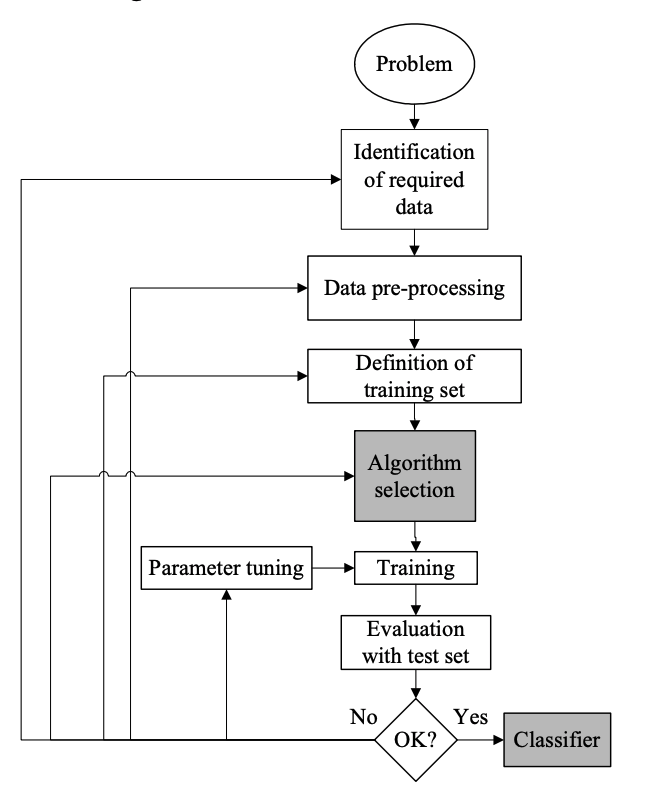
Classification:

The process of applying supervised machine learning to a real-world problem is described in Figure 1. The data processing part and feature selection part is mention are mentioned above. The crucial part for machine learning classifications is algorithm selection. There are many supervised classifiers. In this article, we will use some typical classifiers as well as some modern ensemble learning classifiers:



* Logistic Regression:

Logistic regression is a classification algorithm used to assign observations to a discrete set of classes. Instead of fitting a straight line or hyperplane, the logistic regression model uses the logistic function to squeeze the output of a linear equation between 0 and 1.

* Support Vector Machine:

Support vector machine (SVM) is a discriminative classifier formally defined by a separation hyperplane. That is, given labeled training data, SVM will output the optimal hyperplane that classifies the new instances. SVM can also perform a non-linear classification using kernel trick.

* Naive Bayes:

The Naive Bayes Classifier technique is based on the Bayesian theorem and is particularly suited when the dimensionality of the inputs is high.

* Ensemble methods: Ensemble methods

Ensemble methods are meta-algorithms that combine several machine learning techniques into one predictive model in order to decrease variance (bagging), bias (boosting), or improve predictions (stacking). In this article we will use AdaBoost, Gradient Boosting and Random Forest algorithms.

Then we will build classifiers for each ETF and class (y1 or y2). We say that training a classifier is actually finding suitable parameters for the supervised learning algorithm, which makes it perform best in the test set. We use AUC-ROC curve (Area Under the Receiver Operating Characteristics curve) as our evaluation metrics. It is a performance measurement for classification problem at various thresholds settings. Based on this evaluation metrics, we use cross-validation grid search to find the optimal parameters.

Below are AUC-ROC curves of each classifier for each ETF’s y1 and y2. For convenience, the time period of training set we use here is the whole time period, while actually we use a rolling window basis.