TESIS – IF185401

METODE HIBRIDA *OVERSAMPLING* DAN *UNDERSAMPLING* UNTUK MENANGANI KETIDAKSEIMBANGAN DATA KEGAGALAN AKADEMIK UNIVERSITAS XYZ

SHABRINA CHOIRUNNISA
NRP. 05111750010029

DOSEN PEMBIMBING
Prof. Ir. Joko Lianto Buliali, M.Sc., Ph.D.
NIP. 196707271992031002

PROGRAM MAGISTER
BIDANG KEAHLIAN DASAR TERAPAN KOMPUTASI
DEPARTEMEN TEKNIK INFORMATIKA
FAKULTAS TEKNOLOGI INFORMASI DAN KOMUNIKASI
INSTITUT TEKNOLOGI SEPULUH NOPEMBER
SURABAYA
2019

[Halaman ini sengaja dikosongkan]

LEMBAR PENGESAHAN

Tesis ini disusun untuk memenuhi salah satu syarat memperoleh gelar
Magister Komputer (M. Kom)
di

Institut Teknologi Sepuluh Nopember Surabaya

oleh:
SHABRINA CHOIRUNNISA
Nrp. 05111750010029

Dengan judul:
Metode Hibrida *Oversampling* dan *Undersampling* untuk Menangani Ketidakseimbangan Data Kegagalan Akademik Universitas XYZ

Tanggal Ujian : 18 Januari 2019
Periode Wisuda : 2019 Ganjil

Disetujui oleh:

Prof. Ir. Joko Lianto Buliali, M.Sc., Ph.D. (Pembimbing)
NIP. 196707271992031002

Daniel O. Siahaan, S.Kom., M.Sc., Ph.D. (Penguji 1)
NIP. 19772172003121001

Dr. Eng. Chastine Fatichah, S.Kom., M.Kom. (Penguji 2)
NIP. 198611252018031001

Hadziq Fabroyir, S. Kom., Ph.D. (Penguji 3)
NIP. 1986201911089

Dekan Fakultas Teknologi Informasi dan Komunikasi,

Dr. Agus Zainal Arifin, S.Kom, M.Kom
NIP. 19720809 199512 1 001

[Halaman ini sengaja dikosongkan]

PERNYATAAN KEASLIAN

Dengan ini saya menyatakan bahwa isi sebagian maupun keseluruhan Tesis saya dengan judul:

METODE HIBRIDA *OVERSAMPLING* DAN *UNDERSAMPLING* UNTUK MENANGANI KETIDAKSEIMBANGANDATA KEGAGALAN AKADEMIK UNIVERSITAS XYZ

adalah benar-benar hasil karya intelektual mandiri, diselesaikan tanpa menggunakan bahan-bahan yang tidak diijinkan dan bukan merupakan karya pihak lain yang saya akui sebagai karya sendiri.

Semua referensi yang dikutip maupun dirujuk telah ditulis secara lengkap pada daftar pusaka.

Apabila ternyata pernyataan ini tidak benar, saya bersedia menerima sanksi sesuai dengan peraturan yang berlaku.

Surabaya, 18 Januari 2019

Shabrina Choirunnisa
NRP: 05111750010029

[Halaman ini sengaja dikosongkan]

METODE HIBRIDA *OVERSAMPLING* DAN *UNDERSAMPLING* UNTUK MENANGANI KETIDAKSEIMBANGAN DATA KEGAGALAN AKADEMIK PADA UNIVERSITAS XYZ

Nama mahasiswa : Shabrina Choirunnisa
NRP : 05111750010029
Pembimbing : Prof. Ir. Joko Lianto Buliali, M.Sc., Ph.D

ABSTRAK

Ketidakseimbangan (*Imbalance*) data terjadi pada berbagai macam data termasuk data akademik Universitas XYZ. Apabila terhadap data akademik Universitas XYZ dilakukan proses komputasi (misalnya *klasifikasi*), adanya *imbalance* data tersebut berpotensi menyebabkan terjadinya misklasifikasi karena data mayoritas lebih dominan terhadap data minoritas. Metode kombinasi dari *oversampling* dan *undersampling* dapat menjadi salah satu solusi dalam menyelesaikan kasus *imbalance*. Penelitian ini bertujuan menangani permasalahan *imbalance* data akademik dengan memadukan metode *oversampling* dengan metode *undersampling* sehingga diperoleh data sintetik yang lebih representatif. Pada penelitian ini, *Adaptive Semi-unsupervised Weighted Oversampling* (A-SUWO) digunakan sebagai metode *oversampling.* Dan metode *undersampling* yang digunakan adalah: *Random Undersampling* (RUS), *Neighborhood Cleaning Rule* (NCL), dan Tomek-Link. Setelah dilakukan proses *undersampling* dan *oversampling*, data diklasifikasi menggunakan algoritma *Decision Tree* C4.5. Evaluasi performa diproses menggunakan perhitungan *precision*, *recall*, dan akurasi. Diperoleh nilai rata-rata akurasi tertinggi yang dicapai yaitu 76.55% dengan nilai *precision* dan *recall* sebesar 87.04%, 80.35% untuk gabungan metode A-SUWO-Tomeklinks pada dataset akademik. Sedangkan pada dataset Keel, diperoleh nilai akurasi, *precision*, dan *recall* yakni 85.41%, 93.18%, 90.54%.

Kata kunci: *Imbalance, undersampling, oversampling, RUS, NCL, Tomek Link, A-SUWO, data akademik*

[Halaman ini sengaja dikosongkan]

HYBRID METHOD OF OVERSAMPLING AND UNDERSAMPLING FOR HANDLING ACADEMIC FAILURE DATA AT XYZ UNIVERSITY

Student's Name	:	Shabrina Choirunnisa
Student ID	:	05111750010029
Supervisor	:	Prof. Dr. Ir. Joko Lianto Buliali, M.Sc.

ABSTRACT

Imbalance of data occurs in various kinds of data including XYZ University academic data. If the computation process of the XYZ University academic data is carried out (for example classification), the data imbalance has the potential to cause misclassification because the majority data is more dominant than the minority data. The combination method of oversampling and undersampling can be one solution in solving imbalance cases. This study aims to address the problem of imbalance of academic data by combining the oversampling method with the undersampling method to obtain more representative synthetic data. In this study. Adaptive Semi-unsupervised Weighted Oversampling (A-SUWO) is used as an oversampling method. While the undersampling methods used were: Random Undersampling (RUS), Neighborhood Cleaning Rules (NCL), and Tomek Link. After the undersampling and oversampling process is carried out, the data is classified using the Decision Tree C4.5 algorithm. Performance evaluation is processed using the calculation of precision, recall, and accuracy. Performance evaluation is processed using calculations of precision, memory, F-measure and accuracy. The highest average accuracy value obtained was 76.55% with precision and recall values of 87.04%, 80.35% for the combined A-SUWO-Tomeklinks method in the academic dataset. While in the Keel dataset, the values of accuracy, precision, and recall obtained were 85.41%, 93.18%, 90.54%.

Keyword: *Imbalance, undersampling, oversampling, RUS, NCL, Tomek Link, A-SUWO, academic data*

[Halaman ini sengaja dikosongkan]

KATA PENGANTAR

Puji syukur penulis panjatkan ke hadirat Allah SWT, atas segala rizki, berkah, nikmat serta karunia-Nya yang terlimpahkan kepada penulis, sehingga penulis akhirnya dapat menyelesaikan penelitian dengan judul **"Metode Hibrida *Oversampling* Dan *Undersampling* Untuk Menangani Ketidakseimbangan Data Kegagalan Akademik Pada Universitas XYZ"**

Penulis juga ingin mengucapkan banyak terimakasih karena tanpa bantuan dari berbagai pihak, penelitian ini tidak akan terselesaikan dengan hasil seperti sekarang ini. Oleh sebab itu pada kesempatan ini penulis ingin menyampaikan terima kasih yang sebesar-besarnya, kepada:

1. Allah SWT atas limpahan nikmat iman, islam, kesehatan, waktu, serta berbagai kemudahan dari arah yang tidak pernah diduga sebelumnya, sehingga penulis dapat menyelesaikan penelitian ini dengan baik.
2. Keluarga tercinta (Ibu, Ayah, dan Dik Rifda) yang tidak hentinya memberikan dukungan materil, do'a yang tulus, serta *belief* bahwa penulis harus senantiasa menjadi yang insan terbaik yang bermanfaat untuk orang sekitar dimanapun penulis berada.
3. Mas Khafidurrohman Agustianto, S.Pd., M. Eng., selaku partner hidup yang selalu memberikan doa, *support*, dan arahan untuk selalu sabar, ikhlas, serta senantiasa tawakal dan yakin kepada Allah sehingga thesis ini dapat selesai tepat pada waktu yang ditargetkan.
4. Bapak Prof. Ir. Joko Lianto Buliali, M.Sc., Ph.D. selaku dosen pembimbing yang telah banyak meluangkan waktu dan dengan sangat sabar mendidik, dan membimbing, dalam menyelesaikan penelitian ini.
5. Bapak Daniel O. Siahaan, S.Kom., M.Sc., Ph.D., Ibu Dr. Eng. Chastine Fatichah, S.Kom., M.Kom., Bapak Dr. Eng. Darlis Heru Mukti, S.Kom., M.Kom., Bapak Bagus Jati Santoso, S.Kom., Ph.D., serta Bapak Hadziq Fabroyir, S. Kom., Ph.D selaku dosen penguji yang telah memberikan banyak saran dan arahan agar penulis mampu lebih baik dalam menyelesaikan penelitian.

6. Bapak Waskitho Wibisono, S.Kom., M.Eng. Ph.D., dan Ibu Dr. Eng. Chastine Fatichah, S.Kom., M.Kom., selaku Kaprodi dan Wakaprodi S2 Teknik Informatika ITS Surabaya serta Bapak Dr. Eng. Darlis Heru Mukti dan Bapak Radityo Anggoro selaku Kaprodi dan Wakaprodi S1 yang memfasilitasi mahasiswanya untuk belajar di Lab S2 hingga larut dalam rangka menyelesaikan penelitian.
7. Pak Tora Fahrudin, dan Mas Meyda Cahyo selaku peneliti terdahulu yang telah memberikan segala ilmu, data, arahan, serta bimbingan dalam meneruskan penelitian beliau sehingga menjadi penelitian ini.
8. Seluruh anggota Laboratorium DTK: Bapak Endyk Noviyantono, Bapak Mudjahidin, Bapak Mohammad Yazdi Pusadan, Ibu Eviana Tjatur Putri, Ibu Anggreni, Ibu Myrna Ermawati, Ibu Eva Firdayanti Bisono, Mas Tegar, Mas Achmad Saiful, dan Mas Reza Prasetya Prayogo.
9. Seluruh staf dosen, staf tata usaha dan karyawan perpustakaan Jurusan Teknik Informatika, Institut Teknologi Sepuluh Nopember.
10. Rekan-rekan S2 Teknik Informatika atas bantuan dan diskusi selama penelitian.
11. Serta pihak-pihak yang tidak dapat dituliskan satu per satu oleh penulis, terima kasih banyak atas doa dan dukungannya.

Semoga Allah SWT senantiasa menyayangi, menguatkan, memampukan, dan menunjukan jalan yang terbaik atas semua kebaikan yang telah diberikan. Penulis menyadari bahwa laporan penelitian ini tentunya masih jauh dari kesempurnaan. Oleh sebab itu, saran, dan kritik sangat diharapkan untuk perbaikan dimasa yang akan datang. Semoga laporan penelitian ini dapat bermanfaat bagi penulis dan pembaca pada umumnya.

Surabaya, Januari 2019

Shabrina Choirunnisa

DAFTAR ISI

DAFTAR GAMBAR

[Halaman ini sengaja dikosongkan]

DAFTAR TABEL

DAFTAR LAMPIRAN

[Halaman ini sengaja dikosongkan]

BAB 1
PENDAHULUAN

Pada Bab ini akan dijelaskan mengenai beberapa hal dasar dalam pembuatan proposal penelitian yang meliputi latar belakang, perumusan masalah, tujuan, manfaat, kontribusi penelitian, dan batasan masalah.

1.1. Latar Belakang

Suatu instansi pendidikan, pada umumnya memiliki data akademik mahasiswa yang diambil secara berkala dan selalu bertambah setiap tahun. Tidak terkecuali pada Universitas XYZ. Data akademik Universitas XYZ tersebut diambil dari penelitian *"Deteksi Dini Kegagalan Akademik Serta Multilabelisasi Permasalahan Mahasiswa Dari Data Media Sosial"* (Fahrudin and Fatichah, 2016). Data akademik Universitas XYZ memiliki beberapa atribut diantaranya data presensi, data nilai (kuis, UTS, maupun UAS), dan data aktivitas lainnya seperti media sosial dan organisasi dimana data tersebut diakumulasi setiap bulan dalam 1 semester. Data akumulasi ini dapat dijadikan sebagai acuan maupun bahan analisis untuk menentukan keberhasilan proses akademik mahasiswa maupun institusi. Namun, seringkali terjadi ketidakseimbangan data, dimana jumlah data pada suatu kelas lebih dominan secara signifikan dibandingkan dengan jumlah data pada kelas lainnya. Dalam sistem komputasi, ketidakseimbangan data disebut dengan *imbalance* (Nekooeimehr and Lai-yuen, 2016).

Ketidakseimbangan pada data akademik Universitas XYZ ditandai dengan jumlah mahasiswa lulus tahun pertama secara normal lebih banyak dari pada mahasiswa yang tidak lulus. Mahasiswa yang tidak dapat melewati evaluasi tahap pertama dalam waktu normal (1 tahun) diartikan sebagai data minoritas. Sedangkan, mahasiswa yang mampu melewati tahap pertama disebut data mayoritas. Fenomena *imbalanced* ini dapat menyebabkan akurasi data minoritas menjadi rendah (Jayasree and Gavya, 2015). Selain itu, pendistribusian yang tidak seimbang juga dapat menimbulkan proses klasifikasi akan lebih condong pada kelas mayoritas dibandingkan dengan jumlah data minoritas (Sáez *et al*., 2016). Sedangkan Mellor *et al.* (2015) menyatakan bahwa kasus *imbalanced* ternyata cukup berperan terhadap terjadinya kesalahan pada proses

klasifikasi (*misclassified)*. Selain itu, *imbalance* dapat menyebabkan *overfitting* dan pembuatan model yang buruk (Gong and Kim, 2017).

Berdasarkan permasalahan-permasalahan tersebut, maka kasus *imbalanced* memerlukan penanganan khusus sehingga diperoleh model yang memiliki ketepatan prediksi yang optimal pada semua kelas data (Rivera, 2017; Piri, Delen and Liu, 2018). Beberapa solusi yang dapat menangani kasus *imbalanced* diantaranya: metode *oversampling* (Chawla *et al.*, 2002; Barua *et al.*, 2014; Piri *et al.*, 2018) dan *undersampling* (Purwar and Singh, 2015; John and Jayasudha, 2017), serta *hybrid* dari metode *oversampling* maupun *undersampling*. *Oversampling* dilakukan dengan membuat replika (*resample*) data minoritas, sedangkan metode *undersampling* dilakukan dengan mengurangi data mayoritas sehingga diperoleh data mayoritas dan minoritas yang lebih seimbang (Barua *et al.*, 2014; Fahrudin, Buliali and Fatichah, 2016; Fakhruzi, 2018).

Metode *oversampling* yang berlebihan dapat menyebabkan *overfitting* sedangkan *undersampling* yang berlebihan dapat berpengaruh pada hilangnya beberapa informasi penting yang terdapat pada dataset (Seiffert *et al.*, 2009; Napierała, 2012a). Salah satu metode oversampling yang paling dikenal yaitu *Synthethic Minority Oversampling Technique* (SMOTE) dimana metode ini dapat menangani *overfitting* data sintetik pada *oversampling* melalui pendekatan K-NN (Chawla *et al.*, 2002; Blagus *et al.*, 2013) dengan penggunaan variabel (Blagus *et al.*, 2013). Selain SMOTE, (Barua *et al.*, 2014) mengusulkan *Majority Weighted Minority Oversampling Technique* (MWMOTE) sebagai metode pembuatan data sintetik melalui pembobotan dan *clustering* data minoritas. Namun, MWMOTE masih memiliki beberapa kelemahan diantaranya data sintetik pada kelas minoritas yang dihasilkan oleh MWMOTE seringkali *overlap* dengan data pada kelas mayoritasnya. Overlap pada data sintetis dapat mengakibatkan rendahnya performa pada *classifier* secara signifikan (Nekooeimehr dan Lai-yuen 2016). Oleh karena itu, Nekooeimehr *et al.* mengusulkan suatu metode berbasis *semi-unsupervised hierarchical clustering* untuk mengatasi permasalahan tersebut dan disebut dengan *Adaptive Semi-unsupervised Weighted Oversampling* (A-SUWO). Usulan ini ternyata mampu memperbaiki dan meningkatkan hasil akurasi.

Untuk meningkatkan performa metode *oversampling*, beberapa peneliti menambahkan metode *undersampling* sebagai metode pembersihan (Ramentol,

Caballero and Bello, 2011). Penggabungan dengan metode *undersampling* ini diharapkan agar data yang diproses menjadi lebih bersih dan terhindar dari noise sehingga mampu meningkatkan kemampuan metode *oversampling* dalam membuat data sintetik (menghindari pembuatan data sintetik dengan mereplika data *noise*). Metode-metode *undersampling* yang digunakan pada penelitian ini antara lain: *Neighborhood Cleaning Rule* (NCL) (Laurikkala, 2001), Tomek Link (Tomek, 1976), dan *Random Undersampling* (RUS) (Prusa *et al.*, 2015). Sehingga dapat dikatakan, penelitian ini berfokus pada penanganan data yang tidak seimbang (*imbalanced*) dengan memadukan metode *oversampling (A-SUWO)* serta metode *undersampling* (RUS, NCL, dan Tomek Link). Sedangkan pengukuran performa pada sistem didasarkan pada perhitungan nilai F-*measure*, *recall*, *precision,* dan *accuracy.*

1.2. Perumusan Masalah

Rumusan masalah yang dibahas dalam penelitian ini dipaparkan sebagai berikut:

1. Bagaimana cara menangani *imbalance* pada data kegagalan akademik Universitas XYZ?
2. Bagaimana mencari kombinasi teknik *undersampling* dan *oversampling* yang tepat pada data akademik Universitas XYZ?

1.3. Tujuan Penelitian

Tujuan yang akan dicapai dalam penelitian ini adalah menggabungkan metode untuk mengatasi permasalahan ketidakseimbangan jumlah data pada data akademik dengan mengkombinasikan metode *oversampling (Adaptive Semi-unsupervised Weighted Oversampling* (A-SUWO) dengan metode *undersampling* (*Random Undersampling* (RUS), *Neighborhood Cleaning Rule* (NCL), dan Tomek Link sehingga dapat diperoleh *F-measure*, *Recall*, *Precision,* dan akurasi klasifikasi yang lebih baik.

1.4. Manfaat Penelitian

Manfaat dari penelitian ini diantaranya untuk mengimplementasi metode hibrida *oversampling* dan *undersampling* pada data akademik Universitas XYZ. Hal tersebut diharapkan dapat meningkatkan nilai akurasi dan menghasilkan data yang lebih representatif.

1.5. Kontribusi Penelitian

Kontribusi yang diharapkan dari penelitian ini adalah penggabungan metode *oversampling* dan *undersampling* untuk menghasilkan *F-measure*, *Recall*, *Precision,* dan akurasi klasifikasi pada data akademik yang lebih baik.

1.6. Batasan Masalah

Batasan masalah pada penelitian ini adalah:

- Data yang digunakan merupakan data akademik Universitas XYZ selama 1 semester pada bulang Agustus-Desember tahun ajaran 2014/ 2015.
- Metode sampling yang digunakan yaitu kombinasi dari metode metode *oversampling: Adaptive Semi-unsupervised Weighted Oversampling* (A-SUWO) dengan *undersampling*: *Random Undersampling* (RUS), *Neighborhood Cleaning Rule* (NCL), dan Tomek Link. Jadi uji coba akan dilakukan secara kombinatorial sehingga akan diperoleh 3 pasangan komparasi metode.
- Scope pada penelitian bukan melakukan deteksi dini terhadap kegagalan akademik mahasiswa.
- Metode yang diajukan adalah untuk menangani data akademik yang *imbalance*. Selain tidak seimbang, data akademik yang diproses pada metode yang diusulkan memiliki karakteristik lain yaitu *overlap* antar data pada kelas mayoritas dan minoritas.
- Perhitungan parameter pada uji coba dibatasi menggunakan *Recall*, *Precision*, dan *F-Measure* untuk mengevaluasi performa dari sistem.
- Perbandingan dalam membagi data training dan testing akan ditentukan berdasarkan nilai presentase tertentu.

BAB 2
KAJIAN PUSTAKA

Bab ini merupakan pembahasan dari referensi terkait yang telah dilakukan dalam menyelesaikan permasalahan sesuai dengan uraian pada latar belakang. Bab ini diawali dengan menjabarkan hal-hal yang diterapkan pada metode yang diusulkan, kelemahan yang terdapat pada penelitian sebelumnya, komparasi penelitian sebelumnya. Selanjutnya dilanjutkan dengan kelebihan dari metode yang akan digunakan untuk menyelesaikan permasalahan *imbalance* pada data akademik Universitas XYZ.

2.1. Data Mining

Informasi merupakan salah satu elemen yang sangat penting dalam berbagai bidang. Untuk memeperoleh informasi yang representatif dan akurat seringkali dibutuhkan proses yang panjang dan tidak mudah. Terlebih untuk informasi dalam jumlah yang sangat besar, diperlukan proses penggalian terlebih dahulu agar dapat menyajikan informasi yang sesuai dengan kebutuhan. Selain itu, Analisa juga sangat diperlukan untuk mengolah data agar mampu menghasilkan informasi yang lebih bermanfaat dalam pengambilan keputusan pada suatu masalah. Oleh karena itu, untuk membantu dalam mempermudah proses pengambilan keputusan dan Analisa tersebut diperlukan *data mininig* (Asriningtias *et al.*, 2014).

Jadi, data mining dapat diartikan sebagai suatu proses penggalian informasi dan pengenalan pola dari suatu kumpulan data. Data mining sangat perlu dilakukan terutama dalam mengelola data yang sangat besar untuk memudahkan aktivitas *data recording* agar dapat memberikan informasi yang lebih akurat bagi penggunanya. Data mining juga sudah sangat umum digunakan untuk memecahkan segala permasalahan (yang masih berkaitan dengan data) di dunia industri, akademik, *science*, kesehatan, dan lain sebagainya. Selain itu, *data mining* juga memiliki kaitan erat dengan bidang ilmu yang lain seperti kecerdasan buatan, *machine learning*, statistik, dan *database* (Daniel, 2005). Dengan memanfaatkan bidang ilmu tersebut, pada aplikasinya, data mining dapat lebih berpotensi untuk mendapatkan informasi yang lebih akurat dan presisi.

Terdapat beberapa tahapan secara umum pada proses *data mining* itu sendiri, diantaranya (Han, 2012):

- Pembersihan Data (*Data Cleaning*)

  Pada tahap ini, dilakukan proses pembersihan data terhadap *noise* atau merupakan data yang tidak relevan. Pada umumnya, data yang tidak relevan tersebut dihapus atau dibuang.
- Integrasi Data (*Data Integration*)

  Integrasi data merupakan penggabungan data dari berbagai database ke dalam satu database baru. Proses ini perlu dilakukan dengan cermat agar tidak terjadi kesalahan pada integrasi data yang dapat menghasilkan penyimpangan data.
- Seleksi Data (*Data Selection*)

  Data yang ada pada database sering kali tidak semuanya dipakai, oleh karena itu hanya data yang relevan untuk dianalisis yang akan diambil dari database.
- Transformasi Data (*Data Transformation*)

  Dimana data diubah atau digabungkan menjadi format yang sesuai untuk diproses pada tahapan selanjutnya
- Proses Mining

  Merupakan suatu proses utama saat metode *mining* diterapkan untuk menggali informasi berharga yang tersembunyi dari suatu data.
- Evaluasi Pola (*Pattern Evalution*)

  Tahapan ini bertujuan untuk mengidentifikasi pola yang benar-benar menarik ke dalam knowledge based yang ditemukan. Dalam tahap ini hasilnya berupa pola-pola yang khas untuk menilai apakah hipotesa yang ada memang tercapai.
- Presentasi Pengetahuan (*Knowledge Presentation*)

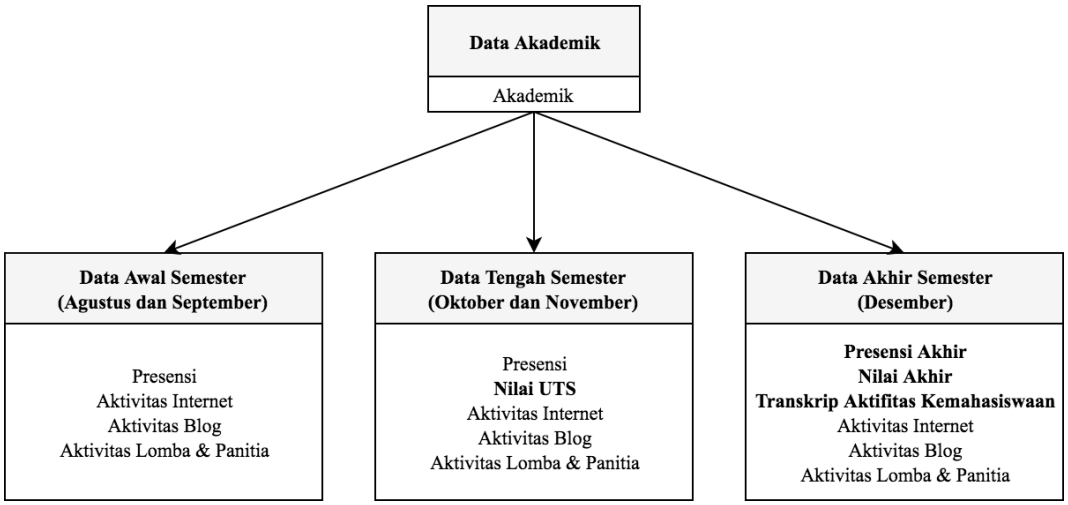
Merupakan visualisasi dan penyajian pengetahuan mengenai metode yang digunakan untuk memperoleh pengetahuan yang diperoleh pengguna.

2.2. Data Akademik

Data akademik diperoleh langsung dari basis data Sistem Informasi Akademik mahasiswa Universitas XYZ angkatan 2014/2015 (Fahrudin and Fatichah, 2016). Data akademik memiliki 12854 baris data dengan jumlah atribut sebanyak 34 hingga 42. Jumlah data mewakili jumlah mahasiswa yang selalu tetap pada setiap bulan yaitu

12854 mahasiswa. Data tersebut diambil dari histori akademik mahasiswa. Data akademik digunakan sebagai inputan yang terbagi menjadi 3 skenario pemodelan: data awal semester, deteksi tengah semester dan deteksi akhir semester. Gambar 2.1 menjelaskan mengenai usulan data akademik yang digunakan di dalam rancangan pemodelan dengan perbedaan berupa atribut data masukan yang berbeda, menyesuaikan dengan proses masuknya data ke dalam Sistem Informasi akademik.

Pada data awal semester, atribut yang digunakan yaitu presensi, aktivitas internet, aktivitas blog, serta aktivitas lomba dan panitia. Pada data tengah semester, terdapat atribut tambahan yaitu nilai UTS. Sedangkan pada data akhir semester, nilai UTS berubah menjadi nilai akhir, presensi berubah menjadi presensi akhir dan terdapat tambahan atribut yaitu transkrip aktivitas kemahasiswaan.



Gambar 2.1. Data Akademik

Data akademik tersebut merupakan akumulasi data akademik pada bulan Agustus, September, Oktober, November, dan Desember. Data awal semester terdiri dari data pada bulan Agustus dan September. Data tengah semester merupakan data bulan awal ditambahkan dengan data pada bulan Oktober maupun November. Sedangan Data akhir semester merupakan akumulasi keseluruhan bulan termasuk bulan Desember.

Data akademik yang digunakan pada penelitian ini sudah dilakukan proses normalisasi terlebih dahulu (Fahrudin and Fatichah, 2016). Namun, penjelasan secara rinci terkait metode normalisasi yang dilakukan tidak dipaparkan karena hal tersebut sudah diluar *scope* dari penelitian ini.

2.3. Fenomena Ketidakseimbangan (Imbalanced)

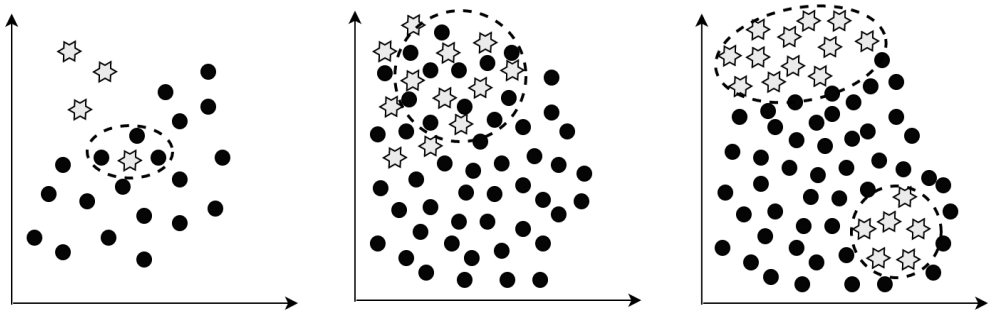
Suatu fenomena pada kumpulan data dapat dikatakan imbalanced atau tidak seimbang apabila jumlah data pada suatu kelas tertentu (kelas mayor) lebih banyak dibandingkan dengan jumlah data pada kelas yang lain (kelas minor)(Guo *et al.*, 2016). Kondisi ini merupakan salah satu kondisi yang krusial untuk ditangani pada kasus-kasus data mining karena dipercaya dapat mempengaruhi nilai akurasi saat proses klasifikasi data (Thanathamathee and Lursinsap, 2013) (Rushi Longadge and Malik, 2013).

Terdapat dua jenis pendekatan untuk menangani kasus *imbalance*, diantaranya pendekatan pada level data dan level algoritma (Mahmood, 2017). Pendekatan pada level data dilakukan dengan melakukan proses *sampling* pada data mayoritas ataupun data minoritas sehingga jumlah data menjadi lebih seimbang. Sedangkan pendekatan pada level algoritma yaitu melakukan improvisasi pada metode-metode *classifier* tanpa memproses atau merubah data awal. Kedua metode tersebut memiliki kelebihan dan kekurangan masing-masing, diantaranya untuk pendekatan level data cenderung tangguh dan stabil terhadap hampir seluruh *classifier*. Namun, kelemahan pada solusi ini yaitu memungkinkan terjadinya *overfitting* atau *missing information* pada data yang telah dilakukan *sampling*. Sedangkan pada solusi level algoritma, data yang diolah merupakan data asli tanpa perubahan apapun namun data tersebut akan sangat bergantung pada *classifier* tertentu. Dengan kata lain, algoritma yang diusulkan belum tentu mampu mencapai performa yang sama baik apabila diimplementasikan pada data lain yang memiliki karakteristik berbeda. Dari kedua jenis solusi ini, riset singkat berdasarkan jumlah penelitian yang ada menyatakan bahwa ternyata solusi pada level data lebih banyak dikembangkan untuk menangani kasus *imbalance* dibandingkan solusi dengan modifikasi algoritma (Napierała, 2012b). Secara umum, solusi pada level data menggunakan teknik sampling dibagi menjadi tiga jenis yaitu *oversampling*, *undersampling*, dan gabungan dari keduanya (*hybrid*).

Metode *oversampling* merupakan metode yang bertujuan untuk menambahkan jumlah data pada kelas minoritas dengan memanfaatkan teknik *sampling* pada data training kelas minoritas sehingga diharapkan rasio antar kelas minoritas dan kelas mayoritas dapat lebih berimbang. Sedangkan sebaliknya, teknik *undersampling* justru mengeliminasi sebagian data yang dianggap kurang relevan pada kelas mayoritas. Sedangkan metode *hybrid* merupakan kombinasi dari kedua teknik sampling tersebut

sesuai dengan kebutuhan dan karakteristik data (Mahmood, 2017). Beberapa permasalahan yang sering muncul pada kasus *imbalanced* yaitu:

1. *Outlier* yaitu ketika data yang bernilai ekstrim atau beda sangat jauh dengan mayoritas kelompoknya. Pada kondisi ini seringkali terjadi misklasifikasi sehingga pada beberapa penelitian, *outlier* tersebut dihapus.
2. Banyaknya data antar kelas yang *overlap.* Apabila terdapat *overlapping*, maka discriminative rule akan sulit untuk diproses. Hal tersebut dapat berdampak pada semakin besarnya kemungkinan terjadi mis-klasifikasi pad akelas minoritas dikarenakan jumlah data yang lebih minim. Apabila overlap yang terjadi dapat diminimalisir, maka metode klasifikasi sederhana apapun akan dapat menghasilkan distribusi kelas dengan sangat baik (V. Garcıa, R. A. Mollineda, 2008). Gambar 2.2 berikut merupakan ilustrasi kondisi dari ketiga permasalahan tersebut (Weiss and Provost, 2003).
3. Terdapat beberapa data pada *sub-cluster* yang memiliki jarak terlalu rapat antar kedua kelas (*small disjunction*). Adanya sub-cluster yang saling berdekatan tersebut dapat menambah kompleksitas dari suatu data dikarenakan pada umumnya jumlah data pada setiap *sub-cluster* tersebut tidak berimbang.

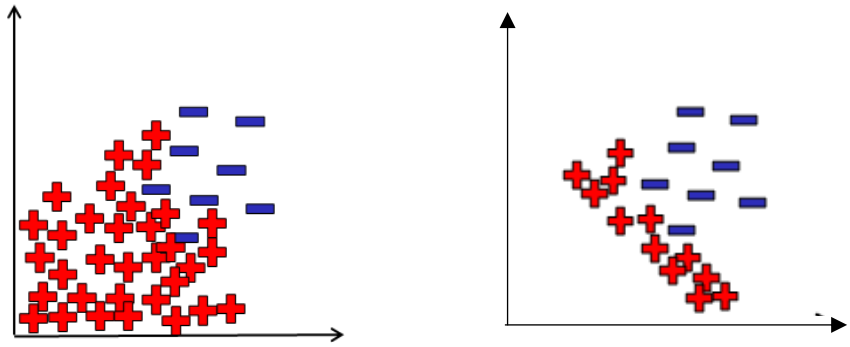


Gambar 2.2 Ilustrasi dari keadaan: (1) *outlier*, (2) *overlapping*, (3) *small disjunction*

2.4. Undersampling

Seperti dijelaskan pada sub bab sebelumnya, teknik *undersampling* merupakan proses sampling yang dilakukan dengan mengurangi atau mengeliminasi sebagian data pada kelas mayoritas pada data. Proses eliminasi tersebut dapat dilakukan secara random (paling sederhana) sehingga biasa disebut dengan *random undersampling.* Selain itu, *undersampling* juga dapat dilakukan dengan menggunakan perhitungan statistik yang biasa disebut dengan *informed undersampling*. Pada teknik ini, metode

iterasi dan teknik *data cleaning* juga diterapkan untuk menyaring data pada kelas mayoritas lebih lanjut.



Gambar 2.3 Ilustrasi proses *undersampling* secara umum

Gambar 2.3 merupakan ilustrasi setelah dilakukan *undersampling* dimana terdapat dua kelas yaitu kelas positif yang memiliki jumlah data mayoritas dan kelas negative yang memiliki jumlah data minoritas. Gambar pada sisi kiri merupakan data original sedangkan gambar pada sisi kanan merupakan data setelah dilakukan pembersihan menggunakan metode *undersampling*. merepresentasikan batas keputusan ideal sedangkan garis biru merupakan hasil aktual yang diperoleh. Metode *undersampling* menyebabkan beberapa informasi pada kelas negatif terhapus dan proporsi jumlah data pada kelas mayoritas serta minoritas lebih berimbang. Teknik eliminasi data yang diusulkan pada metode *undersampling* sangat bermacam-macam. Terdapat metode yang focus menghapus pada area kelas mayoritas saja (seperti pada gambar), ada pula yang berfokus pada kedua kelas namun hanya di area border saja, dan lain sebagainya.

Terdapat beberapa metode *undersampling* yang digunakan pada penelitian ini diantaranya *Neighborhood Cleaning Rule* (NCL) (Laurikkala, 2001), Tomek Link [32], dan *Random Undersampling* (RUS) (Prusa *et al.*, 2015).

2.4.1 *Neighborhood Cleaning Rule* (NCL)

J. Laurikkala menemukan salah satu metode *undersampling* untuk mengatasi distribusi kelas yang *imbalance* dengan mereduksi data berbasis *cleaning*. Salah satu kelebihan pada NCL yaitu NCL sangat mempertimbangkan kualitas dari data yang akan dihapus dengan tidak berfokus pada reduksi data saja melainkan berfokus pada pembersihan data (*cleaning data*). Proses *cleaning data* tersebut diperuntukkan tidak hanya untuk *sample* pada kelas mayoritas namun juga kelas minoritas.

Pada dasarnya, prinisp pada NCL didasarkan pada konsep *one-sided selection* (OSS) yang merupakan salah satu teknik untuk mereduksi data berbasis *instance* untuk mereduksi kelas. Tujuan utamanya yaitu untuk mengurangi data yang tidak relevan secara lebih hati-hati. Proses *cleaning* data pada NCL diterapkan pada *sample* mayoritas dan minoritas secara terpisah. Secara garis besar, NCL mengadopsi metode *Edited Nearest Neighbor* (ENN) untuk membersihkan data pada kelas mayoritas. Sebagai contoh terdapat sample E_1 pada training set, kemudian temukan ketiga tetangga terdekat dari masing-masing sample tersebut. Apabila E_1 termasuk sebagai kelas mayoritas dan hasil klasifikasi ternyata berlawanan dengan kelas original pada E_1, maka E_1 akan dihapus. Sebaliknya, apabila E_1 merupakan kelas minoritas dan ketiga tetangganya terklasifikasi berlawanan (mayoritas), maka tetangga terdekat akan dihapus. Cara kerja *Neighborhood Cleaning Rule* dipaparkan lebih lanjut pada Tabel 2.1.

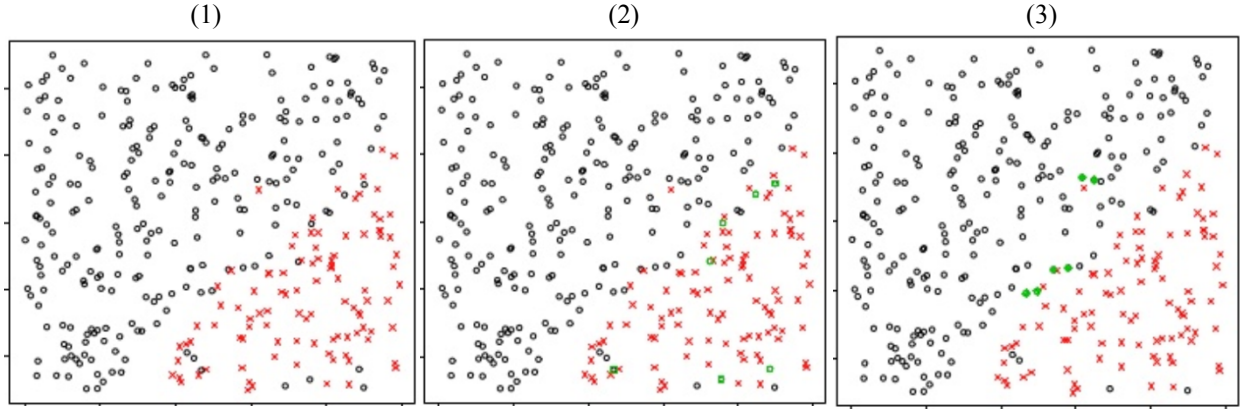
Diasumsikan terdapat suatu dataset *T* dimana *C* merupakan *class of interest* dengan jumlah data kecil dan O merupakan kelas mayoritas yang diperoleh dari pengurangan $O = T - C$. NCL menggunakan aturan *Edited Nearest Neighbor* (ENN) yang diusulkan oleh (More, 2016) untuk mengurangi O dengan menghapus data *noise* A_1 pada O. Selain itu, ENN menghapus data yang memiliki kelas berbeda dengan kelas mayoritasnya (*misclassify*). Kemudian pada metode NCL, proses pembersihan tersebut diimprovisasi dengan menghapus ketiga tetangga terdekat dari data pada *C* yang salah terklasifikasi dan masih merupakan bagian dari *O*. Ketiga tetangga terdekat yang dihapus tersebut dijadikan sebagai himpunan A_2.

Tabel 2.1 Algoritma *Neighborhood Cleaning Rule*

1.	Split data *T* menjadi *interest class C* dan data lain *O*
2.	Identifikasi data *noise* A_1 pada *O* dengan menggunakan *Edited Nearest Neighbor* (ENN)
3.	Pada setiap kelas C_i pada *O*
	Jika ($x \in Ci$)pada 3 tetangga terdekat dari misklasifikasi $y \in C$) dan ($\|Ci\| > 0.5\|C\|$) maka $A_2 = \{x\} \cup A_2$
4.	Data reduksi $S = T - (A_2 \cup A_2)$

Dalam versi asli dari algoritma NCL, hanya sampel dari kelas yang lebih besar atau sama dengan setengah ukuran interest class (0.5 • |C|) yang dipertimbangkan untuk A_2. Ide ini diusulkan demi menghindari pengurangan kelas minoritas secara berlebihan.

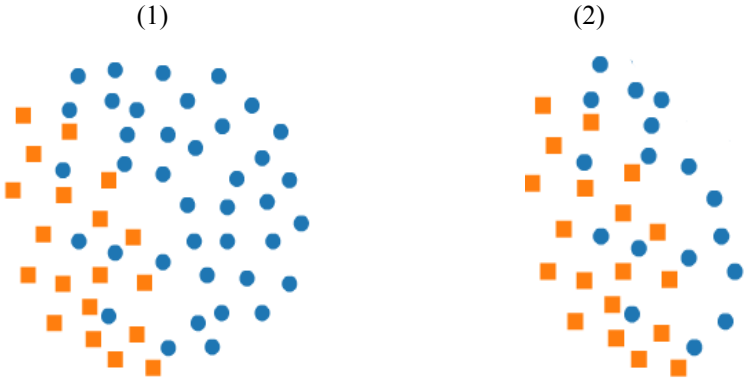
Gambar 2.4 merupakan ilustrasi dari proses ENN dan NCL dimana titik hitam merepresentasikan kelas mayoritas, titik merah merupakan kelas minoritas sedangkan titik tereliminasi disimbolkan dengan warna hijau.



Gambar 2.4 Ilustrasi data: (1) original, (2) setelah proses ENN, (3) setelah proses NCL

2.4.2 Tomek Link

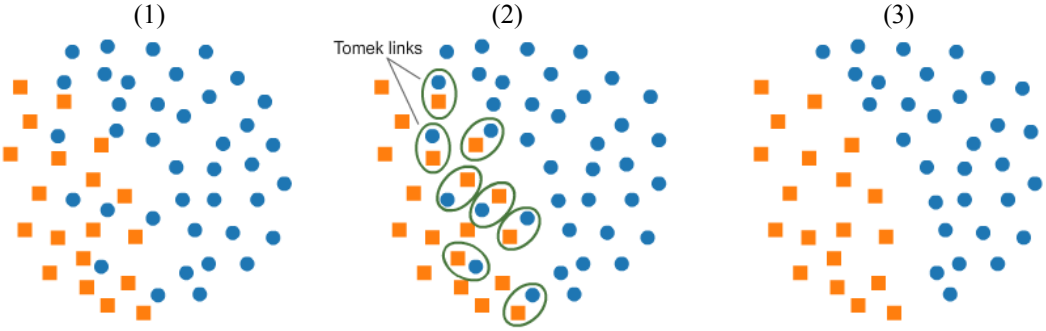
Tomek link merupakan metode *undersampling* hasil pengembangan dari metode Condensed Nearest Neighbor Rule (CNN) (Hart, 1967). Pada CNN, proses eliminasi ditujukan hanya untuk data-data pada kelas mayoritas yang letaknya jauh dari batas keputusan *(decision border)*. Hal ini menyebabkan CNN akan sangat sensitif terhadap *noise*. Padahal justru pada umumnya noise lah yang menyebabkan misklasifikasi sering terjadi. Gambar 2.5 merupakan ilustrasi dari metode CNN.



Gambar 2.5. Ilustrasi data: (1) original, dan (2) setelah proses CNN

Berbeda dengan CNN, tujuan utama dari Tomek link yaitu untuk menghapus noise dan borderline pada data yang dapat mempersulit proses klasifikasi pada data imbalanced. Tomek link didefinisikan sebagai pasangan dari data kelas minoritas dan kelas mayoritas sehingga tidak ada lagi data diantara kedua data Ei an Ej tersebut (Tomek, 1976) . Gambar 2.6 merupakan representasi dari metode Tomek Link.

Diketahui terdapat dua sample E_i dan E_j dengan kelas yang berbeda. Dimana $d(E_i, E_j)$ merupakan jarak antara E_i an E_j. Sepasang (E_i, E_j) disebut sebagai Tomek Link apabila tidak terdapat sample pada E_1, atau dengan kata lain $d(E_i,E_1)<d(E_i,E_j)$ atau $d(E_j,E_1)<d(E_i,E_j)$. Apabila dua bentuk sample membentuk Tomek link, maka salah satu dari sample tersebut merupakan noise atau keduanya merupakan *borderline*. Jadi, Tomek link dapat dikatakan sebagai metode *undersampling* ketika sample dari kedua kelas dihapus.



Gambar 2.6 Ilustrasi data: (1) original, (2) deteksi tomek link, (3) tomek link dihapus

2.4.3 Random Undersamping (RUS)

Random Undersampling merupakan salah satu teknik non-heuristik yang sederhana dalam mereduksi kelas mayoritas untuk mencapai keseimbangan pada distribusi kelas. Pada RUS, seluruh data *training* dari kelas minoritas digunakan. Sesuai dengan namanya, metode eliminasi pada kelas mayoritas dilakukan secara random hingga keseimbangan diperoleh. Namun, metode ini memiliki satu kelemahan yaitu kemungkinan hilangnya beberapa informasi penting pada kelas mayoritas. Akan tetapi, pada beberapa kasus dimana data pada kelas mayoritas berada berdekatan dengan kelas minor, RUS dapat mendistribusi kelas dengan hasil yang baik (More, 2016). Tabel 2.2 merepresentasikan cara kerja algoritma *Random Undersampling* secara umum.

Tabel 2.2 Algoritma *Random Undersampling* (RUS)

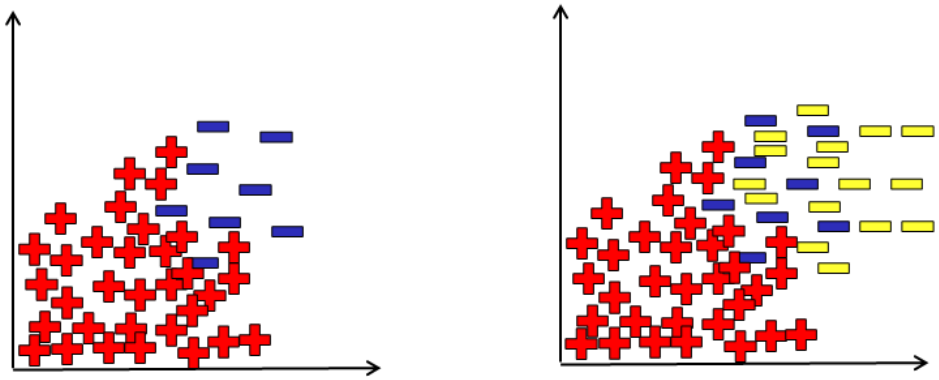
1.	Identifikasi kelas mayoritas dan minoritas

2.	Hitung jumlah data yang dihapus berdasarkan presentase *undersampling*
3.	Identifikasi data random pada kelas mayoritas dan hapus data tersebut dari kelas mayoritas
4.	Ulangi tahap 3 hingga data yang terhapus sebanyak presentase yang ditentukan

2.5. Oversampling

Berkebalikan dengan *undersampling*, *oversampling* justru merupakan metode *sampling* dengan menambahkan jumlah data pada kelas minoritas sehingga dapat mengimbangi atau mendekati jumlah data pada kelas mayoritas. Konsep penambahan data pada *oversampling* dibagi menjadi dua yaitu: *oversampling* menggunakan data asli, seperti metode *Random Oversampling* dan yang kedua yaitu metode penambahan menggunakan data sintetik seperti *Synthetic Minority Over-sampling Technique* (SMOTE) (Chawla *et al.*, 2002), SMOTE Borderline, Safe Level SMOTE, *Adaptive Semi-unsupervised Weighted Oversampling* (A-SUWO) (Nekooeimehr and Lai-yuen, 2016). Namun pada penelitian ini, penulis hanya berfokus pada *oversampling* menggunakan metode penambahan data secara sintetik dimana metode *state-of-the-art* yang dipilih ialah metode A-SUWO.

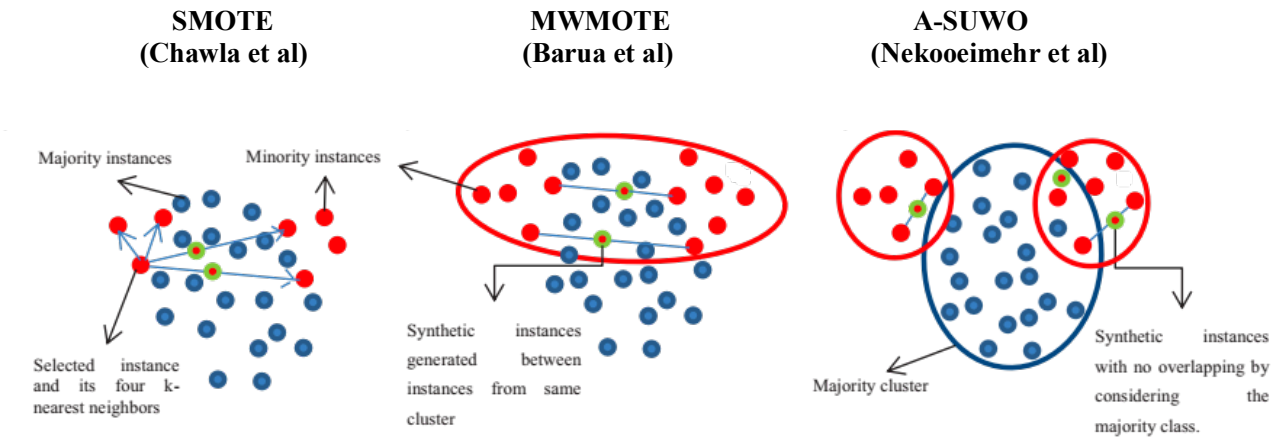
Permasalahan yang umum pada *oversampling* adalah terjadinya *overfitting* dikarenakan penambahan data secara berulang menyebabkan *decision boundary* menjadi lebih ketat. Oleh karena itu, pada perkembangannya, metode oversampling bukan lagi mengopi data yang sama tetapi membuat data baru yang mirip. Hal ini bertujuan untuk memperhalus *boundary decision* seperti terlihat pada Gambar 2.7.



Gambar 2.7 Ilustrasi keadaan data sebelum dan setelah diproses menggunakan *oversampling*

2.5.1 *Adaptive Semi-unsupervised Weighted Oversampling* (A-SUWO)

Metode A-SUWO merupakan pengembangan dari metode MWMOTE. Dalam hal ini, metode yang diusulkan pada A-SUWO mengelompokkan data minoritas menggunakan *semi-unsupervised hierarchical clustering* dan secara adaptif mampu menentukan size untuk melakukan *oversampling* pada setiap sub-klusternya (Nekooeimehr and Lai-yuen, 2016). Sedangkan, data minoritasnya akan dilakukan proses *oversample* berdasarkan jarak Euclidian. Tujuan dari A-SUWO antara lain metode ini dapat mengidentifikasi data yang sulit dipelajari dengan mempertimbangkan data minoritas pada setiap sub-cluster yang berada dekat pada *borderline*. Selain itu, A-SUWO juga mampu mencegah terjadinya overlap terhadap kelas mayoritas ketika menghasilkan data sintetis.



Gambar 2.8 Pembuatan data sintetik pada: SMOTE, MWMOTE, dan A-SUWO.

Dapat dilihat pada Gambar 2.8, pada SMOTE, data sintetik dibuat diantara 2 dari 4 tetangga terdekat. Sedangkan pada MWMOTE, data sintetik dibuat diantara data kandidat dan 1 tetangga terdekat yang berada pada sub-klaster yang sama. Kedua metode tersebut dapat menyebabkan terjadinya kemungkinan data sintetik yang overlap dengan data pad akelas mayoritas. Terlalu banyak overlap dapat menyebabkan menurunnya hasil performa pada proses klasifikasi (Barua et al, 2014; Beyan & Fisher, 2014). Dengan mempertimbangkan kelas mayoritas saat melakukan sub-clustering dan pembuatan data sintetik pada metode A-SUWO, permasalahan *overlap* dapat diminimalisir.

Metode A-SUWO terbagi menjadi tiga tahapan utama yaitu: (1) *Semi-unsupervised clustering*, (2) *Adaptive sub-cluster sizing*, (3) Pembuatan data sintetik (Nekooeimehr and Lai-yuen, 2016). Pada tahap pertama, data pada kelas minoritas

dikusterisasi menggunakan semi-unsupervised hierarchical clustering. Kemudian pada tahap kedua, pada tahap *Adaptive Sub-cluster Sizing*, ukuran setiap sub-klaster minoritas akan dilakukan *oversampling* berdasarkan kompleksitas dalam proses klasifikasi (*misclassification error*). Terdapat usulan perhitungan berdasarkan average error rate untuk menentukan kompleksitas sub-klaster tersebut dan akan dihitung menggunakan *cross validation*. Pada tahap akhir, data sintetik akan dihasilkan menggunakan system pembobotan untuk memberikan bobot pada data minoritas berdasarkan jarak *Euclidian* pada *NN-nearest majority class neighbor*. Cara kerja A-SUWO lebih lanjut dipaparkan pada Tabel 2.3 berikut:

Tabel 2.3. Algoritma A-SUWO

Input: *I* : dataset original c_{thres} : *threshold* untuk *clustering* *NN* : jumlah tetangga terdekat pada setiap data minoritas untuk menentukan bobot *NS* : jumlah tetangga terdekat untuk menentukan data *noise* *K* : jumlah fold dalam *K-fold Cross Validation*	
Output: *O* = dataset setelah dilakukan *oversampling*	
Fase 1: *Semi-unsupervised clustering*	
1.	Tentukan T menggunakan persamaan: $$T = d_{avg} * c_{thresh} \quad (2.1)$$
2.	Lakukan clustering kelas mayoritas, yang akan menghasilkan *m* sub-cluster $Cmaj_{i=1, \ldots, m}$
3.	Tentukan setiap data minoritas menjadi sub-cluster terpisah
4.	Temukan 2 sub-cluster terdekat $Cmin_a$ dan $Cmin_b$
5.	Cek apakah terdapat overlap pada sub-cluster mayoritas antara $Cmin_a$ dan $Cmin_b$
6.	if overlap set jarak menjadi *infinity* dan kembali pada step 5 else merge $Cmin_a$ dan $Cmin_b$ menjadi sub-cluster $Cmin_c$
7.	Ulangi step 5-7 hingga jarak Euclidian antar *sub-cluster* terdekat < *threshold* T
Fase 2: Adaptive sub-cluster sizing	
8.	Secara random *split* setiap *sub-cluster* minoritas menjadi *K* folds
9.	Buat model menggunakan K-1 folds dari setiap *sub-cluster* minoritas dan semua data mayoritas sebagai training set
10.	Tes model menggunakan 1 fold lainnya pada setiap sub-cluster minoritas
11.	Tentukan Standardized Average Minority Error Rate

12.	Ulangi step 2-4 sebanyak K perulangan
13.	Tentukan ukuran akhir dari *Sj* untuk seluruh $Cmin_{j=1,\ldots,n}$ menggunakan persamaan: $$\varepsilon_j^* = \frac{\xi_j^*}{\sum_{j=1}^{n} \xi_j^*} \quad (2.2)$$ $$\frac{S_{L1}}{S_{L2}} = \frac{\varepsilon_{L1}^*}{\varepsilon_{L2}^*} \; \forall \, L1, L2 \in \{1, \ldots, n\} \quad (2.3)$$

Fase 3: *Synthetic instance generation*

Menentukan distribusi probabilitas untuk data dalam setiap sub-kluster minoritas

Pada setiap sub-cluster j=1,2,…n:

14.	Untuk seluruh data minoritas x_{jh} pada sub cluster $Cmin_j$, tentukan NN tetangga terdekat antar data mayoritas
15.	Tentukan $W(x_{jh})$ pada setiap data minoritas $Cmin_j$ menggunakan persamaan berikut dan dengan estimasi TH_j $$W(X_{jh}) = \sum_{v=1}^{k} \Gamma\,(X_{jh}, W_{jh_{(v)}}) \quad (2.4)$$
16.	Transform bobot dengan distribusi probabilitas $P(x_{jh})$ $$P(X_{jh}) = \frac{W(X_{jh})}{\sum_{h=1}^{Rj} W(X_{jh})} \quad (2.5)$$

Oversample Data Minoritas

Inisialisasi *O=I*

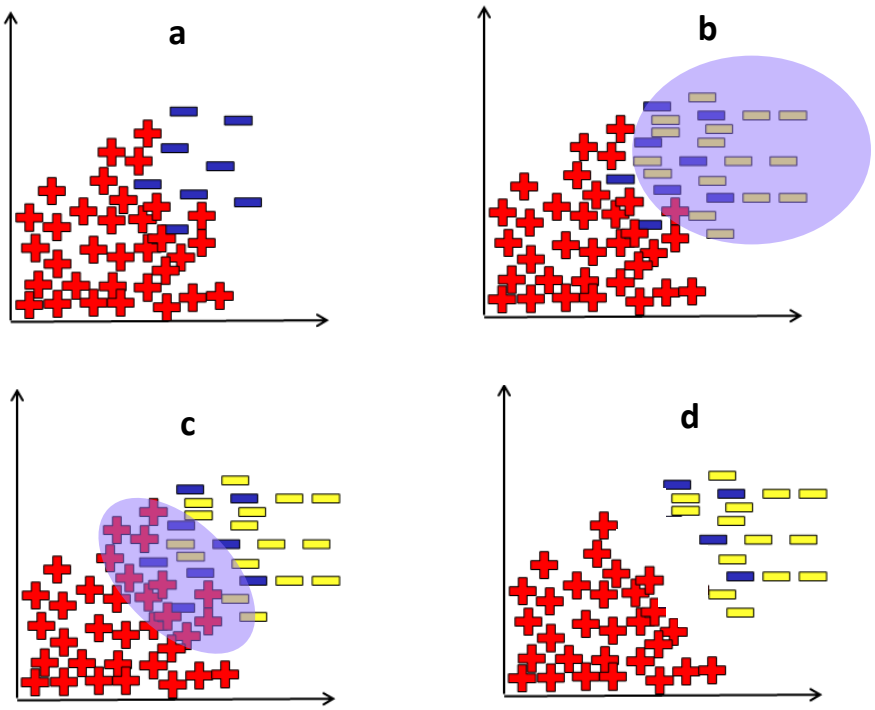
Untuk setiap sub-klaster j=1,2, …, n

17.	Tentukan data minoritas a oada sub klister j dengan sampling dari distribusi probabilitas $P(x_{jh})$
18.	Tentukan secara random 2 dari NN tetangga terdekat b yang berada pada sub-klaster yang sama
19.	Buat data sintetik baru *c* diantara *a* dan *b* menggunakan persamaan: $$\boldsymbol{c = \beta a + (1 - \beta) b} \quad (2.6)$$
20.	Tambahkan *c* pada set *O*
21.	Ulangi tahap 1-4 hingga sub-klaster meraih size *Sj*

2.6. Metode Hybrid Oversampling dan Undersampling

Metode hybrid merupakan metode *balancing* dengan menggabungkan metode *oversampling* dan *undersampling*. Jumlah data pada kelas minoritas ditambahkan dengan metode *oversampling*, begitu juga dengan jumlah data pada kelas mayoritas dikurangi atau dibersihkan dari data noise menggunakan konsep *undersampling*. Beberapa contoh penelitian sebelumnya yang menggunakan metode hybrid diantaranya: SMOTE+Tomek [33], SMOTE+ENN [33], SMOTE-RSB [16], dan SMOTE+NCL [32].

Gustavo, *et al* dalam penelitiannya [33] mengusulkan suatu pengembangan metode SMOTE dengan *undersampling*. Hal yang menjadi dasar adalah meskipun proses *oversampling* pad akelas minoritas dapat menyeimbangkan distribusi kelas, beberapa masalah lainnya biasanya muncul pada dataset dengan kemiringan kelas distribusi yang sulit untuk dipecahkan. Seringkali, kelompok kelas tidak didefinisikan dengan baik karena beberapa contoh kelas mayoritas mungkin menyerang (mendominasi) ruang pada kelas minoritas. Sebaliknya juga bisa benar, karena interpolasi data pada kelas minoritas dapat memperluas kluster kelas minoritas, menambahkan data sintetik pada kelas minoritas yang berlebihan dapat mengganggu ruang kelas mayoritas. Situasi seperti itu dapat menyebabkan overfitting. Untuk membuat kluster kelas yang lebih jelas, Gustavo mengusulkan penerapan *Tomek-link* pada *training set* sebagai data metode pembersihan. Jadi, daripada hanya menghapus contoh kelas mayoritas yang membentuk Tomek-link, contoh dari kedua kelas dihapus. Aplikasi dari metode ini diilustrasikan pada Gambar 2.9 dimana terdapat data set original (a), kemudian dilakukan over-sample dengan Smote (b), lalu Tomek-Link diidentifikasi (c) dan dihapus, sehingga menghasilkan set data yang seimbang dengan kelompok kelas yang terdefinisi dengan baik (d).



Gambar 2.9. Ilustrasi Smote+Tomek-link

2.7. Klasifikasi

Proses mempelajari suatu obyek untuk mengelompokkan obyek tersebut menjadi beberapa kategori atau kelas tertentu merupakan salah satu karakterisitik dari kecerdasan yang sangat popular di dunia riset utamanya *computer science* (Mahmood, 2017). Kemampuan untuk menghasilkan klasifikasi yang akurat akan memudahkan manusia dalam mengambil keputusan. Semakin baik performa suatu sistem untuk mengklasifikasikan data, akan semakin efisien pula proses pengambilan keputusan.

Pada *machine learning*, tahap klasifikasi secara umum terjadi secara *supervised,* dimana biasanya kelas atau kategori dari suatu obyek sudah didefinisikan secara spesifik sejak awal. Pada *supervised learning*, algoritma seolah-olah dilatih terlebih dahulu agar mampu melakukan prediksi maupun klasifikasi. Beberapa contoh algoritma *supervised learning* yaitu *Decision Tree, Random Forest, Naïve Bayes Classifier, Nearest Neighbor Classifier, Artificial Neural Network (ANN) dan Support Vector Machine (SVM).* Sedangkan pada *unsupervised learning*, untuk melakukan prediksi maupun klasifikasi mereka tidak perlu dilatih terlebih dahulu. Berdasarkan model matematisnya, algoritma ini tidak memiliki target variabel. Salah satu tujuan dari algoritma ini adalah mengelompokkan objek yang hampir sama dalam suatu area tertentu. Algoritma yang banyak digunakan untuk kasus unsupervised yaitu K-means, *Hierarchical Clustering*, DBSCAN, dan Fuzzy C-Means.

Namun, pada penelitian ini, penulis menggunakan salah satu metode klasifikasi secara supervised dalam mengolah data akademik. Hal ini diadopsi dari beberapa penelitian sebelumnya yaitu (Fahrudin and Fatichah, 2016)(Cahyo and Lianto, 2018) (Barua *et al.*, 2014).

2.7.1. *Decision Tree Classifier*

Sesuai dengan namanya, algoritma decision tree diadopsi dari bentuk pohon yang secara teknis pohon yang dimaksud merupakan sekumpulan dari beberapa simpul (nodes) dan cabang (branches). Setiap simpul merepresentasikan atribut sedangkan cabang merepresentasikan nilai dari masing-masing atribut [Mahmood]. Node teratas merupakan akar atau root merepresentasikan atribut yang paling berpengaruh terhadap suatu kelas tertentu. Strategi pencarian yang digunakan yaitu secara top-down dengan melacak jalur dari node akar (root) hingga node terbawah (leaves). Manfaat utama dari

penggunaan decision tree adalah kemampuannya untuk mem-break down proses pengambilan keputusan yang kompleks menjadi lebih sederhana sehingga pengambil keputusan akan lebih menginterpretasikan solusi dari permasalahan. Decision tree juga berguna untuk mengeksplorasi data, menemukan hubungan tersembunyi antara sejumlah calon variabel input dengan sebuah variabel target.

Beberapa kelebihan dari algoritma decision tree adalah:

- Ranah pengambilan keputusan yang sebelumnya kompleks dan sangat global, dapat diubah menjadi lebih simpel dan spesifik.
- Eliminasi perhitungan-perhitungan yang tidak diperlukan, karena ketika menggunakan metode pohon keputusan maka sample diuji hanya berdasarkan kriteria atau kelas tertentu.
- Fleksibel untuk memilih fitur dari internal node yang berbeda, fitur yang terpilih akan membedakan suatu kriteria dibandingkan kriteria yang lain dalam node yang sama.
- Dalam analisis multivariat, dengan kriteria dan kelas yang jumlahnya sangat banyak, seorang penguji biasanya perlu untuk mengestimasikan baik itu distribusi dimensi tinggi ataupun parameter tertentu dari distribusi kelas tersebut. Metode pohon keputusan dapat menghindari munculnya permasalahan ini dengan menggunakan kriteria yang jumlahnya lebih sedikit pada setiap node internal tanpa banyak mengurangi kualitas keputusan yang dihasilkan.

Pada penelitian ini, algoritma decision tree yang digunakan adalah algoritma C4.5 yang ditemukan oleh (Ross, Morgan and Publishers, 1994). Algoritma C4.5 merupakan algoritma untuk memperbaiki algoritma ID3 yang memiliki kelemahan dalam menangani atribut kontinu maupun diskrit, *missing value*, dan *pruning*. Tabel 2.4 merupakan gambaran umum proses C4.5.

Tabel 2.4. Algoritma *Decision Tree C4.5*

Input: atribut pada Dataset D	
1.	Tree {}
2.	if D memenuhi stop criteria then terminate
3.	end if
4.	for all atribut a $\in$ D do

5.	Hitung informasi teoritis kriteria jika split pada a
6.	end for
7.	a_{best} = atribut terbaik berdasarkan kriteria yang telah dihitung
8.	Tree = node keputusan yang menguji a_{best} pada root
9.	D_v = subdataset dari D berdasarkan a_{best}
10.	for all
11.	$Tree_v$ = C4.5 (D_v)
12.	Attach $Tree_v$ pada branch Tree
13.	end for
14.	Return Tree

2.8. Evaluasi Kinerja

Pengukuran kinerja yang dibuat menggunakan basis konsep dari *confusion* matrix seperti pada Tabel 2.5. Dimana **kelas (+)** merepresentasikan kelas mahasiswa yang memiliki **kemungkinan gagal** dalam akademik. Mahasiswa tersebut membutuhkan waktu lulus tingkat 1 dalam waktu lebih dari 1 tahun dan kurang sama dengan 2 tahun (4 semester) atau mahasiswa yang tidak dapat menyelesaikan beban tingkat 1 dalam 2 tahun tersebut. Sedangkan **kelas (-)** merupakan kelas mahasiswa yang berhasil atau **tidak memiliki kemungkinan gagal dalam akademik**. Atau dengan kata lain mahasiswa mahasiswa tersebut **mampu** lulus tingkat 1 dalam waktu 1 tahun (2 semester).

Tabel 2.5 *Confussion Matrix*

	Prediksi (+)	**Prediksi (-)**
Aktual (+)	TP	FN
Aktual (-)	FP	TN

Evaluasi kinerja model terhadap 2 kelas yang berbeda diberikan dalam bentuk Akurasi, *Recall*, *Precision*, *F-Measure* untuk masing-masing kelas. Formulasi masing masing ukuran kinerja dapat dilihat dalam bentuk **Persamaan 7- Persamaan 1**0.

Precision merupakan perhitungan ketepatan klasifikasi pada jumlah data berlabel positif atau data kelas minoritas yang memang benar secara actual merupakan kelas positif. Sedangkan, *Recall* (disebut juga *sensitivity*) adalah perhitungan ketepatan

klasifier pada jumlah data positive yang teridentifikasi benar sebagai kelas positif. *F-measure* sendiri adalah *relative importance* antar *precision* dan *recall*.

$$Precision = \frac{TP}{TP+FP} \tag{2.7}$$

$$Recall = \frac{TP}{TP+FN} \tag{2.8}$$

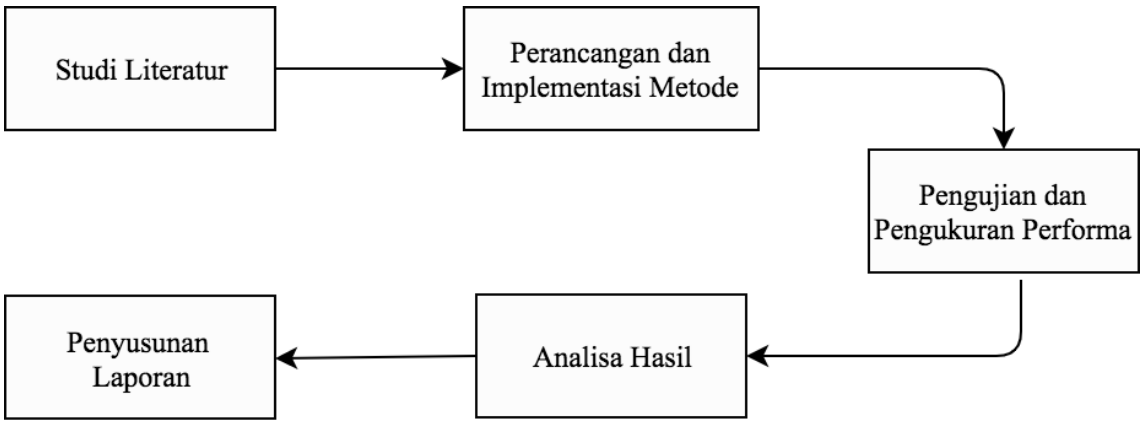
$$F - measure = \frac{(1+\beta^2)\,.\,Recall.\,Precision}{\beta^2\,.\,Recall\,.\,Precision} \tag{2.9}$$

$$Akurasi = \frac{TP+TN}{TP+TN+FP+FN} \tag{2.10}$$

BAB 3
METODE PENELITIAN

Bab ini akan memaparkan terkait metodologi penelitian yang digunakan pada penelitian ini, data input, rancangan metode, scenario uji coba, evaluasi, dan jadwal penelitian yang digunakan.

Tahapan-tahapan yang dilakukan pada penelitian ini antara lain terdiri dari studi literatur, perancangan dan implementasi metode, pengujian, pengukuran performa, analisa hasil, dan dokumentasi berupa penyusunan laporan. Ilustrasi alur metodologi penelitian lebih lanjut dapat dilihat pada Gambar 3.1.



Gambar 3.1 Diagram Alur Penelitian

3.1. Studi Literatur

Pada studi literatur, tahap yang harus dilakukan terlebih dahulu adalah tahap pengkajian yang berkaitan dengan topik pada penelitian ini. Pengkajian dilakukan berdasarkan referensi yang diperoleh dari jurnal maupun publikasi seminar bereputasi yang terkait dengan metode *undersampling* (*Random Undersampling*, Tomek Link, dan *Neighborhood Cleaning Rule*), metode *oversampling (ASUWO)*, dan metode klasifikasi serta metode pengujian performa. Dari studi literatur yang telah dilakukan, diperoleh informasi yang berhubungan dengan penelitian, diantaranya sebagai berikut:

1. Data akademik yang digunakan pada penelitian ini menyimpan informasi yang dapat berguna untuk memajukan Universitas XYZ berdasarkan histori akademis mahasiswa.

2. Data akademik memiliki rasio keberhasilan studi yang tidak berimbang jika dibandingkan dengan kegagalan studi (*imbalanced*).
3. Fenomena *imbalance* pada data akademik ini dapat ditangani dengan beberapa teknik *sampling*: *undersampling oversampling*, dan metode *hybrid* dari *oversampling* maupun *undersampling*.
4. Terdapat beberapa teknik *undersampling* yang digunakan pada penelitian ini diantarnya: *Random Undersampling* (RUS), Tomek Link, dan *Neighborhood Cleaning Rule* (NCL).
5. Metode *oversampling* untuk menyeimbangkan kelas minoritas dengan menambahkan data sintetis sehingga diharapkan rasio data akademik dapat lebih berimbang. Metode *oversampling* yang digunakan pada penelitian ini yaitu metode A-SUWO.
6. Tujuan metode *undersampling* yaitu untuk mengurangi atau mengeliminasi sebagian data yang dianggap kurang relevan pada kelas mayoritas. Selain itu, data akademik juga memiliki tingkat overlap yang cukup tinggi, sehingga dengan proses eliminasi data pada metode *undersampling* diharapkan dapat meminimalisir permasalahan tersebut.
7. Penggabungan kedua teknik *undersampling* dengan *oversampling* sering disebut dengan metode *hybrid* dan akan terdapat 3 pasang metode gabungan yang harapannya dapat saling memperbaiki kekurangan dari masing-masing metode dan dapat menghasilkan performa yang lebih baik.

3.1.1 Perancangan dan Implementasi Metode

Pada tahap perancangan sistem, akan dipaparkan mengenai *step by step* mulai dari data masukan, serta penggambaran alur proses yang terjadi dalam metode untuk menghasilkan data keluaran. Penentuan desain model sistem ini akan memberikan gambaran mengenai apa dan bagaimana suatu penelitian dilaksanakan.

Tahap implementasi metode bertujuan untuk mengimplementasikan rancangan metodeyang diusulkan. Pada penelitian ini rancangan metode yang diusulkan akan diimplementasikan menggunakan aplikasi Matlab dan Bahasa pemrograman python.

3.1.2 Pengujian

Pada tahapan ini dilakukan pengujian dan analisis terhadap hasil dan performa metode yang diusulkan dalam melakukan penelitian dengan topik penggabungan *oversampling* dan *undersampling* untuk data akademik. Gabungan metode pada penelitian ini terdapat 3 pasang metode gabungan.

Pada tahapan ini juga dilakukan pengujian performa terhadap hasil dan performa metode yang diusulkan dalam melakukan proses *sampling* (*hybrid oversampling* dan *undersampling*).

3.1.3 Analisis Hasil

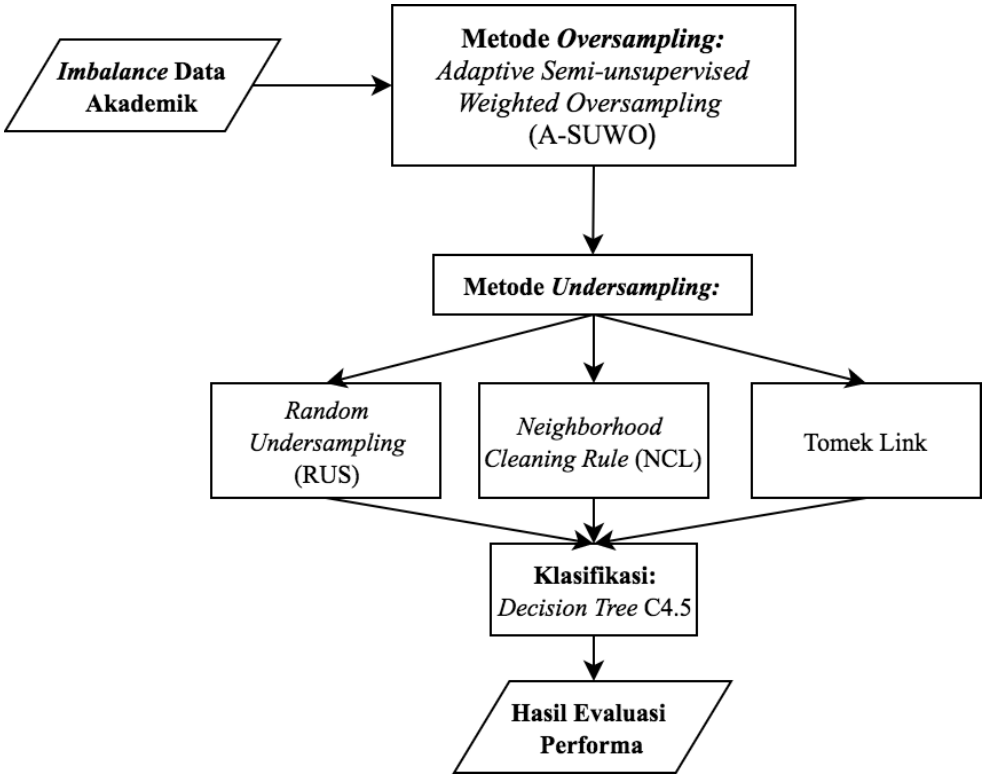
Setelah mendapatkan hasil performa dari metode yang diusulkan, tahap selanjutnya yaitu menganalisis hasil performa tersebut.

3.1.4 Penyusunan Laporan

Pada tahap ini dilakukan penyusunan laporan terhadap penelitian yang telah dilakukan, mulai dari studi literatur, analisis masalah dan desain, hasil implementasi, hingga uji coba dan analisis.

3.2. Perancangan

Penelitian dilakukan dengan menggunakan data akademik mahasiswa angkatan 2014 dan 2015 dalam 1 semester. Tujuan penelitian ini adalah untuk melakukan teknik *sampling* dengan menggabungkan metode *oversampling* dan *undersampling* untuk memperoleh rasio data yang lebih berimbang. Pada teknik *undersampling*, metode yang digunakan yaitu *Random Undersampling* (RUS), *Neighborhood Cleaning Rule* (NCL), dan Tomek Link sedangkan metode *oversampling* yaitu *Adaptive Semi-unsupervised Weighted Oversampling* (A-SUWO). Dari metode-metode tersebut maka akan diperoleh 3 pasang gabungan metode yang akan diuji cobakan kemudian performa dari masing-masing gabungan metode dievaluasi. Gabungan metode dengan performa terbaik akan dipilih sebagai metode gabungan yang paling optimal. Sebelum pengujian performa, diperlukan proses klasifikasi terlebih dahulu menggunakan metode *Decision Tree* C4.5. Adapun rancangan metode yang diusulkan ditampilkan pada Gambar 3.2.



Gambar 3.2 Alur Usulan Metode Penelitian

3.2.1. Data Masukan

Seperti dijelaskan pada sub bab 2.2, data akademik terdiri dari data awal, pertengahan dan akhir semester. Data akademik tersebut merupakan akumulasi data akademik pada bulan Agustus, September, Oktober, November, dan Desember tahun ajaran 2014/2015. Data awal semester terdiri dari data pada bulan Agustus dan September. Data tengah semester merupakan data bulan awal ditambahkan dengan data bulan Oktober dan November. Sedangan Data akhir semester merupakan akumulasi keseluruhan bulan termasuk bulan Desember.

Tabel 3.1 Deskripsi data akademik angkatan 2014 dan 2015 Universitas XYZ

Dataset	Atribut	Jumlah Data	Mayoritas	Minoritas	Persentase Mayoritas: Minoritas	*Imbalanced* ratio Mayoritas:Minoritas
Agustus	34	12854	9482	3372	74%:26%:	0.74: 0.26
September						
Oktober	37					
November						
Desember	42					

Data akademik memiliki dua label yaitu lulus (kelas minoritas) dan tidak lulus (kelas mayoritas). Tabel 3.1 merupakan deskripsi rinci terkait jumlah atribut, jumlah baris keseluruhan, jumlah baris pada kelas mayoritas maupun minoritas serta rasio *imbalance* data masing-masing bulan.

Tabel 3.2 berikut merupakan contoh dataset berupa data akademik Universitas XYZ yang digunakan dalam penelitian ini. Dimana masing-masing data direpresentasikan berdasarkan ID mahasiswa dan memiliki label kelas lulus maupun tidak lulus. Pada mahasiswa lulus disimbolkan menggunakan 0 sebagai kelas mayoritas dan mahasiswa yang diperkiraan tidak dapat lulus tepat waktu disimbolkan dengan 1 sebagai kelas minoritas.

Tabel 3.2 Contoh Data Akademik Universitas XYZ Angkatan 2014 dan 2015

ID Mahasiswa	Absensi	Aktivitas Internet	Aktifitas Blog	Aktivitas Lomba & Panitia	Nilai UTS	Nilai Akhir	Label	Simbol
1	94.41	0.08	0	0	100	100	Lulus	0
2	11.48	0.18	0	0	77.70	61.11	Tidak Lulus	1
3	97.9	0.36	0	0	100	100	Lulus	0
...	…	…	…	…	…	…	…	
…	…	…	…	…	…	…	…	

Selain data akademik, ujicoba juga dilakukan pada 8 jenis Dataset Keel diantaranya data Ecoli1, Glass0, Haberman, Iris0, Segment0, Vehicle0, dan Wisconsin. Masing-masing dataset tersebut memiliki jumlah data, jumlah atribut maupun rasio ketidakseimbangan yang berbeda-beda. Range jumlah atribut berada diantara 3-19, sedangkan pada jumlah data yaitu 150-2308 serta rasio ketidakseimbangan berada diantara 0.1- 3.36. Hal tersebut masih cukup relevan apabila kita bandingkan dengan data akademik yang memiliki rasio ketidakseimbangan sebesar 1:3.

Tabel 3.3. Deskripsi 7 Dataset Keel

Dataset	Atribut	Jumlah Data	*Imbalanced* Ratio Mayor:Minor
Ecolli1	7	336	0.77 : 0.23
Glass0	9	214	0.67 : 0.33
Haberman	3	306	0.74 : 0.26
Iris0	4	150	0.67 : 0.33
Segment0	19	2308	0.86 : 014

Vehicle0	18	846	0.77 : 0.23
Wisconsin	9	683	0.65 : 0.35

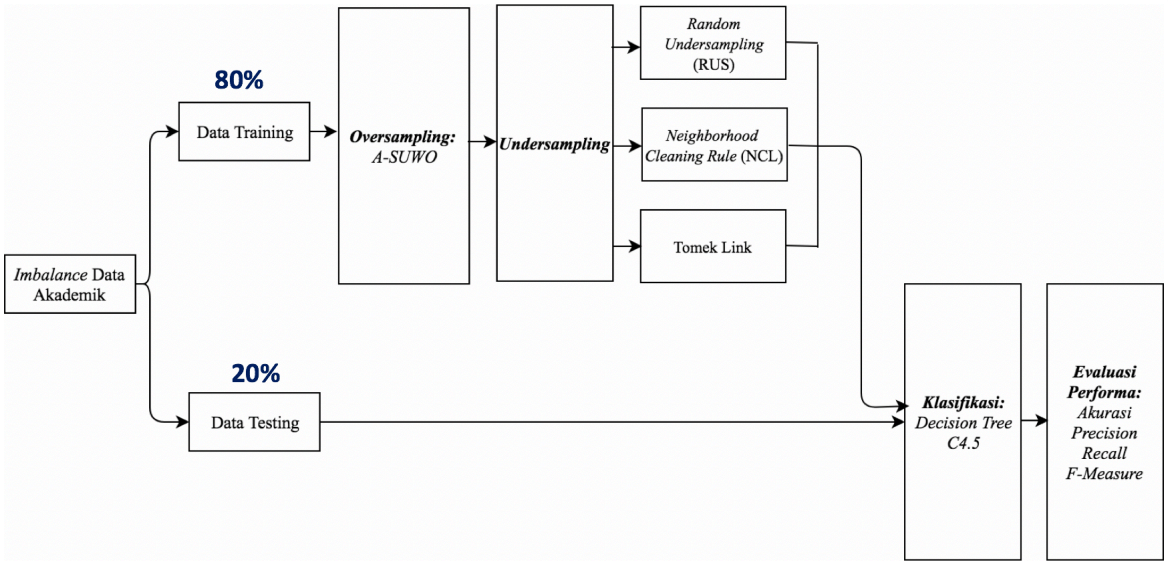
3.2.1. Pra-proses Data Imbalanced

Pra-proses data *imbalanced* bertujuan mempersiapkan data akademik untuk diproses. Proses tersebut terdiri dari pemisahan data mayoritas dan data minoritas berdasarkan label kelas dataset. Pra-proses data yang berlabel lulus disimbolkan dengan angka nol (0), sedangkan angka satu (1) menunjukkan simbol kelas minoritas berlabel tidak lulus.

3.3. Skenario Uji Coba dan Pengujian

Pada tahap pengujian, dilakukan pemisahan terlebih dahulu antara data training dan data testing berdasarkan **presentase 80%:20**%. Kemudian data training diuji menggunakan metode usulan yaitu gabungan metode A-SUWO dan metode RUS, NCL, maupun Tomek Link. Setelah diperoleh data hasil latih tersebut, kemudian dilakukan proses klasifikasi menggunakan metode Decision Tree C4.5 bersamaan dengan data testing untuk proses pengujian. Hasil evaluasi performa akan dihitung berdasarkan nilai **akurasi, *precision*, *recall*** maupun ***F-measure***. Ilustrasi skenario uji coba dapat dilihat pada Gambar 3.3.

Selain evaluasi performa dari masing-masing metode menggunakan perhitungan tersebut, penulis juga melakukan ujicoba parameter pada metode A-SUWO demi memperoleh parameter-parameter yang paling optimal pada masing-masing data (baik pada data akademik maupun pada data keel). Terdapat beberapa parameter yang diujicoba diantaranya nilai K-Fold, nilai Cthresh maupun nilai NN (ketetanggaan). Tabel 3.4 merupakan deskripsi singkat terkait parameter-parameter yang diujicobakan pada metode A-SUWO berdasarkan paper acuan. Pada paper, parameter Ctresh dirasa optimal pada nilai 0.7-2, sedangkan parameter NN 3-7, dan parameter k-fold antara 2-5. Untuk parameter iterasi, pada paper dilakukan percobaan sebanyak 3 kali iterasi.



Gambar 3.3 Diagram Skenario Ujicoba

Uji parameter tersebut dilakukan untuk parameter K-fold, Cthresh dan NN. Kemudian dihitung rata-rata akurasi dari keseluruhan data. Sedangkan parameter iterasi dilakukan dengan tujuan memastikan bahwa dalam setiap iterasi tidak akan terdapat perbedaan nilai akurasi yang perbedaannya terlampau jauh. Perulangan dilakukan sebanyak 5 kali dan 5 iterasi hasil A-SUWO ini akan diproses selanjutnya menggunakan metode undersampling.

Tabel 3.4. Parameter pada Metode A-SUWO

Parameter	Deskripsi Parameter	Nilai optimal pada Paper	Nilai range ujicoba
K-fold	Parameter untuk menentukan fold saat proses cluster sizing	2,3,4,5	Dataset Akademik: 3 dan 5 Dataset Keel: 3 dan 5
Chtresh	Threshold pada *agglomerative hierarchical clustering*	0.7;1;1.5;2	Dataset Akademik: 0,3; 1;1,5 Dataset Keel: 0,3; 1;1,5
NN	Ketetanggaan saat menentukan bobot pada setiap data minoritas	3;4;5;6;7	Dataset Akademik: 3;5;7 Dataset Keel: 3;5;7
NumIteration	Jumlah iterasi yang diinputkan user untuk memastikan bahwa hasil akurasi setiap iterasi tidak berbeda signifikan	3	Dataset Akademik: 5 Dataset Keel: 5

[Halaman ini sengaja dikosongkan]

BAB 4
HASIL PENELITIAN DAN PEMBAHASAN

Bab ini berisi penjelasan mengenai uji coba dan evaluasi penelitian yang telah dilakukan. Bab ini akan dibagi menjadi tiga bagian, yaitu implementasi penelitian, skenario uji coba dan analisis uji coba.

4.1. Implementasi

Metode yang diusulkan pada penelitian ini diimplementasi dengan menggunakan Bahasa pemrograman python 2.7 menggunakan Spyder serta MATLAB 2018a. Spesifikasi perangkat keras terdiri dari system operasi Windows 10 Enterprise 64-bit, RAM 16 GB, Processor Intel(R) Core(TM) i5 – 7400 CPU 3.00GHz.

4.2. Dataset Pengujian

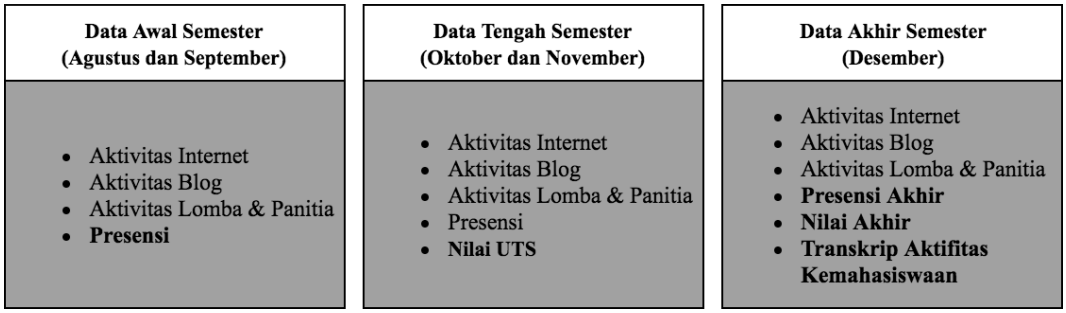
Uji coba dilakukan pada data akademik angkatan 2014 dan 2015 Universitas XYZ sebanyak 12854 baris data dan 8 dataset KEEL (**Tabel 4.1 dan 4.2**). Data tersebut merupakan akumulasi data akademik yang bersifat *imbalanced* pada bulan Agustus, September, Oktober, November, dan Desember. *Oversampling* dilakukan guna menangani ketidakseimbangan pada dataset. Tahap pra-proses terlebih dahulu dilakukan sebelum *oversampling* dengan tujuan pemisahan data menjadi data *training* dan data *testing* dengan perbandingan persentase 80%:20%.

Tabel 4.1. Deskripsi Data Akademik angkatan 2014 dan 2015 Universitas XYZ

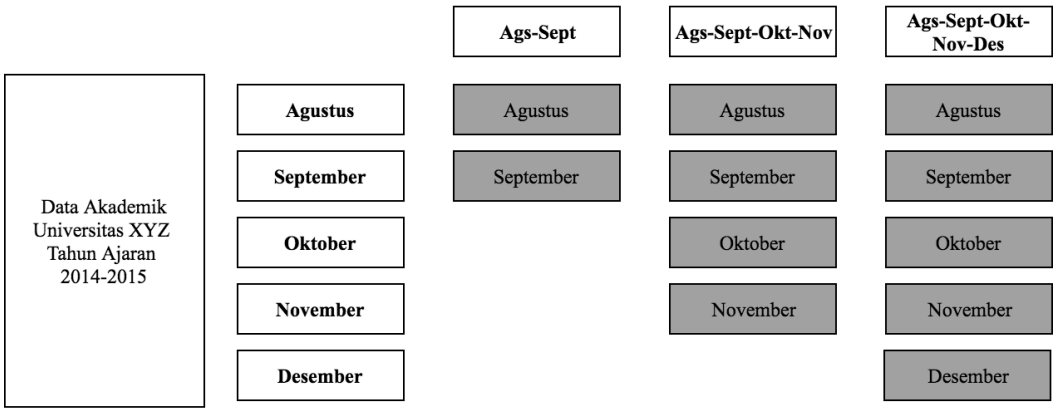
Dataset	Atribut	Jumlah Data	Mayoritas	Minoritas	Persentase Mayoritas:Minoritas	*Imbalanced* ratio Mayoritas:Minoritas
Agustus	34	12854	9482	3372	74%:26%:	0.74: 0.26
September						
Oktober	37					
November						
Desember	42					

Data *training* digunakan untuk pembuatan model terhadap data *testing*. Pada data akademik, kelas terdiri dari dua label yaitu lulus dan tidak lulus, dalam tahap pra-proses label akan dirubah ke dalam angka 0 (lulus) dan 1(tidak lulus) (**Lampiran 1**).

Dataset pengujian disusun ke dalam bentuk akumulasi bulanan (**Gambar 4.1**). Proses pencarian pola mahasiswa bermasalah dengan menggunakan klasifikasi data mining dilakukan terhadap data bulanan yang disusun terakumulasi dari satu bulan ke bulan-bulan berikutnya. Adanya perbedaan atribut bulan Agustus sampai dengan Desember dikarenakan ada pembagian data akademik. Pembagian 3 data berdasarkan data awal semester, data tengah semester dan data akhir semester (**Gambar 4.1** dan **Lampiran 1**). Skenario tengah semester melengkapi skenario data akumulasi bulanan. Hasil nilai UTS dimasukkan ke dalam skenario bulanan akumulasi pada bulan Oktober, November, dan Desember (**Gambar 4.2)**

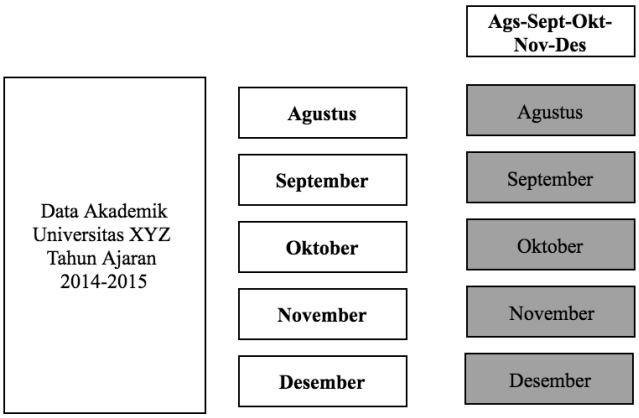


Gambar 4.1 Atribut Data Akademik



Gambar 4.2 Akumulasi Bulan Tengah Semester

Skenario akhir semester melengkapi skenario bulanan dan tengah semester pada bulan Desember yang memiliki tambahan atribut berupa nilai akhir dalam bentuk persentase Satuan Kredit Semester (SKS) lulus di akhir semester. Selain itu, digunakan pula persentase nilai di atas IPK 2, dan persentase nilai di bawah IPK 2 (**Gambar 4.3**)



Gambar 4.3 Akumulasi Bulan Akhir Semester

Sedangkan untuk data ujicoba yang lain yaitu menggunakan 7 data yang terpilih dari Dataset Keel. Dimana masing-masing data memiliki beberapa atribut dan jumlah data serta rasio *imbalance* yang berbeda-beda. Tabel 4.2 merupakan deskripsi Dataset Keel. Kelas pada dataset Keel juga terdiri dari label 0 dan 1 yang menandakan bahwa kelas 0 merupakan kelas yang teridentifikasi negatif dan kelas berlabel 1 merupakan data yang kelasnya teridentifikasi positif.

Tabel 4.2. Deskripsi Data Keel

Dataset	**Atribut**	**Jumlah Data**	***Imbalanced* Ratio Mayor:Minor**
Ecolli1	7	336	0.77 : 0.23
Glass0	9	214	0.67 : 0.33
Haberman	3	306	0.74 : 0.26
Iris0	4	150	0.67 : 0.33
Segment0	19	2308	0.86 : 014
Vehicle0	18	846	0.77 : 0.23
Wisconsin	9	683	0.65 : 0.35

4.3. Evaluasi Rasio Setelah *Oversampling* dan *Undersampling*

Berdasarkan data akademik, terdapat 5 data yang terdiri dari bulan Agustus, September, Oktober, November, dan Desember. Data akademik memiliki rasio yang sama dari ke 5 data tersebut sebesar 0.74:0.26, dimana 0.74 merupakan data mayoritas dan 0.26 adalah data minoritas (**Tabel 4.3**). Data akademik yang memiliki karakteristik data yang berbeda antara Agustus –

Desember, sehingga data mayoritas dan minoritas memiliki hasil evaluasi berbeda.

Terdapat perbedaan yang cukup signifikan ketika melekakukan proses *oversampling* A-SUWO dimana *imbalanced ratio* data awal sebesar 0.74 berubah menjadi 0.5. Hal ini menandakan bahwa A-SUWO terbukti dapat menyeimbangkan jumlah data dengan menambahkan data sintetik pada kelas minoritas. Sedangkan pada metode *undersampling* Tomek dan RUS rasio imbalance yang diperoleh tidak berubah dikarenakan sifat metode *undersampling* hanya sebagai pembersihan data. Sedangkan pada NCL, rasio imbalance terdapat penurunan namun penurunan tersebut tidak banyak dan tidak signnifikan. Maksudnya disini yaitu, rasio 0.47 masih dapat dikatakan seimbang.

Tabel 4.3. Rasio Jumlah Data Akademik Kelas Mayoritas Sebelum dan Sesudah Penanganan Imbalance

***Imbalanced* data**	**IR Tanpa *Oversampling &Undersampling***	**IR A-SUWO**	**IR A-SUWO +NCL**	**IR A-SUWO +Tomek**	**IR A-SUWO +RUS**
Agustus	0.74	0.5	0.47	0.5	0.5
September	0.74	0.5	0.47	0.5	0.5
Oktober	0.74	0.5	0.47	0.5	0.5
November	0.74	0.5	0.46	0.5	0.5
Desember	0.74	0.5	0.49	0.5	0.5

Table 4.4 merupakan table rasio jumlah data pada dataset Keel sesbelum maupun sesudah penangan imbalance. Penanganan imbalance yang dimaksud yaitu metode *oversampling* A-SUWO dan metode *undersampling* NCL, Tomek, dan RUS. Dari hasil tersebut dapat disimpulkan bahwa pada data awal, masing-masing data memiliki karakteristik tidak seimbang yakni rasio imbalance berada di range 0.65-0.86. Pada A-SUWO rasio sudah seimbang secara keseluruhan yaitu sebesar 0.5. Sedangkan ketika metode *undersampling* ditambahkan, terdapat kenaikan maupun penurunan rasio imbalance pada data. Hal ini bukan menjadi suatu masalah karena penurunan dan kenaikan yang terjadi tidak terlalu signifikan dan data pada rasio 0.44-0.5 masih bisa dikatakan bahwa data tersebut seimbang.

Tabel 4.4 Rasio Jumlah Data Keel Kelas Mayoritas Sebelum dan Sesudah Penanganan Imbalance

***Imbalanced* data**	**IR Tanpa *Oversampling &Undersampling***	**IR A-SUWO**	**IR A-SUWO +NCL**	**IR A-SUWO +Tomek**	**IR A-SUWO +RUS**
Ecolli1	0.77	0.5	0.48	0.5	0.5
Glass0	0.67	0.5	0.45	0.49	0.5
Haberman	0.74	0.5	0.44	0.5	0.45
Iris0	0.67	0.5	0.49	0.5	0.5
Segment0	0.86	0.5	0.49	0.5	0.49
Vehicle0	0.77	0.5	0.48	0.5	0.5
Wisconsin	0.65	0.5	0.5	0.5	0.5

4.4. Evaluasi Hasil

Pengukuran kinerja yang dibuat menggunakan basis konsep dari *confusion* matrix seperti pada Tabel 4.5. Dimana **kelas positive (+)** atau kelas minoritas (disimbolkan dengan **1**) merepresentasikan kelas mahasiswa yang memiliki **kemungkinan gagal** dalam akademik. Sedangkan **kelas (-)** atau kelas mayoritas (disimbolkan dengan **0**) merupakan kelas mahasiswa yang berhasil atau **tidak memiliki kemungkinan gagal dalam akademik**.

Tabel 4.5 Contoh *Confussion Matrix* pada Data Bulan Desember

	Prediksi (+)	**Prediksi (-)**
Aktual (+)	1762	133
Aktual (-)	144	531

Evaluasi kinerja model terhadap 2 kelas yang berbeda diberikan dalam bentuk Akurasi, *Recall*, *Precision*, *F-Measure* untuk masing-masing kelas. **Persamaan 4.1- Persamaan 4.4** merupakan contoh perhitungan precision, recall, fmeasure serta akurasi pada ujucoba penelitian ini.

Precision merupakan perhitungan ketepatan klasifikasi pada jumlah data berlabel positif atau data kelas minoritas yang memang benar secara actual merupakan kelas positif. Sedangkan, *Recall* (disebut juga *sensitivity*) adalah perhitungan ketepatan klasifier pada jumlah data positive yang teridentifikasi benar sebagai kelas positif. *F-measure* sendiri adalah *relative importance* antar *precision* dan *recall*.

$$Precision = \frac{TP}{TP+FP} = \frac{1762}{1762+144} = 92.44 \quad (4.1)$$

$$Recall = \frac{TP}{TP+FN} = \frac{1762}{1762+133} = 93{,}08 \quad (4.2)$$

$$F - measure = \frac{(1+\beta^2)\,.\,Recall.\,Precision}{\beta^2\,.\,Recall\,.\,Precision} = \frac{2*92.44*93.08}{2*92.44*93.08} = 92.71 \quad (4.3)$$

$$Akurasi = \frac{TP+TN}{TP+TN+FP+FN} = \frac{1762+531}{1762+531+144+133} = 89.22 \quad (4.4)$$

Evaluasi metode yang diusulkan menggunakan *decision tree classification* (J48) untuk data akademik Universitas XYZ angkatan 2014 dan 2015. Pengujian evaluasi meliputi beberapa kriteria, yaitu nilai *precision, recall, F-Measure,* dan *accuracy*. Tabel 4.6 merupakan nilai *precision*, *recall*, *f-measure*, dan akurasi pada data akademik dari bulan Agustus-Desember. Hasil perhitungan menunjukkan bahwa bulan Desember memiliki nilai akurasi tertinggi yaitu sebesar 85.92%. Hal tersebut masuk akal, karena data pada bulan Desember cenderung lebih matang karena terdapat nilai akhir yang dapat sangat membedakan antar mahasiswa yang berpotensi lulus tepat waktu dan tidak. Selanjutnya, akurasi terbaik kedua diduduki oleh bulan November yakni 73.47%, disusul oleh bulan September dan Agustus yakni masing-masing sebesar 70.21% dan 66,36%.

Tabel 4.6 Data **Akademik Original** sebelum dilakukan *Oversampling*

Dataset	*Precision*	*Recall*	*F-measure*	*Accuracy*
Desember	90.50%	90.40%	90.45%	85.92%
November	81.14%	80.92%	81.03%	73.47%
Oktober	81.36%	80.96%	81.16%	73.21%
September	80.91%	78.01%	79.43%	70.21%
Agustus	77.55%	76.53%	77.04%	66,36%

Table 4.7 merupakan table hasil evaluasi data akademik setelah dilakukan proses *oversampling* A-SUWO. Pada tahap ini, terdapat perbedaan tidak hanya dari sisi akurasi namun juga dari rasio *imbalance* yang sudah dijelaskan pada sesi sebelumnya. Setelah dilakukan proses *oversampling* menggunakan metode A-SUWO, terdapat kenaikan akurasi sebesar 2-3% pada masing-masing metode bulan jika dibandingkan dengan data original. Kenaikan tersebut tidak hanya pada nilai akurasi namun juga pada *precision, recall,* dan juga f-

measure. Seperti pada data original, pada tahap ini, nilai evaluasi terbaik dicapai bulan Desember yakni akurasi sebesar 87.66%, *precision* sebesar 90.72%, *recall* sebesar 92.74%, dan *F-measure* sebesar 91.72%.

Tabel 4.7 Hasil Evaluasi Dataset **Akademik** dengan **A-SUWO**

Dataset	*Precision*	*Recall*	*F-measure*	*Accuracy*
Desember	90.72%	92.74%	91.72%	87.66%
November	82.46%	81.03%	81.74%	74.05%
Oktober	82.36%	83.13%	82.74%	74.34%
September	80.26%	80.82%	80.53%	72.04%
Agustus	78.06%	82.3%	80.12%	69.89%

Tabel 4.8 - Tabel 4.10 merupakan table hasil evaluasi data akademik menggunakan gabungan metode *oversampling* A-SUWO dengan NCL (*Neighborhood Cleaning Rule*), Tomek-Link, dan *Random Undersampling* (RUS). Dari hasil evaluasi menyatakan bahwa pada metode NCL (Table 4.8) terdapat kenaikan akurasi sebesar 0.5-3% apabila dibandingkan dengan hasil akurasi menggunakan metode A-SUWO saja. Akurasi tertinggi diperoleh bulan Desember sebesar 88.97% sedangkan kenaikan tertinggi terdapat pada Bulan Agustus yakni sebesar 3%.

Sedangkan pada Tabel 4.9 merupakan evaluasi menggunakan gabungan metode A-SUWO-Tomek link. Pada metode A-SUWO-Tomek link akurasi tertinggi juga diperoleh pada bulan Desember yakni sebesar 89.22% dan kenaikan terjadi pada bulan agustustus yakni sebesar 3%. Sedangkan range kenaikan apabila dibandingkan dengan metode A-SUWO saja yakni sebesar 0.2-3%.

Tabel 4.10 merupakan hasil evaluasi A-SUWO dengan metode *undersampling* RUS dimana metode ini merupakan metode paling sederhana, dan akurasi tertinggi terdapat pada bulan Desember sebesar 89.22%. Hal ini tidak terlalu jauh perbedaanya dengan metode Tomek-Link yakni 89.29%.

Tabel 4.8 Hasil Evaluasi Dataset **Akademik** dengan **A-SUWO-NCL**

Dataset	*Precision*	*Recall*	*F-measure*	*Accuracy*
Desember	92.31%	92.77%	92.53%	88.97%
November	90.65%	73.43%	81.1%	74.79%
Oktober	88.91%	74.98%	81.34%	74.64%
September	86.07%	66.77%	75.11%	67.52%
Agustus	80.05%	82.92%	81.46%	72.17%

Tabel 4.9 Hasil Evaluasi Dataset Akademik dengan A-SUWO-Tomek Link

Dataset	*Precision*	*Recall*	*F-measure*	*Accuracy*
Desember	92.51%	92.98%	92.72%	89.29%
November	89.98%	73.55%	80.89%	74.41%
Oktober	88.27%	75.86%	81.56%	74.74%
September	84.54%	76.4%	80.23%	72.29%
Agustus	79.89%	82.94%	81.38%	72.02%

Tabel 4.10. Hasil Evaluasi Dataset **Akademi**k dengan **A-SUWO-RUS**

Dataset	*Precision*	*Recall*	*F-measure*	*Accuracy*
Desember	92.46%	92.96%	92.7%	89.22%
November	89.75%	74.07%	81.13%	74.61%
Oktober	87.68%	75.6%	81.17%	74.37%
September	84.55%	76.54%	80.31%	72.37%
Agustus	79.95%	82.78%	81.33%	71.98%

Table 4.11 merupakan table hasil akurasi pada setiap metode dan pada baris paling bawah merupakan nilai rata-rata capaian akurasi dari masing-masing metode. Data yang dibandingkan meliputi data original, data setelah proses A-SUWO, data setelah penggabungan menggunakan NCL, Tomek link dan *Random Undersampling*. Hasil evaluasi menunjukkan bahwa A-SUWO-Tomek memperoleh akurasi rata-rata tertinggi yakni 76.54% sedangkan ASUWO-RUS sebesar 76.51% dan ASUWO-NCL 75.62%. Hal tersebut menandakan bahwa penambahan metode *undersampling* dapat menaikkan akurasi apabila dibandingkan dengan metode original dan metode A-SUWO saja.

Tabel 4.11 Evaluasi **Akurasi** Dataset **Akademik** pada Setiap Metode

Bulan	Original	ASUWO	ASUWO+ NCL	ASUWO+ TOMEK	ASUWO+ RUS
Agustus	66.36%	69.89%	72.17%	72.02%	71.98%
September	70.21%	72.04%	67.52%	72.29%	72.37%
Oktober	73.21%	74.33%	74.64%	74.74%	74.37%
November	73.47%	74.05%	74.79%	74.41%	74.61%
Desember	85.92%	87.66%	88.97%	89.27%	89.22%
Rata-rata	73.80%	75.59%	75.62%	76.55%	76.51%

Table 4.12 merupakan table hasil *precision.* Hasil evaluasi menunjukkan bahwa A-SUWO-Tomek memperoleh *precision* rata-rata tertinggi yakni 87.04% sedangkan ASUWO-RUS sebesar 86.88% dan ASUWO-NCL 87.0%. Table 4.13 merupakan table *recall*. Hasil evaluasi menunjukkan bahwa recall pada A-SUWO saja yang tertinggi yakni sebesar 84%. Sedangkan A-SUWO-Tomek memperoleh *recall* rata-rata tertinggi kedua yakni 80.35% sedangkan ASUWO-RUS sebesar 80.17% dan ASUWO-NCL 78.16%. Pada bulan Agustus dan Desember, baik nilai *precision* maupun *recall* mengalami kenaikan. Sedangkan pada bulan September, Oktober dan November, nilai *precision* naik sedangkan nilai *recall* menurun. Hal tersebut dipengaruhi oleh data yang sangat overlap, sehingga ada kemungkinan terjadi data mayoritas yang dihapus pada bulan September, Oktober dan November tersebut justru mempengaruhi factor kedekatan terhadap data minoritasnya.

Tabel 4.12 Evaluasi ***Precision*** Dataset **Akademik** pada Setiap Metode

Bulan	Original	ASUWO	ASUWO+ NCL	ASUWO+ TOMEK	ASUWO+ RUS
Agustus	77,55%	78,06%	80,06%	79,89%	79,95%
September	80,91%	80,26%	86,07%	84,54%	84,55%
Oktober	81,36%	82,36%	86,91%	88,27%	87,68%
November	81,14%	82,46%	89,65%	89,98%	89,75%
Desember	90,50%	90,72%	92,31%	92,51%	92,46%
Rata-rata	82,29%	82,77%	87,00%	87,04%	86,88%

Tabel 4.13 Evaluasi ***Recall*** Dataset **Akademik** pada Setiap Metode

Bulan	Original	ASUWO	ASUWO+ NCL	ASUWO+ TOMEK	ASUWO+ RUS
Agustus	76,53%	82,3%	82,86%	82,94%	82,78%
September	78,01%	80,82%	66,77%	76,4%	76,44%
Oktober	80,96%	83,13%	74,98%	75,86%	75,60%
November	80,92%	81,03%	73,43%	73,55%	73,07%
Desember	90,40%	92,74%	92,77%	92,98%	92,96%
Rata-rata	81,36%	84,00%	78,16%	80,35%	80,17%

Ujicoba kedua dilakukan menggunakan dataset Keel yang terdiri dari 8 jenis dataset berbeda. Dimana oercobaan juga dibandingkan pada data sebelum dilakukan penanganan imbalanced apapun, data setelah dilakukan *oversampling* (A-SUWO) dan setelah dilakukan penggabungan *oversampling* dan *undersampling* (NCL, Tomek-link, dan RUS). Tabel 4.14 merupakan table hasil evaluasi berupa nilai akurasi, *precision*, *recall* dan *f-measure* pada dataset Keel.

Tabel 4.14. Data **Keel Original** sebelum dilakukan Oversampling

Dataset	*Precision*	*Recall*	*F-measure*	*Accuracy*
Ecoli1	90.74	94.23	92.45	88.06
Glass0	71.43	71.43	71.43	61.9
Haberman	79.55	77.78	78.65	68
Iris0	100	100	100	100
Segment0	99.49	99.24	99.37	98.92
Vehicle0	89.78	95.35	92.48	88.17
Wisconsin	93.33	95.45	94.38	91.65

Table 4.15 merupakan table hasil evaluasi data Keel setelah dilakukan proses *oversampling* A-SUWO. Setelah dilakukan proses *oversampling* menggunakan metode A-SUWO, terdapat kenaikan akurasi sebesar 0.5-9% pada masing-masing metode data jika dibandingkan dengan data original. Kenaikan tersebut tidak hanya pada nilai akurasi namun juga pada *precision, recall*, dan juga *f-measure*.

Tabel 4.15. Hasil Evaluasi Dataset **Keel** dengan **A-SUWO**

Dataset	*Precision*	*Recall*	*F-measure*	*Accuracy*
Ecoli1	94.53	94.90	94.72	91.94
Glass0	75.79	72.14	73.89	66.19
Haberman	78.96	78.22	78.52	68.52
Iris0	100	100	100	100
Segment0	99.70	100	99.846	99.74
Vehicle0	93.02	95.04	94.016	90.75
Wisconsin	92.69	94.99	93.826	91.91

Tabel 4.16 - Tabel 4.18 merupakan table hasil evaluasi data akademik menggunakan gabungan metode oversampling A-SUWO dengan NCL (*Neighborhood Cleaning Rule*), Tomek-Link, dan *Random Undersampling* (RUS). Dari hasil evaluasi menyatakan bahwa pada metode NCL (Table 4.13) terdapat kenaikan akurasi sebesar 1-4% apabila dibandingkan dengan hasil

akurasi menggunakan metode A-SUWO saja. Akurasi dengan kenaikan tertinggi diperoleh data Glass0 yakni sebesar 4%.

Sedangkan pada Tabel 4.14 merupakan evaluasi menggunakan gabungan metode A-SUWO-Tomek link. Pada metode A-SUWO-Tomek link kenaikan akurasi tertinggi juga diperoleh data Glass0 yakni sebesar 4% dimana nilai akurasinya adalah 70.73%. Sedangkan *range* kenaikan apabila dibandingkan dengan metode A-SUWO saja yakni sebesar 2-4%.

Tabel 4.15 merupakan hasil evaluasi A-SUWO dengan metode *undersampling* RUS dimana metode ini merupakan metode paling sederhana, dan akurasi tertinggi terdapat pada data Iris dan Segment0 yakni sebesar 100%.

Tabel 4.16 Hasil Evaluasi **Keel** Dataset dengan **A-SUWO-NCL**

Dataset	***Precision***	***Recall***	***F-measure***	***Accuracy***
Ecoli1	95.83	92	94.28	90.91
Glass0	92.14	60	72.64	70.16
Haberman	76.18	81.36	78.64	67.67
Iris0	100	100	100	100
Segment0	100	100	100	100
Vehicle0	97.60	93.59	95.53	93.33
Wisconsin	94.81	96.55	95.67	94.37

Tabel 4.17 Hasil Evaluasi **Keel** dataset dengan **A-SUWO-Tomek Link**

Dataset	***Precision***	***Recall***	***F-measure***	***Accuracy***
Ecoli1	95.17	94.4	94.78	92.12
Glass0	85.68	66.67	74.94	70.73
Haberman	78.26	81.81	79.99	70
Iris0	100	100	100	100
Segment0	100	100	100	100
Vehicle0	98.37	93.59	95.92	93.93
Wisconsin	94.80	96.55	95.67	94.37

Tabel 4.18 Hasil Evaluasi **Keel** dataset dengan **A-SUWO-RUS**

Dataset	***Precision***	***Recall***	***F-measure***	***Accuracy***
Ecoli1	95.97	95.2	95.58	88.18
Glass0	77.08	64.44	69.98	65.18
Haberman	77.75	79.54	78.61	68.33
Iris0	100	100	100	100
Segment0	100	100	100	100
Vehicle0	98.36	93.44	95.84	93.81
Wisconsin	94.81	96.55	95.67	94.37

Tabel 4.19 Evaluasi **Akurasi** Dataset Keel pada Setiap Metode

Data	Original	ASUWO	ASUWO+ NCL	ASUWO+ TOMEK	ASUWO+ RUS
Ecoli1	88.06	91.94	90.91	92.12	88.18
Glass0	61.9	66.19	70.16	70.73	65.18
Haberman	68	68.52	67.67	70	68.33
Iris	100	100	100	100	100
Segment0	98.92	99.74	100	100	100
Vehicle0	88.17	90.75	93.33	93.93	93.81
Wisconsin	91.65	91.91	94.34	94.37	94.37
Rata-rata	**82.18**	**84.08**	**84.76**	**85.41**	**84.04**

Table 4.19 merupakan table evaluasi akurasi pada setiap metode dan pada baris paling bawah merupakan nilai rata-rata capaian akurasi dari masing-masing metode. Data yang dibandingkan meliputi data original, data setelah proses A-SUWO, data setelah penggabungan menggunakan NCL, Tomek link dan *Random Undersampling*. Hasil evaluasi menunjukkan bahwa A-SUWO-Tomek memperoleh akurasi rata-rata tertinggi yakni 85.407% sedangkan ASUWO-RUS sebesar 84.036% dan ASUWO-NCL 84.075%. Hal tersebut menandakan bahwa penambahan metode *undersampling* dapat menaikkan akurasi apabila dibandingkan dengan metode original dan metode A-SUWO saja. Sedangkan Tabel 4.20 dan Tabel 4.21 merupakan table *precision* dan *recall*.

Tabel 4.20 Evaluasi ***Precision*** Dataset Keel pada Setiap Metode

Data	Original	ASUWO	ASUWO+ NCL	ASUWO+ TOMEK	ASUWO+ RUS
Ecoli1	96	94.534	95.83	95.17	95.97
Glass0	82.35	75.79	92.14	85.68	77.08
Haberman	68	78.96	76.18	78.26	77.75
Iris	99.75	99.70	100	100	100
Segment0	96.09	93.02	97.6	98.37	98.36
Vehicle0	79.09	92.69	94.81	94.81	94.81

Wisconsin	100	100	100	100	100
Rata-rata	**88.75**	**90.67**	**93.79**	**93.18**	**91.99**

Tabel 4.21 Evaluasi ***Recall*** Dataset Keel pada Setiap Metode

Data	Original	ASUWO	ASUWO+ NCL	ASUWO+ TOMEK	ASUWO+ RUS
Ecoli1	71.43	94.9	92	95.17	95.97
Glass0	77.78	72.14	60	66.67	64.44
Haberman	68	78.22	81.36	81.81	79.54
Iris	95.35	100	100	100	100
Segment0	96.63	95.04	93.59	93.59	93.44
Vehicle0	77.73	94.99	96.55	96.55	96.55
Wisconsin	100	100	100	100	100
Rata-rata	**83.5**	**90.76**	**89.07**	**90.54**	**89.99**

4.5. Evaluasi Parameter

Evaluasi parameter pada metode A-SUWO dilakukan untuk memperoleh nilai parameter paling optimal pada setiap data. Tabel 4.22 memaparkan parameter-parameter apa saja yang diujicoba beserta range nilai ujicoba dan nilai optimal berdasarkan hasil ujicoba. Deskripsi hasil ujicoba akan dijelaskan pada tabel-tabel berikutnya yang diperoleh dari nilai rata-rata akurasi.

Tabel 4.22 Parameter A-SUWO pada Dataset Akademik dan Keel

Parameter	Nilai Ujicoba	Nilai Optimal Hasil Ujicoba
K-fold	3 dan 5	Dataset Akademik: 5 Dataset Keel: 3
Chtresh	0.3;1; 1.5	Dataset Akademik: 1 Dataset Keel: 1
NN	3;5;7	Dataset Akademik: 3 Dataset Keel: 5
NumIteration	5	Dataset Akademik: 5 Dataset Keel: 5

4.5.1. Ujicoba Parameter A-SUWO pada Data Akademik

Seperti dijelaskan pada sub bab sebelumnya, bahwa ujicoba parameter dilakukan untuk memperoleh parameter A-SUWO yang paling optimal. Pada data akademik, nilai K-fold yang diuji coba yaitu 3 dan 5. K-fold merupakan suatu parameter untuk menentukan fold pada saat proses cluster sizing atau penentuan size pada klaster. Sedangkan parameter Cthresh merupakan parameter threshold pada proses clustering menggunakan metode agglomerative hierarchical clustering. Range nilai CThresh yang diujicoba yaitu 0.3, 1 dan 5. Semakin besar nilai threshold, maka akan menghasilkan jumlah klaster yang semakin sedikit namun size dari masing-masing cluster semakin besar. Parameter NN yaitu parameter penentu ketetanggaan pada saat menentukan bobot pada kelas minoritas. Sedangkan parameter iterasi yaitu proses perulangan untuk memastikan bahwa setiap iterasi tidak memiliki perbedaan yang signifikan.

a. Parameter K-Fold

Ujicoba Parameter K-fold dilakukan hanya dalam 1 iterasi saja dan masing-masing dihitung nilai akurasinya. Setelah dilakukan pengujian ternyata pada Data Akademik, nilai K-fold=5 yang memiliki akurasi relatif lebih tinggi yaitu 65.29% pada Bulan Agustus, 71.51% pada bulan September, 76.78%, 78.19%, dan 86.89% masing-masing pada Bulan Oktober, November, dan Desember.

Tabel 4.23 Hasil Evaluasi Ujicoba Parameter K-Fold pada Data Akademik

Bulan	K-fold=3	K-fold=5
Agustus	64.95%	**65.29%**
September	68.1%	**71.51%**
Oktober	76.47%	**76.78%**
November	77.43%	**78.19%**
Desember	86.6%	**86.89%**
Rata-Rata	74.71%	**75.73%**

b. Parameter Cthresh

Pada ujicoba parameter CThresh, ujicoba dilakukan juga dengan 1 iterasi saja. Ternyata CThresh=1 merupakan nilai yang paling optimal apabila dibandingkan dengan CThresh=0.3 dan CThresh=1.5. Hal tersebut sangat relevan dengan paper acuan. Perhitungan dilakukan dengan membandingkan nilai rata-rata akurasi, precision, recall, dan f-measure dari masing-masing bulan. Nilai rata-rata akurasi yang diraih yakni 76.56%, 84.62%, 83.29%, dan 83.95% masing-masing untuk nilai akurasi*, precision, recall*, dan *f-measure*.

Tabel 4.24 Hasil Evaluasi Ujicoba Parameter CThresh pada Data Akademik

Chtresh=0.3						
Matriks	**Agustus**	**September**	**Oktober**	**November**	**Desember**	**Rata-Rata**
Akurasi	68.14	69.62	77.47	77.95	87.75	76.19
Precision	78.03	80.68	84.35	84.64	90.85	83.71
Recall	79.06	77.32	85.28	85.66	92.72	84.008
Fmeasure	78.54	78.97	84.82	85.15	91.78	83.852
Chtresh=1						
Matriks	**Agustus**	**September**	**Oktober**	**November**	**Desember**	**Rata-Rata**
Akurasi	66.36	72.77	77.32	79.11	87.24	**76.56**
Precision	78.22	82.46	84.57	86.64	91.22	**84.62**
Recall	75.37	80.12	84.7	84.77	91.51	**83.29**
Fmeasure	76.77	81.27	84.64	85.69	91.36	**83.95**
Chtresh=1.5						
Matriks	**Agustus**	**September**	**Oktober**	**November**	**Desember**	**Rata-Rata**
Akurasi	67.83	71.02	76.81	78.18	86.81	76.13
Precision	77.82	80.47	84.87	85.84	89.74	83.75
Recall	78.85	80.17	83.44	84.34	92.72	83.9
Fmeasure	78.83	80.32	84.15	85.08	91.21	83.92

c. Parameter NN

Sedangkan pada ujicoba parameter NN, nilai NN=3 merupakan nilai yang paling optimal pada Data Akademik apabila dibandingkan dengan NN=5 maupun NN=7. Sama hal nya dengan CThresh, perhitungan juga dilakukan dengan membandingkan nilai rata-rata akurasi, precision, recall, dan f-measure dari masing-masing bulan. Nilai rata-rata akurasi yang diraih yakni 77.31%, 84.06%, 85.45%, dan 84.75% untuk masing-masing untuk nilai akurasi, *precision, recall,* dan *f-measure*.

Tabel 4.25 Hasil Evaluasi Ujicoba Parameter NN pada Data Akademik

NN=3						
Matriks	**Agustus**	**September**	**Oktober**	**November**	**Desember**	**Rata-Rata**
Akurasi	69.39	72.54	78.52	78.61	87.48	**77.31**
Precision	78.05	81.48	84.64	85.62	90.53	**84.06**
Recall	81.38	81.22	86.6	85.35	92.72	**85.45**
Fmeasure	79.68	81.35	85.61	85.48	91.61	**84.75**
NN=5						
Matriks	**Agustus**	**September**	**Oktober**	**November**	**Desember**	**Rata-Rata**
Akurasi	66.36	72.77	77.32	79.11	87.24	76.56
Precision	78.22	82.46	84.57	86.64	91.22	84.62
Recall	75.37	80.12	84.7	84.77	91.51	83.29
Fmeasure	76.77	81.27	84.64	85.69	91.36	83.95
NN=7						
Matriks	**Agustus**	**September**	**Oktober**	**November**	**Desember**	**Rata-Rata**
Akurasi	67.64	70.71	75.84	77.13	85.76	75.42
Precision	78.54	81.28	84.44	86.3	90.18	84.15
Recall	77.22	78.32	82.44	82.02	90.56	82.11
Fmeasure	77.87	79.77	83.43	84.11	90.37	83.11

d. Parameter Iterasi

Seperti dijelaskan sebelumnya, bahwa parameter iterasi digunakan karena pada metode A-SUWO terdapat proses yang random sehingga setiap kali iterasi tidak dapat menghasilkan akurasi yang sama persis. Oleh karena itu dilakukan perulangan untuk memastikan bahwa masing-masing iterasi tidak memberikan dampak yang besar. Hal tersebut dibuktikan dengan selisih dari masing-masing nilai evaluasi setiap iterasi tidak berbeda signifikan. Data setelah proses A-SUWO inilah yang akan menjadi input untuk tahap selanjutnya yakni metode *undersampling*. Jadi, masing-masing data Bulan memiliki 5 jenis data berdasarkan iterasi.

Tabel 4.26 Hasil Evaluasi Ujicoba Parameter Iterasi pada Data Akademik

Agustus (K-FOLD=5, CThresh=1, NN=3)							
Matriks	**Original**	**Iterasi1**	**Iterasi2**	**Iterasi3**	**Iterasi4**	**Iterasi5**	**Rata-rata**
Akurasi	66.36	69.54	71.18	70.32	69	69.43	69.894
Precision	77.55	78.12	78.46	78.14	77.46	78.14	78.064
Recall	76.53	81.54	83.97	82.96	81.75	81.28	82.3
Fmeasure	77.04	79.79	81.12	80.48	79.55	79.68	80.124
September (K-FOLD=5, CThresh=1. NN=3)							
Matriks	**Original**	**Iterasi1**	**Iterasi2**	**Iterasi3**	**Iterasi4**	**Iterasi5**	**Rata-rata**
Akurasi	70.21	71.97	72.5	72.45	72.02	71.28	72.044
Precision	80.91	80.9	80.69	80.32	79.85	79.54	80.26
Recall	78.01	80.33	82.44	81.23	81.23	78.85	80.816
Fmeasure	79.43	80.61	81.55	80.77	80.53	79.2	80.532
Oktober (K-FOLD=5, CThresh=1. NN=3)							
Matriks	**Original**	**Iterasi1**	**Iterasi2**	**Iterasi3**	**Iterasi4**	**Iterasi5**	**Rata-rata**
Akurasi	73.21	74.08	74.32	74.52	74.47	74.25	74.328
Precision	81.36	84.2	81	81.07	81.53	83.98	82.356
Recall	80.96	84.86	82.15	81.97	82.02	84.65	83.13
Fmeasure	81.16	84.53	81.57	81.52	81.78	84.32	82.744

November (K-FOLD=5, CThresh=1, NN=3)							
Matriks	**Original**	**Iterasi1**	**Iterasi2**	**Iterasi3**	**Iterasi4**	**Iterasi5**	**Rata-rata**
Akurasi	73.47	73.53	73.91	74.5	74.46	73.84	74.048
Precision	81.14	81.53	86.46	81.67	81.67	80.96	82.458
Recall	80.92	78.97	85.29	80.35	80.29	80.24	81.028
Fmeasure	81.03	80.23	85.88	81	80.97	80.6	81.736
Desember (K-FOLD=5, CThresh=1, NN=3)							
Matriks	**Original**	**Iterasi1**	**Iterasi2**	**Iterasi3**	**Iterasi4**	**Iterasi5**	**Rata-rata**
Akurasi	85.92	87.05	87.98	87.36	87.75	88.14	87.656
Precision	90.5	90.3	91.01	90.34	90.89	91.07	90.722
Recall	90.4	92.35	92.88	92.77	92.67	93.04	92.742
Fmeasure	90.45	91.32	91.93	91.54	91.77	92.04	91.72

4.5.2. Ujicoba Parameter A-SUWO pada Data Keel

Sama halnya pada Data Akademik, pada data Keel, nilai K-fold yang diuji coba yaitu 3 dan 5. K-fold merupakan suatu parameter untuk menentukan fold pada saat proses cluster sizing atau penentuan size pada klaster. Sedangkan parameter Cthresh merupakan parameter threshold pada proses clustering menggunakan metode agglomerative hierarchical clustering. Range nilai CThresh yang diujicoba yaitu 0.3, 1 dan 5. Semakin besar nilai threshold, maka akan menghasilkan jumlah klaster yang semakin sedikit namun size dari masing-masing cluster semakin besar. Parameter NN yaitu parameter penentu ketetanggaan pada saat menentukan bobot pada kelas minoritas. Sedangkan parameter iterasi yaitu proses perulangan untuk memastikan bahwa setiap iterasi tidak memiliki perbedaan yang signifikan.

a. Parameter K-Fold

Ujicoba Parameter K-fold dilakukan hanya dalam 1 iterasi saja dan masing-masing dihitung nilai akurasinya. Setelah dilakukan pengujian ternyata pada Data Keel, nilai K-fold=3 yang memiliki nilai rata-rata akurasi, precision,

recall, dan fmeasure yang relatif lebih tinggi yaitu 84.84%, 89.62%, 88.88% dan 89.27%.

Tabel 4.27 Hasil Evaluasi Ujicoba Parameter K-Fold pada Dataset Keel

Matriks	**Akurasi**	
	K-fold=3	**K-fold=5**
Akurasi	**84.84**	84.27
Precision	**89.62**	89.45
Recall	**88.88**	88.26
Fmeasure	**89.270**	88.83
Rata-rata	**85.030**	84.53

b. Parameter Cthresh

Pada ujicoba parameter CThresh, ujicoba dilakukan juga dengan 1 iterasi saja. Ternyata CThresh=1 merupakan nilai yang paling optimal apabila dibandingkan dengan CThresh=0.3 dan CThresh=1.5. Perhitungan dilakukan dengan membandingkan nilai rata-rata akurasi, precision, recall, dan f-measure dari masing-masing bulan. Namun dikarenakan dataset Keel yang diujicoba cukup banyak, makan akan disertakan lebih detail pada Lampiran. Tabel 4.28 merupakan nilai rata-rata akurasi yang diraih yakni 87.18%, 91.28%, 90.7%, dan 90.96% masing-masing untuk nilai akurasi, precision, recall, dan f-measure.

Tabel 4.28 Hasil Evaluasi Ujicoba Parameter CThresh pada Dataset Keel

Cthresh=0.3		**Cthresh=1**		**Cthresh=1.5**	
Matriks	**Rata-Rata**	**Matriks**	**Rata-Rata**	**Matriks**	**Rata-Rata**
Akurasi	76.19	**Akurasi**	**87.18**	**Akurasi**	84.88
Precision	83.71	***Precision***	**91.28**	***Precision***	89.97
Recall	84.008	***Recall***	**90.70**	***Recall***	88.86
Fmeasure	83.85	***Fmeasure***	**90.96**	***Fmeasure***	89.39

c. Parameter NN

Sedangkan pada ujicoba parameter NN, nilai NN=5 merupakan nilai yang paling optimal pada Data Akademik apabila dibandingkan dengan NN=3 maupun NN=7. Sama hal nya dengan CThresh, perhitungan juga dilakukan dengan membandingkan nilai rata-rata akurasi, precision, recall, dan f-measure dari masing-masing bulan. Namun dikarenakan dataset Keel yang diujicoba cukup banyak, makan akan disertakan lebih detail pada Lampiran. Tabel 4.24 merupakan nilai rata-rata evaluasi yang diraih yakni 87.18%, 91.28%, 90.7%, dan 90.96% untuk masing-masing untuk nilai akurasi, *precision, recall*, dan *f-measure.*

Tabel 4.29 Hasil Evaluasi Ujicoba Parameter NN pada Dataset Keel

NN =3		**NN=5**		**NN=7**	
Matriks	**Rata-Rata**	**Matriks**	**Rata-Rata**	**Matriks**	**Rata-Rata**
Akurasi	83.91	**Akurasi**	**87.18**	**Akurasi**	84.26
Precision	88.55	**Precision**	**91.28**	**Precision**	78.85
Recall	89.63	**Recall**	**90.70**	**Recall**	87.64
Fmeasure	89.02	**Fmeasure**	**90.96**	**Fmeasure**	88.67

d. Parameter Iterasi

Seperti dijelaskan sebelumnya, bahwa parameter iterasi digunakan karena pada metode A-SUWO terdapat proses yang random sehingga setiap kali iterasi tidak dapat menghasilkan akurasi yang sama persis. Oleh karena itu dilakukan perulangan untuk memastikan bahwa masing-masing iterasi tidak memberikan dampak yang besar. Namun dikarenakan dataset Keel yang diujicoba cukup banyak, makan akan disertakan lebih detail pada Lampiran. Tabel 4.30 merupakan nilai rata-rata dari masing-masing perhitungan akurasi, *precision, recall* dan *f-measure*. Data setelah proses A-SUWO inilah yang akan menjadi input untuk tahap selanjutnya yakni metode *undersampling*. Jadi, masing-masing data memiliki 5 jenis data berdasarkan iterasi.

Tabel 4.30 Hasil Evaluasi Ujicoba Parameter Iterasi pada Dataset Keel

Iterasi = 5				
Data	**Akurasi**	**Precision**	**Recall**	**F-Measure**
Ecoli1	91.94	94.53	94.90	94.72
Glass0	66.19	75.79	72.14	73.89
Haberman	68.52	78.97	78.22	78.52
Iris	100	100	100	100
Segment0	99.74	99.70	100	99.85
Vehicle0	90.75	93.02	95.04	94.02
Wisconsin	91.91	92.69	94.99	93.83

[Halaman ini sengaja dikosongkan]

BAB 5
KESIMPULAN DAN SARAN

5.1. Kesimpulan

Berdasarkan hasil uji coba dari metode usulan memberikan beberapa kesimpulan yaitu:

1. Metode ***oversampling*** **A-SUWO** digunakan untuk menangani permasalahan *imbalanced* data akademik Universitas XYZ angkatan 2014 dan 2015 dengan nilai rata-rata akurasi, *recall*, dan *precision* sebesar **75.59%, 84%, 82.7% untuk Data Akademik** dan **84.08, 90.67%, 90.76% % untuk Dataset Keel.**
2. Sedangkan untuk tahapan penggabungan dengan metode *undersampling*, diperoleh **akurasi, *precision*, dan *recall***: sebesar **76.55%, 87.04%, 80.35%** untuk **ASUWO-Tomeklink,** pada metode **ASUWO-RUS 76.51%, 86,88%, 80.17%** dan **ASUWO-NCL** sebesar **75.59%, 87%, 78.16%.**
3. Sedangkan untuk **Dataset Keel** dalam penelitian ini diperoleh hasil evaluasi rata-rata akurasi, *precision*, *recall* yakni **85.41%, 93.18%,90.54%** untuk **ASUWO-Tomeklink**, **84.08%, 93,79%, 89.07%** pada **ASUWO-NCL**, dan **84.04%, 91.99%, 89.99%** untuk metode gabungan **ASUWO-RUS**.

5.2. Saran

Berdasarkan hasil yang diperoleh pada penelitian ini, diperlukan *improvement* dari segi algoritma A-SUWO maupun *undersampling* sehingga diperoleh hasil peningkatan yang lebih signifikan. Selain itu, analisa pada penelitian ini akan lebih optimal dan lebih berdampak apabila data yang digunakan sangat berfokus terhadap suatu data tertentu saja (tidak terlalu banyak data yang diuji coba).

[Halaman ini sengaja dikosongkan]

DAFTAR PUSTAKA

Asriningtias, Y. *et al.* (2014) 'APLIKASI DATA MINING UNTUK MENAMPILKAN INFORMASI', 8(1), pp. 837–848.

Barua, S. *et al.* (2014) 'MWMOTE — Majority Weighted Minority Oversampling Technique for Imbalanced Data Set Learning', *IEEE Transaction Knowledge Management*, 26(2), pp. 405–425.

Blagus, R. *et al.* (2013) 'SMOTE for high-dimensional class-imbalanced data', *BMC Bioinformatics*, 14(1), p. 106. doi: 10.1186/1471-2105-14-106.

Cahyo, M. and Lianto, J. (2018) 'Penanganan imbalance class data laboratorium kesehatan dengan Majority Weighted Minority Oversampling Technique', 4(1), pp. 14–20.

Chawla, N. V. *et al.* (2002) 'SMOTE: Synthetic minority over-sampling technique', *Journal of Artificial Intelligence Research*, 16, pp. 321–357. doi: 10.1613/jair.953.

Daniel, T. (2005) *An Introduction to Data Mining*.

Fahrudin, T., Buliali, J. L. and Fatichah, C. (2016) 'RANDSHUFF : an Algorithm to Handle Imbalance Class for Qualitative Data', 11(December), pp. 1093–1104. doi: 10.15866/irecos.v11i12.10956.

Fahrudin, T. and Fatichah, C. (2016) 'Predictive Modeling of the First Year Evaluation Based on Demographics Data : Case Study Students of Telkom University , Indonesia', pp. 0–5.

V. Garcıa, R. A. Mollineda, J. S. S. (2008) 'On the k -NN performance in a challenging scenario of imbalance and overlapping', pp. 269–280. doi: 10.1007/s10044-007-0087-5.

Gong, J. and Kim, H. (2017) 'RHSBoost: Improving classification performance in imbalance data', *Computational Statistics and Data Analysis*. Elsevier B.V., 111, pp. 1–13. doi: 10.1016/j.csda.2017.01.005.

Guo, X. *et al.* (2016) 'On the Class Imbalance Problem *', (October). doi: 10.1109/ICNC.2008.871.

Hart, P. E. (1967) 'The Condensed Nearest Neighbor Rule', pp. 1966–1967.

Jayasree, S. and Gavya, A. A. (2015) 'Classification of Imbalance Problem by

MWMOTE and SSO', pp. 1–4.

John, M. and Jayasudha, J. S. (2017) 'Enhancing Performance of Deep Learning Based Text Summarizer', *International Journal of Applied Engineering Research*, 12(24), pp. 15986–15993.

Laurikkala, J. (2001) 'Improving Identification of Difficult Small Classes by Balancing Class Distribution', pp. 63–66.

Mahmood, A. M. (2017) 'Class Imbalance Learning in Data Mining – A Survey Class Imbalance Learning in Data Mining – A Survey', (September 2015). doi: 10.21742/ijctsns.2015.3.2.02.

Mellor, A. *et al.* (2015) 'Exploring issues of training data imbalance and mislabelling on random forest performance for large area land cover classification using the ensemble margin', *ISPRS Journal of Photogrammetry and Remote Sensing*. International Society for Photogrammetry and Remote Sensing, Inc. (ISPRS), 105, pp. 155–168. doi: 10.1016/j.isprsjprs.2015.03.014.

More, A. (2016) 'Survey of resampling techniques for improving classification performance in unbalanced datasets', 10000, pp. 1–7.

Napierała, K. (2012a) 'Improving Rule Classifiers For Imbalanced Data', (October).

Napierała, K. (2012b) 'Improving Rule Classifiers For Imbalanced Data Doctoral Dissertation', (October).

Nekooeimehr, I. and Lai-yuen, S. K. (2016) 'Adaptive semi-unsupervised weighted oversampling (A-SUWO) for imbalanced datasets', *Expert Systems With Applications*. Elsevier Ltd, 46, pp. 405–416. doi: 10.1016/j.eswa.2015.10.031.

Piri, S., Delen, D. and Liu, T. (2018) 'A synthetic informative minority over-sampling (SIMO) algorithm leveraging support vector machine to enhance learning from imbalanced datasets', *Decision Support Systems*. Elsevier B.V, 106, pp. 15–29. doi: 10.1016/j.dss.2017.11.006.

Prusa, J. *et al.* (2015) 'Using Random Undersampling to Alleviate Class Imbalance on Tweet Sentiment Data'. doi: 10.1109/IRI.2015.39.

Purwar, A. and Singh, S. K. (2015) 'Hybrid Prediction Model with missing

value Imputation for medical data', *EXPERT SYSTEMS WITH APPLICATIONS*. Elsevier Ltd. doi: 10.1016/j.eswa.2015.02.050.

Ramentol, E., Caballero, Y. and Bello, R. (2011) 'SMOTE-RS B ∗ : a hybrid preprocessing approach based on oversampling and undersampling for high imbalanced data-sets using SMOTE and rough sets theory'. doi: 10.1007/s10115-011-0465-6.

Rivera, W. A. (2017) 'Noise Reduction A Priori Synthetic Over-Sampling for class imbalanced data sets', *Information Sciences journal*. Elsevier Inc., 408, pp. 146–161. doi: 10.1016/j.ins.2017.04.046.

Ross, J., Morgan, Q. and Publishers, K. (1994) 'Book Review : C4 . 5 : Programs for Machine Learning', 240, pp. 235–240.

Rushi Longadge, S. S. D. and Malik, L. (2013) 'Class Imbalance Problem in Data Mining : Review', 2(1).

Thanathamathee, P. and Lursinsap, C. (2013) 'Handling imbalanced data sets with synthetic boundary data generation using bootstrap re-sampling and AdaBoost techniques', *Pattern Recognition Letters*. Elsevier B.V., 34(12), pp. 1339–1347. doi: 10.1016/j.patrec.2013.04.019.

Tomek, I. (1976) 'Two Modifications of CNN', *IEEE Transactions on System, Man, and Cybernetics,* pp. 769–772.

Weiss, G. M. and Provost, F. (2003) 'Learning When Training Data are Costly : The Effect of Class Distribution on Tree Induction', 19, pp. 315–354.

[Halaman ini sengaja dikosongkan]

LAMPIRAN

Lampiran 1 : Contoh Data Akademik Universitas XYZ Angkatan 2014 – 2015

1. **Data akademik Universitas XYZ Angkatan 2014 – 2015 bulan Agustus**

Instance	A1	A2	A3	A4	A5	A6	A7	A8	A9	A10	A11	A12	A13	A14	A15	A16	A17	A18	A19	A20	A21	A22	A23
1178	0.0	0.0	0.0	0.0	0.0	100.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
1184	0.0	0.0	0.0	0.0	0.0	100.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
1190	0.0	0.0	0.0	0.0	0.0	66.7	0.0	0.0	33.3	0.0	0.0	0.0	50.0	50.0	0.0	0.0	0.0	0.0	50.0	50.0	0.0	0.0	0.0
1192	0.0	0.0	0.0	0.0	0.0	100.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
1196	0.0	0.0	0.0	0.0	0.0	100.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
1199	0.0	0.0	0.0	0.0	0.0	69.2	0.0	0.0	30.8	0.0	0.0	0.0	25.0	0.0	25.0	50.0	50.0	25.0	25.0	0.0	0.0	0.0	0.0
1203	0.0	0.0	0.0	0.0	0.0	100.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
1206	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
1210	0.0	0.0	0.0	0.0	0.0	100.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
1211	0.0	0.0	0.0	0.0	0.0	100.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
1213	0.0	0.0	0.0	0.0	0.0	100.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
1215	0.0	0.0	0.0	0.0	0.0	100.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
1225	0.0	0.0	0.0	0.0	0.0	100.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
1226	0.0	0.0	0.0	0.0	0.0	87.5	0.0	0.0	12.5	0.0	0.0	0.0	0.0	100.0	0.0	0.0	0.0	0.0	100.0	0.0	0.0	0.0	0.0
1228	0.0	0.0	0.0	0.0	0.0	100.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
1229	0.0	0.0	0.0	0.0	0.0	100.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
1230	0.0	0.0	0.0	0.0	0.0	100.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
1232	0.0	0.0	0.0	0.0	0.0	100.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
1233	0.0	0.0	0.0	0.0	0.0	100.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
1236	0.0	0.0	0.0	0.0	0.0	100.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
1240	0.0	0.0	0.0	0.0	0.0	100.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0

1241	0.0	0.0	0.0	0.0	0.0	100.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
1245	0.0	0.0	0.0	0.0	0.0	100.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
1251	0.0	0.0	0.0	0.0	0.0	100.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0

Instance	A24	A25	A26	A27	A28	A29	A30	A31	A32	A33	A34	Label
1178	0.0	0.0	0.0	0.0	0.0	0.0	0.0	43.5	58.2	100.0	0	Tidak Lulus
1184	0.0	0.0	0.1	0.0	0.5	0.0	0.0	60.9	58.2	100.0	0	Tidak Lulus
1190	0.0	0.0	0.0	0.0	0.1	0.0	0.0	14.9	34.4	80.0	0	Tidak Lulus
1192	0.0	0.0	0.0	0.0	0.0	0.0	0.0	10.4	46.7	80.0	1	Lulus
1196	0.0	0.0	0.0	0.0	0.0	0.0	0.0	15.2	41.8	100.0	1	Lulus
1199	0.0	0.0	0.0	0.0	0.0	0.0	0.0	70.9	67.4	100.0	1	Lulus
1203	0.0	0.0	0.1	0.0	0.3	0.0	0.2	60.9	58.2	100.0	0	Tidak Lulus
1206	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0	Tidak Lulus
1210	0.0	0.0	0.1	0.0	0.3	0.2	0.0	25.2	46.1	80.0	1	Lulus
1211	0.0	0.0	0.8	0.0	3.2	0.0	0.0	43.5	58.2	100.0	0	Tidak Lulus
1213	0.0	0.0	0.0	0.0	0.0	0.0	0.0	11.1	58.7	80.0	0	Tidak Lulus
1215	0.0	0.0	0.0	0.0	0.0	0.0	0.0	24.9	57.4	60.0	0	Tidak Lulus
1225	0.0	0.0	0.0	0.0	0.0	0.0	0.0	15.2	36.3	80.0	1	Lulus
1226	0.0	0.0	0.0	0.0	0.0	0.0	0.0	15.2	41.8	100.0	1	Lulus
1228	0.0	0.0	0.0	0.0	0.0	0.0	0.0	25.2	46.1	80.0	1	Lulus
1229	0.0	0.0	0.1	0.0	0.4	0.0	0.0	14.9	34.4	80.0	1	Lulus
1230	0.0	0.0	0.0	0.0	0.0	0.0	0.0	43.9	63.3	80.0	1	Lulus
1232	0.0	0.0	0.0	0.0	0.0	0.0	0.0	60.9	58.2	100.0	1	Lulus
1233	0.0	0.0	0.2	0.0	0.0	0.6	0.0	11.1	58.7	80.0	1	Lulus
1236	0.0	0.0	0.9	0.0	2.6	0.6	0.0	10.4	46.7	80.0	1	Lulus
1240	0.0	0.0	1.4	0.0	3.0	1.7	0.0	14.9	34.4	80.0	1	Lulus
1241	0.0	0.0	0.0	0.0	0.0	0.0	0.0	15.2	36.3	80.0	1	Lulus
1245	0.0	0.0	9.5	0.0	16.8	13.4	1.8	25.2	46.1	80.0	1	Lulus
1251	0.0	0.0	1.4	0.0	5.5	0.0	0.0	60.9	58.2	100.0	1	Lulus

2. Data akademik Universitas XYZ Angkatan 2014 – 2015 bulan September

Instance	A1	A2	A3	A4	A5	A6	A7	A8	A9	A10	A11	A12	A13	A14	A15	A16	A17	A18	A19	A20	A21	A22	A23
1178	0.0	0.0	0.0	0.0	0.0	89.5	0.0	8.8	1.8	0.0	0.0	0.0	0.0	0.0	100.0	0.0	0.0	0.0	100.0	0.0	0.0	0.0	0.0
1184	0.0	0.0	0.0	0.0	0.0	98.3	0.0	0.0	1.8	0.0	0.0	0.0	0.0	0.0	0.0	100.0	100.0	0.0	0.0	0.0	0.0	0.0	0.0
1190	0.0	0.0	0.0	0.0	0.0	19.6	0.0	0.0	80.4	0.0	19.5	34.2	12.2	19.5	0.0	14.6	53.7	0.0	34.2	12.2	0.0	0.0	0.0
1192	0.0	0.0	0.0	0.0	0.0	100.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
1196	0.0	0.0	0.0	0.0	0.0	96.0	0.0	0.0	4.0	0.0	50.0	0.0	0.0	50.0	0.0	0.0	50.0	0.0	50.0	0.0	0.0	0.0	0.0
1199	0.0	0.0	0.0	0.0	0.0	81.3	0.0	0.0	18.8	0.0	0.0	0.0	16.7	16.7	33.3	33.3	33.3	16.7	25.0	25.0	0.0	0.0	0.0
1203	0.0	0.0	0.0	0.0	0.0	98.3	0.0	0.0	1.8	0.0	0.0	0.0	0.0	0.0	0.0	100.0	100.0	0.0	0.0	0.0	0.0	0.0	0.0
1206	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
1210	0.0	0.0	0.0	0.0	0.0	98.1	0.0	0.0	1.9	0.0	100.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	100.0	0.0	0.0	0.0	0.0
1211	0.0	0.0	0.0	0.0	0.0	98.3	0.0	0.0	1.8	0.0	0.0	0.0	0.0	0.0	0.0	100.0	0.0	100.0	0.0	0.0	0.0	0.0	0.0
1213	0.0	0.0	0.0	0.0	0.0	98.3	0.0	0.0	1.8	0.0	100.0	0.0	0.0	0.0	0.0	0.0	100.0	0.0	0.0	0.0	0.0	0.0	0.0
1215	0.0	0.0	0.0	0.0	0.0	100.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
1225	0.0	0.0	0.0	0.0	0.0	100.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
1226	0.0	0.0	0.0	0.0	0.0	98.0	0.0	0.0	2.0	0.0	0.0	0.0	0.0	100.0	0.0	0.0	0.0	0.0	100.0	0.0	0.0	0.0	0.0
1228	0.0	0.0	0.0	0.0	0.0	90.6	3.8	0.0	5.7	0.0	0.0	66.7	33.3	0.0	0.0	0.0	33.3	33.3	0.0	33.3	0.0	0.0	0.0
1229	0.0	0.0	0.0	0.0	0.0	100.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
1230	0.0	0.0	0.0	0.0	0.0	98.1	0.0	0.0	1.9	0.0	0.0	100.0	0.0	0.0	0.0	0.0	100.0	0.0	0.0	0.0	0.0	0.0	0.0
1232	0.0	0.0	0.0	0.0	0.0	100.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
1233	0.0	0.0	0.0	0.0	0.0	100.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
1236	0.0	0.0	0.0	0.0	0.0	100.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
1240	0.0	0.0	0.0	0.0	0.0	100.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
1241	0.0	0.0	0.0	0.0	0.0	98.0	0.0	0.0	2.0	0.0	0.0	0.0	0.0	0.0	100.0	0.0	0.0	0.0	0.0	100.0	0.0	0.0	0.0
1245	0.0	0.0	0.0	0.0	0.0	100.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
1251	0.0	0.0	0.0	0.0	0.0	100.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0

Instance	A24	A25	A26	A27	A28	A29	A30	A31	A32	A33	A34	Label
1178	0.0	0	0.01	0	0	0.02	0	67.12	70.29	83.75	0	Tidak Lulus
1184	0.0	0	0.05	0	0.08	0	0	63.94	70.29	83.75	0	Tidak Lulus
1190	0.0	0	0.01	0	0.02	0	0	31.04	63.24	80.66	0	Tidak Lulus
1192	0.0	0	0	0	0	0	0	42.43	59.88	93.42	1	Lulus
1196	0.0	0	0.01	0	0.02	0	0	39.97	62.66	80.66	1	Lulus
1199	0.0	0	0	0	0	0	0	67.72	67.17	96.71	1	Lulus
1203	0.0	0	0.05	0	0.05	0	0.09	63.94	70.29	83.75	0	Tidak Lulus
1206	0.0	0	0	0	0	0	0	0	0	0	0	Tidak Lulus
1210	0.0	0	0.17	0	0.05	0.5	0	24.54	65.9	80.66	1	Lulus
1211	0.0	0	0.53	0	0.56	0.59	0	67.12	70.29	83.75	0	Tidak Lulus
1213	0.0	0	0	0	0	0	0	46.96	85.9	67.7	0	Tidak Lulus
1215	0.0	0	0	0	0	0	0	44.7	69.02	74.08	0	Tidak Lulus
1225	0.0	0	0	0	0	0	0	42.44	55.95	93.42	1	Lulus
1226	0.0	0	0	0	0	0	0	39.97	62.66	80.66	1	Lulus
1228	0.0	0	0	0	0	0	0	24.54	65.9	80.66	1	Lulus
1229	0.0	0	0.15	0	0.23	0	0	31.04	63.24	80.66	1	Lulus
1230	0.0	0	0	0	0	0	0	63.21	80.69	67.7	1	Lulus
1232	0.0	0	0	0	0	0	0	63.94	70.29	83.75	1	Lulus
1233	0.0	0	3.34	2.26	1.32	3.87	3.01	46.96	85.9	67.7	1	Lulus
1236	0.0	0	0.51	0	0.6	0.44	0	42.43	59.88	93.42	1	Lulus
1240	0.0	0	0.65	0	0.53	1.1	0	31.04	63.24	80.66	1	Lulus
1241	0.0	0	0	0	0	0	0	42.44	55.95	93.42	1	Lulus
1245	0.0	0	8.87	0	3.6	22.3	1.42	24.54	65.9	80.66	1	Lulus
1251	0.0	0	0.68	0	1.06	0	0	63.94	70.29	83.75	1	Lulus

3. Data akademik Universitas XYZ Angkatan 2014 – 2015 bulan Oktober

Instance	A1	A2	A3	A4	A5	A6	A7	A8	A9	A10	A11	A12	A13	A14	A15	A16	A17	A18	A19	A20	A21	A22	A23
1178	0.0	0.0	0.0	0.0	0.0	93.0	0.0	5.8	1.2	0.0	0.0	0.0	0.0	0.0	100.0	0.0	0.0	0.0	100.0	0.0	0.0	0.0	0.0
1184	0.0	0.0	0.0	0.0	0.0	98.8	0.0	0.0	1.2	0.0	0.0	0.0	0.0	0.0	0.0	100.0	100.0	0.0	0.0	0.0	0.0	0.0	0.0
1190	0.0	0.0	0.0	0.0	0.0	16.7	0.0	0.0	83.3	0.0	18.3	30.0	11.7	23.3	0.0	16.7	53.3	0.0	33.3	13.3	0.0	0.0	0.0
1192	0.0	0.0	0.0	0.0	0.0	96.1	0.0	0.0	3.9	0.0	33.3	0.0	0.0	0.0	66.7	0.0	100.0	0.0	0.0	0.0	0.0	0.0	0.0
1196	0.0	0.0	0.0	0.0	0.0	95.8	0.0	0.0	4.2	0.0	33.3	0.0	0.0	66.7	0.0	0.0	33.3	0.0	66.7	0.0	0.0	0.0	0.0
1199	0.0	0.0	0.0	0.0	0.0	80.9	0.0	0.0	19.1	0.0	11.8	0.0	11.8	29.4	23.5	23.5	35.3	29.4	17.7	17.7	0.0	0.0	0.0
1203	0.0	0.0	0.0	0.0	0.0	98.8	0.0	0.0	1.2	0.0	0.0	0.0	0.0	0.0	0.0	100.0	100.0	0.0	0.0	0.0	0.0	0.0	0.0
1206	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
1210	0.0	0.0	0.0	0.0	0.0	94.7	0.0	0.0	5.3	0.0	25.0	25.0	0.0	50.0	0.0	0.0	25.0	0.0	50.0	25.0	0.0	0.0	0.0
1211	0.0	0.0	0.0	0.0	0.0	98.8	0.0	0.0	1.2	0.0	0.0	0.0	0.0	0.0	0.0	100.0	0.0	100.0	0.0	0.0	0.0	0.0	0.0
1213	0.0	0.0	0.0	0.0	0.0	98.8	0.0	0.0	1.2	0.0	100.0	0.0	0.0	0.0	0.0	0.0	100.0	0.0	0.0	0.0	0.0	0.0	0.0
1215	0.0	0.0	0.0	0.0	0.0	100.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
1225	0.0	0.0	0.0	0.0	0.0	98.7	0.0	0.0	1.3	0.0	100.0	0.0	0.0	0.0	0.0	0.0	100.0	0.0	0.0	0.0	0.0	0.0	0.0
1226	0.0	0.0	0.0	0.0	0.0	98.6	0.0	0.0	1.4	0.0	0.0	0.0	0.0	100.0	0.0	0.0	0.0	0.0	100.0	0.0	0.0	0.0	0.0
1228	0.0	0.0	0.0	0.0	0.0	88.2	2.6	0.0	9.2	0.0	14.3	42.9	14.3	28.6	0.0	0.0	42.9	28.6	14.3	14.3	0.0	0.0	0.0
1229	0.0	0.0	0.0	0.0	0.0	100.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
1230	0.0	0.0	0.0	0.0	0.0	97.4	0.0	0.0	2.6	0.0	0.0	50.0	50.0	0.0	0.0	0.0	100.0	0.0	0.0	0.0	0.0	0.0	0.0
1232	0.0	0.0	0.0	0.0	0.0	100.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
1233	0.0	0.0	0.0	0.0	0.0	100.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
1236	0.0	0.0	0.0	0.0	0.0	100.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
1240	0.0	0.0	0.0	0.0	0.0	100.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
1241	0.0	0.0	0.0	0.0	0.0	93.4	0.0	0.0	6.6	0.0	20.0	0.0	0.0	40.0	40.0	0.0	40.0	20.0	0.0	40.0	0.0	0.0	0.0
1245	0.0	0.0	0.0	0.0	0.0	98.7	0.0	0.0	1.3	0.0	0.0	100.0	0.0	0.0	0.0	0.0	0.0	100.0	0.0	0.0	100.0	0.0	0.0
1251	0.0	0.0	0.0	0.0	0.0	100.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0

Instance	A24	A25	A26	A27	A28	A29	A30	A31	A32	A33	A34	A35	A36	A37	Label
1178	0.0	0.0	0.0	0.0	0.0	0.0	0.0	73.7	69.1	79.3	50.0	16.7	33.3	0	Tidak Lulus
1184	0.0	0.0	0.0	0.0	0.1	0.0	0.0	71.0	68.7	79.3	66.7	0.0	33.3	0	Tidak Lulus
1190	0.0	0.0	0.0	0.0	0.0	0.0	0.0	28.5	62.4	73.4	0.0	66.7	33.3	0	Tidak Lulus
1116	0.0	0.0	0.0	0.0	0.0	0.0	0.0	42.7	58.1	78.6	50.0	33.3	16.7	1	Lulus
1051	0.0	0.0	0.0	0.0	0.0	0.0	0.0	37.2	58.7	71.4	66.7	0.0	33.3	1	Lulus
1052	0.0	0.0	0.0	0.0	0.0	0.0	0.0	63.8	60.2	83.9	27.8	55.6	16.7	1	Lulus
1203	0.0	0.0	0.0	0.0	0.1	0.0	0.0	71.0	68.7	79.3	50.0	16.7	33.3	0	Tidak Lulus
1206	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	100.0	0	Tidak Lulus
1050	0.0	0.0	0.1	0.0	0.1	0.3	0.0	25.2	68.9	64.3	50.0	16.7	33.3	1	Lulus
1211	0.0	0.0	0.2	0.0	0.6	0.3	0.0	73.7	69.1	79.3	50.0	16.7	33.3	0	Tidak Lulus
1213	0.0	0.0	0.0	0.0	0.0	0.0	0.0	46.5	83.5	63.6	50.0	16.7	33.3	0	Tidak Lulus
1215	0.0	0.0	0.0	0.0	0.0	0.0	0.0	46.9	65.6	66.1	83.3	0.0	16.7	0	Tidak Lulus
1035	0.0	0.0	0.0	0.0	0.0	0.0	0.0	44.7	52.7	82.1	44.4	38.9	16.7	1	Lulus
1036	0.0	0.0	0.1	0.0	0.0	0.2	0.1	37.2	58.7	71.4	50.0	16.7	33.3	1	Lulus
1037	0.0	0.0	0.0	0.0	0.0	0.0	0.0	25.2	68.9	64.3	33.3	33.3	33.3	1	Lulus
1038	0.0	0.0	0.2	0.0	0.4	0.1	0.3	28.5	62.4	73.4	83.3	0.0	16.7	1	Lulus
1039	0.0	0.0	0.0	0.0	0.0	0.0	0.0	61.0	72.6	60.7	50.0	16.7	33.3	1	Lulus
1040	0.0	0.0	0.1	0.0	0.3	0.3	0.0	71.0	68.7	79.3	66.7	0.0	33.3	1	Lulus
1041	0.0	0.0	3.9	2.8	2.3	5.5	5.2	46.5	83.5	63.6	66.7	0.0	33.3	1	Lulus
1042	0.0	0.0	0.3	0.0	0.7	0.3	0.0	42.7	58.1	78.6	83.3	0.0	16.7	1	Lulus
1043	0.0	0.0	0.5	0.0	0.9	0.7	0.0	28.5	62.4	73.4	66.7	16.7	16.7	1	Lulus
1044	0.0	0.0	0.0	0.0	0.0	0.0	0.0	44.7	52.7	82.1	27.8	55.6	16.7	1	Lulus
1045	0.0	0.0	4.4	0.0	4.4	12.1	0.6	25.2	68.9	64.3	44.4	22.2	33.3	1	Lulus
1046	0.0	0.0	0.5	0.0	1.1	0.5	0.0	71.0	68.7	79.3	66.7	0.0	33.3	1	Lulus
1047		0.0	0.0	0.0	0.0	0.0	0.0	73.7	69.1	79.3	50.0	16.7	33.3	1	Lulus
1049		0.0	0.0	0.0	0.1	0.0	0.0	71.0	68.7	79.3	66.7	0.0	33.3	1	Lulus
1115		0.0	0.0	0.0	0.0	0.0	0.0	28.5	62.4	73.4	0.0	66.7	33.3	1	Lulus

4. **Data akademik Universitas XYZ Angkatan 2014 – 2015 bulan November**

Instance	A1	A2	A3	A4	A5	A6	A7	A8	A9	A10	A11	A12	A13	A14	A15	A16	A17	A18	A19	A20	A21	A22	A23
1178	0.0	0.0	0.0	0.0	0.0	93.8	0.0	3.9	2.3	0.0	0.0	0.0	0.0	33.3	33.3	33.3	0.0	33.3	66.7	0.0	0.0	0.0	0.0
1184	0.0	0.0	0.0	0.0	0.0	98.4	0.0	0.0	1.6	0.0	0.0	0.0	0.0	0.0	0.0	100.0	100.0	0.0	0.0	0.0	0.0	0.0	0.0
1190	0.0	0.0	0.0	0.0	0.0	13.1	0.0	0.0	86.9	0.0	19.4	31.2	11.8	22.6	0.0	15.1	53.8	0.0	34.4	11.8	0.0	0.0	0.0
1192	0.0	0.0	0.0	0.0	0.0	95.0	0.0	0.0	5.0	0.0	16.7	0.0	0.0	0.0	33.3	50.0	83.3	16.7	0.0	0.0	0.0	0.0	0.0
1196	0.0	0.0	0.0	0.0	0.0	92.9	0.0	0.0	7.1	0.0	25.0	0.0	25.0	50.0	0.0	0.0	50.0	0.0	50.0	0.0	0.0	0.0	0.0
1199	0.0	0.0	0.0	0.0	0.0	80.8	0.0	0.0	17.7	1.5	17.4	8.7	8.7	26.1	17.4	21.7	34.8	39.1	13.0	13.0	0.0	0.0	0.0
1203	0.0	0.0	0.0	0.0	0.0	98.4	0.0	0.0	1.6	0.0	0.0	0.0	0.0	0.0	50.0	50.0	50.0	0.0	0.0	50.0	0.0	0.0	0.0
1206	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
1210	0.0	0.0	0.0	0.0	0.0	95.8	0.0	0.0	4.2	0.0	20.0	20.0	20.0	40.0	0.0	0.0	20.0	20.0	40.0	20.0	0.0	0.0	0.0
1211	0.0	0.0	0.0	0.0	0.0	98.5	0.0	0.0	1.6	0.0	0.0	0.0	50.0	0.0	0.0	50.0	50.0	50.0	0.0	0.0	0.0	0.0	0.0
1213	0.0	0.0	0.0	0.0	0.0	98.4	0.0	0.0	1.6	0.0	50.0	0.0	0.0	50.0	0.0	0.0	100.0	0.0	0.0	0.0	0.0	0.0	0.0
1215	0.0	0.0	0.0	0.0	0.0	100.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
1225	0.0	0.0	0.0	0.0	0.0	96.6	0.0	0.9	2.6	0.0	33.3	0.0	0.0	0.0	0.0	66.7	100.0	0.0	0.0	0.0	0.0	0.0	0.0
1226	0.0	0.0	0.0	0.0	0.0	98.2	0.0	0.0	1.8	0.0	0.0	0.0	0.0	50.0	0.0	50.0	50.0	0.0	50.0	0.0	0.0	0.0	0.0
1228	0.0	0.0	0.0	0.0	0.0	87.3	1.7	3.4	7.6	0.0	22.2	33.3	11.1	33.3	0.0	0.0	44.4	22.2	22.2	11.1	0.0	0.0	0.0
1229	0.0	0.0	0.0	0.0	0.0	99.1	0.0	0.0	0.9	0.0	0.0	100.0	0.0	0.0	0.0	0.0	100.0	0.0	0.0	0.0	0.0	0.0	0.0
1230	0.0	0.0	0.0	0.0	0.0	96.4	0.0	0.0	3.6	0.0	0.0	25.0	50.0	0.0	0.0	25.0	100.0	0.0	0.0	0.0	0.0	0.0	0.0
1232	0.0	0.0	0.0	0.0	0.0	100.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
1233	0.0	0.0	0.0	0.0	0.0	96.8	0.0	2.4	0.8	0.0	100.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	100.0	0.0	0.0	0.0	0.0
1236	0.0	0.0	0.0	0.0	0.0	100.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
1240	0.0	0.0	0.0	0.0	0.0	99.1	0.0	0.0	0.9	0.0	100.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	100.0	0.0	0.0	0.0	0.0
1241	0.0	0.0	0.0	0.0	0.0	89.7	0.0	0.0	10.3	0.0	33.3	8.3	0.0	25.0	16.7	16.7	50.0	33.3	0.0	16.7	0.0	0.0	0.0
1245	0.0	0.0	0.0	0.0	0.0	99.2	0.0	0.0	0.9	0.0	0.0	100.0	0.0	0.0	0.0	0.0	0.0	100.0	0.0	0.0	100.0	0.0	0.0
1251	0.0	0.0	0.0	0.0	0.0	97.6	0.0	0.8	1.6	0.0	0.0	0.0	0.0	0.0	50.0	50.0	0.0	50.0	0.0	50.0	0.0	0.0	0.0

Instance	A24	A25	A26	A27	A28	A29	A30	A31	A32	A33	A34	A35	A36	A37	Label
1178	0.0	0.0	0.1	0.0	0.2	0.0	0.0	63.2	69.5	80.9	50.0	16.7	33.3	0	Tidak Lulus
1184	0.0	0.0	0.2	0.0	0.4	0.3	0.1	59.8	69.1	80.9	66.7	0.0	33.3	0	Tidak Lulus
1190	0.0	0.0	0.0	0.0	0.0	0.0	0.0	20.1	59.7	76.7	0.0	66.7	33.3	0	Tidak Lulus
1192	0.0	0.0	0.1	0.0	0.2	0.1	0.0	36.7	57.7	85.1	50.0	33.3	16.7	1	Lulus
1196	0.0	0.0	0.2	0.0	0.2	0.5	0.0	34.8	59.6	77.6	66.7	0.0	33.3	1	Lulus
1199	0.0	0.0	0.1	0.0	0.2	0.3	0.0	46.9	58.7	88.8	27.8	55.6	16.7	1	Lulus
1203	0.0	0.0	0.2	0.0	0.3	0.4	0.0	59.8	69.1	80.9	50.0	16.7	33.3	0	Tidak Lulus
1206	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	100.0	0	Tidak Lulus
1210	0.0	0.0	0.1	0.0	0.1	0.2	0.0	22.4	69.0	70.1	50.0	16.7	33.3	1	Lulus
1211	0.0	0.0	0.3	0.0	0.6	0.5	0.0	63.2	69.5	80.9	50.0	16.7	33.3	0	Tidak Lulus
1213	0.0	0.0	0.0	0.0	0.0	0.1	0.0	42.3	84.1	64.6	50.0	16.7	33.3	0	Tidak Lulus
1215	0.0	0.0	0.0	0.0	0.0	0.0	0.0	40.6	65.8	71.3	83.3	0.0	16.7	0	Tidak Lulus
1225	0.0	0.0	0.0	0.0	0.0	0.0	0.0	36.3	53.7	87.6	44.4	38.9	16.7	1	Lulus
1226	0.0	0.0	0.3	0.0	0.6	0.5	0.2	34.8	59.6	77.6	50.0	16.7	33.3	1	Lulus
1228	0.0	0.0	0.0	0.0	0.1	0.0	0.0	22.4	69.0	70.1	33.3	33.3	33.3	1	Lulus
1229	0.0	0.0	0.3	0.0	0.6	0.3	0.2	20.1	59.7	76.7	83.3	0.0	16.7	1	Lulus
1230	0.0	0.0	0.8	0.0	1.1	1.8	0.3	44.8	71.9	61.3	50.0	16.7	33.3	1	Lulus
1232	0.0	0.0	0.1	0.0	0.3	0.2	0.0	59.8	69.1	80.9	66.7	0.0	33.3	1	Lulus
1233	0.0	0.0	7.1	8.7	3.6	6.9	9.4	42.3	84.1	64.6	66.7	0.0	33.3	1	Lulus
1236	0.0	0.0	0.3	0.0	0.9	0.4	0.0	36.7	57.7	85.1	83.3	0.0	16.7	1	Lulus
1240	0.0	0.0	0.7	0.0	1.7	1.1	0.0	20.1	59.7	76.7	66.7	16.7	16.7	1	Lulus
1241	0.0	0.0	0.3	0.0	0.0	1.3	0.1	36.3	53.7	87.6	27.8	55.6	16.7	1	Lulus
1245	0.0	0.0	4.6	0.0	5.5	12.6	0.8	22.4	69.0	70.1	44.4	22.2	33.3	1	Lulus
1251	0.0	0.0	0.4	0.0	1.1	0.5	0.0	59.8	69.1	80.9	66.7	0.0	33.3	1	Lulus

5. **Data akademik Universitas XYZ Angkatan 2014 – 2015 bulan Desember**

Instance	A1	A2	A3	A4	A5	A6	A7	A8	A9	A10	A11	A12	A13	A14	A15	A16	A17	A18	A19	A20	A21	A22	A23
1178	0.0	0.0	0.0	0.0	0.0	94.4	0.0	3.5	2.1	0.0	0.0	0.0	0.0	33.3	33.3	33.3	0.0	33.3	66.7	0.0	0.0	0.0	0.0
1184	0.0	0.0	0.0	0.0	0.0	98.6	0.0	0.0	1.4	0.0	0.0	0.0	0.0	0.0	0.0	100.0	100.0	0.0	0.0	0.0	0.0	0.0	0.0
1190	0.0	0.0	0.0	0.0	0.0	11.5	0.0	0.0	88.5	0.0	18.5	31.5	13.0	22.2	0.0	14.8	52.8	0.0	35.2	12.0	0.0	0.0	0.0
1192	0.0	0.0	0.0	0.0	0.0	94.7	0.0	0.0	5.3	0.0	14.3	14.3	0.0	0.0	28.6	42.9	85.7	14.3	0.0	0.0	0.0	0.0	0.0
1196	0.0	0.0	0.0	0.0	0.0	93.7	0.0	0.0	6.4	0.0	25.0	0.0	25.0	50.0	0.0	0.0	50.0	0.0	50.0	0.0	0.0	0.0	0.0
1199	0.0	0.0	0.0	0.0	0.0	82.5	0.0	0.0	16.1	1.4	17.4	8.7	8.7	26.1	17.4	21.7	34.8	39.1	13.0	13.0	0.0	0.0	0.0
1203	0.0	0.0	0.0	0.0	0.0	98.6	0.0	0.0	1.4	0.0	0.0	0.0	0.0	0.0	50.0	50.0	50.0	0.0	0.0	50.0	0.0	0.0	0.0
1206	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
1210	0.0	0.0	0.0	0.0	0.0	96.2	0.0	0.0	3.8	0.0	20.0	20.0	20.0	40.0	0.0	0.0	20.0	20.0	40.0	20.0	0.0	0.0	0.0
1211	0.0	0.0	0.0	0.0	0.0	97.9	0.0	0.0	2.1	0.0	0.0	33.3	33.3	0.0	0.0	33.3	66.7	33.3	0.0	0.0	0.0	0.0	0.0
1213	0.0	0.0	0.0	0.0	0.0	98.6	0.0	0.0	1.5	0.0	50.0	0.0	0.0	50.0	0.0	0.0	100.0	0.0	0.0	0.0	0.0	0.0	0.0
1215	0.0	0.0	0.0	0.0	0.0	100.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
1225	0.0	0.0	0.0	0.0	0.0	96.9	0.0	0.8	2.3	0.0	33.3	0.0	0.0	0.0	0.0	66.7	100.0	0.0	0.0	0.0	0.0	0.0	0.0
1226	0.0	0.0	0.0	0.0	0.0	97.6	0.0	0.0	2.4	0.0	0.0	0.0	0.0	33.3	33.3	33.3	33.3	0.0	33.3	33.3	0.0	0.0	0.0
1228	0.0	0.0	0.0	0.0	0.0	85.7	1.5	3.0	9.8	0.0	30.8	30.8	15.4	23.1	0.0	0.0	53.9	15.4	23.1	7.7	0.0	0.0	0.0
1229	0.0	0.0	0.0	0.0	0.0	99.2	0.0	0.0	0.8	0.0	0.0	100.0	0.0	0.0	0.0	0.0	100.0	0.0	0.0	0.0	0.0	0.0	0.0
1230	0.0	0.0	0.0	0.0	0.0	95.2	0.0	1.6	3.2	0.0	0.0	25.0	50.0	0.0	0.0	25.0	100.0	0.0	0.0	0.0	0.0	0.0	0.0
1232	0.0	0.0	0.0	0.0	0.0	100.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
1233	0.0	0.0	0.0	0.0	0.0	96.4	0.0	2.2	1.5	0.0	50.0	0.0	50.0	0.0	0.0	0.0	0.0	0.0	50.0	50.0	0.0	0.0	0.0
1236	0.0	0.0	0.0	0.0	0.0	100.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
1240	0.0	0.0	0.0	0.0	0.0	99.2	0.0	0.0	0.8	0.0	100.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	100.0	0.0	0.0	0.0	0.0
1241	0.0	0.0	0.0	0.0	0.0	88.5	0.0	0.0	11.5	0.0	26.7	13.3	6.7	26.7	13.3	13.3	46.7	40.0	0.0	13.3	0.0	0.0	0.0
1245	0.0	0.0	0.0	0.0	0.0	99.3	0.0	0.0	0.8	0.0	0.0	100.0	0.0	0.0	0.0	0.0	0.0	100.0	0.0	0.0	100.0	0.0	0.0
1251	0.0	0.0	0.0	0.0	0.0	97.9	0.0	0.7	1.4	0.0	0.0	0.0	0.0	0.0	50.0	50.0	0.0	50.0	0.0	50.0	0.0	0.0	0.0

Instance	A24	A25	A26	A27	A28	A29	A30	A31	A32	A33	A34	A35	A36	A37	A38	A39	A40	A41	A42	Label
1178	0.0	0.0	0.1	0.0	0.3	0.1	0.0	53.4	62.9	78.3	0.0	0.0	0.0	0.0	0.0	100.0	100.0	0.0	0	Tidak Lulus
1184	0.0	0.0	0.2	0.0	0.3	0.3	0.1	51.8	62.8	78.3	0.0	0.0	0.0	0.0	0.0	100.0	100.0	0.0	0	Tidak Lulus
1190	0.0	0.0	0.0	0.0	0.0	0.0	0.0	17.4	53.8	75.8	0.0	0.0	0.0	0.0	0.0	0.0	0.0	100.0	0	Tidak Lulus
1192	0.0	0.0	0.1	0.0	0.2	0.1	0.0	32.2	52.7	82.1	0.0	0.0	0.0	0.0	0.0	100.0	100.0	0.0	1	Lulus
1196	0.0	0.0	0.2	0.0	0.2	0.4	0.0	28.3	53.8	75.5	0.0	0.0	0.0	0.0	0.0	100.0	100.0	0.0	1	Lulus
1199	0.0	0.0	0.1	0.0	0.2	0.2	0.0	38.5	52.9	86.5	0.0	0.0	0.0	0.0	0.0	77.8	61.1	38.9	1	Lulus
1203	0.0	0.0	0.4	0.0	0.7	0.8	0.0	51.8	62.8	78.3	0.0	0.0	0.0	0.0	0.0	100.0	100.0	0.0	0	Tidak Lulus
1206	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0	Tidak Lulus
1210	0.0	0.0	0.2	0.0	0.2	0.5	0.0	18.8	61.7	65.9	0.0	0.0	0.0	0.0	0.0	100.0	100.0	0.0	1	Lulus
1211	0.0	0.0	0.4	0.0	0.9	0.6	0.0	53.4	62.9	78.3	0.0	0.0	0.0	0.0	0.0	100.0	100.0	0.0	0	Tidak Lulus
1213	0.0	0.0	0.1	0.0	0.1	0.1	0.0	37.7	75.9	63.1	0.0	0.0	0.0	0.0	0.0	100.0	100.0	0.0	0	Tidak Lulus
1215	0.0	0.0	0.0	0.0	0.0	0.0	0.0	33.8	58.6	70.1	0.0	0.0	0.0	0.0	0.0	100.0	100.0	0.0	0	Tidak Lulus
1225	0.0	0.0	0.0	0.0	0.0	0.0	0.2	30.5	48.9	85.5	0.0	0.0	0.0	0.0	0.0	100.0	100.0	0.0	1	Lulus
1226	0.0	0.0	0.6	0.0	1.4	0.7	0.3	28.3	53.8	75.5	0.0	0.0	0.0	0.0	0.0	100.0	83.3	16.7	1	Lulus
1228	0.0	0.0	0.0	0.0	0.1	0.0	0.0	18.8	61.7	65.9	0.0	0.0	0.0	0.0	0.0	83.3	55.6	44.4	1	Lulus
1229	0.0	0.0	0.3	0.0	0.8	0.3	0.2	17.4	53.8	75.8	0.0	0.0	0.0	0.0	0.0	100.0	100.0	0.0	1	Lulus
1230	0.0	0.0	0.9	0.0	1.2	2.1	0.3	39.6	65.5	60.3	0.0	0.0	0.0	0.0	0.0	100.0	100.0	0.0	1	Lulus
1232	0.0	0.0	0.8	0.0	1.0	2.3	0.0	51.8	62.8	78.3	0.0	0.0	0.0	0.0	0.0	100.0	100.0	0.0	1	Lulus
1233	0.0	0.0	6.7	8.0	3.6	6.5	8.9	37.7	75.9	63.1	0.0	0.0	0.0	0.0	0.0	100.0	100.0	0.0	1	Lulus
1236	0.0	0.0	0.5	0.0	1.2	0.7	0.0	32.2	52.7	82.1	0.0	0.0	0.0	0.0	0.0	100.0	100.0	0.0	1	Lulus
1240	0.0	0.0	0.8	0.0	1.7	1.4	0.0	17.4	53.8	75.8	0.0	0.0	0.0	0.0	0.0	100.0	100.0	0.0	1	Lulus
1241	0.0	0.0	0.3	0.0	0.1	1.2	0.1	30.5	48.9	85.5	0.0	0.0	0.0	0.0	0.0	100.0	83.3	16.7	1	Lulus
1245	0.0	0.0	5.2	0.0	5.7	14.4	1.5	18.8	61.7	65.9	0.0	0.0	0.0	0.0	0.0	100.0	100.0	0.0	1	Lulus
1251	0.0	0.0	0.4	0.0	1.0	0.5	0.0	51.8	62.8	78.3	0.0	0.0	0.0	0.0	0.0	100.0	100.0	0.0	1	Lulus

Lampiran 2
Atribut pada data akademik Universitas XYZ

No	Atribut Bulan Awal	Atribut Bulang Tengah Semester	Atribut Akhir Semester
1	Id_Mahasiswa	Id_Mahasiswa	Id_Mahasiswa
2	Label	Label	Label
3	Post_0_6	Post_0_6	Post_0_6
4	Post_7_12	Post_7_12	Post_7_12
5	Post_13_18	Post_13_18	Post_13_18
6	Post_18_24	Post_18_24	Post_18_24
7	Prosen_Hadir	Prosen_Hadir	Prosen_Hadir
8	Prosen_Ijin	Prosen_Ijin	Prosen_Ijin
9	Prosen_Sakit	Prosen_Sakit	Prosen_Sakit
10	Prosen_Alfa	Prosen_Alfa	Prosen_Alfa
11	Prosen_Dispensasi	Prosen_Dispensasi	Prosen_Dispensasi
12	Prosen_Alfa_Senin	Prosen_Alfa_Senin	Prosen_Alfa_Senin
13	Prosen_Alfa_Selasa	Prosen_Alfa_Selasa	Prosen_Alfa_Selasa
14	Prosen_Alfa_Rabu	Prosen_Alfa_Rabu	Prosen_Alfa_Rabu
15	Prosen_Alfa_Kamis	Prosen_Alfa_Kamis	Prosen_Alfa_Kamis
16	Prosen_Alfa_Jumat	Prosen_Alfa_Jumat	Prosen_Alfa_Jumat
17	Prosen_Alfa_Sabtu	Prosen_Alfa_Sabtu	Prosen_Alfa_Sabtu
18	Prosen_Alfa_Slot_Pagi	Prosen_Alfa_Slot_Pagi	Prosen_Alfa_Slot_Pagi
19	Prosen_Alfa_Slot_Jelang_Siang	Prosen_Alfa_Slot_Jelang_Siang	Prosen_Alfa_Slot_Jelang_Siang
20	Prosen_Alfa_Slot_Siang	Prosen_Alfa_Slot_Siang	Prosen_Alfa_Slot_Siang
21	Prosen_Alfa_Slot_Sore	Prosen_Alfa_Slot_Sore	Prosen_Alfa_Slot_Sore
22	Prosen_Alfa_Internal	Prosen_Alfa_Internal	Prosen_Alfa_Internal
23	Frek_Pemb_Mhs	Frek_Pemb_Mhs	Frek_Pemb_Mhs
24	Frek_Akademik	Frek_Akademik	Frek_Akademik
25	Frek_Kes_Mahasiswa	Frek_Kes_Mahasiswa	Frek_Kes_Mahasiswa
26	Frek_Adm_Akademik	Frek_Adm_Akademik	Frek_Adm_Akademik
27	Durasi_Internet	Durasi_Internet	Durasi_Internet
28	Durasi_Internet_0_6	Durasi_Internet_0_6	Durasi_Internet_0_6
29	Durasi_Internet_6_12	Durasi_Internet_6_12	Durasi_Internet_6_12
30	Durasi_Internet_12_18	Durasi_Internet_12_18	Durasi_Internet_12_18
31	Durasi_Internet_18_24	Durasi_Internet_18_24	Durasi_Internet_18_24
32	Bobot_Mk_Sambung	Bobot_Mk_Sambung	Bobot_Mk_Sambung
33	Bobot_Mk_Per_Hari	Bobot_Mk_Per_Hari	Bobot_Mk_Per_Hari
34	Rerata_Jml_Hari_Kuliah_Seminggu	Rerata_Jml_Hari_Kuliah_Seminggu	Rerata_Jml_Hari_Kuliah_Seminggu
35		Prosen_Mk_Atas2	Jml_Ukm
36		Prosen_Mk_Bawah2	Poin_Ukm
37		Prosen_Mk_Blm_Ada	Jml_Tak
38			Poin_Tak
39			Jml_Prestasi_Mhs
40			Prosen_Sks_Lulus
41			Prosen_Jml_Mk_Atas2
42			Prosen_Jml_Mk_Bawah2

Lampiran 3
Contoh Data yang Digunakan

Contoh Data yang Digunakan Semester 1 Bulan ke 1

StudentId	Aktivitas Internet	Aktivitas Blog	Aktivitas Lomba & Panitia	Presensi	...
1	3.24	7.41	0	97	...
2	1.04	0	0	100	...
3	0	0	0	83.33	...
....	....	....	....	....	....

Data Akademik

Contoh Data yang Digunakan Semester 1 Bulan ke 2

StudentId	Aktivitas Internet	Aktivitas Blog	Aktivitas Lomba & Panitia	Presensi	...
1	3.23	6	0	92.16	...
2	2.05	0	0	96.88	...
3	0	0	0	94.12	...
....	....	....	....	....	....

Data Akademik

Contoh Data yang Digunakan Semester 1 Bulan ke 3

Student Id	Aktivitas Internet	Aktivitas Blog	Aktivitas Lomba & Panitia	Presensi	...	% MK Atas 2	% MK Bawah 2	% MK Belum Ada
1	2,79	5,2	0	94,78	...	55,56	22,22	22,22
2	1,57	0	0	97,78	...	76,47	23,53	0,00
3	0	0	0	94,12	...	56,52	43,48	0,00
...	...	...	...	...	...	...	...	...

Data Akademik + Nilai UTS

Contoh Data yang Digunakan Semester 1 Bulan ke 4

Student Id	Aktivitas Internet	Aktivitas Blog	Aktivitas Lomba & Panitia	Presensi	...	% MK Atas 2	% MK Bawah 2	% MK Belum Ada
1	2,95	5,2	0	96,80	...	55,56	22,22	22,22
2	1,38	0	0	98,57	...	76,47	23,53	0,00
3	0	0	0	93,75	...	56,52	43,48	0,00
...	...	...	...	...	...	...	...	...

Data Akademik + Nilai UTS

Contoh Data yang Digunakan Semester 1 Bulan ke 5

Student Id	Aktivitas Internet	Aktivitas Blog	Aktivitas Lomba & Panitia	Presensi	...	Poin TAK	%SKS Lulus	% MK Bawah 2	% MK Belum Ada
1	3,24	7,4	0	97,01	...	0	77,78	77,78	22,22
2	1,29	0	0	98,73	...	11,11	100	100	0
3	0	0	0	94,52	...	4,44	100	100	0
...	...	...	...	...	...	...	...	...	...

Data Akademik + TAK + Nilai Akhir

Lampiran 4
Hasil Evaluasi Pengujian Parameter pada Dataset Keel

Parameter	Deskripsi Parameter	Nilai default	Nilai range ujicoba	Nilai optimal
Chtresh	Threshold pada *agglomerative hierarchical clustering*	Range: 0.2 – 7 Nilai optimal pada paper: 1	Dataset Keel: 0,3; 1;1,5	Dataset Keel: 1
NN	Ketetanggaan saat menentukan bobot pada setiap data minoritas	Range: 1-15 Nilai optimal pada paper 3-7	Dataset Keel: 3,5,7	Dataset Keel: 5
NumIteration	Jumlah iterasi yang diinputkan user untuk memastikan bahwa hasil akurasi setiap iterasi tidak berbeda signifikan	Nilai pada paper 3	Dataset Keel: 5	Dataset Keel: 5

1. Parameter Cthresh pada Dataset Keel

Cthresh=0.3								
Matriks	**Ecoli1**	**Glass0**	**Haberman**	**Iris0**	**Segment0**	**Vehicle0**	**Wisconsin**	**Rata-rata**
Akurasi	92,00	82,81	78,02	100	99,42	90,51	94,61	85,90
Precision	97,59	88,10	84	100	100	94	96,95	77,25
Recall	93,10	86,05	88,73	100	99,32	94	94,78	91,33
Fmeasure	95,29	87,06	86,3	100	99,66	94	95,85	90,38

Cthresh=1								
Matriks	**Ecoli1**	**Glass0**	**Haberman**	**Iris0**	**Segment0**	**Vehicle0**	**Wisconsin**	**Rata-rata**
Akurasi	**92,00**	**81,25**	**73,63**	**100**	**99,13**	**94,86**	**94,61**	**87,18**
Precision	**97,59**	**89,74**	**81,33**	**100**	**100**	**96,06**	**94,90**	**91,28**
Recall	**93,10**	**81,40**	**85,92**	**100**	**99,00**	**97,50**	**97,00**	**90,70**
Fmeasure	**95,30**	**85,37**	**83,56**	**100**	**99,50**	**96,77**	**95,94**	**90,96**
Cthresh=1.5								
Matriks	**Ecoli1**	**Glass0**	**Haberman**	**Iris0**	**Segment0**	**Vehicle0**	**Wisconsin**	**Rata-rata**
Akurasi	90,00	81,25	71,43	100	98,41	89,72	94,61	84,88
Precision	96,39	87,80	80	100	99,83	94,39	96,95	89,97
Recall	91,95	83,72	84,51	100	98,3	92,5	94,78	88,86
Fmeasure	94,12	85,71	82,19	100	99,06	93,43	95,85	89,39

2. Parameter NN pada Dataset Keel

NN=3								
Matriks	**Ecoli1**	**Glass0**	**Haberman**	**Iris0**	**Segment0**	**Vehicle0**	**Wisconsin**	**Rata-rata**
Akurasi	91,00	76,56	68,13	100	98,27	91,7	94,12	83,91
Precision	97,56	81,82	77,63	100	99,66	95,9	94,85	88,55
Recall	91,95	83,72	83,1	100	98,4	93,5	96,27	89,63
Fmeasure	94,67	82,76	80,27	100	98,97	94,68	95,56	89,02
NN=5								

Matriks	Ecoli1	Glass0	Haberman	Iris0	Segment0	Vehicle0	Wisconsin	Rata-rata
Akurasi	**92,00**	**81,25**	**73,63**	**100**	**99,13**	**94,86**	**94,61**	**87,18**
Precision	**97,59**	**89,74**	**81,33**	**100**	**100,00**	**96,06**	**94,90**	**91,28**
Recall	**93,10**	**81,40**	**85,92**	**100**	**99,00**	**97,5**	**97,00**	**90,70**
Fmeasure	**95,30**	**85,37**	**83,56**	**100**	**99,50**	**96,77**	**95,94**	**90,96**
NN=7								
Matriks	**Ecoli1**	**Glass0**	**Haberman**	**Iris0**	**Segment0**	**Vehicle0**	**Wisconsin**	**Rata-rata**
Akurasi	92,00	79,69	59,34	100	99,13	94,07	93,63	84,26
Precision	97,59	84,09	78,33	100	100	96,48	94,16	78,85
Recall	93,10	86,05	66,2	100	98,98	96	96,27	87,64
Fmeasure	95,29	85,06	71,76	100	99,49	96,24	95,2	88,67

3. Parameter Iterasi pada Dataset Keel

Ecoli1 (K-FOLD=3, Cthresh=1, NN=5)							
Matriks	**Original**	**Iterasi1**	**Iterasi2**	**Iterasi3**	**Iterasi4**	**Iterasi5**	**Rata-rata**
Akurasi	88,06	91,04	92,54	94,03	91,04	91,04	91,938
Precision	90,74	94,12	94,23	96,08	94,12	94,12	94,534
Recall	94,23	94,12	96,08	96,08	94,12	94,12	94,904
Fmeasure	92,45	94,12	95,15	96,08	94,12	94,12	94,718

Glass0 (K-FOLD=3, Cthresh=1, NN=5)							
Matriks	**Original**	**Iterasi1**	**Iterasi2**	**Iterasi3**	**Iterasi4**	**Iterasi5**	**Rata-rata**
Akurasi	61,9	80,95	59,52	64,29	64,29	61,9	66,19
Precision	71,43	85,71	72	74,07	74,07	73,08	75,786
Recall	71,43	85,71	64,29	71,43	71,43	67,86	72,144
Fmeasure	71,43	85,71	67,92	72,73	72,73	70,37	73,892
Haberman (K-FOLD=3, Cthresh=1, NN=5)							
Matriks	**Original**	**Iterasi1**	**Iterasi2**	**Iterasi3**	**Iterasi4**	**Iterasi5**	**Rata-rata**
Akurasi	68	67,21	67,21	73,77	63,93	70,49	68,522
Precision	79,55	77,78	77,78	79,59	76,74	82,93	78,964
Recall	77,78	77,78	77,78	86,67	73,33	75,56	78,224
Fmeasure	78,65	77,78	77,78	82,98	75	79,07	78,522
Iris0(K-FOLD=3, Cthresh=1, NN=5)							
Matriks	**Original**	**Iterasi1**	**Iterasi2**	**Iterasi3**	**Iterasi4**	**Iterasi5**	**Rata-rata**
Akurasi	100	100	100	100	100	100	100
Precision	100	100	100	100	100	100	100

Recall	100	100	100	100	100	100	100
Fmeasure	100	100	100	100	100	100	100
Segment0 (K-FOLD=3, Cthresh=1, NN=5)							
Matriks	**Original**	**Iterasi1**	**Iterasi2**	**Iterasi3**	**Iterasi4**	**Iterasi5**	**Rata-rata**
Akurasi	98,92	99,78	99,57	99,78	99,78	99,78	99,738
Precision	99,49	99,75	99,5	99,75	99,75	99,75	99,70
Recall	99,24	100	100	100	100	100	100
Fmeasure	99,37	99,87	99,75	99,87	99,87	99,87	99,846
Vehicle0 (K-FOLD=3, Cthresh=1, NN=5)							
Matriks	**Original**	**Iterasi1**	**Iterasi2**	**Iterasi3**	**Iterasi4**	**Iterasi5**	**Rata-rata**
Akurasi	88,17	91,12	90,35	90,53	90,53	91,2	90,746
Precision	89,78	93,18	93,8	92,48	92,48	93,18	93,024
Recall	95,35	95,35	93,8	95,35	95,35	95,35	95,04
Fmeasure	92,48	94,25	93,8	93,89	93,89	94,25	94,016
Wisconsin (K-FOLD=3, Cthresh=1, NN=5)							
Matriks	**Original**	**Iterasi1**	**Iterasi2**	**Iterasi3**	**Iterasi4**	**Iterasi5**	**Rata-rata**

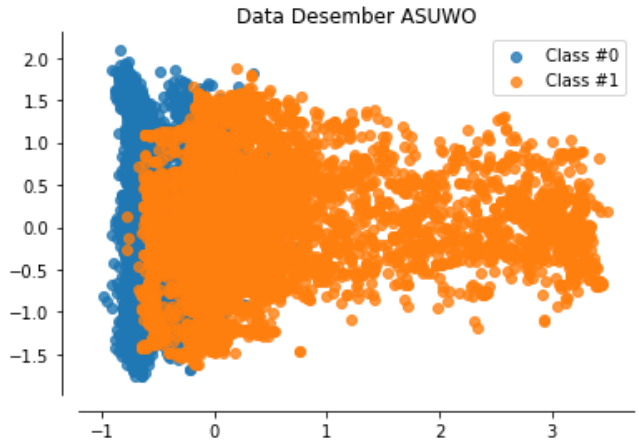
Akurasi	91,65	91,91	91,91	91,91	91,91	91,91	91,91
Precision	92,33	92,31	92,31	92,31	93,26	93,26	92,69
Recall	94,45	95,45	95,45	95,45	94,32	94,32	94,998
Fmeasure	93,38	93,85	93,85	93,85	93,79	93,79	93,826

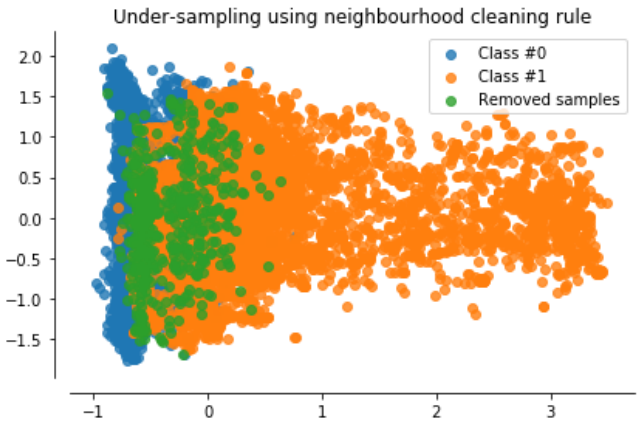
Lampiran 5

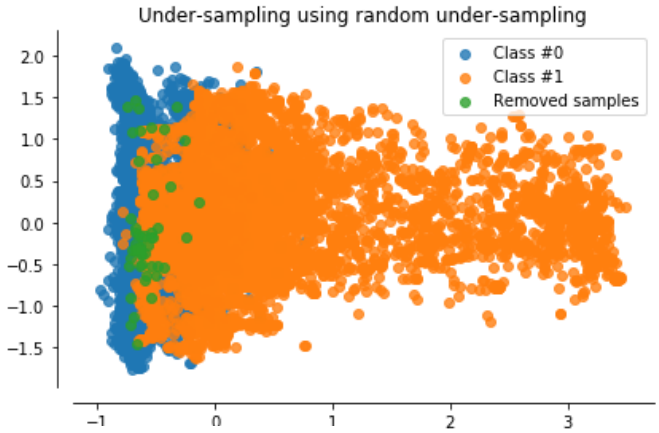
Analisa Data Akademik Lebih Lanjut pada Data Akademik dalam 1 Iterasi

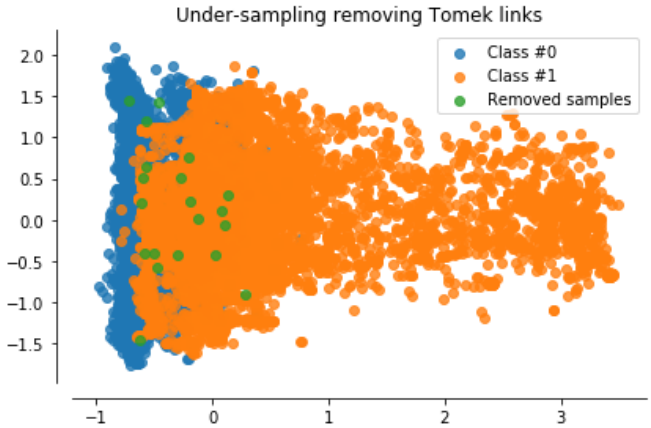
1. Desember

	Data Train Asuwo	Data Train NCL	Data Train Tomek	Data Train RUS
0=Major	7403	6924	7383	7353
1=Minor	7352	7352	7352	7352
Jumlah Data	14755	14276	14735	14705
Data yang dihapus/ ditambahkan	4472	479	20	50



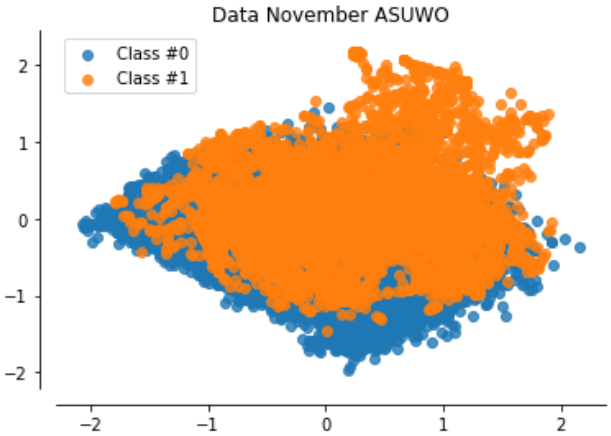


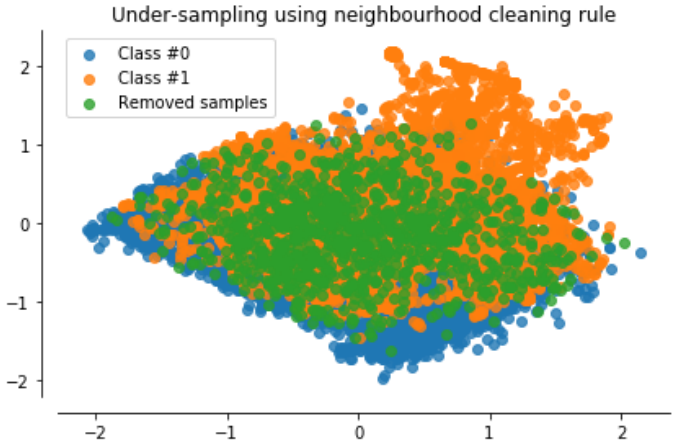


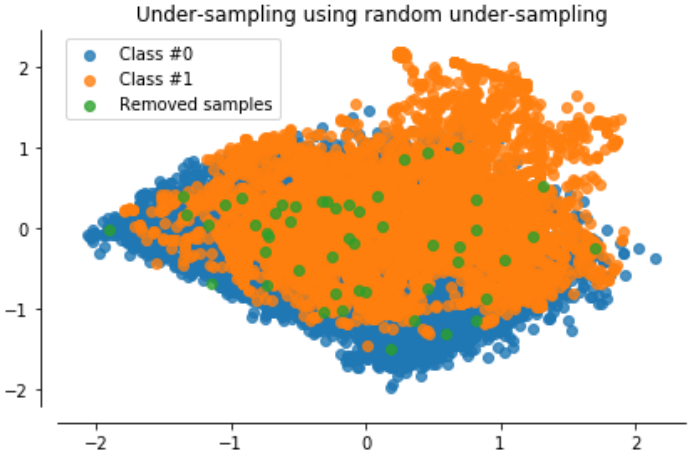


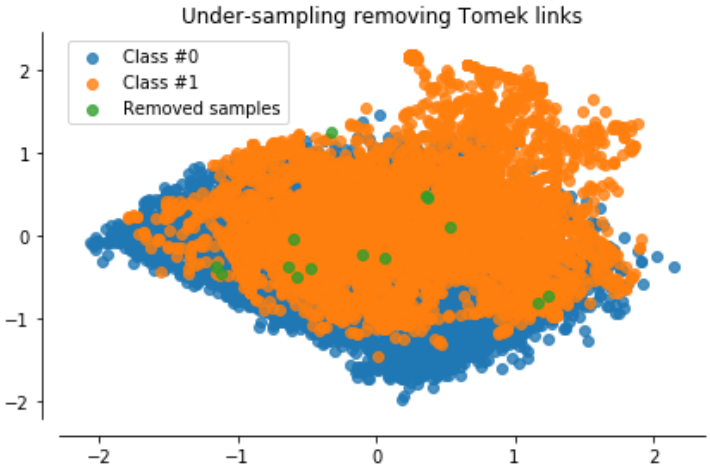
2. November

	Data Train Asuwo	Data Train NCL	Data Train Tomek	Data Train RUS
0=Major	7225	6215	7211	7175
1=Minor	7174	7174	7147	7174
Jumlah Data	14399	13389	14358	14349
Data yang dihapus/ ditambahkan	4116	1011	15	51



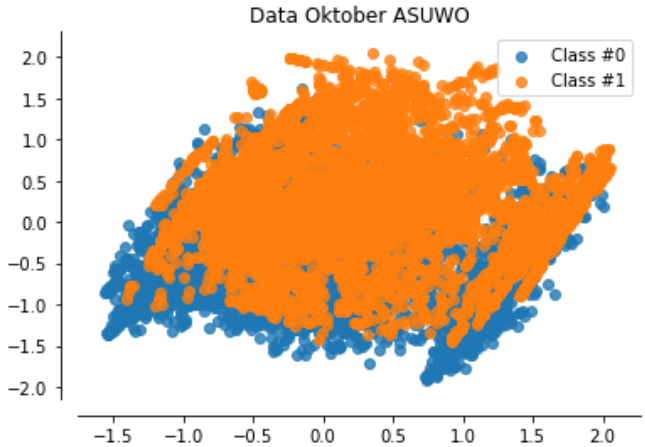


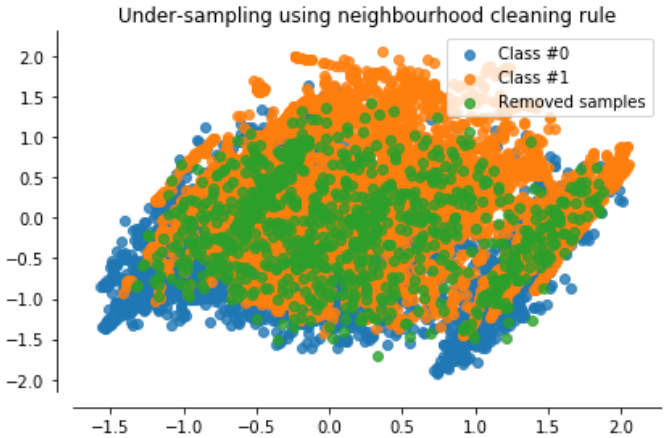


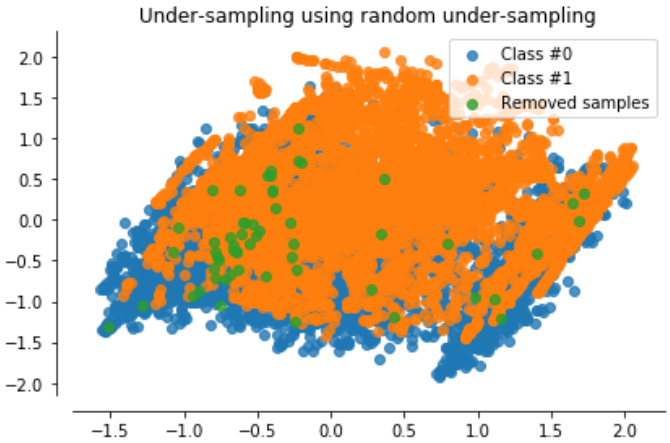


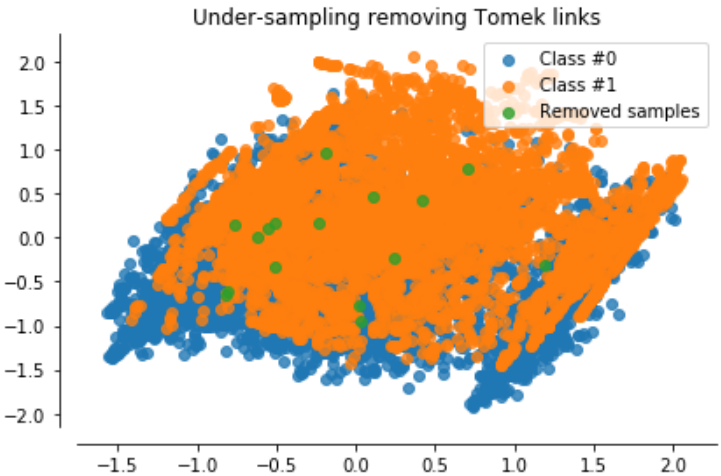
3. Oktober

	Data Train Asuwo	Data Train NCL	Data Train Tomek	Data Train RUS
0=Major	7190	6306	7192	7151
1=Minor	7150	7150	7150	7150
Jumlah Data	14358	13456	14342	14301
Data yang dihapus/ ditambahkan	4075	903	17	58



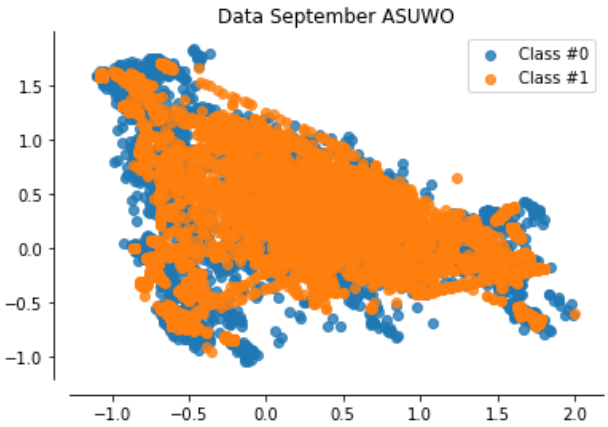


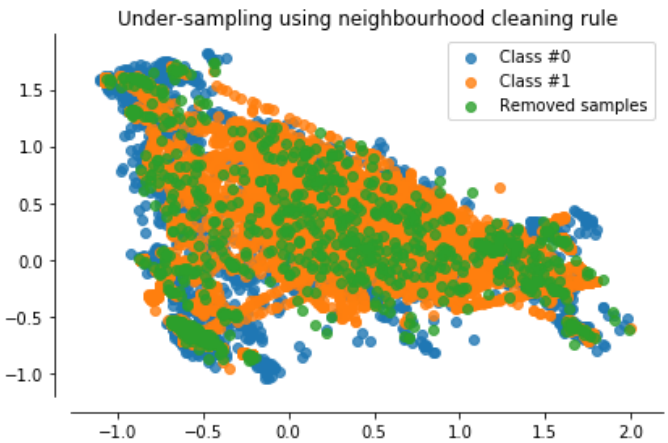


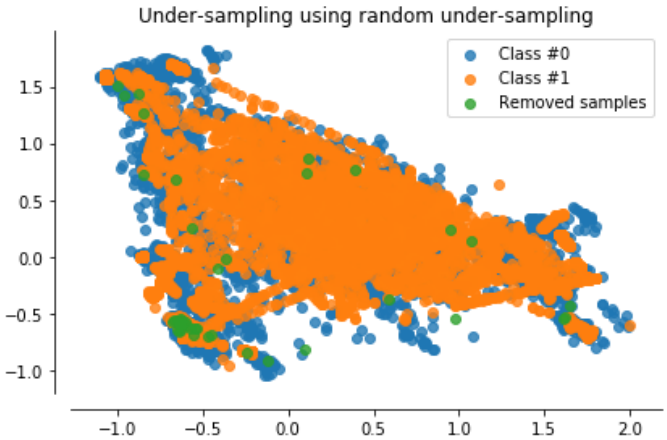


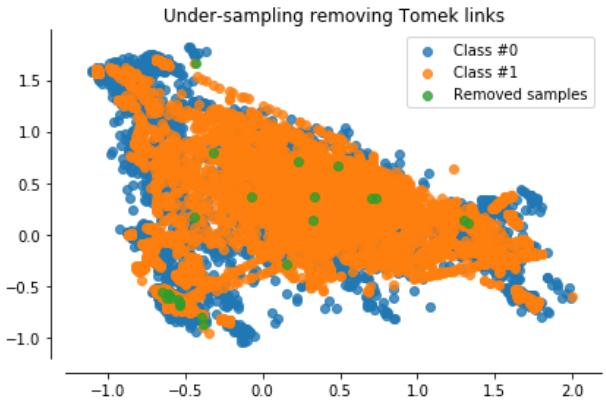
4. September

	Data Train Asuwo	Data Train NCL	Data Train Tomek	Data Train RUS
0=Major	5716	6122	7167	7151
1=Minor	7150	7150	7150	7150
Jumlah Data	13272	14317	14301	2571
Data yang dihapus/ ditambahkan	4058	1070	24	40



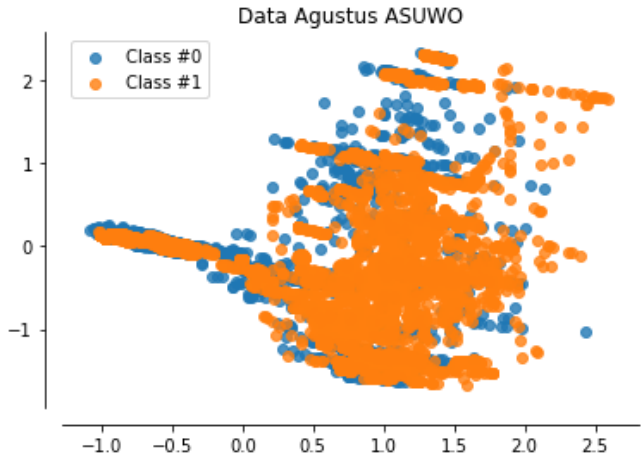


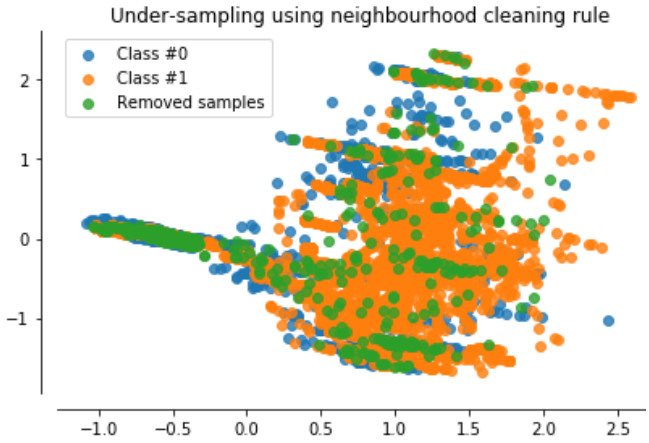


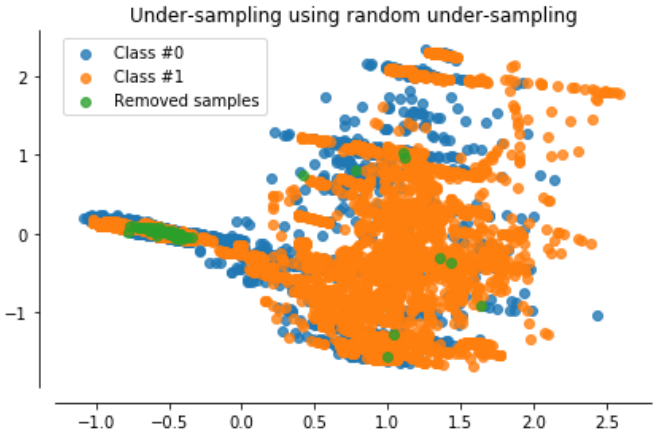


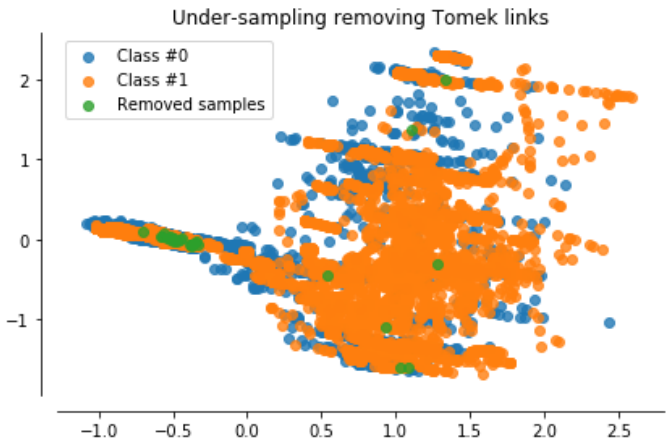
5. Agustus

	Data Train Asuwo	Data Train NCL	Data Train Tomek	Data Train RUS
0=Major	7138	5852	7111	7083
1=Minor	7082	7082	7082	7107
Jumlah Data	14220	13446	14193	14145
Data yang dihapus/ ditambahkan	3937	754	27	75

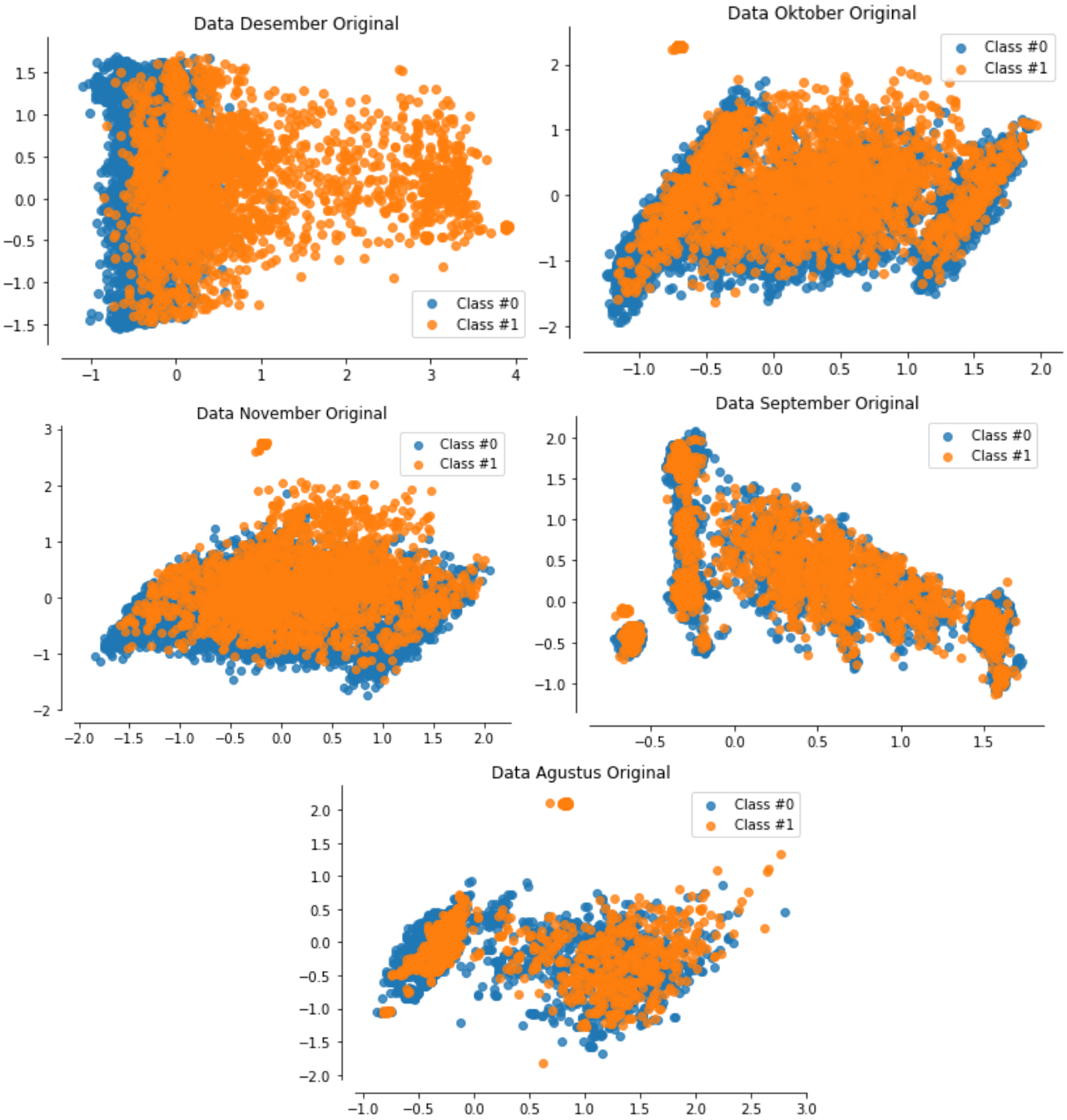








Data Original



[Halaman ini sengaja dikosongkan]

BIODATA PENULIS

Shabrina Choirunnisa, lahir di Jember, 25 April 1993. Penulis merupakan anak pertama dari dua bersaudara dari pasangan Darmadji dan Indah Eko Suryani.

Penulis menempuh pendidikan formal di SDN Mangli 1 Jember (1999 – 2005), SMP Negeri 1 Jember (2005 – 2008), SMA Negeri 1 Jember (2008 – 2011), S-1 Teknik Informatika Institut Teknologi Sepuluh November (ITS) dengan bidang minat Komputasi Cerdas Visual (KCV) pada tahun 2011-2015. Pada tahun 2017 penulis melanjutkan pendidikan Magister di Institut Teknologi Sepuluh Nopember di Departemen Informatika. Pada studi pasca sarjana atau S2, penulis mengambil bidang minat pada topik Dasar dan Terapan Komputasi (DTK). Penulis dapat dihubungi via surel dengan alamat shabrinachnisa@gmail.com