ITSS 4354 Group Assignment
Submission due: 11/30, In Class Presentation: 12/1

Submission Notes:

1. Only one student needs to make the submission of the code and presentation.
    a. Submit your complete code to eLearning by the due date in R and PDF. Name the file in the following manner: Team#.R and Team#.pdf, where # indicates your team's number.
    b. This code should be appropriately commented on to make it easily used by other programmers.
    c. Submit the PPT file with slides. Name the file similar to your R file. These slides should be prepared in a professional manner. They should accurately convey the information from your analysis and provide an overview of the methods you used. This should be presented at a technical level appropriate to the audience.
2. Every student should individually submit peer review feedback.

Overview:

In this project, you will analyze a set of data using **at least two** of the methods you have learned for classification in the course. You will then create a presentation that explains the practical side of using machine learning techniques for classification.

Background:

A group of physicians have become interested in using machine learning techniques to assist in identifying malignant tumors. They have asked your group to demonstrate machine learning techniques.

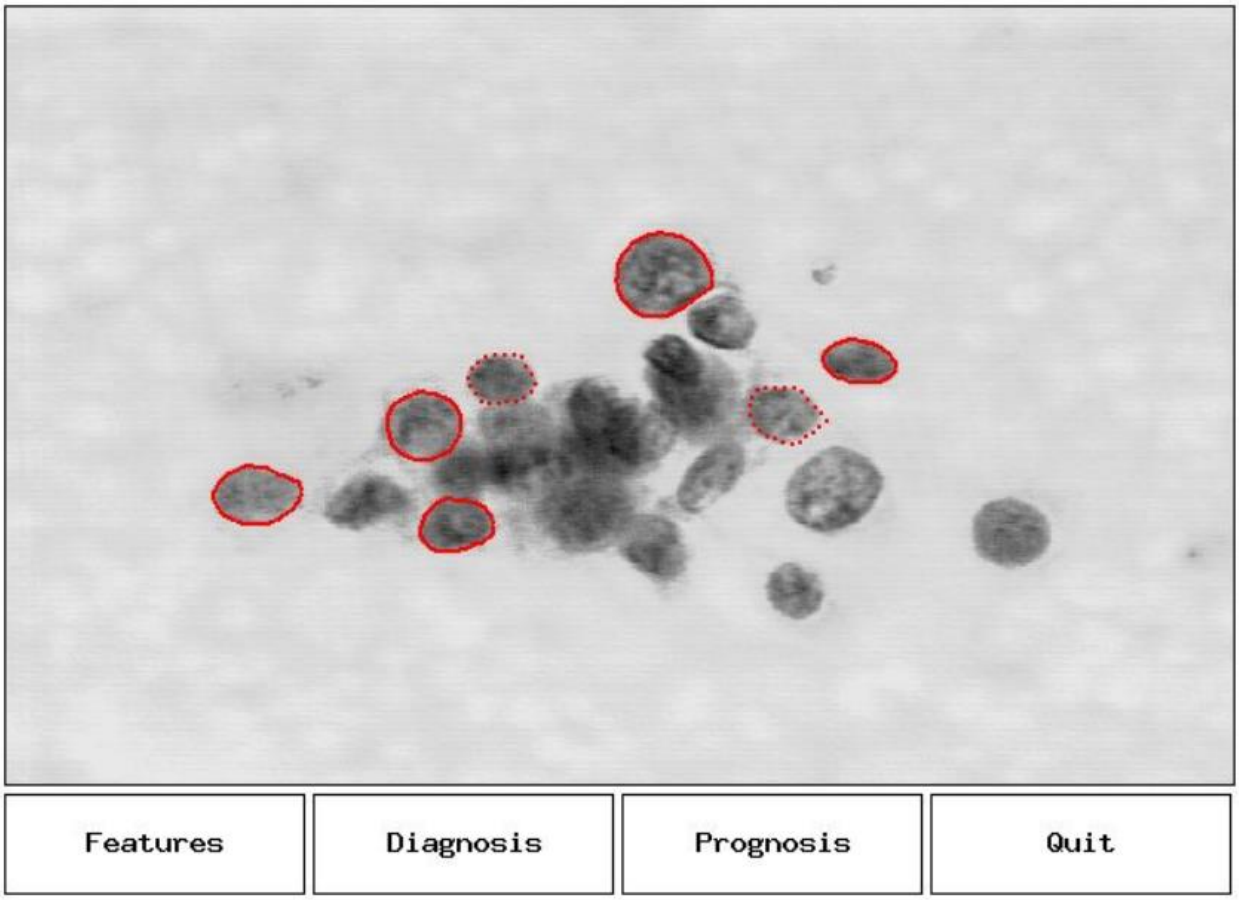
Audience:

Your presentation should be directed to a group of physicians. These individuals are generally well versed in basic science but have a very limited background in statistics and generally no exposure to machine learning techniques.

Data:

The physicians have identified a data set that consists of over 500 measurements from Fine Needle Aspiration (FNA) of breast tissue masses. In an FNA, a small needle is used to extract a sample of cells from a tissue mass. The cells are then photographed under a microscope. The resulting photographs are entered into graphical imaging software. A trained technician uses a

mouse pointer to draw the boundary of the nuclei. The software then calculates each of ten characteristics for the nuclei. An example image is given below.



This process is repeated for most or all of the nuclei in the sample.

The data consists of measurements of the cell nuclei for the following characteristics:

1. radius
2. texture
3. perimeter
4. area
5. smoothness (local variation in radius lengths)
6. compactness (perimeter^2 / area - 1.0)
7. concavity (severity of concave portions of the contour)
8. concave points (number of concave portions of the contour)
9. symmetry
10. fractal dimension ("coastline approximation" - 1)

Measurements of these ten characteristics are summarized for all cells in the sample. The dataset consists of the mean, standard error of the mean, and maximum of the 10 characteristics, for a total of 30 observations for each. Additionally, the data set includes an

identification number and a variable that indicates if the tissue mass is malignant (M) or benign (B).

Task:

You have been asked by the physicians to conduct an analysis of the data using at least two of the classification methods we have seen in this class and provide a presentation that describes those results.

For your analysis, you should:

1. Download the data - FNA_cancer.csv
2. Perform basic exploratory data analysis.
3. Split the data into test and training data.
4. Build at least two classification algorithms.
5. Include metrics that are appropriate for evaluating the performance of your classification models.

For your presentation, you should:

1. Describe the results of your analysis.
2. Describe each algorithm at a level that is appropriate to your audience.
3. Discuss the issue of misclassification. You may assume misclassifying a malignant as benign is worse than misclassifying a benign tumor as malignant.
4. Compare and contrast between the supervised learning algorithms. Discuss the strengths and weaknesses of each algorithm in this context?
5. Discuss the ethical implications of using machine learning for medical diagnosis. How can you ensure fairness and transparency in your model's predictions?

Some notes:

- You do not need to do detailed research into the medical aspects of this project.
- You do not need to use techniques that were not discussed in the course.
- This is not a project that expects you to do "fancy" programming in R.