Perbandingan Algoritme C5.0 dan *Random Forest* menggunakan Data *Bank* Marketing

Rizka Aulia[1], Muhamad Fadli[2]

[12]Universitas PGRI Silampari
e-mail: rizkaaulia1515@gmail.com* ,

Abstract *- Klasifikasi merupakan teknik yang digunakan untuk menentukan item dari dataset ke dalam suatu kategori atau kelas. Tujuan dari klasifikasi yaitu sebagai sarana mencari pola dengan menganalisis sekumpulan dataset yang mendeskripsikan dan membedakan class data. Klasifikasi dapat digunakan untuk memprediksi untuk mengetahui hasil dari penawaran produk, dan perusahaan dapat mengambil keputusan lebih cepat. Metode yang dapat digunakan dalam klasifikasi yaitu C5.0 dan Random Forest. Pada penelitian hasil akurasi pada data Bank Marketing menggunakan algoritme C5.0 dan random forest untuk data latih dan data uji menggunakan K – fold cross validation, menghasilkan tingkat akurasi diatas 90%. Algoritme C5.0 lebih unggul untuk tingkat akurasi data latih sebesar 90.4 % , nilai akurasi terbaik untuk algoritme C5.0 terjadi pada fold ke 10 .Sedangkan random forest lebih unggul untuk data uji sebesar 97%, nilai ini konvergen dari 1-fold sampai 10-fold.*
Keywords: Klasifikasi, C5.0, Random Forest

PENDAHULUAN

Penawaran secara langsung atau *direct marketing* merupakan salah satu metode yang efisien untuk melakukan penawaran suatu produk atau jasa. *Direct marketing* bekerja dengan cara menawarkan secara langsung pada calon *customer* tertentu. Lebih efisien jika dibandingkan dengan promosi melalui iklan di media massa. *Direct marketing* juga dapat mengurangi *cost* karena biaya untuk komunikasi lebih murah jika dibandingkan dengan iklan (Liao *et al*, 2011). Pada tahun 1990, pekerja marketing mulai mengumpulkan informasi tentang konsumen dan menggunakannya untuk keperluan marketing. (Petrison *et al*, 1993).

Setelah masa penawaran produk selesai, bank akan menunggu hasil dari penawaran tersebut. Apakah dari penawaran yang telah dilakukan banyak konsumen yang tertarik dengan produk atau jasa yang telah ditawarkan atau tidak. Berdasarkan deskripsi diatas, dapat disimpulkan bahwa diperlukan kemampuan prediksi dalam perusahaan agar dapat mengetahui hasil dari penawaran produk, dan perusahaan dapat mengambil keputusan lebih cepat. Teknik penelitian tentang prediksi membuat perusahaan memiliki keuntungan karena dapat mengetahui konsumen yang tertarik dengan produk yang ditawarkan (Petrison *et al*, 1993). Dalam penelitian (Moro *et al*, 1993) data telemarketing.

Dalam ilmu komputer klasifikasi merupakan teknik yang digunakan untuk menentukan item dari dataset ke dalam suatu kategori atau kelas. Tujuan dari klasifikasi yaitu sebagai sarana mencari pola dengan menganalisis sekumpulan dataset yang mendeskripsikan dan membedakan *class* data. Sehingga dapat diguakan untuk memprediksi data yang belum diketahui (Kesavaraj *et al*, 2013). Dalam penyelesaian permasalahan klasifikasi terdapat beberapa metode yang dapat digunakan yaitu *classification and assiosiation rule, radom forest ¸desission tree* dan sebagainya.

Bank X dengan menghasilkan akurasi sebesar *C5.0* 87,72%, *CART* 87,27% dan *CAHID* 87,15 %. Dari ketiga metode tingkat akurasi yang lebih tinggi adalah *C5.0.* Mambang *et al* (2017) melakukan penelitian dengan membandingkan algoritme C.45, *Random Forest* dan *Chaid Decision* Penelitian terkait mengenai data *marketing* pernah dilakukan oleh Grzonka *et al.* (2009) dalam penelitiannya melakukan klasifikasi menggunakan *decision tree* untuk mendefiniskan skenario pelanggan bank dalam membuat keputusan pengaktifan deposit. Yogi (2007) melakukan penelitian dengan membandingkan metode *C5.0, CART,* dan *CHAID* pada kasus kredit *Tree* dengan menghasilkan tingkat akurasi untuk ketiga algotitme sebesar 64%, 64%, dan 62,67%.

Berdasarkan penelitian- penelitian sebelumnya, pada penelitian ini akan melakukan perbandingan tingkat akurasi dengan mengunakan algoritme C5.0 dan *Random Forest.*

TINJUAN PUSTAKA

Decision Tree

Decision tree atau pohon keputusan adalah salah satu dari metode klasifikasi dengan mengunakan struktur hirarki atau struktur pohon (Han *et al*. 2012). Fungsi dari *decision tree* yaitu sebagai *decision support tool* yang digunakan dalam pengambilan keputusan (Tsang *et al.* 2009). Kegunaan *decision tree* yaitu melakukan *break down* dalam pengambilan keputusan sehingga mempermundah untuk keputusan yang komplek atau sulit menjadi lebih simple dalam penyelesainnya. Konsep utama pada *decision tree* yaitu mengubah data menjadi *tree* dan mengubah aturan menjadi keputusan (*rule).* Proses *decision tree* terdiri dari tiga bagian yaitu *root node, internal node* dan *leaf node* (Alpaydin 2004).

Algoritme C5.0

Menurut Patil *et al.* (2012) algoritme C5.0 merupakan pengembangan dari ID3 dan algoritme C4.5. Nilai *information gain* (IG) digunakan untuk pemilihan atribut pada algorime C5.0. Pemilihan atribut dalam memecahkan objke yaitu atribut yang memiliki nilai *information gain* yang paling tinggi Nilai IG yang paling tinggi akan dipilih sebagai *parent* atau *node* selanjutnya. Menurut Han *et al*. (2012) algoritme C5.0 menggunakan persamaan *entropy* dan *information gain* sebagai berikut:

$$Info(D) = -\sum_{k=1}^{m} Pi \log_2 (Pi) \text{.....} (1)$$

Keterangan untuk persamaan diatas yaittu info (D) merupakan nilai *entropy* dari sample data D, *m* merupakan jumlah kelas atribut, *Pi* merupakan peluang untuk kelas *i* atau rasio dri kelas. D merupakan partisi *tuple* pada beberapa atribut *A* yang memiliki nilai *v* berbeda $\{a_1, a_2, \dots, a_v\}$ dari data latih. Penggunaan atribut *A* dilakukan untuk memisahkan *D* kedalam *v* partisi $\{D_1, D_2, \dots, D_v\}$. $\frac{|D_j|}{D}$ yaitu bobot partisi ke-*j*.

Menurut Han *et al* (2012) hasil nilai *entropy* digunakan untuk mengkalsifikasikan *tuple D* berdasarkan partisi *A*:

$$Info_A(D) = \sum_{j=1}^{v} \frac{|D_j|}{D} \times Info(D_j) \text{.....} (2)$$

Information gain yang didapatkan pada atribut *A* yaitu :

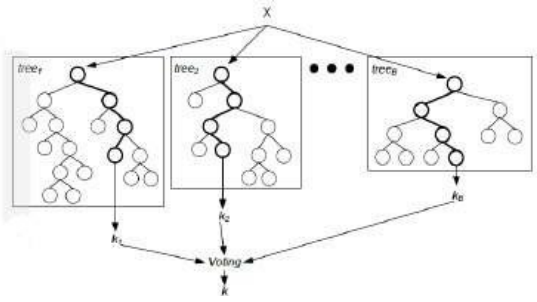
$$Gain(A) = Info(D) - Info_A(D) \text{.....} (3)$$

Gain (A) merupakan banyaknya cabang yang dihasilkan oleh *A.* Atribut *A* dengan *information gain* terbesar akan dipilih sebagai *node* (Han *et al*. 2012).

Algoritme *Random Forest*

Menurut Breiman (2001) metode *Classification and Regression Tree* (CART) mengembangkan sebuah algoritme yang dikenal sebagai algoritme *random forest* (RF) dengan menerapkan metode *bootstrap aggregating* (*bagging*) dan *random* (*ensemble tree)*. *Random forest* membentuk banyak pohon (*tree*) sehingga terbentuklah *forest* (hutan), kemudian dilakukan analisi pada pohon yang terbentuk. *Bagging* digunakan untuk mengatasi sifat ketidakstabilan pada metode klasifikasi tunggal.

Pembentukan *tree* pada RF dilakukan dengan *training* sampel data. *Sampling with replacement* dilakukan untuk mengambil data. Penggunaan variable sebagai *split* diambil secara acak (*random)*. Proses klasifikasi dilakukan setelah semua *tree* terbentuk dan penentuan hasil klasifikasi diambil dari *vote* untuk setiap masing – masing *tree. Vote* dengan nilai tertinggi akan ditetapkan sebagai pemenang. Arsitektur umum RF ditunjukan pada Gambar 1.



Gambar 1 Arsitektur umum *Random Forest* (Verikas *et al*. 2011).

Berikut ini adalah prosedur atau algoritme untuk membangun *Random Forest* pada gugus data yang terdiri dari n amatan dan p peubah penjelas (Breiman 2001; Breiman dan Cutler 2003):

1. Lakukan penarikan contoh acak berukuran n dengan pemulihan pada gugu

data. Langkah ini dinamakan dengan *bootstrap* (bag).

2. Dengan menggunakan contoh *bootstrap*, pohon dibangun sampai mencapai ukuran maksimum yaitu tanpa pemangkasan (pruning). Pembangunan pohon dilakukan dengan menerapkan *random feature selection* yaitu m peubah penjelas dipilih secara acak dengan m << p, selanjutnya pemilah terbaik dipilih berdasarkan m peubah penjelas.
3. Langkah 1 dan 2 diulangi sebanyak k kali untuk membuat sebuah *forest* yang terdiri dari k pohon.

Tahapan pembuatan model klasifikasi menggunakan algoritme *Random Forest* dilakukan setelah membuat pemodelan data latih menggunakan *package randomForest RStudio*.

METODE

Data

Penelitian ini menggunakan data telemarketing bank yang menawarkan produk berupa deposito berjangka, Fokus pada penelitian ini adalah prediksi dari hasil telemarketing. apakah client yang ditawarkan ingin berlangganan deposito berjangka atau tidak, yang nantinya akan diklasifikasikan menggunakan algoritma C5.0 dan *Random Forest* dengan target. berlangganan (yes), tidak berlangganan (No). Dataset penelitian diambil dari UCI machine learning: Bank Marketing dataset dengan jumlah data 45211 record, dan jumlah atribut adalah 17 akan ditampilkan pada Tabel 1.

Tabel 1 Dataset *Bank Marketing*

No	Atribut	Tipe	Nilai
1	Usia (*Age*)	*Numeric*	>18
2	Pekerjaan (*Job*)	Kategorik	unknown(1), administrator(2), unemployed(3), manager(4), servant(5), contractor(6) , student(7), worker(8), self-employed (9), retired(10), technician(11), service personnel(12)
3	Marital Status (*Marital*)	Kategorik	single(1) , divorced(0), married(-1)
4	Pendidikan (*Education*)	Kategorik	unknown(1), primary school(2), secondary school(3), undergraduate(4)
5	*Default*	Kategorik	yes(1), no(-1)
6	*Balance*	*Numeric*	-
7	Kredit rumah (*Housing*)	Kategorik	yes(1), no(-1)
8	Kredit pribadi (*Loan*)	Kategorik	yes(1), no(-1)
9	Kontak (*Contact*)	Kategorik	Cellular (1), telephone(-1)
10	Hari (*Day*)	Kategorik	monday(1)..., sunday(7)
11	Bulan (*Month*)	Kategorik	January(1), february(2),...december(12)
12	Durasi (*Duration*)	*Numeric*	4 sec < ; <3025 sec
13	*Campaign*	*Numeric*	1 - 50
14	*Pdays*	*Numeric*	1-871
15	*Previous*	*Numeric*	0-25
16	*Outcome*	Kategorik	failure(1), nonexist(2), succes(3)
17	y	*Binary*	ya(1), tidak (0)

Data *Preprocesing*

- **Transformasi Data**

Berdasarkan tabel diatas dapat dilihat bahwa data memiliki keragaman kategorik yang banyak, terutama untuk atribut seperti pekerjaan, dan bulan. Sehingga variable masukan berupa data katagorik harus ditransformasikan ke dalam bentuk numeric data agar dapat diproses.

- **Reduksi Data**

Data yang sudah ditransformasikan kemudian dilihat keseimbangan data perkelasnya. Hasil *summary* yang dilakukan pada atribut kelas dapat dilihat pada Tabel 2. Terlihat bahwa jumlah kelas "*no*: dan kelas "*yes*" memiliki ketidakseimbangan jumlah. Hal ini menyebabkan data yang digunakan belum cocok untuk membuat model klasifikasi sehingga memelukan tahap praproses.

Tabel 2 *Summary* atribut kelas

Summary	
No	**Yes**
39922	5289

Teknik yang digunakan untuk menyeimbangkan atribut kelas adalah teknik *undersampling*. Teknik *undersampling* bekerja dengan melakukan *samplinng* pada kelas yang jumlahnya lebih banyak agara jumlahnya sama dengan kelas yang jumlahanya lebih sedikit. Teknik *undersampling* ini dipilih dikarenakan jumlah kelas "*no*" mencapai 40.000 data. Jika dilakukan *oversampling* maka jumlah

keseluruan data akan mencapai 80.000 data. Hal ini akan membuat proses pembuatanmodel klasifikasi menjadi lama. Sehingga digunakanlah teknik *undersampling* yang membuat jumlah datanya menjadi sekitar 10.000 data dengan 5289 data kelas "*no*" dan 5289 data kelas "*yes*".

- **Pembagian data**

Data yang sudah direduksi kemudian dilakukan pembagian data menggunakan *K – fold cross validation* pada perangkat lunak R dengan menggunakan *packaget caret*. *K – fold cross validation* akan melakukan perulangang percobaan sebnayak 10 kali, sehingga didapatkan data uji dan data latih yang berbeda.

- **Lingkungan Pengembangan**

Penelitian ini dilakukan menggunakan perangkat keras dan perangkat lunak sebagai berikut :

a. Perangkat keras :
   - Amd a6.
   - Ram 4 GB dan *hardisk* 1 TB.

b. Perangkat Lunak :
   - Sistem Operasi *Windows* 10
   - Microsoft Excel 2010.
   - Rstudio.

HASIL DAN PEMBAHASAN

Pelatihan algoritme C5.0 dan *Random Forest* menggunakan *package* yang telah disediakan oleh *software* R. *package* yang disediakan yaitu *package* C.50 untuk klasifikasi menggunakan algoritma C5.0 dan *package randomForest* untuk klasifikasi *Random Forest.*

- **Algoritme C5.0**

Algoritme C5.0 menghasilkan model klasifikasi berupa model berbasis aturan dan model *decision tree*. Model berbasis aturan memangkas aturan dan menyerhanakannya sehingga aturan yang dihasilkan lebih sedikit dari model pohon keputusan. Pada implementasinya dilakukan pengujian untuk tingkat akurasi data latih dan data uji tiap *fold.*

a. Akurasi Data Latih

Data latih digunakan untuk membangun model klasifikasi pada Data Bank Marketing. Tabel 3 akan menampilkan hasil akurasi data latih untuk setiap *fold*.

Tabel 3. Akurasi data latih model *decision tree*

Fold	***Kappa***	**Akurasi**
1	0.80	80%
2	0.80	84%
3	0.80	80%
4	0.80	81%
5	0.81	87%
6	0.80	89%
7	0.78	86%
8	0.80	88%
9	0.79	90%
10	0.79	90%
Average	**0.79**	**85.5%**

Tabel 3 menunjukan rata – rata nilai akurasi data latih yang diperoleh menggunakan model *decision tree* untuk 10-*fold* sebesar 85,5 % dengan nilai koefisien *kappa* sebesar 0.79. Table 4 menunjukan rata – rata nilai akurasi menggunakan model bebabasis aturan sebesar 90.4% dengan nilai *kappa* 0.79. Ini menunjukan bahwa algoritme C5.0 tingkat akurasi yang lebih baik menggunakan model berbasis aturan karena pada model ini banyaknya aturan ada akan dipangkas dan disederhanakan, sehingga mendapatkan model keputusan yang lebih baik.

Tabel 4 Akurasi data latih model berbasis aturan

Fold	***Kappa***	**Akurasi**
1	0.79	90%
2	0.77	89%
3	0.79	90%
4	0.78	89%
5	0.79	90%
6	0.78	90%
7	0.76	88%
8	0.78	89%
9	0.78	90%
10	0.76	99%
Average	**0.79**	**90.4%**

b. Akurasi Data Uji

Data uji digunakan untuk menghitung akurasi *decision tree.*

Tabel 5 Akurasi data Uji model *decision tree*

Fold	**Akurasi**
1	80%
2	84%

3	80%
4	81%
5	87%
6	89%
7	86%
8	88%
9	90%
10	90%
Avarage	85.5%

Tabel 6 Akurasi data uji model berbasis aturan

Fold	**Akurasi**
1	81%
2	83%
3	80%
4	82%
5	80%
6	87%
7	89%
8	86%
9	88%
10	91%
Avarage	84.7%

Hasil yang diperoleh untuk akurasi menggunakan model *decision tree* sebesar 85.5% dan untuk nilai akurasi berbasis aturan sebesar 84.7%. dan dapat dilihat dari table ketika mencapai fold ke 10 tingkat akurasi untuk kedua model mencaoai 90%.

- ***Random Forest***

a. Data Latih

Data latih digunakan untuk membangun model klasifikasi pada Data Bank Marketing.

Tabel 7 Akurasi data latih *Random Forest*

Fold	**Akurasi**
1	96%
2	96%
3	95%
4	93%
5	93%
6	92%
7	73%
8	82%
9	85%
10	61%
Avarage	86.6%

Tingkat akurasi pada data latih *random forest* paling tinggi terjadi pada *fold* 1 dan 2 mencapa 96%, tetapi untuk fold selanjutnya megalami penurunan sampai ke *fold 10.* Nilai akurasi paling terkecil terjadi pada *fold* 10. Untuk rata – rata nilai akurasi pada data latih sebesar 86.6%.

a. Data Uji

Data uji digunakan untuk menghitung akurasi *Random Forest.*

Tabel 8 Akurasi data uji *Random Forest*

Fold	**Akurasi**
1	97%
2	97%
3	97%
4	97%
5	97%
6	97%
7	97%
8	97%
9	97%
10	97%
Avarage	97%

Hasil akurasi untuk data uji *random forest* sebesar 97%, nilai ini selalu konvergen dari mulai *fold* awal sampai *fold* ke 10.

- **Perbandingan Model Klasifikasi**

Evaluasi hasil klasifikasi dilakukan untk melihat nilai akurasi. Nilai akurasi yang didapatkan menggunakan *Confusion matrix.* Table 4 menunjuka hasil akurasi menggunakan algoritme C5.0 dan *Random Forest*. Hasil yang ditunjukan merupakan hasil rata – rata yang didapatkan dari 10-*fold.*

Table 9 Perbandingan akurasi data latih algoritme C5.0 dan *Random Forest*

Algoritme	**Akurasi Keseluruhan**
C5.0	90.4%
Random Forest	86.6%

Table 10 Perbandingan akurasi data uji algoritme C5.0 dan *Random Forest*

Algoritme	**Akurasi Keseluruhan**
C5.0	85.5%
Random Forest	97%

Dari hasil perbandingan algoritme C50. dan *Random forest* dengan nilai akurasi data latih dan data uji yang ditunjukan pada table 4. Table 4.1 menunjukan bahwa untuk nilai akurasi data

latih yang lebih baik yaitu algoritme C5.0 sebesar 90.4%, perbedaan nilai akurasi sebesar 3.8%. Sedangkan untuk nilai akurasi data uji yang ditunjukan oleh table 4.2 *Random Forest* lebih unggul sebesar 97%, nilai ini selalu konvergen dari 1-*fold* sampai 10-*fold.* Perbedaan nilai akurasi 11.5%.

Dalam membangun pohon (*tree)* algoritme *random forest* menggunkan indeks gini untuk membagi kiteria. Dalam mengembangkan pohon dilakukan dengan cara *binary split*. Variable yang digunakan *split* dipilih secara acak. Sedangkan algoritme C5.0 dalam membuat pohon menggunakan data *multi-split* yaitu data numeric akan diubah menjadi *binary split* dan untuk klasifikasi menggunakan data kategorikal. Penggunan *binary split* menyebabkan atribut akan muncul beberapa kali pada pohon *(tree).*

KESIMPULAN

Hasil penelitian perbandingan algoritme C5.0 dan *random forest* untuk data latih dan data uji menggunakan *K – fold cross validation*, menghasilkan tingkat akurasi diatas 90%. Algoritme C5.0 lebih unggul untuk tingkat akurasi data latih sebesar 90.4 % , nilai akurasi terbaik untuk algoritme C5.0 terjadi pada *fold* ke 10 .Sedangkan random forest lebih unggul untuk data uji sebesar 97%, nilai ini konvergen dari 1-*fold* sampai 10-*fold.*

REFERENSI

Alpaydin E. 2004. *Introduction to Machine Learning.* Cambridge (MA): The MIT Press.

Breiman L. 2001. *Random Forest*s. *Machine Learning*. 45: 11–13.

Breiman L, Cutler A. 2003. Manual–setting up, using, and understanding *Random Forest*s V4.0.

Han J, Kamber M, Pei J. 2012. Data Mining Concepts and Techniques Second Edition. San Francisco (US): Morgan Kaufmann Publisher.

Han J, Kamber M. 2006. Data Mining Concepts and Techniques. San Fransisco: Morgan Kaufmann Publishers.

Grzonka D, Suchacka G, Borowik B. 2016. Application of selected supervised classification methods to bank marketing campaign. *Information Syatems in Management*. vol.5 (1), pp. 36-48.

Kesavaraj.G , S Sukumaran. 2013. A Study On Classification Techniques in Data Mining. *International conference on computing, communications and networking teknologies.*

Liao shu-shiesn, Hu Da-Chian.2011. *African Journal of Business Management.* 5(34). pp.12929-12938.

Mabang, Byna A. 2017. Analisis perbandingan algoritma C.45, Random Forest dengan Chaid decision tree untuk klasifikasi tingkat kecemasan ibu hamil. *Seminar nasional teknologi informasi dan multimedia.* 2(1), pp 103-108.

Patil N, Lathi R, Chitre V. 2012. Customer card classification based on C5.0 and CART algorithms. International Journal of Enggineering Research.

Petrison Lisa, Wang Paul. 1993. From relationships to relationship marketing: applying database technology to public relation. *Public jurnal Review.*19(3), pp 235-245.

Tsang S, Kao B, Yip K, Ho WS, Lee SD, 2009. Decision trees for uncertain data. Data Engineering, 2009. *ICDE "09. IEEE 25th International Conference on*, 2009, pp. 441–444.

Turban E. 2005. Decision Support Systems and Intelligent Systems (Sistem Pendukung Keputusan dan Sistem Cerdas). Andi Offset: Yogyakarta

Verikas A, Gelzinis A, Becausekiene M. 2011. Mining data with Random Forest: asurvey and result of new tests. Pattern Recognition. 44(2): 330-349