Komparasi Metode SMOTE dan ADASYN dalam Meningkatkan Performa Klasifikasi Herregistrasi Mahasiswa Baru

http://dx.doi.org/10.28932/jutisi.v8i1.4004

Riwayat Artikel
Received: 22 September 2021 | *Final Revision*: 29 Maret 2022 | *Accepted*: 29 Maret 2022

Risky Agung Nurdian[✉ #1], Mujib Ridwan[#2], Ahmad Yusuf[#3]

[#] *Sistem Informasi, Universitas Islam Negeri Sunan Ampel Surabaya*
Jl. Ahmad Yani No.117, Jemur Wonosari, Kec. Wonocolo, Surabaya, Indonesia
[1]riskyagungnurdian@gmail.com
[2]mujibrw@uinsby.ac.id
[3]ahmadyusuf@uinsby.ac.id

Abstract — **Universities annually accept new students at the beginning of the new school year. In the acceptance of prospective students on the Seleksi Prestasi Akademik Nasional Perguruan Tinggi Keagamaan Islam Negeri (SPAN PTKIN) di State Islamic University Of Sunan Ampel Surabaya, many prospective students who do not register will have an impact on income of the State Islamic University Of Sunan Ampel Surabaya institution. If the institution can find out early on the probability of a prospective student who will resign, then the management can take action to retain the prospective student. To overcome this, data mining classification can be carried out. The methods used in this classification are decision trees and naïve bayes. The number of students who did not re register compared to reregister resulted in the data being imbalanced. Data imbalances can affect the accuracy of the classification results. The imbalance of the data used can result in an unsuitable model. The solution to handle the data imbalance is to use the SMOTE and ADASYN oversampling methods. The purpose of this study was to compare performance of the SMOTE and ADASYN methods. The results show that the SMOTE method can balance the data in a balanced way compared to ADASYN. From the test results, the SMOTE method is more suitable to use than the ADASYN method because the ROCAUC SMOTE value is higher than ADASYN.**

Keywords— ***ADASYN; Decision Tree; Naïve Bayes; SMOTE***

I. PENDAHULUAN

Institusi pendidikan merupakan kebutuhan bagi masyarakat, pendidikan merupakan sebuah fenomena kehidupan yang sudah tidak dapat dipisahkan dari kehidupan sosial [1]. Universitas Islam Negeri Sunan Ampel Surabaya merupakan salah satu perguruan tinggi negeri yang berada di kota Surabaya, saat ini memiliki jumlah mahasiswa aktif 20184 yang tersebar pada 9 fakultas dan 64 program studi [2].

Perguruan tinggi setiap tahunnya melakukan penerimaan mahasiswa baru pada awal tahun ajaran baru. Sistem penerimaan mahasiswa baru di Universitas Islam Negeri Sunan Ampel Surabaya saat ini memiliki beberapa jalur seleksi yaitu SPAN PTKIN, SNMPTN, SBMPTN, UMPTKIN, MANDIRI, dan MANDIRI PRESTASI [3]. Dalam penerimaan calon mahasiswa baru jalur SPAN PTKIN di UINSA banyak calon mahasiswa yang tidak melakukan daftar ulang pada tahun 2020 sebanyak 41.9% calon mahasiswa baru yang tidak melakukan daftar ulang [4].

Banyaknya mahasiswa yang tidak melakukan daftar ulang akan berdampak kepada pendapatan lembaga UINSA [5]. Jika, lembaga dapat mengetahui sejak dini dari kemungkinan calon mahasiswa yang akan mengundurkan diri, maka pihak dari manajemen dapat melakukan tindakan untuk mempertahankan calon mahasiswa tersebut. Pendekatan *data mining* dapat dilakukan untuk mengatasi hal tersebut [6].

Klasifikasi merupakan salah satu dari metode *data mining.* Klasifikasi adalah tipe analisis data yang digunakan untuk membantu dalam menentukan kelas label yang ingin di klasifikasi. Klasifikasi merupakan metode *data mining supervised learning* yang mencoba menemukan hubungan antara atribut dan target [7]. Klasifikasi adalah cara pengelompokkan sebuah data berdasarkan dengan karakteristik dan ciri-ciri dari data tersebut. Klasifikasi memiliki beberapa metode *decision*

tree, naïve bayes, support vector machine, fuzzy, dan *neural network* [8]. Algoritma klasifikasi menurut [9] diantaranya adalah *naïve bayes* dan *decision tree.*

Penelitian yang dilakukan oleh [6], dalam penelitian ini menggunakan terdapat ketidakseimbangan data yang mana *class majority* lebih banyak dibandingkan dengan *class minority*. Hasil penelitian memaparkan bahwa algoritma *decision tree C45* memiliki *accuracy* sebesar 80.72%, algoritma *KNN* memiliki *accuracy* 80.46%, dan algoritma *naïve bayes* memiliki *accuracy* sebesar 74.49%. Untuk nilai AUC algoritma *naïve bayes* memiliki nilai AUC sebesar 0.602, algoritma *KNN* memiliki nilai AUC sebesar 0.578, dan algoritma *decision tree C45* memiliki nilai AUC sebesar 0.500.

Penelitian lain yang dilakukan oleh [10], penelitian ini memaparkan bahwa tingkat *accuracy* dari algoritma *bayesian classification* sebesar 78%. Penelitian dengan kasus yang serupa juga dilakukan oleh [11], penelitian ini memaparkan bahwa metode *k-NN* memerlukan proses *testing* yang lama dibandingkan dengan algoritma C4.5. Hal ini disebabkan oleh penentuan bobot variabel dan penentuan dari perbandingan antara nilai di setiap variabel.

Penelitian lain yang dilakukan oleh [5], penelitian ini memaparkan bahwa algoritma *decision tree* memiliki performa yang lebih baik daripada metode lainnya. Dari hasil pengujian yang dilakukan bahwa algoritma *decision tree* memiliki nilai *f-Measure* diatas 80% untuk setiap skenario yang dilakukan. Penelitian yang dilakukan oleh [12], dari hasil penelitian menunjukkan akurasi *decision tree* sebesar 88.74% dan algoritma *naïve bayes* sebesar 87.24%.

Dari penelitian terdahulu mengenai klasifikasi untuk menganalisis heregistrasi mahasiswa yang mana penelitian tersebut menggunakan dataset yang memiliki ketidak seimbangan kelas. Hal tersebut dapat mengakibatkan pada kinerja klasifikasi tiap algoritma. Penelitian terdahulu memaparkan bahwa *preprocessing* data memiliki dampak yang signifikan dalam *performance* metode klasifikasi *data mining* [13], [14]. Seperti menghapus jarak yang terlalu menyimpang, memiliki terlalu banyak nilai nol [14]. *Preprocessing* data mengacu pada pembersihan data, integrasi data, transisi data, dan reduksi data diproses sebelum implementasi algoritma *data mining* [15].

Dalam penelitian ini data daftar ulang calon mahasiswa Universitas Islam Negeri Sunan Ampel Surabaya yang mana data tersebut memiliki *class imbalanced*. Dimana kelas *majority* lebih besar dibandingkan dengan kelas *minority*. *Class imbalanced* dapat mempengaruhi hasil keakuratan klasifikasi. *Imbalanced data* yang digunakan dapat menghasilkan model yang tidak cocok. Jika, *imbalanced data* digunakan dalam algoritma klasifikasi maka akan menghasilkan akurasi yang jauh lebih tinggi untuk kelas mayoritas daripada kelas minoritas [16]. Solusi untuk mengatasi *imbalanced data* yaitu menggunakan metode *oversampling* SMOTE dan ADASYN. SMOTE adalah salah satu algoritma yang menangani *imbalanced data*. Metode *oversampling* pada kelas minoritas dapat membantu mencapai kinerja klasifikasi yang lebih baik [17]. ADASYN (*adaptive synthetic*) merupakan metode pendekatan *sampling* pada pembelajaran *imbalanced data* [18].

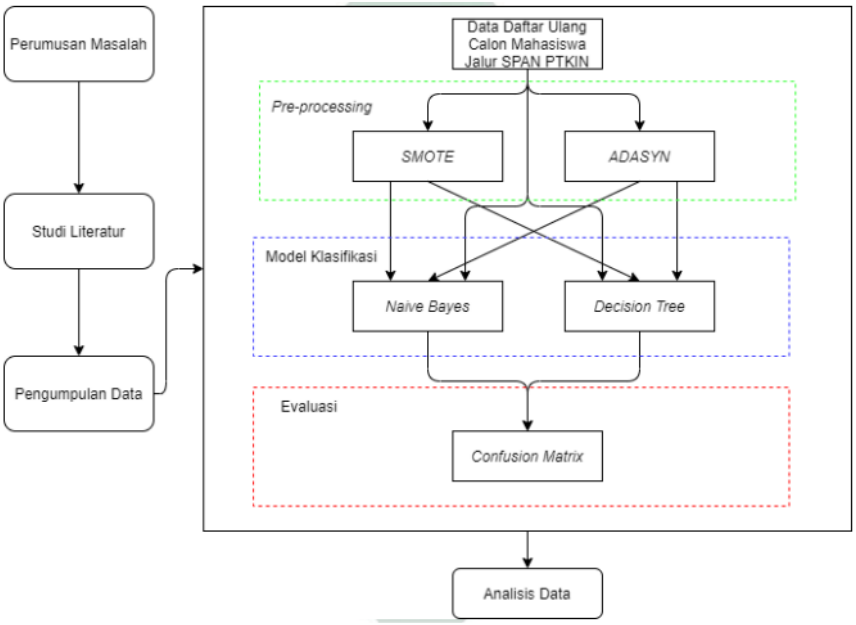
Penelitian yang dilakukan oleh [19], hasil dari penelitian memaparkan bahwa dari semua skenario yang dilakukan *SMOTE + decision tree* memiliki performa klasifikasi yang lebih baik dibandingkan algoritma *decision tree* tanpa adanya *SMOTE* nilai akurasi yang dihasilkan algoritma *decision tree* 65.97% dan nilai akurasi yang dihasilkan algoritma *SMOTE + decision tree* sebesar 71.12%. Penelitian lain yang dilakukan oleh [20], hasil penelitian memaparkan bahwa penggunaan algoritma *SMOTE* dapat digunakan untuk menangani ketidakseimbangan data. Hasil dari pengujian menunjukkan bahwa nilai *AUC naïve bayes* sebesar 0.69 dan nilai *AUC SMOTE + naïve bayes* sebesar 0.75. Penelitian lain yang dilakukan oleh [18], hasil pengujian nilai *confusion matrix* didapatkan yang terbaik setelah dilakukan *ADASYN oversampling* dengan akurasi sebesar 0.9

Penelitian ini mengenai klasifikasi data daftar ulang calon mahasiswa Universitas Islam Negeri Sunan Ampel Surabaya akan membandingkan metode SMOTE dan ADASYN pada algoritma *naïve bayes* dan *decision tree*. Tujuan dari penelitian ini membandingkan hasil klasifikasi performa metode SMOTE dan ADASYN dalam menangani *imbalance data* pada data daftar ulang calon mahasiswa Universitas Islam Negeri Sunan Ampel Surabaya.

II. METODE PENELITIAN

Metode penelitian dimulai dari perumusan masalah, studi literatur dari berbagai sumber jurnal, buku, internet dan lainnya. Pengumpulan data, data yang digunakan untuk penelitian data daftar ulang jalur SPAN PTKIN calon mahasiswa Universitas Islam Negeri Sunan Ampel Surabaya pada tahun 2019, 2020, dan 2021. Data tersebut dilakukan *preprocessing data* menggunakan algoritma SMOTE dan ADASYN untuk mengatasi *class imbalanced* sebelum dilakukan proses klasifikasi.

Dalam penelitian ini proses klasifikasi ada 6 pengujian untuk setiap *dataset* yaitu menggunakan metode *naïve bayes, decision tree,* SMOTE *naïve bayes,* SMOTE *decision tree,* ADASYN *naïve bayes,* dan ADASYN *decision tree.* Untuk mengetahui performa tiap algoritma maka dilakukan evaluasi menggunakan *confusion matrix* untuk melihat algoritma terbaik dalam penelitian ini. Proses terakhir yaitu analisis data untuk mengambil kesimpulan. Metode penelitian ini sesuai dengan Gambar 1.



Gambar 1. Metode Penelitian

Dalam proses klasifikasi hal pertama yang harus dilakukan adalah menentukan atribut data yang hendak dijadikan *class*. Berdasarkan data yang telah dikumpulkan data tersebut memiliki 11 atribut dan atau satu *class label.* Tabel 1 struktur data dari *dataset.*

TABEL 1

STRUKTUR DATA

Atribut	Jenis Variabel	Contoh
Nama Prodi	Nominal	"Ekonomi Syariah","Hukum Ekonomi Syariah"
Skor Peserta	Numerik	"550,612","396,964"
Akreditasi	Ordinal	"A","B"
Pekerjaan Ayah	Nominal	"Petani","Karyawan"
Pekerjaan Ibu	Nominal	"Petani","Tidak Ada"
Penghasilan Ayah	Interval	"1.000.000 s/d 2.000.000"
Penghasilan Ibu	Interval	"1.000.000 s/d 2.000.000"
Jenis Sekolah	Nominal	"SMA","MA"
Kepemilikan	Nominal	"Negeri","Swasta"
Kabupaten	Nominal	"KABUPATEN BOJONEGORO, KABUPATEN LAMONGAN"
Provinsi	Nominal	"JAWA TIMUR","RIAU"
Rumpun Prodi	Nominal	"Bahasa", "Agama"
Daftar Ulang	Biner	"1","0"

Setelah data yang digunakan diperoleh, langkah selanjutnya adalah pengolahan data. Penelitian ini menggunakan 3 tahap pengolahan data yaitu *pre-processing,* klasifikasi, dan *evaluating.*

1. *Pre-processing*

   Pada tahap ini *atribut* yang memiliki korelasi lemah dengan *class* atau *label* berdasarkan dari nilai korelasi yang dihitung dengan menggunakan *pearson correlation* akan dibuang atau tidak digunakan. Selanjutnya akan dilakukan pemilihan fitur dan *class* (label) hal ini dilakukan untuk mempersingkat dalam proses eksekusi model. Tahapan proses kedua yaitu *oversampling* dengan menggunakan dua metode SMOTE dan ADASYN. Algoritma SMOTE menyeimbangkan jumlah distribusi data pada sampel kelas data minoritas dengan jumlah sampel data kelas mayoritas [19]. Algoritma ADASYN menggunakan bobot distribusi pada data kelas minoritas berdasarkan pada tingkat kesulitan pembelajaran data oleh model, yang mana data sintesis dihasilkan dari kelas minoritas yang susah untuk belajar dibandingkan dengan data mayoritas yang lebih mudah untuk melakukan belajar [18]. Dimana data akan

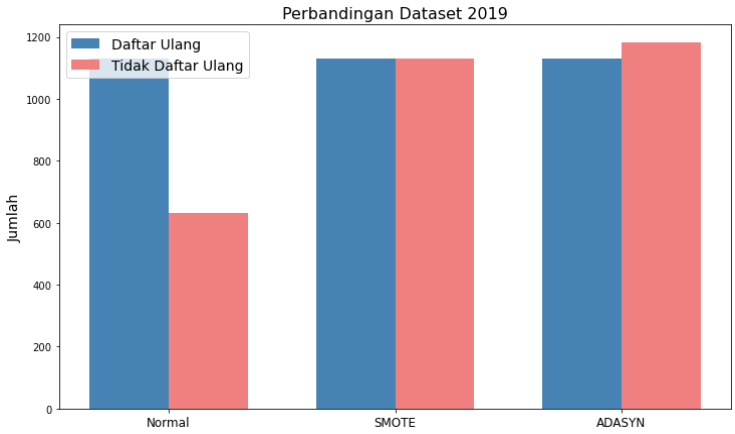
diperbanyak dengan cara membuat sampel baru yang memiliki karakteristik sama dengan sampel yang lama/ sudah ada yang bertujuan untuk menyeimbangkan data tiap *class label*.

2. Proses Klasifikasi
   Pada tahap ini akan menentukan keberhasilan dalam penelitian ini. Tujuan dari proses klasifikasi sendiri adalah melakukan perhitungan *dataset* daftar ulang calon mahasiswa UINSA jalur SPAN PTKIN yang sudah melalui tahap *pre-processing* dengan menggunakan algoritma *naïve bayes* dan *decision tree*. Algoritma *naïve bayes* menghitung satu set probabilitas dengan menghitung frekuensi dan kombinasi nilai dari kumpulan data [21]. Algoritma *decision tree* menggunakan pohon keputusan untuk melakukan klasifikasi [18]. Dalam penelitian ini menggunakan 6 skenario pengujian *naïve bayes, decision tree, SMOTE naïve bayes, SMOTE decision tree, ADASYN naïve bayes,* dan *ADASYN decision tree.*
3. Evaluasi
   Metode yang digunakan dalam proses ini adalah *confusion matrix*, dengan dilakukan 6 skenario untuk kebutuhan pengujian algoritma *naïve bayes* dan *decision tree.*
   1. Skenario pertama akan dilakukan evaluasi terhadap klasifikasi tanpa dilakukan metode *oversampling* menggunakan algoritma *naïve bayes.*
   2. Skenario kedua akan dilakukan evaluasi terhadap klasifikasi tanpa dilakukan metode *oversampling* menggunakan algoritma *decision tree.*
   3. Skenario ketiga akan dilakukan evaluasi terhadap klasifikasi metode *naïve bayes* dengan dilakukan *oversampling SMOTE.*
   4. Skenario keempat akan dilakukan evaluasi terhadap klasifikasi metode *decision tree* dengan dilakukan *oversampling SMOTE.*
   5. Skenario kelima akan dilakukan evaluasi terhadap klasifikasi metode *naïve bayes* dengan dilakukan *oversampling ADASYN.*
   6. Skenario terakhir akan dilakukan evaluasi terhadap klasifikasi metode *decision tree* dengan dilakukan *oversampling ADASYN.*

   Untuk mengetahui performa setiap metode dihitung dari nilai *accuracy, precision, recall,* dan ROC_AUC.
   Perhitungan nilai *accuracy* :
   Accuracy = (TP + TN) / (TP + FP + TN + FN) [12]
   Perhitungan nilai *precision* :
   Precision = TP / (TP + FP) [12]
   Perhitungan nilai *recall* :
   Recall = TP / (TP + FN) [12]
   ROC_AUC digunakan untuk mengukur kinerja diskriminatif dengan memperkirakan output dari sampel yang dipilih secara acak dari populasi negatif atau positif, semakin besar nilai ROC_AUC maka semakin kuat klasifikasi yang digunakan [12].
   Dengan pembagian data menggunakan *k-fold cross validation* dengan nilai k 3,5, dan 7 yang merujuk pada penelitian [21].

III. HASIL PENELITIAN

Dalam menangani ketidakseimbangan data pada *dataset* maka digunakan metode *SMOTE* dan *ADASYN.* Pada Gambar 2. dapat dilihat bahwa data normal memiliki ketidakseimbangan kelas yang signifikan 1131 daftar ulang dan 631 tidak daftar ulang. Metode *SMOTE* dapat melakukan *oversampling* daftar ulang sebanyak 1131 dan tidak daftar ulang sebanyak 1131 yang artinya jumlah daftar ulang dan tidak daftar ulang sama dan metode *ADASYN* dapat melakukan *oversampling* daftar ulang sebanyak 1131 dan tidak daftar ulang sebanyak 1181.



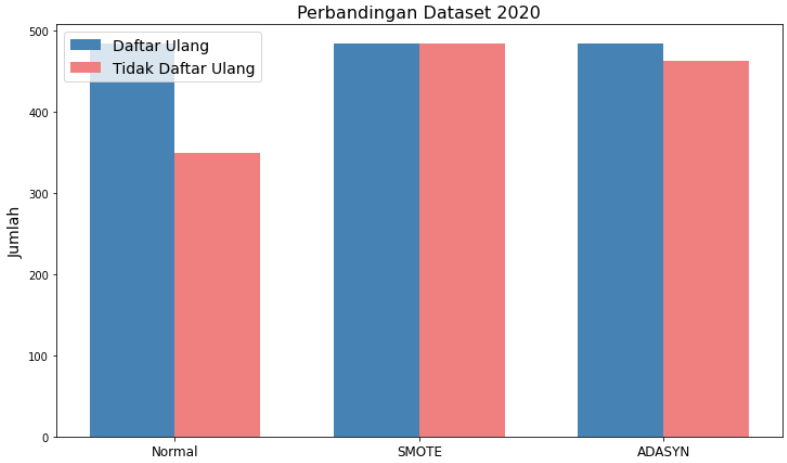
Gambar 2. Perbandingan Data Normal dan Oversampling SPAN PTKIN 2019

Pada Tabel 2 dapat disimpulkan bahwa penggunaan metode *ADASYN* dan *SMOTE* untuk menangani ketidakseimbangan data pada *dataset* SPAN PTKIN tahun 2019 dapat menurunkan nilai *accuracy*, hal ini dikarenakan sebelum dilakukan *oversampling* maka banyak bias yang mengabaikan kelas minoritas hal ini dapat dilihat dari nilai *precision* dan *recall* sebelum dilakukan *oversampling* dan setelah dilakukan *oversampling*. Algoritma *decision tree* dengan nilai *cross validation* = 5 memiliki nilai *accuracy* tinggi dibandingkan dengan algoritma lainnya sebesar 67% dan algoritma *ADASYN decision tree* dan *SMOTE decision tree* dengan nilai *cross validation* 5 memiliki *accuracy* paling rendah sebesar 55%. Namun, penggunaan algoritma *ADASYN* dan *SMOTE* dapat meningkatkan nilai ROC_AUC pada algoritma *naïve bayes* dengan nilai *cross validation* 3, 5 dan 7 seperti terlihat pada Tabel 2 Algoritma *SMOTE* memiliki nilai *accuracy* dan nilai ROC_AUC lebih tinggi dibandingkan dengan algoritma *ADASYN* pada *dataset* SPAN PTKIN tahun 2019. Jadi, algoritma *SMOTE* memiliki kinerja yang lebih baik dibandingkan algoritma *ADASYN*.

TABEL 2
PELATIHAN MODEL DATA SPAN PTKIN 2019

Data Daftar Ulang Tahun 2019						
k-fold	*Decision Tree*	*SMOTE Decision Tree*	*ADASYN Decision Tree*	*Naïve Bayes*	*SMOTE Naïve Bayes*	*ADASYN Naïve Bayes*
Accuracy						
3	64%	54%	54%	66%	55%	54%
5	**67%**	53%	53%	66%	57%	57%
7	66%	54%	55%	66%	59%	57%
Roc_Auc						
3	58%	55%	54%	62%	62%	61%
5	58%	55%	54%	61%	**63%**	62%
7	56%	56%	56%	60%	62%	61%
Precision						
3	66%	60%	59%	67%	53%	50%
5	**67%**	56%	54%	67%	54%	53%
7	66%	58%	55%	67%	55%	54%
Recall						
3	89%	53%	54%	91%	73%	68%
5	96%	60%	66%	94%	79%	74%
7	**97%**	54%	58%	95%	81%	77%

Pada Gambar 3 dapat dilihat bahwa data normal memiliki ketidakseimbangan kelas yang signifikan 484 daftar ulang dan 350 tidak daftar ulang. Metode *SMOTE* dapat melakukan *oversampling* daftar ulang sebanyak 484 dan tidak daftar ulang sebanyak 484 yang artinya jumlah daftar ulang dan tidak daftar ulang sama dan metode *ADASYN* dapat melakukan *oversampling* daftar ulang sebanyak 484 dan tidak daftar ulang sebanyak 463.



Gambar 3. Perbandingan Data Normal dan *Oversampling* SPAN PTKIN 2020

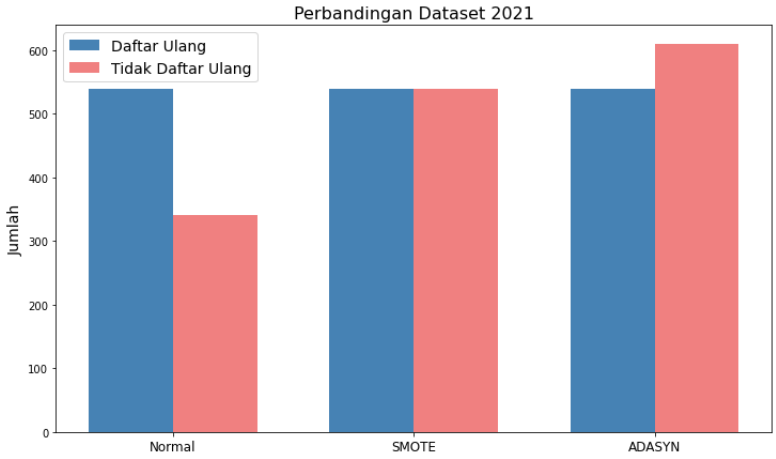
Pada Tabel 3 dapat disimpulkan bahwa penggunaan algoritma *ADASYN* dan *SMOTE* untuk menangani ketidakseimbangan data pada *dataset* SPAN PTKIN tahun 2020 dapat menurunkan nilai *accuracy* hal ini dikarenakan sebelum dilakukan *oversampling* maka banyak bias yang mengabaikan kelas minoritas, hal ini dapat dilihat dari nilai *precision* dan *recall* sebelum dilakukan *oversampling* dan setelah dilakukan oversampling. Algoritma *naïve bayes* pada nilai *cross validation* 3,5,7 dan algoritma *decision tree* dengan nilai *cross validation* 7 memiliki nilai *accuracy* paling tinggi diantara algoritma lainnya sebesar 60%. Algoritma *ADASYN decision tree* memiliki nilai *accuracy* paling rendah yaitu sebesar 50% pada nilai *cross validation* 3. Namun, algoritma *SMOTE* dan *ADASYN* dapat meningkatkan nilai ROC_AUC pada algoritma *decision tree* seperti terlihat pada Tabel 3 Nilai ROC_AUC algoritma *SMOTE* lebih tinggi dibandingkan dengan algoritma *ADASYN* pada *dataset* SPAN PTKIN tahun 2020. Jadi, algoritma *SMOTE* memiliki kinerja yang lebih baik dibandingkan algoritma *ADASYN*.

TABEL 3
PELATIHAN MODEL DATA SPAN PTKIN 2020

Data Daftar Ulang Tahun 2020						
k-fold	*Decision Tree*	*SMOTE Decision Tree*	*ADASYN Decision Tree*	*Naïve Bayes*	*SMOTE Naïve Bayes*	*ADASYN Naïve Bayes*
Accuracy						
3	59%	55%	50%	**60%**	53%	52%
5	58%	53%	52%	**60%**	54%	53%
7	**60%**	54%	53%	**60%**	54%	53%
Roc_Auc						
3	53%	56%	51%	57%	51%	49%
5	52%	54%	54%	**58%**	54%	52%
7	54%	55%	56%	**58%**	54%	52%
Precision						
3	59%	54%	54%	**60%**	52%	52%
5	**60%**	58%	53%	**60%**	51%	53%
7	**60%**	58%	54%	59%	52%	52%
Recall						
3	**95%**	62%	65%	95%	90%	90%
5	87%	80%	89%	96%	90%	92%
7	**95%**	73%	65%	96%	92%	93%

Pada Gambar 4 dapat dilihat bahwa data normal memiliki ketidakseimbangan kelas yang signifikan 539 daftar ulang dan 340 tidak daftar ulang. Metode *SMOTE* dapat melakukan *oversampling* daftar ulang sebanyak 539 dan tidak daftar

ulang sebanyak 539 yang artinya jumlah daftar ulang dan tidak daftar ulang sama dan metode *ADASYN* dapat melakukan *oversampling* daftar ulang sebanyak 539 dan tidak daftar ulang sebanyak 609.



Gambar 4. Perbandingan Data Normal dan *Oversampling* SPAN PTKIN 2021

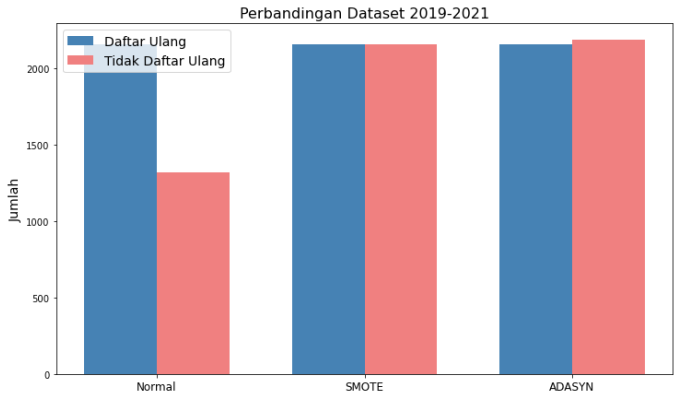
Pada Tabel 4 dapat disimpulkan bahwa penggunaan algoritma *ADASYN* dan *SMOTE* untuk menangani ketidakseimbangan data pada *dataset* SPAN PTKIN tahun 2021 dapat menurunkan nilai *accuracy* hal ini dikarenakan sebelum dilakukan *oversampling* maka banyak bias yang yang mengabaikan kelas minoritas hal tersebut dapat dilihat dari nilai *precision* dan *recall*. Algoritma *decision tree* pada nilai *cross validation* 3 dan 7 memiliki nilai *accuracy* paling tinggi diantara algoritma lainnya sebesar 65%. Algoritma *SMOTE naïve bayes* pada nilai *cross validation* 3 memiliki nilai *accuracy* paling rendah yaitu sebesar 49%. Algoritma *ADASYN* memiliki nilai *accuracy* lebih besar dibandingkan algoritma *SMOTE* di beberapa model dan nilai ROC_AUC *SMOTE* lebih tinggi dibandingkan dengan algoritma *ADASYN* di beberapa model pada *dataset* SPAN PTKIN tahun 2021. Jadi, algoritma *SMOTE* memiliki kinerja yang lebih baik dibandingkan algoritma *ADASYN.*

TABEL 4
PELATIHAN MODEL DATA SPAN PTKIN 2021

Data Daftar Ulang Tahun 2021						
k-fold	*Decision Tree*	*SMOTE Decision Tree*	*ADASYN Decision Tree*	*Naïve Bayes*	*SMOTE Naïve Bayes*	*ADASYN Naïve Bayes*
Accuracy						
3	**65%**	57%	54%	64%	49%	52%
5	64%	51%	53%	64%	50%	52%
7	**65%**	55%	56%	64%	52%	52%
Roc_Auc						
3	61%	56%	53%	**62%**	56%	56%
5	61%	50%	51%	61%	57%	57%
7	61%	55%	53%	60%	56%	54%
Precision						
3	**65%**	55%	43%	64%	53%	50%
5	64%	51%	39%	64%	53%	50%
7	64%	49%	50%	64%	53%	49%
Recall						
3	**95%**	43%	34%	96%	81%	51%
5	**95%**	59%	40%	96%	86%	48%
7	**95%**	59%	43%	96%	90%	51%

Pada Gambar 5 dapat dilihat bahwa data normal memiliki ketidakseimbangan kelas yang signifikan 2154 daftar ulang dan 1321 tidak daftar ulang. Metode *SMOTE* dapat melakukan *oversampling* daftar ulang sebanyak 2154 dan tidak daftar

ulang sebanyak 2154 yang artinya jumlah daftar ulang dan tidak daftar ulang sama dan metode *ADASYN* dapat melakukan *oversampling* daftar ulang sebanyak 2154 dan tidak daftar ulang sebanyak 2183.



Gambar 5 Perbandingan Data Normal dan *Oversampling* SPAN PTKIN 2019-2021

Pada Tabel 5 dapat disimpulkan bahwa penggunaan metode *ADASYN* dan *SMOTE* untuk menangani ketidakseimbangan data pada *dataset* SPAN PTKIN tahun 2019-2021 dapat menurunkan nilai *accuracy* hal ini dikarenakan sebelum dilakukan *oversampling* maka banyak bias yang yang mengabaikan kelas minoritas hal tersebut dapat dilihat dari nilai *precision* dan *recall*. Algoritma *decision tree* dengan nilai *cross validation* = 3,5, dan 7 memiliki nilai *accuracy* tinggi dibandingkan dengan algoritma lainnya sebesar 65% dan algoritma *SMOTE decision tree* dengan nilai *cross validation* 7 memiliki *accuracy* paling rendah sebesar 49%. Algoritma *SMOTE* memiliki nilai *accuracy* dan nilai ROC_AUC lebih tinggi dibandingkan dengan algoritma *ADASYN* pada *dataset* SPAN PTKIN tahun 2019-2021. Jadi, algoritma *SMOTE* memiliki kinerja yang lebih baik dibandingkan algoritma *ADASYN*.

TABEL 5
PELATIHAN MODEL DATA SPAN PTKIN 2019-2021

Data Daftar Ulang Tahun 2019-2021						
k-fold	*Decision Tree*	*SMOTE Decision Tree*	*ADASYN Decision Tree*	*Naïve Bayes*	*SMOTE Naïve Bayes*	*ADASYN Naïve Bayes*
Accuracy						
3	**65%**	53%	51%	64%	54%	53%
5	**65%**	52%	50%	64%	54%	51%
7	**65%**	49%	50%	64%	55%	52%
Roc_Auc						
3	57%	55%	52%	**60%**	58%	55%
5	57%	54%	50%	**60%**	57%	52%
7	56%	53%	49%	**60%**	56%	50%
Precison						
3	**65%**	52%	50%	64%	53%	52%
5	**65%**	52%	50%	64%	52%	50%
7	**65%**	48%	49%	64%	52%	51%
Recall						
3	**95%**	47%	52%	97%	90%	84%
5	**95%**	47%	53%	97%	90%	83%
7	**95%**	41%	48%	97%	91%	85%

Pada semua *dataset* berdasarkan statistik deskriptif semua atribut memiliki pengaruh terhadap calon mahasiswa yang tidak daftar ulang. Hal tersebut dapat dilihat masih banyaknya persentase calon mahasiswa yang tidak daftar ulang di tiap atributnya. Namun, setelah dilakukan perhitungan korelasi dengan *pearson correlation* semua atribut memiliki korelasi yang lemah karena nilai dari *pearson correlation* berada di bawah 0.2. Hal tersebut akan berpengaruh pada hasil *accuracy* dan kinerja model klasifikasi. *Dataset* yang digunakan dalam penelitian ini memiliki *class imbalanced* hal tersebut akan berpengaruh pada keakuratan klasifikasi.

Dalam penelitian ini menggunakan tipe data nominal. Untuk melakukan *oversampling* dengan menggunakan metode *SMOTE* dan *ADASYN* yang mana data sintetis dibuat dengan menghitung jarak terdekat dengan menggunakan metode *k-nn* maka tipe data harus dirubah atau di *labeling* ke numerik.

Hasil pengujian menunjukkan bahwa nilai *precision* dan *recall* sebelum dilakukan *oversampling* dan setelah dilakukan *oversampling* mengalami perubahan. Nilai *precision* dan *recall* sebelum dilakukan *oversampling* terdapat ketidak seimbangan yang mana nilai *recall* lebih besar dibandingkan nilai *precision.* Namun, setelah dilakukan *oversampling* nilai *recall* dan *precision* di hampir semua pengujian memiliki nilai yang seimbang. Hal ini dapat disimpulkan bahwa proses klasifikasi tanpa dilakukan *oversampling* mengabaikan kelas minoritas.

Dengan dilakukan *oversampling* dengan menggunakan metode *SMOTE* dan *ADASYN* pada *dataset* yang memiliki nilai korelasi rendah akan mengakibatkan kinerja yang dihasilkan dalam klasifikasi dapat menurun. Hal ini dikarenakan data sintetis yang dibuat dalam *oversampling* memiliki korelasi yang rendah. Merujuk pada penelitian yang dilakukan oleh [22], penggunaan metode *SMOTE* pada penelitian ini memberikan dampak yang signifikan pada nilai akurasi. Nilai akurasi dari algoritma *naïve bayes* sebesar 89,24% dan *SMOTE + naïve bayes* memiliki nilai akurasi sebesar 74.2%.

IV. SIMPULAN

Dalam menangani ketidakseimbangan data pada keseluruhan *dataset* mulai dari tahun 2019, 2020, 2021, dan 2019-2021. Algoritma *SMOTE* dapat melakukan *oversampling* yang mana jumlah daftar ulang dan tidak daftar ulang seimbang dan metode *ADASYN* masih memiliki jarak antara jumlah daftar ulang dan tidak daftar ulang. Untuk performa klasifikasi berdasarkan hasil evaluasi algoritma *SMOTE* lebih cocok digunakan untuk mengatasi *class imbalanced* pada penelitian ini, dikarenakan algoritma *SMOTE* memiliki kinerja yang lebih baik dibandingkan algoritma *ADASYN.*

UCAPAN TERIMA KASIH

Terima kasih diberikan kepada Universitas Islam Negeri Sunan Ampel Surabaya yang telah mengizinkan penelitian ini.

DAFTAR PUSTAKA

[1] M. Mambang and F. D. Marleny, "Prediksi Calon Mahasiswa Baru Menggunakan Metode Klasifikasi Decision Tree," *CSRID (Computer Science Research and Its Development Journal)*, vol. 7, no. 1, pp.48-56, 2015.
[2] (2021) Pangkalan Data Pendidikan Tinggi. [Online]. Tersedia: https://pddikti.kemdikbud.go.id/data_pt/RTQxRkQ0QjgtQkQ2NC00RTRGLTg0QzQtQzhBQzAzQzc1RjI4/
[3] (2021) Seleksi Penerimaan Mahasiswa Baru - Jalur seleksi. [Online]. Tersedia: https://pmb.uinsby.ac.id/jalur-seleksi/
[4] (2020) Data SPAN PTKIN. [Online]. Tersedia: https://span-ptkin.ac.id/prodi/
[5] M. N. Rabbani, A. Yusuf, and D. Rolliawati, "Komparasi Model Prediksi Daftar Ulang Calon Mahasiswa Baru Menggunakan Metode Decision Tree Dan Adaboost," *SISFOKOM*, vol. 10, no. 1, pp. 18–24, 2021.
[6] D. Aribowo and A. E. H. Setiadi, "Komparasi Algoritma Untuk Klasifikasi Heregistrasi Calon Mahasiswa," *Seminar Nasional Edusaintek*, vol. 1, no. 1, pp. 108-114, 2018.
[7] S. Hendrian, "Algoritma Klasifikasi Data Mining Untuk Memprediksi Siswa Dalam Memperoleh Bantuan Dana Pendidikan," *FaktorExacta*, vol. 11, no. 3, pp. 2018.
[8] A. P. Wibawa, M. G. A. Purnama, M. F. Akbar, and F. A. Dwiyanto, "Metode-metode Klasifikasi," *Prosiding SAKTI*, vol. 3, no. 1, pp. 5, 2018.
[9] X. Wu *et al.*, "Top 10 algorithms in data mining," *Knowledge and Information Systems*, vol. 14, no. 1, pp. 1-37, 2007.
[10] D. Sugianti, "Algoritma Bayesian Classification Untuk Memprediksi Heregistrasi Mahasiswa Baru di STMIK WIDYA PRATAMA," *Jurnal Ilmiah ICTech*, vol. 10, no. 2, pp. 5, 2012.
[11] N. Hartati, "Intellectual Capital Dalam Meningkatkan Daya Saing: Sebuah Telaah Literatur," *Etikonomi*, vol. 13, no. 1, pp. 51-68, 2015.
[12] N. Yahya and A. Jananto, "Komparasi Kinerja Algoritma C.45 dan Naïve Bayes Untuk Prediksi Kegiatan Penerimaan Mahasiswa Baru (STUDI KASUS : UNIVERSITAS STIKUBANK SEMARANG)," *proceeding SENDI_U*, 2019.
[13] P. D. Wicaksana, "Perbandingan Algoritma K-Nearest Neighbors dan Naive Bayes Untuk Studi Data 'WISCONSIN DIAGNOSIS BREAST CANCER,'" FST UNIVERSITAS SANATA DHARMA YOGYAKARTA, YOGYAKARTA, 2015.
[14] S. Kotsiantis, D. Kanellopoulos, and P. Pintelas, "Data Preprocessing for Supervised Learning," *International Journal of Computer Science*, vol. 1, no.2, pp. 111–117, 2006.
[15] L. Xiang-wei and Q. Yian-fang, "A Data Preprocessing Algorithm for Classification Model Based On Rough Sets," *Physics Procedia*, vol. 25, pp. 2025–2029, 2012.
[16] A. M. Ws and R. Nooraeni, "Penerapan Metode Resampling Dalam Mengatasi Imbalanced Data Pada Determinan Kasus Diare Pada Blita di Indonesia (ANALISIS DATA SDKI 2017)," *Jurnal MSA (Matematika dan Statistika serta Aplikasinya)*, vol. 8, no. 1, pp. 19-27, 2020.
[17] T. Cai, "Breast Cancer Diagnosis Using Imbalanced Learning and Ensemble Method," *ACM*, vol. 7, no. 3, pp. 146-154, 2018.
[18] F. S. Dhitama and F. A. Bachtiar, "Penentuan Kelayakan Debitur Menggunakan Metode Decision Tree C4.5 Dan Oversampling Adaptive Synthetic (ADASYN)," *Jurnal Pengembangan Teknologi Informasi dan Ilmu Komputer*, vol. 4, no. 10, pp. 10, 2020.
[19] A. Franseda, W. Kurniawan, and S. Anggraeni, "Integrasi Metode Decision Tree dan SMOTE untuk Klasifikasi Data Kecelakaan Lalu Lintas," *JUSTIN (Jurnal Sistem dan Teknologi Informasi)*, vol. 08, no. 3, pp. 9, 2020.
[20] C. Catal, "Software fault prediction: A literature review and current trends," *Expert Systems with Applications*, vol. 38, no. 4, pp. 4626–4636, 2011.
[21] H. Rhomadhona and J. Permadi, "Klasifikasi Berita Kriminal Menggunakan NaÃ¯ve Bayes Classifier (NBC) dengan Pengujian K-Fold Cross Validation," *JSI*, vol. 5, no. 2, pp. 108–117, 2019.
[22] H. Hairani, "Metode Klasifikasi Data Mining dan Teknik Sampling Smote Menangani Class Imbalance untuk Segmentasi Customer pada industri Perbankan," *Prosiding SNST Fakultas Teknik*, 2019.