PENERAPAN TEKNIK BAGGING PADA ALGORITMA NAIVE BAYES DAN ALGORITMA C4.5 UNTUK MENGATASI KETIDAKSEIMBANGAN KELAS

Achmad Ridwan[1], Annisa Tsani Khoiriyah[2]
[1,2] Sistem Informasi, Universitas Muhammadiyah Kudus
Jln. Ganesha I Purwosari Kudus 59316, Jawa Tengah, Indonesia
[1]achmadridwan@umkudus.ac.id, [2]12019100005@std.umkudus.ac.id

Abstrak

Beberapa dataset yang memiliki dua kelas atau binominal mengalami ketidakseimbangan kelas yang menyebabkan kurangnya akurasi pada klasifikasi. Masalah ketidakseimbangan kelas sangat menghambat kinerja klasifikasi . Oleh karena itu, sejumlah metode seperti bagging dan boosting, telah diusulkan untuk memecahkan masalah ini. Masalah ini menarik banyak perhatian dari para peneliti dari berbagai bidang. Pada penelitian ini diusulkan kombinasi teknik bagging dan algoritma klasifikasi untuk meningkatkan akurasi dari klasifikasi dataset medis. Teknik bagging digunakan untuk menyelesaikan masalah ketidakseimbangan kelas. Metode yang diusulkan diterapkan pada dua algoritma classifier yaitu, algoritma naïve bayes dan algoritma C4.5. Dalam riset ini, data yang digunakan adalah Blogger dataset yang diambil dari UCI repository of machine learning. Pada dataset ini atributnya terdiri : Pendidikan (degree), tingkah politik(caprice), topik, media local turnover (LMT) dan ruang lokal, politik dan sosial (LPSS). Dari hasil penelitian, dengan menerapkan teknik bagging untuk klasifikasi berbasis ensemble pada algoritma C4.5 dapat meningkatkan akurasi sebesar 9 %. Dengan akurasi awal 68 %, setelah diterapkan teknik bagging menjadi 77 %. pada algoritma naïve bayes dapat meningkatkan akurasi sebesar 3,00 %. Dengan akurasi awal 77,00%, setelah diterapkan teknik bagging menjadi 80,00%.

Kata kunci: Data mining, Decision Tree,c4.5,Naïve Bayes , Ensemble, Bagging, blogger profesional.

Abstract

Some datasets that have two classes or binominal experience a class imbalance that causes a lack of accuracy in classification. The problem of class imbalance greatly impedes classification performance. Therefore, a number of methods, such as bagging and boosting, have been proposed to solve this problem. This issue attracts a lot of attention from researchers from various fields. In this study a combination of bagging techniques and classification algorithms is proposed to improve the accuracy of the classification of medical datasets. Bagging techniques are used to solve class imbalance problems. The proposed method is applied to two classifier algorithms, namely the naïve bayes algorithm and the C4.5 algorithm. In this research, the data used is the Blogger dataset taken from the UCI repository of machine learning. In this dataset the attributes consist of: education (degree), political behavior (caprice), topics, local media turnover (LMT) and local, political and social space (LPSS). From the results of the study, by applying bagging techniques for ensemble-based classification on the C4.5 algorithm can increase accuracy by 9%. With an initial accuracy of 68%, after applying the bagging technique to 77%. the naïve bayes algorithm can increase accuracy by 3.00%. With an initial accuracy of 77.00%, after applying the bagging technique to 80.00%.

Keywords: *Data mining, Decision Tree, c4.5, Naïve Bayes, Ensemble, Bagging, professional bloggers.*

I. PENDAHULUAN

Blog memang banyak diminati baik dari kalangan remaja hingga dewasa maupun orang tua. Blogger memiliki latar belakang yang sama yaitu suka menulis dengan memanfaatkan internet. Satu hal yang membuat blog akan disukai pengunjung yaitu blog yang konsisten. Sebagai blogger yang bertanggung jawab, menulis atau mengisi blog haruslah konsisten. Konsisten dalam pengertian waktu dan tema. Konsisten waktu

artinya blogger harus menulis rutin misalnya setiap hari, seminggu sekali atau dua kali, sebulan sekali, dan lainnya. Sebenarnya bebas tetapi jika rutin atau terjadwal maka akan jadi bloger professional

Blog bisa disebut buku harian di internet yang lahir sebagai situs web dinamis dan melihat asalnya pada tahun 1997 berkat **Dave Winer**. Dave Winer adalah pengembang perangkat lunak pertama yang menerbitkan artikel baru secara mandiri. Blog dari segi fungsinya di bagi menjadi blog pribadi dan blog bisnis

Di era teknologi Komputer yang terus berkembang dengan pesat dapat membantu kita untuk mendeteksi bloger profesional secara akurat dan dapat menghemat waktu untuk tujuan sebuah bisnis . Data mining adalah bidang dalam ilmu komputer digunakan untuk prediksi diberbagai bidang kehidupan dan berbagai tujuan. Semua ini adalah proses menemukan dataset baru dari dataset yang sebelumnya diketahui melalui analisis dataset untuk sebuah tujuan (Larose, 2006). Untuk memprediksi bloger professional dan tidak dengan melalui data mining diperlukan unsur – unsur penunjang dalam penentuannya disertai dengan data yang valid.

Salah satu algoritma *data mining* adalah Naïve Bayes. Metode Naïve Bayes digunakan mengklasifikasikan penentuan keprofesionalan seorang bloger.

Penelitian dilakukan untuk menganalisis membandingkan Algoritma Naïve Bayes, C4.5, bagging+C.45 dan bagging+naïve Bayes untuk pengklasifikasian Dataset blogger sehingga mendapatkan akurasi yamg di dapat dari hasil proses evaluasi.

II. LANDASAN TEORI

A. Data Mining

Data Mining (DM) adalah inti dari proses Knowledge Discovery in Database (KDD), melibatkan algoritma yang mengeksploarasi data, mengembangkan model dan menemukan pola yang tidak diketahui sebelumnya. Model ini digunakan untuk memahami fenomena dari data, analisis dan prediksi. Aksesibilitas dan banyaknya data membuat Knowledge Discovery dan Data Mining menjadi masalah yang cukup penting dan dibutuhkan.

Menurut Larose (2005) (Fatmawati, 2016)berdasarkan tugasnya, data mining dibagi menjadi 6 kelompok, yaitu:

Deskripsi

Terkadang peneliti dan analisis secara sederhana mencoba mencari cara untuk menggambarkan pola dan kecenderungan yang terdapat dalam data..

Estimasi

Estimasi hampir sama dengan klasifikasi, tetapi variable target estimasi lebih ke arah numerik daripada kategori.

Prediksi

Prediksi hampir sama dengan estimasi dan klasifikasi, tetapi dalam prediksi akan menghasilkan nilai.

Klasifikasi

Dalam klasifikasi terdapat target variabel kategori.

Cluster

Mengelompokan record, pengamatan dan membentuk kelas obyek-obyek yang memiliki kemiripan. Tujuan dari algoritma cluster adalah dengan memecahkan setiap datadalam dataset menjadi kelompok kelompok yang homogen. Kelompok data ini biasanya disebut sebagi cluster. Setiap cluster yang terbentuk akan terdiri dari data yang sejenis dan berbeda dengan data pada cluster lainnya. Pengelompokkan ini sama dengan cara kerja otak manusia, dimana ilmu pengetahuan dikelompokkan dalam setiap bidangnya. Dengan adanya pengelompokkan, data yang dapat diolah dengan lebih spesifik sesuai dengan tujuan penelitian. Pemecahan data kedalam cluster data juga diterapkan pada tahappengolahan awal data dalam proses data mining, sehingga dapat diterapkan metode data mining data mining kedalam setiap cluster data. Proses clustering juga dapat mengurangi jumlah ataupun dimensi data yang diolah(Larose, 2006).

B. Algoritma C4.5

Algoritma C 4.5 adalah salah satu metode untuk membuat *decision tree* berdasarkan *training data* yang telah disediakan. Algoritma C 4.5 merupakan pengembangan dari ID3. Beberapa pengembangan yang dilakukan pada C 4.5 adalah bisa mengatasi

missing value, bisa mengatasi c*ontinue data*, dan *pruning*(Fatmawati, 2016).

Pohon keputusan merupakan metode klasifikasi dan prediksi yang sangat kuat dan terkenal. Metode pohon keputusan mengubah fakta yang sangat besar menjadi pohon keputusan yang merepresentasikan aturan. Aturan dapat dengan mudah dipahami dengan bahasa alami. Dan mereka juga dapat diekspresikan dalam bentuk bahasa basis data seperti *Structured Query Language* untuk mencari *record* pada kategori tertentu. Pohon keputusan juga berguna untuk mengeksplorasi data, menemukan hubungan tersembunyi antara sejumlah calon variabel *input* dengan sebuah variabel target.

Karena pohon keputusan memadukan antara eksplorasi data dan pemodelan, pohon keputusan sangat bagus sebagai langkah awal dalam proses pemodelan bahkan ketika dijadikan sebagai model akhir dari beberapa teknik lain. Sebuah pohon keputusan adalah sebuah struktur yang dapat digunakan untuk membagi kumpulan data yang besar menjadi himpunan-himpunan *record* yang lebih kecil dengan menerapkan serangkaian aturan keputusan. Dengan masingmasing rangkaian pembagian, anggota himpunan hasil menjadi mirip satu dengan yang lain (Marlina, lim and Utama Siahaan, 2016).

Sebuah model pohon keputusan terdiri dari sekumpulan aturan untuk membagi sejumlah populasi yang heterogen menjadi lebih kecil, lebih homogen dengan memperhatikan pada variabel tujuannya. Sebuah pohon keputusan mungkin dibangun dengan seksama secara manual atau dapat tumbuh secara otomatis dengan menerapkan salah satu atau beberapa algoritma pohon keputusan untuk memodelkan himpunan data yang belum terklasifikasi.

Variabel tujuan biasanya dikelompokkan dengan pasti dan model pohon keputusan lebih mengarah pada perhitungan probability dari tiap-tiap record terhadap kategori-kategori tersebut atau untuk mengklasifikasi record dengan mengelompokkannya dalam satu kelas. Pohon keputusan juga dapat digunakan untuk mengestimasi nilai dari variabel continue meskipun ada beberapa teknik yang lebih sesuai untuk kasus ini.

Banyak algoritma yang dapat dipakai dalam pembentukan pohon keputusan,antara lain ID3, CART, dan C4.5 (Larose, 2006).

Data dalam pohon keputusan biasanya dinyatakan dalam bentuk tabel dengan atribut dan record. Atribut menyatakan suatu parameter yang dibuat sebagai kriteria dalam pembentukan pohon. Misalkan untuk menentukan main tenis, kriteria yang diperhatikan adalah cuaca, angin, dan temperatur.

Salah satu atribut merupakan atribut yang menyatakan data solusi per item data yang disebut target atribut. Atribut memiliki nilai-nilai yang dinamakan dengan instance. Misalkan atribut cuaca mempunyai instance berupa cerah, berawan, dan hujan(Basuki and Suwarno, 2018) .

Proses pada pohon keputusan adalah mengubah bentuk data (tabel) menjadi model pohon, mengubah model pohon menjadi rule, dan menyederhanakan rule.

C. Algoritma Naïve Bayes

Algoritma Naïve Bayes akan mengevaluasi setiap atribut yang mengkontribusi prediksi pada atribut target. Naïve Bayes tidak memperhitungkan relasi antar atribut-atribut kontributor prediksi, tidak seperti Decision Tree yang memperhitungkan relasi antara atribut. Bentuk tugas dasar yang dilakukan oleh algoritma Naïve Bayes adalah hanyalah klasifikasi (Boukenze *et al.*, 2012) . Naïve Bayes merupakan teknik data mining dengan pendekatan teori probabilitas untuk membangun sebuah model klasifikasi berdasarkan pada kejadian masa lalu yang mempunyai potensi membentuk sebuah objek baru yang dikategorikan sebagai kelas yang memiliki probabilitas terbaik.

Naïve Bayes memiliki kemampuan yang cepat dalam membuat model, mempunyai kemampuan memprediksi dan juga menyediakan metode baru dalam mengeksplor dan memahami data. Algoritma Naïve Bayes hanya mendukung pada atribut yang bertipe data discrete atau discretized, atau tidak mendukung atribut yang bernilai continuous (numerik) dan semua atribut dapat menjadi independen, menjadi atribut yang memberi kontribusi kepada atribut yang diprediksi.

Bayesian classification adalah pengklasifikasian statistik yang dapat digunakan untuk memprediski probabilitas keanggotaan suatu class. Bayesian classification didasarkan pada teorema Bayes yang memiliki kemampuan klasifikasi serupa dengan decision tree dan neural network. Bayesian classification terbukti memiliki akurasi dan kecepatan yang tinggi saat diaplikasikan ke dalam database dengan data yang besar (Wu *et al.*, 2008).

D. Rapid Miner

Rapid Miner adalah aplikasi data mining open-source yang terkemuka dan ternama di dunia. Dirancang sebagai aplikasi yang berdiri sendiri untuk analisis data dan sebagai mesin pengolah data mining untuk diintegrasikan ke dalam produk sendiri. Ribuan aplikasi RapidMiner di lebih dari 40 negara memberikan banyak manfaat bagi penggunanya, antara lain : Integrasi data, Analitis ETL, Data Analisis, dan Pelaporan dalam suatu suite tunggal (Naik and Samant, 2016).

RapidMiner merupakan sebuah lingkungan untuk machine learning, data mining, text mining dan predictive analytics.

E. Bagging

Bagging merupakan emseble pada algoritma supervised. Untuk setiap percobaan t = 1,2, ..., T, satu set pelatihan dengan ukuran N diambil sampelnya (dengan penggantian) dari instance asli. Set pelatihan ini memiliki ukuran yang sama dengan data asli, tetapi beberapa contoh mungkin tidak muncul di dalamnya, sementara yang lain muncul lebih dari sekali. Sistem training menghasilkan pengklasifikasi C t dari sampel dan pengklasifikasi akhir C* dibentuk dengan menggabungkan pengklasifikasi T dari uji coba ini. Untuk mengklasifikasikan sebuah instance x, a. suara untuk kelas k dicatat oleh setiap pengklasifikasi yang C ”(Z) = k dan C * (x) maka kelas yang paling banyak data yang muncu menggunakan CART (Breiman, Friedman, Olshen, dan Stone 1984) sebagai sistem pembelajaran, Breiman (1996) melaporkan hasil pengantongan pada tujuh berukuran sedang kumpulan data. Dengan jumlah ulangan T ditetapkan 50, kesalahan rata-rata dari rentang C * pengklasifikasi yang dikantongi dari 0,57 menjadi 0,94 dari kesalahan yang sesuai ketika a pengklasifikasi tunggal dipelajari. Breiman memperkenalkan konsep sistem pembelajaran pengklasifikasi yang benar-benar sebagai salah satu yang, lebih dari banyak set pelatihan, cenderung memprediksi kelas yang benar dari contoh uji lebih sering daripada kelas lainnya. Peserta yang benar urutan mungkin tidak menghasilkan pengklasifikasi yang optimal, tetapi Breiman menunjukkan bahwa menggabungkan pengklasifikasi yang dihasilkan oleh peserta yang benar urutan menghasilkan pengklasifikasi yang optimal. Breiman mencatat: bahwa Unsur vital adalah ketidakstabilan metode prediksi. Jika mengganggu perangkat pembelajaran dapat menyebabkan perubahan signifikan pada prediktor yang dibuat, kemudian bagging dapat meningkatkan akurasi.

Bagging ditemukan oleh Breiman (1996) yang merupakan kepanjangan dari ” boostrap agregating” (Wu and Kumar, 2009). Bagging adalah salah satu teknik dari ensemble method dengan cara memanipulasi data training, data training di duplikasi sebanyak d kali dengan pengembalian (sampling with replacement), yang akan menghasilkan sebanyak d data training yang baru, kemudian dari data d training tersebut akan dibangun classifier-classifier yang disebut sebagai bagged classifier(Altman and Krzywinski, 2017).

F. Evaluasi Algoritma Klasifikasi Data Mining

D.1 Evaluasi Confusion Matrix

Untuk melakukan evaluasi terhadap model klasifikasi berdasarkan perhitungan objek testing mana yang diprediksi benar dan tidak benar. Confusion Matrix berisi informasi tentang aktual (actual) dan prediksi (predicted) pada sistem klasifikasi. Kinerja sistem seperti ini biasanya dievaluasi dengan menggunakan data pada matriks. Perhitungan ini ditabulasikan kedalam tabel yang disebut Confusion Matrix (Luque *et al.*, 2019). Bentuk Confusion Matrix dapat dilihat pada Tabel berikut ini

Tabel 2.1 Confusion Matrix

CLASSIFICATION	PREDICTED CLASS	
CLASS:YES	CLASS:NO	
OBSERVED	CLASS:YES	a
		b

CLASS Negative-FN)	(True	Positive-TP)	(False
CLASS:NO	c	d	
(False Possitive-FP)	(True negative-TN)		

Pada Tabel 2.1 untuk True positive merupakan tupel positif didata set yang diklasifikasikan positif, True negatives merupakan tupel negatif di data set yang diklasifikasikan negatif. False positives adalah tupel positif di data set yang diklasifikasikan negatif False negatives merupakan jumlah tupel negatif yang diklasfikasikan positif.

Setelah dilakukan confusion matrix berikutnya akan dihitung accuracy, sensitivity, specificity, PPV, NPV. Sensitivity digunakan untuk membandingkan jumlah true positives terhadap jumlah tupel yang positives sedangkan specificity adalah perbandingan jumlah true negatives terhadap jumlah tupel yang negatives. Sedangkan untuk PPV(Positives Predictive Value atau nilai prediktif positif) adalah proporsi kasus dengan hasil diagnosa positif, NPV(Negatives Predictive Value atau nilai prediktif negatif) adalah proporsi kasus dengan hasil diagnosa negatif.

Confusion matrix memberikan penilaian kinerja model klasifikasi berdasarkan jumlah objek yang diprediksi dengan benar dan salah (Gorunescu, 2011). Akurasi kelas minoritas dapat menggunakan metrik TP rate/recall (sensitivitas).

Rumus-rumus yang digunakan untuk melakukan penghitungannya adalah:

Keakuratan (*Accuracy)* adalah proporsi jumlah prediksi yang benar. Hal ini ditentukan dengan menggunakan rumus *accuracy berikut:* :

$$Accuracy = \frac{a+b}{a+b+c+d} = \frac{TP+TN}{TP+TN+FP+FN}$$

D.2 Kurva ROC

Kurva ROC banyak digunakan untuk menilai hasil prediksi, kurva ROC adalah teknik untuk memanjemen pengklasifikasian berdasarkan kinerja mereka(Hoo, Candlish and Teare, 2017). Kurva ROC merupakan alat dua dimensi yang digunakan untuk menilai kinerja klasifikasi yang menggunakan dua class keputusan, tiap objek dipetakan ke salah satu elemen dari himpunan pasangan, positif atau negatif. Pada kurva ROC, TP rate diplot pada sumbu Y dan FP rate diplot pada sumbu X. Untuk klasifikasi data mining menurut Gorunescu, nilai AUC dapat dibagi menjadi beberapa kelompok(Ridwan, Andono and Supriyanto, 2018):

1. Nilai 0,90 – 1.00= Excellent Classification
2. Nilai 0,80 – 0,90= Good Classification
3. Nilai 0,70 – 0,80= Fair Classification
4. Nilai 0,60 – 0,70= Poor Classification
5. Nilai 0,50 – 0,60= Failure

III. METODE PENELITIAN

Dalam penelitian ini metode yang digunakan yaitu metode penelitian kuantitatif. Tujuan dari penelitian ini adalah melakukan klasifikasi dari evaluasi model Algoritma Naïve bayes, C4.5 , Bagging+naïve bayes, Bagging+C4.5 untuk mengetahui akurasi masing-masing algoritma dalam mengklasifikasikan blogger Profesional

A. Sumber Data

Sumber data yang digunakan pada penelitian ini adalah data sekunder dari dataset UCI Machine Learning Repository dengan alamat web https://archive.ics.uci.edu/ml/machine-learning-databases/00255 Dataset yang digunakan adalah Blogger dataset dimana file tersebut bernama kohkiloyeh.xls. Variabel yang digunakan pada penelitian ini adalah sebanyak 6 variabel dengan jumlah data sebanyak 100. Pengumpulan informasi untuk membentuk database dilakukan dengan daftar pertanyaan. Kuesioner ini diberikan secara lisan, tertulis dan juga pemrograman situs web yang menyertakan internet kuesioner dan pengguna dapat menjawab pertanyaan sebagaimana mereka inginkan. Data dikumpulkan dari Provinsi Kohkiloye dan Boyer Ahmad di Iran..

IV. HASIL DAN PEMBAHASAN

Training dan Data Testing akan kita proses klasifikasi menggunakan aplikasi Rapidminer Adapun hasil dari Confunsion Matrix.nya :

Tabel 4.1 Tabel Hasil Class Recall dan Precision naive bayes

	true yes	**true no**	**class precision**
pred. yes	63	18	77.78%
pred. no	5	14	73.68%
class recall	92.65%	43.75%	

Dari Tabel 4.1 didapatkan Class Precision = 73,68%, dan Class Recall: 43,75 %

Tabel 4.2 Tabel Hasil Class Recall dan Precision bagging+naïve bayes

	true yes	true no	class precision
pred. yes	62	14	81.58%
pred. no	6	18	75.00%
class recall	91.18%	56.25%	

Dari Tabel 4.2 didapatkan Class Precision = 75%, dan Class Recall: 56,25 %

Tabel 4.3 Tabel Hasil Class Recall dan Precision C4.5

	true yes	true no	class precision
pred. yes	68	32	68.00%
pred. no	0	0	0.00%
class recall	100.00%	0.00%	

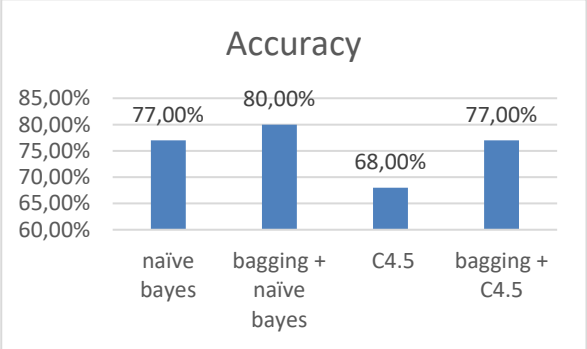
Dari Tabel 4.3 didapatkan Class Precision = 0%, dan Class Recall: 0 %

Tabel 4.4 Tabel Hasil Class Recall dan Precision bagging+C4.5

	true yes	true no	class precision
pred. yes	68	23	74.73%
pred. no	0	9	100.00%
class recall	100.00%	28.12%	

Dari Tabel 4.4 didapatkan Class Precision = 100%, dan Class Recall: 28.12 %

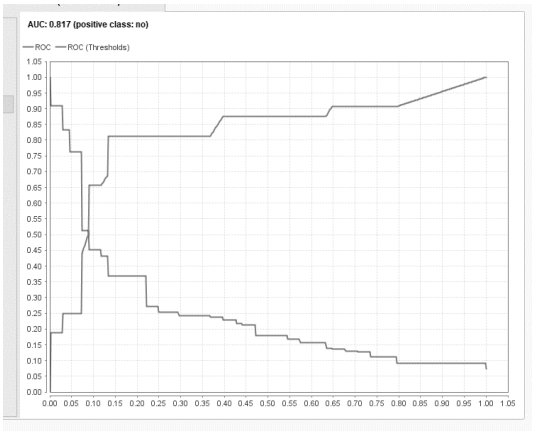
Hasil dari evaluasi pengklasifikasian dengan Algoritma Naïve bayes, C4.5 , Bagging+naïve bayes, Bagging+C4.5 menghasilkan :



Gambar 4.1 Akurasi Naïve Bayes dan bagging+naïve bayes

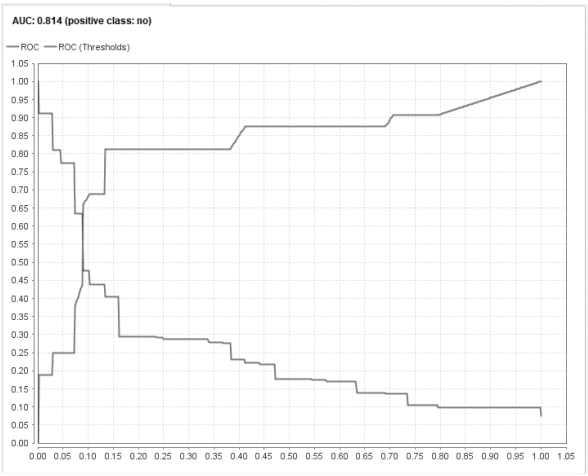
Dari gambar diatas didapatkan akurasi untuk mengklasifisikan Dataset blogger yaitu dengan naïve bayes 77 % sedangkan untuk bagging+naïve bayes 80%. Algoritma Naïve bayes = 77 %, C4.5 = 68% , Bagging+naïve bayes = 80 % , Bagging+C4.5 = 77%

Adapun hasil dari pengujian ini juga menghasilkan ROC sebagai berikut :



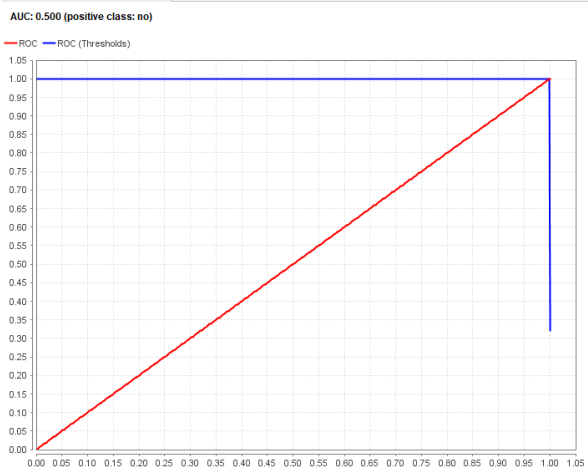
Gambar 4.2 Kurva Roc Model Naive Bayes

Kurva ROC (Reciver Operating Characteristic) diatas menunjukkan algoritma Naïve Bayes memiliki nilai AUC sebesar 0.817



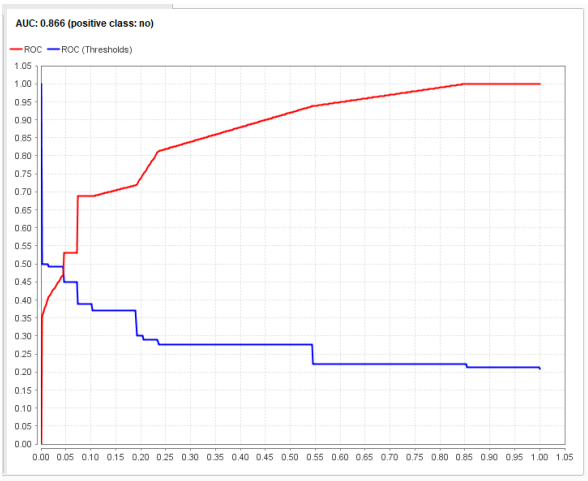
Gambar 4.3 Kurva Roc Model bagging+Naive Bayes

Gambar diatas Kurva ROC (Reciver Operating Characteristic) diatas menunjukkan algoritma bagging +naïve bayes memiliki nilai AUC sebesar 0.826



Gambar 4.3 Kurva Roc Model C4.5

Gambar diatas Kurva ROC (Reciver Operating Characteristic) diatas menunjukkan algoritma C4.5 memiliki nilai AUC sebesar 0.500



Gambar 4.3 Kurva Roc Model Bagging+C4.5

Gambar diatas Kurva ROC (Reciver Operating Characteristic) diatas menunjukkan algoritma Bagging+C4.5 memiliki nilai AUC sebesar 0.866

Dalam penggunaan ensemble bagging pada algoritma naïve bayes dan c4.5 ini sangat berpengaruh dan meningkatkan akurasi dimana akurasi tertinggi pada bagging+naïve bayes yaitu 80 %. Bagging meningkatkan class recall dimana class tertinggi yaitu 56,26 % dan precission tertinggi yaitu 75%

V. KESIMPULAN

Terdapat 5 atribut yaitu terdiri : Pendidikan (degree), tingkah politik(caprice), topik, media local turnover (LMT) dan ruang lokal,politik dan sosial (LPSS) dan penentunya adalah Professional blogger (pb).

Penelitian ini membandingkan Algoritma Naïve Bayes, C4.5, bagging+C.45 dan bagging+naïve Bayes untuk pengklasifikasian Dataset blogger. Data ini terdiri 100 data.

Dari hasil proses Klasifikasi permodelan algoritma naïve bayes terhadap Dataset Blogger Dalam penggunaan ensemble bagging pada algoritma naïve bayes dan c4.5 ini sangat berpengaruh dan meningkatkan akurasi , class Precision, class Recall dan nilai AUC . Ini artinya Teknik bagging dapat meningkatkan class recall dan class precision pada klasifikasi naïve bayes dan c4.5 dan menangani ketidak seimbangan kelas

DAFTAR PUSTAKA

Altman, N. and Krzywinski, M. (2017) 'Points of Significance: Ensemble methods: Bagging and random forests', *Nature Methods*. doi: 10.1038/nmeth.4438.

Basuki, A. and Suwarno (2018) 'Online dissolved gas analysis of power transformers based on decision tree model', in *4th IEEE Conference on Power Engineering and Renewable Energy, ICPERE 2018 - Proceedings*. doi: 10.1109/ICPERE.2018.8739761.

Boukenze, B. *et al.* (2012) 'Top 10 algorithms in data mining', *Knowledge and Information Systems*. doi: 10.1017/S0269888910000032.

Fatmawati, F. (2016) 'PERBANDINGAN ALGORITMA KLASIFIKASI DATA MINING MODEL C4.5 DAN NAIVE BAYES UNTUK PREDIKSI PENYAKIT DIABETES', *None*.

Gorunescu, F. (2011) *Data Mining, Soft Computing*. Berlin, Heidelberg: Springer Berlin Heidelberg. doi: 10.1007/978-3-642-19721-5.

Hoo, Z. H., Candlish, J. and Teare, D. (2017) 'What is an ROC curve?', *Emergency Medicine Journal*. doi: 10.1136/emermed-2017-206735.

Larose, D. T. (2006) *Data Mining Methods and Models, Data Mining Methods and Models*. doi: 10.1002/0471756482.

Luque, A. *et al.* (2019) 'The impact of class imbalance in classification performance

metrics based on the binary confusion matrix', *Pattern Recognition*. doi: 10.1016/j.patcog.2019.02.023.

Marlina, L., lim, M. and Utama Siahaan, A. P. (2016) 'Data Mining Classification Comparison (Naïve Bayes and C4.5 Algorithms)', *International Journal of Engineering Trends and Technology*. doi: 10.14445/22315381/ijett-v38p268.

Naik, A. and Samant, L. (2016) 'Correlation Review of Classification Algorithm Using Data Mining Tool: WEKA, Rapidminer, Tanagra, Orange and Knime', *Procedia Computer Science*, 85, pp. 662–668. doi: 10.1016/j.procs.2016.05.251.

Ridwan, A., Andono, P. N. and Supriyanto, C. (2018) 'Optimasi Klasifikasi Status Gizi Balita Berdasarkan Indeks Antropometri Menggunakan Algoritma Naive', *Teknologi Informasi*.

Wu, X. *et al.* (2008) 'Top 10 algorithms in data mining', *Knowledge and Information Systems*. doi: 10.1007/s10115-007-0114-2.

Wu, X. and Kumar, V. (2009) *The Top Ten Algorithms in Data Mining*. Taylor & Francis Group.