CROSS VALIDATION IN CLASSIFICATION ALGORITHM

WHAT ARE CLASSIFICATION ALGORITHMS?

Classification algorithms in machine learning use input training data to predict the likelihood that subsequent data will fall into one of the predetermined categories. Basically, classification is a form of pattern recognition with classification algorithms applied to the training data to find the same pattern. Some examples of classification algorithms are:

- Logistic Regression
- Naive Bayes
- K-Nearest Neighbors
- Decision Tree
- Support Vector Machines

NEED FOR EVALUATION

Let us consider, you are working on a spam emails data set, which contains 98% of spam emails and 2% of non-spam valid emails. In this case, even if you do not create any model but just classify every input as spam, you will be getting 0.98 accuracy.

Hence, whichever algorithm you have used to build your hypothesis function and train the machine learning model, you have to evaluate its performance before moving forward with it.

The easiest and fastest method to evaluate a model is to split the data set into training and testing set, train the model using the training set data and check its accuracy with the test data set.

CROSS VALIDATION AND IMPORTANCE

Cross validation is a technique for assessing how the statistical analysis generalises to an independent data set.It is a technique for evaluating machine learning models by training several models on subsets of the available input data and evaluating them on the complementary subset of the data.

The accuracy requirement of a machine learning model varies across requirement and problem statement. A final model should never be performed without evaluation of all essential metrics. Once evaluation is done, we must re-use the test data that was initially isolated for testing and train the model with the complete data we have so as to increase the chances for a better prediction.

PARTITIONING METHOD

WHAT IS PARTITIONING?

Partitioning method classifies the data into multiple groups based on the characteristics and similarity of the data. (The number of clusters(K) that has to be generated is specified by the user).

In this method, database(D) that contains multiple(N) objects then the partitioning method constructs user-specified(K) partitions of the data in which each partition represents a cluster.

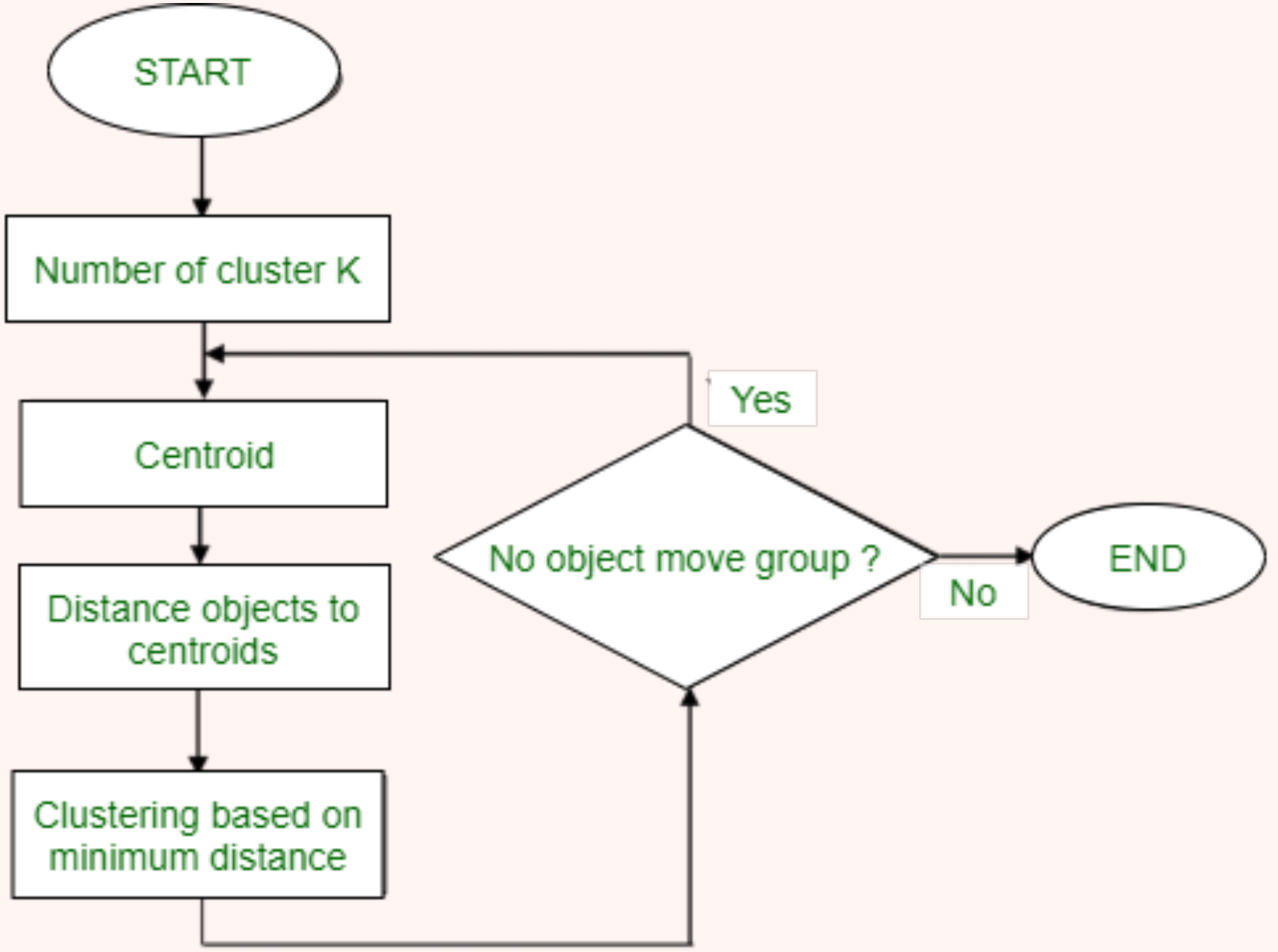
K-MEAN ALGORITHM

The K mean algorithm takes the input parameter K(number of clusters to be generated) from the user and partitions the dataset objects into K clusters so that resulting similarity among the data objects inside the group is high but the similarity of data objects with the data objects from outside the cluster is low. The similarity of the cluster is determined with respect to the mean value of the cluster.

Randomly K objects from the dataset are chosen in which each of the objects represents a cluster mean(centre). For the rest of the data objects, they are assigned to the nearest cluster based on their distance from the cluster mean. The new mean of each of the cluster is then calculated with the added data objects.

METHOD

1. Randomly assign K objects from the dataset(D) as cluster centres(C)
2. (Re) Assign each object to which object is most similar based upon mean values.
3. Update Cluster means, i.e., Recalculate the mean of each cluster with the updated values.
4. Repeat until no change occurs.


START
Number of cluster K
Centroid
Distance objects to centroids
Clustering based on minimum distance
No object move group ?
Yes
No
END

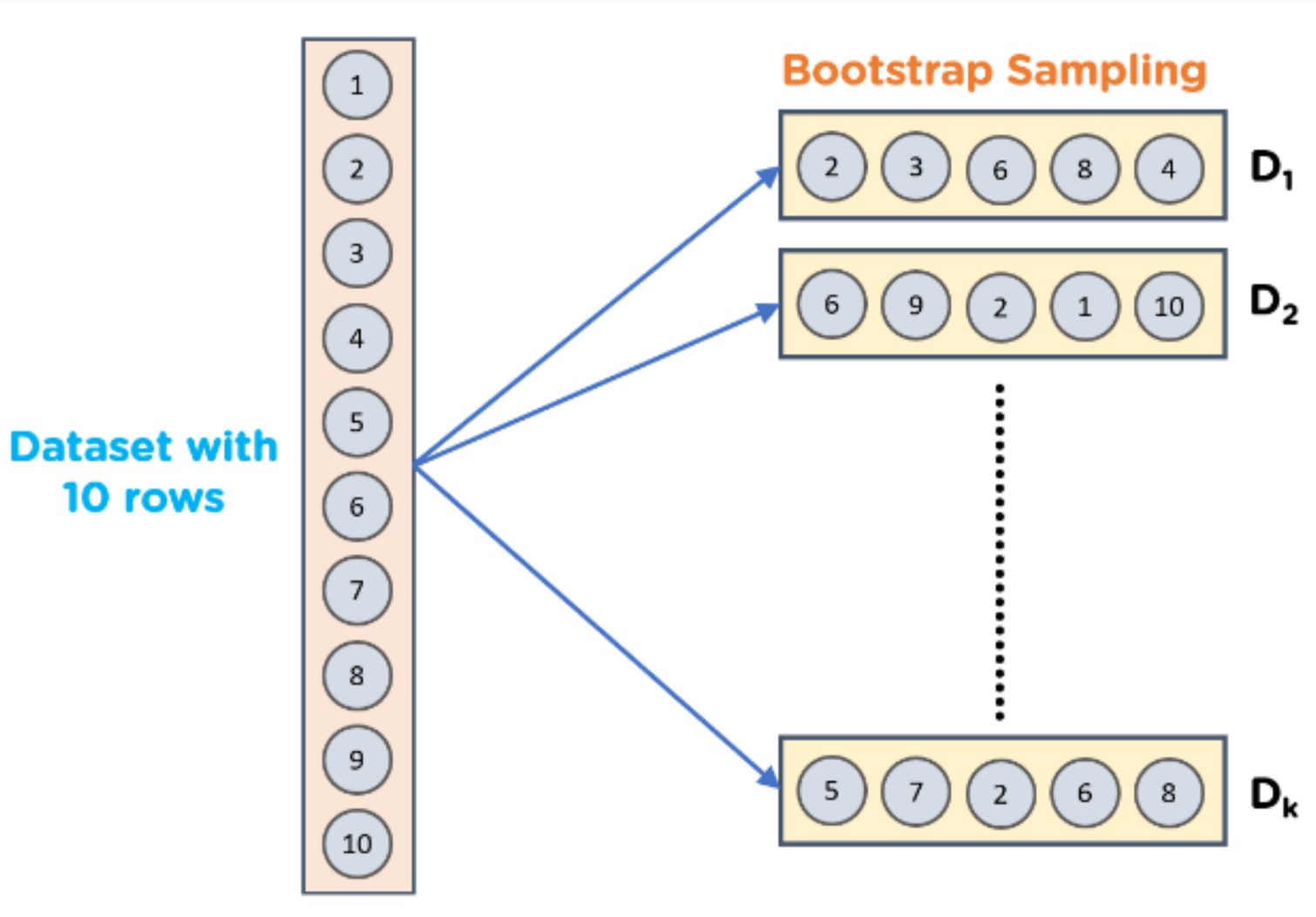
BAGGING

WHAT IS BAGGING?

- Bagging (also known as Bootstrap aggregating) is an ensemble learning technique that helps to improve the performance and accuracy of machine learning algorithms.
- It is used to reduce the variance of a prediction model.
- Avoids overfitting of data.

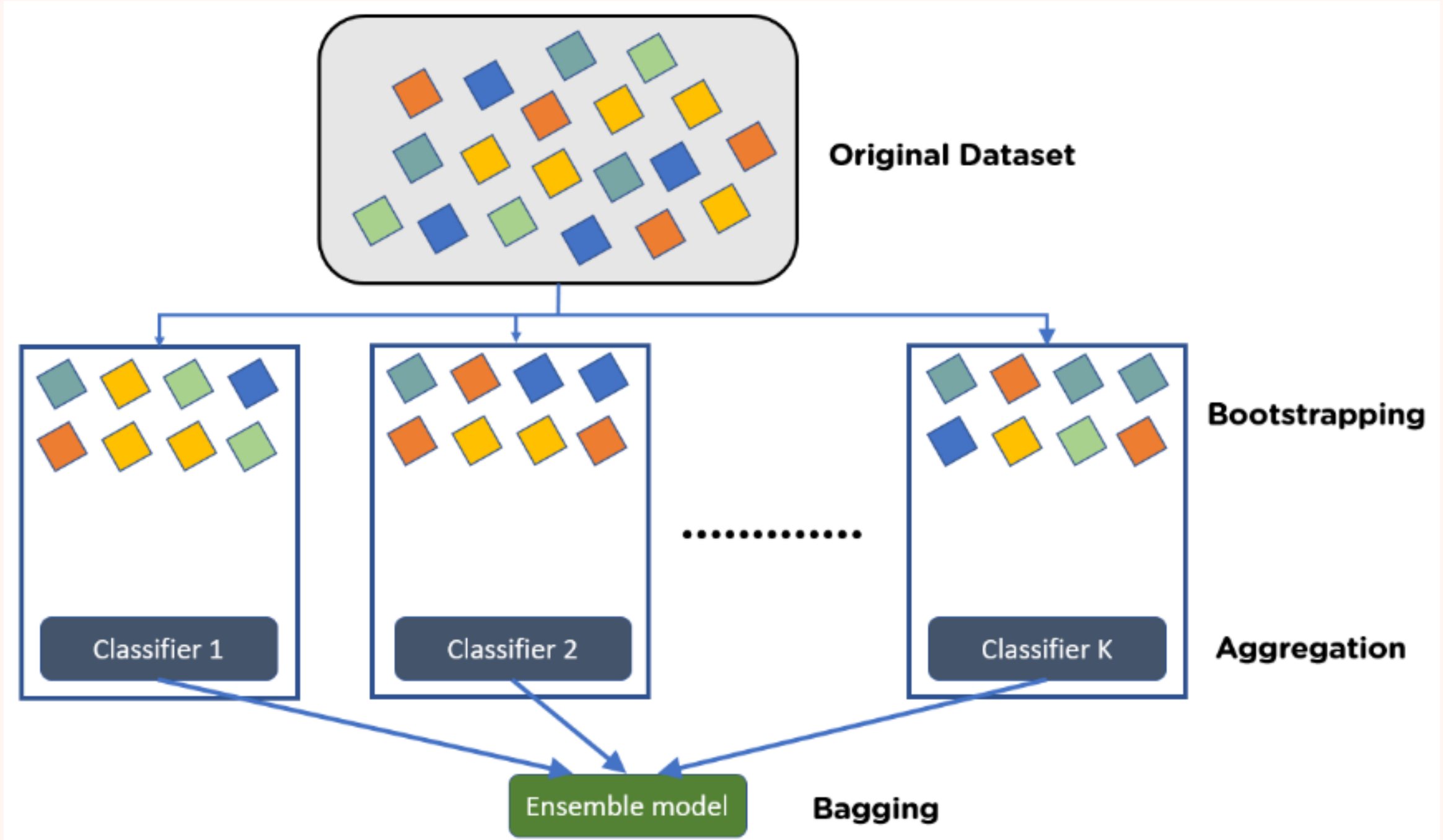
WHAT IS BOOTSTRAPPING?

Bootstrapping is the method of randomly creating samples of data out of a population with replacement to estimate a population parameter.



HOW IS IT DONE?

1. Consider there are n observations and m features in the training set. Select a random sample from the training dataset.
2. A subset of m features is chosen randomly and a model is trained using sample observations.
3. The above step is repeated n times on different models.
4. The aggregate of the outputs of individual models is taken to form a ensemble classifier to give the best prediction.
5. This ensemble classifier has better accuracy and less error rate.


Original Dataset
Bootstrapping
Classifier 1
Classifier 2
Classifier K
Aggregation
Ensemble model
Bagging

ADVANTAGES

- Minimises the overfitting of data.
- It improves the model’s accuracy.
- It deals with higher dimensional data efficiently.

THANK YOU!