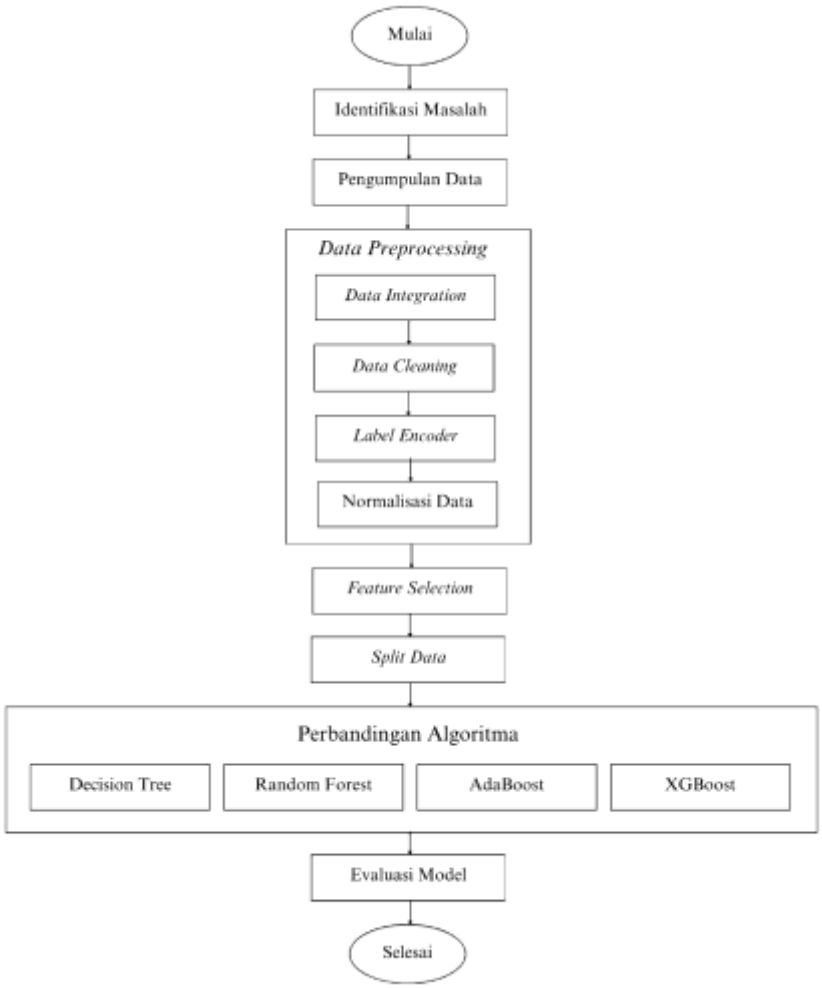
BAB III

METODOLOGI PENELITIAN

Metode penelitian sebagai dasar ilmiah untuk mengumpulkan data dan menemukan jawaban terhadap berbagai pertanyaan yang relevan untuk subjek penelitian. Metodologi penelitian adalah suatu tahapan yang diterapkan dalam penelitian algoritma pada pendekatan *supervised learning* dengan seleksi fitur Chi-Square untuk menentukan ketepatan status kesehatan jemaah haji. Tahapan-tahapan yang dilakukan dalam penelitian ini digambarkan pada Gambar 3.1.



Gambar 3.1 Tahapan Penelitian

3.1 Identifikasi Masalah

Identifikasi masalah merupakan suatu proses menemukan serta mendefinisikan masalah yang akan diteliti sehingga alur penelitian dapat lebih terstruktur. Identifikasi masalah pada penelitian ini dilakukan dengan melalui pengamatan dan penemuan permasalahan yang muncul terkait status kesehatan jemaah haji. Sebagai informasi sebelum keberangkatan jemaah haji, akan dilakukan evaluasi terlebih dahulu apabila terdapat permasalahan yang terjadi pada tingkat kesehatan jemaah haji di wilayah Tasikmalaya pada tahun 2024. Data yang berkaitan dengan pemeriksaan jemaah haji diperlukan dalam proses penerapan algoritma pada pendekatan *supervised learning*.

3.2 Pengumpulan Data

Pengumpulan data merupakan tahapan penting dalam sebuah penelitian untuk membangun dan mengembangkan pemahaman mendalam terkait dengan topik yang akan diteliti (Rifa'i, 2023). Untuk dapat mengetahui informasi tentang kesehatan jemaah haji, yaitu dengan menggunakan studi literatur. Studi literatur dilakukan dengan mengumpulkan data yang relevan yang berkaitan dengan penelitian yang sedang diteliti. Informasi yang didapatkan bisa melalui berbagai sumber seperti jurnal ilmiah, buku, artikel, dan sumber lainnya. Data yang digunakan dalam penelitian ini adalah dataset pemeriksaan kesehatan jemaah haji tahun 2024, yang diambil dari Dinas Kesehatan Kota dan Kabupaten Tasikmalaya, melalui Sistem Komputerisasi Haji Terpadu Bidang Kesehatan (Siskohatkes) dengan total sebanyak 2.522 data. Alasan penggunaan data yang digunakan terbatas hanya pada wilayah Tasikmalaya adalah untuk menguji dan menentukan algoritma

yang paling sesuai untuk klasifikasi status kesehatan jemaah haji sehingga dapat dilakukan sebagai tahapan awal sebelum menerapkan metode tersebut pada data yang yang lebih luas pada penelitian berikutnya. Karena data yang digunakan berasal langsung dari lapangan, hasilnya dapat segera dianalisis untuk memastikan kesesuaiannya.

3.3 *Data Preprocessing*

Data preprocessing merupakan tahapan penting dalam pengolahan data mining. *Data preprocessing* dilakukan untuk memilih data dalam dataset yang digunakan, sehingga data menjadi lebih ringkas dan relevan serta menghilangkan data yang tidak diperlukan, seperti menghapus beberapa bagian yang tidak digunakan (A'yuniyah & Reza, 2023). Terdapat beberapa tahapan *preprocessing* yang dilakukan pada penelitian ini, diantaranya:

1. *Data Integration*

   Data pemeriksaan kesehatan jemaah haji di wilayah Tasikmalaya memiliki dua data yaitu data dari Kota dan Kabupaten Tasikmalaya, untuk menggunakan kedua data tersebut maka dilakukan penggabungan data pemeriksaan kesehatan jemaah haji di Kota dan Kabupaten Tasikmalaya.

2. *Data Cleaning*

   Data cleaning atau pembersihan data merupakan suatu prosedur untuk memastikan kebenaran, konsistensi, dan fungsi pada suatu data yang terdapat dalam dataset (Kartika Sari dkk., 2024). Tujuan dari tahap pembersihan untuk

memastikan dataset tidak mengandung *noise* yang dapat berpengaruh terhadap hasil klasifikasi.

3. *Label Encoder*

Label encoder dilakukan untuk mengubah tipe data *string* pada variabel kategori menjadi tipe numerik agar mudah dipahami model. Dilakukannya *label encoder* dapat lebih efisien dalam memahami dan memproses informasi, karena algoritma umumnya berfungsi lebih optimal saat bekerja dengan data numerik (Septian, 2024).

4. Normalisasi Data

Normalisasi data merupakan proses mengubah skala nilai atribut menjadi rentang yang lebih kecil dengan bobot yang seimbang. Skala baru ini dapat meningkatkan kinerja klasifikasi yang membantu menghapus fitur yang memiliki tingkat *noise* tinggi dan relevansi rendah (Suryanegara dkk., 2021).

3.4 *Feature Selection*

Feature selection atau seleksi fitur merupakan suatu proses menghapus fitur yang berlebihan dan tidak relevan dari dataset yang sebenarnya. Sehingga waktu yang digunakan mengeksekusi dari pengklasifikasi yang memproses data berkurang, dan dapat meningkatkan akurasi juga karena fitur yang tidak relevan dapat memperburuk data mempengaruhi akurasi klasifikasi secara negatif (Rahmansyah dkk., 2018). Seleksi fitur yang digunakan dalam penelitian ini yaitu Chi-Square metode berbasis *filter*. Metode statistik ini mengevaluasi hubungan antara nilai suatu fitur dengan kelas target melalui uji statistik. Fitur

dengan nilai chi-square yang lebih tinggi menunjukkan hubungan yang lebih kuat dengan kelas target (Lumbantobing dkk., 2020)

3.5 *Split Data*

Split Data merupakan pembagian dataset menjadi dua atau lebih subset, seperti *data training* dan *data testing*. Tahap ini termasuk tahapan penting pada pembentukan model yang mampu memberikan hasil klasifikasi dengan tingkat akurasi yang optimal. Penelitian ini dilakukan beberapa skema pembagian data.

3.6 Perbandingan Algoritma

Penelitian ini melakukan perbandingan algoritma pada *pendekatan supervised learning*. Klasifikasi yang diterapkan dalam penelitian yaitu dengan menggunakan algoritma Decision Tree, Random Forest, AdaBoost dan XGBoost. Penelitian klasifikasi diarahkan pada pembagian data terkait status kesehatan jemaah haji ke dalam empat kelas. Implementasi klasifikasi dari keempat algoritma akan dijalankan dengan Google Collaboratory menggunakan bahasa pemrograman Phyton.

3.7 Evaluasi Model

Evaluasi model merupakan proses pengujian model yang bertujuan untuk mengevaluasi performa algoritma yang digunakan. Penelitian ini menggunakan *confusion matrix* dalam evaluasi model. Teknik ini melibatkan parameter evaluasi seperti akurasi*, precision, recall*, dan f1-*score.*

1. Akurasi

Akurasi merupakan penggambaran tingkat efektifitas pada algoritma Decision Tree, Random Forest, Adaboost dan XGBoost yang digunakan pada tahap klasifikasi. Persamaannya adalah (Muhaimin dkk., 2024):

$$Accuracy = \frac{TP + TN}{TP + FP + FN + TN} \quad (5)$$

2. *Precision*

Precision merupakan proses yang melibatkan perhitungan jumlah *record* atau data dengan nilai *true* positif, kemudian dibagi dengan jumlah keseluruhan *record*. Persamaannya adalah:

$$Precision = \frac{TP}{TP + FP} \quad (6)$$

3. *Recall*

Recall merupakan rasio prediksi benar positif dengan keseluruhan jumlah data yang benar positif. Persamaannya adalah:

$$Recall = \frac{TP}{TP + FN} \quad (7)$$

4. F1-*score*

F1-*score* merupakan perhitungan dengan membandingkan rata-rata nilai *recall* dan *precision*. Persamaannya adalah:

$$F1 - score = 2 \times \frac{(Recall \times Precision)}{(Recall + Precision)} \quad (8)$$

Tahap evaluasi dilakukan *confusion matrix* untuk memberikan informasi tentang kinerja model klasifikasi. Evaluasi dilakukan untuk melihat perbandingan

hasil klasifikasi dari algoritma Decision Tree, Random Forest, Adaboost dan XGBoost yang kemudian menemukan model klasifikasi dengan melihat performa masing-masing algoritma.