Introduction (background of the task)

Portugal began to produce and export wine during the Roman Empire, and over the years, Portugal has become a major wine growing country. The data set used in this project relates to red and white wine samples from northern Portugal. This wine is called Vinho Verde and is known for its unique characteristics. This dataset can be processed as a classification and regression task. The ordered and lopsided nature of these categories adds to the complexity, with average wines being significantly abundant compared to fine or poor-quality wines. This imbalance raises the potential application of outlier detection algorithms to identify exceptional or sub-par wines in a dataset. It can be seen from The data set that The number of instances is 1599 for red wine and 4898 for white wine. Due to privacy and logistical constraints, although there is a lack of information on grape types, wine brands and sale prices, this highlights the focus on physicochemical and sensory variables. But there are a total of 11 variables, which are fixed acidity, volatile acidity, lime, acid, residual sugar, chloride, free sulfur dioxide, total sulfur dioxide, density, pH, sulfate, alcohol. The aim is to reveal and predict the sensory quality of Vinho Verde wines based on physicochemical properties by using linear regression, logistics regression, polynomial regression methods.

Problem formulation *(what is input, what is output, where did you get the dataset, number of samples, etc.)*

The input comprises physicochemical attributes obtained through objective tests on red and white variants of Vinho Verde wine. These attributes include: fixed acidity volatile acidity citric acid residual sugar chlorides free sulfur dioxide total sulfur dioxide density pH sulphates alcohol. These attributes represent measurable characteristics of the wines.

the output is the sensory quality evaluation of the wines, expressed as a quality score between 0 (very bad) and 10 (very excellent).
This output is derived from the median of at least three evaluations made by wine experts.

The dataset used for this project is originally from
https://archive.ics.uci.edu/dataset/186/wine+quality

Sources Created by: Paulo Cortez (Univ. Minho), Antonio Cerdeira, Fernando Almeida, Telmo Matos and Jose Reis (CVRVV) @ 2009 Available at: [@Elsevier] http://dx.doi.org/10.1016/j.dss.2009.05.016 [Pre-press (pdf)] http://www3.dsi.uminho.pt/pcortez/winequality09.pdf
[bib] http://www3.dsi.uminho.pt/pcortez/dss09.bib

Approaches and baselines (what are the hyper parameters of each approach/baseline, how do you tune them)?

Linear Regression

Approach: Linear regression is a basic and interpretable model for predicting the sensory quality of a red wine based on its physicochemical properties. It assumes a linear relationship between input features and outputs.

Hyper parameter:
Linear regression has the fewest hyper parameters, making it a simple model. The main hyper parameters are related to the fitting process:

There are no specific hyperparameters to adjust:
Linear regression has no complex hyperparameters such as learning rate or hidden layers.The main concern is the coefficient (weight) assigned to each input feature.

Linear regression is started using scikit-learn's LinearRegression class. Model evaluation involves mean square error, mean absolute error and accuracy. The model is trained on the training set and then selected based on the performance on the verification set. Finally, the selected model is evaluated on the test set. This linear regression implementation can be used as a basic baseline for predicting wine quality.

Logistic Regression

Approach:Logistic regression is used as a classification model to predict binary outcomes (positive or negative) for wine quality based on physicochemical properties.

Hyper parameter:
Logistic regression has hyper parameters that can be adjusted for better performance:

Maximum number of iterations:
Represents the maximum number of iterations in which the solver converges.
Increase to 10,000 to avoid convergence warnings.

Tuning process:
The main hyper parameter to consider is max iteration. Increasing it ensures that the optimization algorithm has enough iterations to converge. This value may need to be adjusted for data set size and complexity.

Polynomial Regression:

Approach:Polynomial regression extends linear regression by introducing polynomial features, allowing the model to capture non-linear relationships between input features and the target variable. In this case, a polynomial of degree 2 is applied.

Hyper parameters:
The primary hyper parameter is the degree of the polynomial (degree). In this script, degree is set to 2, indicating a quadratic polynomial.

Tuning Process:
The tuning process involves selecting an appropriate degree for the polynomial. This can be done through experimentation and validation set performance. Higher degrees can capture more complex relationships but may lead to overfitting.

Evaluation metric (what is the measure of success, is it the real goal of the task, or an approximation? If it's an approximation, why is it a reasonable approximation?)

Measure of Success:
Mean Squared Error (MSE): This metric measures the average squared difference between the predicted and actual values. A lower MSE indicates better predictive performance.

Mean Absolute Error (MAE): It represents the average absolute difference between the predicted and actual values. Similar to MSE, a lower MAE signifies better model accuracy.

Accuracy: This metric calculates the proportion of correctly predicted instances. In the context of regression, where predicting an exact score may be challenging, rounding the predictions and comparing to the true values provides a measure of accuracy.

Close Accuracy (within ±1 score): This is a more lenient measure, considering predictions within a one-point difference from the true values. It accounts for situations where the model may not precisely predict the exact score but is close enough.

Results. (What is the result of the approaches? How is it compared with baselines? How do you interpret the results?)

Linear regression:
In the context of predicting wine quality, linear regression models achieved moderate success.This accuracy, while reasonable, suggests there is room for improvement,

especially given the nature of wine quality scores, which are often integers.Notably, the close accuracy on the test set (93.13%) shows that the model predicts wine quality scores that are typically within 1 point of the true value.

Logistic regression：

Logistic regression does not appear to be suitable for predicting wine quality scores, as evidenced by the extremely high MSE and MAE. An accuracy of 0 indicates that the logistic regression model does not effectively capture the relationships in the data set. The lack of accuracy may stem from regression methods applied to similar classification tasks, where wine quality is scored discretely.

Polynomial regression：

It can be seen from the results that polynomial regression (degree 2) is better than linear regression and is well in line with the expectations for modeling the nonlinear relationship of wine quality. The model achieved higher accuracy and near accuracy on all datasets, indicating improved prediction performance. At the same time, the close accuracy of the test data (96.25%) shows that the wine quality score predicted by the polynomial regression model differs from the true value by no more than 1 point.

Conclusion：

In the context of predicting wine quality scores, polynomial regression with degree 2 emerges as the most effective approach.

Linear regression provides reasonable accuracy but falls short in capturing non-linear patterns inherent in wine quality assessment.

Logistic regression, designed for classification tasks, is inappropriate for predicting wine quality scores, resulting in poor performance.

The success of polynomial regression underscores the importance of considering non-linear relationships when modeling wine quality. The close accuracy metric is particularly relevant, aligning with the practical expectation that predicting wine quality within a narrow range is meaningful in real-world scenarios.