Introduction[¶](https://www.kaggle.com/code/alexisbcook/ai-fairness" \l "Introduction" \t "_self)

Adil bir makine öğrenimi (ML) modelinde ne arayacağımızı tanımlamanın birçok farklı yolu vardır. Örneğin, kredi kartı başvurularını onaylayan (veya reddeden) bir model üzerinde çalıştığımızı varsayalım. Sizce bu durum:

* Cinsiyetler arasında onay oranının eşit olmasıyla mı adil olur, yoksa
* Cinsiyet verisinin veri kümesinden çıkarılıp modelden gizlenmesiyle mi daha iyi olur?

Bu eğitimde, bir modelin adil olup olmadığına karar vermek için kullanabileceğimiz dört kriteri ele alacağız. Ardından, öğrendiklerinizi uygulayacağınız **uygulamalı bir alıştırmada**, birkaç model eğitmek ve her birinin adilliğini analiz etmek için kod çalıştıracaksınız. (Bu kurstaki her derste olduğu gibi, daha önce kodlama deneyimi gerekli değildir.)

# Four fairness criteria[¶](https://www.kaggle.com/code/alexisbcook/ai-fairness" \l "Four-fairness-criteria" \t "_self)

Bu dört adillik kriteri faydalı bir başlangıç noktasıdır, ancak adilliği resmileştirmenin daha fazla yolu olduğunu ve bunları keşfetmenizin teşvik edildiğini belirtmek önemlidir.

Bir sonucu almak üzere bireyleri seçen bir model üzerinde çalıştığımızı varsayalım. Örneğin, model bir krediye onaylanması, bir üniversiteye kabul edilmesi veya bir iş fırsatı sunulması gereken insanları seçebilir. (Bu nedenle, yüz tanıma veya metin oluşturma gibi görevleri yerine getiren modelleri ele almıyoruz.)

## **1. Demographic parity / statistical parity[¶](https://www.kaggle.com/code/alexisbcook/ai-fairness" \l "1.-Demographic-parity-/-statistical-parity" \t "_self)**

**Demografik eşitlik**, model tarafından seçilen kişilerin kompozisyonu, başvuranların grup üyeliği yüzdeleriyle eşleşiyorsa, modelin adil olduğunu söyler.

Uluslararası bir konferans düzenleyen kâr amacı gütmeyen bir kuruluş, 20.000 kişinin katılım için kaydolduğunu görüyor. Organizatörler, konferansta potansiyel olarak ilgi çekici konuşmalar yapabilecek 100 katılımcıyı seçmek için bir makine öğrenimi modeli yazıyor. Katılımcıların %50'si kadın olduğu için (20.000'de 10.000), seçilen konuşmacı adaylarının da %50'sinin kadın olacağı şekilde modeli tasarlıyorlar.

## **2. Equal opportunity[¶](https://www.kaggle.com/code/alexisbcook/ai-fairness" \l "2.-Equal-opportunity" \t "_self)**

**Fırsat eşitliği** adilliği, model tarafından seçilmesi gereken ("pozitifler") ve model tarafından doğru bir şekilde seçilen kişilerin oranının her grup için aynı olmasını sağlar. Bu oranı **doğru pozitif oranı** (true positive rate - TPR) veya modelin **duyarlılığı** (sensitivity) olarak adlandırırız.

Bir doktor, ciddi tıbbi rahatsızlıklar geliştirme riski taşıyabilecek, ekstra bakıma ihtiyacı olan hastaları belirlemek için bir araç kullanır. (Bu araç, yalnızca doktorun uygulamasına destek olmak, ikinci bir görüş olarak kullanılmaktadır.) Her demografik grup için eşit ve yüksek bir TPR'ye sahip olacak şekilde tasarlanmıştır.

## **3. Equal accuracy[¶](https://www.kaggle.com/code/alexisbcook/ai-fairness" \l "3.-Equal-accuracy" \t "_self)**

Alternatif olarak, modelin her grup için **eşit doğruluğa** sahip olduğunu kontrol edebiliriz. Yani, doğru sınıflandırma yüzdesi (reddedilmesi gereken ve reddedilen kişiler ile onaylanması gereken ve onaylanan kişiler) her grup için aynı olmalıdır. Model bir gruptaki bireyler için %98 doğruysa, diğer gruplar için de %98 doğru olmalıdır.

Bir banka, insanlara kredi onayı vermek için bir model kullanır. Model, her demografik grup için eşit derecede doğru olacak şekilde tasarlanmıştır: bu şekilde, banka hem reddedilmesi gereken kişileri onaylamaktan (bu durum hem başvuru sahibi hem de banka için finansal olarak zararlı olur) hem de onaylanması gereken kişileri reddetmekten (bu durum başvuru sahibi için kaçırılmış bir fırsat olur ve bankanın gelirini azaltır) kaçınır.

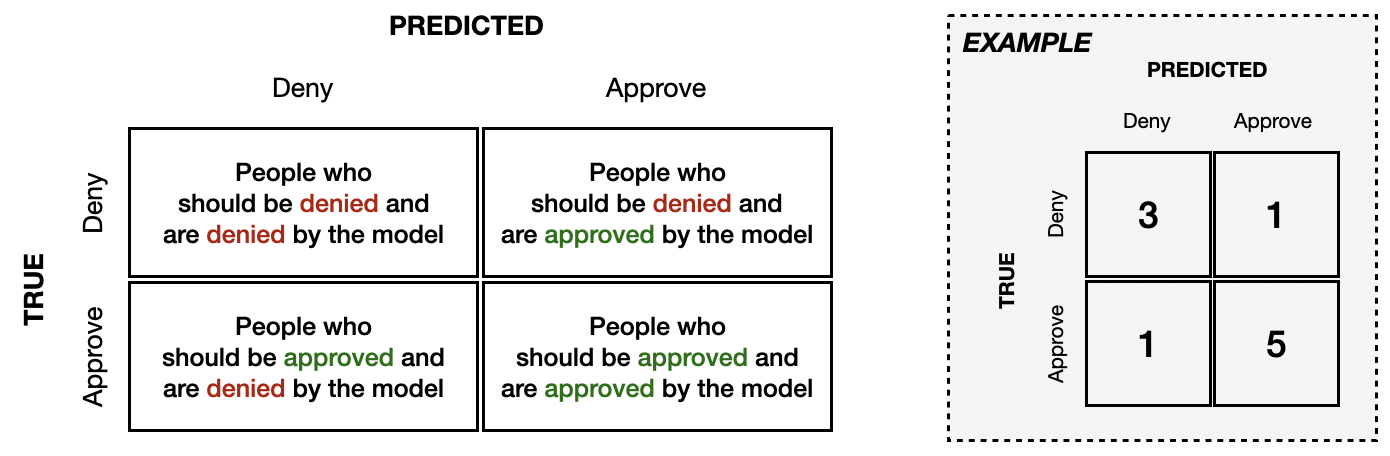
## **4. Group unaware / "Fairness through unawareness"**[**¶**](https://www.kaggle.com/code/alexisbcook/ai-fairness#4.-Group-unaware-/-%22Fairness-through-unawareness%22)

**Grup bilgisinden bağımsız adillik**, tüm grup üyeliği bilgilerini veri kümesinden kaldırır. Örneğin, farklı cinsiyet grupları için modeli adil hale getirmeye çalışmak amacıyla cinsiyet verilerini kaldırabiliriz. Benzer şekilde, ırk veya yaş hakkındaki bilgileri de kaldırabiliriz.

Bu yaklaşımı pratikte uygulamanın zorluklarından biri, grup üyeliği verileri için vekil (proxy) olabilecek değişkenleri dikkatlice belirleyip kaldırmaktır. Örneğin, ırksal olarak ayrılmış şehirlerde, posta kodu ırk için güçlü bir vekildir. Yani, ırk verileri kaldırıldığında, posta kodu verileri de kaldırılmalıdır, aksi takdirde makine öğrenimi uygulaması hala bir bireyin ırkını verilerden çıkarabilir. Ayrıca, grup bilgisinden bağımsız adilliğin tarihsel yanlılık için iyi bir çözüm olması olası değildir.

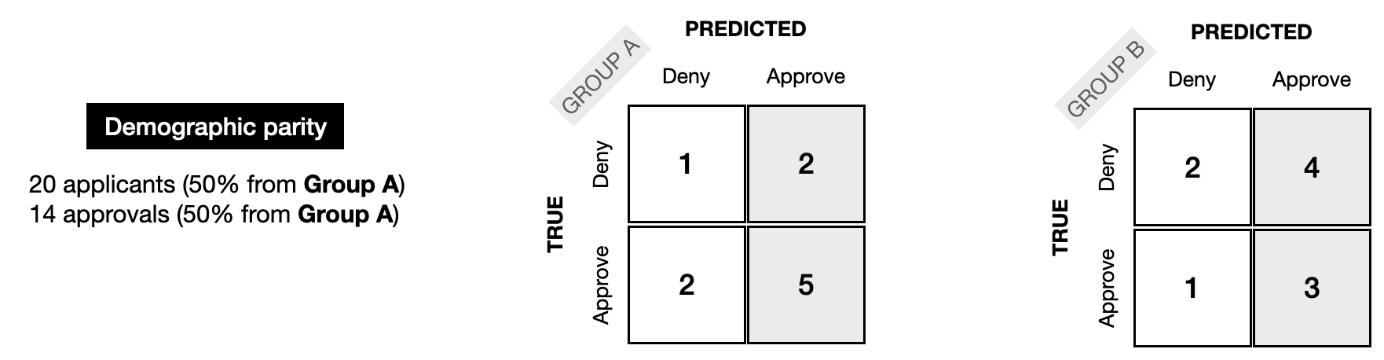
# Example[¶](https://www.kaggle.com/code/alexisbcook/ai-fairness" \l "Example" \t "_self)

Dört farklı adalet türü arasındaki farkları göstermek için küçük bir örnek üzerinde çalışacağız. Bir makine öğrenimi modelinin performansını anlamak için yaygın olarak kullanılan bir araç olan karışıklık matrisini kullanacağız. Bu araç, %80 doğruluk oranına (10 kişiden 8'i doğru sınıflandırıldığı için) ve %83 gerçek pozitif oranına (6 kişiden 5'i doğru sınıflandırıldığı için) sahip bir modeli gösteren aşağıdaki örnekte gösterilmiştir.

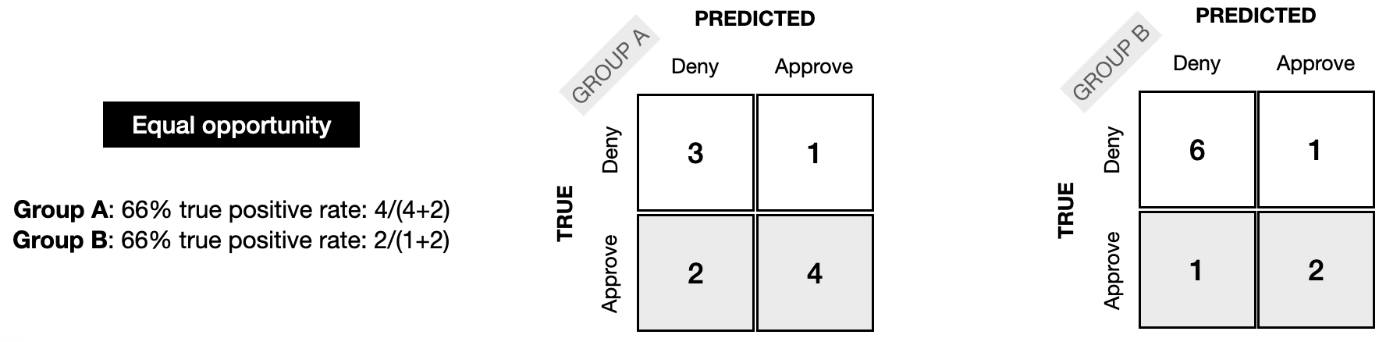


Bir modelin performansının gruplar arasında nasıl değiştiğini anlamak için, her grup için farklı bir karışıklık matrisi oluşturabiliriz. Bu küçük örnekte, yalnızca 20 kişiden, iki grup arasında eşit olarak bölünmüş (Grup A'dan 10 ve Grup B'den 10) veriye sahip olduğumuzu varsayacağız.

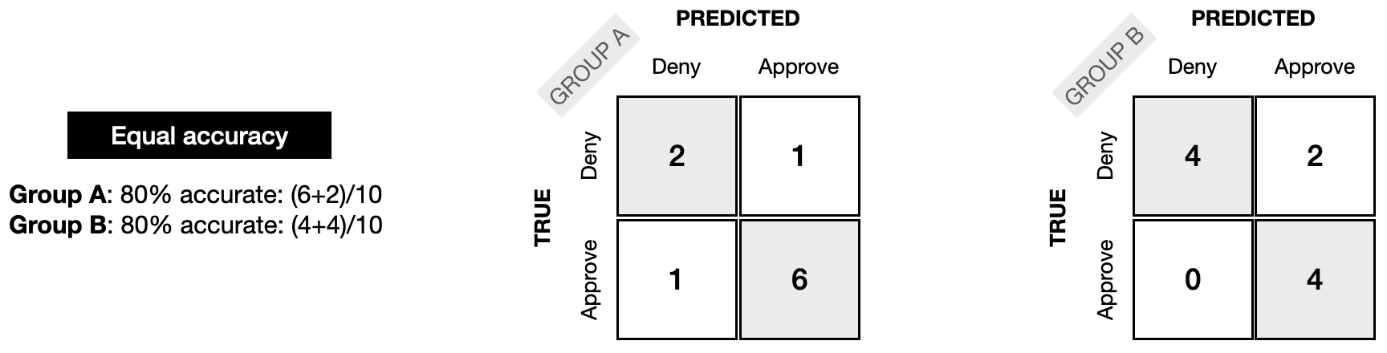
Sonraki görsel, modelin **demografik eşitlik** adilliğini sağladığı durumda karışıklık matrislerinin nasıl görünebileceğini gösteriyor. Her gruptan 10 kişi (Grup A'dan %50 ve Grup B'den %50) model tarafından değerlendirildi. Yine gruplar arasında eşit olarak bölünmüş (Grup A'dan %50 ve Grup B'den %50) 14 kişi model tarafından onaylandı.



Fırsat eşitliği adaleti için her grubun TPR'si aynı olmalıdır; aşağıdaki örnekte her durumda %66'dır.



Ardından, karışıklık matrislerinin eşit doğruluk ve adalet açısından nasıl görünebileceğini görebiliriz. Her grup için model %80 doğruluk oranına sahipti.



Grup farkında olmayan adaletin, karışıklık matrisinden tespit edilemeyeceğini ve veri kümesinden grup üyelik bilgilerinin çıkarılmasıyla daha çok ilgili olduğunu unutmayın.

Şimdi bu oyuncak örnekleri incelemek ve farklı adalet türleri arasındaki farklar için sezgilerinizi geliştirmek için zaman ayırın. A Grubu'nun B Grubu'nun iki katı başvuru sahibi olması durumunda örnek nasıl değişir?

Ayrıca, örneklerin hiçbirinin birden fazla adalet türünü karşılamadığını unutmayın. Örneğin, demografik eşitlik örneği eşit doğruluk veya eşit fırsat sağlamaz. Bunu şimdi doğrulamak için zaman ayırın. Pratikte, bir modeli birden fazla adalet türü için optimize etmek mümkün değildir: bu konuda daha fazla bilgi edinmek için Makine Adaletinin İmkansızlık Teoremi'ni inceleyin. Peki, yalnızca birini karşılayabiliyorsanız, hangi adalet kriterini seçmelisiniz? Çoğu etik soruda olduğu gibi, doğru cevap genellikle basit değildir ve bir kriter seçmek, ekibinizdeki herkesin dahil olduğu uzun bir konuşma gerektirir.

Gerçek bir projeyle çalışırken veriler çok daha büyük olacaktır. Bu durumda, karışıklık matrisleri model performansını analiz etmek için hala kullanışlı bir araçtır. Ancak dikkat edilmesi gereken önemli bir nokta, gerçek dünya modellerinin genellikle herhangi bir adalet tanımını mükemmel bir şekilde karşılamasının beklenemeyeceğidir. Örneğin, adalet ölçütü olarak "demografik eşitlik" seçilirse ve modelin hedefi %50 erkek seçmekse, nihai modelin nihayetinde %50'ye yakın, ancak tam olarak %48 veya %53 gibi bir yüzdeyi seçmesi söz konusu olabilir.

# Learn more[¶](https://www.kaggle.com/code/alexisbcook/ai-fairness" \l "Learn-more" \t "_self)

* Explore different types of fairness with an [interactive tool](http://research.google.com/bigpicture/attacking-discrimination-in-ml/).
* You can read more about equal opportunity in [this blog post](https://ai.googleblog.com/2016/10/equality-of-opportunity-in-machine.html).
* Analyze ML fairness with [this walkthrough](https://pair-code.github.io/what-if-tool/learn/tutorials/walkthrough/) of the What-If Tool, created by the People and AI Research (PAIR) team at Google. This tool allows you to quickly amend an ML model, once you've picked the fairness criterion that is best for your use case.

Bu, farklı adillik türlerini keşfetmek ve bir adillik türü için optimizasyon yaparken ortaya çıkabilecek ödünleşimi göstermek için sadece kısa bir alıştırmadır. Burada model eğitimine odaklandık, ancak pratikte, yanlılığı gerçekten azaltmak veya makine öğrenimi sistemlerini adil hale getirmek için, veri toplamadan nihai bir ürünü kullanıcılara sunmaya kadar sürecin her adımını yakından incelememiz gerekir.

Örneğin, verilere yakından bakarsanız, ortalama olarak B Grubundaki bireylerin A Grubundaki bireylerden daha yüksek gelire sahip olduğunu ve ayrıca bir eve veya arabaya sahip olma olasılıklarının daha yüksek olduğunu fark edeceksiniz. Bunu bilmek, hangi adillik kriterini kullanmanız gerektiğine karar vermeniz ve adilliği sağlamanın yollarını belirlemeniz için çok değerli olacaktır. (Örneğin, verideki tarihsel yanlılığı kaldırmadan ve ardından her grup için eşit doğruluk elde etmek üzere modeli eğitmek kötü bir yaklaşım olacaktır.)

Bu kursta, belirli projelerde yanlılığı tam olarak nasıl en aza indireceğimiz ve adilliği nasıl sağlayacağımız konusunda kasıtlı olarak fikir beyan etmekten kaçınıyoruz. Çünkü yapay zekâda adillik aktif bir araştırma alanı olduğu için doğru cevaplar sürekli olarak gelişmektedir. Bu ders, konuya uygulamalı bir giriş niteliğindeydi ve **Yapay Zekâ Ortaklığı'ndan (Partnership on AI)** blog yazılarını okuyarak veya **ACM Adillik, Hesap Verebilirlik ve Şeffaflık Konferansı (ACM FAccT)** gibi konferansları takip ederek öğreniminize devam edebilirsiniz.