BAB I
PENDAHULUAN

1.1 Latar Belakang Penelitian

Kartu kredit merupakan alat pembayaran menggunakan kartu yang dapat membantu pemegang kartu dalam melakukan transaksi pembayaran atas suatu kegiatan ekonomi. Kegiatan transaksi ini dapat meliputi transaksi penarikan tunai maupun pembelanjaan. Kartu kredit tidak hanya dapat digunakan transaksi belanja di merchant *offline* melalui mesin EDC, namun juga dapat digunakan transaksi pada merchant *online* seperti situs *e-commerce*. Kewajiban pembayaran pemegang kartu dipenuhi terlebih dahulu oleh penerbit kartu yang selanjutnya dilakukan pembayaran pada waktu yang disepakati baik dengan pelunasan sekaligus ataupun dengan pembayaran secara angsuran. Kemudahan yang disediakan oleh kartu kredit tidak menjamin pemegang kartu terhindar dari modus kejahatan siber.

Modus kejahatan siber yang dapat terjadi pada kartu kredit adalah *carding*. Sebuah tindakan kejahatan yang dilakukan dengan melakukan transaksi menggunakan kartu kredit tanpa diketahui pemegang kartu sebenarnya. Pelaku kejahatan umumnya mendapatkan informasi pribadi kartu kredit secara ilegal baik dengan teknik *wiretapping, phishing, counterfeiting* maupun teknik ilegal lainnya. Segala transaksi yang dilakukan oleh pelaku bersifat *fraud* dan merugikan pemegang kartu. Dampak dari tindakan kejahatan ini berakibat pada kerugian finansial. Pemegang kartu harus membayar kewajiban kartu kredit yang dilakukan oleh orang lain dan telah dibayarkan terlebih dahulu oleh penerbit kartu. Selain digunakan untuk transaksi belanja secara ilegal, pelaku kejahatan juga menjual

informasi pribadi kartu kredit kepada pihak lain secara ilegal. Kondisi ini menjadikan potensi transaksi *fraud* meningkat dan mengakibatkan kerugian finansial.

Potensi transaksi *fraud* kartu kredit cukup besar apabila melihat dari kasus-kasus penjualan informasi kartu kredit dan tindakan kejahatan *carding* yang masih banyak terjadi. Kasus yang terjadi di India menunjukan bahwa informasi pribadi kartu kredit dari 70 Juta orang tersebar dan dijual pada situs Dark Web (Dubey, Mundhe, and Kadam 2020). Menurut laporan dari European Central Bank, nilai total transaksi fraud menggunakan kartu yang diterbitkan di kawasan SEPA dan dilakukan transaksi di seluruh dunia mencapai 1,87 miliar Euro pada tahun 2019 (European Central Bank 2021). Dalam laporan tersebut dinyatakan bahwa sebagian besar transaksi fraud kartu kredit terjadi pada kanal layanan digital seperti pembayaran online. Data lain dari Federal Trade Commission menyatakan bahwa pada tahun 2021 telah terjadi 88.354 transaksi *fraud* kartu kredit di Amerika Serikat dengan total nilai transaksi mencapai 181 juta US Dollar (Federal Trade Commission 2022).

Berdasarkan data yang telah dipaparkan, klasifikasi transaksi *fraud* perlu dilakukan untuk mencegah transaksi yang terindikasi sebagai *fraud*. Kerugian yang ditanggung pemegang kartu serta institusi jasa keuangan cukup besar. Disisi lain, pelaku kejahatan terus mengembangkan teknik dan strategi dalam melakukan transaksi *fraud* untuk memperoleh keuntungan. Perkembangan volume transaksi kartu kredit diikuti dengan teknik penipuan yang terus berkembang mendorong adanya pencegahan *fraud* yang lebih cepat dan akurat sesuai pola kejahatan terbaru.

Jika proses deteksi fraud dilakukan dengan proses verifikasi transaksi manual dengan bantuan petugas tentu tidak efektif apabila melihat jumlah transaksi yang terjadi. Maka dari itu, solusi klasifikasi transaksi *fraud* untuk kartu kredit perlu dikembangkan dengan menggunakan model *machine learning*.

Masalah umum yang dihadapi dalam pengembangan sistem deteksi *fraud* berkaitan dengan ketidakseimbangan dalam dataset transaksi *fraud*. Kondisi ini terjadi karena jumlah data transaksi yang bersifat *fraud* sangat kecil dibandingkan dengan jumlah transaksi normal dalam dataset. Hal ini mengakibatkan terjadinya permasalahan ketidakseimbangan data antara transaksi *not fraud* dan transaksi *fraud* yang berpengaruh pada kinerja model prediksi yang dikembangkan. Dari segi performa, dataset yang tidak seimbang dapat menghasilkan akurasi yang tinggi. Namun hal ini bersifat bias dan hasil prediksinya condong ke kelas mayoritas. Untuk mengatasi permasalahan tersebut, diperlukan strategi *resampling* yang mampu melakukan perbaikan distribusi data antara kelas *fraud* dengan *not fraud.*

Selain permasalahan dataset yang tidak seimbang, dataset dengan dimensi yang besar juga menjadi masalah dalam pengembangan model *machine learning*. Kondisi ini berdampak pada kebutuhan sumber daya komputasi yang semakin tinggi karena dimensi data yang diolah semakin besar juga. Maka dari itu, perlu dilakukan teknik pengolahan dataset untuk mengurangi dimensi data dengan tetap mempertahankan fitur yang memiliki kontribusi terhadap kelas prediksi. Dengan pengolahan data ini, diharapkan penggunaan sumber daya komputasi lebih efisien dengan tetap menghasilkan prediksi dengan akurasi yang baik pada dataset yang tidak seimbang.

Permasalahan terkait deteksi *fraud* kartu kredit, dataset tidak seimbang dan dimensi dataset transaksi *fraud* masih menjadi permasalahan yang terus dikembangkan oleh para peneliti. Untuk mengatasi masalah deteksi *fraud* kartu kredit, beberapa peneliti telah mengusulkan model prediksi dengan memanfaatkan algoritma *machine learning* yang berbasis *supervised learning* (Taha and Malebary 2020) (Lin and Jiang 2021) (Dileep, Navaneeth, and Abhishek 2021) (Sumanth et al. 2022) (Alfaiz and Fati 2022) (Madhurya et al. 2022) . Algoritma yang umum digunakan meliputi SVM, Naïve Bayes, Logistic Regression, Decision Tree, KNN, Random Forest, GBM, LightGBM, XGBoost, AdaBoost, serta CatBoost. Ada beberapa penelitian yang berfokus pada penyelesaian permasalahan dataset tidak seimbang dengan memanfaatkan teknik *resampling* agar memiliki kinerja deteksi yang lebih baik serta memperbaiki distribusi data. Teknik *oversampling* yang memiliki tujuan untuk membuat data sintetis pernah digunakan oleh beberapa peneliti untuk melakukan perbaikan rasio menggunakan metode seperti Random Oversampling, SMOTE dan ADASYN (Gupta et al. 2023) (Moreira et al. 2022) (Lu et al. 2020). Selain itu, peneliti lainnya juga mengusulkan penggunaan teknik *undersampling* yang mengurangi distribusi kelas mayoritas hingga seimbang dengan kelas minoritas (Afriyie et al. 2023) (Vuttipittayamongkol and Elyan 2020). Dari segi upaya mengurangi dimensi dataset, terdapat beberapa penelitian terkait deteksi *fraud* kartu kredit yang telah menggunakan beberapa teknik seleksi fitur seperti *Pearson's Correlation Coefficient*, K-Best, SVM-RFE (Mqadi, Naicker, and Adeliyi 2021) (Zhang et al. 2022) (Malik et al. 2022).

Penelitian oleh Berkmans dan Karthick pada tahun 2022 memaparkan bahwa data yang tidak seimbang menjadi suatu tantangan yang signifikan untuk mengembangan model klasifikasi transaksi *fraud* kartu kredit. Jumlah sampel positif (transaksi *fraud*) yang minim mengakibatkan adanya potensi hasil prediksi yang bias. Hal ini mendorong adanya proses *oversampling* yang perlu dilakukan untuk menghasil lebih banyak data sintetis dari kelas minoritas. Dalam penelitian ini, metode berbasis SMOTE direkomendasikan untuk melakukan proses *resampling* dengan kinerja *recall* tertinggi sejumlah 81% dan *precission* sejumlah 86%. Dari aspek *future research*, pendekatan *resampling* lainnya perlu dilakukan kajian lebih lanjut dengan tujuan untuk meningkatkan kinerja dari *classifier* yang digunakan (Berkmans and Karthick 2022). Hal ini sejalan dengan kekurangan metode SMOTE yang berpotensi membuat data sintetis yang *noisy* serta tumpang tindih (Grina, Elouedi, and Lefevre 2020) (Lu et al. 2020).

Permasalahan terkait dimensi dataset transaksi *fraud* pernah dibahas dalam penelitian yang dilakukan oleh Malik dkk. Dalam penelitian tersebut, peneliti memaparkan bahwa dataset transaksi *fraud* yang digunakan memiliki dimensi yang cukup besar dengan 432 variabel independen dan dependen. Untuk menghemat sumber daya komputasi, maka peneliti menggunakan metode seleksi fitur *correlation-based filter* serta SVM-RFE. Model prediksi yang menggunakan seleksi fitur dan *classifier* AdaBoost dan LGBM ini berhasil mendapatkan kinerja sejumlah 82% AUC. Dari segi aspek *future research*, terdapat pekerjaan lanjutan terkait penggunaan metode seleksi fitur lain yang bertujuan untuk mendapatkan

kinerja prediksi yang lebih optimal dalam studi kasus prediksi transaksi *fraud* kartu kredit (Malik et al. 2022).

Berdasarkan pemaparan masalah tersebut, maka peneliti mengusulkan penelitian dengan judul "Klasifikasi Transaksi Fraud Kartu Kredit Menggunakan Teknik Resampling dan Seleksi Fitur SVM-RFECV". Motif kejahatan *fraud* khususnya kartu kredit yang terus berkembang mendorong peneliti untuk terus mengembangkan strategi pengembangan model prediksi transaksi *fraud* kartu kredit. Tantangan terkait data tidak seimbang serta dimensi dataset yang cukup besar perlu disiapkan solusi agar mampu menghasilkan model prediksi yang mampu melakukan klasifikasi transaksi *fraud* dengan handal. Berdasarkan penelitian yang dilakukan oleh Malik dkk, teknik seleksi fitur SVM-RFE memiliki keterbatasan saat ingin mengetahui jumlah fitur paling optimal. Peneliti harus menetapkan jumlah fitur yang ingin disisakan dari keseluruhan jumlah fitur yang ada. Hal ini yang akan membawa ketidakpastian pada hasil akhir dari jumlah yang diperlukan. Untuk mengatasi ini, peneliti mengusulkan penggunaan *cross validation* yang dikemas dalam metode SVM-RFECV untuk seleksi fitur. *Cross validation* mampu mengatasi masalah ini dengan menahan sebagian data dalam data pelatihan sebagai data uji (Xie et al. 2018). Peneliti lain telah menggunakan RFECV untuk melakukan eliminasi fitur dan menemukan subset paling optimal untuk studi kasus prediksi DNA, prediksi penerimaan kembali pasien ICU, serta prediksi hunian (Khanal et al. 2021) (Qiu et al. 2021) (Yang et al. 2021). Hal ini sejalan dengan penelitian sebelumnya yang dilakukan oleh Malik dkk yang mengarahkan penelitian selanjutnya untuk melakukan optimalisasi teknik seleksi

fitur SVM-RFE (Malik et al. 2022). Dalam hal ini, teknik RFECV memanfaatkan algoritma SVM untuk melakukan seleksi subset paling optimal.

Dari segi permasalahan data tidak seimbang, peneliti menggunakan metode ADASYN yang merupakan metode perbaikan dari SMOTE. Hal ini mempertimbangkan kekurangan SMOTE yang memiliki potensi menghasilkan data sintetis yang *overlapping* karena melakukan pengembangan data sintetis pada setiap tetangga terdekat. Metode ini juga dipertimbangkan kaitannya dengan kebutuhan *oversampling* terhadap kelas minoritas berdasarkan sebaran profil transaksi *fraud* yang pernah terjadi. ADASYN melakukan pembuatan data sintetis berdasarkan distribusi tertimbang untuk sampel data berbeda dari kelas minoritas yang bergantung pada tingkat kesulitan sampel tersebut dipelajari dengan model (Moreira et al. 2022). Pemanfaatan metode ini juga sejalan dengan aspek *future research* terkait pemanfaatan metode *resampling* alternatif selain SMOTE yang dipaparkan dalam penelitian sebelumnya oleh Berkmans dan Karthick (Berkmans and Karthick 2022). Selain melakukan *oversampling* yang menghasilkan data sintetis baru, peneliti juga melakukan analisa terkait strategi *oversampling* yang optimal untuk melakukan klasifikasi transaksi *fraud* berdasarkan metrik klasifikasi yang digunakan. Strategi *oversampling* yang diuji meliputi 100%, 50%, 25%, dan 0% (tanpa *oversampling*). Untuk memastikan kualitas data sintetis yang dihasilkan, peneliti melakukan analisa kualitas data berdasarkan visualisasi distribusi yang digambarkan menggunakan histogram. Analisa ini dilakukan untuk memastikan data sintetis yang dihasilkan tetap dapat menggambarkan data asli dari transaksi *fraud*.

1.2 Identifikasi Masalah

Berdasarkan uraian latar belakang di atas, penulis mengidentifikasi masalah penting dalam penelitian ini sebagai berikut.

1. Kejahatan *fraud* kartu kredit cukup masif dan mengakibatkan kerugian finansial bagi pemegang kartu.
2. Dataset transaksi *fraud* kartu kredit memiliki ketidakseimbangan antara kelas *fraud* dan *not fraud* yang mempengaruhi hasil prediksi.
3. Dimensi dataset transaksi *fraud* kartu kredit memiliki dimensi yang besar sehingga mempengaruhi aspek komputasi.
4. Kinerja model prediksi transaksi *fraud* kartu kredit menggunakan teknik *resampling* ADASYN dan seleksi fitur SVM-RFECV belum diketahui.

1.3 Pembatasan Masalah

Untuk memudahkan dalam proses penelitian, penulis membuat batasan masalah yang akan dibahas sebagai berikut.

1. Penelitian ini menggunakan data primer riwayat transaksi dengan metode pembayaran *kartu kredit* dari dataset IEEE-CIS Fraud Detection dengan dimensi awal 148.986 baris data.
2. Penelitian menggunakan metode *oversampling* ADASYN dengan strategi *oversampling* 100%, 50% dan 25%.
3. Teknik seleksi fitur yang digunakan dalam penelitian ini adalah SVM-RFECV.
4. Algoritma *machine learning* yang digunakan dalam penelitian ini adalah algoritma *supervised learning* meliputi AdaBoost dan LGBM.

5. Output dari hasil penelitian ini adalah prediksi status transaksi kartu kredit yang bersifat *binary classification* dengan menggunakan label *Fraud* dan *Not Fraud*.
6. Proses pengolahan data dilakukan menggunakan bahasa pemrograman Python.
7. Metode evaluasi kinerja prediksi yang digunakan dalam model prediksi ini adalah AUC, *precision*, *recall* dan *f1-score*.
8. Analisis kualitas data dilakukan dengan histogram distribusi data dengan membandingkan karakteristik bentuk dan lokasi puncak histogram data sebelum dan sesudah *oversampling*.

1.4 Rumusan Masalah

Berdasarkan uraian pada latar belakang, maka penulis menentukan rumusan masalah, yaitu sebagai berikut:

1. Bagaimana proses seleksi fitur menggunakan metode SVM-RFECV untuk klasifikasi transaksi *fraud* kartu kredit?
2. Bagaimana proses *oversampling* data menggunakan metode ADASYN untuk klasifikasi transaksi *fraud* kartu kredit?
3. Bagaimana kinerja model prediksi transaksi *fraud* kartu kredit menggunakan metode ADASYN dan seleksi fitur SVM-RFECV?

1.5 Tujuan Penelitian

Adapun tujuan dari penelitian ini adalah sebagai berikut.

1. Untuk mengetahui proses seleksi fitur menggunakan metode SVM-RFECV untuk klasifikasi transaksi *fraud* kartu kredit.

2. Untuk mengetahui proses *oversampling* data menggunakan ADASYN untuk klasifikasi transaksi *fraud* kartu kredit.
3. Untuk mengetahui kinerja model prediksi transaksi *fraud* kartu kredit menggunakan metode *oversampling* ADASYN dan seleksi fitur SVM-RFECV.

1.6 Manfaat Penelitian

Hasil dalam penelitian ini diharapkan dapat memberikan suatu manfaat baik secara teoritis maupun manfaat praktis, sebagai berikut.

2.1.1 Manfaat teoritis

Manfaat teoritis yaitu hasil penelitian ini diharapkan dapat memberikan suatu manfaat untuk pengembangan referensi penelitian dalam ilmu komputer yang berkaitan dengan deteksi *fraud, machine learning,* data tidak seimbang serta seleksi fitur untuk menunjang klasifikasi transaksi *fraud* kartu kredit.

2.1.2 Manfaat praktis

Selain manfaat teoritis, penelitian ini juga diharapkan bekontribusi Untuk manfaat praktis, hasil penelitian ini diharapkan dapat memberikan manfaat bagi pihak-pihak sebagai berikut.

1. Bagi penulis

   Dengan adanya penelitian ini, dapat meningkatkan pengetahuan penulis tentang model klasifikasi transaksi *fraud* kartu kredit menggunakan algoritma *machine learning* dengan teknik *resampling* serta seleksi fitur.

2. Bagi penerbit kartu kredit

   Dapat menjadi referensi dalam melakukan pengembangan maupun implementasi *fraud detection system* sebagai tindakan preventif dalam mencegah potensi kejahatan kartu kredit.

3. Bagi pihak terkait secara umum

   Dapat memberikan pengetahuan kepada pihak terkait khususnya dalam melakukan klasifikasi transaksi *fraud* kartu kredit.

4. Bagi peneliti lainnya

   Dapat digunakan sebagai rujukan atau referensi dalam mengembangkan penelitian sejenis pada kesempatan lainnya agar menjadi penelitian berkelanjutan.