Pengelompokkan Data Profil Dosen Kopertis Wilayah IV Menggunakan Metode Two Step Cluster Analysis

Sri Widaningsih[1)], Agus Suheri[2)]
Prodi Teknik Informatika, Fakultas Teknik, Universitas Suryakancana
Jl. Pasir Gede Raya Cianjur, (0263) 270106
sriwida@unsur.ac.id[1)], agussuheri@unsur.ac.id[2)]

Abstrak

Data profil dosen di Kopertis Wilayah IV merupakan suatu kumpulan data dengan jumlah yang sangat besar, dimana terdapat informasi-informasi yang dapat diambil. Dalam penelitian ini dilakukan pengelompokkan data dosen tersebut dengan menggunakan metode two step cluster analysis untuk melihat segmentasi dan pola data yang ada. Metode two step clustering digunakan karena variabel terdiri dari bermacam tipe data yaitu tipe numerik untuk variabel usia, nilai jabatan fungsional, dan jangka waktu, ; tipe nominal untuk status, dan tipe ordinal untuk pendidikan, dan jabatan fungsional Dari hasil perhitungan dengan SPSS 17 dengan auto clustering berdasarkan nilai rasio BIC, dihasilkan empat kelompok. Ukuran kelompok paling besar yaitu terdapat pada kelompok empat sebesar 41,6%. Karateristik kelompok ini yaitu dosen-dosen sebagian besar berpendikan magister dengan jabfung 150 , berusia sekitar 40 tahun dan baru memiliki jabfung selama empat tahun. Dari gambaran pengelompokkan yang ada dapat dilakukan kebijakan pembinaan terutama dari perguruan tinggi dimana dosen bernaung.

***Kata Kunci** : two step cluster, data mining, pengelompokkan, spss, dosen*

1. Pendahuluan

Kumpulan data profil yang terdapat pada suatu institusi baik yang bersifat komersial seperti perusahaan maupun yang bersifat nirlaba seperti organisasi pemerintah merupakan salah satu aset yang dapat digali informasinya. Data profil tersebut umumnya berjumlah besar dan tersimpan dalam suatu basis data. . Salah satu teknik yang digunakan untuk penggalian informasi yaitu dengan *data mining. Data mining* merupakan proses mengesktrasi informasi dan pola yang berguna dari jumlah data yang besar, disebut juga sebagai *knowledge discovery process* [1]. Begitu pula dengan data dosen di Kopertis IV yang meliputi wilayah Jabar dan Banten. Jumlah dosen yang telah memiliki jabatan fungsional sebanyak 13.000 dosen. Dari kumpulan data dosen yang telah memiliki jabatan fungsional tersebut dapat digali lebih lanjut untuk melihat pola dan karakteristik yang tersembunyi dan belum pernah dilakukan analisis lebih lanjut sehingga dapat mendorong para dosen untuk dapat segera menaikkan jabatan fungsionalnya sesuai dengan jangka waktu yang diberikan.

Salah satu teknik data mining yang digunakan untuk mengelompokkan data adalah dengan *cluster analysis. Cluster analysis* merupakan proses mempartisi sekumpulan objek data (atau observasi) ke dalam himpunan bagian. Objek dalam kelompok memiliki kemiripan satu sama lain, namun berbeda dengan objek dalam kelompok lain [2]. Terdapat beberapa prosedur pengelompokkan yang umum digunakan yaitu *partitioning, hierarchical, two step clustering.* Setiap prosedur memiliki pendekatan yang berbeda dalam mengelompokkan objek-objek yang paling mirip untuk dimasukkan ke dalam suatu kelompok[3] .Dalam *algoritma clustering partitioning* terlebih dahulu dihitung partisi data berdasarkan kemiripan dari data, dan kemudian dipilih salah satu yang mengoptimalkan kriteria. Algoritma semacam ini sangat kompleks. Beberapa algoritama yang termasuk ke dalam algoritma partitioning yaitu k-means dan k-medoids. Pengelompokkan dengan pendekatan hirarki atau sering disebut dengan *hierarchical clustering* membuat dekomposisi sekumpulan objek secara hirarki. Algoritma yang termasuk ke dalam jenis ini yaitu *agglomerative (bottom-up)* dan *divisive (top-down)* [4].

Beberapa jenis tipe data digunakan dalam analisis pengelompokkan yaitu numerik , nominal, dan ordinal. Data numerik merupakan nilai data riil yang diperoleh dari hasil pengukuran atau penghitungan seperti berat badan, jumlah mobil, atau kadar kalsium. Sedangkan data nominal merupakan data berupa bilangan atau lambang-lambang untuk mengelompokkan objek seperti jenis kelamin atau golongan darah. Ciri dari data nominal yaitu memiliki tingkatan yang sama. Sedangkan data ordinal mirip dengan data

nominal tetapi memiliki tingkatan tertentu seperti pendidikan, jabatan, atau tingkat kepuasan. Penentuan jenis data harus dilakukan sebelum menentukan prosedur mana yang akan dipilih.

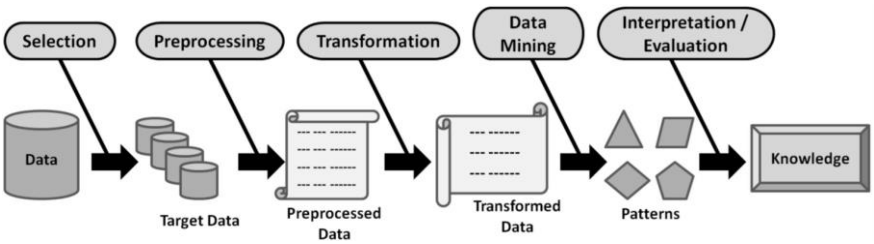
Sebagian besar algoritma pengelompokan tradisional terbatas pada penanganan kumpulan data yang mengandung angka numerik atau atribut kategoris dan algoritma ini umumnya tidak terukur untuk jumlah data yang besar. Tetapi kumpulan data dengan jenis atribut campuran tersebut biasa ditemukan dalam aplikasi data mining pada kehidupan nyata [5]. Sebagai contoh apabila terdapat variabel -variabel dengan data bersifat nominal atau ordinal, jika akan digunakan teknik pengelompokkan seperti k-means maka harus diubah dulu menjadi bentuk numerik. Tetapi metode k-means menggunakan jarak Euclidian dan rata-rata dari atribut objek dalam suatu kelompok. Sehingga algoritma ini lebih cocok untuk tipe data yang dengan skala interval atau rasio. Data dengan jenis kategori atau ordinal bisa dilakukan, tetapi dapat menimbulkan distorsi dan jarang digunakan[3] [6]. Suatu usulan *framework* yang mendukung pengelompokkan data dengan atribut campuran (numerik dan kategori) dengan meminimalkan informasi yang hilang selama pengelompokkan. Proses ini menggunakan entropi untuk menghitung kesamaan yang diekstraksi angka dan bobot [7]. Teknik algoritma pengelompokkan lainnya berdasarkan bobot kesamaan dan paradigma metode filter yang bekerja dengan baik untuk data dengan fitur numerik dan kategoris yang beragam. Deskripsi kelompok dimodifikasi untuk mengatasi keterbatasan data numerik dan memberikan karakterisasi kelompok yang lebih baik[8].

Teknik lain yang dapat digunakan untuk pengelompokkan data campuran yaitu *dengan two- step cluster*. Seperti *k-means,* teknik ini dapat menangani data dengan jumlah yang besar. Jarak yang digunakan dalam metode *two step cluster* adalah jarak *Log-Likelihood* untuk data yang bersifat kategeoris dan jarak Euclidean apabila data terdiri dari nilai numerik. Teknik *two-step clustering* melalui dua tahap: Pada tahap pertama, algoritma melakukan prosedur yang sangat mirip dengan algoritma *k-means*. Pengelompokan awal pengamatan menjadi sub-kelompok kecil dilakukan. Selanjutnya berdasarkan hasil tersebut, dilakukan prosedur pengelompokan aglomeratif hierarkis yang dimodifikasi yang menggabungkan objek secara berurutan untuk membentuk kelompok homogen atau pengelompokkan akhir [3].

Beberapa penelitian menggunakan metode *two-step cluster* dalam membentuk pengelompokkan untuk beberapa tujuan. Penelitian [9] mengelompokkan pengguna internet di kalangan pelajar menjadi enam kelompok karakteristik pengguna internet sesuai dengan kecenderungan karakteristik dominannya. Penelitian mengenai klasifikasi *multiple intelligence* lulusan manajemen juga dikelompokkan menggunakan *two-step clustering* untuk melihat karakteristik setelah bekerja. Variabel yang digunakan merupakan variabel campuran dan terbentuk empat kelompok [10].

2. Metode Penelitian

Penelitian ini menggunakan proses *Knowledge Discovery in Databases* (KDD). Tahapan dalam KDD terdiri dari *selection, pre processing/cleaning, transformation, data mining* dan *evaluation* [11].



Gambar 1. Tahapan Proses KDD [11]

2.1 *Selection (*seleksi*)*

Data yang dipilih merupakan data yang berhubungan dengan fokus penelitian yaitu data dosen-dosen Kopertis Wilayah IV yang telah memiliki jabatan fungsional. Jumlah data sebanyak 13330 data dosen pada tahun 2017. Terdapat beberapa variabel yang terdapat di *database* yang tidak digunakan karena tidak berhubungan dengan penelitian. Berikut ini adalah variabel-variabel dan tipe data yang digunakan.

Tabel 1. Variabel dan Tipe Data

Variabel	Tipe Data	Keterangan
Pendidikan	Ordinal	Pendidikan tertinggi dosen
Jabatan Fungsional	Ordinal	Jabatan Fungsional terakhir dosen
Nilai Kumulatif	Numerik	Nilai kumulatif jabatan fungsional terakhir dosen
Status	Nominal	Status dosen dalam perguruan tinggi
Usia	Numerik	Usia dosen pada tahun 2017
LamaJabfung	Numerik	Jangka waktu jabatan fungsional terakhir hingga saat ini

2.2 *Pre processing* (Pemrosesan Awal)

Pada tahap ini dilakukan pemrosesan awal dan pembersihan data sehingga data siap untuk diolah. Proses pembersihan yang dilakukan diantaranya pemeriksaan pada data yang hilang, duplikasi data, kesalahan penulisan, *outlier*, dan data yang tidak inkonsisten.

2.3 *Transformation* (Trasformasi)

Pada tahap ini dilakukan dua proses transformasi yaitu mengubah nilai dari variabel-variabel yang bersifat ordinal dan nominal menjadi bentuk angka , tanpa mengubah tipe data. Hal ini dilakukan untuk kemudahan pengkodean dalam analisis. Transformasi yang kedua adalah pada jenis data numerik yang distandarisasi dengan teknik normalisasi z score. Variabel cenderung memiliki rentang dan skala satuan yang sangat bervariasi satu sama lain seperti umur dan lama jabfung (dalam tahun) dengan nilai kumultif (tanpa satuan). Untuk beberapa algoritma *data mining*, perbedaan dalam rentang memiliki pengaruh pada hasil. Oleh karena itu, variabel numerik tersebut harus dinormalkan dengan distandarisasikan.

Tabel 2. Transformasi Variabel

Variabel	Nilai	Transformasi
Pendidikan	Doktor(S3)	1
	Spesialis2(SP2)	2
	Magister(S2)	3
	Spesialis1(SP2)	4
	Sarjana(S1)	5
	Diploma(DIV)	6
Jabatan Fungsional	Guru Besar	1
	Lektor Kepala	2
	Lektor	3
	Asisten Ahli	4
Status	Dosen PNS DPK	1
	Dosen Tetap Yayasan	2
	Dosen dengan Perjanjian Kerja	3

2.4 *Data mining*

Teknik pengelompokkan untuk mengolah data dosen yaitu metode *two-step cluster*. Metode ini digunakan karena jumlah data yang besar dan variabel dengan data campuran yaitu numerik, ordinal dan nominal. Tahapan pada *two-step cluster* terdiri dari dua tahap , yaitu [12] :

1. Tahap Pertama/Penggelompokkan awal

   Tahap ini dilakukan dengan pendekatan sekuensial, yaitu objek diamati satu persatu berdasarkan ukuran jarak yang kemudian ditentukan apakah objek tersebut masuk dalam kelompok yang telah terbentuk atau harus membentuk kelompok baru. Algoritma melakukan prosedur yang sangat mirip dengan algoritma *k-means.* Pada langkah ini, diimplementasikan dengan pembentukan *Cluster Feature* (CF) *Tree*. CF-*tree* terdiri dari beberapa tingkat cabang /simpul dan masing-masing cabang berisikan individu objek (*entries*) dari kelompok awal. Tingkatan daun atau daun entri yang terdapat pada cabang merepresentasikan anak kelompok (*subcluster*). Prosedur CF-*tree* dilakukan dengan memilih satu amatan awal secara acak yang akan diukur jaraknya satu persatu dengan amatan lainnya menggunakan ukuran jarak yang telah ditentukan.

   Untuk mendeteksi ada tidaknya pencilan maka dilakukan perhitungan jarak *Log-Likelihood*. Jika terdapat jarak terbesar antar gerombol yang melebihi titik kritis , yaitu:

   C = log(V)

   dengan :

   $$V = \prod_k R_k \prod_m L_m$$

   dan

   R_k = range dari peubah kontinu ke- k

   Lm = banyaknya kategori untuk peubah kategorik ke-m

   Pada jarak Euclidean, data yang memuat pencilan memiliki prosedur yang sama dengan jarak *Log-Likelihood*. Dikatakan pencilan jika jarak Euclidean terbesar antara kelompok tersebut lebih besar dari titik kritis , dengan rumus sebagai berikut :

$$C = 2\left(\sum_{i=1}^{K^A}\frac{\hat{\sigma}_{kl}^2}{K^A}\right)^{\frac{1}{2}}$$

dengan :
R_k= range dari peubah kontinu ke- k
K^A= banyaknya peubah kontinu
σ_{kl}^2= ragam dugaan untuk peubah kontinu ke-i dalam kelompok k .

2. Tahap Kedua/Pengelompokkan Akhir
hasil dari CF-*tree* dikelompokkan dengan analisis kelompok berhierarki dengan metode *agglomerative*. Untuk menghitung banyaknya kelompok dapat dilakukan dengan dua tahapan, yang pertama adalah menghitung *Schwarz's Bayesian Criterion* (BIC) untuk tiap kelompok. Rumus BIC untuk kelompok adalah sebagai berikut:

$$BIC(J) = -2\sum_{j=1}^{J}\xi_j + m_j \log(N)$$

dengan

$$m_j = J\left\{2K^A + \sum_{k=1}^{K^B}(L_k - 1)\right\}$$

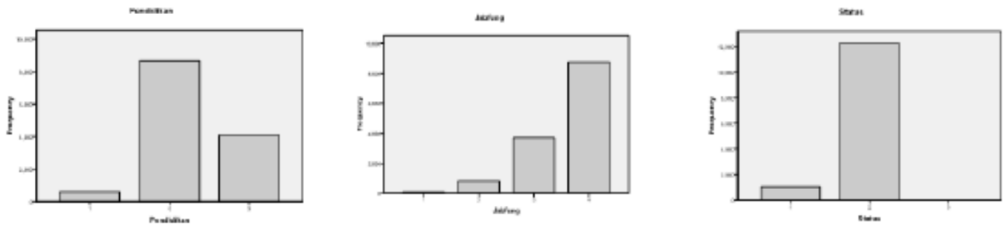
$$\xi_j = -N\left(\sum_{k=1}^{K^A}\frac{1}{2}log\left(\hat{\sigma}_k^2 + \hat{\sigma}_{jk}^2\right)\sum_{k=1}^{K^B}\sum_{l=1}^{L_k}\frac{N_{jkl}}{N_j}log\left(\frac{N_{jkl}}{N_j}\right)\right).$$

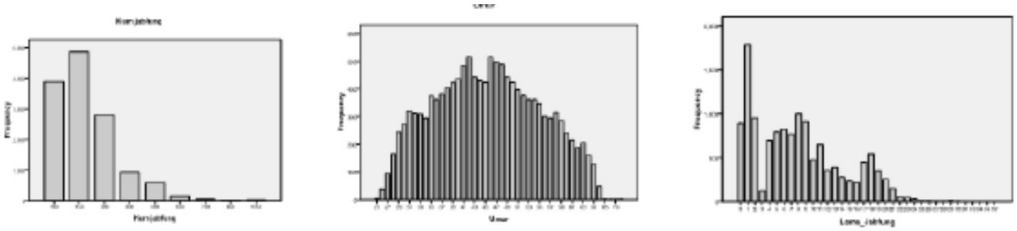
Solusi banyaknya kelompok yang optimal adalah yang memiliki nilai BIC terkecil, tetapi ada beberapa kasus dalam pengelompokkan dimana BIC akan terus meningkat nilainya bila jumlah kelompok semakin meningkat. Maka dalam situasi tersebut, *ratio BIC changes* (rasio perubahan BIC) dan *ratio of distance measure changes* (rasio perubahan jarak) digunakan untuk mengidentifikasi solusi banyaknya kelompok optimal. Solusi untuk banyaknya kelompok optimal akan memiliki *ratio BIC changes* dan *ratio distance measure* yang besar.
Jumlah kelompok yang terbentuk dapat diketahui dengan menggunakan perbandingan antar jarak untuk gerombol, dengan rumus perbandingannya sebagai berikut:

$R(k) = d_{(k-1)}/d_k$
$d_k = l_{k-1} - l_k$
dengan:
$l_v = (r_v \log n - BIC_v)/2$, atau
$l_v = (2r_v - AIC_V)/2$
v = k , k-1
d_{k-1} = jarak jika kelompok k digabungkan dengan k-1 kelompok
R(k) = rasio perubahan jarak.

3. Hasil dan Pembahasan

Sebelum menganalisis hasil pengelompokkan, berikut ini adalah gambaran profil dari dosen-dosen yang telah memiliki jabatan fungsional di lingkungan Kopertis Wilayah IV.





Gambar 2. Statistik Profil Dosen

Analisis kelompok menggunakan software SPSS 17 dengan menggunakan *Auto-clustering* untuk menentukan jumlah kelompok dan kriteria pengelompokkan menggunakan *Schwarz's Bayesian Clustering* (BIC).

Tabel 3. Statistik *Auto Clustering Schwarz's Bayesian Clustering*

Number of Clusters	Schwarz's Bayesian Criterion (BIC)	BIC Change[a]	Ratio of BIC Changes[b]	Ratio of Distance Measures[c]
1	77993.464			
2	52843.312	-25150.152	1.000	1.812
3	39572.023	-13271.289	.528	1.519
4	30877.988	-8694.035	.346	1.887
5	26136.345	-4741.643	.189	1.533
6	23086.761	-3049.584	.121	1.158
7	20469.217	-2617.544	.104	1.540
8	18812.984	-1656.233	.066	1.188
9	17438.646	-1374.339	.055	1.352
10	16454.153	-984.493	.039	1.031
11	15502.899	-951.254	.038	1.281
12	14787.558	-715.341	.028	1.125
13	14165.593	-621.965	.025	1.036
14	13569.363	-596.229	.024	1.106
15	13041.835	-527.528	.021	1.166

Pada tabel 3 nilai BIC terus menurun dan paling kecil adalah jumlah kelompok ke-15, tetapi menurut perhitungan SPPS,jumlah kelompok yang paling optimal adalah empat kelompok, melihat nilai *rasio of distance measure* yang paling tinggi. Distribusi masing-masing kelompok dapat dilihat pada tabel 4 dan tabel pusat kelompok pada tabel 5.

Tabel 4. Distribusi Hasil Pengelompokkan

		N	% of Combined	% of Total
Cluster	1	3059	23.0%	23.0%
	2	3338	25.0%	25.0%
	3	1391	10.4%	10.4%
	4	5540	41.6%	41.6%
	Combined	13328	100.0%	100.0%
Total		13328		100.0%

Tabel 5. Pusat Kelompok Setiap Variabel

	Pendidikan	Jabfung	KumJabfung	Status	Usia	LamaJabfung
Kelompok-1	5(100%)	4(100%)	104,71	2(100%)	46,33	11,74
Kelompok-2	3(66,5%)	3(99,9%)	224,61	2(100%)	48,84	9,38
Kelompok-3	3(63,6%)	2(56,3%)	475,56	1(74,6%)	55,08	11,09
Kelompok-4	3(100%)	4(100%)	140,75	2(100%)	40,48	4,43

Berdasarkan hasil *two-step clustering* pada SPSS dapat terlihat karakteristik setiap kelompok yaitu :

1. Pada kelompok pertama terdiri dari dosen-dosen yang berpendidikan semuanya adalah sarjana, jabfung semuanya asisten ahli dengan kum jabfung 100, status semuanya dosen tetap yayasan, usia sekitar 46 tahun dan jangka waktu jabfung terakhir hingga saat ini yaitu sekitar 11 tahun.
2. Pada kelompok kedua terdiri dari dosen-dosen yang berpendidikan sebagian besar adalah magister, jabfung sebagian besar lektor dengan kum jabfung 200, status semuanya dosen tetap yayasan, usia sekitar 48 tahun dan jangka waktu jabfung terakhir hingga saat ini yaitu sekitar 9 tahun.
3. Pada kelompok ketiga terdiri dari dosen-dosen yang berpendidikan sebagian besar adalah magister, jabfung sebagian adalah lektor kepala dengan kum jabfung 400, status sebagian besar dosen DPK, usia sekitar 55 tahun dan jangka waktu jabfung terakhir hingga saat ini yaitu sekitar 11 tahun.
4. Pada kelompok keempat terdiri dari dosen-dosen yang berpendidikan semuanya adalah magister, jabfung semuanya asisten ahli dengan kum jabfung 150, status semuanya dosen tetap yayasan, usia sekitar 40 tahun dan jangka waktu jabfung terakhir hingga saat ini yaitu sekitar 4 tahun.

4. Simpulan

Dari hasil pengelompokkan dengan menggunakan *two-step clustering* dapat terlihat bahwa jumlah dosen terbanyak terdapat pada kelompok empat yang merupakan dosen-dosen yayasan yang memiliki jabatan fungisonal sebagian besar adalah magister dan baru sekitar 4 tahun memiliki jabatan fungsional. Data-data pengelompkkan ini dapat dijadikan dasar untuk kebijakan pembinaan dosen-dosen tetap yayasan agar secepatnya menaikkan jabatan fungsionalnya sehingga dapat menunjang karir seorang dosen. Selain itu dengan cukup banyaknya dosen dengan pendidikan sarjana, hal tersebut juga harus menjadi perhatian setiap perguruan tinggi.

Daftar Pustaka

[1] K.M Raval. Data Mining Techniques. *International Journal of Advanced Research in Computer Science and Software Engineering*. 2012; Volume 2, Issue 10 : 439-442.

[2] J. Han, M. Kamber, dan J.Pei. Data Mining Concept and Techniques, 3rd ed. Amsterdam: Morgan Kaufmann-Elsevier, 2012 : 444.

[3] M. Sarstedt dan E. Mooi. A Concise Guide to Market Research. Berlin Heidelberg: Springer-Verlag, 2014 : 275.

[4] T. Soni Madhulatha. An Overview On Clustering Methods. *IOSR Journal of Engineering* .2012. Vol. 2(4) : 719-725

[5] H. Mutazinda ,M Sowjanya , dan O.Mrudula. Cluster Ensemble Approach for Clustering Mixed Data. *International Journal of Computer Techniques* .2015. Volume 2 Issue 5: 43-51.

[6] S. Firdaus Md. dan A Uddin. A Survey on Clustering Algorithms and Complexity Analysis. 2015. *International Journal of Computer Science Issues*. Volume 12, Issue 2: 62-85.

[7] J. Lim , J. Jun , S. H Kim dan D. McLeod .A Framework for Clustering Mixed Attribute Type Datasets. 2012. Proceeding of the fourth International Conference on Emerging Databases. Seoul

[8] M. V. Jagannatha R dan B. Kavitha. Clustering the Mixed Numerical and Categorical Dataset using Similarity Weight and Filter Method . *International Journal of Database Theory and Application*. 2012. Vol.5,No. 1 : 121-134

[9] E. M. Husni dan A Fatulloh. Kategorisasi Pengguna Internet di Kalangan Pelajar SD dan SMP Menggunakan Metode Twostep Cluster. Seminar Nasional Aplikasi Teknologi Informasi (SNATi) .2016 .Yogyakarta

[10] Noorazilah, Mohamed dan S. R Awang. The multiple intelligence classification of management graduates using Two-step cluster analysis. *Malaysian Journal of Fundamental and Applied Sciences .2015.Vol.11, No.1 :48-51*

[11] Francesco Gullo. From Patterns in Data to Knowledge Discovery: What Data Mining Can Do. Physics Procedia 62. 2015: 18 – 22.

[12] C.E Mongi. Penggunaan Analisis *Two Step Clustering* untuk Data Campuran. *de CARTESIAN. 2015. Vol. 4, No. 1 : 9-19*