

**Uyku Bozukluğu Tahmini ve Sınıflandırma**

Bu veri seti, uyku bozukluklarını tespit etmeye yönelik birden fazla modelin eğitimini ve performansını değerlendirmek amacıyla kullanılmıştır. Modelin başarısı, uyku bozukluğu olmayanları (None), uykusuzluğu olanları (Uykusuzluk) ve uyku apnesi olanları (Uyku Apnesi) doğru bir şekilde sınıflandırma yeteneği üzerine odaklanmaktadır.

Elde edilen sonuçlar, modelin her bir uyku bozukluğu türünü değerlendirmede ne kadar etkili olduğunu göstermektedir. Başarı oranı %92 olan bu model, farklı metrikler kullanılarak detaylı bir şekilde değerlendirilmiştir. Precision, recall ve f1-score gibi metrikler, modelin her bir sınıf için ne kadar doğru ve eksiksiz tahminler yaptığını ortaya koymaktadır. Bu değerler, modelin performansını anlamak ve geliştirmek için önemli ipuçları sağlamaktadır.

Bu dokümantasyon, modelin nasıl eğitildiğini, performansının nasıl değerlendirildiğini ve elde edilen sonuçları açıklamaktadır. Ayrıca, uyku bozukluklarıyla ilişkili faktörlerin anlaşılmasına yönelik faydalı bilgiler sunmaktadır.

**Projenin adımları**

- Kütüphanelerin eklenmesi

-Benzersiz değerlerin bulunması

-Veri ön temizleme

-Aykırı değerlerin kontrolü ve düzeltilmesi

-Kategorik değişkenleri sayısal değişkenlere dönüştürme

-Model oluşturma ve modeli eğitme

-Algoritmaların Kullanılması

-Verilerin görselleştirilmesi

-Model peroformansını değerlendirme

**Hazırlayan: Medeni Aba**

**VERİ SETİ HAKKINDA**

**Veri Seti Genel Bakışı**

Uyku Sağlığı ve Yaşam Tarzı Veri Seti, uyku ve günlük alışkanlıklarla ilgili çok çeşitli değişkenleri kapsayan 374 satır ve 13 sütundan oluşur. Cinsiyet, yaş, meslek, uyku süresi, uyku kalitesi, fiziksel aktivite seviyesi, stres seviyeleri, (Vücut Kitle İndeksi ) BMI kategorisi, kan basıncı, kalp hızı, günlük adımlar ve uyku bozukluklarının varlığı veya yokluğu gibi ayrıntıları içerir.

**Veri Seti Sütunları**

**Kişi Kimliği:** Her bir birey için bir tanımlayıcı.

**Cinsiyet:** Kişinin cinsiyeti (Erkek/Kadın).

**Yaş:** Kişinin yıl cinsinden yaşı.

**Meslek:** Kişinin mesleği veya mesleği.

**Uyku Süresi (saat):** Kişinin günde uyuduğu saat sayısı.

**Uyku Kalitesi (ölçek: 1-10):** Uyku kalitesinin 1 ile 10 arasında değişen öznel bir değerlendirmesi.

**Fiziksel Aktivite Seviyesi (dakika/gün):** Kişinin günlük olarak fiziksel aktivitede bulunduğu dakika sayısı.

**Stres Seviyesi (ölçek: 1-10):** Kişinin deneyimlediği stres seviyesinin 1 ile 10 arasında değişen öznel bir değerlendirmesi.

**Vücut Kitle İndeksi (BKİ) Kategorisi**: Kişinin BKİ kategorisi (örn. Düşük Kilolu, Normal, Fazla Kilolu,Obez).

**Kan Basıncı (sistolik/diyastolik):** Kişinin kan basıncı ölçümü, sistolik basıncın diyastolik basınca oranı olarak gösterilir.

**Kalp Hızı (bpm):** Kişinin dinlenme kalp hızı, dakikadaki atış sayısı olarak.

**Günlük Adım Sayısı:** Kişinin günde attığı adım sayısı.

**Uyku Bozukluğu:** Kişide uyku bozukluğunun varlığı veya yokluğu (None, Uykusuzluk, Uyku Apnesi).

Uyku Bozukluğu Sütunu Hakkında Ayrıntılar;

None: Kişi herhangi bir spesifik uyku bozukluğu göstermez.

Uykusuzluk: Kişi uykuya dalmakta veya uykuda kalmakta zorluk çeker, bu da yetersiz veya kalitesiz uykuya yol açar.

Uyku Apnesi: Kişi uyku sırasında nefes almada duraklamalar yaşar, bu da uyku düzeninin bozulmasına ve potansiyel sağlık risklerine neden olur.

**Kullanılan Algoritmalar**

Logistic Regression

Lojistik regresyon, isminde “regresyon” geçmesine rağmen bir sınıflandırma algoritmasıdır. Yani görseldeki hayvanın kedi mi, köpek mi olduğu veya verilmiş olan bilgilerin bir erkeğe mi yoksa bir kadına mı ait olduğunu tahmin etme gibi iki sınıflı sınıflandırma problemlerinde sıkça kullanılır

çizgi, öykü gelişim çizgisi; kumpas; grafiğini çıkarma, diyagram, ekran görüntüsü içeren bir resim

Açıklama otomatik olarak oluşturuldu

KNN-En Yakın Komşu Algoritması

KNN en basit anlamı ile içerisinde tahmin edilecek değerin bağımsız değişkenlerinin oluşturduğu vektörün en yakın komşularının hangi sınıfta yoğun olduğu bilgisi üzerinden sınıfını tahmin etmeye dayanır.KNN (K-Nearest Neighbors) Algoritması iki temel değer üzerinden tahmin yapar;

Distance (Uzaklık): Tahmin edilecek noktanın diğer noktalara uzaklığı hesaplanır. Bunun için Minkowski uzaklık hesaplama fonksiyonu kullanılır.

K (komşuluk sayısı): En yakın kaç komşu üzerinden hesaplama yapılacağını söyleriz. K değeri sonucu direkt etkileyecektir. K 1 olursa overfit etme olasılığı çok yüksek olacaktır. Çok büyük olursa da çok genel sonuçlar verecektir. Bu sebeple optimum K değerini tahmin etmek problemin asıl konusu olarak karşımızda durmaktadır.

metin, ekran görüntüsü, diyagram içeren bir resim

Açıklama otomatik olarak oluşturuldu

XGBoost Algoritması

**XGBoost** (Extreme Gradient Boosting), makine öğrenmesinde sıklıkla kullanılan güçlü bir **gradient boosting** algoritmasıdır. Basitçe söylemek gerekirse, birçok zayıf öğrenciyi (genellikle karar ağaçları) bir araya getirerek çok daha güçlü bir tahmin modeli oluşturmayı hedefler.

**Nasıl Çalışır?**

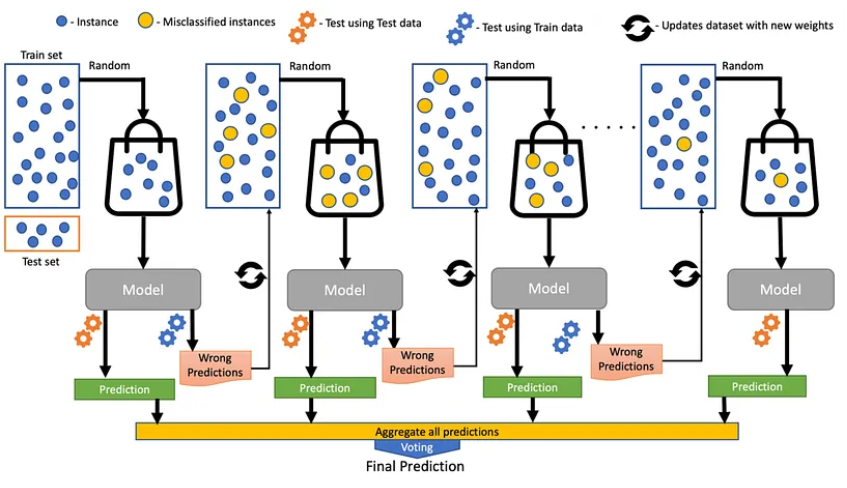
* **İteratif Öğrenme:** XGBoost, her adımda yeni bir model ekleyerek önceki modelin hatalarını düzeltmeye çalışır. Bu süreç, bir merdiven çıkmak gibi düşünebilirsiniz. Her basamak, modeli biraz daha iyi hale getirir.
* **Gradyan İniş:** Yeni bir model eklerken, hangi yönde ilerleyeceğimize karar vermek için gradyan iniş yöntemini kullanır. Bu, en hızlı şekilde en iyi çözüme ulaşmamızı sağlar.
* **Karar Ağaçları:** XGBoost, temel olarak karar ağaçlarını kullanır. Bu ağaçlar, verileri bölerek ve sınıflandırarak kararlar verir.

**Neden XGBoost?**

* **Yüksek Performans:** Büyük ve karmaşık veri setlerinde bile çok iyi sonuçlar verir.
* **Esneklik:** Sınıflandırma, regresyon gibi birçok farklı problem türüne uygulanabilir.
* **Düzenlileştirme:** Modelin aşırı karmaşıklaşmasını önler, böylece yeni verilere daha iyi genelleme yapabilir.

metin, ekran görüntüsü, diyagram, tasarım içeren bir resim

Açıklama otomatik olarak oluşturuldu



|  |  |  |  |
| --- | --- | --- | --- |
| **Özellik** | **KNN (En Yakın Komşular)** | **XGBoost (Extreme Gradient Boosting)** | **Lojistik Regresyon** |
| **Temel Fikir** | Yeni bir veri noktasının sınıfını, en yakın komşularının çoğunluk sınıfına göre belirler. | Birçok zayıf öğrenciyi (genellikle karar ağaçları) birleştirerek güçlü bir model oluşturur. | Bir olayın gerçekleşme olasılığını lojistik bir fonksiyon kullanarak tahmin eder. |
| **Model Eğitimi** | Eğitim verisi üzerinde tüm hesaplamaları yapar. Yeni bir veri geldiğinde, tekrar en yakın komşuları bulur. | İteratif bir süreçle, her adımda yeni bir model ekleyerek modeli geliştirir. | Tüm eğitim verisi kullanılarak tek seferde model eğitimi yapılır. |
| **Model Karmaşıklığı** | K değeri (komşu sayısı) ile kontrol edilir. K küçükse model daha karmaşık, büyükse daha basit olur. | Düzenlileştirme parametreleriyle kontrol edilir. | Modelin karmaşıklığı, kullanılan özellik sayısı ve düzenlileştirme terimleriyle ilişkilidir. |
| **Ölçeklenebilirlik** | Büyük veri setlerinde yavaş olabilir, özellikle yüksek boyutlu verilerde. | Büyük veri setlerinde iyi çalışır. | Genellikle büyük veri setlerinde iyi çalışır. |
| **Yorumlanabilirlik** | Modelin kararlarını anlamak kolaydır (hangi komşular etkiledi). | Modelin kararlarını anlamak daha zordur, çünkü birçok karar ağacının birleşimiyle oluşur. | Modelin kararlarını anlamak daha kolaydır, lojistik regresyon katsayıları özelliklerin önemini gösterir. |
| **Hiperparametreler** | K değeri, uzaklık metriği gibi. | Öğrenme oranı, ağaç derinliği, düzenlileştirme parametreleri gibi birçok hiperparametre. | Düzenlileştirme parametreleri ve özellikleri seçme gibi. |
| **Uygulama Alanları** | Sınıflandırma, regresyon. | Sınıflandırma, regresyon, sıralama gibi birçok alanda kullanılır. | Sınıflandırma, özellikle ikili sınıflandırma problemlerinde kullanılır. |

**Kullanılan Yöntemler**

**Veri Ön İşlemede Kullanılan Yöntemler**

**1-Robust Scaler Nedir?**

Robust Scaler, veri setindeki **aşırı uç değerlerin** etkisini en aza indirmek için kullanılan bir normalizasyon tekniğidir. Bu yöntem, verileri **ortanca** ve **çeyrekler arası aralığa** göre ölçeklendirir. Standart Scaler veya Min-Max Scaler gibi yöntemlerin aksine, aşırı uç değerlerden fazla etkilenmez.

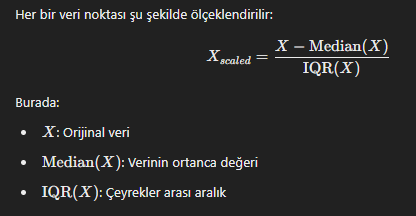
Robust Scaler Nasıl Çalışır?

* Verinin merkezi eğilimi, **ortalama** yerine **ortanca** ile ifade edilir. Ortanca, aşırı uç değerlerden etkilenmez ve veri setinin tam ortasındaki değeri temsil eder.

 Verinin yayılımını ölçmek için **çeyrekler arası aralık (IQR)** kullanılır.

 IQR, verinin 1. çeyrek (Q1, %25) ve 3. çeyrek (Q3, %75) arasındaki farkını ifade eder.

IQR=Q3−Q1



**2-OneHotEncoder**

**Kategorik (niteliksel)** verileri, makine öğrenmesi algoritmalarının işleyebilmesi için **sayısal (niceliksel)** bir forma dönüştüren bir kodlama yöntemidir. Her kategori, **0** ve **1**'den oluşan ayrı bir sütun olarak temsil edilir.

**Her kategori için ayrı bir sütun** oluşturulur ve gözlemde bulunan kategori sütununda 1, diğer sütunlarda 0 yazılır.

**Neden Kullanılır?**

* **Makine öğrenmesi modellerinin** çoğu, kategorik verilerle doğrudan çalışamaz. OneHotEncoder, bu kategorik verileri sayısal bir forma çevirerek modellerin anlamasına olanak tanır.
* Özellikle kategoriler arasında **sıralama ilişkisi olmayan** durumlar için uygundur. (Örneğin: "Renk" kategorisinin kırmızı, mavi ve yeşil arasında sıralama ilişkisi yoktur.)

**Dengesizlik Verilerin İşlenmesi**

**SMOTE (Synthetic Minority Oversampling Technique)**

Makine öğreniminde dengesiz veri setleri ile çalışırken kullanılan bir yeniden örnekleme (resampling) yöntemidir. Özellikle, sınıf dengesizliği olan veri setlerinde azınlık sınıfındaki örneklerin sayısını artırmak için kullanılır.

**Neden SMOTE Kullanılır?**

Sınıf dengesizliği, bir sınıfın diğer sınıflara kıyasla çok daha az sayıda örneğe sahip olduğu durumlarda ortaya çıkar. Bu durum, makine öğrenimi algoritmalarının çoğunluk sınıfa daha fazla odaklanmasına ve azınlık sınıfını ihmal etmesine yol açabilir.

SMOTE, azınlık sınıfındaki örneklerin sayısını artırarak bu sorunu çözmeyi hedefler. Ancak bunu **orijinal örnekleri kopyalamak yerine, yeni örnekler oluşturarak** yapar.

**SMOTE'un Avantajları**

1. **Sınıf Dengesizliğini Giderir:**
   * Azınlık sınıfının öğrenilmesini kolaylaştırır ve modelin performansını artırabilir.
2. **Overfitting'i Azaltır:**
   * Azınlık sınıfındaki örnekleri kopyalamak yerine sentetik örnekler oluşturduğu için aşırı öğrenme (overfitting) riskini azaltır.
3. **Esnek ve Genel Amaçlı:**
   * Birçok makine öğrenimi algoritmasıyla kolayca entegre edilebilir.

**GRAFİKLER**

**-------------------------------------------------------------------------------------------------------------------------**

**metin, ekran görüntüsü, diyagram, çizgi içeren bir resim

Açıklama otomatik olarak oluşturuldu**

**Şekil 1.1** Farklı vücut kitle indeksi (BMI) kategorilerinde uyku bozukluğu türlerinin dağılımını gösteren bir yığılmış sütun grafiğidir.

**metin, ekran görüntüsü, diyagram, renklilik içeren bir resim

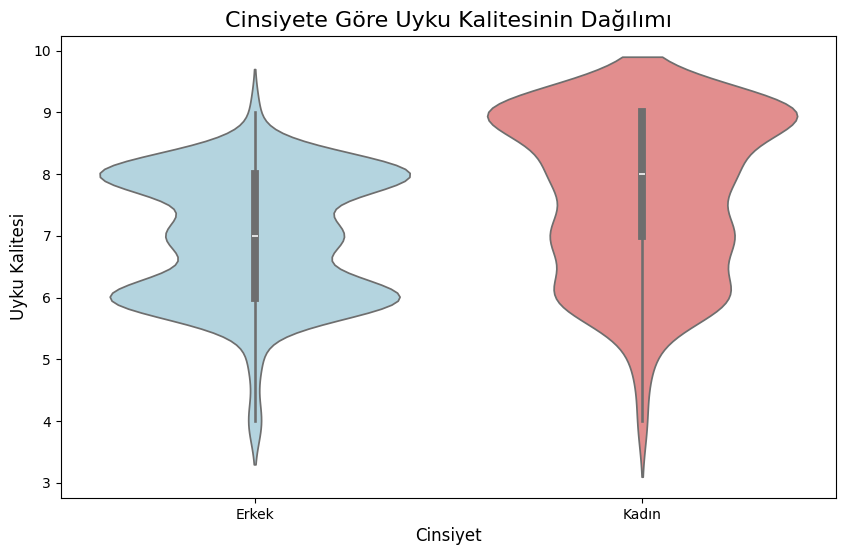
Açıklama otomatik olarak oluşturuldu**

**Şekil 1.2** Uyku kalitesiyle uyku bozukluklarının ilişkilendirilmesini ve kategorilere göre sayısal dağılım grafiği

**metin, ekran görüntüsü, diyagram, öykü gelişim çizgisi; kumpas; grafiğini çıkarma içeren bir resim

Açıklama otomatik olarak oluşturuldu**

**Şekil 1.3** Farklı meslek gruplarındaki bireylerin vücut kitle indeksi kategorilerine göre uyku bozukluğu durumlarını karşılaştıran bir sütun grafiğidir.

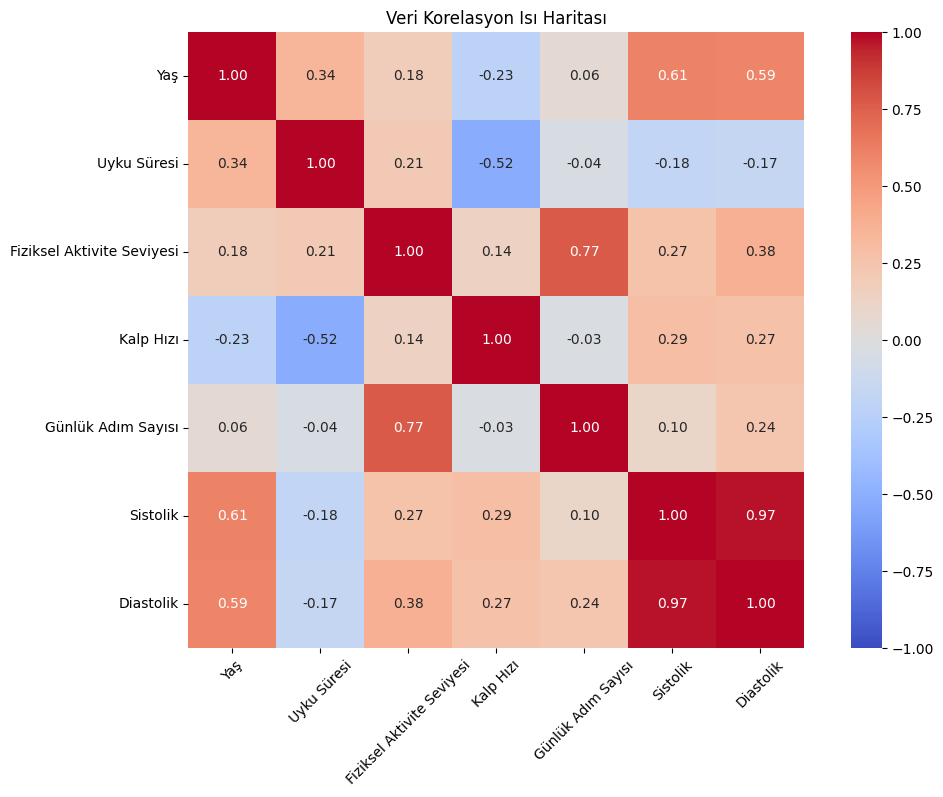


**Şekil 1.4-** Cinsiyete göre uyku kalitesinin dağılımını gösteren bir viyol grafiği

ekran görüntüsü, metin, renklilik, dikdörtgen içeren bir resim

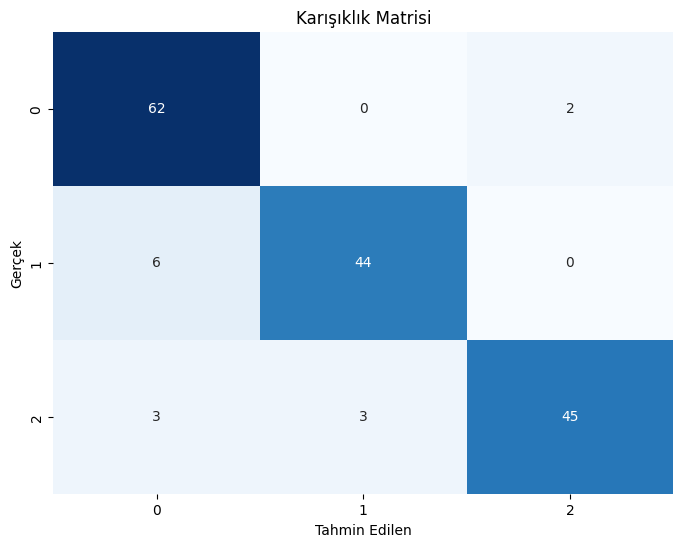
Açıklama otomatik olarak oluşturuldu

**Şekil 2.1 -** Stres seviyesi ile uyku kalitesi arasındaki ilişkiyi gösteren bir çubuk grafiği

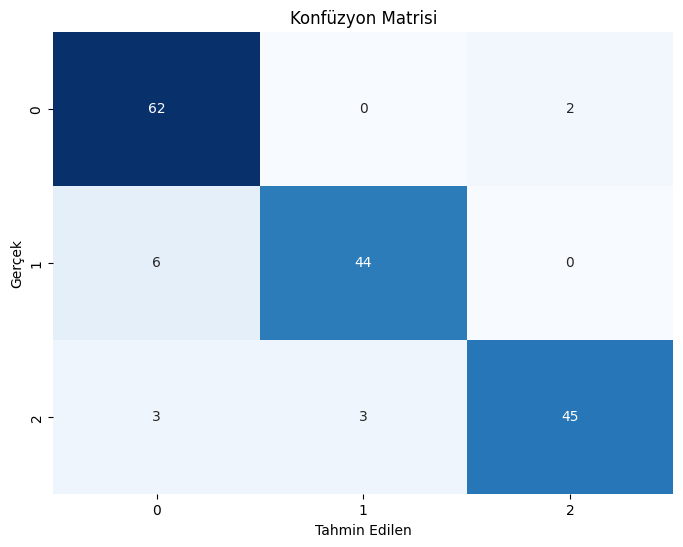
****

**Şekil 2.2-** Farklı özellikler arasındaki korelasyonları gösteren bir ısı haritası grafiği

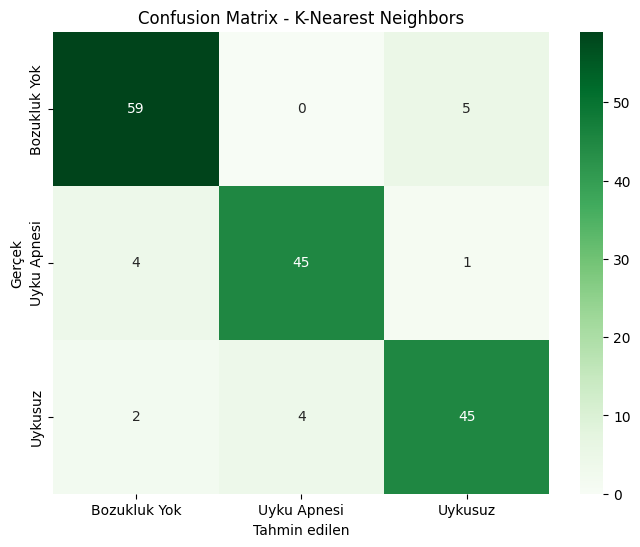
**XGB Classifier Konfüzyon Matrisi**

**Şekil 3.1 -** Makine öğrenmesi modelinin performansını gösteren bir karışıklık matrisi; modelin doğru ve yanlış tahminlerini, her sınıf için gerçekteki değerlerle karşılaştırarak özetler.

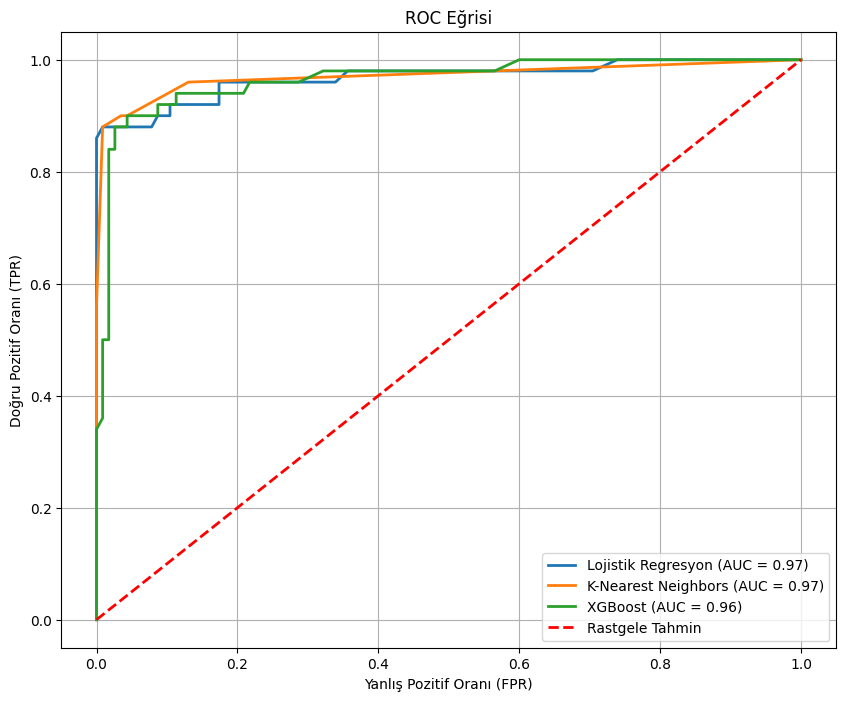
**Lojistik Regresyon Confüzyon Matrisi**

**Şekil 3.2 -** Tahmin başarı oranını gösteren bir konfüzyon matrisi, gerçek sınıflar ile modelin tahmin ettiği sınıflar arasındaki birebir karşılaştırmayı sunar.

**KNN Konfüzyon Matrisi**

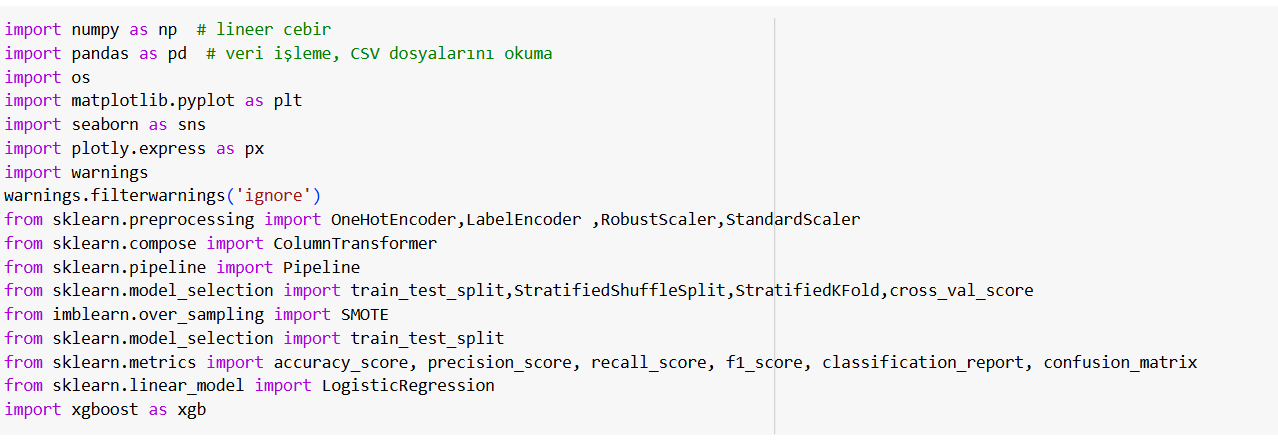
****

**Şekil 3.3** K-En Yakın Komşu (KNN) algoritması kullanılarak yapılan bir sınıflandırma modelinin performansını gösteren bir karışıklık matrisi, modelin gerçek ve tahmin edilen sınıflar arasındaki doğruluk oranlarını özetler.

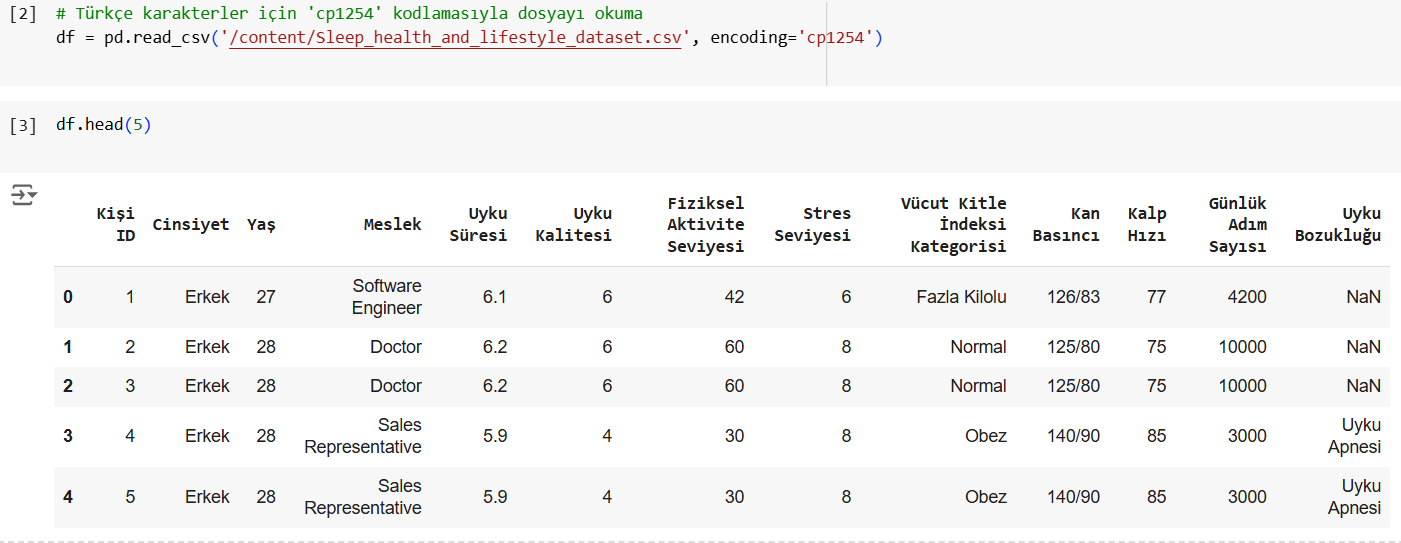
**Şekil 4** Doğru pozitif oran (TPR) ile yanlış pozitif oran (FPR) arasındaki ilişkiyi gösterir ve farklı algoritmaların başarılarını karşılaştırmak için AUC (Eğri Altındaki Alan) değerlerini içerir.

**Lojistik Regresyon** ve **KNN** modellerinin AUC değerleri (0.97 , 0.97) oldukça iyi. Bu modeller, sınıfları çok iyi ayırt edebiliyor. **XGBoost** AUC değeri (0.96) biraz daha düşük, ancak yine de çok iyi bir performansa sahip.

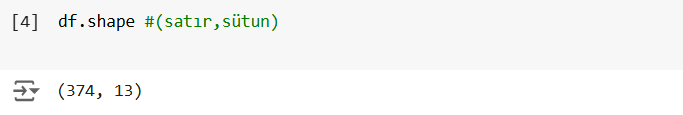
**Kütüphaneleri Ekleme**

****

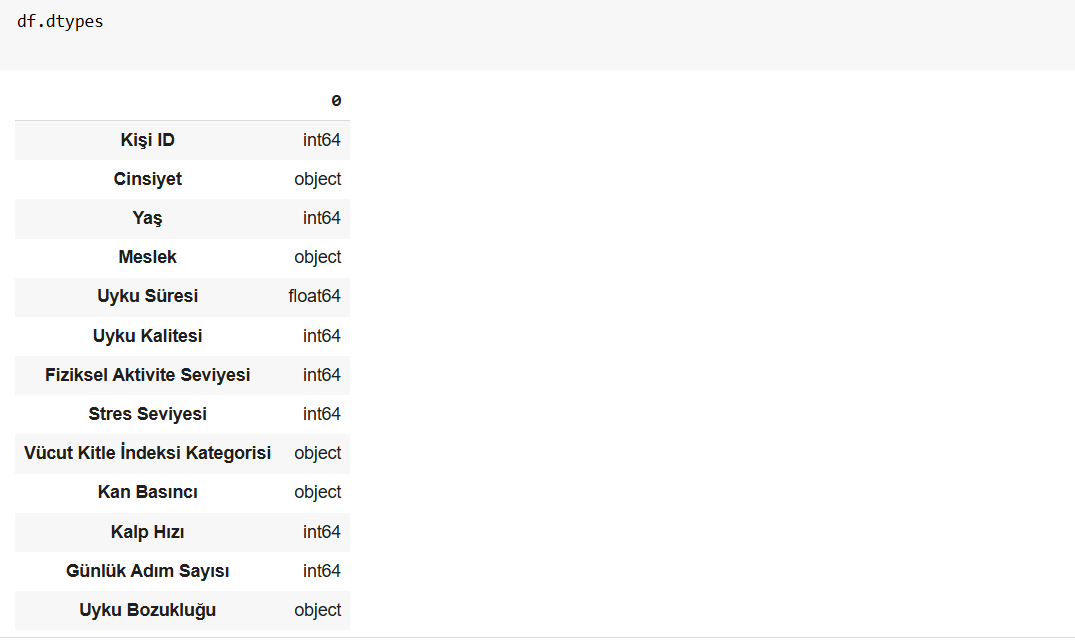
**Veri Setini Okuma İlk 5 Satını Görüntüleme**

****

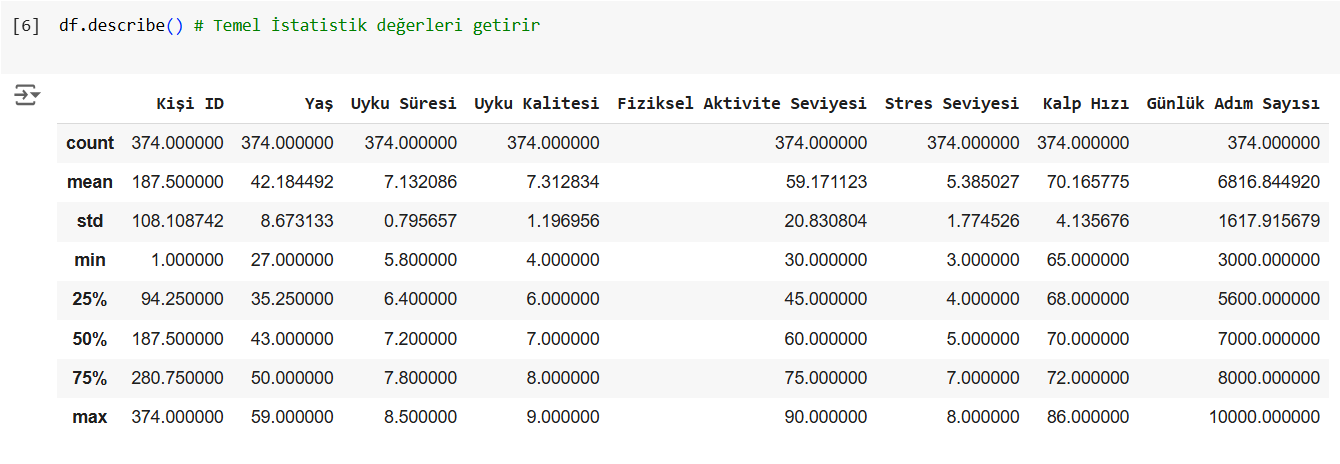
**Veri seti satır ve sütun görüntüleme**

****

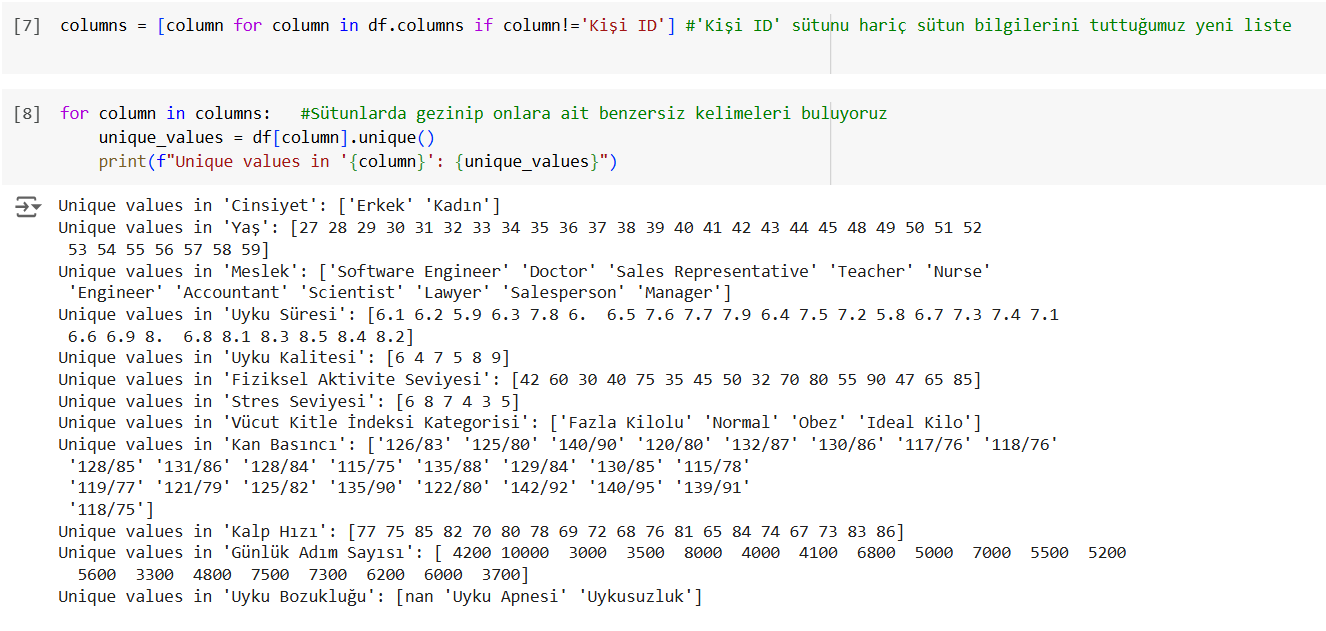
**Veri seti sütun veri tiplerini görüntüleme**

****

**Veri seti temel istatistik değerleri getirme**

****

**Sütun düzenlemesi ve benzersiz kelimelerin bulunması**

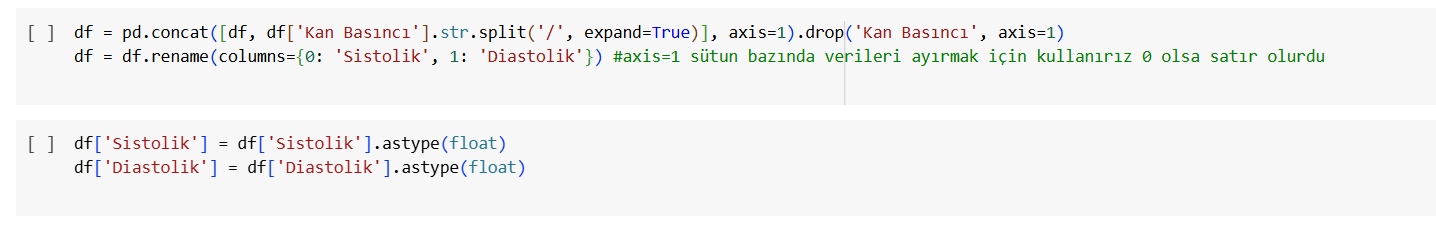
****

****

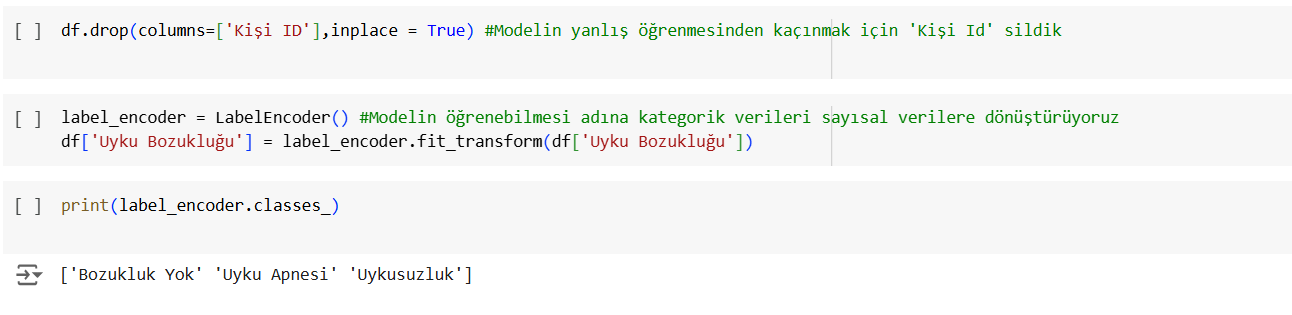
**metin, ekran görüntüsü, yazı tipi, çizgi içeren bir resim

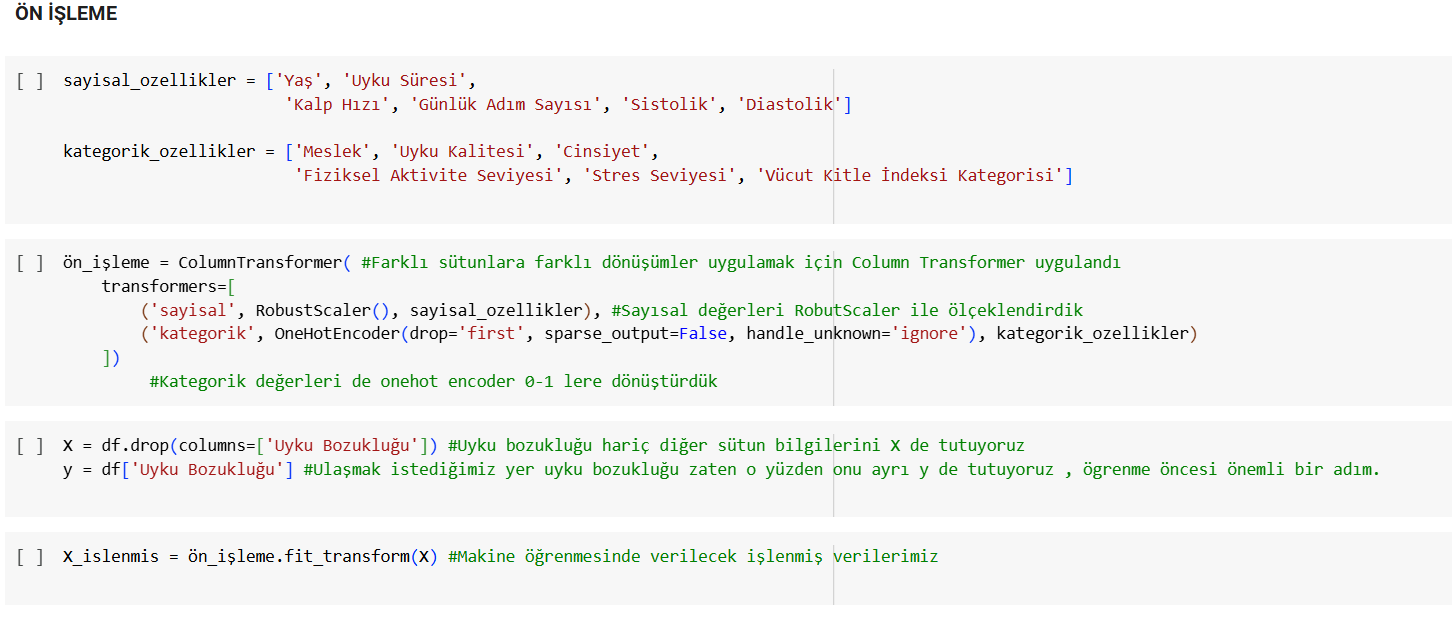
Açıklama otomatik olarak oluşturuldu**

**Kan basıncı değerinin model daha iyi öğrensin diye 2 sütuna ayırılması**

****

**Veri Önişleme**

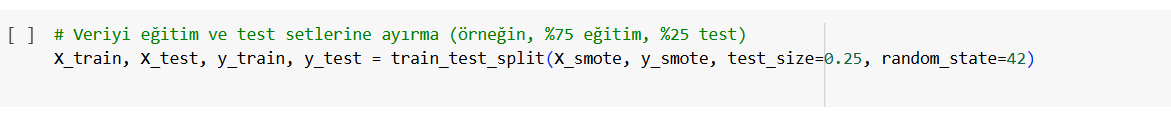
****

****

**Dengesiz Verilerin İşlenmesi**

**metin, yazı tipi, ekran görüntüsü içeren bir resim

Açıklama otomatik olarak oluşturuldu**

****

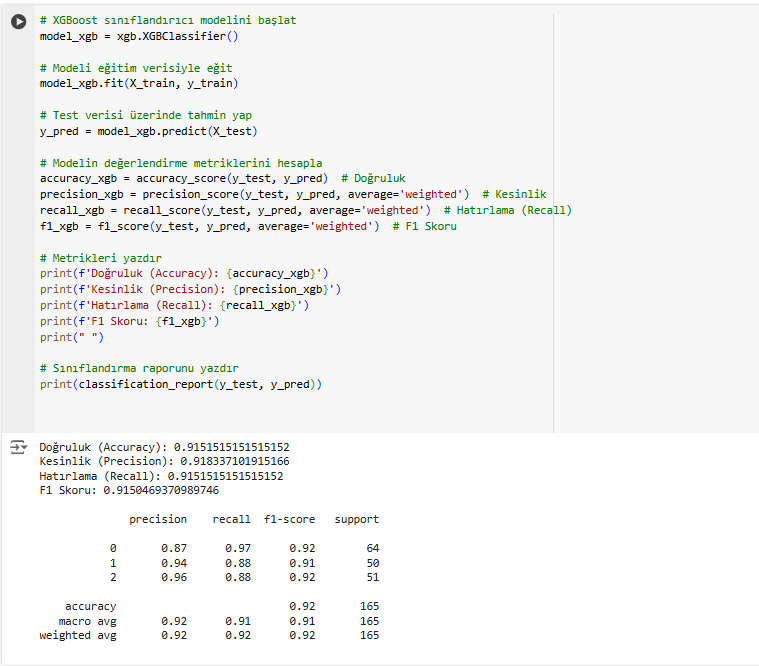
**Model Eğitme**

**Logistic Regresyon**

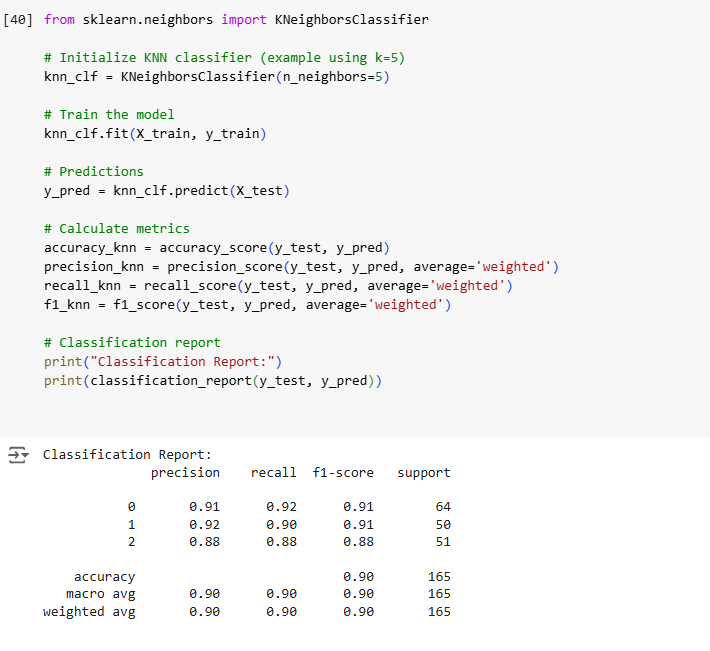
**metin, ekran görüntüsü, yazı tipi, sayı, numara içeren bir resim

Açıklama otomatik olarak oluşturuldu**

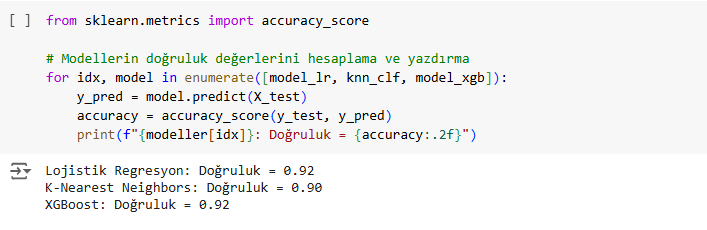
**XGB**

****

**KNN**

****

**Genel Doğruluk**

****