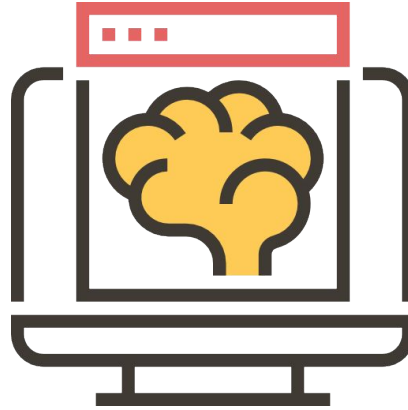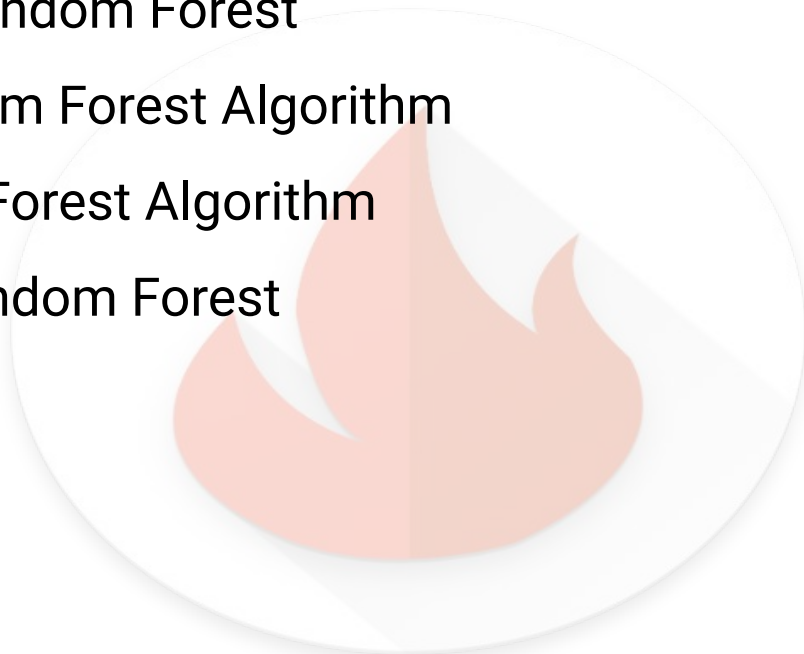# Supervised Learning Classification
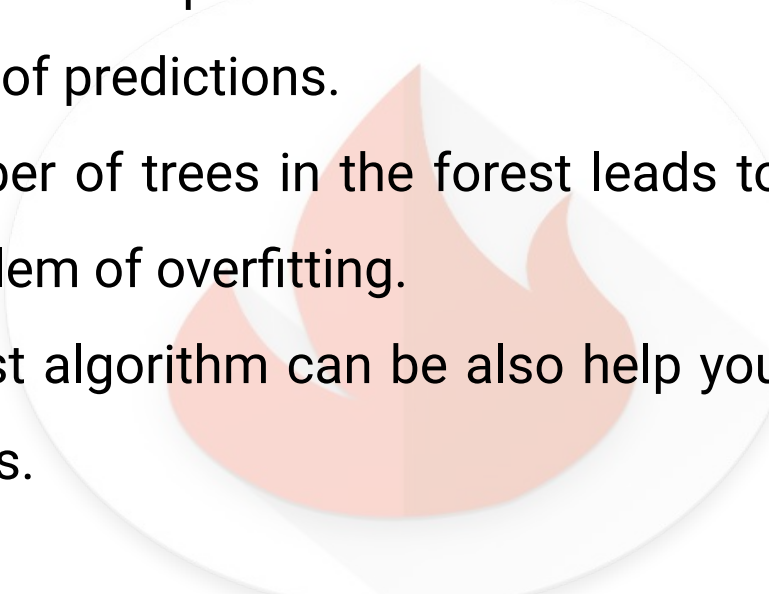
Random Forest

Agenda

- Introduction to Random Forest

- Working of Random Forest Algorithm

- Uses of Random Forest Algorithm

- Application of Random Forest

- Advantages

- Disadvantages

Introduction to Random Forest

1. Tree-based supervised learning algorithms.

2. The algorithm can be used to solve both classification and regression problems.

3. It is based on ensemble learning which is process of combining multiple classifiers to solve a complex problem.

4. Random forest is a collection of decision tree on various subsets of the given dataset and take average to improve the predictive accuracy of that dataset.

Introduction to Random Forest

1. Random forest takes the predictions from each decision tree and based on majority votes of predictions.

2. The greater number of trees in the forest leads to higher accuracy and prevents the problem of overfitting.

3. The random forest algorithm can be also help you find best features or important features.

Working of Random Forest Algorithm

Step 1: The algorithm select random samples from dataset provided (K).

Step 2: The algorithm will create a decision tree for each sample selected. Then will get a prediction result from decision tree created.

Step 3: Voting will then performed of every predicted result.
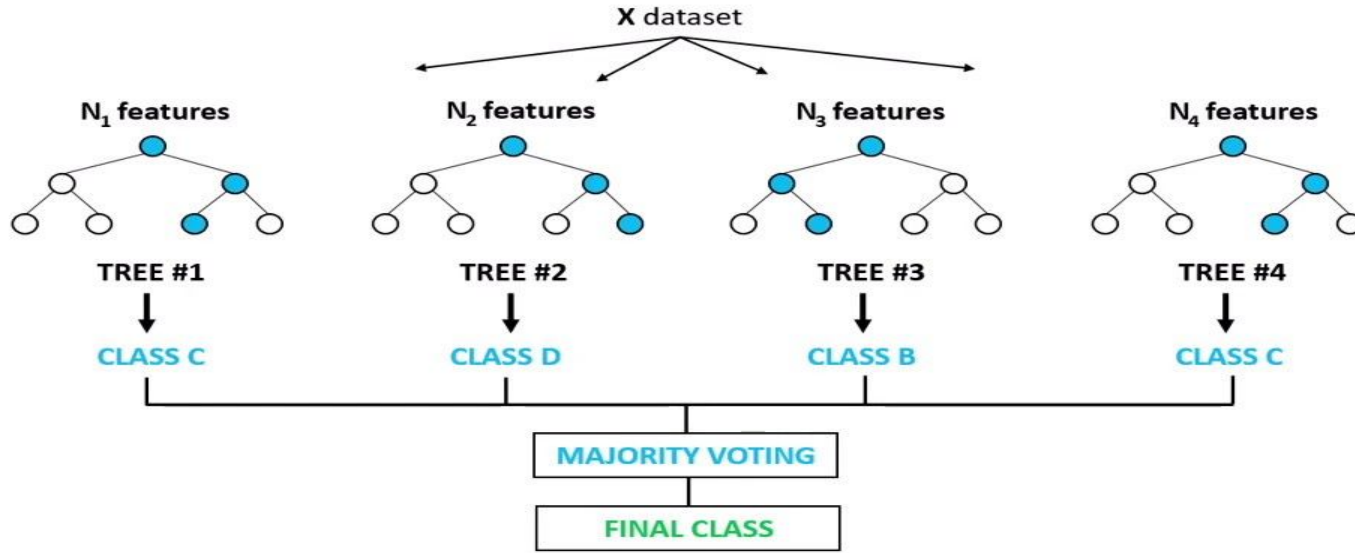
Step 4: For new data points, find the predictions of each decision tree, and assign the new data points to the category that wins the majority votes.

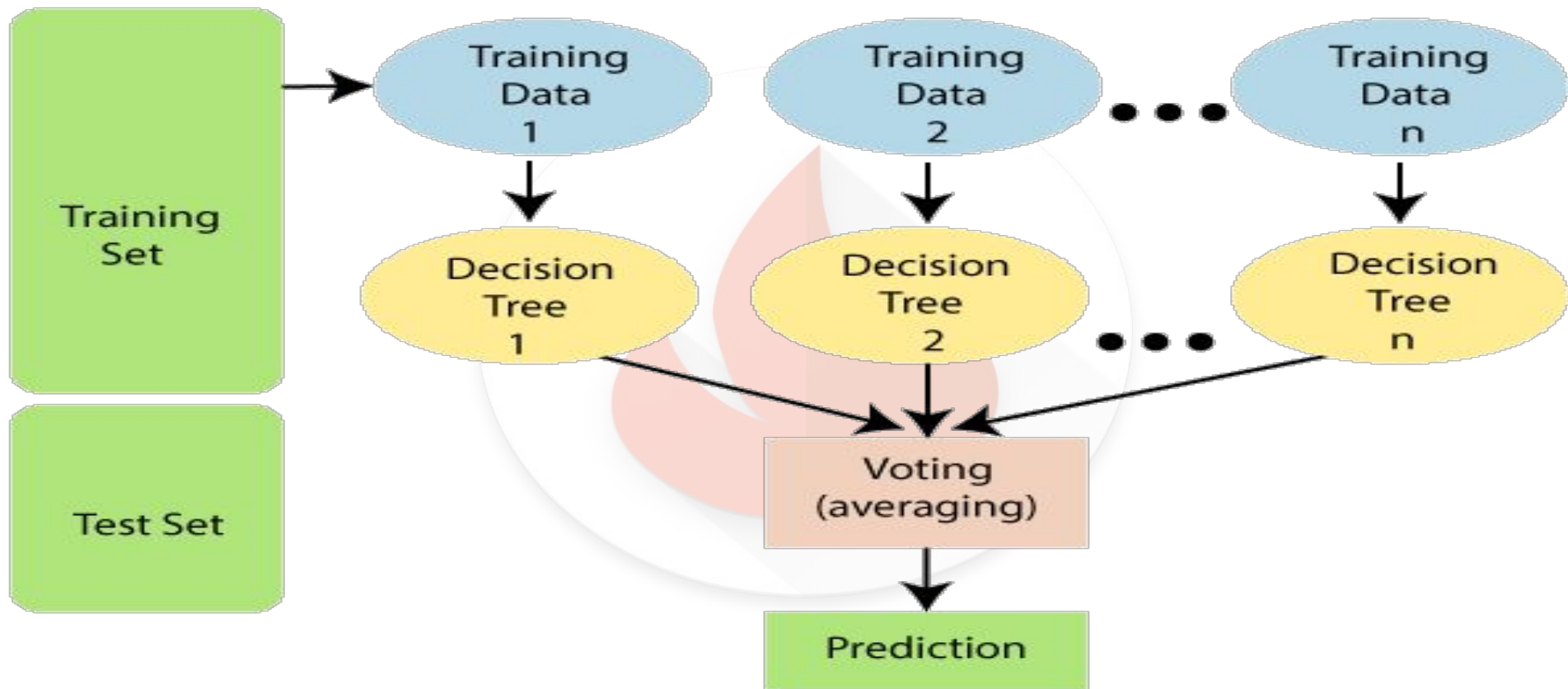**Step 5:** And finally, will selected the most voted prediction result.

Step 6: For classification problem, use **'mode'**, and for regression **'mean'**.

Random Forest Representation
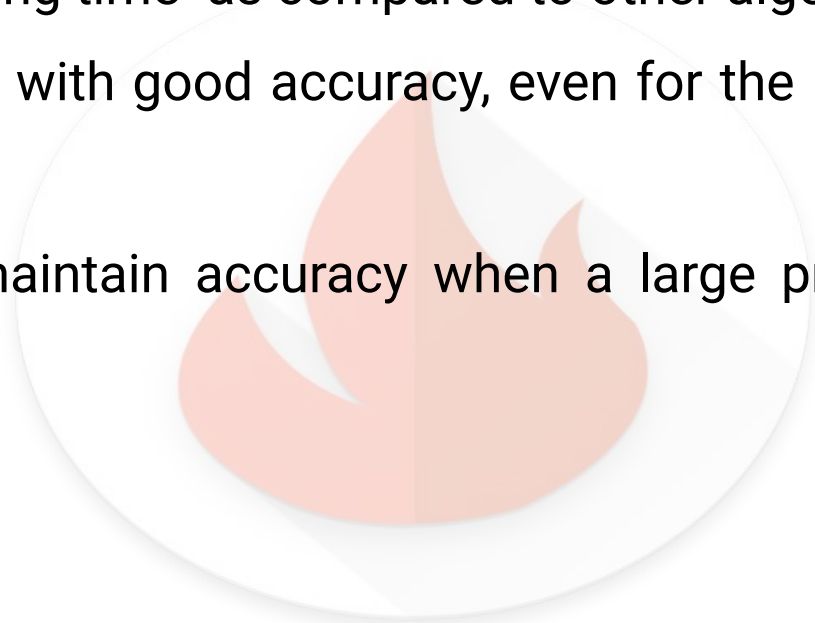

Random Forest Classifier

Random Forest Representation

Why Random Forest?

1. It takes less training time as compared to other algorithms.

2. It predicts output with good accuracy, even for the large dataset it runs efficiently.

3. It can be also maintain accuracy when a large proportion of data is missing.

Assumptions of Decision Tree

1. There should be some actual values in feature variable of the dataset so that the classifier can predict accurate results than a guessed result.

2. The predictions from each tree must have very low corrections.
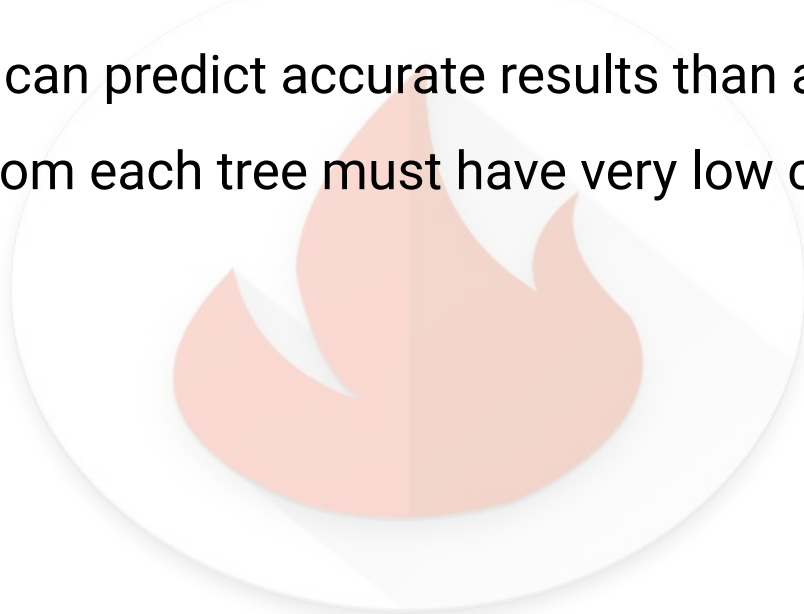
Application

1. **Banking:** Identification of loan risk.

2. **Medicine:** Risk of disease can be identified. Identify the patient's.

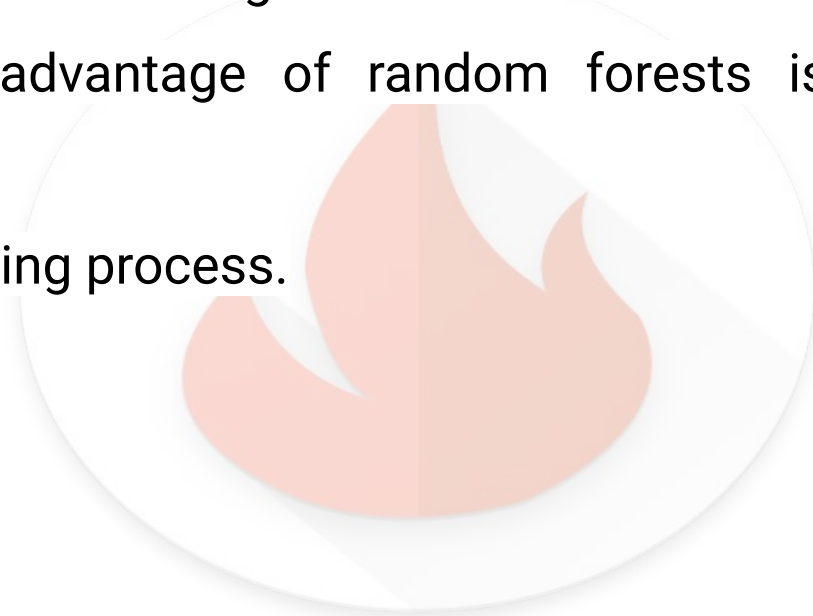3. **Marketing:** Marketing trends can be identified using this algorithm.
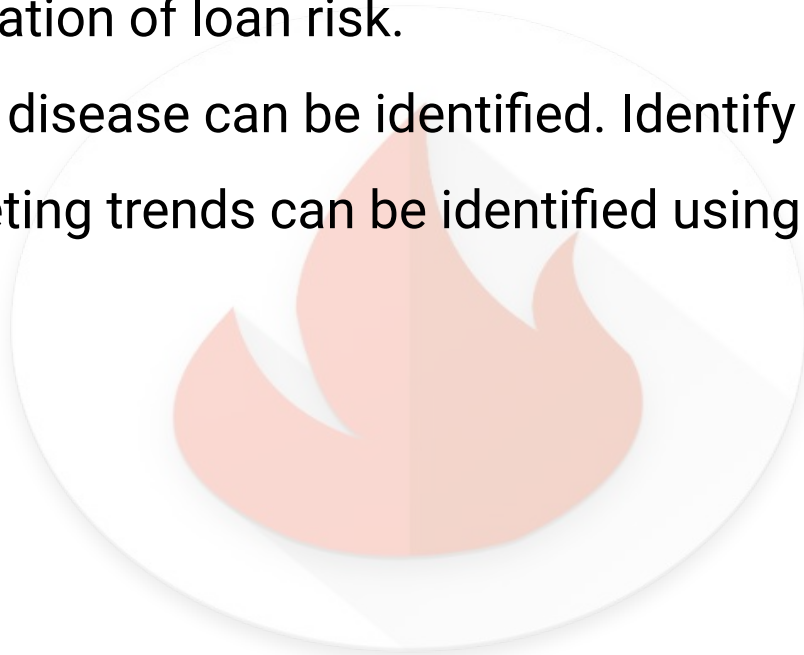
Advantage

1. Random Forest is capable of performing both classification and regression.

2. It is capable handle large datasets with high dimensionality.

3. It is considered as very accurate and robust model because it uses large number of DT to make prediction.

4. It does not suffer from the overfitting.

5. Can handle missing values. There are two ways - first use median value to replace continuous variable and second is to continue the proximity weighted avg of missing value

Disadvantage

1. It is not more suitable for regression tasks.

2. The biggest disadvantage of random forests is its computational complexity.

3. It is time consuming process.

Thank you