BATCH : **B 150** Data Science
LESSON : **Machine Learning**
DATE : 24.07.2023
SUBJECT : **Supervised Learning**
Logistic Regression – Decision Tree - SVM

techproeducation
techproeducation
techproeducation
techproeducation
techproedu

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

MACHINE LEARNING - 4

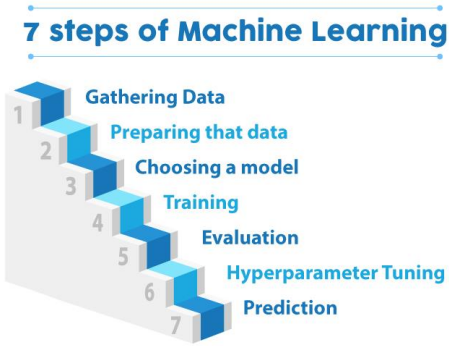
Machine Learning

Makine Öğrenmesi – 4

Overall Table of Contents

General Content

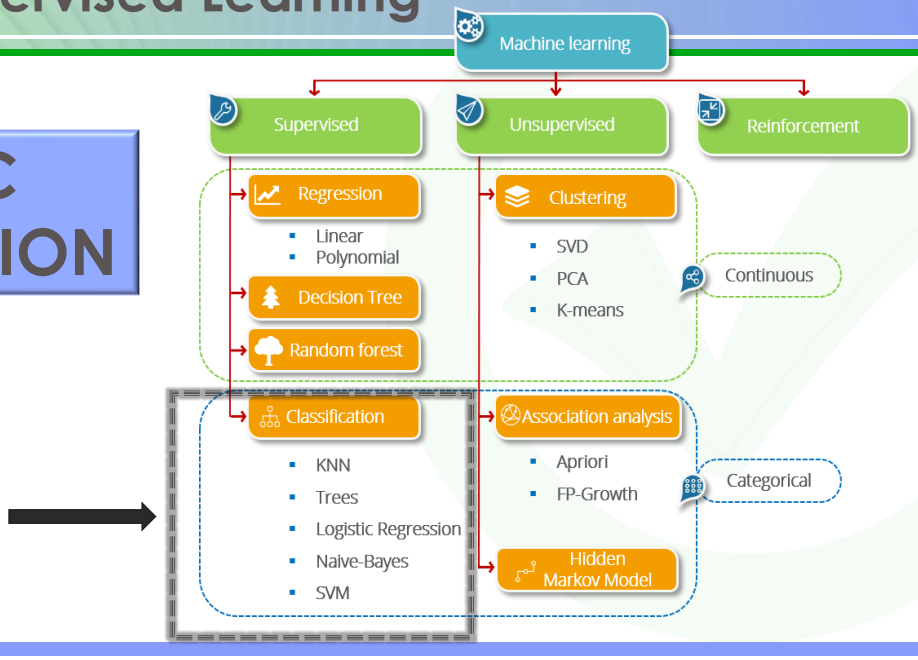
- Supervised Learnig Algorithm - **Classisfication**
- Supervised Algorithm practices Python application
- Projects Solutions



LOGISTIC REGRESSION

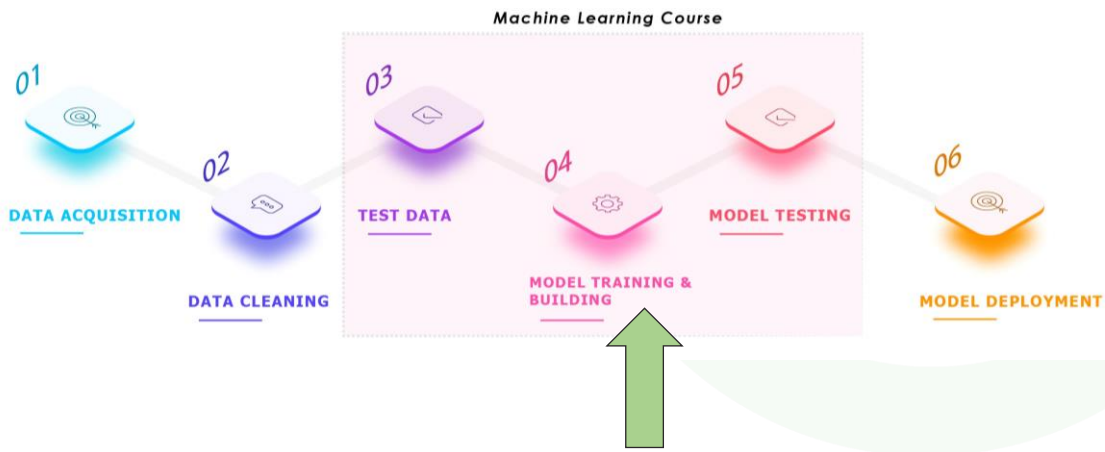
Supervised Learning

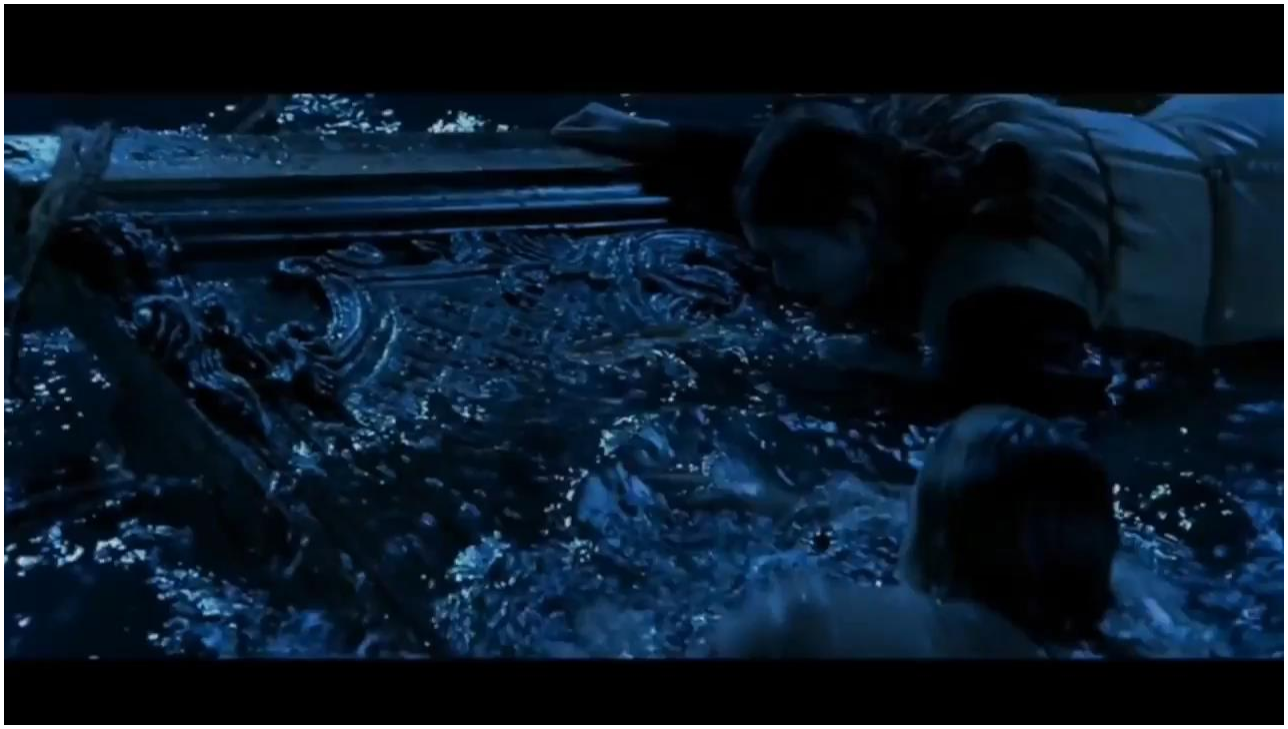
LOGISTIC REGRESSION



Run Navigation

Where are we?





Supervised Learning

LOGISTIC REGRESSION

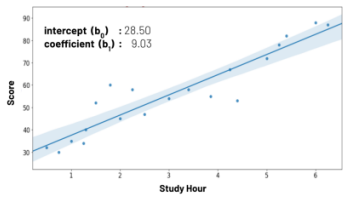
- Titanic filmini hatırlayalım…
- İhtimaller…
- Adı regresyon ama…….
- Avantajları…

İhtimaller	*Rose'un Değerlendirmesi*
Gemi batmadan varır.	*Süper!*
Gemi batar ama ikisi de kurtulur.	*O da olur.*
Gemi batar ve ikisi de kurtulamaz.	*Birkaç dakika önce intihara yeltenen Rose için üzerinde durulacak bir ihtimal değil.*
Gemi batar, Jack kurtulur, Rose kurtulmaz.	
Gemi batar, Rose kurtulur, Jack kurtulamaz.	*Fecaat!*

Supervised Learning

LOGISTIC REGRESSION

Student	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
Study Hour	0.5	0.75	1.00	1.25	1.30	1.50	1.80	2.00	2.25	2.50	3.00	3.40	3.85	4.25	4.40	5.00	5.25	5.40	6.00	6.25
Score	32	30	35	34	40	52	60	45	58	47	54	58	55	67	53	72	78	82	88	87

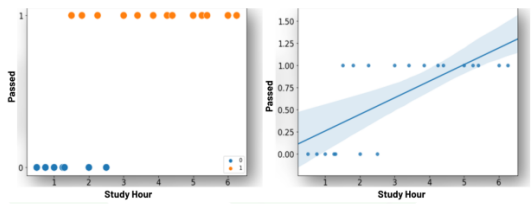


Regression Equation

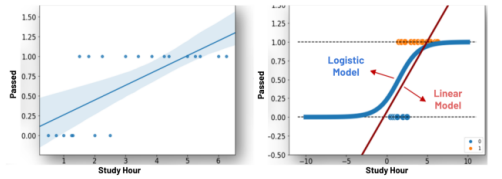
$y = b_0 + b_1 * X$

$y = 28.50 + 9.03 * \text{Study hour}$

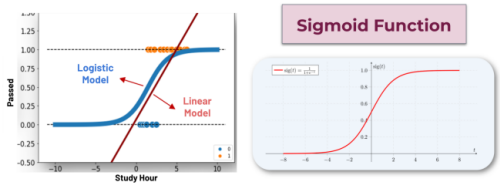
Student	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
Study Hour	0.5	0.75	1.00	1.25	1.30	1.50	1.80	2.00	2.25	2.50	3.00	3.40	3.85	4.25	4.40	5.00	5.25	5.40	6.00	6.25
Passed	0	0	0	0	0	1	1	0	1	0	1	1	1	1	1	1	1	1	1	1



Student	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
Study Hour	0.5	0.75	1.00	1.25	1.30	1.50	1.80	2.00	2.25	2.50	3.00	3.40	3.85	4.25	4.40	5.00	5.25	5.40	6.00	6.25
Passed	0	0	0	0	0	1	1	0	1	0	1	1	1	1	1	1	1	1	1	1



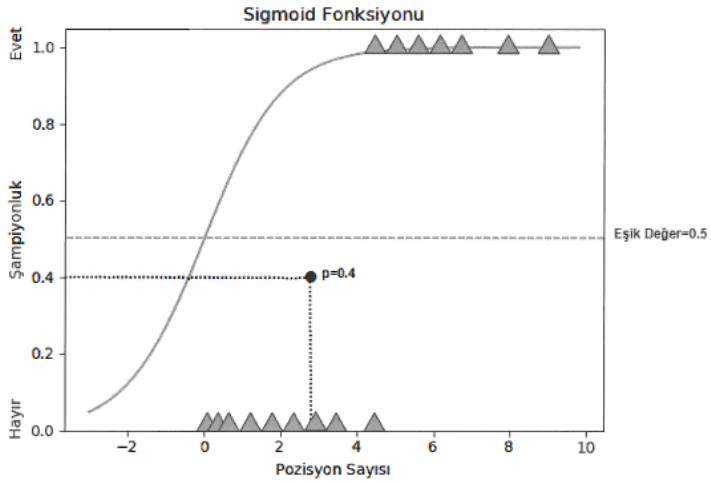
Student	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
Study Hour	0.5	0.75	1.00	1.25	1.30	1.50	1.80	2.00	2.25	2.50	3.00	3.40	3.85	4.25	4.40	5.00	5.25	5.40	6.00	6.25
Passed	0	0	0	0	0	1	1	0	1	0	1	1	1	1	1	1	1	1	1	1



Supervised Learning

LOGISTIC REGRESSION

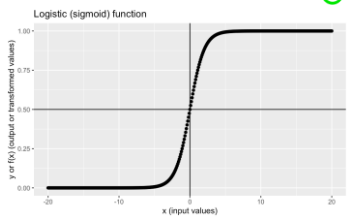
- Örnek

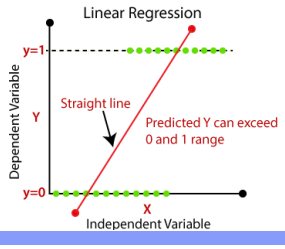


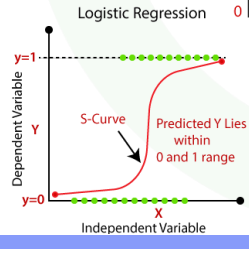
Supervised Learning

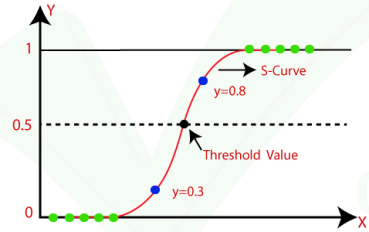
LOGISTIC REGRESSION

- Logit (sigmoid) fonksiyon
- Eşik değer katsayısı (disciriminant threshold) -







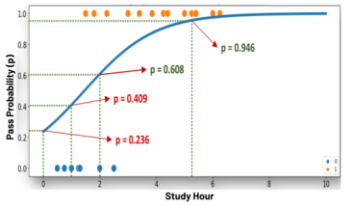


Supervised Learning

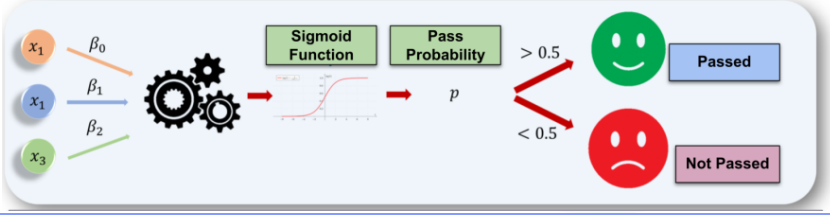
LOGISTIC REGRESSION

- Logit (sigmoid) fonksiyon

Student	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
Study Hour	0.5	0.75	1.00	1.25	1.30	1.50	1.80	2.00	2.25	2.50	3.00	3.40	3.85	4.25	4.40	5.00	5.25	5.40	6.00	6.25
Passed	0	0	0	0	0	1	1	0	1	0	1	1	1	1	1	1	1	1	1	1



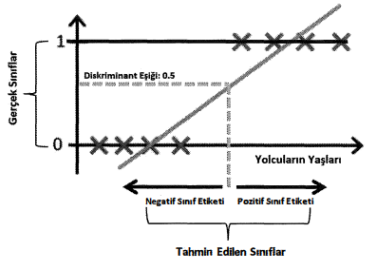
Student	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
Study Hour	0.5	0.75	1.00	1.25	1.30	1.50	1.80	2.00	2.25	2.50	3.00	3.40	3.85	4.25	4.40	5.00	5.25	5.40	6.00	6.25
Passed	0	0	0	0	0	1	1	0	1	0	1	1	1	1	1	1	1	1	1	1



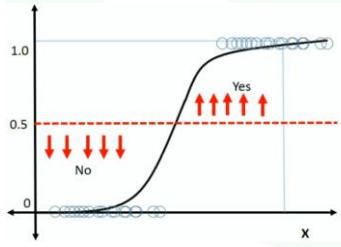
Supervised Learning

LOGISTIC REGRESSION

- Yaş ve hayatta kalma ihtimaliyetini inceleyelim



Gerçek Sınıflar	Tahmin Edilen Sınıflar: Negatif(0)	Pozitif (1)
Negatif (0)	DN	YP
Pozitif (1)	YN	DP



Threshold Value of 50% probability
> 0.5 Success
< 0.5 Failure

Supervised Learning

LOGISTIC REGRESSION

- sigmoid fonksiyonu→ logit fonksiyonu → logistik regresyon süreci
- Titanic veri seti hakkında..

Sigmoid Fonksiyonu → Logit Fonksiyonu → Logistik Regresyon

$$y = \alpha_0 + \beta_1 x$$

Doğrusal regresyonda y hedef değişkeni x öznitelik değişkeni ile tahmin edilir. β katsayısı özniteliğin hedef değişken üzerindeki etkisini ya da hedef değişkendeki değişimlere katkısını gösterir.

$$p(t) = \frac{1}{1 + e^{-t}}$$

Sigmoid fonksiyonu bir öznitelik değişkeninin değerlerini bir olayın olma olasılığına indirger. Değişkenin gözlem değerleri ne olursa olsun sigmoid fonksiyonu bunları 1 ve 0'lara indirger.

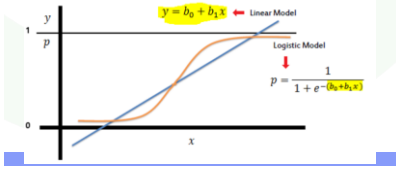
$$p(x) = \frac{1}{1 + e^{-(\alpha_0 + \beta_1 x)}}$$

Doğrusal regresyon fonksiyon kalıbının sigmoid fonksiyonu ile geliştirilmesi ***Logistik fonksiyonu***, $p(x)$, verir. Buna göre hedef değişkene konu olan bir olayın olması öznitelik değişkeninin hedef değişkene katkı yapma olasılığına bağlıdır.

$$g(x) = ln\frac{p(x)}{1 - p(x)} = \alpha_0 + \beta_1 x = y$$

Logit foksiyonu, $g(x)$, ise aslında logistik fonksiyonun tersidir. Buna göre bir olayın olma olasılığı (hedef değişkenin doğru pozitif olması) özniteliklerin doğrusal kombinasyonuna bağlıdır.

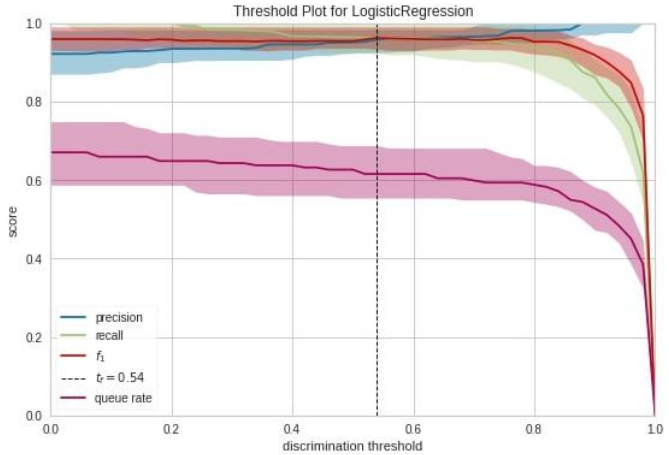
Veri Setinin Açıklamaları:
PassengerId: Yolcu numarası
Survived: Kaza sonucu yolcunun kurtulup kurtulmadığı ile ilgili sınıf etiketi. 1: kurtuldu; 0: kurtulamadı. Bu etiket sadece eğitim setinde mevcuttur.
Pclass: Bilet Sınıfı. Aynı zamanda sosyoekonomik statü indikatörü (1: Üst Sınıf,2: Orta Sınıf, 3: Alt Sınıf).
Names: Yolcunun adı ve soyadı
Sex: Yolcunun cinsiyeti (female: kadın; male: erkek)
Age: Yolcunun yaşı
SibSp: Yolcunun beraber seyahat ettikleri ailesindeki kişi sayısı. Kardeşler ve eşlerin toplam sayısı.
Parch: Yolcunun beraber seyahat ettikleri ailesindeki kişi sayısı. Annesi, babası, çocukları. Eğer çocuk dadısı ile seyahat ediyorsa parch=0 olarak kaydedilmiştir.
Ticket: Bilet numarası
Fare: Bilet ücreti
Cabin: Kabin numarası
Embarked: Yolcunun gemiye bindiği liman (C: Cherbourg, Q:Queenstown, S:Southampton)

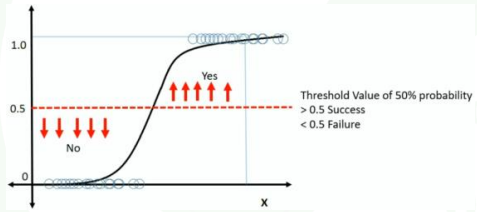


Supervised Learning

LOGISTIC REGRESSION

- Yellowbrick kullanımı

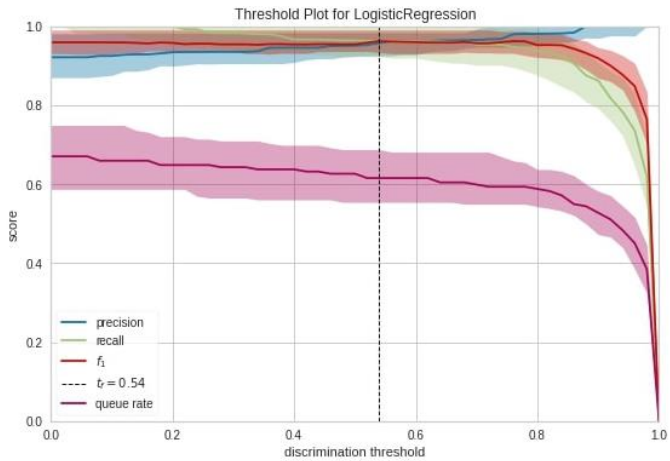


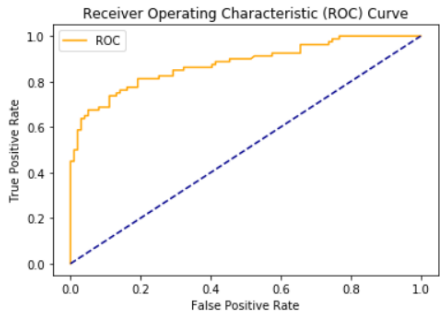


Supervised Learning

LOGISTIC REGRESSION

- Precision-recall curve (PRC) grafiği anlamı

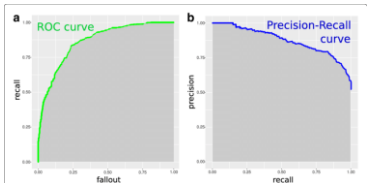


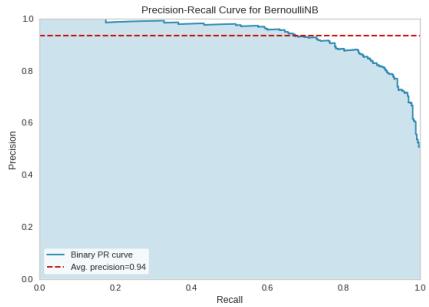


Supervised Learning

LOGISTIC REGRESSION

- Precision-recall curve (PRC) grafiği anlamı





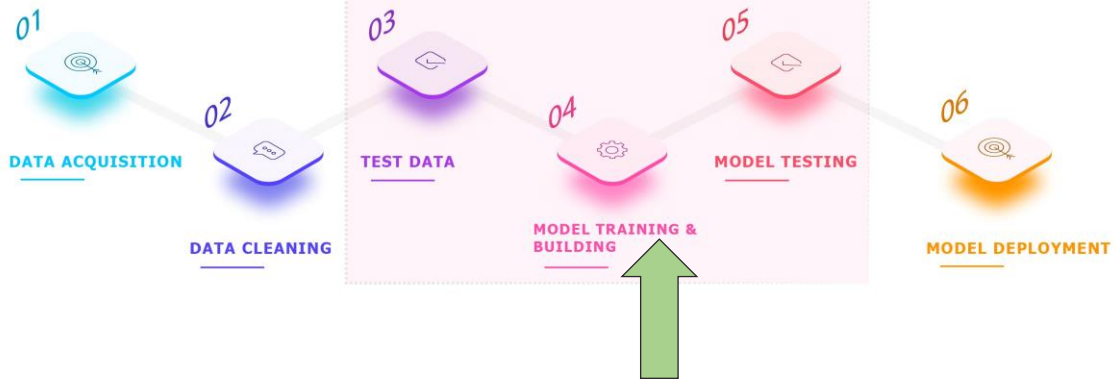
	Predicted class POSITIVE (spam ✉)	Predicted class NEGATIVE (normal 🔒)	
Actual class POSITIVE (spam ✉)	TRUE POSITIVE (TP) ✉✉ 320	FALSE NEGATIVE (FN) ✉🔒 43	$Recall = \frac{TP}{TP+FN} = \frac{320}{320+43} = 0.882$
Actual class NEGATIVE (normal ✉)	FALSE POSITIVE (FP) ✉✉ 20	TRUE NEGATIVE (TN) ✉🔒 538	
	$Precision = \frac{TP}{TP+FP} = \frac{320}{320+20} = 0.941$		

ROC: Balanced Data
PRC: Unbalanced data

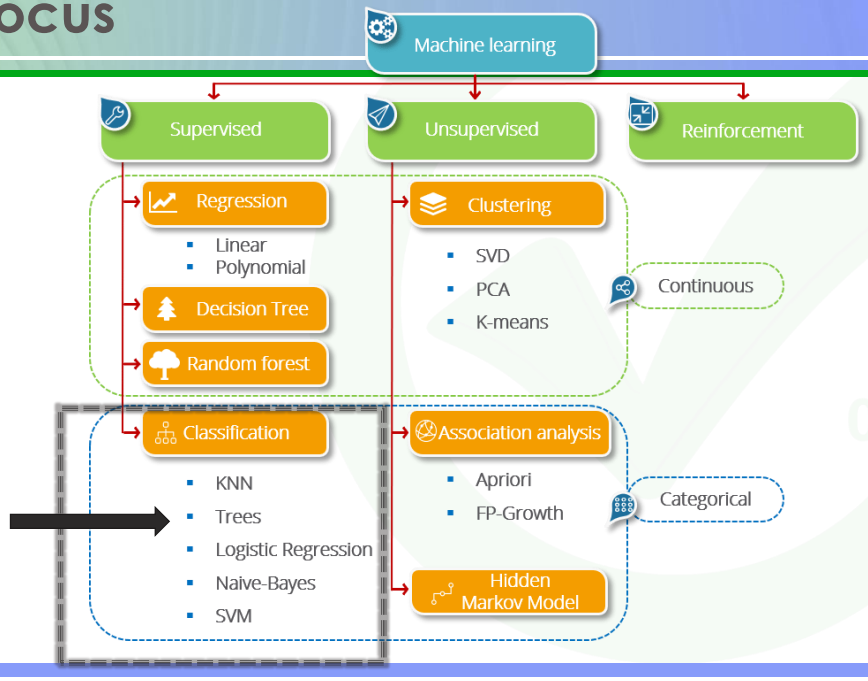
Where are we?

DATA SCIENCE

Machine Learning Course



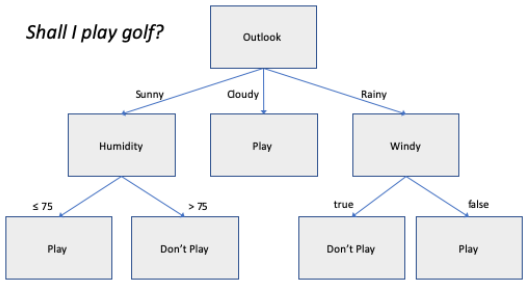
More focus

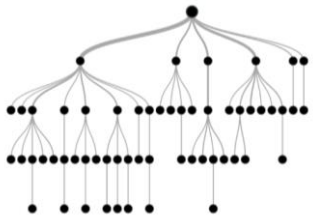


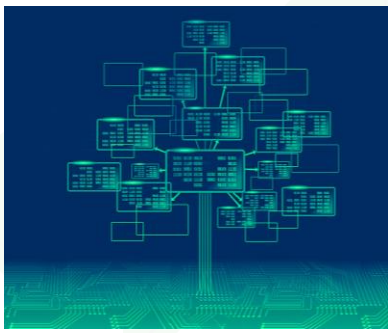
Supervised Learning – Decision Tree

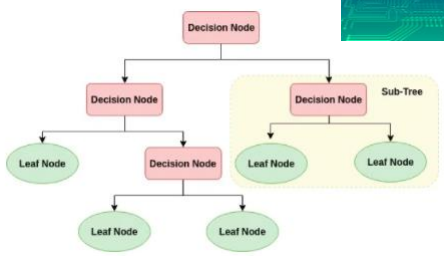
DECISION TREE (Karar Ağaçları Algoritmaları)

- Karar Ağacı ne demek ?
- Eşik değeri, Kök ve yapraklar..







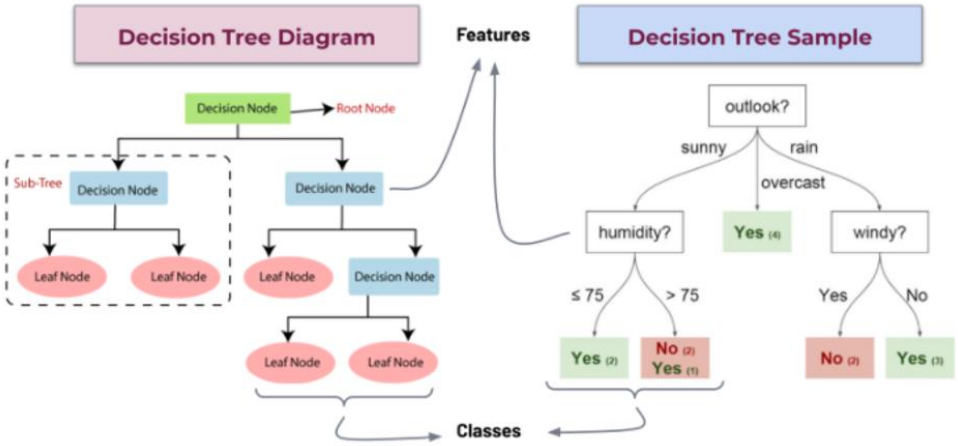


Grubun tüm elemanları aynı sınıf etiketine sahip olana kadar kümeleme işlemi derinlemesine devam eder.

Supervised Learning

DECISION TREE (Karar Ağaçları Algoritmaları)

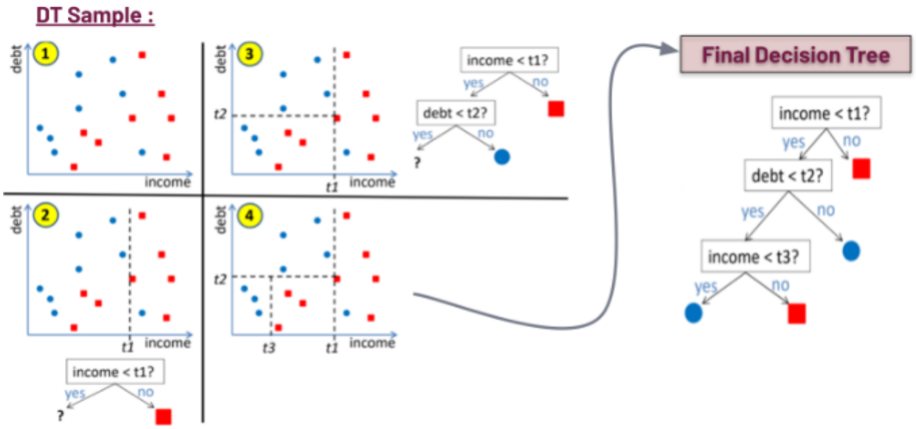
- Karar Ağacı ne demek ?
- Eşik değeri, Kök ve yapraklar..



Supervised Learning

DECISION TREE (Karar Ağaçları Algoritmaları)

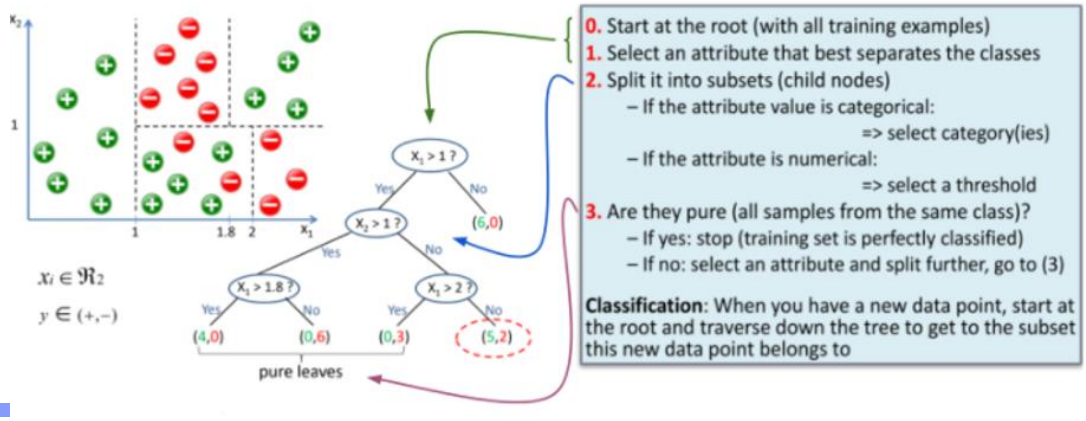
- Karar Ağacı ne demek ?
- Eşik değeri, Kök ve yapraklar..



Supervised Learning

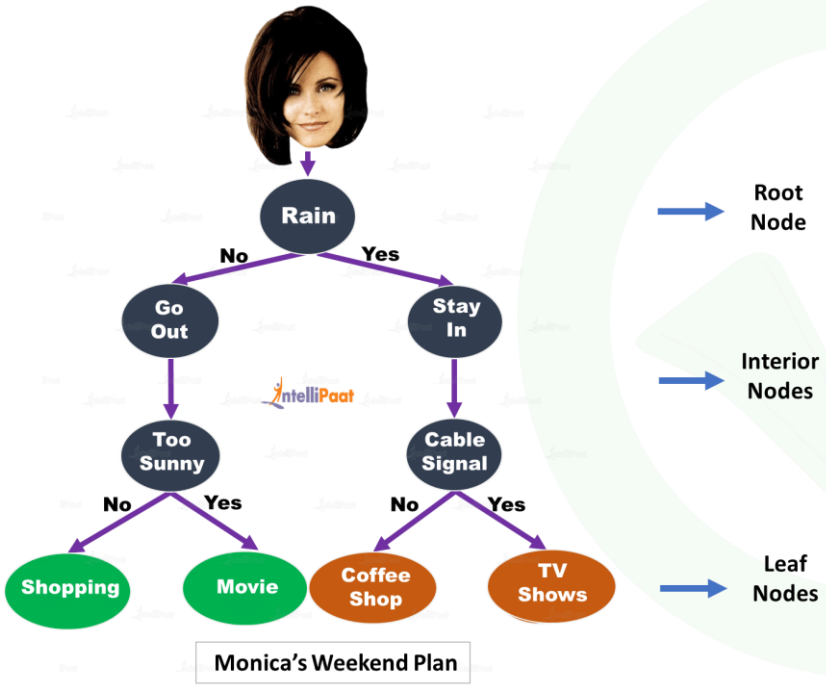
DECISION TREE (Karar Ağaçları Algoritmaları)

- Karar Ağacı İşlem sırası..
- Pür-saf olması için nasıl bölünmeli



0. Start at the root (with all training examples)
1. Select an attribute that best separates the classes
2. Split it into subsets (child nodes)
   - If the attribute value is categorical:
     => select category(ies)
   - If the attribute is numerical:
     => select a threshold
3. Are they pure (all samples from the same class)?
   - If yes: stop (training set is perfectly classified)
   - If no: select an attribute and split further, go to (3)

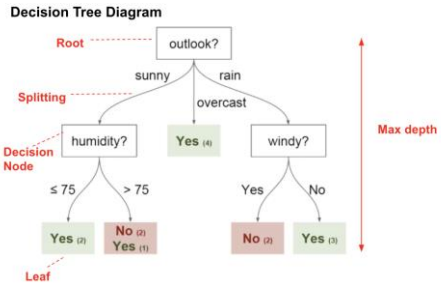
Classification: When you have a new data point, start at the root and traverse down the tree to get to the subset this new data point belongs to

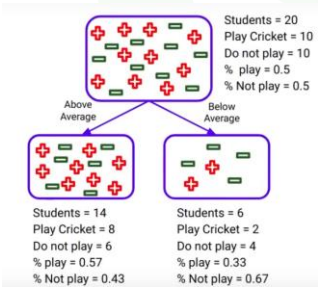


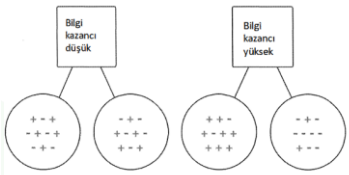
Supervised Learning

DECISION TREE (Karar Ağaçları Algoritmaları)

- Root ve feature seçimi..



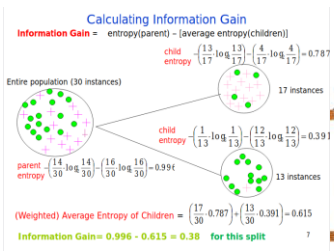


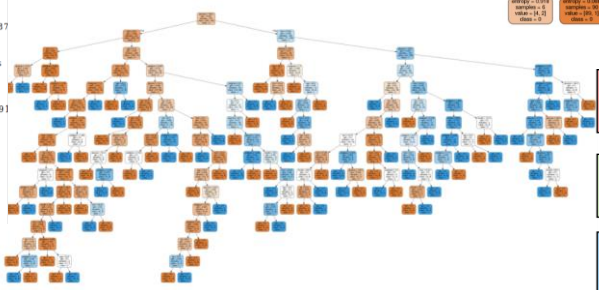


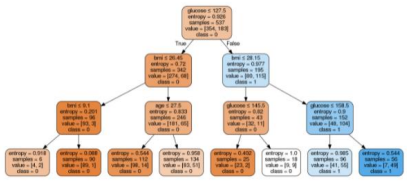
Supervised Learning

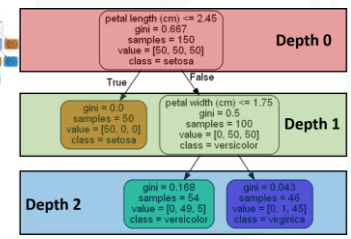
DECISION TREE (Karar Ağaçları Algoritmaları)

- Information gain..
  - Entropy - Gini indexes
  - Max. Depth



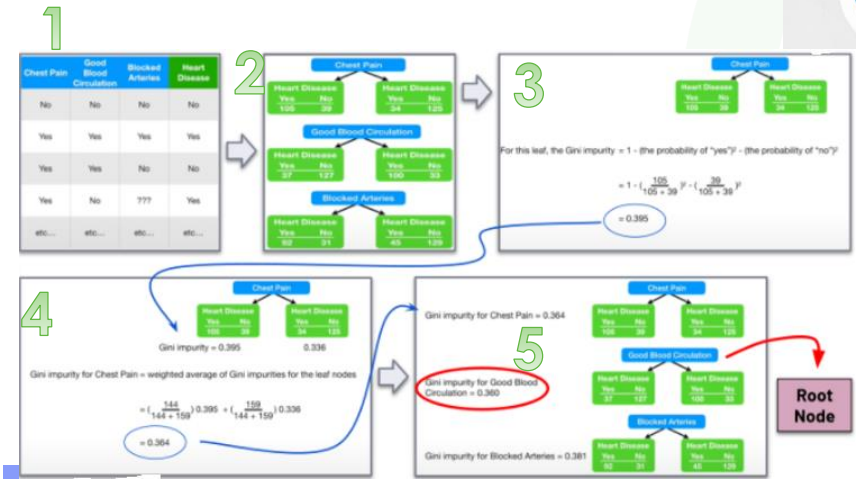


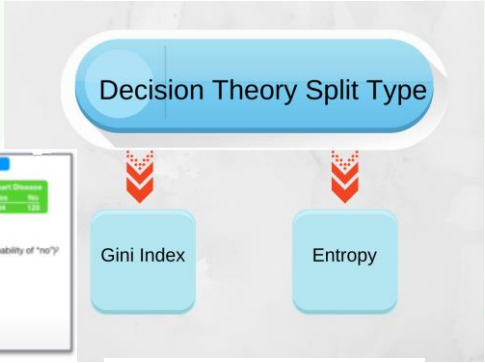




Supervised Learning - DT

- Information Gain Entropy
- Gini index (Gini Impurity Index)

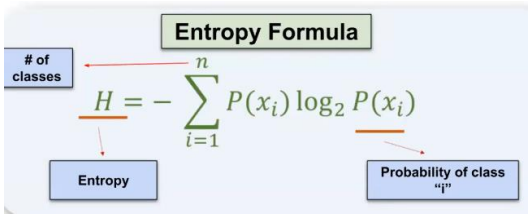


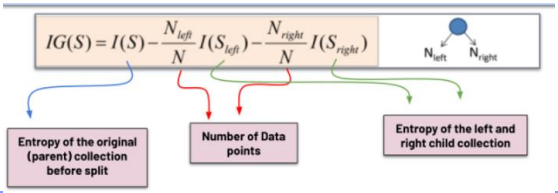


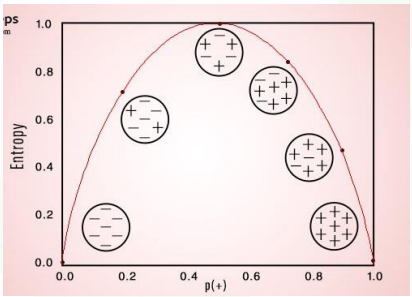
$$Gini = 1 - \sum_{i=1}^{C} (p_i)^2$$

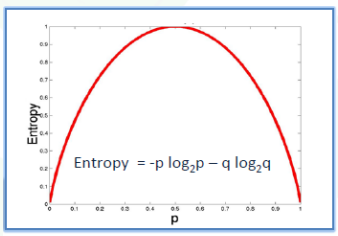
Supervised Learning -DT

- Information Gain Entropy
- Gini index (Gini Impurity Index)





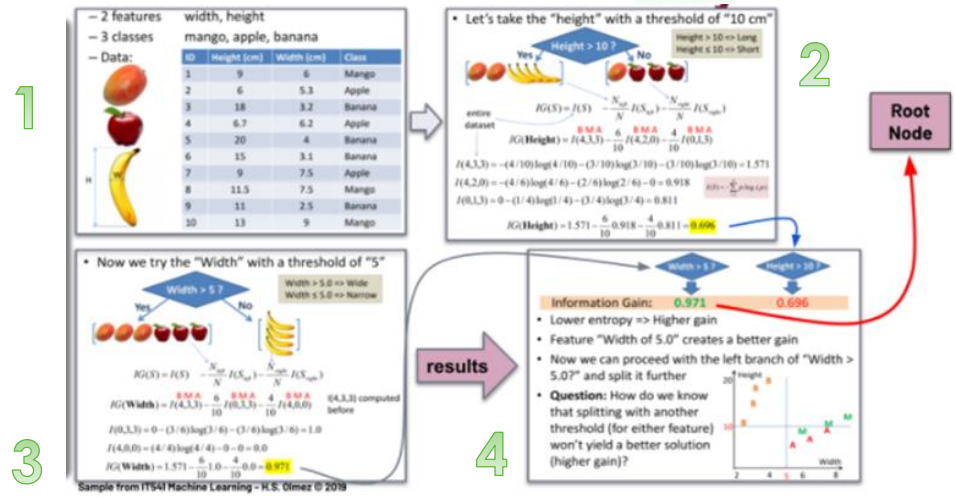




Entropy = -0.5 $\log_2$0.5 – 0.5 $\log_2$0.5 = 1

Supervised Learning –DT

Information Gain Entrophy Örnek



Supervised Learning

DECISION TREE (Karar Ağaçları Algoritmaları)

Hyperparameters (Criterion ve max_depth)

Criterion (Gini – Entrophy)

sklearn.tree.DecisionTreeClassifier

```
sklearn.tree.DecisionTreeClassifier(*, criterion='gini', splitter='best', max_depth=None, min_samples_split=2,
min_samples_leaf=1, min_weight_fraction_leaf=0.0, max_features=None, random_state=None, max_leaf_nodes=None,
min_impurity_decrease=0.0, min_impurity_split=None, class_weight=None, presort='deprecated
```

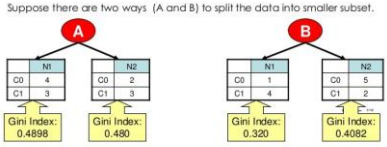
Splitting Binary Attributes (using Gini)

Example :

Parent: C0 6, C1 6, Gini = 0.5

Gini : $1-(6/12)^2-(6/12)^2 = 0.5$

Suppose there are two ways (A and B) to split the data into smaller subset.



By : Mohd. Noor Abdul Hamid, Ph.D (Universiti Utara Malaysia)

```
param_grid = {
    'max_depth': np.arange(3, 10),
    'criterion': ['gini', 'entropy'],
    'min_samples_split' : range(10,500,20)}
# Hiper parametre optimizasyon modeli:
grid_DT = GridSearchCV(DecisionTreeClassifier(), param_grid, cv=cv)
# Model oluşturma:
grid_DT.fit(X, y)
```

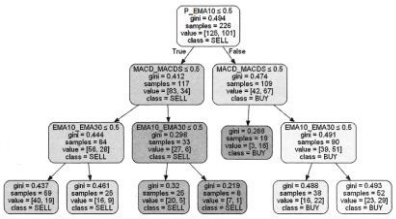
```
En İyi Parametreler: DecisionTreeClassifier(ccp_alpha=0.0, class_weight=None,
criterion='gini',
                       max_depth=3, max_features=None, max_leaf_nodes=None,
                       min_impurity_decrease=0.0, min_impurity_split=None,
                       min_samples_leaf=1, min_samples_split=30,
                       min_weight_fraction_leaf=0.0, presort='deprecated',
                       random_state=None, splitter='best')

Izgara Aramasındaki Optimum Hiperparametreler: {'criterion': 'gini', 'max_depth': 3,
'min_samples_split': 30}
```

Supervised Learning

DECISION TREE (Karar Ağaçları Algoritmaları)

- Hyperparameters (Criterion ve max_depth)
- Grafize edilmesi
- Karar ağacının yorumu



Impurity Criterion

Gini Index

$$I_G = 1 - \sum_{j=1}^{c} p_j^2$$

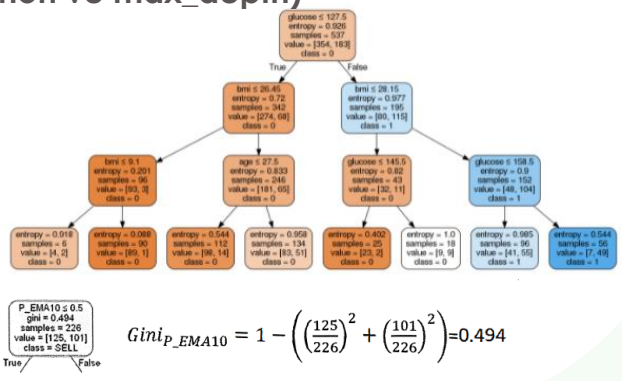
p_j: proportion of the samples that belongs to class c for a particular node

Entropy

$$I_H = -\sum_{j=1}^{c} p_j log_2(p_j)$$

p_j: proportion of the samples that belongs to class c for a particular node.

*This is the the definition of entropy for all non-empty classes ($p \neq 0$). The entropy is 0 if all samples at a node belong to the same class.



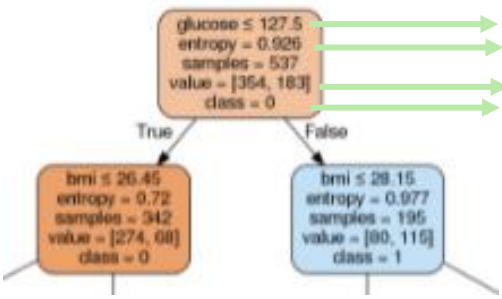
$$Gini_{P\_EMA10} = 1 - \left(\left(\frac{125}{226}\right)^2 + \left(\frac{101}{226}\right)^2\right) = 0.494$$

$$Ent_{P\_EMA10} = -\left(\left(\frac{125}{226}\right) \times \left(log_2\left(\frac{125}{226}\right)\right)\right) + \left(\left(\frac{101}{226}\right) \times \left(log_2\left(\frac{101}{226}\right)\right)\right) = 0.992$$

Supervised Learning

DECISION TREE

- Hyperparameters (Criterion ve max_depth)
- Grafize edilmesi
- Karar ağacının yorumu

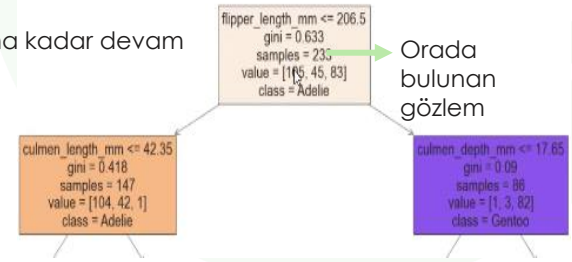


Seçilen feature ve threshold

Seçilen kriter

Aşağıda 0 olana kadar devam

Atanan sınıf

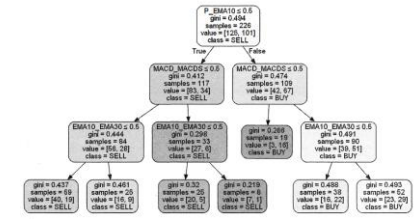


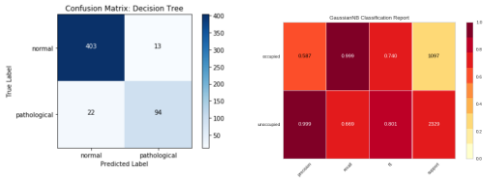
Orada bulunan gözlem

Supervised Learning

DECISION TREE

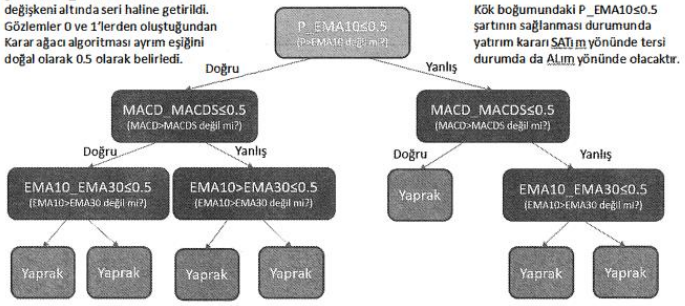
- Karar ağacının yorumu





P>EMA10 ise 1, değilse 0 atandı. 1 Al sinyali, 0 sat sinyali olmak üzere tüm gözlemler P_EMA10 isimli öznitelik değişkeni altında seri haline getirildi. Gözlemler 0 ve 1'lerden oluştuğundan Karar ağacı algoritması ayrım eşiğini doğal olarak 0.5 olarak belirledi.

0.5'in altında sadece SAT sinyalleri ve üzerinde de AL sinyalleri bulunmaktadır.
Kök boğumundaki P_EMA10≤0.5 şartının sağlanması durumunda yatırım kararı SATı m yönünde tersi durumda da ALım yönünde olacaktır.



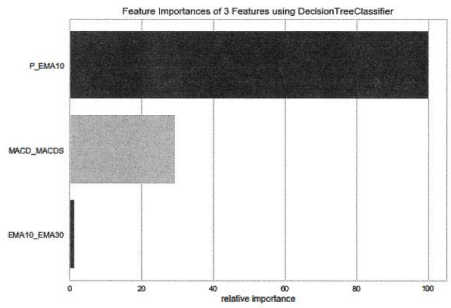
- P_EMA10≤0.5 ise ve MACD_MACDS≤0.5 ise SAT
- P_EMA10≤0.5 ise ve MACD_MACDS≤0.5 ise ve EMA10_EMA30 ≤0.5 ise SAT
- P_EMA10≤0.5 ise ve MACD_MACDS≤0.5 değilse ve EMA10_EMA30 ≤0.5 ise SAT
- P_EMA10≤0.5 ise ve MACD_MACDS≤0.5 değilse ve EMA10_EMA30 ≤0.5 değilse SAT

- P_EMA10≤0.5 değilse ve MACD_MACDS≤0.5 ise AL
- P_EMA10≤0.5 değilse ve MACD_MACDS≤0.5 değilse ve EMA10_EMA30 ≤0.5 ise AL
- P_EMA10≤0.5 değilse ve MACD_MACDS≤0.5 değilse ve EMA10_EMA30 ≤0.5 değilse AL

Supervised Learning

DECISION TREE

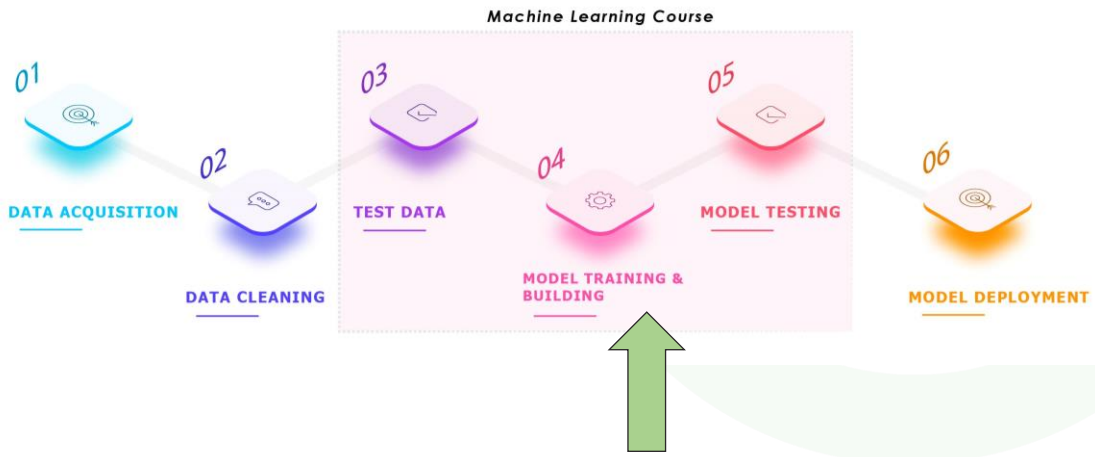
- Feature impoortance'a bakılması
- Decision tree son hususlar
  - Avantaj
  - Dezavatajları



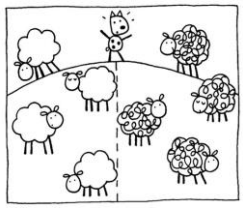
- **Exploratory Data Analysis and Visualization**
- **Machine Learning**
  - **Train | Test Split**
    - X_train, X_test, y_train, y_test = train_test_split()
  - **Scalling (if needed)**
    - scaler = scaler_name()
    - scaler.fit_transform(X_train)
    - scaler.transform(X_test)
  - **Modelling**
    - 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**

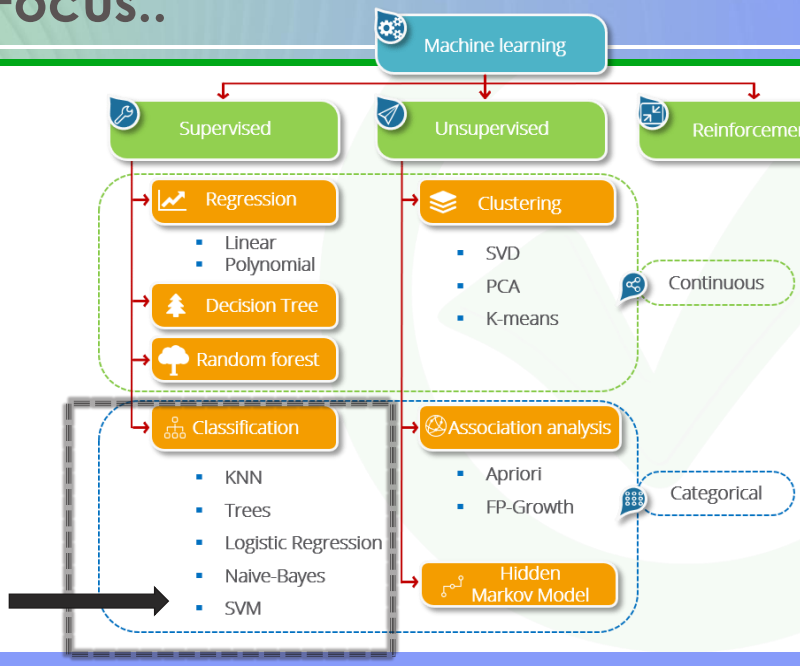
Run Navigation

Where are we?



More Focus..



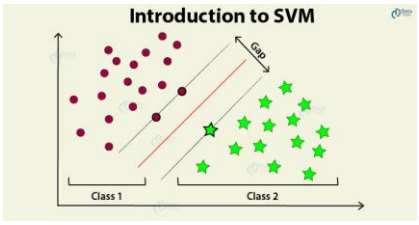


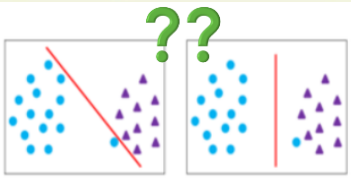
Supervised Learning

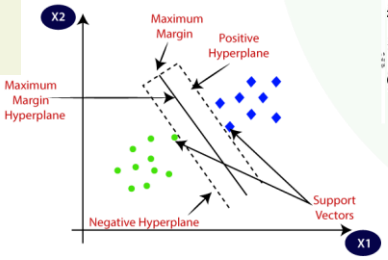
SUPPORT VECTOR MACHINES (SVM)

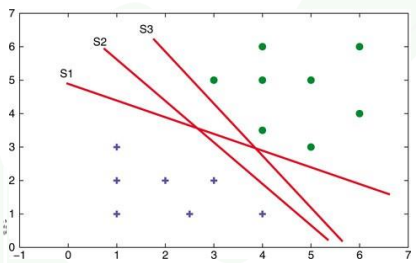
(Karar Destek Makineleri)

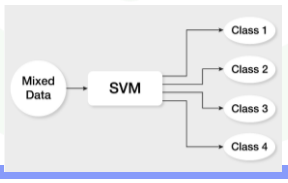
- Datayı 2 sınıfa ayıran en iyi hiperdüzlem nasıl seçilir





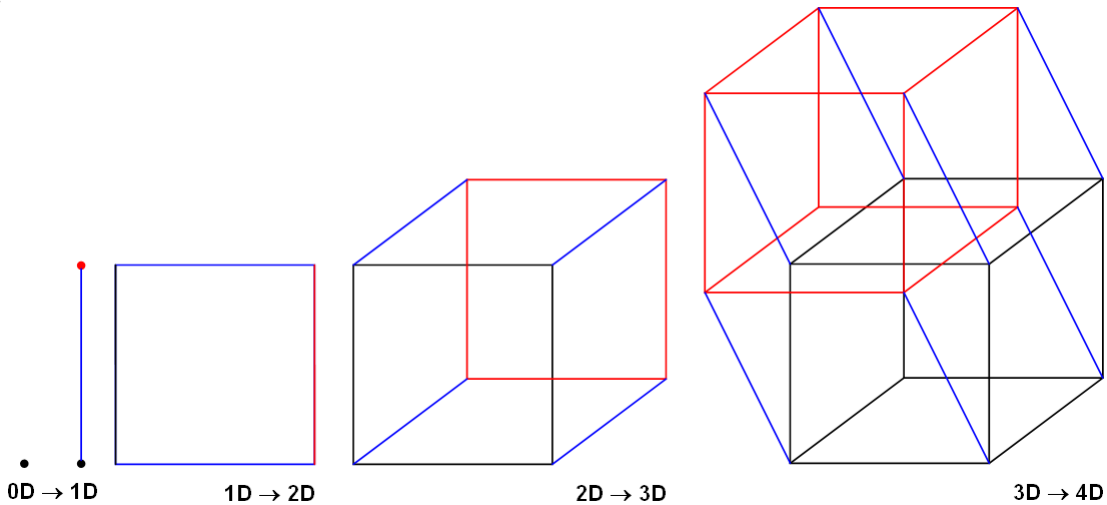







which categorized new examples.
In one dimensional space.
SUBSCRIBE

Boyutlar

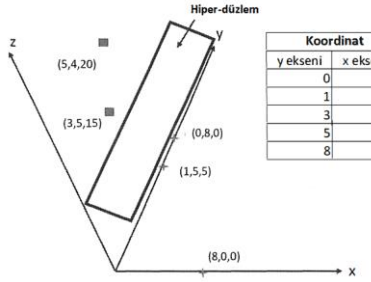

0D → 1D
1D → 2D
2D → 3D
3D → 4D

Supervised Learning

SUPPORT VECTOR MACHINES (SVM)

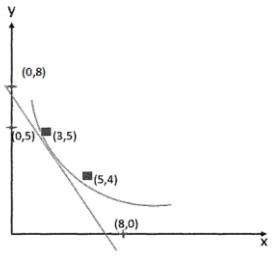
- Kernel Trick adımlar

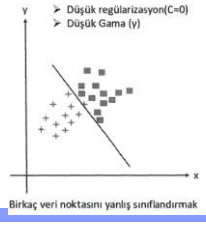
Koordinat		1 Boyutlu	2 Boyutlu	3 Boyutlu
y ekseni	x ekseni	$x+y$	$x \times y$	x^2+y^2
0	8	8	0	64
1	5	6	5	26
3	5	8	15	34
5	4	9	20	41
8	0	8	0	64

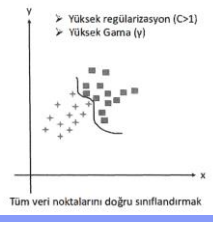


Koordinat		1 Boyutlu	2 Boyutlu	3 Boyutlu
y ekseni	x ekseni	$x+y$	$x \times y$	x^2+y^2
0	8	8	0	64
1	5	6	5	26
3	5	8	15	34
5	4	9	20	41
8	0	8	0	64

Veri noktalarının y ve x koordinatlarına (bunlardan türetilen) üçüncü eksen koordinatı eklenir. Örneğin, birinci gözlemin koordinatı (0,8) iken üç boyutlu koordinatı (0,8,0) olur.





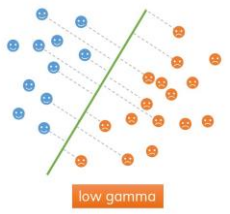


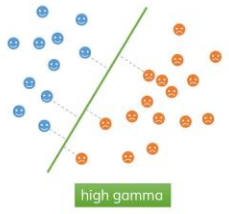
Supervised Learning

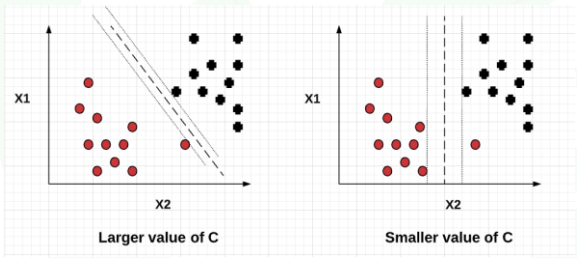
SUPPORT VECTOR MACHINES (SVM)

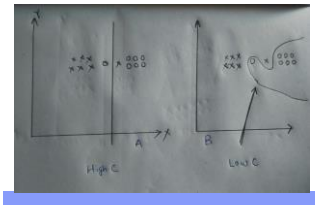
- Kernel Trick adımlar

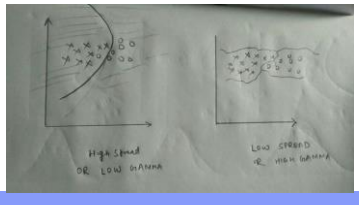
	Large Gamma	Small Gamma	Large C	Small C
Variance	Low	High	High	Low
Bias	High	Low	Low	High

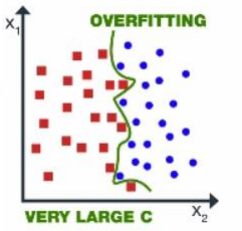


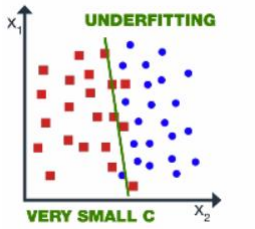


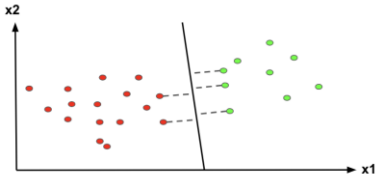






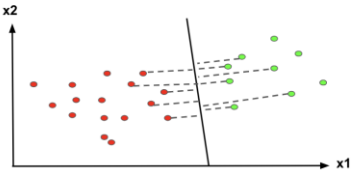






High Gamma

- only near points are considered.



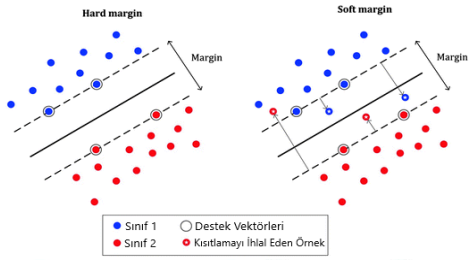
Low Gamma

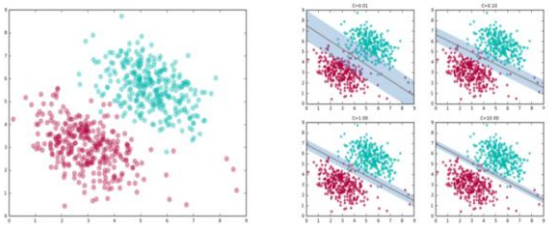
- far away points are also considered

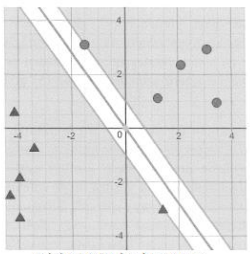
Supervised Learning

SUPPORT VECTOR MACHINES (SVM)

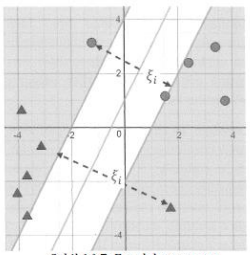
- Soft Margin - C hyperparameter



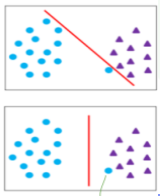




Şekil 16.6: Keskin kenar payı



Şekil 16.7: Esnek kenar payı



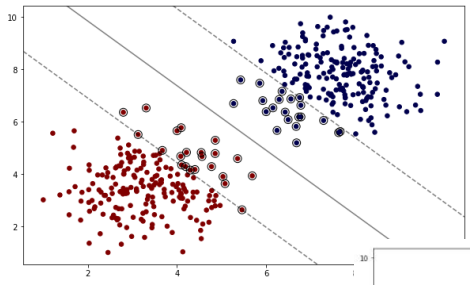
Hard margin: çok dar bir marjinle doğrusal biçimde ayrılmış veriler

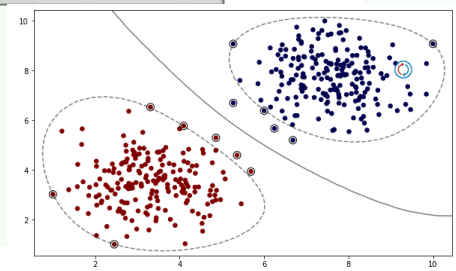
Soft margin: büyük margin, az miktardaki yanlış sınıflandırma durumunda ayırma biçimi olarak daha iyidir

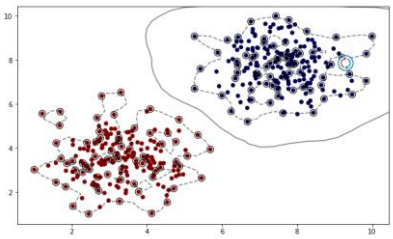
Supervised Learning

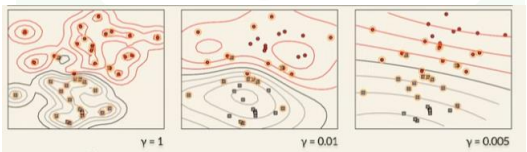
SUPPORT VECTOR MACHINES (SVM)

- Soft Margin - C hyperparameter



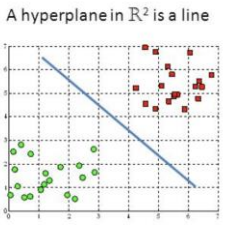


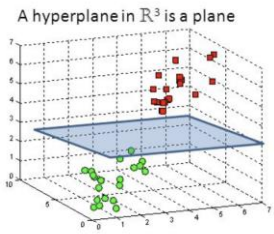


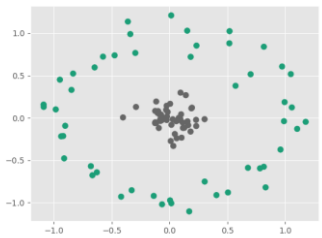


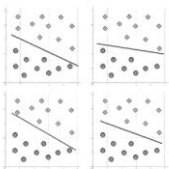
Supervised Learning - SVM

- Kernel Trick ve Soft Margin
- Doğrusal olmayan durumlarda SVM 2 sınıfı nasıl ayırır ?

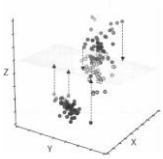




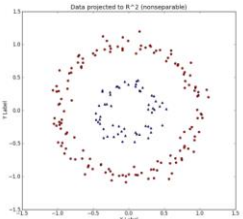


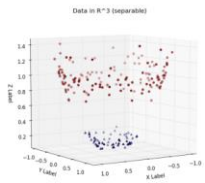


Şekil 16.1: İki boyutlu uzayda doğru ile doğrusal ayırma

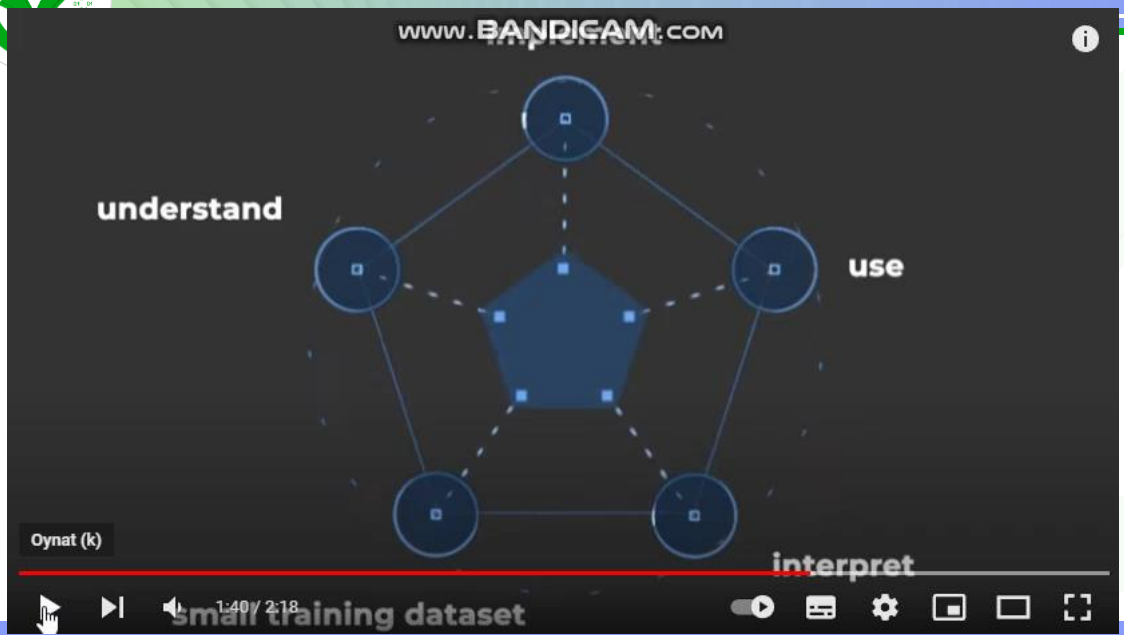


Şekil 16.2: Üç boyutlu uzayda düzlem ile doğrusal ayırma





SVM – Kernel



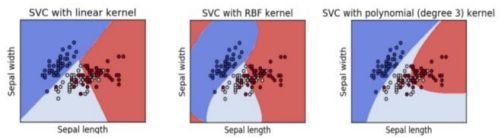
Supervised Learning - SVM

SUPPORT VECTOR MACHINES (SVM)

- Kernel çeşitleri
  - Linear Kernels
  - Polynomial Kernels
  - RBF Kernels

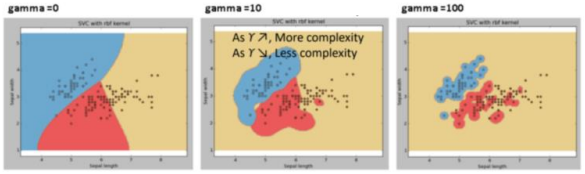
"kernel" *parameter:* (default="rbf")
{'linear', 'poly', 'rbf', 'sigmoid', 'precomputed'}

Specifies the kernel type to be used in the algorithm.



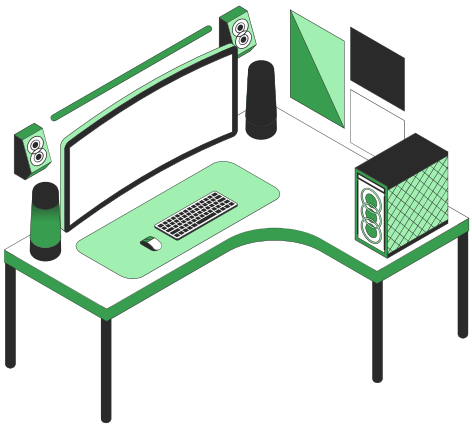
"gamma" *parameter:* (default = 1)

Kernel coefficient for '**rbf**', '**poly**' and '**sigmoid**'.



Bu dersi anladım..

Everythig is clear ?



Do you have any questions?

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