Analisis Sistem Pendeteksi Penipuan Transaksi Kartu Kredit dengan Algoritma *Machine Learning*

Putu Tirta Sari Ningsih[1)*], Muhammad Gusvarizon[2)], Rudi Hermawan[3)]

[1)]Program Studi Akuntansi, Universitas Mohammad Husni Thamrin
[2)]Program Studi Manajemen, Universitas Mohammad Husni Thamrin
[3)]Program Studi Teknik Informatika, Universitas Mohammad Husni Thamrin

Correspondence author: putu_tirtasari@yahoo.com, Jakarta, Indonesia
DOI: https://doi.org/10.37012/jtik.v8i2.1306

Abstrak

Meningkatnya jumlah pengguna kartu kredit di indonesia menimbulkan kekhawatiran akan terjadinya tindak penipuan transaksi kartu kredit. Banyaknya volume transaksi dan cepatnya proses transaksi yang berlangsung, membuat tidak mungkin untuk diawasi secara manual oleh manusia. Pengawasan diperlukan untuk melakukan pencegahan terhadap tindak penipuan transaksi kartu kredit. Cara terbaik yang dapat dilakukan adalah dengan memanfaatkan teknologi *machine learning* dan algoritmanya untuk membuat sebuah sistem yang dapat mendeteksi penipuan transaksi kartu kredit. Dalam *machine learning* terdapat banyak algoritma yang pada dasarnya memiliki tingkat akurasi dan efisiensi berbeda-beda. Untuk memilih algoritma apa yang paling cocok untuk memecahkan suatu masalah perlu dilakukan perbandingan antar beberapa algoritma. Pada penelitian ini akan diukur performa dari beberapa algoritma *machine learning* seperti *decision tree* (DT), *random forest (RF)*, *logistic regression* (LR), dan *support vector machine* (SVM) untuk mendeteksi penipuan transaksi kartu kredit menggunakan data transaksi kartu kredit yang didapatkan dari Kaggle. Data yang digunakan berisi 284807 transaksi kartu kredit yang dilakukan oleh pemegang kartu di eropa selama dua hari dengan bobot transaksi normal sebanyak 99,83% dan fraud sebanyak 0,17%. Adapun langkah-langkah yang dilakukan ialah dengan melakukan *preprocessing* data terlebih dahulu termasuk melakukan *oversampling*, lalu membuat model tanpa menentukan parameter dan dengan parameter yang ditentukan dengan bantuan fungsi GridSearchCV, melatih model dengan data pelatihan, dan melakukan prediksi menggunakan data tes. Hasil dari penelitian ini yaitu didapatkan bahwa model dengan algoritma *random forest* memiliki nilai performa paling tinggi secara keseluruhan. Maka berdasarkan hasil tersebut dapat disimpulkan bahwa algoritma *random forest* adalah algoritma yang paling cocok untuk mendeteksi penipuan transaksi kartu kredit.

Kata Kunci: Penipuan, Transaksi Kartu Kredit, *Machine Learning*, *Decision Tree*, *Random Forest*

Abstract

The increasing number of credit card users in Indonesia raises concerns about credit card transaction fraud. The large volume of transactions and the fast transaction processing that takes place, makes it impossible for them to be monitored manually by humans. Supervision is needed to prevent acts of fraudulent credit card transactions. The best way to do this is to utilize machine learning technology and its algorithms to create a system that can detect fraudulent credit card transactions. In machine learning there are many algorithms that basically have different levels of accuracy and efficiency. To choose which algorithm is most suitable for solving a problem, it is necessary to do a comparison between several algorithms. In this study, the performance of several machine learning algorithms such as decision trees (DT), random forests (RL), logistic regression (LR), and support vector machines (SVM) will be measured to detect fraudulent credit card transactions using credit card transaction data obtained from Kaggle. The data used contains 284807 credit card transactions made by cardholders in Europe for two days with a normal transaction weight of 99.83% and fraud of 0.17%. The steps taken are preprocessing the data first including oversampling, then creating a model without specifying parameters and with the parameters specified with the help of the GridSearchCV function, training the model with training data, and making predictions using test data. The results of this study are that the model with the random forest algorithm has the highest overall performance value. So based on

these results it can be concluded that the random forest algorithm is the most suitable algorithm for detecting fraudulent credit card transactions.

Keywords: *Fraud, Credit Card Transactions, Machine Learning, Decision Tree, Random Forest*

PENDAHULUAN

Kartu Kredit merupakan alat pembayaran berupa kartu pengganti uang tunai yang digunakan untuk melakukan transaksi. Banyak masyarakat memilih menggunakan kartu kredit karena berbagai keuntungan diantaranya adalah pemilik kartu kredit dapat melakukan transaksi tanpa harus memiliki uang tunai. Selain itu, penerbit dan pemberi layanan kartu kredit juga seringkali memberi berbagai penawaran menarik seperti hadiah bila sering bertransaksi menggunakan kartu kredit, pinjaman jangka pendek tanpa bunga dengan ketentuan, dan jaminan transaksi andal dan nyaman sehinggamembuat masyarakat semakin tertarik untuk menggunakan kartu kredit. Sejauh ini peredaran kartu kredit di indonesia mengalami peningkatan. Berdasarkan data Statistik Sistem Pembayaran dan Infrastruktur Pasar Keuangan (SPIP) yang diterbitkan oleh Bank Indonesia pada juni 2022 dapat diketahui bahwa jumlah kartu kredit yang beredar di indonesia mengalami peningkatan sejak 10 tahun terakhir, dimana per Mei 2022 jumlah kartu kredit yang beredar tercatat sebanyak 16.588.263 unit, bertambah hampir 2 juta unit dibandingkan dengan jumlah pada tahun 2012 yang sebanyak 14.817.168 unit. Sedangkan untuk volume transaksi menggunakan kartu kredit di indonesia pada bulan Mei 2022 mencapai 28.360 transaksi, lebih tinggi dibandingkan volume pada Mei 2021 sejumlah 23.452 transaksi.

Namun disamping berbagai keuntungan dan kemudahan yang ditawarkan, transaksi menggunakan kartu kredit tidaklah luput dari tindak kejahatan. Banyaknya pengguna menjadi salah satu faktor yang membuat kartu kredit menjadi target utama tindak kejahatan seperti penipuan transaksi, yaitu transaksi tidak sah yang dilakukan orang tidak dikenal dengan memanfaatkan kebocoran data pribadi pemilik kartu kredit. Keamanan dari data pribadi merupakan tanggungjawab bersama antara pihak yang memiliki dan diberikan akses terhadap data pribadi tersebut. Sehingga pihak-pihak yang bersangkutan wajib dengan sungguh-sungguh menjaga agar data pribadi tersebut tidak bocor dan disalahgunakan oleh pihak tak dikenal. Namun demikian, kasus kebocoran data masih sering terjadi. Berbagai cara untuk menjaga kerahasiaan data dirasa belum cukup untuk mencegah terjadinya transaksi tidak sah menggunakan kartu kredit. Oleh karena itu diperlukan cara lain untuk mencegah penipuan transaksi kartu kredit, seperti melakukan pendeteksian sedini mungkin terhadap transaksi yang berlangsung. Tetapi dengan banyaknya volume transaksi dan cepatnya proses transaksi yang berjalan terus-menerus menjadikannya sangat tidak mungkin

untuk dilakukan pendeteksian secara manual oleh manusia. Maka solusi untuk menangani masalah ini adalah dengan membuat sebuah sistem pendeteksi penipuan transaksi kartu kredit, yaitu sistem yang dapat mendeteksi tindak penipuan transaksi kartu kredi secara otomatis, cepat, akurat, dan bekerja terus-menerus tanpa bantuan tangan manusia, dengan memanfaatkan *machine learning* dan algoritmanya. Sistem pendeteksi penipuan transaksi kartu kredit adalah salah satu bentuk pemanfaatan *machine learning*. Cara kerja dari sistem pendeteksi penipuan berbasis *machine learning* adalah dengan mendeteksi anomali pada data transaksi berdasarkan variabel atau faktor tertentu yang dipelajari oleh mesin dari data yang diberikan untuk latihan, yang berisi berbagai informasi terkait transaksi kartu kredit termasuk kategori dari transaksi tersebut (transaksi normal, transaksi *fraud*). Setelah ditemukan kecocokkan maka sistem akan mengklasifikasikan transaksi ke dalam masing-masing kelas transaksi normal atau *fraud*.

Machine learning merupakan cabang dari ilmu komputer yang menggabungkan berbagai macam disiplin ilmu dalam pengembangannya, seperti ilmu komputer, ilmu statistik dan ilmu matematika. Secara umum *machine learning* bertujuan untuk membuat sebuah program yang dapat belajar sendiri dari pengalaman atau data yang diberikan sebagai bahan ajar mesin atau program tersebut. *Machine learning* memiliki beberapa jenis pendekatan dan algoritma berbeda yang dapat digunakan untuk mengatasi berbagai masalah berbeda. Salah satu contoh pendekatan dari *machine learning* adalah *supervised learning* atau suatu pendekatan di mana program belajar untuk mengklasifikasi atau memprediksi hasil keluaran dari data yang sudah memiliki hasilnya. *Supervised learning* dapat digunakan untuk mengatasi masalah klasifikasi dan regresi. Klasifikasi adalah suatu proses untuk mengklasifikasikan suatu nilai berdasarkan kelas yang sudah ditentukan, sedangkan regresi adalah proses memprediksi niai kontinu. Jenis pendekatan supervised learning memiliki berbagai algoritma seperti *decision tree, support vector machine, linear regression, logistic regression, random forest classifier, random forest regressor*, dan lain-lain.

Setiap algoritma memiliki performa yang berbeda bahkan jika digunakan untuk menyelesaikan masalah yang sama, baik dalam kasus klasifikasi maupun regresi. Oleh karena itu dalam memilih algoritma apa yang akan diterapkan dalam pengembangan model atau sistem berbasis *machine learning* ada beberapa tahap yang perlu dilakukan terlebih dahulu seperti identifikasi masalah, hingga membandingkan algoritma-algoritma berbeda untuk kasus yang sama. Umumnya algoritma yang menghasilkan performa terbaik adalah algoritma yang akan dipilih untuk diterapkan untuk pengembangan lebih lanjut. Perlu dilakukan penelitian untuk mengukur dan membandingkan performa beberapa algoritma

supervised machine learning agar dapat diketahui algoritma mana yang memiliki performa paling tinggi dalam mendeteksi dan mengklasifikasikan transaksi kartu kredit ke dalam kelasnya masing-masing.

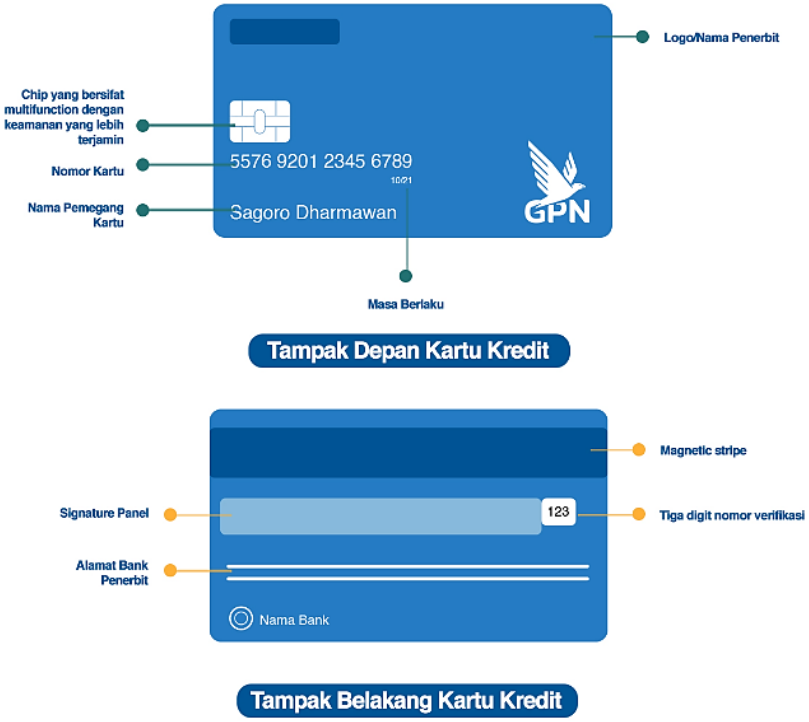
Berdasarkan Peraturan Bank Indonesia Nomor 14/2/PBI/2012 Tentang Perubahan Atas Peraturan Bank Indonesia Nomor 11/11/PBI/2009 Tentang Penyelenggaraan kegiatan Alat Pembayaran Dengan Menggunakan Kartu atau selanjutnya disebut APMK Pasal 1 Angka 4, "Kartu Kredit adalah APMK yang dapat digunakan untuk melakukan pembayaran atas kewajiban yang timbul dari suatu kegiatan ekonomi, termasuk transaksi pembelanjaan dan/atau untuk melakukan penarikan tunai, dimana kewajiban pembayaran pemegang kartu dipenuhi terlebih dahulu oleh *acquirer* atau penerbit, dan pemegang kartu berkewajiban untuk melakukan pembayaran pada waktu yang disepakati baik dengan pelunasan secara sekaligus (*charge card*) ataupun pembayaran secara angsuran". Kartu kredit merupakan alat pembayaran berbentuk kartu untuk melakukan transaksi menggantikan uang tunai, yang saat jatuh tempo pada waktu yang telah disepakati pemilik kartu kredit wajib melakukan pembayaran sekaligus atau dengan jumlah minimum diawal dan sisanya dibayarkan secara angsuran.

Ada dua jenis transaksi menggunakan kartu kredit, yaitu Transaksi Dengan Kartu atau *Card Present* (CP) *Transaction* dan Transaksi Tanpa Kartu atau *Card Not Present* (CNP) *Transaction*. Perbedaan antara kedua jenis transaksi tersebut adalah Transaksi Dengan Kartu yaitu melakukan transaksi dengan menggunakan kartu kredit fisik yang digesek, ditempel, atau dimasukkan ke mesin atau alat pembaca kartu hingga chip dibaca dan diproses, sedangkan Transaksi Tanpa Kartu ialah melakukan transaksi tanpa kartu fisik, melainkan dengan menggunakan informasi-informasi kunci dari kartu kredit seperti Nama Pemegang Kartu, Nomor Kartu, CVV/CVC, dan lain-lain. Karakteristik kartu kredit berdasarkan Bank Indonesia adalah sebagaimanat terdapat pada gambar 1.

Detail Kartu Kredit :

1. Chip pada kartu kredit berada di sisi depan kartu, chip telah ditambahkan berbagai aplikasi yang dapat mengenkripsi data sehingga data dapat tersimpan lebih aman,
2. Nomor Kartu adalah 16 digit angka yang merupakan nomor dari kartu kredit, nomor ini tidak akan pernah sama dengan nomor kartu kredit lain,
3. Nama Pemegang Kartu adalah nama dari pemegang atau pemilik kartu kredit,
4. Nama atau Logo Penerbit adalah nama atau logo perusahaan yang menerbitkan kartu,
5. Masa Berlaku adalah tanggal yang berupa 2 digit bulan dan 2 digit belakang tahun batas masa berlaku kartu kredit,

6. Logo Jaringan Kartu (dalam ilustrasi, GPN) adalah logo dari jaringan kartu kredit,
7. Pada tampak belakang, terdapat *magnetic stripe* yang merupakan garis yang akan dibaca ketika melakukan transaksi dengan cara gesek kartu. *Magnetic stripe* masih dapat digunakan untuk bertransaksi di luar negeri,
8. Signature panel adalah tempat pembubuhan tanda tangan pemilik kartu,
9. Nomor Verifikasi atau CVV/CVC adalah 3 digit di samping signature panel,
10. Alamat Bank Penerbit adalah alamat dari perusahaan/bank yang menerbitkan kartu,
11. Nama atau Logo penerbit kartu.



Gambar 1. Karakteristik kartu kredit. Source: www.bi.go.id

Carding merupakan salah satu tindak *Cybercrime* atau kejahatan siber di bidang perbankan. Istilah *carding* merujuk pada jenis transaksi *Card Not Present* (CNP) *Transaction* atau Transaksi Tanpa Kartu yang dilakukan secara tidak sah oleh orang tak dikenal. Atau transaksi yang dilakukan oleh orang lain menggunakan data pribadi pemilik kartu kredit yang sebenarnya, yang didapatkan secara ilegal seperti melalui jual beli data pribadi oleh oknum, kebocoran data, peretasan, penyadapan, dan bentuk akses ilegal lainnya, yang biayanya dibebankan kepada pemilik kartu kredit yang sebenarnya.

Undang-Undang Republik Indonesia Nomor 11 Tahun 2008 Tentang Informasi Dan Transaksi Elektronik, dan UndangUndang Republik Indoensia Nomor 19 Tahun 2016 Tentang Perubahan Atas Undang-Undang Nomor 11 Tahun 2008 Tentang Informasi Dan Transaksi Elektronik menyatakan bahwa yang termasuk tindak pidana di bidang Teknologi

Informasi dan Transaksi Elektronik adalah aktivitas illegal seperti penyadapan terhadap informasi atau transmisi informasi dan/atau dokumen elektronik yang tidak bersifat publik milik orang lain, mendsitribusikan dan/atau mentransmisikan dan/atau membuat dapat diaksesnya informasi elektronik milik orang lain yang tidak bersifat publik, memfasilitasi perbuatan yang dilarang, dan pemalsuan informasi atau dokumen elektronik, dan lain-lain. Berdasarkan pemaparan dalam undang-undang tersebut maka dapat dikatakan bahwa *carding* merupakan tindakan yang melanggar hukum di indonesia.

METODE

Samuel (dalam Mahesh, 2019) mendefinisikan *machine learning* sebagai bidang studi yang memberikan komputer kemampuan untuk belajar tanpa diprogram secara eksplisit. Menurut Hao dan Ho (2019) *machine learning* merujuk pada kumpulan metodologi yang memungkinkan komputer untuk "mempelajari" hubungan antara representasi numerik dari data dan target nilai tertentu.

Machnie Learning (ML) dalam bahasa indonesia berarti Pembelajaran Mesin, adalah suatu cabang dalam ilmu komputer yang secara umum memiliki tujuan untuk membuat program komputer yang dapat belajar dari data. *Machine learning* merupakan bidang multidisiplin yang artinya selain ilmu komputer, banyak ilmu-ilmu lain yang diterapkan dalam machine learning seperti statistik, matematika, logika, dan lain-lain.

Fungsi utama dari *machine learning* adalah melakukan prediksi. Selayaknya manusia, *machine learning* akan belajar dari pengalaman(data) untuk memprediksi suatu kemungkinan yang akan terjadi, atau membuat keputusan terbaik berdasarkan informasi atau pengetahuan yang didapatkan dari pengalaman atau data yang dipelajari. Selain itu, *machine learning* juga didesain untuk secara otomatis meningkatkan kemampuannya sendiri dalam belajar dan bertugas dari waktu ke waktu, dengan cara disuplai data serta informasi yang menjadi bentuk pengalaman bagi mesin tersebut. Mitchell (dalam Jafar Alzubi et al, 2018) menyatakan program komputer dikatakan belajar dari pengalaman (E/*Experience*) yang sehubungan dengan beberapa tugas (T/*Task*) dan beberapa pengukuran kinerja (P/*Performance measurement*). Dalam pengembangan dan pemanfaatannya, *machine learning* bergantung pada pendekatan dan algoritma berbeda. Setiap pendekatan dan algoritma dalam machine learning memiliki keunggulan tersendiri dan hanya cocok untuk kasus tertentu untuk diselesaikan. Misal, pendekatan *supervised learning* hanya cocok untuk menangani masalah klasifikasi dan regresi, dan contoh algoritmanya yaitu *random forest classifier* atau *logistic regression* hanya dapat menangani masalah klasifikasi, dan *random*

forest regressor atau *linear regression* hanya untuk menangani masalah regresi. Lalu pendekatan *unsupervised learning* hanya cocok untuk masalah *clustering* dan *association*, contoh algoritmanya adalah K-Means Clustering yang hanya cocok untuk masalah pengelompokkan atau *clustering*.

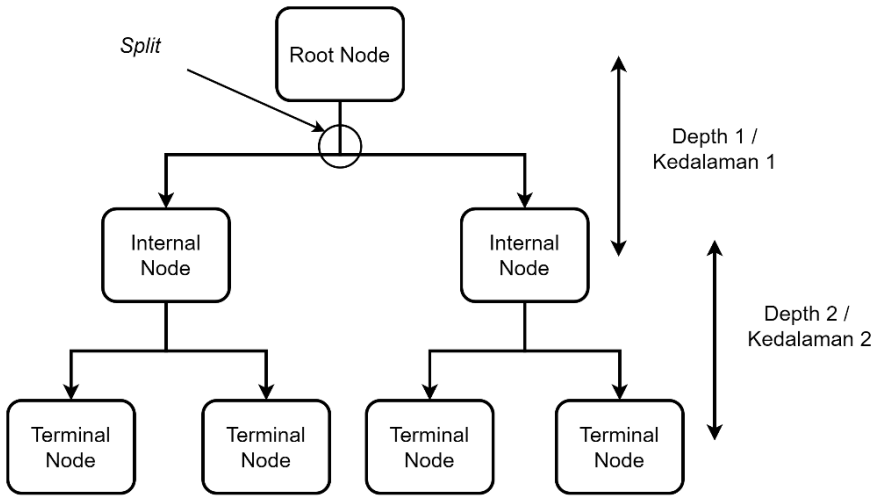
Semisupervised Learning adalah pendekatan yang menggabungkan *supervised* dan *unsupervised learning*. Bila pada *supervised* dan *unsupervised* data memiliki label atau tidak sama sekali, dalam *semisupervised* sebagian data memiliki label tapi sebagian besar data tidak memiliki label karena melabeli data memakan biaya yang tinggi atau kurangnya kemampuan yang dimiliki sumber daya manusia untuk melakukan pelabelan. Maka untuk data yang sebagian memiliki label dan sebagian lagi tidak memiliki label, pendekatan *semisupervised learning* adalah pendekatan yang cocok. *Semisupervised* learning juga dapat digunakan untuk menangani masalah klasifikasi, regresi, dan prediksi.

Reinforcement Learning adalah model learning yang bekerja dengan cara mengamati lingkungan. *Reinforcement learning* adalah bagian dari *machine learning* yang berkaitan dengan bagaimana agen harus mengambil tindakan dalam lingkungan untuk memaksimalkan reward (Mahesh, 2019).

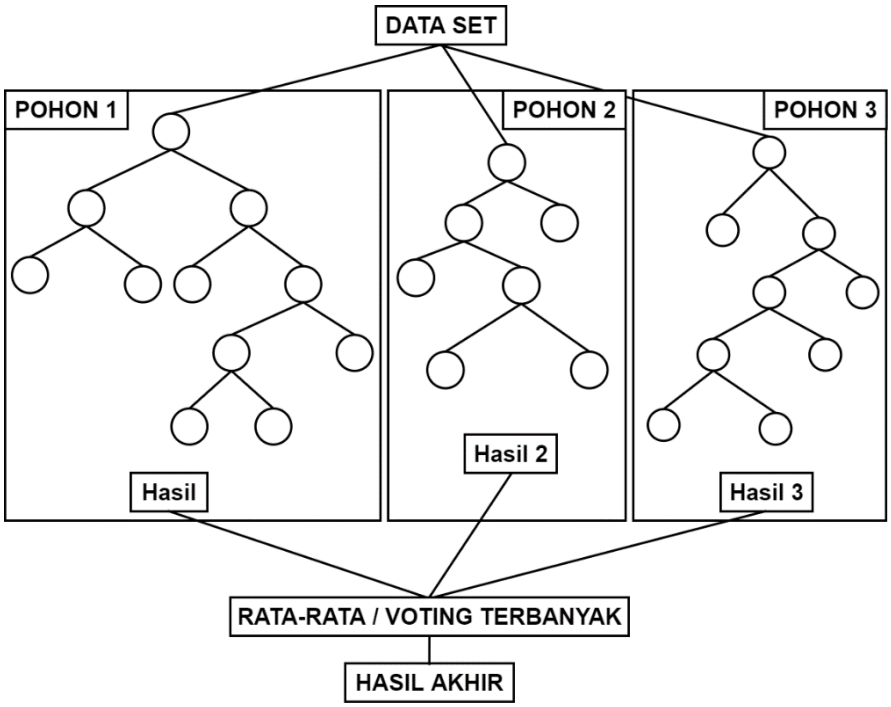
Menggabungkan beberapa algoritma atau model pembelajaran menjadi satu untuk menyelesaikan suatu masalah disebut *Ensemble Learning*. *Ensemble learning* digunakan untuk meningkatkan performa model atau algoritma yang lemah sehingga lebih kuat dan lebih akurat. Menurut Mahesh (2019), *ensemble learning* adalah proses di mana multipel model dibuat dan digabungkan secara strategis untuk menyelesaikan komputasi intelejen tertentu. *Ensemble learning* utamanya digunakan untuk meningkatkan performa model, atau mengurangi kemungkinan pemilihan yang tidak menguntungkan dari model yang lemah. Sementara menurut Jafar Alzubi et al (2018) *ensemble learning* adalah model *machine learning* di mana sejumlah pembelajaran (individual model) dilatih untuk menyelesaikan masalah. Tidak seperti teknik machine learning lainnya yang mempelajari hipotesis tunggal dari data pelatihanan, *ensemble learning* mencoba belajar dengan membangun sekumpulan hipotesis dari data pelatihanan dan dengan mengkombinasikan hipotesis-hipotesis tersebut untuk membuat model prediksi untuk mengurangi *bias* (*boosting*), *varians* (*bagging*), atau meningkatkan prediksi (*stacking*).

Decision Tree (DT) atau pohon keputusan adalah sebuah teknik pengambilan keputusan yang digambarkan dalam bentuk pohon terbalik. Mahesh (2019) menyatakan bahwa *decision tree* adalah grafik untuk merepresentasikan pilihan dan hasilnya dalam bentuk pohon. *Nodes* atau simpul dalam grafik mewakili suatu kejadian atau pilihan dan

ujung dari grafik mewakili aturan keputusan atau kondisi. Setiap pohon terdiri atas simpul-simpul dan cabang. Setiap simpul mewakili atribut dalam sebuah grup yang akan diklasifikasikan dan tiap cabang mewakili nilai yang dapat diambil oleh simpul. Menurut Panda & Sagar B.S. (2022), *decision tree* adalah regresi melalui klasifikasi untuk *data mining* dan aplikasi lain yang diwakili dengan struktur seperti pohon terbalik, di mana akar berada di atas adalah input dan daun di bawah adalah hasil atau keputusan. Konsep Dasar dari decision tree dapat dilihat pada gambar 2 berikut.



Gambar 2. *Decision Tree* (DT)



Gambar 3. *Random Forest* (RF)

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

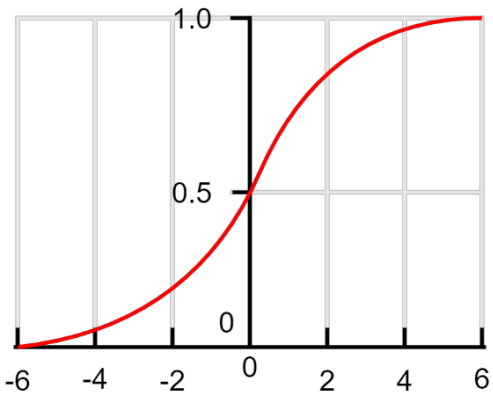
pengembangan dari metode CART (*Classification and Regression Tree*) dengan menerapkan metode *bootstrap aggregating* (bagging) dan *random feature selection* (Breiman 2001). Sedangkan Resende dan Durmond (2018) menjelaskan bahwa model *random forest* adalah gabungan dari *decision tree*, yang dapat 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 rata-rata dari nilai-nilai yang diprediksi *decision tree*.

Logistic Regression (LR) adalah teknik dalam *machine learning* yang berasal dari bidang statistik. *Logistic regression* adalah salah satu algoritma klasifikasi yang digunakan untuk memprediksi nilai biner dalam satu set variabel independen tertentu (1/0, Iya/Tidak, Benar/Salah) (Lakhsmi & Kavila, 2018). Formula dasar pembentuk dari *logistic regression* dapat dinotasikan sebagai berikut.

$$g(X) = sigmoid(\alpha + \beta X)$$

$$sigmoid(x) = \frac{1}{1 + exp\,(-x)}$$

Adapun visualisasi dari model *logistic regression* di atas akan membentuk kurva sigmoid sebagai berikut.



Gambar 4. Kurva Sigmoid *Logistic Regression* (LR)

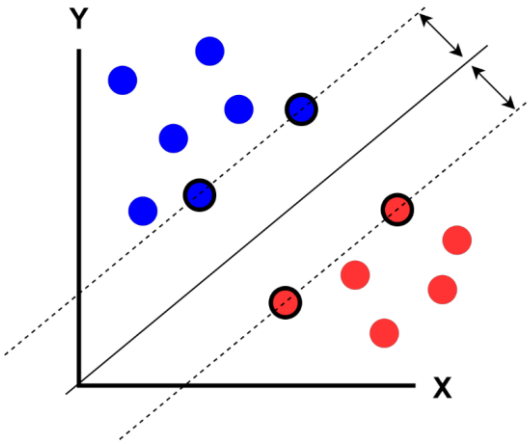
Kurva sigmoid yang dihasilkan akan berbentuk huruf S. Rentang nilai sumbu x tergantung pada kasus atau masalah yang dihadapi dan tidak harus -6 hingga 6 seperti pada ilustrasi, namun bisa berapapun tergantung pada kasus atau data set. Untuk sumbu y pada kurva sigmoid akan selalu berada di rentang 0 dan 1. Rentang nilai y tersebut akan berkorelasi dengan nilai *binomial probability*. Pada dasarnya, *logistic regression* akan memprediksi nilai yang berada dalam rentang 0 dan 1. Dalam penerapannya untuk menangani masalah klasifikasi biner, kelas akan diubah menjadi 0 dan 1. Apabila

probabilitynya mendekati 0, maka akan diklasifikasikan sebagai kelas 0. Dan apabila nilai probabilitynya mendekati 1 maka akan diklasifikasikan sebagai kelas 1.

Support Vector Machine (SVM) adalah model *supervised learning* dengan algoritma pembelajaran tersupervisi yang menganalisis data yang digunakan untuk analisis klasifikasi dan regresi. Selain melakukan klasifikasi linier, SVM dapat secara efisien melakukan klasifikasi non-linier menggunakan apa yang disebut *trik kernel*, secara implisit memetakan inputnya ke dalam ruang fitur berdimensi tinggi. Pada dasarnya, SVM menggambar margin antara kelas-kelas. Margin ditarik sedemikian rupa sehingga jarak antara margin dan kelas maksimum dan karenanya, meminimalkan kesalahan klasifikasi (Mahesh 2019).

Menurut Jafar Alzubi et al (2018) SVM bekerja pada konsep perhitungan margin. Dalam algoritma ini, setiap item data diplot sebagai titik dalam ruang n-dimensi (di mana n adalah jumlah fitur yang dimiliki dataset). Nilai setiap fitur adalah nilai koordinat yang sesuai. SVM mengklasifikasikan data ke dalam kelas yang berbeda dengan menemukan garis (*hyperplane*) yang memisahkan data set latihan ke dalam kelas-kelas. SVM bekerja dengan memaksimalkan jarak antara titik data terdekat (di kedua kelas) dan *hyperplane* yang disebut sebagai margin.

Berdasarkan penjelasan di atas dapat diketahui bahwa SVM adalah teknik klasifikasi dan regresi yang bekerja dengan cara mencari *hyperplane* terbaik. *Hyperplane* adalah batas keputusan atau pemisah antar dua kelas. Jarak antara *hyperplane* dengan data terdekat disebut sebagai margin, sedangkan data yang paling dekat dengan *hyperplane* disebut *support vector*. Ilustrasi dasar dari SVM dapat dilihat di bawah ini.

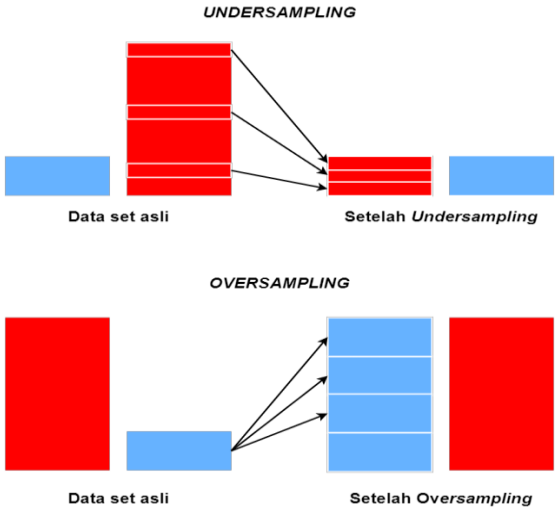


Gambar 5. *Support Vector Machine* (SVM)

Dalam *machine learning* dan statistik pada umumnya terdapat ketidakseimbangan antar kelas pada data yang digunakan yang disebut *class imbalance* atau *data imbalance*. Ketidakseimbangan data adalah kondisi di mana suatu kelas memiliki porsi yang jauh lebih banyak (mayoritas) dibandingkan dengan kelas lainnya (minoritas) pada data. Dengan

menggunakan data yang tidak seimbang, model *machine learning* yang dibangun akan lebih condong untuk mempelajari pola kelas mayoritas dan mengabaikan kelas minoritas. Hal tersebut mengakibatkan model menjadi lemah dan menghasilkan performa yang buruk. Dengan demikian, meskipun nilai akurasi yang tinggi dapat diperoleh, nilai metriks evaluasi yang lain seperti *recall*, *precision*, *F1 score*, dan ROC menjadi tidak cukup baik. Untuk mengatasi masalah ketidakseimbangan pada data dapat digunakan teknik *under/oversampling*.

Undersampling adalah salah satu teknik sampling dengan mengurangi porsi dari kelas mayoritas sehingga memiliki porsi yang seimbang dengan kelas minoritas. Sedangkan *oversampling* adalah teknik untuk mengatasi ketidakseimbangan dengan cara meningkatkan jumlah atau porsi kelas minoritas hingga sama atau hampir sama dengan kelas mayoritas. Ilustrasi dari *undersampling* dan *oversampling* dapat dilihat pada gambar di bawah ini.



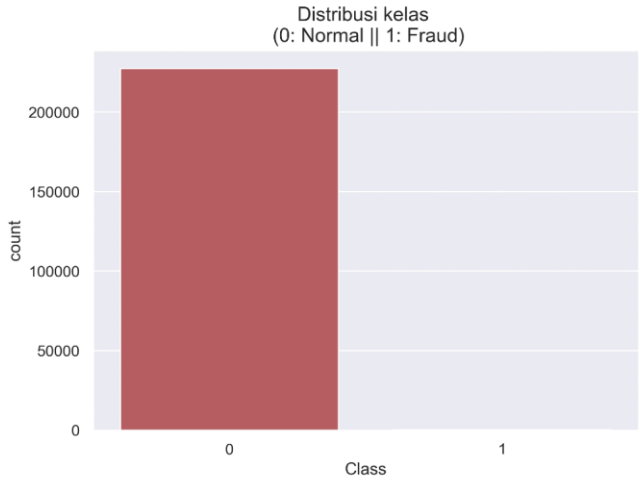
Gambar 6. *Undersampling* dan *Oversampling*

Teknik *under/oversampling* memiliki beberapa metode. Namun dalam penelitian ini, hanya menerapkan teknik *oversampling* dengan SMOTE (*Synthetic Minority Oversampling Technique*). SMOTE adalah teknik *oversampling* dengan membuat sampel sintetis dari kelas minoritas. SMOTE bekerja lebih baik dibandingkan dengan teknik *oversampling* biasa. SMOTE digunakan untuk memperoleh kelas yang seimbang atau setidaknya hampir seimbang dalam data pelatihan yang akan digunakan untuk melatih pengklasifikasi atau model *machine learning*.

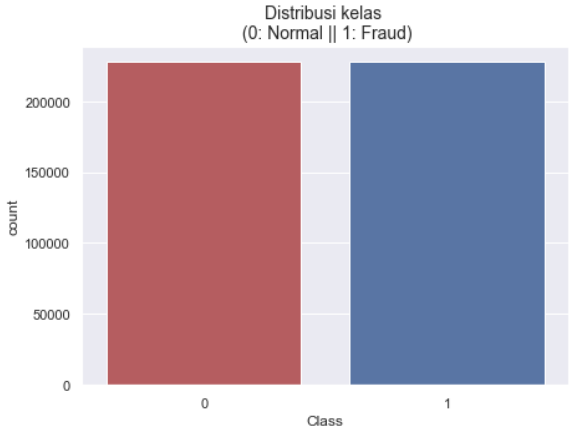
HASIL DAN PEMBAHASAN

Hasil dari persiapan, eksplorasi dan transformasi data diketahui bahwa dalam data tidak terdapat Null, NA, NaN, *Missing Value* dan nilai error lain yang berarti data sudah bersih. Data set merupakan tipe data numerik dan memiliki porsi antar kelas yang sangat tidak seimbang, sehingga dibutuhkan teknik *oversampling* menggunakan SMOTE guna mendapatkan porsi kelas yang seimbang agar mesin dapat mempelajari data dengan lebih baik. Jumlah data pelatihan sebelum dilakukan oversampling adalah 227.845 dengan

transaksi normal [0] 227.454 dan fraud [1] 391. Dan setelah dilakukan *oversampling* jumlah data pelatihan menjadi 454.908 dengan transaksi normal [0] 227.454 dan fraud [1] 227.454. Adapun visualisasinya dapat dilihat di bawah ini.



Gambar 7. Data pelatihan sebelum penerapan *oversampling* dengan SMOTE



Gambar 8. Data pelatihan sebelum penerapan *oversampling* dengan SMOTE

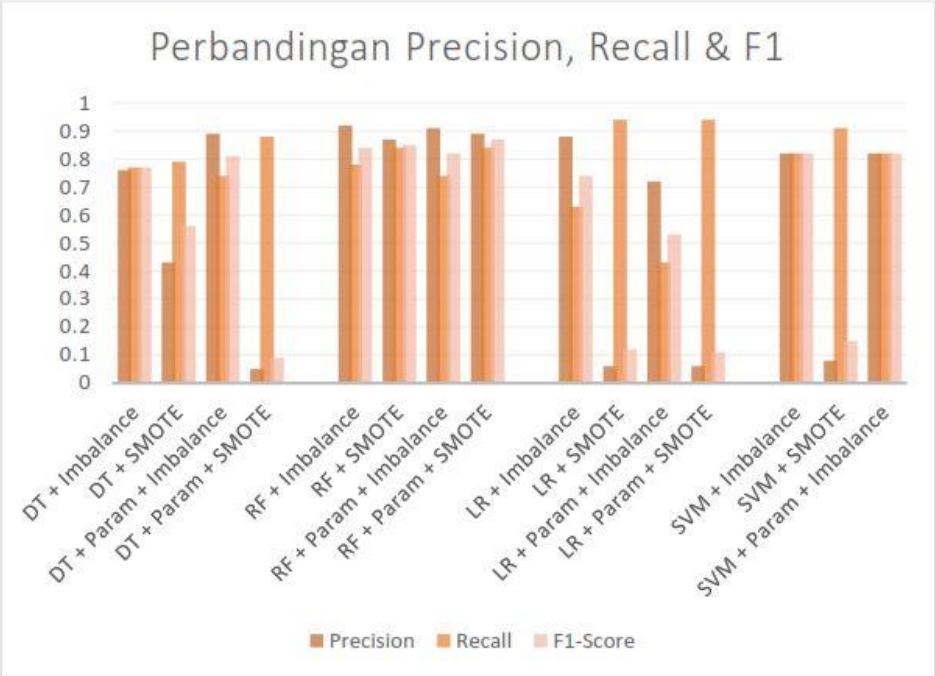
Setelah membuat 15 model dengan 4 algoritma *supervised learning* dan penerapan teknik *oversampling* pada data serta menggunakan parameter terbaik yang ditentukan dengan fungsi GridSearchCV, didapatkan performa masing-masing percobaan sebagaimana ditunjukkan pada tabel 1.

Secara keseluruhan, model yang menggunakan algoritma *logistic regression* (LR) adalah model yang paling lemah setelah *support vector machine* (SVM). Sementara model dengan algoritma *random forest* (RF) memiliki performa paling baik di antara semua model, disusul oleh model dengan algoritma *decision tree* (DT), hal ini pula yang membuktikan bahwa *ensemble learning* adalah teknik untuk meningkatkan performa model yang lemah. Berikut adalah perbandingan nilai performa berdasarkan evaluasi metriks dari seluruh algoritma dan teknik yang digunakan.

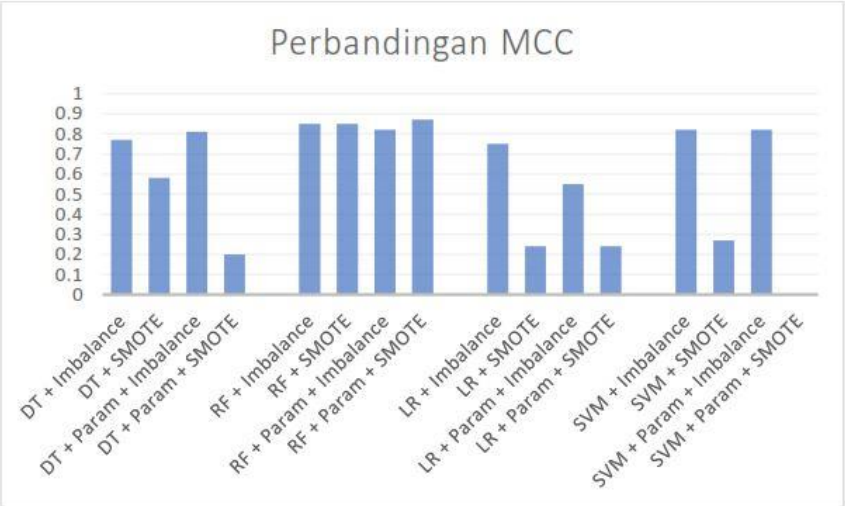
Tabel 1. Perbandingan skor Berbagai Model

Model	Precision	Recall	F1-Score	MCC	ROC	AUC
DT + Imbalance	*0.76*	*0.77*	*0.77*	*0.77*	*0.89*	*0.89*
DT + SMOTE	*0.43*	*0.79*	*0.56*	*0.58*	*0.89*	*0.90*
DT + Param + Imbalance	*0.89*	*0.74*	*0.81*	*0.81*	*0.87*	*0.93*
DT + Param + SMOTE	*0.05*	*0.88*	*0.09*	*0.20*	*0.93*	*0.95*
RF + Imbalance	*0.92*	*0.78*	*0.84*	*0.85*	*0.89*	*0.95*
RF + SMOTE	*0.87*	*0.84*	*0.85*	*0.85*	*0.92*	*0.98*
RF + Param + Imbalance	*0.91*	*0.74*	*0.82*	*0.82*	*0.87*	*0.95*
RF + Param + SMOTE	*0.89*	*0.84*	*0.87*	*0.87*	*0.92*	*0.99*
LR + Imbalance	*0.88*	*0.63*	*0.74*	*0.75*	*0.82*	*0.98*
LR + SMOTE	*0.06*	*0.94*	*0.12*	*0.24*	*0.96*	*0.99*
LR + Param + Imbalance	*0.72*	*0.43*	*0.53*	*0.55*	*0.71*	*0.92*
LR + Param + SMOTE	*0.06*	*0.94*	*0.11*	*0.24*	*0.96*	*0.98*
SVM + Imbalance	*0.82*	*0.82*	*0.82*	*0.82*	*0.91*	*0.95*
SVM + SMOTE	*0.08*	*0.91*	*0.15*	*0.27*	*0.95*	*0.98*
SVM + Param + Imbalance	*0.82*	*0.82*	*0.82*	*0.82*	*0.91*	*0.96*
SVM + Param + SMOTE	*Na*	*Na*	*Na*	*Na*	*Na*	*Na*

Di bawah ini perbandingan berdasarkan nilai yang sama, dalam diagram batang.



Gambar 9. Perbandingan *precision*, *recall* & *f1-score* pada Berbagai Model



Gambar 10. Perbandingan skor *Matthews Correlation Coefficient* (MCC)

KESIMPULAN DAN REKOMENDASI

Berdasarkan hasil penelitian dan pembahasan dapat disimpulkan bahwa algoritma *random forest* adalah algoritma yang paling cocok untuk digunakan dalam pengembangan sistem pendeteksi penipuan transaksi kartu kredit, karena menghasilkan nilai performa keseluruhan paling baik dan stabil dibandingkan dengan algoritma lain. Nilai performa yang dihasilkan model dengan algoritma *random forest* lebih tinggi dibandingkan dengan nilai dengan algoritma *decision tree.*

Pada penelitian ini, beberapa rekomendasi saran yang dapat dipertimbangkan untuk penelitian berikutnya:

1. Menggunakan *cross validation* untuk melakukan evaluasi kinerja model atau algoritma.
2. Melakukan pendekatan lain seperti terlebih dahulu mengelompokkan transaksi berdasarkan jumlah transaksi.
3. Menggunakan perangkat komputasi yang lebih mumpuni agar proses kalkulasi saat melatih model dapat dilakukan dengan lebih cepat.

REFERENSI

Abdou, H., Delamaire, L. & Pointon, J., (2009). Credit Card Fraud And Detection Techniques: A Review. *Banks and Banks Systems*, 4(2), pp. 57-68.

Abdullah, M., Mohammed, R. & Rawashdeh, J., (2020). Machine Learning with Oversamplind and Undersampling Techniques: Overview Study and Experimental Results. *ResearchGate*, pp. 1-6.

Anne, B. L., Probst, P. & Wright, M., (2019). Hyperparameters and Tuning Strategies for Random Forest. *WIREs Data Mining And Knowledge Discovery*, pp. 1-19.

Asha, R. & Kumar, S. K., (2021). Credit Card Fraud Detection Using Artificial Neural Network. *Global Transitions Proceedings*, Volume 2, pp. 35-41.

Blagus, R. & Lusa, L., (2013). SMOTE for High-Dimensional Class-Imbalanced Data. *Blagus and Lusa BMC Bioinformatics*, 14(106), pp. 1-16.

Blondel, M. et al., (2011). Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research*, Volume 12, pp. 2825-2830.

Bowers, A. J. & Zhou, X., (2019). Receiver Operating Characteristic (ROC) Area Under the Curve (AUC): A Diagnostic Measure for Evaluating the Accuracy of Predictors of Education Outcomes. *Journal of Education for Students Placed At Risk* (JESPAR), 24(1), pp. 1-19.

Breiman, L., (2001). Random Forest. *Statistics Department, University of California*, Volume 45, pp. 5-32.

Chicco, D. & Jurman, G., (2020). The Advantages Of The Matthews Correlation Coefficient (MCC) Over F1 Score And Accuracy In Binary Classification Evaluation. *Chicco and Jurman BMC Genomics*, 21(6), pp. 1-13.

Dewi, N. K., Mulyadi, S. Y. & Syafitri, U. D., (2011). Penerapan Metode Random Forest dalam Driver Analysis. *Forum Statistika dan Komputasi*, 16(1), pp. 35-43.

Dornadula, V. N. & Geetha, S., (2019). Credit Card Fraud Detection Using Machine Learning Algorithms. *Procedia Computer Science*, Volume 165, pp. 631-641.

Drummond, A. C. & Resende, P. A. A., (2018). A Survey of Random Forest Based Methods for Intrusion Detection Systems. *ACM Computing Surveys*, 51(3), pp. 1-36.

Elkan, C., Lipton, Z. C. & Naryanaswamy, B., (2014). Thresholding Classifiers to Maximize F1 Score. *University of California*, pp. 1-16.

Ezukwoke, K. & Zareian, S., (2019). Logistic Regression And Kernel Logistic Regression A Comparative Study Of Logistic Regression And Kernel Logistic Regression For Binary Classification. *ResearchGate*, pp. 1-10.

Ghamisi, P. et al., (2020). Support Vector Machine Versus Random Forest for Remote Sensing Image Classification: A Meta-Analysis and Systematic Review. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, Volume 13, pp. 6308-6325.

Hamori, S., Kawai, M., Kume, T. & Murakami, Y., (2018). Ensemble Learning or Deep Learning? Application to Default Risk Analysis. *Journal of Risk and Financial Management*, 11(12), pp. 1-14.

Hao, J. & Ho, T. K., (2019). Machine Learning Made Easy: A Review of Scikit-learn Package in Python Programming Language. *Journal of Educational and Behavioral Statistics*, 44(3), pp. 348-361.

Hendarsyah, D., (2020). Analisis Prilaku Konsumen Dan Keamanan Kartu Kredit Perbankan. *Jurnal Perbankan Syariah*, 1(1), pp. 85-96.

Hermuningsih, S., Irmawati & Rahayu, F. A., (2011). Perkembangan Kartu Kredit di Indonesia. *Jurnal Manajemen*, 1(1), pp. 5-13.

Kavila, K. D. & Lakshmi, S. V. S. S., (2018). Machine Learning For Credit Card Fraud Detection System. *International Journal of Applied Engineering Research*, 13(24), pp. 16819-16824.

Kurniawan, A. & Yulianingsih, (2021). Pendugaan Fraud Detection pada Kartu Kredit dengan Machine Learning. *Kilat*, 10(2), pp. 320-325.

LaValley, M. P., (2008). Logistic Regression. *Circulation*, pp. 2395-2399.

Mahesh, B., (2018). Machine Learning Algorithms - A Review. *International Journal of Science and Research*, 9(1), pp. 381-386.

Mewengkang, F. R., Ratulangi, C. H. & Wahongan, A. S., (2021). Tindak Pidana Cyber Crime dalam Kegiatan Perbankan. *Lex Privatum*, 9(5), pp. 179-187.

Panda, R. M. & Sagar, B. S. D., (2022). Decision Tree. *Encyclopedia of Mathematical Geosciences*, pp. 1-14.

Patel, B. R. & Rana, K. K., (2014). A Survey on Decision Tree Algorithm For Classification. *International Journal of Engineering Development and Research*, 2(1), pp. 1-5.

Powers, D. M., (2011). Evaluation: From Precision, Recall And F-Measure To ROC, Informedness, Markedness & Correlation. *Flinders University*, pp. 37-63.

Prasetyo, A. & Sofyan, S., (2021). Penerapan Synthetic Minority Oversampling Technique (SMOTE) Terhadap Data Tidak Seimbang Pada Tingkat Pendapatan Pekerja Informal Di Provinsi D.I. Yogyakarta Tahun 2019. *Seminar Nasional Official Statistics*, pp. 868-877.

Siringoringo, R., (2018). Klasifikasi Data Tidak Seimbang Menggunakan Algoritma Smote Dan K-Nearest Neighbor. *Jurnal Information System Development*, 3(1), pp. 44-49.

Subekti, A. & Syukron, A., (2018). Penerapan Metode Random Over-Under Sampling dan Random Forest untuk Klasifikasi Penilaian Kredit. *Jurnal Informatika*, 5(2), pp. 175-185.

Yazid & Fiananta, A., (2017). Mendeteksi Kecurangan Pada Transaksi Kartu Kredit Untuk Verifikasi Transaksi Menggunakan Metode SVM. *Indonesian Journal of Applied Informatics*, 1(2), pp. 61-66.