1. **Data Source and Contents**

   https://drive.google.com/file/d/13335ZpTuNi7bE25go_4nNpRXpVm90zxO/view?usp=sharing

   The link to access the dataset.

   Import the packages and functions that are needed for the assignment.

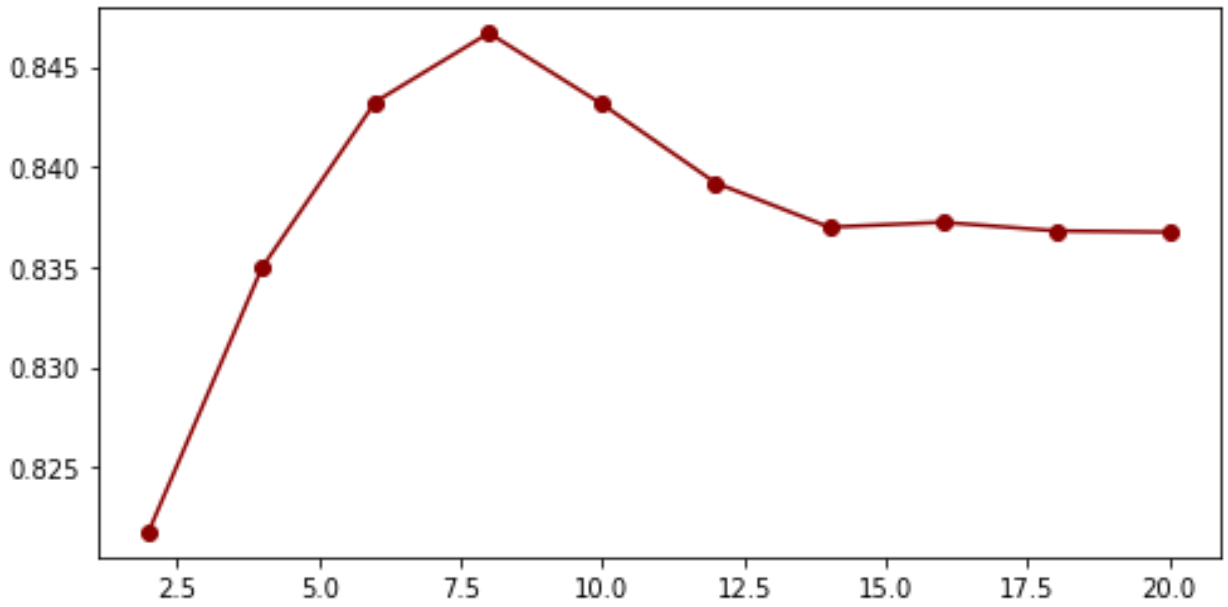
   Apply label encoder.

   Label x-test, y-test, x-train, and y-train.

2. **Finding Optimal Value of a key Hyper-parameter**

   `[2,4,6,8,10,12,14,16,18,20]`

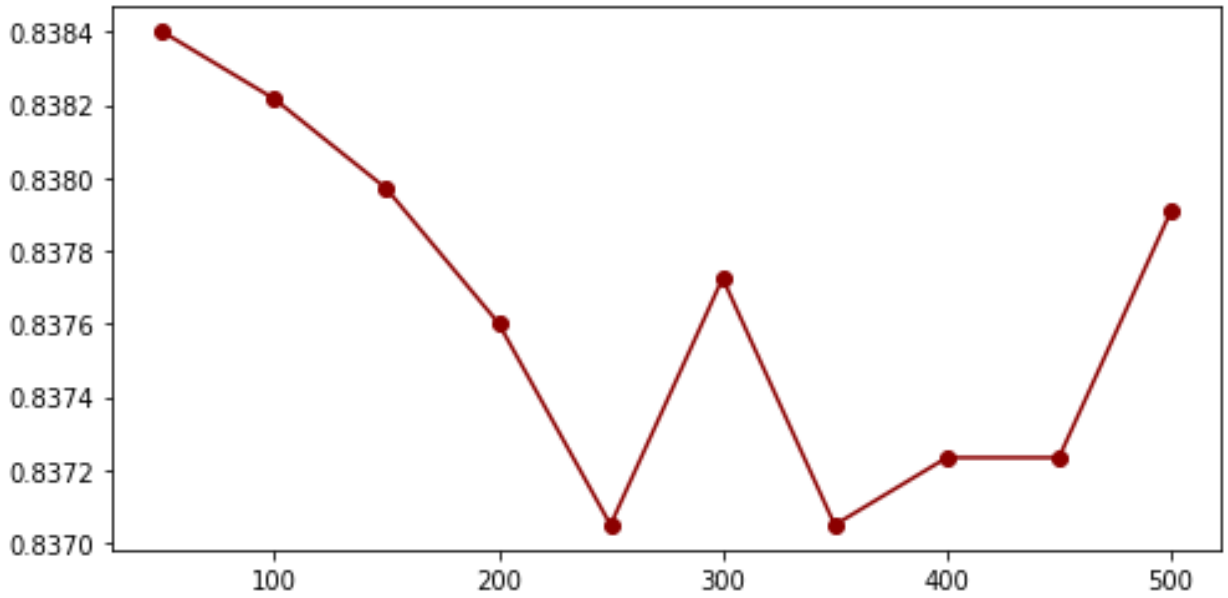
   Run the Decision Tree model with these max depth options, and the graph shows max depth = 8 gives the highest accuracy score.



3. **Building a Random Forest Model**

   [50,100,150,200,250,300,350,400,450,500]

   Run the Random Forest Model with these n_estimators.



Questions and answers:

Write your observations about the Classifier's behavior with respect to the number of estimators.

- The accuracy score is highest at the n_estimator 50. The score decreases when the n_estimator goes from 50 to 250. The score goes up at 300 and goes down again at 350. The score keeps changing when the n_estimator goes up without any clear pattern.

Is there an optimal value of the estimator within the given range?

- The optimal value of the estimator is 50 because it gives the highest accuracy score.

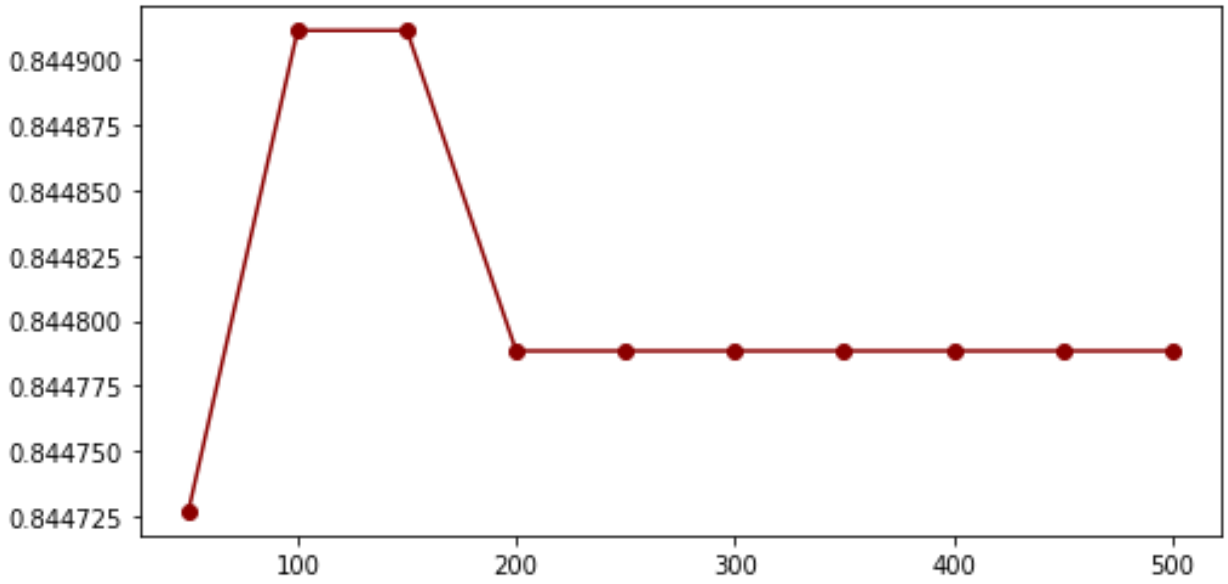
4. Building AdaBoost, Gradient Boost (classifier) and XGB Model

[50,100,150,200,250,300,350,400,450,500]

Apply the conditions of these n_estimators into AdaBoost, Gradient Boost, and XGB model.

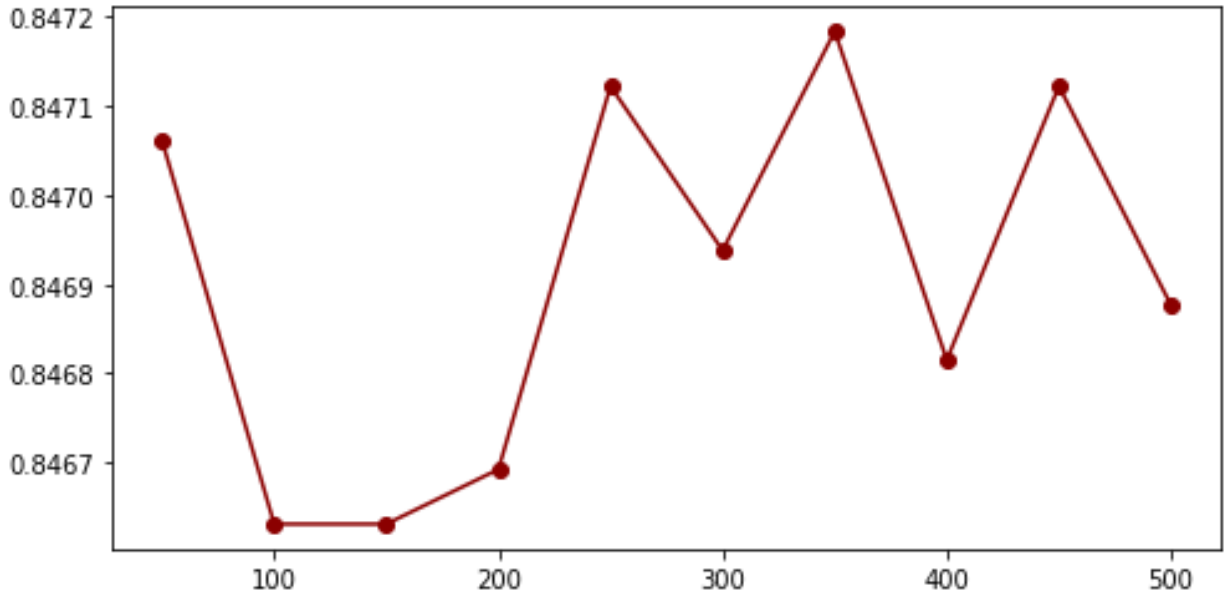
Questions and answers:

AdaBoost



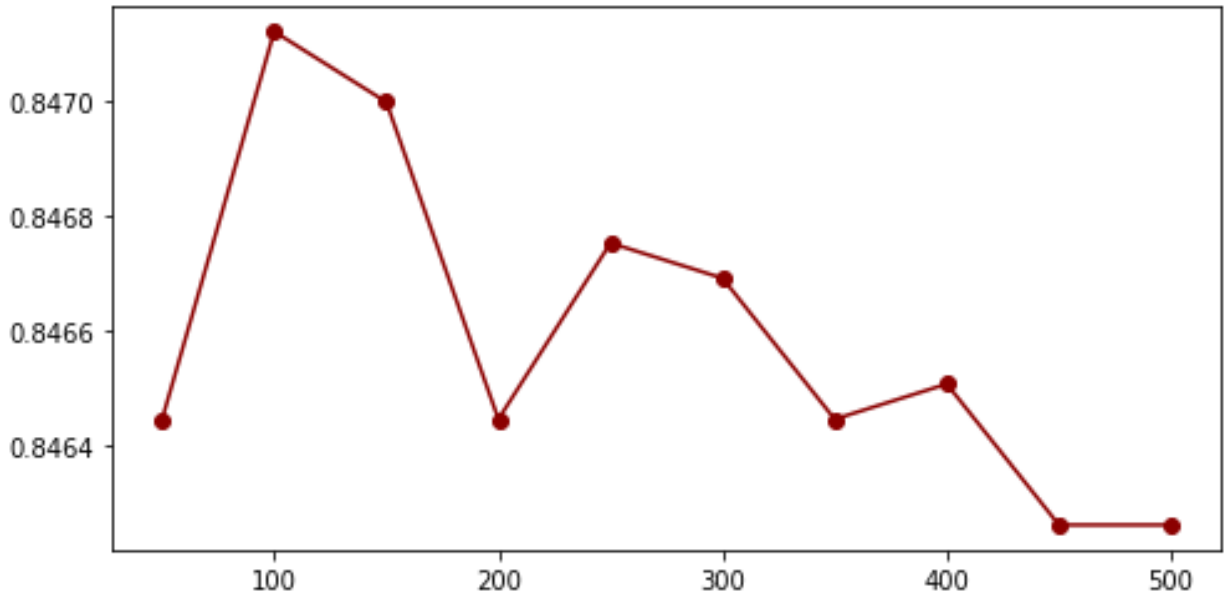
- The accuracy score is low at 50, and it goes up the maximum score at 100 and 150. Then, the score drops at 200, and it does not change anymore when the n_estimator goes up.
- The optimal n_estimators are 100 and 150 because they both have the highest accuracy score.

Gradient Boost



- The accuracy score keeps changing when the n_estimator changes.
- The optimal n_estimator is 350 because it has the highest accuracy score.

XGB Model



- The accuracy goes to the maximum at n_estimator 100, and then it drops. The score goes up and down, but it cannot hit the highest score like when n_estimator = 1000.
- The optimal n_estimator is 100.

5. Compare Performance

Run the comparison of 4 models with the same n_estimator.

When n_estimator = 100, Gradient Boost and XGB gives the highest accuracy score and AUC.

	Random Forest	AdaBoost	Gradient Boost	XGB
Accuracy	0.838216	0.844911	0.846631	0.847122
AUC	0.747542	0.746088	0.756732	0.756785
Hyper-parameters	100.000000	100.000000	100.000000	100.000000

When n_estimator = 50, Gradient Boost provides the highest accuracy score and AUC. The other n_estimators also show that Gradient Boost is the best model for this dataset classification.

	Random Forest	AdaBoost	Gradient Boost	XGB
Accuracy	0.838401	0.844727	0.847061	0.846447
AUC	0.748830	0.739053	0.755936	0.750865
Hyper-parameters	50.000000	50.000000	50.000000	50.000000