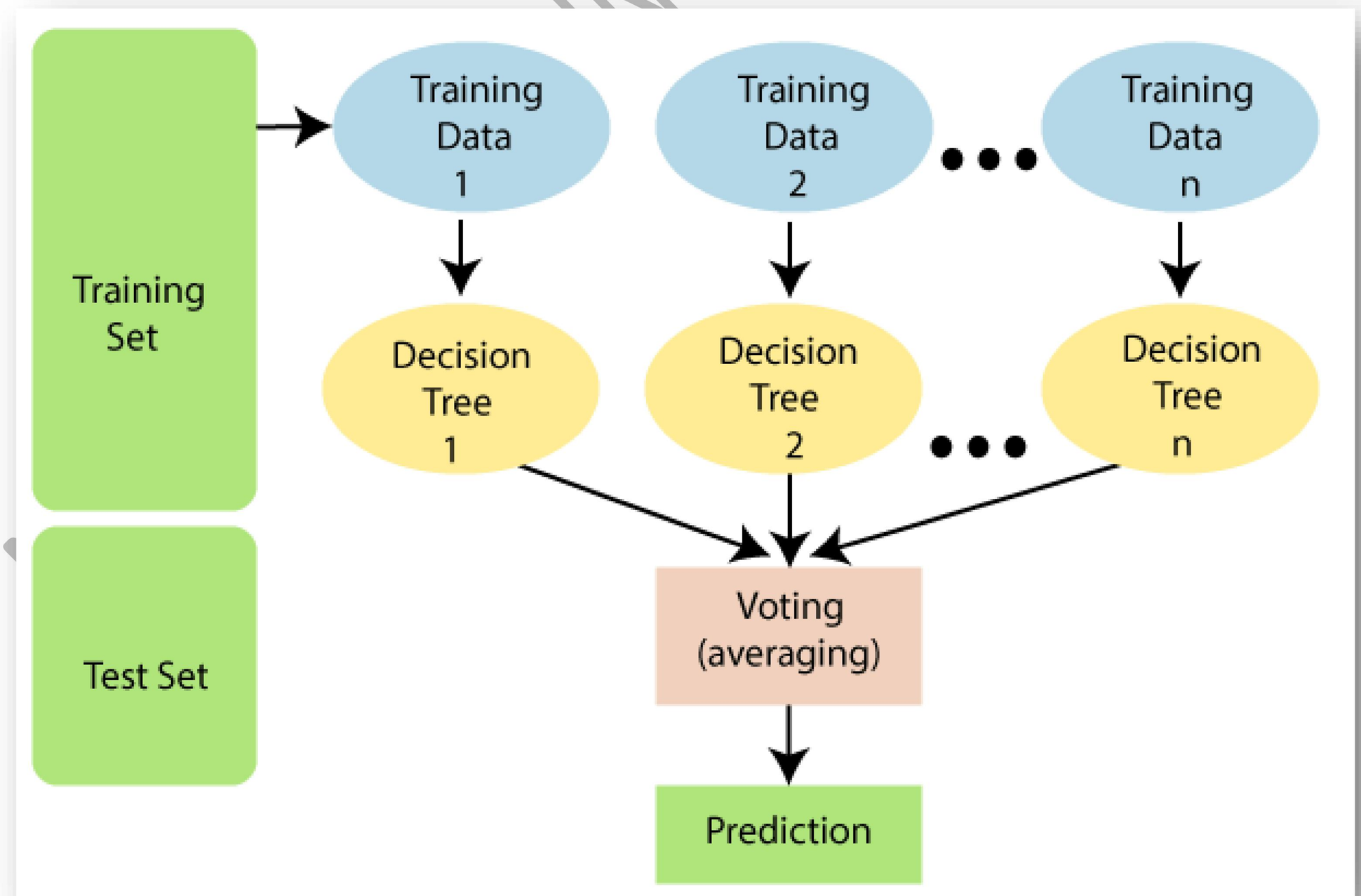
Chapter 10: Random Forest Algorithm

Random Forest Algorithm:

- Random forest is a supervised learning algorithm.
- It has two variations –
  - one is used for classification problems and
  - other is used for regression problems.
- Random forest algorithm combines multiple decision-trees, resulting in a forest of trees, hence the name Random Forest.

Random Forest Classifier:

- It is based on concept of ensemble learning, which is a process of combining multiple classifiers to improve the performance of the model.
- Random forest combines multiple decision trees to predict the class of the dataset.
- The greater number of trees in the forest leads to higher accuracy.



Bagging:

- The objective of bagging is to create several subsets of data from training sample chosen randomly with replacement. Each collection of subset data is used to train their decision trees.
- As a result, we get an ensemble of different models. Average of all the predictions from different trees are used which is more robust than a single decision tree classifier.

For Classification: most occurring class would be selected
For Regression: mean of the output would be the result

Feature Importance:

- Feature importance is calculated as the decrease in node impurity weighted by the probability of reaching that node.
- The node probability can be calculated by the number of samples that reach the node, divided by the total number of samples.
- *The higher the value the more important the feature.*

How to calculate it?

- Scikit-learn calculates a nodes importance using Gini Importance

$$ni_j = w_j C_j - w_{left(j)} C_{left(j)} - w_{right(j)} C_{right(j)}$$

- ✓ ni_j = the importance of node j
- ✓ w_j = weighted number of samples reaching node j
- ✓ C_j = the impurity value of node j
- ✓ left(j) = child node from left split on node j
- ✓ right(j) = child node from right split on node j

- The importance for each feature on a decision tree is then calculated as:
  - ✓ fi_i= the importance of feature i
  - ✓ ni_i= the importance of node j

$$fi_i = \frac{\sum_{j:node\ j\ splits\ on\ feature\ i} ni_j}{\sum_{k \in all\ nodes} ni_k}$$

- These can then be normalized to a value between 0 and 1 by dividing by the sum of all feature importance values:

$$normfi_i = \frac{fi_i}{\sum_{j \in all\ features} fi_j}$$

- The final feature importance, at the Random Forest level, is it's average over all the trees. The sum of the feature's importance value on each trees is calculated and divided by the total number of trees:

$$RFfi_i = \frac{\sum_{j \in all\ trees} normfi_{ij}}{T}$$

- ✓ $RFfi_i$= the importance of feature i calculated from all trees in the Random Forest model
- ✓ $normfi_{ij}$= the normalized feature importance for i in tree j
- ✓ T = total number of trees

Advantages of Random Forest:

- ❑ Random forest algorithm can be used to solve both classification and regression problems.
- ❑ It is considered as very accurate and robust model because it uses large number of decision-trees to make predictions.
- ❑ Random forests takes the average of all the predictions made by the decision-trees, which cancels out the biases. So, it does not suffer from the Overfitting problem.

Disadvantages of Random Forest:

- ❑ The biggest disadvantage of random forests is its computational complexity. Random forests is very slow in making predictions because large number of decision-trees are used to make predictions. All the trees in the forest have to make a prediction for the same input and then perform voting on it. So, it is a time-consuming process.
- ❑ The model is difficult to interpret as compared to a decision-tree, where we can easily make a prediction as compared to a decision-tree.

Application of Random Forest Classifier:

- ✓ **Banking:** Banking sector mostly uses this algorithm for the identification of loan risk.
- ✓ **Medicine:** With the help of this algorithm, disease trends and risks of the disease can be identified.
- ✓ **Land Use:** We can identify the areas of similar land use by this algorithm.
- ✓ **Marketing:** Marketing trends can be identified using this algorithm.