**DATA610 – Introduction to Data Mining**

**Group Project 1 – Classification**

This project consists of two activities:

1. Identify the best classification algorithm to classify credit card defaults.
2. Identify the best classification algorithm to classify the quality of the Wine dataset.

The datasets “WineQualityData.txt” and “CreditCardDefaultData.txt” are available on Canvas. To make sure every group has a different version of the dataset, extract a random sample of 5,000 records from the “CreditCardDefault” dataset and a random sample of 2,000 records from the “WineQuality” dataset. These 5,000 and 2,000 record datasets are the dataset your group will start working on.

**What you need to do (do following tasks for each of the sample dataset you have):**

1. As mentioned above, you are working with two datasets containing 5,000 and 2,000 records.
2. Data Descriptions:
   * 1. Provide numerical descriptions of the variables.
     2. Provide and interpret 4 different multivariate (combining 3 variables) ggplot2 data visuals for the dataset. Use ggplot2 to generate appropriate graphical descriptions of the dataset. ggplot2 plots should conform to Tufft’s guidelines and ACCENT principles. Justify the purpose/use of the graphs with respect to the dataset. Be creative and make sure you are providing 4 different types of plots (not different versions of the same plot time).
3. Preprocess data
4. Check and remove rows of missing values. The dataset is big enough so remove all the rows with missing values.
5. Check and remove outliers.
6. Now create your train (80%) and test (20%) datasets.
7. Apply PCA to train dataset and reduce dimensions. Save this dataset. This is what will move to the next stage.
8. Apply all the classification techniques to your train dataset after step (d).
9. Use the above classification models and perform predictions for the test dataset.
10. Find and interpret the accuracy, sensitivity, AUC of ROC, and create ROC curves for the predictions on each model. (Present these as a table.)
11. Compare and discuss the above results and identify the best classification technique for given dataset.

**Software:** You are required to use R-studio for all the data mining tasks.

**Submission**: You are required to submit a word-processed file containing your report. The report should look like a standard report with a cover page (with a title, course name, group members, date, etc), table of contents, chapters, discussion, and conclusion, etc.

Do not include the dataset in the report. You will need to upload the report (as a pdf), your train dataset (as a txt file) and your test dataset (as a txt file) as 3 different attachments.

**Important**:

* You are required to justify any methods or statistics you are using in this project.
* Do not include R commands in the report. You must include all the appropriate/necessary R outputs to justify your work.
* Provide all the appropriate outputs and interpret them.
* All the graphical and numerical descriptions should accompany **detailed interpretations** of them.

**Grading criteria:**

1. For each dataset (45 X 2 points)
2. Data Preprocessing – 10 points
3. Data Visualization – 8
4. Dimension reduction – 6
5. Details of applying each classification algorithm, analysis and interpretations – 15 points
6. Comparison of performances of the algorithms - 6
7. Quality and organization of the report – 10 points
8. Individual Contribution to the project (peer evaluations) – Each member is expected to know the work of the entire project. Your project score will be based on your contribution to both the project and the group. Contribution includes, but is not limited to, the work you have done, how much support you have given to the other members, how well you communicated with your team members, whether you have been an encouraging group member, whether you have asked help from the other members, etc.

**Academic Honesty**

**“**Butler University is an academic community. It exists for the sake of the advancement of knowledge; the pursuit of truth; the intellectual, ethical and social development of students and the general wellbeing of society. All members of our community have an obligation to themselves, to their peers and to the institution to uphold the integrity of Butler University. In the area of academic integrity, this means that one’s work should be one’s own and that the instructor’s evaluation should be based on the student’s own efforts and understanding.  When the standards of academic integrity are breached, mutual trust is undermined, the ideals of personal responsibility and autonomy are violated, teaching and learning are severely compromised and other goals of the academic community cannot be realized.

Students are responsible for being fully aware of what constitutes academic dishonesty; claims of ignorance cannot be used to justify or rationalize dishonest acts. Academic dishonesty can take a number of forms, including but not limited to cheating, plagiarism, fabrication, facilitation and interference:” Taken from <http://www.butler.edu/media/95131/student-handbook-09-10.pdf>