Statistika, Vol. 22 No. 2, 127 – 132
November 2022

Klasifikasi Varietas Unggul Padi Menggunakan Metode *Bagging, Boosting,* dan *Extremely Randomized Trees*

LUKMANUL HAKIM[1], ASEP SAEFUDDIN[2], SAUSAN NISRINA[3]

[1)]Program Studi Sains Data Universitas Insan Cita Indonesia, Indonesia
[2)]Departemen Statistika FMIPA IPB University, Indonesia
[3)]Program Studi Statistika Universitas Hamzanwadi, Indonesia
e-mail: lukman@uici.ac.id

ABSTRAK

Padi merupakan salah satu komoditas pertanian yang utama di Indonesia. Salah satu faktor pendukung yang berperan sangat penting dalam meningkatkan prodifitas padi yaitu varietas unggul. Varietas unggul padi memiliki karakteristik yang mirip antara varietas satu dengan yang lainnya. Sehingga, dalam menentukan jenis padi yang sesuai, petani perlu memilih varietas yang digunakam melalui proses klasifikasi. Pada tahapan ini terdapat tiga metode yang digunakan yaitu bagging, boosting, dan *extremely randomized trees*. Dari hasil analisis yang dilakukan secara keseluruhan metode extremaly randomized trees memiliki kekmampuan yang lebih optimal jika dibandingkan dengan metode bagging dan boosting. Hal tersebut ditunjukkan dari tiga parameter yang digunakan yaitu sensitifitas, spesifisitas dan akurasi memiliki nilai yang tertinggi. Selanjutnya di susul dengan metode boosting dan bagging.

Kata Kunci: Klasifikasi, Bagging, Boosting, Extremely randomized trees.

ABSTRACT

Rice is one of the main agricultural commodities in Indonesia. One of the supporting factors that play a very important role in increasing the productivity of rice is superior varieties. Rice superior varieties have characteristics that are similar between one variety and another. Thus, in determining the appropriate type of rice, farmers need to choose the variety used through the classification process. At this stage there are three methods used, namely bagging, boosting, and *extremely randomized trees*. From the results of the analysis carried out as a whole the *extremely randomized trees* method has more optimal capabilities when compared to the bagging and boosting methods. It is shown that the three parameters used, namely sensitivity, specificity, and accuracy have the highest values. Then followed by the boosting and bagging methods.

Keywords: Classification, Bagging, Boosting, Extremely randomized trees.

1. PENDAHULUAN

Padi merupakan salah satu produk pertanian yang menjadi komoditas utama di Indonesia (Triyanto, 2013). Padi menghasilkan beras yang merupakan kebutuhan pangan pokok bagi masyarakat. Produksi beras secara berkelanjutan merupakan suatu keharusan untuk memenuhi kebutuhan masyarakat serta menjaga konsistensi baik kondisi sosial maupun ekonomi (Simatupang, 2004). Oleh karena itu, upaya terus dilakukan dalam meningkatkan jumlah produksi padi, hal ini sejalan dengan meningkatnya jumlah penduduk dan gaya hidup masyarakat (Gurning, 2018).

Faktor pendukung yang berperan sangat penting dalam upaya meningkatkan produksi padi yaitu varietas unggul (Rachman dkk, 2019). Sebelum tahun 1970, petani di Indonesia menggunakan varietas padi lokal yang memiliki jumlah melimpah dan menyebar di lingkungan yang sempit dan berbeda (Shinta, 2001). Tahun 2021 Kementrian Pertanian merilis buku mengenai empat variaetas unggul padi yaitu INPARI, HIPA, INPAGO, dan INPARA. Varietas Unggul Padi memiliki karakteristik yang berbeda. Karakteristik padi seperti bentuk tanaman,

daun bendera, bentuk gabah, kerontokan dan sebagainya berpengaruh terhadap ketahanan padi hingga tekstur nasi. Dalam menentukan jenis padi yang sesuai, maka petani perlu memilih varietas yang digunakam melalui proses klasifikasi (Rachman, 2019).

Menurut Han (2012), dalam penelitiannya menjelaskan bahwa klasifikasi adalah suatu analisis yang digunakan untuk membedakan kelas data dengan tujuan agar dapat memprediksi kelas yang belum diketahui. Dalam pengertian yang berbeda, klasifikasi merupakan proses dalam data mining yang digunakan untuk memprediksi kelas data baru berdasarkan record kelas data sebelumnya (Arhami dkk, 2020). Pada analisis klasifikasi mensyaratkan target variabel harus kategorik. Misalnya, pada pendapatan diberi tiga kategori yaitu kategori pertama pendapatan tinggi, kategori kedua pendapatan sedang, dan kategori ketiga pendapatan rendah (Gunadi dan Sensuse, 2016). Dalam penerapannya analisis klasifikasi yang paling banyak digunakan yaitu analisis diskriminan dan regresi logistik.Mengingat tingkat kompleksitas data yang digunakan pada penelitian ini cukup tinggi dengan kelas yang tidak seimbang tentu saja membutuhkan metode khusus dalam melakukan pengklasifikasin. Metode yang mampu memprediksi kelas tidak seimbang pada level algoritma dikenal dengan istilah ensemble. Ensemble adalah menggabungkan klasifikasi tunggal dengan tujuan untuk mendapatkan hasil yang lebih baik (Saifudin, 2015). Metode ensemble yang umum dikenal dikalangan akademisi dan peneliti yaitu *bagging* dan *boosting* (Hakim dkk, 2017). Kedua metode ini diketahui mampu meningkatkan hasil akurasi pada data dengan kasus kelas tidak seimbang (Galar & Fransico, 2012).

Metode ensemble yang akan digunakan pada penelitian in yaitu *bagging*, *boosting*, dan *extremely randomized trees*. *Bootsrap aggregating* atau dikenal dengan istilah *bagging* dalam penerapannya menggunakan sub dataset untuk menghasilkan set pelatihan (Breiman, 1996). Berdasarkan Namanya, metode *bagging* terdiri dari dua tahapan utama dalam analisis, tahapan pertama yaitu resampling pengambilan contoh pada data leraning dan tahapan kedua yaitu aggregating yaitu menggabungkan banyaknya nilai dugaan menjadi satu. (Sartono dan Syafitri, 2010). Dalam penelitian yang ditulis oleh Alpaydin (2010), menjelaskan bahwa metode pembelajaran yang stabil pada perubahan dikenal dengan metode *bagging*.

Selanjutnya metode ensemble yang akan digunakan yaitu *boosting*. *Boosting* merupakan salah satu metode ensemble yang cukup populer dan masih banyak di minati dikalangan peneliti. Ide dasar dari *boosting* yaitu bobot disetiap learning diatur memiliki bobot nonnegatif (Okun, 2011). Konsep sederhana dari *boosting* yaitu memberikan bobot yang sama pada data trainig. Kemudian setelah itu, proses dilanjutkan dengan menentukan base learner (weak learner) yang merupakan suatu fungsi yang mengklasifikasikan data sampel yang telah diboboti (Fernanda, 2012).

Kemudian yang terakhir yaitu metode *extremely randomized trees*. Extra trees terkadang disebut juga sebagai metode *Extremely randomized trees*, metode ini merupakan varian baru dari metode *decision tree* (Laksana dan Suliantika, 2017). Konsep dasar dari algoritma ini yaitu membangun pohon tanpa melakukan pemangkasan dan hanya mengikuti prosedur klasik top down (Geurts dkk, 2006).

Mengingat kemampuan metode ensemble mampu melakukan pengklasifikasian dengan baik, peneliti ingin membuktikan dengan cara membandingkan 3 metode tersebut untuk memprediksi data varietas padi. Data yang digunkan memiliki permasalahan seperti kelas tidak seimbang dan jumlah kelasnya lebih dari dua/multiclass.

2. METODE PENELITIAN

Data pada yang digunakan pada penelitian ini berasal website resmi Balai Besar Penelitian Tanaman Padi Balitbangan Kementrian Pertanian (BBPADI) melalui situs https://bbpadi.litbang.pertanian.go.id/. Jumlah data yang digunakan pada penelitian ini sebanyak 112 dengan empat varietas yaitu: Inbrida Padi Sawah Irigasi (INPARI), Hibrida Padi (HIPA), Inbrida Padi Gogo (INPAGO) dan Inbrida Padi Rawa (INPARA). Masing-masing kelas terdiri dari tujuh atribut yaitu, bentuk tanaman, daun bendera, bentuk gabah, warna gabah, kerontokan, kerebahan dan tekstur nasi. Analisis data yang digunakan pada penelitian ini yaitu menggunakan analisis deskriptif untuk melihat gambaran umum dari data. Kemudian dilanjutkan dengan analisis klasifikaski dengan menggunakan metode *bagging*, *boosting*, dan *extremely randomized trees*. Pada analisis klasifikasi data dibagi menjadi dua bagian yaitu training sebesar 30% dan testing sebesar 70%. Kemudian pengukuran kebagikan model dengan menggunakan confusion matrix. Confusion Matrix digunakan untuk mengetahui seberapa baik classifier dapat mengenali sampel dari kelas yang berbeda (Han dan Kamber 2006). Berikut dibawah ini tabel 1 menjelaskan tentang confusion matrix.

Tabel 1. Confusion Matrix

	Predicted Positive	**Predicted Negative**
Actual Positive	TP	FN
Actual Negative	FP	TN

Keterangan:

Benar Positif (TP): Total prediksi benar dari data positif.

Benar Negatif (FN): Total prediksi benar dari data negatif.

Salah Positif (FP): Total prediksi salah dari data negatif.

Salah Negatif (FN): Total prediksi salah dari data positif

Dari Tabel 1 Confusion Matrix dapat diukur nilai akurasi, sensitivitasdan spesifisitas sebagai berikut:

i. Akurasi (Rahman & Afroz, 2013)

Akurasi adalah tingkat ketepatan prediksi secara keseluruhan, yaitu persentase banyaknya prediksi yang tepat pada seluruh amatan.

$$Akurasi = \frac{TP + TN}{TP + FN + FP + TN} \quad \ldots(1)$$

ii. Sensitivitas (Han & Kamber, 2006)

Sensitivitas adalah persentase ketepatan prediksi pada kelas positif, artinyaamatan yang berada pada kelas positif diprediksi positif (TP).

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

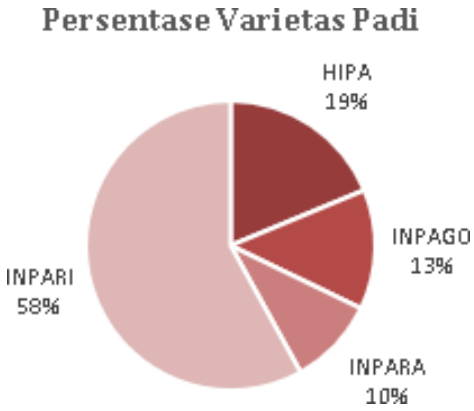
iii. Spesifisitas (Han & Kamber, 2006)

Spesifisitas adalah persentase ketepatan prediksi pada kelas negatif, artinya amatan yang berada pada kelas negatif diprediksi negatif (TN).

$$Spesifisitas = \frac{TN}{TN + FP} \quad \ldots(3)$$

3. HASIL DAN PEMBAHASAN

Untuk melihat gambaran umum dari masing-masing variaetas ditampilkan dalam bentuk grafik berikut ini.



Gambar 1. Persentase Masing-Masing Varietas

Dapat dilihat pada gambar 1 verietas padi didominasi dengan varietas inpari sebesar 58%, varietas hipa sebesar 19%, varietas inpago sebesar 13%< dan terakhir yaitu varietas inpara sebesar 10%. Masing-masing varietas tersebut memiliki ciri yang ditunjukkan pada gambar 2 berikut ini.

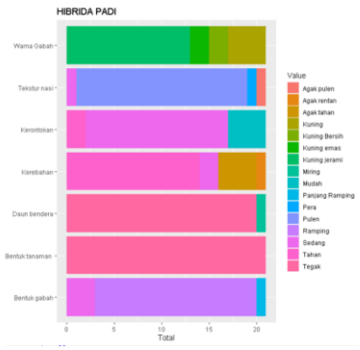


Diagram kelas HIPA

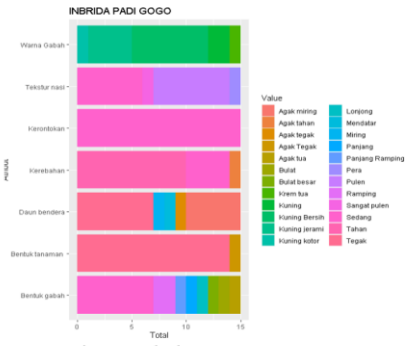


Diagram kelas INPAGO

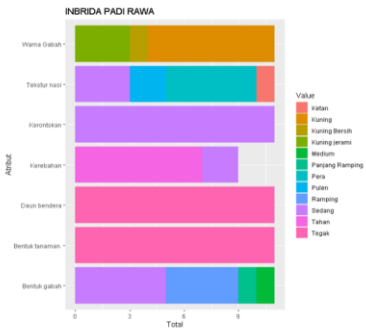


Diagram kelas INPARA

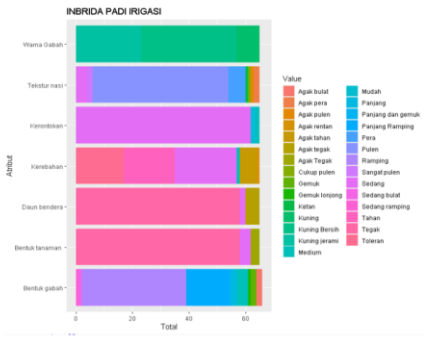


Diagram kelas INPARI

Gambar 2. menampilkan ciri-ciri dari masing varietas

Jika diperhatikan gambar 2 terdapat ciri-ciri pada masing-masing varietas diantaranya:

- Varietas padi hipa dengan ciri-ciri warna gabah kuning jerami, tekstur nasi pulen, kerontokan padi sedang, tingkat kerebahan tahan, daun bendera yang tegak, Bentuk tanaman tegak, dan bentuk gabah yang ramping.
- Varietas padi inpago memiliki ciri-ciri warna gabah yang beragam dan di dominasi dengan warna kuning bersih dan kuning jerami, tekstur nasi yang pulen dan sedang, kerontokan padi sedang, tingkat kerebahan yang tahan, daun bendera yang agak miring dan tegak, bentuk tanaman tegak, dan bentuk gabah sedang.

- Varietas inpara memiliki ciri-ciri warna gabah kuning, tekstur nasi pera, tingkat kerontokan yang sedang, tingkat kerebahan yang tahan, daun bendera tegak, bentuk tanaman tegak, dan bentuk gabah ramping dan sedang.
- Kemudian padi dengan varietas inpari memiliki ciri-ciri warna gabah didominasi dengan warna kuning bersih dan kuning jerami, tekstur nasi pulen, tingkat kerontokan sedang, yang sedang dan toleran, daun bendera yang tegak, bentuk tanaman tegak, dan rampoing dan Panjang ramping.

Dengan melihat ciri-ciri yang sudah dijabarkan pada gambar 2 langkah selanjutnya membandingkan hasil analisis ketiga metode yang digunakan seperti *boosting*, *bagging* dan *extremely randomized trees*. Pada table 2 menampilkan tingkat sensitifitas, spesifisitas, dan akurasi dari masing-masing metode terhadap 4 varietas padi.

Tabel 2. Nilai Sensitifitas, Spesifisitas, dan Akurasi

Metode		***Sensitivity***	***Specificity***	***Accuracy***
Boosting	HIPA	0.6923	0.8571	0.8000
	INPAGO	0.5000	1.0000	
	INPARA	1.0000	0.9800	
	INPARI	0.8170	0.8333	
Bagging	HIPA	0.6250	0.8537	0.7143
	INPAGO	0.4286	0.9524	
	INPARA	0.6667	0.9565	
	INPARI	0.8065	0.7778	
ERT	HIPA	0.9091	1.0000	0.9636
	INPAGO	1.0000	0.9792	
	INPARA	1.0000	0.9792	
	INPARI	0.9667	1.0000	

Dapat dilihjat pada tabel 2 menampilkan nilai sensitifitas, spesifisitas, dan akurasi dari metode *boosting*, *bagging*, dan extremely randomized tress. Nilai akurasi tertinggi terdapat pada metode *extremely randomized trees* sebesar 0.9636 atau 96.36%. Akurasi tertinggi kedua terdapat pada metode *boosting* sebesar 0.8000 atau 80%. Kemudian akurasi terkecil terdapat pada metode *bagging* sebesar 0.7143 atau 71.43%. Jika diperhatikan pada kasus multiclass metode yang mampu memberikan akurasi terbaik yaitu *extremely randomized trees*, hal itu dikarenakan karena pada metode ini tidak menggunakan data replica/boostrap. Sedangkan metode *bagging* diketahui memiliki kelemahan pada data dengan kelas yang banyak/ *multiclass*.

Secara keseluruhan nilai sensitifitas dan spesifisitas pada metode extremaly randomized trees untuk semua varietas memiliki nilai yang tertinggi jika dibandingkan dengan kedua metode lainnya. Berbeda halnya pada metode *bagging* dan *boosting* hanya mampu memberikan nilai tertinggi pada satu parameter yaitu spesifisitas atau kemampuan menebak kelas mayoritas, sedangkan nilai sensitifitasnya atau kemampuan menebak kelas minorotas sangat rendah seperti yang ditunjukkan pada varietas inpago sebesar 0.500 atau 50% pada metode *boosting* dan 0.4286 atau 42.86% pada metode *bagging*. Pada metode *bagging* hanya terdapat satu varietas yang memiliki nilai spesifisitas tertinggi yaitu pada varietas Inpari sebesar 0.8065 atau 80.65% dan sisanya masih dibawah 70%. Selanjutnya pada metode *boosting* terdapat dua varietas dengan nilai spesifisitas yang tinggi yaitu pada varietas Inpara sebesar 1.000 atau 100% dan varietas inpari sebesar 0.8170 atau 81.70% dan sisanya dibawah 70%. Jika diperhatikan metode yang paling lemah terhadap kelas tidak seimbang yaitu terdapat pada metode *bagging*.

4. SIMPULAN DAN SARAN

Kesimpulan yang dapat diambil pada penelitian ini yaitu peforma metode yang terbaik terdapat pada metode *extremely randomized trees*. Hal tersebut ditunjukkan dengan nilai akurasi,

sensitifitas, dan spesifisitas lebih tinggi jika dibandingkan dengan metode *bagging* dan *boosting*. Metode *extremely randomized trees* mampu memecahkan permasalahan klasifikasi seperti kelas tidak seimbang dan multiclass. Pada metode *bagging* dan *boosting* secara keseluruhan hanya mampu menabah kelas mayoritas dibuktikan dengan nilai spesifisitas untuk semua varietas yang tinggi. Sedangkan untuk nilai sensitifitas atau kemampuan menebak kelas minoritas terdapat varietas dengan nilai sensitifitas yang paling rendah seperti pada varietas inpago. Masing-masing memberikan nilai sensitifitas dibawah 60% pada metode *bagging* dan *boosting*. Atrinbya metode *bagging* dan *boosting* tidak cukup kuat untuk memprediksi kelas tidak seimbang.

Saran untuk penelitian selanjutnya agar menggunakan data yang beragam untuk membuktian ketiga metode tersebut. Apakah memang kuat untuk data tidak seiambang atau tidak.

DAFTAR PUSTAKA

Alfaro, E., Gamez, M., dan Garcia, N. (2013). adabag: An R Package for Classification with *Boosting* and *Bagging*. Journal of Statistical Software, 11-35.

Alpaydin, E. (2010). Introduction to Machine Learning. London: The MIT Press.

Arhami dan Nasir,M. S. T. (2020). *Data Mining-Algoritma dan Implementasi*. Penerbit Andi.

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

Fernanda, J. W. (2012). *Boosting* Neural Network dan *Boosting* Cart Pada Klasifikasi Diabetes Militus Tipe II. Jurnal Matematika. Volume 2. Nomor 2. ISSN : 1693-1394

Galar, M dan Fransico. (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: Application And Reviews. Volume 42. Nomor 4.

Geurts, P. Ernst, D dan Wehenkel, L. (2006). Extremely randomized rees. Mach Learn. DOI 10.1007/s10994-006-6226-1

Gunadi, G dan Sensuse, D.I. (2016). Penerapan Metode Data Mining Market Basket Analysis Terhadap Data Penjualan Produk Buku Dengan Menggunakan Algoritma Apriori dan Frequent Pattern Growth (FP-Growth): Studi Kasus Percetakan PT. Gramedia. Jurnal Telematika Mkom. 4(1).

Gurning, I. P., Yuprin, A. D., dan Taufik, E. N. (2019). Trend dan Estimasi Produksi Padi dan Konsumsi Beras di Provinsi Kalimantan Tengah. *Journal Socio Economics Agricultural*, *14*(1), 48-61.

Hakim, L., Sartono, B., dan Saefuddin, A. (2017). *Bagging* Based Ensemble Classification Method on Imbalance Datasets. JCSN - International Journal of Computer Science and Network, Volume 6, Issue 6.

Han, J dan Kamber, M. (2006). Data mining concepts and techniques second edition. San Francisco: Diane Cerra.

Han, J., Kamber, M., dan Pei, J. (2012). Data Mining Concepts and Techniques. San Fransisco: Morgan Kauffman.

Laksana, E. A dan Sulianta, F. (2017). Analisis Dan Studi Komparatif Algoritma Klasifikasi Genre Musik. *Semnasteknomedia Online*, *5*(1), 2-1.

Okun, O.(2011). Feature Selection and Ensemble in Machine Learning Bioinformatics: Algoritmic Classification and Implementations. United States of America: IGI Global.

Rachman, A., Furkon, M.T., dan Ramdani, F. (2019). Klasifikasi Varietas Unggul Padi menggunakan Algoritme C4.5. Jurnal Pengembangan Teknologi Informasi dan Ilmu Komputer. 3(9).

Rahman, R dan Afroz, F. (2013). Comparison of Various Classification Techniques Using Different Data Mining Tools for Diabetes Diagnosis. Journal of Software Engineering and Applications.

Saefudin, A. (2015). Penerapan Teknik Ensemble untuk Menangani Ketidakseimbangan Kelas pada Prediksi Cacat Software. Journal of Software Engineering. Volume 1. Nomor 1.

Sartono, B dan Syafitri. (2010). "Metode pohon gabungan: solusi pilihan untuk mengatasi kelemahan pohon regresi dan klasifikasi tunggal". Forum Statistika dan Komputasi, 15 (1), 1-7.

Shinta, A. (2001). *Ilmu Usaha Tani*. Universitas Brawijaya Press.

Simatupang, P dan Rusastra, I. W. (2004). Kebijakan Pembangunan Sistem Agribisnis Padi. *Ekonomi Padi dan Beras Indonesia*, 31-52.

Tanha, J., Abdi, Y., Samadi, N., Razzaghi, N., dan Asadpour, M. (2020). *Boosting* methods for multi-class imbalanced data classification: an experimental review. Journal of Big Data. 7:70.

Triyanto, A. C dan Hardinto, P. (2013). Analisis Produktivitas Sektor Pertanian Komoditi Tanaman Padi Berbasis Agribisinis Dalam Peningkatan Ekonomi. *JESP*, *5*(1), 53-62.