SIGMA
JURNAL TEKNOLOGI PELITA BANGSA
Vol. 12 No. 3 September 2021 p-ISSN: 2407-3903 e-ISSN: 2407-3903
Diterima, 13 Juli 2021 | Direvisi, 11 Agustus 2021 | Dipublikasikan, 28 Septemer 2021

PENERAPAN DATA MINING UNTUK MEMPREDIKSI MINAT SISWA YANG MENDAFTAR DI SMK AL AMIN CIBARUSAH

Ermanto

Program Studi Teknik Informatika Fakultas Teknik Universitas Pelita Bangsa
ermanto@pelitabangsa.ac.id

Abstraksi

Untuk menentukan tingkat minat siswa yang akan mendaftar di Sekolah Menengah Kejuruan (SMK) Al Amin Cibarusah bisa diperkirakan dengan mengunakan metode data mining, untuk mencari suatu informasi didalam data yang sangat besar diperlukan metode algoritma yang akurat, beberapa metode yang sangat akurat dalam prediksi dan klasifikasi yang sering digunakan adalah metode Naive Bayes, Nearest neighbour dan Decision tree atau J48. Fokus penulis dalam penelitian ini adalah menganasilsa penerapan algoritma Naive Bayes, Nearest neighbour dan algoritma J48 dalam memprediksi minat siswa yang ingin mendaftar di Sekolah Menengah Kejuruan (SMK) Al Amin Cibarusah. Dan hasil yang diperoleh dapat diterapkan dalam pembuatan keputusan dan kebijakan sekolah.

Kata Kunci : Data mining, *Naive bayes*, *Nearest neighbour*, *Decision tree* dan Algoritma *J48*.

Abstract

The level of interest of students who will enroll in Vocational High School (SMK) Al Amin Cibarusah can be estimated by using data mining methods. In order to find some information in a very large data, accurate algorithmic method is required. Several methods that are found very accurate in predicting and classifying and are often used are the Naive Bayes method, Nearest neighbor and Decision tree or J48. The author's focus in this study is to analyze the application of the Naive Bayes algorithm, Nearest neighbor and the J48 algorithm in predicting the interest of students who want to enroll in the Al Amin Cibarusah Vocational High School (SMK). The results obtained can be applied in making certain decisions and school policies.

Keywords: Data mining, Naive bayes, Nearest neighbor, Decision tree and J48 Algorithm.

1. Pendahuluan

Sekolah Kejuruan adalah pendidikan khusus yang mempunyai tujuan membentuk siswa yang siap kerja. Sekolah kejuruan menitikberatkan pada kemampuan ketrampilan siswa pada bidang-bidang tertentu seperti teknik kendaraan ringan, teknik komputer, tata boga, tata busana, pertanian dan lain-lain. Himbauan dari pemerintah tentang pentingnya sekolah Kejuruan membuat animo masyarakat begitu besar untuk menyekolahkan anaknya pada pendidikan fokasi tersebut. Dengan slogan "SMK BISA" pemerintah berhasil menarik hati masyarakat untuk menyekolahkan anaknya di Sekolah Menengah Kejuruan. Meningkatknya minat dari masyarakat untuk masuk sekolah di pendidikan kejuruan tentu menjadi perhatian bagi penyelenggara pendidikan, dimana sekolah dituntut untuk bisa menyelenggarakan pendidikan kejuruan sesuai dengan standar kompetensinya, baik dari sarana dan prasarana yang memadai juga dari sumber daya yang kompeten guna menarik minat siswa untuk masuk sekolah tersebut.

Sekolah Menengah Kejuruan (SMK) Al Amin Cibarusah merupakan salah satu SMK swasta yang terus menerus melakukan perbaikan dengan meningkatkan kulaitas pembelajaran, perekrutan tenagga pendidik yang sesuai kompetensinya menjadi salah satu cara untuk meningkatkan mutu pembelajaran dan juga peningkatan sarana prasarana sekolah seperti membangun Ruang kelas baru dan juga pengadaan alat alat praktek. Dengan meningkatkan kualitas baik SDM dan juga sarana prasarana sekolah itu saja belum

cukup, hal ini bisa kita lihat dari jumlah siswa yang mendaftar tetapi tidak mendaftar ulang, sehingga siswa tersebut tidak masuk terdaftar di SMK Al Amin Cibarusah, banyak faktor yang mempengaruhi hal tersebut sehingga sulit untuk bisa memastikan penyebabnya, SMK Al Amin Cibarusah berdiri sekitar tahun 2006 yang beralamat di Jl Raya Cibarusah – Jonggol Km. 1,3 Kp. Cibedug Desa Cibarusah Jaya, Kecamatan Cibarusah, Kabupaten Bekasi Provinsi Jawa Barat.

Semakin bertambahnya jumlah sekolah kejuruan di sekitar Kecamatan Cibarusah menjadi permasalahan tersendiri bagi setiap sekolah, dimana setiap tahun mereka berlomba-lomba mempromosikan sekolahnya agar bisa mendapatkan siswa sesuai dengan kuota yang tersedia, berbagai cara dilakukan mulai dari pembagian brosur, pamlet, pemasangan spanduk, iklan di media sosial dan juga presentasi ke sekolah-sekolah SMP. Banyaknya siswa yang mendaftar di SMK Al Amin Cibarusah tetapi tidak mendaftar ulang sebagai syarat masuk SMK Al Amin Cibarusah menjadi sebuah pertanyaan, apakah penyebab siswa tidak jadi mendaftar di SMK Al Amin Cibarusah.

Dari data yang di peroleh mulai dari tahun 2012 – 2017 terdapat 1150 siswa yang mendaftar tetapi hanya sekitar 850 siswa yang mendaftar ulang hal ini menjadi permasalahan yang mendasar bagi peneliti untuk melakukan penelitian. Data yang diperoleh tersebut kemudian akan diolah untuk mengetahui suatu pola agar kita bisa mengambil informasi yang berguna dari data yang besar tersebut. Pengolahan data seperti ini bisa dilakukan dengan menggunakan metode data mining.

Pada penelitian ini analisis yang digunakan adalah menggunakan metode algoritma Naive Bayes, Nearest neighbour dan J48. Penggunaan algoritma Naive bayes adalah untuk bisa mengitung probiltas dan statistik agar bisa memprediksi peluang yang akan datang berdasarkan data dimasa lalu dan algoritma J48 dimaksudkan untuk membuat pohon keputusan.

2. Landasan Pemikiran

2.1.1. Penerimaan Siswa Baru (PSB)

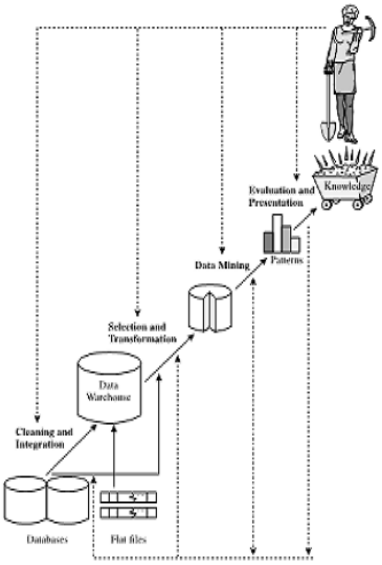
Penerimaan siswa baru adalah langkah awal atau tahap yang paling penting dalam penyelenggaraan pendidikan, dimana dalam tahapan ini adalah prosesa awal penyaringan siswa yang akan belajar di suatu sekolah, suksesnya sebuah sekolah itu bisa dipengaruhi oleh proses penerimaan siswa baru, maksudnya adalah jika kita salah langkah dalam menerima siswa baru itu bisa menyebabkan masalah dan hambatan dalam proses pendidikan.

Menurut Herek (1982:9) pada dasarnya Pendaftaran adalah untuk memperlancar dan mempermudah proses pendataan dan pembagian kelas siswa atau siswi baru, sehingga bisa terorganisir, teratur dengan cepat dan tepat dengan beberapa persyaratan yang telah ditentukan oleh sekolah. Proses pendaftaran siswa baru adalah salah satu kewajiban pihak sekolah dari Dinas Pendidikan setiap tahun ajaran baru.

2.1.2. Data Mining

Menurut Han dan Kamber (2011). Data mining merupakan suatu proses mendapatkan pola yang menarik dan pengetahuan dari data yang berjumlah besar.(1) Menurutnya secara umum data mining dapat dikelompokan menjadi dua kategori utama yaitu predictive dan descriptive. Predictive adalah proses untuk mendapatkan pola dari data dengan menggunakan beberapa variabel. Teknik yang terdapat dalam predictive mining salah satunya adalah klasifikasi, klasifikasi adalah suatu teknik pembentukan model dari data yang belum terklasifikasi untuk digunakan mengklasifikasikan data baru sedangkan descriptive pada data mining adalah suatu proses agar menemukan sesuatu yang penting dari data dalam suatu database. Tujuan dan tugas deskriptif adalah untuk menemukan pola-pola yang meringkas hubungan yang pokok dalam data.

Menurut Turban et al, (2007) Data mining adalah suatu istilah yang digunakan untuk menggambarkan penemuan ilmu pengetahuan dalam bidang database, sebuah bidang analisis informasi yang mencari pola tersembunyi dalam kelompok data yang dapat digunakan untuk meprediksi prilaku masa depan. Adapun fungsi data mining menurut Maclennan, Tang & Crivat (2009) fungsi data mining adalah *classification, clustering, association, regrssion, forecasting dan sequence analysis.*



Gambar 1. Tahapan dalam *Data Mining*

2.1.3. *Naive bayes*

Algoritma naive bayes adalah bagian dari algoritma yang terdapat pada teknik klasifikasi. Sebagimana dikatakan oleh penememunya yaitu Thomas Bayes, adalah memprediksi peluang dimasa depan berdasarkan pengalaman dimasa lalu yang dikenal dengan Teorema *Bayes*, dimana teorema

tersebut dikombinasikan dengan naive yang diasumsikan kodisi antar atribut saling bebas. Teorema bayes memiliki kecepatan yang baik dan akurasi yang tinggi ketika diterapkan pada database yang besar. Karena naive bayes termasuk ke dalam pembelajaran supervised maka pada tahapan pembelajaran diperlukan data awal yaitu berupa data pelatihan untuk bisa mengambil keputusan, selanjutnya pada tahap pengklasifikasian akan dihitung nilai probilitas dari setiap label kelas yang ada terhadap input yang yang diberikan. Adapun persamaan dari *Naive Bayes* adalah sebagai berikut :

$$P(Ci|X) = \frac{P(X|Ci)P(Ci)}{P(xi)}$$

Keterangan :
X : Kriteria suatu kasus berdasarkan masukan
Ci : Kelas solusi pola ke –i, dimana i adalah jumlah label kelas
P(Ci|X) : Probilitas kemunculan label kelas Ci dengan kreteria masukan X
P(X|Ci) : Probiltas kriteria masukan X dengan lebel kelas Ci
P(Ci) : Probilitas label kelas Ci

2.1.4. Nearest neighbour

Nearest Neighbour adalah algoritma pengklasifikasian yang didasarkan pada analogi, yaitu membandingkan data uji dengan data pelatihan yang berada dekat dengan dan memiliki kemiripan dengan data uji tersebut. Kemiripan data uji dengan data pelatihan didasarkan pada jaraknya. Banyak persamaan yang dapat digunakan untuk menghitung jarak antara data uji dan data pelatihan. Tiga diantaranya yang paling sering digunakan adalah:

1. Atribut yang bertipe numeric

Terdapat dua pendekatan perhitungan jarak/kemiripan yang umum digunakan untuk atribut yang bertipe numerik, yaitu *euclidean distance* dengan persamaan berikut:

$$Dist(x1, x2) = \sqrt{\sum_{i=1}^{n} (x_{1i} - x_{2i})} \quad (2)$$

Keterangan:
n : jumlah data
x1 : data uji
x2 : data pembelajaran

Persamaan yang kedua yaitu *Manhattan distance* sebagai berikut :

$$Dist(p_i(an), p_i(nc)) = \frac{p_i(an) - p_i(nc)}{\max\_dist\_i} \quad (3)$$

Keterangan:
pi : atribut ke-i
an : data pembelajaran
nc : data uji

2. Atribut yang bertipe simbolik

Persamaan yang digunakan untuk atribut yang menggunakan istilah eksplisit yaitu ada atau tidak ada, memiliki atau tidak memiliki, ya atau tidak dan sebagainya maka perhitungan kemiripan atau jarak dapat dihitung dengan fungsi sebagai berikut :

$$Sim(Ki(a), Ki(b)) = \begin{cases} 0 & Ki(a) \neq Ki(b) \\ 1 & Ki(a) = Ki(b) \end{cases} \quad (4)$$

Keterangan :
Ki(a) : kriteria ke-i dari kasus a
Ki(b) : kriteria ke-i dari kasus b
Sim(Ki(a),Ki(b) : nilai kemiripan kriteria ke-i antara kasus a dengan kasus b

Perhitungan selanjutnya adalah persamaan untuk mencari kemiripan dengan *nearest neighbour* yaitu :

$$Similarity(T, S) = \frac{\sum_{i=1}^{n} Sim(K_i(T), K_i(S)) x w_i}{\sum_{i=i}^{n} w} \quad (5)$$

Keterangan:
T : data uji
S : data pembelajaran
n : jumlah kriteria
w : bobot kriteria
Sim(*K* (*T*), *K* (*S*)) *i i* : Nilai kemiripan/jarak kriteria kasus target dan target sumber

2.1.5. Decision Tree (Algotirma J48)

Menurut Han dan Kamber, (2001) Metode *decission tree* merupakan suatu struktur *flowchart* yang sama seperti struktur pohon, dimana setiap titik pohon adalah atribut yang telah diuji, setiap cabang adalah hasil uji, dan titik akhir merupakan pembagian kelas yang dihasilkan.

Decission tree adalah metode klasifikasi dan prediksi yang sangat terkenal. Metode pohon keputusan mengubah fakta yang sangat besar menjadi pohon keputusan yang merepresentasikan aturan. Aturan dapat dengan mudah dipahami dengan bahasa alami dan mereka juga dapat diekspresikan dalam bentuk bahasa basis data seperti *Structured Query Language* untuk mencari *record* pada kategori tertentu (kusrini dan emha taufiq luthfi, 2009).

Pemilihan atribut sebagai simpul, baik akar *(root)* atau simpul internal didasarkan pada nilai *Gain* tertinggi dari atribut-atribut yang ada. Perhitungan nilai *Gain* digunakan rumus seperti dalam Persamaan 1.

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

di mana:
S : Himpunan kasus
A : Atribut
n : Jumlah partisi atribut *A*
$|S_i|$: Jumlah kasus pada partisi ke-*i*
$|S|$: Jumlah kasus dalam *S*

Konsep entropi digunakan untuk penentuan pada atribut mana sebuah pohon akan terbagi (split). Semakin tinggi entropy sebuah sampel, semakin tidak murni sampel tersebut. Untuk menghitung nilai *Entropy* dapat dilihat pada Persamaan 2.

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

Di mana:
S : Himpunan kasus
n : Jumlah partisi *S*
pi : Proporsi dari S_i terhadap *S*

Algoritma J48 adalah implementasi dari algoritma C.4.5 yang memproduksi decision tree menggunakan tool weka.. Algoritma J48 dapat mengklasifikasikan data dengan metode pohon keputusan yang memiliki keunggulan bisa memproses data numerik dan diskret, bisa mengatasi nilai atribut yang hilang, menghasilkan rule-rule yang mudah interprestasikan dan cepat. Karena kelebihan inilah diharapkan algoritma J48 mampu menangani studikasus dengan optimal sehingga menghasilkan akurasi dan performan yang baik.

2.1.6. Akurasi

Akurasi diperlukan untuk evaluasi dan mengukur keakuratan dari hasil klasifikasi, semakain besar nilai akurasi maka semakin baik tingkat klasifikasinya:

$$Accuraccy = \left(\frac{jumlah\ dokumen\ yang\ terklasifikasi}{jumlah\ dokumen\ keseluruhan} X\ 100\%\right)$$

3. Metodologi Penelitian

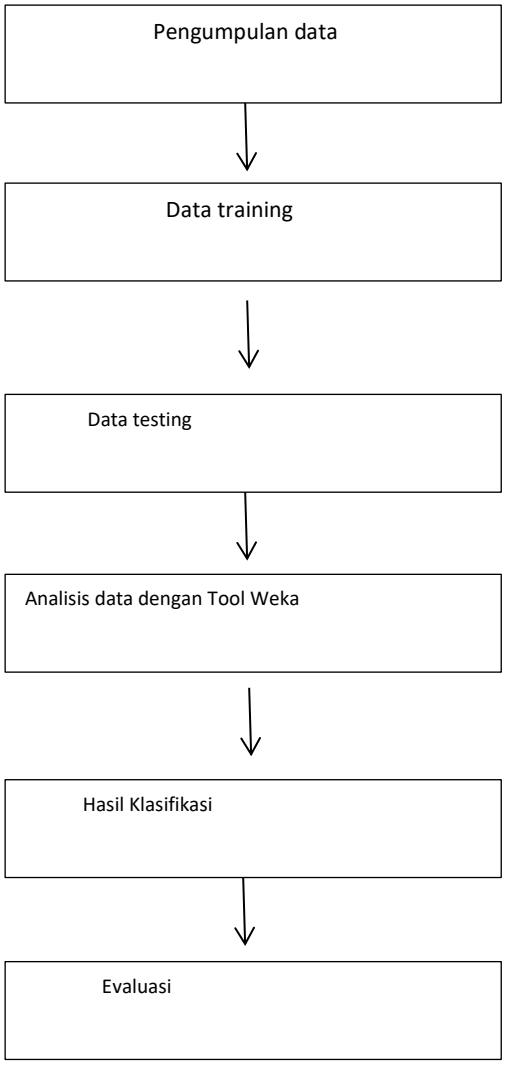
Dalam Bab ini akan penulis akan membahas informasi tentang metode penelitian yang dilakukan dalam tesis ini. Pertama dari objek penelitian, desain penelitian dan teknik pengumpulan data yang dilakukan dari berbagai sumber penelitian, kemudian selanjutnya metode yang diusulkan yaitu melakukan training dan tes data serta penerapan algoritma. Langkah selanjutnya yang akan dilakukan adalah melakukan eksperimen/pengujian terhadap data berupa data siswa baru menggunakan alat bantu sofware Weka.

Penelitian ini memakai pendekatan kualitatif, adapun tempat penelitiannya adalah di SMK Al Amin Cibarusah Kabupaten Bekasi. Target penelitian adalah siswa yang mendaftar di SMK Al Amin Cibarusah Tahun 2012-2017 sebanyak 1150 siswa.

3.1. Disain Penelitian

Disain penelitian melalui tahapan-tahapan sebagai berikut :

1. Mengumpulkan data
2. Pengolahan data
3. Model yang diusulkan
4. Hasil Prediksi
5. Evaluasi



Gambar 2. Kerangka berfikir penelitian

4. Pembahasan

Bab ini membahas tentang hasil penelitian yang dilakukan mulai dari dataset yang digunakan dan hasil akurasi dari klasifikasi yang diperoleh. Dalam penelitian ini dilakukan percobaan terhadap dataset siswa SMK al amin dengan algoritma *decision tree C4.5*, algoritma naïve bayes dan algoritma K-NN.

4.1. Hasil Evaluasi Algoritman Naïve Bayes

Tabel 1. Tabel persentase data training dan testing dengan algortima naïve bayes

Nilai	Training dan testing naïve bayes									
	Training					Testing				
Akurasi	16a	13a	11a	9a	7a	16a	13a	11a	9a	7a
persentase	68	68	69	69	68	68	67.75%	68.625%	68.125%	68.8571%

Dari tabel 1 dihasilkan data training dengan akurasi tertinggi pada 9 atribut dengan akurasi 69.125%, sedangkan untuk data testing nilai kaurasi tertinggi pada 7 atribut dengan akurasi 68.857%.

4.2. Hasil Evaluasi Algoritma J.48

Tabel 2 Tabel persentase data training dan testing dengan algortima J.48

Nilai	Training dan testing J.48									
	Training					Testing				
Akurasi	16a	13a	11a	9a	7a	16a	13a	11a	9a	7a
persentase	75.25 %	74.125%	73.12 %	74.5 %	73.5%	66 %	68.8571 %	68.2857 %	67.7143%	69.428%

Dari tabel 2 dihasilkan data training dengan akurasi tertinggi pada 16atribut dengan akurasi 75.25%, sedangkan untuk data testing nilai kaurasi tertinggi pada 7 atribut dengan akurasi 69.428%.

4.3. Hasil Evaluasi Algoritma KNN

Tabel 3 Tabel persentase data training dan testing dengan algortima KNN

Nilai	Training dan testing KNN									
	Training					Testing				
Akurasi	16a	13a	11a	9a	7a	16a	13a	11a	9a	7a
persentase	92.125 %	88.875 %	86.75 %	83.875 %	80.125 %	61.7143 %	65.1429 %	68.2857 %	65.4286 %	64.571 %

Dari tabel 3 dihasilkan data training dengan akurasi tertinggi pada 16 atribut dengan akurasi 92.125%, sedangkan untuk data testing nilai kaurasi tertinggi pada 11 atribut dengan akurasi 68.2857 %.

Dari tabel ketiga table diatas berdasarkan hasil trining dan testing dari ketiga algoritma yaitu naïve bayes, J.48 dan KNN, nilai akurasi klasifikasi hasil akhir lebih baik menggunakan algoritma KNN klasifkasi yaitu dengan data training 83.875% dan data testing 65.428%.

Tabel 4 tabel perbandingan algoritma klasifikasi dengan akurasi terbaik

Data	Naïve bayes	J.48	KNN
Training	69.12%	73.5%	83.875%
Testing	68.5%	69.428%	65.428%
	9 atribut	7 atribut	9 atribut

5. Penutup

Berdasarkan hasil pengujian yang dilakukan oleh peneliti maka dapat disimpulkan sebagai berikut:

1. Dalam penelitian ini pendekatan dengan 7 Atribut pada hasil akurasi training KNN dengan nilai akurasi 83.875% lebih baik dari J.48 dan naïve bayes dalam penentuan siswa yang mendaftar psb(penerimaan siswa baru).
2. Dalam penelitian ini hasil testing nilai akurasi terbaik pada algotima J.48 dengan percobaan 7 atribut dengan nilai akurasi 69.428%.

Daftar Pustaka

[1] Han, J and Kamber, M. 2012. *Data Mining Concept and Techniques Third Edition.*Morgan Kauffman. San Francisco.

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

[3] Mufizar, T., Anwar, D. S., & Aprianis, E. (2016). Sistem Pendukung Keputusan Pemilihan Jurusan Dengan Menggunakan Metode SAW Di SMA 6 Tasikmalaya. *Voice Of Informatics*, *5*(1).

[4] Andrian, Y., & Wayahdi, M. R. Analisis Kinerja Data Mining Algoritma C4. 5 Dalam Menentukan Tingkat Minat Siswa yang Mendaftar di Kampus ABC.

[5] Naparin, H. (2017). Klasifikasi Peminatan Siswa SMA Menggunakan Metode Naive Bayes. Systemic: Information System And Informatics Journal, 2(1), 25-32.

[6] Kusumadewi, Sri Hartati, Sri Harjoko, Agus dan Wardoyo, Retantyo. (2006) *Fuzzy Multi Attributte Decission Making* (Fuzzy MADM). Graha Ilmu. Yogyakarta .

[7] Turban E., Aronson J. E., Liang T. P., dan Sharda R. (2007). *Dicision Suport and busimess intelligencesystem (Eighth ed.).* Pearson Education.

[8] MacLennan, J., Tang, Z., & Crivat, B. (2008). Data Mining with Microsoft SQL Server. Copyright© 2009 by Wiley Publishing. Inc., Indianapolis, Indiana, 371-387.