Analisis Deteksi dan Pencegahan Penipuan Kartu Kredit Menggunakan Teknik *Data Mining* dan *Machine Learning*

Reva Geryansyah Afqal
Magister Teknik Informatika Universitas Pamulang
e-mail: reva.geryansyah@undira.ac.id

Abstrak*—Dalam proses menganalisis penipuan kartu kredit, sebuah teknik machine learning digunakan untuk membuat sebuah model prediktif yang mampu membedakan antara transaksi yang sah dan kemungkinan penipuan, teknik data mining dan machine learning dapat diterapkan dalam menganalisis, mengidentifikasi, dan mencegah tindakan penipuan yang terjadi pada kartu kredit. menerapkan metode random forest pada klasifikasi penipuan transaksi kartu kredit berdasarkan evaluasi ketepatan klasifikasi dengan harapan metode random forest dapat memberikan hasil klasifikasi fraud atau non-fraud dengan benar.Untuk mengatasi masalah tersebut dapat dipecahkan menggunakan teknik data mining yaitu klasifikasi. Tujuan dari klasifikasi yaitu untuk memprediksi label kelas dari suatu objek berdasarkan atribut yang ada. Teknik ini adalah pemilihan data terlebih dahulu setelahnya melakukan pra- pemrosesan dan memverifikasi sesuai kebutuhan. Metode yang termasuk kedalam klasifikasi diantaranya yaitu metode random forest. Random forest mampu mengatasi masalah non-linier karena didasarkan pada teknik pohon keputusan (decision tree). Konsep dasar random forest yaitu menggunakan lebih dari classifier dari metode yang sama kemudian mengkombinasikannya melalui voting untuk mendapatkan hasil dugaan klasifikasi akhir. Pada data yang di ambil dari Kaggle harus memahami data tersebut, atau Data Understanding, dimana data tersebut mengambil 1000 data pelanggan yang melakukan transaksi kepada 800 pedagang, data sample tersebut dalam periode 1 Januari 2019 - 31 Desember 2020. Jumlah fitur yang digunakan yaitu 2, 3, 6 yang selanjutnya akan dicobakan pada tiap-tiap pohon untuk melihat kombinasi mana yang akan menghasilkan nilai misklasifikasi yang paling kecil untuk menentukan parameter optimal, dari perhitungan tersebut maka akan muncul nilai errorOBB yang bervariasi, maka didapatkan lah nilai F- Measure dan Nilai AUC, jika dikatakan klasifikasi yang sangat baik, maka dari keseluruhan hasil klasifikasi berada pada rentang 90-100%.

Kata Kunci— Data Mining; Decision Tree;* Kartu Kredit; *Machine Learning; Random Forest.

I. PENDAHULUAN

Kartu kredit telah menjadi salah satu metode pembayaran yang sangat diminati di seluruh dunia, memungkinkan konsumen untuk melakukan transaksi secara praktis dan efisien. Tetapi, seiring dengan meningkatnya penggunaan kartu kredit, jumlah penipuan kartu kredit juga mengalami peningkatan yang signifikan. Penyalahgunaan kartu kredit menyebabkan kerugian baik bagi pemilik kartu kredit maupun perusahaan pembayaran, dengan konsekuensi yang dapat berdampak negatif secara finansial dan pada citra mereka untuk menangani isu penipuan kartu kredit, penggunaan teknik data mining dan machine learning telah terbukti sangat efisien dalam mengenali pola dan tingkah laku yang mencurigakan. Dalam beberapa tahun terakhir, perkembangan analisis deteksi dan pencegahan penipuan kartu kredit menggunakan teknik ini telah meningkat dengan cepat.

Data mining merupakan suatu kegiatan pengambilan informasi berharga dari data yang memiliki ukuran besar serta kompleks. Dalam situasi penipuan kartu kredit, penggunaan data mining bertujuan untuk mencari pola-pola transaksi yang tidak normal atau mencurigakan, yang dapat menunjukkan adanya kegiatan penipuan. Berbagai metode seperti metode clustering, asosiasi, dan klasifikasi digunakan untuk menemukan kejanggalan dan aktivitas yang mencurigakan. Machine learning merupakan bagian dari kecerdasan buatan yang memungkinkan sistem untuk mengembangkan kemampuan belajar dari data serta menghasilkan prediksi atau keputusan tanpa memerlukan pemrograman yang spesifik.

Dalam proses menganalisis penipuan kartu kredit, sebuah teknik machine learning digunakan untuk membuat sebuah model prediktif yang mampu membedakan antara transaksi yang sah dan kemungkinan penipuan. Model-model ini secara terus-menerus diperbaharui dengan informasi terbaru agar dapat mengenali pola kecurangan yang baru muncul. Tidak bisa dianggap remeh pentingnya melakukan analisis untuk mendeteksi dan mencegah penipuan yang terjadi pada kartu kredit. Kerugian keuangan dan peringkat yang muncul akibat tindakan penipuan kartu kredit bisa sangat merugikan bagi perusahaan penerbit kartu kredit, penjual, dan pemilik kartu. Karena itu, penerapan metode data mining dan machine learning menjadi sangat krusial dalam upaya mengurangi risiko penipuan kartu kredit.

Dalam tulisan ini, kita akan mempelajari lebih lanjut tentang cara-cara teknik data mining dan machine learning dapat diterapkan dalam menganalisis, mengidentifikasi, dan mencegah tindakan penipuan yang terjadi pada kartu kredit. Kami akan mengamati contoh-contoh penerapan yang nyata dan kegunaan yang bisa didapatkan melalui pendekatan ini. Selain itu, kami juga akan mengulas hambatan dan kesulitan yang terkait dengan penerapan sistem pengenalan dan pencegahan penipuan kartu kredit yang berhasil. Dengan cara mengerti dan mengimplementasikan teknologi ini, kita bisa ikut serta dalam menjaga keamanan dan

keaslian sistem pembayaran elektronik yang semakin krusial pada era digital sekarang. Maka dari itu berdasarkan penelitian sebelumnya maka peneliti ingin menerapkan metode random forest pada klasifikasi penipuan transaksi kartu kredit berdasarkan evaluasi ketepatan klasifikasi dengan harapan metode random forest dapat memberikan hasil klasifikasi fraud atau non-fraud dengan benar.

II. METODE PENELITIAN

Untuk mengatasi masalah tersebut dapat dipecahkan menggunakan teknik data mining yaitu klasifikasi. Tujuan dari klasifikasi yaitu untuk memprediksi label kelas dari suatu objek berdasarkan atribut yang ada. Teknik ini adalah pemilihan data terlebih dahulu setelahnya melakukan pra-pemrosesan dan memverifikasi sesuai kebutuhan. Semua langkah ini melibatkan data yang hilang dan nilai yang direplikasi. Beberapa alat klasifikasi dapat digunakan pada dasar-dasar data yang dicapai melalui hasil dan pengalaman. Data model penambangan memprediksi data untuk meningkatkan kinerja dan mengurangi kemungkinan terjadinya kejahatan. Pembelajaran mesin berlangsung dengan data mining adalah metode terbaik untuk menemukan yang tepat metode untuk mendeteksi penipuan [2].

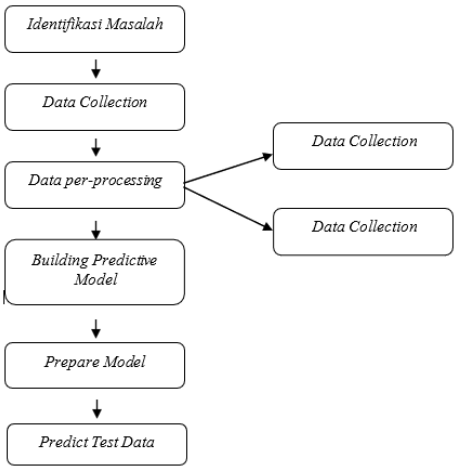
Semakin maju metode tersebut seperti yang dapat dilakukan oleh kecerdasan buatan, pembelajaran mendalam, dan jaringan saraf juga diimplementasikan untuk meningkatkan kinerja. Itu akan meningkatkan algoritma penambang penipuan juga diterapkan, ada berbagai jenis kriteria pembelajaran seperti diawasi, tanpa pengawasan dan semi-terawasi digunakan untuk tujuan tersebut memprediksi transaksi terkait keuangan. Bahkan ini metode yang digunakan di sektor perbankan untuk motif aman data dan sistem mereka. Data Mining adalah proses menggali data, mengekstraknya, serta menganalisis informasi yang berguna dari beragam basis data yang besar, dan melaksanakan operasi ekstraksi pada elemen-elemen tertentu dengan cara atau dalam format yang sesuai. Data Mining adalah proses penemuan pola dalam data (Witten, 2011).

Data Mining adalah proses menemukan korelasi baru yang bermakna, pola dan tren dengan memilah-milah sejumlah besar data yang tersimpan dalam repository, menggunakan teknologi penalaran pola serta teknik- teknik statistic dan matematika (Larose, 2005). Data Mining adalah sebuah proses, yang mana dalam melakukan prosesnya harus sesuai dengan prosedur dari proses tersebut, yaitu CRISP-DM (CrossIndustry Standard Process for Data Mining), yang terdiri dari keseluruhan proses, preprosesing data, pembentukan model, model evaluasi, dan tahap akhir penyebaran model (Larose, 2005).

Metode yang termasuk kedalam klasifikasi diantaranya yaitu metode random forest. Random forest mampu mengatasi masalah non-linier karena didasarkan pada teknik pohon keputusan *(decision tree)*. Konsep dasar random forest yaitu menggunakan lebih dari classifier dari metode yang sama kemudian mengkombinasikannya melalui voting untuk mendapatkan hasil dugaan klasifikasi akhir. Random forest memiliki kelebihan yaitu diantaranya dapat memberikan hasil klasifikasi yang baik disertai dengan hasil residu/error yang lebih rendah, dapat mengatasi data training yang berukuran sangat besar secara efisien, selain itu metode random forest juga dinilai efektif dalam mengatasi masalah missing data.

Metode random forest adalah pengembangan dari metode CART, yaitu dengan menerapkan metode bootstrap aggregating (bagging) dan random feature selection (Breiman 2001). Dalam random forest, banyak pohon ditumbuhkan sehingga terbentuk hutan (forest), kemudian analisis dilakukan pada kumpulan pohon tersebut. Pada gugus data yang terdiri atas n amatan dan ubah penjelas, random forest dilakukan dengan cara (Breiman 2001; Breiman & Cutler 2003):

1. Lakukan penarikan contoh acak berukuran n dengan pemulihan pada gugus data. Tahapan ini merupakan tahapan bootstrap.
2. Dengan menggunakan contoh bootstrap, pohon dibangun sampai mencapai ukuran maksimum (tanpa pemangkasan). Pada setiap simpul, pemilihan pemilah dilakukan dengan memilih m peubah penjelas secara acak, dimana m << p. Pemilah terbaik dipilih dari m peubah penjelas tersebut. Tahapan ini adalah tahapan random feature selection.
3. Ulangi langkah 1 dan 2 sebanyak k kali, sehingga terbentuk sebuah hutan yang terdiri atas k pohon. Respons suatu amatan diprediksi dengan menggabungkan (aggregating) hasil prediksi k pohon. Pada masalah klasifikasi dilakukan berdasarkan majority vote (suara terbanyak).



Gambar 1.
Langkah Penelitian

III. HASIL DAN PEMBAHASAN

Saat ini machine learning sangat banyak digunakan dalam data mining dan big data analitics, pendekatan Random Forest (RF) yang merupakan bagian dari algoritma supervised learning pada machine learning diterapkan dalam penelitian ini untuk memberikan solusi terbaik pada permasalahan fraud detection. Random Forest merupakan metode yang dujikan untuk melakukan prediksi mayoritas adanya pedugaan *fraud detection* pada penggunaan kartu kredit. Data yang digunakan untuk mendukung penelitan bersumber dari Kaggle, dengan variabel pengukuran sebanyak yang diperlukan dalam penelitian, variabel merupakan pola transaksi hasil analisis dengan *Principal Componen Analysis (PCA)* sementara 2 variabel lainnya merupakan waktu dan jumlah transaksi yang miliki dari setiap nasabah kartu kredit.

Metode yang digunakan dalam penelitian ini yaitu metode random forest yang diterapan pada data penipuan transaksi kartu kredit yang berisi transaksi sah dan penipuan dari durasi 1 Januari 2019 - 31 Desember 2020. Ini mencakup kartu kredit dari 1000 pelanggan yang melakukan transaksi dengan kumpulan 800 pedagang yang diperoleh dari situs Kaggle, Variabel yang digunakan yaitu sebanyak 10 variabel dimana 9 variabel merupakan variabel independent dan 1 variabel yang merupakan variabel dependent seperti yang disajikan pada Tabel 1.

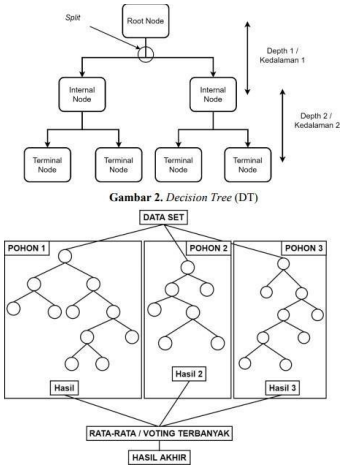
Tabel 1.
Variabel Data Transaksi Penipuan Kartu Kredit

Variable	Skala Pengukuran	Keterangan
Trans_date_trans_time	Nominal	Waktu transaksi
Cc_num	Nominal	Nomor kartu kredit pelanggan
Merchant	Nominal	Nama pedagang
Category	Nominal	Kategori pedagang
Amt	Rasio	Jumlah transaksi
City	Nominal	Nama kota dari pemegang kartu kredit
State	Nominal	Nama provinsi dari pemegang kartu kredit
Job	Nominal	Pekerjaan dari pemegang kartu kredit
Age	Nominal	Usia dari pemegang kartu kredit
Is_fraud	Nominal	Kelas target penipuan atau tidak

A. Tahapan Analisis Data

1. Persiapan data credit card fraud yang didapatkan pada situs Kaggle.
2. Dilakukannya data preprocessing seperti memilih variabel yang dibutuhkan, mengatasi dan membersihkan data yang missing, dan melakukan pengkodean data string ke numerik.
3. Mendeskripsikan data dari variabel-variabel yang informatif.
4. Pembagian data (splitting) menjadi dua bagian yaitu data training untuk melatih model dan data testing untuk evaluasi model. Proporsi pembagian yaitu sebesar 75% untuk data training dan sisanya untuk data testing sebesar 25%.
5. Melakukan pemodelan pengklasifikasian metode random forest dengan menentukan jumlah pohon dan jumlah fitur terbaik berdasarkan nilai error OOB terkecil.
6. Menerapkan model klasifikasi sehingga menghasilkan hasil prediksi akhir. Pada tahapan ini akan memperlihatkan hasil ketepatan klasifikasi seperti confusion matrix.
7. Interpretasi hasil.

Random forest (RF) adalah metode ensemble learning yang merupakan kumpulan dari pohon keputusan (*decision tree*) atau konsep dari pohon keputusan yang dilakukan secara berulang sehingga membentuk suatu hutan atau forest. Random forest adalah pengembangan dari metode CART (Classification and Regression Tree) dengan menerapkan metode bootstrap aggregating (bagging) dan random feature selection (Breiman 2001). Sedangkan Resende dan Durmond (2018) menjelaskan bahwa model random forest adalah gabungan dari decision tree, yang dapat digunakan untuk klasifikasi dan regresi. Prediksi dalam kasus klasifikasi didasarkan pada suara terbanyak dari nilai-nilai yang diprediksi menggunakan decision tree, dan dalam kasus regresi, hasilnya adalah ratarata dari nilai-nilai yang diprediksi decision tree.



Gambar 2. *Decision Tree* (DT)

Gambar 2.
Random Forest (RF)

Adapun penelitian lainnya dari (Niveditha et al., 2019) menjelaskan bahwa random forest bekerja lebih baik dengan jumlah data pelatihan yang lebih besar dan menghasilkan nilai akurasi dan precision yang hampir sempurna yaitu 99%. Pada data yang di ambil dari Kaggle harus memahami data tersebut, atau *Data Understanding*, dimana data tersebut mengambil 1000 data pelanggan yang melakukan transaksi kepada 800 pedagang, data sample tersebut dalam periode 1 Januari 2019 - 31 Desember 2020. Untuk mengetahui lagi tentang dataset tersebut, maka penulis memanggil perintah untuk menampilkan dataset tersebut menggunakan Pandas.

Dataset sudah terbuka dan bisa ditampilkan pada pandas, dengan beberapa variable yang sudah dijelaskan sebelumnya. Selanjutnya, penulis memanggil distribusi ketidakseimbangan kelas pada sebuah dataset transaksi yang berupa label yaitu, Is_fraud. Maka akan menampilkan jumlah kasus terkait. *Data Preparation*, pada tahap ini semua data akan membuat pemodelan agar data tersebut siap untuk digunakan. Data yang pertama adalah data cleaning, dimana data cleaning ini untuk mencari *missing values* pada dataset, dengan menggunakan pandas dalam python, denga menggunakan fungsi isna() dan sum(), fungsi dari isna() adalah untuk mencari nilai yang hilang dan untuk sum() sebagai agresi data agar mudah dipahami.

Analisis Klasifikasi *Random Forest*, dalam proses klasifikasi dengan menggunakan random forest terdapat dua parameter yang harus disesuaikan yaitu dengan menentukan jumlah fitur (mtry) dan jumlah pohon. Penentuan nilai mtry dilakukan dengan 3 cara yaitu dengan menggunakan Persamaan (1), (2), dan (3) dengan jumlah variabel prediktor sebanyak 9 variabel didapatkan perhitungan sebagai berikut:

1. $mtry_1 = \frac{1}{2}\left|\sqrt{9}\right| = 1{,}5 \approx 2$
2. $mtry_2 = \left|\sqrt{9}\right| = 3$
3. $mtry_3 = 2 \times \left|\sqrt{9}\right| = 6$

Jumlah fitur yang digunakan yaitu 2, 3, 6 yang selanjutnya akan dicobakan pada tiap-tiap pohon untuk melihat kombinasi mana yang akan menghasilkan nilai misklasifikasi yang paling kecil untuk menentukan parameter optimal, dari perhitungan tersebut maka akan muncul nilai *error*OBB yang bervariasi, semakin besar nilai *mtry* yang digunakan, nilai *error*OBB semakin menurun.

Evaluasi Model, membuat hasil dari klasifikasi confusion matrix dari metode random forest dengan menggunakan *pyhton*. Maka akan diketahui nilai dari hasil klasifikasi pada dataset tersebut, dengan mengunakan persamaan (4) untuk menghitung akurasi, persamaan (5) untuk menghitung sensitivity, persamaan (6) untuk menghitung precision, persamaan (7) untuk menghitung F-Measure, persamaan (8) untuk menghitung False Positive Rate, dan persamaan (9) untuk menghitung nilai AUC pada data testing, dari hasil tersebut maka didapat nilai *F-Measure* dan Nilai AUC.

IV. KESIMPULAN

Berdasarkan hasil dan pembahasan dalam penelitian ini dapat diambil kesimpulan bahwa metode random forest dapat diterapkan pada data klasifikasi penipuan transaksi kartu kredit dengan parameter optimal yang digunakan yaitu jumlah fitur sebanyak 6 fitur dan jumlah pohon yang digunakan sebanyak yang diperlukan, maka didapatkan lah nilai *F-Measure* dan Nilai AUC, jika dikatakan klasifikasi yang sangat baik, maka dari keseluruhan hasil klasifikasi berada pada rentang 90-100%.

DAFTAR PUSTAKA

[1] T. S. Lestari and D. A. N. Sirodj, "Klasifikasi Penipuan Transaksi Kartu Kredit Menggunakan Metode Random Forest," Jurnal Riset Statistika, vol. 1, pp. 160-167, 2021.
[2] A. Ramadhan and B. Susetyo, "Penerapan Metode Klasifikasi Random Forest Dalam Mengidentifikasi Faktor Penting Penilaian Mutu Pendidikan," Jurnal Pendidikan dan Kebudayaan, vol. 4, no. 2, pp. 169-182, 2019.
[3] P. T. S. Ningsih, M. Gusvarizon, and R. Hermawan, "Analisis Sistem Pendeteksi Penipuan Transaksi Kartu Kredit dengan Algoritma Machine Learning," Jurnal Teknologi Informatika dan Komputer, vol. 8, no. 2, pp. 386- 401, 2022.
[4] H. D. Honesqi, "Klasifikasi Data Mining Untuk Menentukan Tingkat Persetujuan Kartu Kredit," Jurnal Teknologi Informatika Institut Teknologi Padang, vol. 5, no. 2, pp. 57-62, 2017.
[5] Y. Yazid and A. Fiananta, "Mendeteksi Kecurangan Pada Transaksi Kartu Kredit Untuk Verifikasi Transaksi Menggunakan Metode SVM," Indonesian Journal of Applied Informatics, vol. 1, no. 2, pp. 61-66, 2017.
[6] A. Kurniawan and Y. Yulianingsih, "Pendugaan Fraud Detection pada kartu kredit dengan Machine Learning," Kilat, vol. 10, no. 2, pp. 320-325, 2021.
[7] A. Zuhairah, "Penerapan Algoritma Random Forest, Support Vector Machines (Svm) dan Gradient Boosted Tree (Gbt) Untuk Deteksi Penipuan (Fraud Detection) Pada Transaksi Kartu Kredit," Bachelor's thesis, Fakultas Sains dan Teknologi UIN Syarif Hidayatullah Jakarta