Klasifikasi Tingkat Keparahan Kecelakaan Lalu Lintas Menggunakan Random Forest Classifier

I Gusti Ngurah Bagus Lanang Purbhawa[a1], I Gede Arta Wibawa[a2]

Program Studi Informatika, Fakultas Matematika dan Ilmu Pengetahuan Alam,
Universitas Udayana
Jalan Raya Kampus UNUD, Bukit Jimbaran, Kuta Selatan, Badung, Bali, Indonesia
[1]purbhawa.2208561108@student.unud.ac.id
[2]gede.arta@unud.ac.id

Abstract

Traffic accidents are a common problem that often occurs. Many factors cause and determine the severity of traffic accidents. These factors can include road conditions, weather, light conditions, driver age, and the cause of the accident. In this study, researchers will try to apply the Random Forest method to classify the severity of traffic accidents. The Random Forest method was chosen because of its excellent ability to handle high-dimensional data and tolerance for overfitting. The dataset used in this research was taken from Kaggle, consisting of 12316 records and 32 features covering various attributes related to traffic accidents. Before applying random forest, it is necessary to carry out a preprocessing stage on the dataset to remove irrelevant features, fill in empty values and divide the data into training and testing data. The results of this research show that Random Forest can produce a good level of in classifying the severity of traffic accidents with 92% accuracy. This shows the potential of this method as a useful tool in the analysis and prediction of traffic accidents. Therefore, this research makes a significant contribution to efforts to improve road safety."

Keywords: *Random Forest Classifier, Traffic accident, Machine learning, Data Classification, Supervised Learning*

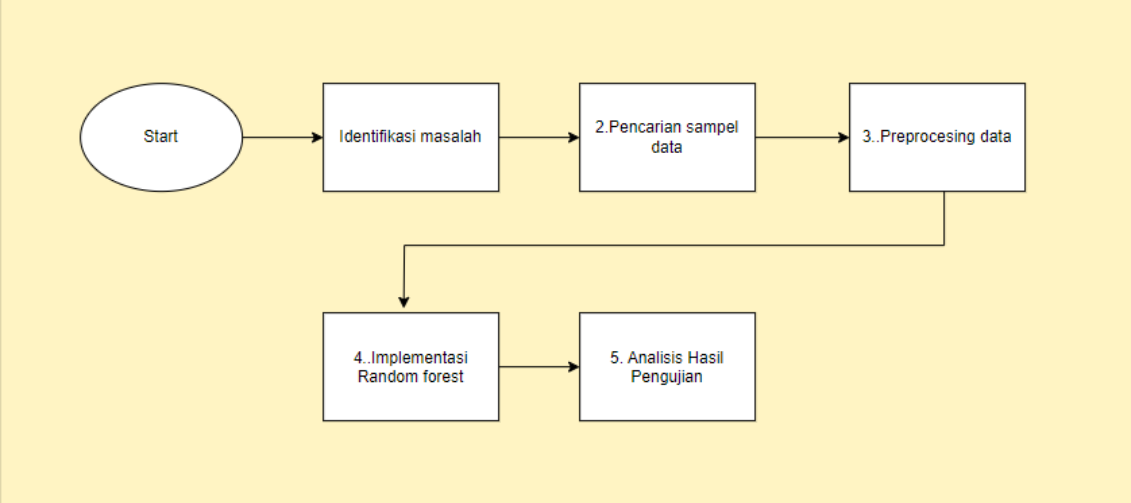
1. Pendahuluan

Kecelakaan lalu lintas merupakan masalah serius yang mempengaruhi keselamatan masyarakat dan mobilitas di jalan raya. Setiap tahunnya, ribuan kecelakaan terjadi di jalan raya, menyebabkan kerugian besar baik dalam hal korban jiwa maupun kerugian materi. Menurut laporan Organisasi Kesehatan Dunia jumlah kematian akibat kecelakaan lalu lintas lebih dari 1,25 juta orang dan setiap tahunnya kecelakaan non-fatal menimpa lebih dari 20–50 juta orang [1].hingga dengan memahami faktor-faktor yang mempengaruhi keparahan kecelakaan lalu lintas merupakan hal penting untuk merancang strategi pencegahan yang lebih efektif.Tingkat keparahan dalam kecelakaan lalu lintas dapat dipengaruhi oleh hal seperti, kondisi lingkungan sekitar, misalnya jalan yang licin, hujan, jalanan yang berbatu hingga faktor internal seperti pengalaman mengemudi hingga umur pengendara. Dalam konteks ini, analisis data dapat menjadi alat yang sangat berguna untuk memahami pola dan tren kecelakaan lalu lintas. Dengan memanfaatkan teknik analisis data dan machine learning, kita dapat mengklasifikasikan faktor-faktor apa yang memiliki pengaruh terhadap tingkat keparahan kecelakaan, sehingga dapat dilakukan tindakan pencegahan yang lebih tepat dan efektif.Algoritma Machine Learning dalam kasus ini dapat menemukan pola tersembunyi untuk memprediksi apakah tingkat keparahan kecelakaan itu fatal, serius, atau ringan [2]. Random forest merupakan algoritma yang sesuai dengan permasalahan klasifikasi ini yang dimana dalam beberapa kasus yang menangani data yang lebih kompleks dengan baik, termasuk data dengan banyak fitur dan non-linearitas jika dibandingkan dengan algoritma lain seperti SVM, algoritma random forest dapat menghasilkan nilai akurasi, presisi, dan recall yang lebih tinggi [3].

p-ISSN: 2986-3929
e-ISSN: 3032-1948

2. Metode Penelitian

2.1 Alur Penelitian



Gambar 1. Alur Penelitian

Adapun alur kerangka dari penelitian klasifikasi jenis kecelakaan berdasarkan keparahannya, yang memiliki penjelasan sebagai berikut:

a. **Identifikasi masalah**
Merupakan tahapan awal dari penelitian, di mana peneliti mengidentifikasi permasalahan yang akan di carikan solusi, dalam kasus ini peneliti akan mengklasifikasi faktor-faktor dalam pengaruh keparahan dalam kecelakaan lalu lintas

b. Pencarian sampel data
Pencarian data dilakukan secara sekunder yang dimana datanya diambil dari kaggle, yang dimana datanya data tingkat keparahan kecelakaan lalu lintas, data yang diambil memiliki format csv dengan 32 features di dalamnya

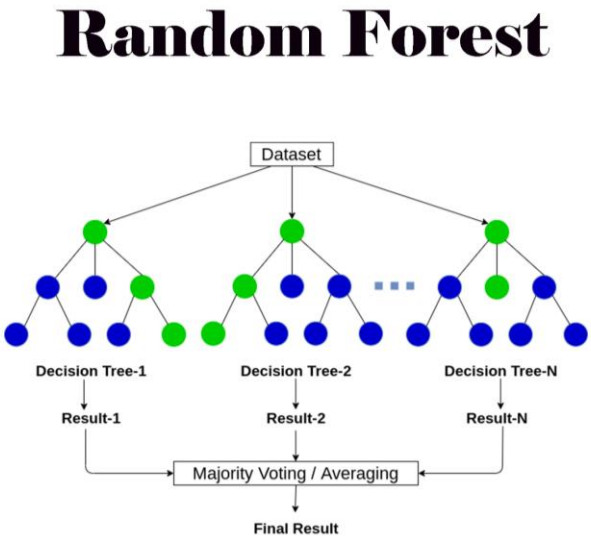
c. Preprocessing data
Preprocessing merupakan proses penting sebelum melakukan pemodelan dengan Random Forest, di tahap ini akan menghilangkan nilai yang tidak perlu, sehingga data yang diuji merupakan data yang baik.

d. Implementasi Random Forest
Pengimplementasian random forest akan menggunakan python dan google collabs sebagai tools.

e. Analisi Hasil Pengujian
Setelah dilakukan implementasi kita perlu melakukan evaluasi apakah hasil dari pengujian mendapatkan akurasi yang baik sehingga dapat menjadi solusi untuk penyelesaian kasus ini

p-ISSN: 2986-3929
e-ISSN: 3032-1948

2.2 Random Forest Classification



Gambar 2. Random forest klasifikasi

Klasifikasi dapat diartikan sebagai suatu pendekatan yang dikenal supervised algorithm dalam ranah data science. Tentu pada teknik klasifikasi memerlukan data yang berlabel [4]. Tipe data pada classification memerlukan data label, yang dapat berupa kategori biner, multi-kelas, nilai numerik, atau diekstrak dari teks, gambar, atau audio.Tanpa adanya label dalam data, maka algoritma classification seperti random forest tidak dapat belajar untuk mengidentifikasi pola dan membuat prediksi yang akurat.Algoritma classification belajar dari data berlabel untuk mengidentifikasi pola dan membuat prediksi, dan beberapa algoritma umum termasuk SVM, KNN, Decision Trees, dan Random Forests. Random Forest merupakan algoritma ensemble learning yang kuat dan populer, digunakan untuk menyelesaikan berbagai tugas machine learning.random Forest dipakai untuk masalah regresi dan klasifikasi dengan kumpulan data yang berukuran besar [5]. Algoritma ini bekerja dengan membangun banyak pohon keputusan acak dari subset data yang berbeda, dan hasil akhir dikonsensuskan untuk menghasilkan prediksi yang lebih akurat dan robust. Keunggulan utama Random Forest terletak pada kemampuannya menangani kumpulan data besar dengan efisien tanpa mengalami overfitting

3. Hasil dan Diskusi

Berdasarkan Alur penelitian dan kajian literatur, berikut adalah hasil yang diperoleh.

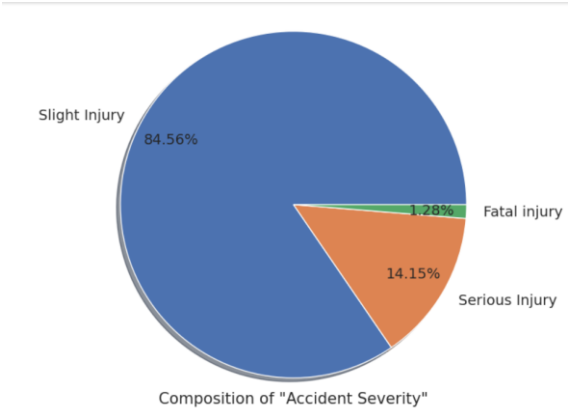
3.1. Pencarian Sampel data

Penelitian ini menggunakan data tingkat keparahan kecelakaan lalu lintas dari Kaggle.com yang di-publish oleh Saurabh Shanane. dataset ini terdiri dari 12316 baris/records dan 32 features/columns.

```
Index(['Time', 'Day_of_week', 'Age_band_of_driver', 'Sex_of_driver',
       'Educational_level', 'Vehicle_driver_relation', 'Driving_experience',
       'Type_of_vehicle', 'Owner_of_vehicle', 'Service_year_of_vehicle',
       'Defect_of_vehicle', 'Area_accident_occured', 'Lanes_or_Medians',
       'Road_allignment', 'Types_of_Junction', 'Road_surface_type',
       'Road_surface_conditions', 'Light_conditions', 'Weather_conditions',
       'Type_of_collision', 'Number_of_vehicles_involved',
       'Number_of_casualties', 'Vehicle_movement', 'Casualty_class',
       'Sex_of_casualty', 'Age_band_of_casualty', 'Casualty_severity',
       'Work_of_casuality', 'Fitness_of_casuality', 'Pedestrian_movement',
       'Cause_of_accident', 'Accident_severity'],
      dtype='object')
```

Gambar 3. Kolom pada dataset

p-ISSN: 2986-3929
e-ISSN: 3032-1948



Gambar 4. Komposisi Accident Severity pada data

Berdasarkan gambar diatas dapat dilihat kalau adanya ketidakseimbangan label tingkat keparahan kecelakaan lalu lintas. pada label fatal injury memiliki persentase yang sangat rendah hanya sekitar 1%. kita perlu menyeimbangkan records pada masing-masing label pada tahap preprocessing dan menghapus fitur-fitur yang tidak perlu.

3.2 Preprocessing Data

Data Preprocessing merupakan tahapan awal dalam data mining, biasanya dilakukan melalui cara eliminasi data yang tidak sesuai.

a. Mengecek nilai null pada kolom

```
null_df = df.isnull().sum().sort_values(ascending=False).to_frame()
null_df.columns= ["No of Null values"]
null_df["% of Null values"] = round(null_df["No of Null values"]/len(df)*100,2)
null_df[null_df["No of Null values"] > 0]
```

Gambar 5. Kode mengecek nilai null

Setelah dilakukan pengecekan null pada setiap kolom kita akan mencari kolom nilai yang memiliki lebih 2500 null

b. Mengecek baris yang duplikat

```
[62] ### Checkin for the duplicate values in the dataset
     df.duplicated().sum()

     0
```

Gambar 6. Mengecek baris yang duplikat

Pada Dataset tingkat keparahan kecelakaan lalu lintas tidak ditemukannya ada baris yang bernilai sama atau duplikat

p-ISSN: 2986-3929
e-ISSN: 3032-1948

c. Mengisi nilai yang kosong

```
#for categorical values we can replace the null values with the Mode of it
for i in categorical:
    df[i].fillna(df[i].mode()[0],inplace=True)
```

Gambar 7. Mengisi kolom kategorikal kosong

potongan kode tersebut akan mengganti nilai-nilai yang hilang (null) dalam fitur-fitur dengan modus (nilai yang paling sering muncul) dari setiap fitur tersebut.

d. Label Encoding

#	Column	Non-Null Count	Dtype
0	Day_of_week	12316 non-null	int64
1	Age_band_of_driver	12316 non-null	int64
2	Sex_of_driver	12316 non-null	int64
3	Educational_level	12316 non-null	int64
4	Vehicle_driver_relation	12316 non-null	int64
5	Driving_experience	12316 non-null	int64
6	Type_of_vehicle	12316 non-null	int64
7	Owner_of_vehicle	12316 non-null	int64
8	Area_accident_occured	12316 non-null	int64
9	Lanes_or_Medians	12316 non-null	int64
10	Road_allignment	12316 non-null	int64
11	Types_of_Junction	12316 non-null	int64
12	Road_surface_type	12316 non-null	int64
13	Road_surface_conditions	12316 non-null	int64
14	Light_conditions	12316 non-null	int64
15	Weather_conditions	12316 non-null	int64
16	Type_of_collision	12316 non-null	int64
17	Vehicle_movement	12316 non-null	int64
18	Casualty_class	12316 non-null	int64
19	Sex_of_casualty	12316 non-null	int64
20	Age_band_of_casualty	12316 non-null	int64
21	Casualty_severity	12316 non-null	int64
22	Pedestrian_movement	12316 non-null	int64
23	Cause_of_accident	12316 non-null	int64

Gambar 8. Label encoding

mengubah nilai-nilai dalam fitur kategorikal menjadi nilai numerik menggunakan metode Label Encoding. Ini dilakukan setelah analisis chi-kuadrat dilakukan pada dataset asli untuk memilih fitur-fitur yang paling relevan.

p-ISSN: 2986-3929
e-ISSN: 3032-1948

e. Fitur Selection

	features	Fscore	Pvalues
14	Light_conditions	16.082824	0.000322
20	Age_band_of_casualty	13.778413	0.001019
16	Type_of_collision	10.096323	0.006421
1	Age_band_of_driver	8.915392	0.011589
12	Road_surface_type	6.994806	0.030276
4	Vehicle_driver_relation	5.345345	0.069067
5	Driving_experience	4.499679	0.105416
8	Area_accident_occured	3.616540	0.163937
9	Lanes_or_Medians	3.281615	0.193824
18	Casualty_class	3.216860	0.200202
23	Cause_of_accident	3.193666	0.202537
11	Types_of_Junction	3.086487	0.213687
17	Vehicle_movement	2.200712	0.332753
15	Weather_conditions	1.149345	0.562889
7	Owner_of_vehicle	1.104262	0.575722
6	Type_of_vehicle	1.077671	0.583427

Gambar 9. Table score nilai fitur

Nilai-nilai ini dapat digunakan untuk menilai signifikansi setiap fitur dalam memprediksi variabel target. Semakin kecil nilai p, semakin signifikan hubungan antara fitur dan variabel targetnya.

f. Dummy Variable

```
#get_dummies
dummy=pd.get_dummies(df2[['Age_band_of_driver', 'Vehicle_driver_relation', 'Driving_experience',
                          'Area_accident_occured', 'Lanes_or_Medians', 'Types_of_Junction', 'Road_surface_type',
                          'Light_conditions', 'Weather_conditions', 'Type_of_collision', 'Vehicle_movement',
                          'Casualty_class', 'Age_band_of_casualty', 'Cause_of_accident']],drop_first=True)
dummy.head()
```

	Age_band_of_driver_31-50	Age_band_of_driver_Over 51	Age_band_of_driver_Under 18	Age_band_of_driver_Unknown	Vehicle_driver_relation_Other	Vehicle_driver_relation_Owner	Vehicle_driver_r
0	False	False	False	False	False	False	
1	True	False	False	False	False	False	
2	False	False	False	False	False	False	
3	False	False	False	False	False	False	
4	False	False	False	False	False	False	

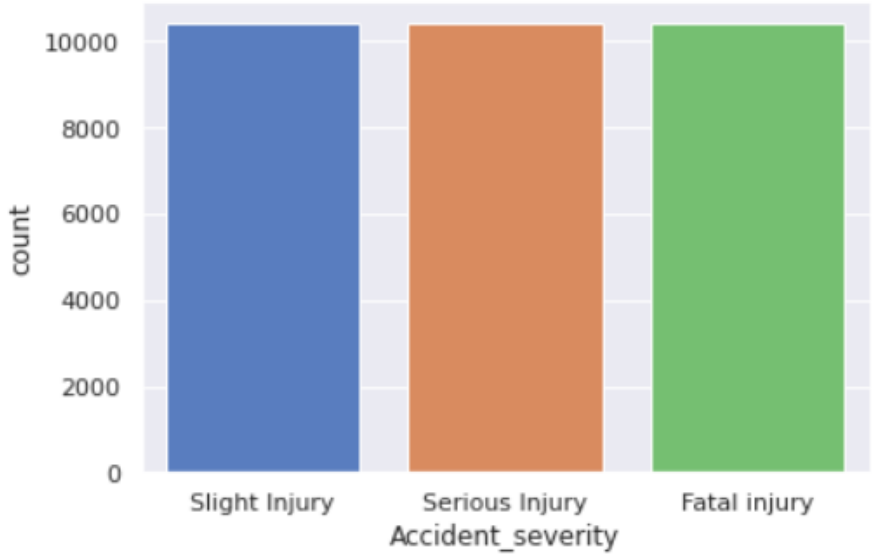
5 rows × 102 columns

Gambar 10. Output Dummy Variable

Setiap fitur kategorikal akan diubah menjadi kolom2 baru berdasarkan jumlah data unik pada kolom nya. Dengan menggunakan variabel dummy, Anda dapat lebih memahami bagaimana fitur kategorikal mempengaruhi variabel target.

p-ISSN: 2986-3929
e-ISSN: 3032-1948

g. Oversampling



Gambar 11. Diagram data setelah oversampling

Oversampling adalah teknik preprocessing data yang digunakan untuk meningkatkan jumlah contoh dari kelas minoritas dalam dataset yang tidak seimbang.Dataset tidak seimbang terjadi ketika terdapat perbedaan signifikan dalam jumlah contoh antara kelas-kelas di dalamnya. Hal ini dapat menyebabkan model machine learning menjadi bias terhadap kelas mayoritas dan menghasilkan performa yang buruk untuk kelas minoritas.dalam dataset ini kelas dengan label fatal injury memiliki jumlah yang sangat kecil sehingga perlu dilakukan oversampling

h. Split Train and test data

```
#converting data to training data and testing data
from sklearn.model_selection import train_test_split
#splitting 70% of the data to training data and 30% of data to testing data
x_train,x_test,y_train,y_test=train_test_split(xo,yo,test_size=0.30,random_state=42)

print(x_train.shape,x_test.shape,y_train.shape,y_test.shape)

(21871, 104) (9374, 104) (21871,) (9374,)
```

Gambar 12. Split data

Langkah terakhir dalam tahap preprocessing adalah membagi dataset kita menjadi data latih dan data uji, disini data akan dibagi menjadi data latih dan data uji, dengan perbandingan 70% banding 3%.

3.3 Random Forest Classifier

Pada tahap ini, algoritma Random Forest akan diimplementasikan untuk mengklasifikasikan data keparahan kecelakaan lalu lintas. Data ini terdiri dari 3 label, yaitu fatal injury, Serious Injury, dan Slight Injury. Random Forest bekerja dengan membangun banyak pohon keputusan acak dari subset data yang berbeda. RandomizedSearchCV disini digunakan untuk melakukan tuning parameter model Random Forest, dengan tujuan menemukan kombinasi parameter terbaik yang menghasilkan kinerja optimal. Data yang digunakan dibagi menjadi data pelatihan (training data) dan data pengujian (test data).

p-ISSN: 2986-3929
e-ISSN: 3032-1948

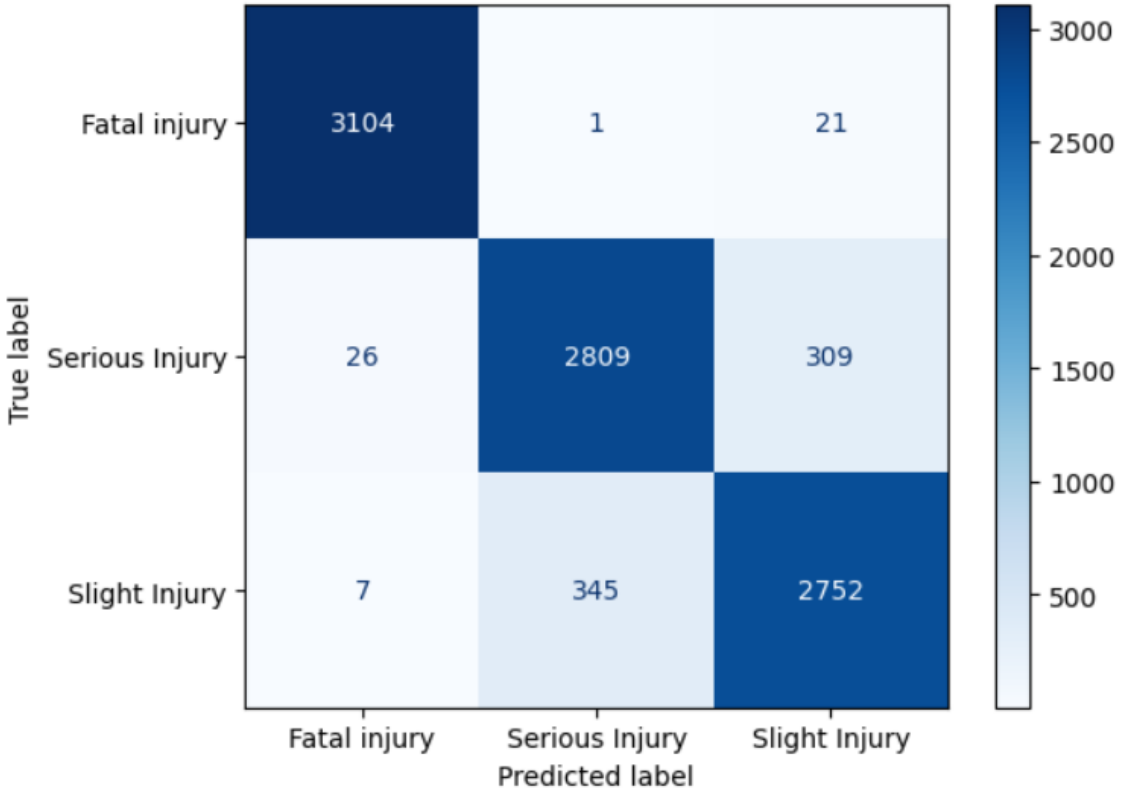
```
Best parameters found:  {'criterion': 'gini', 'max_depth': None, 'min_samples_split': 3, 'n_estimators': 189}
Best cross-validation accuracy:  0.9176537618394409
              precision    recall  f1-score   support

  Fatal injury       0.99      0.99      0.99      3126
Serious Injury       0.89      0.89      0.89      3144
 Slight Injury       0.89      0.89      0.89      3104

      accuracy                           0.92      9374
     macro avg       0.92      0.92      0.92      9374
  weighted avg       0.92      0.92      0.92      9374
```

Gambar 13. Akurasi random forest

Berdasarkan hasil tuning parameter, berikut adalah kombinasi parameter terbaik yang ditemukan adalah criterion: 'entropy', max_depth: None, min_samples_split: 5, n_estimators: 195.



Gambar 14. Confusion matrix display

Hasil akhir dari penelitian ini, Random Forest menunjukkan akurasi yang tinggi dalam mengklasifikasikan tingkat keparahan kecelakaan lalu lintas, yaitu 92%. Hal ini menunjukkan bahwa model mampu memprediksi kategori kecelakaan dengan tepat Implementasi Random Forest diharapkan dapat menghasilkan akurasi klasifikasi yang tinggi untuk membantu memahami pola dan faktor-faktor yang berkontribusi pada tingkat keparahan kecelakaan lalu lintas.

4. Kesimpulan

Hasil akhir dari penelitian ini dapat mengklasifikasi data keparahan kecelakaan lalu lintas dengan baik, hal itu dibuktikan dari akurasi yang didapat pada pengimplementasian algoritma random forest yang dimana menyentuh angka 92%. Algoritma ini menghasilkan akurasi tinggi dan robust terhadap data dengan membangun banyak pohon keputusan acak dan voting mayoritas sebagai prediksi akhir.model ini dapat digunakan untuk mendukung pengambilan keputusan dalam upaya

p-ISSN: 2986-3929
e-ISSN: 3032-1948

peningkatan keselamatan lalu lintas. Hasil evaluasi menunjukkan bahwa model ini sangat efektif dalam mengidentifikasi kecelakaan dengan tingkat keparahan fatal, serta cukup andal dalam mengklasifikasikan kecelakaan dengan tingkat keparahan serius dan ringan. dengan adanya penelitian ini diharapkan dapat membantu banyak orang baik para ahli hingga pihak terkait dalam mengambil keputusan pencegahan kecelakaan lalu lintas yang tepat.

Daftar Pustaka

[1] Tadesse Kebede Bahiru, Dheeraj Kumar Singh, and Engdaw Ayalew Tessfaw, “Comparative Study on Data Mining Classification Algorithms for Predicting Road Traffic Accident Severity,” Apr. 2018, doi: https://doi.org/10.1109/icicct.2018.8473265.

[2] S. Malik, Hesham El Sayed, Manzoor Ahmed Khan, and Muhammad Jalal Khan, “Road Accident Severity Prediction — A Comparative Analysis of Machine Learning Algorithms,” Dec. 2021, doi: https://doi.org/10.1109/gcaiot53516.2021.9693055.

[3] Ayu Aina Nurkhaliza and Arie Wahyu Wijayanto, “Perbandingan Algoritma Klasifikasi Support Vector Machine dan Random Forest pada Prediksi Status Indeks Mitigasi dan Kesiapsiagaan Bencana (IMKB) Satuan Kerja BPS di Indonesia Tahun 2020,” Jurnal Informatika Universitas Pamulang, vol. 7, no. 1, pp. 54–59, 2022, doi: https://doi.org/10.32493/informatika.v7i1.16117.

[4] “Teknik pre-processing dan classification dalam data science,” Master of Industrial Enginering, 2019. https://mie.binus.ac.id/2022/08/26/teknik-pre-processing-dan-classification-dalam-data-science (accessed May 10, 2024).

[5] Ilham Adriansyah, Muhammad Diemas Mahendra, Errissya Rasywir, and Yovi Pratama, “Perbandingan Metode Random Forest Classifier dan SVM Pada Klasifikasi Kemampuan Level Beradaptasi Pembelajaran Jarak Jauh Siswa,” Bulletin of Informatics and Data Science, vol. 1, no. 2, pp. 98–98, Nov. 2022, doi: https://doi.org/10.61944/bids.v1i2.49.

p-ISSN: 2986-3929
e-ISSN: 3032-1948

Halaman ini sengaja dibiarkan kosong