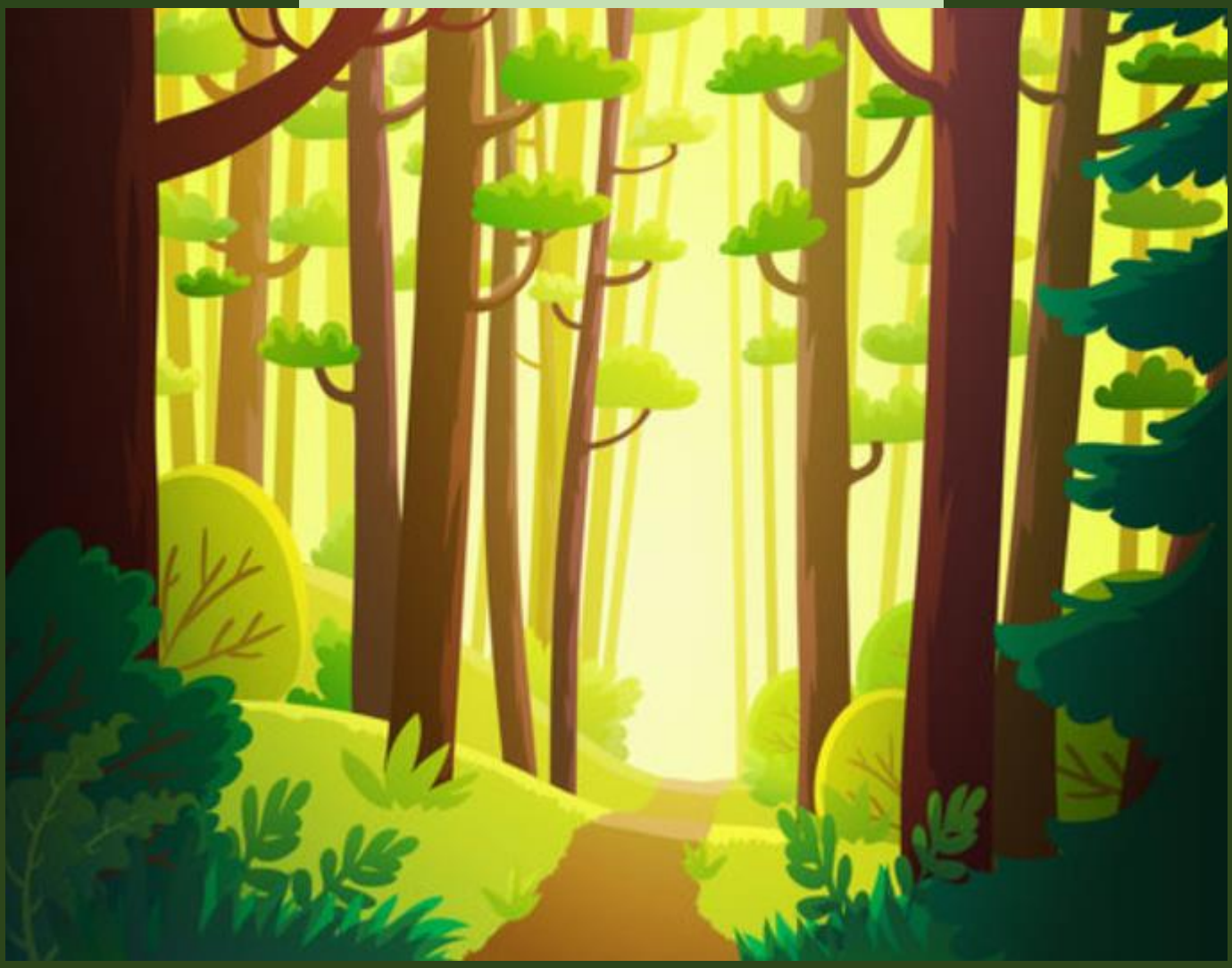
Random Forest



Sunit

What is it?

Ensemble learning method

Consists of multiple decision trees

Each tree is trained on a random subset of the data and features

Final prediction is the average of all tree predictions

Reduces overfitting and increases accuracy

Sunit

Random as…

Creation of a set of decision trees using random subsets of the features and data

Each tree in the ensemble is constructed using a random subset of the features

Random sampling of the data is also used to generate different training sets for each tree

This randomness helps to reduce overfitting and increase generalization performance

Forest as…

Ensemble of decision trees created

Composed of a large number of decision trees, typically hundreds or thousands

Each decision tree in the forest is grown independently

Sunit

Why use it?

Handles high-dimensional and complex data

Can handle missing values and maintain accuracy

Can provide feature importance ranking

Used in various fields such as finance, medicine, and marketing

Sunit

Decision Trees

A tree-based model that splits data based on features to make decisions

Splitting criteria: entropy, Gini impurity, information gain

Limitations: overfitting, instability, prone to bias, sensitive to small changes in data

Solutions: pruning, ensemble methods, feature selection

Sunit

Ensemble Learning

Ensemble learning combines multiple models to improve performance

Bagging creates multiple models with different data subsets

Boosting improves weak models through weighted training

Other ensemble methods include stacking, mixture of experts, and random forests

Ensemble learning reduces overfitting and improves accuracy

Sunit

Algorithm

Creates multiple decision trees using a random subset of features and data

Each tree is trained independently, using a different subset of the training data

The algorithm uses the mode or average prediction of the trees to make the final prediction

The final prediction is made by aggregating the predictions of all the trees

Feature Importance

Measures the relative contribution of each feature in the model's prediction

Calculated by evaluating the decrease in model performance when a particular feature is randomly permuted

Aids in feature selection by identifying the most relevant features for the model's accuracy

Helps in understanding the underlying data by highlighting the key factors that impact the outcome

Similarity Matrix

Measures the proximity of instances in the feature space

Proximity is calculated based on the number of times two instances end up in the same terminal node

Used for missing data evaluation

Visualize the structure of the data and gain insights into the relationships between instances

Proximity measure in random forest can be used to calculate feature importance and aid in feature selection

Compare

High-dimensional data and can handle noise

Use RF

Small datasets with a clear margin of separation

Use SVM

Complex non-linear data and large datasets

Use ANN

Binary classification and small datasets

Use Log Regression

Follow

Sunit Ghosh

to get #interesting
and latest #titbits
on #java, #AiMl, #cloud
technologies