Health Informatics Program, College of Public Health

Spring 2026

Syllabus	
Course information	HI 880: Advanced Health Data Mining
Format	**Online**: Tuesday, 7.20pm, Required Weely Zoom Sessions: https://gmu.zoom.us/j/95961375684?pwd=kwEPyC4vLytH4Y1inyao5Ua6GPimbi.1 The weekly meetings are mandatory and count towards course participation. These are camera-on meetings.
Course placement	**( )** Core **(X)** Concentration **(X)** Elective in HSR PhD (Knowledge Discovery and Health Informatics Concentration), Health Informatics MS (Data analytics Concentration), Health Informatics and Data Analytics graduate certificate. **(X)** Pre-requisite(s): HAP 618, HAP 719, HAP 780 or permission of instructor **This is a follow up course to HAP 780**. It includes advanced topics that require both knowledge of basic data mining techniques and statistical methods. Before enrolling, students must be fluent in SQL, know basic data mining concepts, and have programming skills in Python. Students should understand the basic structure of claims data and EHR data as well as their use in administrative and clinical settings.
Instructor	Dr. Janusz Wojtusiak jwojtusi@gmu.edu Peterson Hall, Room 4425, Fairfax Campus Appointments for office hours by email. Meetings online or in person on Fairfax Campus.
Course description	An advanced course in health data mining. Includes knowledge and skills needed to analyze health data using modern tools. Describes analytics of administrative and

	clinical data. Covers concepts and tools for big data analytics and NoSQL data analytics.
Course objectives	**Upon completion of the course, students will be able to:** 1. Understand application of data mining and analytics techniques applied to administrative and clinical data 2. Analyze data using modern open source analytics tools 3. Understand big data platforms and tools needed for analyzing health data 4. Apply data mining tools to administrative and clinical data 5. Use selected tools for NoSQL data analysis, including graph data
Course requirements	**Computer requirements** This is a computationally intensive course and you are expected to access databases, software tools, and other contents. You will need: • Fast computer (multicore PC or Mac) with at least 200GB+ of free disk space and at least 16GB RAM (32GB+ highly recommended), Windows 11, Mac or Linux. • Fast internet connection • Microsoft office for viewing and preparing files • Ability to install other software provided in class (Python/Anaconda, PostgreSQL, Neo4j) **Expectations:** **Students are responsible for assigned readings, class content and material. Students are also responsible for finding right computer equipment that allows accessing the course materials, using data and software tools, and for checking email/blackboard on daily basis.** **Data mining is a very broad topic, which is condensed here into one semester course. In addition to 3-hour class meetings, this course requires students to participate in lectures and spend at least another 6 hours per week on assignments, reading, and project.** **Evaluation Methods:** If you are taking this course as part of a graduate level course, you will receive a grade. Your grade will depend on your participation, quality of your project work and your team work. Assignments and projects are graded based on multiple criteria that will be discussed in detail. **Always write all answers in own words. Do not copy-and-paste.** It is normal practice to search the internet for source code in discussion boards or tutorials when programming. It is also normal practice to use **AI-based tools** to help with answers. However, the submitted work needs to be yours. If any of the code or response is copied or generated, you must clearly indicate which portions of the code and what are sources. When AI tools are used, clearly state what tools and what prompts you used to arrive at answers. Failing to do so is considered honor code violation.

Late submissions may be penalized up to 20% of the grade.

You can ask questions by sending to the instructor. In most cases you will receive response within 24-48 hours. If you don't get response, please resend your email/message. You may request 1-1 meeting with the TA to help with content. You may request 1-1 meeting with the instructor to discuss projects.

Data Access

Students need to obtain access to MIMIC IV dataset through Physionet. Gaining access takes long time, and you should apply and submit required documents in the first week of class the latest. Please check Blacboard for instructions on how to get access.

Other datasets you may want to use for your individual project also may require several steps to complete. Access to most data hosted at Mason requires completion of CITI training (www.citiprogram.org). After creating an account and selecting GMU as organization, complete the Group 2 Biomedical Research Investigators course.

Several real datasets are available in the DSHI center at GMU and access requires additional background check and security training. This takes time to complete, so start working on data access early. Most of the datasets require additional training and paperwork to access. These data do not leave the computing infrastructure in the center and need to be analyzed within its Linux environment (with all tools covered in class available).

Final Project

Data mining requires combining theoretical knowledge with practical skills. In order to develop skills in the context of health care applications, semester-long project is the most important component of the grade.

The project topics should be related to analyzing healthcare data in order to solve clinical or administrative problems. The project report should include, but be not limited to: (1) problem description; (2) data selection; (3) data pre-processing; (4) selection DM methods; (5) application of methods; (7) detailed model evaluation; (8) analysis of results; (9) review of available literature and related work; (10) conclusions and description of impact on healthcare. Brief description of what you learned in the project. The project should be based on tools in which you need to 'write code' to analyze data, i.e., Python, not tools with GUI such as Weka or RapidMiner. The project is expected to focus on methods covered in class, but can go beyond them.

Direct application of existing software to publicly available datasets is not sufficient. The projects must demonstrate significant efforts in data manipulation, processing, and mining. Projects must also illustrate understanding of applied techniques as well as understanding healthcare problem being solved.

	1. Do not use data from UCI, Kaggle or similar repositories/competition sites. 2. If you use (copy or adapt) any source code from online repositories or written by someone else (such as github) you must note every part of code that copied with appropriate sources. Within your Jupyter Notebooks you should provide exact source of source code you used, even if modified. Use comments to do so. 3. If you use any conde generated by AI-based tools, you need to clearly indicate how this is done and which parts of code are yours ad which are not. 4. If your project follows an online tutorial or similar resource you must clearly indicate what is your contribution to the problem. The project will be graded on your contribution. 5. This is individual project and must be distinct from what other students do. If multiple similar projects are submitted, you will be graded only based on what is unique contribution of your specific work. 6. This is class in health informatics program and the models you create must be connected to real health applications. You must demonstrate understanding on health data and application area to make the model usable. 7. Your project must be based on methods and approaches covered in class. Simple application of data science methods you learned elsewhere will result in a failed project. Your project will be judged based on the material you learned in this class. 8. There are tools to automatically generate source code to solve specific problems. While it is reasonable to use these tools to generate small parts of the source code, these parts need to be clearly marked (within jupyter notebook). 9. You will be asked detailed questions about your project, source code, implementation and results. This includes detailed knowledge of all programs and solutions that you used. For example, if you copied a program from github, or used one generated by an AI-based tool, you must know what it does and how it works.
Required textbook(s) and/or materials	**Required Text:** Class notes and slides. **Additional Readings:** Readings will be provided weekly.
Teaching methods	(X) Lecture ( ) Group work ( ) Independent research ( ) Field work ( )Papers ( ) Guest speakers ( ) Student presentations ( ) Case Studies (X) Lab ( ) Class discussion ( ) Other _____

<table>
<tr><td>Evaluation</td><td><table>
<tr><td>Final Exam</td><td>30%</td><td></td></tr>
<tr><td>Weekly Assignments</td><td>20%</td><td></td></tr>
<tr><td>Semester-long project</td><td>40%</td><td></td></tr>
<tr><td>Participation</td><td>10%</td><td></td></tr>
</table></td></tr>
<tr><td>Grading Scale</td><td>96+ A
90-95 A -
86-89 B +
80-85 B
75-79 B -
70-74 C
0-70 F</td></tr>
<tr><td>Mason Honor Code</td><td>The complete Honor Code is as follows:
To promote a stronger sense of mutual responsibility, respect, trust, and fairness among all members of the George Mason University community and with the desire for greater academic and personal achievement, we, the student members of the university community, have set forth this honor code: Student members of the George Mason University community pledge not to cheat, plagiarize, steal, or lie in matters related to academic work.

(catalog.gmu.edu)</td></tr>
<tr><td></td><td>Posting course materials on outside websites is a violation of honor code. Do not copy materials from Canvas or those sent by the instructor and share with anybody outside the course without permission of the instructor.

Posting your responses to assignments or sharing with others is a violation of honor code.</td></tr>
<tr><td>Individuals with Disabilities</td><td>The university is committed to providing equal access to employment and educational opportunities for people with disabilities. Mason recognizes that individuals with disabilities may need reasonable accommodations to have equally effective opportunities to participate in or benefit from the university educational programs, services, and activities, and have equal employment opportunities. The university will adhere to all applicable federal and state laws, regulations, and guidelines with respect to providing reasonable accommodations as necessary to afford equal employment opportunity and equal access to programs for qualified people with disabilities. Applicants for admission and students requesting reasonable accommodations for a disability should call the Office of Disability Services at 703-993-2474. Employees and applicants for employment should call the Office of Equity and Diversity Services at 703-993-8730. Questions regarding reasonable accommodations and discrimination on the basis of disability should be directed to the Americans with Disabilities Act (ADA) coordinator in the Office of Equity and Diversity Services.

(catalog.gmu.edu)</td></tr>
<tr><td>E-mail Policy</td><td>Mason uses electronic mail to provide official information to students. Examples include notices from the library, notices about academic standing, financial aid information, class materials, assignments, questions, and instructor feedback.</td></tr>
</table>

	Students are responsible for the content of university communication sent to their Mason e-mail account and are required to activate that account and check it regularly. Students are also expected to maintain an active and accurate mailing address in order to receive communications sent through the United States Postal Service. *(catalog.gmu.edu)*

Tentative Weekly Schedule

The schedule below is approximate and may be changed to adapt to students' needs and requests, new material, and for other reasons. In general all assignments are due Sunday 11.59pm before the next class meeting.

Date	Topics
1/20	Review of Data Mining Methods for Healthcare Applications Introduction to Data Mining in Python. Pandas, numpy, and sklearn, libraries. Reading the data: text files, SQL, XML, JSON. Manipulating & exploring the data.
1/27	Classification & regression, medical claims data
2/3	Classification & regression, hyperparameter tuning
2/10	Model evaluation part 1
2/17	Model evaluation part 2
2/24	Clinical data, MIMIC
3/3	Model evaluation part 3 – SHAP, guest lecture
3/10	**Spring Break**
3/17	Clinical data, MIMIC part 2
3/24	More clinical data, MIMIC part 3
3/31	Team Project on MIMIC data
4/7	Team Project on MIMIC data
4/14	Deep Learning, EHR and claims data
4/21	Deep Learning, EHR and claims data
4/28	Graph Databases, Neo4j
5/5	Final Project Presentations, Project reports and source code due on 5/4
5/12	**Final Exam Due**