

MOOC APPROVAL REQUEST

As per KTU B.Tech Regulations 2024, Section 17

KTU COURSE

Code: PECST785

Name: Algorithms for Data Science

NPTEL COURSE

Name: Data Science for Engineers

Instructor: Prof. Ragunathan Rengasamy, Prof. Shankar Narasimhan

Institution: IIT Madras

Duration: 8 Weeks

Course ID: noc26-cs65

Semester: Jan-Apr 2026

Date: December 02, 2025

Document Contents:

1. KTU Course Syllabus (Complete)
2. NPTEL Course Details
3. Syllabus Comparison Report

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	<p>Foundations of Data Science Algorithms</p> <p>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.</p> <p>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</p> <p>Tasks: Handle missing values, outliers, and noisy data. Apply</p>	9

	dimensionality reduction techniques to simplify the dataset. Implement data transformation and normalization processes.	
2	<p>Algorithms for Data Summarization and Visualization :-</p> <p>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.</p> <p>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</p> <p>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.</p>	9
3	<p>Algorithms for Data Modeling :-</p> <p>Regression Algorithms - Linear regression and polynomial regression; Regularization techniques - Ridge, Lasso, Elastic Net; Evaluation metrics - RMSE, MAE, R²; 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.</p> <p>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</p> <p>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.</p>	9
4	<p>Algorithms for Big Data and Scalability :-</p> <p>Introduction to Big Data Algorithms - Overview of big data challenges and</p>	9

	<p>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.</p> <p>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</p> <p>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.</p>	
--	--	--

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

Continuous Internal Evaluation Marks (CIE):

<i>Attendance</i>	<i>Internal Ex</i>	<i>Evaluate</i>	<i>Analyse</i>	<i>Total</i>
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
<ul style="list-style-type: none"> • 2 Questions from each module. • Total of 8 Questions, each carrying 3 marks (8x3 =24 marks) 	<ul style="list-style-type: none"> • 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
1	Algorithms for Data Science Hardcover		Brian Steele, John Chandler, Swarna Reddy	Springer International
2	Mining of Massive Datasets		Jure Leskovec, Anand Rajaraman, Jeff Ullman	Cambridge University Press

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/



DATA SCIENCE FOR ENGINEERS

PROF. RAGUNATHAN RENGASAMY

Department of Chemical Engineering
IIT Madras

PROF. SHANKAR NARASIMHAN

Department Of Chemical Engineering
IIT Madras

PRE-REQUISITES : 10 hrs of pre-course material will be provided, learners need to practise this to be ready to take the course.

INTENDED AUDIENCE : Any interested learner

INDUSTRY SUPPORT : HONEYWELL, ABB, FORD, GYAN DATA PVT. LTD

COURSE OUTLINE :

Learning Objectives :

- Introduce R as a programming language
- Introduce the mathematical foundations required for data science
- Introduce the first level data science algorithms
- Introduce a data analytics problem solving framework
- Introduce a practical capstone case study

Learning Outcomes:

- Describe a flow process for data science problems (Remembering)
- Classify data science problems into standard typology (Comprehension)
- Develop R codes for data science solutions (Application)
- Correlate results to the solution approach followed (Analysis)
- Assess the solution approach (Evaluation)
- Construct use cases to validate approach and identify modifications required (Creating)

ABOUT INSTRUCTOR :

Prior to joining IIT Madras as a professor, **Prof.Rengaswamy** was a professor of Chemical Engineering and Co-Director of the Process Control and Optimization Consortium at Texas Tech University, Lubbock, USA. He was also a professor and associate professor at Clarkson University, USA and an assistant professor at IIT Bombay. His major research interests are in the areas of fault detection and diagnosis and development of data science algorithms for manufacturing industries.

Prof.Shankar Narasimhan is currently a professor in the department of Chemical Engineering at IIT Madras. His major research interests are in the areas of data mining, process design and optimization, fault detection and diagnosis and fault tolerant control. He has co-authored several important papers and a book titled Data Reconciliation and Gross Error Detection: An Intelligent Use of Process Data which has received critical appreciation in India and abroad.

COURSE PLAN :

Week 1: Course philosophy and introduction to R

Week 2: Linear algebra for data science

1. Algebraic view - vectors, matrices, product of matrix & vector, rank, null space, solution of over-determined set of equations and pseudo-inverse)
2. Geometric view - vectors, distance, projections, eigenvalue decomposition

Week 3: Statistics (descriptive statistics, notion of probability, distributions, mean, variance, covariance, covariance matrix, understanding univariate and multivariate normal distributions, introduction to hypothesis testing, confidence interval for estimates)

Week 4: Optimization

Week 5: 1. Optimization

2. Typology of data science problems and a solution framework

Week 6: 1. Simple linear regression and verifying assumptions used in linear regression

2. Multivariate linear regression, model assessment, assessing importance of different variables, subset selection

Week 7: Classification using logistic regression

Week 8: Classification using kNN and k-means clustering

SYLLABUS COMPARISON

KTU: PECST785 - Algorithms for Data Science

NPTEL: Data Science for Engineers

KTU SYLLABUS TOPICS	NPTEL SYLLABUS TOPICS	OK
Module 1: Linear Algebra, Matrix Operations	Linear Algebra, Matrix Computations	■
Module 2: Probability, Statistics	Probability, Statistics, Hypothesis Testing	■
Module 3: Regression, Classification	Regression Analysis, Classification	■
Module 4: Clustering, Dimensionality Reduction	Clustering, PCA	■

CONTENT OVERLAP: $\geq 70\%$

The above comparison confirms that the NPTEL course content matches at least 70% of the KTU syllabus as required by R 17.4.