**Hatasız Model Olmaz: Denetimli Öğrenmede Metrikleri Anlamak**

[[Feyza Keskin](https://medium.com/@keskinfeyza28?source=post_page---byline--f527b2cb84a2---------------------------------------)](https://medium.com/@keskinfeyza28?source=post_page---byline--f527b2cb84a2---------------------------------------)

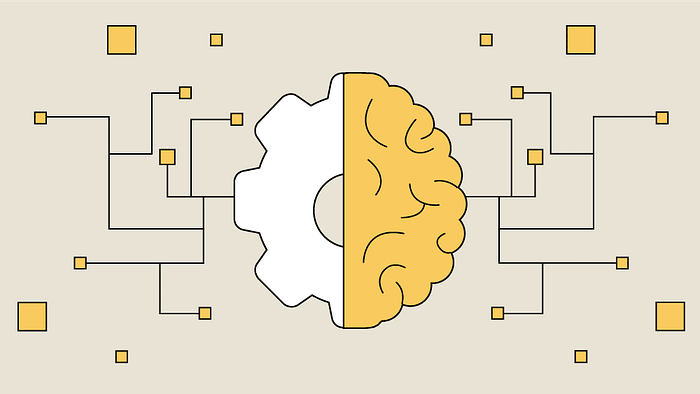
[Feyza Keskin](https://medium.com/@keskinfeyza28?source=post_page---byline--f527b2cb84a2---------------------------------------)

7 min read

·

4 days ago

Zoom image will be displayed



Şu anki piyasa koşullarına bakarak bir evin fiyatını tahmin etmek istediğinizi düşünün. En fazla kaç evin fiyatını karşılaştırıp bir tahminde bulunabilirsiniz? Bir makine öğrenmesi modeli sizin için binlercesine bakarak tahmin edebilir. Bu örnek yapabileceği şeylerin sadece biri. Peki hata yapabilir mi? Elbette.

Pratik anlamda gerçek dünya verileri ile **hatasız bir model mümkün değildir**. Bunun sebepleri şunlardır:

1. Verilerin genellikle eksik, hatalı veya çelişkili yani **gürültülü** olması. Bu sebeple %100 doğruluk imkansızlaşır.
2. Modeli eğittiğimiz veriler ile test ettiğimiz **verilerin dağılımının farklı olması** veya modelin hiç görmediği bir örnek ile karşılaştığı zaman hata yapabilmesi.
3. Modelin karmaşık verileri öğrenmek yerine **ezberlemesi** veya basitliği sebebi ile **öğrenememesi**. Bu iki durumda da genelleme hataları oluşabilir.

Amacımız hatasızlık değil, **optimum performans** olmalıdır.

Tabii ki modelimizi değerlendirmek için bazı metrikler mevcut. Biz bu yazıda **denetimli öğrenme** için kullanılan metriklerden bahsedeceğiz. Basit bir tanım olarak denetimli öğrenme, **bilgisayara bir dizi giriş (X) ve bu girişlere karşılık gelen doğru cevaplar (etiketler, Y) vererek bir model oluşturma sürecidir**.

**Denetimli Öğrenme Problem Türleri**

Denetimli makine öğrenmesi problemleri regresyon problemi ve sınıflandırma problemiolarak ifade edilir.

**Regresyon Problemi:**

Tahmin edilecek değerin sayısal olduğu problemlerdir. Ev fiyatı tahmini bu probleme uygundur. Sürekli değerler tahmin edilir (ondalıklı olabilecek değerler).

**Sınıflandırma Problemi:**

Tahmin edilecek değerin kategorik olduğu yani bir kategoriye atanması gerektiği problemlerdir. Bir e-postanın spam olması (birinci kategori) veya olmaması (ikinci kategori) bu probleme bir örnektir.

**Regresyon Değerlendirme Metrikleri**

En yaygın metrikler şunlardır :

1. **Ortalama Mutlak Hata (Mean Absolute Error — MAE)**

A mathematical equation with numbers and symbols

AI-generated content may be incorrect.

Ortalama mutlak hata, veri kümesindeki gerçek ve tahmin edilen değerler arasındaki mutlak farkın ortalamasını temsil eder. Buradaki y değerleri, gerçek değer ile tahmin edilen değerin farkı alınması için kullanılır. Küçük ve büyük hataları eşit değerlendiren bir metrik gerektiğinde kullanabiliriz. **Birimler aynı kalır ve hatalar eşit ağırlıkta değerlendirilir**. Finansal tahminler gibi **ortalama sapmanın önemli olduğu** durumlarda kullanışlıdır.

2. **Ortalama Karesel Hata (Mean Squared Error — MSE)**

A mathematical equation with numbers and symbols

AI-generated content may be incorrect.

Tahmin edilen değerlerin gerçek değerlerden ne kadar sapma gösterdiğini ölçer, ancak hataları karesini alarak büyütür. Yani büyük hatalara karşı daha hassas bir metriktir. Sonuç birimin karesi cinsinden olduğu için yorumlamak daha zor olabilir.

3. **Kök Ortalama Kare Hatası (Root Mean Squared Error — RMSE)**

A math equations with numbers and symbols

AI-generated content may be incorrect.

MSE’nin kareköküdür, modelin genel hata düzeyini gösterir. Hata biriminde sonuç alındığı için daha kolay yorumlanabilir.

4. **R-Kare Skoru (R² Score)**

A mathematical equation with numbers and symbols

AI-generated content may be incorrect.

Modelin ne kadar açıklayıcı olduğunu tespit etmeye yarar. Genel performansın bir özetidir ve 0 ile 1 aralığında sonuç alınır (1 = model mükemmel, 0 = model hiç öğrenmemiş). Büyük veri setlerinde daha faydalıdır.

Bu metrikler ile ilgili yorumlar şunlar olabilir :

* Farklı doğrusal regresyon modelleri arasındaki doğruluğu karşılaştırmak için RMSE, R Squared’den daha iyi bir seçimdir.
* MSE, MAE gibi türevlenemeyen bir fonksiyona kıyasla matematiksel işlemleri kolaylaştıran türevlenebilir bir fonksiyondur. Bu nedenle, birçok modelde RMSE, yorumlanması MAE’den daha zor olmasına rağmen Kayıp Fonksiyonu’nu hesaplamak için varsayılan bir ölçüt olarak kullanılır.
* MAE, MSE ve RMSE değerlerinin daha düşük olması, regresyon modelinin daha yüksek doğrulukta olduğunu gösterir. Ancak, daha yüksek bir R kare değeri tercih edilir.
* Hem RMSE hem de R-Kare, doğrusal bir regresyon modelinin bir veri kümesine ne kadar iyi uyduğunu ölçer. RMSE, bir regresyon modelinin bir yanıt değişkeninin değerini mutlak değerlerle ne kadar iyi tahmin edebildiğini gösterirken, R-Kare, tahmin edici değişkenlerin yanıt değişkenindeki değişimi ne kadar iyi açıklayabildiğini gösterir.

Örnek bir senaryoyu inceleyelim. Bir bankanın kredi riski modeli, bir kişinin 1 yıl içinde geri ödeme yapamama ihtimaline karşı beklenen zararı (TL bazında) tahmin ediyor.

Değerlendirme sonuçları şu şekilde :

* **MAE:** 1.200 TL
* **RMSE:** 5.400 TL
* **R²:** 0.62

Ortalama hata düşük gibi görünüyor olsa da RMSE’nin çok yüksek olması büyük tahmin hatalarının olduğunu gösteriyor. Bu gibi alanlarda büyük hatalı tahminler ciddi zarara neden olabildiğinden, RMSE burada çok daha anlamlıdır çünkü uç/aykırı değerlerleri cezalandırır diyebiliriz. R karenin nispeten düşük olması ise modelin riski yeterince açıklayamadığı anlamına gelir.

**Sınıflandırma Değerlendirme Metrikleri**

En yaygın metrikler şunlardır :

1. **Confusion Matrix (Hata Matrisi)**

Confusion Matrix, sınıflandırma problemlerinin temel değerlendirme aracı olup, modelin sınıf bazında doğru ve yanlış tahminlerini görselleştirir.

A diagram of a yellow square with black text

AI-generated content may be incorrect.

* **TP (Doğru Pozitif):** Pozitif sınıfın doğru tahmini. Yani bir kişinin kendisinde hastalık olduğunu “tahmin” edip gerçekten hasta olması gibi.
* **TN (Doğru Negatif):** Negatif sınıfın doğru tahmini. Kişinin hasta olmadığını doğru tahmin etmesi gibi.
* **FP (Yanlış Pozitif):** Negatif sınıfın pozitif olarak yanlış tahmini. (Tip I Hata olarak da bilinir). Kişinin gerçekten hasta olmadığı halde, hasta olduğunu tahmin etmesi gibi.
* **FN (Yanlış Negatif):** Pozitif sınıfın negatif olarak yanlış tahmini. (Tip II Hata olarak da bilinir). Kişinin hasta olduğu halde, hasta olmadığını tahmin etmesi gibi.

2.**Accuracy (Doğruluk)**

A black text with black lines

AI-generated content may be incorrect.

Bir sınıflayıcının performansını ölçmenin en basit yolu bu metriği kullanmaktır. Her veri noktasının gerçek ve tahmin edilen değerleri karşılaştırılır. Doğruluk, doğru tahmin sayısının toplam tahmin sayısına bölünmesiyle elde edilir. Tek başına kesinlikle yeterli değildir, çünkü FP ve FN gruplarını göz ardı ederek hesaplama yapar. Verilerde sınıflar arası dengesizlik olması durumunda yanıltıcı olabilir.

3. **Precision (Hassasiyet)**

A black text with black letters

AI-generated content may be incorrect.

Pozitif olarak tahmin edilenlerin, gerçekten pozitif olma oranını ölçer. Yanlış pozitiflerin kritik olduğu durumlarda önemlidir (örneğin, spam tespiti). Kesinlik, odaklandığı gruplar sebebiyle doğruluğun gözden kaçırdıklarının bir temsili gibidir.

4. **Recall (Duyarlılık / TPR — True Positive Rate)**

A mathematical equation with black text

AI-generated content may be incorrect.

Recall, Precision ile aynı prensibi kullanır ancak odak noktası artık yanlış pozitifler yerine yanlış negatiflerdir. Yanlış negatiflerin önemli olduğu durumlarda (örneğin, kanser teşhisi) kullanılır. Gerçek pozitiflerin, model tarafından ne kadar iyi tahmin edildiğini gösterir. Yine, gerçek negatifler dikkate alınmaz.

5. **F1 Score**

A black text with black text

AI-generated content may be incorrect.

Bu metrik hem Precision hem de Recall metriklerini dikkate alır, ikisi arasında bir denge kurma söz konusudur. Verilerin dengesiz olduğu durumlarda modelin genel performansını ölçmek için kullanılır.

6. **ROC-AUC (Receiver Operating Characteristic — Area Under Curve / Kesinlik-Hatırlama Eğrisi ve ROC Eğrisi)**

A diagram of a curve

AI-generated content may be incorrect.

* **ROC Eğrisi:** TPR (Recall) ile FPR (False Positive Rate) arasındaki ilişkiyi gösterir.
* **AUC (Eğri Altındaki Alan):** ROC eğrisinin altındaki alanı temsil eder. Modelin genel ayırt edicilik performansını ölçer. Bu metrik ROC eğrisinin sayısal bir şekilde ifade edilişidir. AUC, tüm olası sınıflandırma eşikleri için toplu bir performans ölçüsüdür. **AUC değerinin yüksek olması sınıflandırmanın iyi yapıldığı anlamına gelir**.

Precision ve Recall arasında bir denge söz konusudur. Her ikisi için aynı anda optimizasyon yapmak zordur, çünkü yanlış pozitifleri optimize etmek (dolayısıyla Precision’ı iyileştirmek) yanlış negatifleri (dolayısıyla Recall’u kötüleştirmek) pahasına olur ve bunun tersi de geçerlidir.

Bu durumda hangi metriğe öncelik verilmeli? Cevap, görevinizin niteliğine bağlıdır.

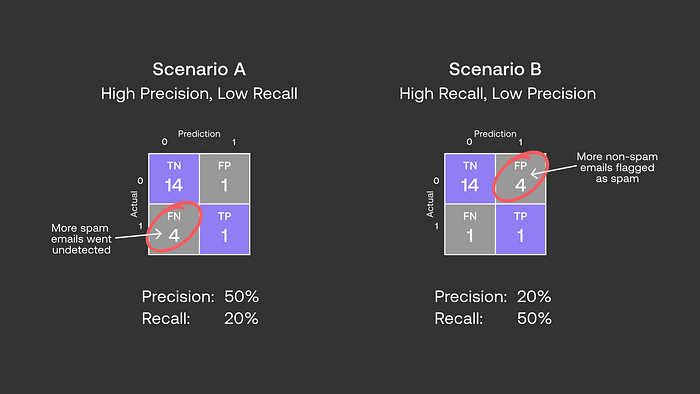
Örnek bir senaryo ile buna bakalım:

Spam sınıflandırıcının yüksek Precision ve düşük Recall (Senaryo A) elde ettiğini varsayalım. Bu, spam olmayan e-postaların spam olarak işaretlenmesiyle (FP) sonuçlanacaktır. Ancak bu aynı zamanda gerçek spam e-postalarının daha fazlasının tespit edilememesi (FN) anlamına da gelecektir.

Tersine, sınıflandırıcı yüksek Recall ve düşük Precision elde ederse (Senaryo B), daha az tespit edilemeyen spam e-postası olur (FN), ancak bu, daha fazla spam olmayan e-postanın spam olarak işaretlenmesi (FP) pahasına olur.

Spam sınıflandırma görevi için, önemli e-postaların spam klasörüne taşınmasını önlemek, ara sıra gelen spam e-postaların gelen kutusuna gitmesini önlemekten muhtemelen daha iyidir. Bu nedenle, bu görev için Precision’ı Recall’dan daha öncelikli hale getirmek istiyoruz.

Zoom image will be displayed



Benzer bir örneği F1 Score ve Accuracy karşılaştırması için verelim. Örneğin Bir e-posta servis sağlayıcısı, gelen e-postaları **“spam”** ve **“normal”** olarak sınıflandırmak istiyor. Ancak verinin %95’i **normal**, sadece %5’i **spam** olsun (yani dengesiz veri seti).

Eğer model hiç spam tespit etmeyip tüm e-postaları “normal” olarak tahmin etse bile doğruluk oranı %95 olarak görünür. Yani model aslında hiçbir spam e-postayı tespit edememiştir çünkü gerçek pozitifler (spam’ler) tamamen göz ardı edilmiştir.

Özellikle spam, sahtekarlık hastalık gibi önemli konularda F1 Score modelin “gerçek iş başarısını” daha doğru gösterir.

**Kaynakça**

**[MAE, MSE, RMSE, Coefficient of Determination, Adjusted R Squared — Which Metric is Better?](https://medium.com/analytics-vidhya/mae-mse-rmse-coefficient-of-determination-adjusted-r-squared-which-metric-is-better-cd0326a5697e?source=post_page-----f527b2cb84a2---------------------------------------" \t "_blank)**

[The objective of Linear Regression is to find a line that minimizes the prediction error of all the data points.](https://medium.com/analytics-vidhya/mae-mse-rmse-coefficient-of-determination-adjusted-r-squared-which-metric-is-better-cd0326a5697e?source=post_page-----f527b2cb84a2---------------------------------------" \t "_blank)

[medium.com](https://medium.com/analytics-vidhya/mae-mse-rmse-coefficient-of-determination-adjusted-r-squared-which-metric-is-better-cd0326a5697e?source=post_page-----f527b2cb84a2---------------------------------------" \t "_blank)

**[Classification Metrics Explained: Accuracy, Precision, Recall & F1 Score](https://cohere.com/blog/classification-eval-metrics?utm_source=chatgpt.com&source=post_page-----f527b2cb84a2---------------------------------------" \t "_blank)**

[How do you evaluate the performance of a Classifier? These are the four most commonly used classification evaluation…](https://cohere.com/blog/classification-eval-metrics?utm_source=chatgpt.com&source=post_page-----f527b2cb84a2---------------------------------------" \t "_blank)

[cohere.com](https://cohere.com/blog/classification-eval-metrics?utm_source=chatgpt.com&source=post_page-----f527b2cb84a2---------------------------------------" \t "_blank)

**[Makine Öğrenmesi(Machine Learning)](https://medium.com/machine-learning-t%C3%BCrkiye/makine-ogrenmesi-7cfbb3d859db?source=post_page-----f527b2cb84a2---------------------------------------" \t "_blank)**

[Makine öğrenmesi nedir? Hangi alanlarda kullanılır? Makine öğrenmesine 360 derece bakış.](https://medium.com/machine-learning-t%C3%BCrkiye/makine-ogrenmesi-7cfbb3d859db?source=post_page-----f527b2cb84a2---------------------------------------" \t "_blank)

[medium.com](https://medium.com/machine-learning-t%C3%BCrkiye/makine-ogrenmesi-7cfbb3d859db?source=post_page-----f527b2cb84a2---------------------------------------" \t "_blank)

**[Regresyon ve Sınıflandırmada Hata Metrikleri](https://medium.com/deep-learning-turkiye/regresyon-ve-s%C4%B1n%C4%B1fland%C4%B1rmada-hata-metrikleri-143a40c6b656?source=post_page-----f527b2cb84a2---------------------------------------" \t "_blank)**

[Merhabalar, bu yazıda makine öğrenmesi modellerinin doğruluğunu ölçmek için kullanılan metriklerden bahsedeceğim. Bu…](https://medium.com/deep-learning-turkiye/regresyon-ve-s%C4%B1n%C4%B1fland%C4%B1rmada-hata-metrikleri-143a40c6b656?source=post_page-----f527b2cb84a2---------------------------------------" \t "_blank)

[medium.com](https://medium.com/deep-learning-turkiye/regresyon-ve-s%C4%B1n%C4%B1fland%C4%B1rmada-hata-metrikleri-143a40c6b656?source=post_page-----f527b2cb84a2---------------------------------------" \t "_blank)