SKRIPSI

KAJIAN PENGGUNAAN SMOTE-NC PADA MODEL PEMBELAJARAN MESIN UNTUK KLASIFIKASI TRANSAKSI PENIPUAN

Jennifer Lorenza

NPM: 2017710010

PROGRAM STUDI MATEMATIKA
FAKULTAS TEKNOLOGI INFORMASI DAN SAINS
UNIVERSITAS KATOLIK PARAHYANGAN
2022

LEMBAR PENGESAHAN

KAJIAN PENGGUNAAN SMOTE-NC PADA MODEL PEMBELAJARAN MESIN UNTUK KLASIFIKASI TRANSAKSI PENIPUAN

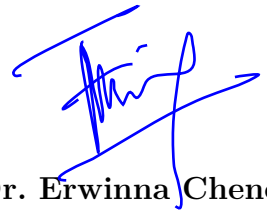
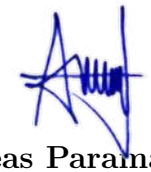
Jennifer Lorenza

NPM: 2017710010

Bandung, 10 Januari 2022

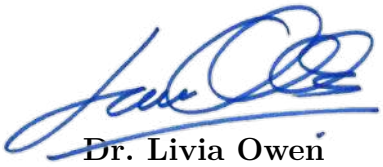
Menyetujui,

Pembimbing 1	Pembimbing 2
	
Agus Sukmana, M.Sc.	**Liem Chin, M.Si.**

Ketua Tim Penguji	Anggota Tim Penguji
	
Dr. Erwinna Chendra	**Dr. Andreas Parama Wijaya**

Mengetahui,

Ketua Program Studi



Dr. Livia Owen

PERNYATAAN

Dengan ini saya yang bertandatangan di bawah ini menyatakan bahwa skripsi dengan judul:

KAJIAN PENGGUNAAN SMOTE-NC PADA MODEL PEMBELAJARAN MESIN UNTUK KLASIFIKASI TRANSAKSI PENIPUAN

adalah benar-benar karya saya sendiri, dan saya tidak melakukan penjiplakan atau pengutipan dengan cara-cara yang tidak sesuai dengan etika keilmuan yang berlaku dalam masyarakat keilmuan.

Atas pernyataan ini, saya siap menanggung segala risiko dan sanksi yang dijatuhkan kepada saya, apabila di kemudian hari ditemukan adanya pelanggaran terhadap etika keilmuan dalam karya saya, atau jika ada tuntutan formal atau non-formal dari pihak lain berkaitan dengan keaslian karya saya ini.

Dinyatakan di Bandung,
Tanggal 10 Januari 2022

Jennifer Lorenza
NPM: 2017710010

ABSTRAK

Model-model pembelajaran mesin tersupervisi umumnya diasumsikan untuk dilatih dengan dataset seimbang, padahal banyak permasalahan klasifikasi di dunia nyata berasal dari dataset timpang, seperti contohnya deteksi transaksi penipuan (*fraud*). Dataset timpang dapat menyebabkan model klasifikasi biner menjadi kurang sensitif terhadap kategori minoritas (*fraud*) – yang tentunya tidak diharapkan terjadi pada detektor *fraud*. Salah satu cara untuk mengatasi masalah tersebut adalah dengan menggunakan teknik pengambilan ulang sampel pada kategori minoritas yang disebut SMOTE-NC (*Synthetic Minority Oversampling Technique for Nominal and Continuous*) – khususnya untuk data tipe numerik dan kategorial. Eksperimen dilakukan untuk menguji efek penggunaan SMOTE-NC dalam meningkatkan performa dan efisiensi data *training* pada empat jenis model pembelajaran mesin – Regresi Logistik dan SVC (*Support Vector Classifier*) Linier sebagai model linier, serta Pohon Keputusan dan *Random Forest* sebagai model non-linier. Secara umum SMOTE-NC meningkatkan performa model dengan *trade-off* antara *precision* dan *recall*, sehingga model mampu mendeteksi lebih banyak transaksi *fraud* yang sesungguhnya, tetapi juga lebih banyak salah memprediksi transaksi *non-fraud* sebagai *fraud*.

Kata-kata kunci: SMOTE-NC, dataset timpang, deteksi penipuan, pembelajaran mesin tersupervisi, klasifikasi biner

ABSTRACT

Supervised machine learning models is usually assumed to be trained with balanced dataset, whereas lots of real-world classification problem – including fraud detection – are manifested as imbalanced datasets. Imbalanced dataset could cause model's insensitivity of the minority class (fraud) – which is certainly least expected from a fraud detector. One alternative to overcome those problem is by resampling the minority class samples and generate new ones from them, namely the *Synthetic Minority Oversampling Technique* for nominal and continous data (SMOTE-NC). Experiment is done to examine the effect of SMOTE-NC to increase performance and training-set efficiency in four machine learning model types – Logistic Regression and Linear Support Vector Classifier as two linear models, along with Decision Tree and Random Forest as two non-linear models. In general, SMOTE-NC does increase model's performance, that is by trading-off precision score with recall score. Therefore, the model could detect more actual frauds accurately, at the expense of misjudging actual non-frauds as frauds.

Keywords: SMOTE-NC, imbalanced dataset, fraud detection, supervised machine learning, binary classification

KATA PENGANTAR

Puji syukur kepada Tuhan yang (penulis asumsikan) termanifestasi melalui orang-orang baik dan berbagai peristiwa keberuntungan, sedemikian sehingga mendukung penulis untuk dapat menyelesaikan skripsi ini. Skripsi yang berjudul "Kajian Penggunaan SMOTE-NC pada Model Pembelajaran Mesin untuk Klasifikasi Transaksi Penipuan" ini disusun sebagai salah satu syarat wajib untuk menyelesaikan studi Strata-1 Program Studi Matematika, Fakultas Teknologi Informasi dan Sains, Universitas Katolik Parahyangan, Bandung. Proses penyusunan skripsi ini adalah bagian dari tahap pendewasaan diri penulis, dan sepertinya tidak mungkin penulis dapat melaluinya tanpa dukungan dari berbagai pihak. Oleh karena itu, penulis menyampaikan terima kasih berlimpah kepada:

- Bapak Agus Sukmana, M.Sc. sebagai dosen pembimbing utama dan Bapak Liem Chin, M.Si. sebagai dosen pembimbing serta yang telah memberikan kebebasan dan kepercayaan penuh sehingga penulis berkesempatan untuk belajar menjadi pribadi mandiri.
- Ibu Dr. Erwinna Chendra selaku dosen penguji pertama atas waktu dan masukannya untuk skripsi ini, juga selaku dosen wali yang menasihati penulis untuk berani menentukan pilihan sendiri, bukan sekadar ikut arus.
- Bapak Dr. Andreas Parama Wijaya selaku dosen penguji kedua atas waktu dan masukannya untuk skripsi ini, juga selaku dosen yang memberikan kesempatan berharga bagi penulis untuk magang menjadi asisten peneliti.
- Bapak Janto Vincent Sulungbudi, Drs. selaku dosen pengajar MK-MK terkait *big data* dan *machine learning*, yang telah menerima penulis sebagai muridnya selama 4 semester berturut-turut dan selalu berusaha menyelenggarakan kelas yang *fun*. Sosok beliau telah menginsipirasi penulis untuk terus mengembangkan diri, berani mencoba dan belajar dari kesalahan, serta bermanfaat bagi orang lain.
- Seluruh dosen FTIS khususnya para dosen Program Studi Matematika UNPAR yang telah mendidik penulis.
- Para *YouTuber* dengan *channel* berisi video-video penjelasan sederhana yang mudah dimengerti namun tetap berbobot perihal *machine learning*, statistika, dan matematika; dua di antaranya yaitu Joshua Starmer (StatQuest) dan Grant Sanderson (3Blue1Brown). Berkat jasa para *YouTuber* tersebut, banyak orang (termasuk penulis) bisa merdeka belajar.
- Ms. Nia, Ms. Ayu, Mrs. Maryati, dan semua tenaga pengajar di Kumon EE dan Matematika Taman Surya yang selalu sabar membimbing dan memaklumi kondisi penulis saat dulu sulit bicara karena terlalu pemalu. Berkat jasa mereka, penulis memiliki modal dan fondasi untuk seterusnya mandiri belajar.
- Mami dan Papa yang telah membiayai pendidikan, menasihati dengan lembut namun tegas, dan mendukung keputusan penulis, juga Koko yang (meski kadang galak) selalu berusaha menjadi teladan, membagikan cerita tentang dunia kerja, dan meyakinkan penulis bahwa adiknya ini mampu menjadi mandiri.
- Nicho, Ii Yanny (alm.) dan Icang, 3 sepupu krucil (Carlyn, Jordan, Jolene), Ii Mina, Ii Bing, dan seluruh anggota Tio's clan yang telah memberikan dukungan moril.
- Gezia dan Ata, dua sahabat yang selalu berusaha memahami dan menerima penulis selama awal masa perkuliahan hingga entah kapan. Dari mereka, penulis belajar tentang sosok yang kreatif, humoris, percaya diri, tulus, peduli, berjiwa pemimpin, dan selalu berusaha

memberikan yang terbaik dari diri.
- Semua orang baik yang membantu menyalakan (kembali) lilin-lilin harapan bagi penulis dan tidak dapat disebutkan satu per satu namanya.

Penulis menyadari bahwa skripsi ini masih jauh dari sempurna, sehingga segala kritik dan saran yang konstruktif dari pembaca akan penulis terima dengan tangan terbuka. Meski awalnya penulis hanya berharap agar karya ini sungguh mencerminkan hasil belajar selama 4.5 tahun, pada akhirnya penulis juga berharap agar karya ini dapat bermanfaat bagi para pembaca.

Bandung, Januari 2022

Penulis

DAFTAR ISI

DAFTAR GAMBAR

DAFTAR TABEL

BAB 1

PENDAHULUAN

1.1 Latar Belakang

Metode pembayaran non-tunai dengan kartu (kartu kredit dan kartu debet/ATM) semakin menjadi andalan bagi masyarakat Indonesia dalam bertransaksi. Berdasarkan statistik Bank Indonesia, jumlah alat pembayaran menggunakan kartu (APMK) mengalami tren peningkatan, per tahun 2020 telah mencapai 230,5 juta – sudah mendekati jumlah penduduk Indonesia di tahun yang sama sebesar 270,2 juta jiwa. Walaupun volume transaksi di Indonesia masih didominasi metode pembayaran tunai, selisih persentasenya dengan metode pembayaran non-tunai terus menipis, yaitu dari tahun 2015 sebesar 19,7% hingga tahun 2020 sebesar 15,5%. Namun, kemajuan teknologi ini tidak bisa terlepas dari risiko keamanan dalam penggunaannya. Modus pencurian uang tunai berevolusi menjadi penipuan (*fraud*), yaitu pencurian atau pemalsuan data pelanggan (pemilik kartu) untuk mengakses dan menggunakan rekening pelanggan tersebut, demi keuntungan sang penipu sendiri. Tindakan kriminal yang sering juga disebut *carding* ini merugikan terutama keuangan pihak pelanggan, reputasi pihak bank, dan akhirnya dapat berdampak buruk bagi ekonomi negara bila tidak dilakukan upaya pencegahan untuk menekan jumlah kasusnya.

Sebagaimana sekarang adalah era maha data dan pembelajaran mesin, salah satu upaya pencegahan kejahatan *carding* yang dapat dilakukan adalah dengan merancang algoritma untuk mendeteksi transaksi *fraud*. Meskipun bukan satu-satunya faktor penentu, kualitas dan kuantitas data *training* adalah faktor penting yang mempengaruhi performa algoritma pembelajaran mesin. Seringkali jumlah kejadian transaksi *fraud* jauh lebih sedikit dibandingkan *non-fraud*. Dari perspektif pihak pelanggan dan pihak bank tentu ini adalah hal yang patut disyukuri, tetapi hal tersebut menyebabkan timpangnya distribusi data (*imbalanced dataset*) dalam konteks klasifikasi biner, yaitu adanya kelas positif sebagai minoritas (*fraud*) dan kelas negatif sebagai mayoritas (*non-fraud*). Umumnya, algoritma pembelajaran mesin klasifikasi biner yang dilatih dengan *imbalanced dataset* cenderung berperforma lebih buruk dalam mengklasifikasi kelas minoritas daripada yang dilatih dengan *balanced dataset* [1]. Semakin timpang rasio kelas minoritas terhadap kelas mayoritas, semakin besar rasio *error* prediksi kelas minoritas terhadap prediksi kelas mayoritas.

Salah satu cara untuk mengatasi masalah dataset timpang adalah teknik pengambilan ulang sampel data, yaitu *Synthetic Minority Oversampling Technique* (SMOTE) [2], yang bertujuan mengurangi ketimpangan rasio kelas minoritas terhadap kelas mayoritas, dengan cara membangkitkan sampel baru (sintetis) sehingga memperbanyak sampel dari kelas minoritas. Varian dari SMOTE yang dirancang khusus untuk *oversampling* data dengan tipe fitur numerik (kontinu) dan kategorik (nominal) adalah SMOTE-*Nominal Continuous* (SMOTE-NC) [2], dengan contoh penerapannya seperti yang dilakukan dalam penelitian oleh Sifa dkk. [3], penelitian oleh Akyon dan Kalfaoglu [4], serta penelitian oleh Gök dan Olgun [5]. Namun, ketiga penelitian tersebut tidak berfokus pada detail pembahasan cara kerja SMOTE-NC. Selain itu, sejauh ini belum ada penelitian terkait efek penggunaan SMOTE-NC dalam memampukan model pembelajaran mesin untuk belajar tanpa memerlukan banyak data *training*.

Oleh karena itu, selain membangun *pipeline* pembelajaran mesin untuk klasifikasi transaksi *fraud* dan menggunakan *oversampler* SMOTE-NC sebagai upaya mengatasi masalah dataset timpang,

dalam skripsi ini juga akan dirancang eksperimen untuk meneliti efek SMOTE-NC terhadap performa dan efisiensi penggunaan data *training* dari model-model pembelajaran mesin. Cara kerja SMOTE-NC juga akan dikaji lebih lanjut, dengan mengamati efek penggunaan SMOTE-NC terhadap fungsi keputusan model. Diharapkan temuan penelitian ini dapat menjadi bahan pertimbangan dalam mengatasi masalah dataset timpang – terutama dataset dengan fitur bertipe numerik dan kategorik yang jumlah sampelnya terbatas.

1.2 Rumusan Masalah

Masalah-masalah yang sudah diidentifikasi di bagian 1.1 dan akan diselesaikan dalam skripsi ini adalah:

1. Bagaimana membangun model pembelajaran mesin tersupervisi untuk mendeteksi transaksi *fraud*?
2. Bagaimana pengaruh penggunaan SMOTE-NC terhadap peningkatan performa dari model pembelajaran mesin dalam mendeteksi transaksi *fraud* pada kasus dataset timpang?

1.3 Tujuan

Tujuan yang hendak dicapai dalam skripsi ini antara lain adalah:

1. Membangun *pipeline* pembelajaran mesin dan menggunakan model-model linier (Regresi Logistik dan SVC Linier) serta non-linier (Pohon Keputusan dan *Random Forest*) untuk menyelesaikan masalah pembelajaran tersupervisi.
2. Melakukan eksperimen terhadap *pipeline* pembelajaran mesin – yang performanya diukur dari skor F_1 – dengan faktor-faktor yang diuji adalah (1) tipe model pembelajaran mesin, (2) penggunaan *oversampler* SMOTE-NC, dan (3) fraksi data *training*.
3. Menggunakan metode ANOVA untuk menyelidiki keberadaan efek interaksi antar-faktor dan efek utama dari masing-masing faktor.
4. Mengkaji cara kerja SMOTE-NC dengan mengamati efek penggunaannya terhadap skor *precision*, *recall*, dan *accuracy* dari model pembelajaran mesin, serta terhadap fungsi keputusan dari model.

1.4 Batasan Masalah

Kesulitan dalam memperoleh data faktual lengkap dari transaksi dengan kartu, dikarenakan adanya prinsip kerahasiaan data perbankan. Oleh karena itu, dalam penelitian ini digunakan data sintetis (yang tersedia di Kaggle) hasil simulasi dari simulator multi-agen BankSim [6] yang meniru pola perilaku para nasabah nyata dari suatu bank di Spanyol dalam bertransaksi sehari-hari.

1.5 Metodologi

Langkah-langkah penelitian dalam skripsi ini dapat dibagi menjadi 4 bagian, yaitu:

1. Persiapan Data
   Data yang digunakan untuk *training* dan *testing* model pembelajaran mesin adalah data sintetis berlabel hasil simulasi dari simulator multi-agen BankSim [6] dengan fitur-fitur numerik dan kategorik. Pada bagian ini, dilakukan eksplorasi, pembersihan, serta penambahan fitur baru berdasarkan fitur waktu dan nomor ID.
2. *Pipeline* Pembelajaran Mesin
   Pipeline dirancang untuk mengotomatisasi alur kerja yang diperlukan untuk menghasilkan

model pembelajaran mesin, mencakup antara lain standardisasi untuk fitur numerik, *oversampling* (untuk model yang dikenakan SMOTE-NC), dan *one-hot encoding* untuk fitur kategorik. Selain itu, dilakukan juga penentuan *hyperparameter* dan metrik evaluasi model.

3. Eksperimen
   Pada bagian ini, dilakukan perancangan eksperimen dengan menentukan variabel bebas, variabel terikat, serta nilai respons yang diukur dari objek eksperimen. Kemudian dilakukan proses K-*fold cross validation* yang dimodifikasi demi kepentingan pengujian hipotesis dengan ANOVA.
4. Evaluasi Hasil Eksperimen
   Kajian lebih lanjut tentang efek dan cara kerja SMOTE-NC dalam meningkatkan performa model dan efisiensi data *training*.

1.6 Sistematika Pembahasan

Bab 1 berisi pendahuluan yang mencakup latar belakang, rumusan masalah, tujuan penelitian, batasan masalah, dan metodologi penelitian. Bab 2 akan diisi landasan teori dari model-model pembelajaran mesin yang digunakan, masalah klasifikasi dengan *imbalanced dataset*, algoritma *oversampler* SMOTE-NC, dan metode ANOVA. Bab 3 akan diisi rincian metodologi, sementara bab 4 akan diisi pembahasan hasil eksperimen. Bab 5 akan diisi dengan kesimpulan dan saran untuk penelitian selanjutnya.