Perbandingan Kinerja Tool Data Mining Weka dan Rapidminer Dalam Algoritma Klasifikasi

Mochammad Faid
Program Studi Teknik Informatika
Universitas Nurul Jadid
mfaid@unuja.ac.id

Moh. Jasri
Program Studi Teknik Informatika
Universitas Nurul Jadid
jasri@unuja.ac.id

Titasari Rahmawati
Program Studi Sistem Informasi
Institut Informatika Indonesia
tita@ikado.ac.id

Abstrak – *Data mining* merupakan ilmu yang membahas tentang bagaimana menambang pengetahuan dari sebuah data. Klasifikasi merupakan salah satu bagian dari *data mining*. Algoritma klasifikasi dalam *data mining* bermacam-macam model. Karena setiap model yang ada di algoritma klasifikasi tidak sama, maka akurasinya tentu akan berubah. Untuk mengetahui baik tidaknya sebuah algoritma klasifikasi, indikatornya adalah tingkat akurasi. Dengan perhitungan-perhitungan yang rumit dan membutuhkan waktu yang sangat lama, diciptakan sebuah *tools data mining* sehingga proses dan pengolahan *data mining* lebih mudah. *Tools data mining* dalam penelitian ini menggunakan Weka dan Rapidminer. Adapun tujuan dari penelitian ini adalah untuk mengetahui kinerja dari *tools data mining* Weka dan Rapidminer.

Kata Kunci: Klasifikasi, Rapidminer, Weka.

I. PENDAHULUAN

Dalam kaidah keilmuan, fakta dikumpulkan untuk mendapatkan sebuah data. Data kemudian diolah sehingga dapat dideskripsikan secara jelas dan tepat sehingga dapat dimengerti oleh orang lain yang tidak langsung mengalaminya sendiri. Pemilihan banyak data sesuai dengan persamaan atau perbedaan yang dikandungnya dinamakan klasifikasi. Proses pengolahan data menjadi sebuah informasi ini sangat berguna sebagai dasar pengambilan keputusan yang tepat. Namun seiring perkembangan zaman, data yang telah terkumpul sekian tahun lamanya dimana hampir semua data tersebut dimasukkan dengan menggunakan aplikasi komputer sehingga menumpuk seperti gunung dan tidak berguna lalu dibuang. Ternyata data yang sudah dianggap kadaluarsa dan tak berguna masih bisa diperas/diekstraksi lagi menjadi sebuah pengetahuan baru yang sangat bermanfaat bagi organisasi yang menggunakannya. Ilmu yang membahas tentang ekstraksi/penambangan data disebut *data mining*. *Data mining* mengeksplorasi basis data untuk menemukan pola-pola yang tersembunyi, mencari informasi guna memprediksi yang mungkin saja terlupakan oleh pelaku yang data tersebut diluar ekspektasi mereka. Dengan ditemukan ilmu *data mining*, bermunculanlah algoritma-algoritma *data mining* yang memiliki kekurangan dan kelebihan. Untuk mengetahui kinerja dari sebuah algoritma *data mining* dibuatlah *tool data mining*. Dalam penelitian ini mencoba untuk melihat kinerja algoritma *data mining* dengan menggunakan *tool data mining*. Adapun *tool data mining* yang akan digunakan adalah Weka dan Rapidminer.

Berdasarkan penelitian yang dilakukan oleh Siti Masripah dengan judul penelitiannya "Komparasi Algoritma Klasifikasi *Data Mining* untuk Evaluasi Pemberian Kredit" dilakukan dengan menggunakan *tool Data Mining* yaitu Rapidminer. Penelitian ini mencoba mengkomparasikan dua algoritma klasifikasi yaitu C4.5 dan Naive Bayes. Setelah komparasi dapat disimpulkan bahwa C4.5 memprediksi lebih akurat dari pada Naive Bayes [1]. Pada penelitian selanjutnya yang dilakukan oleh Vinita Chandani dengan judul penelitiannya yaitu "Komparasi Algoritma Klasifikasi *Machine Learning* dan *Feature Selection* Pada Analisis Sentimen *Review* Film". Adapun algoritma yang komparasikan ada 3 yaitu ANN, SVM, dan NB. Dari hasil uji coba didapatkan kesimpulan bahwa algoritma SVM memiliki kinerja terbaik dengan nilai akurasi sebesar 81,10% [2]. Kemudian penelitian selanjutnya dilakukan oleh Rizal Amegia Saputra dengan judul penelitiannya "Komparasi Algoritma Klasifikasi *Data Mining* Untuk Memprediksi Penyakit Tuberculosis (TB) Studi Kasus Puskesmas Karawang Sukabumi". Pada penelitian ini algoritma yang dikomparasikan ada 4 yaitu C4.5, Naive Bayes, Neural Network, dan Logistic Regression. Dari hasil penelitian ini hasil terbaik terdapat pada algoritma Naive Bayes sebesar 91,61%[3]. Dari semua penelitian sebelumnya untuk menghitung akurasi semuanya menggunakan *tool data mining*. Perbedaan penelitian ini dengan penelitian sebelumnya adalah penelitian ini mencoba mengkomparasikan *tool data mining* yang sering digunakan oleh para peneliti yaitu Rapidminer dan Weka.

Tujuan dari penelitian ini untuk mengetahui kinerja dan membandingkan *tool data mining* Weka dan Rapidminer sehingga bisa digunakan untuk keperluan penelitian tentang data mining, dan sama sekali tidak ada niatan untuk menyudutkan keduanya. Karena masing-masing *tool data mining* memiliki kekurangan dan kelebihannya sendiri.

II. STUDI PUSTAKA

A. *Data Mining*

Data Mining adalah suatu istilah yang digunakan untuk menguraikan penemuan pengetahuan di dalam *database*. *Data mining* merupaka proses yang menggunakan teknik statistik, Matematika, kecerdasan buatan, dan *machine*

learning untuk mengektraksi dan mengindentifikasi informasi yang bermanfaat dan pengetahuan yang terakit dari berbagai *database* besar [4]. Sedang menurut Gatner Grup, *data mining* adalah suatu proses menemuan hubungan yang berarti, pola dan kecenderungan dengan memeriksa dalam sekumpulan besar data yang tersimpan dalam penyimpanan, dengan menggunakan teknik pengenalan pola seperti teknik Statistik dan Matematika [5]

B. Klasifikasi *Data Mining*

Klasifikasi adalah teknik yang dilakukan untuk memprediksi *class* atau properti dari setiap *instance* data, Proses klasifikasi dilakukan setelah data selesai melewati tahap *preprocessing.* Pada tahapan *preprocessing* dilakukan pengecekan duplikasi data dengan Levenshteain *string metric*, pemilihan fitur, dan penanganan *missing value* [6].

Metode-metode yang ada pada *Data Mining Classification* antara lain adalah:

1. C4.5

C4.5 diperkenalkan Quinlan (1996) sebagai versi terbaru dari ID3. Dalam induksi *tree* hanya bisa dilakukan pada fitur bertipe kategorikal (nominal atau ordinal). Sedangkan tipe numerik (interval atau rasio) tidak dapat digunakan. Perbaikan yang membedakan algoritma C4.5 dan ID3 adalah dapat menangani fitur dengan tipe numerik, melakukan pemotongan *decision tree*. Algoritma C4.5 juga menggunakan kriteria dalam menentukan fitur yang menjadi pemecah pada pohon yang diinduksi [7].

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

Keterangan :
- S : Himpunan Kasus
- A : Atribut
- N : Jumlah partisi atribut A
- |Si| : Jumlah kasus pada partisi ke i
- |S| : Jumlah kasus dalam S

Sedangkan perhitungan nilai *Entropy* dapat dilihat pada persamaan 2-3:

$$\boldsymbol{e}ntropy(A) = -\sum_{i=1}^{n} \frac{|Si|}{|S|} \log_2 \frac{|Si|}{|S|} \qquad (2)$$

Keterangan :
- S : Himpunan kasus
- A : Atribut
- N : Jumlah partisi atribut A
- |Si| : Jumlah kasus pada partisi ke i
- |S| : Jumlah kasus dalam S

2. Naive Bayes

Naive Bayes mendasarkan pada asumsi penyederhanaan dimana nilai atribut secara kondisional saling bebas apabila diberikan nilai *output*. Metode ini merupakan sebuah metode yang berakar pada teorema *Bayes*. Persamaan (3) merupakan persamaan Teorema *Bayes* yang menyatakan bahwa:

$$P(B|A) = \frac{P(A|B)P(B)}{P(B)} \qquad (3)$$

Keterangan:
- P(B|A) : Probabilitas *posterior*, probabilitas muncul B jika diketahui A
- P(B|A) : Probabilitas *posterior*, probabilitas muncul A jika diketahui B
- P(A) : Probabilitas *prior*, probabilitas kejadian A
- P(B) : Probabilitas *prior*, probabilitas kejadian B

3. K-NN

K-*Nearest Neighbor* merupakan sebuah metode untuk melakukan klasifikasi terhadap objek berdasarkan data pembelajaran yang jaraknya paling dekat dengan objek tersebut. Data pembelajaran diproyeksikan ke ruang berdimensi banyak, dimana masing-masing dimensi merepresentasikan fitur dari data. Ruang ini dibagi menjadi bagian-bagian berdasarkan klasifikasi data pembelajaran. Nilai k yang terbaik untuk algoritma ini tergantung pada data. Secara umumnya, nilai k yang tinggi akan mengurangi efek *noise* pada klasifikasi, tetapi membuat batasan antara setiap klasifikasi menjadi lebih kabur. Ada banyak cara untuk mengukur jarak kedekatan antara data baru dengan data lama (*data training*), diantaranya euclidean *distance* dan manhattan *distance* (*city block distance*), yang paling sering digunakan adalah *euclidean distance* [8], yaitu:

$$R = \sqrt{(a_1 - b_1)^2 + (a_2 - b_2)^2 + \cdots + (a_n - b_n)^2} \qquad (4)$$

Keterangan:
- R : Jarak data
- a_1 : Atribut pertama data *testing*
- b_1 : Atribut pertama data *training*
- a_2 : Atribut kedua data *testing*
- b_2 : Atribut kedua data *training*
- a_n : Atribut ke-n data *training*
- b_n : Atribut ke-n data *training*

4. Random Forest

Random forest adalah sekumpulan *classifier* yang terdiri dari banyak pohon keputusan dan melakukan klasifikasi berdasarkan keluaran dari hasil klasifikasi setiap pohon keputusan anggota. Metode *random forest* adalah pengembangan dari metode CART, yaitu dengan menerapkan metode *bootstrap aggregating* (*bagging*) dan *random feature selection* [9]. Dalam *random forest*, banyak pohon ditumbuhkan sehingga terbentuk hutan (*forest*), kemudian analisis dilakukan pada kumpulan pohon tersebut. Pada gugus data yang terdiri atas n amatan dan peubah penjelas, *random forest* dilakukan dengan cara [10]:

1. Lakukan penarikan contoh acak berukuran n dengan pemulihan pada gugus data. Tahapan ini merupakan tahapan *bootstrap*.
2. Dengan menggunakan contoh *bootstrap*, pohon dibangun sampai mencapai ukuran maksimum (tanpa pemangkasan). Pada setiap simpul, pemilihan pemilah dilakukan dengan memilih m peubah penjelas secara acak, dimana m << p. Pemilah terbaik dipilih dari m peubah penjelas tersebut. Tahapan ini adalah tahapan *random feature selection*.
3. Ulangi langkah 1 dan 2 sebanyak k kali, sehingga terbentuk sebuah hutan yang terdiri atas k pohon.

C. *Tool Data Mining*

Tool data mining adalah *software* yang digunakan untuk mempermudah seorang peneliti, akademis, dan pihak manapun dalam hal mengolah dan menambang sebuah data, Selain itu juga bisa dijadikan sebuah *library* sehingga bisa ditanam dalam sebuah program, sehingga program bisa melakukan apa yang bisa *tool* lakukan.

1. Weka

Weka merupakan rangkaian perangkat lunak pembelajaran mesin yang ditulis dalam bahasa Java, dikembangkan di Universitas Waikato, Selandia Baru. Perangkat lunak ini memiliki banyak algoritma *machine learning* untuk keperluan *data mining*. Weka juga memiliki banyak *tool* untuk pengolahan data, mulai dari *pre-processing*, *classification*, *association rules*, dan *visualization*.

2. Rapidminer

Rapidminer adalah salah satu *software* untuk pengolahan *data mining*. Pekerjaan yang dilakukan oleh rapidminer *text mining* adalah berkisar dengan analisis teks, mengekstrak pola-pola dari *dataset* yang besar dan mengkombinasikannya dengan metode statistika, kecerdasan buatan, dan *database*.

3. *Confusion Matrix*

Confusion adalah suatu metode yang biasanya digunakan untuk melakukan perhitungan akurasi pada konsep *data mining* [11]. Dengan menggunakan *split ratio* 0.1 serta data yang digunakan adalah data *playing* tenis kita akan coba membandingkan *confusion matrix* antara Weka dan Rapidminer. Berikut ini adalah *confusion Matrix* Weka dan Rapidminer.

```
accuracy 69.2308 %
=== Confusion Matrix ===

 a b   <-- classified as
 0 4 | a = No
 0 9 | b = Yes
```

Gambar 1. *Confusion Matrix* Weka.

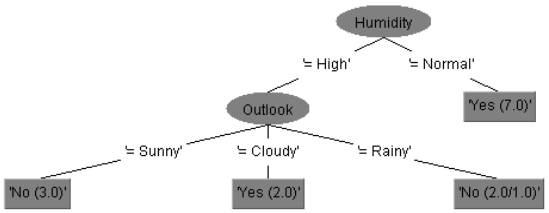
accuracy: 69.23%

	true No	true Yes
pred. No	0	0
pred. Yes	4	9
class recall	0.00%	100.00%

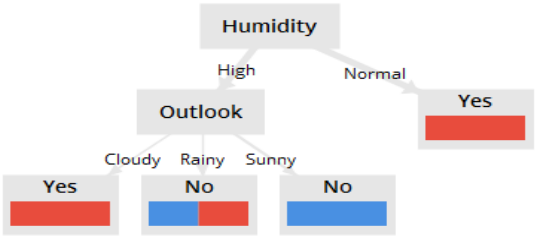
Gambar 2. *Confusion Matrix* Rapidminer.

4. *Decision Tree*

Decision Tree merupakan pohon keputusan yang dihasilkan dari hasil Ekstraksi sebuah data dengan menggunakan algoritma klasifikasi.



Gambar 3. *Decision Tree* Weka.



Gambar 4. *Decision Tree* Rapidminer.

III. METODOLOGI PENELITIAN

Instrumen-instrumen yang digunakan dalam penelitian ini dibagi menjadi dua bagian. Yang pertama adalah instrumen pengumpulan data dan instrumen penelitian.

1. Instrumen Pengumpulan Data

Dalam tahap ini dilakukan pengumpulan data sebagai bahan untuk menganalisis kinerja algoritma *data mining* Sistem yang bisa digunakan untuk *data testing* pada *tool data mining* yaitu Weka dan Rapidminer. Adapun *dataset* yang digunakan adalah sebagai berikut:

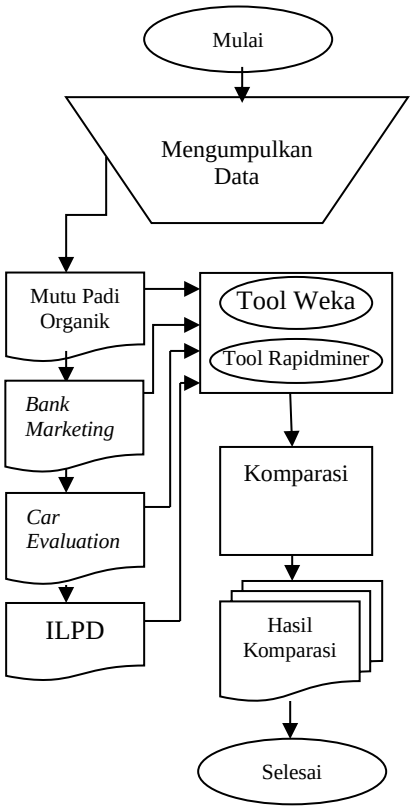
- Mutu Padi Organik
- *Bank Marketing*
- *Car Evaluation*
- *Indian Liver Patient Dataset (ILPD)*

2. Instrumen Penelitian

Untuk mendesain sebuah sistem tentunya kita juga membutuhkan instrumen diantaranya adalah *software* yang mendukung untuk menganalisis dan mengkomparasi algoritma klasifikasi:

- *Tool Data Mining* (Weka, Rapidminer)
- Excel
- *Timer*

Adapun tahapan penelitian bisa dilihat seperti *Flowchart* dibawah ini.

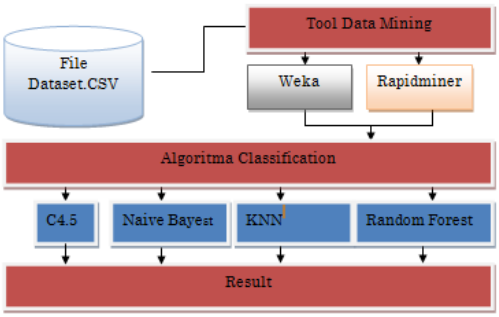


Gambar 5. *Flowchart* Tahapan Penelitian.

IV. HASIL DAN PEMBAHASAN

A. Prosedur Penelitian

Procedur penelitian ini adalah mengumpulkan 2 tools *data mining* yaitu Weka dan Rapidminer serta menentukan kumpulan data yang akan digunakan, dan memilih satu set algoritma klasifikasi untuk menguji kinerja *tools data mining.* Kemudian mencatat hasil akurasi dari masing masing *tools data mining* dalam bentuk tabel dan grafik



Gambar 6. Skema Penelitian

Hasil percobaan data dalam penelitian ini menggunakan data padi organik, *Indian Liver Patient Dataset* (ILPD), *Bank Marketing*, dan *Car Evaluation*. Untuk atribut yang digunakan dalam *dataset* nampak pada Tabel 1. Sedangkan untuk hasil akurasinya dan kecepatan pemrosesannya nampak pada Tabel 2. Untuk mengetahui kinerja terbaik dari sebuah *tools data mining* maka dalam penelitian ini akan membuat rata-rata selisih dari masing-masing algoritma klasifikasi. Adapun *detail* dari performa bisa dilihat pada Tabel 3.

Data yang dijadikan *testing* diantaranya adalah mutu padi organik Dinas Pertanian Bondowoso, *bank marketing, car evaluation*, dan *Indian Liver Patient Dataset* (ILPD). *Dataset* tersebut dipilih karena sesuai dengan algoritma yang digunakan diantaranya C4.5/J48, Naive Bayes, K-NN, serta *Random Forest*. Adapun Gambar 7-10 merupakan representasi dari Tabel 2.

Tabel 1**.** Atribut Yang Ada Pada *Dataset*.

Dataset	**Atribut**	**Sumber**
Mutu Padi Organik	*Varietas* *Panjang* *Bentuk* *Warna* *Rasa* *Teknik* *Musim* *Hama* *Ph*	*Dinas Pertanian Bondowoso*
Bank Marketing	*Age* *Job* *Marital* *Education* *Default* *Balance* *Housing* *Loan* *Contact* *Day* *Month* *Duration* *Campaign* *Pdays* *Previous* *Poutcome*	*archive.ics.uci.edu*
Car Evaluation	*Buying* *Maint* *Doors* *Persons* *Ug_boot* *Safety* *Varian*	*archive.ics.uci.edu*

DOI: 10.34148/teknika.v8i1.95

Indian Liver Patient Dataset (ILPD)	*Age* *Gender* *Tb_total_bilirubin* *Db_direct_bilirubin* *Alkphos_alkaline_phosphotase* *Sgpt_alamine_aminotransferase* *Sgot_aspartate_aminotransferase* *Tp_total_protiens* *Alb_albumin* *A/g_ratio_albumin_and_globulin ratio* *Patient*	*archive.ics.uci.edu*

Tabel 2. Detail Performa *Tool Data Mining*.

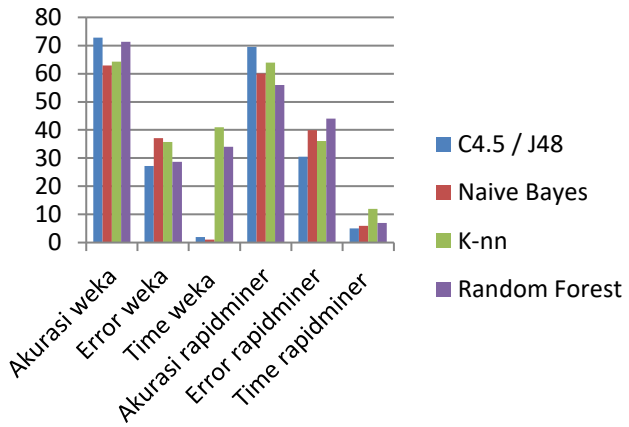
Dataset	Algoritma *Classfication*	Weka			Rapidminer		
		Akurasi	*Error*	*Time*	Akurasi	*Error*	*Time*
Mutu Padi Organik	C4.5 / J48	72,79	27,21	2	69,47	30,53	5
	Naive Bayes	62,90	37,10	1	60,06	39,94	6
	K-nn	64,29	35,71	41	63,94	36,06	12
	Random Forest	71,27	28,73	34	55,95	44,05	7
Bank Marketing	*C4.5 / J48*	87,81	12,19	3	88,28	11,72	3
	Naive Bayes	88,82	11,18	1	87,44	12,56	4
	K-nn	87,05	12,95	60	85,82	14,18	8
	Random Forest	89,01	10,99	19	88,47	11,53	15
Car Evaluation	*C4.5 / J48*	80,51	19,49	1	79,36	20,64	5
	Naive Bayes	79,04	20,96	1	80,26	19,74	3
	K-nn	77,75	22,25	4	76,78	23,22	3
	Random Forest	83,15	16,85	8	70,68	29,32	4
Indian Liver Patient Dataset (ILPD)	*C4.5 / J48*	67,62	32,38	1	64,12	35,88	6
	Naive Bayes	59,24	40,76	1	56,68	43,32	4
	K-nn	64,38	35,62	3	67,94	32,06	5
	Random Forest	70,86	29,14	3	69,66	30,34	3
Rata Rata Akurasi dan Waktu		75,41		11,44	72,80688		5,8125

Tabel3. Rata Akurasi dan Waktu penggunaan *Tool Data Mining*.

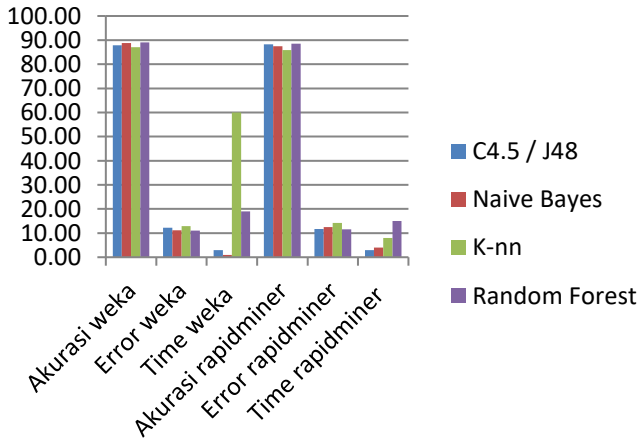
Weka		Rapidminer	
Akurasi	*Time*	Akurasi	*Time*
75,41	11,44	72,80688	5,8125

Jika dibandingkan antara Rapidminer dan Weka dari Tabel 2 bisa dilihat hasil rata-ratanya. Rata-rata kecepatan Rapidminer 5,8125 detik dan Weka 11,44 detik. Selisih rata-rata kecepatan antara Weka dan Rapidminer adalah 5,625 dan dilihat dari tingkat akurasi algoritma klasifikasinya, Weka lebih unggul dari *Rapidminer* dengan selisih rata-rata sebesar 2,60 %.

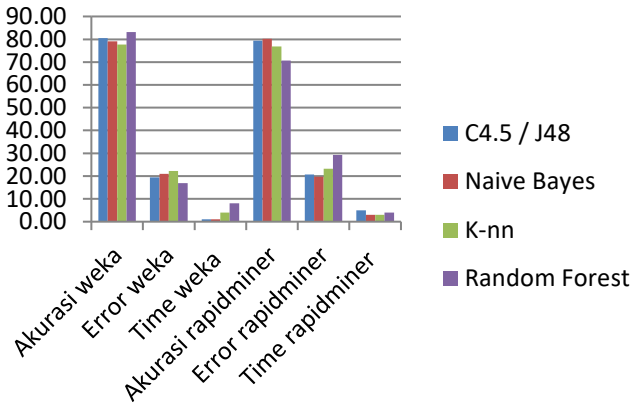
Tabel 2 kemudian direpresentasikan dalam bentuk grafik untuk memudahkan membaca kinerja *tool data mining* serta algoritmanya. Hal ini nampak pada Gambar 7-10 sebagai berikut



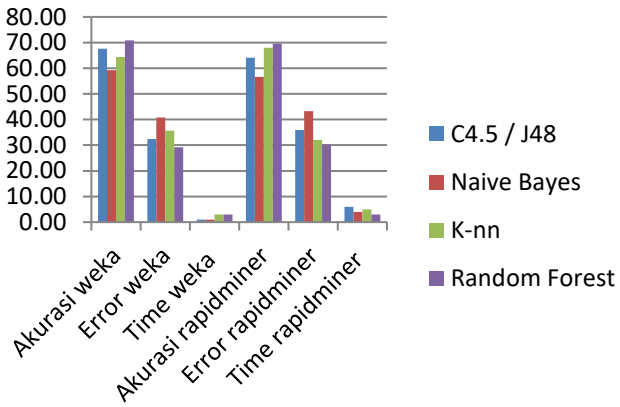
Gambar 7. Grafik Performa Dengan *Dataset* Data Mutu Padi Organik.



Gambar 8. Grafik Performa Dengan *Dataset Bank Marketing*.



Gambar 9. Grafik Performa Dengan *Dataset Car Evaluation.*



Gambar 10. Grafik Performa Dengan *Dataset Indian Liver Patient Dataset (ILPD).*

V. KESIMPULAN

Mengenai kesimpulan dari hasil penelitian dalam hal kecepatan pemprosesan data dari algoritma klasifikasi *Tool Data mining Rapidminer* memiliki kecepatan yang lebih unggul dari pada *Tool Data Mining Weka*, Sedangkan disisi lain Tingkat akurasi *Tool* yang lebih Unggul adalah *Weka* ketimbang *Rapidminer*, Dari dataset yang digunakan *dataset Indian Liver Patient Dataset (ILPD)* yang memiliki waktu pemprosesan tercepat dari yang lain

REFERENSI

[1] Masripah, S. (2016). Komparasi Algoritma Klasifikasi Data Mining untuk Evaluasi Pemberian Kredit. *Bina Insani ICT Journal*, Vol. 3, No. 1, pp. 187-193.

[2] Chandani, V., Wahono, R.S. & Purwanto. (2015). Komparasi Algoritma Klasifikasi Machine Learning dan Feature Selection pada Analisis Sentimen Review Film. *Journal of Intelligent Systems*, Vol. 1, No. 1, pp. 56-60.

[3] Amegia, R. (2014). Komparasi Algoritma Klasifikasi Data Mining untuk memprediksi Penyakit Tuberculosis (TB) Studi Kasus Puskesmas Karawang Sukabumi. *Proceedings SNIT 2014*.

[4] Turban, E., dkk., (2005). *Decision Support Systems and Intelligent Systems*. Yogyakarta: Andi Offset.

[5] Furnkranz, J. (1994). A Comparison of Pruning Methods for Relational Concept Learning. Austria: AAAi.

[6] Han, J. & Kamber, M. (2006). *Data Mining Concept and Tehniques*. San Fransisco: Morgan Kauffman.

[7] Nithya, A. & Sundaram, V. (2011). Classification Rules for Indian Race Diseases, *IJCSI*.

[8] Bramer, M. (2007). *Principles of Data Mining*. London: Springer.

[9] Breiman, L., (2001). *Random Forests*. California: University of California Berkeley.

[10] Larose, D.T. (2005). Discovering Knowledge in Data. Canada. Wiley-Interscience.

[11] Berry & Linoff (2004). *Data Mining Techniques for Marketing, Sales and CRM*. Wiley.