Random Forest: Sınıflandırma ve Regresyon

Atıl Samancıoğlu

1 Giriş

Random Forest, **bagging (Bootstrap Aggregating)** yaklaşımını temel alan ve birçok **karar ağacının** birleşiminden oluşan güçlü bir topluluk (ensemble) yöntemidir. Hem **sınıflandırma** hem de **regresyon** problemlerini çözmek için kullanılır.

Bagging'deki temel fikir, farklı veri alt kümeleriyle birden fazla model eğitmek ve sonuçları birleştirmektir. Random Forest bu prensibi kullanırken tüm modellerini **karar ağaçları** olarak belirler.

2 Random Forest'ın Yapısı

- **Veri Seti:** D boyutunda toplam veri seti, m tane özellik $(f_1, f_2, ..., f_m)$.
- **Base Learners:** Tüm modeller **karar ağaçlarıdır.** Örneğin:
  - Karar Ağacı 1: DT_1
  - Karar Ağacı 2: DT_2
  - Karar Ağacı 3: DT_3
  - ...
- **Sampling:**
  - **Row Sampling:** D'den rastgele satırlar seçilir $(D' < D)$.
  - **Feature Sampling:** Rastgele özellikler seçilir $(F' \subset F)$.

Her karar ağacına farklı **veri ve özellik kombinasyonları** verilir. Bu işlem genellikle **bootstrapping (replacement ile sampling)** ile yapılır. Replacement sayesinde bazı row ve feature'lar kesişse bile, karar ağaçlarına verilen verilerin hepsinin aynı olmaması ve farklı örneklemleri incelemesi sağlanır.

3 Tahmin Aşaması

- **Sınıflandırma Problemlerinde:** Her bir karar ağacı kendi tahminini yapar. Nihai karar:

$$\text{Sonuç} = \text{mode}(\hat{y}_1, \hat{y}_2, ..., \hat{y}_n)$$

  yani **çoğunluk oylaması** ile belirlenir.

- **Regresyon Problemlerinde:** Tahminler sürekli sayılar olduğu için:

$$\text{Sonuç} = \frac{1}{n} \sum_{i=1}^{n} \hat{y}_i$$

  yani **modellerin tahminlerinin ortalaması** alınır.

4 Neden Random Forest?

Bir **karar ağacı** varsayılan parametrelerle kullanıldığında **overfitting** yapma riski taşır:

- Eğitim başarımı çok yüksektir (low bias).
- Test başarımı düşer (high variance).

Random Forest:

- **Variance**'ı düşürür (çoklu karar ağacı ile).
- Daha **genelleştirilebilir** sonuçlar sunar.

5 Şema: Random Forest Süreci

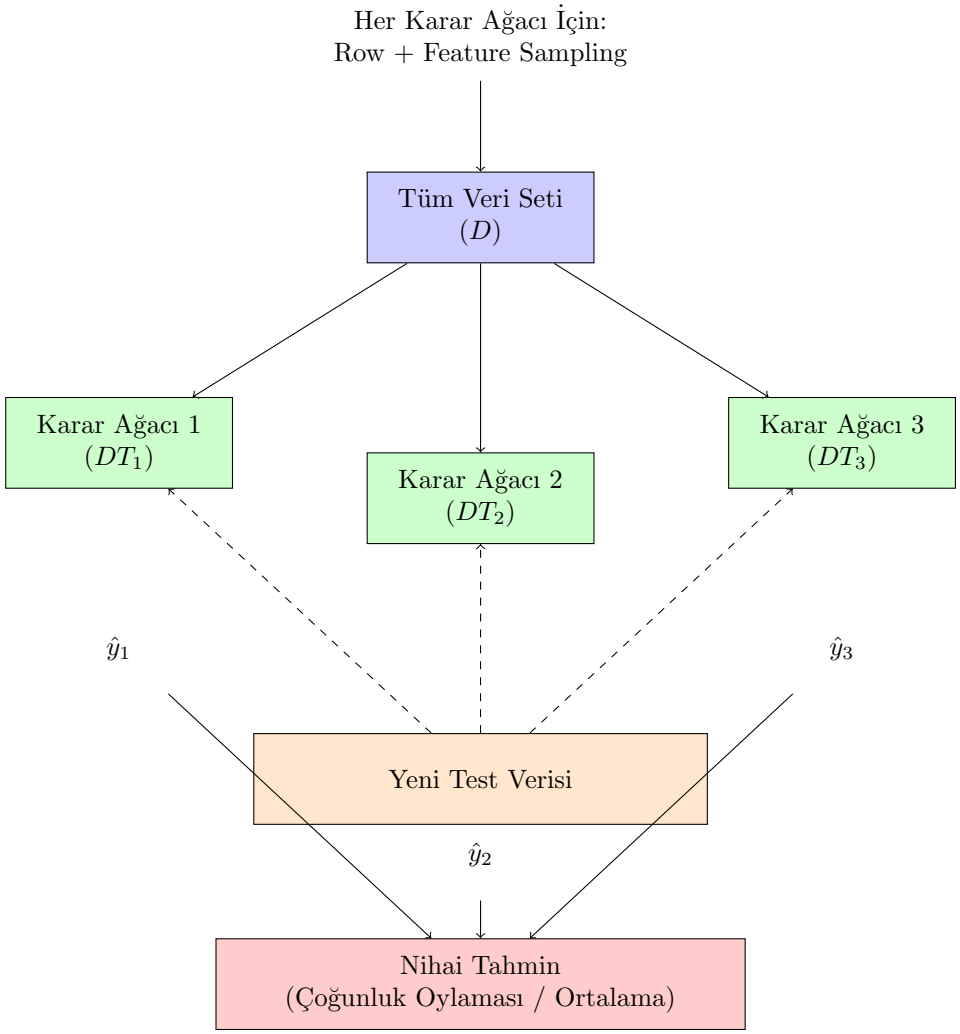

Figure 1: Random Forest sürecinin düzenlenmiş şematik gösterimi

6 Avantajlar ve Dezavantajlar

Avantajlar:

- Overfitting riskini azaltır.
- Karar ağaçlarının güçlü ve zayıf yönlerini dengeler.
- Hem sınıflandırma hem de regresyon için uygundur.

Dezavantajlar:

- Eğitim süresi uzun olabilir.
- Yorumlanabilirlik düşüktür (tek karar ağacı kadar kolay yorumlanamaz).