Wine Quality Prediction

Stijn de Preter (852726504) [1] Arjan Broer (850166428) [1]

1. Brief Introduction

The assignment is based on the article (Dahal et al., 2021) and focuses on the reproduction of the article results. The reproduction is limited to the Support Vector Regression (SVR) and Artificial Neural Network (ANN) models. The Ridge Regression (RR) and Gradient Boosting Regression models are not discussed in this report. The goal of the assignment is to predict the quality of wine based on its physical and chemical properties and understand the methods used to generate the predictions. The dataset will be analyzed to find the features that most influence the quality of wine and suggestions are provided to improve the hyperparameters of the model for best results.

The research questions discussed in this report are:

1. Reproduce the results of the article (Dahal et al., 2021) using the same dataset and models.
2. Analyze the dataset to find the features that most influence the quality of wine.
3. Provide suggestions to improve the hyperparameters of the model for best results.
4. Explore the improvements that an ensemble model can bring.

2. Methods

The methods used for this assignment are based on the article (Dahal et al., 2021) and include Support Vector Regression (SVR) and Artificial Neural Network (ANN) models. The impact of the features is analyzed using SHAP (SHapley Additive exPlanations) values and the Pearson correlation coefficient. Improvements on the both the SVR and ANN models are investigated and reported.

2.1. Reproduce the results

In order to compare the results of the article with our results, a visualization of the performance metrics was made. The values of the article (Dahal et al., 2021) are compared with our results repeatedly. Our results are reproduced repeatedly with different test and training data each time. This way we are sure that it is not a lucky shot but we get a realistic picture of how our models perform. Also, standardization has been applied to the data features of the dataset so not to the quality column.

2.2. Impact of the features

The article analyses the impact of individual features on the quality of the wine. This analysis is done by looking at the correlation between the features and the quality of the wine. The correlation is calculated in the article using the Pearson correlation coefficient. The Pearson correlation coefficient is a measure of the linear correlation between two variables.

By using SHAP (SHapley Additive exPlanations) values, the impact of the features on the model can be analyzed in more detail. The explanation of the model should reflect a similar result as compared to the pearson correlation coefficient. Shapley values are created by using the Shapley value method, which is a method from cooperative game theory. Effectively it will calculate the contribution of each feature to the model output by looking at the impact of each feature on the model output. (Lundberg & Lee, 2017) describes in detail how the impact of individual features is determined by retraining the model with this a feature excluded. Then the impact of the feature on the model output is calculated by looking at the difference between the model output with and without the feature. Diagrams used here to explain the impact of features are the summary plot and dependency plot. The summary plot shows the impact of all features on the model output ordered from strongest impact to weaker impact. The dependency plot shows the impact of a single feature on the model output.

2.3. Improve the model

For optimizing the hyperparameters, BayesSearch is used for SVR. For ANN, gridsearch was used to examine the impact of the hyperparameters. The optimized models will be evaluated and compared with the values from the article and with the reproduced models.

2.4. Ensemble model

The article (Dahal et al., 2021) also mentions the use of ensemble models. The ensemble model is a combination of multiple models to improve the performance of the model.

Specifically, the article mentions the use of Gradient Boosting Regression (GBR) model as ensemble model. For this report an ensemble model is created by combining the SVR and ANN models, by averaging the predictions. The ensemble model is compared with the individual models to see if the ensemble model performs better than the individual models.

3. Experimental Results

This chapter on experimental results is divided into four sections: a brief overview, an analysis of SVR and ANN, and a part on the ensemble model.

3.1. Datasets

When exploring the data it was found that the information given in the article about the data did not match the information about the data. The article indicates that the red wine dataset was used. Based on the number of rows and the statistics (mean, std, min, max, median) we see that the white wine dataset was used. In the rest of this report we will work with the white wine dataset.

The correlation between all the features has also been visualized in figure Figure 1 and investigated. The following findings emerged:

- Alcohol has the greatest correlation with quality.
- Density and residual sugar are inversely proportional, which can also be explained. Dissolved sugar increases the density.
- Density and alcohol have a directly proportional relationship. This can also be explained. Alcohol is lighter than water, more alcohol makes the liquid less dense.
- Residual sugar and alcohol are inversely proportional. This can be explained by the fact that the sugar is converted into alcohol. The more residual sugar is left, the less alcohol will have been made.

3.2. SVR

3.2.1. Datasets

The article indicates that the data is split into training data and testing data with a ratio of 3:1. This is also the ratio of the splits that was applied in further research.

3.2.2. Implementation Details

Reproduce the results
The results of the article are simulated by using the parameters mentioned in the article. The following parameters are mentioned: kernel=rbf, cost = 0.95 and Gamma = 0.13. For all other parameters, the default parameters of sklearn are chosen, except for the Tolerance for stopping criterion, which is set to 0.0001 to obtain the best possible result and because the cost is negligible given the small dataset. How the training data is chosen is unknown, so the split between training and test data is performed multiple times. In this way, there is a realistic picture of the models that generate these parameters.

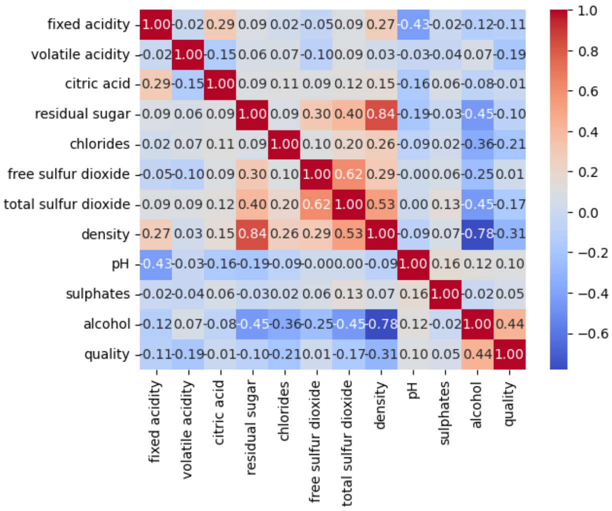

Figure 1. Correlation between the features of the white wine dataset

Impact of the features
The trained model used for analyzing the impact of features is the same as the model used for reproducing the results. When reproducing the results it was shown that the selection of a train and test set has an impact of on the model performance. The training set split used for the feature analysis results from applying `random_state=1`.

For the shap explainer the KernelExplainer was used. The KernelExplainer is a model-agnostic explainer that can be used to explain the output of any machine learning model. As background data a random sample of 200 rows from the training data was used. The KernelExplainer uses the background data to estimate the expected value of the model output. The KernelExplainer is a computationally expensive method, so a smaller sample size is used for the background data. Then a random sample of 50 rows as test data has been taken from the test data to explain the model output. The SHAP values are calculated for these 50 test samples. For higher accuracy both the background data and test samples

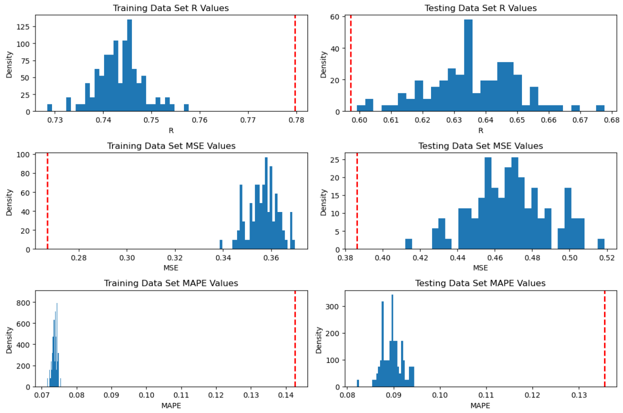

Figure 2. Comparison between the SVR model from the article (red dashed line) and the reproduced model (blue bars)

can be increased, but this will increase the computation time. Having only a home laptop available, the sample size is kept small to keep the computation time reasonable.

Both the summary plot and the dependency plot are used to visualize the impact of the features on the model output. These plots are both generated using the same SHAP values calculated for the test samples.

Improve the model

To optimize the model, the cost and gamma were examined. BayesSearch was used to find a model with the lowest possible mean squared error. The result is visualized in a scatter plot where green points indicate a good model and red points indicate a less good model. The new parameters will be used to train a model 100 times. Each time with different train and test data, but the ratio (3:1) will be maintained. These 100 trained models will be visually compared with the article (Dahal et al., 2021) model.

3.2.3. Results

Reproduce the results

The results are visualized in Figure 2. The important part is how the results are compared to the test data. The correlation and the MAPE give better results than the article (Dahal et al., 2021). While the MSE of the reproduced model are slightly worse.

Impact of the features

In Figure 3 the impact of all features is visualized. From this diagram it is clear that density has the most impact on the model output. The blue values indicate a low value of the feature and the red values indicate a high value of the feature. The impact on the model output is shown on the x-axis. For alcohol the red values are on the left side of the plot and the blue values are on the right side of the plot, meaning the impact is negative. Residual sugar and alcohol have a positive impact on the model output, but the impact is slightly smaller than for density. Volatile acidity has a negative impact on the model output. The features with a very weak correlation to the output are: chlorides, sulphates and fixed acidity. The impact of these features is very small compared to the other features. The SHAP values for those features are more chaotic and do not show a clear trend.

In figure Figure 4 the impact of the individual features are visualized. The dependency plot shows the impact of each feature on the model output. The color of the dots indicates the value of the interacting feature, with red indicating a high value and blue indicating a low value. The y-axis shows the impact of the feature on the model output.

The plot for alcohol shows that the impact of alcohol on the model output is positive. This means that a higher value of alcohol leads to a higher quality of wine. The trend of the values is clearly visible in the plot, showing a positive correlation between the feature and the model output. This is also confirmed by the Pearson correlation coefficient, which shows a positive correlation between alcohol and quality.

The plot for density show that the impact of density on the model output is negative. This means that a higher value of density leads to a lower quality of wine. Just as with alcohol, the trend of the values is clearly visible in the plot, showing a strong negative correlation between the feature and the model output.

One of the feature with the least impact on the model output is fixed acidity. The plot shows that the impact of fixed acidity on the model output is positive. This means that a higher value of fixed acidity leads to a higher quality of wine. However, the trend of the values is not clearly visible in the plot, showing a weak correlation between the feature and the model output. This is also confirmed by the Pearson correlation coefficient, which shows a weak positive correlation between fixed acidity and quality. Notable is that Free sulfur dioxide has a strong positive trend for the lower values. This suggests that the feature has a minimum value for the quality of wine. This was not visible in the pearson correlation coefficient because the correlation coefficient is calculated over the entire dataset.

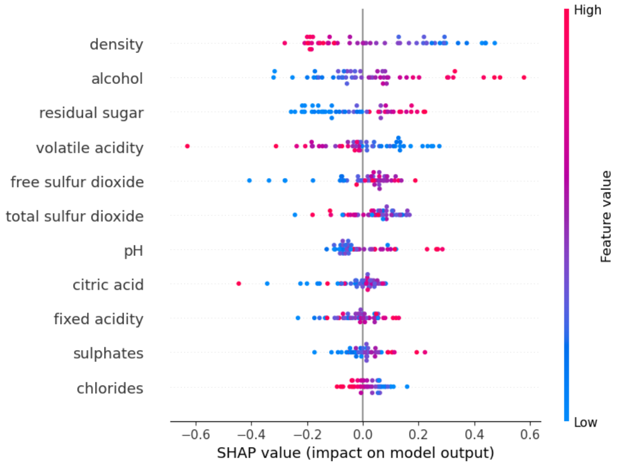

Figure 3. Summary plot of the features for SVR

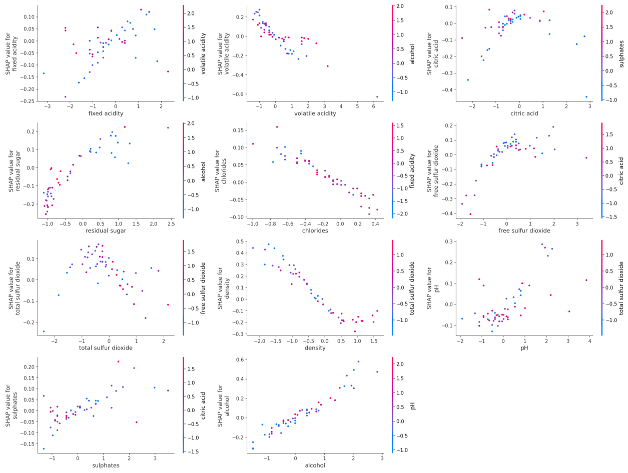

Figure 4. Dependency plot of the features for SVR

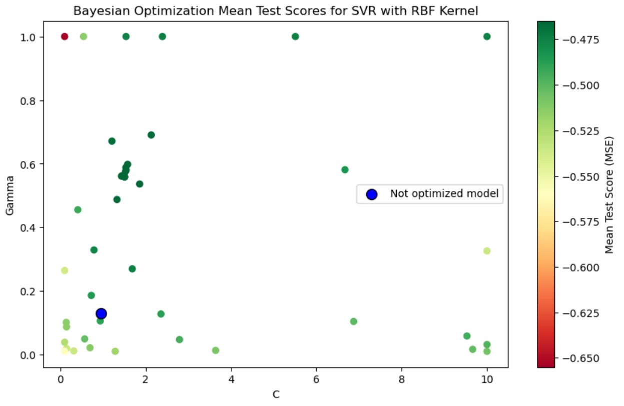

Figure 5. Bayesian search for optimizing gamma and cost. Green indicates better hyperparameters. Blue indicates the parameters from the article.

Improve the model

In Figure 5 the best possible values for gamma and cost are examined. The non-optimized model (the model with the parameters as indicated in the article (Dahal et al., 2021)) is visualized by a blue sphere.

The most optimal parameters are: cost=1.51 and gamma=0.558. In Figure 6 we see how the model performs compared to the values of the model of the article (Dahal et al., 2021) (dashed line) and compared to the non-optimized model (dotted line). Based on the three evaluation criteria (correlation, RMSE and MAPE) the model consistently outperforms the non-optimized model. The parameters perform significantly better based on the training data suggesting overfitting, but the results from the testing data are also (slightly) better indicating no overfitting.

3.2.4. DISCUSSION

The results of the article have been reproduced and are close to the reported performance based on the metrics. The performance was not the same, but this can be a result of different implementations used or different library versions. The impact of the features has been investigated. The results are in line with the pearson correlation coefficient. There is also some insights gained from the dependency plots which show non-linear relationships. Further research can be done by excluding features with less impact on the model output. All this while the article does not seem fully reliable because of the incorrect naming of the data set (namely red wine instead of white wine). Not having access to the original code used to generate the results makes it difficult to check the validity of the hyperparameters.

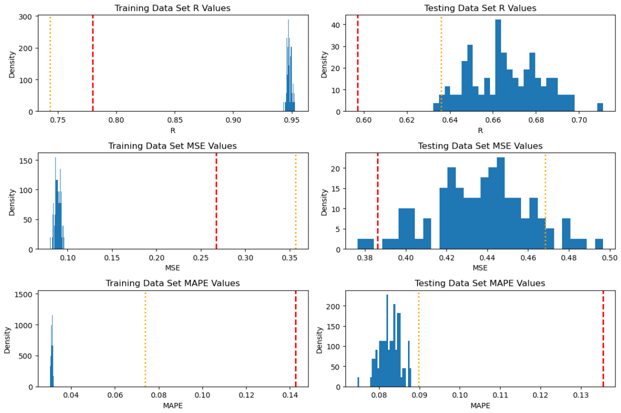

Figure 6. An optimized SVR model (blue bars) compared to the model from the article (red dashed line) and the reproduced model (orange dotted line)

3.3. ANN

3.3.1. Datasets

In the article (Dahal et al., 2021) it is stated that the train-validation-test ratio is 60-20-20. This ratio has been maintained for reproducing and optimizing the model.

3.3.2. Implementation Details

Reproduce the results

The article (Dahal et al., 2021) mentions the following parameters: 1 input layer, 3 hidden layers (each with 15 neurons) and one output layer with the Adam optimizer. As activation function relu was chosen and the mean squared error as loss to optimize. The following parameters were not specified so they were filled in with the following values: number of epochs = 50 and batch size = 16. How the training data is chosen is also unknown, so the split between training and test data is performed multiple times. In this way, there is a realistic picture of the models that generate these parameters.

Impact of the features

The analysis of the impact of the feature is done in the same way as with the SVR model. We need to be able to compare the results of the feature impact analysis, so the same KernelExplainer is used as with the SVR model. The KernelExplainer is a model-agnostic explainer that can be used to explain the output of any machine learning model. In contrast the DeepExplainer is a model-specific explainer that can be used to explain the output of a deep learning model. The performance of the DeepExplainer is far better compared to the KernelExplainer and seems to have a similar output as the KernelExplainer. The KernelExplainer is prioritized because it is model-agnostic and as such eliminates differences in result due to different Explainer implementations.

Improve the model

In a perfect world, all hyperparameters and all possible models would be tested. However, due to limited time and access to resources, it was decided to look at the number of hidden layers and the number of neurons. Gridsearch will be used to see if the model can be optimized by adjusting the number of neurons in the hidden layers. It was also considered to reduce the number of hidden layers to one. The reasoning behind this is that the data is not that complex and that the complexity of the three layers would be unnecessary. With one hidden layer, the impact of the number of neurons is also investigated using Gridsearch. And once an optimum has been found, the impact of batch size and epochs is also examined.

3.3.3. Results

Reproduce the results

In Figure 7 the original model is indicated by a red dashed line. 100 models (with the selected parameters) were trained and shown as blue bar charts. The average of the bar charts is shown by an orange dotted line. The R values are very similar. The MSE results are worse than those of the model from the article while the MAPE results are better than those from the article. The combination of a higher MSE and a lower MAPE may indicate that most forecasts are accurate, but there are some exceptions with large errors that increase the MSE. This must be put into perspective, because we are talking about very small differences compared to the original model. If we know that the quality has no significant figures, the differences compared to the original model are negligible.

Impact of the features

The summary plot for the ANN model is shown in figure Figure 8. In this summary plot it is shown that the most important feature is the alcohol as this one is shown on top. The impact of the alcohol is positive, which means that a higher value of alcohol leads to a higher quality of wine. Next are residual sugar and density. The impact of residual sugar is positive, while the impact of density is negative.

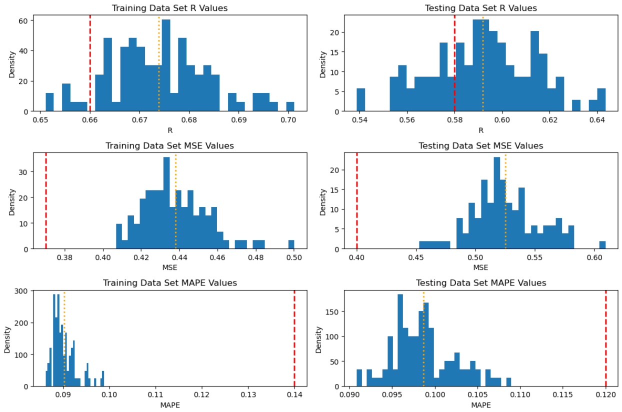

Figure 7. An reproduced ANN model (blue bars) compared to the model from the article (red dashed line). The average of the reproduced model is also indicated with an orange dotted line.

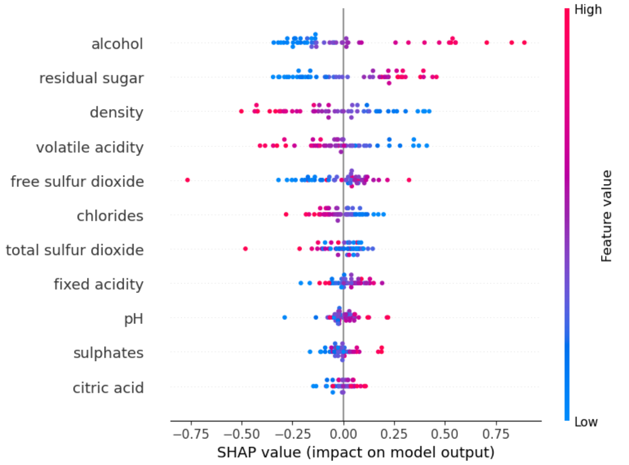

Figure 8. Summary plot of the features for ANN

This is in line with the expectations from the dataset analysis. Residual suger and density are naturally correlated. Other features can be read in the same way as being positive (blue to red) or negative (red to blue), but are less important than the three features mentioned above. The features lower in the graph have a very small impact on the model output. The impact of these features is very small compared to the other features.

Fixed acidity, pH, sulphates and citric acid show more chaotic values and do not show a clear trend. This can also be seen int the dependency plot in figure Figure 9. Where other diagrams how a clear upward or downward trend, the dependency plot for these features shows a more chaotic picture. This choatic picture results in a low correlation with the model output.

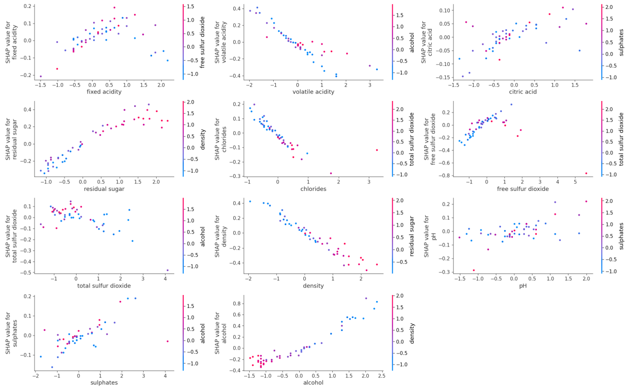

Figure 9. Dependency plot of the features for ANN

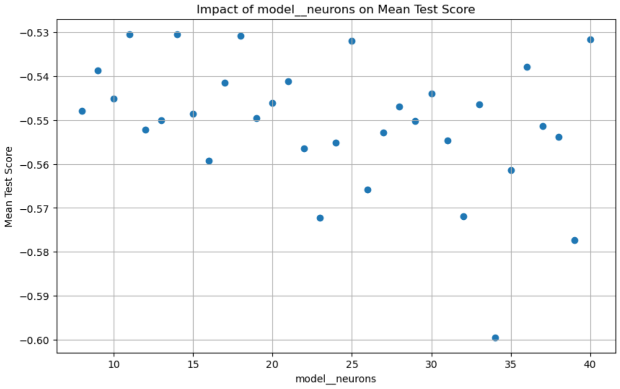

Figure 10. The effect of number of neurons on an ANN model with 3 hidden layers

Improve the model

Based on Figure 10 it can be concluded that the model with 15 neurons for each layer is quite well chosen. Increasing the number of neurons will certainly not give better results.

With one hidden layer the story is different, see Figure 11. The model improves as the number of neurons increases with a flattening starting around 70 neurons. With these 70 neurons the impact of batch size and epochs was investigated with the aim of finding a better model.

Figure 12 and Figure 13 show the results of a grid search for a better model around number of epochs and batch size. As expected we see that we get better models at higher epochs and smaller batch sizes. For the final model epochs = 40 and batch size = 16 were chosen.

Figure 14 shows the model from the article with a red dashed line and the reproduced model with an orange dotted line.

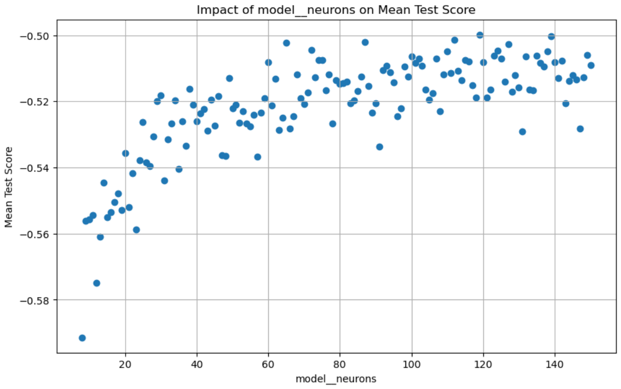

Figure 11. The effect of number of neurons on an ANN model with one hidden layer

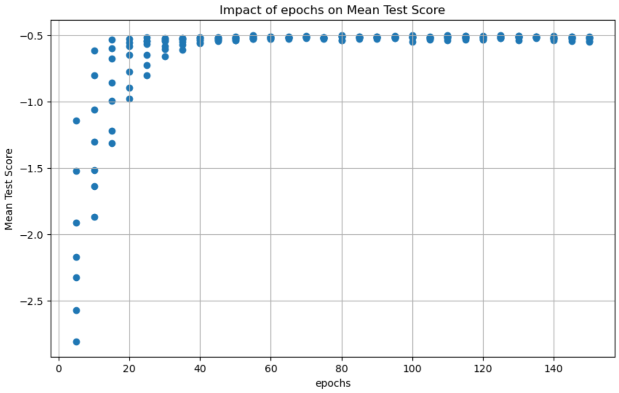

Figure 12. The effect of number of epochs on an ANN model with one hidden layer

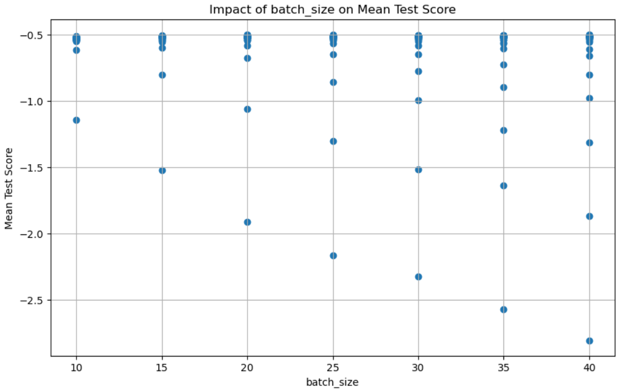

Figure 13. The effect of number of batch size on an ANN model with one hidden layer

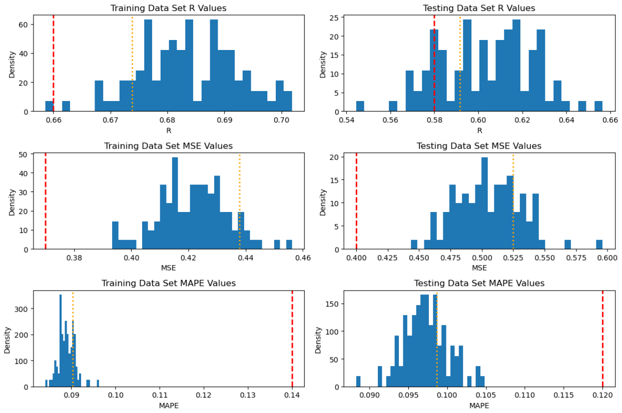

Figure 14. An optimized ANN model (blue bars) compared to the model from the article (red dashed line) and the reproduced model (orange dotted line)

The blue bars show an optimized model with one layer (with 70 neurons). Both the R values, MSE and MAPE give better results than the reproduced model. This improvement is only minimal.

3.3.4. Discussion

The article has been reproduced and the results are close to the reported performance based on the metrics. The performance was not the same, but this can be a result of different implementations used or different library versions. The impact of the features has been investigated. The results are in line with the pearson correlation coefficient. When searching for improvements on the model, it was found that the number of neurons in the hidden layers had a significant impact on the model performance. Setting the number of neurons to 70 in the hidden layer with one hidden layer gives a better model than the model with three hidden layers and 15 neurons.

3.4. Ensemble model

3.4.1. Datasets

For the ensemble model, the same dataset is used as for the individual models. Again using the white wine dataset. The data is split into training, testing and validation data with a ratio of 60-20-20. The same training and testing data is used to train both the SVR and ANN models.

3.4.2. Implementation Details

The ensemble model is created by combining the scilearn SVR model and the keras/tensorflow Sequential ANN model. These models are not compatible, so the scikit facilities for combining and training two models can not be used.

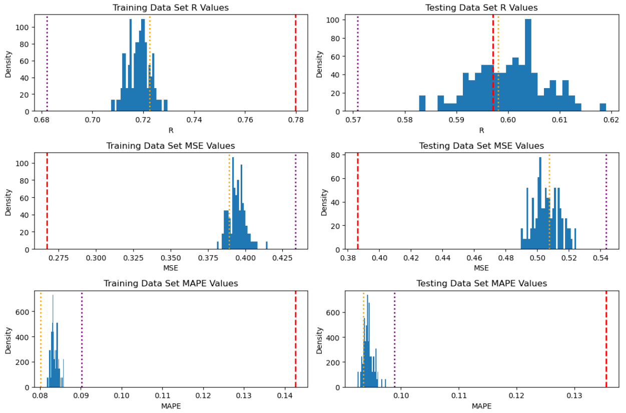

Figure 15. An ensemble model (blue bars) compared to the SVR model (orange dotted line) and the ANN model (purple dotted line)

Each model is trained separately using the same train and test data. The keras/tensorflow model uses also the validation data. The predictions of both models are combined by averaging the predictions. To ensure that the model performance is affected by the provided training set, the training, testing and validation data is split multiple times.

3.4.3. Results

in figure Figure 15 the ensemble model is compared with the individual models. The individual model results are shown as a red dashed line for the article results, an orange dotted line for the reproduced SVR model and a purple dotted line for the ANN model. As expected the results for the ensemble model are between the results of the individual models.

Discussion

The ensemble model can be usefull to combine the strengths of both models. But the implemented ensemble model with averaging the predictions does not show a significant improvement compared to the individual models. It can be usefull to eliminate extreme values for one of the models. Further research can be done to find performance difference between the individual models on subsets of the data. Then an ensemble model could benefit from those differences. For example, if one model performs better for low quality wines and the other better for high quality wines, the ensemble model could be trained to use the best model for each subset of the data. In this report no such observations were made.

4. Citing references

References

Dahal, K. R., Dahal, J., Banjade, H., and Gaire, S. Prediction of wine quality using machine learning algorithms. *Open Journal of Statistics*, 11(2):278–289, 2021.

Lundberg, S. M. and Lee, S.-I. A unified approach to interpreting model predictions. *Advances in neural information processing systems*, 30, 2017.