

Python Exercise.

Theme 3

Wine Quality Analysis

*Máster en Big Data*

*Curso 2019-2020*

***Data Science & Advanced Analytics***

*Miguel Rooney*

*Objective*

Apply a logistic regression over a wine quality dataset to draw conclusions.

*Tasks to perform*

**1.- Import the needed Python packages**

In this practice you will be using the following libraries/Packages:

* [pandas](https://pandas.pydata.org/) – used to create the DataFrame to capture the dataset in Python and make ETL if needed.
* [sklearn](https://scikit-learn.org/stable/) – used to build the logistic regression model.
* [seaborn](https://seaborn.pydata.org/) – used to display the results via a [Confusion Matrix](https://datatofish.com/confusion-matrix-python/)

**2.- Input Data and build the Dataframe**

You will be using the “winequality-red.csv” provided dataset

* Load the dataset into a dataframe and show the first 5 rows
* Print to show:
  + describe
  + shape
  + nulls
  + NaN
  + dtype

**3.- Make the dataframe ETL tasks**

Perform the appropriate Dataframe’s transformation to ensure that it is ready for logistic regression (think of what the conditions are for running a logistic regression on a dataframe).

* Eliminate outliers (standardize)
* Delete columns, only if needed and justify the decision if you do so
* Others…?

**4.- Plot variable’s correlation**

Using a seaborn plot, create a heatmap plot showing the correlation between variables. Use “Correlation between variables” as a title).

* Hint: heatmap(corr,ax=ax, annot= True)

Guide: <https://seaborn.pydata.org/generated/seaborn.heatmap.html?highlight=heatmap#seaborn.heatmap>

**4.- Perform the logistic regression**

* set the independent variables (represented as X) and the dependent variable (represented as y)
* Apply train\_test\_split setting the test size to 0.25, and therefore the model testing will be based on 25% of the dataset, while the model training will be based on 75% of the dataset:
* Apply the logistic regression
* Get the confusion Matrix
* Print the accuracy

**6.- Plot with matplotlib**

Plot two graphs that you consider are the most suitable to show the obtained results. Justify it.

**7.- Check the prediction for a new dataset.**

Apply the logistic regression obtained in the previous steps to the following dataset:

* winequality-white.csv

**8.- Explain the results**

Draw two conclusions about the results you got as a result of the developed model so far, comparing white and red wines.

Maximum number of people: 3

Deliverable:

* A file containing:
  + Python code in Jupyter Notebook (\*.ipynb)
  + Name of the file: <Apellidos+InicialNombres>\_Exercise\_3.ipynb
  + Example: LlosM\_RooneyM\_Exercise\_3.ipynb
* A pdf document with your names and photos
  + Example: LlosM\_RooneyM\_Ejercicio\_3.pdf

Deadline: 2020, February 10th at 24h

Mode of delivery: Upload on campus in the corresponding evaluation activity.