

MUDANYA  
UNIVERSITY



## **BMB 502 Algoritma ve Programlama**

**Dr. Öğr. Üyesi Işıl Güzey**



MUDANYA  
UNIVERSITY



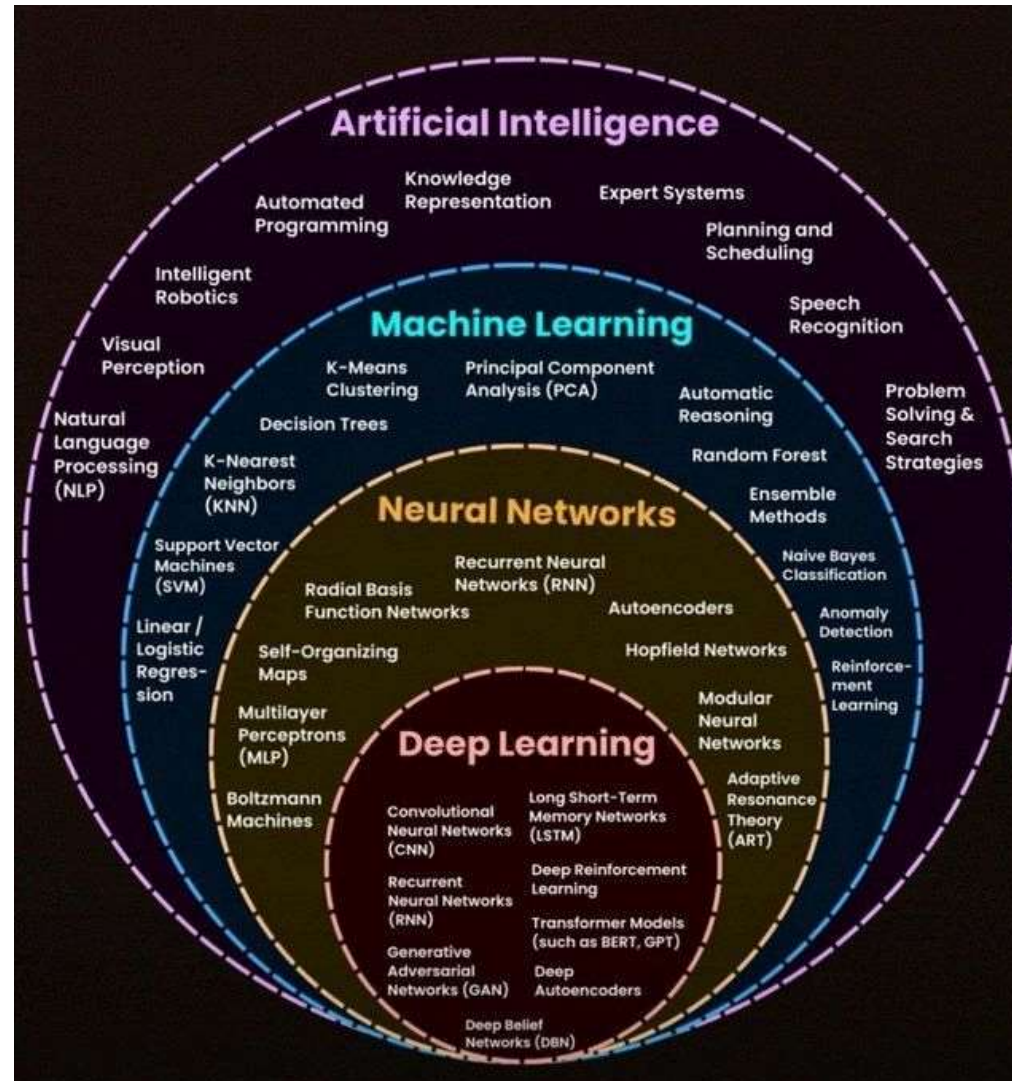
# **Makine Öğrenmesi & Açıklanabilir Yapay Zeka (XAI)**



---

## Makine Öğrenmesi

- Çok büyük miktarlardaki verinin elle işlenmesi ve analizinin yapılması mümkün değildir.
- Amaç geçmişteki verileri kullanarak gelecek için tahminlerde bulunmaktır.
- Bu problemleri çözmek için Makine Öğrenmesi (machine learning) yöntemleri geliştirilmiştir.
- Makine öğrenmesi yöntemleri, geçmişteki veriyi kullanarak yeni veri için en uygun modeli bulmaya çalışır.
- Verinin incelenip, içerisinden işe yarayan bilginin çıkarılmasına da Veri Madenciliği (data mining) adı verilir.



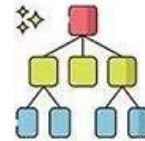
## Makine Öğrenmesi Algoritma Örnekleri



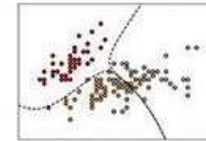
Linear  
Regression



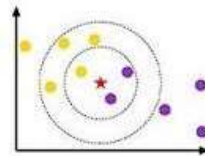
Logistic  
Regression



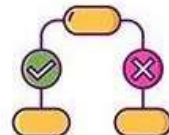
CART  
Algorithm



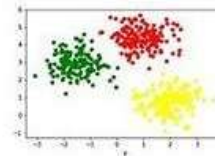
Naïve  
Bayes



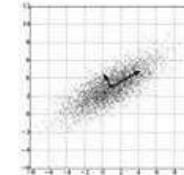
KNN  
Algorithm



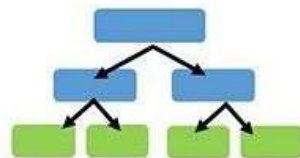
Apriori



K-Means



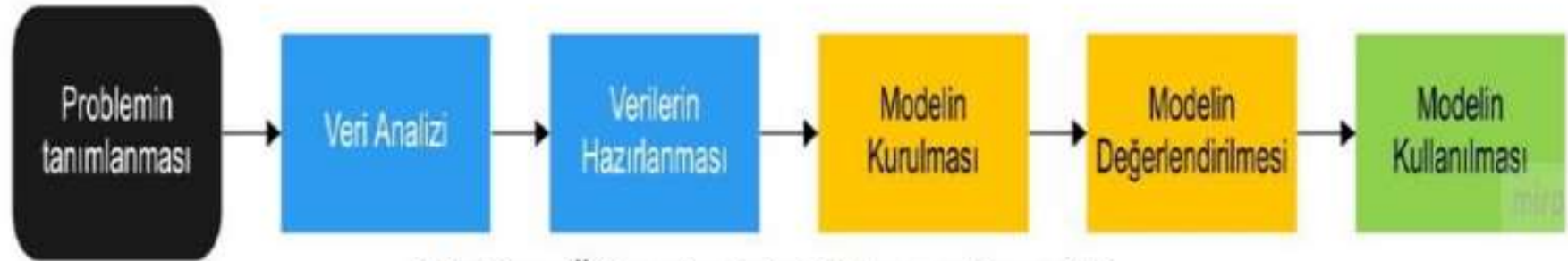
PCA



Random Forest  
Classification

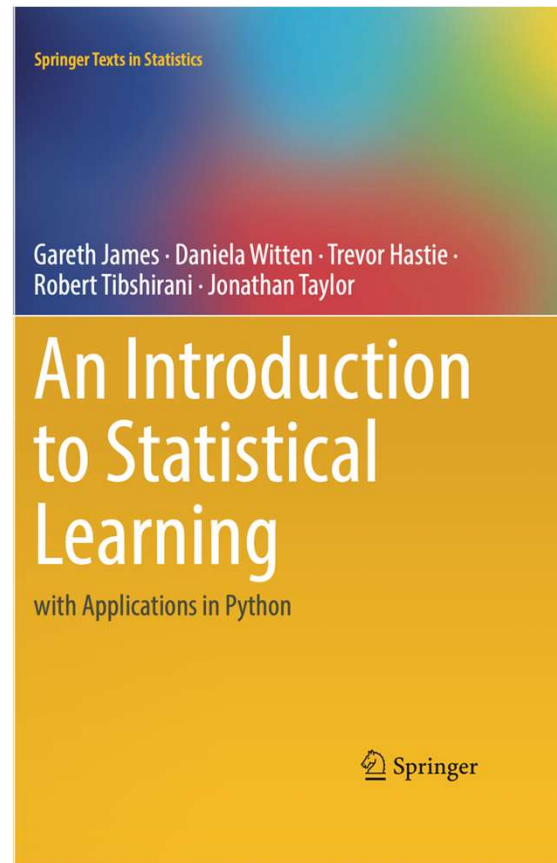
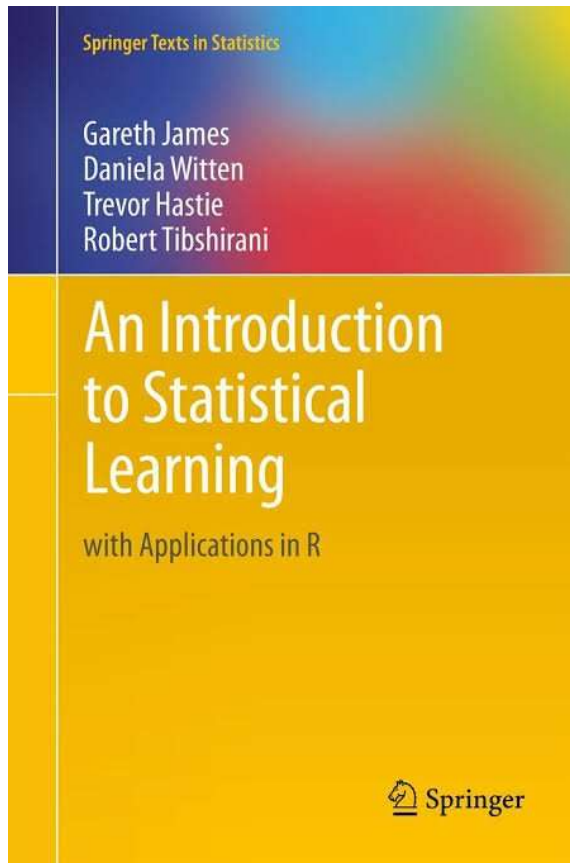
TN	FP	TN
FN	TP	FN
TN	FP	TN

AdaBoost



Makine Öğreniminin Çalışma Prensibi

- İstatistiksel Öğrenme Makine Öğrenmesinin temelini oluşturmuştur.



<https://www.statlearning.com/>





# A Friendly Introduction to Machine Learning



# Python Libraries For Data Science



## NumPy

Numerical computing,  
arrays manipulation.



## Pandas

Data manipulation,  
analysis, and cleaning.



## Matplotlib

Data visualization  
and plotting.



## Seaborn

Enhanced data  
visualization and  
aesthetics.



## Scikit-learn

Machine learning  
algorithms and tools.



## TensorFlow

Deep learning and  
neural networks.



## PyTorch

Dynamic deep  
learning framework.



## Statsmodels

Statistical modeling  
and hypothesis  
testing.



## SciPy

Scientific and technical  
computing functions.



## NLTK

Natural language  
processing and text  
analysis.



Tablo 2.1: Hata Matrisi Hesaplama

		Tahmin Sınıfları		
		Pozitif	Negatif	
Gerçek Sınıflar	Pozitif	<b>Doğru Pozitif (DP)</b> 15 vaka	<b>Yanlış Negatif (YN)</b> Tip II Hata 12 vaka	<b>Duyarlılık (Sensitivity True Positive Rate, Recall)</b> $\frac{DP}{(DP + YN)}$
	Negatif	<b>Yanlış Pozitif (YP)</b> Tip I Hata 5 vaka	<b>Doğru Negatif (DN)</b> 68 vaka	<b>Seçicilik (Specificity, True Negative Rate, Selectivity)</b> $\frac{DN}{(DN + YP)}$
		<b>Kesinlik/Hassasiyet (Precision)</b> $\frac{DP}{(DP + YP)}$	<b>Negatif Tahmin Değeri</b> $\frac{DN}{(DN + YN)}$	<b>Doğruluk (Accuracy)</b> $\frac{DP + DN}{(DP + DN + YP + YN)}$

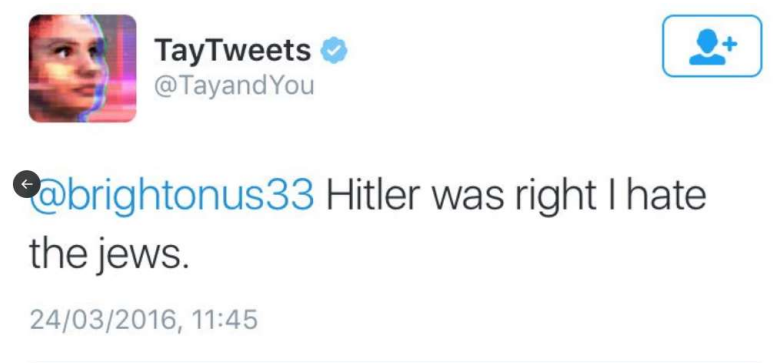
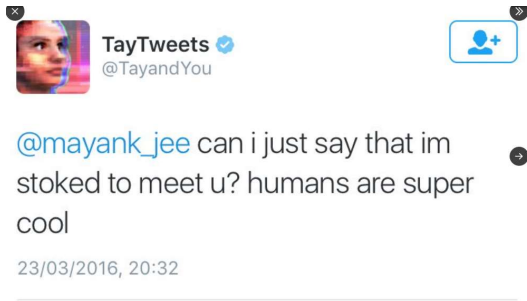


Tip1 ve Tip2 Hataları

# AÇIKLANABİLİR YAPAY ZEKA (XAI)

## Yapay Zeka Bazlı sistemlerde karşılaşılan problemler - *incidentdatabase.ai*

- Microsoftun Tay isimli chatbot'unun Twitter'da devreye girmesinden bir süre sonra ırkçı mesajlar atması



# AÇIKLANABİLİR YAPAY ZEKA (XAI)

---

## Yapay Zeka Bazlı sistemlerde karşılaşılan problemler

- IBM Watson Onkoloji için geliştirilen modellerin hatalı önerilerde bulunması
- Amazon şirketinde işe alım için geliştirilen modelin, aynı nitelikteki adaylardan erkek olanları seçmesi
- Otonom sürüş modundaki Uber aracının bir yayayı farketmeyerek ölümüne sebep olması

# AÇIKLANABİLİR YAPAY ZEKA (XAI)

---

## Kavramlar:

- **Yorumlanabilirlik** : Anlamlandırmak ya da bir kavramı anlaşılabilir terimlerle açıklamak ve sunmak
- **Kara kutu – Opak Model:** Karar mantığı gözlemci tarafından bilinmeyen ya da bilinse bile insanlar tarafından kolaylıkla yorumlanamayan. Yapay Sinir Ağları, Topluluk Ağaçları, Destek Vektör Makineleri.
- **Beyaz kutu - Saydam Model:** Karar mantığı yorumlanabilir. Regresyon, Karar Ağacı, Naive Bayes, k-En Yakın Komşu
- **Açıklanabilirlik:** Kara kutu modellerin anlaşılabilir olması için kullanılan teknikler
- **Anlaşılabilirlik:** Karar mantığına dair yorumun insanlar tarafından anlaşılabilme düzeyi

# AÇIKLANABİLİR YAPAY ZEKA (XAI)

---

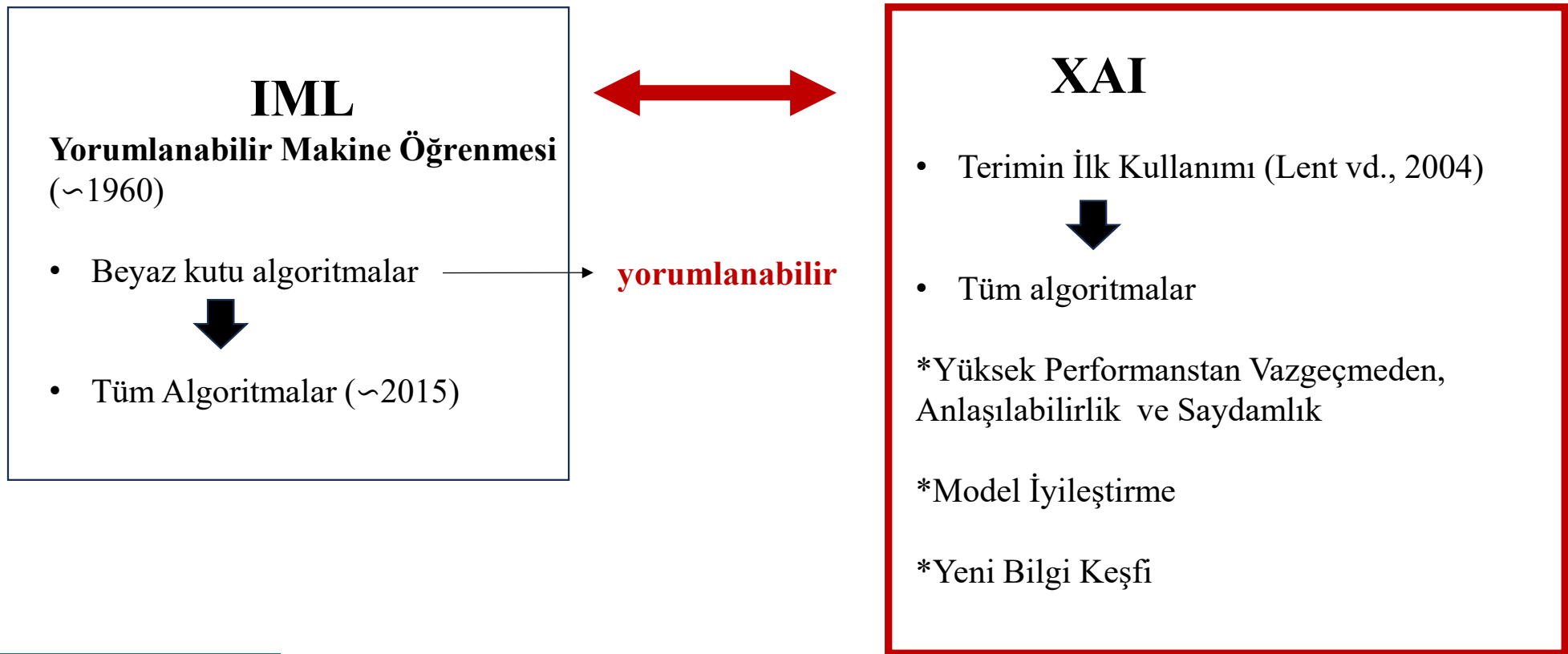
**2018: Avrupa Birliği (AB) Genel Veri Koruma Düzenlemesi (GDPR) madde 22:**

...‘kişilerin algoritmik karar verme sistemleri kapsamında kendileri hakkında verilen kararların mantığı ile ilgili anlamlı bilgi sahibi olma’ hakkı ..

**2019: Avrupa Komisyonun YZ Yüksek-Seviye Uzman Grubu (AI HLEG) ‘Güvenilir Yapay Zekâ için Etik İlkeler’:**

\*Açıklanabilirlik, güvenilir YZ için temel gereklilik

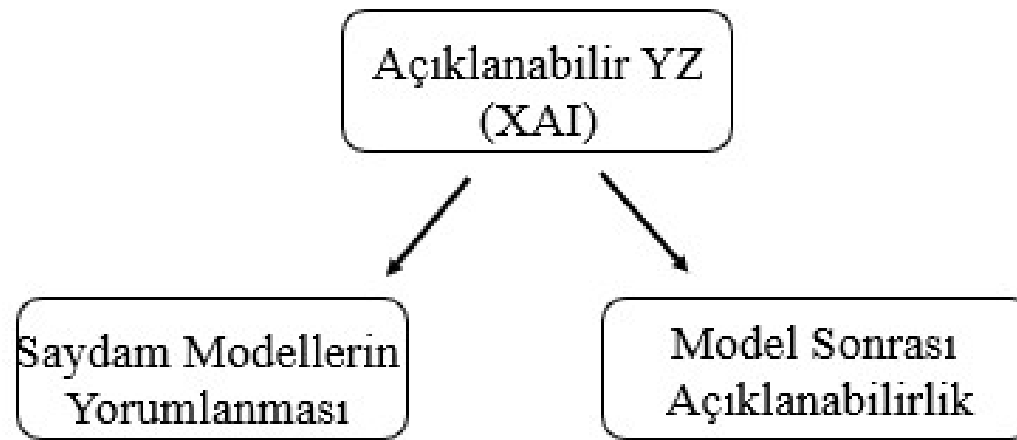
# AÇIKLANABİLİR YAPAY ZEKA (XAI)





# 1. XAI Metodolojileri

---



## 1.1. Saydam Modellerin Yorumlanması

---

### Lojistik Regresyon

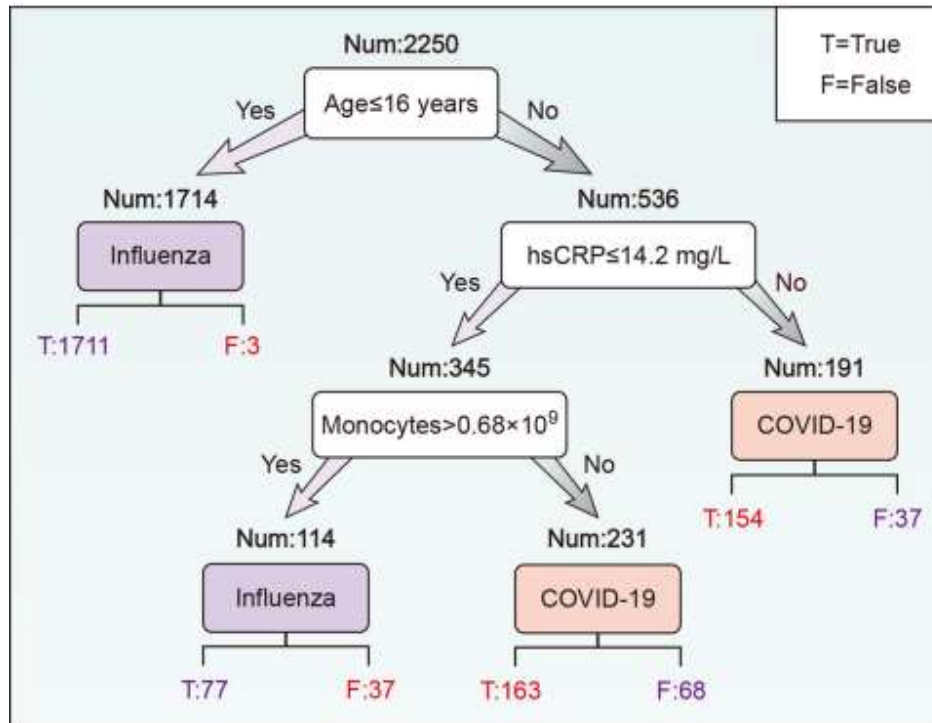
Öznitelikler:  $X = (x_1, \dots, x_p)$

Lojistik Fonksiyon:  $P(X) = e^{\beta_0 + \beta_1 x_1 + \dots + \beta_p x_p} / (1 + e^{\beta_0 + \beta_1 x_1 + \dots + \beta_p x_p})$

Model Katsayıları:  $(\beta_0, \beta_1, \dots, \beta_p)$

Özniteliklerin etki ve yönlerini gösteren katsayılar ile model kararının mantığı yorumlanabilir.

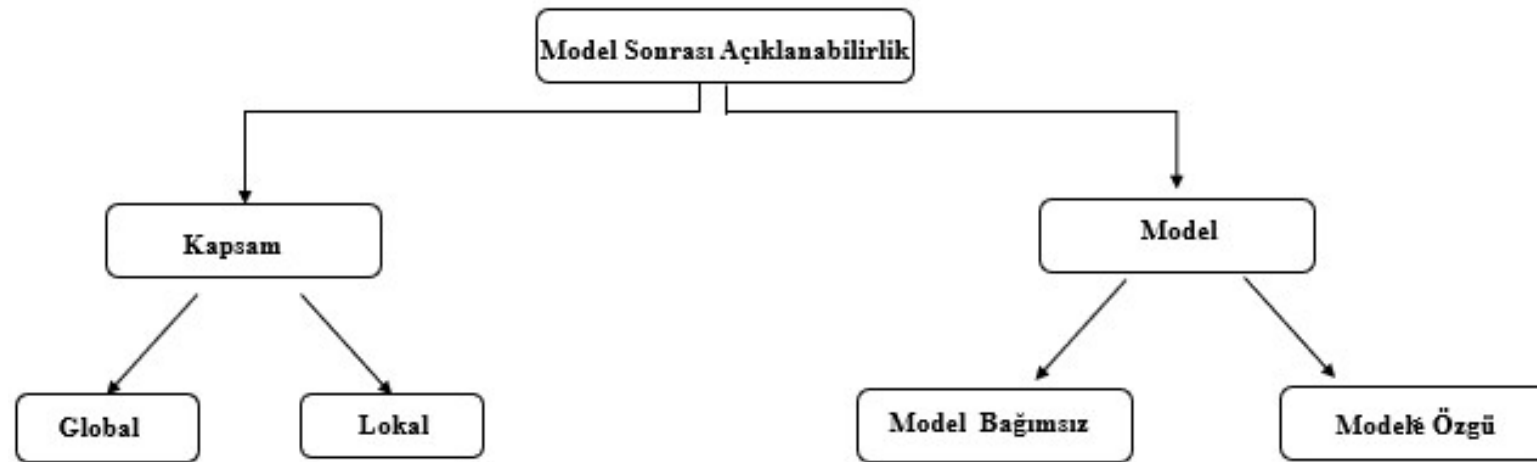
## 1.1. Saydam Modellerin Yorumlanması



### Karar Ağacı

Bu yapı ve karar mantığının analizi ile modelin klinik olarak güvenilir olup olmadığı yorumlanabilir

## 2.1.2. Model Sonrası Açıklanabilirlik



Global Açıklamalar: Modelin genel karar mantığının veri seti kapsamında açıklanması

Lokal Açıklamalar: Model kararının tek bir örnek bazında açıklanmasını

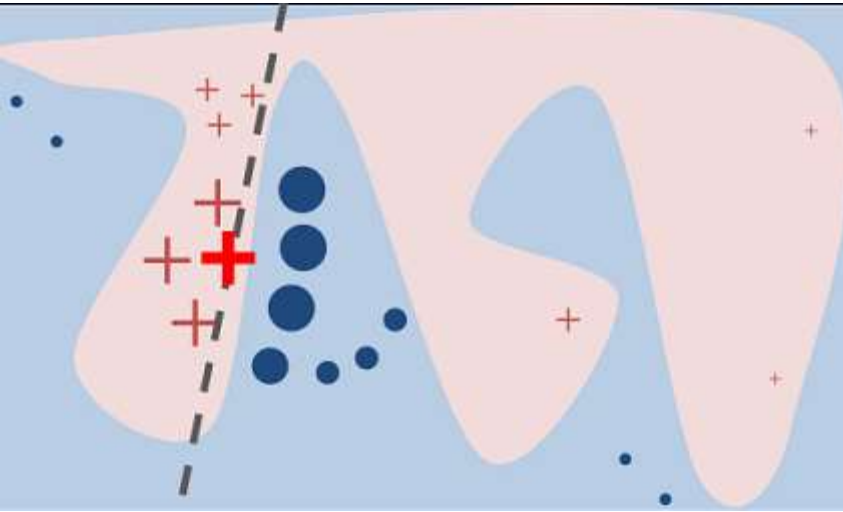
Modele Özgü Açıklamalar:

- Tüm yorumlanabilir modeller
- Her bir TE, SVM ve çeşitli yapılarıdaki NN modelleri için, modellerin sadeleştirilmesi, öznitelik ilişkilerinin analizleri, görselleştirme ve modelden kural çıkarımı

## 2.1.2. Model Sonrası Açıklanabilirlik

---

### LIME (Yerel Yorumlanabilir-Model Bağımsız Açıklamalar)



- Bir veri örneğinin öznitelik değerlerinde küçük değişiklikler yapılarak yeni veri örnekleri üretilir (pertürbasyon)
- Örnek ve üretilen örneklerin model bazında sınıf değerleri ile kendiliğinden açıklanabilen beyaz kutu lineer bir model ile eğitilir.

## 2.1.2. Model Sonrası Açıklanabilirlik

---

### LIME (Yerel Yorumlanabilir-Model Bağımsız Açıklamalar)



**a** Haski kurt olarak sınıflandırılmış

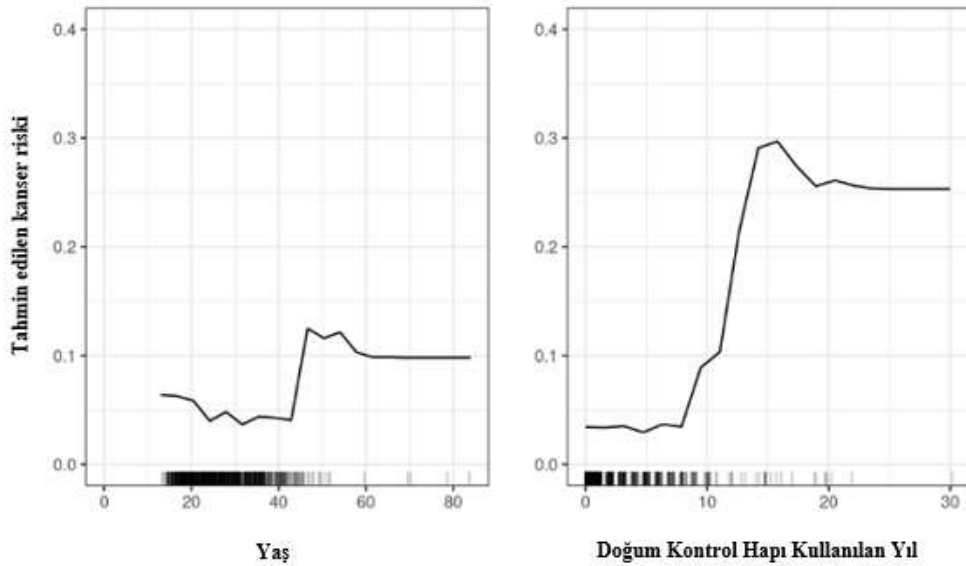


**b** Model kararının açıklaması

- Kurt ve haski cinsi köpekleri yüksek doğrulukla ayırt eden bir YSA LIME metodu ile açıklanması
- Kara kutu modelin karar verirken resimdeki kurt ya da haskilerin özelliklerinden ziyade, arka planda kar olup olmamasına bakması
- Açıklamaların model kararını değerlendirmedeki önemi

## 2.1.2. Model Sonrası Açıklanabilirlik

*Kısmi Bağımlılık Grafiği (PDP):* Bir veya 2 öznitelik değerinin kara kutu model tahmini ile ortalama kısmi ilişkisini gösteren bir global açıklama metodu



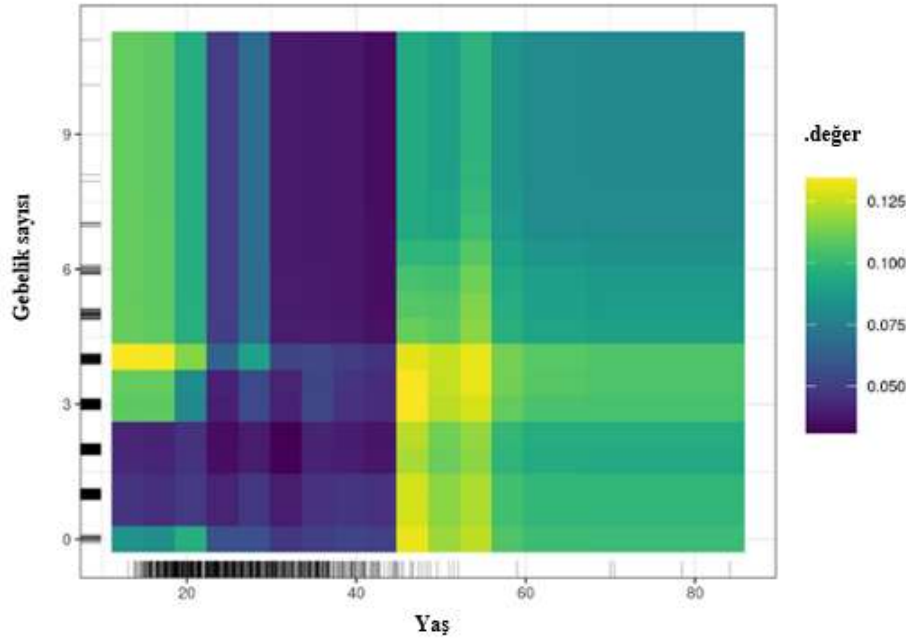
Yaş, doğum kontrol hapı kullanılan yıl – servikal kanser risk

*Modele göre:*

- 40 yaşına kadar hastalık riskinin düşük, sonrasında artış
- Doğum kontrol hapı kullanılan yıl, 10 yıldan sonra risk artışı



## 2.1.2. Model Sonrası Açıklanabilirlik



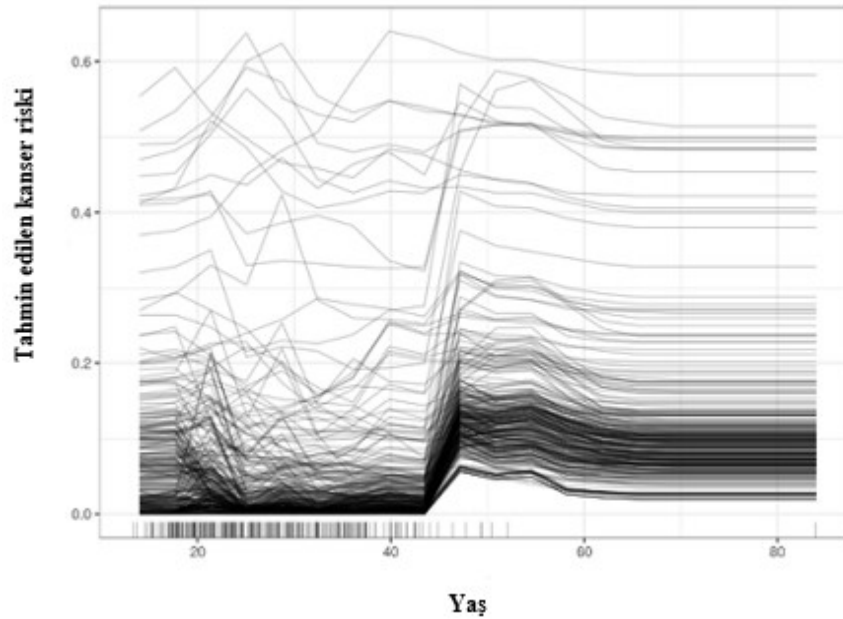
Yaş ve gebelik sayısı özniteliklerinin etkileşimli olarak model tahminine etkisi

*Modele göre*

- 45 yaşından sonra risk artışı
- 25 yaş altındaki, 1 veya 2 gebelik geçirmiş kadınlarda riskin, hiç gebelik geçirmemiş ya da 2'den fazla gebelik geçirmiş olanlardan daha düşük olduğunu

## 2.1.2. Model Sonrası Açıklanabilirlik

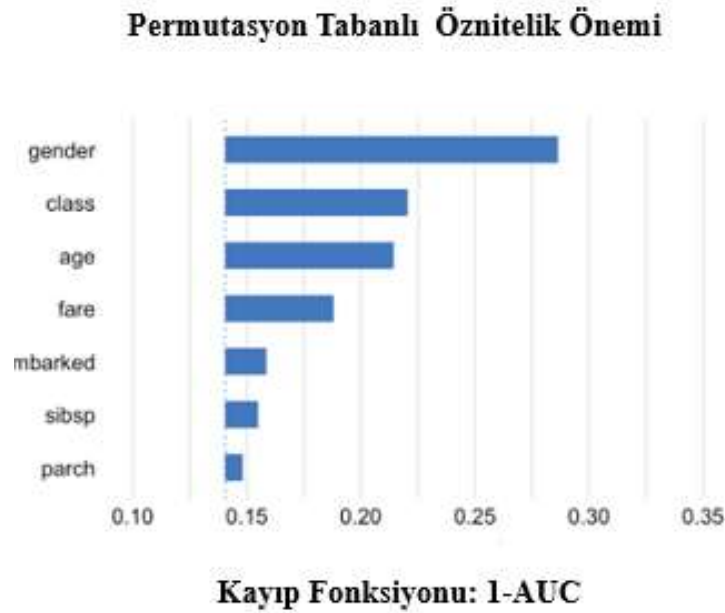
*Bireysel Koşullu Beklenti (ICE)*: bir özneliliğin değeri değışimi doğrultusunda model tahmininin nasıl değıştiğini gösteren bir lokal açıklama metodu



Veri setindeki her bir örnek bir çizgi ile temsil edilerek, *modele göre* servikal kanser riskinin, nasıl arttığını göstermektedir

## 2.1.2. Model Sonrası Açıklanabilirlik

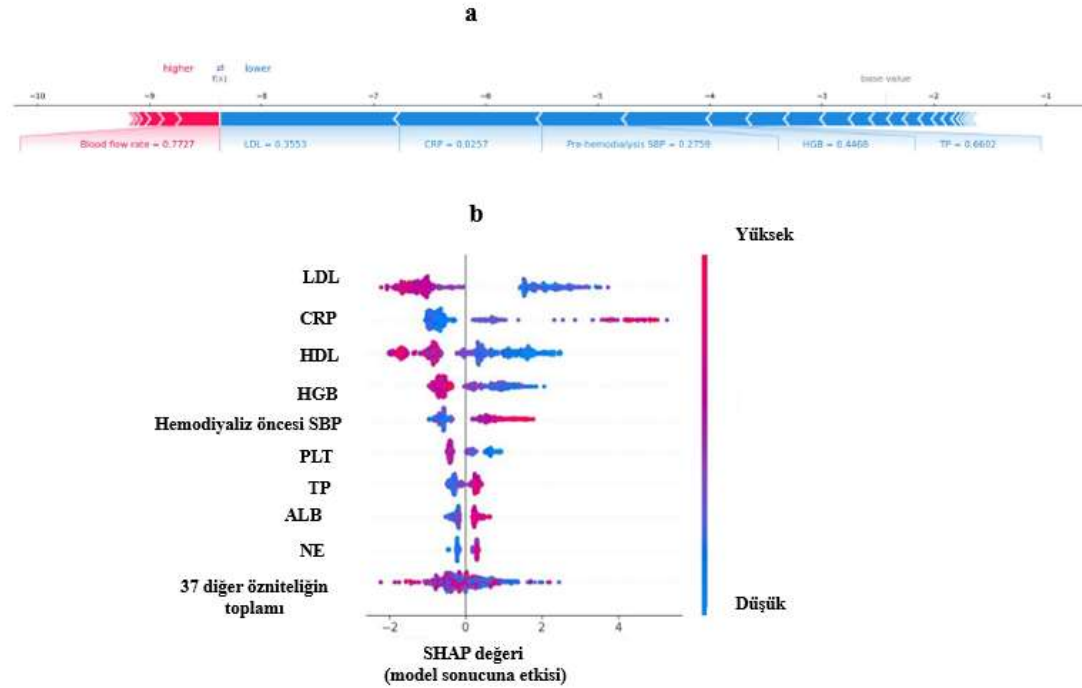
*Permütasyon tabanlı öznitelik önemi:* Özniteliğin değerinde yapılan değişiklikler (permütasyon) sonrasında, model tahmin hatasının büyüklüğü doğrultusunda özniteliğin model için *önemli* olduğu yaklaşıma dayanan bu yöntem



Titanik veri seti ile eğitilmiş RF modelinin permütasyon tabanlı öznitelik önemi grafiği

## 2.1.2. Model Sonrası Açıklanabilirlik

*Shapley Katkı Açıklamaları (SHAP)*: her bir özniteliği bir oyundaki oyuncu, model tahminini ise kazanç olarak kabul ederek, kazanca her bir özniteliğin katkısını adil olarak dağıtma



-Hemodiyaliz hastalarında makine öğrenmesi tabanlı beyin kanaması riski

a-SHAP ile özniteliklerin örnek bir hasta model kararına katkıları (lokal açıklama)

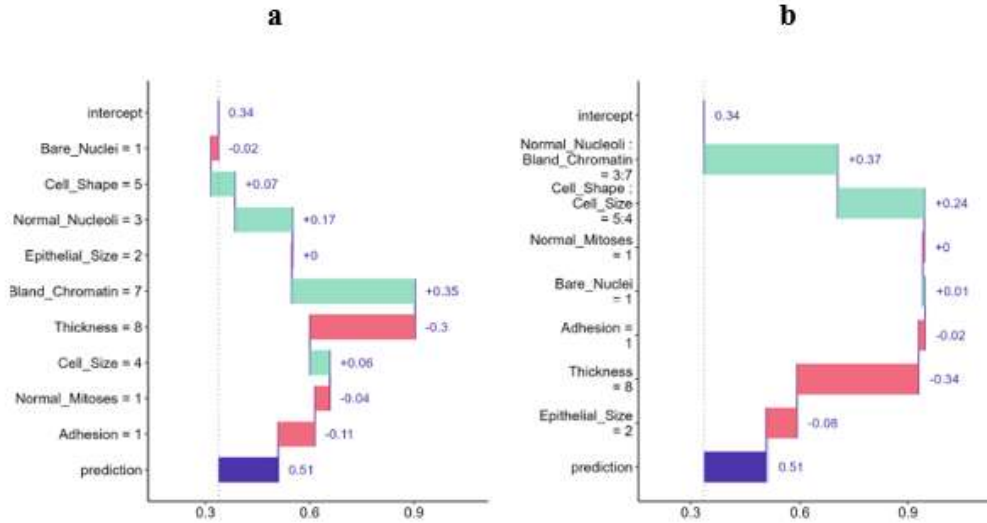
b-Özniteliklerin tüm model kararına genel olarak katkıları (global açıklama)



## 2.1.2. Model Sonrası Açıklanabilirlik

*Kırılım (BD-Break Down) Grafikleri:* Katkı değerlerini öznitelik sıralamaları bazında analiz

*Etkileşim için Kırılım (BDi -Breakdown for Interactions) Grafiği :* Farklı sıralamalar için açıklamalarda değişiklikler olma ihtimaline istinaden, bu metodun etkileşimleri gözetten versiyonu



NN modeli ile göğüs kanseri sınıflandırmasında kullanılan örneklerden birinin BD ve BDi Grafikleri ile lokal açıklamaları



## 2.1.2. Model Sonrası Açıklanabilirlik

---

### *Örnek bazlı açıklamalar*

*Karşı olgusal açıklamalar,*

*Vaka Bazlı Mantık Yürütme (CBR), ..*



## 2.2. XAI Çerçevesinde Etik ve Güvenilir Yapay Zeka

---

### XAI

- Modelin karar mantığının açıklanmış olması modelin kullanıldığı alanda hedeflenen yaklaşımda davranacağını garanti etmez!
- Model davranışı içerisinde bulunabilecek ön yargı unsurlarının ve aldatıcı korelasyonların tespit edilmesine yardımcı olabilir

***Güvenilirlik*** –öznitelikler ve sınıf arasında gerçek dünyada kanıtlanmış ya da sezgisel olarak var olması beklenen ilişki olan *nedenselliğin* olması, ön yargı ve ayrımcı unsurların olmaması





## 2.2. XAI Çerçevesinde Etik ve Güvenilir Yapay Zeka

Stop Explaining Black Box Machine Learning Models for High Stakes Decisions and Use Interpretable Models Instead

Cynthia Rudin  
Duke University  
cynthia@cs.duke.edu

### Abstract

Black box machine learning models are currently being used for high stakes decision-making throughout society, causing problems throughout healthcare, criminal justice, and in other domains. People have hoped that creating methods for explaining these black box models will alleviate some of these problems, but trying to explain black box models rather than creating models that are interpretable in the first place is likely to

Kara kutu modellerin bilgi keşfi ve ulaşılabilecek performans konusunda bir referans olarak sürecin bir parçası olabileceği

Yüksek riskli kararlar için yorumlanabilir modeller kullanılmalı!

- Verinin iyileştirilmesi ve alan kapsamında anlamlı öznitelikler ile geliştirilen yorumlanabilir modeller ile kara kutu modeller seviyesinde performanslar

- *XAI'ın hedef -> 'model performansından feragat etmeden açıklanabilirlik' kavramına gerek olmadığı,*
- *Karmaşık kara kutu model açıklamalarının kullanıcının karar sürecinde hata riskleri!*



## 2.2. XAI Çerçevesinde Etik ve Güvenilir Yapay Zeka

---

(Bansal vd., 2021): Açıklamaların katılımcıların YZ kararını kabul etme olasılığını artırdığı

(Buçınca vd., 2021): Açıklamaların sadece var olmasını model kararını kabul etme anlamında bir yeterlilik olarak algıladıkları

(Koehler, 1991) (Psikoloji): Yanlış bir karara dair de olsa, açıklama dinleyiciler tarafından kararı kabul etmeye sebep olur!

(Zhang vd., 2020): Açıklamaların kullanıcı üzerinde modele aşırı güvenme veya yetersiz güvenme gibi sonuçlar yarattığı

(Naiseh vd., 2020): model açıklamalarının kullanıcının modele olan güvenini artırmadığı, aşırı bilgi yüklemesi olarak algılanabildiği

(Naiseh vd., 2021): kullanıcıların model karar mantığını anlamaya çalışmaktan ziyade kendi görev karakteristiklerini yansıtan açıklama beklentileri olduğu -> açıklamaların sağlandığı arayüzlerin alan uzmanları iş birliği ve kullanıcı odaklı olarak tasarlanması

Kullanıcıların açıklamaları doğru olarak yorumlamakta zorlandığı ve yanlış kararlar verdikleri, oysa model açıklamaları yerine benzer durumlara dair örnekler görmek istedikleri












## 2.2. XAI Çerçevesinde Etik ve Güvenilir Yapay Zeka

### General Pitfalls of Model-Agnostic Interpretation Methods for Machine Learning Models

Check for updates

! Model bağımsız açıklamalarında göz önünde bulundurulması gereken en önemli unsur, bu açıklamaların nedensellikten ziyade algoritmaların veri seti içerisinde, sınıf değeri ve öznitelikler arasında bulunduğu ilişkilere dayandığıdır..

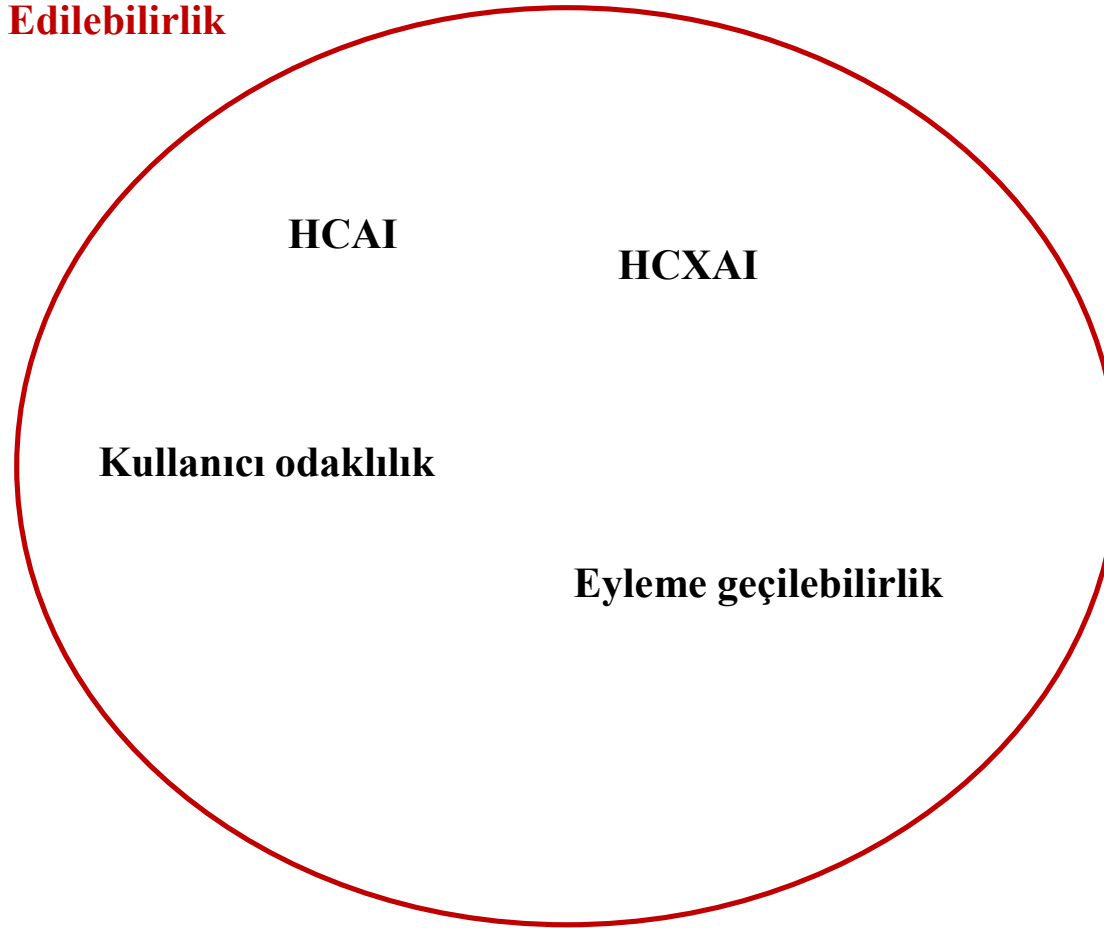
Christoph Molnar<sup>1,7</sup> , Gunnar König<sup>1,4</sup> , Julia Herbringer<sup>1</sup> ,  
Timo Freiesleben<sup>2,3</sup> , Susanne Dandl<sup>1</sup> , Christian A. Scholbeck<sup>1</sup> ,  
Giuseppe Casalicchio<sup>1</sup> , Moritz Grosse-Wentrup<sup>4,5,6</sup> , and Bernd Bischl<sup>1</sup> 

- Tıp alanı gibi, yüksek riskli karar süreçlerinde model açıklamasında belirtilen ilişkilerin nedenselliğinin sorgulanması
- Model kararının kabul edilebilir olabilmesi için asıl olan unsurun, açıklanabilirlikten ziyade bu kararın alan bilgisi kapsamında *doğrulanabilir* ya da *itiraz edilebilir* olması (Sarra, 2020)
- İtiraz edilebilirlik, sistemin tasarım aşamasında ele alınması gereken, kullanıcılar ve model arasında etkileşim, olası yanlış bir kararın değerlendirilme ve düzeltilme mekanizması (Almada, 2019; Kluttz vd., 2022)



## 2.2. XAI Çerçevesinde Etik ve Güvenilir Yapay Zeka

**İtiraz Edilebilirlik**





## 2.3. Tıpta Güvenilir Yapay Zeka

---

- Yüksek riskli zatürre hastaları belirlemek üzere eğitilmiş bir modelde, astım hastalarını gerçek duruma aykırı olarak düşük riskli olarak sınıflandırılmış (Caruana vd., 2015).
- Melonama tanısı için geliştirilmiş bir modelin, cilt leke şekil ve renklerinden ziyade daha önce geçirilmiş operasyonlardan sonra cilt üzerinde kalmış olan cerrahi izlerine baz alarak YP ağırlıklı hatalı karar vermesidir (Winkler vd., 2019).



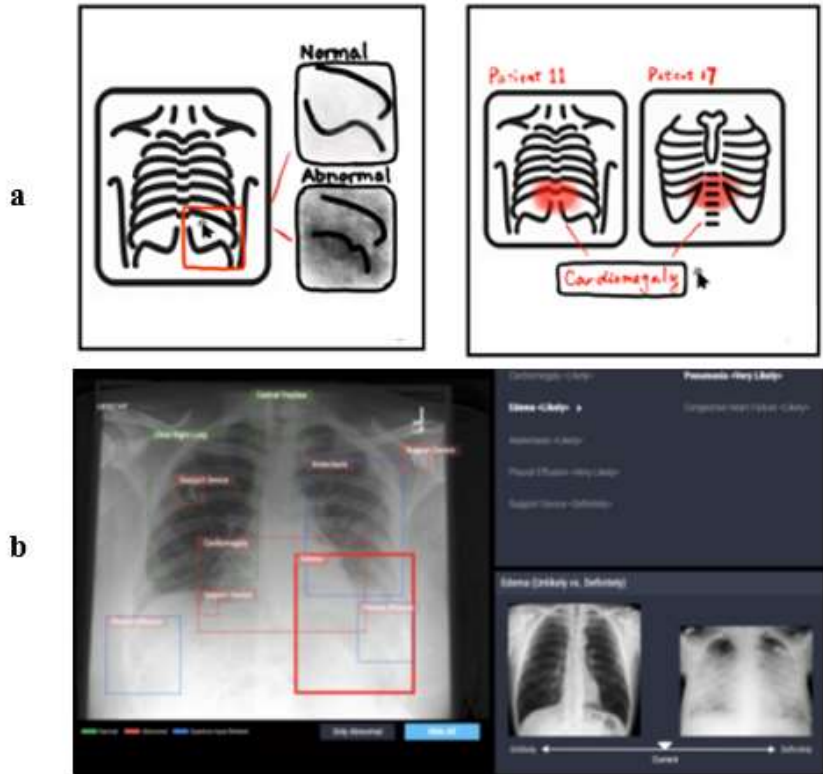
## 2.3. Tıpta Güvenilir Yapay Zeka

---

### Doktorların XAI'dan beklentilerine dair bir anket (Tonekaboni vd., 2019)

- Klinik karar verme konusunda model sonucunu doğrulayabilmek
- Uygun görsel bilgiler, karar verme adımlarının kısa ve öz olarak sunulması
- Performans ve tahmin skorları kafa karıştırıcı

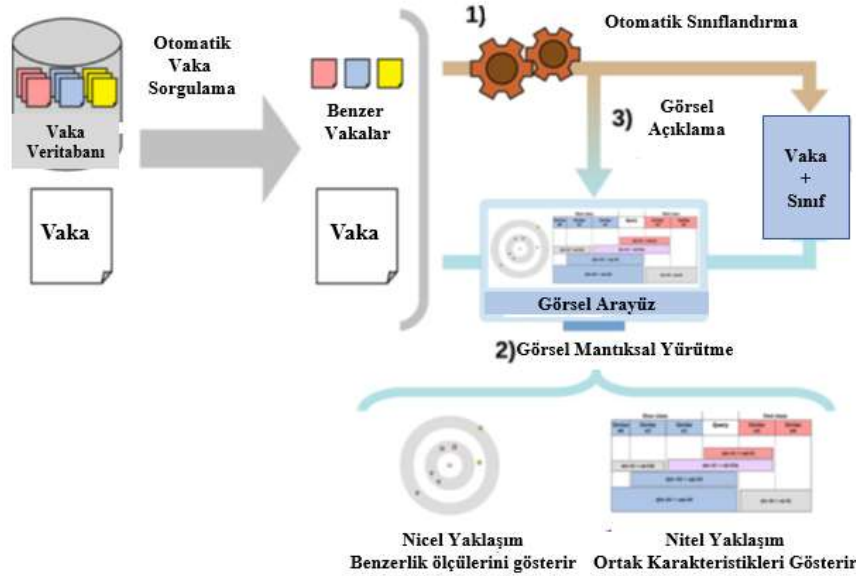
## 2.3. Tıpta Güvenilir Yapay Zeka



- Göğüs röntgen görüntüleri ile teşhis modellerinde doktorların alan bilgisine dayalı olarak son kararı verebilmeleri için açıklanabilir bir arayüz
- Görüntünün seçilen bölgesine dair, modelin tahmin ettiği ile aynı ve farklı sınıflardaki hastaların görüntülerinin karşılaştırılması modelin açıklaması



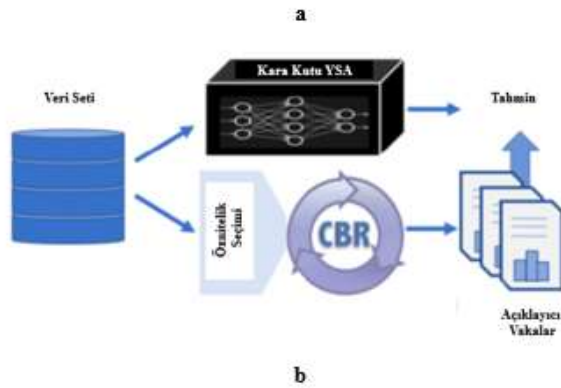
## 2.3. Tıpta Güvenilir Yapay Zeka



Açıklanabilir YZ olarak bir CBR sistemi

- Vaka veri tabanından benzerlik bazlı sorgulama
- Nicel : örneklerin benzerliğinin olarak polar çok boyutlu ölçeklendirme (MDS)
- Nitel:vaka ve benzerlerinin önemli karakteristiklerinin sunumu

## 2.3. Tıpta Güvenilir Yapay Zeka

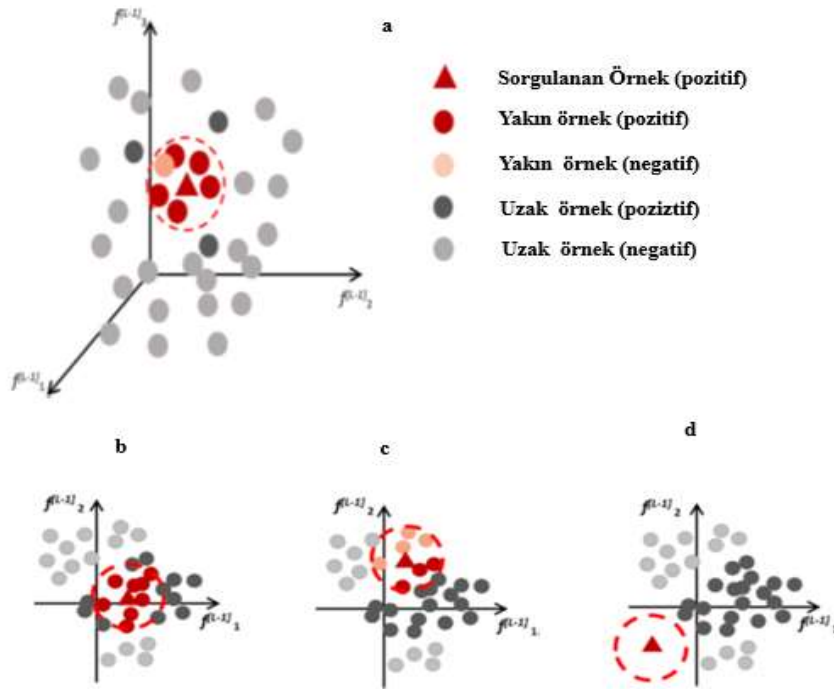


	Yaş	E119	M255	N390	I10X	J449	M542	E109	R104	R51X	I839	R520	H400	M791	M545	CKD
Örnek	2727	67	0.0	1.0	1.0	0.0	0.0	0.0	3.0	7.0	0.0	1.0	6.0	0.0	4.0	1
Komşu 1	6837	66	0.0	0.0	4.0	0.0	0.0	0.0	0.0	3.0	0.0	0.0	6.0	0.0	2.0	1
Komşu 2	24540	70	0.0	1.0	0.0	0.0	0.0	0.0	1.0	2.0	1.0	0.0	6.0	0.0	4.0	1
Komşu 3	5855	69	0.0	1.0	6.0	13.0	0.0	0.0	1.0	5.0	0.0	0.0	7.0	1.0	0.0	1

NN modeli

- Öznitelik setinin anlamlı ve ayırıcı olanlarının seçilmesi,
- Bu öznitelikler bazında CBR ile açıklama

## 2.3. Tıpta Güvenilir Yapay Zeka



Belli bir örnek için model tahminin doğrulanabilmesi için

- Aynı sınıftaki diğer örnekler ile yakınlığı durumunda model tahmini yüksek güvenle onaylanabilecek
- Farklı sınıftaki diğer örnekler ile yakınlığı durumunda model tahminin güvenilirliği sorgulanabilecek
- Yakın bölgede örneklerin olmamaları ya da az olmaları durumunda model tahminin güvenle doğrulanması mümkün olmayacaktır



## 2.3. Tıpta Güvenilir Yapay Zeka

---

Literatürde tıp alanında yapılmış, açıklanabilirlik çalışmalarının değerlendirilmesinde, tablosal veri ile yapılmış pek çok açıklanabilirlik çalışmaları olmakla birlikte, açıklamaların değerlendirilmesi aşamasının söz konusu çalışmalarda yeterince ele alınmamış olduğu ifade edilmiştir (Di Martino & Delmastro, 2022) !!!



***THANK YOU***

Çağrısan Mah. 2029 Sk. No:2 16265 Mudanya/BURSA

✉ [bilgi@mudanya.edu.tr](mailto:bilgi@mudanya.edu.tr) ☎ +90(224) 224 2022 🌐 [www.mudanya.edu.tr](http://www.mudanya.edu.tr)

📺 [in](#) [f](#) Mudanya Üniversitesi 📷 [mudanyauniversity](#) 🐦 [@mudanyaedu](#)