**Model Comparison**

1. **Introduction**

**This programming study aims to compare the performance of SVM, Naïve Bayes**

1. **Data Set Description:** Pima Indians dataset

The UC Irvine machine learning data repository hosts a famous collection of data on whether a patient has diabetes originally owned by the National Institute of Diabetes and Digestive and Kidney Diseases and donated by Vincent Sigillito.

Link to data: <http://archive.ics.uci.edu/ml/datasets/Pima+Indians+Diabetes>

This data has a set of attributes of patients, and a categorical variable telling whether the patient is diabetic or not.

1. **Model Description:**

**Model (a) : Implementation of Naïve Bayes classifier with own code**

**Model (b): Implementation of Naïve Bayes classifier by considering value of ‘0’ as missing value**

**Model (c): Using** e caret and klaR packages to build a naive bayes classifier

**To building Naïve Bayes classifer**

**Model (d): Train and evaluate an SVM to classify the data**

1. **Result: Accuracy of the 4 classifiers (a) to (d)**

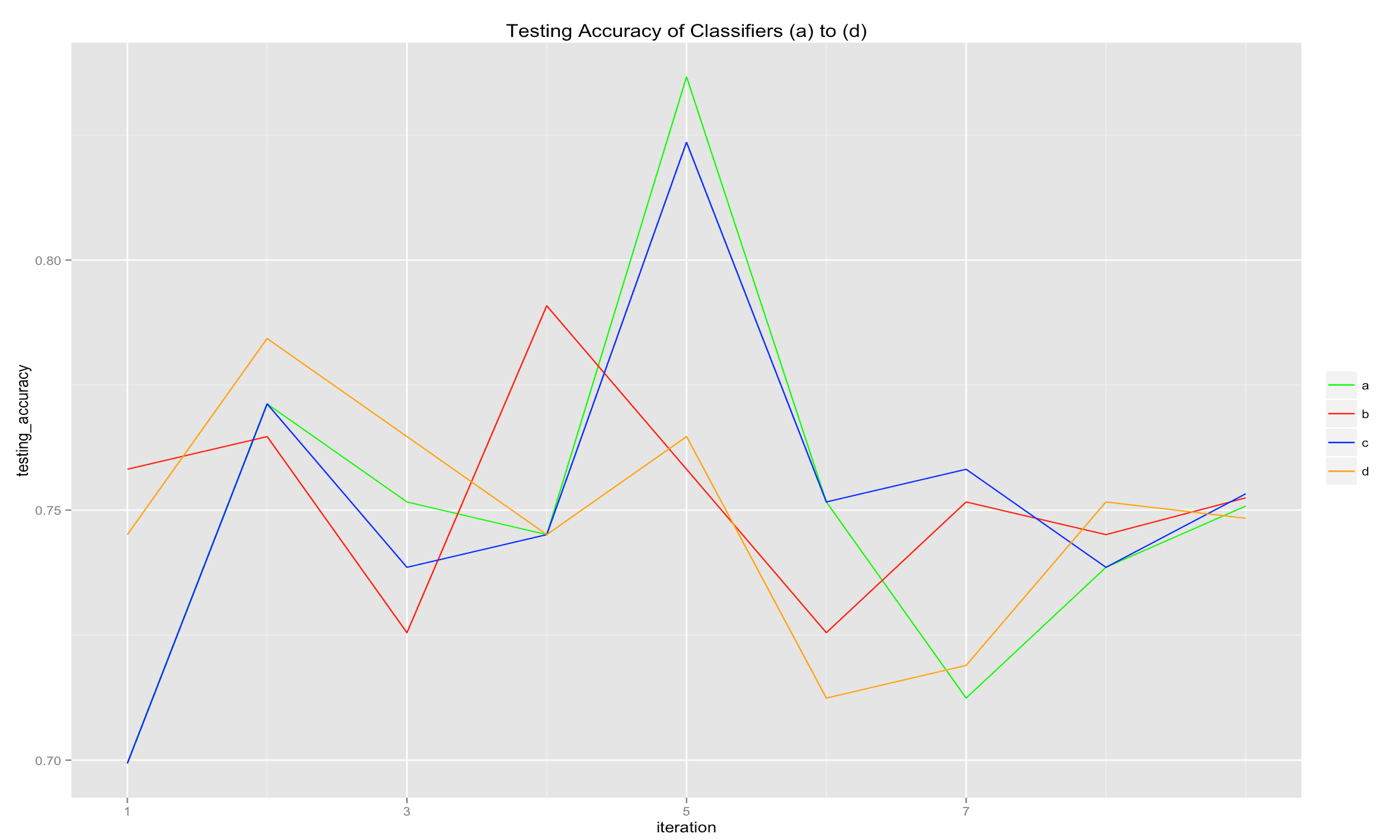
Table 1: Training Accuracy



Table 2: Testing Accuracy



Figure 1: Comparison of Testing Accuracy



1. **Discussion**
2. Performance

In general, the 4 classifiers achieve similar testing and training accuracy. Their minor discrepancies are discussed below:

* *Comparing (a) and (b):*

From Table 1, there is a slight drop of 2% in average accuracy from (a) and (b). However for testing data, there is no trend of whether (a) or (b) is more accurate over the 8 iterations, while the average accuracy is very similar (0.751 vs 0.752).

* + *Comparing (a) and (c)*

From Table 2, (c) achieves the highest accuracy. This could be due to cross-validation performed in (c) which avoids overfitting on the training data.

* *Comparing Naïve Bayes and SVM*

From Graph 1, naïve Bayes seems to work better than SVM. One possible reason could be the assumption of independence of the variables are somewhat satisfied.

1. Issues

For (c), there are warnings of “Numerical 0 probability for all classes with observation (number)”. This could be mainly due to observations with 0s, resulting in the probability being close to 0 in the model.