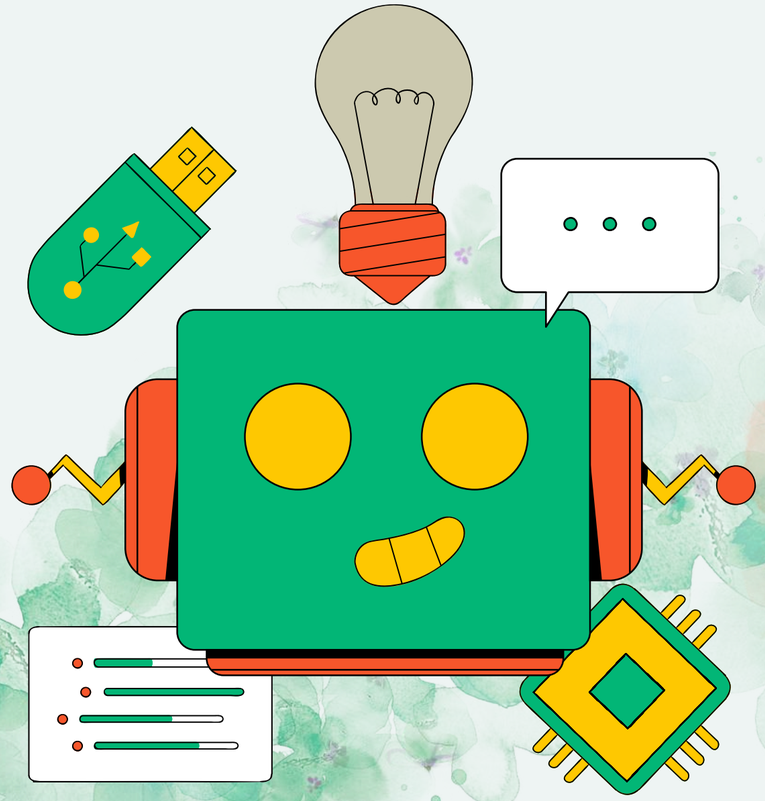
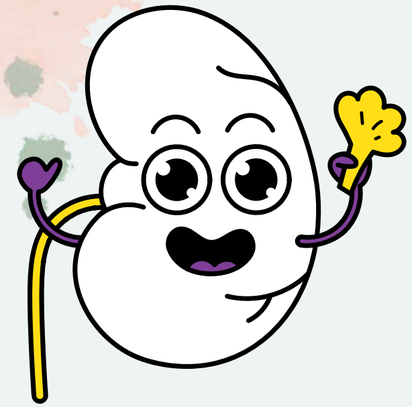
ÇAĞLA ÖKMEN
ÖZLEM AKINCI

MAKİNE ÖĞRENMESİ

CHRONİC KİDNEY DİSEASE

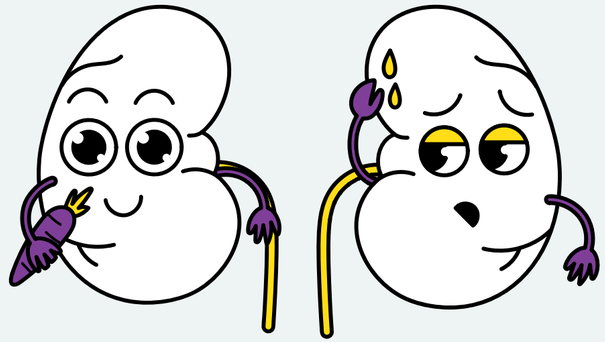




Kaynak:

University of California Irvine Machine Learning Repository

(Kaggle'da Mansoor Iqbal tarafından yayımlanmıştır – Iqbal, 2017)

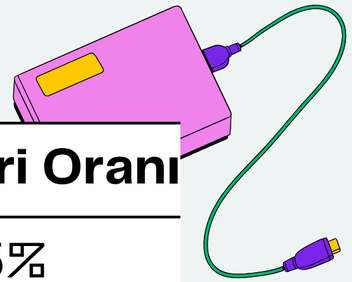


Veri Seti Özellikleri:

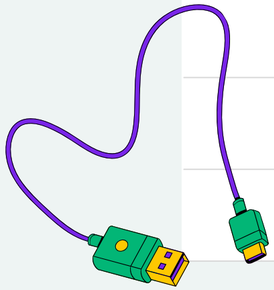
- Toplam 400 örnek
- Her örnekte 24 özellik (değişken)
  - 11 sayısal (numerical)
  - 13 kategorik (nominal)

Sınıflar:

- ckd (Kronik Böbrek Hastalığı olanlar) – 250 örnek
- notckd (Hastalığı olmayanlar) – 150 örnek



Variable	Açıklama	Tür (Class)	Ölçek (Scale)	Eksik Veri Oranı
age	Age	Numerical	age in years	225%
bp	Blood Pressure	Numerical	in mm/Hg	3%
sg	Specific Gravity	Nominal	1.010, 1.015, 1.020	1.175%
al	Albumin	Nominal	(0, 1, 2, 3, 4, 5)	115%



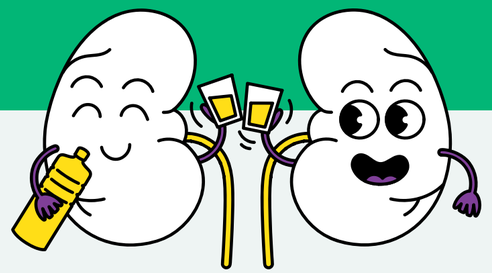
VERİ ÖNİŞLEME

Veri ön işleme, makine öğrenimi modellerinin başarısı için kritik bir adımdır. Gerçek dünya verileri genellikle:

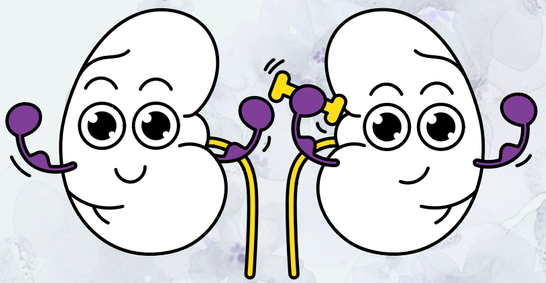
- Gürültü (noise)
- Eksik değerler (Missing Values – MVs)
- Tutarsız ve gereksiz veriler
- Yüksek boyutluluk (örnek ve özellik fazlalığı) gibi sorunlar içerir.

UYGULADIĞIMIZ ADIMLAR

- Kategorik Verilerin Sayısal Formatlara Dönüştürülmesi:
  - Örn: One-Hot Encoding
  - ML algoritmaları için daha uygun hale getirme
- Eksik Değerlerin Giderilmesi:
  - Kategorik değişkenler: En sık görülen değerle (mod) dolduruldu
  - Sayısal değişkenler: Ortalama (mean) değeriyle dolduruldu



KULLANILAN SINIFLANDIRMA YÖNTEMLERİ



KNN

LİNEAR SVM

POLİNOMAL SVM

Random Forest

NATİVE BAYES

KARAR AĞAÇLARI

RBF SVM

LOJISTİK REGRESYON

MLP

KARMAŞIKLIK MATRİSİ

Karmaşıklık Matrisi, sınıflandırıcıların tahmin performansını ölçmek için kullanılır.
Gerçek değerler: Doğru (1) veya Yanlış (0)
Tahmin edilen değerler: Pozitif (1) veya Negatif (0)

Class designation		Actual class	
		True (1)	False (0)
Predicted class	Positive (1)	TP	FP
	Negative (0)	FN	TN

PERFORMANS METRİKLERİ

ACCURACY

$$\frac{TP + TN}{P + N}$$

PRECISION

$$\frac{TP}{TP + FP}$$

RECALL

$$\frac{TP}{P}$$

SPECIFICITY

$$\frac{TN}{N}$$

F-1 SCORE

Kesinlik ve duyarlılığa bağlı olarak hesaplanır.

MCC

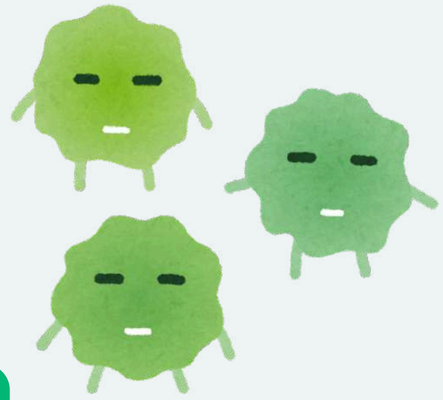
Tahmin edilen sınıflar ile temel gerçek arasındaki korelasyondur

10 KATLI ÇAPRAZ DOĞRULAMA İLE GARFİK YORUMLAMA

	1. Parça	2. Parça	3. Parça	4. Parça	5. Parça	6. Parça	7. Parça	8. Parça	9. Parça	10. Parça
1. Adım	Test	Eğitim	Eğitim	Eğitim	Eğitim	Eğitim	Eğitim	Eğitim	Eğitim	Eğitim
2. Adım	Eğitim	Test	Eğitim	Eğitim	Eğitim	Eğitim	Eğitim	Eğitim	Eğitim	Eğitim
3. Adım	Eğitim	Eğitim	Test	Eğitim	Eğitim	Eğitim	Eğitim	Eğitim	Eğitim	Eğitim
4. Adım	Eğitim	Eğitim	Eğitim	Test	Eğitim	Eğitim	Eğitim	Eğitim	Eğitim	Eğitim
5. Adım	Eğitim	Eğitim	Eğitim	Eğitim	Test	Eğitim	Eğitim	Eğitim	Eğitim	Eğitim
6. Adım	Eğitim	Eğitim	Eğitim	Eğitim	Eğitim	Test	Eğitim	Eğitim	Eğitim	Eğitim
7. Adım	Eğitim	Eğitim	Eğitim	Eğitim	Eğitim	Eğitim	Test	Eğitim	Eğitim	Eğitim
8. Adım	Eğitim	Eğitim	Eğitim	Eğitim	Eğitim	Eğitim	Eğitim	Test	Eğitim	Eğitim
9. Adım	Eğitim	Eğitim	Eğitim	Eğitim	Eğitim	Eğitim	Eğitim	Eğitim	Test	Eğitim
10. Adım	Eğitim	Eğitim	Eğitim	Eğitim	Eğitim	Eğitim	Eğitim	Eğitim	Eğitim	Test

Sınıflandırıcının değişkenliğini azaltmak için, eğitim ve test alt örneklerinin farklı bölümlerini kullanarak birçok döngüde gerçekleştirilir. En son döngülerin sonuçlarının ortalaması alınır

10 KATLI ÇAPRAZLAMA SONUÇLARI



En yüksek performans:

Random Forest ve

MLP'dir.

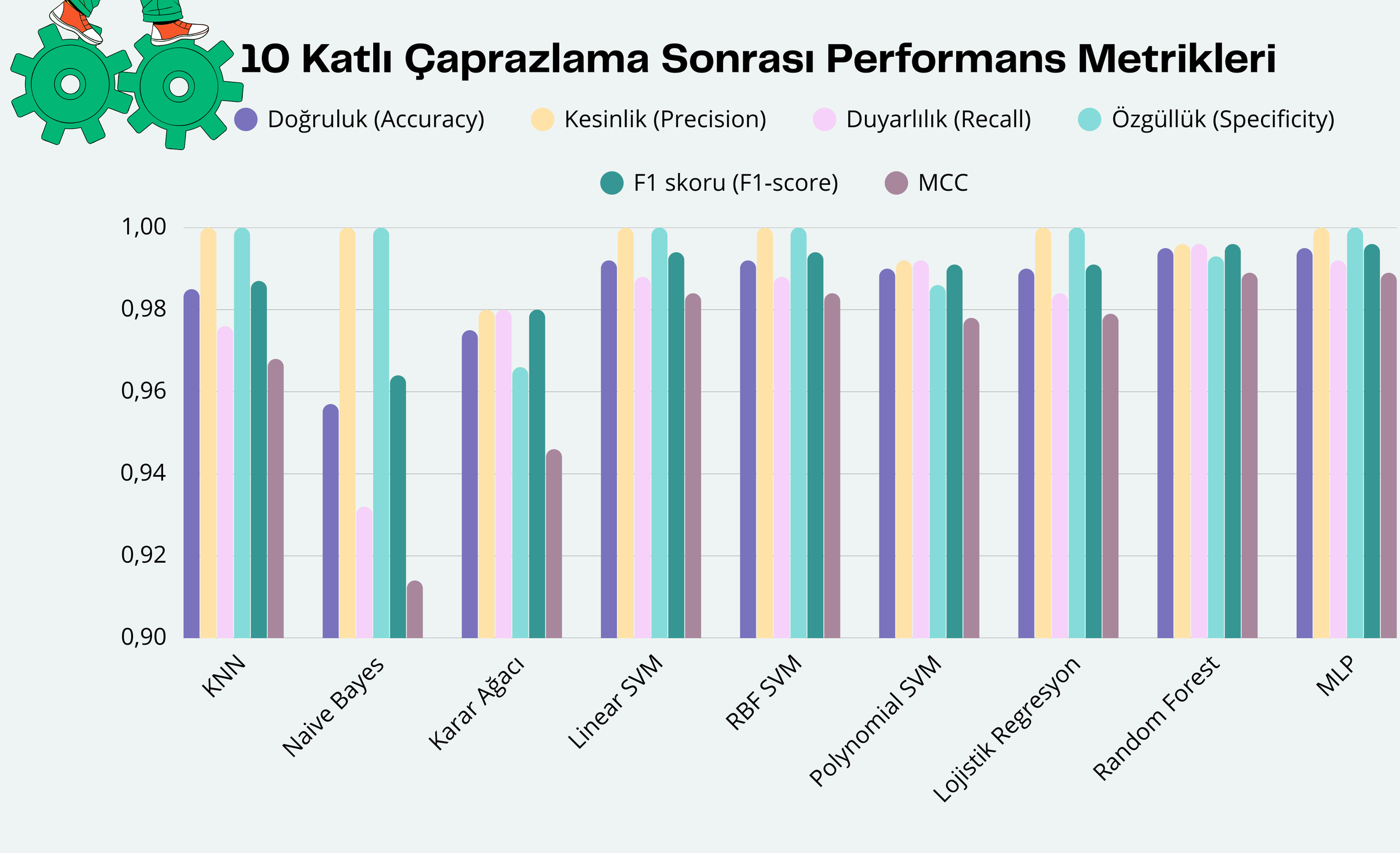
Bu algoritmalar, Doğruluk, F1 Skoru ve MCC gibi metriklerde 0,99'un üzerinde değerler elde etmiştir.

En düşük performans:

Naive Bayes

Bu algoritma, diğerlerine kıyasla daha düşük Doğruluk (0,957) ve F1 Skoru (0,964) elde etmiştir.

10 Katlı Çaprazlama Sonrası Performans Metrikleri



Sütun1	Doğruluk (Accu	Kesinlik (Precis	Duyarlılık (Reca	Özgüllük (Spec	F1 skoru (F1-sc	MCC
KNN	0,985	1	0,976	1	0,987	0,968
Naive Bayes	0,957	1	0,932	1	0,964	0,914
Karar Ağacı	0,975	0,98	0,98	0,966	0,98	0,946
Linear SVM	0,992	1	0,988	1	0,994	0,984
RBF SVM	0,992	1	0,988	1	0,994	0,984
Polynomial SVI	0,99	0,992	0,992	0,986	0,991	0,978
Lojistik Regres	0,99	1	0,984	1	0,991	0,979
Random Forest	0,995	0,996	0,996	0,993	0,996	0,989
MLP	0,995	1	0,992	1	0,996	0,989

GENETİK ALGORİTMALAR İLE ÖZNİTELİK SEÇİMİ

- Yüksek boyutlu veri kümelerinde öznitelik seçimi için etkili bir sezgisel yöntemdir.
- Öznitelik seçimi ve sınıflayıcı eğitimi eş zamanlı gerçekleştirilir.
- Evrimsel süreçle çalışır:
  - Populasyon oluşturulur.
  - Uygunluk değeri yüksek bireyler seçilir.
  - Çaprazlama & mutasyon işlemleriyle yeni nesiller oluşturulur.
  - Süreç en iyi birey bulunana kadar devam eder.

PROJEMİZDE GA SÜRECİ

Başlangıç Popülasyonu: 50 birey

Kromozomlar: 0–1 ikili dizi (öznitelik seçimi)

Çaprazlama oranı: 0.7

Mutasyon oranı: 0.2

Toplam nesil sayısı: 40

GENETİK ALGORİTMA SONUÇLARI

Bazı algoritmaların performansını artırmıştır, bazılarını ise önemli ölçüde etkilememiştir.
Random Forest ve MLP, GA öncesinde de yüksek performansa sahipti.
GA sonrası bu iki modelde küçük ama olumlu iyileşmeler gözlenmiştir.

- **KNN**
- **RBF SVM**
- **Polynomial SVM**

- **Random Forest**
- **MLP**

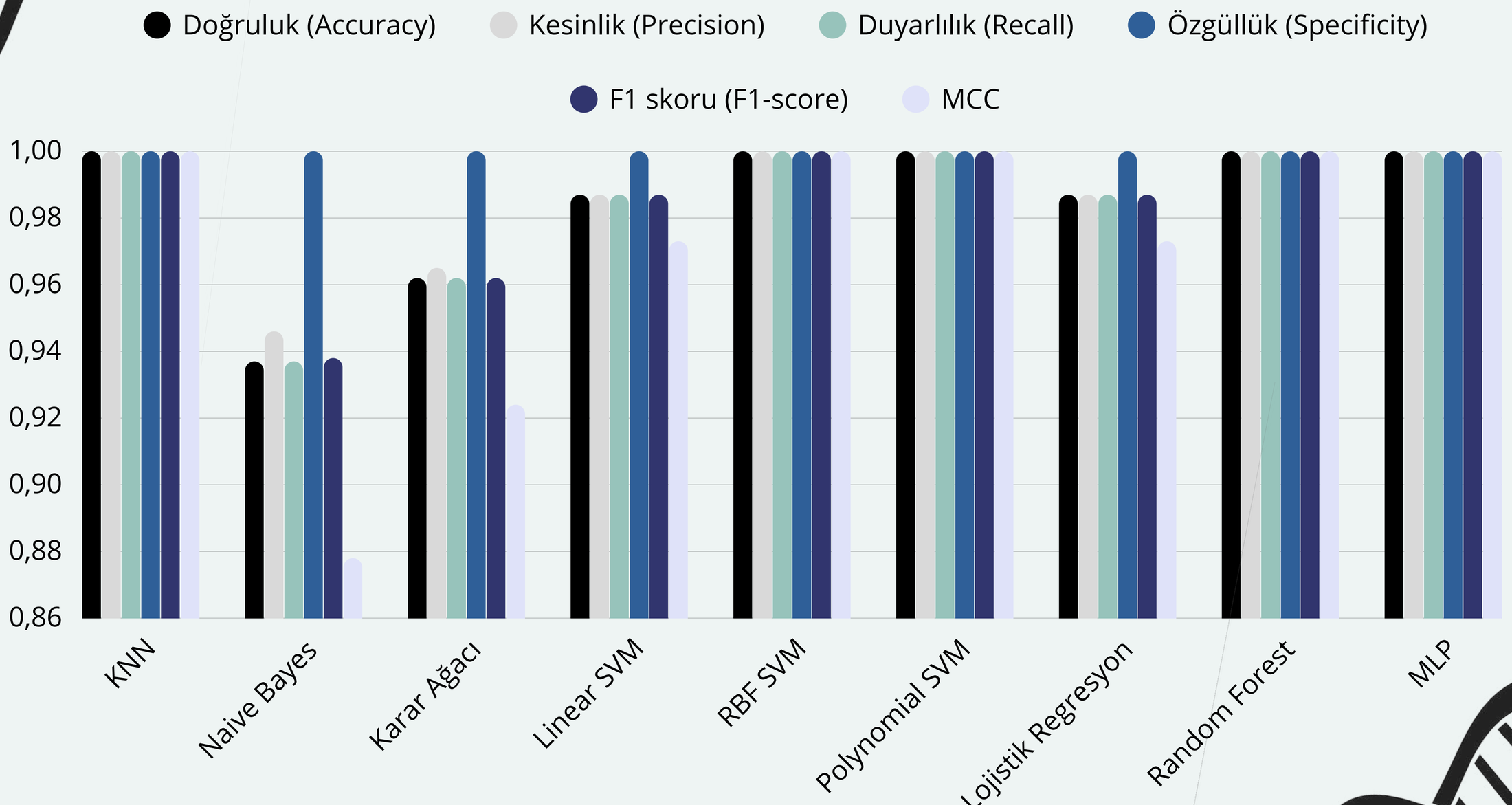
Bu algoritmalar, GA sonrası en yüksek başarıyı gösteren modeller olmuştur.

Naive Bayes Değerlendirmesi

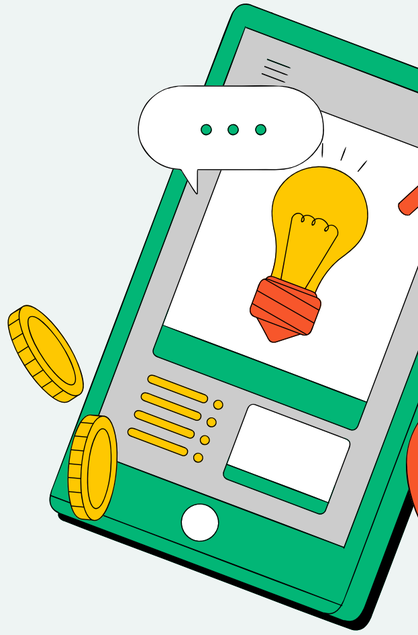
- GA ile performansı artmıştır.
- Ancak diğer modellere kıyasla hala daha düşük başarı göstermektedir.

	Doğruluk (Acc	Kesinlik (Preci	Duyarlılık (Re	Özgüllük (Spe	F1 skoru (F1-sc	MCC
KNN	1	1	1	1	1	1
Naive Bayes	0,937	0,946	0,937	1	0,938	0,878
Karar Ağacı	0,962	0,965	0,962	1	0,962	0,924
Linear SVM	0,987	0,987	0,987	1	0,987	0,973
RBF SVM	1	1	1	1	1	1
Polynomial SV	1	1	1	1	1	1
Lojistik Regres	0,987	0,987	0,987	1	0,987	0,973
Random Fores	1	1	1	1	1	1
MLP	1	1	1	1	1	1

Genetik Algoritma ile Metriklerin Karşılaştırılması



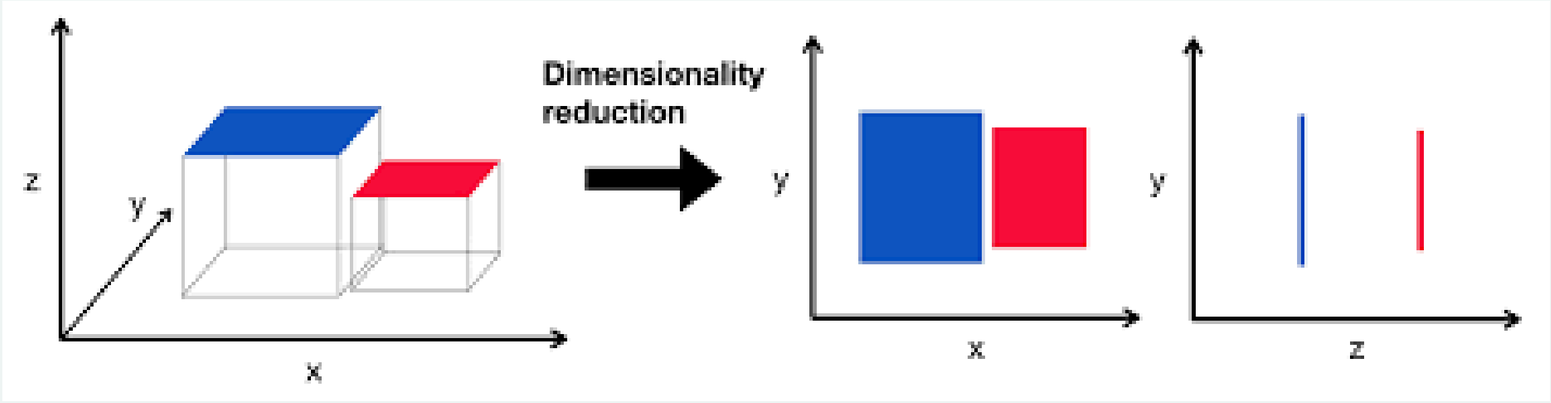
PCA İLE FEATURE REDUCTION



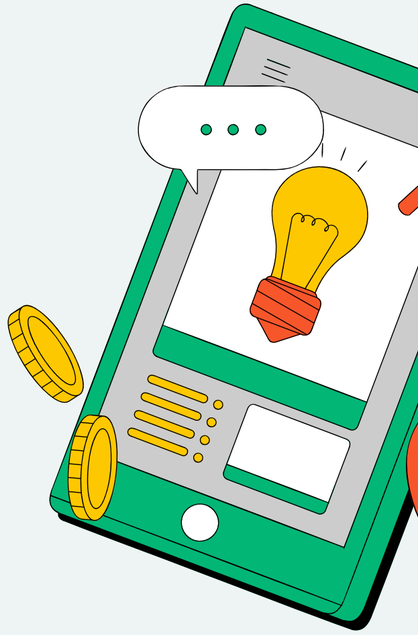
PCA’nın asıl amacı özniteliklerin birbirlerine olan bağımlılığını ortadan kaldırarak, öznitelikleri bağımsız hale getirmektir.

Yüksek boyutlu verileri daha düşük boyutlu bir alt uzaya projekte ederek en yüksek varyansa sahip yönleri koruyan doğrusal bir dönüşümdür.

Verinin temel yapısı korunur



PROJEMİZDE PCA İLE FEATURE REDUCTION



Uygulama Adımları:

- PCA tüm bileşenler üzerinden analiz edildi.
- Her bileşenin açıkladığı varyans oranı hesaplandı.
- Kümülatif varyans incelendi.
- %95'ten fazla varyans sağlayan minimum bileşen sayısı belirlendi.

Sonuç:

- 24 öznitelik → 20 bileşene indirildi.
- Toplam varyansın %95.90'ı korunmuş oldu.

PCA İŞLEMLERİ SONRASI SONUÇLAR

- **KNN**
- **Karar Ağacı**
- **Polynomial SVM**
- **Lojistik Regresyon**
- **Random Forest**

PCA, algoritmaların performansını genel olarak yüksek tutmuştur.

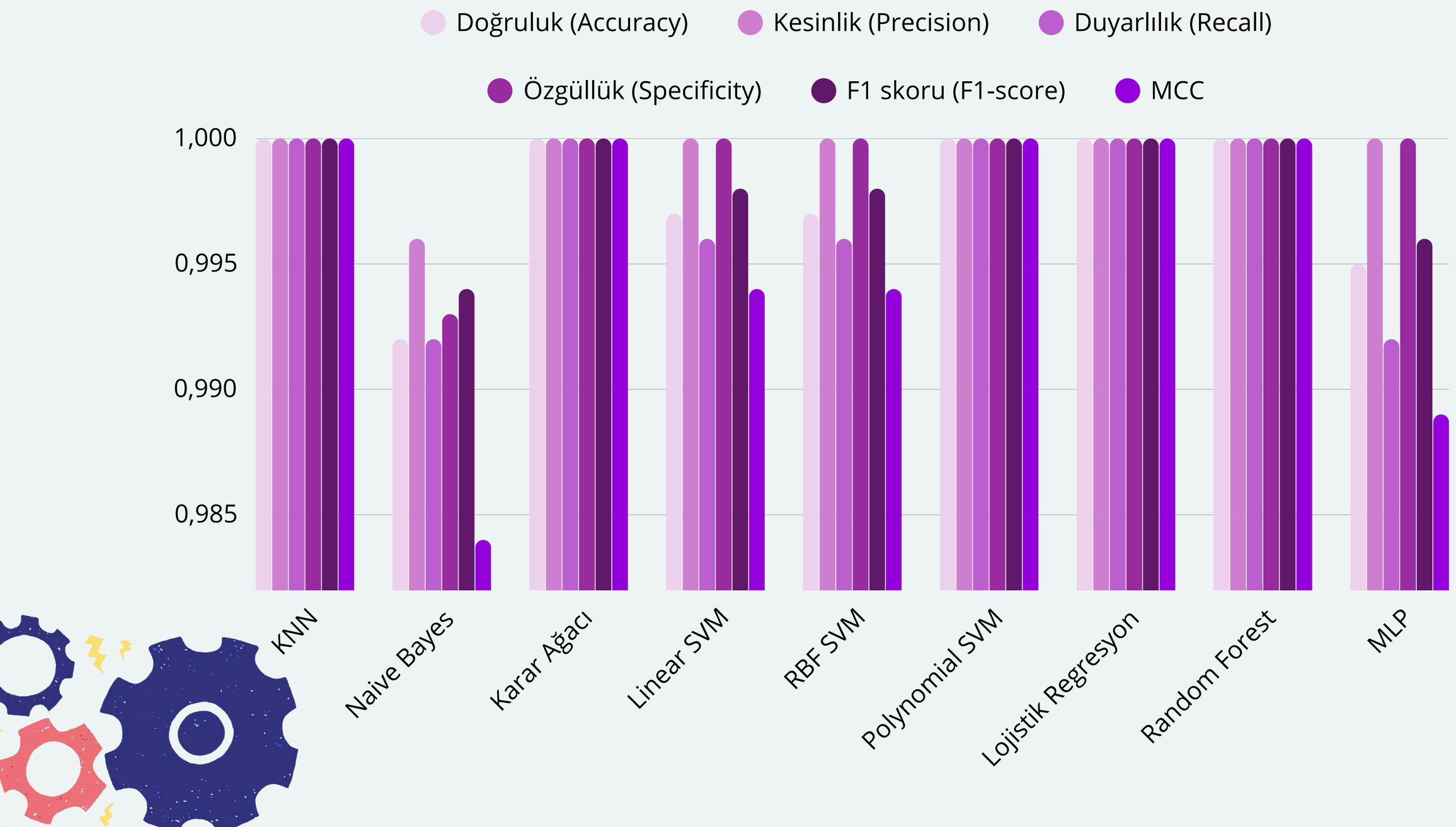
En düşük performans:

Naive Bayes

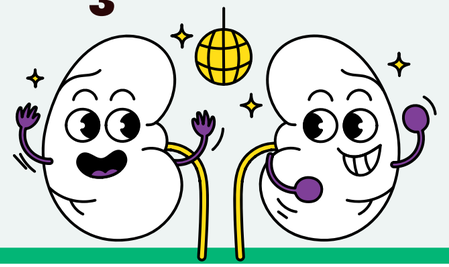
Naive Bayes diğer algoritmalar kadar iyi performans göstermese de PCA ile en iyi performansına sahiptir.

Sütun1	Doğruluk (Accu	Kesinlik (Precis	Duyarlılık (Reca	Özgüllük (Speci	F1 skoru (F1-sc	MCC
KNN	1	1	1	1	1	1
Naive Bayes	0,992	0,996	0,992	0,993	0,994	0,984
Karar Ağacı	1	1	1	1	1	1
Linear SVM	0,997	1	0,996	1	0,998	0,994
RBF SVM	0,997	1	0,996	1	0,998	0,994
Polynomial SVM	1	1	1	1	1	1
Lojistik Regresy	1	1	1	1	1	1
Random Forest	1	1	1	1	1	1
MLP	0,995	1	0,992	1	0,996	0,989

PCA ile Performans Metriklerinin Karşılaştırılması

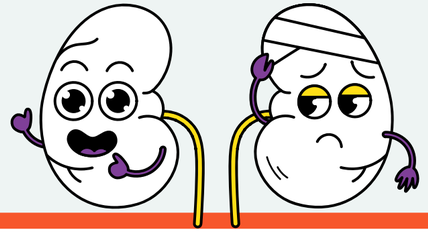


SONUÇLARIMIZ



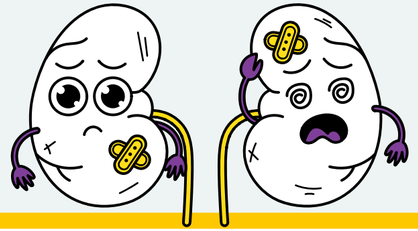
EN İYİ PERFORMANSLAR

Random Forest ve MLP, her üç senaryoda da (10 katlı çapraz doğrulama, GA ile öznitelik seçimi, PCA ile boyut indirgeme) en iyi performansı göstermiştir.



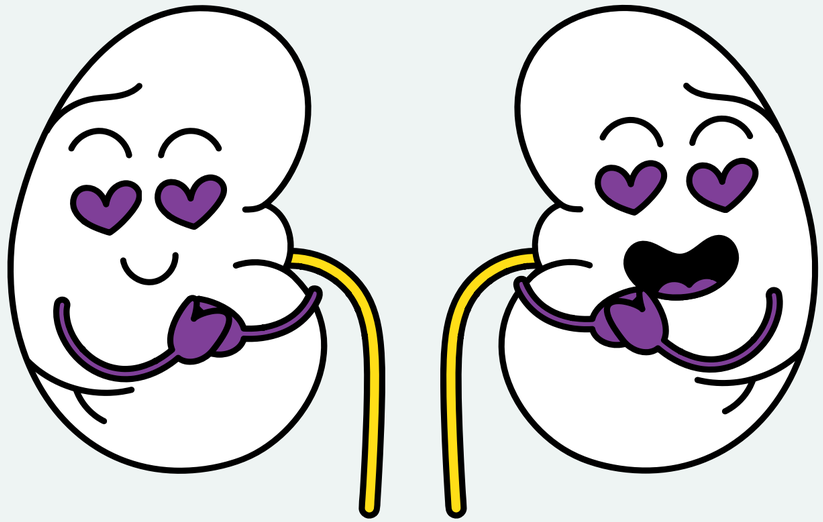
GA VE PCA'NIN ETKİSİ

GA ve PCA, bazı algoritmaların performansını artırsa da Random Forest ve MLP gibi zaten yüksek performans gösteren algoritmalar üzerinde önemli bir etkisi olmamıştır



EN DÜŞÜK PERFORMNAS

Naive Bayes en düşük performansı göstermiştir



DİNLEDİĞİNİZ İÇİN TEŞEKKÜRLER