDSE1	Introduction to Data Science	3	-	-	3	30	70	100
2	23MRDSE2	Statistical Learning for Data Science	3	-	-	3	30	70	100
3	23MRDSE3	Data Engineering	3	-	-	3	30	70	100
4	23MRDSE4	Machine learning fundamentals	3	-	-	3	30	70	100
5	23MRDSE5	Data Science applications	3	-	-	3	30	70	100
6	23MRDSE6	Data Science Practice Lab (R/Python)	-	-	3	1.5	30	70	100
7	23MRDSE7	Statistical Learning & ML Lab	-	-	3	1.5	30	70	100

SREENIVASAINSTITUTEOFTECHNOLOGYANDMANAGEMENTSTUDIES
(Autonomous)

CSD Department

DATA SCIENCE

23MRDSE1	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

SREENIVASAINSTITUTEOFTECHNOLOGYANDMANAGEMENTSTUDIES
(Autonomous)

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)

23MRDSE2	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

SREENIVASAINSTITUTEOFTECHNOLOGYANDMANAGEMENTSTUDIES
(Autonomous)

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)

23MRDSE3	DATA ENGINEERING	L	T	P	C
		3	0	0	3

PRE-REQUISITES:

COURSE EDUCATIONAL OBJECTIVES:

- Understand the data engineering lifecycle, tools, and platforms.
- Gain proficiency in building scalable data pipelines using batch and streaming systems.
- Explore data storage formats, data lakes, and warehouses.
- Learn about workflow orchestration, scheduling, and data quality.
- Apply data engineering techniques in real-time and big data environments.

Unit I: Introduction to Data Engineering: **(9)**

What is Data Engineering?, Data Engineering vs Data Science, Data Lifecycle and Architectures, Roles and Responsibilities of a Data Engineer, Tools Overview: Hadoop, Spark, Kafka, Airflow, Hive.

Unit II: Data Storage and Formats: **(9)**

Structured, Semi-structured, and Unstructured Data, File Formats: CSV, JSON, Parquet, Avro, ORC, Data Warehousing Concepts (Star/Snowflake Schema), Data Lakes vs Data Warehouses, Cloud Storage Solutions (S3, Azure Blob, GCS)

Unit III: Batch and Stream Processing: **(9)**

ETL vs ELT, Batch Processing with Apache Spark, Stream Processing with Apache Kafka and Spark Streaming, Windowing, Late Events, and Watermarks, Lambda and Kappa Architectures.

Unit IV: Workflow Orchestration & Data Quality: **(9)**

Introduction to Apache Airflow, DAGs, Operators, Scheduling, Data Lineage and Observability, Data Profiling, Cleaning, Validation, Ensuring Data Quality with GreatExpectations

Unit V: End-to-End Pipeline and Case Studies: **(9)**

Designing a Data Pipeline: Source to Sink, Real-time Analytics with Kafka + Spark + Cassandra, Monitoring and Logging, CI/CD in Data Engineering (using GitHub Actions or Jenkins), Case Studies: Fraud Detection, Log Analysis, and Recommendation Engine.

Total Hours: 45

SREENIVASAINSTITUTEOFTECHNOLOGYANDMANAGEMENTSTUDIES
(Autonomous)

COURSE OUTCOMES:

On successful completion of the course- students will be able to		Bloom's Level
CO1	Understand the foundations of data engineering tools, platforms, and data lifecycle.	Understand (L2)
CO2	Design and implement batch and streaming data pipelines.	Apply (L3)
CO3	Analyze data storage and querying using file formats, data lakes, and warehouses.	Analyze (L4)
CO4	Evaluate orchestration and workflow management tools.	Evaluate (L5)
CO5	Build end-to-end data engineering workflows and troubleshoot data quality issues.	Create (L6)

TEXT BOOKS:

1. **Andreas François Vermeulen**, *The Data Engineering Cookbook* – Packt
2. **Joe Reis and Matt Housley**, *Fundamentals of Data Engineering* – O'Reilly, 2022
3. **AnuragBhardwaj**, *Data Engineering with Apache Spark, Delta Lake, and Lakehouse* – Packt

REFERENCE BOOKS:

1. *Designing Data-Intensive Applications* – Martin Kleppmann (O'Reilly)
2. *Streaming Systems: The What, Where, When, and How of Large-Scale Data Processing* – Tyler Akidau

REFERENCE WEBSITE:

NPTEL / SWAYAM

- NPTEL: Cloud Computing and Distributed Systems
- NPTEL: Big Data Computing

SREENIVASAINSTITUTEOFTECHNOLOGYANDMANAGEMENTSTUDIES
(Autonomous)

23MRDSE4	MACHINE LEARNING FUNDAMENTALS	L	T	P	C
		3	0	0	3

PRE-REQUISITES:

COURSE EDUCATIONAL OBJECTIVES:

- Understand core concepts of supervised, unsupervised, and reinforcement learning.
- Learn foundational algorithms for classification, regression, and clustering.
- Analyze model performance using evaluation metrics and tuning methods.
- Gain practical knowledge in implementing ML models using datasets.
- Understand the mathematical and probabilistic foundations of ML algorithms.

UNIT I: INTRODUCTION TO MACHINE LEARNING: (9)

Definition and Scope of ML, Types of Learning: Supervised, Unsupervised, Reinforcement, Basic Terminology: Instance, Feature, Label, Training and Testing Sets, Applications and Challenges, Python Libraries: scikit-learn, pandas, numpy

UNIT II: SUPERVISED LEARNING – REGRESSION & CLASSIFICATION: (9)

Linear Regression, Polynomial Regression, Logistic Regression, k-Nearest Neighbors (k-NN), Decision Trees and Random Forests, Naïve Bayes Classifier

UNIT III: UNSUPERVISED LEARNING & DIMENSIONALITY REDUCTION: (9)

k-Means Clustering, Hierarchical Clustering, Principal Component Analysis (PCA), t-SNE, Autoencoders (Intro Only). Association Rule Mining,

UNIT IV: MODEL EVALUATION AND TUNING: (9)

Confusion Matrix, Precision, Recall, F1-Score, ROC-AUC, Bias-Variance Tradeoff, Cross-Validation, Hyperparameter Tuning: Grid Search & Random Search, Overfitting&Underfitting

UNIT V: ADVANCED TOPICS & REAL-TIME APPLICATIONS: (9)

Introduction to Neural Networks (Perceptron), Introduction to SVMs, Feature Engineering & Selection, Real-world Use Cases: Spam Detection, Credit Scoring, Medical Diagnosis, Ethical AI and Model Explainability (XAI – Intro Only).

Total Hours: 45

SREENIVASAINSTITUTEOFTECHNOLOGYANDMANAGEMENTSTUDIES
(Autonomous)

COURSE OUTCOMES:

On successful completion of the course- students will be able to		**Bloom's Level**
CO1	Understand the concepts and assumptions behind key ML algorithms.	Understand (L2)
CO2	Apply various supervised and unsupervised learning techniques.	Apply (L3)
CO3	Analyze the strengths and limitations of different machine learning models.	Analyze (L4)
CO4	Evaluate model performance using metrics and validation techniques.	Evaluate (L5)
CO5	Design and implement complete ML solutions for real-world problems.	Create (L6)

TEXT BOOKS:

1. **Tom M. Mitchell**, *Machine Learning*, McGraw-Hill, 1997.

SREENIVASAINSTITUTEOFTECHNOLOGYANDMANAGEMENTSTUDIES
(Autonomous)

23MRDSE5	DATA SCIENCE APPLICATIONS	L	T	P	C
		3	0	0	3

PRE-REQUISITES:

COURSE EDUCATIONAL OBJECTIVES:

- Understand key domains where data science plays a transformative role.
- Learn how data is collected, preprocessed, and analyzed in different domains.
- Explore application-driven problem-solving using real-world datasets.
- Evaluate models and interpret data science outputs effectively.
- Gain hands-on exposure to tools and techniques used in data science applications.

UNIT –1: Introduction to Data Science & Domains (9)

Definition and Workflow of Data Science. Lifecycle of a Data Science Project, Overview of Applications: Healthcare, Finance, Retail, Manufacturing, Agriculture, Education, and Smart Cities, Ethical Considerations in Data Science.

UNIT –2: Data Science in Healthcare and Finance: (9)

Use cases: Disease prediction, EHR data, medical imaging, outbreak modelling, Predictive modeling for diagnosis (e.g., Diabetes prediction), Applications in FinTech: Fraud detection, Risk scoring, Credit analytics, Time series analysis in financial markets.

UNIT –3: Data Science in Retail, Marketing & Customer Analytics (9)

Market basket analysis, Recommendation Systems, Customer Segmentation using Clustering, Churn Prediction Models, Sentiment Analysis for Customer Feedback.

UNIT –4: Data Science in Smart Cities, Agriculture & Education (9)

Traffic prediction, Waste management using ML, IoT-enabled Smart Farming and Yield prediction, Adaptive Learning and Student Performance Prediction, Case studies with real-world datasets.

UNIT –5: Tools, Case Studies, and Deployment (9)

Case Studies on End-to-End Projects, Tools: Python, Jupyter, Power BI, Tableau, Deploying Models: APIs, Streamlit, Flask, Interpreting Results and Storytelling, Recent Advances and Future Directions

Total Hours: 45

COURSE OUTCOMES:

On successful completion of the course- students will be able to		Bloom's Level
CO1	Explain the role and impact of data science across various application areas.	Understand (L2)
CO2	Apply suitable data science methods to domain-specific datasets.	Apply (L3)
CO3	Analyze domain-specific problems using exploratory and predictive techniques.	Analyze (L4)
CO4	Evaluate model effectiveness and insights in practical use cases.	Evaluate (L5)
CO5	Design data-driven solutions for real-world problems.	Create (L6)

SREENIVASAINSTITUTEOFTECHNOLOGYANDMANAGEMENTSTUDIES
(Autonomous)

TEXT BOOKS:

1. **Joel Grus**, *Data Science from Scratch*, O'Reilly, 2nd Edition.
2. **Cathy O'Neil & Rachel Schutt**, *Doing Data Science*, O'Reilly Media.
3. V. K. Jain**, *Data Science and Big Data Analytics*, Khanna Publishing**

REFERENCE BOOKS:

1. **Wes McKinney**, *Python for Data Analysis*, O'Reilly.
2. **Sebastian Raschka**, *Machine Learning with Python Cookbook*, Packt.
3. **Bill Franks**, *Taming the Big Data Tidal Wave*, Wiley.

REFERENCE WEBSITE:

NPTEL / SWAYAM

- NPTEL – Data Science for Engineers
- SWAYAM – Data Science Applications

SREENIVASAINSTITUTEOFTECHNOLOGYANDMANAGEMENTSTUDIES
(Autonomous)

23MRDSE6	DATA SCIENCE PRACTICE LAB (R/PYTHON)	L	T	P	C
		-	-	3	1.5

PRE-REQUISITES: Nil.

COURSE EDUCATIONAL OBJECTIVES:

- To introduce basic data analysis using R/Python.
- To understand data wrangling, visualization, and preprocessing.
- To develop skills in exploratory data analysis and simple ML models.

Experiments:

1. Introduction to Python/R and IDE setup
2. Data Types, Variables, and Operators
3. Reading and writing CSV/Excel data
4. Data Cleaning and Wrangling (nulls, types, etc.)
5. Exploratory Data Analysis with Matplotlib/ggplot2
6. Correlation and Covariance Calculation
7. Linear Regression implementation
8. Classification using KNN
9. Clustering using K-Means
10. Data Visualization using Seaborn
11. Building simple ML pipeline using scikit-learn
12. Data analysis on real-world dataset

Course Outcomes:

- Apply Python/R for real-time data handling.
- Perform data visualization and preprocessing.
- Build basic machine learning models.

SREENIVASAINSTITUTEOFTECHNOLOGYANDMANAGEMENTSTUDIES
(Autonomous)

23MRDSE7	STATISTICAL LEARNING & ML LAB	L	T	P	C
		-	-	3	1.5

PRE-REQUISITES: Nil.

COURSE EDUCATIONAL OBJECTIVES:

- To explore statistical methods for data science.
- To apply statistical learning to predictive modeling.

Experiments:

1. Introduction to probability distributions
2. Descriptive statistics on datasets
3. Hypothesis testing
4. Linear Regression and visualization
5. Logistic Regression for classification
6. Decision Trees
7. Random Forests
8. Confusion Matrix and Accuracy metrics
9. Cross-validation techniques
10. Feature scaling and encoding
11. Principal Component Analysis
12. Regression or Classification on dataset

Course Outcomes:

- Utilize statistical models in data science tasks.
- Implement regression and classification algorithms.