BATCH : B 150 Data Science
LESSON : Machine Learning
DATE : 07.10.2022
SUBJECT : Supervised Learning
AdaBoost-Gradient Boost - XGBoost

techproeducation
techproeducation
techproeducation
techproeducation
techproedu

techproeducation.com info@techproeducation.com +1 (917) 768-7466

MACHINE LEARNING - 5

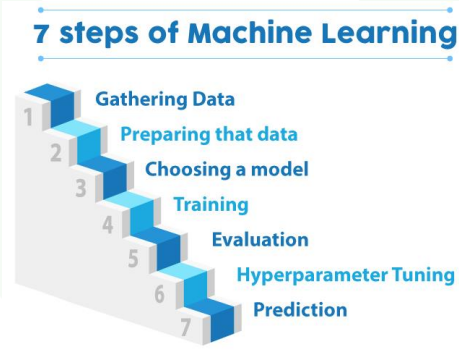
Machine Learning

Makine Öğrenmesi – 5

Overall Table of Contents

General Content

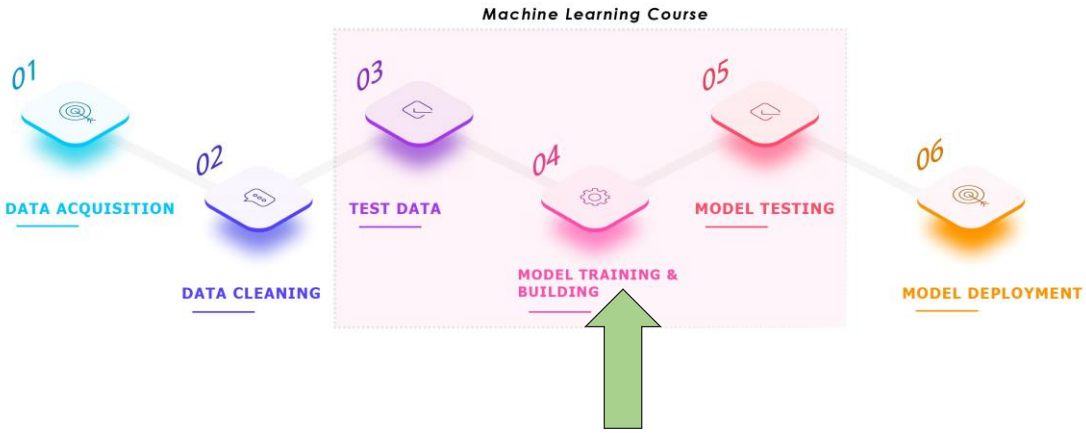
- Supervised Learnig Algorithm – **Bagging – Boosting Methods**
- Supervised Algorithm practices Python application
- Projects Solutions



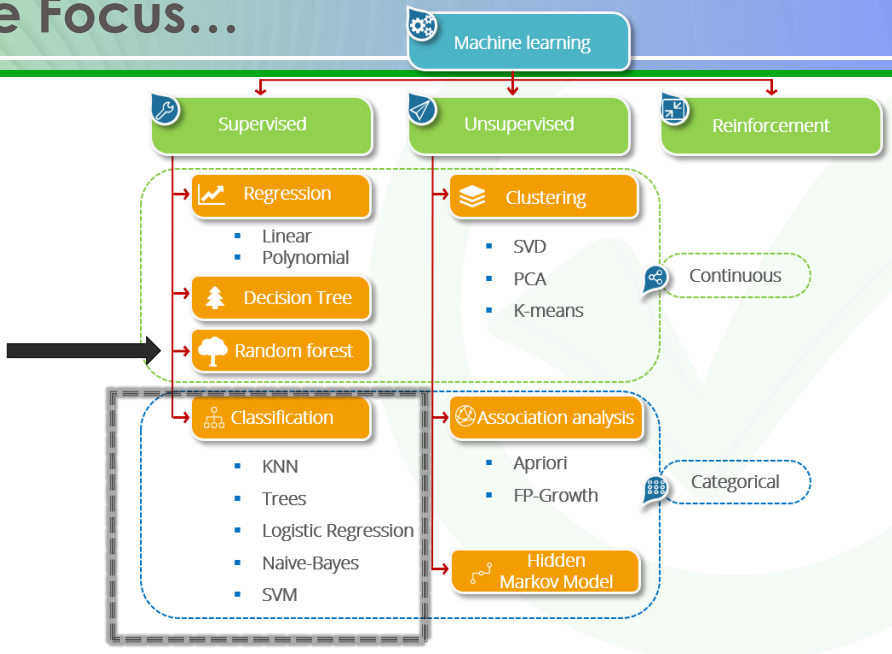
ENSEMBLE LEARNING

Where are we?

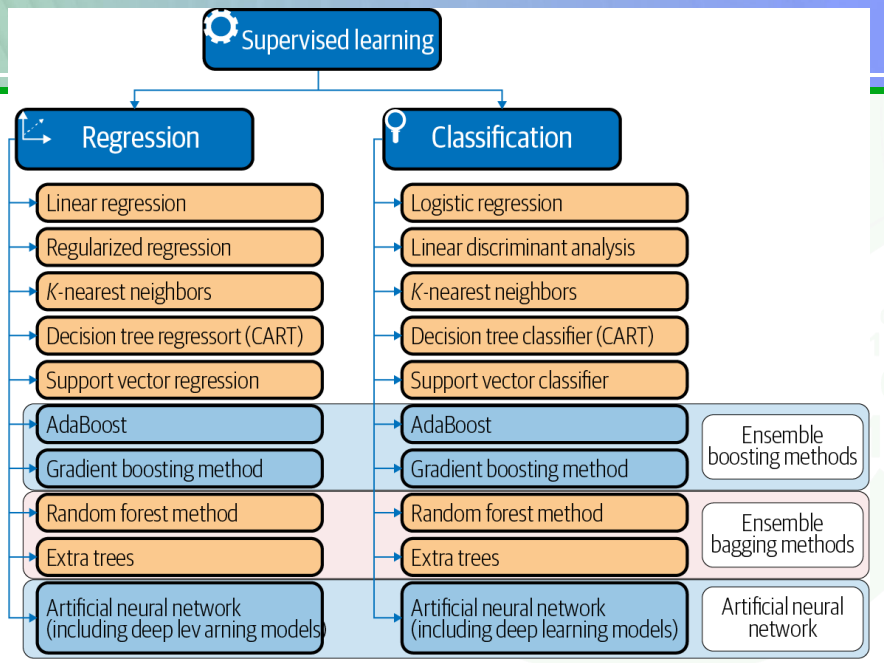
DATA SCIENCE



More Focus...



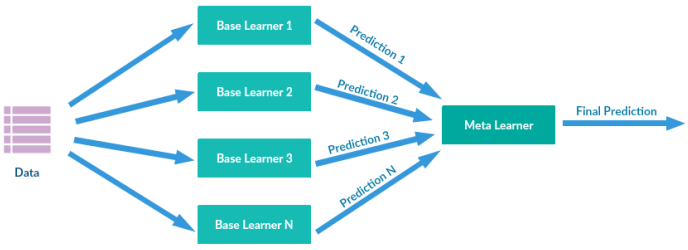
More..

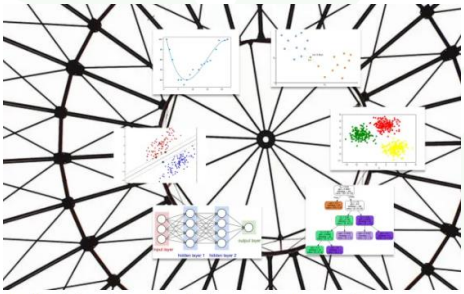


Supervised Learning

ENSEMBLE LEARNING METHODS (Kolektif Öğrenme Metotları)

- Farklı makine öğrenmesi algoritmalarını daha iyi bir tahmin için birlikte kullanmak mümkün mü?

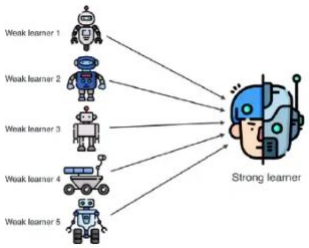




Ensemble methods combine several machine learning models to improve results

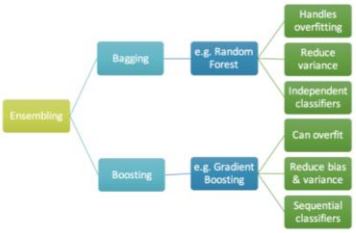
Supervised Learning

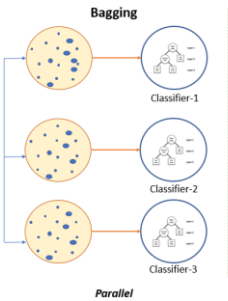
Özetle…

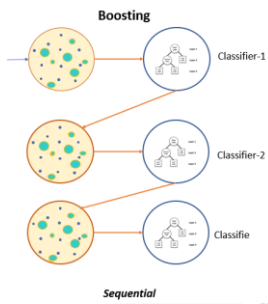


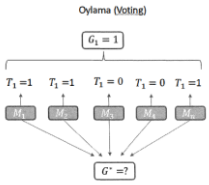
Supervised L. - ENSEMBLE LEARNING METHODS

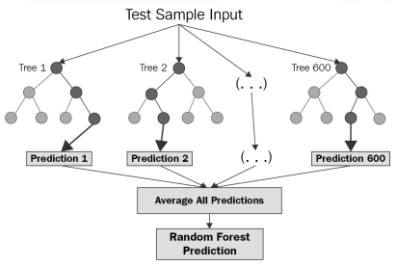
- ✓ Bagging (Bootstrap Aggregating)
- ✓ Boosting

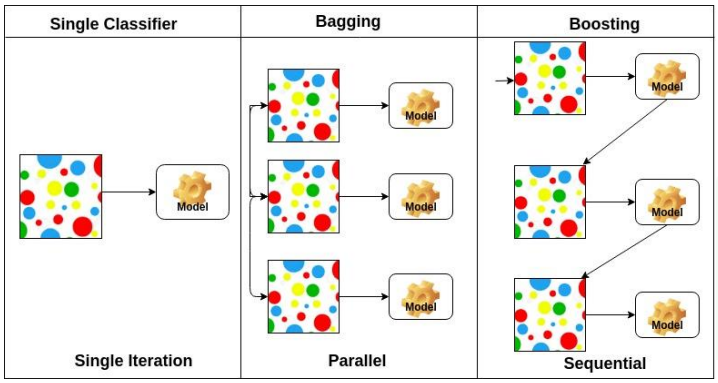


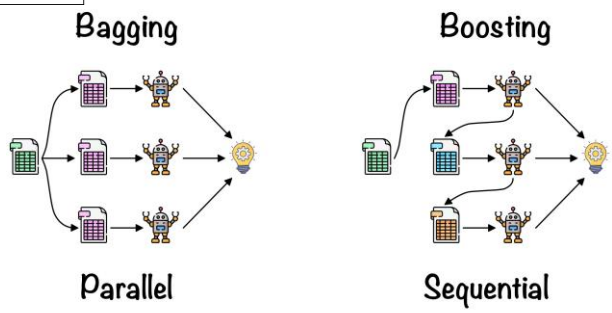






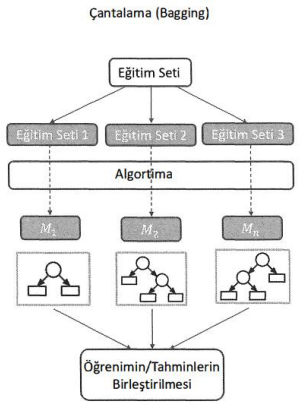


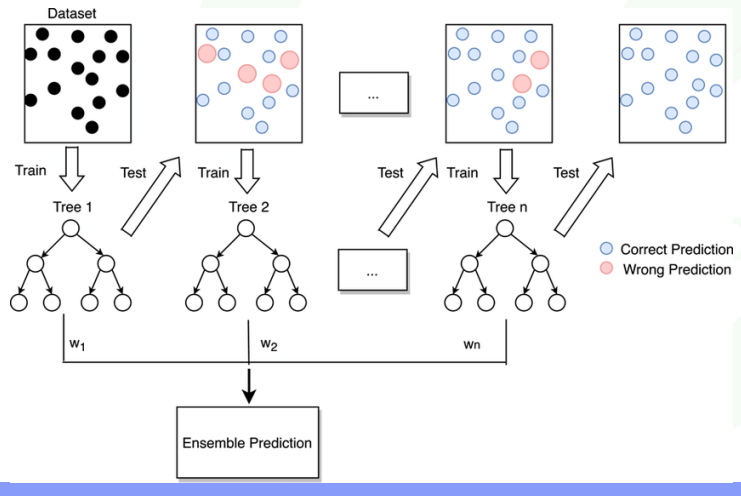




Supervised L. - ENSEMBLE LEARNING METHODS

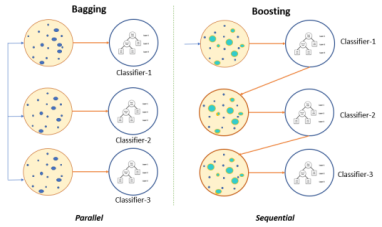
✓ Bagging (Random Forest-RF)

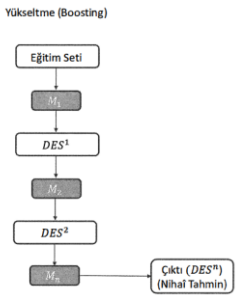


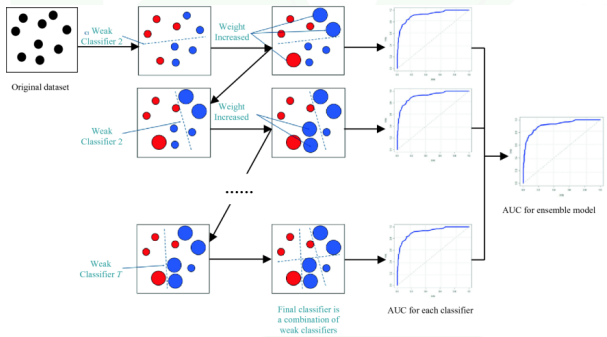


Supervised L. - ENSEMBLE LEARNING METHODS

- ✓ Bagging
- ✓ Boosting (Ada Boost-Gradient Boost-XGBoost)

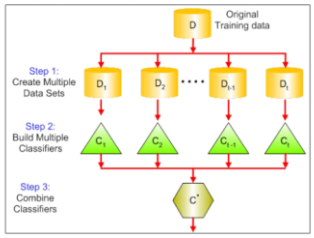






Bagging

Building multiple models (typically of the same type) from different subsamples of the training dataset.



	Bagging	Boosting
Similarities	• Uses voting • Combines models of the same type	
Differences	Individual models are built separately	Each new model is influenced by the performance of those built previously
	Equal weight is given to all models	Weights a model's contribution by its performance

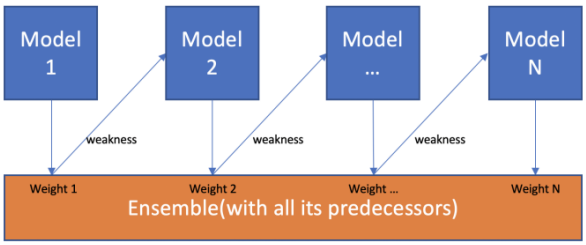
⇨ BOOSTING -->AdaBoost

⇨ BAGGING --> R F

Boosting

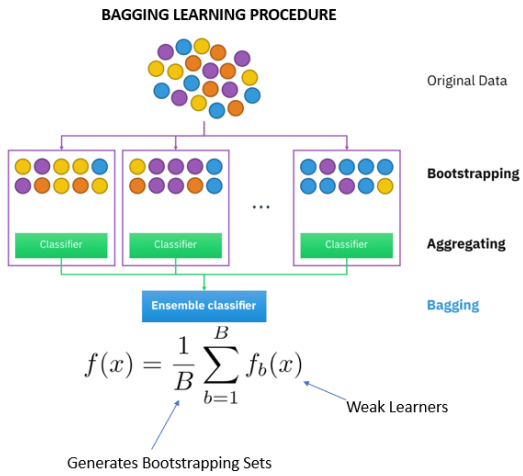
Building multiple models (typically of the same type) each of which learns to fix the prediction errors of a prior model in the chain.

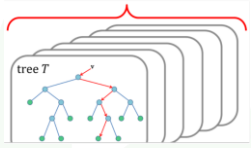
Model 1,2,..., N are individual models (e.g. decision tree)

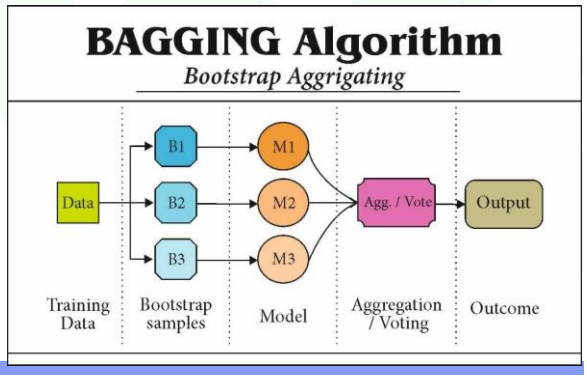


Supervised L. - ENSEMBLE LEARNING METHODS

✓ Bagging (Bootstrap Aggregating)

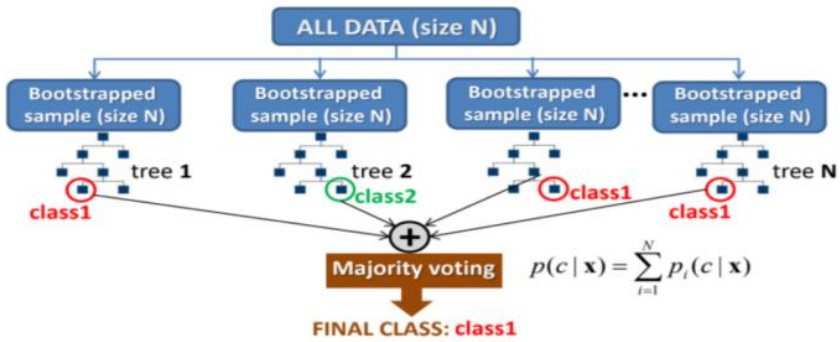






Supervised L. - ENSEMBLE LEARNING METHODS

✓ Random Forest (RF)



Hyperparameters:

"n_estimators" *parameter:* (default=100)

The number of trees in the forest.

The more # of trees, the better accuracy. But CPU intensive.

"max_depth" *parameter:* (default=None)

The maximum depth of the tree.

If None, then nodes are expanded until all leaves are pure.

"max_features" *parameter:*

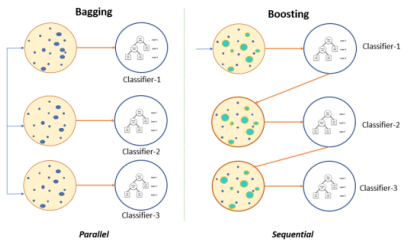
Number of features to consider when looking for the best split.

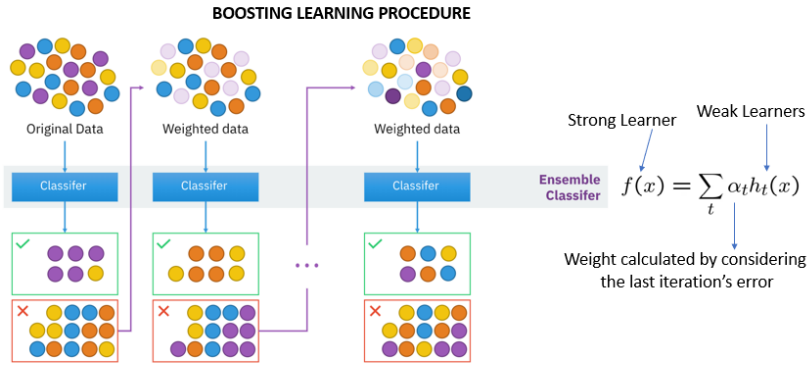
Increase will improve the performance but results in a correlation between the trees.

Avantaj - Dezavantajları

Supervised L. - ENSEMBLE LEARNING METHODS

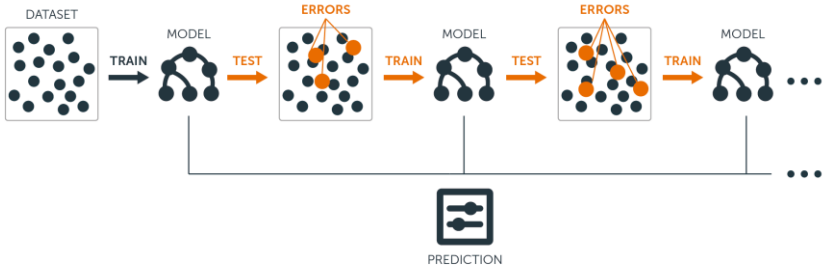
✓ Boosting





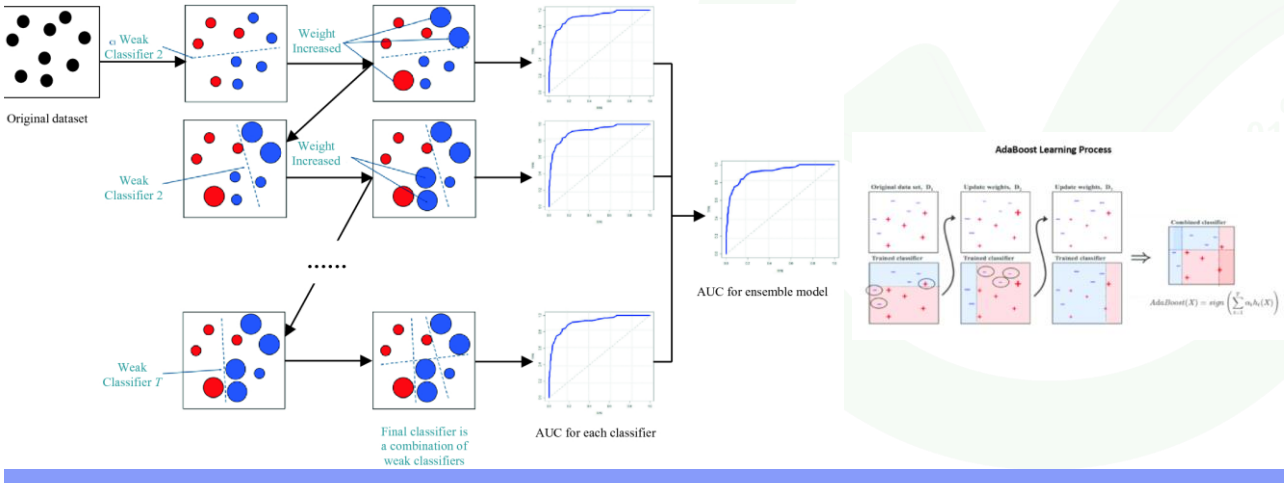
Supervised L. - ENSEMBLE LEARNING METHODS

✓ Boosting (Yükseltme)



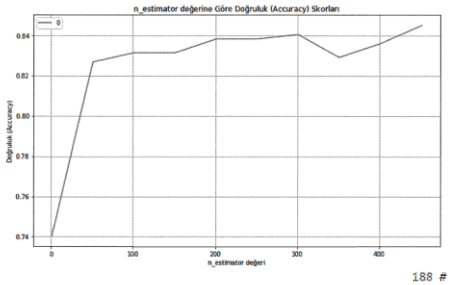
Supervised L. - ENSEMBLE LEARNING METHODS

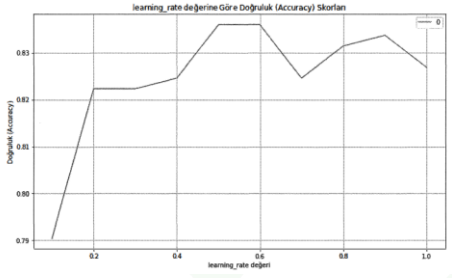
✓ Ada Boosting (Adaptive Boosting)



Supervised L. - ENSEMBLE LEARNING METHODS

✓ Ada Boosting (Adaptive Boosting)



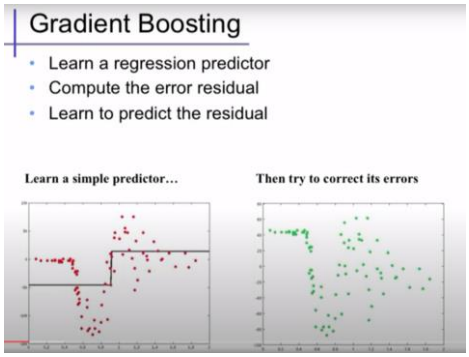


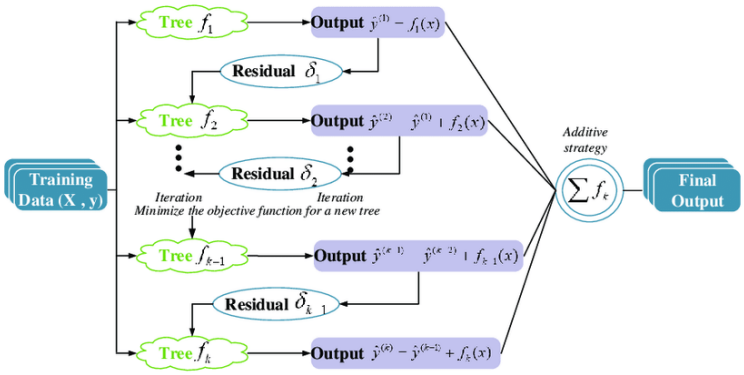

```
188 # Varyans sorunu var mı?
189 from sklearn import metrics
190 # Modelin eğtim seti  için doğruluk oranı:
191 print('Accuracy:%0.3f'%metrics.accuracy_score(y_train, y_pred_train))
192 # Modelin test seti için doğruluk oranı:
193 print('Accuracy:%0.3f'%metrics.accuracy_score(y_test, y_pred_test))
194
```

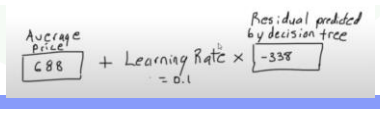
```
Accuracy:0.863
Accuracy:0.827
```

Supervised L. - ENSEMBLE LEARNING METHODS

✓ Gradient Boosting



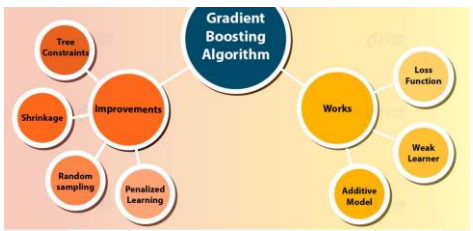




https://youtu.be/3zEqUSf5duw

Supervised L. - ENSEMBLE LEARNING METHODS

✓ Gradient Boosting



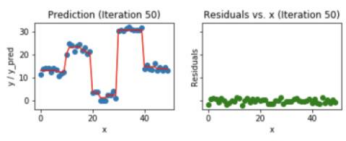


Fig 7. Visualization of gradient boosting prediction (iteration 50th)

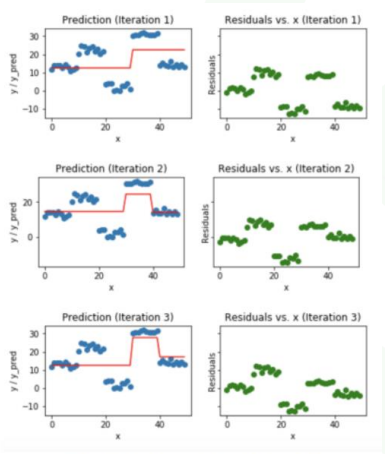


Fig 5. Visualization of gradient boosting predictions (First 4 iterations)

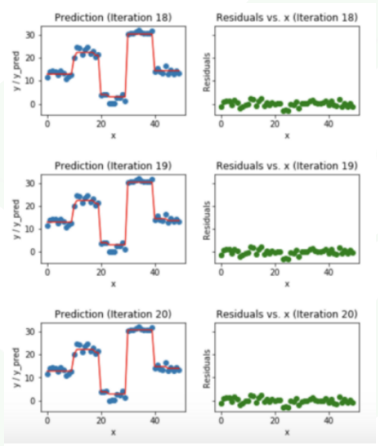


Fig 6. Visualization of gradient boosting predictions (18th to 20th iterations)

https://blog.mlreview.com/gradient-boosting-from-scratch-1e317ae4587d

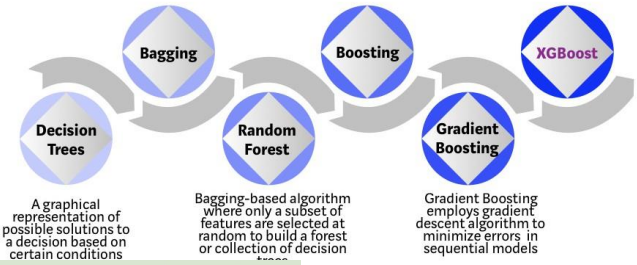
Supervised L. - ENSEMBLE LEARNING METHODS

✓ XG Boost (EXtreme Gradient Boosting)

Bootstrap aggregating or Bagging is a ensemble meta-algorithm combining predictions from multiple-decision trees through a majority voting mechanism

Models are built sequentially by minimizing the errors from previous models while increasing (or boosting) influence of high-performing models

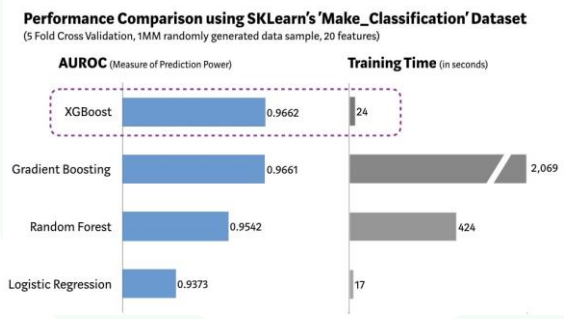
Optimized Gradient Boosting algorithm through parallel processing, tree-pruning, handling missing values and regularization to avoid overfitting/bias



A graphical representation of possible solutions to a decision based on certain conditions

Bagging-based algorithm where only a subset of features are selected at random to build a forest or collection of decision

Gradient Boosting employs gradient descent algorithm to minimize errors in sequential models



"n_estimators" *parameter:* (default=100)

The number of boosting stages to perform.

"max_depth" *parameter:* (default=6)

*Max. depth of the individual estimators. This parameter **limits the number of nodes** in the tree.*

"subsample" *parameter:* (default=1.0)

The fraction of samples to be used for fitting the individual base learners. Choosing subsample < 1.0 leads to a reduction of variance and an increase in bias

"learning_rate" *parameter:* (default=0.3)

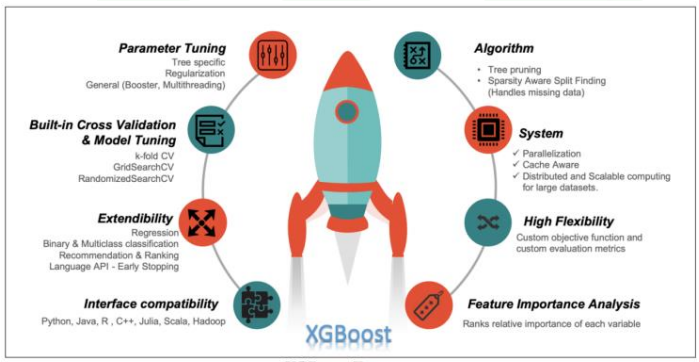
Learning rate shrinks the contribution of each tree by learning_rate. There is a trade-off between learning_rate and n_estimator

Supervised L. - ENSEMBLE LEARNING METHODS

✓ XG Boost (EXtreme Gradient Boosting)

✓ XGB neden diğerlerinden daha iyi?

- ✓ ***Regularization***
- ✓ ***Parallel Processing***
- ✓ ***High Flexibility***
- ✓ ***Handling Missing Values***
- ✓ ***Tree Pruning***
- ✓ ***Built-in Cross-Validation***



XGBoost Features

Supervised L. - ENSEMBLE LEARNING METHODS

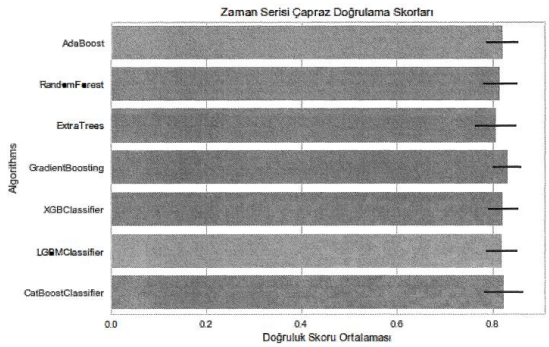
✓ ENSEMBLE KOLEKTİF ÖGRENME YÖNTEMLERİ ARASINDAN MODEL SEÇİMİ

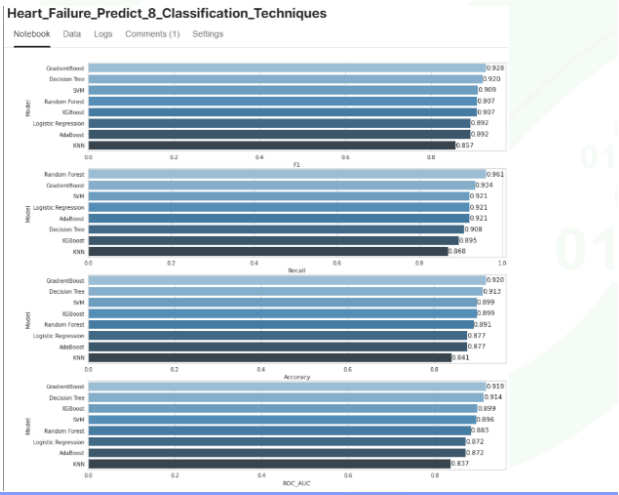
✓ Zaman-performans ikilemi

- ✓ **i)** ortaya çıkacak performans bir tek algoritmanın ürettiği modelinkinden daha yüksek olmalı;
- ✓ **ii)** işlem süresi algoritmaların tek tek uygulanmasından daha kısa sürmeli.

Supervised L. - ENSEMBLE LEARNING METHODS

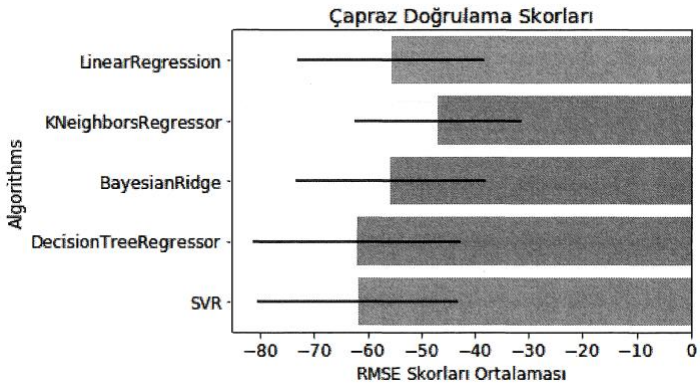
✓ XG BOOST (eXtram Gradient Boosting)





Supervised L. - ENSEMBLE LEARNING METHODS

REGRESYON OLARAK KULLANILABİLEN SINIFLANDIRMA ALGORTİMALARI



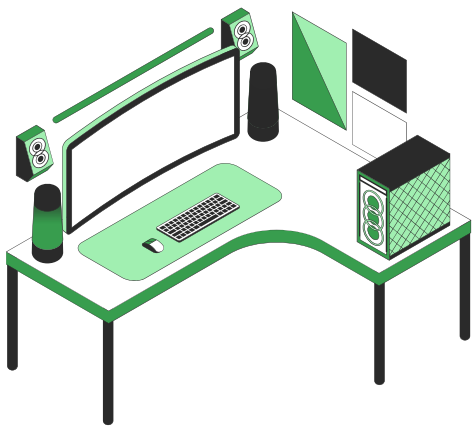

```
Eğitim Seti için RMSE:36.87
Test Seti için RMSE:77.11

  Eğitim Seti için R2:0.88
  Test Seti için R2:0.56
```

- **Exploratory Data Analysis and Visualization**
- **Machine Learning**
  - **Train | Test Split**
    - X_train, X_test, y_train, y_test = train_test_split()
  - **Scalling (if needed)**
    - scaler = scaler_name()
    - scaler.fit_transform(X_train)
    - scaler.transform(X_test)
  - **Modelling**
    - model = model_name().fit(X_train, y_train)
    - y_pred = model.predict(X_test)
    - y_pred_proba = model.predict_proba(X_test)
  - **Model Performance**
    - Regression => r2_score, MAE, MSE, RMSE
    - Classification => accuracy, recall, precision, f1_score (confusion_matrix, classification_report)
    - Cross Validate => cross_val_score, cross_validate
  - **Tunning (if needed)**
    - grid_param = {}
    - GridsearchCV(grid_param)
  - **Final Model**
    - model = model_name().fit(X, y)
- **Model Deployment**

Bu dersi anladım..

Everythig is clear ?



Do you have any questions?

Send it to us! We hope you learned something new.

TECHPROED