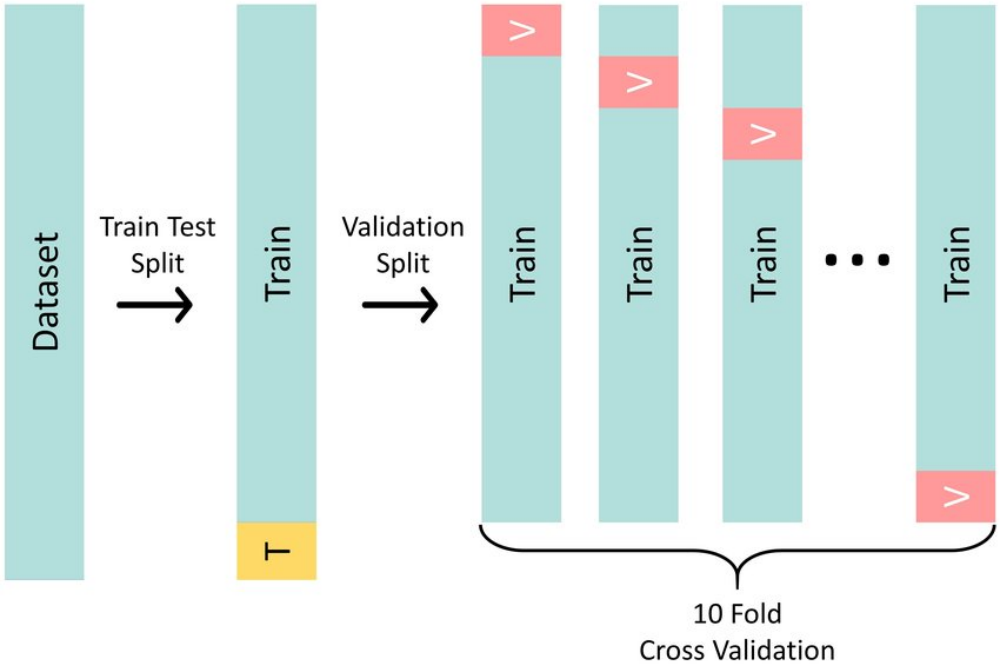
# Çapraz Doğrulama (Cross Validation) Yöntemleri

Atıl Samancıoğlu

## 1 Giriş

Makine öğrenmesi modelleri oluşturulurken, modelin performansını daha sağlam bir şekilde değerlendirmek için **çapraz doğrulama (cross validation)** yöntemleri kullanılır. Bu yöntemler, modelin farklı veri alt kümeleriyle test edilmesini sağlar ve genelleme yeteneğini ölçer.




Şekil 1: Train validation test ayrımına ve kfold'a örnek.

Bu dökümanda en yaygın çapraz doğrulama yöntemlerini açıklayacağız:

- Leave-One-Out (LOOCV)
- Leave-P-Out
- K-Fold Cross Validation
- Stratified K-Fold Cross Validation
- Time Series Cross Validation

## 2 1. Leave-One-Out (LOOCV)

Her deneyde verisetinden bir gözlem çıkarılır, kalan verilerle model eğitilir, çıkarılan gözlem ile test yapılır.
**Özellikleri:**

- Eğitim için $n-1$ gözlem, test için 1 gözlem kullanılır.
- $n$ gözlem varsa $n$ defa eğitim yapılır.

**Dezavantajları:**

- Büyük veri setlerinde oldukça yavaş çalışır.
- Aşırı öğrenmeye neden olabilir.

## 3 2. Leave-P-Out (LPOCV)

Bu yöntemde, her seferinde $p$ adet örnek doğrulama (validation) için ayrılır. Geri kalan verilerle eğitim yapılır. Bu işlem tüm olası kombinasyonlar için tekrar edilir.

**Avantajı:** Daha fazla veriyle test yapılarak güvenilirlik artar.

**Dezavantajı:** Kombinasyon sayısı çok arttığı için hesaplama maliyeti yüksektir.

2 gözlem doğrulama, 8 gözlem eğitim için (örnek)

Şekil 2: Leave-2-Out örneği: Kırmızı bloklar doğrulama verisi.

## 4 3. K-Fold Cross Validation

Veri seti $k$ eşit parçaya bölünür. Her seferinde bir parça doğrulama için, kalanlar eğitim için kullanılır. Bu işlem $k$ kez tekrarlanır.

**Avantajları:**

- Dengeli doğrulama sağlar.
- Hesaplama maliyeti LOOCV'ye göre düşüktür.

Fold 1: İlk 2 örnek doğrulama, geri kalan eğitim

Şekil 3: K-Fold (k=5) örneği: Her adımda farklı grup doğrulama için kullanılır.

## 5 4. Stratified K-Fold Cross Validation

**K-Fold'un sınıflandırma problemleri için iyileştirilmiş versiyonudur.** Katmanlardaki etiket oranlarını (ör. %60 1 ve %40 0) her katmanda korumaya çalışır.

**Avantajı:** Dengesiz veri setlerinde, doğrulama setinin sınıf dağılımını koruyarak daha doğru değerlendirme yapılmasını sağlar.

## 6 5. Time Series Cross Validation

Zaman serisi problemlerinde veriler ardışık olduğu için **rasgele bölme yapılamaz**. Bu yüzden, eğitim ve doğrulama verisi zaman sırasına göre ayrılır.

**Özellikleri:**

- Geçmiş veriler eğitimde, ileri tarihli veriler doğrulamada kullanılır.
- Veri sırası mutlaka korunmalıdır.


Eğitim (Mavi): Gün 1-4, Doğrulama (Kırmızı): Gün 5-8

Şekil 4: Time Series Cross Validation: Zamana duyarlı veri bölme örneği