

BAB I

PENDAHULUAN

1.1 Latar Belakang

Penipuan klaim asuransi mobil merupakan salah satu bentuk kejahatan yang merugikan perusahaan asuransi. Penipuan ini dapat terjadi melalui berbagai cara, seperti klaim palsu, klaim yang dibuat secara berlebihan, atau klaim yang dibuat untuk kerusakan yang sebenarnya sudah ada sebelumnya [1]. Menurut *Insurance Information Institute* pada tahun 2019, total klaim kendaraan yang diajukan di Amerika Serikat adalah sekitar 33,3 juta klaim. Dari total klaim tersebut, sekitar 3,0% atau sekitar 995.000 klaim ditolak karena adanya indikasi penipuan dalam klaim kendaraan [2]. Penipuan dalam klaim kendaraan dapat terjadi dengan cara memalsukan dokumen klaim atau melaporkan klaim untuk kerusakan yang sebenarnya tidak terjadi. Seiring dengan semakin banyaknya jumlah kecelakaan dan kerusakan mobil, risiko kecurangan juga semakin meningkat. Ketidakseimbangan kelas memberikan dampak yang tidak baik pada hasil klasifikasi dimana kelas minoritas sering disalah klasifikasikan sebagai kelas mayoritas. Hal ini dapat menurunkan nilai *accuracy* hasil klasifikasi [3]. Data klaim asuransi mobil memiliki 33 atribut pada atribut target yaitu *fraudfound* memiliki 14447 No dan 923 Yes dan berisi 15.420 data. *Data mining* dapat menjadi alat yang berguna dalam menganalisis data klaim asuransi mobil.

Data mining adalah proses penggalian dan analisis data besar-besaran untuk mengidentifikasi pola-pola atau hubungan yang bermanfaat dalam data [4]. *Data mining* telah menjadi topik yang penting dalam bidang teknologi informasi karena dapat membantu organisasi atau perusahaan untuk memahami lebih dalam tentang data yang mereka miliki, termasuk data asuransi. Dengan menggunakan teknik klasifikasi dalam data mining, perusahaan asuransi mobil dapat memanfaatkan potensi data yang mereka miliki, mengidentifikasi pola atau hubungan yang bermanfaat, membuat prediksi yang akurat, mengambil

keputusan yang terinformasi, dan mengurangi risiko kerugian finansial akibat kecurangan yang dilakukan oleh pelanggan mereka [5].

Beberapa penelitian terdahulu yang dilakukan untuk mendukung perbandingan algoritma *machine learning*. Pada tahun 2022 dengan menggunakan data *Vehicle Claim Fraud* yang membandingkan algoritma *Logistic Regression*, *K-Nearest Neighbor*, *Random Forest* dan *XGBoost* menunjukkan bahwa algoritma *Random Forest* memiliki tingkat akurasi tertinggi yaitu 98.5% dalam menangani *imbalance* data pada data *Vehicle Claim Fraud*. [5]. Kemudian pada tahun 2018 dalam menangani *imbalance* data menggunakan data *Acceptability Car* yang membandingkan teknik sampling *Random Oversampling* dan *Random Undersampling* menunjukkan teknik *Random Oversampling* dengan algoritma *Decision Tree* menghasilkan akurasi tertinggi yaitu sebesar 97.1% [6]. Selain itu pada tahun 2023 dengan menggunakan data *Algerian Forest Fires* dan teknik *Random Undersampling* dengan algoritma *Decision Tree* menghasilkan akurasi sebesar 94.52% dan nilai ROC sebesar 0,950 dalam menangani *imbalance* data [7]. Selanjutnya pada tahun 2018 dengan menggunakan data German kredit yang membandingkan *Random Forest*, *K-Nearest Neighbor*, *Naïve Bayes* dan *J48* menunjukkan bahwa algoritma *Random Forest* memiliki akurasi sebesar 90.10% dalam menangani *imbalance* data [8].

Dalam penelitian ini, akan dianalisis dan dibandingkan performa masing-masing algoritma dengan menggunakan teknik sampling *SMOTE Oversampling* dan Sampel (*Undersampling*) berdasarkan matrik *confusion matrix* seperti akurasi, presisi, *recall* dan T-test. Ketidakseimbangan kelas dapat memiliki dampak negatif terhadap hasil klasifikasi karena seringkali kelas minoritas akan salah diklasifikasikan sebagai kelas mayoritas [3] sehingga dalam penelitian ini bertujuan untuk membandingkan teknik sampling *SMOTE Oversampling* dan Sampel (*Undersampling*) pada algoritma *Decision Tree*, *Random Forest*, *Naïve Bayes*, *K-Nearest Neighbors* dan *Logistic Regression*. Dengan memilih algoritma *machine learning* yang paling efektif dalam

mengklasifikasikan label *fraud* pada data klaim asuransi mobil, perusahaan asuransi mobil dapat meningkatkan efektivitas memprediksi *fraud* dan mengurangi risiko kerugian finansial.

1.2 Rumusan Masalah

Langkah awal yang penting dalam penelitian komparasi algoritma adalah merumuskan masalah secara jelas dan spesifik agar peneliti dapat fokus pada topik yang relevan. Rumusan masalah pada penelitian ini memberikan panduan yang terperinci mengenai tujuan penelitian dan pertanyaan penelitian yang ingin dijawab. Rumusan masalah pada penelitian ini yaitu:

1. Bagaimana membangun model klasifikasi data klaim asuransi mobil dengan menggunakan algoritma *Naïve Bayes*, *Decision Tree*, *Random Forest*, *K-Nearest Neighbors* dan *Logistic Regression*?
2. Bagaimana perbandingan kinerja performa algoritma *Naïve Bayes*, *Decision Tree*, *Random Forest*, *K-Nearest Neighbors* dan *Logistic Regression* dalam klasifikasi data klaim asuransi mobil?
3. Bagaimana hasil visualisasi data yang merupakan hasil perbandingan performa dapat digunakan untuk membandingkan performa algoritma *Naïve Bayes*, *Decision Tree*, *Random Forest*, *K-Nearest Neighbors* dan *Logistic Regression*?

1.3.1 Batasan Masalah

Penting bagi peneliti untuk menetapkan batasan masalah yang jelas dalam penelitiannya. Dengan menetapkan batasan masalah peneliti dapat fokus pada aspek penting dari penelitian dan menghindari permasalahan yang tidak perlu yang dapat menghambat pencapaian tujuan penelitian. Dalam melakukan penelitian komparasi algoritma terdapat beberapa batasan masalah yaitu:

1. Data yang digunakan dalam penelitian ini adalah dataset klaim asuransi mobil yang diambil dari Kaggle.
2. Algoritma yang akan dibandingkan yaitu *Naïve Bayes*, *Decision Tree*, *Random Forest*, *K-Nearest Neighbors* dan *Logistic Regression*.

3. Penelitian menggunakan RapidMiner dengan versi 10.1 sebagai *tools* untuk klasifikasi data klaim asuransi mobil.
4. Penelitian ini menggunakan Tableau dengan versi 2022.2.0 sebagai *tools* untuk memvisualisasikan hasil performa algoritma.
5. Pada penelitian ini menggunakan teknik sampling SMOTE *Oversampling* dan *Undersampling*.

1.4 Tujuan Penelitian

Penelitian komparasi algoritma bertujuan untuk membandingkan kinerja beberapa algoritma. Penelitian ini sangat berguna untuk memilih algoritma terbaik yang dapat digunakan untuk memprediksi kecurangan. Dengan menetapkan tujuan penelitian, peneliti dapat fokus pada aspek penting dari penelitian dan menghindari permasalahan yang dapat menghambat pencapaian tujuan penelitian. Tujuan dalam penelitian ini yaitu:

1. Membangun model klasifikasi algoritma *Naïve Bayes*, *Decision Tree*, *Random Forest*, *K-Nearest Neighbors* dan *Logistic Regression* menggunakan data klaim asuransi dengan tools RapidMiner.
2. Membandingkan performa algoritma *Naïve Bayes*, *Decision Tree*, *Random Forest*, *K-Nearest Neighbors* dan *Logistic Regression* berdasarkan accuracy, precision, recall dan AUC.
3. Membandingkan performa berbagai algoritma *Naïve Bayes*, *Decision Tree*, *Random Forest*, *K-Nearest Neighbors* dan *Logistic Regression* dalam memprediksi klaim asuransi kendaraan dengan menggunakan visualisasi.

1.5 Manfaat Penelitian

Manfaat yang dapat dihasilkan dari penelitian ini antara lain adalah meningkatkan pemahaman kita tentang performa berbagai algoritma dalam memprediksi kecurangan serta membantu memilih algoritma yang terbaik dalam memprediksi manfaat penelitian ini, diantaranya.

1.5.1 Manfaat Penelitian Teoritis

Penelitian teoritis atau penelitian yang berfokus pada eksplorasi konsep, teori, dan prinsip-prinsip dasar tanpa melibatkan pengumpulan data empiris memiliki beberapa manfaat penting. Berikut adalah beberapa manfaat teoritis:

1. Membantu perusahaan membangun model klasifikasi untuk memprediksi kecurangan pada klaim asuransi mobil.
2. Memberikan solusi alternatif bagi perusahaan asuransi mobil dalam memprediksi kecurangan pada klaim asuransi mobil dengan menggunakan model algoritma *Naïve Bayes*, *Decision Tree*, *Random Forest*, *K-Nearest Neighbors* dan *Logistic Regression* yang optimal.
3. Memberikan perbandingan performa algoritma menggunakan visualisasi agar lebih mudah dipahami.

1.5.2 Manfaat Penelitian Praktis

Manfaat praktis dari penelitian adalah dampak langsung atau penerapan nyata dari konsep, penemuan, atau penelitian yang dilakukan. Penelitian yang dilakukan dengan tujuan mendapatkan pemahaman lebih dalam tentang suatu fenomena atau untuk menghasilkan solusi baru dapat memberikan manfaat praktis yang berharga. Manfaat praktis dari penelitian adalah sebagai berikut:

1. Membantu perusahaan khususnya kepada departemen investigasi atau departemen keamanan dalam memilih algoritma yang optimal dalam melakukan analisis data dan pengambilan keputusan dalam menentukan premi.
2. Mengurangi risiko kerugian dan penipuan yang dapat berdampak pada keuangan perusahaan asuransi dan keputusan perusahaan untuk mengurangi premi yang dibayar nasabah.

1.6 Sistematika Penulisan

Sistematika penulisan pada penelitian ini dirancang secara khusus untuk memudahkan pembaca dalam memahami isi penelitian dan mengikuti langkah-

langkah yang dilakukan oleh peneliti. Sistematika penulisan yang disusun secara terstruktur memberikan kerangka yang jelas bagi pembaca untuk mengikuti alur penelitian dengan mudah. Sistematika penulisan pada penelitian ini.

BAB 1 PENDAHULUAN

Bab 1 merupakan bagian yang penting karena berisi latar belakang penelitian yang menjelaskan konteks dan kebutuhan yang melatarbelakangi penelitian. Selain itu, bab ini juga berisi rumusan masalah yang secara jelas mengidentifikasi permasalahan yang ingin diselesaikan atau diteliti dalam penelitian ini. Pada bab 1 berisi latar belakang, rumusan masalah, tujuan penelitian, manfaat penelitian yang dibagi menjadi teoritis dan praktis dan sistematika penulisan.

BAB II LANDASAN TEORI

Bab 2 merupakan bab yang berisi referensi dan teori yang relevan dengan masalah yang dibahas. Dalam bab ini, peneliti menyajikan tinjauan pustaka yang mendalam mengenai beberapa konsep penting, termasuk machine learning, data mining, teknik sampling, teknik evaluasi dan validasi, serta framework dan algoritma yang relevan. Pada bab 2 berisi referensi dan teori yang relevan dengan masalah yang dibahas dalam penelitian ini. Terdiri dari machine learning, data mining, teknik sampling, teknik evaluasi dan validasi, framework dan algoritma.

BAB III METODOLOGI PENELITIAN

Bab 3 penelitian ini merupakan bab yang berfokus pada desain penelitian yang digunakan. Pada bab 3 menjelaskan tentang desain penelitian yaitu terdiri dari gambaran umum penelitian, metode penelitian, teknik pengumpulan data, populasi dan sampel, variabel penelitian dan teknik analisis.

BAB IV ANALISIS DAN HASIL PENELITIAN

Bab 4 merupakan bab yang berfokus pada hasil penelitian yang telah diperoleh. Dalam bab ini, peneliti menyajikan hasil penelitian secara detail dan memberikan interpretasi terhadap hasil tersebut, serta menjelaskan hubungannya dengan penelitian terdahulu yang telah dilakukan. Pada bab 4 berisi hasil penelitian yang diperoleh dan dijelaskan secara detail dan menjelaskan tentang interpretasi hasil penelitian dan hubungannya dengan teori.

BAB V SIMPULAN DAN SARAN

Bab 5 merupakan bab yang berfokus pada kesimpulan dari hasil penelitian yang telah dilakukan. Dalam bab ini, peneliti menyajikan kesimpulan secara singkat dan jelas, merangkum temuan utama yang telah diperoleh dari penelitian. Pada bab 5 berisi Kesimpulan dari hasil penelitian secara singkat dan jelas, serta memberikan rekomendasi atau saran untuk penelitian selanjutnya.

UMMN

UNIVERSITAS
MULTIMEDIA
NUSANTARA