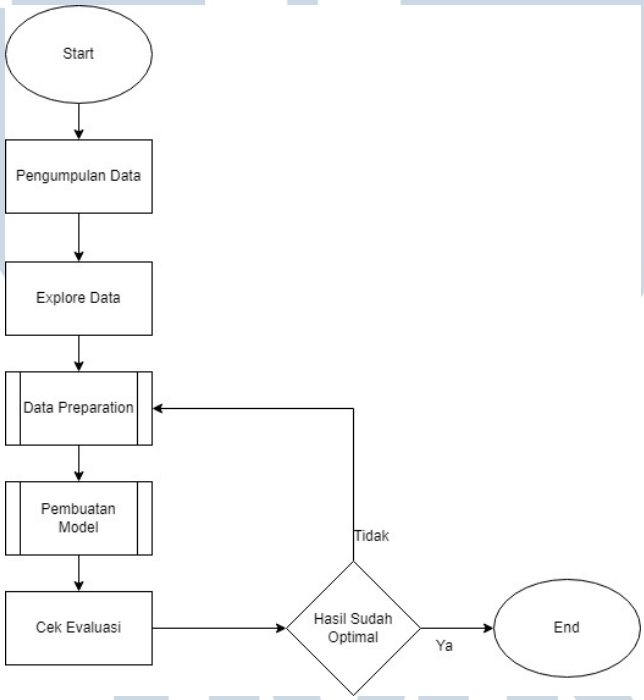
BAB 3
METODOLOGI PENELITIAN

3.0.1 Alur penelitian

berikut pada gambar 3.1 merupakan alur dari penelitian yang diakukan.



Gambar 3.1. Flowchart penelitian

Flowchart pada gambar 3.1 menggambarkan alur metode penelitian yang dilakukan dalam proyek ini. Berikut adalah tahapan-tahapan yang terdapat dalam flowchart

1. Pengumpulan Data: Proses pengumpulan data dimulai dengan permintaan resmi kepada Rumah Sakit Pusat Pertamina untuk mendapatkan data Medical Check-Up (MCU). Setelah disetujui, pada tanggal 27 Agustus 2024, peneliti mengunjungi bagian MCU untuk mengakses dan mengunduh data dari aplikasi web rumah sakit. Dataset yang diperoleh berupa file Excel yang terdiri atas dua bagian utama, yaitu ”REPORT MCU JAN-AGS 2024” dengan jumlah 10.305 data dan ”Excel 2023” dengan jumlah 3.707 data, sehingga

totalnya mencapai 13.628 data. Dataset ini mencakup berbagai informasi medis pasien, seperti kadar glukosa darah, HbA1c, kolesterol, trigliserida, lingkar pinggang, dan status gizi, serta memuat data pribadi seperti NIK, Alamat, Nama, Tanggal Lahir, NoPEK, dan lainnya. Untuk menjaga kerahasiaan, informasi pribadi tidak digunakan dalam analisis kecuali No Rekam Medis, yang dijadikan primary key untuk mengidentifikasi data secara anonim. Setelah data diperoleh, dilakukan pemeriksaan kelengkapan dan validitas untuk memastikan data siap digunakan pada tahap eksplorasi dan preprocessing.

2. Eksplorasi Data: Pada tahap ini, data yang telah dikumpulkan dianalisis untuk memahami karakteristiknya secara lebih mendalam. Analisis dilakukan untuk mengetahui struktur data, format masing-masing kolom, serta distribusi label, khususnya perbandingan antara jumlah data dengan label diabetes dan non-diabetes. Tahap ini bertujuan untuk mendapatkan gambaran awal mengenai kualitas dan pola data yang tersedia sebagai dasar untuk langkah-langkah pemrosesan lebih lanjut.

3. Data Preparation: Pada tahap ini, data yang telah dieksplorasi dipersiapkan untuk analisis lebih lanjut dengan melakukan serangkaian proses pembersihan dan transformasi. Proses ini mencakup penghapusan data duplikat untuk memastikan tidak ada informasi ganda yang dapat memengaruhi hasil analisis, serta penanganan nilai yang hilang atau tidak valid. Transformasi data dilakukan, seperti normalisasi atau standarisasi nilai numerik, agar lebih konsisten. Selain itu, label tambahan ditambahkan untuk informasi penyakit lain yang relevan, guna memperkaya analisis.

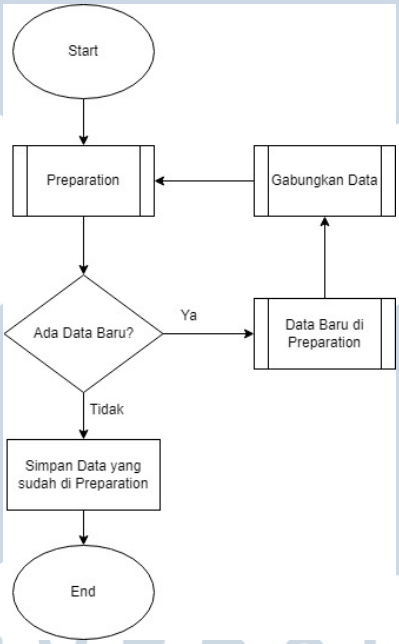
4. Pembuatan Model: Setelah data siap, langkah berikutnya adalah pembuatan model untuk menganalisis data dan memperoleh hasil penelitian. Dalam proses ini, dilakukan beberapa upaya untuk meningkatkan performa model, seperti penggunaan teknik SMOTE (Synthetic Minority Oversampling Technique) untuk menangani ketidakseimbangan data antara label diabetes dan non-diabetes. Selain itu, dilakukan pula proses hyperparameter tuning pada algoritma untuk mengoptimalkan parameter model, sehingga hasil evaluasi seperti akurasi, recall, precision, dan F1-score dapat lebih maksimal. Pendekatan ini memastikan model yang dihasilkan lebih akurat dan efektif dalam mendeteksi diabetes.

5. Cek Evaluasi: Setelah model dibuat, tahap evaluasi dilakukan untuk menilai kinerja model. Pertanyaan yang diajukan adalah apakah hasil evaluasi sudah mencapai optimal, mendekati 1.

    - **Jika Tidak**: Jika hasil evaluasi belum memenuhi kriteria, maka akan dilakukan percobaan dengan mencobakan skenario baru untuk meningkatkan kinerja model.
    - **Jika Ya**: Jika hasil evaluasi telah memenuhi kriteria, maka penelitian dapat dilanjutkan menuju tahap akhir.

3.0.2 Alur Preprocessing Data

Pada gambar 3.2 ditampilkan alur preprocessing untuk data baru serta proses penambahan data.



Gambar 3.2. Flowchart Data Preparation

1. Preparation: Tahap awal di mana data yang diperlukan untuk analisis disiapkan.

2. Gabungkan Data: Data dari berbagai sumber digabungkan untuk membentuk satu set data yang utuh.

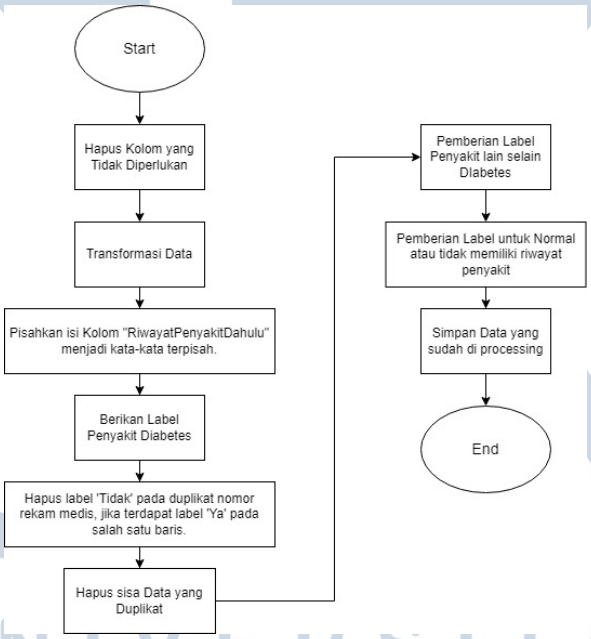
3. Ada Data Baru?: Pada tahap ini, sistem memeriksa apakah ada data baru yang perlu diproses.

    - Ya: Jika ada data baru, data tersebut akan dipindahkan ke tahap *Data Baru di Preparation* untuk diproses lebih lanjut.
    - Tidak: Jika tidak ada data baru, proses akan berlanjut ke langkah berikutnya.

4. Simpan Data yang sudah Preparation: Setelah semua data diproses dan digabungkan, data yang telah dipersiapkan disimpan untuk digunakan dalam analisis selanjutnya.

3.1 Alur Preprocessing Data

Pada gambar 3.3 merupakan flowchart isi dari proses preprocessing



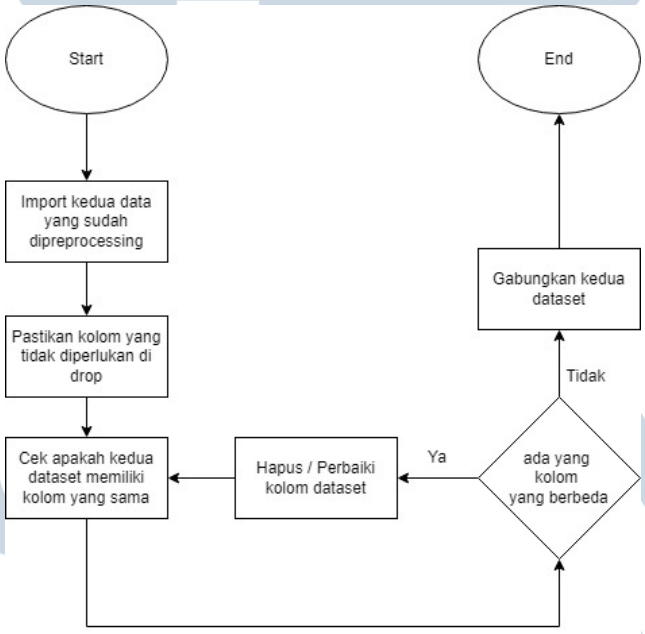
Gambar 3.3. Flowchart Preprocessing Data

1. Hapus Kolom yang Tidak Diperlukan: Kolom yang tidak relevan untuk analisis dihapus dari dataset.

2. Transformasi Data: Data yang ada diubah menjadi format yang sesuai untuk analisis

3. Hapus Data yang Duplikat: Data duplikat dihapus berdasarkan nomor rekam medis. Jika ada data dengan label *Ya*, namun terdapat duplikat, salah satu dari data tersebut dihapus.
4. Berikan Label: Label diberikan pada data berdasarkan kriteria tertentu. Dalam hal ini, data yang berisi informasi tentang riwayat penyakit diberikan label sesuai dengan apakah mereka memiliki riwayat diabetes atau tidak.
5. Pemberian Label untuk Normal atau tidak memiliki riwayat penyakit: Data yang tidak memiliki riwayat penyakit di-label sebagai *Normal*.
6. Simpan Data untuk Processing: Data yang telah dilabeli dan diproses disimpan untuk langkah analisis selanjutnya.

3.2 Alur Penggabungan Dataset



Gambar 3.4. Flowchart Penggabungan Data

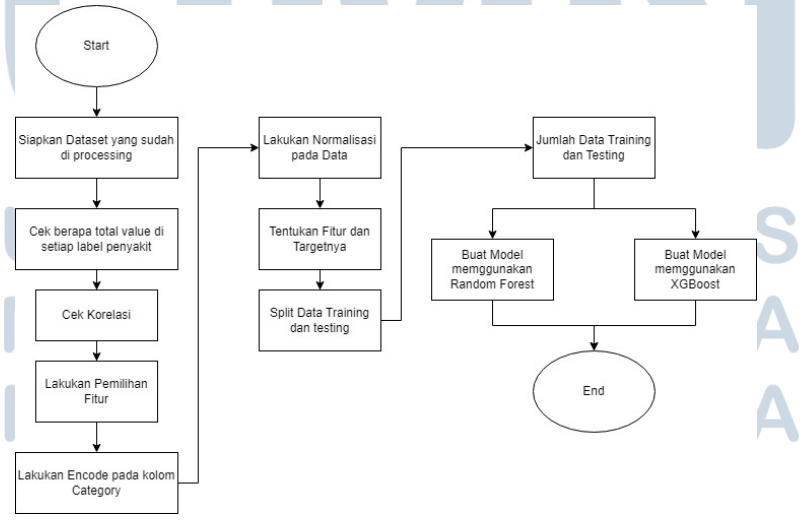
Proses penggabungan dataset seperti pada gambar 3.4 dilakukan melalui beberapa tahapan seperti berikut:

1. Import Kedua Data yang Sudah Diproses Kedua dataset yang akan digabungkan diimpor terlebih dahulu ke dalam sistem analisis data. Dataset tersebut telah melalui tahap preprocessing seperti normalisasi, encoding, dan penanganan *missing values*.

2. Memastikan Kolom yang Tidak Diperlukan Dihapus Kolom yang tidak relevan atau tidak mendukung proses analisis dihapus dari kedua dataset. Langkah ini dilakukan untuk memastikan hanya fitur yang penting yang dipertahankan dalam model prediksi diabetes.
3. Pemeriksaan Konsistensi Kolom Dilakukan pengecekan apakah kedua dataset memiliki kolom dengan nama dan tipe data yang sama. Hal ini penting untuk memastikan integrasi yang benar antara kedua dataset.
4. Penanganan Kolom yang Berbeda
   (a) Jika ditemukan perbedaan kolom antara kedua dataset, langkah selanjutnya adalah memperbaiki atau menghapus kolom tersebut agar sesuai.
   (b) Jika semua kolom sudah seragam, proses dilanjutkan ke tahap penggabungan.
5. Penggabungan Kedua Dataset Setelah dipastikan bahwa kedua dataset memiliki kolom yang sama, langkah berikutnya adalah menggabungkan kedua dataset menjadi satu kesatuan data yang siap untuk digunakan dalam proses analisis lebih lanjut.

3.3 Alur Pembuatan Model

Pada gambar 3.5 ditampilkan flowchart alur pembuatan model



Gambar 3.5. Flowchart Pembuatan Model

1. Siapkan dataset yang sudah diproses (pembersihan data hilang, penghapusan duplikat, penyamaan format data).
2. Cek total value di setiap label penyakit (cek distribusi kelas, lakukan SMOTE jika tidak seimbang).
3. Cek korelasi antara fitur numerik dengan label target.
4. Lakukan pemilihan fitur berdasarkan hasil analisis korelasi.
5. Lakukan encoding pada kolom kategori (*StatusGizi*, *OlahRaga*, *Merokok*) dengan *one-hot encoding*.
6. Lakukan normalisasi pada data numerik menggunakan *Min-Max Scaler*.
7. Tentukan fitur dan target (X sebagai variabel independen, Y sebagai target atau label).
8. Split data menjadi data pelatihan dan pengujian dengan rasio seperti 80:20 atau 70:30.
9. Cek jumlah data pada data pelatihan dan pengujian serta distribusi kelasnya.
10. Buat model menggunakan algoritma *Random Forest*.
11. Buat model menggunakan algoritma *XGBoost*.

3.4 Spesifikasi Sistem

Penelitian ini dilakukan menggunakan perangkat keras dan perangkat lunak dengan spesifikasi sebagai berikut:

3.4.1 Spesifikasi Perangkat Keras

- **Prosesor:** Intel(R) Core(TM) i5-6200U CPU @ 2.30GHz 2.40GHz
- **RAM Terinstal:** 8,00 GB (7,85 GB dapat digunakan)
- **Penyimpanan:** SSD/HDD dengan ruang yang cukup untuk menyimpan dataset sebesar 13.628 data dan hasil analisis
- **Jenis Sistem:** 64-bit operating system, x64-based processor
- **Sistem Input:** Tidak tersedia pen atau layar sentuh

3.4.2 Spesifikasi Perangkat Lunak

- **Sistem Operasi:** Windows 10 Home Single Language, versi 22H2
- **Bahasa Pemrograman:** Python 3.9.12
- **Lingkungan Pemrograman:** Jupyter Notebook
- **Library Python yang Digunakan:**
  - **Scikit-learn:** Untuk implementasi algoritma Random Forest dan XGBoost
  - **Pandas** dan **NumPy:** Untuk manipulasi dan analisis data
  - **Matplotlib** dan **Seaborn:** Untuk visualisasi data
  - **Imbalanced-learn:** Untuk penyeimbangan data dengan SMOTE
  - **XGBoost Library:** Untuk implementasi algoritma XGBoost
- **Jupyter Core Packages:**
  - **IPython:** 8.2.0
  - **ipykernel:** 6.9.1
  - **jupyter_client:** 6.1.12
  - **nbformat:** 5.3.0