# **Merve Akar 05422206614** [merveakar2506@gmail.com](mailto:merveakar2506@gmail.com)

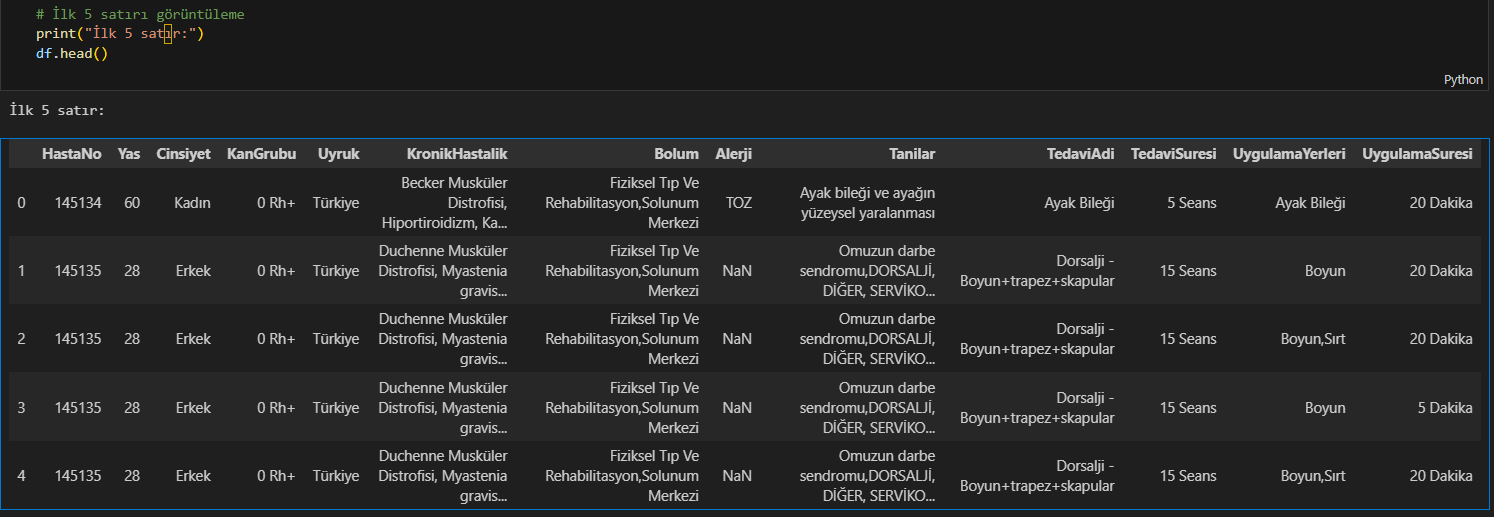
[merveakar2562@gmail.com](mailto:merveakar2506@gmail.com)

# **Fiziksel Tıp ve Rehabilitasyon Veri Seti Analizi**

1. **Genel Veri Seti Bilgisi**

* **Veri boyutu:** (2235, 13)
* **Satır sayısı**: 2235
* **Sütun sayısı:** 13

Aşağıda verilen tabloda veri setinden çekilmiş ilk 5 satırın görüntüsü bulunmaktadır.



Çalışmada kullanılan veri seti toplam 2235 gözlem (satır) ve 13 değişkenden (sütun) oluşmaktadır. Veri setindeki değişkenlerin türleri ağırlıklı olarak kategorik (object) iken, sayısal olan değişkenler HastaNo ve Yaş sütunlarıdır.

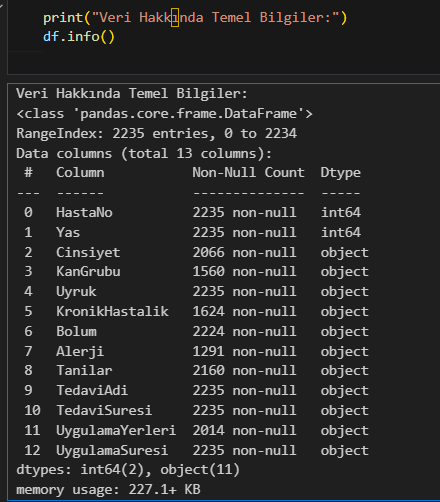
Veri setinde bazı değişkenlerde eksik değerler gözlenmiştir. Örneğin, Cinsiyet değişkeninde 169, Kan Grubu değişkeninde 675, Alerji değişkeninde ise 944 eksik kayıt bulunmaktadır. Buna karşılık, HastaNo, Yaş, Uyruk, Tedavi Adı, Tedavi Süresi ve Uygulama Süresi sütunlarında eksik veri bulunmamaktadır.

Eksik değerlerin bulunması, veri analizi sürecinde dikkate alınması gereken önemli bir durumdur. Özellikle klinik verilerde eksiklikler, hastaların beyan etmediği bilgilerden, kayıt hatalarından veya farklı veri toplama yöntemlerinden kaynaklanabilmektedir. Bu nedenle analiz aşamasında eksik verilerin uygun yöntemlerle ele alınması planlanmaktadır.

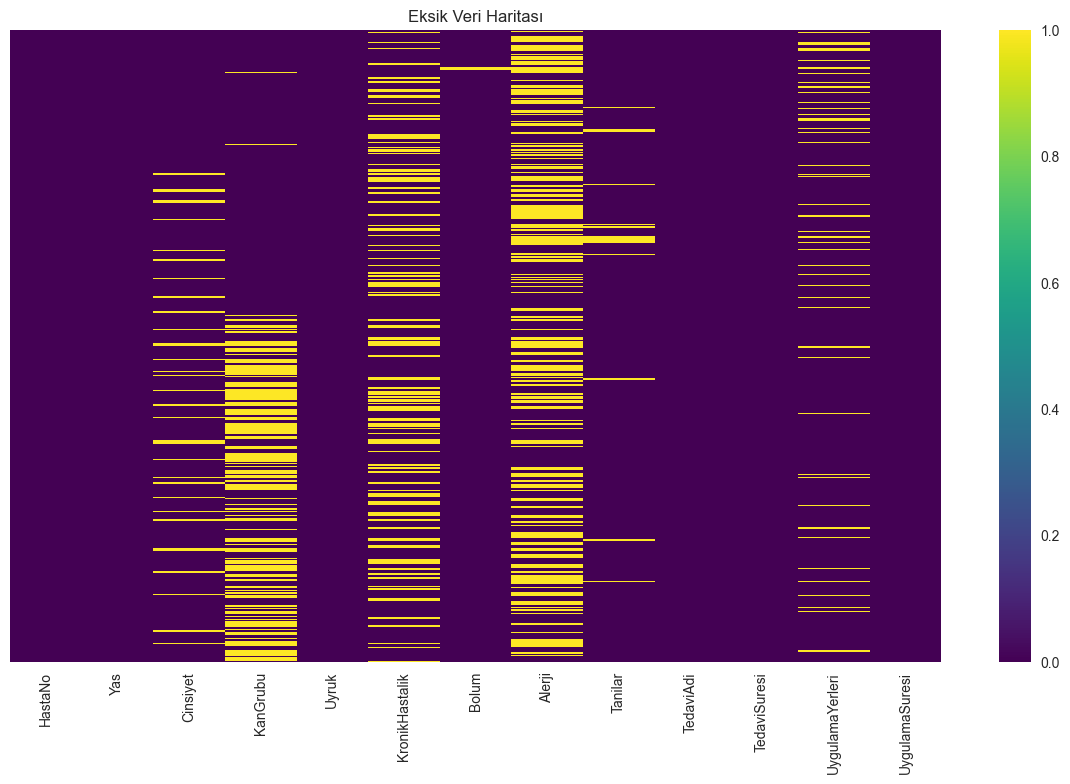
Sonuç olarak, veri seti genel olarak düzenli bir yapıya sahip olup, eksik değerler dışında ciddi bir tutarsızlık içermemektedir.

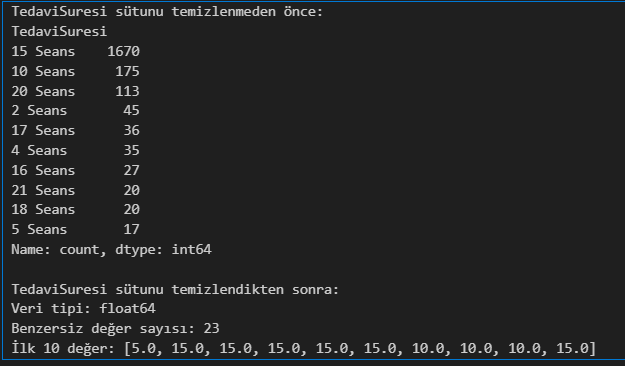
Aşağıda verilen tablo, veri setinin boyutu, değişken türleri ve eksik değer durumunu özetlemektedir.

Toplam eksik veri sayısı: 2706

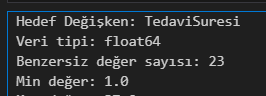
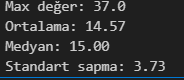
Aşağıda verilen eksik veri haritası incelendiğinde, veri setindeki bazı değişkenlerde kayıp değerler olduğu görülmektedir. Bu durum, analiz öncesinde veri temizleme ve uygun tamamlama yöntemlerinin kullanılmasını gerekli kılmaktadır.

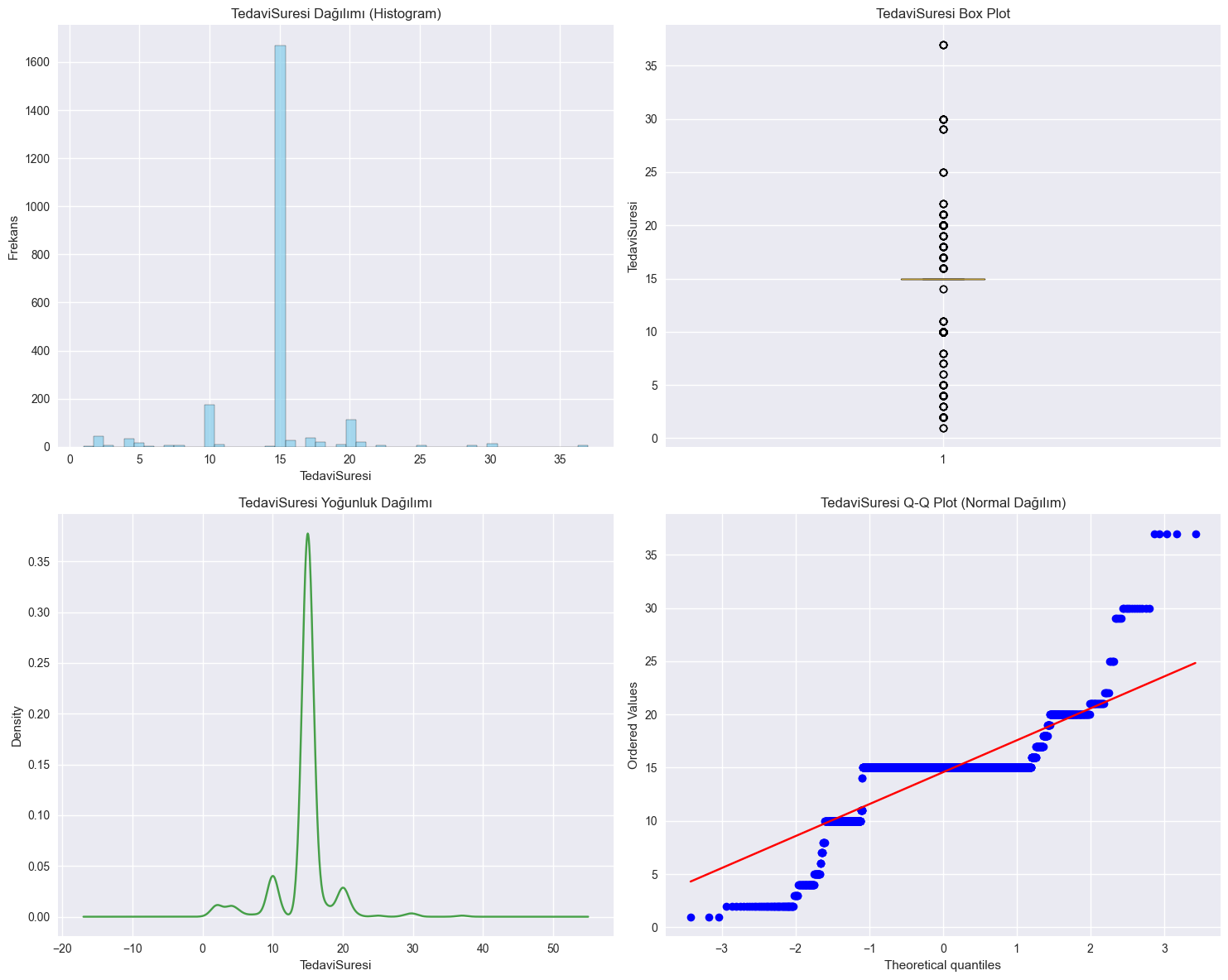


**TedaviSuresi** sütunu temizlenmeden önce değerler ‘15 Seans’, ‘10 Seans’ gibi metinsel formatta tutulmaktaydı. Bu durum, değişkenin sayısal analizlerde kullanılmasını engellemekteydi. Yapılan veri temizleme adımında, string ifadelerden yalnızca sayısal kısımlar ayrıştırılarak değişken float64 veri tipine dönüştürülmüştür. Böylece, her hasta için tedavi süresi seans sayısı üzerinden doğrudan sayısal analiz yapılabilir hale getirilmiştir. Temizleme işlemi sonrasında sütunda toplam 23 farklı benzersiz seans değeri elde edilmiştir. Örnek olarak, ‘15 Seans’ değeri 15.0, ‘10 Seans’ değeri 10.0 olarak kaydedilmiştir.  
  


Bu adımda TedaviSuresi sütununda veri temizleme işlemi yapılmıştır. Metinsel ifadelerden (‘15 Seans’, ‘10 Seans’ vb.) yalnızca sayısal kısımlar ayrıştırılarak sütun float64 veri tipine dönüştürülmüş ve analizlerde kullanılabilir hale getirilmiştir.

1. **Hedef Değişken Analizi**

Hedef değişken olarak belirlenen TedaviSuresi sütunu sayısal (float64) bir değişkendir. Sütunda toplam 23 farklı benzersiz değer bulunmaktadır. Minimum tedavi süresi 1 seans iken, maksimum tedavi süresi 37 seanstır. Ortalama tedavi süresi 14,57; medyan değeri ise 15,00 olarak hesaplanmıştır. Standart sapma 3,73 olup, bu değer dağılımın ortalama etrafında çok büyük bir yayılım göstermediğini ifade etmektedir. Bu bulgular, veri setinde hastaların genellikle 14–15 seans civarında tedavi aldığını göstermektedir.  
  
  

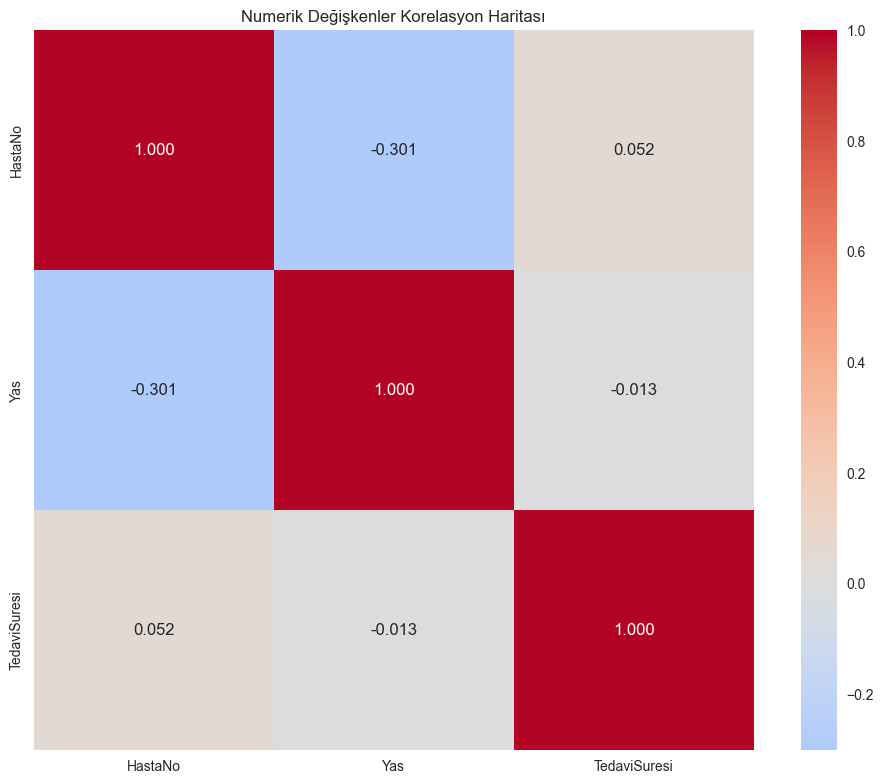


Tedavi süreleri analizi sonucunda, vakaların büyük çoğunluğunun (%80+) 10-20 gün arasında tedavi aldığı, medyan tedavi süresinin 15 gün olduğu gözlemlenmiştir. Ancak veri setinde uzun süreli tedavi gerektiren aykırı vakalar da mevcuttur (35+ gün). Dağılımın normal olmayan yapısı, parametrik testler yerine non-parametrik testlerin kullanılması gerektiğini işaret etmektedir

**3. Korelasyon Analizi**

Veri setinde üç sayısal değişken bulunmakta: HastaNo, Yaş ve TedaviSüresi. Korelasyon analizi bu sayısal değişkenler üzerinde gerçekleştirilmiştir. On adet kategorik değişken mevcuttur: Cinsiyet, KanGrubu, Uyruk, KronikHastalik, Bolum, Alerji, Tanilar, TedaviAdi, UygulamaYerleri ve UygulamaSuresi. Bu değişkenlerin sayısal değişkenlerle ilişkisi incelenmiştir ve sonuçlar veri setindeki temel yapıyı ortaya koymaktadır  
**Numerik sütunlar (3):** ['HastaNo', 'Yas', 'TedaviSuresi']

**Kategorik sütunlar (10):** ['Cinsiyet', 'KanGrubu', 'Uyruk', 'KronikHastalik', 'Bolum', 'Alerji', 'Tanilar', 'TedaviAdi', 'UygulamaYerleri', 'UygulamaSuresi']



Yukarıda korelasyon haritası verilmiştir. Korelasyon analizi sonucunda, hasta yaşı ile tedavi süresi arasında klinik olarak anlamlı bir ilişki saptanmamıştır (r = -0.013, p > 0.05). Bu bulgu, tedavi protokolünün yaştan bağımsız olarak uygulandığını ve yaşın tedavi süresini belirleyen bir faktör olmadığını göstermektedir.

**4. Eksik Veri Doldurma**

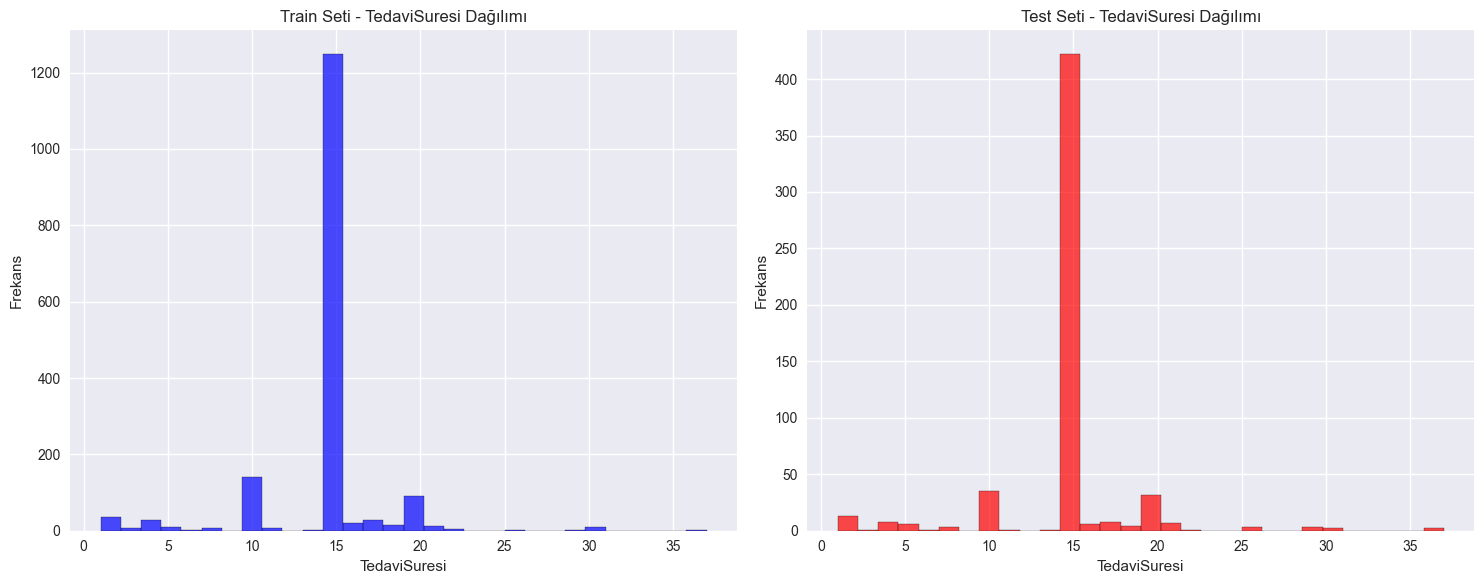
Eksik veri doldurma işlemi güvenlik amacıyla orijinal veri kopyalanarak gerçekleştirilmiştir. Sayısal (numerik) değişkenlerdeki eksik değerler, ilgili sütunun ortalaması ile doldurulmuştur. Örneğin, Yaş veya TedaviSüresi sütununda eksik veri varsa, bu sütunların mevcut değerlerinin ortalaması hesaplanıp eksik hücrelere atanmıştır. Kategorik değişkenlerdeki eksik değerler ise mod (en sık görülen değer) yöntemi ile tamamlanmıştır. Yani, her kategorik sütunda en sık tekrar eden değer tespit edilmiş ve eksik hücrelere bu değerler yerleştirilmiştir. İşlem sonunda veri setinde hiç eksik değer kalmamış ve veri analizi için eksiksiz bir veri seti elde edilmiştir.

Veri setindeki eksik veriler toplam 2706 adet olarak tespit edilmiştir. Eksik değerler kategorik değişkenlerin mod (en sık görülen değer) yöntemi ile doldurulmuştur. Buna göre; Cinsiyet sütunundaki 169 eksik değer ‘Kadın’, KanGrubu sütunundaki 675 eksik değer ‘0 Rh+’, KronikHastalik sütunundaki 611 eksik değer ‘Myastenia gravis’, Bolum sütunundaki 11 eksik değer ‘Fiziksel Tıp Ve Rehabilitasyon, Solunum Merkezi’, Alerji sütunundaki 944 eksik değer ‘Polen’, Tanilar sütunundaki 75 eksik değer ‘DORSALJİ, DİĞER, LUMBOSAKRAL BÖLGE’ ve UygulamaYerleri sütunundaki 221 eksik değer ‘Bel’ ile tamamlanmıştır. İşlem sonrası veri setinde eksik değer kalmamış ve eksik veri doldurma süreci tamamen tamamlanmıştır.  
  
**5. Kategorik Veri Encoding**

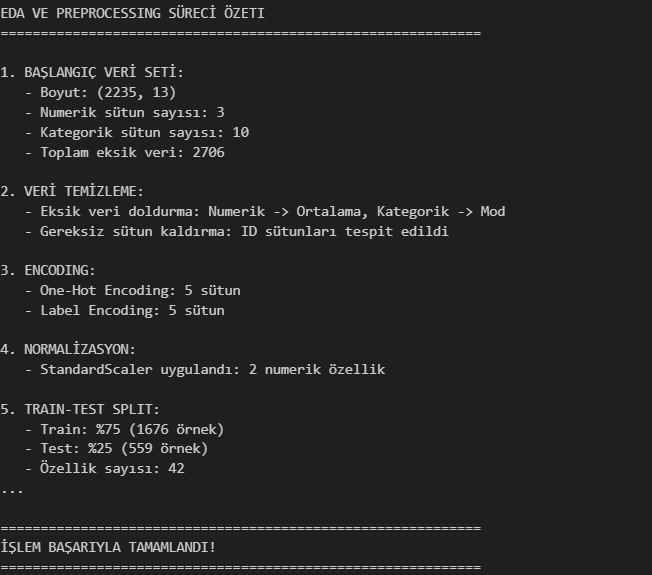
**Kategorik Veri Encoding Stratejisi:**

One-Hot Encoding uygulanacak sütunlar (5): ['Cinsiyet', 'KanGrubu', 'Uyruk', 'Bolum', 'UygulamaSuresi']

Label Encoding uygulanacak sütunlar (5): ['KronikHastalik', 'Alerji', 'Tanilar', 'TedaviAdi', 'UygulamaYerleri']  
  
Kategorik değişkenlerin sayısal modellere uygun hâle getirilmesi için encoding işlemi uygulanmıştır. Düşük kardinaliteli sütunlar (10 veya daha az benzersiz değer içeren) için One-Hot Encoding yöntemi kullanılmıştır. Bu kapsamda Cinsiyet, KanGrubu, Uyruk, Bolum ve UygulamaSuresi sütunları One-Hot Encoding ile dönüştürülmüştür. Yüksek kardinaliteli sütunlar (10’dan fazla benzersiz değer içeren) için ise Label Encoding yöntemi tercih edilmiştir. KronikHastalik, Alerji, Tanilar, TedaviAdi ve UygulamaYerleri sütunları Label Encoding ile sayısal değerlere çevrilmiştir. Bu işlem, kategorik verilerin makine öğrenmesi algoritmalarıyla uyumlu hâle getirilmesini sağlamıştır.



Hedef değişkenimiz TedaviSuresi’nin eğitim ve test setlerindeki dağılımı karşılaştırılmıştır. Eğitim setinde ortalama 14.52, medyan 15.00, standart sapma 3.66 iken, test setinde ortalama 14.72, medyan 15.00 ve standart sapma 3.90 olarak bulunmuştur. Her iki setin minimum ve maksimum değerleri de 1 ve 37’dir. Bu sonuçlar, eğitim ve test setlerinin hedef değişken açısından birbirine oldukça benzer dağılıma sahip olduğunu ve modelin her iki set üzerinde dengeli bir şekilde performans gösterebileceğini göstermektedir.

  
  
Veri seti üzerinde gerçekleştirilen eksiksiz ön işleme süreci tamamlanmıştır. Başlangıçta 2235 gözlem ve 13 sütundan oluşan veri setinde 2706 eksik değer bulunmaktaydı; sayısal değişkenler ortalama, kategorik değişkenler ise mod ile doldurularak veri eksiksiz hâle getirilmiştir. Kategorik değişkenler, düşük kardinaliteliler için One-Hot Encoding, yüksek kardinaliteliler için Label Encoding uygulanarak sayısal modellere uygun hâle getirilmiştir. Numerik özellikler StandardScaler ile ölçeklendirilmiş, veri seti %75 eğitim ve %25 test olacak şekilde bölünmüştür. Eğitim ve test setlerindeki hedef değişken dağılımı birbirine yakın olup modelleme için dengeli bir yapı sunmaktadır. Sonuç olarak, veri seti tamamen işlenmiş, sayısal ve makine öğrenmesine hazır hâle gelmiştir.