Jurnal Sains Komputer & Informatika (J-SAKTI)
Volume (1) No. 2 September 2017
ISSN:2548-9771/EISSN: 2549-7200
http://tunasbangsa.ac.id/ejurnal/index.php/jsakti

ANALISA TERHADAP PERBANDINGAN ALGORITMA *DECISION TREE* DENGAN ALGORITMA *RANDOM TREE* UNTUK PRE-PROCESSING DATA

Saifullah[1], Muhammad Zarlis[2], Zakaria[3], Rahmat Widia Sembiring[4]
[1]STIKOM Tunas Bangsa Pematangsiantar, Jln Jend. Sudirman Blok A No.1/2/3
[2]Fasilkom TI USU, Jl. Universitas No. 9A Kampus *USU*, Medan
[3]Universitas Methodist, Kampus I Jl. Hang Tuah No. 8 *Medan*
[4]Politeknik Negeri Medan, Jl. Almamater No. 1, Kampus USU Medan

Abstract

Preprocessing data is needed some methods to get better results. This research is intended to process employee dataset as preprocessing input. Furthermore, model decision algorithm is used, random tree and random forest. Decision trees are used to create a model of the rule selected in the decision process. With the results of the preprocessing approach and the model rules obtained, can be a reference for decision makers to decide which variables should be considered to support employee performance improvement.

Keywords*: Pre-processing Data, Decision Tree, Random Tree, Random Forest.*

Abstrak

Preprocessing data sangat dibutuhkan beberapa metode untuk mendapatkan hasil yang lebih baik. Penelitian ini ditujukan mengolah dataset karyawan sebagai inputan *preprocessing*. Selanjutnya digunakan model algoritma *decision tree*, *random tree* dan *random forest.* Pohon keputusan digunakan untuk membuat model aturan yang dipilih dalam proses mengambil keputusan. Dengan hasil pendekatan *preprocessing* dan model aturan yang didapat, dapat menjadi referensi bagi pengambil keputusan untuk mengambil keputusan variabel mana yang harus diperhatikan untuk mendukung peningkatan kinerja karyawan.

Kata Kunci: *Pre-processing Data, Decision Tree, Random Tree, Random Forest.*

1. PENDAHULUAN

Dengan meningkatnya teknologi informasi (TI) jumlah data semakin tinggi yang akan diproses dan disimpan dalam database, sehingga tingkat kesulitannya dalam memprosesan cukup tinggi. Para peneliti banyak menggunakan *data mining* untuk mengatasi masalah pengelompokan dan pengolahan database yang sangat besar. Teknik *data mining* secara garis besar dapat dibagi dalam dua kelompok: verifikasi dan *discovery*. Metode verifikasi umumnya meliputi teknik-teknik statistik seperti *goodness of fit*, dan *analisis variansi*. Metode *discovery* lebih lanjut dapat dibagi atas model prediktif dan model deskriptif. Teknik prediktif melakukan prediksi terhadap data dengan menggunakan hasil-hasil yang telah diketahui dari data yang berbeda. Sementara itu, model deskriptif bertujuan mengidentifikasi pola-pola atau hubungan antar data dan memberikan cara untuk mengeksplorasi karakteristik data yang diselidiki [1].

Dalam pengolahan data, penulis ingin membuat perbandingan metode dalam memprosesnya, diantaranya menggunakan model *preproceesing* data *Handle missing value as category* dan *Missing value replenishment* yang dipaplikasikan pada pohon keputusan *decision tree*, *random tree* dan *random forest*. Dengan menggunakan perbandingan model ini, penelitian ini akan memberikan aturan preprocessing mana yang paling efisien untuk diaplikaksikan pada *decision tree, random tree* dan *random forest*.

2. METODOLOGI PENELITIAN

2.1. Pengertian Data Mining

Data mining adalah suatu istilah yang digunakan untuk menemukan pengetahuan yang tersembunyi di dalam database. Data mining merupakan proses semi otomatik yang menggunakan teknik statistik, matematika, kecerdasan buatan, dan *machine learning* untuk mengekstraksi dan mengidentifikasi informasi pengetahuan potensial dan berguna yang bermanfaat yang tersimpan di dalam database besar [2].

Enam fase CRISP-DM (*Cross Industry Standard Process for Data Mining*) [3].

1. Fase Pemahaman Bisnis (*Business Understanding Phase*)
2. Fase Pemahaman Data (*Data Understanding Phase*)
3. Fase Pengolahan Data (*Data Preparation Phase*)
4. Fase Pemodelan (*Modeling Phase*)
5. Fase Evaluasi (*Evaluation Phase*)
6. Fase Penyebaran (*Deployment Phase)*

2.2. Pengertian Decision Tree

Decision tree merupakan salah satu metode klasifikasi yang menggunakan representasi struktur pohon (*tree*) di mana setiap node merepresentasikan atribut, dimana cabangnya merepresentasikan nilai dari atribut, dan daun merepresentasikan kelas. *Node* yang paling atas dari *decision tree* disebut sebagai *root*. *Decision tree* merupakan metode klasifikasi yang paling populer digunakan. Selain karena pembangunannya relatif cepat, hasil dari model yang dibangun mudah untuk dipahami. Pada decision tree terdapat 3 jenis node, yaitu *Root Node, Internal Node, Leaf node.*

2.3. Pengertian *Random Tree*

Operator ini mempelajari tentang sebuah pohon keputusan. Operator ini hanya menggunakan subset acak atribut untuk setiap perpecahan. Operator ini mempelajari tentng pohon keputusan yakni data nominal dan numerik. Pohon keputusan adalah metode klasifikasi yang kuat yang dapat dengan mudah dipahami. Operator pohon Random bekerja sama dengan Quinlan C4.5 atau CART memilih subset acak atribut sebelum diterapkan. Ukuran subset ditentukan oleh parameter rasio bagian.

2.4. Pengertian *Random Forest*

Operator ini menghasilkan satu set sejumlah tertentu pohon random yaitu menghasilkan forest (hutan; kumpulan pohon) acak. Model yang dihasilkan adalah model suara pilihan dari semua pohon. Operator *Random Forest* menghasilkan satu set pohon acak. Pohon-pohon acak yang dihasilkan dengan cara yang persis sama seperti

operator Acak Pohon menghasilkan pohon. Model hutan yang dihasilkan mengandung sejumlah tertentu dari model pohon acak. Jumlah pohon parameter menentukan jumlah yang diperlukan pohon. Model yang dihasilkan adalah model suara pilihan dari semua pohon acak. Untuk informasi lebih lanjut tentang pohon acak silakan mempelajari operator *random Tree.*

2.4. Preprocessing data

Pre-processing data adalah proses mengubah data ke dalam format yang sederhana, lebih efektif, dan sesuai dengan kebutuhan pengguna. Indikator yang dapat digunakan sebagai referensi adalah hasil lebih akurat, waktu komputasi yang lebih pendek, juga data menjadi lebih kecil tanpa mengubah informasi di dalamnya.

2.4.1. Jenis-Jenis metode *Preprecessing* data

Ekstraksi fitur adalah perubahan dari data dimensi tinggi ke dimensi rendah. Transformasi data dapat linier dan nonlinier dimensi data, tujuannya adalah pemetaan data ke dimensi yang lebih rendah. Beberapa algoritma telah lakukan, untuk supervised learning: LDA, CCA, PLS, LSI, SVD, dan *unsupervised learning*: PCA, ICA, FastICA][4][5].

2.5. Handle Missing Value as Category

Operator ini memetakan nilai-nilai tertentu dari atribut yang dipilih ke nilai baru. Operator ini dapat diterapkan pada kedua atribut numerik dan nominal. Operator ini dapat digunakan untuk menggantikan nilai nominal (misalnya mengganti nilai 'hijau' dengan nilai ‘warna_hijau’) serta nilai-nilai numeric.

Tapi, salah satu penggunaan operator ini dapat melakukan pemetaan untuk atribut hanya satu jenis. Sebuah pemetaan tunggal dapat ditentukan dengan menggunakan parameter menggantikan *what* dan *replace by* seperti dalam operator *replace*.

2.6. Missing Value Replenishment

Operator ini menggantikan nilai-nilai yang hilang dalam contoh atribut yang dipilih oleh pengganti yang ditentukan. Operator ini menggantikan nilai-nilai yang hilang dalam contoh atribut yang dipilih oleh pengganti yang ditentukan. Nilai-nilai yang hilang dapat diganti dengan nilai minimum, maksimum atau rata-rata atribut tersebut. Nol juga dapat ditempatkan di tempat nilai-nilai yang hilang. Setiap nilai pengisian juga dapat ditentukan sebagai pengganti nilai-nilai yang hilang [6].

2.7. Metode Penelitian

Rancangan penelitian ini pertama kali dilakukan dengan memahami data *(observasi)* untuk mempelajari klasifikasi data yang di gunakan untuk proses preprocessing data. Hasil pengamatan kemudian dibuat menjadi scenario implementasi pohon keputusan yang mendukung, kemudian mendapatkan aturan yang sesuai untuk digunakan. Data yang sudah diolah merupakan data input pada proses pohon keputusan. Selanjutnya data input diproses dengan menggunakan *Decision Tree, Random Tree dan Random Forest.*

3. HASIL DAN PEMBAHASAN

Adapun hasil percobaan training dan testing data dapat dilihat pada bagian berikut ini.

3.1. Sampel Data

Dalam pengujian data set ini yang terdiri dari 10 data dengan rincian sebagai berikut:

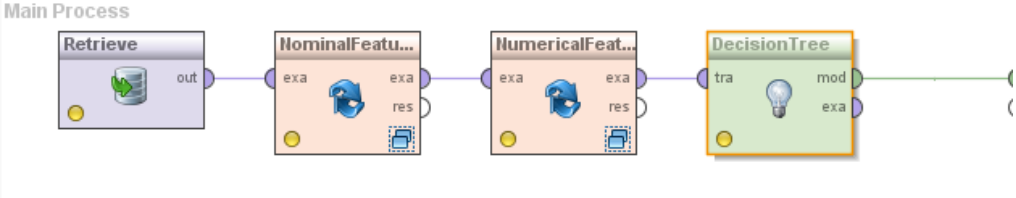
Tabel 1. Deskripsi data

Tipe	Nama	Tipe data	Deskripsi	Uraian	Missing value
label	Class	nominal	mode = good (26), least = bad (14)	bad (14), good (26)	0
regular	Duration	integer	avg = 2.103 +/- 0.754	[1.000 ; 3.000]	1
regular	wage-inc-1st	real	avg = 3.621 +/- 1.331	[2.000 ; 6.900]	1
regular	wage-inc-2nd	real	avg = 3.913 +/- 1.281	[2.000 ; 7.000]	10
regular	wage-inc-3rd	real	avg = 3.767 +/- 1.415	[2.000 ; 5.100]	28
regular	col-adj	nominal	mode = none (14), least = tcf (4)	tcf (4), none (14), tc (6)	16
regular	working-hours	integer	avg = 37.811 +/- 2.717	[27.000 ; 40.000]	3
regular	Pension	nominal	mode = none (8), least = ret_allw (3)	none (8), empl_contr (7), ret_allw (3)	22
regular	standby-pay	integer	avg = 6.143 +/- 4.845	[2.000 ; 13.000]	33

3.2. Hasil *Preprocessing* dengan *Handle missing value as category*

3.2.1. DecisionTree

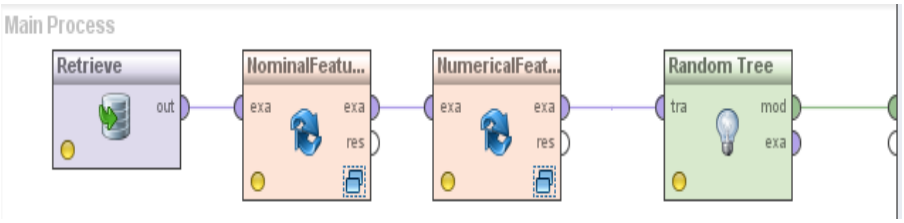
Model *preprocessing* dengan grafik dari *software* rapidminer yang akan digunakan dapat dilihat pada gambar 1.



Gambar 1. Model Preprocessingnya *Handle missing value as category* dengan implementasi *decision tree*

3.2.2. Random Tree

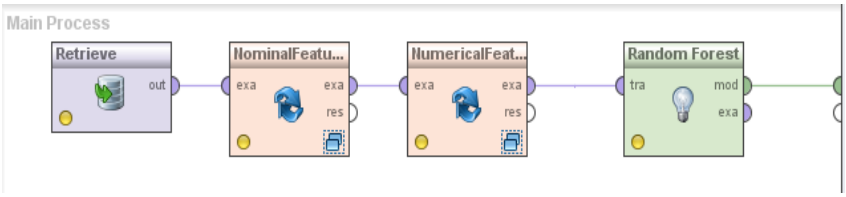
Model *preprocessing* dengan grafik dari *software* rapidminer yang akan digunakan dapat dilihat pada gambar 2.



Gambar 2. Model Preprocessingnya *Handle missing value as category* dengan implementasi *random tree*

3.2.3. Random Forest

Model *preprocessing* dengan grafik dari *software* rapidminer yang akan digunakan dapat dilihat pada gambar 3.

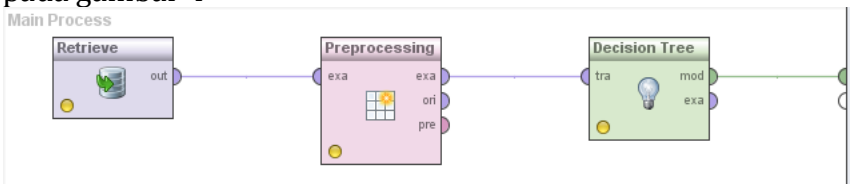


Gambar 3. Model Preprocessingnya *Handle missing value as category* dengan implementasi *random forest*

3.3. Preprocessing dengan *Missing value replenishment*

3.3.1. DecisionTree

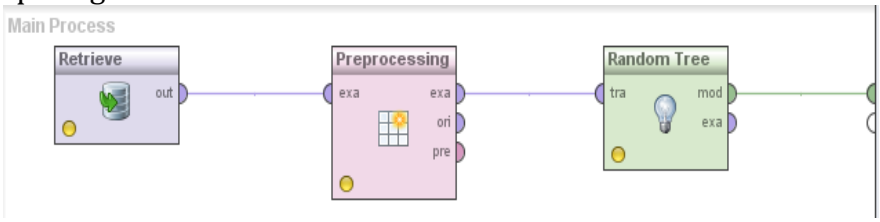
Model *preprocessing* dengan grafik dari *software* rapidminer yang akan digunakan dapat dilihat pada gambar 4



Gambar 4. Model Preprocessingnya *Missing value replenishment* dengan implementasi *decision tree*

3.3.2. Random Tree

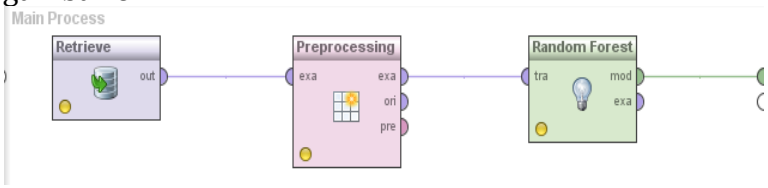
Model *preprocessing* dengan grafik dari *software* rapidminer yang akan digunakan dapat dilihat pada gambar 5.



Gambar 5. Model Preprocessingnya *Missing value replenishment* dengan implementasi *random tree*

3.3.3. Random Forest

Model *preprocessing* dengan grafik dari *software* rapidminer yang akan digunakan dapat dilihat pada gambar 6.



Gambar 6. Model Preprocessingnya *Missing value replenishment* dengan implementasi *random tree*

4. SIMPULAN

Penelitian ini menghasilkan beberapa kesimpulan sebagai berikut :

a. Dengan menerapkan model *preprocessing* data *Handle missing value as category* dan *Missing value replenisment* data hasil *pre-processing* dapat diaplikasikan pada pohon keputusan *Decision tree, random tree* dan *Random Forest.*
b. Diperoleh suatu model aturan yang dapat memperlihatkan aturan keterhubungan antara *wage_inc_1s*t dengan *staturoty holidays* dan *working hours*
c. Dalam studi kasus *labour realtion* ditemukan bahwa jika *statutrory holidays* akan diberikan jika *wage_inc_1st* lebih besar dari 2.0.
d. *Preprocessing* ternyata memberi efek pada efisiensi implementasi pohon keputusan.

DAFTAR PUSTAKA

[1] Dunham, M.H.2003. Data Mining Introductory and advanced topics. News Jersey: Prentice Hall.
[2] Turban, E., Aronson, J. E. & Liang, T., 2005, *Decision Support Sistems and Intellegent Sistems (Sistem Pendukung Keputusan dan Sistem Cerdas).*
[3] Larose D, T., 2006, *Data Mining Methods and Models,* Jhon Wiley & Sons, Inc. Hoboken New Jersey.
[4] Sembiring R dan Zain J, 2010, Rancangan *Pre-Processing* Data Multidimensi Berdasarkan Analisa Komponen, Proceeding The 5th IMT-GT International Conference on Mathematics, Statistic, and Their Application.
[5] Sembering S, Embong A, Mohammad, M. A, Furqan M, "*Improving Student Academic Performace by An Application of Data Mining Techniques*", Proceeding The 5th IMT-GT International Conference on Mathematics, Statistic, and Their Application (ICMSA 2009).
[6] Juan S, Xi-Zhao W., (2005), *An Initial Comparison on Noise Resisting Between Crisp and Fuzzy Decision Trees,* IEEE 2005 Proceeding of the Fourth International Conference on Machine Learning and Cybernetics.