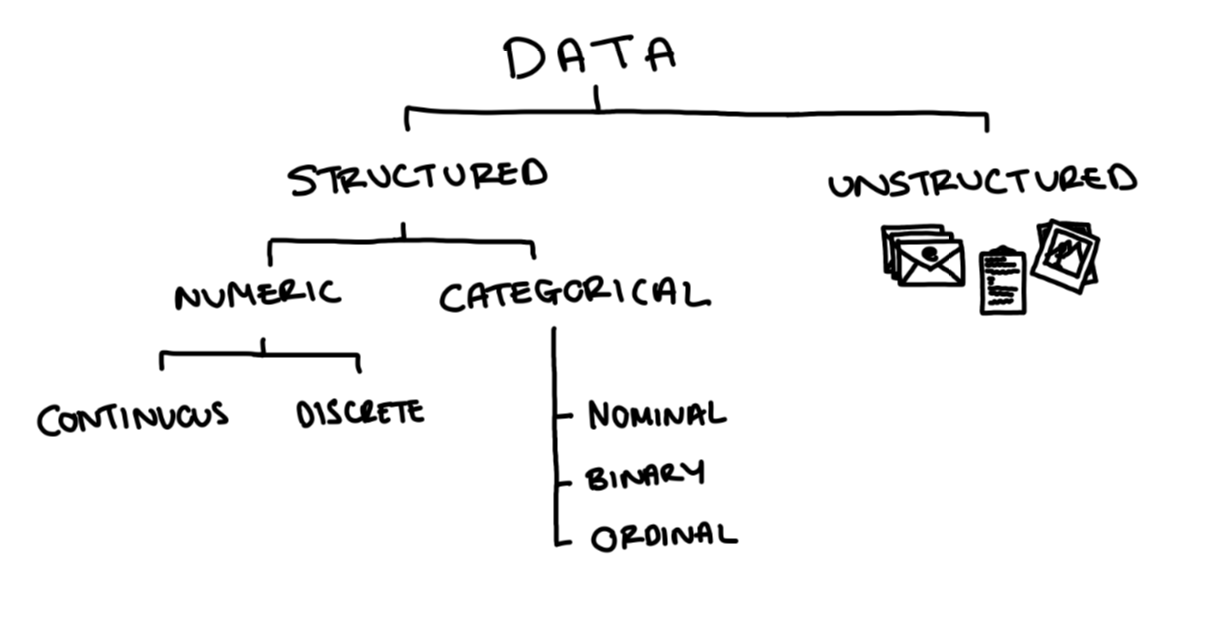
**1. Aşama**

Makine öğrenmesi sürecinde, modellerin tahminler yapmasını veya kararlar almasını sağlamak için verilerden bilgiler öğrenilir. Bu süreçte veriler, modelin öğrenebilmesi için en önemli kaynaktır. Eğer veri yoksa, makine öğrenmesi modeli hiçbir şey öğrenemez. Verinin kalitesi, modelin doğruluğunu, genelleme gücünü ve genel performansını doğrudan etkiler. Temiz ve kaliteli veri, modelin daha doğru sonuçlar üretmesini sağlarken, eksik ya da hatalı veriler modelin yanlış öğrenmesine ve yanıltıcı tahminler yapmasına yol açabilir.

Yapısal veri, düzenli bir formatta saklanan ve genellikle tablo şeklinde organize edilen verilerdir, her sütun belirli bir veri türünü temsil eder. Yapısal olmayan veri ise, belirli bir düzeni olmayan ve genellikle metin, resim, ses veya video gibi öğelerden oluşur. Bu veriler genellikle işlenmeden önce ön işleme adımlarına tabi tutulur.

Sayısal veri, sayısal değerlerden oluşur ve sürekli ya da kesikli olabilir. Kategorik veri ise, sınırlı sayıda belirli kategoriye ait olan verilerdir ve özellikleri temsil eder.



**2. Aşama**

Yapay zekanın varlığının temeli veriye dayanır ve makine öğrenmesi modellerinin performansı, veri kalitesiyle doğrudan ilişkilidir.

Temizlenmemiş veriler şu sorunlara yol açar:

* Hatalı veriler, modelin yanlış öğrenmesine yol açar.
* Eksik veriler, makinenin yanlış öğrenmesine yol açabilir, bazı durumlarda algoritmaların tamamen çalışmasını engeller.
* Kategorik verilerin dönüştürülmemesi, algoritmalar tarafından verinin anlaşılmamasına ve algoritmanın o satır veya sütundan faydalanamamasına sebep olabilir.

Veri temizleme ve ön işleme, modelin doğruluğu ve performansını doğrudan etkiler.

**Veri Temizleme Adımları**

1–) **Eksik Verilerle Başa Çıkma**

**Silme:** Eksik değer içeren satırlar dropna() fonksiyonu ile silinir ancak bu yöntemin veri kaybı riski bulunmaktadır.

**Doldurma:** Eksik veriler, seçilen yönteme göre doldurulur. Seçilebilecek yöntemler şunlar olabilir: Ortalama (mean), medyan (median), mod veya K-NN algoritması.

2-) **Tekrarlanan Verilerin Silinmesi**

Tekrarlanan veriler tespit edilerek drop\_duplicates() fonksiyonu ile silinir.

3-) **Aykırı Değerlerin İşlenmesi**

Veri setinin %75. değer ile %25. değer arasındaki mesafe (IQR) bulunur ve %75. değer ile %25. değerden IQR mesafesinin 1.5 katı kadar uzaklığına sınırlar konur. Bu sınırların dışındaki değerler aykırı (outlier) olarak kabul edilir ve aykırı değerlerin sayısı ve durumuna göre işlemler yapılır. İşlemlere birkaç örnek: Aykırı değerleri silme, sayısal verilerin dönüşümü (logaritma gibi) veya sınırlanması (winsorization).

**Kategorik Verilerin Dönüştürülmesi**

1-) **One-Hot Encoding**

Belli bir sıralamaya sahip olmayan her kategorik veri için yeni bir binary sütuna dönüştürülerek makine için kullanılmaya hazır hale getirilir.

2-) **Label Encoding**

Belli bir sıralamaya sahip olan her kategorik veri sayısal sütuna dönüştürülerek makine için kullanılmaya hazır hale getirilir. Bu Sayısal değerler bir sıralamaya sahiptir, örneğin “2 > 1” şeklinde.



**Sayısal Verilerde Ölçeklendirme**

1-) **Normalizasyon (Min-Max Scaling)**

Verideki sayısal verileri 0 ile 1 arasına sıkıştırır.

2-) **Standartlaştırma (Z-Score Normalization)**

Verideki sayısal verileri ortalaması 0, standart sapması 1 olacak şekilde dönüştürür.

**3. Aşama**

Model eğitimi, veri seti üzerindeki verilerin birbiri üzerindeki ilişkileri inceleyerek aralarındaki bağlantıyı öğrenip örüntüler çıkararak tahmin yapabilen bir sistem oluşturma sürecidir ve temel adımları şunlardır:

1. **Veri Toplama ve Hazırlama**:  
   a-) Veri seti toplanır, temizlenir ve ön işlemlerden geçirilir.  
   b-) Veri seti, eğitim ve test setlerine ayırılır.
2. **Model Seçimi**  
   a-) Problem türüne uygun bir algoritma seçilir.
3. **Model Eğitimi**a-) Eğitim verisi kullanılarak modelin parametreleri öğrenilir.  
   b-) Veri setindeki değerlerin ağırlıkları (weights) optimize edilir.
4. **Hiperparametre Ayarlaması**a-) Modelin sahip olduğu performansı arttırmak için modelin parametrelerinin arttırılması, Örneğin K-NN algoritması için k değerinin optimize edilmesi.
5. **Model Değerlendirme**a-) Test verisi üzerinde modelin performansı ölçülür.  
   b-) Test verisi üzerinde başarı ölçütleri hesaplanır.
6. **Modelin Kullanılması**a-) Model artık yeni veriler için çalışmaya hazır durumdadır.

**Denetimli (Supervised) ve Denetimsiz (Unsupervised) Öğrenme Arasındaki Farklar**

**Denetimli Öğrenme (Supervised)**

Etiketli veriler kullanır.

Sınıflandırma veya regresyon tahmini yapmak için kullanılır.

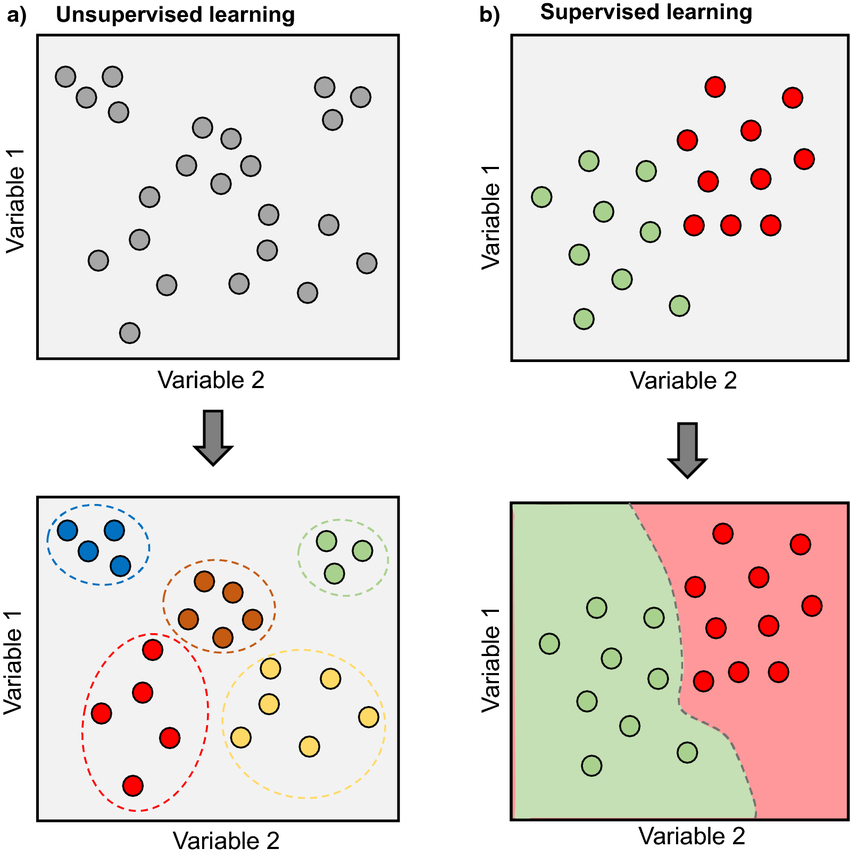
Örnek Algoritmalar: Doğrusal Regresyon, Karar Ağaçları, SVM

**Denetimsiz Öğrenme (Unsupervised)**

Etiketsiz veriler kullanır.

Kümeleme veya boyut indirgeme kullanarak verideki örüntüleri keşfetmek için kullanılır

Örnek Algoritmalar: K-Means, PCA, DBSCAN



**Temel Makine Öğrenmesi Algoritmaları**

1. **Doğrusal Regresyon**  
   Bağımsız değişkenler ile bağımlı değişkenler arasındaki ilişkiyi modelleyip sürekli değişkenlerin tahmininde kullanılır.
2. **Karar Ağaçları**  
   Veriyi istenilen sayıda düğümlere bölerek dallara ayırılır. Yeni veriler bu düğümler üzerinden tahmin edilir.
3. **K-NN**  
   Yeni gelen bir veri, en yakınındaki k sayısı kadar veriye bakılarak yeni verinin sınıflandırılması yapılır.

**Model Değerlendirme Metrikleri**

1. **Doğruluk**Sınıflandırma problemlerinde, modelin doğru tahmin ettiği örneklerin, toplam örneğe oranıdır.
2. **MSE**Regresyon problemlerinde, tahmin edilen değerler ile gerçek değerler arasındaki farkların karelerinin ortalamasıdır.
3. **Precision & Recall**Precision, modelin pozitif olarak doğru tahmin ettiği değerin modelin pozitif olarak tahmin ettiği değerlere oranı iken, Recall, modelin pozitif olarak doğru tahmin ettiği değerlerin gerçekte pozitif olanlara oranıdır.