Semester VI

Course Code	Course Title	Credits	Lectures /Week
USCS601	**Data Science**	**2**	**3**

About the Course: This course introduces the field of Data Science, covering the fundamental concepts, techniques, and tools used for data analysis, machine learning, and data visualization. Students will learn how to preprocess and analyze data, build predictive models, evaluate model performance, and effectively communicate insights through visualizations. The course also explores data management principles and practices. Practical hands-on exercises and projects using popular Data Science libraries and technologies are included to reinforce the concepts learned.

Course Objectives:

- Understand the foundations and scope of Data Science, including its applications and comparison to related fields like Business Intelligence and Artificial Intelligence.
- Develop skills in data preprocessing, including cleaning, transforming, selecting, and merging data, to ensure data quality and suitability for analysis.
- Gain knowledge of machine learning algorithms and techniques, such as regression, classification, clustering, and ensemble learning, to build predictive models and make data-driven decisions.
- Learn how to evaluate and select models using appropriate evaluation metrics and cross-validation techniques to ensure reliable and robust model performance.
- Develop proficiency in data visualization techniques and tools to effectively communicate insights and tell compelling stories using data.

Learning Outcomes:

After successful completion of this course, students would be able to

- Apply data preprocessing techniques to clean and transform raw data, handle missing values and outliers, and merge datasets.
- Implement machine-learning algorithms to perform tasks such as regression, classification, clustering, and ensemble learning.
- Evaluate and compare different machine learning models using appropriate evaluation metrics and cross-validation techniques.
- Create informative and visually appealing data visualizations to communicate insights and patterns in data.
- Understand the principles and practices of data management, including data governance, data quality assurance, and data privacy considerations.

Unit	Topics	No of Lectures
I	**Introduction to Data Science and Data Preprocessing** **What is Data Science?:** Definition and scope of Data Science, Applications and domains of Data Science, Comparison with other fields like Business Intelligence (BI), Artificial Intelligence (AI), Machine Learning (ML), and	**15**

	Data Warehousing/Data Mining (DW-DM) **Data Types and Sources:** Different types of data: structured, unstructured, semi-structured, Data sources: databases, files, APIs, web scraping, sensors, social media **Data Preprocessing:** Data cleaning: handling missing values, outliers, duplicates, Data transformation: scaling, normalization, encoding categorical variables, Feature selection: selecting relevant features/columns, Data merging: combining multiple datasets **Data Wrangling and Feature Engineering:** Data wrangling techniques: reshaping, pivoting, aggregating, Feature engineering: creating new features, handling time-series data Dummification: converting categorical variables into binary indicators, Feature scaling: standardization, normalization **Tools and Libraries:** Introduction to popular libraries and technologies used in Data Science like Pandas, NumPy, Sci-kit Learn, etc.	
II	**Data Analysis and Machine Learning** **Exploratory Data Analysis (EDA):** Data visualization techniques: histograms, scatter plots, box plots, etc., Descriptive statistics: mean, median, mode, standard deviation, etc., Hypothesis testing: t-tests, chi-square tests, ANOVA, etc. **Introduction to Machine Learning:** Supervised learning: classification and regression, Unsupervised learning: clustering and dimensionality reduction, Bias-variance tradeoff, underfitting, and overfitting **Regression Analysis:** Simple linear regression, Multiple linear regression, Stepwise regression, Logistic regression for classification **Model Evaluation and Selection:** Techniques for evaluating model performance: accuracy, precision, recall, F1-score, Confusion matrix and ROC curve analysis, Cross-validation: k-fold cross-validation, stratified cross-validation, Hyperparameter tuning and model selection **Machine Learning Algorithms:** Decision Trees and Random Forests, Support Vector Machines (SVM), Artificial Neural Networks (ANN), Ensemble Learning: Boosting and Bagging, K-Nearest Neighbors (K-NN), Gradient Descent for optimization	**15**
III	**Model Evaluation, Data Visualization, and Management** **Model Evaluation Metrics:** Accuracy, precision, recall, F1-score, Area Under the Curve (AUC), Evaluating models for imbalanced datasets **Data Visualization and Communication:** Principles of effective data visualization, Types of visualizations: bar charts, line charts, scatter plots, etc. Visualization tools: matplotlib, seaborn, Tableau, etc. Data storytelling: communicating insights through visualizations **Data Management:** Introduction to data management activities, Data pipelines: data extraction, transformation, and loading (ETL), Data governance and data quality assurance, Data privacy and security considerations	**15**

Textbook(s):

1. Data Science from Scratch First Principles with Python- Joel Grus O'reilly, 2nd Edition
2. Advancing into Analytics From Excel to Python and R, George Mount, Oreilly, First Edition
3. Introduction to Machine Learning with Python, Andreas C. Muller, Sarah Guido, Oreilly, First Edition

Additional Reference(s):

1. Doing Data Science, Rachel Schutt and Cathy O'Neil, O'Reilly,2013
2. Mastering Machine Learning with R, Cory Lesmeister, PACKT Publication,2015
3. Hands-On Programming with R, Garrett Grolemund,1st Edition, 2014
4. An Introduction to Statistical Learning, James, G., Witten, D., Hastie, T., Tibshirani, R.,Springer,2015