Ratthamnoon Prakitpong, 63205165
Tutorial 5, MECH306

Looking at the Corrplot, I've selected 5 predictors with the highest correlation score to var12 (quality).

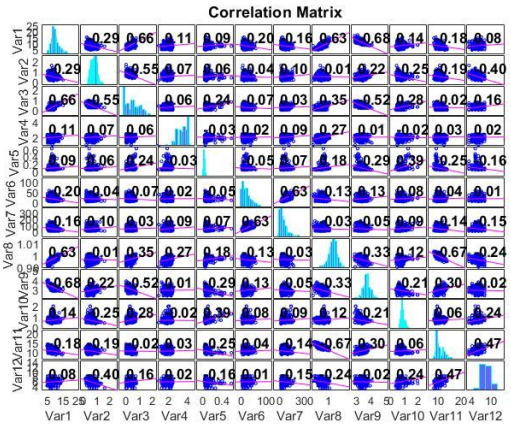

Figure 1: Correlation Matrix

The 5 predictors with highest correlation are v11, v2, v10, v8, v3, so I've trained the following models using there predictors:

Model #	Predictors Selected	Model Type	Validation Option	Accuracy (%)	Comment
1	V2, V3, V8, V10, V11	Fine Tree	Cross Validation, 10 folds	58.5	
2	V2, V3, V8, V10, V11	Linear Discriminant	Cross Validation, 10 folds	58.8	
3	V2, V3, V8, V10, V11	Fine Gaussian SVM	Cross Validation, 10 folds	66.4	Comparatively high score
4	V2, V3, V8, V10, V11	Cubic SVM	Cross Validation, 10 folds	60.7	
5	V2, V3, V8, V10, V11	Fine KNN	Cross Validation, 10 folds	66.9	Comparatively high score
6	V2, V3, V8, V10, V11	Fine Tree	Hold Out, 20%	58.0	
7	V2, V3, V8, V10, V11	Linear Discriminant	Hold Out, 20%	58.0	
8	V2, V3, V8, V10, V11	Fine Gaussian SVM	Hold Out, 20%	69.0	Comparatively high score
9	V2, V3, V8, V10, V11	Cubic SVM	Hold Out, 20%	58.5	

10	V2, V3, V8, V10, V11	Fine KNN	Hold Out, 20%	62.5	Comparatively high score

Table 1: Models and their accuracies

From the two validation types, Fine Gaussian and Fine KNN provides the highest prediction accuracy.

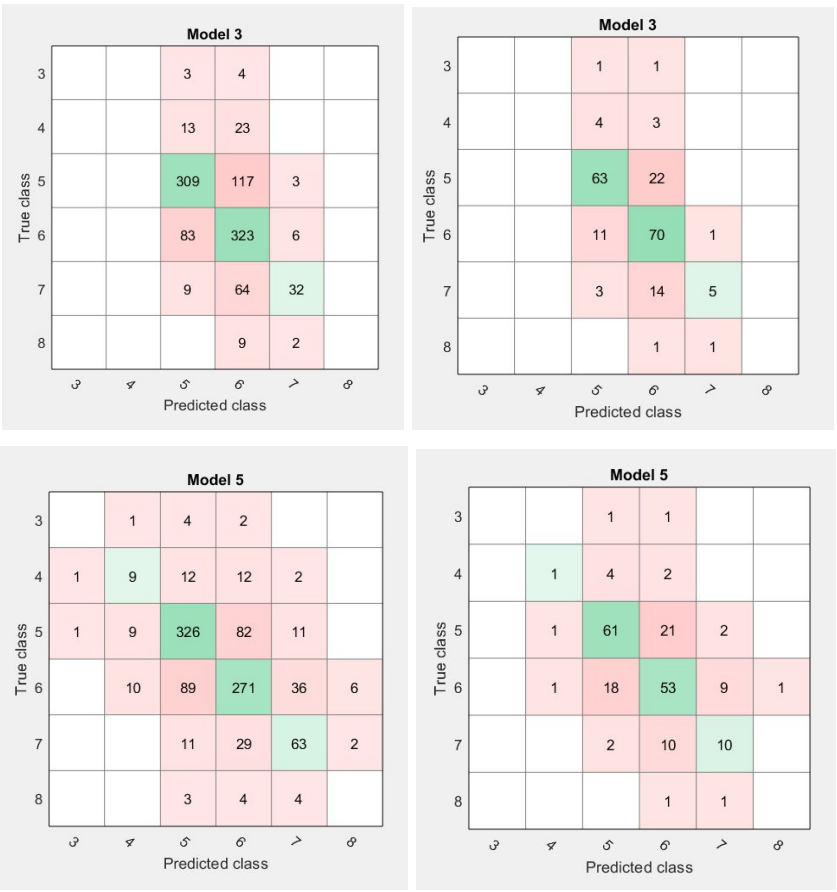

Figure 2.1-2.4: Confusion matrix for models with top accuracy; Fine Gaussian, cross validation (top left); Fine Gaussian, hold out (top right); Fine KNN, cross validation (bottom left); Find KNN, hold out (bottom right)

To select the model, I chose Fine KNN, cross validation. While the accuracy of these four models are quite similar, Fine KNN provides greater range for accurately predicting quality score compared to Fine Gaussian. Note that Fine Gaussian (Figure 2.1, 2.2) did not predict anything outside of 5, 6, and 7. This may be problematic if the dataset comes from where outliers are likelier to appear (ex: good climate year leading to better grape yields, skewing scores higher) and the model would not be able to predict those scores at all. Comparing Fine KNN between cross validation (Figure 2.3) and hold out validation method (Figure 2.4), cross validation provides greater range of correct prediction. Hence, it will be chosen.

This model would work well with the validation data because it could predict a wider range of values. By being able to predict most values well, rather than three specific score very well, I could validate a wider range of data with my model and I am not confined to a dataset with a similar distribution to my training dataset.