ANALISIS BALANCING DATA UNTUK MENINGKATKAN AKURASI DALAM KLASIFIKASI

Yolanda Eveline Ardiningtyas[1], Paulina Heruningsih Prima Rosa[2]
[1,2]Program Studi Informatika, Universitas Sanata Dharma
e-mail: [1]yolanda.eveline@gmail.com, [2]rosa@usd.ac.id

ABSTRACT

Classification is a function used to estimate the class of an object whose data label is unknown. In classification using machine learning, data that has an unbalanced class (imbalance) will affect the results of the classification, because the classification process only work in the majority data class. This results in the sample from the majority class data being classified as good, while the sample from the minority data will tend to be wrong or considered as noise data. In this study, the random forest method was used, which is a decision tree based machine learning method. The data used in this study is the quality of red wine. This record has 1599 rows (records) and 12 attributes. In this data, there is a quality attribute which represents an assessment of the quality of wine on a scale of 1-10, where the greater the value, the better the wine. This data has 6 classes that have distribution from class 3 (undrinkable) is 10 data, 4 (undrinkable) is 53 data, 5 (pretty bad) is 681 data, 6 (fair) is 635 data, 7 (quaffable) is 199 data, and 8 (very good) with 18 data. Preprocessing is done to prepare data so that it can be processed into algorithms. Preprocessing that is done is checking for noise and missing values, and transforming the data using the min-max normalization method. After the preprocessing is done, then the minority data class is balancing using the SMOTE (Synthetic Minority Oversampling Technique) method. The process of sharing testing data and training data is carried out using the K-Fold Cross Validation method. In the testing phase, 3-fold, 5-fold, 7-fold, 9-fold, and 11-fold were used with a combination of $2trees^n$, where n = 1,2,..., 11.

Keywords: *classification, imbalancing, red wine*

INTISARI

Klasifikasi merupakan fungsi yang digunakan untuk memperkirakan kelas dari suatu objek yang label datanya tidak diketahui. Pada klasifikasi menggunakan pembelajaran mesin, data yang memiliki kelas yang tidak seimbang (imbalance) akan mempengaruhi hasil dari klasifikasi, karena proses klasifikasi hanya akan berjalan di kelas data mayoritas. Hal ini mengakibatkan sampel dari data kelas mayoritas tergolong baik, sedangkan sampel dari data minoritas akan cenderung salah atau dianggap sebagai noise data. Pada penelitian ini, digunakan metode random forest yang merupakan metode pembelajaran mesin berbasis pohon keputusan. Data yang digunakan pada penelitian ini adalah data kualitas red wine. Data ini memiliki 1599 baris (records) dan 12 atribut. Pada data ini terdapat atribut quality yang merepresentasikan penilaian kualitas wine dengan skala 1-10, di mana semakin besar nilainya maka akan semakin baik anggur tersebut. Data ini memiliki 6 kelas yang memiliki distribusi kelas 3 (undrinkable) sebanyak 10 data, 4 (undrinkable) sebanyak 53 data, 5 (pretty bad) sebanyak 681 data, 6 (fair) sebanyak 635 data, 7 (quaffable) sebanyak 199 data, dan 8 (very good) sebanyak 18 data. Preprocessing dilakukan untuk mempersiapkan data agar dapat diolah ke dalam algoritma. Preprocessing yang dilakukan adalah pengecekan terhadap noise dan missing value, dan transformasi data menggunakan metode normalisasi min-max. Setelah preprocessing dilakukan, maka selanjutnya dilakukan balancing kelas data minoritas dengan menggunakan metode SMOTE (Synthetic Minority Oversampling Technique). Proses pembagian data testing dan data training dilakukan dengan metode K-Fold Cross Validation. Pada tahap pengujian, digunakan 3-fold, 5-fold, 7-fold, 9-fold, dan 11-fold dengan kombinasi pohon 2^n, dimana n = 1,2, ..., 11.

Kata kunci: *ketidakseimbangan, klasifikasi, red wine*

1. PENDAHULUAN

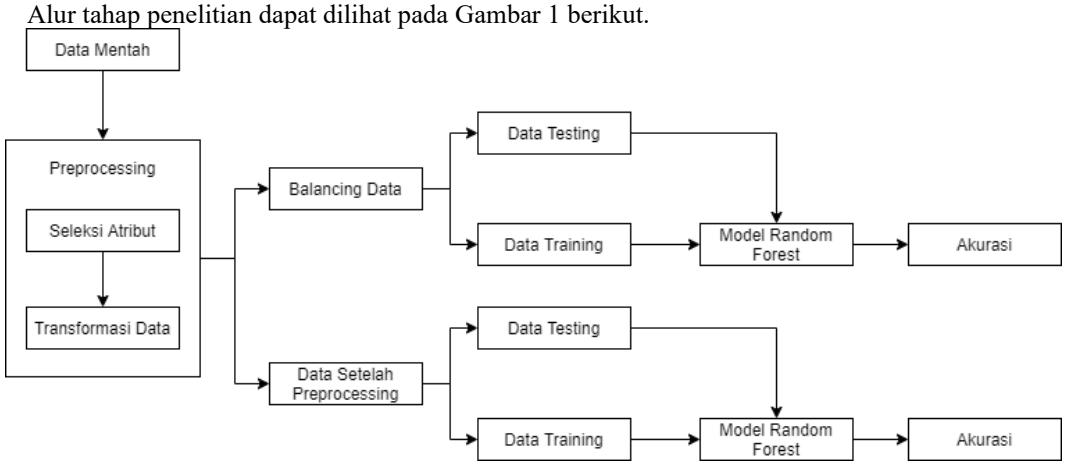
Klasifikasi merupakan fungsi yang menjelaskan atau membedakan konsep atau kelas data, dengan tujuan untuk dapat memperkirakan kelas dari suatu objek yang labelnya tidak diketahui. Sebelum melakukan proses klasifikasi menggunakan *random forest*, terlebih dahulu dilakukan *balancing* dengan menggunakan metode SMOTE (*Synthetic Minority Oversampling Technique*) untuk menyeimbangkan kelas data dengan cara membuat replikasi data dari kelas minoritas, karena pada pembelajaran mesin, data yang memiliki kelas yang tidak seimbang (*imbalance*) dapat membuat pengklasifikasian berkinerja buruk karena klasifikasinya hanya berjalan di kelas mayoritas (Poolsawad *et al.,* 2014). Proses klasifikasi dilakukan dengan menggunakan metode *random forest* yang merupakan salah satu metode pembelajaran mesin berbasis pohon keputusan yang banyak digunakan sejak diperkenalkan oleh Breiman. *Random Forest* dianggap sebagai salah satu pembelajaran yang akurat, cepat dan mudah diterapkan, juga dapat menghasilkan prediksi yang kuat. Data yang digunakan pada penelitian ini merupakan data kualitas *red wine* yang didapatkan dari website Kaggle, dan merupakan data yang bersifat *public*.

Data ini memiliki 1599 baris (*records*), dan memiliki 12 atribut. Pada atribut ke-12, yang merupakan atribut *quality* berisikan penilaian *wine* yang memiliki skala 1-10, dimana semakin besar nilainya maka semakin baik kualitas anggur tersebut.

Klasifikasi kualitas *red wine* sebelumnya juga dilakukan dengan menggunakan metode KNN untuk membandingkan metode normalisasi, dengan hasil 63,10% untuk *decimal scaling*, 65,92% untuk *min-max*, dan 65,85% untuk *z-score* (Nasution *et al.,* 2019). Selain itu pada penelitian lainnya dilakukan klasifikasi kualitas *red wine*, dan menghasilkan akurasi 58% untuk metode SVM, 64,8% untuk metode k-NN, dan 71,2% untuk metode *Random Forest* (Er dan Atasoy, 2016). Berbeda dengan dua penelitian sebelumnya, penelitian ini bertujuan untuk melihat pengaruh *balancing* data terhadap hasil akurasi pada proses klasifikasi kualitas *red wine* menggunakan *random forest*. Hal ini akan dapat dilihat pada proses pengujian menggunakan 3-*fold*, 5-*fold*, 7-*fold*, 9-*fold* dan 11-*fold* yang akan berfokus untuk membandingkan hasil akurasi antara data yang tidak dilakukan proses *balancing* dan data yang dilakukan proses *balancing*.

2. METODE PENELITIAN

Alur tahap penelitian dapat dilihat pada Gambar 1 berikut.



Gambar 1. Alur Tahap Penelitian

2.1 Dataset Penelitian

Pada penelitian ini untuk menguji *performance* dari *Random Forest* dan SMOTE, digunakan data kualitas *red wine* yang didapatkan melalui website Kaggle. Data ini memiliki kelas data yang merupakan penilaian anggur menurut ahli (Cicchetti dan Cicchetti, 2009) yang dapat dilihat pada Tabel 1 berikut.

Tabel 1. Skala Penilaian *Wine*

Peringkat Numerik	Peringkat Kata
10	*Excellent*
9	*Delicious*
8	*Very Good*
7	*Quaffable*
6	*Fair*
5	*Pretty Bad*
1-4	*Undrinkable*

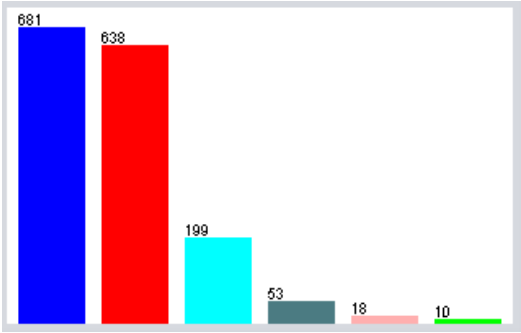
2.2 Preprocessing

Preprocessing yang dilakukan adalah dengan transformasi data menggunakan metode *min-max*. Metode *min-max* merupakan metode normalisasi dengan melakukan transformasi linier terhadap data asli. Rumus dari metode *min-max* dapat dilihat pada persamaan (1) dibawah ini:

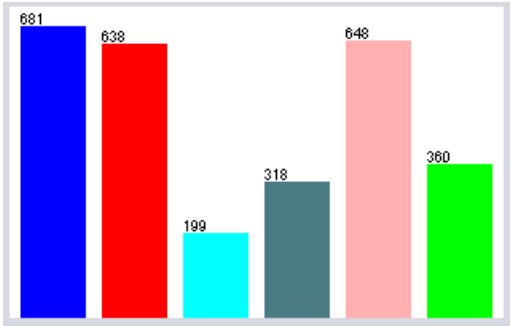
$$XB = \frac{X-Min(X)}{Max(X)-Min(X)}\left(Max_{new_A} - Min_{new_A}\right) + Min_{new_A} \quad (1)$$

2.3 Balancing Data

Pada tahap *balancing* data, dilakukan dengan menggunakan metode SMOTE atau *Synthetic Minority Oversampling Technique*. Metode ini merupakan pendekatan yang bekerja dengan membuat replikasi dari data minoritas. SMOTE bekerja dengan mencari *k-nearest neighbors* atau ketetanggan terdekat (Siringoringo, 2018). Distribusi kelas data *imbalance* dapat dilihat pada Gambar 2, dan distribusi kelas data setelah *balancing* dapat dilihat pada Gambar 3 berikut ini.



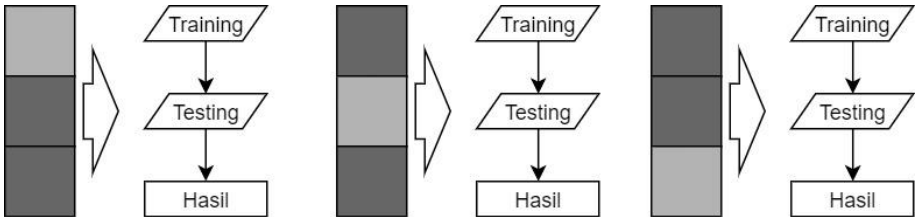
Gambar 2. Distribusi Data *Imbalance*



Gambar 3. Distribusi Data Setelah *Balancing*

2.4 Data *Testing* dan *Training*

Metode yang digunakan untuk mengevaluasi algoritma dengan membagi dua segmen menjadi *training* dan *testing* yaitu *k-fold cross validation*. Langkah pertama dalam *K-Fold Cross Validation* adalah data akan dipartisi ke dalam segmen atau *fold* yang sama atau identik. Selanjutnya adalah melakukan iterasi ke-k dari *training* dan validasi sedemikian rupa sehingga dalam setiap iterasi *fold* data yang berbeda dimunculkan untuk validasi. Sementara sisa *fold* k-1 digunakan untuk *training*. Pada Gambar 4 di bawah ini menggambarkan contoh 3-*fold validation*.



Gambar 4. Prosedur 3-Fold Validation

2.5 Random Forest

Random Forest adalah salah satu metode pembelajaran mesin berbasis pohon keputusan yang banyak digunakan sejak diperkenalkan pertama kali oleh Breiman, karena memiliki dimensi yang tinggi, dan pemrosesan yang lebih cepat berfungsi pada fitur subset (Au, 2018). *Random Forest* merupakan metode pengembangan lanjutan dari pohon keputusan CART dengan menerapkan metode *bootstrap aggregating (bagging)* dan *random feature selection* (Breiman, 2001). Algoritma *Random Forest* adalah sebagai berikut (Cutler *et al*., 2011):

1. Buat subset data dari data set menggunakan *bootstrap*.
2. Menggunakan sampel *bootstrap* sebagai data latih, dan buat pohon menggunakan partisi rekursif biner:
   a. Mulai dengan satu node.
   b. Ulangi langkah-langkah berikut untuk setiap node hingga kriteria terpenuhi:
      i. Pilih m prediktor secara acak dari prediktor yang tersedia.
      ii. Temukan pemisah biner terbaik pada m prediktor dari langkah i.
      iii. Pisahkan node menjadi dua node turunan menggunakan pemisah dari langkah ii.

Dalam membangun pohon keputusan dalam *Random Forest*, digunakan metode CART. Dimulai dengan menghitung nilai *entropy* sebagai penentu tingkat ketidakmurnian atribut dan nilai *information gain*. Untuk menghitung nilai *entropy*, digunakan rumus pada persamaan (2):

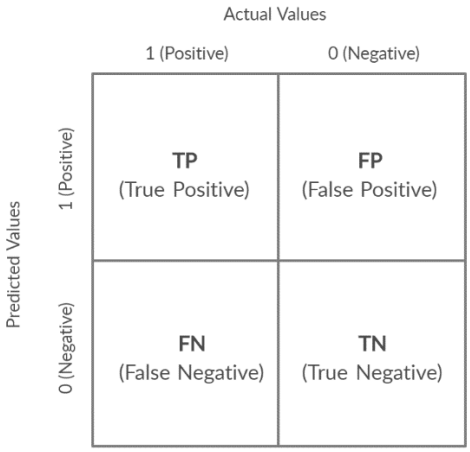
$$Entropy(Y) = -\sum ip(Y)log_2p(Y)Entropy(Y) = -\sum ip(Y)log_2p(Y) \quad (2)$$

Sedangkan rumus untuk menghitung information gain dapat dilihat pada persamaan (3) berikut ini:

$$Information\ Gain\ (Y,a) = Entropy(Y) - \sum values(a)\frac{|Yv|}{|Ya|}Entropy(Yv)) \quad (3)$$

2.6 Akurasi

Metode yang digunakan untuk menghitung akurasi adalah *confusion matrix*. *Confusion matrix* berbentuk tabel matriks yang menggambarkan kinerja model klasifikasi pada serangkaian data uji yang nilai sebenarnya diketahui. Prosedur *confusion matrix* dapat dilihat pada Gambar 5 berikut ini.



Gambar 5. Prosedur *Confusion Matrix*
(Han *et al*., 2011)

Adapun rumus yang digunakan untuk menghitung *confusion matrix* dapat dilihat pada persamaan (4) berikut ini:

$$Akurasi = \frac{TP+TN}{TP+TN+FP+FN} \quad (4)$$

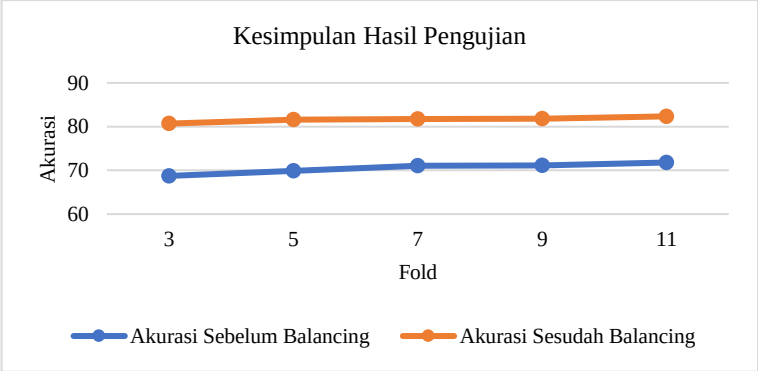
3. HASIL DAN PEMBAHASAN

Pada penelitian ini dilakukan pengujian dengan membandingkan hasil akurasi antara data yang tidak dilakukan *balancing* dengan data yang dilakukan proses *balancing*. Pengujian dilakukan dengan menggunakan kombinasi pohon 2^n, di mana n = 1,2, …, 11, dan model yang digunakan adalah 3-*fold*, 5-*fold*, 7-*fold*, 9-*fold*, dan 11-*fold*. Hasil dari pengujian dapat dilihat pada Tabel 2 di bawah ini:

Tabel 2. Kesimpulan Hasil Pengujian

Fold	Akurasi Sebelum *Balancing*	Akurasi Sesudah *Balancing*
3	68,7305%	80,6962%
5	69,9177%	81,5746%
7	71,0439%	81,7162%
9	71,1071%	81,7862%
11	71,7971%	82,3478%

Grafik dari pengujian dapat dilihat pada Gambar 6 berikut ini.



Gambar 6. Kesimpulan Hasil Pengujian

Pada Tabel 2 dan Gambar 6 merupakan kesimpulan dari hasil percobaan menggunakan 3-*fold*, 5-*fold*, 7-*fold*, 9-*fold*, dan 11-*fold*. Dari gambar di atas, dapat diketahui jika hasil akurasi terbaik adalah 82,3478% yang didapatkan dari data yang dilakukan proses *balancing* dengan membangkitkan kelas data minoritas. Sedangkan hasil akurasi terbaik yang diperoleh menggunakan data yang tidak dilakukan proses *balancing* adalah 71,7971%. Berdasarkan hal ini dapat diketahui jika proses *balancing* dapat memperbaiki hasil klasifikasi.

4. KESIMPULAN

Kesimpulan yang didapatkan dari penelitian ini adalah *balancing* data dengan menggunakan SMOTE dapat menaikkan hasil akurasi pada proses klasifikasi menggunakan *random forest*. Hasil akurasi tertinggi yang didapatkan adalah 82,3478% dengan menggunakan 11-*fold* dan jumlah pohon 1024.

UCAPAN TERIMA KASIH

Saya sangat berterima kasih kepada Drs. Hari Suparwito, S.J. M.App.IT dan Dr. Ridowati Gunawan, S.Kom., M.T. atas dukungan serta bimbingan yang telah diberikan.

DAFTAR PUSTAKA

Au, T. C. (2018). Random Forest, Decision Tree, and Categorical Predictors: The "Absent Level" Problem. *Journal of Machine Learning Research*, 19, 1-30.

Breiman, L. (2001). Random Forest, *Machine Learning*, 45, 5-32.

Cicchetti, V.D., Cicchetti, A.F. (2009). Wine rating scales: assessing their utility for producers, consumers, and oenologic researchers. *International Journal of Wine Research*, 1, 73-83.

Cutler, A. Cutler, D.R. Stevens, J.R. (2011). *Random Forest*. Berlin: Springer.

Er, Y., Atasoy, A. (2016). *The* Classification of White Wine and Red Wine According to Their Physicochemical Qualities. *International Journal of Intelligent System and Application in Engineering*, 4, 23-26.

Han, J., Kamber, M., Pei, J. (2011). *Data Mining Concepts and Techniques (edisi 3)*. Burlington: Morgan Kaufmann.

Nasution, D.A., Khotimah, H.H., Chamidah, N. (2019). Perbandingan Normalisasi Data Untuk Klasifikasi Wine Menggunakan Algoritma K-NN. *Journal of Computer Engineering System and Science*, 4 (1), 78-82.

Poolsawad, N., Kambhampati, C., Cleland, J.G.F. (2014). Balancing Class for Performance of Classification with a Clinical Dataset. *Proceedings of the World Congress on Engineering* (pp. 237-242), Hong Kong: Newswood.

Siringoringo, R. (2018). Klasifikasi Data Tidak Seimbang Menggunakan Algoritma SMOTE dan k-Nearest Neighbor. *Journal Information System Development* , 3(1), 44-49.