**Final Exam CSC-260**

For this assignment, we want you to demonstrate the skills/techniques and analytical thinking we have explored this semester. So, focus on what you think is the best way to analyze the data based on the business issue we are trying to solve, as well as how to best communicate your findings with explanatory text (not just program comments), graphs, images and context explanations.

Write your findings as if you are writing to a person who does not know anything about AI/ML/NLP and analytics. At the end of your document, you should include the source code you are using for your analysis (do NOT include this in your overall response. Just present the findings, images, graphs and recommendations).

You response should be a WORD document as an email attachment. The exam is due by Monday May 4rd at noon (11:59am) as email submission to **bjarne.berg@lr.edu** with the email title as follows: "CSC-260 Final Exam - Student Name.docx"

**1. CLASSIFICATION – 100 points**

Using the dataset called “smallsurvey.csv”, build the best classification model you can for classifying the political affiliation found in the field “politicalparty”. Feel free to explore logistic regression, decision trees, or Random Forest as your classification method. Conduct an EDA first and include images, findings, graphics in your finding, as well as evaluation of the model you built. Make sure you use the appropriate sampling techniques for the respective model technique you decide to build and test.

**2. PREDICTION (100 points)**

We are trying to find the best type of red wine to make. This include ‘food engineering” a variety of factors that may be related to the quality of wines. Historically variables assumed to be good indicators for red wine quality include:

* fixed.acidity
* volatile.acidity
* residual.sugar
* chlorides
* total.sulfur.dioxide
* density
* sulphates
* alcohol

In the dataset on Canvas, called “redwines.csv”, we find 1,600 different red wines that have been judged by Wine Enthusiast Magazine to be good (Yes/No). This was determined by looking at wines that scored above 90 on a 100 point scale (already done in the dataset). This is real data.

a) Complete an EDA of this dataset, with graphs.

b) Build a backwards stepwise logistic regression to determine what are the best predictors/ factors for making a good wine.

c) Make a recommendation to the manufacturer on how they should ‘engineer’ their red wines to get a good rating from this magazine.

d) Make sure you evaluate the model based on standard criteria from the class and use appropriate sample/build/test techniques.

**3. ANALYSIS OF GROUPS (100 POINTS)**

You are working for a retailer who also has a web-site sales presence (i.e. Target, Wal-Mart, and BestBuy). The company has captured clickstream web log data of 400 customers during Saturday mornings.

The overall goal is to find out what ads to display to different people on their website, depending on the usage patterns/behaviors. They have 4 potential ad campaigns they can use. This include express shipping (availability), low cost products (for ‘value customers’), offer a 10% discounts to buyers with potential high sales amounts, or choose to display no ads if the person is unlikely to buy anything.

They have given you the “Web\_Purchasing.csv” dataset and are asking you to find meaningful groups of users, based on the captured web server log data. They want an in-depth recommendation on:

a) How many groups to create

b) What users are in each of the proposed groups?

c) A behavioral analysis of each of the groups

d) A set of recommendations on what ads to display to each of the groups

e) An analysis of potential other groupings and a discussion why you picked a certain number of groups.

Use graphs, images and a discussion on what you think the management will find helpful.