SEMESTER S7

ALGORITHMS FOR DATA SCIENCE

(Common to CS/AM/CM)

Course Code	**PECST785**	**CIE Marks**	40
Teaching Hours/Week (L: T:P: R)	3:0:0:0	**ESE Marks**	60
Credits	5/3	**Exam Hours**	2 Hrs. 30 Mins.
Prerequisites (if any)	PCCST303 PCCST502	**Course Type**	Theory

Course Objectives:

1. To equip students with the ability to design, analyze, and implement advanced algorithms that are fundamental to data science, enabling them to process and analyze large-scale datasets efficiently and effectively.
2. To provide hands-on experience through real-world projects that require students to apply algorithmic techniques to solve data science problems, strengthen the development of practical skills in data manipulation, analysis, and interpretation.

SYLLABUS

Module No.	Syllabus Description	Contact Hours
1	**Foundations of Data Science Algorithms** Introduction to Data Science and Algorithms - Overview of data science and its significance, Role of algorithms in data science; Data Preprocessing Techniques - Data cleaning, transformation, and normalization, Handling missing data, outliers, and data imputation techniques; Dimensionality reduction techniques - Principal Component Analysis (PCA), t-Distributed Stochastic Neighbor Embedding (t-SNE); Algorithmic Approaches to Data Sampling - Random sampling, stratified sampling, and bootstrapping, Importance of representative sampling in data analysis. **Project 1**: Data Cleaning and Preprocessing - Develop a pipeline for cleaning and preprocessing a large, messy dataset like UCI Machine Learning Repository - Adult Data Set **Tasks:** Handle missing values, outliers, and noisy data. Apply	**9**

	dimensionality reduction techniques to simplify the dataset. Implement data transformation and normalization processes.	
2	**Algorithms for Data Summarization and Visualization :-** Data Summarization Techniques - Central tendency measures: mean, median, mode; Dispersion measures - variance, standard deviation, Interquartile range (IQR), Quantiles, percentiles, and outlier detection; Visualization Algorithms - Basics of data visualization, histograms, bar charts, scatter plots; Advanced visualization techniques - heatmaps, correlation matrices, and pair plots; Visualization tools and libraries - Matplotlib, Seaborn, Plotly; Algorithmic Approaches to Data Grouping - Clustering: k-means, hierarchical clustering, DBSCAN; Association rule learning - Apriori, FP-Growth. **Project 2**: Exploratory Data Analysis and Visualization Perform exploratory data analysis (EDA) and create visualizations to uncover patterns and insights in the dataset like Kaggle - Titanic Dataset ***Tasks:*** Summarize the dataset using statistical measures. Create various visualizations to explore relationships and patterns in the data. Implement clustering algorithms to identify natural groupings within the data.	**9**
3	**Algorithms for Data Modeling :-** Regression Algorithms - Linear regression and polynomial regression; Regularization techniques - Ridge, Lasso, Elastic Net; Evaluation metrics - RMSE, MAE, R^2; Classification Algorithms - Logistic regression, decision trees, and k-Nearest Neighbors (k-NN); Performance metrics - accuracy, precision, recall, F1-score, ROC-AUC; Algorithmic Optimization Techniques - Gradient descent and its variants: stochastic, mini-batch; Hyperparameter tuning - grid search, random search, Bayesian optimization. **Project 3**: Predictive Modeling and Evaluation - Build and evaluate predictive models using regression and classification algorithms using datasets like Kaggle - House Prices: Advanced Regression Techniques ***Tasks:*** Implement linear and polynomial regression models to predict house prices. Apply classification algorithms to classify houses into different categories. Evaluate the models using appropriate performance metrics and fine-tune them for better accuracy.	**9**
4	**Algorithms for Big Data and Scalability :-** Introduction to Big Data Algorithms - Overview of big data challenges and	**9**

	processing techniques; Distributed computing frameworks - Hadoop, Spark; MapReduce paradigm - concepts and applications; Scalable Data Processing Algorithms - Algorithms for large-scale data processing : sorting, searching, filtering; Data partitioning and shuffling techniques in distributed systems; Handling data with memory constraints - external memory algorithms. **Project 4**: Scalable Data Processing with Spark - Implement scalable algorithms using Apache Spark to process large datasets efficiently using datasets like Kaggle - Google Analytics Customer Revenue Prediction ***Tasks:*** Set up a Spark environment for large-scale data processing. Implement scalable algorithms for sorting, searching, and filtering the dataset. Analyze the performance of your algorithms on different dataset sizes and optimize for scalability.	

Course Assessment Method
(CIE: 40 marks, ESE: 60 marks)

Continuous Internal Evaluation Marks (CIE):

Attendance	*Internal Ex*	*Evaluate*	*Analyse*	*Total*
5	**15**	**10**	**10**	**40**

Criteria for Evaluation(Evaluate and Analyse): 20 marks

Assignment evaluation pattern:

- Correctness and Accuracy (30%) - Correct Solution and Implementation.
- Effectiveness and Efficiency (25%) - Algorithm Efficiency and Performance Metrics.
- Analytical Depth (25%) - Problem Understanding and Solution Analysis.
- Justification and Comparisons (20%) - Choice Justification and Comparative Analysis.

End Semester Examination Marks (ESE):

In Part A, all questions need to be answered and in Part B, each student can choose any one full question out of two questions

Part A	Part B	Total
• 2 Questions from each module. • Total of 8 Questions, each carrying 3 marks **(8x3 =24 marks)**	• 2 questions will be given from each module, out of which 1 question should be answered. • Each question can have a maximum of 3 subdivisions. • Each question carries 9 marks. **(4x9 = 36 marks)**	**60**

Course Outcomes (COs)

At the end of the course students should be able to:

Course Outcome		Bloom's Knowledge Level (KL)
CO1	Implement data preprocessing and cleaning techniques to prepare raw data for analysis, ensuring the quality and reliability of the datasets.	**K3**
CO2	Perform exploratory data analysis (EDA) and create insightful visualizations that help in understanding the underlying patterns and trends in the data.	**K4**
CO3	Develop predictive models using various regression and classification algorithms, and optimize them for better performance, applying appropriate evaluation metrics.	**K5**
CO4	Implement scalable algorithms using distributed computing frameworks like Apache Spark to process large datasets efficiently.	**K6**

Note*: K1- Remember, K2- Understand, K3- Apply, K4- Analyse, K5- Evaluate, K6- Create*

CO-PO Mapping Table (Mapping od Course Outcomes to Program Outcomes)

	PO1	PO2	PO3	PO4	PO5	PO6	PO7	PO8	PO9	PO10	PO11	PO12
CO1	3	3	3		3							2
CO2	3	3	3	3								2
CO3	3	3	3		3							2
CO4	3	3	3		3							2
CO5	3	3	3		3							2

Note: *1: Slight (Low), 2: Moderate (Medium), 3: Substantial (High), -: No Correlation*

Text Books				
Sl. No	**Title of the Book**	**Name of the Author/s**	**Name of the Publisher**	**Edition and Year**
1	Algorithms for Data Science Hardcover	Brian Steele, John Chandler, Swarna Reddy	Springer International	1/e, 2016
2	Mining of Massive Datasets	Jure Leskovec, Anand Rajaraman, Jeff Ullman	Cambridge University Press	2/e, 2020

Reference Books				
Sl. No	**Title of the Book**	**Name of the Author/s**	**Name of the Publisher**	**Edition and Year**
1	Foundations of Data Science	Avrim Blum, John Hopcroft and Ravi Kannan	Cambridge University Press	1/e, 2020
2	The Elements Of Statistical Learning: Data Mining, Inference, And Prediction	Trevor Hastie, Robert Tibshirani and Jerome Friedman	Springer	9/e, 2017
3	Data Mining: Concepts and Techniques	Jiawei Han, Micheline Kamber and Jian Pei Professor	Morgan Kaufmann	3/e, 2011
4	Data Mining and Predictive Analytics	Daniel T. Larose	Wiley	2/e, 2015
5	Hadoop for Dummies	Dirk Deroos, Paul C. Zikopoulos, Roman B. Melnyk, Bruce Brown, Rafael Coss	Wiley	1/e, 2014

Video Links (NPTEL, SWAYAM…)	
Module No.	**Link ID**
1	https://archive.nptel.ac.in/courses/106/104/106104189/ https://onlinecourses.nptel.ac.in/noc20_cs92/preview
2	https://archive.nptel.ac.in/courses/106/104/106104189/ https://onlinecourses.nptel.ac.in/noc20_cs92/preview
3	https://archive.nptel.ac.in/courses/106/104/106104189/ https://onlinecourses.nptel.ac.in/noc20_cs92/preview
4	https://archive.nptel.ac.in/courses/106/104/106104189/ https://nptel.ac.in/courses/106105186 https://archive.nptel.ac.in/courses/106/106/106106142/