Techno.COM, Vol. 20, No.3, Agustus 2021: 400-410

Komparasi Kinerja Algoritma C4.5, Random Forest, dan Gradient Boosting untuk Klasifikasi Komoditas

Performance Comparison Between C4.5 Algorithm, Random Forests, and Gradient Boosting for Commodity Classification

Edi Ismanto[1], Melly Novalia[2]
[1,2]Pendidikan Informatika, Universitas Muhammadiyah Riau
E-mail: [1]edi.ismanto@umri.ac.id, [2]melly_novalia@umri.ac.id

Abstrak

Penentuan komoditas unggulan pada suatu daerah merupakan hal yang sangat penting untuk dilakukan, salah satunya di Provinsi Riau. Memahami mengenai prioritas perencanaan pengembangan wilayah yang diarahkan pada pengembangan komoditas unggulan. Sejauh ini Provinsi Riau memiliki potensi komoditas disektor perkebunan yang sangat menjajikan, data yang ada sebelumnya banyak digunakan sebagai laporan, dalam bentuk data *excel*. Data komoditas bisa digali dengan teknik *data mining* untuk mendapatkan pola klasifikasi, sehingga lebih memudahkan Pemerintah Provinsi Riau dalam mendapatkan informasi komoditas unggulannya. Pada penelitian ini, dilakukan pengujian kinerja algoritma klasifikasi yang banyak digunakan dalam *data mining*, agar mendapatkan algoritma yang memiliki kinerja paling baik untuk klasifikasi data komoditas. Beberapa penelitian mengatakan algoritma klasifikasi C4.5 memiliki kinerja kurang baik dibandingkan dengan algoritma yang lain seperti *random forest*, dan *gradient boosting*. Dalam penelitian ini dilakukan perbandingan antara algoritma C4.5, *random forest*, dan *gradient boosting*, untuk mengukur kinerja terbaik dalam melakukan klasifikasi data komoditas. Data yang digunakan dalam penelitian ini yaitu data komoditas perkebunan Provinsi Riau pada tahun 2019. Hasil dari penelitian ini, algoritma yang memiliki kinerja terbaik untuk klasifikasi adalah algoritma *random forest* dengan syarat menggunakan *shuffle sampling*. Dan mayoritas *linear sampling* menghasilkan kinerja kurang baik. Sedangkan *shuffle sampling* memiliki kinerja sangat baik untuk algoritma berbasis *tree*.

Kata kunci: *Data Mining, Algoritma C4.5, Random Forest, Gradient Boosting*, Klasifikasi

Abstract

Determination of superior commodities in an area is very important to do, one of which is in Riau Province. Understand the priority of regional development planning that is directed at developing superior commodities. So far, Riau Province has the potential for commodities in the plantation sector which is very promising, the existing data is widely used as a report, in the form of excel data. Commodity data can be extracted with data mining techniques to obtain classification patterns, making it easier for the Riau Provincial Government to obtain information on its superior commodities. In this research, we tested the performance of the classification algorithm which is widely used in data mining, in order to obtain the best performing algorithm for the classification of commodity data. Some studies say the C4.5 classification algorithm has less performance than other algorithms such as random forest and gradient boosting. In this study, a comparison between the C4.5 algorithm, random forest, and gradient boosting was carried out to measure the best performance in classifying commodity data. The data used in this study is the plantation commodity data of Riau Province in 2019. The results of this study show that the algorithm that has the best performance for classification is the random forest algorithm on the condition that it uses shuffle sampling. And the majority of linear sampling results in poor performance. Meanwhile, shuffle sampling has a very good performance for tree-based algorithms.

Keywords: Data Mining, C4.5 Algorithm, Random Forest, Gradient Boosting, Classification

1. PENDAHULUAN

Penentuan komoditas unggulan pada suatu daerah merupakan hal yang sangat penting untuk dilakukan. Komoditas unggulan berdasarkan kontek kepentingan ekonomi secara makro, memiliki tujuan khusus untuk mendukung tercapainya peningkatan pertumbuhan ekonomi, dalam kerangka penciptaan lapangan kerja, dan meningkatkan daya saing perekonomian ditengah kompetisi yang makin mengglobal. Terbitnya peraturan menteri dalam negeri republik indonesia nomor 9 tahun 2014 tentang pedoman pengembangan produk unggulan daerah (PUD), mengamanatkan kepada pemerintah daerah untuk menyusun dan menetapkan produk unggulan daerah (PUD) setiap tahun (pasal 2 Ayat 1). Provinsi Riau dalam mengembangkan sektor perekonomian sebagai salah satu modal pembangunan, tidak terlepas dari berbagai masalah yang bersifat umum maupun yang bersifat strategis kewilayahan seperti; ketahanan pangan,kemiskinan, dan pembangunan daerah tertinggal. Berdasarkan data yang ada, provinsi Riau memiliki potensi komoditas disektor perkebunan, data yang ada ini, bisa digali dengan teknik data mining untuk mendapatkan pola klasifikasi, sehingga lebih memudahkan dalam mendapatkan informasi komoditas unggulan.

Data mining dapat diartikan sebagai serangkaian proses untuk menggali nilai tambah berupa informasi yang selama ini tidak diketahui secara manual dari suatu basis data dengan melakukan proses ekstraksi dan menggali pola penting dari data yang ada [1]. Klasifikasi adalah proses menemukan model atau fungsi yang menggambarkan dan membedakan kelas atau konsep data. Algoritma yang sering digunakan untuk penyelesaian klasifikasi data seperti Naïve Bayes, Decission Tree, Linear Regresion, k-NN (k Nearest Neighbors), Neural Network, Support Vector Machine, dan Logistic Regresion [1].

Pengembangan algoritma decission tree seperti C45, random forest, dan gradient boosting merupakan algoritma yang banyak digunakan dalam menangani beberapa kasus klasifikasi dan prediksi. Berdasarkan penelitian[2] algoritma C45 mampu menangani masalah klasifikasi data hama tanaman padi, tetapi pada penelitian ini, tidak melakukan pengukuran kinerja algoritma C45. Penelitian yang dilakukan [3] menghasilkan data klasifikasi jenis pekerjaan alumni dengan menggunakan 259 kasus data, penelitian ini menfokuskan pengukuran atribut gain ratio (GR) yang digunakan oleh algoritma C45, hasil atribut gain ratio memperoleh nilai yang cukup baik. Pada penelitian[4] algoritma C45 mampu menyelesaikan permasalahan klasifikasi data penjualan produk makanan dengan menghasilkan nilai gain ratio yang baik. Berdasarkan[5][6][7][8] algoritma C45 mampu menyelesaikan permasalahan prediksi dan klasifikasi data dengan menghasilkan nilai kinerja akurasi algoritma C45 yang baik.

Algoritma random forest merupakan algoritma yang dapat digunakan untuk penyelesaian permasalahan klasifikasi data. Seperti penyelesaian masalah klasifikasi penyakit daun tomat yang dilakukan pada penelitian[9] algoritma random forest dikombinasikan dengan fitur eskstraksi hu-moment dan haralick diperoleh nilai akurasi yang sangat baik. Penggunaan random forest pada[10] untuk menyelesaikan permasalahan klasifikasi juga mampu menghasilkan nilai akurasi yang sangat baik, dengan melakukan kombinasi fitur scala invariant feature transform (sift). Begitu juga penelitian yang dilakukan[11] penyelesaian permasalahan klasifikasi data kelayakan kredit telah diperoleh kinerja akurasi algoritma random forest yang sangat baik. Algoritma random forest juga mampu menyelesaikan permasalahan prediksi data seperti penelitian yang dilakukan pada[12][13][14] dimana algoritma random forest mampu melakukan prediksi data harga bitcoin, data harga ponsel, dan data curah hujan dengan hasil nilai akurasi yang baik diatas 85%.

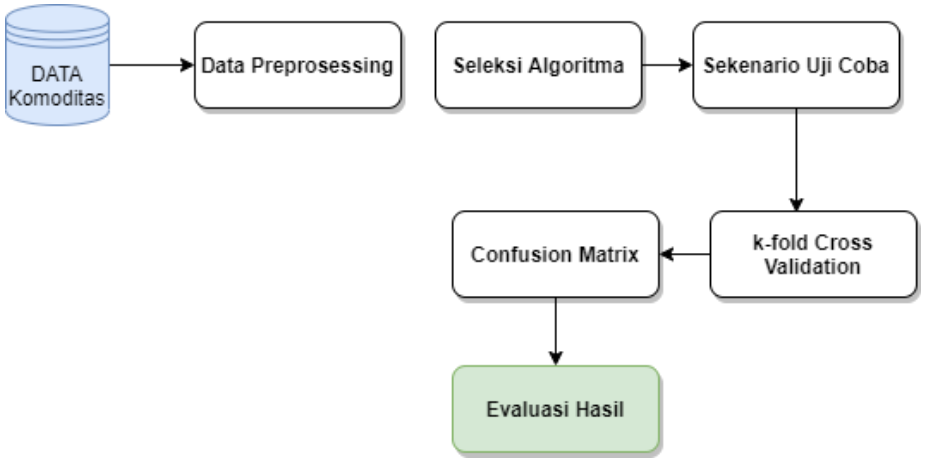
Gradient boosting adalah algoritma yang menggunakan teknik ensamble dari decision tree, algoritma ini mampu menyelesaikan persoalan klasifikasi dan prediksi data. Seperti pada penelitian[15] untuk klasifikasi kebakaran hutan dan lahan kombinasi antara XGboost dan feature importance algoritma ini mampu menghasilkan akurasi sebesar 89.52%. Sedangkan pada penelitian[16] algoritma gradient boosting untuk menyelesaikan persoalan deteksi kelulusan mahasiswa dengan menambahkan teknik smote dan bagging telah diperoleh nilai akurasi sebesar 80.57% dengan kategori klasifikasi yang baik. Klasifikasi penerimaan sinyal GPS[17]

klasifikasi gradient boosting memberikan hasil yang cukup baik. Begitu juga untuk persoalan prediksi seperti pada penelitian[18] algoritma gradient boosting mampu melakukan prediksi data kecelakaan lalu lintas dengan nilai Root Mean Square Error (RMSE) diatas 4. Penyelesaian permasalahan klasifikasi data juga bisa diatasi pada algoritma gradient boosting dengan menggunakan teknik bagging seperti pada penelitian[19] dimana mampu meningkatkan akurasi prediksi sebesar 6% pada kasus prediksi kanker payudara.

Dari beberapa hasil review penelitian sebelumnya ketiga algoritma klasifikasi C45, random forest, dan gradient boosting. Algoritma C4.5 masih memiliki kinerja kurang baik jika dibandingkan dengan *random forest*, dan *gradient boosting* untuk kasus dengan dataset yang berbeda. Sehingga pada penelitian ini akan dilakukan komparasi dari tiga algoritma klasifikasi *C4.5, random forest*, dan *gradient boosting* untuk mendapatkan kinerja akurasi yang paling baik untuk menyelesaikan persoalan klasifikasi data komoditas.

2. METODE PENELITIAN

Metode yang dilakukan dalam penelitian ini yaitu seperti pada gambar 1. Tahap pertama melakukan pengumpulan data komoditas perkebunan di Provinsi Riau. Data asli berjumlah 720 data komoditas perkebunan selama tahun 2019, yang terdiri dari atribut nama kabupaten/kota, nama komoditi, luas areal tanaman, jumlah produksi.



Gambar 1 Metode penelitian yang digunakan

2.1 Data Preprocessing

Pada tahapan ini melakukan pembersihan dan perapian dataset untuk mengatasi missing value, data noise, dan data yang tidak konsisten. Tahapan yang dilakukan meliputi proses (1) data *cleaning*, (2) data *integration*, (3) data *selection*, dan (4) data *transformation* untuk menjadi dataset komoditas yang siap dilakukan pengujian. Dari proses preprosessing didapat dataset dengan jumlah data sebanyak 60 data komoditas dengan rincian atribut nama kabupaten/kota, tahun, kode komoditas, jenis komoditas, nama komoditas, luas area, jumlah produksi, dan potensi. Adapun rincian tipe datanya seperti pada tabel 1 berikut.

Tabel 1 Komposisi dataset komoditas

No	Nama atribut	Tipe Data
1	Kab/Kota	Polynominal
2	Tahun	Integer
3	Kode Komoditas	Polynominal
4	Jenis Komoditas	Polynominal
5	Nama Komoditas	Polynominal
6	Luas Area (Ha)	Integer
7	Jumlah Produksi (ton)	Integer
8	Potensi	Polynominal

Atribut potensi memiliki skala yaitu sangat unggul, unggul, sangat baik, baik, cukup, dan kurang. Dengan *rule based* sebagai berikut jika luas area > 6.500 Ha dan jumlah produksi > 18729 ton, maka memiliki potensi sangat unggul, jika luas area > 6.500 Ha dan jumlah produksi ≤ 187.29 ton , maka memiliki potensi unggul, jika luas area > 4.370 Ha dan jumlah produksi ≤ 187.29 ton, maka memiliki potensi sangat baik, jika luas area < 145 Ha dan jumlah produksi ≤ 102.500 ton, maka memiliki potensi baik, jika luas area < 145 Ha dan jumlah produksi ≤ 102.500 ton, maka memiliki potensi cukup. Hasil dataset yang digunakan dapat dilihat pada gambar 2.

Row No.	Kab/kota	Tahun	Kode	Jenis Komoditas	Nama Komoditas	Luas Area (Ha)	Jumlah Produksi (Ton)	Potensi
1	Kuantan Sing...	2019	KS01	Perkebunan	Kelapa Sawit	128750	450804	Sangat Unggul
2	Kuantan Sing...	2019	KP02	Perkebunan	Kelapa	2760	1924	Sangat Baik
3	Kuantan Sing...	2019	KT03	Perkebunan	Karet	139202	83983	Unggul
4	Kuantan Sing...	2019	KI04	Perkebunan	Kopi	13	5	Cukup
5	Kuantan Sing...	2019	KO05	Perkebunan	Kakao	2181	660	Baik
6	Indragiri Hulu	2019	KS01	Perkebunan	Kelapa Sawit	118969	469273	Sangat Unggul
7	Indragiri Hulu	2019	KP02	Perkebunan	Kelapa	1828	250	Sangat Baik
8	Indragiri Hulu	2019	KT03	Perkebunan	Karet	61370	32306	Unggul
9	Indragiri Hulu	2019	KI04	Perkebunan	Kopi	348	44	Cukup
10	Indragiri Hulu	2019	KO05	Perkebunan	Kakao	638	117	Baik
11	Indragiri Hilir	2019	KS01	Perkebunan	Kelapa Sawit	227802	731009	Sangat Unggul
12	Indragiri Hilir	2019	KP02	Perkebunan	Kelapa	351526	361348	Unggul
13	Indragiri Hilir	2019	KT03	Perkebunan	Karet	5653	4616	Sangat Baik

ExampleSet (60 examples, 0 special attributes, 8 regular attributes)

Gambar 2 Dataset komoditas yang terbentuk

2.2 Seleksi Algoritma

Seleksi algoritma yaitu memilih dan melakukan pengujian algoritma klasifikasi yang paling baik untuk digunakan dalam mengklasifikasi data komoditas, tiga algoritma decision tree yang telah di review dan banyak digunakan dalam penyelesaian kasus klasifikasi dengan kinerja yang sangat baik yaitu algoritma C4.5, random forest, dan gradient boosting. Kriteria pengukuran algoritma yang akan digunakan dalam komparasi yaitu gaint ratio, information gain, gini index, dan akurasi.

2.3 Pengujian Algoritma

Data komoditas dibagi menjadi 2 bagian yaitu dataset komoditas sebagai training dan dataset komoditas sebagai testing. Dimana dalam tahap pengujian dilakukan pengukuran kinerja dari setiap algoritma C4.5, random forest, dan gradient boosting. Adapun teknik pengujian model yang akan digunakan yaitu random sampling dengan teknik linear, shuffle, stratified. Dan kriteria pengukuran yaitu gaint ratio, information gain, gini index, dan akurasi.

2.3.1 Tahapan penyelesaian dengan Algoritma C4.5

Algoritma C4.5 merupakan pengembangan dari algoritma ID3 yang menggunakan entropi informasi, atribut kontinyu dan diskret, atribut kategorial dan numerik, dan missing values[20].

Tahapan pengujian algoritma C4.5 dilakukan dengan menggunakan langkah-langkah sebagai berikut.

1. Akar dipilih dari salah satu atribut.
2. Tiap nilai dibuat cabang.
3. Pada cabang dbagi kasus.
4. Hingga setiap kasus memiliki kelas yang sama pada cabang maka ulangi proses tersebut.

Pemilihan akar dari atribut, berdasarkan pada atribut yang memiliki nilai gain tertinggi. Dengan menggunakan formula sebagai berikut;

$$Gain\ (S, A) = Entropy\ (S) - \sum_{1=1}^{n} \frac{|Si|}{|S|} * Entropy\ (Si) \quad (1)$$

Keterangan:
S = kelompok kasus
A = atribut
n = banyaknya bagian atribut A
|Si| = banyaknya kasus pada bagian ke-i
|S| = banyaknya kasus dalam S

Kemudian lakukan perhitungan nilai entropi dengan formula sebagai berikut:

$$Entropy\ (S) = \sum_{i=1}^{n} - pi * log_2\ pi \quad (2)$$

Keterangan:
S = kelompok kasus
n = banyaknya partisi S.
pi = perbandingan dari Si terhadap S

2.3.2 Tahapan penyelesaian dengan Algoritma Random Forest

Random Forest memetakan atribut dari kelas sehingga dapat digunakan untuk menemukan klasifikasi terhadap data yang belum muncul. Dinamakan Random Forest karena merupakan keturunan dari pendekatan ID3 untuk membangun pohon keputusan [20].

Berikut tahapan pengujian kinerja Algoritma Random Forest:

1. Perhatikan label pada data, jika sudah sama semua, maka akan dibentuk daun dengan nilai label data keseluruhan.
2. Menghitung nilai informasi dengan menggunakan semua data yang ada, dengan formula sebagai berikut:

$$info\ (D) = -\sum_{i=1}^{m} pi\ log^{2(pi)} \quad (3)$$

Keterangan:

Formula diatas merupakan probabilitas tuple dalam D yang menjadi kelas dengan asumsi atau disebut juga entropy dari D merupakan rata-rata informasi yang diperlukan untuk identifikasi tuple dalam D.

Jika nilai A adalah nilai diskrit maka data D akan dipisahkan sejumlah nilai data A sehingga nilai setiap cabang akan murni dan sejenis. Setelah percabangan pertama, jumlah percabangan yang mungkin terjadi diukur dengan formula sebagai berikut:

$$info\ A(D)\ \sum_j^v \frac{|Dj|}{|D|} * info\ A(Dj) \quad (4)$$

3. Menghitung nilai informasi dengan formula sebagai berikut
4. Untuk setiap atribut dengan memperhatikan isi data dari atribut. Dimana $\frac{|Dj|}{|D|}$ merupakan bobot dari partisi j. $info_A(D)$ merupakan informasi yang diperlukan untuk mengklasifikasikan tuple dari D pada partisi A. Semakin kecil hasil persamaan ini, semakin baik pula partisi yang dihasilkan. Nilai dari sebuah atribut menentukan penting tidaknya atribut tersebut dalam penyusunan pohon keputusan. Jika atribut bernilai kontinyu, maka akan dicari split_point dengan cara mengurutkan seluruh data menurut atribut tersebut dari kecil ke besar, lalu di rata-rata antar satu data dengan data setelahnya. Nilai informasi akan dihitung menurut satu persatu calon split_point dan nilai split_point yang akan dipilih yang terkecil. (4) Nilai gain untuk setiap atribut akan diperhitungkan dengan formula (2.3), nilai dengan gain tertinggi akan dijadikan cabang dalam pohon keputusan.

$$gain\ (A) = Info\ (D) - InfoA(D) \quad (5)$$

5. Setelah cabang pohon keputusan terbentuk, perhitungan dilakukan kembali seperti pada tahap 1 sampai 4. Namun jika cabang telah mencapai maksimal cabang yang diperbolehkan, daun akan terbentuk dengan nilai mayoritas dari nilai data.

2.3.3 Tahapan penyelesaian dengan Algoritma Gradient Boosting

Algoritma Gradient Boosting mampu membangun decision tree berdasarkan peningkatan dalam struktur pohon pada pembelajaran yang lemah untuk memperbaiki kesalahan pohon dan mencegah terjadinya potensi overfitting. Dalam membangun decision tree, dapat dilakukan penambahan jumlah iterasi yang sangat konservatif yang dapat menghasilkan dan meningkatkan kinerja model yang lebih baik. Gradient Boosting mampu memecahkan masalah dengan menyesuaikan pembelajaran lemah dengan gradien negatif dari fungsi kerugian (loss function) dan meningkatkan pohon (trees) dengan parameter yang mewakili variabel split yang dipasang pada setiap node terminal pohon [20].

Tahapan pengujian Algoritma Gradient Boosting yang dilakukan dengan menggunakan prosedur berikut:

1. Tentukan dataset training D :

$$D = \{(X_1, Y_1), \ldots\ldots (X_n, Y_n)\}, \ (6)$$

2. Sampel T set elemen n dari D (replacement)

$$D_1, D_2, \ldots\ldots\ldots\ldots\ldots D_T \quad \rightarrow T\ dataset\ pelatihan \ (7)$$

3. Training di setiap

$$D_1, i = 1, \ldots\ldots, T\ dan\ urutan\ T\ output\ F_1(x), \ldots\ldots\ F_T(x) \ (8)$$

4. Aggregate classifier dapat digunakan untuk regresi maupun klasifikasi dengan formula pada (9) sampai (11).

Regresi:

$$f(x) = \sum_{i=1}^{T} F_i\ (x) \quad (9)$$
$$rata - rata\ f_{1\ untuk\ i=1,\ldots\ldots,T}$$

Klasifikasi:

$$f(x) = sign\ \left(\sum_{i=1}^{T} f_i\ (x)\right) \quad (10)$$

atau

$$f(x) = sign\ \left(\sum_{i=1}^{T} sign\ (f_i\ (x))\right) \quad (11)$$

2.4 Tahap validasi

Metode validasi yang digunakan pada penelitian ini seperti penelitian pada umumnya yang menggunakan K-Fold Cross Validation. K-Fold Cross Validation adalah teknik validasi dengan membagi data secara acak kedalam k bagian dan masing-masing bagian akan dilakukan proses klasifikasi[21]. Dengan menggunakan cross validation akan dilakukan percobaan sebanyak k. Data yang digunakan dalam percobaan ini adalah data training komoditas untuk mencari nilai error rate secara keseluruhan. Secara umum, pengujian nilai k dilakukan sebanyak 10 kali untuk memperkirakan akurasi estimasi. Dalam penelitian ini nilai k yang digunakan berjumlah 10 atau 10-fold Cross Validation.

2.5 Tahap Evaluasi Kinerja Algoritma

Evaluasi kinerja algoritma klasifikasi hanya menggunakan confusion matrix. Confusion matrix memberikan keputusan yang diperoleh dalam traning dan testing, confusion matrix memberikan penilaian performance klasifikasi berdasarkan objek dengan benar atau salah[21]. Confusion matrix berisi informasi aktual dan prediksi pada sistem klasifikasi.

Dari kriteria pengukuran yang digunakan, sudah dapat menggambarkan kinerja dari tiga algoritma klasifikasi. Uji coba dilakukan sebanyak 100 kali, di mana skenario uji coba berdasarkan tipe numerik dan polinomial sebanyak 50 kali. Masing-masing algoritma C4.5, random forest, dan gradient boosting dilakukan uji coba berdasarkan tiga sampling techniques yaitu linear sampling, shuffle sampling dan stratified sampling. Parameter kriteria yang digunakan terdiri dari gain ratio, information gain, gini index dan accuracy[22].

2.6 Tahap Membandingkan Hasil Evaluasi

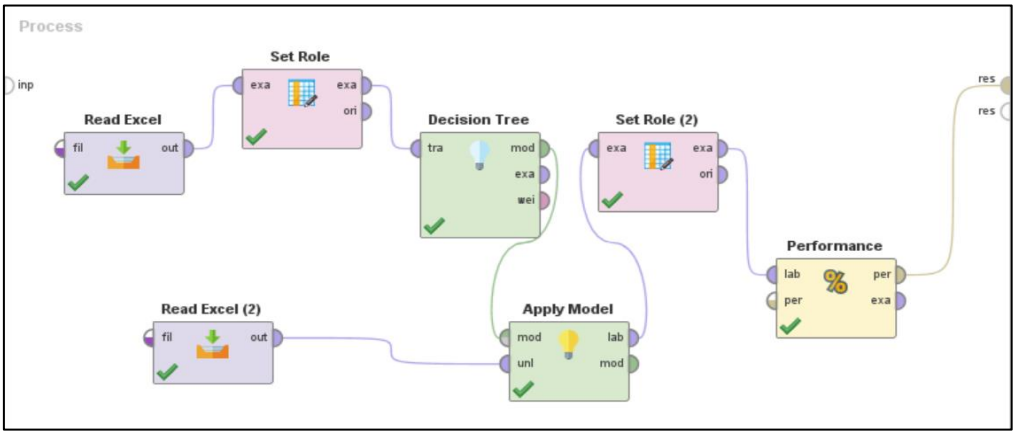
Melakukan simpulan hasil untuk membandingkan masing-masing algoritma dari berbagai pengujian, agar mengetahui algoritma mana yang memiliki kinerja terbaik dan yang lemah guna mengatasi permasalahan yang ada dalam penelitian ini.

3 HASIL DAN PEMBAHASAN

Berdasarkan hasil uji coba yang telah dilakukan dalam penelitian ini, dengan membandingkan beberapa algoritma *C4.5*, *random forest*, dan *gradient boosting* untuk klasifikasi data komoditas dengan menggunakan beberapa parameter algoritma sehingga didapat hasil sebagai berikut.

3.1 Pengujian Model Algoritma C4.5

Adapun uji coba model algoritma C4.5 dengan menggunakan tipe data *polynominal* dan *integer* dataset komoditas menggunakan software RapidMiner dilakukan dengan tahapan seperti pada gambar 3.



Gambar 3 Pengujian model algoritma C4.5

Dari beberapa kali hasil pengujian model klasifikasi algoritma C4.5 dengan teknik random sampling model linear, shuffle, dan stratified dengan menggunakan kriteria gain ratio, information gain, gini index, accuracy telah didapat hasil seperti yang tertera pada tabel 2.

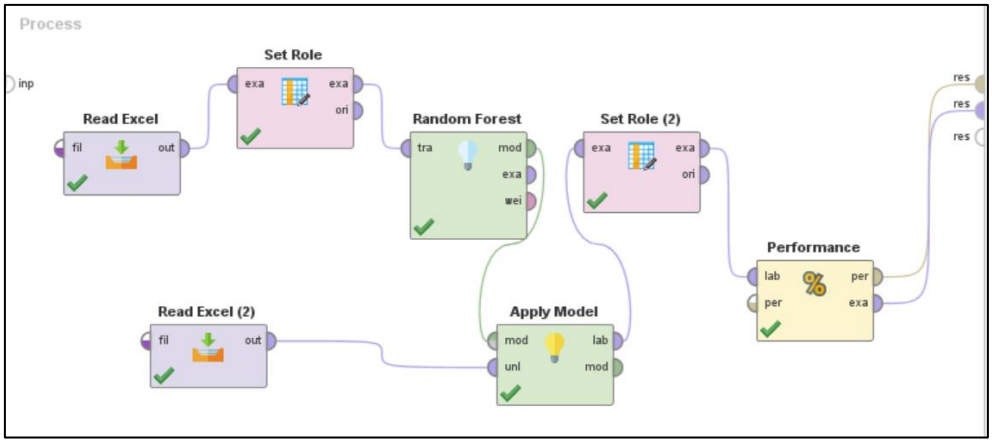
Tabel 2 Hasil uji coba algoritma C4.5

Random Sampling	***Criterion***	***Accuracy (%)***	***Weighted Mean Recall (%)***	***Weighted Mean Presisi (%)***
Linear	*Gain Ratio*	90.00	90.50	92.00
	Information Gain	88.33	87.09	89.70
	Gini Index	88.33	87.09	89.70
	Accuracy	**86.67**	**85.70**	**88.71**
Shuffle	*Gain Ratio*	**90.00**	**90.50**	**92.00**
	Information Gain	88.00	70.00	83.33
	Gini Index	88.00	80.00	91.67
	Accuracy	86.00	90.91	90.00
Stratified	*Gain Ratio*	90.00	83.33	92.31
	Information Gain	88.33	91.67	87.50
	Gini Index	88.00	90.00	90.00
	Accuracy	87.00	91.67	87.50

Dari tabel.2 dapat terlihat bahwa algoritma C4.5 memiliki kinerja terbaik ketika menggunakan teknik *shuffle sampling* dengan kriteria *gain ratio.* Dan algoritma C4.5 memiliki hasil kurang baik ketika menggunakan teknik *linear sampling* dengan kriteria *accuracy*.

3.2 Pengujian Model Algoritma Random Forest

Adapun uji coba model algoritma Random Forest dengan menggunakan tipe data *polynominal* dan *integer* dataset komoditas menggunakan software RapidMiner dilakukan dengan tahapan seperti pada gambar 4.



Gambar 4 Pengujian model algoritma random forest

Dari beberapa kali hasil pengujian model klasifikasi algoritma Random Forest dengan teknik random sampling model linear, shuffle, dan stratified dengan menggunakan kriteria gain ratio, information gain, gini index, accuracy seperti yang telah dilakukan pengujian pada algoritma C4.5 maka didapat hasil seperti yang tertera pada tabel 3.

Tabel 3 Hasil uji coba algoritma random forest

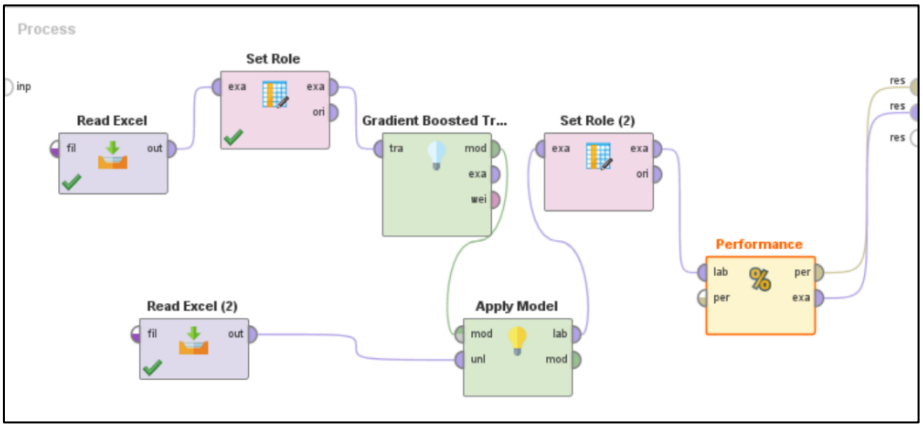
Random Sampling	***Criterion***	***Accuracy (%)***	***Weighted Mean Recall (%)***	***Weighted Mean Presisi (%)***
Linear	*Gain Ratio*	**90.00**	**91.00**	**92.00**
	Information Gain	95.00	97.00	96.04
	Gini Index	90.03	95.50	96.00

	Accuracy	94.00	97.30	93.02
Shuffle	*Gain Ratio*	**98.96**	**97.86**	**98.55**
	Information Gain	98.00	95.00	96.00
	Gini Index	96.00	98.00	95.00
	Accuracy	97.04	96.50	95.02
Stratified	*Gain Ratio*	98.00	95.00	96.00
	Information Gain	95.40	97.30	93.25
	Gini Index	93.00	92.00	90.00
	Accuracy	97.00	93.00	91.00

Dari tabel.3 dapat terlihat bahwa algoritma Random Forest memiliki kinerja terbaik ketika menggunakan teknik *shuffle sampling* dengan kriteria *gain ratio.* Dan Algoritma *random forest* memiliki hasil kurang bagus ketika menggunakan *linear sampling* dengan kriteria *Gini Index*.

3.3 Pengujian Model Algoritma Gradient Boosting

Adapun uji coba model algoritma Gradient Boosting dengan menggunakan tipe data *polynominal* dan *integer* dataset komoditas menggunakan software RapidMiner dilakukan dengan tahapan seperti pada gambar 5.



Gambar 5 Hasil pengujian model algoritma *gradient boosting*

Dari beberapa kali hasil pengujian model klasifikasi algoritma *Gradient Boosting* dengan teknik random sampling model linear, shuffle, dan stratified seperti yang telah dilakukan pengujian pada algoritma C4.5 dan Random Forest maka didapat hasil seperti yang tertera pada tabel 4.

Tabel 4 Hasil uji coba algoritma *gradient boosting*

Random Sampling	***Accuracy (%)***	***Weighted Mean Recall (%)***	***Weighted Mean Presisi (%)***
Linear	**75.00**	**91.67**	**91.67**
Shuffle	**91.00**	**79 .04**	**81.42**
Stratified	83.33	91.67	91.67

Berdasarkan hasil uji coba pengukuran dengan menggunakan *random sampling*, *linear*, *shuffle*, dan *starified* menggunakan dataset komoditas. Algoritma *gradient boosting* memiliki kinerja terbaik ketika menggunakan *shuffle sampling*. Dan Algoritma *gradient boosting* memiliki hasil kurang bagus ketika menggunakan *linear sampling*.

3.4 Perbandingan algoritma untuk klasifikasi data komoditas

Berdasarkan hasil pengujian algoritma C4.5, *random forest*, dan *gradient boosting* dapat dibandingkan berdasarkan tabel 5.

Tabel 5 Perbandingan algoritma C4.5, *random fores,* dan *gradient boosting*

Algoritma	Perbandingan kinerja terbaik dan kurang baik	Acuraccy (%)	*Weighted Mean Recall (%)*	*Weighted Mean Presisi (%)*
C4.5	*shuffle sampling (gain ratio)*	90.00	90.50	92.00
	linear sampling (accuracy)	86.67	85.70	88.71
Random Forest	*shuffle sampling (gain ratio)*	**98.96**	**97.86**	**98.55**
	linear sampling (Gini Index)	90.00	91.00	92.00
Gradient Boosting	*shuffle sampling*	91.00	79.04	81.42
	linear sampling	75.00	91.67	91.67

Berdasarkan semua uji coba algoritma didapatkan kesimpulan bahwa algoritma yang bagus untuk karakteristik data klasifikasi pada penelitian ini adalah algoritma *Random Forest* dengan menggunakan *shuffle sampling (gain ratio).*

4 KESIMPULAN DAN SARAN

Penelitian ini menghasilkan kesimpulan bahwa algoritma decision tree yang memiliki kinerja terbaik dalam melakukan klasifikasi adalah algoritma *random forest* dengan syarat menggunakan *shuffle sampling*. Mayoritas *linear sampling* menghasilkan kinerja kurang baik. Sedangkan *shuffle sampling* memiliki kinerja bagus untuk algoritma berbasis tree.

Penerapan *cretarion gain ratio* pada C4.5 dan *random forest* memiliki kinerja sangat baik, berbading terbalik ketika menggunakan *cretarion accuracy*. Kinerja bagus didasarkan pada hasil evaluasi *accuracy*, *weighted mean precision*, dan *weighted mean recall. Shuffled sampling* menggambarkan distribusi antar kelas seimbang baik pada data *training* maupun data *testing*, begitu sebaliknya dengan *stratified sampling*.

Berdasarkan hasil penelitian algoritma berbasis tree yang sangat baik untuk digunakan dalam melakukan klasifikasi data komoditas yaitu algoritma *random forest*. Sedangkan algoritma C4.5 masih kurang baik dalam melakukan klasifikasi data komoditas.

Penelitian selanjutnya diharapkan dapat menganalisis secara detail terkait pengaruh parameter data yang digunakan, sehingga dapat mempengaruhi hasil *accuracy, recall, dan presisi* dan bisa menambahkan perbandingan dengan beberapa algoritma klasifikasi yang lainnya.

DAFTAR PUSTAKA

[1] S. Agarwal, *Data mining: Data mining concepts and techniques*. 2014.

[2] S. Sularno and P. Anggraini, "Penerapan Algoritma C4.5 Untuk Klasifikasi Tingkat Keganasan Hama Pada Tanaman Padi," *Jurnal Sains dan Informatika*, vol. 3, no. 2, p. 161, 2017, doi: 10.22216/jsi.v3i2.2779.

[3] A. Asroni, B. Masajeng Respati, and S. Riyadi, "Penerapan Algoritma C4.5 untuk Klasifikasi Jenis Pekerjaan Alumni di Universitas Muhammadiyah Yogyakarta," *Semesta Teknika*, vol. 21, no. 2, pp. 158–165, 2018, doi: 10.18196/st.212222.

[4] E. P. Cynthia and E. Ismanto, "Metode Decision Tree Algoritma C.45 Dalam Mengklasifikasi Data Penjualan," *Jurnal Riset Sistem Informasi Dan Teknik Informatika (JURASI)*, vol. (3) Juli, no. July, pp. 1–13, 2018, [Online]. Available: http://tunasbangsa.ac.id/ejurnal/index.php/jurasik/article/download/60/pdf.

[5] E. Sugiatna, A. M. Ibrahim, and I. Abdul Hadi, "Implementasi Algoritma Klasifikasi C4.5 Untuk Memprediksi Kelayakan Pembelian Kendaraan," *JTIM : Jurnal Teknologi Informasi dan Multimedia*, vol. 1, no. 2, pp. 124–132, 2019, doi: 10.35746/jtim.v1i2.26.

[6] N. Nanni and A. Sudransyah, "Perbandingan Kinerja Algoritma Naive Bayes dan C4.5 Untuk Klasifikasi Harga Pangan," *PROtek : Jurnal Ilmiah Teknik Elektro*, vol. 7, no. 1, pp. 20–24, 2020, doi: 10.33387/protk.v7i1.1710.

[7] A. Yuliana and D. B. Pratomo, "Memprediksi Kepuasan Mahasiswa Terhadap Kinerja Dosen Politeknik TEDC Bandung," *Semnasinotek 2017*, pp. 377–384, 2017.

[8] G. Lukhayu Pritalia, "Penerapan Algoritma C4.5 untuk Penentuan Ketersediaan Barang E-commerce," *Indonesian Journal of Information Systems*, vol. 1, no. 1, pp. 47–56, 2018, doi: 10.24002/ijis.v1i1.1727.

[9] U. Khultsum and A. Subekti, "Penerapan Algoritma Random Forest dengan Kombinasi Ekstraksi Fitur Untuk Klasifikasi Penyakit Daun Tomat," *Jurnal Media Informatika*, vol. 5, pp. 186–193, 2021, doi: 10.30865/mib.v5i1.2624.

[10] S. Devella, Y. Yohannes, and F. N. Rahmawati, "Implementasi Random Forest Untuk Klasifikasi Motif Songket Palembang Berdasarkan SIFT," *JATISI (Jurnal Teknik Informatika dan Sistem Informasi)*, vol. 7, no. 2, pp. 310–320, 2020, doi: 10.35957/jatisi.v7i2.289.

[11] D. S. S. Wuisan, "Penerapan Algoritma Klasifikasi Random Forest Untuk Penentuan Kelayakan Pemberian Kredit di Koperasi Mitra Sejahtera," *Journal Of Technology Information*, vol. 6, no. 1, pp. 29–34, 2020, [Online]. Available: https://ojs.uajy.ac.id/index.php/IJIS/article/view/1704/1195.

[12] T. Online, S. Kasus, D. Acak, P. Awal, and M. Pandemic, "Prediksi Harga Bitcoin Menggunakan Metode Random Forest," *Jurnal Komputer Terapan*, vol. 7, no. 1, pp. 24–32, 2021.

[13] V. W. Siburian and I. E. Mulyana, "Prediksi Harga Ponsel Menggunakan Metode Random Forest," *Prosiding Annual Research Seminar*, vol. 4, no. 1, pp. 144–147, 2018.

[14] A. Primajaya and B. N. Sari, "Random Forest Algorithm for Prediction of Precipitation," *Indonesian Journal of Artificial Intelligence and Data Mining*, vol. 1, no. 1, p. 27, 2018, doi: 10.24014/ijaidm.v1i1.4903.

[15] I. Muslim, K. Karo, F. Informatika, and U. Telkom, "Implementasi Metode XGBoost dan Feature Importance untuk Klasifikasi pada Kebakaran Hutan dan Lahan," *Journal of Software Engineering, Information and …*, vol. 1, no. 1, pp. 10–16, 2020, [Online]. Available: https://ejournal.upi.edu/index.php/SEICT/article/view/29347.

[16] A. Bisri and R. Rachmatika, "Integrasi Gradient Boosted Trees dengan SMOTE dan Bagging untuk Deteksi Kelulusan Mahasiswa," *Jurnal Nasional Teknik Elektro dan Teknologi Informasi (JNTETI)*, vol. 8, no. 4, p. 309, 2019, doi: 10.22146/jnteti.v8i4.529.

[17] R. Sun, G. Wang, W. Zhang, L. T. Hsu, and W. Y. Ochieng, "A gradient boosting decision tree based GPS signal reception classification algorithm," *Applied Soft Computing Journal*, vol. 86, no. xxxx, p. 105942, 2020, doi: 10.1016/j.asoc.2019.105942.

[18] N. Nyoman, P. Pinata, I. M. Sukarsa, N. Kadek, and D. Rus jayanthi, "Prediksi Kecelakaan Lalu Lintas di Bali dengan XGBoost pada Python," *Jurnal Ilmiah Merpati*, vol. 8, no. 3, pp. 188–196, 2020.

[19] S. R. Putri, T. Informatika, U. Pamulang, T. Selatan-indonesia, and G. B. Trees, "Teknik Bagging untuk Mengurangi Kesalahan Klasifikasi Algoritma Gradient Boosted Tress (GBT) Pada Prediksi Kanker Payudara," *Prosiding Seminar Nasional Informatika dan Sistem Informasi*, vol. 3, pp. 566–572, 2020.

[20] S. Agarwal, *Data mining: Data mining concepts and techniques*. 2014.

[21] M. Bramer, *Introduction to Data Mining*. 2016.

[22] V. Kotu and B. Deshpande, *Predictive Analytics and Data Mining: Concepts and Practice with RapidMiner*. 2014.