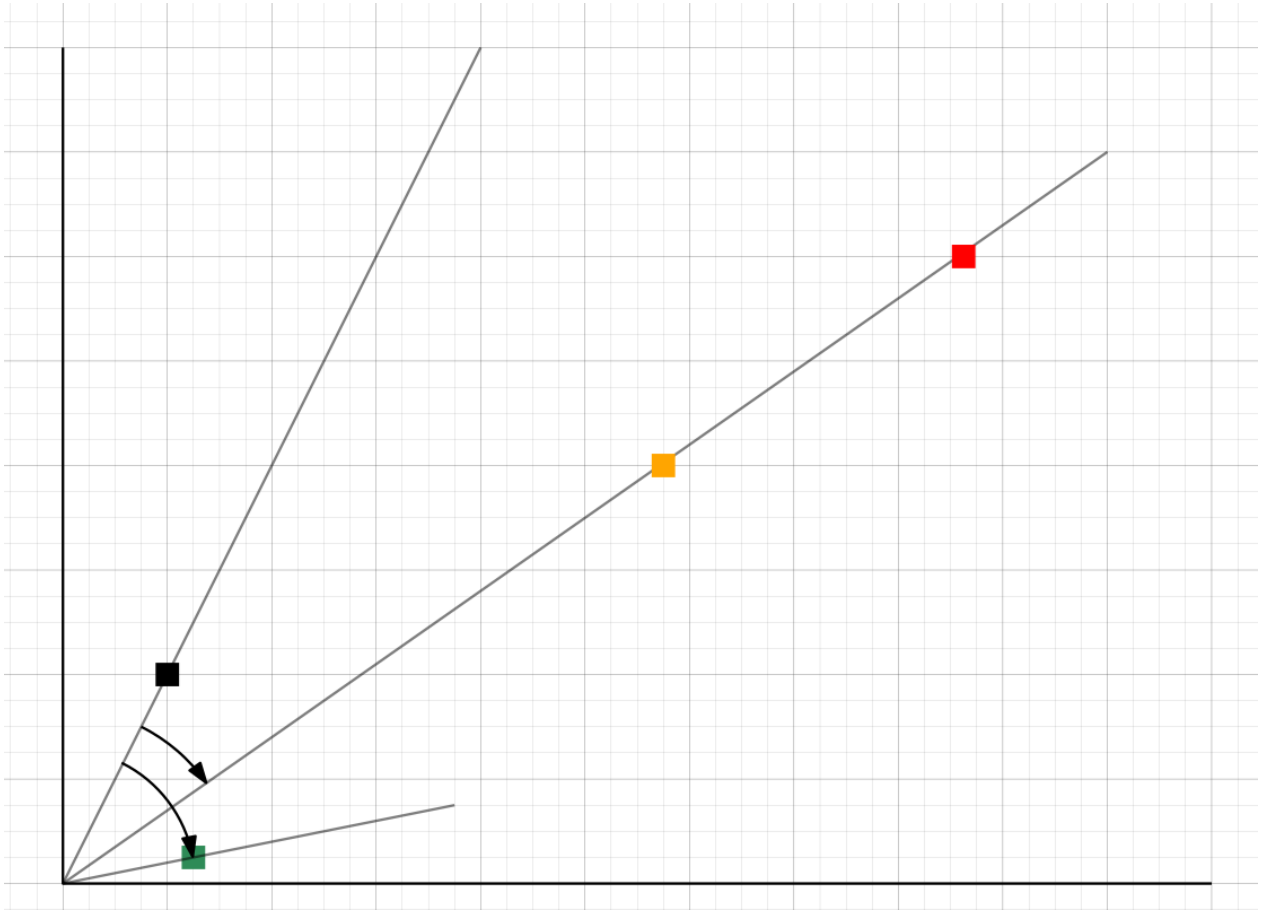
BİL 475 Örüntü Tanıma

Hafta-4:

Bayes Karar Teorisi-2

K-En Yakın Komşu Sınıflandırma (k-NN)

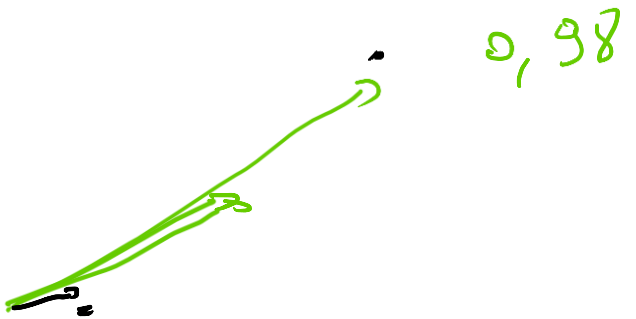


$$\frac{\langle x_1, x_2 \rangle}{|x_1||x_2|}$$

$$\mathbf{u}.\mathbf{v} = u_1v_1 + u_2v_2 + u_3v_3$$

$$\mathbf{u}.\mathbf{v} = ||\mathbf{u}||||\mathbf{v}|| \cos\theta$$

$\cos(\theta)$ $-1, 1$



K-En Yakın Komşu Sınıflandırma (k-NN)

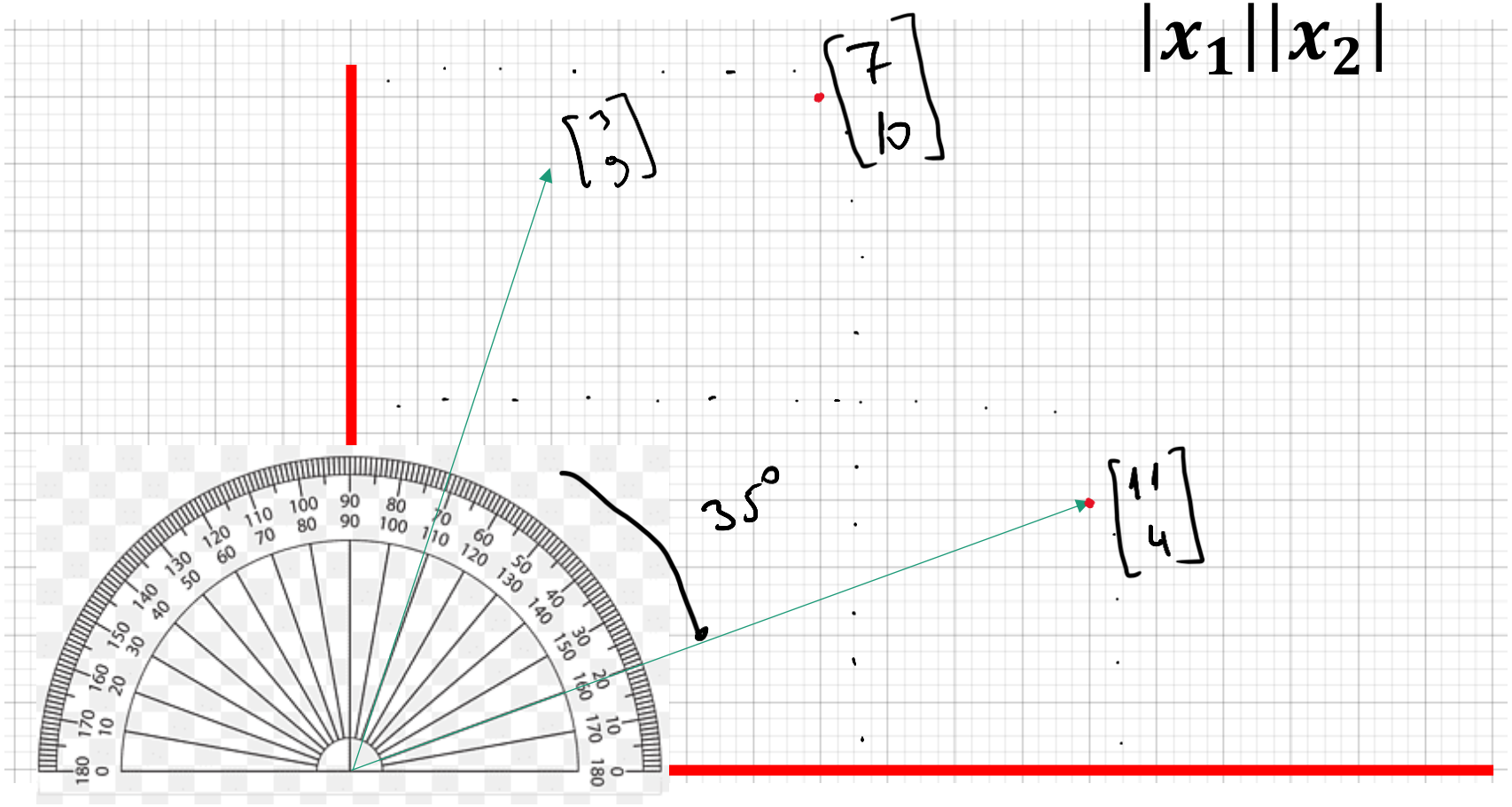
$$\frac{\langle x_1, x_2 \rangle}{|x_1||x_2|} = \alpha$$

$$\mathbf{u}.\mathbf{v} = u_1v_1 + u_2v_2 + u_3v_3$$

$$\mathbf{u}.\mathbf{v} = ||\mathbf{u}||||\mathbf{v}|| \cos\theta$$

$$\alpha = \cos(\theta)$$

$$\arccos(\alpha) = \theta$$



metric : *str or callable, default='minkowski'*

Metric to use for distance computation. Default is "minkowski", which results in the standard Euclidean distance when p = 2. See the documentation of scipy.spatial.distance and the metrics listed in `distance_metrics` for valid metric values.

If metric is "precomputed", X is assumed to be a distance matrix and must be square during fit. X may be a sparse graph, in which case only "nonzero" elements may be considered neighbors.

If metric is a callable function, it takes two arrays representing 1D vectors as inputs and must return one value indicating the distance between those vectors. This works for Scipy's metrics, but is less efficient than passing the metric name as a string.

sklearn.metrics.pairwise.distance_metrics

`sklearn.metrics.pairwise.distance_metrics()` [source]

Valid metrics for pairwise_distances.

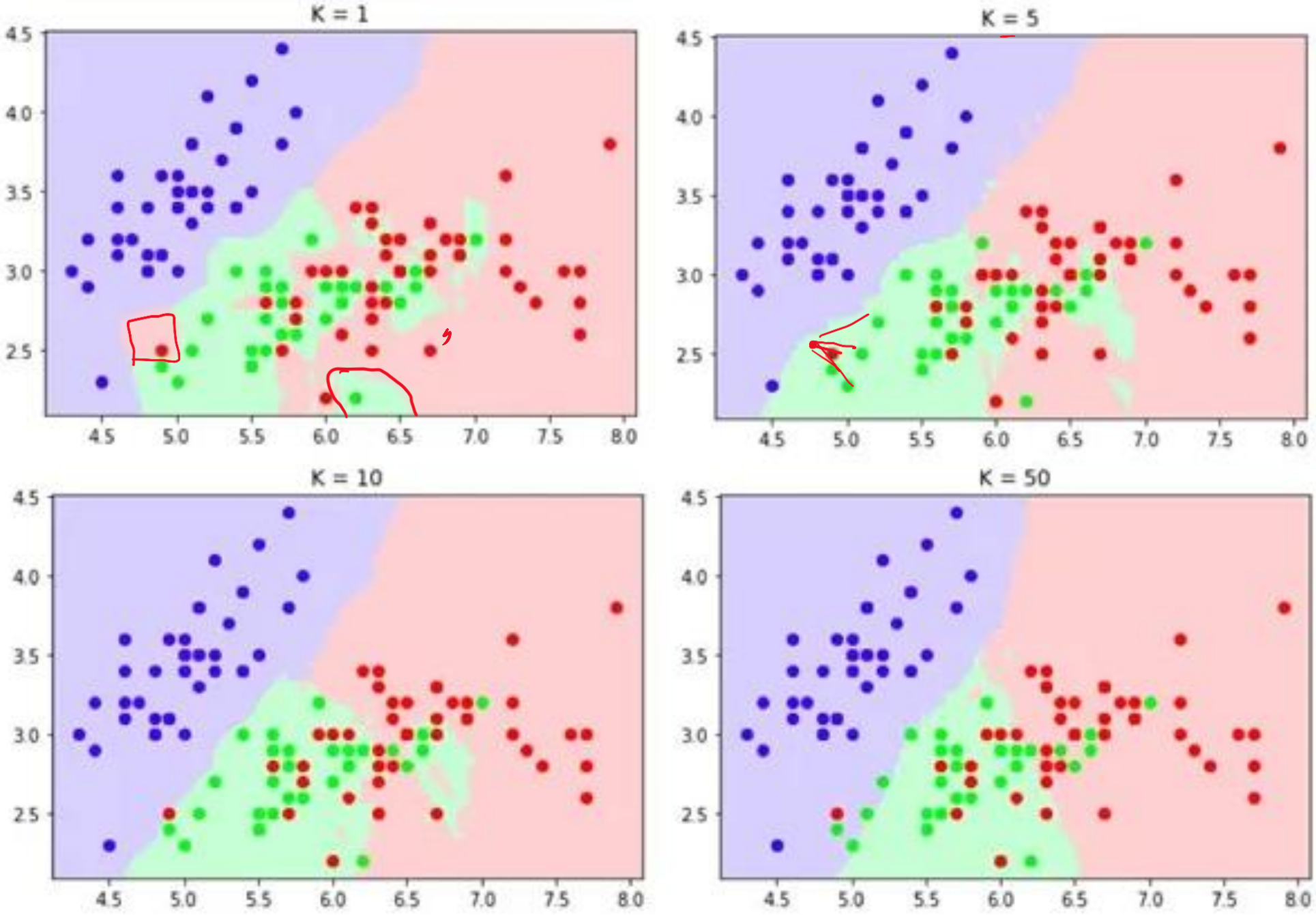
This function simply returns the valid pairwise distance metrics. It exists to allow for a description of the mapping for each of the valid strings.

The valid distance metrics, and the function they map to, are:

metric	Function
'cityblock'	metrics.pairwise.manhattan_distances
'cosine'	metrics.pairwise.cosine_distances
'euclidean'	metrics.pairwise.euclidean_distances
'haversine'	metrics.pairwise.haversine_distances
'l1'	metrics.pairwise.manhattan_distances
'l2'	metrics.pairwise.euclidean_distances
'manhattan'	metrics.pairwise.manhattan_distances
'nan_euclidean'	metrics.pairwise.nan_euclidean_distances

K-NN algoritması

- Avantajları
  - Gerçekleme kolaylığı
  - Herhangi bir ön kabule ihtiyacı yoktur.
  - Eğitim yok
  - Yeni örnekler geldiğinde hızlı adaptasyon sağlar
  - Hem sınıflandırma hem de regresyon için kullanılır.
  - Birkaç parametre (k ve norm)
  - Doğrusal olmayan veriler sınıflandırılabilir
- Dezavantajları
  - Yavaş bir algoritmadır (büyük veri)
  - Homojen öznitelikler olması gerekir
  - Aykırı örneklere takılabilir.
  - K sayısının tespiti
  - RAM ihtiyacı

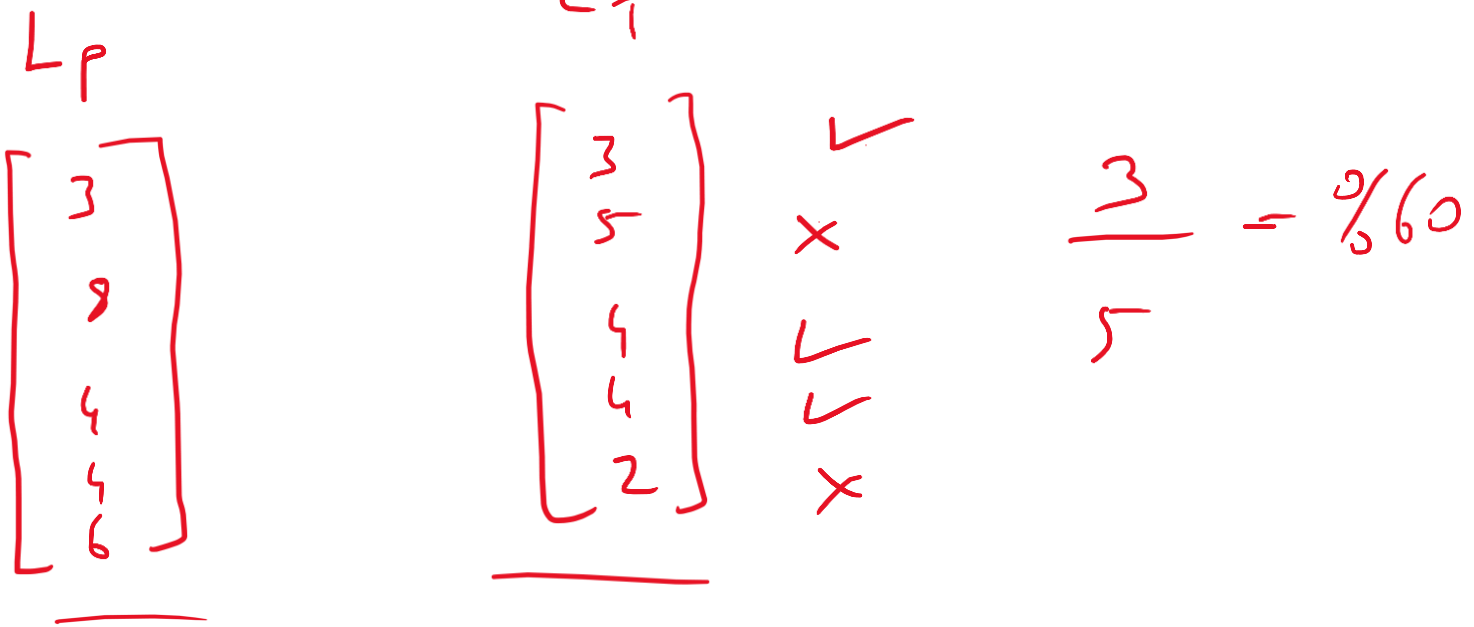


https://kdagiit.medium.com/k-nearest-neighbor-knn-algorithm-9a0eefe1f148

K-NN algoritması

- Doğruluk Hesaplaması

Doğruluk (Accuracy)



- DENETİMLİ
  - Sınıflandırma
    - k-NN : K, norm
  - Regresyon
- DENETİMSİZ

Başarım Kriteri

- Doğruluk

K-NN algoritması

- İleri Konular (Büyük Veri)

PAPER • OPEN ACCESS

Analysis of KNN Algorithm with Mapreduce Technique on Big Data

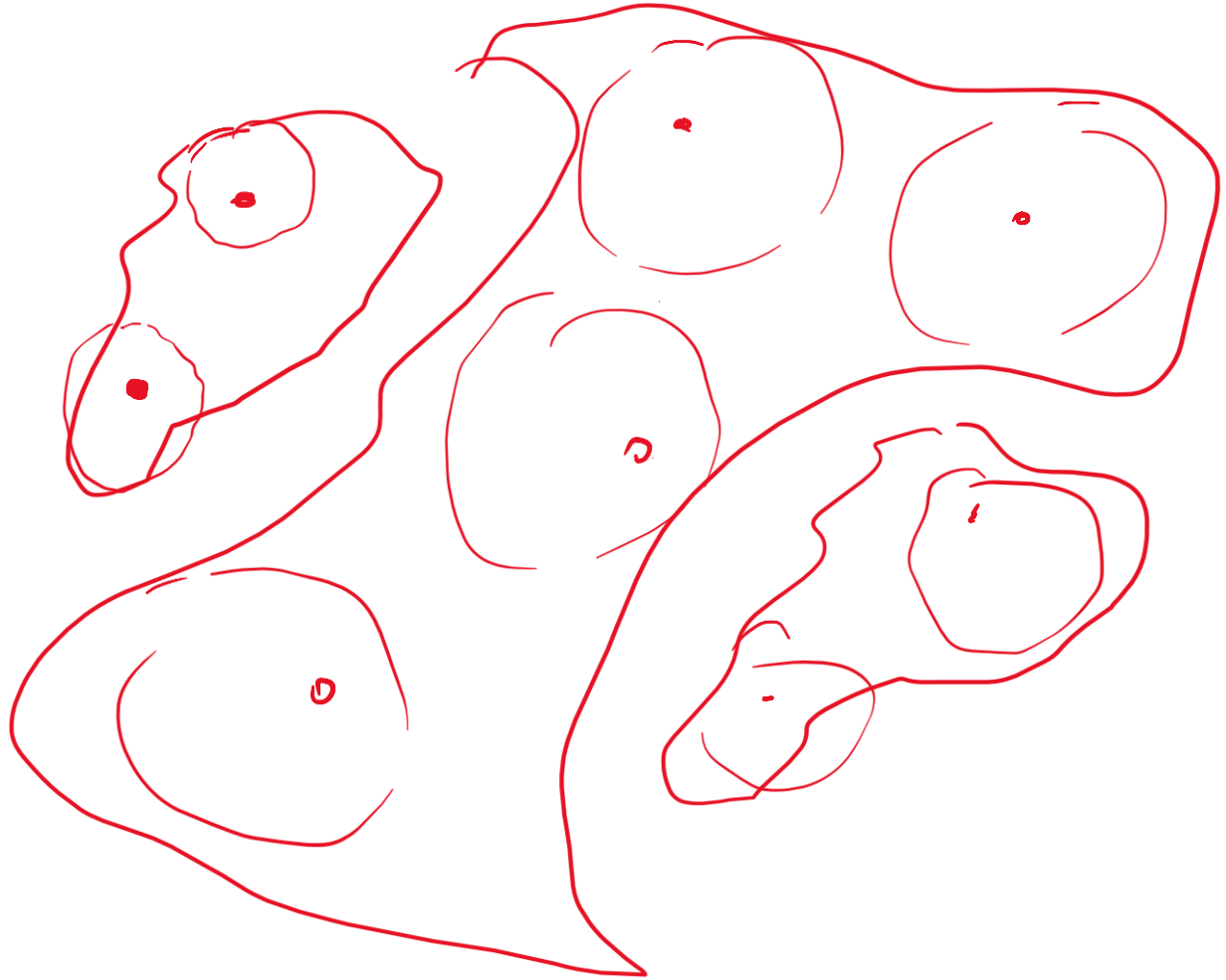
Tatikonda Bhavana[1], J. Padmavathy[1], R. Sethuraman[2] and J.K. Jeevitha[3]

IOP Conference Series: Materials Science and Engineering, Volume 590, International Conference on Frontiers in Materials and Smart System Technologies 10 April 2019, Tamil Nadu, India

Citation Tatikonda Bhavana *et al* 2019 *IOP Conf. Ser.: Mater. Sci. Eng.* **590** 012028

DOI 10.1088/1757-899X/590/1/012028

Article PDF



K-NN algoritması – Son Örnek

Bayes Karar Teorisi

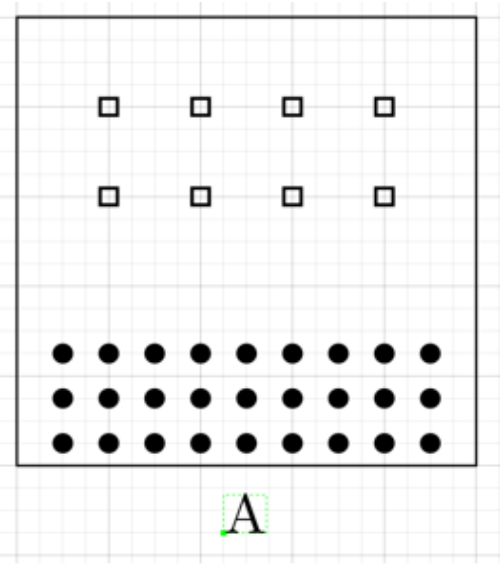
Olasılık 101

- Olasılık nedir?
  - ❑Bir şeyin olmasına ait matematiksel yüzdesi (wiki)
  - ❑Popülasyonu betimleyen sayısal bilgiler

- Yazı tura
- Zar atma
- Okula varma süresi
- 5 günlük hava raporunun

sonunda meteoroloji tahmini

Olasılık 101



Kare ve Daire

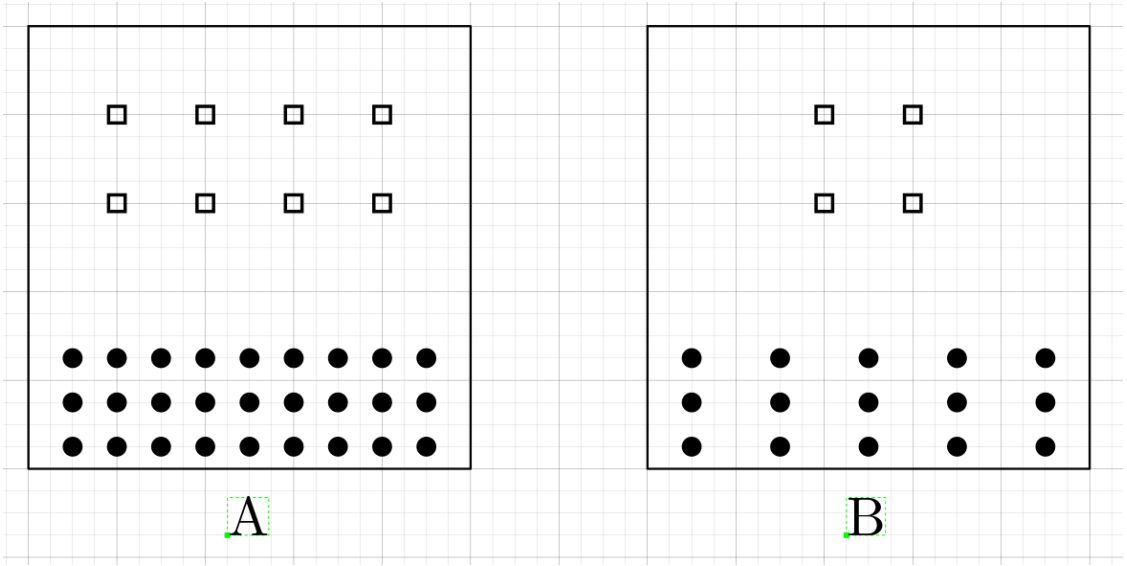
S: 27 D , 8 K

Axioms of Probability:

- Axiom 1: For any event A, $P(A) \geq 0$.
- Axiom 2: Probability of the sample space S is $P(S) = 1$.
- Axiom 3: If $A_1, A_2, A_3, \cdots$ are disjoint events, then $P(A_1 \cup A_2 \cup A_3 \cdots) = P(A_1) + P(A_2) + P(A_3) + \cdots$

https://www.probabilitycourse.com/chapter1

Olasılık 101 – Bayes Teoremi

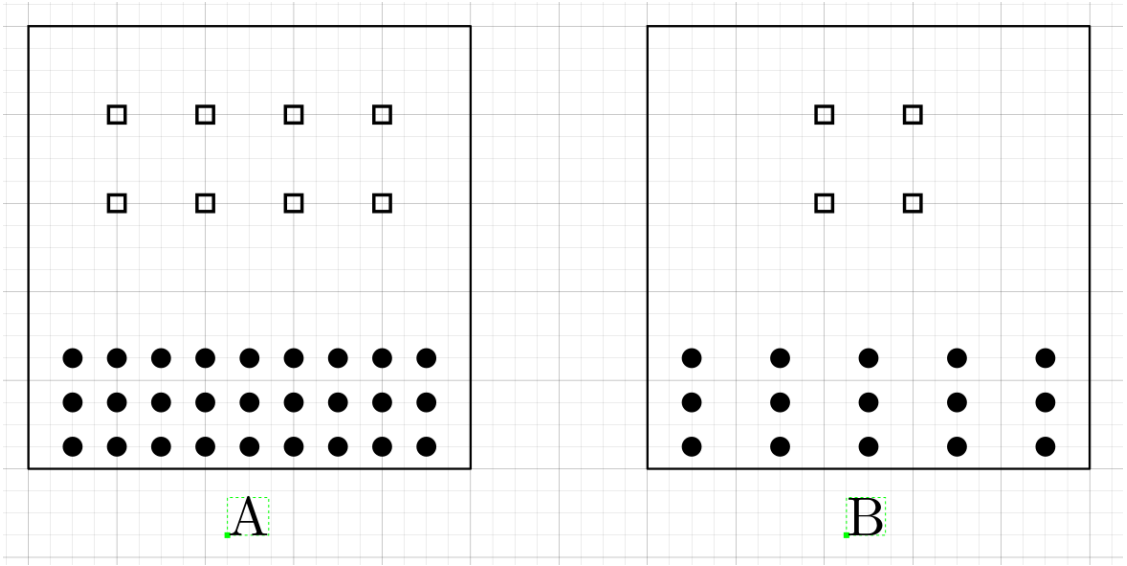


Kare ve Daire

A: 27 D , 8 K

B: 15 D , 4 K

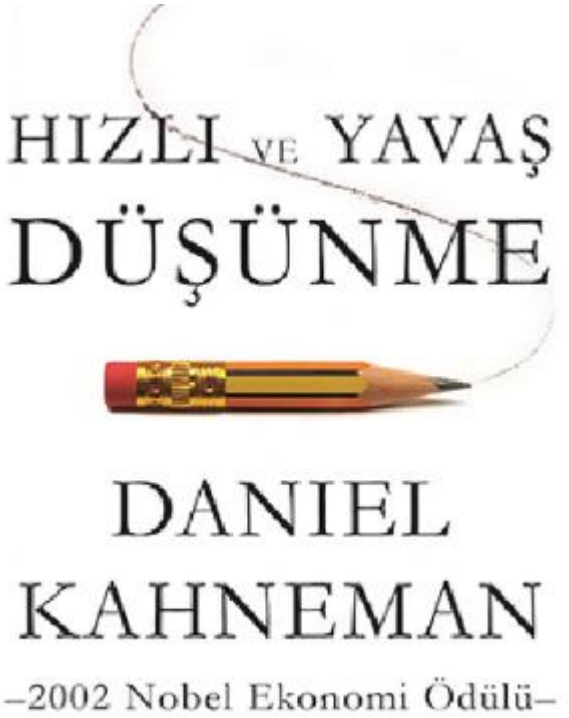
Olasılık 101 – Bayes Teoremi



Kare ve Daire

A: 27 D , 8 K

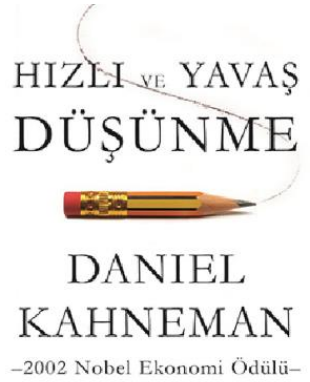
B: 15 D , 4 K



If A and B are two events in a sample space S, then the **conditional probability of** A **given** B is defined as

$$P(A|B) = \frac{P(A \cap B)}{P(B)}, \text{ when } P(B) > 0.$$

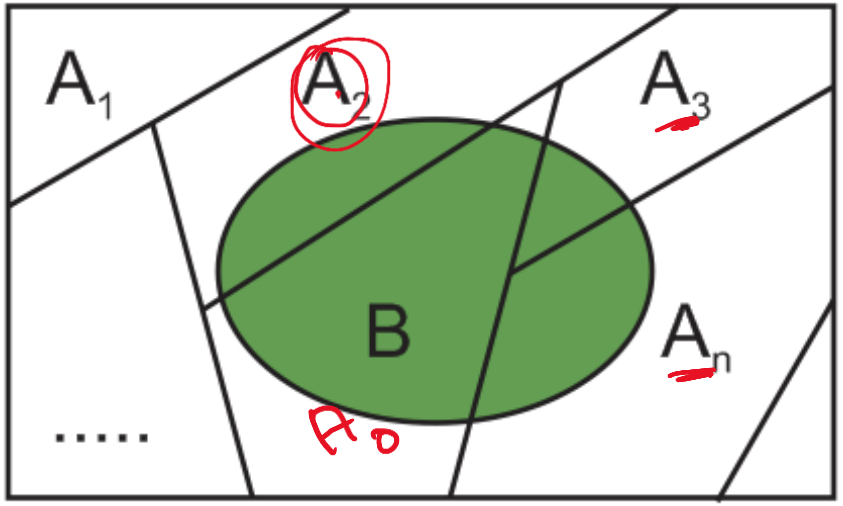
Olasılık 101 – Bayes Teoremi



Kim Bu? Çiftçi yada Kütüphaneci

- Kendisi içine kapanık ve duygusal biriydi.
- Sosyal çevresi pek yok, keni ağır işleriyle ilgilenmeyi sever.
- Oldukça entelektüel ve derinlikli fikir sahibi.

Olasılık 101 – Bayes Teoremi



$$P(B) = P(B|A_1) + P(B|A_3) + P(B|A_4) + P(B|A_0)$$

Olasılık 101 – Bayes Teoremi

- Bir toplumda kanser vakaları %0.1 olsun (0.001)
- Bir test cihazı hasta (C) iken %98 (+), (C') iken %95 (-) hassasiyetle hastaları tespit edebiliyor.
- Eğer cihaz bir kişiye (+) demiş ise bu kişinin kanser olma (C) ihtimali nedir?

$$P(C|+) = \frac{P(+|C)\,P(C)}{P(+)}$$

$P(+|C) = 0,98$

$P(C) = 0,001$

$P(+)$

$$P(+) = \boxed{P(+|C')\,P(C') + P(+|C)\,P(C)}$$

$0,05 \quad 0,999 \quad 0,98 \quad 0,001$

$$\frac{0,98 \times 0,001}{0,98 \times 0,001 + 0,05 \times 0,999}$$

Olasılık 101 – Bayes Teoremi

- Bir toplumda kanser vakaları %0.1 olsun (0.001)
- Bir test cihazı hasta (C) iken %98 (+), (C’) iken %95 (-) hassasiyetle hastaları tespit edebiliyor.
- Eğer cihaz bir kişiye (+) demiş ise bu kişinin kanser olMAma (C’) ihtimali nedir?

$$P(C'|+) = \frac{P(+|C') \cdot P(C')}{P(+|C') \cdot P(C') + P(+|C) \cdot P(C)} =$$

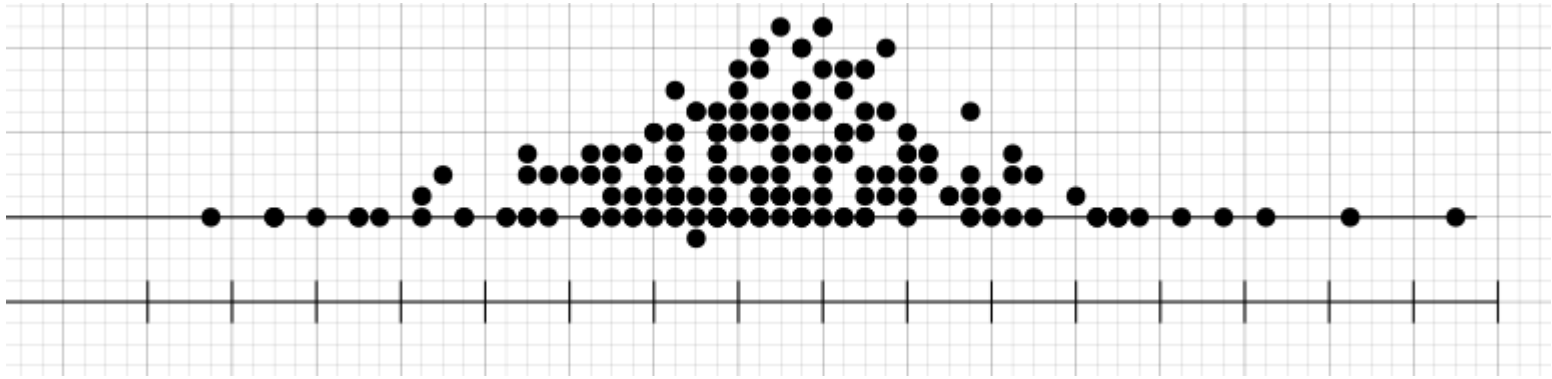
Olasılık 101 – Rassal Değişken ve İstatistikler

Olasılık 101 – Rassal Değişken ve İstatistikler

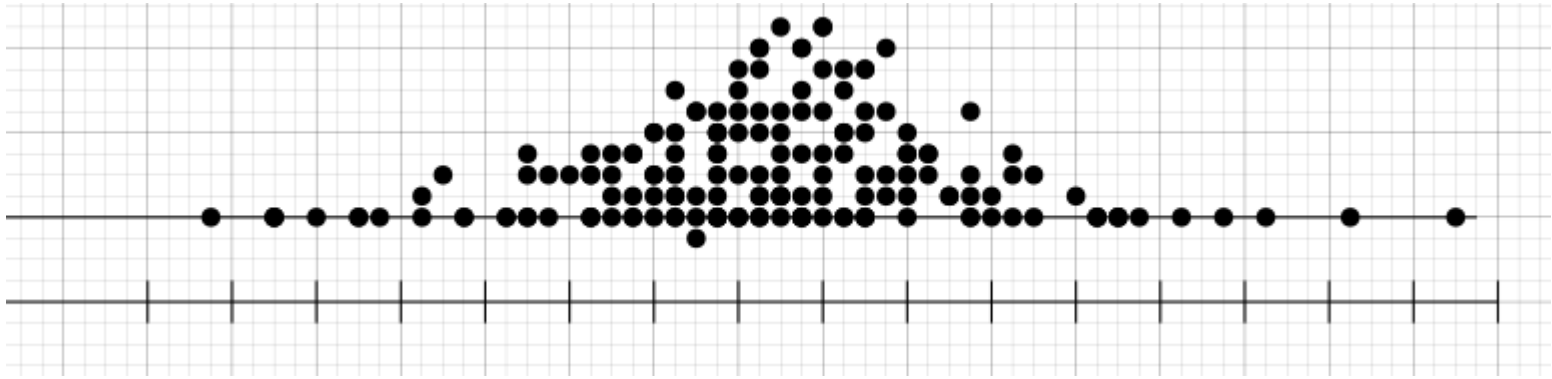
Olasılık 101 – Rassal Değişken ve İstatistikler

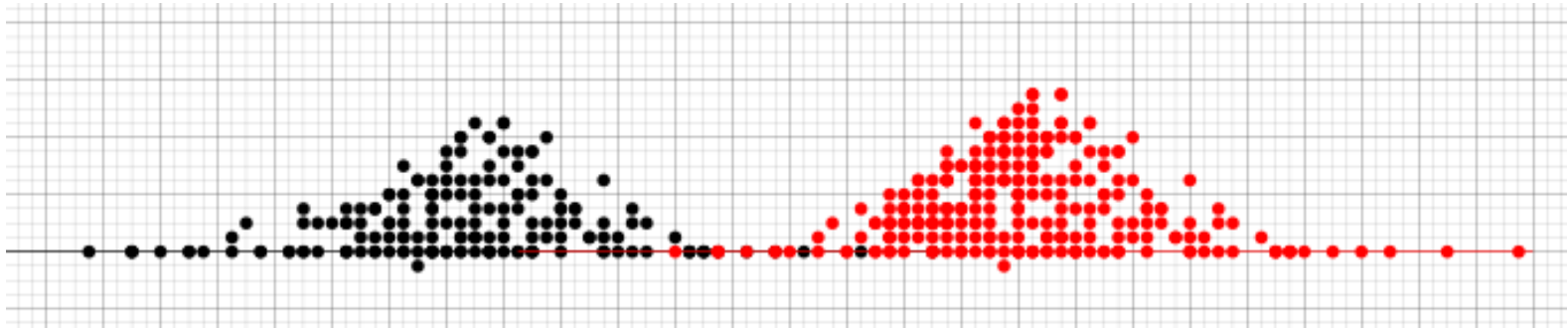
- Elimizde p(x) yoksa

Olasılık 101 – Rassal Değişken ve İstatistikler



Olasılık 101 – Rassal Değişken ve İstatistikler





Olasılık 101 – Rassal Değişken ve İstatistikler

- Denizden palamut çıkma olasılığı P(A) = 0.6
- Hamsi çıkma olasılığı P(B) = 0.4
- P(10 cm | B) = 0.5 , P(10 cm | A) = 0.2
- Eğer 10 cm bir balık geldiyse H mi P mi?

Olasılık 101 – Rassal Değişken ve İstatistikler

$$P(\text{class} \mid \text{data}) \qquad P(\text{data} \mid \text{class}) \times P(\text{class})$$

Olasılık 101 – Rassal Değişken ve İstatistikler

- Merhaba,
- Bugün gerçekleştirilen bir çekilişte tam 3M TL kazandınız.
- TC kimlik ve IBAN adresinizi xxxx adresine gönderdiğiniz takdirde size büyük ödül iletilecektir.

Uskumru (C1)		Palamut (C2)	
Uzunluk (cm)	Adet	Uzunluk (cm)	Adet
05-10	5	15-20	5
10-15	15	20-25	10
15-20	20	25-30	20
20-25	15	30-35	30
25-30	5	35-40	10
30-35	0	40-45	5
35-40	0	45-50	0

1. Sınıf olasılıklarını bul
2. Sınıf içi olasılıkları bul
3. Formülde yerine koy

$$P(C1|u = 22) = ?$$

Uskumru (C1)		Palamut (C2)	
Uzunluk (cm)	Adet	Uzunluk (cm)	Adet
05-10	5	15-20	5
10-15	15	20-25	10
15-20	20	25-30	20
20-25	15	30-35	30
25-30	5	35-40	10
30-35	0	40-45	5
35-40	0	45-50	0

1. Sınıf olasılıklarını bul
2. Sınıf içi olasılıkları bul
3. Formülde yerine koy

$$P(C1|u = 22) = ?$$

Playing Golf				
Weat.	**Temp.**	**Hum.**	**Windy**	**Class**
Rainy	Hot	High	False	No
Rainy	Hot	High	True	No
Sunny	Cool	Normal	True	No
Rainy	Mild	High	False	No
Sunny	Mild	High	True	No

Playing Golf				
Weat.	**Temp.**	**Hum.**	**Windy**	**Class**
Overcast	Hot	High	False	Yes
Sunny	Mild	High	False	Yes
Sunny	Cool	Normal	False	Yes
Overcast	Cool	Normal	True	Yes
Rainy	Cool	Normal	False	Yes
Sunny	Mild	Normal	False	Yes
Rainy	Mild	Normal	True	Yes
Overcast	Mild	High	True	Yes

1. Sınıf olasılıklarını bul
2. Sınıf içi olasılıkları bul
3. Formülde yerine koy

Weat	No
R	
S	
O	

Temp.	No
H	
M	
C	

Hum.	No
High	
Norm.	

Windy	No
True	
False	

https://www.geeksforgeeks.org/naive-bayes-classifiers/

Playing Golf				
Weat.	**Temp.**	**Hum.**	**Windy**	**Class**
Rainy	Hot	High	False	No
Rainy	Hot	High	True	No
Sunny	Cool	Normal	True	No
Rainy	Mild	High	False	No
Sunny	Mild	High	True	No

Playing Golf				
Weat.	**Temp.**	**Hum.**	**Windy**	**Class**
Overcast	Hot	High	False	Yes
Sunny	Mild	High	False	Yes
Sunny	Cool	Normal	False	Yes
Overcast	Cool	Normal	True	Yes
Rainy	Cool	Normal	False	Yes
Sunny	Mild	Normal	False	Yes
Rainy	Mild	Normal	True	Yes
Overcast	Mild	High	True	Yes

1. Sınıf olasılıklarını bul
2. Sınıf içi olasılıkları bul
3. Formülde yerine koy

Weat	Yes
R	
S	
O	

Temp.	Yes
H	
M	
C	

Hum.	Yes
High	
Norm.	

Windy	Yes
True	
False	

https://www.geeksforgeeks.org/naive-bayes-classifiers/

Playing Golf				
Weat.	**Temp.**	**Hum.**	**Windy**	**Class**
Rainy	Hot	High	False	No
Rainy	Hot	High	True	No
Sunny	Cool	Normal	True	No
Rainy	Mild	High	False	No
Sunny	Mild	High	True	No

Playing Golf				
Weat.	**Temp.**	**Hum.**	**Windy**	**Class**
Overcast	Hot	High	False	Yes
Sunny	Mild	High	False	Yes
Sunny	Cool	Normal	False	Yes
Overcast	Cool	Normal	True	Yes
Rainy	Cool	Normal	False	Yes
Sunny	Mild	Normal	False	Yes
Rainy	Mild	Normal	True	Yes
Overcast	Mild	High	True	Yes

1. Sınıf olasılıklarını bul
2. Sınıf içi olasılıkları bul
3. Formülde yerine koy

```
today = (Sunny, Hot, Normal, False)
```

https://www.geeksforgeeks.org/naive-bayes-classifiers/

- Bayes Sınıflandırıcısı +/-

Advantages of Naive Bayes

The **Naive Bayes is a popular algorithm** due to its following advantages:

- This algorithm works very fast and can easily predict the class of a test dataset.
- You can use it to solve multi-class prediction problems as it's quite useful with them.
- Naive Bayes classifier performs better than other models with less training data if the assumption of independence of features holds.
- If you have categorical input variables, the Naive Bayes algorithm performs exceptionally well in comparison to numerical variables.
- It can be used for Binary and Multi-class Classifications.
- It effectively works in Multi-class predictions.

https://www.upgrad.com/blog/naive-bayes-classifier/

- Bayes Sınıflandırıcısı +/-

Disadvantages of Naive Bayes

- If your test data set has a categorical variable of a category that wasn't present in the training data set, the Naive Bayes model will assign it zero probability and won't be able to make any predictions in this regard. This phenomenon is called 'Zero Frequency,' and you'll have to use a smoothing technique to solve this problem.
- This algorithm is also notorious as a lousy estimator. So, you shouldn't take the probability outputs of 'predict_proba' too seriously.
- It assumes that all the features are independent. While it might sound great in theory, in real life, you'll hardly find a set of independent features.

https://www.upgrad.com/blog/naive-bayes-classifier/

Sınıflandırma Metrikleri

Hata Matrisi (Confusion Matrix)

Sınıflandırma Metrikleri

Doğruluk (Accuracy) Kesinlik (Precision) Hassasiyet (Recal)

Sınıflandırma Metrikleri

Doğruluk (Accuracy) Kesinlik (Precision) Hassasiyet (Recal)

Sınıflandırma Metrikleri

F Skor