

PENERAPAN TEKNIK BAGGING UNTUK MENINGKATKAN AKURASI KLASIFIKASI PADA ALGORITMA C4.5 DALAM MENENTUKAN BLOGGER PROFESIONAL

Taftazani Ghazi Pratama, Agung prihandono, Achmad Ridwan
 Universitas Muhammadiyah Kudus
 Jln. Ganesha I Purwosari Kudus 59316, Jawa Tengah, Indonesia
 agungdono@umkudus.ac.id

Abstrak

Ada beberapa model data mining salah satunya fungsinya sebagai klasifikasi. Di bidang klasifikasi, ada banyak cabang yang berkembang yaitu pohon keputusan (decision tree). Salah satu decision tree yang populer adalah C4.5. tetapi algoritma ini kurang maksimal dalam menangani kesimbangan kelas. Masalah ketidakseimbangan kelas telah dilaporkan sangat menghambat kinerja algoritma klasifikasi dan telah menarik banyak perhatian dari para peneliti dari berbagai bidang. Oleh karena itu telah diusulkan untuk memecahkan masalah ini maka diperlukan sebuah algoritma ensemble yaitu bagging untuk meningkatkan Akurasi. 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 %.

Kata kunci: *Data mining, Decision tree, C4.5, Ensemble, Bagging, blogger profesional.*

Abstract

There are several data mining models, one of which functions as a classification. In the field of classification, there are many branches that develop, namely the decision tree. One popular decision tree is C4.5. but this algorithm is less optimal in handling class balance. The problem of class imbalance has been reported to severely hamper the performance of classification algorithms and has attracted much attention from researchers from various fields. Therefore it has been proposed to solve this problem that we need an ensemble algorithm that is bagging to improve Accuracy. 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%.

Keywords: *Data mining, Decision tree, C4.5, Ensemble, Bagging, professional blogger.*

I. PENDAHULUAN

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

Blog pribadi adalah Anda bebas untuk mengungkapkan pikiran Anda untuk membiarkan kata-kata mengalir di selembar kertas. Blogging pribadi adalah cara spektakuler untuk menyuarakan kreativitas Dan untuk membuat orang membaca karya Anda. Dengan cara ini Anda dapat

mengekspos bakat, Anda dapat memastikan bahwa ada umpan balik dari pembaca .

Setiap bentuk kreativitas membutuhkan perbandingan. Blogger jika ingin tahu apakah yang ditulis itu berharga dan berhasil di baca para pembaca , itu cara melihat bahwa blogger kita berhasil atau tidak.

Blog perusahaan adalah investasi . Dalam hal ini banyak menyebut tentang aktivitas freelancer, freelance, yang ingin menggunakan blog untuk mempromosikan bisnisnya . Uang yang di dikeluarkan bukan biaya tapi cara untuk menghasilkan uang lebih banyak dalam jangka panjang. Dalam sebuah bloger dapat terlihat blog yang banyak disukai oleh pengunjung biasanya pembuat atau blogernya sudah professional. Blogger yang professional selalu konsisten. Kekonsistennya dapat berupa pengisian konten dan waktu. Pengunjung akan selalu terikat akan konten yang diminatinya dan tergantung pada keprofesionalan bloggernya.

Di era teknologi Komputer yang terus berkembang dengan pesat dapat membantu kita untuk mendeteksi blogger 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 blogger profesional dan tidak dengan melalui data mining diperlukan unsur – unsur penunjang dalam penentuannya disertai dengan data yang valid.

Salah satu algoritma *data mining* adalah C4.5. Metode C4.5 digunakan mengklasifikasikan penentuan keprofesionalan seorang blogger.

Penelitian dilakukan untuk menganalisis melalui algoritma C4.5 untuk mengklasifikasi penentuan keprofesionalan seorang blogger sehingga mendapatkan akurasi yang di dapat dari hasil proses evaluasi.

II. LANDASAN TEORI

A. Data Mining

Data mining adalah sebuah ilmu yang mempelajari metode untuk menemukan sebuah pola dari suatu data yang besar ekstraksi dari data ke pengetahuan dan dapat disimpulkan mempunyai 3 kunci (Larose, 2006) yaitu :

Data karena data tersebut adalah fakta yang terekam dan tidak membawa arti

Informasi merupakan rangkuman penjelasan dan statistik dari data

Pengetahuan dimana pengetahuan ini pola rumus dan aturan atau model yang muncul dari data

Nama lain dari data mining adalah Knowledge Discovery in Data, ada juga tentang Big data atau bussines intelijen, Knowledge Extraction dan pattern analisis atau information harvesting .

Konsep dari data mining yaitu himpunan data yang telah terhimpun banyak sekali kemudian dengan metode data mining yang ada diproses dengan rumus/algoritma data-data tersebut diambil atau diekstraksi untuk menjadi di sebuah pengetahuan. Pengetahuan Itulah bisa untuk diambil untuk sebuah tujuan tertentu. Pengaturan tersebut sangat berguna untuk berbagai keperluan .

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 continue data, dan pruning(Quinlan, 1996).

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 masing-masing rangkaian pembagian, anggota himpunan hasil menjadi mirip satu dengan yang lain (Berry and Linoff, 1997).

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).

C. Rapid Miner

Rapidminer adalah aplikasi perangkat lunak yang menyediakan lingkungan yang terintegrasi untuk persiapan data, pembelajaran mesin, penggalian data, penggalian teks, dan menganalisa prediksi. Rapid Miner sebelumnya dikenal dengan nama YALE (Yet Another Learning Environment) dikembangkan mulai tahun 2001 oleh Ralf Klinkenberg, Ingo Mierswa, dan Simon Fischer dari bagian kecerdasan buatan Universitas Teknik Dortmund. Pengembangan Rapidminer di dorong oleh rapid-I mulai tahun 2006, rapid-I adalah perusahaan berdiri dibawah Ingo Mierswa dan

Ralf Klinkenberg . Tahun 2007 namanya berubah dari YALE menjadi Rapidminer. Aplikasi ini untuk tujuan bisnis, penelitian, pendidikan , pelatihan, visualisasi Prototipe, dan pengembangan platform aplikasi serta mendukung semua proses pembelajaran mesin diantaranya preparing data, visualisasi, validitas modeling, dan pengembangan. Rapidminer dikembangkan dengan model core terbuka. Dengan Rapid Studio Free Edition yang dibatasi 1 processor logika dan 10.000 baris data terintegrasi.(Naik and Samant, 2016).

D. Bagging

Bagging ditemukan oleh Breiman (1996) yang merupakan kepanjangan dari ” bootstrap 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). Karena data pada tehnik bagging dilakukan sampling with replacement, ukuran data bagging sama dengan data aslinya, tetapi distribusi data dari tiap data bagging berbeda, beberapa data dari data training bisa saja muncul beberapa kali atau mungkin tidak muncul sama sekali (Galar *et al.*, 2012). Hal inilah yang menjadi kunci kenapa bagging bisa meningkatkan akurasi karena dengan sampling with replacement dapat memperkecil variance dari dataset(Mordelet and Vert, 2014)

E. 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:

CLASSIFICATION		PREDICTED CLASS
CLASS:YES	CLASS:NO	

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

Pada Tabel 2.1 untuk True positive merupakan tupel positif di data 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 diklasifikasikan 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(Positive Predictive Value atau nilai prediktif positif) adalah proporsi kasus dengan hasil diagnosa positif, NPV(Negative 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 :

$$Accuracy = \frac{a + b}{a + b + c + d} \quad (2)$$

$$= \frac{TP + TN}{TP + TN + FP + FN}$$

D.2 Kurva ROC

Kurva ROC banyak digunakan untuk menilai hasil prediksi, kurva ROC adalah teknik untuk manajemen 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

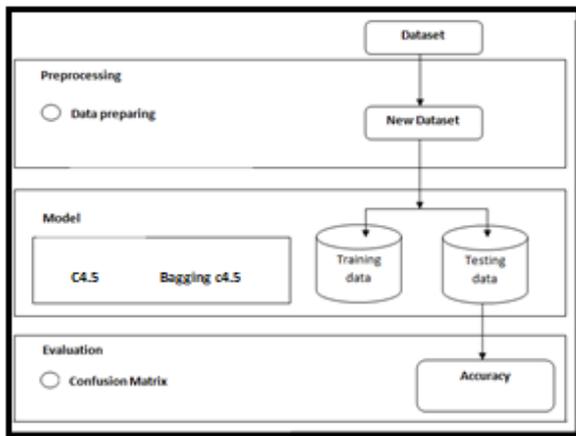
Pendekatan utama dalam penelitian ini yaitu pendekatan kualitatif dan pendekatan kuantitatif. Dalam penelitian ini metode yang digunakan yaitu metode penelitian kuantitatif. Tujuan dari penelitian ini adalah melakukan klasifikasi dan evaluasi model algoritma C4.5 dan Bagging+C4.5 untuk mengetahui akurasi 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 data. 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..

B.Kerangka Pemikiran

Kerangka penelitian dari penelitian bisa kita lihat pada Gambar dibawah ini :



Gambar 3.1 Kerangka Pemikiran

Adapun kerangka pemikiran pada Gambar 3.1 dapat dijelaskan sebagai berikut :

1. Preprocessing

Preprocessing data dilakukan dengan cara menangani nilai yang hilang mengikuti teknik mengabaikan tupel dengan nilai yang tidak lengkap. Setelah praproses, total 100 instans telah tersisa.

2. Model

Tool RapidMiner diterapkan pada Dataset yang baru untuk di Training maupun di Testing pada Algoritma C4.5 dan juga mentestingnya pada Algoritma Bagging+C4.5

3. Evaluasi

Kemudian Dataset diuji/dievaluasi dengan Confusion Matrix serta diukur tingkat akurasinya.

IV. HASIL DAN PEMBAHASAN

Variabel data yang digunakan pada penelitian ini disajikan pada Tabel 4.1 yakni sebagai berikut.

Tabel 4.1 Deskripsi Variable Dataset

No.	Atribut	Value
1	Degree	1.High, 2.Medium, 3.Low
2	Caprice	1.Left, 2.middle, 3. Right
3	Topic	1.Impression, 2.political, 3.news, 4. Tourism, 5.Sceintific
4	lmt	1.Yes, 2.No.
5	lpss	1.Yes, 2.No.

6	pb	1.Yes, 2.No.
---	----	--------------

Dari Tabel 4.1 Ada 5 variabel dataset gejala dan 1 variabel class penentu klasifikasi. Pada dataset ini atributnya terdiri : Pendidikan (degree), tingkah politik(caprice), topik, media local turnover (LMT) dan ruang lokal, politik dan sosial (LPSS) dan penentunya adalah professional blogger (pb).

A. Preprocessing

Dari dataset yang akan di analisis dijadikan sebagai Data Training dan Data Testing yang ada di klasifikasikan oleh algoritma C4.5 dan bagging+C4.5

Adapun contoh data training dan data testing dapat di lihat pada tabel 4.2 :

Tabel 4.2 Data Training dan Testing

No.	Atribut	Value	
		1	2
1	Degree	high	high
2	Caprice	left	left
3	Topic	Impression	political
4	lmt	yes	yes
5	lpss	yes	yes
6	pb	yes	yes

Setelah preproses, total 100 data masih valid

B. Model

Setelah Praprocessing, data Training dan Data Testing akan kita proses klasifikasi menggunakan aplikasi Rapidminer Adapun hasil dari Confusion Matrix.nya :

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 %

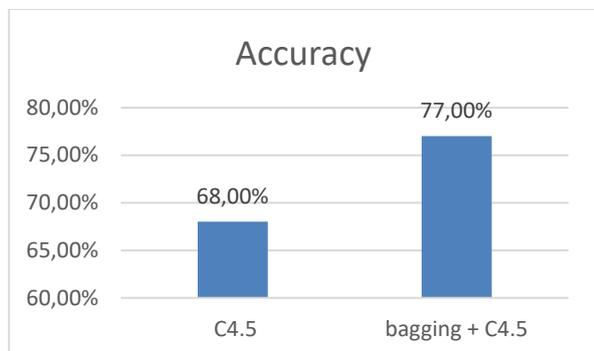
Tabel 4.4 Tabel Hasil Class Recall dan Precision bagging+C4.5

	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 %

C. Evaluasi

Hasil dari evaluasi pengklasifikasian dengan C4.5 dan bagging+C4.5 menghasilkan :



Gambar 4.1 Akurasi C4.5 dan bagging+C4.5

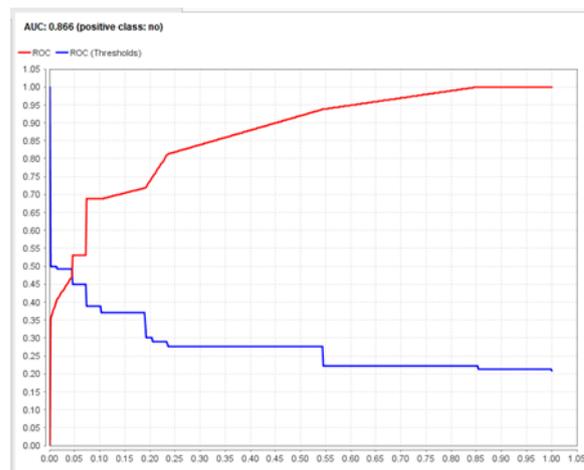
Dari gambar diatas didapatkan akurasi untuk mengklasifikasikan Dataset blogger yaitu dengan C4.5 = 68% sedangkan untuk bagging+C4.5 = 77%.

Adapun hasil dari pengujian ini juga menghasilkan gambar 4.2 dibawah ini



Gambar 4.2 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+Naive Bayes

Gambar diatas Kurva ROC (Reciver Operating Characteristic) diatas menunjukkan algoritma Bagging+C4.5 memiliki nilai AUC sebesar 0.866 yang artinya Good Classification (Bagus) ini artinya bagging meningkatkan nilai AUC

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 C4.5 dan bagging+C4.5 untuk pengklasifikasian Dataset blogger. Data ini terdiri 100 data.

Dari hasil proses Klasifikasi permodelan algoritma C4.5 terhadap Dataset Blogger menghasilkan Class Precision = 0%, dan Class Recall: 0 % sedangkan untuk algoritma bagging+C4.5 yaitu Class Precision = 100%, dan Class Recall: 28,12 % ini artinya Teknik bagging dapat meningkatkan class recall dan class precision pada klasifikasi C4.5

Hasil evaluasi klasifikasi ini menghasilkan akurasi untuk mengklasifikasikan Dataset blogger pada C4.5 yaitu 68% dan untuk bagging +C4.5 meningkat yaitu 77%

Kurva ROC (Reciver Operating Characteristic) Klasifikasi permodelan algoritma C4.5 memiliki nilai AUC sebesar 0.500 dan untuk bagging+C4.5 meningkat sebesar 0.866

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.

Berry, M. J. and Linoff, G. (1997) 'Data Mining Techniques: For Marketing, Sales, and Customer Support', *John Wiley & Sons, Inc.*

Galar, M. *et al.* (2012) 'A review on ensembles for the class imbalance problem: Bagging-, boosting-, and hybrid-based approaches', *IEEE Transactions on Systems, Man and Cybernetics Part C: Applications and Reviews*. doi: 10.1109/TSMCC.2011.2161285.

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/emered-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.

Mordelet, F. and Vert, J. P. (2014) 'A bagging SVM to learn from positive and unlabeled examples', *Pattern Recognition Letters*. doi: 10.1016/j.patrec.2013.06.010.

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.

Quinlan, J. R. (1996) 'Bagging, boosting, and C4.5', in *Proceedings of the National Conference on Artificial Intelligence*.

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

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