MSBD 6000B Project 1
Name: Chan Ngae Chau
Student ID: 20411891

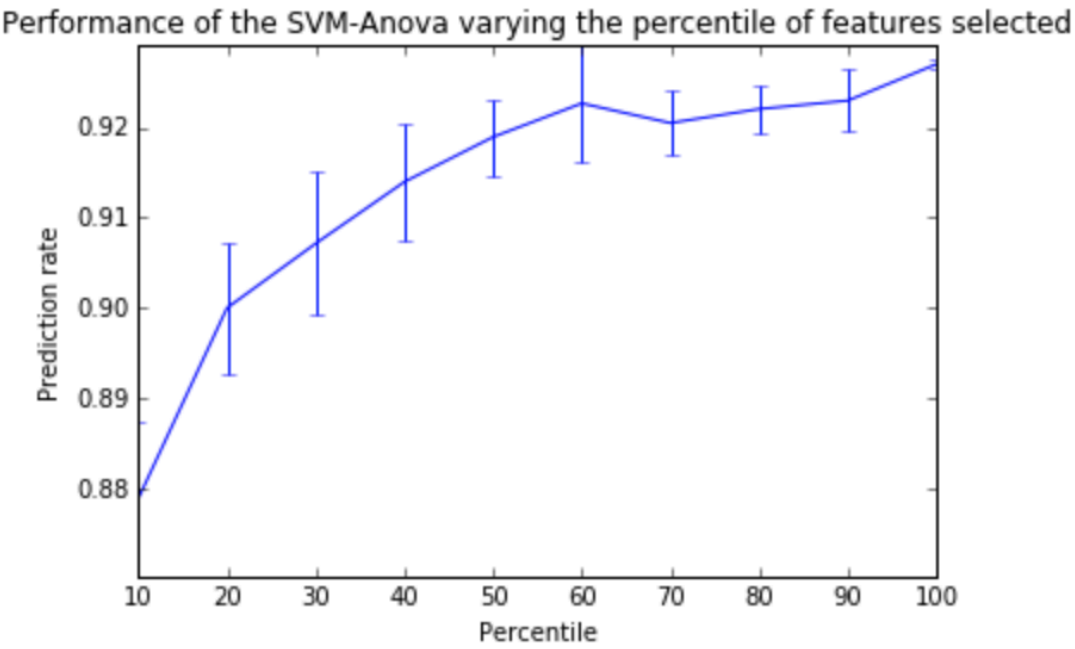
Training Dataset Analysis

The dataset has 3220 rows samples and 57 features. The target has two classes which are being 0 or 1. This is a binary classification problem.

Preprocessing

The data is first standardized by sklearn.preprocessing.StandardScaler().

There is a question to whether to select all 57 features for training. sklearn.feature_selection.SelectPercentile is used to study the impact of the features. The values to be evaluated are: [10, 20, 30, 40, 50, 60, 70, 80, 90, 100]. The result is:



It suggests that all features should be selected as it gets highest precision rate at 100% percentile.

The traindata.csv is splitted into two sets of data. 70% of data is the training set. 30% of data is the validation set.

Model Selection

SVM model of Sklearn.svm.SVC is used. GridSearchCV is used to search for the best parameter for the SVM between different kernels (linear, RBF, polynominal) and between different Cs. The kernel suggested is 'RBF' and C parameter is suggested to be 5.

```
GridSearchCV took 0.74 seconds for 9 candidate parameter settings.
Model with rank: 1
Mean validation score: 0.929 (std: 0.016)
Parameters: {'C': 5}

Model with rank: 2
Mean validation score: 0.928 (std: 0.011)
Parameters: {'C': 7}

Model with rank: 3
Mean validation score: 0.924 (std: 0.008)
Parameters: {'C': 2}
```

The performance of this model is:

```
Train set accuracy: 0.93 (+/- 0.05)
Validation set accuracy: 0.93 (+/- 0.03)
```

Test data prediction

The test data has 1380 rows. The test data is first standardized by the same sklearn.preprocessing.StandardScaler object obtained during training. A SVM model with kernel = RBF and C = 5 is used to predict the target.