

Penerapan *Modified ADASYN* untuk Meningkatkan Akurasi Pendeteksian Pola *Fraud* pada Transaksi Kartu Kredit

Ebhen Haezer Sitohang, Djoni Haryadi Setiabudi, Stephanus A. Ananda

Program Studi Informatika Fakultas Teknologi Industri Universitas Kristen Petra

Jl. Siwalankerto 121 – 131 Surabaya 60236

Telp. (031) – 2983455, Fax. (031) – 8417658

E-mail: ebhensitohang583@gmail.com, djonihs@petra.ac.id, ananda@petra.ac.id

ABSTRAK

Salah satu faktor yang paling berpengaruh dalam akurasi adalah ketidakseimbangan *class*, hal ini diulas dalam penelitian yang dilakukan oleh Gameng [4]. Pada penelitian Bagga [1], metode *Pipelining* dikombinasikan dengan *ADASYN* akurasi bisa mencapai 0.99999. Masalah yang ada pada penelitian tersebut adalah akurasi belum tentu bisa mencapai 0.99999 jika menggunakan *dataset* di luar *dataset* yang mereka gunakan dan jika menggunakan algoritma klasifikasi selain *Pipelining*. Pada penelitian yang dilakukan Dornadula & Geetha [2], akurasi paling tinggi hanya 0.9994. Pada penelitian yang dilakukan Makki [8], model klasifikasi yang menggunakan metode penyeimbang *class* justru akurasi menjadi lebih rendah.

Pada penelitian ini, *Modified ADASYN* digunakan karena pada penelitian Gameng [4] akurasi, presisi dan *f1-score*nya berhasil melampaui *ADASYN* dan *SMOTE*. Metode *Pipelining* digunakan karena pada penelitian Bagga [1], *Pipelining* dapat membuat akurasi klasifikasi hingga mencapai 0.99999.

Sebagai hasil dari pengujian, maka penelitian ini mendapatkan kesimpulan bahwa *Modified ADASYN* belum mampu mendapatkan akurasi 0.999999 pada dua *dataset* berbeda. Pada penelitian ini *Modified ADASYN* mampu meningkatkan akurasi *K-NN* hingga 0.9995148 dan 0.97617554 menggunakan *dataset* pertama dan kedua. *Modified ADASYN* dapat melebihi performa *SMOTE*, *ADASYN*, *One-Class Classification* dan *Cost Sensitive*. Pada penelitian ini didapatkan bahwa nilai *K* yang optimal pada *Modified ADASYN* bisa berbeda-beda tergantung banyak parameter dan sampel datanya.

Kata Kunci: *credit-card fraud; modified ADASYN; pipelining; imbalanced classification; class imbalance.*

ABSTRACT

One of the most influential factors in accuracy is class imbalance, this is reviewed in a study conducted by Gameng [4]. In the study of Bagga [1], the Pipelining method combined with ADASYN the accuracy can reach 0.99999. The problem in this study is that accuracy may not necessarily reach 0.99999 if using a dataset outside the dataset they are using and if using a classification algorithm other than pipelining. In a study conducted by Dornadula & Geetha [2], the highest accuracy was only 0.9994. In the research conducted by Makki [8], the classification model that uses the class balancing method has lower accuracy.

In this research, Modified ADASYN is used because in the research of Gameng [4] its accuracy, precision and f1-score surpassed ADASYN and SMOTE. Pipelining method is used

because in the study of Bagga [1], Pipelining can make classification accuracy up to 0.99999.

As a result of testing, this research concludes that Modified ADASYN has not been able to obtain an accuracy of 0.999999 on two different datasets. In this research, Modified ADASYN is able to increase the accuracy of K-NN to 0.9995148 and 0.97617554 using the first and second datasets. Modified ADASYN can outperform SMOTE, ADASYN, One-Class Classification and Cost Sensitive. In this research, it is found that the optimal K value in Modified ADASYN can vary depending on many parameters and sample data.

Keywords: *credit-card fraud; modified ADASYN; pipelining; imbalanced classification; class imbalance.*

1. PENDAHULUAN

Saat ini pengguna kartu kredit sudah semakin banyak ditemukan, namun disisi lain kasus penipuan juga semakin banyak terjadi [2]. Menurut Cifa, 33.305 kasus penipuan kartu dilaporkan antara Januari dan Juni 2018 [8]. Hal ini menyebabkan kerugian yang cukup besar baik pada lembaga keuangan maupun nasabah. Oleh karena itu, kegiatan pendeteksian dan pencegahan penipuan yang efektif sangatlah penting untuk mengamankan sistem keuangan suatu institusi [8].

Pada *paper* Bagga [1], dijelaskan bahwa akurasi merupakan hal yang penting dalam pendeteksian pola *fraud* pada transaksi kartu kredit. Ini dikarenakan semakin baik akurasi pendeteksian maka kerugian yang ditimbulkan oleh transaksi *fraud* akan semakin kecil. Pada *paper* Gameng [4], dijelaskan bahwa salah satu faktor yang paling berpengaruh dalam akurasi adalah ketidakseimbangan *class*. Ketidakseimbangan *class* terjadi ketika setidaknya salah satu *class* pada data, jumlahnya sangat sedikit jika dibandingkan dengan *class* yang lain. Dalam penelitian ini *class* yang mendominasi adalah *class* negatif (*non-fraud*) sedangkan *class* positif (*fraud*) jumlahnya sangat sedikit.

Pada *paper* Bagga [1] dijelaskan bahwa ketidakseimbangan ini dapat mengurangi kemampuan algoritma klasifikasi untuk mengenali kasus positif atau negatif yang menyebabkan kesalahan klasifikasi sampel positif sebagai kelas negatif atau bisa disebut negatif palsu. Pendapat yang sama juga tertulis pada *paper* Johnson & Khoshgoftaar [6], Ketika ketidakseimbangan *class* ada dalam data *training*, model biasanya akan lebih cenderung berpihak ke data yang termasuk dalam *class* mayoritas. Akibatnya, data yang termasuk ke dalam *class* minoritas lebih sering salah diklasifikasikan daripada data yang termasuk dalam *class* mayoritas.

Seperti yang tertulis pada *paper* Bagga [1], *Machine Learning* yang dikombinasikan dengan solusi ketidakseimbangan *class* bisa membuat pendeteksian transaksi yang tidak sah menjadi lebih efektif. Pada penelitian ini, metode *Pipelining* dikombinasikan dengan *ADASYN* yang dimana akurasi bisa mencapai 0.99999. Masalah yang ada pada penelitian tersebut adalah akurasi belum tentu bisa mencapai 0.99999 jika menggunakan *dataset* di luar *dataset* yang mereka gunakan dan jika menggunakan algoritma klasifikasi selain *Pipelining*. Selain itu, presisi pada *class* 1(*Fraud*) pada penelitian Bagga [1] hanya sampai 0.85. Pada penelitian yang dilakukan Dornadula & Geetha [2], *SMOTE* digunakan untuk mengatasi ketidakseimbangan data, dan didapatkan bahwa *SMOTE* dapat meningkatkan akurasi *Random forest* hingga 0.9994. Namun pada penelitian tersebut akurasi nya paling tinggi hanya sampai 0.9994. Pada penelitian yang dilakukan Makki [8], *Random Oversampling*, *One-Class Classification* dan *Cost Sensitive* digunakan untuk mengatasi ketidakseimbangan data. Pada penelitian tersebut algoritma *C5.0*, *SVM* dan *ANN* adalah metode terbaik dilihat dari 3 kriteria. Dimana kriterianya adalah akurasi, sensitivitas dan *AUPRC*. Namun pada penelitian ini metode pendekatan yang dirancang khusus untuk mengatasi masalah ketidakseimbangan *class* tidak cukup efektif. Ini dikarenakan metode *C5.0*, *SVM* dan *ANN* yang menggunakan *Random Oversampling*, *One-Class Classification* dan *Cost Sensitive* justru akurasi nya menjadi lebih rendah.

Secara garis besar penelitian ini akan menggunakan *Modified ADASYN* untuk menyeimbangkan data. Data sintesis akan dibuat untuk meningkatkan jumlah data minoritas kemudian akan dilakukan klasifikasi menggunakan *Pipelining*. Pada pengujian pertama, performa metode *Modified ADASYN* dan *Pipelining* menggunakan *dataset* pertama akan dibandingkan dengan metode *Modified ADASYN* dan *Pipelining* menggunakan *dataset* kedua. Pada pengujian kedua, performa metode *Modified ADASYN* dan *K-NN* menggunakan *dataset* pertama akan dibandingkan dengan performa metode *Modified ADASYN* dan *K-NN* menggunakan *dataset* kedua. Pada pengujian ketiga, performa metode *Modified ADASYN* dan *Pipelining* menggunakan *dataset* pertama akan dibandingkan dengan performa *SMOTE* dan *Random Forest*, *ADASYN* dan *Pipelining* pada penelitian Dornadula & Geetha [2] dan Bagga [1] Pada pengujian keempat, akurasi metode *Modified ADASYN* dan *Pipelining* akan dibandingkan dengan *Pipelining* tanpa metode penyeimbang *class*. Pada penelitian ini *Modified ADASYN* digunakan untuk menyeimbangkan data dikarenakan pada penelitian Gameng [4] akurasi, presisi dan *f1-score* nya berhasil melampaui *ADASYN* dan *SMOTE*. Metode *Pipelining* digunakan untuk klasifikasi dikarenakan pada penelitian Bagga [1] metode ini dapat membuat akurasi klasifikasi hingga mencapai 0.99999. *Pipelining* yang digunakan Pada penelitian ini menggunakan *SelectKBest* untuk pemilihan fitur dan menggunakan *f-regression* untuk perhitungan pemilihan fiturnya. *K* merupakan jumlah fitur data yang akan dipakai dalam *Pipelining* dan Pada penelitian ini nilai *K=28*. Metode *Modified ADASYN* dan *Pipelining* digunakan untuk memperbaiki kelemahan yang ada pada 3 penelitian sebelumnya. *Modified ADASYN* akan menggunakan jarak *Manhattan* untuk menggantikan jarak *Euclidean* pada proses *K-NN* yang digunakan [4].

1.1 Rumusan Masalah

Dikutip dari *paper* Bagga [1], Akurasi adalah ukuran jumlah prediksi yang benar dibagi dengan jumlah total prediksi. Akurasi dicari untuk mendapat jawaban dari pertanyaan “Berapa persen transaksi yang benar diprediksi *fraud* dan *tidak fraud* dari

keseluruhan transaksi. Akurasi merupakan aspek paling penting dalam pendeteksian pola transaksi *fraud* pada kartu kredit. Ini karena semakin tinggi akurasi maka semakin kecil kemungkinan untuk salah prediksi. Berdasarkan latar belakang yang telah diuraikan diatas, maka rumusan masalahnya adalah:

1. Apakah akurasi *Modified ADASYN* untuk menyeimbangkan data dan *Pipelining* untuk melakukan pendeteksian tetap bisa 0.99999 jika dilakukan pada dua *dataset* yang berbeda?
2. Apakah akurasi *Modified ADASYN* untuk menyeimbangkan data dan *K-NN* untuk melakukan pendeteksian bisa lebih dari 0.9494 jika dilakukan pada dua *dataset* yang berbeda?
3. Apakah akurasi klasifikasi menggunakan *Modified ADASYN* dan *Pipelining* bisa lebih tinggi daripada menggunakan *SMOTE* dan *Random Forest*, *ADASYN* dan *Pipelining* pada penelitian sebelumnya, yaitu sebesar 0.9998 dan 0.99999?
4. Apakah penerapan *Modified ADASYN* pada *Pipelining* bisa membuat akurasi lebih besar daripada *Pipelining* yang tidak menggunakan metode penyeimbang *class*?
5. Seberapa tinggi presisi *Modified ADASYN* pada *K-NN* dan *Pipelining*?

2. DASAR TEORI

2.1 K-Nearest Neighbor

Algoritma *k-nearest neighbor* adalah sebuah metode untuk melakukan klasifikasi terhadap objek berdasarkan data pembelajaran yang jaraknya paling dekat dengan objek tersebut. Data pembelajaran diproyeksikan ke ruang berdimensi yang banyaknya sesuai dengan jumlah fitur data, dimana masing-masing dimensi merepresentasikan fitur dari data. [4].

2.2 Pipelining

Pipelining mengacu pada penerapan rangkaian transformasi yang diikuti oleh pengklasifikasi akhir. *Pipelining* digunakan untuk merakit beberapa proses ganda untuk tujuan memvalidasi silang mereka bersama-sama, melakukan pengaturan secara bersamaan hingga parameter yang berbeda [11]. Dalam penelitian ini, *selectKBest* akan digunakan yang memungkinkan pemilihan fitur berdasarkan skor tertinggi *K*. *K* merupakan jumlah fitur data yang akan dipakai dalam *Pipelining* [10]. *F-regresi* akan digunakan untuk tujuan menjalankan fitur pilihan. Ini didasarkan pada uji linier (regresi) yang *uni-variate*. Ini dilakukan untuk tujuan menemukan pengaruh dari setiap *regressor*. Ini sebenarnya adalah fungsi penilaian. Pada tahap akhir, *Random Forest Classifier* akan digunakan untuk tujuan klasifikasi dan prediksi [1].

2.3 Binary Classification

Klasifikasi biner adalah klasifikasi yang terdiri dari satu grup *class*. Klasifikasi ini terdiri dari grup positif dan satu grup negatif. Klasifikasi ini biasanya digunakan untuk diagnosis penyakit(kanker, tidak kanker), Pendeteksian transaksi penipuan(*fraud*, *non fraud*). Di dalam klasifikasi biner sering kali ditemukan masalah *class imbalance* ketika jumlah suatu *class*(negatif) lebih banyak dari jumlah *class* yang lain(positif) [6].

2.4 Random Forest

Random Forest adalah salah satu algoritma *Supervised Learning*. Ini digunakan baik untuk klasifikasi dan regresi. Algoritma

Random Forest memilih sampel acak dari *dataset* yang diberikan. Kemudian membangun *Decision Tree* untuk setiap sampel dan kemudian mendapatkan prediksi dari masing-masing *Decision Tree*. Pemungutan suara dilakukan untuk setiap hasil prediksi dan keputusan dengan suara terbanyak dipilih sebagai prediksi akhir [1].

2.5 Data Mining

Secara analogi, data mining seharusnya lebih tepat dinamai "penambangan pengetahuan dari data", yang terdengar masih terlalu panjang. Namun, secara singkat, penambangan pengetahuan mungkin tidak mencerminkan penekanan pada menambang dari data dalam jumlah besar. Namun demikian, kata "penambangan" sudah mencerminkan sangat jelas mengenai proses yang menemukan sekumpulan kecil bongkahan berharga dari banyak bahan mentah. Banyak orang mengenal *data mining* sebagai sinonim untuk istilah lain dari, penemuan *pattern/knowledge* dari sebuah kumpulan data, yang juga biasanya disingkat *KDD (Knowledge Discovery)* [5].

2.6 Kartu Kredit

Kartu kredit umumnya mengacu pada kartu yang diberikan kepada pelanggan (*cardholders*), biasanya memungkinkan mereka untuk melakukannya membeli barang dan jasa dalam batas kredit atau menarik uang tunai di muka. Kartu kredit memberikan *cardholders* sebuah memanfaatkan waktu, yaitu, memberikan waktu bagi pelanggan mereka untuk membayar nanti dalam waktu yang ditentukan, dengan membawanya ke siklus penagihan berikutnya [2].

2.7 Penipuan Kartu Kredit

Penipuan kartu kredit terjadi ketika pelaku menggunakan kartu kredit seseorang untuk keuntungan pribadi, bahkan bank penerbit tidak sadar bahwa kartu tersebut sedang digunakan. Selain itu, pelaku tidak memiliki hubungan dengan pemegang kartu atau penerbit, dan tidak berniat untuk memberi tahu pemilik kartu tentang kartu yang hilang dan melakukan pembayaran kembali atas transaksi yang dilakukan [7]. Penipuan kartu kredit merupakan tindak pidana. Menurut Cifa, 33.305 kasus *fraud* kartu dilaporkan antara Januari dan Juni di 2018 (dalam [8]). Penipuan (*fraud*) kartu kredit adalah sasaran empuk. Ini dikarenakan jumlah yang signifikan dapat ditarik tanpa sepengetahuan pemilik, dalam waktu singkat. Penipu selalu berusaha membuat setiap transaksi *fraud* menjadi seolah-olah transaksi *non-fraud*, yang membuat setiap transaksi *fraud* menjadi tugas yang sangat menantang dan sulit untuk dideteksi. Pada tahun 2017, ada 1.579 *data breaches* dan hampir 179 juta catatan diantaranya penipuan kartu kredit. 133.015 laporan dengan bentuk penipuan yang paling umum, kemudian dengan penipuan dengan modus *employment* atau *tax-related* dengan 82.051 laporan, penipuan dengan modus *fraud phones*. dengan 55.045 laporan diikuti oleh penipuan *bank* dengan 50.517 laporan dari statika yang dirilis oleh FTC (dalam [2]). *COVID-19* telah mempercepat tingkat pertumbuhan *e-commerce* sebesar 28% untuk tahun 2020, dengan penjualan mencapai lebih dari \$4 triliun. Dengan peningkatan penjualan yang cepat ini juga muncul peningkatan yang signifikan dalam penipuan kartu kredit. *Digital analyst house, Juniper Research*, memperkirakan bahwa kerugian penipuan pembayaran *online* akan meningkat 50% di Eropa pada tahun 2024, meskipun perusahaan di Eropa menerapkan biometrik dan metode lain dari *Strong Customer Authentication (SCA)*. Penipuan di situs *retailer web* pengecer Inggris saja menelan

biaya lebih dari £251 juta per tahun, membuktikan bahwa pencegahan tidak hanya bersifat preferensial tetapi juga kritis[3].

2.8 Ketidakseimbangan Class

Ketidakeimbangan *class* didefinisikan sebagai sebuah *dataset* yang ingin diklasifikasikan memiliki jumlah suatu *class* yang lebih banyak daripada jumlah *class* lainnya. Dimana ini bisa membuat performa pengklasifikasian menjadi kurang baik [8]. Sebagai contoh, misalnya jika ada 10 transaksi pada *class fraud* per 1.00.000 transaksi *class non-fraud*, maka meskipun modelnya memprediksi negatif untuk semua data, akurasiya masih 99,999%. Ini dikarenakan, model akan belajar memprediksi semua transaksi sebagai *non-fraud* meskipun sebenarnya ada transaksi yang *fraud* [1].

2.9 Modified ADASYN

Modified ADASYN adalah *ADASYN* yang dimodifikasi menggunakan jarak *Manhattan* untuk menggantikan jarak *Euclidean* pada proses *K-NN* yang digunakan [4].

1. Kalkulasi degree of class imbalance menggunakan:

$$d_c = \frac{m_s}{m_l} \quad (1)$$

Di mana $d \in [0, 1]$.

2. Jika $d_c < d_{th}$ (d_{th} adalah penetapan *threshold* untuk derajat toleransi maksimum dari rasio ketidakseimbangan kelas):

- a. Hitung jumlah *instance* data sintesis yang perlu digeneralisasi untuk kelas minoritas:

$$G_c = (m_l - m_s) \times \beta \quad (2)$$

Di mana $\beta \in [0, 1]$ adalah parameter yang digunakan untuk menetapkan level balance yang diinginkan. Dimana $\beta = 1$ berarti data yang sepenuhnya seimbang.

- b. Untuk setiap *instance* $x_i \in$ minoritas *class*, temukan *k-nearest neighbors* berdasarkan *Manhattan distance* pada n *dimensional space*, dan kalkulasi rasio r_i yang didefinisikan sebagai:

$$r_i = \frac{\#majority_i}{k}, \quad i = 1, \dots, m_s \quad (3)$$

Di mana $\#majority$ adalah jumlah *instance* pada *nearest neighbor* yang termasuk kelas y_c (mayoritas) atau termasuk semua kelas kecuali y_{kc} (minoritas), oleh karena itu $x_i \in [0, 1]$, y_{kc} adalah kelas yang dievaluasi

- c. Nilai r_i dinormalisasi, di mana total semua nilai r sama dengan 1:

$$\hat{r}_i = \frac{r_i}{\sum r_i} \quad (4)$$

- d. Hitung jumlah dari *instance* data sintesis yang perlu dihasilkan pada setiap *instance* minoritas x_i :

$$g_i = \hat{r}_i \times G_c \quad (5)$$

Dimana G_c adalah total jumlah dari *instance* data sintesis yang perlu untuk dihasilkan untuk kelas minoritas ke- yang dijelaskan pada rumus (2).

- e. Hasilkan data g_i untuk setiap lingkungan. Pertama, ambil contoh minoritas untuk lingkungannya, x_i Kemudian, pilih secara acak contoh minoritas lain dalam lingkungan tersebut yang akan menjadi xz_i , λ adalah bilangan acak antara 0–1. s_i adalah contoh sintetik baru, x_i dan xz_i adalah dua contoh minoritas dalam lingkungan yang sama.

$$s_i = x_i + (xz_i - x_i) \times \lambda \quad (6)$$

Dimana $(xz_i - x_i)$ adalah vektor perbedaan ruang-ruang berdimensi, dan λ adalah bilangan acak: $\lambda \in [0,1]$ [4].

2.10 Manhattan Distance

Jarak Manhattan juga disebut sebagai "city block distance" yang merupakan jumlah jarak dari semua parameter. Untuk dua titik data atau lebih dari x dan y dan seterusnya dalam dimensi ruang- d , jarak *Manhattan* antara titik-titik tersebut didefinisikan sebagai berikut [4]:

$$d = \sum_{k=1}^n |x_i - y_i| \quad (7)$$

2.11 Performance Evaluation(Classification Report)

Eksperimen dievaluasi menggunakan 4 metrik dasar - *True Negatif* (TN), *True Positif* (TP), *False Negatif* (FN) dan *False Positif* (FP). Kinerja 7 metode dibandingkan berdasarkan akurasi, presisi, *F1 score* dan *recall*.

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN} \quad (8)$$

$$\text{Precision} = \frac{TP}{TP + FP} \quad (9)$$

$$\text{Recall} = \frac{TP}{TP + FN} \quad (10)$$

$$\text{F1 Score} = \frac{2 * \text{Precision} * \text{Recall}}{\text{Precision} + \text{Recall}} \quad (11)$$

Akurasi adalah ukuran jumlah prediksi yang benar dibagi dengan jumlah total prediksi. Presisi adalah rasio prediksi positif dengan jumlah total kelas positif yang diprediksi. *Recall* adalah rasio prediksi positif dengan jumlah nilai kelas positif dalam data *testing*. Skor *F1* menggambarkan keseimbangan antara presisi dan *Recall* [1].

2.12 F-Regression

Model linier untuk menguji efek individu dari masing-masing banyak regresi. Ini adalah fungsi penilaian untuk digunakan dalam prosedur pemilihan fitur, bukan prosedur pemilihan fitur yang berdiri sendiri. Metode ini dilakukan dengan 2 langkah. Pertama, Korelasi antara setiap regressor dan target dihitung dengan rumus:

$$((X[:, i] - \text{mean}(X[:, i])) * (y - \text{mean}(y))) / (\text{std}(X[:, i]) * \text{std}(y)) \quad (12)$$

Kedua, Hasil perhitungan langkah pertama akan dikonversikan ke *f-Score* kemudian ke *p-value* [9].

3. DATASET

3.1 Anonymized credit card transactions labeled as fraudulent or genuine (Dataset Pertama)

Dataset ini didapat dari link:

<https://www.kaggle.com/mlg-ulb/creditcardfraud>

Pada *dataset* ini, didapatkan hasil analisa dan masalah yang ada. Banyak *row* dari *dataset* ini adalah 284.808, dimana *class* 1 menandakan transaksi *fraud* dan 0 menandakan transaksi *non fraud*. Masalah yang ada pada *dataset* ini adalah data pada *class* masih berbentuk *string* untuk itu data pada *class* perlu diubah menjadi *integer* terlebih dahulu. *Dataset* ini memiliki parameter *Time*, *V1-V28*, *Amount*, *Class*. *Time* merupakan penjelasan waktu dan *V1-V28* merupakan hasil *PCA* untuk menjaga identitas nasabah. *Class* menandakan parameter yang menyimpan data *fraud* atau *non-fraud*. Dimana penelitian ini akan menggunakan parameter *V1-V28* dan *Class*.

3.2 Abstract data set for Credit card fraud detection (Dataset Kedua)

Dataset ini didapat dari link:

<https://www.kaggle.com/shubhamjoshi2130of/abstract-data-set-for-credit-card-fraud-detection>

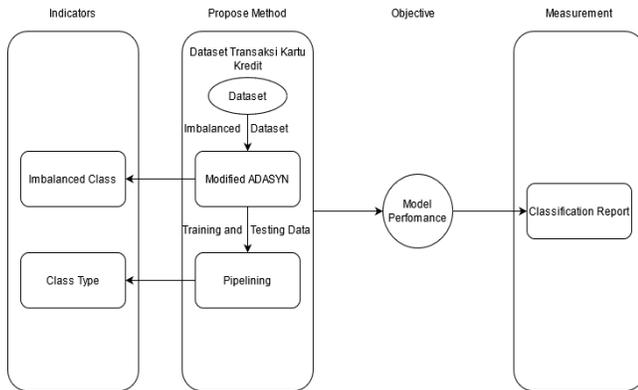
Pada *dataset* ini, didapatkan beberapa hasil analisa. Banyak *row* dari dataset ini adalah 3.076, dimana *isFraudulent Y* menandakan transaksi *fraud* dan *N* menandakan transaksi *non fraud*. Pada *dataset* ini, *Y* pada parameter *Risk Country* menandakan transaksi berasal dari *risk country* dan *N* menandakan transaksi tidak berasal dari *risk country*. Pada dataset ini, *Y* pada parameter *Foreign Transaction* menandakan transaksi berasal dari *foreign country* dan *N* menandakan transaksi tidak berasal dari *foreign country*. Pada parameter *Isdeclined*, *Y* menandakan transaksi ditolak dan *N* menandakan transaksi tidak ditolak. Selain itu permasalahan yang ada pada *dataset* ini adalah sebagian data masih dalam bentuk *char* sehingga harus diubah ke dalam *integer* terlebih dahulu contohnya nilai *Y* dan *N* pada beberapa parameter akan diganti menjadi 1 dan 0.

Dataset ini memiliki 12 parameter dan yang akan digunakan pada penelitian ini adalah sebagai berikut:

- *AverageAmount/transaction/day*
- *Transaction_amount* (Jumlah transaksi)
- *Isdeclined*
- *TotalNumberofdeclines/day* (Jumlah transaksi ditolak per hari)
- *isForeignTransaction*
- *isHighRiskCountry*
- *Daily_chargeback_avg_amt* (Rata-rata *chargeback* per hari)
- *6_month_avg_chbk_amt* (Rata-rata *chargeback* dalam 6 bulan)
- *6-month_chbk_freq* (Rata-rata *chargeback* frekuensi dalam 6 bulan)
- *isFraudulent* (Termasuk *fraud* atau tidak *fraud*).

4. DESAIN PENGUJIAN

4.1 Kerangka Pemikiran

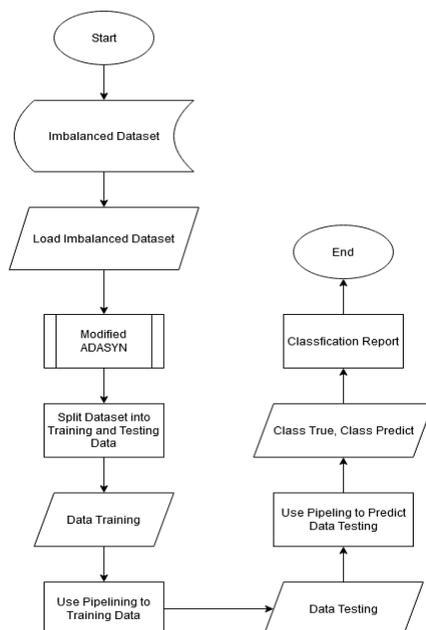


Gambar 1. Model Perbaikan Metode

Pada bagian ini akan dijelaskan tentang gambar 1. Pada penelitian ini, *dataset* yang digunakan adalah *dataset* transaksi kartu kredit. Sedangkan metode yang diusulkan adalah menggunakan metode *Pipelining*, di mana pada proses untuk menyeimbangkan datanya digunakan *Modified ADASYN*. Indikator yang berpengaruh pada *Pipelining* adalah *class type* yaitu non fraud dan fraud, dimana ini memiliki kemungkinan *class 0* dan *class 1*. Tujuan Pada penelitian ini adalah adanya peningkatan performa pada model (*Model Performance*), performa yang paling dilihat adalah akurasi dan presisi. Pengukuran peningkatan performanya akan menggunakan *Performance Evaluation*. *Performance Evaluation* didapatkan melalui laporan performa *class* positif maupun negatif pada saat pengujian.

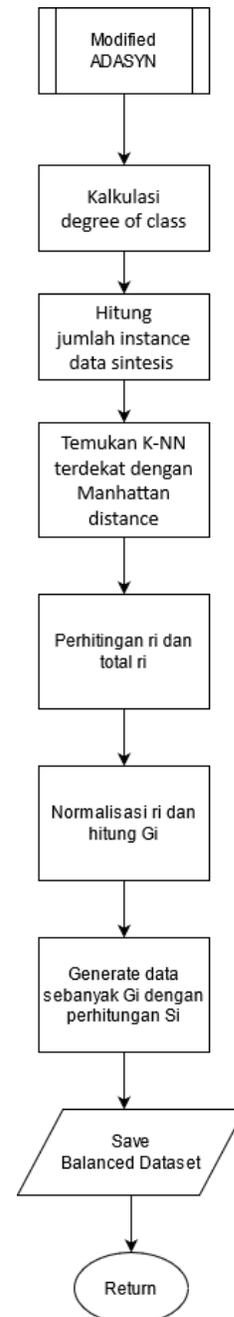
4.2 Desain

Pada subbab ini akan dijelaskan tentang desain bagaimana penelitian ini dapat mengatasi masalah yang ada diatas. Mulai dari masalah keamanan maupun masalah yang ada pada penelitian sebelumnya.



Gambar 2. Flowchart pengujian

Pada gambar 2 dijelaskan tentang *flowchart* pengujian yang akan dilakukan Pada penelitian ini. Pertama, data akan di seimbangkan menggunakan *Modified ADASYN*. Kedua, data akan dipisahkan antara training dan testing. Ketiga, data akan *training* menggunakan *Pipelining*. Keempat, data akan *testing* menggunakan *Pipelining*. Terakhir, *Class true* dan *class predict* akan dihitung performanya menggunakan *Classification Report*. Dimana *Class true* adalah *class* sebenarnya dari *data testing* sedangkan *class predict* adalah *class* hasil prediksi dari *data testing*.



Gambar 3. sub-routine dari Modified ADASYN

Pada gambar 3 ini dijelaskan tentang detail dari proses *Modified ADASYN*. Dimana *sub-subroutine* ini terdiri dari 6 langkah yang setiap langkahnya memiliki rumus masing-masing, seperti yang tertulis pada sub-bab 2.9.

5. HASIL PENGUJIAN

5.1 Perangkat Pengujian

Tabel 1. Perangkat yang digunakan

No	Nama perangkat	Jenis
1	Prosesor	Intel CORE i5
2	Ram	4 GB
3	Kartu grafis	AMD Radeon Graphics

Pada tabel 1. menjelaskan mengenai perangkat yang digunakan untuk melakukan pengujian pada program. Pada penelitian ini, perangkat harus mampu menjalankan *Windows 10*, *Anaconda Navigator*, *Jupyter Notebook* dan *Python* versi terbaru.

5.2 Software Pengujian

Tabel 2. Software yang digunakan

No	Nama Software	Penggunaan
1	<i>Windows 10</i>	<i>Operating System</i>
2	<i>Anaconda Navigator</i>	<i>Cloud Jupyter Notebook</i>
3	<i>Jupyter Notebook</i>	<i>Compiler Python</i>

Dari hasil pengujian yang didapatkan, perangkat pada tabel 1. mampu menjalankan software yang ada pada tabel 2. Dengan versi terbaru. Namun waktu yang dibutuhkan untuk *Run* program cukup lama, sehingga butuh *Ram* yang lebih besar.

5.3 Hasil Pengujian Pertama sampai Empat

Tabel 3. Pengujian 1 sampai 4

No. Pengujian	Dataset	Presisi		Akurasi
		Class 0	Class 1	
1, 3 dan 4	<i>Anonymized credit card transactions + Modified ADASYN + Pipelining</i>	1.00	1.00	1.00
1, 3 dan 4	<i>Anonymized credit card transactions + Modified ADASYN + Pipelining</i>	1.00	1.00	1.00

Tabel 3. Lanjutan

1, 3 dan 4	<i>Anonymized credit card transactions + Modified ADASYN + Pipelining</i>	1.00	1.00	0.999908
1	<i>Abstract dataset for Credit card fraud + Modified ADASYN + Pipelining</i>	0.99	0.99	0.991930
2	<i>Anonymized credit card transactions + Modified ADASYN + KNN</i>	1.00	1.00	0.9995148
2	<i>Abstract dataset for Credit card fraud + Modified ADASYN + K-NN</i>	0.99	0.96	0.97617554
4	<i>Anonymized credit card transactions + Tanpa Penyeimbang Class + Pipelining</i>	1.00	1.00	0.999090
4	<i>Anonymized credit card transactions + Tanpa Penyeimbang Class + Pipelining</i>	1.00	1.00	0.999595
4	<i>Anonymized credit card transactions + Tanpa Penyeimbang Class + Pipelining</i>	1.00	1.00	0.999636

5.4 Diskusi

Dari 4 pengujian yang dilakukan, menunjukkan bahwa performa *Modified ADASYN* dan *Pipelining* yang menggunakan $K=28$ pada fungsi *selectKbest* di *dataset* pertama sangat baik. Ini dibuktikan dengan semua presisinya mendapat nilai 1.00 sedangkan untuk akurasi hanya turun pada data sample 50.000. Ini dikarenakan nilai K optimal pada K -NN dalam *Modified ADASYN* sangat sulit ditemukan. Pada penelitian ini nilai K optimal pada K -NN dalam *Modified ADASYN* ditentukan dengan cara *trial and error*. Sedangkan performa *Modified ADASYN* dan *Pipelining* yang pada *dataset* kedua kurang baik karena akurasi dan presisinya hanya sampai 0.99.

6. KESIMPULAN DAN SARAN

Bagian ini akan menjelaskan mengenai kesimpulan yang didapat dalam penerapan *Modified ADASYN* untuk pendeteksian *fraud* pada transaksi kartu kredit. Kesimpulan didapatkan dari 4 pengujian dan saran adalah kesimpulan untuk membuat penelitian selanjutnya berdasarkan 4 yang sudah dilakukan.

Metode *Modified ADASYN* yang diterapkan pada penelitian ini sudah cukup baik untuk mendeteksi penipuan kartu kredit jika dilihat dari akurasi, presisi, *recall*, *F1* karena hampir semua nilainya 1.00. Namun penelitian ini masih memiliki masalah dalam menentukan nilai *K* yang optimal. Selain itu program yang ada pada penelitian ini masih belum terlalu optimal dikarenakan waktu untuk *run* program membutuhkan waktu yang cukup lama, yaitu sekitar 20 menit untuk 50.000 data sampel. Untuk itu dibutuhkan perbaikan program agar data sampel yang digunakan bisa *full* dari jumlah *row* pada *dataset* yang ada.

6.1 Kesimpulan

- Untuk menjawab rumusan masalah no.1, maka disimpulkan bahwa akurasi *Modified ADASYN* dan *Pipelining* hanya mampu 1.00 pada *dataset* pertama namun tidak pada *dataset* kedua.
- Nilai *K* yang optimal pada *Modified ADASYN* tergantung dari *dataset*.
- Sebagai jawaban dari menjawab rumusan masalah no.2, maka disimpulkan bahwa *Modified ADASYN* mampu membuat akurasi *K-NN* hingga 0.9995148 dan 0.97617554 pada *dataset* pertama dan kedua.
- Sebagai jawaban dari menjawab rumusan masalah no.3, maka disimpulkan bahwa *Modified ADASYN* dan *Pipelining* mampu mengungguli *SMOTE* dan *Random Forest*. *ADASYN* dan *Pipelining* pada penelitian sebelumnya.
- Nilai *K* yang optimal pada *Modified ADASYN* tergantung dari jumlah sampel datanya.
- Untuk menjawab rumusan masalah no.4, maka disimpulkan bahwa *Modified ADASYN* terbukti tidak membuat performa *Pipelining* menurun jika dibandingkan dengan performa *Pipelining* yang tidak menggunakan metode penyeimbang *class*.
- Untuk menjawab rumusan masalah no.5 maka disimpulkan bahwa presisi *Modified ADASYN* cukup tinggi. Presisi *Modified ADASYN* dan *Pipelining* pada *dataset* pertama adalah 1.00, 1.00, 1.00 dengan data sampel 10.000, 15.000, 50.000. Sedangkan pada *dataset* kedua presisinya adalah 0.99. Presisi *Modified ADASYN* dan *K-NN* pada *dataset* pertama adalah 1.00 untuk *class 0* dan *class 1*. Sedangkan pada *dataset* kedua presisinya adalah 0.99 dan 0.96 untuk *class 0* dan *class 1*.

6.2 Saran

- Dalam penelitian ini nilai *K* pada *Modified ADASYN* sangat sulit ditemukan, untuk itu disarankan untuk mencari metode yang bisa menemukan nilai *K* yang optimal untuk macam-macam *dataset* dan jumlah sampel data pada penelitian selanjutnya.
- Dalam penelitian ini untuk *run* program membutuhkan waktu yang cukup lama, untuk itu disarankan untuk membuat program dengan mengurangi *loop* pada penelitian selanjutnya.

7. DAFTAR PUSTAKA

- Bagga, S., Goyal, A., Gupta, N., & Goyal, A. 2020. Credit Card Fraud Detection using Pipeling and Ensemble Learning. *Procedia Computer Science*, 173, 104-112. DOI= <https://doi.org/10.1016/j.procs.2020.06.014>
- Dornadula, V. N., & Geetha, S. 2019. Credit card fraud detection using machine learning algorithms. *Procedia Computer Science*, 165, 631-641. DOI= <https://doi.org/10.1016/j.procs.2020.01.057>
- Fraugster. *The State Of Credit Card Fraud* 2021. 2021. Google Book Search. URI= https://fraugster.cdn.prismic.io/fraugster/9cedffd2-9339-4111-9b0d-d54a148de932_Fraugsters+State+Of+Credit+Card+Fraud+2021.pdf
- Gameng, H. A., Gerardo, B.B., & Medina, R.P. 2019. Modified Adaptive Synthetic SMOTE to Improve Classification Performance in Imbalanced Datasets. 2019 *IEEE 6th International Conference on Engineering Technologies and Applied Sciences (ICETAS)*, pp. 1-5. DOI= <https://doi.org/10.1109/ICETAS48360.2019.9117287>.
- Han, J., Kamber, M., & Pei, J. 2011. *Data Mining: Concepts and Techniques 3rd edition*. Elsevier Science. URI= <http://myweb.sabanciuniv.edu/rdehkharghani/files/2016/02/The-Morgan-Kaufmann-Series-in-Data-Management-Systems-Jiawei-Han-Micheline-Kamber-Jian-Pei-Data-Mining.-Concepts-and-Techniques-3rd-Edition-Morgan-Kaufmann-2011.pdf>
- Johnson, J. M., & Khoshgoftaar, T.M. 2019. Survey on Deep Learning with Class Imbalance. *Journal Of Big Data*, 6:27. DOI= <https://doi.org/10.1186/s40537-019-0192-5>
- Kho, J. R., & Vea, L. A. 2017. Credit card fraud detection based on transaction behavior. *TENCON 2017 - 2017 IEEE Region 10 Conference*. DOI= <https://doi.org/10.1109/TENCON.2017.8228165>
- Makki, S., Assaghir, Z., Taher, Y., Haque, R., Hacid, M. S., & Zeineddine, H. 2019. An experimental study with imbalanced classification approaches for credit card fraud detection. *IEEE Access*, 7, 93010-93022. DOI= <https://doi.org/10.1109/ACCESS.2019.2927266>
- Sklearn.feature_selection.f_regression*. 2020. scikit.(n. d). URI= https://scikit-learn.org/stable/modules/generated/sklearn.feature_selection.f_regression.html
- Sklearn.feature_selection.selectKBest*. 2020.scikit. (n.d). URI= https://scikit-learn.org/stable/modules/generated/sklearn.feature_selection.SelectKBest.html
- Sklearn.pipeline.Pipeline*. 2020. scikit. (n. d). URI= <https://scikit-learn.org/stable/modules/generated/sklearn.pipeline.Pipeline.html>.