Pengaruh Komposisi Data Training dan Testing terhadap Akurasi Algoritma C4.5

Wilem Musu[1], Abdul Ibrahim[2], Heriadi[3]
[1,3]Program Studi Teknik Informatika, Universitas Dipa Makassar
[2]Program Studi Manajemen Informatika, Universitas Dipa Makassar
[1,2,3]Jl. Perintis Kemerdekaan Km. 9 Makassar, Tlp. 0411-587194
e-mail: [1]wilem.musu@dipanegara.ac.id, [2]abdulibrahim@dipanegara.ac.id, [3]heriadi@dipanegara.ac.id

Abstrak

Akurasi adalah tolak ukur yang digunakan untuk mengetahui seberapa tepat suatu pola klasifikasi memprediksi kelas data dari data yang akan datang. Dalam praktek data mining pengujian akurasi dari sebuah pola klasifikasi menggunakan data testing, sementara untuk menemukan pola itu sendiri, menggunakan data training. Pembagian presentasi jumlah data training dan data testing dari sebuah dataset menjadi salah satu faktor penentu besaran nilai akurasi. Sehingga kesalahan menentukan komposisi antara kedua jenis data tersebut akan mempengaruhi nilai akurasi yang diperoleh. Penelitian ini menguji tingkat akurasi empat dataset menggunakan algoritma C4.5 dengan sembilan komposisi jumlah data training dan testing. Hasil yang diperoleh menunjukan bahwa untuk dataset dengan tingkat sebaran data yang baik, maka komposisi jumlah data training dan testing tidak akan memberikan nilai akurasi yang fluktuatif. Tetapi kondisi sebaliknya memperlihatkan nilai akurasi yang fluktuatif sehingga diperlukan pengujian untuk semua komposisi untuk menemukan akurasi yang maksimum.

Kata kunci: *Data training, Data testing, Akurasi, Komposisi.*

Abstract

The accuracy is a measure used to find out how precisely a classification pattern predicts the class of data from future data. In data mining practice, testing the accuracy of a classification pattern uses testing data, while finding the pattern itself uses training data. Sharing the presentation of the amount of training data and testing data from a dataset is one of the determining factors for the accuracy value. So that errors in determining the composition between the two types of data will affect the accuracy value obtained. This study tested the level of accuracy of four datasets using the C4.5 algorithms with nine compositions of the amount of training and testing data. The results obtained show that for a dataset with a good data distribution level, the composition of the amount of training and testing data will not provide a fluctuating accuracy value. But the opposite condition shows fluctuating accuracy values so that testing for all compositions is needed to find the maximum accuracy.

Keywords*: Data training, Data testing, Accuracy, Composition.*

1. Introduction

Algoritma C4.5 merupakan algoritma klasifikasi yang banyak digunakan untuk menemukan pengetahuan atau pola dari sekumpulan data yang telah terjadi. Dalam melakukan proses klasifikasi, algoritma ini membentuk pohon keputusan yang merupakan sebuah struktur yang berfungsi menggolongkan data sesuai aturan pengelompokan sampai pada bagian terkecil [1]. Ditinjau dari sudut pandang machine learning, C4.5 tergolong algoritma supervised learning, yaitu teknik atau cara yang digunakan komputer untuk mengenali pola dari dataset menggunakan sebuah atribut sebagai pengenal/pengajar terhadap karakteristik data, dimana atribut tersebut pada umumnya diberi nama label atau kelas [2]. Selain memiliki atribut label/kelas, karakteristik utama dari supervised learning adalah wajib memiliki dua jenis data, yaitu data training dan data testing.

Data training merupakan sekumpulan data yang memiliki atribut label/kelas yang digunakan oleh mesin untuk mengenal karaktersitik kumpulan data sehingga menghasilkan sebuah pola/model data. Sementara data testing adalah sekumpulan data yang juga memiliki label/kelas yang digunakan untuk

menguji ketepatan pola/model dalam mengklasifikasikan data testing. Pada saat melakukan proses testing model, atribut label dari data testing disembunyikan selama proses klasifikasi berlangsung dan akan digunakan untuk membandingkan hasil klasifikasi sebagai tolak ukur seberapa besar ketepatan/akurasi model tersebut melakukan klasifikasi.

Dalam praktek penemuan pola dari kumpulan data, seluruh data yang akan dianalisis dikumpulkan menjadi bentuk dataset yang terdiri dari atribut-atribut yang mengandung informasi penting terhadap pola yang hendak di temukan [3]. Dataset tersebut kemudian dibagi menjadi dataset untuk traing dan dataset untuk testing. Hal-hal yang mempengaruhi tingkat akurasi pola klasifikasi yang dihasilkan selain karakteristik dari data yang dimiliki, komposisi atau presentasi pembagian jumlah data untuk training dan testing juga merupakan suatu hal penting yang harus menjadi pertimbangan, secara kusus ketika menggunakan algoritma-algoritma supervised learning. Ketidaktepatan dalam penentuan komposisi jumlah dataset training dan testing akan mempengaruhi akurasi terhadap pola klasifikasi yang di temukan [4].

Penelitian ini bertujuan untuk mengungkapkan bahwa komposisi dataset training dan testing mempengarui tingkat akurasi pola klasifikasi. Untuk membatasi ruang lingkup masalah kami menggunakan algoritma C4.5 sebagai bahan ekperimen. Diharapkan melalui penelitian ini dapat memberikan informasi yang dapat digunakan untuk menemukan komposisi yang tepat terhadap dataset dan testing sehingga tingkat akurasi yang diperoleh dalam setiap analisis menjadi maksimal terutama ketika menggunakan algoritma-algoritma supervised learning.

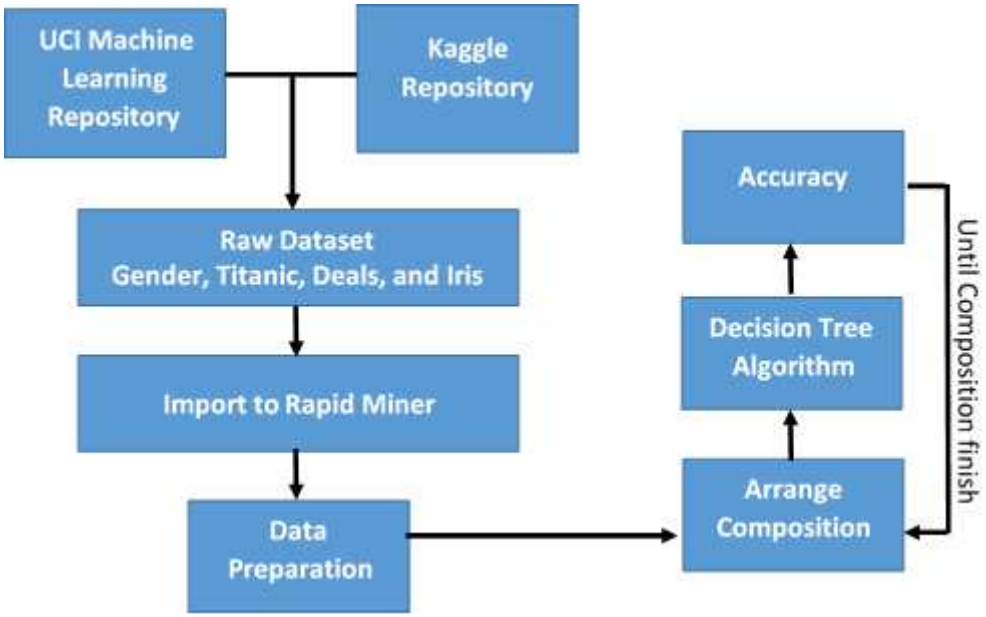
Beberapa penelitian yang telah dilakukan yang berhubungan dengan penelitian ini adalah sebagai berikut: Dalam penelitian M. F. Rahman, dkk. (2017) yang berjudul Klasifikasi Untuk Diagnosa Diabetes Menggunakan Metode Bayesian Regularization Neural Network (RBNN) melakukan ekperimen untuk mengevaluasi komposisi jumlah dataset training dan testing untuk mendiaknosis penyakit diabets. Hasil ekperiment tersebut menunjukan tingkat akurasi bertambah secara linier terhadap komposisi data training yang semakin besar [5]. Selain itu penelitian yang dilakukan oleh Novantiran, dkk. (2015) dengan judul Analisis Sentimen pada Twitter untuk Mengenai Penggunaan Transportasi Umum Darat Dalam Kota dengan Metode Support Vector Machine menyimpulkan bahwa, komposisi jumlah data training dan testing mempengaruhi akurasi dari Support Vector Machine. Walaupun demikian dalam ekperimennya tidak merincikan komposisi secara jelas dalam bentuk presentasi [6]. Iswara, dkk. (2019) melakukan penelitian dengan judul Rekomendasi Pengambilan Mata Kuliah Pilihan Untuk Mahasiswa Sistem Informasi Menggunakan Algoritme Decision Tree, dimana penelitian ini melakukan ekperimen terhadap lima jenis algoritma *Decision Tree*, yaitu ID3, *Random Forest*, CHAID, dan *Rule Induction*. Pada penelitian ini komposisi jumlah data training dan testing berbeda setiap algoritma, dan tidak menguji perubahan komposisi yang sama pada setiap algoritma yang ada [7]. Ketiga penelitian di atas menggunakan satu dataset dengan teknik masing-masing. Pada penelitian ini menggunakan empat dataset dengan perbedaan komposisi yang sama.

2. Research Method

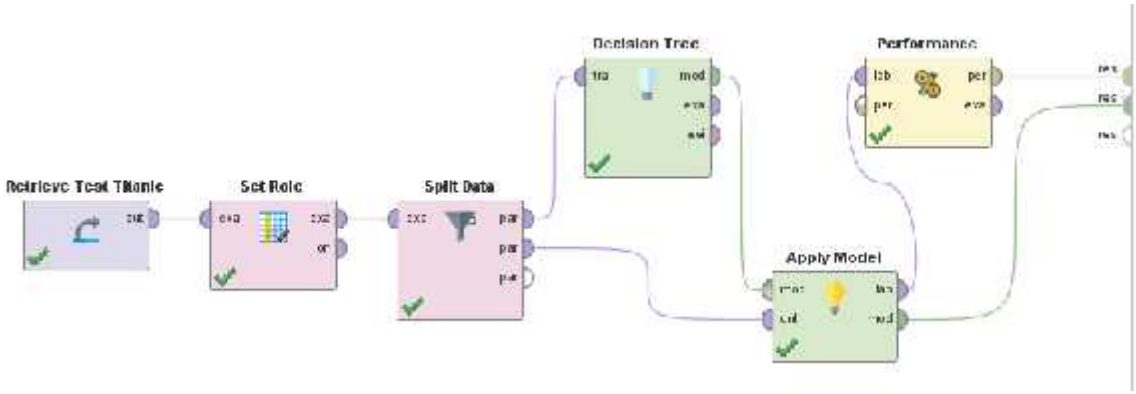
Pada penelitian ini jenis penelitian yang digunakan adalah penelitian ekperimental. Data yang digunakan adalah data yang berasal dari UCI *Machine Learning Repository* dan Kaggle dalam bentuk dataset yang terdiri dari empat dataset, yaitu Gender, Titanic, Deals, dan Iris. Dataset Gender terdiri dari 5001 baris data dengan delapan atribut, dataset Titanic terdiri dari 1024 baris data dengan lima atribut, dataset Deals terdiri dari 1000 baris data dengan empat atribut, sedangkan dataset Iris terdiri 150 baris data dengan lima atribut.

Eksperimen dilakukan menggunakan tool data mining, yaitu Rapid Miner dengan tahapan proses seperti pada Gambar 1. dengan penjelasan sebagai berikut:

1. Download empat dataset dari UCI Machine Learning Repository dan Kaggle.
2. Import dataset ke repository RapidMiner.
3. Memeriksa statistik setiap dataset dan melakukan perbaikan dataset terhadap missing data.
4. Membangun workflow dengan operator-operator yang dapat digunakan untuk memeriksa akurasi dan pengaturan komposisi data training dan testing dengan menggunakan algoritma decision tree. Workflow yang sama diguna untuk semua pengujian, kecuali pada operator Retrieve dataset yang berubah sesuai dataset yang digunakan (Gambar 2.).
5. Menguji komposisi data training dan testing yang dibagi dalam presentasi antara kedua jenis data tersebut. Komposisi-komposisi yang diuji adalah sebagai berikut: training/testing dalam persen; 10/90, 20/80, 30/70, 40/60, 50/50, 60/40, 70/30, 80/20, dan 90/10. Setiap dataset mendapat perlakuan yang sama pada saat pengujian.
6. Hasil pengujian dicatat dan selanjutnya di analisis untuk memperoleh kesimpulan.



Gambar 1. Tahapan Penelitian.



Gambar 2. Workflow Rapid Miner untuk Menguji Akurasi Komposisi Data Training dan Testing.

3. Results and Discussion

Persiapan data (*Data Preparation*) merupakan bagian terpenting dari proses data mining untuk menyiapkan data valid yang akan diolah menggunakan teknik atau metode tertentu. Dataset yang belum bersih dari noise, missing value atau data yang hilang/tidak sesuai, dan data yang terduplikasi akan sangat mempengaruhi hasil dari proses yang dilakukan [8]. Menurut S. K. Kwak and J. H. Kim (2017) yang paling sering terjadi pada dataset yang berkualitas rendah adalah *missing values.* Hal ini dimukninkan terjadi saat melakukan pengumpulan data/pengukuran terjadi kesalahan-kesalahan sehingga ada data yang hilang. Pendekatan statistik merupakan cara yang ampuh untuk memeriksa dan memperbaiki *missing value tersebut* [9]. Oleh karena itu sebelum masuk pada pengujian akurasi, maka kami memaparkan hasil *data preparation* yang dilakukan terhadap empat dataset yang digunakan. Proses persiapan data menggunakan fasilitas *cleansing* dari Rapid Miner dengan operator *Replace Missing Value*.

3.1. The Statistic of Dataset

Dengan pendekatan statistik untuk mempersiapkan empat dataset maka hasil yang diperoleh adalah sebagai berikut:

1. Dataset Gendre, seperti yang terlihat pada Table 1. dataset ini memiliki delapan atribut dengan dengan tiga jenis tipe atribut, yaitu binominal, integer, dan real. Nilai missing setelah perlakuan menunjukan semua atribut terbebas dari missing values. Tipe atribut pada label yaitu binominal dengan dua nilai, yaitu Male dengan jumlah data sebanyak 2500 dan Female sebanyak 2501. Untuk atribut bertipe integer nilai terendah adalah nol dan nilai tertinggi satu. Sementara, untuk atribut Forehead_width_cm yang bertipe real memiliki range 11.4 sampai 15.5 dan untuk atribut Forehead_width_cm juga bertipe real memiliki range nilai 5.1 sampai 7.1.

Tabel 1. Statistik Dataset Genre Hasil Proses Persiapan Data.

Nama Atribut	Tipe Atribut	Missing	Statistik		
			Nilai Terendah	Nilai Tertinggi	Rata-rata
Gender (sebagai label)	Binominal	0	Male	Female	Female (2501), Male (2500)
Long_hair	Integer	0	0	1	0.870
Forehead_width_cm	Real	0	11.4	15.5	13.181
Forehead_height_cm	Real	0	5.1	7.1	5.946
Nose_wide	Integer	0	0	1	0.494
Nose_long	Integer	0	0	1	0.508
Lips_thin	Integer	0	0	1	0.493
Distance_nose_to_lips_long	Integer	0	0	1	0.499

2. Dataset Titanic, pada Table 2. dipaparkan bahwa dataset ini memiliki lima atribut dengan dua jenis tipe atribut, yaitu binominal dan integer. Nilai missing nol pada semua atribut menunjukan semua atribut terbebas dari missing values. Atribut Survived disetting sebagai label dengan tipe binominal yang memiliki dua nilai, yaitu Yes dengan jumlah data sebanyak 614 dan No sebanyak 410. Atribut yang bertipe binominal lainnya adalah Sex dengan nilai Male berjumlah 645 data sementara yang bernilai Female jumlah data sebanyak 379. Untuk atribut bertipe integer, yaitu atribut Age rentang nilai antara 2 sampai 80 dengan rata-rata 30.5, untuk atribut No of Sibling rentang nilai antara 0 sampai 8 dengan rata-rata 0.448, dan atribut terakhir adalah No of Parents memiliki rentang nilai antara 0 sampai 6 dengan rata-rata 0.398.

Table 2. Statistik Dataset Titanic Hasil Proses Persiapan Data.

Nama Atribut	Tipe Atribut	Missing	Statistik		
			Nilai Terendah	Nilai Tertinggi	Rata-rata
Survived (sebagai label)	Binominal	0	Yes	No	No (614), Yes (410)
Sex	Binominal	0	Female	Male	Male (645), Female (379)
Age	Integer	0	2	80	30.5
No of Siblings	Integer	0	0	8	0.448
No of Parents	Integer	0	0	6	0.398

3. Dataset Deals, pada Table 3. dipaparkan bahwa dataset ini memiliki empat atribut dengan tiga jenis tipe atribut, yaitu binominal, polynominal, dan integer. Seluruh atribut setelah melalui tahap persiapan data memiliki nilai missing nol. Hal tersebut menunjukan bahwa tidak ditemukan lagi kesalahan nilai pada semua atribut. Pada dataset ini atribut Future Cusomer ditetapkan sebagai sebagai label dengan tipe binominal yang memiliki dua nilai, yaitu Yes dengan jumlah data sebanyak 423 dan No sebanyak 527. Atribut yang bertipe binominal lainnya adalah Gender dengan nilai Male berjumlah 550 data sementara yang bernilai Female jumlah data sebanyak 450. Atribut bertipe nominal lainnya adalah Peyment Method dengan tipe polynominal dengan nilai Credit Card sebanyak 652 data, Cash sebanyak 280 dan Cheque 68 data. Hanya terdapat satu atribut numerik yaitu atribut Age dengan type integer dimana rentang nilai mulai dari 17 sampai 91. Titik tengah dari sebaran data pada atribut ini berada pada nilai 45.7.

Table 3. Statistik Dataset Deals Hasil Proses Persiapan Data.

Nama Atribut	Tipe Atribut	Missing	Statistik		
			Nilai Terendah	Nilai Tertinggi	Rata-rata
Future Customer (sebagai label)	Binominal	0	Yes	No	No (527), Yes (423)
Age	Integer	0	17	91	45.7
Gender	Binominal	0	Male	Female	Male (550), Female (450)
Payment Method	Polynominal	0	Cheque (68)	Credit Card (652)	Credit Card (652), Cash (280), dan Cheque (68)

4. Dataset Iris, seperti yang terlihat pada Table 4. bahwa dataset Iris memiliki lima atribut dengan dua jenis tipe atribut, yaitu polynominal dan real. Nilai missing dari data untuk kesemua atribut pada dataset ini bernilai nol. Hal tersebut menunjukan bahwa tidak ditemukan lagi kesalahan nilai pada semua atribut. Atribut Sepsies ditetapkan sebagai label dimana atribut ini bertipe polinominal yang berarti data bertipe nominal dengan tiga jenis nilai atau lebih. Nilai dari atribut Species terdiri dari Sentosa sebanyak 50 data, Versicolor 50 data, dan Virginica juga terdiri dari 50 data. Atribut yang bertipe real adalah Sepal_Lenght dengan rentang nilai 4.3 sampai 7.9 dengan nilai rata-rata 5.8, selanjutnya untuk atribut Sepal_Width rentang nilai antara 2 sampai 4.4 dengan nilai rata-rata sebesar 3.1. Atribut Petal_Lenght dengan rentang nilai 1 sampai 6.9 memiliki nilai rata-rata 3.8, atribut terakhir adalah Petal_Width dengan rintang nilai antara 0.1 sampai 2.5 dengan rata-rata sebesar 1.2.

Table 4. Statistik Dataset Iris Hasil Proses Persiapan Data.

Nama Atribut	Tipe Atribut	Missing	Statistik		
			Nilai Terendah	Nilai Tertinggi	Rata-rata
Species (sebagai label)	Polinominal	0	Virginica (50)	Sentosa (50)	Sentosa (50), Versicolor (50), Virginica (50)
Sepal_Lenght	Real	0	4.3	7.9	5.8
Sepal_Width	Real	0	2	4.4	3.1
Petal_Lenght	Real	0	1	6.9	3.8
Petal_Width	Real	0	0.1	2.5	1.2

Hasil statistik dari keempat dataset tersebut dipaparkan untuk memperoleh gambaran tentang sebaran data yang ada pada masing-masing dataset tersebut. Pentingnya gambaran sebaran data yaitu untuk mengetahui factor-faktor lain yang mempengarui akurasi selain komposisi jumlah data training dan testing. Menurut Mishara, dkk. (2019) bahwa *central tendency* merupakan salah satu penjelasan statistik yang umum digunakan untuk menjelaskan tentang sebaran data dala sebuah dataset [10]. Oleh karena itu melalui Rapid Miner sebagai tools data mining dapat digunakan untuk memperoleh gambaran tentang sebaran data dalam bentuk grafis terhadap *central tencency* pada setiap atribut dataset.

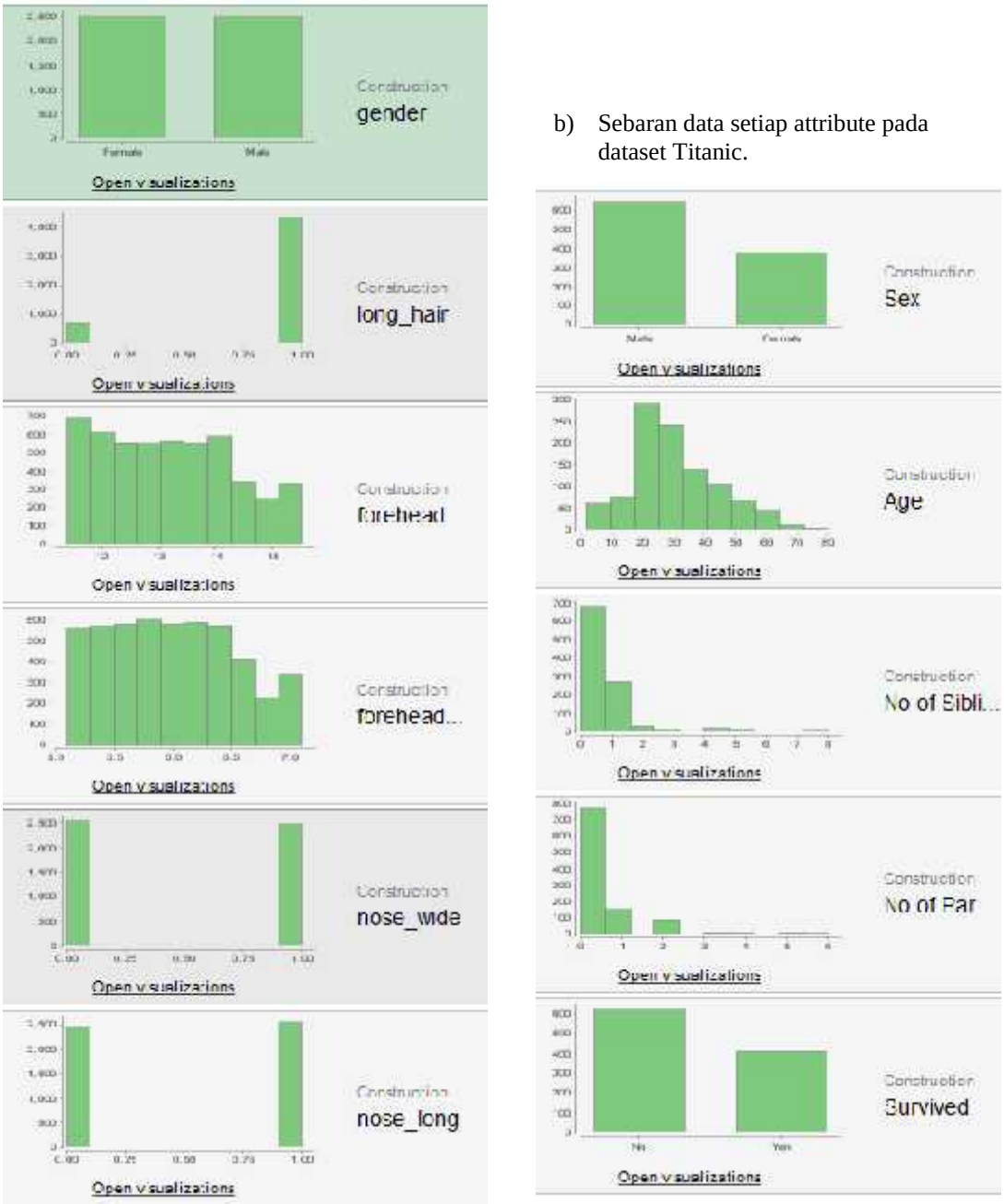
Pada Gambar 3a. terlihat sebaran data pada setiap atribut dataset Gender. Dari gambar tersebut menjelaskan bahwa sebaran data pada setiap atribut nampak seimbang disetiap atributnya kecuali pada atribut long_hair dimana konsentrasi data pada nilai sekitar satu sangat mendominasi dengan perbedaan lebih dari 75% terhadap sebaran data disekitar nilai nol.

Pada Gambar 3b. terlihat sebaran data pada setiap atribut dataset Titanic sangat bervariasi dan cenderung tidak seimbang. Pada atribut Survived sebagai label dan Atribut Sex sebaran data yang ada sekitar 3:2. Sementara, pada atribut No_of_Siblings dan No_of_Parents sebaran data cenderung kekiri dengan frekuensi sekitar 90% data berada pada nilai 0-2 dengan dominasi signifikan berada pada rentang nilai nol. Untuk atribut Age sebaran data cenderung seimbang walaupun jika diliat dari grafik *central*

tendensi bentuk data cenderung ke arah kiri. Kondisi ini dapat diprediksi akan mempengaruhi akurasi dari proses klasifikasi yang akan dilakukan.

a) Sebaran data setiap attribute pada dataset Gender.

b) Sebaran data setiap attribute pada dataset Titanic.



Gambar 3. Sebaran data setiap atribut pada dataset, a) Sebaran data setiap attribute pada dataset Gender, b) Sebaran data setiap attribute pada dataset Titanic.

Pada Gambar 4a. terlihat sebaran data pada atribut-atribut dataset Deals cenderung seimbang khususnya pada atribut yang bertipe binominal, sementara pada atribut Payment_Method yang bertipe polynominal terlihat sebaran data hampir 50% didominasi oleh data yang bernilai *credit card*. Sebaran data pada atribut Age yang bertipe integer terlihat merata walaupun dari sudut *central tendency* cenderung kea rah kiri.

Pada Gambar 4b. yang merupakan visualisasi dari sebaran data pada dataset Iris terlihat seimbang terkhusus pada atribut Species sebagai atribut label. Untuk atribut Sepal_Length dan Sepal_Width dari sudut pandang *central tendency* terlihat posisi center terhadap sebaran data. Namun, pada atribut Petal_Length dan Petal_Width *central tendency* cenderung kekanan walaupun sebaran data disekitar nilai satu dan 0.1 – 0.3 mendominasi.

a) Sebaran data setiap attribute pada dataset Deals.

b) Sebaran data setiap attribute pada dataset Iris.

Gambar 4. Sebaran data setiap atribut pada dataset, a) Sebaran data setiap attribute pada dataset Deals, b) Sebaran data setiap attribute pada dataset Iris.

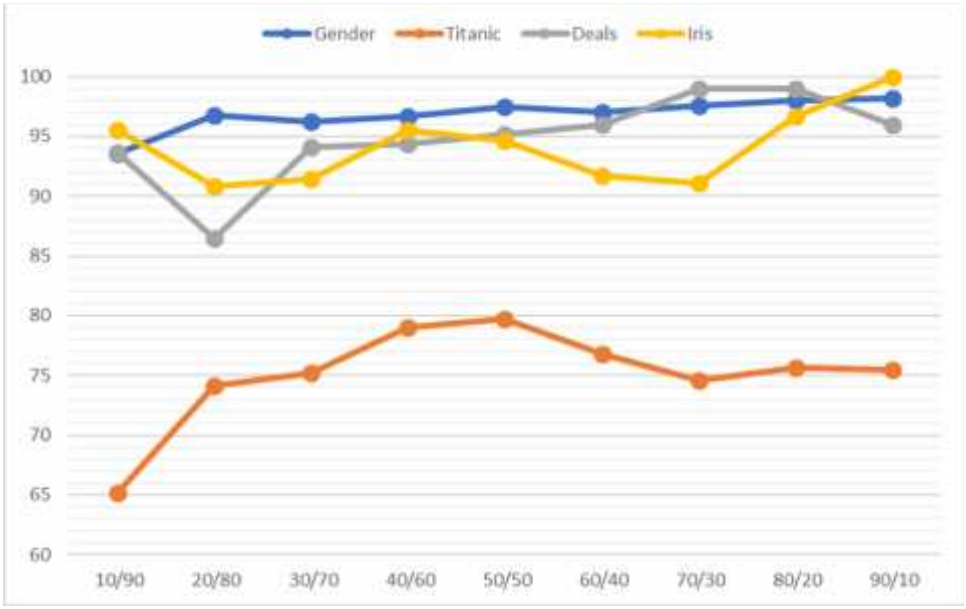
3.2. Accuracy and Composition

Hasil ekperimen dalam uji akurasi untuk setiap komposisi jumlah data training dan data testing terlihat pada Tabel 5. Pada table tersebut terlihat jumlah data pada setiap presentasi jumlah data training dan data testing yang langsung dibandingkan dengan jumlah data berdasarkan komposisi persentasi beserta hasil akurasi yang diperoleh. Penjelasan pada table tersebut memperlihatkan bahwa dari komposisi jumlah data bervariasi disetiap dataset yang digunakan. Hal tersebut bertujuan untuk melihat pengaruh jumlah data dan kondisi data terhadap nilai akurasi yang diperoleh di setiap komposisi yang digunakan. Menurut Maksim, dkk. (2019) bahwa jumlah data sangat menentukan hasil analisis pada algoritma-algoritma Artificial Inteligent [11]. Oleh karena itu dalam penelitian ini perbedaan jumlah data dibutuhkan selain keragaman sebaran data untuk melihat pengaruh akurasi dari algoritma C4.5.

Tabel 5. Presentasi dan Komposisi Data disetiap Dataset terhadap Nilai Akurasi.

Persentasi	Gender		Titanic		Deals		Iris	
	Komposisi Data	Akurasi	Komposisi Data	Akurasi	Komposisi Data	Akurasi	Komposisi Data	Akurasi
10/90	500/4500	93.53	102/921	65.18	100/900	93.67	15/135	95.56
20/80	1000/4000	96.73	204/819	74.11	200/800	86.5	30/120	90.83
30/70	1500/3500	96.26	307/716	75.17	300/700	94.14	45/105	91.43
40/60	2000/3000	96.67	409/614	78.99	400/600	94.33	60/90	95.56
50/50	2500/2500	97.44	512/512	79.69	500/500	95.2	75/75	94.67
60/40	3000/2000	97	614/409	76.83	600/400	96	90/60	91.67
70/30	3500/1500	97.6	716/307	74.59	700/300	99	105/45	91.11
80/20	4000/1000	98	819/204	75.67	800/200	99	120/30	96.67
90/10	4500/500	98.2	921/102	75.49	900/100	96	135/15	100

Pada Gambar 5. terlihat hasil pengujian akurasi terhadap dataset Gender, Titanic, Deals dan Iris. Dari keempat dataset tersebut nilai akurasi dataset Titanic paling rendah untuk semua komposisi jumlah data training dan testing yaitu dengan akurasi di bawah 80%, sementara keriga dataset lainnya nilai akurasi berada di atas 85%.



Gambar 5. Nilai Akurasi Dataset Gender, Titanic, Deals, dan Iris pada Sembilan Komposisi Jumlah Data Training dan Data Testing.

Dataset Gender dengan jumlah data 5001, nilai akurasi terlihat stabil pada nilai sekitar 95%. Hasil yang diperoleh menunjukan walaupun nilai akurasi tertinggi di peroleh pada komposisi 90/10 namun tetap terjadi fluktuatif nilai akurasi yang rendah pada komposisi tertentu. Pada komposisi 20/80 merupakan komposisi yang lebih tinggi dibandingkan komposisi 10/90 dan 30/70 walaupun nilai akurasinya masih tetap berada dibawah nilai akurasi pada komposisi 95%. Kondisi tersebut juga terjadi pada komposisi 50/50 dimana nilai akurasinya lebih tinggi dari nilai akurasi komposisi 40/50 dan 60/50. Hal ini mengindikasikan bahwa komposisi jumlah data training mempengaruhi perhitungan nilai entropi dan gain dari algoritma C4.5 walaupun pada dataset Gender tidak terlihat signifikan. Jumlah data pada dataset dan variasi data mempengaruhi akan hal tersebut.

Hasil nilai akurasi pada dataset Titanic yang memiliki jumlah data 1024 walaupun nampak rendah tetapi kenaikan nilai akurasi terlihat naik seiring meningkatnya jumlah data training. Pada Gambar

5. terlihat dari komposisi 10/90 dengan nilai akurasi sekitar 65% terjadi kenaikan sampai pada komposisi 50/50 hingga mencapai nilai akurasi 80%. Namun setelah itu pada komposisi 60/40 nilai akurasi menurun sampai pada komposisi 90/10 dengan nilai akurasi sekitar 75%. Hasil ini menjelaskan bahwa komposisi jumlah data training dan data testing memiliki nilai akurasi tertinggi pada komposisi 50/50 dengan pola data yang dimiliki oleh dataset Titanic. Perhitungan nilai entropi dan gain pada C4.5 sangat dipengaruhi oleh karakteristik data pada atribut No_of_Siblings dan No_of_Parents.

Untuk dataset Deals yang memiliki jumlah data 1000 komposisi jumlah data training dan data testing juga mempengaruhi nilai akurasi. Pada komposisi 10/90 nilai akurasi diperoleh sebesar 93.67%, namun pada komposisi 20/80 turun drastis sampai pada nilai 86.5% dan kemudian mengalami kenaikan mulai pada komposisi 30/70 sampai mencapai puncak pada komposisi 70/30 dan 80/20 sebesar 99%. Pada komposisi 90/10 nilai akurasi kembali menurun pada nilai 96%. Pada dataset ketiga ini kembali memperlihatkan bahwa komposisi jumlah data training tertinggi tidak menunjukan hasil akurasi yang tinggi pula.

Dataset Iris dengan jumlah data terendah yaitu sebanyak 150 data memiliki grafik fluktuatif yang sama dengan dataset sebelumnya. Hasil akurasi pada dataset ini lebih bergelombang dimana pada komposi 10/90 dimulai dengan nilai akurasi sebesar 95.56% lalu turun pada komposisi 20/80 sebesar 90.83. Walaupun mengalami kenaikan sampai pada komposisi 40/60 dengan nilai 95.56% tetapi mengalami penurunan lagi sampai pada komposisi 70/80. Pada komposisi 80/20 nilai akurasi kembali naik dan mencapai puncak pada komposisi 90/10 dengan nilai akurasi 100%.

Dari seluruh hasil yang diperoleh memperlihatkan bahwa jumlah data pada dataset tidak berpengaruh terhadap fluktuasi nilai akurasi pada setiap komposi jumlah data training dan data testing. Penerapan algoritma C4.5 harus memperhatikan komposisi data training dan testing, disarankan untuk memeriksa semua komposisi yang ada untuk mendapatkan nilai akurasi yang maksimum sehingga informasi pencapaian dilai akurasi pada setiap eksperimen menjadi valid dan dapat di percaya.

4. Conclusion

Dari hasil eksperimen yang dilakukan dalam penelitian ini yang menggunakan empat dataset dengan karakteristik masing-masing, maka disimpulkan:

1. Nilai akurasi pada algoritma C4.5 fluktuatif dari nilai komposisi data training terendah sampai dengan komposisi data training ter tinggi. Oleh karena itu dalam meng implementasikan algoritma ini perlu untuk melakukan pengujian di setiap komposisi data training dan data tesing sehingga nilai akurasi yang di peroleh menjadi maksimal.
2. Nilai akurasi di pengaruhi oleh sebaran data dari dataset yang digunakan. Oleh karena itu proses persiapan data dan metode pengumpulan data harus diperhatikan untuk mendapatkan kualitas dataset yang dapat memberikan nilai akurasi terbaik.
3. Eksperiment tentang komposisi jumlah data training dan dataset ini dapat digunakan untuk mengevaluasi kualitas data dari dataset yang di miliki.

References

[1] P. B. N. Setio, D. R. S. Saputro, and Bowo Winarno, “Klasifikasi Dengan Pohon Keputusan Berbasis Algoritme C4.5,” *Prism. Pros. Semin. Nas. Mat.*, vol. 3, pp. 64–71, 2020.

[2] G. M. D. Godaliyadda, D. H. Ye, M. D. Uchic, M. A. Groeber, G. T. Buzzard, and C. A. Bouman, “A Supervised Learning Approach for Dynamic Sampling,” *Electron. Imaging*, vol. 2016, no. 19, pp. 1–8, Feb. 2016, doi: 10.2352/ISSN.2470-1173.2016.19.COIMG-153.

[3] A. Fernández, V. López, M. Galar, M. J. Del Jesus, and F. Herrera, “Analysing the classification of imbalanced data-sets with multiple classes: Binarization techniques and ad-hoc approaches,” *Knowledge-Based Syst.*, vol. 42, pp. 97–110, 2013, doi: 10.1016/j.knosys.2013.01.018.

[4] A. Nair, B. D. Kuban, N. Obuchowski, and D. Geoffrey Vince, “Assessing spectral algorithms to predict atherosclerotic plaque composition with normalized and raw intravascular ultrasound data,” *Ultrasound Med. Biol.*, vol. 27, no. 10, pp. 1319–1331, 2001, doi: 10.1016/S0301-5629(01)00436-7.

[5] M. F. Rahman, D. Alamsah, M. I. Darmawidjadja, and I. Nurma, “Klasifikasi Untuk Diagnosa Diabetes Menggunakan Metode Bayesian Regularization Neural Network (RBNN),” *J. Inform.*, vol. 11, no. 1, p. 36, 2017, doi: 10.26555/jifo.v11i1.a5452.

[6] A. Novantirani, M. K. Sabariah, and V. Effendy, “Analisis Sentimen pada Twitter untuk Mengenai Penggunaan Transportasi Umum Darat Dalam Kota dengan Metode Support Vector Machine,” *e-Proceeeding Eng.*, vol. 2, no. 1, pp. 1–7, 2015.

[7] I. P. P. Iswara, F. Farhan, W. Kumara, and A. A. Supianto, “Rekomendasi Pengambilan Mata

Kuliah Pilihan Untuk Mahasiswa Sistem Informasi Menggunakan Algoritme Decision Tree," *J. Teknol. Inf. dan Ilmu Komput.*, vol. 6, no. 3, pp. 341–348, 2019, doi: 10.25126/jtiik.2019.6892.

[8] S. Stieglitz, M. Mirbabaie, B. Ross, and C. Neuberger, "Social media analytics – Challenges in topic discovery, data collection, and data preparation," *Int. J. Inf. Manage.*, vol. 39, pp. 156–168, Apr. 2018, doi: 10.1016/j.ijinfomgt.2017.12.002.

[9] S. K. Kwak and J. H. Kim, "Statistical data preparation: management of missing values and outliers," *Korean J. Anesthesiol.*, vol. 70, no. 4, p. 407, 2017, doi: 10.4097/kjae.2017.70.4.407.

[10] P. Mishra, C. Pandey, U. Singh, A. Gupta, C. Sahu, and A. Keshri, "Descriptive statistics and normality tests for statistical data," *Ann. Card. Anaesth.*, vol. 22, no. 1, p. 67, 2019, doi: 10.4103/aca.ACA_157_18.

[11] K. Maksim *et al.*, "Classification of Wafer Maps Defect Based on Deep Learning Methods With Small Amount of Data," in *2019 International Conference on Engineering and Telecommunication (EnT)*, Nov. 2019, pp. 1–5, doi: 10.1109/EnT47717.2019.9030550.