# Veri Kümesi Hakkında

## Veri Kümesindeki Değişkenler ve Özellikleri

### Demografik Bilgiler

**Hasta Numarası (hasta\_no):** Hasta kimlik numarası (anonimleştirilmiş).

**Yaş (yas):** Hastanın yaşı.

**Cinsiyet (cinsiyet):** Hastanın cinsiyeti (1: Erkek, 2: Kadın).

Eğitim Düzeyi (egitim\_duzeyi): Hastanın eğitim düzeyi.

Meslek (meslek): Hastanın mesleği.

### Yaşam Tarzı ve Alışkanlıklar

Sigara Kullanımı Durumu (sigara\_kullanimi):

1: Hiç kullanmamış,

2: Kullanmış ama bırakmış,

3: Aktif kullanıyor.

Sigara İçme Süresi (sigara\_birakan\_ne\_kadar\_gun\_icmis): Sigara bırakanların toplam sigara içme süresi (gün).

Sigara İçme Adeti (sigara\_birakan\_gunde\_kac\_adet\_icmis): Sigara bırakanların günde ortalama içtikleri sigara sayısı.

Sigara Bırakma Süresi (ne\_zaman\_birakmis\_gun): Sigara bırakılmasından bu yana geçen süre (gün).

### Tıbbi Durumlar ve Aile Öyküsü

Ailede KOAH veya Astım Tanısı (ailede\_koah\_veya\_astim\_tanili\_hasta\_var\_mi):

1: Evet,

0: Hayır.

Hasta Aile Bireyleri (varsa\_kimde\_anne/baba/kardeş/diger): Eğer varsa tanı almış aile bireyleri.

Akciğer Fonksiyon Testleri

**FEV1 Değeri (FEV1):** Zorlu ekspiratuvar volüm (bir saniyede çıkartılan hava miktarı).

**FEV1 Yüzdesi (FEV1\_yuzde):** FEV1 değerinin normal popülasyona göre yüzdesi.

PEF Değeri (PEF): Pik ekspiratuvar akış hızı.

PEF Yüzdesi (PEF\_yuzde): PEF değerinin normal popülasyona göre yüzdesi.

FEV1/FVC Oranı (FEV1\_FVC\_sDeğeri): FEV1'in FVC'ye (zorlu vital kapasite) oranı.

# Veri Ön İşleme

## 2.1 Eksik Gözlem Analizi

## metin, ekran görüntüsü, menü, yazı tipi içeren bir resim Açıklama otomatik olarak oluşturuldu

**(missing\_values\_report.csv)**

**Sigara Kullanımına İlişkin Veriler**

Sigara kullanımıyla ilgili değişkenlerde önemli eksiklikler bulunmaktadır. Özellikle, **sigara\_birakan\_ne\_kadar\_gun\_icmis**, **sigara\_birakan\_gunde\_kac\_adet\_icmis**, ve **ne\_zaman\_birakmis\_gun** değişkenlerinde eksik veri oranı %55.8'dir. **sigara\_devam\_eden\_gunde\_kac\_adet\_iciyor** değişkeninde eksik veri oranı daha da yüksek olup %78.2 seviyesindedir.

Eksik verilerin bu denli yüksek olmasının nedeni, sigara kullanmayan veya hiç bırakmamış bireylerden kaynaklanıyor. Bu tür eksiklikler için, "yok" anlamına gelen bir kategori eklenerek doldurma işlemi gerçekleştireceğiz.

**Solunum Fonksiyon Testlerine İlişkin Veriler**

Solunum fonksiyon testleriyle ilgili değişkenlerde de eksik veriler bulunmaktadır. **FEV1** ve **PEF** değişkenlerindeki eksiklik oranı %32.93 olup, oldukça dikkate değerdir. Ayrıca, **PEF\_yuzde** değişkenindeki eksiklik oranı daha düşüktür (%0.40). Eksik verilerin doldurulması için daha doğru ve istatistiksel olarak sağlam yöntemler kullanılacaktır. Bu bağlamda:

* Eksik veriler başlangıçta çıkarılarak eksiksiz veriler üzerinde modelleme yapılacaktır.
* **Çapraz doğrulama** ile desteklenen **lineer regresyon** ve **random forest** regresyon modelleri, eksik değerlerin tahmini için kullanılacaktır.
* Bu yöntemler, mevcut değişkenler arasındaki ilişkilerden yararlanarak eksik değerlerin tahmin edilmesini sağlayacaktır.

**Ailede Tanılı Hasta Bilgileri**

**varsa\_kimde\_anne**, **varsa\_kimde\_baba**, **varsa\_kimde\_kardes**, ve **varsa\_kimde\_diger** değişkenlerinde eksik veri oranı oldukça düşüktür (%0.2 - %0.6). Eksik verilerin olmasının nedeni, ailede hasta olmamasından dolayı eksik bırakılmasından kaynaklanıyor. Bu tür eksiklikler için "yok" anlamına gelen bir kategori değişken ile doldurma işlemi gerçekleştireceğiz.

**Tanı Süresi Verileri**

tani\_suresi\_yil ve tani\_suresi\_ay değişkenlerinde eksik veri oranı %1.20 düzeyindedir. Bu oran, diğer değişkenlere kıyasla düşük olup, eksikliklerin doldurulmasında basit yöntemler kullanılabilir. Eksiklikler, tanının yakın geçmişte konmuş olmasından kaynaklanıyor olabilir.