**JOUTICA Volume 5 No.1 2020** **ISSN: 2503-071X**
**E-ISSN: 2621-511X**

# KOMBINASI METODE ENSEMBLE, CFS DAN POHON KEPUTUSAN UNTUK PREDIKSI KINERJA PETUGAS STUDI KASUS: SURVEY PODES BADAN PUSAT STATISTIK

**Eko Hardiyanto**
*Pranata Komputer, Badan Pusat Statistik Provinsi Jawa Timur*
*Jl. Kendangsari Industri No. 43-44, Surabaya*
*E-mail:eko.hardi@bps.go.id*

**ABSTRAKS**
*Kegiatan rilis data pada pendataan Potensi Desa (Podes) Badan Pusat Statistik pada kurun waktu sembilan tahun terakhir, selalu mengalami keterlambatan. Untuk meminimalisir agar keterlambatan tidak terjadi secara berulang, penelitian bertujuan memprediksi kinerja petugas berdasarkan faktor internal dan eksternal sebagai informasi kegiatan di Badan Pusat Statistik (BPS). Pemilihan atribut pada penelitian ini menggunaan nilai gain informasi dan Correlation feature selection (CFS), selanjutnya dilakukan pemodelan dengan algoritma Pohon Keputusan. Hasil penelitian ini menunjukkan akurasi prediksi pada petugas organik BPS meningkat sebesar 10.57 % dari 63,19 % menjadi 69,87 % dengan faktor penentu keterlambatan adalah beban kerja, track record dan kemudahan lokasi, sedangkan pada petugas mitra dengan menggunakan metode CFS akurasi prediksi meningkat sebesar 24,11 % dari 65,78 % menjadi 81,643 % dengan faktor penentu keterlambatan adalah nilai pendalaman,kemampuan bekerja dalam tim, dan profesionalitas*.

*Kata Kunci: prediksi, kinerja petugas, survey bps, podes, badan pusat statistik*

***ABSTRACTS***
*Data release activities in the Data Collection Village Potential (Podes) of the Central Statistics Agency in the past nine years, always experiencing delays. To minimize delays that do not occur repeatedly, the study aims to predict the performance of officers based on internal and external factors as information on activities in the Central Statistics Agency (BPS). The selection of attributes in this study uses information gain values and Correlation feature selection (CFS), then modeling is done using the Decision Tree algorithm. The results of this study indicate the accuracy of prediction on BPS organic officers increased by 10.57% from 63.19% to 69.87% with the determinants of delays were workload, track record and location convenience, while in partner officers using the CFS method the prediction accuracy increased by 24.11% from 65.78% to 81.663% with the determinants of delay being the value of deepening, the ability to work in teams, and professionalism.*

*Keywords: prediction, officer performance, survey bps, podes, statistical central agency*

## 1. PENDAHULUAN

Badan Pusat Statistik (BPS) merupakan lembaga nonkementrian yang bertugas menyediakan data statistik yang berkualitas dan terpercaya bagi pemerintah dan masyarakat. Penyampaian data yang terpercaya dan peningkatkan layanan kualitas data, memerlukan informasi yang akurat dan rilis data yang tepat waktu (BPS, 2014). Namun, Ketepatan waktu dalam rilis data merupakan salah satu masalah yang sering dihadapi di BPS. Pelaksanaan pendataan survei maupun sensus sering kali mengalami keterlambatan yang menyebabkan kegiatan rilis data terlambat. Berdasarkan hal ini, BPS membutuhkan prediksi keterlambatan kegiatan.

Salah satu faktor penyebab keterlambatan pelaksanaan survei adalah kinerja petugas dari survei yang kurang bagus. Oleh karena itu, untuk mengetahui petugas dengan kinerja kurang bagus perlu dilakukan prediksi terhadap kinerja petugas. Pemberian prediksi umumnya dilakukan dengan metode naive bayes (Witten, dkk., 2011). Pada metode na¨ıve bayes ini masih memiliki beberapa kekurangan diantaranya, belum memberikan penjelasan tentang variabel yang dapat memberikan nilai akurasi tinggi saat digunakan dalam prediksi kinerja petugas (Al-Radaideh, dkk., 2012).

Karena nilai prediksi yang dihasilkan berupa peluang, maka akan sangat mungkin berubah saat ada pengurangan serta penambahan variabel dalam pemberian prediksi. Oleh karena itu, dibutuhkan metode lain yang mampu memberikan prediksi lebih tepat berdasarkan variabel yang berpengaruh pada kinerja petugas.

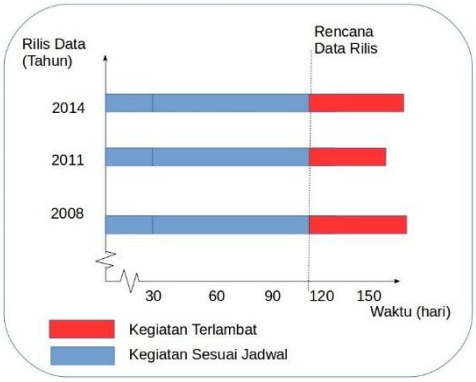
Klasifikasi dengan menggunakan pohon keputusan merupakan metode klasifikasi yang paling banyak digunakan pada penelitian sebelumnya. Teknik klasifikasi pohon keputusan ini menggunakan pemilihan atribut dengan memberikan bobot sesuai peringkat dari penghitungan nilai gain informasi. Pemilihan atribut pada pohon keputusan sangat bergantung pada nilai informasi yang dihasilkan oleh tiap atribut. Pemilihan atribut mengambil peranan yang sangat penting dalam penentuan nilai akurasi dari pembentukan model pohon keputusan.

Prediksi dengan menggunakan teknik pohon keputusan menghasilkan nilai akurasi prediksi kinerja petugas BPS yang masih rendah yaitu 60 persen. Hal ini disebabkan penggunaan atribut dengan gain informasi belum merupakan atribut yang kuat di dalam melakukan prediksi petugas. Pada disertasi Hall (Hall, 2008) menyatakan bahwa atribut yang baik adalah atribut yang berhubungan kuat dengan kelasnya dan tidak saling berhubungan antar tiap atribut penyusunnya. Pada disertasinya, Mark mengusulkan *Correlation feature selection* (CFS). Oleh karena itu, pada penelitian ini menggunakan CFS untuk menentukan atribut terbaik dalam memprediksi kinerja petugas.

Penelitian ini bertujuan untuk memprediksi kinerja petugas survei berdasarkan faktor internal dan eksternal sebagai informasi keterlambatan kegiatan di BPS. Petugas dalam penelitian ini merupakan petugas lapangan dan kegiatan dalam penelitian ini merupakan aktivitias outdoor.

## 2. PENELITIAN SEBELUMNYA

Kegiatan pada Badan Pusat Statistik sering mengalami keterlambatan dalam hal rilis data. Padahal, ketika terjadi keterlambatan akan berdampak kepada dua hal, yaitu 1) mengganggu pengambilan keputusan para pengguna data dan 2) bagi internal BPS sendiri menyebabkan kegiatan overlap sehingga beban kerja pegawai BPS menjadi bertumpuk yangakhirnya menggangu kegiatan yang lain. Salah satu kegiatan yang sering mengalami keterlambatan adalah kegiatan pendataan Potensi Desa (Podes). berdasarkan evaluasi kegiatan dalam kurun waktu sembilan tahun terakhir (3 periode kegiatan Podes) (BPS, 2014) selalu mengalami keterlambatan yang ditunjukkan Gambar 1.




**Gambar 1. Lama keterlambatan rilis data kegiatan pendataan podes**

Keterlambatan ini terjadi pada tahap pengolahan yaitu tahapan "check kewajaran" data Podes yang membutuhkan cross-check ulang di lapangan. Berdasarkan wawancara dengan subject matter di BPS, keterlambatan tersebut disebabkan oleh dua faktor yaitu: faktor internal dan faktor eksternal. Faktor internal terdiri dari jenis kelamin, jabatan petugas, status petugas, tingkat pendidikan, tingkat profesionalitas, integritas, kemampuan bekerja di dalam tim, pengalaman survei, *track record*, nilai pendalaman materi pelatihan serta beban kerja petugas. Faktor eksternal terdiri dari jarak lokasi pencacahan, ketersediaan jenis transportasi, kondisi medan (topografi) dan tingkat kemudahan lokasi dimana lokasi yang sulit dijangkau menyebabkan proses cross-check memerlukan waktu lama. Untuk mengatasi permasalahan tersebut diperlukan evaluasi di awal kegiatan.

Saat ini evaluasi yang telah dilakukan hanya pada tahapan administrasi dan belum menyentuh evaluasi kinerja petugas. Berdasarkan hal ini, dibutuhkan prediksi kinerja petugas berdasarkan faktor internal dan eksternal sebagai informasi untuk meminimalisir keterlambatan kegiatan di Badan Pusat Statistik (BPS). Ketersediaan informasi mengenai kinerja petugas digunakan sebagai dasar pengawasan lebih intensif terhadap petugas yang berpotensi terlambat sehingga keterlambatan dapat diminimalkan. Pemberian prediksi yang tepat terhadap karakteristik petugas yang berpotensi mengalami keterlambatan dapat menjadi dasar untuk evaluasi di awal kegiatan pada kegiatan Badan Pusat Statistik.

Pemberian prediksi untuk preferensi dari pengguna saat ini sudah banyak digunakan di berbagai sektor baik itu sektor pariwisata, penyedia akomodasi, perusahaan ritel, industri perfilman hingga perusahaan besar seperti Google, Ebay yang menyediakan layanan kepada pengguna (Ricci, dkk., 2011). Prediksi juga diterapkan dalam mengurangi resiko penyakit (Lin dan Yang, 2012), risiko kerusakan lingkungan (Kretser, 2015) dan mengukur tingkat produktivitas ataupun keuntungan perusahaan (Wang dan Chuang, 2015). Pemberian prediksi kinerja petugas pada (Al-radaideh dkk, 2012) yang dilakukan di suatu perusahaan ternyata memberikan keuntungan dalam hal efektifitas biaya dan waktu. Penelitian tersebut mengevaluasi performa karyawan perusahaan (Al-radaideh dkk, 2012) dengan cara memberikan informasi mengenai kinerja serta kemungkinan pegawai yang akan keluar dengan menggunakan klasifikasi. Teknik yang sesuai untuk mementukan prediksi performa pegawai adalah dengan menggunakan pohon keputusan (Al-radaideh dkk, 2012; Peter dan Somasundaram, 2012; Sohn dan Kim, 2012; Magesh, dkk, 2013; Patidar, dkk, 2015). Metode pohon keputusan telah dilakukan antara lain oleh Al-radaideh (Al-radaideh dkk, 2012), Wang (Wang dan Chuang, 2015), Peter dan Somasundaram (2012) menggunakan pohon keputusan C4.5 yang dikembangkan oleh Ross Quinlan (Quinlan, 1993) dalam menentukan prediksi kinerja petugas di perusahaan, yang hasilnya memberikan kemudahan di dalam mengelola pegawai pada perusahaan. Wang (Wang dan Chuang, 2015) dengan memprediksi kinerja yang diterapkan pada perusahaan pembangkit listrik tenaga surya sehingga dapat

memprediksi total produksi listrik pada pembangkit listrik. Peter dan Somasundaram (2012) menggunakan untuk memprediksi suatu penyakit dengan menghitung asupan nutrisi dan pola makan pada pasien yang hasilnya dapat mengetahui pasien yang berpotensi untuk terserang penyakit.

Prediksi kinerja petugas dengan pohon keputusan menggunakan rumus naive bayes lihat persamaan:

$$\Pr(H|E) = \frac{\Pr(E|H)\Pr(H)}{\Pr(E)} \quad (1)$$

Penggunaan Pr(E1|H) pada persamaan 1memiliki artiPeluang kejadian E1 dengan kondisional kejadian H. Notasi E diartikan sebagai keterangan sedangkan H diartikan sebagi hipotesis (Witten, dkk, 2011). Pada penelitian ini H adalah kejadian kemungkinan petugas tepat, terlambat atau sangat terlambat dalam hal pengumpulan data saat pelaksanaan kegiatan BPS. Untuk notasi E sendiri adalah kombinasi dari atribut yang digunakan untuk memprediksi kinerja, antara lain *track record*, beban kerja, profesionalitas dan integritas. Pada Rumus 2 secara lebih lengkap penerapan pada suatu kejadian menjadi seperti ini:

$$\Pr(tepat|E) = \frac{\prod_{i=1}^{k} \Pr(E_i|tepat \ldots \times \Pr(H)}{\Pr(E)} \quad (2)$$

Penelitian sebelumnya (Al-radaideh dkk, 2012)(Witten, dkk, 2011) menunjukkan penghitungan probabilitas dengan bayes dan pembentukan model dengan pohon keputusan sangat dipengaruhi oleh pemilihan atribut. Ketepatan pemilihan atribut sangat berpengaruh pada prediksi suatu kejadian. Pemilihan atribut yang mempengaruhi dalam memberikan keputusan kepada seorang pegawai atau karyawan telah diteliti Chien (Chien dkk, 2008) dan klasifikasi dengan menggunakan data mining berdasarkan karakteristik pribadi dan luaran kinerja oleh Shuangcheng (Shuangcheng dan Ping, 2009)(Vecchione dkk, 2012). Penggabungan model klasifikasi oleh Shuangcheng (Shuangcheng dan Ping, 2009) memang memberikan nilai akurasi yang cukup, namun pemberian model prediksi pada penelitian tersebut menghasilkan akurasi yang rendah saat diujikan pada kinerja petugas, karena belum dilakukan pengujian terhadap atribut yang saling berhubungan atau berkorelasi. Oleh karena itu untuk memperbaiki permasalahan tersebut dilakukan penghitungan terhadap kemungkinan atribut masih terdapat korelasi antar atribut yang tinggi, dimana menyebabkan akurasi pada algoritma pembelajaran mesin menjadi rendah. Pada disertasi Hall (Hall, 2008) mengusulkan metode *Correlation feature selection* (CFS) untuk mengetahui atribut mana yang masih memiliki hubungan antar tiap atribut independen. Pemilihan atribut/fitur berpengaruh terhadap akurasi sehingga harus dihilangkan atribut yang saling berhubungan. Algoritma pada pembelajaran mesin akan sulit membuat keputusan jika data yang digunakan memiliki hubungan sehingga attribut yang saling berhubungan dan data yang berulang harus diminimalisir. Pengukuran evaluasi subset dari fitur berdasarkan pada "Subset atribut yang baik merupakan atribut yang berkorelasi tinggi dengan klasifikasinya, namun tidak memiliki korelasi antar atribut-atribut penyusunnya" (Hall, 2008). Penggunaan seleksi fitur pada penelitian (Kohavi dan John, 1996) menunjukkan bahwa seleksi fitur memberikan peningkatan pada informasi yang dilakukan pengekstrakan. Beberapa metode stastistik digunakan dalam mengevaluasi fitur/atribut yang memiliki karakteristik dengan data numerik. Pengukuran dari data numerik cenderung monotonik (meningkatkan ukuran dari subset fitur yang tidak akan mengurangi kinerja dari algoritma) sehingga beberapa algoritma pencarian tidak dapat diimplementasikan (Narendra dan Fukunaga, 1977).

**Tabel 1. Perbandingan atribut penelitian sebelumnya dengan penelitian saat ini**

<table>
<tr><th>Chien (2008)</th><th>Vecchione (2012)</th><th>Jantan (2010)</th><th>Al-radaideh (2012)</th><th colspan="2">Atribut Penelitian terpilih</th></tr>
<tr><td rowspan="2">Umur, Status Pernikahan</td><td rowspan="2">Semangat</td><td rowspan="2">Kualifikasi pendidikan</td><td rowspan="3">Usia,Status Pernikahan, Jabatan, Tingkat Gaji</td><td>Jenis Kelamin</td><td>Jabatan petugas</td></tr>
<tr><td>Integritas</td><td>Status petugas</td></tr>
<tr><td>Jenis Kelamin</td><td rowspan="2">Kecocokan</td><td>Jenis Kelamin</td><td>profesionalitas</td><td>kerja tim</td></tr>
<tr><td>Pendidikan</td><td>Kategori</td><td>Jenis Kelamin</td><td>Pendidikan</td><td>Track record</td></tr>
<tr><td>Pengalaman, Sekolah/ Universitas</td><td>Sifat Berhati-hati (Persisten, perhatian)</td><td rowspan="3">Output pekerjaan,KemampuanIndividu, pengalaman, Nilai evaluasi pimpinan</td><td rowspan="3">Pendidikan,Universitas Spesialisasi, Pengalaman (tahun)</td><td>Pengalaman survei</td><td>Nilai pendalaman materi</td></tr>
<tr><td rowspan="2">Asal Perekrutan,BidangKeilmuan</td><td rowspan="2">Kestabilan emosional</td><td colspan="2">Beban kerja</td></tr>
<tr><td>Jarak lokasi</td><td>topografi</td></tr>
</table>

Pada Tabel 1 merupakan perbandingan atribut penelitian pada prediksi untuk kinerja pegawai di suatu perusahaan (Al-radaideh dkk, 2012)(Vecchione dkk, 2012)(Jantan dkk, 2010)(Chien dkk, 2008). Pada penelitian ini kami menambahkan atribut yang telah disesuaikan dengan studi kasus penelitian ini yaitu pada kegiatan yang dilakukan di lapangan, sehingga mempertimbangkan faktor eksternal seperti jarak, transportasi dan topografi dari lokasi.

## 3. METODE

Data yang digunakan dalam penelitian ini menggunakan data set yang diperoleh dari kegiatan pendataan Podes 2014 pada Unit BPS Provinsi Sulawesi Selatan. Ada dua jenis data yang digunakan di dalam penelitian ini:

a. Data primer, yaitu data yang secara langsung dikumpulkan oleh peneliti. Data primer dalam penelitian ini meliputi: nilai integritas, kerja tim, profesionalitas, Pendidikan, *Track record* dan beban kerja petugas Podes 2014 yang diperoleh dengan melakukan wawancara kepada subject matter dan kuesioner. Rincian dari responden penelitian dapat dilihat pada Tabel 2.
b. Data sekunder yaitu data yang diperoleh dari data yang telah tersedia pada data BPS meliputi data evaluasi Podes tiap kabupaten di Provinsi Sulawesi Selatan, daftar petugas dan wilayah kerja pada kegiatan Podes di BPS Provinsi Sulawesi Selatan, data hasil pengolahan Podes, rekapitulasi nilai pendalaman materi dan daftar status petugas mitra dan pegawai BPS.

Ada dua macam variabel dalam penelitian ini, 1) variabel dependen yaitu ketepatan pengumpulan data dari petugas Podes dan merupakan variabel yang akan diprediksi pada penelitian ini, 2) variabel independen yaitu variabel yang didasarkan pada faktor eksternal dan faktor internal yang berpengaruh pada keterlambatan kegiatan Podes. Penentuan variabel independen selain dari wawancara terhadap subjectmatter juga mempertimbangkan hasil penelitian sebelumnya tentang variabel yang mempengaruhi kinerja pegawai (Al-radaideh dkk, 2012; Vecchione dkk, 2012; Shuangcheng dan Ping, 2009; Jantan dkk, 2010). Pada penelitian ini, selanjutnya variabel independen disebut dengan atribut. Adapun daftar atribut seperti tercantum pada Tabel 3.

**Tabel 2. Responden dataset penelitian**

| *Narasumber* | *Jumlah* |
|---|---|
| BPS RI | |
| Kepala Seksi Ketahanan Sosial | 1 |
| Staf | 2 |
| BPS Provinsi | |
| Kepala Bidang Stat. Sosial | 1 |
| Kepala Seksi | 2 |
| BPS Kabupaten/Kota Kepala | |
| Seksi Sosial / IPDS | 24 |
| Staf | 18 |

Pada pengisian kuesioner selanjutnya atribut yang telah dikumpulkan diberikan nilai dengan keterangan seperti yang tercantum pada Tabel 4. Penelitian ini menggunakan alat bantu aplikasi Waikato Environment for Knowledge Analysis (WEKA) dalam pengolahan data serta analisa dari akurasi dari algoritma

**Tabel 3. Atribut penelitian berdasarkan faktor internal dan faktor eksternal**

| *Faktor* | *Atribut* |
|---|---|
| Eksternal | 1. Jarak lokasi ke kab. |
| | 2. Topografi wilayah |
| | 3. Transportasi lokasi |
| | 4. Tingkat kemudahan lokasi |
| Internal | 1. Jenis Kelamin |
| | 2. Integritas |
| | 3. Kerja tim |
| | 4. Profesionalitas |
| | 5. Pendidikan |
| | 6. *Track record* petugas |
| | 7. Pengalaman survey |
| | 8. Hasil pendalaman materi pelatihan Podes |
| | 9. Jabatan Petugas |
| | 10. Status Petugas |
| | 11. Beban Kerja pegawai |

**Tabel 4. Keterangan nilai dari atribut pada pelaksanaan podes**

| No | Atribut | Range Nilai Atribut |
|---|---|---|
| 1. | Jarak dari Lokasi ke Kab | Jarak lokasi pencacahan ke kabupaten (Merupakan nilai rata-rata dari jarak desa yang menjadi lokasi pencacahan) |
| 2. | Topogafi Wilayah | Nilai topografi wilayah 1 = Pegunungan, 2=Lembah, 3=Dataran |
| 3. | Transportasi lokasi | Alat transportasi utama yang digunakan menuju lokasi pencacahan: 1 dan 2=Sepeda motor dan Mobil, 1=Mobil, 2=Sepeda Motor, 3=Perahu, 4=Lainnya(jalan kaki, kuda, gerobak) |
| 4. | Tk. kemudahan lokasi | Nilai tingkat kemudahan lokasi (Mudah, Menengah, Sulit, Sangat Sulit) |
| 5. | Jenis Kelamin | Jenis kelamin dari petugas (L=laki-laki, P=Perempuan) |

| | | |
|---|---|---|
| 6. | Integritas | Nilai integritas yang diberikan oleh responden pencacahan untuk petugas yang berada di bawahnya (Skala 1-5, 1=kurang bagus, 5=sangat bagus) |
| 7. | Kerja tim | Nilai Kerja tim petugas yang diberikan oleh responden pencacahan untuk petugas yang berada di bawahnya (Skala 1-5, 1=kurang bagus, 5=sangat bagus) |
| 8. | Profesionalitas | Nilai profesionalitas petugas yang diberikan oleh responden pencacahan untuk petugas yang berada di bawahnya (Skala 1-5, 1=kurang bagus, 5=sangat bagus) |
| 9. | Pendidikan | Latar belakang pendidikan tertinggi yang pernah dimiliki oleh petugas (SMA, DI, DII, DIII, S1/DIV, S2) |
| 10. | *Track record* petugas | Nilai evaluasi *track record* petugas yang diberikan oleh responden pencacahan untuk petugas yang berada di bawahnya (kurang bagus, cukup, bagus) |
| 11. | Pengalaman survey | Merupakan nilai dari pengalaman survey dari petugas (Belum pernah (Podes yang pertama kali), Pernah 1 survei, Pernah 2 Survei, Pernah lebih dari 2 survei) |
| 12. | Hasil pendalaman materi pelatihan Podes | Nilai pendalaman materi petugas petugas yang diadakan saat pelatihan petugas Podes skala (0-100) |
| 13. | Jabatan Petugas | Jabatan petugas yaitu pencacah atau pemeriksa |
| 14. | Status Petugas | Status petugas merupakan Petugas tersebut Mitra BPS atau Pegawai Organik BPS |
| 15. | Beban Kerja pegawai | Beban kerja pegawai adalah banyaknya pekerjaan (baik itu survei atau sensus) yang sedang dilakukan bersamaan dengan kegiatan Podes |
| 16. | Ketepatan pengumpulan data sesuai jadwal (kelas) | Ketepatan pengumpulan data dari petugas: tepat waktu (tepat berdasarkan jadwal yang ditentukan), Terlambat (yaitu keterlambatan yang tidak lebih dari 5 hari dari jadwal), Sangat terlambat (keterlambatan lebih dari 5 hari dari jadwal pendataan) |

Tahapan analisis pada penelitian ini melalui beberapa tahap, yaitu:

a. Seleksi atribut, yaitu pemilihan variabel atau fitur yang akan digunakan di dalam pembuatan model klasifikasi dari algoritma machine learning. Pada tahap awal atribut akan dihitung dengan menggunakan gain informasi dengan menggunakan Rumus 4 untuk melihat peringkat nilai informasi dari atribut. Selanjutnya, seleksi atribut dengan metode diskretisasi untuk atribut dengan tipe kontinu atau numerik (Fayyad dan Irani, 1993) dan penghitungan derajat CFS/merits (Hall, dkk, 2012). Penentukan atribut pada penelitian ini menggunakan dua penghitungan yaitu information gain dengan melihat nilai entropy dari setiap atribut dan *correlation feature selection* (CFS) untuk melihat hubungan korelasi diantara atribut-atribut independen dengan dependen. Rumus untuk menghitungan entropy (Witten, dkk, 2011) dan CFS (Hall, 2008):

$$H(p_1, p_2, \dots, p_n) = \sum_{i=0}^{n}(-p_i log p_i) \quad (3)$$

$$Gain(S, A) = H(S) - \sum_{V \in Values\ (A)}^{n} \left(\frac{S_v}{S}\right) S_v \quad (4)$$

$$CFS = max \frac{r_{cf_1} + r_{cf_2} \dots + r_{cf_k}}{\sqrt{k + 2(r_{f_1 f_2} + \dots + r_{f_k f_1})}} \quad (5)$$

Keterangan:

max : nilai yang tertinggi yang terpilih

k : jumlah kelas

$r_{cf_k}$ : Korelasi antara atribut terhadap kelas

$r_{f_1 f_2}$ : Korelasi antar atribut dalam suatu kelas

Rumus rcfi dan $r_{fkfi}$ menggunakan rumus korelasi pearson sebagai berikut:

$$r_{f_1 f_2} = \frac{n \sum_{i=1}^{n}(x_i - \bar{x})(x_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2 \sum_{i=1}^{n}(y_i - \bar{y})^2}} \quad (6)$$

Keterangan:

n : jumlah data set

x dan y : atribut yang akan dihitung korelasi

Seleksi atribut dengan menggunakan CFS, atribut yang telah dipilih berdasarkan nilai information gain akan dilihat nilai derajatnya dengan menggunakan persamaan 5. Nilai dari CFS ini dipengaruhi oleh korelasi yang terjadi antar tiap atribut dengan kelas penyusunnya. Penghitungan untuk korelasi dalam CFS menggunakan korelasi pearson pada persamaan 6. Nilai yang terpilih sebagai CFS adalah nilai yang tertinggi. Hal ini berdasarkan pada Hall (Hall, 2008) dimana : Semakin besar nilai CFS berarti atribut penyusunnya memiliki korelasi yang semakin besar terhadap kelasnya. Semakin kecil nilai CFS mengindikasikan bahwa antar atribut memiliki korelasi yang tinggi karena berdasarkan persamaan 5 nilai dari korelasi antar atribut menjadi pembagi. Semakin besar nilai korelasi antar atribut penyusun maka akan menyebabkan nilai CFS rendah.

b. Penentuan prediksi kinerja petugas menggunakan algoritma pohon keputusan (Al-radaideh dkk, 2012; Peter dan Somasundaram, 2012; Sohn dan Kim, 2012; Magesh, dkk, 2013; Patidar, dkk, 2015). Penggunaan pohon keputusan pada penelitian ini didasarkan pada penelitian

sebelumnya yang memberikan tingkat akurasi tinggi serta mudah dipahami dalam penggunaannya.

c. Analisis akurasi data, pada tahap ini dilakukan analisis terhadap model pohon keputusan yang memberikan nilai akurasi paling tinggi. Selanjutnya dilakukan peningkatan algoritma dengan metode ensemble (Witten, dkk, 2011). *Ensemble* method merupakan penggabungan dari algoritma dasar sehingga keunggulan dari tiap algoritma digunakan untuk meningkatkan akurasi.

## 4. HASIL

### 4.1 Pengujian 1

a. Seleksi atribut pada data petugas Podes

Data diolah menggunakan aplikasi *Waikato Environment for Knowledge Analysis* (WEKA). Hasil penghitungan nilai gain informasi dari tiap atribut. Peringkat dan nilai gain informasi dari tiap atribut yang digunakan untuk memprediksi kinerja petugas. Setelah dilakukan pemberian peringkat pada tiap atribut dan dilihat nilai gain informasi hanya terpilih 12 atribut dari 16 atribut. Kedua belas atribut tersebut terpilih karena memiliki nilai informasi yang lebih besar dari nol. Atribut dengan nilai gain informasi lebih besar dari nol memiliki pengaruh saat digunakan dalam pembentukan pohon keputusan sedangkan yang nilainya rendah atau nol atribut tersebut tidak memberikan informasi saat digunakan dalam pohon keputusan. Kedua belas atribut yang terpilih dari proses penghitungan gain informasi digunakan di dalam penghitungan CFS dengan persamaan 5.

Hasil perhitungan CFS menunjukkan bahwa nilai CFS tertinggi pada atribut yang berjumlah 8 dengan nilai 0,3578. Kedelapan atribut yang terpilih dengan seleksi CFS adalah *track record* petugas, kemampuan bekerja dalam tim, integritas, profesionalitas, hasil pengolahan di kabupaten, transportasi, tingkat kemudahan lokasi dan pendidikan. Kedelapan atribut ini merupakan atribut yang memiliki pengaruh kuat terhadap kelas yang akan diprediksi dan memiliki hubungan korelasi antar atribut yang paling minimal dibandingkan dengan kombinasi dari atribut-atribut lain.

b. Penentuan prediksi kinerja dengan Pohon keputusan:

1) Penghitungan *gain* informasi

Atribut yang terpilih dengan CFS selanjutnya akan digunakan untuk membentuk model dalam pohon keputusan. Hasil penghitungan gain informasi dalam menentukan percabangan pertama. Pada tabel tersebut tampak bahwa nilai gain informasi tertinggi terdapat pada atribut *track record* petugas dengan nilai gain info 0,09379. Hal ini berarti, atribut *track record* petugas memiliki informasi paling banyak dibandingkan atribut lain. Oleh karena itu, atribut *track record* petugas akan dipilih sebagai percabangan pertama dalam membentuk pohon keputusan.

2) Penentuan percabangan

Pada tahap 1) telah didapatkan nilai gain informasitertinggi pada atribut *track record*, sehingga digunakan sebagai percabangan pertama pohon keputusan. *Track record* petugas memiliki tiga percabangan yaitu baik, cukup dan kurang. Pada *track record* baik, terdapat 153 petugas yang mengumpulkan data tepat waktu, 25 petugas terlambat dan 8 petugas sangat terlambat. Adapun pada percabangan *track record* cukup dan kurang.

3) Menghitung nilai entropy dari setiap percabangan

Nilai entropy pada percabangan yang memiliki nilai (nol) tidak dilanjutkan untuk dipecah. Namun pada percabangan dengan nilai lebih besar dari nol maka akan dipecah. Nilai entropy dari setiap percabangan dari *track record*, dimana nilai entropy dari ketiga percabangan lebih besar dari nol. Oleh karena itu setiap percabangan akan dilihat nilai gain informasinya untuk menentukan percabangan selanjutnya dengan penghitungan pada tahap 1).

4) Hasil pembentukan Pohon keputusan

Setelah dihitung nilai informasi untuk setiap percabangan, maka percabangan akan berhenti saat kelompok pada percabangan sudah homogen yaitu nilai gain informasimendekati nol.

c. Pengujian akurasi

Setelah mendapatkan atribut yang paling kuat dan berpengaruh maka selanjutnya dilakukan pemilihan algoritma dengan akurasi tertinggi berdasarkan penelitian sebelumnya (Wang dan Chuang, 2015; Al-radaideh dkk, 2012; Magesh, dkk, 2013; Patidar, dkk, 2015) dengan algoritma Decision Tree J48, CART dan NBTree. Tabel 5 menunjukkan perbandingan akurasi dengan menggunakan aplikasi WEKA. Pada aplikasi WEKA disediakan menu untuk melihat perbandingan dari beberapa algoritma pada menu WEKA-experimenter. Dengan memasukkan data pada tiap tahapan yaitu dari atribut yang belum dilakukan seleksi atribut, atribut setelah seleksi dengan gain informasi dan dengan menggunakan CFS. Pada semua jenis algoritma terlihat bahwa dengan menggunakan CFS dapat meningkatkan akurasi dari algoritm. Pengujian 1 memberikan kesimpulan bahwa penggunaan seleksi atribut berpengaruh dalam meningkatkan akurasi.

**Tabel 5. Perbandingan Akurasi Algoritma Seleksi Fitur Pada Data Podes**

| Algoritma | Seleksi atribut | Akurasi Cross Validation Folds (10) |
|---|---|---|
| J48 | Original | 63.1958 % |
| | CFS | 67.7346 % |
| CART | Original | 64.3021 % |
| | CFS | 64.5309 % |
| NBTree | Original | 63.386 % |
| | CFS | 63.5561 % |

d. Analisis akurasi dan model pohon keputusan

Pada percabangan untuk nilai *track record* akan menimbulkan masalah jika diterapkan pada evaluasi petugas Podes, dimana informasi *track record* didapatkan setelah melaksanakan kegiatan di BPS. Untuk karakteristik pengambilan keputusan pada pegawai BPS model ini tidak akan mengalami masalah karena ukuran atribut *track record* dapat ditentukan dengan melihat kegiatan petugas tersebut di masa lalu. Namun, masalah akan muncul pada petugas mitra yang belum pernah memiliki *track record* di kegiatan BPS sebelumnya. Petugas mitra pada kegiatan BPS adalah petugas yang direkrut hanya saat BPS akan melakukan kegiatan survei atau sensus, dimana *track record* dari petugas tidak dapat secara terus-menerus untuk dinilai.

Selain dari atribut *track record* faktor yang seharusnya muncul pada petugas namun pada pemodelan ini tidak terpilih, yaitu banyaknya beban kerja pada pegawai. Beban kerja pegawai sangat berpengaruh kepada tingkat ketepatan pengumpulan data. Beban kerja tidak muncul karena beban kerja pada pegawai BPS dan mitra memiliki karakteristik yang berbeda. Pada pegawai BPS beban kerja dapat lebih dari satu kegiatan, sedangkan untuk petugas mitra rata-rata seragam yaitu hanya memiliki satu beban pekerjaan yaitu kegiatan podes saja. Berdasarkan analisis data, ditemukan bahwa karakteristik untuk petugas BPS dan petugas Mitra berbeda sehingga evaluasi terhadap data petugas dipisahkanantara petugas pegawai BPS dan petugas Mitra.

Pada tahap ini didapatkan kesimpulan awal yaitu dengan menggunakan seleksi fitur dengan CFS dapat meningkatkan akurasi dari algoritma. Algoritma dengan nilai tertinggi yaitu J48 akan digunakan di dalam pengujian selanjutnya. Pada pengujian selanjutnya akan dilakukan pemisahan antara data status petugas organik BPS dan petugas mitra karena berdasarkan analisis awal pada data gabungan (mitra dan pegawai) memberikan informasi yang bias.

### 4.2 Pengujian 2

1) Seleksi atribut pada data petugas mitra dan pegawai BPS

Hasil perbandingan seleksi atribut pada data Pegawai BPS dan mitra pada pendataan Podes dapat dilihat dari penghitungan nilai. Nilai gain informasi dan penghitungan atribut yang terpilih dengan menggunakan CFS seperti yang telah dijabarkan pada pengujian 1. Perbedaan dengan Pengujian 1, data pengujian 2 telah dipisahkan antara Pegawai organik BPS dan PetugasMitra.

Perbandingan hasil seleksi atribut Berdasarkan seleksi atribut pada Tabel 4 didapatkan atribut topografi wilayah, transportasi, kerja tim, *track record* dan beban kerja pegawai merupakan atribut yang berpengaruh kuat terhadap kinerja pengumpulan data pada pegawai BPS. Sedangkan pada mitra yang berpengaruh adalah kemudahan lokasi, integritas, kerja tim, profesionalitas, *track record* dan nilai pelatihan yang akan digunakan dalam pemodelan dengan pohon keputusan.

2) Penentuan Prediksi kinerja dengan Pohon Keputusan pada data petugas mitra dan pegawai BPS

Selanjutnya pembentukan model dengan algoritma pohon keputusan DT-J48 pada petugas Podes dengan statuspetugas mitra. Dari representasi gambar dengan pohon keputusan dapat diketahui dengan mudah bahwa sebagian besar keterlambatan yang terjadi di petugas mitra karena memilikinilai pendalaman saat pelatihan dibawah angka 57,2 yang terjadi sebanyak 11 kejadian. Selain itu petugas yang memiliki profesionalitas biasa atau nilai tiga namun lokasi pencacahan sulit akan berpotensi terlambat sebanyak 16 kejadian. Faktor kerja tim yang kurang bagus yaitu dibawah skala nilai dua akan mengakibatkan kegiatan tersebut terlambat dan sangat terlambat, hal ini muncul sebanyak 6 kejadian.

3) Pengujian Akurasi

Berdasarkan Tabel 5 bahwa dari pemilihan atribut ini terlihat bahwa atribut yang mempengaruhi kinerja pengumpulan data dari pegawai BPS dan Mitraberbeda. Selanjutnya setelah model pohon keputusan pada data mitra maupun Petugas Podes maka dilakukan pengujian pada algoritma dengan akurasi tertinggi berdasarkan penelitian sebelumnya (Wang dan Chuang, 2015; Al-radaideh dkk, 2012; Magesh, dkk, 2013; Patidar, dkk, 2015) dengan algoritma Decision Tree J48, CART dan NBTree. Hasil penghitungan pada akurasi data yang sudah dipisah antara petugas BPS dan petugas mitra.

Berdasarkan metode decision tree J48 (DT-J48) memberikan tingkat akruasi yang tertinggi dimana untuk akurasi pada data pegawai BPS sebesar 69,09 % dan pada petugas mitra sebesar 70,374 %. Jika dibandingkan dengan akurasi pada pengujian 1 maka pada pengujian 2 akurasi untuk setiap algoritma bertambah. Hal ini mengindikasikan bahwa dengan pemisahan data Petugas dan Mitra meningkatkan akurasi untuk memprediksi kinerja petugas di kedua kelompok.

Klasifikasi dari banyaknya kejadian yang terjadi pada tiap rule yang dihasilkan dari model pohon keputusan. Petugas dengan klasifikasi terlambat terjadi sebanyak 33 petugas. Yang berarti 15 % dari keseluruhan petugas mitra masih mengalami keterlambatan.

Pada petugas organik BPS sendiri, representasi dari pohon keputusan dalam bentuk gambar secara hierarkis. Dari representasi gambar dapat diketahui denganmudah bahwa sebagian besar keterlambatan yang terjadi pada petugas BPS dimana petugas tersebut memiliki beban pekerjaan diatas dua pekerjaan pada waktu yang bersamaan dan riwayat *track record* yang biasa saja sebanyak 29 kejadian dengan klasifikasi terlambat. Selain itu, transport

lokasi yang sulit hanya bisa ditempuh dengan perahu atau sepeda motor dan *track record* dari pegawai yang kurang bagus serta total sejumlah 22 kejadian.

Berdasarkan klasifikasi kinerja pada pegawai BPS diketahui bahwa masih terdapat 49 kejadian dengan klasfikasi terlambat dan 2 kejadian dengan klasifikasi sangat terlambat. Sebagian besar dari kinerja dari pegawai BPS masih belum maksimal sehingga masih terdapat keterlambatan yang terjadi dan angka keterlambatan yang lebih tinggi dibanding dengan petugasyang direkrut dengan status mitra.

Petugas dengan status mitra pada kegiatan di BPSmerupakan petugas yang direkrut untuk membantu kegiatan di BPS. Pada penelitian sebelumnya padasebesar 69,32 % serta metode stacking algoritma J48 dan fungsi logistik regresi sebesar 68,80 %. Pengujian evaluasi akurasi menggunakan data yang berbeda dari data yang digunakan sebagai data training. Metode ensemble bagging dipilih untuk digunakan menguji pada data petugas podes untuk mitra dan pegawai. Hasil pengujian dengan menggunakan data pengujian atau data testing untuk menguji dari keakuratan dari algoritma pada Tabel 5 yang menunjukkan hasil dari akurasi dari ensemble method.

Tabel 5 menunjukkan akurasi dengan metode yang sama yaitu Bagging-classifier: J48, pada data petugas mitra memberikan hasil lebih signifikan dibandingkan dengan akurasi pada data pegawai BPS. Hal ini disebabkan karena sejak seleksi atribut, nilai gain informasi pada data petugas mitra lebih tinggi pada kisaran 0,1478 - 0,09094 dibandingkan dengan nilai gain informasi pada pegawai BPS yaitu pada kisaran 0,09823 – 0,032. Nilai tersebut mengindikasi pemilihan atribut yang terpilih untuk pegawai BPS di dalam penelitian ini belum kuat sehingga belum bisa mencerminkan penyebab keterlambatan pegawai BPS.

Berdasarkan studi lanjut dan diskusi dengan subject matter pada pemaparan hasil penelitian, ditemukan faktor lain yang mempengaruhi kinerja petugas yaitu faktor motivasi kerja. Dimana di dalam penelitian ini motivasi kerja dari dua kelompok petugas (mitra dan pegawai BPS) belum dipertimbangkan, sehingga kemungkinan menjadi penyebab dari akurasi prediksi untuk pegawai BPS yang masih rendah (belum signifikan). Motivasi dari dua kelompok petugas sangat berbeda, dimana motivasi kerja pada petugas mitra cenderung seragam yakni faktor finansial. Sedangkan motivasi pada pegawai BPS sangat beragam antara lain motivasi untuk diperhatikan pimpinan, menambah angka kredit, mutasi pegawai, faktor idealis hingga pengaruh pemikiran bahwa untuk pegawai BPS walaupun tidak bekerja tetap dibayarkan gajinya.

Pengujian akurasi algoritma pada Tabel 5. dengan menggunakan pemisahan dari data petugas podes yang telah dilakukan dengan menggunakan 90 data test untuk petugas didapatkan nilai akurasi sebesar 81,642 %. Detil akurasi dari prediksi yang dilakukan dapat dilihat dimana akurasi dari algoritma dalam memprediksikan kinerja petugas. Prediksi probabilitas pada tiap data testing menggunakan formula probabilias bayes yaitu Rumus (4).

## 5. KESIMPULAN

Kesimpulan pada penelitian ini adalah sebagai berikut:

1. Prediksi kinerja petugas dengan menggunakan algoritmabagging decision tree J48 dengan menggunakan seleksi fitur dengan metode CFS memberikan output akurasi untuk prediksi pada petugas organik BPS tanpa menggunakan CFS sebesar 63,19 % setelah menggunakan CFS sebesar 69,87 % dan pada petugas mitra akurasi sebesar 65,78 % dan setelah menggunakan CFS meningkat menjadi 81,643 %. Dimana teknik decision tree ini mampu untuk memberikan keputusan dengan variasi data pada data nominal dan numerik.
2. Atribut-atribut utama yang digunakan sebagai prediksi penentu keterlambatan untuk pegawai organik BPS adalah beban kerja yang lebih dari dua kegiatan dan *track record* petugas yang masih kurang bagus. Sedangkan pada petugas mitra, atribut utama yang digunakan adalah petugas yang memiliki nilai kerja tim rendah dan petugas dengan nilai pendalaman dibawah 57.2. Selain itu, potensi keterlambatan yang terjadi karena faktor topografi yang sulit dapat diantisipasi dengan persiapan alternatif transportasi, persiapan yang lebih matang dan lebih intensif untuk daerah sulit.

## PUSTAKA

Buku Pedoman Pencacah Podes 2014, Jakarta: BPS, 2014.

Al-radaideh, Q. A., dan Nagi, E. Al., "Using Data Mining Techniques to Build a Classification Model for Predicting Employees Performance", International Journal of Advanced Computer Science and Application, vol. 3, issue 2, 2012, pp. 144 – 151.

Vecchione, M., Alessandri, G., dan Barbaranelli, C., "The Five Factor Model in personnel selection: Measurement equivalence between applicant and non-applicant groups", Personality and Individual Differences, vol. 52, no.4, 2012, pp. 503–508.

Jantan, H., Hamdan, A. R., dan Othman, Z. A. Human Talent Prediction in HRM using C4.5 Classification Algorithm. International Journal on Computer Science and Engineering, vol. 2, no. 1, 2010, pp. 2526–2534.

Chien, C., dan Chen, L., "Data mining to improve personnel selection and enhance human capital: A case study in high-technology industry",

Expert Systems with Applications, vol. 34, no. 1, 2008, 280–290.

Hall, M. A., based Feature Selection for Machine Learning, Thesis for the Doctor of Philosophy, The University of Waikato, 1999, 25-50.

Ricci, F. Rokach, L., Shapira, B., Kantor, P.B., Recommender Systems Handbook, Springer, New York, 2011, 3 – 15.

Lin, E., dan Yang, D., "System Design of an Intelligent Nutrition Consultation and Recommendation Model", Proceeding of International Conference on Ubiquitous Intelligence and Computing, vol. 9, 2012, pp. 740 – 745.

Kretser, H. E., Wong, R., Roberton, S., "Mobile decision-tree tool technology as a means to detect wildlife crimes and build enforcement networks", Biological Conservation Journal, Elsevier, vol. 189, no. 1, 2015, pp. 33–38.

Wang, C., dan Chuang, J., "Integrating decision tree with back propagation network to conduct business diagnosis and performance simulation for solar companies." Decision Support Systems Journal, vol. 81, no.1, 2015, pp. 1 – 8.

Peter, T. J., dan Somasundaram, K. "An empirical study on prediction of heart disease using classification data mining techniques", IEEE International Conference on Advances In Egineering, Science and Managment (ICAESM), vol.1, no.1, 2012, pp. 514–518.

Sohn, S. Y., dan Kim, J. W., "Decision tree-based technology credit scoring for start-up firms: Korean case". Expert Systems With Applications, vol 39, no.4, 2012, pp. 4007–4012.

Magesh, N., Thangaraj, P., Sivagobika, S., Praba, S., dan Priya, R. M., "Evaluating The Performance Of An Employee Using Decision Tree Algorithm", International Journal of Engineering Research & Technology (IJERT), vol. 2, no.1, 2013, pp. 2814–2830.

Patidar, P., Dangra, J., dan Rawar, M. K., "Decision Tree C4.5 algorithm and its enhanced approach for Educational Data Mining", Enginering Universe for Scientific Research and Management, vol. 7, no. 1, 2015, pp. 1–14.

Shuangcheng, L., dan Ping, W., "Study on the Data preprocessing of the Questionnaire Based on the Combined Classification Data Mining Model." Proceeding of International Conference on E-Learning, EBusiness, Enterprise Information Systems, and E-Government, 2009, pp. 217–220.

Fayyad, U.M. dan Irani, K.B., "Multi-interval discretisation of continuous- valued attributes for classification learning", In Proceedings of the Thirteenth International Join Conference on Artificial Intelligence, Morgan Kaufmann, 1993, pp. 1022-1027.

Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P., dan Witten, I. H., "The WEKA Data Mining Software: An Update", SIGKDD Explorations, vol. 11, Issue 1, 2012.

Witten, I. H., Frank, E., dan Hall, M. a., Data Mining: Practical Machine Learning Tools and Techniques, Third Edition, Elsevier, San Francisco, 2011, pp. 41 – 283.

Quinlan, J. R. C4.5: Programs for Machine Learning, Morgan Kaufmann Publishers, San Francisco, CA, USA, 1993.

Kohavi, R dan John, G. "Wrappers for feature subset selection. Artificial Intelligence", special issue on relevance, vol. 97 no.(1–2), 1996, pp 273–324.

Narendra, P.M. and Fukunaga, K., "A branch and bound algorithmfor feature subset selection". IEEE Transactions on Computers, vol 26 no. 9, 1977, pp. 28-35.