



"ANALISIS FAKTOR RISIKO GAYA HIDUP, UMUR, DAN JENIS KANKER BERDASARKAN SKOR RISIKO DAN KORELASI"

Kelompok 18

Natan Obet Nego Hutahaean (11423054)
Leoni Nazwa Friskilla Sibuea (11423055)
Stefani Margareth Rajagukguk (11423066)

Latar Belakang

Kanker merupakan salah satu penyebab utama kematian global, dengan diperkirakan 19,3 juta kasus baru dan 9,9 juta kematian pada 2020 (Sung et al., 2020). Prevalensinya terus meningkat di seluruh dunia, terutama akibat faktor risiko gaya hidup seperti merokok, konsumsi alkohol, pola makan tidak sehat, dan kurangnya aktivitas fisik, serta faktor usia, di mana sebagian besar kanker ditemukan pada usia lanjut. Meskipun banyak penelitian yang meneliti faktor risiko ini, belum ada model yang menggabungkan gaya hidup dan usia dalam skor risiko kanker yang komprehensif. Penelitian ini bertujuan untuk menganalisis hubungan antara gaya hidup, usia, dan jenis kanker, serta mengembangkan model skor risiko berbasis data untuk mendukung kebijakan kesehatan dan program pencegahan kanker yang lebih efektif dan berbasis bukti.

Tujuan dan Manfaat

Tujuan:

Adapun tujuan yang ingin dicapai dalam penelitian ini adalah sebagai berikut:

1. Menganalisis pengaruh faktor gaya hidup terhadap risiko kanker.
2. Mengidentifikasi peran usia dalam peningkatan risiko kanker.
3. Menganalisis korelasi antara berbagai faktor risiko dengan jenis kanker.
4. Menyusun strategi pencegahan yang berbasis data.

Manfaat:

1. Memberikan pemahaman yang lebih mendalam mengenai faktor-faktor yang berisiko terhadap kanker.
2. Mengembangkan model skor risiko kanker berbasis data yang dapat digunakan untuk prediksi risiko kanker secara lebih akurat.
3. Mendukung kebijakan kesehatan masyarakat dengan menyediakan data yang kuat untuk merancang program pencegahan kanker yang lebih efektif.
4. Membantu individu dalam mengambil langkah-langkah pencegahan yang tepat untuk mengurangi risiko kanker.

Metode

1. Data Collection

Data Collection adalah langkah pertama dalam proses data science. Pada tahap ini, data dikumpulkan dan dicari dari berbagai sumber dataset, seperti API, file CSV, data publik, atau lainnya. Untuk proyek analisis ini, digunakan data yang berupa angka dan teks yang dibuat ke dalam file csv. Data ini diambil dari sumber dataset Kaggle yang bernama cancer.csv. Dataset ini berisi informasi terkait berbagai faktor yang mempengaruhi risiko kanker pada pasien. Setiap baris data mewakili satu pasien dan mencakup informasi tentang jenis kanker yang diderita, usia, jenis kelamin, kebiasaan merokok, konsumsi alkohol, tingkat obesitas, dan riwayat keluarga yang mengidap kanker.

Metode

2. Preprocessing Data

Proses ini terdiri dari:

- 1.Cek Jumlah Missing Values sebelum Imputasi, pada tahap ini akan diperiksa jumlah dari missing values di setiap kolom sebelum melakukan imputasi. Proses ini dilakukan dengan mengambil semua kolom numerik yaitu kolom yang berisi data yang dapat dihitung dalam dataset.
- 2.Handling Missing Values, Pada proses ini, digunakan scikit-learn yaitu sebuah library Python yang digunakan untuk machine learning dan data preprocessing. Library ini digunakan sebagai teknik untuk mengatasi missing values (nilai yang hilang) dalam dataset.
- 3.Handling Outliers using IQR method, Tujuan dari proses ini adalah untuk mengatasi outliers yang dapat memengaruhi hasil analisis atau model. Dengan mengurangi pengaruh nilai ekstrim, data menjadi lebih konsisten dan representatif. Selain itu, dengan mengurangi outliers, kualitas model pun meningkat, sehingga model yang dibangun menjadi lebih akurat, stabil, dan dapat diandalkan.

Metode

2. Preprocessing Data

Proses ini terdiri dari:

4. Feature Scaling using StandardScaler, Proses ini dilakukan dengan menggunakan StandardScaler dari scikit-learn untuk standarisasi kolom numerik dalam dataset. Standarisasi fitur numerik memastikan bahwa semua fitur berada dalam skala yang sama, yang sangat penting untuk algoritma machine learning yang sensitif terhadap perbedaan skala, seperti regresi linier, SVM, dan K-means clustering.

5. Analisis Komponen Utama (PCA)

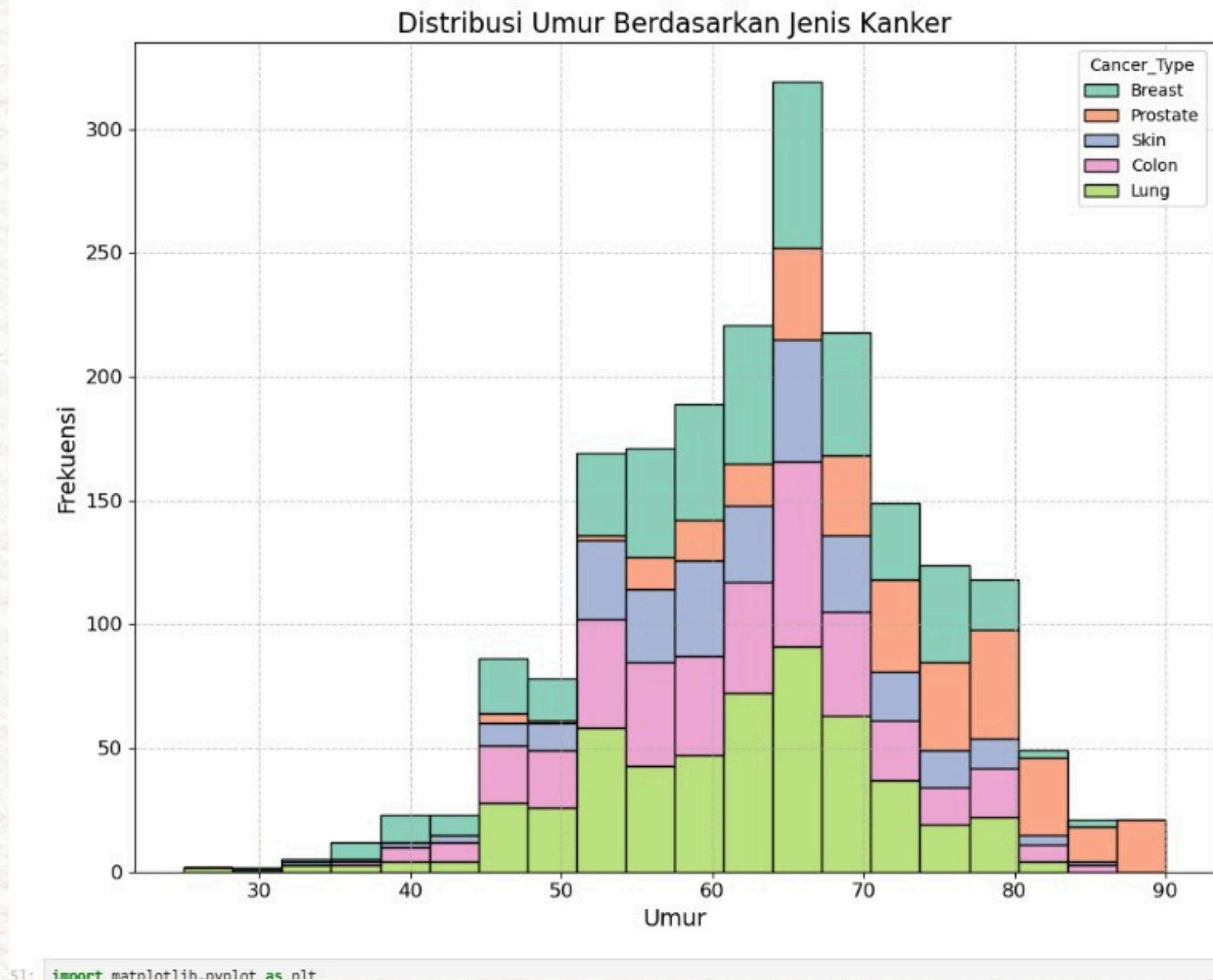
Principal Component Analysis (PCA) adalah teknik yang digunakan dalam tahap preprocessing untuk reduksi dimensi dalam dataset. Tujuan dari PCA adalah mengurangi jumlah fitur yang digunakan dalam model sambil tetap mempertahankan informasi penting yang ada dalam data.

Metode

3. Visualization (Visualisasi Data)

Visualisasi 1: Visualisasi Distribusi Umur Berdasarkan Jenis Kanker

Pada gambar ini, terdapat stacked histogram yang menunjukkan distribusi umur berdasarkan jenis kanker. Setiap batang mewakili frekuensi (jumlah) individu pada rentang umur tertentu yang didiagnosis dengan jenis kanker yang berbeda. Tujuan dari visualisasi ini adalah untuk menggambarkan bagaimana distribusi umur bervariasi antara jenis kanker yang berbeda. Ini memberikan pemahaman tentang pola usia pada pasien dengan masing-masing jenis kanker

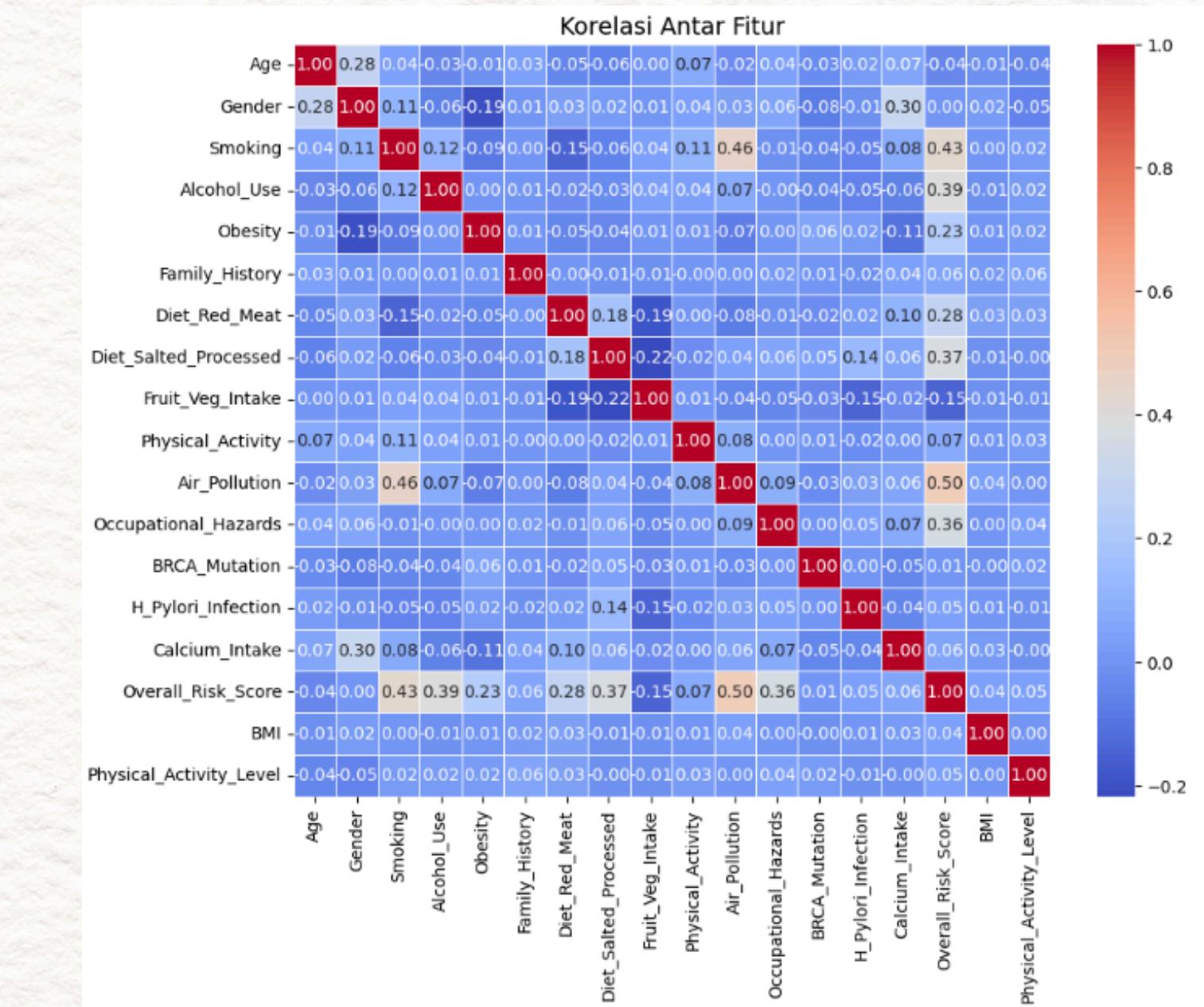


Metode

3. Visualization (Visualisasi Data)

Visualisasi 2: Visualisasi Korelasi Antar Fitur

Gambar ini menampilkan heatmap korelasi antar fitur (variabel) dalam dataset cancer.csv. Setiap kotak dalam heatmap menunjukkan nilai korelasi antara dua variabel numerik. Tujuan dari visualisasi ini adalah untuk menunjukkan hubungan antar variabel dalam dataset, agar kita dapat melihat korelasi yang signifikan antara fitur yang satu dengan yang lainnya.

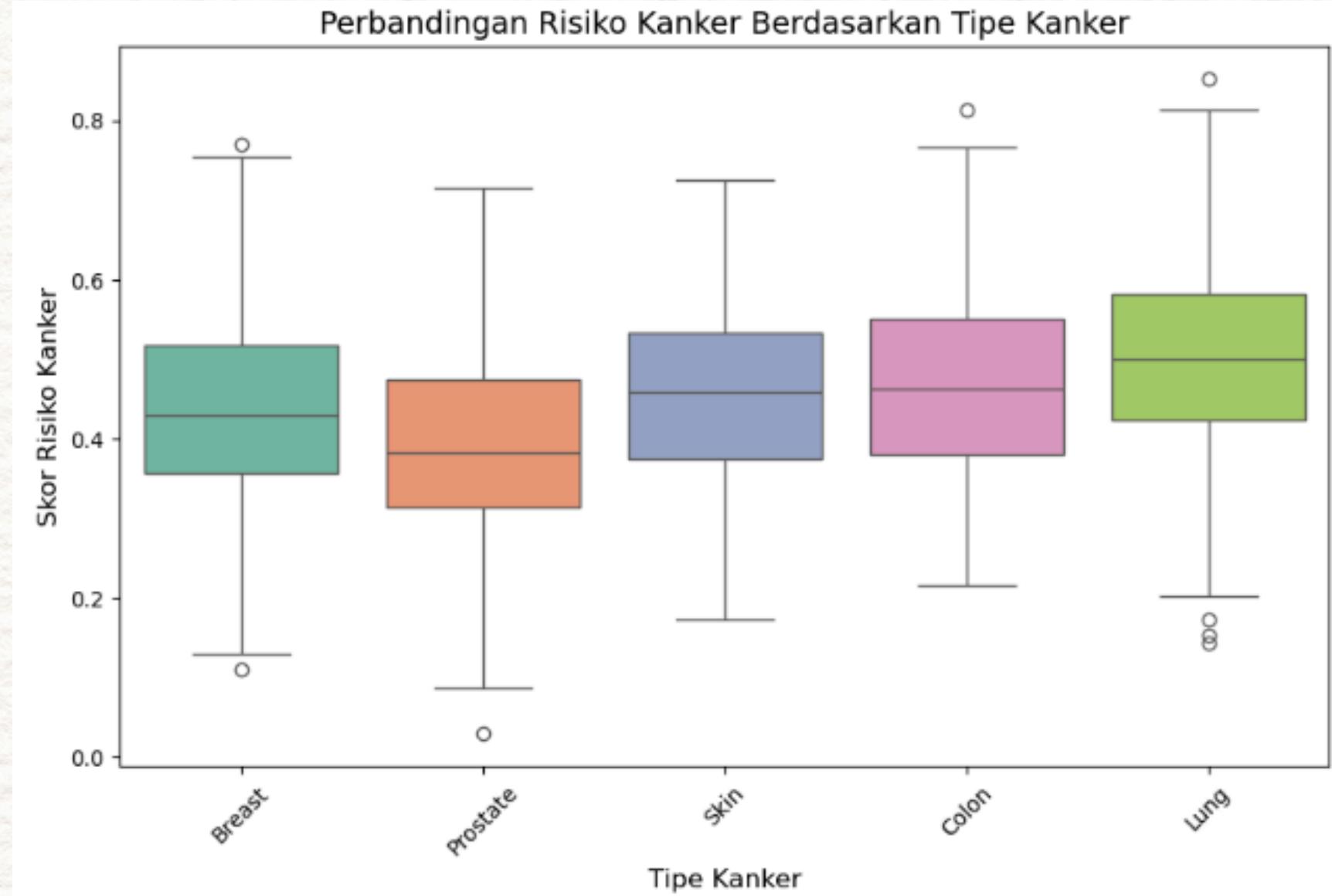


Metode

3. Visualization (Visualisasi Data)

Visualisasi 3: Visualisasi Perbandingan Risiko Kanker Berdasarkan Tipe Kanker

Gambar ini menampilkan boxplot yang menggambarkan perbandingan skor risiko kanker berdasarkan jenis kanker. Setiap boxplot mewakili distribusi skor risiko kanker untuk jenis kanker tertentu. Tujuan dari visualisasi ini adalah untuk membandingkan skor risiko kanker di antara berbagai jenis kanker. Boxplot ini menunjukkan Median, yaitu posisi garis di dalam kotak yang menunjukkan nilai tengah dari skor risiko. Interquartile Range (IQR), bagian kotak yang menggambarkan rentang skor antara kuartil pertama (25%) dan kuartil ketiga (75%).



Metode

4. Analysis (Analisis Data)

Dalam tahap analisis ini, kami menggunakan Exploratory Data Analysis (EDA) sederhana untuk memahami karakteristik dan hubungan antar variabel dalam dataset cancer.csv. Tujuan utama dari tahap ini adalah untuk memberikan pemahaman awal tentang data yang kami miliki. Tujuan lain dari tahap ini adalah memahami karakteristik data, memvisualisasikan data, dan menganalisis korelasi antar variabel.

Statistical Analysis

Uji Parametrik

Tujuan: Untuk mengukur kekuatan dan arah hubungan linier antara dua variabel numerik, yaitu BMI dan Overall_Risk_Score, dan menentukan apakah hubungan tersebut signifikan secara statistik.

Uji Non-Parametrik

Tujuan: Untuk menguji apakah terdapat perbedaan signifikan dalam distribusi skor risiko kanker antara lebih dari dua tipe kanker yang berbeda, tanpa mengasumsikan distribusi normal pada data.

```
Uji Parametrik: Pearson Correlation
Pearson Correlation: 0.037
p-value: 0.097
95% Confidence Interval: (-0.007, 0.081)
```

```
Uji Non-Parametrik: Kruskal-Wallis Test
Kruskal-Wallis H-statistic: 167.212
p-value: 0.000
```

TERIMA KASIH

