ENGR 301 – Engineering Applications of Data Science

Otto H. York Department of Chemical and Materials Engineering
New Jersey Institute of Technology

Instructor: Dr. Joshua Young, Assistant Professor of Chemical and Materials Engineering
- email: jyoung@njit.edu
- office: York 322

Office Hours: Tuesdays (3pm to 4pm) and Thursdays (11am to noon), starting January 23, 2024.
- Tuesday office hours will be held virtually in Professor Young's Webex room: https://njit.webex.com/meet/jyoung
- Thursday office hours will be held in Professor Young's office (please email me if you would like to come as the door to the hallway is locked if you don't have keycard access).

Date, Time, and Location: Class will meet twice a week, once for a lecture (2 hours/week) and once for a hands-on laboratory (2 hours/week).

Date: Monday and Wednesday
Time: 11:30am to 1:30pm EST.
Location: Kupfrian 210

Monday sessions will consist of a longer lecture and in-class discussion time, while Wednesday sessions will consist of a shorter lecture and hands-on coding time.

The Canvas website will be the primary place for finding and submitting assignments and for grades.

Course Description: **ENGR 301 – Engineering Applications of Data Science (2:2:0), 3 credits.** This is a course for junior level undergraduates in any engineering discipline focusing on the use of data science techniques to solve problems in engineering. We will first discuss the Python programming language and how it can be used to access, manipulate, explore, and visualize scientific datasets. We will discuss statistics and probability as it applies to engineering problems such as safety factors and probability of part failure; this includes conditional probability, probability distributions, hypothesis testing, and Bayesian inference. We will then discuss more advanced statistical models ("machine learning"), including linear and logistic regression, decision trees, and clustering. Possible applications of these methods will be demonstrated in such disciplines and topics as (but not limited to): chemical, mechanical and electrical engineering (optimization and controls), materials engineering (structure and property databases), biomedical engineering (medical diagnosis and medical imaging) and electrical and computer engineering (signal processing, target tracking, robotic navigation). 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: This course is intended for engineering majors.

Prerequisite: Any ONE of the following: CS 100 (Roadmap to Computing); CS 101 (Computer Programming and Problem Solving); CS 106 (Roadmap to Computing for Engineers); CS 113 (Introduction to Computer Science I); CS 115 (Introduction to Computer Science); BME 210 (Computing for Biomedical Engineers)

Prerequisite OR corequisite: Any ONE of the following: MATH 225 (Survey of Probability and Statistics); MATH 244 (Introduction to Probability Theory), MATH 279 (Statistics and Probability for Engineers);

MATH 305 (Statistics for Technology), MATH 333 (Probability and Statistics); ECE 321 (Random Signals and Noise)

Course Objectives: At the end of this course, students will be able to:

1. access, read, construct, and manipulate datasets and databases using Python.
2. visualize data in a variety of forms such as bar charts and scatter plots using matplotlib.
3. implement statistical models and learning algorithms in Python 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 application in decision support in engineering design, medical diagnostics, industrial manufacturing and radar.
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. analyze unlabeled data through the use of unsupervised learning algorithms (*i.e.*, clustering), with applications in nondestructive testing.
10. perform cross-validation to prevent overtraining of models.
11. prepare an effective technical report describing design project goals, progress, and results.
12. disseminate results through oral presentations to classmates.

This course addresses the following ABET student outcomes: 1, 2, 3, 5, 7

Learning Materials:

Textbook: Practical Statistics for Data Scientists, 2nd Edition by Bruce, Bruce, and Gedeck.

Hardware: A working computer is required.

Software: Python3 will be used throughout the class, and assignments can be written in Google Collaboratory or a desktop Jupyter notebook. Week 1 will describe how to set these up, and they are required to be set up by Monday of Week 2.

Grading: The final grade for the course is divided as follows:

- Laboratory Reports = 40% of grade (7 due throughout the semester, ~ 6% each)
- Class participation = 5%
- Interim design update reports = 25% of grade (2 due throughout the semester)
  - Report 1, selection and justification of engineering topic and challenges = 15%
  - Report 2, discussion of methodology and initial results = 15%
- Final Project = 30% of grade
  - Written report = 15%
  - Oral presentation = 15%

Grades will be assigned with the following rubric:

90% and above	A
85-89%	B+
80-84%	B

75-79%	C+
70-74%	C
60-69%	D
Below 60%	F

Class Participation: Occasionally during the lecture, there will be a short discussion session. You will sometimes be required to read an article as part of the homework. Participation in these sessions counts towards your final grade.

Laboratory Reports: On weeks a laboratory report is assigned, the assignment will be published Tuesday night before the Wednesday lab session. Wednesday in-class sessions will occasionally consist of a short lecture, with the rest of the time dedicated to working on the lab alone or in groups. The instructor will be available to answer questions and clear up topics. Some laboratory sessions will be reserved for working on your final project.

Laboratory reports should be submitted to Canvas as a Word document, and code submitted as a Jupyter notebook (.ipynb extension). The first cell of the laboratory notebook should include your name, UCID, and any people you collaborated with. The Jupyter notebook laboratory report should be fully runnable without errors. An example report and code will be provided on the first day of class.

The laboratory report is due the FOLLOWING TUESDAY AT 11:59pm. The code for the report should be easily digestible and well commented, and any discussion should be clearly written.

Design Project: The final project consists of an individual project in which you choose 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.

Two interim progress reports will be due throughout the semester. 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. 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.

A final written report detailing selection the challenge, methodology, data science approach selected, and results is required at the end. A 15 minute presentation defending your selection and methodology will also be required: 10 minutes for presenting and 5 minutes for answering questions. More details will be provided about this after interim report 2. Everyone will present either Monday or Wednesday of Week 15. The final report is due Monday of Finals Week.

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 academic code of integrity policy that is found at: http://www5.njit.edu/policies/sites/policies/files/academic-integrity-code.pdf.

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.

Code may be provided to you for this course but under no circumstances should it be distributed outside of the course without the express written consent of the instructor. This includes uploading to generative AI models.

Generative AI Policy: The use of generative AI as a tool to advance learning is allowed for this course. However, the use of generative AI as a shortcut to bypass the learning process is not permitted. Specifically, using content (ideas, words, processes, code, and results) not written by oneself (including by generative AI) and sharing it as one's work is considered plagiarism in the context of this course and will be reported as detailed under Academic Integrity. Furthermore, generative AI is not considered a source of information; specific information cited should be accompanied with by a peer-reviewed academic publication or equivalent.

Detailed Schedule:

Week	Topics	Assignment
1	- What is data science? - Data science across engineering disciplines - Introduction to Python and Jupyter - Manipulating data: introduction to *pandas* - NO CLASS MONDAY	No assignment
2	- Data visualization: introduction to *matplotlib* - Statistics: Central tendencies - Statistics: Correlation and outliers - Probability Basics	Laboratory Report 1 assigned
3	- Hypothesis generation and testing - The p-test - Confidence intervals	Laboratory Report 1 due Laboratory Report 2 assigned
4	- What is machine learning? - Overfitting and underfitting - Training and test sets - Linear regression	Laboratory Report 2 due Laboratory Report 3 assigned
5	- Advanced linear regression - Regularization - Feature selection	Laboratory Report 3 due Interim Report 1 assigned
6	- Logistic regression - Evaluating classification models	Interim Report 1 due Laboratory Report 4 assigned
7	- Naïve Bayes classification - Principal components analysis	Laboratory Report 4 due Laboratory Report 5 assigned
8	- K-Nearest Neighbors algorithm - Variable encoding	Laboratory Report 5 due No assignment
9	SPRING BREAK, NO CLASS MONDAY OR WEDNESDAY	
10	- Decision Trees	Laboratory Report 6 assigned

	- Random Forest - Bagging and Boosting	
11	- Introduction to unsupervised learning - Clustering algorithms	Laboratory Report 6 due Interim Report 2 assigned
12	- Neural networks	Interim Report 2 due Laboratory Report 7 assigned
13	- Text and data mining - Databases - Large language models - Introduction to *BeautifulSoup*	Laboratory 7 due
14	- Time series analysis - Relational databases (SQL) - Open questions	
15	- Final presentations	Final Presentation due
16	- Industry and visit day - NO CLASS WEDNESDAY	
17 - FINALS	- Final report due	Final Report due