Introduction

Makine öğrenimi (ML), hayatları iyileştirme potansiyeline sahiptir ancak aynı zamanda bir zarar kaynağı da olabilir. ML uygulamaları, ırk, cinsiyet, din, sosyoekonomik durum ve diğer kategorilere dayanarak bireylere karşı ayrımcılık yapmıştır.

Bu eğitimde, özellikle sonuçlar belirli grupları orantısız şekilde etkiliyorsa, ML uygulamalarının olumsuz, istenmeyen sonuçlarını ifade eden **yanlılık (bias)** hakkında bilgi edineceksiniz.

Herhangi bir ML uygulamasını etkileyebilecek altı farklı yanlılık türünü ele alacağız. Ardından, gerçek dünyadan bir senaryodaki yanlılığı tanımlayacağınız **uygulamalı bir alıştırmada** yeni bilgilerinizi kullanacaksınız.

# Bias is complex[¶](https://www.kaggle.com/code/alexisbcook/identifying-bias-in-ai" \l "Bias-is-complex" \t "_self)

Birçok makine öğrenimi (ML) uzmanı, "yanlı veriler" ve "ne ekersen onu biçersin" (garbage in, garbage out) kavramına aşinadır. Örneğin, Yahudi karşıtı çevrimiçi konuşmalar içeren bir veri kümesiyle bir sohbet robotu eğitiyorsanız ("garbage in"), sohbet robotu muhtemelen Yahudi karşıtı yorumlar yapacaktır ("garbage out"). Bu örnek, tanınması ve ele alınması gereken önemli bir yanlılık türünü (aşağıda göreceğiniz gibi **tarihsel yanlılık** olarak adlandırılır) detaylandırmaktadır.

Ancak yanlılık, ML uygulamalarını mahveden tek şey bu değildir.

Verideki yanlılık karmaşıktır. Bir grup eğitim verilerinde yetersiz temsil ediliyorsa, hatalı veriler **temsil yanlılığına** (bu eğitimde daha sonra ele alınacaktır) da yol açabilir. Örneğin, bir yüz tanıma sistemi eğitilirken, eğitim verileri çoğunlukla açık tenli bireyleri içeriyorsa, koyu tenli kullanıcılarda iyi performans gösteremez. Eğitim verilerinden kaynaklanabilecek üçüncü bir yanlılık türü ise, aşağıda öğreneceğiniz **ölçüm yanlılığıdır**.

Ve yalnızca yanlı veriler, adil olmayan ML uygulamalarına yol açmaz: öğreneceğiniz gibi, yanlılık aynı zamanda ML modelinin tanımlanma şeklinden, modelin diğer modellerle karşılaştırılma şeklinden ve günlük kullanıcıların modelin nihai sonuçlarını yorumlama şeklinden de kaynaklanabilir. ML sürecinin her aşamasından zarar gelebilir.

# Six types of bias[¶](https://www.kaggle.com/code/alexisbcook/identifying-bias-in-ai" \l "Six-types-of-bias" \t "_self)

Farklı yanlılık türlerinin farkına vardığımızda, makine öğrenimi projelerinde onları tespit etme olasılığımız artar. Dahası, ortak bir kelime dağarcığı ile yanlılığın nasıl hafifletileceği (veya azaltılacağı) hakkında verimli konuşmalar yapabiliriz.

2020 yılının başlarından, altı farklı yanlılık türünü karakterize eden bir **araştırma makalesini** yakından takip edeceğiz.

## **Historical bias[¶](https://www.kaggle.com/code/alexisbcook/identifying-bias-in-ai" \l "Historical-bias" \t "_self)**

**Tarihsel yanlılık**, verinin oluşturulduğu dünya durumunun kusurlu olduğu zaman ortaya çıkar.

2020 itibarıyla, Fortune 500 CEO'larının yalnızca %7,4'ü kadındır. Araştırmalar, kadın CEO'lara veya CFO'lara sahip şirketlerin genellikle aynı pozisyondaki erkeklerin olduğu şirketlerden daha kârlı olduğunu göstermektedir, bu da kadınların işe alımda erkeklerden daha yüksek standartlara tabi tutulduğunu düşündürmektedir. Bunu düzeltmek için, insan girdisini ortadan kaldırmayı ve işe alım sürecini daha adil hale getirmek için yapay zekâyı kullanmayı düşünebiliriz. Ancak, geçmiş işe alım kararlarından elde edilen veriler bir modeli eğitmek için kullanılırsa bu verimsiz olabilir, çünkü modelin veride mevcut olan aynı yanlılıkları (önyargıları) öğrenmesi muhtemeldir.

## **Representation bias[¶](https://www.kaggle.com/code/alexisbcook/identifying-bias-in-ai" \l "Representation-bias" \t "_self)**

**Temsil yanlılığı**, bir modeli eğitmek için veri kümeleri oluşturulurken, bu veri kümelerinin modelin hizmet edeceği insanları kötü bir şekilde temsil etmesi durumunda ortaya çıkar.

Akıllı telefon uygulamaları aracılığıyla toplanan veriler, akıllı telefon sahibi olma olasılığı daha düşük olan grupları yetersiz temsil edecektir. Örneğin, ABD'de veri toplanıyorsa, 65 yaş üstü bireyler yetersiz temsil edilecektir. Veriler, bir şehir ulaşım sisteminin tasarımına bilgi sağlamak için kullanılırsa, bu feci olur, çünkü yaşlı insanların sistemin erişilebilir olmasını sağlamak için önemli ihtiyaçları vardır.

## **Measurement bias[¶](https://www.kaggle.com/code/alexisbcook/identifying-bias-in-ai" \l "Measurement-bias" \t "_self)**

**Ölçüm yanlılığı**, verinin doğruluğunun gruplar arasında farklılık göstermesi durumunda ortaya çıkar. Bu, doğrudan ölçülemeyen bir değişkenin yerine geçen **vekil değişkenlerle** çalışırken, vekil değişkenin kalitesinin farklı gruplarda değişmesi durumunda meydana gelebilir.

Yerel hastaneniz, geçmiş teşhisler, ilaçlar ve demografik veriler gibi bilgilere dayanarak, ciddi rahatsızlıklar geliştirmeden önce yüksek riskli hastaları belirlemek için bir model kullanıyor. Model, bu bilgiyi sağlık hizmetleri maliyetlerini tahmin etmek için kullanıyor ve buradaki fikir, daha yüksek maliyetli hastaların yüksek riskli hastalara denk geldiği yönünde. Modelin özellikle ırkı dışlamasına rağmen, ırksal ayrımcılık gösterdiği görülüyor: algoritma, uygun Siyah hastaları seçme olasılığı daha düşük. Bu nasıl olabilir? Çünkü maliyet, risk için bir vekil olarak kullanılmıştır ve bu değişkenler arasındaki ilişki ırka göre değişmektedir: Siyah hastalar, sağlık hizmetlerine erişimde daha fazla engelle karşılaşmakta, sağlık sistemine daha az güven duymakta ve bu nedenle, aynı sağlık koşullarına sahip Siyah olmayan hastalara kıyasla ortalama olarak daha düşük tıbbi maliyetlere sahip olmaktadır.

## **Aggregation bias[¶](https://www.kaggle.com/code/alexisbcook/identifying-bias-in-ai" \l "Aggregation-bias" \t "_self)**

**Kümeleme (Agregasyon) yanlılığı**, grupların uygunsuz bir şekilde birleştirilmesi sonucunda, hiçbir grup için iyi performans göstermeyen veya yalnızca çoğunluk grubu için iyi performans gösteren bir modelin ortaya çıkması durumunda meydana gelir. (Bu genellikle bir sorun değildir, ancak en sık tıbbi uygulamalarda ortaya çıkar.)

Hispaniklerin, Hispanik olmayan beyazlara göre daha yüksek diyabet ve diyabete bağlı komplikasyon oranları vardır. Diyabeti teşhis veya izlemek için yapay zekâ oluşturuluyorsa, ya etnik kökeni veride bir özellik olarak dahil ederek ya da farklı etnik gruplar için ayrı modeller oluşturarak sistemi bu etnik farklılıklara karşı duyarlı hale getirmek önemlidir.

## **Evaluation bias[¶](https://www.kaggle.com/code/alexisbcook/identifying-bias-in-ai" \l "Evaluation-bias" \t "_self)**

**Değerlendirme yanlılığı**, bir model değerlendirilirken, kıyaslama verileri (modeli benzer görevleri yerine getiren diğer modellerle karşılaştırmak için kullanılan) modelin hizmet edeceği popülasyonu temsil etmediğinde ortaya çıkar.

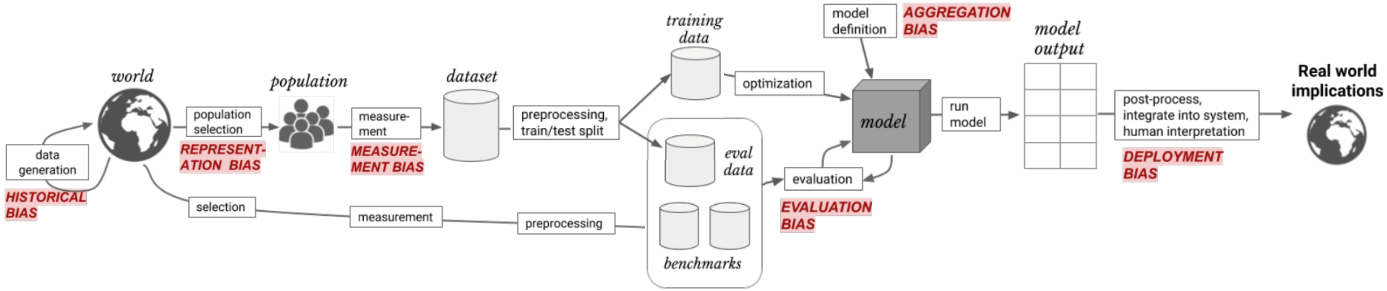
**Gender Shades** makalesi, yaygın olarak kullanılan iki yüz analizi kıyaslama veri kümesinin (IJB-A ve Adience) esas olarak açık tenli deneklerden oluştuğunu (sırasıyla %79,6 ve %86,2) ortaya koymuştur. Ticari cinsiyet sınıflandırma yapay zekâsı, bu kıyaslama verilerinde en iyi performansı gösterirken, **ten rengi koyu olan insanlarda orantısız şekilde yüksek hata oranları** yaşamıştır.

## **Deployment bias[¶](https://www.kaggle.com/code/alexisbcook/identifying-bias-in-ai" \l "Deployment-bias" \t "_self)**

**Yaygınlaştırma (Deployment) yanlılığı**, modelin çözmek için tasarlandığı problemin, gerçekte kullanıldığı yoldan farklı olması durumunda ortaya çıkar. Son kullanıcılar modeli amaçlandığı şekilde kullanmazlarsa, modelin iyi performans göstereceğine dair bir garanti yoktur.

Ceza adalet sistemi, hüküm giymiş bir suçlunun suç davranışına yeniden başlama olasılığını tahmin etmek için **araçlar** kullanır. Tahminler, yargıçların ceza zamanında uygun cezaları belirlemesi **için tasarlanmamıştır**.

ML iş akışının farklı aşamalarında ortaya çıkan bu farklı önyargı türlerini görsel olarak temsil edebiliriz:



Bunların birbirini dışlayan olmadığını unutmayın: yani, bir makine öğrenimi uygulaması birden fazla önyargı türünden kolayca etkilenebilir. Örneğin, Rachel Thomas'ın yakın zamanda yaptığı bir araştırma konuşmasında açıkladığı gibi, giyilebilir fitness cihazlarındaki makine öğrenimi uygulamaları şunlardan etkilenebilir:

Temsil önyargısı (modelleri eğitmek için kullanılan veri kümesi koyu ten renklerini hariç tutuyorsa),

Ölçüm önyargısı (ölçüm cihazı koyu ten renklerinde düşük performans gösteriyorsa) ve

Değerlendirme önyargısı (modeli kıyaslamak için kullanılan veri kümesi koyu ten renklerini hariç tutuyorsa).