**COMP 4112 Introduction to Data Science**

**Assignment 2, Regression**

**5 marks, 5 percent total**

**Submission:**

**Submit a single .py file named “<username>\_A2.py” on myCourselink.**

**Grading:**

**1/5 – Does not function**

**2/5 – Partially functioning**

**3/5 – Basic features, no evidence of experimentation**

**4 and 5 / 5 – Working and evidence of some experimentation**

In this assignment, you will work with a dataset about wine quality. You will select/develop features and train three regression models, one for red wines, and one for white wines, and a third to try and predict wine quality based on attributes available to consumers only (you will need to select appropriate features to do this). Your goal with these 3 regression models is to predict the quality of wine on a scale of 1-10.

Number of Instances: red wine - 1599; white wine - 4898.

Number of Attributes: 11 + output attribute.

A Hint from the dataset: several of the attributes may be correlated. You can explore these in R with the cor function.

**For this assignment (Python only):**

1. **Read in the CSV dataset. You can do this how you like; Python lists are totally acceptable but you might have to convert to other formats for scikit-learn sometimes. If you want, you could use a pandas dataframe or a numpy array.**
2. **Fit the three regression models using the LinearRegression from sklearn.**
3. **Report on the performance of these models in Python with R^2 or MSE. Other measures can be used as well.**

**Other Hints:**

**The Multiple Regression example code can be re-used and modified for this assignment. The code should be sufficiently refactored.**

**The citation of the associated paper for this dataset developed by Cortez et al. is** [**https://www.sciencedirect.com/science/article/abs/pii/S0167923609001377**](https://www.sciencedirect.com/science/article/abs/pii/S0167923609001377)