

Presentation by

---

**Keiko & Delisa**

CCIT FTUI

ITE | 2023

# *Analisis Data Pasien Kanker Paru-Paru*

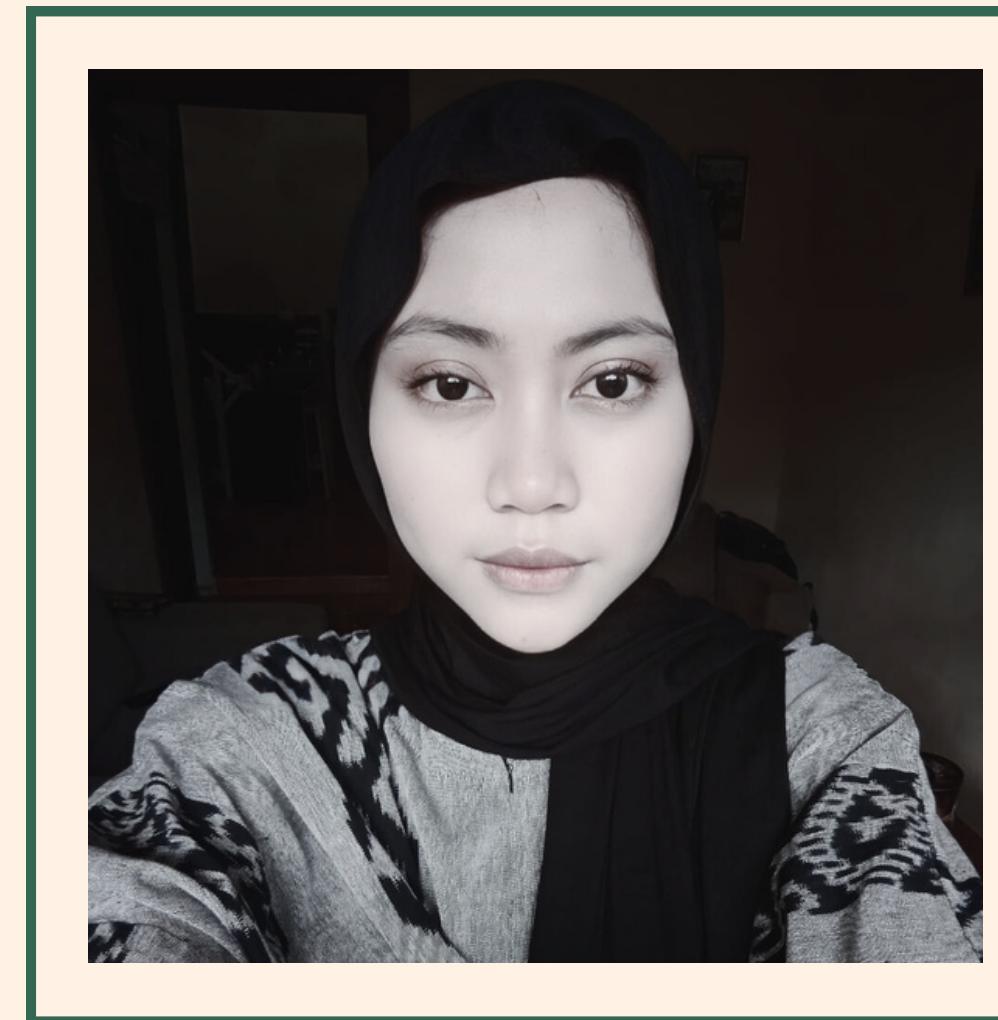


# Our Team

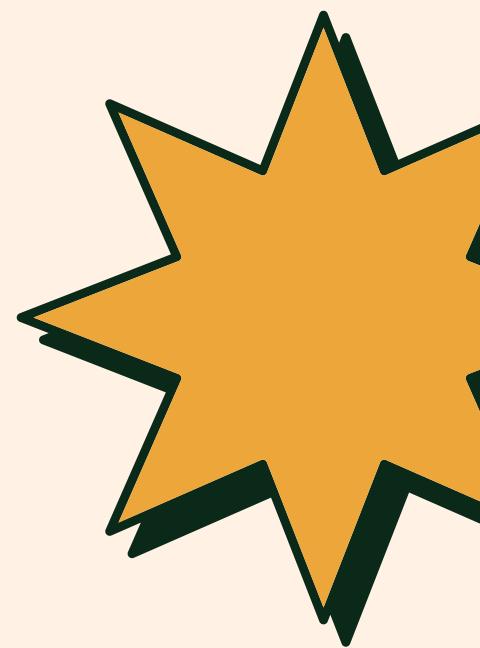
---



**Keiko Joceliandita**  
**2213020201**



**Delisa Febriana**  
**2213020195**



---

# *Table of Content*

---

**Latar Belakang**

**Rumusan Masalah**

**Penjelasan Data**

**Tujuan Penelitian**

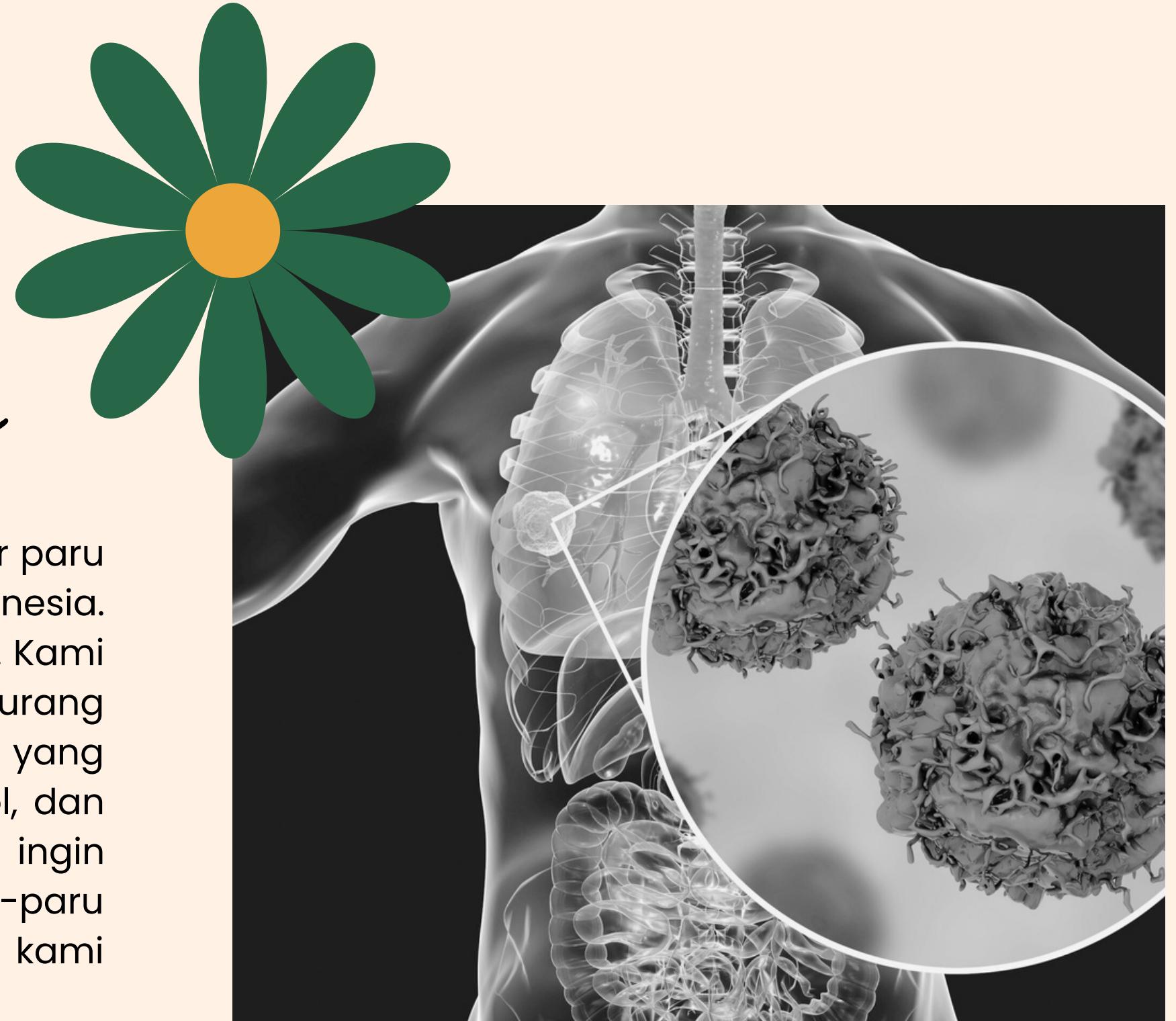
**Implementasi dan Hasil**

**Kesimpulan**



# *Latar Belakang*

Dikutip dari website Media Indonesia, jumlah kasus kanker paru di Indonesia semakin meningkat tiap tahunnya di Indonesia. Selain itu, usia penderita kanker paru pun semakin muda. Kami menyadari bahwa mungkin generasi muda sekarang kurang peduli dengan kesehatan diri karena menjalani lifestyle yang kurang sehat seperti kebiasaan merokok, minum alkohol, dan pola makan junk food berlebihan. Maka dari itu kami ingin mencari tahu faktor terbesar penyebab kanker paru-paru dengan data pasien kanker paru-paru yang sudah kami temukan dari website kaggle.



# Kanker Paru-Paru



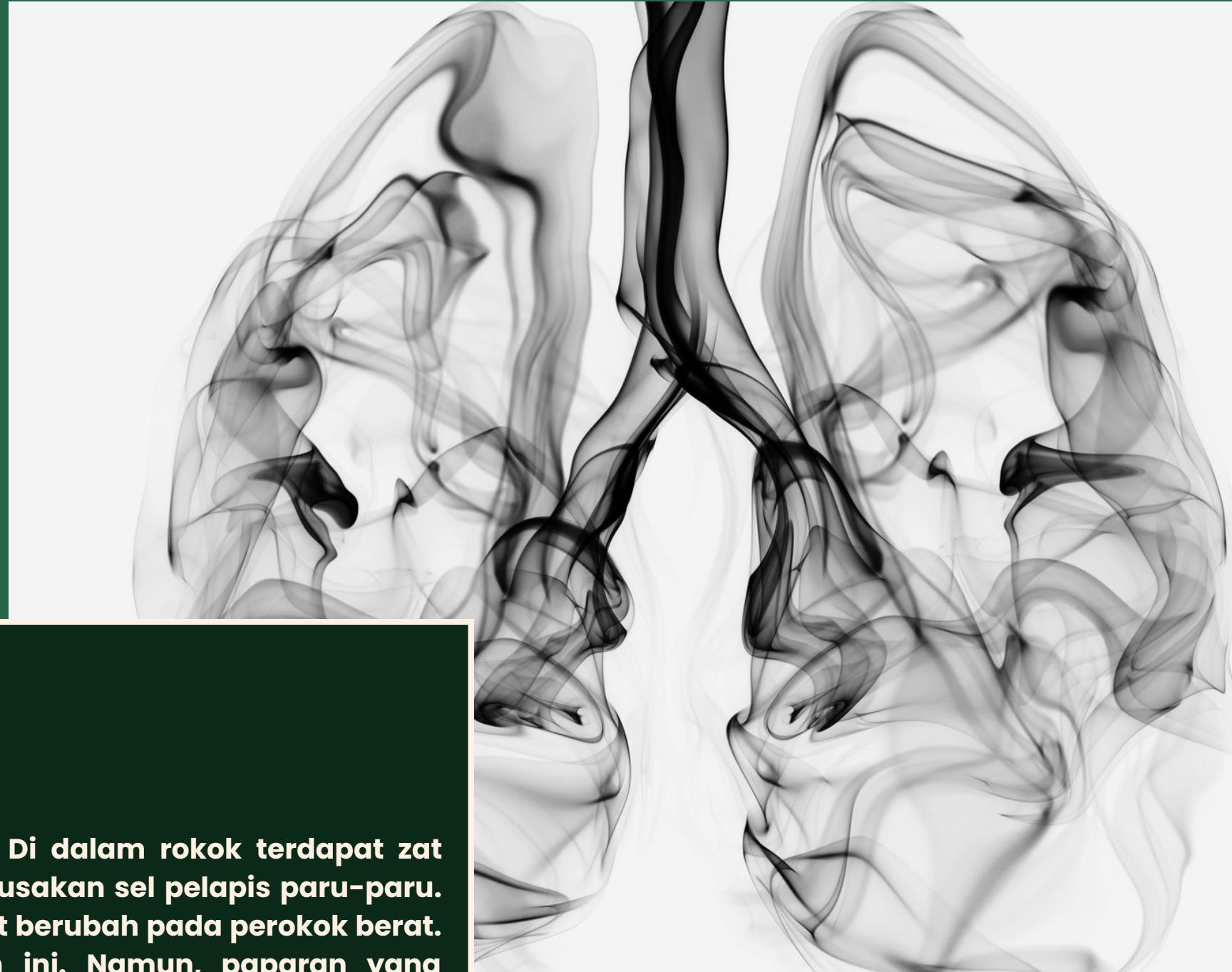
## Pengertian

Suatu kondisi di mana sel-sel tumbuh secara tidak terkendali di dalam organ paru-paru



## 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 dapat 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.

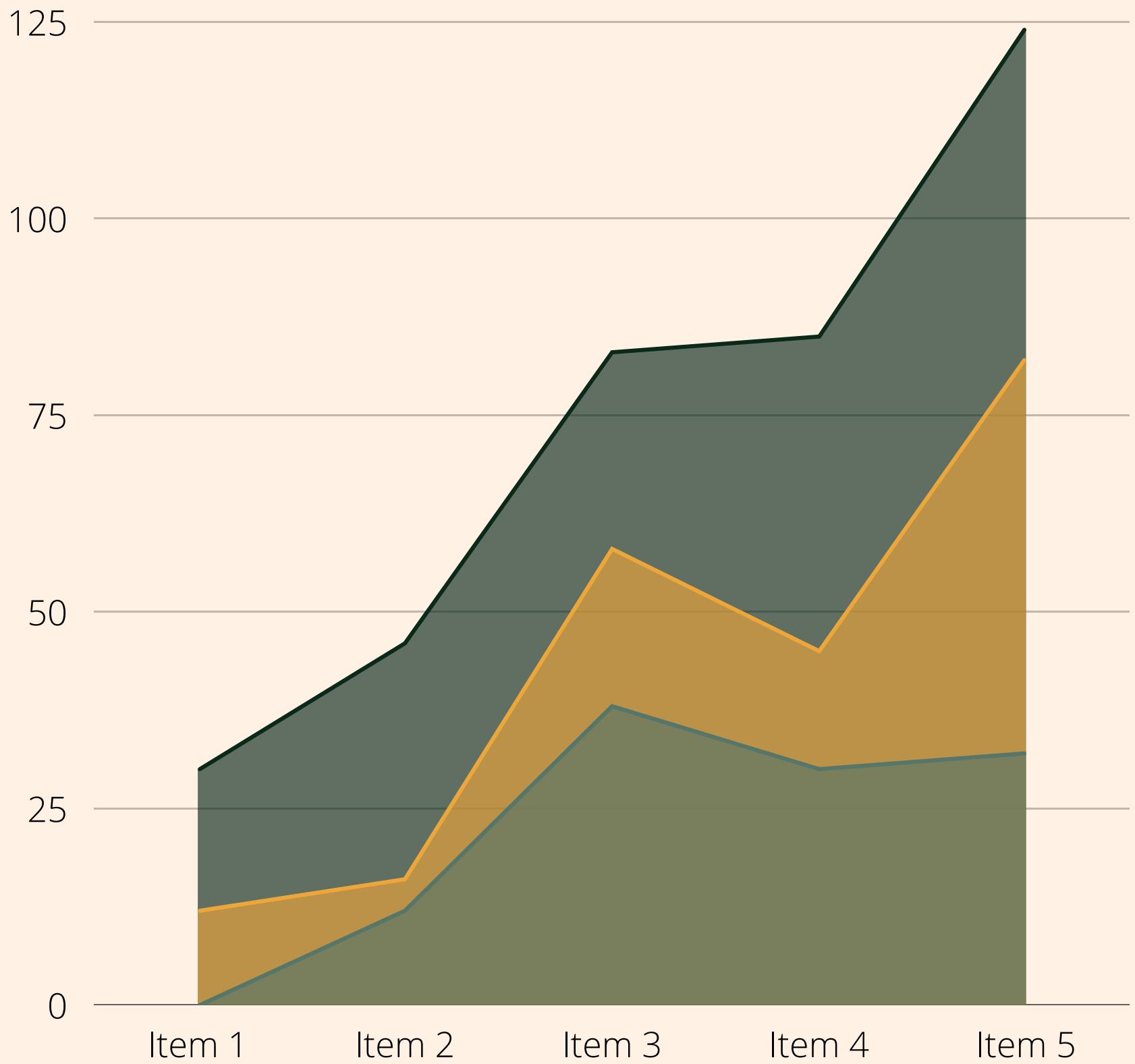


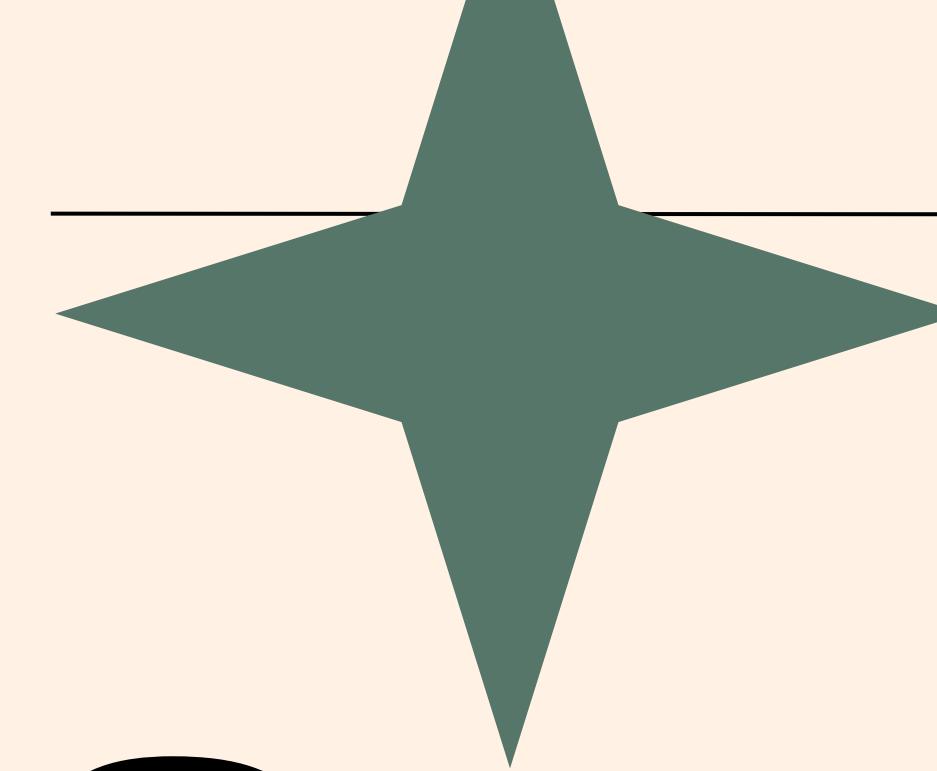
# Data Pasien Kanker Paru-Paru

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
2. Age
3. Gender
4. Air pollution
5. Alcohol use
6. Dust allergy
7. Occupational hazards
8. Genetic risk
9. Chronic lung disease
10. Balanced diet
11. Obesity
12. Smoking
13. Passive smoker
14. Chest pain
15. Coughing of blood
16. Fatigue
17. Weight loss
18. Shortness of breath
19. Wheezing
20. Swallowing difficulty
21. Clubbing of finger nails
22. Frequent cold
23. Dry cough
24. Snoring
25. Level

Kami mengkategorikan data tersebut menjadi 4 bagian berdasarkan faktor yang relevan dengan risiko kanker paru-paru, lingkungan, gaya hidup, dan gejala yang dialami pasien.





# Rumusan Masalah

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

# Tujuan Penelitian

01

Untuk menentukan apakah ada korelasi antara faktor-faktor seperti usia, jenis kelamin, dan risiko genetik dengan tingkat risiko kanker paru-paru pada pasien.

02

Untuk menentukan pengaruh gaya hidup, seperti merokok, konsumsi alkohol, pola makan seimbang, dan obesitas, terhadap tingkat risiko kanker paru-paru.

03

Untuk mengidentifikasi dampak faktor lingkungan, termasuk polusi udara, bahaya pekerjaan, terhadap tingkat risiko kanker paru-paru.

04

Untuk membangun model prediktif yang dapat memprediksi tingkat risiko kanker paru-paru berdasarkan gejala-gejala yang dialami pasien, serta mengevaluasi performa model tersebut.

# *Implementasi*

# Pemrosesan Data

```
//PEMROSESAN DATA
View(cancer_patient_data_sets)
str(cancer_patient_data_sets)
sapply(cancer_patient_data_sets, class)
cancer_patient_data_sets <- cancer_patient_data_sets[-1]
table(cancer_patient_data_sets$Level)
round(prop.table(table(cancer_patient_data_sets$Level)) * 100, digits = 1)
```

Kode untuk menganalisis struktur dataset, meriksa tipe data dari setiap variable, menghitung frekuensi dan proporsi relatif dari nilai-nilai dalam kolom Level

# *Visualisasi Data Histogram*

**Menampilkan distribusi frekuensi dari suatu variable numerik.**  
**Histogram berguna untuk memahami pola distribusi data dan melihat apakah terdapat kecondongan atau anomali.**

```
//VISUALISASI DATA
hist(cancer_patient_data_sets$Level)
hist(cancer_patient_data_sets$Age)
hist(cancer_patient_data_sets$Gender)
hist(cancer_patient_data_sets$'Air Pollution')
hist(cancer_patient_data_sets$'Alcohol use')
hist(cancer_patient_data_sets$'Dust Allergy')
hist(cancer_patient_data_sets$'Occupational Hazards')
hist(cancer_patient_data_sets$'Genetic Risk')
hist(cancer_patient_data_sets$'chronic Lung Disease')
hist(cancer_patient_data_sets$'Balanced Diet')
hist(cancer_patient_data_sets$Obesity)
hist(cancer_patient_data_sets$'Smoking')
hist(cancer_patient_data_sets$'Passive Smoker')
hist(cancer_patient_data_sets$'Chest Pain')
hist(cancer_patient_data_sets$'Coughing of Blood')
hist(cancer_patient_data_sets$'Fatigue')
hist(cancer_patient_data_sets$'Weight Loss')
hist(cancer_patient_data_sets$'Shortness of Breath')
hist(cancer_patient_data_sets$'Wheezing')
hist(cancer_patient_data_sets$'Swallowing Difficulty')
hist(cancer_patient_data_sets$'Clubbing of Finger Nails')
hist(cancer_patient_data_sets$'Frequent Cold')
hist(cancer_patient_data_sets$'Dry Cough')
hist(cancer_patient_data_sets$'Snoring')
```

# Normalisasi

```
//NORMALISASI DATA
normalize <- function(x) {
  return ((x - min(x)) / (max(x) - min(x)))
}
cancer_patient_data_sets_n <- as.data.frame(lapply(cancer_patient_data_sets[2:23], normalize))
View(cancer_patient_data_sets_n)
summary(cancer_patient_data_sets_n)
```

Normalisasi data membantu menghilangkan perbedaan skala dan memungkinkan perbandingan yang adil antara variabel. Sebetulnya data kami sudah oke jadi tidak perlu menggunakan normalisasi data, kami hanya mencoba untuk mengimplementasikan kode nya saja di RStudio.



# Korelasi

Fungsi korelasi digunakan untuk menentukan sejauh mana dua variabel bergerak bersama-sama atau saling terkait.

```
//RUMUSAN MASALAH 1 dengan CORRELATION
Faktor_Berhubungan <- cor(cancer_patient_data_sets[c("Age", "Gender", "Genetic Risk", "Level")])
Faktor_Berhubungan["Age", "Level"]
Faktor_Berhubungan["Gender", "Level"]
Faktor_Berhubungan["Genetic Risk", "Level"]
Faktor_Berhubungan_Gabungan <- mean(c(Faktor_Berhubungan["Age", "Level"], Faktor_Berhubungan["Gender", "Level"], Faktor_Berhubungan["Genetic Risk", "Level"]))
print (Faktor_Berhubungan_Gabungan)
```

Gambar 4.1 Rumusan Masalah 1 dengan Correlation

```
//RUMUSAN MASALAH 2 dengan CORRELATION
Faktor_Gaya_Hidup <- cor(cancer_patient_data_sets[c("Alcohol use", "Smoking", "Balanced Diet", "Obesity")])
Faktor_Gaya_Hidup["Alcohol use", "Level"]
Faktor_Gaya_Hidup["Smoking", "Level"]
Faktor_Gaya_Hidup["Balanced Diet", "Level"]
Faktor_Gaya_Hidup["Obesity", "Level"]
Faktor_Gaya_Hidup_Gabungan <- mean(c(Faktor_Gaya_Hidup["Alcohol use", "Level"], Faktor_Gaya_Hidup["Smoking", "Level"], Faktor_Gaya_Hidup["Balanced Diet", "Level"], Faktor_Gaya_Hidup["Obesity", "Level"]))
print (Faktor_Gaya_Hidup_Gabungan)
```

Gambar 4.2 Rumusan Masalah 2 dengan Correlation

```
//RUMUSAN MASALAH 3 dengan CORRELATION
Faktor_Lingkungan <- cor(cancer_patient_data_sets[c("Air Pollution", "Occupational Hazards", "Radiation Exposure", "Water Contamination")])
Faktor_Lingkungan["Air Pollution", "Level"]
Faktor_Lingkungan["Occupational Hazards", "Level"]
Faktor_Lingkungan_Gabungan <- mean(c(Faktor_Lingkungan["Air Pollution", "Level"], Faktor_Lingkungan["Occupational Hazards", "Level"], Faktor_Lingkungan["Radiation Exposure", "Level"], Faktor_Lingkungan["Water Contamination", "Level"]))
print (Faktor_Lingkungan_Gabungan)
```

Gambar 4.3 Rumusan Masalah 3 dengan Correlation

# Regression

```
//RUMUSAN MASALAH 1 dengan REGRESSION  
data <- cancer_patient_data_sets  
model <- lm(Level ~ Age + Gender + `Genetic Risk`, data = data)  
summary(model)
```

Gambar 5.1 Rumusan Masalah 1 dengan Regression

```
//RUMUSAN MASALAH 2 dengan REGRESSION  
data2 <- cancer_patient_data_sets  
model2 <- lm(Level ~ `Alcohol use` + Smoking + `Balanced Diet` + Obesity, data = data2)  
summary(model2)
```

Gambar 5.2 Rumusan Masalah 2 dengan Regression

```
//RUMUSAN MASALAH 3 dengan REGRESSION  
data3 <- cancer_patient_data_sets  
model3 <- lm(Level ~ `Air Pollution` + `Occupational Hazards`, data = data3)  
summary(model3)
```

Gambar 5.3 Rumusan Masalah 3 dengan Regression

Memodelkan hubungan antara satu atau lebih variabel independen (variabel prediktor) dengan satu variabel dependen.

# Decision Tree

Menggambarkan serangkaian keputusan atau kondisi yang membantu dalam mengklasifikasikan atau memprediksi nilai variabel target



```
//RUMUSAN MASALAH 1 dengan DESICION TREE
library(rpart)
data <- cancer_patient_data_sets
model_tree <- rpart(Level ~ Age + Gender, data = data)
summary(model_tree)
plot(model_tree)
text(model_tree)
new_data <- data.frame(Age = 44, Gender = 1)
prediction <- predict(model_tree, newdata = new_data)
print(prediction)
```

Gambar 6.1 Rumusan Masalah 1 dengan Decision Tree

```
//RUMUSAN MASALAH 2 dengan DESICION TREE
data2 <- cancer_patient_data_sets
model_tree2 <- rpart(Level ~ `Alcohol use` + Smoking + `Balanced Diet` + Obesity, data = data2)
summary(model_tree2)
plot(model_tree2)
text(model_tree2)
new_data2 <- data.frame(`Alcohol use` = 4, Smoking = 3, `Balanced Diet` = 2, Obesity = 4)
prediction2 <- predict(model_tree2, newdata2 = new_data2)
print(prediction2)
```

Gambar 6.2 Rumusan Masalah 2 dengan Decision Tree

```
//RUMUSAN MASALAH 3 dengan DESICION TREE
data3 <- cancer_patient_data_sets
model_tree3 <- rpart(Level ~ `Air Pollution` + `OccuPational Hazards` , data = data3)
summary(model_tree3)
plot(model_tree3)
text(model_tree3)
new_data3 <- data.frame(`Air Pollution` = 2, `OccuPational Hazards` = 4)
prediction3 <- predict(model_tree3, newdata3 = new_data3)
print(prediction3)
```

Gambar 6.3 Rumusan Masalah 3 dengan Decision Tree

# *Analisis Prediksi Menggunakan Linear Regression*

```
//RUMUSAN MASALAH 4
set.seed(123) # Untuk hasil yang dapat direplikasi
train_index <- sample(1:nrow(data), nrow(data)*0.7) # Mengambil 70% data sebagai data train
train_data <- data[train_index, ]
test_data <- data[-train_index, ]
rownames(test_data) <- NULL
predictions_train <- predict(model, newx = X_train)
print(predictions_train)
print(predictions_test)
model4 <- lm(Level ~ Wheezing + `Dry Cough` + `Clubbing of Finger Nails` + `Dust Allergy` +
    `chronic Lung Disease` + `Chest Pain` + `Coughing of Blood` + Fatigue +
    `Weight Loss` + `Shortness of Breath` + `Swallowing Difficulty` + `Frequent Cough` +
    Snoring, data = train_data)
predictions_train <- predict(model4, newdata = train_data)
predictions_test <- predict(model4, newdata = test_data)
print(predictions_train)
print(predictions_test)
summary(predictions_train)
summary(predictions_test)
summary(cancer_patient_data_sets$Level)
```

*Gambar 7.1 Rumusan Masalah 4 dengan Linear Regression*

# Menghitung Tingkat Akurasi Model dengan Linear Regression

```
# Menghitung tingkat akurasi untuk data pelatihan  
actual_train <- train_data$Level  
accuracy_train <- sum(round(predictions_train) == actual_train) / length(actual_train) * 100  
  
# Menghitung tingkat akurasi untuk data uji  
actual_test <- test_data$Level  
accuracy_test <- sum(round(predictions_test) == actual_test) / length(actual_test) * 100  
  
# Menampilkan tingkat akurasi  
print(paste("Tingkat Akurasi untuk Data Pelatihan:", accuracy_train, "%"))  
print(paste("Tingkat Akurasi untuk Data Uji:", accuracy_test, "%"))
```

Gambar 8.1 Menghitung Tingkat Akurasi Model dengan Linier Regression

# Evaluasi Performa Model dengan Linear Regression

```
//EVALUASI PERFORMA  
train_mse <- mean((train_data$Level - predictions_train)^2)  
test_mse <- mean((test_data$Level - predictions_test)^2)  
train_r2 <- summary(model4)$r.squared  
test_r2 <- summary(model4, newdata = test_data)$r.squared  
print(train_mse)  
print(test_mse)  
print(train_r2)  
print(test_r2)  
predictions_test <- predict(model4, newdata = test_data)  
actual_test <- test_data$Level  
r_squared_test <- 1 - sum((actual_test - predictions_test)^2) / sum((actual_test - mean(act  
print(r_squared_test))
```

Gambar 9.1 Menghitung Tingkat Akurasi Model dengan Linier Regression

# Analisis Prediksi Menggunakan Decision Tree

```
//RUMUSAN MASALAH 4 DENGAN DESICION TREE
data4 <- cancer_patient_data_sets
model_tree4 <- rpart(Level ~ Wheezing + `Dry Cough` + `Clubbing of Finger Nails` + `Dust Allergy` + `chronic Lung Disease` + `Chest Pain` + `Coughing of Blood` + Fatigue + `Weight Loss` + `Shortness of Breath` + `Swallowing Difficulty` + `Frequent Snoring , data = data4)
summary(model_tree4)
//Mengatur Margin Plot Desicion Tree
par(mar = c(5, 3, 3, 3))
plot(model_tree4, main = "Decision Tree", margin = 0.2, uniform = TRUE)
text(model_tree4, use.n = TRUE, all = TRUE, cex = 0.7)
new_data4 <- data.frame(wheezing = 2, `Dry Cough` = 3, `Clubbing of Finger Nails` = 1, `Dust Allergy` = 0, `chronic Lung Disease` = 0, `Chest Pain` = 0, `Coughing of Blood` = 0, Fatigue = 0, `Weight Loss` = 0, `Shortness of Breath` = 0, `Swallowing Difficulty` = 0, `Frequent Snoring` = 0, data = new_data4)
prediction4 <- predict(model_tree, newdata4 = new_data4)
print(prediction4)
summary(prediction4)
```

Gambar 10.1 Rumusan Masalah 4 dengan Decision Tree

```
# Memisahkan data menjadi data pelatihan dan data uji
set.seed(123)
train_indices <- sample(1:nrow(data4), 0.7 * nrow(data4))
train_data <- data4[train_indices, ]
test_data <- data4[-train_indices, ]
# Membangun model decision tree dengan data pelatihan
model_tree4 <- rpart(Level ~ Wheezing + `Dry Cough` + `Clubbing of Finger Nails` + `Dust Allergy` + `chronic Lung Disease` + `Chest Pain` + `Coughing of Blood` + Fatigue + `Weight Loss` + `Shortness of Breath` + `Swallowing Difficulty` + `Frequent Snoring , data = train_data)
# Memprediksi tingkat level kanker paru-paru dengan data uji
predictions <- predict(model_tree4, newdata = test_data)
print(predictions)
```

Gambar 10.2 Rumusan Masalah 4 dengan Decision Tree

```
# Menghitung tingkat akurasi model
actual <- test_data$Level
accuracy <- sum(predictions == actual) / length(actual) * 100
# Menampilkan tingkat akurasi
print(paste("Tingkat Akurasi Model Decision Tree:", accuracy, "%"))
```

Gambar 11.1 Menghitung Tingkat Akurasi Model dengan Decision Tree

*Menghitung Tingkat Akurasi Model dengan Decision Tree*

# Hasil

# Rumusan Masalah Pertama

## Dengan Correlation

```
> Faktor_Berhubungan <- cor(cancer_patient_data_sets[c("Age", "Gender", "Genetic Risk", "Level")])
> Faktor_Berhubungan["Age", "Level"]
[1] 0.06004781
> Faktor_Berhubungan["Gender", "Level"]
[1] -0.1649852
> Faktor_Berhubungan["Genetic Risk", "Level"]
[1] 0.7013027
> Faktor_Berhubungan_Gabungan <- mean(c(Faktor_Berhubungan["Age", "Level"], Faktor_Berhubungan["Gender", "Level"], Faktor_Berhubungan["Genetic Risk", "Level"]))
> print (Faktor_Berhubungan_Gabungan)
[1] 0.1987885
```

## Dengan Linear Regression

```
> data <- cancer_patient_data_sets
> model <- lm(Level ~ Age + Gender + `Genetic Risk`, data = data)
> summary(model)

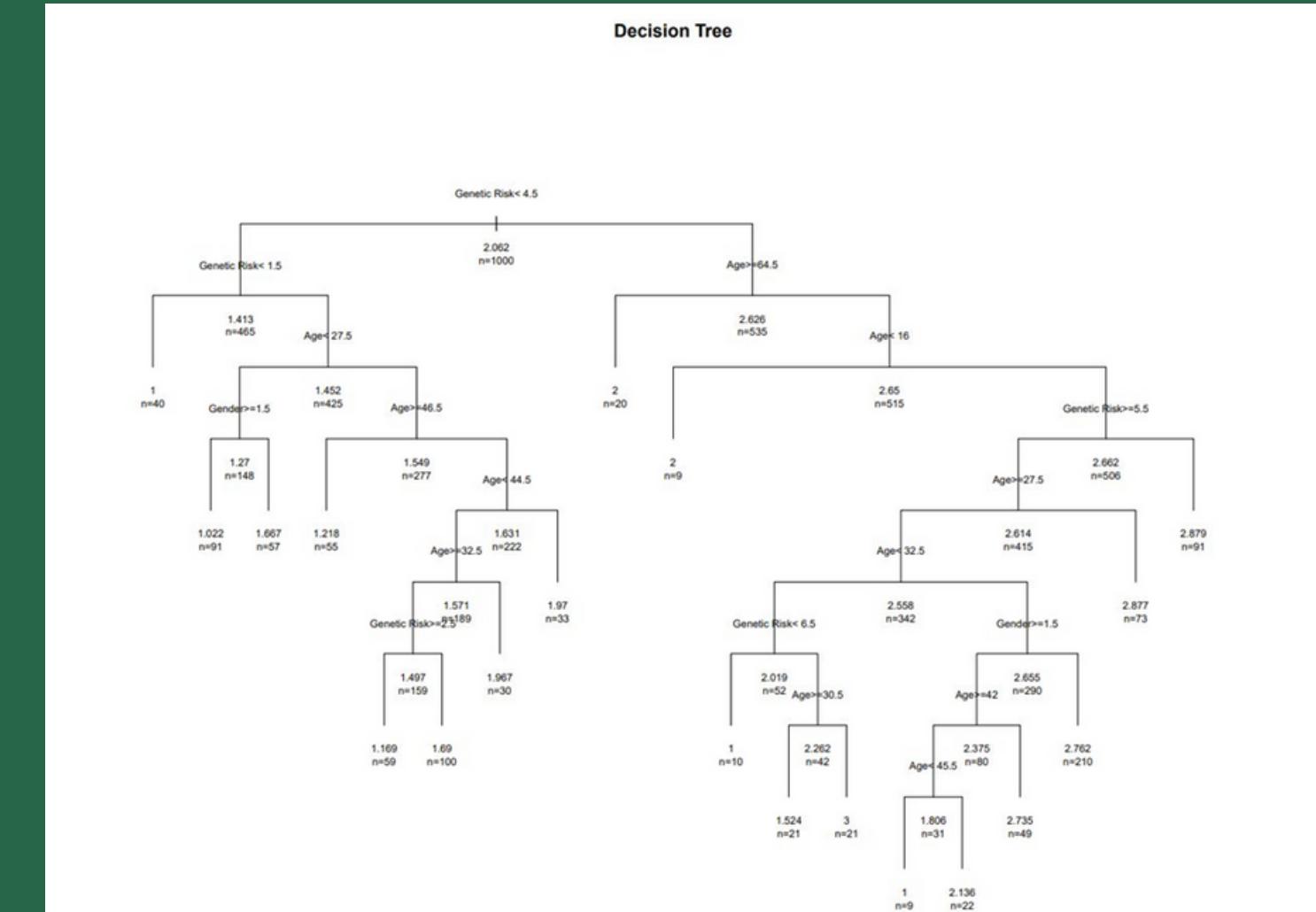
Call:
lm(formula = Level ~ Age + Gender + `Genetic Risk`, data = data)

Residuals:
    Min      1Q  Median      3Q     Max 
-1.7131 -0.4656  0.2799  0.3722  0.8393 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept) 0.8343999  0.1022252   8.162 9.86e-16 ***
Age          0.0004982  0.0015669   0.318   0.751    
Gender       -0.0129821  0.0392300  -0.331   0.741    
`Genetic Risk` 0.2679654  0.0088847  30.160 < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.582 on 996 degrees of freedom
Multiple R-squared:  0.492,    Adjusted R-squared:  0.4904 
F-statistic: 321.5 on 3 and 996 DF,  p-value: < 2.2e-16
```

## Dengan Desicion Tree



> print(prediction)

1 2 3  
1.169492 1.666667 2.879121

# Rumusan Masalah Kedua

## Dengan Correlation

```
> Faktor_Gaya_Hidup <- cor(cancer_patient_data_sets[c("Alcohol use", "Smoking", "Balanced Diet", "Obesity", "Level")])
> Faktor_Gaya_Hidup["Alcohol use", "Level"]
[1] 0.7187103
> Faktor_Gaya_Hidup["Smoking", "Level"]
[1] 0.5195301
> Faktor_Gaya_Hidup["Balanced Diet", "Level"]
[1] 0.706273
> Faktor_Gaya_Hidup["Obesity", "Level"]
[1] 0.8274351
> Faktor_Gaya_Hidup_Gabungan <- mean(c(Faktor_Gaya_Hidup["Alcohol use", "Level"], Faktor_Gaya_Hidup["Smoking", "Level"], Faktor_Gaya_Hidup["Balanced Diet", "Level"], Faktor_Gaya_Hidup["Obesity", "Level"]))
> print (Faktor_Gaya_Hidup_Gabungan)
[1] 0.6929871
```

## Dengan Linear Regression

```
> data2 <- cancer_patient_data_sets
> model2 <- lm(Level ~ `Alcohol use` + Smoking + `Balanced Diet` + Obesity, data = data2)
> summary(model2)

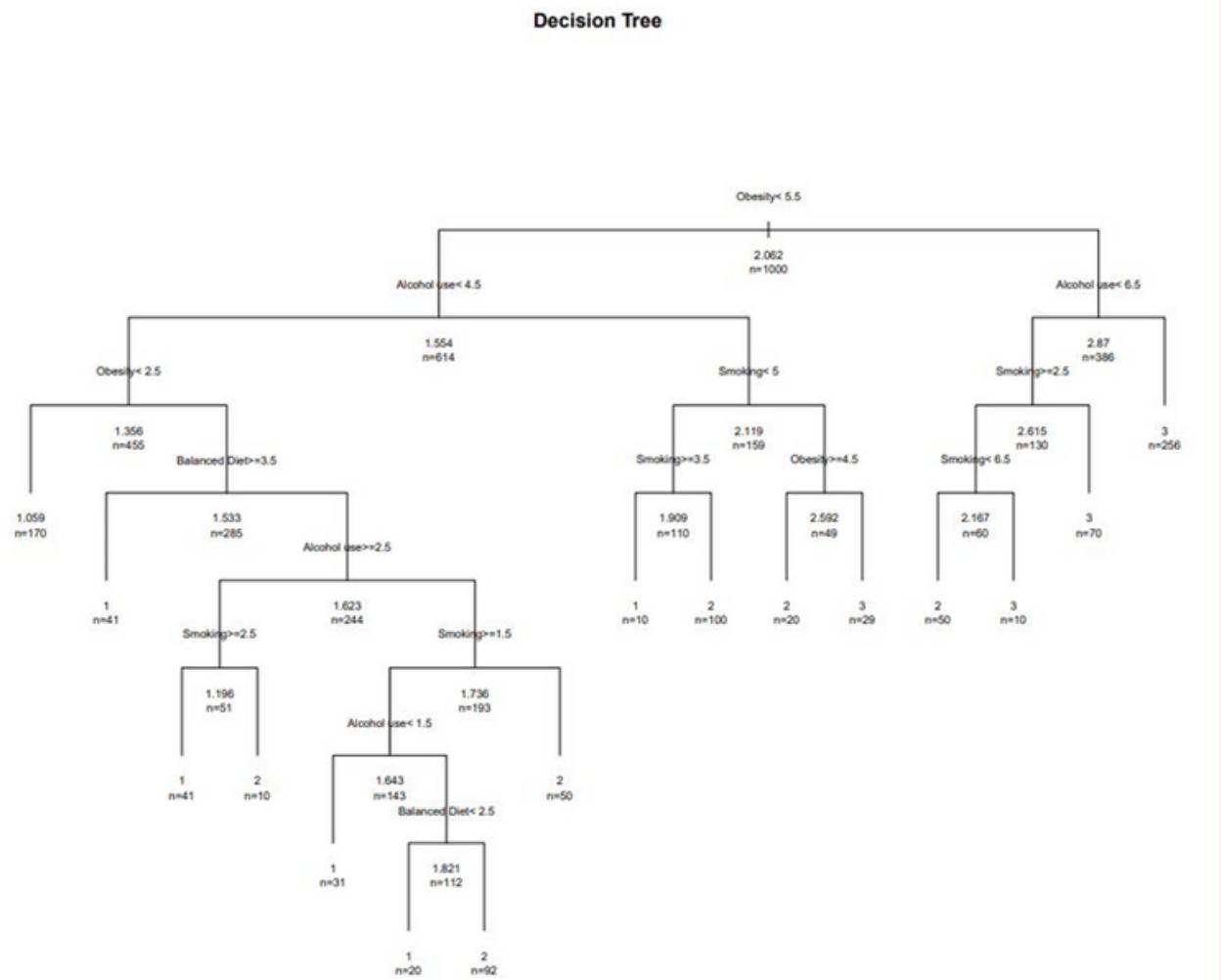
Call:
lm(formula = Level ~ `Alcohol use` + Smoking + `Balanced Diet` +
    Obesity, data = data2)

Residuals:
    Min      1Q      Median      3Q      Max 
-0.85492 -0.25752 -0.02035  0.30458  0.87988 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept) 0.495725  0.032652 15.182 < 2e-16 ***
`Alcohol use` 0.075387  0.007307 10.317 < 2e-16 ***
Smoking     0.009190  0.007021  1.309   0.191    
`Balanced Diet` 0.052819  0.010143  5.208 2.32e-07 ***
Obesity     0.212495  0.009431 22.531 < 2e-16 ***  
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.4128 on 995 degrees of freedom
Multiple R-squared:  0.7447,    Adjusted R-squared:  0.7436 
F-statistic: 725.5 on 4 and 995 DF,  p-value: < 2.2e-16
```

## Dengan Desicion Tree



```
> print(prediction2)
```

	1	2	3
1.169492	1.666667	2.879121	

# Rumusan Masalah Ketiga

## Dengan Correlation

```
> Faktor_Lingkungan <- cor(cancer_patient_data_sets[c("Air Pollution", "OccuPational Hazards", "Level")])
> Faktor_Lingkungan["Air Pollution", "Level"]
[1] 0.6360385
> Faktor_Lingkungan["OccuPational Hazards", "Level"]
[1] 0.6732549
> Faktor_Lingkungan_Gabungan <- mean(c(Faktor_Lingkungan["Air Pollution", "Level"], Faktor_Lingkungan["OccuPational Hazards", "Level"]))
> print (Faktor_Lingkungan_Gabungan)
[1] 0.6546467
```

## Dengan Linear Regression

```
> data3 <- cancer_patient_data_sets
> model3 <- lm(Level ~ `Air Pollution` + `OccuPational Hazards`, data = data3)
> summary(model3)

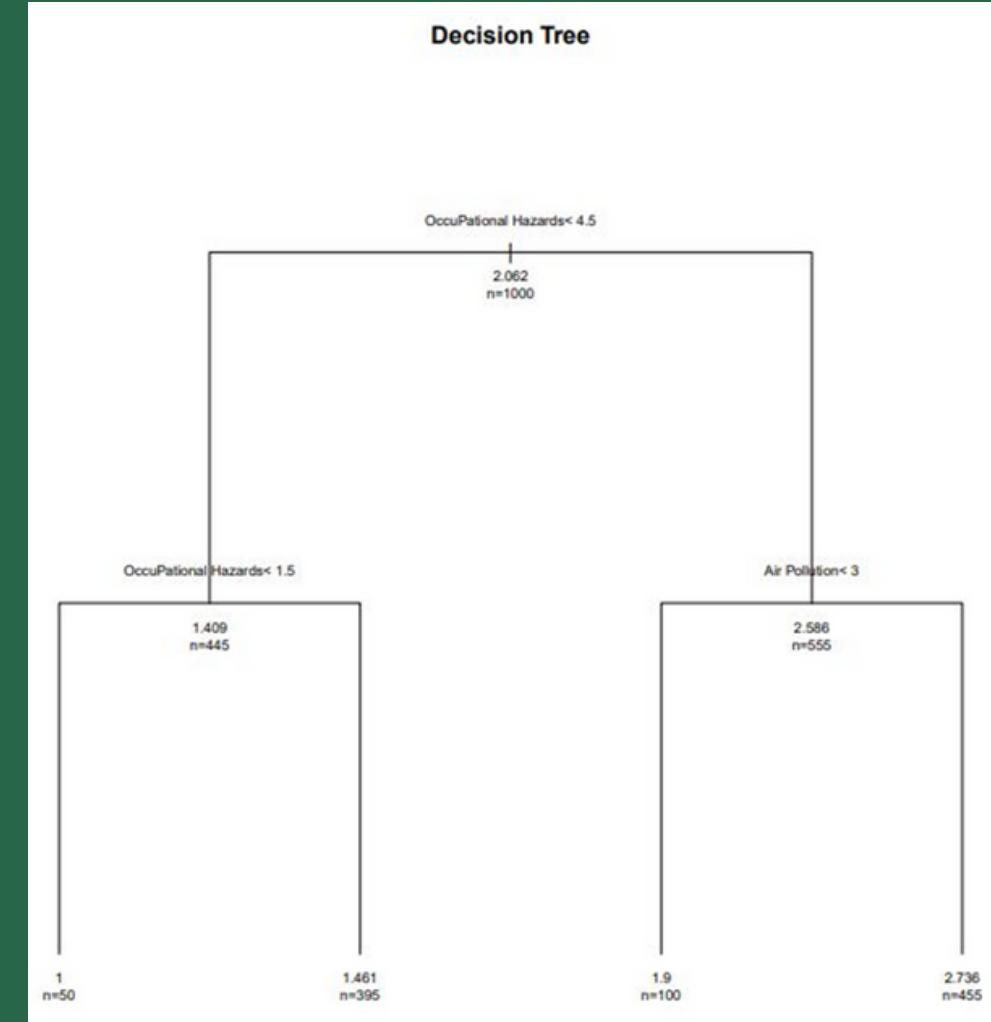
Call:
lm(formula = Level ~ `Air Pollution` + `OccuPational Hazards`,
    data = data3)

Residuals:
    Min      1Q      Median      3Q      Max 
-1.7534 -0.4730   0.2466   0.3827   1.5270 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept)  0.65705   0.04546 14.45   <2e-16 ***
`Air Pollution` 0.14429   0.01094 13.19   <2e-16 ***
`OccuPational Hazards` 0.17580   0.01054 16.68   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.5569 on 997 degrees of freedom
Multiple R-squared:  0.5345,    Adjusted R-squared:  0.5336 
F-statistic: 572.4 on 2 and 997 DF,  p-value: < 2.2e-16
```

## Dengan Desicion Tree



```
> print(prediction3)
```

	1	2	3
1.169492	1.666667	2.879121	

# Rumusan Masalah Keempat

## Prediksi dengan Linear Regression

```
> print(predictions_train)
      1       2       3
1.9658355 3.4973422 2.1558647

> print(predictions_test)
      1       2       3
1.4530391 2.9418259 3.0801412
```

```
> summary(predictions_train)
   Min. 1st Qu. Median Mean 3rd Qu. Max.
0.6346 1.4172 2.0157 2.0500 2.7688 3.4973
> summary(predictions_test)
   Min. 1st Qu. Median Mean 3rd Qu. Max.
0.6346 1.5480 1.9909 2.1002 2.7804 3.4973
> summary(cancer_patient_data_sets$Level)
   Min. 1st Qu. Median Mean 3rd Qu. Max.
1.000 1.000 2.000 2.062 3.000 3.000
```

## Tingkat Akurasi Model

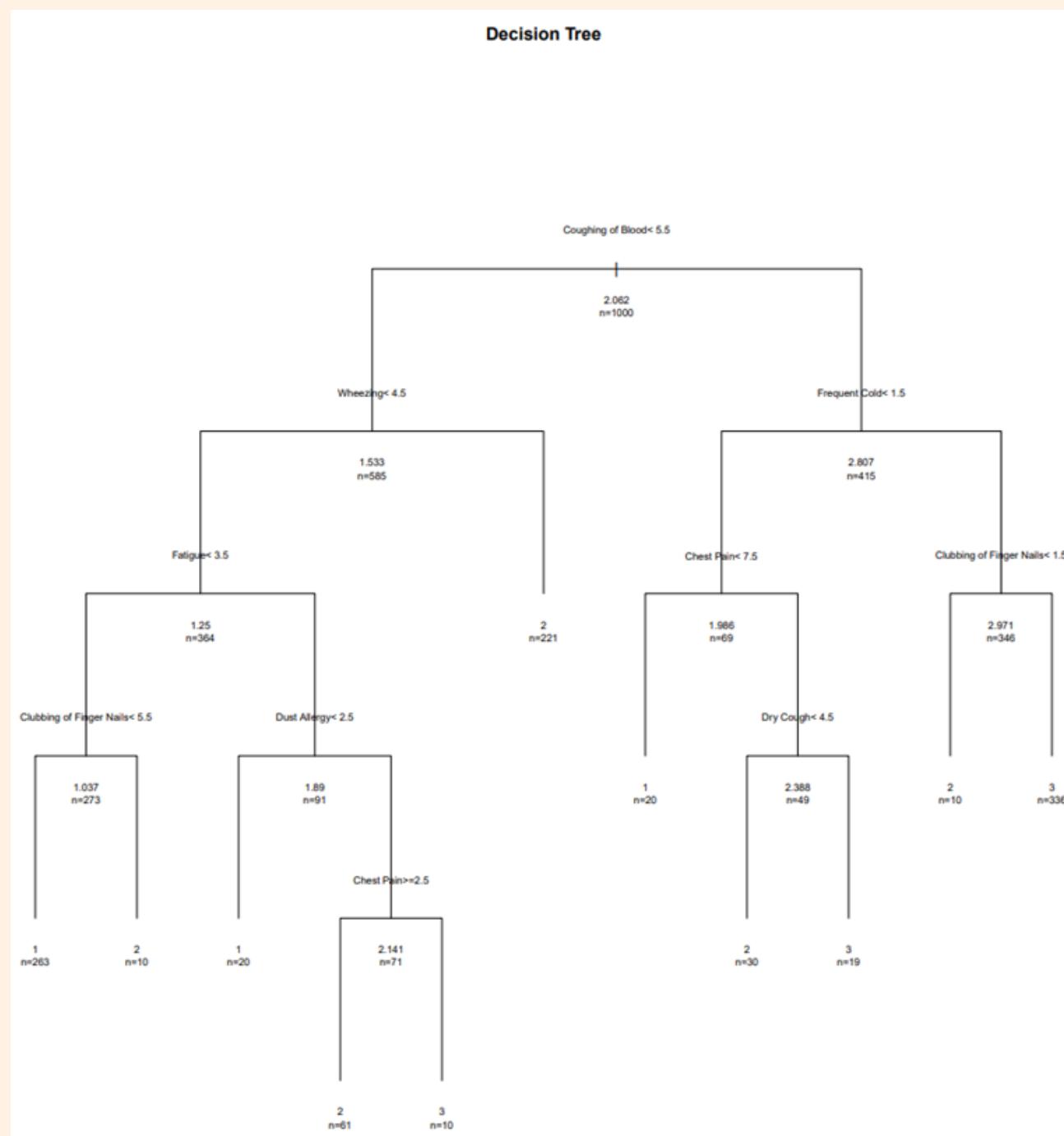
```
> actual_train <- train_data$Level
> accuracy_train <- sum(round(predictions_train) == actual_train) / length(actual_train) * 100
> actual_test <- test_data$Level
> accuracy_test <- sum(round(predictions_test) == actual_test) / length(actual_test) * 100
> print(paste("Tingkat Akurasi untuk Data Pelatihan:", accuracy_train, "%"))
[1] "Tingkat Akurasi untuk Data Pelatihan: 96.7142857142857 %"
> print(paste("Tingkat Akurasi untuk Data Uji:", accuracy_test, "%"))
[1] "Tingkat Akurasi untuk Data Uji: 97.6666666666667 %"
```

## Evaluasi Performa Model

```
> train_mse <- mean((train_data$Level - predictions_train)^2)
> test_mse <- mean((test_data$Level - predictions_test)^2)
> train_r2 <- summary(model4)$r.squared
> test_r2 <- summary(model4, newdata = test_data)$r.squared
> print(train_mse)
[1] 0.0792177
> print(test_mse)
[1] 0.07717894
> print(train_r2)
[1] 0.8842934
> print(test_r2)
[1] 0.8842934
> predictions_test <- predict(model4, newdata = test_data)
> actual_test <- test_data$Level
> r_squared_test <- 1 - sum((actual_test - predictions_test)^2) / sum((actual_test - mean(actual_test))^2)
> print(r_squared_test)
[1] 0.8745534
```

# Rumusan Masalah Keempat

## Prediksi dengan Desicion Tree



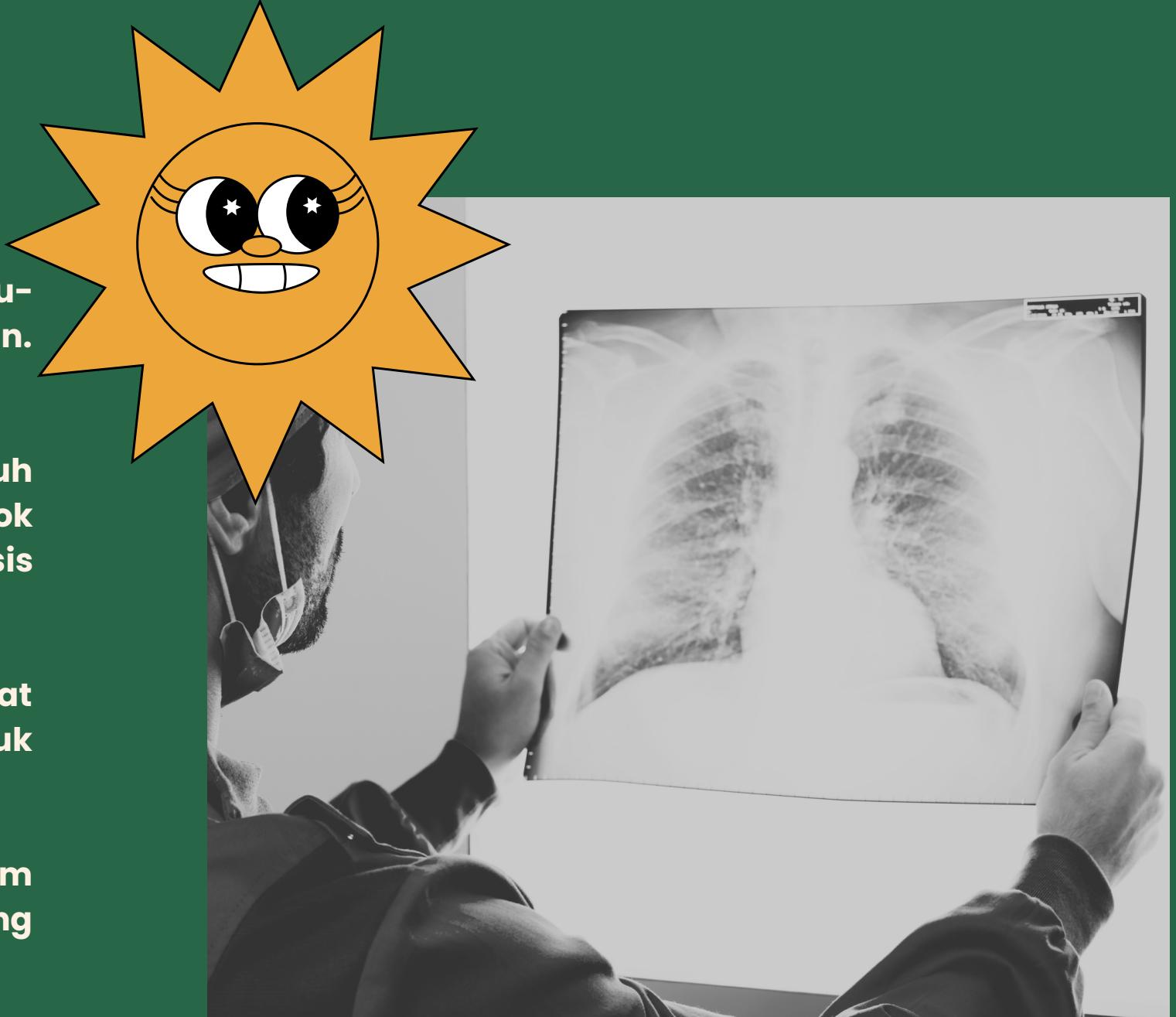
```
> print(predictions)
      1          2          3
1.025126 3.000000 3.000000
```

## Tingkat Akurasi Model

```
> print(paste("Tingkat Akurasi Model Decision Tree:", accuracy, "%"))
[1] "Tingkat Akurasi Model Decision Tree: 75.333333333333 %"
```

# Kesimpulan

- Faktor risiko genetik berpengaruh signifikan terhadap tingkat risiko kanker paru-paru, sedangkan usia dan jenis kelamin tidak memiliki pengaruh yang signifikan. Algoritma linear regression digunakan untuk menganalisis hubungan ini.
- Penggunaan alkohol, pola makan seimbang, dan kegemukan memiliki pengaruh signifikan terhadap tingkat risiko kanker paru-paru, sedangkan kebiasaan merokok tidak signifikan. Algoritma linear regression lebih cocok untuk menganalisis hubungan ini.
- Polusi udara dan bahaya pekerjaan memiliki pengaruh signifikan terhadap tingkat risiko kanker paru-paru. Algoritma linear regression dapat digunakan untuk menganalisis hubungan ini.
- 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, faktor lingkungan, dan gejala-gejala pasien berpengaruh terhadap risiko kanker paru-paru. Algoritma linear regression lebih cocok dalam menganalisis hubungan variabel-variabel tersebut.



Presentation by

---

**Estelle Darcy**

Timmerman University

Biology I 2023

*Thank  
You!*

