BAB I PENDAHULUAN

PENDAHULUAN

1.1. Latar Belakang

Kemunculan teknologi internet telah sangat mempengaruhi aktivitas konsumen sehari-hari, dan banyak aktivitas *offline* kini telah bermigrasi ke lingkungan *online* (Athapaththu and Kulathunga, 2018). Abad ke-21 saat ini menjadi lebih terhubung dan mengubah perilaku belanja kita. Sejak dua dekade terakhir, orang cenderung lebih banyak membeli barang di *platform e-commerce* karena lebih nyaman dan memakan waktu lebih sedikit daripada belanja tradisional (Wagner, Chaipoopirutana and Combs, 2019). Kecenderungan dalam memenuhi kebutuhan secara *online* telah meningkat baru-baru ini. Telah menjadi kebutuhan pula bagi penjual untuk mengetahui pola dan niat dari berbagai jenis pelanggan *online* dengan cara menganalisis riwayat pelanggan (Kabir, Ashraf and Ajwad, 2019). Memahami perilaku dan niat dari pelanggan *online* sangat penting untuk pemasaran, meningkatkan pengalaman pelanggan, dan meningkatkan penjualan. Analisis niat beli *online* dari data pengalaman pembeli telah menjadi bidang penelitian yang muncul di bidang komputasi dan penambangan data (*data mining*) (Kabir, Ashraf and Ajwad, 2019). Data mining merupakan proses penggalian dan pencarian pengetahuan dan informasi yang bermanfaat dengan menggunakan algoritma/metode/teknik tertentu sesuai dengan pengetahuan atau informasi yang dicari (Buulolo, 2020). Algoritma/metode/teknik penggalian atau pencarian pengetahuan atau informasi mempunyai fungsi dan tujuan yang berbeda-beda, salah satunya adalah klasifikasi. C5.0 adalah algoritma klasifikasi pohon keputusan yang merupakan versi lanjutan dari C4.5 dengan kinerja *superior*. Algoritma C5.0 lebih baik daripada C4.5 pada akurasi, kecepatan, dan memori (Ross Quinlan, 2017).

Salah satu masalah dalam klasifikasi data mining adalah ketika kelas tidak diwakili secara seimbang. Masalah ini disebut ketidakseimbangan dataset (Japkowicz, 2013). Dataset terbagi menjadi kelas mayoritas dan kelas minoritas, dimana kelas mayoritas memiliki jumlah *instance* yang lebih banyak dibanding

kelas minoritas. Dataset niat beli online (*Online Shoppers Purchasing Intention Dataset*, 2018) adalah tidak seimbang, yang berarti bahwa hanya sebagian kecil yang berakhir dengan pembelian dan sebagian besarnya tidak. Masalah ketidakseimbangan kelas menjadi tantangan dalam proses klasifikasi dan menarik perhatian banyak peneliti (Haixiang *et al.*, 2017). Permasalahan ketidakseimbangan kelas (*class imbalance*) membuat proses belajar *classifier* sulit (Japkowicz, 2013). Selain itu, ketidakseimbangan kelas akan mempengaruhi kualitas data dalam hal kinerja klasifikasi (Gao, Khoshgoftaar and Wald, 2014) dan menyebabkan kinerja klasifikasi menjadi tidak optimal (Nurmasani and Pristyanto, 2021). Dalam proses klasifikasi, kelas minoritas sering salah diklasifikasikan, karena *machine learning* memprioritaskan kelas mayoritas dan mengabaikan kelas minoritas (Santoso, Wibowo and Himawati, 2019).

Salah satu cara untuk menangani dataset yang tidak seimbang adalah pendekatan level data yaitu *resampling*, seperti metode *random undersampling* dan metode *random oversampling*. Metode *random undersampling* terkadang dapat memberikan hasil yang tidak diinginkan karena sifatnya yang acak (Yildirim, 2016). Sedangkan metode *oversampling* dapat meningkatkan *precision* dan *recall* (Goodfellow, Bengio and Courville, 2016), namun terlihat terjadi penurunan pada akurasi pengklasifikasi (Sakar *et al.*, 2019). Pendekatan lain dalam menangani dataset yang tidak seimbang adalah pendekatan menggabungkan atau memasangkan (*ensemble*) metode, ada dua algoritma *ensemble-learning* paling populer, yaitu *boosting* dan *bagging* (Yap *et al.*, 2014). Algoritma *boosting* telah dilaporkan sebagai meta-teknik untuk mengatasi masalah ketidakseimbangan kelas (*class imbalance*) (Sun *et al.*, 2007). Algoritma *adaptive boosting* (*adaboost*) membangun pengklasifikasi kuat dengan cara mengombinasi-kannya dengan sejumlah pengklasifikasi sederhana (lemah) (Mulyati, Yulianti and Saifudin, 2017). Cara kerja algoritma *adaboost* adalah memberikan bobot yang sama pada setiap sampel pada awalnya. Setelah setiap klasifikasi, bobot hasil yang benar berkurang, dan bobot hasil yang salah bertambah. Proses ini diulangi hingga mencapai ambang batas atau jumlah siklus maksimum (Hao *et al.*, 2020).

Seleksi fitur dilakukan untuk mengidentifikasi *subset* atribut yang optimal (Thaseen, Kumar and Ahmad, 2019). Seleksi fitur dapat mengurangi dimensi data dan waktu pelatihan (AlShboul *et al.*, 2018). (Thakkar and Lohiya, 2021) menjelaskan bahwa mengikutsertakan seleksi fitur dapat memberikan dua keuntungan yaitu dapat membantu mengurangi *curse of dimensionality* dan mengkomputasikan fitur yang penting dapat membantu interpretasi data. Contoh metode untuk seleksi fitur seperti *wrapper* dan *filter*. *Wrapper based feature selection* berinteraksi dengan pengklasifikasi untuk mengevaluasi kegunaan fitur sehingga menghasilkan *subset* fitur yang lebih baik. Metode ini lebih lambat dibandingkan dengan metode *filter based feature selection* (Balasaraswathi, Sugumaran and Hamid, 2017). Salah satu metode *filter based feature selection* adalah *chi-square feature selection*, merupakan bagian dari *supervised feature selection*. *Chi-square* menggunakan uji independensi untuk menilai apakah fitur tersebut tidak bergantung pada label kelas (Li *et al.*, 2017). Untuk menyeleksi fitur, skor X^2 optimal yang dipilih. Jika suatu fitur memiliki skor X^2 yang rendah, maka fitur tersebut tidak bergantung pada kelas target yang menyiratkan bahwa fitur tersebut tidak informatif untuk mengklasifikasikan sampel data (Thakkar and Lohiya, 2021).

Berdasarkan uraian latar belakang di atas, maka penelitian ini diberi judul "**Optimalisasi Kinerja Klasifikasi Melalui Seleksi Fitur dan AdaBoost dalam Penanganan Ketidakseimbangan Kelas**"

1.2. Masalah Penelitian

Berdasarkan uraian pada latar belakang di atas, maka masalah dalam penelitian ini dapat diuraikan menjadi dua bagian yaitu identifikasi masalah dan rumusan masalah.

1.2.1. Identifikasi Masalah

Permasalahan di dalam klasifikasi data mining adalah ketidakseimbangan kelas, dimana kelas mayoritas memiliki jumlah *instance* yang lebih banyak

dibanding kelas minoritas. Dataset niat beli online (*Online Shoppers Purchasing Intention Dataset*, 2018) memiliki kelas yang tidak seimbang, yang berarti hanya sebagian kecil yang berakhir dengan pembelian, dan sebagian besarnya tidak. Permasalahan ketidakseimbangan kelas menyebabkan kinerja klasifikasi menjadi tidak optimal. Ketidakseimbangan kelas menyebabkan kelas minoritas sering salah diklasifikasikan, karena *machine learning* memprioritaskan kelas mayoritas dan mengabaikan kelas minoritas.

1.2.2. Rumusan Masalah

Berdasarkan identifikasi masalah yang ada, maka rumusan masalah dalam penelitian ini adalah bagaimana menghasilkan model yang dapat mengklasifikasi niat beli *online* (*Online Shoppers Purchasing Intention*) dengan kinerja optimal.

1.3. Tujuan dan Manfaat Penelitian

Adapun tujuan dan manfaat dari penelitian diuraikan pada dua poin di bawah ini.

1.3.1. Tujuan Penelitian

Tujuan dari penelitian ini adalah untuk membandingkan kinerja metode klasifikasi algoritma C5.0 dengan dan tanpa metode *chi-square feature selection* dan *adaptive boosting* (*adaboost*) dalam menangani ketidakseimbangan kelas pada klasifikasi niat beli *online* (*Online Shoppers Purchasing Intention)*.

1.3.2. Manfaat Penelitian

Hasil penelitian ini dapat memberikan manfaat yaitu sebagai berikut.

1. Sebagai referensi untuk menangani ketidakseimbangan kelas pada klasifikasi niat beli *online*.
2. Sebagai referensi untuk penelitian lebih lanjut baik di bidang yang sama maupun di bidang lainnya.

1.4. Ruang Lingkup (Batasan Masalah)

Dalam upaya untuk mencapai tujuan dalam penelitian ini, maka ruang lingkup penelitian ini adalah sebagai berikut.

1. Studi kasus (dataset) yang digunakan dalam penelitian ini adalah *Online Shoppers Purchasing Intention* yang berasal dari *UCI Machine Learning Repository*. Dataset tersebut dapat diakses pada *link* di bawah ini. https://archive.ics.uci.edu/ml/datasets/Online+Shoppers+Purchasing+Intention+Dataset
2. Adapun fitur yang digunakan akan diseleksi dengan metode *chi-square* menggunakan IBM SPSS Statistics 23.
3. Pengklasifikasian menggunakan C5.0 dengan dan tanpa *adaboost* menggunakan IBM SPSS Modeler 18.0.
4. Evaluasi berdasarkan *confusion matrix* untuk penilaian *accuracy*, *precision*, *recall/sensitivity/TPR*, *selectivity/specificity/TNR*, dan *f1 score*.

1.5. Metodologi Penelitian

Metodologi dalam penelitian ini adalah sebagai berikut.

1. Studi literatur
   Pada tahap ini dilakukan proses untuk memahami bagaimana proses klasifikasi pada beberapa penelitian sebelumnya.
2. Analisis masalah
   Pada tahap ini dilakukan analisis berdasarkan hasil studi literatur untuk mengidentifikasi masalah yang harus diselesaikan, data yang dibutuhkan, dan menentukan metode yang diusulkan untuk menyelesaikan masalah.
3. Perancangan model
   Pada tahap ini dilakukan perancangan model yang menggambarkan proses dari mengumpulkan dataset *Online Shoppers Purchasing Intention*, melakukan seleksi fitur dengan metode *chi-square* menggunakan IBM

SPSS Statistics 23. Kemudian menerapkan algoritma C5.0 dengan dan tanpa *adaboost* menggunakan IBM SPSS Modeler 18.0.

4. Pengujian

   Pengujian dilakukan untuk membandingkan algoritma C5.0 dengan dan tanpa *chi-square* dan *adaboost* dengan melihat nilai dari *confusion matrix*.

5. Menarik kesimpulan dari hasil pengujian
6. Menyusun laporan Tesis

1.6. Sistematika Penulisan

Sistematika penulisan laporan penelitian ini terdiri dari 5 bab, dimana secara garis besar masing-masing bab membahas hal-hal berikut ini. Bab 1 berisi penjelasan umum, masalah dan solusi yang sudah ada dan akan dilakukan. Bab 2 berisi studi literatur dan tinjauan pustaka terkait masalah dan metode yang berhubungan dengan penelitian yang akan dilakukan. Bab 3 berisi identifikasi masalah, langkah-langkah dari metode yang diusulkan, data yang digunakan, alat-alat penelitian dan metode analisis. Bab 4 berisi hasil yang diperoleh dari model yang dibangun dan pengujian yang dilakukan. Bab 5 berisi kesimpulan yang diperoleh dari hasil dan pengujian penelitian yang dilakukan dan saran yang dapat dilakukan untuk hasil yang lebih baik pada penelitian selanjutnya.