COMP 4448: Data Science Tools II Assignment 7

**Directions:** Do this assignment in Jupyter Notebook and provide screenshots of code and output in this word document wherever required. You will upload this word document containing screenshots of code and answers as well as your Jupyter Notebook to Canvas. All assignments will be submitted and graded through canvas and grades will be transferred to the 2U platform.

**Goal:** The goal of this assignment is to give you the opportunity to implement clustering and logistic regression using your own data.

**Packages:** Core packages you may need for this assignment include numpy, pandas, sklearn, statsmodels, matplotlib.pyplot and/or seaborn.

Question 1.

Find some dataset with at least four variables and at least 100 cases and apply clustering to the dataset. Make sure to address the following:

* Data cleaning where necessary
* Scale the data using an appropriate scaler
* Graphical user interface, text, application, email

  Description automatically generated
* Conduct a principal component analysis
* Graphical user interface, text, application

  Description automatically generated
* Construct a dendrogram using agglomerative clustering to see how many clusters will be optimal to specify in the kmeans clustering.
* Text

  Description automatically generated
* Implement a kmeans clustering to find the clusters in the data
* Graphical user interface, text, application

  Description automatically generated
* Predict the clusters
* 
* Visualize the clusters
* Chart, scatter chart

  Description automatically generated
* Also use a loop and a plot to tune the number of clusters. Does the number of clusters obtain parallel the number of clusters obtained using the dendrogram?
* Chart, scatter chart

  Description automatically generated
* Add the cluster values to your original dataset to be the labels
* 
* Using tools in sklearn, run a logistic regression on the original dataset with new labels to classify cases into the clusters or labels.
* Tune the cutoff value, c, of the logistic regression constructor in sklearn and plot the error rates and the corresponding settings of c. note that the cutoff c ranges from 0 to 1.
* Chart, scatter chart

  Description automatically generated
* Use the optimal cut off to compute the accuracy of your logistic regression.
* Text, letter

  Description automatically generated
* Do you think the clustering algorithm found good clusters that can be used for classification?
* They seem to be divided evenly but I’m not sure how to interpret it.
* Try if you can interpret or describe what the clusters represent based on the pattern of values in the data set and the cluster. For example, if the data was only age and income, maybe clusters one is young and richer, cluster two may be rich seniors and cluster three are poor youths. Your visualization can also help with this interpretation.

Question 2

Find another dataset that is suitable for logistic regression. Run a logistic regression on the data using the statsmodel package. Print the results and interpret the parameter coefficients for each input variable: <https://www.statsmodels.org/stable/index.html>. Evaluate the model as well.

Alcohol content, sulphates and free sulfur dioxide are the most statistically significant factors in determining wine quality.

Graphical user interface, application

Description automatically generated

Graphical user interface, table

Description automatically generated