

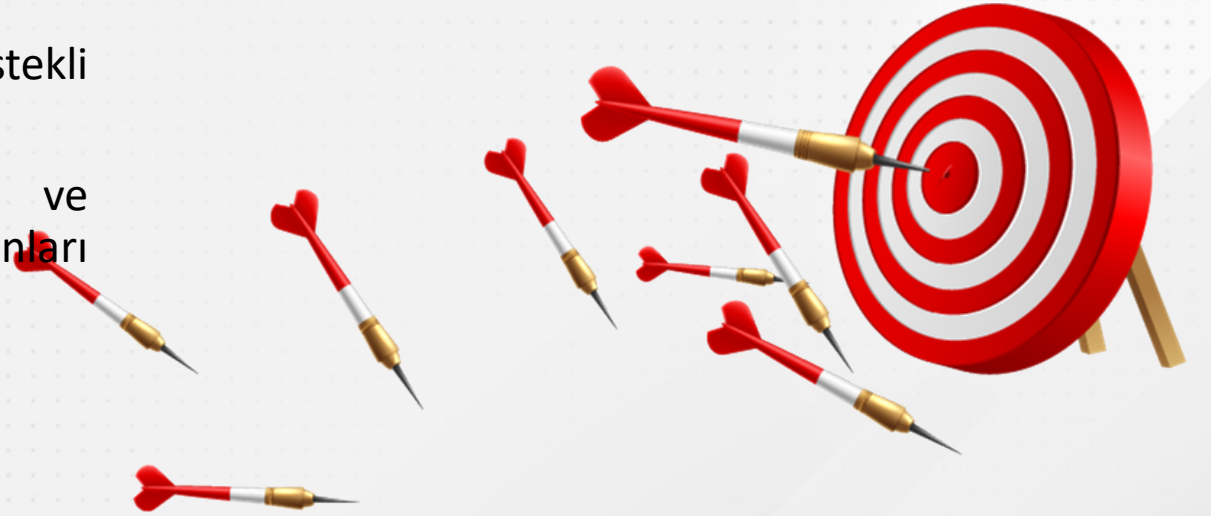


ANKARA
HACI BAYRAM VELİ ÜNİVERSİTESİ
**UZAKTAN EĞİTİM UYGULAMA VE
ARAŞTIRMA MERKEZİ**

YAPAY ZEKÂ TEKNOLOJİLERİ VE YÖNETİMİ

Dersin Genel Hedefleri

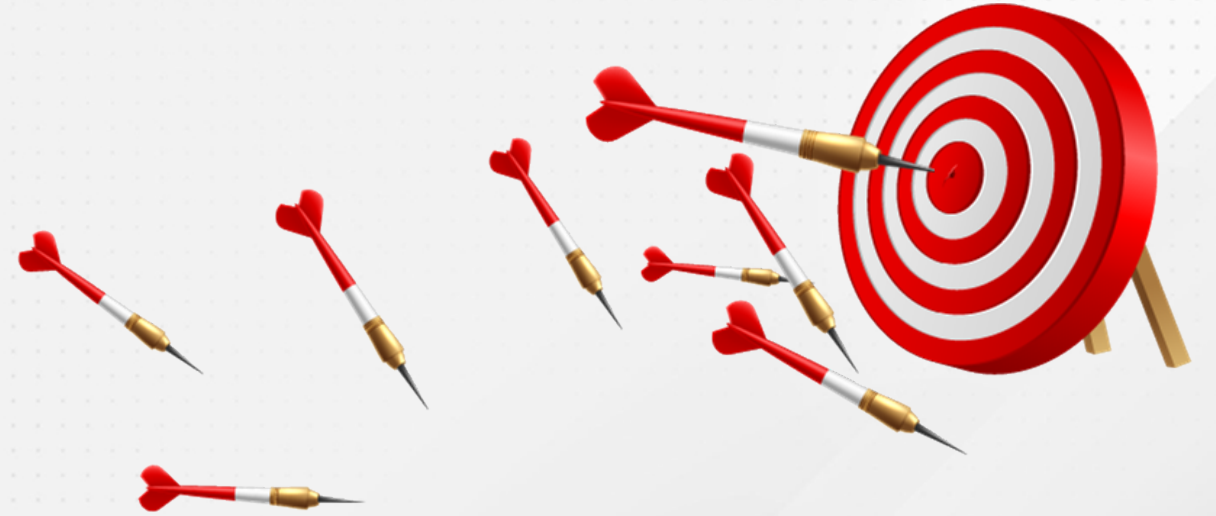
1. Yapay zekâ türlerinin işleyişlerini, güçlü zayıf yönlerini ve kullanım alanlarını açıklar.
2. Hazır kütüphaneleri kullanarak sınıflandırma, kümeleme ve tahmin amaçlarını yerine getiren scriptler yazabilir.
3. Veri analizi için istatistiksel çözümleri uygunluk ve etkililik açısından karşılaştırır.
4. İşletme problemlerine yönelik yapay zekâ destekli çözümler geliştirebilir.
5. YBS alanındaki güncel yapay zekâ teknoloji ve uygulamalarından hareketle gelecek projeksiyonları yapar.



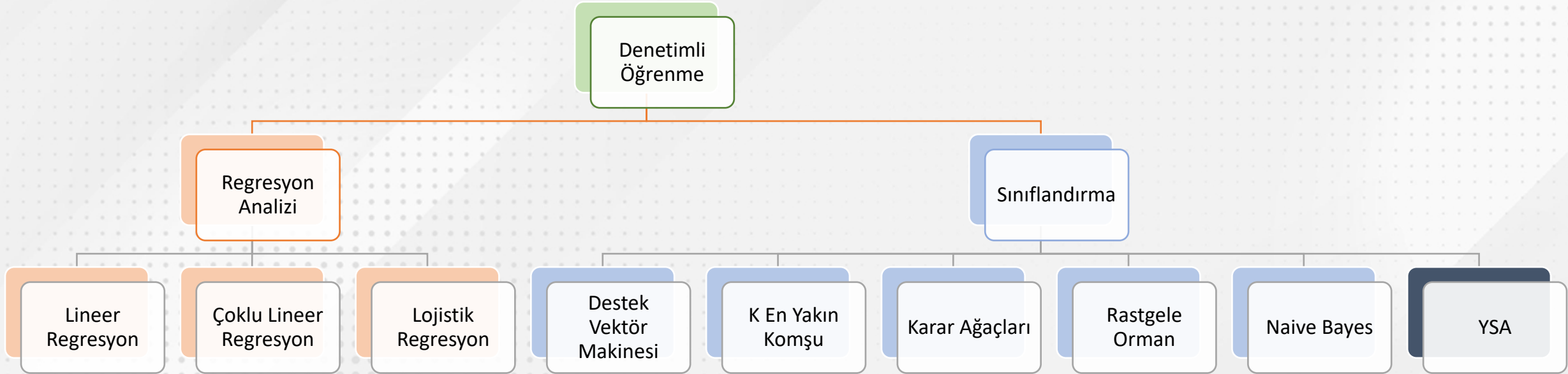
HAFTA 5. MAKİNE ÖĞRENMESİ

Bölüm Hedefleri

1. Karar Ağacı, Rastgele Orman, Naive Bayes yaklaşımlarına hakim olur.
2. Karışıklık matrisi oluşturabilir.
3. Performans ölçütlerini ve farklarını bilir.
4. Overfitting ve outlier kavramlarını tanımlayabilir.

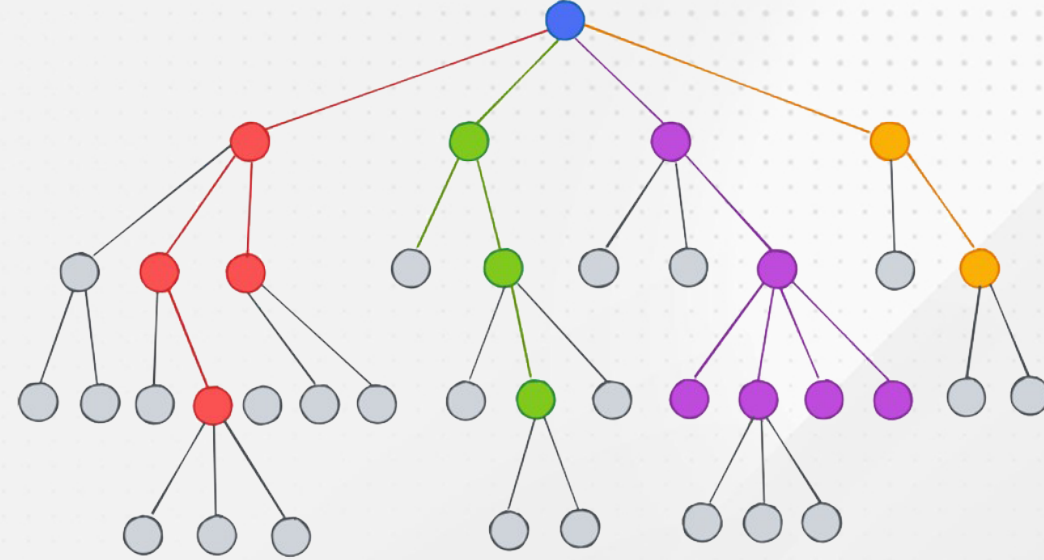


5.1. Denetimli Öğrenme Yaklaşımları



5.1.1. Karar Ağaçları (Decision Trees)

- Karar ağaçları, makine öğrenmesinde sınıflandırma ve regresyon problemlerini çözmek için yaygın olarak kullanılan bir modelleme tekniğidir.
- Bu teknik, bir ağaç yapısı kullanarak veri setindeki özelliklerin değerlerine göre kararlar alır.
- Her iç düğüm, bir özellik ve bu özelliğin değerine göre bir ayırım noktası içerirken, yaprak düğümleri sonuçları veya tahminleri temsil eder.
- Karar ağaçları, kolayca yorumlanabilir olmaları ve anlaşılır bir karar yapısına sahip olmaları nedeniyle tercih edilirler.
- Ayrıca, kayıp verilerle başa çıkmak ve sayısal ve kategorik verileri bir arada işlemek gibi avantajlara sahiptirler.
- Ancak, aşırı uyum eğiliminde olabilirler, bu nedenle ağacın derinliği ve dal bölünmelerinin sayısı gibi hiperparametrelerin iyi ayarlanması önemlidir.



5.1.1. Karar Ağaçları (Decision Trees)

Örnek:

- Veri Seti: Elimizde, bir e-ticaret şirketinin müşteri verilerini içeren bir veri seti olduğunu düşünelim. Her müşteri için yaş, cinsiyet, üyelik süresi ve satın alma geçmişi gibi özellikler bulunmaktadır. Amacımız, müşterilerin bir indirim kampanyasına katılıp katılmayacaklarını tahmin eden bir karar ağacı oluşturmaktır.

1. Adım Entropi Hesaplama

- İlk adımda, her bir özellik için veri setinin entropisi hesaplanır. Entropi, belirsizlik seviyesini ölçen bir terimdir. ID3 ve C4.5 gibi algoritmalar, entropiyi kullanarak en iyi bölünmeyi bulmaya çalışır.

2. Adım En İyi Bölünme Seçimi

- En iyi bölünme, veri setini mümkün olan en homojen alt gruplara ayıran bölünmedir. Entropi azaldıkça, bölünme daha iyidir. Örneğin, yaş özelliği için 30 yaşından büyük veya küçük diye bir bölünme yapılabilir.

3. Adım Özellik Seçimi

- Özellik seçiminde, her bir özelliğin entropi azaltma miktarını değerlendiririz. Entropi azaldıkça, özellik daha iyi bir bölünme sağlar.

4. Adım Ağacın Oluşturulması

- Bu bölünme işlemi, her bir alt gruptaki veri setleri için tekrarlanır. Her adımda en iyi bölünme seçilerek ağaç derinleştirilir. Bu işlem, bir alt grup belirli bir kriterle uymadığında veya belirli bir derinlik seviyesine ulaşıldığında durur.

5. Adım Yaprak Düğümlere Sınıfların Atanması

- Her yaprak düğümü, o düğüme ait olan veri noktalarının çoğunluğuna dayanarak bir sınıfa atanır. Örneğin, bir yaprak düğümü "Katılır" sınıfına aitse, o düğümdeki müşterilere "Katılır" denir.

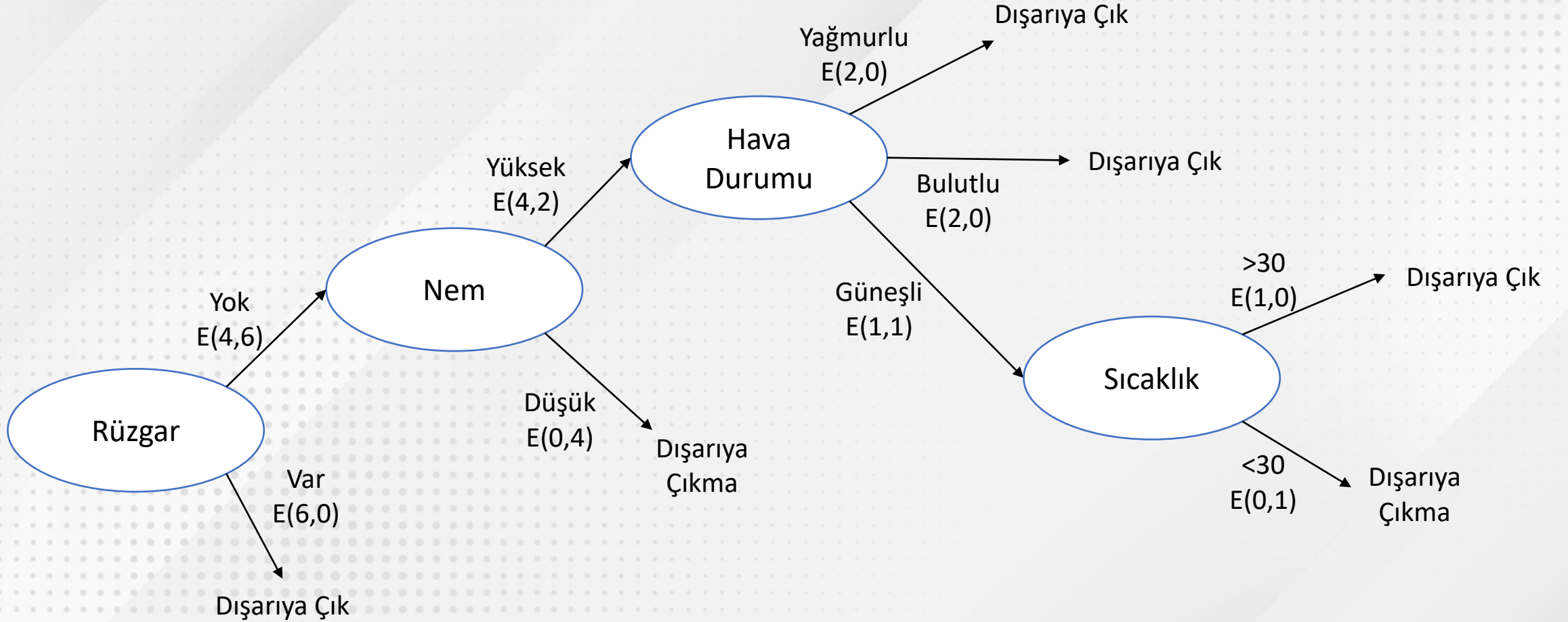
5.1.1. Karar Ağaçları (Decision Trees)

Örnek:

Hava Durumu	Sıcaklık	Nem	Rüzgar	Yürüyüşe Çıkma
Güneşli	30	Düşük	Var	1
Güneşli	30	Düşük	Var	1
Yağmurlu	22	Düşük	Yok	0
Bulutlu	28	Düşük	Yok	0
Yağmurlu	22	Düşük	Yok	0
Bulutlu	28	Düşük	Yok	0
Güneşli	24	Yüksek	Var	1
Bulutlu	27	Yüksek	Var	1
Güneşli	24	Yüksek	Var	1
Bulutlu	27	Yüksek	Var	1
Güneşli	25	Yüksek	Yok	0
Bulutlu	20	Yüksek	Yok	1
Yağmurlu	18	Yüksek	Yok	1
Güneşli	25	Yüksek	Yok	0
Bulutlu	20	Yüksek	Yok	1
Yağmurlu	18	Yüksek	Yok	1

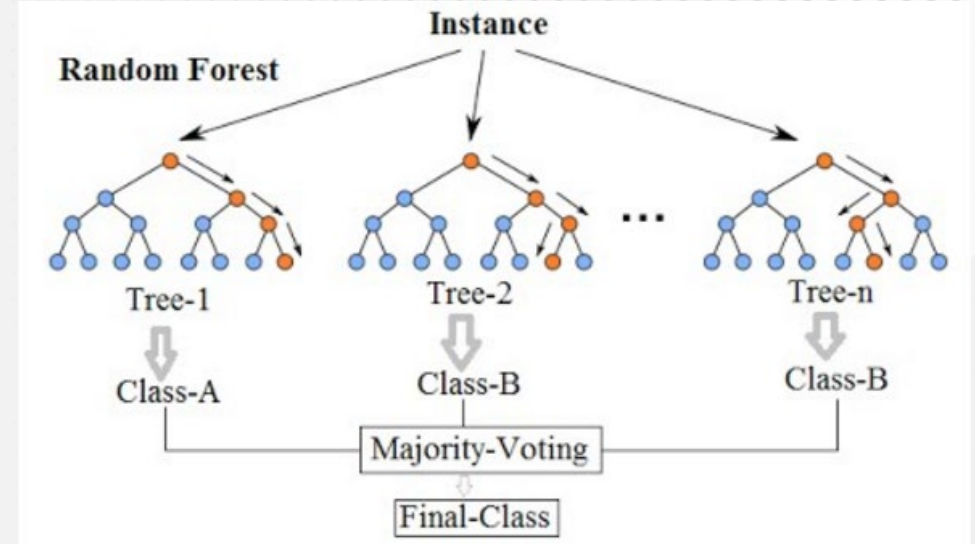
5.1.1. Karar Ağaçları (Decision Trees)

Örnek:



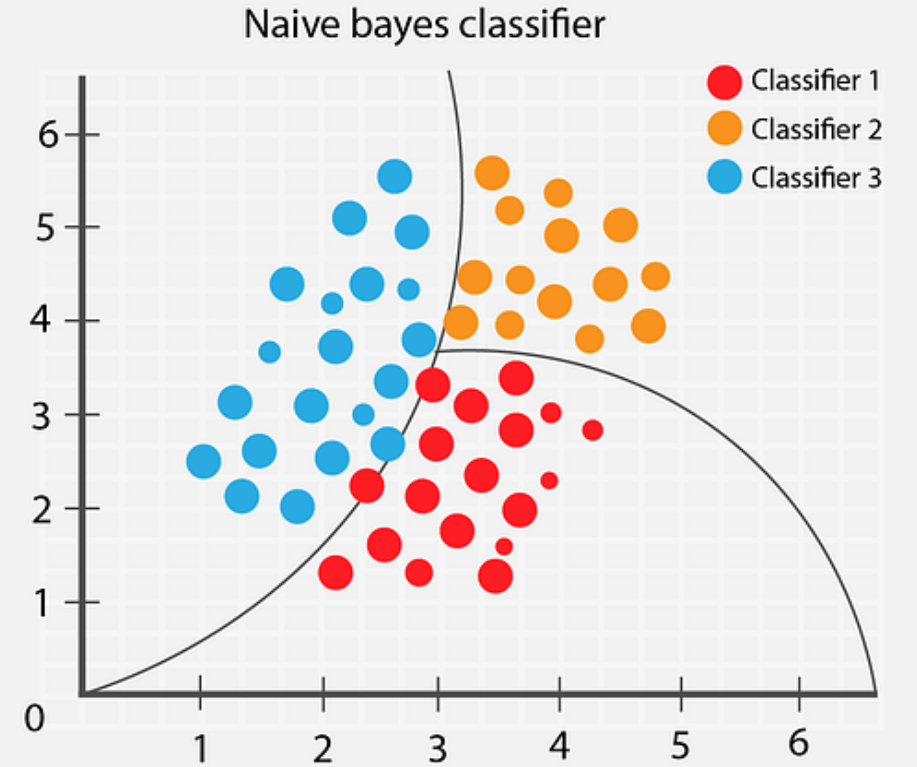
5.1.2. Rastgele Orman (Random Forest)

- Rastgele Orman, makine öğrenmesinde sınıflandırma ve regresyon problemlerini çözmek için kullanılan güçlü bir öğrenme algoritmasıdır.
- Rastgele Orman, bir dizi karar ağacının bir araya gelmesiyle oluşur ve her bir ağacın eğitim veri setinin farklı bir alt kümesi üzerinde eğitilmesini sağlar.
- Her ağaç, rastgele özelliklerle oluşturulan bir alt kümeyle oluşturulur ve genellikle veri setindeki değişkenlik ve aşırı uyumu azaltır.
- Rastgele Orman, her bir ağacın tahminini alırken, sınıfların veya değerlerin bir ortalamasını birlikte kullanır.
- Rastgele Orman, yüksek boyutlu veri setlerinde ve büyük özellik uzaylarında etkili bir şekilde çalışır, ayrıca genellikle hiperparametre ayarına duyarlı olmayan doğrusal olmayan ilişkileri ele alabilir.



5.1.3. Naive Bayes

- Naive Bayes, makine öğrenmesinde sınıflandırma problemlerini çözmek için kullanılan bir olasılık temelli bir sınıflandırma algoritmasıdır.
- Bu algoritma, Bayes teoreminin temel prensiplerine dayanır ve "naive" (saf) olarak adlandırılır, çünkü sınıf etiketleri arasındaki ilişkileri basitleştirir ve özellikler arasındaki bağımsızlık varsayımını kabul eder.
- Özellikle metin sınıflandırması gibi problemlerde başarılı sonuçlar verir.
- Naive Bayes, hesaplama açısından hızlıdır ve genellikle küçük boyutlu veri kümeleri üzerinde iyi performans gösterir. Ancak, bağımsızlık varsayımının gerçek dünyadaki veri setlerinde her zaman geçerli olmadığı durumlar söz konusu olabilir.



5.1.4. Denetimli Öğrenme Yaklaşımlarının Karşılaştırılması

Yöntem	Avantajları	Dezavantajları
Karar Ağacı	- Anlaşılması ve yorumlanması kolaydır. - Kategorik ve sayısal verilerle çalışabilir. - Overfitting'e karşı dirençlidir.	- Çok karmaşık modeller oluşturabilir. - Küçük veri setlerinde aşırı öğrenme eğilimindedir.
Destek Vektör Makineleri	- Yüksek boyutlu veri setlerinde etkili. - Non-lineer sınıflandırma sorunlarına uyarlanabilir. - Outlier'lara karşı dirençlidir.	- Hesaplama maliyeti yüksek olabilir. - Çok sayıda özellik içeren veri setlerinde performans düşebilir.
K En Yakın Komşu	- Basit ve anlaşılır bir algoritmadır. - Yeni veri setleriyle kolayca güncellenebilir.	- Büyük veri setlerinde hesaplama maliyeti yüksek olabilir. - Etkili bir şekilde çalışabilmesi için özellik mühendisliği önemlidir.
Naive Bayes	- Basit, hızlı ve etkili bir algoritmadır. - Çok sayıda özellik içeren veri setlerinde iyi performans gösterir.	- Özellikler arasındaki bağımlılıkları dikkate almaz. - Duyarlılık ve özgüllük dengesini iyi ayarlamak zordur.

5.2. Karışıklık Matrisi (Confussion Matrix)

- Karışıklık matrisi, bir sınıflandırma modelinin performansını değerlendirmek için kullanılan bir araçtır.
- Genellikle makine öğrenimi ve istatistiksel sınıflandırma problemlerinde kullanılır.
- Bir karışıklık matrisi, gerçek ve tahmin edilen sınıflar arasındaki ilişkiyi görselleştirmek için kullanılır.
- Matris, doğru ve yanlış sınıflandırılmış örneklerin sayısını dört farklı kategoriye ayırır: doğru pozitif (TP), yanlış pozitif (FP), doğru negatif (TN) ve yanlış negatif (FN).
- Karışıklık matrisi, modelin performansını değerlendirmek ve olası hataları belirlemek için kullanılabilir, örneğin hassasiyet, özgüllük, doğruluk ve F1 puanı gibi metriklerin hesaplanmasında temel oluşturur.

Gerçek Değerler

Tahmini Değerler

	Pozitif (1)	Negatif (1)
Pozitif (1)	TP	FP
Negatif (0)	FN	TN

5.2. Karışıklık Matrisi (Confussion Matrix)

- TP (True Positive): Gerçek pozitiflerin sayısını ifade eder. Yani, modelin doğru şekilde pozitif olarak sınıflandırdığı örneklerin sayısıdır.
- TN (True Negative): Gerçek negatiflerin sayısını ifade eder. Modelin doğru şekilde negatif olarak sınıflandırdığı örneklerin sayısıdır.
- FP (False Positive): Yanlış pozitiflerin sayısını ifade eder. Yani, modelin pozitif olarak sınıflandırdığı ancak gerçekte negatif olan örneklerin sayısıdır.
- FN (False Negative): Yanlış negatiflerin sayısını ifade eder. Modelin negatif olarak sınıflandırdığı ancak gerçekte pozitif olan örneklerin sayısıdır.
- Örneğin, bir hastalığı tespit etmek için bir model kullanıyorsanız, TP hastalığı doğru şekilde tespit edilen hastaları, TN sağlıklı bireyleri doğru şekilde teşhis edilenleri, FP sağlıklı bireyleri hastalıklı olarak yanlışlıkla sınıflandırılanları ve FN hastalıklı bireyleri sağlıklı olarak yanlışlıkla sınıflandırılanları temsil eder.

Gerçek Değerler

Tahmini Değerler

	Pozitif (1)	Negatif (1)
Pozitif (1)	TP	FP
Negatif (0)	FN	TN

5.2.1. Karışıklık Matrisi Örneği

- Bir işletmenin rekabet durumunu düşük ve yüksek olmak üzere iki sınıfta tahmin eden bir sınıflandırma algoritması için sağ tarafta bir karışıklık matrisi verilmiştir.
- Gerçek veriler ve algoritmanın tahmin ettiği veriler aşağıdaki gibidir.
- Gerçek veri: Y,Y, Y,Y,Y,Y,D,D,D
- Tahmini Veri: Y,D,D,Y,Y,Y,D,D,Y,D
- Burada karışıklık matrisi yüksek sınıfına göre kurulacaktır.

Tahmini Veri Gerçek Veri	Yüksek	Düşük
Yüksek	TP	FP
Düşük	FN	TN

5.2.1. Karışıklık Matrisi Örneği

- İlk olarak TP'lerin sayısını alırız. Yani Gerçek verisi ve Tahmini Verisi «Yüksek» sınıfına ait olanların sayısı:
- Gerçek veri: Y, Y, Y, Y, Y, Y, Y, D, D, D
- Tahmini Veri: Y, D, D, Y, Y, Y, D, D, Y, D
- Daha sonra TN'lerin sayısını alırız. Yani Gerçek verisi ve Tahmini verisi «Düşük» sınıfına ait olanların sayısı:
- Gerçek veri: Y, Y, Y, Y, Y, Y, Y, D, D, D
- Tahmini Veri: Y, D, D, Y, Y, Y, D, D, Y, D

Tahmini Veri Gerçek Veri	Yüksek	Düşük
Yüksek	4	FP
Düşük	FN	2

5.2.1. Karışıklık Matrisi Örneği

- Daha sonra FP'lerin sayısını alırız. Yani Gerçek verisi «Yüksek», Tahmini Verisi «Düşük» sınıfına ait olanların sayısı:
- Gerçek veri: Y, Y, Y, Y, Y, Y, Y, D, D, D
- Tahmini Veri: Y, D, D, Y, Y, Y, D, D, Y, D
- Daha sonra FN'lerin sayısını alırız. Yani Gerçek verisi «Düşük», Tahmini verisi «Yüksek» sınıfına ait olanların sayısı:
- Gerçek veri: Y, Y, Y, Y, Y, Y, Y, D, D, D
- Tahmini Veri: Y, D, D, Y, Y, Y, D, D, Y, D

Tahmini Veri Gerçek Veri	Yüksek	Düşük
	Yüksek	Düşük
Yüksek	4	3
Düşük	1	2

5.2.2. Çok Sınıflı Karışıklık Matrisi Örneği

- Bir işletmenin rekabet durumunu düşük, orta ve yüksek olmak üzere üç sınıfta tahmin eden bir sınıflandırma algoritması için sağ tarafta değerleri yazılmış bir karışıklık matrisi verilmiştir.
- Burada karışıklık matrisi yüksek sınıfına göre kurulacaktır.
- TP: «Yüksek» sınıfına ait gerçek ve «Yüksek» sınıfına ait tahmini verilerin kesişim sayısı (Turuncu Renk)
- TN: «Yüksek» sınıfı harici diğer sınıflarda olup kesişenlerin sayısı (Mavi Renk)
- FP: Diğer sınıflarda olduğu tahmin edilen ama gerçekte «Yüksek» sınıfına ait olanların kesişim sayısı (Yeşil Renk)
- FN: «Yüksek» sınıfında olduğu tahmin edilen ama gerçekte diğer sınıflara ait olanların sayısı

Tahmini Veri \ Gerçek Veri	Yüksek	Orta	Düşük
Yüksek	5	1	1
Orta	1	5	0
Düşük	0	0	6

5.3. Performans Ölçütleri

Hassasiyet (Precision):

- Hassasiyet, pozitif olarak sınıflandırılan örneklerin ne kadarının gerçekten pozitif olduğunu gösterir. Yani, $TP/(TP + FP)$ formülüyle hesaplanır. Yüksek hassasiyet, modelin pozitif olarak tahmin ettiği örneklerin gerçekten de pozitif olduğunu belirtir. Hassasiyet ne kadar yüksek olursa, modelin yanlış pozitiflerle karşılaşma olasılığı o kadar düşüktür.

Özgüllük (Specificity):

- Özgüllük, negatif olarak sınıflandırılan örneklerin ne kadarının gerçekten negatif olduğunu gösterir. Yani, $TN/(TN + FP)$ formülüyle hesaplanır. Yüksek özgüllük, modelin negatif olarak tahmin ettiği örneklerin gerçekten de negatif olduğunu belirtir. Özgüllük ne kadar yüksek olursa, modelin yanlış negatiflerle karşılaşma olasılığı o kadar düşüktür.

Doğruluk (Accuracy):

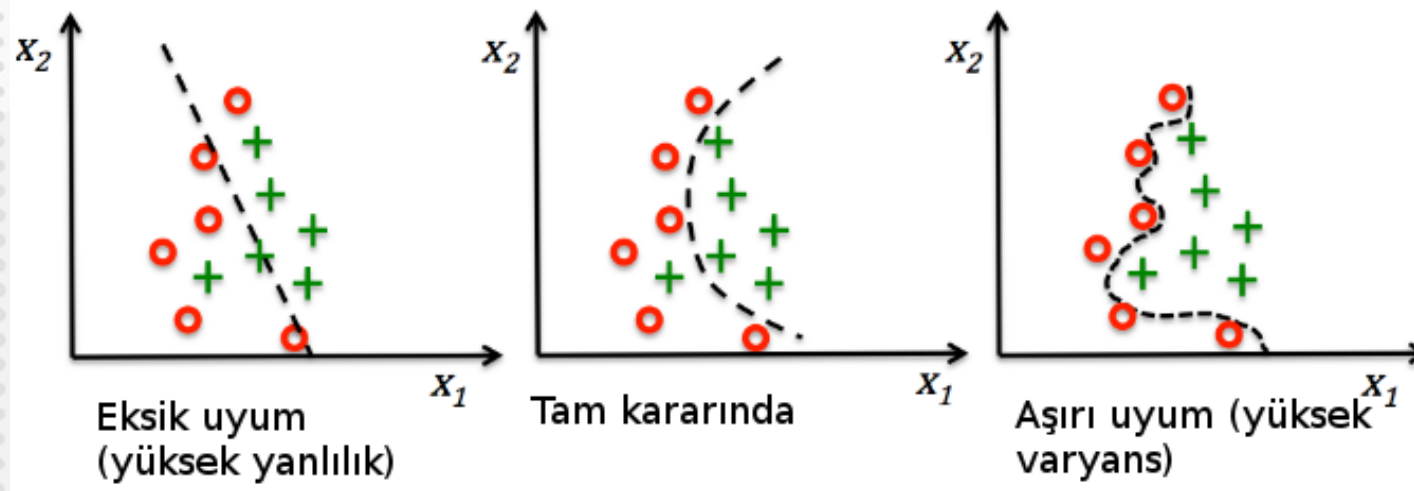
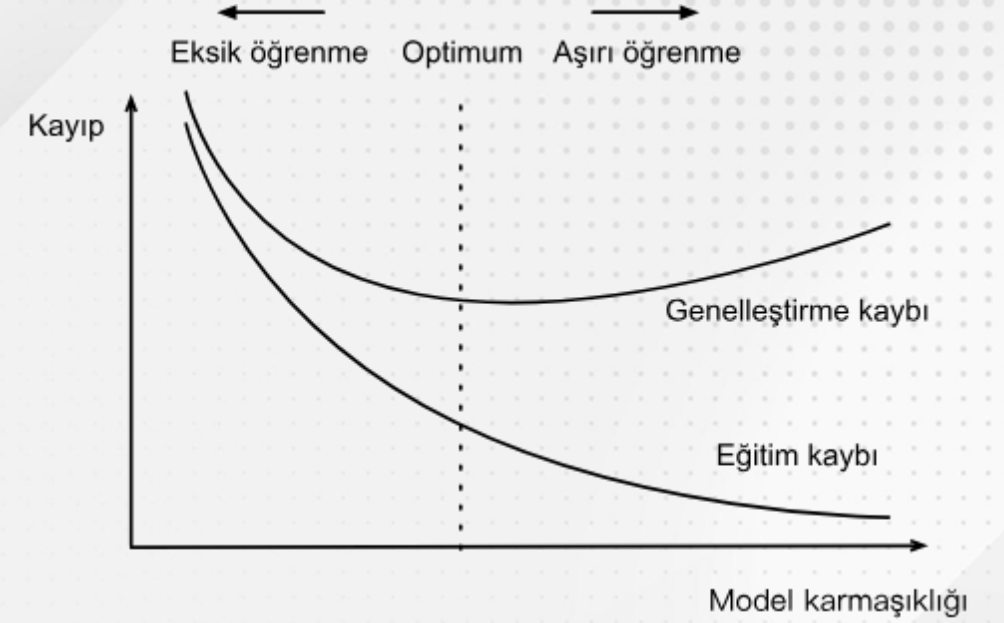
- Doğruluk, toplam doğru tahminlerin toplam örnek sayısına oranını gösterir. Yani, $(TP + TN) / (TP + TN + FP + FN)$ formülüyle hesaplanır. Yüksek doğruluk, modelin genel olarak doğru tahminlerde bulunduğunu gösterir. Ancak, dengesiz sınıf dağılımlarında doğruluk yanıltıcı olabilir.

F1 Puanı:

- F1 puanı, hassasiyet ve geri çağrı (recall) değerlerinin harmonik ortalamasını ifade eder. Bu, hem hassasiyetin hem de geri çağrının yüksek olması gereken durumlar için ideal bir metriktir. Özellikle dengesiz sınıf dağılımlarında kullanışlıdır. F1 puanı, $2 * (Hassasiyet * Geri Çağrı) / (Hassasiyet + Geri Çağrı)$ formülüyle hesaplanır.

5.4. Overfitting (Aşırı Uyum, Aşırı Öğrenme)

- Overfitting, bir makine öğrenmesi modelinin eğitim verilerinde tüm kırılımları çalıştırarak sonuçları ezberlemesine denir.
- Bütün kırılımları çalıştırması sayesinde, o veriler üzerinde başarı elde eder. Böylece genelleme yeteneği düşer.
- Eğitim verisini ezberleyen model, test verisi üzerinde düşük başarı elde edebilir.



5.4.1. Aşırı Uyumun Sebepleri

Veri Azlığı – Eksikliği

- Küçük veya yetersiz bir verisetinde çıkan sonuçların genellenebilirliği düşüktür, yüksek başarı elde edemez.

Veri Çokluğu – Gereksiz Veriler

- Model öğrenmek için,, gereksiz veya fazla özellikleri olan verilerdeki rastgele ilişkileri kullanabilir ve genelleme becerisi düşebilir.

Veri Önişleme Hataları

- Veriseti işlenmeye başlamadan önce model ile alakasız olan veriler, çıkarılmazsa yanlış sonuçlara ve aşırı uyuma neden olabilir.

Parametrelerin Uygun Olmaması

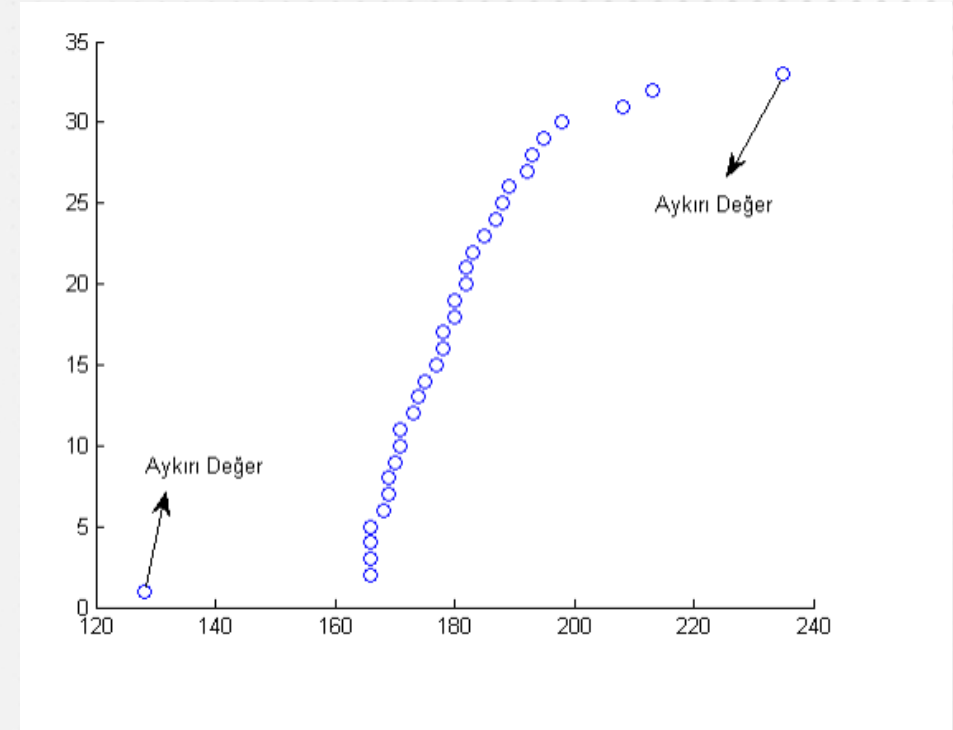
- Kullanılan modellerdeki parametrelerin yanlış belirlenmesi, eğitim modeline aşırı uyum sağlamasına neden olur. Model parametrelerinin beklenen sonuca göre ayarlanması; makinenin gerçek verileri ve verisetindeki gürültüyü tanımasını engeller.

Eğitim ve Doğrulama Verisinin Ayrımı

- Modelin elde etmesi gereken sonuç verilerinin, eğitimde kullanılması aşırı öğrenmeye ve genelleme zorluğuna yol açacaktır.

5.5. Aykırı Değer (Outlier)

- Aykırı değerler; verisetinde; kendisi dışındaki gözlemlerden önemli ölçüde farklılık gösteren değerlerdir. Aykırı değer; gerçek aykırı bir performanstan kaynaklandığı gibi, veri girişindeki hatalar ya da verideki bozulmalardan da ortaya çıkabilir.
- Aykırı değerlerin makine öğrenmesinde bazı avantaj ve dezavantajları vardır.
- Avantajları; modelin istisnai durumları tanıması ve değerlendirmesi, gerçek bir aykırı değer gerçeğe sonuçlara yol açması şeklindedir.
- Dezavantajları; modelin hassasiyetini, genelleme kabiliyetini, performansını düşürmesi olarak özetlenecektir.



5.5.1. Aykırı Değerlerin Tespit Edilmesi

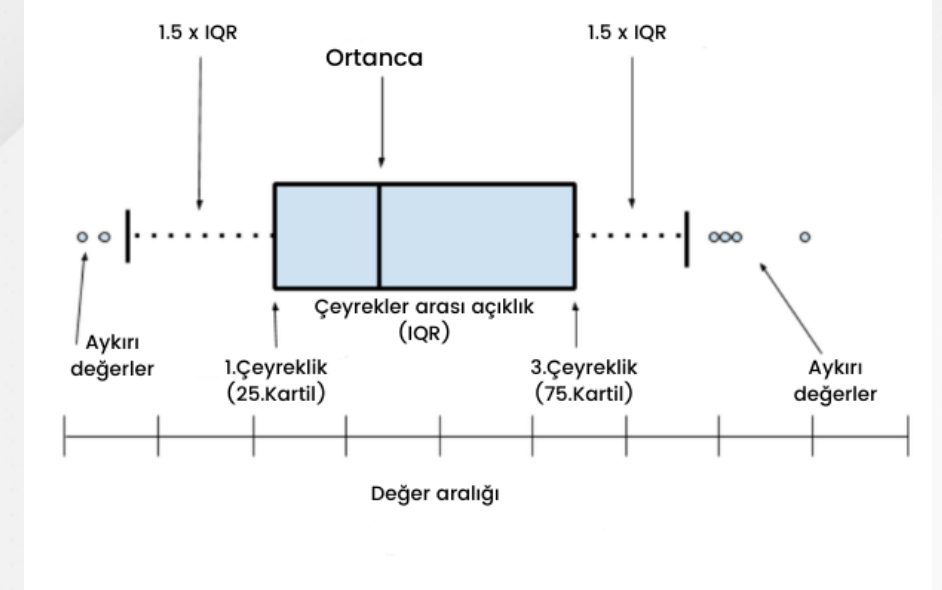
- Aykırı değerlerin belirlenmesinde görselleştirme, istatistik, makine öğrenmesi yöntemleri kullanılabilir.



5.5.1.1. Görselleştirme Yöntemleri

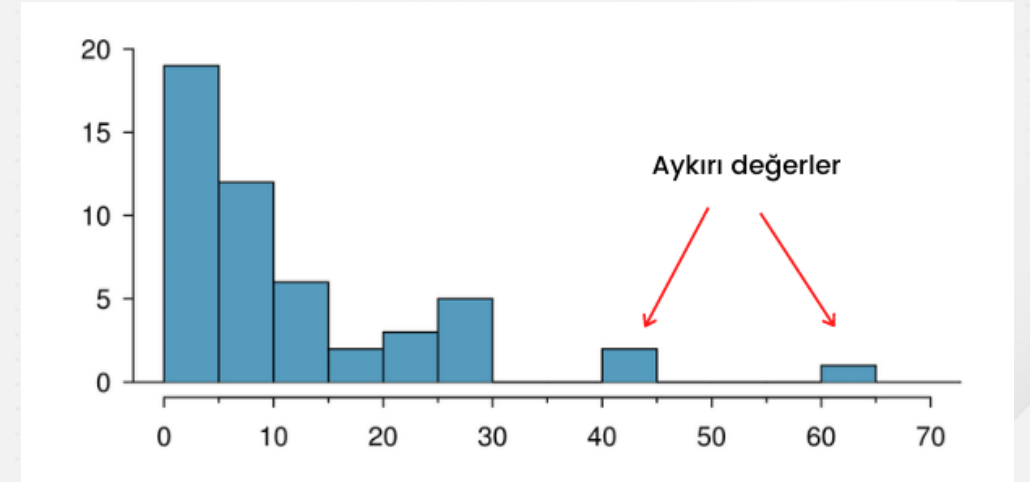
Kutu Grafiği:

- Kutu grafiği I. ve III. çeyrekler ile bunların arasındaki mesafeyi aykırı değer eşiği belirlemede kullanır. Bu eşiğin dışındakiler aykırıdır.



Histogram:

- Sayısal verilerin dağılımı histogram grafiğinde görselleştirilir ve aykırı değerler grafikte uçlarda görünür.



5.5.1.2. İstatistikî Yöntemler

Z Skoru:

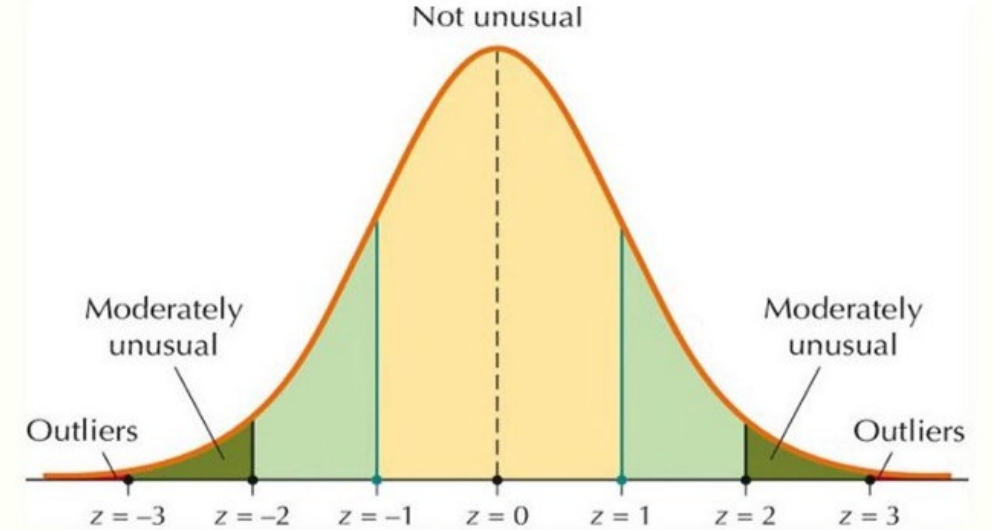
- Z-skoru, veri noktasının ortalamaya olan standart sapma birimleri cinsinden uzaklığını ölçer. Genellikle Z-skoru 3 veya -3'ten büyük olan noktalar aykırı değer olarak kabul edilir.

Ortanca Mutlak Sapma (Median Absolute Deviation – MAD):

- Aykırı değerler, verisetinin standart sapma ve ortanca (medyan) değerlerini etkiler. Bunu onarmak için aşağıdaki adımlar izlenir:
- Verisetinin ortanca değeri bulunur.
- Bulunan ortancanın her bir veriye olan mutlak sapması hesaplanır.
- Bu mutlak sapmaların ortancası alınır.

$$MAD = \text{ortanca}(\text{ortanca}(X)|)$$

Detecting Outliers with z-Scores



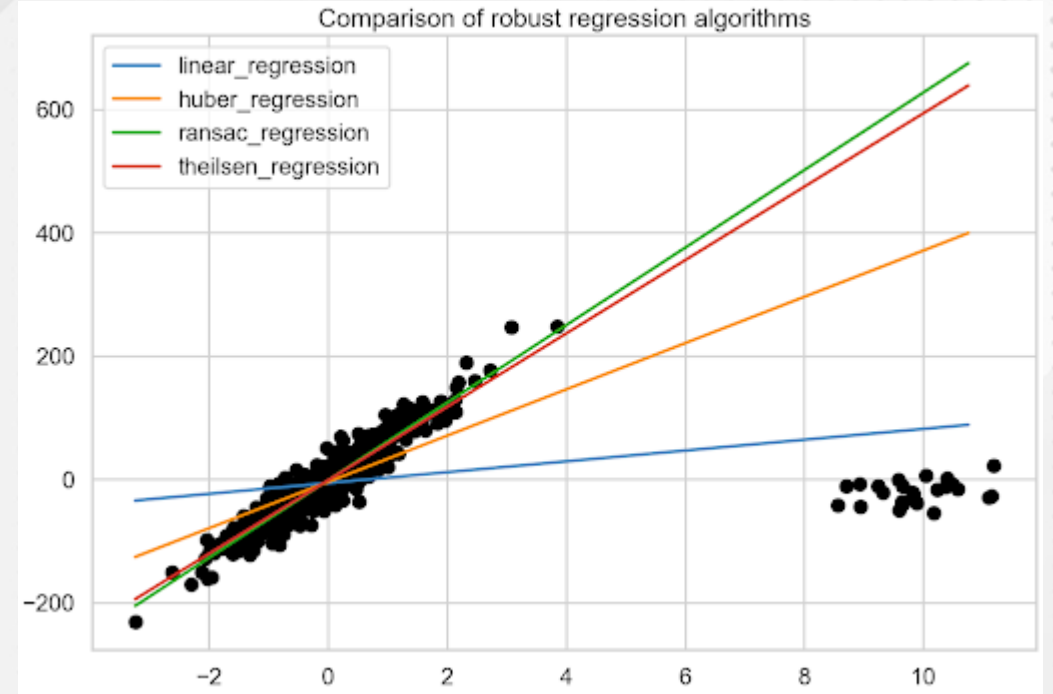
5.5.1.3. Makine Öğrenmesi Yöntemleri

Baskılayıcı (Suppressor) Modeller:

- Baskılayıcı (Suppressor) Modeller, aykırı değerlerin etkisini azaltmada. Örneğin, Robust Regression gibi baskılayıcı modeller, aykırı değerlere karşı daha dirençli olabilir.

Gözlemler Arası Uzaklık (Observation Distance):

- Bazı algoritmalar, özellikle kümeleme (clustering) ve anomali tespiti (anomaly detection) gibi, gözlemler arasındaki uzaklıkları hesaplayarak aykırı değerleri tespit edebilir.



HAFTA 5. MAKİNE ÖĞRENMESİ

Bölüm Özeti

- Karar ağacı, sınıflandırma ve regresyon problemlerini çözmek için kullanılan bir makine öğrenmesi algoritmasıdır. Veri setini, belirli özellikler üzerinden dallara ayrılan bir ağaç yapısına böler. Her bir iç düğüm, bir özelliğin testini temsil eder, her dal testin sonucunu ve her yaprak düğüm ise bir sınıf etiketini veya sürekli bir değeri temsil eder.
- Rastgele Orman (Random Forest), birden fazla karar ağacının birleşiminden oluşan bir topluluk öğrenme yöntemidir. Her bir karar ağacı, eğitim veri setinin rastgele alt örnekleri ve özellikleri kullanılarak oluşturulur.
- Karışıklık matrisi, sınıflandırma modellerinin performansını değerlendirmek için kullanılan bir tablodur. Matris, modelin tahmin ettiği ve gerçek sınıf etiketlerini karşılaştırarak doğru ve yanlış tahminlerin sayısını gösterir. Karışıklık matrisi dört ana bileşenden oluşur: True Positive, True Negative, False Positive, False Negative. Bu bileşenler, doğruluk, hassasiyet, özgüllük ve F1 skoru gibi performans metriklerinin hesaplanmasında kullanılır.
- Overfitting (aşırı uyum), bir makine öğrenmesi modelinin eğitim verilerine fazla uyum sağlaması durumudur. Bu, modelin eğitim verilerindeki gürültü ve rastlantısal desenleri öğrenmesine yol açar, bu da modelin yeni, görülmemiş verilere genelleme yapma yeteneğini azaltır.
- Outlier (aykırı değer), veri kümesindeki diğer gözlemlerden önemli ölçüde farklı olan veri noktalarını ifade eder. Aykırı değerler, modelin performansını ve doğruluğunu olumsuz etkileyebilir, çünkü model bu uç değerler nedeniyle yanlış desenler öğrenebilir.

Değerlendirme Soruları

- Karar ağacının yapısında her bir iç düğüm, hangi tür bilgiyi temsil eder ve her bir yaprak düğüm neyi temsil eder?
- Rastgele Orman yönteminde, her bir karar ağacının nasıl oluşturulduğunu ve bu ağaçların nasıl bir araya getirildiğini açıklayınız.
- Karışıklık matrisinde yer alan dört ana bileşeni tanımlayın ve her bir bileşenin sınıflandırma performans metriklerinin hesaplanmasındaki rolünü açıklayınız.
- Overfitting nedir ve bir modelin aşırı uyum sağladığı durumlarda ne gibi sorunlar ortaya çıkabilir? Ayrıca, bu sorunu önlemek için hangi stratejiler kullanılabilir açıklayınız.