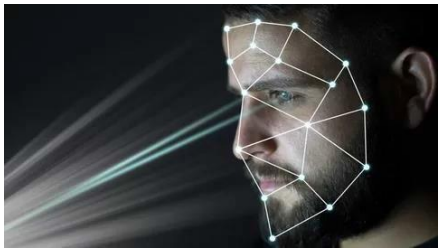
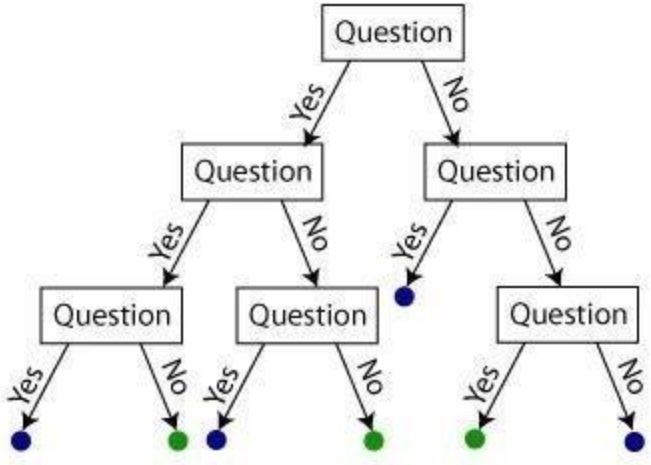


Makine Öğrenmesi ve İmge Tanıma

Birleştirme Modelleri

BÖLÜMLEME KARARI



BAŞARILI - BAŞARISIZ

Araştırma: Gelir Dağılımındaki Eşitsizlik Değerleri (Gini Katsayıları)

1. Gini Endeksi

Gini: «Bir popülasyondan rastgele iki öğe seçildiğinde, aynı sınıfta olmaları gerekir»
Bu durumda popülasyonlardan birinin saf olması olasılığı yüksektir.

"Başarı" (success) veya "Başarısızlık" (failure) gibi kategorik hedef değişkenle çalışır.

Yalnızca İkili bölümleme yapabilir.

Gini değeri arttıkça, homojenlik değeri de artar.

Gini yöntemi, CART (Sınıflandırma ve Regresyon Ağacı) tarafından ikili bölünmeler yapmak için kullanılır.

GİNİ ENDEKSİ

BÖLÜMLEME KARARI

1. Gini Endeksi

Başarı ve başarısızlık için olasılık karesinin toplamını hesapla, yani (p^2+q^2)

Şimdi oluşturulan bölümün her bir düğümünün ağırlıklı Gini bölünümü için Gini'yi hesaplayın.

	elma	muz	brokoli
Adet	= 4	4	4
p	= 4/12	4/12	4/12
	= 1/3	1/3	1/3

GI = 1 - [(1/3)^2 + (1/3)^2 + (1/3)^2]
= 1 - [1/9 + 1/9 + 1/9]
= 1 - 1/3
= 2/3
= 0.667

GİNİ ENDEKSİ

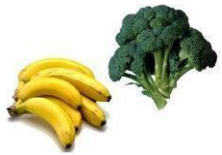
BÖLÜMLEME KARARI

Gini değeri arttıkça, homojenlik değeri de artar.

1. Gini Endeksi

Başarı ve başarısızlık için olasılık karesinin toplamını hesapla, yani (p^2+q^2)

Şimdi oluşturulan bölümün her bir düğümünün ağırlıklı Gini bölünümü için Gini'yi hesaplayın.



	elma	muz	brokoli
Adet =	4	4	4
p =	4/12	4/12	4/12
=	1/3	1/3	1/3

GI = 1 - [(1/3)^2 + (1/3)^2 + (1/3)^2]
= 1 - [1/9 + 1/9 + 1/9]
= 1 - 1/3
= 2/3
= 0.667

	elma	muz	brokoli
Adet =	3	3	6
p =	3/12	3/12	6/12
=	1/4	1/4	1/2

GI = 1 - [(1/4)^2 + (1/4)^2 + (1/2)^2]
= 1 - [1/16 + 1/16 + 1/4]
= 1 - 6/16
= 10/16
= 0.625

GİNİ ENDEKSİ

BÖLÜMLEME KARARI

1. Gini Endeksi

Başarı ve başarısızlık için olasılık karesinin toplamını hesapla, yani (p^2+q^2)

Şimdi oluşturulan bölümün her bir düğümünün ağırlıklı Gini bölünümü için Gini'yi hesaplayın.

	elma	muz	brokoli
Adet =	4	4	4
p =	4/12	4/12	4/12
=	1/3	1/3	1/3

GI = 1 - [(1/3)^2 + (1/3)^2 + (1/3)^2]
= 1 - [1/9 + 1/9 + 1/9]
= 1 - 1/3
= 2/3
= 0.667

	elma	muz	brokoli
Adet =	3	3	6
p =	3/12	3/12	6/12
=	1/4	1/4	1/2

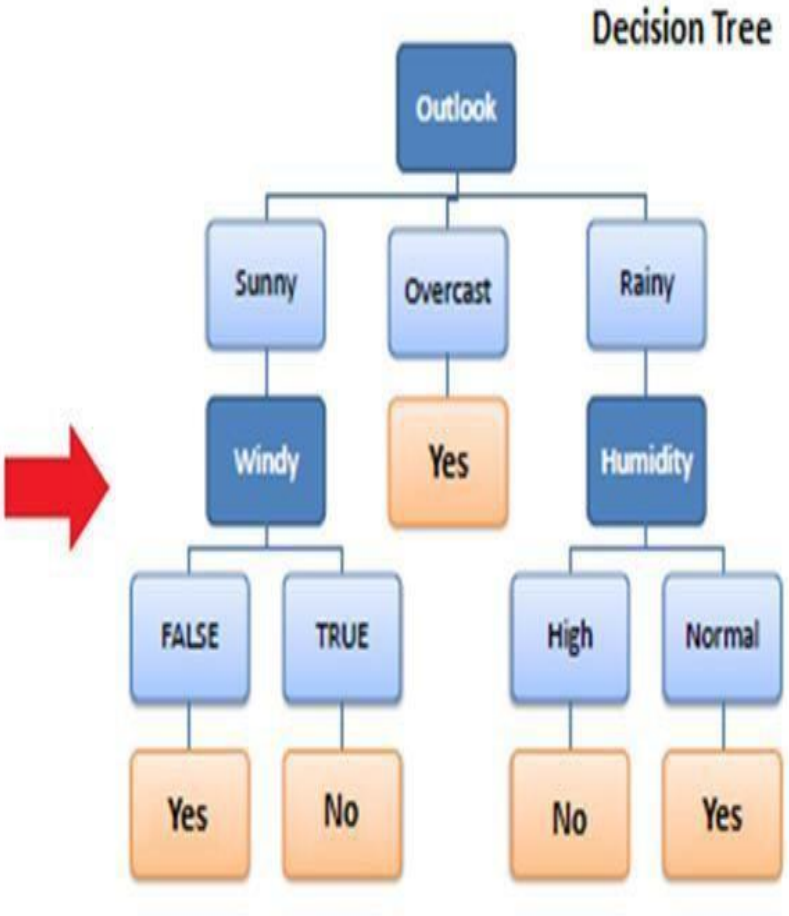
GI = 1 - [(1/4)^2 + (1/4)^2 + (1/2)^2]
= 1 - [1/16 + 1/16 + 1/4]
= 1 - 6/16
= 10/16
= 0.625

	elma/brokoli	muz
Adet =	0	12
p =	0/12	12/12
=	0	1

GI = 1 - [0^2 + 1^2 + 0^2]
= 1 - [0 + 1 + 0]
= 1 - 1
= 0.00

GİNİ ENDEKSİ

Predictors				Target
Outlook	Temp	Humidity	Windy	Play Golf
Rainy	Hot	High	False	No
Rainy	Hot	High	True	No
Overcast	Hot	High	False	Yes
Sunny	Mild	High	False	Yes
Sunny	Cool	Normal	False	Yes
Sunny	Cool	Normal	True	No
Overcast	Cool	Normal	True	Yes
Rainy	Mild	High	False	No
Rainy	Cool	Normal	False	Yes
Sunny	Mild	Normal	False	Yes
Rainy	Mild	Normal	True	Yes
Overcast	Mild	High	True	Yes
Overcast	Hot	Normal	False	Yes
Sunny	Mild	High	True	No



GİNİ ENDEKSİ

Predictors				Target
Outlook	Temp	Humidity	Windy	Play Golf
Rainy	Hot	High	False	No
Rainy	Hot	High	True	No
Overcast	Hot	High	False	Yes
Sunny	Mild	High	False	Yes
Sunny	Cool	Normal	False	Yes
Sunny	Cool	Normal	True	No
Overcast	Cool	Normal	True	Yes
Rainy	Mild	High	False	No
Rainy	Cool	Normal	False	Yes
Sunny	Mild	Normal	False	Yes
Rainy	Mild	Normal	True	Yes
Overcast	Mild	High	True	Yes
Overcast	Hot	Normal	False	Yes
Sunny	Mild	High	True	No

		play		
		yes	no	total
	sunny	3	2	5
Outlook	overcast	4	0	4
	rainy	2	3	5
				14

https://sefiks.com/2018/08/27/a-step-by-step-cart-decision-tree-ex ample/

GİNİ ENDEKSİ

	Predictors			Target
Outlook	Temp	Humidity	Windy	Play Golf
Rainy	Hot	High	False	No
Rainy	Hot	High	True	No
Overcast	Hot	High	False	Yes
Sunny	Mild	High	False	Yes
Sunny	Cool	Normal	False	Yes
Sunny	Cool	Normal	True	No
Overcast	Cool	Normal	True	Yes
Rainy	Mild	High	False	No
Rainy	Cool	Normal	False	Yes
Sunny	Mild	Normal	False	Yes
Rainy	Mild	Normal	True	Yes
Overcast	Mild	High	True	Yes
Overcast	Hot	Normal	False	Yes
Sunny	Mild	High	True	No

		play		
		yes	no	total
Outlook	sunny	3	2	5
	overcast	4	0	4
	rainy	2	3	5
				14

Gini(Outlook=Rainy)=

$$1 - (2/5)^2 - (3/5)^2 = 1 - 0.16 - 0.36 = 0.48$$

(!)DİKKAT(!)

Gini Impurity = 1-Gini

https://sefiks.com/2018/08/27/a-step-by-step-cart-decision-tree-ex ample/

GİNİ ENDEKSİ

Predictors				Target
Outlook	Temp	Humidity	Windy	Play Golf
Rainy	Hot	High	False	No
Rainy	Hot	High	True	No
Overcast	Hot	High	False	Yes
Sunny	Mild	High	False	Yes
Sunny	Cool	Normal	False	Yes
Sunny	Cool	Normal	True	No
Overcast	Cool	Normal	True	Yes
Rainy	Mild	High	False	No
Rainy	Cool	Normal	False	Yes
Sunny	Mild	Normal	False	Yes
Rainy	Mild	Normal	True	Yes
Overcast	Mild	High	True	Yes
Overcast	Hot	Normal	False	Yes
Sunny	Mild	High	True	No

		play		
		yes	no	total
Outlook	sunny	3	2	5
	overcast	4	0	4
	rainy	2	3	5
				14

Gini(Outlook=Rainy)=

$$1 - (2/5)^2 - (3/5)^2 = 1 - 0.16 - 0.36 = 0.48$$

$$\text{Gini(Outlook=Overcast)} = 1 - (4/4)^2 - (0/4)^2 = 0$$

Gini(Outlook=Sunny)=

$$1 - (3/5)^2 - (2/5)^2 = 1 - 0.36 - 0.16 = 0.48$$

https://sefiks.com/2018/08/27/a-step-by-step-cart-decision-tree-ex ample/

GİNİ ENDEKSİ

Predictors				Target
Outlook	Temp	Humidity	Windy	Play Golf
Rainy	Hot	High	False	No
Rainy	Hot	High	True	No
Overcast	Hot	High	False	Yes
Sunny	Mild	High	False	Yes
Sunny	Cool	Normal	False	Yes
Sunny	Cool	Normal	True	No
Overcast	Cool	Normal	True	Yes
Rainy	Mild	High	False	No
Rainy	Cool	Normal	False	Yes
Sunny	Mild	Normal	False	Yes
Rainy	Mild	Normal	True	Yes
Overcast	Mild	High	True	Yes
Overcast	Hot	Normal	False	Yes
Sunny	Mild	High	True	No

		play		
		yes	no	total
Outlook	sunny	3	2	5
	overcast	4	0	4
	rainy	2	3	5
				14

Gini(Outlook=Rainy)=

$$1 - (2/5)^2 - (3/5)^2 = 1 - 0.16 - 0.36 = 0.48$$

$$\text{Gini(Outlook=Overcast)} = 1 - (4/4)^2 - (0/4)^2 = 0$$

Gini(Outlook=Sunny)=

$$1 - (3/5)^2 - (2/5)^2 = 1 - 0.36 - 0.16 = 0.48$$

Ağırlıklı hesap

$$\text{Gini(Outlook)} = (5/14) \times 0.48 + (4/14) \times 0 + (5/14) \times 0.48 = 0.171 + 0 + 0.171 = 0.342$$

https://sefiks.com/2018/08/27/a-step-by-step-cart-decision-tree-ex ample/

GİNİ ENDEKSİ

Predictors				Target
Outlook	Temp	Humidity	Windy	Play Golf
Rainy	Hot	High	False	No
Rainy	Hot	High	True	No
Overcast	Hot	High	False	Yes
Sunny	Mild	High	False	Yes
Sunny	Cool	Normal	False	Yes
Sunny	Cool	Normal	True	No
Overcast	Cool	Normal	True	Yes
Rainy	Mild	High	False	No
Rainy	Cool	Normal	False	Yes
Sunny	Mild	Normal	False	Yes
Rainy	Mild	Normal	True	Yes
Overcast	Mild	High	True	Yes
Overcast	Hot	Normal	False	Yes
Sunny	Mild	High	True	No

Temperature	Yes	No	Number of instances
Hot	2	2	4
Cool	3	1	4
Mild	4	2	6

$$Gini(Temp=Hot) = 1 - (2/4)^2 - (2/4)^2 = 0.5$$

$$Gini(Temp=Cool) = 1 - (3/4)^2 - (1/4)^2 = 1 - 0.5625 - 0.0625 = 0.375$$

$$Gini(Temp=Mild) = 1 - (4/6)^2 - (2/6)^2 = 1 - 0.444 - 0.111 = 0.445$$

$$Gini(Temp) = (4/14) \times 0.5 + (4/14) \times 0.375 + (6/14) \times 0.445 = 0.142 + 0.107 + 0.190 = 0.439$$

$$Gini(Outlook) = (5/14) \times 0.48 + (4/14) \times 0 + (5/14) \times 0.48 = 0.171 + 0 + 0.171 = 0.342$$

https://sefiks.com/2018/08/27/a-step-by-step-cart-decision-tree-ex ample/

GİNİ ENDEKSİ

Predictors | Target

Outlook	Temp	Humidity	Windy	Play Golf
Rainy	Hot	High	False	No
Rainy	Hot	High	True	No
Overcast	Hot	High	False	Yes
Sunny	Mild	High	False	Yes
Sunny	Cool	Normal	False	Yes
Sunny	Cool	Normal	True	No
Overcast	Cool	Normal	True	Yes
Rainy	Mild	High	False	No
Rainy	Cool	Normal	False	Yes
Sunny	Mild	Normal	False	Yes
Rainy	Mild	Normal	True	Yes
Overcast	Mild	High	True	Yes
Overcast	Hot	Normal	False	Yes
Sunny	Mild	High	True	No

Humidity	Yes	No	Number of instances
High	3	4	7
Normal	6	1	7

$Gini(Humidity=High) = 1 - (3/7)^2 - (4/7)^2 = 1 - 0.183 - 0.326 = 0.489$

$Gini(Humidity=Normal) = 1 - (6/7)^2 - (1/7)^2 = 1 - 0.734 - 0.02 = 0.244$

$Gini(Humidity) = (7/14) \times 0.489 + (7/14) \times 0.244 = 0.367$

Wind	Yes	No	Number of instances
Weak	6	2	8
Strong	3	3	6

$Gini(Wind=Weak) = 1 - (6/8)^2 - (2/8)^2 = 1 - 0.5625 - 0.062 = 0.375$

$Gini(Wind=Strong) = 1 - (3/6)^2 - (3/6)^2 = 1 - 0.25 - 0.25 = 0.5$

$Gini(Wind) = (8/14) \times 0.375 + (6/14) \times 0.5 = 0.428$

$Gini(Temp) = (4/14) \times 0.5 + (4/14) \times 0.375 + (6/14) \times 0.445 = 0.142 + 0.107 + 0.190 = 0.439$

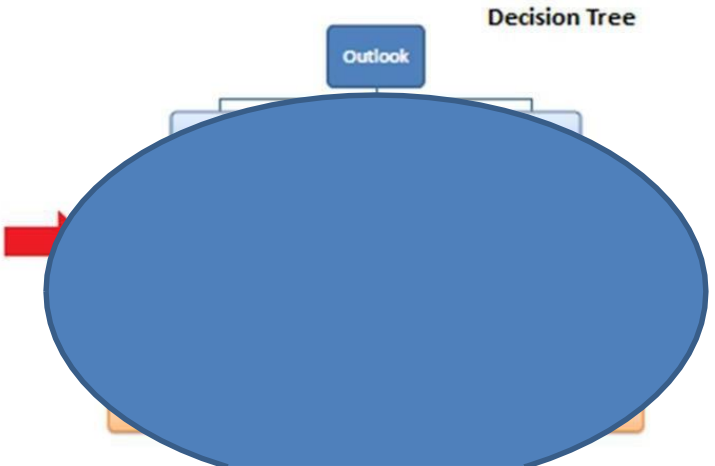
$Gini(Outlook) = (5/14) \times 0.48 + (4/14) \times 0 + (5/14) \times 0.48 = 0.171 + 0 + 0.171 = 0.342$

https://sefiks.com/2018/08/27/a-step-by-step-cart-decision-tree-ex ample/

GİNİ ENDEKSİ

Predictors				Target
Outlook	Temp	Humidity	Windy	Play Golf
Rainy	Hot	High	False	No
Rainy	Hot	High	True	No
Overcast	Hot	High	False	Yes
Sunny	Mild	High	False	Yes
Sunny	Cool	Normal	False	Yes
Sunny	Cool	Normal	True	No
Overcast	Cool	Normal	True	Yes
Rainy	Mild	High	False	No
Rainy	Cool	Normal	False	Yes
Sunny	Mild	Normal	False	Yes
Rainy	Mild	Normal	True	Yes
Overcast	Mild	High	True	Yes
Overcast	Hot	Normal	False	Yes
Sunny	Mild	High	True	No

Gini index
0.342
0.439
0.367
0.428



https://sefiks.com/2018/08/27/a-step-by-step-cart-decision-tree-ex ample/

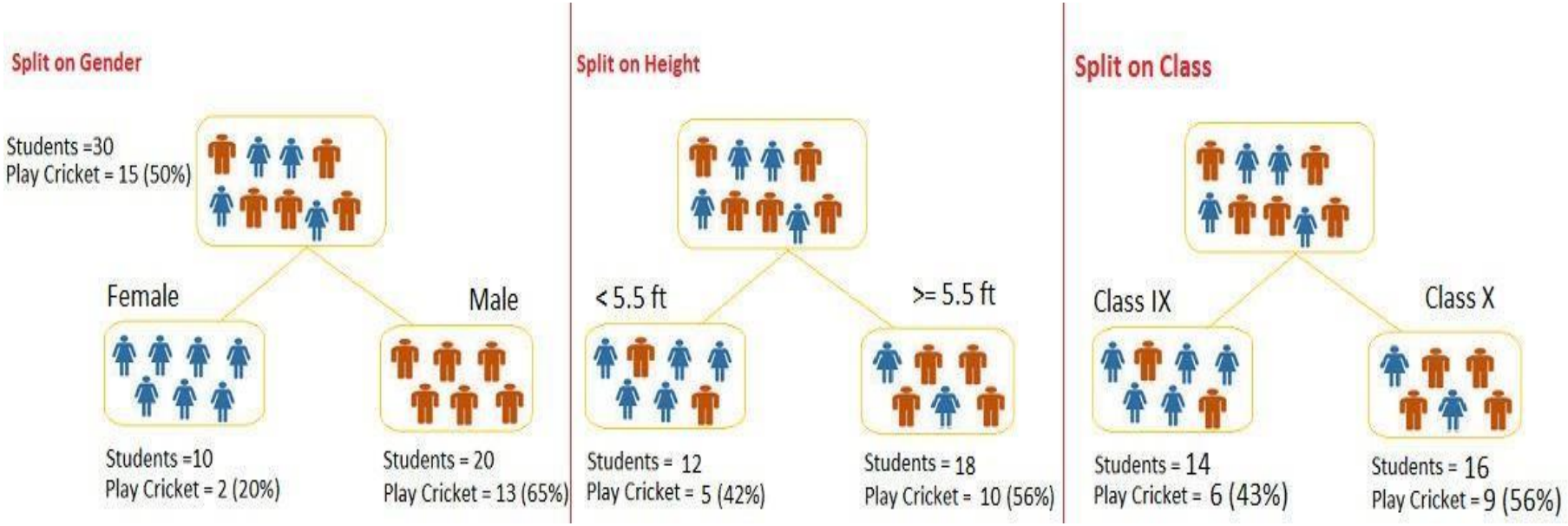
Decision Making Quotes

It is better to be approximately right than precisely wrong

Warren Buffet

KARAR AĞAÇLARI

- Üç farklı özellikle tanımlanan 30 çocuğa bakalım.
  - *Cinsiyet (E veya K),*
  - *Sınıf(IX. veya X.)*
- *Boy (1.50 – 1.80 cm)*.
- Peki üç öznitelikten en belirgin/belirleyici olan girdi değişkeni hangisi?



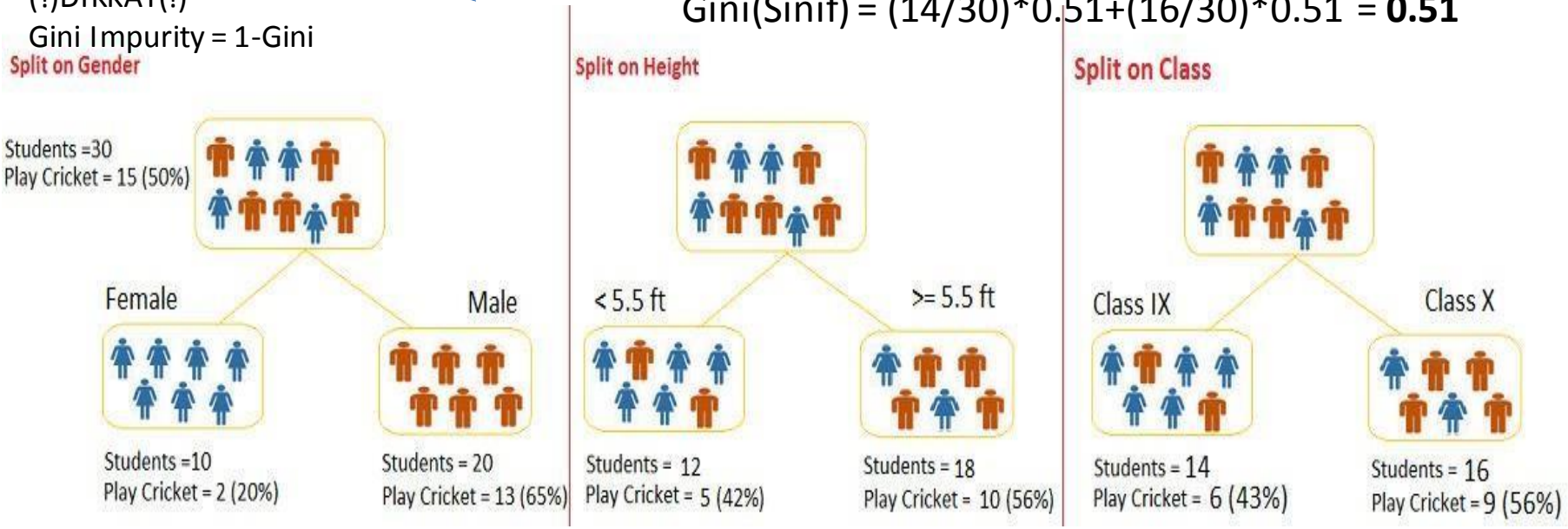
KARAR AĞAÇLARI

- Üç farklı özellikle tanımlanan 30 çocuğa bakalım.
  - *Cinsiyet (E veya K),*
  - *Sınıf(IX. veya X.)*
  - *Boy (1.50 – 1.80 cm).*
- Peki üç öznitelikten en belirgin/belirleyici olan girdi değişkeni hangisi?

(!)DİKKAT(!)
Gini Impurity = 1-Gini

Cinsiyet:
Gini(Kız) = (0.2)*(0.2)+(0.8)*(0.8)=0.68
Gini(Erkek) = (0.65)*(0.65)+(0.35)*(0.35)=0.55
Gini (Cinsiyet)= (10/30)*0.68+(20/30)*0.55 = **0.59**

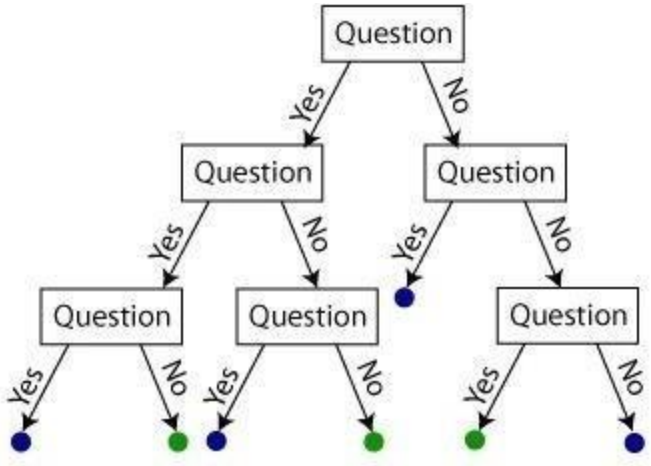
Sınıf:
Gini(IX) = (0.43)*(0.43)+(0.57)*(0.57)=0.51
Gini(X) = (0.56)*(0.56)+(0.44)*(0.44)=0.51
Gini(Sınıf) = (14/30)*0.51+(16/30)*0.51 = **0.51**



BÖLÜMLEME KARARI

1. **Gini Endeksi**
2. **Ki-Kare(Chi-Square):**

- üst düğüm ve alt düğüm arasındaki istatistiksel farkı bulmak için kullanılır. Hedef değişkenin beklenen frekansları ile hedef değişkenin gözlemlenen/elde edilen frekansları arasındaki tüm farkların karelerinin toplamı ile hesaplanır.
- Ki-Karenin **değeri arttıkça**, ana düğüm ile alt düğüm arasındaki **istatistiksel fark değeri** de artar.
- **Ki kare = ((Gerçek - Beklenen) ^{2} / Beklenen)$^{1/2}$**



BAŞARILI - BAŞARISIZ

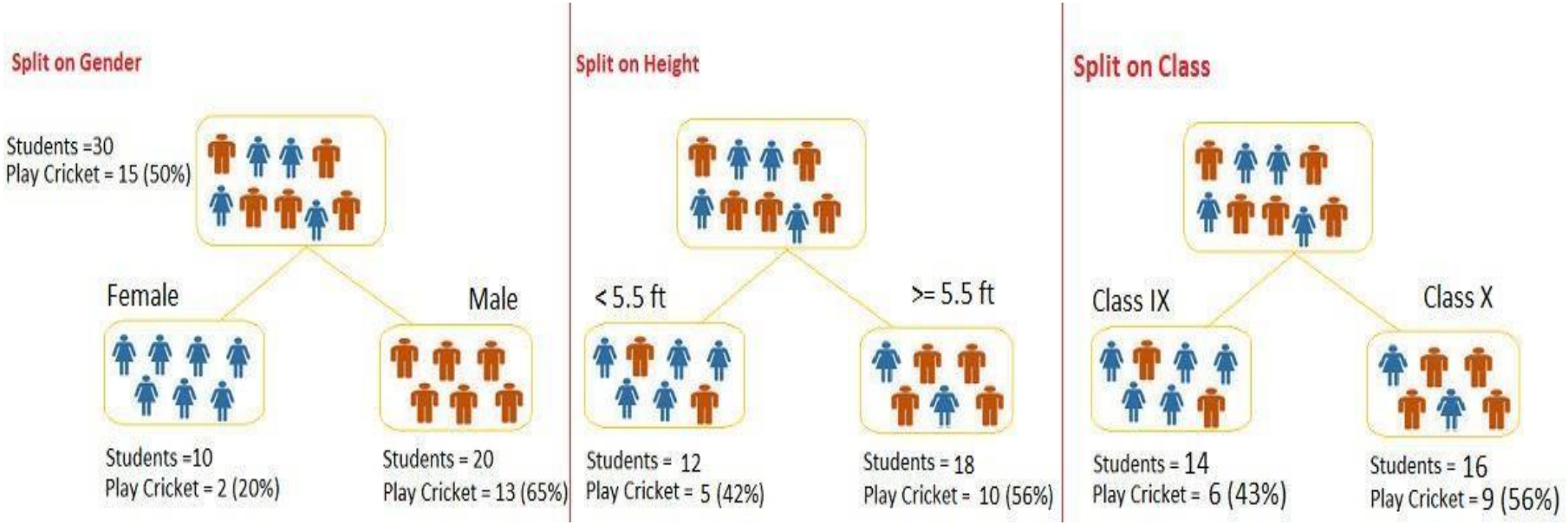
3. **Varyans**

$$\text{Variance} = \frac{\Sigma(X - \overline{X})^2}{n}$$

$\bar{X}$: X in ortalaması

Kİ-KARE

- Üç farklı özellikle tanımlanan 30 çocuğa bakalım.
  - *Cinsiyet (E veya K),*
  - *Sınıf(IX. veya X.)*
  - *Boy (1.50 – 1.80 cm).*
- Peki üç öznitelikten en belirgin/belirleyici olan girdi değişkeni hangisi?



Kİ-KARE

- Üç farklı özellikle tanımlanan

30 çocuğa bakalım.

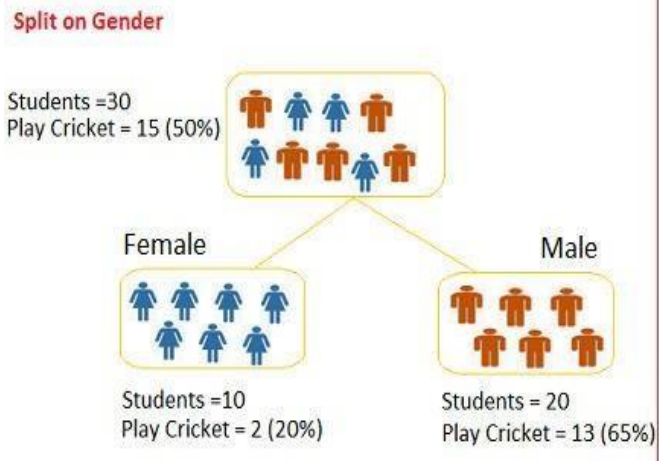
  - *Cinsiyet (E veya K),*
  - *Sınıf(IX. veya X.)*
  - *Boy (1.50 – 1.80 cm).*

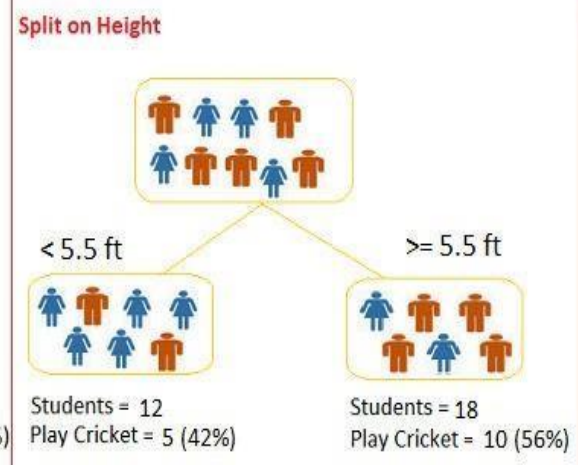
- Peki üç öznitelikten en belirgin/belirleyici olan girdi değişkeni hangisi?

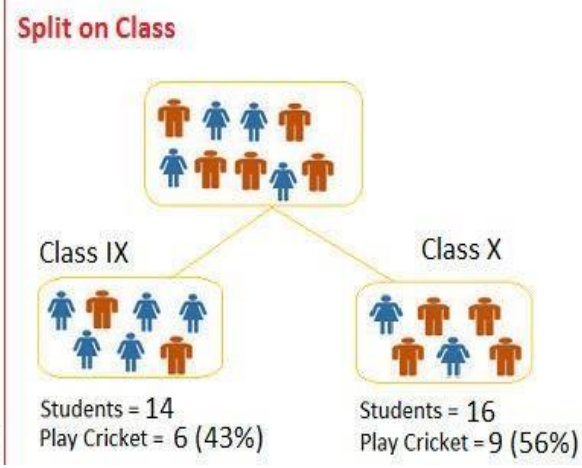
Node	Play Cricket	Not Play Cricket	Total	Expected Play Cricket	Expected Not Play Cricket	Deviation Play Cricket	Deviation Not Play Cricket	Chi-Square	
								Play Cricket	Not Play Cricket
Female	2	8	10	5	5	-3	3	1.34	1.34
Male	13	7	20	10	10	3	-3	0.95	0.95
							Total Chi-Square	4.58	

Node	Play Cricket	Not Play Cricket	Total	Expected Play Cricket	Expected Not Play Cricket	Deviation Play Cricket	Deviation Not Play Cricket	Chi-Square	
								Play Cricket	Not Play Cricket
IX	6	8	14	7	7	-1	1	0.38	0.38
X	9	7	16	8	8	1	-1	0.35	0.35
							Total Chi-Square	1.46	

Ki kare = ((Gerçek - Beklenen)^2 / Beklenen) $^{1/2}$







Kİ-KARE

Node	Play Cricket	Not Play Cricket	Total	Expected Play Cricket	Expected Not Play Cricket	Deviation Play Cricket	Deviation Not Play Cricket	Chi-Square	
								Play Cricket	Not Play Cricket
Female	2	8	10	5	5	-3	3	1.34	1.34
Male	13	7	20	10	10	3	-3	0.95	0.95
							Total Chi-Square	4.58	

1

2

- **Ki kare = ((Gerçek - Beklenen)^2 / Beklenen) $^{1/2}$**

Node	Play Cricket	Not Play Cricket	Total	Expected Play Cricket	Expected Not Play Cricket	Deviation Play Cricket	Deviation Not Play Cricket	Chi-Square	
								Play Cricket	Not Play Cricket
IX	6	8	14	7	7	-1	1	0.38	0.38
X	9	7	16	8	8	1	-1	0.35	0.35
							Total Chi-Square	1.46	

=(2-5)^2/ Beklenen) ^{½}

= (9/5)$^{1/2}$ =1,8 ^{½} =1,34

https://medium.com/@rishabhjain_22692/decision-trees-it-begins-here-93ff54ef134

AVANTAJLARI

Anlaması kolay:

Analitik altyapısı olmayan insanlar için bile, karar ağacı algoritmasının anlaşılması çok kolaydır. Bir kullanıcının ağaçları incelemesi, okuması ve yorumlaması için herhangi bir istatistiksel bilgi veya bilgiye sahip olması gerekmez. Kullanıcılar verileri kolayca okuyabilir.

Grafik **gösterimi** son derece sezgisel (bütünsel) ve kullanıcı dostudur .

Veri araştırmalarında faydalı:

Karar ağacı, en hızlısı olmasa bile, kesinlikle en önemli değişkenin tanımlanmasının en hızlı yönteminden olduğuna inanılmaktadır. Karar ağacı, kullanıcıların, özelliklerin yanı sıra yeni değişkenler oluşturmalarına yardımcı olabilir. Bu yeni özellikler, hedef değişkeni tahmin etmek için daha fazla güce sahip olacaktır. Veri arama aşamasında da kullanılabilir.

AVANTAJLARI

Daha az veri temizliği gerekiyor:

makine öğreniminde, bir kullanıcı zamanının çoğunu veri temizlemeye ve iyi verileri kötü verilerden ayırmaya harcamak zorundadır. Bununla birlikte, karar ağacı söz konusu olduğunda, bu süreç oldukça kolaydır ve çok zaman almaz. Uç değerlerin yanı sıra aykırı değerlerden etkilenmez ve böylece temizleme işlemi kolaylaşır.

Veri türü bir kısıtlama değil:

Karar ağacı çok yönlü bir algoritmadır ve kategorik ve sayısal veri değişkenlerini kolaylıkla işleyebilir.

Parametrik Olmayan Yöntem:

Parametrik olmayan bir yöntem, sınıflandırıcı yapıları veya uzamsal dağılım ile ilgili varsayımları olmayan bir yöntem anlamına gelir. Karar ağacı parametrik olmayan bir yöntemdir.

NOT: parametrik yöntemler sınırlı sayıda parametre alır parametrik olmayan yöntemlerde veri arttıkça parametre artabilir.

DEZAVANTAJI

Aşırı uyum (overfitting) gösterme:

Karar ağacı modellerinde, aşırı uyumun en yaygın ve karşılaşılan problemdir. Modeli eldeki veriyle son derece yüksek başarılı sonuç üretecek şekilde oluşturur ama hiç görmediği yeni bir veriye çok yüksek hata verir. (Soruları ezberleme örneği) Bununla birlikte, bu sorun budamanın yanı sıra model parametreleri üzerindeki kısıtlamalar kullanılarak çözülebilir.

Sürekli değişkenler için uygun değil:

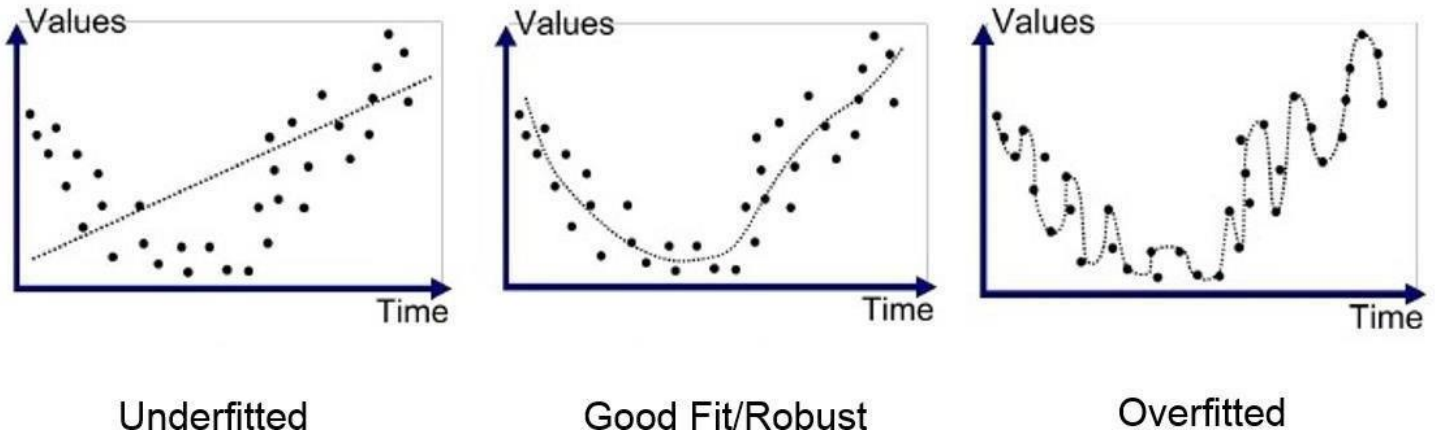
Sürekli değişkenlerle çalışabilse de, hiç uygun değildir. Karar ağacı, de değişkenleri gittikçe daha fazla sınıflandırmaya başladığında kategoride bilgileri kaybetmeye başlar.

NOT: Bağımlı değişken sürekli olduğunda, regresyon ağaçları kullanılırken bağımlı değişken kategorik olduğunda; sınıflandırma ağaçları kullanılır

AŞIRI UYUMU NASIL ÖNLERİZ

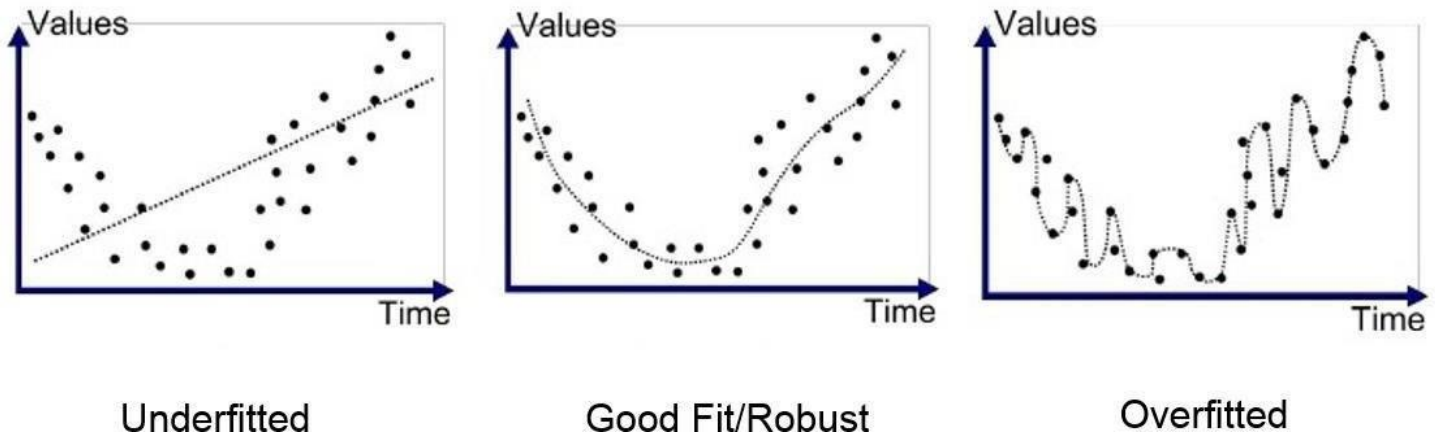
Aşırı uyum (overfitting) gösterme:

Karar ağacı modellerinde, aşırı uyumun en yaygın ve karşılaşılan problemdir. Modeli eldeki veriyle son derece yüksek başarılı sonuç üretecek şekilde oluşturur ama hiç görmediği yeni bir veriye çok yüksek hata verir. (Soruları ezberleme örneği) Bununla birlikte, bu sorun budamanın yanı sıra model parametreleri üzerindeki kısıtlamalar kullanılarak çözülebilir.



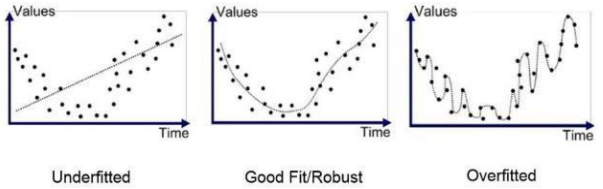
AŞIRI UYUMU NASIL ÖNLERİZ

- Ağaç boyutuna ilişkin kısıt koyma
- Ağaç budaması



AŞIRI UYUMU NASIL ÖNLERİZ

- Ağaç boyutuna ilişkin kısıt koyma



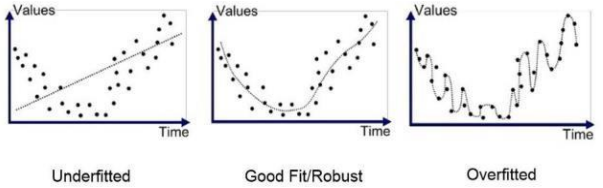
- Ağaç budaması

- Düğüm bölümlemesi için Minimum Örnekler (minimum samples).
- Bölümlemek için gereken gözlem adedi tanımlanabilir.
- Bir yaprak için minimum örnek tanımlanabilir.
- Azınlık sınıfının çoğunluk sınıfı olduğu bölgeler çok sınırlı olduğundan, dengesiz sınıf problemleri için daha düşük değerler seçilir.

Ağacın Maksimum Derinliği (maximum depth):

- Bu, bir ağacın derinliğini kontrol etmek için kullanılabilir.
- Çapraz kontrol (cross validation) kullanılarak ayarlanması gerekir.
- Çok sayıda terminal düğüm oluşur.

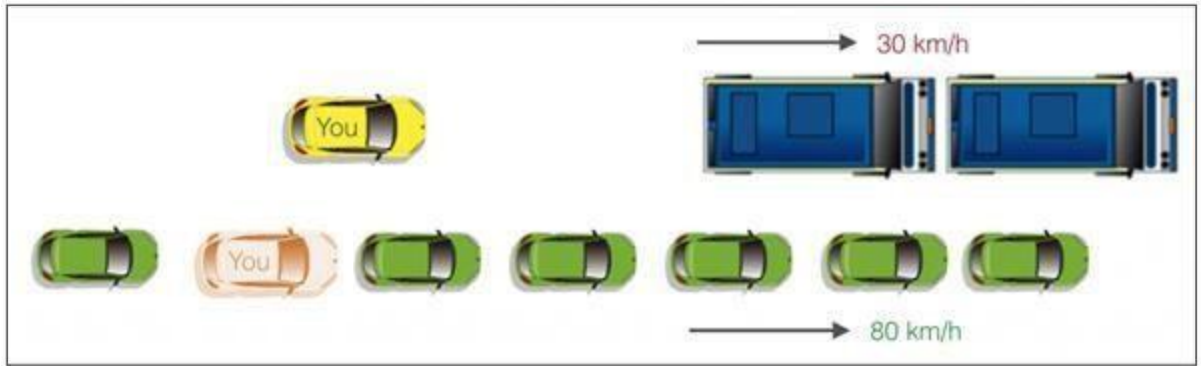
AŞIRI UYUMU NASIL ÖNLERİZ

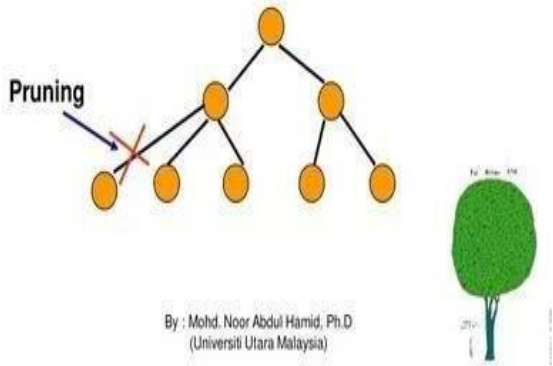


- Ağaç budaması

Açgözlü yaklaşım (greedy)

Yöntem, verilen en iyi durma koşullarından biri elde edilene kadar sadece en iyi ayrımı kontrol edecektir.





Gerçek hedef ne?
-geçmek mi?
-Mesafe almak mı?
- Maksimum yol kat etmek mi?

budama olsaydı, yöntem birkaç adım geri alacak ve üzerinde işlem yapmadan önce durumu düşünme şansına sahip olacaktı.


DATA BASE
OF USEFUL
INFORMATION
5K
DATA BASE
OF USELESS
INFORMATION
500,000,000 GB
CARTOONSTOCK
Search ID: tmcn188

TANIMLAR

Birleştirme (ensemble) yöntemleri:

Birleştirme yöntemleri, daha iyi kararlılık (stability)ve doğruluk(accuracy) elde edebilen bir grup öngörücü modelden oluşur.

Karar ağacı temelli modellerinin sapma ve varyans problemleri bulunmaktadır. Modelin karmaşıklığı arttığında, modeldeki düşük sapma sayesinde tahmin hatasındaki azalmayı görmek mümkündür.
Bununla birlikte, yavaş yavaş daha karmaşık bir model oluşturmaya devam ettiğinizde, modeliniz ezberleyebilir ve bu nedenle modeliniz yüksek varyans gösterebilir.

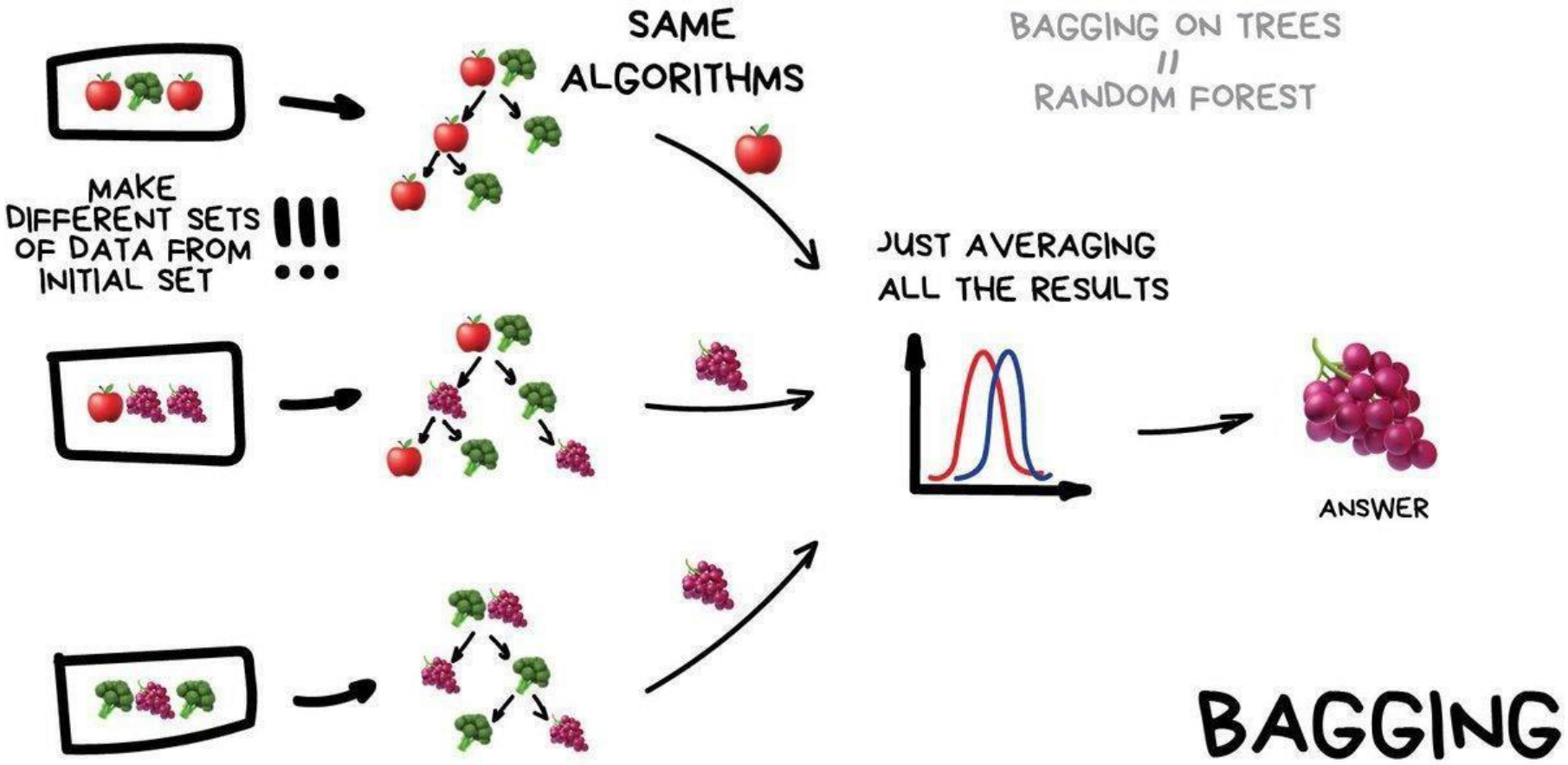
(Sapma, gerçek değerlerden tahmin edilen ortalama değerler arasındaki farkı ifade eder. Varyans, numuneler aynı popülasyondan alınırsa, aynı noktada modellerin tahminlerinin çeşitliliğini ifade eder.)

Birleştirme (ensemble) yöntemleri:

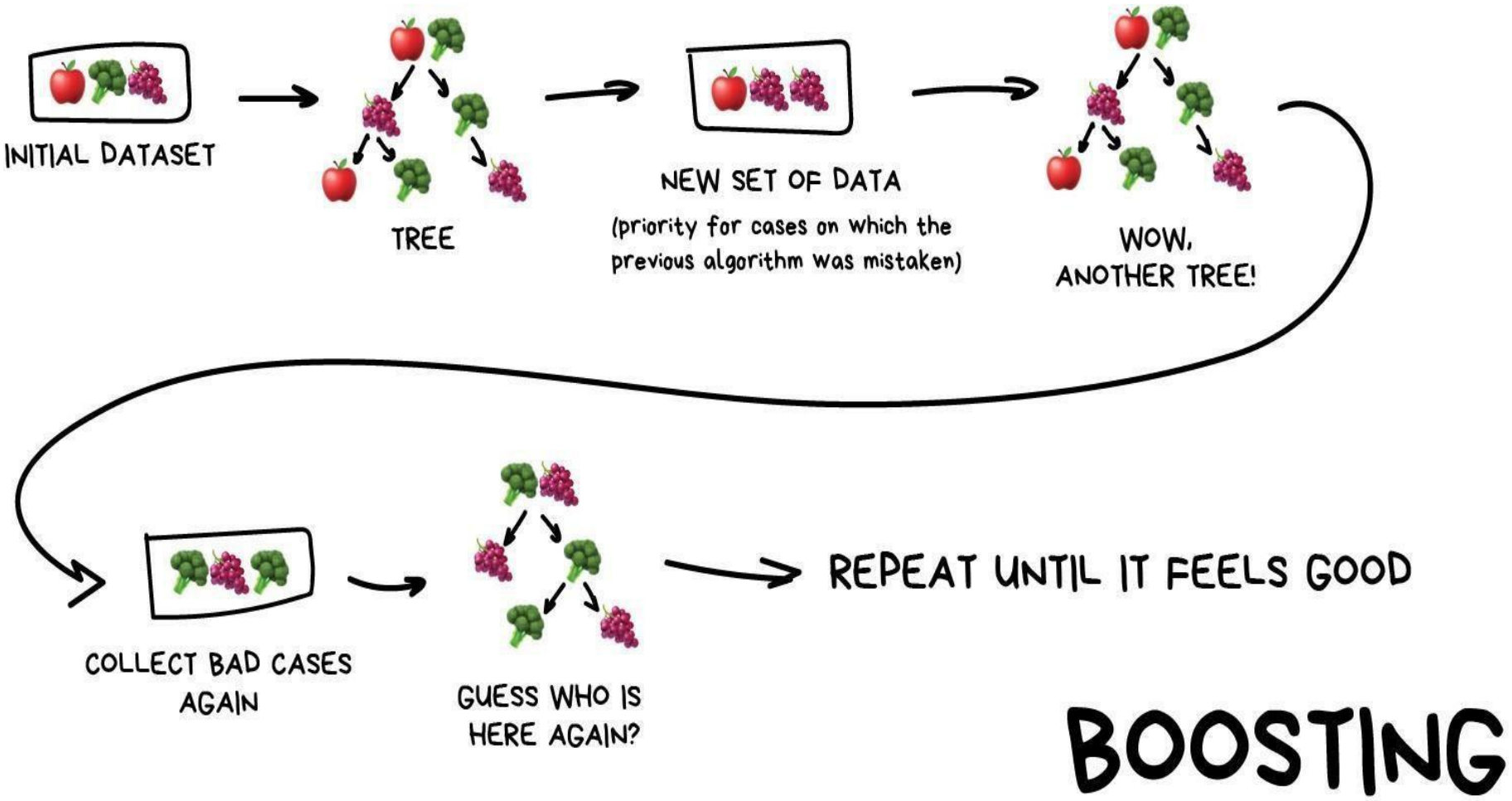
Birleştirme öğreniminin en yaygın yöntemleri şunlardır:

- Torbalama (bagging)
- Artırma (boosting)
- İstifleme (stacking)

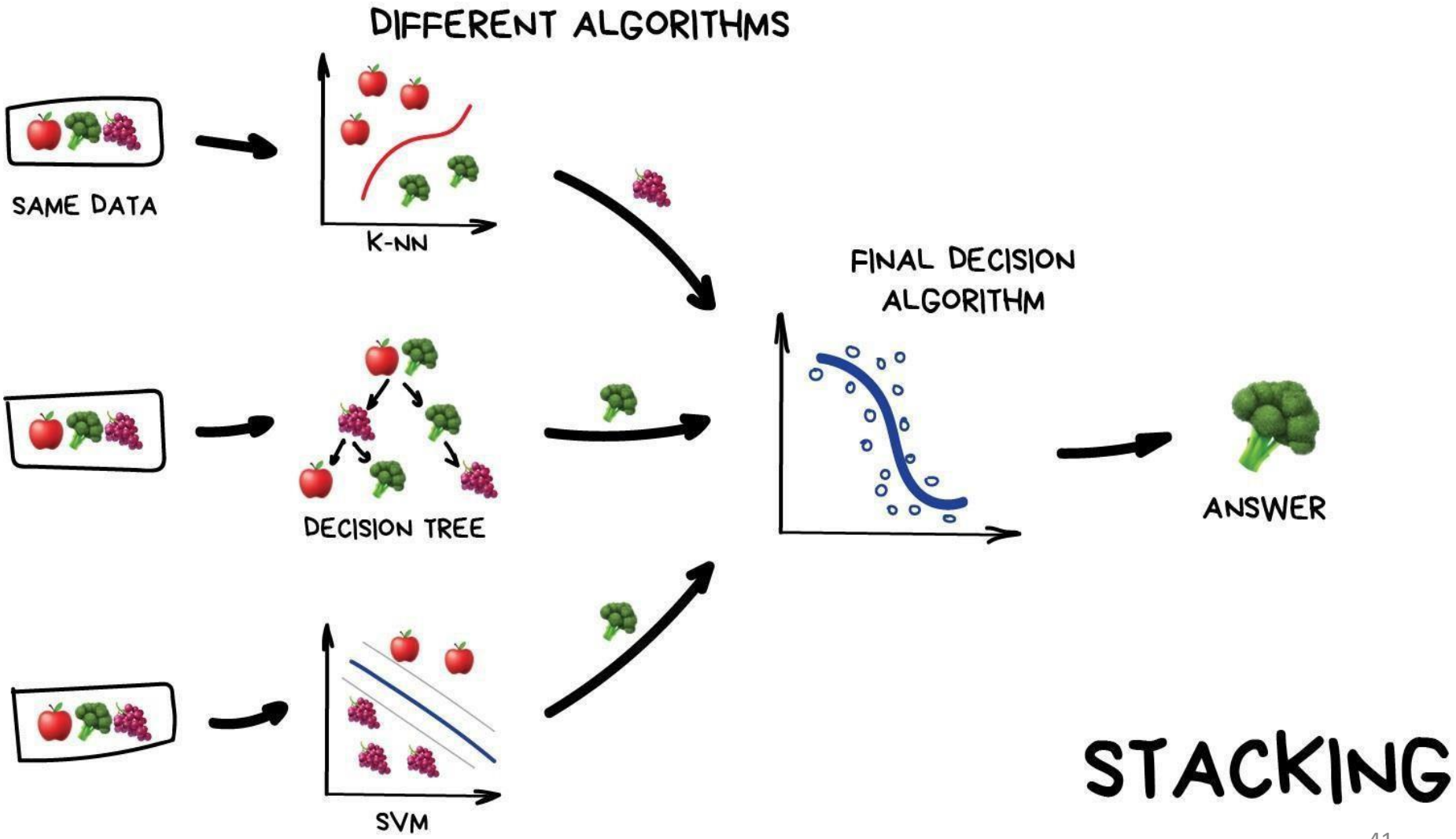
Birleştirme (ensemble) yöntemleri:



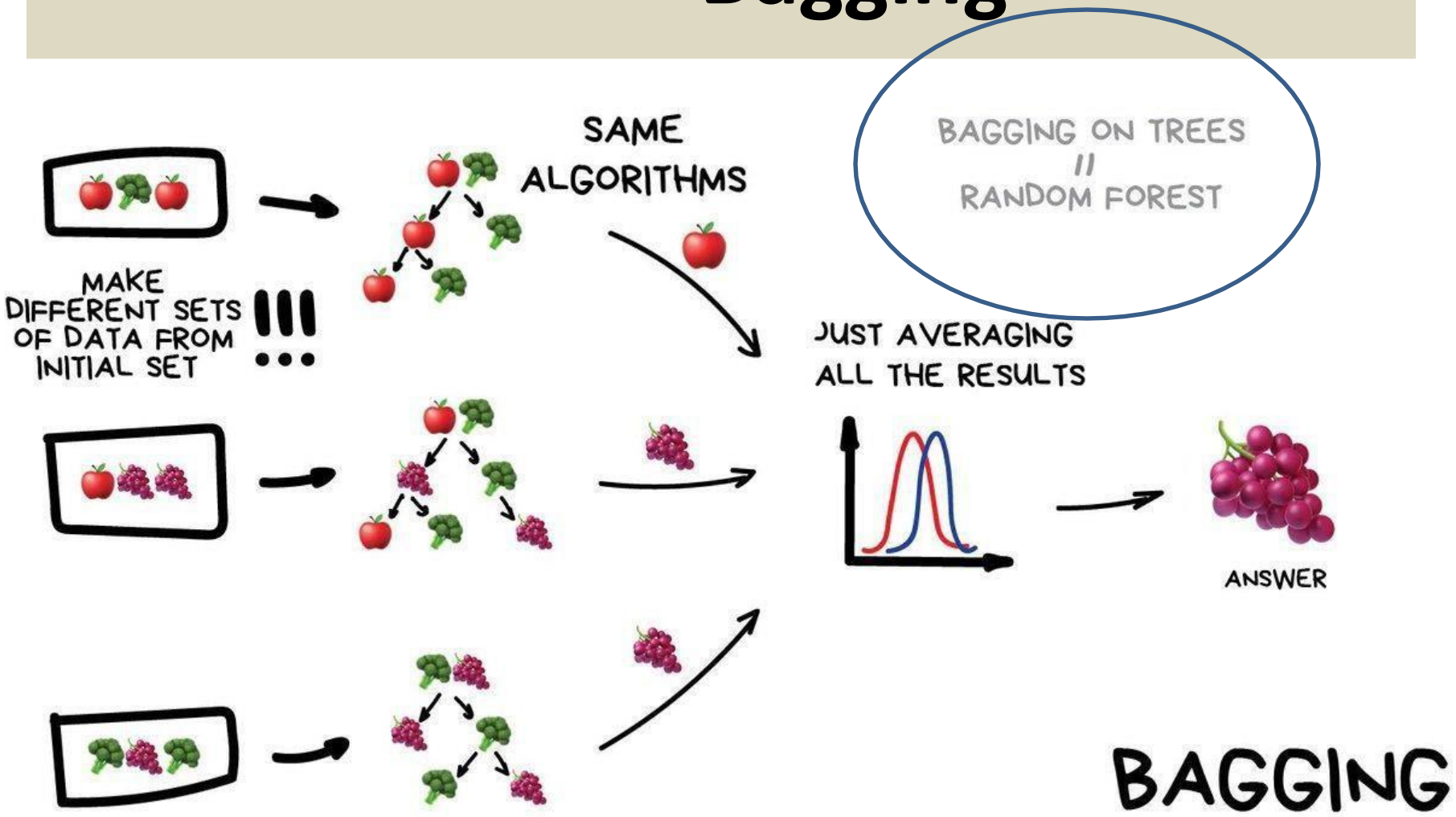
Birleştirme (ensemble) yöntemleri:



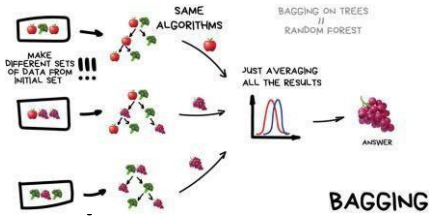
Birleştirme (ensemble) yöntemleri:



Bagging


BAGGING ON TREES
=
RANDOM FOREST
SAME
ALGORITHMS
MAKE
DIFFERENT SETS
OF DATA FROM
INITIAL SET
!!!
JUST AVERAGING
ALL THE RESULTS
ANSWER
BAGGING

TANIMLAR



Rassal Orman(Random Forest)

hangi algoritmayı kullanacağınızı bilmiyorsanız(?)

Rassal orman, hem sınıflamaları hem de regresyon görevlerini yerine getirebilen çok yönlü ve akıllı bir makine öğrenme yöntemi olarak tanımlanabilir.

Bazı veri araştırma adımlarını da içerir:

- boyut küçültme (dimension reduction),
- aykırı değerler(outlier),
- eksik değerler gibi

Bagging den farklı olarak rassal orman **verinin altkümesini** girdi olarak kullanır. (bagging tamamını kullanır)

Bir güçlü model oluşturmak için bir grup zayıf modelin bir araya getiren bir topluluk öğrenme yöntemi olarak bilinir.

Tüm ağaçlar büyüyebildikleri kadar büyür ve budama yapılmaz.

Rassal Orman

Rassal Ormanın Avantajları

- Rassal orman algoritması regresyon ve sınıflandırmada kullanılabilir.
- Yüksek boyutlu büyük miktarda veri kümesini işleyebilir ve aralarındaki önemli değişkenleri çok iyi tanımlayabilir. Bu nedenle önemli bir boyutsal küçültme yöntemi olarak kabul edilir.
- Eksik verileri etkin bir şekilde tahmin edebilir ve çok miktarda veri beslenmiş olsa bile doğruluğu kolayca koruyabilir.
- Veri kümesindeki hataları dengelemek için kullanılabilecek çeşitli yöntemleri vardır.

Yukarıdaki özellikler etiketlenmemiş verilerle de kullanılabilir. Böylece denetimsiz çalışabilir.

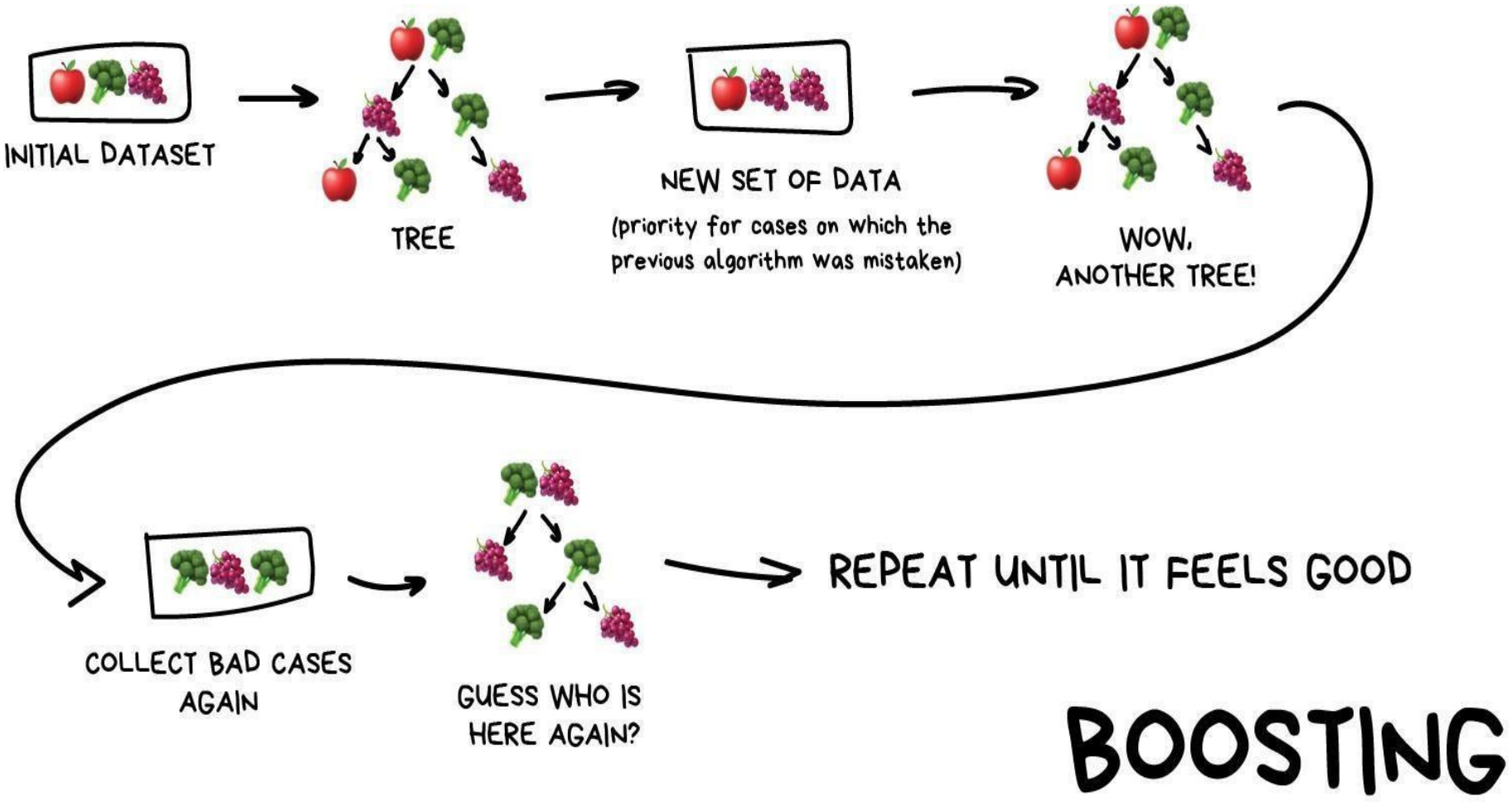
- Giriş verilerini değiştirerek örnekler. Bu işleme bootstrap örneklemesi denir

Rassal Orman

Rassal Ormanın Dezavantajları

- Regresyonda, sınıflandırmada olduğu kadar iyi değildir. Sürekli tahminlemede kesin sonuç elde edemez. Regresyon durumunda sağlanan eğitim verisi aralığının ötesinde tahminlerde bulunamaz.
- Örnek veriler çok gürültülü ise veriler aşırı uyum gösterebilir (ezberler).

Boosting(Artırma)



Boosting(Artırma)

Zayıf modelleri güçlü öğrenici modellere dönüştüren bir algoritma ailesidir.

Ancak, bu ölçütlerin e-postaları ~~spam değil~~ /spam olarak sınıflandırmak için yeterli değildir.

Örnek:

Spam ve normal postaları tanımlamanız istenirse,

- E-postada yalnızca bir tane promosyon resmi varsa,: SPAM
- Yalnızca bağlantılar varsa: SPAM
- E-postanın gövdesi yalnızca "- olarak bir ödül para kazandınız" gibi cümleler içerirse: SPAM
- Resmi bir alan adından alınan e-posta: SPAM değ i l
- Bilinen bir gönderenden alınan e-posta: SPAM değ i l

Boosting(Artırma)

Zayıf kuralı bulmak için önce makine-öğrenme algoritmasını değişik bir dağılımla uygulamak gerekir.

Bu algoritma her uygulandığında yeni bir zayıf tahmin kuralı oluşur. Bu tekrarlayan bir süreçtir ve birçok kez tekrarlanır.

Algoritma bu kuralları tek bir güçlü tahmin kuralında birleştirir.

Arttırma, daha önce gelen yanlış sınıflandırılmış ve yüksek hata örneklerine odaklanır.