Ronaldlee Ejalu
Student ID: 2020637

1) **(20 points):** The Excel spreadsheet heart.csv contains one sheet named heart. These are data from a sample of 1,988 and consists of four databases from Cleveland, Hungary, Switzerland, and Long Beach with 14 variables for each patient. These are:

    1) age
    2) sex
    3) chest pain type (4 values)
    4) resting blood pressure
    5) serum cholestoral in mg/dl
    6) fasting blood sugar > 120 mg/dl
    7) resting electrocardiographic results (values 0,1,2)
    8) maximum heart rate achieved
    9) exercise induced angina
    10) oldpeak = ST depression induced by exercise relative to rest
    11) the slope of the peak exercise ST segment
    12) number of major vessels (0-3) colored by flourosopy
    13) thal: 0 = normal; 1 = fixed defect; 2 = reversable defect

    Develop a Linear Discriminant Analysis model to classify the heart disease from the other variables.

    a) What is the performance of the classifier using cross-validation?

       It is 0.8283,which is equivalent to 82.83%

    b) What is the performance of the classifier using training and testing?

       It is 0.8565, which is equivalent to 85.65%

    c) Would certain misclassification errors be worse than others?

       Yes, certain misclassification errors are worse than other because of the cost involved especially in a class imbalance problem, such as in medical data, there may be a rare class like heart disease. Suppose that you have trained a classifier to classify medical data tuples, where the class label attribute is heart disease and the possible class values are yes and no. The accuracy rate of 86.5% may make the classifier seem quite accurate, but what if only 13.5% of the training tuples are heart disease? This shows that an accuracy rate of 86.5% may not be acceptable, the classifier could be correctly labeling only the nonheart disease tuples for instance and misclassifying all the heart disease tuples.

       If so, how would you suggest measuring this?

       We need other measures that assess how well the classifier can recognize the positive tuples and how well it can recognize the negative tuples: sensitivity and specificity

measures can be used respectively for this purpose. Sensitivity is the proportion of the positive tuples that are correctly identified and from the screenshot below, this is 91.3%. Specificity is the proportion of negative tuples that are correctly identified, and this is defined as 82.3% from the screenshot below.

There is the Positive predicted value, which is the probability that subjects with a positive screening test truly the heart disease; this is shown as 81.2% and Negative Predicted value is the probability that subjects with a negative screening test truly don't have the heart disease and this is shown as 92%.

Also, there is a balanced classifier, which is a metric used in imbalanced classes. It is used to evaluate how good a binary classifier is and this is given as 86.9%.

```
#And then predict this data on the training data and come up with a confusion matrix.
```{r}
p <- predict(heartModelFit, train)
cm <- confusionMatrix(train$heartdisease, p, dnn=c("Actual Group", "Predicted Group"))
cm
```
```

```
Confusion Matrix and Statistics

            Predicted Group
Actual Group   0   1
           0 272  63
           1  26 298

               Accuracy : 0.8649
                 95% CI : (0.8365, 0.8901)
    No Information Rate : 0.5478
    P-Value [Acc > NIR] : < 2.2e-16

                  Kappa : 0.7303

 Mcnemar's Test P-Value : 0.0001356

            Sensitivity : 0.9128
            Specificity : 0.8255
         Pos Pred Value : 0.8119
         Neg Pred Value : 0.9198
             Prevalence : 0.4522
         Detection Rate : 0.4127
   Detection Prevalence : 0.5083
      Balanced Accuracy : 0.8691

       'Positive' Class : 0
```

3) **(10 points-Cluster Analysis): Using Google Scholar**, locate **a journal article**, which uses cluster analysis in your field of interest. **Write a summary** of the journal article and how it utilizes the cluster analysis in **two to three paragraphs**. **Cite the paper in APA format.**

The paper I am summarizing is highlighted in green in the reference section.

Blank et al.(2010) and Klemm et al. (2003) describe how online support groups offer

the same services that face-to-face groups don't offer: greater accessibility in terms of time and geographic proximity, anonymity, and the ability to obtain information without face-to-face interaction. Also, online support group provide people a sense of empowerment, always reminding that you are not alone and this increases self confidence and optimism, which enhances one's social wellbeing, nevertheless there has been little attempt to employ automated methods to analyze online health-related discussion content amongst the three conditions. In order to represent a range of possible physiological and psychological illness experiences, the study employed cluster analysis to identify similar posts for the three different conditions: breast cancer, type1 diabetes and fibromyalgia ; thus, resulting in differing types of information exchange (Chen, 2012, p.1).

Although there has been limited literature review on online support groups on type 1 diabetes and fibromyalgia, breast cancer has been widely studied.

According to Ma et al. (2005), "the internet can serve as a medium for patient education through which necessary knowledge for self-management can be transmitted" (p.577). This explains why so many people use social media platforms for emotional support and encouragement as they seek for treatment.

Three online support groups on a health-related social networking site were analyzed. Through using an application, a crawl was used to download separate document collections for each of the three conditions from the sections of the public discussion content on the website. While researchers have employed document clustering techniques, such as text clustering, text summarization, ranking and organization in previous research in biomedical information retrieval, the author used Vector Space Model to model the discussion content (Chen, 2012, p.2). They applied a procedure for refining the text by removing stop words, punctuation, high and low frequency terms and then used these deliverables to generate frequency matrices where each row represented a single document and each column represented a term.

The author proposed to use repeated bisecting k-means clustering algorithm as this algorithm has been shown to perform better than both direct k -means and hierarchical methods (Steinback, 2000, p.4). With repeated bisecting k-means, they performed cluster analysis, which yielded 20-cluster solutions for each of the three conditions for comparative purposes and through examining the most frequent terms in each cluster, a descriptive cluster label was assigned.

The author was able to come up with a suitable size post sample for each condition. Amongst all the three conditions, the number of posts was higher for fibromyalgia than other condition. The comparison of cluster solutions resulted into a set of common categories for the three conditions and these categories are Generic, Support, Patient Centered, Experimental knowledge, Treatments/Procedures, Medications and Condition

Management. The clusters identified through studies of online support groups share similarities with those found in this study.

Right from emotional aspects, patient education and intervention, the clusters serve as a guide in improving the overall patient experience in dealing with all the three different conditions. This study encourages the integration of interfaces with clustering engines, which can be helpful in complementing regular search engines for fast subtopic information retrieval. In the health care domain, cluster analysis can be used to identify topics of interest to forum participants, and this enables easy access of information to patients especially during the times when they it. In spite of all this, the author needs to conduct other studies to explore whether the nature of discussion on other sites is similar. The author suggests that in future, it might be useful to employ a "soft" clustering method in which threads could be assigned to more than one cluster, such as the algorithm employed by Chen et al.(2010). This needs to be done because there are many factors, which the author never discussed that affect cluster formation. This study used cluster analysis to analyze online support group discussion content, which raised the need to consider people, emotions and temporal aspects of the illness experience.

References

Blank, T. O., Schmidt, S. D., Vangsness, S. A., Monteiro, A. K., & Santagata, P. V. (2010). Differences among breast and prostate cancer online support groups. *Computers in Human Behavior*, *26*(6), 1400-1404.

Chen, A. T. (2012). Exploring online support spaces: using cluster analysis to examine breast cancer, diabetes and fibromyalgia support groups. *Patient education and counseling*, *87*(2), 250-257.

Chen, C. L., Tseng, F. S., & Liang, T. (2010). An integration of WordNet and fuzzy association rule mining for multi-label document clustering. *Data & Knowledge Engineering*, *69*(11), 1208-1226.

Klemm, P., Bunnell, D., Cullen, M., Soneji, R., Gibbons, P., & Holecek, A. (2003). Online cancer support groups: a review of the research literature. *CIN: Computers, Informatics, Nursing*, *21*(3), 136-142.

Ma, C., Warren, J., Phillips, P., & Stanek, J. (2006). Empowering patients with essential information and communication support in the context of diabetes. *International Journal of Medical Informatics*, *75*(8), 577-596.

Steinbach, M., Karypis, G., & Kumar, V. (2000). A comparison of document clustering techniques.

Extra Credit (10 points)

An academic paper from a conference or Journal will be posted to the Homework 4 content section of D2L. Review the paper and evaluate their usage of FA and Latent Dirichlet Allocation (the other LDA). In particular address the following: **(See article on Comparison of Latent Dirichlet Modeling and Factor Analysis for Topic Extraction A Lesson of History)**

- What is the application of this paper?

This paper compares the perceived coherence of topics extracted on three different datasets using Factor Analysis (FA) and Latent Dirichlet Allocation (LDA).

- What is the research question the authors wish to answer in this paper?
  The author wishes to do an evaluation of the topic models, which is still an active area of research and it suffers from a lack of widely accepted evaluation methods.
- What is Natural Language Processing (NLP) and what can we learn from it?
  Natural Language Processing is branch of artificial intelligence that helps computers understand, interpret and manipulate human language. In addition to accurately extract information and insights contained in the documents, NLP can also categorize and organize documents themselves.
- How does this paper utilize FA and LDA in Natural Language Processing?
  The paper utilizes FA and LDA to compare the different topics generated for a given data set by using these two approaches for topic modeling.
- What are the results and conclusions from this paper?
  The results show that topics produced using FA are considered by participants to be more coherent than those obtained using LDA and this was true on all three measures. Further comparatives studies involving both FA and LDA as well as more recent topic modeling techniques should be undertaken to identify conditions under which one technique performs better than the others. Also, FA seems to offer additional benefits over LDA.
- What other areas or fields do you think would benefit from LDA?
  Sentiment analysis, information retrieval and text summarization
- What other thoughts do you have on topic modeling, NLP, and LDA?
  We didn’t find any contemporary study on topic modeling comparing the performance of techniques such as LDA to topic models extracted using Factor Analysis and raises some legitimate questions about the reason why such a technique is broadly ignored today.