KREA-TIF: JURNAL TEKNIK INFORMATIKA
http://ejournal.uika-bogor.ac.id/krea-tif
p-ISSN: 2338-2910, e-ISSN: 2658-583
Vol 7, No 2, November 2019, pp. 80 - 87
DOI : 10.32832/kreatif.v7i2.2696

Model *Fiture Selection* dalam Penentuan Parameter Pengelompokan Kompetensi SDM IG

Budi Susetyo[1], Puspa Eosina[2], Immas Nurhayati[3], Indupurnahayu[4]

[1,2]Program Studi Teknik Informatika, Fakultas Teknik dan Sains, Universitas Ibn Khaldun Bogor, Indonesia
[3]Program Studi Akuntansi, Fakultas Ekonomi dan Manajemen, Universitas Ibn Khaldun Bogor, Indonesia
[4]Program Studi Manajemen, Fakultas Ekonomi dan Manajemen, Universitas Ibn Khaldun Bogor, Indonesia
*e-mail koresponden: budi.susetyo@uika-bogor.ac.id

Abstrak

Industri geospasial memiliki prospek bisnis yang berkembang pesat di Indonesia, khususnya di sektor swasta. Untuk mengetahui seberapa besar potensi sumberdaya manusia sesuai dengan kompetensi bidang informasi geospasial tersebut dibutuhkan survey dan analisis terkait parameter beberapa parameter kompetensi. Tujuan penelitian ini adalah mencari pengukuran parameter yang paling mempengaruhi pengelompokan kompetensi sumberdaya manusia bidang informasi geospasial. Penelitian ini menggunakan data profil yang telah diolah menjadi 5 kategori index yaitu WEI, EFI, ENI, CFI, dan CPI. dengan jumlah sampel 46 data. Metode yang digunakan adalah k-means clustering untuk pembentukan cluster kompetensi yang selanjutnya dibandingkan di antara 4 ,5 dan 6 cluster. Evaluasi cluster yang dipilih adalah menggunakan Mean intercluster dissimilarity dengan rumus jarak Euclidean. Dihasilkan bahwa pengelompokan paling optimal adalah 4 cluster dengan nilai intercluster terbesar, yaitu 0.45699. Fature subset selection dilakukan terhadap data yang sudah membentuk 4 cluster untuk melihat parameter yang paling berpengaruh. Untuk hal ini, digunakan metode Decision Tree Induction dengan skema Binary Tree. Diperoleh nilai Impurity terkecil pada atribut EFI, yaitu sebesar 0.6857 yang menunjukkan bahwa atribut EFI adalah parameter yang paling berpengaruh dalam menentukan label sebuah data.

Kata kunci: *Decision tree induction, Euclidean distance, Feature subset selection, K-means, mean intercluster dissimilarity*

Abstract

The geospatial industry has a rapidly developing business prospect in Indonesia, especially in the private sector. To find out how much the potential of human resources in accordance with the competence of the geospatial information field, surveys and analyzes are needed regarding the parameters of several competency parameters. The purpose of this study is to find the measurement parameters that most influence the grouping of human resource competencies in the geospatial information field. This study uses profile data that has been processed into 5 index categories, namely WEI, EFI, ENI, CFI, and CPI. with a sample of 46 data. The method used is k-means clustering for the formation of competency clusters which are then compared between 4, 5 and 6 clusters. The chosen cluster evaluation uses Mean Inter-cluster dissimilarity with the Euclidean distance formula. The result is that the most optimal grouping is 4 clusters with the largest inter-cluster value, which is 0.45699. Feature

subset selection is performed on data that has formed 4 clusters to see the most influential parameters. For this case, the Decision Tree Induction method is used with the Binary Tree scheme. Obtained the smallest Impurity value on the EFI attribute, which is equal to 0.6857 which indicates that the EFI attribute is the most influential parameter in determining the label of a data.

Keywords: *Decision tree induction, Euclidean distance, Feature subset selection, K-means, mean intercluster dissimilarity*

PENDAHULUAN

Pemilihan fitur (fiture selection) pada data mining, merupakan bagian praproses penting dalam melakukan mining data agar dihasilkan nilai akurasi yang cukup tinggi dalam proses klasifikasi[1]. Fiture selection, biasa dilakukan pada dataset dengan dimensi yang besar, yaitu yang memiliki fiture data yang banyak [2]. Untuk menghasilkan sebuah model yang baik dengan tujuan tertentu, tidak semua feature berpengaruh, maka harus dilakukan reduksi dimensi dengan cara pemilihan fiture yang paling berpengaruh dalam proses memining data [3,4]. Feature selection banyak diterapkan pada bidang data mining, machine learning, image processing, anomaly detection, bioinformatics, dan natural language processing [5]. Alasan melakukan seleksi fiture antara lain untuk mengurangi beban proses pembelajaran data, membuang fiture yang kurang relevan dan dapat mempengaruhi hasil penambangan data serta meningkatkan kinerja pembelajaran data. Ada dua cara dalam mereduksi feature, yaitu rangking selection dan subset selection[3]. Metode subset selection, selain digunakan untuk mereduksi feature, dapat juga digunakan dalam menganalisis atribut yang paling berpengaruh pada penentuan kelompok sebuah data. Dikaitkan dengan industri geospasial yang saat ini berkembang pesat di Indonesia, serta melihat masih kurangnya SDM-IG di Indonesia, maka dibutuhkan analisis terkait pemetaan kemampuan SDM-IG tersebut tehadap faktor yang paling berpengaruh. Maka penelitian ini memiliki tujuan mencari pengelompokan kemampuan SDM-IG di Indonesia menggunakan metode K-means dalam mencari pengelompokan yang optimal dan menganalisis atribut yang paling berpengaruh terhadap pengelompokan tersebut menggunakan metode *Decision Tree Induction* dengan skema *Binary Tree*. Dalam hal ini, dilakukan pengukuran derajat impurity pada setiap atribut, dimana atribut dengan derajat impurity terendah merupakan faktor yang paling berpengaruh.

Feature Selection memiliki beberapa tipe, yaitu tipe Filter, tipe wrapper dan tipe Embedded. *Feature selection* dengan tipe filter hampir sama dengan selection tipe wrapper dengan menggunakan *intrinsic statistical properties* dari data, namun berbeda dari tipe wrapper dalam hal pengkajian feature yang tidak dilakukan bersamaan dengan pemodelan yang dilakukan. Selection ini dilakukan dengan memanfaatkan salah satu dari beberapa jenis filter yang ada, misalnya: *Individual Merit-Base Feature Selection* dengan *selection criterion: Fisher Criterion, Bhattacharyya, Mahalanobis Distance* atau *Divergence*, *Kullback-Leibler Distance, Entropy* dan lain-lain. Pemilihan metode filter ini umumnya dilakukan pada tahapan *preprocessing* dan mempunyai *computational cost* yang rendah. Sedangkan *tipe Embedded* memanfaatkan suatu *learning machine* dalam proses *feature selection*, di mana feature secara natural dihilangkan, apabila *learning machine* menganggap *feature* tersebut tidak begitu berpengaruh. Beberapa *learning machine* yang bisa digunakan antara lain: *Decision Trees, Random Forests*, dan lain-lain[6].

METODE PENELITIAN

Berbagai metode untuk klasifikasi dalam data mining seperti bayesian, pohon keputusan, rule base, jaringan saraf tiruan dll. memerlukan penyaringan atribut yang tidak relevan [7]. Penyaringan tersebut dapat dilakukan dengan menggunakan teknik pemilihan

fitur. Ada banyak teknik yang dapat dilakukan untuk pemilihan fitur seperti enkapsulasi, filtering, serta teknik embedded [6]. Pemilihan fitur pada data mining, merupakan bagian praproses penting dalam melakukan mining data agar dihasilkan nilai akurasi yang cukup tinggi dalam proses klasifikasi. Fiture selection, biasa dilakukan pada dataset dengan dimensi yang besar, yaitu yang memiliki fiture data yang banyak. Alasan melakukan seleksi fiture antara lain untuk mengurangi beban proses pembelajaran data, membuang fiture yang kurang relevan dan dapat mempengaruhi hasil penambangan data serta meningkatkan kinerja pembelajaran data. Ada dua cara dalam mereduksi feature, yaitu rangking selection dan subset selection [8].

Feature Selection memiliki beberapa tipe, yaitu tipe Filter, tipe wrapper dan tipe Embedded. Feature selection dengan tipe filter hampir sama dengan selection tipe wrapper dengan menggunakan intrinsic statistical properties dari data, namun berbeda dari tipe wrapper dalam hal pengkajian feature yang tidak dilakukan bersamaan dengan pemodelan yang dilakukan. Selection ini dilakukan dengan memanfaatkan salah satu dari beberapa jenis filter yang ada, misalnya: Individual Merit-Base Feature Selection dengan selection criterion: Fisher Criterion, Bhattacharyya, Mahalanobis Distance atau Divergence, Kullback-Leibler Distance, Entropy dan lain-lain. Pemilihan metode filter ini umumnya dilakukan pada tahapan preprocessing dan mempunyai computational cost yang rendah. Sedangkan tipe Embedded memanfaatkan suatu learning machine dalam proses feature selection, di mana feature secara natural dihilangkan, apabila learning machine menganggap feature tersebut tidak begitu berpengaruh. Beberapa machine learning yang bisa digunakan antara lain: Decision Trees, Random Forests, dan lain-lain [9].

a. Data

Data sampel yang digunakan pada penelitian berupa data profil SDM-IG berjumlah 46 record, bersumber dari LSP MAPIN tahun 2019. Data profil yang dikelola, dikelompokan menjadi data dasar SDM-IG, data pendidikan formal, pengalaman kerja, pengalaman proyek, keahlian secara umum, serta keahlian khusus. Kelompok data tersebut kemuadian diolah secara statistic[10]. Data pendidikan formal diolah menjadi WEI-Index, data pengalaman kerja diolah menjadi EFI-Index, pengalaman proyek diolah menjadi ENI-Index, keahlian secara umum diolah menjadi CFI, serta keahlian khusus diolah menjadi CPI. Seluruh data index berskala 0-100. Sebelum dilakukan pemilihan fiture, dilakukan standardisasi [0,1] terhadap data yaitu mengubah skala menjadi bernilai antara 0 dan 1 seperti terlihat pada Tabel 1.

b. Standarisasi

Sebelum data diolah lebih lanjut, pertama-tama dilakukan praproses terhadap data, yaitu standardisasi data dengan mengubah skala data ke skala [0,1]. Standardisasi [0,1] ini, biasa disebut juga dengan *Min-Max Normalization* [11], dilakukan terhadap skala indeks WEI, EFI, ENI, CFI, dan CPI, menggunakan rumus sebagai berikut:

$$x'_i = \frac{x_{i-min_A}}{max_A - min_A}(n\_max_A - n\_min_A) + n\_min_A \quad (1)$$

di mana A menyatakan sebuah field untuk nilai Indeks; x_i adalah data yang ada pada field A;

min_A *dan* max_A masing-masing menyatakan nilai minimum dan nilai maksimum pada field indeks; n_min_A dan n_max_A masing-masing menyatakan nilai minimum dan maksimum baru setelah data pada field A digeser dengan rumus $x_i - min_A$; x'_i menyatakan hasil dari

standardisasi data. Data yang sudah dipraproses selanjutnya dikluster menggunakan metode K-means [12]. Evaluasi hasil kluster dilakukan dengan pengukuran jarak inter-cluster. Semakin besar nilai inter klaster menunjukan pembagian klaster yang baik.

Tabel 1. Indexes sample of HR-GI in Indonesia

NO	HR-Code	WEI Index	EFI Index	ENI Index	CFI Index	CPI Index
1	HR-01	0.42	0.17	0.00	0.40	0.00
2	HR-02	0.08	1.00	0.80	1.00	0.20
3	HR-03	0.11	0.67	0.20	0.00	0.20
4	HR-04	0.13	0.33	0.35	0.00	0.80
5	HR-05	0.08	0.67	0.00	0.60	0.00
6	HR-06	0.25	0.67	0.70	0.20	0.80
7	HR-07	0.21	0.67	0.00	0.20	0.40
8	HR-08	0.32	0.17	0.00	0.00	0.20
9	HR-09	0.37	0.17	0.00	0.00	0.20
10	HR-10	0.71	0.17	0.00	0.00	0.80
11	HR-11	0.00	0.67	0.00	0.00	0.80
12	HR-12	0.00	0.67	0.00	0.00	0.20
13	HR-13	0.00	0.67	0.00	0.00	0.80
14	HR-14	0.14	0.67	0.15	0.00	0.40
15	HR-15	0.22	0.67	0.00	0.00	0.60
16	HR-16	0.14	0.67	0.25	0.40	0.80
17	HR-17	0.28	0.17	0.20	0.00	1.00
18	HR-18	0.15	0.67	0.00	0.20	0.60
19	HR-19	0.66	0.67	0.20	0.00	0.40
20	HR-20	0.19	0.67	0.65	0.20	0.40
21	HR-21	0.06	0.67	0.75	0.00	0.60
22	HR-22	0.04	0.57	0.00	0.00	0.20
23	HR-23	0.07	0.67	0.40	0.20	1.00
24	HR-24	0.46	0.57	0.00	0.00	0.20
25	HR-25	0.19	0.57	0.00	0.00	0.40
26	HR-26	0.22	0.50	0.00	0.00	0.20
27	HR-27	0.01	0.67	0.00	0.20	1.00
28	HR-28	0.08	0.50	0.00	0.00	0.60
29	HR-29	0.13	0.57	0.15	0.20	1.00
30	HR-30	0.32	0.57	0.00	0.00	0.40
31	HR-31	0.09	0.67	0.00	0.40	0.80
32	HR-32	0.13	0.57	0.60	0.00	0.00
33	HR-33	0.34	0.83	0.48	0.00	0.00
34	HR-34	0.36	0.67	0.00	0.00	0.00
35	HR-35	0.41	0.23	0.15	0.00	0.20
36	HR-36	0.17	0.67	0.00	0.20	1.00
37	HR-37	0.11	0.57	0.50	0.00	0.00
38	HR-38	0.03	0.50	0.50	0.00	0.00
39	HR-39	0.16	0.50	0.00	0.00	0.00
40	HR-40	0.92	0.17	1.00	0.00	0.20
41	HR-41	0.31	0.33	1.00	0.00	0.20
42	HR-42	1.00	0.17	0.00	0.00	0.00
43	HR-43	0.41	0.50	0.60	0.00	0.80
44	HR-44	0.07	0.67	0.00	0.40	0.20
45	HR-45	0.53	0.17	1.00	0.00	0.20
46	HR-46	0.92	0.17	0.00	0.20	0.00

Sumber: MAPIN, 2019.

HASIL DAN PEMBAHASAN

Tabel data hasil analisis index kinerja telah dilakukan pada penelitian sebelumnya[13] sebagaimana disajikan pada Tabel 2.

Tabel 2. Nilai Indeks Sampel SDM IG di Indonesia

No	HR-Code	WEI Index	EFI Index	ENI Index	CFI Index	CPI Index	No	HR-Code	WEI Index	EFI Index	ENI Index	CFI Index	CPI Index
1	HR-01	0.42	0.17	0	0.4	0	24	HR-24	0.46	0.57	0	0	0.2
2	HR-02	0.08	1	0.8	1	0.2	25	HR-25	0.19	0.57	0	0	0.4
3	HR-03	0.11	0.67	0.2	0	0.2	26	HR-26	0.22	0.5	0	0	0.2
4	HR-04	0.13	0.33	0.35	0	0.8	27	HR-27	0.01	0.67	0	0.2	1
5	HR-05	0.08	0.67	0	0.6	0	28	HR-28	0.08	0.5	0	0	0.6
6	HR-06	0.25	0.67	0.7	0.2	0.8	29	HR-29	0.13	0.57	0.15	0.2	1
7	HR-07	0.21	0.67	0	0.2	0.4	30	HR-30	0.32	0.57	0	0	0.4
8	HR-08	0.32	0.17	0	0	0.2	31	HR-31	0.09	0.67	0	0.4	0.8
9	HR-09	0.37	0.17	0	0	0.2	32	HR-32	0.13	0.57	0.6	0	0
10	HR-10	0.71	0.17	0	0	0.8	33	HR-33	0.34	0.83	0.48	0	0
11	HR-11	0	0.67	0	0	0.8	34	HR-34	0.36	0.67	0	0	0
12	HR-12	0	0.67	0	0	0.2	35	HR-35	0.41	0.23	0.15	0	0.2
13	HR-13	0	0.67	0	0	0.8	36	HR-36	0.17	0.67	0	0.2	1
14	HR-14	0.14	0.67	0.15	0	0.4	37	HR-37	0.11	0.57	0.5	0	0
15	HR-15	0.22	0.67	0	0	0.6	38	HR-38	0.03	0.5	0.5	0	0
16	HR-16	0.14	0.67	0.25	0.4	0.8	39	HR-39	0.16	0.5	0	0	0
17	HR-17	0.28	0.17	0.2	0	1	40	HR-40	0.92	0.17	1	0	0.2
18	HR-18	0.15	0.67	0	0.2	0.6	41	HR-41	0.31	0.33	1	0	0.2
19	HR-19	0.66	0.67	0.2	0	0.4	42	HR-42	1	0.17	0	0	0
20	HR-20	0.19	0.67	0.65	0.2	0.4	43	HR-43	0.41	0.5	0.6	0	0.8
21	HR-21	0.06	0.67	0.75	0	0.6	44	HR-44	0.07	0.67	0	0.4	0.2
22	HR-22	0.04	0.57	0	0	0.2	45	HR-45	0.53	0.17	1	0	0.2
23	HR-23	0.07	0.67	0.4	0.2	1	46	HR-46	0.92	0.17	0	0.2	0

Sumber: Susetyo, et. al. 2019.

a. Standarisasi

Sebelum data diolah lebih lanjut, pertama-tama dilakukan praproses terhadap data, yaitu standardisasi data dengan mengubah skala data ke skala [0,1]. Standardisasi [0,1] ini dilakukan terhadap skala index WEI, EFI, ENI, CFI, dan CPI, menggunakan rumus sebagai berikut:

Data yang sudah dipraproses selanjutnya dikluster menggunakan K-means methods. Evaluasi hasil kluster dilakukan dengan pengukuran jarak inter-cluster dan antar-cluster. Semakin besar nilai inter klaster dan semakin kecil nilai intra klaster menunjukan pembagian klaster yang baik.

b. K-means Clustering

Pada penelitian ini, dilakukan proses mining clustering dengan 4 cluster ($k = 4$) , 5 cluster ($k = 5$) dan 6 cluster ($k = 6$) serta dicari pembagian cluster yang paling optimal. Pemilihan nilai k ini didasarkan pada teori clustering, bahwa jumlah cluster terbaik untuk pengelompokan data adalah berkisar pada nilai $\sqrt{n}$, di mana n adalah banyaknya data yang akan dikelompokkan. Hasil yang paling optimal diperoleh dengan cara membandingkan nilai inter cluster, intra cluster dari ketiga cara pembagain cluster ini. Proses pembagian cluster, dimulai dengan memberikan nilai inisialisasi untuk sejumlah pusat cluster sesuai dengan jumlah cluster yang diinginkan. Karena data sudah di standardisasi [0,1], maka nilai inisialisasi yang diberikan pada pusat cluster berkisar antara 0 – 1 secara random. Selanjutnya terhadap masing-masing data dilakukan pengukuran jarak terhadap seluruh pusat data dengan rumus sebagai berikut:

$$d(x_i, P_j) = \sqrt{(x_i - P_j)^2} \quad (3)$$

Di mana x_i adalah data dengan $i = 1..n$; n banyaknya data yang akan di cluster;
P_j adalah pusat cluster dengan $j = 1..K$; K adalah jumlah cluster yang akan dibentuk.

Selanjutnya data diberi label terhadap jarak terdekatnya ke pusat dengan label tertentu. Langkah selanjutnya, setiap cluster dihitung ulang pusat datanya terhadap pengelompokan data yang diperoleh dari proses sebelumnya. Proses perhitungan jarak data terhadap pusat data, menentukan jarak terdekat terhadap pusat data dan pemberian label data terhadap jarak terdekatnya terhadap label pusat data tertentu dilakukan berulang sampai data tidak ada yang berubah lagi pelabelannya.

c. Evaluasi cluster

Setelah proses klaster selesai, dilakukan proses evaluasi terhadap cluster yang terbentuk, yaitu dengan mengukur jarak inter cluster. Evaluasi yang dipilih adalah *Mean intercluster dissimilarity* menggunakan rumus jarak Euclidean, dengan rumus sebagai berikut:

$$\underset{C_1, \ldots, C_k}{minimize} \left\{ \sum_{k=1}^{k} \frac{1}{|C_k|} \sum_{i,i' \in C_k} \sum_{j=1}^{p} \left(x_{ij} - x_{i'j} \right)^2 \right\} \quad (4)$$

di mana x_{ij} menyatakan data pada suatu cluster dan $x_{i'j}$ menyatakan data di cluster lain; P adalah banyaknya data pada tiap cluster. Hasil pengukuran inter-cluster terlihat pada Tabel 3.

Tabel 3. Hasil pengukuran nilai intercluster.

Number of cluster	4-cluster (C0, C1, C2, C3)	5-cluster (C0, C1, C2, C3, C4)	6-cluster (C0, C1, C2, C3, C4, C5)
Jumlah data per cluster	(5, 16, 18, 7)	(5, 4, 18, 4, 15)	(4, 2, 18, 7, 11, 4)
Intercluster	**0.45699**	**0.36110**	**0.31876**

Nilai intercluster tertinggi 0.45699 dengan pembagian data menjadi 4 cluster. Maka untuk pengukuran parameter pada langkah selanjutnya, diterapkan pada 4 cluster yang telah terbentuk.

d. Fiture Selection

Pembagian cluster optimal untuk sampel data penelitian telah diperoleh, yaitu data terkluster optimal dengan pembagian cluster sebanyak 4 cluster. Setiap cluster yang terbentuk diberi label class. Selanjutnya, penelitian dilakukan pada analisis parameter yang berpengaruh secara terurut terhadap penentuan pengelompokan sebuah data profil. Teknik yang digunakan adalah fiture selection, sering disebut subset selection. Ada beberapa metode subset selection, di antanyanya adalah Stepwise Forward Selection, Stepwise Backward Elimination, Combination of Forward Selection and Backward Elimination, dan Decision Tree Induction. Pada penelitian ini, kami menggunakan metode Decision Tree Induction dengan skema Binary Tree untuk mencari nilai Impuriry terkecil. Adapun nilai Impurity dihitung menggunakan perhitungan GINI Index dengan rumus sebagai berikut:

$$I_G = 1 - \sum_{j=1}^{c} P_j^2 \quad (4)$$

di mana P_j merupakan nilai peluang dari setiap class yang terbentuk.

Hasil penghitungan derajat Impurity terlihat pada Tabel 3. Dari hasil perhitungan GINI Index, diperoleh nilai Impurity terkecil adalah pada atribut EFI, yaitu sebesar 0.6857 (Lihat

Tabel 4). Hal ini menunjukkan bahwa atribut EFI adalah parameter yang paling berpengaruh dalam menentukan label sebuah data.

Tabel 4. Derajat Impurity

Feature	Impurity
WEI	0.690926
EFI	0.685728
ENI	0.690926
CFI	0.690926
CPI	0.690926

e. Pembahasan

Dengan melihat hasil analisis dari pengolahan data terhadap 46 sampel SDM-IG di Indonesia, ternyata faktor yang masih sangat mempengaruhi nilai kompetensi adalah faktor pendidikan formal. Hal ini dapat terlihat dari hasil pengolahan data bahwa derajat impurity terkecil diperoleh dari nilai indeks EFI sebesar 0.6857. Ini menggambarkan bahwa keahlian dengan kompetensi khusus yang dapat diperoleh melalui training mau pun sertifikasi keahlian masih belum populer di Indonesia. Masih sedikit SDM-IG yang memiliki keahlian khusus di bidang geospasial. Dengan informasi tersebut, dapat dibuat strategi untuk meningkatkan kemampuan dan daya saing SDM-IG di Indonesia terhadap dunia Internasional. Untuk ke depannya, penelitian ini dapat dikembangkan dengan memikirkan cara pengumpulan data SDM-IG yang ada di seluruh Indonesia dan mengcluster ulang data tersebut. Selain itu untuk menghitung ulang derajat Impurity dapat dilakukan menggunakan metode lain untuk mendapatkan hasil yang lebih optimal.

KESIMPULAN

Dari pengelompokkan data sampel SDM-IG, yang berjumlah 46, menggunakan metode K-means, diperoleh bahwa jumlah optimal cluster yang terbentuk adalah 4 cluster dengan nilai intercluster sebesar 0.45699. Dari 4 cluster yang terbentuk ini, dianalisis atribut yang paling mempengaruhi sebuah data untuk masuk pada kelompok dengan label tertentu. Menggunakan metode Decision Tree Induction dengan skema Binary Tree , dari lima atribut yang dianalisis berupa 5 kategori index yaitu WEI, EFI, ENI, CFI, dan CPI, diperoleh bahwa atribut yang paling mempengaruhi adalah EFI dengan derajat impurity sebesar 0.6857.

DAFTAR PUSTAKA

[1] Chandrashekar, G., & Sahin, F. (2014). A survey on feature selection methods. *Computers and Electrical Engineering*, *40*(1), 16–28. https://doi.org/10.1016/j.compeleceng.2013.11.024

[2] Leela Sandhya Rani, Y., Sucharita, V., Bhattacharyya, D., & Kim, H. J. (2016). Performance evaluation of feature selection methods on large dimensional databases. *International Journal of Database Theory and Application*, *9*(9), 75–82. https://doi.org/10.14257/ijdta.2016.9.9.07

[3] Beniwal, S., & Arora, J. (2012). Classification and Feature Selection Techniques in Data Mining, *1*(6), 1–6.

[4] Shang, R., Zhang, Z., Jiao, L., Liu, C., & Li, Y. (2016). Self-representation based dual-graph regularized feature selection clustering. *Neurocomputing*, *171*, 1242–1253. https://doi.org/10.1016/j.neucom.2015.07.068

[5] Bennasar, M., Hicks, Y., & Setchi, R. (2015). Feature selection using Joint Mutual

Information Maximisation. *Expert Systems with Applications*, *42*(22), 8520–8532. https://doi.org/10.1016/j.eswa.2015.07.007

[6] Kumbhar, P. (2016). A Survey on Feature Selection Techniques and Classification Algorithms for Efficient Text Classification, *5*(5), 1267–1275.

[7] Zhao, Z., Wang, L., Liu, H., & Ye, J. (2013). On similarity preserving feature selection. *IEEE Transactions on Knowledge and Data Engineering*, *25*(3), 619–632. https://doi.org/10.1109/TKDE.2011.222

[8] Casella, G., Fienberg, S., & Olkin, I. (2013). *An Introduction to Statistical Learning*. *Springer Texts in Statistics*. https://doi.org/10.1016/j.peva.2007.06.006

[9] Kittler, J. "Feature Selection & Extraction", in Handbook of Pattern Recognition and Image Processing, Tzay Y. Young, King Sun Fu Ed. Academic Press, 1986.

[10] MAPIN. Profil SDM-IG, Masyarakat Penginderaan Jauh Indonesia. 2019.

[11] Han, J., Kamber, M., & Pei, J. (2012). *Data mining: Data mining concepts and techniques* (Third). Morgan Kaufmann. https://doi.org/10.1109/ICMIRA.2013.45

[12] Sullivan, R. (2012). *Introduction to Data Mining for the Life Sciences*. *Journal of Chemical Information and Modeling* (Vol. 53). Springer. https://doi.org/10.1017/CBO9781107415324.004

[13] Susetyo, B., I. Nurhayati, I. Purnahayu, P. Eosina, (2017), "Model Evaluasi Kinerja SDM Geospasial Menggunakan Metode CPI dan CPD Berbasis WebGIS". Prosiding Seminar Nasional XII "Rekayasa Teknologi Industri dan Informasi 2017 Sekolah Tinggi Teknologi Nasional Yogyakarta