Machine Learning Models on the wines dataset

Qingsong Tan, *grkp39, 001055343*
E-mail: grkp39@durham.ac.uk

Abstract—This is a report on a study of the quality of white and red wines, analysing their quality scores in relation to their physico-chemical properties (including fixed acidity, citric acid, etc.). What's interesting about this project is its ability to judge the taste of a wine by its specific physico-chemical properties, which is very friendly to the average consumer who can judge whether a wine is good or not by simply judging the quality score without having to delve into a list of ingredients.

Index Terms—Computer Science, AI, Machine Learning Models, Machine Learning Systems, Wines' Quality, Decision Tree, Multivariate Linear Regression, Regression.

1 Introduction: Background

THE fact is that alcohol is everywhere in our lives, whether it's at parties, festivals or other occasions, we can't live without it. On these happy days, alcohol is a catalyst to get strangers talking, while in some extremely sad situations, such as a lost love, alcohol can also paralyse us and make us feel less miserable. At this point, there is nothing the average person wants more than to pick a great tasting alcohol, but they don't want to bother picking one, so this quality score will help them choose.

So I wanted to use a machine learning regression algorithm to achieve this, where as long as you can provide the 11 physical-chemical properties within the data, it will automatically output a quality score, with higher scores indicating better quality wines, and lower scores indicating poorer quality.

Qingsong
Jaunary 1, 2024

1.1 Making a Machine Learning System

To effectively address this issue, we must first understand the comprehensive workflow involved in machine learning. This process begins with the meticulous preparation and pre-processing of data, ensuring it is primed for analysis. The next critical step is selecting the most appropriate algorithms tailored to our specific objectives. Following this, we embark on training the model, utilizing our prepared data to develop its predictive capabilities. After training, the model undergoes rigorous testing to evaluate its accuracy and reliability. The final stage involves fine-tuning the model's parameters based on its performance during testing. This iterative process of adjustment is crucial to enhancing the model's precision. By diligently following these steps, we can develop a robust and high-quality wine prediction model, capable of delivering reliable and insightful results.

1.1.1 Data Preparation

After defining the research question, I began the data preparation phase. The dataset I chose was 'Wine Quality', which included separate datasets for red and white wines (I will cite the web page where I chose the data in the citation). Initially, I checked the data for missing and duplicate values and, reassuringly, I did not find any missing and duplicated values. Although no missing or duplicate values were found, some outliers were detected. Since it's unclear whether these data values are accurate or anomalies, I plan to leave the outliers unprocessed in order to maintain the integrity of the data. In order to compare the characteristics of the data more easily and in depth, I used Seaborn's visualisation tool to create 12 violin plots, It shows the distribution of all the data, it seems that all of the features data conform to a normal distribution. (as shown in Figure 1).

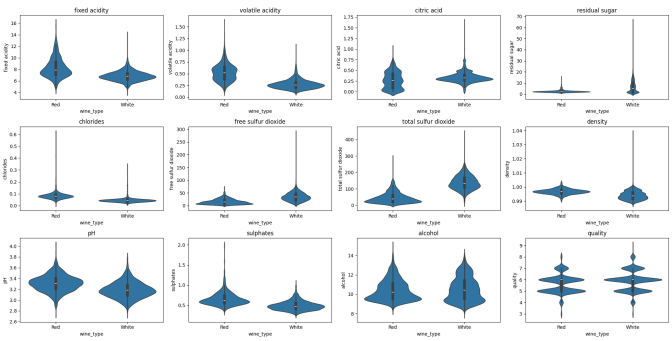

Fig. 1. Here's a violin chart of the physical-chemical characteristics and quality scores of red and white wines

In addition, to explore the intricate relationship between the various physical-chemical properties and the quality of wine, heat maps for both red and white wines were constructed, as depicted (in Figures 2 and 3). These visual representations uncovered significant correlations. Most notably, there exists a discernible positive correlation between alcohol content and wine quality in both types of wine. This indicates that wines with higher alcohol content tend to be of better quality. Conversely, for attributes that exhibit a negative correlation with wine quality, such as density, the relationship is inverse; wines with higher density are often found to be of lower quality. Of course there is more than just alcohol and density, you can also find other characteristics and quality correlations among these two heat maps.

This thorough preparation and data preprocessing pro-

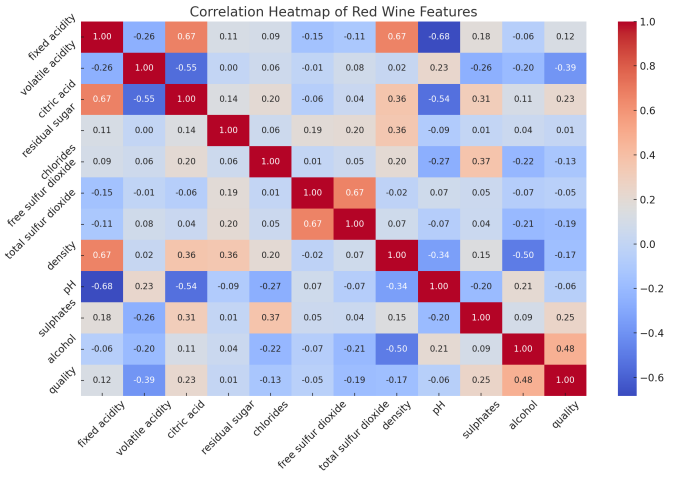

Fig. 2. This is a heat map of the correlation between the quality of red wines and 11 other physical and chemical characteristics.

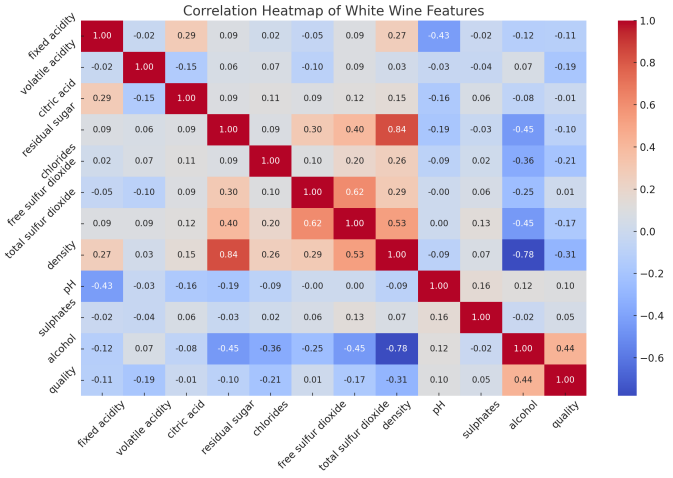

Fig. 3. This is a heat map of the correlation between the quality of white wines and 11 other physical and chemical characteristics.

vided a solid foundation for the subsequent analysis phase.

1.1.2 *Algorithm Selection*

After selecting and preprocessing the data is done, next it is time to choose a suitable machine learning regression model to build a regression to predict the quality of the wine, here I found that 11 physical-chemical characteristics are involved, so Multivariate Linear Regression is a good choice because Multivariate Linear Regression is specially designed for more than variables. You can observe that this equation in Figure 4 is a multivariate linear regression prediction function, in this wine column n=11 and the final h(x) is the predicted quality of the wine.

In fact, we looked at the other features of the heat map and the wine quality correlation and found that there was actually not much of a linear relationship between them, so I was able to make use of another machine learning model, the decision tree. Decision tree models are great for dealing with non-linear relationships between variables [3]. Decision trees capture these complex relationships by

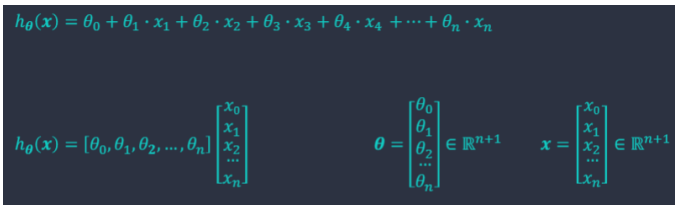

Fig. 4. This is the Hypothesis Function [2]

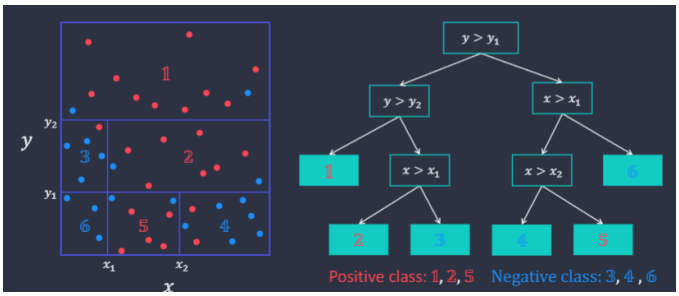

Fig. 5. This is the process of decision tree model [2]

segmenting the data. Figure 5 below shows the flowchart of a decision tree regression.

I've only considered these two regression models at the moment, we can build the model and then test their performance separately to see whose predictability is more accurate so as to update the model, of course there should be more than just these two models that can be predictive, perhaps there are better models to choose from, I will try more model choices if there will be a chance to continue the research in the future.

1.1.3 *Model Training and Testing*

After selecting the regression model, I set the multivariate linear regression model as the baseline method and the decision tree model as the control method to compare the performance with the former. Next I need to divide the data into two parts, i.e. test data (30 percentage), and training data (70 per cent). Of course the red and white wines have to be modelled separately. So I start by calculating their root mean square error:

I calculated the RMSE (Root Mean Squared Error is a commonly used metric to assess the accuracy of regression models. It represents the standard deviation of the differences between predicted values and actual values.) of the multivariate linear regression model for red wine to be 0.641, compared to 0.795 for the decision tree, and 0.745 for white wine, compared to 0.841 for the decision tree. Based on the RMSEs alone, it seems that the multivariate linear regression model outperforms the decision tree model, with a lower RMSE, but let's take the test of the data. However, bringing in the other thirty percent of the data used for testing, we found that the multivariate linear regression model for red wines had a prediction accuracy of only 65.42 percentage, whereas the decision tree model had an accuracy of 77.50 percentage; and the decision tree model had a higher accuracy than the multivariate linear regression model for white wines too, with an accuracy of 77.62 percentage and 65.10 percentage, respectively. (Figure

Fig. 6. The calculation formula for RMSE is:

$$\text{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}$$

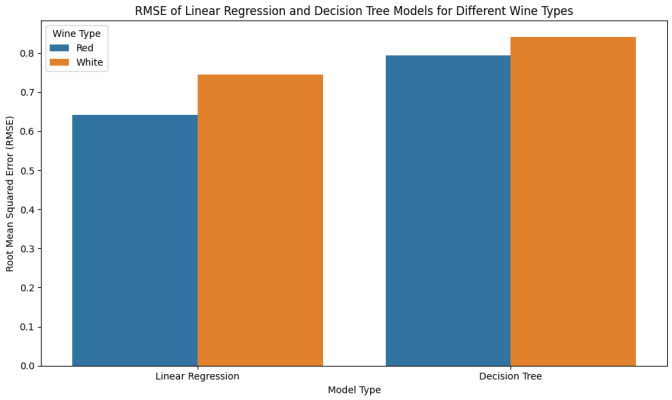

Fig. 7. This graph is a histogram created after calculating the RMSE, which facilitates the comparison of the two types of wines and different regression models.

7 and 8 are histograms of RMSE and prediction accuracy, respectively, allowing for a straightforward comparison.)

These results suggest that although the multivariate linear regression model performed better in terms of predictive accuracy in terms of RMSE metrics, the decision tree model had a slight advantage in terms of accuracy on these particular datasets, which means the experiment shows that the prediction accuracy of the decision tree model is still higher than that of the multivariate linear regression. . It is important to note that accuracy may not be the best metric for evaluating these types of regression models because it requires the conversion of continuous predictions into discrete categories.

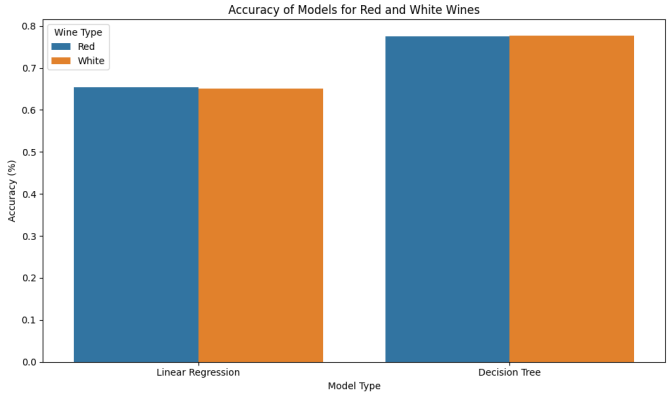

Fig. 8. This graph is a histogram created after calculating the prediction accuracy, which facilitates the comparison of the two types of wines and different regression models.

1.1.4 Hyperparameter Tuning and Prediction

The choice of these two models is just the beginning. The multivariate linear regression model, due to its simplicity, offers limited scope for hyperparameter tuning. However, when the performance of the linear model is not particularly strong, one option is to include higher-order terms of the original features in the model. This can help the model capture non-linear relationships in the data. Alternatively, using Ridge Regression (L2 regularization) or Lasso Regression (L1 regularization) can prevent overfitting. Ridge Regression tends to shrink the parameter values, whereas Lasso Regression can reduce some parameters entirely to zero, thus facilitating feature selection.

However, the decision tree model offers more room for parameter tuning. 1. Tree's maximum depth ('max_depth'), the hyperparameter max_depth controls the overall complexity of the decision tree. This hyperparameter allows for a trade-off between underfitting and overfitting decision trees; 2. Minimum number of samples required to split a node ('min_samples_split'), the minimum number of samples a node must have before it can split. Increasing this value can prevent the model from learning patterns that are too specific; 3. Minimum number of samples required at a leaf node ('min_samples_leaf'), the minimum number of samples a leaf node must have. Increasing this value can smooth model predictions and help prevent overfitting. I can use for the grid search tuning method, which involves traversing a predefined grid of parameters to find the best parameter combination [1].

After finding the optimal parameters, we recalculated the RMSE for red wine to be 0.683 and for white wine to be 0.757. Compared to the original RMSEs of 0.795 for red wine and 0.841 for white wine, there was a significant reduction in the root mean square errors, indicating that our hyperparameter tuning was successful.

From the perspective of both root mean square error and prediction accuracy, the decision tree model outperforms the multivariate linear regression model. Therefore, among these two models, the decision tree model, after hyperparameter tuning, is the most suitable choice. It's noteworthy that although the decision tree model is quite apt for predicting the quality of red and white wines, if one aims for more in-depth research, such as studying how to categorize wines into high and low quality for differentiated marketing, logistic regression can be considered as the primary model for accomplishing the classification task.

2 Conclusion

The narrative presented here encapsulates my comprehensive journey through a machine learning research project. It commenced with the meticulous selection of the dataset, followed by a rigorous preparation and preprocessing phase. This phase involved thorough checks for duplicates and missing values, ensuring data integrity and reliability. To facilitate an easier comparison of the various features, I employed Seaborn's visualization tool to craft 12 detailed violin plots. These plots not only enhanced the clarity of the data's distribution but also allowed for a straightforward comparison across different attributes. Furthermore, I generated a heatmap to gain a more intuitive grasp of the correlations

between the diverse physicochemical characteristics of wine and their overall quality.

The next is that I delved into the selection of suitable machine learning models. Given the unique characteristics of the data and the specific objectives of my study, I chose to focus on two models: the multivariate linear regression and the decision tree models. My approach to evaluating these models was twofold. Firstly, I assessed their performance based on the root mean square error (RMSE), a metric that quantifies the model's accuracy in predicting outcomes. Secondly, I examined their prediction accuracy, a crucial metric that measures the model's ability to correctly predict outcomes in real-world scenarios.

Through a careful and systematic comparison of these two models from both the RMSE and prediction accuracy standpoints, I arrived at a significant conclusion. The decision tree model, particularly after thorough hyperparameter tuning, emerged as the superior choice. This finding underscores the importance of not just model selection but also the fine-tuning of models to optimize performance. My research journey in machine learning, thus, culminates with the validation of the decision tree model as the most effective tool for predicting wine quality, illustrating the intricate interplay between data preparation, model selection, and hyperparameter optimization in the realm of machine learning.

3 Self-Evaluation Report

3.1 learned from the lectures

This lecture opened a door to a new world for me, I learnt about the three types of machine learning and the workflow of machine learning. According to my understanding, the first step is to process the data, which I think is the key step after I learnt it. If we compare machine learning to building a house, then the data is equivalent to the bricks used to build the house. So after the initial data selection and pre-processing, you must get a clean data. The second part is the selection of the machine learning model, which will be decided according to the research objectives, but there are mainly two parts one is supervised learning which includes regression and classification, which involves linear regression, logistic regression, decision tree etc. The other part is the unsupervised learning which includes both clustering and dimensionality reduction. The third step is to get a complete model after training and testing. The last step is the hyperparameter tuning of the model, which can be understood as refining the model.

3.2 learned from the coursework

In fact, when I was doing this coursework I strictly followed the machine learning workflow, so I didn't learn anything new, but through this assignment I built two models by myself hands-on, which deepened my understanding of the specific methods of machine learning (Multiple Linear Regression and Decision Tree Modelling), which was very meaningful to me, and meant that I was able to try to independently start to try to build a complete machine learning process.

3.3 difficulties in the module

I think the harder part in this module is when constructing a model I need to learn the code of the model in question, for example before writing this assignment I didn't know how to build a multivariate linear regression model and a decision tree, I learnt how to build the model after consulting a lot of information and web sites, I think even if I learnt this it would be a long term job as there are still many more models waiting for me to try and that's the reason why I said that after I go through this assignment I can try to build my own machine learning models.

3.4 differently if do again

As I said after building the model, if I had the opportunity I would want to try a different direction of research for example studying the classification of high grade level quality wines and low grade quality wines for sale, a situation where logistic regression could obviously be used to do the classification, or more other models. I think that's something I might be interested in if I redo it.

3.5 unique contributions or novel ideas

There doesn't seem to be any unique contributions or novel ideas for me as a beginner, but I'll be exploring more as I try to do more research, so maybe I'll have a lot of new ideas when that time comes.

References

[1] FermatSavant. Decision Tree High ACC using GridSearchCV [Internet]. Kaggle. [cited 2024 Jan 1]. Available from: https://www.kaggle.com/code/fermatsavant/decision-tree-high-acc-using-gridsearchcv

[2] University Blackboard. Available from: https://blackboard.durham.ac.uk/ultra/courses/_54394_1/outline/file/_1853829_1

[3] scikit-learn developers. DecisionTreeClassifier [Internet]. scikit-learn. [cited 2024 Jan 1]. Available from: https://scikit-learn.org/stable/modules/generated/sklearn.tree.DecisionTreeClassifier.html

[4] UCI Machine Learning Repository. Wine Quality Data Set [Internet]. [cited 2024 Jan 1]. Available from: https://archive.ics.uci.edu/ml/datasets/wine+quality