Assignment 3
Name: Chakrya Ros

Part 1

1a.

- Logistic Regression:
  - Data A has 0.97 accurate
  - Data B has 0.42 accurate
  - Data C has 0.87 accurate
- Naïve Bayes:
  - Data A has 0.97 accurate
  - Data B has 0. 94 accurate
  - Data C has 0.89 accurate
- SVM
  - Data A has 0.97 accurate
  - Data B has 0.97 accurate
  - Data C has0.96 accurate
- Decision Tree
  - Data A has 0.96 accurate
  - Data B has 0.95 accurate
  - Data C has 0.94 accurate
- Random Forest
  - Data A has 0.95 accurate
  - Data B has 0.94 accurate
  - Data C has 0.96 accurate

1b. From various classifiers, the support vector machine, decision tree and random forest have high accurate on three different of dataset and have best decision boundary that separates the data points into regions which are actually classes in which they belong. In these three datasets, these classifiers classified the class A and class B separately by decision boundary. Because they had percent between 94 to 97 of accuracy score, it means that they accurately classified the class A and class B between 94 to 97 percent, and they had 3 to 6 percent misclassified that some data points of class A were in the class B and via versa.

Part 2
2a.1

| **Classifiers** | **Mean AUROC Cross 10 Folds** | **Standard Deviation AUROC Cross 10 Folds** |
|---|---|---|
| Logistic Regression | 0.8464 | 0.0624 |

| | | |
|---|---|---|
| Naïve Bayes | 0.7676 | 0.0520 |
| SVM | 0.4983 | 0.0052 |
| Decision Tree | 0.7825 | 0.0261 |
| Random Forest | 0.8250 | 0.0432 |
| Linear Regression | 0.9106 | 0.0522 |
| Gradient Boosting | 0.8459 | 0.0622 |

2a.2. Linear Regression is the best overall model because it has higher mean AUROC among these seven classifiers, and it has 0.0522 standard deviation AUROC away from mean AUROC. Thus, linear regression could classify these two class correctly.

2a.3 I chose linear regression and Gradient boosting.

- Linear regression is used to predict a dependent variable value based on independent variables which are features. Each feature is continuous variable. We use linear regression to find the best fit line which is known as regression line and its linear equation is Y = a*X + b. X is independent variable and Y is dependent variable, a is the slop and b is intercept.
- Gradient boosting is a boosting algorithm for regression and classification that make a prediction model in the form of an ensemble of weak prediction model, typically decision tree.

2b.1

I used two different technique for each classifier. One technique, I manually try different combination of hyperparameters. For SVC, I set C = [1, 10, 100, 1000] and gamma = [0.001, 0.0001], I found the best parameters are C=1000 and gamma = 0.0001 for high mean AUROC. For random forest, I set max_depth = np.linspace(1, 32, 32, endpoint=True), and n_estimators = [1, 2, 4, 8, 16, 32, 64, 100, 200]. I found the max_depth = 9 and n_estimators = 200 for high mean AUROC. Another technique, I used GridSearchCV from SciKit-Learn to implement a fit and score method. I used GridSearchCV which take estimator that is classifier, param_grid that is C, gamma for SVC, max_depth and n_estimator for random forest. The best parameter

- SVC: {'C': 1000, 'gamma': 0.0001}
  - The mean AUROC of SVC : 0.7042
  - The Standard Deviation of SVC : 0.0589
- Random Forest: {'max_depth': 9.0, 'n_estimators': 200}
  - The mean AUROC of Random Forest: 0.8657
  - The Standard Deviation of Random Forest: 0.0553

2b.2

- C is regularization parameter that we want to void misclassifying each training example. The large values of C allow optimization to get a smaller margin hyperplane to classify the all training points correctly. However, a small value of C will make optimization to

look for a larger margin that separate the hyperplane, and hyperplane misclassify more points.

- Gamma is parameter that refer to how far the influence of a single training reach, with low values meaning far and with high values meaning close.
- Max_depth is the depth of each tree in the forest. The deeper tree, the more splits it has and get more information about data.
- n_estimators is the number of the tree in the forest. The higher number of trees capture more about data.

2b.3

| **Classifiers** | **Mean AUROC Cross 10 Folds** | **Standard Deviation AUROC Cross 10 Folds** |
|---|---|---|
| SVM | 0.7042 | 0.0589 |
| Random Forest | 0.8601 | 0.0432 |
| Linear Regression | 0.9106 | 0.0522 |

It's improved from 0.4983 to 0.7042 mean AUROC for SVM, from 0.8250 to 0.8601 mean AUROC for Random Forest, but it's still lower than Linear Regression that had 0.9106 mean AUROC. Thus, Linear Regression is the best overall model.

2c.1
Confusion matrix for my model
[210 61]
[ 14 215]
Accuracy = 85.0
Precision = 77.89
Recall = 93.88
AUROC Score = 85.69

2c.2 I conclude that the person has good credit because accuracy about 85.0%, and recall is about 93.88% which means that we have more true positive and less false negative, and precision value are acceptable.
2d.1. my model is Linear regression with no hyperparameter.