DMG Assignment - 2

Kaustav Vats | 2016048

Data description

Given dataset contains buy-product of information from customer clicks monitored by online fashion platform. Every attribute is the given dataset has categorical nature. Last columns which contains class label is of binary form [True, False]. Goal to predict whether user will view another page or not.

Pre-processing

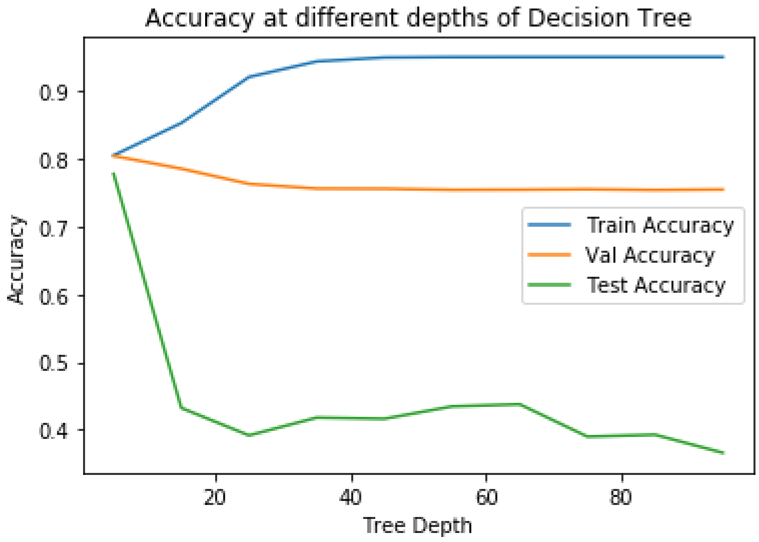
1. Given datasets contain lot of missing values in a form of “?” (None), “NULL”.
2. I replaced all ? and Null value with np.NaN.
3. I dropped columns having NaN > 50%. After this step I got 139 columns in which last column is the class Label.
4. I did one_hot_encoding for first 138 columns and also added extra nan column for each attribute. This gave total 817 columns.
5. I applied label encoder on class label and converted it to [0, 1] representation.

Question 1

For testing dataset I selected the same columns that were present in the training dataset.
Some results obtained after training decision tree with default parameters on whole training dataset.

Training Accuracy	94.82 %
Testing Accuracy	53.26 %

Now to find the best parameters for decision tree, I used stratified Kfold, where i created 5 folds to calculate validation accuracy, testing accuracy and training accuracy.
Performed K fold with [5-100, +10] depths of decision tree.



As we can clearly see that testing accuracy is significantly decreasing after certain depth of the tree. It seems like on increasing depth of the tree train accuracy is increasing but, for unseen data accuracy is decreasing on increasing **model complexity**. Overall decision tree is not able to differentiate between both the class properly. Due to class bias, its predicting every data point in one class.

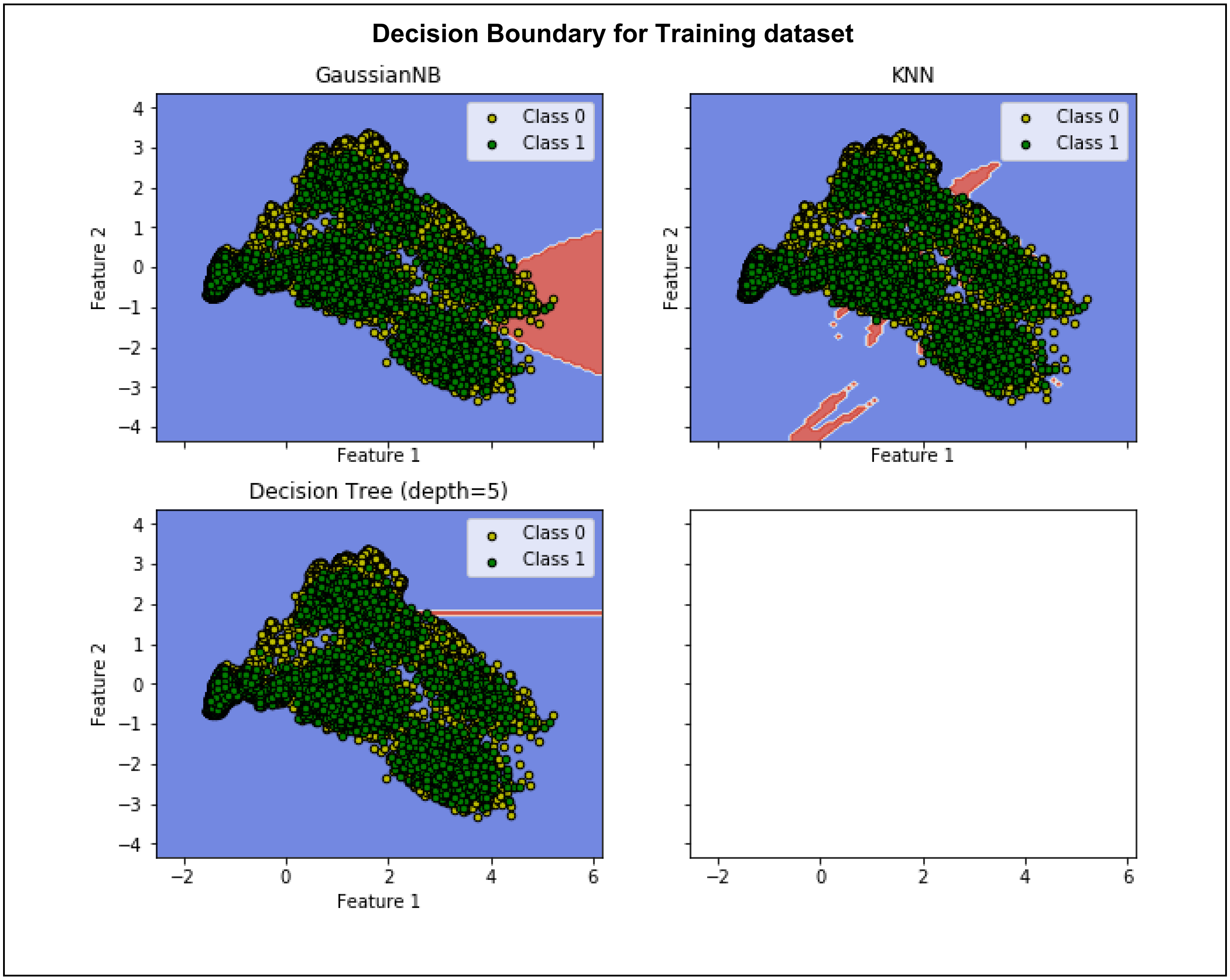
Best Accuracy	80.48
Best Depth	5

I also observed that there is not much difference in accuracy with depth in range [1-5].

Question 2

In order to two decision boundary in 2D, we need to plot data points in 2D.
I performed PCA with n_components=2, which will give me best 2 features based on which data has the best separation.
Then i created all classifier with default parameters except decision tree.

Using meshgrid i found decision boundary of all classifier on train and testing dataset.

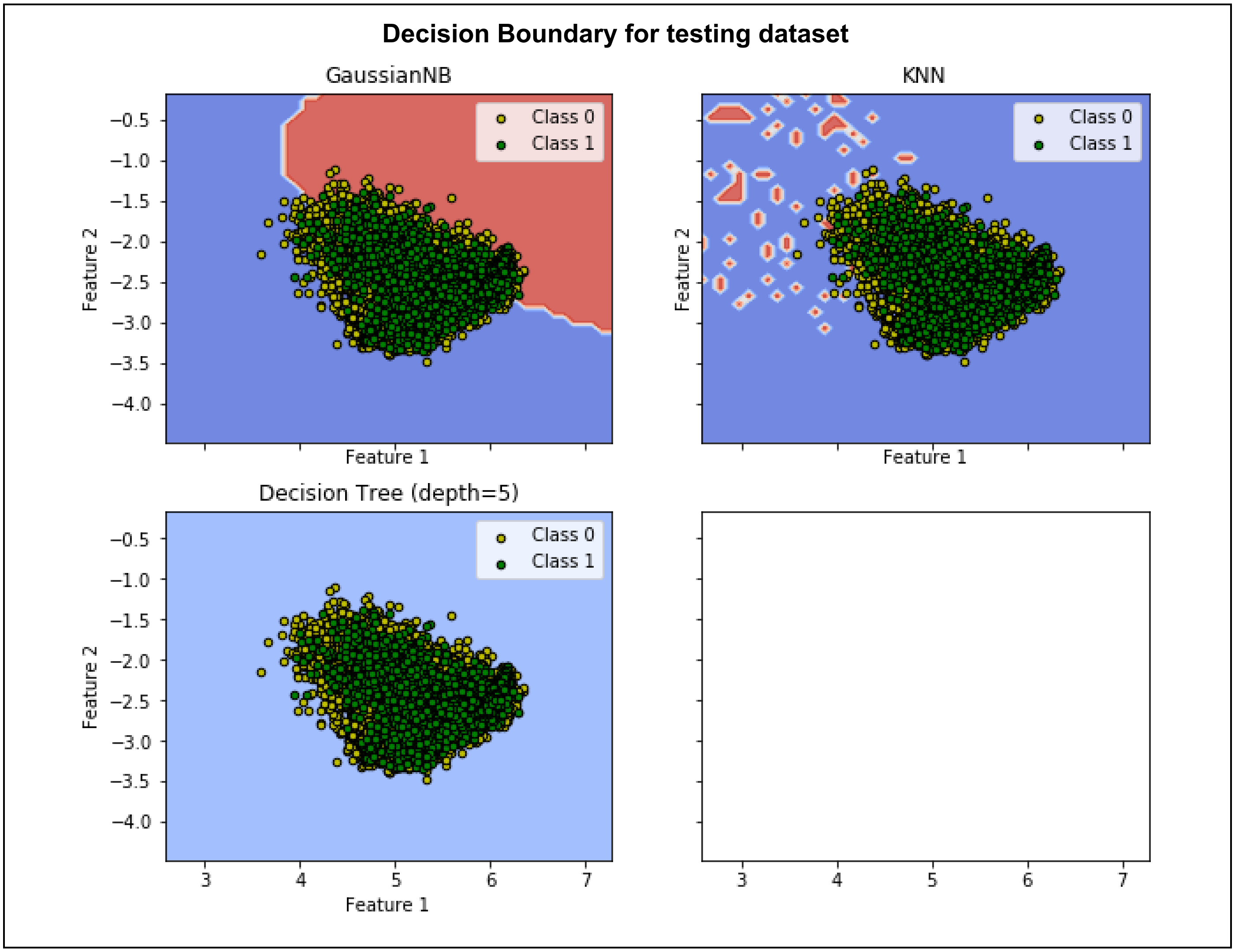


Decision Boundary for Training dataset

GaussianNB

KNN

Decision Tree (depth=5)



Decision Boundary for testing dataset

	GaussianNB	KNeighborsClassifier	DecisionTreeClassifier
Training Accuracy	80.29	83.27	80.50
Testing Accuracy	55.81	85.30	85.44

From decision boundary we can see that, decision tree is classifying all test data points in one class, also the testing accuracy of the decision tree is highest as compared to other classifier. This has happened because of class bias data and also the 2 features selected is not enough to fully classify dataset correctly.

Assumptions

1. Null value is basically a missing value or an unknown value. Which i'm treated as a NaN value.
2. I can assume NaN to be of different attribute value in my dataset.
3. There are no duplicate records in the dataset.
4. Stratified K fold is allowed.
5. Test accuracy need to be calculated on classifier which is trained on the training set of the fold.
6. No hyper parameter tuning required except decision tree.
7. No need to remove class bias, as observing disadvantages of class bias is also a task of this assignment.
8. Its ok to take only 2 features using PCA, in order to plot decision boundary.

9. Plot decision boundary for complete dataset