# Statistics, Machine Learning, and Data Mining

# Syllabus for Ma322 Section 802 Spring 2025

## Info

3 Credits
Monday 3:10am - 6:00pm
Room

## Instructor Information

calvin_williamson@fitnyc.edu
office: B831 Science and Math
office hours: T 5-6, W 11-12, R 10-12

## Description

This is an introduction to statistical techniques for machine learning and data mining. It emphasizes mathematical methods and computer applications related to automated learning for prediction, classification, knowledge discovery and forecasting in modern data science. Special emphasis will be given to the collection, mining, and analysis of massive data sets. (G2: Mathematics) Prerequisite(s): MA 222 and mathematic proficiency (see beginning of Mathematics section)

## Outcomes

Upon completion of this course students will be able to:

1. Describe the concepts of machine learning and identify examples of its use in data science.
2. Employ statistical software to collect data, create training and test sets, and perform predictions.
3. Create regression models for predicting outcome variables in terms of predictors.
4. Explain the contributions of Google in understanding web scale data and the structure of the internet.
5. Identify the characteristics of massive data sets and describe the tools needed to analyze them.
6. Analyze decision tree models and display them with appropriate graphics.
7. Use recommendation systems software and understand how it makes suggestions based on similarity measures.

8. Perform classifications for data sets using nearest neighbor and probabilistic algorithms.
9. Collect text data and use text mining software to perform sentiment analysis.

## Course Materials

We will be using Google Colab and Google Sheets for all work in this course. Since these are web-based applications there is NO OTHER SOFTWARE required for the course besides a web browser.

## Topics

Regression

- Simple Regression
- Multiple Regression
- Applications
- Conjoint Analysis

Introduction to Python

- Google Colab Notebook
- Using LLM as Coding Assistant
- Calculations
- Variables
- DataTypes
- Lists
- Dictionaries
- Functions
- Dataframes
- f-Strings

Introduction to Large Language Models (LLMs)

- LLM Examples
- ChatGPT, GPT-4o, Gemini, Claude, Mistral
- Completions, APIs

Prompt Engineering

- Prompting
- Prompt Chaining
- Roles and Personas
- Chain of thought

- Few-shot and zero-shot Learning

Machine Learning

- Classification, Accuracy
- Training, Testing
- Decision Trees

## Evaluation

Your grade will come from these parts:

- Quizzes (85%)
- InClass/Homework Credits (15%)

Each of these parts is described in more detail below

### Quizzes

Your quiz grade will come from 5 quizzes roughly covering 2 or 3 weeks material eachThis quizzes are 30-45 minutes each and are usually 5 or 6 questions each.These quizzes are with no notes, no internet, no phone, no software, no AI tools.Pen and paper and calculator only. They are some multiple choice, some short answer, some true false.

### Problem Credits (2 or 3 per class)

Problem credits are credits you obtain for demostrating you have completed assigned problems. Some of these will come from homework assignments that you show me at the beginning of the class, some of these will come from in class assignments that are done during class and you show as you complete them. You will earn 1 credit for each successful problem completion. You must be in attendance to earn problem credits.

### DataCamp Courses

This course will use some material at datacamp.com, a website for learning data science. You will have free login to the site for 6 months beginning the first day of our course. You will be completeing 2 or 3 required courses in datacamp for topics related to our course. But you can also take any of the courses in datacamp for free, related to our course or not. There are many courses at all levels from beginner to more advanced about data science, programming, AI, etc.

There is NO FINAL EXAM.

### AI Policy

All uses of chatbots are encouraged, and there is no restriction on their use. This is especially for topics about large language models (ChatGPT, Gemini, Claude, etc).