**PENERAPAN *EXPLORATORY DATA ANALYSIS (EDA)***

**UNTUK DETEKSI DINI KANKER PARU-PARU STUDI KASUS PADA DATASET MEDIS**

Disusun Untuk Memenuhi Tugas Mata Kuliah Kecerdasan Buatan

Dosen Pengampu : Kodrat Mahatma S.Kom., M.Kom.



**Disusun oleh:**

Bayu Fajri Noorjamil (20423004)

Wida Waryani (20423012)

Nita Andiyani (20423013)

Namia Resca Kirana (20423010)

Risna De Sinta (20125008)

**PROGRAM STUDI MANAJEMEN INFORMATIKA**

**UNIVERSITAS TEKNOLOGI DIGITAL**

**2024**

Daftar Isi

[BAB I PENDAHULUAN 2](#_Toc195792077)

[1.1 Latar Belakang 2](#_Toc195792078)

[BAB II PEMBAHASAN 3](#_Toc195792079)

[2.1 Kasus (Penerapan EDA untuk Deteksi Dini Kanker Paru-Paru: Studi Kasus pada Dataset Medis 3](#_Toc195792080)

[2.2 Sumber Data 3](#_Toc195792081)

[2.3 Analisis 3](#_Toc195792082)

[Langkah-langkah analisis yang dilakukan meliputi : 3](#_Toc195792083)

[BAB III HASIL 6](#_Toc195792084)

[3.1 Hasil Analisis EDA 6](#_Toc195792085)

[3.2 Visualisasi Data: Distribusi Umur Pasien 8](#_Toc195792086)

[3.3 Visualisasi Data: Heatmap Korelasi Antar Fitur 9](#_Toc195792087)

[3.4 Link Github EDA Untuk Deteksi Dini Kanker Paru-Paru Studi Kasus Pada Dataset Medis 10](#_Toc195792088)

[BAB IV PENUTUP 11](#_Toc195792089)

[4.1 Kesimpulan 11](#_Toc195792090)

[4.2 Saran 11](#_Toc195792091)

[DAFTAR PUSTAKA 12](#_Toc195792092)

# **BAB I PENDAHULUAN**

## **Latar Belakang**

Kanker paru-paru merupakan salah satu penyebab utama kematian akibat kanker di dunia. Menurut data Global Cancer Observatory dari Organisasi Kesehatan Dunia, pada tahun 2020, terdapat sekitar 1.796.144 kematian akibat kanker paru-paru secara global, dengan 30.843 kasus di Indonesia.

Dalam upaya meningkatkan deteksi dini, pendekatan berbasis data mining dan machine learning telah banyak digunakan untuk memprediksi risiko kanker paru-paru. Algoritma seperti Random Forest, *Support Vector Machine* (SVM), dan *K-Nearest Neighbors* (KNN) telah diterapkan untuk menganalisis data klinis dan gejala pasien guna mengidentifikasi individu dengan risiko tinggi. Misalnya, penelitian oleh Permana et al. (2023) menggunakan algoritma Random Forest untuk menganalisis data klinis dan gejala pasien dalam memprediksi kanker paru-paru.

Penggunaan dataset terbuka, seperti yang tersedia di Kaggle, telah memfasilitasi pengembangan model prediksi yang lebih akurat dan dapat diakses oleh berbagai pihak. Misalnya, dataset "Cancer Patients and Air Pollution" mencakup informasi tentang pasien kanker paru-paru, termasuk usia, jenis kelamin, dan paparan polusi udara. Model prediksi yang dibangun dengan menggunakan data ini dapat membantu dalam identifikasi dini dan pengambilan keputusan klinis.

Namun, tantangan masih ada, terutama dalam hal ketersediaan data yang berkualitas tinggi dan representatif, serta kebutuhan akan validasi klinis yang ketat untuk memastikan keandalan model prediksi. Oleh karena itu, kolaborasi antara peneliti, profesional medis, dan pengembang teknologi sangat penting untuk mengoptimalkan penggunaan data mining dan machine learning dalam deteksi dini kanker paru-paru.

# **BAB II PEMBAHASAN**

## **Kasus (Penerapan EDA untuk Deteksi Dini Kanker Paru-Paru: Studi Kasus pada Dataset Medis**

Kasus ini membahas bagaimana teknik *Exploratory Data Analysis* (EDA) digunakan untuk memahami pola dan karakteristik data medis pasien yang berisiko terkena kanker paru-paru. Dataset yang digunakan berisi informasi seperti usia, jenis kelamin, kebiasaan merokok, riwayat keluarga, serta gejala klinis seperti batuk kronis, nyeri dada, dan sesak napas.

Melalui EDA, dilakukan visualisasi dan analisis statistik awal untuk mengidentifikasi fitur-fitur penting yang berkorelasi dengan kemunculan kanker paru-paru. Hasil analisis ini menjadi dasar untuk pengembangan model prediksi dan pengambilan keputusan medis yang lebih tepat sasaran dalam upaya deteksi dini.

## **Sumber Data**

<https://www.kaggle.com/datasets/shantanugarg274/lung-cancer-prediction-dataset>

## **Analisis**

Analisis Data Eksploratif (*Exploratory Data Analysis*/EDA) merupakan langkah awal dalam proses analisis data yang bertujuan untuk memahami struktur, pola serta kemungkinan adanya kejanggalan pada data sebelum masuk ke tahap pemodelan (Shafa et al., 2024). Dalam bidang medis, khususnya dalam deteksi kanker paru paru, EDA memegang peran penting untuk :

* Menganalisis keterkaitan antara gejala klinis dengan hasil diagnosis
* Menemukan fitur-fitur yang paling berkontribusi terhadap potensi risiko kanker
* Mempersiapkan data agar dapat diolah secara optimal oleh model *machine learning.*

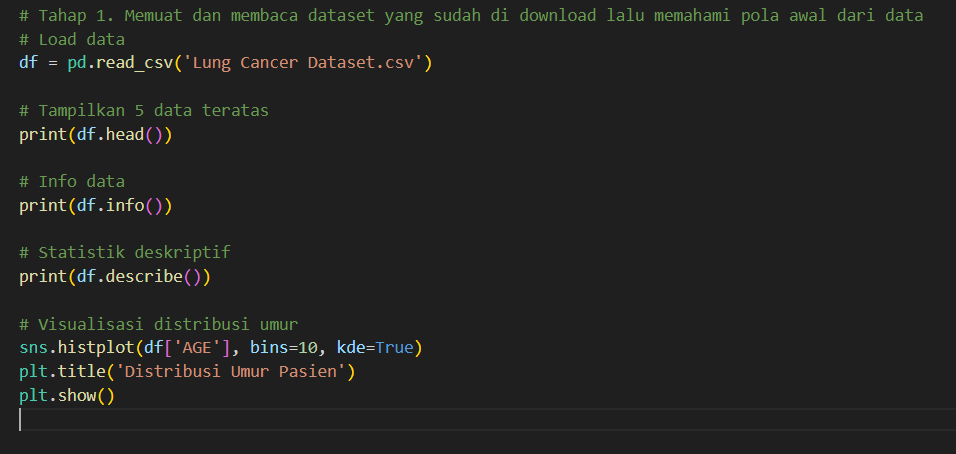
Analisis yang dilakukan dengan pendekatan EDA untuk memahami distribusi, pola dan hubungan antar fitur dalam dataset. Dataset yang digunakan adalah *“Lung Cancer Prediction Dataset”* dataset ini memuat sejumlah atribut penting seperti usia, jenis kelamin dan juga informasi mengenai riwayat terhadap penyakit kanker.

### Langkah-langkah analisis yang dilakukan meliputi :

#### Visualisasi data awal

Visualisasi data awal adalah proses menampilkan data dalam bentuk grafik atau diagram dengan tujuan untuk memahami pola, distribusi dan karakteristik umum dari data sebelum dilakukan analisis lebih lanjut. Tahap ini sangat penting karena membantu dalam mengenali informasi tersembunyi di balik angka dan mendeteksi potensi masalah pada data seperti distribusi yang tidak merata.

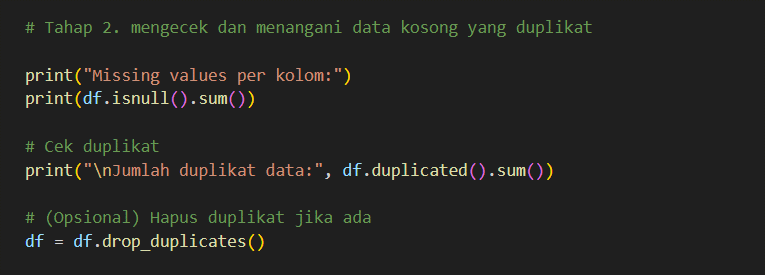
Code visualisasi data



#### Pembersihan data

Pembersihan data (*data cleaning*) adalah proses sistematis untuk mendeteksi, memperbaiki atau menghapus data yang tidak valid, tidak konsisten atau duplikat, guna meningkatkan kualitas data dan memastikan hasil analisis yang akurat. Pada tahap pembersihan data ini sangat penting terutama dalam penerapan EDA untuk deteksi dini kanker paru-paru, karena bisa menjamin akurasi hasil, mencegah adanya error pada model dan meningkatkan konsisten data.

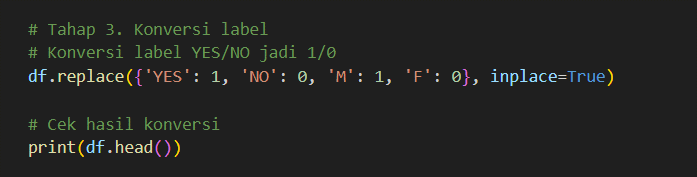
Code pembersihan data



#### Konversi tabel

Konversi tabel disini proses mengubah data kategorikal menjadi data numerik agar bisa diproses oleh algritma statistic atau machine learning. Konversi label proses krusial dalam EDA yang menjembatani antara data mentah dan algoritma analisis. Tanpa konversi tabel, data medis seperti pada kasus kanker paru-paru tidak bisa digunakan untuk prediksi maupun visualisasi korelasi.

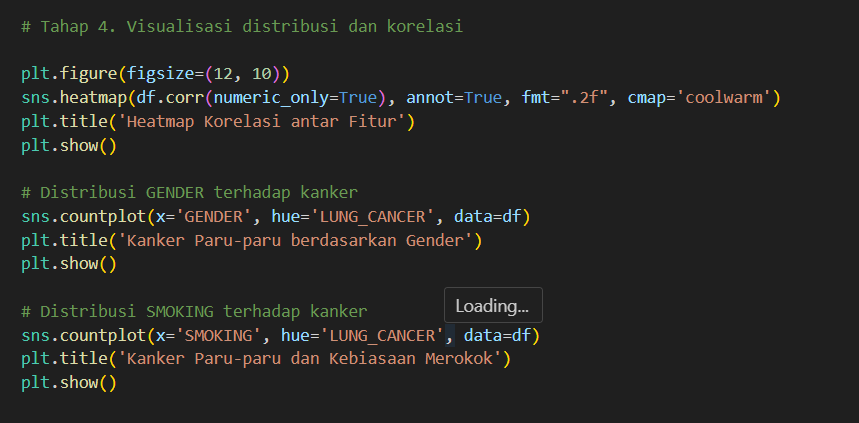
Code konversi tabel



#### Visualisasi data lanjutan

Visualisasi lanjutan disini dimana proses menyajikan data dengan grafik komplek dan interaktif untuk memahami pola tersembunyi, menjelaskan hubungan antar variable dan membantu pengambilan keputusan berbasis data. Visualisasi lanjutan bukan hanya soal “grafik” tapi alat diagnosis awal dalam data science medis, dan visualisasi membantu mendeteksi hubungan klinis penting, memilih fitur terbaik untuk model dan menyampaikan temuan ke tim medis secara jelask dan efektikf.

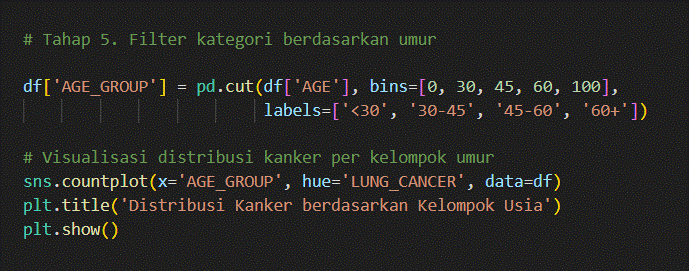
Code visualisasi data lanjutan



#### Metrik turunan

Metrik turunan adalah hasil analisis awal berupa angka atau statistic yang diperoleh dari data mentah dan digunakan untuk memahami karakteristik pasien, mengidentifikasi fitur penting, menilai kualitas data dan membentuk dasar untuk pemodelan selanjutnya. Tahap ini sangat penting dalam memahami data sebelum pemodelan, menentukan fitur yang relevan dan mendeteksi masalah seperti data kosong atau tidak seimbang.

Code metrik turunan



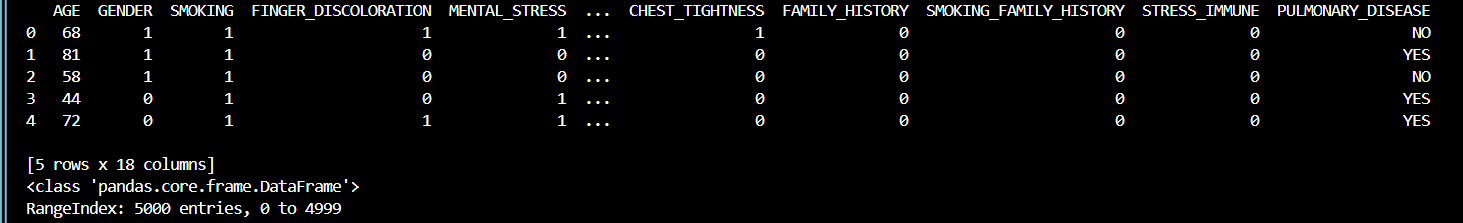
# **BAB III HASIL**

## **Hasil Analisis EDA**

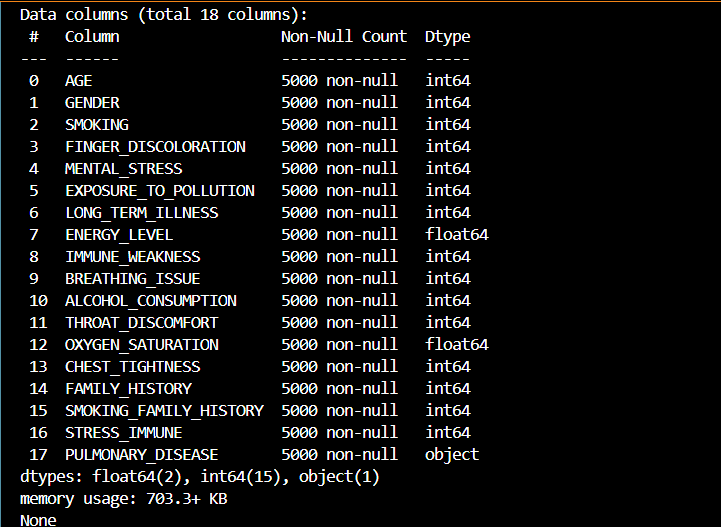
Pada tahap ini, dilakukan analisis eksploratif (Exploratory Data Analysis/EDA) terhadap dataset yang digunakan untuk mendeteksi risiko kanker paru-paru. Tujuan dari analisis ini adalah untuk memahami karakteristik data, mengidentifikasi pola awal, serta mendeteksi potensi hubungan antar variabel yang dapat membantu dalam proses pemodelan prediktif. Analisis dilakukan terhadap keseluruhan entri dan variabel yang tersedia untuk mendapatkan gambaran umum mengenai kondisi kesehatan pasien yang menjadi subjek penelitian.Berikut adalah hasil analisis yang diperoleh dari dataset:

1. Deskripsi Umum Dataset

Dataset yang dianalisis terdiri dari 5.000 entri dan 18 kolom, yang mencakup berbagai atribut penting terkait kondisi kesehatan pasien, seperti usia, jenis kelamin, kebiasaan merokok, dan sejumlah gejala klinis.

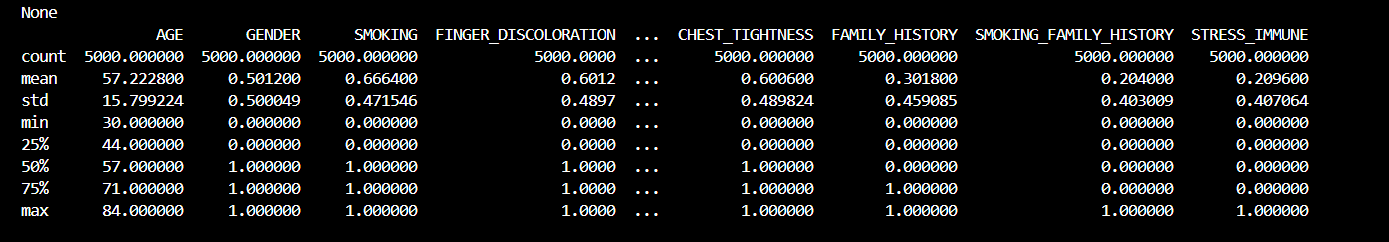
Tipe data yang terdapat dalam dataset ini didominasi oleh tipe numerik, yaitu 15 kolom bertipe *integer* dan 2 kolom bertipe *float*, serta satu kolom bertipe *object* yang merepresentasikan status penyakit paru-paru.

Struktur dataset yang bersih dan konsisten memungkinkan proses eksplorasi data dilakukan secara efisien tanpa memerlukan banyak proses pembersihan data.

Setiap atribut telah diberi label yang jelas, sehingga dapat langsung digunakan dalam proses analisis.

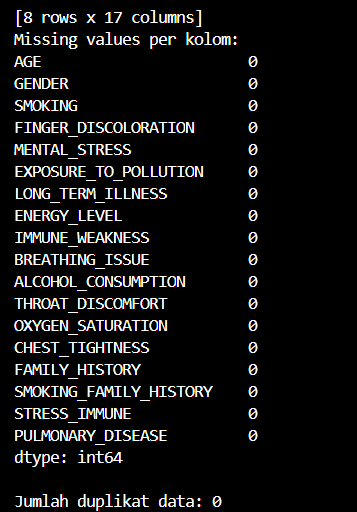
1. Statistik Deskriptif

Berdasarkan statistik deskriptif, rata-rata usia pasien adalah 57,22 tahun dengan rentang usia dari 30 hingga 84 tahun. Hal ini menunjukkan bahwa populasi yang dianalisis cukup beragam dan mencakup kelompok usia dewasa hingga lanjut usia, yang memang merupakan kelompok rentan terhadap penyakit kanker paru-paru. Sekitar 66,64% pasien tercatat sebagai perokok. Persentase ini menunjukkan tingginya prevalensi kebiasaan merokok di antara pasien, sekaligus mengindikasikan adanya hubungan yang signifikan antara paparan asap rokok dan meningkatnya risiko terkena kanker paru-paru. Data ini juga dapat menjadi landasan penting dalam penentuan variabel prediktif utama pada tahap pemodelan.

Selain itu, sekitar 30,18% pasien memiliki riwayat keluarga dengan penyakit kanker, sebagaimana ditunjukkan oleh nilai mean pada variabel family\_history. Variabel kombinasi antara kebiasaan merokok dan riwayat keluarga. (smoking\_family\_history) menunjukkan nilai rata-rata sebesar 0,204, mengindikasikan bahwa 20,4% pasien memiliki dua faktor risiko tersebut secara bersamaan. Di sisi lain, stres yang memengaruhi sistem imun (stress\_immune) tercatat pada sekitar 21% pasien, yang menegaskan bahwa faktor psikologis juga berperan dalam menurunkan ketahanan tubuh terhadap penyakit, termasuk kanker paru-paru. Seluruh hasil ini memberikan gambaran awal yang kuat terhadap karakteristik data dan akan menjadi landasan dalam tahap pemodelan prediktif selanjutnya.

1. Kualitas Data

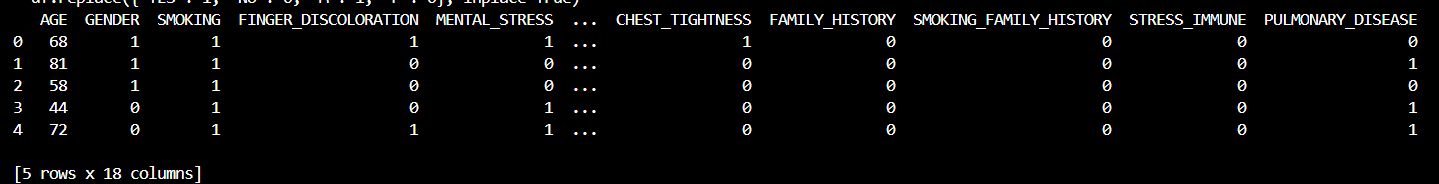
Berdasarkan hasil pemeriksaan terhadap kualitas data, diketahui bahwa tidak terdapat nilai yang hilang pada seluruh kolom dalam dataset. Setiap atribut memiliki jumlah entri yang lengkap, yaitu sebanyak 5.000 data, tanpa adanya kekosongan nilai (*missing values*). Kondisi ini merupakan indikator penting yang menunjukkan bahwa dataset memiliki struktur yang baik dan siap digunakan dalam proses analisis lebih lanjut.



Ketersediaan data yang lengkap ini memberikan keuntungan besar dalam proses pemodelan dan eksplorasi data, karena tidak diperlukan langkah tambahan untuk melakukan imputasi, pengisian nilai, atau penghapusan baris data yang tidak lengkap. Dengan demikian, seluruh entri dalam dataset dapat dimanfaatkan secara optimal tanpa kehilangan informasi apa pun.

Secara keseluruhan, tidak adanya nilai yang hilang mencerminkan kualitas dataset yang tinggi. Hal ini berdampak langsung pada keandalan hasil analisis dan akurasi model prediktif yang akan dibangun. Dataset yang bersih dan lengkap memungkinkan proses analisis berjalan lebih efisien dan hasil yang diperoleh pun memiliki tingkat validitas yang lebih tinggi.

1. Entri Dataset

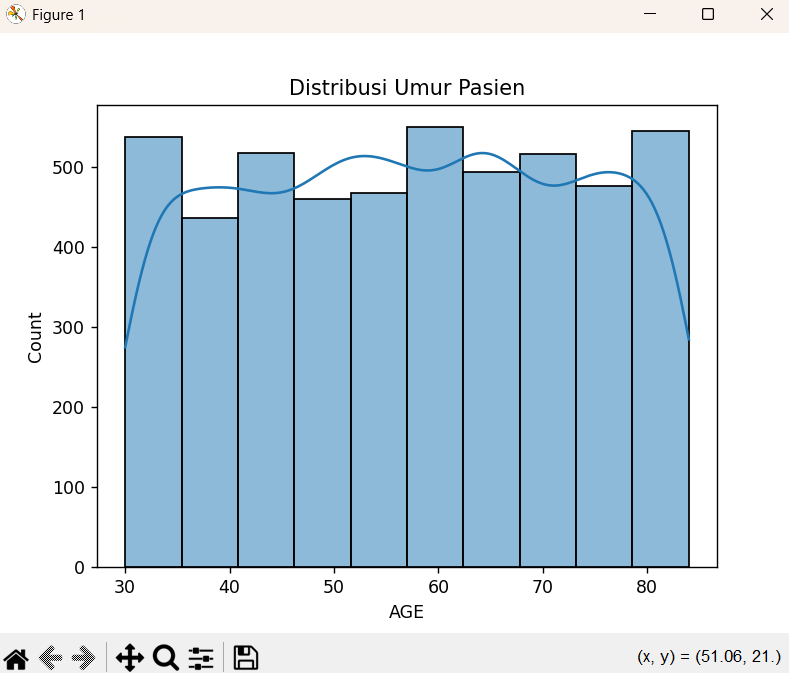
Gambar di atas menampilkan lima entri contoh dari dataset yang digunakan dalam analisis risiko kanker paru-paru. Setiap entri memuat informasi lengkap mengenai pasien, mencakup variabel seperti usia (age), jenis kelamin (gender), kebiasaan merokok (smoking), perubahan warna pada jari (finger\_discoloration), stres mental (mental\_stress), sesak dada (chest\_tightness), riwayat keluarga (family\_history), serta kondisi lain yang berhubungan dengan paru-paru. penyajian data ini memberikan gambaran awal yang jelas mengenai struktur dataset dan jenis informasi yang tersedia.

sebagian besar variabel dalam dataset direpresentasikan dalam bentuk data kategorikal biner, dengan nilai 1 menunjukkan keberadaan suatu kondisi atau karakteristik, dan nilai 0 menandakan ketidakhadiran. misalnya, nilai 1 pada kolom smoking menandakan bahwa pasien tersebut merupakan perokok, sementara nilai 0 berarti sebaliknya.

Format representasi ini memudahkan proses analisis karena memungkinkan penggunaan teknik pemodelan prediktif secara langsung tanpa perlu konversi data tambahan. Contoh entri ini juga menunjukkan adanya variasi data yang cukup luas, baik dari segi usia, jenis kelamin, maupun kondisi kesehatan. Hal ini mencerminkan keragaman profil pasien yang menjadi objek analisis, yang penting dalam membangun model prediksi yang bersifat general dan tidak bias terhadap kelompok tertentu. Dengan struktur data yang rapi dan informatif, contoh ini menjadi pijakan penting untuk memahami konteks dataset sebelum melangkah ke tahap analisis dan pemodelan selanjutnya.

## **Visualisasi Data: Distribusi Umur Pasien**

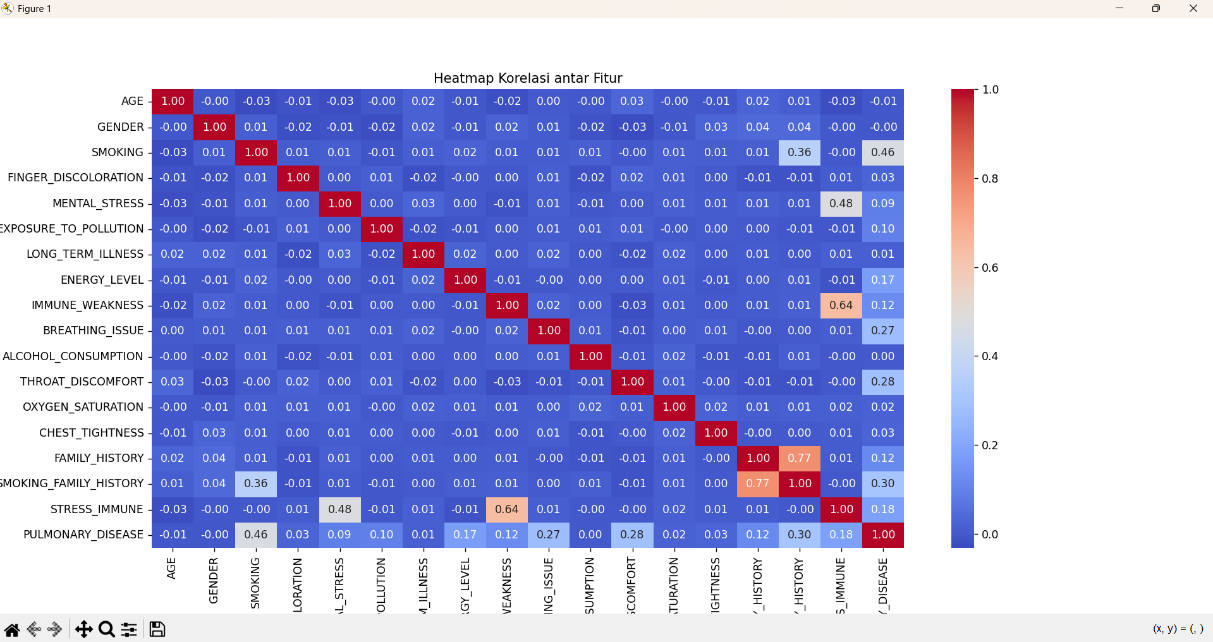
Analisis distribusi umur pasien dalam dataset ini memberikan wawasan yang signifikan mengenai profil populasi yang berisiko terhadap kanker paru-paru. Grafik yang ditampilkan menunjukkan beberapa hal penting yang perlu diperhatikan. Rentang umur pasien dalam dataset ini berkisar antara 30 hingga 84 tahun, dengan pengelompokan interval yang memudahkan pemahaman distribusi data.



Secara umum, distribusi jumlah pasien pada setiap interval umur cukup merata, meskipun terdapat penurunan sedikit pada kelompok usia yang lebih tua. Hal ini menunjukkan bahwa dataset mencakup pasien dari berbagai rentang usia, dengan kecenderungan lebih banyak pasien berada pada kelompok umur antara 50 hingga 70 tahun. Garis biru yang melintasi histogram menggambarkan tren frekuensi pasien berdasarkan umur, yang meskipun mengalami fluktuasi, menunjukkan bahwa sebagian besar pasien berada pada kelompok umur menengah hingga lanjut, yang cenderung memiliki risiko lebih tinggi terhadap kanker paru-paru. Pemahaman mengenai distribusi umur pasien ini sangat penting bagi tenaga medis dalam melakukan diagnosis serta pengambilan keputusan terkait deteksi dini kanker paru-paru, di mana kelompok usia yang lebih tua memerlukan perhatian lebih dalam hal skrining dan evaluasi kesehatan. Dengan analisis ini, tenaga medis dan peneliti dapat lebih memahami karakteristik populasi pasien yang dapat mendukung pengembangan strategi deteksi dini yang lebih efektif.

## **Visualisasi Data: Heatmap Korelasi Antar Fitur**

Pada bagian ini, disajikan heatmap korelasi yang menggambarkan hubungan antara berbagai fitur dalam dataset yang digunakan untuk analisis risiko kanker paru-paru. Heatmap ini memberikan wawasan visual yang memudahkan pemahaman mengenai keterkaitan antar fitur. Gambar heatmap ini akan membantu dalam mengidentifikasi pola-pola korelasi yang ada, yang nantinya dapat digunakan dalam pemodelan risiko kanker paru-paru.



Skala korelasi pada heatmap ini berkisar antara -1 hingga 1, di mana nilai mendekati 1 menunjukkan hubungan positif yang kuat, nilai mendekati -1 menunjukkan hubungan negatif yang kuat, dan nilai 0 menunjukkan tidak ada korelasi. Beberapa fitur dalam dataset menunjukkan korelasi yang signifikan, seperti smoking, yang memiliki korelasi positif kuat dengan fitur finger\_discoloration dan chest\_tightness, menandakan bahwa pasien yang merokok cenderung mengalami gejala-gejala tersebut. Selain itu, family\_history juga menunjukkan korelasi dengan beberapa fitur lainnya, mengindikasikan adanya hubungan antara riwayat keluarga dan faktor risiko lainnya. sebaliknya, beberapa fitur seperti mental\_stress dan alcohol\_consumption menunjukkan korelasi yang sangat rendah, yang bisa menandakan bahwa faktor-faktor tersebut tidak berkontribusi signifikan terhadap risiko kanker paru-paru dalam dataset ini. Pemahaman tentang korelasi antar fitur ini sangat penting dalam proses pemodelan. Fitur dengan korelasi tinggi dapat saling mendukung, sementara fitur dengan korelasi rendah mungkin perlu dipertimbangkan kembali dalam pemilihan fitur untuk model prediksi. Visualisasi melalui heatmap ini membantu peneliti memilih fitur-fitur yang relevan untuk analisis lebih lanjut dan pengembangan model prediksi yang lebih efektif.

## **Link Github EDA Untuk Deteksi Dini Kanker Paru-Paru Studi Kasus Pada Dataset Medis**

https://github.com/enNamgil/kecerdasanbuatan

# **BAB IV PENUTUP**

## **Kesimpulan**

Berdasarkan hasil penelitian, dapat disimpulkan bahwa penerapan Exploratory Data Analysis (EDA) sangat efektif dalam membantu proses deteksi dini kanker paru-paru. Melalui EDA, peneliti dapat mengenali pola, hubungan antar gejala, serta faktor-faktor risiko yang paling berpengaruh terhadap penyakit ini. Dataset yang digunakan juga telah terbukti berkualitas baik dan tidak memiliki data kosong, sehingga memungkinkan analisis dilakukan secara menyeluruh dan akurat. Hasil analisis menunjukkan bahwa faktor seperti merokok, riwayat keluarga, dan stres merupakan penyumbang risiko terbesar terhadap kanker paru-paru. Selain itu, visualisasi seperti distribusi usia dan korelasi antar variabel menunjukkan bahwa usia lanjut dan kebiasaan merokok memiliki hubungan yang kuat dengan kemungkinan terjadinya kanker. EDA juga memberikan wawasan awal yang sangat penting sebelum dilakukan pemodelan machine learning, terutama dalam proses pemilihan fitur yang relevan dan penyampaian hasil kepada tim medis. Dengan demikian, EDA berperan penting sebagai dasar yang kuat dalam pengembangan sistem pendukung keputusan untuk deteksi dini kanker paru-paru.

## **Saran**

Penelitian ini sudah sangat bagus dalam menunjukkan bagaimana analisis data (EDA) bisa membantu mengenali tanda-tanda awal kanker paru-paru. Namun, ke depannya ada beberapa hal yang bisa ditingkatkan. Pertama, akan lebih baik jika penulis menambahkan cara lain untuk memastikan hasil, misalnya menggunakan metode statistik sederhana untuk menguji hubungan antar gejala. Kedua, hasil dari penelitian ini bisa dibandingkan dengan penelitian lain yang mirip, agar kita tahu apakah hasilnya sejalan atau berbeda. Ketiga, penelitian ini juga bisa dilanjutkan dengan membuat model prediksi menggunakan kecerdasan buatan (machine learning), agar bisa digunakan untuk membantu diagnosa secara otomatis. Keempat, gambar-gambar atau grafik yang digunakan sebaiknya diberi penjelasan singkat agar mudah dipahami oleh pembaca umum. Selain itu, penulis juga bisa menjelaskan bagaimana hasil ini bisa digunakan oleh dokter atau rumah sakit dalam pengambilan keputusan. Terakhir, jika memungkinkan, sistem analisis ini bisa dibuat dalam bentuk aplikasi atau website sederhana agar bisa digunakan oleh banyak orang.

# **DAFTAR PUSTAKA**

Shafa, B., Handayani, H. H., Lestari, S. A. P., & Cahyana, Y. (2024). Prediksi Kanker Paru dengan Normalisasi menggunakan Perbandingan Algoritma Random Forest, Decision Tree dan Naïve Bayes. *Decode: Jurnal Pendidikan Teknologi Informasi*, *4*(3), 1057-1070.

Idris, J. F., Ramadhani, R., & Mutoffar, M. M. (2024). Klasifikasi Penyakit Kanker Paru Menggunakan Perbandingan Algoritma Machine Learning. *Jurnal Media Akademik (JMA)*, *2*(2).

Wahid, M. A. R., Nugroho, A., & Anshor, A. H. (2023). Prediksi Penyakit Kanker Paru-Paru Dengan Algoritma Regresi Linier. *Bulletin of Information Technology (BIT)*, *4*(1), 63-74.