ALGORITMA C4.5 UNTUK PENILAIAN KINERJA KARYAWAN

Windy Julianto[1], Rika Yunitarini[2], Mochammad Kautsar Sophan[3]
Universitas Trunojoyo Madura
windy.julianto@gmail.com

Abstrak. S*emakin berkembang perusahaan, semakin dibutuhkan juga kebutuhan tenaga manusiayang produktif. Perusahaan umumnya ingin memberikan kinerja yang terbaik dalam berinteraksi dengan pihak yang terkait, salah satunya yaitu konsumen. Untuk mewujudkan itu, suatu perusahaan akan mencari sumber daya manusia yang melaksanakan pekerjaannya dengan baik. Selama ini penyeleksian pada penilaian kerja masih dilakukan secara manual dengan memberikan kriteria tertentu oleh Human Resource Development. Banyaknya kriteria dan karyawan yang dipekerjakan cukup menyulitkan divisi HRD dalam penyeleksian karyawan terbaik. Selain itu, seringkali pengambil keputusan dinilai secara subyektif karena suatu hubungan tertentu antara pelamar dan pengambil keputusan. Kualitas kerja karyawan juga seringkali berubah-ubah. Ini sangat berpengaruh pada prospek perusahaan kedepannya. Kesibukan seorang manager membuat tatap muka dengan karyawan kurang. Tentu saja ini membuat pengawasan pada kinerja karyawan sulit dilakukan. Untuk itu dibuat suatu sisstem pendukung keputusan dalam melakukan penilaian kinerja karyawan. Sistem ini menggunakan algoritma C4.5 yang menggunakan teknik data mining untuk membuat pohon keputusan, algoritma ini dimulai dengan memasukkan data training ke dalam simpul akar pada pohon keputusan. Data training adalah sampel yang digunakan untuk membangun model classifier dalam hal ini pohon keputusan. Pada proses monitoring karyawan akan dipantau untuk penentuan kenaikan gaji atau promosi jabatan.*

***Kata Kunci**:sistem pendukung keputusan,algoritma C4.5, monitoring kinerja.*

Keputusan seorang manager adalah hal yang sangat vital pada perusahaan untuk dampak selanjutnya. Pada perusahaan kegiatan penilaian kinerja karyawan sulit dilaksakan karena frekuensi tatap muka antara pihak manager dan karyawan sangat minim. Karena itulah dalam perusahaan ini dibutuhkan sistem yang dapat membantu dalam penilaian kinerja karyawan agar proses seleksi dapat dilakukan secara obyektif dan efisien.

Penyeleksian ini dilakukan sesuai kriteria-kriteria klasifikasi utama dengan menggunakan metode C4.5 yang dapat mengelola nilai inputan yang sesuai dengan kriteria-kriteria pada penilaian kinerja pegawai yang mempunyai nilai prioritas tertentu. Algoritma ini menggunakan konsep data mining yang melakukan proses penggalian informasi pada keputusan-keputusan sebelumnya, untuk kemudian dijadikan informasi yang membentuk pola pohon keputusan. Proses tersebut dimulai dengan pengolahan data training, kemudian data tersebut dicari prioritas yang paling besar dan dijadikan sebuah *root* pada pohon. Proses tersebut diulang dengan mencari *gain* terbesar sampai bertemu perbandingan yang bersih dan pohon berhenti dibentuk.

Tujuan dalam penelitian ini yaitu untuk membuat sebuah system pendukung keputusan dalam penilaian kinerja karyawan. Sistem ini dibuat untuk mempermudah pihak HRD maupun manager dalam melakukan penilaian kinerja karyawan. Selain itu sistem juga diharapkan memberikan manfaat untuk merekap data kinerja maupun mengolah data untuk menghasilkan suatu keputusan. Dan sebagai alternatif HRD untuk menilai karyawan yang berkompeten pada perusahaan.

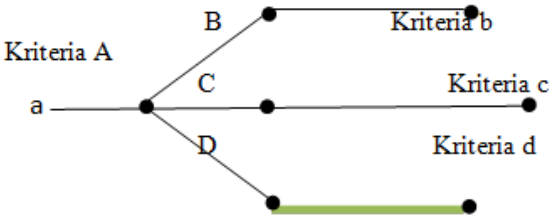
Sistem yang dibangun menggunakan beberapa kriteria dalam penilaian kinerja, antara lain: komunikasi, orientasi prestasi, inisiatif, pemikiran analitis, kepedulian terhadap tugas dan kualitas, kerja sama, orientasi pelayanan pelanggan, kerapian administrasi, pengaturan kerja, kemampuan teknis dan Fungsionalitas.

I. METODOLOGI

Decision Tree

Decision Tree adalah sebuah struktur pohon, dimana setiap node pohon merepresentasikan atribut yang telah diuji, setiap cabang merupakan suatu pembagian hasil uji, dan node daun (*leaf*) merepresentasikan

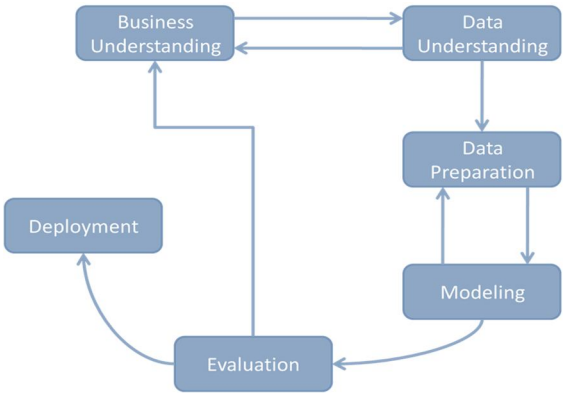
kelompok kelas tertentu. Level node teratas dari sebuah *Decision Tree* adalah node akar (root) yang biasanya berupa atribut yang paling memiliki pengaruh terbesar pada suatu kelas tertentu[1]. Pada umumnya *Decision Tree* melakukan strategi pencarian secara top-down untuk solusinya. Pada proses mengklasifikasi data yang tidak diketahui, nilai atribut akan diuji dengan cara melacak jalur dari node akar (*root*) sampai node akhir (daun) dan kemudian akan diprediksi kelas yang dimiliki oleh suatu data baru tertentu.



Gambar 1 Susunan pohon keputusan

Data Mining

Data mining adalah serangkaian proses untuk menggali nilai tambah berupa pengetahuan yang selama ini tidak diketahui secara manual dari suatu kumpulan data[2]. Dalam melakukan proses *data mining* melalui beberapa tahap seperti pada gambar berikut :



Gambar 2 Tahapan *data mining* berdasarkan CRISP-DM

Algoritma C4.5 merupakan generasi baru dari algoritma ID3 yang dikembangkan oleh J.Ross Quinlan pada tahun 1983. Untuk membuatsebuah pohon keputusan, algoritma ini dimulai dengan memasukkan training samples ke dalam simpul akar pada pohon keputusan. Training samples adalah sampel yang digunakan untuk membangun model classifier dalam hal ini pohon keputusan. Kemudian sebuah atribut dipilih untuk mempartisi sampel ini. Untuk tiap nilai yang dimiliki atribut ini, sebuah cabang dibentuk. Setelah cabang terbentuk maka subset darihimpunan data yang atributnya memiliki nilai yang bersesuaian dengan cabang tersebut dimasukkan ke dalam simpul yang baru.

Algoritma ini mempunyai prinsip dasar kerja yang sama dengan algoritma ID3. Perbedaan utama C4.5 dari ID3 adalah:

- C4.5 dapat menangani atribut kontinyu dan diskrit.
- C4.5 dapat menangani training data dengan missing value.
- Hasil pohon keputusan C4.5 akan dipangkas setelah dibentuk.
- Pemilihan atribut yang dilakukan dengan menggunakan *Gain Ratio*. *Information gain* pada ID3 lebih mengutamakan pengujian yang menghasilkan banyak keluaran.

Entropy, Gain, dan Pruning

Sebuah obyek yang diklasifikasikan dalam pohon harus dites nilai entropinya. Entropy adalah ukuran dari teori informasi yang dapat mengetahui karakteristik dari impuryt,dan homogenity dari kumpulan data. Dari nilai *entropy* tersebut kemudian dihitung nilai *information gain*(IG) masing-masing atribut.

$$Entropy(S) = -p_+ log_2 p_+ - p_- log_2 p_- \quad (1)$$

dimana :

- S adalah ruang (data) sample yang digunakan untuk training.
- P+ adalah jumlah yang bersolusi positif(mendukung) pada data sample untuk kriteria tertentu.
- P- adalah jumlah yang bersolusi negatif(tidak mendukung) pada data sample untuk kriteria tertentu. Dari rumus entropy diatas dapat disimpulkan bahwa definisi entropy(S) adalah jumlah bit yang diperkirakan dibutuhkan untuk dapat mengekstrak suatu kelas(+ atau -) dari sejumlah data acak pada suatu ruang sampel S. Entropy bisa dikatakan sebagai kebutuhan bit untuk menyatakan suatu kelas. Semakin kecil nilai entropy maka semakin baik digunakan dalam mengekstraksi suatu kelas.

Panjang kode untuk menyatakan informasi secara optimal adalah $-\log_2 p$ bits untuk *messages* yang mempunyai probabilitas p.

Information Gain

Setelah mendapat nilai *entropy* untuk suatu kumpulan data, maka kita dapat mengukur efektivitas suatu atribut dalam mengklasifikasikan data. Ukuran efektifitas ini disebut *information gain*.Secara matematis, *infomation gain* dari suatu atribut A,dituliskan sebagai berikut :

$$Gain(S, A) - Entropy(S) - \sum ve\ Value(A) \frac{|S_v|}{S} Entropy(S_v) \quad (2)$$

dimana :
A : atribut
V : suatu nilai yang mungkin untuk atribut A
Values (A) : himpunan yang mungkin untuk atribut A
$|S_v|$: jumlah sampel untuk nilai v
|S| : jumlah seluruh sampel data
Entropy(S_v): entropy untuk sampel-sampel yang memilki nilai v

Gain Ratio

Perhitungan *information gain* masih memiliki sejumlah kekurangan. Salah satu kekurangan yang mungkin terjadi adalah pemilihan atribut yang tidak relevan sebagai pemartisi yang terbaik pada suatu simpul. *Gain ratio* merupakan normalisasi dari *information gain* yang memperhitungkan entropi dari distribusi probabilitas subset setelah dilakukan proses partisi. Secara matematis, gain ratio dihitung sebagai berikut [3]:

$$GainRatio(X) = \frac{Gain(X)}{SpitInfo(X)} \quad (3)$$

Dimana SplitInfo(X) merupakan entropy dari seluruh distribusi probabilitas *subset* setelah dilakukan pemartisian(*splitting*).

$$SplitInfo(x) = -\sum(\left(\frac{|T_i|}{T}\right) * log_2\left(\frac{|T_i|}{|T|}\right)) \quad (4)$$

(4)

Dimana persamaan $4|T_i|$ adalah kardinalitas dari subset T_i yang berada dalam *training data* T.

Penanganan Atribut *Continuos*

Untuk mengatasi atribut ini yaitu dengan mengurutkan *value* dari *training* data terlebih dahulu sehingga diperoleh {v1, v2, ..., Vm}. Cari setiap *threshold* yang berada di antara vi dan vi+1 dengan menggunakan persamaan $\frac{V_i+V_{i+1}}{2}$. Dengan demikian hanya akan ada m-1 kemungkinan *threshold*. Untuk masing-masing threshold ini dilakukan perhitungan *Gain* [3]. Gain dengan nilai terbaik akan digunakan sebagai batas dari rentang nilai.

Berikut adalah bagaimana memilih nilai terbaik dari *threshold* c:

Contoh kriteria yang bertipe *continuous*

Table 1 Data kontinyu

Temperature	*20*	*22*	*24*	*26*	*28*	*30*
Enjoy Sport	No	No	Yes	Yes	Yes	No

(2)

$$C_1 = \frac{22+24}{2}=23,\ C_2 = \frac{28+30}{2}=29$$

Mencari *Gain* dari tiap-tiap nilai diatas

Tabel 2 Perhitungan data kontinyu

	ya	*Ya*				*Entropy ≤*	*Entropy >*	*Gain*
	tidak	*Jumlah*	*Tidak*	*jumlah*				
1	0	2	3	1	4	0	0,81	0,4
2	3	2	0	1	1	0,97	0	0,2

Dari data pada tabel 2 dapat diperoleh *Gain* terbesar yaitu pada C_1 yang bernilai 23.Sehingga dapat digunakan sebagai batas *threshold* dari nilai-nilai continous pada atribut tersebut.Perhitungan diatas adalah contoh dari perhitungan dua *threshold.*Untuk perhitungan yang lebih lanjut, *Gain* dihitung pada masing-masing *threshold* pada setiap nilai yang sudah dilakukan perhitungan *criterion*.

(3)

Pruning

Pada saat pembangunan pohon keputusan, apabila terdapat banyaknya cabang mungkin mencerminkan adanya *noise* atau *outlier* pada *training data* [4]. Pemangkasan pohon dapat dilakukan untuk mengenali dan menghapus cabang-cabang tersebut. Pohon yang dipangkas akan menjadi lebih kecil dan lebih mudah dipahami. Pohon semacam itu biasanya juga menjadi lebih cepat dan lebih baik dalam melakukan klasifikasi. Terdapat dua pendekatan utama dalam pemangkasan pohon yaitu, *prepruning* dan *postpruning*.

Pada pendekatan *prepruning*, sebuah pohon dipangkas dengan cara menghentikan pembentukan cabangnya. Pada pendekatan *postpruning* pemangkasan cabang dilakukan setelah pohon terbentuk.

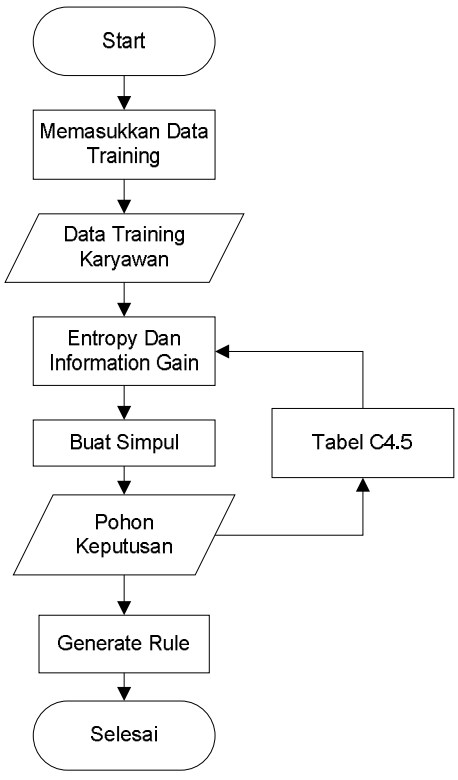
Bagian ini menjelaskan metode-metode yang digunakan dalam penelitian dan termasuk

juga rancangan sistem dan prosedur penelitian (dalam bentuk algoritma atau yang lainnya).Dalam penelitian ini menggunakan pendekatan *prepruning* dan *postpruning* untuk melakukan pemangkasan pohon. Berikut ini formula yang digunakan untuk pendekatan *prepruning* :

$$e = \frac{r + \frac{z^2}{2n} + z\sqrt{\frac{r}{n} - \frac{r^2}{n} + \frac{z^2}{4n^2}}}{1 + \frac{z^2}{n}} \quad (5)$$

Keterangan :
r = Nilai perbandingan *error rate*
n = total *sample*
$z = \Phi^{-1}(c)$
c = *confidence level*

Pada pendekatan postprunning menggunakan metode *Reduced Error Pruning* [5]. *Reduced Error Pruning* merupakan salah satu algoritma *postpruning*. Algoritma ini membagi data menjadi dua, yaitu *training data* dan *test data.Training data* adalah data yang digunakan untuk membentuk pohon keputusan, sedangkan *test data* digunakan untuk menghitung nilai *error rate* pada pohon setelah dipangkas.



Gambar 3 *Flowchart* pembentukan pohon

II. HASIL DAN PEMBAHASAN

Sistem yang dibuat melakukan penerapan untuk proses seleksi dan kinerja karyawan. Pada kasus ini proses seleksi menggunakan metode C4.5. Metode ini memberikan ouput yang mencocokan inputan dengan pohon keputusan yang dibuat. Proses pembentukan pohon akan memerlukan data *training* atau data *history* perusahaan. Jadi penilaian yang dilakukan dari data sebelumnya akan menyesuaikan dalam pembentukan pohon yang menjadi aturan pengambilan keputusan.

Flowchart

Flowchart merupakan gambaran dari alur sistem yang dikerjakan secara keseluruhan maupun secara terpisah dalam suatu proses tertentu dan menjelaskan prosedur – prosedur yang ada dalam sistem. Pada system ini data training akan diolah untuk menjadi aturan atau *Rule*, kemudian aturan tersebut akan dijadikan sebuah aturan baku yang digunakan sebagai prediksi keputusan data-data yang baru.

Dalam proses yang dilakukan metode C4.5, pohon dibentuk dengan melalui beberapa tahap. Yang pertama data *Training* yang ada diolah, sehingga menghasilkan *entropy* dan *information gain*. Setelah itu *information gain* terbesar yang diperoleh dari masing-masing atribut, akan digunakan sebagai simpul akan. Dan proses itu akan terus berulang sampai tidak bisa dibentuk dengan perolehan *information gain*. Proses selanjutnya yaitu membuat *rule* yang akan digunakan untuk melakukan proses pengambilan keputusan pada data selanjutnya.

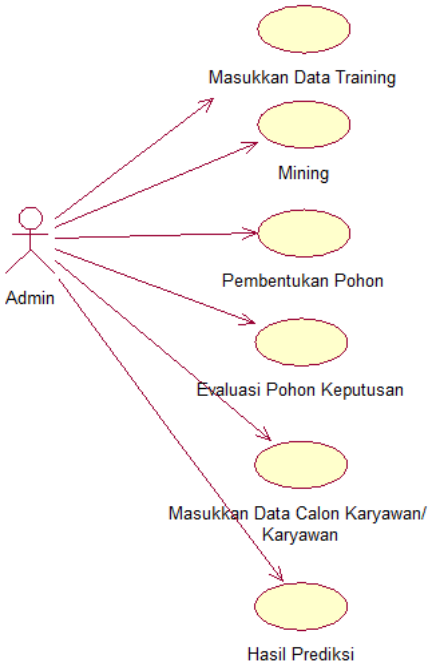
Use Case Diagram

Dalam sistem pendukung keputusan penempatan dan kinerja pegawai ini, hanya terdapat satu pengguna yaitu, admin. Admin tersebut dapat melakukan tugas sebagai berikut :

a. *Input data training*, yaitu melakukan proses menambah, data yang akan dijadikan data training. Data yang dimasukkan memiliki kelas direkomendasikan dan tidak direkomendasikan.
b. *Mining*, yaitu proses pengolahan data dari data training yang kemudian akan membentuk sebuah aturan.
c. *Pembentukan pohon*, yaitu proses proses pembuatan pohon keputusan yang terbentuk dari proses mining dengan algoritma C4.5. Pada pohon tersebut akan terbentuk sebuah

rule yang akan menjadi aturan dari data yang akan diprediksi.

d. Evaluasi pohon keputusan, yaitu menghitung nilai *accuracy*, *precision*, *recall*, dan *error rate* dari masing – masing pohon keputusan yang terbentuk.
e. Input data Calon Karyawan atau Karyawan, yaitu melakukan proses input data yang akan diprediksi hasilnya.
f. Hasil Prediksi, yaitu prediksi yang yang diberikan oleh sistem setelah melakukan pencocokan dengan pohon keputusan.



Gambar 4 Use Case Diagram

Penelitian ini menggunakan data karyawan yang diambil dari Gajah Mada Lumajang. Data terdiri dari total data karyawan yang sudah dipekerjakan maupun pelamar. Yang pertama yaitu data *Employee of the Mounth* yang terdiri dari 364. data tersebut adalah data yang diperoleh dari remokendasi manager untuk karyawan terbaik selama 4 bulan, yaitu pada bulan januari sampai april 2012.

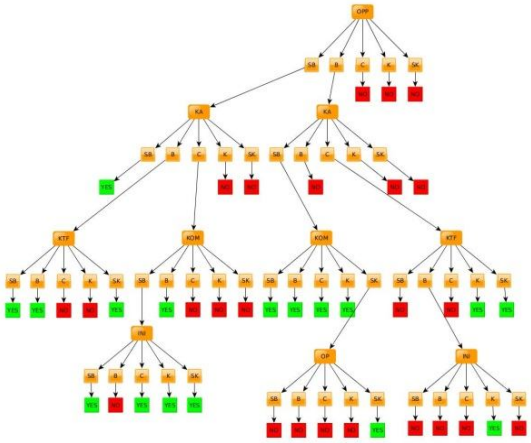
Selanjutnya dilakukan penyeleksian data yang kurang atau tidak lengkap. Dan dilakukan pembersihan data sehingga didapatkan data yang mempuyai nilai yang lengkap. Setelah itu dapat dilakukan proses penggalian informasi.

Semua data yang siap akan dibagi menjadi 3 partisi. Pada tiap-tiap partisi tersebut ada data training dan testing. Data training akan digunakan untuk membentuk pohon, kemudian akan dilakukan pengetesan dengan menggunakan data testing untuk perbandingan. Perbandingan akan digunakan untuk melakukan menentukan akurasi pohon keputusan.

Tabel 3 Pengujian data

	Data
Data Training	192
Data Testing	152
Uji *Pruning*	20
Jumlah	364

Pada data di tabel 3 akan dilakukan proses mining untuk membentuk aturan dengan pohon keputusan. Pada proses awal pohon dibentuk, pohon mempunyai bentuk seperti Gambar 5.



Gambar 5 Pohon Keputusan Penilaian kinerja karyawan terbaik

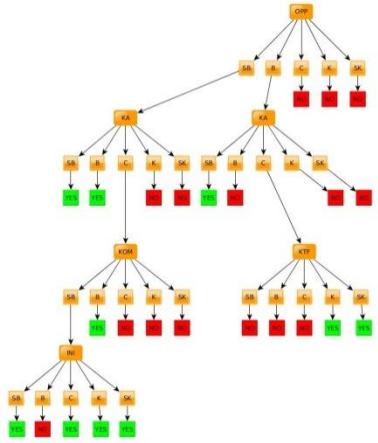
Pembentukan pohon di gambar 5 dilakukan dengan menghitung *entropy* masing-masing atribut, kemudian mencari information *Gain* terbesar.Dan yang terbesar itu digunakan sebagai simpul. Kemudian dilakukan proses iterasi sampai pohon berhenti terbentuk. Pada contoh diatas diketahui OPP menjadi simpul pada perhitungan *Information Gain* awal.Kemudian diikuti dengan kerapian administrasi pada iterasi pertama, dan kemampuan teknis dan fungsionalitas, dan komunikasi pada itersai ke dua.Setelah itu inisiatif dan orientasi prestasi pada iterasi ketiga. Pada pohon keputusan diatas menggunakan partisi A dengan 192 data training dari 364 data. Dan dilakukan pengetesan dengan 152 data dan menghasilkan data akurasi Tabel 4. Dengan nilai yang dihitung dengan *confusion matrix*:

- *Precision* : 60%
- *Recall* : 88,24%
- *Accuracy* :92,05%
- *Errot Rate* : 7,96%

Tabel 4 Tabel data akurasi

Keputusan/ Identifikasi	*Tidak Direkomendasikan*	*Direkomendasikan*
Tidak Direkomendasikan = 134	124	10
Direkomendasikan = 17	2	15

Kemudian dilakukan pemangkasan pohon dengan metode pruning. Yang pertama dilakukan yaitu dengan cara prepruning yaitu melakukan pemangkasan pohon ketika nilai *error rate* cabang pohon lebih kecil dari nilai Uji *pruning*. Prepruning mengecek perbandingan *error* tersebut ketika pohon mulai terbentuk. Berikut adalah pohon keputusannya:



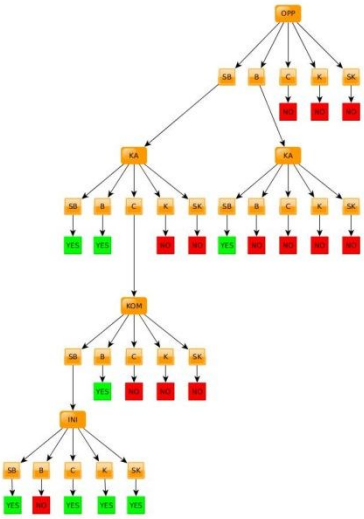
Gambar 6 Pohon Keputusan Prepruning Karyawan terbaik

Pada gambar 6 terjadi pemangkasan cabang pohon yaitu pada atribut KTF yang mempunyai nilai KA baik dan OPP Sangat Baik, Kemudian KOM dengan KA Sangat Baik dan OPP Baik. Setelah itu juga terjadi pemangkasan pada INI dengan parent KTF Baik, KA Cukup dan OPP Baik. Sehingga terbentuk aturan keputusan seperti di atas. Pada akurasi terdapat perbedaan dengan pohon yang belum dipruning yaitu sebagai berikut:

- *Precision* : 61,54%
- *Recall* : 94,12%
- *Accuracy* :92,75%
- *Errot Rate* : 7,28%

Jika dillihat dari hasil tersebut, ada peningkatan akurasi sebesar 0,70%. Peningkatan ini terjadi karena cabang yang tidak perlu sudah dipangkas untuk mempercepat komputasi.

Tahap *pruning* yang kedua yaitu proses postpruning dimana proses pemangkasan dilakukan setelah semua pohon terbentuk. Berikut ini adalah pohon keputusan yang dibentuk setelah melalui proses *postpruning*:



Gambar 7 Pohon Keputusan Kinerja Karyawan *Postpruning*

Pada gambar 7 terjadi pemangkasan yang cukup banyak pada pohon keputusan *postpruning*.Pemangkasan yang terjadi pada *prepruning* juga terjadi pada *postpruning*.Tetapi satu cabang pohon juga terpangkas pada pohon ini, yaitu cabang KTF yang memiliki parent KA dengan nilai Cukup.

Pada struktur pohon yang berbeda, akan terjadi perbedaan pada level akurasinya juga. Berikut adalah nilai dari pohon keputusan ini setelah dihitung dengan menggunakan teknik *confusion matrix*:

- *Precision* : 58,33%
- *Recall* : 82,35%
- *Accuracy* : 91,39%
- *Errot Rate* : 8,61%

Dengan hasil ini, dapat dibandingkan bahwa proses *prepruning* memiliki sedikit lebih besar akurasinya daripada yang lain. Dengan jarak 0,7% dari pohon tanpa pruning dan 2,73% dengan *postpruning*.

III. SIMPULAN

Dari penelitian ini dihasilkan beberapa kesimpulan, antara lain :

1. Berdasarkan evaluasi yang dilakukan dapat diketahui bahwa proses pembentukanpohon menggunakan teknik *pruning* memiliki kecepatan yang lebih tinggi karena penyederhanaan pohon, tetapi tidak selalu memliki akurasi yang lebih besar.

2. Perbedaan pohon keputusan yang dihasilkan disebabkan oleh perbedaan jumlah *data training* yang digunakan pada masing-masing partisi.
3. Pohon keputusan Partisi A menggunakan teknik *pruning* dengan jumlah *data training* lebih besar daripada *data testing* memiliki akurasi tertinggi dibandingkan dengan pohon keputusan yang lain, yaitu mencapai 90 %.
4. Dengan adanya system ini, pengolahan data meliputi penempatan dan kinerja karyawan dapat dilakukan dengan mudah dan cepat.

Saran-saran yang dapat disampaikan antara lain :

1. SPK dapat dikembangkan dengan lebih dinamis dengan mengurangi dan menambah atribut lain.
2. Dapat dilakukan komputasi lebih cepat ketika mengolah data kontinyu untuk memaksimalkan waktu proses.

IV. DAFTAR PUSTAKA

[1] Sabna, Eka,. 2010. Aplikasi Data Mining Untuk Menganalisis Track Record Penyakit Pasien Dengan Menggunakan Teknik Decision Tree.Fakultas Ilmu Komputer Universitas Putra Indonesia

[2] Chapman, P., Clinton, J., Kerber, R., Khabaza, T., Reinartz, T., Shearer, C., and Wirth, R. *CRISP-DM 1.0 Step-by-step data mining guide.* CRISP-DM Consortium. 2000.

[3] Indriyani, Novi, 2009. Penerapan Metode Literatur Sistem Pengukung Keputusan. Universitas Indonesia

[4] Khairana, Indah Kuntum., 2009. Penggunaan Pohon Keputusan untuk Data Mining, institut Teknologi Bandung, 2009

[5] Nugroho, F. S., Kristanto, H., dan Oslan, Y. Validitas Suatu Alamat menggunakan Pohon keputusan dengan Algoritma ID3. *Jurnal Informatika, Volume 3 Nomor 2 April 2007* : 2. 2007.