**UE Machine Learning: Supervised Techniques**

Exercise 7 Report

Name: Omar Amr

Matrikel-Nr: k11776960

**Sequence Data (One-hot encoded Data):**

Using 10 Cross Validation to train the RBF SVM:

|  |  |  |  |
| --- | --- | --- | --- |
| C Parameter | Sigma Parameter | Cross Validation Error | Cross Validation Run Time |
| 5 | 0.001 | 11.95 % | 10.57044 secs |
| 5 | 0.01 | 15.15 % | 15.93966 secs |
| 5 | 0.1 | 34.05 % | 16.1011 secs |
| 5 | 1 | 34.2 % | 15.9258 secs |
| 5 | 10 | 34.05 % | 15.4783 secs |
| 5 | 100 | 34.15 % | 15.2985 secs |
|  |  |  |  |
| 10 | 0.001 | 11.7 % | 10.109 secs |
| 10 | 0.01 | 14.7 % | 15.6695 secs |
| 10 | 0.1 | 34.4 % | 15.8919 secs |
| 10 | 1 | 34.15 % | 16.1955 secs |
| 10 | 10 | 34.65 % | 15.4999 secs |
| 10 | 100 | 34.55 % | 15.7556 secs |
|  |  |  |  |
| 50 | 0.001 | 11.85 % | 10.3031 secs |
| 50 | 0.01 | 15.85 % | 15.3035 secs |
| 50 | 0.1 | 34 % | 17.9066 secs |
| 50 | 1 | 34.25 % | 16.8074 secs |
| 50 | 10 | 34.65 % | 15.7029 secs |
| 50 | 100 | 34.1 % | 15.7569 secs |
|  |  |  |  |
| 100 | 0.001 | 11.15 % | 9.8851 secs |
| 100 | 0.01 | 15.35 % | 15.6489 secs |
| 100 | 0.1 | 34.05 % | 15.6377 secs |
| 100 | 1 | 34.65 % | 15.8796 secs |
| 100 | 10 | 34.65 % | 15.3996 secs |
| 100 | 100 | 34.45 % | 15.1953 secs |
| 200 | 0.001 | 12.3 % | 10.097 secs |
| 200 | 0.01 | 15.75 % | 16.0958 secs |
| 200 | 0.1 | 34.2 % | 15.9296 secs |
| 200 | 1 | 33.85 % | 16.3399 secs |
| 200 | 10 | 34.15 % | 15.9439 secs |
| 200 | 100 | 34.25 % | 15.7323 secs |

Using 10 Cross Validation to train the random forest and comparing results to out-of- bag estimate:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Cross Validation | No. of trees | OOB Error Estimate | Cross Validation Error | Cross Validation Run Time |
| 10 | 10 | 19.32 % | 14.6 % | 3.764281 secs |
| 10 | 100 | 12.05 % | 1.1 % | 34.61006 secs |
| 10 | 1000 | 11 % | 0 % | 5.739604 mins |
| 10 | 10000 | 10.9 % | 0 % | 57.26711 mins |

The best model trained by RBF SVM has a Cross-Validation error = 11.15% with run time equivalent to 9.8 seconds. On the other hand, the best random forest model has a cross validation error of 0% and OOB of 10.9%, however the runtime is considerably worse as it takes the model 57.26711 minutes to finish training.

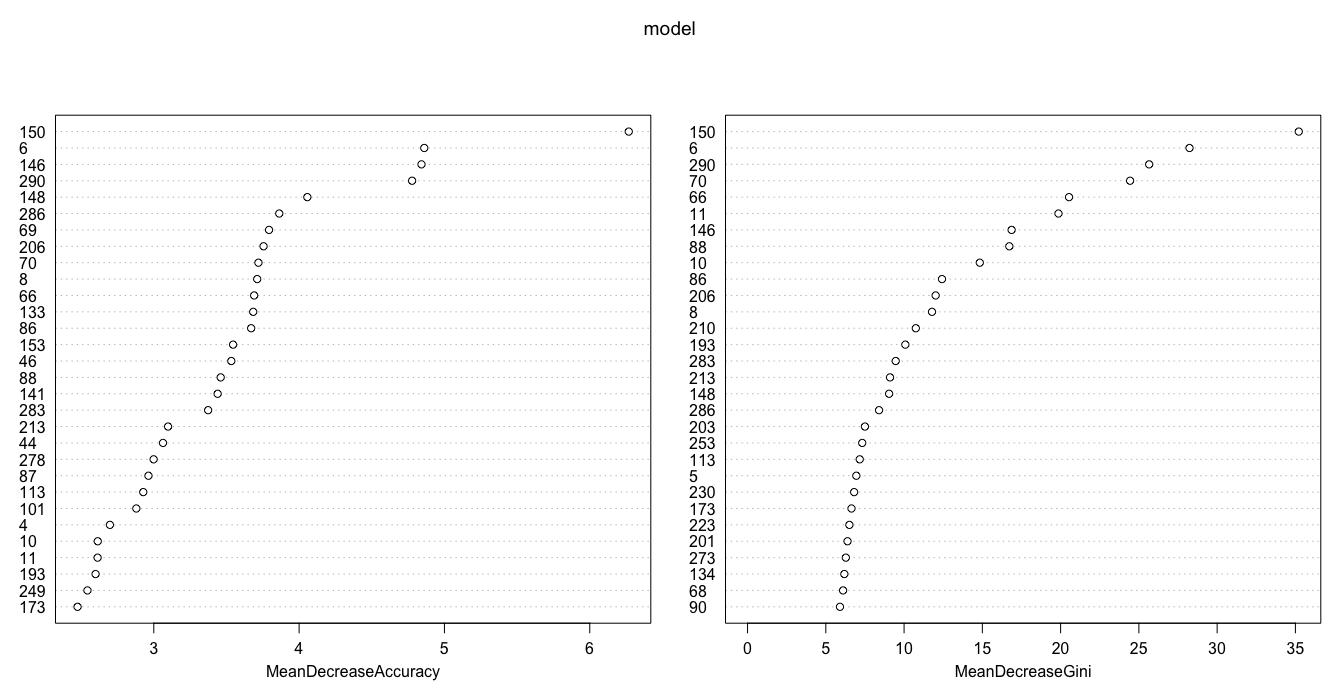
Training Time overview:

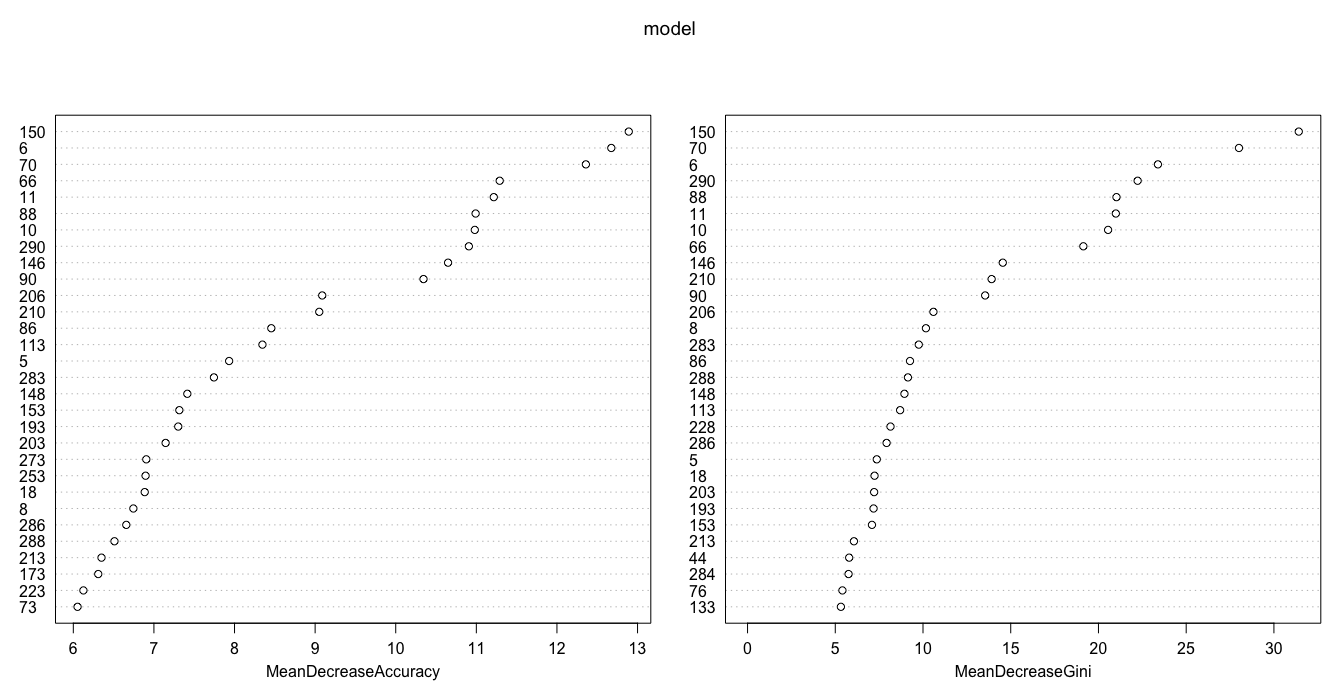
The runtime of random forest is comparable to those of the SVM as far as the number of trees is small. However, when the number of trees increases, the run time increases too.

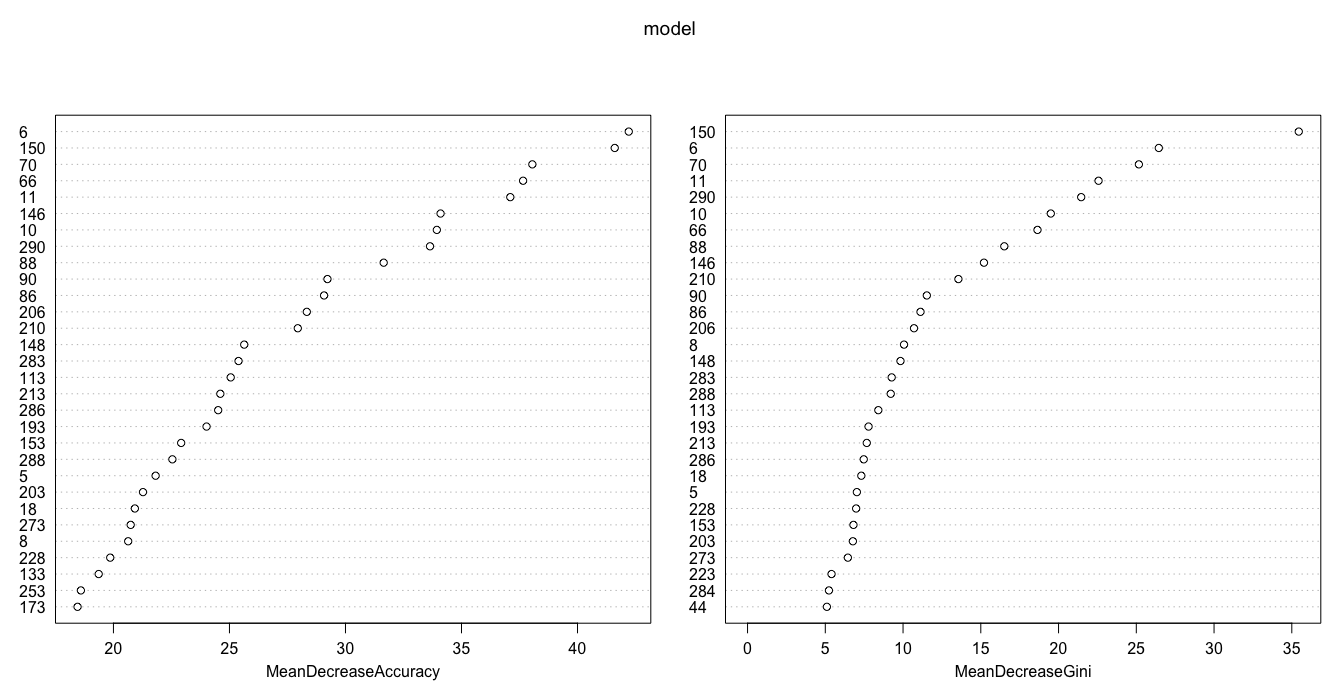
Parameter selection overview:

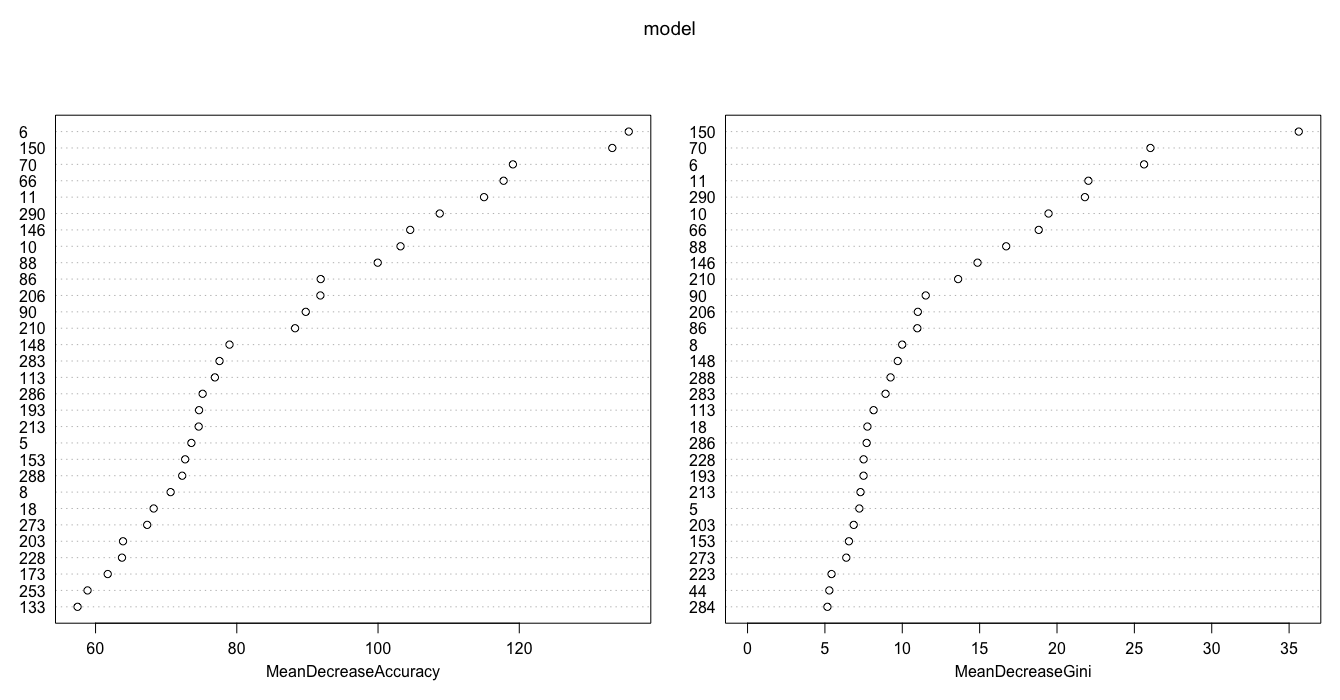
RBF SVM depends on 2 parameters (c and sigma), that is why there is more combinations for models. The search strategy followed here is grid search where each value of C is tested with every value for Sigma. Random forests depend mainly on one parameter which is the number of trees. Which mean there is less combinations for the model compared to RBF SVM. The random forest is trained with each value for the number of trees.

Sequence Dataset Variable Importance for Random Forests:

The following figure shows the variable importance for a random tree with 10 trees:

The following figure shows the variable importance for a random tree with 100 trees:

The following figure shows the variable importance for a random tree with 1000 trees:

The following figure shows the variable importance for a random tree with 10000 trees:

From the last 4 figures, one can see that some of the dataset variables has noticeable effect on the accuracy of the tree classification (e.g. variables 150, 71, 11, 290, 60, 88) seem to play a vital role in the random forest classification process. The insight that can be deduced from this observation is that the existence of a certain acid composition in a specific position is very important in the process of defining the type of the acid.

Preferred Method for Acid Sequences Dataset:

I would prefer using random forest on this dataset rather that SVMs for the following reasons:

1. The best SVM has cross-validation error of 11.15 %.
2. Random forests with 1000 trees (cross error = 0 %, OOB Error = 11 %) or 10000 trees (cross error = 0 %, OOB Error = 10.9 %) perform better than the best SVM in terms of cross-validation error and OOB Estimate Error too.
3. The runtime for random forests with 1000 or 10000 trees is considerably more that SVM, however accuracy is better.

**Breast Cancer Dataset:**

Using 10 Cross Validation to train the RBF SVM:

|  |  |  |  |
| --- | --- | --- | --- |
| C Parameter | Sigma Parameter | Cross Validation Error | Cross Validation Run Time |
| 5 | 0.001 | 2.92 % | 0.0616 secs |
| 5 | 0.01 | 3.22 % | 0.0649 secs |
| 5 | 0.1 | 4.1 % | 0.605 secs |
| 5 | 1 | 4.24 % | 0.1217 secs |
| 5 | 10 | 12.75 % | 0.174 secs |
| 5 | 100 | 25.02 % | 0. 1712 secs |
|  |  |  |  |
| 10 | 0.001 | 3.07 % | 0.0523 secs |
| 10 | 0.01 | 2.92 % | 0.05 secs |
| 10 | 0.1 | 4.39 % | 0.0697 secs |
| 10 | 1 | 3.8 % | 0.1258 secs |
| 10 | 10 | 12.6 % | 0.1746 secs |
| 10 | 100 | 25.17 % | 0.1616 secs |
|  |  |  |  |
| 50 | 0.001 | 3.07 % | 0.0485 secs |
| 50 | 0.01 | 3.07 % | 0.0678 secs |
| 50 | 0.1 | 4.53 % | 0.0654 secs |
| 50 | 1 | 4.39 % | 0.1081 secs |
| 50 | 10 | 12.57 % | 0.1773 secs |
| 50 | 100 | 24.87 % | 0.1727 secs |
|  |  |  |  |
| 100 | 0.001 | 3.51 % | 0.0501 secs |
| 100 | 0.01 | 3.22 % | 0.065 secs |
| 100 | 0.1 | 4.54 % | 0.0736 secs |
| 100 | 1 | 3.95 % | 0.1219 secs |
| 100 | 10 | 13.16 % | 0.1745 secs |
| 100 | 100 | 25.48 % | 0.1709 secs |
| 200 | 0.001 | 3.36 % | 0.0579 secs |
| 200 | 0.01 | 3.81 % | 0.0794 secs |
| 200 | 0.1 | 4.24 % | 0.0661 secs |
| 200 | 1 | 3.5 % | 0.1148 secs |
| 200 | 10 | 13.04 % | 0.169 secs |
| 200 | 100 | 25.34 % | 0.1646secs |

Using 10 Cross Validation to train the random forest and comparing results to out-of- bag estimate:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Cross Validation | No. of trees | OOB Error Estimate | Cross Validation Error | Cross Validation Run Time |
| 10 | 10 | 4.88 % | 4.55 % | 0.0547 secs |
| 10 | 100 | 3.22 % | 0.294 % | 0.3429 secs |
| 10 | 1000 | 2.64 % | 0 % | 2.6231 secs |
| 10 | 10000 | 2.64 % | 0 % | 25.9456 secs |

The best model trained by RBF SVM has a Cross-Validation error = 2.92% with run time equivalent to 0.05 seconds. On the other hand, the best random forest model has a cross validation error of 0% and OOB of 2.64%, however the runtime is worse as it takes the model 2.6231 seconds to finish training but it is not very bad of course.

Training Time overview:

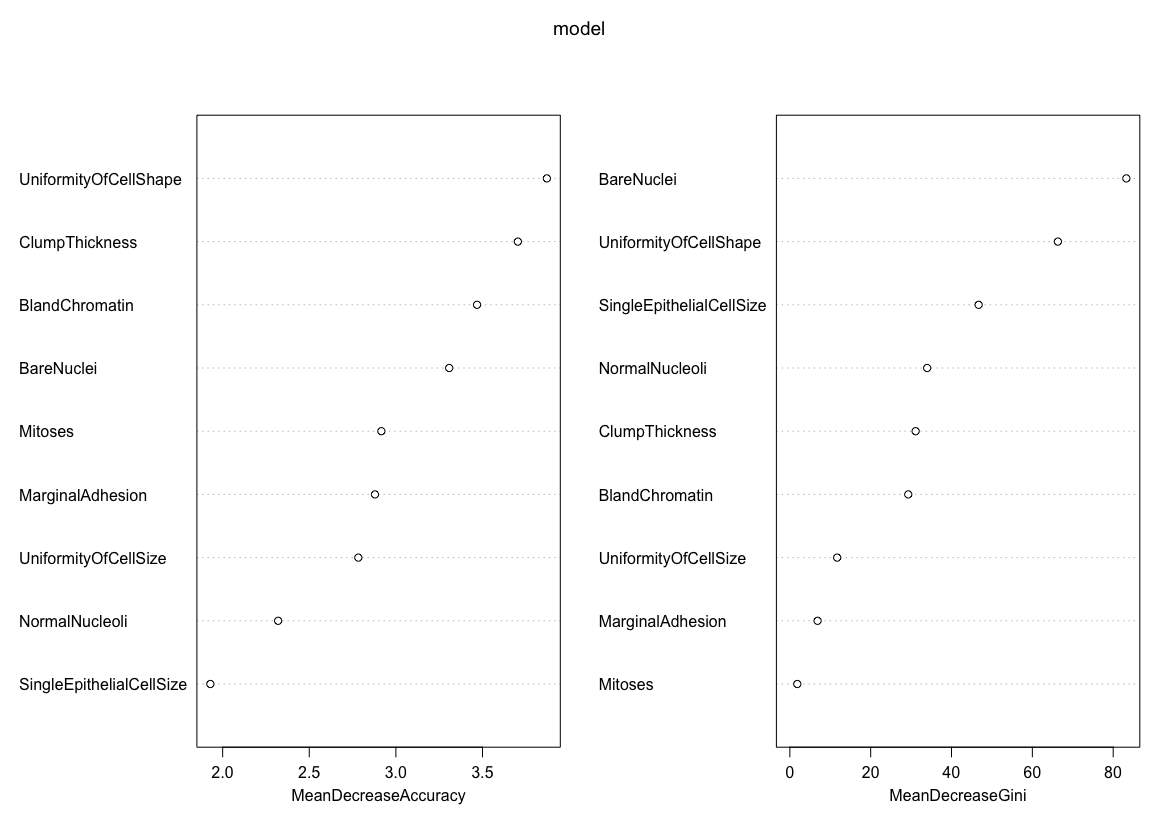
The runtime of random forest is very comparable to those of the SVM even for random forest trees with 1000 or 10,000 trees.

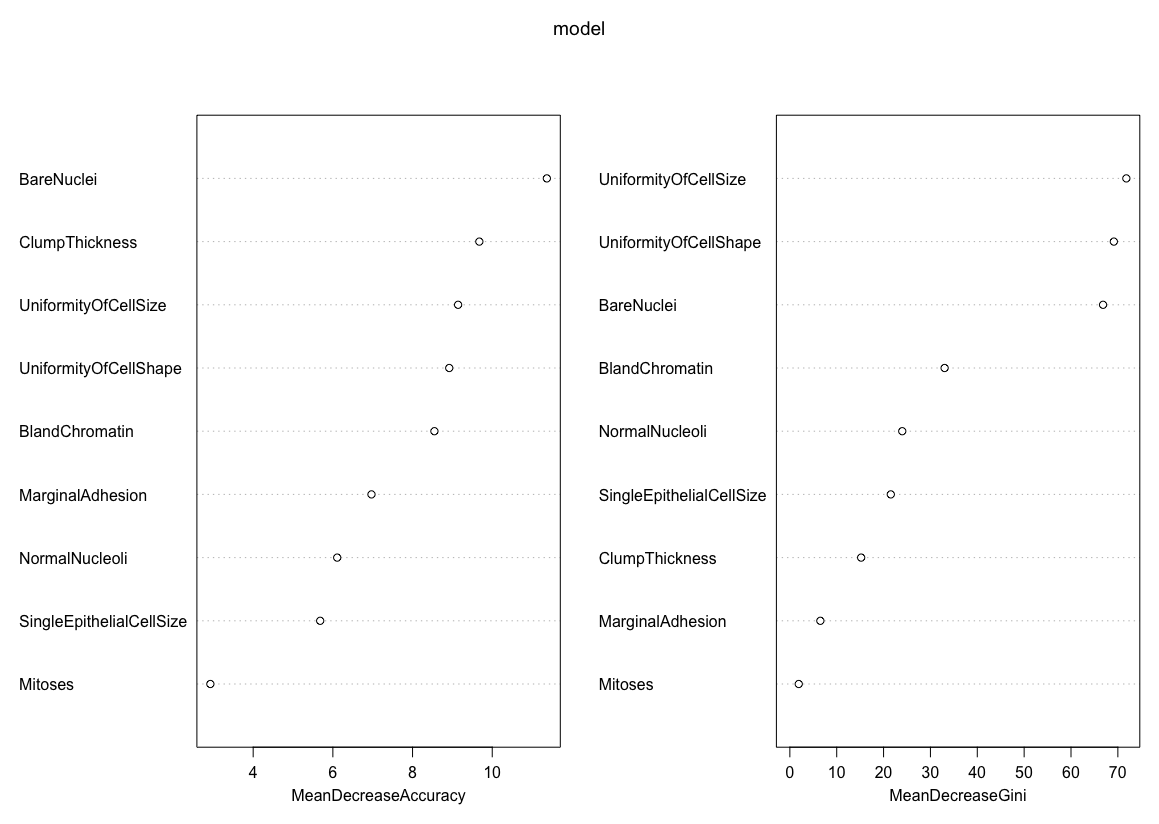
Parameter selection overview:

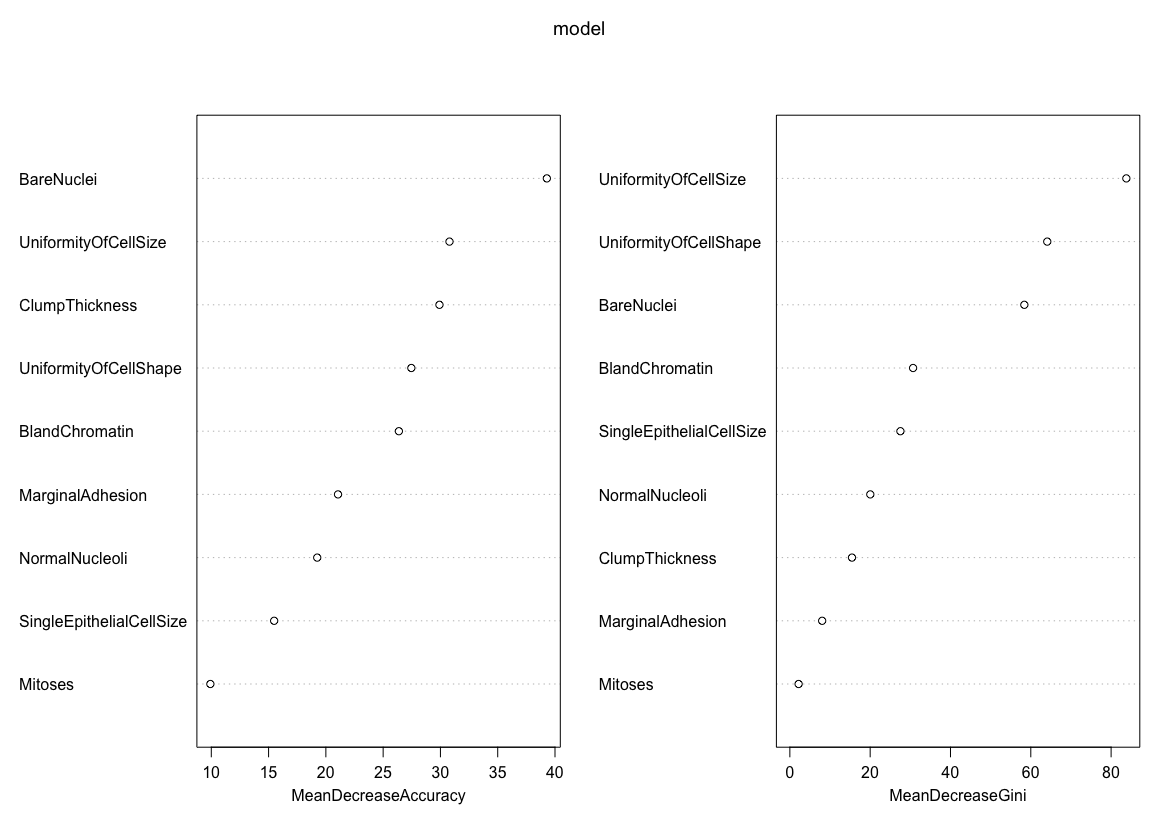
RBF SVM depends on 2 parameters (c and sigma), that is why there is more combinations for models. The search strategy followed here is grid search where each value of C is tested with every value for Sigma. Random forests depend mainly on one parameter which is the number of trees. Which mean there is less combinations for the model compared to RBF SVM. The random forest is trained with each value for the number of trees.

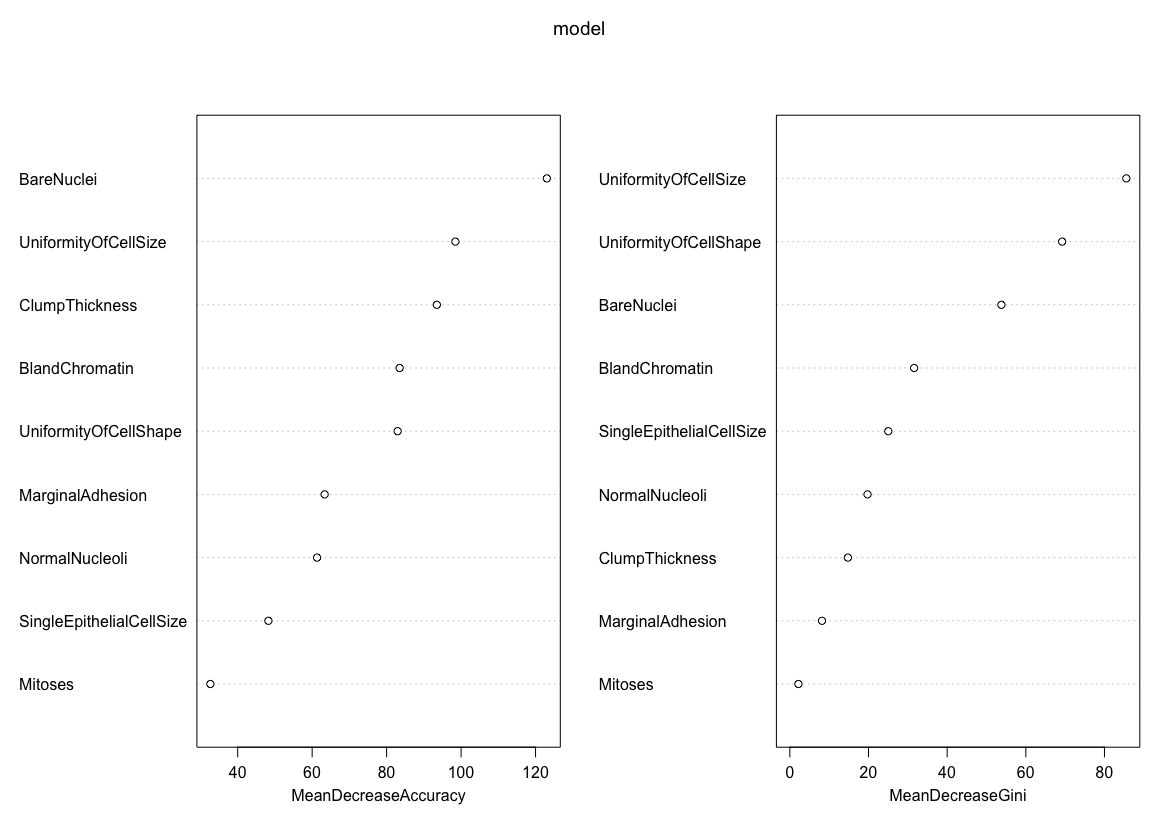
Breast Cancer Dataset Variable Importance for Random Forests:

The following figure shows the variable importance for a random tree with 10 trees:



The following figure shows the variable importance for a random tree with 100 trees:

The following figure shows the variable importance for a random tree with 1000 trees:

The following figure shows the variable importance for a random tree with 10000 trees:

From the last 4 figures, one can see that some of the dataset variables has noticeable effect on the accuracy of the tree classification (e.g. Uniformity of Cell Size & Bare Nuclei & Uniformity of Cell Shape) seem to play a vital role in the random forest classification process. The insight that can be deduced from this observation is that the values of such attribute as the shape and size of a cell are a good indicator whether a patient has breast cancer or not.

Preferred Method for Breast Cancer Dataset:

I would prefer using random forest on this dataset rather that SVMs for the following reasons:

1. The best SVM has cross-validation error of 2.92 %.
2. Random forests with 1000 trees (cross error = 0 %, OOB Error = 2.64 %) or 10000 trees (cross error = 0 %, OOB Error = 2.64 %) perform better than the best SVM in terms of cross-validation error and OOB Estimate Error too.
3. The runtime for random forests with 1000 or 10000 trees in this dataset is very low that it is comparable to that of SVM, however accuracy is better.

**Wine Dataset:**

Using 10 Cross Validation to train the RBF SVM:

|  |  |  |  |
| --- | --- | --- | --- |
| C Parameter | Sigma Parameter | Cross Validation Error | Cross Validation Run Time |
| 5 | 0.001 | 2.22 % | 0.0471 secs |
| 5 | 0.01 | 1.11 % | 0.0402 secs |
| 5 | 0.1 | 1.66 % | 0.0412 secs |
| 5 | 1 | 34.2 % | 0.0749 secs |
| 5 | 10 | 60.09 % | 0.0653 secs |
| 5 | 100 | 60.06 % | 0.0555 secs |
|  |  |  |  |
| 10 | 0.001 | 2.22 % | 0.0433 secs |
| 10 | 0.01 | 2.25 % | 0.0387 secs |
| 10 | 0.1 | 1.11 % | 0.0433 secs |
| 10 | 1 | 34.86 % | 0.0552 secs |
| 10 | 10 | 60.29 % | 0.0612 secs |
| 10 | 100 | 60.29 % | 0.0618 secs |
|  |  |  |  |
| 50 | 0.001 | 2.28 % | 0.038 secs |
| 50 | 0.01 | 3.92 % | 0.0388 secs |
| 50 | 0.1 | 1.69 % | 0.0423 secs |
| 50 | 1 | 35.52 % | 0.0528 secs |
| 50 | 10 | 60.22 % | 0.0588 secs |
| 50 | 100 | 60.06 % | 0.0702 secs |
|  |  |  |  |
| 100 | 0.001 | 3.36 % | 0.0409 secs |
| 100 | 0.01 | 3.36 % | 0.0366 secs |
| 100 | 0.1 | 1.11 % | 0.0468 secs |
| 100 | 1 | 33.26 % | 0.0572 secs |
| 100 | 10 | 60.19 % | 0.0635 secs |
| 100 | 100 | 60.13 % | 0.0586 secs |
| 200 | 0.001 | 2.84 % | 0.0404 secs |
| 200 | 0.01 | 5.13 % | 0.0399 secs |
| 200 | 0.1 | 1.14 % | 0.0419 secs |
| 200 | 1 | 34.93 % | 0.057 secs |
| 200 | 10 | 60.16 % | 0.0584 secs |
| 200 | 100 | 59.9 % | 0.0555 secs |

Using 10 Cross Validation to train the random forest and comparing results to out-of- bag estimate:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Cross Validation | No. of trees | OOB Error Estimate | Cross Validation Error | Cross Validation Run Time |
| 10 | 10 | 3.39 % | 2.94 % | 0.0256 secs |
| 10 | 100 | 2.25 % | 0 % | 0.1434 secs |
| 10 | 1000 | 1.69 % | 0 % | 0.8892 secs |
| 10 | 10000 | 1.69 % | 0 % | 8.4717 secs |

The best model trained by RBF SVM has a Cross-Validation error = 1.11% with run time equivalent to 0.0433 seconds. On the other hand, the best random forest model has a cross validation error of 0% and OOB of 1.69%, however the runtime is a little bit worse as it takes the model 0.8892 seconds to finish training but it is not very bad of course.

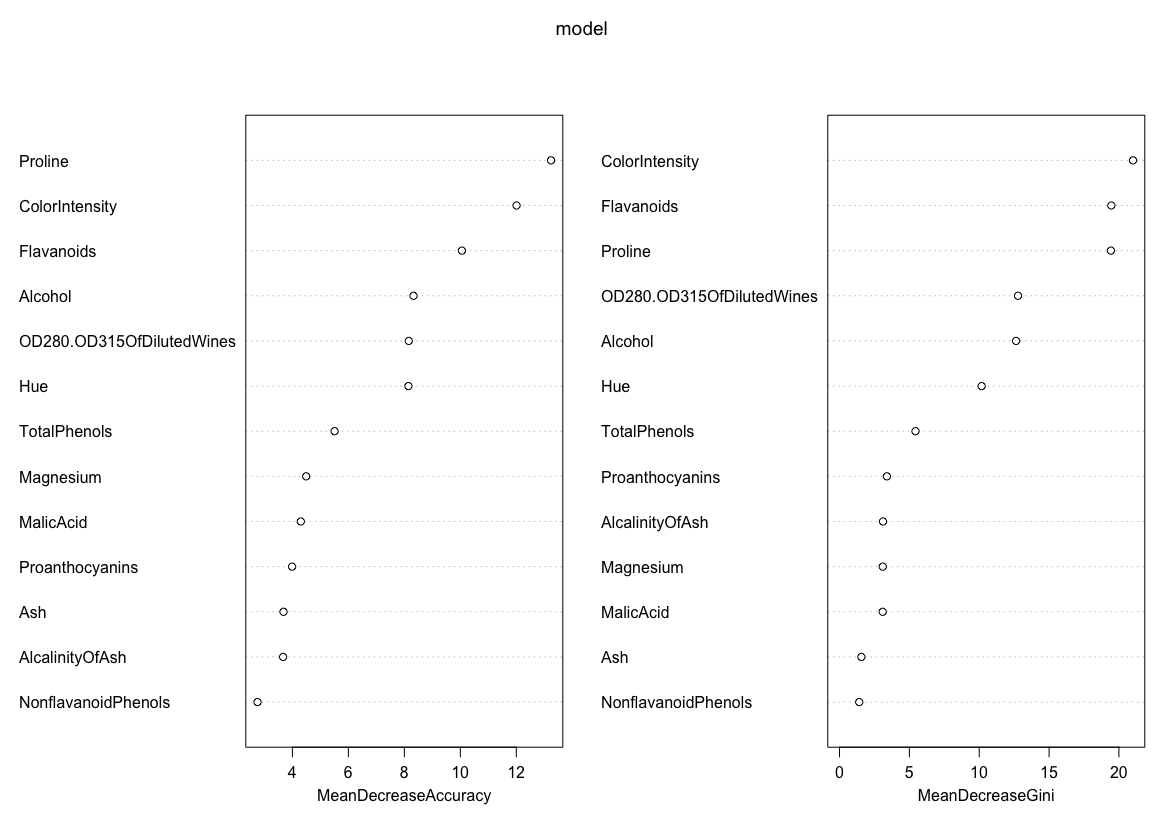
Training Time overview:

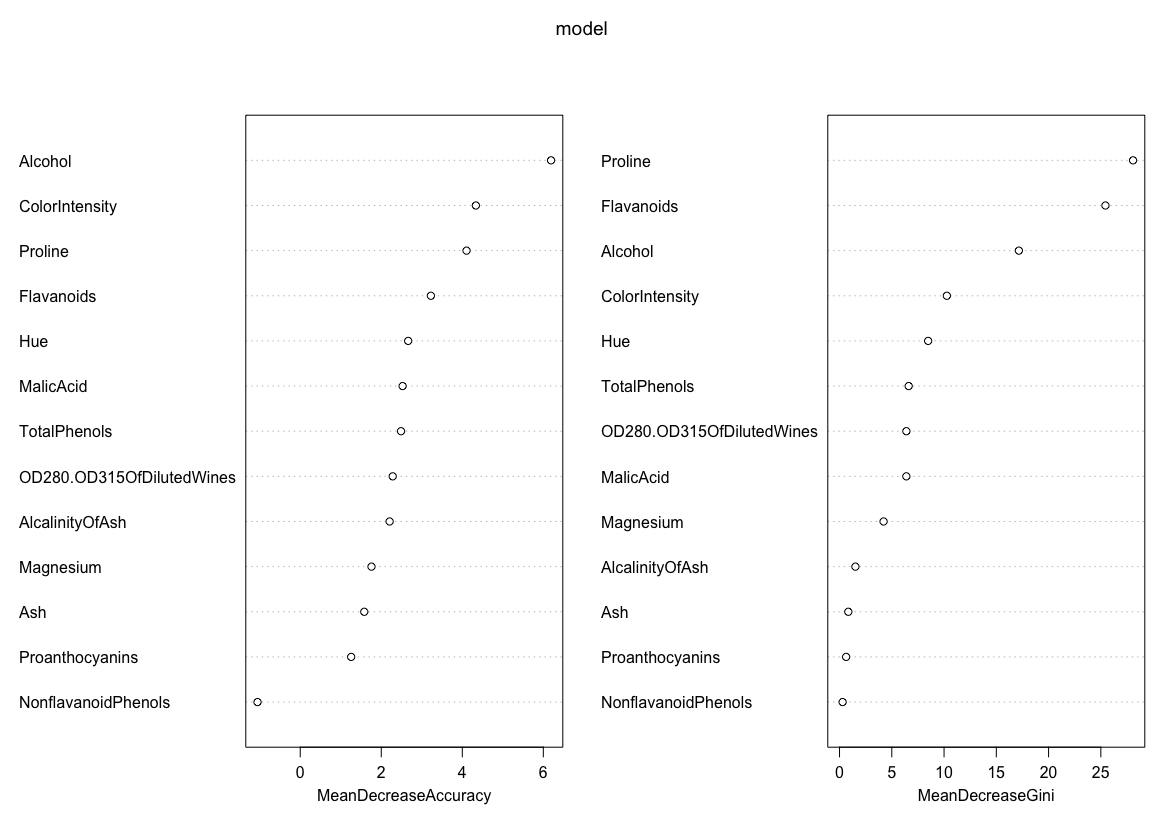
The runtime of random forest is very comparable to those of the SVM even for random forest trees with 1000 or 10,000 trees.

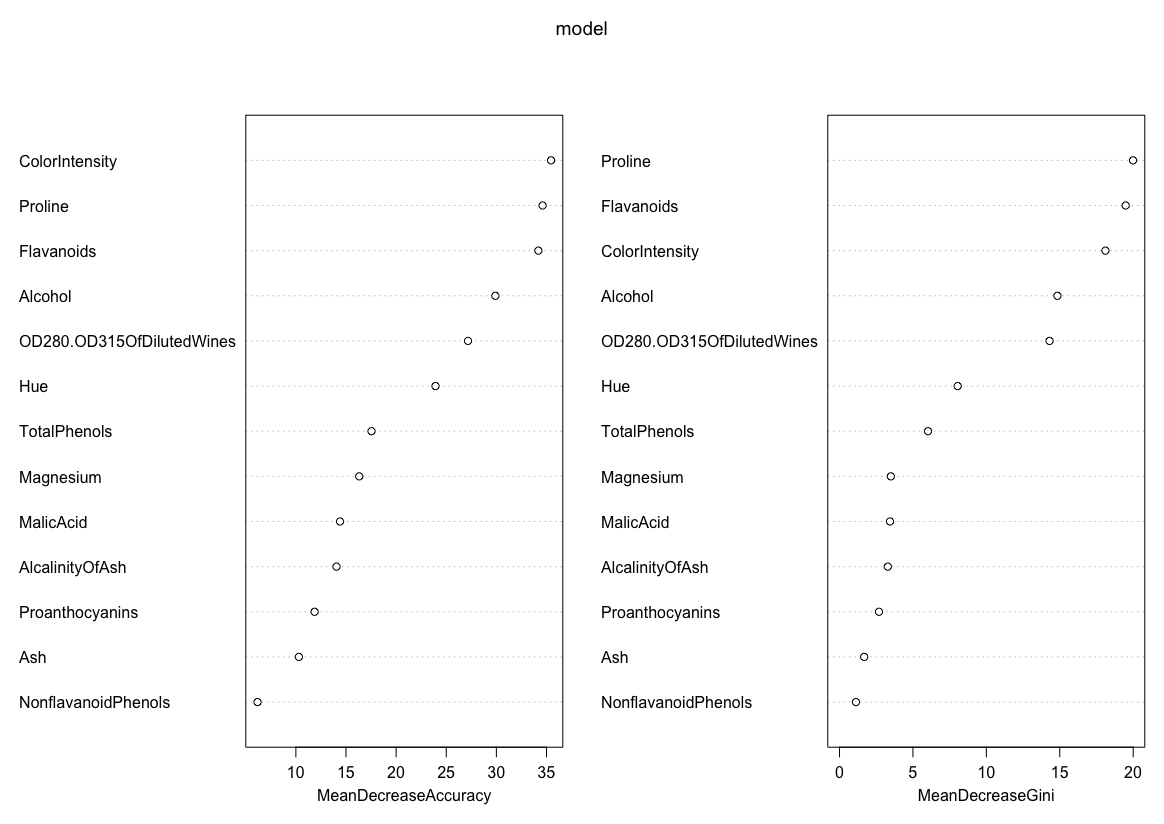
Parameter selection overview:

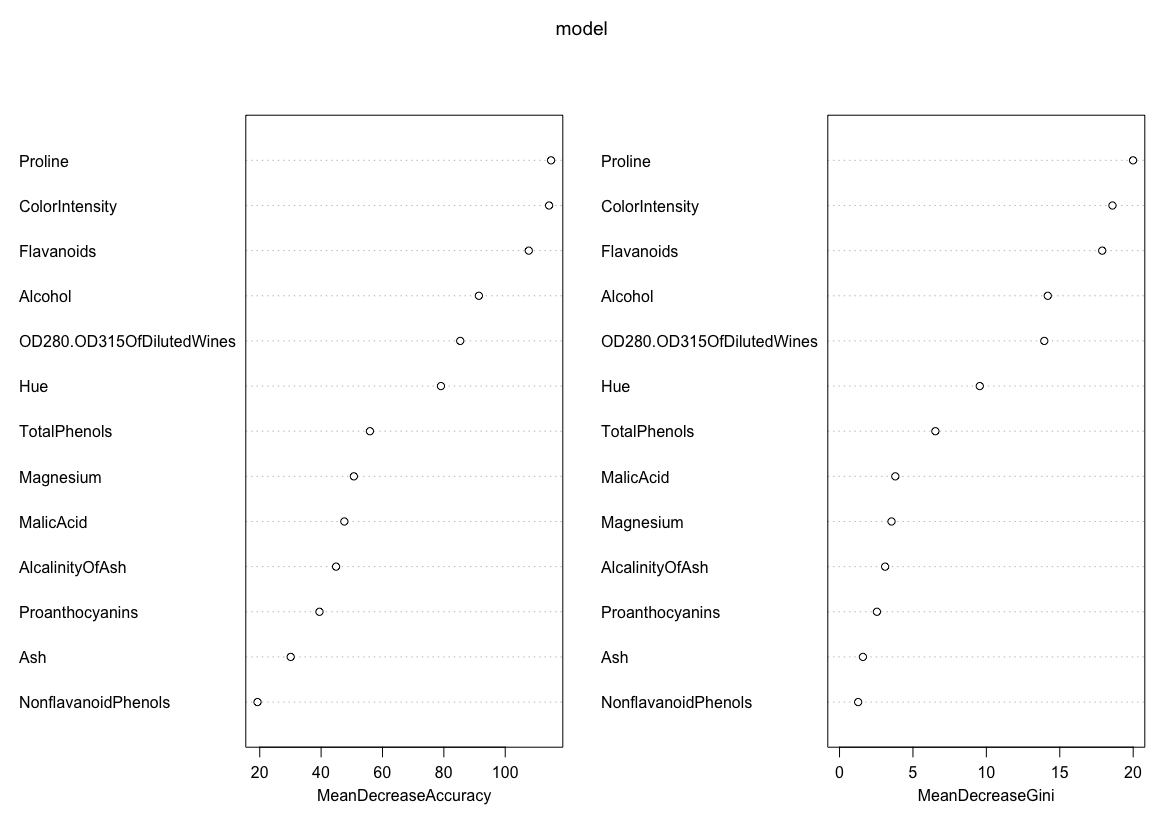
RBF SVM depends on 2 parameters (c and sigma), that is why there is more combinations for models. The search strategy followed here is grid search where each value of C is tested with every value for Sigma. Random forests depend mainly on one parameter which is the number of trees. Which mean there is less combinations for the model compared to RBF SVM. The random forest is trained with each value for the number of trees.

Wine Dataset Variable Importance for Random Forests:

The following figure shows the variable importance for a random tree with 10 trees:

The following figure shows the variable importance for a random tree with 100 trees:

The following figure shows the variable importance for a random tree with 1000 trees:

The following figure shows the variable importance for a random tree with 10000 trees:

From the last 4 figures, one can see that some of the dataset variables has noticeable effect on the accuracy of the tree classification (e.g. Proline & Flavanoids & Alcohol & Color Intensity) seem to play a vital role in the random forest classification process. The insight that can be deduced from this observation is that the values of such attribute as Alcohol and Color Intensity are a good indicator of the class of the wine.

Preferred Method for Wine Dataset:

Deciding which is better for this dataset is tough as both SVM and Random Forest perform very well on this dataset. The best models of each type have very close error values and also the runtime for both is very low. However the best RBF SVM has a very low validation error which might be a plus for choosing SVM.