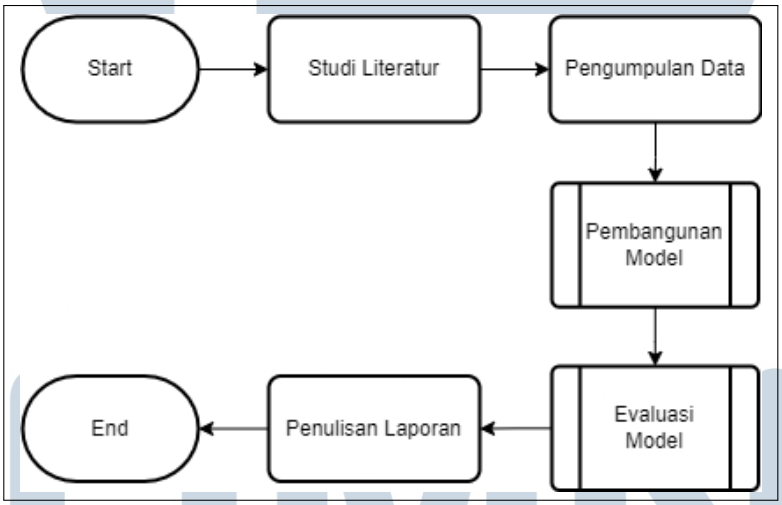
BAB 3
METODOLOGI PENELITIAN

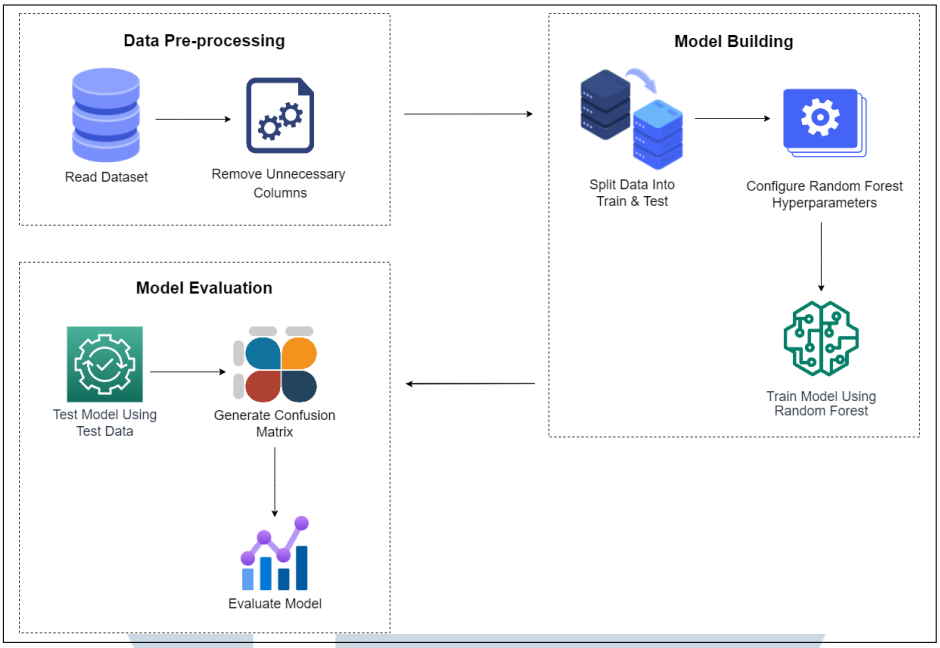
3.1 Gambaran Umum

Langkah awal dalam penelitian melibatkan pengumpulan dan pengolahan data. Pada tahap ini, *dataset* diunduh dan diolah terlebih dahulu sebelum digunakan dalam uji coba. Setelah itu, dilakukan proses pembangunan model yang melibatkan pelatihan model dengan menggunakan *dataset* yang telah diolah. Model yang telah dilatih kemudian dievaluasi menggunakan *confusion matrix* untuk mengukur metrik evaluasi seperti *accuracy*, *precision*, *recall*, dan *F1 score*. Gambar 3.1 adalah *flowchart* gambaran umum penelitian.



Gambar 3.1. *Flowchart* Gambaran Umum Penelitian

Alur kerja dalam penelitian ini dapat dilihat pada Gambar 3.2. Dalam proses *data pre-processing*, dari 26 kolom dilakukan penyaringan kolom yang tidak diperlukan seperti *Patient Id*. Kolom yang tersisa akan dijadikan sebagai fitur untuk pembangunan model. Sebelum mulai membangun model, *dataset* akan dibagi menjadi *train* dan *test data*. Tahap selanjutnya, model *Random Forest* dikonfigurasi dan dilatih menggunakan *train data*. Pada tahap evaluasi model, model diuji menggunakan data tes yang asli dan *confusion matrix* terbuat dari hasil pengujian. Berdasarkan hasil dari *confusion matrix*, performa model dapat dievaluasi lebih lanjut menggunakan beberapa metrik performa.



Gambar 3.2. Alur Kerja Utama Penelitian

3.1.1 Spesifikasi Perangkat

Untuk menjalankan penelitian ini, digunakan beberapa peralatan yang dapat dikategorikan sebagai perangkat keras dan perangkat lunak. Berikut rincian *hardware* dan *software* yang digunakan adalah sebagai berikut.

1. Hardware
    - Prosesor: Processor Intel(R) Core(TM) i5-9300H CPU @ 2.40GHz.
    - Memori: 24576MB RAM (DDR4).
    - VGA: NVIDIA GeForce GTX 1050 Ti.
    - Media Penyimpanan: HDD 1TB.
2. Software
    - Sistem Operasi: Windows 11 Pro 64-bit (10.0, Build 22621).
    - Web Browser: Brave.
    - Bahasa Pemrograman: LaTeX (Overleaf), Python.
    - Text Editor: Visual Studio Code.

3.2 Studi Literatur

Telaah literatur menjadi langkah awal dalam penyusunan laporan penelitian ini, bertujuan untuk menggali informasi terkait dengan tema yang menjadi fokus penelitian. Informasi yang diperlukan ditemukan dan dikumpulkan melalui proses membaca serta memahami berbagai materi dari sumber-sumber terverifikasi, seperti jurnal ilmiah, artikel ilmiah, buku, dan sumber lainnya.

3.3 Pengumpulan Data

Dataset yang digunakan dalam penelitian adalah *Cancer Patient Datasets* yang dapat ditemukan diwebsite *Kaggle* dengan format CSV yang memiliki 26 atribut yang berbeda dan mencakup 1000 sampel data. Untuk mengetahui gambaran mengenai *dataset* yang digunakan pada penelitian ini, berikut Gambar 3.3 dan Gambar 3.4 adalah gambaran *dataset* yang digunakan.

index	Patient Id	Age	Gender	Air Pollution	Alcohol use	Dust Allergy	OccuPational Hazards	Genetic Risk	chronic Lung Disease	Balanced Diet	Obesity	Smoking
0	P1	33	1	2	4	5	4	3	2	2	4	3
1	P10	17	1	3	1	5	3	4	2	2	2	2
2	P100	35	1	4	5	6	5	5	4	6	7	2
3	P1000	37	1	7	7	7	7	6	7	7	7	7
4	P101	46	1	6	8	7	7	7	6	7	7	8
5	P102	35	1	4	5	6	5	5	4	6	7	2
6	P103	52	2	2	4	5	4	3	2	2	4	3
7	P104	28	2	3	1	4	3	2	3	4	3	1
8	P105	35	2	4	5	6	5	6	5	5	5	6
9	P106	46	1	2	3	4	2	4	3	3	3	2
10	P107	44	1	6	7	7	7	7	6	7	7	7

Gambar 3.3. *Cancer Patient Datasets* I

Passive Smoker	Chest Pain	Coughing of Blood	Fatigue	Weight Loss	Shortness of Breath	Wheezing	Swallowing Difficulty	Clubbing of Finger Nails	Frequent Cold	Dry Cough	Snoring	Level
2	2	4	3	4	2	2	3	1	2	3	4	Low
4	2	3	1	3	7	8	6	2	1	7	2	Medium
3	4	8	8	7	9	2	1	4	6	7	2	High
7	7	8	4	2	3	1	4	5	6	7	5	High
7	7	9	3	2	4	1	4	2	4	2	3	High
3	4	8	8	7	9	2	1	4	6	7	2	High
2	2	4	3	4	2	2	3	1	2	3	4	Low
4	3	1	3	2	2	4	2	2	3	4	3	Low
6	6	5	1	4	3	2	4	6	2	4	1	Medium

Gambar 3.4. *Cancer Patient Datasets* II

3.4 Pembangunan Model

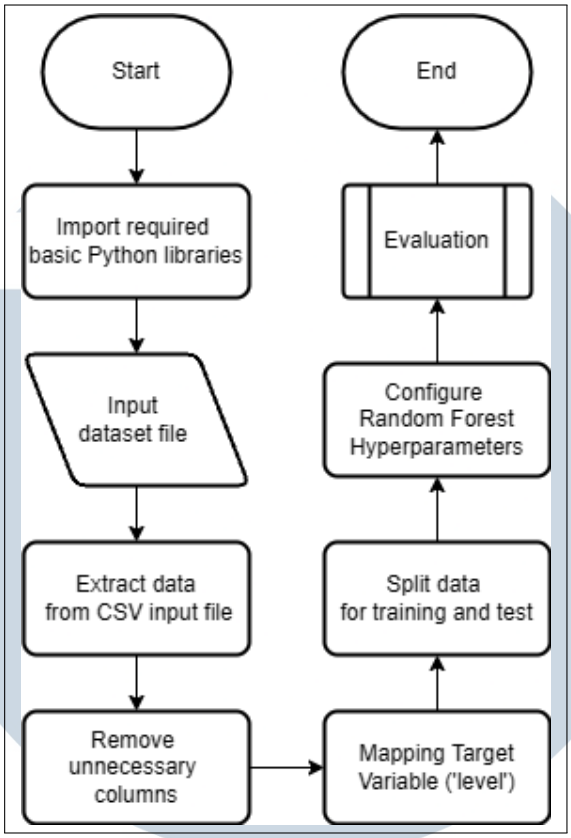
Flowchart mengenai pembangunan model dalam penelitian ini dapat dilihat pada Gambar 3.5. Pada tahap ini, langkah pertama yang dilakukan adalah melakukan *import library* yang dibutuhkan untuk mendukung pembangunan model. Selanjutnya, *dataset* di*import* untuk mendukung proses pelatihan dan pengujian model kemudian *dataset* diekstrak menggunakan *library DataFrame* dari *library pandas* dan ditampilkan dalam bentuk tabel untuk memudahkan analisis.

Setelah melihat hasil ekstrak *dataset*, tahap selanjutnya adalah membersihkan *dataset* dengan menghapus kolom yang tidak diperlukan untuk analisis dan menggunakan *correlation matrix* untuk mengetahui hubungan antara variabel-variabel dalam sebuah *dataset*. Dengan menggunakan *correlation matrix* dapat dibuat suatu fungsi untuk memilih fitur-fitur yang mempunyai korelasi kuat dengan variabel target (*"level"*) sehingga dapat menghilangkan fitur yang tidak relevan dengan variabel target (*"level"*) atau menghilangkan fitur yang ganda *duplicate*.

Setelah menghapus kolom yang tidak diperlukan, tahap berikutnya adalah menerapkan proses *"Mapping"* pada variabel target (*"level"*). *Mapping* ini bertujuan untuk mengubah nilai kategori teks menjadi nilai numerik agar lebih mudah diolah. Setelah proses *mapping*, nilai *'High'* akan menjadi 2, *'Medium'* akan menjadi 1, dan *'Low'* akan menjadi 0.

Selanjutnya, *dataset* dibagi menjadi data pelatihan dan data uji. Data pelatihan digunakan untuk melatih model *Random Forest*, sementara data uji digunakan untuk mengevaluasi kinerja model *Random Forest*. Pembagian *dataset* dilakukan dengan rasio 80:20, yang berarti 80% data akan digunakan untuk melatih model *Random Forest* (data pelatihan), sementara 20% data akan digunakan untuk menguji kinerja model *Random Forest* (data uji).

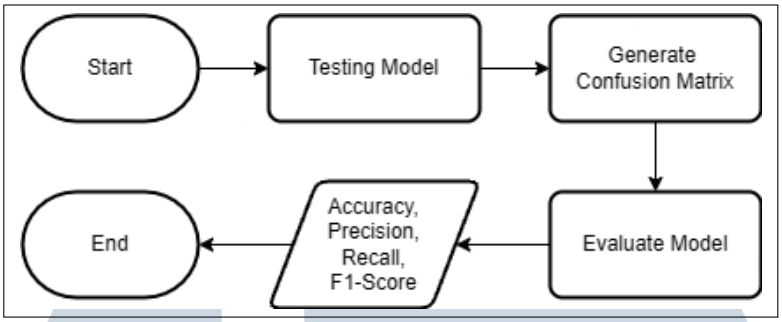
Sebelum melatih model, *hyperparameter* algoritma *Random Forest* perlu dikonfigurasi. Konfigurasi *hyperparameter* akan dibuat dengan variasi nilai *hyperparameter* tertentu untuk melihat apakah perubahan nilai *hyperparameter* tersebut mempengaruhi kinerja model yang dibangun. Setelah melakukan konfigurasi, model akan dilatih menggunakan data pelatihan dan hasil kinerja model dievaluasi menggunakan *confusion matrix*. Dari *confusion matrix*, berbagai metrik evaluasi seperti *accuracy*, *precision*, *recall*, dan *F1 score* dapat dihitung untuk mengevaluasi performa model.



Gambar 3.5. *Flowchart* Pembangunan Model

3.5 Evaluasi

Pada tahap ini, evaluasi dilakukan untuk mengukur sejauh mana model dapat menghasilkan klasifikasi yang akurat. Model akan dievaluasi menggunakan data uji dengan lima skenario konfigurasi *hyperparameter* untuk mengukur kinerja dan akurasi klasifikasi model terhadap tingkat risiko mengidap kanker paru-paru. Sesudah itu *confusion matrix* dibuat untuk setiap hasil dari lima skenario konfigurasi *hyperparameter*. Dari *confusion matrix* dievaluasi apakah perubahan nilai *hyperparameter* tersebut mempengaruhi kinerja model yang dibangun atau tidak. Hasil evaluasi (*accuracy*, *precision*, *recall* dan *F1 score*) ini akan memberikan gambaran tentang seberapa baik model dapat mengklasifikasikan tingkat risiko mengidap kanker paru-paru. Berikut Gambar 3.6 adalah *Flowchart* evaluasi model yang dilakukan.



Gambar 3.6. *Flowchart* Evaluasi Model

3.6 Penulisan Laporan dan Konsultasi

Pada tahap ini dilakukan penyusunan laporan dengan tujuan untuk mendokumentasikan hasil penelitian, perancangan, dan pengembangan aplikasi. Laporan tersebut bertujuan memberikan informasi yang dapat berguna untuk penelitian serupa di masa yang akan datang. Selain itu, konsultasi dengan dosen pembimbing juga dilaksanakan untuk memastikan bahwa penelitian ini terarah dan menghasilkan hasil yang memuaskan.