DATA PREPARATION

Kelompok 2

Anggota Kelompok

01

Farah Nasywa

2208107010051

02

Iwani Khairina

2208107010078

03

Dinda Maharani

2208107010081

Dataset

01

Dataset yang digunakan dalam laporan ini merupakan dataset The_Cancer_data_1500_V2.csv yang mana kami dapatkan dari Kaggle.com. Berisi informasi mengenai faktor-faktor risiko kanker pada individu. Setiap baris dalam dataset merepresentasikan seorang individu dengan berbagai atribut yang dapat mempengaruhi kemungkinan terkena kanker.

link dataset :
https://www.kaggle.com/datasets/rabieelkharoua/cancer-prediction-dataset

Data Loading

02

Membaca Dataset Dataset diunggah dalam format CSV (Comma-Separated Values), sehingga kita dapat menggunakan pustaka pandas untuk memuatnya ke dalam DataFrame.

```
import numpy as np  # Untuk operasi aljabar linear dan komputasi numerik
import pandas as pd  # Untuk pemrosesan data dan membaca file CSV
import matplotlib.pyplot as plt  # Untuk visualisasi data dasar
import seaborn as sns  # Untuk visualisasi data yang lebih kompleks dan menarik

import os
for dirname, _, filenames in os.walk('/kaggle/input'):
    for filename in filenames:
        print(os.path.join(dirname, filename))
```

DATA DESCRIPTION

01

Age

Usia individu

02

BMI

Indeks Massa Tubuh (Body Mass Index).

03

Smoking

Status merokok individu (0 = Tidak Merokok, 1 = Merokok).

04

Alcohol Intake

Konsumsi alkohol individu

05

Genetic risk

Risiko genetik terkait kanker (0 = Tidak Ada Risiko, 1 = Risiko Sedang, 2 = Risiko Tinggi).

06

Physical Activity

Frekuensi aktivitas fisik individu.

07

Cancer History

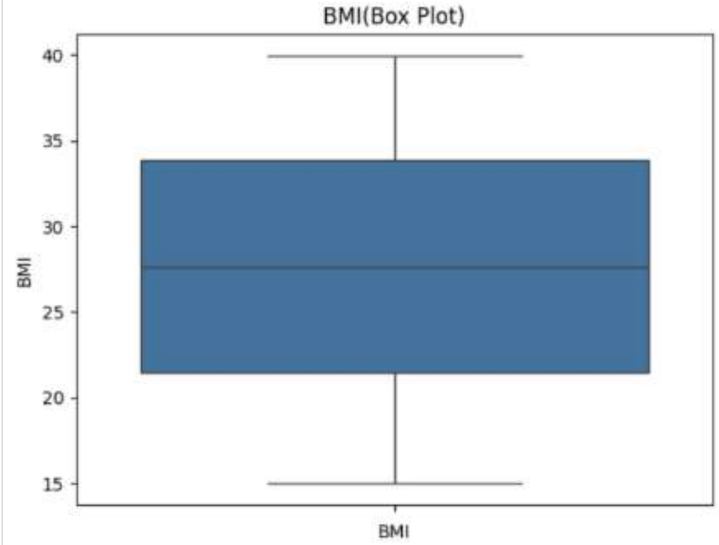
Riwayat kanker dalam keluarga (0 = Tidak, 1 = Ya).

08

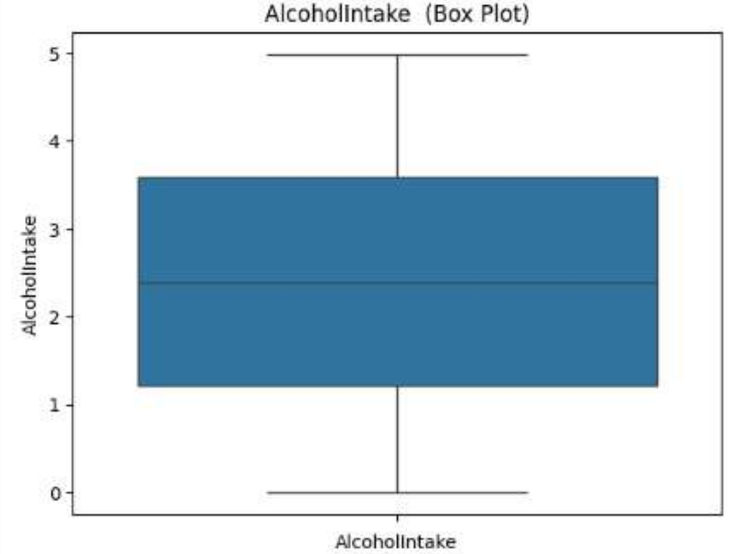
Diagnosis

Diagnosis kanker pada individu (0 = Tidak Terdiagnosis, 1 = Terdiagnosis).

DATA CLEANING



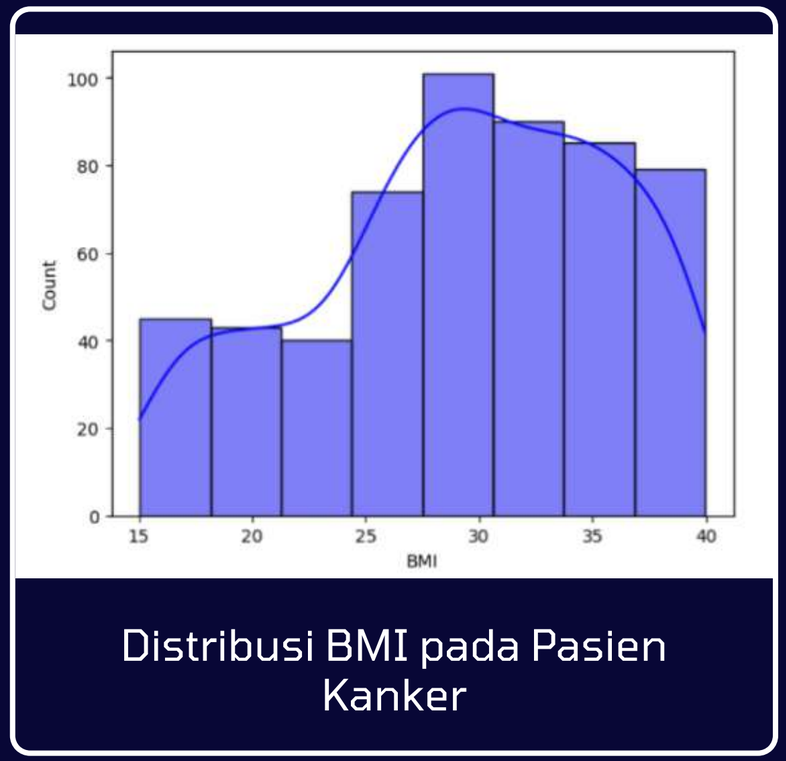
Distribusi nilai BMI pada dataset.



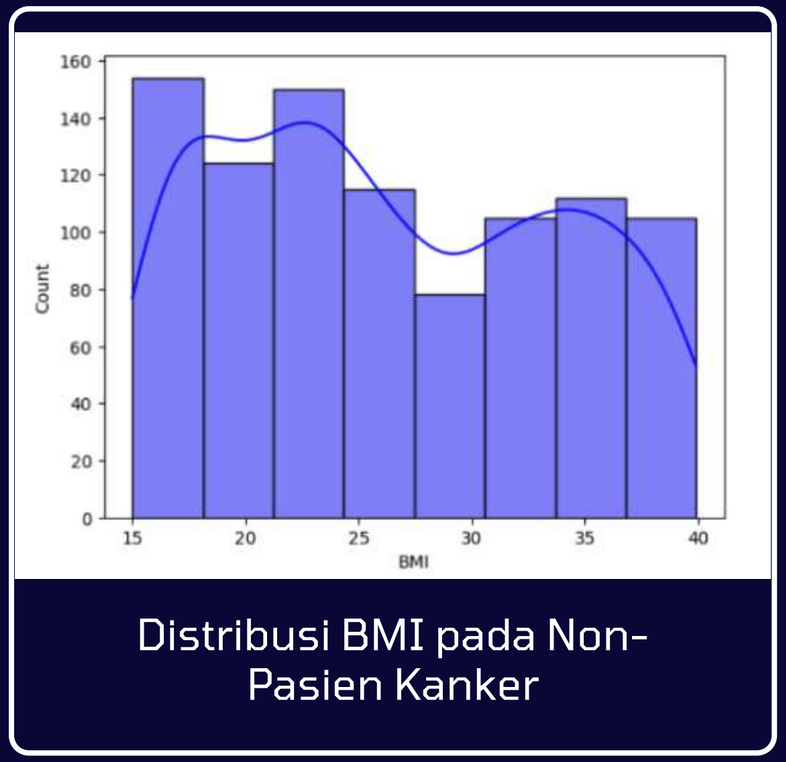
Distribusi konsumsi alkohol pada dataset.

DATA UNDERSTANDING

Histogram Distribusi BMI pada Pasien Kanker dan Non-Kanker

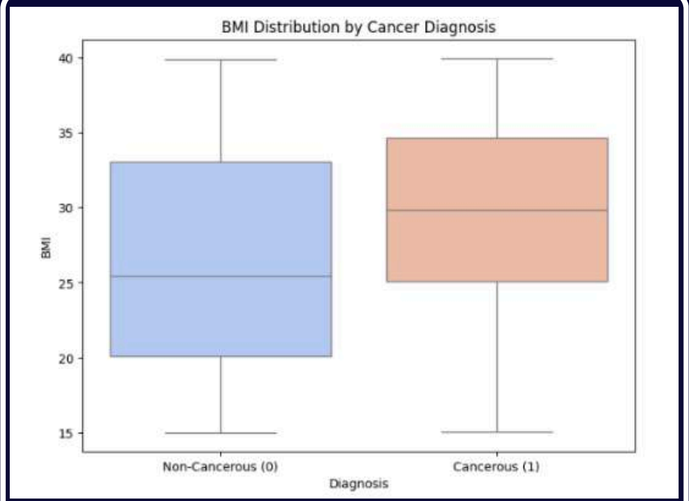


Distribusi BMI pada Pasien Kanker

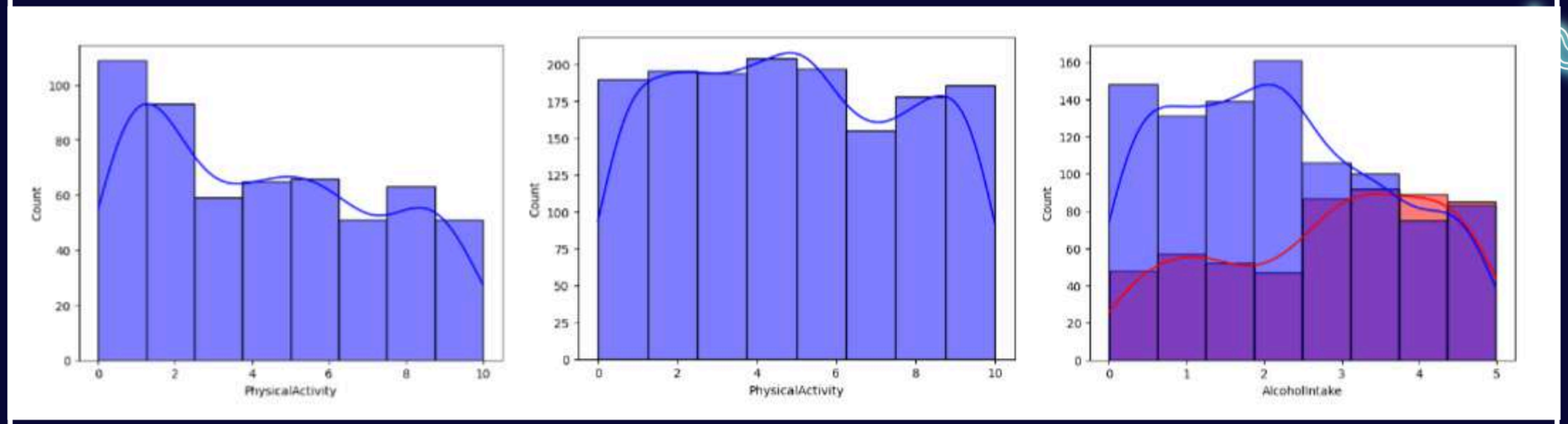


Distribusi BMI pada Non-Pasien Kanker

DATA UNDERSTANDING

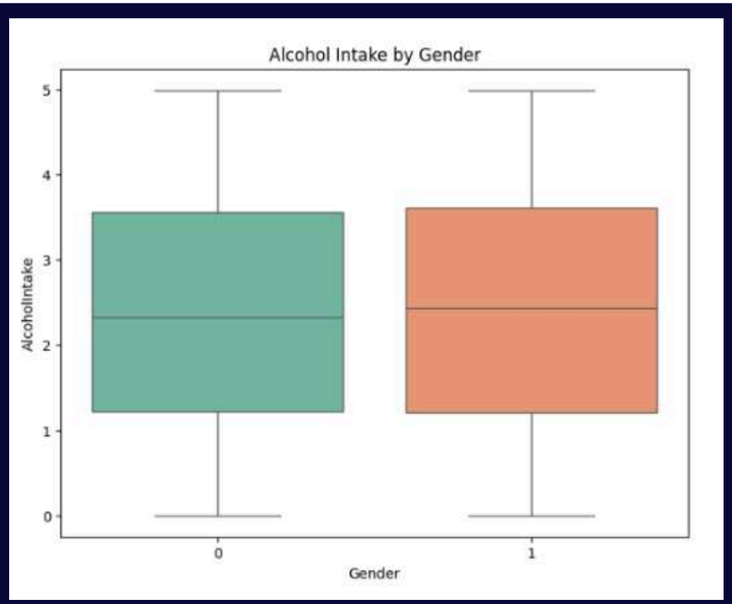


Boxplot Distribusi BMI
Berdasarkan Diagnosis Kanker



Histogram Distribusi Aktivitas
Fisik pada Pasien

VISUALISASI DATA



Boxplot Konsumsi Alkohol Berdasarkan Jenis Kelamin

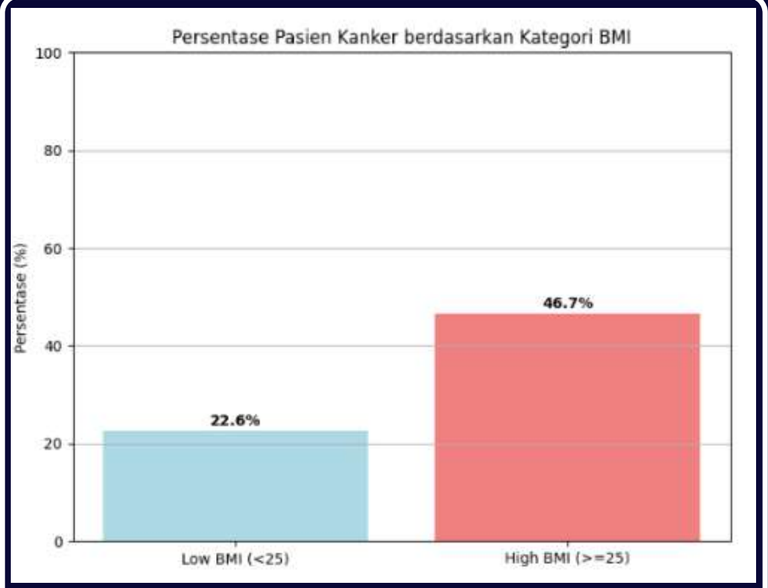


Diagram Batang Persentase pasien kanter berdasarkan kategori BMI

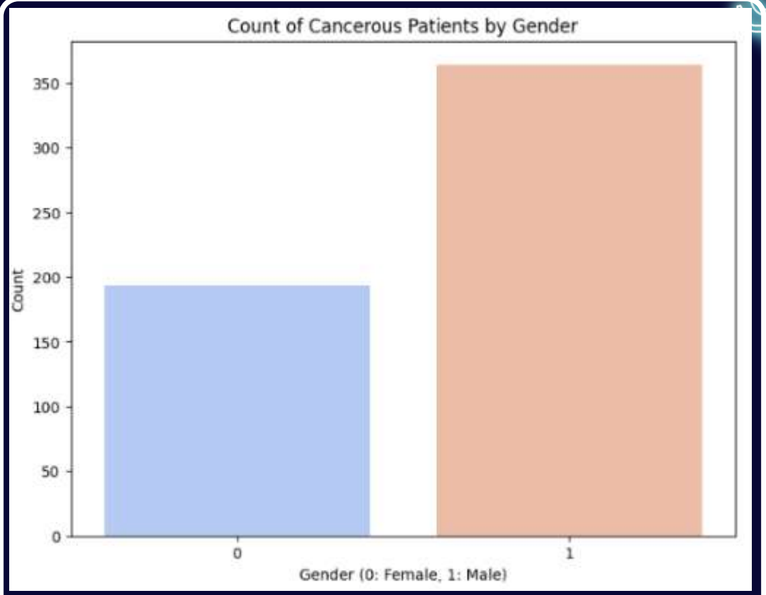
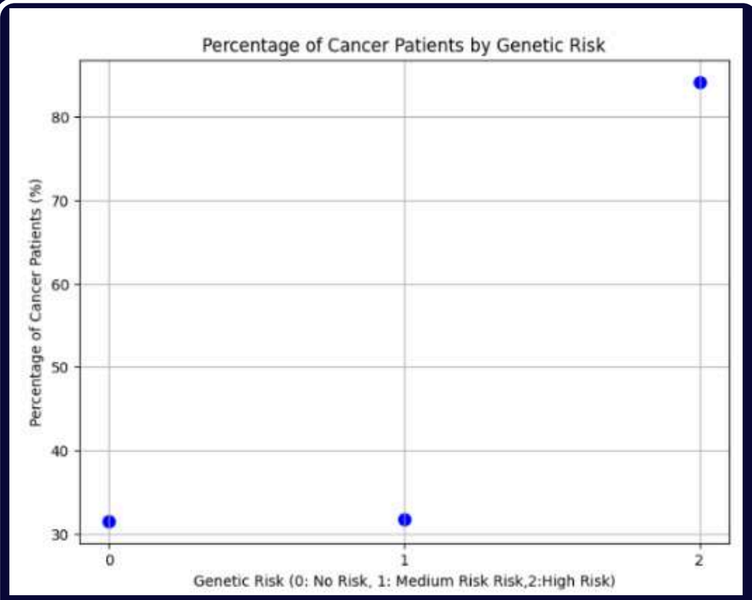


Diagram Batang Persentase pasien kanter berdasarkan kategori Gender

VISUALISASI DATA



Scatter plot pasien kanker berdasarkan risiko genetik

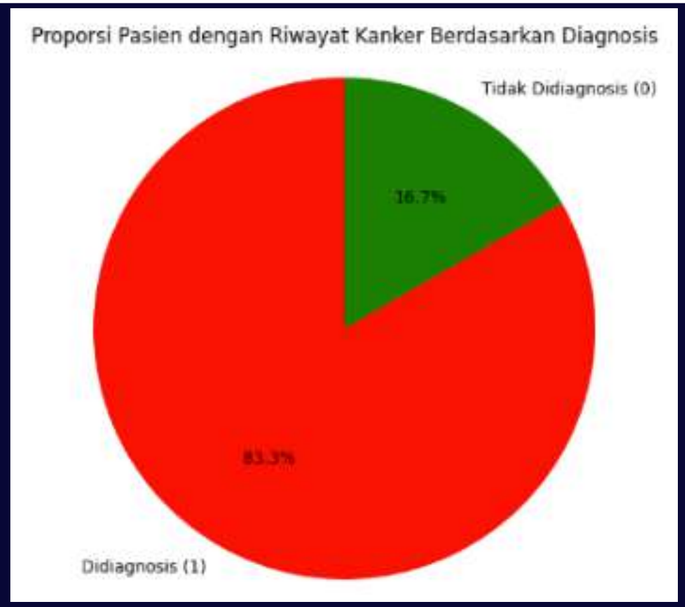
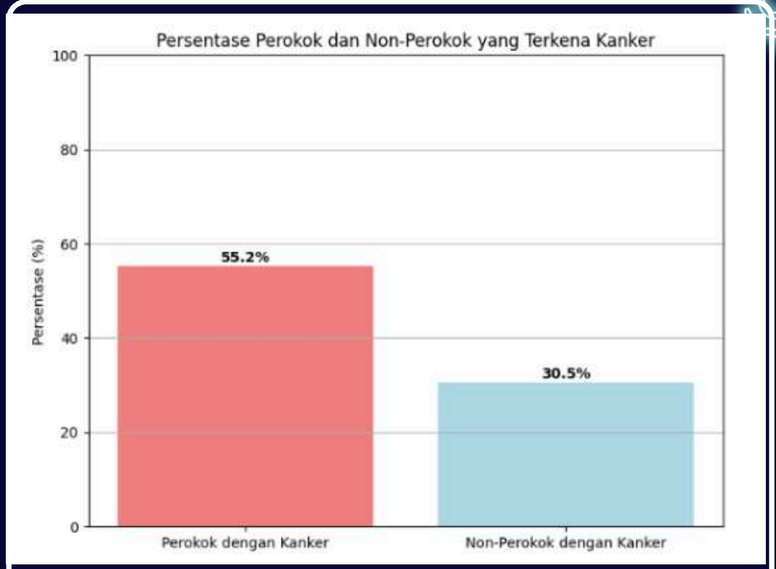


Diagram Lingkaran Pasien dengan Riwayat Kanker berdasarkan diagnosis



Persentase perokok dan Non-Perokok yang terkena kanker