Integrasi Metode Klasifikasi Dan Clustering dalam Data Mining

Conference Paper · January 2015

CITATIONS
17

READS
46,183

2 authors:

Sofi Defiyanti
University of Singaperbangsa Karawang
24 PUBLICATIONS **187** CITATIONS
SEE PROFILE

Mohamad Jajuli
University of Singaperbangsa Karawang
16 PUBLICATIONS **175** CITATIONS
SEE PROFILE

Integrasi Metode Klasifikasi Dan Clustering dalam Data Mining

Sofi Defiyanti
Teknik Informatika Fakultas Ilmu Komputer
Universitas Singaperbangsa Karawang
Jl. HS. Ronggowaluyo Teluk Jambe Timur
Karawang
Sofi.defiyanti@unsika.ac.id

Mohamad Jajuli
Teknik Informatika Fakultas Ilmu Komputer
Universitas Singaperbangsa Karawang
Jl. HS. Ronggowaluyo Teluk Jambe Timur
Karawang
Mohamad.jajuli@staff.unsika.ac.id

Abstrak - Data mining atau knowledge discovery in database merupakan salah satu teknik yang digunakan untuk mendapatkan pengetahuan baru dengan memanfaatkan jumlah data yang sangat besar. Beberapa teknik di dalam data mining adalah klasifikasi, clustering, asosiasi dan prediksi. Teknik klasifikasi digunakan untuk menemukan model untuk kepentingan tertentu. Sedangkan clustering merupakan teknik data mining untuk mengelompokkan data berdasarkan kemiripan.
Tingkat akurasi pada masing masing teknik memiliki perbedaan dari setiap model yang dihasilkan. Tingkat akurasi yang baik terjadi jika mendekati angka 100% dengan arti bahwa model yang dihasilkan menunjukkan hasil yang tepat dalam pembangunan modelnya. Integrasi metode klasifikasi dan klustering dalam data mining diharapkan dapat meningkatkan akurasi yang didapat.
Integrasi metode klasifikasi dan clustering dalam data mining dengan memanfaatkan data yang bervolume besar yaitu dengan 1.025.010 record dan memiliki 10 atribut input dilakukan dengan memanfaatkan 10 fold cross validation untuk pengujicobaan evaluasi model yang telah dilakukan. Terdapat perbedaan hasil antara integrasi metode klasifikasi dan clustering dalam data mining terhadap keakurasian, waktu pembentukan model, kehandalan model dan kinerja model yang telah dihasilkan.

Kata Kunci : Data Mining, Klasifikasi, Clustering, integrasi

I. Pendahuluan

Data mining atau lebih di kenal juga dengan sebutan knowledge discovery in databases (KDD). Data mining merupakan salah satu cara yang digunakan untuk mendapatkan pengetahuan baru dengan memanfaatkan jumlah data yang sangat besar. Beberapa teknik telah dikembangkan dan diimplementasikan untuk mengekstrak pengetahuan dan informasi untuk menemukan pola pengetahuan yang mungkin berguna untuk pengambilan keputusan. Teknik-teknik yang digunakan untuk pengekstrakan pengetahuan dalam data mining adalah pengenalan pola, clustering, asosiasi, prediksi dan klasifikasi.

Klasifikasi adalah pemprosesan untuk menemukan sebuah model atau fungsi yang menjelaskan dan mencirikan konsep atau kelas data, untuk kepentingan tertentu.

Clustering digunakan untuk pengelompokkan data berdasarkan kemiripan pada objek data dan sebaliknya meminimalkan kemiripan terhadap kluster yang lain.

Mengintegrasikan metode clustering dengan klasifikasi didapat hasil model yang didapat memiliki akurasi dan robustbess yang lebih baik jika hanya dilakukan dengan metode klasifikasi saja (1).

Algoritma ID3 dianggap metode yang lebih baik jika dibandingkan oleh algoritma C4.5 untuk kasus *spam mail dataset* (2).

Algoritma *decision tree, naive bayes* dan *artificial neural network* digunakan untuk memprediksi prestasi belajar mahasiswa berdasarkan data akademik pada perguruan tinggi didapat bahwa algoritma *naive bayes* merupakan teknik yang memiliki akurasi yang lebih tinggi dan lebih cepat jika dibandingkan kedua algoritma yang lain (3).

Tingkat akurasi dari masing-masing model memiliki perbedaan dari setiap model yang telah di lakukan pembelajaran. Tingkat akurasi yang baik terjadi jika tingkat akurasi mendekati 100% artinya model yang dihasilkan menunjukkan hasil yang tepat. Model yang memiliki tingkat akurasi paling tinggi yang akan dijadikan perbaikan model terhadap model yang sudah ada

Berdasarkan latar belakang tersebut, maka akan dilakukan penelitian untuk mengintegrasikan metode clustering dengan metode klasifikasi dengan memanfaatkan data yang bervolume besar. Dengan meilihat pada survey yang telah dilakukan oleh Wu, algoritma clustering yang akan dipakai adalah K-means sedangkan algoritma klasifikasi yang akan dipakai adalah algoritma ID3 dan *naive bayes*. Dari penelitian yang akan dilakukan diduga bahwa terdapat perbedaan hasil yang didapat dari perlakuan kalasifikasi dengan integrasi metode clustering dan klasifikasi. Sesuai dengan penelitian yang sudah dilakukan oleh Varun Kumar bahwa integrasi metode clustering dan klasifikasi

memberikan tingkat akurasi yang maksimal dan memiliki *robustness* terhadap *missing data.*

II. Tinjauan Pustaka

Iterative Dichotomiser 3 (ID3)

Algoritma ID3 merupakan algoritma yang digunakan untuk membuat pohon keputusan. Algoritma ini menggunakan konsep entropi infoasi. Dengan langkah kerja sebagai berikut :

1. Perhitungan *information gain* dari setiap atribut dengan menggunakan

$$Gain(S, A) = Entropy(S) - \sum_{v \text{ nilai}(A)} \frac{|S_v|}{|S|} Entropy(S_v)$$

Dimana

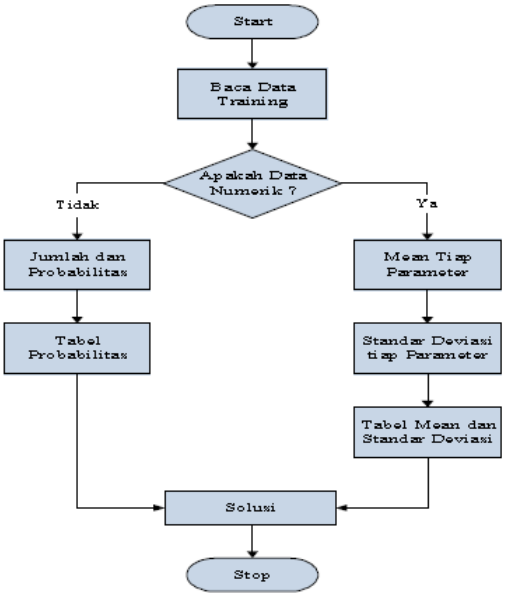
$$Entropy(S) = -P_+ \log_2 P_+ - P_- \log_2 P_-$$

2. Pemilihan atribut yang memiliki nilai information gain terbesar
3. Pembentukan simpul yang berisi atribut tersebut
4. Ulangi proses perhitungan *information gain* akan terus dilakukan sampai semua data telah dimasukkan dalam kelas yang sama. Atribut yang telah dipilih tidak diikutkan lagi dalam perhitungan *information gain.*

Naïve Bayes

Model *Naïve Bayes* adalah klasifikasi statisttik yang dapat digunakan untuk memprediksi suatu kelas. Model *Naïve Bayes* dapat diasumsikan bahwa efek dari suatu nilai atribut sebuah kelas yang diberikan adalah bebas dari atribut-atribut lain.

Naive Bayes memiliki alur seperti pada gambar 1.



Gambar 1. Alur *Naive Bayes* (4)

Kelebihan yang dimiliki oleh *Naïve Bayes* adalah dapat menangani data kuantitatif dan data diskrit, *Naïve Bayes* kokoh terhadap *noise, Naïve Bayes* hanya memerlukan sejumlah kecil data pelatihan untuk mengestimasi parameter yang dibutuhkan untuk klasifikasi, *Naïve Bayes* dapat menangani nilai yang hilang dengan mengabaikan instansi selama perhitungan estimasi peluang, *Naïve Bayes* cepat dan efisiensi ruang

K-means

Algoritma K-means merupakan metode clustering yang paling sederhana dan umum. Ha ini dikarenakan k-means mempunyai kemampuan mengelompokkan data dalam jumlah cukup besar dengan waktu komputasi yang cepat dan efisien. K-means merupakan salah satu algoritma clustering dengan metode partisia yang berbasis titik pusat.

Algoritma K-means memiliki alur sebagai berikut (5) :

Langkah 1: Tentukan berapa banyak *cluster* **k** dari dataset yang akan dibagi.

Langkah 2: teteapkan secara acak data **k** menjadi pusat awal lokasi klaster.

Langkah 3: untuk masing-masing data, temukan pusat *cluster* terdekat. Sehingga masing-masing pusat *cluster* memiliki sebuah subset dari dataset. Yang mewakili bagian dari subset.

Langkah 4: untuk masing-masing *cluster* **k**, temukan pusat luasan *cluster*, dan perbaharui lokasi dari masing-masing pusat *cluster* ke nilai baru dari pusat luasan.

Langkah 5: ulangi langkah ke-3 dan ke-5 hingga data-data pada tiap *cluster* menjadi terpusat atau sesuai.

Teknik Pengujian

Confusion Matrix

Confusion matrix merupakan metode yang menggunakan tabel matriks seperti pada Tabel 2.2, jika data set hanya terdiri dari dua kelas, kelas yang satu dianggap sebagai positif dan yang lainnya negatif (Olson & Yong, 2008)

Tabel 1. Model *Confusion Matrix* (6)

		True Class	
		Positive	Negative
Predicted Class	Positive	true positives count (TP)	false negatives count (FP)
	Negative	false positives count (FN)	true negatives count (TN)

True positives adalah jumlah *record* positif yang diklasifikasikan sebagai positif, *false positives* adalah jumlah *record* negatif yang diklasifikasikan sebagai positif, *false negatives* adalah jumlah *record* positif yang diklasifikasikan sebagai negatif, *true negatives* adalah jumlah *record* negatif yang diklasifikasikan sebagai negative, kemudian masukkan data uji. Setelah data uji dimasukkan ke dalam *confusion matrix*. Setelah data-data telah masuk ke dalam *confusion matrix*

maka dapat dihitung nilai-nilai *sensitivity* (*recall*), *specificity*, *precision* dan *accuracy*. Untuk menghitung digunakan persamaan di bawah ini (Olson & Yong, 2008):

$$Accuracy = \frac{TP + TN}{TP + TN + FN + FP}$$

Ketika klasifikasi tidak menghasilkan binary class, maka *Confusion Matrix* akan berubah menjadi lebih konpleks. Contoh untuk klasifikasi non binary diperlihatkan pada Table 2.3 untuk tiga class dibawah ini :

Tabel 2. *Confusion Matrix untuk Tiga Class:* (a) Actual dan (b) Expected (7)

		Predicted Class			
		a	b	c	Total
Actual Class	a	88	10	2	100
	b	14	40	6	60
	c	18	10	12	40
	Total	120	60	20	

(a)

		Predicted Class			
		a	b	c	Total
Actual Class	a	60	30	10	100
	b	36	18	6	60
	c	24	12	4	40
	Total	120	60	20	

(b)

ROC Curve

ROC (Receiver Operating Characteristics) curve adalah pengujian berdasarkan performanya. ROC mengekspresikan *confusion matrix*. Nilai dari *ROC curve* hanya terdiri dari 0 sampai 1. Semakin nilai *ROC curve* mendekati 1 maka akan semakin baik seperti diperlihatkan pada Tabel 2.4

Tabel 3. *Classfying the accuracy of diagnostic tes* (8)

0,90-1,00	*Excellent Classification*
0,80-0,90	*Good Classification*
0,70-0,08	*Fair Classsification*
0,60-0,70	*Poor Classification*
0,50-0,60	*Failure*

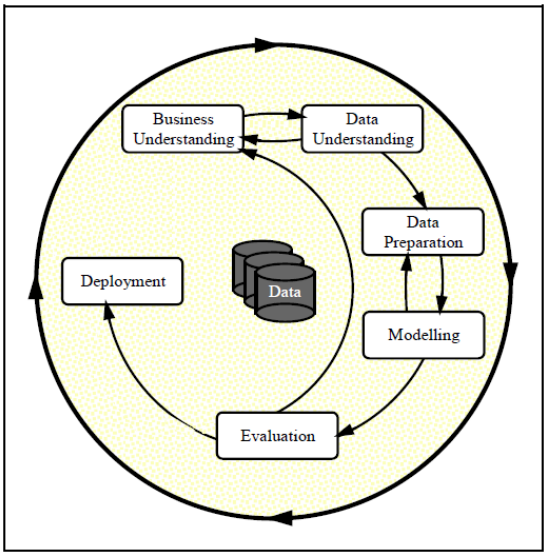
Ukuran Kehandalan Model

Uji coba adalah salah satu cara untuk mengevaluasi keandalan model. Uji coba dapat dilakukan dengan membandingkan nilai prediksi model dengan nilai sebenarnya. Model yang baik adalah model yang mampu memberikan nilai estimasi yang akurat, yaitu nilai y dugaan mendekati nilai y observasi sehingga error mendekati nol. Nilai RMSE (*Root Mean Square Error)* yang semakin kecil menunjukkan model semakin andal dalam memberikan prediksi (9).

$$RMSE = \sqrt{\sum_{i=1}^{n} \frac{\left(Y_i - \hat{Y}_i\right)^2}{n}}$$

III. Metodologi Penelitian

Metode CRISP-DM terdapat enam proses *data mining* seperti tergambarkan dalam Gambar 1 berikut ini :



Gambar 1 Model Crips-DM (10)

a. *Bussiness Understanding*

Pada fase ini berfokus pada pemahaman dan perspektif bisnis proses dari suatu sistem. Yaitu penentuan tujuan proyek, menerjemahkan tujuan, dan menyiapkan strategi untuk penyampaian tujuan.

b. *Data Understanding*

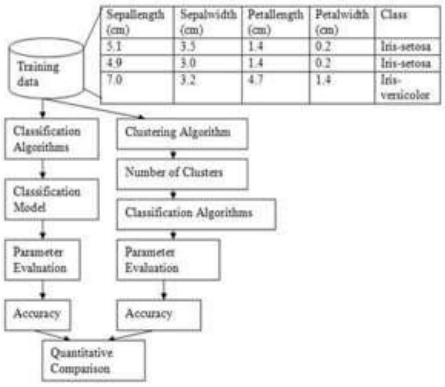
Pada fase ini berfokus pada pembelajaran data yang sudah ada, pengumpulan dan penyeleksian data.

c. *Data Preparation*

Fase persiapan data adalah fase yang terdiri dari pemilihan data, pembersihan data, mengintegrasikan data, dan transformasi data agar dapat dilanjutkan kedalam tahap pemodelan.

d. *Modeling*

Pada fase ini proses yang terjadi adalah pemilihan model yang sesuai. Pemodelan disini dapat dikalibrasi agar mengoptimalkan hasil. Model dengan klasifikasi dan integrasi clustering dengan klasifikasi seperti terlihat pada gambar 3.2 dibawah ini.



Gambar 2 Framework metodelogi penelitian (1)

Model klasifikasi dimulai dari dataset akan dilakukan pemodelan dengan algoritma klasifikasi maka dihasilkan model klasifikasi dari model klasifikasi ini akan muncul parameter evauasi

beserta akurasi dari model algoritma yang diterapkan.

Model integrasi clustering dengan klasifikasi dimulai dari dataset akan dilakukan pemodelan dengan algoritma clustering lalu menentukan banyaknya jumlah cluster setelah didapat model clustering maka dilanjutkan dengan algoritma klasifikasi setelah modelnya keluar maka akan muncul parameter evauasi beserta akurasinya.

e. *Evaluasi*

pada phase ini akan dilakukan proses evaluasi dari phase sebelumnya. Phase evaluasi ini akan dilakukan perbandingan quantitatif dengan mempertimbangkan nilai akurasi, *root mean square error (RMSE)*, ROC *Curve* dan waktu pembangunan model

f. *Deployment*

Pada phase ini proses yang terjadi adalah penyusunan laporan atau presentasi dari pengetahuan yang didapat dari evaluasi pada proses data mining (11).

IV. Hasil yang Dicapai

Bussiness Understanding

Pada tahap ini dilakukan pada pemahaman bisnis proses dari suatu sistem, yaitu penentuan tujuan proyek, menterjemahkan tujuan dan menyiapkan strategi untuk penyampaian tujuan.

Akurasi merupakan kedekatan antara nilai yang diamati serta nilai referensi atau standar yang diterima. Nilai akurasi yang kurang mencerminkan adanya bias sistematis dalam sistem pengukuran diantaranya adalah penggunaan yang tidak sepantasnya. Akurasi diukur dengan jumlah kesalahan dalam pengukuran dibandingkan dengan proporsi pengukuran total (12).

Integrasi metode clastering dan klasifikasi dalam metode data mining dengan menggunakan volume data yang berbeda-beda bertujuan mendapatkan hasil akurasi dan presisi yang sama-sama besar sehingga dapat memenuhi standar yang diterima ataupun yang diingnkan yaitu memiliki nilai akurasi dan presisi yang lebih besar dibandingkan dengan metode klasifikasi saja.

Data Understanding

Data diperoleh dari UCI Mechine Learning Repository (13). Data yang dipergunakan adalah data set yang memiliki jumlah volume yang besar lebih dari 1000 *instance* yaitu *Poker Hand Data Set. Poker Hand Data Set* disumbangkan oleh Robert Cattral pada Universitas Carleton Kanada (14) dengan jumlah instance adalah 1.025.010, Jumlah Atribut 11, tidak memiliki *missing value* didonasikan tanggal 1 Januri 2007 dengan tipe atribut adalah kategori dan integer. Setiap instance didapat dari kartu yang didapat setiap dari lima orang pemain. Jumlah kartu terdiri dari 52 jenis yang terdiri dari jenis dan urutan. Berdasar jenis kartu terdiri dari 4 jenis yaitu *spade*, *heart*, *diamond*, dan *club.* Sedangkan urutan kartu dimulai dari as, 2, 3, dan seterusnya sampai *king. Class* yang dihasilkan dari dataset ini adalah sebanyak 10 *class.*

Data Preparation

Poker Hand Data Set memiliki sepuluh atribut dan satu class output. Lima atribut dengan tipe data numerik dan lima atribut dengantipe data ordinal. Semua atribut yang ada didalam *Poker Hand Data Set* akan dipergunakan keseluruhannya yaitu sepuluh atribut dan satu atribut output.

Tahapan data *preparation* akan dilakukan konversi tipe data agar data set dapat masuk ke tahap selanjutnya yaitu tahap modelling. Yaitu merubah atribut dengan tipe data numerik menjadi ordinal.

Modelling

Fase *modelling* akan dilakukan sesuai dengan gambar 3.2 yaitu dibagi kedalam kempat buah skenario yaitu skeario pertama pemodelan dengan menggunakan klasifikasi yaitu dengan memanfaatkan algoritma *decision tree* (ID3), skenario kedua menggunakan klasifikasi dengan memanfaatkan algoritma *naive bayes*, skenario keempat menggunakan integrasi clustering dengan memanfaatkan algoritma k-mean dan klasifikasi dengan memanfaatkan algoritma *decision tree* (ID3) dan yang terakhir adalah skenario keempat yaitu integrasi clustering dengan menggunakan algortima K-means dan klasifikasi dengan memaanfaatkan algotima *naive bayes.* Skenario tersebut dilakukan agar dapat mengetahui hasil keakurasian.

Evaluation

Setetelah tahapan pemodelan dilakukan maka selanjutnya adalah tahapan evaluasi dengan melihat hasil dari tahapan pemodelan yang telah dilakukan. Evaluasi yang dilakukan menggunakan hasil akurasi, *root mean square error (RMSE)*, ROC *curve* dan waktu pembangunan model.

Tabel 4. Hasil Evaluasi

		Akurasi dalam %	Waktu dalam Detik	RMSE	Roc Curve
Klasifikasi	ID3	97,80%	19,69	0,042	0,81
	Naive Bayes	98,62%	0,67	0,0471	0,986
Integrasi Clustering & Klasifikasi	ID3	76,69%	43,72	0,1555	0,88
	Naive Bayes	90,41%	0,73	0,1371	0,996

Deployment

Tahap terakhir dari metodologi CRIPS-DM adalah tahap *Deployment.* Pada tahap ini akan dilakukan pembuatan laporan hasil kegiatan yang sudah dilakukan. Laporan akhir mengenai pengetahuan yang didapat.

Dari hasil analisis yang telah dilakukan dan telah dievaluasi maka didapatkan hasil bahwa teknik klasifikasi dengan menggunakan algoritma *naive bayes* memiliki akurasi yang lebih tinggi, waktu pembangunan model lebih cepat, kehandalan yang baik dan memiliki kinerja yang baik jika dibandingkan dengan algoritma ID3. Hal ini dikarenakan algoritma naive bayes merupakan algoritma yang memiliki kelebihan kesederhanaan dan memiliki akurasi yang tinggi (15), selain itu naive bayes juga memiliki kecepatan dalam pembuatan model (16), naive bayes juga adalah salah satu algoritma klasifikasi yang mudah digunakan (17).

Sedangkan integrasi teknik clustering dan klasifikasi yang telah dilakukan dengan algoritma K-means untuk clustering dan algoritma ID3 dan *Naive bayes* didapatkan hasil yang tidak terlalu bagus jika dibandingkan dengan teknik klasifikasi saja. Terlihat didalam hasil evaluasi akurasi yang dihasilkan lebih menurun, waktu yang dibutuhkan dalam pembangunan model juga memiliki waktu yang lebih lama, kehandalan model yang dihitung dengan menggunakan RMSE juga memiliki nilai yang lebih tinggi, dan untuk kinerja klasifikasi yang dihitung dengan ROC Curve hanya di algoritma *naive bayes* saja yang mengalami peningkatan, tetapi untuk algoritma ID3 justru mengalamai penurunan. Ini disebabkan karena algoritma ID3 ketidakstabilannya dalam melakukan klasifikasi data apabila terjadi sedikit perubahan pada data training (18). Selain itu penggunaan teknik clustering menggunakan algoritma k-means tidak terlalu cocok dengan jenis data yang di gunakan dalam penelitian ini dikarenakan penelitian ini menggunakan data kategorikal sedangkan k-means lebih unggul digunakan untuk data yang numerik, ini merupakan kelemahan dari algoritma k-means yaitu bekerja baik pada atribut numerik (19).

V. Kesimpulan

Integrasi metode klasifikasi dan clustering dalam data mining dengan memanfaatkan algoritma ID3 dan *naive bayes* untuk teknik klasifikasi dan algoritma K-means untuk teknik clustering menggunakan *poker hand dataset* dengan 11 atribut yang diantaranya adalah 10 sebagai atribut input dan 1 atribut sebagai atribut output atau class. Dengan jumlah data sebesar 1.025.010 dan tidak ada nilai *missing value.* Dilakukan dengan baik dengan memanfaatkan metodologi CRIPS-DM.

Pengujicobaan terhadap penelitian integrasion metode clustering dengan klasfikasi menggunakan *10 fold cross validation*.

Evaluasi model yang telah dihasilkan didapat bahwa yang memiliki akurasi tertinggi adalah skenario kedua yaitu metode klasifikasi dengan algoritma *naive bayes* dengan 98,62%, untuk waktu pembentukan model yang paling cepat adalah skenario keempat yaitu metode integrasi clustering dengan algoritma K-means dan metode klasifikasi dengan algoritma *naive bayes* dengan kecepatan 0,52 detik, untuk kehandalan model dengan memanfaatkan perhitungan RMSE didapat model yang paling handal adalah skenario pertama yaitu klasifikasi dengan metode ID3 dengan nilai RMSE adalah 0,042, dan yang terakhir evaluasi untuk mengetahui kinerja model yang telah dihasilkan didapat bahwa skenario keempat yaitu integrasi metode clustering dengan algoritma k-means dengan metode klasifikasi dengan algoritma *naive bayes* merupakan model yang memiliki kinerja terbaik diantaraskenario yang lain dihitng dari nilai ROC *curve* yaitu dengan nilai 0,995 yang berada di kelompok *Excellent Classification.*

Terdapat perbedaan hasil antara integrasi metode clustering dengan klasifikasi terhadap keakurasian, waktu pembentukan model, kehandalan model dan kinerja model yang telah dihasilkan

Daftar Pustaka

1. *Knowledge Discovery from Database Using an Integration of Clustering and Classification.* **Varun, Kumar and Nisha, Rathee.** 2011, International Journal of Advanced Computer Science and Applications (IJACSA), Vol. 2 No. 3, p. 29.
2. **Defiyanti, Sofi and Pardede, D. L. Crispina.** *Perbandingan Kinerja Algoritma ID3 dan C4.5 Dalam Klasifikasi Spam-Mail.* Depok : Universitas Gunadarma, 2010.
3. **Defiyanti, Sofi.** *Analisis dan Prediksi Prestasi Belajar Mahasiswa Menggunakan Teknik Data Mining Study Kasus Fasilkom UNSIKA.* Karawang : Universitas Singaperbangsa Karawang, 2013.
4. *Penerapan Algoritma Naive Bayes untuk Mengklasifikasikan Data Nasabah Asuransi.* **Bustami.** Aceh : TECHSE Jurnal Penelitian Teknik Informatika.
5. **Irwansyah, Edy.** *Advance Clustering : Teori dan Aplikasi.* jakarta : Bina Nusantara University, 2015. 978-602-280-500-7.
6. **Olson, David and Yong, Shi.** *Pengantar Ilmu Penggalian Data Bisnis (Chriswan Sungkono, Penerjemah).* Jakarta : Selemba Empat, 2008.
7. **Witten, Ian H. Eibe, Frank and Hall, Mark A.** *Data Mining Practical Machine Learning Tools and Techniques Third Edition.* s.l. : The morgan Kaufman Series in data Management Systems, 2011.

8. **Gorunescu, Florin.** *Data Mining Concepts, Model and Techniques.* s.l. : Springer, 2011.
9. **Nawari.** *Analisis Regresi.* Jakarta : PT Elex Media Komputindo, 2010.
10. **Nisbet, Robet, Elder IV, John and Liner, Gary.** *Handbook of Statictical Analysis and Data Mining Applications.* s.l. : Elsevier Inc, 2009.
11. **Budiman, Irwan.** *Data Clustering Menggunakan Metodologi CRIPS-DM untuk pengenalan Pola Proporsi Pelaksanaan Thidharma.* Semarang : Program Pasca Sarjana Universitas Diponegoro , 2012.
12. **Evans, James R and Lindsay, William M.** *An Introduction to Six Sigma & Process Improvment.* Singapore : Thomson, 2007.
13. **Aha, David, Asuncion, Arthur and Newman, David.** UCI Machine Learning Repository. *UCI Machine Learning Repository.* [Online] April Rabu, 2015. http://archive.ics.uci.edu/ml/index.html.
14. **Cattral, Robert.** Poker Hand Data Set. *UCI Machine Learning Repository.* [Online] April Rabu, 2015. https://archive.ics.uci.edu/ml/datasets/Poker+Hand.
15. **Rish, Irna.** *IBM Research Report "An Empirical Study of the Naive Bayes Classifier".* s.l. : IBM Research DIvision, 2001.
16. *Klasifikasi Teks dengan Naive Bayes Classifier (NBC) untuk Pengelompokan Teks Berita dan Abstract Akademis.* **Hamzah, Amir.** Yogyakarta : akprind Yogyakarta, 2012. Prosiding Seminar Nasional Aplikasi Sains & Teknologi (SNAST) Periode III. pp. B-269.
17. **Dunham, Margret.** *Data Mining : Introductory and Advance Topic.* s.l. : Pearson Education, 2006.
18. *Perancangan Sistem Pendukung Keputusan (SPK) untuk Menentukan Kelaiklautan Kapal.* **Setiawan, Bambang, Widjaja, Raden Sjarief and Nugroho, Setyo.** Surabaya : Program Studi MMT-ITS, 2009. Prosiding Seminar Nasional Manajemen Teknologi X. pp. C-14-1.
19. **Berkhin, Pavel.** A Survey of Clustering Data Mining Techniques. *Grouping Multidimensional Data.* s.l. : Springer, 2006, pp. 25-71.
20. **Turban, E, Aronson, J.E and Liang, T.** *Decision Support System and Intelligent System.* Upper Saddle River : Pearson Education Inc, 2005.
21. **Matteucci, Mateo.** *Clustering An Introduction.* Milan : Politecnico, Milano, 2002.
22. *Root Mean Square Error (RMSE) or Mean Absolute Error (MAE)? Argument Against Avoiding RMSE in The Literature.* **Chai, T and Draxler, R.** 2014, Geoscientific Model Development, p. 1247.
23. *Perancangan Sistem Pendukung Keputusan (SPK) untuk Menentukan Kelaiklautan Kapal.* **Bambang Setiawan, Raden Sjarief Widjaja, Setyo Nugroho.** Surabaya : Prosiding Seminar Nasional Manajemen Teknologi X, 2009. 978-979-99735-8-0.