

**Sağlık Sigortası Maliyet Tahmini**

Bu veri seti, bireylerin sağlık sigortası maliyetlerini tahmin etmek için kullanılmış ve iki farklı makine öğrenimi algoritmasıyla değerlendirilmiştir ; Linear Regression (Doğrusal Regresyon) ve Random Forest (Rastgele Orman). Amaç, sigorta maliyetlerini etkileyen faktörlerin (örneğin, yaş, cinsiyet, vücut kitle indeksi, sigara kullanımı, çocuk sayısı ve bölge) sağlık sigortası ücretleri üzerindeki etkisini analiz ederek bireysel tıbbi maliyetleri tahmin etmektir. Bu analiz, sağlık sigortası maliyetlerini etkileyen faktörlerin detaylı bir şekilde anlaşılmasını ve gelecekteki maliyetlerin daha iyi tahmin edilmesini sağlamaktadır.

Elde edilen modeller:

1. Sigorta şirketleri tarafından maliyet analizi ve fiyatlandırma stratejileri oluşturulmasında,
2. Sağlık politikalarının belirlenmesinde,
3. Kişisel maliyet öngörülerinin geliştirilmesinde kullanılabilir.

Bu dokümantasyon, uygulanan modellerin performansını ve sağlık sigortası maliyetleriyle ilişkili faktörlerin analizine yönelik değerli bilgiler sunmaktadır.

**Hazırlayan: Medeni Aba, Ahmet Can İzgi**

**Veri Setine Genel Bakış**

Bu veri seti, ABD'deki bireylerin sağlık sigortası maliyetleriyle ilgili demografik, yaşam tarzı ve sağlık bilgilerini içermektedir. Toplamda 1338 satır ve 7 sütun bulunmaktadır. Veri seti, bireylerin yaşları, cinsiyetleri, sigara kullanımları, çocuk sayıları, yaşadıkları bölge ve sağlık sigortası maliyetleri gibi detayları içermektedir.

**Veri Seti Sütunları**

1. **Yaş** : Bireyin yaşı (Tam sayı).
2. **Cinsiyet** : Sigorta sahibinin cinsiyeti (Kadın/Erkek).
3. **Vücut Kitle İndeksi** : Bireyin vücut kitle indeksi (kg/m²). 18.5 ile 24.9 arası ideal kabul edilir.
4. **Çocuk Sayısı** : Sağlık sigortası kapsamındaki çocuk sayısı (Tam sayı).
5. **Sigara Kullanımı** : Bireyin sigara kullanıp kullanmadığı (Evet/Hayır).
6. **Bölge** : Sigorta sahibinin ABD'deki yaşadığı bölge (Kuzeybatı, Güneydoğu, Güneybatı, Kuzeydoğu).
7. **Sigorta Ücretleri** : Sağlık sigortası tarafından faturalandırılan bireysel tıbbi maliyet (Ondalıklı sayı).

**Sigara Kullanımı Sütunu Hakkında Ayrıntılar:**

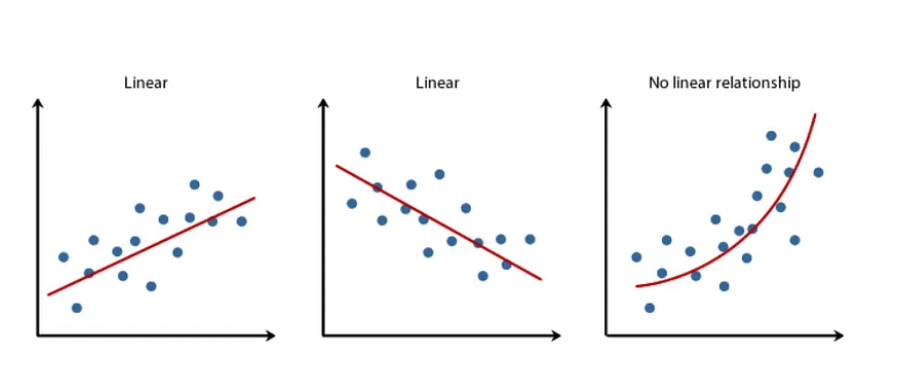
* **Evet** : Bireyin sigara içtiğini gösterir. Genellikle sağlık sigortası maliyetlerinde artışa yol açar.
* **Hayır** : Bireyin sigara içmediğini gösterir.

Bu veri seti, sağlık sigortası maliyetlerini etkileyen çeşitli faktörleri analiz etmek ve bu faktörlerin sağlık giderleri üzerindeki etkilerini incelemek için uygundur.

**Kullanılan Algoritmalar**

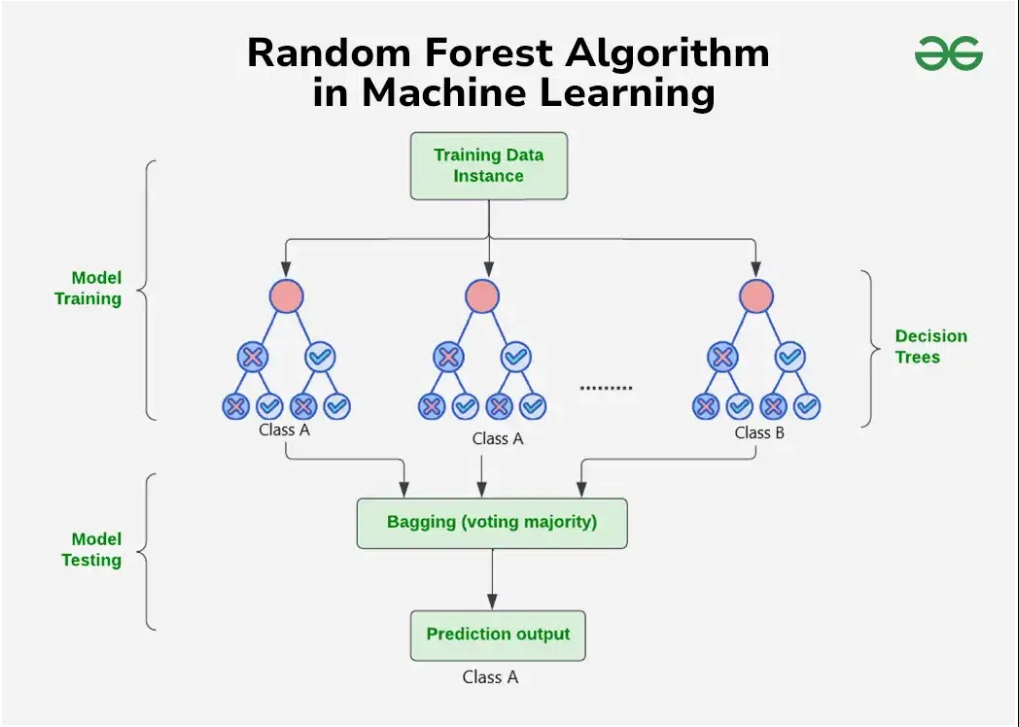
Lineer Regresyon

Lineer Regresyon, bir bağımlı değişken ile bir veya daha fazla bağımsız değişken arasındaki ilişkiyi modellemek için kullanılan basit bir istatistiksel regresyon yöntemidir. Bu yöntem, doğrusal bir ilişkiyi ifade eden bir denklem oluşturarak, bağımsız değişkenlerin değerleri ile bağımlı değişken arasındaki ilişkiyi açıklamaya çalışır.



Random Forest

Random Forest, birçok karar ağacının bir araya gelerek oluşturduğu bir ansambil (ensemble) öğrenme algoritmasıdır. Her bir karar ağacı, rastgele örneklem alınmış veri alt kümesi üzerinde eğitilir ve bu ağaçlar rastgele özelliklerle (değişkenlerle) oluşturulur.



|  |  |  |  |
| --- | --- | --- | --- |
| **Özellik** |  | **Random Forest** | **Lineer Regresyon** |
| Temel Fikir |  | Birçok karar ağacının bir araya gelmesiyle oluşturulan bir ansambl öğrenme yöntemi. Her bir karar ağacı, veri kümesinin farklı bir alt kümesi üzerinde eğitilir ve farklı özellikler seçilir. | Bağımlı değişken ile bir veya daha fazla bağımsız değişken arasındaki doğrusal ilişkiyi modellemek. |
| Model Karmaşıklığı |  | Genellikle daha karmaşık modeller oluşturabilir. Çok sayıda etkileşim ve doğrusal olmayan ilişkileri yakalayabilir. | Daha basit modellerdir. Sadece doğrusal ilişkileri modelleyebilir. |
| Veri Tipi |  | Hem sayısal hem de kategorik verilerle çalışabilir. | Genellikle sayısal verilerle çalışır. Kategorik veriler dönüştürülmelidir. |
| Aşırı Uyum (Overfitting) |  | Daha az eğilimlidir. Birden fazla ağaç kullanarak aşırı uyum riskini azaltır. | Daha fazla eğilimlidir. Özellikle veri kümesi küçükse veya çok fazla özellik varsa. |
| Eğitim Süresi |  | Genellikle daha uzun sürer. Birden fazla ağaç eğitmek gerekir. | Daha hızlıdır. Tek bir doğrusal denklem çözülür. |
| Tahmin Süresi |  | Daha yavaştır. Birden fazla ağaçta tahmin yapmak gerekir. | Daha hızlıdır. Tek bir doğrusal denklem kullanılır. |
| Özellik Önemliliği |  | Her bir özelliğin modelde ne kadar önemli olduğunu gösteren bir metrik sağlar. | Özelliklerin önemini belirlemek için ek analizler yapmak gerekebilir. |
| Aykırı Değerlere Duyarlılık |  | Daha az duyarlıdır. Birçok ağacın ortalaması alındığı için aykırı değerlerin etkisi azalır. | Daha duyarlıdır. Aykırı değerler modeli önemli ölçüde etkileyebilir. |
| İnteraksiyonlar |  | Özellikler arasındaki etkileşimleri yakalayabilir. | Sadece doğrusal etkileşimleri yakalayabilir. |
| Kullanım Alanları |  | Sınıflandırma, regresyon, özellik seçimi gibi birçok alanda kullanılır. | Genellikle regresyon problemlerinde kullanılır. |

**Kullanılan Yöntemler**

**Aykırı Değer Tespitinde Kullanılan Yöntemler**

Aykırı değer tespiti, bir veri setindeki diğer gözlemlerden önemli ölçüde farklı olan değerleri belirleme işlemidir. IQR (Interquartile Range) yöntemi, aykırı değerleri tespit etmek için yaygın olarak kullanılan istatistiksel bir tekniktir. IQR, bir veri setinin orta %50'lik kısmını temsil eder ve çeyrekler arası aralık olarak bilinir.

**IQR Nedir?**

IQR, bir veri setinin dağılımını ölçen bir metriktir ve şu şekilde hesaplanır:

IQR=Q3−Q1

* Q1 (Birinci Çeyrek): Verilerin %25'inin altında olduğu değerdir (alt çeyrek).
* Q3 (Üçüncü Çeyrek): Verilerin %75'inin altında olduğu değerdir (üst çeyrek).
* IQR: Verilerin ortadaki %50'lik kısmının genişliğini temsil eder.

veri setindeki aykırı değerleri tespit etmek ve temizlemek için güvenilir bir yöntemdir. Ancak, aykırı değerlerin her zaman "hatalı" olduğunu varsaymadan, veri analizi bağlamına göre değerlendirilmesi gerekir.

**Z-Skoru**

**Z-skoru (Z-score)**, bir veri noktasının, bir dağılımın ortalamasından kaç standart sapma uzaklıkta olduğunu gösteren bir ölçüdür. Veri analizi ve istatistikte yaygın olarak kullanılan bir yöntemdir ve özellikle **standart normal dağılım** ile ilgili hesaplamalarda kullanılır.

**Z-skoru Formülü:**



Burada:

* Z: Z-skoru (standartlaştırılmış değer)
* X: İncelenen veri noktası
* μ: Veri setinin ortalaması
* σ: Veri setinin standart sapması

**Z-skorunun Kullanım Alanları:**

1. **Aykırı Değer Tespiti:**
   * Z-skoru belirli bir eşik değeri (genelde Z>3 veya Z<−3) aşarsa, o veri noktası **aykırı değer** olarak kabul edilir.
2. **Veri Normalizasyonu:**
   * Z-skoru, farklı ölçekteki veri setlerini standardize etmek için kullanılır. Bu, özellikle **makine öğrenmesi** ve istatistiksel modelleme gibi uygulamalarda faydalıdır.
3. **Olasılık Hesaplamaları:**
   * Z-skoru, standart normal dağılım tablosu kullanılarak belirli bir veri noktasının olasılığını hesaplamak için kullanılır.
4. **Karşılaştırmalar:**
   * Farklı ölçekteki verileri doğrudan karşılaştırmak için Z-skoru kullanılır.

Z-skoru, bir veri noktası ile dağılımın geri kalanı arasındaki farkı standartlaştırılmış bir şekilde ifade eder. Bu, veri analizi ve istatistiksel modellemede güçlü bir araçtır.

**Isolation Forest (İzolasyon Ormanı)**

Isolation Forest (İzolasyon Ormanı), anormal veya aykırı değerleri tespit etmek için kullanılan bir denetimsiz makine öğrenmesi algoritmasıdır. Temel fikri, aykırı değerlerin veri kümesinde daha kolay izole edilebilmesidir, çünkü aykırı değerler genellikle diğer verilere kıyasla daha seyrek bulunur ve farklı özellikler taşır. Bu yöntem özellikle yüksek boyutlu verilerde etkili ve hesaplama açısından verimli bir algoritmadır.

**Nasıl Çalışır?**

1. **Rastgele Alt Küme Oluşturma:**
   * Algoritma, veri kümesinden rastgele alt küme seçer ve bu alt kümelerde çalışır.
2. **Rastgele Ayrıştırma (Partitioning):**
   * Veriler, rastgele seçilen bir özellik ve bu özelliğin rastgele bir değeri kullanılarak bölünür.
   * Bölme işlemi, verileri belirli bir sınırdan (threshold) ayırarak bir **ikili ağaç (binary tree)** oluşturur.
3. **İzolasyon Yolu (Path Length):**
   * Ağaç yapısı oluşturulduktan sonra her veri noktası için kökten yaprağa kadar olan yolun uzunluğu hesaplanır.
   * Aykırı değerler, genellikle daha az bölünmeyle (kısa yol) izole edilebilir. Çünkü, aykırı değerler diğer veri noktalarından daha uzakta yer alır.
4. **Skor Hesaplama:**
   * İzolasyon yolları kullanılarak her bir veri noktasına bir **anormallik skoru** atanır.
   * Skor yüksekse (örneğin, score>0.5), veri noktası bir aykırı değer olarak değerlendirilir.

**Avantajları:**

Verimli**:** Çalışma süresi O(n logn) olduğu için, büyük veri setlerinde hızlı çalışır.

Yüksek Boyutlu Veriler**:** Yüksek boyutlu veri kümeleri üzerinde iyi performans gösterir.

Özelleştirilmiş Aykırılık Skoru**:** Her veri noktası için bir skor ürettiğinden, aykırı değerlerin derecesini belirlemek kolaydır.

**Dezavantajları:**

Rastgelelik: Rastgele ağaç oluşturma işlemi nedeniyle sonuçlar her çalıştırmada biraz farklı olabilir.

Karmaşık Dağılımlar: Karmaşık ve yoğun veri kümelerinde performansı düşebilir.

Dengeli Aykırılık: Veri setindeki aykırı değerlerin dağılımı çok dengesizse performans azalabilir.

**Örnek:**

Bir veri kümesi düşünün: [10,12,13,14,15,16,1000]

* Normal değerler yoğun olarak 10 ile 16 arasında.
* 1000, diğer değerlerden oldukça uzakta ve aykırı olarak tespit edilir.

Isolation Forest, 1000 değerini az sayıda bölünme ile izole eder ve buna yüksek bir anormallik skoru atar.

**Local Outlier (Yerel Aykırı Değer)**

Local Outlier , bir veri noktasının çevresindeki komşu veri noktalarına göre anormal bir davranış sergilemesi durumudur. Bu yaklaşım, global (küresel) dağılımdan ziyade, veri noktasının bulunduğu yerel çevreye odaklanır.

**LOF'un Çalışma Adımları:**

* + Her veri noktası için k-en yakın komşuları belirlenir. Bu, veri noktalarının birbirine olan mesafeleri kullanılarak yapılır.
  + Bir veri noktasının, komşularına olan ortalama erişim mesafesi hesaplanır.
  + Eğer bir veri noktası komşularına çok uzaksa, bu mesafe yüksek olacaktır.
  + LOF, bir veri noktasının yerel yoğunluğunu komşularının yerel yoğunluğuna kıyaslayarak hesaplar.



LOF(p)>1: Aykırı değer olma ihtimali yüksek.

LOF(p)≈1: Normal bir veri noktası.

LOF(p)<1: Yoğunluğu komşularından daha yüksek olan noktalar.

**Random Search**

Random Search, makine öğrenmesi modellerinde hiper parametre optimizasyonu için kullanılan bir tekniktir. Hiper parametre optimizasyonu, model performansını artırmak için en uygun hiper parametrelerin aranması sürecidir. Random Search, bu aramayı rastgele hiper parametre kombinasyonlarını deneyerek gerçekleştirir.

**Nasıl Çalışır?**

1. **Hiper parametre Alanı Belirlenir:** Optimizasyon yapılacak hiper parametrelerin aralıkları veya olası değerleri tanımlanır.
2. **Rastgele Seçim:**
   * Hiper parametre alanından rastgele bir kombinasyon seçilir.
3. **Model Eğitimi:**
   * Seçilen hiper parametre kombinasyonu ile model eğitilir ve performansı değerlendirilir.
4. **Tekrar:**
   * Belirlenen sayıda rastgele kombinasyon denenir.
5. **En İyisini Seçme:**
   * En iyi performansı sağlayan hiper parametre kombinasyonu seçilir.

**Random Search ile Grid Search Karşılaştırması**

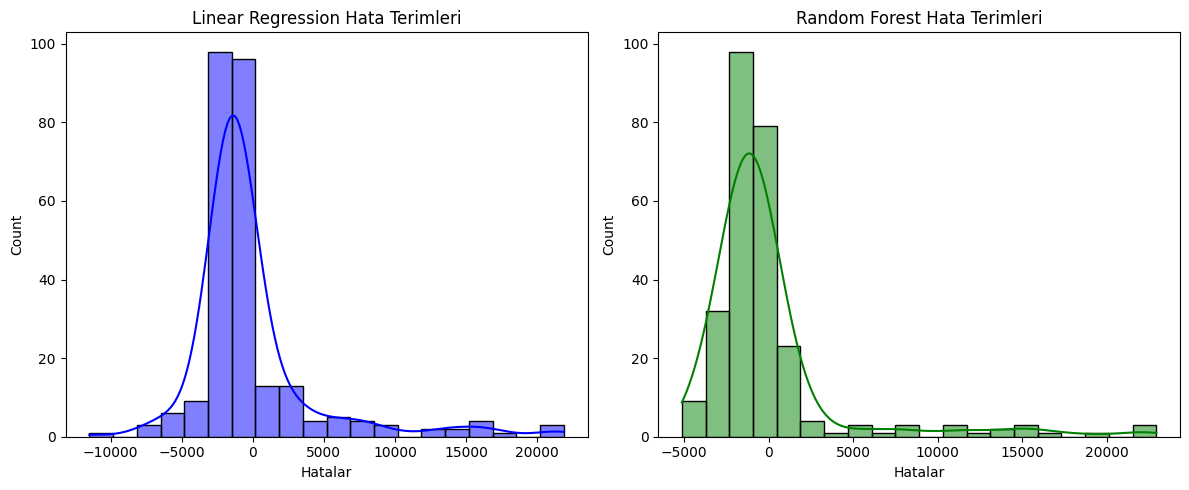
|  |  |  |
| --- | --- | --- |
| **Özellik** | **Grid Search** | **Random Search** |
| Arama Yöntemi | Tüm olası hiperparametre kombinasyonlarını sistematik olarak deniyor. | Hiperparametre uzayında rastgele noktalar seçerek değerlendiriyor. |
| Hesaplama Maliyeti | Yüksek. Özellikle hiperparametre sayısı arttıkça hesaplama süresi önemli ölçüde artar. | Genellikle Grid Search'e göre daha düşük. Ancak, şans faktörü nedeniyle optimum değere ulaşmamak riski vardır. |
| Optimum Değere Ulaşma Olasılığı | Tüm olası kombinasyonları denediği için teorik olarak en iyi değeri bulma olasılığı daha yüksek. | Tüm olası kombinasyonları denemediği için en iyi değeri bulma olasılığı daha düşüktür. Ancak, iyi bir başlangıç noktası sağlayabilir. |
| Kullanım Kolaylığı | Uygulaması nispeten kolaydır. | Uygulaması kolaydır. |
| Verimlilik | Düşük verimlidir. Özellikle yüksek boyutlu hiperparametre uzaylarında zaman alıcı olabilir. | Daha yüksek verimlidir. Özellikle yüksek boyutlu uzaylarda Grid Search'e göre daha hızlı sonuçlar verir. |
| Ne Zaman Kullanılır? | Tüm olası kombinasyonları denemek istendiğinde veya hesaplama kaynakları sınırlı değilse. | Hızlı bir şekilde iyi bir sonuç elde etmek istendiğinde veya yüksek boyutlu hiperparametre uzaylarında çalışıldığında. |

**Kullanım Alanları:**

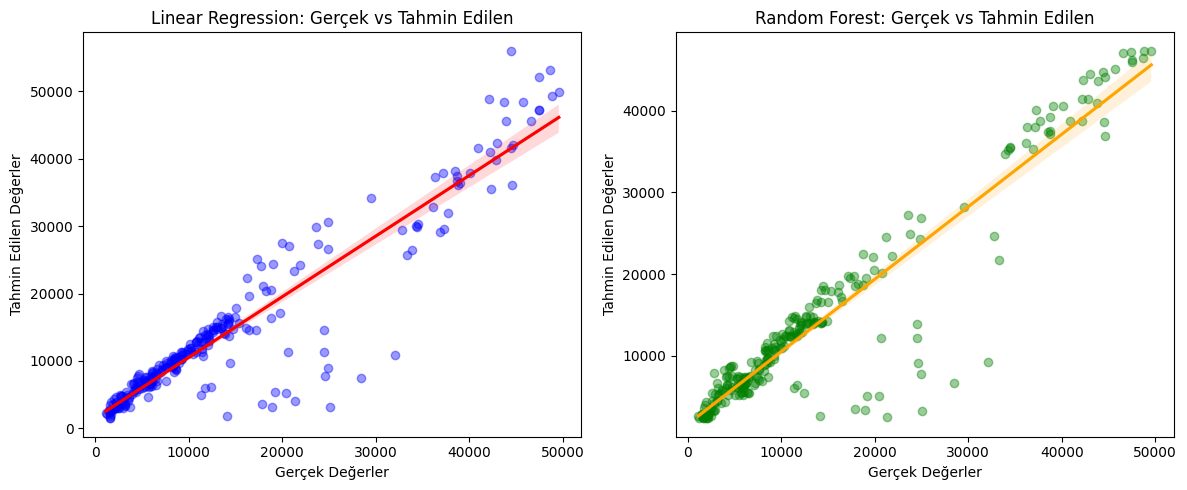
* Makine Öğrenmesi Modelleri: Hiperparametre optimizasyonu (örneğin, Random Forest, XGBoost).
* Derin Öğrenme Modelleri: Sinir ağı hiperparametrelerini optimize etmek.
* Genel Optimizasyon Problemleri: Karmaşık optimizasyon süreçlerinde hızlı çözümler.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Metrik | Anlamı | Değer Aralığı | Ne İfade Eder | Kullanım Alanları |
| RMSE | Kök Ortalama Kare Hata | 0 - ∞ | Tahmin hatalarının ortalama büyüklüğü | Regresyon, Sınıflandırma |
| R-Kare | Belirleme Katsayısı | 0 - 1 | Modelin verileri açıklama gücü | Regresyon |
| Açıklanan Varyans | Belirleme Katsayısı | 0 - 1 | Modelin verileri açıklama gücü | Regresyon |

**Grafikler**

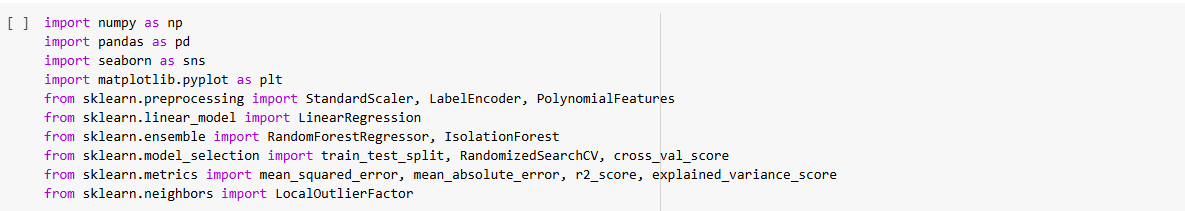
****

Şekil 1.0

****

Şekil 1.1

**Kütüphaneleri Ekleme**



**Veri Setini Okuma İlk 5 Satını Görüntüleme**

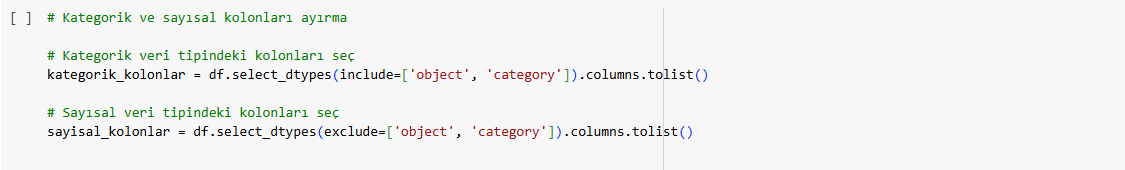
metin, yazı tipi, sayı, numara, ekran görüntüsü içeren bir resim

Açıklama otomatik olarak oluşturuldu

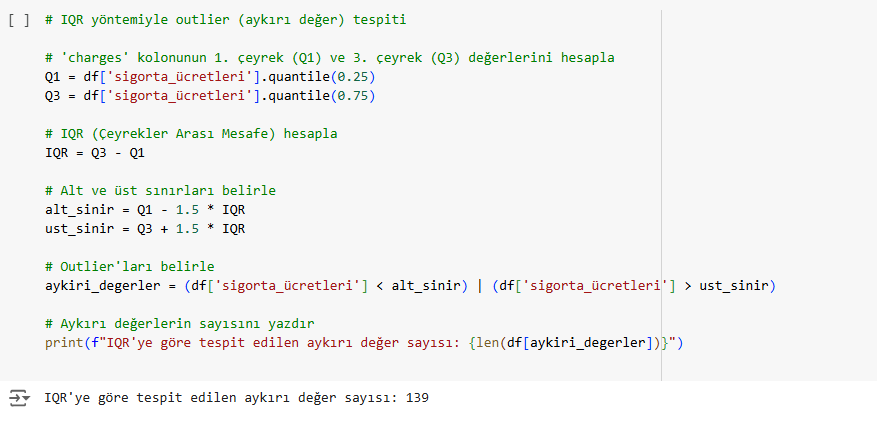
**Unique değerlerin bulunması**

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

Açıklama otomatik olarak oluşturuldu**

****

**Aykırı Değer Tespiti ( IQR)**

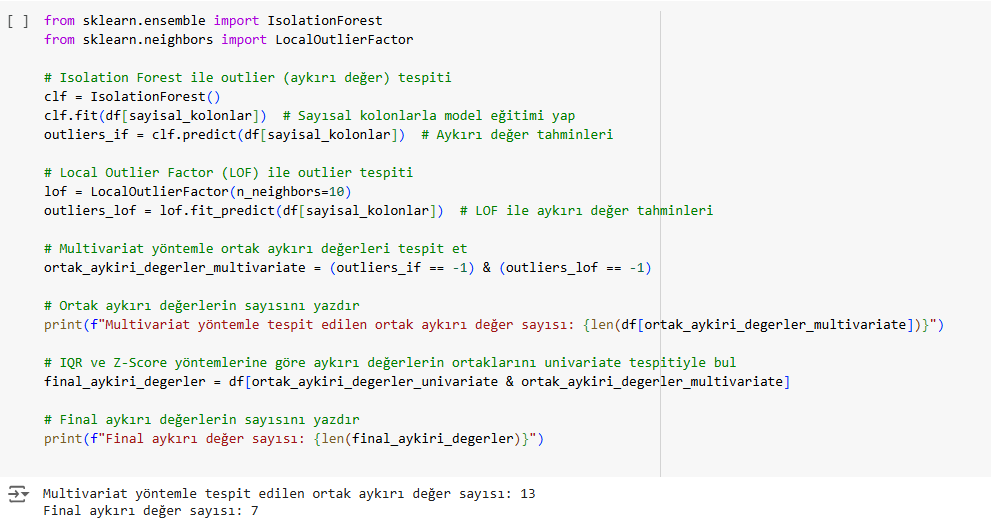
****

**Aykırı Değer Tespiti (Z-Score)**

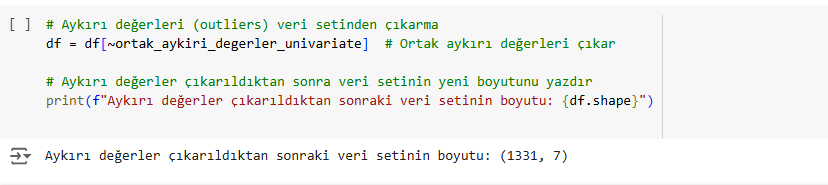
**metin, ekran görüntüsü, yazı tipi içeren bir resim

Açıklama otomatik olarak oluşturuldu**

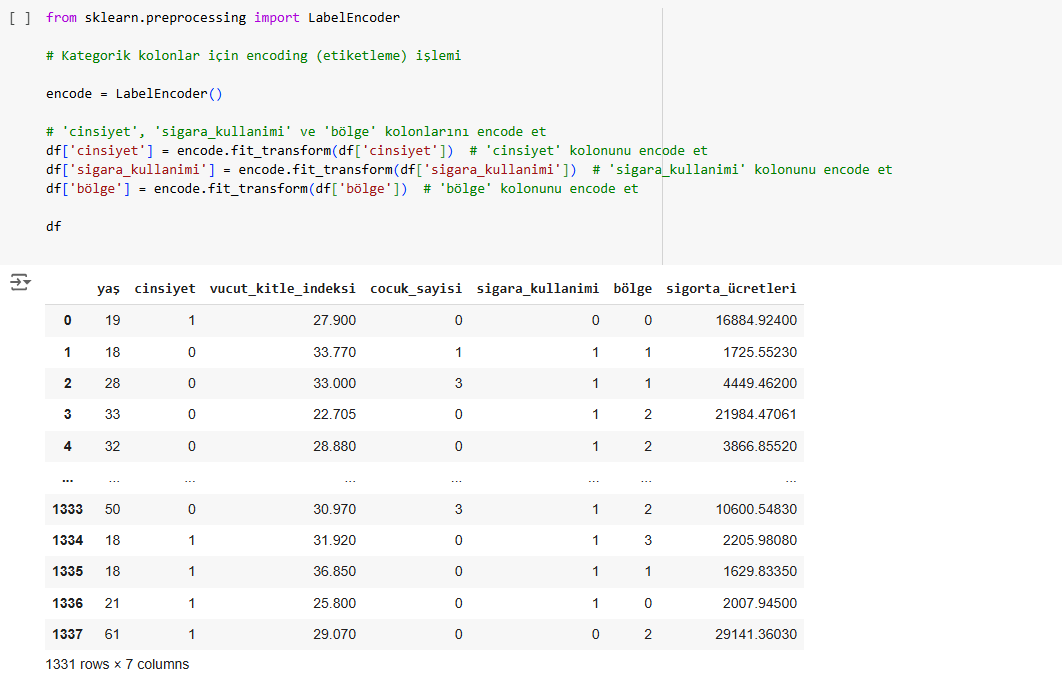
**Aykrı Değer Tespiti (Isolation Forest,** **Local Outlier)**

****

**Aykırı Değerlerin Veri Setinden Çıkarılması**

****

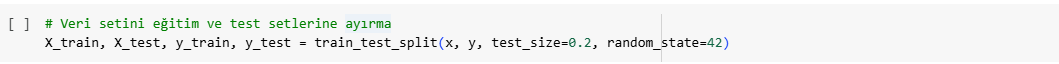
**Veri Ön İşleme**

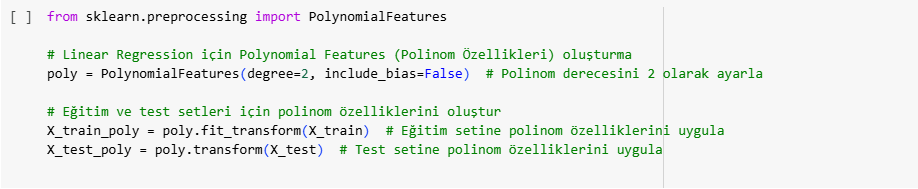
****

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

Açıklama otomatik olarak oluşturuldu**

**Model Ayarları**

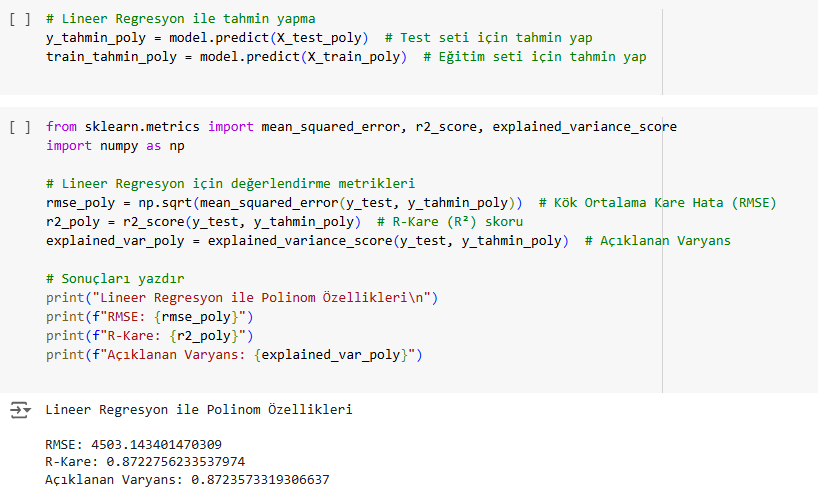
****

****

**Lineer Regresyon**

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

Açıklama otomatik olarak oluşturuldu**

****

**Random Forest**

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

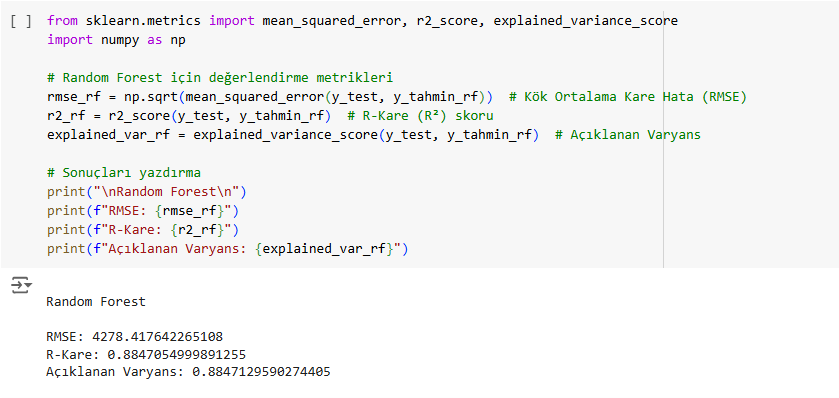
Açıklama otomatik olarak oluşturuldu**

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

Açıklama otomatik olarak oluşturuldu**

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

Açıklama otomatik olarak oluşturuldu**

****

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

Açıklama otomatik olarak oluşturuldu**