

***UNDERSAMPLING MAJORITY CLASS PADA KASUS IMBALANCED
DATASET DAN APLIKASINYA PADA DETEKSI ANOMALI TRANSAKSI
KARTU KREDIT***

SKRIPSI

**Diajukan untuk memenuhi persyaratan mendapatkan gelar Strata Satu Program Studi
Informatika**



Disusun Oleh:

MUHSIN AHADI

M0515026

PROGRAM STUDI INFORMATIKA

FAKULTAS MATEMATIKA DAN ILMU PENGETAHUAN ALAM

UNIVERSITAS SEBELAS MARET

2019

HALAMAN PERSETUJUAN

SKRIPSI

***UNDERSAMPLING MAJORITY CLASS PADA KASUS IMBALANCED
DATASET DAN APLIKASINYA PADA DETEKSI ANOMALI
TRANSAKSI KARTU KREDIT***

Disusun Oleh:

MUHSIN AHADI

M0515026

Skripsi ini telah disetujui untuk dipertahankan di hadapan dewan penguji pada
tanggal 20 September 2019

Pembimbing I



Heri Prasetyo S.Kom, M.Sc.Eng., Ph.D.

NIP. 1983030220161001

Pembimbing II



Dr. Anto Satriyo Nugroho, M.Eng.

NIP. 197010211989111001

HALAMAN PENGESAHAN**SKRIPSI****UNDERSAMPLING MAJORITY CLASS PADA KASUS IMBALANCED
DATASET DAN APLIKASINYA PADA DETEKSI ANOMALI
TRANSAKSI KARTU KREDIT****Disusun Oleh:****MUHSIN AHADI****M0515026**

Telah dipertahankan di hadapan Dewan Penguji
pada tanggal 20 September 2019

Susunan Dewan Penguji

1. Heri Prasetyo S.Kom, M.Sc.Eng., Ph.D.
NIP. 1983030220161001
2. Dr. Anto Satriyo Nugroho, M.Eng.
NIP. 197010211989111001
3. Dr. Wisnu Widiarto S.Si., M.T.
NIP. 197006012008011009
4. Dr. Umi Salamah S.Si., M.Kom.
NIP. 197002171997022001

**Disahkan Oleh****Kepala Program Studi Informatika**

MOTTO

“Usaha yang kita tanam pada hari kemarin dan sekarang, adalah buah yang akan dipetik
kemudian hari” - *anonymous*



PERSEMBAHAN

Skripsi ini saya persembahkan untuk:

Ibu, Bapak, Bang Waji, Bang Uti, Mpo Dada, Bang Ibnu, dan semua orang yang telah memberikan dukungan moril dan materiil.



UNDER-SAMPLING MAJORITY CLASS PADA KASUS IMBALANCED DATASET DAN APLIKASINYA PADA DETEKSI ANOMALI TRANSAKSI KARTU KREDIT

MUHSIN AHADI

Program Studi Informatika

Fakultas Matematika dan Ilmu Pengetahuan Alam

Universitas Sebelas Maret

ABSTRAK

Kartu kredit adalah salah satu sistem pembayaran yang banyak digunakan di seluruh dunia. Tetapi sistem keamanan kartu kredit masih rentan terjadi *fraud*. Salah satu cara untuk mengatasi *fraud* adalah menggunakan *machine learning* dengan metode *anomaly detection*. *Anomaly detection* adalah metode pencarian pola yang janggal pada suatu keadaan. Metode tersebut dapat dilakukan dengan menggunakan algoritma klasifikasi. Masalah yang sering terjadi dalam klasifikasi adalah terjadinya *imbalanced dataset*, yaitu keadaan *class* tujuan yang diklasifikasi tidak seimbang rasionya. Masalah ini dapat menyebabkan hasil klasifikasi menjadi bias karena *classifier* lebih sering mendeteksi kelas mayoritas dibanding kelas minoritas. *Imbalanced dataset* dapat diatasi dengan menggunakan metode *undersampling*. Penelitian ini bertujuan untuk menangani *imbalanced dataset* dengan metode *undersampling K Nearest Neighbors-Undersampling* pada deteksi anomali transaksi kartu kredit dengan *classifier Support Vector Machine* dan *K-Nearest Neighbors* agar menghasilkan performa prediksi yang optimal. *K Nearest Neighbors-Undersampling* mengurangi kelas mayoritas dari 199013 menjadi 367, 661, 949, 1244, dan 1574 dengan variasi nilai K 4, 8, 12, 16, dan 20 secara berturut-turut. Metode *K Nearest Neighbors* mampu meningkatkan nilai *true positive* dan mengurangi nilai *false negative*. Hasil penelitian menunjukkan performa terbaik dihasilkan 20NN-Und + *Support Vector Machine* parameter C = 1 bekerja paling optimal dengan nilai *precision* 0.95 dan *recall* 0.75 dalam waktu 4.51 detik.

Kata kunci : *anomaly detection, undersampling, imbalanced dataset, fraud*

***UNDER-SAMPLING THE MAJORITY CLASS ON IMBALANCED DATASET
AND ITS APPLICATION ON ANOMALY DETECTION CREDIT CARD
TRANSACTION***

MUHSIN AHADI

Department of Informatics

Faculty of Mathematics and Natural Sciences

Sebelas Maret University

ABSTRACT

Credit card is one of the most widely used payment system all over the world. But credit card's security system still vulnerable against fraud. One way to handle fraud is using machine learning with anomaly detection method. Anomaly detection is a method to find strange pattern in some circumstances. It can be done using classification algorithm. The problem that often occurs in classification is imbalanced dataset, which is state of the objective class is in unbalanced ratio. This problem could lead the classification result biased because the classifier detects majority class more often than the minority class. Imbalanced dataset could be handled by using undersampling method. This research's goal is to handle imbalanced dataset using K Nearest Neighbors-Undersampling in anomaly detection credit card transaction using Support Vector Machine and K-Nearest Neighbors classifier in order to produce optimal predictive performance. K Nearest Neighbors-Undersampling reduced majority class from 199013 to 367, 661, 949, 1244, and 1574 with variance of K 4, 8, 12, 16, and 20 consecutively. K Nearest Neighbors could improved true positive value and reduced false negative value. The result showed best performance was produced by 20NN-Und + Support Vector Machine parameter $C = 1$, with precision 0.95 and recall 0.75 in 4.51 seconds.

Keyword: *anomaly detection, undersampling, imbalanced dataset, fraud*

KATA PENGANTAR

Puji syukur kepada Allah SWT atas segala limpahan rahmat dan karunia-Nya, penulis dapat menyelesaikan penulisan skripsi ini. Sholawat dan salam senantiasa penulis haturkan kepada Rasulullah SAW sebagai pembimbing seluruh umat manusia

Skripsi ini tidak akan selesai tanpa adanya bantuan dari banyak pihak, karena itu penulis menyampaikan terima kasih kepada:

1. Bapak Dr. Wiharto, S.T., M.Kom. selaku Kepala Program Studi Informatika yang telah memberikan dukungan selama proses perkuliahan dan penyusunan Tugas Akhir.
2. Bapak Heri Prasetyo, S.Kom., M.Sc., Ph.D selaku dosen pembimbing I atas ilmu, waktu, dan masukan yang diberikan, serta kesediaan untuk membimbing penulis selama pelaksanaan Tugas Akhir.
3. Bapak Dr. Anto Satriyo Nugroho, M.Eng. selaku dosen pembimbing II atas ilmu, waktu, dan masukan yang diberikan, serta kesediaan membimbing penulis selama penulisan Tugas Akhir.
4. Orang Tua dan keluarga penulis yang senantiasa memberikan dukungan dan doa kepada penulis untuk menyelesaikan Tugas Akhir.
5. Teman-teman Informatika 2015 atas segala kebersamaan dan dukungannya selama perkuliahan di Program Studi Informatika.
6. Semua pihak yang tidak dapat penulis sebutkan satu persatu.

Penyusunan skripsi ini tentunya masih ada beberapa kekurangan. Oleh sebab itu, kritik dan saran pembaca sangat diperlukan. Semoga skripsi ini bisa bermanfaat bagi pembaca maupun penulis sendiri.

Surakarta, Agustus 2019

Penulis

DAFTAR ISI

HALAMAN PERSETUJUAN.....	ii
HALAMAN PENGESAHAN	iii
MOTTO	iv
PERSEMBAHAN.....	v
ABSTRAK.....	vi
ABSTRACT.....	vii
KATA PENGANTAR	viii
DAFTAR ISI.....	ix
DAFTAR TABEL.....	xi
DAFTAR LAMPIRAN.....	xii
DAFTAR GAMBAR.....	xiii
BAB I PENDAHULUAN.....	1
1.1. Latar Belakang	1
1.2. Rumusan Masalah	3
1.3. Batasan Masalah.....	3
1.4. Tujuan Penelitian.....	4
1.5. Manfaat Penelitian.....	4
1.6. Sistematika Penulisan.....	4
BAB II TINJAUAN PUSTAKA	6
2.1. Dasar Teori.....	6
2.1.1. Anomaly Detection	6
2.1.2. Undersampling.....	9
2.1.3. K-Nearest Neighborss-Undersampling (KNN-Und)	10

2.1.4.	Support Vector Machine (SVM).....	14
2.1.5.	K-Nearest Neighborss Classifier.....	17
2.1.6.	Confusion Matrix	18
2.2.	Penelitian Terkait	20
BAB III METODOLOGI PENELITIAN		24
3.1.	Pengambilan Data	24
3.2.	Data Preprocessing.....	25
3.3.	Pembagian Training Set dan Testing Set	26
3.4.	Undersampling KNN-Und pada Training Set.....	26
3.5.	Klasifikasi menggunakan SVM dan KNN	26
3.6.	Evaluasi	26
BAB IV HASIL DAN PEMBAHASAN.....		27
4.1.	Dataset.....	27
4.2.	Hasil Data Preprocessing	28
4.3.	Pembagian Training Set dan Testing Set	31
4.4.	Undersampling dengan KNN-Und pada Training Set	31
4.5.	Klasifikasi Menggunakan SVM dan KNN.....	37
4.6.	Evaluasi	43
BAB V PENUTUP		51
5.1.	Kesimpulan.....	51
5.2.	Saran.....	51
DAFTAR PUSTAKA		52
LAMPIRAN.....		54

DAFTAR TABEL

Tabel 1 Euclidean Distance Minority Data Terhadap Majority Data	12
Tabel 2 3-Nearest Neighbors Minority Data Terhadap Majority Data	12
Tabel 3 Confusion Matrix Binary Classification	18
Tabel 4 10 Baris dan 8 Kolom Pertama pada Dataset	28
Tabel 5 Rangkuman Statistik Data Kolom Time – V5	28
Tabel 6 Rangkuman Statistik Data Kolom V5 – V11	28
Tabel 7 Rangkuman Statistik Data Kolom V12 – V17	29
Tabel 8 Rangkuman Statistik Data Kolom V19 – V24	29
Tabel 9 Rangkuman Statistik Data Kolom V25 - Class	29
Tabel 10 Kolom Amount Sebelum Scaling	30
Tabel 11 Kolom Amount Setelah Scaling	31
Tabel 12 Distribusi Training Set dan Testing Set	31
Tabel 13 Euclidean Distance Minority Data Terhadap Majority Data	32
Tabel 14 Urutan 4 Nearest Neighbors Terdekat	32
Tabel 15 Urutan 8 Nearest Neighbors Terdekat	33
Tabel 16 Urutan 12 Nearest Neighbors Terdekat	34
Tabel 17 Urutan 16 Nearest Neighbors Terdekat	35
Tabel 18 Urutan 20 Nearest Neighbors Terdekat	36
Tabel 19 Hasil Undersampling dengan Berbagai Nilai K	36
Tabel 20 Running Time Pelatihan SVM	37
Tabel 21 Running Time Pelatihan KNN	38
Tabel 22 Running Time Pengujian dengan SVM	39
Tabel 23 Running Time Pengujian dengan KNN	39
Tabel 24 False Negative dan True Positive pada SVM	46
Tabel 25 False Negative dan True Positive pada KNN	47
Tabel 26 Performa Klasifikasi pada SVM	48
Tabel 27 Performa Klasifikasi pada KNN	49

DAFTAR LAMPIRAN

Lampiran 1 Metode untuk Menangani Imbalanced Dataset 54
Lampiran 2 Confusion Matrix Semua Model 55



DAFTAR GAMBAR

Gambar 1 Objek di dalam Lingkaran adalah Outlier (Han et al., 2011).....	6
Gambar 2 Objek Hitam Membentuk Collective Outlier (Han et al., 2011).....	8
Gambar 3 Ilustrasi Undersampling	10
Gambar 4 Dataset Awal	11
Gambar 5 Proses 3NN-Und	13
Gambar 6 Hasil 3NN-Und	13
Gambar 7 SVM Berusaha Mencari Hyperplane Terbaik (Nugroho et al., 2003).....	14
Gambar 8 Margin pada Hyperlane (Nugroho et al., 2003).....	15
Gambar 9 Metodologi Penelitian.....	24
Gambar 10 Distribusi Class pada Dataset.....	27
Gambar 11 Decision Boundary SVM Tanpa Sampling.....	40
Gambar 12 Decision Boundary SVM Dengan Sampling	41
Gambar 13 Decision Boundary KNN Tanpa Sampling.....	42
Gambar 14 Decision Boundary KNN Dengan Sampling	42
Gambar 15 Confusion Matrix SVM C 0.1	43
Gambar 16 Confusion Matrix 8NN-Und + SVM C 0.1.....	44