HLA 622: Modeling and Predictive Analytics in Healthcare

Instructor:	Chris Malone
Office:	Gildemeister 141 | Zoom: http://minnstate.zoom.us/my/chrismalone
Phone:	Office: 507-457-2989
Email:	cmalone@winona.edu
Websites:	D2L: https://winona.learn.minnstate.edu/
	StatsClass.org: www.statsclass.org

Text: There is no text required for this course. The following references are provided.

- Nelson, J, Felgen, J, Hozak, M. (2021); Using Predictive Analytics to Improve Healthcare Outcomes, Wiley
- Sharda, R, Delen, D. and Turban, E. (2018); Business Intelligence, Analytics, and Data Science: A Managerial Perspective, Pearson.
- Hastie, T., Tibshirani, R., Friedman, J. (2009); The Elements of Statistical Learning: Data Mining, Inference, and Prediction (2nd ed.), Springer.
- James, G., Witten, D., Hastie, T., Tibshirani, R.; An Introduction to Statistical Learning: with Applications in R. (2013)

Course Description: The focus of this course will be on supervised and unsupervised learning methods with applications to healthcare and medical research. The student will be exposed to a variety of machine learning algorithms for uncovering structure in data and predicting outcomes of interest. Unsupervised learning methods covered include clustering and dimension reduction methods. A variety of supervised learning methods for predicting both numeric and non-numeric outcomes will be covered such as tree-based methods and neural networks. Students will gain hands-on experience through course assignments and a final project.

Prerequisites: STAT 601: Statistical Methods for Health Care Research or STAT/HLA 301: Statistical Thinking for Healthcare

Learning Outcomes: A student who has successfully completes this course will be able to:

1. The student will be able to discuss the proper use of various statistical models and machine learning algorithms.
2. The student will apply unsupervised learning methods to an application in healthcare.
3. The student will apply supervised learning methods, i.e. predictive models, to an application in healthcare.
4. The student will assess the quality of statistical models and/or machine learning algorithms.
5. The student will interpret outcomes from statistical models and/or machine learning algorithms.
6. The student will practice taking outcomes from statistical models and/or machine learning algorithms and applying these into decision-making and/or policy development in healthcare.

Assessments:

Interactive Course Notes (Approximately 33% of grade)
A set of interactive course notes will be available for this course. The course notes are divided into modules and parts. These interactive notes may require that you complete various tasks and/or answer questions that are embedded within the notes. After you submit your notes, the answers to questions and/or tasks are provided so that you can check your understanding of the content being covered.

Homework Assignments (Approximately 67% of grade)
Most modules will include an associated homework assignment. All homework assignments will be distributed via Google Docs and completed assignments must be saved to your Google Drive.

Grades:

Your grade is determined by the completion and performance of the required work for this course.

You must complete the interactive course notes for each module / part. The course notes will be scored as follows:

- 0 pts: Did not complete tasks and/or answer questions posed
- 5 pts: Attempted to successfully complete tasks and/or answer questions posed

After completing the set of course notes for a module / part, you will be required to complete the associated homework assignment for that module. Most assignments will include two parts: 1) a technical component, and 2) a second component for the write up.

- The technical component for each assignment will be worth 5 points and will be graded as follows:
    - 0 pts: Did not make a reasonable attempt to complete the technical component of the assignment
    - 3 pts: Made a reasonable attempt to sufficiently complete the technical component of assignment by the due date
    - 5 pts: Sufficiently completed the technical component of the assignment by the due date
- The Write Up component of each assignment will be worth about 10 points. This component will be graded in a more typical fashion with credit/partial credit given. Clear and concise communication is an important element of this component.

Note: You cannot submit homework assignments after solutions have been posted; thus, it is important that you submit your work before the specified deadline for the assignment.

Final grades will be determined using the following scale

- F: Less than 55%
- D: 55% - 64.99%
- C: 65% - 74.99%
- B: 75% - 94.99%
- A: 85% and above

Topic Outline:

1. Introduction to Unsupervised and Supervised Learning
    a. Types of unsupervised learning problems
    b. Types of supervised learning problems

2. Measuring Distance and Similarity
    a. Measuring distance between observations
    b. Measuring distance between variables
    c. Metrics for use with numeric variables
    d. Metrics for use with ordinal/nominal variables
    e. Metrics with mixed variable types

3. Cluster Analysis
    a. Introduction and motivation
    b. Hierarchical cluster analysis (HCA)
        i) distance between clusters
        ii) linkage methods
    c. Interpreting the results from HCA
    d. K-means clustering
    e. Methods for visualizing and describing clusters
    f. Applications of clustering methods

4. Dimension Reduction Methods
    a. Introduction and motivation
    b. Principal Component Analysis (PCA)
    c. Visualizing and interpreting the results from PCA
    d. Multiple Correspondence Analysis (MCA)
    e. Visualizing and interpreting the results from MCA.
    f. Applications of dimension reduction methods.

5. Predicting a Numeric Response
    a. Concepts of bootstrap sampling and cross-validation
    b. Multiple linear regression
    c. Regularized regression

   d. Tree-based models, e.g. regression trees, random forests, variable importance, concept of boosting
   e. Neural networks
   f. Ensemble methods
   g. Applications, communication, and case studies

6. Predicting a Non-numeric Response
   a. Introduction and motivation
   b. Naïve Bayes
   c. Nearest neighbors
   d. Logistic regression
   e. Regularized logistic regression
   f. Discriminant analysis
   g. Tree-based models
   h. Neural networks
   i. Ensemble methods
   j. Applications, communication, and case studies