Veri Temizleme (Data Cleaning)

Veri temizleme Makine öğrenmesi için neden çok önemlidir?

Makine öğreniminin iş akışının temeli temizlenmiş verilere bağlıdır

- Gözlemler : Verinin bir örneği. (Genelde bir veri kümesindeki nokta veya satır). Eğer veri kümesindeki bir satır veya nokta temiz değilse makineye verdiğimiz veri sorunlu demektir.
- Etiketler : Çıktının doğru tahmin edilebilmesi için verilerin doğru etiketlenmiş olması gerekir. Yanlış etiketli veriler doğal olarak yanlış tahminlere yol açar.
- Özellikler : Veriler hakkındaki bilgiler. Doğru özellik seçimi algoritma ve tahmin açısından son derece önemlidir.
- Algoritmalar
- Model

Şirketlerin verilerde en çok karşılaşıtığı sorunlar

- Veri eksikliği (Lack Of Data)
- Aşırı fazla veri

- Kötü veri - Makine öğreniminde kullanılacak verinin eksiksiz ve özellik mühendisliği uygun şekilde yapılmılş olması gerekir.

Veriler Nasıl Dağınık olabilir?

- Gereksiz veya tekrar edilen veriler barındırma.
- Verilerdeki yazım hataları. (Aynı etiketi büyük ve küçük harflerle yazmak iki farklı etiket oluşmasına yol açar)
- Eksik veriler
- Aykırı veriler (Veri kümesinin tamamından sapmış olan ve algoritmayı saptırma potansiyeli olan verilerdir.)
- Veri kaynağından Kaynaklanan durumlar. (Farklı veritabanı tipleri vs.)

Eksik veri durumunda izlenebilecek yöntemler

1. Tüm kolonu çıkarmak

Eksik veri bulunan kolondaki tüm verileri çıkararak eksik veri sorunu ortadan kaldırılabilir ancak bu kolaya kaçmak olabilir. Elimizdeki kolon değerli olabilecek verileri içeriyorsa diğer yöntemleri kullanmak daha mantıklı olablir.

2. Eksik verileri doldurmak

Eksik verileri Doldurmak kolonda varolan verilerin kaybolmaması açısından önemli bir durumdur. Dolsurmak içi,n birden fazla çeşidimiz mevcuttur. Örneğin kategorik bir veri ise moda göre, Sayısal bir veri ise ortalamaya göre doldurma (imputing) işlemleri yapılabilir.

3.Kolonu maskelemek

Aykırı değerlerin analizi

🧐 Aykırı değerler (Outliers)