Destek Vektör Makineleri (SVM)

Sınıflandırma Nedir?

- **Sınıflandırma**: Etiketli verilerle, örneklerin önceden belirlenmiş sınıflara ayrılmasıdır.
- **Binary (İkili) Sınıflandırma**: İki sınıf arasında ayrım yapılır (örn. "spam" vs "spam değil").
- **Çoklu Sınıflandırma**: Üç ya da daha fazla sınıf bulunur (örn. meyve türleri).

SVM (Destek Vektör Makineleri)

- **Amacı**: Verileri en büyük **marjin** ile ayıran bir **hiperdüzlem** bulmak.
- **Doğrusal Sınıflandırma**: Veriler doğrusal olarak ayrılabiliyorsa, basit bir hiperdüzlem ile çözülür.
- **Doğrusal Olmayan Sınıflandırma**: Veriler doğrusal değilse, **kernel fonksiyonları** kullanılarak veriyi daha yüksek boyutlu uzaya dönüştürür.

Kernel Fonksiyonları

- **Kernel**: Veri noktaları arasındaki benzerliği hesaplayan fonksiyonlardır.
- Yaygın Kernel Türleri:
    - **Doğrusal Kernel**: Doğrusal sınıflandırmalar için kullanılır.
    - **RBF Kernel (Radial Basis Function)**: Doğrusal olmayan verilerle iyi çalışır.
    - **Polinomsal Kernel**: Daha az tercih edilir, verimliliği düşüktür.

SVM Hiperparametreleri

- **C Parametresi**:
    - **Büyük C**: Daha az marjin hatası, overfitting riski.
    - **Küçük C**: Daha fazla tolerans, underfitting riski.
- **Gamma** (RBF için): Modelin karmaşıklığını ayarlar.

Avantajlar

- **Yüksek Boyutlu Verilerle İyi Performans**: Özellikle çok boyutlu uzaylarda etkilidir.
- **Aykırı Değerlere Dayanıklıdır**.
- **Küçük Veri Kümesinde Başarılı**: Az veriyle yüksek doğruluk elde edilebilir.

Dezavantajlar

- **Büyük Veri Kümesinde Yavaşlık**: Hesaplama maliyeti yüksektir.
- **Hiperparametre Ayarı Zor**: C ve kernel seçimleri genellikle zor ve zaman alır.
- **Dengesiz Veri Kümesinde Zayıf Performans**.

One-vs-Rest (OvR) vs One-vs-One (OvO)

- **One-vs-Rest (OvR)**: Her sınıf için bir ikili model eğitilir. Yüksek veri boyutlarında etkili, fakat dengesiz veriyle sorun yaşayabilir.
- **One-vs-One (OvO)**: Her sınıf çifti için ayrı bir model eğitilir. Dengesiz veri kümelerinde daha iyi çalışır, ancak hesaplama maliyeti yüksektir.

Kernel Seçimi

- **Doğrusal Kernel**: Veriler doğrusal olarak ayrılabiliyorsa kullanılır.
- **RBF Kernel**: Veriler doğrusal değilse, karmaşık sınıflandırmalar için uygundur

Soru 1:

SVM algoritmasının temel amacı nedir?

- A) Verileri her sınıfa en yakın merkeze yerleştirmek
- B) Veriler arasındaki benzerliği en yüksek olan noktaları bulmak
- C) Verileri en büyük marjin ile ayıran hiperdüzlemi bulmak
- D) Verilerin her birini farklı kümeleme alanlarına yerleştirmek
- E) En küçük doğrusal hata ile verileri sınıflandırmak

Doğru Cevap: C

- **Açıklama**: SVM, verileri en büyük **marjin** ile ayıran bir **hiperdüzlem** bulmayı amaçlar. Bu marjin, sınıflar arasında en büyük mesafe ile ayrılmayı sağlar.

Soru 2:

Aşağıdakilerden hangisi SVM'nin **kernel** fonksiyonları ile ilgili doğru bir açıklamadır?

- A) Kernel fonksiyonları sadece doğrusal sınıflandırmalar için kullanılır.
- B) Kernel fonksiyonları, veriyi daha yüksek boyutlu bir uzaya dönüştürerek doğrusal olmayan verileri ayırmak için kullanılır.
- C) Kernel fonksiyonları, sadece doğrusal olmayan sınıflandırmalar için uygundur.

- D) Kernel fonksiyonları sadece dengesiz veri kümeleri için uygundur.
- E) Kernel fonksiyonları, yalnızca modelin eğitim sürecinde kullanılır.

Doğru Cevap: B

- **Açıklama**: Kernel fonksiyonları, doğrusal olmayan verileri daha yüksek boyutlu bir uzaya dönüştürmek için kullanılır ve böylece doğrusal olmayan verilerin ayrılmasına imkan tanır.

Soru 3:

Aşağıdaki veri kümelerinden hangisi için doğrusal bir SVM modelinin başarılı bir şekilde uygulanması beklenir?

- A) Karmaşık, çok katmanlı bir sınıflandırma problemi
- B) Veriler arasında doğrusal olmayan bir ilişki bulunan küme
- C) Veriler doğrusal olarak ayrılabiliyorsa
- D) Verilerde büyük aykırı değerler bulunduğunda
- E) Veri kümeleri dengesiz olduğunda

Doğru Cevap: C

- **Açıklama**: Doğrusal bir SVM modeli, **doğrusal** olarak ayrılabilen veri kümelerinde etkili çalışır. Eğer veriler doğrusal olarak ayrılabiliyorsa, doğrusal hiperdüzlem ile başarılı sınıflandırma yapılabilir.

Soru 4:

"One-vs-Rest" (OvR) stratejisi hangi durumda kullanılır?

- A) Verilen her sınıf için bir ikili sınıflandırma modeli eğitilir.
- B) Her sınıf, diğer sınıflarla karşılaştırılarak bir model oluşturulur.
- C) Tüm sınıflar için tek bir model eğitilir.
- D) Her sınıf çifti için ayrı bir model eğitilir.
- E) En yakın komşu sınıfı kullanılır.

Doğru Cevap: A

- **Açıklama**: One-vs-Rest (OvR) stratejisinde, her sınıf için ayrı bir ikili sınıflandırma modeli eğitilir. Her model, o sınıfı pozitif, diğerlerini negatif olarak değerlendirir.

Soru 5:

Aşağıdakilerden hangisi **SVM'nin avantajları** arasında yer almaz?

- A) Yüksek boyutlu verilerde etkili çalışır.
- B) Aykırı değerlere karşı dayanıklıdır.
- C) Küçük veri kümelerinde iyi sonuçlar verir.
- D) Hesaplama maliyeti çok düşüktür.
- E) Doğrusal olmayan verilerle başa çıkabilir.

Doğru Cevap: D

- **Açıklama**: SVM, özellikle büyük veri kümelerinde hesaplama maliyetinin yüksek olabileceği bir algoritmadır. Bu nedenle hesaplama maliyetinin çok düşük olduğu söylenemez.

Soru 6:

Aşağıdaki kernel fonksiyonlarından hangisi, doğrusal olmayan verilerin daha yüksek boyutlu bir uzaya dönüştürülmesinde yaygın olarak kullanılır?

- A) Doğrusal Kernel
- B) RBF (Radial Basis Function) Kernel
- C) Polinomsal Kernel
- D) Bessel Kernel
- E) Anova Kernel

Doğru Cevap: B

- **Açıklama**: **RBF kernel**, doğrusal olmayan verileri daha yüksek boyutlu bir uzaya dönüştürmek için yaygın olarak kullanılır ve karmaşık sınıflandırma problemleriyle başa çıkmada etkilidir.

Soru 7:

SVM modelinde **C parametresi** ile ilgili aşağıdaki ifadelerden hangisi doğrudur?

- A) Küçük C, daha az hata ile sınıflandırma yapar ancak overfitting riskini artırır.
- B) Büyük C, daha fazla hata ile sınıflandırma yapar ancak overfitting'i engeller.

- C) Küçük C, daha fazla hata ile sınıflandırma yapar ancak underfitting'i engeller.
- D) Büyük C, daha az marjin hatası sağlar ancak overfitting riskini artırır.
- E) C parametresi modelin doğruluğunu etkilemez.

Doğru Cevap: D

- **Açıklama**: **Büyük C**, marjin hatalarını azaltmaya çalışır ancak bu, **overfitting**'e yol açabilir. Küçük C ise daha fazla hata kabul eder ve genellikle **underfitting**'e yol açar.

Soru 8:

Aşağıdaki durumlardan hangisinde **One-vs-One** (OvO) stratejisi daha avantajlıdır?

- A) Düşük boyutlu veri setlerinde
- B) Veri kümeleri çok dengesiz olduğunda
- C) Çok sayıda model oluşturmak gereksizse
- D) Sınıflar arasındaki ayrım netse
- E) Her modelin yalnızca iki sınıfı dikkate alması gerektiğinde

Doğru Cevap: E

- **Açıklama**: **One-vs-One** stratejisinde her model yalnızca iki sınıfı dikkate alır, bu da daha iyi ayrım sağlar. Ancak, çok sayıda model gerektiği için hesaplama maliyeti artar.

Soru 9:

Aşağıdaki kernel türlerinden hangisi, genellikle daha verimli olmayan ve çok fazla tercih edilmeyen bir türdür?

- A) Doğrusal Kernel
- B) RBF Kernel
- C) Polinomsal Kernel
- D) Bessel Kernel
- E) Anova Kernel

Doğru Cevap: C

- **Açıklama**: **Polinomsal kernel**, diğer kernel türlerine göre daha az verimli olabilir ve genellikle yüksek hesaplama maliyetine neden olur. Bu yüzden daha az tercih edilir.

Soru 10:

SVM'nin **dezavantajlarından** biri nedir?

- A) Sınıflar arasındaki ayrım her zaman net olur.
- B) Yüksek boyutlu verilerle başa çıkmakta zorlanır.
- C) Büyük veri kümesi ile çalışırken eğitim süresi uzun olabilir.
- D) Aykırı değerlere karşı duyarlı değildir.
- E) Verilerin doğrusal olarak ayrılabilir olduğu her durumda mükemmel sonuçlar verir.

Doğru Cevap: C

- **Açıklama**: SVM, **büyük veri kümelerinde** eğitim süreci oldukça **yavaş** olabilir ve hesaplama maliyeti yüksektir. Bu da pratikte bir dezavantaj oluşturur.

KNN (K-En Yakın Komşu) Algoritması

Tanım:

- KNN, denetimli öğrenme algoritmalarından biridir.
- Hem sınıflandırma hem de regresyon problemleri için kullanılabilir.
- Temel prensip: Yeni bir veri noktasının sınıfı, en yakın "K" komşusunun çoğunluğuna göre belirlenir.

Çalışma Prensibi:

- Veriler arasındaki mesafeler hesaplanarak, yeni bir veri noktasının hangi sınıfa ait olduğu bulunur.
- Regresyonda, komşuların değerlerinin ortalaması alınır.

Avantajlar:

- **Basitlik:** Anlaşılması ve uygulanması kolaydır.
- **Parametrik Olmayan Yapı:** Veri dağılımı ile ilgili varsayım yapmaz.
- **Esneklik:** Hem sınıflandırma hem de regresyon için uygundur.

Dezavantajlar:

- **Hız Sorunu:** Büyük veri setlerinde yavaş çalışabilir.
- **Hafıza Tüketimi:** Tüm eğitim verisini saklar.
- **Boyut Laneti:** Yüksek boyutlu verilerde performans düşer.
- **Özellik Ölçeklendirme Gerekliliği:** Özellikler aynı ölçek üzerinde olmalıdır.

Mesafe Ölçümleri:

1. **Öklid Uzaklığı:** Sayısal veriler için en yaygın kullanılan mesafe ölçümüdür.
2. **Manhattan Uzaklığı:** Dikdörtgen düzlemdeki mesafeleri ölçer.
3. **Minkowski Uzaklığı:** Öklid ve Manhattan'ı genelleştirir; parametre "p" ile özelleştirilebilir.
4. **Hamming Mesafesi:** İkili veya kategorik veriler için uygundur, farklı bit sayısını ölçer.
5. **Mahalanobis Uzaklığı:** Korelasyon ve varyansı dikkate alır; yüksek boyutlu verilerde kullanılır.

Doğru "K" Değeri:

- **K Küçükse:** Aşırı öğrenme (overfitting) riski vardır.

- **K Büyükse:** Model daha genel olur ve ayrıntıları kaçırabilir.
- **Çapraz doğrulama** ile en iyi K değeri seçilmelidir.

Özellik Ölçeklendirme:

- KNN, mesafelere dayalı bir algoritma olduğu için, özelliklerin aynı ölçekte olması önemlidir. Aksi halde, büyük ölçekli özellikler diğerlerini domine edebilir.

Soru 1:
KNN algoritmasının temel prensibi nedir?
a) Veri kümesindeki en uzak komşuyu dikkate alır
b) En yakın "K" komşusunun etiketlerine göre sınıflandırma yapar
c) Veri kümesindeki her nokta için rastgele etiketler atar
d) Verilerin sırasını dikkate alarak karar verir
Doğru Cevap: b) En yakın "K" komşusunun etiketlerine göre sınıflandırma yapar
Açıklama: KNN algoritması, yeni bir veri noktasının sınıfını en yakın "K" komşusunun çoğunluğuna göre belirler.

Soru 2:
KNN algoritması hangi tür öğrenme problemleri için kullanılabilir?
a) Sadece sınıflandırma
b) Sadece regresyon
c) Hem sınıflandırma hem de regresyon
d) Sadece kümeleme
Doğru Cevap: c) Hem sınıflandırma hem de regresyon
Açıklama: KNN, hem sınıflandırma hem de regresyon problemleri için kullanılabilir.

Soru 3:
KNN algoritmasının hangi avantajı vardır?
a) Verilerin sırasını dikkate alır
b) Parametrik olmayan bir yapıya sahiptir
c) Hızlı çalışır ve düşük bellek tüketir
d) Yüksek boyutlu verilerde iyi performans gösterir
Doğru Cevap: b) Parametrik olmayan bir yapıya sahiptir
Açıklama: KNN, parametrik olmayan bir algoritmadır ve veri dağılımı hakkında herhangi bir varsayımda bulunmaz.

Soru 4:
KNN algoritması hangi durumda performans sorunları yaşayabilir?

a) Küçük veri kümesi ile çalışırken
b) Boyut laneti nedeniyle yüksek boyutlu verilerde
c) Özellikler ölçeklendirilmişse
d) Veriler birbirine uzak olduğunda
Doğru Cevap: b) Boyut laneti nedeniyle yüksek boyutlu verilerde
Açıklama: Yüksek boyutlu verilerde, veriler arasındaki mesafeler birbirine yakın hale gelir ve KNN'nin performansı düşer.

Soru 5:
KNN algoritmasında mesafelerin hesaplanmasında hangi faktör önemlidir?
a) Özelliklerin aynı ölçekte olması
b) Verilerin sıralanması
c) Veri noktalarının ortalaması
d) Veri kümesinin büyüklüğü
Doğru Cevap: a) Özelliklerin aynı ölçekte olması
Açıklama: KNN, mesafelere dayalı bir algoritma olduğu için, özelliklerin aynı ölçekte olması gereklidir.

Soru 6:
Aşağıdaki mesafe ölçümlerinden hangisi, sayısal veriler için en yaygın olarak kullanılır?
a) Hamming Mesafesi
b) Mahalanobis Uzaklığı
c) Manhattan Uzaklığı
d) Öklid Uzaklığı
Doğru Cevap: d) Öklid Uzaklığı
Açıklama: Öklid uzaklığı, genellikle sayısal veriler için kullanılan en yaygın mesafe ölçüsüdür.

Soru 7:
KNN algoritmasında "K" parametresi nasıl seçilmelidir?
a) Küçük K değeri aşırı öğrenmeye yol açar, büyük K değeri ise modelin ayrıntıları kaçırmasına sebep olur
b) K değeri her zaman 1 olmalıdır
c) K değeri veri kümesinin boyutuna göre otomatik seçilir
d) K değeri her zaman büyük olmalıdır
Doğru Cevap: a) Küçük K değeri aşırı öğrenmeye yol açar, büyük K değeri ise modelin ayrıntıları kaçırmasına sebep olur
Açıklama: Küçük K, aşırı öğrenmeye (overfitting) neden olabilirken, büyük K, modelin

yeterince ayrıntılı öğrenmesini engelleyebilir. Doğru K değeri çapraz doğrulama ile seçilmelidir.

Soru 8:
KNN algoritmasında Manhattan uzaklığı hangi tür veri için daha uygundur?
a) Kategorik veriler
b) Sürekli sayısal veriler
c) Yüksek boyutlu veriler
d) İkili (binary) veriler
Doğru Cevap: b) Sürekli sayısal veriler
Açıklama: Manhattan uzaklığı, özellikle büyük farklılıkları olan veri setlerinde daha uygun bir mesafe ölçümüdür.

Soru 9:
KNN algoritmasında hangi mesafe ölçüm yöntemi ikili (binary) veriler için uygundur?
a) Öklid Uzaklığı
b) Mahalanobis Uzaklığı
c) Minkowski Uzaklığı
d) Hamming Mesafesi
Doğru Cevap: d) Hamming Mesafesi
Açıklama: Hamming mesafesi, ikili (binary) verilerde farklı bit sayısını ölçen bir mesafe ölçümüdür.

Soru 10:
KNN algoritmasında **Mahalanobis uzaklığı** hangi durumu dikkate alır?
a) Verilerin sırasını dikkate alır
b) Veri noktalarının korelasyonunu ve varyansını dikkate alır
c) Her veriyi aynı şekilde değerlendirir
d) Verilerin ortalamalarını dikkate alır
Doğru Cevap: b) Veri noktalarının korelasyonunu ve varyansını dikkate alır
Açıklama: Mahalanobis uzaklığı, özellikle yüksek boyutlu ve korelasyonlu verilerde kullanılır ve veri noktalarının korelasyonunu ve varyansını dikkate alır.

Soru 1:
KNN algoritması hangi tür öğrenme yöntemine aittir?
a) Denetimsiz Öğrenme
b) Denetimli Öğrenme

c) Pekiştirmeli Öğrenme
d) Derin Öğrenme
e) Sürükleyici Öğrenme
Doğru Cevap: b) Denetimli Öğrenme
Açıklama: KNN, etiketli veri kullanarak yeni verilerin sınıflandırılması için kullanılan denetimli bir öğrenme algoritmasıdır.

Soru 2:
KNN algoritması hangi tür problemlerde kullanılabilir?
a) Sadece sınıflandırma
b) Sadece regresyon
c) Hem sınıflandırma hem de regresyon
d) Zaman serisi analizi
e) Sadece kümeleme
Doğru Cevap: c) Hem sınıflandırma hem de regresyon
Açıklama: KNN, hem sınıflandırma hem de regresyon problemlerinde kullanılabilen bir algoritmadır.

Soru 3:
KNN algoritmasında "K" değeri neyi ifade eder?
a) Eğitim verisinin boyutunu
b) En yakın komşu sayısını
c) Verinin doğruluğunu
d) Eğitim veri setinin çeşidini
e) Test veri setinin sayısını
Doğru Cevap: b) En yakın komşu sayısını
Açıklama: "K", yeni bir veri noktasının sınıflandırılmasında dikkate alınacak en yakın komşu sayısını ifade eder.

Soru 4:
KNN algoritmasının "Boyut Laneti" problemi hangi durumla ilgilidir?
a) Çok sayıda veri noktasının olması
b) Verinin boyutunun artması ile mesafelerin birbirine yakın hale gelmesi
c) Özelliklerin eksik olması
d) Eğitim verisinin olmaması
e) Modelin hiperparametrelerinin yanlış seçilmesi
Doğru Cevap: b) Verinin boyutunun artması ile mesafelerin birbirine yakın hale gelmesi

Açıklama: Boyut laneti, veri boyutunun artmasıyla, özellikler arasındaki mesafelerin anlamlı farklar yaratmaması durumunu ifade eder.

Soru 5:
KNN algoritmasında mesafe ölçüm yöntemlerinden hangisi genellikle sayısal veriler için kullanılır?
a) Hamming Mesafesi
b) Manhattan Uzaklığı
c) Mahalanobis Uzaklığı
d) Minkowski Uzaklığı
e) Öklid Uzaklığı
Doğru Cevap: e) Öklid Uzaklığı
Açıklama: Öklid uzaklığı, özellikle sayısal veriler için yaygın olarak kullanılan bir mesafe ölçümüdür.

Soru 6:
KNN algoritmasında hangi mesafe ölçüm yöntemi ikili veriler için uygundur?
a) Mahalanobis Uzaklığı
b) Hamming Mesafesi
c) Minkowski Uzaklığı
d) Manhattan Uzaklığı
e) Öklid Uzaklığı
Doğru Cevap: b) Hamming Mesafesi
Açıklama: Hamming mesafesi, ikili ya da kategorik veriler için uygundur ve iki veri noktası arasındaki farklı bit sayısını ölçer.

Soru 7:
KNN algoritmasında "K" değeri çok küçük seçildiğinde ne olabilir?
a) Model çok genelleştirilmiş olur
b) Model aşırı öğrenme (overfitting) riski taşır
c) Model yavaş çalışır
d) Model doğru sonuç vermez
e) Model daha hızlı çalışır
Doğru Cevap: b) Model aşırı öğrenme (overfitting) riski taşır
Açıklama: Küçük "K" değerleri, modelin aşırı öğrenmesine (overfitting) ve eğitim verisine fazla uyum sağlamasına yol açabilir.

Soru 8:
KNN algoritmasında mesafe ölçüm yöntemlerinden hangisi, verilerin varyansını dikkate alır?
a) Mahalanobis Uzaklığı
b) Manhattan Uzaklığı
c) Öklid Uzaklığı
d) Minkowski Uzaklığı
e) Hamming Mesafesi
Doğru Cevap: a) Mahalanobis Uzaklığı
Açıklama: Mahalanobis uzaklığı, veri noktalarının korelasyonunu ve varyansını dikkate alarak mesafeyi hesaplar.

Soru 9:
KNN algoritmasında doğru "K" değerini nasıl seçebilirsiniz?
a) Rastgele seçilir
b) Ağırlıklı ortalama ile belirlenir
c) Çapraz doğrulama ile belirlenir
d) Eğitim verisiyle optimize edilir
e) Modelin hızına göre seçilir
Doğru Cevap: c) Çapraz doğrulama ile belirlenir
Açıklama: Doğru "K" değeri çapraz doğrulama ile seçilir. Farklı "K" değerleri denenerek en iyi performans sağlanır.

Soru 10:
KNN algoritmasında, hangi durumda özelliklerin ölçeklendirilmesi gereklidir?
a) Yalnızca büyük veri setlerinde
b) Kategorik verilerde
c) Özellikler farklı ölçeklere sahipse
d) Yalnızca regresyon problemlerinde
e) Yalnızca küçük veri setlerinde
Doğru Cevap: c) Özellikler farklı ölçeklere sahipse
Açıklama: KNN, mesafelere dayalı bir algoritmadır, bu yüzden özellikler aynı ölçek üzerinde olmalıdır. Aksi takdirde, daha büyük ölçekli özellikler diğerlerini domine edebilir.

Karar Ağaçları - Kısa ve Öğretici Notlar

1. Karar Ağaçları Nedir?

- **Tanım**: Karar ağaçları, sınıflandırma ve regresyon problemlerinde kullanılan güçlü bir algoritmadır.
- **Kullanım Alanı**: Veri kümesi karmaşık değilse, anlaşılabilirlik, yorumlanabilirlik ve hızlı prototipleme gerektiğinde tercih edilir.
- **Görsel Temsil**: Ağacın kök, dallar ve yaprak düğümleri gibi bileşenlerden oluşur.

2. Karar Ağacının Yapısı

- **Kök Düğüm (Root Node)**: Veriyi ilk kez bölen, en önemli özelliği içeren düğüm. Kök düğüm seçiminde bilgi kazancı veya Gini indeksi gibi ölçütler kullanılır.
- **Dallar (Branches)**: Kökten veya iç düğümlerden çıkan, veriyi alt gruplara ayıran yollar.
- **Yaprak Düğümler (Leaf Nodes)**: Sonuçların verildiği son düğümler. Sınıflandırma için etiketler, regresyon için tahmin edilen değerler içerir.

3. Ölçütler

- **Bilgi Kazancı (Information Gain)**: Entropi ile ölçülür; düşük entropi, verinin daha homojen olduğunu gösterir.
- **Gini İndeksi**: Veri kümesindeki saflığı ölçer. Homojen veri kümesi düşük Gini indeksine, heterojen veri kümesi ise yüksek Gini indeksine sahiptir.

4. Karar Ağacı Örneği

- **Özellik Örneği**: Bir müşterinin kredi alıp almayacağını tahmin eden karar ağacında, ilk özellik olarak "Gelir" seçilebilir.
- **Dallar**:
    - Gelir > 50K
    - Gelir ≤ 50K
- **Yaprak Düğümler**:
    - Gelir > 50K ve Kredi Geçmişi İyi → "Kredi Verilebilir"
    - Gelir > 50K ve Kredi Geçmişi Kötü → "Kredi Verilemez"

5. Karar Ağaçları Algoritmaları

- **ID3**: Bilgi kazancı kullanır ve yalnızca kategorik verilerle çalışır.

- **C4.5**: ID3'ün geliştirilmiş versiyonudur, sürekli değerleri destekler ve eksik veriyi işleyebilir.
- **CART (Classification and Regression Trees)**: Hem sınıflandırma hem regresyon problemleri için uygundur. Gini indeksi veya MSE (Mean Squared Error) kullanır.

6. Scikit-learn Uygulaması

- **DecisionTreeClassifier**: Sınıflandırma için kullanılır, CART algoritmasını kullanır.
- **DecisionTreeRegressor**: Regresyon için kullanılır.
- **Özellikler**:
    - **gini**: Gini indeksi (sınıflandırma)
    - **entropy**: Entropi bazlı bilgi kazancı
    - **MSE**: Regresyon problemleri için

7. Aşırı Öğrenme (Overfitting) ve Budama

- **Aşırı Öğrenme (Overfitting)**: Karar ağaçları, eğitim verisine çok iyi uyum sağlarsa genelleme yeteneğini kaybedebilir.
- **Budama (Pruning)**: Ağaçtaki gereksiz dalların kaldırılması işlemi. İki türü vardır:
    - **Ön Budama (Pre-Pruning)**: Ağacın büyümesini önceden durdurur.
    - **Sonradan Budama (Post-Pruning)**: Ağaç oluşturulduktan sonra gereksiz dallar kaldırılır.

8. Scikit-learn'de Budama

- **Pre-Pruning**: max_depth, min_samples_leaf gibi parametrelerle ağacın büyümesi kontrol edilir.
- **Post-Pruning**: ccp_alpha parametresi ile ağacın karmaşıklığına bir maliyet eklenir, böylece gereksiz dallar budanır.

9. Karar Ağaçlarının Avantajları ve Dezavantajları

- **Avantajlar**:
    - Kolay anlaşılır ve görselleştirilebilir.
    - Hem kategorik hem de sürekli verilerle çalışabilir.
- **Dezavantajlar**:
    - Aşırı öğrenme (overfitting) eğilimi vardır.
    - Karmaşık veri setlerinde verimsiz olabilir.

Özetle, karar ağaçları, veri setinin kolayca analiz edilmesine yardımcı olan güçlü ve esnek bir yöntemdir. Ancak, büyük veri setlerinde ve karmaşık problemler için aşırı öğrenmeyi engellemek adına uygun parametrelerle dikkatlice ayarlanmalıdır.

Soru 1: Karar ağaçları hangi tür problemleri çözmek için kullanılır?
a) Sınıflandırma
b) Regresyon
c) Her ikisi
d) Kümeleme
e) Dimensionality reduction

Doğru Cevap: c) Her ikisi
Açıklama: Karar ağaçları hem sınıflandırma hem de regresyon problemleri için kullanılabilir.

Soru 2: Karar ağacında, veriyi ilk kez bölen düğüm hangi düğümdür?
a) Yaprak düğüm
b) İç düğüm
c) Kök düğüm
d) Dal
e) Terminal düğüm

Doğru Cevap: c) Kök düğüm
Açıklama: Kök düğüm, karar ağacının başlangıç noktasıdır ve veriyi ilk kez bölen özelliği temsil eder.

Soru 3: Karar ağacında, veriyi alt gruplara ayıran yolları hangi terim ifade eder?
a) Kök düğüm
b) Yaprak düğüm
c) Dallar
d) Özellikler
e) Kriterler

Doğru Cevap: c) Dallar
Açıklama: Dallar, veriyi alt gruplara ayıran yollardır ve özelliklerin değerlerine dayanır.

Soru 4: Gini indeksi hangi durumu ölçer?
a) Verinin doğruluğu
b) Veri kümesinin saflığı
c) Veri kümesinin çeşitliliği

d) Özelliklerin önem sırasını
e) Modelin performansını

Doğru Cevap: b) Veri kümesinin saflığı
Açıklama: Gini indeksi, veri kümesinin saflığını ölçer. Düşük Gini değeri homojen veri kümesini, yüksek Gini değeri ise heterojen veri kümesini gösterir.

Soru 5: Karar ağaçlarında hangi özellik veri kümesini bölerken kullanılır?
a) Hedef değişken
b) Kök düğümdeki özellik
c) Özellikler rastgele seçilir
d) Eğitim setindeki örnekler
e) Ağırlıklı ortalama

Doğru Cevap: b) Kök düğümdeki özellik
Açıklama: Kök düğümde, veri kümesini bölen özellik genellikle bilgi kazancı veya Gini indeksi gibi ölçütlerle seçilir.

Soru 6: Karar ağacında hangi parametre, ağacın derinliğini sınırlayarak aşırı öğrenmeyi engeller?
a) max_depth
b) min_samples_split
c) ccp_alpha
d) max_features
e) criterion

Doğru Cevap: a) max_depth
Açıklama: max_depth parametresi, ağacın büyümesini sınırlandırarak aşırı öğrenmeyi engeller.

Soru 7: Karar ağaçlarında hangi ölçüt, sınıflandırma problemi için kullanılır?
a) Mean Squared Error (MSE)
b) Bilgi Kazancı
c) Entropi
d) Gini İndeksi
e) Her biri

Doğru Cevap: d) Gini İndeksi
Açıklama: Gini indeksi, sınıflandırma problemleri için yaygın olarak kullanılır ve veri kümesinin saflığını ölçer.

Soru 8: Karar ağacında, veriyi bölerken kullanılan kriterlerden biri nedir?
a) Prediktif doğruluk
b) Hedef fonksiyon
c) Bilgi Kazancı
d) Regresyon hatası
e) Derinlik oranı

Doğru Cevap: c) Bilgi Kazancı
Açıklama: Bilgi kazancı, entropiyi (belirsizlik ölçüsünü) kullanarak, veri kümesindeki bölme noktasını belirler.

Soru 9: Karar ağacı algoritmalarında hangi işlem, modelin gereksiz dallardan kurtulmasına yardımcı olur?
a) Budama
b) Çapraz doğrulama
c) Ağırlıklı ortalama
d) Hiperparametre ayarlama
e) Dallanma

Doğru Cevap: a) Budama
Açıklama: Budama, karar ağacındaki gereksiz dalları keserek aşırı öğrenmeyi (overfitting) engellemeye yardımcı olur.

Soru 10: Karar ağacında "Pre-pruning" işlemi ne zaman yapılır?
a) Ağaç tamamen oluşturulmadan önce
b) Ağaç tamamen oluşturulduktan sonra
c) Eğitim veri kümesi azsa
d) Özellikler homojense
e) Model başarısızsa

Doğru Cevap: a) Ağaç tamamen oluşturulmadan önce
Açıklama: Pre-pruning, karar ağacının büyümesini önceden durdurur ve bu sayede aşırı öğrenmeyi engeller.

Karar Ağaçları ve Random Forest Algoritması - Kısa Notlar

1. **Karar Ağaçları**: Veriyi dallara ayırarak sonuca ulaşmayı sağlar. Tek bir karar ağacı, *overfitting* (aşırı öğrenme) yapabilir. Her düğümde veriyi bir özellik ile böler.
2. **Random Forest**: Birden fazla karar ağacını bir arada kullanarak tahmin doğruluğunu artırır. Çoğunlukla *bootstrapping* kullanılarak veri setinden alt kümeler alınır ve her ağaç farklı alt kümelerle eğitilir.
3. **Bootstrapping**: Eğitim verisinin rastgele örneklenmesi, her örneklemede veri setinin %63-67'si kullanılır. Geri kalan veri ise *Out-of-Bag* (OOB) hatası için saklanır.
4. **Out-of-Bag (OOB) Hatası**: Modelin doğruluğunu test verisi olmadan ölçer ve her ağaca dahil edilmeyen verilerle hesaplanır.
5. **Ağaçların Çeşitliliği**: Her ağaçta yalnızca özniteliklerin bir alt kümesi kullanılır. Bu, farklı ağaçların çeşitlenmesini sağlar ve *overfitting* riskini azaltır.
6. **Random Forest Tahmin Yöntemi**:
    - **Sınıflandırma**: Çoğunluk oylaması (majority vote).
    - **Regresyon**: Ağaçların ortalama tahmini.
7. **Scikit-learn Parametreleri**:
    - **n_estimators**: Kaç adet karar ağacı oluşturulacağı.
    - **max_depth**: Ağaçların maksimum derinliği.
    - **min_samples_split**: Bir düğümün bölünmesi için gereken minimum örnek sayısı.
    - **min_samples_leaf**: Yaprak düğümdeki minimum örnek sayısı.
    - **max_features**: Her ağaçta kullanılacak özellik sayısı.
8. **Avantajlar**:
    - Daha az *overfitting* yapar, genellikle daha iyi genelleştirme sağlar.
    - Hem sınıflandırma hem regresyon için kullanılabilir.
    - Gürültüye karşı dayanıklıdır.
9. **Dezavantajlar**:
    - Büyük veri setlerinde yavaş olabilir.
    - Karar ağaçlarına göre daha karmaşık bir yapıya sahiptir.

Random Forest, birden fazla karar ağacının birleşimiyle daha güçlü ve genel geçerli tahminler yapar. Ancak, daha fazla ağaç kullanıldıkça hesaplama maliyeti artar.

Soru 1: Karar ağaçlarının en büyük dezavantajı nedir?
a) Zor anlaşılabilir olması
b) Overfitting (aşırı öğrenme) riski
c) Hızlı çalışması
d) Hem sınıflandırma hem de regresyonu desteklememesi
e) Gürültüye dayanıklı olmamaları

Doğru Cevap: b) Overfitting (aşırı öğrenme) riski
Açıklama: Tek bir karar ağacı genellikle eğitim verisine çok iyi uyum sağlar, ancak bu genelleme yeteneğini azaltır. Bu da *overfitting* sorununa yol açar.

Soru 2: Random Forest algoritmasında, her ağaç oluşturulurken hangi işlem yapılır?
a) Eğitim verisi yalnızca bir kez kullanılır
b) Tüm veriler aynı alt küme ile kullanılır
c) Verinin bir alt kümesi rastgele seçilir (Bootstrapping)
d) Her ağaç için aynı özellikler kullanılır
e) Ağaçlar birbirini etkilemez

Doğru Cevap: c) Verinin bir alt kümesi rastgele seçilir (Bootstrapping)
Açıklama: Random Forest, verinin farklı alt kümelerini rastgele seçerek birden fazla karar ağacı oluşturur.

Soru 3: Random Forest algoritmasında tahmin yapılırken hangi yöntem kullanılır?
a) En küçük tahmin
b) Ortalama tahmin
c) Çoğunluk oylaması (majority vote)
d) Kümelenmiş tahmin
e) Ortalama hata

Doğru Cevap: c) Çoğunluk oylaması (majority vote)
Açıklama: Sınıflandırma problemlerinde, Random Forest çoğunluk oylaması kullanarak ağaçların tahminlerini birleştirir.

Soru 4: Out-of-Bag (OOB) hatası nedir?
a) Modelin eğitim verisiyle test edilmesi
b) Modelin doğruluğunu test verisi olmadan ölçme
c) Her ağacın doğruluğunun ölçülmesi

d) Modelin yalnızca belirli verilerle test edilmesi
e) Eğitim verisinin tamamı ile doğrulama yapılması

Doğru Cevap: b) Modelin doğruluğunu test verisi olmadan ölçme
Açıklama: OOB hatası, her ağaç için eğitim verisine dahil edilmeyen verilerle modelin doğruluğunu test etmeyi sağlar.

Soru 5: Random Forest algoritmasında hangi parametre, her ağaç için kullanılacak özellik sayısını belirler?
a) n_estimators
b) max_depth
c) max_features
d) min_samples_split
e) min_samples_leaf

Doğru Cevap: c) max_features
Açıklama: *max_features*, her ağaç için kullanılacak öznitelik sayısını belirler ve ağaçların çeşitliliğini artırır.

Soru 6: Random Forest algoritmasında "n_estimators" parametresi neyi kontrol eder?
a) Ağaçların derinliğini
b) Ağaç sayısını
c) Ağaçların hızı
d) Özellik sayısını
e) Düğüm sayısını

Doğru Cevap: b) Ağaç sayısını
Açıklama: *n_estimators*, Random Forest modelinde kaç tane karar ağacı oluşturulacağını belirler.

Soru 7: Random Forest, tek bir karar ağacına göre hangi avantajları sunar?
a) Daha hızlı çalışır
b) Daha az *overfitting* yapar
c) Kararları daha basit hale getirir
d) Daha düşük hesaplama maliyeti sunar
e) Sadece sınıflandırma problemleri için uygundur

Doğru Cevap: b) Daha az *overfitting* yapar
Açıklama: Birden fazla ağaç kullanarak genellikle daha iyi genelleştirme sağlar ve *overfitting* riskini azaltır.

Soru 8: Random Forest algoritmasında hangi parametre, her düğümün bölünebilmesi için gereken minimum örnek sayısını belirler?
a) max_depth
b) n_estimators
c) min_samples_split
d) min_samples_leaf
e) max_features

Doğru Cevap: c) min_samples_split
Açıklama: *min_samples_split*, her düğümün bölünebilmesi için gereken minimum örnek sayısını belirler.

Soru 9: Random Forest algoritmasında *overfitting* nasıl önlenir?
a) Daha fazla ağaç kullanarak
b) Daha az özellik seçerek
c) Her ağaç için daha derin bir yapı oluşturularak
d) Ağaçların büyümesi durdurularak (Budama)
e) Tüm verilerle eğitim yapılarak

Doğru Cevap: d) Ağaçların büyümesi durdurularak (Budama)
Açıklama: Random Forest, *overfitting*'i engellemek için *pre-pruning* veya *post-pruning* yöntemlerini kullanabilir.

Soru 10: Random Forest algoritmasının büyük veri setlerinde neden yavaş olabileceği söylenebilir?
a) Daha fazla ağaç eğitildiği için
b) Her ağaçta tüm veriler kullanıldığı için
c) Ağaçların her biri bağımsız olarak eğitildiği için
d) Ağaçlar paralel çalıştığı için
e) Her ağaçta kullanılan özellik sayısı çok fazla olduğu için

Doğru Cevap: a) Daha fazla ağaç eğitildiği için
Açıklama: Daha fazla ağaç eğitildiğinde hesaplama maliyeti artar ve bu, özellikle büyük veri setlerinde yavaşlamaya neden olabilir.