STAT 724: Intro in Data Science and Machine Learning, **Spring 2020**

General Information. *Class meets*: W 5:35-7:25pm, ONLINE
Instructor: Iordan Slavov
Office Hours: Th 5:30-7:30pm, or by appointment.
Email: iordan.slavov@hunter.cuny.edu

Description. The course gives an overview, from Statistical point of view, of Data Science and its core Machine Learning models and algorithms. It provides detailed knowledge of how these methods work and how statistical models are applied to analyze large data sets. The focus is on the important tasks of classification and regression (so called Supervised Learning) and clustering and anomaly detection (Unsupervised Learning). The data analysis friendly approach requires the use of a good statistical software. Python, which is the most popular open source scripting language for Data Science at the moment, will be used for analysis and modeling. The syntax and the environment will be discussed in enough detail. Code in Python (and R) will also be provided.

Required textbooks
[**ISLR**] An Introduction to Statistical Learning with Applications in R by Gareth James, Daniela Witten, Trevor Hastie and Robert Tibshirani, Springer (2013), ISBN 978-1-4614-7137-0
http://www-bcf.usc.edu/~gareth/ISL/ISLR%20Seventh%20Printing.pdf

[**ESLII**] The Elements of Statistical Learning - Data Mining, Inference, and Prediction by Trevor Hastie and Robert Tibshirani and Jerome Fridman, Springer (2016), ISBN-13**:** 978-0387848570.
https://web.stanford.edu/~hastie/ElemStatLearn/printings/ESLII_print12.pdf

Recommended:
[**HOML**] Hands-On Machine Learning with SciKit-Learn, Keras and TensorFlow by Aurelien Geron, 2nd Edition, O'Reilly Media (2019), ISBN: 978-1-492-03264-9

Prerequisites. STAT 706 (General Linear Models I) or instructor's permission

Grading. You will be graded on in-class participation (5%), about 5 homework assignments and 1 mini project (40%), an "in-class" midterm (20%) and a final take-home exam (35%). The final paper (exam) will involve the analysis of a non-trivial dataset of your choice (but approved by me). The lowest homework score will be dropped so late homework will not be accepted.

Policy on Cheating. Hunter College regards acts of academic dishonesty (e.g., plagiarism, cheating on examinations, obtaining unfair advantage, and falsification of records and official documents) as serious offenses against the values of intellectual honesty. The college is committed to enforcing the CUNY Policy on Academic Integrity and will pursue cases of academic dishonesty according to the Hunter College Academic Integrity Procedures.

Important Dates. No classes . Last class: . Final projects due .

Material Covered (based on chapters in [**ISLR**])
[1] Linear Regression, K-Nearest Neighbors; [2] Classification – Logistic Regression, Linear Discriminant Analysis; [3] Resampling Methods – Cross-Validation, Bootstrap; [5] Model Selection – Subset Selection, Shrinkage Methods, Dimension Reduction Methods; [6] Tree-Based Methods; [7] Support Vector Machines; [9] Neural Networks; [10] Unsupervised Learning

Time permitting, additional topics will be added. (Syllabus subject to change, 8/26/20)