

PEMODELAN KLASIFIKASI DALAM MENINGKATKAN PROSES PEMILIHAN CALON KARYAWAN DENGAN METODE C4.5 DAN JARINGAN SYARAF TIRUAN

Henny Leidiyana¹, Angga Aditya Permana²

¹ Program Studi Teknik Informatika, Fakultas Teknologi Informasi
Universitas Bina Sarana Informatika

² Program Studi Teknik Informatika, Fakultas Teknik
Universitas Muhammadiyah Tangerang
Jl. Perintis Kemerdekaan 1/33 Cikokol Kota Tangerang TLP. 55793251

Email : henny.hnl@bsi.ac.id¹, anggaamt@gmail.com²

ABSTRAK

Bagi manajemen SDM dalam sebuah perusahaan, pemilihan sumber daya manusia yang memiliki kualifikasi dan kinerja yang tinggi merupakan tugas yang penting. Pelamar kerja dengan jumlah yang sangat banyak membuat manajemen SDM harus melakukan pemilihan calon karyawan dengan teliti dan membutuhkan waktu yang tidak sedikit. Bidang penambangan data yang saat ini sedang populer, dapat dimanfaatkan untuk meningkatkan kinerja pemilihan karyawan. Dalam penulisan ini, dibahas tentang penerapan algoritma C4.5 dan jaringan syaraf tiruan untuk membuat model klasifikasi pemilihan karyawan. Untuk menguji model yang dihasilkan, beberapa percobaan dilakukan dengan menggunakan data yang dikumpulkan dari sebuah cabang perusahaan. Model yang dihasilkan dapat digunakan untuk membantu SDM dalam meningkatkan hasil pemilihan pelamar baru.

1.1. Pendahuluan

Hampir semua manajer dalam perusahaan, sumber daya manusia menjadi salah satu perhatian utama. Oleh sebab itu perusahaan tertarik untuk melakukan pemilihan karyawan secara tepat. Setelah merekrut karyawan, manajemen memperhatikan dengan kinerja karyawan ini untuk ditempatkan pada tempat yang sesuai. Penambangan data adalah bidang penemuan informasi dan pengetahuan yang baru dan handal (Han et al., 2011). Penambangan data memiliki beberapa tugas seperti asosiasi rule mining, klasifikasi dan prediksi, dan clustering.

Penambangan Data (Witten, 2011) didefinisikan sebagai proses penemuan pola dalam data. Berdasarkan tugasnya, penambangan data dikelompokkan menjadi (Larose, 2005) deskripsi, estimasi, prediksi, klasifikasi, *clustering*, Asosiasi. Proses dalam tahap *data*

klasifikasi. Proses kedua adalah klasifikasi, dimana data tes digunakan untuk memperkirakan akurasi dari *rule* klasifikasi. (Han, 2006). Proses klasifikasi didasarkan pada empat komponen: (Gorunescu, 2011)

- Kelas**
Variabel dependen yang berupa kategorikal yang merepresentasikan 'label' yang terdapat pada objek. Contohnya: resiko penyakit jantung, resiko kredit, *customer loyalty*, jenis gempu.
- Predictor**
Variabel independen yang direpresentasikan oleh karakteristik (atribut) data. Contohnya: merokok, minum alkohol, tekanan darah, tabungan, aset, gaji.
- Training dataset**
Satu set data yang berisi nilai dari kedua komponen di atas yang digunakan untuk menentukan kelas yang cocok berdasarkan *predictor*.

d. *Testing dataset*

Berisi data baru yang akan diklasifikasikan oleh model yang telah dibuat dan akurasi klasifikasi dievaluasi

1. Metode C4.5

J. Ross Quinlan, seorang peneliti di bidang *machine learning*, membuat sebuah algoritma *decision tree* yang dikenal dengan ID3 (*Iterative Dichotomiser*). Quinlan kemudian membuat algoritma C4.5 (sering disebut dengan pohon keputusan) yang merupakan pengembangan dari algoritma ID3 (Han, 2006). Algoritma ini memiliki kelebihan, yaitu mudah dimengerti, fleksibel, dan menarik karena dapat divisualisasikan dalam bentuk gambar (pohon keputusan) (Gorunescu, 2011). Algoritma C4.5 merupakan struktur pohon dimana terdapat simpul yang mendeskripsikan atribut-atribut, setiap cabang menggambarkan hasil dari atribut yang diuji, dan setiap daun menggambarkan kelas. Algoritma C4.5 secara rekursif mengunjungi setiap simpul keputusan, memilih pembagian yang optimal, sampai tidak bisa dibagi lagi. Algoritma C4.5 menggunakan konsep *information gain* atau *entropy reduction* untuk memilih pembagian yang optimal (Han, 2006).

2. Jaringan Syaraf Tiruan

Neural network adalah (Han, 2006) satu set unit input/output yang terhubung dimana tiap relasinya memiliki bobot. Sejak tahun 1950-an, neural network telah digunakan untuk tujuan prediksi, bukan hanya klasifikasi tapi juga untuk regresi dengan atribut target kontinu (Vercellis, 2009).

Multilayer perceptron (MLP) disebut juga multilayer feedforward neural network merupakan algoritma yang paling luas digunakan. Menurut Wong, Bodnovich dan Selvi (1997), sekitar 95% aplikasi bisnis yang menggunakan neural network, memakai algoritma ini (Vercellis, 2009). Salah satu kelebihan neural network adalah cukup baik dalam menangani data yang mengandung noise (Larose, 2005).

MLP terdiri dari input layer, satu atau lebih hidden layer, dan output layer. Berikut penjelasan masing-masing layer: (Vercellis, 2009)

1. Input layer

Input layer untuk menerima nilai masukan dari tiap record pada data. Jumlah simpul input sama dengan jumlah variabel prediktor.

2. Hidden layer

Hidden layer mentransformasikan nilai input di dalam network.

3. Output layer

Garis yang terhubung dengan Output layer berasal dari hidden layer atau input layer dan mengembalikan nilai keluaran yang bersesuaian dengan variabel prediksi.

3. Metode Evaluasi dan Validasi Algoritma Klasifikasi Penambangan Data

Untuk mengukur akurasi algoritma klasifikasi, metode yang dapat digunakan antara lain *cross validation* dan *confusion matrix* Untuk mengembangkan aplikasi (*development*) berdasarkan model yang dibuat, digunakan Rapid Miner.

1. *Cross Validation*

Cross validation adalah pengujian standar yang dilakukan untuk memprediksi *error rate*. Data *training* dibagi secara random ke dalam beberapa bagian dengan perbandingan yang sama kemudian *error rate* dihitung bagian demi bagian, selanjutnya hitung rata-rata seluruh *error rate* untuk mendapatkan *error rate* secara keseluruhan.

2. *Confusion matrix*

Metode ini menggunakan tabel matriks seperti pada Tabel 1 jika data set hanya terdiri dari dua kelas, kelas yang satu dianggap sebagai positif dan yang lainnya negatif (Bramer, 2007).

Tabel 1 Model *Confusion Matrix* (Bramer, 2007)

Klasifikasi yang benar	Diklasifikasikan sebagai	
	+	-
+	true positives	false negatives
-	false positives	true negatives

True positives adalah jumlah *record* positif yang diklasifikasikan sebagai positif, *false positives* adalah jumlah *record* negatif yang diklasifikasikan sebagai positif, *false negatives* adalah jumlah *record* positif yang diklasifikasikan sebagai negatif, *true negatives* adalah jumlah *record* negatif yang diklasifikasikan sebagai negative, kemudian masukkan data uji. Setelah data uji dimasukkan ke dalam *confusion matrix*, hitung nilai-nilai yang telah dimasukkan tersebut untuk dihitung jumlah *sensitivity (recall)*, *specificity*, *precision* dan *accuracy*. *Sensitivity* digunakan untuk membandingkan jumlah TP terhadap jumlah *record* yang positif sedangkan *specificity* adalah perbandingan jumlah TN terhadap jumlah *record* yang negatif.

1.2. Hasil Penelitian

Atribut-atribut beserta nilainya mengenai pemilihan karyawan terdapat pada tabel 2, terdiri dari tujuh atribut prediktor dan satu atribut kelas. Pemodelan menggunakan aplikasi rapidminer.

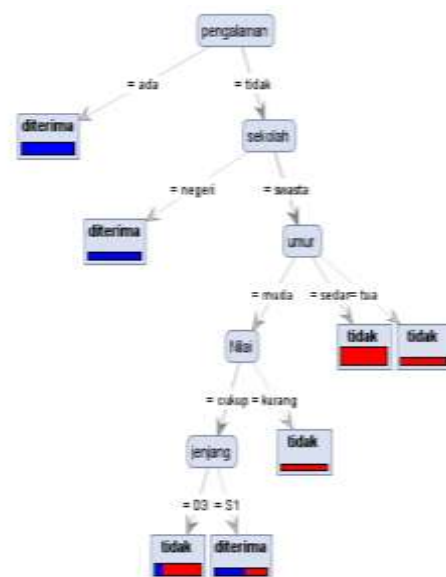
Tabel 2. Atribut dan Nilai

No	Atribut	Nilai
1	Umur	muda, sedang, tua
2	Status	belum menikah, menikah, duda/janda
3	Sekolah	negeri, swasta
4	Jenjang	D3, S1
5	Nilai	cukup, kurang
6	Pengalaman	ada, tidak ada
7	domisili	sesuai, tidak diterima, tidak
8	Kesimpulan	diterima

Penelitian ini dilakukan terhadap pemilihan karyawan calon *surveyor* yang tugasnya menganalisa calon pelanggan. Penerimaan karyawan hanya untuk

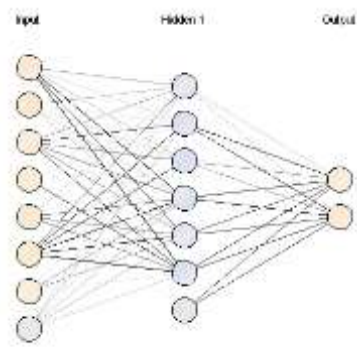
karyawan yang berjenis kelamin laki-laki karena pekerjaan ini merupakan pekerjaan lapangan dan dilakukan pada waktu yang tidak menentu sehingga atribut jenis kelamin tidak masuk sebagai atribut predictor. Jenjang pendidikan yang diinginkan oleh perusahaan untuk bidang pekerjaan ini adalah lulusan D3 atau S1 baik dari perguruan tinggi negeri ataupun swasta. Perusahaan ini memiliki banyak cabang sehingga atribut domisili juga menjadi atribut predictor. Pemilihan calon karyawan saat ini dilakukan dengan cara seleksi administrasi dan wawancara. Tidak ada tes lainnya. Pewawancara adalah kepala cabang. Jika memenuhi kualifikasi menurut pewawancara maka diterima. dari Data yang diperoleh sebanyak 56 data pelamar dan hasil seleksinya.

Penelitian ini menerapkan dua algoritma yaitu C4.5 dan jaringan syaraf tiruan. Kemudian dua model yang dihasilkan dari dua algoritma dibandingkan akurasi. Gambar 1 adalah model yang dihasilkan dari algoritma C4.5. Gambar 2 adalah model yang dihasilkan dari jaringan syaraf tiruan.



Gambar 1 Model Algoritma C4.5

Pohon yang dihasilkan dengan algoritma C4.5 adalah *rule* yang dapat digunakan menentukan calon karyawan yang layak diterima. Hasil penerapan dengan jaringan syaraf tiruan dapat dilihat pada keterangan di bawah Gambar 2



Gambar 2 Model Jaringan Syaraf Tiruan

Hidden 1
=====

Node 1 (Sigmoid)

umur: 1.109
status: 0.857
sekolah: 1.148
jenjang: 0.286
Nilai: 0.628
pengalaman: -1.977
Domisili: -0.321
Threshold: -0.648

Node 2 (Sigmoid)

umur: -0.422
status: -0.226
sekolah: -2.075
jenjang: -0.249
Nilai: -0.131
pengalaman: 2.888
Domisili: -0.245
Threshold: 0.882

Node 3 (Sigmoid)

umur: 0.862
status: 0.442
sekolah: 1.608
jenjang: 0.203
Nilai: 0.294
pengalaman: -2.498
Domisili: 0.020
Threshold: -0.401

Node 4 (Sigmoid)

umur: 3.314
status: 0.618
sekolah: 1.783
jenjang: 0.030

Nilai: 1.719
pengalaman: -3.020
Domisili: -1.415
Threshold: -0.887

Node 5 (Sigmoid)

umur: 2.927
status: 0.669
sekolah: 1.470
jenjang: 0.080
Nilai: 1.546
pengalaman: -2.691
Domisili: -1.184
Threshold: -0.903

Node 6 (Sigmoid)

umur: 3.512
status: 0.206
sekolah: 2.746
jenjang: -2.655
Nilai: 1.627
pengalaman: -3.295
Domisili: -0.266
Threshold: -0.608

Output

Class 'diterima' (Sigmoid)

Node 1: -0.656
Node 2: 2.701
Node 3: -1.242
Node 4: -2.609
Node 5: -2.192
Node 6: -2.536
Threshold: 2.419

Class 'tidak' (Sigmoid)

Node 1: 0.671
Node 2: -2.694
Node 3: 1.243
Node 4: 2.579
Node 5: 2.203
Node 6: 2.540
Threshold: -2.427

1.3. Pengujian Algoritma

1. Cross Validation

Dalam penelitian ini digunakan 10 fold-cross validation dimana 56 data training dibagi secara random ke dalam 10 bagian dengan

perbandingan yang sama kemudian *error rate* dihitung tiap bagiannya.

2. Confusion Matrix

Tabel *confusion matrix* yang dihasilkan dapat dilihat pada Gambar 3, yang merupakan hasil penerapan algoritma C4.5 dan Gambar 4 hasil penerapan jaringan syaraf tiruan.

	Actual	Predicted
Actual	10 (100%)	0 (0%)
Predicted	10 (100%)	0 (0%)

Gambar 3 Confusion Matrix untuk Algoritma C4.5

Akurasi model yang dihasilkan dengan menerapkan algoritma C4.5 sebesar 87,5% sedangkan akurasi penerapan dengan jaringan syaraf tiruan lebih besar sebesar 89,29%

	Actual	Predicted
Actual	10 (100%)	0 (0%)
Predicted	10 (100%)	0 (0%)

Gambar 4 Confusion Matrix untuk Jaringan syaraf tiruan

1.4 Kesimpulan

Penelitian bertujuan untuk untuk meningkatkan kualitas pemilihan calon karyawan khususnya untuk jenis pekerjaan surveyor dengan menerapkan penggalian data yang merupakan bidang yang saat ini sedang populer. Tugas penambahan data di sini adalah klasifikasi dengan membandingkan dua algoritma klasifikasi yaitu C4.5 dan jaringan syaraf tiruan. Pengujian model yang dihasilkan menggunakan metode *Cross Validation* dan *Confusion Matrix*. Berdasarkan pengujian, penerapan metode jaringan syaraf tiruan menghasilkan model yang lebih akurat dibanding C4.5. Agar model dapat digunakan secara umum, khususnya pada bidang pekerjaan ini maka sebaiknya data dikumpulkan dari berbagai cabang dari perusahaan yang sama atau bahkan dari perusahaan yang berbeda di berbagai wilayah.

DAFTAR REFERENSI

- [1] Bramer, Max. (2007). *Principles of Data Mining*. London: Springer
- [2] Gorunescu, Florin (2011). *Data Mining: Concepts, Models, and Techniques*. Verlag Berlin Heidelberg: Springer
- [3] Han, J., & Kamber, M. (2006). *Data Mining Concept and Tehniques*. San Fransisco: Morgan Kauffman.
- [4] Kusriani & Luthfi, E.T. (2009). *Algoritma Data Mining*. Yogyakarta: Andi Publishing.
- [5] Larose, D. T. (2005). *Discovering Knowledge in Data*. New Jersey: John Willey & Sons, Inc.
- [6] Vercellis, Carlo (2009). *Business Intelligent: Data Mining and Optimization for Decision Making*. Southern Gate, Chichester, West Sussex: John Willey & Sons, Ltd.
- [7] Witten, I. H., Frank, E., & Hall, M. A. (2011). *Data Mining: Practical Machine Learning and Tools*. Burlington: Morgan Kaufmann Publisher.
- [8] Al-Radaideh, Qasem A. Using Data Mining Techniques to Build a Classification Model for Predicting employees Performance. *IJACSA*. Vol 2 No 3. 2012 page 144