## SINIFLANDIRMA

Sınıflandırma problemi regresyon problemi ile beraber en yaygın gözetimli (supervised) öğrenme problemidir. Burada amaç eldeki verilere göre bir örneğin kategorisini bulmaktır. Bu alanda en çok kullanılan veri seti MNIST veri setidir. El yazısıyla yazılmış rakamların ne olduğunu bulmayı amaçlar. Bu proje için sınıflandırma probleminin "hello world'ü denebilir.

Yukarıdaki veri setinde her bir rakamın kaç olduğunu tahmin etmeye çalıştığımız problemlere multiclass classification denir. Veri setindeki rakam fotoğrafları 0-9 arası 10 kategoriye sahip şekilde etiketlenmiş olur.

Tüm rakamları tespit etmek yerine bir rakamın örneğin '5' olup olmadığını bulmaya çalıştığımızda buna binary classification denir. Bu durumda veri '5' veya '5 değil' şeklinde etiketlenir.

# Sınıflandırma Problemlerinde Performans Metrikleri

## Cross Validation

Model performansını ölçmenin en iyi yollarından bir k fold cross validation yöntemidir. Bu yöntemde veri k parçaya ayrılır. Her bir parça bir kez test k-1 defa eğitim setinde yer alacak şekilde model k defa eğitilir. Bu k eğitimin isabet oranlarının ortalaması sonucu verir.

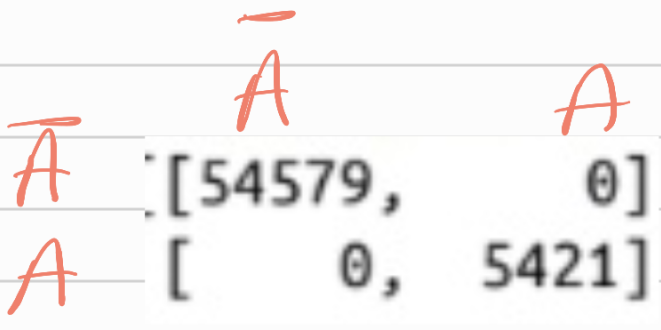
Burada bahsedilen isabet oranı (accuracy score) doğru tahmin edilen örneklerin tüm test örneklerine oranıdır. Rakamları tahmin etme probleminde rakamların %90'ı 5 değildir. Bu durumda tüm örneklere "5 değil" dediğimiz takdirde %90 accuracy score elde ederiz. Bu yanıltıcılığından dolayı accuracy score genellikle tercih edilmez.

Daha doğru sonuç verebilecek performans metriği confusion Matrix'dir.

## Confusion Matrix

Geliştirdiğimiz modelden aldığımız tahminleri değerlendirirken 5 olan fotoğraflara 5 değil dediğimiz ya da 5 olmayan fotoğraflara 5 dediğimiz tahminlerin oranını karşılaştırmak için konfüzyon matrisleri kullanılır.

Matristeki her satır bir etiketi temsil ederken sütunlar. tahminlenen etiketi temsil eder.



## Precision

Doğru pozitif tahminlerin tüm pozitif tahminlere oranıdır.

$$\text{Precision} = \frac{TP}{TP+FP}$$

| | Negatif | Pozitif |
|---|---|---|
| Negatif | TN | FP |
| Pozitif | FN | TP |

## Recall

Hassasiyet (sensitivity) olarak da bilinen recall doğru pozitif tahminlerin testteki tüm pozitif örneklere oranı denebilir.

$$\text{Recall} = \frac{TP}{TP+FN}$$

5 tahminleme modelimizi precision ve recall ile değerlendirecek olursak doğru 5'leri bulma oranının %83'e düştüğünü görüyoruz. Bu da aslında 5 içeren fotoğrafların sadece %65'ini tespit edebildiğimizi gösteriyor.

Precision ve Recall'u daha rahat anlayabilmek adına birleştiririz ve F1 scoru oluştururuz.

$$2x\ \frac{\text{precision}\ X\ \text{recall}}{\text{precision}\ +\ \text{recall}} = \frac{TP}{TP + \frac{FP+FN}{2}}$$

F1 score genellikle precision ve recall'un birbirine yakın olmasını ister. Gerçekte ise bazen precision'ın bazen de recall'un daha iyi olmasını tercih ederiz.Örneğin çocuklara zararlı videoları tespit eden bir modelin precisionının daha yüksek olmasını tercih ederiz. Dükkanda hırsızlık yapan insanları tespit eden modelimizin ise recallunun yüksek olmasını isteriz böylece daha az sayıda yanlış alarm vermiş ve daha az insanı hırsızlıkla suçlamış oluruz.

## The ROC Curve