p-ISSN 2301–4156 | e-ISSN 2460–5719

Pengolahan Data Sensor Gerak Ponsel untuk Klasifikasi Karakteristik Mengemudi

Lisa Dinda Yunita[1], Ema Utami[2], Ainul Yaqin[3]

[1,2,3] Magister Teknik Informatika Universitas Amikom Yogyakarta, Jl. Ring Road Utara, Sleman 55281 INDONESIA (email: [1]lisadinday@students.amikom.ac.id, [2]ema.u@amikom.ac.id, [3]ainulyaqin@amikom.ac.id)

[Diterima: 7 Maret 2023, Revisi: 23 Juni 2023]
Corresponding Author: Lisa Dinda Yunita

INTISARI — Perilaku berkendara memiliki pengaruh signifikan terhadap keselamatan di jalan raya. Perilaku mengemudi yang tidak aman, seperti mengemudi dalam keadaan mabuk, kecepatan tinggi, dan penggunaan ponsel (*smartphone*), dapat menyebabkan kecelakaan serius dan kematian. Penelitian ini bertujuan untuk mengamati karakteristik berkendara dengan memanfaatkan data sensor gerak ponsel. Metode pengumpulan data melibatkan perekaman sensor gerak ponsel pengemudi selama perjalanan dan data tersebut diekspor dari sistem untuk diolah lebih lanjut. Tujuan utama penelitian ini adalah mengolah data dengan membuat model klasifikasi yang memiliki kinerja terbaik dalam mengolah data sensor gerak ponsel. Hasil penelitian ini diharapkan dapat menjadi model yang dapat diimplementasikan dalam mengatasi permasalahan keselamatan berkendara di masa mendatang. Selain itu, dengan memanfaatkan teknologi pendeteksi karakteristik pengemudi, kesadaran dalam berkendara dengan aman akan meningkat. Metode yang digunakan dalam penelitian ini adalah penambangan data (*data mining*) dengan menggunakan pemodelan klasifikasi pembelajaran mesin (*machine learning*) menggunakan metode *random forest* (RF), *support vector machine* (SVM), dan *decision tree* (DT). Hasil pengujian menunjukkan bahwa model RF memiliki kinerja terbaik dengan akurasi mencapai 91,22%. Selain itu, penelitian ini menemukan bahwa kecepatan merupakan faktor yang paling berpengaruh dalam mengidentifikasi perilaku berkendara yang aman atau tidak aman. Model klasifikasi yang dikembangkan menunjukkan potensi untuk meningkatkan efisiensi pengaturan lalu lintas dan berkontribusi pada transportasi yang lebih aman. Dengan memanfaatkan teknologi pendeteksi karakteristik pengemudi, diharapkan kesadaran akan praktik berkendara yang aman dapat meningkat, sehingga menciptakan lingkungan jalan yang lebih aman.

KATA KUNCI — Penambangan Data, CRISP-DM, Perilaku Mengemudi, Pembelajaran Mesin, Klasifikasi.

I. PENDAHULUAN

Kecelakaan lalu lintas dari tahun ke tahun mengalami peningkatan yang sangat signifikan. Kurang memadainya sarana transportasi menyebabkan masyarakat memilih menggunakan kendaraan pribadi, sehingga menyebabkan peningkatan jumlah kendaraan. Naiknya jumlah kendaraan ini meningkatkan risiko kecelakaan lalu lintas. Berdasarkan laporan bidang keselamatan jalan di masing-masing negara, World Health Organization (WHO) merilis beberapa faktor utama penyebab kematian [1]. Laporan tersebut menyatakan bahwa kecelakaan lalu lintas menduduki peringkat ke delapan sebagai penyebab tingkat kematian tertinggi. Dilaporkan juga bahwa pada tahun 2018 terjadi lebih dari satu juta kecelakaan lalu lintas [1].

Berdasarkan beberapa penelitian, di antara perilaku mengemudi yang tergolong tidak aman adalah mengemudi dalam keadaan mabuk, mengemudi dengan kecepatan tinggi, dan mengemudi sambil menggunakan ponsel. Jenis berkendara yang tidak aman ini sering kali menjadi faktor utama hilangnya konsentrasi, sehingga perhatian pengemudi tidak terfokus. Penelitian terkait pengawasan perilaku mengemudi dengan cara mengawasi pengemudi juga telah dilakukan. Hasil penelitian menunjukkan bahwa pengemudi yang diawasi cenderung mengemudi dengan relatif aman. Oleh karena itu, untuk mengurangi kecelakaan di jalan raya, penting untuk dilakukan pemantauan menggunakan komputer secara sistematis agar pemantauan dapat dilakukan secara *real-time* dan berkesinambungan. Pemantauan ini dapat membantu meningkatkan kesadaran pengemudi dan memastikan keselamatan berkendara yang lebih baik. Beberapa penelitian juga telah mengembangkan teknologi deteksi. Namun, implementasinya masih dapat dioptimalkan lagi karena bergantung pada aplikasi massal yang ada [2]. Selain mengakibatkan kerugian material, kecelakaan lalu lintas juga dapat menimbulkan korban jiwa. Oleh karena itu, diperlukan tindakan yang tepat untuk mengimplementasikan deteksi perilaku mengemudi untuk mengurangi risiko kecelakaan dengan harapan dapat memenuhi target global di masa depan yang mungkin ditetapkan dan mengurangi jumlah kecelakaan lalu lintas.

Organization for Economic Cooperation and Development (OECD) pada tahun 1988 menyatakan bahwa setiap tahun negara-negara mengeluarkan lebih dari 1 miliar dolar hanya di bidang transportasi [3]. *Intelligent Vehicle Highway System* (IVHS) melaporkan bahwa pada tahun 1991, di Amerika Serikat terdapat 41 ribu orang meninggal dan lebih dari 5 juta orang luka-luka akibat kecelakaan lalu lintas. Laporan tersebut juga menyatakan bahwa kecelakaan lalu lintas menyebabkan Amerika Serikat kehilangan 100 miliar dolar setiap tahunnya. Setelah mempelajari karakteristik permasalahan, IVHS telah menjadi *intelligent transportation system* (ITS) [4].

ITS merupakan teori yang diharapkan dapat mengatasi permasalahan transportasi dengan memanfaatkan teknologi. Peran teknologi bertujuan untuk meningkatkan keamanan lalu lintas, pemantauan data, dan sistem pengambilan keputusan secara *real-time*. Dalam banyak hal, ITS dapat mendukung sistem transportasi massal terdistribusi [5]. Penelitian ini dimulai dengan pemantauan transaksi dan operasional. Studi ini berfokus untuk mengatasi masalah pemantauan operasional. Model klasifikasi untuk mengemudi akan dihasilkan dengan memanfaatkan data perjalanan. Data yang dibutuhkan untuk mewujudkan teknologi ITS adalah data sensor gerak pada ponsel (*smartphone*) pengemudi. Metode statistik dapat digunakan untuk menghasilkan beberapa informasi penting

p-ISSN 2301–4156 | e-ISSN 2460–5719

pada ITS, antara lain informasi deteksi perilaku mengemudi dan deteksi mode kendaraan yang digunakan [6]. Data *global positioning system* (GPS), akselerometer, dan giroskop digunakan sebagai data penelitian. Penelitian ini melakukan proses pengklasifikasian deteksi keselamatan berkendara, yaitu pengemudi mengemudi secara biasa atau agresif, sehingga dapat disimpulkan bahwa pengemudi tersebut mengemudikan kendaraan dengan aman atau berbahaya. Pada pengolahan data, diperlukan mekanisme komputasi yang tinggi agar data menjadi pengetahuan bagi sistem dan menjadi parameter untuk mendeteksi cara mengemudi. Metode tersebut dapat menggunakan teknologi pembelajaran mesin (*machine learning*) [7].

Pembelajaran mesin adalah teknik komputasi yang mempelajari data menjadi informasi dan kemudian membentuk pengetahuan. Pembelajaran mesin memiliki beberapa proses, antara lain impor data, prapemrosesan, pemodelan, dan validasi. Penelitian ini mengimplementasikan pembelajaran mesin menggunakan teknik klasifikasi. Teknik ini dipilih karena berdasarkan karakteristik data yang diberikan, *dataset* memiliki label, sehingga dapat dijadikan sebagai nilai acuan [8]. Proses pelabelan *dataset* juga sudah melalui proses yang benar, mulai dari pembuatan aturan pelabelan hingga eksekusi pelabelan. Pada penelitian ini, data yang diterima telah diberi label dan tidak ada informasi mengenai proses pelabelan. *Dataset* yang digunakan siap diproses, dengan ekstensi *.csv*. Fitur yang tersedia dalam *dataset* antara lain *booking ID*, waktu, sensor giroskop x, y, dan z, sensor akselerometer x, y, dan z, serta data kecepatan. Jenis keseluruhan *dataset* adalah data teks. Teknik pembelajaran mesin adalah teknik klasifikasi dengan mengelompokkan data berdasarkan objek yang telah diberi label secara manual, sehingga model klasifikasi dapat mendeteksi cara berkendara pada perjalanan yang akan datang [9]. Penelitian terkait sebelumnya juga meneliti cara berkendara. Dari beberapa penelitian yang ada, kebanyakan diggunakan sensor giroskop, sedangkan metode yang digunakan adalah metode klasifikasi dengan menerapkan beberapa algoritma *support vector machines* (SVM) dan *random forest* (RF).

Pada evaluasi kinerja *machine learning algorithms* (MLA) sebelumnya, digunakan *artificial neural networks* (ANN), SVM, RF, dan *Bayesian networks* (BN) untuk mendeteksi cara berkendara menggunakan tiga sensor ponsel [10]. Sementara itu, fitur yang digunakan adalah empat fitur dataset (akselerometer, akselerometer linear, magnetometer, dan giroskop). Penelitian ini mengambil sampel perjalanan secara acak dengan menggunakan dua pengemudi yang berbeda. Berdasarkan hasil pengujian, giroskop merupakan sensor terbaik untuk mendeteksi perilaku berkendara, dengan RF merupakan MLA dengan kinerja terbaik, sehingga pada penelitian ini digunakan MLA [10].

Dataset penelitian yang digunakan adalah data perjalanan dengan jarak tempuh bervariasi dan cara mengemudi yang diambil secara acak. *Dataset* diperoleh dari data transaksi taksi *online* yang disediakan oleh salah satu perusahaan teknologi yang beroperasi di Indonesia untuk melayani transportasi melalui layanan pemesanan *online*. Penyedia aplikasi memperoleh data tersebut dengan mengambil rekaman data sensor secara acak pada ponsel pengemudi. Data sensor ponsel meliputi data sensor akselerometer, sensor giroskop, dan sensor GPS. Dalam penelitian ini, data dikumpulkan selama perjalanan individu, mencakup seluruh perjalanan dari lokasi penjemputan hingga lokasi penurunan. Data yang tercatat mencakup detail penting, seperti jarak yang ditempuh, rute spesifik yang diambil, dan total durasi setiap perjalanan. Semua informasi ini dikonsolidasikan dan dihubungkan dengan nomor identifikasi unik (*booking ID*) yang mewakili setiap pemesanan. Perlu ditekankan bahwa proses pengumpulan data terjadi secara *real-time* dan mengikuti pendekatan pengambilan sampel acak. Akibatnya, ukuran data untuk setiap pemesanan dapat bervariasi secara signifikan karena karakteristik yang berbeda dari tiap perjalanan. Mengingat variasi yang signifikan dalam ukuran data di antara tiap pemesanan, perlu dilakukan analisis data menggunakan metode statistik [11]. Analisis statistik memungkinkan peneliti untuk mendapatkan wawasan yang bermakna dari kumpulan data dan mengungkap pola, tren, dan korelasi penting di antara variabel yang tercatat. Dengan menggunakan teknik statistik yang tepat, penelitian ini bertujuan untuk membuat interpretasi berdasarkan informasi dan menyimpulkan hasil yang valid dari data yang terkumpul, berkontribusi pada pemahaman yang lebih komprehensif tentang perilaku dan karakteristik mengemudi.

Setelah dilakukan analisis awal terhadap *dataset*, ditemukan bahwa *dataset* yang tersedia tidak seimbang dan fitur yang tersedia terbatas. Hal ini dapat menyebabkan kinerja dari pengklasifikasi (*classifier*) RF mendapatkan nilai sensitivitas yang rendah. Masalah ketidakseimbangan *dataset*, dalam beberapa kasus, memengaruhi keakuratan data secara dramatis. Namun, beberapa algoritma klasifikasi, seperti RF, memiliki karakteristik yang cocok untuk mengatasi ketidakseimbangan *dataset* ini [12]. Untuk menguji reliabilitas klasifikasi RF dalam penelitian ini, dilakukan simulasi pemodelan. Jika hasil akurasi kurang memuaskan, beberapa teknik prapemrosesan akan diimplementasikan terlebih dahulu untuk mengatasi masalah ketidakseimbangan tersebut. Beberapa penelitian telah melakukan percobaan pada *dataset* yang tidak seimbang dan solusi yang ditawarkan adalah *random sampling*. Beberapa hasil eksperimen berhasil, tetapi beberapa tidak.

Seperti yang telah dijelaskan sebelumnya, *dataset* yang tidak seimbang dan terbatasnya fitur yang tersedia mengakibatkan sensitivitas yang rendah dan ketidakstabilan akurasi selama pengujian. Selain itu, masih banyak kesalahan dalam pendeteksian data perjalanan. Oleh karena itu, ekstraksi fitur juga dapat dilakukan, selain digunakannya metode *resampling*. Ekstraksi fitur bertujuan untuk menggabungkan beberapa fitur menjadi sebuah fitur yang diharapkan dapat mewakili semua karakteristik data [13]. Oleh karena itu, metode *resampling* dan ekstraksi fitur dicoba untuk mengatasi permasalahan yang terdapat pada basis data.

Analisis dilakukan dengan membandingkan hasil klasifikasi perilaku berkendara menggunakan *dataset* asli dan *dataset* yang telah melewati prapemrosesan data mulai dari pembersihan data yang memiliki nilai *null*, normalisasi nilai data, dan ekstraksi fitur. Kemudian, langkah selanjutnya adalah mencoba menerapkan teknik *resampling*. Hasil prapemrosesan *dataset* yang diperoleh selanjutnya digunakan dalam proses klasifikasi menggunakan beberapa algoritma klasifikasi yang umum digunakan. Hasilnya diharapkan dapat menjadi langkah awal penerapan ITS untuk mengatasi masalah kelalaian berkendara. Setelah didapatkan, model klasifikasi dapat diimplementasikan secara massal pada aplikasi yang sudah ada dan telah digunakan secara massal. Untuk tahap implementasi, perlu dilakukan penelitian lebih lanjut mengenai implementasi model dan interkoneksi langsung, dengan mengambil data pengujian secara *real-time* dari aplikasi transportasi *online*

p-ISSN 2301–4156 | e-ISSN 2460–5719

yang sedang beroperasi. Di sisi lain, penerapan ITS tidak hanya menyangkut teknologi, tetapi juga harus memperhatikan regulasi di suatu negara.

Kecerdasan buatan (*artificial intelligence*, AI) adalah kecerdasan yang ditambahkan ke sistem yang dapat dikelola dalam konteks ilmiah. Mesin mampu meniru pemikiran mirip manusia dengan meniru berbagai aktivitas manusia, seperti pengambilan keputusan, penyelesaian masalah, dan pembelajaran. Pendekatan kognitif ini memungkinkan mesin menyimulasikan proses pemikiran mirip manusia. Dalam domain pemikiran manusia, terdapat dua metode utama, yaitu introspeksi dan eksperimen. Introspeksi melibatkan bertanya dan berpikir kritis untuk mengungkap solusi saat dihadapkan dengan tantangan. Di sisi lain, eksperimen melibatkan percobaan berbagai pendekatan, metode, atau strategi untuk mengeksplorasi dan menemukan solusi. Dengan memanfaatkan introspeksi maupun eksperimen, mesin dan manusia dapat meningkatkan kemampuan penyelesaian masalah dan beradaptasi secara efektif dengan situasi yang berubah. Berperilaku seperti manusia dapat diartikan sebagai meniru kebiasaan manusia dalam menghadapi masalah. Konsep berperilaku seperti manusia ini menghadirkan uji Turing, yaitu menguji kemampuan manusia untuk mengenali mesin dengan membuat manusia berkomunikasi dengan entitas (mesin) melalui *teletype*. Misalnya, dalam 5 menit, seorang manusia tidak dapat mengenali entitas yang diinterogasi adalah manusia atau mesin. Maka, dalam hal ini, entitas tersebut lolos uji Turing dan dapat dikatakan sebagai sistem cerdas. Namun, entitas ini setidaknya harus memiliki kemampuan untuk mengenali suara, memahami bahasa manusia, melakukan sintesis ucapan dan representasi pengetahuan, merespons secara otomatis, melakukan pembelajaran mesin, melakukan penilaian, dan membuat keputusan. Penalaran memungkinkan komputer membuat persepsi, memberikan tanggapan, dan mengatasi suatu masalah. Cara untuk mencapai hal ini bagi sistem kecerdasan buatan adalah dengan memodelkan cara manusia berpikir dan merespons dalam kondisi ideal. Berperilaku rasional berarti membangkitkan sikap rasional dalam hal proses komputasi. Bersikap rasional adalah bertindak untuk mencapai tujuan dengan tetap mempertimbangkan kondisi dan pemahaman diri. Dalam hal ini, agen cerdas sebagai sistem komputer berperan dalam pengambilan keputusan terbaik dengan tetap mempertimbangkan situasi. Misalnya, ketika agen cerdas bermain catur, agen cerdas diharapkan melakukan langkah terbaik untuk memenangkan pertandingan.

Big data adalah istilah umum untuk setiap kumpulan data yang luas dan kompleks yang tidak dapat dianalisis oleh alat pemrosesan data tradisional. Namun, makin signifikan data yang dikumpulkan, makin besar kemungkinan dihasilkannya informasi baru yang belum pernah diketahui sebelumnya. Beberapa karakteristik *big data* dijelaskan sebagai berikut. *Big data* merupakan kumpulan data dengan volume yang sangat tinggi. Pada penelitian ini, *dataset* yang digunakan berjumlah sekitar 16 juta baris data yang akan diproses dengan kurang lebih 2 GB data. Analisis data sebanyak ini dapat lebih efisien jika dilakukan menggunakan cara tradisional. Aliran data harus dapat menerima dan mengolah data dengan kecepatan tinggi dan secara *real-time*. Metode kecepatan masih perlu digunakan pada tahap penelitian ini karena teknik pembelajaran masih menggunakan data *batch*. Namun, aliran data tersebut akan langsung dianalisis untuk tahap implementasi. *Variety* atau variasi adalah banyaknya jenis data yang beredar berdasarkan bentuk dan jenisnya. Pada data tradisional, data yang dikumpulkan umumnya berupa data yang terstruktur dan fit. Namun, pada teknologi *big data*, data yang diperoleh umumnya tidak terstruktur dan berasal dari berbagai sumber. Pada penelitian ini, variasi data terlihat karena sensor yang digunakan adalah beberapa sensor perjalanan, seperti giroskop, akselerometer, dan sensor GPS. Ketiga sensor ini memiliki format penulisan data yang berbeda.

Penambangan data (*data mining*) adalah bidang ilmu interdisipliner yang menggabungkan teknik pembelajaran mesin menggunakan beberapa algoritma unik untuk pengenalan pola, statistik, basis data, dan visualisasi untuk mengekstraksi informasi atau pengetahuan yang bermakna dari kumpulan data yang besar. Dengan teknik tertentu, informasi yang dihasilkan dari proses ini dapat digunakan untuk memprediksi hasil. Umumnya, tahapan penambangan data terkait dengan proses *knowledge discovery and data mining* (KDD). KDD adalah proses yang dibantu komputer untuk mengenali dan menganalisis kumpulan data besar dan mengekstraksi informasi dan pengetahuan yang berguna. Salah satu tahapan dalam keseluruhan proses KDD adalah penambangan data itu sendiri. Beberapa fungsi implementasi penambangan data, seperti metode klasifikasi, memiliki fungsi pengelompokan objek berdasarkan kelompok yang ada. Metode klasifikasi berbeda dengan *clustering*. Karakteristik metode klasifikasi membutuhkan data latih yang telah diberi label atau diklasifikasikan. Metode klasifikasi digunakan dalam penelitian ini.

ITS mengadopsi beberapa teknologi, seperti penentuan posisi, komunikasi, sistem informasi, dan kendali elektronik. Mengenai teknologi penunjang ITS, biasanya GPS berperan sebagai teknologi penentuan posisinya, dan *geographic information system* (GIS) berperan sebagai teknologi sistem informasinya. Sistem navigasi ITS dapat diklasifikasikan menjadi empat jenis, yaitu *autonomous ITS*, *fleet management ITS*, *advisory ITS*, dan *inventory ITS*. ITS memadukan faktor manusia (*people*), jalan (*roads*), dan kendaraan (*vehicles*) dengan memanfaatkan teknologi informasi mutakhir. ITS bertujuan untuk menerapkan teknologi canggih pada fasilitas transportasi agar lebih aman, lebih efisien, lebih berkembang, dan lebih ramah lingkungan. Berikut ini adalah ruang lingkup ITS. *Fleet management* mengelola kendaraan dari *dispatch center* melalui jalur komunikasi. Pada sistem ini, kendaraan yang bersangkutan dilengkapi dengan sistem penentuan posisi dan umumnya tidak dilengkapi dengan sistem peta elektronik. Kendaraan ini melaporkan posisinya ke pusat kendali, sehingga pusat kendali mudah mengatur pergerakan kendaraan tersebut. Selain memberikan petunjuk arah, pusat kendali juga bertugas menyediakan informasi yang dibutuhkan pengemudi kendaraan, seperti informasi cuaca dan kondisi lalu lintas. Sistem ini menggabungkan aspek pemosisisan dan sistem peta elektronik dari sistem *autonomous ITS* dengan aspek komunikasi dari arsitektur *fleet management ITS*. ITS bersifat *autonomous*, yang berarti bahwa *dispatch center* tidak mengendalikan sistem ini. Namun, pada saat yang sama, sistem ini merupakan bagian dari armada kendaraan yang menerima layanan dari pusat informasi lalu lintas. Dalam beberapa ITS, kendaraan tertentu berdiri sendiri sebagai *probe* lalu lintas, menyediakan bagi kendaraan lain (tidak ditentukan oleh pusat informasi lalu lintas) informasi terbaru tentang lalu lintas dan kondisi cuaca. Sistem ini biasanya terdiri atas kendaraan-kendaraan yang berdiri sendiri dan dilengkapi dengan kamera video digital untuk mengumpulkan data (lengkap dengan

p-ISSN 2301–4156 | e-ISSN 2460–5719

koordinat dan waktu pengambilan) yang berkaitan dengan jalan. Data ini diperlukan antara lain untuk inventarisasi jalan, pemeliharaan jalan, dan investigasi objek gangguan lalu lintas. Kendaraan yang digunakan juga dilengkapi dengan peranti pemosisisan (*positioning device*), *data logger*, dan tampilan data berupa peta elektronik. Sistem bantuan untuk pengemudian yang aman ini merupakan bentuk ITS yang sangat canggih. Kendaraan memiliki beberapa sensor yang dapat mengarahkan pengemudi untuk berkendara dengan aman. Penelitian tentang *soft computing* dalam berbagai ilmu cukup banyak, salah satunya dalam bidang keselamatan berkendara. Beberapa metode penambangan data sering digunakan untuk klasifikasi. Metode ini bertujuan untuk memprediksi hasil dari beberapa atau semua variabel untuk memprediksi kelas yang berisi dua nilai atau lebih [14]. Penelitian ini menggabungkan beberapa teknologi untuk mendapatkan *research gap* yang dapat dilakukan. Beberapa teori yang telah diadopsi antara lain adalah teori kecerdasan buatan, *big data*, penambangan data, dan ITS. Ditemukan celah antara irisan teori yang saling berhubungan. Oleh karena itu, selain teori pemetaan, penelitian ini juga mengacu pada penelitian-penelitian sebelumnya. Sebuah penelitian melakukan kajian deteksi mengemudi [10]. Evaluasi kuantitatif dilakukan terhadap kinerja algoritma klasifikasi MLA, antara lain SVM, RF, ANN, dan BN, untuk mendeteksi cara berkendara menggunakan data dari empat fitur sensor ponsel, yaitu akselerometer, akselerator linear, magnetometer, dan giroskop [10]. Dalam penelitian lainnya, *human activity recognition* (HAR) telah diteliti, dengan data yang digunakan adalah data sensor akselerometer dan sensor suara pada ponsel [15]. Penelitian ini mengklasifikasikan aktivitas seseorang dengan kondisi duduk, berjalan, atau berlari. Kinerja pengklasifikasi yang digunakan dalam penelitian ini, yang meliputi *multi-layer perceptron*, *decision tree*, dan SVM, dibandingkan. Penelitian ini juga mengangkat topik ketidakseimbangan *dataset*, yaitu dari data yang dipaparkan, terdapat perbandingan data duduk sebesar 26,4%, data berlari sebesar 1,9%, dan data berlari normal sebesar 45,91%. Dari masalah ini, diusulkan metode *oversampling*. Hasil penelitian menunjukkan bahwa metode *oversampling* dapat mengatasi masalah ketidakseimbangan *dataset*, dengan pengklasifikasi terbaik adalah *multi-layer perceptron*. Perbandingan kinerja klasifikasi untuk *F*1-*score multi-layer perceptron* dengan penerapan *oversampling* adalah sekitar 15% [15]. Selanjutnya, telah diteliti juga metode pemantauan berbasis pembelajaran mesin dengan metode klasifikasi multikelas untuk mengidentifikasi moda transportasi (mobil, sepeda, bus, jalan kaki, dan lari) [16]. Metode klasifikasi yang digunakan adalah *k-nearest neighbor* (KNN), SVM, dan RF, dengan mengolah data sensor ponsel, yaitu akselerometer, giroskop, dan sensor cahaya. Proses ekstraksi fitur dilakukan dari data sensor yang diberikan, dengan hasil akhir diperoleh 165 fitur. Metode pengklasifikasi RF menghasilkan kinerja terbaik berdasarkan hasil pengujian. Selain itu, karakteristik pengklasifikasi RF sangat bermanfaat dalam pemrosesan *dataset* dengan banyak fitur [16]. Penelitian lainnya meneliti aktivitas manusia, dengan data yang digunakan adalah data sensor dari ponsel berdasarkan sensor akselerometer dua sumbu (x,y) [17]. Penelitian ini melakukan klasifikasi mengemudi normal dan agresif dengan algoritma *fuzzy*. Dari hasil yang diperoleh, disimpulkan bahwa sistem klasifikasi *fuzzy* dapat menyajikan data grafik perjalanan berdasarkan klasifikasi yang diharapkan. Namun, penelitian tersebut juga menyatakan bahwa sensor akselerometer lebih dibutuhkan dalam mendeteksi cara berkendara. Oleh karena itu, perlu dilakukan pengujian dengan menggunakan sensor lain, seperti sensor giroskop [17]. Sebuah sistem dengan metode pembelajaran mesin untuk mendeteksi dan mengidentifikasi jenis perilaku mengemudi yang tidak normal berdasarkan data giroskop tiga sumbu dan akselerometer tiga sumbu pada sensor ponsel juga telah diteliti [18]. Penelitian ini menggunakan 20 sampel perjalanan. Dalam penelitian ini, yang diukur adalah cara berkendara yang tidak aman berdasarkan enam kriteria pergerakan, yaitu berliku (*weaving*), berbelok (*swerving*), tergelincir (*side slipping*), putar balik dengan radius sempit (*fast U-turn*), putar balik dengan radius lebar, dan pengereman mendadak. Metode klasifikasi yang digunakan adalah SVM dan jaringan saraf. Berdasarkan hasil penelitian, sensor giroskop memiliki tingkat akurasi pengklasifikasi yang paling baik di antara sensor lainnya dalam mendeteksi cara berkendara [18].

II. METODOLOGI

Secara umum, metode yang digunakan dalam penelitian ini adalah metodologi penambangan data. Penambangan data adalah penggalian atau ekstraksi informasi dari sejumlah besar data. Penambangan data mencakup berbagai analisis statistik dan teknik pembelajaran mesin yang digunakan untuk mencari pola dan hubungan dalam data yang mungkin tidak terlihat oleh mata telanjang. Tujuan penambangan data adalah untuk menemukan informasi bermanfaat yang dapat digunakan dalam membuat keputusan bisnis yang lebih baik dan meningkatkan efisiensi operasi data. Penambangan data juga dapat digunakan untuk memprediksi tren atau pola masa depan berdasarkan data historis [19]. Teknik pemrosesan data untuk akselerometer dan sensor giroskop untuk mendeteksi cara mengemudi dengan aman atau berbahaya bukanlah hal baru. Ada beberapa penelitian dengan tujuan serupa. Namun, metode atau skema yang digunakan dalam penelitian di bidang ini sangat bervariasi. Karena menyesuaikan dengan karakteristik *dataset* yang digunakan, pengembangan metode atau skema baru untuk mendeteksi jalur berkendara masih terbuka untuk penelitian lebih lanjut. Berdasarkan hasil tinjauan dari beberapa penelitian sebelumnya, hal ini memberikan kontribusi terhadap penelitian sebelumnya sebagai pembanding dalam sebuah penelitian. Penelitian dengan topik yang sama tetapi menggunakan *dataset* berbeda akan memberikan hasil dan kesimpulan yang berbeda. Penelitian ini menggunakan *dataset* dengan karakteristik yang berbeda dengan penelitian sebelumnya, seperti menggunakan data perjalanan transportasi *online*; belum pernah ada penelitian tentang keselamatan berkendara menggunakan *dataset* ini; dan memiliki sampel data perjalanan terbanyak dari studi penelitian sebelumnya, yaitu 20 ribu perjalanan dengan jumlah data lebih dari 16 juta baris data.

A. DATASET

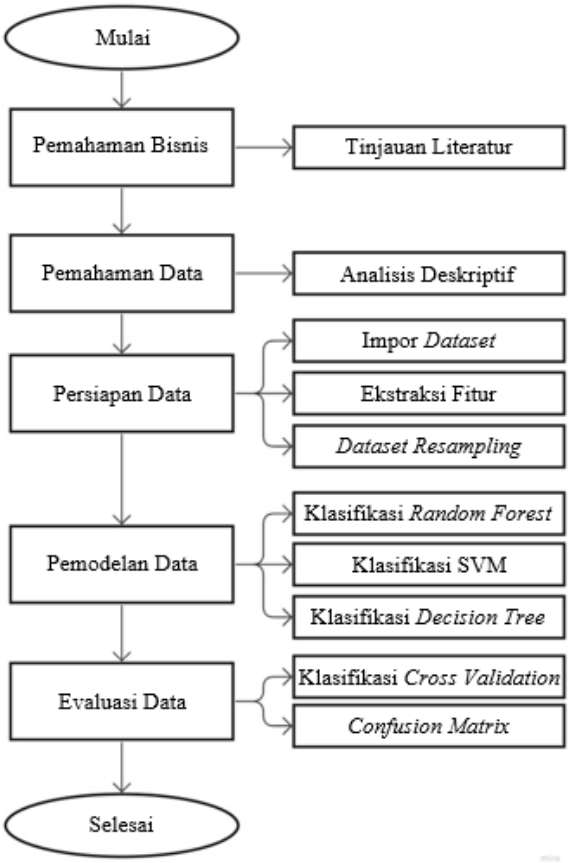
Pada penelitian ini, data yang digunakan adalah *dataset* dari sensor ponsel yang sudah diekspor ke dalam ekstensi *.csv*. *Dataset* tersebut merupakan data dari sistem transaksi transportasi *online* yang beroperasi di Indonesia. Beberapa hal krusial yang diperoleh dari tahap awal pemahaman antara lain adalah data diambil dari sistem transportasi *online*; data transaksi perjalanan menggunakan mobil yang di dalamnya terdapat data sensor ponsel, yaitu data sensor giroskop x, y, dan x, akselerometer x, y, dan z, data kecepatan, dan data waktu; terdapat jenis data pelatihan dan data pengujian; dan berdasarkan kapasitasnya, terdapat kurang lebih 16 juta baris data dari 20 ribu perjalanan.

p-ISSN 2301–4156 | e-ISSN 2460–5719

B. ALUR PENELITIAN

Secara umum, alur penelitian ini menggunakan model yang sudah ada dalam metodologi *cross-industry standard process for data mining* (CRISP-DM). CRISP-DM adalah salah satu model proses yang paling terkenal dan banyak digunakan dalam penambangan data. CRISP-DM menyediakan struktur dan panduan untuk mengelola proyek penambangan data dari awal hingga akhir, termasuk perencanaan, pengumpulan data, pemrosesan data, pengembangan model, dan evaluasi hasil. CRISP-DM sangat berguna dalam membantu mengelola proyek penambangan data secara efektif dan terstruktur. Namun, model proses ini fleksibel dan dapat diadaptasi sesuai dengan kebutuhan proyek tertentu [20].

Proses CRISP-DM terdiri atas lima tahap utama, yaitu pemahaman bisnis, pemahaman data, persiapan data, pemodelan data, dan evaluasi data. Teknik evaluasi yang digunakan adalah pengukuran hasil *confusion matrix* yang menghasilkan nilai akurasi dari model yang dibangun. Hasil akurasi model klasifikasi dibandingkan dengan kinerja terbaiknya, seperti yang digambarkan pada diagram alir dalam Gambar 1. Dalam percobaan pemodelan, dilakukan dua percobaan yang berbeda. Uji coba awal bertujuan untuk memodelkan tanpa menerapkan teknik prapemrosesan, sedangkan uji coba kedua melibatkan penerapan teknik prapemrosesan yang digunakan dalam penelitian sebelumnya. Di samping itu, selain menggunakan metode prapemrosesan dalam penelitian ini, beberapa algoritma klasifikasi yang umum digunakan juga dibandingkan.



Gambar 1. Alur penelitian.

1) PEMAHAMAN BISNIS

Pada tahap ini, dilakukan penggalian teori atau kajian pustaka terhadap pengolahan data sensor gerak ponsel. Meskipun penelitian sebelumnya yang relevan ditemukan selama tahap tinjauan literatur, penelitian ini melibatkan skenario eksperimental untuk membandingkan kinerja MLA pada *dataset* perjalanan. Kumpulan data ini memiliki fitur yang mirip dengan penelitian ini, seperti data akselerometer, data giroskop, dan data GPS. Karakteristik *dataset* yang digunakan dalam penelitian ini mirip dengan *dataset* dalam penelitian sebelumnya, sehingga memungkinkan adanya perbandingan antara model eksperimen yang digunakan pada penelitian ini dengan model eksperimen pada penelitian sebelumnya.

2) PEMAHAMAN DATA

Tahap ini meliputi pengumpulan data dan analisis deskriptif terhadap data penelitian. *Dataset* terdiri atas data sensor gerak ponsel yang dikumpulkan dari sistem transportasi *online*. Fitur yang tersedia antara lain *booking ID*, waktu, giroskop, sensor akselerometer, dan fitur kecepatan. Kumpulan data diambil dengan mengekspornya ke dalam format *.csv* dari data yang disediakan oleh aplikasi penyedia layanan. *Dataset* juga menyertakan label. Setiap perjalanan direkam secara *real-time* dan menghasilkan ukuran data yang bervariasi. Oleh karena itu, perlu dilakukan analisis data dengan menggunakan metode statistik.

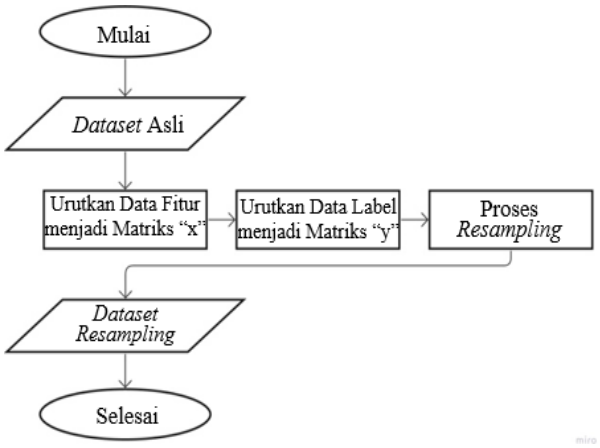
3) PERSIAPAN DATA

Tujuan utama dari tahap ini adalah untuk melakukan penyiapan data dengan prapemrosesan data lebih lanjut karena *dataset* yang digunakan sangat luas. *Dataset* yang tersedia dan akan diproses memiliki jumlah baris lebih dari 16 juta baris data dari 20 ribu perjalanan. Secara bertahap, data pada *dataset* mentah akan diproses. Data yang diproses adalah data sensor ponsel pengemudi dari setiap transaksi. Setiap transaksi memiliki karakteristik waktu dan panjang data yang berbeda, tergantung pada jarak rute perjalanan. Pada tahap penyiapan data dilakukan beberapa hal sebagai berikut.

Langkah awal dalam persiapan data adalah dengan mengimpor data dari *dataset* yang telah didapatkan. Seperti yang telah dijelaskan sebelumnya, terdapat dua jenis data yang digunakan, yaitu data fitur dan data label. Data label memiliki dua kelas, yaitu label 0 dan 1, dengan label 0 merepresentasikan perjalanan yang aman, sementara label 1 merepresentasikan perjalanan yang relatif berbahaya. Dalam proses ini, *dataset* fitur akan digabungkan dengan *dataset* label, sehingga setiap data fitur berdasarkan *ID* transaksi telah dikategorikan untuk setiap kelas kategori. Selanjutnya, untuk memfasilitasi penggunaan kembali dan mengurangi upaya komputasi, hasil gabungan data fitur dengan data label disimpan. Nantinya, ketika data tersebut akan digunakan, tidak perlu lagi dilakukan penggabungan *dataset*.

Setelah tahap penggabungan *dataset* fitur dan label selesai, langkah selanjutnya adalah ekstraksi fitur. Berdasarkan data yang tersedia, sensor giroskop x, y, dan z dapat digabungkan untuk membuat fitur baru yang disebut “giroskop”. Demikian pula fitur akselerometer x, y, dan z dapat digabungkan untuk membuat fitur baru yang disebut “akselerometer”. Terakhir, data waktu dan jarak dapat digunakan untuk membuat fitur baru yang disebut “kecepatan”. Fitur-fitur baru ini berasal dari fitur-fitur yang sudah ada. Sebelum masuk ke tahap klasifikasi, dilakukan ekstraksi ciri pada setiap perjalanan yang teridentifikasi. Mesin pengklasifikasi akan menggunakan fitur yang diekstrak ini untuk membedakan antara dua jenis label, yaitu aman dan berbahaya. Seperti disebutkan sebelumnya, kategorisasi data perjalanan ke dalam dua label ini memfasilitasi integrasi data selama tahap ekstraksi fitur. Data setelah mengalami proses ekstraksi fitur akan membentuk fitur

p-ISSN 2301–4156 | e-ISSN 2460–5719



Gambar 2. Alur *dataset resampling*.

baru yang memungkinkan pengklasifikasi mengidentifikasi dengan cepat label perjalanan yang sesuai.

Setelah ekstraksi fitur, langkah selanjutnya adalah *dataset resampling*. Tahapan ini merupakan proposal yang diajukan untuk mengatasi masalah ketidakseimbangan *dataset* yang ditemukan pada analisis deskriptif di awal penelitian. Dari *dataset* yang telah diperiksa, terlihat ketidakseimbangan jumlah label data untuk perjalanan aman dan berbahaya. Jumlah label data 0 lebih banyak daripada label data 1. *Dataset* yang tidak seimbang akan menjadi masalah bagi model karena jumlah data perjalanan aman yang dominan memudahkan objek perjalanan berbahaya dikenali sebagai perjalanan aman.

Gambar 2 menunjukkan alur penerapan *dataset resampling* yang digunakan dalam penelitian ini. Metode yang digunakan adalah metode reduksi dimensi yang umum digunakan dalam pembelajaran mesin, yaitu *principal component analysis* (PCA). Metode PCA digunakan untuk memetakan objek pada bidang koordinat. Dengan metode ini, sebaran data dapat diketahui. Maka, dengan menggunakan metode *resampling* yang sesuai pada setiap *dataset* beserta karakteristik sebaran data pada *dataset* penelitian dari hasil analisis berdasarkan metode PCA, dapat disimpulkan bahwa metode *resampling* pada *dataset* dalam penelitian ini dapat dicoba untuk melihat pengaruh ketidakseimbangan pada *dataset*, yaitu memengaruhi hasil klasifikasi pembelajaran mesin atau tidak.

4) PEMODELAN DATA

Pada tahap prapemrosesan *dataset*, didapatkan fitur-fitur yang memiliki signifikansi tinggi dalam membedakan dua kelas, yaitu perjalanan aman dan perjalanan berbahaya. Fitur-fitur ini nantinya akan digunakan dalam percobaan klasifikasi yang diusulkan. Saat ini, tidak ada model yang dapat mendeteksi keamanan berkendara. Dengan penelitian ini, model klasifikasi karakteristik berkendara dapat menjadi model yang dapat mendeteksi perjalanan yang tidak aman. Oleh karena itu, tahapan ini mencoba mengurangi kesalahan pendeteksian dengan menerapkan teknik klasifikasi. Hal ini akan meningkatkan kinerja dan akurasi deteksi pengenalan mengemudi dalam metode eksperimental yang diusulkan. Model klasifikasi ini memanfaatkan beberapa MLA, yaitu RF, SVM, dan *decision tree* (DT). Dilihat dari karakteristik sebaran menggunakan metode PCA, ketiga algoritma klasifikasi ini cocok digunakan pada *dataset* dengan karakteristik yang ada.

5) EVALUASI DATA

Tahapan pengujian dilakukan secara bertahap. Pada penelitian ini, dilakukan tiga skema percobaan, yaitu sebagai berikut.

- Menguji model klasifikasi RF pada kumpulan data dengan menerapkan metode pengolahan data.
- Menguji model klasifikasi SVM pada kumpulan data dengan menerapkan metode pengolahan data.
- Menguji model klasifikasi DT pada kumpulan data dengan menerapkan metode pengolahan data.

Setelah ketiga skema eksperimen metode klasifikasi tersebut dijalankan, dilakukan proses validasi untuk menilai kinerja model klasifikasi. Proses menggunakan validasi sebayak lima kali, yang umumnya dikenal sebagai *k-fold cross-validations*. Telah diamati dalam beberapa penelitian bahwa *k-fold cross-validation* menunjukkan varians yang lebih rendah dibandingkan dengan pemisahan data latih dan data uji sederhana. Varians yang lebih rendah ini menyiratkan bahwa titik data cenderung lebih dekat dengan nilai yang diharapkan. Karakteristik ini sangat berharga ketika berhadapan dengan data yang terbatas. Dalam *k-fold cross-validation*, data dibagi menjadi *k fold*, biasanya menggunakan nilai 5. Setiap *fold* data berfungsi sebagai data uji satu kali, sedangkan *fold* lainnya digunakan sebagai data latih. Proses ini diulang sebanyak *k* kali, dengan setiap *fold* berfungsi sebagai data uji tepat satu kali. *Confusion matrix* yang dihasilkan memberikan hasil akhir dari proses validasi ini. Dengan membandingkan *confusion matrix* dari beberapa eksperimen, dapat ditentukan model klasifikasi yang berkinerja terbaik dalam mengklasifikasikan data sensor gerak ponsel dan mengidentifikasi karakteristik mengemudi.

III. HASIL DAN DISKUSI

Data yang diolah dalam penelitian ini adalah data sensor, seperti giroskop, akselerometer, dan sensor GPS, yang dipasang pada ponsel pengemudi. *Dataset* tersebut diperoleh dari hasil pencatatan perjalanan yang dilakukan oleh penyedia jasa transportasi *online* yang beroperasi di Indonesia. Sebelum memasuki tahap pengolahan data, penelitian ini melakukan analisis data eksploratif. Tahapan ini dimaksudkan untuk mengenali dan mempelajari lebih jauh karakteristik data. Data yang akan diproses harus dipastikan tidak mengandung *noise* dan *missing value*. Dengan demikian, pengolahan dapat dilakukan dengan pendekatan yang tepat untuk mendapatkan hasil yang diinginkan. Hasil analisis data eksploratif telah memberikan representasi yang jelas. Data perjalanan memiliki 14 fitur. Informasi lebih lanjut mengenai data tersebut disajikan pada Tabel I.

1) PEMODELAN DATA

Parameter uji kinerja yang diukur dalam penelitian ini dijelaskan sebagai berikut. Akurasi adalah matriks yang mengukur banyaknya prediksi model benar dari jumlah total data yang diprediksi. Matriks ini umumnya digunakan ketika kelas target memiliki distribusi yang seimbang. Presisi mengukur banyaknya prediksi optimis yang benar dari total prediksi positif. Metrik ini memastikan bahwa model tidak mengklasifikasikan data berbahaya sebagai positif. *Recall* menghitung banyaknya hasil prediksi positif yang valid dari data positif lengkap. Metrik ini berguna untuk memastikan bahwa model dapat mengidentifikasi sebanyak mungkin data positif. *F*1-*score* adalah rata-rata harmonik antara presisi dan daya ingat. Matriks ini digunakan untuk memperhitungkan akurasi dan daya ingat secara bersamaan. Tahap validasi model dilakukan untuk mengevaluasi kinerja model statistik atau pembelajaran mesin terhadap data yang tidak digunakan dalam proses pelatihan model. Validasi model penting untuk memastikan model dapat menggeneralisasi dengan baik data

p-ISSN 2301–4156 | e-ISSN 2460–5719

TABEL I
DATASET FITUR

Data	Deskripsi
Accuracy	Nilai pengukuran akurasi GPS
Bearing	Pengukuran putaran dalam kompas digital
acc_x	Pengukuran percepatan digital pada sumbu x
acc_y	Pengukuran percepatan digital pada sumbu y
acc_z	Pengukuran percepatan digital pada sumbu z
Vector_accelerometer	Ektraksi fitur percepatan pada sumbu x, y, dan z.
gyroscope_x	Pengukuran giroskop digital pada sumbu x
gyroscope_y	Pengukuran giroskop digital pada sumbu y
gyroscope_z	Pengukuran giroskop digital pada sumbu z
Vector_gyro	Ektraksi fitur giroskop pada sumbu x, y, dan z.
second	Data waktu dalam satuan detik
Speed	Data kecepatan
Distance	Ektraksi fitur jarak nilai kalkulasi dari kecepatan dan waktu.
Label	Label *dataset* fitur sudah tersedia oleh pemilik data. Metode pelabelan dilakukan dengan pengambilan sampel berdasarkan penilaian kepuasan pelanggan dalam satu kali transaksi berjalan dari titik penjemputan hingga titik penurunan, dengan parameter *review* penumpang. Kategori kelas cara mengemudi yang dilakukan secara manual pada setiap perjalanan dengan asumsi kelas '0' aman dan kelas '1' berbahaya

yang tidak diketahui. Dalam penelitian ini, pendekatan *split validation* dan *cross-validation* mengikuti hasil pengujian kinerja model.

2) HASIL PENGUJIAN *SPLIT VALIDATION*

Metode ini memisahkan data menjadi dua bagian, yaitu data latih dan validasi data (data uji). Model dilatih pada data latih dan dievaluasi pada validasi data. Rasio pemisahan adalah 80:20. Hasil pengujian kinerja menggunakan metode *split validation* pada SVM, RF, dan DT ditunjukkan pada Tabel II.

Dari hasil pengujian model menggunakan *split validation*, diperoleh nilai akurasi, presisi, *recall*, dan *F*1-*score*. Berdasarkan perbandingan nilai kinerja seluruh parameter uji, RF memiliki kinerja terbaik. Hasil pengujian menujukkan nilai akurasi 76% dan nilai *recall* 100% pada model SVM, yang artinya model SVM memiliki kinerja yang baik dalam mengidentifikasi kelas positif (*true positive*), tetapi memiliki banyak *false negative*. Hal ini berarti bahwa model tidak dapat mengingat semua perjalanan kelas aman sebagaimana mestinya. Hal tersebut dapat terjadi jika model memilih untuk mengklasifikasikan perjalanan sebagai berbahaya (*false negative*) untuk memaksimalkan keakuratan data pelatihan.

3) HASIL PENGUJIAN *CROSS-VALIDATION*

Metode ini membagi data menjadi beberapa bagian (biasanya 5-10), yang disebut *fold*. Model dilatih pada data *k-fold* (*k*-1 *fold* sebagai data latih dan 1-*fold* sebagai validasi data)

TABEL II
HASIL PENGUJIAN *SPLIT VALIDATION*

Model	Akurasi	Presisi	*Recall*	*F*1-*Score*
SVM	76,00%	76,00%	100,00%	87,00%
RF	91,00%	92,00%	97,00%	94,00%
DT	88,00%	92,00%	92,00%	92,00%

TABEL III
HASIL PENGUJIAN *CROSS-VALIDATION*

Hasil Akurasi Model			
K-Folds	**SVM**	**RF**	**DT**
1	76,61%	91,30%	87,42%
2	76,31%	91,39%	87,40%
3	76,47%	91,45%	88,01%
4	76,70%	91,02%	87,93%
5	76,39%	90,94%	87,57%
Rata-rata	76,50%	91,22%	87,67%

dan dievaluasi pada *k-fold* yang tidak digunakan sebagai data latih maupun validasi data. Hal ini dilakukan sebanyak *k* kali, sehingga setiap bagian data digunakan sebagai set validasi. Terakhir, diambil nilai rata-rata hasil evaluasi *k* sebagai hasil akhir. Parameter *cross-validation* yang digunakan adalah 5. Dari hasil *cross-validation* didapatkan hasil seperti pada Tabel III. Dari hasil pengujian model menggunakan *cross-validation*, didapatkan rata-rata nilai akurasi SVM adalah 76,50%, rata-rata akurasi DT adalah 87,67%, dan rata-rata akurasi RF adalah 91,22%, yang merupakan kinerja terbaik.

4) HASIL PENGUJIAN STANDAR DEVIASI

Standar deviasi adalah ukuran distribusi data yang menunjukkan jauhnya data dari nilai rata-rata. Dalam model klasifikasi, standar deviasi dapat memberikan informasi tentang variabel atau sebaran data pada setiap kelas. Misalnya, jika standar deviasi fitur pada tipe tertentu kecil, nilai fitur tersebut cenderung saling mendekati dan kurang bervariasi. Sebaliknya, jika standar deviasi elemen pada kelas tertentu signifikan, nilai fitur tersebut lebih bervariasi. Dalam membuat model klasifikasi, informasi tentang variabilitas fitur di setiap kategori dapat membantu memilih dan mengevaluasi model yang sesuai. Misalnya, jika ada perbedaan yang signifikan dalam distribusi data antarkelas, model yang dibangun dapat mempertimbangkan untuk menggunakan metode yang berbeda untuk masing-masing kelas tersebut. Pada penelitian ini, diperoleh perbandingan standar deviasi pada *dataset* latih yang dilakukan dengan menggunakan SVM, RF, dan DT dengan menggunakan *tuning parameter* standar dan $k = 5$. Hasil perbandingan tersebut disajikan dalam Tabel IV.

Berdasarkan perhitungan, diperoleh nilai standar deviasi SVM adalah 0,0014, RF sebesar 0,0020, dan DT sebesar 0,0026. Hasil tersebut menunjukkan bahwa ketiga model menghasilkan nilai yang baik secara keseluruhan, dengan parameter di bawah 0,1. Tidak ada nilai deviasi standar "baik" yang universal untuk hasil *cross-validation*. Nilai ini dapat bervariasi, tergantung pada banyak faktor, seperti ukuran *dataset*, jumlah *fold* dari *cross-validation*, dan kompleksitas model. Namun, makin kecil standar deviasi, makin baik kinerja model. Sebagai patokan, standar deviasi kurang dari 0,1 menunjukkan bahwa model memiliki kinerja yang konsisten di seluruh validasi. Sebaliknya, standar deviasi yang lebih besar

p-ISSN 2301–4156 | e-ISSN 2460–5719

TABEL IV
HASIL PENGUJIAN STANDAR DEVIASI

Model	Standar Deviasi
SVM	0,0014
RF	0,0020
DT	0,0026

TABEL V
HASIL PENGUJIAN FITUR PENTING

Fitur	Kepentingan Fitur	Korelasi Fitur
Speed	0,14580	0,1175
acc_z	0,10225	0,0833
Accuracy	0,07678	0,0649
acc_y	0,06993	0,0509
second	0,06780	0,0918
Bearing	0,06721	0,0032
Distance	0,05161	0,1188
acc_x	0,03925	0,0300
Vector_Gyro	0,03147	0,0611
Vector_Acc	0,02027	0,0280
gyroscope_z	0,01681	0,0036
gyroscope_x	0,01600	0,0179
gyroscope_y	0,01563	0,0032

dari 0,1 menunjukkan bahwa model memiliki variasi kinerja yang lebih signifikan antar *cross-validation*. Namun, penting untuk diingat bahwa standar deviasi harus dinilai dengan skor akurasi dan metrik evaluasi lainnya untuk mendapatkan gambaran kinerja model yang lebih holistik. Selain itu, penyetelan parameter model juga dapat memengaruhi nilai standar deviasi dan kinerja model secara keseluruhan, sehingga perlu dilakukan dengan teliti.

5) FITUR PENTING

Berdasarkan hasil pemodelan, dapat dilihat juga fitur yang paling berpengaruh dalam mengklasifikasikan karakteristik berkendara. Penilaian fitur memilih subset elemen *dataset* yang paling relevan dan signifikan untuk analisis atau pembuatan model pembelajaran mesin. Ada beberapa alasan yang menyebabkan pemilihan fitur sangat penting. Jika kumpulan data memiliki terlalu banyak fitur yang tidak relevan, model pembelajaran mesin dapat menjadi terlalu rumit dan *overfitting*. Pemilihan fitur membantu mengurangi kompleksitas model dan mencegah *overfitting*. Model pembelajaran mesin yang digunakan dalam kumpulan data dengan fitur yang relevan dan signifikan biasanya memiliki kinerja lebih baik dibandingkan dengan fitur yang tidak relevan. Kumpulan data dengan lebih sedikit fitur akan membutuhkan lebih sedikit waktu dan biaya untuk diproses dan dianalisis. Kumpulan data dengan fitur yang relevan dan signifikan juga lebih mudah diinterpretasikan daripada yang tidak relevan. Hal ini memungkinkan penggunaan hasil analitik yang lebih mudah dan lebih akurat. Secara keseluruhan, pemilihan fitur sangat penting untuk menghasilkan model pembelajaran mesin yang akurat dan efisien untuk memproses dan menganalisis kumpulan data. Beberapa metode yang digunakan untuk mengetahui fitur-fitur penting dalam *dataset* penelitian antara lain sebagai berikut. Metode pertama adalah menggunakan model pembelajaran mesin untuk mengevaluasi pentingnya setiap fitur dalam kumpulan data dan memilih komponen yang paling relevan dengan variabel target, disebut kepentingan fitur. Lalu, metode kedua yaitu melakukan analisis korelasi antara setiap pasangan elemen dalam *dataset* dan menyukai fitur dengan korelasi rendah atau tidak signifikan, disebut korelasi fitur. Tabel V menunjukkan hasil perhitungan kedua fitur tersebut.

Hasil pada Tabel V menunjukkan bahwa fitur *Speed* merupakan fitur yang paling berpengaruh terhadap kelas. Hal ini menunjukkan bahwa kecepatan merupakan faktor utama yang memengaruhi karakteristik keselamatan berkendara.

IV. KESIMPULAN

Berdasarkan hasil percobaan pengolahan data yang dilakukan, *dataset* penelitian ini tergolong tidak terstruktur. Setelah digabungkan dengan data label, ternyata tidak semua perjalanan diberi label, sehingga dilakukan tahap prapemrosesan dengan membuang data yang tidak memiliki label. Kemudian, ditemukan pula bahwa *dataset* tidak seimbang, dengan komposisi 20:80 cenderung berisi label data aman. Dengan karakteristik tersebut, metode klasifikasi menghasilkan nilai kinerja yang cukup baik tanpa melakukan proses *resampling*. Hasil akhir dari penelitian ini menunjukkan bahwa RF memiliki hasil kinerja terbaik, dengan akurasi sebesar 91,22%. Hasil penelitian juga menunjukkan fitur atau variabel data yang memengaruhi karakteristik berkendara. Diperoleh hasil bahwa kecepatan merupakan faktor utama yang memengaruhi karakteristik berkendara. Selain kecepatan dan akselerasi, perilaku berkendara juga merupakan faktor kedua yang dapat menangkap karakteristik pengoperasian dengan aman atau serius. Kesimpulan yang diangkat dalam penelitian ini dapat digunakan untuk mengetahui pengaruh ketidakseimbangan *dataset* dan pengaruh ketidakseimbangan *dataset* tersebut terhadap kinerja klasifikasi pembelajaran mesin. Kemudian, dari pengolahan data sensor gerak pada ponsel, diharapkan karakteristik berkendara dapat diketahui. Akhirnya, model yang dibuat diharapkan dapat diimplementasikan untuk mendeteksi jenis perjalanan, aman atau berbahaya. Penelitian ini menjadi tahap awal penerapan ITS untuk keselamatan berkendara.

KONFLIK KEPENTINGAN

Penulis menyatakan bahwa tidak ada konflik kepentingan dalam penelitian dan penyusunan makalah ini.

KONTRIBUSI PENULIS

Konseptualisasi, Lisa Dinda Yunita, Ema Utami, dan Ainul Yaqin; metodologi, Lisa Dinda Yunita dan Ema Utami; analisis model, Lisa Dinda Yunita dan Ainul Yaqin; validasi, Lisa Dinda Yunita, Ema Utami, dan Ainul Yaqin; analisis, Lisa Dinda Yunita; validasi model, Lisa Dinda Yunita; kurasi data, Lisa Dinda Yunita; penulisan—draf asli, Lisa Dinda Yunita; penulisan—penelaahan dan penyuntingan, Lisa Dinda Yunita, Ema Utami, dan Ainul Yaqin; visualisasi, Lisa Dinda Yunita.

REFERENSI

[1] World Health Organization, "Global Status Report on Road Safety 2018," 2018, [Online], https://www.who.int/publications/i/item/9789241565684

[2] A. Pirayre, P. Michel, S.S. Rodriguez, dan A. Chasse, "Driving Behavior Identification and Real-World Fuel Consumption Estimation with Crowdsensing Data," *IEEE Trans. Intell. Transp. Syst.*, Vol. 23, No. 10, hal. 18378-18391, Okt. 2022, doi: 10.1109/TITS.2022.3169534.

[3] P. Kukuca dan M. Chlebovec, "Vehicle Location System," *2006 Int. Conf. Appl. Electron.*, 2006, hal. 101-104, doi: 10.1109/AE.2006.4382974.

[4] J.F. McLellan, M.A. Abousalem, dan T.R. Porter, "Quality Control in DGPS Separation Vector Systems," *Proc. 1994 IEEE Position Locat., Navig. Symp. - PLANS'94*, 1994, hal. 726-732, doi: 10.1109/PLANS.1994.303410.

[5] L. Zhu dkk., "Big Data Analytics in Intelligent Transportation Systems: A Survey," *IEEE Trans. Intell. Transp. Syst.*, Vol. 20, No. 1, hal. 383-398, Jan. 2019, doi: 10.1109/TITS.2018.2815678.

[6] M. Adnane, B.-H. Nguyễn, A. Khoumsi, dan J.P.F. Trovão, "Driving Mode Predictor-Based Real-Time Energy Management for Dual-Source Electric Vehicle," *IEEE Trans. Transp. Electrific.*, Vol. 7, No. 3, hal. 1173-1185, Sep. 2021, doi: 10.1109/TTE.2021.3059545.

[7] A. Fang, C. Qiu, L. Zhao, dan Y. Jin, "Driver Risk Assessment Using Traffic Violation and Accident Data by Machine Learning Approaches," *2018 3rd IEEE Int. Conf. Intell. Transp. Eng. (ICITE)*, 2018, hal. 291-295, doi: 10.1109/ICITE.2018.8492665.

[8] S. Agarwal, "Data Mining: Data Mining Concepts and Techniques," *2013 Int. Conf. Mach. Intell., Res. Adv.*, 2013, hal. 203-207, doi: 10.1109/ICMIRA.2013.45.

[9] J.A. Talingdan, "Performance Comparison of Different Classification Algorithms for Household Poverty Classification," *2019 4th Int. Conf. Inf. Syst. Eng. (ICISE)*, 2019, hal. 11-15, doi: 10.1109/ICISE.2019.00010.

[10] G. Castignani, R. Frank, dan T. Engel, "Driver Behavior Profiling Using Smartphones," *16th Int. IEEE Conf. Intell. Transp. Syst. (ITSC 2013)*, 2013, hal. 552-557, doi: 10.1109/ITSC.2013.6728289.

[11] T.-Y. Lee dan H.-W. Shen, "Efficient Local Statistical Analysis via Integral Histograms with Discrete Wavelet Transform," *IEEE Trans. Vis., Comput. Graph.*, Vol. 19, No. 12, hal. 2693-2702, Des. 2013, doi: 10.1109/TVCG.2013.152.

[12] L. Hakim dan S. Rochimah, "Oversampling Imbalance Data: Case Study on Functional and Non Functional Requirement," *2018 Elect. Power Electron. Commun. Controls, Inform. Seminar (EECCIS)*, 2018, hal. 315-319, doi: 10.1109/EECCIS.2018.8692986.

[13] F.P. Shah dan V. Patel, "A Review on Feature Selection and Feature Extraction for Text Classification," *2016 Int. Conf. Wirel. Commun. Signal Process., Netw. (WiSPNET)*, 2016, hal. 2264-2268, doi: 10.1109/WiSPNET.2016.7566545.

[14] Y. Xiao, Y. Liu, Y. Deng, dan H. Li, "Enhancing Multi-Class Classification in One-Versus-One Strategy: A Type of Base Classifier Modification and Weighted Voting Mechanism," *2021 Int. Conf. Commun. Inf. Syst., Comput. Eng. (CISCE)*, 2021, hal. 303-307, doi: 10.1109/CISCE52179.2021.9445948.

[15] K.T. Nguyen, F. Portet, dan C. Garbay, "Dealing with Imbalanced Data Sets for Human Activity Recognition Using Mobile Phone Sensors," *3rd Int. Workshop Smart Sens. Syst.*, 2018, hal. 1-10.

[16] A. Jahangiri dan H.A. Rakha, "Applying Machine Learning Techniques to Transportation Mode Recognition Using Mobile Phone Sensor Data," *IEEE Trans. Intell. Transp. Syst.*, Vol. 16, No. 5, hal. 2406-2417, Okt. 2015, doi: 10.1109/TITS.2015.2405759.

[17] A. Aljaafreh, N. Alshabatat, dan M.S.N. Al-Din, "Driving Style Recognition Using Fuzzy Logic," *2012 IEEE Int. Conf. Veh. Electron., Safety (ICVES 2012)*, 2012, hal. 460-463, doi: 10.1109/ICVES.2012.6294318.

[18] J. Yu dkk., "Fine-Grained Abnormal Driving Behaviors Detection and Identification with Smartphones," *IEEE Trans. Mobile Comput.*, Vol. 16, No. 8, hal. 2198-2212, Agu. 2017, doi: 10.1109/TMC.2016.2618873.

[19] S.M. Dol dan P.M. Jawandhiya, "Use of Data mining Tools in Educational Data Mining," *2022 Fifth Int. Conf. Comput. Intell., Commun. Technol. (CCICT)*, 2022, hal. 380-387, doi: 10.1109/CCiCT56684.2022.00075.

[20] F. Schäfer, C. Zeiselmair, J. Becker, dan H. Otten, "Synthesizing CRISP-DM and Quality Management: A Data Mining Approach for Production Processes," *2018 IEEE Int. Conf. Technol. Manage. Oper., Decis. (ICTMOD)*, 2018, hal. 190-195, doi: 10.1109/ITMC.2018.8691266.