Introduction

In this report, we are going to analyse two data sets related to red and white variants of the Portuguese "Vinho Verde" wine from UC Irvine's Machine Learning Repository. It will be interesting to observe if we will be able to distinguish which wine is of which color wine based on other physicochemical features when we combined the datasets. The summary of data set is included in Appendix A.

The Data

Both the data sets have the same variables and for the purpose of analysis we are combining them into a single data frame and have added a variable color.

The new data set has *17 variables* and a total of *6497 observations* with *no missing data*. It contains:

- fixed.acidity is a variable of type num and has the range between 3.8 to 15.9
- volatile.acidity is a variable of type num and has ranges between 0.080 to 1.580
- citric.acid is a variable of type num and has ranges between 0.0 to 1.66
- residual.sugar is a variable of type num and has ranges between 0.60 to 65.80
- chlorides is a variable of type num and has ranges between 0.009 to 0.611
- free.sulfur.dioxide is a variable of type num and has ranges between 1 to 289.0
- total.sulfur.dioxide is a variable of type num and has ranges between 6 to 440.0
- density is a variable of type num and has ranges between 0.987110 to 1.03898
- pH is a variable of type num and has ranges between 2.72 to 4.01
- sulphates is a variable of type num and has ranges between 0.22 to 2.00
- alcohol is a variable of type num and has ranges between 8 to 14.9000
- quality is a variable of type int and has values between 1 to 10
- color is a factor variable with two levels "red" and "white"

Planning the analysis

As we are trying to find the color of the wine based on different physiochemical feature variables, Figure 1 shows the correlation among those variables. We can see that alcohol and density is highly correlated each other, **but can we distinguish Red vs White Wine by sight from this relationship?**

Yes! If we will just ignore the few outlier points, we can definitely see there is clear separation between white and red wine in Figure 2. It looks like we can build a model to distingush wine color based on these features.

Building Logistic Regression Model

As we do not have any prior hypothesis to select particular factors resulting in the color of wine, let's start with all variables as our predictors and then use the backward stepwise to select the variables. As color is a categorical variable, we use the **logistic regression** and **red is the first level (baseline category) of variable color**.

The output of the **Model 1** has shown that the coefficient of pH variable is the least significant as p value is 0.1689 as in Appendix A. AIC value of our initial model is 450.23. Hence we remove pH variable from our analysis and proceed with Model 2 where we retain all other variables as predictor variables **except pH**.

Figure 1: The Correlation of all variables except Color

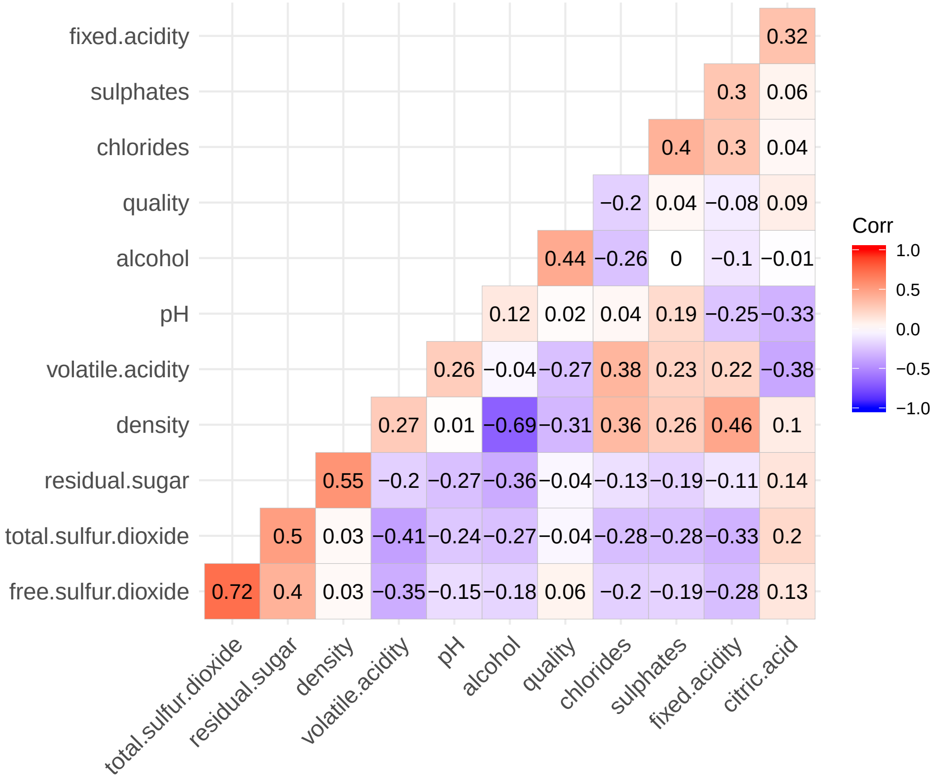
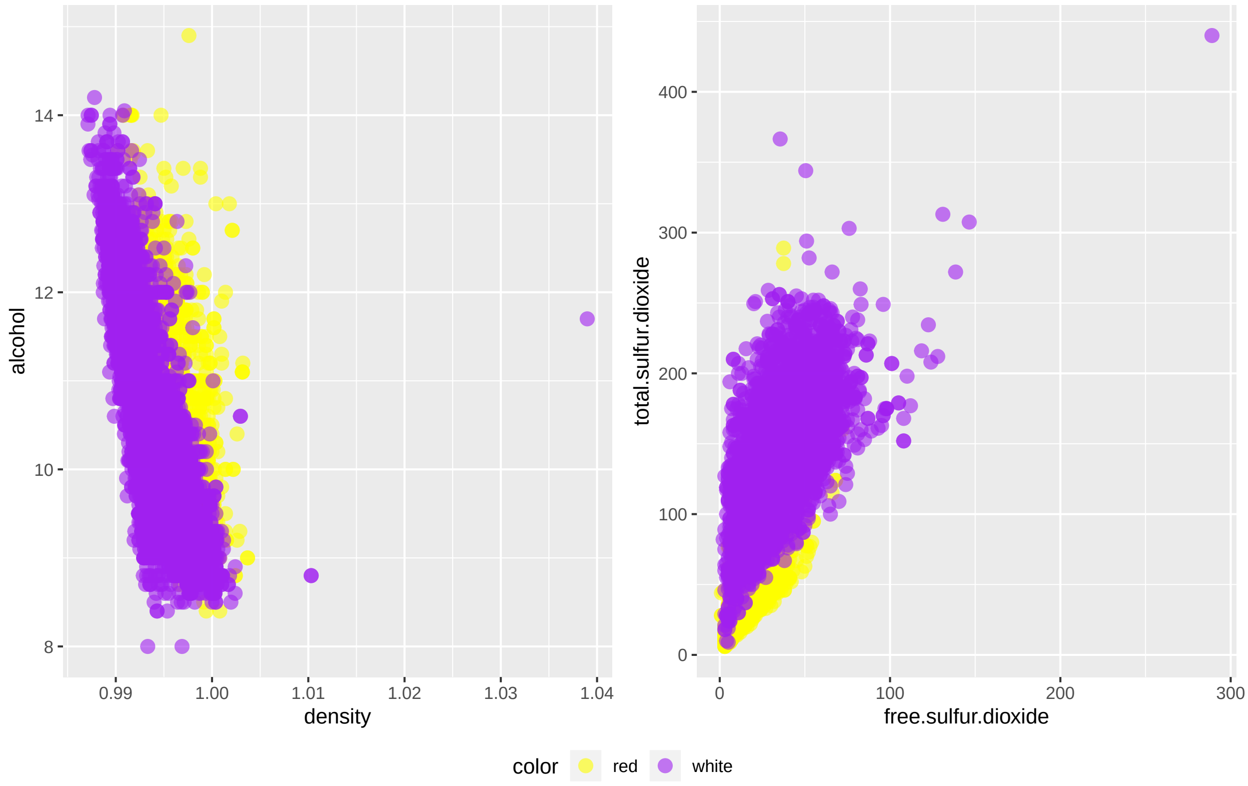


Figure 2: Alcohol vs Density

Free SD vs Total SD



The output of **Model 2** shows that the AIC is lower than Model 1 (450.23 to 450.15), indicating that model

2 is the better model (see Appendix A). We also observe that **fixed.acidity** variable is not significant with a pvalue of 0.2985, hence we have removed this variable and proceeded with Model 3.

Model 3 has all variables as predictor variables **excepts fixed.acidity and pH variables** (see Appendix A). Intercepts of all the variables look statistically significant at all the levels. **AIC of Model 3 is 449.23** which is lower than Model 1 and Model 2.

There are two important questions to ask after generating a model [1].

- **Does the model fit the observed data well?**: we check the model outliers and influential points.
- **Can the model generalize to other samples?**: we check the regression model assumptions - linearity, multicollinearity and independence of errors.

Outliers and influential points

Now, as we come up with a model where all the predictors are significant individually, we want verify if the model fits our observed data well by looking at the possible outliers and influential points if any.

We found 20 residuals are above or below 1.96 standard deviations. As this represents **0.3% of the observations**, expected if the residuals are normal, we do not consider any of these observations as outliers and continued with all 6497 observations included in the model [2].

Next, we have checked for any influential cases by calculating Cook's distance.

Our output shows that there is **one value that is greater than 1** with value 2.128554 (see Appendix A). We consider it may be cause for concern and decide to remove that influential point. Computing the logistics model again without one influential point. The output in Appenddix shows that **Model 4** has AIC value of *390.68* which is lower comparing with Model 3 (449.23). Therefore, we can conclude Model 4 is a better fit to the data. Next, we interpret our logistic regression using the confidence interval of odds ratio.

Interpreting the Model

To interpret model 4, we use the *odds ratio* calculated for each predictor (see Appendix A). The output shows that the intercept is between (inf) to (inf), which overlaps one. This means there is not a significant difference between the odds of white wine and red wine in general, at the 5% level of significance.

The coefficients of citric.acid, residual.sugar, total.sulfur.dioxide are greater than 1 and hence we can say that the odds of wine being white increases with the increase in the values of these predictor variables at the 5% level of significance.

The coefficients of predictor variables volatile.acidity, chlorides, free.sulphur.dioxide, chlorides, sulphates, alcohol, quality are all less than 1 and hence we can say that the odds of wine being white decreases with the increase in the values of these predictor variables at the 5% level of significance.

Next, we want to check if Model 4 can generalize to other samples.

Checking Assumptions of the Model

We need to check and make sure that all the assumptions of logistic regression are followed.

- **Multi-collinearity**: The vif value of all the predictor variables are less than 10 (see Appendix A) and hence we can say that there is no strong correlation among the variables and the assumption of multi-collinearity is not violated.
- **Linearity**: We need to check the linear relationship between continuous predictors and the logit of the outcome variable (see Appendix A). Our output shows that all the interaction variables are significant at the 0.05% confidence level and hence we proceed that the linearity assumption is satisfied as well.

- **Independence errors**: We use Durbin–Watson test for correlations between errors. Our D-W Statistics value is 1.592426 which is closer to 2 which means the assumption is likely met (see Appendix A).

Cross Validation of Model

Now once we have finalised our model and have checked already that it does fit well with our obseved data, next thing we may want to to check how this model will act as generalized - whether it will fit to other samples as well by doing cross validation.

We use the 'caret' package here to perform the 10 fold cross validation which automatically partition our sample dataset into train and test data (see Appendix A).

The cross validation output shows that 1582 observations have been correctly predicted as red wine where 17 cases have been incorrectly predicted as white wine where they are actually Red wine. Also 4885 cases have been correctly predicted as whitewine where they are actually white and only 12 cases are incorrectly predicted as Red where in the sampled data they were actually White. Therefore, in total **6467 cases have been correctly predicted** and remaining 29 cases were predicted incorrectly. So our model actually performs well overall with accuracy level **99%**.

Conclusion

We have tried to come up with a model to predict wine color (white/red) from two separate datasets of Red and White Wine data consisting variable features. At the beginning we started with all the variables as our predictors and then gradually remove variables stepwise based on their significant level and moving to a better model based on AIC value. We have come up with Model 3 where all the predictors are significant. Then to verify that the how well our model fits in the observed dataset,we check for outliers or influential cases. When checking the influential point, we removed one value and generate Model 4 which has better fit. We then question if Model 4 can generalize for other samples by checking its linearity, multicoliniearity, and independence of errors. Finally, in the purpose of verifying how well our model can be generalized as a whole, we have performed crossverification and obtained the confusion matrix which represents that our model is quite well with accuracy level as 99%.

Model limitations

When we first observed the individual dataset Red Wine Dataset contains 1599 observations which is smaller in compared to the White Wine Datasets which has 4898 observations. When we merged the two datasets, this causes imbalanced between data size for each wine types. This can attribute to our high accuracy value in confusion matrix.

As part of our future work, we can apply different sampling methods like oversampling , undersampling or both as will be applicable to our dataset and adjusted our models accordingly.

Reference

[1] Field, A., Miles, J., & Field, Z. (2012). Discovering statistics using R. Sage publications.

[2] Schneider, O. (2020). Seminar 10: Outliers and assumptions, lecture notes, Statistical Methods for Data Analytics MSCI 718, University of Waterloo, delivered in Mar 2020.