Final project – machine learning - OVERSAMPLING

IMPORTANCE OF SCALING

This project utilises two datasets related to red and white wine samples from the UC Irvine Machine learning database. These wine samples are vinho verde ('green wine) from the northwestern Minho region of Portugal. The Vinho Verde Viticulture Commission (CVRVV) oversees regulation and control of wine quality from this region. The datasets used by this model were collected from May 2004 to February 2007 using samples tested by CVRVV. Each sample was evaluated by a minimum of three testers, with a score given from 0-10 with 0 being 'very bad' and 10 being 'excellent', this data was recorded by a computerised system and exported onto a csv.

This project aims to use random forest regressors and support vector regressors to accurately predict wine quality and most important features, examine differences between the models, and to address the consequences of unbalanced data.

Wine quality monitoring is important to determine pricing, ensure safety, and regulatory compliance. Wine quality changes are particularly important, given climate change will affect key factors (temperature, precipitation, disease) that control grape quality and yield. Further, given that wine is culturally, historically, and economically significant to many wine-growing regions, a decrease in wine quality may drastically impact the region. The wine sector in Portugal is a major employer, with approximately 150,000 (15% of people employed in agriculture) employees involved, globally Portugal is one of the top 10 wine exporters http://166.62.7.99/assets/default/article/2024/04/15/article_1713235616.pdf . Therefore, monitoring wine quality, and the dominant features that predict quality is essential.

Data

This project utilises two datasets related to red and white wine samples from the UC Irvine Machine learning database. These wine samples are vinho verde ('green wine) from the northwestern Minho region of Portugal. The Vinho Verde Viticulture Commission (CVRVV) oversees regulation and control of wine quality from this region. The datasets used by this model were collected from May 2004 to February 2007 using samples tested by CVRVV. Each sample was evaluated by a minimum of three testers, with a score given from 0-10 with 0 being 'very bad' and 10 being 'excellent', this data was recorded by a computerised system and exported onto a csv.

Before the modelling, stage, data frames and data descriptions were printed (Figure 1) to aid in choosing the most suitable model, and to determine whether reprocessing is required. The wine datasets contain the target variable, quality, and 10 continuous

features: fixed_acidity, volatile_acidity, citric_acid, residual_sugar, chlorides, free_sulfur_dioxide, total_sulfur_dioxide, density, pH and sulphates. There are no missing values in these datasets. Given that wine quality is given as an integer between 0-10, this data will lend itself to regression models. Both datasets have the same features. Therefore, it is useful to combine the datasets, providing more data to the models and increasing the number of extreme values. When combining the datasets, another feature 'is_red' is added, with red wine instances given a score of 1, and white wine instances given a score of 0. This enables investigation into whether wine colour affects quality.

A

	fixed acidity	volatile acidity	citric acid	residual sugar	chlorides	free sulfur dioxide	total sulfur dioxide	density	pH	sulphates	alcohol	quality	is_red
0	7.4	0.700	0.00	1.9	0.076	11.0	34.0	0.99780	3.51	0.56	9.4	5	1.0
1	7.8	0.880	0.00	2.6	0.098	25.0	67.0	0.99680	3.20	0.68	9.8	5	1.0
2	7.8	0.760	0.04	2.3	0.092	15.0	54.0	0.99700	3.26	0.65	9.8	5	1.0
3	11.2	0.280	0.56	1.9	0.075	17.0	60.0	0.99800	3.16	0.58	9.8	6	1.0
4	7.4	0.700	0.00	1.9	0.076	11.0	34.0	0.99780	3.51	0.56	9.4	5	1.0
...	...	...	...	...	...	...	...	...	...	...	...	...	...
1594	6.2	0.600	0.08	2.0	0.090	32.0	44.0	0.99490	3.45	0.58	10.5	5	1.0
1595	5.9	0.550	0.10	2.2	0.062	39.0	51.0	0.99512	3.52	0.76	11.2	6	1.0
1596	6.3	0.510	0.13	2.3	0.076	29.0	40.0	0.99574	3.42	0.75	11.0	6	1.0
1597	5.9	0.645	0.12	2.0	0.075	32.0	44.0	0.99547	3.57	0.71	10.2	5	1.0
1598	6.0	0.310	0.47	3.6	0.067	18.0	42.0	0.99549	3.39	0.66	11.0	6	1.0

B

	fixed acidity	volatile acidity	citric acid	residual sugar	chlorides	free sulfur dioxide	total sulfur dioxide	density	pH	sulphates	alcohol	quality	is_red
0	7.0	0.27	0.36	20.7	0.045	45.0	170.0	1.00100	3.00	0.45	8.8	6	0.0
1	6.3	0.30	0.34	1.6	0.049	14.0	132.0	0.99400	3.30	0.49	9.5	6	0.0
2	8.1	0.28	0.40	6.9	0.050	30.0	97.0	0.99510	3.26	0.44	10.1	6	0.0
3	7.2	0.23	0.32	8.5	0.058	47.0	186.0	0.99560	3.19	0.40	9.9	6	0.0
4	7.2	0.23	0.32	8.5	0.058	47.0	186.0	0.99560	3.19	0.40	9.9	6	0.0
...	...	...	...	...	...	...	...	...	...	...	...	...	...
4893	6.2	0.21	0.29	1.6	0.039	24.0	92.0	0.99114	3.27	0.50	11.2	6	0.0
4894	6.6	0.32	0.36	8.0	0.047	57.0	168.0	0.99490	3.15	0.46	9.6	5	0.0
4895	6.5	0.24	0.19	1.2	0.041	30.0	111.0	0.99254	2.99	0.46	9.4	6	0.0
4896	5.5	0.29	0.30	1.1	0.022	20.0	110.0	0.98869	3.34	0.38	12.8	7	0.0
4897	6.0	0.21	0.38	0.8	0.020	22.0	98.0	0.98941	3.26	0.32	11.8	6	0.0

Figure 1A and 1B. printed data frames for red and white wine data, both have the same features meaning they can be combined with ease

	quality
count	6497.000000
mean	5.818378
std	0.873255
min	3.000000
25%	5.000000
50%	6.000000
75%	6.000000
max	9.000000

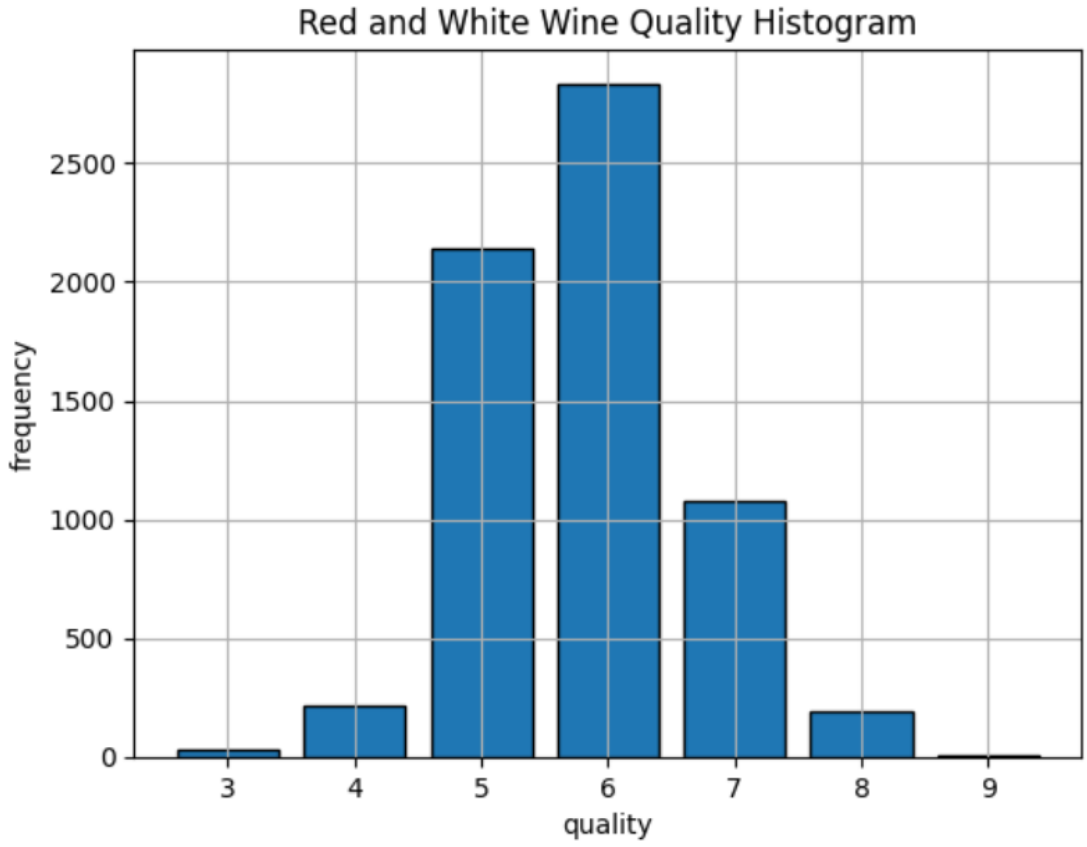
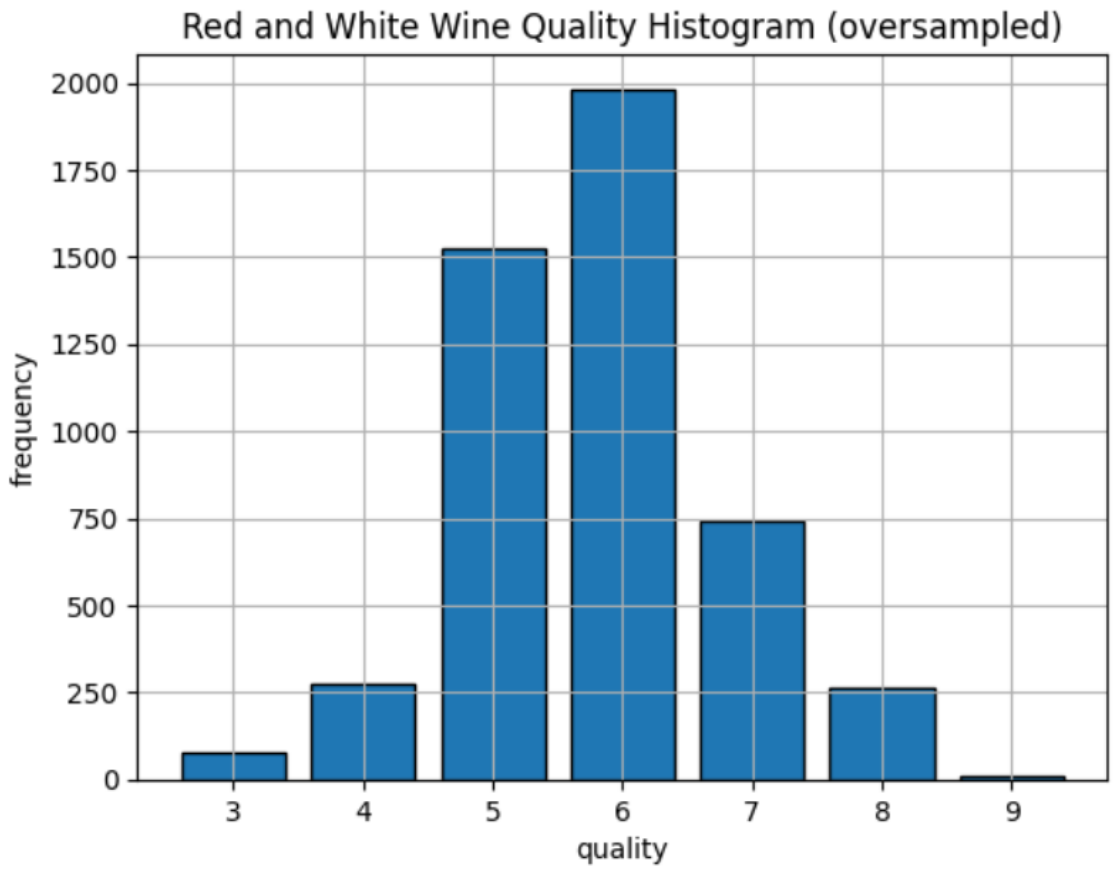


Figure 1. Histogram of red and white wine after combining both datasets. The unbalanced nature of the dataset is evident, with very little data for quality ratings of 3 and 9

To increase the weight of the underpredicted values, 'rare' quality values, were doubled (qualities 4, and 8), and quadrupled (qualities 3 and 9), to produce a new dataset, 'df_duplicated_data.' This followed splitting the data into train and test data to ensure duplicated values are only present in the training data, preventing data leakage. This

data frame will be analysed separately to evaluate the response of the models to more balanced data.



Modelling:

Two models, random forest regressor and support vector regression were used for this project. For modelling, 30% of the data was used for testing the model, and 70% was used for training, striking the balance between size of the training data and potential overfitting https://scholarworks.utep.edu/cs_techrep/1209/ .

Random forest Regressor is a supervised ensemble learning method, built from a collection of many decision trees. Each tree is tree is trained on a bootstrap sample, and a random subset of features is used at each split. This increases model diversity and reduces overfitting. Hence, random forest regressors are robust to noise, capable at handling complex relationships, and generally have high classification accuracy. https://link.springer.com/chapter/10.1007/978-3-642-34062-8_32. A benefit of using random forests is the ease of determining feature importances, which were computed for the 'best_model.'

As this model underpredicted(for high qualities) and overpredicted (for low qualities) the data, I modified the following hyperparameters/parameters: oob_score, max_depth,

and n_estimators, on a trial and error approach, choosing the best combination by observing the change in error metrics as I changed these parameters. I found that increasing the max_depth to 1000 and increasing n_estimators to 300 helped reduce the RMSE by approximately 0.05, beyond which, any changes were minimal (<0.01).

Support vector regression (SVR) is a supervised learning method that predicts continuous target values by fitting a function within a specific tolerance. Support vectors define a large margin around the predictions, making them robust to noise. SVR is well-suited to small to medium datasets (up to 10,000 samples) making it an ideal choice for the wine data.

To help fine-tune this model, I decided to modify C, where increasing C fits makes the model fit data more closely (risks overfitting), whereas decreasing C increases regularisation (default = 1) to help reduce underfitting. However little change was observed in the error metrics upon this modification and therefore the default parameters were used.

For this report, Scikit-learn's pipelines were used to scale the data set, the pipelines were then instantiated and trained on the data. To reduce overfitting, I used cross validation, which repeatedly trains and tests the model on different subsets of the data, thus producing the 'best_model.' Several error metrics: RMSE, R^2, mean absolute error and regression were then calculated enabling evaluation of the model performance. Regression slopes for each model were then plotted, with translucent data points to enhance visual clarity. These slopes were compared to a line where all predictions made by the model correspond exactly to the actual y values to visualise how closely each model follows the ideal perfect prediction line.

Next, REC curves were plotted for each model, showing for a given error, the percentage of correct predictions. Following this, feature ranking was conducted to determine feature importances. The exact same steps were conducted on the oversampled data.

Results and Discussion

The following results were obtained from the error metrics. For the random forest, the root mean square error (RMSE) and mean absolute errors were 0.63 and 0.46 respectively, showing that, on average the model predictions deviate by around 0.63 or 0.46 from the actual values depending on the chosen error metric. The R^2 metric was 0.48 showing that 48% of the variance in wine quality is explained by the model. For the SVR, the root mean square error (RMSE) and mean absolute errors were 0.68 and 0.51 respectively, The R^2 metric was 0.41 showing that 41% of the variance in wine quality is explained by the model.

For both regressions, the slope differs significantly from the 'perfect prediction.' This is particularly significant for the lowest and highest qualities, where low qualities are overpredicted and high qualities are underpredicted.

The REC curves for both SVR and random forest regressor are very similar, with the random forest regressor having a slightly higher prediction accuracy. At an error tolerance of 1, both models predict the quality within an error of 1, at a tolerance of 2, almost 100% of the data is predicted within an error of 2.

The most important features by a significant margin is alcohol, and volatile acidity, it is evident that wine colour has very little effect on wine quality (FIGURE)

Very similar results were obtained from the error metrics for the oversampled data. For the random forest, the root mean square error (RMSE) and mean absolute errors were 0.63 and 0.46 respectively, showing that, on average the model predictions deviate by around 0.63 or 0.46 from the actual values depending on the chosen error metric. The R^2 metric was 0.48 showing that 48% of the variance in wine quality is explained by the model. For the SVR, the root mean square error (RMSE) and mean absolute errors were 0.68 and 0.51 respectively, The R^2 metric was 0.41 showing that 41% of the variance in wine quality is explained by the model.

The steps above were conducted on the 'df_duplicated_data,' without feature ranking which is based off the true data.

The REC curves for the model ran on the oversampled data show the REC curve for the random forest regressor is slightly improved with approximately 90% of wine quality correctly predicted within an erorr interval of 1. The SVR model ran on the oversampled data has a similar curve to the real dataset.

In this context, the RMSE score will be used as it is the preferred metric for normally distributed data https://gmd.copernicus.org/articles/15/5481/2022/#bib1.bibx5 .

This value is significant, as quality rankings are based on taste, which is complex and varies from person to person. https://books.google.com/books?hl=en&lr=&id=VaytbeaLY0AC&oi=fnd&pg=PA41&dq=is+taste+subjective&ots=Ryi4pbdpwN&sig=nj-CbgrW7bHO6pa_D8EIeI7To0I#v=onepage&q=is%20taste%20subjective&f=false.

Using the random forest regressor proved very useful, as feature ranking could be employed, this was conducted using the 'best_model' obtained from cross validation.

The steps above were conducted on the 'df_duplicated_data,' without feature ranking which is based off the true data.

The REC curves for the model ran on the oversampled data show the REC curve for the random forest regressor is slightly improved with approximately 90% of wine quality correctly predicted within an erorr interval of 1. The SVR model ran on the oversampled data has a similar curve to the real dataset.

Dicussion

This is likely due to the balance of data shown in the intial histogram, where there is a central tendency of data of qualities from 5-7. These qualities are generally predicted accurately (fig XX)

Overprediction of low wine qualities and underprediction of high wine qualities is due to the unbalanced nature of the selected dataset, where quality values are normally distributed resulting little datapoints for the extremes. This makes up the majority of the error (see figure REGRESSION VS PERFECT) in predictions.

Upon running the models on the dataset containing doubled/quadrupled datapoints, the random forest regressor is improved, COMPARE RMSE R^2 ETC.. ,

Conclusion