NUANSA INFORMATIKA ***p-ISSN :1858-3911* , *e-ISSN : 2614-5405***
Volume 17 Nomor 2, Juli 2023 https://journal.fkom.uniku.ac.id/ilkom

Variabel Non Akademik Untuk Memprediksi Prestasi Siswa Dengan Data Mining Menggunakan Metoda Naïve Bayes

Arnold Ropen Sinaga
Sistem Informasi, Fakultas Teknologi dan Informatika
Sistem Informasi Universitas Informatika dan Bisnis Indonesia
E-mail: arnoldropen@unibi.ac.id

Abstrak

Penelitian ini bertujuan untuk mengetahui tingkat akurasi serta hasil presisi dan recall dengan penerapan data mining untuk memprediksi hasil belajar siswa menengah pertama (SMP) berdasarkan status jenis kelamin, asal sekolah, pendidikan dan pekerjaan orang tua. Penentuan hasil belajar dari peserta didik merupakan hal penting dalam dunia pendidikan. Hal ini merupakan hal yang penting karena sulitnya menentukan faktor atau variabel yang mempengaruhi hasil belajar peserta didik.

Pengelolaan data mining yang tepat dapat mengenali dan mengekstrak pola pengetahuan menawarkan solusi untuk meningkatkan kualitas pendidikan dalam mengelola peserta didik ini sehinga dapat memaksimalkan prestasi mereka.

Ada beberapa model klasifikasi dalam data mining yaitu algoritma ID3 dan C4.5 juga Naïve Bayes yang dapat digunakan untuk memprediksi prestasi peserta khususnya peserta menengah pertama. Penelitian ini akan menggunakan metode klasifikasi naïve bayes untuk memprediksi prestasi belajar peserta didik SMP Santa Maria dengan harapan mendapat akurasi yang lebih baik

Kata Kunci*— Data Mining, Naive Bayes, Prestasi*

Abstract

The aim of this research is to measure not only the accuracy rate but also the precision result and the recall of the data mining application to predict Junior High School students' learning outcomes based on their gender, the origin of the school, parents' education and occupation. Determination of the students' learning outcomes are very important in the education world. it becomes important because of the difficulty in determining the factors and variables which can affect the students' learning outcomes.

The accurate process of the data mining can recognize and extract the pattern of knowledge in order to offer solutions to increase the education quality where it can help the students maximize their achievement.

There are some classification models in data mining: ID3 algorithm, C4.5 and Naïve Bayes which can be used to predict the students' achievement, specifically, in Junior High School. This research uses Naïve Bayes classification mode to predict the Saint Mary Junior High School students' achievement in order to get a better accuracy.

Keywords*— Data Mining, Naive Bayes, Achievment*

Diajukan: 20 June 2023 *Disetujui: 25 June 2023* *Dipublikasi: 11 July 2023*

1. PENDAHULUAN

Efektifitas proses pembelajaran sangat dipengaruhi oleh faktor internal maupun faktor eksternal. Faktor internal

p-ISSN :1858-3911* , *e-ISSN : 2614-5405
https://journal.fkom.uniku.ac.id/ilkom

yang terdapat atau terjadi di dalam lingkungan sekolah dapat terlihat oleh para tenaga pendidik atau tenaga non kependidikan, sehingga dengan cepat dapat diatasi. Tetapi untuk faktor eksternal tidak dapat langsung terlihat sehingga sukar ditangani dengan cepat.

Proses pembelajaran dalam periode tertentu akan terkumpul sejumlah data yang besar. Kumpulan data tersebut dapat diproses lebih lanjut untuk menghasilkan pola informasi yang baru untuk dapat digunakan dalam meningkatkan efektifitas dalam proses pembelajaran, hal tersebut sangat berpengaruh pada peningkatan mutu siswa yang dihasilkan oleh sekolah, akan meningkatkan kecerdasan intelektual dan perilaku. Aspek-aspek yang yang dapat digunakan untuk data mining antara lain, Pendidikan orang tua, pekerjaan orang tua, kondisi keluarga, jarak tempat tinggal dengan sekolah dan berbagai aspek yang lain..

2. METODE PENELITIAN

1. Metoda Penelitian

Peneltian yang penulis lakukan di SMP Santa Maria menggunakan pendekatan kuantitatif. Target penelitian adalah siswa kelas 7 mulai tahun pelajaran 2013/2014 – 2017/2018 sejumlah 455 siswa yang diperoleh dari bagian Tata Usaha. Data yang digunakan adalah dalam bentuk comma separated values (.csv) yang dikonversi dari data yang diambil dari 5 buah file Excel
Penelitian ini menggunakan metode yang diantaranya adalah:

a. Pengamatan (Observasi) Melakukan pengamatan langsung ke bagian Tata Usaha SMP Santa Maria Bandung untuk mendapatkan data yang dibutuhkan.

b. Wawancara (Interview) Mengadakan wawancara dengan pihak-pihak yang berkaitan langsung dengan permasalahan yang sedang di bahas pada tugas akhir ini untuk memperoleh gambaran dan penjelasan secara mendasar.

c. Studi Pustaka Penulis mengumpulkan berbagai referensi dan literatur pendukung penelitian berupa buku, jurnal dan artikel yang berasal dari berbagai sumber yang erat kaitannya dengan objek permasalahan

2. Tinjauan Pustaka

a. Data mining

Adalah aktivitas yang menggambarkan sebuah proses analisis yang terjadi secara literatif pada database yang besar, dengan tujuan mengekstrak informasi dan knowledge yang akurat dan berpotensial berguna untuk knowledge workers yang berhubungan dengan pengambilan keputusan dan pemecahan masalah.

Menurut Han dan Kamber, (2011, p24), secara garis besar data mining dapat dikelompokkan menjadi 2 kategori utama, yaitu:

1. Predictive

Suatu proses penemuan pola dari data yang besar dengan menggunakan beberapa variabel prediksi. Klaisifikas merupakan salah satu teknik dalam predictive mining. Tujuan proses ini adalah memprediksi suatu nilai dari atribut (variable target / terikat) yang ditentukan dari nilai atribut yang lain (Variabel bebas)

2. Descriptive

Bertujuan menemuan karakteristik data penting data dalam sekumpulan data yang besar dengan cara menurunkan pola-pola (korelasi, trend, cluster, teritori, dan anomali). Tugas data deskriptif adalah menyelidiki pola-pola dalam hal validasi dan penjelasan hasil mengunakan teknik post-processing.

b. Naïve Bayes

Proses klasifikasi menggunakan metode statistik dan kemungkinan. Setiap kelas keputusan Naïve didapat dari perhitungan kemungkianan dengan asumsi kelas keputusan adalah benar karena merupakan vektor informasi obyek.

Dibanding dengan model klasifikasi yang lan, klasifikasi Bayesian mempunyai tingkat akurasi yang tinggi. Tetapi, seringkali dalam pelaksanaannya jarang terjadi, diakibatkan kurangnya akurasi asumsi yang dibuat dan

p-ISSN :1858-3911* , *e-ISSN : 2614-5405
https://journal.fkom.uniku.ac.id/ilkom

minimnya data yang digunakan dalam probabilitas yang ada.

c. Weka

Syarat yang harus dipenuhi dalam menggunakan Weka adalah bahwa data harus merupakan file tunggal berbentuk flat file,dimana atribut biasanya berjenis numeric atau nominal, tetapi dapat juga dengan jenis yang lain.

Format Input WEKA

Format file untuk input pendukungnya, antara lain:

a. Comma Separated Values (CSV): Merupakan file teks dengan pemisah tanda koma (,) dapat dibuat menggunakan MS Excel ataupun Notepad
b. Format C45: Hanya dapat digunakan menggunakan aplikasi WEKA.
c. Attribute-Relation File Format (ARFF): File berjenis teks sebagai instance data yang berhubungan dengan suatu set atribut data.
d. SQL Server/MySql Server: Dapat mengakses database dengan menggunakan SQL Server/MySql Server.

Objek dan metoda penelitian

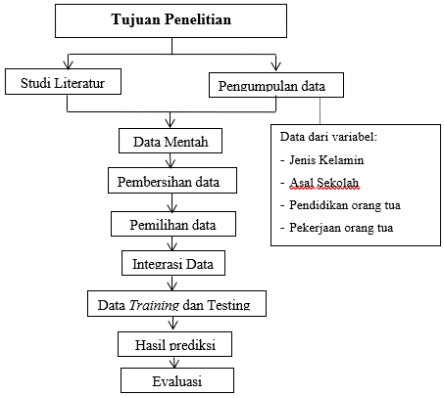
SMP Santa Maria Bandung merupakan salah satu sekolah katolik swasta yang berada di jalan Ahmad Yani No.273 Bandung. Sekolah ini merupakan sekolah yang berada dibawah naungan Yayasan Salib Suci yang berlamat di jalan Van Deventer no.18 Bandung. Perkembangan penerimaan siswa baru setiap tahunnya relatif stabil dikarenakan pangsa pasar untuk siswa baru sudah terdata dari rekapitulasi setiap tahunnya dari bagian panitia penerimaan siswa baru dan membatasi untuk setiap tahunnya.

Penelitian yang penulis lakukan di SMP Santa Maria menggunakan pendekatan kuantitatif. Target penelitian adalah siswa kelas 7 mulai tahun pelajaran 2013/2014 – 2017/2018 sejumlah 455 siswa yang diperoleh dari bagian Tata Usaha. Data yang digunakan adalah dalam bentuk comma separated values (.csv) yang dikonversi dari data yang diambil dari 5 buah file Excel

Metoda yang digunakan

Penelitian ini adalah menggunakan beberapa metoda:

1. Pengamatan (Observasi) Melakukan pengamatan langsung ke bagian Tata Usaha SMP Santa Maria Bandung untuk mendapatkan data yang dibutuhkan.
2. Wawancara (Interview), melakukan tanya jawab dengan pihak yang terlibat langsung dengan permasalahan yang sedang di bahas.
3. Studi Pustaka Penulis mengumpulkan berbagai referensi dan literatur pendukung penelitian berupa buku, jurnal dan artikel yang berasal dari berbagai sumber yang erat kaitannya dengan objek permasalahan



Gambar 1. Kerangka pikir penelitian

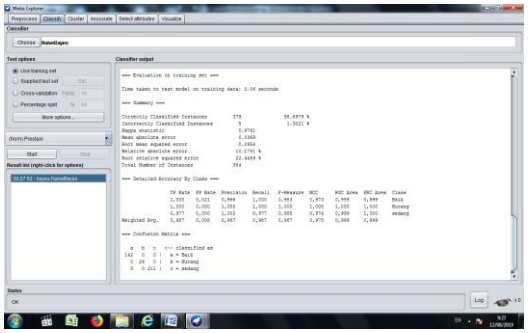
PENGOLAHAN DATA

Data Training dan Testing

1. Data training

Data yang akan di training menggunakan data alumni SMP Santa Maria yang memiliki 384 record/instances yang dijadikan data training merupakan data siswa yang diambil dari data siswa tahun pelajaran 2013/2014 sampai dengan tahun pelajaran 2016/2017. Hasil data training dapat dilihat pada gambar 2

p-ISSN :1858-3911* , *e-ISSN : 2614-5405
https://journal.fkom.uniku.ac.id/ilkom



Gambar 2. Data Training

Hasil pada data training yang menggunakan file Data-master-siswa - training-1316.ARFF, maka akan terlihat bahwa terlihat dari pada tabel 1

Tabel 1. *Confusion Matrix Data Training*

A	b	C	←*Classified Class*
142	0	0	a ← Baik
0	26	0	b ← Kurang
5	0	211	c ← Sedang

Baris pertama "142 0 0" menunjukkan bahwa ada "142" (142+0+0) instances class "Baik" dalam file Data-master-siswa - training-1316.ARFF dan semua benar diklasifikasikan sebagai instances class "Baik".

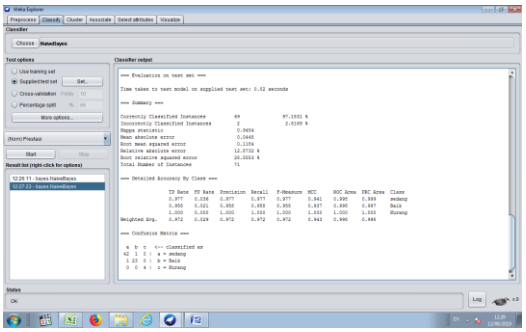
a. Baris kedua "0 26 0" menunjukkan bahwa ada "26" (0+26+0) instances class "Kurang", dalam file Data-master-siswa - training-1316.ARFF dan semua benar diklasifikasikan sebagai instances class "Kurang".
b. Baris ketiga "5 0 211" menunjukkan bahwa hanya ada "211" (5+0+211) instances class "Sedang", dalam file Data-master-siswa - training-1316.ARFF dan 5 diantaranya salah diklasifikasikan sebagai instances class "Baik".

Dari hasil confusion matrix dapat dilihat bahwa tingkat akurasi benar untuk data training adalah 98,7 % (379 instance) sedangkan kesalahan prediksi sebesar 1.3% (5 instance)

2. Data Testing

Data yang akan di testing menggunakan data alumni SMP Santa Maria yang memiliki 71 record/instances yang dijadikan data testing merupakan data siswa yang diambil dari data siswa tahun pelajaran 2017/2018.

Hasil data training dapat dilihat pada gambar 3.



Gambar 3. Data Testing

Hasil pada data testing yang menggunakan file Data-master-siswa - testing17 sebagai file data testing, akan terlihat bahwa terlihat dari pada tabel 2

Tabel 2. Confusion Matrix Data Testing

A	b	C	←*Classified Class*
42	1	0	a ← Sedang
1	23	0	b ← Baik
0	0	4	c ← Kurang

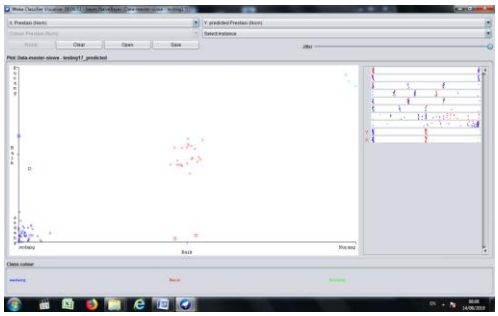
a. Baris pertama "42 1 0" menunjukkan bahwa ada "42" (42+1+0) instances class "Sedang" dalam file Data-master-siswa - training-1316.ARFF dan ada 1 diantaranya salah diklasifikasikan sebagai instances class "Baik".
b. Baris kedua "1 23 0" menunjukkan bahwa ada "23" (1+23+0) instances class "Baik", dalam file Data-master-siswa - training-1316.ARFF dan ada 1 diantaranya salah diklasifikasikan sebagai instances class "Sedang".
c. Baris ketiga "0 0 4" menunjukkan bahwa hanya ada "4" (0+0+4) instances class "Kurang", dalam file Data-master-siswa - training-1316.ARFF dan semua benar diklasifikasikan sebagai instances class "Kurang".

Dari hasil confusion matrix dapat ahwa tingkat akurasi benar untuk data testing adalah 97,2 % (69 instance) sedangkan kesalahan prediksi sebesar 2.8% (2 instance)

p-ISSN :1858-3911* , *e-ISSN : 2614-5405
https://journal.fkom.uniku.ac.id/ilkom

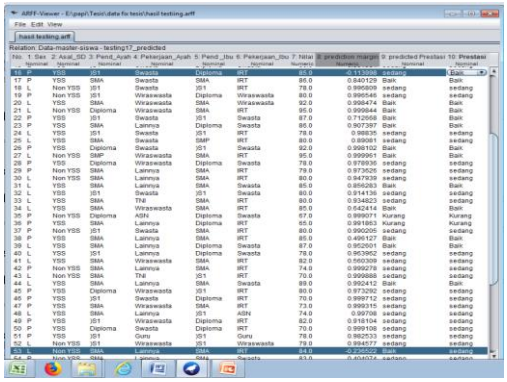
3. HASIL DAN PEMBAHASAN

Setelah melakukan pengujian/testing menggunakan file Data-master-siswa-testing17 menggunakan metoda naïve bayes, maka terlihat sebanyak 2 instance/record atau sebesar 2.8% yang salah prediksi. Hasil visualize classiffier errors dapat dilihat pada gambar 4. (instances/record yang salah prediksi diberi hightlight)



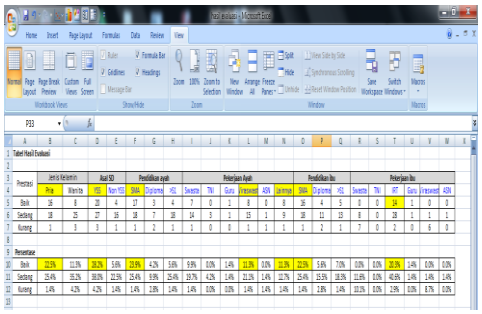
Gambar 4. Visualisasi error data testing

Dari data Visualisasi error data testing maka dilakukan penyimpanan visualisasi tersebut ke file ARFF, ini dimaksudkan agar mudah dalam melihat melakukan analisa variabel/atribut yang berpengaruh terhadap prestasi siswa.Hasil konversi file ARFF ke file excel dapat terlihat pada gambar 5.



Gambar 5. Hasil visualisasi data

Dari data Visualisasi error data testing file ARFF tersebut maka dilakukan penyimpanan visualisasi dari file ARFF ke file excel, ini dimaksudkan agar mudah dalam melakukan analisa variabel/atribut yang berpengaruh terhadap prestasi siswa.Hasil konversi file ARFF ke file excel



dapat terlihat pada gambar 6.
Gambar 6. Hasil data konversi

ANALISA DATA

Atribut/variabel yang berpengaruh terhadap prediksi prestasi siswa berdasarkan metoda naïve bayes untuk masing-masing nilai adalah:

1. Prestasi Baik, mempunyai kecenderungan dimiliki oleh atribut yang memiliki persentase yang tinggi dari masing-masing atribut/variabel, yaitu siswa berjenis kelamin laki-laki (22,5%), yang berasal dari Sekolah Dasar lingkungan Yayasan Salib Suci (YSS) sebanyak 28,2%, Pendidikan ayah adalah SMA (23,9%), pekerjaan ayah adalah wiraswasta (11,3%), pendidikan ibu SMA (22,5%) dan pekerjaan ibu adalah Ibu Rumah Tangga (20,3)%.
   Untuk atribut jenis kelamin mempunyai peluang berprestasi baik adalah laki-laki, dan yang berasal dari Yayasan Salib Suci).
2. Prestasi Sedang, mempunyai kecenderungan dimiliki oleh atribut yang memiliki persentase yang tinggi dari masing-masing atribut/variabel, yaitu siswa berjenis kelamin perempuan (35,2%) yang berasal dari Sekolah Dasar lingkungan Yayasan Salib Suci (YSS) sebanyak 38%, Pendidikan ayah adalah SMA dan >S1 mempunyai presentase yang sama sebesar (25,4%), pekerjaan ayah adalah wiraswasta (21,1%), pendidikan ibu SMA (25,4%) dan pekerjaan ibu adalah Ibu Rumah Tangga (40,6%).
   Untuk atribut jenis kelamin perempuan dan pekerjaan ibu sebagai ibu rumah

p-ISSN :1858-3911* , *e-ISSN : 2614-5405
https://journal.fkom.uniku.ac.id/ilkom

tangga mempunyai faktor dominan dalam prestasi siswa yang sedang.

3. Prestasi Kurang, mempunyai kecenderungan dimiliki oleh atribut yang memiliki persentase yang tinggi dari masing-masing atribut/variabel, yaitu siswa berjenis kelamin perempuan (4,2%) yang berasal dari Sekolah Dasar lingkungan Yayasan Salib Suci (YSS) sebanyak 4,2%, Pendidikan ayah adalah diploma mempunyai presentase sebesar (2,9%), hampir pekerjaan ayah mempunyai persentase sebesar 1,4%, pendidikan ibu Diploma (2,9%) dan pekerjaan ibu adalah swasta (10,1)%
Untuk atribut pekerjaan ibu sebagai swasta dan pendidkan ibu yang diploma mempunyai faktor dominan dalam prestasi siswa kurang.

4. KESIMPULAN

Untuk pengujian klasifikasi hasil belajar siswa pada penelitian ini dapat disimpulkan:

a. Metoda Naïve bayes memiliki kinerja yang cukup baik dari segi akurasi, yaitu sebesar 97,2 %
b. Atribut-atribut yang digunakan dalam memprediksi tingkat keberhasilan prestasi siswa cukup mencerminkan prediksi hasil.
C. Untuk siswa berjenis kelamin laki-laki, pendidikan orang tua SMA cukup mempengaruhi prediksi prestasi siswa yang bernilai baik.

5. SARAN

Saran yang dapat dikemukakan adalah penggunaan data yang lebih besar lagi agar hasilnya dapat lebih valid, dan juga dapat menggunakan metoda yang berbeda apakah menghasilkan suatu keimpulan yang relevan dengan metoda sebelumnya

DAFTAR PUSTAKA

[1] Angga Raditya, Implementasi Data Mining Classification untuk Menacrai Pola Prediksi Hujan dengan Menggunakan Algoritma C4.5, Jurusan Teknik Informatika, Fakultas Teknologi Industri, Universitas Gunadarma

[2] Ariana Azimah, Yudho Giri Sucahyo, Penggunaan Data Warehouse Dan Data Mining Untuk Data Akademik Sebuah Studi Kasus Pada Universitas Nasional, 2007

[3] Daniel T, Larose, 2005. “Discovering Knowledge in Data: An Introduction to Data Mining”. John Wiley & Sons, Inc

[4] Djamarah, Syaiful Bahri. 1994. Prestasi Belajar dan kompetensi Guru. Surabaya: Usaha Nasional.

[5] Depdiknas. (2003). Undang-Undang Republik Indonesia Nomor 20 Tahun 2003, ten¬tang Sistem Pendidikan Nasional

[6] Depdiknas. (1990). Peraturan Pemerintah RI No. 29, Tahun 1990, tentang Pendidikan Menengah

[7] Han, J. And Kamber, M, 2011, “Data Mining Concept and Techniques Second Edition ”. Morgan Kauffman, San Francisco.

[8] Iin Ernawati, 2008, “Prediksi Status Keaktifan Studi Mahasiswa dengan Algoritma C5.0 dan K-Nearest Neighbor”, Institut Pertanian Bogor

[9] Kass G.V. (1980). An exploratory technique for investigating large quantities of cat¬egorical data. Appl. Statist. 29 No.2. pp 119-127

[10] Kurniawan, Deny. (2008). Regresi linier (lin¬ear regression). Vienna, Austria: R Foundation for Statistical Computing

[11] Larose, & Daniel T. (2005). Discovering knowledge in data: an introduction to data mining. USA: John Wiley and Sons

[12] Lior Rokach, & Oded Maimon. (2005). Data mining with decision tree. World Sci¬entific Publishing Co. Pte. Ltd. Series in Machine Perception Artificial Intel¬ligence Volume 69

p-ISSN :1858-3911* , *e-ISSN : 2614-5405
https://journal.fkom.uniku.ac.id/ilkom

[13] Nurkencana. 2005. Evaluasi Hasil Belajar Mengajar. Surabaya: Usaha Nasional.

[14] Rainardi, Vincent, 2008, "Building Data Warehouse with Examples in SQL Server", Springer, New York.

[15] Slameto. 2003. Belajar dan Faktor-Faktor yang Mempengaruhinya. Jakarta: Rineka Cipta.

[16] Tan S, Kumar P, Steinbach M. 2005. "Introduction To Data Mining". Addison Wesley

[17] Tu'u,Tulus.2004.Peran Disiplin pada Perilaku dan Prestasi Siswa.Jakarta:Rineka Cipta.

[18] Sudjana, Nana. 1989. Cara Belajar Siswa Aktif-Dalam Proses Belajar Mengajar. Bandung:Sinar Baru.

[19] Tulus. (2004). Peran disiplin pada perilaku dan prestasi siswa. Jakarta: Grasindo

[20] Umaedi. (2001). Manajemen peningkatan mutu berbasis sekolah. Jakarta: Departemen Pendidikan Nasional Direktorat Jendral Pendidikan Dasar dan Menengah Direk¬torat Sekolah Lanjutan Tingkat Pertama

[21] Xin Yan, & Xiao Gang Su. (2009). Linear re¬gression analysis. London: World Sci¬entific Publishing Co. Pte. Ltd., Covent Garden

[22] https://www.gurupendidikan.co.id/pengertian-prestasi-menurut-para-ahli-beserta-macamnya/

-