PERBANDINGAN HASIL METODE *SUPPORT VECTOR MACHINE* (SVM) DENGAN *ENSEMBLE SMOTE BAGGING* DAN *SMOTE BOOSTING* PADA DATA KELULUSAN MAHASISWA UNIMUS

JURNAL ILMIAH

Diajukan sebagai salah satu syarat untuk memperoleh gelar Sarjana Statistika

Oleh

NABILA AFFAH MUMTAZAHH

B2A017037

PROGRAM STUDI S1 STATISTIKA
FAKULTAS MATEMATIKA DAN ILMU PENGETAHUAN ALAM
UNIVERSITAS MUHAMMADIYAH SEMARANG
2021

HALAMAN PENGESAHAN

Skripsi dengan Judul "**Perbandingan Hasil Metode *Support Vector Machine* (SVM) dengan *Ensemble Smote Bagging* dan *Smote Boosting* Pada Data Kelulusan UNIMUS**" yang disusun oleh :

Nama : Nabila Affah Mumtazah

NIM : B2A017037

Program Studi : S-1 STATISTIKA

telah disetujui oleh dosen pembimbing pada tanggal : 1 Mei 2021

Pembimbing Utama	Pembimbing Pendamping
Tiani Wahyu Utami, S.Si., M.Si	Dr. Rochdi Wasono, M.Si
NIK. 28.6.1026.341	NIK. 28.6.1026.119

Mengetahui,
Ketua Program Studi Statistika

Indah Manfaati Nur, S.Si., M.Si
NIK. 28.6.1026.221

SURAT PERNYATAAN
PUBLIKASI KARYA ILMIAH

Yang bertandatangan di bawah ini, saya :

Nama	: Nabila Affah Mumtazah
NIM	: B2A017037
Fakultas/Jurusan	: Matematika dan Ilmu Pengetahuan Alam/Statistika
Jenis Penelitian	: Skripsi
Judul	: Perbandingan Hasil Metode *Support Vector Machine* (SVM) dengan *Ensemble Smote Bagging* dan *Smote Boosting* Pada Data Kelulusan UNIMUS
Email	: nabilaaffah11@gmail.com

Dengan ini menyatakan bahwa saya menyetujui untuk :

1. Memberikan hak bebas royalti kepada Perpustakaan Unimus atas penulisan karya ilmiah saya, demi pengembangan ilmu pengetahuan.
2. Memberikan hak menyimpan, mengalih mediakan/ mengalih formatkan, mengelola dalam bentuk pangkalan data (*database*), mendistribusikannya, serta menampilkannya dalam bentuk *softcopy* untuk kepentingan akademis kepada Perpustakaan Unimus, tanpa perlu meminta ijin dari saya selama tetap mencantumkan nama saya sebagai penulis/pencipta.
3. Bersedia dan menjamin untuk menganggung secara pribadi tanpa melibatkan pihak perpustakaan Unimus, dari semua bentuk tuntutan hukum yang timbul atas pelanggaran hak cipta dalam karya ilmiah ini.

Demikian pernyataan ini saya buat dengan sesungguhnya dan semoga dapat digunakan sebagaimana mestinya.

Semarang, 2 Mei 2021
Yang Menyatakan

(Nabila Affah Mumtazah)
NIM. B2A.017.037

PERBANDINGAN HASIL METODE *SUPPORT VECTOR MACHINE* (SVM) DENGAN *ENSEMBLE SMOTE BAGGING* DAN *SMOTE BOOSTING* PADA DATA KELULUSAN MAHASISWA UNIMUS

Oleh: Nabila Affah Mumtazah
Program Studi Statistika, Univeristas Muhammadiyah Semarang
e-mail: nabilaaffah11@gmail.com

Article history
Submission :
Revised :
Accepted :

Keyword:
Bagging, Boosting, SMOTE, SVM.

Abstract

Imbalanced data is a condition in which the amount of data that represents one class (minority) is very small compared to other classes (majorities). The data used in this research is the graduation data of the 2010-2020 unimus students which are large in size, namely 87% of the data are in the on Timecategory,while only 13% are not on time so they are categorized data. The SVM classification method cannot solve this problem because it will be biased towards the major class and have low performance in the minor class. So that a new method is needed to solve the problem of imbalanced data, so we introduce the smote bagging and smote boosting methods. The purpose of using the smote bagging and smote boosting methods is to solve the imbalanced problem of the SVM classification results obtained. In addition, this study divides training and testig data using 10-fold cross validation. The results of the classification can provide input to the Muhammadiyah University of Semarang in the formulation of policies in order to minimize the length of student graduation studies based on the factors that influence it. The result showed that the selected smote bagging model had a G-mean value of 0.87721, which means that the model was good at explaining the accuracy of student graduation compared to smote boosting with a G-mean of 0.86779.

PENDAHULUAN

Pendidikan merupakan salah satu cara dalam meningkatkan kualitas Sumber daya manusia (SDM) dan aset penting bagi kemajuan bangsa. Melalui pendidikan, SDM yang berkualitas baik dari segi spiritual, intelegensi serta keterampilan dapat disiapkan. Adanya SDM yang berkualitas, sosok-sosok individu ini diharapkan dapat berperan dalam proses pembangunan bangsa dan Negara. Salah satu cara agar dapat membentuk manusia berkualitas ialah berdasarkan tingkat pendidikan yang semakin tinggi (Profil Anak Indonesia, 2019).

Universitas Muhammadiyah Semarang (UNIMUS) merupakan salah satu perguruan tinggi swasta yang berada di Semarang, Jawa Tengah, Indonesia. UNIMUS didirikan pada tanggal 4 agustus 1999 dengan program studi yang memperoleh ijin operasional pada awal pembukaan tahun 1999 sebanyak 14 program studi. Namun seiring berjalannya waktu UNIMUS telah tumbuh berkembang menjadi tempat pembelajaran yang terpilih. Hingga pada akhirnya UNIMUS membuka beberapa program studi baru, pada tahun 2020 ini Universitas Muhammadiyah Semarang memiliki total 8 fakultas, dengan 4 (empat) program Diploma tiga, 1 (satu) program Diploma empat, 2 (dua) program pascasarjana, 1 (satu) program studi profesi ners, 1 (satu) program studi profesi doktor gigi, 1 (satu) program studi profesi dokter dan 15 program sarjana. Setiap tahun Universitas Muhammadiyah Semarang mengadakan perayaan wisuda untuk mahasiswa yang dinyatakan lulus, tetapi juga masih banyak mahasiswa Universitas Muhammadiyah Semarang di 8 fakultas tersebut yang tidak lulus tepat waktu

Teknik klasifikasi bertujuan untuk menemukan suatu fungsi keputusan yang secara akurat memprediksi kelas dari data *testing* yang berasal dari fungsi distribusi yang sama dengan data untuk *training*. Oleh karena itu terdapat dua kondisi himpunan kelas data, yaitu *balance* dan *imbalanced class* data. *Imbalanced class* terjadi ketika satu kelas melebihi jumlah kelas lainnya. Kelas data banyak disebut kelas mayoritas (kelas negatif) sedangkan kelas data sedikit disebut kelas minoritas (kelas positif). Dalam kondisi seperti data ketepatan waktu kelulusan UNIMUS, sebagian besar *classifier* bias terhadap kelas mayor, karena mesin klasifikasi akan condong memprediksi ke kelas mayor dan mengabaikan kelas minor (Japkowicz dan Stephen, 2002).

Salah satu metode dalam klasifikasi adalah *Support Vector Machine* (SVM). SVM merupakan metode klasifikasi non parametrik yang tidak harus memenuhi asumsi dan distribusi tertentu (Sahitayakti & Fithriasari, 2015). Teknik SVM digunakan untuk menemukan fungsi pemisah (*classifier*) yang optimal yang bisa memisahkan dua set data dari dua kelas yang berbeda. Kelebihan utama SVM adalah dengan *support vectornya* sudah mewakili semua data yang ingin diklasifikasikan berbeda dengan metode lain yang mengharuskan semua data diinput untuk diklasifikasikan sehingga menghasilkan *performance* yang baik. Selain itu SVM juga baik dalam hal prediksi untuk klasifikasi (Sain & Purnami, 2015). Sementara itu kelebihan lainnya yaitu dalam menentukan jarak menggunakan *support vector* sehingga proses komputasi menjadi cepat (Octaviani, et al., 2014). Pada umumnya data dalam dunia nyata jarang yang bersifat *linier separable*, kebanyakan bersifat *non-liniear*. Untuk menyelesaikan masalah *non-liniear*, SVM dimodifikasi dengan memasukkan fungsi kernel (Johra, 2018). Metode ini sudah baik dalam melakukan klasifikasi ketika jumlah kelas dari variabel label dalam data *balanced* akan tetapi jika data yang digunakan *imbalanced*, akan berdampak pada sulitnya mendapatkan model prediksi yang baik dan bermakna karena adanya ketidakcukupan informasi dari kelas minor (Yap dkk, 2014). Metode klasifikasi ini akan bias terhadap kelas mayor dan memiliki kinerja rendah pada kelas minor (Batuwita & Palade, 2012).

Secara umum pendekatan berbasis sampling dibagi menjadi dua yaitu metode *oversampling* dan *undersampling*. Metode *undersampling* menyeimbangkan data dengan cara menghapus beberapa pengamatan pada kelas mayor hingga keseimbangan data *training* yang diinginkan tercapai. Pendekatan berbasis sampling yang kedua yaitu *oversampling*. Metode ini bekerja untuk menyeimbangkan data *training*, jika data yang digunakan untuk membuat model tidak seimbang maka akan meningkatkan kesalahan dalam klasifikasi kelas minor dengan cara meningkatkan jumlah data pada kelas minor. Oleh karena itu, salah satu alternatif paling efektif untuk meningkatkan akurasi model adalah melakukan *Synthetic Minority Oversampling Technique* (SMOTE) pada praproses (Barro, et al., 2013). Selain itu, alternatif lain untuk meningkatkan nilai akurasi kelas *imbalanced* adalah dengan menggunakan metode ensemble. Metode *ensemble* pada prinsipnya mengkombinasikan sekumpulan *classifier* yang dilatih dengan tujuan untuk membuat model klasifikasi (*classifier*) campuran yang terimprovisasi sehingga membuat *classifier ensemble* yang terbentuk lebih akurat dari pada classifier asalnya dalam melakukan suatu pengklasifikasian (Han dkk, 2012).

Boosting (Freud dan Schapire, 1997) dan *Bagging* (Breiman, 1996) merupakan metode *ensemble* yang paling popular digunakan. *Boosting* dan *Bagging* adalah salah satu metode *ensemble* yang berbasis variasi data ensemble, yang terdiri dari memanipulasi data *training* sedemikian rupa sehingga masing-masing *classifier* dilatih dengan data *training* yang berbeda. Metode *Bagging* didasarkan pada gagasan membuat berbagai sampel dari data *training*. Untuk variasi dari data *training* akan dihasilkan model klasifikasi tertentu, kemudian hasilnya akan diberikan sebagai kombinasi atau gabungan model.

Pada prinsipnya *Boosting* membentuk satu *classifier* yang kuat dengan mengkombinasikan sekumpulan

classifier.*Boosting* mempertahankan sekumpulan bobot pengamatan pada saat *training* pengamatan dan secara adaptif menyesuaikan (updating) bobot-bobot ini pada akhir tiap iterasi *boosting*. Bobot-bobot dari pengamatan yang salah terklasifikasikan pada saat training akan dinaikkan sementara bobot-bobot pengamatan yang terklasifikasikan dengan benar akan diturunkan nilainya, dengan kata lain *Boosting* memaksa suatu *classifier* untuk memberi perhatian yang lebih pada pengamatan yang salah diklasifikasikan (Li dkk, 2008). Namun, karena desainnya yang berorientasi pada akurasi, algoritma metode *ensemble* yang secara langsung diterapkan ke data yang *imbalanced* tidak bisa menyelesaikan masalah pengklasifikasi. Dengan mengkombinasikan *ensemble* dengan teknik lain untuk mengatasi masalah *imbalanced* data telah dilakukan dan menghasilkan nilai yang positif (Galar dkk, 2011).

Chawla dkk (2003) menggunakan metode *smote boosting* untuk pengklasifikasian *imbalanced data* dengan rasio *imbalanced* sebesar 71% untuk mayoritas kelasnya dan 29% untuk minoritas kelasnya, hasil penelitiannya menunjukkan bahwa beberapa set data yang tidak seimbang menunjukkan bahwa 86% 14% Sekolah Putus Sekolah 5 algoritma *smote boosting* yang diusulkan dapat menghasilkan prediksi yang lebih baik dari kelas minoritas daripada *AdaBoost*, *AdaCost*. *SMOTEBoost* secara implisit meningkatkan bobot pada kelas minoritas yang salah diklasifikasikan (*false negative*) dalam distribusi jumlah kelas minoritas akan meningkat menggunakan algoritma SMOTE. Oleh karena itu, dalam ukuran resample *boosting* peningkatan *SMOTEBoost* mampu membuat wilayah keputusan yang lebih luas untuk kelas minoritas dibandingkan dengan *boosting* standar. Chawla dkk menyimpulkan bahwa *SMOTEBoost* dapat membangun pengklasifikasian dan mengurangi bias pada pengklasifikasian. *SMOTEBoost* menggabungkan kekuatan SMOTE untuk meningkatkan nilai recall dan *boosting* untuk meningkatkan nilai *precision*. Secara keseluruhan didapatkan ukuran *performance F-value* yang lebih baik.

Begitu halnya dengan *SMOTE-Bagging* yang menambahkan algoritma SMOTE di tiap prosedur resampling-nya. Tujuan dari adanya SMOTE yaitu untuk menambah *probabilitas* terpilihnya sampel-sampel yang sulit diklasifikasikan yang berasal dari kelas minor ke dalam data training di tiap iterasi sehingga membuat *base classifier* lebih fokus pada pengamatan kelas minor. Hal ini tentunya akan meningkatkan ketepatan klasifikasi pada kelas minoritas. Kemudian SMOTE yang dikombinasikan dengan prosedur *Bagging* memberikan kinerja keseluruhan (G-Mean) mengalami peningkatan (Wang, 2009).

Setelah mempelajari dan memahami beberapa penelitian terdahulu yang berkaitan dengan metode dan objek yang digunakan pada penelitian ini, maka dapat diketahui perbedaan yang dimiliki dari penelitian ini dengan penelitian- penelitian sebelumnya yang terletak pada objek yang digunakan dengan membandingkan beberapa metode klasfifikasi. Objek yang digunakan pada penelitian ini adalah data kelulusan mahasiswa UNIMUS yang diklasifikasikan menggunakan 3 metode klasifikasi yaitu *SMOTE bagging* SVM dan *SMOTE boosting* SVM yang kemudian hasil dari ketiga metode tersebut dibandingkan metode mana yang menghasilkan akurasi paling baikSebelum memasuki dunia investasi diperlukan pengetahuan keuntungan dan risiko yang akan didapatkan ketika mengambil langkah selanjutnya. Risiko investasi disini diartikan sebagai kemungkinan terjadinya perbedaan antara keuntungan yang aktual dengan keuntungan yang diharapkan. Risiko maupun harapan keuntungan dalam berinvestasi selalu ada dan berdampingan. Dalam berinvestasi disamping menghitung keuntungan yang diharapkan investor juga perlu memperhatikan risiko yang akan ditanggung (Abdul Halim, 2005).

berdasarkan permasalahan dan penjelasan yang telah diuraikan, peneliti tertarik untuk melakukan penelitian tentang perbandingan hasil metode *support vector machine* (svm) dengan *ensemble smote bagging*

dan *smote boosting* pada data kelulusan mahasiswa unimus

LANDASAN TEORI

Kelulusan Mahasiswa

Berdasarkan ketetapan Kementerian Pendidikan dan Kebudayaan Direktorat Jenderal Pendidikan Tinggi tentang Sistem Pendidikan Tinggi disebutkan bahwa untuk memenuhi standar kompetensi lulusan bagi mahasiswa program sarjana (S1) beban wajib yang harus ditempuh adalah paling sedikit 144-160 satuan kredit semester (sks) dengan masa studi selama 8-12 semester atau 4-6 tahun. Sedangkan Waktu lama studi untuk jenjang D3 lama studinya yaitu 6 semester (36 bulan), untuk jenjang S1 lama studinya yaitu 8 semester (48 bulan), dan untuk jenjang S2 lama studinya yaitu 4 semester (24 bulan). Seorang mahasiswa dinyatakan lulus program apabila telah menyelesaikan minimal SKS sesuai dengan kurikulum masing-masing Program Studi dengan Indeks Prestasi Kumulatif (IPK) minimal 2.00 dan menyelesaikan Tugas Akhir dan/atau skripsi dan telah mempublikasi-kan karya ilmiah untuk program S1 atau telah menyelesaikan Tugas Akhir atau Karya Tulis.

Klasifikasi

Menurut Bramer (2007) klasifikasi merupakan tugas yang sering terjadi dalam kehidupan sehari-hari. Pada dasarnya klasifikasi membagi objek kesalah satu dari sejumlah kategori yang disebut kelas, sedangkan menurut Han dkk (2012) klasifkasi adalah proses menemukan sebuah model atau fungsi yang menggambarkan dan membedakan kelas data atau konsep. Model didasarkan pada 10 analisis set data training yaitu objek data untuk label kelas yang sudah diketahui, model ini digunakan untuk memprediksi kelas label dari objek label kelas yang belum diketahui. Masalah prediksi merujuk pada masalah dimana variabel yang akan diprediksi memiliki jumlah nilai yang terbatas yaitu variabel kategorik.

Dalam pengklasifikasian data terdapat dua proses yang akan dilakukan yaitu:

1. Proses *Training*
   Pada proses ini digunakan training set yang telah diketahui labelnya untuk membangun model atau fungsi.
2. Proses *Testing*
   Pada proses ini untuk mengetahui akurasi model atau fungsi yang akan dibangun pada proses training, maka digunakan data yang disebut dengan testing set untuk memprediksi labelnya.

Support Vector Machine (SVM)

SVM merupakan metode klasifikasi non parametrik yang tidak harus memenuhi asumsi dan distribusi tertentu (Sahitayakti & Fithriasari, 2015). Teknik SVM digunakan untuk menemukan fungsi pemisah (*classifier*) yang optimal yang bisa memisahkan dua set data dari dua kelas yang berbeda. Kelebihan utama SVM adalah dengan *support vectornya* sudah mewakili semua data yang ingin diklasifikasikan berbeda dengan metode lain yang mengharuskan semua data diinput untuk diklasifikasikan sehingga menghasilkan *performance* yang baik. Selain itu SVM juga baik dalam hal prediksi untuk klasifikasi (Sain & Purnami, 2015). Sementara itu kelebihan lainnya yaitu dalam menentukan jarak menggunakan *support vector* sehingga proses komputasi menjadi cepat (Octaviani, et al., 2014). Pada umumnya data dalam dunia nyata jarang yang bersifat *linier separable*, kebanyakan bersifat *non-liniear*. Untuk menyelesaikan masalah *non-liniear*, SVM dimodifikasi dengan memasukkan fungsi kernel (Johra, 2018). Metode ini sudah baik dalam melakukan klasifikasi ketika jumlah kelas dari variabel label dalam data *balanced* akan tetapi jika data yang digunakan *imbalanced*, akan berdampak pada sulitnya mendapatkan model prediksi yang baik dan bermakna karena adanya ketidakcukupan informasi dari kelas minor (Yap dkk, 2014). Metode klasifikasi ini akan bias terhadap kelas mayor dan memiliki kinerja rendah pada kelas minor (Batuwita & Palade, 2012).

Imbalanced Data

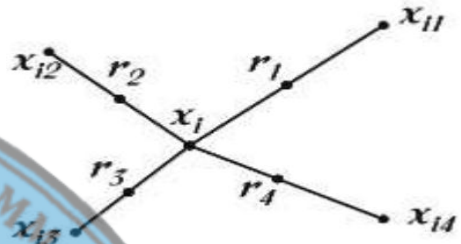
Imbalanced class terjadi ketika satu kelas melebihi jumlah kelas lainnya. Kelas data banyak disebut kelas mayoritas (kelas negatif) sedangkan kelas data sedikit disebut kelas minoritas (kelas positif). Permasalahan *imbalanced class* banyak terjadi pada klasifikasi data dalam dunia nyata dan masalahnya terdiri dari jumlah disproporsi kelas yang berbeda. Imbalanced merupakan suatu kondisi dimana jumlah dari data yang mempresentasikan satu kelas (minoritas) sangatlah kecil dibandingkan kelas lain (mayorits). Sebagai ilustrasi misalkan sebuah data imbalanced dengan rasio 1:100 untuk masing-masing objek kelas minoritas, terdapat 100 objek kelas mayoritas (Pangastuti, 2018).

Beberapa teknik yang telah dikembangkan untuk mengatasi masalah *imbalanced data*, teknik ini dikategorikan kedalam empat kelompok yaitu sebagai berikut (Fernandez dkk, 2018):

1. *Algorithm Level* bekerja dengan mencoba untuk menyesuaikan algoritma dengan memperhitungkan kelas minoritas (positif).
2. Data Level dengan menyeimbangkan distribusi kelas dengan resampling ruang data.
3. *Cost-Sensitive* adalah menggabungkan algorithm level dan data level dengan mengasumsikan kesalahan biaya klasifikasi pada kelas minoritas dan berusaha meminimalkan total kesalahan biaya dari dua kelas.
4. *Ensemble Method* adalah kombinasi antara algoritma *ensemble* dan salah satu teknik sebelumnya, khusunya data level dan cost-sensitive.

Synthetic Minority Oversampling Technique (SMOTE)

Teknik SMOTE merupakan salah satu metode oversampling yang bekerja dengan menerapkan metode sampling untuk meningkatkan jumlah kelas minoritas (positif) melalui replikasi data secara acak, sehingga sejumlah data minoritas sama dengan data mayoritas. Algoritma SMOTE pertama kali diperkenalkan oleh Nithes V. Chawla (2002), ide utama pendekatan ini bekerja dengan membangun data 23 *synthetic* replikasi data minor. Algoritma SMOTE dilakukan untuk mendefinisikan *k-nearest neighbor* (KNN) untuk setiap kelas minoritas, kemudian menyusun duplikasi data synthetic sebanyak persentase yang diinginkan antara kelas minoritas dan KNN yang dipilih secara acak (Sain & Purnami, 2015).



Gambar 2. 1 Contoh Pembangkit SMOTE

(Sumber: Fernandez dkk, 2018)

Pada Gambar 2.4 terlihat bahwa xi adalah data ke-i dari kelas minoritas yang dipilih sebagai dasar untuk membangkitkan data *synthetic* baru, sehingga didapatkan *synthetic* baru yaitu $r1$ sampai $r4$. Chawla dkk (2002) menunjukkan bahwa pendekatan SMOTE dapat meningkatkan akurasi klasifikasi untuk kelas minoritas. Kombinasi SMOTE dan metode *ensemble* memiliki *performance* yang lebih baik daripada metode SMOTE standar.

Bagging

Breiman memperkenalkan konsep *bootstrap aggregating* dengan tujuan untuk memanipulasi data training dengan mengganti data training asli T secara acak menjadi N item (Yongqing dkk, 2014). Algoritma ini juga digunakan untuk membangun model *ensembles*. Model ensemble ini terdiri dari beberapa pelatihan pengklasifikasian dengan *bootstrap* replika yang berbeda dari training

dataset asli untuk melatih setiap classifier. Dataset baru dibentuk secara acak (dengan penggantian) objek dari dataset asli. Ketika sebuah objek yang tidak diketahui akan diklasifikasikan maka voting mayoritas atau bobot yang akan digunakan untuk mendapatkan kelasnya. Voting mayoritas dilakukan dengan menggunakan confidence yang diberikan oleh masing-masing pengklasifikasian dalam prediksi. Salah satu kelebihan bagging adalah kesederhanaannya. Selain itu, *bagging* dapat mengurangi varians. Karena efek voting mirip dengan rata-rata dalam regresi di mana pengurangan *overfitting* menjadi lebih mudah untuk diatasi. Didalam kasus klasifikasi, algoritma klasifikasi akan membentuk *classifier* $H{:}D \rightarrow \{-1, +1\}$ sebagai dasar dari data training. Metode ini membuat urutan classifier Ht , dimana $t = 1, \cdots, T$ sebagai modifikasi dari data training. Classifier tersebut kemudian digabungkan sehingga menjadi satu classifier. Singkatnya, teknik *ensemble bagging* membantu meningkatkan akurasi klasifikasi. Tetapi untuk hasil yang lebih baik dapat dilakukan pada pengkombinasian dengan metode lain untuk lebih meningkatkan akurasi. Machova dkk (2006) menunjukkan bahwa metode bagging adalah cara yang cocok untuk meningkatkan efisiensi *algoritma machine learning*. Jumlah minimum pengklasifikasian yang diperlukan untuk mencapai efisiensi ini dapat ditemukan.

Boosting

*Boostin*g adalah salah satu metode ensemble untuk meningkatkan performance pada suatu algoritma dengan mengkombinasikan classifier yang lemah menjadi classifier yang kuat. Ide utama didalam proses boosting adalah memilih sekumpulan data training dengan beberapa cara untuk kemudian dipelajari oleh suatu base learner, dimana *base learner* dipaksa menarik sesuatu yang baru tentang sampel tersebut setiap kali base learner itu dipanggil. Prinsip kerja boosting yaitu mempekerjakan sekumpulan *classifier* yang dilatih secara iteratif. Salah satu algoritma boosting yang paling populer adalah *Adaptive Boosting* (AdaBoost) yang diperkenalkan oleh Freund & Schapire (1995).

K-Fold Cross Validation

Teknik *K-Fold Cross Validation* adalah salah satu metode yang digunakan untuk mempartisipasi data menjadi data *training* dan data *testing*. Pada penelitian ini digunakan metode tersebut karena dapat mengurangi bias yang terjadi dalam pengambilan sampel. Metode ini digunakan secara berulang-ulang membagi data menjadi dua yaitu data training dan testing, setiap data memperoleh kesempatan menjadi data *testing* (Ibrahim, N. dan Wibowo, A. 2014). K disini merupakan besar angka partisi data yang akan digunakan untuk pembagian antara *training* dan *testing*.

Evaluasi Performance Metode

Kriteria evaluasi merupakan hal yang paling penting dalam penilaian performance klasifikasi. Cara yang paling mudah untuk mengetahui performance klasifikasi dengan melakukan tabulasi silang antara kelas aktual dan prediksi. Hasil dari tabulasi silang, disebut *confusion matrix*. Pada masalah dua kelas, *confusion matrix* digunakan sebagai informasi bahwa pada kolom menjelaskan jumlah dari kelas prediksi dan pada baris menjelaskan jumlah dari kelas aktual (Han dkk, 2012).

Tabel 1 Confussion Matrix

Confusion Matrix		**Predicted Class**	
		Positive	***Negative***
Actual Class	***Positive***	*TP (True Positive)*	*FN (False Negative)*
	Negative	*FP (False Positive)*	*TN (True Negative)*

Sumber: Chawla dkk, 2003

1. *True Positive* (TP) menunjukkan bahwa kelas yang dihasilkan dari prediksi klasifikasi adalah positif dan kelas aktualnya adalah positif.
2. *True Negative* (TN) menunjukkan bahwa kelas yang dihasilkan dari prediksi klasifikasi adalah negatif dan kelas aktualnya adalah negatif.

3. *False Positive* (FP) menunjukkan bahwa kelas yang dihasilkan dari prediksi klasifikasi adalah negatif dan kelas aktualnya adalah positif.
4. *False Negative* (FN) menunjukkan bahwa kelas yang dihasilkan dari prediksi klasifikasi adalah positif dan kelas aktualnya adalah negatif.

Dalam imbalanced, hasil klasifikasi akan mencapai akurasi tinggi karena hanya berfokus pada kelas mayoritas. Jelas bahwa akurasi tidak cukup kuat jika dijadikan sebagai tolak ukur untuk ukuran kriteria performance. Sehingga untuk mengukur performa algoritma pada kelas minoritas digunakan kriteria sensitivity

$$Recall\ /sensitivity\ = \frac{TP}{TP+FN} x\ 100\%$$

$$Specificity = \frac{TN}{TN+FP}\ x\ 100\%$$

Untuk melakukan evaluasi secara keseluruhan, dapat digunakan kriteria seperti *Geometric Mean (G-mean)* dan analisis AUC.

$$G - mean = \sqrt{Sensitivity * Specificity}$$

$$FPR = 1 - Specificity$$

$$AUC = \frac{1+sensitivity-FPR}{2}$$

Geometric Mean (G-mean) adalah rata-rata *geometrik sensitivity* dan *specificity*. G-mean akan bernilai tinggi apabila TP dan TN tinggi dan perbandingan antara TP dan TN kecil. Apabila semua kelas positif tidak dapat diprediksi maka G-mean akan bernilai nol sehingga diharapkan suatu algoritma klasifikasi mencapai nilai G-mean yang tinggi.

METODE PENELITIAN

Sumber Data

Data yang digunakan dalam penelitian ini merupakan data sekunder yang bersumber dari *database* LP3M Rektorat Universitas Muhammadiyah Semarang. Data yang digunakan merupakan data seluruh alumni UNIMUS dari tahun 2010-2020. Jumlah data yang digunakan pada penelitian ini sebanyak 5315 data individu.

Struktur Data

Tabel 2. Struktur Data Penelitian

n	X_1	X_2	…..	X_6	Y
1	$X_{1,1}$	$X_{2,1}$	…….	$X_{6,1}$	0
2	$X_{1,2}$	$X_{2,2}$	…….	$X_{6,2}$	0
⋮	⋮	⋮	…..	⋮	⋮
366	$X_{1,366}$	$X_{2,366}$	…..	$X_{6,366}$	0
367	$X_{1,367}$	$X_{2,367}$	…..	$X_{6,367}$	1
⋮	⋮	⋮	…..	⋮	⋮
5315	$X_{1,5315}$	$X_{2,5315}$	…..	$X_{6,5315}$	1

Langkah Penelitian

Dalam penelitian ini menggunakan *software R* dan langkah-langkah dalam penelitian ini adalah sebagai berikut:

1. Menentukan jumlah ukuran resample dengan melakukan proses *grid search* (5,10,20,30,40,50) dan kernelnya (*Linier, Polynomial, Radial Basis Function, Sigmoid*). Dimana dalam masing-masing kernel memiliki parameter yaitu nilai *cost* (0.01, 0.1, 1, 10) , *gamma* (0.01, 0.1, 1, 10) serta *degree* (2,4,6,8).
2. Membagi data ke dalam data *training* dan *testing* dengan menggunakan *10- fold cross validation* dimana komposisi dari masing-masing fold berisi 10% dari jumlah data mayoritas (negatif) dan 10% dari jumlah data minoritas (positif). Kemudian dibentuk 10-fold untuk masing-masing kelas sehingga ke 10-fold untuk kelas mayoritas. Proses pemilihan anggota fold dilakukan dengan acak dan pengamatan-pengamatan disetiap fold tidak tumpang tindih.
3. Menerapkan metode *smote bagging* dengan SVM sebagai *base classifier* pada data kelulusan Mahasiswa UNIMUS.
   a. Membangkitkan data *synthetic* untuk menyeimbangkan komposisi kelas mayoritas dan kelas minoritas pada setiap data training menggunakan

algoritma SMOTE dengan 5 tetangga terdekat dalam proses pembangkitan data synthetic berdasarkan rekomendasi Chawla dkk (2002).

b. Membangun model *smote bagging SVM* menggunakan data training pada tahap (a) yang telah diseimbangkan oleh algoritma SMOTE.
c. Mengklasifikasikan pengamatan-pengamatan pada data testing menggunakan fungsi klasifikasi smote bagging SVM yang diperoleh pada tahapan (b).
d. Membentuk *confussion matrix* dan menghitung performance klasifikasi dengan G-mean dari metode *smote bagging SVM*.
e. Kembali ketahap (a) sampai ketahap (d) untuk validasi kedua, ketiga, keempat sampai kesepuluh. Kemudian menghitung nilai rata-rata performance metode klasifikasi.

4. Menerapkan metode *smote boosting* dengan SVM sebagai base classifier pada data kelulusan Mahasiswa UNIMUS.
   a. Membangkitkan data *synthetic* untuk menyeimbangkan komposisi kelas mayoritas dan kelas minoritas pada setiap data training menggunakan algoritma SMOTE dengan 5 tetangga terdekat dalam proses pembangkitan data synthetic berdasarkan rekomendasi Chawla dkk (2002).
   b. Membangun model *smote boosting SVM* menggunakan data training pada tahap (a) yang telah diseimbangkan oleh algoritma SMOTE.
   c. Mengklasifikasikan pengamatan-pengamatan pada data testing menggunakan fungsi klasifikasi *smote boosting SVM* yang diperoleh pada tahapan (b).
   d. Membentuk *confussion matrix* dan menghitung performance klasifikasi dengan G-mean dari metode *smote boosting SVM*.
   e. Kembali ketahap (a) sampai ketahap (d) untuk validasi kedua, ketiga, keempat sampai kesepuluh. Kemudian menghitung nilai rata-rata performance metode klasifikasi.
5. Mendapatkan performance *smote bagging-boosting* berdasarkan ukuran resample.
6. Mendapatkan metode dan fungsi kernel berdasarkan nilai g-mean

HASIL PENELITIAN dan PEMBAHASAN

Membagi Data Menjadi Training dan Testing

Pada klasifikasi, terdapat dua proses yang dilakukan sebagai berikut:

1. Proses *training*, dilakukan dengan menganalisis serangkaian *training data* yang label kelasnya diketahui untuk memebangun model klasifikasi.
2. Proses *testing*, dilakukan untuk validasi dan menguji model yang telah dibangun pada langkah awal. Langkah ini juga digunakan untuk mengetahui *performance* dari model klasifikasi.

Berdasarkan penjelasan sebelumnya, maka perlu dilakukan pembagian data yaitu *training* dan *testing*. Teknik yang digunakan untuk memebagi data yaitu *cross validation*, dimana pada penelitian ini menggunakan 10 *fold cross validation*. Komposisi dari masing-masing *fold* berisi 10% dari jumlah data mayoritas (negatif) dan 10% dari jumlah data minoritas (positif). Proses validasi pada ketepatan kelulusan mahasiswa unimus dengan jumlah kelas mayoritas sebesar 4640 dan jumlah kelas minoritas sebesar 675. Kemudian dibentuk *10-fold* untuk masing-masing kelas sehingga ke *10-fold* untuk kelas mayoritas dan minoritas. Proses pemilihan anggota *fold* dilakukan dengan acak dan pengamatan-pengamatan disetiap *fold* tidak tumpang tindih.

Membangun Model Smote Bagging

Pemodelan klasifikasi menggunakan *smote bagging* dilakukan berdasarkan data *training*. Beberapa kemungkinan model klasifikasi *smote bagging* yang akan dibangun berdasarkan beberapa ukuran *resample* yang digunakan yaitu 5, 10, 20, 30, 40, 50. Sedangkan fungsi *kernel* yang digunakan yaitu *kernel linear, polynomial, radial* dan *sigmoid*. Sehingga didapatkan hasil klasifikasi pada *confussion matrix* berdasarkan data training,

yang selanjutnya dapat dihitung nilai G-mean dan AUC.

Tabel 3. Membangun model smote bagging dengan data training

fungsi kernel	ukuran resample	G-mean	AUC
Linear	5	0.7645	0.834
	10	0.7619	0.83
	20	0.7619	0.832
	30	0.766	0.834
	40	0.7657	0.835
	50	0.7653	0.833
Polynomial	5	0.8717	0.941
	10	0.8724	0.944
	20	0.8758	0.946
	30	0.877	0.946
	40	0.8764	0.946
	50	0.8772	0.947
Radial Basis Function	5	0.8635	0.926
	10	0.8709	0.932
	20	0.8714	0.931
	30	0.8712	0.933
	40	0.8703	0.933
	50	0.8724	0.933
Sigmoid	5	0.6541	0.678
	10	0.651	0.679
	20	0.6533	0.682
	30	0.6561	0.682
	40	0.6525	0.683

Berdasarkan tabel 3 menunjukkan bahwa fungsi kernel *linear, polynomial dan radial* cenderung stabil dibandingkan dengan kernel lainnya sehingga nilai ukuran resample tidak begitu mempengaruhi, tetapi berbeda untuk fungsi kernel *sigmoid* cenderung fluktuatif dibandingkan kernel lainnya sehingga ukuran *resample* yang terpilih mempengaruhi kualitas pengklasifikasian. Hal ini mengindikasikan bahwa semakin tinggi ukuran *resample* maka nilai untuk *g-mean* cenderung semakin konvergen. Nilai g-mean tertinggi yang diperoleh yaitu 0.87723 dan *g-mean* terendah diperoleh yaitu 0.65104 sedangkan untuk nilai AUC tertinggi diperoleh 0,94658 dan AUC terendah diperoleh yaitu 0, 0.67824.

Membangun Model Smote Boosting

Pemodelan klasifikasi menggunakan *smote boosting* dilakukan berdasarkan data *training*. Beberapa kemungkinan model klasifikasi *smote boosting* yang akan dibangun berdasarkan beberapa ukuran *resample* yang digunakan yaitu 5, 10, 20, 30, 40, 50. Sedangkan fungsi *kernel* yang digunakan yaitu *kernel linear, polynomial, radial* dan *sigmoid*. Sehingga didapatkan hasil klasifikasi pada *confussion matrix* berdasarkan data training, yang selanjutnya dapat dihitung nilai G-mean dan AUC.

Tabel 4. Membangun model smote boosting dengan data testing

fungsi kernel	ukuran resample	G-mean	AUC
Linear	5	0.7534	0.823
	10	0.7508	0.819
	20	0.7508	0.821
	30	0.7549	0.823
	40	0.7546	0.824
	50	0.7542	0.822
Polynomial	5	0.8606	0.93
	10	0.8612	0.933
	20	0.8647	0.935
	30	0.8679	0.947
	40	0.8653	0.935
	50	0.8679	0.935
Radial Basis Function	5	0.8524	0.915
	10	0.8598	0.921
	20	0.8603	0.92
	30	0.8601	0.922
	40	0.8592	0.921
	50	0.8612	0.922
Sigmoid	5	0.643	0.667

	10	0.6399	0.668
	20	0.6422	0.671
	30	0.645	0.671
	40	0.6414	0.671
	50	0.6406	0.67

Berdasarkan tabel 4 menunjukkan bahwa fungsi kernel *linear, polynomial dan radial* cenderung stabil dibandingkan dengan kernel lainnya sehingga nilai ukuran resample tidak begitu mempengaruhi, tetapi berbeda untuk fungsi kernel *sigmoid* cenderung fluktuatif dibandingkan kernel lainnya sehingga ukuran *resample* yang terpilih mempengaruhi kualitas pengklasifikasian. Hal ini mengindikasikan bahwa semakin tinggi ukuran *resample* maka nilai untuk *g-mean* cenderung semakin konvergen. Nilai g-mean tertinggi yang diperoleh yaitu 0.867917 dan *g-mean* terendah diperoleh yaitu 0.639929 sedangkan untuk nilai AUC tertinggi diperoleh 0.946626 dan AUC terendah diperoleh yaitu 0.667131.

Menguji Model Smote Bagging

Untuk menguji klasifikasi dari *smote bagging* maka dilakukan analisis berdasarkan data testing. Model klasifikasi terbaik dibentuk dengan ukuran resample dan fungsi kernel yang telah ditentukan sebelumnya untuk memperoleh model yang paling optimum. Ukuran evaluasi performance klasifikasi yang digunakan yaitu g-mean. G-mean mampu memberikan informasi tentang seberapa baik model yang dihasilkan dalam memprediksi kelas mayoritas (negatif) dan kelas minoritas (positif). Sehingga performance yang dihasilkan cocok untuk kasus imbalanced data.

Tabel 5.Ukuran performance klasifikasi model *smote bagging* Berdasarkan Ukuran Resample dan Fungsi Kernel

fungsi kernel	ukuran resample	G-mean	AUC
Linear	5	0.76439	0.834151
	10	0.761805	0.830135
	20	0.761792	0.831668
	30	0.765938	0.833678
	40	0.7656	0.834772
	50	0.765175	0.832638
Polynomial	5	0.871722	0.940921
	10	0.872286	0.944123
	20	0.875765	0.945686
	30	0.877007	0.945725
	40	0.876376	0.946112
	50	0.877217	0.946573
Radial Basis Function	5	0.863425	0.926087
	10	0.870843	0.932477
	20	0.8713	0.931361
	30	0.871142	0.933054
	40	0.870226	0.932494
	50	0.872245	0.933473
sigmoid	5	0.653991	0.67823
	10	0.650929	0.678622
	20	0.653174	0.682475
	30	0.655999	0.681995
	40	0.652376	0.68256

Berdasarkan tabel 5 menunjukkan bahwa fungsi kernel *linear, polynomial dan radial* cenderung stabil dibandingkan dengan kernel lainnya sehingga nilai ukuran resample tidak begitu mempengaruhi, tetapi berbeda untuk fungsi kernel *sigmoid* cenderung fluktuatif dibandingkan kernel lainnya sehingga ukuran *resample* yang terpilih mempengaruhi kualitas pengklasifikasian. Hal ini mengindikasikan bahwa semakin tinggi ukuran *resample* maka nilai untuk *g-mean* cenderung semakin konvergen. Nilai g-mean tertinggi yang diperoleh yaitu 0.8772171 dan *g-mean* terendah diperoleh yaitu 0.6509293 sedangkan untuk nilai AUC tertinggi diperoleh 0.946573 dan AUC terendah diperoleh yaitu 0, 0.67823.

Menguji Model Smote Boosting

Pemodelan klasifikasi menggunakan *smote boosting* dilakukan berdasarkan data *testing*. Beberapa kemungkinan model klasifikasi *smote boosting* yang akan dibangun berdasarkan beberapa ukuran *resample* yang digunakan yaitu 5, 10, 20, 30, 40, 50. Sedangkan fungsi *kernel* yang digunakan yaitu *kernel linear, polynomial, radial* dan *sigmoid*. Sehingga didapatkan hasil klasifikasi pada

confussion matrix berdasarkan data training, yang selanjutnya dapat dihitung nilai G-mean dan AUC.

Tabel 6.Ukuran performance klasifikasi model *smote boosting* Berdasarkan Ukuran Resample dan Fungsi Kernel

fungsi kernel	ukuran resample	G-mean	AUC
Linear	5	0.75327	0.822941
	10	0.750685	0.818925
	20	0.750672	0.820458
	30	0.754818	0.822468
	40	0.75448	0.823562
	50	0.754055	0.821428
Polynomial	5	0.860512	0.929712
	10	0.861076	0.932914
	20	0.864555	0.934477
	30	0.867797	0.946516
	40	0.865166	0.934903
	50	0.866007	0.935364
Radial Basis Function	5	0.852305	0.914877
	10	0.859723	0.921267
	20	0.86018	0.920151
	30	0.860022	0.921844
	40	0.859106	0.921284
	50	0.861125	0.922263
sigmoid	5	0.642881	0.667021
	10	0.639819	0.667413
	20	0.642064	0.671266
	30	0.644889	0.670786
	40	0.641266	0.671351
	50	0.640508	0.670087

Berdasarkan Gambar 4.11 menunjukkan bahwa fungsi kernel *linear, polynomial dan radial* cenderung stabil dibandingkan dengan kernel lainnya sehingga nilai ukuran resample tidak begitu mempengaruhi, tetapi berbeda untuk fungsi kernel *sigmoid* cenderung fluktuatif dibandingkan kernel lainnya sehingga ukuran *resample* yang terpilih mempengaruhi kualitas pengklasifikasian. Hal ini mengindikasikan bahwa semakin tinggi ukuran *resample* maka nilai untuk *g-mean* cenderung semakin konvergen. Nilai g-mean tertinggi yang diperoleh yaitu 0.87723 dan *g-mean* terendah diperoleh yaitu 0.94658 sedangkan untuk nilai AUC tertinggi diperoleh 0,95658 dan AUC terendah diperoleh yaitu 0, 0.67824.

Perbandingan smote bagging dan smote boosting SVM

Model untuk metode *smote bagging-boosting* berdasarkan fungsi kernel terpilih dan ukuran *resample* yang optimum dengan rata-rata *g-mean* dapat dilihat sebagai berikut:

Tabel 7 Mendapatkan Model Smote Bagging-Boosting dan Fungsi Kernel berdasarkan Ukuran Resample

Model	Fungsi Kernel	Ukuran Resample	G-mean
Smote Bagging	Polynomial	50	0.87721
Smote Boosting	polynomial	30	0.86779

Berdasarkan Tabel 7 menunjukkan perbandingan Ukuran resample dari klasifikasi *smote bagging-boosting*. Ukuran performance yang digunakan adalah g-mean, dimana *smote bagging* menunjukkan hasil performance sebesar 0.87721 dengan ukuran resample 50 dan fungsi kernel polynomial dengan parameter nilai *cost* 0.1 , *gamma* 0,1 dan *degree* 8. Sedangkan untuk smote boosting sebesar 0.86779 dengan ukuran resample 30 dan fungsi kernel polynomial dengan parameter nilai *cost* 0.1 , *gamma* 0,1 dan *degree* 8.

SIMPULAN dan SARAN

Simpulan

Berdasarkan penelitian yang telah dilakukan maka dapat disimpukan Hasil analisis klasifikasi dengan menggunakan metode *smote bagging SVM* dan *smote boosting SVM* pada data kelulusan mahasiswa UNIMUS 2010-2020 diperoleh bahwa Ukuran performance yang digunakan adalah *G-mean*, dimana *smote bagging* menunjukkan hasil

performance sebesar 0.87721 dengan ukuran resample 50 dan fungsi kernel *polynomial* yang artinya model sudah baik dalam menjelaskan ketepatan lama studi kelulusan mahasiswa unimus tahun 2010-2020 dibandingkan *smote boosting* nilai *G-mean* sebesar 0.86779 dengan ukuran resample 30 dan fungsi kernel *polynomial*. Berdasarkan hasil klasifikasi dengan menggunakan metode *smote bagging SVM* dan *smote boosting SVM*, metode terbaik dalam melakukan klasifikasi data kelulusan mahasiswa UNIMUS 2010-2020 adalah metode *smote bagging SVM* dengan hasil performance sebesar 0.87721 dengan ukuran resample 50 dan fungsi kernel *polynomial*

Saran

Saran yang diberikan oleh peneliti untuk penelitian selanjutnya tentang metode *smote bagging SVM* dan *smote boosting* SVM yaitu diharapkan dapat menggunakan metode-metode klasifikasi lain sebagai bentuk perbandingan dalam melihat tingkat akurasi klasifikasi yang paling baik.

Daftar Pustaka

Attaran. M., dan Deb. P., 2018. Machine Learning: The New "Big Thing" for Competitive Advantage. Int. J. Knowledge Engineering and Data Mining, Vol. 5 No. 4, School of Business and Public Administration, California State University.

Batuwita. R., dan Palade. V., 2012. Class Imbalance Learning Methods For Support Vector Machines, Singapore MIT Alliance for Research and Technology Centre, University Oxford.

Bramer. M., 2007. Principles of Data Mining, Springer, Verlag London.

Breiman, L., Friedman, J. H., Olshen, R., dan Stone, C.1983. Classification and Regression Trees. Wadsworth.

Breiman, L. 1996. Bagging Predictors. Machine Learning, pp. 123-140.

Breiman, L. 2001. Random Forests. Machine Learning, Vol. 45, pp. 5-32.

Bramer. M., 2007. Principles of Data Mining, Springer, Verlag London.

Chawla. N. V., Bowyer. K. W., Hall. L. O., dan Kegelmeyer. W. P., 2002. SMOTE: Synthetic Minority Oversampling Technique, Journal of Artificial Intelligence Research 16 (1): 321-357, Notre Dame University.

Chawla, N.V., Lazarevic, A., Hall, L.O dan Bowyer, K.W., 2002. SMOTEBoost: Improving prediction of the minority class in boosting. Proc. Knowl. Discov. Databases, pp. 107–119.

Chawla. N. V., Lazarevic. A., Hall. L. O., dan Bowyer. K. W., 2003. SMOTEBoost: Improving Prediction of the Minority Class in Boosting, Springer, Verlag Berlin Heidelberg.Abdul, Halim. (2005). Analisis Investasi. Edisis Kedua. Jakarta: Salemba Empat. Alfabeta.

Freund, Y., dan Schapire, R. E. 1995. A desicion-theoretic generalization of online learning and an application to boosting. In European conference on computational learning theory (pp. 23-37). Springer, Berlin, Heidelberg

Freund, Y., dan Schapire, R. E. 1996. Experiments with a New Boosting Algorithm. Machine Learning: Proceedings of the Thirteenth International Conference. Morgan Kaufmann, Italy.

Furey TS, C. N. 2000. Support Vector Machine Classification dan Validation Of Cancer Tissue Samples Using Microarray Expression Data . Bioinformatics.

Gunn, S. 1998. Support vector Machines for Classification and Regression. Technical Report, ISIS.

Han J., M. K. 2000. Data Mining: Concepts And Techniques. New York: MorganKaufman.

Han, J., Kamber, dan M., Pei, J. 2006. Data Mining Concepts and Techniques 2nd Edition. Kaufman Publisher, USA.

Han, J., Kamber, dan M., Pei, J. 2012. Data Mining Concepts and Techniques 3rd Edition. Kaufman Publisher, USA.

Japkowicz, N., dan Stephan, S. 2002. The Class Imbalance Problem: A Systematic Study. Intelligent Data Analysis, pp. 203-231.

Johra. M. B., 2018. Perbandingan Kernel Trick pada Non-linier Support Vector Machine (Studi Kasus: Pemilihan Penolong Persalinan di Provinsi Maluku Utara 2016), Universitas Padjajaran.

Kim, M. J., Kang, D. K., dan Kim, H. B. 2015. Geometric mean based boosting algorithm with over-sampling to resolve data imbalance problem for bankruptcy prediction. Expert Systems with Applications, 42(3), pp. 1074-1082.

Kotsianis, S. B., Kanellopoulos, D., dan Pintelas, P. 2006. Handling Imbalance Dataset: A Review. GESTS International Transaction on Computer Science ang Engineering, 25-36.

Li, K. & Hu, Y., 2019. Research on unbalanced training samples based on SMOTE algorithm. Journal of Physics, X(1303), pp. 1742-6596.

Li, X., Wang, L., dan Sung, E. 2008. AdaBoost with SVM-based component classifiers. Engineering Applications of Artificial Intelligence, 21(5), 785- 795.

Machova. K., Barcak. F., dan Bednar. P., 2006. A Bagging Method using Decision Trees in the Role of Base Classifiers, Department of Cybernetics and Arti ficial Intelligence, Technical University.

Mohammed. M., Khan. M. B., dan Bashier. E. B. M., 2017. Machine Learning Algorithms and Applications, Taylor & Francis Group, LLC, Boca Raton.

Pangastuti. S. S., 2018. Perbandingan Metode Ensemble Random Forest dengan SMOTE-Boosting dan SMOTE-Bagging pada Klasifikasi Data Mining untuk Kelas Imbalance (Studi Kasus: Data Beasiswa Bidikmisi Tahun 2017 di Jawa Timur), Institut Teknologi Sepuluh Nopember.

Patankar. B., dan Chavda. Dr V., 2015. Improving Classification Accuracy through Ensemble Technique in Data Mining, Vol. 1 Issue 6, ISSN: 2395-1990, Hemchandracharya North Gujarat University.

Pratama. R. F. W., 2018. Boosting Support Vector Machine pada Data Microarray yang Imbalance, Institut Teknologi Sepuluh Nopember.

Primartha, R. 2018. *Belajar Machine Learning Teori Dan Praktik.* Bandung: Informatika Bandung.

Profil Anak Indonesia. 2019. Pendidikan Anak, Kementerian Pemberdayaan Perempuan dan Perlindungan Anak (KPPPA) Jakarta.

R, Shyara Taruna dan Saroj Hiranwal. 2013. Enhanced Naive Bayes Algorithm for Intrusion Detection in Data Mining. International Journal of Computer Science and information Technologies, Vol. 4, 2013.

Sain, H., dan Purnami, S.W. 2015. Combine Sampling Support Vector Machine for Imbalance Data Classification. Procedia Computer Science, 72, 771-780.

Santosa, B. 2007. Data Mining: Teknik Pemanfaatan Data untuk Keperluan Bisnis. Graha Ilmu. Yogyakarta.

Shigei. N., and Miyajima. H., 2009. Bagging and Boosting Algorithms for Support Vector Machine Classifiers, Proceedings of The 8th WSEAS International Conference on Artificial Intelligence Knowledge Engineering and Database, Cambridge United Kingdom.

Suyanto. 2018. Machine Learning: Tingkat Dasar dan Lanjut. Bandung: Informatika. Vo, A. T., Tran, H. S., & Le, T. H. 2017. Advertisement Image Classification Using Convolutional Neural Network. 9th International Conference on Knowledge and Systems Engineering(KSE), 197–202.

Syauqi Amri, Y. 2018. Klasifikasi Ketepatan Lama Studi Mahasiswa Menggunakan Metode Support Vectore Machine dan Random Forest. Skripsi. Program Studi Statistika Fakultas Matematika dan Ilmu Pengetahuan Alam Universitas Islam Indonesia.

Turban, Efraim et al and Bonczek et al.. Decision Support System and Intelligent System. Penerbit Andi : Yogyakarta. 2005.

Turban, E., Aronson, J.E. dan Liang, T.P. 2005. Decision Support Systems and Intelligent Systems 7th Ed. New Jersey : Pearson Education.

Utami, Tiani W., Fauzi, Fatkhurokhman. 2017. smooth support vector machine (ssvm) untuk pengklasifikasian indeks pembangunan manusia kabupaten/kota se-indonesia. Skripsi. Program Studi Statistika Fakultas Matematika dan Ilmu Pengetahuan Alam Universitas Muhammadiyah Semarang.

Wang. S., dan Yao. X., 2009. Diversity Analysis on Imbalanced Data Sets by Ensemble Models, IEEE, Birmingham University.

Vapnik, V., dan Cortes, C. 1995. Support Vector Networks. Machine Learning, Volume 20, 273-297.

Vapnik, V N. 1999. An Overview of Statistical Learning Theory. IEEE Trans. on Neural Network, 10, 988-999.

Yongqing. Z., Min. Z., Danling. Z., Gang. M., and Daichuan. M., 2014. Improved SMOTEBagging and its Application in Imbalanced Data Classification, Sichuan University.