MAKİNE ÖĞRENMESİ

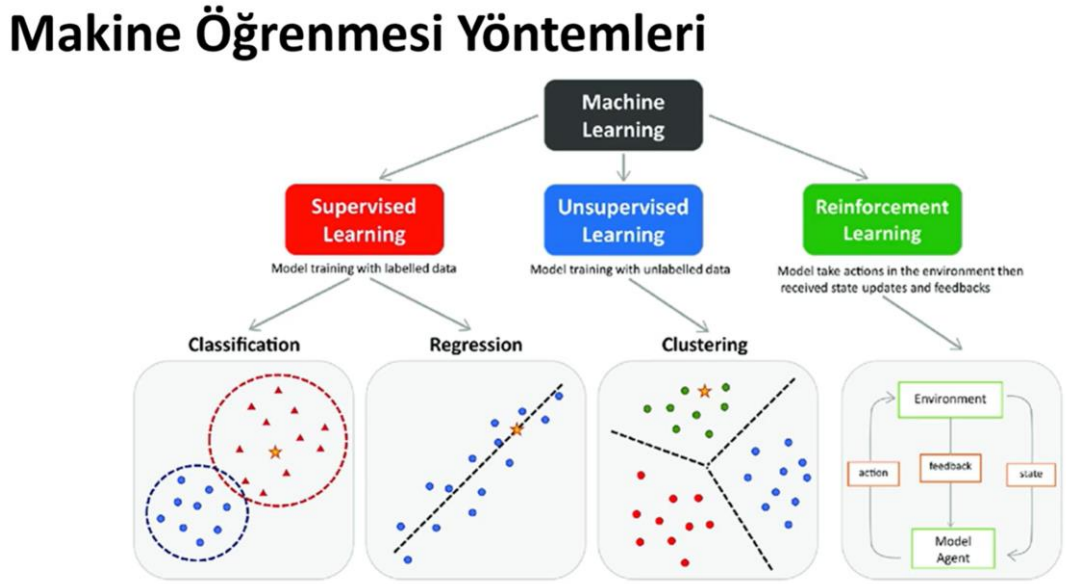
Makine öğrenmesi (ML), bilgisayar sistemlerinin açıkça programlanmadan veri ve deneyim yoluyla öğrenmesini sağlayan bir yapay zeka dalıdır. Makine öğrenmesinde, bilgisayara büyük miktarda veri verilir ve bu verilerden kalıpları, ilişkileri öğrenerek gelecekteki veriler hakkında tahminler yapması beklenir.

Makine öğrenmesi, istatistiksel yöntemlere dayanır ve algoritmalar, verilerle beslenerek zamanla performanslarını iyileştirir. Genelde üç ana kategoriye ayrılır:

Denetimli Öğrenme (Supervised Learning): Bu yöntemde sistem, etiketlenmiş veri setleri ile eğitilir. Sistem, girdiler ve bu girdilere karşılık gelen doğru çıktıları öğrenir. Yeni bir veri geldiğinde, sistem bu öğrendiği bilgiyi kullanarak tahmin yapar. Örnek: E-posta spam filtreleme.

Denetimsiz Öğrenme (Unsupervised Learning): Etiketlenmemiş veriler kullanılır ve sistemin verideki gizli kalıpları keşfetmesi amaçlanır. Örnek: Müşteri segmentasyonu, veri kümelerinin gruplandırılması.

Pekiştirmeli Öğrenme (Reinforcement Learning): Bu yöntemde, bir ajan (yani öğrenen sistem), bir ortamla etkileşime girer ve belirli eylemler için ödül veya ceza alarak öğrenir. Amaç, uzun vadede en yüksek ödülü sağlayacak eylemleri bulmaktır. Örnek: Oyun oynayan yapay zeka sistemleri.



MAKİNE ÖĞRENMESİNDA KULLANILAN TERMİNOLOJİLER:

1. Veri Seti (Dataset)

Eğitim Verisi (Training Data): Makine öğrenme modelini eğitmek için kullanılan etiketli veri.

Test Verisi (Test Data): Modelin performansını değerlendirmek için kullanılan veri.

2. Özellik (Feature)

Veri setindeki her bir gözlem hakkında bilgi veren bireysel veri parçaları. Örneğin, bir araba için "hız", "ağırlık", "renk" gibi bilgiler özelliklerdir.

3. Özellik Mühendisliği (Feature Engineering)

Verilerden anlamlı ve etkili özellikler oluşturma sürecidir.

4. Etiket (Label)

Denetimli öğrenmede, her bir girdiye karşılık gelen doğru çıkış (sonuç). Örneğin, bir spam tespiti sisteminde e-postaların "spam" ya da "spam değil" etiketi olabilir.

5. Model

Eğitilmiş algoritmanın, girdilere göre bir tahmin ya da karar yapmasını sağlayan matematiksel yapı.

6. Algoritma

Verileri kullanarak modelin eğitilmesini sağlayan matematiksel işlem dizisidir. Makine öğrenmesinde kullanılan popüler algoritmalar şunlardır:

Doğrusal Regresyon (Linear Regression): Bir bağımlı değişken ile bir veya daha fazla bağımsız değişken arasındaki ilişkiyi modellemek için kullanılır.

Destek Vektör Makineleri (Support Vector Machines): Verileri, en iyi sınıflandıran doğruyu bulmak için kullanılır.

Karar Ağaçları (Decision Trees): Karar yapma süreçlerini modellemek için kullanılan ağaç yapısındaki algoritmadır.

7. Hiperparametre (Hyperparameter)

Modelin eğitimi başlamadan önce ayarlanan ve modelin yapısını kontrol eden parametrelerdir. Örneğin, bir karar ağacındaki maksimum derinlik, bir hiperparametredir.

8. Aşırı Uydurma (Overfitting)

Modelin eğitim verilerine çok fazla uyum sağlaması ve bu nedenle yeni verilerde kötü performans göstermesi durumudur.

9. Az Uydurma (Underfitting)

Modelin verilerdeki ana kalıpları bile öğrenemediği durumdur. Yetersiz model karmaşıklığı veya yeterince eğitilmemiş bir model sonucu ortaya çıkar.

10. K-Fold Cross Validation

Veriyi eğitim ve test setlerine bölerek modeli eğitme ve değerlendirme yöntemidir. Veriler K farklı bölüme ayrılır ve her bir bölme test verisi olarak kullanılırken diğer bölümler eğitim için kullanılır.

11. Doğruluk (Accuracy)

Modelin doğru tahmin sayısının toplam tahmin sayısına oranıdır. Yani, modelin ne kadar isabetli çalıştığını gösteren bir ölçümdür.

12. Kesinlik (Precision) ve Geri Çağırma (Recall)

Kesinlik (Precision): Modelin doğru pozitif tahminlerinin, toplam pozitif tahminlere oranıdır. Yani, modelin doğru pozitif sınıflandırmaları ne kadar isabetli bulduğunu gösterir.

Geri Çağırma (Recall): Doğru pozitif tahminlerin, gerçek pozitif değerlerin tamamına oranıdır. Yani, gerçek pozitifleri bulmada modelin başarısını gösterir.

13. Gradyan İnişi (Gradient Descent)

Öğrenme algoritmalarının, hata fonksiyonunu minimize etmek için parametrelerini adım adım optimize ettiği yöntemdir. Bu yöntem, modelin en iyi ağırlıklarını bulmak için kullanılır.

14. Maliyet Fonksiyonu (Cost Function)

Modelin hata miktarını ölçen bir fonksiyondur. Modelin yaptığı tahmin ile gerçek sonuç arasındaki farkı hesaplar.

15. Epoch

Eğitim veri setinin modelden tam olarak bir kez geçmesi sürecine denir. Model, her epoch'ta öğrenmesini günceller.

16. Ağırlık (Weight)

Modelin tahmin yaparken her bir özelliğe ne kadar önem verdiğini gösteren parametrelerdir.

17. Sınıflandırma (Classification)

Sınıflandırma, denetimli öğrenme kategorisinde yer alır ve modelin, verilen verileri belirli kategorilere veya sınıflara ayırmasıdır. Girdi verileri etiketlenmiştir ve model, bu etiketlere göre sınıflandırmayı öğrenir.

Örnekler:

Bir e-postanın spam olup olmadığını sınıflandırmak.

Bir hastalığın pozitif ya da negatif olduğunu tespit etmek.

Görüntüde bir nesnenin kedi mi köpek mi olduğunu ayırt etmek.

18.Kümeleme (Clustering)

Kümeleme, denetimsiz öğrenme tekniklerinden biridir ve modelin verileri önceden belirlenmiş bir etikete göre değil, verilerin benzerliklerine göre gruplara ayırmasını içerir. Kümeleme, verilerdeki doğal yapıyı ve kalıpları ortaya çıkarmaya çalışır. Bu yöntem, sınıflandırmada olduğu gibi önceden belirlenmiş etiketlere ihtiyaç duymaz.

Örnekler:

Müşteri segmentasyonu (benzer davranışlara sahip müşterileri gruplandırma).

Görüntü veri kümesinde benzer nesneleri gruplandırma.

19. Regresyon (Regression)

Regresyon, denetimli öğrenmenin başka bir alt dalıdır ve sürekli (numerik) verilerde tahmin yapmak için kullanılır. Modelin amacı, bir girdiye karşılık gelen sürekli bir değeri tahmin etmektir.

Örnekler:

Ev fiyatlarının tahmini.

Bir ürünün gelecekteki satışlarını tahmin etme.

Bu sondaki üç yöntem de farklı problemlere yönelik makine öğrenme çözümleri sunar.Her biri farklı veri tiplerine ve probleme göre tercih edilir.

VERİ ÖN İŞLEME

Veri setini hazırlama sürecidir.

Aşamaları:

- Veri temizleme
- Veri standardizasyonu
- Öznitelik seçimi
- Veri dönüşümü
- Aykırı değerlerin işlenmesi

KAYIP VERİ PROBLEMİ

Kayıp veri problemi (missing data problem), bir veri setinde bazı değerlerin eksik veya hatalı olması durumunda ortaya çıkar .Kayıp veri, genellikle verilerin yanlış toplanması, sensör hataları veya anketlerde yanıtlanmayan sorular gibi nedenlerle meydana gelir. Kayıp veri problemini çözmek için yöntemler :

1. Kayıp Veriyi Göz Ardı Etme (Removing Missing Data)

Satır veya sütunları silme: Eğer kayıp veri sayısı az ise, eksik veriye sahip satırları veya sütunları tamamen silmek yaygın bir çözümdür. Ancak bu yöntem, özellikle çok sayıda eksik veri olduğunda veri kaybına neden olabilir ve modelin genelleme kapasitesini olumsuz etkileyebilir.

2. Ortalamayla Doldurma (Mean/Median/Mode Imputation)

Ortalamayı kullanma: Sayısal verilerde kayıp veriyi sütunun ortalaması ile doldurmak yaygın bir yöntemdir. Örneğin, bir veri setinde bir sütunda maaş bilgileri eksikse, eksik değerler o sütunun ortalama maaşı ile doldurulabilir.

Medyan kullanma: Aykırı değerlerin olduğu durumlarda, medyan daha iyi bir doldurma yöntemidir.

Mod kullanma: Kategorik verilerde en sık tekrarlanan değeri (mod) kullanarak eksik veriyi doldurmak mantıklı olabilir.

3. İleri Doldurma ve Geri Doldurma

(Forward/Backward Filling)

İleri doldurma (forward fill): Zaman serisi verilerinde önceki değeri eksik olan yerin yerine yazmak.

Geri doldurma (backward fill): Sonraki değeri eksik olan yere yazmak.

Bu yöntemler, özellikle zamana dayalı veri setlerinde sık kullanılır.

4. Tahmin Edici Modeller (Predictive Models)

Eksik veriyi tahmin etmek için diğer sütunları kullanarak bir makine öğrenmesi modeli eğitmek mümkündür. Bu yöntemle, eksik olan veriler regresyon, k-NN (k-En Yakın Komşu) veya karar ağacı gibi algoritmalar kullanılarak tahmin edilebilir.

Avantajı: Modelin kayıp veriyi tahmin edebilme yeteneğini geliştirir ve daha karmaşık verilerle başa çıkılabilir.

Dezavantajı: Karmaşıklığı artırır ve daha fazla hesaplama gücü gerektirir.

5. Çoklu Atama (Multiple Imputation)

Bu yöntem, eksik veriler için birden fazla olası değer oluşturarak bu değerler arasında rastgele seçim yapar. Farklı doldurma setleri üretilir ve her bir doldurulmuş veri seti ile model eğitilir.

Avantajı: Veri kaybı olmaz ve belirsizlikle daha iyi başa çıkar.

Dezavantajı: Hesaplama açısından yoğundur ve daha karmaşık bir sürece sahiptir.

6. KNN ile Doldurma (K-Nearest Neighbors Imputation)

Kayıp veriyi, komşu (k-NN algoritması kullanarak) en benzer verilere dayanarak doldurmak mümkündür. Bu yöntemde, eksik veri noktasına en yakın k komşu gözlemler bulunur ve bu komşuların değerlerinin ortalaması alınarak eksik veri doldurulur.

Avantajı: Doldurulan değer, veri setindeki diğer verilere dayandığı için daha gerçekçi olabilir.

Dezavantajı: Büyük veri setlerinde hesaplama açısından maliyetlidir.

7. Veri Artırma (Data Augmentation)

Eksik verileri artırmak veya yeni veri noktaları oluşturmak, veri setini genişletme stratejisi olabilir. Genelde derin öğrenme ile ilişkilendirilir, ancak bazı durumlarda eksik verileri dengelemek için kullanılır.

8. Eksik Veri İçin Bayes Yaklaşımları

Bayes istatistiklerini kullanarak eksik verilerin olasılığını modelleyebilir ve bu olasılık değerlerine göre veriyi doldurabilirsiniz. Bu, belirsizliği hesaba katan güçlü bir yöntemdir, ancak daha ileri düzey istatistik bilgisi gerektirebilir.

Supervised Learning:

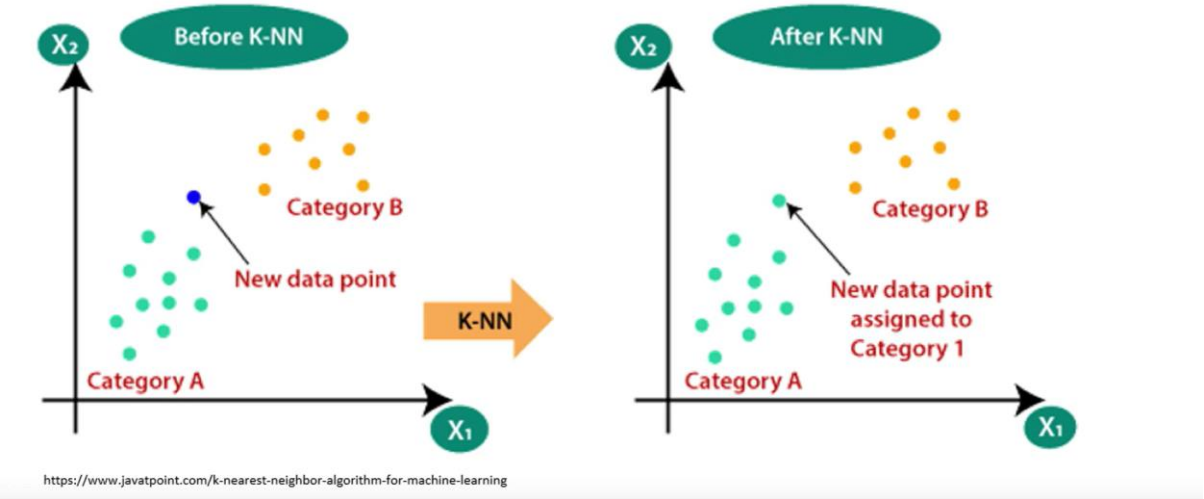
Gözetimli Öğrenme Algoritmaları

- Sınıflandırma
  - K-En Yakın Komşu
  - Karar Ağaçları
  - Rasgele Orman
  - Logistic Regresyon
  - Destek Vektör Makinesi
  - Naive Bayes
- Regresyon
  - Lineer Regresyon
  - Çoklu Lineer Regresyon
  - Polinom Regresyon

CLASIFICATION(SINIFLANDIRMA)

1. K-En Yakın Komşu (K-Nearest Neighbors, KNN):

KNN, sınıflandırma problemlerinde, bir veriyi sınıflandırırken ona en yakın "k" komşusunun sınıflarına bakar ve çoğunluğa göre sınıflandırma yapar.

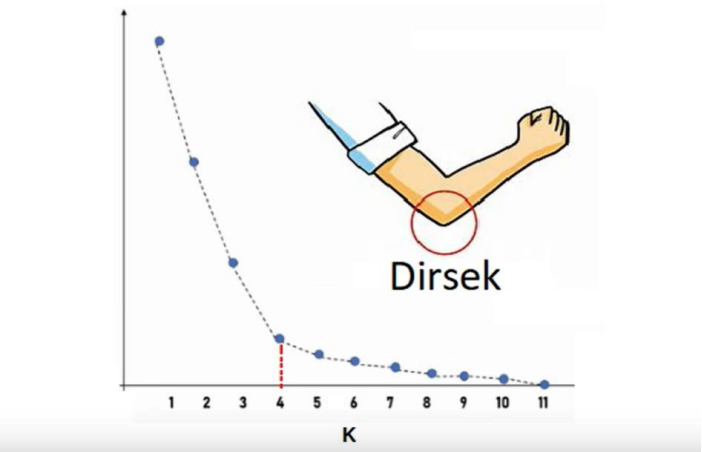


KNN Algoritmasının Adımları:

- Parametre Olarak K'nin Belirlenmesi
- Mesafe Ölçümü
- En Yakın K Komşunun Bulunması

- Çoğunluk Oyu ile Sınıflandırma

Elbow method, KNN gibi kümeleme ve sınıflandırma algoritmalarında K(komşu sayısı) değerini seçmek için kullanılan bir tekniktir. Bu yöntem, farklı K değerlerini deneyerek modelin performansını değerlendirir ve optimum K değerini belirlemeye çalışır. Elbow methodu, K değerinin artmasıyla birlikte modelin performansındaki azalışın hızının azalması ve eğrinin dirsek gibi bükülmesi prensibine dayanır.



Uygulama Alanları:

- Sağlık(Hastalık teşhisinde risk tahmini)
- E-ticaret(Kullanıcını geçmiş alıverişine dayanarak ürün önerisi sunma)
- Finans(Bir müşterinin kredi başvurusu değerlendirilirken risk analizi)

KNN'nin Avantajları:

- Basit ve Anlaşılır
- Eğitim Gerektirmez

KNN'nin Dezavantajları:

- Hafıza Tüketimi
- Yüksek Hesaplama Maliyeti
- Kayıp Verilere Duyarlılık
- K Değerine Duyarlılık

KNN ile Sınıflandırma Yapılan Veri Setleri

- İris veri seti: Çiçek türlerinin (setosa, versicolor, virginica) sınıflandırıldığı bir veri setidir. Her veri noktası, sepal uzunluğu, sepal genişliği, petal uzunluğu ve petal genişliği gibi ölçüler içerir.
- MNIST veri seti: El yazısı rakamlarının (0-9) sınıflandırıldığı bir veri setidir. Her veri noktası, bir el yazısı rakamının piksel değerlerini içerir.

- Hasta verileri: Tıbbi veri setlerinde hastaların bazı özelliklerine (yaş, kan basıncı, kolesterol) göre hastalıklarının (kalp hastası olup olmadığı gibi) sınıflandırıldığı **veri setleri kullanılır.**

KNN ile Kullanılan Veri Tipleri

- Numerik (Sayısal) Veriler: KNN algoritması genellikle sayısal veri tipleri ile çalışır. Mesafeye dayalı bir algoritma olduğu için veri noktalarının karşılaştırılabilmesi amacıyla sayısal özelliklerin bulunması önemlidir. Örneğin, ağırlık, boy, yaş gibi sayısal veriler.
- Kategorik Veriler: KNN'de genellikle numerik verilerle çalışılsa da kategorik veriler de kullanılabilir. Örneğin, cinsiyet veya sınıf etiketleri gibi.
- Görüntü ve Ses Verileri: Özellikle piksel temelli görüntü verileri, KNN algoritmasıyla işlenebilir. El yazısı tanıma veya yüz tanıma gibi problemler için kullanılabilir.
- Metin Verileri: KNN algoritması, metin verilerinin sayısal temsilleriyle (TF-IDF veya kelime vektörleri gibi) de çalışabilir. Örneğin, e-posta sınıflandırması yaparken metin verisi kullanılabilir.

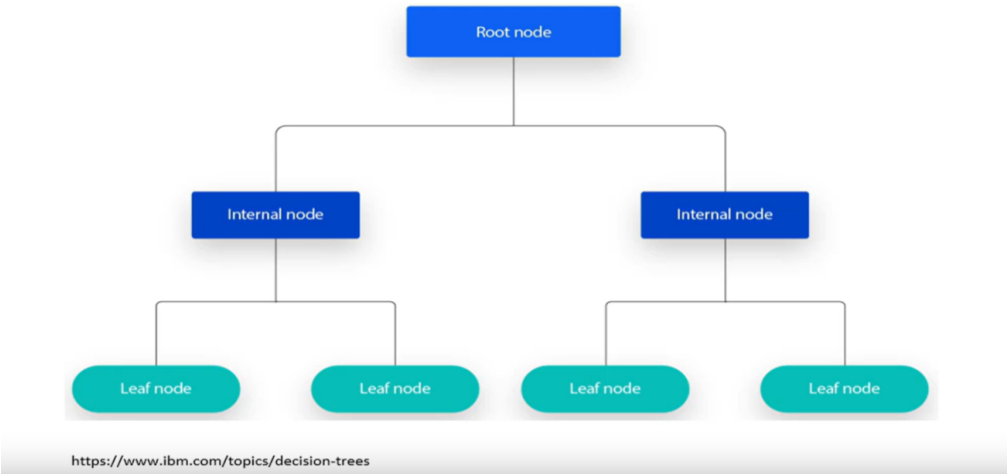
2.Karar Ağaçları

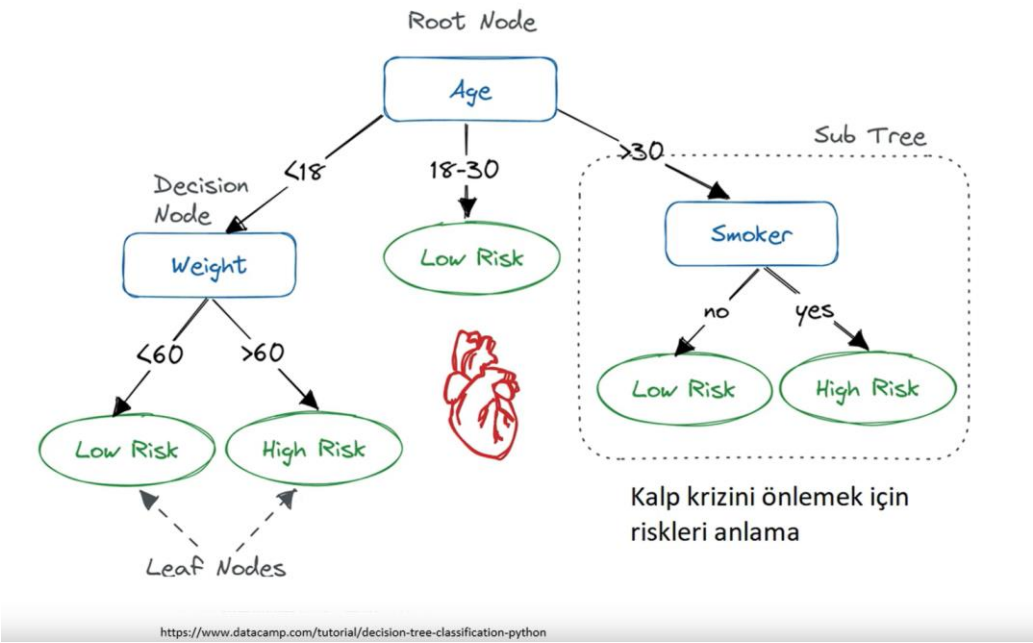
Karar ağaçları, bir ağaca benzeyen yapıda çalışır ve veriyi dallara ayırarak sonuca ulaşır.

Karar Ağaçlarının Yapısı

Karar ağaçları, hiyerarşik bir yapı kullanır. Bu yapı üç temel bileşenden oluşur:

- Kök Düğüm (Root Node)
- Dahili Düğümler (Internal Nodes)
- Yaprak Düğümler (Leaf Nodes)





Bilgi Kazancı: Bir özelliğin bölünme noktasının ne kadar iyi olduğunu belirlemek için kullanılır.

Entropi (Karışıklık): Belirli bir düğümdeki veri noktalarının sınıflarının ne kadar karışık olduğunu ölçen bir metriktir.

Karar Ağacı Türleri:

- ID3 (Iterative Dichotomiser 3):
  Bu algoritma, aday bölünmeleri değerlendirmek için entropi ve bilgi kazancını kullanır.

- C4.5 (ID3'ün gelişmiş versiyonu):
  Karar ağaçlarında bölme noktalarını değerlendirmek için bilgi kazancı veya kazanç oranlarını kullanabilir.
- CART (Classification and Regression Trees):
- Bu algoritma, genellikle ideal bölünme özelliğini belirlemek için Gini karışıklığını kullanır. Gini karışıklığı, rastgele seçilen bir özelliğin ne sıklıkla yanlış sınıflandırıldığını ölçer.

Uygulama Alanları:

- Tıbbi hastalık teşhislerinde
- Müşteri segmentasyon işlemlerinde
- Pazarlama ve satış analizinde
- Portföy yönetimi
- Endüstriyel üretim ve kalite kontrol

Karar Ağaçlarının Avantajları:

- Kolay Anlaşılabilir ve Yorumlanabilir
- Az Veri Hazırlığı
- Kategorik ve Sayısal Verilerle Çalışabilir
- Aykırı Değerlere Karşı Dayanıklı

Karar Ağaçlarının Dezavantajları:

- Aşırı Uydurma (Overfitting)
- Dengesiz Veri Setlerine Karşı Hassasiyet
- Hesaplama Maliyeti

Karar Ağaçlarının Kullanıldığı Veri Setleri

Sınıflandırma Problemleri:

- İris veri seti: Çiçek türlerini sınıflandırmak için kullanılır (setosa, versicolor, virginica). Karar ağacı, her bir çiçek örneğinin özelliklerine göre çiçek türünü belirler.
- Titanic veri seti: Yolcuların özelliklerine (yaş, cinsiyet, bilet sınıfı) göre hayatta kalma durumunu (hayatta kaldı/kalmadı) tahmin etmek için kullanılır.
- Hasta verileri: Tıbbi veri setlerinde hastaların belirli hastalıklara yakalanma olasılığını sınıflandırmak için kullanılır. Örneğin, kalp hastalığı riskini belirlemek için yaş, kan basıncı, kolesterol gibi özellikler kullanılır.

Regresyon Problemleri:

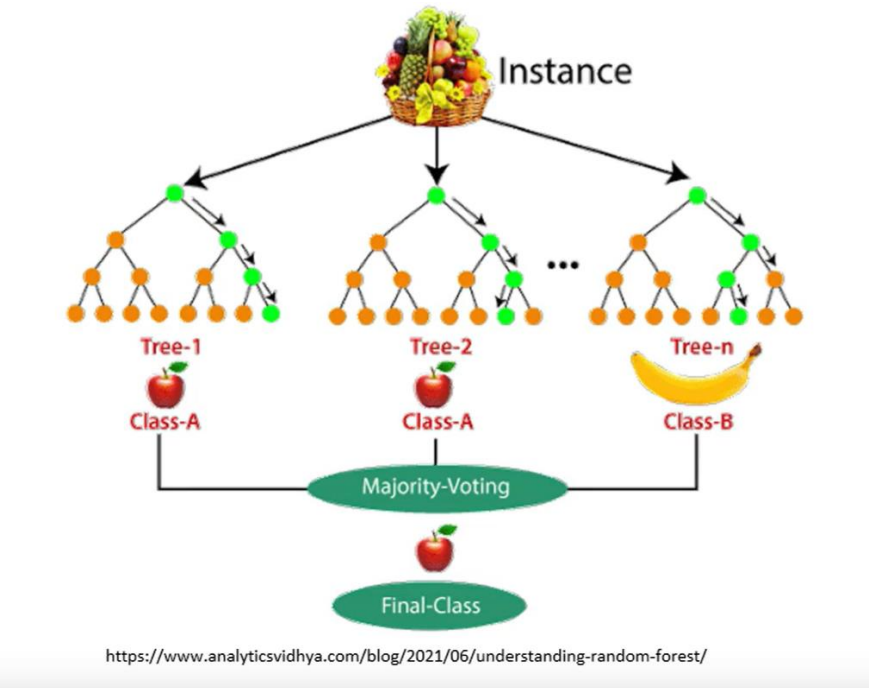
- Boston Housing veri seti: Evlerin çeşitli özelliklerine (oda sayısı, konum, arazi büyüklüğü) göre fiyatlarının tahmin edilmesi için kullanılır.
- Diğer ekonomik veri setleri: Örneğin, bir ülkedeki işsizlik oranını etkileyen faktörler (yaş, eğitim düzeyi, sanayi sektörü) analiz edilip, gelecekteki işsizlik oranı tahmin edilebilir.

Karar Ağaçlarında Kullanılan Veri Tipleri

- Numerik (Sayısal) Veriler:
  Karar ağaçları sayısal verilerle kolayca çalışabilir. Örneğin, yaş, gelir, sıcaklık gibi sürekli değerler karar ağacındaki düğümlerde kullanılabilir.
- Kategorik Veriler: Karar ağaçları kategorik verilerle de çalışabilir. Cinsiyet, renk, eğitim düzeyi gibi kategorik değişkenler, karar ağacında dallanmayı sağlar.

3. Rastgele Orman (Random Forest)

Birden fazla karar ağacından oluşan güçlü bir topluluk öğrenme (ensemble learning) algoritmasıdır. Rastgele ormanlar, karar ağaçlarının zayıf yönlerini gidermek amacıyla geliştirilmiştir. Algoritmanın temel mantığı, çok sayıda karar ağacını bir araya getirerek her ağacın sonucuna göre bir tahmin yapmaktır. Daha fazla ağaç, daha doğru tahminler anlamına gelir.



Rastgele Ormanın Avantajları:

- Aşırı Uydurmaya Karşı Dirençli
- Yüksek Doğruluk

- Dengesiz Veri Setleriyle İyi Çalışma
- Veri Hazırlığı Azdır

Rastgele Ormanın Dezavantajları:

- Hesaplama Maliyetli
- Yorumlanabilirlik Düşük
- Bellek Kullanımı Fazla

Random Forest Algoritmasının Kullanıldığı Veri Setleri

Sınıflandırma Problemleri:

- MNIST veri seti: El yazısı rakamların sınıflandırılmasında Random Forest kullanılarak yüksek doğruluk oranları elde edilebilir.
- İris veri seti: Çiçek türlerini (setosa, versicolor, virginica) sınıflandırmak için kullanılır. Karar ağaçlarının birleşimiyle türlerin doğru bir şekilde sınıflandırılması sağlanır.
- Kredi riski veri setleri: Bankacılık ve finans sektöründe, kredi başvurularını değerlendirmek için kullanılır. Özellikler arasında gelir, kredi notu, çalışma süresi gibi değişkenler bulunabilir.

Regresyon Problemleri:

- Boston Housing veri seti: Evlerin özelliklerine (bölge, oda sayısı, suç oranı) göre fiyatlarının tahmin edilmesi amacıyla kullanılır.
- Borsa veri setleri: Hisse senedi fiyatlarının tahmin edilmesinde kullanılabilir. Hisse senetlerinin geçmiş performansları ve ekonomik göstergelere dayalı olarak tahmin yapar.

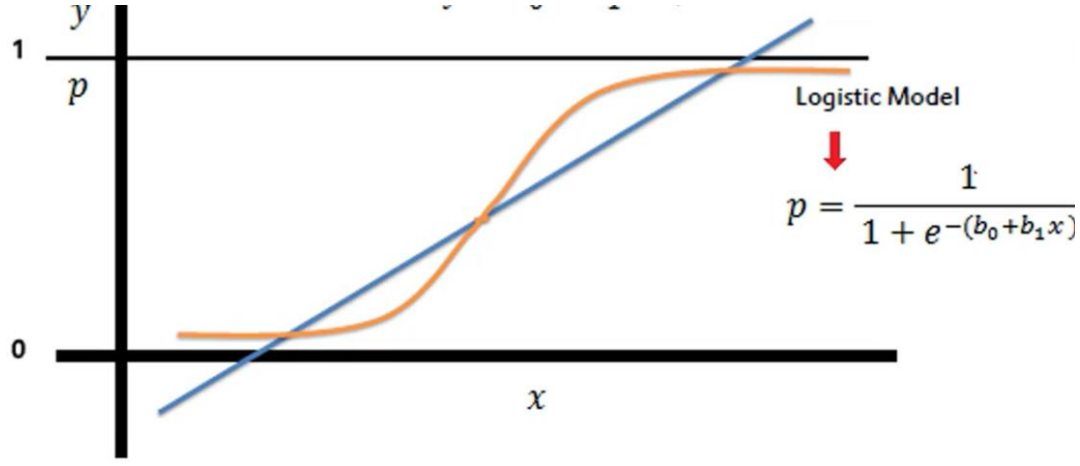
Random Forest ile Kullanılan Veri Tipleri

- Numerik Veriler: Yaş, gelir, sıcaklık, fiyat gibi sayısal veriler Random Forest algoritması için uygundur.
- Kategorik Veriler: Cinsiyet, eğitim düzeyi, meslek gibi kategorik veriler sınıflandırma problemlerinde kullanılır. Karar ağaçları gibi Random Forest da kategorik verilerle çalışabilir.
- Karmaşık Veriler: Görüntü, metin ve genetik veri gibi daha karmaşık veri türleri de, uygun ön işleme adımlarıyla Random Forest için uygun hale getirilebilir.

4.Lojistik Regresyon (Logistic Regression)

Adında "regresyon" geçmesine rağmen, aslında sınıflandırma problemlerinde kullanılan bir makine öğrenmesi algoritmasıdır. Özellikle **ikili sınıflandırma (binary classification)** problemlerinde sıklıkla tercih edilir. Lojistik regresyon, bir olayın gerçekleşme olasılığını tahmin eder ve bu tahmini 0 ile 1 arasında bir değer olarak sunar.

Lojistik regresyonun amacı, veriler arasındaki ilişkileri temel alarak bir olasılık tahmini yapmaktır. Doğrusal regresyon gibi bir modelleme yapılır, ancak doğrusal regresyon sınırsız sonuçlar üretebilir. Lojistik regresyon ise, sonuçları bir olasılık olarak almak istediğimiz için, tahmin edilen değeri 0 ile 1 arasına sıkıştırır. Bu işlem, logistik (sigmoid) fonksiyonu kullanılarak yapılır.



https://saedsayad.com/logistic_regression.htm

Uygulama Alanları:

- Sağlık(Hastalık teşhisinde)
- İnsan Kaynakları (Personel performans değerlendirmesi)
- Finans

Lojistik Regresyonun Avantajları:

- Yorumlanabilirlik
- Basitlik
- Aşırı Uydurmaya Karşı Dayanıklı
- Olasılık Tahmini
- Aykırı Değerler ve Gürültüye Karşı Dayanıklı

Lojistik Regresyonun Dezavantajları:

- Doğrusal Varsayım
- Veri Dengesizliği
- Karmaşık İlişkiler

Lojistik Regresyonun Kullanıldığı Veri Setleri

Lojistik regresyon genellikle ikili sınıflandırma problemlerinde kullanılır. Yani, iki sınıfa ayrılması gereken veri setlerinde etkilidir:

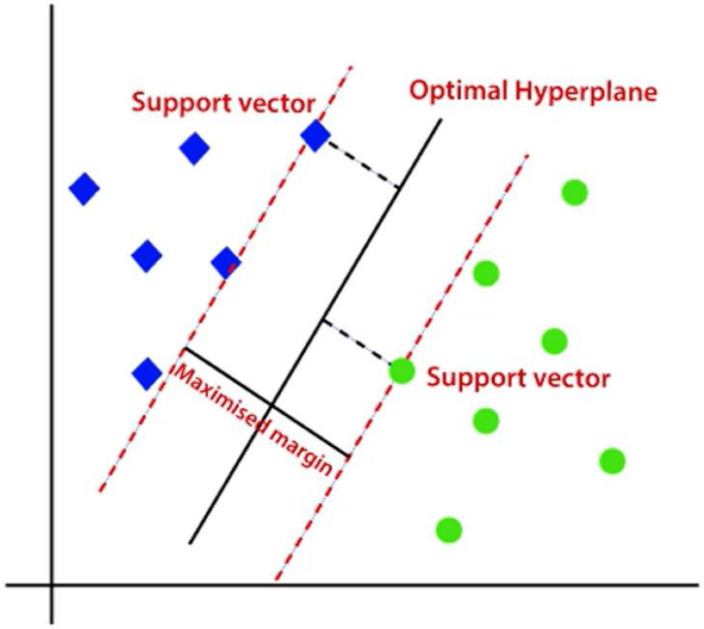
- Hastalık Tahmini-Hasta verileri: Hastaların belirli hastalıklara yakalanma olasılıklarını tahmin eder. Örneğin, yaş, kilo, kan basıncı gibi özelliklere göre bir hastanın kalp hastalığına yakalanma olasılığı hesaplanabilir.
- Müşteri Davranışı-Kredi riski veri seti: Bir müşterinin kredi başvurusunun onaylanıp onaylanmayacağını tahmin etmek için kullanılabilir. Kişinin kredi geçmişi, gelir durumu gibi bilgiler kullanılarak, bu müşterinin kredi riskine dair bir olasılık hesaplanır.
- Pazarlama ve Tıklama Tahminleri: Reklam tıklama oranları: Online reklamlara tıklayıp tıklamayacaklarını tahmin etmek için lojistik regresyon kullanılabilir. Özellikler arasında kullanıcının yaşı, cinsiyeti, ilgi alanları olabilir.
- Spam Filtreleme: E-posta sınıflandırma veri seti: E-postaların spam olup olmadığını sınıflandırmak için lojistik regresyon kullanılabilir. E-postanın uzunluğu, belirli kelimelerin varlığı gibi özellikler spam olma olasılığını belirler.

Lojistik Regresyon ile Kullanılan Veri Tipleri

- Numerik Veriler: Yaş, gelir, sıcaklık, eğitim seviyesi gibi sayısal veriler lojistik regresyonda kullanılabilir. Özellikle sürekli ve sayısal veriler, modelin doğruluğunu artırabilir.
- Kategorik Veriler: Cinsiyet, medeni durum, meslek gibi kategorik değişkenler de lojistik regresyonla işlenebilir. Bu tür veriler, genellikle one-hot encoding veya label encoding ile sayısal hale getirilir.
- Dönüştürülmüş Veriler: Metin, görüntü veya ses gibi veri tipleri, özellik mühendisliği yapılarak lojistik regresyona uygun hale getirilebilir. Örneğin, metin verileri TF-IDF veya kelime vektörleri ile sayısallaştırılabilir.

5. Destek Vektör Makineleri (SVM - Support Vector Machines)

SVM, veri noktalarını bir hiper düzlem (hyperplane) kullanarak ayırır ve bu hiper düzlem, sınıflar arasındaki farkı maksimum olacak şekilde konumlandırılır. Temel amacı, farklı sınıflara ait veri noktalarını en iyi ayıran düzlemi bulmaktır. SVM,özellikle doğrusal olmayan veri kümeleri için etkilidir.



https://www.analyticsvidhya.com/blog/2021/10/support-vector-machinessvm-a-complete-guide-for-beginners/

Uygulama Alanları:

- Tıp
- Görüntü İşleme
- Finans
- Biyoinformatik (Gen sınıflandırması)
- Metin ve Dil İşleme

SVM'nin Avantajları:

- Etkili Sınıflandırma
- Genelleme Yeteneği
- Hassasiyet

SVM'nin Dezavantajları:

- Büyük Veri Setlerinde Yavaşlık
- Hiperparametre Ayarı
- Karmaşıklık

SVM'nin Kullanıldığı Veri Setleri

İkili Sınıflandırma Problemleri:

- İris veri seti: Çiçek türlerini sınıflandırmak için kullanılır. SVM, özellikle doğrusal olarak ayrılamayan veri setlerinde kernel fonksiyonlarıyla iyi sonuçlar verir.

- Kredi riski veri seti: Bir müşterinin kredi başvurusunun onaylanıp onaylanmayacağını tahmin etmek için kullanılabilir.
- E-posta sınıflandırması: E-postaların spam olup olmadığını sınıflandırmada SVM kullanılabilir. Özellikle büyük boyutlu metin verileri için iyi sonuçlar verebilir.

Çok Sınıflı Sınıflandırma Problemleri:

- MNIST veri seti: El yazısı rakamları tanımak için kullanılır. SVM, çok sınıflı sınıflandırma problemlerinde de kernel yöntemleriyle yüksek doğruluk oranları sağlar.
- Yüz tanıma veri seti: Bir kişinin yüzünün tanınması gibi yüksek boyutlu veri setlerinde SVM kullanılarak etkili sonuçlar alınabilir.

Regresyon Problemleri:

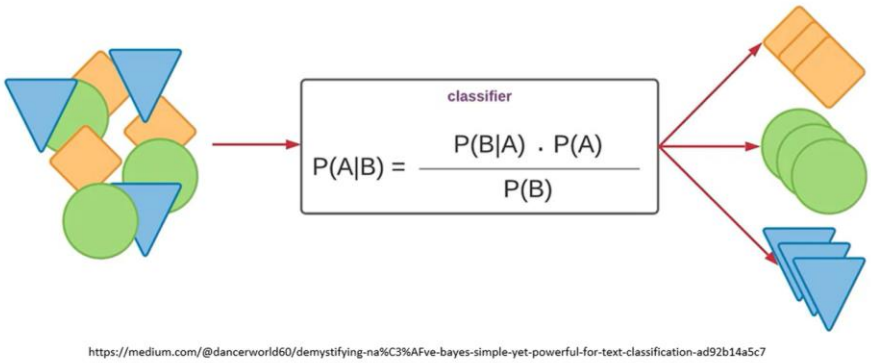
- Boston Housing veri seti: Ev fiyatlarını tahmin etmek için kullanılır. SVM'nin bir versiyonu olan Support Vector Regression (SVR), doğrusal olmayan regresyon problemleri için de kullanılabilir.

SVM ile Kullanılan Veri Tipleri

- Numerik Veriler: Sayısal verilerle çalışmak SVM için uygundur. Özellikle yüksek boyutlu veri kümelerinde SVM oldukça etkilidir.
- Kategorik Veriler: Doğrudan sayısal hale getirilebilen kategorik veriler SVM'de kullanılabilir. Kategorik veriler genellikle label encoding veya one-hot encoding ile sayısal hale getirilir.
- Metin Verileri: Özellikle metin sınıflandırma problemlerinde, metin verileri sayısal vektörlere dönüştürüldükten sonra SVM algoritması başarılı sonuçlar verir. Örneğin, spam e-posta tespiti gibi problemler için uygundur

6.Naive Bayes:

Olasılık temelli bir sınıflandırma yöntemidir ve genellikle metin sınıflandırma, spam tespiti, duygu analizi gibi uygulamalarda kullanılır. Naive Bayes'in temelini, Bayes Teoremi ve özelliklerin bağımsız olduğu varsayımı (naive - saf varsayım) oluşturur. Adından da anlaşılacağı gibi, "naive" olarak her özelliğin birbirinden bağımsız olduğunu varsayar ve bu basit varsayıma dayanarak çalışır. Gerçekte bu varsayım genellikle geçerli olmasa da, Naive Bayes pratikte oldukça başarılı sonuçlar verir.



https://medium.com/@dancerworld60/demystifying-na%C3%AFve-bayes-simple-yet-powerful-for-text-classification-ad92b14a5c7

Uygulama Alanları:

- Tıp
- Pazarlama
- Biyoinformatik
- Metin Sınıflandırma
- Duygu Analizi

Naive Bayes'in Avantajları:

- Basit ve Hızlı
- İyi Genelleme Yeteneği
- Düşük Hafıza Kullanımı
- Sınıflandırma Performansı
- Çoklu Sınıflar

Naive Bayes'in Dezavantajları:

- Bağımsızlık Varsayımı
- Veri Dengesizliği

Naive Bayes'in Kullanıldığı Veri Setleri

- Metin Sınıflandırma:

Spam Filtreleme: E-postaların spam olup olmadığını tespit etmek için Naive Bayes sıklıkla kullanılır. E-postadaki kelimelerin sıklığına göre spam olup olmadığına karar verilir.

Haber Kategorisi Tespiti: Bir haberin hangi kategoriye (örneğin, spor, politika, ekonomi) ait olduğunu sınıflandırmak için Naive Bayes kullanılabilir.

- Doküman Sınıflandırma:

Sentiment Analizi: Bir dokümanın (örneğin, bir film eleştirisi) olumlu mu yoksa olumsuz mu olduğunu belirlemek için kullanılabilir. Naive Bayes, kelimelerin sıklığına dayanarak sınıflandırma yapar.

- Tıbbi Teşhis:

Hastalık Teşhisi: Naive Bayes, belirli bir hastalığın belirtilerini göz önünde bulundurarak hastalık teşhisi yapabilir. Örneğin, belirli bir yaş, cinsiyet ve semptomlar verildiğinde, kişinin belirli bir hastalığa sahip olma olasılığı hesaplanır.

- Müşteri Segmentasyonu:

Pazarlama Kampanyaları: Müşterilerin belirli ürünlere olan ilgisini tahmin etmek ve onları farklı segmentlere ayırmak için Naive Bayes kullanılabilir. Bu, satın alma davranışlarını tahmin etmekte etkili olabilir.

Naive Bayes ile Kullanılan Veri Tipleri

- Metin Verileri: Özellikle doğal dil işleme (NLP) problemlerinde çok yaygın olarak kullanılır. Metin verileri sayısal hale getirilip kelime sıklıkları veya var/yok bilgileri olarak temsil edilir .
- Sayısal Veriler: Gaussian Naive Bayes, sürekli sayısal verilere uygun bir şekilde çalışır. Örneğin, yaş, kilo, gelir gibi sürekli özellikler bu algoritmayla işlenebilir.
- Ayrık Veriler: Multinomial Naive Bayes, ayrık veri türleriyle iyi çalışır. Kelime sayıları, ürün kategorileri veya müşteri segmentleri gibi ayrık veriler bu algoritmada kullanılabilir.

Ödev Soruları:

1-Cross Validation (Çapraz Doğrulama), makine öğrenmesinde bir modelin genelleme yeteneğini test etmek için kullanılan bir yöntemdir. En yaygın türü k-katlı çapraz doğrulamadır. Bu yöntemde veri kümesi k alt kümeye bölünür; her seferinde bir alt küme test verisi, diğerleri ise eğitim verisi olarak kullanılır. İşlem k kez tekrarlanarak sonuçlar ortalanır.

Faydaları:

Aşırı uyumu (overfitting) önler.

Daha doğru model değerlendirmesi sağlar.

Veri kullanımını artırır.

Bu sayede, modelin genel performansı daha güvenilir bir şekilde ölçülür.

2- Label çeşidine göre sınıflandırma metodu değişir mi?
2- Evet, etiket çeşidine göre sınıflandırma metodu değişir. Makine öğrenmesinde etiketler, modelin öğrenme sürecini etkiler.

3-Hangi labelde hangi sınıflandırma kullanılır?
3- Farklı etiket türleri, belirli sınıflandırma yöntemleriyle ilişkilidir ve seçilecek yöntem, problem türüne ve veri yapısına göre değişir:
1. İkili Sınıflandırma (Binary Classification)
İki sınıfa (örneğin, olumlu/olumsuz) ayrılan etiketler.
Kullanılan Yöntemler:

- Lojistik Regresyon
- Destek Vektör Makineleri (SVM)
- Karar Ağaçları
- Random Forest
- Naive Bayes
- Sinir Ağları

2. Çoklu Sınıflandırma (Multi-class Classification)
Üç veya daha fazla sınıfa ayrılan etiketler.
Kullanılan Yöntemler:

- Karar Ağaçları
- Random Forest
- Destek Vektör Makineleri (SVM) (one-vs-all veya one-vs-one)
- Softmax Regresyon
- Sinir Ağları (Çok katmanlı yapılarla)
- K-Nearest Neighbors (KNN)

3. Çok Etiketli Sınıflandırma (Multi-label Classification)
Her örnek birden fazla etiket alabilir.
Kullanılan Yöntemler:

- KNN
- Random Forest
- Destek Vektör Makineleri (SVM)
- Sinir Ağları
- Binary Relevance (her etiket için ayrı bir sınıflandırıcı kullanma)

4. Regresyon Problemleri (Regression Problems)
Sürekli bir değer tahmin edilir; genellikle etiket yoktur ama etiketler sürekli değerlerdir.
Kullanılan Yöntemler:

- Lineer Regresyon
- Ridge/Lasso Regresyon
- Destek Vektör Regresyonu (SVR)
- Karar Ağaçları (regresyon için)
- Sinir Ağları (regresyon için)

5. Etiketsiz Öğrenme (Unsupervised Learning)
Etiket yok, verilerdeki kalıpları keşfetmek amaçlanır.
Kullanılan Yöntemler:

- K-Means
- Hiyerarşik Kümeleme
- DBSCAN
- PCA (Principal Component Analysis)

4- Sınıf (class) çeşidine göre (nominal veya numeric) sınıflandırma metodu değişir mi ?

4- Evet, sınıf (class) çeşidine göre (nominal veya numeric) sınıflandırma metodu değişir.

> **Nominal sınıflar:** Kategorik verilere dayalı sınıflandırma yöntemleri (lojistik regresyon, karar ağaçları, vb.) kullanılır.
>
> **Numeric sınıflar:** Sürekli değerleri tahmin eden regresyon yöntemleri (lineer regresyon, SVR, vb.) kullanılır.

5- KNN sınıflandırma ve KNN kümeleme arasındaki fark ve benzerlikler nelerdir?
5- KNN (K-Nearest Neighbors), hem sınıflandırma hem de kümeleme (clustering) için kullanılan bir algoritmadır. Ancak bu iki kullanım arasında bazı temel farklar ve benzerlikler bulunmaktadır:

Benzerlikler

- Temel Prensip: Her iki yöntem de komşuluk ilişkisine dayanır. Yani, bir nesnenin sınıfını veya kümesini belirlemek için en yakın komşularının bilgilerini kullanır.
- Mesafe Ölçümü: Her iki yaklaşımda da benzerlik veya uzaklık ölçümünde genellikle Euclidean mesafesi gibi mesafe ölçütleri kullanılır.
- Veri Yapısı: Her ikisi de veri noktalarını n-boyutlu bir uzayda işler, bu da yüksek boyutlu verilerle çalışmayı mümkün kılar.
- Non-parametrik: Hem KNN sınıflandırma hem de KNN kümeleme, belirli bir dağılım varsayımına dayanmaz (non-parametrik yöntemlerdir).

Farklar

- Amaç:

KNN Sınıflandırma: Veri noktalarını belirli sınıflara ayırmak için kullanılır. Yani, verilen bir veri noktasının hangi sınıfa ait olduğunu belirlemeye çalışır.

KNN Kümeleme: Veri noktalarını gruplamak için kullanılır. Yani, benzer veri noktalarını bir araya getirerek küme oluşturur.

- Sonuç:

KNN Sınıflandırma: Her bir veri noktasına bir etiket (sınıf) atar.
KNN Kümeleme: Veri noktalarını etiketlemez, sadece benzerliklerine göre gruplar (kümeler).

- Uygulama Alanları:

KNN Sınıflandırma: Sınıflandırma problemleri (örneğin, e-posta spam tespiti, hastalık teşhisi gibi).
KNN Kümeleme: Veri keşfi ve segmentasyon (örneğin, müşteri segmentasyonu, pazar analizi gibi).

- Girdi ve Çıktı:

KNN Sınıflandırma: Girdisi etiketli veriler, çıktısı sınıf etiketleridir.
KNN Kümeleme: Girdisi genellikle etiketlenmemiş veriler, çıktısı küme gruplarıdır.