

IIT Mandi

Proposal for a New Course

Course Number	: IC272
Course Name	: Data Science 3 (Introduction to data preprocessing and modeling)
Credits	: 2-0-2-3 (L-T-P-C)
Prerequisites	: IC111 – Linear Algebra, IC152 – Computing and Data Science, IC252 - Probability and Statistics (Data Science 2)
Intended for	: Second year B.Tech
Distribution	: Institute core for B. Tech
Semester	: Even/Odd

Preamble: Data science involves using the scientific methods to process the data to extract knowledge and insights from data in various forms, both structured and unstructured. Vast amounts of data are being generated in many fields, and the data analyst's job is to make sense of it all: to extract important patterns and trends. This involves learning from data. This course intends to provide fundamental ideas and techniques in learning from data.

Objective: After data science - 1 (computation & data science), and data science – 2 (probability and statistics), the students acquire familiarity with statistical and mathematical properties for analyzing the data and computation involved with it. This introductory course aims to discuss about data collection from different sources, data preprocessing and provide foundations to different techniques for learning from useful information/patterns and trends from data with practical applications in various fields. This course is only at the user level, aiming to train the students in using tools for different learning techniques to predict useful information from data. The learning outcome of this course is at novice level.

Primary objective of this course are:

- Understanding different data preprocessing techniques
- Understand various machine learning (ML) algorithms such as supervised and unsupervised algorithms for tasks such as regression, classification and clustering.
- Learn to analyse data to gain insights using an appropriate ML algorithm under a given task and context.
- Analyse & solve problems from different areas and disciplines such as finance, environment, agriculture, networking, manufacturing, sports, bioinformatics, healthcare etc using ML algorithms.
- Learn to use python-based toolbox scikit-learn, a simple and efficient tool for machine learning and data modeling

Course Contents:

1. **Introduction to data collection:** Data collection process, gridding of data, data formats [4 Lectures]
2. **Data preprocessing:** Data cleaning – missing values, noisy data; data integration and transformation – normalization, standardization; outlier removal, data reduction – dimension reduction and principal component analysis (PCA) [7 Lecture]
3. **Introduction to machine learning:** Supervised and unsupervised learning, different pattern analysis tasks – classification, regression and clustering [2 Lecture]
4. **Supervised learning with applications in classification problems:** Bayes classifier with unimodal and multimodal density - maximum likelihood estimation, expectation-maximization (EM) algorithm (only at idea level), K-nearest neighbor methods, logistic regression [6 Lectures]

5. **Supervised learning - regression:** Linear regression and polynomial regression, Auto-regression [5 Lectures]
6. **Unsupervised learning algorithms - Clustering:** K-means and hierarchical clustering, density based clustering (DBSCAN) [4 Lectures]

Lab Exercises:

- Lab to be conducted on a 3-hour slot. It will be conducted in tandem with the theory course so the topics for problems given in the lab are already initiated in the theory class. The topics taught in the theory course should be appropriately be sequenced for synchronization with the laboratory.
 - o Lab1 : Data visualization – statistics from data
 - o Lab2 : Data preprocessing – missing values, normalization and standardization
 - o Lab3 : Data preprocessing – outlier identification and removal
 - o Lab4 : Data preprocessing – data reduction using PCA
 - o Lab5-7 : Classification using different techniques
 - o Lab8-9: Regression using different techniques
 - o Lab9-10: Clustering using different techniques

Textbooks:

1. C. Muller and S. Guido, *Introduction to Machine Learning with Python: A Guide for Data Scientists*, O'Reilly, 2017
2. Manaranjan Pradhan and U Dinesh Kumar, *Machine Learning Using Python*, Wiley, 2019

Reference books:

1. J. Han and M. Kamber, *Data Mining: Concepts and Techniques*, Third Edition, Morgan Kaufmann Publishers, 2011
2. S. Theodoridis and K. Koutroumbas, *Pattern Recognition*, Academic Press, 2009.
3. R. O. Duda, P. E. Hart and D. G. Stork, *Pattern Classification*, John Wiley, 2001.

Instructor:

Dileep A. D (e-mail: addileep@iitmandi.ac.in, Phone number: 9459477409)

Lead TA:

Muralikrishna H (Email: D17021@students.iitmandi.ac.in), Phone number: 9482282346

Evaluation:

- **Theory (60%):** Quiz-1 (20% of 60%), Quiz-2 (20% of 60%), Endsem exam (40% of 60%), Assignments (20% of 60%)
- **Laboratory (40%):** Weekly evaluations (20% of 40%), Lab exam-1 (30% of 40%), Lab exam-2 (30% of 40%), Mini project (20% of 40%)

Class Hours and Slot: H slot - Tuesday (11:00 – 11:50 AM), Thursday (10:00 – 10:50 AM), Friday (1:00 – 1:50)

Lab Hours: Tuesday, Thursday and Friday, 2:00 PM – 5:00 PM

Attendance Requirement: 75%

Other Requirements: 1. Switch off your mobile phone inside the class room.
2. Be punctual to the class. Late arrival is not entertained.