# PENERAPAN ALGORITMA NAIVE BAYES DAN METODE CRISP-DM DALAM PREDIKSI HASIL TES KEMAMPUAN BAHASA INGGRIS MAHASISWA

**Sany Santiastry, Asriyanik, Winda Apriandari**
Teknik Informatika, Universitas Muhammadiyah Sukabumi
Jl. R. Syamsudin, S.H. No. 50, Cikole, Kec. Cikole, Kota Sukabumi, Jawa Barat 43113
*santiastri25@gmail.com*

**ABSTRAK**

Bahasa Inggris memiliki peran penting dalam sistem pendidikan di Indonesia, terutama karena fungsinya sebagai bahasa internasional. Pentingnya hal ini terlihat dari banyaknya informasi ilmiah dan teknologi di berbagai bidang yang ditulis dalam bahasa Inggris. Meskipun demikian, belum ada penelitian di Universitas Muhammadiyah Sukabumi yang berfokus pada memprediksi tingkat keberhasilan mahasiswa dalam tes kecakapan bahasa Inggris. Penelitian ini bertujuan untuk membangun model prediksi menggunakan algoritma Naive Bayes di *platform* Google Collaboratory untuk memperkirakan hasil tes kecakapan bahasa Inggris mahasiswa di Universitas Muhammadiyah Sukabumi. Naïve Bayes adalah salah satu algoritma *machine learning* yang banyak digunakan dalam klasifikasi dan prediksi. Penelitian ini menghasilkan model prediktif serta situs web sederhana yang dapat mencatat nilai mata kuliah bahasa Inggris mahasiswa dan memprediksi keberhasilan mereka dalam tes kemahiran bahasa Inggris. Evaluasi model ini menggunakan metrik akurasi, presisi, *recall*, dan nilai F1. Model yang dibuat menunjukkan akurasi sebesar 87,94%, dengan rata-rata presisi makro 0,82, *recall* 0,91, dan nilai F1 0,84, serta rata-rata presisi berbobot 0,89, *recall* 0,88, dan nilai F1 0,88. Model ini dapat secara akurat memprediksi hasil kelulusan mahasiswa berdasarkan nilai mereka.

***Kata kunci :*** *Prediksi, Naive Bayes, Tes Bahasa Inggris, Kelulusan, Machine Learning.*

## 1. PENDAHULUAN

Pendidikan tinggi memainkan peran kunci dalam mengembangkan individu dengan meningkatkan wawasan mereka dan meningkatkan kualitas sumber daya manusia untuk menciptakan tenaga kerja yang professional [1].

Di tengah persaingan global, keterampilan berbahasa Inggris menjadi semakin penting, terutama dalam era globalisasi [2].

Penguasaan bahasa asing dapat mendukung mahasiswa dalam mencapai tujuan akademis dan karir mereka, memperluas kesempatan kerja, dan memungkinkan komunikasi yang lebih efektif dengan orang-orang dari berbagai negara. Oleh karena itu, kemampuan memahami dan mempelajari bahasa asing sangat penting dalam mempersiapkan diri untuk dunia kerja [3].

Sebagai bahasa internasional, bahasa Inggris memiliki peran yang signifikan dalam dunia pendidikan di Indonesia [4].

Ini tidak bisa dihindari karena banyak pengetahuan ilmiah dan teknologi di berbagai bidang yang terdokumentasi dalam bahasa Inggris atau bahasa asing lainnya [5].

Oleh karena itu, kemampuan dalam bahasa Inggris atau bahasa asing lainnya menjadi alat penting bagi masyarakat Indonesia untuk berkontribusi dalam kemajuan pengetahuan dan penyebaran informasi di dalam negeri [6].

Selain itu, dalam konteks masyarakat global yang dipengaruhi oleh Revolusi Industri 4.0, kemajuan teknologi informasi telah menghilangkan batasan jarak dan waktu [7].

Dalam dunia yang semakin terhubung ini, penguasaan bahasa asing, terutama bahasa Inggris, menjadi kunci bagi masyarakat Indonesia untuk ikut serta dalam interaksi internasional dan berperan sebagai warga dunia [8].

Menurut laporan English Proficiency Index (EPI) 2023 oleh EF Education First, Indonesia menempati peringkat ke-79 dari 113 negara, turun enam posisi dari tahun sebelumnya, dengan skor EPI 473, di bawah rata-rata global 503. Ini menunjukkan rendahnya kemampuan bahasa Inggris di Indonesia [9].

Dibandingkan dengan negara-negara tetangga seperti Singapura, Filipina, dan Malaysia, yang berada di posisi teratas di Asia, kinerja Indonesia dalam bahasa Inggris relatif lemah [10].

Universitas Muhammadiyah Sukabumi, sebagai institusi pendidikan tinggi yang kompetitif, berupaya meningkatkan keterampilan bahasa Inggris mahasiswanya. Mata kuliah bahasa asing diwajibkan bagi mahasiswa tahun pertama, dan tes kemampuan bahasa Inggris dilakukan oleh Lembaga Kerjasama dan Hubungan Internasional (LKHI). Mengingat pentingnya kemampuan bahasa Inggris untuk menghadapi tuntutan global, memprediksi keberhasilan mahasiswa dalam tes ini menjadi krusial. Namun, hingga kini belum ada sistem prediktif di universitas untuk menilai keberhasilan mahasiswa dalam tes tersebut.

Belum ada penelitian di Universitas Muhammadiyah Sukabumi yang dapat memprediksi keberhasilan mahasiswa dalam bahasa Inggris. Oleh karena itu, penelitian ini akan menerapkan algoritma Naive Bayes untuk memprediksi keberhasilan mahasiswa dalam tes kemampuan bahasa Inggris di

universitas. Data yang digunakan adalah dataset yang berisi nilai mahasiswa dari Lembaga Kerjasama dan Hubungan Internasional (LKHI) Universitas. Penelitian ini, yang berjudul "PREDIKSI HASIL TES KEMAMPUAN BAHASA INGGRIS MAHASISWA MENGGUNAKAN ALGORITMA NAIVE BAYES DAN METODE CRISP-DM" diharapkan dapat membantu LKHI dan Universitas Muhammadiyah Sukabumi dalam memprediksi keberhasilan mahasiswa dalam tes kemampuan bahasa Inggris.

## 2. TINJAUAN PUSTAKA

### 2.1. *Machine Learning*

*Machine Learning* adalah cabang ilmu komputer yang bertujuan untuk memungkinkan komputer "belajar" tanpa harus diprogram secara eksplisit. Berasal dari gerakan kecerdasan buatan (AI) pada tahun 1950-an, *Machine Learning* difokuskan pada tujuan praktis seperti prediksi dan optimalisasi [11].

Teknologi ini digunakan untuk memprediksi dan menganalisis data, terutama data baru yang belum diketahui. Prediksi data dapat meningkatkan kinerja komputer, sementara analisis data membantu memperoleh pengetahuan baru. Tujuan utama *Machine Learning* adalah untuk mengembangkan model yang dapat secara akurat memprediksi dan menganalisis data [12].

### 2.2. CRISP-DM

CRISP-DM adalah kerangka kerja yang banyak digunakan untuk menangani masalah bisnis dan penelitian. Metodologi ini dibagi menjadi enam tahap: *Business Understanding, Data Understanding, Data Preparation, Modelling, Evaluation,* dan *Deployment* [10]. Berikut adalah penjelasan rinci dari setiap tahap:

a. *Business Understanding*: Berisi pemahaman terhadap tujuan dan kebutuhan bisnis.
b. *Data Understanding*: Berisi pengumpulan dan eksplorasi data awal untuk mendapatkan pemahaman terhadap data.
c. *Data Preparation*: Tahap ini mencakup beberapa hal seperti membersihkan data (Data Cleaning), memilih data yang relevan (Data Selection), dan mentransformasi data (Data Transformation) untuk digunakan dalam tahap pemodelan.
d. *Modelling*: Berisi pembuatan dan penerapan model pada data yang telah dipersiapkan.
e. *Evaluation*: Tahap ini berfokus pada pengujian dan penilaian model untuk memastikan mereka memenuhi tujuan bisnis.
f. *Deployment*: Tahap akhir ini melibatkan penerapan model ke dalam lingkungan operasional untuk digunakan secara praktis.

### 2.3. Naïve Bayes

Naive Bayes adalah metode klasifikasi yang sederhana namun efektif. Algoritma ini memprediksi kelas suatu data berdasarkan probabilitas dan asumsi independen antar atribut. Dengan mengasumsikan bahwa setiap fitur data tidak saling terkait, algoritma ini menghitung probabilitas kelas suatu data berdasarkan nilai fitur-fiturnya. Meski sederhana, Naive Bayes terbukti efektif dalam berbagai aplikasi seperti filter spam, analisis sentimen, dan prediksi berdasarkan data historis [13].

Persamaan algoritma Naïve Bayes [14] adalah sebagai berikut:

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)} \quad (1)$$

Keterangan:
A = Hipotesis data B merupakan suatu kelas spesifik
B = Data dengan kelas yang belum diketahui
P(A) = Probabilitas dari A
P(B) = Probabilitas dari B
P(A|B) = Probabilitas hipotesis A berdasarkan kondisi B
P(B|A) = Probabilitas hipotesis B berdasarkan kondisi B

### 2.4. Prediksi

Klasifikasi adalah metode dalam *Machine Learning* yang digunakan untuk mengategorikan data ke dalam kelas-kelas tertentu. Proses ini melibatkan pembelajaran dari data berlabel (data training) untuk membuat prediksi terhadap data baru yang belum diberi label (data testing) dengan memanfaatkan data historis dan indikator tertentu untuk meramalkan peristiwa di masa depan [15].

### 2.5. Python

Python dikenal sebagai bahasa pemrograman *open source* yang menawarkan fleksibilitas dan kemudahan luar biasa. Kesederhanaan Python sebagai bahasa skrip serta aplikasinya yang serbaguna menjadikannya pilihan ideal untuk pengembangan program prediksi dalam berbagai konteks [16].

Selain itu, Python memiliki banyak koleksi *package* atau modul yang siap digunakan untuk berbagai kebutuhan, seperti pengembangan *artificial intelligence (AI), big data, data mining, Machine Learning,* hingga *deep learning* [17].

### 2.6. Google Collaboratory

Google Colaboratory (Google Colab) adalah *platform online* yang memungkinkan pembuatan prototipe model *Machine Learning* menggunakan perangkat keras canggih seperti GPU dan TPU. Sebagai *Integrated Development Environment* (IDE) berbasis Python, Google Colab memproses data di cloud menggunakan *server* Google yang kuat. Platform ini juga menyediakan berbagai library penting untuk analisis data, termasuk Keras, TensorFlow, NumPy, Pandas, dan Matplotlib [18], [19].

### 2.7. Penelitian Terdahulu

Penelitian oleh Widaningsih mengungkapkan bahwa algoritma Naïve Bayes unggul dibandingkan

dengan algoritma lainnya dalam semua aspek performa [20].

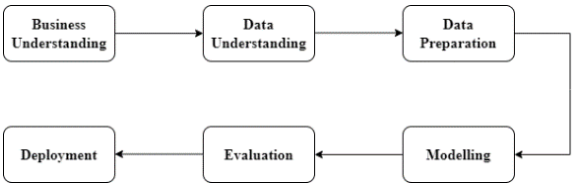
Algoritma ini menunjukkan kinerja yang lebih baik dalam hal akurasi dan nilai AUC, di mana nilai yang lebih tinggi lebih diinginkan, serta dalam tingkat kesalahan, di mana nilai yang lebih rendah lebih diharapkan. Nilai AUC untuk algoritma Naïve Bayes dan C4.5 dikategorikan sebagai "baik", sementara nilai AUC untuk SVM dan kNN masuk dalam kategori "cukup". Secara keseluruhan, hasil penelitian ini menunjukkan bahwa Naïve Bayes adalah algoritma paling efektif untuk memprediksi kelulusan mahasiswa tepat waktu dan pencapaian IPK ≥ 3, dengan tingkat akurasi 76,79%, tingkat kesalahan 23,17%, dan nilai AUC 0,850.

Penelitian lainnya yang dilakukan oleh Rovidatul Hikmah Tanjung, Yuhandri Yunus, dan Gunadi Widi Nurcahyo menunjukkan bahwa algoritma Naïve Bayes lebih unggul dibandingkan C4.5 dalam memprediksi kelulusan mahasiswa [21].

Hasil penelitian ini mengungkapkan bahwa Naïve Bayes memiliki akurasi yang konsisten sebesar 81,58%, sementara akurasi algoritma C4.5 mengalami penurunan dari 77,19% menjadi 73,68%.

## 3. METODE PENELITIAN

Metodologi penelitian ini menjelaskan kerangka dan rencana pelaksanaan penelitian, menggunakan metode CRISP-DM. Tahapan ini mencakup detail proses pengumpulan data dari instansi terkait, penjelasan mengenai alat bantu yang digunakan, serta jadwal kegiatan yang merinci langkah-langkah penelitian. Berikut adalah alur metode CRISP-DM yang digunakan untuk mengarahkan setiap langkah penelitian ini.




Gambar 1. Alur Metode CRISP-DM [22]

### 3.1. *Business Understanding*

Pada tahap *Business Understanding*, penelitian ini bertujuan untuk memprediksi kelulusan mahasiswa dalam tes kemahiran bahasa Inggris di Universitas Muhammadiyah Sukabumi. Selain itu, kriteria evaluasi seperti akurasi dan presisi prediksi ditetapkan untuk mengukur keberhasilan model.

### 3.2. *Data Understanding*

Di tahap *Data Understanding*, dilakukan beberapa langkah, termasuk pengumpulan, deskripsi, eksplorasi, dan verifikasi kualitas data. Data dikumpulkan dari Lembaga Kerjasama dan Hubungan Internasional (LKHI) dan Program Studi Teknik Informatika di Universitas Muhammadiyah Sukabumi, mencakup nilai *speaking*, *writing*, *listening*, *reading*, dan ujian akhir Bahasa Inggris. Data ini dieksplorasi untuk memahami struktur, karakteristik, serta kualitasnya, termasuk identifikasi data yang hilang, duplikasi, dan anomali.

### 3.3. *Data Preparation*

*Data Preparation* melibatkan pembersihan, pemilihan, dan transformasi data sebelum pemodelan untuk memastikan pemodelan yang optimal. Proses ini menangani masalah pada data agar informasi yang diperlukan dapat diperoleh.

### 3.4. *Modelling*

Pada tahap modeling, metode Naive Bayes digunakan untuk membangun model prediksi kelulusan mahasiswa, dilatih dengan data yang telah dipersiapkan. Proses ini juga mencakup pengoptimalan parameter model untuk meningkatkan kinerja prediksi.

### 3.5. *Evaluation*

Tahap *Evaluation* melibatkan pengukuran kinerja model menggunakan metrik seperti akurasi, presisi, *recall*, dan F1-score. Model diuji dengan data yang belum pernah dilihat sebelumnya untuk memastikan performa yang optimal dan bahwa model memenuhi tujuan penelitian.

### 3.6. *Deployment*

*Deployment* akhir berupa website yang dibangun menggunakan Visual Studio Code. Editor teks ini digunakan untuk menyusun model, mengatur struktur, dan logika program, serta memungkinkan pengelolaan dan pengeditan kode secara efisien.

## 4. HASIL DAN PEMBAHASAN

### 4.1. *Business Understanding*

Fokus dari penelitian ini adalah memprediksi kelulusan mahasiswa dalam tes kemahiran bahasa Inggris di Universitas Muhammadiyah Sukabumi, yang meliputi keterampilan *speaking*, *writing*, *listening*, *reading*, serta nilai ujian akhir (PTS dan PAS). Kriteria evaluasi seperti akurasi, presisi, *recall*, dan F1-score ditetapkan untuk menilai keberhasilan model prediksi.

### 4.2. *Data Understanding*

Pada tahap *Data Understanding*, langkah-langkah yang dilakukan meliputi pengumpulan, deskripsi, eksplorasi, dan verifikasi kualitas data. Data dikumpulkan dari Universitas Muhammadiyah Sukabumi, khususnya dari Lembaga Kerjasama dan Hubungan Internasional (LKHI) dan Program Studi Teknik Informatika, mencakup nilai *speaking*, *writing*, *listening*, *reading*, dan ujian akhir Bahasa Inggris (PTS dan PAS).

Tabel 1. Detail Atribut Dataset Nilai

| Atribut | Penjelasan |
|---|---|
| NIM | Identifikasi unik mahasiswa. |
| Nama | Nama lengkap mahasiswa. |
| Nilai Tugas Speaking | Nilai untuk kategori berbicara dalam Bahasa Inggris. |
| Nilai Tugas Reading | Nilai untuk kategori membaca dalam Bahasa Inggris. |
| Nilai Tugas Listening | Nilai untuk kategori mendengarkan dalam Bahasa Inggris. |
| Nilai Tugas Writing | Nilai untuk kategori menulis dalam Bahasa Inggris. |
| Rerata | Rata-rata nilai |
| Prosen Nilai | Persentase nilai dari keseluruhan nilai |
| PTS | Nilai penilaian tengah semester |
| Prosen PTS | Persentase nilai PTS dari keseluruhan nilai |
| PAS | Nilai penilaian akhir semester |
| Prosen PAS | Persentase nilai PAS dari keseluruhan nilai |
| Kehadiran | Nilai kehadiran mahasiswa di mata kuliah Bahasa Inggris |
| Prosen Kehadiran | Persentase nilai kehadiran dari keseluruhan nilai |
| Nilai Angka | Total nilai Bahasa Inggris dalam angka. |
| Nilai Huruf | Total nilai Bahasa Inggris dalam huruf (A, B, C, D, E). |

Data ini diproses dengan algoritma Naive Bayes, yang digunakan untuk membangun model prediksi kelulusan mahasiswa berdasarkan analisis data. Algoritma ini membantu menemukan pola dan faktor yang mempengaruhi kelulusan mahasiswa dalam tes kemahiran Bahasa Inggris, dengan data dari tahun 2021 hingga 2023.

### 4.3. *Data Preparation*

Tahapan *pengolahan* data awal mencakup pemilihan atribut data yang akan digunakan dan penghapusan atribut yang tidak diperlukan untuk proses pemodelan. Pada tahap persiapan data, langkah-langkah berikut dilakukan:

a. Membaca semua *sheet* dari file Excel, kecuali sheet yang berjudul "Catatan."
b. Menghapus baris yang mengandung metadata atau informasi tambahan yang tidak relevan, seperti nama file, fakultas, dan universitas, dengan menghapus 11 baris pertama.
c. Menghilangkan kolom yang tidak relevan berdasarkan indeks, seperti nilai 1, 2, dan 3 dari kategori *speaking, reading, listening, writing*, dan persentase dari setiap kategori.
d. Menyesuaikan nama kolom agar sesuai dengan format yang diinginkan, seperti *'Speaking', 'Reading', 'Listening', 'Writing'*, 'PTS', 'PAS', 'Kehadiran', 'Nilai Angka', dan 'Nilai Huruf.'
e. Mengubah karakter *non-numeric* menjadi NaN *(Not a Number)* pada kolom tertentu dan mengganti nilai NaN dengan nilai *default* (0 untuk kolom *numeric* dan 'E' untuk kolom kategori).
f. Menghapus baris jika kolom *'Reading'* atau *'Listening'* kosong (NaN).
g. Menyesuaikan tipe data kolom agar sesuai dengan tipe yang diinginkan (misalnya, int64 untuk kolom *numeric* dan *category* untuk kolom kategori).
h. Menggabungkan semua DataFrame yang telah dibersihkan menjadi satu DataFrame.

### 4.4. *Modelling*

Penelitian ini menggunakan dua jenis data: data pelatihan (*data*_training.csv) dan data pengujian (data_testing.csv). Kedua set data ini mencakup informasi tentang nilai mahasiswa dalam berbagai kategori seperti *'speaking', 'reading', 'listening', 'writing'*, 'PTS', dan 'PAS', serta nilai akhir dalam format angka dan huruf. Model yang dikembangkan mampu memprediksi grade mahasiswa dengan akurasi sebesar 87,94% berdasarkan data nilai dari berbagai kategori mata kuliah Bahasa Inggris tersebut.

Tabel 2. Tampilan Dataset setelah Cleaning Data

| NIM | Nama | speaking | reading | listening | writing | PTS | PAS | Nilai Angka | Nilai Huruf |
|---|---|---|---|---|---|---|---|---|---|
| 2130221001 | ASRI RANTI NURAENI | 75 | 75 | 70 | 85 | 76 | 90 | 79 | B |
| 2130221002 | RANGGA SRI TRESNA | 80 | 85 | 70 | 85 | 84 | 80 | 81 | B |
| 2130111001 | BINTANG SUHARMAN PUTRA | 75 | 75 | 70 | 73 | 68 | 70 | 72 | C |
| 2130111002 | RIKAM PRAJAMUSTI | 75 | 75 | 70 | 73 | 80 | 73 | 74 | C |
| 2130111003 | TASYA AZZAHRA | 0 | 0 | 0 | 0 | 0 | 0 | 0 | E |
| 2130111004 | DIMAS SUROATMOJO | 60 | 75 | 70 | 65 | 80 | 80 | 72 | C |
| 2130111005 | RIZKI RENDI PUTRA | 80 | 75 | 70 | 75 | 73 | 60 | 72 | C |
| 2130111006 | WILDAN HAFIZH APRIANDY HAQ | 80 | 80 | 70 | 75 | 84 | 68 | 76 | B |
| 2130111007 | RUKBANDI SANJAYA | 0 | 0 | 0 | 0 | 0 | 0 | 0 | E |
| 2130111009 | FAHMI IDRIS RAMADHANI | 70 | 75 | 70 | 70 | 84 | 65 | 72 | C |

Data yang sudah dibersihkan berisi data csv dengan entri sebanyak 459 data, kemudian akan dibagi menjadi *data training* sebanyak 260 data dan *data testing* sebanyak 199 data.

### 4.5. Detail Data *Training*

Dataset *training* mencakup 260 entri dan memiliki *10* kolom. Semua kolom dalam dataset ini mengandung 260 nilai non-null, yang menunjukkan tidak ada data yang hilang. Dari sisi tipe data, terdapat 8 kolom bertipe int64 (integer) dan 2 kolom bertipe object (string).

```
Training Data Information:
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 260 entries, 0 to 259
Data columns (total 10 columns):
 #   Column       Non-Null Count  Dtype
---  ------       --------------  -----
 0   NIM          260 non-null    int64
 1   Nama         260 non-null    object
 2   speaking     260 non-null    int64
 3   reading      260 non-null    int64
 4   listening    260 non-null    int64
 5   writing      260 non-null    int64
 6   PTS          260 non-null    int64
 7   PAS          260 non-null    int64
 8   Nilai Angka  260 non-null    int64
 9   Nilai Huruf  260 non-null    object
dtypes: int64(8), object(2)
memory usage: 20.4+ KB
```

Gambar 2. Detail Data Training

### 4.6. Detail Data *Testing*

Dataset *testing* berisi 199 entri dan 10 kolom. Sama *seperti* dataset pelatihan, semua kolom dalam dataset ini memiliki 199 nilai non-null, menunjukkan bahwa tidak ada data yang hilang. Tipe data terdiri dari 8 kolom dengan tipe int64 (integer) dan 2 kolom dengan tipe object (string).

```
Testing Data Information:
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 199 entries, 0 to 198
Data columns (total 10 columns):
 #   Column       Non-Null Count  Dtype
---  ------       --------------  -----
 0   NIM          199 non-null    int64
 1   Nama         199 non-null    object
 2   speaking     199 non-null    int64
 3   reading      199 non-null    int64
 4   listening    199 non-null    int64
 5   writing      199 non-null    int64
 6   PTS          199 non-null    int64
 7   PAS          199 non-null    int64
 8   Nilai Angka  199 non-null    int64
 9   Nilai Huruf  199 non-null    object
dtypes: int64(8), object(2)
memory usage: 15.7+ KB
```

Gambar 3. Detail Data Testing

### 4.7. Standarisasi Data Fitur

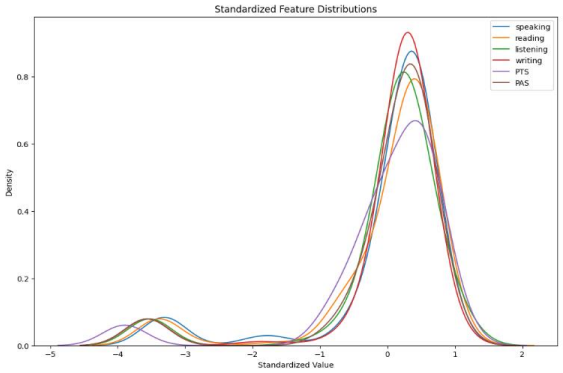
Untuk memastikan semua fitur memiliki skala yang *seragam* (yang penting bagi banyak algoritma *Machine Learning*) data fitur distandarisasi sebelum model dilatih. Proses standarisasi ini menggunakan StandardScaler dari sklearn.

```
from sklearn.preprocessing import
StandardScaler

scaler = StandardScaler()
X_train_scaled =
scaler.fit_transform(X_train)
X_test_scaled =
scaler.transform(X_test)
```

### 4.8. Distribusi Fitur yang Distandarisasi

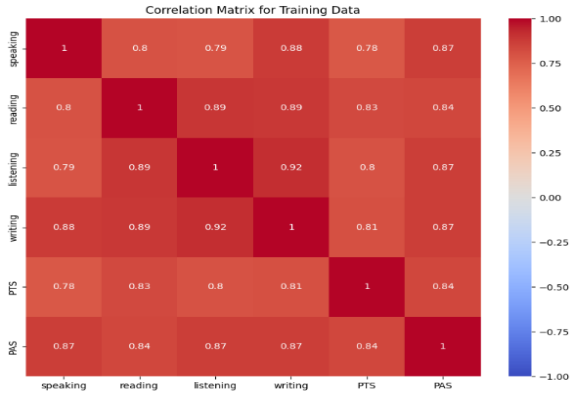
Setelah distandarisasi, distribusi semua fitur *cenderung* simetris dan mendekati normal, dengan puncak di sekitar nilai 0. Fitur-fitur ini memiliki distribusi serupa dengan variasi kecil, menunjukkan bahwa skala fitur telah disamakan dengan baik dan siap digunakan dalam model *Machine Learning*.




Gambar 4. Visualisasi Fitur Standarisasi

### 4.9. Matriks Korelasi

Matriks korelasi visualisasi menunjukkan *hubungan* antar fitur dalam data pelatihan, dengan semua fitur saling berkorelasi positif. Peningkatan pada satu fitur biasanya diikuti oleh peningkatan pada fitur lainnya. Korelasi tertinggi ditemukan antara *'writing'* dan *'listening'*, menandakan bahwa kemampuan menulis dan mendengarkan cenderung berkembang bersama. Korelasi tinggi antara fitur lainnya juga mengindikasikan adanya keterkaitan erat dan kemungkinan faktor-faktor umum yang mempengaruhi kemampuan belajar mahasiswa. Matriks ini penting untuk memahami hubungan antar fitur dan membangun model *Machine Learning* yang efektif.




Gambar 5. Matriks Korelasi

### 4.10. Training Model

*Model* yang digunakan yaitu Naïve Bayes akan dilatih (training) menggunakan data training yang sudah distandarisasi.

```
from sklearn.naive_bayes import
GaussianNB

model = GaussianNB()
model.fit(X_train_scaled, y_train)
```

### 4.11. Penyimpanan Model

*Model* yang sudah dilatih disimpan untuk digunakan dalam website prediksi.

```
import pickle

with open('naivebayes-model.pkl',
'wb') as file:
    pickle.dump((model, scaler),
file)
```

### 4.12. *Evaluation*

Kinerja model akan diukur menggunakan metrik evaluasi seperti akurasi, presisi, *recall*, dan F1-score, untuk menilai seberapa baik model memprediksi kelulusan tes Bahasa Inggris. Evaluasi ini juga membantu memahami kehandalan model dalam membuat prediksi. Model diuji dengan data pengujian, dan hasil prediksi dibandingkan dengan nilai sebenarnya untuk menghitung akurasi serta menghasilkan laporan klasifikasi yang mencakup precision, *recall*, dan F1-score. Berikut hasil evaluasi model:

```
Akurasi model: 0.8793969849246231
Classification Report:
              precision    recall  f1-score   support

           A       1.00      0.86      0.92         7
           B       0.94      0.88      0.91       118
           C       0.76      0.83      0.80        54
           D       0.40      1.00      0.57         2
           E       1.00      1.00      1.00        18

    accuracy                           0.88       199
   macro avg       0.82      0.91      0.84       199
weighted avg       0.89      0.88      0.88       199
```

Gambar 6. Hasil Evaluasi Model

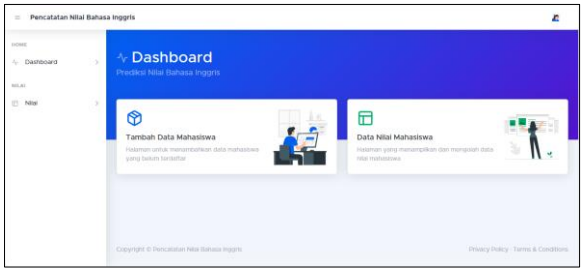
*Model* ini menunjukkan akurasi sebesar 0.88, dengan Macro Average Precision 0.82, *Recall* 0.91, dan F1-Score 0.84. Weighted Average-nya mencatat Precision 0.89, *Recall* 0.88, dan F1-Score 0.88. Menggunakan metode Naive Bayes, model ini mencapai akurasi 87.94% pada data pengujian, terutama efektif untuk memprediksi nilai huruf 'B' dan 'E'. Namun, performa pada kelas 'D' rendah, kemungkinan karena jumlah sampel yang sedikit. Secara keseluruhan, model ini adalah alat yang andal untuk memprediksi kelulusan mahasiswa.

### 4.13. *Deployment*

*Deployment* dilakukan kedalam sebuah website sederhana yang dirancang untuk memprediksi kemungkinan grade kelulusan ujian akhir Bahasa Inggris bagi mahasiswa Universitas Muhammadiyah Sukabumi. Prediksi didasarkan pada nilai-nilai mata kuliah Bahasa Inggris, seperti nilai speaking, reading, listening, writing, PTS, PAS, serta nilai akhir dalam bentuk angka dan huruf.

### 4.14. Halaman Dashboard

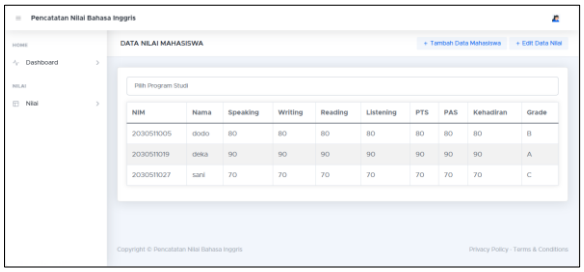
*Halaman* ini merupakan halaman awal *website*.




Gambar 7. Halaman Dashboard

### 4.15. Halaman Data Nilai

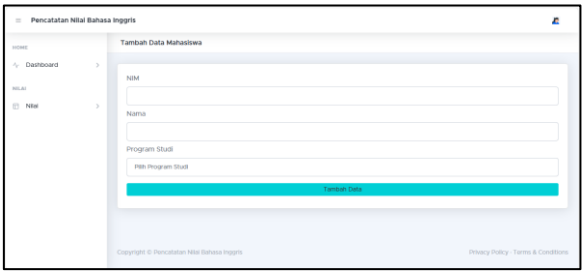
Halaman ini berisi data nilai mahasiswa *dalam* berbagai kategori dan prediksi *grade* tes kemahiran yang didapat.




Gambar 8. Halaman Data Nilai

### 4.16. Halaman Tambah Data Mahasiswa

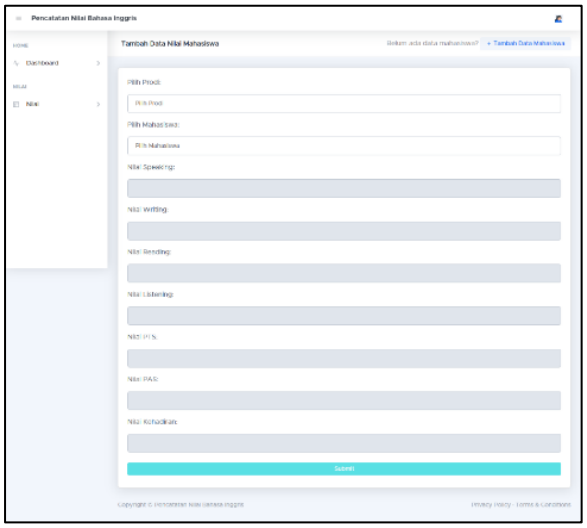
Halaman ini berisi form untuk menambah data mahasiswa.




Gambar 9. Halaman Tambah Data Mahasiswa

### 4.17. Halaman Tambah Data Nilai

*Halaman* ini berisi form untuk mengisi data nilai mahasiswa.




Gambar 10. Halaman Tambah Data Nilai

## 5. KESIMPULAN DAN SARAN

Penelitian ini berhasil mengimplementasikan algoritma Naive Bayes untuk memprediksi kelulusan mahasiswa dalam tes Bahasa Inggris di Universitas Muhammadiyah Sukabumi, menghasilkan model prediktif yang diakses melalui website sederhana. Website ini mencatat nilai Bahasa Inggris mahasiswa dan memprediksi kelulusan mereka. Evaluasi kinerja model dengan metrik akurasi, presisi, *recall*, dan F1-score menunjukkan bahwa model memiliki akurasi 0.88, dengan Macro Average precision 0.82, *recall* 0.91, dan F1-score 0.84, serta Weighted Average precision 0.89, *recall* 0.88, dan F1-score 0.88. Meskipun akurasi model pada data pengujian mencapai 87.94%, performa pada kelas 'D' masih rendah karena jumlah sampel yang sedikit. Secara keseluruhan, model ini efektif dan dapat diandalkan untuk memprediksi kelulusan mahasiswa berdasarkan nilai yang diperoleh.

## DAFTAR PUSTAKA

[1] G. Kruss, S. McGrath, I. Petersen, and M. Gastrow, "Higher education and economic development: The importance of building technological capabilities," *Int. J. Educ. Dev.*, vol. 43, pp. 22–31, 2015.

[2] K. Abdul Kadir and W. S. Wan Mohd Noor, "Students' awareness of the importance of English language proficiency with regard to future employment," *World Rev. Bus. Res.*, vol. 5, no. 3, pp. 259–272, 2015.

[3] V. A. Hernanda, A. Y. Azzahra, and F. Alfarisy, "Pengaruh penerapan bahasa Asing dalam kinerja pendidikan," *J. Indones. Sos. Teknol.*, vol. 3, no. 02, pp. 287–292, 2022.

[4] R. E. Hamel, "The dominance of English in the international scientific periodical literature and the future of language use in science," *Aila Rev.*, vol. 20, no. 1, pp. 53–71, 2007.

[5] R. Roinah, "Penggunaan Bahasa Inggris Pada Masyarakat Ekonomi ASEAN Di Masa Pandemi Covid-19," *J. Cakrawala Ilm.*, vol. 1, no. 12, pp. 3625–3634, 2022.

[6] J. Rymarczyk, "The impact of industrial revolution 4.0 on international trade," *Entrep. Bus. Econ. Rev.*, vol. 9, no. 1, pp. 105–117, 2021.

[7] F. Alfarisy, "Kebijakan Pembelajaran Bahasa Inggris di Indonesia dalam Perspektif Pembentukan Warga Dunia dengan Kompetensi Antarbudaya," *J. Ilm. Profesi Pendidik.*, vol. 6, no. 3, pp. 303–313, 2021, doi: 10.29303/jipp.v6i3.207.

[8] M. R. Aini and P. Nohantiya, "Peningkatan kemampuan bahasa inggris sebagai bahasa kedua bagi siswa desa jatinom," *Jmm (jurnal Masy. mandiri)*, vol. 4, no. 3, pp. 338–347, 2020.

[9] L. Diana, "Hambatan Pembelajaran Bahasa Inggris Mahasiswa Fakultas Pertanian Universitas Pembangunan Nasional Veteran Jawa Timur," *Agridevina Berk. Ilm. Agribisnis*, vol. 7, no. 1, pp. 93–101, 2018.

[10] M. A. Hasanah, S. Soim, and A. S. Handayani, "Implementasi CRISP-DM Model Menggunakan Metode Decision Tree dengan Algoritma CART untuk Prediksi Curah Hujan Berpotensi Banjir," *J. Appl. Informatics Comput.*, vol. 5, no. 2, pp. 103–108, 2021.

[11] Q. Bi, K. E. Goodman, J. Kaminsky, and J. Lessler, "What is machine learning? A primer for the epidemiologist," *Am. J. Epidemiol.*, vol. 188, no. 12, pp. 2222–2239, 2019, doi: 10.1093/aje/kwz189.

[12] H. Li, *Machine Learning Methods*. Haidian: Tsinghua University Press, 2024.

[13] M. K. Sari and I. Wisnubhadra, "Pembangunan Aplikasi Klasifikasi Mahasiswa Baru untuk Prediksi Hasil Studi Menggunakan Naïve Bayes Classifier," pp. 135–142, 2015.

[14] J. A. Nurcahyo and T. B. Sasongko, "Hyperparameter Tuning Algoritma Supervised Learning untuk Klasifikasi Keluarga Penerima Bantuan Pangan Beras," *Indones. J. Comput. Sci.*, vol. 12, no. 3, 2023.

[15] R. G. Wardhana, G. Wang, and F. Sibuea, "Penerapan Machine Learning Dalam Prediksi Tingkat Kasus Penyakit Di Indonesia," *J. Inf. Syst. Manag.*, vol. 5, no. 1, pp. 40–45, 2023.

[16] A. A. Rizal, L. P. I. Kharisma, and F. Fahrurrozi, "Peningkatan Efektifitas Programming Dengan Pelatihan Python for Data Science Bagi Komunitas Programming Pondok Pesantren Nahdlatul Wathan Anjani," *J. Widya Laksmi J. Pengabdi. Kpd. Masy.*, vol. 1, no. 1, pp. 13–19, 2021, doi: 10.59458/jwl.v1i1.3.

[17] I. L. Mulyahati, "IMPLEMENTASI MACHINE LEARNING PREDIKSI HARGA SEWA APARTEMEN MENGGUNAKAN

ALGORITMA RANDOM FOREST MELALUI FRAMEWORK WEBSITE FLASK PYTHON (Studi Kasus: Apartemen di DKI Jakarta Pada Website mamikos.com )," 2020.

[18] P. Kanani and M. Padole, "Deep learning to detect skin cancer using google colab," *Int. J. Eng. Adv. Technol.*, vol. 8, no. 6, pp. 2176–2183, 2019, doi: 10.35940/ijeat.F8587.088619.

[19] R. G. Guntara, "Pemanfaatan Google Colab Untuk Aplikasi Pendeteksian Masker Wajah Menggunakan Algoritma Deep Learning YOLOv7," *J. Teknol. Dan Sist. Inf. Bisnis*, vol. 5, no. 1, pp. 55–60, 2023.

[20] S. Widaningsih, "Perbandingan Metode Data Mining Untuk Prediksi Nilai Dan Waktu Kelulusan Mahasiswa Prodi Teknik Informatika Dengan Algoritma C4,5, Naïve Bayes, Knn Dan Svm," *J. Tekno Insentif*, vol. 13, no. 1, pp. 16–25, 2019, doi: 10.36787/jti.v13i1.78.

[21] Y. Yunus and G. W. Nurcahyo, "Perbandingan algoritma c4. 5 dan naive bayes dalam prediksi kelulusan mahasiswa," *J. CoSciTech (Computer Sci. Inf. Technol.*, vol. 4, no. 1, pp. 193–199, 2023.

[22] Y. Suhanda, I. Kurniati, and S. Norma, "Penerapan Metode Crisp-DM Dengan Algoritma K-Means Clustering Untuk Segmentasi Mahasiswa Berdasarkan Kualitas Akademik," *J. Teknol. Inform. dan Komput.*, vol. 6, no. 2, pp. 12–20, 2020, doi: 10.37012/jtik.v6i2.299.