BATCH : **B 84** Data Science

LESSON : **Machine Learning**

DATE : 12.09.2022

SUBJECT : **Supervised Learning**

techproeducation

techproeducation

techproeducation

techproeducation

techproedu

MACHINE LEARNING - 3

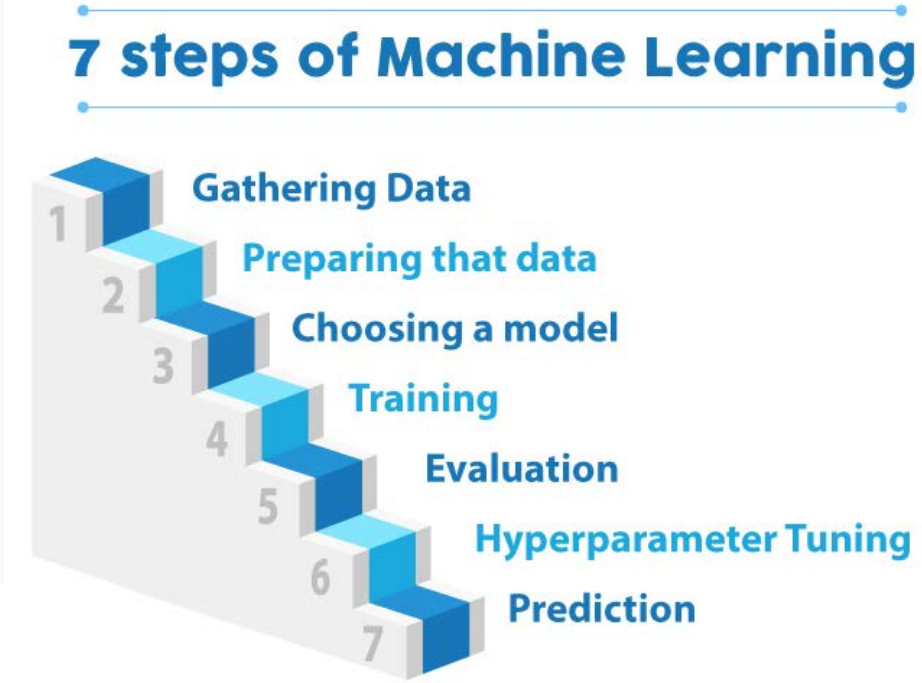
Machine Learning

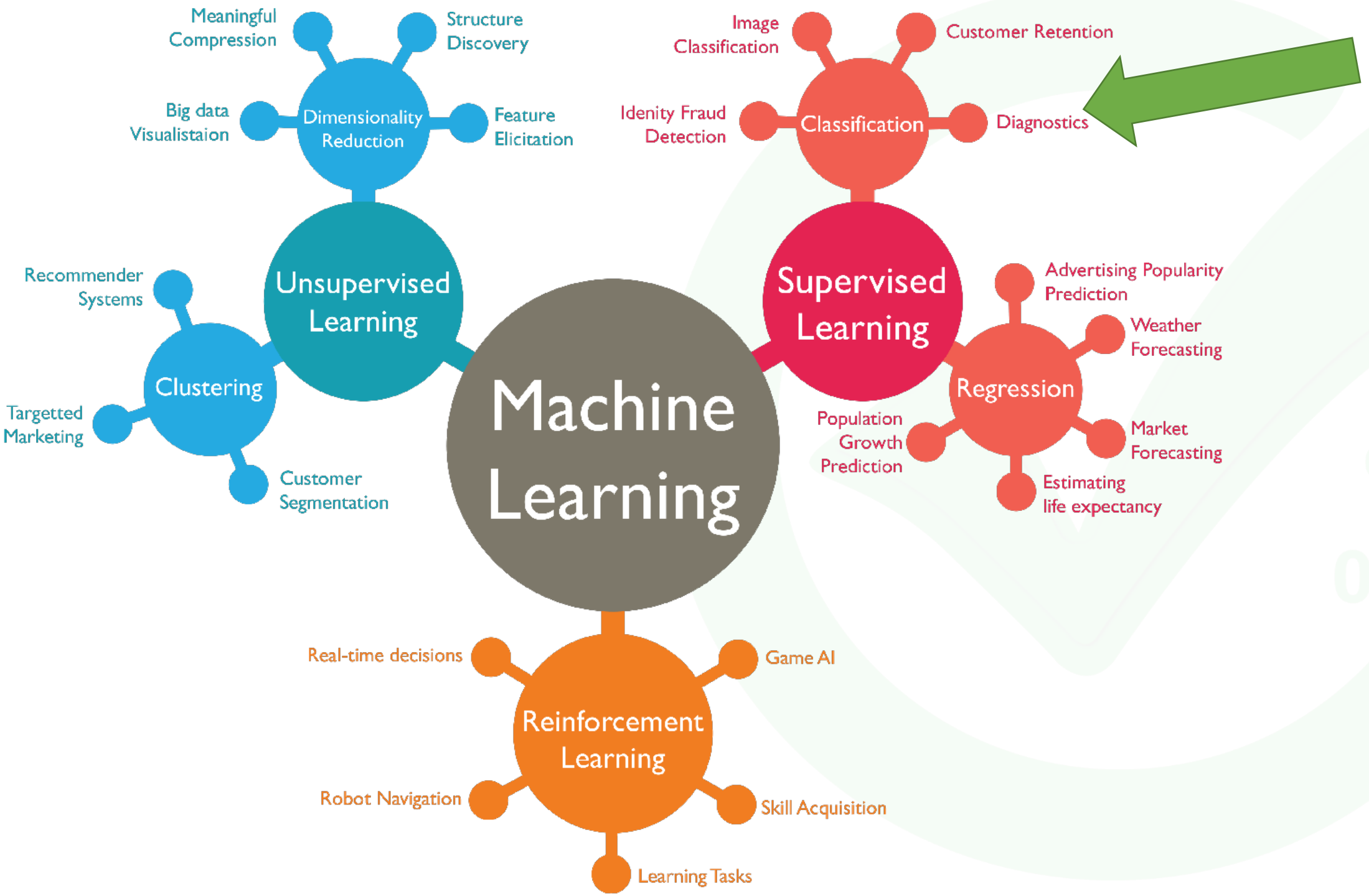
Makine Öğrenmesi – 3

Overall Table of Contents

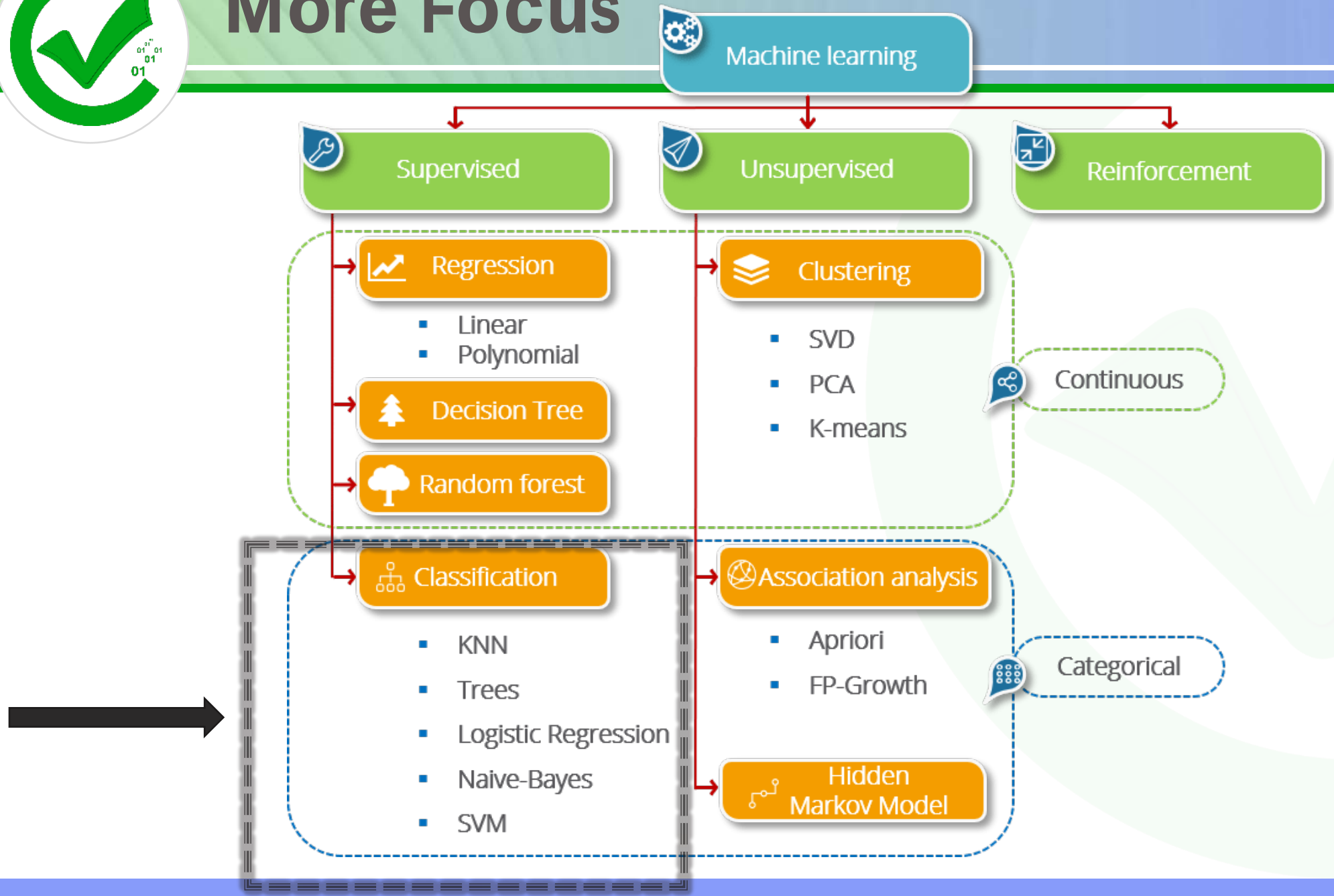
General Content

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




Machine Learning
Unsupervised Learning
Supervised Learning
Reinforcement Learning
Dimensionality Reduction
Clustering
Classification
Regression
Meaningful Compression
Structure Discovery
Big data Visualistaion
Feature Elicitation
Recommender Systems
Targetted Marketing
Customer Segmentation
Image Classification
Customer Retention
Idenity Fraud Detection
Diagnostics
Advertising Popularity Prediction
Weather Forecasting
Population Growth Prediction
Market Forecasting
Estimating life expectancy
Real-time decisions
Game AI
Robot Navigation
Skill Acquisition
Learning Tasks

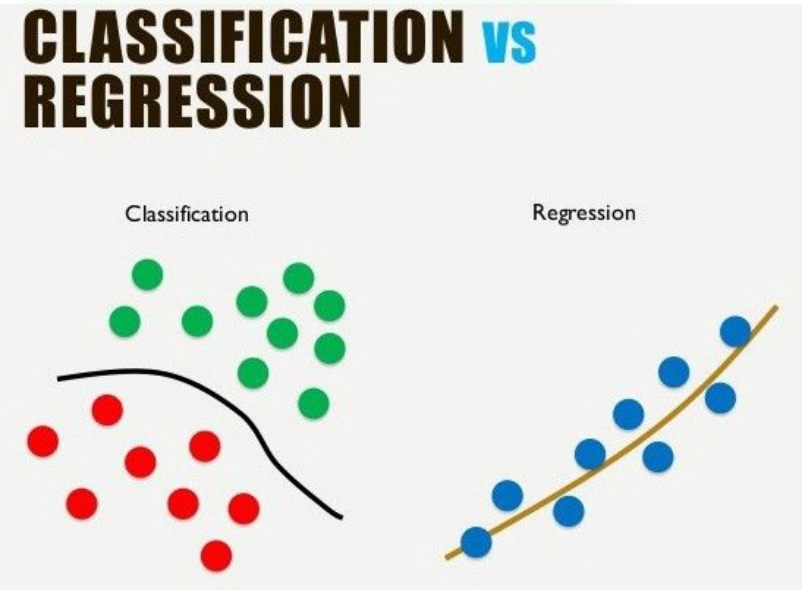
More Focus

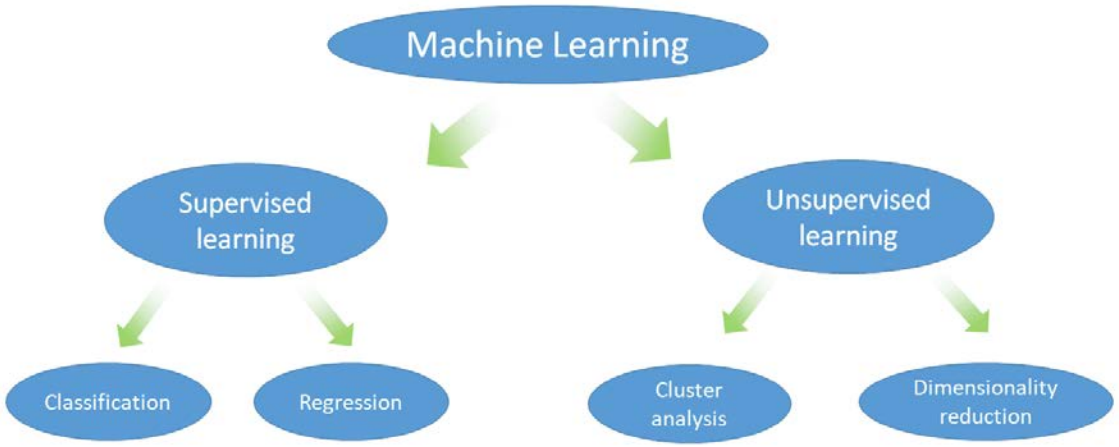

Machine learning
Supervised
Unsupervised
Reinforcement
Regression
Linear
Polynomial
Decision Tree
Random forest
Clustering
SVD
PCA
K-means
Continuous
Classification
KNN
Trees
Logistic Regression
Naive-Bayes
SVM
Association analysis
Apriori
FP-Growth
Categorical
Hidden Markov Model

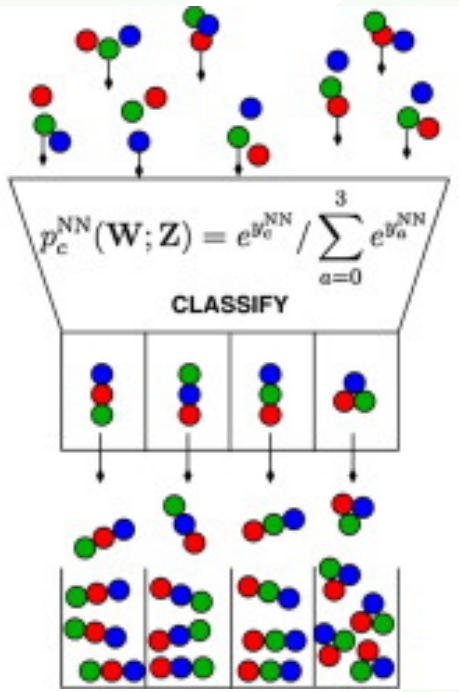
SUPERVISED LEARNING - Classification

Sınıflandırma


PAPER
METAL
PLASTIC
E-WASTE
GLASS
ORGANIC


CLASSIFICATION VS REGRESSION
Classification
Regression

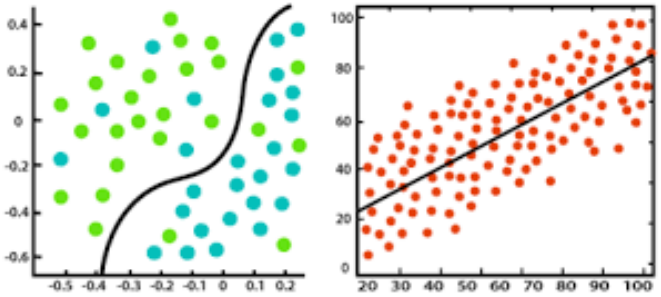

Machine Learning
Supervised learning
Unsupervised learning
Classification
Regression
Cluster analysis
Dimensionality reduction

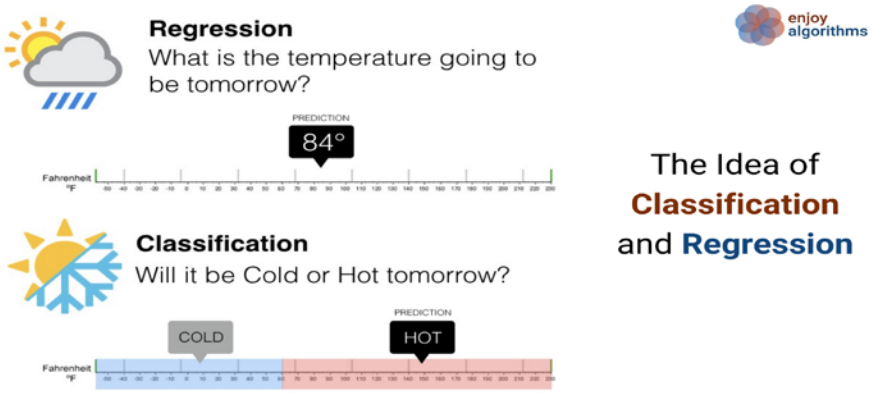

$p_c^{NN}(\mathbf{W};\mathbf{Z}) = e^{y_c^{NN}} / \sum_{a=0}^{3} e^{y_a^{NN}}$
CLASSIFY

Supervised Learning

Classification ‘a Giriş (Sınıflandırma)

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



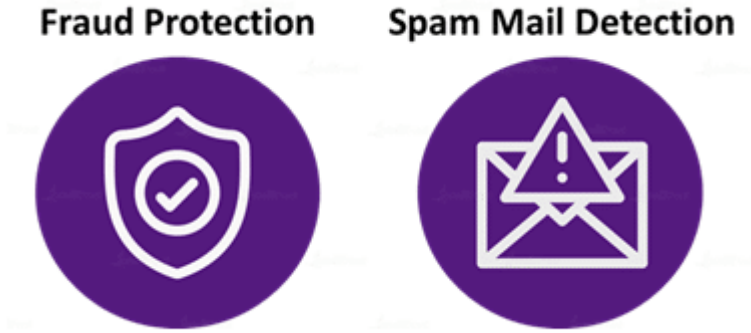


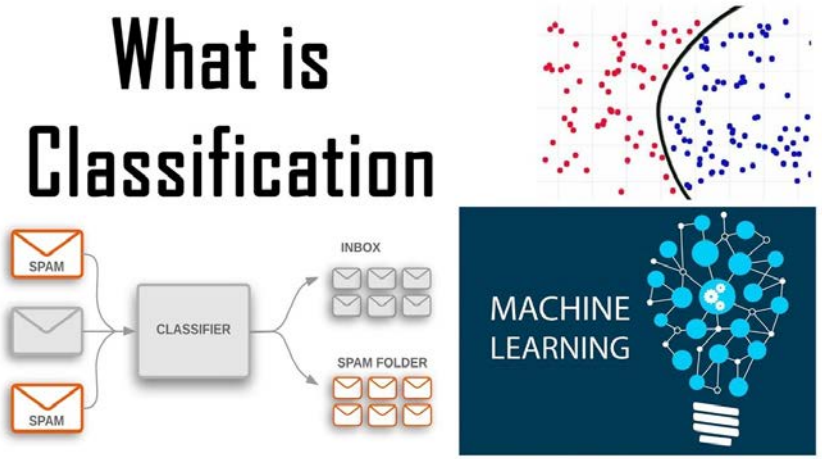
!! **Regresyonda** target hedef değişkenin **sayısal değerlerini**; **sınıflandırmada** ise target değişkenin ait olduğu **sınıfları (ya da "etiketi")** tahmin eden modelleri oluşturmaya çalışırız.

Supervised Learning

Classification ‘a Giriş

- Classification un hayattaki kullanım alanları
- Binary Classification



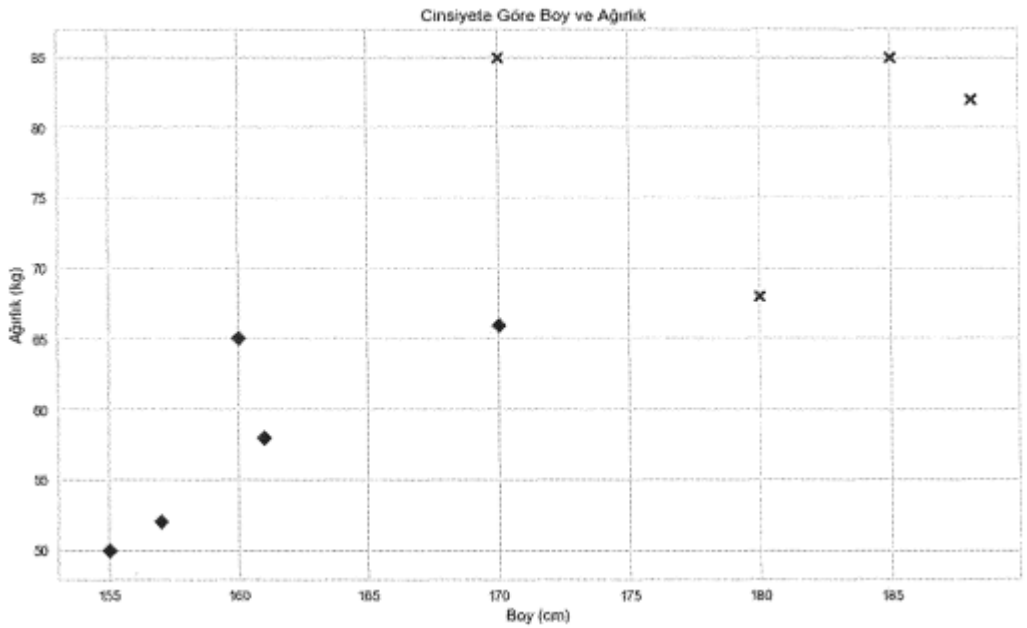


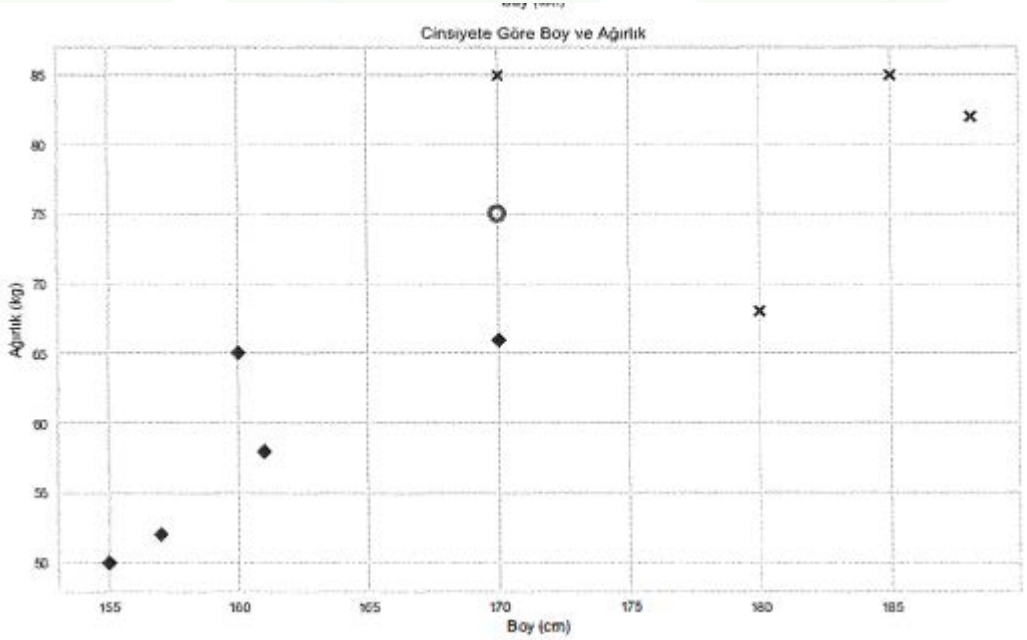
Boy	Kilo	Cinsiyet
160	65	K
170	85	E
185	85	E
188	82	E
155	50	K
161	58	K
180	68	E
157	52	K
170	66	K

Supervised Learning

Classification ‘a Giriş

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

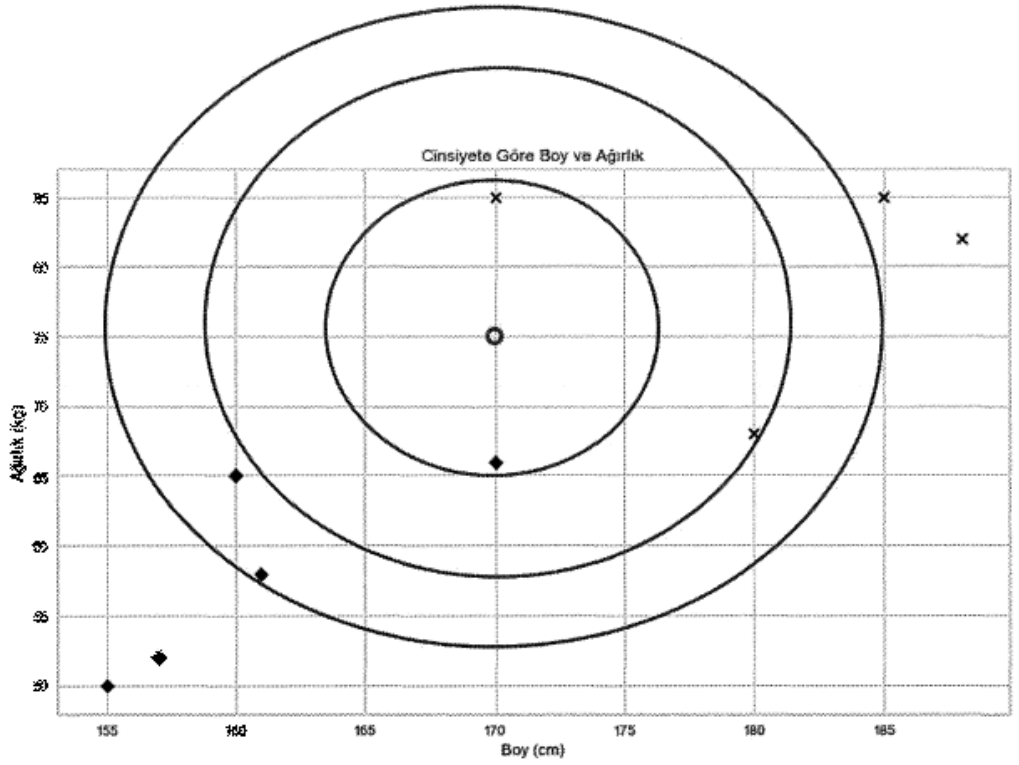


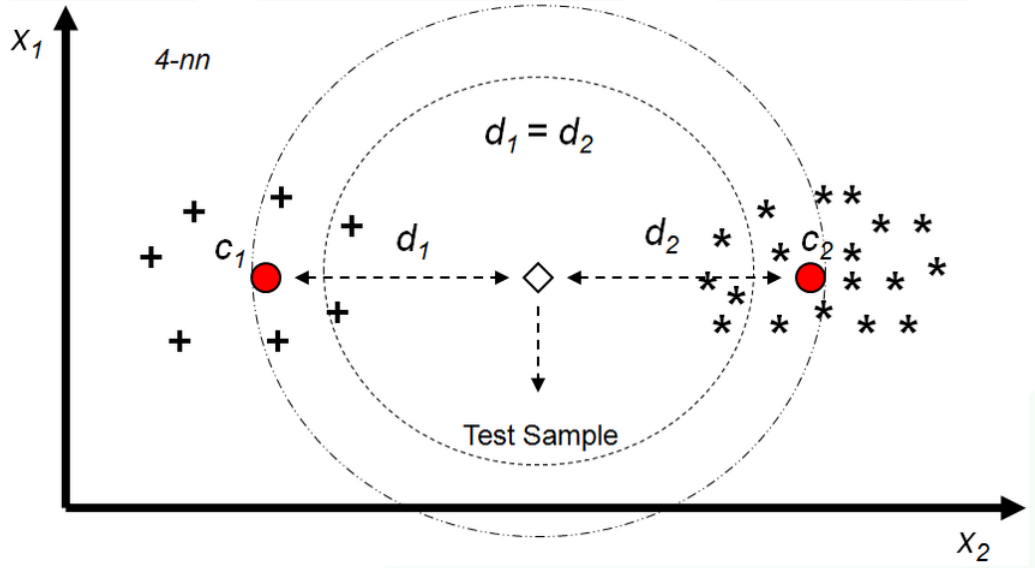


Supervised Learning

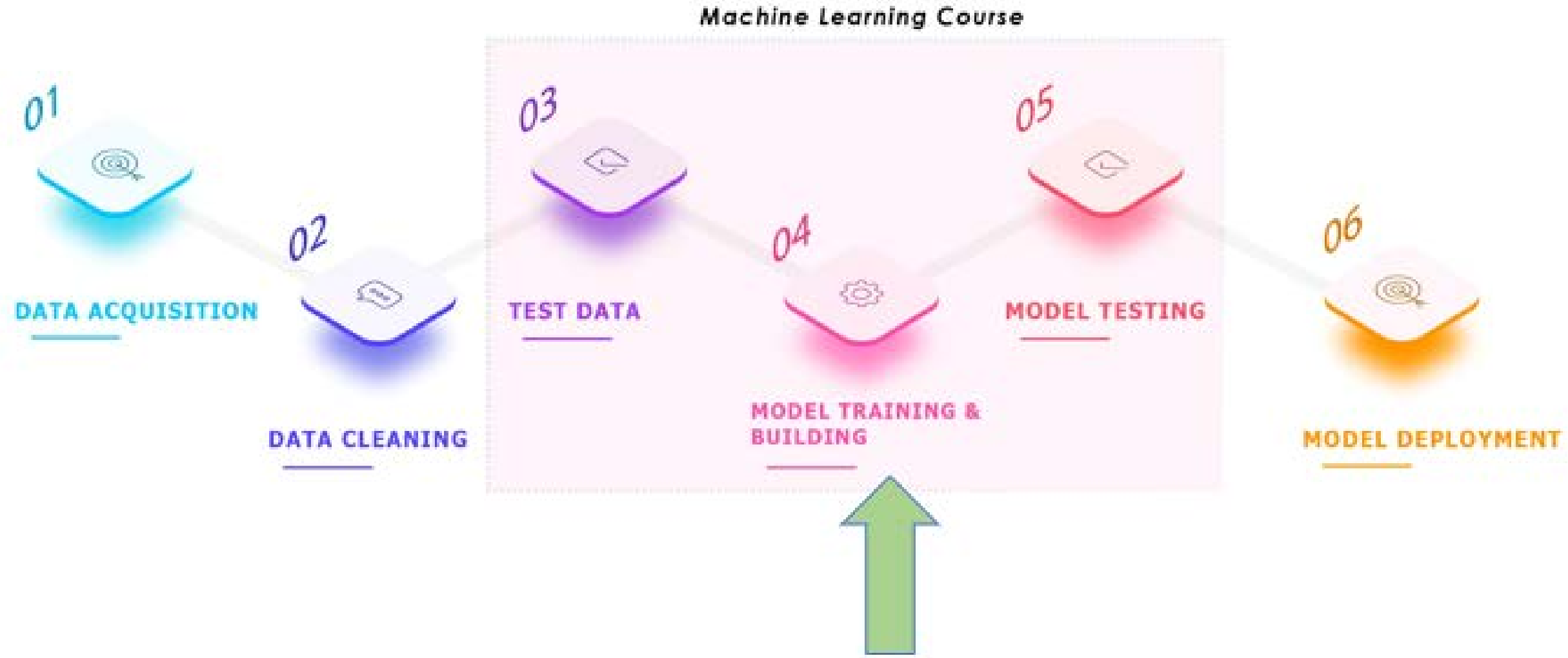
Classification ‘a Giriş

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





Where are we?

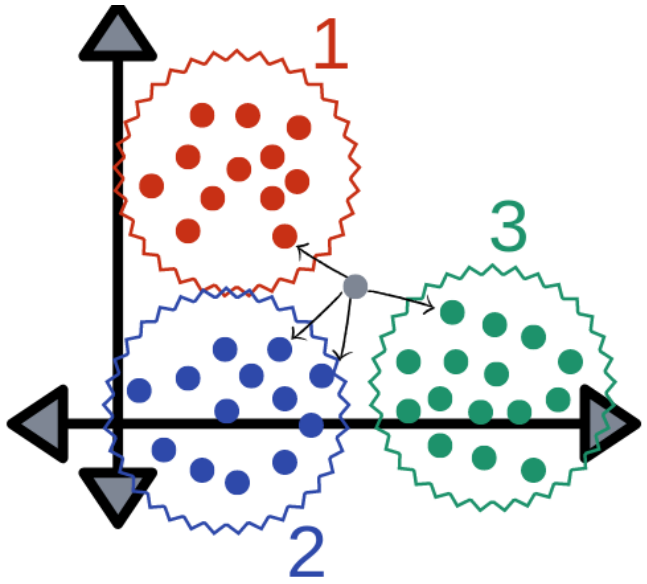

Machine Learning Course
01
DATA ACQUISITION
02
DATA CLEANING
03
TEST DATA
04
MODEL TRAINING & BUILDING
05
MODEL TESTING
06
MODEL DEPLOYMENT

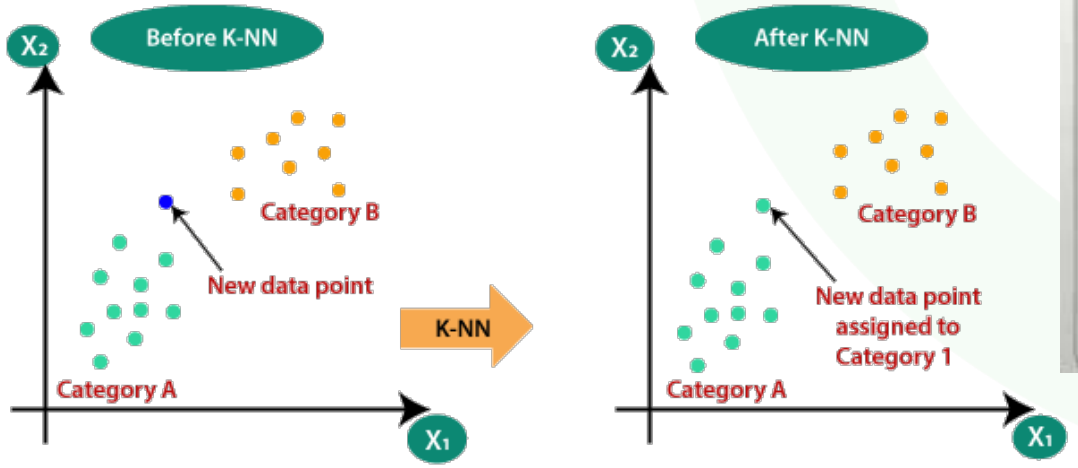
Supervised Learning

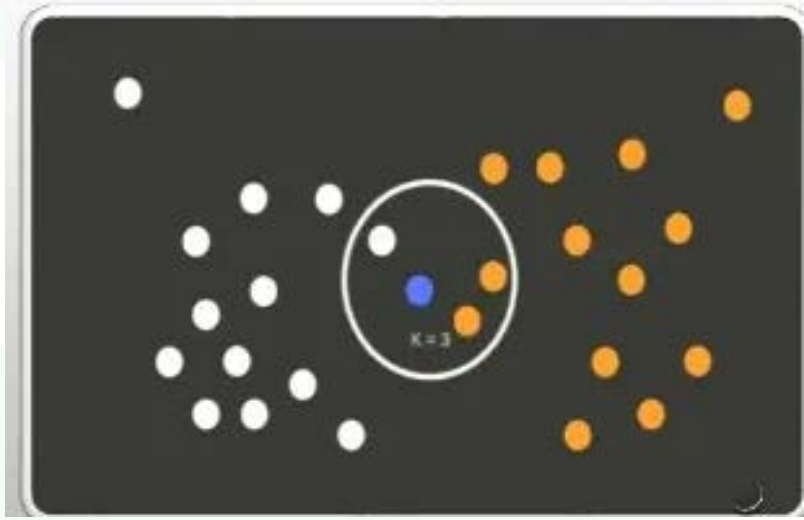
K Nearest Neighbour-KNN Algoritması

K-EN YAKIN KOMŞU Algoritması

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



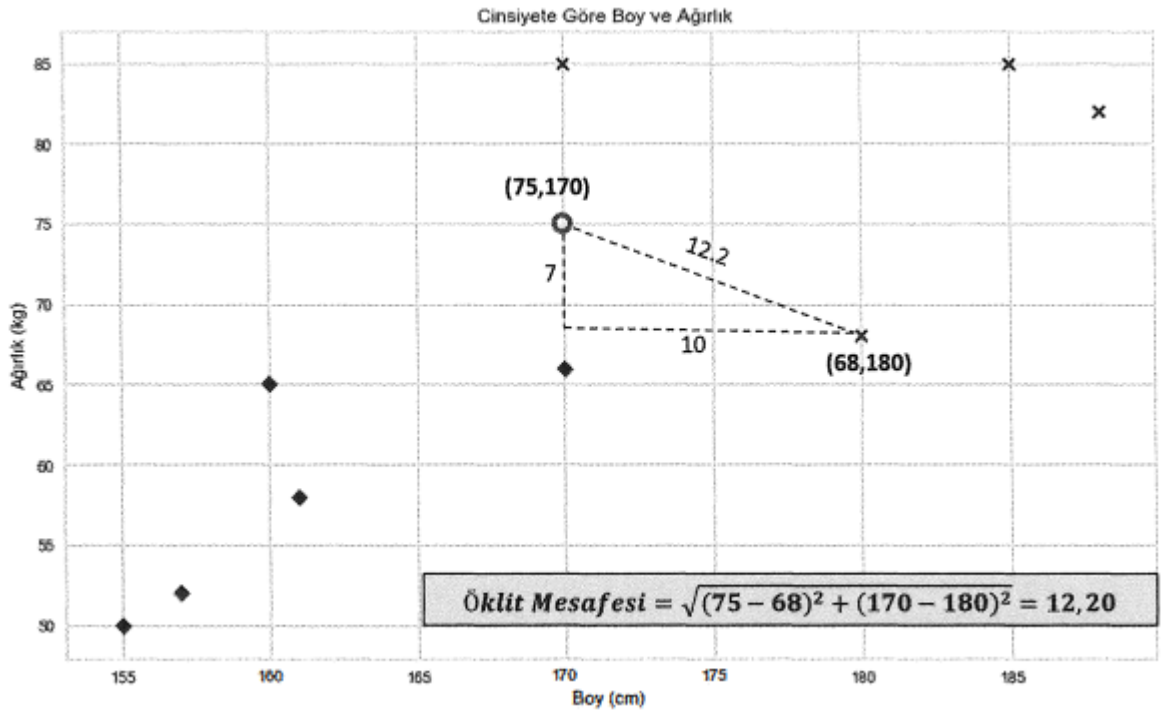


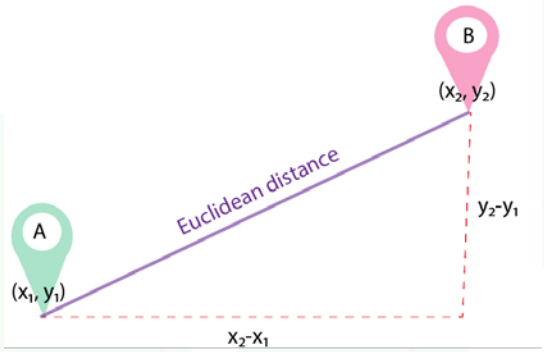


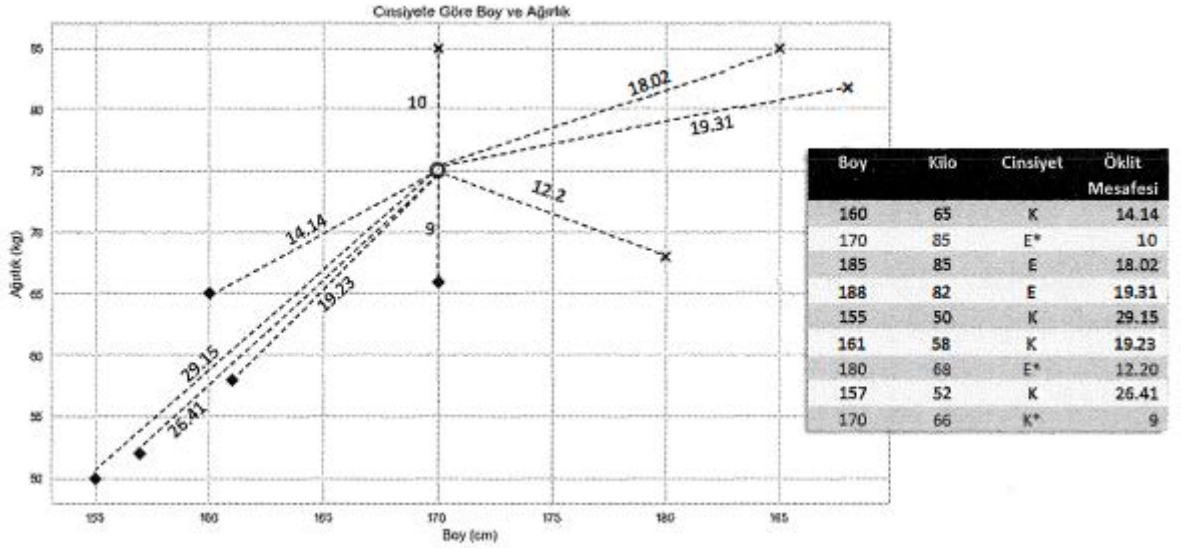
Supervised Learning

K Nearest Neighbour-KNN Algoritması

- Öklid mesafesi





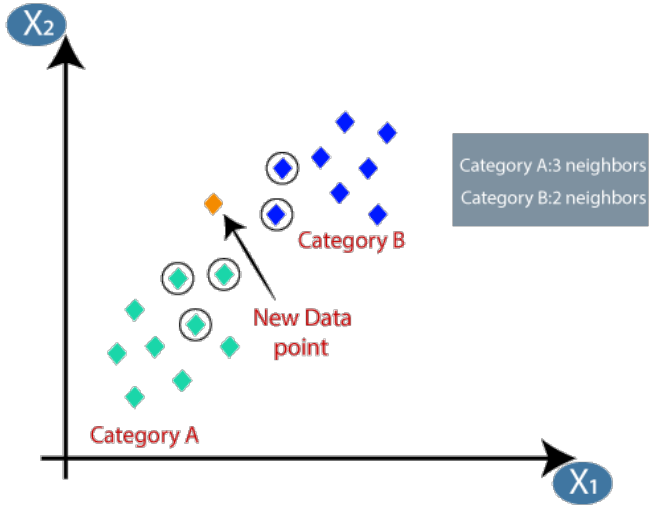


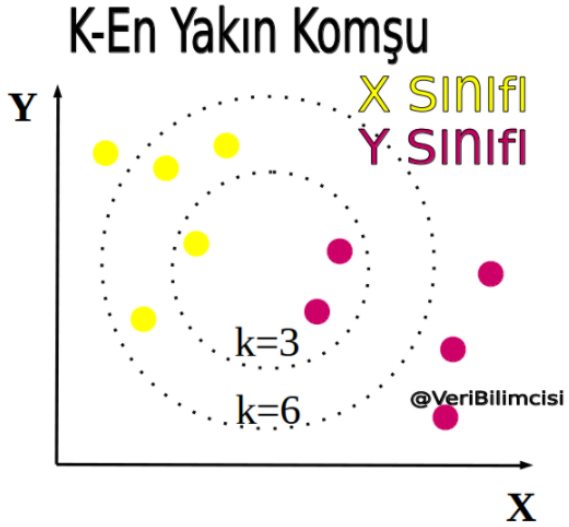
Boy	Kilo	Cinsiyet	Öklit Mesafesi
160	65	K	14.14
170	85	E*	10
185	85	E	18.02
188	82	E	19.31
155	50	K	29.15
161	58	K	19.23
180	68	E*	12.20
157	52	K	26.41
170	66	K*	9

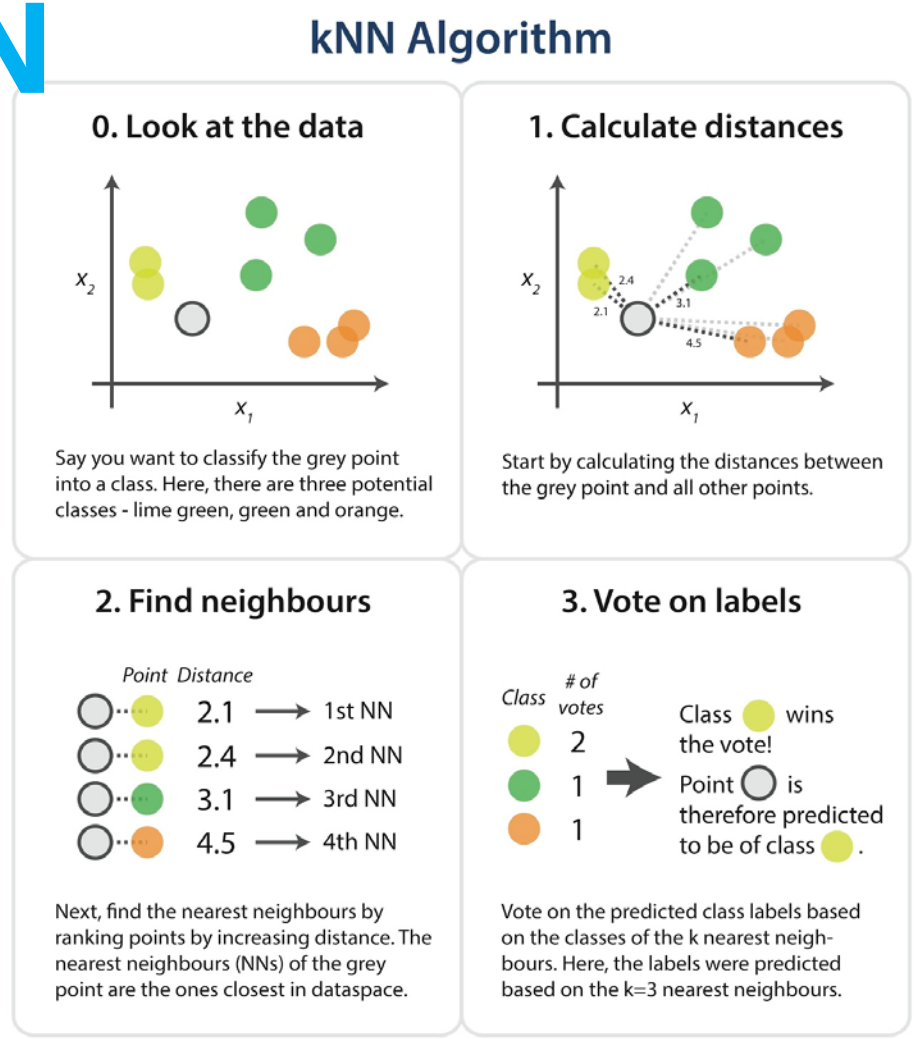
Supervised Learning

K Nearest Neighbour-KNN Algoritması

- KNN avantajları
- KNN dezavantajları

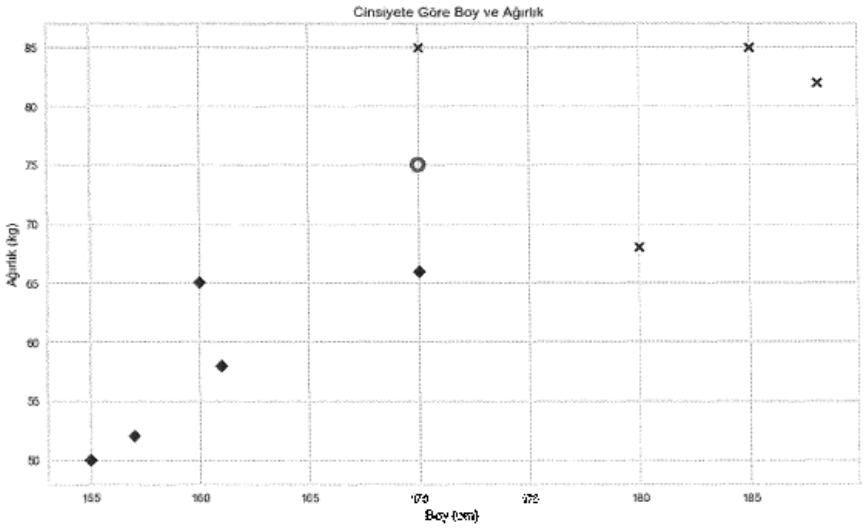


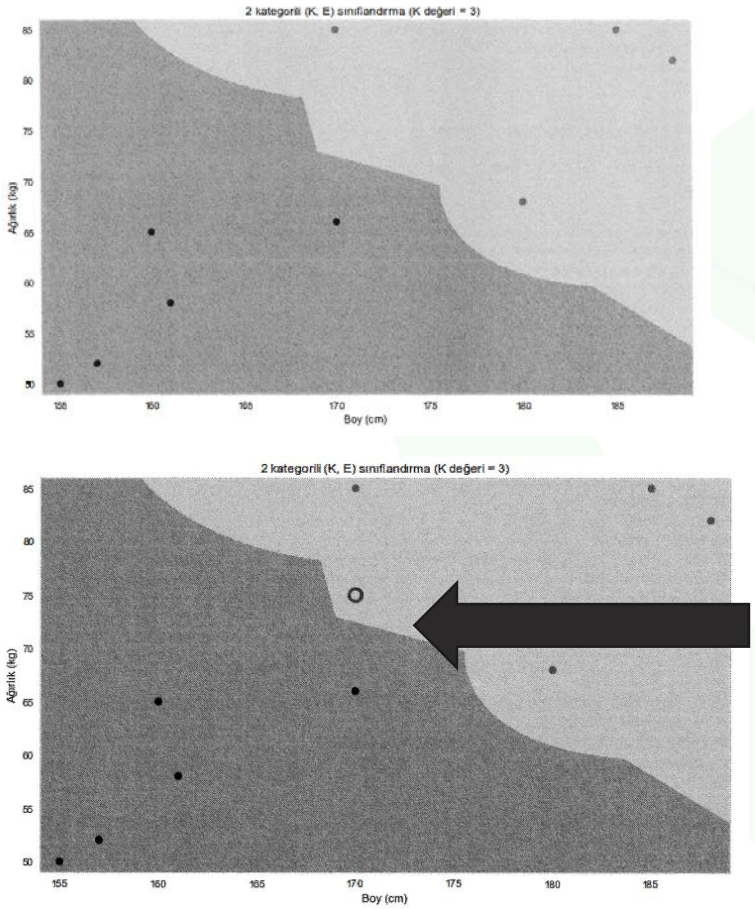


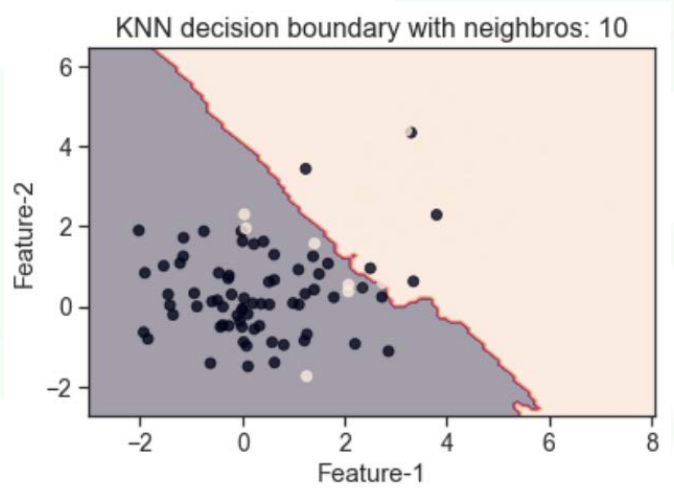


K Nearest Neighbour-KNN Algoritması

Decision boundary kavramı

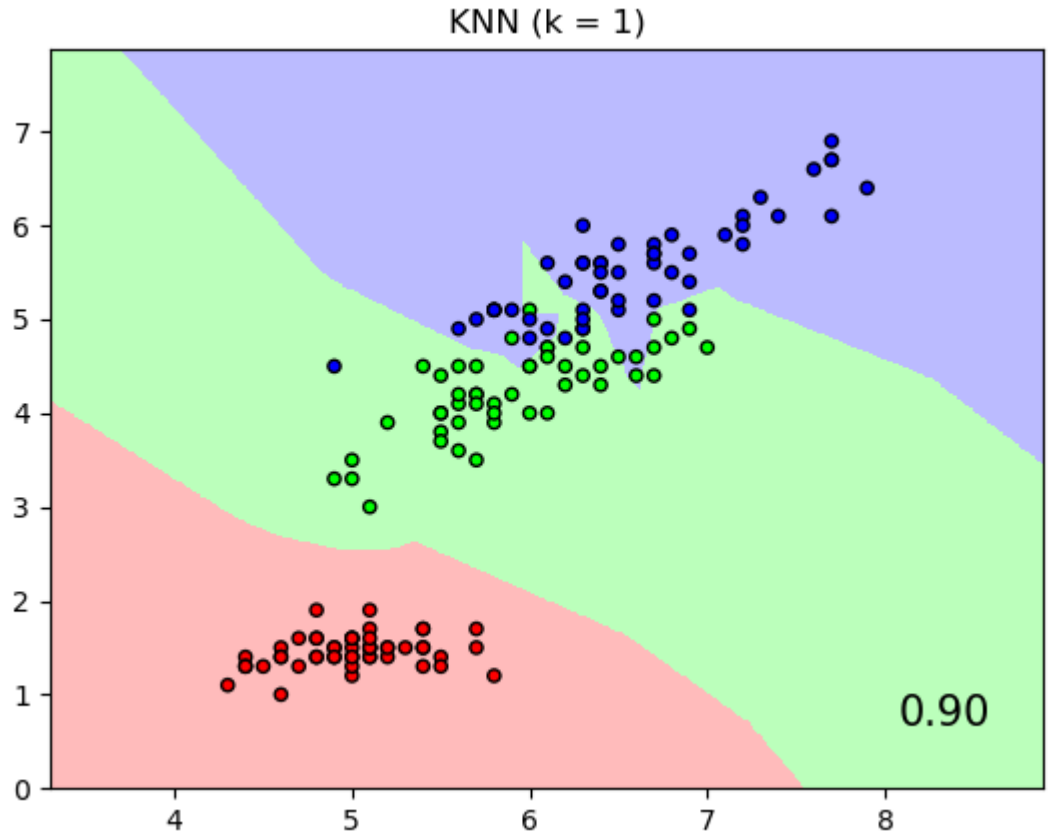


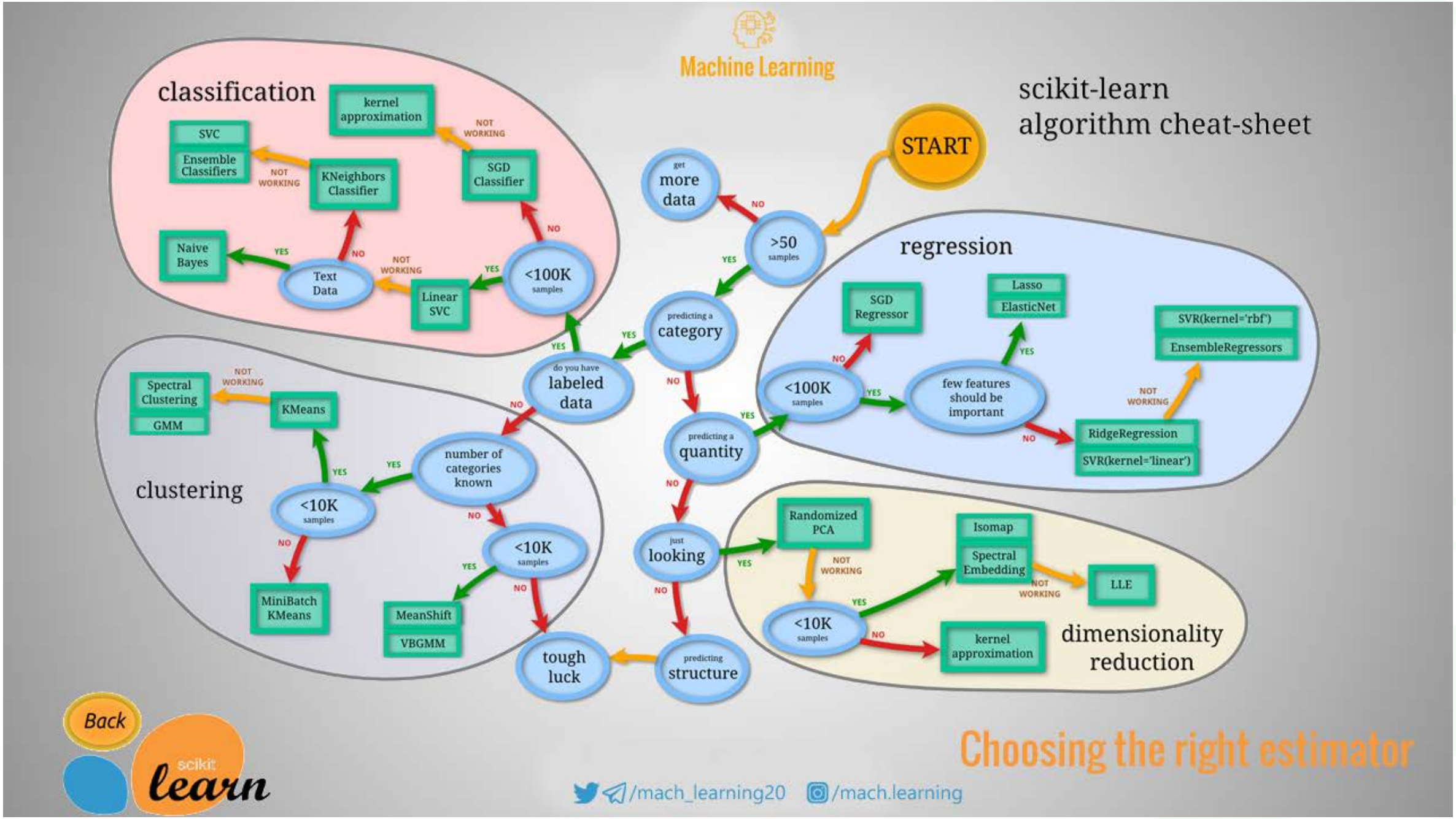




Supervised Learning

Multi Label KNN (Çok sınıflı KNN)




Machine Learning
scikit-learn
algorithm cheat-sheet
START
classification
kernel approximation
NOT WORKING
SVC
Ensemble Classifiers
NOT WORKING
KNeighbors Classifier
SGD Classifier
NO
Naive Bayes
YES
NO
Text Data
NOT WORKING
Linear SVC
YES
<100K samples
NO
YES
get more data
NO
>50 samples
YES
regression
SGD Regressor
Lasso
ElasticNet
SVR(kernel='rbf')
EnsembleRegressors
predicting a category
YES
NO
<100K samples
NO
YES
few features should be important
YES
NOT WORKING
RidgeRegression
NO
SVR(kernel='linear')
do you have labeled data
YES
NO
predicting a quantity
YES
NO
Spectral Clustering
GMM
NOT WORKING
KMeans
YES
clustering
number of categories known
YES
NO
<10K samples
NO
MiniBatch KMeans
<10K samples
YES
NO
MeanShift
VBGMM
just looking
YES
NO
Randomized PCA
NOT WORKING
Isomap
Spectral Embedding
NOT WORKING
LLE
YES
<10K samples
NO
kernel approximation
dimensionality reduction
tough luck
predicting structure
Back
scikit learn
Choosing the right estimator
/mach_learning20
/mach.learning

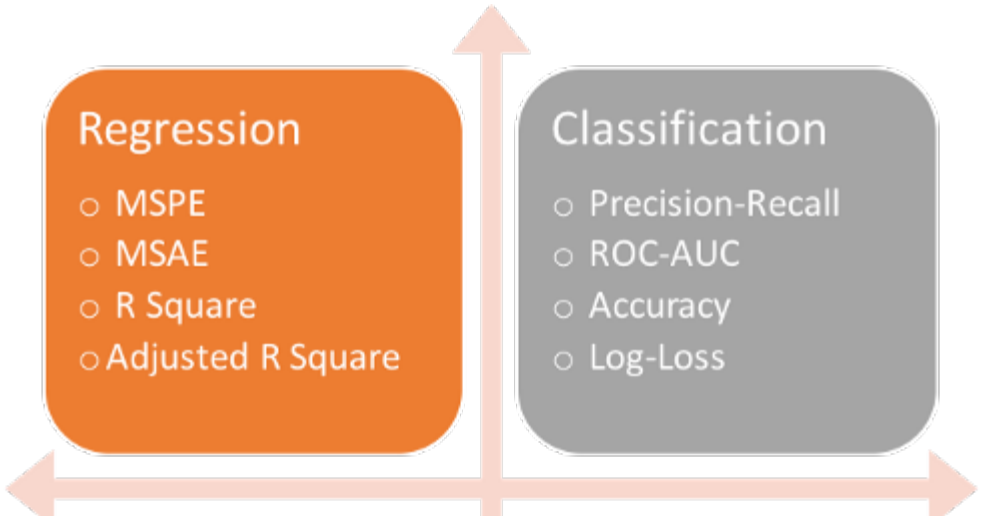
Supervised Learning

Evaluation Metrics for Classification Problems

(Performans Ölçütleri)

very Important

- Regresyon kriterleri nasıl olurdu burada?
- Confusion metrics kavramı



		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

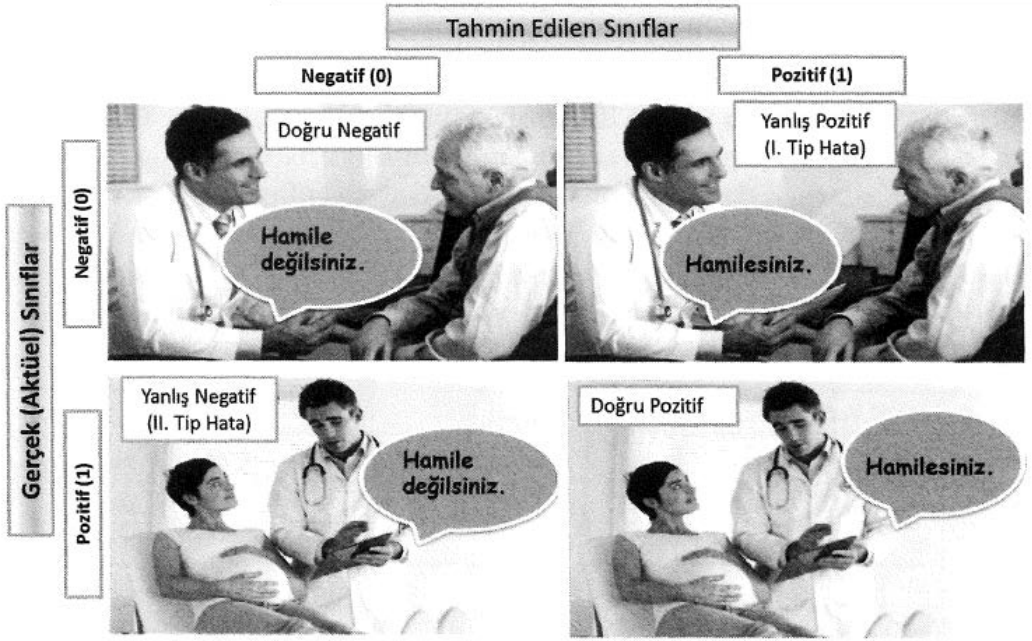
2-SINIFLI (CLASS) SINIFLANDIRMA MODELİ İÇİN HATA MATRİSİ

Gerçek Sınıflar	Tahmin Edilen Sınıflar: Negatif (0)	Tahmin Edilen Sınıflar: 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)	Tahmin Edilen Sınıflar: 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.

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

Performance metrics associated with Class 1

Predicted Labels \ Actual Labels	1	0
1	True Positive	False Positive
0	False Negative	True Negative

(Is your prediction correct?) (What did you predict)

True Negative

(Your prediction is correct) (You predicted 0)

$$\text{Precision} = \frac{TP}{TP + FP}$$

$$\text{F1 score} = 2x \frac{(\text{Prec x Rec})}{(\text{Prec} + \text{Rec})}$$

$$\text{Specificity} = \frac{TN}{TN + FP}$$

$$\text{False +ve rate} = \frac{FP}{TN + FP}$$

$$\text{Accuracy} = \frac{TP + TN}{TP + FN + FP + TN}$$

$$\text{Recall, Sensitivity} = \text{True +ve rate} = \frac{TP}{TP + FN}$$

Actual class \ Predicted class	+	-
+	TP True Positives	FN False Negatives Type II error
-	FP False Positives Type I error	TN True Negatives

In Python

```
15]: confusion_matrix(y, y_pred)

15]: array([[448,  52],
            [121, 147]], dtype=int64)
```

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
	✗	✓	✗	✓	✓	✓	✓	✓	✓	✗

$$\frac{\text{All correctly predicted values (7)}}{\text{All predicted values (10)}} \times 100 \Rightarrow \text{Accuracy = 70 \%}$$

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 ....
	✗	✓	✗	✓	✓	✓	✓	✓	✓	✗

$$\frac{\text{All correctly predicted values (60)}}{\text{All predicted values (63)}} \times 100 \Rightarrow \text{Accuracy = 95 \% (PERFECT ????)}$$

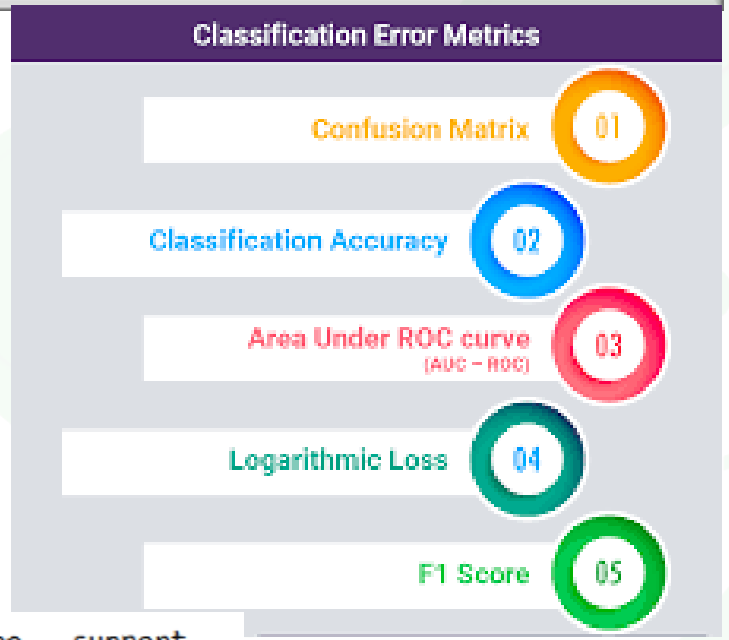
Accuracy is very high, but missed 2 actual patient.

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



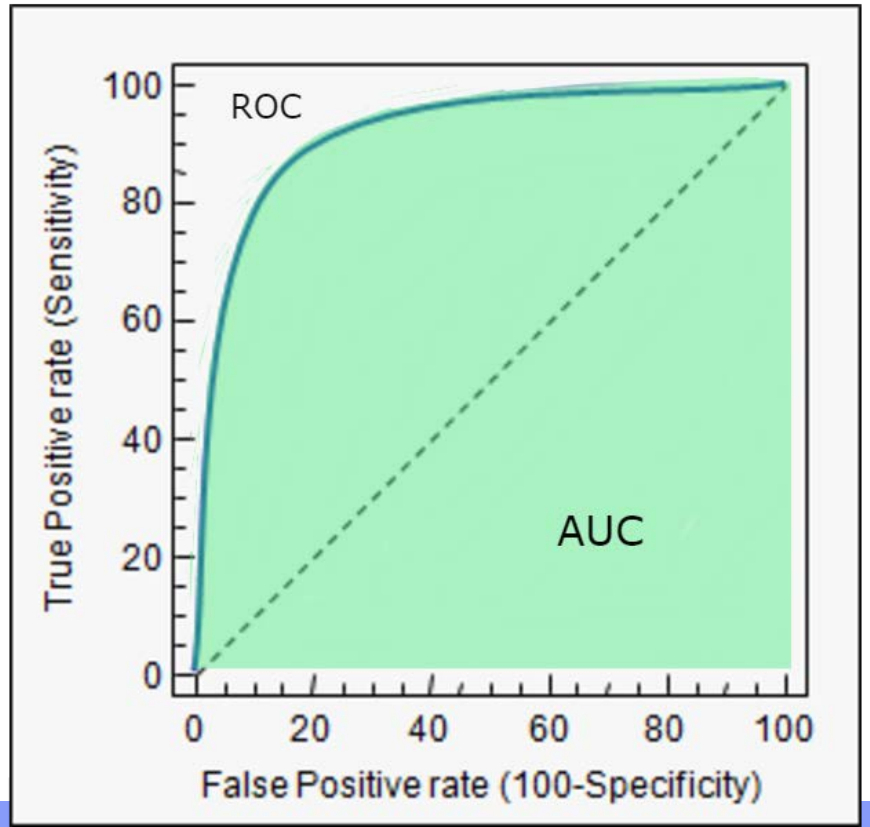
Hata Matrisi		Tahmin Edilen Etiketler	
Negatif (0)		Pozitif (1)	
Gerçek/Aktüel Veri	Negatif (0)	DN=2	YP=0
	Pozitif (1)	YN=1	DP=2

	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

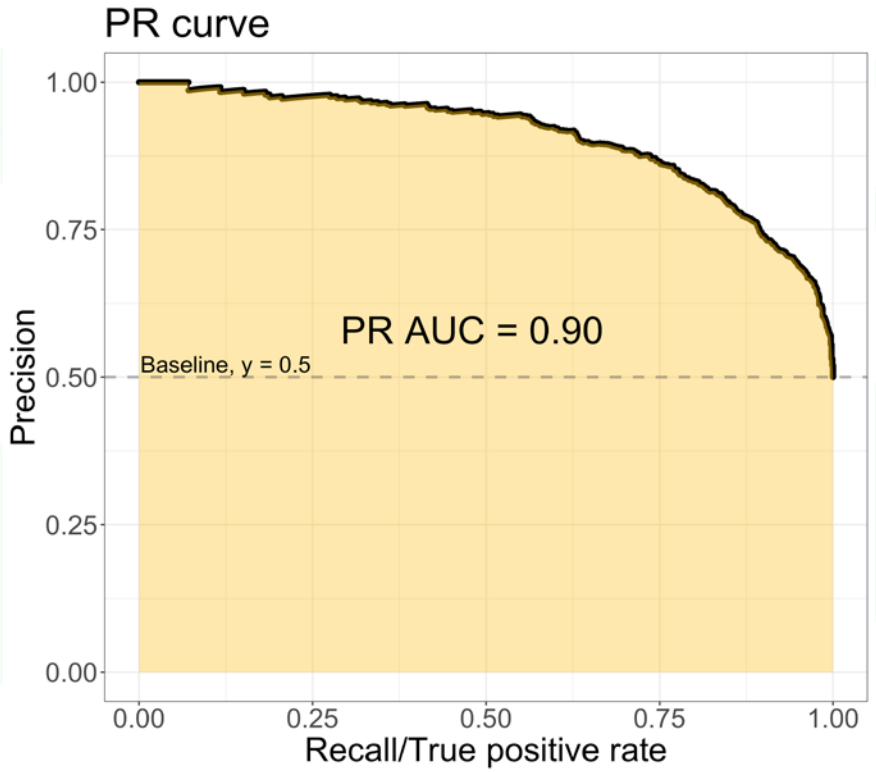
Supervised Learning

Evaluation Metrics for Classification Problems

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



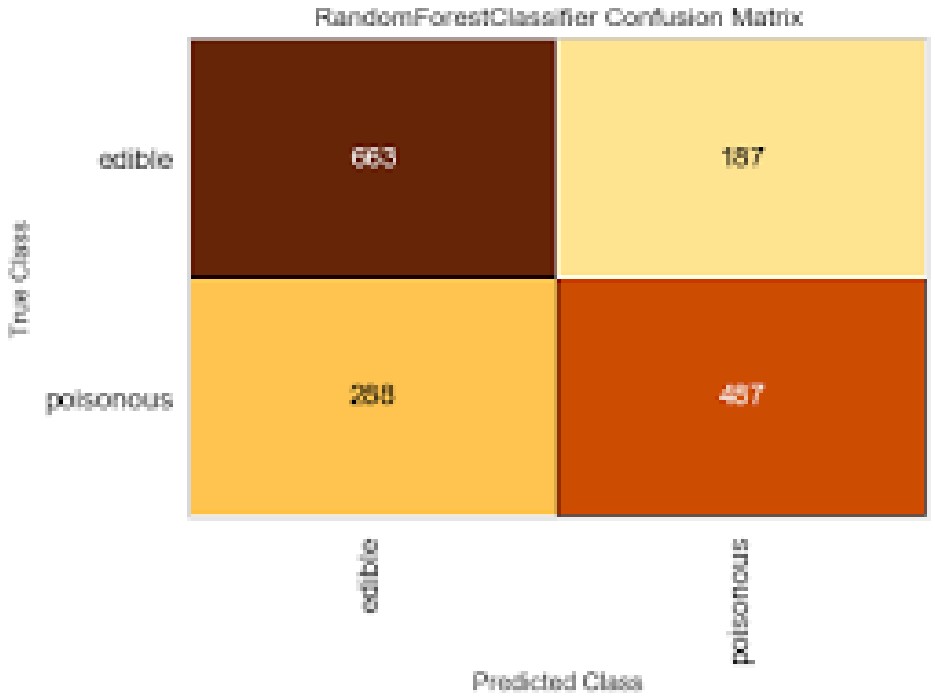
AUC'un olabildiğinde yüksek olması (1 'e yakın olmasını) istiyoruz
AUC ne kadar yüksekse model o kadar negatif (0) durumları negatif; pozitif (1) durumları da pozitif öngörüyor demektir.

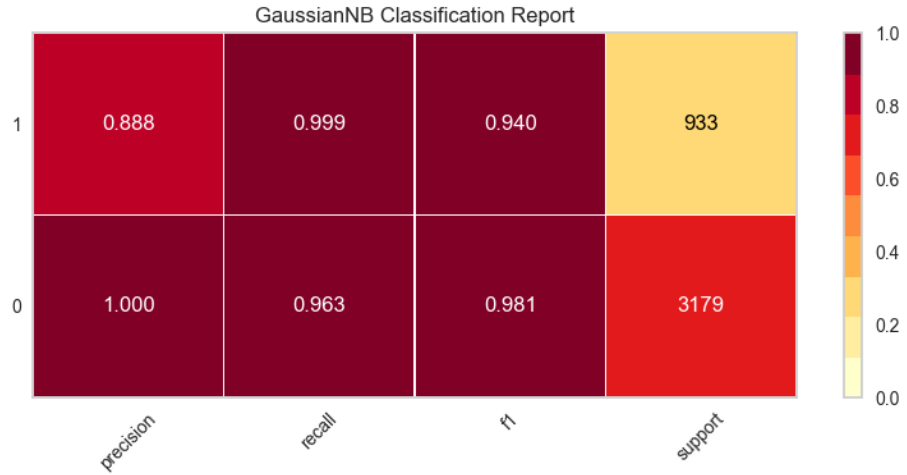


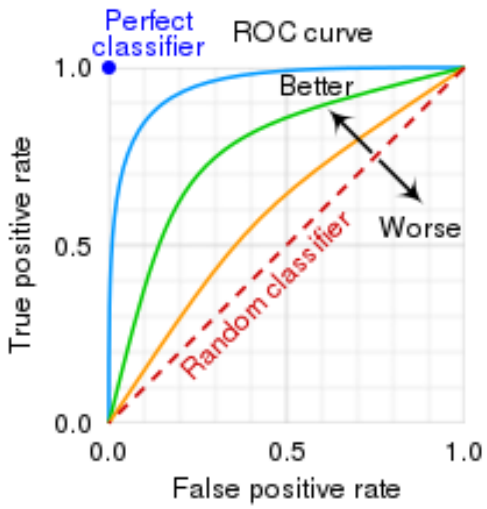
Supervised Learning

Evaluation Metrics for Classification Problems

- Yellowbrick ile confusion matrics

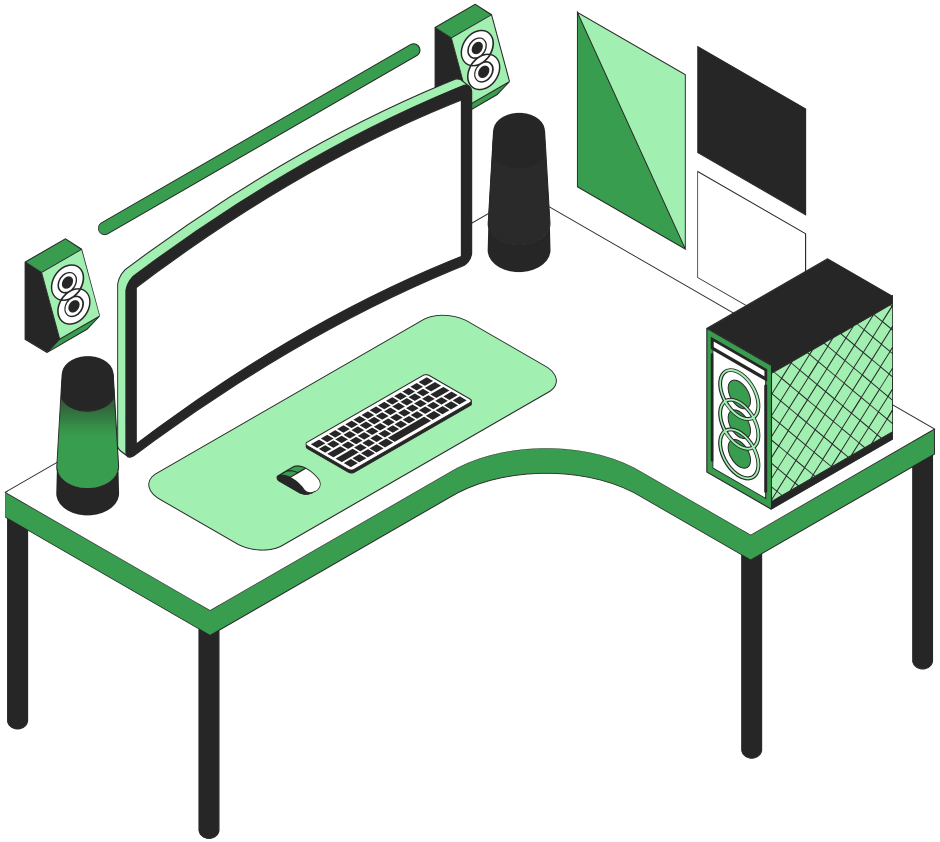






Bu dersi anladım..

Everythig is clear ?



Do you have any questions?

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