Assignment 4

Using the wine_properties.csv data set, implement:

- a linear regression (https://www.statsmodels.org/stable/regression.html) to predict the percentage of alcohol of the wine, using the other available variables
- a logistical regression (as we did during the last lecture) to predict if a wine is “strong” or “weak”. To do that, create a dummy variable using a threshold to discretize the alcohol variable, and use this dummy variabile as target. The threshold is arbitrary (hint: maybe you can use a percentile of the alcohol distribution, like the 75th percentile).
- The same as the previous points, but using as regressors the first principal components after PCA (do not include all the components, but a subset, for example the one that explains around the 75% of the variance).

Provide some indication about the overall fitting for the first problem (as R squared) and about the accuracy of the prediction for the second problem. Compare the results you get with the two approaches (non-PCA vs PCA).