BATCH : **B 150** Data Science
LESSON : **Machine Learning**
DATE : 15.07.2023
SUBJECT : **Supervised Learning**
Classification - KNN

techproeducation
techproeducation
techproeducation
techproeducation
techproedu

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

MACHINE LEARNING - 3

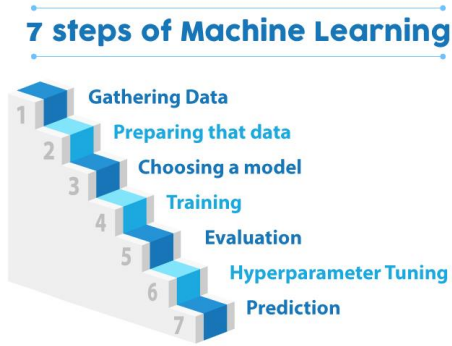
Machine Learning

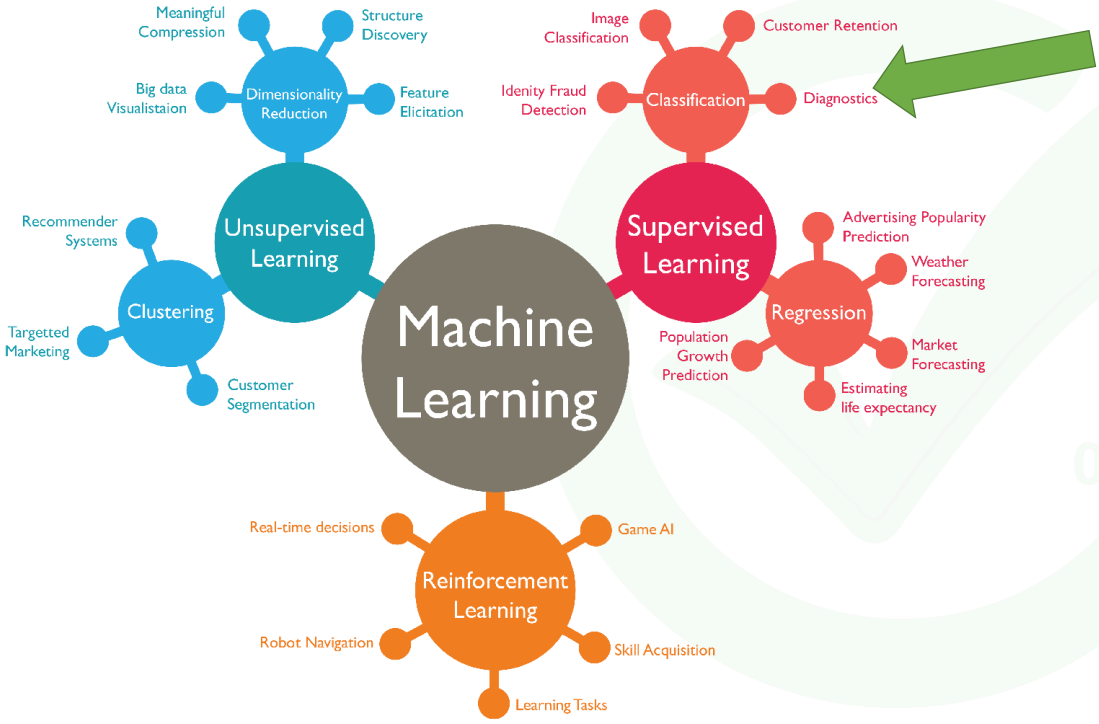
Makine Öğrenmesi – 3

Overall Table of Contents

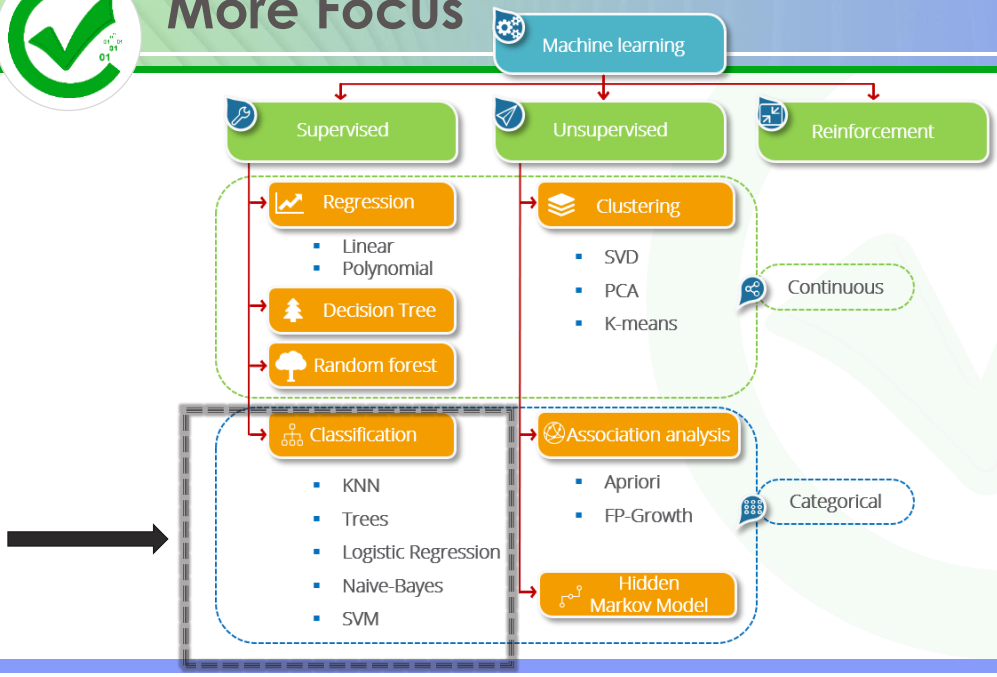
General Content

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



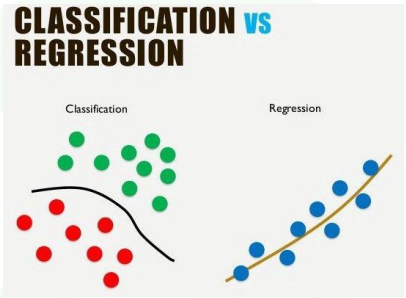


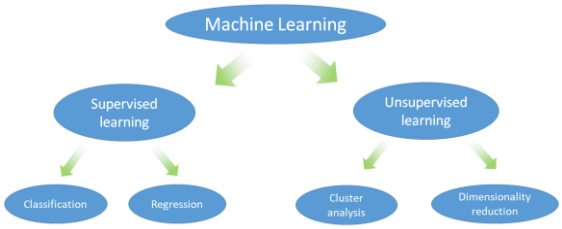
More Focus

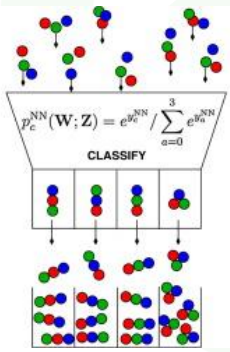


SUPERVISED LEARNING - Classification

Sınıflandırma



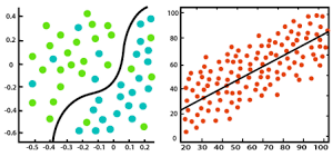


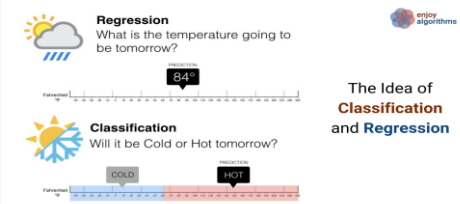


Supervised Learning

Classification 'a Giriş (Sınıflandırma)

- Regresyonda neler gördük
- Supervised Learning in 2. tipi olarak Classification
- Regression vs Classification
- Target için Kategorik Sınıflandırma vardır



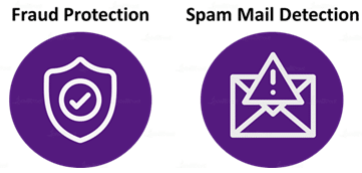


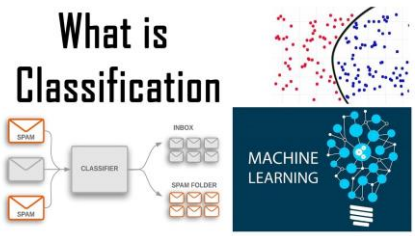
!! **Regresyonda** target hedef değişkenin **sayısal değerlerini**; **sınıflandırmada** ise target değişkenin ait olduğu **sınıfları (ya da "etiketi")** tahmin eden modelleri oluşturmaya çalışırız.

Supervised Learning

Classification ‘a Giriş

- Classification un hayattaki kullanım alanları
- Binary Classification



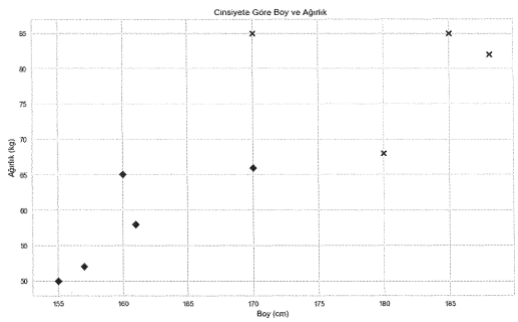


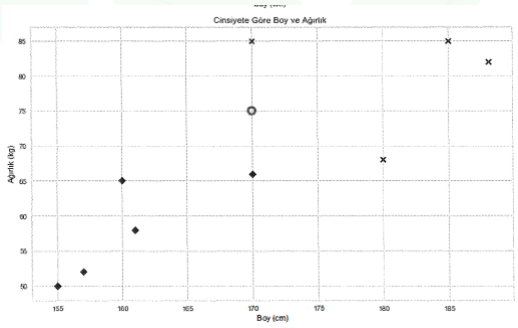
Boy	Kilo	Cinsiyet
160	65	K
170	85	E
185	85	E
188	82	E
155	50	K
161	58	K
180	68	E
157	52	K
170	66	K

Supervised Learning

Classification ‘a Giriş

- Binary Classification
- Kategori tahmini
- «Yakınlık» kavramı - ‘ara mesafe ölçümü’

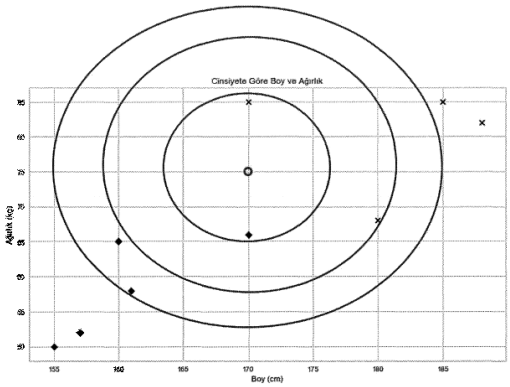


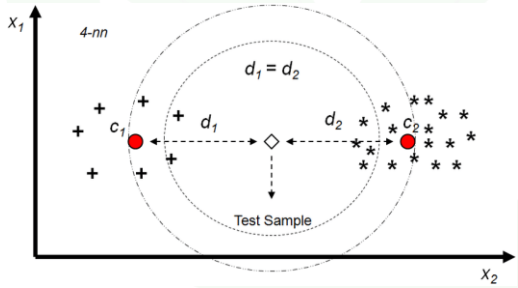


Supervised Learning

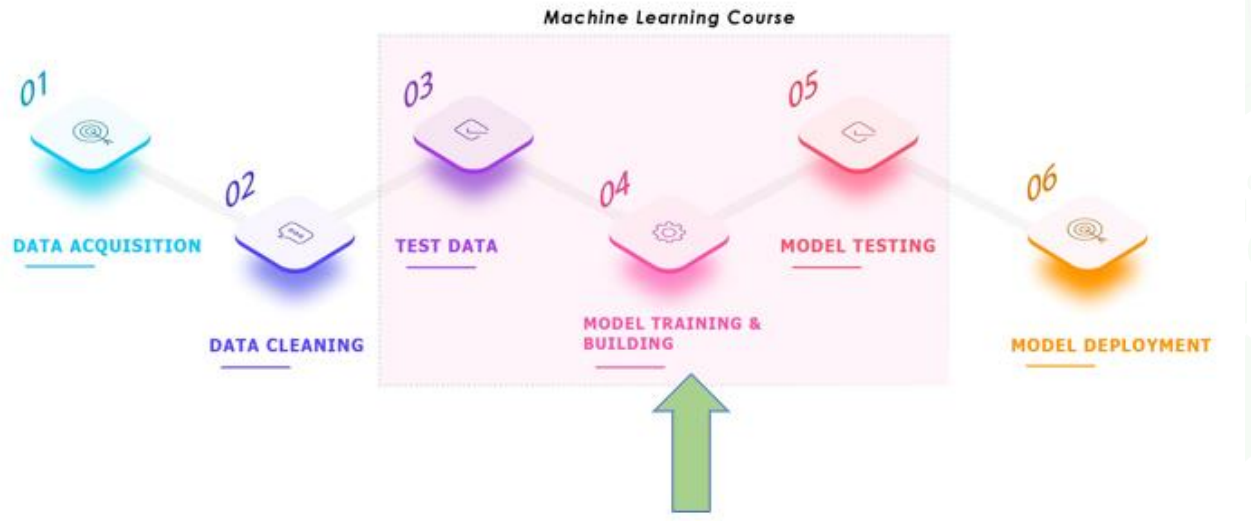
Classification 'a Giriş

- «**Yakınlık**» kavramı - «**ara mesafe ölçümü**»
- En yakın komşu kavramı





Where are we?

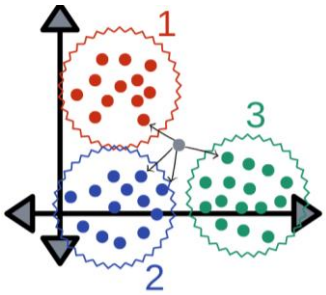


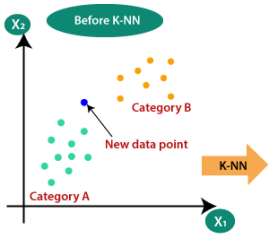
Supervised Learning

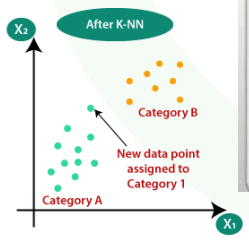
K Nearest Neighbour-KNN Algoritması

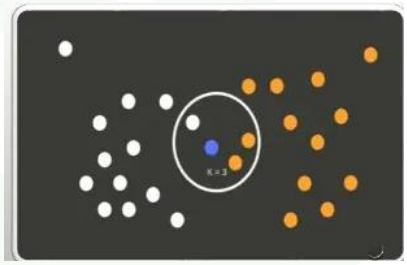
K-EN YAKIN KOMŞU Algoritması

- Logistic Regression ' dan bir adım öncesinde
- Classification için en basit yol olarak KNN
- En yakın komşu sayısı





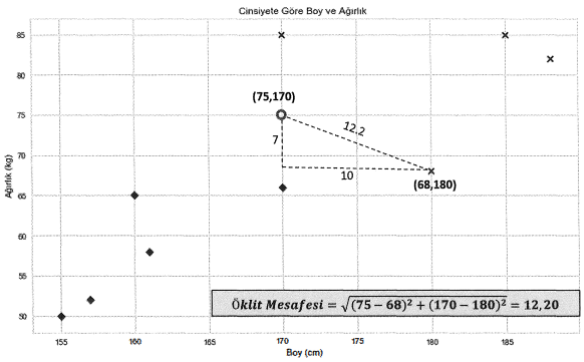


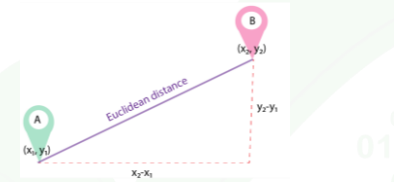


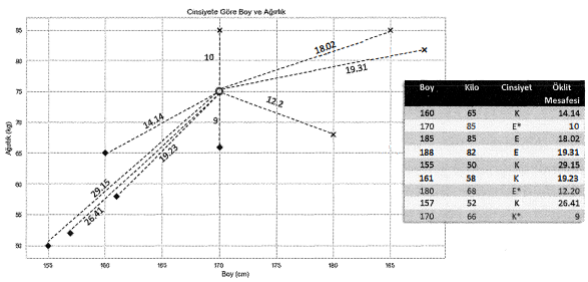
Supervised Learning

K Nearest Neighbour-KNN Algoritması

- Öklid mesafesi





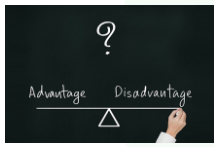


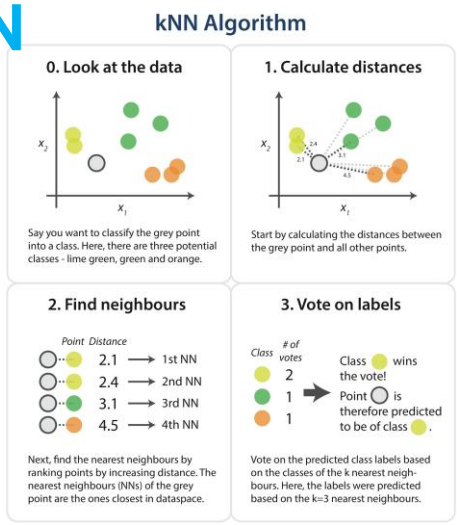
Boy	Kilo	Cinsiyet	Öklit Mesafesi
160	65	K	14.14
170	85	E*	10
185	85	E	18.02
188	82	E	19.31
155	50	K	29.15
161	58	K	19.23
180	68	E*	12.20
157	52	K	26.41
170	66	K*	9

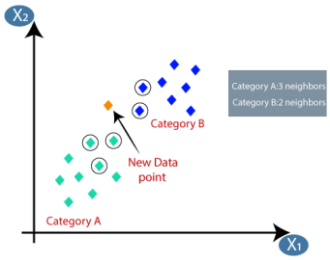
Supervised Learning

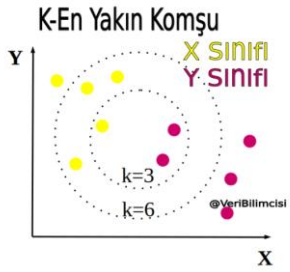
K Nearest Neighbour-KNN Algoritması

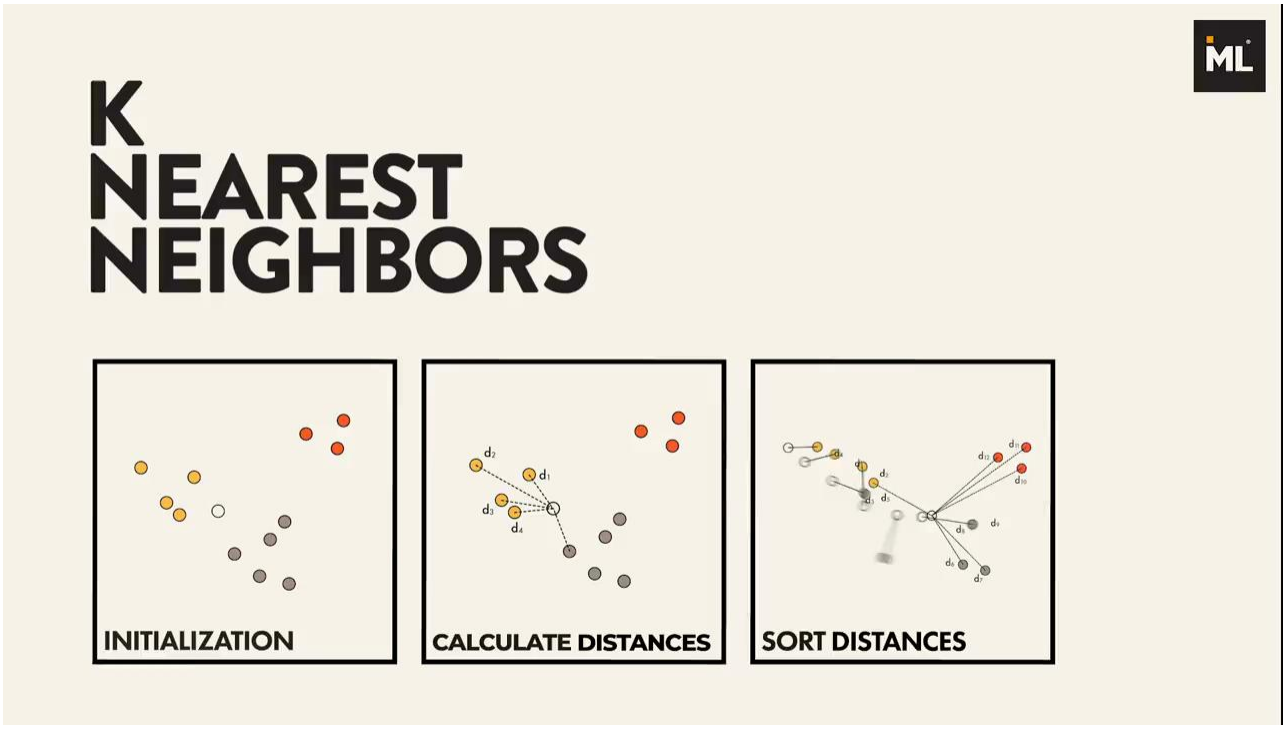
- KNN avantajları
- KNN dezavantajları





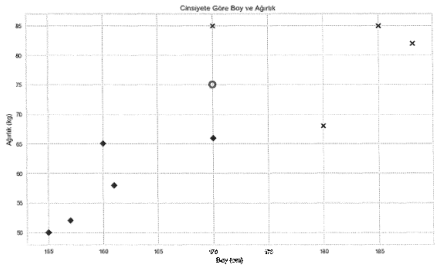


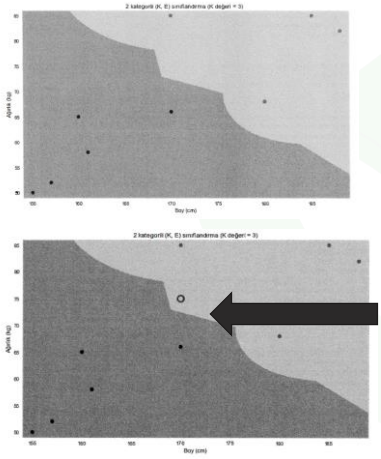


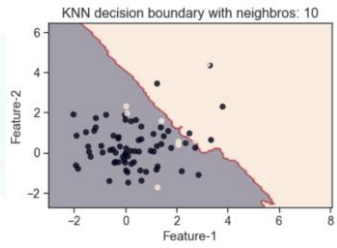


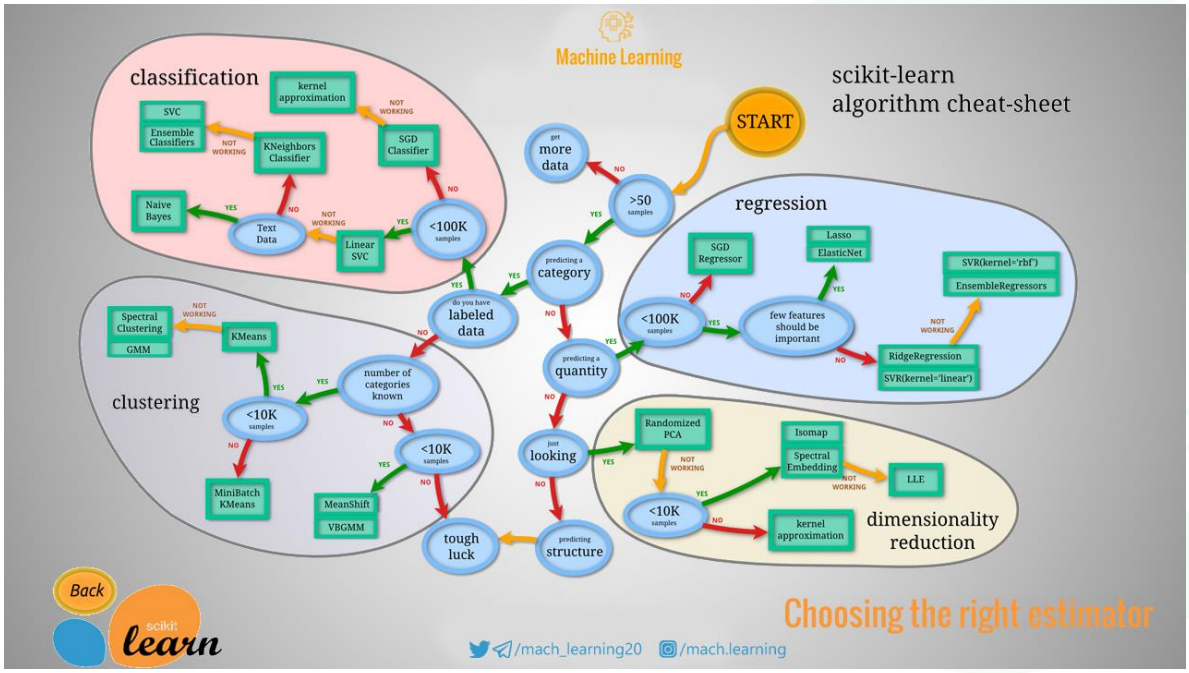
K Nearest Neighbour-KNN Algoritması

Decision boundary kavramı









Supervised Learning

Evaluation Metrics for Classification Problems
(Performans Ölçütleri)

very Important

- Regresyon kriterleri nasıl olurdu burada?
- Confusion metrics kavramı

Regression
- MSPE
- MSAE
- R Square
- Adjusted R Square

Classification
- Precision-Recall
- ROC-AUC
- Accuracy
- Log-Loss

		Predicted Class: Positive	Predicted Class: Negative	
Actual Class	Positive	True Positive (TP)	False Negative (FN) Type II Error	Sensitivity $\frac{TP}{(TP+FN)}$
	Negative	False Positive (FP) Type I Error	True Negative (TN)	Specificity $\frac{TN}{(TN+FP)}$
		Precision $\frac{TP}{(TP+FP)}$	Negative Predictive Value $\frac{TN}{(TN+FN)}$	Accuracy $\frac{TP+TN}{(TP+TN+FP+FN)}$

Supervised Learning

Evaluation Metrics for Classification Problems
- Confusion metrics kavramı: TN-FN-FP-TP

2-SINIFLI (CLASS) SINIFLANDIRMA MODELİ İÇİN HATA MATRİSİ

Veri setindeki hedef değişkenin gerçek sınıfları: Negatif, Pozitif

Hedef değişkenin tahmin edilen sınıfları: Negatif, Pozitif

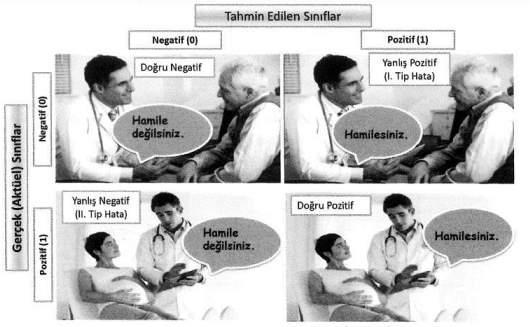
2-SINIFLI (CLASS) SINIFLANDIRMA MODELİ İÇİN HATA MATRİSİ

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

Supervised Learning

Evaluation Metrics for Classification Problems

Confusion metrics kavramı: TN-FN-FP-TP



Predicted

Actual	0	1
0	30	12
1	8	56

SINIFLANDIRMA MODELLERİ İÇİN PERFORMANS DEĞERLENDİRME ÖLÇÜTLERİ

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

Doğruluk (Accuracy): Doğru tahmin edilen hedef değişkenlerin tüm hedef değişkenlerine oranıdır. Model hedef değişkenleri ne kadar doğrulukla tahmin ediyor?

$$Doğruluk(Accuracy) = \frac{DP + DN}{DP + DN + YP + YN}$$

Kesinlik (Precision): Doğru pozitif olarak tahmin edilen gözlemlerin tüm pozitif gözlemlere oranıdır. Doğru bir şekilde pozitif tahmin edilen gözlemlerin gerçekte ne kadarı doğrudur?

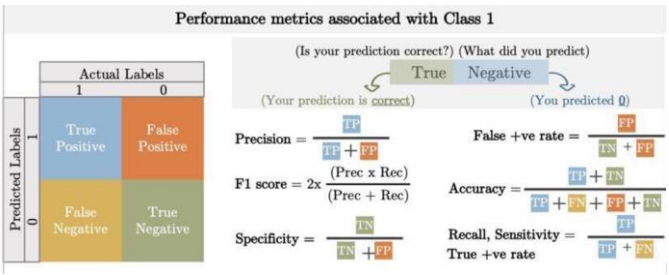
$$Kesinlik\,(Precision) = \frac{DP}{DP + YP}$$

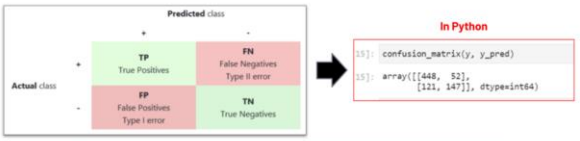
Duyarlılık (Recall): Doğru bir şekilde pozitif olarak tahmin edilen gözlemlerin ne kadar başarılı tahmin edildiğini gösterir.

$$Duyarlılık\,(Recall) = \frac{DP}{DP + YN}$$

F1 Skoru: Kesinlik (precission) ve Duyarlılığın (Recall) harmonik ortalamasıdır.

$$F1 = 2 * \frac{(Kesinlik * Duyarlılık)}{(Kesinlik * Duyarlılık)}$$





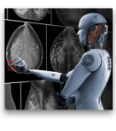
Actual Cancer: 1 1 0 0 1 0 0 0 0 1 ....
Predicted Cancer: 0 1 1 0 1 0 0 0 0 0 ....
FN TP FP TN TP TN TN TN TN FN...

Actual Labels: 1 1 0 1 0 0 0 1 1 0
Predicted Labels: 0 1 1 1 0 0 0 1 1 1

All correctly predicted values (7) / All predicted values (10) X 100 → Accuracy = 70 %

Why is Accuracy not a good metric?

Cancer Detection Example:



Actual Cancer: 1 1 0 0 1 0 0 0 0 1 ....
Predicted Cancer: 0 1 1 0 1 0 0 0 0 0 ....

All correctly predicted values (60) / All predicted values (63) X 100 → Accuracy = 95 % (PERFECT ????)

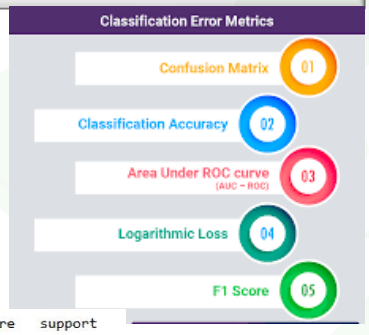
Accuracy is very high, but missed 2 actual patient.

Supervised Learning

Evaluation Metrics for Classification Problems

- Confusion metrics kavramı: TN-FN-FP-TP

Gerçek/Actual Veri				Tahmin		
Boy	Kilo	Cinsiyet	Etiket	Tahmin Edilen Cinsiyet	Etiket	Hata Türü
170	75	E	0	E	0	DN
180	95	E	0	E	0	DN
160	50	K	1	K	1	DP
165	62	K	1	K	1	DP
167	88	K	1	E	0	YN



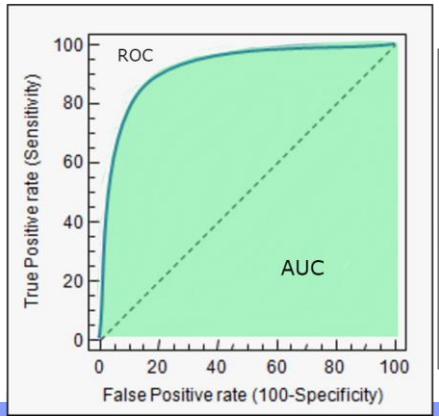
Hata Matrisi		Tahmin Edilen Etiketler	
Negatif (0)		Pozitif (1)	
Gerçek/Aktüel Veri	Negatif (0)	DN=2	YP=0
	Pozitif (1)	YN=1	DP=2

```
              precision    recall  f1-score   support

           E       0.67      1.00      0.80         2
           K       1.00      0.67      0.80         3

   micro avg       0.80      0.80      0.80         5
   macro avg       0.83      0.83      0.80         5
weighted avg       0.87      0.80      0.80         5
```

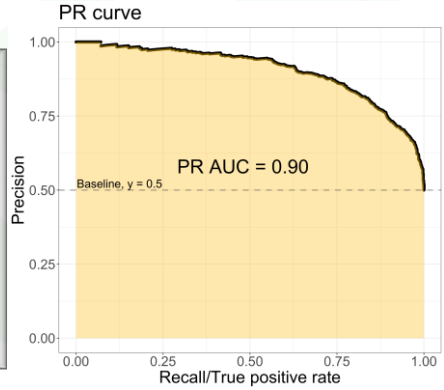
Supervised Learning

Evaluation Metrics for Classification Problems

- ROC Curve (Receiver Operating Characteristics)
- AUC Area (Area Under the Curve)
- (ROC Eğrisi ve AUC Alanı)



AUC'un olabildiğinde yüksek olması (1 'e yakın olmasını) istiyoruz
AUC ne kadar yüksekse model o kadar negatif (0) durumları negatif; pozitif (1) durumları da pozitif öngörüyor demektir.



$$\text{True Positive Rate} = \text{Sensitivity} = \frac{\text{True Positives}}{\text{True Positives} + \text{False Negatives}}$$

Receiver
Operator
Characteristics

Area
Under
Curve

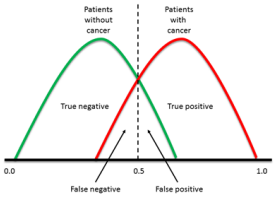
True Positive Rate (Sensitivity)

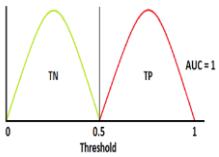
The x-axis shows the **False Positive Rate**, which is the same thing as **1 - Specificity**.

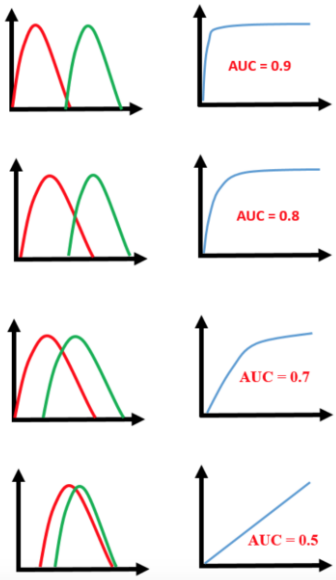
False Positive Rate (1 - Specificity)

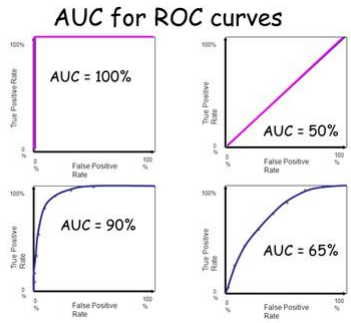
$$\text{False Positive Rate} = (1 - \text{Specificity}) = \frac{\text{False Positives}}{\text{False Positives} + \text{True Negatives}}$$

Picture credit :StatQuest





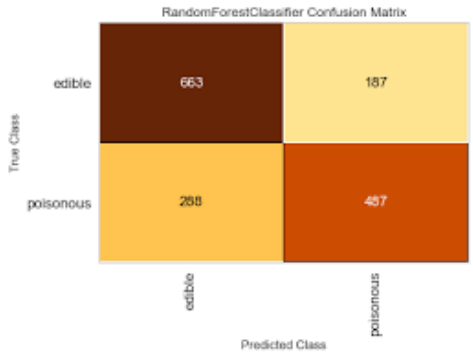


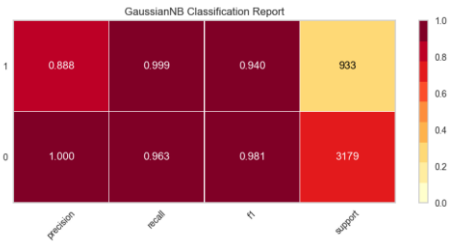


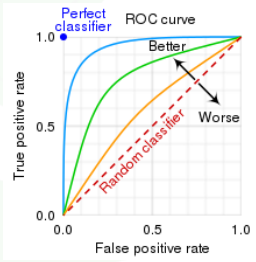
Supervised Learning

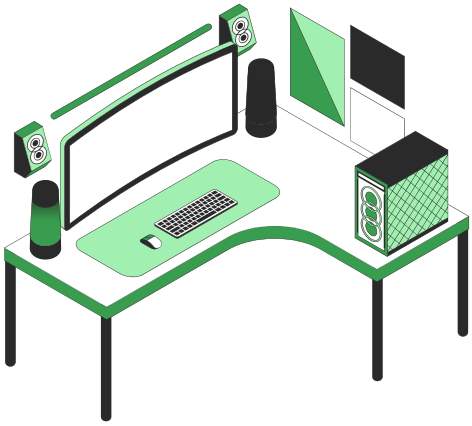
Evaluation Metrics for Classification Problems

- Yellowbrick ile confusion matrics









Do you have any questions?

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