BAB III

METODOLOGI PENELITIAN

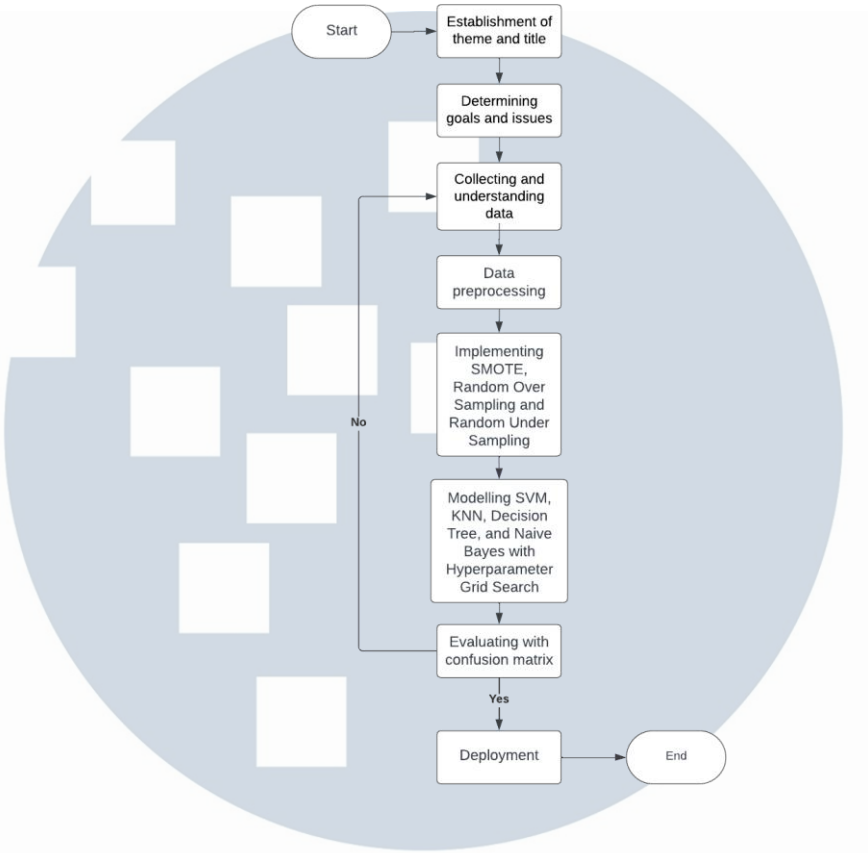
3.1 Objek Penelitian

Penelitian berfokuskan terhadap mengembangkan dan membandingkan algoritma KNN, SVM, Decision Tree, dan Naïve Bayes terhadap kasus kecelakaan lalu lintas Kepolisian Daerah Jawa Timur. Penelitian menggunakan data kasus kecelakaan lalu lintas Polda Jatim tahun 2024. Tantangan terhadap klasifikasi data yang memiliki 3 kategori, yaitu Ringan, Sedang, dan Berat, dimana kelas tingkat kecelakaan sedang memiliki 413 data. Hal ini menyebabkan ketidakstablian pada data yang cukup signifikan, makan dibentuk kelas baru yang menggabung kategori sedang dan berat menjadi tingkat kecelakaan Darurat. Pengolahan data juga dilakukan untuk mendapatkan faktor utama yang mempengaruhi peningkatan tingkat kecelakaan pada data Polda Jatim. Pada data Polda Jatim, jumlah data yang terkumpul mencapai 28.000 data kecelakaan lalu lintas tercatat di tahun 2024. Mengembangkan model KNN, SVM, Decision Tree, dan Naïve Bayes, akan menghasilkan perbandingan akurasi keempat model baik menggunakan atau tidak menggunakan teknik *sampling* (SMOTE, ROS, dan RUS), serta menentukan kecocokan keempat algoritma terkait dengan klasifikasi Tingkat kecelakaan. Model yang memiliki akurasi terbaik akan dibentuk menjadi aplikasi web sederhana untuk melakukan klasifikasi terhadap kasus kecelakaan berdasarkan faktor utama, juga dapat menjadi referensi untuk dikembangkan pada penelitian terhadap data kecelakaan Polda Jatim.

3.2 Metode Penelitian

3.2.1 Alur Penelitian

Metode penelitian ini digambarkan menggunakan *flowchart*, yang menjelaskan alur proses dari penelitian ini.



Gambar 3.1 Diagram Alur Penelitian

Pada gambar 3.1, ditunjukan alur penelitian yang dimulai dengan penetapan judul. Mencakup identifikasi permasalahan, penetapan manfaat dan tujuan penelitian dengan melakukan riset terhadap latar belakang tema yang telah ditetapkan. Dilanjutkan dengan proses penetapan model yang ingin digunakan untuk mengatasi permasalahan, yang pada penelitian ini akan dilakukan klasifikasi terhadap data kecelakaan lalu lintas Polda Jatim tahun 2024 dengan menggunakan algoritma KNN, SVM, Decision Tree, dan Naïve Bayes.

Data yang telah ditetapkan akan melalui proses *preprocessing* atau data *preparation*. Pada tahapan ini, dilaksanakan proses *encoding* terhadap fitur, juga mencari faktor yang memiliki korelasi tinggi terhadap peningkatan tingkat kecelakaan. Faktor yang mendapatkan korelasi terbaik terhadap tingkat kecelakaan adalah cuaca, kondisi cahaya, dan

kondisi permukaan jalan, kemudian menampilkan informasi terkait keempat fitur, tidak didapatkan *missing values* yang perlu ditangani. Pada tahapan ini juga perubahan kategori tingkat kecelakaan dilakukan karena jumlah distribusi tingkat sedang yang memiliki angka signifikan sedikit hanya mencapai 413 data dibandingkan tingkat ringan dan berat. Setelah melalui proses percobaan dan *sampling* menggunakan augmentasi, hasil yang diberikan tidak optimal dikarenakan tingginya angka *noise* atau duplikasi data. Maka dari itu, penelitian ini mengambil pendekatan dengan menggabungkan tingkat sedang dengan berat menjadi darurat, menghasilkan 5.000 data tingkat darurat dan 22.000 data tingkat ringan.

3.2.2 Metode Data Mining

Tabel 3.1 Perbandingan *framework data mining*

Framework	**CRISP-DM**	**SEMMA**
Proses	*Business Understanding*	*Sample*
	Data Understanding	*Explore*
	Data Preparation	*Modify*
	Data Modeling	*Model*
	Evaluation	*Assess*
	Deployment	
Kelebihan	- Fleksibel dan iteratif - Dapat diterapkan di berbagai domain	- Fokus pada analisis data yang mendalam. - Memungkinkan iterasi dan perbaikan model. - Dapat diterapkan pada berbagai jenis data dan masalah.
Kekurangan	- Kurang terstruktur untuk proyek besar - Tidak selalu mencakup aspek pemeliharaan model	- Memerlukan pemahaman statistik yang kuat. - Proses yang bisa memakan waktu, terutama dalam tahap eksplorasi dan modifikasi. - Tidak selalu cocok untuk proyek dengan batasan waktu yang ketat.
Kompleksitas	Menengah; mudah dipahami tetapi dapat menjadi rumit dalam implementasi besar	Menengah hingga tinggi; tergantung pada ukuran dan kompleksitas dataset serta tujuan analisis.

Penelitian ini akan dikerjakan secara individu, maka dari itu TDSP tidak dapat diaplikasikan sebagai metode *data mining* pada penelitian kali ini. CRISP-DM dan SEMMA merupakan opsi yang seimbang untuk menjadi metode pada penelitian. pada tabel 3.1, CRISP-DM memberikan pendekatan yang lebih struktural, dan proses dari SEMMA akan memakan waktu yang lama terutama tahapan eksplorasi dan modifikasi. Pada penelitian ini, digunakan proses CRISP-DM juga, dikarenakan proyek akan melalui tahapan *deployment* untuk menampilkan performa model yang dipilih terhadap klasifikasi tingkat kecelakaan

Metode yang digunakan dalam penelitian ini adalah CRISP-DM. CRISP-DM merupakan metode *data mining* yang sering digunakan dan memiliki tahapan yang terstruktur, juga mudah dipahami. CRISP-DM memiliki 6 tahapan, berikut tahapannya:

1. *Business understanding*

   Dalam tahapan ini, penelitian harus menentukan permasalahan dan tujuan dari penelitian, penelitian juga harus dapat mengidentifikasi kebutuhan yang dapat menjadi solusi terhadap permasalahan. Penelitian ini ingin menentukan model algoritma terbaik diantara KNN, SVM, Decision Tree, dan Naïve Bayes untuk melakukan klasifikasi terhadap Tingkat kecelakaan lalu lintas di Jawa Timur.
2. *Data understanding*

   Dalam mengembangkan model *facial recognition*, penelitian menggunakan data kecelakaan lalu lintas dari Kepolisian Polda Jawa Timur dengan periode Tahun 2020-2024. Berisikan catatan laporan kecelakaan yang dikumpulkan dari seluruh Kepolisian Resort Jawa Timur ke Kepolisian Daerah.
3. *Data preparation*

   Proses *data preparation* mencakup pembersihan data, koreksi format, dan lainnya. Data kecelakaan lalu lintas Polda Jatim,

dikumpulkan dari 5 data yang terpisah, yaitu data kecelakaan tahun 2020 hingga 2024. Ke-lima data akan digabungkan menjadi 1 *big data* yang akan dipangkas lagi, mulai dari pembersihan *null,* konversi format tanggal dan waktu, pelabelan, sampai dapat dikelola ke dalam proses pemodelan algoritma.

4. *Modeling*

   Menerapkan model algoritma KNN, SVM, Decision Tree, dan Naïve Bayes dengan *hyperparameter* Grid Search untuk mengklasifikasi Tingkat kecelakaan lalu lintas. Menggunakan data *training* untuk melatih model yang akan di uji menggunakan data *testing*. Model juga dikombinasikan menggunakan teknik *sampling* (SMOTE, ROS, atau RUS) yang akan dibandingkan juga terhadap performa model tanpa *sampling*.

5. *Evaluation*

   Model yang telah dikembangkan menggunakan kedua algoritma yang akan dievaluasi agar menghasilkan nilai akurasi yang dapat diukur. Penelitian ini menggunakan *confusion matrix,* dengan menghasilkan *accuracy, precision, recall,* dan *f1-score.* Model dengan nilai pengujian yang tinggi akan dijadikan

6. *Deployment*

   Model dengan akurasi tertinggi akan dibentuk menjadi sebuah aplikasi web menggunakan Streamlit yang dapat mengklasifikasi tingkat kecelakaan berdasarkan faktor utama.

3.2.3 Metode Pengolahan Data

Tabel 3.2 Perbandingan Algoritma Pengolahan Data

Algortima	Kelebihan	Kekurangan
KNN	- Sederhana dan mudah dipahami. -Tidak memerlukan pelatihan model yang rumit. - Efektif untuk dataset kecil dan sederhana.	- Rentan terhadap noise dan outlier. -Memerlukan waktu komputasi yang tinggi untuk dataset besar. - Kinerja tergantung pada pemilihan nilai k.
SVM	- Kinerja baik pada data berdimensi tinggi. - Efektif dalam kasus klasifikasi non-linier dengan kernel trick. - Robust terhadap overfitting, terutama pada dataset kecil.	- Memerlukan pemahaman yang baik tentang parameter dan kernel. - Waktu pelatihan bisa lama untuk dataset besar. - Kurang efektif jika jumlah fitur jauh lebih besar dari jumlah sampel.
Decision Tree	- Mudah diinterpretasikan dan divisualisasikan. - Dapat menangani data kategorikal dan numerik. - Tidak memerlukan normalisasi data.	- Rentan terhadap overfitting jika tidak dipangkas dengan baik. - Kinerja dapat menurun pada dataset yang tidak seimbang. - Sensitif terhadap perubahan kecil dalam data.
Naïve Bayes	- Sederhana dan cepat dalam pelatihan. - Efektif untuk data dengan fitur independen. - Dapat menangani data dengan ukuran besar.	- Asumsi independensi fitur sering tidak terpenuhi. - Tidak efektif jika fitur-fitur memiliki korelasi kuat. - Kurang akurat dibandingkan algoritma lain dalam beberapa kasus.

Penelitian menggunakan keempat algoritma untuk melakukan klasifikasi terhadap data kecelakaan lalu lintas Polda Jatim. Model SVM, KNN, Decision Tree, dan Naïve Bayes akan dikombinasikan menggunakan *hyperparameter* Grid Search dan teknik *sampling* (SMOTE, ROS, dan RUS) juga akan dilakukan perbandingan performa model untuk dikembangkan menjadi aplikasi web sederhana dalam mengklasifikasi tingkat kecelakaan. Evaluasi dari setiap model akan dilakukan menggunakan *confusion matrix* menampilkan *accuracy, precision, recall,* dan *f1-score accuracy, precision, recall,* dan *f1-score*.

3.3 Teknik Pengumpulan Data

Pengumpulan data dilakukan secara studi dokumentasi dan berpartisipasi di Polda Jatim. Data yang diperoleh untuk dilakukan analisis merupakan data kecelakaan lalu lintas Kepolisian Daerah Jawa Timur tahun 2024. Jumlah data mencapai 28.000 data kasus kecelakaan lalu lintas yang tercatat.

3.4 Variabel Penelitian

3.4.1 Variabel Tergantung (Dependent Variable)

Dependent variable pada penelitian ini berupa tingkat kecelakaan lalu lintas yang dikategorikan sebagai Ringan, Sedang, dan Berat. Akan dibentuk sebuah program yang menggunakan model dengan akurasi prediksi klasifikasi tertinggi dari setiap model yang telah dibentuk. Akurasi ditentukan menggunakan *confusion matrix* yang terdiri dari *accuracy, precision, recall,* dan *f1-score.*

3.4.2 Variabel Bebas (Independent Variable)

Independent variable yang dibutuhkan untuk menjadi parameter pendukung *dependent variable* pada penelitian ini berupa faktor-faktor yang dapat meningkatkan tingkat kecelakaan lalu lintas, kondisi cahaya, cuaca, dan kondisi permukaan jalan.

3.5 Teknik Analisis Data

Data dikumpulkan dengan teknik studi dokumentasi secara partisipasi, dengan melakukan program magang dan melaksanakan kegiatan peng-*input*an data kecelakaan di Polda Jawa Timur pada tahun 2024. Menggunakan data dari Polda Jawa Timur terkait klasifikasi tingkat kecelakaan yang bersifat kategorikal, pendekatan analisis data menggunakan teknik kualitatif untuk mengkategorikan kombinasi faktor-faktor yang mengakibatkan peningkatan tingkat kecelakaan agar dapat menghasilkan pola klasifikasi. Data yang telah dikumpulkan akan dikelola melalui metode CRIPS-DM, dimulai dengan memahami topik permasalahan dari data yang telah dikumpulkan sampai dengan pembentukan model klasifikasi SVM, KNN, Decision Tree, dan Naïve

Bayes yang akan dikembangkan menjadi aplikasi *web* sederhana menggunakan Streamlit. Analisis data ini dilakukan agar dapat mengklasifikasi tingkat kecelakaan di Jawa Timur dengan upaya membantu pengambilan keputusan terhadap pencegahan peningkatan tingkat kecelakaan dan efisiensi kinerja pendataan yang dilakukan Instansi Polda Jatim.

Pada proses analisis data digunakan sebuah *tools* yang dapat mengelola sebuah dataset kecelakaan yang disimpan dengan format Microsoft Excel. Terdapat perbandingan *tools* yang telah dipertimbangkan dalam memenuhi kriteria untuk melaksanakan analisis klasifikasi data menggunakan bahasa pemrograman Python.

Tabel 3.3 Perbandingan *tools* analisis data

Software	**Google Colab**	**Apache Zeppelin**	**JupyterNotebook**
Bahasa	Mendukung Python, R, dan beberapa bahasa lainnya	Mendukung banyak bahasa seperti Python, Scala, R, SQL, dan lainnya	Mendukung berbagai bahasa (Python, R, Julia, SQL, dll.)
Antarmuka	Berbasis web, antarmuka sederhana dan mudah digunakan	Mendukung eksplorasi data dan visualisasi data intuitif	Mudah untuk digunakan dan interaktif dalam menjalankan kode
Visualisasi	Matplotlib, Seaborn, Plotly, dll.	Highcharts, Plotly, dll	Matplotlib, Seaborn, Plotly, dll.
Komputasi	Menyediakan GPU/TPU gratis untuk komputasi berat	Mendukung integrasi dengan Spark, Flink, dan lainnya untuk komputasi terdistribusi	Bergantung pada sumber daya lokal atau server yang dihosting (tidak ada GPU/TPU terintegrasi secara default)
Fleksibilitas	Fleksibilitas terbatas pada ekosistem Google, lebih cocok untuk analisis data dan machine learning	Fleksibel pada tahap evaluasi dan pemodelan algoritma	Fleksibel dalam pemilihan dan pengembangan algoritma juga teknologi

Berdasarkan tabel 3.3, Penelitian ini menggunakan *software* berupa Google Colab yang mendukung penggunaan Python sebagai bahasa pemrogramannya.

Penggunaan *software* dapat dilakukan berbasis *website* tanpa menggunakan GPU atau CPU yang dapat membatasi komputasi data besar. Hal ini mempermudah proses *tuning parameter* yang akan dapat memakan waktu cukup lama terhadap GPU yang digunakan. Google Colab juga merupakan *software* yang umum digunakan, agar dapat mudah mencari solusi terkait permasalahan *code*.