Kayıp Veri

Makine öğrenmesinde "kayıp veri" (missing data), bir veri setindeki bazı verilerin eksik olması durumunu ifade eder. Bu eksiklik, çeşitli nedenlerle ortaya çıkabilir ve makine öğrenmesi modellerinin performansını olumsuz yönde etkileyebilir. Kayıp verilerle başa çıkmak, makine öğrenmesi ve veri bilimi uygulamalarında önemli bir konudur.

Kayıp Veri Türleri

1. **MCAR (Missing Completely at Random - Tamamen Rastgele Eksik):** Veri kaybı tamamen rastgeledir ve herhangi bir sistematik nedene bağlı değildir. Örneğin, bir anket formunun bazı katılımcılar tarafından doldurulmamış olması.
2. **MAR (Missing at Random - Rastgele Eksik):** Veri kaybı, gözlemlenen diğer değişkenlerle ilişkilidir, ancak kayıp olan veriyle doğrudan ilişkili değildir. Örneğin, bir tıbbi araştırmada, yaşlı katılımcıların bazı test sonuçlarının eksik olması, ancak bu eksikliğin hastalığın şiddetiyle değil yaşla ilişkili olması.
3. **MNAR (Missing Not at Random - Rastgele Olmayan Eksik):** Veri kaybı, doğrudan kayıp olan veriyle ilişkilidir. Örneğin, bir gelir anketinde, yüksek gelirli bireylerin gelirlerini belirtmekten kaçınmaları.

Kayıp Verilerle Başa Çıkma Yöntemleri

1. **Eksik Verileri Göz Ardı Etme:**
   - **Liste Kayıt (Listwise Deletion):** Eksik veriye sahip tüm satırları veri setinden çıkarır. Kolay ve yaygın bir yöntemdir ancak önemli bilgi kaybına neden olabilir.
   - **Çift Kayıt (Pairwise Deletion):** Eksik veriye sahip olan satırları yalnızca ilgili analizlerde göz ardı eder. Veri setindeki mümkün olan en fazla bilgiyi kullanır.
2. **Veri İmputasyonu:**
   - **Ortalama/Medyan/Mod İmputasyonu:** Eksik değerler, ilgili sütunun ortalama, medyan veya mod değeri ile doldurulur. Basit ancak verilerin dağılımını bozabilir.
   - **K-En Yakın Komşu (KNN) İmputasyonu:** Eksik değerler, en yakın komşularının (k komşu) değerlerinin ortalamasıyla doldurulur.
   - **Regresyon İmputasyonu:** Eksik değerler, gözlemlenen diğer değişkenler kullanılarak regresyon modeli ile tahmin edilir.
   - **Multiple İmputation:** Eksik veriler, birden fazla kez imputasyon yapılarak doldurulur ve sonuçlar birleştirilir. Bu yöntem, imputasyonun belirsizliğini de dikkate alır.
3. **Gelişmiş Teknikler:**
   - **Maksimum Olabilirlik (Maximum Likelihood):** Eksik veri problemini çözmek için olasılık modelleri kullanılır. Veri setinin olasılıksal özelliklerini tahmin eder.
   - **Bayesyen Yöntemler:** Eksik veriler, bayesyen istatistikler kullanılarak tahmin edilir. Ön bilgiler ve gözlemler birleştirilir.

Kayıp Verilerin Etkileri

- **Model Performansı:** Eksik veriler, modelin doğruluğunu ve genelleme yeteneğini düşürebilir. İmputasyon yöntemleri doğru kullanılmazsa, yanıltıcı sonuçlar elde edilebilir.
- **Veri Dağılımı:** Eksik veriler yanlış veya hatalı imputasyon yöntemleri kullanıldığında veri dağılımını bozabilir ve bu da modelin performansını olumsuz etkileyebilir.
- **Önyargı:** Eksik verilerin sistematik olması durumunda, modelde önyargı oluşabilir. Örneğin, sadece belirli bir grup veriyi eksik bıraktıysa, model bu grubu doğru şekilde öğrenemeyebilir.

Sonuç

Makine öğrenmesinde kayıp veri, veri analizi ve modelleme sürecinde dikkat edilmesi gereken önemli bir konudur. Kayıp veri türlerinin ve uygun başa çıkma yöntemlerinin bilinmesi, daha doğru ve güvenilir modeller geliştirmeye yardımcı olur. Bu nedenle, kayıp verilerle başa çıkma stratejileri iyi anlaşılmalı ve veri setinin özelliklerine uygun yöntemler seçilmelidir.