Jurnal Mantik Penusa

Volume 3, No. 1.1, Agustus 2019

e-ISSN 2580-9741
p-ISSN 2088-3943

PREDIKSI KEGAGALAN SISWA DALAM DATA MINING DENGAN MENGGUNAKAN METODE NAÏVE BAYES

Rumini[1], Norhikmah[2]

[1,2]IlmuKomputer,
Universitas AMIKOM Yogyakarta, Jl. Ringroad Utara, Condong Catur, Depok, Sleman., Yogyakarta

Email: rumini*@amikom.ac.id[1], norhikmah@amikom.ac.id[2]*

Abstrak

Dalam upaya meningkatkan kualitas suatu bangsa, tidak ada cara lain kecuali melalui peningkatan mutu pendidikan. Keberhasilan dan kegagalankelas siswa dalam studi ditentukan banyak faktor yang mempengaruhinya. Hasil penelitian ini adalah menggunakan metode data mining naïve bayes yang digunakan untuk memprediksi kegagalan siswa dalam studinya serta faktor-faktor berpengaruh diantaranya failure, traveltime, internet, romantic, freetime, go-out, health, dan absence. Algoritma naïve bayes merupakan salah satu algoritma yang dapat digunakan untuk memprediksi dengan menggunakan teori probabilitas dengan tingkat akurasi tinggi. Pengujian algoritma Naïve Bayes menggunakan tool WEKA yang menghasilkan tingkat akurasi sebesar 77.97 dari 395 dataset. Algoritma ini digunakan untuk memprediksi kegagalan kelassiswa.

Keywords: Data Mining, KegagalanSiswa, Naïve Bayes

Abstract

In an effort to improve the quality of a nation, there is no other way except through improving the quality of education. The success and failure of students' classes in the study determined many factors that influence it. The results of this study are using the naïve bayes data mining method which is used to predict student failures in their studies as well as influential factors including failure, traveltime, internet, romantic, freetime, go-out, health, and absence. Naïve Bayes algorithm is an algorithm that can be used to predict using probability theory with a high degree of accuracy. Naïve Bayes algorithm testing uses WEKA tool which produces an accuracy of 77.22 from 395 datasets. This algorithm is used to predict student class failures.

Keywords: Data Mining, Student Failure, Naïve Bayes

1. Pendahuluan

Pendidikan merupakan hal penting dan tidak dapat dilepaskan dalam kehidupan manusia. Kita semua sebagaimanusia membutuhkan pendidikan untuk mewujudkan apa yang ingin dicapai. Pendidikan yang telah dirancang dan dibangun harus kuat dan tidak mudah tergoyahkan, bangunan pendidikan harus diperkokoh dengan pilar-pilar yang kuat. UNESCO sebagai lembaga PBB yang bergerak di bidang pendidikan, ilmu pengetahuan, dan budaya telah merumuskan empat pilar pendidikan, hal ini dimaksudkan agar tujuan pendidikan dapat terwujud. Empat pilar pendidian tersebut adalah belajar untuk mengetahui (*learning to know*), belajar untuk melakukan (*learning to do*), belajar untuk menjadi (*learning to be*), belajar untuk hidup bersama (*learning to live together*).

Pendidikan dalam ruang lingkup pembelajaran dilingkungan sekolah tidak terlepas dari peran guru. Guru yang memberikan dan mengatur pembelajaran dengan seksama berdasarkan kurikulum yang telah ditentukan oleh lembaga pendidikan negara. Guru juga berperan dalam upaya meningkatkan kualitas suatu bangsa, melalui peningkatan mutu pendidikan. Selama masa studipembelajarandikelas ada siswa yang berhasil naik kelas dalam akademiknya, tetapi tidak jarang para siswa juga ada yang gagal kelas (tidak naik kelas). Keberhasilan dan kegagalan siswa dalam studi ditentukan banyak faktor yang mempengaruhinya. Faktor-faktor tersebut dapat datang dari internal maupun eksternal.

2. Teori

a. Pengertian Data Mining

Data mining merupakan proses semi otomatik yang menggunakan teknik statistik, matematika, kecerdasan buatan, dan machine learning untuk mengekstraksi dan mengidentifikasi informasi pengetahuan potensial dan berguna yang bermanfaat yang tersimpan di dalam database besar. Salah satu

teknik data mining adalah klasifikasi (*classification*).Klasifikasi adalah menentukan recorddata baru kesalah satu dari beberapa kategori (*class*) yang telah didefinisikan sebelumnya. Biasanya hal ini disebut juga dengan *supervised learning*. Klasifikasi merupakan penempatan objek – objek kesalah satu dari beberapa kategori yang telah ditetapkan sebelumnya. Klasifikasi sekarang ini telah banyak digunakan dalam berbagai aplikasi, sebagai contoh pendeteksian pesan email, spam berdasarkan header dan isi atau mengklasifikasikan galaksi berdasarkan bentuk-bentuknya. Pada proses klasifikasi data yang diinputkan adalah data record atau data sampel. Pada setiap record dikenal sebagai instance atau atau contoh yang ditentukan oleh sebuah tuple (x,y). Dimana x adalah himpunan atribut dan y adalah atribut tertentu yang menyatakan sebagai label class (Turban E, dkk. 2005).

Berdasarkanfungsionalitasnya, tugas-tugas data mining dikelompokkan kedalam enam kelompok berikut ini (Fayyad et al. 1996):

1) Klasifikasi (*Classification*), men-generali sasis truktur yang diketahui untuk diaplikasikan pada data-data baru. Misalkan, klasifikasi penyakit kedalam sejumlah jenis.
2) Klasterisasi (*clustering*), mengelompokkan data yang tidak diketahui label kelsnya, kedalam sejumlah kelompok tertentu sesuai ukuran kemiripannya.
3) regresi, menemukan suatufungsi memodelkan data dengan galat (kesalahan prediksi) seminimal mungkin.
4) Deteksi anomali (*anomalydetection*), mengidentifikasi data yang tidak umum, dapat berupa outlier (pencilan), perubahan atau deviasi yang munkgin sangat penting dan perluinvestigasilebihlanjut.
5) Pembelajaran aturan asosiasi (*assosiationrule learning*) atau pemodelan kebergantungan (*dependency modeling*), mencari relasi/hubungan antar variabel.
6) Perangkuman (*summarization*), menyediakanrepresentasi data yang lebihsederhana, meliputivisualisasi dan pembuatanlaporan.

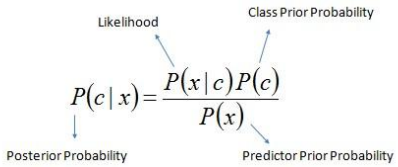
b. Metode Naïve Bayes

Naive Bayes merupakan sebuah pengklasifikasian probabilistik sederhana yang menghitung sekumpulan probabilitas dengan menjumlahkan frekuensi dan kombinasi nilai dari dataset yang diberikan. Algoritma mengunakan teorema Bayes dan mengasumsikan semua atribut independen atau tidak saling ketergantungan yang diberikan oleh nilai pada variabel kelas(Patil *and* Sherekar.2013).

Definisi lain mengatakan *Naive Bayes* merupakan pengklasifikasian dengan metode probabilitas dan statistik yang dikemukan oleh ilmuwan Inggris Thomas Bayes, yaitu memprediksi peluang di masa depan berdasarkan pengalaman di masa sebelumnya (Bustami. 2013).

Naive Bayes didasarkan pada asumsi penyederhanaan bahwa nilai atribut secara kondisional saling bebas jika diberikan nilai *output*. Dengan kata lain, diberikan nilai *output*, probabilitas mengamati secara bersama adalah produk dari probabilitas individu(Pattekari and Parveen. 2012). Keuntungan penggunaan *Naive Bayes* adalah bahwa metode ini hanya membutuhkan jumlah data pelatihan (*TrainingData*) yang kecil untuk menentukan estimasi paremeter yang diperlukan dalam proses pengklasifikasian. *Naive Bayes* sering bekerja jauh lebih baik dalam kebanyakan situasi dunia nyata yang kompleks dari pada yang diharapkan.

Bayes merupakan pengklasifikasian statistik yang dapat digunakan untuk memprediksi probabilitas keanggotaan suatu *class*. *Bayes* memiliki akurasi dan kecepatan yang sangat tinggi saat diaplikasi ke dalam *database* dengan data yang besar. Berikut teorema *bayes*(Turban E, dkk. 2005) :



Keterangan :

x : data dengan class yang belum diketahui
c : hipotesis data x merupakan suatu class spesifik
P(c|x) : probabilitas hipotesis c berdasar kondisi x (*posteriori probability*)
P(c) : probabilitas hipotesis c (*prior probability*)
P(x|c) : probabilitas x berdsar kondisi pada hipotesis c
P(x) : probabilitas dari x

3. Hasil Penelitian DanPembahasan

Berikut adalah contoh dataset dari UCI *Student Performance* untukstudent_mat dengan

variable *failure, traveltime, internet, romantic, freetime, go-out, health, dan absence*.

Tabel 1. Sample Dataset Faktor Kegagalan Siswa

No.	Travel time	Internet	Romantic	Freetime	Goout	Health	Absences	Failures
1	2	no	no	3	4	3	6	0
2	1	yes	no	3	3	3	4	0
3	1	yes	no	3	2	3	10	3
26	1	yes	no	2	2	5	14	2
41	2	yes	yes	3	3	3	25	1
395	1	yes	no	2	3	5	5	0

Tabel 2 dibawah ini, merupakan tabel data training untuk perhitungan manual dengan metode naïve bayes untuk memprediksi status berhasil atau gagal di kelas.

Tabel 2. Data Training CalonSiswa yang BerhasilatauGagal Kelas

No.	Traveltime	Internet	Romantic	Freetime	Go-Out	Health	Absences	Status (gagal/tidak)
396	3	No	No	4	3	5	20	?

a. Klasifikasi

Pada kegiatan ini, peneliti melakukan analisa terhadap atribut class yang menjadi faktorkegagalan dan keberhasilankelassiswa. Klasifikasi faktor-faktor dalam tabel 1 diatas merupakan akan menjadi keputusan kegagalan studi dengan level kegagalan rendah, sedang dan tinggi. Jumlahdataset yang digunakan adalah395 data dengan 7 kriteria. Berikut klasifikasi pada setiap kriteria yang ada :

1) Klasifikasi Traveltime (Waktu perjalanan rumah ke sekolah)

Waktu	Keterangan
(1) <15 Menit	Dekat
(2) 15 - 30 Menit	Cukup Jauh
(3) 30 – 60 Menit	Jauh
(4) > 60 Menit	Sangat Jauh

2) KlasifikasiInternet (akses Internet dirumah)

JaringanInternet	Keterangan
Yes (Ada)	YI
No (Tidak ada)	NI

3) KlasifikasiRomantic Hubungan Romantis)

Hubungan Romantis	Keterangan
Yes (Ada)	YR
No (Tidak Ada)	NR

4) KlasifikasiFreetime (Waktu Luang sepulang sekolah

Waktu Luang	Keterangan
1	Sangat Rendah
2	Rendah
3	Sedang
4	Tinggi
5	Sangattinggi

5) KlasifikasiGo-Out (Pergi Bersama teman)

Waktu	Keterangan
1	Sangat Jarang
2	Jarang
3	Sedang
4	Sering
5	Sangat Sering

6) KlasifikasiHealth (Status Kesehatan)

Status Kesehatan	Keterangan
1	Sangat buruk
2	Buruk
3	Sedang
4	Baik
5	Sangat Baik

7) KlasifikasiAbsences (Jumlah absensi sekolah)

Jumlah Absensi	Keterangan
0 - 10	Sangat Rendah
11 – 20	Rendah
21 – 30	Sedang
>31	Tinggi

b. Perhitungan Terhadap Data Training Menggunakan Metode Naïve Bayes

1) Menghitung jumlah class/label
P(class=tidak gagal kelas/berhasil) = 312/395 = 0.79, "Jumlah data tidak gagal kelas (berhasil) pada kolomstatus dibagi dengan jumlah keseluruhan data".
P(class=gagal kelas) = 83/395 = 0.21, "Jumlah data Gagal kelas pada kolom status dibagi dengan jumlah keseluruhan data".

2) Menghitung jumlah kasus dengan kasus yang sama
P(*traveltime* = jauh) | class = berhasil) = 17/312 = 0.0545
P(*traveltime* = jauh) | class = gagal) = 6/83 = 0.0723
P(internet=NI | class = berhasil) = 47/312 = 0.151
P(internet=NI | class = gagal) = 19/83 = 0.229
P(*romantic* = NR | class = berhasil) = 216/312 = 0.692
P(*romantic* = NR | class = gagal) = 47/83 = 0.566
P(freetime = tinggi | class = berhasil) = 86/312 = 0.276
P(freetime = tinggi | class = gagal) = 29/83 = 0.349
P(go-out = tinggi | class = berhasil) = 111/312 = 0.356
P(go-out = tinggi | class = gagal) = 19/83 = 0.229
P(*health* = sangat baik | class = berhasil) = 110/312 = 0.353
P(*health* = sangat baik | class = gagal) = 36/83 = 0.434
P(*absences* = rendah | class = berhasil) = 32/312 = 0.103
P(*absences* = rendah | class = gagal) = 20/83 = 0.241

3) Mengalikan semua hasil variable berhasil dan gagal kelas
(P | berhasil) = { P(*traveltime* = jauh) | class = berhasil) * P(internet=NI | class = berhasil) * P(*romantic* = NR | class = berhasil) * P(*freetime* = tinggi | class = berhasil) * P(*go-out* = tinggi | class= berhasil) * P(*health* = sangat baik | class = berhasil) * P(*absences* = rendah | class = berhasil)
= 0.0545 * 0.151 * 0.692 * 0.276 * 0.356 * 0.353 * 0.103= 2.034

Berikut adalah hitungan untuk status gagal:
(P | gagal) = { P(traveltime = jauh) | class = gagal) * P(internet=NI | class = gagal) * P(romantic = NR | class = gagal) * P(freetime = tinggi | class = gagal) * P(go-out = tinggi | class = gagal) * P(health = sangat baik | class = gagal) * P(absences = rendah | class = gagal)
= 0.0723 * 0.229 * 0.566 * 0.349 * 0.229 * 0.434 * 0.241= 7.834

4) Membandingkan hasil class, berhasil dan gagal kelas
Perbandingan hasil perhitungan class, adalah bahwa hasil (P | gagal) lebih besar dari pada (P | berhasil), maka hasil keputusannya dari data training tabel 2 diatas adalah statusnya gagal kelas.

c. Klasifikasi dengan tools WEKA

Pada gambar 1 dibawah ini merupakan output peng-klasifikasian pada dataset.

```
=== Evaluation on training set ===

Time taken to test model on training data: 0 seconds

=== Summary ===

Correctly Classified Instances         308               77.9747 %
Incorrectly Classified Instances        87               22.0253 %
Kappa statistic                          0.0572
Mean absolute error                      0.2973
Root mean squared error                  0.3899
Relative absolute error                 89.3206 %
Root relative squared error             95.7008 %
Total Number of Instances              395

=== Detailed Accuracy By Class ===
```

Gambar 1. Classifier Output

Berikut nilai akurasi algoritma Naïve Bayes menggunakan aplikasi WEKA:
Dari gambar hasil perhitungan metode naïvebayes diatas dengan menggunakan bantuan tools wek apada presentase untuk correctly Classified Instance adalah sebesar 77.97% sementara persentase untukuntuk incorrectly Classified Instance adalah sebesar 22.03%.
Dari dataset 395 studi siswa, adase banyak 308 data siswa berhasil diklasifikasikan dengan benar dan sebanyak 87 data tidak berhasil diklasifikasikan dengan benar.

d. Pengujian Kinerja Algoritma

Pengujian dilakukan untuk mengukur keakuratan hasil dari tiap model yang diusulkan. Akurasi didefinisikan sebagai tingkat kedekatan antara nilai prediksi dengan nilai aktual. Pengukuran akurasi terhadap model dengan menggunakan confusion matrix yang menitik beratkan pada kelasnya. Confusion matrix merupakan tabel untuk mencatat hasil kerja klasifikasi ditunjukkan pada gambar 2.
Sehingga diketahui dari 395 data, 312 diklasifikasikan sebagai class berhasil, lalu 83 data diklasifikasikan sebagai class gagal.

```
=== Confusion Matrix ===

   a   b   <-- classified as
 302  10 |   a = Berhasil
  77   6 |   b = Gagal
```

Gambar 2. Confusion Matrix

4. Kesimpulan

Data mining dengan metode Naïve Bayes menggunakan data training untuk menghasilkan probabilitas dalam setiap kriteria untuk class yang berbeda, sehingga nilai-nilai probabilitas dari kriteria tersebut dapat dioptimalkan untuk memprediksi suatu kondisi berdasarkan proses klasifikasi yang dilakukan oleh metode Naïve Bayes.

Kesimpulannya adalah bahwa metode tersebut dapatdigunakan untuk memprediksi kegagalan kelas pada siswa dengan tingkat persentase keakuratan perhitungan algoritma naïve bayes menggunakan tool WEKA dapat dilihat menunjukkan 77.97% algoritma naïve bayes tepat digunakan untuk memprediksi kegagalan kelas pada siswa, sedangkan 22.03% tidak dapat membantu dalam pengambilan keputusan.

5. Reerensi

[1] Fayyad, U., Piatetsky-Shapiro, G., & Smyth, P. 1996. From data mining to knowledge discovery in database. AI magazine: p.37-54. Available at: http://www.aaai.org/ojs/index.php/aimagazine/article/viewArticle/123O.

[2] Bustami.2013. Penerapan Algoritma Naive Bayes Untuk Mengklasifikasi Data Nasabah Asuransi, TECHSI : Jurnal Penelitian Teknik Informatika, Vol. 3, No.2, Hal. 127-146.

[3] Manalu, E., dkk. 2017. PenerapanAlgoritma Naïve Bayes untuk Mempredikasi Jumlah Produksi barang Berdasarkan Data Persediaan dan Jumlah Pemesanan Pada CV. Papa dan Mama Pastries. Jurnal Mantik Penusa, Vol.1 No.2, Hal. 16-21.

[4] Patil, T. R., Sherekar, M. S., (2013), Performance Analysis of Naive Bayes and J48 Classification Algorithm for Data Classification, International Journal of Computer Science and Applications, Vol. 6, No. 2, Hal 256-261.

[5] Pattekari, S. A., Parveen, A., (2012), Prediction System for Heart Disease Using Naive Bayes, International Journal of Advanced Computer and Mathematical Sciences, ISSN 2230-9624, Vol. 3, No 3, Hal 290-294.

[6] Turban, E., dkk. Decision Support System and Inteligent System edisi 7 jilid1.Yogyakarta : Penerbit Andi, 2005.