

# BAB I

## PENDAHULUAN

### 1.1 LATAR BELAKANG

Kanker paru-paru merupakan suatu keadaan dimana terdapat sel kanker yang berkembang secara tidak terkontrol di dalam organ paru-paru. Ada berbagai faktor yang menyebabkan seseorang mengidap penyakit kanker paru-paru, tetapi sebab utamanya adalah merokok. Kanker paru-paru juga menjadi sebab utama dari kematian akibat kanker pada pria (Siloam Hospitals, 2022). Menurut WHO (*World Health Organization*), tingkat kematian dari kanker paru-paru mencapai 30.843 atau sekitar 13,2% dari total kematian akibat penyakit kanker, dan ini membuat kanker paru-paru menjadi kanker dengan tingkat kematian tertinggi dari total kematian akibat penyakit kanker (Rizaty, 2022). Gejala pada kanker paru-paru umumnya tidak tampak di stadium awal, dan baru dapat terdeteksi ketika memasuki stadium lanjut, sehingga deteksi dini terhadap penyakit kanker paru-paru diperlukan agar peluang sembuh semakin besar (Azizah, 2022). Oleh karena itu, agar dapat melakukan deteksi dini terhadap penyakit kanker paru-paru adalah dengan mengetahui apakah seseorang memiliki risiko untuk terkena penyakit kanker paru-paru.

Solusi yang dapat diterapkan agar dapat mengetahui apakah seseorang memiliki risiko untuk terkena kanker paru-paru adalah dengan memanfaatkan algoritma dari *machine learning*, khususnya dengan metode klasifikasi. Algoritma untuk metode klasifikasi yang sering untuk digunakan adalah *Naïve Bayes*,

*Decision Tree, Support Vector Machines, Random Forest, dan K-Nearest Neighbor (KNN)* (Wibisono & Fahrurrozi, 2019). Ada beberapa penelitian yang telah dilakukan dengan menggunakan algoritma dari *machine learning* untuk melakukan klasifikasi terhadap suatu penyakit. Ada penelitian yang melakukan perbandingan terhadap metode *K-Nearest Neighbor* dan *Naïve Bayes* untuk melakukan klasifikasi terhadap penyakit jantung dengan tingkat akurasi sebesar 67% pada *K-Nearest Neighbor* dengan menggunakan pengukuran jarak *Manhattan*, tingkat akurasi sebesar 65% pada *K-Nearest Neighbor* dengan menggunakan pengukuran jarak *Euclidean*, dan tingkat akurasi sebesar 58% pada *Naïve Bayes* (Sahar, 2020). Ada penelitian lainnya yang membandingkan algoritma *Naïve Bayes* dan *K-Nearest Neighbor* untuk melakukan klasifikasi terhadap penyakit Hepatitis C Virus, dengan hasil tingkat akurasi sebesar 90,98% pada algoritma *Naïve Bayes*, dan tingkat akurasi sebesar 91,80% pada algoritma *K-Nearest Neighbor* (Raharja et al., 2021).

Penelitian ini akan menggunakan algoritma *Random Forest* dan *K-Nearest Neighbor*. Penggunaan kedua algoritma dipilih karena kedua algoritma ini termasuk dalam algoritma dengan *supervised learning*. Algoritma dengan pendekatan *supervised learning* sesuai untuk digunakan dalam melakukan klasifikasi dan regresi (A.Vincent, 2022). Kelebihan dari algoritma *K-Nearest Neighbor* adalah algoritma ini dapat bekerja secara maksimal terhadap data yang memiliki kesalahan acak (*data noise*) (Setiawan & Triayudi, 2022). *Random Forest* memiliki kelebihan yaitu mempunyai hasil klasifikasi yang baik, hasil *error* yang lebih rendah, dapat menangani *data training* dengan jumlah yang besar secara efisien, serta efektif dalam mengatasi masalah dalam *missing data* (Lestari & Sirodj, 2022).

Oleh karena itu, tujuan dari penelitian ini adalah untuk melakukan perbandingan antara algoritma *Random Forest* dan *K-Nearest Neighbor*, dengan tujuan untuk mengetahui tingkat akurasi kedua algoritma tersebut dalam melakukan klasifikasi terhadap risiko seseorang dalam terkena penyakit kanker paru-paru, serta untuk mengetahui algoritma manakah yang lebih baik dalam melakukan klasifikasi terhadap risiko seseorang dalam terkena penyakit kanker paru-paru.

## **1.2 RUMUSAN MASALAH**

Berdasarkan latar belakang tersebut, ada beberapa rumusan masalah yang ada di dalam penelitian ini, yaitu sebagai berikut:

- 1) Bagaimana cara melakukan klasifikasi terhadap risiko seseorang dalam terkena penyakit kanker paru-paru dengan menggunakan algoritma *K-Nearest Neighbor* (KNN) dan *Random Forest*?
- 2) Bagaimana hasil perbandingan tingkat akurasi dari klasifikasi terhadap risiko seseorang dalam terkena penyakit kanker paru-paru dengan menggunakan algoritma *K-Nearest Neighbor* (KNN) dan *Random Forest*?

## **1.3 TUJUAN PENELITIAN**

Berdasarkan rumusan masalah tersebut, ada beberapa tujuan penelitian dari penelitian ini, yaitu sebagai berikut:

- 1) Mengetahui cara melakukan klasifikasi terhadap risiko seseorang dalam terkena penyakit kanker paru-paru dengan menggunakan algoritma *K-Nearest Neighbor* (KNN) dan *Random Forest*.

- 2) Mengetahui perbandingan tingkat akurasi dari klasifikasi terhadap risiko seseorang dalam terkena penyakit kanker paru-paru dengan menggunakan algoritma *K-Nearest Neighbor* (KNN) dan *Random Forest*.

#### **1.4 BATASAN MASALAH**

Adapun batasan masalah yang ada di dalam penelitian ini, yaitu sebagai berikut:

- 1) Penggunaan algoritma *machine learning* pada penelitian ini hanya menggunakan metode *K-Nearest Neighbor* (KNN) dan *Random Forest*.
- 2) Data pada penelitian ini akan menggunakan *dataset* penyakit kanker paru-paru yang diperoleh dari *Kaggle*.

#### **1.5 MANFAAT PENELITIAN**

Manfaat yang dapat diperoleh dari penelitian ini adalah memberikan pengetahuan mengenai *machine learning* dan mencari tahu perbandingan tingkat akurasi yang lebih baik diantara algoritma *K-Nearest Neighbor* dan *Random Forest* dalam kasus klasifikasi terhadap risiko seseorang dalam terkena penyakit kanker paru-paru.

#### **1.6 SISTEMATIKA PENULISAN**

Penulisan tugas akhir ini akan terbagi menjadi lima bab, yaitu:

- 1) BAB I PENDAHULUAN, berisi tentang latar belakang permasalahan, rumusan masalah, tujuan penelitian, batasan masalah, manfaat penelitian, dan sistematika penulisan.

- 2) BAB II LANDASAN TEORI, berisi tentang pengertian dari kanker paru-paru, *machine learning*, algoritma *K-Nearest Neighbor*, dan algoritma *Random Forest*, serta berisi tentang penelitian penelitian terdahulu.
- 3) BAB III METODOLOGI PENELITIAN, berisi tentang kerangka pikir dari penelitian ini, metode pengumpulan data, serta tahapan penelitian.
- 4) BAB IV HASIL DAN PEMBAHASAN, berisi tentang hasil dan analisis data yang diperoleh dari hasil penelitian.
- 5) BAB V KESIMPULAN DAN SARAN, berisi tentang kesimpulan dari penelitian dan saran saran yang dapat diberikan untuk perbaikan terhadap proses pada penelitian selanjutnya.

