Glass Identification

Introduction

Different glasses have different chemical properties. Classifying the type of glass that you have, is an important problem that can help give insights and direction for many fields. This project aims to do this efficiently and accurately for use in forensic investigation. Forensic scientists can use this to classify glass found in a crime scene, compare it with other glass found on suspected individuals and other crime scenes, and use this information to further their investigations. This coursework aims to train and compare and contrast different supervised classification models for use in classifying glass based on their chemical properties. I aim to find out which is the most accurate mode, which is the fastest model, and which is the overall best model for the general use case of glass classification.

Data and Preparation

The models are to be trained and tested on a dataset of 214 entries each with 10 features and of 7 classes. Each entry passed into these models need to have the features that appear in the training dataset. When processing the dataset I first removed the column containing the ID of each glass then split the data into an independent (containing the features) dataset, and dependent (containing the classes) dataset. I then randomised the order of the rows in these two datasets and split them further into training and testing datasets on a 80:20 split. I also created a standardised version of the independent dataset and split them in the same way as for the unstandardised ones.

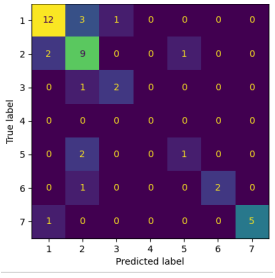
Methodology

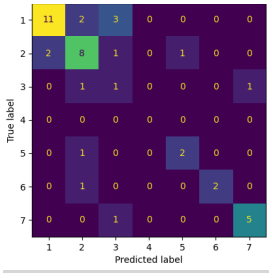
I trained a k nearest neighbour (KNN) model, a classification and regression tree (CART decision tree) model, a naive bayes model, and a support vector machine (SVM) model. I trained the KNN, naive bayes and SVM models on the standardized trainig datasets and the CART model on the unstandardised training dataset. For evaluation, I used each model to make predictions with the testing datasets and recorded their accuracy, training time, prediction time, and confusion matrices.

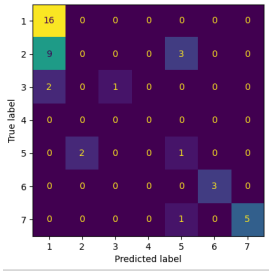
Results

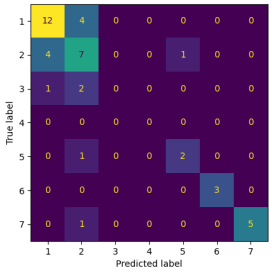
Model	Accuracy	Precision	Recall	F1 Score	Training time	Testing time
KNN	0.721	0.755	0.667	0.699	0.0000205	0.0807
CART	0.674	0.688	0.642	0.654	13.9	0.000944
Naive Bayes	0.605	0.632	0.583	0.567	0.0000758	0.0683
SVM	0.674	0.640	0.639	0.637	3.47	0.00329

Below are the Confusion matrices for KNN, CART, Naive Bayes, and SVM respectively:









CART and SVM have higher training times as they parse the dataset at training time, whereas KNN and Naive Bayes do this every time they make a prediction. CART and SVM have lower prediction/testing times than KNN and Naive Bayes for the same reason. KNN, CART and SVM have higher accuracies than Naive Bayes and SVM though this could be because of overfitting. Naive Bayes's lower accuracy is, likely due to the assumption of feature independence which is not true in this case. The Precision, Recall, and F1 score of all models are similar showing that potential bias has more to do with the training data than with the models.

Conclusion

I assert KNN as the most accurate and efficient model for glass classification. It has shown to be the most accurate, least prone to overfitting, and has the fastest training time, and overall training & prediction of the tested algorithms. One limitation, however, with this model is its prediction time. While it is fast enough to be effective for most situations, it is slower than CART and SVM. Further research could be spent on determining the effect of: different distance metrics for KNN such as maximum distance, or Mahalanobis distance; and of dimensionality reduction for these models, to reduce the computational load of them.