Instructions: Please submit your answers as 2 files uploaded to Courseworks: a Jupyter Notebook (.ipynb) file & a pdf export. Please double check that all pages exported properly, sometimes they get cut off! In answering each of the following questions please include (a) the question as a markdown header in your Jupyter notebook, (b) the raw code that you used to generate any results, tables, or figures, and (c) the top ten or fewer rows of the dataframe (do not include more than ten rows for any table in your report). Include any plots or figures generated from your code as well.

Part A: Regression on California test scores

1. Find the url for the California Test Score Data Set from the following website: https://vincentarelbundock.github.io/Rdatasets/datasets.html

Links to an external site. .Read through the "DOC" file to understand the variables in the dataset, then use the following url to import the data. The target data (i.e. the dependent variable) is named "testscr". You can use all variables in the data except for "readscr" and "mathscr" in the following analysis (those two variables were used to generate the dependent variable). https://vincentarelbundock.github.io/Rdatasets/csv/Ecdat/Caschool.csv

Links to an external site.

2. Visualize the univariate distribution of the target feature and each of the three continuous explanatory variables that you think are likely to have a relationship with the target feature.

3. Visualize the dependency of the target on each feature you just plotted.

4. Split the data into training and test sets. Build models that evaluate the relationship between all available quantitative X variables in the California test dataset and the target variable. Evaluate KNN (for regression), Linear Regression (OLS), Ridge, and Lasso using cross-validation with the default parameters. How different are the results?

5. Try running your models from the previous question with and without StandardScaler. Does using StandardScaler help?

6. Tune the parameters of the models where possible using GridSearchCV. Do the results improve?

7. Compare the coefficients of your two best linear models (not KNN). Do they agree on which features are important?

8. Discuss which final model you would choose to predict new data.

Part B: Classification on red and white wine characteristics

First, import the red and the white wine csv files into separate pandas dataframes from the following website. Note that you'll need to adjust the argument for read_csv() from sep=',' to sep=';'
https://archive.ics.uci.edu/ml/machine-learning-databases/wine-quality/winequality-white.csv

Links to an external site.

https://archive.ics.uci.edu/ml/machine-learning-databases/wine-quality/winequality-red.csv

Links to an external site.

2.Add a new column to each data frame called "winetype". For the white wine dataset label the values in this column with a 0, indicating white wine. For the red wine dataset, label values with a 1, indicating red wine. Combine both datasets into a single dataframe. The target data (i.e. the dependent variable) is "winetype".

3.Visualize the univariate distribution of the target feature and each of the three explanatory variables that you think are likely to have a relationship with the target feature.

4.Split data into training and test sets. Build models that evaluate the relationship between all available quantitative X variables in the dataset and the target variable. Evaluate Logistic Regression, Penalized Logistic Regression, and KNN (for classification) using cross-validation. How different are the results?

5.Try running your models from the previous question with and without StandardScaler. Does using StandardScaler help?

6.Tune the parameters of the models where possible using GridSearchCV. Do the results improve?

7.Compare the coefficients for Logistic Regression and Penalized Logistic Regression. Do they agree on which features are important?

8.Discuss which final model you would choose to predict new data.