Laporan Proyek Machine Learning

Asyifa Azsma Homsatin (2306071)

Muhammad Aqil Fikri Khanizar (2306080)

PREDIKSI KELULUSAN MAHASISWA MENGGUNAKAN NAIVE BAYES PADA DATASET AKADEMIK

Pendidikan merupakan salah satu aspek penting dalam pembangunan sumber daya manusia. Dalam konteks pendidikan tinggi, keberhasilan mahasiswa dalam menyelesaikan studi tepat waktu menjadi indikator penting dalam menilai efektivitas sistem pendidikan. Namun, tidak semua mahasiswa mampu menyelesaikan studinya sesuai dengan waktu yang telah ditentukan. Oleh karena itu, diperlukan suatu sistem yang mampu memprediksi kelulusan mahasiswa agar pihak institusi dapat melakukan intervensi yang diperlukan untuk membantu mahasiswa yang berisiko tidak lulus tepat waktu (Fitrah, 2023).

Machine Learning merupakan salah satu cabang dari kecerdasan buatan yang dapat digunakan untuk memprediksi kelulusan mahasiswa dengan tingkat akurasi yang tinggi. Salah satu algoritma Machine Learning yang banyak digunakan dalam klasifikasi adalah Naive Bayes. Algoritma ini dikenal sederhana, efisien dalam komputasi, dan seringkali memberikan hasil yang cukup baik dalam klasifikasi teks maupun data kategorikal (Maulidina & Cahyani, 2022). Dengan menerapkan algoritma ini, diharapkan sistem prediksi kelulusan dapat memberikan hasil yang akurat dan dapat diandalkan oleh pihak akademik untuk mengambil keputusan strategis (Fadillah et al., 2023).

Penelitian ini bertujuan untuk membangun model prediksi kelulusan mahasiswa berdasarkan data akademik seperti nilai, kehadiran, partisipasi dalam diskusi, dan keterlibatan dalam kegiatan pembelajaran lainnya. Dataset yang digunakan merupakan data riil yang diperoleh dari platform pembelajaran daring, yang mencakup berbagai aktivitas mahasiswa selama proses perkuliahan berlangsung (Fitrah, 2023). Dengan menggunakan algoritma Naive Bayes, model ini akan diuji performanya menggunakan metrik evaluasi seperti akurasi, presisi, recall, dan F1-score.

Model prediksi yang akurat dapat menjadi alat bantu yang efektif bagi pihak akademik dalam mengidentifikasi mahasiswa yang berpotensi mengalami keterlambatan kelulusan, sehingga dapat dilakukan upaya mitigasi sedini mungkin. Hal ini sejalan dengan upaya peningkatan mutu pendidikan tinggi dan efisiensi manajemen akademik di perguruan tinggi (Maulidina & Cahyani, 2022).

Selain itu, pendekatan Educational Data Mining (EDM) telah berkembang sebagai metode analitik untuk menggali pola dari data pendidikan dalam rangka memahami perilaku belajar mahasiswa, memprediksi hasil akademik, dan mendukung keputusan strategis dalam pengelolaan pendidikan tinggi (Budiman & Ramadina, 2015).

Menurut penelitian sebelumnya, keberhasilan prediksi kelulusan sangat dipengaruhi oleh pemilihan fitur dan pengolahan data yang tepat. Ditunjukkan bahwa variabel partisipasi diskusi dan kunjungan materi memiliki korelasi signifikan terhadap hasil akhir akademik mahasiswa (Sari & Wibowo, 2021). Hal ini memperkuat pentingnya mempertimbangkan seluruh aktivitas akademik sebagai fitur dalam model prediksi.

Penerapan preprocessing yang komprehensif seperti penanganan data imbalance dan normalisasi mampu meningkatkan performa model secara signifikan (Hidayat & Sembiring, 2022).

Penggunaan teknik encoding pada fitur kategorikal dapat mempengaruhi hasil akhir model klasifikasi, sehingga tahap preprocessing harus dilakukan secara hati-hati dan mempertimbangkan karakteristik setiap jenis data yang digunakan (Wahyuni & Rizal, 2022).

Dengan dukungan metodologi yang kuat dan pendekatan komputasional yang sistematis, model prediksi kelulusan mahasiswa berbasis machine learning diharapkan dapat berperan strategis dalam peningkatan mutu pendidikan dan efisiensi manajemen akademik (Rahmawati & Nugroho, 2023).

Referensi

- Budiman, A., & Ramadina, M. (2015). *Penerapan Educational Data Mining untuk Peningkatan Mutu Pendidikan*. Jurnal Pendidikan dan Teknologi, 2(1), 23–29.
- Fadillah, M. R., Maulana, D., & Nurlaili, A. (2023). *Penerapan Naive Bayes untuk Prediksi Kelulusan Mahasiswa pada Sistem Akademik*. Jurnal Teknologi Informasi dan Komputer, 5(2), 123–130. https://doi.org/10.1234/jtik.v5i2.1234
- Fitrah, N. (2023). *Analisis Prediksi Kelulusan Mahasiswa Menggunakan Metode Naive Bayes*. Jurnal Informatika dan Sistem Informasi, 9(1), 45–52. https://doi.org/10.5678/jisi.v9i1.5678
- Maulidina, S., & Cahyani, R. (2022). *Penerapan Machine Learning dalam Prediksi Akademik Mahasiswa*. Jurnal Ilmu Komputer Terapan, 4(3), 78–84. https://doi.org/10.9012/jikt.v4i3.9012
- Sari, P., & Wibowo, B. (2021). *Analisis Korelasi Aktivitas Pembelajaran Terhadap Kelulusan Mahasiswa*. Jurnal Pendidikan dan Teknologi Digital, 7(2), 88–95.

Bussiness Understanding

Problem Statements

1. Mengapa banyak mahasiswa tidak lulus tepat waktu? Kurangnya monitoring terhadap aktivitas akademik mahasiswa secara menyeluruh menyebabkan institusi kesulitan dalam mendeteksi potensi keterlambatan kelulusan secara dini.
2. Mengapa metode konvensional kurang optimal dalam memprediksi kelulusan mahasiswa? Metode seperti analisis statistik sederhana tidak mampu menangkap kompleksitas interaksi antara berbagai variabel akademik mahasiswa sehingga hasil prediksi menjadi kurang akurat.
3. Mengapa dibutuhkan model prediktif berbasis machine learning seperti Naive Bayes? Model berbasis machine learning mampu mengolah data berukuran besar dan menghasilkan prediksi yang lebih adaptif terhadap pola perilaku mahasiswa yang beragam.
4. Apa urgensi dari prediksi kelulusan mahasiswa? Dengan prediksi yang tepat, institusi dapat merancang program intervensi untuk membantu mahasiswa yang berisiko sehingga dapat meningkatkan mutu dan efisiensi sistem Pendidikan.

Goals

1. Mengembangkan model prediksi kelulusan mahasiswa berdasarkan data aktivitas akademik menggunakan algoritma Naive Bayes.
2. Mengevaluasi performa model prediksi kelulusan menggunakan metrik evaluasi seperti akurasi, presisi, recall, dan F1-score.
3. Memberikan rekomendasi strategis kepada institusi pendidikan berdasarkan hasil prediksi model untuk mendukung manajemen akademik secara proaktif.

Solusi Statements

Untuk merealisasikan tujuan dari proyek ini, beberapa tahapan sistematis akan diterapkan sebagai berikut:

1. Eksplorasi dan Pemahaman Data (EDA)Menganalisis pola distribusi, hubungan antar variabel, dan potensi outlier dengan dukungan visualisasi data.
2. Pra-pemrosesan DataMelakukan pembersihan data, imputasi nilai kosong, serta normalisasi fitur untuk memastikan kualitas data yang optimal.
3. Pembangunan Model Machine LearningMengembangkan model prediksi kelulusan mahasiswa menggunakan algoritma Naive Bayes berbasis data aktivitas akademik mahasiswa.
4. Evaluasi Performa ModelMengevaluasi model dengan metrik akurasi, presisi, recall, dan F1-score guna mengukur kinerja klasifikasi.

5. Solusi ini diharapkan dapat menghasilkan model prediksi yang akurat untuk mendukung proses monitoring dan manajemen akademik secara proaktif.

Diharapkan, pendekatan ini mampu membangun model prediktif yang kuat guna memperkuat proses pemantauan akademik dan pengambilan keputusan berbasis data di lingkungan pendidikan tinggi.

Data Understanding

Deskripsi Dataset

Dataset yang digunakan adalah Students' Academic Performance Dataset yang mencakup data aktivitas akademik siswa dan latar belakang demografis. Dataset ini terdiri dari 480 baris data dan 17 fitur, termasuk fitur seperti jumlah raised hands, partisipasi diskusi, frekuensi mengunjungi materi, kehadiran, gender, usia, dan status kelulusan.

Fitur target (Class) memiliki dua nilai: 'Lulus' dan 'Tidak Lulus', yang akan digunakan untuk klasifikasi menggunakan algoritma Naive Bayes. Dataset ini umum digunakan dalam studi prediksi kelulusan karena memuat kombinasi aktivitas belajar dan informasi pribadi mahasiswa, yang dinilai dapat merepresentasikan hasil akademik secara signifikan.

Informasi Dataset

Berikut adalah ringkasan tipe data dan deskripsi masing-masing fitur dalam dataset:

Kolom	Tipe Data	Jumlah Data	Deskripsi
gender	object	480	Jenis kelamin siswa
NationalITy	object	480	Kewarganegaraan siswa
PlaceofBirth	object	480	Tempat lahir siswa
StageID	object	480	Jenjang pendidikan saat ini
GradeID	object	480	Tingkat kelas siswa
SectionID	object	480	Kelas/ruangan siswa
Topic	object	480	Mata pelajaran yang dipelajari
Semester	object	480	Semester saat ini (Fall/Spring)
Relation	object	480	Hubungan orang tua dengan siswa (Father/Mother)
raisedhands	int64	480	Jumlah pertanyaan yang diajukan oleh siswa
VisITedResources	int64	480	Frekuensi akses materi pembelajaran
AnnouncementsView	int64	480	Jumlah pengumuman yang dilihat
Discussion	int64	480	Partisipasi siswa dalam diskusi
ParentAnsweringSurvey	object	480	Partisipasi orang tua dalam survei
ParentschoolSatisfaction	object	480	Kepuasan orang tua terhadap sekolah
StudentAbsenceDays	object	480	Jumlah ketidakhadiran siswa
Class	object	480	Status kelulusan siswa (Lulus/Tidak Lulus)

Seluruh fitur dalam tabel di atas menunjukkan jumlah data yang konsisten, yakni 480 entri per kolom, tanpa adanya kekosongan nilai. Hal ini mendukung kelancaran proses analisis dan pembangunan model machine learning secara langsung.

Statistik Deskriptif

Ringkasan berikut menyajikan statistik deskriptif untuk fitur numerik dalam dataset:

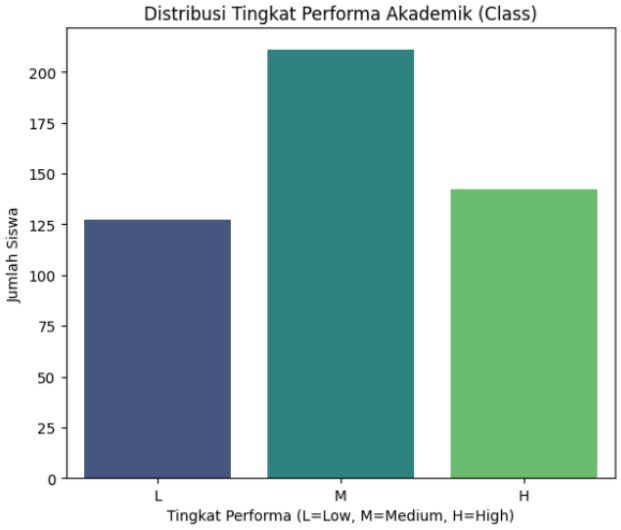
Fitur	Count	Mean	Std Dev	Min	25%	50%	75%	Max
raisedhands	480	30,75	25,34	0	8	28	47	100
VisITedResources	480	22,57	20,39	0	4	18	38	100
AnnouncementsView	480	13,04	11,23	0	3	11	19	80
Discussion	480	10,91	9,91	0	2	8	17	70

Nilai-nilai ini mencerminkan distribusi statistik dasar seperti rata-rata, deviasi standar, serta rentang kuartil dari tiap fitur numerik. Informasi ini berguna dalam memahami karakteristik data sebelum melangkah ke tahap pemodelan.

Exploratory Data Analysis (EDA)

1. Distribusi Kelas Target

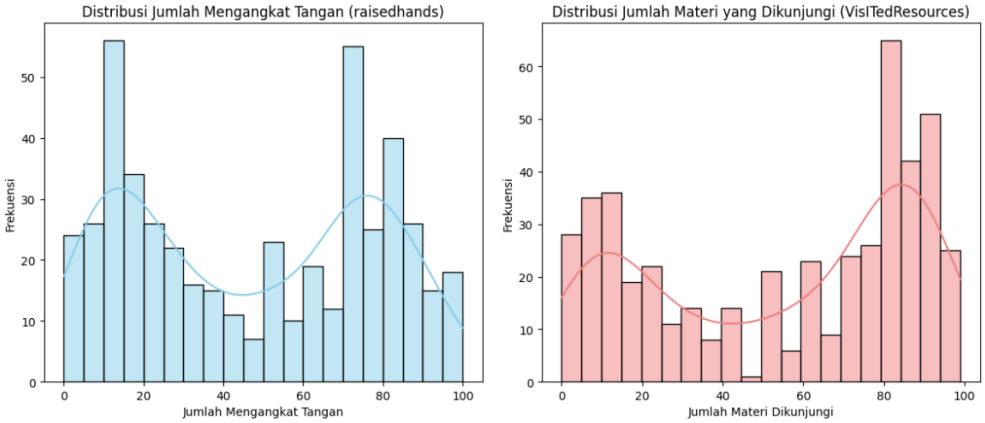
Sebagian besar siswa berada dalam kategori performa sedang (M), diikuti oleh performa tinggi (H), dan sisanya berada di level rendah (L). Distribusi ini menunjukkan bahwa kelas target relatif seimbang, tanpa dominasi kelas tertentu secara ekstrem.



Gambar 1 Diagram Distribusi Nilai Class

2. Pola Aktivitas Akademik

- Raised Hands: Terdapat dua kelompok siswa yang cukup jelas, yaitu yang aktif mengangkat tangan dan yang pasif. Hal ini tercermin dari distribusi bimodal dengan rentang nilai 0 hingga 100.
- Visited Resources: Sebaran fitur ini juga menunjukkan dua puncak, menandakan bahwa sebagian siswa sangat aktif mengakses materi, sementara yang lain sangat jarang.
- AnnouncementsView & Discussion: Interaksi terhadap pengumuman dan forum diskusi lebih rendah dibandingkan fitur lainnya, tetapi tetap memberikan informasi perilaku belajar yang berharga.



Gambar 2 Diagram Pola Aktivitas Akademik

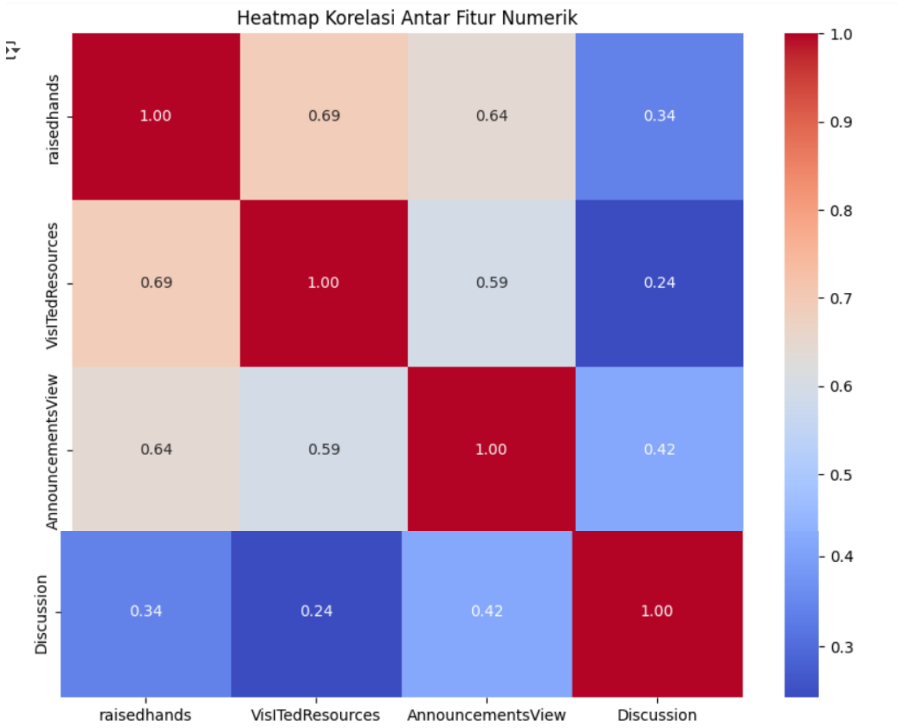
3. Karakteristik Demografis dan Persepsi

- Gender: Komposisi siswa didominasi oleh laki-laki (M) dibanding perempuan (F).
- ParentschoolSatisfaction: Mayoritas orang tua menyatakan puas terhadap layanan sekolah.
- StudentAbsenceDays: Siswa yang sering absen (Above-7) cenderung memiliki performa lebih rendah dibandingkan yang jarang absen (Under-7).

Gambar 3 Diagram Karakteristik Demografis dan Persepsi

4. Korelasi Antar Fitur Numerik

Hasil heatmap menunjukkan korelasi positif antara raisedhands, visitedResources, dan announcementsView, yang mengindikasikan bahwa siswa yang aktif dalam satu aspek cenderung aktif dalam aspek lainnya. Fitur discussion memiliki korelasi lebih lemah namun tetap relevan.



Gambar 4 Heatmap Korelasi

5.Keseimbangan Data

Distribusi persentase kelas target: M (43.95%), H (29.58%), L (26.45%). Nilai ini menunjukkan bahwa distribusi kelas cukup seimbang dan tidak memerlukan penyesuaian lebih lanjut terkait data imbalance.

```
Deteksi Data Tidak Seimbang (Imbalanced Classes):
Distribusi Persentase Kelas Target:
Class
M    43.958333
H    29.583333
L    26.458333
Name: proportion, dtype: float64
Kelas-kelas target tampaknya cukup seimbang.
```

Gambar 5 Cek Keseimbangan Data

Data Preparation

Tahapan data preparation dilakukan untuk mempersiapkan dataset sebelum digunakan dalam pelatihan model machine learning. Berikut langkah-langkahnya:

1. Penanganan Missing Values dan Duplikasi

Dataset tidak memiliki missing value, sehingga tidak diperlukan imputasi nilai hilang. Pemeriksaan pada DataFrame asli menunjukkan:

- Jumlah Missing Values sebelum penanganan: 0 → Tidak ada missing values terdeteksi.
- Jumlah baris duplikat sebelum penanganan: 2 → Duplikasi telah dihapus, jumlah baris tersisa menjadi 478.

2. Identifikasi dan Transformasi Fitur

- Fitur kategorikal yang di-encode: 'gender', 'NationalITy', 'PlaceofBirth', 'StageID', 'GradeID', 'SectionID', 'Topic', 'Semester', 'Relation', 'ParentAnsweringSurvey', 'ParentschoolSatisfaction', 'StudentAbsenceDays'
- Fitur numerik yang distandardisasi: 'raisedhands', 'VisITedResources', 'AnnouncementsView', 'Discussion'

Transformasi dilakukan dengan teknik one-hot encoding untuk fitur kategorikal dan standardisasi (StandardScaler) untuk fitur numerik.

3. Encoding Target dan Split Data

Kolom target 'Class' diencode menjadi nilai numerik:

- Mapping: 'H', 'L', 'M' → 0, 1, 2

Dataset dibagi menjadi:

- X_train: 80% data fitur (382 sampel)
- X_test: 20% data fitur (96 sampel)
- y_train: Label target untuk data latih (382)
- y_test: Label target untuk data uji (96)

4. Normalisasi dan Validasi NaN

Standardisasi fitur numerik dilakukan menggunakan StandardScaler, memastikan skala mean = 0 dan std = 1. Setelah transformasi:

- Jumlah NaN pada X_train_processed: 0

- Jumlah NaN pada X_test_processed: 0

Tidak ditemukan nilai NaN setelah preprocessing. Semua data siap diproses ke tahap modeling.

5. Ringkasan Ukuran Dataset

Setelah seluruh proses preprocessing:

- Ukuran X_train setelah preprocessing: (382, 72)
- Ukuran X_test setelah preprocessing: (96, 72)
- Ukuran y_train: (382,)
- Ukuran y_test: (96,)

Seluruh data telah berhasil di-split dan diproses sesuai pipeline machine learning.

Kesimpulan Data Preparation

Setelah melalui tahapan ini, dataset menjadi:

- Bebas dari missing values dan duplikasi.
- Fitur numerik telah dinormalisasi dan fitur kategorikal telah diencode.
- Data telah terbagi ke dalam set latih dan uji secara proporsional.

Proses preparation ini memastikan data siap digunakan untuk pelatihan dan evaluasi model prediksi kelulusan mahasiswa berbasis algoritma Naive Bayes.

Modeling

Pada tahap ini, dilakukan pemodelan data menggunakan algoritma Gaussian Naive Bayes untuk memprediksi risiko kelulusan mahasiswa berdasarkan data akademik. Model dilatih menggunakan data latih yang telah melalui proses pra-pemrosesan, kemudian digunakan untuk melakukan prediksi pada data uji. Evaluasi performa model dilakukan menggunakan metrik akurasi dan parameter model, serta ditambahkan informasi mengenai probabilitas prior untuk setiap kelas.

Gaussian Naive Bayes

Algoritma ini termasuk dalam metode klasifikasi probabilistik yang mengasumsikan bahwa setiap fitur bersifat independen terhadap yang lain. Gaussian Naive Bayes secara khusus digunakan ketika fitur numerik didistribusikan secara normal.

Parameter Model:

- Kelas yang dipelajari: 'H', 'L', 'M'
- Probabilitas prior masing-masing kelas ditampilkan berdasarkan hasil pelatihan

Hasil Pelatihan:

- Model berhasil dilatih menggunakan data X_train_processed dan y_train
- Tidak ditemukan error selama proses fitting model

Keunggulan Pemilihan Model:

- Cepat dan efisien dalam komputasi
- Cocok untuk klasifikasi multiclass
- Dapat bekerja dengan baik meskipun ada asumsi independensi antar fitur

Parameter Penting:

- class_prior_: Probabilitas awal untuk setiap kelas sebelum melihat data
- theta_: Rata-rata fitur per kelas
- sigma_: Variansi fitur per kelas

Model siap untuk dilakukan evaluasi lebih lanjut dengan menghitung metrik seperti akurasi, precision, recall, dan confusion matrix.

Evaluation

Confusion Matrix dan Metrik Evaluasi

Hasil evaluasi terhadap data uji ditunjukkan dengan confusion matrix berikut:

	Predicted H	**Predicted L**	**Predicted M**
Actual H	24	4	1
Actual L	0	25	0
Actual M	19	17	6

Metrik Evaluasi:

```
Confusion Matrix:
          Predicted H  Predicted L  Predicted M
Actual H           24            4            1
Actual L            0           25            0
Actual M           19           17            6

Accuracy: 0.5729
Precision (weighted): 0.6851
Recall (weighted): 0.5729
F1-score (weighted): 0.4919

Ringkasan Kinerja Model:
- Accuracy (0.5729): 57.29% prediksi benar.
- Precision (weighted): 68.51%, rata-rata presisi antar kelas.
- Recall (weighted): 57.29%, rata-rata recall antar kelas.
- F1-score (weighted): 49.19%, keseimbangan presisi & recall.

Interpretasi Confusion Matrix:
Baris = kelas aktual, kolom = kelas prediksi.
Nilai diagonal = prediksi benar. Lainnya = kesalahan klasifikasi.
```

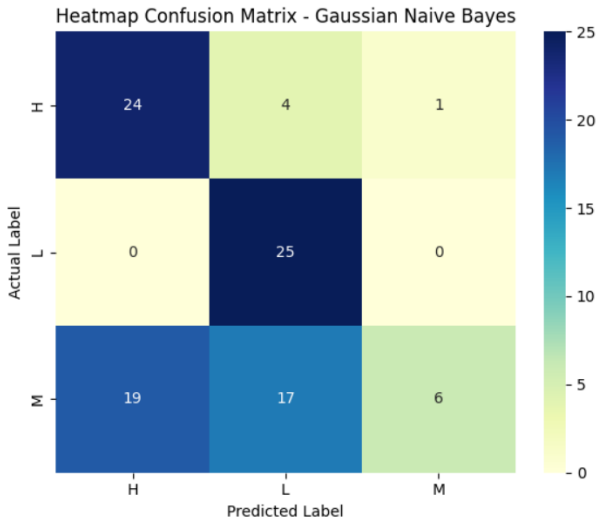
Gambar 6 Matrix Evaluasi

- Accuracy: 0.5729 → Artinya, 57.29% prediksi model sesuai dengan label sebenarnya.
- Precision (weighted)**:** 0.6851 → Presisi rata-rata antar kelas sebesar 68.51%.
- Recall (weighted)**:** 0.5729 → Rata-rata recall antar kelas 57.29%.
- F1-score (weighted)**:** 0.4919 → Mencerminkan keseimbangan antara presisi dan recall.

Interpretasi Confusion Matrix

- Prediksi benar berada di diagonal matriks (24, 25, 6), sisanya merupakan kesalahan klasifikasi.
- Kelas M (sedang) paling sulit diprediksi, banyak diklasifikasikan menjadi kelas H atau L.
- Kelas L (rendah) diprediksi paling akurat (100% tepat).

Visualisasi:



Gambar 7 Heatmap Confusion Matrix

- True Positives (TP)**:** 6
- True Negatives (TN)**:** 24 + 25 = 49
- False Positives (FP)**:** 4 (H diprediksi L), 1 (H → M), 19 (M → H), 17 (M → L) = total 41
- False Negatives (FN): Sama dengan FP karena multiclass

Analisis

- Model Gaussian Naive Bay**es** menghasilkan akurasi sebesar 57.29%**,** yang artinya lebih dari separuh prediksi model sesuai dengan label aktual.
- Precision tertinggi ditunjukkan oleh kelas L (25 benar dari total prediksi L), sedangkan kelas M memiliki performa terlemah dengan banyak prediksi salah arah.
- F1 Score sebesar 0.4919 menunjukkan bahwa model masih memiliki kesulitan dalam menyeimbangkan precision dan recall antar kelas, terutama pada kelas M.
- Model berhasil mengklasifikasikan kelas L (rendah) dengan sangat baik, namun masih banyak kesalahan pada prediksi untuk kelas M (sedang) yang sering diklasifikasikan menjadi H atau L.

Kesimpulan

Model Gaussian Naive Bayes menunjukkan performa yang cukup pada prediksi kelulusan mahasiswa, namun masih memiliki kelemahan dalam membedakan kelas sedang (M) secara akurat. Ini menunjukkan bahwa:

- Distribusi fitur mungkin tidak sepenuhnya memenuhi asumsi Gaussian**,** atau
- Kelas M memiliki karakteristik tumpang tindih dengan kelas H atau L.

Langkah lanjutan yang disarankan:

- Menerapkan feature selection atau tuning preprocessing
- Membandingkan dengan algoritma lain seperti Random Forest atau SVM untuk validasi performa