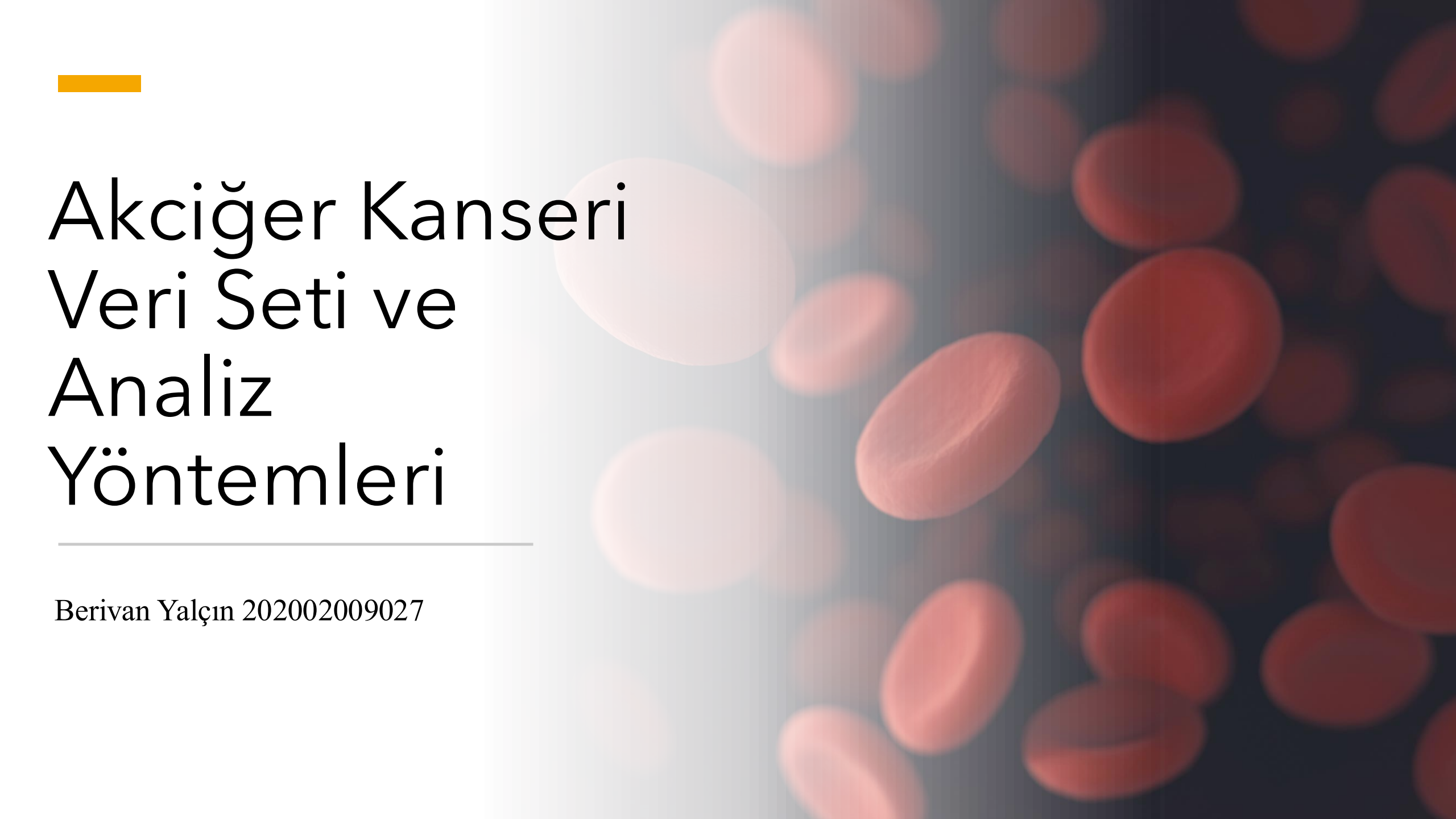
Akciğer Kanseri Veri Seti ve Analiz Yöntemleri

Berivan Yalçın 202002009027



İçerik

- Giriş
- Veri Analizi Nedir ?
- Veri Seti Tanıtımı
- Analizde Kullanılan Testler
- Kapanış

Veri Analizi Nedir ?

Veri analizi, büyük miktardaki verileri inceleyerek, örüntüler, ilişkiler ve anlamlı bilgiler çıkarmak için istatistiksel ve matematiksel yöntemleri kullanma sürecidir. Veri analizi genellikle bilgiye dayalı kararlar almak, tahminler yapmak veya bir sorunu çözmek için kullanılır. Bu süreç, veri toplama, veriyi temizleme, veriyi modelleme, analiz etme ve sonuçları yorumlama gibi aşamalardan oluşur.

Veri Seti Tanıtımı

- Bu sunumda, kanser hastalarının demografik, tıbbi ve tedavi süreçlerine dair bilgilerini içermektedir. Her satır bir hastayı temsil etmektedir ve çeşitli özellikler hakkında bilgi sunmaktadır.

- Aşağıda, veri setindeki her bir sütunun ne anlama geldiğini ve içerdiği bilgileri açıklanmaktadır.

Veri Setinin İçerdiği Sütunlar

- **id**: Hastaya atanmış benzersiz kimlik numarası.
- **age**: Hastanın yaşı (yıl olarak).
- **gender**: Hastanın cinsiyeti (Female: Kadın, Male: Erkek).
- **country**: Hastanın yaşadığı ülke.
- **diagnosis_date**: Kanser teşhisinin konulduğu tarih.
- **cancer_stage**: Kanserin evresi (Stage I, Stage II, Stage III, Stage IV).
- **beginning_of_treatment_date**: Tedaviye başlanan tarih.
- **family_history**: Ailede kanser öyküsü olup olmadığı (Yes: Var, No: Yok).
- **smoking_status**: Hastanın sigara içme durumu (Never Smoked: Hiç içmemiş, Former Smoker: Daha önce içmiş, Passive Smoker: Pasif içici,Current Smoker:Sigara içen).

Veri Setinin İçerdiği Sütunlar

- **bmi**: Vücut Kitle İndeksi (BMI - Body Mass Index).
- **cholesterol_level**: Hastanın kolesterol seviyesi (mg/dL).
- **hypertension**: Hipertansiyon durumu (1: Var, 0: Yok).
- **asthma**: Astım durumu (1: Var, 0: Yok).
- **cirrhosis**: Siroz durumu (1: Var, 0: Yok).
- **other_cancer**: Başka bir kanser türü olup olmadığı (1: Var, 0: Yok).
- **treatment_type**: Uygulanan tedavi türü (Chemotherapy: Kemoterapi, Radiation: Radyoterapi, Surgery: Cerrahi müdahale, Combined: Kombine tedavi).
- **end_treatment_date**: Tedavinin sona erdiği tarih.
- **survived**: Hastanın hayatta kalma durumu (1: Hayatta, 0: Hayatta değil).

Kullanılan Kütüphaneler

Pandas

Veri analizi ve veri manipülasyonu için kullanılan güçlü bir Python kütüphanesidir. Yapısal veri setleri ile çalışmayı kolaylaştırır.

NumPy

Sayısal hesaplamalar ve büyük, çok boyutlu diziler ve matrislerle çalışmak için kullanılan temel bir Python kütüphanesidir.

Matplotlib

2D grafikler ve görselleştirmeler oluşturmak için kullanılan kapsamlı bir Python kütüphanesidir.

SciPy

Bilimsel ve teknik hesaplamalar için kullanılan açık kaynaklı bir Python kütüphanesidir. NumPy üzerine inşa edilmiştir ve daha geniş bir fonksiyon yelpazesi sunar.

Kullanılan Kütüphaneler

Seaborn: istatistiksel veri görselleştirme için kullanılan bir Python kütüphanesidir. Matplotlib üzerine inşa edilmiştir ve daha çekici grafikler sunar.

Missigno: Veri eksikliği ve düzensizlikleriyle başa çıkmak için kullanılan bir Python kütüphanesidir.

Folium: Python için harita görselleştirme kütüphanesidir. OpenStreetMap veya diğer harita sağlayıcılarını kullanarak interaktif haritalar oluşturmanızı sağlar.

Geopy: Geopy, coğrafi konumlarla ilgili işlemleri gerçekleştirmek için kullanılan bir Python kütüphanesidir.

Veri seti Kaynağı

- Bu veri seti, hastaların yaş, cinsiyet, yaşadıkları ülke, sigara içme durumu, vücut kitle indeksi, kolesterol seviyesi, hipertansiyon, astım gibi sağlık durumları ve uygulanan tedavi türleri gibi bilgileri içerir. Ayrıca, hastaların tedavi süreci ve hayatta kalma durumları hakkında da bilgi sunmaktadır. Bu tür veriler, kanser araştırmaları, tedavi planlaması ve hastaların sağlık durumlarının analizi için kullanılabilir.

- Veri seti kaynağı:https://www.kaggle.com/datasets/masterdatasan/lung-cancer-mortality-datasets-v2/data

Veri Setinde Uygulanacak Testler

- **Kolmogorov-Smirnov Testi:** Büyük bir veri seti ile çalıştığımız için bu testi kullanmak daha mantıklı olacaktır. Bu test ile normallik durumunu kontrol edeceğiz.

- **T-Testi:** T-testi, iki grup arasında ortalamalar arasındaki farkın istatistiksel olarak anlamlı olup olmadığını belirlemek için kullanılan bir hipotez testidir.

- **Levene Testi:** Levene testi, grupların varyanslarının homojen olup olmadığını kontrol eden bir istatistiksel testtir. Grupların varyansları eşitse, parametrik testlerin güvenilirliğini artırır. Eğer varyanslar homojen değilse, parametrik olmayan alternatifler değerlendirilebilir.

Veri Setinde Uygulanacak Testler

- **Chi-kare testi:** Chi-kare testi (χ^2 testi), iki kategorik değişken arasındaki ilişkiyi test etmek için kullanılan istatistiksel bir testtir.

- **Korelasyon Testi:** iki sürekli değişken arasındaki doğrusal ilişkiyi ölçmek için kullanılan bir testtir. Pearson korelasyon katsayısı en yaygın olanıdır ve -1 ile 1 arasında bir değer alır; bu değerler değişkenler arasındaki ilişkinin yönü ve gücü hakkında bilgi verir.

- **Smirnov Testi:** Kolmogorov-Smirnov (K-S) testi, iki örnek dağılımı veya bir örnek dağılımı ile teorik bir dağılım arasındaki farkı test etmek için kullanılan non-parametrik bir testtir. En büyük mutlak farkı hesaplayarak dağılımlar arasındaki uyumu değerlendirir.

Veri Setinde Uygulanacak Testler

- **Kaplan-Meier**: Kaplan-Meier yöntemi, sağkalım analizinde kullanılan ve belirli bir zaman diliminde olayın gerçekleşme olasılığını tahmin eden bir istatistiksel yöntemdir. Sağkalım eğrisi, zaman içinde sağ kalan bireylerin oranını gösterir.

- **ANOVA (Analysis of Variance)**: ANOVA, birden fazla grup arasındaki ortalama farklarını karşılaştırmak için kullanılan bir istatistiksel testtir. Gruplar arasındaki farkların istatistiksel olarak anlamlı olup olmadığını belirler.

Beni Dinlediğiniz İçin Teşekkür Ederim