ANALISA AKURASI PERMODELAN SUPERVISED DAN UNSUPERVISED LEARNING MENGGUNAKAN DATA MINING

Warnia Nengsih

Sistem Informasi, Politeknik Caltex Riau
Jl Umban Sari No 1 Rumbai Pekanbaru 28265
E-mail : warnia@pcr.ac.id

ABSTRAK

Data Mining merupakan salah satu proses yang menggunakan teknik statistik, matematika, kecerdasan buatan, *machine learning* untuk mengekstraksi dan mengidentifikasi informasi yang bermanfaat dan pengetahuan yang terkait dari berbagai database besar. Data mining memiliki dua jenis pembelajaran diantaranya *supervised learning* dan *unsupervised learning*. Tentunya setiap pembelajaran memiliki teknik dan algoritma tersendiri. Penelitian ini bertujuan untuk melakukan permodelan dari setiap *learning* dengan mengukur akurasi dari kedua jenis *learning* tersebut menggunakan beberapa metode pengujian. Sementara untuk rancang sistem menggunakan bahasa pemograman matlab. Belum adanya pengukuran akurasi dari kedua *learning* menjadi hal yang melatarbelakangi penelitian ini. Dari hasil pengujian akurasi menggunakan *confusion matrix* dan *lift ratio* diperoleh hasil bahwa perbandingan rata-rata akurasi untuk *supervised learning* adalah 82,33% dan *unsupervised learning* sebesar 78% dengan selisih nilai akurasi sebesar 4,33%. Nilai akurasi dipengaruhi oleh jumlah serta keberagaman dimensi data. Jadi dengan kasus dan jumlah serta dimensi yang berbeda akan menghasilkan nilai akurasi yang beragam pula.

Kata Kunci: *Supervised Learning, Unsupervised Learning, Akurasi, Data Mining*

1. PENDAHULUAN

Data Mining merupakan bidang ilmu yang menyatukan teknik pembelajaran, pengenalan pola, statistik, *database*, serta visualisasi untuk mengatasi masalah ekstraksi informasi dari basis data yang besar. Terdapat dua jenis *learning* pada *data mining* yaitu *supervised learning dan unsupervised learning* (Lumbantoruan, 2015). Setiap *learning* memiliki metode dan algoritma masing masing. Metode yang termasuk ke dalam S*upervised learning* diantaranya *regression,classification,predictive,summarization.*Seme ntara itu *unsupervised learning* terdiri dari metode *clustering, association, knowledge discovery* dan sebagainya (Mabrur, 2012).

Belum adanya pengukuran untuk mengetahui keakuratan dari salah satu permodelan pada saat menggunakan teknik tertentu untuk permasalahan dengan studi kasus yang berbeda menjadi hal yang melatarbelakangi penelitian ini. Rumusan permasalahan dapat diuraikan sebagai berikut: Bagaimana memecahkan permasalahan pada studi kasus yang berbeda dengan menggunakan teknik *data mining* yang ada, Bagaimana mengukur akurasi dari *supervised* dan *unsupervised learning* menggunakan studi kasus dan teknik yang beragam menggunakan metode pengujian yang sudah ditentukan

Tujuan dari penelitian ini adalah untuk mengukur akurasi dari *supervised* dan *unsupervised learning*. Pengukuran akurasi menggunakan metode *confusion matrix* untuk *supervised learning dan pengujian akurasi lift ratio* untuk *unsupervised learning*. Adapun jumlah variabel yang digunakan disesuaikan dengan variabel sebab (x) dari *data set* yang digunakan. Sistem dibangun menggunakan *matlab* sedangkan pengujian akurasi dilakukan di luar sistem. *Output* yang diperoleh dari pengukuran akurasi adalah diperolehnya sebuah pengetahuan baru berupa perbandingan akurasi dari kedua kategori *learning*, sehingga bermanfaat untuk pengembangan penelitian selanjutnya. Beberapa review penelitian terdahulu yang terkait dengan penelitian ini adalah penelitian dengan judul "*Comparison of Supervised and Unsupervised Learning Algorthms for Pattern Classification pada International Journal of Advanced Research ini Artificial Intelligence oleh Sathya Annamma.* Dimana penelitian ini menunjukkan perbandingan antara *supervised* dan *unsupervised learning* serta penentuan pola klasifikasi untuk studi kasus Pendidikan Tinggi. Kemudian review penelitian selanjutnya adalah penelitian dengan judul *"Comparison of Supervised and Unsupervised Learning Classifier for Travel Recommendations*" oleh Zohreh Bahman Isfahani, Shiraz University Iran. Penelitian ini membandingkan tentang pengelompokkan data pariwisata menggunakan dua metode *supervised* dan *unsupervised*. Sementara penelitian yang dilakukan sekarang adalah perbandingan dua metode *learning* menggunakan teknik dan studi kasus yang beragam. Perbandingan review penelitian ini dapat terlihat pada tabel berikut :

Tabel 1. Review Penelitian terdahulu

Judul	Metode Learning	Scope Riset	Output
Comparison of Supervised and Unsupervised Learning Algorthms for Pattern Classification	*Supervised and Unsupervised Learning*	*Classification* Teknik dengan permasalahan pendidikan tinggi	Membandingkan metode *supervised* dengan *unsupervised* pada satu studi kasus yang sama
Comparison of Supervised and Unsupervised Learning Classifier for Travel Recommendations	*Supervised and Unsupervised Learning*	*Classification* Teknik dengan permasalahan pariwisata	Membandingkan metode *supervised* dengan *unsupervised* pada satu studi kasus yang sama
Modeling Accuracy Analysis Supervised dan Unsupervised Learning Using Data Mining	*Supervised and Unsupervised Learning*	Semua teknik data mining dengan kasus yang beragam	Menentukan dan membandingkan akurasi dari kedua permodelan menggunakan teknik-teknik yang ada pada data mining dengan studi kasus yang berbeda.

2. RUANG LINGKUP

Berikut merupakan lingkup penelitian:

1. Penelitian ini membandingkan dua permodelan yaitu *supervised learning* dan *unsupevised learning*
2. Metode yang digunakan pada *supervised learning adalah Decision tree, Support Vector Machine dan Regresi Linear.* Sementara metode yang digunakan untuk untuk *unsupervised learning adalah k-means, single linkage dan apriori*
3. Studi kasus yang digunakan untuk setiap permodelan berbeda

3. BAHAN DAN METODE

Disajikan kajian teori dan metodologi dalam penelitian ini

3.1 *Analisis system*

Permasalahan prediksi, *forecasting, clustering*, deteksi anomali dapat menggunakan teknik data mining yang ada. Setiap teknik data mining merupakan turunan dari beberapa permodelan *supervised* dan *unsupervised.* (Prasetyo, 2014) Setiap permasalahan yang ada akan diselesaikan dengan menggunakan permodelan yang sudah dibuat dengan menggunakan teknik yang tepat untuk mendapatkan *knowledge* atau pengetahuan. Namun perlu adanya pengukuran akurasi dari setiap model pembelajaran yang digunakan, sehingga dapat disimpulkan persentase akurasi dari setiap permasalahan dengan teknik yang digunakan pada setiap permodelan.

Secara konsep, jika akurasi tinggi dan nilai kesalahan rendah maka menunjukkan hasil yang baik terhadap keakuratan permodelan yang dihasilkan. Namun sebaliknya jika nilai akurasi rendah dan dan nilai kesalahan tinggi maka tentunya teknik yang dipilih belum tepat untuk menyelesaikan permasalahan tersebut.

Pengukuran akurasi dari kedua jenis *learning* tersebut menggunakan beberapa metode pengujian disesuaikan dengan metode *learning* yang digunakan. *Dataset* diambil dari studi kasus yang berbeda disesuaikan dengan permodelan yang digunakan.

3.2 Implementasi Sistem

Selanjutnya dilakukan permodelan sistem menggunakan matlab sesuai dengan metode pembelajaran yang digunakan.

3.3 Pengujian

Pengujian digunakan untuk mengukur akurasi dari sebuah kasus dengan metode yang digunakan. Adapun pengujian yang digunakan adalah *confusion matrix* dan *lift ratio* untuk masing – masing *learning*.

3.4 Confusion Matrix

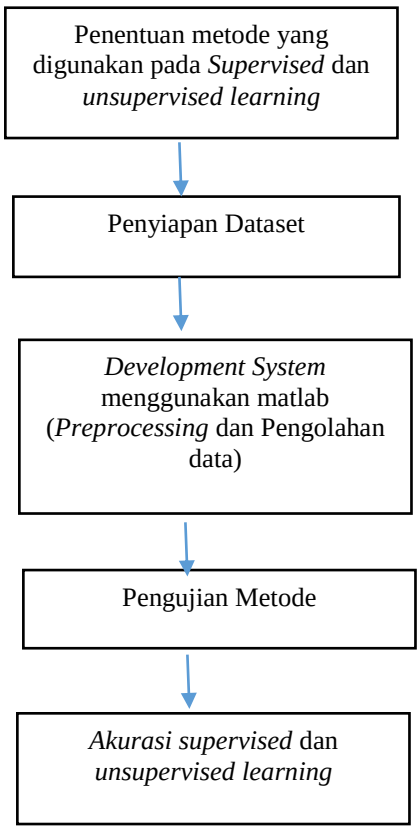
Confusion matrix merupakan salah satu metode yang dapat digunakan untuk mengukur kinerja suatu metode klasifikasi. Pada dasarnya *confusion matrix* mengandung informasi yang membandingkan hasil klasifikaasi yang dilakukan oleh sistem dengan hasil klasifikasi yang seharusnya (Afrisawati, 2002)

3.5 Lift Ratio

Lift Ratio adalah parameter penting selain *suppor*t dan *confidence* dalam *association rule. Lift Ratio* mengukur seberapa penting *rule* yang telah terbentuk berdasarkan nilai *support* dan *confidence.*

Penelitian ini menggunakan beberapa metode pada *supervised learning* diantaranya *regresi linear, decision tree* dan *Support Vector Machine*. Serta metode *unsupervised learning* diantaranya *K-Means , single linkage dan apriori.* Interface sistem untuk kasus menggunakan bahasa pemograman matlab. Penelitian ini menggunakan data dengan studi kasus yang berbeda karena harus disesuaikan dengan kebutuhan metode yang digunakan. Metode pengujian menggunakan *confusion matrix* untuk *supervised learning* dan *lift ratio* untuk metode-metode yang terdapat pada *unsupervised*

learning. Berikut gambar 1 merupakan gambaran umum dari penelitian.



Gambar 1. metodologi penelitian

4. PEMBAHASAN

Disajikan hasil dan pembahasan dalam penelitian ini

4.1 Metode *K-Means*

K-Means merupakan salah satu algoritma *clustering* (al, 2016). Tujuan algoritma ini yaitu untuk membagi data menjadi beberapa kelompok. Algoritma ini menerima masukan berupa data tanpa label kelas. (F. E. M. Agustin, 2013). K-means merupakan salah satu algoritma *clustering*. Tujuan algoritma ini yaitu untuk membagi data menjadi beberapa kelompok. Algoritma ini menerima masukan berupa data tanpa label kelas. Hal ini berbeda dengan *supervised learning* yang menerima masukan berupa vektor (x_1, y_1) , (x_2, y_2) , …, (x_i, y_i), di mana x_i merupakan data dari suatu data pelatihan dan y_i merupakan label kelas untuk x_i (K. Rajalakshmi). Berikut merupakan *dataset* yang digunakan.

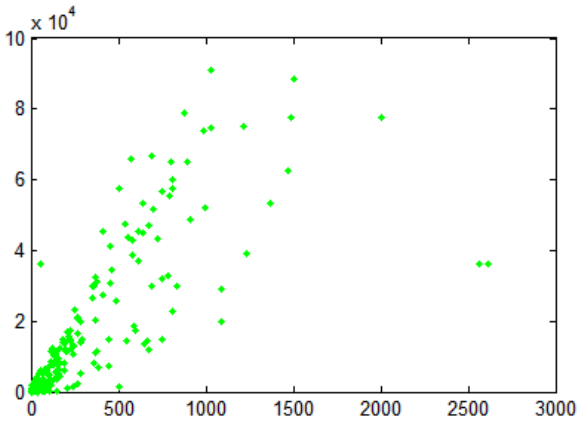
5.34	211.4	39.588
1.16	40.8	35.1724
6.5	252.2	38.8
9.66	221.47	22.9265
9.66	221.47	22.9265
9.66	221.47	22.9265
50	200	20
60	100	19
70	150	18
30	130	17
40	140	16
50	160	14
40	135.33	3.38325
110	24	0.21818
150	159.33	1.0622
13	78	15
12	88	14
28.74	3278.61	114.078
22.49	3090.11	137.399
6.25	188.5	30.16
6.25	188.5	30.16
6.25	188.5	30.16
6.25	188.5	30.16
6.25	188.5	30.16
6.25	188.5	30.16
6.25	188.5	30.16
6.25	188.5	30.16
6.48	1793.09	276.711
54.52	1576	28.9068

Gambar 2. Data set menggunakan K-Means

Dari data ini terdapat 3 variabel sebab (X), yaitu Produktivitas, Luas Lahan , Produksi. *Output dataset* menggunakan 2 *cluster*

```
x=handles.x;
k=handles.k;
[cidx,ctrs] = kmeans (x,k, 'dist' , 'sqeuclidean');
set(handles.listbox1, 'String', cidx);


X = x(:,1);

y = x(:,2);

hold on;
colors = 'rgb';
for num = 1:k
    plot(ax2,x(cidx==num), y(cidx==num) , [colors(num) '.']);
end

plot(ctrs(:,1),ctrs(:,2), '+k', 'MarkerSize', 14,'LineWidth', 3);
grid;
```

Berikut merupakan hasil *cluster* dari *dataset* yang diolah.



Gambar 3. Hasil pengclusteran menggunakan K-Means

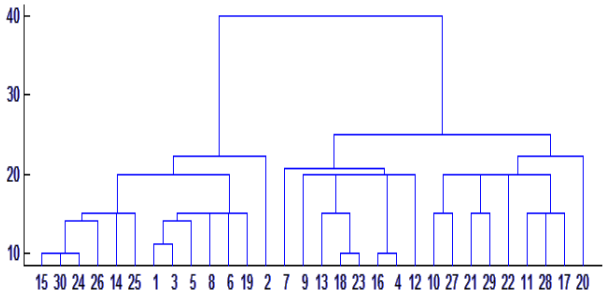
4.2 *Metode Single linkage*

Single linkage merupakan metode yang terdapat pada kategori *clustering* hirarki yang jumlah clusternya diperoleh setelah pengolahan data dilakukan. Berikut ini *dataset* yang digunakan.

100	130	110
150	125	100
120	120	100
110	120	100
130	110	100
100	110	110
100	120	175
100	145	110
100	150	175
100	160	150
150	160	150
100	160	150
150	145	175
150	150	175
150	160	150
130	125	175
100	160	175
100	150	175
100	150	110
140	125	150
120	160	150
120	150	175
100	150	175

Gambar 4. *Data set* menggunakan single linkage

Dataset yang digunakan untuk mengetahui *cluster* pegawai tidak memenuhi dan memenuhi dengan menggunakan 3 kriteria yaitu nilai TWK (Tes Wawancara Kebangsaan),TIU (Tes Inteligent Umum) dan TKP (Tes Karakteristik Kepribadian).



Gambar 5. Hasil pengclusteran menggunakan *Single Linkage*

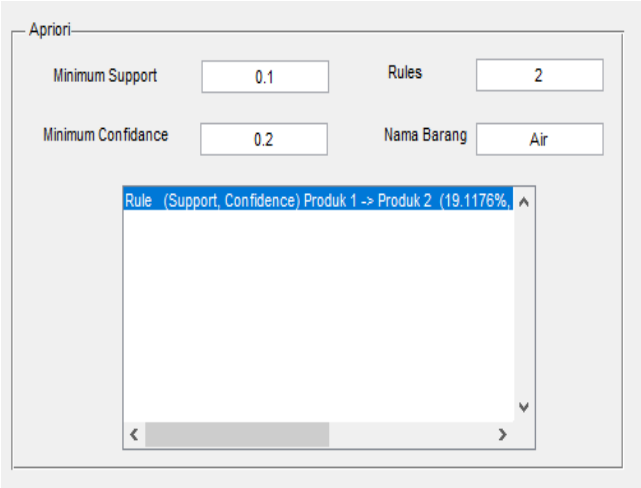
4.3 Metode *Apriori*

Berikut merupakan *dataset* yang digunakan pada *apriori*

Klorida	Asetat	Fluorida	Sulfat	Natrium K Air	Asam Ami	Metabolit	Pegtida	Protein	Natrium K	Kalium Di	Mangan D	Nikel Klor	Pentahidr	Kobalt Klc	Kalium	Ferisinid
1	0	1	0	0	1	1	0	1	0	1	0	0	1	0	1	0
1	0	1	0	0	0	1	0	1	0	1	0	1	1	1	0	1
0	0	1	0	0	0	1	0	1	0	1	0	1	1	1	1	1
1	0	0	0	1	0	0	0	1	0	1	0	1	0	1	1	1
0	0	0	0	1	0	0	1	1	0	0	0	0	0	0	0	0
0	1	1	1	1	1	0	1	1	0	0	1	0	0	0	0	0
0	1	1	1	0	1	1	1	1	0	0	1	0	0	0	1	0
1	0	1	0	0	1	1	0	0	1	1	1	1	1	1	1	1
0	0	0	0	0	1	1	0	0	1	1	1	1	1	1	1	1
0	1	0	0	0	0	0	1	0	1	1	1	1	1	1	1	1
1	1	0	1	1	0	0	1	1	1	1	1	1	0	1	1	0
1	0	0	1	0	0	0	1	1	1	1	1	0	0	1	1	0
1	0	0	1	1	0	0	1	1	1	0	0	0	0	0	1	0
0	0	1	0	0	1	1	0	1	1	0	0	0	0	0	1	1
0	0	1	0	0	1	1	0	0	0	0	0	0	0	0	1	0
1	1	1	1	0	1	1	0	0	0	1	1	1	1	0	1	1
1	1	1	1	1	1	0	0	0	0	1	1	1	1	0	1	1
0	0	0	1	1	0	1	0	1	0	1	1	1	1	0	0	0
0	0	0	1	1	0	1	1	0	1	1	1	1	1	0	0	0
0	0	0	0	1	0	1	1	1	0	1	1	1	0	1	0	0
0	1	1	0	0	1	1	1	1	0	0	1	0	0	1	0	1

Gambar 6. *Dataset* menggunakan apriori

Algoritma apriori adalah suatu metode untuk mencari pola hubungan antar satu atau lebih item dalam suatu dataset (K. Tampubolon, 2013).Pada metode ini tentukan nilai *minimum support* dan *minimum confidence* yang digunakan untuk mengetahui berapa nilai keterkaitan antara item.Gambar 7 menunjukkan *rule* dari nilai keterkaitan yang diperoleh



Gambar 7. Hasil asosiasi menggunakan *matlab*

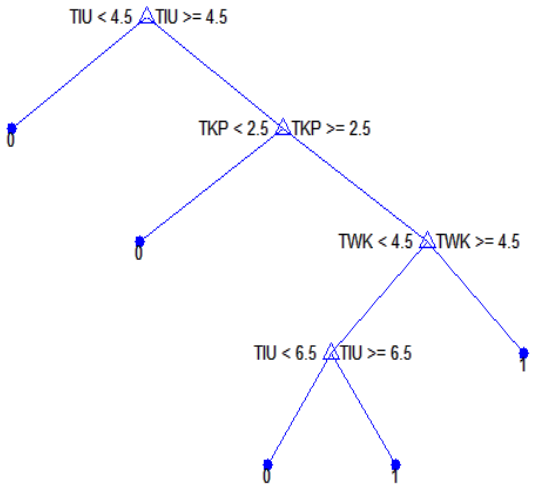
4.4 Metode *Decision Tree*

Decision tree merupakan salah satu metode teknik klasifikas dimana kelas data sudah diketahui sebelumnya.

6	5	3
1	6	4
4	7	4
5	7	4
3	8	4
6	8	3
6	7	1
6	4	3
6	3	1
6	2	2
1	2	2
6	2	2
1	4	1
1	3	1
1	2	2
3	6	1
6	2	1
6	3	1
6	3	3
5	6	2
3	2	2
3	3	1
1	3	1

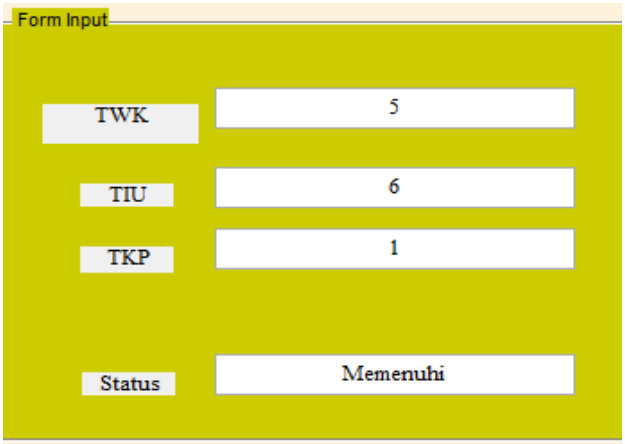
Gambar 8. Dataset mengunakan Decision Tree

Berikut merupakan *rule* yang digunakan dari hasil pengolahan *dataset*.



Gambar 9. hasil pohon keputusan

Dari pohon keputusan tersebut didapat informasi bahwa jika TIU (Tes Inteligent Umum) kecil dari 4.5 maka hasilnya memenuhi. Dan jika TIU (Tes Inteligent Umum) besar dari 4.5 dan TKP besar dari 2.5 serta TWK besar dari 4.5 maka label yang diperoleh tidak memenuhi. Jika TIU besar dari 4.5 dan TKP (Tes Karakteristik Kepribadian) kecil dari 2.5 maka memenuhi. Selanjutnya jika TIU besar dari 4.5,TKP besar dari 2.5 dan TWK kecil dari 4.5 serta TIU < 6.5 maka label yang digunakan memenuhi. Jika TIU besar dari 4.5,TKP besar dari 2.5 dan TWK kecil dari 4.5 serta TIU> 6.5 maka label yang digunakan tidak memenuhi



Gambar 10. Hasil prediksi menggunakan Decision Tree

4.5 Metode *Support Vector Machine(SVM)*

Teknik SVM digunakan untuk menemukan *fungsi pemisah*(klasifier) yang optimal yang bisa memisahkan dua set data dari dua kelas yang berbeda (Kurniawan Defri, 2013). Berikut merupakan hasil *interface* yang digunakan untuk metode SVM. Studi kasus yang digunakan adalah untuk melakukan prediksi jenis pinjaman untuk nasabah.

Gambar 11. Hasil prediksi menggunakan SVM

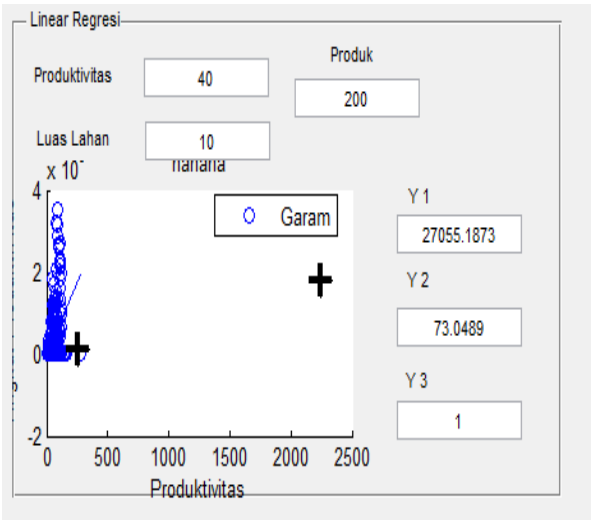
4.6 Metode *linear regresi*

Berikut merupakan data yang digunakan untuk *regresi linear*.

5.34	211.4	39.588
1.16	40.8	35.1724
6.5	252.2	38.8
9.66	221.47	22.9265
9.66	221.47	22.9265
9.66	221.47	22.9265
50	200	20
60	100	19
70	150	18
30	130	17
40	140	16
50	160	14
40	135.33	3.38325
110	24	0.21818
150	159.33	1.0622
13	78	15
12	88	14
28.74	3278.61	114.078
22.49	3090.11	137.399
6.25	188.5	30.16
6.25	188.5	30.16
6.25	188.5	30.16
6.25	188.5	30.16
6.25	188.5	30.16
6.25	188.5	30.16
6.25	188.5	30.16
6.25	188.5	30.16
6.48	1793.09	276.711
54.52	1576	28.9068

Gambar 12. Dataset untuk metode linear regresi

Dataset yang digunakan untuk menentukan jumlah produksi garam. Adapun variabel yang digunakan adalah luas lahan, produktivitas, produksi.

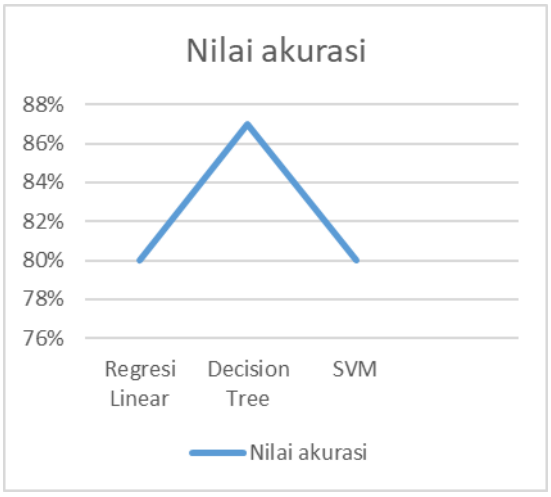


Gambar 13. Hasil dari Metode Linear Regresi

Berikut merupakan hasil pengujian akurasi unuk setiap *learning*:

Tabel 2 Hasil akurasi metode supervised learning

No	Metode Supervised Learning	Nilai Akurasi
1	Regresi Linear	80%
2	Decision Tree	87%
3	Support Vector Machine	80%
Rata-rata akurasi		82,33%

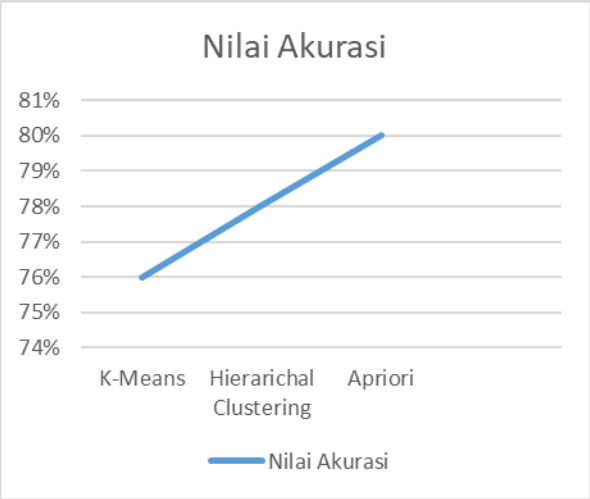


Gambar 14. visualisasi akurasi supervised learning

Untuk hasil akurasi metode *unsupervised* ditunjukkan pada tabel 2.

Tabel 3. Hasil akurasi metode unsupervised learning

No	Unsupervised Learning	Nilai Akurasi
1	K-means	76%
2	Hierarichal clustering	78%
3	Apriori	80%
Rata-rata akurasi		78%



Gambar 15. visualisasi akurasi unsupervised learning

Tabel 4. Perbandingan rata-rata akurasi

Supervised Learning(akurasi)	Unsupervised Learning(akurasi)
82,33%	78%

Berikut merupakan rata-rata akurasi untuk *supervised learning* 82,33% dan unsupervised learning sebesar 78% dengan selisih nilai akurasi sebesar 4,33 %. Dimana nilai akurasi dipengaruhi oleh jumlah serta keberagaman dimensi data. Jadi dengan kasus dan jumlah serta dimensi yang berbeda akan menghasilkan nilai akurasi yang beragam pula.

5. KESIMPULAN

Penelitian ini menggunakan beberapa metode pada *supervised learning* diantaranya regresi linear, *decision tree* dan *Support Vector Machine*. Serta metode *unsupervised learning* diantaranya *K-Means*, *single linkage* dan *apriori*. *Interface* sistem untuk kasus menggunakan bahasa pemograman *matlab*. Penelitian ini menggunakan data dengan studi kasus yang berbeda karena harus disesuaikan dengan kebutuhan metode yang digunakan. Dari hasil pengujian akurasi menggunakan *confusion matrix* dan *lift ratio* diperoleh hasil bahwa Perbandingan rata-rata akurasi untuk *supervised learning* adalah 82,33% dan *unsupervised learning* sebesar 78% dengan selisih nilai akurasi sebesar 4,33 %. Nilai akurasi dipengaruhi oleh jumlah serta keberagaman dimensi data. Jadi dengan kasus dan jumlah serta dimensi yang berbeda akan menghasilkan nilai akurasi yang beragam pula.Metode *supervised* dan *unsupervised learning* yang digunakan harus lengkap sehingga dapat mewakili semua metode pada masing-masing learning yang digunakan

6. SARAN

Saran untuk penelitian berikutnya menggunakan dataset yang sama untuk setiap permodelan, sehingga dapat dilihat perbandingan akurasinya.

7. DAFTAR PUSTAKA

Afrisawati. 2002. Implementasi Data Mining Pemilihan Pelanggan Potensial menggunakan Algoritma K-Means, *Pelita Inform. Budi Darma, vol. V, no. 12110955, (2013):pp. 157–162.*

E. Prasetyo.2014. Data Mining: Konsep dan Aplikasi menggunakan Matlab, *1 ed. Yogyakarta: Andi Offset.*

F. E. M. Agustin, A. Fitria, and A. H. S. 2013. " Implementasi Algoritma K-Means Untuk Menentukan Kelompok Pengayaan Materi Mata Pelajaran Ujian Nasional. (Studi Kasus Smp Negeri 101 Jakarta), *vol. 8:pp. 73–78.*

G.Abdillah et al.2016. Penerapan Data Mining Pemakaian Air Pelanggan Untuk Menentukan Klasifikasi Potensi Pemakaian Air Pelanggan Baru Di PDAM Tirta Raharja Menggunakan Algoritma K-Means,*:pp. 18–19.*

Kurniawan Defri, Catur Suprianto. 2013. Optimasi Algoritma Support Vector Machine untuk penilaian Risiko Kredit, *Universitas Dian Nuswantoro*.

K. Tampubolon, H. Saragih, B. Reza, K. Epicentrum, .2013. "Implementasi Data Mining Algoritma Apriori Pada Sistem Persediaan Alat-Alat Kesehatan," *Inf. dan Teknol. Ilm.,: pp. 93–106.*

K. Rajalakshmi, S. S. Dhenakaran, and N. Roobini, "Comparative Analysis of K-Means Algorithm in Disease Prediction," *Int. J. Sci. Eng. Technol. Res., vol. 4, no. 7, pp. 2697– 2699*

Lumbantoruan, Rutman dan Posma Sariguna Johnson Kennedy.2015. "Analisis Data Mining Dan Warehousing". *Jurnal Ilmiah Buletin Ekonomi, Volume 19 no.1*.

L. R. Angga Ginanjar Mabrur.2012. "Penerapan Data Mining Untuk Memprediksi Kriteria Nasabah Kredit," *J. Komput. dan Inform., vol. 1, no. 1:pp. 53–57*