Dawatuna: Journal of Communication and Islamic Broadcasting
Volume 4 Nomor 4 (2024) 1587 – 1598 E-ISSN 2798-6683 P-ISSN 2798-690X
DOI: 10.47467/dawatuna.v4i4.1975

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

Novan Ponco Laksono*[1], Prima Dina Atikah*[2], Ahmad Fathurrozi*[3]
[1,2,3]Prodi Informatika, Fakultas Ilmu Komputer Universitas Bhayangkara Jakarta Raya
novan.ponco.lasksono19@mhs.ubharajaya.ac.id[1]

ABSTRAK

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,

PENDAHULUAN

Tingginya tingkat keberhasilan siswa merupakan ceminan daripada kualutas dunia Pendidikan. Belakangan ini dunia pendidikan dituntut untuk memiliki daya saing yang tinggi dengan memanfaatkan kualitas serta kuantitas sumber daya manusia (SDM) yang ada, seperti halnya siswa berprestasi. Namun, realitanya sesorang mempunyai kemampuan dalam hal penguasaan dan pemahaman yang berbeda-beda. Pendidikan merupakan salah satu upaya yang dilakukan untuk menjadi penentu keberhasilan suatu bangsa. Perkembangan individual setiap siswa memiliki kemampuan, minat, dan kecepatan belajar yang berbeda-beda. Oleh karena itu, prestasi siswa juga harus dilihat sebagai refleksi dari perkembangan individual mereka. Dalam menentukan prestasi siswa, penting untuk empertimbangkan faktor-faktor ini dan memberikan penilaian yang adil dan akurat berdasarkan kemampuan dan perkembangan siswa secara keseluruhan, prestasi siswa juga terkait erat dengan tujuan pendidikan secara umum. Tujuan pendidikan mencakup pengembangan pengetahuan, keterampilan, sikap dan nilai – nilai siswa.

Permasalahan yang terjadi dilapangan adalah proses penentuan siswa berprestasi masih terpaku pada nilai akhir siswa yang didapatkan yaitu nilai akademik sedangkan

Volume 4 Nomor 4 (2024) 1587 – 1598 E-ISSN 2798-6683 P-ISSN 2798-690X
DOI: 10.47467/dawatuna.v4i4.1975

nilai non-akademik. Sehingga dirasa kurang adil dalam penetuan siswa berprestasi. Dangan mempertimbangkan latar belakang ini, para pendidik dapat mengambil pendekatan yang holistic dan komperhensif dalam menetapkan prestasi siswa.[1]

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

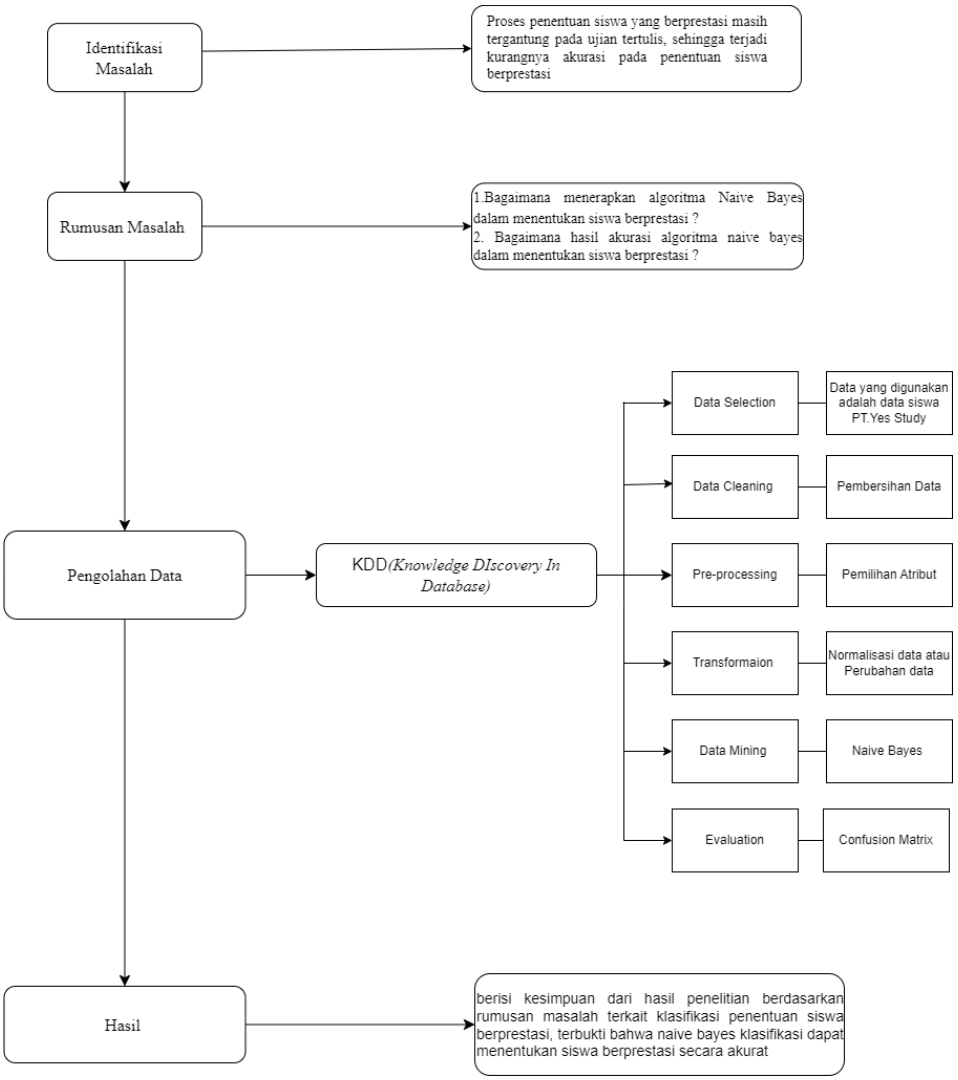
Dawatuna: Journal of Communication and Islamic Broadcasting
Volume 4 Nomor 4 (2024) 1587 – 1598 E-ISSN 2798-6683 P-ISSN 2798-690X
DOI: 10.47467/dawatuna.v4i4.1975

H P(X) : Probabilitas X

METODOLOGI PENELITIAN

1.1 Kerangka penelitian

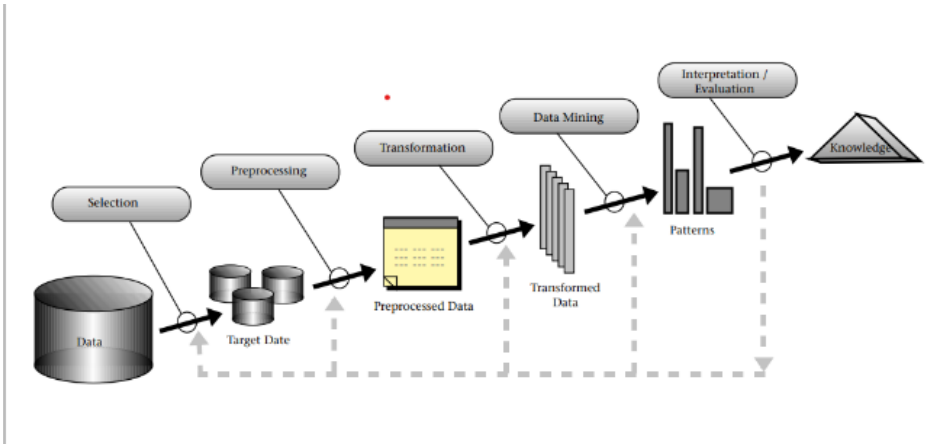
Dalam penelitian ini, peneliti membuat kerangka penelitian yang berguna sebagai dasar pemikiran dalam klasifikasi penentuan siswa berprestasi menggunakan algoritma naïve bayes, berikut ini merupakan gambar kerangka penelitian, yaitu sebagai berikut :



Dawatuna: Journal of Communication and Islamic Broadcasting
Volume 4 Nomor 4 (2024) 1587 – 1598 E-ISSN 2798-6683 P-ISSN 2798-690X
DOI: 10.47467/dawatuna.v4i4.1975

1.2 **Knowledge Discovery In Database (KDD)**

Gambar Proses KDD (*Knowledge Discovery In Database*)



*Knowledge Discovery in Databas*es (KDD) merupakan sekumpulan proses untuk menemukan pengetahuan yang bermanfaat dari data. 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] :

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 *Preprocessing***
   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*

Volume 4 Nomor 4 (2024) 1587 – 1598 E-ISSN 2798-6683 P-ISSN 2798-690X
DOI: 10.47467/dawatuna.v4i4.1975

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

   1.4 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 [13].
5. *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

3.4 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*. A*curacy* 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* [15], digunakan *confusion matrix*, yaitu tes cari tahu sejauh mana klasifikasi berlaku untuk kelas yang berbeda. *Confusion matrix* dapat dilihat pada dilihat sebagai berikut :

Dawatuna: Journal of Communication and Islamic Broadcasting
Volume 4 Nomor 4 (2024) 1587 – 1598 E-ISSN 2798-6683 P-ISSN 2798-690X
DOI: 10.47467/dawatuna.v4i4.1975

Nilai Prediksi	Nilai Aktual: Positive	Nilai Aktual: Negative
Positive	TP	FP
Negative	FN	TN

Keterangan:

- True Positive (TP): yaitu jumlah data dengan kelas positif yang diklasifikasikan positif.
- True Negative (TN): yaitu jumlah data dengan kelas negative yang diklasifikasikan negatif.
- False Positive (FP): yaitu jumlah data dengan kelas positif yang diklasifikasikan negatif.
- 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. Anda dapat menemukan persamaan untuk nilai akurasi dengan:

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN} \quad (2.2)$$

2. *Precision*, menentukan seberapa baik sistem mencocokkan kueri pengguna dengan data yang diambil dan dikembalikannya. Persamaan berikut dapat digunakan untuk menentukan akurasi:

$$Precision = \frac{TP}{TP+FP} \quad (2.3)$$

3. *Recall* adalah proporsi dari setiap informasi yang akan ditemukan dari label [16], rumus yang akan digunakan perhitungan *recall* sebagai berikut:

$$Recall = \frac{TP}{TP+FN} \quad (2.4)$$

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* [16], rumus untuk mencari *F1-Score* sebagai berikut:

$$F - measure = \frac{2.(recall.precision)}{(recall+precision)} \quad (2.5)$$

3.1 Python

Dawatuna: Journal of Communication and Islamic Broadcasting
Volume 4 Nomor 4 (2024) 1587 – 1598 E-ISSN 2798-6683 P-ISSN 2798-690X
DOI: 10.47467/dawatuna.v4i4.1975

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.

IV HASIL DAN PEMBAHASAN

4.1 Data selection

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 4.1 sebagai berikut.:

Tabel 4.1 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

Dawatuna: Journal of Communication and Islamic Broadcasting
Volume 4 Nomor 4 (2024) 1587 – 1598 E-ISSN 2798-6683 P-ISSN 2798-690X
DOI: 10.47467/dawatuna.v4i4.1975

```
#memasukan data testing ke datalatih
data = pd.read_csv("dataset_nilai.csv")
data.head(11)
#Prestasi Siswa = 1 == Ya
#Prestasi Siswa = 2 == Tidak
#Prestasi Sekolah = 1 == Cukup
#Prestasi Sekolah = 2 == Baik
```

	Nama siswa	Jenis Kelamin	Listening	Reading	Writing	Speaking
0	Aaron samuel	Laki-laki	5,5	5	7,5	6
1	Abdullah bambang	Laki-laki	7,5	7,5	7,5	8
2	Adhisya Prisca Nadhiya	Perempuan	7,5	5	8	7,5
3	Adhitya Khemal Rachmadi	Laki-laki	7,5	5	7	7
4	Aditya bisma putra	Laki-laki	7,5	5	5	7,5
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 4.2 hasilnya sebagai berikut :

```
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
```

Masing-masing data terdapat 200 data siswa berarti total dari jumlah data yang akan diteliti berjumlah 200 data siswa yang dihasilkan pada saat observasi.

2.4 4.2 Data 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

Volume 4 Nomor 4 (2024) 1587 – 1598 E-ISSN 2798-6683 P-ISSN 2798-690X
DOI: 10.47467/dawatuna.v4i4.1975

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

Dan langkah selanjutnya yaitu data preprocessing yaitu membersikan nilai-nilai yang berbentuk koma menjadi nilai bilangan bulat, langkah ini dilakukan tidak menggunakan bahasa *python,* melainkan dengan menggunakan perhitungan manual dengan Microsoft Excel dan berikut ini adalah perhitungan dengan menggunakan Microsoft Excel dan hasil juga beserta hasil dari perhitungan tersebut akan ditampilkan pada Tabel diatas. Menunjukkan bahawa ada sedikit perubahan pada nilai listening, reading, writing dan nilai speaking yang sebelumnya dari nilai koma menjadi nilai puluhan, dan dari hasil nilai puluhan dari setiap kolom tersebut akan diolah ke tahap selanjutnya yaitu tahap data transformation.

3.4 4.3 *Data Tranformation*

Dawatuna: Journal of Communication and Islamic Broadcasting
Volume 4 Nomor 4 (2024) 1587 – 1598 E-ISSN 2798-6683 P-ISSN 2798-690X
DOI: 10.47467/dawatuna.v4i4.1975

```
# 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  \
0             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
0          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

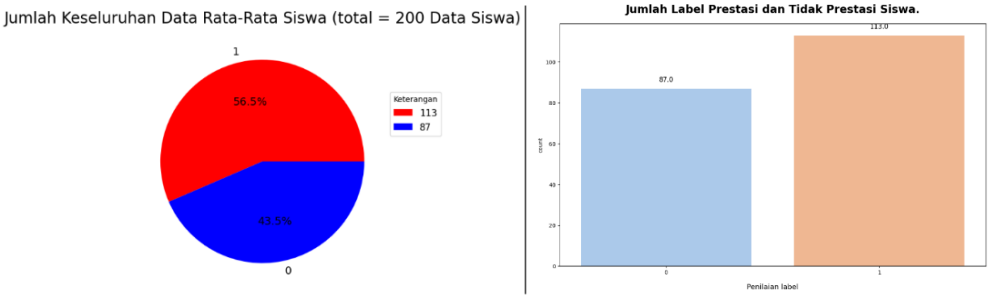
Dan setelah menentukan nilai rata -rata hasil dari data siswa dari nilai *listening, reading*, *writing dan nilai dari speaking* , nantinya data nilai rata-rata akan dilakukan dimana hasil dari penentuan tersebut dari data nilai rata-rata tersebut.

Nama siswa	Listening	Reading	Writing	Speaking	Rata-Rata
Aditya bisma putra	75	50	50	75	62,5
Aflah Fikri Mahmud	80	80	60	60	70
Aini Maryani	70	75	70	65	70
Firly Oktaviani	70	75	70	75	72,5
Frans David Matthew Siallagan	75	70	55	75	68,75
I MADE BRIANTA GAVIN PUTRA	50	55	80	60	61,25

Nantinya apabila nilai rata-rata siswa dibawah nilai 70,00 maka hasil dari pelabelan tersebut mendapatkan status “Tidak Prestasi”, begitu pula sebaliknya jika nilai rata-rata siswa itu diatas nilai 70,00 maka data siswa tersebut akan dinyatakan “Prestasi

Dawatuna: Journal of Communication and Islamic Broadcasting
Volume 4 Nomor 4 (2024) 1587 – 1598 E-ISSN 2798-6683 P-ISSN 2798-690X
DOI: 10.47467/dawatuna.v4i4.1975

". Dan berikut ini adalah source code dan hasil dari penentuan prestasi nilai rata-rata tersebut.



Hasil penentuan data siswa memperoleh sebanyak 87 siswa yang mendapat nilai kelas Prestasi dan sedangkan kelas tidak prestasi memperoleh data sebanyak 113 siswa dengan banyaknya mendapatkan nilai tidak prestasi menjadi dampak pada pihak PT. *Yes Study Education Group* Indonesia, yang disajikan pada Gambar diatas

4.4 Data mining

Sebelum melakukan tahapan untuk perhitungan algoritma *Naïve Bayes Classifier* terlebih dahulu harus membagikan menjadi data latih dan data uji, hal ini digunakan untuk melatih model klasifikasi (pelatihan) yang berisikan pengetahuan yang kemudian akan digunakan untuk memprediksi kelas sentimen yang baru. Semakin banyak pemodelan yang dilatih tentang data, maka semakin juga baik algoritma tersebut memahami data. Data Uji yang akan dibagi dengan perbandingan yang berada di Tabel 4.5 sebagai berikut.

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

PENUTUP

Kesimpulan

Dalam penelitian ini dapat disimpulkan dalam hasil pelabelan data siswa memperoleh sebanyak 87 siswa yang mendapat nilai kelas Prestasi dan sedangkan kelas tidak prestasi memperoleh data sebanyak 113 siswa dengan banyaknya mendapatkan nilai tidak prestasi menjadi dampak pada pihak PT. *Yes Study Education Group* Indonesia. Dan selanjutnya pada hasil pengujian dengan algoritma *Naïve Bayes Classifier,* hasil evaluasi yang diketahui oleh *confusion matriks* memiliki nilai accuracy cukup tinggi dengan rasio 90:10 dengan nilai sebesar 85%. Dengan demikian, algoritma Naïve Bayes merupakan metode yang cukup baik dalam menentukan calon

Volume 4 Nomor 4 (2024) 1587 – 1598 E-ISSN 2798-6683 P-ISSN 2798-690X
DOI: 10.47467/dawatuna.v4i4.1975

siswa berpreastasi PT. *Yes Study Education Group* Indonesia secara lebih efektif dan efisien.

Saran

Agar penelitian ini bisa ditingkatkan, berikut adalah saran-saran yang diusulkan :

1. Menambahkan jumlah data yang lebih besar dan atribut yang lebih banyak, sehingga hasil pengukuran yang akan didapatkan lebih baik lagi.
2. Melakukan pengembangan dengan *feature selection* seperti *genetic algorithm*, *chi square*, dan metode *feature selection* lainnya. untuk menyeleksi atribut yang berpengaruh kuat, sehingga atribut yang dipakai hanya sedikit namun tidak mengurangi akurasi dari algoritma yang digunakan.
3. Penelitian ini dapat dikembangkan dengan mengoptimalkan parameter dengan *Particle Swarm Optimization, Genetic Algorithm* dan lainnya.