KILAT
Vol. 10, No. 2, Oktober 2021, P-ISSN 2089-1245, E-ISSN 2655-4925
DOI: https://doi.org/10.33322/kilat.v10i2.1482

Pendugaan Fraud Detection pada kartu kredit dengan Machine Learning

Arief Kurniawan[1]; Yulianingsih[2]

[1] Pusat Pelaporan dan Analisis Transaksi Keuangan
[2] Teknik Informatika, Fakultas Teknik dan Ilmu Komputer, Universitas Indraprasta PGRI
[1] ariefk@gmail.com

ABSTRACT

The increasing volume and potential losses due to credit card fraud requires solutions that allows credit card service providers to analyze each transaction quickly and accurately. This is almost impossible to do manually by humans. In this study, we tried to implement the Random Forest method to find the best solution to predict the problem and obtained a result of 0.85 which indicates that the developed model has good accuracy.

Keywords: *credit card fraud, random forest, machine learning*

ABSTRAK

Meningkatnya volume dan potensi kerugian akibat tindak pidana penipuan kartu kredit memerlukan solusi yang memungkinkan pihak penyelenggara layanan kartu kredit dapat melakukan analisis pada setiap transaksi secara cepat dan akurat. Hal ini hampir mustahil dilakukan secara manual oleh manusia. Dalam penelitian ini kami coba melakukan implementasi metode *Random Forest* untuk menemukan solusi terbaik melakukan prediksi pada permasalahan dan diperoleh hasil sebesar 0,85 yang menunjukan bahwa model yang dikembangkan memilik akurasi yang baik.

Kata kunci: penipuan kartu kredit, random forest, machine learning

Vol. 10, No. 2, Oktober 2021, P-ISSN 2089-1245, E-ISSN 2655-4925
DOI: https://doi.org/10.33322/kilat.v10i2.1482

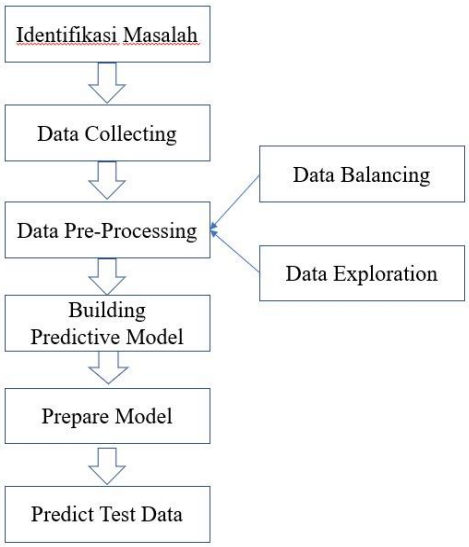
1. PENDAHULUAN

Kemajuan teknologi dan kondisi pandemi yang mendorong perubahan gaya hidup manusia menuju *Cashless society* dimana semakin banyak transaksi yang dilakukan secara virtual dengan sistem pambayaran menggunakan kartu kredit hingga mata uang kripto. Kartu kredit banyak digunakan karena kemudahan layanan dan interaksi langsungnya dengan sistem perbankan sehingga mudah diakses oleh siapapun. Seiring dengan itu kejahatan seputar kartu kredit juga makin meningkat. Dari total transaksi menggunakan kartu di tahun 2019 sejumlah 42.275 trilyun dolar (meningkat 4,2% dari tahun 2018), terjadi kerugian mencapai 28,65 milyar dolar (meningkat 2.9% dari tahun 2019)[1].
Dengan volume transaksi yang sangat banyak sementara kebutuhan perolehan informasi yang cepat dengan jumlah data historis yang banyak, diharapkan adanya metode yang dapat mendukung pendeteksian tindak penipuan kartu kredit.

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*. Algoritma Machine Learning digunakan untuk menganalisis semua transaksi dan melaporkan hal yang mencurigakan. Laporan-laporan ini kemudian diselidiki oleh para profesional yang menghubungi pemegang kartu untuk mengonfirmasi apakah transaksi itu asli atau curang. Para penyelidik memberikan umpan balik ke sistem otomatis yang digunakan untuk melatih dan memperbarui algoritme hingga akhirnya meningkatkan kinerja deteksi penipuan dari waktu ke waktu [2].

2. METODE/PERANCANGAN PENELITIAN

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 [3] dengan variabel pengukuran sebanyak 31, dimana 29 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 dari sejumlah 284.807 data. Dalam melakukan analisis performa NVIDIA Tesla K80 dengan memori 11 GB RAM 12,69 GB memberikan dukungan yang cukup baik.



Gambar 1. Langkah Penelitian

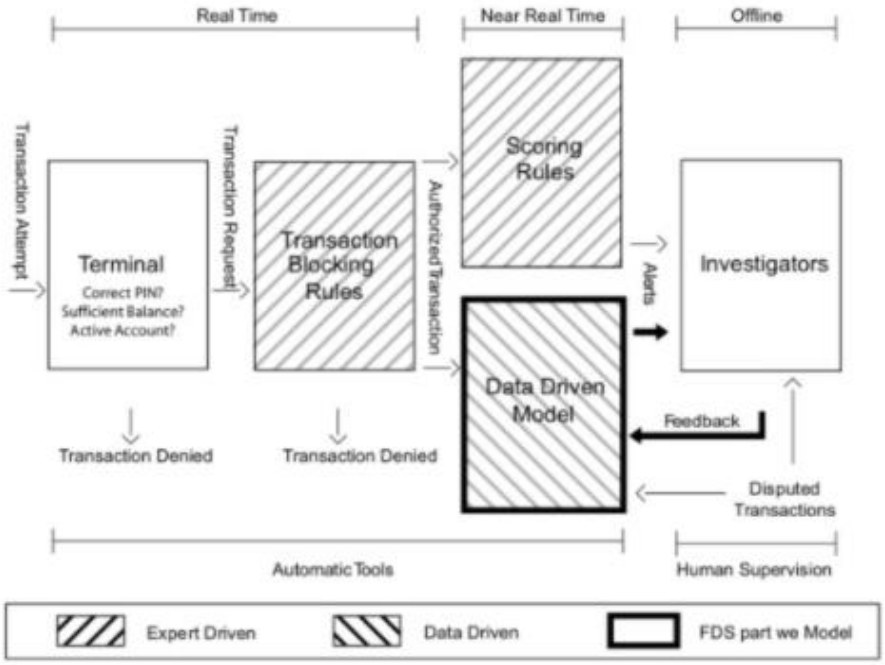
Vol. 10, No. 2, Oktober 2021, P-ISSN 2089-1245, E-ISSN 2655-4925
DOI: https://doi.org/10.33322/kilat.v10i2.1482

Gambar 1 langkah penelitian merupakan alur proses yang dilakukan untuk melakukan prediksi terjadinya fraud dengan membangun model Random Forest dengan Phyton. Praproses data dilakukan kedalam dua tahapan yaitu dengan penyeimbangan data dan eksplorasi data.

2.1 SISTEM PENDETEKSIAN PENIPUAN PADA KARTU KREDIT

Pada tahun 2017, ada 1.579 pelanggaran data dan hampir 179 juta catatan diantaranya dimana penipuan kartu kredit merupakan tindak pidana paling umum dengan 133.015 laporan, kemudian penipuan terkait pekerjaan atau pajak dengan 82.051 laporan, penipuan telepon dengan 55.045 laporan diikuti oleh penipuan bank dengan 50.517 laporan dari statika yang dirilis oleh FTC [4].

Fraud detection mempunyai beberapa tahapan proses pendeteksian diilustrasikan pada gambar 1 berikut dari setiap lapisan kendali



Gambar 2. Lapisan-lapisan kendali dalam sistem pendeteksian penipuan kartu kredit [*Machine Learning for Credit Card Fraud Detection*, 2021]

Dua lapisan pertama *(Terminal and Transaction Blocking Rules)* dijalankan secara *real-time* (yaitu dalam milidetik dan sebelum tahap otorisasi). Dua lapisan berikutnya (*Scoring Rules and Data-Driven Model* (DDM)), dieksekusi hampir secara real-time untuk antisipasi pemblokiran kartu dan mencegah penipuan tambahan. Terakhir, lapisan terakhir (Investigator) adalah satu-satunya yang membutuhkan campur tangan manusia dan dilakukan secara offline [5].

2.2. DECISION TREE

Decision Tree merupakan teknik clasifier dengan *root node* untuk mengumpulkan data, *inner node* berisi pertanyaan dan *leaf node* sebagai pengambil keputusan untuk sejumlah data yang belum diketahui kelasnya kedalam kelas-kelas yang ada. *Decision tree* classifier dikenal karena kinerjanya yang telah ditingkatkan. Karena presisi yang kuat, splitting parameter yang dioptimalkan, dan teknik *tree pruning* yang telah disempurnakan (ID3, C4.5, CART, CHAID, dan QUEST) biasanya digunakan oleh semua pengklasifikasi data yang dikenal luas. Dataset yang terpisah digunakan untuk

Vol. 10, No. 2, Oktober 2021, P-ISSN 2089-1245, E-ISSN 2655-4925
DOI: https://doi.org/10.33322/kilat.v10i2.1482

melatih sampel dari kumpulan data yang sangat besar, yang pada gilirannya mempengaruhi presisi test set [6].

Kumpulan dari beberapa tree terbaik membentuk satu model yang dikenal dengan *Random Forest* (RF). Adanya *Random forest* diyakini dapat memberikan solusi untuk permasalahan *overlap* pada metode *decison tree* ketika menggunakan kriteria dan kelas yang sangat banyak.

2.3. RANDOM FOREST

Random Forest adalah pengembangan dari metode CART, analisis dilakukan pada n amatan dan p peubah penjelas, dengan tahapan bootstrap dan random feature selection yang dilakukan secara berulang sebanyak k kali, sehingga terbentuk sebuah hutan yang terdiri atas k pohon [7]. Hasil pertumbuhan sekelompok tree dan pemilihan kelas yang paling populer dianggap sebagai peningkatan yang cukup besar dalam performa dalam melakukan prediksi [8]. Pada tahapan *bootstrap* kondisi yang harus diperhatikan adalah data yang digunakan tidak terlalu kecil, tidak memiliki banyak *outlier* dan bukan data *time series.*

2.4. GINI IMPURITY

Merupakan salah satu metode pemisahan optimal simpul akar dan simpul berikutnya pada *decision tree*. *Gini impurity* adalah ukuran seberapa sering elemen yang dipilih secara acak dari himpunan akan diberi label yang salah jika diberi label secara acak sesuai dengan distribusi label di dalam *subset*. Penggunaan indeks gini dan perolehan informasi sebelum maupun sesudah data set balanced tidak mempengaruhi performa model [9].

Gini *impurity* dihitung menggunakan rumus:

$$\text{Gini Impurity} = 1 - \sum_{i=1}^{c} P_i^2. \tag{1}$$

Langkah untuk memilih node [10]:
Langkah 1: Hitung Indeks Gini untuk setiap atribut
Langkah 2: Berikan bobot pada setiap fitur
Langkah 3: Pilih atribut dengan nilai indeks Gini terendah.
Langkah 4: Ulangi 1,2,3 hingga pohon yang digeneralisasi selesai dibuat.

3. HASIL DAN PEMBAHASAN

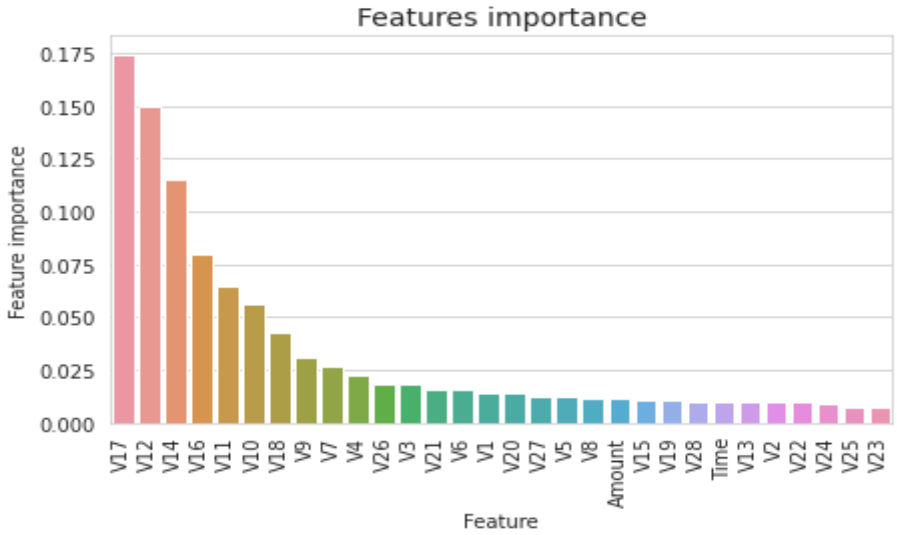
Dalam ujicoba ini digunakan algoritma *RandomForest Classifier* pada Python dengan menggunakan parameter sebagai berikut:

```
RandomForestClassifier(bootstrap=True, ccp_alpha=0.0, class_weight=None,
                       criterion='gini', max_depth=None, max_features='auto',
                       max_leaf_nodes=None, max_samples=None,
                       min_impurity_decrease=0.0, min_impurity_split=None,
                       min_samples_leaf=1, min_samples_split=2,
                       min_weight_fraction_leaf=0.0, n_estimators=100, n_jobs=4,
                       oob_score=False, random_state=2018, verbose=False,
                       warm_start=False)
```

Dari 31 *field* data atau *feature* yang digunakan perlu diketahui *feature* mana yang paling relevan atau yang paling mempengaruhi hasil prediksi sehingga dapat diketahui prioritas langkah

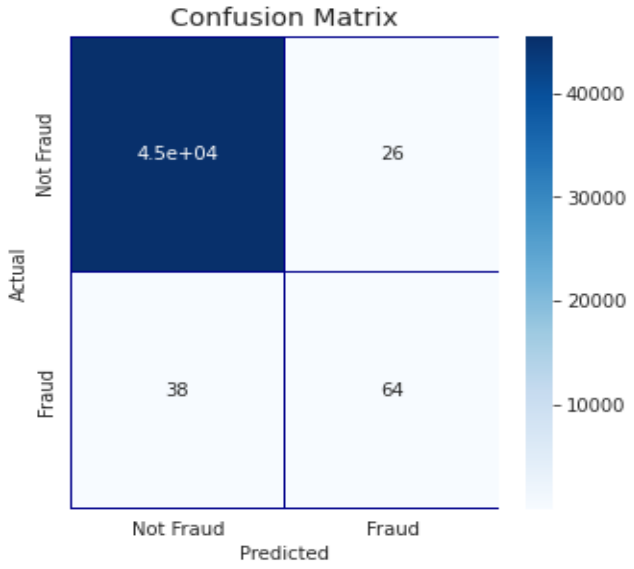
Vol. 10, No. 2, Oktober 2021, P-ISSN 2089-1245, E-ISSN 2655-4925
DOI: https://doi.org/10.33322/kilat.v10i2.1482

yang harus diambil untuk menyelesaikan masalah. Dari hasil perhitungan menggunakan metode *Random Forest Built-in Fature Importance* diperoleh hasil seperti gambar berikut:



Gambar 3. *Feature Importance*

Dari grafik gambar 3 feature Importance dapat disimpulkan bahwa *feature* V17, V12, V14, V10, V11, V16 merupakan unsur data yang paling berpengaruh terhadap hasil prediksi. Dengan model yang digunakan dapat divisualisasikan hasil pengujian akurasinya malalui visualisasi *Confusion Matrix* dengan hasil seperti gambar berikut:



Gambar 4. *Confusion Matrix* dari Model Klasifikasi

Dari grafik gambar 4 dapat diperoleh kesimpulan bahwa:

- Jumlah *True Negatif* (klasifikasi 'bukan penipuan' yang diprediksi secara tepat) sekitar 45.000 data

Vol. 10, No. 2, Oktober 2021, P-ISSN 2089-1245, E-ISSN 2655-4925
DOI: https://doi.org/10.33322/kilat.v10i2.1482

- Jumlah *False Positive* (klasifikasi 'penipuan' yang diprediksi secara salah) sebanyak 26 kali.
- Jumlah *False Negative* (klasikasi 'bukan penipuan' yang diprediksi secara salah) sebanyak 38 kali.
- Jumlah *True Positive* (klasifikasi 'penipuan' yang diprediksi secara tepat) sebanyak 64 kali.

Berdasarkan penghitungan nilai prediksi menggunakan metode *roc_auc_score* diperoleh nilai akurasi dari model yang digunakan adalah 0,85. Dengan nilai yang diperoleh dapat disimpulkan bahwa model yang digunakan telah cukup baik untuk dapat melakukan klasifikasi terhadap kemungkianan tindak pidana penipuan kartu kredit.

4. KESIMPULAN DAN SARAN

Berdasarkan serangkaian pengujian terhadap model yang telah dikembangkan menggunakan algoritma *Random Forest Classification* diperoleh kesimpulan bahwa *feature* yang paling relevan adalan V17, melalui visualisasi akurasi dengan *confusion matriks* dan perhitungan akurasi diperoleh nilai akurasi yang baik. Dengan demikian dapat disimpulkan bahwa algoritma *Random Forest Classifier* dengan model yang dikembangkan mampu melakukan klasifikasi untuk melakukan analisis tindak pidana penipuan kartu kredit.

DAFTAR PUSTAKA

[1] Nilsonreport, "Card Fraud Worldwide," 2020.
[2] S. P. Maniraj, A. Saini, S. D. Sarkar, and A. Shadab, "Credit Card Fraud Detection using Machine Learning and Data Science," Int. J. Res. Appl. Sci. Eng. Technol., vol. 9, no. VII, pp. 3788–3792, 2021.
[3] "Data Kaggle." [Online]. Available: www.kaggle.com/mlg-ulb/creditcardfraud .
[4] V. N. Dornadula and S. Geetha, "Credit Card Fraud Detection using Machine Learning Algorithms," Procedia Comput. Sci., vol. 165, pp. 631–641, 2019.
[5] Le Borgne and Yann-A, Machine Learning for Credit Card Fraud Detection. Bruxelles: Universite Libre de Bruxelles, 2021.
[6] B. Charbuty and A. Abdulazeez, "Classification Based on Decision Tree Algorithm for Machine Learning," J. Appl. Sci. Technol. Trends, vol. 2, no. 01, pp. 20–28, 2021.
[7] Leo Breiman, "Random Forests," Kluwer Academic, 2011.
[8] N. Kousika, G. Vishali, S. Sunandhana, and M. A. Vijay, "Machine Learning based Fraud Analysis and Detection System," J. Phys. Conf. Ser., vol. 1916, no. 1, 2021.
[9] S. Tangirala, "Evaluating the impact of GINI index and information gain on classification using decision tree classifier algorithm," Int. J. Adv. Comput. Sci. Appl., no. 2, pp. 612–619, 2020.
[10] T. Daniya, M. Geetha, and K. S. Kumar, "Classification and regression trees with gini index," Adv. Math. Sci. J., vol. 9, no. 10, pp. 8237–8247, 2020.