CSE 5243

Instructor: Jason Van Hulse

Homework 4 (Classification)

Due Date: 3/22/2017 5:30 pm. Hand in the printed report to me in class.

In this lab, you will experiment with multiple classification algorithms on the Wine_Quality dataset. **This is an individual assignment - you are not permitted to work in a team.**

Dataset: Wine_quality

Additional information on the wine_quality dataset can be found at https://archive.ics.uci.edu/ml/datasets/Wine+Quality. This homework uses the *winequality-red.csv* dataset. Further, the data is converted to a binary class by transforming the *quality* attribute to a *class* attribute using the following rule:

quality <= 5 —> class = “Low”, quality > 5 —> class = “High”

I have posted the **wine** dataset in Carmen as a *.csv file. Note that this dataset has both an ID variable and the quality attribute, which is what was used to create the class.

Your ultimate objective is to build a classification model to predict the class given the independent variables.

You can use any software (or combination of software) for this assignment that you would like. Please consider all aspects of the knowledge discovery cycle in this assignment, including preliminary data exploration, data processing/ transformation, model building and model evaluation. The report that you turn in should describe what you did throughout the project - you do not need to turn in any code.

Your report should have (at minimum) the following sections:

1) **Preliminary data analysis** (15% of grade) - discuss any observed trends with the data, include interesting graphs (histograms, scatterplots), summary statistics, correlations among features, etc. I would suggest including some of the interesting output of this preliminary data analysis to help augment the presentation.
2) **Data Transformations** (5% of grade) - What if any transformation or processing of the data is needed (prior to actually building the models) e.g.,

treatment of outliers and missing values, discretizing, feature subset selection? Even if you decide not to handle these issues, please discuss why.

3) **Model development** (50% of grade)- Test each of the following modeling approaches in this section. For each approach, discuss the parameters you experimented with and why you ended up choosing what you finally chose.
   - **A.** Decision Tree
   - **B.** A rules-based classifier
   - **C.** Naive Bayes
   - **D.** Artificial Neural Network
   - **E.** Support Vector Machine
   - **F.** Ensemble learner (such as Adaboost, RandomForest, etc)
4) **Model evaluation** (30% of grade) - Describe your approach to evaluating and comparing the performance of the different models that you have built in part 3 (e.g., 10-fold cross validation). For each model include performance statistics (such as accuracy, F-measure, ROC Curve, etc). The conclusion should be to select your preferred modeling approach - please justify this choice, and state any pros and cons of this choice.

Throughout the report, state what software packages you used for that part. For example, you may use R or Python for preliminary analysis and transformations and Weka for Model Development. Describe the packages used, and list out settings attempted.

What you need to turn in:

You should hand in a written report, maximum of 8 pages in length. *You do not need to turn in any code or data for this assignment.*

The report should be well-written. Please proof-read and remove spelling and grammar errors and typos. *Writing and presentation will be part of your grade for this assignment.*