DS605: Fundamentals of Machine Learning
Academic Year 2025-26 (Autumn)

Instructor:	Dr. Arpit Rana (arpit_rana@daiict.ac.in)
Prerequisites:	Programming in Python (Ref.: Python for Programmers by Deitel & Deitel 2021 – Ch. 1 to 10)
Slot:	M.Sc. Data Science I Semester
Category:	Core
Course Credits:	4 Credits (3-0-2)
Lectures (1 x 3):	Monday, Tuesday, Wednesday (16:00 – 17:00 hrs.) Venue: Lecture Theatre 2
Lab Session (1 x 2):	14:00 – 16:00 hrs., Friday (Venue: M.Sc. Data Science Lab)
TA contact info:	Dhruv Panchal (202511042@daiict.ac.in) Ritwik Agrawal (202511067@daiict.ac.in)

Course Description:

This course provides a comprehensive introduction to the fundamentals of machine learning, which are essential for first-year M.Sc. Data Science students. The curriculum covers classical machine learning techniques, including supervised and unsupervised learning, decision trees, support vector machines, clustering, and outlier analysis. Students will learn the foundational concepts of learning frameworks, model evaluation, and validation. The course emphasizes practical applications and hands-on experience through projects and lab sessions, preparing students to understand and implement basic machine learning algorithms and methodologies.

Suggested Books:

- Data Mining and Machine Learning: Fundamental Concepts and Algorithms by Zaki and Meira, 2nd Edition, Cambridge University Press [2020].
- Machine Learning with Python Cookbook by Kyle Gallatin and Chris Albon, 2nd Edition.
- Understanding Machine Learning: From Theory to Algorithms by Shwartz and David [2022]

Course Outcomes:

After completing the course, the student should be able to -

- Understand the foundations of Machine Learning (ML).
- Develop skills in using recent ML techniques to solve practical problems.
- Gain experience in doing an independent study and research
- Apply all of the above in Python (using Python ML Stack: Scikit-learn, Numpy, Pandas)

P1	P2	P3	P4	P5	P6	P7	P8	P9	P10	P11	P12
X	X	X		X				X			X

DS605: Fundamentals of Machine Learning
Academic Year 2025-26 (Autumn)

Evaluation Scheme:

Ability to Understand and Apply Theoretical Concepts		Ability to implement and Solve Real Problems and Present the Solution			
In-Sem Exam (20%) [Individual Contribution]	**Endterm Exam (25%)** [Individual Contribution]	**Assignments (10%)** [Individual Contribution]	**Challenges (15%)** [Group Contribution]	**Course Project (15%)** [Group Contribution]	**Chalk Talk (15%)** [Individual Contribution]

Tentative Course Plan (this course plan is subject to change without notice.):

Units	Topics	# Lectures
Introduction	Definition, objectives, components, approaches, challenges, and case studies. Data: Structured, Semi-structured, and Unstructured. Algebraic, Geometric, and Probabilistic View of Data	3
Data Analysis Foundations	Data objects and attribute types. Measures of central tendency, dispersion, and proximity.	4
Data Preprocessing	Data Cleaning, Data Integration, Data Reduction, Data Transformation, and Discretization	2
Fundamentals of Predictive Analytics (supervised learning framework)	(i) Representation: hypothesis class—bias-variance trade-off and expressiveness-model complexity trade-off; (ii) Evaluation: consistency and generalizability; data splits; and measures of model performance; and (iii) Optimization: learning as a search: hyperparameter and parameter tuning;	9
Predictive Tasks	Regression and classification (binary, multiclass, and multi-label); Lazy learners: kNN-based regression and classification. Eager learners: Linear regression. Regularization and Polynomial regression. Logistic regression; Class-imbalance: Undersampling, Oversampling, and SMOTE	9
Classification Techniques	Linear and Non-linear Support Vector Machines: Classification using Hinge Loss; SVR. Decision Trees; Dimensionality Reduction using PCA and SVD; Ensemble Techniques: Voting Classifiers, Bagging and Pasting, Random Forests, Boosting;	9
Cluster Analysis	Basic concepts, Partitioning-based methods, Hierarchical methods, and Density-based methods; Evaluation of Clustering	3
Advanced Topics	Online learning, Contrastive learning, and Self-Supervised Learning: Active Learning with Data Augmentation.	3

DS605: Fundamentals of Machine Learning
Academic Year 2025-26 (Autumn)

Preliminary Schedule of the Course:

Week	Lecture	Lab	Assignments/ Project
Week-1 [21 July 2025]	Introduction and Data Analysis Foundations	– No lab –	-
Week-2 [28 July 2025]	Data objects and attribute types. Measures of central tendency, dispersion, and proximity	Getting familiar with Git and Python Refresher	Group formation and domain finalization through a Google form
Week-3 [04 Aug 2025]	Fundamentals of Predictive Analytics-I: Representation	Data Scraping using **`Scrapy`** and Data Preprocessing	**Lab Assignment-1:** Data Scraping and Preprocessing
Week-4 [11 Aug 2025]	Fundamentals of Predictive Analytics–II: Evaluation Holiday: 15 Aug (Friday)	-	-
Week-5 [18 Aug 2025]	Fundamentals of Predictive Analytics–III: Optimization	Data Wrangling with **`Pandas`**; Array-Oriented (Vectorized) Programming with **`NumPy`**	**Lab Assignment-2:** Vectorized Programming
	First In-Semester Exams: 20 Aug to 23 Aug 2025 (Wednesday to Saturday) – No Exam of DS605: Fundamentals of Machine Learning During this Period –		
Week-6 [25 Aug 2025]	Lazy Learners: Regression Techniques Holiday: 27 Aug (Wednesday)	Getting familiar with the **`Scikit-learn`**: Data loading to Model's Performance Evaluation	**Lab Assignment-3:** Data Preprocessing using Scikit-learn
Week-7 [01 Sept 2025]	Regression Techniques Continued... Holiday: 05 Sept (Friday)	100-Minute Development Challenge 01 Sept (Monday ⇒ Friday)	
Week-8 [08 Sept 2025]	Dimensionality Reduction	Creating a Pipeline in Scikit-learn for an End-to-End ML Project 09 Sept (Tuesday ⇒ Friday):	**Lab Assignment-4:** End-to-End Model Development using **`Scikit-learn`**
		12 Sept (Friday): 100-Minute Development Challenge	
Week-9 [15 Sept 2025]	Linear Support Vector Machines	Handling Text (Sentiment Analysis) using Classical ML: **`NLTK`**	

DS605: Fundamentals of Machine Learning
Academic Year 2025-26 (Autumn)

Week-10 [22 Sept 2025]	Non-linear Support Vector Machines and Regressor	Handling Images (MNIST and F-MNIST) using Classical ML	
Week-11 [29 Sept 2025]	Decision Tree Classifier and Regressor Holiday: 02 Oct (Thursday)	100-Minute Development Challenge	
	Second In-Semester Exams: 3 Oct to 7 Oct 2025 (Friday to Tuesday) – DS605: Fundamentals of Machine Learning Exam will be Conducted –		
Week-12 [06 Oct 2025]	Ensemble Techniques	Ensembling using **`Scikit-learn`**	**Lab Assignment-5:** End-to-End Ensemble Model using **`Scikit-learn`**
Week-13 [13 Oct 2025]	Ensemble Techniques Contd...	100-Minute Development Challenge **Release of the Course Project: (13 Oct 2025)**	
Week-14 [20 Oct 2025]	In-Semester Break: 20 Oct to 24 Oct 2025 (Monday to Friday)		
Week-15 [27 Oct 2025]	Cluster Analysis–I: Partitioning-based methods, Hierarchical methods;	100-Minute Development Challenge	
Week-16 [3 Nov 2025]	Cluster Analysis–II: Density-based methods; Evaluation of Clustering Holiday: 05 Nov (Wednesday)	Cluster Analysis using **`Scikit-learn`**	**Lab Assignment-6:** Clustering using **`Scikit-learn`**
Week-17 [10 Nov 2025]	Advanced Topics of ML	**Project Submission Deadline (15 Nov 2025)**	
Week-18 [17 Nov 2025]	Buffer/ Doubt-Clearing Sessions/ Any Remaining Topics		
Week-19 [24 Nov 2025]	End-semester Examination: 21 Nov to 29 Nov 2025 (Friday to Next Monday) – Chaik Talk will be conducted after the End-of-Semester Exams –		

Guidelines for Lab Assignments, Development Challenges, and Course Project

Lab Assignments

- Each student must select a domain of personal interest and collect a large dataset related to it.
- Additional data scraping is required to enrich the dataset for use in building models during the semester.
- By the **second week**, students must register their **domain of interest** and **team members**.
    - **Note:** No two team members can select the same domain.
    - Students from different teams may work on similar domains, but **no two students may use the same dataset**.
- Each lab session will be structured as follows:
    - The **first hour** will include a concept explanation and demo by the TAs.
    - Assignments will be released after the demo.
    - Students will have **2 hours** to complete and submit their work (Colab/Jupyter notebooks) on Google Classroom.
- **Submissions must be made from within the M.Sc. Data Science lab only**. Remote submissions are not permitted.
- **Late submissions** will be penalized:
    - Submissions beyond **6 hours** of assignment release will receive **no credit**.
- A total of **six lab assignments** will be conducted during the course.

Development Challenges and Course Project

- Students will form teams of **four members**. Each team will:
    - Participate in **five timed ML Development Challenges** (each lasting **100 minutes**).
    - Complete an **End-to-End Course Project** based on one of the domains previously worked on individually: **define their project problem**, gather relevant data, and develop a solution.
- **Team evaluation will account for 30%** of the total course marks. Choose team members carefully.
- Teams will remain **fixed for the entire semester**; no changes will be allowed after registration.
    - With 65 students, there will be **16 teams** (one team may have five members).
    - Team registration will open in **Week 2** and remain open for **three days**. Registration will be via a Google Form.
- During the challenges, **relative grading** will be used based on the highest-performing team.

Course Project Details

- Each team will choose **one thematic domain**, such as:
    - Finance, E-commerce, Healthcare, Pharmaceuticals, Sports, Entertainment, Renewable Energy, Oil & Gas, Automobile, Agriculture, FMCG, Security, Social Media, Supply Chain, or any other relevant and engaging domain.
- Each group must:
    - Define a **unique problem** within the selected domain.
    - Use **reliable and preferably public data sources**.
    - Ensure **no other group** is working on the **same dataset within the same domain**.
    - Enrich the dataset, if possible, through additional data collection.
    - Every team member must be able to explain all aspects of their course project (concepts, code, methodology, etc.).

Submission and Evaluation

- One group member must submit the **final Course Project** on **Google Classroom or GitHub Classroom**.
- Evaluation will be conducted through a **Chalk Talk**, which includes:
    - Problem definition
    - Code review
    - Dataset scale and quality
    - Tools and technologies used
    - Methodology
    - Novelty and originality of the approach
    - All group members should be ready to answer any questions related to the project.
- **Late project submissions** (within 24 hours of the deadline) will incur a **20% penalty**.

Plagiarism Policy

Plagiarism will result in **zero marks** for Lab Assignments, Challenges, and the Course Project. Plagiarism includes:

- Copying code from any external source.
- Submitting work not written personally by the student.
- Any suspected violation will be strictly penalized.