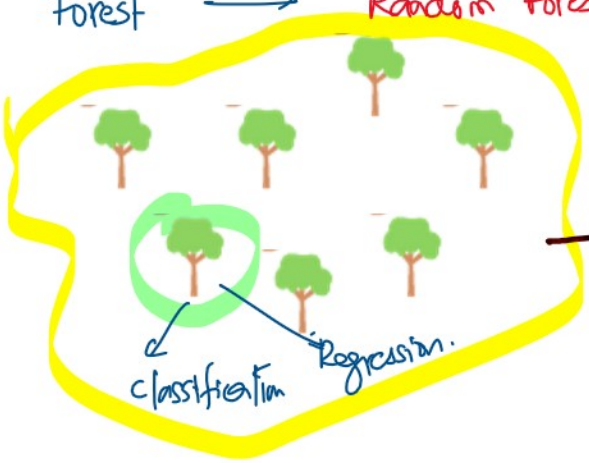
Random Forest

12 January 2024 07:55

Context → Algorithm

Tree → Decision Tree.

Forest → Random Forest.



→ collection of decision trees

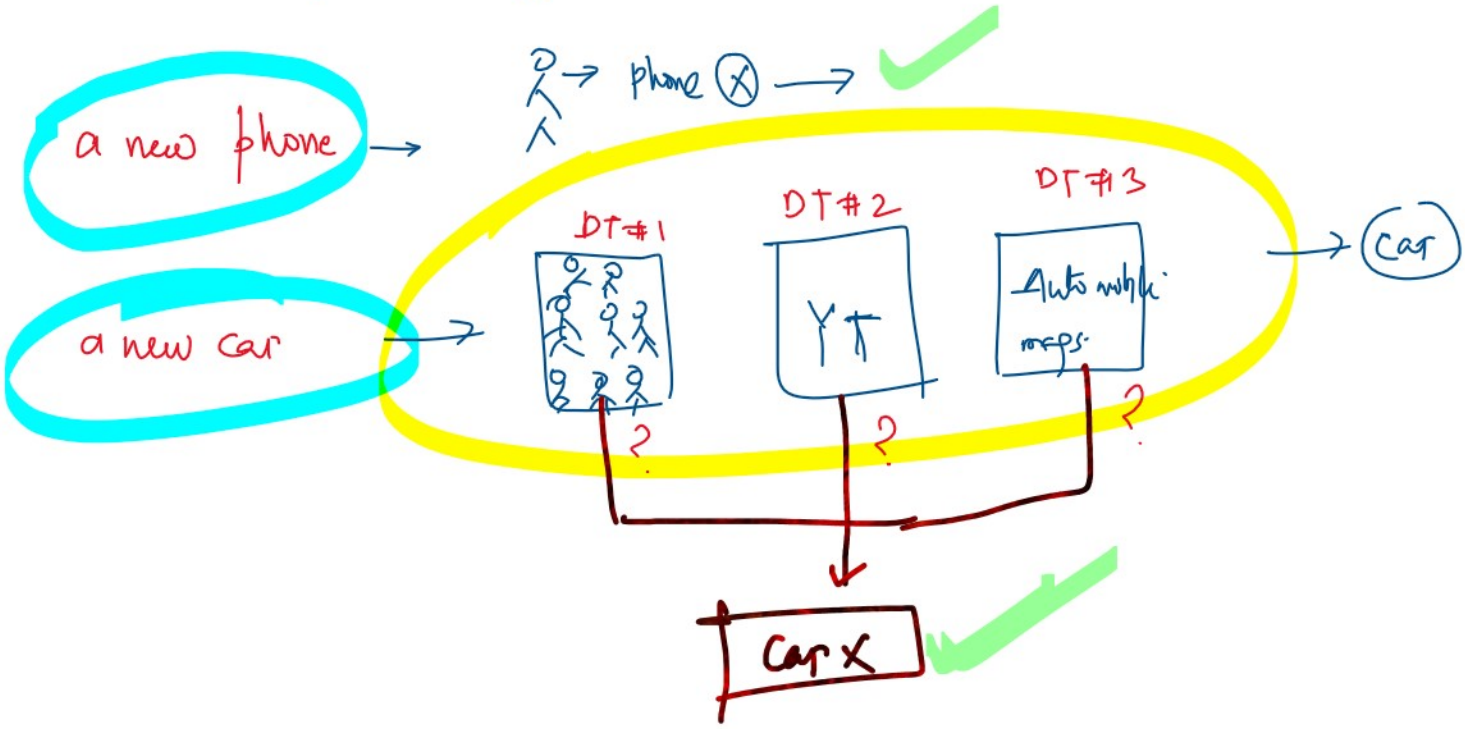
↓

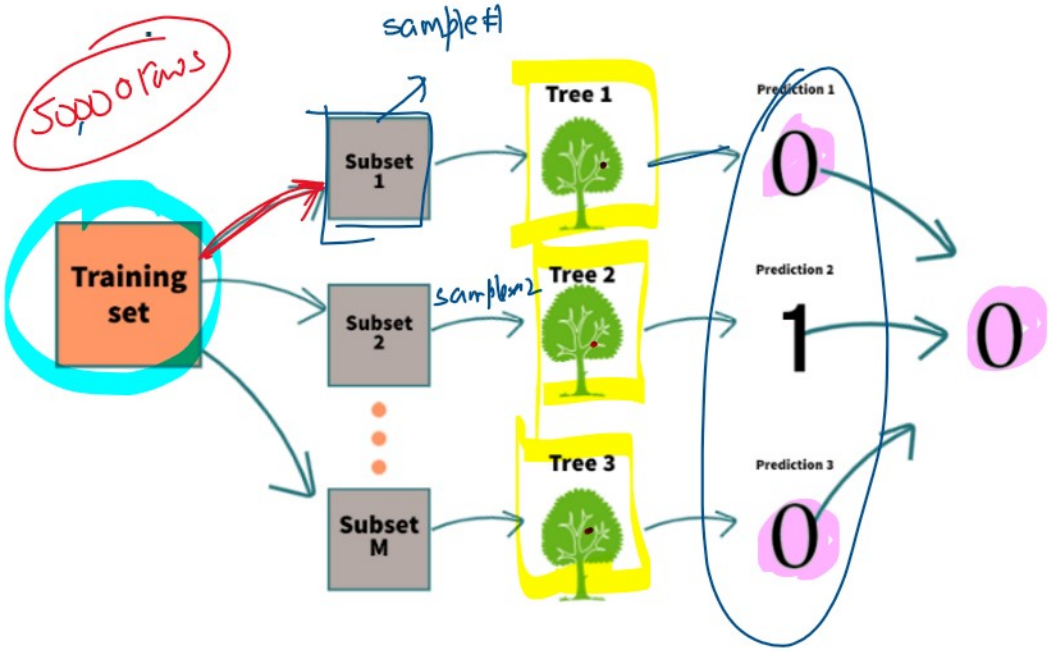
ensemble.

Random Forest is an ensemble learning algorithm that constructs many decision trees during

Random Forest is an ensemble learning algorithm that constructs many decision trees during the training.

- can be used for both regression and classification





Majority voting

https://www.datasciencecentral.com/

What is a random forest?

A random forest is like a group decision-making team in machine learning which combines the opinions of many trees (decision trees) to make better predictions, creating a more robust and accurate overall model.

- It can tackle both classification and regression problems effectively.
- It can handle complex datasets while mitigating overfitting, that makes it one of the valuable tools for various predictive tasks in machine learning.

= RF can deal with categorical input variables.

Working of Random Forest Algorithm

Ensemble means combining multiple models
↳ a collection of models which is used to make predictions rather an individual model.

Ensemble uses two types of methods:

Ensemble uses two types of methods:

a) Bagging b) Boosting

Bagging

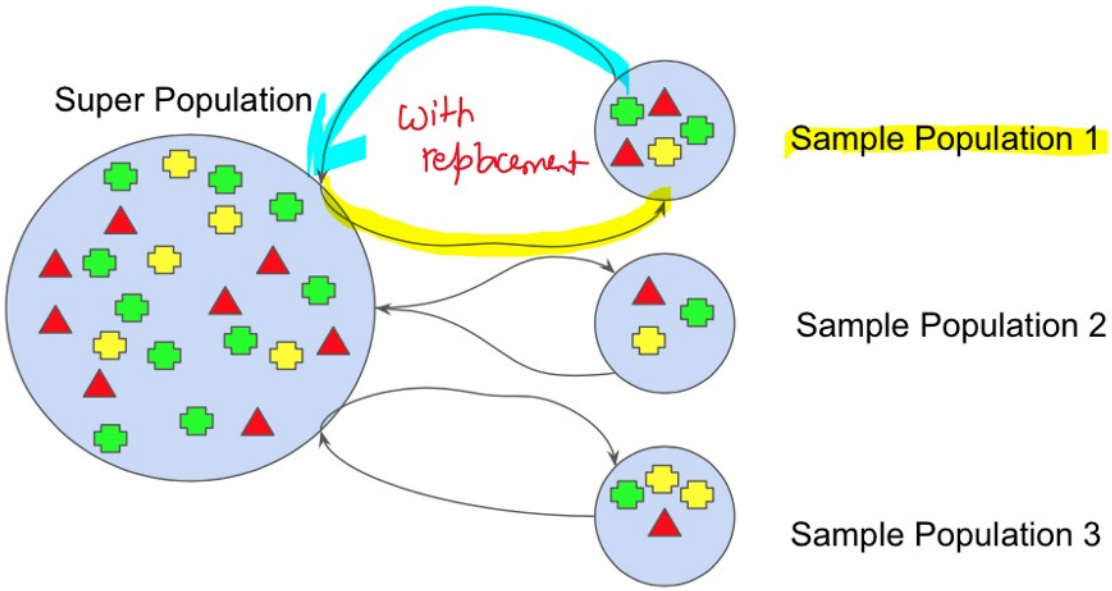
Bagging (Bootstrap aggregation) is a simple and very powerful ensemble method which is applied to a high-variance machine learning algo like decision trees.

- Bagging helps to decrease the model's variance.

Bootstrap:

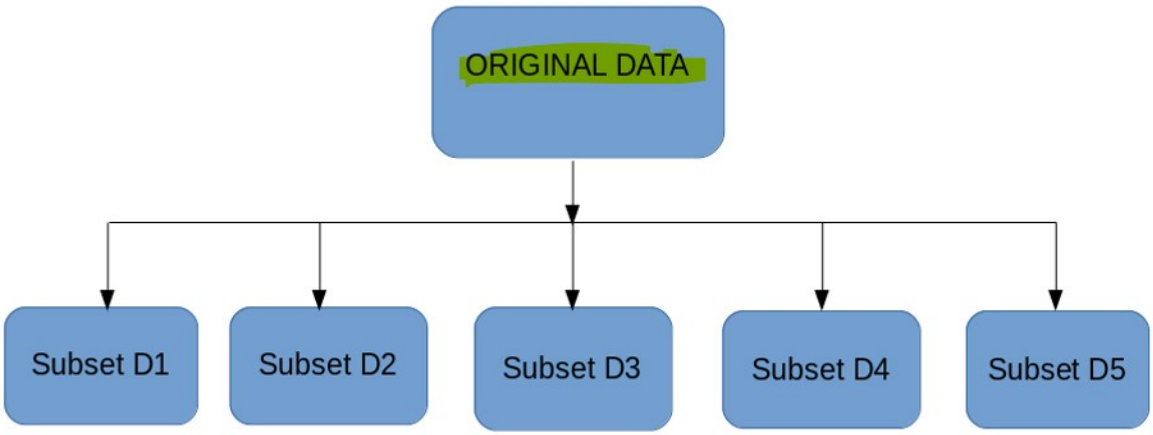
Bootstrap refers to a random sampling with replacement. It allows us to better understand the bias and variance with dataset.

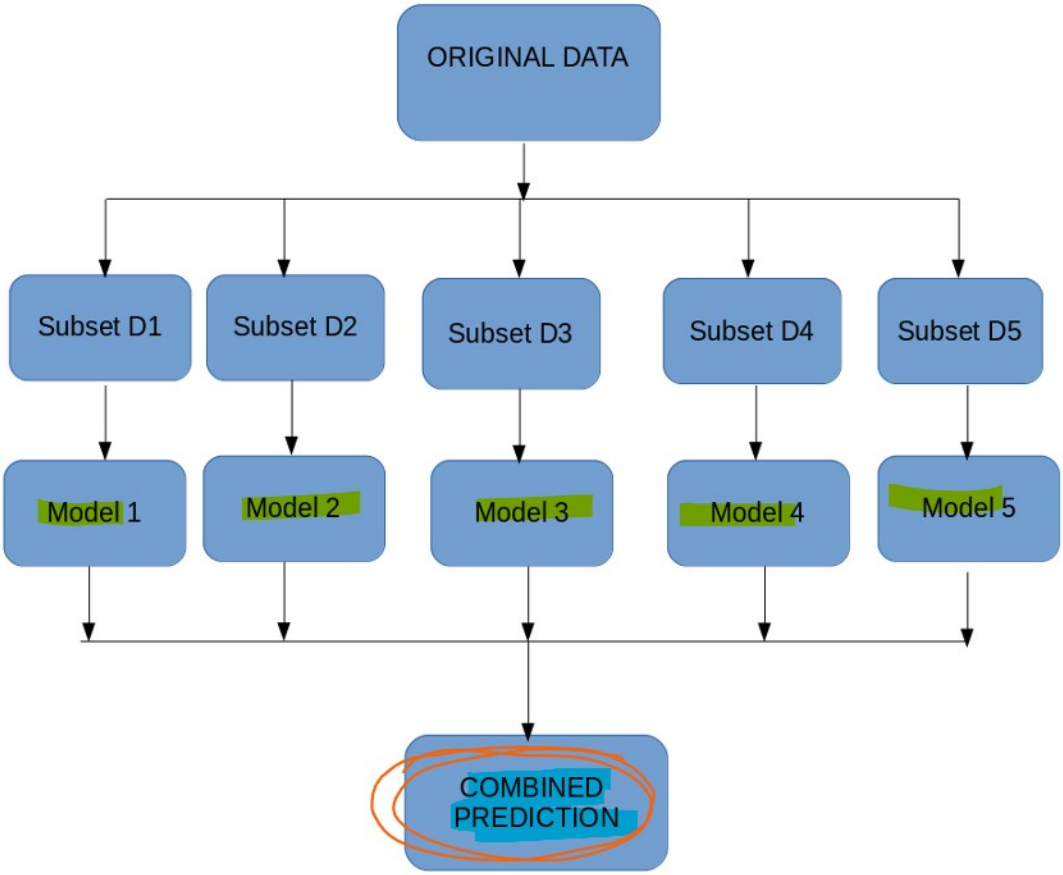
- It is a sampling technique we can create subsets of observations from the original dataset with replacement.



— selection of all the datapoints has equal probability.

In bagging;





Bagging works as follows:

1. Multiple subsets are created from the original dataset, selecting observations with replacement.
2. A base model is create on each of these subsets.
3. Models run in parallel and are independent of each other.
4. Final predictions are determined by combining the predictions from all the models.

In Boosting,

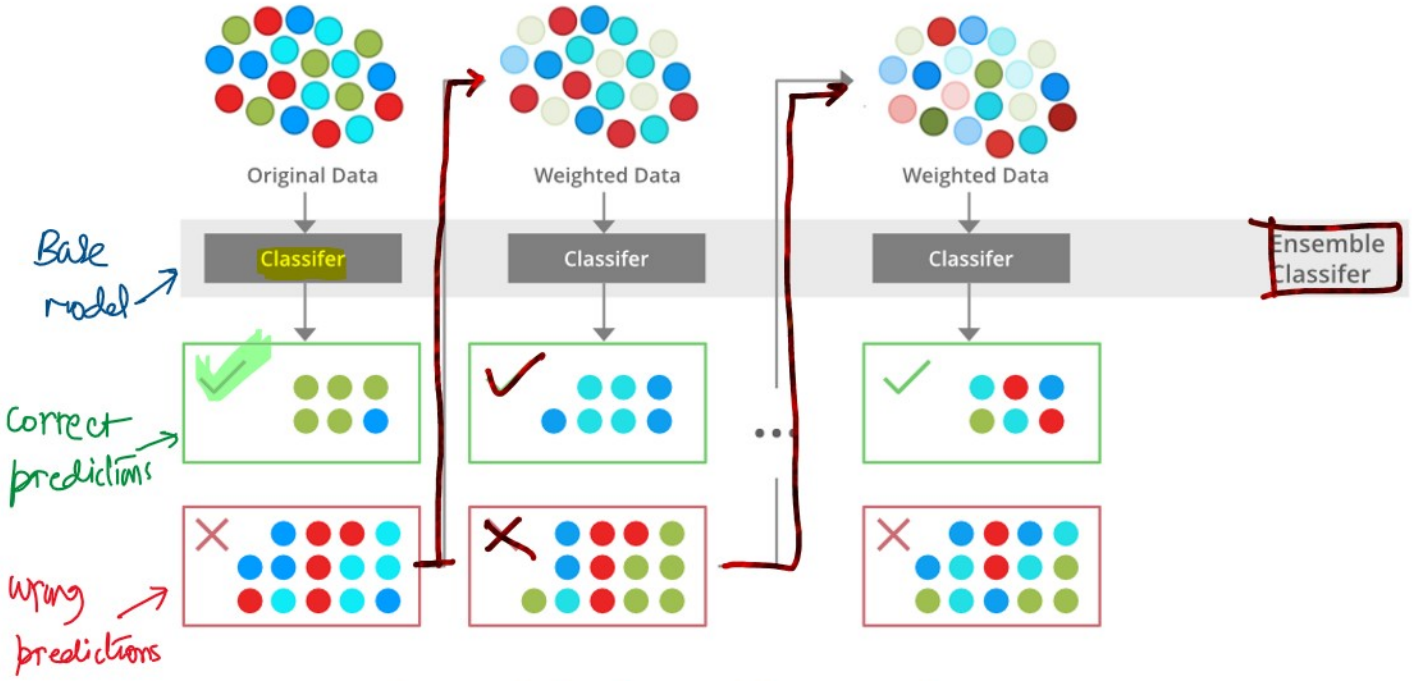
it reduces the model's bias

In Boosting,

— it reduces the model's bias

Boosting is a sequential process, where each subsequent model attempts to correct errors of the previous model.

The succeeding models are dependent on the previous model.



1. A subset is created from the original dataset and initially all data points are given equal weights

2. A base model is created on this dataset.

3. Observations which are incorrectly predicted are given higher weights.

4. In next iteration, another model is created and again predictions are made on this dataset.

4. In next iteration, another model is created
and again predictions are made on this dataset.
- the next model tries to correct the errors from the previous model.

5. Similarly, multiple models are created in sequential manner, each correcting the errors of the previous model.

6. The final model (strong learner) is the weighted mean of all the models

- GBM: Gradient Boosting Model
- XGBM - Extreme " " "
- Adaboost
- Light GBM -

→ Both bagging and boosting algorithms, such as AdaBoost, GBM, Random Forest are generally not strictly dependent on the assumption of the normality of distribution of data points. These algorithms are considered as non-parametric and work well with a variety of data distributions.

Parametric models make assumptions about the functional form of the underlying data distribution

functional form of the underlying data distribution

- Linear Regression, Logistic Regression.

Non parametric models make fewer assumptions about the underlying data distribution.

- Decision Trees, Random forest