Syllabus for MA 5755
Data Analysis & Visualization in R/Python/SQL, Spring 2026

Instructor Information

Instructor:	Rakhi Singh
Class Meeting:	**Mondays from 2 to 4:45 pm (15 minutes break in between)**
Class Location:	RMN 302 (moved from KCB 504)
Office:	KCB 642
E-mail:	rakhi@smail.iitm.ac.in

Course Information

Description. This course is a lab-intensive introduction to modern data analytics, statistical learning, and data visualization techniques. The emphasis is on practical understanding, exploratory data analysis, model selection, evaluation, and effective visual communication of results. Students will work with real-world datasets using Python and will gain hands-on experience with unsupervised and supervised learning methods, ensemble models, neural networks, and visualization tools.

The course assumes prior coursework in probability, and statistics. The course also assumes sufficient knowledge of Python as this is not an "Introduction to Python" course. The focus is on modern statistical learning methods and computational workflows rather than theoretical derivations.

By the end of the course, you should expect to

- Independently perform exploratory data analysis on complex, real-world datasets using principled visualization techniques.
- Select, implement, and evaluate appropriate machine learning models for supervised and unsupervised learning tasks.
- Apply modern statistical learning methods, including ensemble models and neural networks, while accounting for generalization and overfitting.
- Interpret model outputs critically and communicate uncertainty, limitations, and insights effectively.
- Present analytical findings through clear, static, and animated visualizations suitable for technical and non-technical audiences.

Textbook(s)

I will primarily be referring to the following book, an ecopy of which is available for free download at `https://hastie.su.domains/ElemStatLearn/`

Hastie, T., Tibshirani, R., Friedman, J. (2009). The Elements of Statistical Learning: Data Mining, Inference and Prediction. Springer.

You can also use the references mentioned on our webpage (`https://math.iitm.ac.in/program-mtech-new.php`)

Richard O. Duda, Peter E. Hart, and David G. Stork. 2000. Pattern Classification (2nd Edition). Wiley- Interscience, New York, NY, USA.

Christopher M. Bishop. 2006. Pattern Recognition and Machine Learning (Information Science and Statistics). Springer-Verlag, Berlin, Heidelberg. Walpole, Ronald E.. (2012). Probability and Statistics for Engineers and the Scientists 9th ed. (9th ed.). Prentice Hall.

Prerequisites

I understand that students in this course come from diverse backgrounds and bring varying levels of experience and skill. That said, all students are expected to possess a foundational understanding of programming in general, and of Python—the language used in this course—in particular. Although we will introduce several new Python features, this course is not designed as an introductory Python class. Students without prior Python experience may find the course challenging.

Additionally, the course assumes a basic familiarity with introductory statistics concepts. These ideas will be used throughout the term, and while brief reviews will be provided, students with no prior exposure to statistics may also find the course difficult.

Course Format

This is how the course will generally be formatted.

- Weekly 3-hour laboratory session
- Each session consists of:
  - $\sim$ 50 minutes of deep-dive into the topic;
  - $\sim$ 25 minutes of guided lab;
  - $\sim$ 15 minutes break;
  - $\sim$ 75 minutes of guided lab.
- Tools include Jupyter, NumPy, pandas, scikit-learn, PyTorch/Keras, seaborn, ggplot2, and related libraries. Please setup your Google Colab and be ready to code.

Course Plan

Time permitting, this is the plan for the semester:

Week	Main Topic
1	Statistical Learning Perspective and Exploratory Data Analysis
2	Supervised learning and difference with unsupervised learning
3	Distance Geometry and K-means Clustering
4	Probabilistic Clustering and Gaussian Mixture Models
5	Hierarchical Clustering and Cluster Evaluation
6	Visualization of High-Dimensional Data
7	Regression as Prediction
8	Regularized Regression and Model Selection
9	Nonparametric Regression with Gaussian Processes
10	Distance- and Probabilistic-Based Classification
11	Tree-Based Methods and Support Vector Machines
12	Artificial Neural Networks
13	Deep Neural Networks and Convolutional Neural Networks
14	SQL and advanced visualization
15	Course Presentations

Grading

Your grade will be based on your performance on

- small **quizzes** (**10 percent**) will be conducted about five times during the class,
- about five **in-class homeworks** (**20 percent**), you cannot miss these; if they happen on the day you miss the class, you will not be allowed to submit the homework. (The dates can be worked out one week in advance)
- about six **take-home homeworks** (**30 percent**) for which about one week will be provided so that you can work on the problems at your own pace.
- **final project** (**40 percent**), more details about which will be provided in the first week of February, so that you can get started early. These will be in groups of 2 (or at most 3 students)

You are allowed to miss one quiz, one in-class homework, and one take-home homework. The best of $n-1$ exams in each of three categories (quiz, in-class homework, and take-home work) will account for the final score in that category. All of this is subject to Class Committee approval and may change based on the Class Committee's suggestions.

Academic Integrity and Use of Large Language Models

Students are expected to adhere to the highest standards of academic integrity in all coursework. This includes honesty in the preparation and submission of assignments, laboratory work, projects, and examinations.

Permissible Use of Large Language Models Large language models (LLMs) and similar AI-based tools may be used **only as learning aids**, comparable to textbooks, documentation, or online tutorials. Acceptable uses include:

- Clarifying concepts discussed in class or in assigned readings
- Understanding syntax, function usage, or documentation of programming languages and libraries
- Obtaining high-level explanations of algorithms, methods, or terminology

The final solution should reflect your own understanding. In all cases, students are responsible for ensuring that they understand and can explain any material they submit.

Prohibited Use of Large Language Models The following uses of LLMs constitute academic misconduct:

- Copying, paraphrasing, or submitting AI-generated solutions to homework, lab assignments, quizzes, or projects
- Using LLMs to generate complete or partial answers to assignment questions without independent reasoning
- Submitting code or text produced primarily by an AI system as one's own work

Work submitted must reflect the student's **own reasoning, implementation, and interpretation**, not that of an automated system.

Attribution and Responsibility If AI-based tools are used for conceptual clarification or debugging, students may be asked to explain their solution orally or in writing, modify or extend their submitted work during evaluation, demonstrate understanding during in-lab checks or discussions. Inability to do so may be treated as evidence of inappropriate use.

Enforcement Violations of this policy will be treated in accordance with institutional academic integrity regulations and may result in penalties ranging from loss of credit to disciplinary action.

Guiding Principle Students should treat large language models as **tools for understanding**, not as **sources of answers**. The goal of this course is learning and skill development; outsourcing reasoning or problem-solving defeats that purpose.