Açıklanabilir Yapay Zekâ ve Yapay Zekâ Etiği

Yapay Zekâ ve Veri Dersleri Rehberi · Kısım II: Yapay Zekânın Dalları

Neden açıklama? | Açıklama yöntemleri | Yanlılık ve adalet | Sorumluluk | Düzenlemeler | Komşu dersler

Sercan KÜLCÜ · Giresun Üniversitesi · Bilgisayar Mühendisliği

© Sercan Külcü. Tüm hakları saklıdır. · Sunumun etkileşimli HTML sürümü:
https://sercankulcu.github.io/files/ai_courses/sunumlar/Bolum_15_Aciklanabilir_YZ_ve_Etik.html

GİRİŞ

Ders Hangi Soruyu Sorar?

> **Bir yapay zekâ sisteminin kararına nasıl güvenilir ve sonucundan kim sorumludur?**

Bu ders, bir yapay zekâ sisteminin kararını insanların anlayabileceği biçimde **açıklamayı** ve sistemin adil, güvenilir ve hesap verebilir olmasını inceler. Teknik soru "model neden bu kararı verdi?", etik soru "bu karar kime zarar verebilir ve sorumlusu kim?" sorusudur.

Bir bakışta

YAKLAŞIM insan merkezli değerlendirme	**GİRDİ** model, verisi ve kullanım bağlamı	**ÇIKTI** açıklama, denetim raporu, ilke
BAŞARI ÖLÇÜTÜ anlaşılırlık, adalet, hesap verebilirlik	**MATEMATİK YÜKÜ** düşük-orta	**VERİ İHTİYACI** modelin kendi verisi

GİRİŞ

Kısa Tarihçe

2016
ProPublica, COMPAS suç riski puanının siyahi sanıklar aleyhine yanlı olduğunu öne süren analizi yayımlar.

2016
Ribeiro, Singh ve Guestrin LIME yöntemini tanıtır: tek bir tahmini yerel, basit bir modelle açıklamak.

2017
Lundberg ve Lee SHAP yöntemini yayımlar: oyun kuramındaki Shapley değerleriyle öznitelik katkısı.

2018
AB Genel Veri Koruma Tüzüğü (GDPR) yürürlüğe girer; otomatik kararlara itiraz hakkı tanınır.

2018
Gender Shades çalışması, ticari yüz analizi sistemlerinin koyu tenli kadınlarda çok daha sık yanıldığını gösterir.

2024
AB Yapay Zekâ Yasası (AI Act) kabul edilir: sistemler risk düzeyine göre sınıflandırılır.

DERS

Temel Kavramlar

Kara kutu model
İç işleyişi insan tarafından izlenemeyen model.

Yorumlanabilirlik
Doğası gereği okunabilen model: karar ağacı, doğrusal.

Yerel açıklama
Tek bir tahmini açıklamak: LIME, SHAP değerleri.

Genel açıklama
Modelin bütün davranışı: öznitelik önemi, PDP.

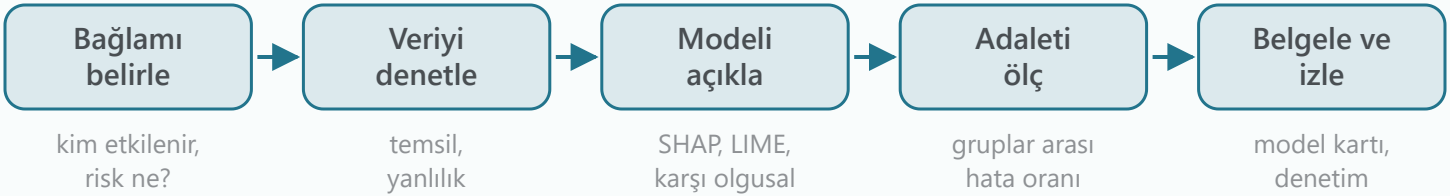
Karşı olgusal
"Gelirin şu kadar olsaydı kredi onaylanırdı."

Veri yanlılığı
Eğitim verisindeki dengesizliğin karara taşınması.

Adalet ölçütleri
Demografik eşitlik, eşit fırsat, kalibrasyon.

Hesap verebilirlik
Kararın denetlenmesi, kaydı ve sorumlusu.

Tipik İş Akışı



Örnek: Bir bankanın kredi modeli başvuru reddeder. SHAP değerleri reddin en çok borç oranından kaynaklandığını gösterir; gruplar arası karşılaştırma bazı semtlerde ret oranının belirgin yüksek olduğunu ortaya çıkarır. Ekip semti dolaylı yoldan temsil eden özniteliği çıkarır ve kararı model kartında belgeler.

DERS

Derste Neler İşlenir?

1. Yorumlanabilir modeller: doğrusal modeller, ağaçlar
2. Öznitelik önemi ve kısmi bağımlılık (PDP, ICE)
3. Yerel açıklamalar: LIME
4. Shapley değerleri ve SHAP
5. Derin ağlarda belirginlik haritaları
6. Karşı olgusal açıklamalar ve başvuru yolu
7. Açıklamaları değerlendirmek: sadakat ve kararlılık
8. Adalet ölçütleri ve aralarındaki çelişkiler
9. Yanlılığı azaltmak: ön, iç ve son işleme
10. Mahremiyet: anonimleştirme, diferansiyel mahremiyet
11. Hesap verebilirlik: model kartı, veri sayfası, denetim
12. Düzenlemeler: GDPR, KVKK, AB Yapay Zekâ Yasası

Kazandırdığı beceriler

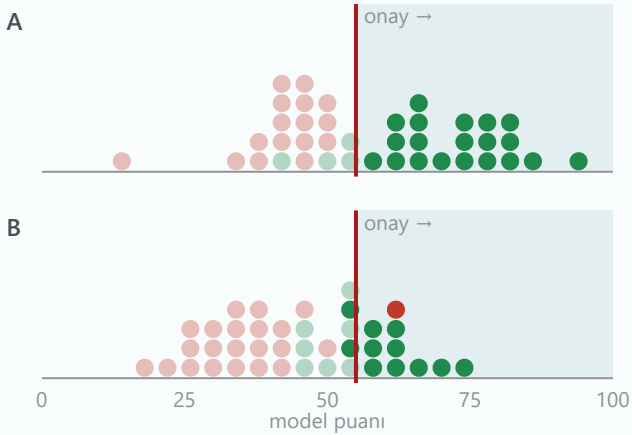
SHAP ve LIME ile açıklama üretmek | Fairlearn ile grup adaletini ölçmek | karşı olgusal açıklama üretmek | model kartı yazmak | bir sistemin risklerini değerlendirmek

Tipik bir dönemin akışıdır; sıra ve kapsam üniversiteden üniversiteye değişir.

DENE

Kredi Modeli İki Gruba Adil mi?

Model her başvuruya 0–100 arası puan veriyor; eşiğin üstü onaylanır. Puanlar B grubunda tarihsel veriden dolayı düşük. Eşikleri kaydırın, iki adalet ölçütünün nasıl değiştiğine bakın.



Ortak eşik | Gruba özel eşik

A grubu eşiği = **55**

B grubu eşiği = **55**

Onay oranlarını eşitle | Fırsatı eşitle

	onay	ödeyene onay (GPO)
A grubu	%50	%83
B grubu	%30	%61

Demografik eşitlik farkı: **20 puan**
Fırsat eşitliği farkı: **22 puan**
Doğruluk: %85

Aynı eşik iki gruba farklı davranıyor; birini sıfırlayınca öteki fark açılıyor. Ders, bu ölçümleri yapmayı, hangi adalet tanımının seçildiğini açıkça yazmayı ve modeli denetlenebilir kılmayı öğretir.

KONUM

Ön Koşullar

Bu dersten önce bilinmesi yararlı olan dersler ve neden gerektikleri.

Ders	Neden gerekir
Programlama	Python ile model eğitme ve sonuçları çözümleme
Algoritmik Oyun Kuramı	Shapley değeri: SHAP açıklamalarının kuramsal kökeni
Felsefe ya da hukuk	etik kuramlar, kişisel veri hakları ve sorumluluk kavramı

KONUM

Bu Ders Sana Göre mi?

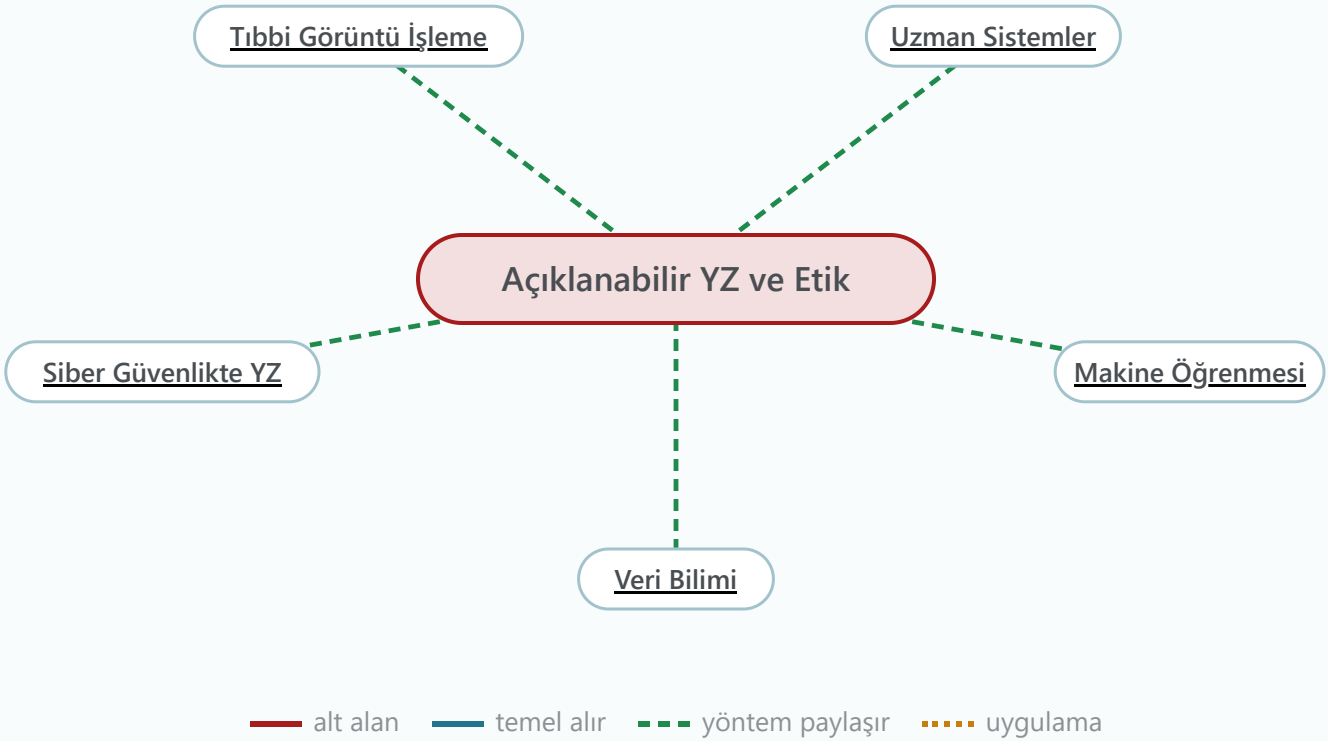
Kendine sor; "hayır" dediklerin için sağdaki sütun nereden başlayacağını gösterir.

Kendine sor	Hayırsa önce
Python ile bir sınıflandırıcı eğitip test kümesinde değerlendirebiliyor musun?	Makine Öğrenmesi
Yanlış pozitif ve yanlış negatif oranlarını hesaplayabiliyor musun?	Karışıklık matrisi ve sınıflandırma metrikleri üzerine bir okuma
Koşullu olasılığı ve taban oranı açıklayabiliyor musun?	Olasılık ve İstatistik
Shapley değerinin oyun kuramındaki anlamını açıklayabiliyor musun?	Algoritmik Oyun Kuramı dersinin Shapley değeri konusu

Çoğuna "evet" diyebiliyorsan derse hazırsın; bir ya da iki "hayır" dönem içinde kapanabilir.

KONUM

Dersin Komşuları



ÖRTÜŞME

Örtüştüğü Dersler

İlişki	Ders	Ortak konu
yöntem paylaşır	Makine Öğrenmesi ile yöntem paylaşır	model açıklama, öznitelik önemi
yöntem paylaşır	Uzman Sistemler ve Bilgi Temsili ile yöntem paylaşır	okunabilir kurallarla açıklanabilir karar
yöntem paylaşır	Tıbbi Görüntü İşleme ile yöntem paylaşır	tanıda açıklanabilirlik ve sorumluluk
yöntem paylaşır	Veri Bilimi ile yöntem paylaşır	veri yanlılığı, adalet ölçütleri
yöntem paylaşır	Siber Güvenlikte Yapay Zekâ ile yöntem paylaşır	güvenilirlik, düşmanca örnekler

AYRIŞMA

Ayrıştığı Noktalar (1/2)

Ölçüt	Açıklanabilir YZ ve Etik	Makine Öğrenmesi	Veri Bilimi
Temel soru	Bir yapay zekâ sisteminin kararına nasıl güvenilir ve sonucundan kim sorumludur?	Bir program örneklerden nasıl genelleme yapar?	Veriden doğru soruları sorup karara dönüşen sonuçlar nasıl çıkarılır?
Yaklaşım	insan merkezli değerlendirme	istatistiksel öğrenme	soru, veri, analiz, anlatı
Girdi	model, verisi ve kullanım bağlamı	etiketli ya da etiketsiz örnekler	ham, dağınık veri
Çıktı	açıklama, denetim raporu, ilke	tahmin yapan model	rapor, görselleştirme, model
Başarı ölçütü	anlaşılırlık, adalet, hesap verebilirlik	görülmemiş veride hata	kararlara katkı, doğruluk
Matematik yükü	düşük-orta	yüksek: doğrusal cebir, olasılık, optimizasyon	orta: istatistik
Veri ihtiyacı	modelin kendi verisi	orta ve çok	orta ve çok

- **Makine Öğrenmesi:** MÖ modelin doğru tahmin etmesine odaklanır; bu ders tahminin nedenini ve kimi nasıl etkilediğini sorgular.
- **Veri Bilimi:** Veri bilimi veriden sonuç çıkarır; bu ders o sonucun yanlı, adil ya da açıklanabilir olup olmadığını denetler.

AYRIŞMA

Ayrıştığı Noktalar (2/2)

Ölçüt	Açıklanabilir YZ ve Etik	Siber Güvenlikte YZ	Uzman Sistemler
Temel soru	Bir yapay zekâ sisteminin kararına nasıl güvenilir ve sonucundan kim sorumludur?	Yapay zekâ saldırıları nasıl yakalar ve kendisi nasıl saldırıya uğrar?	Bir uzmanın bilgisi nasıl yazıya dökülür ve çıkarım için kullanılır?
Yaklaşım	insan merkezli değerlendirme	güvenlik bilgisi ve MÖ	sembolik mantık ve kurallar
Girdi	model, verisi ve kullanım bağlamı	ağ trafiği, kayıtlar, dosyalar	olgular ve uzman kuralları
Çıktı	açıklama, denetim raporu, ilke	uyarı, sınıf, risk puanı	sonuç ve gerekçesi
Başarı ölçütü	anlaşılırlık, adalet, hesap verebilirlik	tespit oranı, yanlış alarm oranı	uzmanla uyum, açıklanabilirlik
Matematik yükü	düşük-orta	orta	düşük: önerme ve yüklem mantığı
Veri ihtiyacı	modelin kendi verisi	dengesiz ve sürekli değişen	az; bilgi tabanı

- **Siber Güvenlikte YZ:** Siber güvenlik saldırgana karşı korur; bu ders sistemin saldırı olmadan da insanlara verdiği zararı inceler.
- **Uzman Sistemler:** Uzman sistemler kurallarıyla zaten açıktır; bu ders kara kutu modelleri sonradan açıklamanın yollarını arar.

UYGULAMA

Uygulama Alanları

Kredi kararları
Ret gerekçesini başvurana açıklamak.

Tıbbi tanı
Hekime modelin neye baktığını göstermek.

İşe alım
Özgeçmiş elemede cinsiyet ve yaş yanlılığı.

Yargı ve kamu
Risk puanlarının adil olup olmadığını denetlemek.

Yüz tanıma
Gruplar arası hata farklarını ölçmek.

Mevzuata uyum
AB YZ Yasası ve KVKK için belgeleme.

UYGULAMA

Riskler ve Etik

Yanıltıcı açıklama

LIME ve SHAP yaklaşıktır; kötü niyetli bir model açıklama yöntemine masum görünüp asıl karar kuralını gizleyebilir.

Etik yıkama

Bir etik ilkeler belgesi ya da tek bir adalet puanı, gerçek denetimin yerine vitrin olarak kullanılabilir.

Yanlış adalet ölçütü

Ölçüt seçimi bir değer yargısıdır: COMPAS bir ölçüte göre adil, ötekine göre yanlı çıkabildi.

Açıklamaya aşırı güven

Makul görünen bir açıklama, kullanıcının yanlış bir karara da sorgusuz uymasına yol açabilir.

UYGULAMA

Araçlar ve Örnek Veriler

Araçlar ve kütüphaneler

SHAP | LIME | Captum | InterpretML | Fairlearn | AI Fairness 360 | What-If Tool

Tanınmış veri setleri

COMPAS (ProPublica) | Adult (gelir) | German Credit | CelebA | Pilot Parliaments Benchmark

Bu derste veri setleri çoğu zaman **bir sorunu göstermek** için kullanılır: COMPAS ve Adult, yanlılığın nasıl ölçüldüğünün klasik örnekleridir.

UYGULAMA

Sık Karıştırılanlar

Yaygın düşünce	Doğrusu
Korunan özniteliği (cinsiyet, ırk) çıkarmak yeter.	Posta kodu ya da isim gibi vekil öznitelikler aynı bilgiyi taşıyabilir; yanlılık sürer.
Algoritma sayılarla çalıştığı için tarafsızdır.	Model, geçmiş kararlardaki yanlılığı veriden öğrenir ve ölçekler.
Açıklama modelin gerçek işleyişini gösterir.	LIME ve SHAP gibi yöntemler yaklaşıktır; açıklamanın kendisi de yanıltıcı olabilir.
Bütün adalet ölçütleri birlikte sağlanabilir.	Gruplar arasında taban oranlar farklıysa kalibrasyon ve eşit hata oranları çoğu zaman birlikte sağlanamaz.

SONRASI

Bu Dersten Sonra

Bu dersi temel alan ya da genişleten dersler

Bu ders çoğunlukla uç noktadır; öteki dersleri temel alır.

İlgili meslekler

Sorumlu YZ uzmanı | Model denetçisi | YZ politika ve uyum uzmanı | Veri bilimci | Araştırmacı

Bu ders bütün uygulama derslerine bir mercek ekler: tıbbi görüntü işlemede tanının açıklanması, siber güvenlikte modelin aldatılması, MLOps'ta modelin izlenmesi aynı sorulara dayanır. Teknik tarafta nedensel çıkarım ve düşmanca dayanıklılık, toplumsal tarafta hukuk ve politika ile devam edilebilir.

Kaynaklar

- Molnar, *Interpretable Machine Learning*, 2. baskı, kendi yayını (çevrim içi), 2022.
- Barocas, Hardt ve Narayanan, *Fairness and Machine Learning: Limitations and Opportunities*, MIT Press, 2023.
- O'Neil, *Weapons of Math Destruction*, Crown, 2016.
- Ribeiro, Singh ve Guestrin, "Why Should I Trust You?: Explaining the Predictions of Any Classifier", *KDD*, 2016.
- Lundberg ve Lee, "A Unified Approach to Interpreting Model Predictions", *NeurIPS*, 2017.

Son

Bölüm 15: Açıklanabilir Yapay Zekâ ve Yapay Zekâ Etiği

← Önceki bölüm · Alanın haritası · Sonraki bölüm →