Analisis Performa Algoritma K-NN Dan C4.5 Pada Klasifikasi Data Penduduk Miskin

Femi Dwi Astuti[*1], Mohammad Guntara[2]
[1,2] Teknik Informatika, STMIK AKAKOM, Yogyakarta
e-mail: [*1]femi@akakom.ac.id , [2]guntara@akakom.ac.id

Abstrak

Status kemiskinan penduduk di Kecamatan Bantul diklasifikasikan melalui 11 aspek. Jumlah nilai dari keseluruhan aspek akan menentukan kelas kemiskinan diantaranya kelas miskin, sangat miskin dan rawan miskin. Klasifikasi dengan model tersebut membuat hasil pengelompokan kurang akurat sehingga perlu dicoba klasifikasi dengan model yang lain. Analisis performa klasifikasi data penduduk miskin pada penelitian ini dikerjakan menggunakan metode klasifikasi K-NN dan C4.5. Kedua algoritma klasifikasi akan dibandingkan performanya melalui uji akurasi, precision dan recall.
Hasil analisis perbandingan performa algoritma K-NN dengan parameter setting k=1 memiliki performa yang paling baik dibandingkan dengan nilai k=10, 100, 1000 maupun algoritma C4.5. Hasil nilai Accuracy sebesar 94,71%, precision sebesar 84,96% dan recall sebesar 83,6%.

Kata kunci*— Klasifikasi, C4.5, K-NN, Kemiskinan*

1. PENDAHULUAN

Klasifikasi data penduduk miskin sudah dilakukan oleh BKKPPKB (Badan Kesejahteraan Keluarga Pemberdayaan Perempuan dan Keluarga Berencana) Kabupaten Bantul, melalui peraturan Bupati. Penentuan klasifikasi saat ini dilakukan berdasarkan beberapa aspek penentu kemiskinan diantaranya aspek sandang, pangan, papan, kesehatan, pendidikan, kekayaan 1, kekayaan 2, air bersih, listrik dan jumlah jiwa. Proses penentuan status miskin, rawan miskin atau miskin sekali ditentukan berdasarkan jumlah nilai keseluruhan aspek. Semakin tinggi nilainya maka sebuah keluarga dianggap semakin miskin, begitu juga sebaliknya. Dengan prosedur seperti itu, maka penentuan status kemiskinan menjadi tidak akurat sehingga penyaluran bantuan menjadi tidak tepat sasaran. Hal ini dikarenakan semua keluarga dianggap memiliki status yang sama jika masuk dalam range yang sama.

Klasifikasi penduduk miskin melalui teknik data mining sudah banyak dilakukan, tetapi belum ada yang menganalisa teknik dengan algoritma apa yang paling sesuai dan akurat. Analisa berbagai algoritma klasifikasi sudah pernah dilakukan oleh beberapa peneliti tetapi dengan obyek yang berbeda. Perbandingan performance algoritma K-NN, Decision tree dan Naïve Bayes pernah dilakukan melalui performa *accuracy*, MAE, Kappa Statistic, algoritma terbaik didapatkan dari model algoritma Naive Bayes [1]. Perbandingan metode Decision Tree, Nearest Neighbour dan Naïve Bayes yang lain menunjukkan Decision Tree memiliki tingkat akurasi tertinggi [2]. Algoritma Naive Bayesian memiliki akurasi terbaik jika dibandingkan dengan metode Lazy-IBK, Zero-R dan Decision Tree [3]. Algoritma C4.5 dan K-NN dengan parameter k=100 menghasilkan nilai accuracy, kappa statistic dan recall terbaik dibandingkan dengan penggunaan naive bayes pada klasifikasi benih gandum. Nilai accuracy yang diperoleh sebesar 95,24%, kappa statistic 0.929 dan recall 95,24 [4]. Algoritma C4.5

lebih akurat dibandingkan K-NN dan Neural Network pada kasus penerimaan kredit kendaraan bermotor [5]. Selain kasus-kasus tersebut, C4.5 juga terbukti lebih akurat dibanding K-NN pada kasus klasifikasi penyakit diabetes pada penderita jantung [6]. Prediksi lahan kritis dengan metode C4.5, K-NN dan naive Bayes juga menunjukkan C4.5 memiliki akurasi tertinggi [7].

Berdasarkan permasalahan tersebut maka dalam penelitian ini akan di lakukan klasifikasi dengan metode K-NN dan Decision Tree kemudian akan dianalisis performanya berdasarkan nilai *accuracy* terbaik, *precision* terbaik dan *recall*.

2. METODE PENELITIAN

Beberapa metode dan pustaka yang digunakan dalam penelitian ini diantaranya :

2.1. Data mining

Data Mining adalah proses yang mempekerjakan satu atau lebih teknik pembelajaran komputer (*machine learning*) untuk menganalisis dan mengekstraksi pengetahuan (*knowledge*) secara otomatis. Data mining berisi pencarian trend atau pola yang diinginkan dalam database besar untuk membantu pengambilan keputusan diwaktu yang akan datang [8].

2.2. Klasifikasi

Klasifikasi merupakan suatu pekerjaan menilai objek data untuk memasukkannya ke dalam kelas tertentu dari sejumlah kelas yang tersedia. Dalam klasifikasi ada dua pekerjaan utama yang dilakukan, yaitu (1) pembangunan model sebagai prototipe untuk disimpan sebagai memori dan (2) penggunaan model tersebut untuk melakukan pengenalan/ klasifikasi/ prediksi pada suatu objek data lain agar diketahui di kelas mana objek data tersebut dalam model yang sudah disimpannya.

2.3. K-NN

Algoritma k-nearest neighbor (K-NN) adalah sebuah metode untuk melakukan klasifikasi terhadap objek berdasarkan dari data pembelajaran yang jaraknya paling dekat dengan objek tersebut. Algoritma metode k-nearest neighbor (KNN) bekerja berdasarkan jarak terpendek dari *query instance* ke *training sample* untuk menentukan KNN-nya. *Training sample* diproyeksikan ke ruang berdimensi banyak, dimana masing-masing dimensi merepresentasikan fitur dari data. Ruang ini dibagi menjadi bagian-bagian berdasarkan klasifikasi training sample.

Sebuah titik pada ruang ini ditandai oleh kelas jika kelas (c) merupakan klasifikasi yang paling banyak ditemui pada K (tetangga terdekat dari titik tersebut). Dekat atau jauhnya lokasi (jarak) biasanya dihitung berdasarkan jarak Euclidean [9] dengan rumus seperti persamaan (1).

$$d(X_i, X_j) = \sqrt{\sum_{l=1}^{N} \left(\text{diff}(X_{il}, X_{jl}) \right)^2} \quad (1)$$

Dengan :

X_{il} = data testing ke-i pada variabel ke-l
X_{jl} = data training ke-i pada variabel ke-l
$d(X_i,X_j)$ = jarak
N = dimensi data variabel bebas
$diff(X_{il},X_{jl})$ = difference atau ketidaksamaan

Langkah yang digunakan dalam metode K-Nearest Neighbor :

1. Tentukan parameter K (jumlah tetangga paling dekat).
2. Hitung kuadrat jarak euclid masing – masing objek terhadap data sample yang diberikan.
3. Urutkan objek – objek kedalam kelompok yang memiliki jarak terkecil.

4. Kumpulkan kategori Y (Klasifikasi nearest neighbor).
5. Dengan kategori nearest neighbor yang paling banyak, maka dapat diprediksikan nilai query instance yang telah dihitung.

2.4. Decision tree

Algoritma decision tree yang digunakan untuk membangun pohon keputusan pada penelitian ini adalah C4.5. Algoritma C4.5 menurut [10] :

1. Pilih atribut sebagai root
2. Buat cabang untuk masing-masing nilai
3. Bagi kasus dalam cabang
4. Ulangi proses untuk masing-masing cabang sampai semua kasus pada cabang memiliki kelas yang sama.

Untuk memilih atribut sebagai akar, didasarkan pada nilai gain tertinggi dari atribut-atribut yang ada. Untuk menghitung gain digunakan persamaan (2) :

$$Gain(S, A) = Entropy(S) - \sum_{i=1}^{n} \frac{|S_i|}{|S|} * Entropy(S_i) \quad (2)$$

Dengan :

$\{S_1, S_2, S_3, \ldots, Sn\}$ = partisi S, sesuai dengan nilai atribut A
A = Atribut
n = jumlah partisi atribut A
$|S_i|$ = jumlah kasus pada partisi S_i
$|S|$ = jumlah kasus dalam S

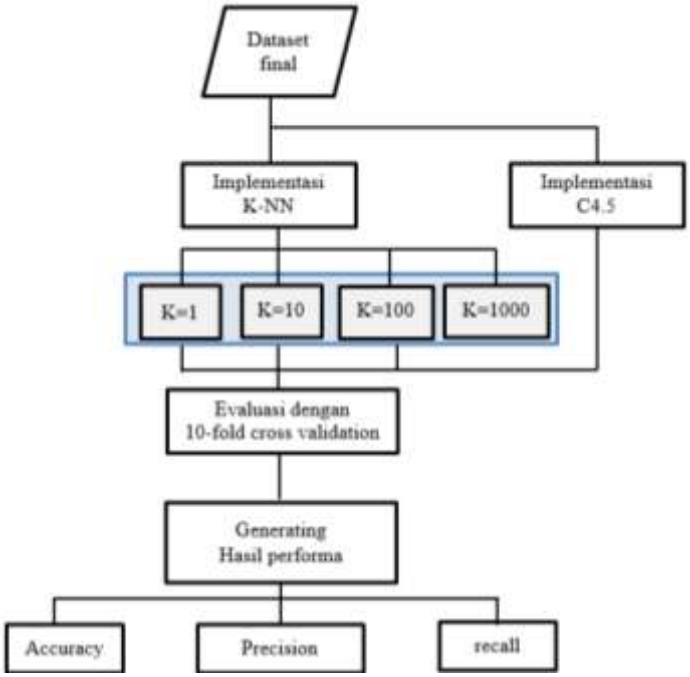
Sedangkan perhitungan nilai entropy menggunakan persamaan (3) :

$$Entropy(S) = \sum_{i=1}^{n} -pi * log_2 pi \quad (3)$$

Dengan :
S = Himpunan kasus
n = jumlah kasus pada partisi S
pi = proporsi $S_i$terhadap S

2.5. Gambaran umum penelitian

Gambaran umum penelitian dapat dilihat pada Gambar 1.



Gambar 1. Gambaran umum penelitian

Mengacu pada Gambar 1., dataset final yang digunakan adalah data penduduk miskin Kecamatan Bantul Kabupaten Bantul yang terdiri dari 1313 record dan 11 atribut. Pengujian pada peneltian ini dimulai dengan menginputkan dataset kemudian dataset di - import ke *machine learning* Rapidminer 7.0.0. Dataset di uji coba dengan mengimplementasikan beberapa algoritma klasifikasi seperti K-NN dengan parameter k yang digunakan adalah k=1,10,100, 1000 dan *Decision Tree* C4.5 dengan *parameter setting gain ratio*. Setelah diterapkan ke algoritma KNN dan C4.5, selanjutnya algoritma divalidasi menggunakan teknik 10-fold *cross validation*. Setelah divalidasi kemudian *machine learning* Rapidminer akan menghasilkan output berupa performa algoritma. Performa yang dibandingkan yaitu accuracy, precision dan recall.

2.6. Dataset Penduduk Miskin

Data yang digunakan dalam penelitian ini adalah dataset yang diperoleh dari BKKPPKB dengan indikator keluarga miskin sebanyak 11 aspek diantaranya : aspek sandang, pangan, papan, penghasilan, pendidikan, kesehatan, kekayaan 1, kekayaan 2, air bersih, listrik dan jumlah jiwa. Contoh dataset dapat dilihat pada Tabel 1.

Tabel 1. Data Penduduk Miskin

No	Nama	Dukuh	RT	1	2	3	4	5	6	7	8	9	10	11
1	Abdullah	Ringinharjo	3	0	9	0	35	0	0	5	0	0	0	5
2	Abdur Rahim Salman	Ringinharjo	4	0	9	0	35	0	0	5	6	0	0	0
3	Ade Kurniawan	Ringinharjo	1	0	9	0	35	0	0	5	6	0	0	0
4	Adi Asrori	Bantul	4	0	0	0	35	6	6	5	0	0	0	0
5	Adi Pawiro	Palbapang	4	0	9	0	35	0	0	5	6	0	0	0
6	Adi Sucipto	Bantul	4	0	0	0	35	0	6	5	6	4	0	5
7	Adi Sumarto	Palbapang	2	0	9	0	35	0	0	5	6	0	0	0
8	Sumidi	Bantul	2	0	9	0	35	0	0	5	6	0	0	0
9	Adi Suwarno	Trirenggo	7	0	9	9	35	0	0	5	0	0	0	0
..	...	..	..	..	..	..	..	..	..	..	..	..	..	..
..	...	..	..	..	..	..	..	..	..	..	..	..	..	..
..	...	..	..	..	..	..	..	..	..	..	..	..	..	..
1313	Boinah	Palbapang	7	12	9	0	35	6	0	5	0	0	0	0

Keterangan :
1. Pangan
2. Sandang
3. Papan
4. Penghasilan
5. Kesehatan
6. Pendidikan
7. Kekayaan 1
8. Kekayaan 2
9. Air Bersih
10. Listrik
11. Jumlah Jiwa

3. HASIL DAN PEMBAHASAN

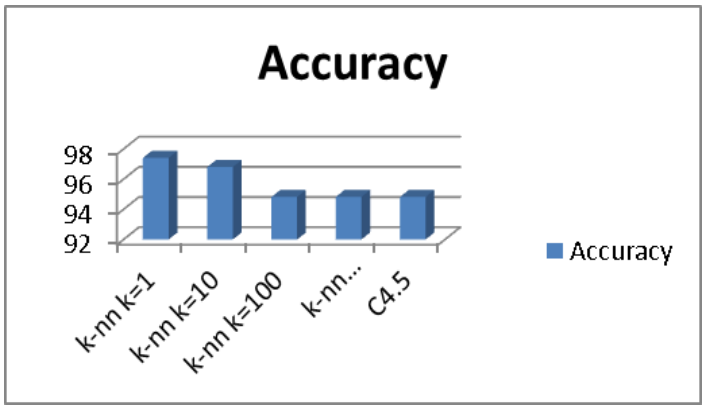
3.1. Hasil perbandingan Accuracy K-NN dan C4.5

Hasil perbandingan *accuracy* K-NN dan C4.5 dapat dilihat pada Tabel 2. Berdasarkan Tabel 2. tersebut, dapat dilihat nilai *accuracy* dari dua algoritma yaitu K-NN dengan *parameter setting* yang berbeda-beda dan algoritma C4.5.

Tabel 2. Hasil perbandingan Accuracy

Metode		Accuracy
K-NN	k=1	97,41
	k=10	96,8
	k=100	94,82
	k=1000	94,82
C4.5		94,82

Algoritma K-NN dengan *parameter setting* k=1 menghasilkan nilai *accuracy* sebesar 97,41%. *Parameter setting* k=10 menghasilkan nilai accuracy sebesar 96,8%. *Parameter setting* k=100 menghasilkan nilai *accuracy* sebesar 94,82%. *Parameter setting* k=1000 menghasilkan nilai *accuracy* sebesar 94,2%. *Accuracy* untuk k=100 dan k=1000 memiliki nilai yang sama. Sedangkan nilai *accuracy* untuk algoritma C4.5 sebesar 94,82%. Grafik perbandingan nilai *accuracy* dengan beberapa metode dapat dilihat pada Gambar 2.



Gambar 2. Grafik perbandingan Accuracy

Berdasarkan Gambar 2. dapat dilihat perbandingan nilai accuracy dengan lebih mudah. Dari gambar tersebut, dapat dilihat bahwa nilai accuracy tertinggi diperoleh saat menggunakan algoritma klasifikasi K-NN dengan parameter setting k=1. Nilai accuracy terendah diperoleh ketika menggunakan algoritma K-NN dengan parameter setting k=100 dan k=1000 serta algoritma C4.5 yaitu sebesar 94,82%. Berdasarkan hasil tersebut dapat disimpulkan bahwa algoritma K-NN dengan parameter setting k=1 memiliki performa terbaik dibandingkan dengan algoritma C4.5 maupun algoritma yang sama dengan parameter lain dari sisi *accuracy*.

3.2. Hasil Perbandingan Precision K-NN dan C4.5

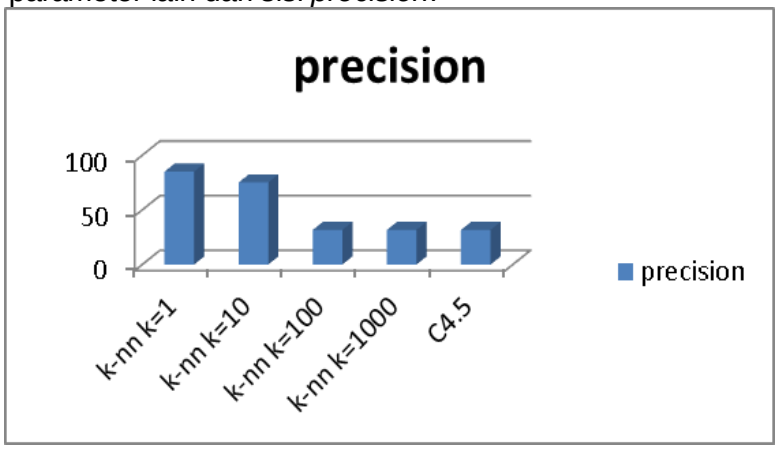
Hasil perbandingan *Precision* K-NN dan C4.5 dapat dilihat pada tabel 3. Berdasarkan tabel 3 tersebut, dapat dilihat nilai *Precision* dari dua algoritma yaitu K-NN dengan *parameter setting* yang berbeda-beda dan algoritma C4.5.

Tabel 3. Hasil perbandingan Precision

Metode		Precision
K-NN	k=1	84,96
	k=10	75,05
	k=100	31,61
	k=1000	31,61
C4.5		31,61

Algoritma K-NN dengan *parameter setting* k=1 menghasilkan nilai *precision* sebesar 84,96%. *Parameter setting* k=10 menghasilkan nilai *precision*sebesar 75,05%. *Parameter setting* k=100 menghasilkan nilai *precision* sebesar 31,61%. *Parameter setting* k=1000 menghasilkan nilai *precision* sebesar 31,61%. Sedangkan nilai *precision* untuk algoritma C4.5 sebesar 31,61%. *Precision* untuk K-NN dengan k=100 dan k=1000 memiliki nilai yang sama dengan nilai precision C4.5. Grafik perbandingan nilai *precision* dengan beberapa metode dapat dilihat pada Gambar 3.

Dari Gambar 3. dapat dilihat bahwa nilai *precision* tertinggi diperoleh saat menggunakan algoritma klasifikasi K-NN dengan parameter setting k=1. Nilai *precision* terendah diperoleh ketika menggunakan algoritma K-NN dengan parameter setting k=100 dan k=1000 serta algoritma C4.5 yaitu sebesar 31,61%. Berdasarkan hasil tersebut dapat disimpulkan bahwa algoritma K-NN dengan parameter setting k=1 memiliki performa terbaik dibandingkan dengan algoritma C4.5 maupun algoritma yang sama dengan parameter lain dari sisi *precision*.



Gambar 3. Grafik perbandingan Precision

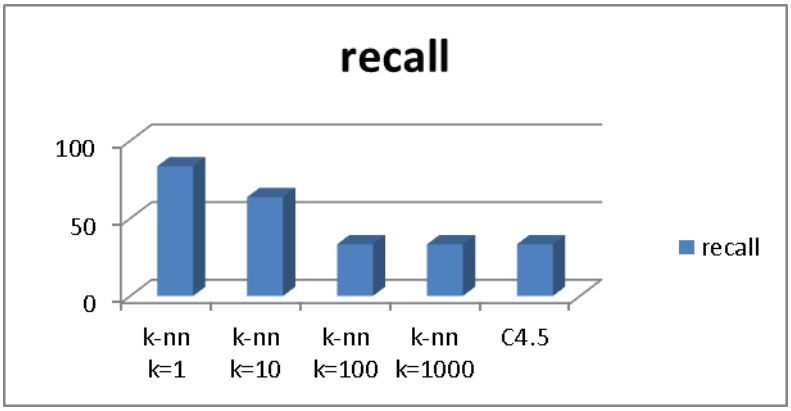
3.3. Hasil Perbandingan Recall K-NN dan C4.5

Hasil perbandingan r*ecall* K-NN dan C4.5 dapat dilihat pada Tabel 4. Berdasarkan Tabel 4 tersebut, dapat dilihat nilai r*ecall* dari dua algoritma yaitu K-NN dengan *parameter setting* yang berbeda-beda dan algoritma C4.5. Algoritma K-NN dengan *parameter setting* k=1 menghasilkan nilai r*ecall* sebesar 83,6%. *Parameter setting* k=10 menghasilkan nilai r*ecall* sebesar 63,73%. *Parameter setting* k=100 menghasilkan nilai r*ecall* sebesar 33,33%. *Parameter setting* k=1000 menghasilkan nilai r*ecall* sebesar 33,33%.

Tabel 4. Hasil Perbandingan Recall

Metode		**Recall**
K-NN	k=1	83,6
	k=10	63,73
	k=100	33,33
	k=1000	33,33
C4.5		33,33

Sedangkan nilai r*ecall* untuk algoritma C4.5 sebesar 33,33%. *Recall* untuk K-NN dengan k=100 dan k=1000 memiliki nilai yang sama dengan nilai r*ecall* C4.5. Grafik perbandingan nilai r*ecall* dengan beberapa metode dapat dilihat pada Gambar 4.



Gambar 4. Grafik perbandingan Recall

Dari Gambar 4 dapat dilihat bahwa nilai r*ecall* tertinggi diperoleh saat menggunakan algoritma klasifikasi K-NN dengan *parameter setting* k=1. Nilai r*ecall* terendah diperoleh ketika menggunakan algoritma K-NN dengan parameter setting k=100 dan k=1000 serta algoritma C4.5 yaitu sebesar 33,33%. Berdasarkan hasil tersebut dapat disimpulkan bahwa algoritma K-NN dengan *parameter setting* k=1 memiliki performa terbaik dibandingkan dengan algoritma C4.5 maupun algoritma yang sama dengan parameter lain dari sisi r*ecall*.

4. KESIMPULAN

Berdasarkan hasil dan pembahasan pada bab-bab sebelumnya maka dapat diambil kesimpulan bahwa algoritma K-NN dengan parameter *setting* k=1 memiliki performa yang lebih baik dibandingkan dengan nilai k=10, 100, 1000 maupun algoritma C4.5 dengan nilai *accuracy* 94,71%, *precision* sebesar 84,96% dan *recall* sebesar 83,6%.

5. SARAN

Penelitian selanjutnya diharapkan dapat membandingkan metode yang lain agar diperoleh metode yang paling tepat akurat. Metode yang sudah digunakan dicoba untuk diterapkan pada dataset yang lain.

UCAPAN TERIMA KASIH

Penulis mengucapkan terima kasih kepada **STMIK AKAKOM** yang telah memberi **"dukungan financial"** terhadap penelitian ini.

DAFTAR PUSTAKA

[1] Yusa M, Utami E, Luthfi E, T, 2016, Analisis Komparatif Evaluasi Performa Algoritma Klasifikasi pada Readmisi Pasien Diabetes, *Jurnal Buana Informatika, Volume 7, Nomor 4, Oktober 2016, Hal : 293-302*.

[2] Sartika Dewi, Sensuse D, I, 2017, Perbandingan Algoritma Klasifikasi Naive Bayes, Nearest Neighbour dan Decision Tree pada Studi Kasus Pengambilan

Keputusan Pemilihan Pola Pakaian, *Jatisi, Vol.1 No.2 Maret 2017, ISSN : 1978-1520*, hal 151-160

[3] Fitri S, 2014, Perbandingan Kinerja Algoritma Klasifikasi Naive Bayesian, Lazy-IBK, Zero-R dan Decision Tree-J48, *Jurnal DASI Vol.15 No.1 Maret 2014, ISSN:1411-3201*, hal : 33-37

[4] Fajri, I.N., 2017, Analisis Performa Algoritma Klasifikasi pada Pengelompokan Benih Gandum, *Jurnal Ilmiah DASI vol.18 No.3 September 2017, ISSN : 1411-3201, hlm 11-15.*

[5] Astuti P, 2016, Komparasi Penerapan Algoritma C4.5, K-NN dan Neural Network dalam Proses Kelayakan Penerimaan Kredit kendaraan Bermotor, *Faktor Exacta 991) : 87-101*, ISSN : 1979-276X, Hal 87-101.

[6] Viswanathan K, Mayilvahanam P, Christy P, Performance Comparison for C4.5 and K-NN Techniques on Diabetic in heart problem, *International Journal of Engineering Science and Computing, Volume 6 Issue No.9*, ISSN : 231 3361, Hal : 3095 – 3098

[7] Khotimah N, istiawan D, 2018, Perbandingan algoritma C4.5, Nive bayes dan K-Nearest Neighbour untuk Prediksi Lahan Kritis di Kabupaten Pemalang, *The 7th University Research Colloqium 2018*, STIKES PKU Muhammadiyah, Surakarta, Hal 41-50.

[8] Hermawati, F. A., 2013, *Data Mining*, Andi Offset, Yogyakarta

[9] Kamber, M., & Han, J., 2006, *Datamining; Concepts and Techniques SecondEdition,* San Francisco, MorganKaufmann Publishers.

[10] Larose, Daniel, T., 2005, *Discovering Knowledge in Data: an Introduction to Data Mining*, John Wiley and Sons, USA.