SNS COLLEGE OF TECHNOLOGY

Coimbatore-35
An Autonomous Institution

Accredited by NBA – AICTE and Accredited by NAAC – UGC with 'A++' Grade
Approved by AICTE, New Delhi & Affiliated to Anna University, Chennai

Redesigning Common Mind & Business Towards Excellence

3P urpose rocess eople Culture

sns INSTITUTIONS www.snsgroups.com

1st GenAI Powered Design Thinking FrameWork

Build an Entrepreneurial Mindset Through Our Design Thinking FrameWork

DEPARTMENT OF ARTIFICIAL INTELLIGENCE AND MACHINE LEARNING

23AMB201 - MACHINE LEARNING

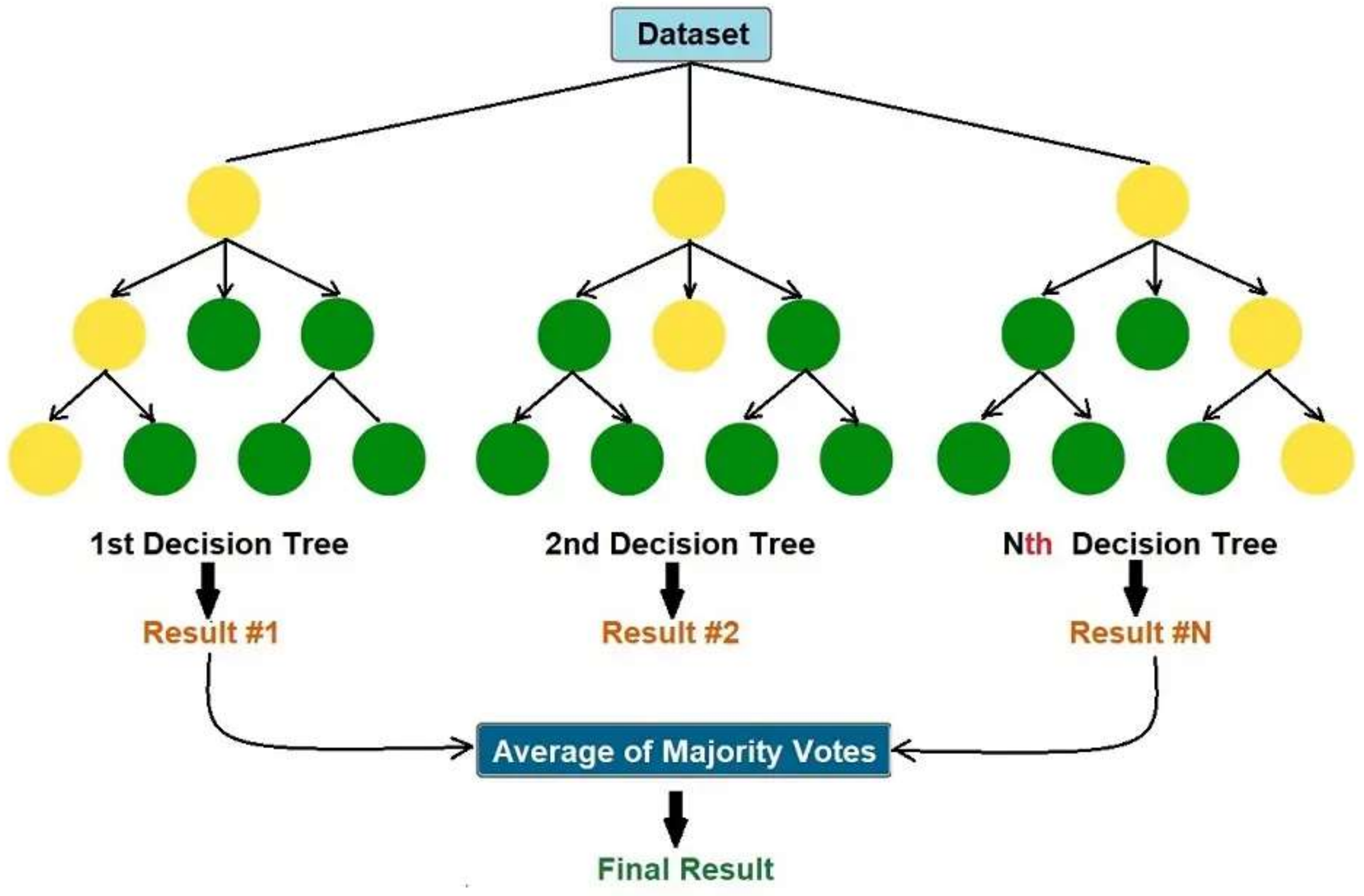
II YEAR IV SEM

UNIT III – GENERATIVE MODELS AND BOOSTING

TOPIC 18, 19, 20 – Random Forest, Ensemble learning and Boosting

A Random Forest Algorithm is a supervised machine learning algorithm that is extremely popular and is used for Classification and Regression problems in Machine Learning.

Random Forest is a classifier that contains several decision trees on various subsets of the given dataset and takes the average to improve the predictive accuracy of that dataset.



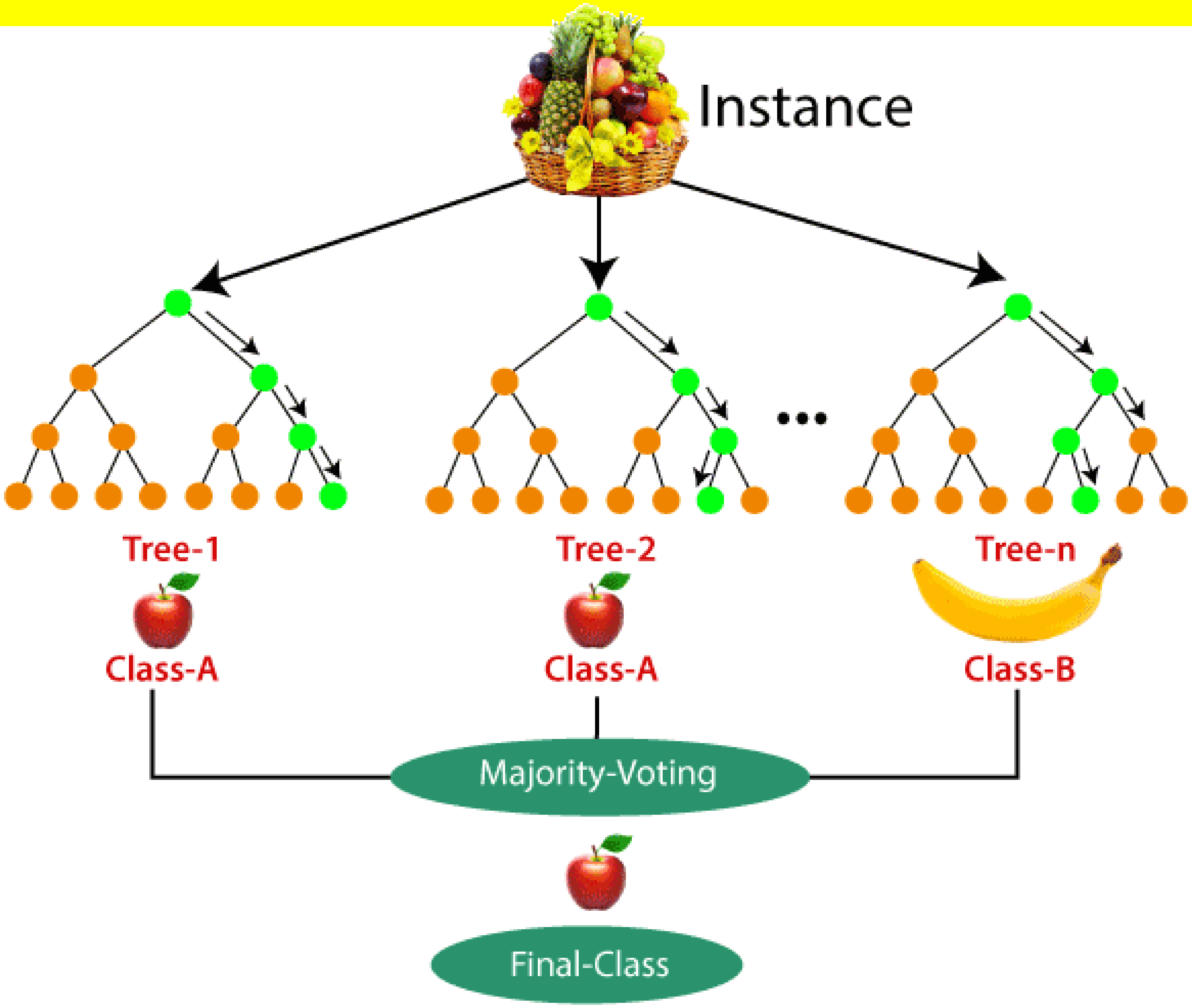
Why Random Forest?

1. Reduced risk of overfitting
2. Provides flexibility
3. Easy to determine feature importance

How does Random Forest algorithm work?

1. Step-1: Select random K data points from the training set.
2. Step-2: Build the decision trees associated with the selected data points (Subsets).
3. Step-3: Choose the number N for decision trees that you want to build.
4. Step-4: Repeat Step 1 & 2.
5. Step-5: For new data points, find the predictions of each decision tree, and assign the new data points to the category that wins the majority votes.

Example

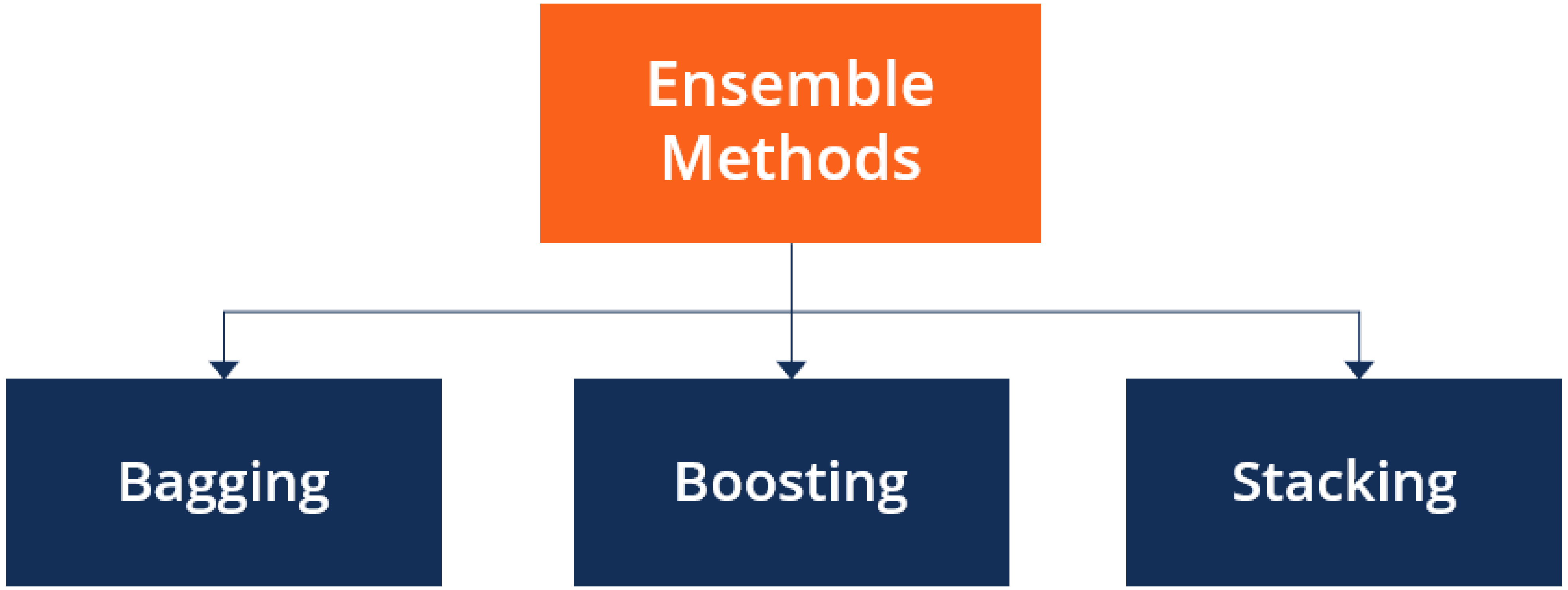


Instance
...
Tree-1
Class-A
Tree-2
Class-A
Tree-n
Class-B
Majority-Voting
Final-Class

Program

```
from sklearn.ensemble import RandomForestClassifier
from sklearn.model_selection import train_test_split
from sklearn.metrics import accuracy_score

# Load dataset
iris = load_iris()
X_train, X_test, y_train, y_test = train_test_split(iris.data, iris.target, test_size=0.2)

# Train Random Forest
clf = RandomForestClassifier(n_estimators=100, random_state=42)
clf.fit(X_train, y_train)

# Predictions
y_pred = clf.predict(X_test)

# Accuracy
print("Accuracy:", accuracy_score(y_test, y_pred))
```

Ensemble Learning: Combine the decisions from multiple models to improve the overall performance.



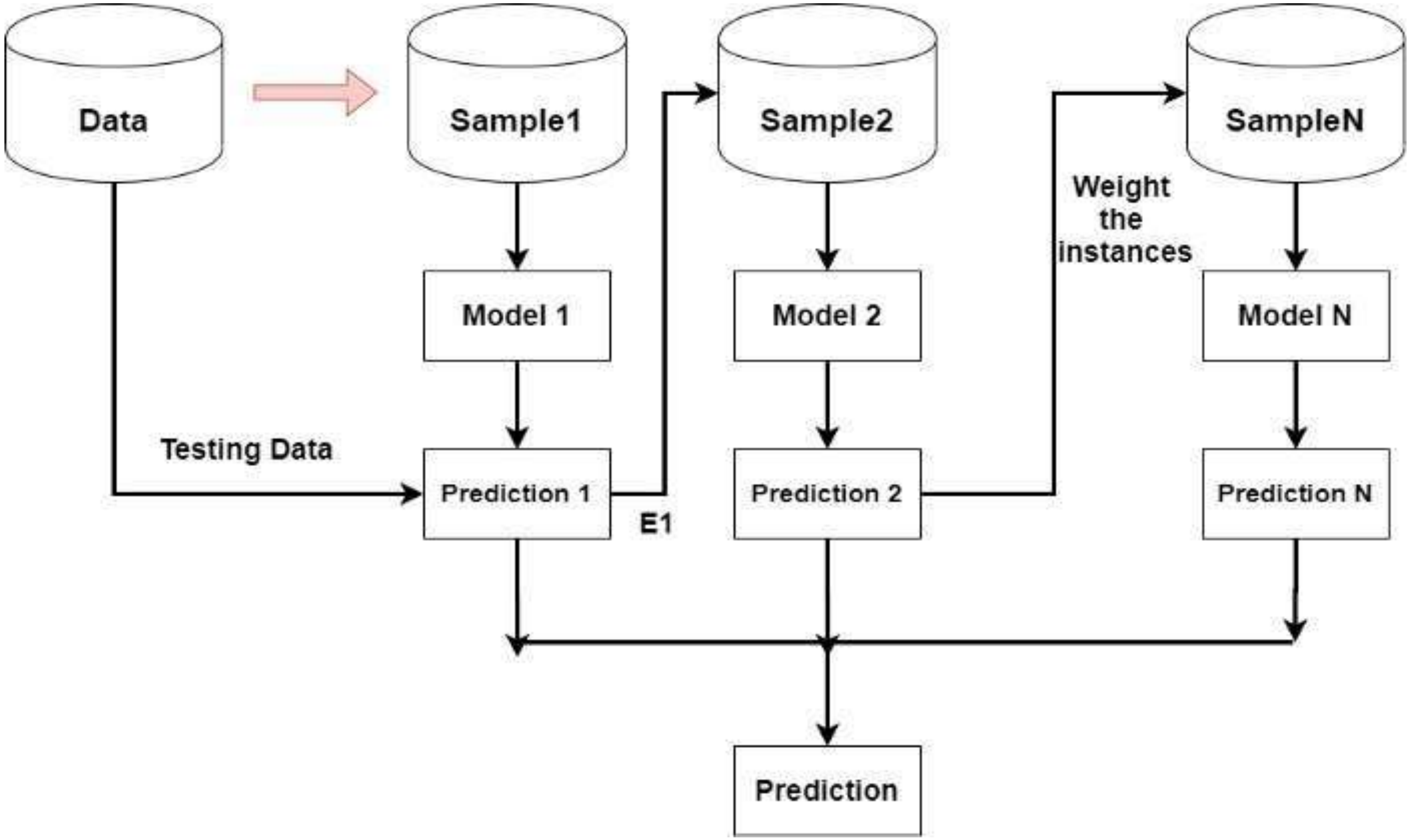
Ensemble

1. An ensemble is itself a supervised learning algorithm because it can be trained and then used to make predictions.
2. It combine several decision trees classifiers to produce better predictive performance than a single decision tree classifier.
3. Main principle is to group of **weak learners come together to form a strong learner**
4. To increasing the accuracy of the model.
5. Ensemble helps to reduce noise, variance and bias

Total error can be expressed as follows:

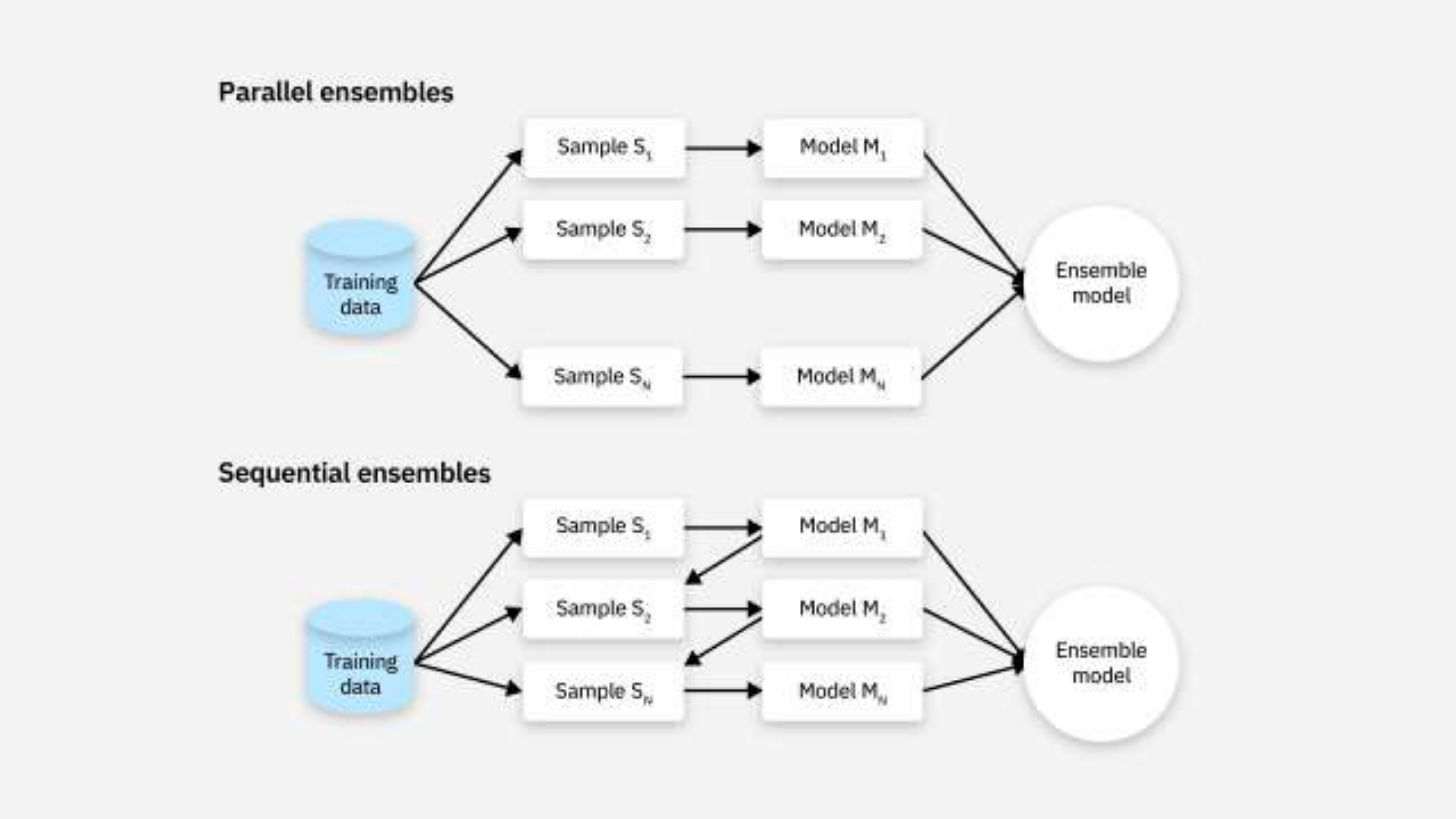
Total Error = Bias + Variance + Irreducible Error

Ensemble



Data
Sample1
Sample2
SampleN
Weight the instances
Model 1
Model 2
Model N
Testing Data
Prediction 1
E1
Prediction 2
Prediction N
Prediction

Types of Ensemble



Basic Ensemble Techniques and Methods

1. Max Voting
2. Averaging
3. Weighted Average

1. Bootstrap
2. Bagging
3. Stocking

Max Voting

PassengerId	Survived	Pclass	Name	Sex	Age	SibSp	Parch
1	0	3	Braund, Mr. Owen Harris	male	22.0	1	0
2	1	1	Cumings, Mrs. John Bradley (Florence Briggs Th...	female	38.0	1	0
3	1	3	Heikkinen, Miss. Laina	female	26.0	0	0
4	1	1	Futrelle, Mrs. Jacques Heath (Lily May Peel)	female	35.0	1	0
5	0	3	Allen, Mr. William Henry	male	35.0	0	0

M1 M2 M3

Pred 1	Pred 2	Pred 3	Final Pred
0	1	0	?
0	1	1	?
1	1	0	?
1	0	1	?
0	0	0	?

PassengerId	Survived
1	0
2	1
3	1
4	1
5	0

M1 M2 M3

Pred 1	Pred 2	Pred 3
0	1	0
0	1	1
1	1	0
1	0	1
0	0	0

M1: logistic regression
M2: KNN
M3: SVM

M1 M2 M3 Vote

Pred 1	Pred 2	Pred 3	Final Pred
0	1	0	0
0	1	1	1
1	1	0	1
1	0	0	**0**
0	0	0	0

Averaging

	Item_Outlet_Sales	M1 Predicted Values 1	M2 Predicted Values 2	M3 Predicted Values 3
0	3735.1380	3900	3000	3500
1	443.4228	390	340	500
2	2097.2700	2000	1900	2600
3	732.3800	700	600	750
4	994.7052	950	800	1060

M1 Predicted Values 1	M2 Predicted Values 2	M3 Predicted Values 3	Average
3900	3000	3500	3466.66
390	340	500	410.00
2000	1900	2600	2166.66
700	600	750	683.33
950	800	1060	936.66

M1: logistic regression
M2: KNN
M3: SVM

Weighted Average

M1	0.6	2
M2	0.4	1
M3	0.6	2

2

ID	Actual Values	Predicted Values 1	Predicted Values 2	Predicted Values 3
0	3735.13	7800	3000	3500
1	443.422	780	340	500
2	2097.27	4000	1900	2600
3	732.380	1400	600	750
4	994.705	1900	800	1060

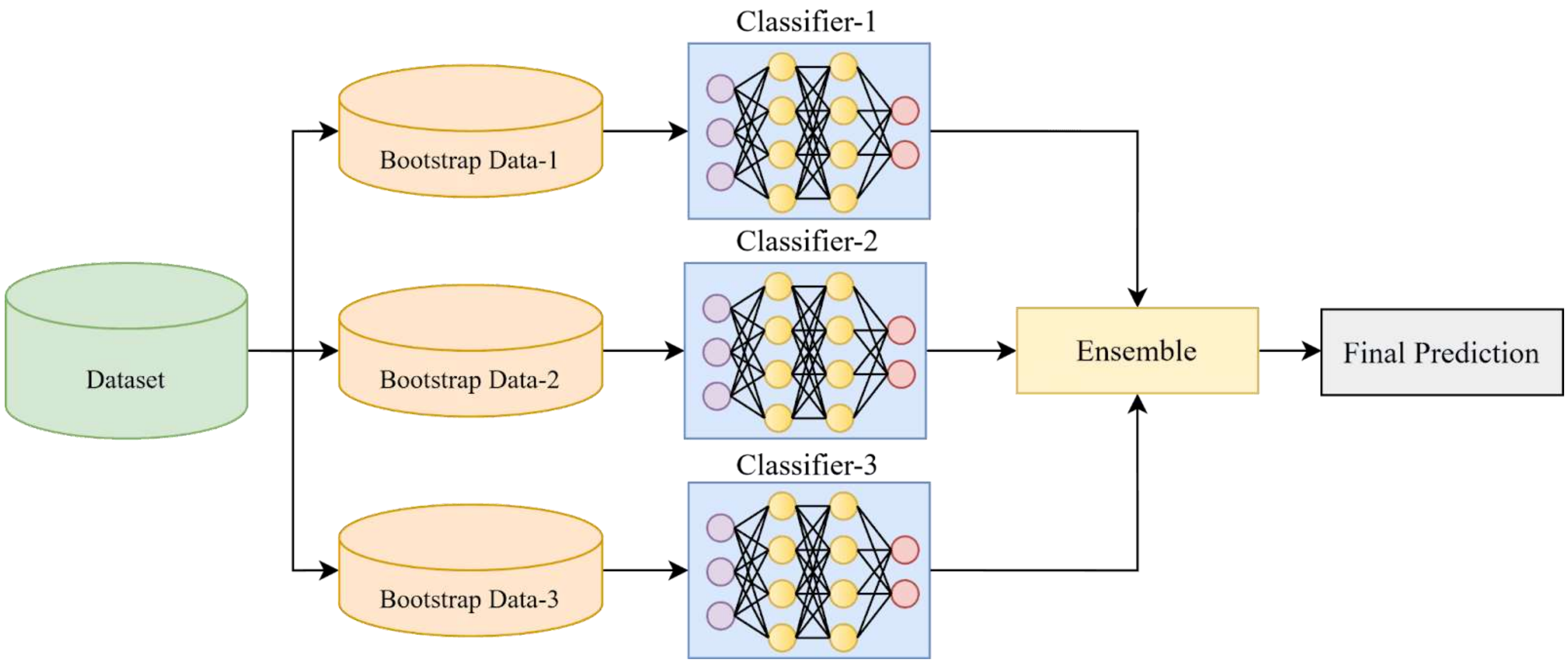
2

ID	Actual Values	Predicted Values 1	Predicted Values 2	Predicted Values 3
0	3735.13	7800	3000	7000
1	443.422	780	340	1000
2	2097.27	4000	1900	5200
3	732.380	1400	600	1500
4	994.705	1900	800	2120

ID	Actual Values	Predicted Values 1	Predicted Values 2	Predicted Values 3	Average
0	3735.13	7800	3000	7000	3560
1	443.422	780	340	1000	424
2	2097.27	4000	1900	5200	2220
3	732.380	1400	600	1500	700
4	994.705	1900	800	2120	964

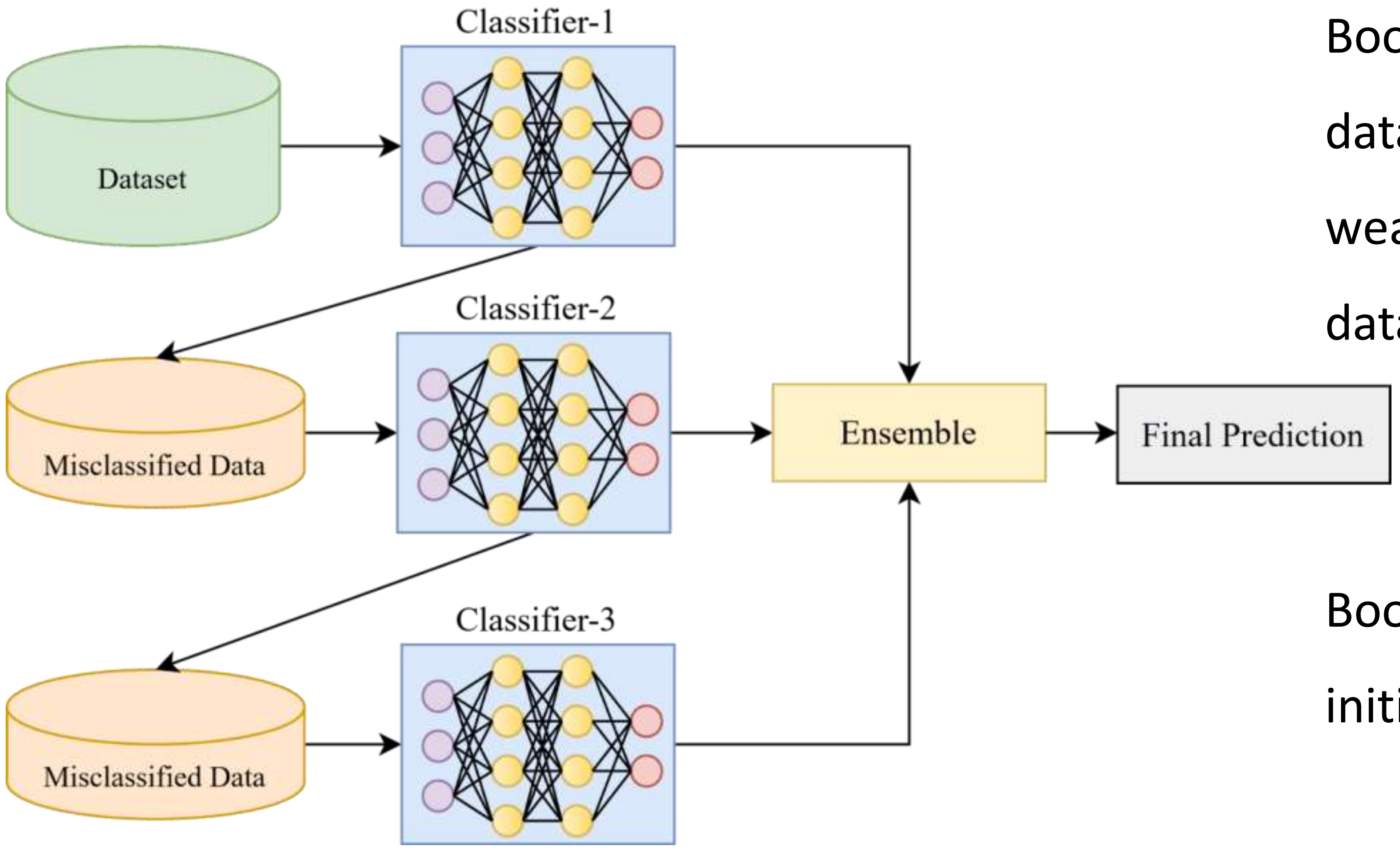
M1: logistic regression
M2: KNN
M3: SVM

Ensemble Learning Methods – Bagging



Subsamples from a dataset are created and they are called "bootstrap sampling." To put it simply, random subsets of a dataset are created using replacement, meaning that the same data point may be present in several subsets.

Ensemble Learning Methods - Boosting

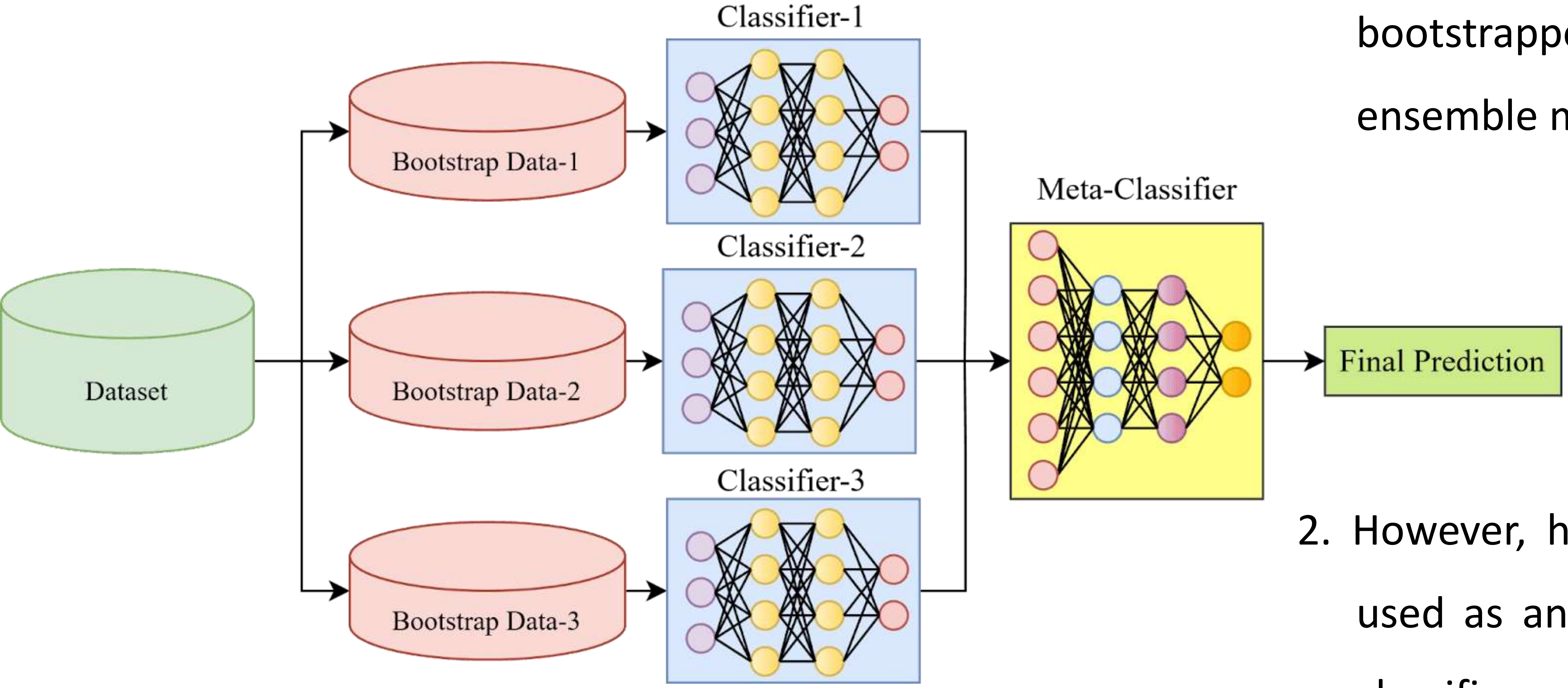


Boosting trains a learner on some initial dataset, *d*. The resultant learner is typically weak, misclassifying many samples in the dataset.

Boosting then samples instances from the initial dataset to create a new dataset (d_2)

Boosting prioritizes misclassified data instances from the first model or learner. A new learner is trained on this new dataset d_2.

Ensemble Learning Methods - Stacking



1. The stacking ensemble method also involves creating bootstrapped data subsets, like the bagging ensemble mechanism for training multiple models.

2. However, here, the outputs of all such models are used as an input to another classifier, called meta-classifier, which finally predicts the samples.

References

Y. S. Abu-Mostafa, M. Magdon-Ismail, and H.-T. Lin, ―Learning from Data, AML Book Publishers, 2012.

P. Flach, ―Machine Learning: The art and science of algorithms that make sense of data‖, Cambridge University Press, 2012.

https://sefiks.com/2018/08/27/a-step-by-step-Random Forest-decision-tree-example/#google_vignette

https://medium.com/@singhakshay.etw69/decision-tree-algorithm-Random Forest-e61032794927