Jurnal Riset Informatika dan Teknologi Informasi (JRITI) ISSN (Print) : 3024-8167
Jejaring Penelitian dan Pengabdian Masyarakat (JPPM)
Vol. 2, No. 3, April – Juli 2025, Hal. 244 - 252
Web : https://ejurnal.jejaringpppm.org/index.php/jriti

Klasifikasi Penentuan Siswa Berprestasi Menggunakan Algoritma Naïve Bayes Classifier DI PT.Yes Study Education Group Indonesia

Novan Ponco Laksono[1], Achmad Akbar Syaaifullah[2], Ajif Yunizar Pratama Yusuf[3]
Program Studi Informatika, Fakultas Ilmu Komputer, Universitas Bhayangkara Jakarta Raya
Email: novan.ponco.laksono19@mhs.ubharajaya.ac.id[1], achmad.akbar.syaaifullah19@mhs.ubharajaya.ac.id[2] .,
ajif.yunizar@dsn.ubharajaya.ac.id[3]

ABSTRAKSI
PT.Yes Study Education Group Indonesia merupakan Lembaga konsultan Pendidikan luar negeri yang didirikan oleh para alumni internasional dan berpusat di Toronto Kanada, yang berpengalaman membantu ribuan siswa dari berbagai belahan dunia untuk menggapai mimpi bersekolah diluar negeri. Namun, tidaklah mudah untuk dapat bersekolah diluar negeri karena ada beberapa faktor dan dokumen yang harus dipersiapkan seperti paspor, visa dan sertifikat tes Bahasa inggris seperti Test Of English Forgein Lenguage (TOEFL) dan International English Language Testing System (IELTS) untuk mendapatkan hasil yang maksimal dibutuhkan hasil belajar yang baik, berikutnya tentu hasil belajar adalah indicator prestasi dari peserta didik sehingga dibutuhkan algoritma yang dapat menentukan prestasi siswa, tujuannya adalah sebagai alat pendukung dalam mengevaluasi proses pembelajaran, dan hasil belajar menggunakan algoritma naïve bayes classifier dengan data uji coba 200 nama siswa berserta dengan nilainya masing – masing, dengan jumlah data uji sebanyak 80 yang didapatkan. Dari perhitungan ini permodelan Gauusien NB split validation 50 : 50 , dengan hasil akurasi sebesar 73%. , scenario 2 dengan rasio 60:40 dengan hasil akurasi 75%, scenario 3 dengan rasio 70:30 dengan akurasi 76,6%, scenario 4 dengan rasio 80:20 dengan akurasi 82,2%, dengan scenario 5 dengan rasio 90 : 10, dengan akurasi 85%.

Kata Kunci: *naïve bayes classifier, klasifikasi naïve bayes, penentuan prestasi siswa.*

ABSTRACT
PT. Yes Study Education Group Indonesia is an overseas education consultancy founded by international alumni and based in Toronto, Canada, with experience helping thousands of students from various parts of the world to achieve their dream of studying abroad. However, it is not easy to study abroad because there are several factors and documents that must be prepared, such as passports, visas, and English test certificates like the Test Of English as a Foreign Language (TOEFL) and the International English Language Testing System (IELTS). To achieve optimal results, good learning outcomes are required; furthermore, of course, learning outcomes are indicators of student achievement, so an algorithm is needed to determine student performance, with the aim of serving as a supporting tool in evaluating the learning process and outcomes using the naïve bayes classifier algorithm with a trial dataset of 200 student names along with their respective scores, from which 80 test records were obtained. From these calculations, the Gaussian NB model with a 50:50 split validation yielded an accuracy of 73%, scenario 2 with a 60:40 ratio yielded 75% accuracy, scenario 3 with a 70:30 ratio yielded 76.6% accuracy, scenario 4 with an 80:20 ratio yielded 82.2% accuracy, and scenario 5 with a 90:10 ratio yielded 85% accuracy.

Keywords: *naïve bayes classifier, naïve bayes classification, determination of student achievement*

Penulis Korespondensi
Novan Ponco Laksono

Tanggal Submit : 03/07/2024
Tanggal Diterima : 17/07/2024
Tanggal Terbit : 25/07/2025

I. PENDAHULUAN

Tingginya tingkat keberhasilan siswa merupakan cerminan langsung dari kualitas dunia pendidikan yang diselenggarakan. Di era modern ini, lembaga pendidikan dituntut untuk memiliki daya saing tinggi dengan memaksimalkan kualitas dan kuantitas sumber daya manusia (SDM) yang ada. Siswa berprestasi menjadi salah satu indikator keberhasilan tersebut, namun persaingan global menuntut agar proses pembelajaran dan penilaian terus diperbaiki agar relevan dengan tuntutan zaman.

Setiap individu siswa memiliki kemampuan, minat, dan kecepatan belajar yang berbeda-beda. Ada yang cepat menguasai konsep teoretis, sementara yang lain lebih unggul pada praktik atau kreativitas. Variasi inilah yang menjadikan pendidikan sebagai tantangan tersendiri: bagaimana menyesuaikan metode dan penilaian agar dapat menjangkau dan memotivasi seluruh spektrum kemampuan siswa.

Pendidikan sendiri sejatinya adalah upaya kolektif untuk mempersiapkan generasi penerus sebagai penentu masa depan bangsa. Melalui kurikulum, metode pengajaran, serta lingkungan belajar yang kondusif, siswa diarahkan untuk mengembangkan potensi intelektual, emosional, dan sosialnya. Keberhasilan pendidikan suatu bangsa diukur bukan hanya dari angka rata-rata nilai akademik, melainkan juga dari kemampuan lulusan untuk beradaptasi, berinovasi, dan berkontribusi positif di masyarakat.

Dalam menentukan prestasi siswa, penting untuk mempertimbangkan faktor-faktor tersebut secara menyeluruh. Penilaian yang adil dan akurat harus berbasis evaluasi atas keseluruhan kemampuan—mulai dari penguasaan materi, kerjasama tim, hingga sikap dan nilai yang ditunjukkan siswa. Pendekatan ini selaras dengan tujuan pendidikan yang ingin mengembangkan pengetahuan, keterampilan, sikap, serta nilai-nilai luhur pada setiap individu.

Permasalahan yang terjadi di lapangan menunjukkan bahwa proses penetapan siswa berprestasi masih terfokus pada nilai akademik semata, tanpa mempertimbangkan capaian non-akademik seperti kreativitas, kepemimpinan, maupun keaktifan sosial. Akibatnya, banyak siswa dengan potensi unggul di bidang lain terabaikan, dan standar penilaian menjadi kurang inklusif serta kurang mencerminkan keberagaman bakat peserta didik.

Mengingat hal tersebut, para pendidik perlu mengambil pendekatan yang lebih holistik dan komprehensif dalam menetapkan prestasi siswa. Model penilaian terpadu—misalnya dengan memadukan portofolio, observasi kompetensi, serta indikator RFM (Recency, Frequency, Monetary) atau metode klasifikasi berbasis data—dapat membantu menghasilkan gambaran yang lebih utuh tentang perkembangan setiap siswa. Dengan demikian, penentuan prestasi tidak hanya adil, tetapi juga mampu memotivasi semua siswa untuk berkembang sesuai kekuatan dan minat masing-masing.

II. LANDASAN TEORI

1. Penelitian terkait

Hasil penelitian data mining yang telah dilakukan dengan menerapakan algoritma naïve bayes classifier diperoleh dengan mengambil data mahasiswa secara acak mulai dari Angkatan 2010 sampai dengan Angkatan 2012 yang telah dinyatakan lulus. Kemudian dilakukan tahap cleaning data, data selection dan penentuan data yang akan menjadi data training, pengolahan data, implementasi algoritma naïve bayes dari hasi evaluasi yang ditunjukkan dari data training dan data uji prestasi mahasiswa menghasilkan 70 : 30 yang berarti kelulusan tepat dan tidak tepat menunjukkan akurasi sebesae 66,6%[2], peneletian juga dilakukan untuk mengelompokkan siswa berprestasi, berdasarkan evaluasi dan validasi hasil indeks daviesbouldin menggunakan dataset sebanyak 414, dapat disimpulkan bahwa metode Clustering K-means memiliki kinerja yang cukup baik. hasil pengelompokkan pada Microsoft Excel dan RapidMinner memperoleh hasil yang sama, yakni sebanyak 107 siswa termasuk kedalam siswa kurang berprestasi, 51 siswa termasuk kedalam siswa berprestasi, dan sebanyak 256 siswa termasuk kedalam siswa berpotensi berprestasi[3]. Dan pada penelitian ini berbasis Naïve Bayes sebagai classifier, sehingga setiap parameter dianggap sama pentingnya. Dari penelitian yang telah dilakukan, hasil analisa menunjukan bahwa model yang diusulkan memiliki tingkat akurasi sebesar 77,5%, dan hasil yang lebih rendah sebesar 69% bila tidak menggunakan outlier detection[2].

2. Naïve Bayes

Pengklasifikasi Bayesian adalah pengklasifikasi statistik yang dapat memprediksi probabilitas bahwa tupel tertentu milik kelas tertentu. Naïve Bayes Classifier menunjukkan akurasi dan kecepatan tinggi saat digunakan dengan database besar.

Algoritme Naïve Bayes Classifier memiliki beberapa keunggulan, termasuk mampu tampil lebih baik dalam kasus dunia nyata yang kompleks dan tidak harus memiliki data latih dalam jumlah besar untuk menentukan parameter atau pola selama klasifikasi [4]

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

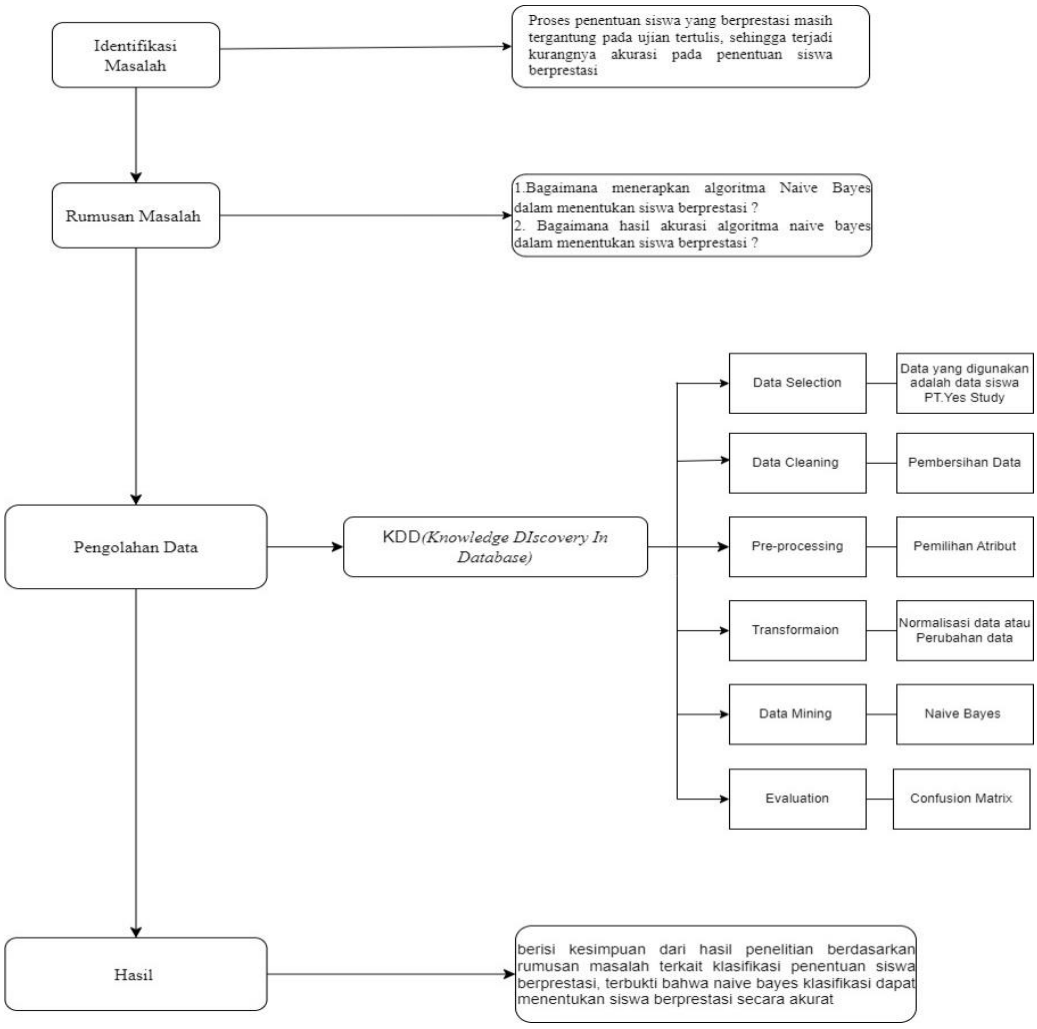
Keterangan :

X : Data dengan class yang belum diketahui
H : Hipotesis data merupakan suatu class spesifik
P(H|X) : Probabilitas hipotesis terhadap H berdasar kondisi X (posteriori probabilitas)
P(H) : Probabilitas hipotesis terhadap H (prior probabilitas)
P(X|H) : Probabilitas X berdasarkan kondisi pada hipotesis H P(X) : Probabilitas X

III. METODOLOGI

1. Kerangka penelitian

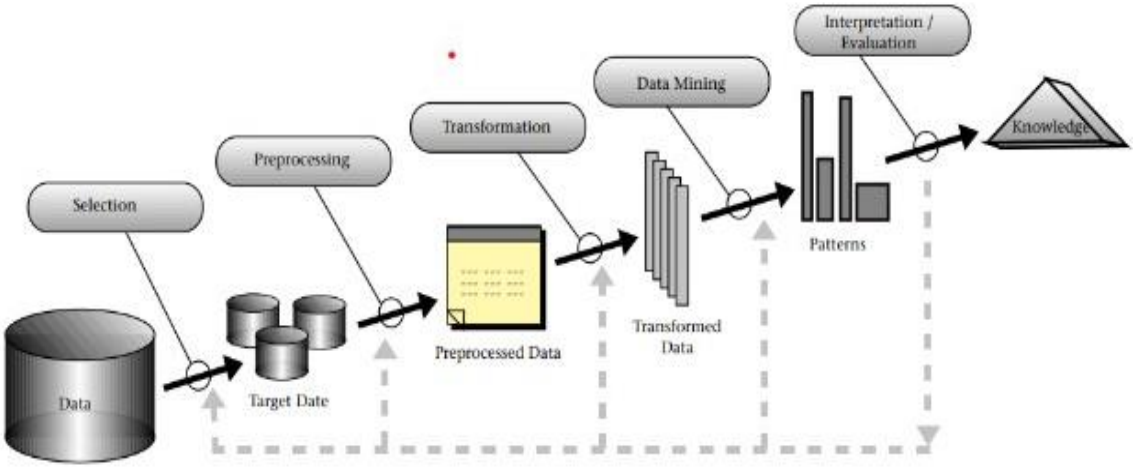
Dalam penelitian ini, peneliti membuat kerangka penelitian yang berguna sebagai dasar pemikiran dalam klasifikasi penentuan siswa berprestasi menggunakan algoritma naïve bayes, berikut ini merupakan gambar kerangka penelitian, yaitu sebagai berikut :



Gambar 1. Kerangka Penelitian

2. Knowledge Discovery In Database (KDD)

Knowledge Discovery in Databases (KDD) merupakan sekumpulan proses untuk menemukan pengetahuan yang bermanfaat dari data (lihat Gambar 2). KDD terdiri dari serangkaian langkah perubahan, termasuk data preprocessing dan juga post processing. Data propecessing merupakan langkah untuk mengubah data mentah menjadi format yang sesuai untuk tahap analisis berikutnya. Selain itu data preprocessing juga digunakan untuk membantu dalam pengenalan atribut dan data segmen yang relevan dengan task data mining. Istilah Data mining dan Knowledge Discovery in Databases (KDD) sering kali digunakan secara bergantian untuk menjelaskanproses penggalian informasi tersembunyi dalam suatu basis data yang besar. Sebenarnya kedua istilah tersebut memiliki konsep yang berbeda, tetapi berkaitan satu sama lain. Dan salah satu tahapan dalam keseluruhan proses KDD adalah Data mining [5]. Proses Knowledge Discovery in Databases (KDD) secara garis besar dapat dijelaskan sebagai berikut.[4] :



Gambar 2. Proses KDD (Knowledge Discovery In Database)

1) Data Selection

Data Selection Pemilihan atau seleksi data dari sekumpulan data operasional perlu dilakukan sebelum tahap penggalian informasi dalam knowledge data discovery dimulai. Data hasil seleksi yang akan digunakan untuk proses Data mining, disimpan dalam suatu berkas, terpisah dari basis data operasional Sebelum proses Data mining dapat dilaksanakan, perlu dilakukan proses cleaning pada data yang menjadi fokus KDD. Proses cleaning mencakup antara lain membuang duplikasi data, memeriksa data yang inkosisten, dan memperbaiki kesalahan pada data, seperti kesalahan cetak (tipografi).

2) Data Pre-processing

Pra-pemprosesan data (preprocessing data) merupakan langkah kritis dalam melakukan analisis klasifikasi, yang bertujuan untuk membersihkan data dari elemen-elemen yang tidak diperlukan guna mempercepat proses klasifikasi.

3) Data Tranformation

Transformation adalah proses tranformasi pada data yang telah dipilih, sehingga data tersebut sesuai untuk proses data mining. Proses coding dalam KDD merupakan proses kreatif dan sangat tergantung pada jenis atau pola informasi yang akan dicari dalam basis data.

4) Data Mining

Data Mining adalah proses mencari pola atau informasi menarik dalam data terpilih dengan menggunakan Teknik atau metode tertentu. Teknik, metode, atau algoritma dalam data mining sangat bervariasi. Pemilihan metode atau algoritma yang tepat sangat bergantung pada tujuan dan proses KDD secara keseluruhan

5) Naïve Bayes Classifier

Pengklasifikasi Bayesian adalah pengklasifikasi statistik yang dapat memprediksi probabilitas bahwa tupel tertentu milik kelas tertentu. Naïve Bayes Classifier menunjukkan akurasi dan kecepatan tinggi saat digunakan dengan database besar. Algoritme Naïve Bayes Classifier memiliki beberapa keunggulan, termasuk mampu tampil lebih baik dalam kasus dunia nyata yang kompleks dan tidak harus memiliki data latih dalam jumlah besar untuk menentukan parameter atau pola selama klasifikasi [4].

6) Interpretation/ Evaluation

Interpretation/ Evalution pola infromasi yang dihasilkan dari proses data data mining yang perlu ditampilkan dalam bentuk yang mudah dimengerti oleh pihak yang berkepentingan. Tahap ini merupakan bagian dari proses KDD yang disebut interpretation. Tahap ini mencakup pemerikasaan apakah pola atau infomasi yang ditemukan bertentangan dengan fakta atau hipotesis yang ada sebelumnnya

3. Confusion Matrix

Confusion matrix menunjukkan hubungan antara nilai aktual dan nilai prediksi. Tiga faktor penting yang mempengaruhi kinerja suatu algoritme dan hasilnya adalah acuracy, precision, dan recall. Acuracy merupakan tingkat ketepatan suatu model dalam melakukan klasifikasi data dengan benar. Precision merupakan tingkat ketepatan hasil prediksi benar yang diinginkan oleh pengguna dengan hasil prediksi yang diberikan oleh suatu model. Recall merupakan tingkat ketepatan suatu model dalam memprediksi data kelas posotif berdasakna keseluruhan data dengan nilai aktual positif. Untuk menguji akurasi atau mengevaluasi algoritme pengklasifikasi Naïve Bayes Classifier [6], digunakan confusion matrix, yaitu tes cari tahu sejauh mana klasifikasi berlaku untuk kelas yang berbeda. Confusion matrix dapat dilihat pada dilihat sebagai berikut :

Keterangan:

1. True Positive (TP): yaitu jumlah data dengan kelas positif yang diklasifikasikan positif.
2. True Negative (TN): yaitu jumlah data dengan kelas negative yang diklasifikasikan negatif.
3. False Positive (FP): yaitu jumlah data dengan kelas positif yang diklasifikasikan negatif.
4. False Negative (FN): yaitu jumlah data dengan kelas negatif yang diklasifikasikan positif.

Ukuran besaran accuracy, precision, biasanya diberi nilai dalam bentuk presentase antara 1 sampai 100%. Sebuah sistem akan dianggap baik jika tingkat precision, dan accuracynya tinggi. Berikut adalah persamaan model confusion matrix:

1) Accuracy, ukuran seberapa baik klasifikasi dibuat, dinyatakan sebagai persentase dari semua kemungkinan data yang berhasil diklasifikasikan.
2) Precision, menentukan seberapa baik sistem mencocokkan kueri pengguna dengan data yang diambil dan dikembalikannya.
3) Recall adalah proporsi dari setiap informasi yang akan ditemukan dari label [7].
4) Precision dan Recall bisa digunakan untuk mendapatkan proporsi pengukuran lain yaitu F1-Score. Sedangkan F1-Score merupakan harmonic mean untuk perhitungan Precision dan Recall [7].

4. Python

Python merupakan bahasa pemrograman interpretatif multiguna dengan filosofi perancangan yang masih berfokus dengan tingkat keterbacaan kode. Python bisa diklaim sebagai bahasa penggabungkan kapabilitas, kemampuan, yang sintaksis kode yang sangat jelas, dan dilengkapi dengan fungsionalitas pustaka standar yang besar serta komprehensif. Python juga bisa dibilang dengan bahasa pemrograman dengan tujuan umum yang akan dikembangkan secara khusus untuk membuat source code mudah dibaca. Python juga akan memiliki library yang sangat lengkap sehingga memungkinkan programmer bisa untuk membuat aplikasi yang paling mutakhir dengan menggunakan source code yang tampak sederhana.

III. HASIL DAN PEMBAHASAN

1. Pemilihan Data

Pada penelitian ini mengumpulkan data yang akan digunakan dengan cara mengambil data dengan cara observasi kepada pihak PT. Yes Study Education Group Indonesia. Atribut yang akan dipakai dari data siswa tersebut terdiri antara lain data nama-nama siswa serta nilai ujian, nilai Listening, nilai reading, nilai writing, nilai speaking. Berikut ini adalah hasil data hasil observasi yang akan ditampilkan pada Tabel 1.

Tabel 1. Data Sampel data Siswa Hasil Observasi

Nama siswa	Jenis Kelamin	Listening	Reading	Writing	Speaking
Aaron samuel	Laki-laki	5,5	5	7,5	6
Abdullah bambang	Laki-laki	7,5	7,5	7,5	8
Adhisya Prisca Nadhiya	Perempuan	7,5	5	8	7,5
Adhitya Khemal Rachmadi	Laki-laki	7,5	5	7	7
Aditya bisma putra	Laki-laki	7,5	5	5	7,5
Aflah Fikri Mahmud	Laki-laki	8	8	6	6

Setelah menampilkan data, langkah selanjutnya dari yaitu melihat tipe data pada masing- masing kolom. Hal ini bertujuan untuk mengetahui apabila ada tipe data siswa yang berbeda atau tidak ada kolom yang sama sebelum data lanjut berikut ini pada Gambar 3.

```
1 data.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 200 entries, 0 to 199
Data columns (total 6 columns):
 #   Column         Non-Null Count  Dtype
---  ------         --------------  -----
 0   Nama siswa     200 non-null    object
 1   Jenis Kelamin  200 non-null    object
 2   Listening      200 non-null    int64
 3   Reading        200 non-null    int64
 4   Writing        200 non-null    int64
 5   Speaking       200 non-null    int64
dtypes: int64(4), object(2)
memory usage: 9.5+ KB
```

Gambar 3. Hasil Info Dataset

Dataset penelitian ini terdiri dari 200 data siswa yang dikumpulkan secara menyeluruh selama periode observasi. Setiap entri mewakili satu individu siswa dengan berbagai atribut seperti identitas, nilai akademik, dan parameter penilaian non-akademik. Total 200 data siswa ini mencerminkan populasi sampel yang dianggap representatif untuk analisis klasifikasi dan segmentasi prestasi. Proses pengumpulan data dilakukan melalui observasi langsung dan pencatatan sistematis pada setiap sesi pembelajaran maupun kegiatan evaluasi. Dengan kelengkapan informasi yang terekam, kualitas data terjamin bebas dari missing values dan kesalahan pencatatan. Oleh karena itu, hasil analisis nantinya diharapkan akurat dan dapat diandalkan untuk mengidentifikasi pola prestasi siswa secara komprehensif.

2. Data Preprocessing

Pada tahap ini, setiap nilai yang semula ditulis dengan desimal menggunakan koma (misalnya 5,5; 7,5) diubah menjadi bilangan bulat puluhan (misalnya 55; 75) secara manual menggunakan Microsoft Excel. Proses konversi meliputi penggantian tanda koma menjadi angka puluhan dengan mengalikan nilai desimal tersebut dengan 10, lalu dibulatkan bila diperlukan. Dengan demikian, kolom Listening, Reading, Writing, dan Speaking yang sebelumnya memuat nilai desimal menjadi praktis untuk diolah: nilai seperti 5,5 menjadi 55; 7,5 menjadi 75; dan seterusnya. Seluruh langkah ini dilakukan tanpa bantuan skrip Python, melainkan melalui rumus Excel sederhana untuk memastikan tidak ada kehilangan data dan menghindari kesalahan imputasi otomatis (lihat Tabel 2).

Tabel 2. Hasil Preprocessing

Nama siswa	Jenis Kelamin	Listening	Reading	Writing	Speaking	Hasil Listening	Hasil Reading	Hasil Writing	Hasil Speaking
Aaron samuel	Laki-laki	5,5	5	7,5	6	55	50	75	60
Abdullah bambang	Laki-laki	7,5	7,5	7,5	8	75	75	75	80
Adhisya Prisca Nadhiya	Perempuan	7,5	5	8	7,5	75	50	80	75
Adhitya Khemal Rachmadi	Laki-laki	7,5	5	7	7	75	50	70	70
Aditya bisma putra	Laki-laki	7,5	5	5	7,5	75	50	50	75

Setelah nilai puluhan diperoleh, data dicek kembali untuk memastikan tidak ada sel kosong atau kesalahan format. Baris-baris yang memiliki missing values pada kolom kunci (nama siswa maupun nilai keempat keterampilan)

dibersihkan atau diperbaiki sesuai catatan observasi. Hasil akhir preprocessing adalah dataset lengkap dengan kolom Nama Siswa, Listening (puluhan), Reading (puluhan), Writing (puluhan), dan Speaking (puluhan), yang siap memasuki proses transformasi selanjutnya. Semua perubahan dan asumsi perhitungan terdokumentasi dalam lembar kerja Excel untuk keperluan audit dan replikasi.

3. Data Transformation

Pada tahap transformasi, kolom nilai Listening, Reading, Writing, dan Speaking yang telah berupa bilangan puluhan digunakan untuk menghitung nilai rata-rata per siswa. Untuk masing-masing baris, nilai rata-rata dihitung dengan menjumlahkan keempat nilai dan membaginya dengan empat. Misalnya, Aditya Bisma Putra dengan skor [75, 50, 50, 75] mendapatkan rata-rata 62,5, sedangkan Aflah Fikri Mahmud dengan skor [80, 80, 60, 60] memperoleh rata-rata 70. Proses ini dilakukan langsung di Excel menggunakan fungsi AVERAGE, sehingga setiap siswa kini memiliki atribut baru "Rata-Rata" yang menggambarkan capaian komprehensif.

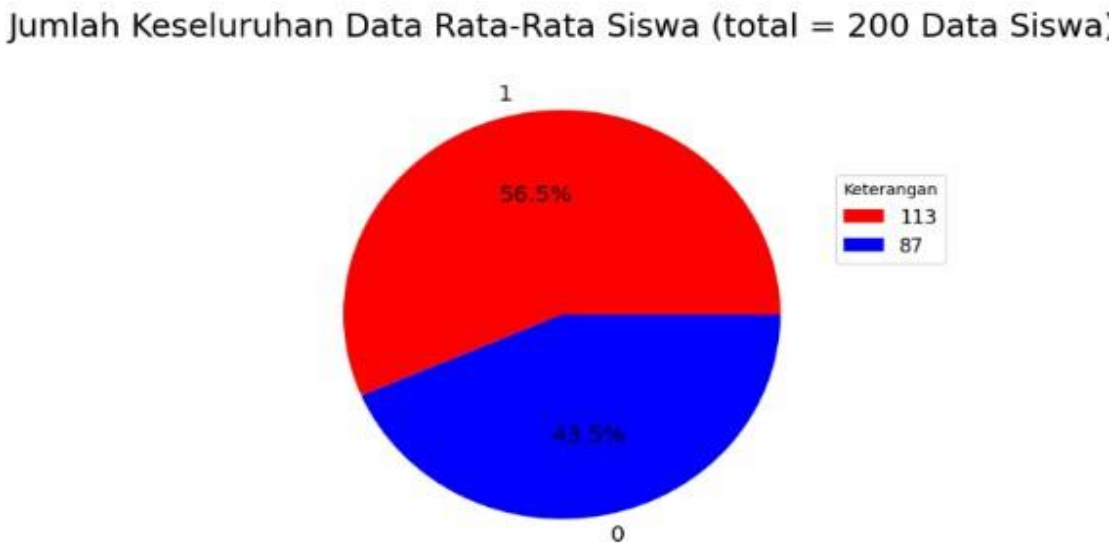
Berdasarkan nilai rata-rata tersebut, pemberian label "Prestasi" atau "Tidak Prestasi" dilakukan dengan aturan ambang 70,00. Siswa dengan rata-rata ≥ 70,00 dikategorikan "Prestasi", sedangkan siswa dengan rata-rata < 70,00 dinyatakan "Tidak Prestasi". Dari total 200 siswa, diperoleh 87 siswa berstatus "Prestasi" dan 113 siswa "Tidak Prestasi". Label ini menjadi kolom target untuk model klasifikasi Naïve Bayes selanjutnya. Semua langkah transformasi dan kriteria pelabelan direkam dalam lembar kerja agar hasil analisis dapat terverifikasi (lihat Gambar 4, Gambar 5 dan Gambar 6).

```
# Menambahkan kolom 'Rata-rata'
data['Rata'] = data[['Listening','Reading','Writing','Speaking']].mean(axis=1)

# Menampilkan DataFrame
print(data)
data.head()
```

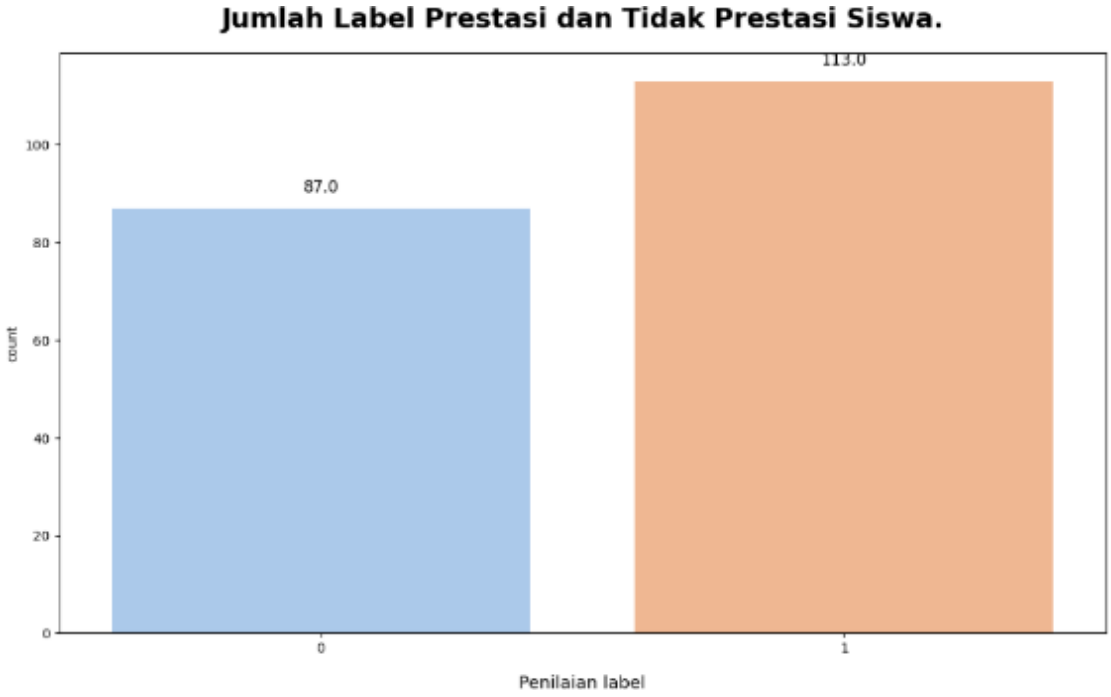
```
                Nama siswa Jenis Kelamin  Listening  Reading  Writing  \
3              Aaron samuel     Laki-laki         55       50       75
1           Abdullah bambang     Laki-laki         75       75       75
2     Adhisya Prisca Nadhiya     Perempuan         75       50       80
3    Adhitya Khemal Rachmadi     Laki-laki         75       50       70
4         Aditya bisma putra     Laki-laki         75       50       50
..                       ...           ...        ...      ...      ...
195            Yoga saputra     Laki-laki         72       65       80
196        Yeremia immanuel     Laki-laki         76       65       82
197       Yudhatama Algozali     Laki-laki         82       75       72
198           Zefanya zulian     Laki-laki         80       75       68
199          Zhiva Wicaksono     Laki-laki         85       85       90

     Speaking   Rata
3          60  60.00
1          80  76.25
2          75  70.00
3          70  66.25
4          75  62.50
..        ...    ...
195        68  71.25
196        75  74.50
197        80  77.25
198        75  74.50
199        89  87.25

[200 rows x 7 columns]
```

	Nama siswa	Jenis Kelamin	Listening	Reading	Writing	Speaking	Rata
0	Aaron samuel	Laki-laki	55	50	75	60	60.00
1	Abdullah bambang	Laki-laki	75	75	75	80	76.25
2	Adhisya Prisca Nadhiya	Perempuan	75	50	80	75	70.00
3	Adhitya Khemal Rachmadi	Laki-laki	75	50	70	70	66.25
4	Aditya bisma putra	Laki-laki	75	50	50	75	62.50

Gambar 4. Hasil Transformasi Data



Gambar 5. Keseluruhan Data



Gambar 6. Label Dataset

4. Data Mining

Pada tahap Data Mining, dataset akhir yang telah berisi kolom nilai rata-rata dan label "Prestasi"/"Tidak Prestasi" dibagi menjadi data latih dan data uji menggunakan fungsi train_test_split di Python. Pembagian ini diterapkan dalam lima skema rasio berbeda—50:50, 60:40, 70:30, 80:20, dan 90:10—dengan parameter random_state=42 untuk memastikan bahwa pembagian acak dapat direproduksi. Setiap skema dirancang untuk menguji bagaimana variasi proporsi data latih memengaruhi kemampuan model dalam mengenali pola prestasi siswa (lihat Tabel 3).

Tabel 3. Data Untuk Menguji Model

Data Latih	Data Uji
50%	50%
60%	40%
70%	30%
80%	20%
90%	10%

Selanjutnya, algoritma Naïve Bayes Classifier diimplementasikan pada data training untuk mempelajari distribusi nilai rata-rata dan label prestasi. Setelah model selesai dilatih, prediksi dilakukan pada data testing, dan hasil prediksi kemudian dianalisis menggunakan Confusion Matrix. Matriks ini memetakan True Positives (TP), True Negatives (TN), False Positives (FP), dan False Negatives (FN) sehingga dapat dihitung metrik kinerja utama: Accuracy, Precision, Recall, dan F1-Score (lihat Tabel 4).

Tabel 4. scenario pembagian data sesuai rasio dari 200 data siswa

Skenario *Rasio* Perbandingan	Data *Training*	Data *Testing*
50:50	100	100
60:40	120	80
70:30	140	60
80:20	160	40
90:10	180	20

Berdasarkan Tabel 4 tampak bahwa seiring peningkatan proporsi data latih, nilai akurasi dan F1-Score meningkat secara signifikan—dari 73% dan 80.57% pada skenario 50:50 hingga 85% dan 90.32% pada skenario 90:10. Precision juga membaik dari 67.46% menjadi 82.35%, menandakan bahwa model semakin sedikit melakukan kesalahan positif. Namun, nilai Recall konsisten sangat rendah (1%) di semua skenario, menunjukkan model sulit menangkap semua siswa berprestasi dalam prediksi.

Secara keseluruhan, skenario 90:10 terbukti optimal untuk dataset ini, menghasilkan akurasi dan F1-Score tertinggi. Meski recall yang rendah mengindikasikan bahwa model cenderung konservatif dalam menandai "Prestasi", tingginya precision dan F1-Score menunjukkan bahwa ketika model memberikan prediksi positif, kemungkinan besar prediksi tersebut benar. Dengan demikian, Naïve Bayes Classifier dengan data latih memadai dapat menjadi alat bantu yang andal untuk mengevaluasi prestasi siswa di PT. Yes Study Education Group Indonesia (lihat Tabel 5).

Tabel 5. Hasil Perfoma Evaluasi 5 Skenario

Skenario	Accuracy	Precision	Recall	F1-Score
50:50	73%	67,46%	1%	80,57%
60:40	75%	69,69%	1%	82,14%
70:30	76,66%	72,54%	1%	84,09%
80:20	82,5%	80,55%	1%	89,23%
90:10	**85%**	**82,35%**	**1%**	**90,32%**

IV. KESIMPULAN

Berdasarkan pelabelan data rata-rata, dari 200 siswa diperoleh 87 siswa berstatus "Prestasi" dan 113 siswa "Tidak Prestasi," mencerminkan sebaran kemampuan yang memengaruhi strategi evaluasi di PT. Yes Study Education Group Indonesia. Saat diuji menggunakan Naïve Bayes Classifier dengan skema split 90% data latih dan 10% data uji, model menunjukkan akurasi tinggi sebesar 85% menurut Confusion Matrix, menandakan bahwa mayoritas prediksi—baik "Prestasi" maupun "Tidak Prestasi"—cukup tepat. Hasil ini memperkuat bahwa Naïve Bayes mampu menjadi alat bantu efektif dan efisien untuk menentukan calon siswa berprestasi.

Untuk peningkatan kualitas penelitian selanjutnya, penting menambah volume dataset menjadi jauh lebih besar dan memperkaya atribut—misalnya menambahkan data non-akademik (kepemimpinan, keaktifan ekstrakurikuler, atau skor psikometrik)—agar model mendapatkan informasi lebih variatif. Selain itu, penerapan feature selection seperti Chi-Square, Genetic Algorithm, atau metode wrapper lain dapat menyaring atribut paling berpengaruh, sehingga hanya sekumpulan fitur paling relevan yang digunakan tanpa menurunkan akurasi.

Lebih jauh, penelitian ini dapat dikembangkan dengan mengoptimalkan parameter model Naïve Bayes menggunakan teknik Particle Swarm Optimization, Genetic Algorithm, atau hyperparameter tuning berbasis grid/random search. Pendekatan ini diharapkan meningkatkan performa metrik—terutama recall—sehingga model tidak hanya akurat dalam memprediksi "Prestasi," tetapi juga mampu menangkap lebih banyak kasus positif dengan lebih baik.

DAFTAR PUSTAKA

[1] R. Doharma dan D. Mafiroh, "Perancangan_Sistem_Informasi_Penilaian_Prestasi_Si, " *Infotech*, ISSN, vol. 4, no. 2, hal. 34–43, 2018,

[2] R. Hasudungan dan W. J. Pranoto, "Implementasi Teorema Naïve Bayes Pada Prediksi Prestasi Mahasiswa," *J. Rekayasa Teknol. Inf.,* vol. 5, no. 1, hal. 10, 2021, doi: 10.30872/jurti.v5i1.4996.

[3] N. Ismaya et al., "Penentuan Siswa Berprestasi Menggunakan Metode K-Means Clustering Di Smp Takhassus Al Qur ' an," *J. Tek. Inform*., vol. 1, no. 1, hal. 64–68, 2022.

[4] S. Mita, Y. Yamazoe, T. Kamataki, dan R. Kato, Data Mining, vol. 14, no. 3. 1981. doi: 10.1016/0304-3835(81)90152-X.
[5] N. Bloom dan J. Van Reenen, *Data Mining : Concepts and Technique*. 2013. [Daring]. Tersedia pada: http://www.nber.org/papers/w16019
[6] S. S. Syarat, M. Gelar, S. Komputer, F. Ilmu, K. Universitas, dan S. Karawang, Sosial Media Tiktok Menggunakan Metode Naïve Bayes Classifier. 2022.
[7] T. Fawcett, "An introduction to ROC analysis," Pattern Recognit. Lett., vol. 27, no. 8, hal. 861–874, 2022, doi: 10.1016/j.patrec.2005.10.010.