

**CEP-CCIT**

**FAKULTAS TEKNIK**

**UNIVERSITAS INDONESIA**

**LAPORAN PROJEK AKHIR MATA KULIAH**

**DAI DAN IML**

**YANG BERJUDUL**

**ANALISIS DATA PASIEN KANKER PARU-PARU**

**MENGGUNAKAN RSTUDIO**

**Oleh:**

**Keiko Joceliandita (2213020201)**

**Delisa Febriana (2213020195)**

**PROGRAM INFORMATION TECHNOLOGY FOR ECONOMIC**

**CEP-CCIT FAKULTAS TEKNIK**

**UNIVERSITAS INDONESIA**

**DEPOK**

**2023**

2023

# KATA PENGANTAR

Puji syukur kehadirat Tuhan Yang Maha Esa. Atas Rahmat dan Hidayah-Nya, kami dapat menyelesaikan laporan makalah yang berjudul “Analisis Data Pasien Kanker Paru-Paru menggunakan RStudio” dengan tepat waktu.

Makalah disusun untuk memenuhi tugas akhir semester mata kuliah Data Analytic Introduction dan Introduction to Machine Learning. Selain itu, makalah ini bertujuan meninjau dan menambah wawasan tentang keamanan pangan bagi para pembaca dan juga bagi Kami penulis.

Kami mengucapkan terima kasih kepada Bapak Riza Iqbal selaku Dosen Mata Kuliah DAI dan IML, yang telah memberikan pengajaran dan juga bimbingan. Ucapan terima kasih juga disampaikan kepada semua pihak yang telah membantu diselesaikannya makalah ini.

Kami menyadari makalah ini jauh dari sempurna. Oleh sebab itu, saran dan kritik yang membangun diharapkan demi kesempurnaan makalah ini.

Depok, 5 July 2023

Penulis

# DAFTAR ISI

[KATA PENGANTAR 1](#_Toc139925864)

[DAFTAR ISI 2](#_Toc139925865)

[BAB I 4](#_Toc139925866)

[PENDAHULUAN 4](#_Toc139925867)

[1.1 Latar Belakang 4](#_Toc139925868)

[1.2 Rumusan Masalah 5](#_Toc139925869)

[1.3 Tujuan 5](#_Toc139925870)

[1.4 Manfaat 5](#_Toc139925871)

[1.5 Sistematika Pembahasan 6](#_Toc139925872)

[BAB I PENDAHULUAN 6](#_Toc139925873)

[BAB II LANDASAN TEORI 6](#_Toc139925874)

[BAB III IMPLEMENTASI 6](#_Toc139925875)

[BAB IV HASIL PERCOBAAN 6](#_Toc139925876)

[BAB V KESIMPULAN 6](#_Toc139925877)

[BAB II 7](#_Toc139925878)

[LANDASAN TEORI 7](#_Toc139925879)

[2.1 Konsep Dasar Bahasa Pemrograman R dan RStudio 7](#_Toc139925880)

[2.1.1 Bahasa Pemrograman R 7](#_Toc139925881)

[2.1.2 RStudio 8](#_Toc139925882)

[2.2 Jenis Analisis Data 9](#_Toc139925883)

[2.2.1 Analisis Deskriptif 9](#_Toc139925884)

[2.2.2 Analisis Korelasi 9](#_Toc139925885)

[2.2.3 Analisis Regresi 9](#_Toc139925886)

[2.2.4 Analisis Prediktif 9](#_Toc139925887)

[2.3 Metode Yang Digunakan 10](#_Toc139925888)

[2.3.1 Normalisasi 10](#_Toc139925889)

[2.3.2 Korelasi 10](#_Toc139925890)

[2.3.3 Regresi Linier 10](#_Toc139925891)

[2.3.4 Pohon Keputusan (Decision Tree) 10](#_Toc139925892)

[2.4 Konsep Dasar Kanker Paru-Paru 11](#_Toc139925893)

[2.5 Penjelasan Terkait Data Yang Digunakan 12](#_Toc139925894)

[BAB III 14](#_Toc139925895)

[IMPLEMENTASI 14](#_Toc139925896)

[3.1 Pemrosesan Data 14](#_Toc139925897)

[3.2 Visualisasi Data Histogram 14](#_Toc139925898)

[3.3 Normalisasi Data 14](#_Toc139925899)

[3.4 Menggunakan Correlation 15](#_Toc139925900)

[3.5 Menggunakan Regression 15](#_Toc139925901)

[3.6 Menggunakan Decision Tree 16](#_Toc139925902)

[3.7 Analisis Prediksi Menggunakan Linear Regression 17](#_Toc139925903)

[3.8 Menghitung Tingkat Akurasi Model dengan Linear Regression 17](#_Toc139925904)

[3.9 Evaluasi Performa Model dengan Linear Regression 17](#_Toc139925905)

[3.10 Analisis Prediksi Menggunakan Decision Tree 18](#_Toc139925906)

[3.11 Menghitung Tingkat Akurasi Model dengan Decision Tree 18](#_Toc139925907)

[BAB IV 19](#_Toc139925908)

[HASIL ANALISIS 19](#_Toc139925909)

[4.1 Hasil Correlation 19](#_Toc139925912)

[4.2 Hasil Regression 20](#_Toc139925913)

[4.3 Hasil Decision Tree 24](#_Toc139925914)

[4.4 Hasil Prediksi Linear Regression 30](#_Toc139925915)

[4.6 Hasil Evaluasi Perfroma Model 31](#_Toc139925916)

[4.7 Hasil Prediksi Decision Tree 32](#_Toc139925917)

[4.8 Menghitung Tingkat Akurasi Model Decision Tree 34](#_Toc139925918)

[BAB V 36](#_Toc139925919)

[KESIMPULAN 36](#_Toc139925920)

[DAFTAR PUSTAKA 37](#_Toc139925921)

# BAB I

# PENDAHULUAN

## Latar Belakang

Kanker paru-paru adalah salah satu jenis kanker yang paling umum dan berbahaya di dunia. Kanker ini terjadi ketika sel-sel di dalam paru-paru mengalami pertumbuhan yang tidak terkendali, membentuk tumor yang dapat menyebar ke bagian tubuh lainnya. Kanker paru-paru dapat menyebabkan gejala yang serius dan dapat berakibat fatal jika tidak dideteksi dan ditangani dengan cepat.

Tingginya angka kejadian kanker paru-paru dan risiko kematian yang tinggi telah mendorong para peneliti dan ahli kesehatan untuk mengidentifikasi faktor-faktor risiko yang terkait dengan penyakit ini. Dalam hal ini, terdapat beberapa faktor risiko yang menjadi fokus dalam analisis ini, yaitu usia, jenis kelamin, risiko genetik, gaya hidup, dan faktor lingkungan.

Faktor usia dan jenis kelamin telah terkait dengan risiko kanker paru-paru. Risiko ini cenderung meningkat seiring dengan bertambahnya usia, sementara perbedaan dalam risiko antara pria dan wanita juga telah diamati.

Risiko genetik juga telah diidentifikasi sebagai faktor yang mempengaruhi kemungkinan seseorang terkena kanker paru-paru. Beberapa faktor genetik dapat meningkatkan kerentanan individu terhadap kanker paru-paru, dan pemahaman lebih lanjut tentang peran faktor genetik dapat membantu dalam pencegahan dan manajemen penyakit.

Selain itu, gaya hidup juga dapat berkontribusi terhadap risiko kanker paru-paru. Kebiasaan merokok dan konsumsi alkohol yang berlebihan telah terbukti menjadi faktor risiko utama. Pola makan seimbang dan tingkat obesitas juga dapat mempengaruhi risiko kanker paru-paru.

Faktor lingkungan juga memiliki peran penting dalam risiko kanker paru-paru. Paparan polusi udara, bahan kimia berbahaya, dan zat karsinogen dapat meningkatkan risiko terkena kanker paru-paru.

Dalam konteks ini, analisis yang dilakukan bertujuan untuk memahami hubungan antara faktor-faktor risiko yang telah disebutkan dengan tingkat risiko kanker paru-paru pada pasien. Dengan memperoleh wawasan yang lebih mendalam tentang faktor-faktor risiko ini, diharapkan dapat membantu dalam upaya pencegahan, deteksi dini, dan pengelolaan kanker paru-paru secara lebih efektif.

## Rumusan Masalah

Berikut masalah yang diangkat pada laporan ini, yaitu :

1. Apakah faktor-faktor seperti usia, jenis kelamin, dan risiko genetik berhubungan dengan tingkat risiko kanker paru-paru pada pasien?
2. Apakah terdapat hubungan antara gaya hidup, seperti merokok, konsumsi alkohol, pola makan seimbang, dan obesitas, dengan tingkat risiko kanker paru-paru?
3. Apakah faktor lingkungan, seperti polusi udara, dan bahaya pekerjaan, mempengaruhi tingkat risiko kanker paru-paru?
4. Dapatkah gejala-gejala yang dialami pasien digunakan untuk memprediksi tingkat risiko kanker paru-paru?

## Tujuan

Tujuan dari riset yang dilakukan adalah:

1. Untuk menentukan apakah ada korelasi antara faktor-faktor seperti usia, jenis kelamin, dan risiko genetik dengan tingkat risiko kanker paru-paru pada pasien.
2. Untuk menentukan pengaruh gaya hidup, seperti merokok, konsumsi alkohol, pola makan seimbang, dan obesitas, terhadap tingkat risiko kanker paru-paru.
3. Untuk mengidentifikasi dampak faktor lingkungan, termasuk polusi udara, bahaya pekerjaan, terhadap tingkat risiko kanker paru-paru.
4. Untuk membangun model prediktif yang dapat memprediksi tingkat risiko kanker paru-paru berdasarkan gejala-gejala yang dialami pasien, serta mengevaluasi performa model tersebut.

## Manfaat

Dalam riset ini kita dapat mengetahui hasil data yang dibutuhkan, menggunakan data yang telah diolah menggunakan aplikasi RStudio.

## Sistematika Pembahasan

Dalam penelitian ini terdapat beberapa sistematika pembahasan dan penyusunan laporan penelitian, dapat diuraikan sebagai berikut:

## BAB I PENDAHULUAN

Pada BAB ini menjelaskan tentang latar belakang, rumusan masalah, tujuan, manfaat, masalah dan sistematika pembahasan dari percobaan.

## BAB II LANDASAN TEORI

Dimana pada BAB ini menjelaskan dan menerangkan tentang landasan teori yang terkait dengan percobaan yang dilakukan.

## BAB III IMPLEMENTASI

Pada BAB ini akan dilakukan implementasi dengan bahasa R di aplikasi RStudio

## BAB IV HASIL PERCOBAAN

Pada BAB ini akan ditampilkan hasil dari percobaan data pada RStudio.

## BAB V KESIMPULAN

Pada BAB ini akan dijelaskan kesimpulan tentang hasil percobaan yang dilakukan.

# BAB II

# LANDASAN TEORI

Pada BAB ini akan dijelaskan tentang landasan yang digunakan untuk melakukan implementasi

# 2.1 Konsep Dasar Bahasa Pemrograman R dan RStudio

## 2.1.1 Bahasa Pemrograman R

1. **Pengertian**

Bahasa R adalah sebuah bahasa pemrograman yang kuat dan fleksibel yang digunakan secara luas dalam statistik, ilmu data, dan analisis data. Sebagai sebuah bahasa pemrograman yang banyak digunakan untuk keperluan analisa data, R dapat dioperasikan pada berbagai sistem operasi pada komputer. Adapun sistem operasi yang didukung antara lain: UNIX, Linux, Windows, dan MacOS.

1. **Fitur**

Bahasa R menawarkan berbagai fitur, sebagai berikut:

1. Bahasa R bersifat open-source, artinya kode sumbernya dapat diakses oleh siapapun secara gratis.
2. Bahasa R bersifat case sensitif. Maksudnta adalah dalam proses input R huruf besar dan kecil sangat diperhatikan.
3. Mendukung export dan import berbagai format file, seperti: TXT, CSV, XLS, dll.
4. Mudah ditingkatkan melalui penambahan fungsi atau library. Penambahan Paket dapat dilakukan secara online melalui CRAN.
5. Interpreted language atau script. Bahasa R memungkinkan pengguna untuk melakukan kerja pada R tanpa perlu kompilasi kode program menjadi bahasa mesin.
6. Mendukung proses loop, decision making, dan menyediakan berbagai jenis operstor (aritmatika, logika, dll).

## 2.1.2 RStudio

1. **Pengertian**

RStudio adalah Integrated Development Environment (IDE) yang dirancang khusus untuk bahasa R. IDE ini menyediakan antarmuka yang intuitif dan user-friendly untuk menulis, menjalankan, dan mengelola kode R serta melakukan analisis data dengan lebih efisien. RStudio dirancang dengan tujuan memudahkan pengguna dalam pengembangan dan eksplorasi data menggunakan bahasa R.

1. **Fitur**

RStudio menyediakan berbagai fitur yang memudahkan pengguna dalam menulis, menjalankan, dan mengelola kode R, serta melakukan analisis data secara efisien. Berikut ini adalah beberapa fungsi utama yang disediakan oleh RStudio:

1. **Editor Kode** : RStudio memiliki editor kode yang lengkap dengan fitur sintaksis highlighting, autocompletion, dan indentasi otomatis. Fungsi ini memudahkan pengguna dalam menulis, mengedit, dan memformat kode R dengan cepat dan akurat.
2. **Console R** : RStudio menyediakan console R yang terintegrasi di mana pengguna dapat menjalankan perintah R secara interaktif. Console ini memungkinkan pengguna untuk melakukan eksperimen, menjalankan fungsi, dan melihat hasil dengan cepat.
3. **Environment dan History** : RStudio menyediakan panel Environment yang menampilkan daftar objek dan variabel yang ada di lingkungan kerja R saat ini. Panel History mencatat perintah yang telah dieksekusi, memungkinkan pengguna untuk mengakses dan mengeksekusi kembali perintah sebelumnya dengan mudah.

# 2.2 Jenis Analisis Data

## 2.2.1 Analisis Deskriptif

Analisis ini melibatkan penjelasan dan pemahaman tentang karakteristik data, seperti statistik ringkasan (mean, median, dan sebagainya), distribusi variabel, dan visualisasi data dengan menggunakan grafik atau diagram.

## 2.2.2 Analisis Korelasi

Analisis ini bertujuan untuk mengevaluasi hubungan antara variabel-variabel yang ada dalam dataset. Misalnya, dapat dilakukan analisis korelasi antara usia, jenis kelamin, risiko genetik, gaya hidup, faktor lingkungan dengan tingkat risiko kanker paru-paru.

## 2.2.3 Analisis Regresi

Analisis regresi dapat digunakan untuk memahami pengaruh variabel-variabel independen, seperti usia, jenis kelamin, risiko genetik, gaya hidup, faktor lingkungan, terhadap variabel dependen yaitu tingkat risiko kanker paru-paru. Ini dapat membantu mengidentifikasi faktor-faktor yang berkontribusi secara signifikan terhadap risiko kanker paru-paru.

## 2.2.4 Analisis Prediktif

Analisis prediktif melibatkan pengembangan model statistik atau model pembelajaran mesin yang dapat memprediksi tingkat risiko kanker paru-paru berdasarkan fitur-fitur yang ada dalam dataset. Misalnya, dapat menggunakan algoritma klasifikasi seperti regresi logistik atau pohon keputusan untuk memprediksi risiko kanker paru-paru berdasarkan gejala-gejala yang dialami pasien.

# 2.3 Metode Yang Digunakan

## 2.3.1 Normalisasi

Normalisasi adalah proses untuk mengubah skala data ke dalam rentang yang lebih konsisten atau standar. Normalisasi dapat diterapkan pada variabel numerik untuk menghindari bias yang disebabkan oleh perbedaan skala. Hal ini memungkinkan pembandingan yang adil antara atribut yang memiliki rentang nilai yang berbeda.

## 2.3.2 Korelasi

Korelasi adalah ukuran statistik yang digunakan untuk menggambarkan hubungan antara dua variabel. Analisis korelasi digunakan untuk mengevaluasi hubungan atau keterkaitan antara variabel-variabel tertentu dengan kemungkinan terjadinya kanker.

## 2.3.3 Regresi Linier

Regresi linier adalah metode statistik yang digunakan untuk memodelkan hubungan linier antara variabel dependen dan satu atau lebih variabel independen.Regresi linier digunakan untuk memodelkan hubungan linier antara variabel dependen (variabel yang ingin diprediksi) dan satu atau lebih variabel independen (variabel yang digunakan untuk melakukan prediksi).

## 2.3.4 Pohon Keputusan (Decision Tree)

Pohon keputusan adalah metode pembelajaran mesin yang menggambarkan struktur pohon dengan pemilihan keputusan berdasarkan atribut tertentu. Pohon keputusan digunakan dalam prediksi kanker karena kemampuannya untuk membuat keputusan berdasarkan atribut-atribut pasien dan menghasilkan prediksi tentang kemungkinan terjadinya kanker.

# 2.4 Konsep Dasar Kanker Paru-Paru

**Pengertian**

Kanker paru-paru adalah suatu kondisi di mana sel-sel tumbuh secara tidak terkendali di dalam organ paru-paru. Organ ini berfungsi untuk menyebarkan oksigen ke dalam darah saat menghirup napas dan membuang karbondioksida saat menghela napas.

**Penyebab**

Penyebab utama kanker paru adalah merokok. Di dalam rokok terdapat zat penyebab kanker (karsinogen) yang memicu kerusakan sel pelapis paru-paru. Perubahan sel dan jaringan pada paru-paru cepat berubah pada perokok berat. Awalnya tubuh dapat memperbaiki kerusakan ini.

Namun, paparan yang berulang membuat sel-sel normal pelapis paru-paru semakin rusak. Kerusakan menyebabkan perubahan abnormal sel, yang berujung pada perkembangan kanker. Di luar dari merokok, mungkin belum ditemukan penyebab pasti dari kanker yang dialami.

**Faktor**

Sejumlah faktor dapat meningkatkan risiko kanker paru. Beberapa faktor risiko dapat dikendalikan, misalnya dengan berhenti merokok. Sedangkan faktor lain seperti riwayat dalam keluarga tidak dapat dikendalikan. Berikut ini beberapa kondisi yang kemungkinan menjadi faktor risiko kanker:

* Merokok. Risiko kanker semakin meningkat seiring dengan jumlah rokok yang dihisap. Berhenti pada usia berapa pun secara signifikan menurunkan risiko kanker.
* Paparan asap rokok. Tidak merokok pun bukan berarti terbebas dari risiko kanker. Risikonya meningkat saat sering terpapar asap rokok.
* Terapi radiasi. Terapi radiasi untuk mengatasi jenis kanker lain dapat meningkatkan risiko terkena kanker.
* Paparan asbes dan karsinogen. Paparan asbes dan zat penyebab kanker, seperti arsenik, kromium, dan nikel meningkatkan risiko kanker.

**Gejala**

Sebagian besar gejala kanker terjadi dalam organ paru, tapi juga mungkin dialami di beberapa bagian tubuh lain jika sel abnormal sudah menyebar. Gejala juga dibedakan berdasarkan intensitas keparahan penyakit. Berikut ini beberapa gejala yang dialami oleh pengidap:

* Ketidaknyamanan atau nyeri di dada.
* Batuk yang tidak hilang atau semakin memburuk dari waktu ke waktu.
* Masalah pernapasan.
* Darah dalam dahak.
* Suara serak.
* Masalah dalam menelan.
* Kehilangan selera makan.
* Kehilangan berat badan tanpa alasan.
* Merasa sangat lelah.
* Peradangan atau sumbatan di paru-paru.
* Pembengkakan atau pembesaran kelenjar getah bening di daerah paru - paru.

## 2.5 Penjelasan Terkait Data Yang Digunakan

Data yang kami gunakan berasal dari dataset pasien kanker yang ditemukan di situs web Kaggle. Dataset ini terdiri dari 25 kolom dengan informasi sebagai berikut:

1. ID pasien: Nomor identifikasi unik untuk setiap pasien.
2. Age (Umur): Umur pasien.
3. Gender (Jenis Kelamin): Jenis kelamin pasien.
4. Air pollution (Polusi udara): Tingkat polusi udara di lingkungan pasien.
5. Alcohol use (Penggunaan minum alkohol): Frekuensi dan jumlah konsumsi alkohol pasien.
6. Dust allergy (Alergi debu): Kondisi pasien terkait alergi terhadap debu.
7. Occupational hazards (Bahaya pekerjaan): Paparan risiko pekerjaan yang mungkin mempengaruhi kesehatan pasien.
8. Genetic risk (Risiko genetik): Faktor risiko genetik yang mungkin mempengaruhi tingkat risiko kanker paru-paru.
9. Chronic lung disease (Penyakit paru-paru kronis): Kondisi pasien terkait penyakit paru-paru kronis yang diderita.
10. Balanced diet (Pola makan seimbang): Tingkat kepatuhan pasien terhadap pola makan seimbang.
11. Obesity (Obesitas): Kondisi pasien terkait kelebihan berat badan.
12. Smoking (Merokok): Status merokok pasien.
13. Passive smoker (Perokok pasif): Paparan asap rokok bagi pasien yang bukan perokok aktif.
14. Chest pain (Nyeri dada): Kondisi pasien terkait nyeri dada.
15. Coughing of blood (Batuk darah): Kondisi pasien terkait batuk berdarah.
16. Fatigue (Kelelahan): Tingkat kelelahan pasien.
17. Weight loss (Penurunan berat badan): Jumlah penurunan berat badan pasien.
18. Shortness of breath (Kesulitan bernapas): Tingkat kesulitan pasien dalam bernapas.
19. Wheezing (Napas berbunyi): Kondisi pasien terkait napas berbunyi saat bernapas.
20. Swallowing difficulty (Kesulitan menelan): Kondisi pasien terkait kesulitan menelan.
21. Clubbing of finger nails (Perubahan bentuk kuku jari): Perubahan bentuk kuku jari pasien.
22. Frequent cold (Pilek sering terjadi): Kondisi pasien terkait seringnya mengalami pilek.
23. Dry cough (Batuk kering): Kondisi pasien terkait batuk kering.
24. Snoring (Sering mendengkur): Kondisi pasien terkait seringnya mendengkur.
25. Level: Tingkat stadium kanker paru-paru pada pasien.

Dalam analisis kami, kami mengkategorikan data tersebut menjadi 4 bagian berdasarkan faktor yang relevan dengan risiko kanker paru-paru, lingkungan, gaya hidup, dan gejala yang dialami pasien. Hal ini membantu kami memfokuskan analisis pada faktor-faktor yang mungkin memiliki hubungan dengan risiko kanker paru-paru dan memahami pengaruhnya terhadap tingkat keparahan kanker paru-paru pada pasien. Data yang kami gunakan juga pasti nya tidak 100% valid/akurat jadi kami hanya menganalisis dan membaca data sesuai dengan data yang kami punya jikalau terdapat perbedaan antara hasil analisis kami dengan hasil analisis kedokteran yang valid.

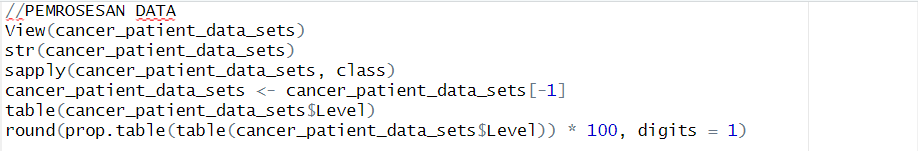
# BAB III

# IMPLEMENTASI

Pada BAB ini akan dijelaskan dan ditampilkan implementasi dari Bahasa R menggunakan aplikasi RStudio.

## Pemrosesan Data

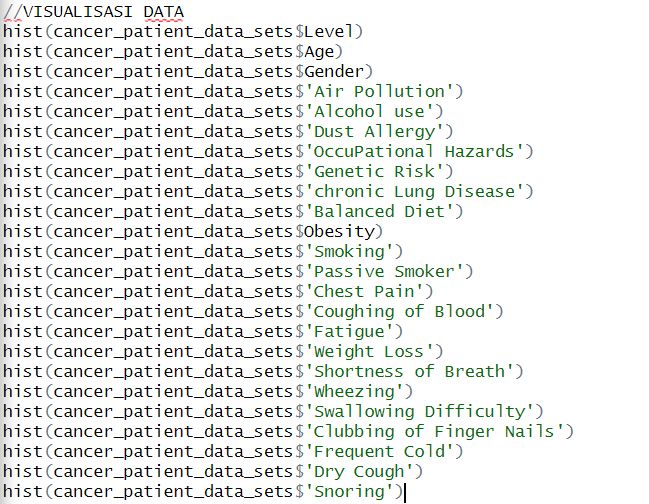
Berikut adalah kode yang kami gunakan untuk menganalisis struktur dataset, memeriksa tipe data dari setiap variabel, menghitung frekuensi dan proporsi relatif dari nilai-nilai dalam kolom Level, dan menghapus kolom tertentu dari dataset jika diperlukan.



*Gambar 1.1 Pemrosesan Data*

## Visualisasi Data Histogram

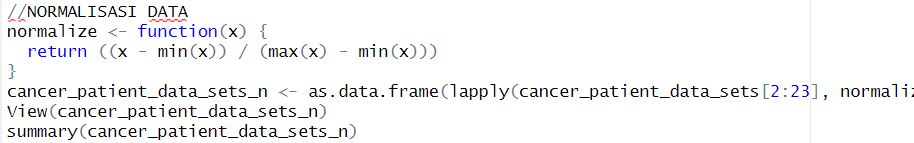
Digunakan untuk menampilkan distribusi frekuensi dari suatu variabel numerik. Histogram berguna untuk memahami pola distribusi data dan melihat apakah terdapat kecondongan atau anomali.



*Gambar 2.1 Visualisasi Data dengan Histogram*

## Normalisasi Data

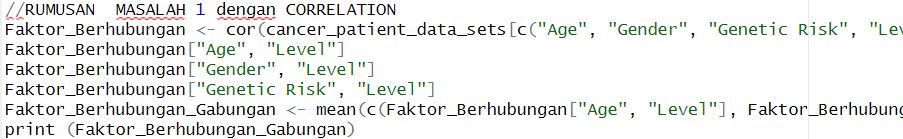
Berikut adalah kode yang kami gunakan untuk melihat data yang telah dinormalisasi dan mendapatkan ringkasan statistiknya.



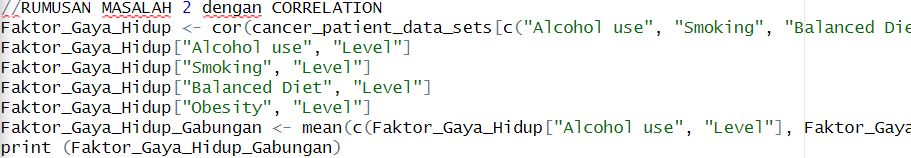
*Gambar 3.1 Normalisasi Data*

## Menggunakan Correlation

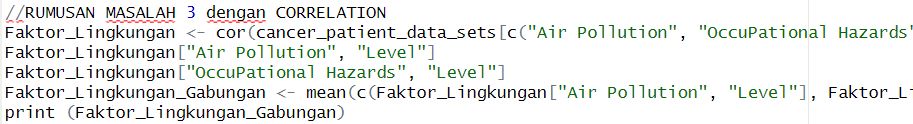
Berikut adalah kode yang kami pakai untuk analisis Correlation di RStudio. Fungsi dari kode ini ialah mencari korelasi antara variabel x dan y sebagaimana yang telah kami tulis di kode berikut.



*Gambar 4.1 Rumusan Masalah 1 dengan Correlation*



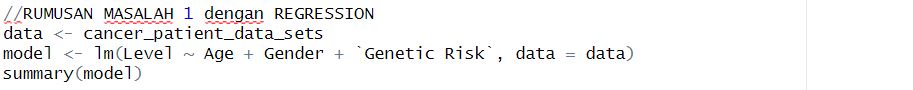
*Gambar 4.2 Rumusan Masalah 2 dengan Correlation*



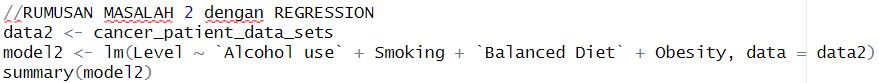
*Gambar 4.3 Rumusan Masalah 3 dengan Correlation*

## Menggunakan Regression

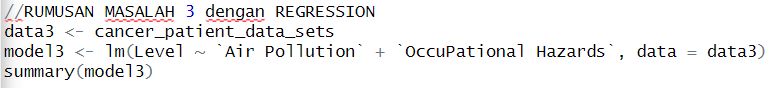
Berikut adalah kode yang kami pakai untuk analisis Regression di RStudio. Fungsi dari kode ini ialah untuk memodelkan hubungan antara satu atau lebih variabel sebagaimana yang telah kami tulis di kode berikut.



*Gambar 5.1 Rumusan Masalah 1 dengan Regression*

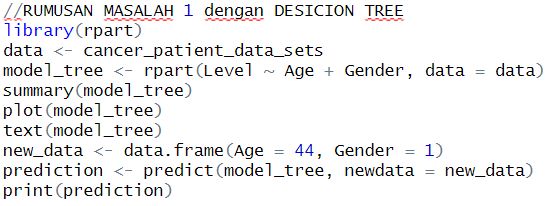


*Gambar 5.2 Rumusan Masalah 2 dengan Regression*

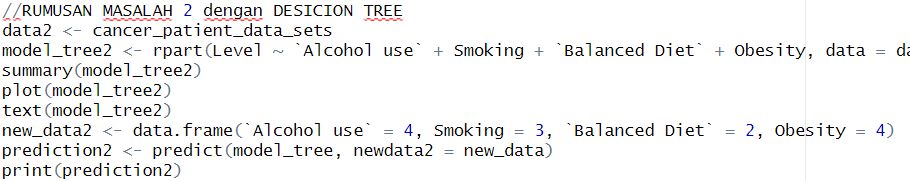


*Gambar 5.3 Rumusan Masalah 3 dengan Regression*

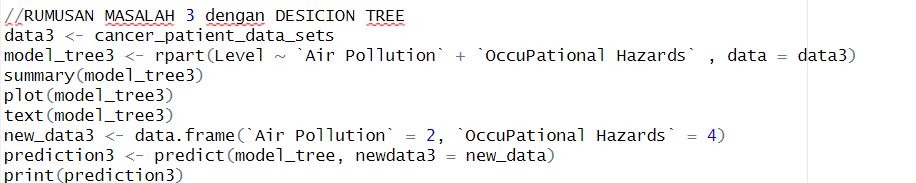
## Menggunakan Decision Tree



*Gambar 6.1 Rumusan Masalah 1 dengan Decision Tree*

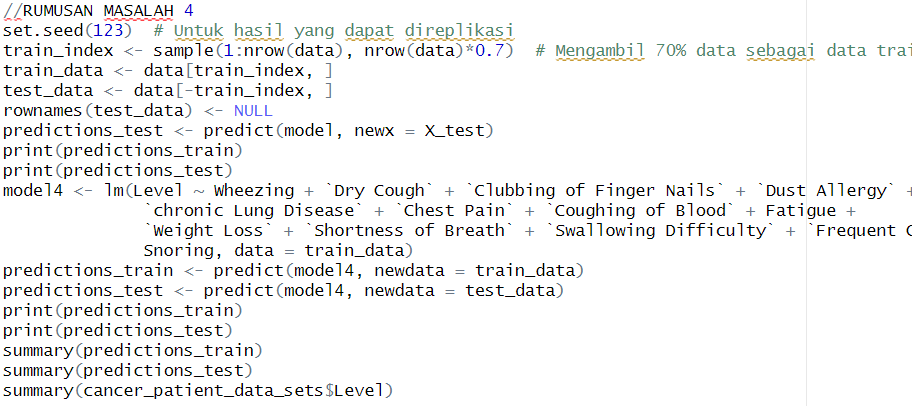


*Gambar 6.2 Rumusan Masalah 2 dengan Decision Tree*



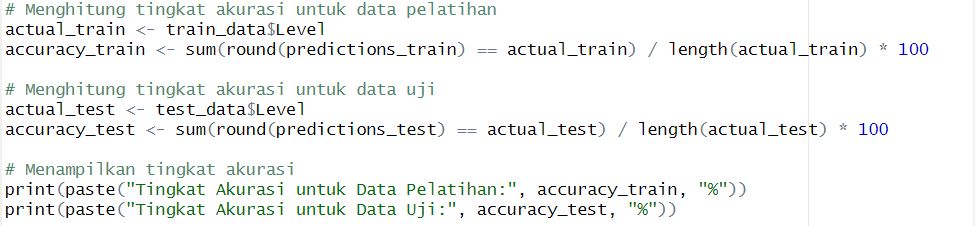
*Gambar 6.3 Rumusan Masalah 3 dengan Decision Tree*

## Analisis Prediksi Menggunakan Linear Regression



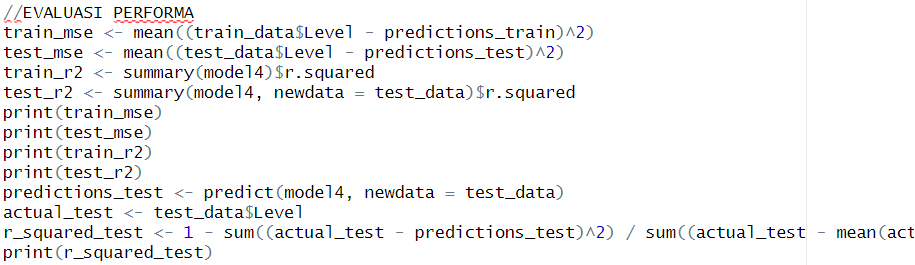
*Gambar 7.1 Rumusan Masalah 4 dengan Linear Regression*

## Menghitung Tingkat Akurasi Model dengan Linear Regression



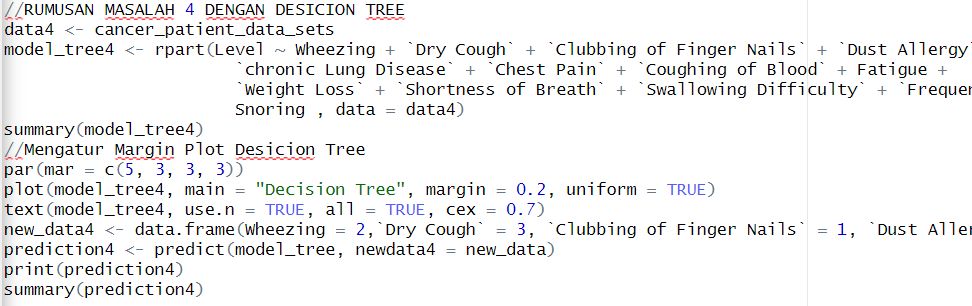
*Gambar 8.1 Menghitung Tingkat Akurasi Model dengan Linier Regression*

## Evaluasi Performa Model dengan Linear Regression

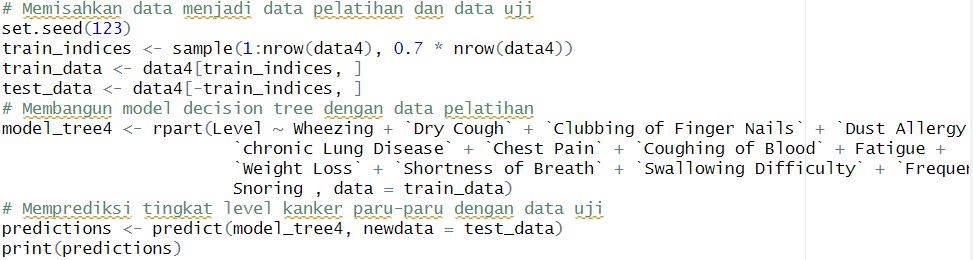
**

*Gambar 9.1 Menghitung Tingkat Akurasi Model dengan Linier Regression*

## Analisis Prediksi Menggunakan Decision Tree

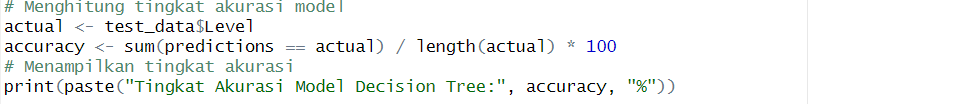


*Gambar 10.1 Rumusan Masalah 4 dengan Decision Tree*



*Gambar 10.2 Rumusan Masalah 4 dengan Decision Tree*

## Menghitung Tingkat Akurasi Model dengan Decision Tree



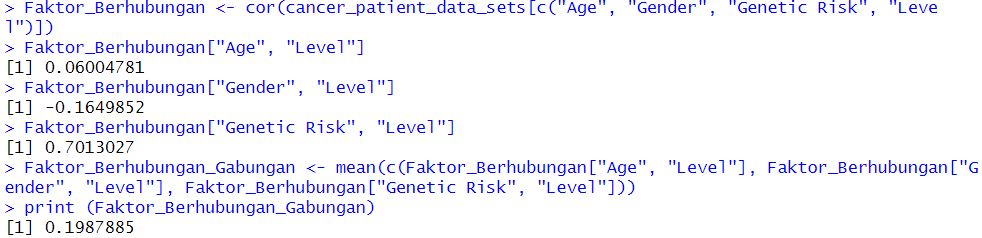
*Gambar 11.1 Menghitung Tingkat Akurasi Model dengan Decision Tree*

# BAB IV

# HASIL ANALISIS

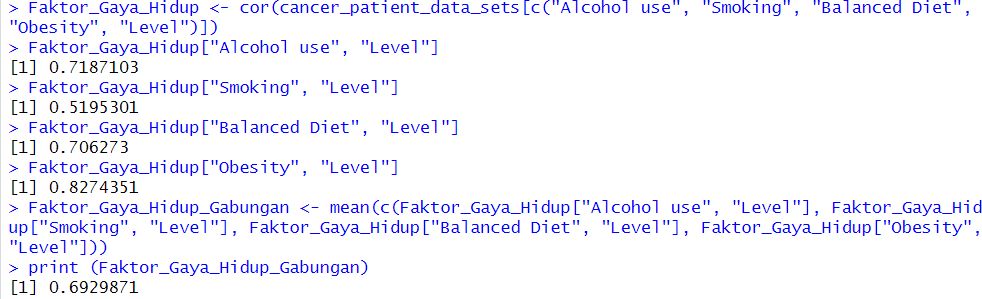


## Hasil Correlation



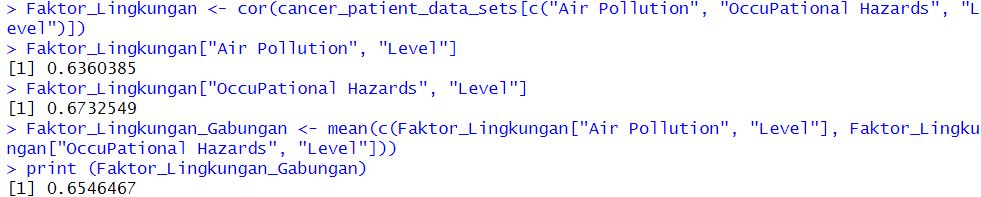
*Gambar 9.1 Hasil Correlation Rumusan Masalah Pertama*

Dari hasil diatas dapat diambil kesimpulan bahwa Umur dan Gender tidak berpengaruh pada level Kanker paru-paru, yang sangat berpengaruh ialah faktor Genetic Risk dengan angka 0.701, jika kita gabungkan ketiga faktor tersebut menjadi grup faktor berhubungan gabungan maka akan memperoleh nilai 0.1987885 yaitu tidak berpengaruh. Hal ini disebabkan angka korelasi Age dan Gender sangat rendah dan bertolak belakang sehingga menghasilkan nilai rata-rata gabungan yang sangat kecil.



*Gambar 9.2 Hasil Correlation Rumusan Masalah Kedua*

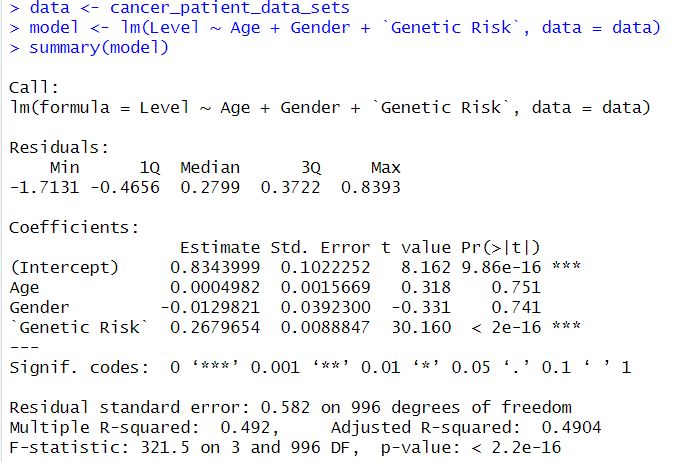
Dari hasil diatas dapat diambil kesimpulan bahwa penggunaan minuman beralkohol , pola makan sehat, dan kondisi pasien terkait berat badan berlebih sangat mempengaruhi level kanker paru-paru, dari analisis kami justru faktor merokok tidak terlalu berdampak pada level kanker paru-paru. Jika kita gabungkan ke empat faktor tersebut menjadi faktor gaya hidup gabungan maka akan memperoleh nilai 0.6929871 yaitu berpengaruh. Perlu diingat bahwa data yang kami gunakan belum terbukti valid jadi kami hanya membaca data sesuai analisis kami dengan data yang ada saja.



*Gambar 9.3 Hasil Correlation Rumusan Masalah Ketiga*

Dari hasil diatas dapat diambil kesimpulan bahwa kualitas udara yang buruk dan Paparan risiko pekerjaan yang mungkin mempengaruhi kesehatan pasien. dapat berpengaruh terhadap level kanker paru-paru. Jika kita gabungkan kedua faktor tersebut menjadi faktor lingkungan gabungan maka akan memperoleh nilai 0.6546467 yaitu berpengaruh.

## Hasil Regression

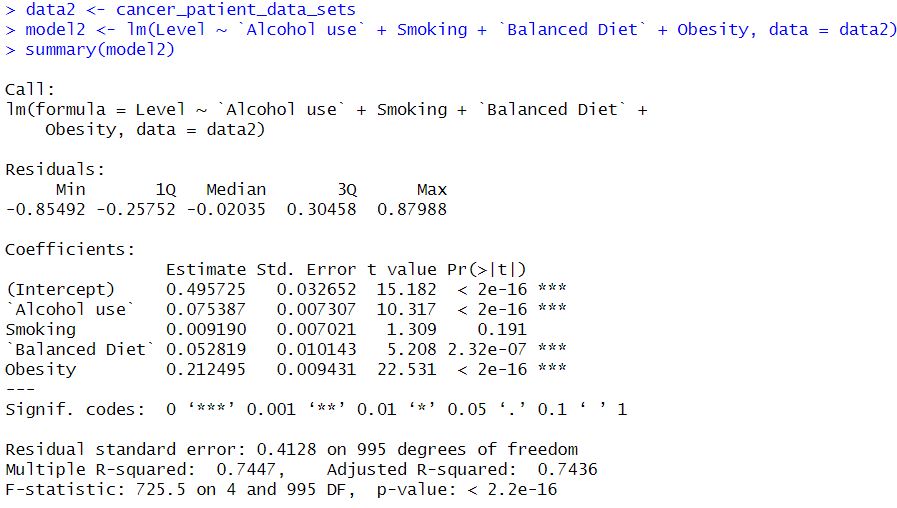


*Gambar 10.1 Hasil Regression Rumusan Masalah Pertama*

Dalam model regresi ini, kita menggunakan variabel umur (`Age`), jenis kelamin (`Gender`), dan risiko genetik (`Genetic Risk`) sebagai prediktor untuk memprediksi tingkat kanker. Berikut adalah hasil ringkasan model:

* + 1. **Koefisien:**
* Untuk setiap tahun peningkatan usia (`Age`), tingkat kanker cenderung meningkat sebesar 0.0004982. Namun, perubahan ini tidak signifikan secara statistik.
* Untuk jenis kelamin (`Gender`), tidak ada pengaruh yang signifikan terhadap tingkat kanker.
* Untuk risiko genetik (`Genetic Risk`), setiap peningkatan sebesar 1 unit pada skala risiko, tingkat kanker cenderung meningkat sebesar 0.2679654.
  + 1. **Tingkat signifikansi:**
* Koefisien untuk `Age` dan `Gender` tidak signifikan secara statistik, yang berarti perubahan dalam variabel ini tidak memiliki hubungan yang kuat dengan tingkat kanker.
* Koefisien untuk `Genetic Risk` sangat signifikan secara statistik, menunjukkan bahwa risiko genetik memiliki hubungan yang kuat dengan tingkat kanker.
  + 1. **Tingkat variasi:**
* Multiple R-squared menunjukkan bahwa model dapat menjelaskan sekitar 49.2% variasi dalam tingkat kanker.
* Adjusted R-squared adalah versi penyesuaian dari R-squared yang mempertimbangkan jumlah prediktor dan ukuran sampel. Nilai ini adalah 49.04%, menunjukkan tingkat variasi yang dijelaskan oleh model dengan mempertimbangkan kompleksitasnya.
  + 1. **Kesimpulan:**
* Usia dan jenis kelamin tidak memiliki pengaruh yang signifikan terhadap tingkat kanker dalam model ini.
* Namun, risiko genetik memiliki pengaruh yang signifikan terhadap tingkat kanker. Semakin tinggi risiko genetik, semakin tinggi tingkat kanker.

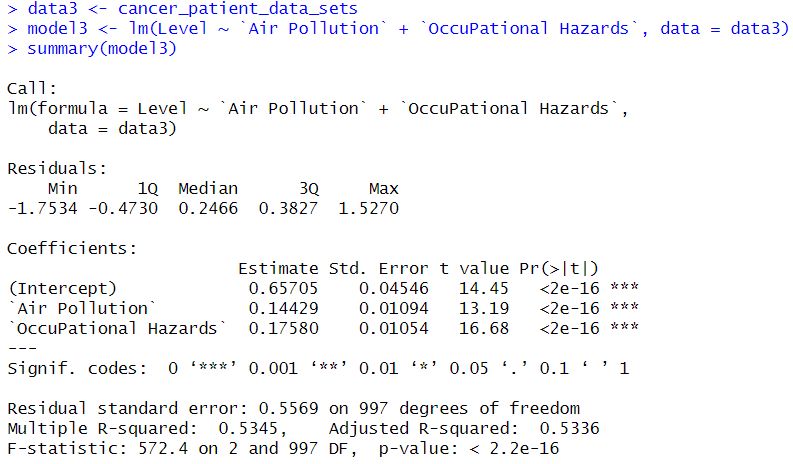
Perlu diingat bahwa hasil ini didasarkan pada analisis statistik dari data kanker pasien yang kita peroleh. Penting untuk diingat bahwa hasil ini hanya mencerminkan hubungan yang ditemukan dalam sampel data tersebut dan tidak dapat digeneralisasi secara langsung ke populasi yang lebih luas.



*Gambar 10.2 Hasil Regression Rumusan Masalah Kedua*

Dalam model regresi ini, kita menggunakan variabel penggunaan alkohol (`Alcohol use`), kebiasaan merokok (`Smoking`), pola makan seimbang (`Balanced Diet`), dan kegemukan (`Obesity`) sebagai prediktor untuk memprediksi tingkat kanker. Berikut adalah hasil ringkasan model:

* + 1. **Koefisien:**
* Penggunaan alkohol (`Alcohol use`) memiliki pengaruh positif yang signifikan terhadap tingkat kanker. Setiap peningkatan dalam penggunaan alkohol sebesar 1 unit, tingkat kanker cenderung meningkat sebesar 0.075387.
* Kebiasaan merokok (`Smoking`) tidak memiliki pengaruh yang signifikan terhadap tingkat kanker.
* Pola makan seimbang (`Balanced Diet`) memiliki pengaruh positif yang signifikan terhadap tingkat kanker. Setiap perbaikan dalam pola makan seimbang, tingkat kanker cenderung meningkat sebesar 0.052819.
* Kegemukan (`Obesity`) memiliki pengaruh positif yang signifikan terhadap tingkat kanker. Setiap peningkatan dalam tingkat kegemukan sebesar 1 unit, tingkat kanker cenderung meningkat sebesar 0.212495.
  + 1. **Tingkat signifikansi:**
* Koefisien untuk `Alcohol use`, `Balanced Diet`, dan `Obesity` sangat signifikan secara statistik, menunjukkan hubungan yang kuat antara variabel-variabel ini dengan tingkat kanker.
* Koefisien untuk `Smoking` tidak signifikan secara statistik, yang berarti kebiasaan merokok tidak memiliki hubungan yang kuat dengan tingkat kanker dalam model ini.
  + 1. **Tingkat variasi:**
* Multiple R-squared menunjukkan bahwa model ini dapat menjelaskan sekitar 74.47% variasi dalam tingkat kanker.
* Adjusted R-squared adalah versi penyesuaian dari R-squared yang mempertimbangkan jumlah prediktor dan ukuran sampel. Nilai ini adalah 74.36%, menunjukkan tingkat variasi yang dijelaskan oleh model dengan mempertimbangkan kompleksitasnya.
  + 1. **Kesimpulan:**
* Penggunaan alkohol, pola makan seimbang, dan kegemukan memiliki pengaruh yang signifikan terhadap tingkat kanker dalam model ini.
* Kebiasaan merokok tidak memiliki pengaruh yang signifikan terhadap tingkat kanker.
* Model ini dapat memberikan pemahaman yang baik tentang faktor-faktor yang mempengaruhi tingkat kanker berdasarkan variabel yang digunakan.



*Gambar 10.3 Hasil Regression Rumusan Masalah Ketiga*

Dalam model regresi ini, kita menggunakan variabel polusi udara (`Air Pollution`) dan bahaya pekerjaan (`OccuPational Hazards`) sebagai prediktor untuk memprediksi tingkat kanker. Berikut adalah hasil ringkasan model:

1. **Koefisien:**

* Polusi udara (`Air Pollution`) memiliki pengaruh positif yang signifikan terhadap tingkat kanker. Setiap peningkatan dalam polusi udara sebesar 1 unit, tingkat kanker cenderung meningkat sebesar 0.14429.
* Bahaya pekerjaan (`OccuPational Hazards`) juga memiliki pengaruh positif yang signifikan terhadap tingkat kanker. Setiap peningkatan dalam tingkat bahaya pekerjaan sebesar 1 unit, tingkat kanker cenderung meningkat sebesar 0.17580.

1. **Tingkat signifikansi:**

Koefisien untuk `Air Pollution` dan `OccuPational Hazards` sangat signifikan secara statistik, menunjukkan hubungan yang kuat antara variabel-variabel ini dengan tingkat kanker.

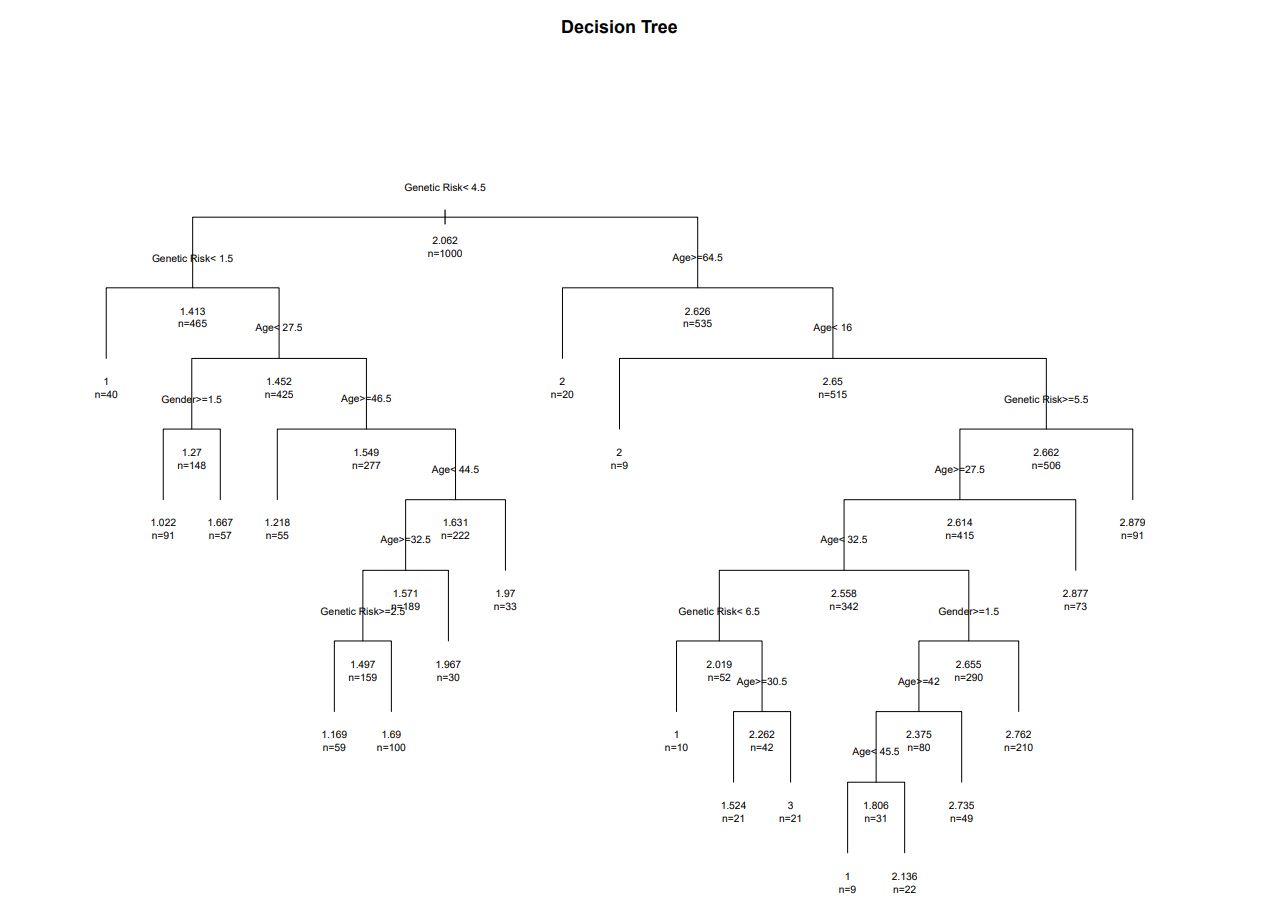
1. **Tingkat variasi:**

* Multiple R-squared menunjukkan bahwa model ini dapat menjelaskan sekitar 53.45% variasi dalam tingkat kanker.
* Adjusted R-squared adalah versi penyesuaian dari R-squared yang mempertimbangkan jumlah prediktor dan ukuran sampel. Nilai ini adalah 53.36%, menunjukkan tingkat variasi yang dijelaskan oleh model dengan mempertimbangkan kompleksitasnya.

1. **Kesimpulan:**

* Polusi udara dan bahaya pekerjaan memiliki pengaruh yang signifikan terhadap tingkat kanker dalam model ini.
* Model ini dapat memberikan pemahaman yang baik tentang faktor-faktor yang mempengaruhi tingkat kanker berdasarkan variabel yang digunakan.

## Hasil Decision Tree



*Gambar 11.1 Hasil Decision Tree Rumusan Masalah Pertama*

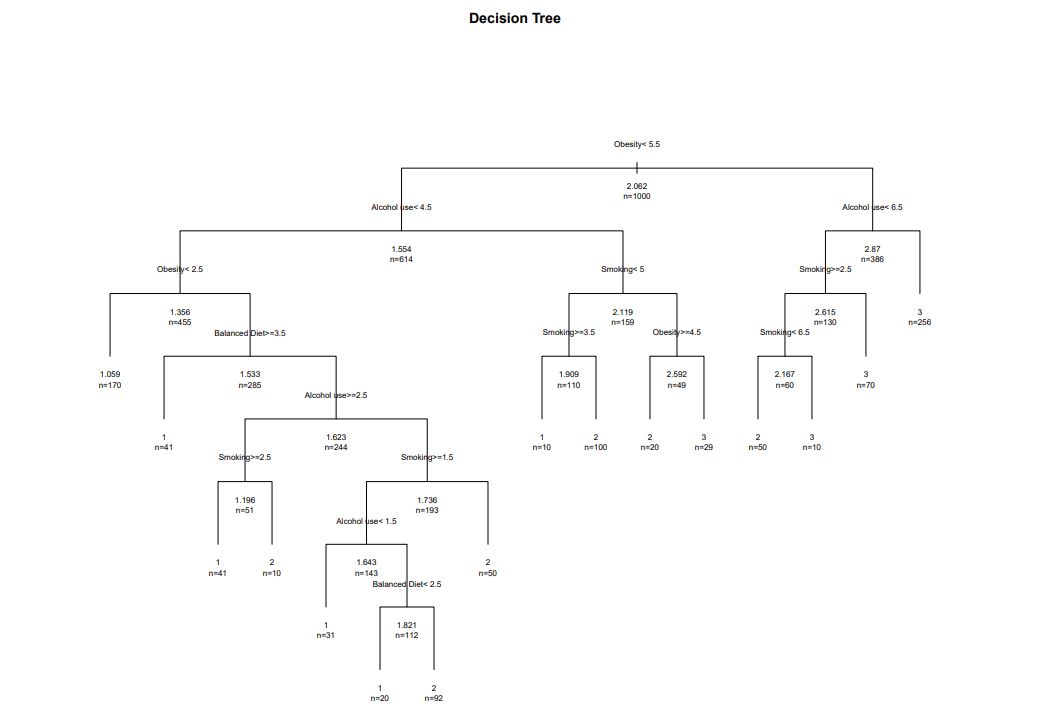
Dalam kasus ini, variabel "Genetic Risk" memiliki tingkat penting sebesar 63%, diikuti oleh "Age" sebesar 25% dan "Gender" sebesar 12%. Terdapat beberapa node penting dalam model decision tree rumusan masalah pertama:

* + 1. Node number 2 memiliki 465 observasi. Jika nilai atribut Genetic Risk kurang dari 1.5, observasi akan diklasifikasikan ke anak kiri. Jika tidak, observasi akan diklasifikasikan ke anak kanan. Rata-rata tingkat Level pada node ini adalah 1.412903 dengan Mean Squared Error (MSE) sebesar 0.2424142.
    2. Node number 3 memiliki 535 observasi. Jika nilai atribut Age kurang dari 64.5, observasi akan diklasifikasikan ke anak kiri. Jika tidak, observasi akan diklasifikasikan ke anak kanan. Rata-rata tingkat Level pada node ini adalah 2.626168 dengan MSE sebesar 0.3462311.
    3. Node number 5 memiliki 425 observasi. Jika nilai atribut Age kurang dari 27.5, observasi akan diklasifikasikan ke anak kiri. Jika tidak, observasi akan diklasifikasikan ke anak kanan. Rata-rata tingkat Level pada node ini adalah 1.451765 dengan MSE sebesar 0.2476734.
    4. Node number 11 memiliki 277 observasi. Jika nilai atribut Age kurang dari 46.5, observasi akan diklasifikasikan ke anak kiri. Jika tidak, observasi akan diklasifikasikan ke anak kanan. Rata-rata tingkat Level pada node ini adalah 1.548736 dengan MSE sebesar 0.2476248.
    5. Node number 15 memiliki 506 observasi. Jika nilai atribut Age kurang dari 16, observasi akan diklasifikasikan ke anak kiri. Jika tidak, observasi akan diklasifikasikan ke anak kanan. Rata-rata tingkat Level pada node ini adalah 2.662055 dengan MSE sebesar 0.3423151.
    6. Node number 31 memiliki 91 observasi. Node ini merupakan node terakhir dalam model decision tree. Rata-rata tingkat Level pada node ini adalah 2.879121 dengan MSE sebesar 0.1062674.



*Gambar 11.2 Hasil Decision Tree Rumusan Masalah Pertama*

Prediksi ini menghasilkan angka yang menunjukkan tingkat keparahan kanker yang mungkin untuk setiap pasien dalam rentang angka. Contohnya, prediksi pertama memiliki tingkat keparahan kanker sebesar 1.169492, sedangkan prediksi kedua memiliki tingkat keparahan sebesar 1.666667, dan seterusnya. Prediksi tingkat keparahan kanker untuk data baru dapat memberikan informasi yang berguna dalam mengevaluasi tingkat keparahan kanker pasien-pasien yang belum diketahui sebelumnya. Namun, untuk memberikan kesimpulan yang lebih mendalam, perlu dianalisis lebih lanjut mengenai performa model, ukuran dataset, dan validitas dari fitur-fitur yang digunakan.



*Gambar 11.3 Hasil Decision Tree Rumusan Masalah Kedua*

Dalam kasus ini, variabel " Obesity" memiliki tingkat penting sebesar 32%, diikuti oleh " Balanced Diet" sebesar 26%, “Alcohol use” sebesar 23% dan "Smoking" sebesar 20%.

Berikut adalah penjelasan singkat mengenai beberapa node yang terdapat dalam model:

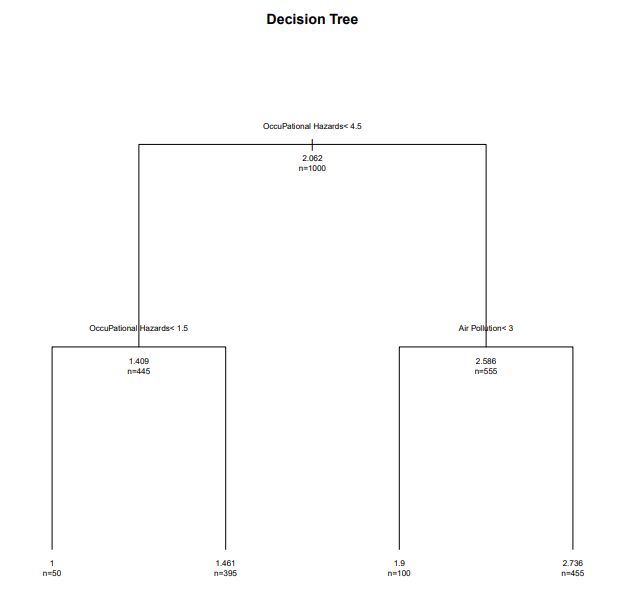
* + 1. Node number 1 (Akar): Node ini merupakan titik awal dari pohon keputusan. Data pada node ini dibagi berdasarkan nilai atribut Obesity, Alcohol use, Balanced Diet, dan Smoking. Node ini memberikan gambaran tentang kondisi awal sebelum memasuki pemilihan cabang berikutnya.
    2. Node number 2: Node ini membagi data berdasarkan nilai atribut-alcohol use. Jika nilai atribut Alcohol use kurang dari 4.5, observasi diklasifikasikan ke anak kiri (Node 4), jika tidak, diklasifikasikan ke anak kanan (Node 5).
    3. Node number 3: Pada node ini, data dibagi berdasarkan nilai atribut Smoking. Jika nilai atribut Smoking kurang dari 6.5, observasi akan diklasifikasikan ke anak kiri (Node 6), jika tidak, akan diklasifikasikan ke anak kanan (Node 7).
    4. Node number 4: Node ini membagi data berdasarkan nilai atribut Obesity. Jika nilai atribut Obesity kurang dari 2.5, observasi akan diklasifikasikan ke anak kiri (Node 8), jika tidak, akan diklasifikasikan ke anak kanan (Node 9).

1. Node number 5: Pada node ini, data dibagi berdasarkan nilai atribut Smoking. Jika nilai atribut Smoking kurang dari 5, observasi akan diklasifikasikan ke anak kiri (Node 10), jika tidak, akan diklasifikasikan ke anak kanan (Node 11).
2. Node number 6: Node ini membagi data berdasarkan nilai atribut Smoking. Jika nilai atribut Smoking kurang dari 2.5, observasi akan diklasifikasikan ke anak kanan (Node 12), jika tidak, akan diklasifikasikan ke anak kanan (Node 13).
3. Node number 7: Node ini tidak melakukan pemisahan data lebih lanjut dan tidak memiliki anak. Hal ini menunjukkan bahwa observasi yang masuk ke node ini akan diklasifikasikan sebagai tingkat Level yang lebih tinggi (mean=3).



*Gambar 11.4 Hasil Decision Tree Rumusan Masalah Kedua*

Setiap angka dalam output merupakan prediksi nilai target untuk data tersebut. Misalnya, prediksi untuk data pertama adalah 1.169492, prediksiuntuk data kedua adalah 1.666667, dan seterusnya.



*Gambar 11.5 Hasil Decision Tree Rumusan Masalah Ketiga*

Terdapat dua variabel penting dalam model ini, yaitu "OccuPational Hazards" (resiko pekerjaan) dengan tingkat kepentingan 55% dan "Air Pollution" (polusi udara) dengan tingkat kepentingan 45%. Berikut adalah penjelasan singkat mengenai node yang terdapat dalam model:

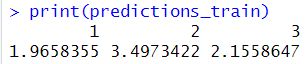
1. Node number 1 (Akar): Node ini merupakan titik awal dari pohon keputusan. Data pada node ini dibagi berdasarkan nilai atribut Obesity, Alcohol use, Balanced Diet, dan Smoking. Node ini memberikan gambaran tentang kondisi awal sebelum memasuki pemilihan cabang berikutnya.
2. Node number 2: Node ini membagi data berdasarkan nilai atribut Alcohol use. Jika nilai atribut Alcohol use kurang dari 4.5, observasi diklasifikasikan ke anak kiri (Node 4), jika tidak, diklasifikasikan ke anak kanan (Node 5).
3. Node number 3: Pada node ini, data dibagi berdasarkan nilai atribut Smoking. Jika nilai atribut Smoking kurang dari 6.5, observasi akan diklasifikasikan ke anak kiri (Node 6), jika tidak, akan diklasifikasikan ke anak kanan (Node 7).
4. Node number 4: Node ini membagi data berdasarkan nilai atribut Obesity. Jika nilai atribut Obesity kurang dari 2.5, observasi akan diklasifikasikan ke anak kiri (Node 8), jika tidak, akan diklasifikasikan ke anak kanan (Node 9).
5. Node number 5: Pada node ini, data dibagi berdasarkan nilai atribut Smoking. Jika nilai atribut Smoking kurang dari 5, observasi akan diklasifikasikan ke anak kiri (Node 10), jika tidak, akan diklasifikasikan ke anak kanan (Node 11).
6. Node number 6: Node ini membagi data berdasarkan nilai atribut Smoking. Jika nilai atribut Smoking kurang dari 2.5, observasi akan diklasifikasikan ke anak kiri (Node 12), jika tidak, akan diklasifikasikan ke anak kanan (Node 13).
7. Node number 7: Node ini tidak melakukan pemisahan data lebih lanjut dan tidak memiliki anak. Hal ini menunjukkan bahwa observasi yang masuk ke node ini akan diklasifikasikan sebagai tingkat Level yang lebih tinggi (mean=3).



*Gambar 11.6 Hasil Decision Tree Rumusan Masalah Ketiga*

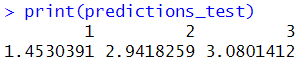
Setiap angka dalam output merupakan nilai prediksi untuk satu data point. Misalnya, prediksi pertama memiliki nilai 1.169492, prediksi kedua memiliki nilai 1.666667, dan seterusnya.

## Hasil Prediksi Linear Regression



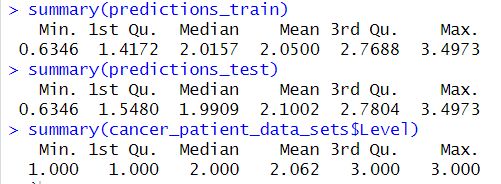
*Gambar 12.1 Hasil Prediksi Linear Regression*

Setiap angka dalam daftar merupakan prediksi nilai yang diberikan oleh model untuk satu data. Misalnya, prediksi pertama memiliki nilai 1.9658355, prediksi kedua memiliki nilai 3.4973422, dan seterusnya. Terdapat total 700 prediksi dalam daftar tersebut.



*Gambar 12.2 Hasil Prediksi Linear Regression*

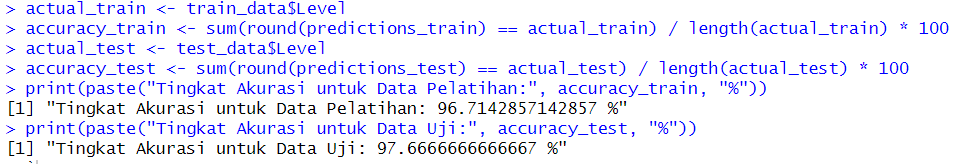
Prediksi pertama memiliki nilai 1.4530391, prediksi kedua memiliki nilai 2.9418259, dan seterusnya. Nilai-nilai ini dapat digunakan untuk mengevaluasi performa model atau untuk pengambilan keputusan berdasarkan prediksi yang diberikan.



*Gambar 12.3 Hasil Prediksi Linear Regression*

Pada data pelatihan (predictions\_train), nilai prediksi Level memiliki rentang antara 0.6346 hingga 3.4973, dengan nilai median sebesar 2.0157 dan mean (rata-rata) sebesar 2.0500. Sedangkan pada data uji (predictions\_test), nilai prediksi Level memiliki rentang yang sama dengan data pelatihan, yaitu antara 0.6346 hingga 3.4973, dengan median sebesar 1.9909 dan mean sebesar 2.1002. Sedangkan pada data asli (cancer\_patient\_data\_sets$Level), tingkat Level memiliki rentang dari 1 hingga 3, dengan median sebesar 2 dan mean sebesar 2.062.

* 1. **Menghitung Tingkat Akurasi Model dengan Linear Regression**

****

*Gambar 13.1 4.5 Menghitung Tingkat Akurasi Model dengan Linear Regression*

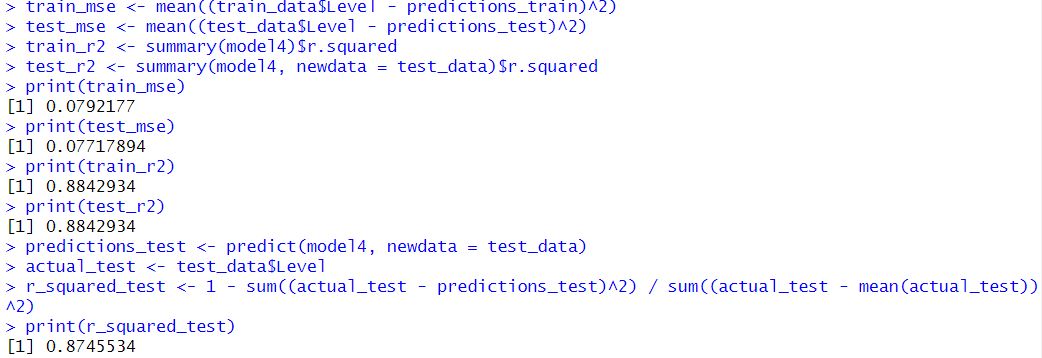
Hasil output di atas menunjukkan tingkat akurasi dari model yang telah dilatih menggunakan data pelatihan dan diuji menggunakan data uji. Tingkat akurasi dinyatakan dalam persentase.

Untuk data pelatihan, tingkat akurasi mencapai 96.71%, yang berarti model berhasil memprediksi dengan benar sekitar 96.71% dari data pelatihan yang digunakan.

Sedangkan untuk data uji, tingkat akurasi mencapai 97.67%, yang berarti model berhasil memprediksi dengan benar sekitar 97.67% dari data uji yang digunakan.

Hasil ini menunjukkan bahwa model memiliki kinerja yang baik dalam memprediksi klasifikasi data, dengan tingkat akurasi yang tinggi pada kedua set data.

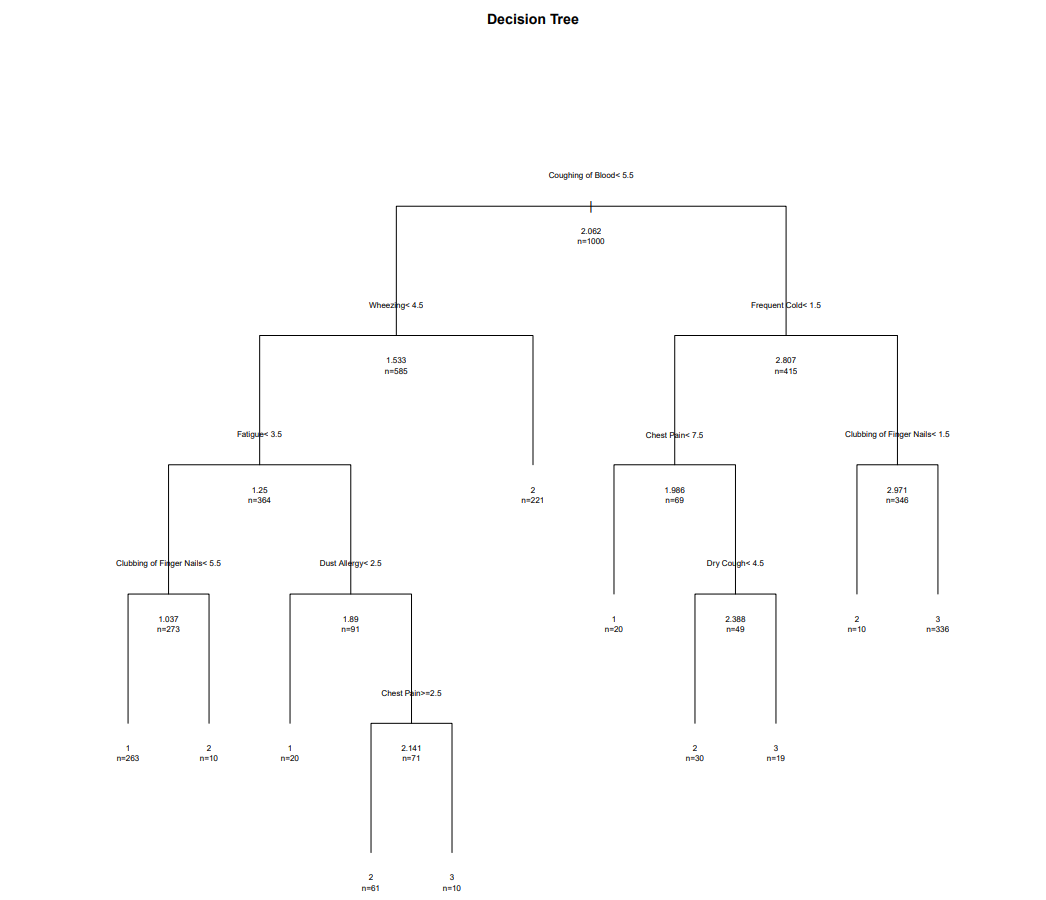
## Hasil Evaluasi Perfroma Model



*Gambar 13.1 Hasil Evaluasi Performa Model*

Hasil evaluasi model menunjukkan bahwa Mean Squared Error (MSE) untuk data pelatihan (train\_mse) adalah sebesar 0.0792177 dan untuk data uji (test\_mse) adalah sebesar 0.07717894. Koefisien determinasi (R-squared) untuk data pelatihan (train\_r2) dan data uji (test\_r2) adalah 0.8842934, yang menunjukkan bahwa model berhasil menjelaskan sekitar 88.43% variabilitas dalam data. Selain itu, koefisien determinasi (r\_squared\_test) yang dihitung secara manual untuk data uji adalah 0.8745534, yang juga mengindikasikan tingkat kecocokan yang baik antara prediksi model dengan data aktual.

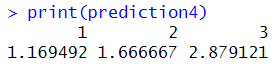
## Hasil Prediksi Decision Tree

****

*Gambar 14.1 Hasil Prediksi Menggunakan Decision Tree*

Beberapa faktor gejala memiliki tingkat penting yang berbeda terhadap variabel Level dalam model pohon keputusan. Faktor Coughing of Blood memiliki tingkat penting sebesar 19%, diikuti oleh Chest Pain dengan tingkat penting 17%, dan Dust Allergy dengan tingkat penting 15%. Faktor lainnya seperti Chronic Lung Disease memiliki tingkat penting 12%, Fatigue 10%, dan Wheezing 9%. Sementara itu, faktor Clubbing of Finger Nails, Frequent Cold, Weight Loss, Shortness of Breath, Snoring, Swallowing Difficulty, dan Dry Cough memiliki tingkat penting di bawah 5%. Hal ini menunjukkan bahwa faktor-faktor tersebut berkontribusi secara berbeda dalam mempengaruhi tingkat Level pada pasien kanker, dengan beberapa faktor memiliki pengaruh yang lebih besar daripada yang lain. Berikut adalah penjelasan singkat mengenai node yang terdapat dalam model:

* + 1. Node number 1 (Akar): Node ini merupakan titik awal dari pohon keputusan. Data pada node ini dibagi berdasarkan nilai atribut Obesity, Alcohol use, Balanced Diet, dan Smoking. Node ini memberikan gambaran tentang kondisi awal sebelum memasuki pemilihan cabang berikutnya.
    2. Node number 2: Node ini membagi data berdasarkan nilai atribut alcohol use. Jika nilai atribut Alcohol use kurang dari 4.5, observasi diklasifikasikan ke anak kiri (Node 4), jika tidak, diklasifikasikan ke anak kanan (Node 5).
    3. Node number 3: Pada node ini, data dibagi berdasarkan nilai atribut Smoking. Jika nilai atribut Smoking kurang dari 6.5, observasi akan diklasifikasikan ke anak kiri (Node 6), jika tidak, akan diklasifikasikan ke anak kanan (Node 7).
    4. Node number 4: Node ini membagi data berdasarkan nilai atribut Obesity. Jika nilai atribut Obesity kurang dari 2.5, observasi akan diklasifikasikan ke anak kiri (Node 8), jika tidak, akan diklasifikasikan ke anak kanan (Node 9).
    5. Node number 5: Pada node ini, data dibagi berdasarkan nilai atribut Smoking. Jika nilai atribut Smoking kurang dari 5, observasi akan diklasifikasikan ke anak kiri (Node 10), jika tidak, akan diklasifikasikan ke anak kanan (Node 11).
    6. Node number 6: Node ini membagi data berdasarkan nilai atribut Smoking. Jika nilai atribut Smoking kurang dari 2.5, observasi akan diklasifikasikan ke anak kanan (Node 12), jika tidak, akan diklasifikasikan ke anak kanan (Node 13).
    7. Node number 7: Node ini tidak melakukan pemisahan data lebih lanjut dan tidak memiliki anak. Hal ini menunjukkan bahwa observasi yang masuk ke node ini akan diklasifikasikan sebagai tingkat Level yang lebih tinggi (mean=3).



*Gambar 14.2 Hasil Prediksi Menggunakan Decision Tree*

Variabel ini digunakan untuk menyimpan hasil prediksi dari model decision tree (model\_tree4) terhadap new\_data4. Setiap angka dalam daftar merupakan prediksi nilai yang diberikan oleh model untuk satu data. Misalnya, prediksi pertama memiliki nilai 1.169492, prediksi kedua memiliki nilai 1.666667, dan seterusnya. Terdapat total 1000 prediksi dalam daftar tersebut.



*Gambar 14.3 Hasil Prediksi Menggunakan Decision Tree*

Variabel ini digunakan untuk menyimpan hasil prediksi dari model decision tree (model\_tree4) terhadap data uji (test\_data). Prediksi ini dilakukan berdasarkan fitur-fitur yang terdapat pada test\_data. Setiap angka dalam daftar merupakan prediksi nilai yang diberikan oleh model untuk satu data. Misalnya, prediksi pertama memiliki nilai 1.025126, prediksi kedua memiliki nilai 3.000000, dan seterusnya. Terdapat total 300 prediksi dalam daftar tersebut.

## Menghitung Tingkat Akurasi Model Decision Tree

****

*Gambar 15.1 Menghitung Tingkat Akurasi Model dengan Decision Tree*

Dalam konteks ini, tingkat akurasi mengindikasikan seberapa baik model decision tree dapat memprediksi tingkat level kanker paru-paru pada data uji (test\_data). Dalam kasus ini, tingkat akurasi sebesar 75.33% menunjukkan bahwa model decision tree berhasil memprediksi dengan benar sekitar 75.33% dari keseluruhan data uji yang digunakan.

Penting untuk diingat bahwa tingkat akurasi hanya memberikan informasi tentang kinerja model pada data uji tertentu yang digunakan dalam percobaan. Tingkat akurasi yang tinggi tidak selalu menjamin performa yang baik pada data yang belum pernah dilihat sebelumnya. Evaluasi lebih lanjut, seperti penggunaan metrik evaluasi yang lebih lengkap dan validasi silang, mungkin diperlukan untuk mendapatkan gambaran yang lebih komprehensif tentang kinerja model.

# BAB V

# KESIMPULAN

Kesimpulan dari isi makalah yang kami tuangkan dalam laporan tugas akhir ini bermaksud untuk menambah ilmu pengetahuan dan wawasan bagi kami dan teman-teman sekalian.

Kesimpulan yang dapat ditarik dari projek ini adalah sebagai berikut:

1. Rumusan Masalah 1: Terdapat pengaruh signifikan faktor risiko genetik terhadap tingkat risiko kanker paru-paru. Namun, usia dan jenis kelamin tidak memiliki pengaruh signifikan. Algoritma linear regression merupakan pilihan yang cocok untuk menganalisis hubungan ini.

2. Rumusan Masalah 2: Penggunaan alkohol, pola makan seimbang, dan kegemukan memiliki pengaruh signifikan terhadap tingkat risiko kanker paru-paru. Namun, kebiasaan merokok tidak memiliki pengaruh yang signifikan. Algoritma linear regression lebih cocok untuk menganalisis hubungan ini.

3. Rumusan Masalah 3: Polusi udara dan bahaya pekerjaan memiliki pengaruh signifikan terhadap tingkat risiko kanker paru-paru. Algoritma linear regression dapat digunakan untuk menganalisis hubungan ini.

4. Rumusan Masalah 4: Algoritma linear regression memberikan tingkat akurasi yang lebih tinggi dalam memprediksi tingkat risiko kanker paru-paru berdasarkan gejala-gejala yang dialami pasien, dibandingkan dengan decision tree.

Berdasarkan kesimpulan di atas, dapat disimpulkan bahwa faktor risiko genetik, gaya hidup (seperti penggunaan alkohol, pola makan seimbang, dan kegemukan), faktor lingkungan (seperti polusi udara dan bahaya pekerjaan), serta gejala-gejala yang dialami pasien dapat berpengaruh terhadap tingkat risiko kanker paru-paru. Algoritma linear regression merupakan pilihan yang lebih cocok dalam menganalisis hubungan variabel-variabel tersebut.

# DAFTAR PUSTAKA

R Programming – Student Guide Part 1. (2016). Birmingham, Inggris: Packt Publishing Ltd.

Machine Learning with R – Student Guide. (2015). Birmingham, Inggris: Packt Publishing Ltd.

Machine Learning with R – Student Guide. (2013). Birmingham, Inggris: Packt Publishing Ltd.

RISHI DAMARLA, 2020 Kaggle. (n.d.). Cancer Patients Data. Retrieved from https://www.kaggle.com/datasets/rishidamarla/cancer-patients-data

Siswantini Suryandari, 2022. Pengidap Kanker Paru di Indonesia Didominasi Usia Muda. Retrieved from https://mediaindonesia.com/humaniora/470270/pengidap-kanker-paru-di-indonesia-didominasi-usia-muda

Wikipedia Indonesia. (n.d.). Kanker Paru-Paru. Retrieved from https://id.wikipedia.org/wiki/Kanker\_paru-paru

RStudio Team. (n.d.). What is RStudio? Retrieved from https://rstudio.com/about/

Quick-R. (n.d.). R Functions. Retrieved from https://www.statmethods.net/management/functions.html

DataCamp. (n.d.). Introduction to Data Analysis in R. Retrieved from https://www.datacamp.com/courses/data-analysis-in-r-introduction