JURNAL SWABUMI, Vol. 7 No.1 Maret 2019, pp. 62~66
ISSN: 2355-990X
E-ISSN**:**2549-5178

ANALISIS KOMPARASI KLASIFIKASI ALGORITMA C4.5 DAN NAïVE BAYES PADA PREDIKSI KEBERHASILAN SOFTWARE REUSE

Ita Yulianti[1]

[1] Universitas Bina Sarana Informatika
e-mail: ita.iyi@bsi.ac.id

Abstraksi

Software reuse (penggunaan kembali perangkat lunak) diyakini sebagai salah satu pendekatan yang paling efektif untuk memperbaiki proses perangkat lunak secara signifikan, meningkatkan kualitas perangkat lunak dan konsistensi penerapan, dan mengurangi biaya pengembangan dan perawatan. Keberhasilan *software reuse* ditentukan pada kemampuan untuk memprediksi variabilitas yang dibutuhkan dalam aset masa depan. Oleh karena itu, untuk memastikan berhasil atau tidaknya penerapan *software reuse* diperlukan suatu analisis yang dapat memprediksi permasalahan tersebut sebagai cara pendekatan terbaik yaitu salah satunya dengan menggunakan data mining. Ada dua metode data mining yang digunakan dalam penelitian ini, yaitu Algoritma C4.5 dan Naïve Bayes. Berdasarkan hasil yang diperoleh, Algoritma C4.5 menjadi metode klasifikasi terbaik dalam memprediksi keberhasilan *software reuse.*
Kata Kunci : Algoritma C4.5, Naïve Bayes, *Software Reuse*

Abstract

Software reuse is believed to be one of the most effective approaches to significantly improve software processes, improve software quality and consistency of implementation, and reduce development and maintenance costs. The success of software reuse is determined by the ability to predict the variability needed in future assets. Therefore, to ensure the success or failure of the application of reuse software, an analysis that can predict these problems is needed as the best approach, one of which is by using data mining. There are two data mining methods used in this study, namely C4.5 and Naïve Bayes Algorithms. Based on the results obtained, C4.5 algorithm becomes the best classification method in predicting the success of software reuse.
Keywords: C4.5 Algorithm, Naïve Bayes, Software Reuse

1. Pendahuluan

Dengan meningkatnya perkembangan sistem perangkat lunak dalam perusahaan menyebabkan kompleksitas industri TI dipaksa untuk menemukan cara mempersingkat proses pembangunan. *Software reuse* (penggunaan kembali perangkat lunak) diyakini sebagai salah satu pendekatan yang paling efektif untuk memperbaiki proses perangkat lunak secara signifikan, meningkatkan kualitas perangkat lunak dan konsistensi penerapan, dan mengurangi biaya pengembangan dan perawatan (Antovski & Florinda Imeri, 2013).

Software reuse telah menjadi topik yang sangat diminati oleh komunitas perangkat lunak karena potensi manfaatnya, yang mencakup peningkatan kualitas produk, penurunan biaya dan jadwal produk (Jalender, et al., 2010). Kualitas dapat ditingkatkan dengan menciptakan sistem perangkat lunak dari segala bentuk sistem yang telah digunakan sebelumnya, termasuk produk dan proses, serta model kualitas dan produktivitas. Dengan mengimplementasikan *software reuse* sebesar 1% saja, Departemen pertahanan AS sendiri bisa menghemat $300 juta setiap tahun (Prakash, et al., 2012). Dari hal ini dapat disimpulkan bahwa *software reuse* dapat mengurangi biaya, waktu, usaha dan resiko, serta meningkatkan produktivitas, kualitas, kinerja dan interoperabilitas.

Namun, tidak sedikit resiko yang harus dihadapi dalam melakukan penerapan *software reuse*. Keberhasilan *software reuse* ditentukan pada kemampuan untuk memprediksi variabilitas yang dibutuhkan dalam aset masa depan. Penelitian diperlukan untuk mengidentifikasi dan memvalidasi ukuran *reusabilitas*, termasuk cara yang baik untuk memperkirakan jumlah potensi penggunaan kembali (Prakash, et al., 2012).

Diterima Maret 1, 2019; Revisi Maret 9, 2019; Disetujui Maret 19, 2019

Kurang optimalnya metode prediksi akan memperlambat dan menyebabkan resiko kegagalan *software reuse* menjadi lebih tinggi. Oleh karena itu, untuk memastikan berhasil atau tidaknya penerapan *software reuse* diperlukan suatu analisis yang dapat memprediksi permasalahan tersebut sebagai cara pendekatan terbaik untuk pencegahan dan penanggulangan pada sistem yang dibangun. Saat ini, data mining menjadi alat yang semakin penting untuk mentransformasikan data ini menjadi informasi (Prakash, et al., 2012). Dengan bantuan teknik data mining, dapat diketahui pola suatu permasalahan berdasarkan data yang sudah ada sehingga jika sudah diketahui faktor-faktor yang mempengaruhi suatu prediksi permasalahan tersebut maka akan memudahkan untuk klasifikasi pola keputusan suatu prediksi.

Algoritma data mining yang diusulkan pada penelitian ini adalah Algoritma C4.5 yang merupakan algoritma klasifikasi pohon keputusan yang banyak digunakan karena memiliki kelebihan utama dari algoritma yang lainnya. Kelebihan Algoritma C4.5 dapat menghasilkan pohon keputusan yang mudah diinterprestasikan, memiliki tingkat akurasi yang dapat diterima, efisien dalam menangani atribut bertipe diskret dan dapat menangani atribut bertipe diskret dan numerik (Kamagi dan Seng Hansun, 2014). Tidak hanya sampai disitu, untuk memperoleh hasil yang maksimal, Klasifikasi Naïve Bayes juga diusulkan untuk dapat membandingkan, model klasifikasi mana yang memberikan hasil yang lebih baik untuk prediksi keberhasilan *software reuse.*

2. Metode Penelitian

A. Jenis Penelitian

Jenis penelitian yang dilakukan pada penelitian ini adalah jenis penelitian eksperimen. Metode ini menguji kebenaran sebuah hipotesis dengan statistik dan menghubungkannya dengan masalah penelitian. Model eksperimen ini bertujuan untuk memprediksi keberhasilan *software reuse* berdasarkan kumpulan-kumpulan data dan variabel yang sudah ditetapkan.

B. Teknik Pengumpulan Data

Dalam penelitian ini metode pengumpulan data untuk mendapatkan sumber data yang digunakan adalah metode pengumpulan data sekunder. Data utama diperoleh dari PROMISE *Repository*, sedangkan data pendukung didapatkan dari buku, jurnal dan publikasi.

Penelitian ini dilakukan dengan mengusulkan model, melakukan eksperimen dengan menguji model yang diusulkan, evaluasi dan validasi, serta membandingkan hasil dari pengujian model tersebut untuk mendapatkan model terbaik yang digunakan dalam objek penelitian tersebut. Dataset yang digunakan adalah dataset NASA (*National Aeronautics and Space Administration*) MDP (*Metrics Data Program*) pada *reuse/predicting successful reuse* yang berisi 27 atribut dan 1 kelas yang terdiri dari kelas *success* dan *failure* dengan spesifikasi dataset sebagai berikut:

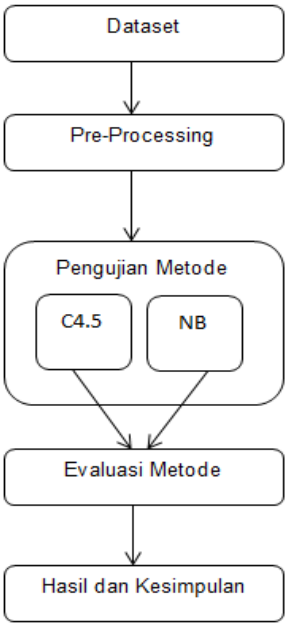
Tabel 1. Spesifikasi Dataset *reuse/predicting successful reuse*

Attribute	**Value**
Project ID	A i
Software Staff	L,M,S
Overall Staff	L,X M,S
Type of Software Production	product-family,isolated
Software and Product	product,alone,process,NA
SP Maturity	high,middle,low
Application Domain	TLC, SE-Tools, Engine_Controller, FMS, ATC, TS, Space, Manufacturing, Measurement, Finance, Book-Keeping
Type of Software	Technical,Business,Embedded-RT,Non-Embedded-RT
Size of Baseline	L,M,S,not_available
Development Approach	OO,proc,not_available
Staff Experience	high,middle,low,not_available
Top Management Commitment	yes,no
Key Reuse Roles Introduced	yes,no,NA
Reuse Processes Introduced	yes,no,NA
Non-Reuse Processes Modified	yes,no,NA
Repository	yes,NA
Human Factors	yes,no
Reuse Approach	tight,loose,NA
Work Products	D+C,C,R+D+C,NA
Domain Analysis	yes,no,NA
Origin	ex-novo,as-is,reeng,NA
Independent Team	yes,no,NA

When Assests Developed	before,justintime,NA
Qualification	yes,no,NA
Configuration Management	yes,no,NA
Rewards Policy	no,yes
# assests	51_to_100,21_to_50,100+,1_to_20,NA
Success or Failure	success,failure

Sumber: PROMISE Repository

Sedangkan untuk langkah-langkah yang digunakan dalam penelitian ini dapat dilihat pada gambar dibawah ini:



Gambar 1. Langkah-Langkah Penelitian

Langkah pertama, pengumpulan dataset dilakukan melalui alamat web http://promise.site.uottawa.ca/SERepository/datasets/reuse.arff. Selanjutnya, untuk dapat mempermudah pembacaan data pada tools RapidMiner, dilakukan penulisan data kembali dalam bentuk yang sama dan disimpan dengan format Excel 97-2003 Worksheet (.xls).

Tahapan berikutnya yaitu pre-processing data dengan melakukan pengisian *missing value* menggunakan teknik *Replace Data Missing* dalam tools RapidMiner. Selanjutnya, dataset diuji menggunakan klasifikasi Algoritma C4.5 dan Naïve Bayes dengan pemilihan data training dan data testing secara otomatis menggunakan metode 10 *fold cross validation.* Terakhir, hasil pengujian kemudian dianalisis sehingga diperoleh metode terbaik untuk prediksi keberhasilan *software reuse* berdasarkan komparasi dari evaluasi model yang dihasilkan pada *tools* RapidMiner.

3. Hasil dan Pembahasan

3.1. Eksperimen dan Pengujian Model

Algoritma C4.5 merupakan algoritma yang digunakan untuk membentuk pohon keputusan. Didalam algoritma C4.5 ini, pohon-pohon keputusan yang dibentuk berdasarkan kriteria-kriteria pembentuk keputusan (Nofriansyah, 2014). Pada tahapan ini, pembuatan pohon keputusan dilakukan dengan cara menghitung jumlah kelas yang *success* dan *failure* dari masing-masing kelas. Kemudian, hitung jumlah Entropy dan Gain menggunakan persamaan berikut:

Persamaan 1.

$$Entropy(S) = \sum_{i=1}^{n} - pi * log_2\ pi$$

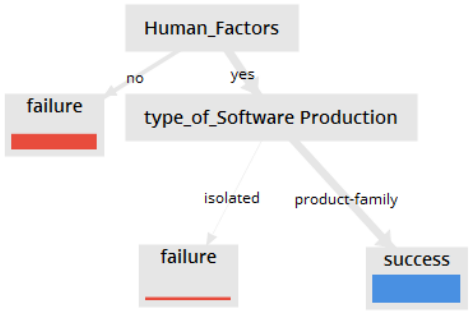
Keterangan:
S : Himpunan Kasus
n : Jumlah partisi S
pi : proporsi dari Si terhadap S

Persamaan 2.

$$Gain(S,A) = Entropy(S) - \sum_{i=1}^{n} \frac{|S_i|}{|S|} * Entropy(Si)$$

Keterangan:
S : Himpunan Kasus
A : Atribut
n : Jumlah partisi atribut A
|Si| : Jumlah kasus pada partisi ke-i
|S| : Jumlah kasus dalam S

Berdasarkan hasil perhitungan tersebut dapat diperoleh pohon keputusan yang terbentuk (lihat Gambar 2.), dimana pohon keputusan tersebut terbentuk dengan memanfaatkan Framework RapidMiner Studio versi 9.2.000.



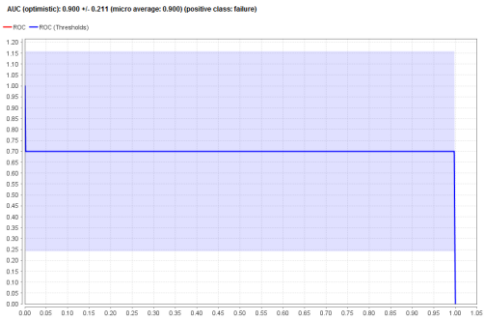
Gambar 2. Pohon Keputusan yang Dihasilkan Model Algoritma C4.5

Selain pohon keputusan, hasil dari uji coba yang dilakukan dalam Framework RapidMiner juga menghasilkan nilai accuracy dan nilai AUC (Area Under Curve) yang disajikan dalam gambar berikut:

accuracy: 95.00% +/- 15.81% (micro average: 95.83%)

	true success	true failure	class precision
pred. success	15	1	93.75%
pred. failure	0	8	100.00%
class recall	100.00%	88.89%	

Gambar 3. *Confusion Matrix* Algoritma Klasifikasi C4.5



Gambar 4. Nilai AUC Algoritma C4.5 dalam Grafik ROC

Berdasarkan Gambar 3. nilai akurasi yang dihasilkan *Confusion Matrix* Algoritma Klasifikasi C4.5 dalam prediksi keberhasilan *software reuse* diperoleh sebesar 95.00%, sedangkan untuk nilai AUCnya (Lihat Gambar 4.) sebesar 0,900 yang termasuk kedalam tingkat Excellent Classification

Berbeda dengan Algoritma C4.5, Klasifikasi Naïve Bayes menghasilkan model berupa nilai probabilitas yang dinyatakan dalam tabel distribusi. Berikut tabel yang dihasilkan oleh *tools* RapidMiner dari model Naïve Bayes untuk prediksi keberhasilan *software reuse*.

SimpleDistribution

```
Distribution model for label attribute class


Class success (0.625)
26 distributions

Class failure (0.375)
26 distributions
```

Gambar 5. *Simple Ditribution* yang dihasilkan Klasifikasi Naïve Bayes

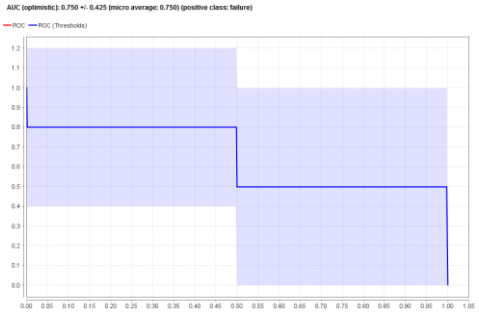
Dari tabel distribusi yang dihasilkan (Lihat Gambar 5.) dapat disimpulkan bahwa jika nilai probabilitas yang dihasilkan 0.625 maka hasil prediksinya "*success"*, sebaliknya jika nilai probabilitasnya 0.375 maka hasil prediksinya "*failure"*.

Sama halnya dengan Algoritma C4.5, hasil dari uji coba Naïve Bayes yang dilakukan juga menghasilkan nilai accuracy dan nilai AUC (Area Under Curve).

accuracy: 91.67% +/- 18.00% (micro average: 91.67%)

	true success	true failure	class precision
pred. success	14	1	93.33%
pred. failure	1	8	88.89%
class recall	93.33%	88.89%	

Gambar 6. *Confusion Matrix* Algoritma Klasifikasi Naïve Bayes



Gambar 7. Nilai AUC Naïve Bayes dalam Grafik ROC

3.2. Hasil Analisis Evaluasi

Berdasarkan hasil eksperimen yang telah dilakukan dalam penelitian ini, hasil uji pada model C4.5 dibandingkan dengan Naïve Bayes pada prediksi keberhasilan *software reuse* dapat dilihat pada tabel berikut:

Tabel 2. Pengujian Algoritma Klasifikasi C4.5 dan Naïve Bayes

	Accuracy	**AUC**
C4.5	95%	0,900
Naïve Bayes	91,67%	0,750

Dari hasil pengujian yang disajikan dalam tabel 4., dengan dilakukan evaluasi baik secara *confusion matrix* maupun *ROC curve* terbukti bahwa pengujian yang dilakukan menggunakan Klasifikasi Algoritma C4.5 memiliki nilai akurasi yang lebih tinggi dibandingkan dengan Naïve Bayes. Akurasi yang dihasilkan untuk model algoritma C4.5 sebesar 95% dengan nilai AUC 0,900 dan akurasi model Naïve Bayes sebesar 91,67% dengan AUC 0,750. Berdasarkan nilai tersebut dapat diketahui bahwa selisih akurasi sebesar 3,33% dan selisih AUC sebesar 0,15.

4. Kesimpulan

Berdasarkan hasil eksperimen yang diperoleh, dapat ditarik kesimpulan bahwa penggunaan model klasifikasi dalam prediksi keberhasilan *software reuse* lebih akurat menggunakan Algoritma C4.5 dibandingkan dengan Naïve Bayes. Hal tersebut dapat dilihat dari nilai akurasi dan AUC yang dihasilkan untuk model algoritma C4.5 sebesar 95% dengan nilai AUC 0,900 sedangkan akurasi model Naïve Bayes sebesar 91,67% dengan AUC 0,750.

Tetapi, dari hasil tersebut penggunaan klasifikasi pada prediksi keberhasilan *software reuse* dapat dilakukan pengembangan untuk penelitian selanjutnya, diantaranya:

a. Untuk pemilihan atribut pada penelitian lanjutan dapat menggunakan seleksi optimasi sehingga atribut yang digunakan tidak terlalu banyak tetapi bisa meningkatkan kinerja pengklasifikasi.
b. Pada penelitian lanjutan dapat melibatkan sejumlah data yang lebih besar dan menggunakan pengklasifikasi lainnya, seperti Neural Network, SVM dan Logistic Regression serta mengimplementasikan model yang dihasilkan menjadi GUI sehingga hasil pengukuran yang akan didapatkan lebih baik lagi.

Referensi

Antovski, Ljupcho dan Florinda Imeri. 2013. "Review of Software Reuse Processes". *IJCSI – International Journal of Computer Science Issues,* Vol. 10, *Issue* 6, No. 2 – ISSN (Print)*:* 1694-0814, ISSN (Online): 1694-0784 – www.IJCSI.org, p. 83-88.

E. Fayad, Mohamed dan Charles A. Flood III. 2016. "*Unified Software Engineering Reuse (USER) using Stable Analysis, Design and Architectural Patterns*". FTC - *Future Technologies Conference,* IEEE, p. 706-711.

Jalender, B., Dr. A. Govardhan dan Dr. P Premchand. 2010. "*A Pragmatic Approach To Software Reuse*". JATIT - *Journal of Theoretical and Applied Information Technology*, p. 87-96.

Kamagi, David Hartanto dan Seng Hansun. 2014. Implementasi Data Mining dengan Algoritma C4.5 Memprediksi Tingkat Kelulusan Mahasiswa. ISSN: 2085-4552. Tangerang: UTLIMATICS, Vol. VI, No. 1, Juni 2014.

Nofriansyah, D. 2014. "Konsep Data Mining VS Sistem Pendukung Keputusan". Yogyakarta: Deepublish.

Prakash, B.V. Ajay, D V Ashoka dan V N Manjunath Aradhya. 2012. "*Application of Data Mining Techniques for Software Reuse Process*". Procedia Technology 4 – Elsevier Ltd. p. 384 – 389.

PROMISE Repository. 2004. Reuse/Predicting Successful Reuse. Dipetik January 16, 2019, dari PROMISE Repository: https://promise.site.uottawa.ca/SERepository/datasets-page.html

Putri, Sukmawati Anggraeni. 2017. "*Integrasi Teknik Smote Bagging Dengan Information Gain Pada Naive Bayes Untuk Prediksi Cacat Software"*. Jurnal Ilmu Pengetahuan dan Teknologi Komputer, Vol. 2. No. 2 Februari 2017, E-ISSN: 2527-4864 Hal. 22-31.

Xin. TAO dan LIU Yang. 2017. " *A Framework of Software Reusing Engineering Management*" SERA IEEE, p. 277-282.