Data Mining Labwork 2

Abdullah Al Thaki, Roll: 12

Introduction

Naive Bayes classifier and decision tree classifier are two of the popular classifiers. They are known for their simplicity. In Lab 2, I have implemented these two classifiers and made a comparison between them in real life data. Data can be discrete or continuous value.

Dataset statistics, train-test split and other information

Several dataset are chosen from the UCI machine learning repository. Datasets statistics are given in the table as well as train-test split size.

	total	training	testing
tic-tac-toe.data	957	765	192
wine.data	177	141	36
heart.data	269	215	54
iris.data	149	119	30
kr-vs-kp.data	3195	2556	639
balance-scale.data	624	499	125
car.data	1727	1381	346
nursery.data	12959	10367	2592
agaricus-lepiota.data	8123	6498	1625
krkopt.data	28055	22444	5611

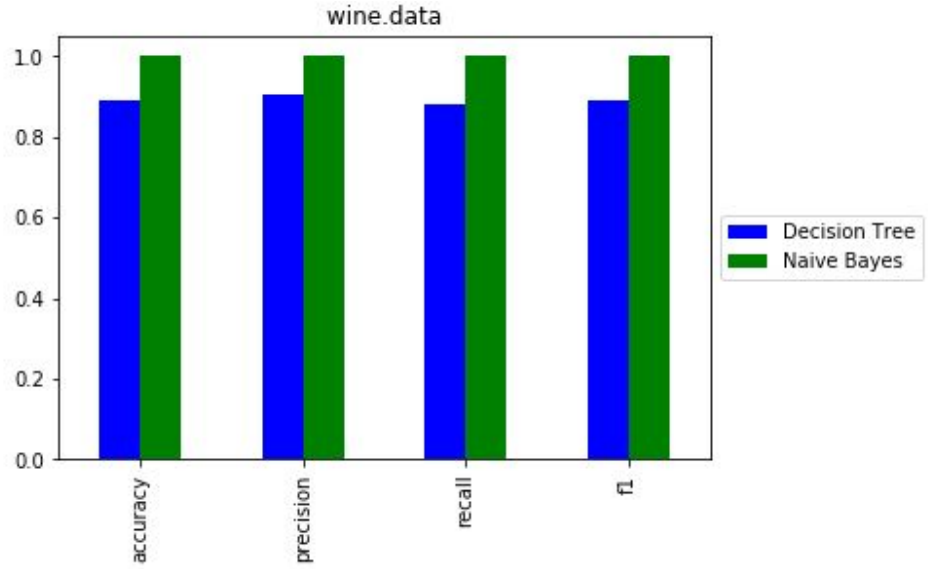
Programming Language: Python
Train-test ratio: 80%:20%

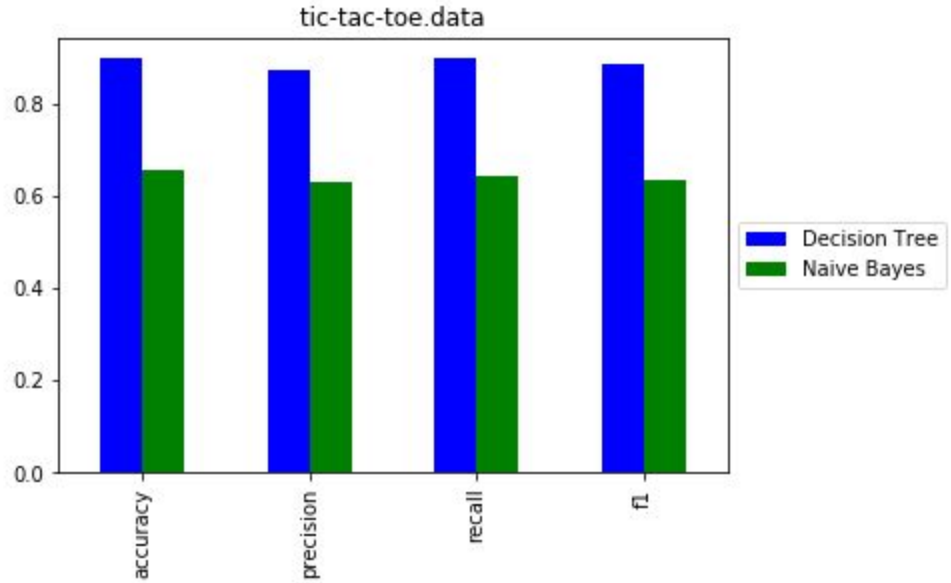
Implementation Notes:

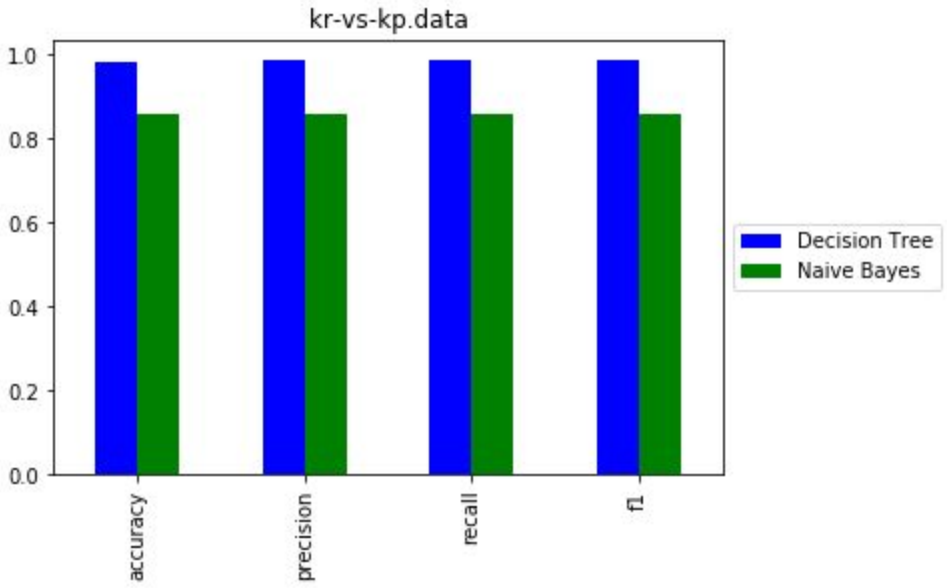
Most of the special cases are handled for both of the classifiers. Gini index has been incorporated for decision tree classifier. Both classifiers can handle discrete or continuous values. As performance measures accuracy, precision, recall, f1-score are calculated.

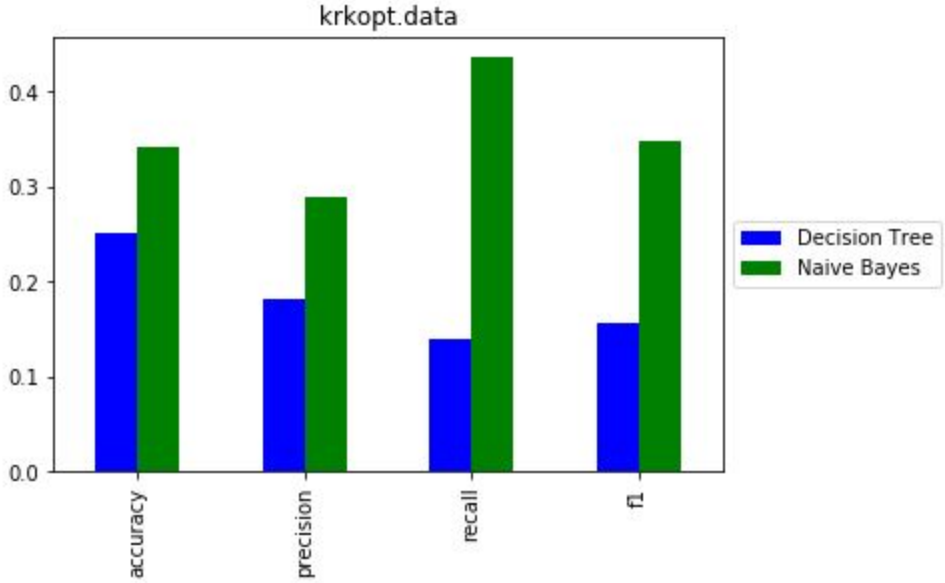
Comparison

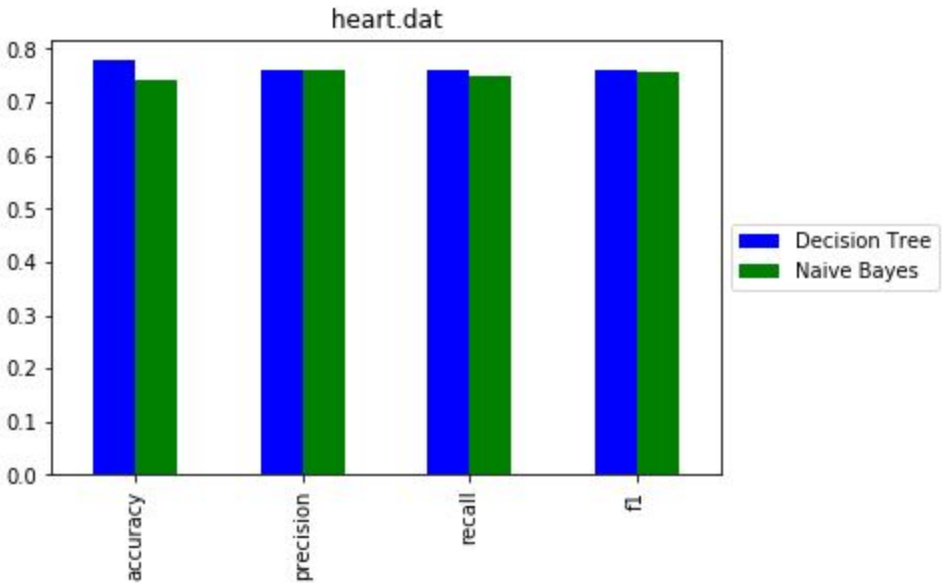
Performance comparison graphs are shown below for different datasets.

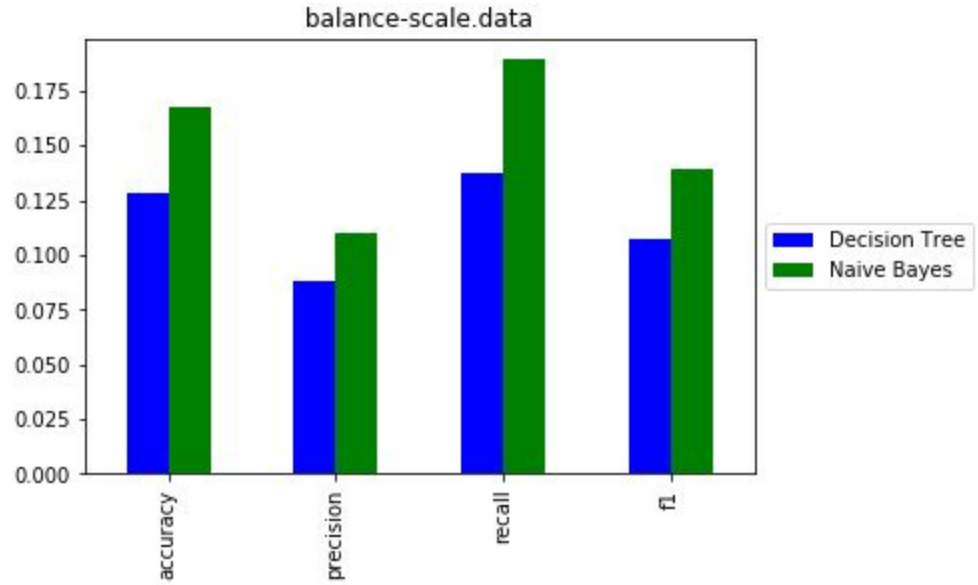


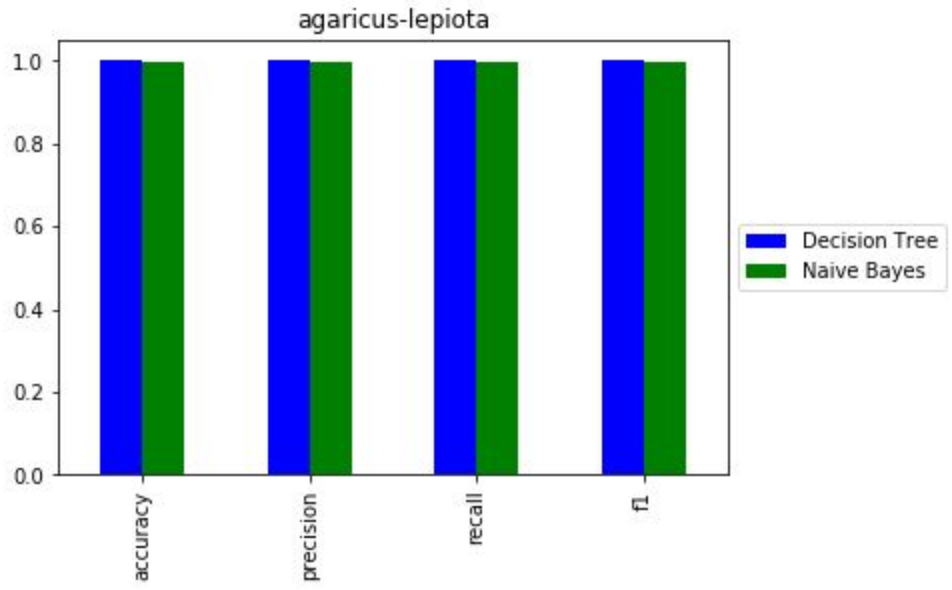


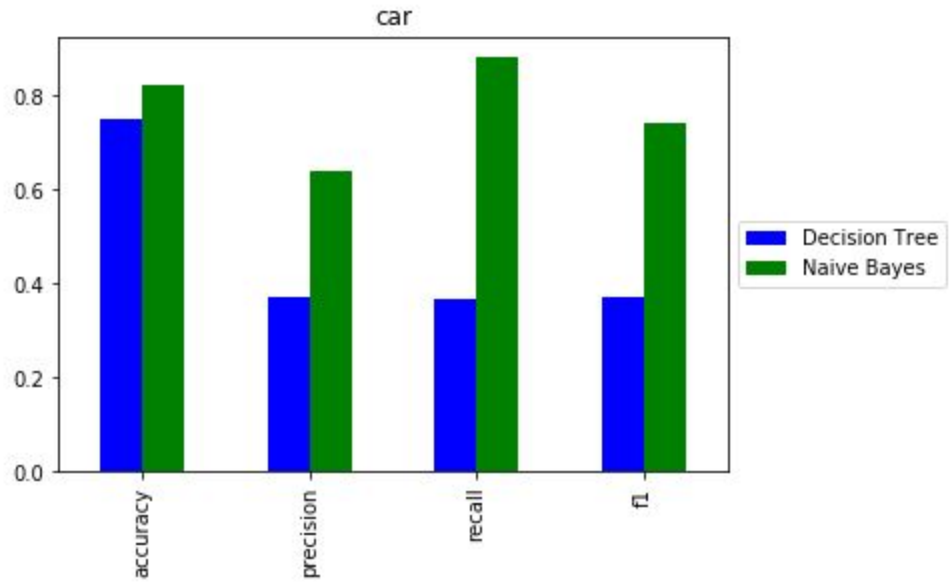

kr-vs-kp.data
1.0
0.8
0.6
0.4
0.2
0.0
accuracy
precision
recall
f1
Decision Tree
Naive Bayes

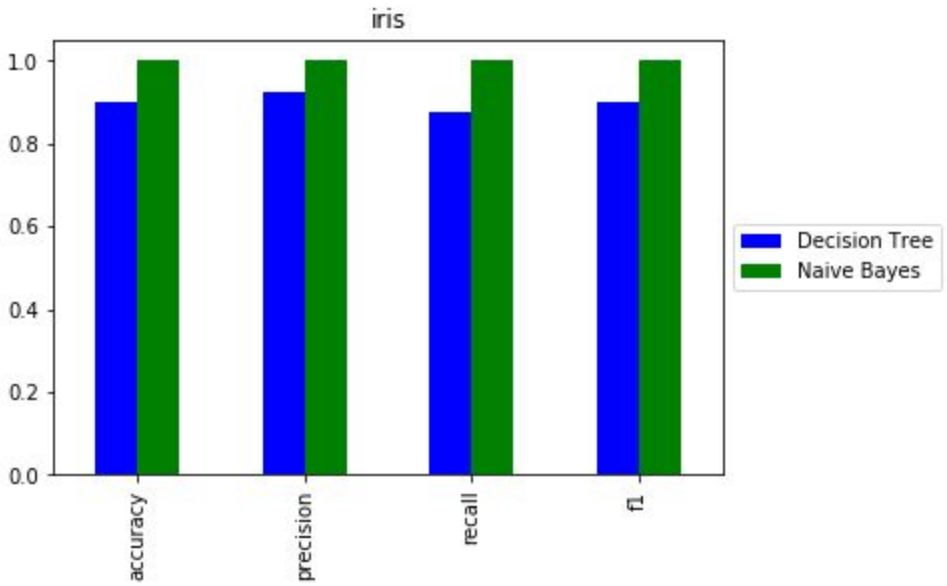

krkopt.data
0.4
0.3
0.2
0.1
0.0
accuracy
precision
recall
f1
Decision Tree
Naive Bayes

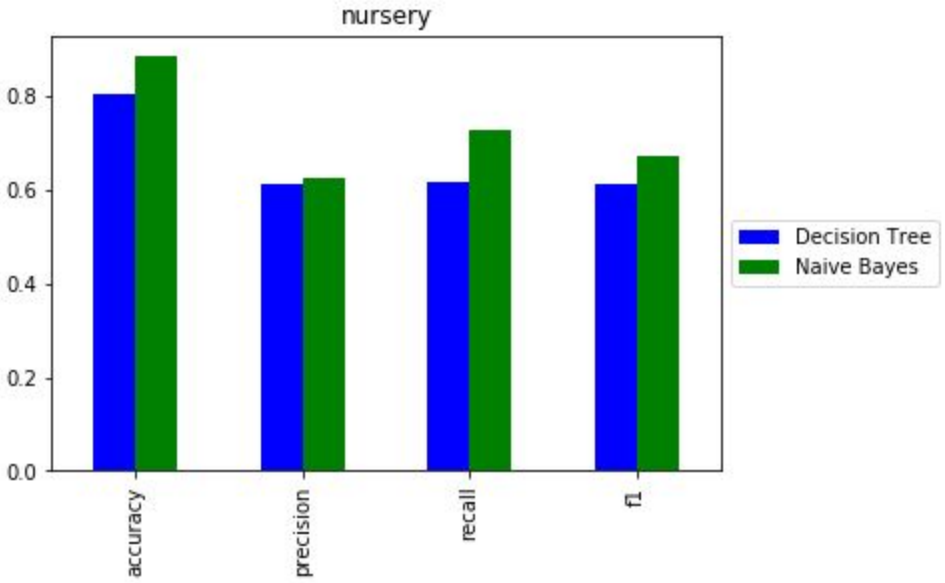

heart.dat
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0.0
accuracy
precision
recall
f1
Decision Tree
Naive Bayes


balance-scale.data
0.175
0.150
0.125
0.100
0.075
0.050
0.025
0.000
accuracy
precision
recall
f1
Decision Tree
Naive Bayes


agaricus-lepiota
1.0
0.8
0.6
0.4
0.2
0.0
accuracy
precision
recall
f1
Decision Tree
Naive Bayes


car
0.8
0.6
0.4
0.2
0.0
accuracy
precision
recall
f1
Decision Tree
Naive Bayes





Finding

According to these empirical data, we can see naive bayes classifier has performed well compared to decision tree classifier. But we can not declare that naive bayes classifier is better than decision tree classifier. In this particular case naive bayes have performed well because I may have implemented both algorithms in a particular way, dataset choice helped naive bayes classifier in a positive way. There could be better implementation or other dataset choice for which result could be otherwise.