

PENERAPAN ALGORITMA NAÏVE BAYES UNTUK PREDIKSI PENERIMAAN SISWA BARU DI SMK AL-AMIEN WONOREJO

¹Saiful Rizal, ²Moch. lutfi

^{1,2}Program Studi Teknik Informatika Universitas Yudharta

¹saifulriz@gmail.com, ²luthfi@yudharta.ac.id

Received : April 2018

Accepted : Mei 2018

Published : Juni 2018

ABSTRACT

One aspect of the first high school quality indicator is the level of acceptance of students in senior high schools or public vocational schools. Some junior high school students' data were analyzed to determine the level of acceptance of students in senior high schools or state vocational high schools. The process of analyzing student data using data mining techniques. The purpose of this research is to know the application of naïve bayes classification on the average value of report cards and the value of national examination on the acceptance level of students in high school or public vocational school using the result of classification model that formed.

In this study the data used is the new student data force 2016 Vocational High School AL-AMIEN Wonorejo. Data mining process is assisted by WEKA software using naïve bayes classification and 10-fold cross validation. Furthermore the naïve bayes classification model is used to process prediction data.

The results of the testing of the Naïve Bayes algorithm in predicting the acceptance of new students at the AL-AMIEN Wonorejo Vocational High School in 196 data of the students tested in this study, showed that the Naïve Bayes algorithm has an accuracy rate of 86.22%.

Keywords: Data mining, classification, naïve bayes, software WEKA

I PENDAHULUAN

1.1. Latar Belakang

Kualitas pendidikan pada siswa merupakan salah satu tujuan dari sekolah. Sekolah akan meningkatkan kualitasnya dari tahun ke tahun sebagai tolak ukur dalam menentukan keberhasilan system pendidikannya. Salah satu aspek sebagai indikator kualitas di sekolah adalah tingkat kelulusan siswa dan banyaknya lulusan siswa sekolah menengah pertama yang diterima dalam memotivasi siswa untuk memberikan jaminan kesempatan kerja yang lebih baik di masa mendatang.

Setiap tahun, jumlah data siswa di sekolah bertambah. Data tersebut tersimpan dalam bentuk hard file misalnya hasil print out data, catatan buku dan dalam bentuk soft file pada computer. Hal tersebut mengakibatkan menumpuknya data yang disimpan oleh sekolah. Keadaan seperti ini digambarkan sebagai sebuah situasi "*data rich but information poor*" yang artinya data yang disimpan berlimpah namun informasi yang diperoleh kurang. Bertambahnya jumlah data yang cepat dan besar yang dikumpulkan dan disimpan setiap tahunnya jauh melampaui kemampuan manusia dalam menganalisis data tanpa suatu teknik yang tepat. Sebagai akibatnya, data yang dikumpulkan menjadi sesuatu yang tidak bermanfaat. Untuk mengatasi hal tersebut diperlukan suatu teknik dalam *data mining* yang dapat mengubah sesuatu yang tidak bermanfaat menjadi sesuatu yang berharga dan dapat memberikan suatu informasi yang penting.

Beberapa penelitian terkait antara lain : Muchamad Hasan Bisri pada tahun 2015. Penelitian ini membahas tentang implementasi algoritma naïve bayes untuk memprediksi penjurusan siswa di SMA Kesatrian 1 Semarang. Pada penelitian ini dibuat suatu aplikasi system pengambilan keputusan untuk memprediksi jurusan siswa di SMA kesatrian 1 Semarang. Fungsi untuk memprediksi dibuat dengan menggunakan Matlab yang memberikan akurasi 83,8798% dengan tingkat kesalahan 16,1202%. [1]. Castaka Agus Sugianto pada tahun 2015. Penelitian ini membahas penerapan teknik data mining untuk menentukan hasil seleksi masuk SMAN 1 Gibeber untuk siswa baru menggunakan decision tree. Uji yang dilakukan menggunakan 3 (tiga) jenis algoritma : pengujian dengan menggunakan Algoritma C4.5 (Decision Tree), adalah Accuracy 99.05%, Precisian 96,33%, pengujian dengan Algoritma Naïve Bayes, Accuracy 90.62%, Precision unknown, sedangkan pengujian dengan Algoritma Neural Network, Accuracy 95.02%, Precision

73.10% [2]. Asrul Ashari Muin pada tahun 2016. Penelitian ini menggunakan metode algoritma naïve bayes untuk melakukan prediksi peluang kelulusan mahasiswa baru. Nilai presentasi akurasi menunjukkan dataset efektivitas penerimaan diterapkan pada metode klasifikasi naïve bayes yang mencapai 94%. [3]

Pada penelitian ini dibuat suatu aplikasi sistem pengambilan keputusan untuk memprediksi penerimaan siswa baru di SMK, dalam hal ini SMK yang akan menjadi objek penelitian adalah SMK AL-AMIEN WONOREJO yang terletak di Jalan Ponpes Riyadhul Banin Wal Banat, Jatiguting, Wonorejo, Pasuruan. Berdasarkan beberapa hal diatas dilakukannya penelitian sebuah aplikasi yang diharapkan dapat menampilkan hasil dari data mining dengan keputusan untuk membantu dalam memprediksi penerimaan siswa baru di SMK AL-AMIEN Wonorejo, Pasuruan.

II TINJAUAN PUSTAKA

2.1. Landasan Teori

2.2.1 Database

Sebuah system *database* atau yang disebut *Database Management system (DBMMS)* adalah sekumpulan data yang saling berhubungan yang dikenal sebagai basis data dan sekumpulan data yang saling berhubungan yang dikenal sebagai basis data dan sekumpulan program perangkat lunak untuk mengatur dan mengakses data (Han, et al, 2012: 9).

2.2.2 Data Mining

Data mining merupakan metode untuk menemukan suatu pengetahuan dalam suatu database. Data mining adalah proses menggali dan menganalisa sejumlah data yang sangat besar, untuk memperoleh sesuatu yang benar, baru, sangat bermanfaat dan akhirnya dapat menemukan suatu corak atau pola dalam data tersebut (Ham, J and Camber, M,2006). [4]

2.2.3 Klasifikasi

Klasifikasi adalah proses untuk menentukan sebuah record data baru pada salah satu dari beberapa kategori (klas) yang sudah di definisikan sebelumnya. Disebut juga ‘supervised learning’. Metode klasifikasi mengutamakan pengelompokan data dengan menerapkan salah satu algoritma. Metode klasifikasi ini berguna untuk proses bisnis yang membutuhkan informasi kategoris.

Klasifikasi juga termasuk metode analisis data yang digunakan sebagai pembentuk model yang memprediksi trend data. Data pada pohon keputusan sering dinyatakan dalam bentuk tabel dengan atribut dan record. Atribut menyatakan parameter sebagai kriteria dalam pembentukan pohon (Hermawati.A.Fajar. 2013.) [6]

2.2.4 Algoritma Naive Bayes

Algoritma Naive Bayes merupakan salah satu algoritma yang terdapat pada teknik klasifikasi. Naive Bayes merupakan pengklasifikasian dengan metode probabilitas dan statistik yang dikemukakan oleh ilmuwan Inggris Thomas Bayes, yaitu memprediksi peluang di masa depan berdasarkan pengalaman dimasa sebelumnya sehingga dikenal sebagai Teorema Bayes. Teorema tersebut dikombinasikan dengan Naive dimana diasumsikan kondisi antar atribut saling bebas. Klasifikasi Naive Bayes diasumsikan bahwa ada atau tidak ciri tertentu dari sebuah kelas tidak ada hubungannya dengan ciri dari kelas lainnya (Bustami,2014). [9]

$$P(H | X) = \frac{P(X | H)P(H)}{P(X)}$$

Dimana :

X = data dengan class yang belum diketahui

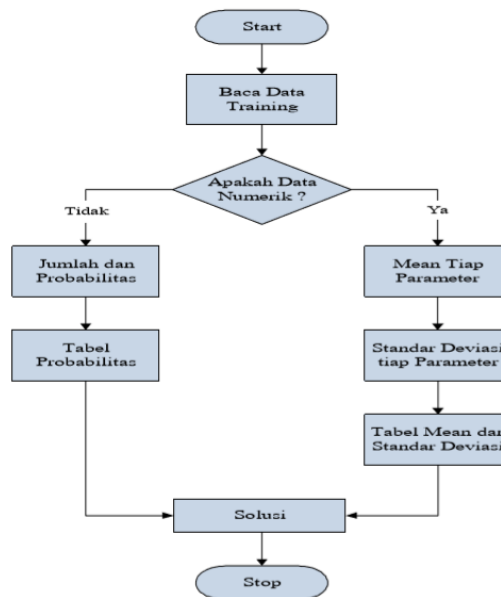
H = hipotesis data X, merupakan suatu class spesifik

P(H|X) = probabilitas hipotesis H berdasarkan kondisi X (posteriori probability)

P(H) = probabilitas hipotesis H (prior probability)

$P(X|H)$ = probabilitas X berdasarkan kondisi pada hipotesis H

$P(X)$ = probabilitas dari X



Gambar 1. Skema Naïve Bayes

2.2.5 Confusion Matrix

Confusion matrix merupakan visualisasi untuk mengevaluasi dari kinerja model klasifikasi. Untuk melakukan klasifikasi evaluasi komparatif, maka dalam penelitian ini menggunakan Confusion Matrix.

Confusion matrix ini meliputi informasi tentang kelas yang sebenarnya dan kelas prediksi. Hal ini akan ditemukan pada kolom matriks yang mewakili kelas yang diprediksi, sedangkan setiap baris mewakili kejadian pada kelas tersebut.

Confusion matrix adalah salah satu alat ukur berbentuk matrik 2x2 yang digunakan untuk mendapatkan jumlah ketepatan algoritma yang dipakai.

Confusion matrix disajikan pada table dibawah ini :

Tabel 1. Confusion Matrix

| | | Actual Class | |
|-----------------|-----------|------------------------|------------------------|
| | | Class=Yes | Class=No |
| Predicted Class | Class=Yes | TP (True Positive) | FP (False Positive) |
| | Class=No | FN (False Negative) | TN (True Negative) |

Keterangan :

True Positive (TP) : Jika data yang diprediksi bernilai positif dan sesuai dengan nilai aktual (positif).

False Positive (FP) : Jika data yang diprediksi tidak sesuai dengan nilai actual.

False Negative (FN) : Jika yang diprediksi bernilai negatif dan aktualnya positif.

True Negative (TN) : Jika benar antara prediksi negatif dan kenyataannya negatif.

Untuk mengukur performance dari hasil data mining salah satunya menggunakan akurasi, adapun rumusnya bila dilihat di bawah ini :

$$accuracy = \frac{\text{number of correct}}{\text{total number of prediction}} \times 100\%$$

Selain akurasi, metric evaluasi lainnya yaitu *precision* dan *recall*. Metrik yang sering digunakan untuk alat penilaian diantaranya, yaitu:

$$Recall = \frac{TP}{TP + FN} \times 100\%$$

$$Precision = \frac{TP}{TP + FP} \times 100\%$$

1.2.6 Waikato Environment for Knowledge Analysis (WEKA)

WEKA adalah sebuah paket tools machine learning praktis. WEKA merupakan singkatan dari Waikato Environment for Knowledge Analysis, yang dibuat di Universitas Waikato, New Zealand untuk penelitian, pendidikan dan berbagai aplikasi. WEKA mampu menyelesaikan masalah-masalah data mining di dunia nyata, khususnya klasifikasi yang mendasari pendekatan-pendekatan machine learning. Perangkat lunak ini ditulis dalam hirarki class Java dengan metode berorientasi objek dan dapat berjalan hampir di semua platform (Dimov, Rossen. 2007) [10]

2.2. Kerangka Pemikiran

Tabel 2. Kerangka Pemikiran

| | | |
|---|---|-----------------------|
| MASALAH | | |
| Menerapkan Algoritma Naïve Bayes pada sistem prediksi penerimaan siswa baru di SMK Al-Amien Wonorejo Pasuruan | | |
| Tools | Data | Metode |
| WEKA | Data Siswa SMK AL-AMIEN Wonorejo Pasuruan angkatan 2016 | Algoritma Naïve Bayes |
| Eksperimen/Analisa* | | |
| Pengujian dan validasi hasil | | |
| Menguji kinerja algoritma Naïve Bayes dalam memprediksi penerimaan siswa baru | | |
| Hasil | | |
| Dapat menghasilkan prediksi penerimaan siswa baru dengan tepat | | |

III METODE PENELITIAN

3.1. Kebutuhan Sistem

Berdasarkan permasalahan yang telah diuraikan pada bab sebelumnya, maka bahan dan peralatan yang diperlukan untuk penelitian ini meliputi :

A. Bahan

Dalam penelitian ini bahan yang dibutuhkan adalah Data yang diperoleh dari Sekolah Menengah Atas Shalahuddin Pasuruan. Data yang dikumpulkan yaitu data siswa baru tahun ajaran 2015/2016.

B. Peralatan

Peralatan dalam penelitian ini meliputi kebutuhan perangkat lunak dan kebutuhan perangkat keras.

Kebutuhan Perangkat Lunak

1. Microsoft Windows 7 sebagai OS notebook.
2. Microsoft Excel 2010 sebagai media penulisan datashet.

3. Software data mining Weka 3.9.2

Kebutuhan perangkat keras

1. Perangkat keras notebook dengan spesifikasi :

- a. Prosesor intel core i3 2.1 GHz
- b. RAM 2GB
- c. Hardisk 160GB

3.2. Pengumpulan Data

Data yang digunakan adalah dataset dari Pendaftaran peserta didik SMK Al-Amien Wonorejo Pasuruan. Maka metode pengumpulan data yang digunakan adalah dengan mengambil dataset disekolahan.

Tabel 3. Dataset

| NO | NAMA | MAT | IPA | B.IND | B.ING | NA | H.Akhir |
|-----|--------------------|------|------|-------|-------|-------|---------|
| 1 | FIRDA NINGSIH | 90 | 90 | 90 | 90 | 73.7 | Y |
| 2 | RYAN REYNALDI | 90 | 92.5 | 90 | 96 | 73.34 | Y |
| | PRATAMA | | | | | | |
| 3 | MUHAMMAD MUHAIMIN | 90 | 85 | 86 | 88 | 72.76 | Y |
| | MUHAMMAD RIZKI | | | | | | |
| 4 | PRATAMA | 85 | 87.5 | 92 | 86 | 72.22 | Y |
| | ANDIKA DWI RACHMAN | | | | | | |
| 5 | PUTRA | 80 | 90 | 92 | 86 | 71.32 | Y |
| | FIRDA SAFITRI | | | | | | |
| 6 | ARAUFY PRATAMA | 75 | 85 | 90 | 88 | 71.32 | Y |
| | ARFIANSYAH | | | | | | |
| 7 | FEBRYLIANA DEWI | 87.5 | 97.5 | 86 | 88 | 70.96 | Y |
| | WAHYUNINGRAT | | | | | | |
| 8 | MARSHAL SHEVA | 52.5 | 67.5 | 92 | 66 | 70.72 | Y |
| | ALJUNZA | | | | | | |
| 9 | | 85 | 92.5 | 92 | 96 | 70.62 | Y |
| | | | | | | | |
| 196 | | 87.5 | 92.5 | 68 | 86 | 13.36 | N |
| | FARID REZA FAHREFI | | | | | | |

Variabel yang terdapat pada dataset terdiri dari 4 atribut dan 1 label,yaitu :

1. Atribut Matematika
 - a. Nilai mapel matematika
2. Atribut IPA
 - a. Nilai mapel ilmu pengetahuan alam (IPA)
3. Atribut Bhs. Indonesia
 - a. Nilai mapel Bhs. Indonesia
4. Atribut Bhs.Ingggris
 - a. Nilai mapel Bhs. Inggris

IV HASIL DAN PEMBAHASAN

4.1. Data yang digunakan

Tabel 4. Data Testing

| NO | NAMA | MAT | IPA | B.IND | B.ING | NA | H.Akhir |
|----|-------------------|-----|------|-------|-------|-------|---------|
| 1 | FIRDA NINGSIH | 90 | 82.5 | 90 | 90 | 73.7 | Y |
| 2 | RYAN | 90 | 92.5 | 90 | 96 | 73.34 | Y |
| | REYNALDI PRATAMA | | | | | | |
| 3 | MUHAMMAD MUHAIMIN | 90 | 85 | 86 | 88 | 72.76 | Y |

Nama – nama tersebut diatas yang akan diuji coba apakah siswa tersebut termasuk siswa yang diterima atau tidak.

Adapun rumus untuk menghitung rata-rata (*mean*) sebagai berikut :

$$\mu (\text{mean}) = \frac{\sum x}{n}$$

Keterangan :

n = Jumlah siswa diterima atau tidak diterima

$\sum x$ = Akumulasi nilai

Contoh : $\mu = \frac{\sum x}{n}$

$$= \frac{90+90+90+85+80+75+87,5+52,5+85+75+\dots+47,5}{70}$$

$$= \frac{5282,501}{70}$$

$$= 75,4643$$

Rumus untuk menghitung standar deviasi sebagai berikut :

$$\sigma (\text{standar deviasi}) = \sqrt{\frac{\sum (x - \mu)^2}{n}}$$

Keterangan :

X = nilai

μ = Rata - rata

n = Jumlah siswa diterima atau tidak diterima

Contoh :

$$\sigma = \sqrt{\frac{\sum (x - \mu)^2}{n}}$$

$$= \frac{(90-75,4643)^2 + (90-75,4643)^2 + \dots + (47,5-75,4643)^2}{70} = 14,7214$$

Tabel 5. perhitungan mean dan stdev atribut

| VARIABEL | MEAN | | STDEV | |
|-------------|------------|-------------|------------|-------------|
| | Masuk | Tidak Masuk | Masuk | Tidak Masuk |
| Matematika | 75.4643 | 65.853 | 14.7214 | 21.816 |
| IPA | 80.71429 | 74.325 | 12.3809 | 17.614 |
| B.Indonesia | 85 | 79.476 | 9.06498 | 10.323 |
| B. Inggris | 83.8571429 | 76.444 | 9.57848694 | 15.259 |
| Nilai Akhir | 66.301429 | 53.19 | 3.2081858 | 7.858 |

Tabel diatas merupakan hasil perhitungan variabel, dihitung mean dan stdev baik yang masuk maupun yang tidak masuk.

4.2. Hasil dan Evaluasi Metode Naïve Bayes

Pada penelitian ini, hasil pengujian yang didapatkan pada metode naive bayes dapat dilihat pada tabel confusion matriks dibawah ini.

Tabel 6. Confusion Matriks Pada Naive Bayes

| | True Y | True N |
|--------------|--------|--------|
| Prediction Y | Tp=63 | Fn=20 |
| Prediction N | Fp=7 | Tn=106 |

Pada tabel confusion matriks diatas dapat diketahui 5 persamaan persamaan model confusion matriks yaitu nilai akurasi, sensitivity dan specificity, positif predictive value, negative predictive value.

1. Nilai Akurasi

$$Tp = 63$$

$$Fn = 20$$

$$Fp = 7$$

$$Tn = 106$$

$$\begin{aligned} akurasi &= \frac{tp + tn}{tp + tn + fp + fn} \\ &= \frac{63 + 106}{63 + 106 + 7 + 20} \\ &= 86,22 \% \end{aligned}$$

2. Nilai Senitivity

$$\begin{aligned} sensitivity &= \frac{tp}{tp + fn} \\ &= \frac{63}{63 + 20} = 75,90 \% \end{aligned}$$

3. Nilai Specificity

$$\begin{aligned} specificity &= \frac{tn}{tn + fp} \\ &= \frac{106}{106 + 7} = 93,80 \% \end{aligned}$$

4. Positif predictive value

$$\begin{aligned} PPV &= \frac{tp}{tp + fp} \\ &= \frac{63}{63 + 7} = 90 \% \end{aligned}$$

5. Negative Predictive Value

$$\begin{aligned} NPV &= \frac{tn}{tn + fn} \\ &= \frac{106}{106 + 20} = 84,13 \% \end{aligned}$$

5.1. Kesimpulan

Hasil pengujian algoritma *Naïve Bayes* dalam prediksi penerimaan siswa baru di Sekolah Menengah Kejuruan (SMK) pada 196 data siswa yang diuji dalam penelitian ini, menunjukkan bahwa algoritma *Naïve Bayes* memiliki tingkat akurasi 86.22 %.

5.2. Saran

Penelitian ini hanya menggunakan variabel siswa dalam bentuk nilai mata pelajaran pada ujian nasional sebagai variabel komputasi dengan *Naïve Bayes* sesuai dengan metode yang diterapkan pada sekolah menengah kejuruan. Oleh karena itu, disarankan agar penelitian ini dikembangkan lagi dalam menambahkan variabel lain. Untuk meningkatkan akurasi yang lebih baik dalam penerapan *Naïve Bayes* untuk prediksi penerimaan siswa baru serta memberikan kontribusi yang lebih besar didunia riset, disarankan agar hasil penelitian ini dikembangkan dengan cara memodifikasi dengan menggabungkan algoritma *Naïve Bayes* dengan algoritma yang lain.

DAFTAR PUSTAKA

- [1] M. H. Bisri, "Implementasi Algoritma Naive Bayes untuk Memprediksi Penjurusan Siswa di SMA Kesatria 1 Semarang," *Jurnal Informatika*, pp. 1-6, 2015.
- [2] C. A. Sugianto, "Penerapan Teknik Data Mining Untuk Menentukan Hasil Seleksi Masuk SMA 1 Gibeber Untuk Siswa Baru Menggunakan Decision Tree," *Jurnal Informatika*, pp. 39-43, 2015.
- [3] A. A. Muin, "Metode Naive Bayes Untuk Prediksi Kelulusan," *Jurnal Ilmiah Ilmu Komputer*, vol. II, pp. 1-5, 2016
- [4] Han, J., and Kamber, M., "Data Mining Concept and Technique," *Morgan Kaufman Publisher, San Fransisco*, 2006
- [5] Larose, Daniel T., "Discovering Knowledge in Data an Introduction to Data Mining," *Wiley Interscience*, pp. 11-17, 2005.
- [6] Hermawati.A.Fajar, "Data Mining," *Andi, Yogyakarta*, 2013.
- [7] Kusriani, dan Emha Taufiq Luthfi "Algoritma Data Mining," *Andi Offset*, 2009.
- [8] M. F. R. C. M. H. A. & K. Kabir, "Enhanced Classification Accuracy on Naive Bayes Data mining Models," *International Journal of Computer Applications*, vol. 28, 2011.
- [9] Bustami, "Penerapan Algoritma Naive Bayes Untuk Mengklasifikasi Data Nasabah Asuransi," *Jurnal Informatika*, vol. VIII, pp. 884-898, 2014.
- [10] Dimov, Rossen. "WEKA : Practical machine learning tools and techniques with java implementations," *AI Tools Seminar*, 2007.
- [11] Bouckaert, Remco R, Eibe Frank, Mark Hall et al, "WEKA manual for version 3-6-0," *New Zealand: University of Waikato*, 2