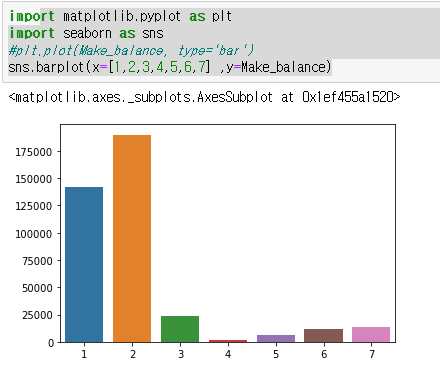
Process:

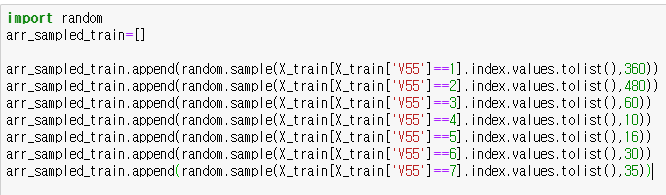
1. **Read data** : I read ‘covertype dataset.
2. **Made the Train and Test data**: I devided 70% of data into train and the 30 % of data into test.
3. **Data imbalance:** Check the ‘Y’ ratio (1 to 7) and make balance for the training data to avoid data imbalance problem.
4. **Find importance features:** I used the best subset selection with the training data. And select the features based on CP and BIC value. And modified the training data.
5. **Made models with the modified training data:** I made Logistic model, LDA, QDA, KNN, TREE
6. **Check accuracy:** I made confusion matrix and check each model’s accuracy
7. **Ensemble model:** I made the ensemble model combine with high accuracy models and show confusion matrix and check the ensemble model’s accuracy

Output

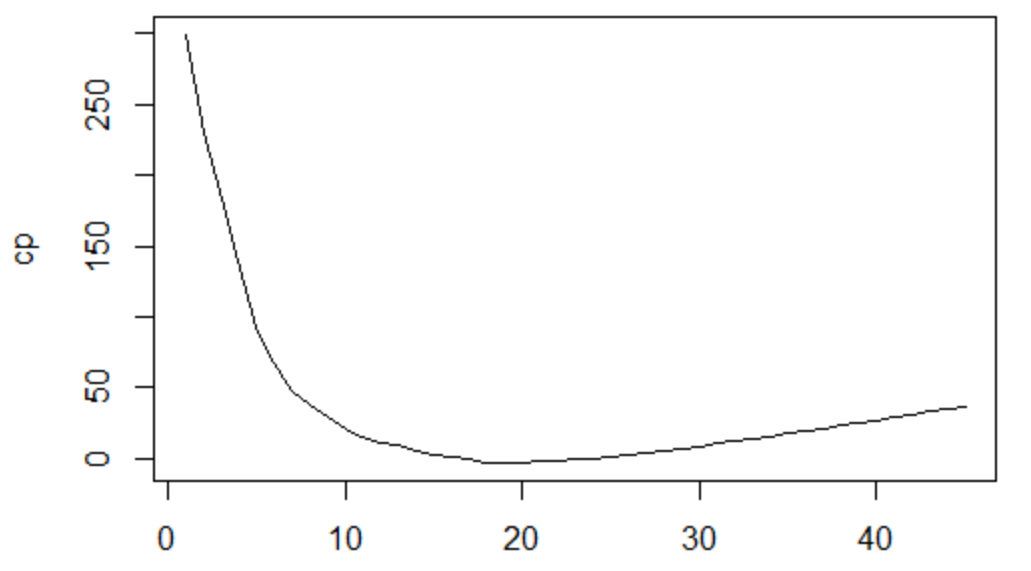
**Chece the class ration for ‘Y’ in the training data.**

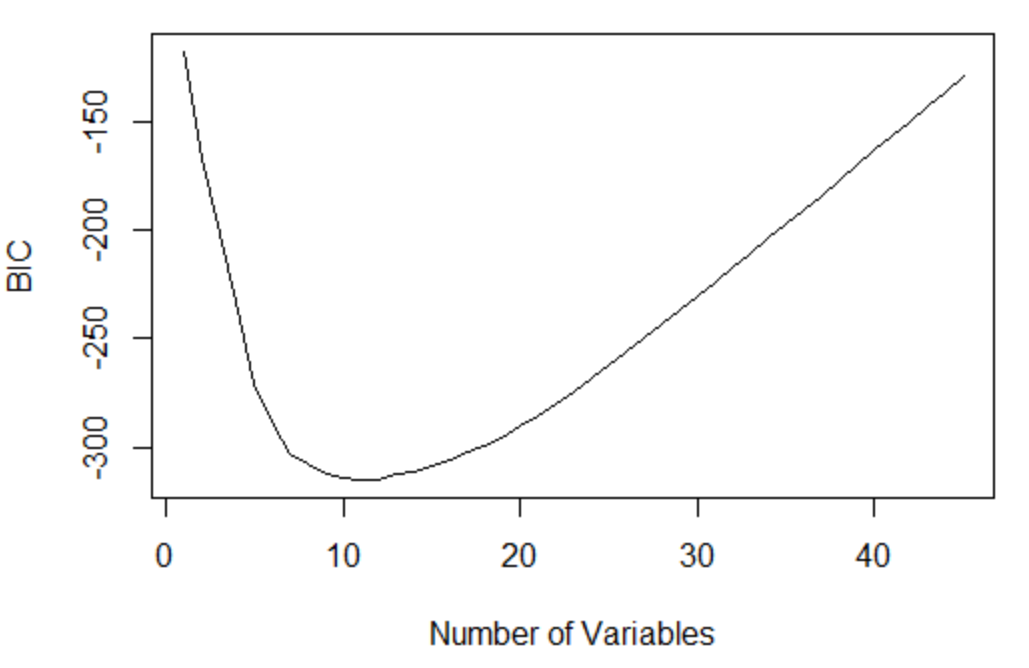


**Made the modified training set.**



**CP valure and BIC values for the training data by the number of variables (the summry of best subset selection)**

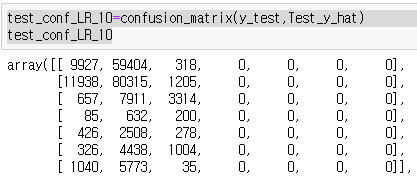


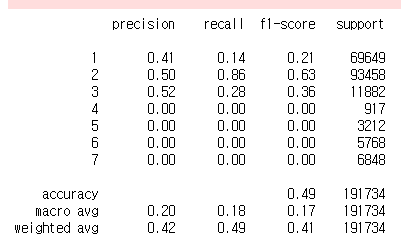


**Selected features**

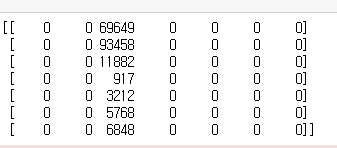


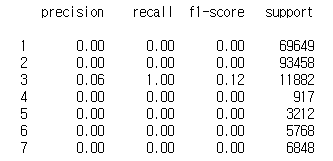
**The Rogistic model result**



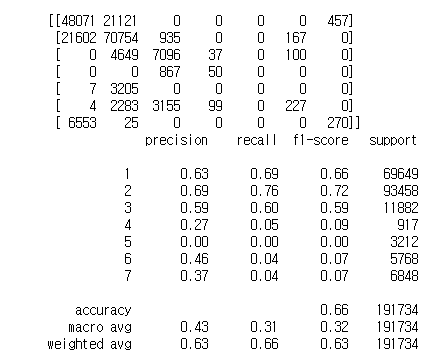


**The QDA model result**

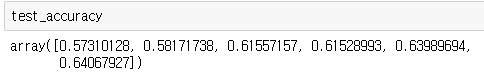


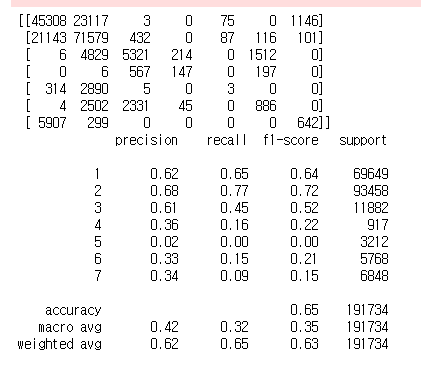


**The LDA model result**

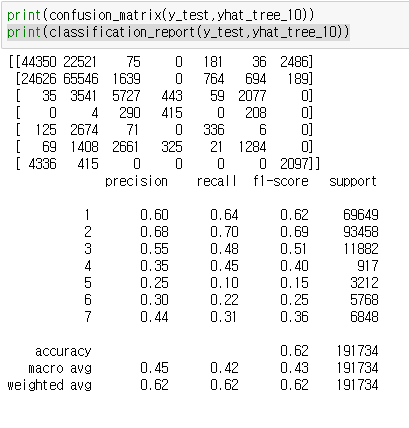


**The KNN model result**

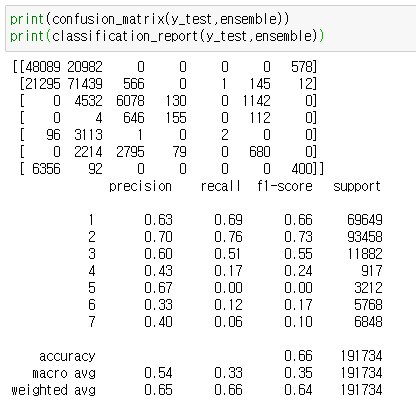




**The tree model result**



**Ensemble model result (KNN+TREE+LDA)**



Considersation

Data imbalace problem

To solve the data imbalance problem, I modified the training data ratio. However, if you make same ratio for all the training data set, I will cause different kind of data imbalance proble. So, I just change my training data ratio slightly.

**Feature selection.**

With the summery of the best subset selection for the training data, I made cp and bic table by the number of variables. Since there are large and so many variables, I used min bic to the feature selection.

**Result for each model.**

The Rogistic and the QDA model was not appropriate for this training data classfication. Their decision boundry masked some ‘Y’ values.

The tree, LDA and KNN shows bout 60-66 percentage of the accuracy.

**Ensemble model**

I made an ensemble model combine the LDA, tree, and KNN which follwing the ‘Votting’ system. I expected better performance than 66 percentage of accuracy, however, the performance was just same as 66.

**What should I do to get better model?**

I though that the data is so enough that I don’t have to do the Cross-validation method. I will try CV and compare with my result.

Also, change some feature selection possible to change the result.

And other model such as the Random forest can enhance the performances.