Pendeteksian *Fraud E-Channel* menggunakan Algoritma Pembelajaran Mesin

ADONG PURBA
Bank Jawa Barat (BJB)
Indonesia

ABSTRAK

Perbankan adalah sektor yang paling sering menjadi korban penipuan menggunakan transaksi *e-channel*, salah satunya adalah menggunakan *Automatic Teller Machine* (ATM). *Fraud* adalah tindakan penyimpangan atau kelalaian yang sengaja dilakukan untuk menipu atau memanipulasi pelanggan, atau pihak lain, yang terjadi di bank atau menggunakan fasilitas bank sehingga menyebabkan pihak lain menderita kerugian dan pelaku penipuan mendapatkan keuntungan finansial baik langsung atau tidak langsung. Untuk mengendalikan penipuan, bank wajib memiliki dan menerapkan strategi anti-*fraud* yang efektif dengan menganalisis data transaksi untuk mencari pola yang mencurigakan sehingga memudahkan identifikasi transaksi sebagai transaksi yang sah atau tidak. Pada bidang data sains, kasus transaksi *fraud* dipandang sebagai permasalahan klasifikasi pembelajaran mesin. Pada penelitian ini telah dilakukan analisis data dan implementasi algoritma pembelajaran mesin untuk mendeteksi transaksi *fraud*. Tahapan eksplorasi data dan pra proses merupakan bagian yang sangat penting sebelum implementasi algoritma. Dari 6 algoritma yang diuji *Random Forest* (RF) memberikan hasil yang paling baik diikuti *Logistic Regression* (LR), *Linear Discriminant Analysis* (LDA) dan *Support Vector Machine* (SVM).

Kata Kunci: *bank, fraud detection, e-channel, algoritma pembelajaran mesin*

ABSTRACT

Banking is a sector that is the most frequent victim of fraud using e-channel transactions, one of which is using an Automatic Teller Machine (ATM). Fraud is an act of irregularity or negligence that is intentionally carried out to deceive or manipulate customers, or other parties, which occurs at a bank or uses bank facilities so as to cause the other party to suffer losses and the fraudster to get financial gain either directly or indirectly. To control fraud, banks are required to have and implement an effective anti-fraud strategy by analyzing transaction data to look for suspicious patterns to make it easier to identify transactions as legitimate or not. In the field of data science, fraud transaction cases are seen as a classification problem for machine learning. In this study, data analysis and machine learning algorithms were implemented to detect fraudulent transactions. The data exploration and preprocessing stages are very important parts before algorithm implementation. Of the 6 algorithms tested, Random Forest (RF) gave the best results followed by Logistic Regression (LR), Linear Discriminant Analysis (LDA) and Support Vector Machine (SVM).

Keywords*: banking, fraud detection, e-channel, machine learning algorithm*
JEL Classification: C38

1. Pendahuluan

Transaksi elektronik pada sektor perbankan yang biasa dikenal dengan *e-banking* atau *e-channel* telah berkembang pesat dengan berbagai *channel* seperti *Automated Teller Machine* (ATM), *Electronic Data Capture* (EDC), internet banking, SMS banking dan mobile banking. Pada tahun 2014 Otoritas Jasa Keuangan (OJK) mencatat volume *e-banking* di Indonesia sudah mencapai 6.447 triliun dan pada tahun 2016 jumlah pengguna *e-banking* di Indonesia mencapai 50,4 juta nasabah, dengan frekuensi 405,4 juta transaksi. Perkembangan tersebut dalam praktiknya disamping telah memberikan berbagai kemudahan bagi nasabah namun juga menimbulkan berbagai bentuk modus operandi tindak pidana yang menimbulkan kerugian bagi nasabah.

Dari semua transaksi *e-channel* tersebut diatas sering terjadi masalah penipuan (*fraud*) yang menggunakan rekening bank sebagai media untuk menerima hasil kejahatan, sehingga menuntut bank untuk dapat bertindak cepat dalam rangka melindungi kepentingan nasabah yang menjadi korban penipuan. Untuk dapat melindungi kepentingan nasabah korban penipuan diperlukan tindakan bank untuk segera melakukan deteksi *fraud* (*fraud detection*) untuk mengurangi dampak risiko khususnya risiko reputasi, risiko operasional (karena ada kerugian untuk mengganti uang nasabah) dan risiko hukum (atas tuntutan nasabah karena terdapat kelemahan sistem di internal bank). Dari permasalahan diatas diperlukan sebuah cara yang efektif untuk mencegah *fraud* yang terjadi pada transaksi perbankan. Saat ini ada ilmu baru yang bernama *data science* yang bisa dimanfaatkan untuk memberikan solusi atas permasalahan diatas.

Data science adalah suatu teknik analisa yang membutuhkan keterampilan (Asniar & Surendro, 2014) rekayasa perangkat lunak. *Data science* juga bisa disebut mengekstrak suatu data agar bisa difilter dan ditemukan data yang benar adanya agar bisa menghasilkan produk data yang sebenarnya. Selain

bidang pemrograman seseorang harus memiliki beberapa kemampuan dan pegetahuan yang cukup dibidang lain seperti matematika dan statistik agar bisa menyaring data dengan cara yang cepat, agar bisa menganalisis data dengan baik dan benar melalui model probabilitas, program komputer dan hal yang berkaitan dengan ilmu sains. *Data science* merupakan sebuah proses yang akan mengubah data dengan memanfaatkan ilmu matematika dan statistika sehingga menghasilkan wawasan, keputusan dan produk yang berguna bagi individu maupun perusahaan dalam mengambil keputusan yang strategis (Asniar & Surendro, 2014).

Fraud detection dapat dikategorikan sebagai permasalahan klasifikasi biner karena *output* dari metode ini ada 2 yaitu *fraud* dan *non-fraud*. Pada keilmuan *data science* terdapat beberapa metode yang umum digunakan untuk menyelesaikan masalah klasifikasi biner, antara lain *Logistic Regeression* (LR), *Linear Discriminant Analysis* (LDA), *K-Nearest Neighbours* (KNN), *Classification and Regression Tree* (CART), *Support Vector Machine* (SVM) dan *Random Forest* (RF). Namun penggunakan metode-metode ini untuk kasus *fraud detection* pada sektor perbankan memiliki tantangan tersendiri. Pertama, adanya *imbalance* kelas pada dataset transaksi perbankan dimana data *non-fraud* memiliki jumlah yang jauh lebih banyak dibandingkan data *fraud*. Kedua, adanya variasi perilaku penipuan. Ketiga, masalah sensitif biaya dimana kerugian kesalahan klasifikasi memberikan dampak berbeda untuk *false positive* dan *false negative*. Keempat, metrik evaluasi yang digunakan. Pada paper ini akan disajikan langkah utuh implementasi pembelajaran mesin untuk kasus pendeteksian transaksi *fraud* pada perbankan mulai dari tahap ekplorasi data, pra proses, implementasi *code* dan pengujian pada beberapa algoritma yang sudah dipilih.

Berdasarkan penjelasan diatas, penelitian ini bertujuan untuk melakukan analisis data dan implementasi algortima pembelajaran mesin untuk mendeteksi transaksi *fraud* pada sektor perbankan.

2. Data dan Metodologi

2.1. Data

Pada data yang digunakan terdapat 28 atribut pada dataset, ke 28 atribut data itu yang ditunjukkan pada Tabel 1:

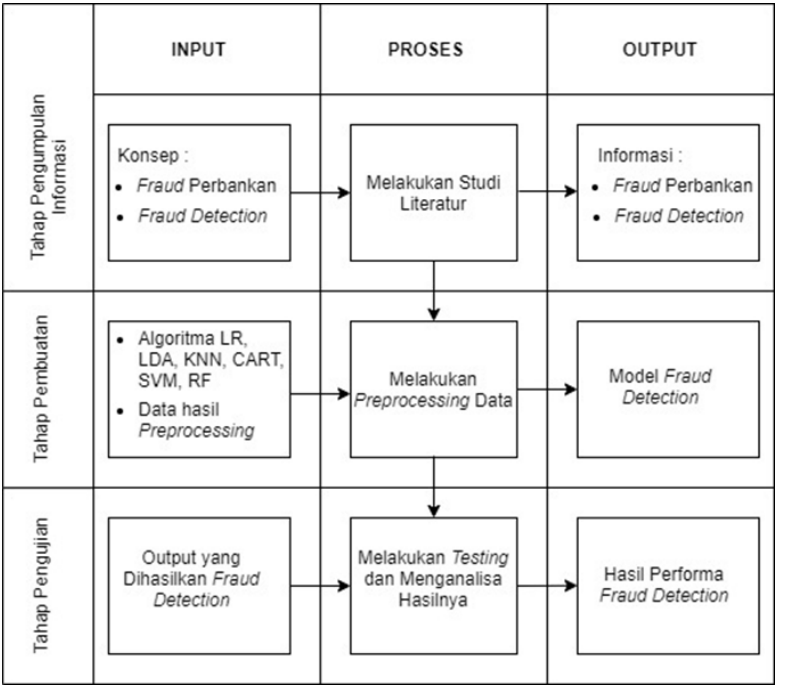
Tabel 1 Atribut dan Deskripsi

No	Atribut	Deskripsi
1.	Id_Transaksi atau Id_Tx	Serangkaian huruf dan angka yang digunakan untuk mengidentifikasi setiap transaksi yang terjadi.
2.	Id_tanggal_transaksi_awal	Huruf dan angka yang digunakan untuk mengidentifikasi pertukaran mata uang digital antara bank dan nasabah.
3.	Tanggal_transaksi_awal	Tahun, bulan dan tanggal yang tercatat pada transaksi digital di sistem bank.
4.	Tipe_kartu	Jenis produk kartu yang dimiliki bank (seperti *gold*, *silver*, *platinum*, *classic*) yang memiliki fitur limited amount, limited frekuensi penarikan, dan biaya administrasi.
5.	Id_*merchant*	Id yang terdapat pada mesin baik itu Debit/Kredit yang memiliki *uniqe number* pada mesin.
6.	Nama_*merchant*	Nama daerah *merchant* berada (misal: daerah A, B, C, D), berdasarkan *uniqe number* pada mesin.
7.	Tipe_mesin	Tipe dari mesin. (misal ATM: NDC, DDC, CRM).
8.	Tipe_transaksi	Bisa juga disebut sebagai jenis transaksi yang berkaitan erat dengan kode transaksi. (misal : tarik tunai, setor tunai, *transfer*, *payment* (PBB, PLN, PDAM)).
9.	Nama_transaksi	Fitur transaksi yang bisa dilakukan oleh nasabah.
10.	Nilai_transaksi	Maksimum atau minimum nilai transaksi yang bisa dilakukan oleh nasabah berdasarkan kebijakan internal bank.
11.	Id_negara	Multi *currency* transaksi yang bisa dilakukan di sistem bank (boleh atau tidak boleh multi *currency* beda negara) berdasarkan jarak negara A ke B.
12.	Nama_negara	Kode mata uang yang diberikan untuk masing-masing negara. (misal: Indonesia = 360 IDR).
13	Nama_kota	Nama tempat *e-channel* berada (melakukan operasional).

No	Atribut	Deskripsi
14.	Lokasi_mesin	Keberadaan mesin yang di daftarkan di google maps dan lokasi menggunakan GPS.
15.	Pemilik_mesin	Unit kerja (kantor cabang) pemilik untuk penghitungan jurnal. (misal: Kas ATM > < No Rekening nasabah Kredit > < Debet Tarikan tunai.)
16.	Waktu_transaksi	Waktu transaksi selama 24 jam (Waktu transaksi < Waktu tempuh minimal).
17.	Kuartal_transaksi	Durasi transaksi per 4 (empat) bulan (KW I, KW II, KW III dan KW IV).
18.	Kepemilikan_kartu	Kepemilikan kartu berdasarkan perorangan dan perusahaan.
19.	Nama_*channel*	Bisa disebut juga tipe *channel*. (misal: ATM, EDC, SMS Banking, Internet Banking, Mobile Banking)
20.	Id_channel	Bisa disebut juga kode *channel* atau kode *merchant*. (misal: 6011: ATM, 6010: Teller, 6012: Internet Banking)
21.	Flag_transaksi_finansial	Pendefenisian yang diberikan terhadap sukses atau gagalnya suatu transaksi. (misal: -*False* = Gagal, -*True* (*Good*) = Sukses)
22.	Status_transaksi	Keterangan transaksi berdasarkan flag_transaksi_finansial. (misal: -*False* = Gagal -*True* (*Good*) = Sukses)
23.	Bank_pemilik_kartu	Bank penerbit kartu yang digunakan oleh nasabah.
24.	Rata_rata_nilai_transaksi (Rata-rata *amount*)	Jumlah transaksi dalam 1 (satu) hari. (misal : 1 hari = 1 juta, 10 juta)
25.	Maksimum_nilai_transaksi	Nilai transaksi berdasarkan tipe kartu dan limit kartu.
26.	Minimum_nilai_transaksi	Transaksi minimum yang dapat dilakukan sesuai ketentuan internal bank dan tidak berpengaruh ke tipe kartu.
27.	Rata_rata_jumlah_transaksi	Jumlah transaksi yang dilakukan per hari. (misal : 1 kali sehari dalam sebulan)
28.	Flag_transaksi_*fraud*	Label diberikan untuk menyatakan transaksi itu *Fraud* atau *Non Fraud*.

2.2 Metodologi

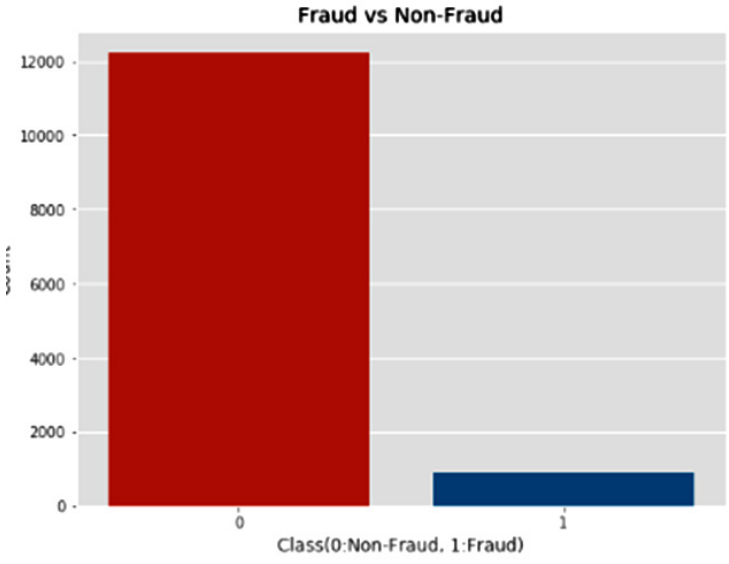
Pada Gambar 1 menggambarkan tahapan langkah penelitian yang dilakukan dalam studi tentang deteksi *fraud* dalam *data science*.



Gambar 1 Metodologi Penelitian

Exploratory Data Analysis (EDA) merupakan salah satu metode pada statistika deskriptif yang digunakan untuk penyajian data yang nantinya akan bermanfaat menjadi informasi yang berguna. EDA ini dapat digunakan untuk mengetahui pola data serta bentuk sebarannya. Untuk mengidentifikasi pola data dan sebarannya ada beberapa metode yang dapat digunakan diantaranya diagram batang dan histogram.

Persentase transaksi yang *fraud* adalah 6.933% dan tidak *fraud* 93.067% dari total keseluruhan jumlah transaksi 12215.



Gambar 2 Distribusi *Fraud* VS *Non Fraud*

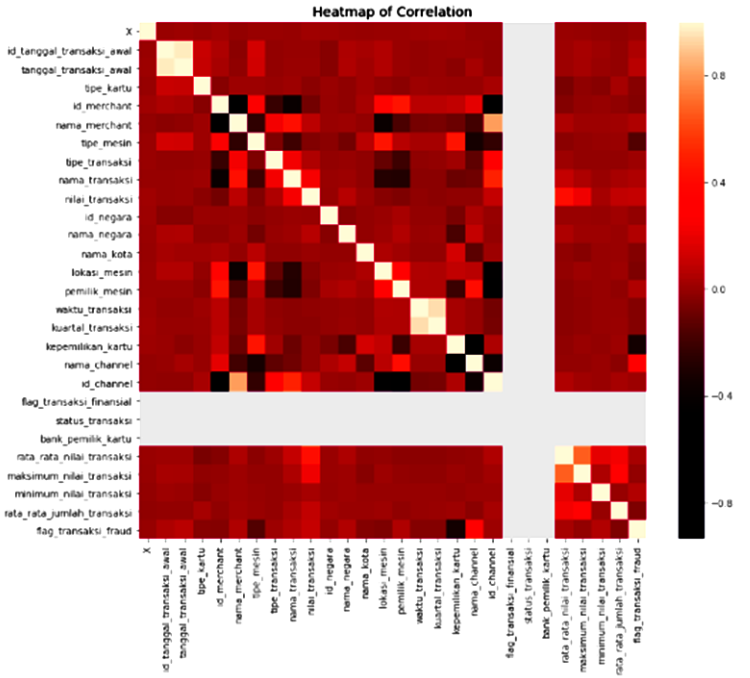
Heatmap of Correlation adalah cara untuk menghitung kovarian antara variable, metrik kovarian kemudian dapat dengan mudah divisualisasikan sebagai *heatmap*. *Heatmap* secara efektif merupakan *pseudocolor* plot dengan baris dan kolom berlabel dengan nilai korelasi berkisar antara -1 dan 1. Ada dua komponen kunci dari nilai korelasi:

1. *Magnitude*

   Semakin besar magnitude (semakin dekat ke 1 atau -1) semakin kuat korelasinya (pada gambar ditunjukan warna putih dan hitam).

2. *Sign*

   Jika negatif ada korelasi terbalik, jika positif korelasi berbanding lurus.

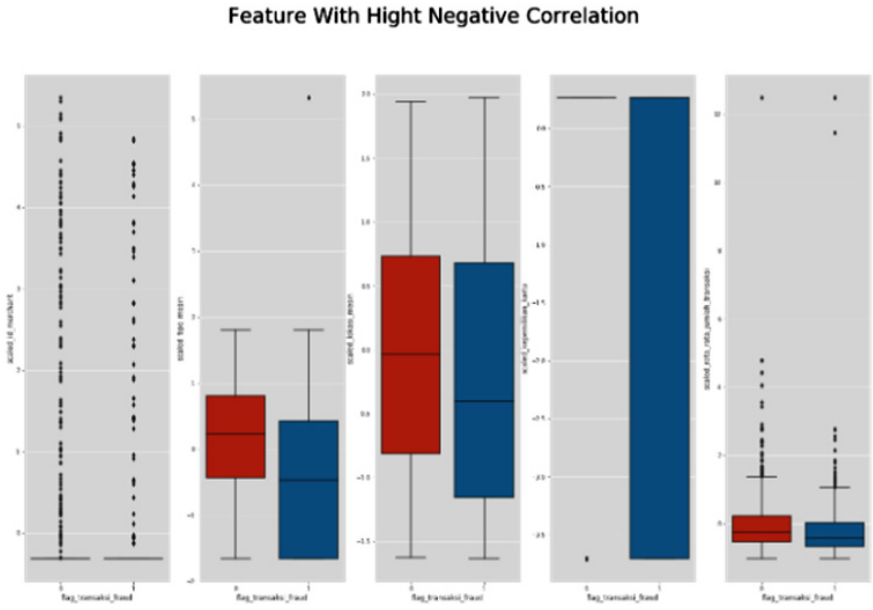


Gambar 3 Heatmap Korelasi Antar Atribut

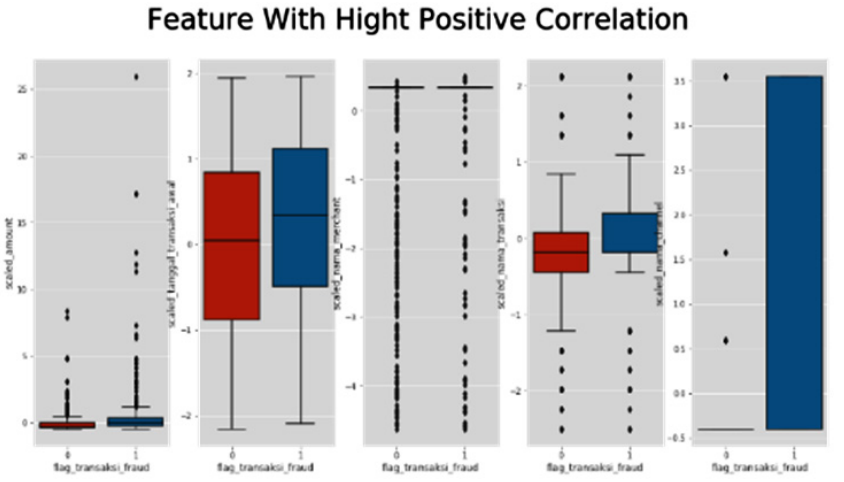
Skew adalah ukuran dari asimetri distribusi probabilitas dari variabel acak bernilai *riil* tentang nilai tengahnya. Nilai *skewness* bisa positif atau negatif, atau tidak terdefinisi. Untuk model distribusinya, condong negatif biasanya menunjukan bahwa ekor grafik berada di sisi kiri distribusi, dan condong positif menunjukan bahwa ekor berada di sebelah kanan. Sedangkan untuk distribusi simetris ekor di kedua sisi rata-rata menyeimbangkan keseluruhan.

Pra proses data adalah suatu proses pengubahan bentuk data yang belum terstruktur menjadi data yang terstruktur sesuai dengan kebutuhan, untuk proses *mining* yang lebih lanjut. Berikut *preprocessing* data yang dilakukan pada penelitian ini:

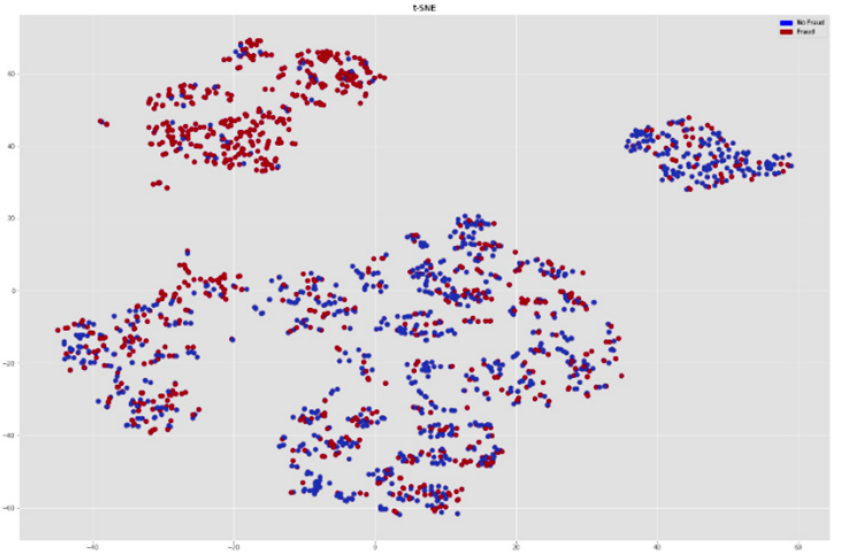
Dari 28 atribut dari dataset maka dipilih 22 fitur berdasarkan korelasi antara *flag_transaksi_fraud* dengan atribut lainnya (lihat gambar IV.5 Heatmap Korelasi antar Atribut). Dari hasil tabulasi korelasi antara *flag_transaksi_fraud* dengan atribut lainnya terdapat 3 (tiga) atribut dengan nilai korelasi NaN, yaitu flag_transaksi_finansial, status_transaksi, dan bank_pemilik_kartu. Hal ini dikarenakan semua nilai pada atribut tersebut sama. Oleh karena itu ketiga attribut ini di *drop* (dibuang). Kemudian attribut dengan nilai korelasi antara -0.1 s.d 0.1 juga di *drop*.



Gambar 4 Lima Fitur dengan Korelasi Negatif Terbesar



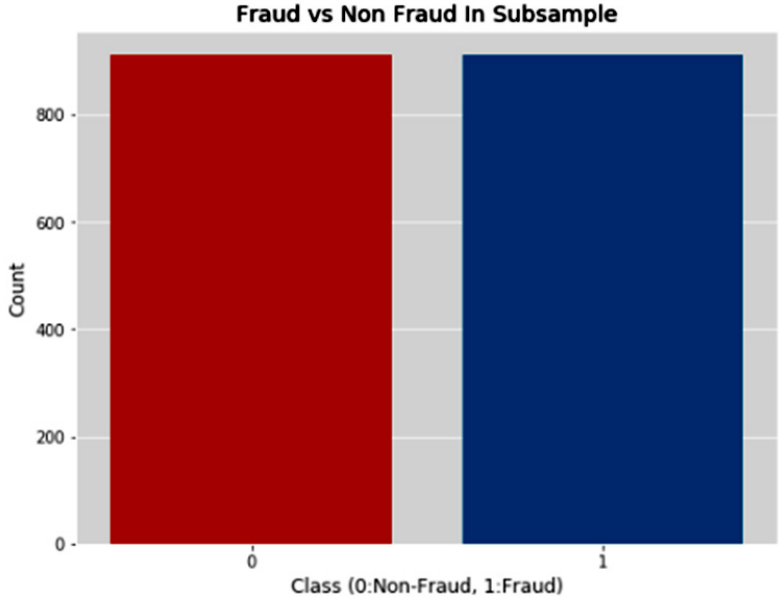
Gambar 5 Lima Fitur dengan Korelasi Positif Terbesar



Gambar 6 t-SNE *Scatter Plot*

Scaling atau standarisasi fitur adalah langkah *pre processing* data yang diterapkan pada variabel independen atau fitur data. Ini dilakukan pada dasarnya membantu untuk menormalkan data dalam rentang tertentu, terkadang juga membantu mempercepat perhitungan dalam suatu algoritma. Proses *scaling* diatas sangat penting dilakukan untuk meningkatkan nilai *accuracy* dari *precision*, *recall* dan *f1-score* pada table *predictive* dan *actual confusion matrix*.

Seperti pada *point* sebelumnya, bahwa *fraud* : *non fraud* = 910 : 12215. Ketimpangan ini dapat menyebabkan model menjadi kurang maksimal sehingga untuk menyeimbangkan jumlah data dilakukan pengambilan sampel secara acak dari data *non fraud* sebanyak 910. Dengan demikian jumlah data *fraud* : *non fraud* = 1: 1.



Gambar 7 Distribusi *Fraud* VS *Non Fraud* Setelah Pra Proses

Pada eksperimen ini dipilih 6 algoritma yang paling sering digunakan dalam kasus pendeteksian *fraud*:

Logistic Regression

Algoritma Regresi Logistik adalah metode analisis statistik yang digunakan untuk memprediksi nilai data berdasarkan pengamatan sebelumnya atas suatu kumpulan data (Puh & Brkić, 2019). Regresi logistik telah menjadi alat penting dalam disiplin pembelajaran mesin. Pendekatan tersebut memungkinkan algoritma digunakan dalam aplikasi pembelajaran mesin untuk mengklasifikasikan data yang masuk berdasarkan data historis. Saat data yang lebih relevan masuk, algoritma harus menjadi lebih baik dalam memprediksi klasifikasi dalam kumpulan data. Regresi logistik juga dapat berperan dalam aktivitas persiapan data dengan memungkinkan kumpulan data dimasukkan ke dalam *bucket* yang telah ditentukan sebelumnya secara khusus selama proses ekstrak, transformasi, *load* (ETL) untuk menyusun informasi untuk analisis. Model regresi logistik memprediksi variabel data dependen dengan menganalisis hubungan antara satu atau lebih variabel independen yang ada.

Linear Discriminant Analysis

Algoritma Analisis Diskriminan Linier (LDA) adalah jenis kombinasi linier pada proses matematika yang menggunakan berbagai item data dan menerapkan fungsi ke himpunan untuk menganalisis secara terpisah beberapa kelas objek atau item. Analisis diskriminan linier dapat berguna di berbagai bidang seperti pendeteksian, pengenalan gambar dan analisis prediktif dalam pemasaran (Mahmoudi & Duman, 2015). Analisis diskriminan linier membantu merepresentasikan data lebih dari dua kelas, ketika regresi logika tidak cukup. Analisis diskriminan linier mengambil nilai rata-rata untuk setiap kelas dan mempertimbangkan varian untuk membuat prediksi dengan asumsi distribusi

Gaussian. Ini adalah salah satu dari beberapa jenis algoritma yang merupakan bagian dari pembuatan model *machine learning* yang kompetitif.

K-Nearest Neighbours

K Nearest Neighbors atau KNN adalah salah satu algoritma pembelajaran mesin untuk melakukan klasifikasi terhadap objek baru berdasarkan sejumlah k tetangga terdekatnya. Untuk tetangga terdekatnya ditentukan oleh analis yang dinyatakan dengan k. Misal nilai k=3, maka setiap data testing dihitung jaraknya terhadap data training dan dipilih 3 data training yang jaraknya paling dekat dengan data *testing*. Tujuan penggunaan KNN adalah untuk memprediksi objek, apakah objek tersebut masuk dalam satu golongan tertentu atau golongan yang lain. Pada KNN data akan dinyatakan dalam ruang *vector.* Sesuai dengan namanya, "*nearest neighbor*", KNN menggunakan klasifikasi berdasarkan "kedekatan" dengan tetangga. KNN merupakan algoritma pembelajaran mesin yang paling sederhana dan dapat digolongkan sebagai *supervised learning*, *lazy learning* algorithm, dan *instance-based learning* atau *memory-based learning*.

Disebut *lazy* karena KNN tidak menggunakan sampel data latih untuk keperluan pembelajaran (generalisasi) atau hanya sedikit sekali tahapan pembelajaran. Sebagian besar waktu hanya dipakai untuk melakukan klasifikasi. Semua data diperlukan dan harus disimpan (tidak boleh dihapus). Disebut *instance-based* karena KNN tidak menggunakan asumsi/model apapun, sebagai gantinya KNN akan membentuk hipotesis secara langsung berdasarkan data latih yang disediakan, artinya semakin bertambah data akan semakin kompleks juga proses pencapaian hipotesis (ROSA, PRIMARTHA, & WIJAYA, 2020).

Classification and Regression Tree

Algoritma *Classifcation and Regression Tree* (CART) adalah metode teknik eksplorasi data berupa pohon keputusan. CART dikembangkan untuk

melakukan analisis klasifkasi pada variabel respon, baik yang nominal, ordinal, maupun kontinu. CART menghasilkan suatu pohon klasifikasi jika peubah responnya kategorikal dan menghasilkan pohon regresi jika peubah responnya kontinu (Lucaroni et al., 2019). Nilai tingkat kesalahan yang paling kecil pada pohon klasifikasi yang dihasilkan akan cenderung membuat pohon keputusan digunakan untuk memperkirakan respon. Prinsip dari metode pohon klasifikasi ini adalah memilah seluruh pengamatan menjadi dua gugus pengamatan dan memilah kembali gugus pengamatan tersebut menjadi dua gugus pengamatan berikutnya, hingga diperoleh jumlah pengamatan minimum pada tiap-tiap gugus pengamatan berikutnya.

Support Vector Machine

Algoritma *Support Vector Machine* (SVM) adalah algoritma pembelajaran mesin yang menganalisis data untuk klasifikasi dan analisis regresi. SVM merupakan metode pembelajaran yang diawasi yang melihat data dan mengurutkannya menjadi salah satu dari dua kategori. SVM mengeluarkan peta dari data yang diurutkan dengan margin di antara keduanya sejauh mungkin. SVM digunakan dalam kategorisasi teks, klasifikasi gambar, pengenalan tulisan tangan, dan sains (Sahin & Duman, 2011b).

Pembelajaran yang diawasi dalam hal ini artinya mengurutkan data menjadi dua kategori, dilatih dengan serangkaian data yang sudah diklasifikasikan ke dalam dua kategori dan membangun model seperti yang awalnya dilatih. Tugas dari algoritma SVM adalah untuk menentukan di kategori mana sebuah titik data baru berada. Hal ini membuat SVM menjadi semacam pengklasifikasi linier non-biner. Algoritma SVM seharusnya tidak hanya menempatkan objek ke dalam kategori, tetapi juga memiliki margin di antara objek tersebut pada grafik selebar mungkin.

Random Forest

Algoritma *Random Forest* merupakan modifikasi dari model *decision tree* yang cukup populer digunakan untuk berbagai masalah pembelajaran mesin karena mudah digunakan dan diinterpretasikan. Setiap pohon keputusan dengan sendirinya sensitif terhadap *overfitting*, tetapi jika digabungkan, mereka akan bekerja dengan baik. *Random Forest* adalah pengklasifikasi *bagging* dan menerapkan dua tingkat keputusan stokastik dalam proses pembelajarannya untuk setiap pohon keputusan individu dalam ansambel memilih subset sampel serta subset fitur untuk pelatihan (Kumar, Soundarya, Kavitha, Keerthika, & Aswini, 2019b).

3. Hasil dan Pembahasan

Skenario pengujian dilakukan dengan membagi dataset menjadi data latih dan data uji dengan perbandingan 0.8 : 0.2. Data latih digunakan untuk membentuk model dan data uji digunakan untuk evaluasi. Skenario pengujian terdiri dari 4 skenario yaitu:

1. Fitur *scaling* + *under sampling*
2. *Under sampling*
3. Fitur *scaling*
4. Tanpa *scaling* dan tanpa *sampling*

Semua algoritma diimplementasikan menggunakan bahasa pemograman python. Pengujian dilakukan dengan 10 *fold cross validation* kemudian diambil nilai rata-rata (*mean*) dan simpangan baku (SB). Hasil perhitungan AUC untuk keempat skenario disajikan pada Tabel 2.

Tabel 2 AUC Pada Berbagai Algortima Pembelajaran Mesin

Algoritma	Fitur *Scaling* + *Under Sampling*		*Under Sampling*		Fitur *Scaling*		Tanpa *Scaling* dan *Sampling*	
	Mean	SB	Mean	SB	Mean	SB	Mean	SB
LR	0.79	0.028	0.668	0.026	0.779	0.028	0.672	0.033
LDA	0.78	0.021	0.78	0.013	0.786	0.032	0.78	0.033
KNN	0.72	0.056	0.669	0.046	0.685	0.027	0.657	0.013
CART	0.677	0.033	0.66	0.044	0.622	0.022	0.616	0.029
SVM	0.775	0.046	0.5	0	0.704	0.041	0.5	0.001
RF	0.793	0.032	0.782	0.029	0.757	0.017	0.755	0.024

Dari Tabel 2 diatas dapat dilihat bahwa fitur *scaling* dapat meningkatkan AUC pada semua algoritma. Dilain sisi *under sampling* dapat meningkatkan AUC pada algoritma KNN, CART dan RF. Sementara pada LR dan LDA, *under sampling* malah menurunkan AUC namun tidak signifikan. Hal ini diakibatkan karena LR dan LDA tidak sensitif terhadap masalah *imbalance* kelas namun reduksi data karena adanya *under sampling* mengakibatkan data latih yang diigunakan menjadi lebih sedikit sehingga terjadi sedikit penurunan nilai AUC.

Pada skenario terbaik dari masing-masing algoritma, nilai AUC yang paling tinggi adalah RF namun tidak terlalu signifikan dibandingkan dengan LR, LDA dan SVM. Sementara CART dan KNN memiliki nilai AUC yang paling rendah.

4. Kesimpulan dan Rekomendasi

Dengan mengacu pada hasil penelitian maka adapun kesimpulan yang dapat diambil adalah sebagai berikut:

1. Telah dilakukan analisis data dan implementasi algoritma pembelajaran mesin untuk mendeteksi transaksi *fraud* pada perbankan. Tahap eksplorasi data dan pra proses merupakan tahapan yang sangat penting dalam implementasi pembelajaran mesin untuk kasus pendeteksi transaksi *fraud*.

2. Masalah *imbalance* kelas tidak sensitive pada algortima berbasis regresi seperti LR dan LDA namun penanganan masalah *imbalance* kelas dapat meningkatkan kinerja pendeteksian *fraud* dengan algoritma SVM, KNN, CART dan RF.
3. Dari 6 algoritma yang diajukan RF memiliki nilai yang paling baik diikuti LR, LDA dan SVM.

Adapun saran dan rekomendasi yang dapat dipertimbangkan untuk penelitian selanjutnya adalah sebagai berikut:

1. Perlu melakukan uji coba dengan dataset yang lebih besar.
2. Apabila memungkinkan, algortima dikembangkan dengan prinsip *adaptive base learner* dan diujikan langsung menggunakan data perbankan sehingga dapat merepresentasikan kondisi nyata.

Referensi

Asniar, & Surendro, K. (2014). Using data science for detecting outliers with k Nearest Neighbors graph. *Proceedings - 2014 International Conference on ICT for Smart Society: "Smart System Platform Development for City and Society, GoeSmart 2014", ICISS 2014*, (i), 300–304. https://doi.org/10.1109/ICTSS.2014.7013191

Besenbruch, J. (2018). *Fraud Detection Using Machine Learning*. Retrieved from https://beta.vu.nl/nl/Images/werkstukbesenbruch_tcm235-910176.pdf%0Ahttp://www.yannispappas.com/Fraud-Detection-Using-Machine-Learning/

Gómez, J. A., Arévalo, J., Paredes, R., & Nin, J. (2018). End-to-end neural network architecture for fraud scoring in card payments. *Pattern Recognition Letters*, *105*, 175–181. https://doi.org/10.1016/j.patrec.2017.08.024

Hoens, T. R., Polikar, R., & Chawla, N. V. (2012). Learning from streaming data with concept drift and imbalance: An overview. *Progress in Artificial Intelligence*, *1*(1), 89–101. https://doi.org/10.1007/s13748-011-0008-0

Kumar, M. S., Soundarya, V., Kavitha, S., Keerthika, E. S., & Aswini, E. (2019b). Credit Card Fraud Detection Using Random Forest Algorithm. *2019 Proceedings of the 3rd International Conference on Computing and Communications Technologies, ICCCT 2019*, 149–153. https://doi.org/10.1109/ICCCT2.2019.8824930

Lucaroni, F., Cicciarella Modica, D., MacIno, M., Palombi, L., Abbondanzieri, A., Agosti, G., … Vinci, A. (2019). Tree-Cart. *BMJ Open*, *9*(12), 1–6. https://doi.org/10.1136/bmjopen-2019-030234

Mahmoudi, N., & Duman, E. (2015). Detecting credit card fraud by Modified Fisher Discriminant Analysis. *Expert Systems with Applications*, *42*(5), 2510–

2516. https://doi.org/10.1016/j.eswa.2014.10.037

Malini, N., & Pushpa, M. (2017). Analysis on credit card fraud identification techniques based on KNN and outlier detection. *Proceedings of the 3rd IEEE International Conference on Advances in Electrical and Electronics, Information, Communication and Bio-Informatics, AEEICB 2017*. https://doi.org/10.1109/AEEICB.2017.7972424

Puh, M., & Brkić, L. (2019). Detecting credit card fraud using selected machine learning algorithms. *2019 42nd International Convention on Information and Communication Technology, Electronics and Microelectronics, MIPRO 2019 - Proceedings*, 1250–1255. https://doi.org/10.23919/MIPRO.2019.8757212

ROSA, T., PRIMARTHA, R., & WIJAYA, A. (2020). *Comparison of Distance Measurement Methods on K-Nearest Neighbor Algorithm For Classification*. *172*(Siconian 2019), 358–361. https://doi.org/10.2991/aisr.k.200424.054

Sahin, Y., & Duman, E. (2011a). *Detecting Credit Card Fraud by Decision Trees and Support Vector Machine*. *I*.

Sahin, Y., & Duman, E. (2011b). *IMECS2011_pp442-447*. *I*.

Zareapoor, M., & Shamsolmoali, P. (2015). Application of credit card fraud detection: Based on bagging ensemble classifier. *Procedia Computer Science*. https://doi.org/10.1016/j.procs.2015.04.201

❖❖❖