e-ISSN 2549-7472
Edumatic: Jurnal Pendidikan Informatika
Vol. 5 No. 1, Juni, 2021, **Hal. 11-20**
DOI: 10.29408/edumatic.v5i1.3125

Pengaruh Algoritma ADASYN dan SMOTE terhadap Performa Support Vector Machine pada Ketidakseimbangan Dataset Airbnb

Wahyu Hidayat*[1], Mursyid Ardiansyah[2], Arief Setyanto[3]
[1,2,3] Program Studi Teknik Informatika, Universitas Amikom Yogyakarta
email: wahyu.1181@students.amikom.ac.id*[1], mursyid.27@students.amikom.ac.id[2], arief_s@amikom.ac.id[3]

(Received: 27 Januari 2021 / Accepted: 7 Maret 2021 / Published Online: 20 Juni 2021)

Abstrak

Aktivitas travelling semakin banyak dilakukan oleh masyarakat di dunia. Beberapa tempat wisata sulit di jangkau oleh hotel karena beberapa lokasi tempat wisata jauh dari pusat kota, Airbnb merupakan platform yang menyediakan penyewaan berbasis rumah atau apartemen. Dalam penawaran penginapan terdapat dua jenis host, yaitu non-super host dan super host, badge super host didapatkan apabila pemilik penginapan memiliki reputasi yang baik dan memenuhi persyaratan. Ada keuntungan menjadi super host seperti memiliki lebih banyak visibilitas, meningkatkan potensi penghasilan, dan hadiah eksklusif. Proses klasifikasi algoritma Support Vector Machine (SVM) dengan mengolah data kriteria tersebut. Dataset tidak seimbang, populasi host super jauh lebih kecil daripada non-super host. Mengatasi ketidakseimbangan maka dilakukan teknik oversampling menggunakan ADASYN dan SMOTE. Tujuan penelitian untuk mengetahui performa teknik oversampling ADASYN dan SMOTE pada algoritma SVM. Teknik yang digunakan untuk menganalisis data adalah oversampling yang bertujuan menangani dataset yang tidak seimbang, dan confusion matrik digunakan untuk pengujian Precision, Recall, dan F1-SCORE, serta Accuracy. Hasil temuan kami menunjukkan bahwa SMOTE SVM meningkatkan tingkat akurasi sebesar 1 persen dari 80% menjadi 81% yang dipengaruhi dari kenaikan hasil pengujian label True (minoritas) dan terjadi penurunan pada hasil pengujian label False (mayoritas), SMOTE SVM tersebut paling unggul dibandingkan ADASYN SVM dan SVM tanpa oversampling.

Kata Kunci: ADASYN, Klasifikasi, SMOTE, SVM, Travel

Abstract

Traveling activities are increasingly being carried out by people in the world. Some tourist attractions are difficult to reach hotels because some tourist attractions are far from the city center, Airbnb is a platform that provides home or apartment-based rentals. In lodging offers, there are two types of hosts, namely non-super host and super host. The super-host badge is obtained if the innkeeper has a good reputation and meets the requirements. There are advantages to being a super host such as having more visibility, increased earning potential and exclusive rewards. Support Vector Machine (SVM) algorithm classification process by these criteria data. Data set is unbalanced. The super host population is smaller than the non-super host. Overcoming the imbalance, this over sampling technique is carried out using ADASYN and SMOTE. Research goal was to decide the performance of ADASYN and sampling technique, SVM algorithm. Data analyses used over sampling which aims to handle unbalanced data sets, and confusion matrix used for testing Precision, Recall, and F1-SCORE, and Accuracy. Research shows that SMOTE SVM increases the accuracy rate by 1 percent from 80% to 81%, which is influenced by the increase in the True (minority) label test results and a decrease in the False label test results (majority), the SMOTE SVM is better than ADASYN SVM, and SVM without over sampling.

Keywords: *ADASYN, Classification, SMOTE, SVM, Traveling*

PENDAHULUAN

Objek wisata selalu meningkat di beberapa negara, *traveller* tidak hanya datang dari masyarakat yang tinggal di negara yang memiliki wisata tersebut, akan tetapi banyak juga traveller yang berasal dari luar negeri atau biasa disebut turis asing. Biasanya saat traveller kesana untuk menginap di hotel atau homestay, namun ketika saat travelling mengunjungi beberapa daerah yang tidak ada hotel sama sekali karena jauh dari pusat kota atau akses yang sulit ke lingkungan wisata, biasanya para *traveller* tersebut menggunakan layanan Airbnb yang memfasilitasi penggunanya untuk dapat menyewa penginapan berupa rumah atau apartemen, pada dasarnya Airbnb dapat menyediakan sewa rumah secara penuh atau hanya menyediakan layanan sewa kamar tidur yang satu rumah dengan pemilik rumah.

Penelitian ini menggunakan objek Airbnb karena layanan tersebut telah membuka dan memfasilitasi berbagi persewaan rumah hingga jutaan pasar global dan menjadi *sharing economy superstar*. Layanan tersebut merupakan *sharing economy* yang dalam artian teknologi layanan Airbnb digunakan oleh pengguna yang sangat besar dan memiliki layanan lokasi dan tempat yang memungkinkan antar kelompok saling bertransaksi dengan metode berbagi atau *stranger sharing* (Crommelin, Troy, Martin, & Pettit, 2018). Airbnb memiliki jangkauan penginapan yang luas karena terdiri dari penyedia tempat sewa yang flexibel tidak bergantung pada satu perusahaan seperti hotel namun mencakup penginapan skala kecil dan skala besar.

Pengguna layanan Airbnb sangat tinggi tercata pada website Airbnb https://news.airbnb.com/about-us/ pada bulan September 2020 lebih dari 800 juta pengguna dan lebih dari 4 juta daftar penginapan. Pengguna tersebut memilih layanan Airbnb dibandingkan hotel karena terdapat faktor pendorong seperti interaksi yang lebih baik, manfaat rumah yang dapat digunakan sepenuhnya, alasan ekonomi yang flexibel, serta keaslian lokal terhadap lokasi penyewaan (Guttentag, Smith, Potwarka, & Havitz, 2018). Berdasarkan dari berbagai faktor pendorong tersebut terdapat banyaknya pengguna layanan Airbnb dibandingkan dengan penyewaan hotel pada saat wisata karena terdapat alasan faktor tersebut.

Berdasarkan banyaknya pengguna Airbnb, maka banyak juga jasa persewaan rumah pada layanan tersebut. Secara umum, traveler akan mendapatkan daftar pencarian sewa rumah terdekat dan diurutkan berdasarkan super host karena salah satu kelebihan super host adalah visibilitas yang baik. Jadi dapat direkomendasikan untuk persewaan rumah yang bagus berdasarkan penawaran yang diprioritaskan berdasarkan super host. Oleh karena itu penting untuk mengetahui apakah host tersebut sudah menjadi super host atau tidak karena jika sudah menjadi super host maka memiliki rating yang baik dan pelayanan yang baik.

Chen & Chang, (2018) meneliti niat pengguna di Airbnb berdasarkan faktor-faktor tertentu berdasarkan ulasan konsumen, kualitas informasi, dan kekayaan media. Menggunakan analisis ANOVA didapatkan hasil bahwa jumlah dan tipe rating sangat mempengaruhi perhatian pengguna untuk membuat sebuah kamar / rumah kontrakan. Kualitas informasi pada rating ternyata memiliki pengaruh positif dan signifikan terhadap kepuasan rental. Peringkat pencarian Airbnb (Pucci & Rooman, 2017) dengan daftar embedding yang membandingkan hasil pencarian Airbnb dengan personalisasi real-time dapat mengatasi pencarian tempat persewaan Airbnb dengan cepat karena bisa dilakukan secara offline, tetapi daftar rekomendasi yang dihasilkan hanyalah daftar pencarian tersimpan yang serupa.

Algoritma Support Vector Machine (SVM) merupakan algoritma untuk klasifikasi, menurut penelitian (Kusumawati, D'Arofah, & Pramana, 2019) dalam penelitiannya yaitu menguji algoritma klasifikasi SVM dan Naïve Bayes dengan tujuan untuk mengklasifikasikan layanan Tokopedia yang tersedia di Twitter menjadi dua klasifikasi positif dan negatif, hasil penelitian menunjukkan bahwa algoritma Support Vector Machine memiliki performansi

klasifikasi yang paling baik dibandingkan dengan algoritma Naïve Bayes. Dalam penelitiannya SVM memiliki hasil akurasi 83,34% sedangkan Naïve Bayes memiliki hasil akurasi 75%.

Selain SVM dan Naïve Bayes, terdapat beberapa algoritma untuk klasifikasi lainnya (Alsmadi & Hoon, 2019) yaitu decision tree, K-Nearest Neighbor (KNN), dan Logistic Regression. Pada penelitian yang dilakukan oleh Alsmadi dengan membandingkan algoritma klasifikasi yang ada dengan kasus pemrosesan teks di Twitter yang diubah menjadi Term Weighting menggunakan Simple Weighting, diperoleh hasil bahwa penerapan algoritma SVM memiliki tingkat akurasi yang paling tinggi dibandingkan algoritma lainnya. Penelitian terkait SVM lainnya yaitu (Rimal, Rijal, & Kunwar, 2020) dengan membandingkan SVM dan Maximum Likelihood (ML) untuk klasifikasi pemetaan urbanisasi menunjukkan hasil bahwa SVM terbukti efektif dalam menentukan klasifikasi tutupan lahan khususnya perkotaan / bangunan.

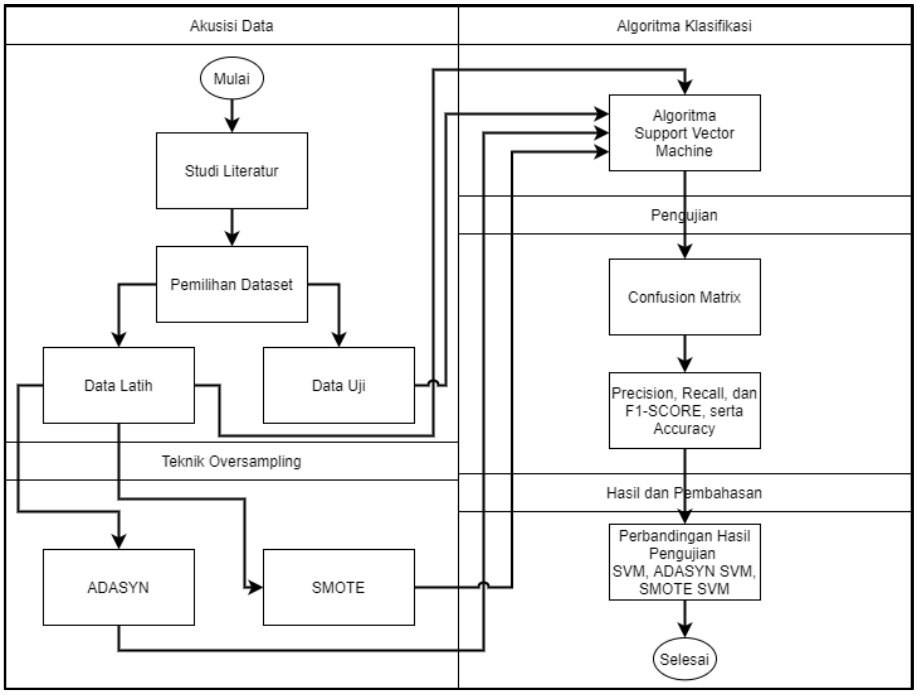
Algoritma untuk mengatasi oversampling menggunakan algoritma ADASYN, pada penelitian sebelumnya (Fico et al., 2018) menggunakan model SVM Linear, RBF, ADASYN, Weighting, ADASYN + Weighting yang memiliki kesimpulan bahwa model SVM ADASYN mendapatkan Mean sebesar 0.946, Median, dan IQR 0.919-0.985. Dimana SVM dengan model ADASYN memiliki nilai Mean, Median, dan IQR terbaik sehingga dapat mengatasi oversampling pada klasifikasi menggunakan algoritma SVM.

Evaluasi yang digunakan untuk menguji kualitas hasil klasifikasi (Düntsch & Gediga, 2020) adalah Confusion matrix, banyak hasil statistik yang diperoleh pada penelitian selanjutnya dikaitkan dengan Confusion matrix, seperti sensitivitas TP / (TP + FN), spesifisitas TN / (TN + FP) , presisi (TP) / (FP + TP). Pengujian pada algoritma SVM dapat dilakukan dengan Confusion Matrix (Rustam & Audia Ariantari, 2018) pengujian dilakukan untuk menempati peran penting bagi kinerja pengklasifikasi karena dari hasil model klasifikasi yang telah didapatkan jumlah prediksi benar dan salah dibandingkan dengan hasil dataset yang sebenarnya sehingga dapat ditunjukkan tingkat akurasi dalam pengujian algoritma SVM.

Berdasarkan penelitian sebelumnya, pemilihan algoritma SVM akan diterapkan untuk klasifikasi dalam mengidentifikasi host yang akan memiliki label superhost atau tidak, yang kemudian menerapkan algoritma ADASYN dan SMOTE untuk mengatasi oversampling sehingga dengan pengujian presisi, recall, Skor F1-Score dan akurasi berdasarkan Confusion Matrix dapat diketahui hasil dari pengujian yang dilakukan. Tujuan dari penelitian yang dilakukan yaitu untuk mengetahui performa dari penerapan teknik oversampling dengan metode ADASYN dan SMOTE pada klasifikasi menggunakan algoritma SVM. Sehingga dari penerapan teknik oversampling tersebut dapat diketahui hasil dan perbandingan antara penerapan algoritma SVM, ADASYN SVM, SMOTE SVM. Sehingga penelitian ini dapat bermanfaat untuk mengetahui performa dari masing-masing teknik oversampling pada algoritma klasifikasi SVM.

METODE

Pada penelitian ini akan dilakukan klasifikasi untuk mengetahui apakah host merupakan super host atau non-super host. Berdasarkan studi literatur yang telah dilakukan, algoritma ADASYN dan SMOTE akan digunakan untuk mengatasi ketidakseimbangan dataset, sedangkan untuk mengklasifikasikannya menggunakan SVM, sedangkan dalam pengujiannya akan dibandingkan penerapan SVM, ADSYN SVM dan SMOTE SVM. Berikut ditampilkan diagram alir penelitian untuk membuat penelitian yang dilakukan lebih tertuju untuk menyelesaikan masalah dan mencapai tujuan penelitian. gambar 1 menampilkan diagram alir penelitian yang dilakukan.



Gambar 1. Diagram alir penelitian.

Berdasarkan Gambar 1 terdapat tahapan penelitian dari tahap akusisi data, penerapan teknik oversampling, penerapan algoritma klasifikasi, pengujian algoritma serta pemaparan hasil dan pembahasan. Berikut detail tahapan yang dilakukan berdasarkan Gambar 1 :

Akuisisi Data

Dataset yang diperoleh berasal dari https://www.kaggle.com/samyukthamurali/Airbnb-ratings-dataset yang merupakan data daftar Airbnb. Kriteria yang memengaruhi host untuk menjadi super host dijelaskan di situs web utama Airbnb yang mencakup reservasi, penilaian, tanggapan, dan pembatalan. Berdasarkan kriteria tersebut maka digunakan kolom dataset yang berhubungan dengan kriteria yang ada yaitu kolom *Host Response Rate*, *Number of reviews, Review Scores Rating*, dan *Host total listing count*.

Teknik Oversampling

Adaptive Synthetic (ADASYN) (He, Bai, Garcia, & Li, 2008) adalah algoritma untuk menangani dataset yang tidak seimbang dalam klasifikasi data. ADASYN dapat menghasilkan sampel secara adaptif dalam data sintetik terhadap kelas minoritas yang dibentuk oleh distribusi data untuk mengurangi bias yang disebabkan oleh distribusi data yang tidak merata pada data pada label lain yang memiliki kelas mayoritas.

Synthetic Minority Over-sampling Technique (SMOTE) dapat meningkatkan akurasi pengklasifikasian data yang memiliki distribusi label yang tidak merata, dan terdapat label data minoritas. SMOTE ditujukan untuk mengelola teknik pengambilan sampel yang difokuskan pada pembelajaran terutama pada distribusi data minoritas dan memperkenalkan bias terhadap kelas minoritas (Chawla, Bowyer, Hall, & Kegelmeyer, 2002).

Algoritma Klasifikasi *Support Vector Machine*

Support Vector Machine atau yang dikenal dengan algoritma SVM merupakan algoritma yang berfokus pada masalah klasifikasi (Zubrinic, Milicevic, & Zakarija, 2013). SVM pada awalnya dikemukakan oleh Vapnik untuk pemecahan masalah klasifikasi dan analisis regresi, algoritma ini telah banyak digunakan untuk masalah klasifikasi data linier dan non linier (Ahmad, Basheri, Iqbal, & Rahim, 2018).

Pengujian

Pengujian dilakukan dengan alat uji yaitu Confusion Matrix untuk mengetahui sebaran kebenaran data prediksi terhadap data aktual. Serta pengujian algoritma klasifikasi dapa dilakukan menggunakan pengujian Precision, Recall, dan F1-SCORE, serta Accuracy (Harianto, Sunyoto, & Sudarmawan, 2020). Model pengujian tersebut dapat dibandingkan dengan hasil pada beberapa skenario yang dilakukan (Sari, Firdausi, & Azhar, 2020). Presisi adalah rasio prediksi true positif dibandingkan dengan hasil prediksi positif secara keseluruhan. Recall adalah rasio true positif dibandingkan dengan semua data positif. F1-SCORE adalah perbandingan presisi dan perolehan rata-rata tertimbang (Patil & Nemade, 2017). Confusion Matrix memberikan perbandingan antara hasil klasifikasi yang dibuat oleh model dengan hasil klasifikasi yang sebenarnya. Akurasi adalah rasio true prediksi (positif dan negatif) terhadap keseluruhan data. Tabel Confusion Matrix ditampilkan pada tabel 1.

Tabel 1. Confusion Matrix

Label	Definisi
True Positive (TP)	Jumlah positif yang dianggap positif.
True Negative (TN)	Jumlah negatif yang dianggap negatif.
False Positive (FP)	Jumlah positif yang dianggap negatif.
False Negatif (FN)	Jumlah negatif yang dianggap positif.

$$Precision = \frac{TP}{TP+FP} \quad (1)$$

$$Recall = \frac{TP}{TP+FN} \quad (2)$$

$$F1 = \frac{2*Precision*Recall}{Precision+Recall} \quad (3)$$

$$Accuracy = \frac{TP+FN}{TP+TN+FP+FN}\ 100\% \quad (4)$$

HASIL DAN PEMBAHASAN

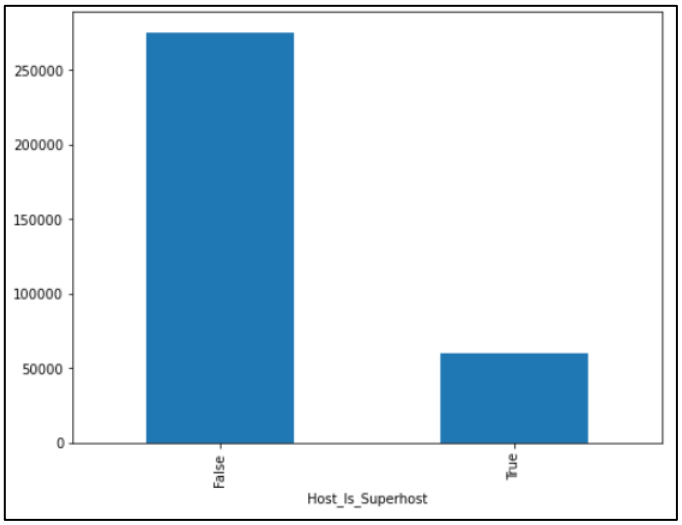
Hasil

Hasil yang diperoleh dari pengujian yang dilakukan menggunakan Confusion Matrix, Precision, Recall, dan F1-SCORE, serta Accuracy selanjutnya dilakukan perbandingan hasil terhadap penerapan algoritma SVM, penerapan metode ADASYN pada algoritma SVM (ADASYN SVM) dan penerapan metode SMOTE pada algoritma SVM (SMOTE SVM). Berdasarkan hasil tersebut akan dibandingakan dan dibahas hasil performa klasifikasi yang paling maksimal dari ketiganya. Dari dataset Airbnb yang diperoleh terdapat 1048575 data dengan beberapa kolom yang berisi nilai Nan (kosong). Data mentah ditunjukkan pada tabel 2.

Tabel 2. Data Mentah

No	Listing ID	Host ID	Host Response Rate	…	Reviews per month
0	13588243	4386046	90%	…	NaN
1	5534229	28697142	100%	…	0
2	18417052	81453949	100%	…	NaN
…	…	…	…	…	…
1048574	11225183.0	42821904.0	100%	…	0.38

Setelah mendapatkan data mentahnya, selanjutnya data yang memiliki nilai kosong akan diabaikan dan data yang digunakan berasal dari kolom *Host Response Rate*, *Number of reviews*, *Review Scores Rating*, dan *Host total listing count*. Setelah dilakukan pemeriksaan data diperoleh data latih pada 2 label, 274911 label data false dan 59571 label data true. Gambar 2 menunjukkan jumlah data pelatihan pada dua label tersebut.



Gambar 2. Jumlah data latih pada dua label.

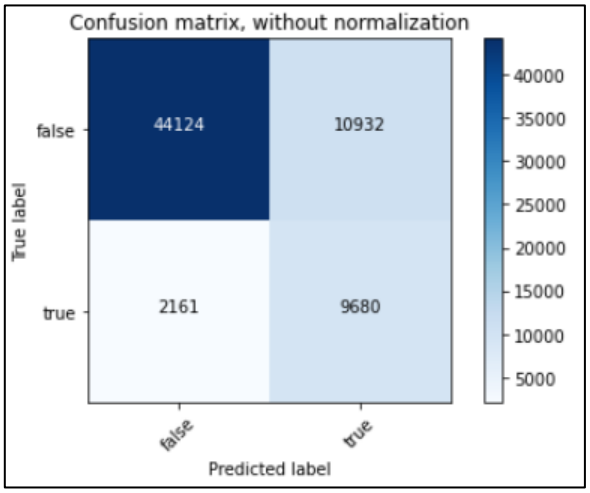
Berikut sebaran data setelah dilakukan pembersihan data pada kolom terpilih berdasarkan kriteria yang mempengaruhi apakah suatu host adalah super host atau non-super host seperti yang ditunjukkan pada Tabel 3. Data dibagi menjadi 80% untuk data latih dan 20% untuk data uji. Dari pembagian data tersebut diolah menjadi algoritma SVM dengan kernel Radial Basis Function (RBF) dan menerapkan parameter class_weight = 'balanced' untuk menangani kelas yang tidak seimbang. Hasil Confusion Matrix dari klasifikasi tersebut dijelaskan untuk memberikan informasi tentang TP, FP, TN, dan FN. Gambar 3 menampilkan Confusion Matrix SVM. Berikut normalisasi Confusion Matrix yang ditunjukkan pada gambar 4.

Tabel 3. Data Latih

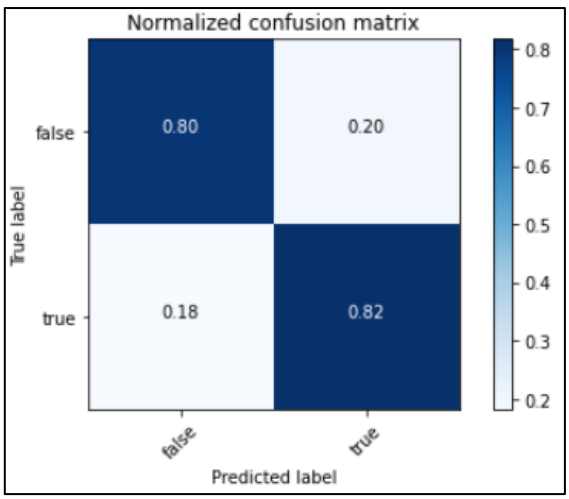
No	*Host Response Rate*	*Number of reviews*	*Review Scores Rating*	*Host total listings count*	*Host Is Super host*
0	1	2	90	5	0
1	0.88	3	87	2	0
2	1	6	100	1	0
…	…	…	…	…	…
334481	1	5	90	3	0

Karena terjadi *oversampling* mengakibatkan hasil yang tidak seimbang pada kedua label, maka diterapkan ADASYN dan SMOTE untuk mengatasi *oversampling* tersebut. Setelah menerapkan ADASYN data latih menjadi 274911 data label false dan 261120 data label *true*. Distribusi data latih pada penerapan ADASYN ditunjukkan pada tabel 4. Dalam penerapan SMOTE pada data latih tidak seimbang, diperoleh sebaran data 274911 data label *false* dan 274911 data untuk label true yang ditunjukkan pada tabel 5.

Pada penerapan ADASYN dan SMOTE pada *dataset imbalanced*, dilakukan penerapan SVM pada data latih ADASYN dan SMOTE yang dibentuk untuk mengklasifikasikan apakah suatu host merupakan super *host* atau non-super *host*. Untuk menghasilkan perbandingan antara presisi, recall, F1-Score dan akurasi pada setiap implementasi, berikut hasil pengujian pada setiap implementasi ditunjukkan pada tabel 6. Hasil ini menunjukkan bahwa pengujian dari implementasi SVM, ADASYN SVM, dan SMOTE SVM dan akurasi terbaik diperoleh pada SMOTE SVM dengan akurasi 0.812.



Gambar 3. Confusion Matrix dari SVM



Gambar 4. Normalisasi Confusion Matrix dari SVM

Tabel 4. Training Data dari ADASYN

No	Host Response Rate	Number of reviews	Review Scores Rating	Host total listings count	Host Is Super host
0	1	2	90	5	0
1	0.88	3	87	2	0
2	1	6	100	1	0
…	…	…	…	…	…
536030	1	82	99.64	3	1

Tabel 5. Training Data dari SMOTE

No	Host Response Rate	Number of reviews	Review Scores Rating	Host total listings count	Host Is Super host
0	1	2	90	5	0
1	0.88	3	87	2	0
2	1	6	100	1	0
…	…	…	…	…	…
549821	1	44	95	2	1

Tabel 6. Perbandingan Hasil Pengujian dari SVM, ADASYN SVM dan SMOTE SVM

Algoritma	Label	Precision	Recall	F1-Score	Accuracy
SVM	False	0.953	0.810	0.871	0.804
	True	0.470	0.817	0.597	
ADSYN	False	0.795	0.728	0.760	0.764
SVM	True	0.737	0.802	0.768	
SMOTE	False	0.819	0.802	0.810	0.812
SVM	True	0.806	0.823	0.815	

Pembahasan

Berdasarkan hasil implementasi algoritma Support Vector Machine (SVM) dengan kernel Radial Basis Function (RBF) yang diaplikasikan pada klasifikasi super host Airbnb diperoleh hasil pengujian yang tidak seimbang, pada Tabel 6 pengujian nilai presisi untuk label false 0.953 sedangkan untuk label true 0.470, dan nilai F1-Score untuk label false adalah 0.871 sedangkan untuk label true 0.597, namun hasil recall memiliki nilai yang baik pada kedua label dengan nilai 0.810 untuk label false dan 0.817 untuk label true dan Confusion Matrix memiliki nilai akurasi sebesar 0.804. Ketidakseimbangan hasil pengujian disebabkan dataset yang tidak seimbang, maka diterapkan algoritma ADASYN dan SMOTE.

ADASYN mempengaruhi hasil pengujian persamaan pada Tabel 6. Hasil pengujian pada nilai presisi mendapatkan nilai 0.795 untuk label false dan 0.737 untuk label true, hasil nilai recall pada label false 0.728 dan 0.802 untuk label true, dan hasil nilai F1-Score 0.760 untuk label false dan 0.768 untuk label true. Berdasarkan Confusion Matrix didapatkan nilai akurasi sebesar 0.764, hasil yang didapatkan membuat persamaan hasil pada kedua label tetapi menurunkan nilai akurasinya dibandingkan tidak menggunakan algoritma ADASYN.

Algoritma SMOTE diterapkan untuk mengetahui perbandingan penggunaan algoritma ADASYN dan SMOTE, penerapan algoritma ADASYN memiliki hasil pengujian ditunjukkan pada tabel 6, hasil presisi mendapatkan 0.819 untuk label false dan 0.806 untuk label true, recall untuk label false 0.819 dan label true 0.823 dan F1-Score 0.810 untuk label false dan 0.815 untuk label true. Hasil akurasi Confusion Matrix yang didapat yaitu 0.812.

Berdasarkan tabel 6 diperoleh hasil terbaik dengan menerapkan algoritma SMOTE pada SVM, terjadi kenaikan pada label True (minoritas) pada pengujian presisi dari 0.470 menjadi 0.806, pada *recall* dari 0.817 menjadi 0.823, pada F1-Score dari 0.597 menjadi 0.815. Namun terjadi penurunan pada label False (mayoritas) pada pengujian presisi dari 0.953 menjadi 0.819, pada *recall* dari 0.810 menjadi 0.802 dan pada F1-Score dari 0.871 menjadi 0.810. Sehingga pada perhitungan akurasi mengalami kenaikan dari 0.804 menjadi 0.812 atau menambahkan tingkat akurasi sebesar 0.008. Pada penelitian ini didapatkan bahwa penerapan SMOTE pada SVM lebih baik dibandingkan dengan penerapan ADASYN pada SVM karena sebaran data yang diperoleh setelah penerapan SMOTE memiliki distribusi data

yang seimbang, data latih *false* label sebanyak 274911 data dan *true* label sebanyak 274911 data. Sedangkan dengan penerapan ADASYN dapat menyeimbangkan *dataset* pada kedua label tetapi tidak memiliki jumlah yang sama pada data latih, label *false* memiliki data 274911 sedangkan label true memiliki data 261120.

SIMPULAN

Temuan paling jelas yang muncul dari penelitian ini adalah algoritma SVM dengan mengolah imbalanced dataset menghasilkan ketepatan dan F1-score yang tidak seimbang untuk setiap label. Penerapan algoritma ADASYN dan algoritma SMOTE dapat memberikan hasil keseimbangan yang lebih baik dari pengujian sebelumnya, penerapan algoritma ADASYN pada algoritma SVM memperoleh kesetaraan hasil pengujian dari label true dan false tetapi pada label false menyebabkan penurunan presisi dan hasil uji F1-Score dibandingkan tanpa menggunakan algoritma ADASYN. Penerapan algoritma SMOTE pada algoritma SVM memberikan hasil pengujian presisi, recall, F1-Score, dan akurasi yang lebih baik dibandingkan dengan ADASYN-SVM. Dapat disimpulkan bahwa penerapan algoritma ADASYN dan algoritma SMOTE dapat memberikan kesetaraan hasil pengujian antar label klasifikasi, namun penerapan algoritma tersebut berpengaruh pada penurunan hasil pengujian terutama pada label yang memiliki dataset paling banyak tetapi memberikan hasil yang baik untuk label yang memiliki dataset terkecil. Penerapan algoritma SMOTE pada SVM meningkatkan tingkat akurasi didapatkan dari kenaikan pada hasil perhitungan pada pengujian label True (minoritas) pada pengujian presisi dari 0.470 menjadi 0.806, pada *recall* dari 0.817 menjadi 0.823, pada F1-Score dari 0.597 menjadi 0.815 dan berdasarkan penurunan label *False* (mayoritas) pada pengujian presisi dari 0.953 menjadi 0.819, pada recall dari 0.810 menjadi 0.802 dan pada F1-Score dari 0.871 menjadi 0.810. Sehingga pada perhitungan akurasi mengalami kenaikan dari 0.804 menjadi 0.812 atau menambahkan tingkat akurasi sebesar 0.008.

REFERENSI

Ahmad, I., Basheri, M., Iqbal, M. J., & Rahim, A. (2018). Performance Comparison of Support Vector Machine, Random Forest, and Extreme Learning Machine for Intrusion Detection. *IEEE Access*, *6*, 33789–33795. https://doi.org/10.1109/ACCESS.2018.2841987

Alsmadi, I., & Hoon, G. K. (2019). Term weighting scheme for short-text classification: Twitter corpuses. *Neural Computing and Applications*, *31*(8), 3819–3831. https://doi.org/10.1007/s00521-017-3298-8

Chawla, N. V., Bowyer, K. W., Hall, L. O., & Kegelmeyer, W. P. (2002). SMOTE: Synthetic Minority Over-sampling Technique. *Journal of Artificial Intelligence Research 16*, 321–357. https://doi.org/10.1002/eap.2043

Chen, C. C., & Chang, Y. C. (2018). What drives purchase intention on Airbnb? Perspectives of consumer reviews, information quality, and media richness. *Telematics and Informatics*, *35*(5), 1512–1523. https://doi.org/10.1016/j.tele.2018.03.019

Crommelin, L., Troy, L., Martin, C., & Pettit, C. (2018). Is Airbnb a Sharing Economy Superstar? Evidence from Five Global Cities. *Urban Policy and Research*, *36*(4), 429–444. https://doi.org/10.1080/08111146.2018.1460722

Düntsch, I., & Gediga, G. (2020). Indices for rough set approximation and the application to confusion matrices. *International Journal of Approximate Reasoning*, *118*, 155–172. https://doi.org/10.1016/j.ijar.2019.12.008

Fico, G., Montalva, J., Medrano, A., Liappas, N., Cea, G., & Arredondo, M. T. (2018). EMBEC & NBC 2017. *IFMBE Proceedings*, *65*, 1089–1090. https://doi.org/10.1007/978-981-10-5122-7

Guttentag, D., Smith, S., Potwarka, L., & Havitz, M. (2018). Why Tourists Choose Airbnb: A Motivation-Based Segmentation Study. *Journal of Travel Research*, *57*(3), 342–359. https://doi.org/10.1177/0047287517696980

Harianto, H., Sunyoto, A., & Sudarmawan, S. (2020). Optimasi Algoritma Naïve Bayes Classifier untuk Mendeteksi Anomaly dengan Univariate Fitur Selection. *Edumatic: Jurnal Pendidikan Informatika*, *4*(2), 40–49. https://doi.org/10.29408/edumatic.v4i2.2433

He, H., Bai, Y., Garcia, E. A., & Li, S. (2008). ADASYN: Adaptive synthetic sampling approach for imbalanced learning. *Proceedings of the International Joint Conference on Neural Networks*, (3), 1322–1328. https://doi.org/10.1109/IJCNN.2008.4633969

Kusumawati, R., D'Arofah, A., & Pramana, P. A. (2019). Comparison Performance of Naive Bayes Classifier and Support Vector Machine Algorithm for Twitter's Classification of Tokopedia Services. *Journal of Physics: Conference Series*, *1320*(1), 0–10. https://doi.org/10.1088/1742-6596/1320/1/012016

Patil, N. M., & Nemade, M. U. (2017). Music Genre Classification Using MFCC , K-NN and SVM Classifier. *International Journal of Computer Applications*, *4*(2), 43–47.

Pucci, F., & Rooman, M. (2017). Airbnb recsys. *Kdd*, 311–320. https://doi.org/10.1145/3219819.3219885

Rimal, B., Rijal, S., & Kunwar, R. (2020). Comparing Support Vector Machines and Maximum Likelihood Classifiers for Mapping of Urbanization. *Journal of the Indian Society of Remote Sensing*, *48*(1), 71–79. https://doi.org/10.1007/s12524-019-01056-9

Rustam, Z., & Audia Ariantari, N. P. A. (2018). Support Vector Machines for Classifying Policyholders Satisfactorily in Automobile Insurance. *Journal of Physics: Conference Series*, *1028*(1). https://doi.org/10.1088/1742-6596/1028/1/012005

Sari, V., Firdausi, F., & Azhar, Y. (2020). Perbandingan Prediksi Kualitas Kopi Arabika dengan Menggunakan Algoritma SGD, Random Forest dan Naive Bayes. *Edumatic: Jurnal Pendidikan Informatika*, *4*(2), 1–9. https://doi.org/10.29408/edumatic.v4i2.2202

Thanh Noi, P., & Kappas, M. (2017). Comparison of Random Forest, k-Nearest Neighbor, and Support Vector Machine Classifiers for Land Cover Classification Using Sentinel-2 Imagery. *Sensors (Basel, Switzerland)*, *18*(1). https://doi.org/10.3390/s18010018

Zubrinic, K., Milicevic, M., & Zakarija, I. (2013). Comparison of Naïve Bayes and SVM Classifiers in Categorization of Concept Maps. *International Journal of Computers*, *7*(3), 109–116.