BAB II
TINJAUAN PUSTAKA DAN DASAR TEORI

2.1 Tinjauan Pustaka

Dalam penelitian ini mengenai prediksi tingkat keparahan cedera kecelakaan lalu lintas dengan Random Forest didasarkan pada suatu penelitian terdahulu yaitu :

Ita Rakhmawati (2015) meneliti klasifikasi faktor-faktor yang mempengaruhi korban kecelakaan lalu lintas di Surabaya menggunakan metode Random Forest. Penelitian ini menerapkan pembagian data training dan testing dengan rasio 95% : 5% serta berbagai kombinasi jumlah pohon. Hasil menunjukkan bahwa kombinasi 50 pohon memberikan akurasi klasifikasi tertinggi pada data testing. Ketepatan klasifikasi untuk data training mencapai 78,69% sementara untuk data testing sebesar 58,33% *(Ita Rakhmawati, 2015)*.

Diluar bidang transportasi, Random Forest telah banyak diterapkan dalam berbagai sektor, terutama di dunia kesehatan dan berbankan. Dibidang kesehatan algoritma ini digunakan untuk mendukung diagnosis penyakit seperti :

Jalis Dwi Muthohhar & Agus prihanto (2023) penelitian tentang analisis perbandingan klasifikasi untuk penyakit jantung. Dalam penelitian ini menggunakan beberapa algoritma yaitu naive bayes, decision tree dan random forest. Hasil menunjukkan bahwa random forest paling unggul dengan perolehan nilai perfoma f1-score dari random search mendapatkan 0,852 dan grid search mendapatkan nilai 0,868 *(Dwi Muthohhar & Prihanto, 2023)*.

Maria Artati Eka Setyorini (2020) melakukan penelitian tentang analisis deteksi kanker paru-paru. Hasil penelitian menunjukkan bahwa metode Random Forest menghasilkan akurasi klasifikasi sebesar 90,32%. Setelah diterapkan teknik PCA-Random Forest akurasi meningkat hingga 100%. Selain itu visualisasi hasil klasifikasi menggunakan PCA-Random Forest mampu memetakan data dari dua kelas kanker paru-paru yaitu NO dan YES dengan cepat *(Maria Artati Eka S, 2020)*.

Nawang Anggita W, Dwi Puspa M, Candra Ayunda AS, Ummi Athiyah (2023) melakukan penelitian tentang analisis klasifikasi citra kanker kulit. Penelitian menunjukkan bahwa klasifikasi dengan random forest mendapatkan hasil akurasi sebesar 100% yang termasuk ke dalam excelent classification dan dapat mengklasifikasikan setiap kelas dengan benar *(Winanti et al., 2023)*.

Yuri Yuliani (2022) melakukan penelitian memprediksi kelangsungan hidup pasien gagal jantung dengan menggunakan 3 algoritma yaitu Random Forest, Random Subspace dan Logitboost. Penelitian menunjukkan bahwa performa algoritma Random Forest lebih unggul dengan menambahkan metode percentage split 80% yang menghasilkan accuracy sebesar 91,45%, precision 0.915, recall 0.914, AUC 0.953. Temuan ini menegaskan bahwa Random Forest unggul dalam menangani permasalahan klasifikasi karena kemampuannya dalam mengolah data dengan stuktur yang kompleks dan memberikan hasil prediksi yang lebih akurat dibandingkan algoritma lainnya *(Yuliani, 2022)*.

Sementara itu dalam sektor perbankan sering menggunakan metode random forest diantaranya :

Budi Prasojo, Emy Haryatmi (2021) melakukan penelitian menganalisa prediksi kelayakan pemberian kredit pinjaman dengan metode random forest. Penelitian menjelaskan bahwa hasil dari pengujian mendapatkan nilai performa dengan menggunakan AUC mendapatkan nilai accuracy sebesar 0,83 (83%). Dan itu termasuk ke dalam kategori Very Good Model (*Prasojo & Haryatmi, 2021*).

Irvan Mangolo (2022) melakukan penelitian menganalisis prediksi kelayakan nasabah kredit menggunakan algoritma Random Forest. Penelitian ini menyimpulkan bahwa random forest sangat cocok digunakan untuk melakukan analisa prediksi kelayakan data dengan output data dengan akurasi yang baik dikarenakan sudah melampaui tahap regresi dan klasifikasi menggunakan decision tree. Dengan mendapatkan precision 0.815, recall 0.782 dan akurasi sebesar 89% *(Panggabean, 2022)*.

2.2 Dasar Teori

2.2.1 Kecelakaan Lalu Lintas

Berdasarkan Undang-Undang Republik Indonesia No.22 Tahun 2009 tentang Lalu Lintas dan Angkutan jalan, kecelakaan lalu lintas didefinisikan sebagai kejadian di jalan yang terjadi secara tiba-tiba dan tanpa disengaja, melibatkan kendaraan dengan atau tanpa pengguna jalan lainnya dan juga dapat mengakibatkan korban jiwa maupun kerugian materil *(Muhammad Fakhruriza P et al., 2019), (M Iqbal Ryamizard, 2021)*.

2.2.2 Kaggle

Kaggle adalah sebuah platform berbasis cloud yang menyediakan lingkungan bagi para data scientist dan praktisi machine learning untuk berkolaborasi, berbagi dataset serta mengikuti kompetisi analisis data. Kaggle juga menjadi wadah bagi komunitas global untuk berbagi pengetahuan melalui forum diskusi, tutorial dan juga proyek open-source. Dengan berbagai fitur yang disediakan kaggle banyak dimanfaatkan dalam penelitian dan pengembangan model prediksi diberbagai bidang, termasuk analisis kecelakaan lalu lintas.

2.2.3 Google Colab

Google Colab (*Google Colaboratory*) ialah layanan berbasis cloud yang disediakan oleh google yang dipergunakan untuk pengguna menjalankan kode Python dalam lingkungan notebook Jupyter. Google Colab banyak digunakan dalam penelitian data science, machine learning dan kecerdasan buatan (AI) karena kemampuannya dalam menangani komputasi berbasis cloud tanpa perlu instalasi perangkat lunak tambahan dikomputer pengguna.

2.2.4 Python

Python ialah bahasa pemrograman tingkat tinggi yang dikembangkan oleh Guido Van Rossum dan pertama kali dirilis pada tahun 1991. Saat ini python menjadi salah satu bahasa pemrograman yang sangat populer. Keunggulannya terletak pada fleksibilitasnya sebagai bahasa pemrograman serbaguna, termasuk dalam penerapan machine learning dan deep learning. Python juga dilengkapi

dengan berbagai pustaka yang mendukung serta didukung oleh komunitas yang aktif dan luas, mengingat sifatnya yang open source *(Riziq sirfatullah Alfarizi et al., 2023)*.

Python juga bahasa pemrograman yang banyak dimanfaatkan sebagai sarana dalam menyelesaikan berbagai permasalahan terkait pengolahan data. Salah satu keunggulan python adalah ketersediaannya berbagai pustaka (library) yang dapat digunakan secara gratis. Salah satu penerapannya adalah dalam penyelesaian permasalahan optimasi khususnya pemrograman linier dengan metode simpleks *(Zahara et al., 2024)*.

2.2.5 Data Mining

Data mining merupakan teknologi yang memungkinkan otomatisasi dalam menemukan pola-pola penting dan tersembunyi dari kumpulan data berukuran besar. Proses ini membantu manusia dalam memahami pola yang ada dengan memanfaatkan teknik yang dapat diskalakan. Teknik data mining dapat digunakan untuk dua tujuan utama yaitu mining deskriptif, yang berfokus pada pengelompokan karakteristik umum dan untuk mining prediktif yang bertujuan membuat prediksi berdasarkan hasil klasifikasi data dalam jumlah besar *(Siregar et al., 2018)*. Data mining dapat didefinisikan sebagai serangkaian langkah yang dilakukan untuk mengekstrak dan menemukan pola penting dari suatu basis data, guna mengungkap informasi bernilai tambah yang sebelumnya tidak diketahui melalui cara manual (*Ismanto & Novalia, 2021*).

2.2.6 Analisis Korelasi

Analisis korelasi merupakan teknik statistik yang digunakan untuk menilai kekuatan dan arah hubungan antara dua variabel. Tujuannya adalah untuk mengetahui seberapa besar perubahan pada satu variabel terkait dengan perubahan pada variabel lainnya. Prosesnya meliputi pengujian normalitas setiap fitur guna menentukan apakah perhitungan akan menggunakan metode parametrik atau non-parametrik, penghitungan korelasi, serta pemilihan fitur berdasarkan ketentuan yang berlaku dalam analisis korelasi.

2.2.7 Machine Learning

Machine learning adalah suatu pendekatan berbasis komputer yang memungkinkan sistem untuk belajar secara mandiri melalui data tanpa memerlukan pemrograman eksplisit sebelumnya *(Nurkholifah et al., 2023)*. Machine learning merupakan sekumpulan algoritma pemrograman yang digunakan untuk meningkatkan kinerja komputer atau sistem dengan memanfaatkan data contoh yang telah ada sebelumnya (*Putra & Santika, 2020*).

Ada 7 langkah dalam machine learning yaitu mengumpulkan data, menyiapkan data masukan, menganalisis data masukan, melibatkan manusia, melatih algoritma, menguji algoritma dan memanfaatkannya (*Putra & Santika, 2020*). Prinsip dasar machine learning adalah menggunakan data untuk membangun model statistik yang biasanya dipakai oleh sistem untuk memprediksi masa depan berdasarkan data masa lalu yang dimasukkan atau untuk mempelajari pola yang ada dalam data dan salah satu keunggulan utama machine learning adalah kemampuannya untuk dimodifikasi dan beradaptasi dalam menanggapi perubahan data (*Putra & Santika, 2020*).

Pembangunan model machine learning melibatkan beberapa tahapan utama :

a. **Pengumpulan Data ,** Merupakan tahap yang melibatkan pengambilan data misalkan data kecelakaan lalu lintas dari berbagai sumber seperti database kepolisian, instansi transportasi atau open data pemerintah. Kualitas data yang diperoleh sangat menentukan keberhasilan hasil model prediksi.

b. **Pemilihan Fitur dan Target,** Proses seleksi atribut-atribut yang dianggap paling berpengaruh terhadap hasil prediksi dalam model dan pemilihan target yaitu pemilihan variabel dependen yang nantinya akan diprediksi. Pemilihan fitur yang tepat dapat meningkatkan akurasi dan efisiensi model dengan menghilangkan informasu yang tidak relevan atau berlebihan.

c. **Preprocessing,** Tahapan persiapan data sebelum digunakan dalam model machine learning. Proses ini mencakup pembersihan data dengan mengatasi nilai yang hilang (*missing value*), menangani data yang tidak konsisten, mengapus duplikasi dan juha melakukan normalisasi agar semua fitur memiliki skala yang seragam. Selain itu data kategori perlu dikonversi kedalam bentuk numerik menggunakan teknik encoding agar dapat digunakan oleh algoritma machine learning.

d. **Data Splinting,** Tahapan pembagian dataset menjadi dua terdiri dari data latih (*training data*) dan data uji (*testing data*). Data latih digunakan untuk melatih model, sedangkan data uji digunakan untuk mengukur performa model terhadap data yang belum pernah dilihat sebelumnya. pembagian ini bertujuan untuk menghindari overfitting dan memastikan bahwa model dapat menggeneralisasi dengan baik terhadap data baru.

e. **Model Training,** Tahapan yang menerapkan algoritma machine learning pada data latih untuk mengenali pola dan membangun aturan prediksi. Dalam proses ini model belajar dari hubungan antara fitur dan targer untuk menghasilkan prediksi yag paling akurat. Berbagai algoritma seperti Random Forest, Neural Network atau Decision Tree dapat digunakan tergantung dari jenis dan kompleksitas data yang diolah.

f. **Model Evaluation,** Performa model diukur menggunakan berbagai metrik evaluasi. Beberapa metrik yang umum digunakan antara lain accuracy, precision, recall dan F1-Score. Proses ini bertujuan memastikan bahwa model dapat memberikan hasil prediksi yang andal dan mengidentifikasi kemungkinan peningkatan melalui tuning parameter atau pemilihan fitur yang lebih optimal.

2.2.8 Random Forest

Random forest merupakan pengembangan dari metode Decision Tree yang memanfaatkan beberapa Decision Tree. Setiap Decision Tree dilatih menggunakan sampel individu dan setiap atribut dibagi pada pohon yang dipilih

dari subset atribut secara acak. Metode ini memiliki beberapa keunggulan, seperti meningkatkan akurasi hasil ketika terdapat data hilang dan tahan terhadap outliers serta efisien dalam penyimpanan data *(Arisusanto et al., 2023)*.

Prosedur ini digunakan untuk membuat pohon keputusan yang terdiri dari node root, ineer dan leaf dengan memilih atribut dan data secara acak berdasarkan aturan yang ditetapkan, node root ialah node yang terletak dipuncak pohon keputusan dan sering disebut sebagai akar (root) *(Pratama Yudha et al., n.d.)*.

Random forest membangun pohon keputusan dengan menggunakan sampel bootstrap yang berbeda untuk setiap pohonnya. Berbeda dengan metode regresi tradisional yang membagi setiap node berdasarkan pemisahan terbaik dari seluruh variabel, random forest memilih subset variabel secara acak disetiap node dan menggunakan pemisahan terbaik dari subset tersebut, algoritma ini memiliki dua parameter utama yaitu variabel yang dipilih dalam setiap subset acak disetiap node dan jumlah total pohon yang dibangun dalam model *(Fitri & Riana, 2022)*.

2.2.9 Teknik SMOTE

SMOTE (*Metode Synthetic Minority Over-Sampling Technique*) digunakan untuk mengatasi ketidakseimbangan kelas dalam data, Teknik ini bekerja dengan menghasilkan sampel baru untuk kelas minoritas dengan cara mensistensis instance baru berdasarkan kombinasi konveks dari sampel yang berdekatan, selanjutnya tingkat oversampling yang diperlukan ditentukan secara acak guna menyeimbangkan distribusi data *(Fahlapi et al., 2022)*. Metode SMOTE diterapkan untuk menangani ketidakseimbangan data serta mengurangi resiko overfitting yang terjadi akibat kecenderungan model dalam memprediksi kelas mayoritas *(Aryanti et al., 2023)*.

2.2.10 Evaluasi

Dalam evaluasi performa model klasifikasi metrik evaluasi berperan penting dalam menentukan tingkat keakuratan model dalam melakukan prediksi. Beberapa metrik yang umum digunakan dalam klasifikasi ialah Accuracy, Precision, Recall dan F1-Score. Metrik-metrik tersebut membantu dalam

memahami sejauh mana model dapat mengklasifikasikan data dengan benar serta mengukur keseimbangan antara kesalahan positif dan negatif.

a. Akurasi (*Accuracy*)

Merupakan metrik yang mengukur proporsi prediksi yang benar terhadap seluruh data yang diuji. Akurasi dihitung dengan rumus :

$$\text{Accuracy} = \frac{TP+TN}{TP+TN+FP+FN}$$

Yang dimana :

- TP (*True Positive*) ialah jumlah data positif yang diprediksi benar.
- TN (*True Negative*) ialah jumlah data negatif yang diprediksi benar.
- FP (*False Positive*) ialah jumlah data negatif yang diprediksi sebagai positif.
- FN (*False Negative*) ialah jumlah data positif yang diprediksi sebagai negatif.

b. Precision

Untuk mengukur proporsi prediksi positif yang benar dibandingkan dengan seluruh prediksi positif yang dilakukan model. Precision dihitung dengan rumus :

$$\text{Precision} = \frac{TP}{TP+FP}$$

c. Recall

Disebut juga Sensitivity atau True Positive Rate untuk mengukur seberapa banyak data positif yang benar-benar berhasil dideteksi oleh model dibandingkan dengan seluruh data positif yang ada. Adapun rumusnya :

$$\text{Recall} = \frac{TP}{TP+FN}$$

d. F1-Score

Merupakan metrik yang menggabungkan precision dan recall dalam satu nilai dengan menggabungkan rata-rata harmonik. Dengan rumus :

$$\text{F1-Score} = 2 \text{ x } \frac{Precision\ x\ Recall}{Precision+Recall}$$