In [1]:
```
import numpy as np
import pandas as pd
from sklearn.datasets import fetch_openml
from sklearn.model_selection import train_test_split
from sklearn.ensemble import RandomForestClassifier
from sklearn.ensemble import BaggingClassifier
from sklearn.tree import DecisionTreeClassifier
from sklearn.ensemble import VotingClassifier
from sklearn.metrics import accuracy_score

mnist = fetch_openml('mnist_784', version=1)
```

Q.no. 1

In [11]:
```
X = mnist['data']
y = mnist['target']
X_train, X_test_val, y_train, y_test_val = train_test_split(X, y, test_size = 0.2
X_val, X_test, y_val, y_test = train_test_split(X_test_val, y_test_val, test_size
print('The number of instances in train set:', X_train.shape[0])
print('The number of instances in validation set:', X_val.shape[0])
print('The number of instances in test set:', X_test.shape[0])
```

```
The number of instances in train set: 50000
The number of instances in validation set: 10000
The number of instances in test set: 10000
```

Q.no. 2

In [12]:
```
rnd_clf = RandomForestClassifier(n_estimators=100, random_state=42)
rnd_clf.fit(X_train, y_train)
y_pred_rf = rnd_clf.predict(X_val)
print("Random Forest Accuracy: ", accuracy_score(y_val, y_pred_rf))
```

```
Random Forest Accuracy:  0.9677
```

In [13]:
```
bag_clf = BaggingClassifier(
    DecisionTreeClassifier(random_state=42), n_estimators=500,
    max_samples=100, bootstrap=True, random_state=42)
bag_clf.fit(X_train, y_train)
y_pred_bag = bag_clf.predict(X_val)
print("Bagging Classifier Accuracy: ", accuracy_score(y_val, y_pred_bag))
```

```
Bagging Classifier Accuracy:  0.8394
```

In [14]:
```
tree_clf = DecisionTreeClassifier(random_state=42)
tree_clf.fit(X_train, y_train)
y_pred_tree = tree_clf.predict(X_val)
print("Decision Trees Accuracy: ", accuracy_score(y_val, y_pred_tree))
```

```
Decision Trees Accuracy:  0.8714
```

Q.no. 3

In [15]:
```
voting_clf = VotingClassifier(
    estimators=[('bag', bag_clf), ('rf', rnd_clf), ('dt', tree_clf)],
    voting='hard') # hard voting
voting_clf.fit(X_train, y_train)
y_pred_voting = voting_clf.predict(X_val)
print("Voting Classifier Accuracy: ", accuracy_score(y_val, y_pred_voting))
```

```
Voting Classifier Accuracy:  0.9334
```

Q.no. 4

Since the output class is balanced, accuracy is used to measure the efficiency of the models. While the ensemble outperforms Bagging(0.84) and Decision Trees Classifier(0.87), it fails to perform better than the Random Forest Classifier as it obtained an accuracy of 0.93 on the validation set which is less than that of the Random Forest (0.97) on the same set.

Q.no. 5 and 6

In [16]:
```
rnd_clf = RandomForestClassifier(n_estimators=100, random_state=42)
rnd_clf.fit(X_train, y_train)
y_pred_rf = rnd_clf.predict(X_test)
print("Random Forest Accuracy: ", accuracy_score(y_test, y_pred_rf))

bag_clf = BaggingClassifier(
    DecisionTreeClassifier(random_state=42), n_estimators=500,
    max_samples=100, bootstrap=True, random_state=42)
bag_clf.fit(X_train, y_train)
y_pred_bag = bag_clf.predict(X_test)
print("Bagging Classifier Accuracy: ", accuracy_score(y_test, y_pred_bag))

tree_clf = DecisionTreeClassifier(random_state=42)
tree_clf.fit(X_train, y_train)
y_pred_tree = tree_clf.predict(X_test)
print("Decision Trees Accuracy: ", accuracy_score(y_test, y_pred_tree))

voting_clf_updated = VotingClassifier(
    estimators=[('rf', rnd_clf), ('dt', tree_clf)],
    voting='hard') # hard voting
voting_clf_updated.fit(X_train, y_train)
y_pred_voting = voting_clf_updated.predict(X_test)
print("Voting Classifier(Revised) Accuracy: ", accuracy_score(y_test, y_pred_voti
```

```
Random Forest Accuracy:  0.9672
Bagging Classifier Accuracy:  0.846
Decision Trees Accuracy:  0.8692
Voting Classifier(Revised) Accuracy:  0.9188
```

Q.no. 7

The Random Forest, Bagging Classifier, and Decision Trees achieved a accuracy of 0.97, 0.85, 0.87 respecitively while the Voting Classifier achieved an accuracy of 0.92. The Voting Classifier exceeded the accuracy of Bagging Classifier and Decision Trees by 0.7 (7%) and 0.5 (5%) respectively on the test set. However, its accuracy is still outperformed by the Random Forest Classifier by 0.5 (5%) in the same set. Since the classifiers used in the voting algorithm are all tree-based classifiers, their errors are not independent hence the voting classifier was not able to outperform all the individual classifiers.