23 Aralık 2025

# Açıklanabilir Yapay Zeka

## 11. Hafta: Adillik (*Fairness*)

**Mustafa Cavus, Ph.D.**
🏢 Eskişehir Teknik Üniversitesi - İstatistik Bölümü
📧 mustafacavus@eskisehir.edu.tr
🌐 linktr.ee/mustafacavus

# Giriş

- Modeller veriden öğrenir; veri geçmişteki toplumsal eşitsizlikleri içeriyorsa, model bu önyargıları yansıtabilir.
- Bu nedenle, bir model genel olarak iyi performansa sahip olsa da belirli gruplar için sistematik olarak düşük performans gösterir.
- Bu sorun işe alım, yüz tanıma sistemleri ve hukuki değerlendirmelerde ortaya çıkabilir.

# Örnek Olay

ABD’de birçok eyalette kullanılan **COMPAS** adlı bir sistem, sanıkların “yeniden suç işleme riski”ni tahmin eder. Bu skor, Kefalet, Denetimli serbestlik, Ceza süresinin tespiti gibi kararlarda kullanılır.

**ProPublica (2016)** analizine göre:

- Siyah sanıklar, **yeniden suç işlemeyecekleri hâlde** yüksek riskli etiketleniyor.
- Beyaz sanıklar, **yeniden suç işleyecekleri hâlde** düşük riskli etiketleniyor.

Masum biri, ten rengine bağlı olarak farklı muamele görebilir!

# Örnek Olay

- Büyük bir teknoloji şirketi, binlerce CV'yi otomatik elemek için ML tabanlı bir sistem geliştiriyor.
- Eğitim verisi, geçmişte işe alınan çalışanlardan (büyük çoğunluğu **erkek)** oluşuyor. Model, kadın adayları sistematik olarak eliyor. "Women's chess club", "female leadership" gibi ifadeler **negatif sinyal** oluyor.
- Kadın adaylar **mülakat aşamasına bile ulaşamıyor**. Ayrımcılık: Sessiz, Ölçeklenebilir ve İnsan fark etmeden gerçekleşiyor.
- İnsan hatası tekilken, algoritmik hata binlerce kişiyi etkileyebilir.

# Örnek Olay

- Ticari yüz tanıma sistemleri, çeşitli güvenlik sistemleri ve özellikle kullanıcı girişi gereken dijital sistemlerde kimlik doğrulama amacıyla kullanılmaktadır.
- Sistemlerin hata oranları, açık tenli erkekler için **%1'den az,** koyu tenli kadınlar için **%30+** olarak ortaya çıkıyor.
- Sonuç olarak aynı sistem, aynı ortamda farklı gruplar için çok farklı doğruluk oranlarıyla çalışıyor.
- Eğitim verilerinin açık tenli erkek ağırlıklı olması nedeniyle model "az gördüğünü" grupları iyi tanımıyor.

# Giriş

Karar destek sistemleri geçmiş verilerdeki bazı yanlılıkları öğrenerek ayrımcılığa neden olabilir. **Sorumlu Makine Öğrenmesi,** modellerin sadece performans açısından değil, aynı zamanda adillik (ya da potansiyel ayrımcılık) açısından da doğrulanması gerekir.

# Adillik Ölçütleri

# Adillik Ölçütleri

Modellerin grup adilliğini ölçmek için kullanılan temel metrikler:

1. **İstatistiksel Eşitlik (*Statistical Parity*):** Gruplar arasında pozitif etiket atanma oranlarının eşitliği.
2. **Eşit Fırsat (*Equal Opportunity*):** Gruplar arasında "Gerçek Pozitif Oranı" (TPR) eşitliği.
3. **Tahmin Eşitliği (*Predictive Equality*):** Gruplar arasında "Yanlış Pozitif Oranı" (FPR) eşitliği.
4. **Doğruluk Eşitliği (*Accuracy Equality*):** Gruplar arasında toplam model doğruluğunun eşitliği.

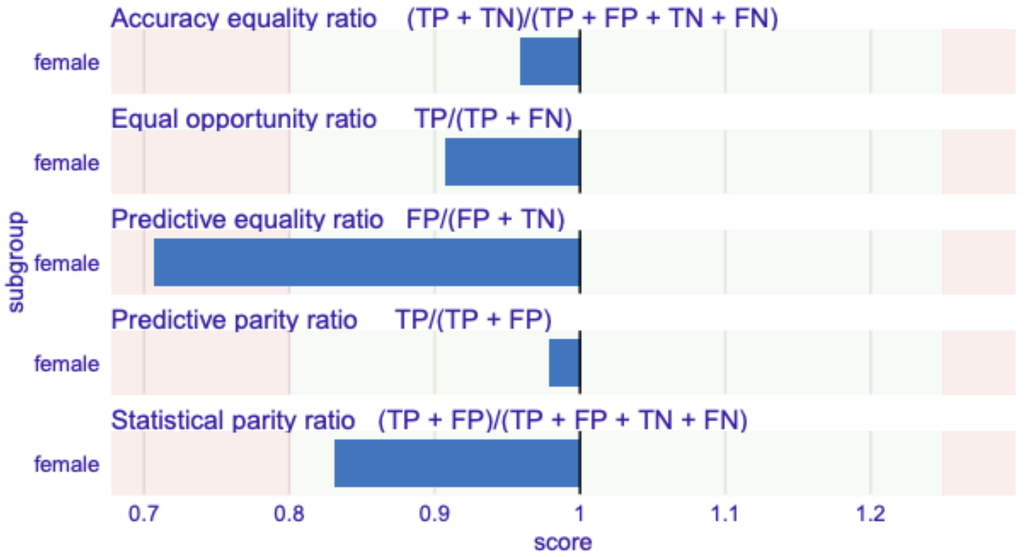
| $A = a$ | $Y = 1$ | $Y = 0$ | |
|---|---|---|---|
| $\hat{Y} = 1$ | $TP_a = P(Y = 1, \hat{Y} = 1 \mid A = a)$ | $FP_a = P(Y = 0, \hat{Y} = 1 \mid A = a)$ | $P(\hat{Y} = 1 \mid A = a)$ |
| $\hat{Y} = 0$ | $FN_a = P(Y = 1, \hat{Y} = 0 \mid A = a)$ | $TN_a = P(Y = 0, \hat{Y} = 0 \mid A = a)$ | $P(\hat{Y} = 0 \mid A = a)$ |
| | $P(Y = 1 \mid A = a)$ | $P(Y = 0 \mid A = a)$ | |

# 4/5 Kuralı

- Bir modelin, alt gruplar arasındaki ilişkileri tamamen aynı tutması oldukça zordur.
- Bu nedenle, "mükemmel uyum" çevresinde bazı tolerans marjlarına ihtiyaç duyulur.
- Bu sorunu ele almak için, ayrımcılık oranı kıstası olarak **dörtte beş (4/5) kuralı** (Federal Düzenlemeler Kanunu, 1978) kabul edilmiştir.
- Bu kurala göre: “Herhangi bir ırk, cinsiyet veya etnik grup için belirlenen seçim oranının, en yüksek orana sahip grubun oranının dörtte beşinden (%80) az olması, genel olarak federal icra kurumları tarafından **olumsuz etki** kanıtı olarak değerlendirilecektir”.

# Adillik Ölçütleri

- **Adillik kontrolü** grafiği, imtiyazsız ve imtiyazlı alt gruplar arasındaki adillik ölçümlerinin oranını özetler.
- Grafikteki açık yeşil alanlar, adillik ölçütlerinde kabul edilebilir fark alanlarını gösterir.
- **4/5 kuralına** uymayan değerleri gösteren kırmızı dikdörtgenlerle sınırlandırılmıştır.
- Buradaki oran; ölçüt değerleri hesaplandıktan sonra imtiyazsız gruba (kadın) ait değerlerin, imtiyazlı alt gruba (erkek) ait değerlere bölünerek hesaplanmıştır.
- Bu örnekte, **tahmin eşitliği oranı (*predictive equality ratio*)** hariç diğer tüm ölçütler ayrımcı olmayan düzeydedir.

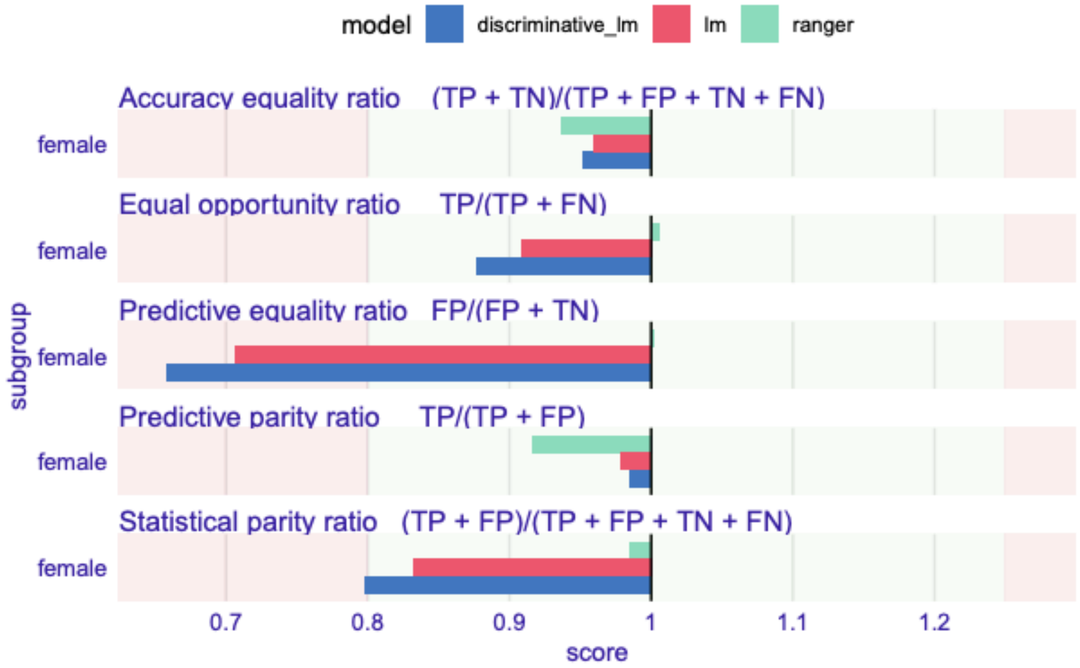


# Uygulama

# Uygulama

`germancredit` veriseti üzerinde müşterilerin kredi riskini tahmin eden bir modelde cinsiyet ayrımcılığının incelenmesi:

- **discriminative_lm** ve **lm** modelleri bu metrikte 0.8 sınırının oldukça altında kalmışlardır. Bu durum, modellerin kadınlar ve erkekler arasında "Yanlış Pozitif Oranı" açısından adil olmadıklarını gösterir.
- **Random forest** modeli, tüm metriklerde 1 değerine en yakın modeldir.
- İstatistiksel eşitlik (*Statistical parity*) açısından her üç model de 0.8 sınırının üzerinde kalarak kadın ve erkeklere benzer oranlarda pozitif kredi sonucu atamıştır.



# Çözüm Yöntemleri

# Çözüm Yöntemleri

Eğer model adil değilse uygulanabilecek üç farklı süreçte çeşitli yaklaşımlar vardır:

1. **Ön İşleme:** Verideki istenmeyen korelasyonları düzeltir (örn. *Reweighting*, *Resampling*).
2. **İşlem Sırası:** Algoritmayı eğitirken adillik ölçütlerine göre optimize eder.
3. **Son İşleme:** Model çıktılarını gruplar arasında daha adil olacak şekilde değiştirir (örn. *Cutoff Manipulation*, *ROC Pivot*).

# Kaynaklar

Bu materyalin hazırlanmasında,

*Wiśniewski, J., & Biecek, P. (2022). fairmodels: a Flexible Tool for Bias Detection, Visualization, and Mitigation in Binary Classification Models. R Journal, 14(1)*

çalışmasından yararlanılmıştır.



## fairmodels: a Flexible Tool for Bias Detection, Visualization, and Mitigation in Binary Classification Models

*by Jakub Wiśniewski and Przemysław Biecek*

**Abstract** Machine learning decision systems are becoming omnipresent in our lives. From dating apps to rating loan seekers, algorithms affect both our well-being and future. Typically, however, these systems are not infallible. Moreover, complex predictive models are eager to learn social biases present in historical data that may increase discrimination. If we want to create models responsibly, we need tools for in-depth validation of models also from potential discrimination. This article introduces an R package **fairmodels** that helps to validate fairness and eliminate bias in binary classification models quickly and flexibly. The **fairmodels** package offers a model-agnostic approach to bias detection, visualization, and mitigation. The implemented functions and fairness metrics enable model fairness validation from different perspectives. In addition, the package includes a series of methods for bias mitigation that aim to diminish the discrimination in the model. The package is designed to examine a single model and facilitate comparisons between multiple models.

### 1 Introduction

Responsible machine learning and, in particular, fairness is gaining attention within the machine learning community. This is because predictive algorithms are becoming more and more decisive and influential in our lives. This impact could be less or more significant in areas ranging from user feeds on social platforms, displayed ads, and recommendations at an online store to loan decisions, social scoring, and facial recognition systems used by police and authorities. Sometimes it leads to automated systems that learn some undesired bias preserved in data for some historical reason. Whether seeking a job (Lahoti et al., 2019) or having one's data processed by court systems (Angwin et al., 2016), sensitive attributes such as sex, race, religion, ethnicity, etc., might play a significant role in the decision. Even if such variables are not directly included in the model, they are often captured by proxy variables such as zip code (a proxy for the race and wealth), purchased products (a proxy for gender and age), eye colour (a proxy for ethnicity). As one would expect, they can give an unfair advantage to a privileged group. Discrimination takes the form of more favorable predictions or higher accuracy for a privileged group. For example, some popular commercial gender classifiers were found to perform the worst on darker females (Buolamwini and Gebru, 2018). From now on, such unfair and harmful decisions towards people with specific sensitive attributes will be called biased.

The list of protected attributes may depend on the region and domain for which the model is built. For example, the European Union law is summarized in the Handbook on European non-discrimination law European Union Agency for Fundamental Rights and Council of Europe (2018), which lists the following protected attributes that cannot be the basis for inferior treatment: sex, gender identity, sexual orientation, disability, age, race, ethnicity, nationality or national origin, religion or belief, social origin, birth, and property, language, political or other opinions. This list, though long, does not include all potentially relevant items, e.g. in the USA, a protected attribute is also pregnancy, the status of a war veteran, or genetic information.

While there are historical and economic reasons for this to happen, such decisions are unacceptable in society, where nobody should have an unfair advantage. The problem is not simple, especially when the only criterion set for the system is performance. We observe a trade-off between accuracy and fairness in some cases where lower discrimination leads to lower performance (Kamiran and Calders, 2011). Sometimes labels, which are considered ground truth, might also be biased (Wick et al., 2019), and when controlling for that bias, the performance and fairness might improve simultaneously. However fairness is not a concept that a single number can summarize, so most of the time, when we want to improve fairness from one perspective, it becomes worse in another (Barocas et al., 2019).

The bias in machine learning systems has potentially many different sources. Mehrabi et al. (2019) categorized bias into its types like historical bias, where unfairness is already embedded into the data reflecting the world, observer bias, sampling bias, ranking and social biases, and many more. That shows how many dangers are potentially hidden in the data itself. Whether one would like to act on it or not, it is essential to detect bias and make well-informed decisions whose consequences could potentially harm many groups of people. Repercussions of such systems can be unpredictable. As argued by Barocas et al. (2019), machine learning systems can even aggravate the disparities between groups, which is called by the authors' feedback loops. Sometimes the risk of potential harm resulting

Ders notlarına dersin **GitHub** reposu üzerinden ulaşabilirsiniz.

Ders ile ilgili sorularınız için mustafacavus@eskisehir.edu.tr adresi üzerinden benimle iletişime geçebilirsiniz.

**Mustafa Cavus, Ph.D.**
🏢 Eskişehir Teknik Üniversitesi - İstatistik Bölümü
📧 mustafacavus@eskisehir.edu.tr
🌐 linktr.ee/mustafacavus