**JURNAL INFORMATIKA**

**Optimasi Prediksi Keputusan Calon Nasabah Potensial menggunakan** **Algoritma C 4.5 berbasis Particle Swarm Optimization**

**Saruni Dwiasnati1 , Yudo Devianto2**

Universitas Mercubuana Jakarta

Fakultas Ilmu Komputer

Jl. Meruya Selatan 1 No. 1, Kembangan, 11650, Jakarta Barat

[Saruni.Dwiasnati@mercubuana.ac.id1](mailto:Saruni.Dwiasnati@mercubuana.ac.id1)

[Yudo.Devianto@mercubuana.ac.id2](mailto:Yudo.Devianto@mercubuana.ac.id2)

**Abstrak**

Dalam era globalisasi, persaingan bisnis menjadi sangat tajam dan berwarna, baik dipasar nasional maupun di pasar internasional. Meningkatnya permintaan pasar mengakibatkan daya saing dan jumlah pesaing pelaku bisnis juga meningkat, setiap perusahaan di tuntut untuk selalu memperhatikan kebutuhan dan keinginan konsumen serta berusaha memenuhi apa yang mereka harapkan dengan cara yang lebih memiliki nilai jual atau variatif dari pada yang dilakukan oleh para pesaing lainnya. Calon Nasabah Asuransi memiliki definisi satu atau beberapa yang mengambil manfaat dari produk fisik dengan berbagai variabel yang digunakan dalam penentuan seperti harga, kualitas, jumlah produk, tempat dan layanan yang dapat disediakan berdasarkan hasil untuk mencapai hasil yang ditargetkan. Berdasarkan hal tersebut, data mining di nilai dapat membantu permasalahan dalam penentuan Calon Nasabah Potensial untuk melakukan prediksi menggunakan tekhnik klasifikasi. Salah satu tekhnik klasifikasi yang digunakan pada penulisan ini yaitu decision tree khususnya C4.5. Pada penelitian sebelumnya yang dilakukan oleh penulis bahwa C4.5 memiliki hasil accuracy yang kurang maksimal karena algoritma C4.5 merupakan algoritma yang lemah, maka di perlukanlah peningkatan akurasi. Particle Swarm Optimization (PSO) merupakan algoritma optimasi yang dapat digunakan untuk meningkatkan algoritma lainnya agar lebih optimal hasil yang diketahui. Penulisan ini menggunakan tools Rapidminner versi 8 sebagai media untuk menguji data yang akan diolah untuk mendapatkan hasil accuracy.

**Kata kunci : Datamining, Calon Nasabah Asuransi, C4.5, Particel Swarm Optimization.**

***Abstract***

In the era of globalization, business competition has become very sharp and colorful, both in the national market and in the international market. Increasing market demand has resulted in increased competitiveness and the number of competitors for business, every company is demanded to always pay attention to the needs and desires of consumers and try to fulfill what they expect in a way that has more selling value or variability than what other competitors do. Prospective Insurance Customers have a definition of one or several who take advantage of physical products with various variables used in the determination such as price, quality, number of products, places and services that can be provided based on results to achieve targeted results. Based on this, data mining in value can help problems in determining potential customer candidates to make predictions using classification techniques. One classification technique used in this writing is the decision tree especially C4.5. In a previous study conducted by the author that C4.5 has a less than optimal accuracy because the C4.5 algorithm is a weak algorithm, it requires an increase in accuracy. Particle Swarm Optimization (PSO) is an optimization algorithm that can be used to improve other algorithms for more optimal known results. This writing uses Rapidminner version 8 tools as a medium to test data that will be processed to get accuracy results.

**Keywords: Datamining, Prospective Insurance Customers, C4.5, Particel Swarm Optimization.**

------------------------------------------------------------------------------------------------------------------

[http://ejournal.bsi.ac.id/ejurnal/index.php/j](http://ejournal.bsi.ac.id/ejurnal/index.php/)i/article/view/4389

1. **Pendahuluan**

Pertumbuhan bisnis di era globalisasi seperti saat ini dapat di bilang cukup pesat karena semua serba terbuka dalam penginformasiaan tentang data atau alur bisnis bahkan ada sebuah perusahaan yang memberi tahukan tentang keuntungan yang diperoleh oleh perusahaan sebagai wadah keterbukaan terhadap pelanggannya. Era globalisasi seperti saat ini membawa dampak yang cukup besar terhadap kemajuan disegala bidang tidak terkecuali dibidang teknologi informasi seperti bidang industri maupun bidang jasa. Era globalisasi mau tidak mau juga membawa dampak perubahan yang besar terhadap tingkat persaingan antara perusahaan satu dengan perusahaan lain dalam berbagai bidang yang ada, sehingga perusahaan-perusahaan tersebut selalu memberikan value-value yang lebih variatif untuk mendapatkan hasil yang terbaik agar dapat berkembangan di tengahnya persaingan pasar bebas.

Ketersediaan data yang banyak dalam sebuah bidang dan kebutuhan akan informasi atau pengetahuan dalam bidang tertentu dapat menjadi acuan sebagai pendukung pengambilan keputusan untuk membuat solusi bisnis dan dukungan infrastruktur di segala bidang merupakan cikal- bakal dari lahirnya teknologi *data mining*. Penggunaan teknik *data mining* diharapkan dapat membantu percepatan proses pengambilan keputusan, memungkinkan perusahaan untuk mengelola informasi yang terkandung didalam data nasabah dan menjadi sebuah pengetahuan (*knowledge*) yang baru untuk menentukan keputusan yang dibutuhkan.

Berawal dari beberapa disiplin ilmu yang ada, data mining memiliki tujuan untuk memperbaiki teknik tradisional yang dapat menghasilkan sesuatu yang jauh lebih informatif sehingga bisa menangani: Jumlah data yang sangat besar, Dimensi data yang tinggi dan Data yang heterogen dan berbeda sifat [ Ratniasih, 2019 ].

Setiap metode tersebut memiliki algoritma-algoritma yang terbaik yang dapat digunakan untuk memproses dataset yang ada. Namun dalam penelitian ini akan dilakukan optimasi hasil prediksi dari algoritma yang telah dilakukan dipenelitian sebelumnya yaitu C4.5 yang berbasis Particle Swarm Optimization untuk konversi data training Keputusan Calon Nasabah Potensial.

Data mining merupakan suatu langkah dalam melakukan Knowledge Discovery in Databases (KDD). Knowledge discovery

sebagai suatu proses terdiri atas pembersihan data (data cleaning), integrasi data (data integration), pemilihan data (data selection), transformasi data (data transformation), data mining, evaluasi pola (pattern evaluation) dan penyajian pengetahuan (knowledge presentation). Data mining mengacu pada proses untuk menambang (mining) pengetahuan dari sekumpulan data yang sangat besar untuk menghasilkan sebuah pengetahuan baru dari bidang tertentu [Firdaus, 2017].

Pada Data Mining terdapat 3 bagian yaitu: Assosiation, Klasifikasi dan Clustering. Assosiation memiliki definisi sebuah proses yang digunakan untuk menemukan suatu hubungan yang terdapat pada nilai atribut dari sekumpulan data yang dimiiki, sedangkan Klasifikasi adalah teknik yang dilakukan untuk memprediksi class atau properti dari setiap instance data, dan Clustering sendiri memliki makna pengelompokan data tanpa berdasarkan kelas data tertentu ke dalam kelas objek yang sama sesuai dengan topik yang diangkat.

Prediksi memiliki kemiripan definisi dengan klasifikasi, akan tetapi data yang digunakan pada klasifikasi berdasarkan perilaku atau nilai yang diperkirakan pada masa yang akan datang yang dilihat dari pola/nilai pada masa lalu. Contoh dari tugas prediksi misalnya untuk memprediksikan adanya pengurangan jumlah pelanggan dalam beberapa waktu yang lalu dan prediksi harga saham dalam empat bulan ke depan.

Salah satu metode utama data mining adalah Supervised Learning ( Pembelajaran dengan Guru ), algoritma ini melakukan proses belajar berdasarkan nilai dari indikator target yang terhubung dengan indikator prediktor. Sedangkan metode analisis yang terdapat pada data mining salah satu nya adalah Klasifikasi. Klasifikasi sendiri memiliki definisi sebuah proses yang digunakan untuk menemukan model atau fungsi yang mendeskripsikan dan membedakan data ke dalam kelas-kelas yang bertujuan untuk menemukan pola yang lebih memiliki nilai dari data yang berukuran relatif besar hingga sangat besar. Klasifikasi melibatkan proses pemeriksaan karakteristik dari objek dan memasukkan objek ke dalam salah satu kelas yang sudah didefinisikan sebelumnya. Dalam klasifikasi hanya terdapat satu atribut dari sekian banyaknya atribut yang bisa menjadi kemungkinan yang disebut atribut target, sedangkan atribut lainnya yang terdapat disebut atribut prediktor. Tiap kemungkinan

nilai yang dimiliki oleh atribut target menunjukkan class yang diprediksi berdasarkan nilai-nilai dari atribut prediktor.

Algoritma C4.5 merupakan pengembangan dari algoritma ID3 yang termasuk salah satu algoritma yang masuk dalam bagian klasifikasi yang dapat digunakan untuk melakukan klasifikasi atau pengelompokan data yang bersifat prediktif untuk membentuk pohon keputusan (Decision Tree). Pohon keputusan merupakan metode klasifikasi dan prediksi yang cukup familiar di kalangan penggiat Data Mining. Pohon keputusan (Decision Tree) berguna untuk mengekspolari data, menemukan hubungan yang tidak saling terlihat antara sebuah calon variabel input dengan sebuah variabel target lainnya. Proses pada decision treeyaitu mengubah bentuk data dalam bentuk tabel menjadi model pohon(tree), mengubah model pohon menjadi rule yang dapat disederhanakan.

Saat mencari hasil dari alur sebuah decision tree langkah awal yang harus dilakukan adalah tentukan dahulu atribut mana yang akan menjadi simpul parents dan atribut mana yang akan menjadi simpul child. Pemilihan atribut yang baik adalah atribut yang memungkinkan untuk mendapatkan decision treeyang paling kecil ukurannya dan unik atau atribut yang bisa memisahkan objek menurut kelasnya masing-masing menjadi lebih detail. Untuk dapat menghasilkan sebuah pohon keputusan, maka nilai entropynya harus di ketahui dahulu hasilnya dengan menggunakan persamaan:

*Entropy (S) =* *pi \*log2 pi*

Keterangan:

S : himpunan kasus

A : atribut

N : jumlah partisi S

Pi : proporsi dari Si terhadap

Setelah langkah pertama selesai dan mendapat hasil himpunan kasusnya, maka selanjutnya bagaimana mencari nilai information gain. Information Gainmerupakan sebuah kriteria yang paling populer untuk pemilihan atribut yang ingin di cari. Information gain adalah salah satu attribute selection measure yang digunakan untuk memilih *test* attribute tiap node pada tree. Atribut dengan informasi gain tertinggi dipilih sebagai testatribut dari suatu node [Sunjana, 2010]. Dalam prosesnya perhitungan gain bisa terjadi atau

tidak suatu missing value. Information gain dapat dihitung dari output data atau variabel dependen *y* yang dikelompokkan berdasarkan atribut *A*, dinotasikan dengan gain (*y,A)*. Gain

*(y,A)* dari atribut *A* relatif terhadap output data *y* adalah:



Particle Swarm Optimization pertama kali dicetuskan oleh [kennedy dan Eberhart](http://www.cs.tufts.edu/comp/150GA/homeworks/hw3/_reading6%201995%20particle%20swarming.pdf" \t "_blank) pada tahun 1995 yang membahas mengenai perilaku kerumunan makhluk hidup. PSO terdiri dari sekumpulan partikel yang mencari posisi terbaik, yang merupakan posisi terbaik untuk masalah optimasi dalamr uang fitur. Penerapan PSO sebagai penentu parameter regulasi akan memberikan pengaruh pada accuracy klasifikasi yang telah ada [Widiasri,Justitia, & Arifin, 2011]. Persamaan untuk menghitung perpindahan posisi dan kecepatan partikel yaitu:

*Vi (t) = Vi (t-1) + c1 r1 [Xpbest i – Xi (t)] + c2 r2 [XGbest – Xi (t)] Xi (t) = Xi (t-1) + Vi (t)*

Keterangan :

Vi (t) : kecepatan partikel i saat iterasi t

Xi (t) : posisi partikel i saat iterasi t

c1 dan c2 : learning rate untuk kemampuan

individu (cognitive) dan pengaruh

hasil.

R1 dan r2 : bilangan random yang berdistribudi uniformal dalam interval 0 dan 1

Xpbest i : posisi terbaik partikel i

XGbest  : posisi terbaik global

Data yang akan digunakan pada penulisan ini adalah 150 dataset. Penulisan ini dilakukan atas kemajuan dari penelitian sebelumnya yang membandingkan 2 alagoritma yaitu C4.5 dan Naïve Bayes, setelah salah satu algoritma tersebut di dapatkan hasil accuracy tertingginya yaitu C4.5 maka hasil accuracy di optimasikan agar mendapatkan nilai accuracy yang jauh lebih baik. Hasil yang didapatkan dari algoritma C4.5 yang belum di optimasikan sebesar 76,21%, sedangkan menggunakan algoritma C4.5 yang sudah dioptimasikan menggunakan Particle Swarm Optimization sebesar 82,92% masih mungkin ditingkatkan lagi jika menggunakan algoritma optimasi lainnya,seperti WMA.

1. **Metode Penelitian**

Penelitian ini akan mengusulkan hasil accuracy baru untuk sebuah permasalahan memprediksi keputusan Calon Nasabah Potensial pada sebuah perusahaan asuransi Swasta ternama. Data yang digunakan dalam penelitian ini di ambil pada tahun 2018 bulan Januari-Desember, data yang di ambil kisaran 100.000 an namun untuk data testing untuk penelitian ini penulis hanya mengambil sampel data sebanyak 150. Pada tahap Klasifikasi akan dilakukan validasi silang (Cross Validation) dengan 10 kali validasi untuk mendapatkan hasil accuracy yang jauh lebih optimal dibandingkan dengan penelitian sebelumnya yang hanya menggunakan algoritma saja bukan algoritma yang optimasi.

Model desain ini akan melakukan pemrosesan data training dan data testing untuk menguji metode algoritma yang digunakan. Tahapan yang akan dilalui dibagi menjadi 3 bagian, yaitu preprocessing, seleksi fitur (Feature Selection) dan validation yang didalamnya berisi sub proses training dan testing.

**a. Pengumpulan Data**

Data yang digunakan pada penelitian kali ini merupakan data sekunder, karena sumber data diperoleh melalui media perantara atau secara tidak langsung yang berupa buku, catatan, bukti yang telah tervalidasi, atau arsip baik yang dipublikasikan maupun yang tidak dipublikasikan secara umum yaitu 100 ribu data yang akan dipecah menjadi 80 persen data training dan 20 persen data testing. Data testing yang digunakan sekitar 150 data yang telah diklasifikasikan berdasarkan variable yang ada.

Masalah yang harus dipecahkan pada penelitian kali ini adalah, bagaimana menghasilkan nilai accuracy yang jauh lebih optimal untuk mengklasifikasikan prediksi Calon Nasabah Potensial untuk ditawari produk lainnya. Dataset yang digunakan pada penelitian ini ada pada Gambar di bawah ini:



Gambar 1. Dataset

**b. Pengolahan Data**

Dataset ini dalam tahap preprocessing harus melalui 3 proses, yaitu:

Tahap ini akan melibatkan sub proses antara lain**:**

1. Membuang duplikasi data
2. Memeriksa data yang inkonsisten
3. Memperbaiki kesalahan pada data.

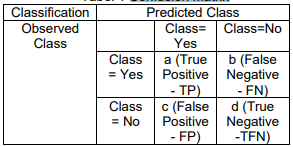
Dan Hasil akhir dari data ini berupa kumpulan data yang sudah bersih atau tidak ada missing valuenya.

**c. Evaluasi dan Validasi Data**

Validasi dilakukan menggunakan 10 fold cross validation. Dimana dengan menggunakan teknik ini dengan membagi secara acak ke dalam tiap bagian dimana terdiri dari 10 bagian untuk setiap bagian akan dilakukan proses klasifikasi terlebih dahulu, Sedangkan pengukuran akurasi diukur dengan confusion matrix dan kurva ROC (Receiver Operating Characteristics) untuk mengukur nilai AUC.

AUC digunakan untuk mengukur kinerja diskriminatif dengan memperkirakan probabilitas output yang sudah di dapatkan hasilnya dari sampel yang dipilih secara acak dari populasi positif atau negatif, semakin besar nilai AUC, semakin kuat klasifikasi yang dihasilkan. Karena AUC merupakan bagian dari daerah unit persegi, nilainya yang dihasilkan akan selalu sama yang dihasilkannya, antara 0,0 dan 1,0.

Tabel 1. Confusion Matrix



Sumber : (Gorunescy, 2011)

1. **Hasil dan Pembahasan**

Data yang digunakan dalam pengklasifikasiaan Calon Nasabah Potensial terdiri dari 100.000 dataset,150 data yang digunakan untuk data testing berdasarkan variabel yang tersedia. Dalam penentuan rekomendasi Calon Nasabah Potensial menggunakan atribut yang terdiri dari 5 atribut dan 1 atribut yang digunakan untuk class keterangan. Atribut yang digunakan yaitu Sex, Marital, Age, Wilayah, Occupation dan 1 atribut keterangan yaitu Keterangan.

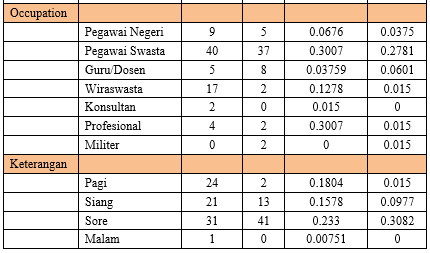
Klasifikasi dilakukan dengan menggunakan Software Rapidminer dengan

rule berbeda sesuai dengan model yang digunakan (Jungermann, 2009).

**3.1. Algoritma C4.5**

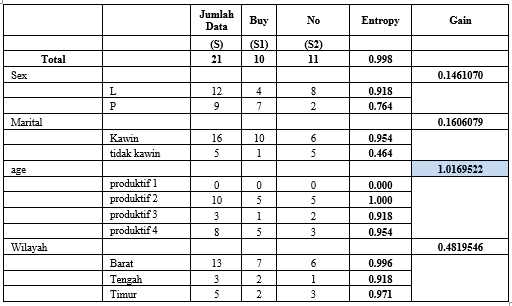
Algoritma C4.5 merupakan sebuah model yang dihasilkan dengan membentuk pohon keputusan. Untuk membuat pohon keputusan, ada beberapa langkah-langkah yang harus dilalui terlebih dahulu, yaitu:

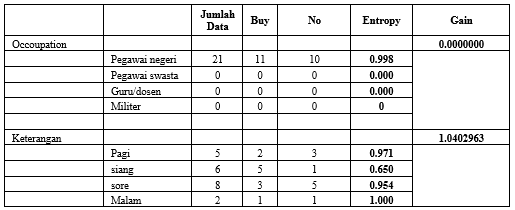
1. Tentukan akar dari pohon keputusan. Akar tersebut diambil dari atribut terpilih dengan cara menghitung nilai entropy masing-masing atribut dengan persamaan sebagai berikut:



*Entropy (S) =* *pi \*log2 pi*

Tabel 2. di bawah ini merupakan tabel jumlah kasus:



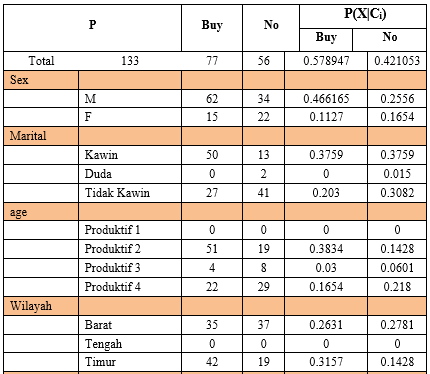


Tabel 2. Jumlah Kasus

2. Setelah mendapat hasil nilai entropy dari masing-masing atribut yang ingin diketahui, selanjutnya mencari nilai gain dengan persamaan terdapat pada gambar 2 sebagai berikut ini:



Tabel 3. berikut merupakan tabel nilai gain:



Tabel 3. Perhitungan Probabilitas Prior

3. Ulangi langkah 1 dan 2 untuk mengisi entropy dan nilai gain sampai semua atribut terisi

4. Proses partisi pohon keputusan akan berhenti saat:

a. Semua tupel dalam node N mendapat kelas yang sama

b. Tidak ada atribut di dalam tupel yang dipartisi lagi

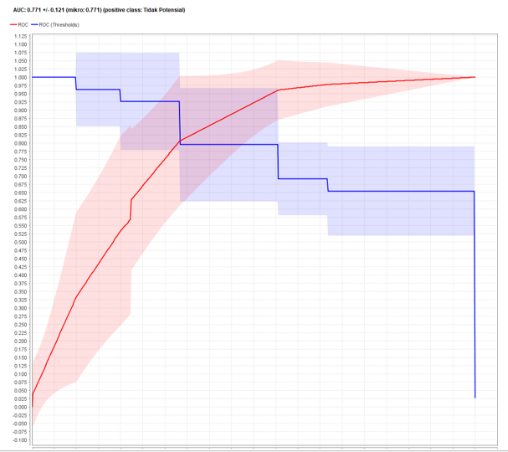
c. Tidak ada tupel di dalam cabang yang kosong

Gambar 2 merupakan perhitungan accuracy data *training* menggunakan algoritma C4.5 yang menghasilkan accuracy 76,21%. Diketahui data *training* terdiri dari 150 *record* data, 92 data diklasifikasikan POTENSIAL dan 59 data diprediksi TIDAK POTENSIAL.



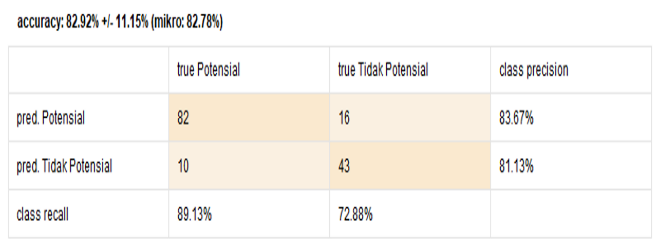
Gambar 2. Model *Confusion Matrix* algoritma C4.5

Sedangkan hasil yang di peroleh dari pengolahan kurva ROC dapat dilihat pada Gambar 3 menghasilkan nilai AUC sebesar 0,771 yang masuk pada kategori Fair Classification.



Gambar 3. Kurva ROC algoritma C4.5

Agar nilai accuracy data yang menggunakan algoritma C4.5 dapat meningkat, maka diperlukan algoritma optimasi agar dapat menghasilkan nilai accuracy yang lebih optimal, yang salah satunya dapat menggunakan algoritma PSO. Pemodelan menggunakan algoritma optimasi PSO terdapat pada gambar 3 berikut ini



Gambar 3. Model Confusion Matrix Algoritma C4.5 berbasis PSO

1. **Kesimpulan**

Kesimpulan yang diperoleh setelah melakukan eksperimen tersebut diperolehlah hasil dari accuracy yang lebih akurat pada prediksi Keputusan Calon Nasabah Potensial menggunakan algoritma C4.5 berbasis Particle Swarm Optimization memiliki nilai accuracy sebesar 82.92%, sedangkan prediksi Keputusan Calon Nasabah Potensial yang hanya menggunakan algoritma C4.5 saja menghasilkan nilai accuracy sebesar 76,51%. Dapat dilihat dari hasil eksperimen tersebut maka nilai accuracy yang dihasilkan dari kedua algoritma tersebut C4.5 berbasis Particle Swarm Optimization terjadi peningkatan yang cukup meningkat, peningkatkan accuracy dari kedua algoritma tersebut mencapai 6.41%.

**Referensi**

Abadi. (2016). “Penentuan Penerima Bantuan Dana untuk Sekolah Menengah Di Kab. Banjar Menggunakan Metode AHP- TOPSIS”. Jurnal Speed- Sentra Penelitian Engeneering dan Edukasi, Vol. 8 No. 1.

Darmanto. (2014). “Penerapan Metode

AHP (*Analytic Hierarchy Process ) untuk Menentukan Kualitas Gula Tumbuk* ”. Jurnal SIMETRIS, Vol.

5 No. 1, April 2014, ISSN :

2252-4983.

Firdaus,Diky (2017). “Penggunaan Data

Mining dalam Kegiatan Sistem Pembelajaran Berbantuan Komputer”.Jurnal FORMAT, Vol 6 Nomor 2, ISSN : 2089-5615

O’Brien. 2008. *Introduction to Information System, edition 12. (Pengantar Sistem Informasi Perspektif Bisnis dan Manajerial.* Jakarta: PT. Salemba Empat (Emban Patria).

Montgomery, D.C. 2009. Introduction to Statistical Quality Control. Mc Graw Hill.

Prayoga. (2017). Sistem Pendukung Keputusan Pemilihan Supplier Botol Galon menggunakan Metode AHP”. Jurnal Teknologi dana Manajemen Industri, Vol. 2

No. 2. Agustus 2017.

Saaty. 2018. The Analityc Hierarchy

Process. McGraw Hill International.

Saragih. (2013). “Penerapan Metode

AHP (*Analytic Hierarchy Process )* Pada Sistem Pendukung Keputusan Pemilihan Laptop ”. Pelita Informatika Budi darma, Vol. IV No. 2, Agustus 2013, ISSN : 2301-9425.

Sestri. (2013). “Penilaian Kinerja Dosen Dengan Menggunakan Metode AHP ( Studi kasus : STIE Ahmad Dahlan Jakarta)”.Jurnal Liquidity, Vol. 2 No. 1. Januari - juni 2013, Hal. 100-109.

Ratniasih, Ni Luh, 2019. Optimasi Data Mining Menggunakan Algoritma Naïve Bayes dan C4.5 Untuk Klasifikasi Kelulusan Mahasiswa. Jurnal Teknologi Informasi dan Komputer, Volume 5, Nomor 1, Januari 2019.

Sunjana. (2010). “Klasifikasi Data Nasabah Sebuah Asuransi menggunakan Algoritma C4.5”.Seminar Nasional Aplikasi Tekhnologi Informasi, 19-20 Juni 2019, ISSN: 1907-5022.

Widiasri, M., Justitia, A., & Arifin, A. Z. (2011).Penerapan Particle Swarm Optimization untuk Penentuan Parameter Regularisasi pada Kernel Regularized Discriminant Analysis. Industrial Electronics Seminar, 61-66, 26 Oktober 2011.