Machine Learning Assignment 1 - Wine Quality Prediction

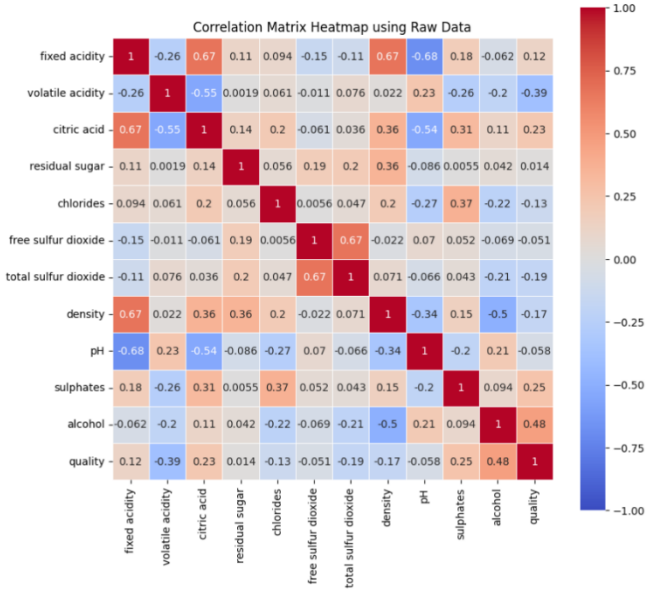
Introduction

In this report, a data set on red wine quality will be explored, with the aim being to create a model to predict the quality of said red wine to the highest degree of accuracy. This is useful for many reasons, as these methods can be used for wine producers to create higher quality wines. Methods used will include a multitude of regression models, and some analysis of the data set provided. The quality of each model will also be analysed.

Part 1 – Data Set Analysis

The first necessity is to analyse our data set, checking for anomalous points and erroneous data. To do this, some simple programming was used, to remove any row from the data set which contained data that violated the 1.5IQR rule. (That is, if the data point is more than 1.5 times the IQR above or below the 3rd and 1st quartile respectively, that result will be ignored).

Doing this resulted in 405 rows being removed from the original 1599 rows (approx. 25.3%). Following this, two heat maps were created, showing the correlation between each variable. The targeted values here were the correlations between each column and the column "quality".



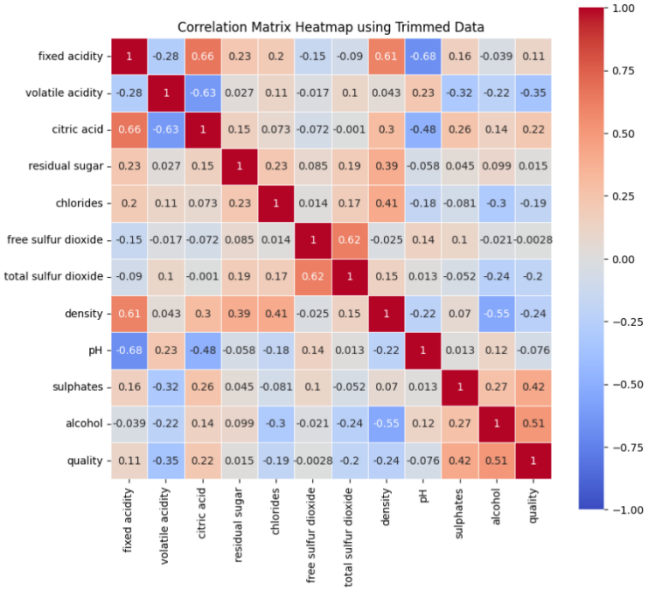


Fig 1 (Left) showing correlations between the raw data

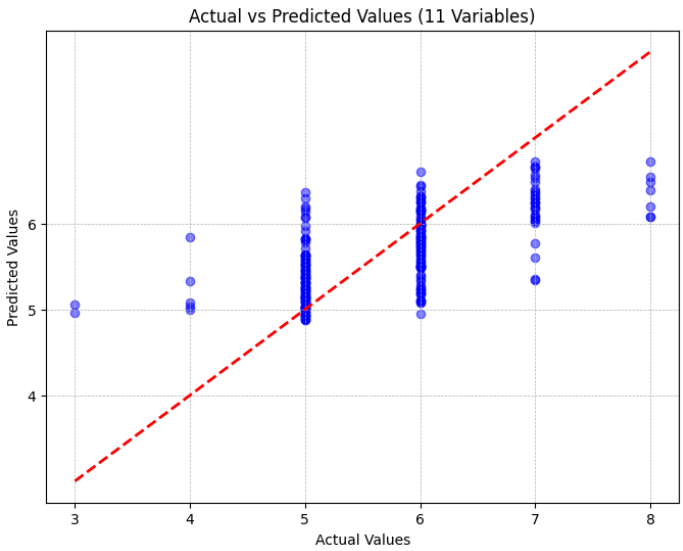
Fig 2 (Right) showing correlations between the trimmed data

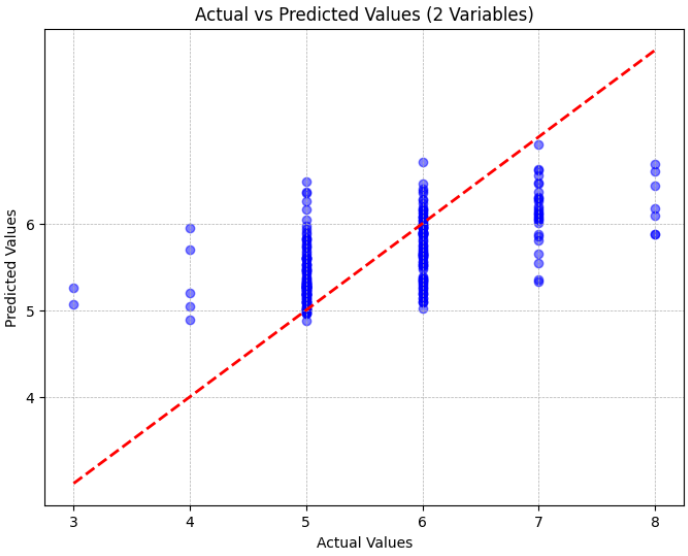
Using this, it becomes clear the two variables with the highest correlation to quality are "alcohol" (0.51) and "sulphates" (0.42) – (Using Trimmed Data / Fig 2).

Part 2 – Regression Models

Following the analysis made, two linear regression models were run, both using the trimmed data set. One used all 11 variables that affect quality, one used only 2 (sulphates and alcohol). The two models are displayed below in

graphs (figures 3 and 4). The linear regression model processes the training set 1000 times (epochs), and the model uses gradient descent to approach the lowest cost function possible. A linear model uses the assumption





that each variable is independent, and a linearly related to the dependent variable (quality in this case). Lowering the cost function is lowering the sum of the differences between observed and predicted values.

Figure 3 (left) Figure 4 (right) – graphs showing actual vs predicted values for each model.

The following error results were returned for each model as well:

11 Variable model (Figure 3): Mean Squared Error = 0.426, Root Mean Squared Error = 0.652,

Mean Absolute Error = 0.490, R-squared = 0.417

2 Variable model (Figure 4): Mean Squared Error = 0.478, Root Mean Squared Error = 0.691,

Mean Absolute Error = 0.517, R-squared = 0.345

All values to 3 decimal places.

Following this, 5 other regression models were implemented. Ordinal, Random Forrest, XGBoost, Polynomial, and SVR. An outline of the methodology of these models follows.

Ordinal Regression:

An ordinal regression model uses proportional odds, estimating the chance of each row of variables being in each class (in this case, the probability of each set of 11 variables resulting in each different quality output).

Random Forrest Regression:

Random forest does not presume any information about the relationship between the independent variables and dependent variables. It uses “decision trees”, constructing decision trees using binary splitting, and averaging all predictions from each individual tree.

XGBoost Regression:

Similar to Random Forrest in its use of decision trees, but built sequentially, with each new tree being built to correct the errors of the previous trees. Also uses gradient descent to optimize similar to linear regression.

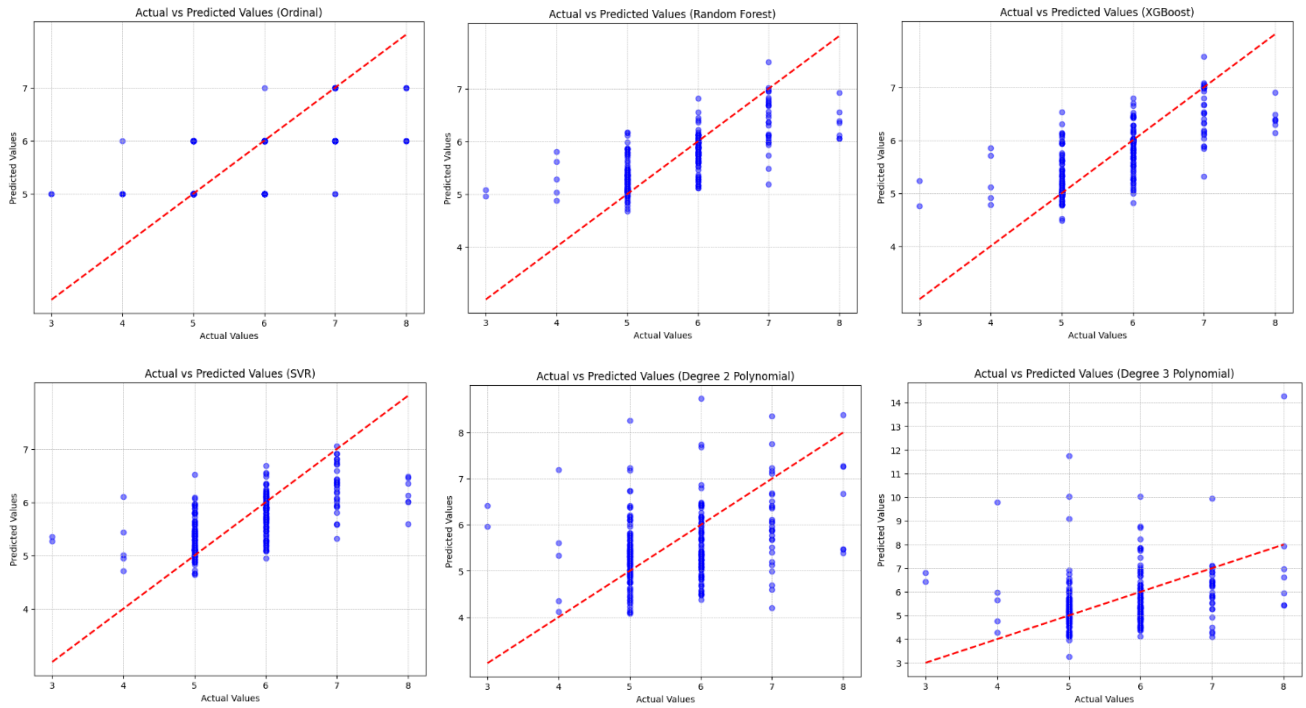
Support Vector Regression (SVR):

Uses higher dimensions to find a hyperplane that best fits the data and attempts to minimize both the complexity of the function and the error term.

Polynomial Regression:

Similar to linear regression, but assuming a non-linear relationship, and modelling the relationship as a polynomial function. The approach is very similar to linear regression, but as a polynomial function, not a linear one.

The results from each of these models is as follows. For the comparison of each, the main error measurement used will be Mean Squared Error. This decision was made based off a paper by Gaudette[1]. (Gaudette, 2009)



From left to right, top to bottom. Figures 5,6,7,8,9,10 – Showing actual values Vs Predicted values for Ordinal, Random Forest, XGBoost, SVR, Degree 2 polynomial, and Degree 3 polynomial regression models respectively.

Part 3 – Discussion of results

Comparing the results from the seven models in order of least to most accurate (using MSE as specified earlier) –

Polynomial Degree 3 = 1.848

Polynomial Degree 2 = 1.005

Ordinal = 0.502

Linear (2 Variables) = 0.478

SVR = 0.431

---

[1] Gaudette, L., Japkowicz, N. (2009). Evaluation Methods for Ordinal Classification. In: Gao, Y., Japkowicz, N. (eds) Advances in Artificial Intelligence. Canadian AI 2009. Lecture Notes in Computer Science(), vol 5549. Springer, Berlin, Heidelberg. https://doi.org/10.1007/978-3-642-01818-3_25

Linear (11 Variables) = 0.426

XG Boost = 0.369

Random Forrest = 0.360

Reading from this, the random forest model is the best predictor of wine quality. If we reverse the mean square system, it's average prediction is roughly 0.6 points away from the correct quality. It is not fair to say it is 100% the best, as many factors could affect trials (random state of model, more trials, etc), and this does not mean it would be the best model with a new set of data, especially given how close in accuracy it is to some of the other models.

A main point to take is that the model is certainly not polynomial. Both polynomial models are considerably more inaccurate than any other model.

Part 4 – Conclusion and potential further analysis

Concluding from this that the random forest model is most accurate in this situation, but that any of the nonpolynomial models are reasonably accurate. There is also a possibility that a better model could be found, methods for that could include trailing more, different combinations of variables, trailing different models not used in this specific project, or more rigorous analysis of the data set.

Statement of AI Usage

In the process of completing this assignment, I have used AI tools in the following ways.

1. Programming and Debugging: I have used AI to improve and optimize some sections of code, along with help with errors I encountered along the ways of programming.
2. Clarification of certain concepts: I have used AI to clarify some specifics about certain methodologies of specific regression models that I did not 100% understand.

All work is my own, the AI was simply used as a tool to enhance my learning and implementation of my own knowledge