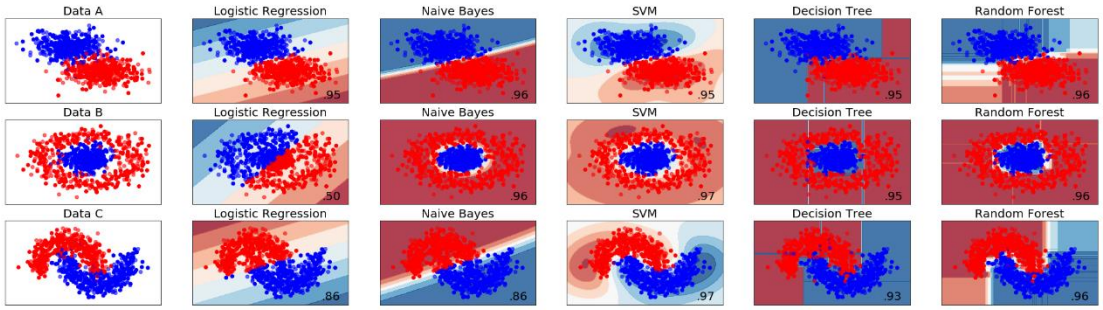
Report Assignment 3
Liyang Ru liru4968

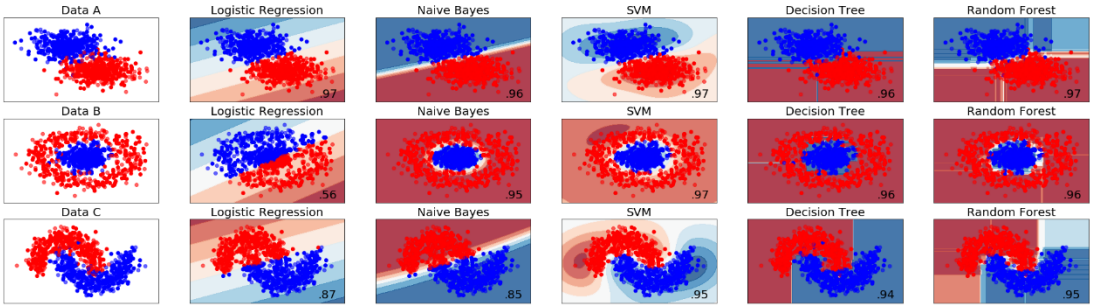
Part 1

a. For each classifier, we tested for three times and compared the accuracies.

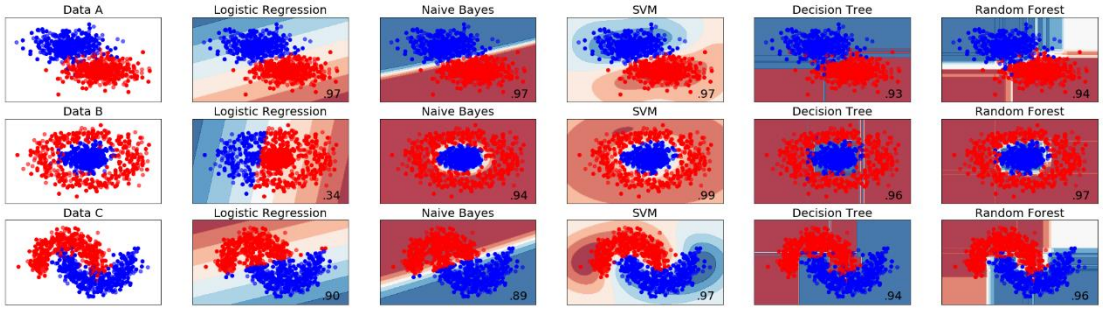
1.



2.



3.



So we find:

1. For Logistic Regression, it always has very high accuracy on Data A, very low accuracy on Data B and high accuracy on Data C but lower than SVM, Decision Tree and Random Forest on C.
2. For Naïve Bayes, it always has very high accuracy on Data A, very high accuracy on Data B and high accuracy on Data C but lower than SVM, Decision Tree and Random Forest on C.
3. For SVM, it always has very high accuracy on Data A, highest accuracy on Data B and highest accuracy on Data C (a little higher than others).
4. For Decision Tree, it always has very high accuracy on Data A, B, C.

5. For Random Forest, it always has very high accuracy on Data A, B, C and a little higher than Decision Tree on 3 datasets.

b.

1. For Logistic Regression, it performs well on linear separability, but bad on nonlinear separable data sets. And the decision boundary of it is always linear and tends to be underfitting.
2. For Naïve Bayes, it performs well on simple separable data set (like data A and B), but no very good on data set like data C. And the decision boundary of it is like to be linear or curvy and it has a good balance.
3. For SVM, it performs very well on 3 data set models but the decision boundary of it is often overfitting.
4. For Decision Tree, it performs very well on 3 data set models and the decision boundary of it is always a good balance.
5. For Random Forest, it performs very well on 3 data set models but the decision boundary of it is sometimes overfitting.

Part 2

a.

1.

```
Logistic Regression :  0.8572360380923156   0.06311500789067366

Naive Bayes :  0.7640004440277093   0.06769865309865102

SVM :  0.49827586206896546   0.005452202862359268

Decision Tree :  0.8204131330803721   0.04634489893741456

Random Forest :  0.8489330341857508   0.05781972037112053

KNN :  0.6403548830904761   0.08675145930283096

Gaussian Process Classifier :  0.47677986666836913
0.050617180725880395
```

	Logistic Regression	Naïve Bayes	SVM	Decision Tree	Random Forest	KNN	Gaussian Process Classifier
Mean	0.8572	0.7640	0.4983	0.8204	0.8489	0.6404	0.4768
SD	0.0631	0.0677	0.0055	0.0453	0.0578	0.0868	0.0506

2. I will choose Random Forest as best overall model. Although the mean of RF is lower than that of Logistic Regression, it has lower standard deviation than LR. Similarly, the SD of Decision Tree is lower than RF, but RF has higher mean than that of DT. To some degrees, Logistic Regression, Decision Tree and Random Forest are all acceptable as best overall model.

3. KNN: The k-nearest neighbors algorithm (k-NN) is a non-parametric method used for classification and regression. In both cases, the input consists of the k closest training examples in the feature space. The output depends on whether k-NN is used for classification or regression.

Gaussian Process Classifier: Gaussian process is a stochastic process (a collection of random variables indexed by time or space), such that every finite collection of those random variables has a multivariate normal distribution. The distribution of a Gaussian process is the joint distribution of all those (infinitely many) random variables, and as such, it is a distribution over functions with a continuous domain.

b.

```
Best score:  0.7310033565468349
For SVC
Best C is 1000
Best gamma is 0.0001

Best score:  0.93280429031516
For Random Forest
Best max_depth is 20
Best n_estimators is 800
```

1. We use GridSearchCV function in sklearn library. Firstly, we give lists for all the parameters as options. Then we call the GridSearchCV function to do the k-fold cross validation while filling the parameters with Classifier, parameter choices, scoring standard and k (of k-fold cross validation). Then we fit it with our data (X and y). It will use different parameter combinations from the choices and calculate AUC. Then we can find the best score (highest AUC) and corresponding parameters.

2. C: C is the regularization parameter that controls the trade-off between the misclassification penalty and margin width.
Gamma: Margin, i.e., how close the separating boundary is to the points.
Max_depth: Number of level splits according to attributes
N_estimator: The number of trees to be used in the forest.

3. Yes. We find a better model than 2.a.2 with hyperparameter tuning since AUC of Random Forest (0.9328) is larger than before.

c.

1.

```
[[20  0]
 [ 3 27]]
Precision is  1.0
Recall is  0.9
Accuracy is  0.94
AUC is  0.95
```

```
[[20  7]
 [ 1 22]]
Precision is  0.7586206896551724
Recall is  0.9565217391304348
Accuracy is  0.84
AUC is  0.8486312399355878
```

```
[[20  4]
 [ 8 18]]
Precision is  0.8181818181818182
Recall is  0.6923076923076923
Accuracy is  0.76
AUC is  0.7628205128205129
```

Since I shuffle the data when testing the best model, I get different confusion matrix each time.

2. If the accuracy is high, then the person has a good credit.

d.

1. RandomForestClassifier(n_estimators=800, max_depth=20)