STEVENS
INSTITUTE of TECHNOLOGY
THE INNOVATION UNIVERSITY

Ethics Pledge

Consistent with the above statements, all homework exercises, tests and exams that are designated as individual assignments MUST contain the following signed statement before they can be accepted for grading.

I pledge on my honor that I have not given or received any unauthorized assistance on this assignment/examination. I further pledge that I have not copied any material from a book, article, the Internet or any other source except where I have expressly cited the source.

Signature: Haodong Zhao Date: Mar 11th 2019

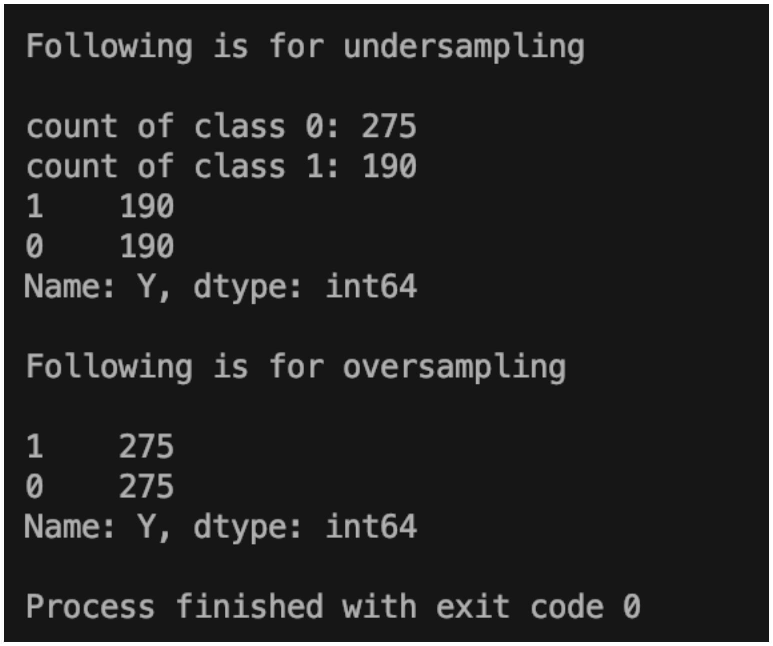
Please note that assignments in this class may be submitted to www.turnitin.com, a web- based anti-plagiarism system, for an evaluation of their originality.

1. Balanced the dependent variable (Y) using the resampling method (either oversampling or undersampling)

Answer:

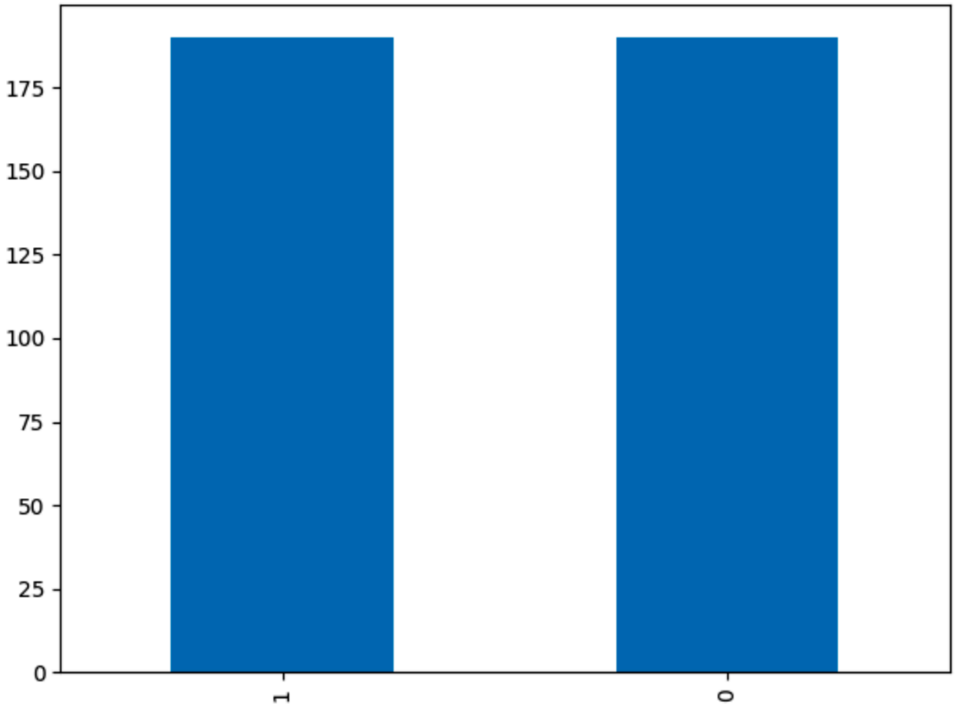
Split the dataset to 75% training data and 25% validation data with random_state=0

Result for undersampling and oversampling

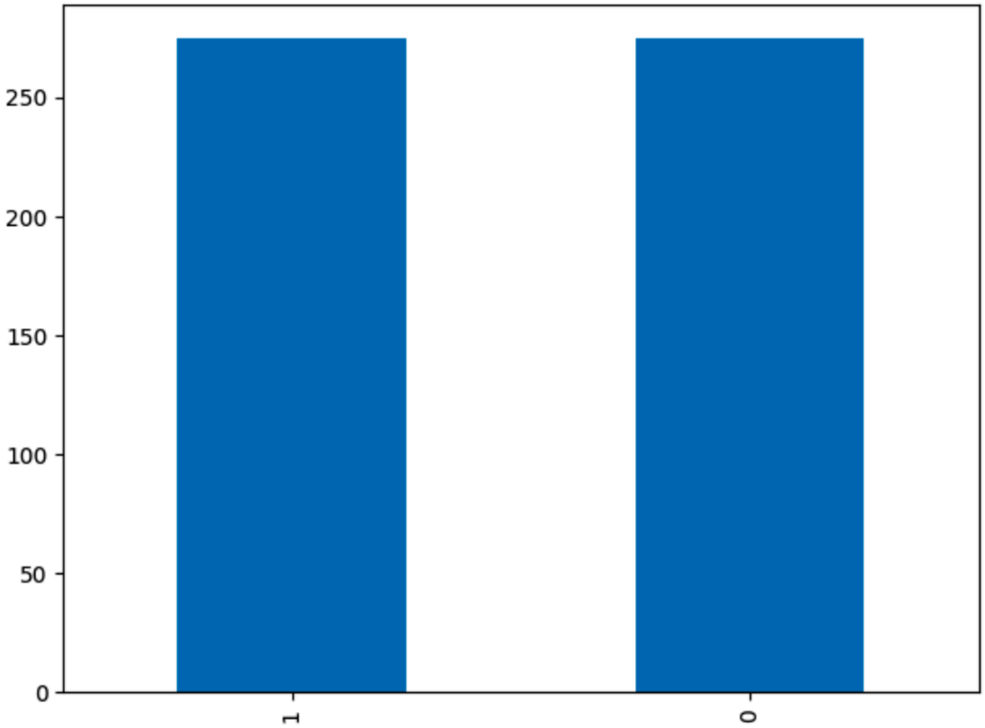


```
Following is for undersampling

count of class 0: 275
count of class 1: 190
1    190
0    190
Name: Y, dtype: int64

Following is for oversampling

1    275
0    275
Name: Y, dtype: int64

Process finished with exit code 0
```

Plot for undersampling:



Plot for oversampling:



2. **Develop Logistic Regression, Linear Discriminant Analysis, K-Nearest Neighbors, and Naïve Bayes models to classify Y using Xs (you can select some or use them all)**

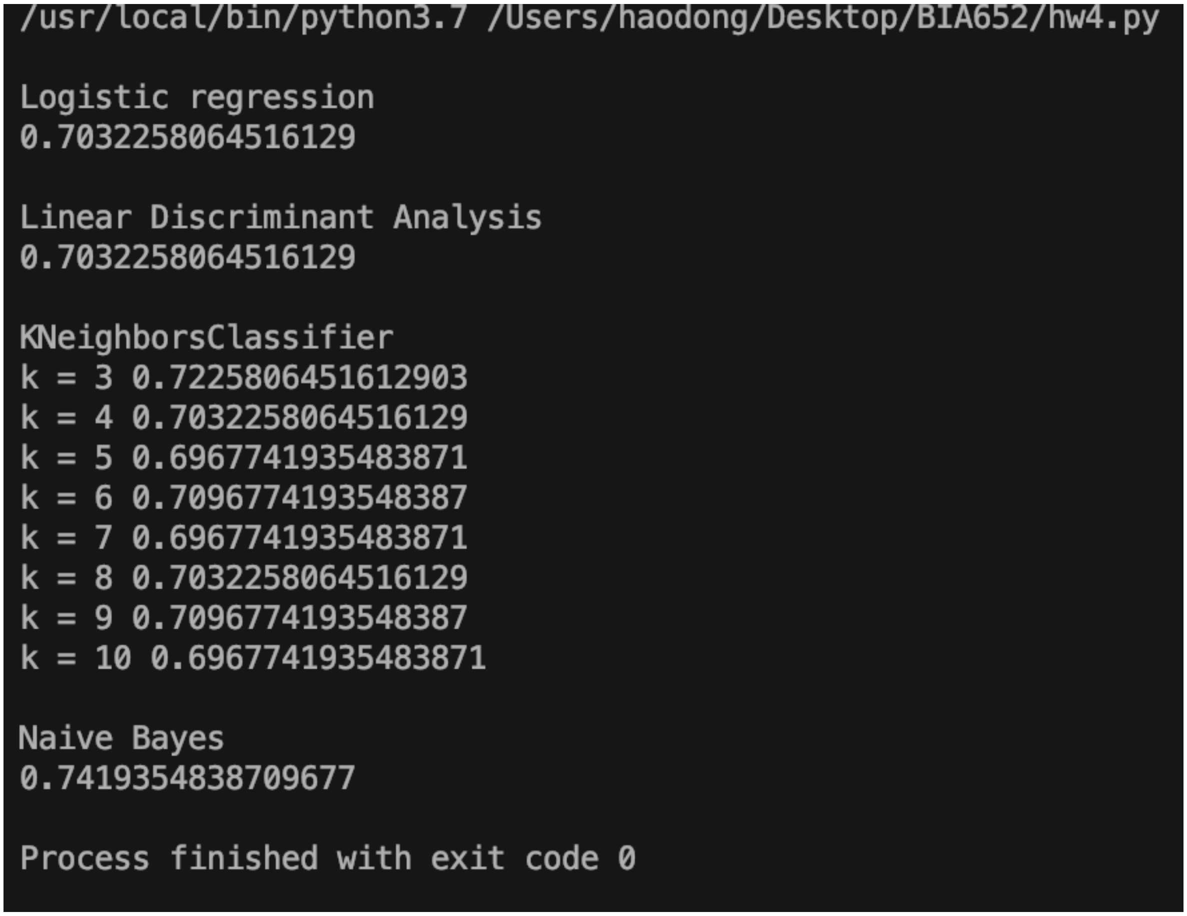
Answer:

Split dataset to 75% training data and 25% validation data.

Develop four kind of models and get their scores.

For logistic regression, use ‘lbfgs’ solver.

For KNN model, test different k from 3 to 10, and we can find when k = 3, KNN model provide the most accurate model.

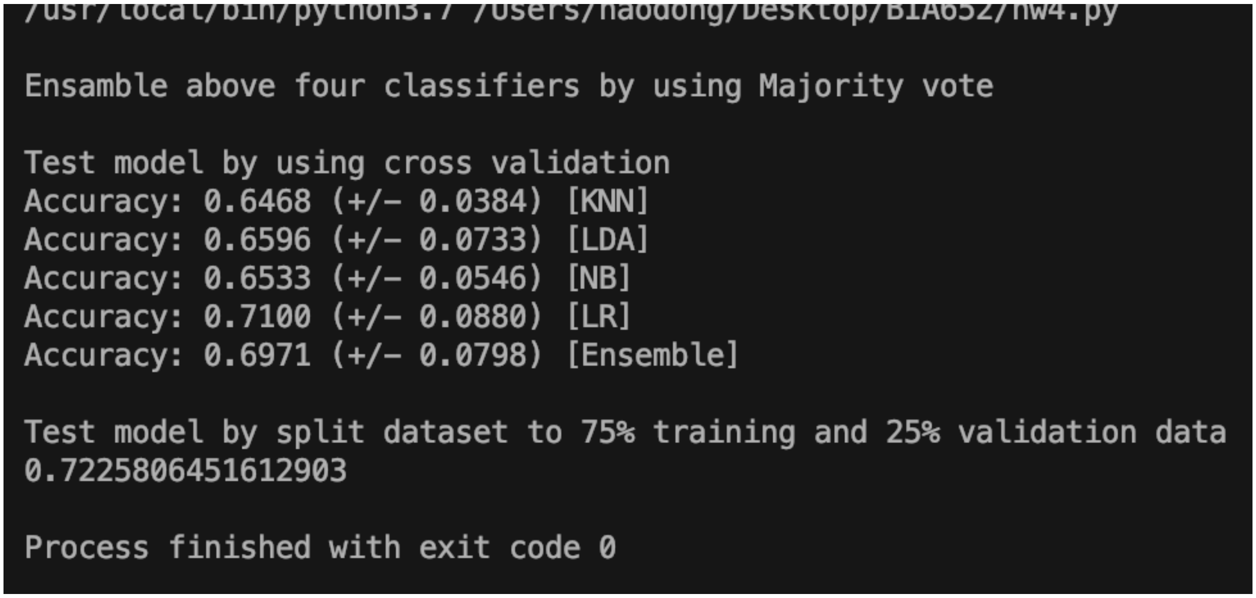


```
/usr/local/bin/python3.7 /Users/haodong/Desktop/BIA652/hw4.py

Logistic regression
0.7032258064516129

Linear Discriminant Analysis
0.7032258064516129

KNeighborsClassifier
k = 3 0.7225806451612903
k = 4 0.7032258064516129
k = 5 0.6967741935483871
k = 6 0.7096774193548387
k = 7 0.6967741935483871
k = 8 0.7032258064516129
k = 9 0.7096774193548387
k = 10 0.6967741935483871

Naive Bayes
0.7419354838709677

Process finished with exit code 0
```

3. Develop an ensemble of these four classifiers using the committee approach

Answer:

Ensemble the four classifiers by using Majority vote and then use 2 different way to test the ensemble model.



```
/usr/local/bin/python3.7 /Users/haodong/Desktop/BIA652/hw4.py

Ensamble above four classifiers by using Majority vote

Test model by using cross validation
Accuracy: 0.6468 (+/- 0.0384) [KNN]
Accuracy: 0.6596 (+/- 0.0733) [LDA]
Accuracy: 0.6533 (+/- 0.0546) [NB]
Accuracy: 0.7100 (+/- 0.0880) [LR]
Accuracy: 0.6971 (+/- 0.0798) [Ensemble]

Test model by split dataset to 75% training and 25% validation data
0.7225806451612903

Process finished with exit code 0
```