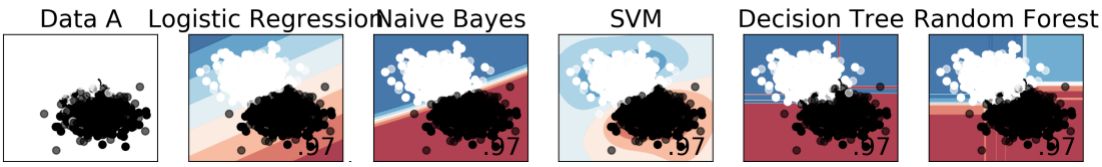
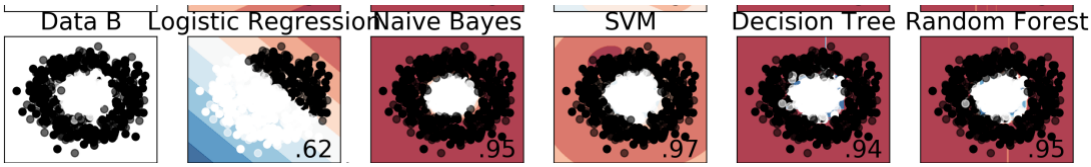
Part 1

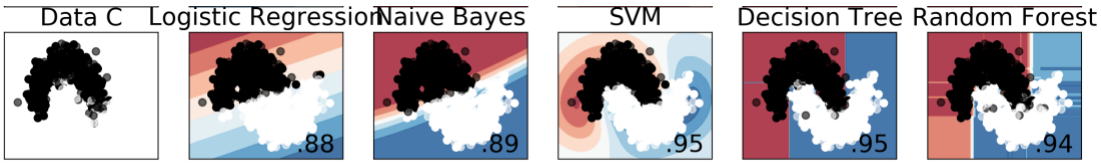
1. For data set A, logic regression, all the learning methods have the same accuracy.



For data set B, logistic regression yields a significant low accuracy compare to all the other learning methods. And all the other learning method have similar accuracy. SVM has the highest accuracy.



For data set C, logic regression, Naïve Bayes have lower accuracy compare to all the other methods. SVM, Decision Tree and Random Forest have similar accuracy.



2.

a) Linear separability:

i. For data set A, it because the two groups of data are separated more smoothly. It is easy to find a line that separate the two groups.

ii. For data set B, because the two group of data are like a circle surround with another. Therefore, drawing a line that separate two data sets would contain a lot of inaccuracy. Also, linear regression is highly dependent on the separation line (because the model of linear regression is a line), therefore, for dataset B, linear regression has the least accuracy.

iii. For data sets C, because the shape of the two data is more bending compare to data set A, so it is harder (it is still doable, but less accurate) to draw a line that separate two group of data sets. The line that separate two groups of data will eventually have some inaccuracy. Therefore, for logistic regression. Same reason with dataset B, drawing a separation line will contain some inaccuracy. For Naive Bayes because it added Slack

Variable to allow misclassification. Therefore, there is tradeoff between the misclassification penalty and margin width.

b) Decision Boundaries

i. SVM

For dataset B and C, because it is hard to use a line to separate two group of data. SVM could map data to higher dimensional space. Therefore, it could still yield a good result.

ii. Decision Tree

Because decision tree is a non-linear mapping x to y. The decision boundary of Decision tree is all points on y coordinates equal to the threshold. Therefore, for dataset B and C. It could still have good accuracy.

iii. Random Forest

For random forest, same with decision tree. It can model nonlinear relationship. So, it could yield a good result.

Part2

A

1.

	Average	Standard Deviation
Logistic Regression	0.8525	0.0573
Naive Bayes	0.7676	0.052
SVM (SVC)	0.4983	0.0052
Decision Tree	0.7951	0.039
Random Forest	0.8462	0.0458
Neural Network	0.717	0.0751
Orthogonal Matching Pursuit (OMP)	0.8438	0.0633

2. Since Logistics Regression model have the highest AUROC average and relatively low standard deviation value (with AUROC average=0.8525 and Standard Deviation=0.0573), therefore, logic regression is the best overall model.

3.

a) For Neural Network, it simulates the data by reassemble it to neurons. Neural Network have 3 layers, input layers (use units represents input filed), output layers and hidden layers. It uses multiple hidden layer to differentiate the input and output (what different between logistic regression). Also, the layers in neural network is arranged by layers. Neural Network learn by testing and gathering prediction of each records. If there is incorrect prediction, it adjusts the weight. Therefore, by doing this process, the prediction become higher [2].

b) For Orthogonal Matching Pursuit, like greedy algorithm, in each iteration, it selects the best fitting column of the sensing matrix using an orthogonal projection. It approximately fit a linear model with limitations on the number of non-zero coefficients.

B

1. For both SVC and Random Forest. I list some value and append the average AUROC score to a list, then choose the best AUROC score and get the hyperparameter selection. For SCV, the C values are [0,1, 1, 10, 100, 100], the gamma values are [1, 0.1, 0.001, 0.0001] and do "crossover" for each C values and gamma values. For example, C=0.1 and gamma=1, C=0.1 and gamma=0.1, C=0.1 and gamma=0.01, C=0.1 and gamma=0.001, C=0.1 and gamma=0.0001. For Random Forest, same with SCV, the max depth values are [2,4,6,8] and the n estimator values are [20, 40,60,80,100.
2. SVC
   a) C is the penalty parameter of the error term. It controls how much you want to avoid misclassifying each training example.
   b) Gamma parameter controls how much the far of a single training example reaches. With low value means far, the high value means close [1].

   Random Forest
   a) Max depth controls the max depth of the tree structure.
   b) N estimators controls the number of trees in each forest.
3. Yes, by changing the penalty of logistic regression to "l1" (default is l2) and increase the max iteration number. It will slightly increase the average AUROC score.

C

1. Confusion Matrix

Test	Present	Absent
Positive	204	36
Negative	235	25

Accuracy Matrix

Accuracy	Precision	Recall	AUROC
0.8851	0.85	0.8908	0.879

2. Base on the calculation above, the accuracy of logistic regression is 0.8851. Therefore, to determine if a person have a good or bad credit. It could have 0.1149 error rate.

D

1. Name of model: Logistics Regression.
   Hyperparameter: default.

References

1. https://scikit-learn.org/stable/auto_examples/svm/plot_rbf_parameters.html
2. https://www.ibm.com/support/knowledgecenter/en/SS3RA7_15.0.0/com.ibm.spss.modeler.help/neuralnet_model.htm