New Jersey Institute of Technology
Digital Commons @ NJIT

Chemical and Materials Engineering Syllabi | NJIT Syllabi

Fall 2024

MTEN 631-851: Data Science for CME

Rajesh Dave

Follow this and additional works at: https://digitalcommons.njit.edu/cme-syllabi

Recommended Citation
Dave, Rajesh, "MTEN 631-851: Data Science for CME" (2024). *Chemical and Materials Engineering Syllabi*. 298.
https://digitalcommons.njit.edu/cme-syllabi/298

This Syllabus is brought to you for free and open access by the NJIT Syllabi at Digital Commons @ NJIT. It has been accepted for inclusion in Chemical and Materials Engineering Syllabi by an authorized administrator of Digital Commons @ NJIT. For more information, please contact digitalcommons@njit.edu.

MTEN 631 Syllabus

Fall 2024

Course Modality:

This is an online course, which will be conducted fully online, asynchronously via Canvas. For more information on using Canvas and other supported learning tools, visit the IST Service Desk Knowledgebase.

Instructor Information

Instructor	Email	Office Hours
Rajesh Dave	dave@njit.edu	Office hours will be by appointment via Webex. Please email the instructor directly to set up a time.

*I will typically respond to direct communications, such as email, within 36 hours. Allow up to 2 weeks for feedback on submitted assignments. This feedback will be provided in Canvas.

General Information

Course Description

This is a course for graduate level students in chemical engineering, materials engineering, pharmaceutical engineering, or a related discipline. The focus is on the use of data science techniques to solve problems in chemical engineering. We will first discuss the Python programming language and how it can be used to manipulate, clean, explore, and visualize scientific datasets using the pandas package. We will cover the statistics and probability as it applies to engineering problems; this includes conditional probability, probability distributions, hypothesis testing, and Bayesian inference. Basic supervised machine learning models will be introduced, including linear and logistic regression, decision trees and random forest, and support vector machines. Students will then learn different analytical techniques and how to combine all of these skills to solve engineering problems involving large amounts of data and make predictions. Finally, we will cover how to access

data and create your own datasets; topics include databases (including relational databases such as SQL) and basic data mining and web scraping. Applications of these methods will be demonstrated in chemical engineering (e.g., optimization and controls, sensor analysis), materials engineering (e.g., structural databases and property selection), and pharmaceutical engineering (e.g., drug selection). Students will gain hands-on experience in implementing and utilizing these various methods through computational laboratory assignments and reports and a semester-long engineering design project.

Prerequisites/Co-requisites

BS degree in chemical, mechanical, electrical and biomedical engineering or in physics or chemistry.

Course Learning Outcomes

By the end of the course, students will be able to:

1. Construct and manipulate datasets and databases using Python.
2. Analyze and plot data in a variety of forms such as bar charts and scatter plots using Matplotlib.
3. Implement statistical models and learning algorithms to analyze datasets, with application to engineering systems.
4. Describe the properties of datasets using central tendencies.
5. Analyze probabilities using statistical distributions such as the normal (“Gaussian”), Poisson, and binomial distribution, with application in detection, estimation, and tracking.
6. Form statistical hypotheses and test them using p-test, constructing confidence intervals, and using Bayesian inference, with applications in engineering design.
7. Measure the strength of and describe the nature of relationships between data using linear and logistic regression.
8. Classify data and predict outcomes using decision tree methods such as random forest, with applications to robotic vision and automated navigation.
9. Perform cross-validation to prevent overtraining of models.
10. Train and evaluate unsupervised learning models to solve complex problems using unlabelled data.
11. Access, extract, and manipulate data from different types of databases.
12. Communicate effectively about data science, machine learning, and engineering through weekly discussion posts and conversations.
13. Prepare effective technical reports describing design project goals, progress, and results.
14. Disseminate results through oral presentations.

Required Materials

Textbooks
Required:

1. Practical Statistics for Data Scientists, 2nd Edition, by Bruce, Bruce, and Gedeck
2. Hands on Machine Learning with Scikit-Learn, Keras, and Tensorflow, 2nd Edition, by Geron

Grading Policy

NJIT Grading Legend

Final Grade Calculation

Final grades for all assignments will be based on the following percentages:

Laboratory Reports	**35%**
Discussion Forums	**10%**
Final Project (Interim Report 1 = 10%, Interim Report 2 = 15%, Final Written Report = 15%, Final Presentation = 15%)	**55%**
Total	**100%**

Course Work

Discussion Forums (10%): Every week a short topic for discussion about a data science topic, article, or problem will be posted on the discussion board. You need to write a response as well as respond to 2 of your classmates' posts.

Laboratory Reports (35%, 4.38% each): 8 laboratory reports are due throughout the semester. Laboratory reports should be submitted as Jupyter notebooks, INCLUDING the discussion portion, written as markup text in the cells. The first cell of the laboratory notebook should include your name and any people you collaborated with on the report. The Jupyter notebook laboratory report should be fully runnable without errors. The Jupyter notebook laboratory report is due each Sunday. The code for the report should be easily digestible and well commented, and any discussion should be clearly written.

Design Project (55%): The final project consists of students selecting an engineering application that is reliant on analyzing large data sets. A list of potential topics will be provided to you, but you are of course free to come up with your own as well. If you have difficulty finding a useful data set, the instructor can assist you. The project is an individual project. Two interim progress reports will be due throughout the semester.

- **Interim Report 1 (10%)**: The first report should summarize the problem you will tackle, why it is important, where you will find the data set, discussion regarding the suitability and useability of the data set, and any potential problems. You should also make a short video describing these topics as well in order to get practice with the video recording software you will use for your final presentation.

- **Interim Report 2 (15%)**: The second report should consist of the methodology you are using/will use to analyze this data set, why it is appropriate, and any initial results/preliminary investigations of the data set. These two reports should NOT be written in Jupyter, but instead submitted as a PDF. Code used to generate figures and results for the report should be submitted as a separate Jupyter notebook. More detail will be given later about the contents and layout of these reports.

- **Final Report (15% written report, 15% oral report)**: A final written report detailing selection of the topic, methodology, data science approach selected, and results is required at the end. A 20 minute presentation defending your selection and methodology will also be required. More details will be provided about this after interim report 2.

Feedback

I will deliver feedback on each laboratory report and project update using the comments feature in Canvas.

Letter to Number Grade Conversions

A	90-100
B+	85-89
B	80-84
C+	75-79
C	70-74
F	0-69

Exam Information and Policies

This course does not have any exams. Per the NJIT Online Course Exam Proctoring Policy, this course will use authentic assessment, meaning you will be assessed and graded on your ability to deliver real-world outputs as well as your participation and feedback to other students.

Policy for Late Work

Late assignments are allowed, but will be penalized by 10% for each day past the deadline. This applies to laboratory reports, discussion posts, and project assignments.

Academic Integrity

"Academic Integrity is the cornerstone of higher education and is central to the ideals of this course and the university. Cheating is strictly prohibited and devalues the degree that you are working on. As a member of the NJIT community, it is your responsibility to protect your educational investment by knowing and following the NJIT academic code of integrity policy.

Please note that it is my professional obligation and responsibility to report any academic misconduct to the Dean of Students Office. Any student found in violation of the code by cheating, plagiarizing or using any online software inappropriately will result in disciplinary action. This may include a failing grade of F, and/or suspension or dismissal from the university. If you have any questions about the code of Academic Integrity, please contact the Dean of Students Office at dos@njit.edu"

Netiquette

Throughout this course, you are expected to be courteous and respectful to classmates by being polite, active participants. You should respond to discussion forum assignments in a timely manner so that your classmates have adequate time to respond to your posts. Please respect opinions, even those that differ from your own, and avoid using profanity or offensive language.

Weekly Expectations

This course is organized in weekly modules. Each week, students must watch all lecture videos, complete the chapter or article readings, provide a response to the weekly discussion topic in the class forum by Thursday at 11:59pm, respond to two classmates discussion posts by Sunday at 11:59pm, and, if assigned, complete a laboratory report or project report by Sunday at 11:59pm.

Course Schedule

Week	Topic	Reading/Assignment	Due Dates
1	- What is data science? - Data science in chemical and materials engineering - Introduction to Python and Jupyter notebooks - Manipulating data: introduction to pandas	- Module 1: Discussion assigned - Read Bruce Ch 1 pp. 1-19 - Conda and Jupyter installation notes	R: Module 1: Discussion initial post Sun: Module 1: Discussion-Peer Response Sun: Module 1: Check-In Survey

Week	Topic	Reading/Assignment	Due Dates
		- Module 1: Check-In Survey	
2	- Data analytics with pandas and other packages - How can data be used to find solutions? - Data cleaning - Data visualization: introduction to matplotlib - Statistics: Central tendencies, Correlation and outliers - Practice with pandas	- Module 2: Discussion assigned - Read Bruce Ch 2 pp. 47-61, Ch 2 pp. 65-71 - Module 2: Laboratory Report 1 assigned	R: Module 2: Discussion initial post Sun: Module 2: Discussion-Peer Response Sun: Module 2: Laboratory Report 1 due
3	- Frequentist statistics - Hypothesis generation and testing - The p-test - Confidence intervals	- Module 3: Discussion assigned - Read Bruce Ch 2, pp. 69-82, Ch 3, pp. 93-139 - Module 3: Laboratory Report 2 assigned	R: Module 3: Discussion initial post Sun: Module 3: Discussion-Peer Response Sun: Module 3: Laboratory Report 2 due
4	- What is machine learning? - Linear regression - Data scaling - Overfitting and underfitting - Training and test sets - Project: Choosing a data science topic and finding data sets	- Module 4: Discussion assigned - Read Interim Report 1 topic selection and guidelines - Read Bruce Ch. 4, pp. 141-161, Ch 4, pp. 169-185 - Module 4: Laboratory Report 3 assigned	R: Module 4: Discussion initial post Sun: Module 4: Discussion-Peer Response Sun: Module 4: Laboratory Report 3 due Sun: Module 4: Reflection due
5	- Advanced regression topics - Bias and variance - Feature selection - Regularization - Hyperparameter tuning	- Module 5: Discussion assigned - Read Geron, pp. 134-141, pg. 62-79 - Interim Report 1 assigned	R: Module 5: Discussion initial post Sun: Module 5: Discussion-Peer Response

Week	Topic	Reading/Assignment	Due Dates
			Sun: Interim Report 1 due (report and video)
6	- Introduction to classification algorithms - Logistic regression - Metrics for model evaluation	- Module 6: Discussion assigned - Read Bruce Ch. 5, pp. 208 - 229 - Module 6: Laboratory Report 4 assigned	R: Module 6: Discussion initial post Sun: Module 6: Discussion-Peer Response Sun: Module 6: Laboratory Report 4 due
7	- Bayesian statistics - K-nearest neighbors - Principal components analysis - Imbalanced data	- Module 7: Discussion assigned - Read Bruce Ch 5, pp. 195-201, Ch 5, pp. 230-234, Ch 6, pp. 237-246, Ch 7, pp. 283-294 - Module 7: Laboratory Report 5 assigned	R: Module 7: Discussion initial post Sun: Module 7: Discussion-Peer Response Sun: Module 7: Laboratory Report 5 due
8	- Decision trees - Random forest - Boosting - Dimensionality reduction	- Module 8: Discussion assigned - Read Bruce Ch 6, pp. 249-279 - Module 8: Laboratory Report 6 assigned	R: Module 8: Discussion initial post Sun: Module 8: Discussion-Peer Response Sun: Module 8: Laboratory Report 6 due Sun: Module 8: Reflection due
9	- Support vector machines - Parallelization - Wrap up of supervised learning	- Module 9: Discussion assigned - Read Geron pp. 153-172	R: Module 9: Discussion initial post Sun: Module 9: Discussion-Peer Response
10	- Introduction to unsupervised learning - K Means Clustering	- Module 10: Discussion assigned	R: Module 10: Discussion initial post

Week	Topic	Reading/Assignment	Due Dates
	- Neural networks	- Read Geron, pp. 235-258, Geron, pp. 279-325 - Read Interim Report 2 guidelines	Sun: Module 10: Discussion-Peer Response
11	- Time series analysis - stationarity - ARIMA - forecasting	- Module 11: Discussion assigned - Read “Forecasting: Principles and Practice” by Hyndman and Athanasopoulos, Sections 8.1-8.5 - Interim Report 2 assigned	R: Module 11: Discussion initial post Sun: Module 11: Discussion-Peer Response Sun: Interim Report 2 due
12	- Text and data mining - Web scraping - Relational databases (SQL) - Non-relational databases - Introduction to BeautifulSoup	- Module 12: Discussion assigned - Read “Comparing database types: how database types evolved to meet different needs” by Justin Ellingwood - Module 12: Laboratory Report 7 assigned	R: Module 12: Discussion initial post Sun: Module 12: Discussion-Peer Response Sun: Module 12: Laboratory Report 7 due Sun: Module 12: Reflection due
13	- Convolutional neural networks - Natural language processing - Introduce final project	- Module 13: Discussion assigned - Read Geron, pp. 445-458, Geron, pp. 525-542 - Read Final Presentations guidelines - Module 13: Laboratory Report 8 assigned	R: Module 13: Discussion initial post Sun: Module 13: Discussion-Peer Response Sun: Module 13: Laboratory Report 8 due
14	- Discuss final project - Telling a story with data science (what makes a good report / presentation) - Interviews with data science professionals	- Module 14: Discussion assigned - Read Final Report guidelines	R: Module 14: Discussion initial post Sun: Module 14: Discussion-Peer Response

Week	Topic	Reading/Assignment	Due Dates
15	- Final presentations	- Module 15: Discussion assigned - Final Presentation assigned - Final Report assigned	R: Module 14: Discussion initial post Sun: Module 14: Discussion-Peer Response-Watch and comment on two other final presentations R: Final Presentation due
16	- Final report		R: Final Report due

Additional Information and Resources

Accessibility:
This course is offered through an accessible learning management system. For more information, please refer to Canvas's Accessibility Statement.

Requesting Accommodations:

The Office of Accessibility Resources and Services works in partnership with administrators, faculty, and staff to provide reasonable accommodations and support services for students with disabilities who have provided their office with medical documentation to receive services.

If you are in need of accommodations due to a disability, please contact the Office of Accessibility Resources and Services to discuss your specific needs.

Resources for NJIT Online Students
NJIT is committed to student excellence. To ensure your success in this course and your program, the university offers a range of academic support centers and services. To learn more, please review these Resources for NJIT Online Students, which include information related to technical support.