THE GEORGE WASHINGTON UNIVERSITY
WASHINGTON, DC

DATS 6103 Introduction to Data Mining
CRN 73984
Tuesday 6:10 PM 8:40 PM

INSTRUCTOR:
Name: Amir Jafari, PhD
Term: Spring 2020
Campus address: ROME 352
E-mail: ajafari@gwu.edu
Office hours: Before Class - Samson Hall Room 315

COURSE DESCRIPTION:

This course is an introductory course on data mining. It introduces the basic concepts, principles, methods, implementation techniques, and applications of data mining, with a focus on Python and data mining algorithms for the Data Science Program. The objective of the course is to give students an overview of data mining techniques and skills to explore, analyze, and leverage data. Due to the diversity of subjects that comprise this emerging field, the class will necessarily have more breadth than depth. At the beginning of the course we will cover python to perform pre-processing and data wrangling then in the next half of the course we will cover 'core' data mining topics, such as regression and classification techniques. Students will use Python to complete the homework, assignments and projects through the course.

LEARNING OUTCOMES: Students will be able to:

1. explain data mining algorithms and concepts.
2. demonstrate knowledge of data mining techniques (pre-processing, feature section) using Python.
3. apply generalizations and model validations on practical data sets .
4. explain the core concept of regression and classification.
5. demonstrate visualization techniques for variety of applications.

RESOURCES:

A- Think Python - Author: Allen Downey , Ebook Web Link.
B- An Introduction to Statistical Learning - Author: G. Casella- Free Ebook Web Link.
C- The Elements of Statistical Learning Data Mining - Author: Trevor Hastie, Ebook Web Link
D- Data Preprocessing in Data Mining - - Author:Garca, Salvador, Ebook Web Link

SOFTWARE:

Python is required for all homework assignments and the project. Pycharm professional will be used as the editor for Python. Students require to install Pycharm professional (get the licence by edu email).

TENTATIVE COURSE OUTLINE (SUBJECT TO CHANGE):

Week	Topic	Comments
January 14, 2020	Software Setup, Python Programming (Basic)	
January 21, 2020	Python Programming (Intermediate)	
January 28, 2020	Python Programming (Advance)	Quiz 1
February 4, 2020	Python Visualizations	
February 11, 2020	Data Wrangling	Quiz 2
February 18, 2020	Numerical and Scientific Python	Quiz 3
February 25, 2020	Preprocessing	Quiz 4
March 3, 2020	Exam 1	
March 10, 2020	Decision Tree & Random Forest	
March 17, 2020	Spring Break (no classes)	
March 24, 2020	Support Vector Machine	Quiz 5
March 31, 2020	K Nearest Neighbor - Naive Bayes	
April 7, 2020	Clustering: K-means,hierarchical	Quiz 6
April 14, 2020	Exam 2	
April 21, 2020	**Final Project Presentation** and Submission	

PREREQUISITES: DATS 6101 or equivalent - Introduction to Data Science

GRADING AND EXAMINATION POLICY:

- 2 Exam - 25 pts each
- Homework/Labs - 25 pts
- Quizzes - 25 pts
- 1 Final project - 25 pts

The top three scores of quizzes, total homeworks/labs and 2 exams will be added to the final project score to obtain the total grade for the course (out of a total of 100 pts). All exams and quizzes may be in class or take home. I may collect homeworks or give a quiz (most probably there is a quiz after every 2 weeks). No make-up exams unless previous arrangements have been made. Students will be expected to attend class and prepare assignments. Habitual failure to do so will result in a reduced grade. An incomplete grade will only be given when a student misses a portion of the semester because of illness or accident. Cheating on examinations, plagiarism and other forms of academic dishonesty are serious offenses and may subject the student to penalties ranging from failing grades to dismissal.

AVERAGE AMOUNT TIME LEARNING PER WEEK:

Students are expected to spend a minimum of 100 minutes of out-of-class work for every 50 minutes of direct instruction, for a minimum total of 2.5 hours a week. A 3-credit course should include 2.5 hours of direct instruction and a minimum of 5 hours of independent learning or 7.5 hours per week.

ASSIGNMENT DESCRIPTION:

The labs and homework will be associated with each module; there will be lab exercises for each module that covers a framework; and there may be one project for each deep network type. The exam will cover all the LAB exercises and homework and quizzes. George Washington University has a Amazon Web Services (AWS) cloud account with NVIDIA compatible GPUs and I will give 50 dollar credit for Google Cloud Platform (GCP). The mini and final projects should be done on these systems.

SECURITY:

In the case of an emergency, if at all possible, the class should shelter in place. If the building that the class is in is affected, follow the evacuation procedures for the building. After evacuation, seek shelter at a predetermined rendezvous location.

DISABILITY SUPPORT SERVICES (DSS):

Any student who may need an accommodation based on the potential impact of a disability should contact the Disability Support Services office at 202-994-8250 in the Marvin Center, Suite 242, to establish eligibility and to coordinate reasonable accommodations.
The University Counseling Center (UCC Phone: 202-994-5300) offers 24/7 assistance and referral to address students' personal, social, career, and study skills problems Web Link. Services for students include:

- crisis and emergency mental health consultations
- confidential assessment, counseling services (individual and small group), and referrals

ACADEMIC INTEGRITY:

The code of academic integrity applies to all courses in the George Washington School ("Academic dishonesty is defined as cheating of any kind, including misrepresenting one's own work, taking credit for the work of others without crediting them and without appropriate authorization, and the fabrication of information."). In the spirit of the code, a student's word is a declaration of good faith acceptable as truth in all academic matters. Cheating and attempted cheating, plagiarism, lying, and stealing of academic work and related materials constitute Honor Code violations. These will not be tolerated. Please become familiar with the code. All students are expected to maintain the highest level of academic integrity throughout the course of the semester. Please note that acts of academic dishonesty during the course will be prosecuted and harsh penalties may be sought for such acts. Students are responsible for knowing what acts constitute academic dishonesty. The code may be found at Web Link

UNIVERSITY POLICIES:

Students should notify faculty during the first week of the semester of their intention to be absent from class on their day(s) of religious observance. Faculty should extend to these students the courtesy of absence without penalty on such occasions, including permission to make up examinations. Faculty who intend to observe a religious holiday should arrange at the beginning of the semester to reschedule missed classes or to make other provisions for their course-related activities.

EMAIL ETIQUETTE:

In the age of technology, when most forms of communication are electronic, it is important to adopt a proper etiquette to communicate with one another. It is asked that students use salutation when sending emails to their instructors and also make sure to SIGN their name and include their class/section at the end of the email. The instructor reserves the right NOT to reply to emails that are not properly addressed or do not have a signature. Students should also use their GWU email for any correspondence with the instructors. Students are required to check their emails daily and especially the morning before class.

COURSE CONTENT (SUBJECT TO CHANGE):

Week	Module	Topics
January 14, 2020	Software Setup, Python Basic	Pycharm Editor Variable, Expersions Statements Conditionals and Recursions Iterations
January 21, 2020	Python Programming Intermediate	Functions Writting Functions List, Tuple and Dictionaries List Processing
January 28, 2020	Python Programming Advance	Files Classes and Objects Classes and Functions Classes and Methods
February 4, 2020	Python Data Visualization	Numpy Matplotlib Seaborn
February 11, 2020	Data Wrangling	Pandas Join Data Frames Statistics (Mean, Variance) Imputation Categorical Encoding
February 18, 2020	Numerical and Scientific Python	Scipy Multidimensional Array Linear Algebra Geopands
February 25, 2020	Pre processing	Cleaning Missing Values Noise Data Reduction Feature Selection
March 3, 2020	Exam1	
March 10, 2020	Decision Tree & Random Forest	Gini Entropy Boosting Sampling with Replacement Ensemble Average
March 17, 2020	Spring Break (no classes)	

Week	Module	Topics
March 24, 2020	Support Vector Machine	Linear Kernel Lagrange Multiplyers Radial Base kernel
March 31, 2020	K Nearest Neighbor - Naive Bayes	Euclidian Distance Metrics Probabilty Joint Distribution
April 7, 2020	Clustering,K-means,hierarchical	Distance Metrics Curse of Dimentionality
April 14, 2020	Exam2	
April 21, 2020	Final Project Presentation and Submission	