IMPLEMENTASI TEKNIK SELEKSI FITUR FORWARD SELECTION PADA ALGORITMA KLASIFIKASI DATA MINING UNTUK PREDIKSI MASA STUDI MAHASISWA POLITEKNIK INDONUSA SURAKARTA

Wiwit Supriyanti[1)], Norma Puspitasari[2)]

Politeknik Indonusa Surakarta
Jl. KH. Samanhudi No. 31 Mangkuyudan Surakarta
Email : [1]wiwitsupriyanti@poltekindonusa.ac.id , [2]normasari@poltekindonusa.ac.id

Abstrak

Berlimpahnya data mahasiswa dan data jumlah kelulusan mahasiswa, informasi yang tersembunyi dapat diketahui dengan cara melakukan pengolahan terhadap data mahasiswa sehingga berguna bagi pihak perguruan tinggi. Pengolahan data mahasiswa perlu dilakukan untuk mengetahui informasi penting berupa pengetahuan baru (*knowledge discovery*), misalnya informasi mengenai pengklasifikasian data mahasiswa berdasarkan profil dan data akademik. Pengetahuan baru tersebut dapat membantu pihak perguruan tinggi untuk melakukan klasifikasi mengenai tingkat kelulusan mahasiswa guna menentukan strategi untuk meningkatkan kelulusan pada tahun-tahun berikutnya.

Diketahui bahwa Politeknik Indonusa Surakarta belum memanfaatkan *database* tersebut dan dalam menentukan prediksi kelulusan masih menggunakan metode manual dengan tingkat subyektifitas yang tinggi. Algoritma klasifikasi data mining dapat diusulkan sebagai salah satu pendekatan yang dapat dilakukan untuk memprediksi masa studi mahasiswa berdasarkan data akademik mahasiswa yang tersedia. Implementasi seleksi fitur *forward selection* pada algoritma klasifikasi bertujuan untuk mencari atribut-atribut yang signifikan dalam prediksi masa studi serta menghilangkan atribut-atribut yang tidak signifikan, sehingga dapat meningkatkan akurasi hasil penghitungan.

Hasil penelitian didapat bahwa algoritma *k-nearest neighbor* menunjukkan nilai akurasi tertinggi dibandingkan algoritma klasifikasi data mining yang lain, yaitu sebesar 59,52% (tanpa tambahan seleksi fitur *forward selection*) dan 58,19% (menggunakan tambahan seleksi fitur *forward selection*).

Kata Kunci : data mining, algoritma klasifikasi, seleksi fitur, masa studi

1. PENDAHULUAN

Perguruan tinggi saat ini dituntut untuk memiliki keunggulan bersaing dengan memanfaatkan semua sumber daya yang tersedia. Bukan hanya sumber daya sarana, prasarana dan manusia, sistem informasi merupakan salah satu sumber daya yang dapat digunakan untuk memperoleh, mengolah dan menyebarkan informasi agar dapat menunjang operasional sehari-hari sekaligus menunjang kegiatan pengambilan keputusan strategis.

Politeknik dalam pendidikan di Indonesia merupakan salah satu bentuk perguruan tinggi selain akademi, institut, universitas, dan sekolah tinggi. Politeknik terdiri atas sejumlah program studi yang menyelenggarakan pendidikan vokasi pada sejumlah ilmu pengetahuan, teknologi, seni. Politeknik adalah suatu institusi pendidikan tinggi dan penelitian, yang memberikan gelar akademik dalam berbagai bidang. Politeknik didirikan untuk mengarahkan lulusannya menjadi tenaga profesional siap kerja. Pada umumnya program yang ditawarkan di salah satu Politeknik adalah program ahli madya dan sarjana terapan.

Berlimpahnya data mahasiswa dan data jumlah kelulusan mahasiswa, informasi yang tersembunyi dapat diketahui dengan cara melakukan pengolahan terhadap data mahasiswa sehingga berguna bagi pihak perguruan tinggi. Pengolahan data mahasiswa perlu dilakukan untuk mengetahui informasi penting berupa pengetahuan baru (*knowledge discovery*), misalnya informasi mengenai pengklasifikasian data mahasiswa berdasarkan profil dan data akademik. Pengetahuan baru tersebut dapat membantu pihak perguruan

tinggi untuk melakukan klasifikasi mengenai tingkat kelulusan mahasiswa guna menentukan strategi untuk meningkatkan kelulusan pada tahun-tahun berikutnya.

Betha Nurina Sari (2016) dalam penelitiannya yang berjudul "Implementasi Teknik Seleksi Fitur Information Gain Pada Algoritma Klasifikasi *Machine Learning* untuk Prediksi Performa Akademik Siswa" berpendapat bahwa masalah utama dalam dalam proses *discovering knowledge* dari data di bidang pendidikan adalah mengidentifikasi data yang representatif. Penelitian ini diterapkan pada beberapa algoritma klasifikasi *machine learning*, yaitu *Decision Tree*, *Random Forest*, *ANN*, *SVM*, dan *Naïve Bayes* agar bisa dilakukan komparasi performa dari hasil klasifikasi sebelum dan sesudah dilakukan seleksi fitur pada data akademik siswa. Adapun kesimpulan yang didapatkan dari penelitian adalah bahwa dengan implementasi teknik pemilihan fitur *information gain* dapat mempengaruhi tingkat akurasi algoritma klasifikasi *machine learning* (J48, *Random Forest*, *MLP*, *SVM (SMO)*, dan *Naïve Bayes*) untuk memprediksi performa akademik siswa pada mata pelajaran matematika.

Sukardi, Abd Syukur, dan Catur Supriyanto (2014) dalam penelitian mereka yang berjudul "Klasifikasi Spam Email Menggunakan Algoritma C4.5 dengan Seleksi Fitur" mengungkapkan *spam messages* membanjiri internet dengan mengirimkan salinan pesan-pesan yang sama untuk memaksa agar pesan-pesan tersebut sampai kepada pemakai yang tidak memilih untuk menerimanya. Akibatnya banyak pemakai yang merasa terganggu oleh banyaknya waktu yang dihabiskan untuk menghapus pesan *spam*, besarnya biaya yang harus dikeluarkan, dan besarnya *bandwidth* jaringan. Hasil eksperimen ditarik kesimpulan bahwa algoritma algoritma C4.5 dengan menggunakan tiga model kriteria yakni *gain ratio, information gain* dan *gini index*, hasil akurasi yang paling tinggi terdapat pada model kriteria *gini index* yakni 92,18%. Selanjutnya model kriteria gini index dilakukan seleksi fitur *chi square, information gain, information gain ratio* dan untuk meningkatkan hasil akurasi. Hasil yang paling tinggi dari ketiga seleksi fitur yakni *information gain ratio* dengan nilai p=0,6 dan hasil akurasinya menjadi 92,46. Serta memiliki nilai AUC rata-rata antara 0,9-1,0 dan ini termasuk klasifikasi sangat baik.

Rizal Amegia Saputra (2014) dalam penelitian "Komparasi Algoritma Klasifikasi *Data Mining* untuk Memprediksi Penyakit *Tuberculosis* (TB): Studi Kasus Puskesmas Karawang Sukabumi" membandingkan beberapa metode klasifikasi data mining, diantaranya yaitu Algoritma C4.5, *Naïve Bayes*, *Neural Network*, dan *Logistic Regression*, keempat metode tersebut digunakan dalam memprediksi diagnosis penyakit TB dengan tujuan agar algoritma terpilih merupakan algoritma yang paling akurat sehingga dapat melakukan diagnosa penyakit TB secara dini, ke empat metode tersebut merupakan sepuluh klasifikasi *data mining* paling popular. Hasil evaluasi dan validasi, diketahui bahwa *Naïve Bayes* memiliki nilai *accuracy* dan *AUC* paling tinggi diantara metode yang dikomprasikan, diikuti oleh algoritma C4.5, *neural network*, dan *logistic regression* memiliki akurasi yang paling rendah.

Pada penelitian ini penulis bermaksud melakukan implementasi teknik seleksi fitur *forward selection* pada algoritma klasifikasi data mining untuk mendapatkan hasil pengujian algoritma terbaik dalam mengolah informasi data akademik mahasiswa serta mengidentifikasi atribut-atribut dominan yang mempengaruhi masa studi mahasiswa khususnya di Politeknik Indonusa Surakarta.

2. TINJAUAN PUSTAKA

a. Data Mining

Data mining adalah ekstraksi informasi atau pola yang penting atau menarik dari data yang ada di *database* yang besar (Sudiyatno dan Susanto, 2014). Data mining adalah serangkaian proses untuk menggali nilai tambah berupa informasi yang selama ini tidak diketahui secara manual dari suatu basis data dengan melakukan penggalian pola-pola dari data dengan tujuan untuk memanipulasi data menjadi informasi yang lebih berharga yang diperoleh dengan cara mengekstraksi dan mengenali pola yang penting atau menarik dari data yang terdapat dalam basis data.

Alasan utama mengapa data mining sangat menarik perhatian industri informasi dalam beberapa tahun belakangan ini adalah karena tersedianya data dalam jumlah yang

besar dan semakin besarnya kebutuhan untuk mengubah data tersebut menjadi informasi dan pengetahuan yang berguna karena sesuai fokus bidang ilmu ini yaitu melakukan kegiatan mengekstraksi atau menambang pengetahuan dari data yang berukuran / berjumlah besar, informasi inilah yang nantinya sangat berguna untuk pengembangan.

b. Algoritma Klasifikasi

Sejumlah teknik atau algoritma yang digunakan untuk pemodelan pada klasifikasi antara lain adalah seperti berikut :

a. *Decision Tree Analysis* (Analisa Pohon Keputusan)
   Decision tree analysis (atau analisa pohon keputusan adalah suatu teknik yang termasuk keluarga *machine-learning*) bisa dibilang teknik klasifikasi yang paling popular pada area data mining.
b. *Statistical Analysis* (Analisa Statistik)
   Teknik-teknik statistik pada awalnya adalah algoritma klasifikasi yang populer selama bertahun-tahun sampai dengan kemunculan teknik-teknik '*machine-learning'*. Teknik-teknik klasifikasi statistik antara lain '*logistic regression*' (regresi logistik) dan *discriminant analysis* (analisa diskriminan), keduanya berasumsi bahwa hubungan antara variabel input dan output pada dasarnya adalah linear, data terdistribusi normal, dan variabel-variabel tidak saling terkait dan tidak tergantung satu sama lain. Sifat-sifat dasar asumsi yang diragukan ini akhirnya membawa pergeseran ke arah teknik-teknik '*machine-learning*'.
c. *Neural Networks* (Jaringan Syaraf Tiruan)
   Ini adalah salah satu diantara teknik-teknik dalam '*machine-learning*' yang paling popular yang bisa digunakan untuk problem-problem klasifikasi.
d. *Case-Based Reasoning* (Penalaran Berbasis Kasus)
   Pendekatan ini menggunakan kasus historis untuk mengenali berbagai kesamaan untuk menentukan suatu kasus baru ke dalam kategori yang paling mungkin.
e. *Bayesian Classifiers* (Klasifikasi Bayesian)
   Pendekatan ini menggunakan teori probabilitas untuk membuat model-model klasifikasi berdasarkan kejadian-kejadian di masa lalu yang bisa untuk menempatkan suatu instans baru ke dalam kelas (atau kategori) yang paling mungkin.
f. *Genetic Algorithms* (Algoritma Genetik)
   Penggunaan analogi terhadap evolusi alami untuk membuat mekanisme berbasis pencarian yang terarah untuk mengklasifikasikan sampel-sampel data.

c. Seleksi Fitur *Forward Selection*

Seleksi fitur adalah proses memilih fitur yang tepat untuk digunakan dalam proses klasifikasi atau *clustering*. Tujuan dari seleksi fitur ini adalah untuk mengurangi tingkat kompleksitas dari sebuah algoritma klasifikasi, meningkatkan akurasi dari algoritma klasifikasi tersebut, dan mampu mengetahui fitur-fitur yang paling berpengaruh terhadap tingkat akurasi.

Dalam metode *forward selection*, pemodelan dimulai dari nol peubah (*empty model*), kemudian satu persatu peubah dimasukan sampai kriteria tertentu dipenuhi. Langkah-langkah metode *forward selection* adalah sebagai berikut (Suyono, 2015):

Metode *forward selection* adalah pemodelan dimulai dari nol peubah (*empty* model), kemudian satu persatu peubah dimasukan sampai kriteria tertentu dipenuhi. Langkah-langkah metode *forward selection* adalah sebagai berikut (Draper dan Smith, 1992) :

a. Gunakan regresi linier sederhana
   $Y = \beta_0 + \beta_1 X + \varepsilon$ ……..…..(2.1)
   terhadap semua variabel independen (X_1, X_2,…, X_i) untuk mengetahui seberapa besar pengaruh dari setiap variabel independen. Uji hipotesis $H_0 : \beta_1 = 0$ lawan $H_0 : \beta_1 \neq 0$ dengan menggunakan uji *t*. Variabel independen yang memberi nilai *t* paling besar diambil sebagai X_1 asalkan H_0 ditolak. Jika H_0 diterima proses selesai.
b. Gunakan regresi linier dengan dua variabel independen
   $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \varepsilon$……....(2.2)
   dimana untuk variabel independen yang kedua diambil dari variabel independen yang tersisa. Uji hipotesis $H_0 : \beta_2 = 0$ lawan $H_0 : \beta_2 \neq 0$ dengan menggunakan uji *t*. Variabel yang memberi nilai *t* terbesar diambil sebagai X_2 asalkan H_0 ditolak. Jika H_0 diterima proses selesai.
c. Gunakan regresi linier dengan tiga variabel independen

$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \varepsilon$……...(2.3)

dimana untuk variabel independen yang ketiga diambil dari variabel independen yang tersisa. Uji hipotesis $H_0 : \beta_3 = 0$ lawan $H_0 : \beta_3 \neq 0$ dengan menggunakan uji *t*. Variabel yang memberi nilai *t* terbesar diambil sebagai X_3. Proses ini dilanjutkan sampai tidak ada lagi variabel independen yang hasil uji parameternya secara individual tidak signifikan.

d. RapidMiner

RapidMiner merupakan perangkat lunak yang bersifat terbuka (*open source*). RapidMiner adalah sebuah solusi untuk melakukan analisis terhadap data mining, text mining dan analisis prediksi. RapidMiner menggunakan berbagai teknik deskriptif dan prediksi dalam memberikan wawasan kepada pengguna sehingga dapat membuat keputusan yang paling baik. RapidMiner memiliki kurang lebih 500 operator data mining, termasuk operator untuk *input, output, data preprocessing* dan visualisasi. RapidMiner merupakan *software* yang berdiri sendiri untuk analisis data dan sebagai mesin data mining yang dapat diintegrasikan pada produknya sendiri. RapidMiner ditulis dengan menggunakan bahasa pemrograman java sehingga dapat bekerja di semua sistem operasi.

3. METODE PENELITIAN

Penelitian yang dilaksanakan adalah jenis penelitian eksperimen, yaitu melakukan pengujian tingkat akurasi terbaik diantara empat algoritma klasifikasi yang populer yaitu *decision tree, naive bayes, k-nearest neighbor* dan *support vector machine* dalam masa studi mahasiswa. Data eksperimen diambil dari data alumni Politeknik Indonusa Surakarta.

a. Metode Pengumpulan Data

1) Metode Observasi

Observasi merupakan salah satu teknik pengumpulan data yang tidak hanya mengukur sikap dari responden (wawancara dan angket) namun juga dapat digunakan untuk merekam berbagai fenomena yang terjadi (situasi, kondisi). Teknik ini digunakan bila penelitian ditujukan untuk mempelajari perilaku manusia, proses kerja, gejala-gejala alam dan dilakukan pada responden yang tidak terlalu besar. Dalam penelitian ini peneliti melakukan observasi pada data akademik mahasiswa di Politeknik Indonusa Surakarta.

2) Metode Wawancara

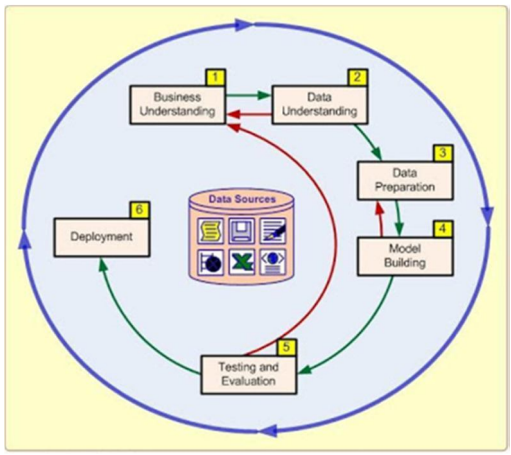
Wawancara merupakan teknik pengumpulan data yang dilakukan melalui tatap muka dan tanya jawab langsung antara pengumpul data maupun peneliti terhadap nara sumber atau sumber data. Dalam penelitian ini peneliti melakukan wawancara dengan pengelola Unit Teknologi Informasi (UTI) Politeknik Indonusa Surakarta.

3) Metode Studi Pustaka

Studi kepustakaan adalah teknik pengumpulan data dengan mengadakan studi penelaahan terhadap buku-buku, literatur-literatur, catatan-catatan, dan laporan-laporan yang ada hubungannya dengan masalah yang dipecahkan. Dalam penelitian ini peneliti melakukan studi pustaka dengan mengambil referensi buku-buku yang berkaitan dengan topik penelitian.

b. Metode Analisis Data

Penelitian ini didesain dengan menggunakan model CRISP-DM (*Cross Industy Standard Process for Data Mining*), dalam metode ini terdapat enam tahapan (Larose, 2005). Gambar 1 menjelaskan tentang siklus hidup pengembangan data mining yang telah ditetapkan dalam CRISP-DM.

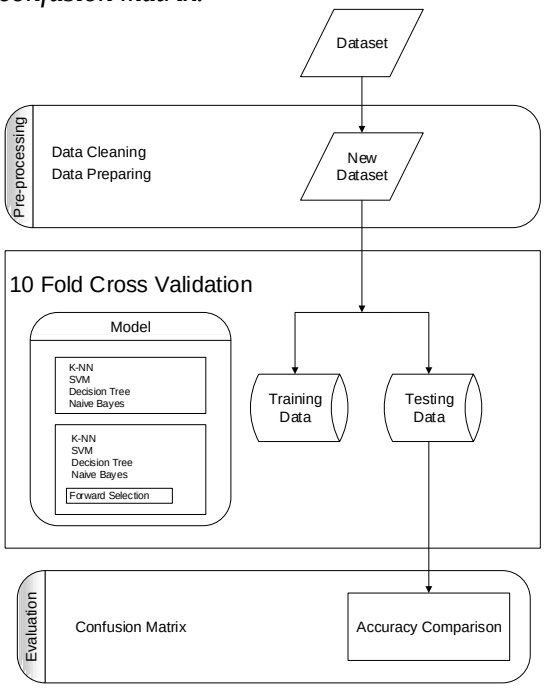


Gambar 1. Enam Tahap Proses CRISP-DM dalam Data Mining

c. Alur Penelitian

Secara umum alur penelitian yang dilakukan mengacu pada kerangka penelitian seperti pada Gambar 2.

Pada tahap pertama yaitu *pre-processing* dengan tahapan *data cleaning* dan *data preparing* pada dataset sampel sehingga diperoleh dataset baru. Tahapan selanjutnya pengujian dataset dengan teknik *10 fold cross validation* pada algoritma klasifikasi data mining serta implementasi teknik seleksi fitur *forward selection* pada masing-masing algoritma klasifikasi. Tahap terakhir adalah pengukuran tingkat akurasi masing-masing algoritma klasifikasi dengan membandingkan hasil evaluasi dari performa menggunakan *confusion matrix.*



Gambar 2. Alur Penelitian

4. HASIL DAN PEMBAHASAN

a. Tahapan *Pre-Processing* Data

Pada tahapan ini, penulis melakukan penyaringan data induk yang berasal dari data akademik Politeknik Indonusa Surakarta agar data tersebut layak untuk digunakan dalam proses penggalian informasi khususnya pada data alumni, hasil penyaringan data yang layak diperoleh 1045 data alumni yang berasal dari tahun 2006 sampai dengan tahun 2017, adapun atribut-atribut yang akhirnya dipilih 4 (empat) atribut dengan 3 (tiga) variabel terikat sebagai masukan (jenis kelamin, kelas, indeks prestasi kumulatif) dan satu variabel bebas sebagai keluaran (lama studi), adapun perinciannya sebagai berikut :

a. Jenis Kelamin
   Pada kategori jenis kelamin hanya ada dua tipe data yaitu ‘L’ untuk alumni dengan jenis kelamin laki-laki, sedangkan ‘P’ untuk alumni dengan jenis kelamin perempuan.
b. Kelas
   Kelas yang dimaksud dalam kategori ini adalah pada saat mendaftar sebagai mahasiswa baru kemudian mengikuti kegiatan perkuliahan sampai dengan lulus, alumni yang bersangkutan mengambil kelas reguler ataukah kelas karyawan.
c. Indeks Prestasi Kumulatif
   Kategori IPK pada penelitian ini adalah nilai akhir yang didapatkan oleh alumni dari semester pertama sampai dengan lulus dengan skala antara 0 (nol) sampai dengan 4 (empat).
d. Lama Studi
   Pada kategori lama studi terbagi menjadi dua kategori, yaitu ‘Tepat Waktu’ bagi alumni yang mampu menyelesaikan studinya maksimal 3 (tiga) tahun, serta ‘Terlambat’ apabila alumni tersebut menyelesaikan studinya lebih dari 3 (tiga) tahun.

NO	GENDER	KELAS	IPK	LAMA_STUDI
1	L	Reguler	3,09	Terlambat
2	L	Reguler	3,43	Terlambat
3	L	Reguler	2,84	Terlambat
4	L	Reguler	2,95	Tepat Waktu
5	L	Reguler	3,31	Tepat Waktu
6	L	Reguler	3,09	Tepat Waktu
7	L	Reguler	2,53	Terlambat
8	L	Reguler	2,83	Terlambat
9	L	Reguler	3,12	Tepat Waktu
10	L	Reguler	3,12	Tepat Waktu
11	P	Reguler	3,25	Tepat Waktu
12	L	Reguler	3,06	Terlambat
13	L	Reguler	3,58	Tepat Waktu
14	L	Reguler	3,5	Tepat Waktu

Gambar 3. Potongan Data Alumni

b. Pengujian Menggunakan RapidMiner

Tahapan berikutnya yaitu pengujian data alumni yang telah melalui tahapan *pre-processing* menggunakan beberapa algoritma klasifikasi data mining yang populer sehingga didapatkan nilai akurasi tertinggi dari salah satu algoritma yang dipilih.

Hasil pengujian data menggunakan *software* rapidminer didapatkan seperti pada Tabel 1 berikut:

Tabel 1. Hasil Pengujian Algoritma Menggunakan RapidMiner

No	Model Algoritma	Hasil Akurasi
1	Decision Tree	50,24%
2	Decision Tree + Forward Selection	50,24%
3	Naïve Bayes	57,42%
4	Naïve Bayes + Forward Selection	57,98%
5	k-Nearest Neighbor (k-NN)	59,52%
6	k-NN + Forward Selection	58,19%
7	Support Vector Machine (SVM)	56,65%
8	SVM + Forward Selection	56,65%

5. KESIMPULAN DAN SARAN

a. Kesimpulan

Berdasarkan hasil implementasi dapat ditarik beberapa kesimpulan sebagai berikut :

1) Pengujian terhadap algoritma klasifikasi data mining yang memiliki kinerja terbaik untuk menyelesaikan masalah prediksi masa studi mahasiswa di Politeknik Indonusa Surakarta adalah algoritma *k-Nearest Neighbor*.
2) Dengan penambahan teknik seleksi fitur *forward selection*, algoritma klasifikasi data mining yang memiliki tingkat akurasi terbaik untuk menyelesaikan masalah prediksi masa studi mahasiswa di Politeknik Indonusa Surakarta masih dipegang oleh algoritma *k-Nearest Neighbor* meskipun nilai akurasi yang dihasilkan mengalami penurunan.

b. Saran

Walaupun penelitian ini telah menghasilkan temuan awal, penulis masih harus mengembangkan analisis dan hasil lebih lanjut, khususnya memperdalam analisis agar hasil akurasi yang dihasilkan dapat meningkat sehingga informasi yang dihasilkan dari proses penggalian data alumni dapat dimanfaaatkan oleh instansi yang bersangkutan.

6. REFERENSI

Carlo Vercellis, 2011, *Business Intelligence: Data Mining and Optimization for Decision Making*, John Wiley & Sons, Inc Publication

Betha Nurina Sari, 2016, *Implementasi Teknik Seleksi Fitur Information Gain Pada Algoritma Klasifikasi Machine Learning untuk Prediksi Performa Akademik Siswa*, Seminar Nasional Teknologi Informasi dan Multimedia 2016, ISSN : 2302-3805

Larose T. Daniel, 2005, *Discovering Knowledge in Data: An Introduction to Data Mining*, John Wiley & Sons, Inc Publication

RapidMiner, 2016, RapidMiner Documentation, http://docs.rapidminer.com/

Rizal Amegia Saputra, 2014, *Komparasi Algoritma Klasifikasi Data Mining untuk Memprediksi Penyakit Tuberculosis (TB): Studi Kasus Puskesmas Karawang Sukabumi*, Seminar Nasional Inovasi dan Tren (SNIT) 2014

Sukardi, Abd Syukur, dan Catur Supriyanto, 2014, *Klasifikasi Spam Email Menggunakan Algoritma C4.5 dengan Seleksi Fitur*, Jurnal Teknologi Informasi, Volume 10 Nomor 1, April 2014, ISSN 1414-9999

Susanto, H., Sudiyatno, 2014, *Data Mining untuk Memprediksi Prestasi Siswa Berdasarkan Sosial Ekonomi, Motivasi, Kedisiplinan dan Prestasi Masa Lalu*, Jurnal Pendidikan Vokasi, Vol. 4 No. 2 Juni 2014

Suyono, 2015, Analisis Regresi untuk Penelitian, Penerbit Deepublish, Yogyakarta