USC Viterbi
School of Engineering

ISE 599: Introduction to Health Analytics

3 Units
Day/Time: Thursday 2:00 – 4:50 PM
Location: TBA

Instructor: Abigail Horn
Office:

Office Hours: TBA (in person)
Virtual office hours by appointment

Contact Info: Abigail.Horn@usc.edu

Teaching Assistants:
TBA

Catalog Course Description
A case-study based approach to learning about descriptive and predictive data analytical tools to use in combination with various health data types, to derive inspiring solutions to real-world problems in various health domain application areas.

Expanded Course Description
Health decisions can have life-altering outcomes for patients and populations. The increased ability to collect health-related data, combined with innovations in data analytics techniques, are revolutionizing the ability of decision-makers to make evidence-based choices that can dramatically improve patient and population health, well-being, and experience, and impact healthcare costs. Health data analytics, at the intersection of statistics, programming, and healthcare, provides a foundation for transforming data into insights and predictions to inform sound, impactful decision-making for problems across many health-domain problems.

Through a survey of health analytics case studies, this course will overview different types of health domain applications, health data, and descriptive and predictive analytical techniques available to address health-domain problems, using the R programming language. The primary objective will be to provide students with the skills and experience to implement existing data analytics techniques to gain insights and make predictions with health data, to interpret results to determine possible decisions and their advantages and disadvantages, and to communicate these findings in the context of achieving results in health.

The course is intended for M.S. students in an analytics or data science master's program who are interested in the health domain, or a health sciences master's program and have experience in data analytics.

Learning Objectives

The overall course objective is to learn to identify, apply, and interpret appropriate analytic methods and health data types to address real-world health-domain problems.

- Students will become familiar with health data types including electronic medical records, medical text, medical claims data, disease registry data, healthcare survey data, and social and behavioral data. They will understand the heterogeneity in these data sources and the diversity of analytical techniques necessary to approach them.

- They will be able to apply existing descriptive and predictive analytics techniques and algorithms to gain insights and make predictions with health data including data exploration and visualization, linear regression and extensions, logistic regression, decision and classification trees, random forests, clustering, text analytics, and causal inference.
- Students will develop an advanced understanding of how to interpret results from the application of these existing analytics methods to determine possible decisions, to identify the advantages and disadvantages of these directions in terms of their health impacts, and to communicate these findings to health sector decision-makers.

Prerequisite(s):
None.

Recommended Preparation:
An undergraduate-level or above course in statistics will be helpful preparation for this course. A basic familiarization and comfort with programming concepts, and in particular the R language, will be helpful. This course will illustrate analytics concepts taught in ISE-529 and ISE-535 using examples from health. Previously or concurrently taking these courses is therefore very helpful, but is not required.

Course Overview
The beginning of the course will provide an introduction and overview of health domain sectors and challenges, heath domain data, and analytics methods that can be applied to address challenges using this data. The remainder of the course is structured into a series of case studies, each addressing a different health domain problem, health domain data type, and data analytics method. Case studies will be analyzed during back-to-back lecture and lab components of the class. The lecture component will involve a conceptual introduction to the methods followed by demonstrations of how to apply the methods to analyze the data. This will provide all of the content the students need to implement the analytics method being developed. Class discussion will be emphasized throughout the lecture. The lab session students will provide students with hands-on experience working with the data to replicate the implementation of the methods demonstrated during the lecture, and to attempt extensions to these solutions.

The course will involve a team project of a topic of students' choosing relating to health analytics (details below) which will begin half-way through the semester and end with final reports and a short team-to-team interview-based presentation. The project is a significant aspect of the class, as reflected in the grading scheme.

The course will also involve assignments designed to develop interpretation of health analytics concepts and their implementation in practice to drive health results, through a critique of a paper or report, and a summary of a guest lecture.

Technological Proficiency and Hardware/Software Required
The R programming language and R Studio will be used in all class demonstrations and assignments. This software freely available.

Required Readings and Supplementary Materials
All required course content will be included in the lecture notes. We will be posting research papers of significance to particular health analytics topics to complement the materials covered in class via Blackboard. To go deeper or for additional reference, you may find it useful to refer to the following textbook:

- James, et. al., *An Introduction to Statistical Learning with Applications in R,* 2nd edition, Springer, 2021 (ISLR2) (this book is available for free download on the authors' website at: https://www.statlearning.com/)

Description and Assessment of Assignments

Homework Assignments

There will be 6 homework assignments given throughout the semester, approximately 1 every 2 weeks. These will aim to reinforce your understanding of the methods covered in class and their computational implementation, and interpretation of results. Some assignments will also cover broader data visualization, problem framing, and data-driven communication skills. You will be given 2 weeks for each assignment.

Assignments will be posted on Blackboard and submitted using GradeScope, a grading system that allows for detailed feedback (instructions will be provided). Assignments should be submitted as R Markdown (.Rmd) and their corresponding html files.

Please make sure to submit on time. Assignments turned in after the due date will be penalized by 25%. Assignments not turned in within 48 hours of the due date will not be graded.

All homework assignments are individual. You may find it useful to discuss the problems with one another, however individual solutions must be submitted and copying will not be tolerated.

Labs

Submission of labs conducted in class is mandatory. Labs must be completed in R Markdown, and both .Rmd and html files should be submitted. Labs will be graded for completeness and clarity rather than accuracy.

Paper Critique

Several research papers showcasing recent health domain research using data analytics methods will be shared. Students will choose one paper and write a short critique. The critique should (i) outline the health domain application problem and explain why it is important; (ii) summarize the analytics method(s) used and the results; (iii) interpret the findings in the context of who/what they apply to (patients, population health, insurers, etc.); and (iv) discuss limitations of the approaches and suggest avenues for future work. Your critique should be a maximum of 1 page long.

Final Projects

This will be a project-based class. The project will provide the opportunity to identify a realistic health challenge and to apply analytics methods to address the problem. You will work in self-formed pairs, teams of two. Each team will choose a health application area and identify the decision-making problem, gather the relevant data, use analytics to conduct the analysis, and interpret and communicate findings and recommendations. Findings will be communicated in the form of a report and an interview-based presentation to a fictitious health sector client. While the choice of health application area is flexible, each team should be able to clearly articulate how the application and decision-making problem relate to the health sector. The project will reinforce both technical skills (e.g., programming, model development, model interpretation, data visualization and reporting) and "soft" skills that are very important for working in the health industry (e.g., teamwork, collaboration, communication, project management).

Deliverables throughout the semester will guide students through conducting the project (dates TBD):

- **By Week 6**: Each team needs to be registered with a topic. (5% of the project grade)
- **By Week 10**: Each team submits an interim report (1-2 pages, excluding appendices) (20% of project grade).
    - The report should describe the health application area, the decision-making problem, the data, the analytics approach, initial results, the next steps, and the expected impact.
- **By Week 14.5**: Each team submits the following deliverables (all in pdf):
    - An executive summary including recommendations written for a senior decision-makers (around 300 words) (15% of project grade).
    - A report (around 4 pages, excluding appendices) that includes a description of the health application area, the problem, your analysis, your results, and implications for health or health industry (30% of project grade).
- **On Week 15**: We will hold interviews to showcase your work (30% of project grade). Each team will be interviewed by another team. The presenting team will first be given an opportunity to present its work (2–

3 minutes, 2–3 slides). The interviewing team will play the role of a key decision-maker, in charge of taking and implementing the recommendations.

Participation

Class participation will be evaluated based on engagement in class discussion. Meaningful engagement may include participation in discussion Q&A (asking or answering questions from the instructor or other students), asking or answering questions during lecture, and engaging in guest lectures. For students who miss lecture, a 0.5-1 page summary of the key concepts taught in the class and the key points from the in-class discussion can be contributed after watching the lecture video. At least 5 meaningful class interactions are needed for full participation points.

Grading Breakdown

Assignment	% of Grade
Homework Assignments (6)	30
Lab Assignments (12)	15
Paper Critique	5
Guest Speaker Summary	5
Final Project	40
Class Participation	5

Grading Scale

Course final grades will be determined using the following scale

A	95-100
A-	90-94
B+	87-89
B	83-86
B-	80-82
C+	77-79
C	73-76
C-	70-72
D+	67-69
D	63-66
D-	60-62
F	59 and below

Assignment Submission Policy

Assignments will be posted on Blackboard and submitted using GradeScope, a grading system that allows for detailed feedback (instructions will be provided). Assignments turned in after the due date will be penalized by 25%. Assignments not turned in within 48 hours of the due date will not be graded.

Course Schedule

The broad outline of topics and timelines is summarized below.

The course will focus on descriptive and predictive analytics, focusing on both key methods and modern developments:

Descriptive and predictive analytics (Sessions 1 – 9): These sessions will focus on the key analytics methods used to extract insights and predictions from health data, including: linear and non-linear regression, model selection, logistic regression, support vector machines, classification and regression trees, and random forests.

Modern analytics developments (Sessions 10 – 13): These sessions will focus on health applications of recent analytics developments from novel techniques or existing techniques that are now receiving increased attention in the machine learning and analytics communities.

Introduction

- Introduction to the 3 dimensions we will explore in this course: health domain applications, health data types, and data analytics approaches
- Introduction to R for data analysis

Case-study lectures demonstrating application in interpretation of the following data analytics methods (see description of this class format under Course Overview above):

1. Exploratory data analysis and data visualization
2. Linear regression
3. Non-linear regression
4. Regularization
5. Resampling and cross-validation
6. Logistic regression
7. Regression trees (CART)
8. Classification trees (CART)
9. Random forests
10. Clustering
11. Text analytics
12. Survival analysis and fairness
13. Causal inference
    - Guest lecture

Closure

- Final Project Interviews (1 week)

A detailed week-by-week course breakdown is found on the following pages.

Please note that as this is the first time this course is offered, the schedule is likely to change as we go. A live Schedule Page will be created to share the latest schedule of case studies, homeworks, and project assignments.

Week	Method	Case	Work Assigned	Work Due
1 1/11	**Introduction to R:** Brief introduction including RStudio, rmarkdown	**Introduction to 3 dimensions of this course:** Health data types and sources, Health domain problems, Data analytics methods	**HW0** (graded for completeness)	
Descriptive and Predictive Analytics				
2 1/18	**Exploratory data analysis (EDA) and data visualization**	**Objective**: Explore patterns in relationships between environmental factors and air pollution on children's respiratory health **Purpose**: Inform hypothesis generation, relationships to explore in estimation and prediction models **Data types**: Health survey (*The USC Children's Health Study*), spatial data	**HW1**: EDA, linear regression	**HW0**
3 1/25	**Linear regression**	**Objective**: Predict children's forced expiratory volume (FEV) from routine clinical characteristics, compare to observed value **Purpose**: Clinical tool for diagnosing lung disease **Data types**: [*Repeated data*] Health survey (*The USC Children's Health Study*), spatial data		
4 2/1	**Logistic regression**	**Objective**: Predict future heart disease diagnosis **Purpose**: Identify patients who should be prescribed preventive medication **Data types**: Health survey data (*Framingham Heart Study*)	**HW2**: Logistic regression	**HW1**
5 2/8	**Non-linear regression**	**Objective**: Forecast pm2.5 level by neighborhood **Purpose**: Support hospital planning for acute cardiovascular events **Data types**: Spatial, environmental		
6 2/15	**Linear regression with regularization**	**Objective**: Predict future diabetes from medical records **Purpose**: Identify patients to prioritize for early treatment **Data types**: Electronic medical records (demographic characteristics, clinical measurements, and lab results)	**HW3**: Regularization, resampling	**HW2** **Final Project:** Teams register with topic
7 2/22	**Resampling and cross-validation**	**Objective**: [*Repeated case*] Predict heart disease diagnosis **Purpose**: Identify patients who should be prescribed preventive medication **Data types**: Health survey data (*Framingham Heart Study*)	**Paper Critique**	
8 3/1	**Regression trees (CART)**	**Objective**: Predict future expenditures from healthcare claims data **Purpose**: Identify high-risk patients to mitigate future risks **Data types**: Healthcare claims data (demographics, expenditures, eligibility, diagnoses, procedures)	**HW4**: CART, random forests	**HW3**

9 3/8	**Classification trees (CART)**	**Objective**: Identify neighborhood predictors of neighborhood-level prevalence of stroke **Purpose**: Prioritize communities for stroke-related policy interventions **Data types**: Spatial data (*500 Cities Project*); administrative data (*U.S. Census*)		
3/15	SPRING BREAK			
10 3/22	**Random forests**	**Objective**: [*Repeat case*] Predict future expenditures from healthcare claims data **Purpose**: Identify high-risk patients to mitigate future risks **Data types**: Healthcare claims data (demographics, expenditures, eligibility, diagnoses, procedures)		**Final project**: Interim report
Modern Analytics Developments				
11 3/29	**Clustering**	**Objective**: Cluster groups of similar foods based on nutritional content **Purpose**: Identify foods meeting certain dietary characteristics (high protein, low carb, etc.) to support dietary intervention studies **Data types**: Recipe and nutritional data	**HW5**: Clustering	**HW4** **Paper Critique**
12 4/5	**Text analytics**	**Objective**: Infer sentiment from drug review text and predict drug review rating **Purpose**: Understand patient sentiment on drugs to inform clinical prescription and/or marketing improvements **Data types**: Text data	**HW6**: Text analytics	
13 4/12	**Survival analysis and algorithmic fairness**	**Objective**: Assess fairness of predictive model of recidivism following parole **Purpose**: Illuminate tradeoffs among fairness criteria, and guide model selection accordingly **Data types**: Administrative data (demographics, criminal history)		**HW5**
14 4/19	**Causal inference** ***Guest Lecture***	**Objective**: Understand differences in requirements and use of causal models vs. estimation and prediction models; become familiar with several methodologies for causal reasoning inference developed within computer science **Purpose**: Be able to apply models for future reasoning **Data types**: TBD		**HW6**
15 4/26	–	**Final project presentations**		**Final project**: Final report due (3 days before presentation)

Statement on Academic Conduct and Support Systems

Academic Conduct:

Plagiarism – presenting someone else's ideas as your own, either verbatim or recast in your own words – is a serious academic offense with serious consequences. Please familiarize yourself with the discussion of plagiarism in SCampus in Part B, Section 11, "Behavior Violating University Standards" policy.usc.edu/scampus-part-b. Other forms of academic dishonesty are equally unacceptable. See additional information in Campus and university policies on scientific misconduct, policy.usc.edu/scientific-misconduct.

Support Systems:

Counseling and Mental Health - (213) 740-9355 – 24/7 on call
studenthealth.usc.edu/counseling
Free and confidential mental health treatment for students, including short-term psychotherapy, group counseling, stress fitness workshops, and crisis intervention.
National Suicide Prevention Lifeline - 1 (800) 273-8255 – 24/7 on call
suicidepreventionlifeline.org
Free and confidential emotional support to people in suicidal crisis or emotional distress 24 hours a day, 7 days a week.
Relationship and Sexual Violence Prevention Services (RSVP) - (213) 740-9355(WELL), press "0" after hours – 24/7 on call
studenthealth.usc.edu/sexual-assault
Free and confidential therapy services, workshops, and training for situations related to gender-based harm.

Office of Equity and Diversity (OED) - (213) 740-5086 | Title IX – (213) 821-8298
equity.usc.edu, titleix.usc.edu
Information about how to get help or help someone affected by harassment or discrimination, rights of protected classes, reporting options, and additional resources for students, faculty, staff, visitors, and applicants.

Reporting Incidents of Bias or Harassment - (213) 740-5086 or (213) 821-8298
usc-advocate.symplicity.com/care_report
Avenue to report incidents of bias, hate crimes, and microaggressions to the Office of Equity and Diversity |Title IX for appropriate investigation, supportive measures, and response.

The Office of Disability Services and Programs - (213) 740-0776
dsp.usc.edu
Support and accommodations for students with disabilities. Services include assistance in providing readers/notetakers/interpreters, special accommodations for test taking needs, assistance with architectural barriers, assistive technology, and support for individual needs.

USC Campus Support and Intervention - (213) 821-4710
campussupport.usc.edu
Assists students and families in resolving complex personal, financial, and academic issues adversely affecting their success as a student.

Diversity at USC - (213) 740-2101
diversity.usc.edu
Information on events, programs and training, the Provost's Diversity and Inclusion Council, Diversity Liaisons for each academic school, chronology, participation, and various resources for students.

USC Emergency - UPC: (213) 740-4321, HSC: (323) 442-1000 – 24/7 on call
dps.usc.edu, emergency.usc.edu