ANALISIS DAN IMPLEMENTASI PERBANDINGAN ALGORITMA C.45 DENGAN NAÏVE BAYES UNTUK PREDIKSI PENAWARAN PRODUK

Comparative Analysis and Implementation of C4.5 and Naïve Bayes Algorithm

For Product Offering Prediction

Asri Khoirunnisa[1], Budhi Irawan[2], R. Rumani M.[3]

Program Studi S1 Sistem Komputer, Fakultas Teknik Elektro, Universitas Telkom

[1]nisaasrikhoirunnisa@gmail.com, [2]budhi.ira1@gmaill.com, [3]rumani@telkomuniversity.ac.id

Abstrak

Metode *telemarketing* sering digunakan oleh banyak perusahaan untuk mencapai target keuntungan perusahaan karena efisien dari sisi waktu dan biaya yang dikeluarkan. Akan tetapi, pelaksanaan metode telemarketing tidaklah efektif jika tidak diimbangi dengan prediksi penjualan produk yang baik.
Penelitian tugas akhir ini menerapkan metode *data mining* untuk membandingkan 2 metode algoritma yang berbeda, yaitu algoritma Naïve Bayes dan algoritma C4.5 yang dikaji untuk mendapatkan nilai akurasi yang terbaik berdasarkan Data *History* Penawaran Produk periode November 2014 sebagai data latih. Kedua algoritma diterapkan dalam aplikasi berbasis *desktop* yang dibangun dengan bahasa pemrograman Java.
Hasil pengujian pada aplikasi perbandingan algoritma menunjukkan nilai akurasi rata-rata dari algoritma Naïve Bayes adalah 57% dan algoritma C4.5 adalah 47.7% yang berarti bahwa algoritma Naïve Bayes adalah algoritma terbaik untuk diterapkan dalam sistem prediksi penawaran produk Kredit Tanpa Agunan.

Kata Kunci: *Data Mining*, Naïve Bayes, C4.5, prediksi penawaran produk.

Abstract

Many companies had been use the telemarketing method to achieve their profit target because of the time efficiency and less of cost. But, telemarketing method implementation was not effective sometimes if prediction of the product offering did not implemented in it.
This final project research is implementing data mining method to compare 2 different algorithm methods, Naïve Bayes and C4.5 which are being examined to get the best accuracy point based on History of Product Offering Data in November 2014 as the data learning. Both of the algorithms were implemented to a desktop based application that was built with Java programming language.
The result of the algorithm comparing application shows us that Naïve Bayes algorithm's accuracy point is 57% and C4.5 algorithm is 47.7% which means that Naïve Bayes algorithm is the best algorithm that should be implemented to predict the product offering of Kredit Tanpa Agunan.

Keywords: Data Mining, Naïve Bayes, C4.5, product offering predicition.

1. Pendahuluan

Bank XYZ merupakan salah satu bank di Indonesia yang memiliki bagian *telemarketing* yang bertugas melakukan penjualan melalui layanan telepon. Data *history* penawaran produk *telemarketing* pada Bank XZY sudah sangat banyak, sehingga sulit untuk melakukan prediksi penawaran produk dengan metode konvensional. Dengan demikian, dibutuhkan metode baru untuk mengatasi masalah tersebut. Dalam hal ini, ada dua metode yang akan diuji untuk memprediksi penawaran produk yaitu metode dengan algoritma Neive Bayes dan algoritma C4.5. Kedua algoritma tersebut akan dikaji untuk mendapatkan akurasi terbaik dalam memprediksi penawaran produk dalam satu periode, sehingga Bank XYZ dapat mengoptimalkan penjualan produknya.

Prediksi penawaran produk dengan dua algoritma ini diharapkan dapat mendukung dan membantu kinerja bagian *telemarketing* dalam pengoptimalisasian penjualan produk dari sisi waktu dan biaya.

2. Tinjauan Pustaka

2.1 *Data Mining*

Data mining merupakan teknik yang digunakan untuk mengeksplorasi data secara menyeluruh secara otomatis. Data mining berkonotasi sebagai pencarian informasi bisnis yang berharga dari sumber data yang besar [11].

2.2 CRISP-DM

Merupakan metode data mining yang menyediakan proses standar pelaksanaan dara mining untuk menyelesaikan masalah dalam sebuah bisnis atau penelitian. CRISP-DM tidak menentukan standar tertentu karena setiap data yang akan dianalisis akan diproses kembali pada fase-fase di dalamnya [6].

2.3 Prediksi

Prediksi adalah proses peramalan kejadian yang akan datang berdasarkan parameter tertentu untuk mengurangi ketidakpastian dalam suatu kondisi serta membat suatu tolak ukur untuk memperkirakan kejadian yang akan datang berdasarkan pola yang telah terjadi di masa lampau [5].

2.4 Algoritma Klasifikasi

Klasifikasi adalah sebuah proses untuk mencari model atau fungsi yang menjelaskan dan membedakan kelas atau konsep dari data, dengan tujuan untuk menggunakan model dan melakukan prediksi dari kelas suatu objek dimana tidak diketahui label dari kelas tersebut [9].

2.4.1 Algoritma Naïve Bayes

Algoritma Naïve Bayes merupakan pengklasifikasi dengan metode probabilitas dan statistik yang dapat digunakan untuk memprediksi probabilitas ekanggotaan suatu kelas [15]. Algoritma Naïve Bayes didasarkan pada teorema Bayes yang memiliki kemampuan klasifikasi serupa dengan decision tree dan neural network [7].

Persamaan teorema Naïve Bayes dinyatakan pada rumus di bawah ini [15]:

$$P(C|V) = \frac{P(V|C) \times P(C)}{P(V)} \quad \text{.....(2.1) [15]}$$

Keterangan:

V : data parameter

C : hipotesis keputusan

P(C|V) : probabilitas keputusan berdasarkan komdisi data parameter

P(V|C) : probabilitas parameter berdasarkan kondisi hipotesis keputusan

P(C) : probabilitas jumlah keputusan

P(V) : probabilitas parameter

2.4.2 Algoritma ID3

Algoritma C4.5 yang digunakan dalam penulisan ini merupakan algoritma yang dikembangkan dari algoritma ID3 (*Iterative Dychotomizer version* 3) yang kemudian telah dapat dikembangkan menjadi algoritma C5.0 atau yang lebih dikenal dengan sebutan *See*5. Algoritma ID3 adalah algoritma pembelajaran pohon keputusan (*decision tree learning*) yang paling dasar. Algoritma ID3 melakukan pencarian secara menyeluruh (*greedy*) pada semua kemungkinan pohon keputusan. Pertumbuhan cabang-cabang pohon keputusan pada algoritma ID3 dilakukan sampai pohon keputusan tersebut mampu mengkliasifikasikan sampel data secara akurat dengan tingkat kebenaran 100% [16].

2.4.3 Algoritma C4.5

Algoritma C4.5 merupakan pengembangan algoritma ID3 yang digunakan untuk membentuk pohon keputusan. Pohon keputusan merupakan cara memprediksi atau mengklarifikasi yang sangat kuat karena pohon keputusan dapat membagi kumpulan data yang besar menajdi himpunan-himpunan *record* yang lebih kecil dengan menerapkan serangkaian aturan keputusan. Algoritma C4.5 dapat mengolah data numerik (kontinyu) dan diskret dan mampu menangani atribut yang hilang, serta menghasilkan aturan-aturan yang mudah diinterpretasikan. Algoritma C4.5 akan mengubah data menjadi pohon keputusan dan atura-aturan keputusan [13].

Penentuan akar pohon keputusan ditentukan dari hasil evaluasi nilai *entropy* yang dinyatakan dengan rumus sebagai berikut ini [1]:

$$Entropy(x) = \sum_{x} -x \times \log_2 x \quad \text{..... (2.2) [1]}$$

Dimana:

x : himpunan kasus

n : jumlah partisi S
S : proporsi dari partisi S terhadap S

Setelah mendapatkan nilai *entropy* untuk suatu kumpulan sampel data, selanjutnya dilakukan pengukuran efektifitas suatu parameter dalam mengklasifikasikan dara. Ukuran efektifitas ini disebut dengan *information gain* yang secara matematis dirumuskan sebagai berikut [1]:

$$Gain(S,A) \equiv Entropy(S) - \sum_{i=0}^{n} \frac{|S_i|}{|S|} \times Entropy(S_i) \quad \text{..... (2.3) [1]}$$

Dimana:

S : himpunan kasus
A : parameter (atribut)
n : jumlah partisi parameter A
$|S_x|$: proporsi S_x terhdap S
|S| : jumlah kasus dalam S

Information gain akan mengalami masalah ketika parameter memiliki nilai yang sangat bervariasi yang akan menghasilkan pohon keputusan yang *overfit*. Untuk mengatasi hal tersebut, dilakukan alternative pengukuran lain yang dinamakan *Gain Ratio*. *Gain Ratio* dihitung berdasarkan *Split Information* yang dirumuskan sebagai berikut [16]:

$$SplitInformation(S,A) = \sum_{i=1}^{n} -\frac{|S_i|}{|S|} \log_2 \frac{|S_i|}{|S|} \quad \text{..... (2.4) [16]}$$

Dimana:

S : himpunan kasus
A : parameter
n : jumlah partisi parameter A
$|S_x|$: proporsi S_x terhadap S
|S| : jumlah kasus dalam S

Selanjutnya, *Gain Ratio* dirumuskan sebagai *Information Gain* dibagi dengan *Split Information* sebagai berikut [16]:

$$GainRatio(S,A) = \frac{Gain(S,A)}{SplitInformation(S,A)} \quad \text{..... (2.5) [16]}$$

2.5 *Confusion Matrix*

Confusion Matrix adalah *tools* yang digunakan untuk mengevaluasi model klasifikasi yang digunakan untuk memperkirakan objek yang benar dan yang salah. Hasil prediksi akan dibandingkan dengan kelas asli dari data tersebut. Confusin Matrix mengevaluasi kinerja mdel berdasarkan pada kemampuan akurasi prediktif suatu model. Akurasi prediktif merupakan parameter yang mengukur ketepatan aturan klasifikasi yang dihasilkan dalam mengklasifikasikan test set berdasarkan atribut yang ada ke dalam kelasnya [2].

Tabel 2.1 Confusion Matrix

Classification	Predicted class	
	Class = Yes	Class = No
Class = Yes	True Positives (TP)	False Negatives (FN)
Class = No	False Positives (FP)	True Negatives (TN)

True Positives adalah jumlah data yang diklasifikasikan *Yes* dan diprediksi *Yes* dengan benar. *False Positives* adalah jumlah data yang diklasifikasikan *No* dan diprediksi *Yes*. *False Negatives* adalah jumlah data yang diklasifikasikan *Yes* dan diprediksi *No*. *True Negatives* adalah jumlah data yang diklasifikasikan *No* dan diprediksi *No* dengan benar [2].

Evaliasi dan validasi hasil dihitung dengan menggunakan rumus *accuracy*, *precision*, *sensitifity*, dan *spesificity*. Nilai akurasi dinyatakan dalam persentase. Jika akurasi mencapai angka 100%, hal tersebut berarti bahwa semua kasus yang diprediksi diklasifikasikan seluruhnya sengan benar. Nilai akurasi, presisi, sensitifitas, dan spesifisitas yang dihitung dengan rumus sebegai berikut [10]:

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \quad \text{.....(2.6) [10]}$$

$$Precision = \frac{TP}{TP + FP} \quad \text{..... (2.7) [10]}$$

$$Sensitivity = \frac{TP}{TP + FN} \quad \text{..... (2.10) [10]}$$

$$\text{[illegible]} = \frac{\text{[illegible]}}{\text{[illegible]}} \quad \text{..... (2.11) [10]}$$

3. Perancangan Sistem

3.1 Deskripsi Sistem

Jika implemetasi dari perancangan berhasil dilakukan, sistem akan memiliki spesifikasi sebagai berikut ini:

1. Mampu melakukan proses "Learning", dimana sistem mempelajari alur algoritma C4.5 dan Naïve Bayes dalam mengklasifikasikan data;
2. Mampu melakukan proses "Forecasting", dimana sistem mampu melakukan prediksi dari data yang disajikan berdasarkan proses "Learning" sebelumnya;
3. Mampu membandingkan performa kedua algoritma melalui nilai akurasi.

Implementasi dilakukan dengan membuat sebuah aplikasi yang dapat digunakan pengguna untuk memprediksi penawaran produk berdasarkan data penawaran produk periode Novermber 2014. Masukan aplikasi berupa file csv yang berisi 667 data penawaran produk yang telah dilakukan oleh divisi Telemarketing sebelumnya. Keluaran yang diharapkan dari sistem berupa daftar prediksi penawaran produk yang telah diolah oleh algoritma dengan nilai akurasi terbaik.

3.2 Perancangan Sistem

3.2.1 Perancangan Parameter dan Nilai

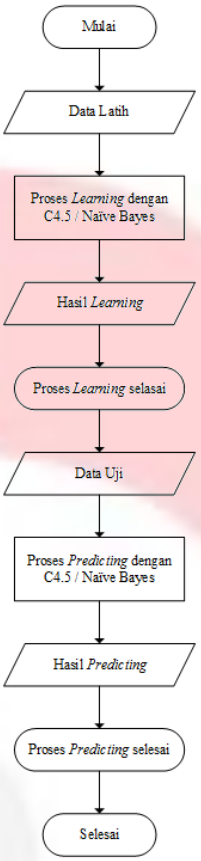
Di bawah ini merupakan daftar parameter beserta nilai-nilainya yang digunakan pada sistem.

Tabel 3.1 Parameter dan nilai yang digunakan dalam sistem

No.	Parameter	Nilai	Keterangan
1	Usia	21-30	Rentang usia dalam satuan tahun
		31-40	Rentang usia dalam satuan tahun
		41-50	Rentang usia dalam satuan tahun
		51-60	Rentang usia dalam satuan tahun
		>60	Rentang usia di atas 60 tahun
2	Keanggotaan	<1	Masa keanggotaan kartu kredit kurang dari satu tahun
		1-3	Masa keanggotaan kartu kredit 1 sampai 3 tahun
		>3	Masa keanggotaan kartu kredit lebih dari 3 tahun
3	Penggunaan Kartu Kredit	Baik	Penggunaan kartu kredit dinilai baik
		Sedang	Penggunaan kartu kredit dinilai sedang
		Buruk	Penggunaan kartu kredit dinilai buruk
4	Pekerjaan	BUMN	Bekerja di perusahaan Badan Usaha Milik Negara
		BUMS	Bekerja di perusahaan Badan Usaha Milik Swasta
		Wiraswasta	Penyelenggaraan usaha secara mandiri
		Pensiunan	Tidak lagi bekerja di BUMN atau BUMS
5	Penghasilan	3-5	Penghasilan dalam juta rupiah per bulan
		5-10	Penghasilan dalam juta rupiah per bulan
		>10	Rentang penghasilan lebih dari 10 juta rupiah per bulan
6	Tanggungan	0	Anggota keluarga yang ditanggung tidak ada
		1-3	Keluarga yang ditanggung 1 sampai 3 orang
		>3	Keluarga yang ditanggung lebih dari 3 orang
7	Produk	KKTAR	Kredit Tanpa Agunan Regular
		KKTAM	Kredit Tanpa Agunan Mitra Karya
		KKTAP	Kredit Tanpa Agunan Payroll

3.2.2 Diagram Alir Sistem

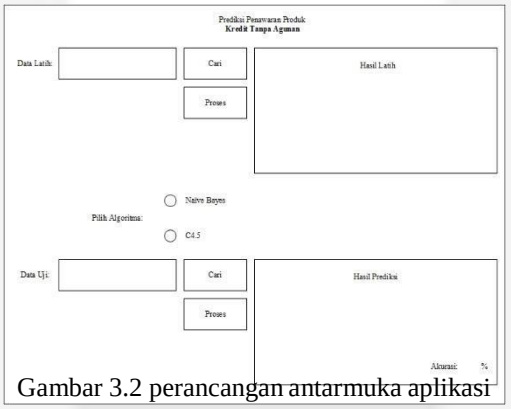
Cara kerja sistem yang dibangun dapat dilihat melalui gambar diagram alir sistem di atas.



Gambar 3.1 diagram alir system

3.2.3 Perancangan Antarmuka

Berdasarkan kebutuhan masukan dan keluaran pada gambaran umum aplikasi, perancangan antarmuka dari aplikasi dapat digambarkan sebagai berikut ini:



Gambar 3.2 perancangan antarmuka aplikasi

4 Implementasi dan Pengujian

4.1 Implementasi Sistem

Dari hasil perancangan, pembuatan implementasi sistem membutuhkan perangkat keras dan perangkat lunak yang mampu mengoperasikan aplikasi Java untuk desktop. Pembuatan implementasi dilakukan dengan menggunakan aplikasi Eclipse Luna pada sistem operasi Windows 7, yang didukung dengan prosesor 2.27 GHz. Penggunaan perangkat keras atau pun lunak dengan spesifikasi lebih rendah masih dapat dilakukan, dengan persyaratan minimum sebagai berikut.

4.1.1 Perangkat Keras Minimum

Perangkat keras yang dapat digunakan memiliki persyaratan minimum sebagai berikut:

- Prosesor 1.66GHz

- RAM 1 GB

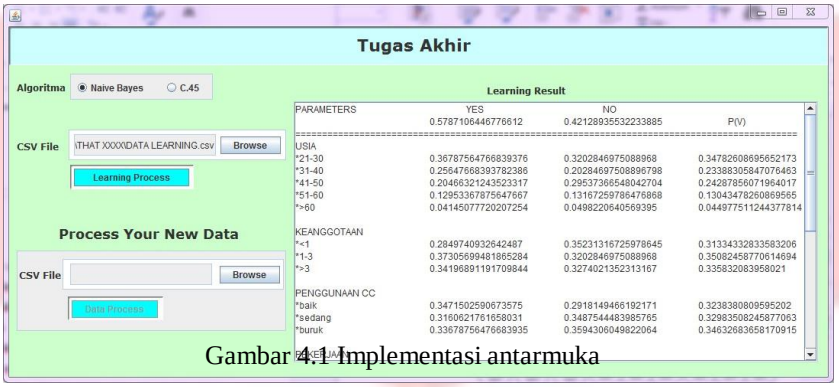
4.1.2 Perangkat Lunak Minimum

Perangkat lunak yang dapat digunakan memiliki persyaratan minimum sebagai berikut ini:

- Sistem operasi Windows XP
- JDK 1.6.0

4.1.3 Implementasi Antarmuka

Sistem dibuat dalam bentuk aplikasi desktop dengan implementasi sebagai berikut:



Gambar 4.1 Implementasi antarmuka

4.2 Pengujian Sistem

Untuk menguji performa dari kedua algoritma yang diimplementasikan ke dalam sistem, maka pengujian sistem dilakukan dengan melakukan pengujian akurasi dengan dengan jumlah data yang berbeda-beda.

Tabel 4.1 Pengujian algoritma Naïve Bayes dengan 10 data uji

Kasus Uji	Jumlah Data Latih	Jumlah Data Uji
I	667	50
II	667	120
III	667	277

Tabel 4.2 Confusion Matrix untuk kasus uji III

Naïve Bayes	Yes	No
Yes	95	33
No	109	40
C4.5	**Yes**	**No**
Yes	63	64
No	82	68

Tabel 4.3 performa algoritma untuk kasus uji III

Algoritma	Sensitifity	Specificity	Precision	Accuracy
Naïve Bayes	74%	27%	47%	49%
C4.5	50%	45%	43%	47%

5 Kesimpulan dan Saran

5.1 Kesimpulan

Dari pengujian yang telah dilakukan, ditarik kesimpulan sebagai berikut ini:

1. Implementasi kedua algoritma klasifikasi berhasil dilakukan, dimana kedua algoritma memberikan performa yang cukup dilihat dari akurasi 60% untuk algoritma Naïve Bayes, dan 48% untuk algoritma C4.5, pada kasus uji I dengan data uji terendah, yaitu 50 data. Pada kasus uji II dengan data uji sebanyak 120 data, performa algoritma Naïve Bayes mengalami peningktan menjadi 62%, sedangkan algoritma C4.5 menunjukan performa yang sema dengan sebelumnya yaitu 48%. Pada kasus uji III dengan data uji sebanyak 227 data, performa algoritm Naïve Bayes mengalami penurunan menjadi 49%, beigut juga dengan algoritma C4.5 menajdi 47%.
2. Performa tertinggi algoritma Naïve Bayes terjadi pada kasus uji II dengan 120 data uji. Performa tertinggi algoritma C4.5 terjadi pada kasus uji I dan II dengan 50 dan 120 data uji.

3. Dari ketiga kasus uji yang dilakukan, didapatkan akurasi rata-rata algoritma Naïve Bayes yaitu 57% dan algoritma C4.5 yaitu 47.7%. Maka, algoritma Naïve Bayes adalah algortima terbaik yang dapat diterapkan untuk studi kasus prediksi penawaran produk Kredit Tanpa Agunan yang sudah dibahas dikeseluruhan tugas akhir ini.

5.2 **Saran**

Dari aplikasi yang dibangun tentunya masih perlu pengembangan agar bisa lebih baik. Saran untuk melakukan pengembangan pada aplikasi ini adalah sebagai berikut:

1. Pengembangan aplikasi berikutnya dapat dikembangakan dengan implementasi algoritma selain algoritma Naïve Bayes dan C4.5.
2. Dalam penelitian ini, penulis menggunakan 7 parameter sebagai masukan untuk proses klasifikasi. Untuk penelitian lebih lanjut dapat menggunakan jumlah atribut yang berbeda.

DAFTAR PUSTAKA

[1] Abdillah, Sigit. Penerapan Algoritma *Decision Tree* C4.5 untuk Diagnosa Penyakit *Stroke* dengan Klasifikasi *Data Mining* pada Rumah Sakit Santa Maria Pemalang. Semarang. Program Studi Teknik Informatika Universitas Dian Nuswantoro.

[2] Andriani, Anik. 2012. Penerapan Algoritma C4.5 pada Program Klasifikasi Mahasiswa *Dropout*. Seminar Nasional Matematika. Jakarta: AMIK BSI.

[5] Hartatik. 2015. Penerapan Algoritma *Learning Vector Quantization* untuk Prediksi Nilai Akademis Menggunakan Instrumen AMS (*Academic Motivation Scale)*. Yogyakarta. Teknik Informatika STMIK AMIKOM Yogyakarta.

[6] Imtiyaz, Muhammad Zain. 2015. Analisis dan Implementasi *Framework* CRISP-DM untuk Mengetahui Perilaku Data Transaksi Pelanggan. Bandung. Program Studi Sistem Komputer Universitas Telkom.

[7] Jananto, Arif. 2013. Algoritma Naïve Bayes untuk Mencari Perkiraan Waktu Studi Mahasiswa. Program Studi Sistem Informasi Universitas Stikubank. Jurnal Teknologi Informasi Dinamik Volume 18.

[9] Leidiyana, Henny. 2013. Penerapan Algoritma K-NN (*K-Nearest Neighbors*) untuk Penentuan Resiko Kredit Kepemilikan Kendaraan Bermotor. Program Pasca Sarjana Magister Ilmu Komputer Stmik Nusa Mandiri.

[10] Maurina, Dina Dan Ahmad Zainul Fanani S.Si, M.Kom. Penerapan *Data Mining* untuk Rekomendasi Beasiswa Pada SMA Muhammadiyah Gubug Menggunakan Algoritma C4.5. Semarang. Jurusan Teknik Informatika FIK Udinus.

[11] Moertini, Veronika S. 2002. *Data Mining* Sebagai Solusi Bisnis. Bandung. Jurnal Integral VOL. 7 NO. 1. Universitas Katolik Parahyangan.

[13] Prastya, Faid Ari. Penereapan Algoritma C4.5 untuk Prediksi Jurusan Siswa SMAN 3 Rembang. Semarang. Jurusan Teknik Informatika Fasilkom Udinus.

[15] Sugianti, Devi. Algoritma Bayesian *Classifiction* untuk Memprediksi Heregistrasi Mahasiswa Baru di STMIK Widya Pratama. Pekalongan. Program Studi Sistem Informasi STMIK Widyapratama.

[16] Suyanto, ST, MSc. 2011. *Artificial Intelligence – Searching*, *Reasoning*, *Planning*, dan *Learning*. Bandung. Informatika.