

IMPLEMENTASI DATA MINING UNTUK KLASIFIKASI KELULUSAN MENGUNAKAN METODE *NAÏVE BAYES* BERBASIS WEB PADA SISWA SMKS YAJ DEPOK

Tito Ponda^{1*}, Dewi Kusumaningsih²

^{1,2} Teknik Informatika, Fakultas Teknologi Informasi, Universitas Budi Luhur, Jakarta, Indonesia

Email: ¹*2011501554@student.budiluhur.ac.id, ²dewi.kusumaningsih@budiluhur.ac.id
(* : corresponding author)

Abstrak- Siswa SMKS YAJ Depok sering mengalami kesulitan dalam melanjutkan pendidikan ke jenjang berikutnya akibat kendala pada nilai yang menjadi persyaratan. Hal ini menyebabkan banyak siswa yang ditolak saat mendaftar. Untuk mengatasi masalah ini, dikembangkan aplikasi yang mengklasifikasikan nilai siswa agar mereka memiliki peluang lebih besar untuk diterima. Penelitian ini memanfaatkan data kelulusan siswa yang diolah menggunakan teknik data mining untuk menghasilkan klasifikasi kelulusan. Metode yang digunakan adalah Naïve Bayes. Atribut yang digunakan dalam prediksi kelulusan meliputi nilai agama, nilai kewarganegaraan, nilai matematika, nilai kejuruan dan prestasi. Data yang digunakan dalam pengujian ini diambil dari arsip nilai sekolah dengan total 264 data. Hasil pengujian dievaluasi menggunakan confusion matrix dan holdout validation. Berdasarkan hasil pengujian, sistem ini mencapai akurasi sebesar 92%, recall sebesar 84,69%, dan presisi sebesar 67,76%.

Kata Kunci: *Naive bayes, Data Mining, Tingkat Kelulusan*

IMPLEMENTATION OF DATA MINING FOR GRADUATION CLASSIFICATION USING THE WEB-BASED *NAÏVE BAYES* METHOD IN SMKS YAJ DEPOK STUDENTS

Abstract- Students of SMKS YAJ Depok often experience difficulties in continuing their education to the next level due to constraints on the grades that are required. This causes many students to be rejected when applying. To address this problem, an app was developed that classifies students' grades so that. This study utilizes student graduation data processed using data mining techniques to produce graduation classifications. The method used is Naïve Bayes. The attributes used in graduation prediction include religious values, citizenship values, mathematics scores, vocational values and achievements. was taken grade archive with a total of 264 data. Based on the test results, this system achieved an accuracy of 92%, recall of 84.69%, and precision of 67.76%.

Keywords: *Naive bayes, Data Mining, Graduation Rate*

1. PENDAHULUAN

Kemajuan teknologi informasi telah mengubah berbagai aspek kehidupan, termasuk bidang pendidikan. Informasi yang akurat dan cepat sangat diperlukan untuk mendukung proses pengambilan keputusan yang tepat. Namun, sering kali informasi tersebut tersembunyi dalam tumpukan data yang besar dan kompleks. Di bidang pendidikan, teknologi informasi sangat dibutuhkan untuk mengoptimalkan proses belajar mengajar dan evaluasi hasil belajar siswa. Salah satu yang bisa digunakan adalah penggunaan data mining, yang mampu menggali informasi berharga dari data yang besar. Pertumbuhan yang begitu pesat dari akumulasi data sudah membuat situasi kaya akan data namun minim informasi [1].

Dalam dunia pendidikan, standarisasi kelulusan siswa merupakan indikator penting untuk mengevaluasi keberhasilan proses pendidikan. Standarisasi ini umumnya didasarkan pada nilai Ujian Sekolah (US) sekolah. Proses penilaian kelulusan yang manual dan subjektif sering kali menimbulkan masalah dalam pengambilan keputusan yang adil dan objektif. Maka dari itu, dibutuhkan sistem yang bisa memberikan bantuan pada sekolah untuk menganalisis data siswa dan memprediksi kelulusan mereka dengan lebih akurat dan efisien [2].

Penelitian ini mempunyai tujuan dalam menjadikan sistem berbasis web yang memakai metode Naive bayes untuk klasifikasi kelulusan siswa SMKS YAJ Depok. Metode Naive bayes ialah teknik klasifikasi dalam data mining yang memakai prinsip probabilitas. Dengan menggunakan data historis siswa, yaitu nilai Ujian Sekolah (US).

Penelitian terdahulu yang sudah dilaksanakan sama [3] bertujuan untuk memprediksi kelayakan tenaga kerja Indonesia menggunakan data uji dari 17 orang yang dibagi kedalam dua kelas. Kelas pertama adalah "Layak" dengan 10 orang, dan kelas kedua adalah "Tidak Layak" dengan 7 orang. Pengujian data dilakukan menggunakan Rapidminer 5.3, yang berhasil menghasilkan dua kelas dari hasil klasifikasi atas tingkat keakuratan 90,00%. Precision untuk kelas "Layak" mencapai 90,91%, sementara untuk kelas "Tidak Layak" mencapai 100%.

Berbeda dengan penelitian sebelumnya. Pada penelitian ini yang dengan judul "Implementasi Data Mining Untuk Klasifikasi Kelulusan memakai Metode Naive bayes Berbasis Web Pada Siswa SMKS YAJ Depok" dimana tidak menggunakan Rapidminer, tetapi menggunakan sistem berbasis web. Sistem ini dapat digunakan langsung dan memiliki keunggulan dengan adanya autentikasi melalui email dan password, hingga tidak seluruh orang bisa mengakses data yang ada dalam program. Meskipun demikian, program ini juga memiliki kelemahan, yaitu semakin besar ukuran file .xls, semakin lama waktu pemrosesannya .

2. METODE PENELITIAN

2.1 Data Mining

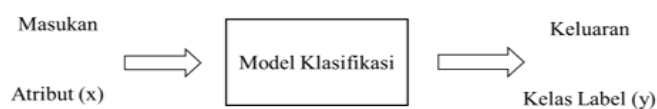
Data mining ialah tahapan mengekstraksi informasi penting dari data yang kompleks yang melibatkan penggunaan kecerdasan buatan, teknik statistik, ilmu matematika, pembelajaran mesin, dan metode lainnya. Tujuannya adalah untuk membantu dalam pengambilan keputusan, prediksi, dan pengelompokan data ke dalam kategori yang relevan.[4].

2.2 Klasifikasi

Klasifikasi ialah tahapan dalam data mining yang bertujuan dalam mengklasifikasikan data pada kategori ataupun kelas yang sudah ditetapkan sebelumnya. Tahapan ini melibatkan pembuatan model atau algoritma yang bisa memprediksi kelas dari data baru menurut data yang telah dilabeli sebelumnya. Klasifikasi digunakan untuk mengenali pola serta hubungan pada data sehingga bisa dipakai dalam membuat keputusan [5].

Klasifikasi merupakan proses di mana sebuah sistem memprediksi label atau kelas yang sesuai untuk suatu data berdasarkan informasi yang ada. Tujuan utama dari klasifikasi adalah untuk memprediksi kategori atau kelas dari data baru berdasarkan pola yang telah dikenali dari data latih yang ada. [6]

Klasifikasi bisa dibuat dengan berbagai metode, baik dengan cara manual oleh manusia maupun sama bantuan teknologi, seperti menggunakan algoritma machine learning. Secara manual, klasifikasi dilakukan dengan cara mengidentifikasi karakteristik atau pola tertentu dalam data dan mengambil keputusan berdasarkan pengalaman atau pengetahuan yang dimiliki. Sementara itu, dengan bantuan teknologi, klasifikasi sering kali menggunakan pendekatan yang lebih terstruktur dan otomatis. Ini dapat mencakup penggunaan algoritma seperti Naive Bayes, Random Forests, dan lainnya. Algoritma-algoritma ini mampu belajar dari data yang ada dan menggunakan informasi tersebut untuk membuat klasifikasi atau prediksi terhadap data baru[7].



Gambar 1. Tahapan data mining

2.3 Metode Naive Bayes

Naive bayes ialah diantara algoritma klasifikasi yang sangat populer pada bidang pembelajaran mesin serta data mining. Naive bayes ialah klasifikasi kemungkinan sederhana yang bisa menilai semua kemungkinan lewat penggabungan sejumlah kombinasi serta frekuensi suatu nilai dari basis data yang diperoleh [7]. Pada Teorema Bayes, jika ada 2 peristiwa yang terpisah (misalkan X dan H), untuk itu Teorema Bayes dirumuskan antara lain:

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

Keterangan:

X : Data atas parameter yang tidak diketahui

H : Hipotesis keputusan

P(H|X) : Probabilitas keputusan H menurut kondisi parameter X

- $P(H)$: Probabilitas total keputusan H
 $P(X|H)$: Probabilitas parameter X berdasarkan kondisi keputusan H
 $P(X)$: Probabilitas Parameter (X).

2.4 Confusion Matrix

Confusion matrix dipakai dalam menilai performa model klasifikasi lewat metode perbandingan hasil prediksi model dengan nilai aktual dari data uji, serta mengidentifikasi true positive, true negative, false positive, serta false negative.

2.5 Holdout Validation

Holdout Validation adalah sebuah metode yang bertujuan agar setiap data yang ada memiliki peluang untuk menjadi data latih dan data uji. Holdout validation dilakukan dengan membagi data yang didapatkan jadi data latih serta data uji serta perbandingannya [9].

2.6 Dataset

Dataset adalah kumpulan data yang digunakan dalam analisis atau pembelajaran mesin [10]. Dataset dapat terdiri dari berbagai jenis data seperti teks, angka, gambar, atau suara, tergantung pada tujuan dan domain aplikasi. Dataset biasanya terbagi menjadi 2 bagian utama: data latih serta data uji.

a. Komponen Dataset

- 1) Fitur (*Attributes/Features*): Kolom dalam dataset yang mewakili karakteristik atau variabel dari data.
- 2) Label (*Class/Target*): Kolom dalam dataset yang menunjukkan kategori atau nilai yang ingin diprediksi. Dalam kasus klasifikasi, label adalah kategori; dalam kasus regresi, label adalah nilai numerik.

b. Pembagian Dataset

- 1) Data Latih (*Training Data*) dipakai dalam melatih model pembelajaran mesin
- 2) Data Uji (*Testing Data*) dipakai dalam menguji kinerja model yang sudah dilatih.

c. Jenis Dataset Berdasarkan Sumbernya

1) Private Dataset

Private dataset adalah kumpulan data yang memiliki informasi sensitif dan bersifat rahasia. Contoh data dari private dataset meliputi data bank, data rumah sakit, data sekolah, dan data Perusahaan. Private dataset memerlukan izin untuk digunakan sebagai objek penelitian karena sifatnya yang rahasia. Hanya penulis atau pemilik dataset yang dapat memodifikasinya, berbeda dengan public dataset yang lebih terbuka untuk modifikasi oleh siapa saja.

2) Publik Dataset

Public dataset adalah dataset yang dapat diakses secara bebas dari repository publik yang disepakati oleh pakar peneliti data mining. Keunggulan public dataset meliputi akses mudah dataset dapat di unduh dan digunakan oleh siapa saja tanpa memerlukan izin khusus, dan Transparansi Tinggi Penelitian yang menggunakan public dataset dapat diverifikasi ulang dan dibandingkan kembali oleh peneliti lain.

d. Jenis Dataset Berdasarkan Tipe Datanya

1) Dataset Numerik

Dataset numerik mencakup atas data yang berbentuk angka serta bisa diukur. Tipe data ini sering digunakan dalam analisis statistik dan matematis. Contoh dataset numerik meliputi pendapatan tahunan, usia, skor ujian, pengukuran fisik (berat, tinggi, suhu).

2) Dataset kategorikal

Dataset kategorikal terdiri dari data yang dibagi ke dalam kategori atau kelompok yang berbeda tanpa urutan yang melekat.

3) Dataset Ordinal

Dataset ordinal terdiri dari data kategorikal yang mempunyai urutan ataupun tingkatan. Contoh dataset ordinal meliputi peringkat (pertama, kedua, ketiga..

4) Dataset Biner

Dataset biner terdiri dari data yang memiliki dua kemungkinan nilai. Contoh dataset biner meliputi jawaban ya/tidak, lulus/gagal, aktif/nonaktif.

5) Dataset Text

Dataset teks terdiri dari data yang berupa teks atau kalimat. Dataset ini kerap dipakai pada pemrosesan bahasa alami (NLP). Contoh dataset teks meliputi Artikel berita, komentar media sosial, review produk

6) Dataset Waktu

Dataset waktu terdiri dari data yang diambil atau diukur pada berbagai titik waktu. Contoh dataset waktu meliputi Data cuaca harian, Harga saham harian, Pengukuran sensor periodic.

7) Dataset Gambar

Dataset gambar terdiri dari data visual dalam bentuk gambar atau foto. Dataset ini sering digunakan dalam bidang pengenalan gambar dan visi komputer. Contoh dataset gambar meliputi Dataset wajah manusia, Dataset objek (mobil, hewan, bangunan), Dataset medis (gambar MRI, CT scan)

8) Dataset Audio

Dataset audio terdiri dari data suara atau rekaman audio. Dataset ini sering digunakan dalam analisis suara dan pemrosesan sinyal audio. Contoh dataset audio meliputi Rekaman percakapan, music, suara lingkungan.

9) Dataset Video

Dataset video terdiri dari serangkaian gambar yang diambil pada interval waktu tertentu, membentuk video. Dataset ini digunakan dalam analisis video dan pengenalan aktivitas. Contoh dataset video meliputi rekaman keamanan, klip video olahraga, video tutorial

2.7 Penerapan Metode

a. Pengumpulan Data

1) Wawancara

Metode ini dikerjakan melalui wawancara sama kepala sekolah serta guru di SMKS YAJ Depok untuk mengumpulkan data dan mengenai nilai siswa. Data yang diperoleh akan dijadikan referensi untuk membuat sistem ini.

2) Studi Kepustakaan

Metode ini dikerjakan lewat metode mencari referensi skripsi maupun jurnal yang ada di internet maupun perpustakaan Universitas Budi Luhur.

b. Transformasi Data

Data kriteria yang telah diseleksi yaitu Nilai Agama, Bahasa Indonesia, Kewarganegaraan, Matematika, Nilai Kejuruan kemudian akan ditransformasikan menjadi kategorikal. Berikut skala untuk setiap kriteria.

Tabel 1. Transformasi Nilai Menjadi Kategorikal

Skala Nilai	Kategorikal
81-100	Tinggi sekali
61-80	Tinggi
31-60	Sedang
0-30	Rendah
Skala Prestasi	Kategorikal
Tingkat Nasional	Tinggi sekali
Tingkat Provinsi	Tinggi
Tingkat Kota	Sedang
Tingkat Sekolah	Rendah

c. Pembagian Data

Pada tahap ini atribut yang telah diberi label akan dibagi jadi 2 set data, ialah 80% Data Latih serta 20% Data Uji.

a. Data Latih berfungsi sebagai model pelatihan yang dipakai untuk membangun proses klasifikasi.

b. Data Uji digunakan untuk mengukur akurasi hasil klasifikasi menurut model pelatihan.

d. Rancangan Pengujian

1) Metode Pengujian

Untuk menguji program ini, maka akan dilakukan simulasi. Berdasarkan pola yang telah ditentukan dari data latih, kemudian akan dihitung nilai probabilitas setiap atribut dari data uji. Klasifikasi siswa akan ditentukan berdasarkan nilai probabilitas terbesar akan dihitung nilai

probabilitas setiap atribut dari data uji. Klasifikasi siswa akan ditentukan berdasarkan nilai probabilitas terbesar.

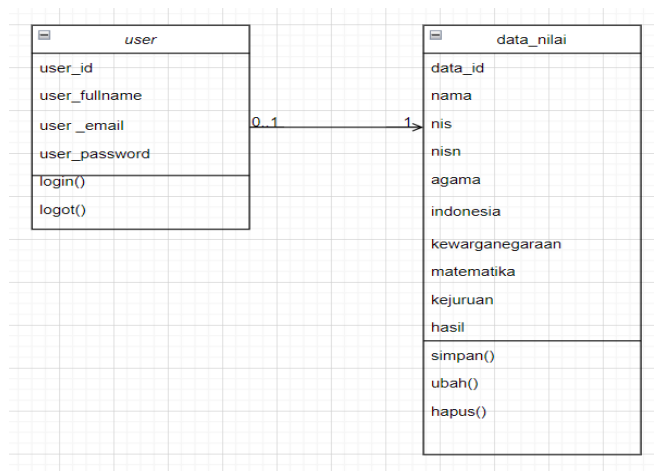
2) Labeling

Terdapat 2 label yaitu lulus dan tidak lulus yang akan diterapkan secara manual di dataset yang tersedia.

e. Rancangan Basis Data

Rancangan basis data ialah tahapan yang melibatkan penentuan konten, struktur, serta pengorganisasian data untuk membuat model basis data yang mendorong banyak kebutuhan sistem. Tujuannya adalah memastikan bahwa data tersimpan dengan cara yang efisien dan dapat diakses serta dikelola dengan mudah sesuai dengan persyaratan sistem yang dirancang.

Diagram yang berupa kelas - kelas memiliki hubungan dengan yang lainnya serta terdapat atribut dan operasi. Berikut diagram yang telah dibuat.



Gambar 2. Class Diagram

3. HASIL DAN PEMBAHASAN

Pada bagian pembahasan berisikan analisis dan metode, implementasi metode dengan melakukan perhitungan, serta melakukan pengujian terkait penelitian ini.

3.1 Implementasi Metode

Data masukkan yang di input melalui import file excel sejumlah 264 data dengan atribut, nama, tempat tanggal lahir, nis, nisan, jurusan, nilai agama, nilai bahasa indonesia, nilai kewarganegaraan, nilai matematika, nilai jurusan dan prestasi siswa. Dimana data ini akan ditransformasi untuk menghilangkan tempat tanggal lahir dan jurusan, sehingga tersisa nilai dan kelas asli. Nilai itu sendiri akan diubah kategorikal rendah, sedang, tinggi, sangat tinggi. Tahap selanjutnya data hendak diubah jadi 2 bagian, 80% data training dan 20% data testing.

a. Transformasi Data

Proses transformasi adalah menggunakan data nilai menjadi kategori. Dimana disesuaikan sama atribut yang hendak dipakai pada tahapan data mining.

Tabel 2. Transformasi

Nilai Agama	Nilai Indonesia	Nilai Kewarganegaraan	Nilai Matematika	Nilai Jurusan
rendah	rendah	rendah	rendah	rendah
Sedang	Sedang	Sedang	Sedang	Sedang
Tinggi	Tinggi	Tinggi	Tinggi	Tinggi
Tinggi sekali	Tinggi sekali	Tinggi sekali	Tinggi sekali	Tinggi sekali

b. Hitung Atribut dan Probabilitas

Probabilitas kelas Asli

Berdasarkan data training sebanyak 211 data, diperoleh :

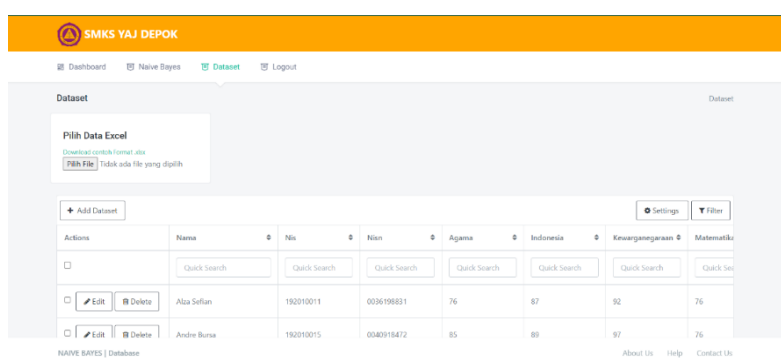
Tabel 3. Probabilitas Kelas Asli

Kelas Asli	Jumlah Kelas Asli	Probabilitas
Lulus	147	0,697
Tidak Lulus	64	0,303
Total	211	1

3.2 Tampilan Layar

a. Tampilan Layar Menu Dataset

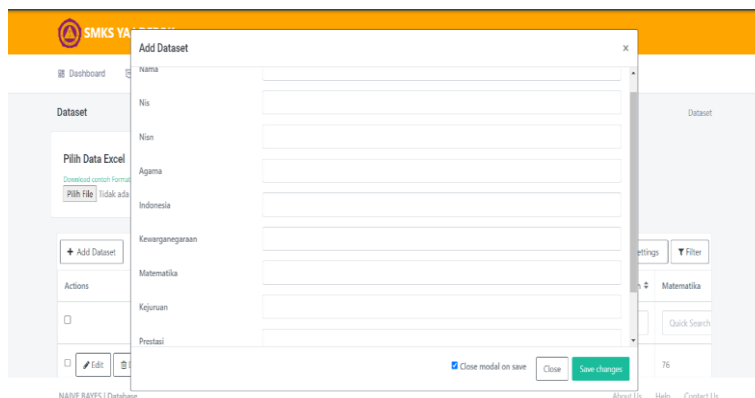
Pada Gambar 3. adalah tampilan menu dataset



Gambar 3. Tampilan Menu Dataset

b. Tampilan Layar Add

Pada Gambar 4. adalah tampilan layar add dataset

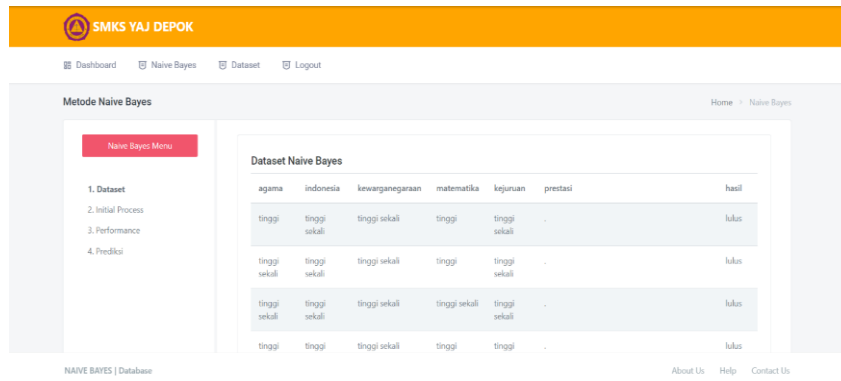


Gambar 4. Tampilan Add Dataset

Tampilan ini adalah layar add dataset, dimana dapat menambah data dan akan muncul didalam tabel menjadi satu dengan data yang telah diimport

c. Tampilan Layar Naive bayes

Pada Gambar 5. adalah tampilan layar add dataset

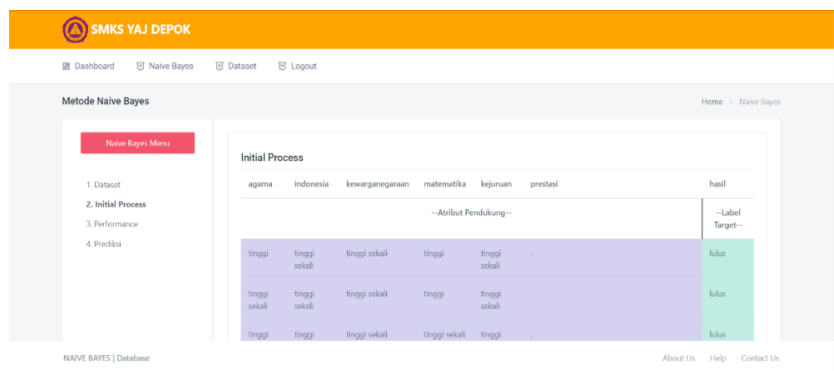


Gambar 5. Tampilan Menu *Naive Bayes*

Tampilan ini adalah menu *Naive Bayes*, dimana pengguna dihadapkan beberapa menu, yaitu dataset, initial process, performance, dan prediksi. Saat pengguna menekan menu Naive bayes, pengguna akan dihadapkan ke tampilan dataset. Tampilan dataset yang terlihat merupakan dataset yang telah di transformasi ke format yang telah ditentukan.

d. Tampilan Layar *Initial Process*

Pada Gambar 6. adalah tampilan *Initial Process*

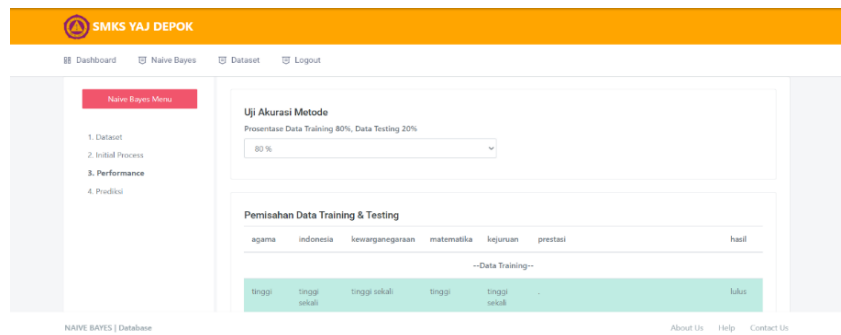


Gambar 6. Tampilan *Initial Process*

Tampilan ini adalah layar Initial Process, dimana dataset yang telah ditransformasi kemudian di pisahkan antara atribut pendukung dan label target.

e. Tampilan Layar *Performance*

Pada Gambar 7. adalah tampilan *Performance*

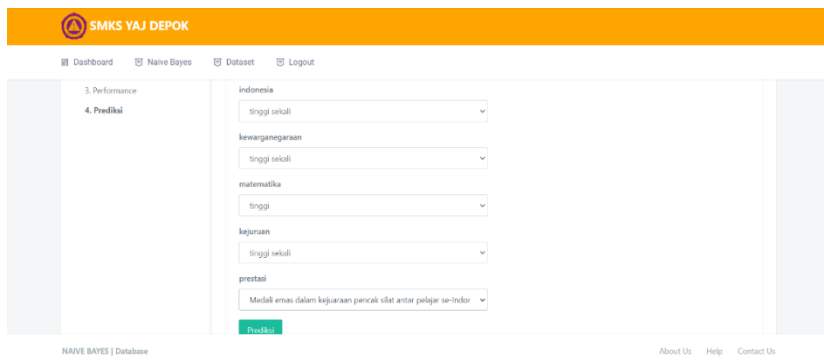


Gambar 7. Tampilan *Performance*

Tampilan ini adalah layar Performance, dimana dataset yang telah ditransformasi tadi, kemudian dipisahkan antara data training dan data testing, pemisahan ini berdasarkan persentase uji akurasi yang user pilih. Setelah pemisahan data, maka tahap selanjutnya adalah proses testing, persentase akurasi akan muncul di bagian paling bawah.

f. Tampilan Layar Prediksi

Pada Gambar 8. adalah tampilan layar prediksi



Gambar 8. Tampilan prediksi

Tampilan ini ialah layar Prediksi, dimana pengguna akan memasukkan kategori nilai yang dimiliki kedalam form yang telah ditampilkan. Saat user menekan prediksi, maka akan muncul hasil dan probabilitasnya.

4. KESIMPULAN

Berdasarkan penelitian yang sudah dikerjakan terkait masalah pada sistem yang dibuat, bisa ditarik kesimpulan dan saran yang mungkin dibutuhkan dalam pengembangan sistem ke tahap yang lebih baik di masa mendatang. Sistem akan menampilkan data yang dimasukkan oleh pengguna baik dari modal import maupun inputan. Nilai akurasi yang dicapai sebanyak 92.5%, recall sebesar 84,69%, dan presisi sebesar 67,76%.

DAFTAR PUSTAKA

- [1] M. Jannah, "Penerapan Data Mining Prediksi Nilai UN Siswa SMP Menggunakan Metode Naive Bayes," *INFORMATIKA*, vol. 12, no. 2, hal. 1, 2020, doi: 10.36723/juri.v12i2.220.
- [2] L. Lisawita dan M. Devega, "Analisis Perbandingan Algoritma Apriori Dan Algoritma Eclat Dalam Menentukan Pola Peminjaman Buku Di Perpustakaan Universitas Lancang Kuning," *INOVTEK Polbeng - Seri Inform.*, vol. 3, no. 2, hal. 118, 2018, doi: 10.35314/isi.v3i2.753.
- [3] D. Yunita, P. Rosyani, dan R. Amalia, "Analisa Prestasi Siswa Berdasarkan Kedisiplinan, Nilai Hasil Belajar, Sosial Ekonomi dan Aktivitas Organisasi Menggunakan Algoritma Naive Bayes," *J. Inform. Univ. Pamulang*, vol. 3, no. 4, hal. 209, 2018, doi: 10.32493/informatika.v3i4.2032.
- [4] J. Dongga, A. ` Sarungallo, N. Koru, dan G. Lante, "Implementasi Data Mining Menggunakan Algoritma Apriori Dalam Menentukan Persediaan Barang (Studi Kasus: Toko Swapen Jaya Manokwari)," *G-Tech J. Teknol. Terap.*, vol. 7, no. 1, hal. 119–126, 2023, doi: 10.33379/gtech.v7i1.1938.
- [5] Syarli dan A. A. Muin, "Metode Naive Bayes Untuk Prediksi Kelulusan," *J. Ilm. Ilmu Komput.*, vol. 2, no. 1, hal. 22–26, 2016, [Daring]. Tersedia pada: <https://media.neliti.com/media/publications/283828-metode-naive-bayes-untuk-prediksi-kelulu-139fcfea.pdf>.
- [6] L. Febriyanti dan H. Zakaria, "Implementasi Data Mining Untuk Memprediksi Produktivitas Pada Tanaman Kacang Tanah Menggunakan Metode Naive Bayes (Studi Kasus : Perkebunan Kacang Tanah Di Kota Bogor)," *Log. J. Ilmu Komput. dan Pendidik.*, vol. 1, no. 2, hal. 105–118, 2023, [Daring]. Tersedia pada: <https://journal.mediapublikasi.id/index.php/logic/article/view/2353>.
- [7] R. Rachman dan R. N. Handayani, "Klasifikasi Algoritma Naive Bayes Dalam Memprediksi Tingkat Kelancaran Pembayaran Sewa Teras UMKM," *J. Inform.*, vol. 8, no. 2, hal. 111–122, 2021, doi: 10.31294/ji.v8i2.10494.
- [8] N. Riyanah dan F. Fatmawati, "Penerapan Algoritma Naive Bayes Untuk Klasifikasi Penerima Bantuan Surat

- Keterangan Tidak Mampu,” *JTIM J. Teknol. Inf. dan Multimed.*, vol. 2, no. 4, hal. 206–213, 2021, doi: 10.35746/jtim.v2i4.117.
- [9] R. Rizki, R. Athallah, I. Cholissodin, dan P. P. Adikara, “Prediksi Potensi Pengidap Penyakit Diabetes berdasarkan Faktor Risiko Menggunakan Algoritme Kernel K-Nearest Neighbor,” *J. Pengemb. Teknol. Inf. dan Ilmu Komput.*, vol. 6, no. 8, hal. 3777–3785, 2022, [Daring]. Tersedia pada: <https://j-ptiik.ub.ac.id/index.php/j-ptiik/article/view/11439>.
- [10] Amrin Amrin, “Data Mining Dengan Algoritma Apriori untuk Penentuan Aturan Asosiasi Pola Pembelian Pupuk,” *Paradigma*, vol. XIX, no. 1, hal. 74–79, 2017, doi: <https://doi.org/10.31294/p.v19i1.1836>.