Penerapan *Machine Learning* Dalam Klasterisasi Perguruan Tinggi Berdasarkan *Tuition Fee* Dan *Living Cost* Mahasiswa Internasional Menggunakan Metode *Gaussian Mixture Models*

SKRIPSI

DISUSUN OLEH

NADYA AULYA PUTRI

2209020054

PROGRAM STUDI TEKNOLOGI INFORMASI

FAKULTAS ILMU KOMPUTER DAN TEKNOLOGI INFORMASI

UNIVERSITAS MUHAMMADIYAH SUMATERA UTARA

MEDAN

2026

Penerapan *Machine Learning* Dalam Klasterisasi Perguruan Tinggi Berdasarkan *Tuition Fee* Dan *Living Cost* Mahasiswa Internasional Menggunakan Metode *Gaussian Mixture Models*

SKRIPSI

Diajukan sebagai salah satu syarat untuk memperoleh gelar Sarjana Komputer (S. Kom) dalam Program Studi Tenologi Infromasi. pada Fakultas Ilmu Komputer dan Teknologi Informasi, Universitas Muhammadiyah Sumatera Utara.

NADYA AULYA PUTRI
2209020054

PROGRAM STUDI TEKNOLOGI INFROMASI
FAKULTAS ILMU KOMPUTER DAN TEKNOLOGI INFORMASI
UNIVERSITAS MUHAMMADIYAH SUMATERA UTARA
MEDAN
2026

LEMBAR PENGESAHAN

Judul Skripsi	:	Penerapaan Machine Learning Dalam Klasterisasi Perguruan Tinggi Berdasarkan Tuition Fee Dan Living Cost Mahasiswa Internasional Menggunakan Metode Gaussian Mixture Models
Nama Mahasiswa	:	NADYA AULYA PUTRI
NPM	:	2209020054
Program Studi	:	TEKNOLOGI INFORMASI

Menyetujui
Komisi Pembimbing

(Hevlie Winda Nazry S, S. Pd., M. Si)
NIDN. 0129079301

Ketua Program Studi

(Fatma Sari Hutagalung, S.Kom, M.Kom)
NIDN. 0117019301

Dekan

(Dr. Al-Khowarizmi, S.Kom., M.Kom.)
NIDN. 0127099201

PERNYATAAN ORISINALITAS

Penerapan *Machine Learning* Dalam Klasterisasi Perguruan Tinggi Berdasarkan *Tuition Fee* Dan *Living Cost* Mahasiswa Internasional Menggunakan Metode *Gaussian Mixture Models*

SKRIPSI

Saya menyatakan bahwa karya tulis ini adalah hasil karya sendiri, kecuali beberapa kutipan dan ringkasan yang masing-masing disebutkan sumbernya.

Medan, 02 April 2026

Yang membuat pernyataan

Nadya Aulya Putri

NPM. 2209020054

PERNYATAAN PERSETUJUAN PUBLIKASI

Sebagai sivitas akademika Universitas Muhammadiyah Sumatera Utara, saya bertanda tangan dibawah ini:

Nama	: Nadya Aulya Putri
NPM	: 2209020054
Program Studi	: Teknologi Informasi
Karya Ilmiah	: Skripsi

Demi pengembangan ilmu pengetahuan, menyetujui untuk memberikan kepada Universitas Muhammadiyah Sumatera Utara Hak Bedas Royalti Non-Eksekutif *(Non-Exclusive Royalty free Right)* atas penelitian skripsi saya yang berjudul:

Penerap Penerapan *Machine Learning* Dalam Klasterisasi Perguruan Tinggi Berdasarkan *Tuition Fee* Dan *Living Cost* Mahasiswa Internasional Menggunakan Metode *Gaussian Mixture Models*

Beserta perangkat yang ada (jika diperlukan). Dengan Hak Bebas Royalti Non-Eksekutif ini, Universitas Muhammadiyah Sumatera Utara berhak menyimpan, mengalih media, memformat, mengelola dalam bentuk database, merawat dan mempublikasikan Skripsi saya ini tanpa meminta izin dari saya selama tetap mencantumkan nama saya sebagai penulis dan sebagai pemegang dan atau sebagai pemilik hak cipta.

Demikian pernyataan ini dibuat dengan sebenarnya.

Medan, 02 April 2026

Yang membuat pernyataan

Nadya Aulya Putri

2209020054

RIWAYAT HIDUP

DATA PRIBADI

Nama Lengkap	: Nadya Aulya Putri
Tempat dan Tanggal Lahir	: TanjungBalai, 03 November 2002
Alamat Rumah	: Jl. Singosari Lk. III
Telepon/Faks/HP	: 0852-7757-5159
E-mail	: nadyaaulyaputri66@gmail.com
Instansi Tempat Kerja	: -
Alamat Kantor	: -

DATA PENDIDIKAN

SD	: SD Negeri 132405 Kota Tanjungbalai	TAMAT: 2015
SMP	: SMP Negeri 10 Kota Tanjungbalai	TAMAT: 2018
SMA	: SMA Negeri 2 Kota Tanjungbalai	TAMAT: 2021

KATA PENGANTAR

Pendahuluan

Penulis tentunya berterima kasih kepada berbagai pihak dalam dukungan serta doa dalam penyelesaian skripsi. Penulis juga mengucapkan terima kasih kepada:

1. Allah SWT, karena berkat limpahan rahmat dan karunia-Nya, penulis dapat menyelesaikan skripsi yang berjudul "Penerapan Machine Learning Dalam Klasterisasi Perguruan Tinggi Berdasarkan Tuition Fee Dan Living Cost Mahasiswa Internasional Menggunakan Metode Gaussian Mixture Models".
2. Bapak Prof. Dr. Akrim, M.Pd., Rektor Universitas Muhammadiyah Sumatera Utara (UMSU)
3. Bapak Dr. Al-Khowarizmi, S.Kom., M.Kom. Dekan Fakultas Ilmu Komputer dan Teknologi Informasi (FIKTI) UMSU.
4. Ibu Dr. Firahmi Rizky, M.Kom, selaku Wakil Dekan I Ilmu Komputer dan Teknologi Informasi
5. Bapak Mhd. Basri, S.Si., M.Kom, selalu Wakil Dekan III Ilmu Komputer dan Teknologi Informasi
6. Ibu Fatma Sari Hutagalung, S.Kom., M.Kom Ketua Program Studi Teknologi Informasi
7. Bapak Okvi Nugroho, S.Kom., M.Kom Sekretaris Program Studi

8. Dosen Pembimbing Ibu Hevlie Winda Nazry S, S.Pd., M.Si. yang telah membimbing dan memberikan arahan kepada penulis sehingga penelitian ini dapat diselesaikan dengan baik
9. Teruntuk Ayah penulis, terimakasih penulis ucapkan karna telah memberikan yang terbaik, menjadi sosok yang tak kenal lelah mendoakan, mengusahakan dan memberikan dukungan baik secara moral maupun finansial kepada penulis.
10. Teruntuk pintu surgaku tersayang, terimakasih karna telah memilih menjadi sosok ibu yang selalu kuat, yang tak pernah menyerah, terimakasih karna telah mewujudkan sosok ibu yang sempurna bagi penulis, terimakasih karna selalu menyemangati penulis disaat penulis ingin menyerah, terimakasih karna telah menjadikan rumah sebagai tempat terhangat bagi penulis untuk pulang, terimakasih atas setiap pengorbanan, semangat yang tak pernah lelah, dan terimakasih karna selalu melangitkan doa untuk penulis.
11. Teruntuk kakak dan abang penulis, Nanda Julia, Zira Adhani, Reza Agung dan Dikki Kurniawan, terimakasih karna telah menjadi sosok teladan bagi penulis, terimakasih karna selalu menjadi tempat pulang bagi penulis disaat penulis tidak ingin bertemu siapa pun, terimakasih karna selalu sabar menghadapi tingkah laku penulis, terimakasih karna telah memberikan dukungan secara moral dan finansial selama penulis menjalankan studi ini hingga selesai.
12. Teruntuk teman teman seperjuangan, Gaizka Pasya Dermawan, Siti Nurisma Siregar, Aida Fadhila, Nabila Azura Putri, Syafa Takbira Aisafitri, Fauzi Lativah, Az-zahra Natasya, M. Ariq Adrian, dan Trifahmi Rivaldo.

Terimakasih atas setiap waktu yang diluangkan, memberikan dukungan, motivasi, semangat serta menjadi rekan yang membersamai penulis sampai perkuliahan ini selesai.

13. Teruntuk adik penulis, Anggi Yolanda Galingging, terimakasih telah menjadi adik yang selalu mendengar keluh kesah hidup penulis, terimakasih karna selalu mendukung penulis dalam keadaan apapun, terimakasih karna tidak pernah menghakimi keputusan penulis, meskipun tidak memiliki aliran darah yang sama, terimakasih karna telah tumbuh dan besar menjadi sosok yang kuat dan penyayang, semoga kita bisa menyaksikan TDS bersama sama.
14. Teruntuk seluruh keluarga besar B1 TI 22, baik yang dapat disebutkan maupun tidak terimakasih atas momen indah yang pernah dilalui bersama.

Penerapan *Machine Learning* Dalam Klasterisasi Perguruan Tinggi Berdasarkan *Tuition Fee* Dan *Living Cost* Mahasiswa Internasional Menggunakan Metode *Gaussian Mixture Models*

ABSTRAK

Tingginya variasi tuition fee dan living cost antar perguruan tinggi di berbagai negara menjadikan proses pengambilan keputusan bagi calon mahasiswa internasional semakin kompleks. Penelitian ini menggunakan dataset dari Kaggle dengan total 1.408 data. Tahapan penelitian meliputi pengumpulan data, preprocessing dengan normalisasi Z-Score, penentuan jumlah klaster optimal menggunakan kriteria Bayesian Information Criterion (BIC) dan Akaike Information Criterion (AIC), serta estimasi parameter melalui algoritma Expectation-Maximization (EM). Penerapan metode GMM didasarkan pada karakteristiknya yang mampu memodelkan distribusi data yang kompleks dan multivariat melalui pendekatan probabilistik. Hasil penelitian menunjukkan bahwa model GMM dengan K=7, cukup efektif dalam mengidentifikasi tujuh kelompok perguruan tinggi dengan profil biaya yang berbeda dan bermakna, dengan model mencapai konvergensi pada iterasi ke-139. Temuan ini menunjukkan bahwa pendekatan soft clustering berbasis GMM cukup memadai dalam menemukan pola tersembunyi pada data biaya mahasiswa internasional yang bersifat right-skewed dan tidak simetris.

Kata Kunci*: Gaussian Mixture Models, Machine Learning, Tuition Fee & Living Cost, Mahasiswa Internasional, Bayesian Information Criterion*.

Application of Machine Learning in the Clustering of Higher Education Institutions Based on Tuition Fees and Living Costs of International Students Using the Gaussian Mixture Models Method

ABSTRACT

The wide variation in tuition fees and living costs among higher education institutions across different countries makes it increasingly difficult for prospective international students to make informed decisions. This study used a dataset from Kaggle, with a total of 1,408 records. The research steps included data collection, preprocessing using Z-Score normalization, determining the optimal number of clusters based on the Bayesian Information Criterion (BIC) and Akaike Information Criterion (AIC), and estimating model parameters using the Expectation-Maximization (EM) algorithm. The GMM method was applied due to its ability to model complex and multivariate data distributions through a probabilistic approach. The results show that the GMM model with K=7 was reasonably effective in identifying seven groups of higher education institutions with distinct and meaningful cost profiles, with the model reaching convergence at iteration 139. These results suggest that the soft clustering approach of GMM is adequately capable of uncovering latent patterns in international student cost data that is right-skewed and asymmetric.

Keywords*: Gaussian Mixture Models, Machine Learning, Tuition Fee & Living Cost, International Students, Bayesian Information Criterion,*

DAFTAR ISI

DAFTAR TABEL

DAFTAR GAMBAR

BAB I

PENDAHULUAN

1.1 Latar Belakang Penelitian

Perkembangan pendidikan global telah mendorong peningkatan mobilitas mahasiswa lintas negara secara signifikan. Berdasarkan data UNESCO institute for statistic jumlah mahasiswa internasional di seluruh dunia telah mencapai 6,9 juta jiwa pada tahun 2022, meningkat sebesar 176% selama dua dekade terakhir dibandingkan angka 2,5 juta jiwa pada tahun 2002. Tiga negara tujuan utama, yaitu Amerika Serikat, Kanada, dan Inggris, menerima hampir 39% dari total mahasiswa internasional secara global, dengan Amerika Serikat mencatat lebih dari 1,1 juta mahasiswa asing pada tahun akademik 2023-2024 (UNESCO, 2024). Berdasarkan laporan (OECD, 2025), mahasiswa dari kawasan Asia mendominasi arus mobilitas internasional dengan mencakup 58% dari total mahasiswa asing yang terdaftar di negara-negara OECD pada tahun 2023, di mana China dan India menjadi dua negara pengirim mahasiswa internasional terbesar di dunia.

Salah satu kesulitan yang dihadapi mahasiswa internasional adalah kompleksitas dalam memperkirakan dan merencanakan biaya studi di luar negeri. Biaya tersebut mencakup dua komponen utama, yaitu educational cost yang meliputi biaya kuliah, pendaftaran, dan kebutuhan akademik, serta living cost yang mencakup biaya akomodasi, konsumsi, transportasi, dan kebutuhan sehari-hari. Kedua komponen ini memiliki variasi yang sangat beragam bergantung pada lokasi geografis, jenis institusi, dan program studi yang dipilih, sehingga menyulitkan proses perencanaan finansial secara akurat. Meskipun informasi mengenai biaya studi mahasiswa internasional tersedia secara umum, belum terdapat sistem yang

secara otomatis dan terstruktur mampu mengelompokkan informasi tersebut ke dalam kategori biaya yang jelas dan dapat dibandingkan. Informasi yang ada masih bersifat tersebar, tidak terstandarisasi, dan sulit untuk dianalisis secara komprehensif. Kondisi ini menyebabkan calon mahasiswa internasional maupun pemangku kebijakan kesulitan dalam membuat keputusan yang berbasis data terkait perencanaan finansial studi ke luar negeri.

Perkembangan teknologi machine learning membuka peluang untuk mengatasi permasalahan tersebut secara lebih efektif. Salah satu pendekatan yang relevan adalah metode klasterisasi atau pengelompokan data secara otomatis berdasarkan kemiripan karakteristik. Metode klasterisasi telah banyak diterapkan dalam berbagai bidang seperti segmentasi pelanggan, analisis pasar, hingga sistem rekomendasi, dan terbukti mampu mengidentifikasi pola tersembunyi dalam data berdimensi tinggi tanpa memerlukan label kelas yang telah ditentukan sebelumnya.

Di antara berbagai algoritma klasterisasi yang ada, Gaussian Mixture Models (GMM) memiliki keunggulan dibandingkan metode konvensional seperti K-Means. GMM merupakan model probabilistik yang mengasumsikan data dari campuran beberapa distribusi Gaussian. Tidak seperti K-Means yang menghasilkan pengelompokan keras (hard clustering), GMM menghasilkan pengelompokan lunak (soft clustering) di mana setiap titik data memiliki probabilitas keanggotaan terhadap masing-masing klaster. Hal ini menjadikan GMM lebih fleksibel dan mampu menangani distribusi data yang kompleks dan tidak simetris, yang sangat relevan untuk data biaya mahasiswa internasional yang bervariasi.

Beberapa penelitian terdahulu telah menerapkan metode machine learning berbasis clustering untuk analisis data di bidang pendidikan dan keuangan.

(Kurniawan et al., 2020) dalam jurnal Journal of Applied Computer Science and Technology (JACOST) menerapkan algoritma K-Means Clustering untuk mengelompokkan besaran Uang Kuliah Tunggal (UKT) mahasiswa baru di Universitas Negeri Padang ke dalam 5 kategori berdasarkan kondisi sosial ekonomi orang tua, dan berhasil mengidentifikasi pola pengelompokan yang signifikan meskipun metode K-Means memiliki keterbatasan pada klaster berbentuk non-sferis. (Mohamed Nafuri et al., 2022) dalam jurnal Applied Sciences (MDPI) mengusulkan pendekatan clustering berbasis machine learning menggunakan Python dan scikit-learn untuk mengklasifikasikan performa akademik mahasiswa kelompok berpenghasilan rendah (B40) di institusi pendidikan tinggi Malaysia, dengan dataset berjumlah 248.568 record mahasiswa, dan menemukan bahwa metode clustering mampu mengidentifikasi pola tersembunyi yang tidak dapat ditemukan secara manual. Namun demikian, penerapan Gaussian Mixture Models secara spesifik untuk klasterisasi tuition fee level dan living cost mahasiswa internasional dari berbagai negara dan kota secara bersamaan belum dieksplorasi dalam literatur yang ada. Penelitian ini bertujuan untuk mengisi celah tersebut dengan mengimplementasikan GMM pada dataset biaya mahasiswa internasional menggunakan bahasa pemrograman Python melalui Visual Studio Code.

Berdasarkan uraian di atas, penelitian yang berjudul *"Penerapan Machine Learning Dalam Klasterisasi Perguruan Tinggi Berdasarkan Tuition Fee Level dan Living Cost Mahasiswa Internasional Menggunakan Metode Gaussian Mixture Models"* bertujuan untuk menerapkan metode Gaussian Mixture Models dalam melakukan klasterisasi tuition fee level dan living cost mahasiswa internasional. Dengan menggunakan pendekatan machine learning berbasis Python, diharapkan

penelitian ini dapat menghasilkan klaster biaya yang informatif, akurat, dan dapat dijadikan referensi bagi mahasiswa internasional maupun institusi pendidikan dalam perencanaan dan pengambilan keputusan terkait biaya studi di luar negeri.

1.2 Rumusan Masalah

Berdasarkan latar belakang yang telah diuraikan, maka rumusan masalah dalam penelitian ini adalah sebagai berikut:

1. Bagaimana penerapan metode Gaussian Mixture Models dalam melakukan klasterisasi tuition fee level dan living cost mahasiswa internasional?
2. Bagaimana karakteristik dan profil biaya dari setiap klaster yang dihasilkan oleh model Gaussian Mixture Models?
3. Bagaimana performa model Gaussian Mixture Models yang diterapkan berdasarkan metrik evaluasi Silhouette Score dan Bayesian Information Criterion (BIC)?

1.3 Batasan Masalah

Agar penelitian ini lebih terfokus dan terarah, maka ditetapkan batasan-batasan masalah sebagai berikut:

1. Data yang digunakan dalam penelitian ini adalah dataset yang telah tersedia yang bersumber dari kaggle dengan nama dataset Internasional Education Costs yang berjumlah 1000 data.
2. Implementasi dan pengembangan model dilakukan menggunakan bahasa pemrograman Python dengan Visual Studio Code.
3. Variabel yang dianalisis terbatas pada fitur-fitur yang terdapat dalam dataset, yaitu tuition fee, living cost, city, country, dsj.

4. Evaluasi performa model dilakukan menggunakan Davies-Bouldin Index, dan Bayesian Information Criterion (BIC).

1.4 Tujuan Penelitian

Berdasarkan rumusan masalah yang telah dipaparkan, tujuan yang ingin dicapai dalam penelitian ini adalah:

1. Menerapkan metode Gaussian Mixture Models untuk melakukan klasterisasi tuition fee level dan living cost mahasiswa internasional menggunakan bahasa pemrograman Python.
2. Menganalisis dan mendeskripsikan karakteristik serta profil biaya dari setiap klaster yang dihasilkan oleh model Gaussian Mixture Models.
3. Mengevaluasi performa model Gaussian Mixture Models menggunakan metrik Davies-Bouldin Index dan Bayesian Information Criterion (BIC) untuk mengukur kualitas klasterisasi yang dihasilkan.

1.5 Manfaat Penelitian

Penelitian ini diharapkan dapat memberikan manfaat secara teoritis maupun praktis. Secara teoritis, penelitian ini berkontribusi pada pengembangan ilmu machine learning, khususnya dalam memperkaya literatur mengenai penerapan metode Gaussian Mixture Models pada domain analisis biaya pendidikan internasional yang selama ini belum banyak dieksplorasi. Selain itu, penelitian ini dapat menjadi referensi ilmiah bagi peneliti selanjutnya yang ingin mengembangkan atau membandingkan metode klasterisasi probabilistik lainnya dalam konteks pendidikan tinggi, serta memperluas wawasan mengenai penerapan algoritma unsupervised learning pada data numerik multivariat di bidang keuangan pendidikan.

Secara praktis, hasil klasterisasi yang dihasilkan penelitian ini diharapkan dapat memberikan gambaran yang terstruktur dan berbasis data kepada mahasiswa internasional mengenai level biaya studi di berbagai negara dan kota tujuan, sehingga membantu mereka dalam merencanakan kebutuhan finansial secara lebih akurat sebelum berangkat studi. Bagi institusi pendidikan dan lembaga pemberi beasiswa, profil klaster biaya yang dihasilkan dapat dijadikan acuan dalam menetapkan besaran bantuan finansial yang lebih tepat sasaran dan proporsional sesuai kondisi biaya di lokasi studi masing-masing. Bagi pemerintah dan pemangku kebijakan, temuan penelitian ini dapat menjadi bahan pertimbangan dalam merancang program dukungan finansial mahasiswa internasional yang lebih inklusif dan berorientasi data.

BAB II

LANDASAN TEORI

2.1 Tuition Fee Mahasiswa Internasional

Tuition fee atau Biaya pendidikan merujuk pada keseluruhan pengeluaran yang dibutuhkan oleh seorang mahasiswa untuk mengikuti program pendidikan tinggi di suatu institusi. Biaya pendidikan secara umum mencakup biaya kuliah (tuition fee) sebagai komponen utama, biaya pendaftaran dan administrasi akademik, biaya ujian dan sertifikasi, biaya buku teks dan materi pembelajaran, serta biaya akses laboratorium atau fasilitas akademik lainnya.

Menurut (Ahmad et al., 2024) mahasiswa internasional pada umumnya dikenakan tarif biaya kuliah yang lebih tinggi dibandingkan mahasiswa domestik karena tidak mendapatkan subsidi pendidikan dari pemerintah negara tujuan. Perbedaan besaran biaya pendidikan antar negara dan institusi inilah yang menjadi salah satu variabel utama yang akan dianalisis dan diklasterisasi dalam penelitian ini.

2.2 Living Cost Mahasiwa Internasional

Living cost atau biaya hidup adalah total pengeluaran yang dibutuhkan oleh seorang mahasiswa untuk memenuhi kebutuhan sehari-hari selama menjalani studi di luar negeri. Living cost mencakup seluruh pengeluaran di luar tuition fee, yaitu biaya akomodasi, konsumsi (makan dan minum), transportasi, serta bahan studi tambahan. Di beberapa negara, biaya akomodasi menjadi komponen terbesar dalam living cost mahasiswa, misalnya di Jerman, biaya sewa tempat tinggal rata-rata menyumbang 48,7% dari total pengeluaran bulanan mahasiswa. Living cost juga sangat bervariasi tergantung lokasi geografis. perbandingan biaya hidup antar kota

tujuan studi sangat kompleks karena dipengaruhi oleh fluktuasi nilai tukar mata uang, perbedaan harga akomodasi berdasarkan jenis dan lokasi, serta variasi gaya hidup masing-masing mahasiswa. Kota-kota metropolitan seperti London, Tokyo, Sydney, dan New York umumnya memiliki living cost yang jauh lebih tinggi dibandingkan kota-kota kecil atau negara-negara berkembang. Variabilitas tinggi pada living cost inilah yang menyulitkan mahasiswa internasional dalam memperkirakan kebutuhan finansial secara akurat sebelum berangkat studi.

2.3 Data Mining

Data mining merupakan suatu proses analitis yang bertujuan menggali informasi serta pengetahuan yang bernilai dari kumpulan data yang besar dan kompleks. Secara konseptual, data mining dapat dipahami sebagai kegiatan menemukan pola, hubungan, korelasi, maupun anomali yang relevan dalam suatu dataset melalui pendekatan komputasi, sehingga menghasilkan wawasan yang mendukung proses pengambilan keputusan (Pangastuti et al., 2021)

Dalam penerapannya, data mining menggunakan berbagai algoritma dan metode yang dapat dikategorikan ke dalam beberapa kelompok utama, yakni klasifikasi, klasterisasi, regresi, asosiasi, dan deteksi anomali. Metode klasifikasi seperti Decision Tree, Naïve Bayes, dan Support Vector Machine (SVM) digunakan untuk memprediksi kategori suatu data berdasarkan atribut-atribut yang dimilikinya. Sementara itu, metode klasterisasi seperti K-Means digunakan untuk mengelompokkan data tanpa label ke dalam kelompok-kelompok yang memiliki kesamaan karakteristik. Selain K-Means, metode Gaussian Mixture Model (GMM) juga banyak digunakan dalam klasterisasi, di mana GMM memodelkan distribusi data sebagai kombinasi dari beberapa distribusi Gaussian sehingga mampu

menangkap struktur klaster yang lebih kompleks dan tumpang tindih dibandingkan metode berbasis jarak konvensional (Widiarina et al., 2024)

2.4 Tahapan Data Mining

Proses data mining mengikuti metodologi sistematis yang dikenal sebagai Knowledge Discovery in Databases (KDD) atau dalam praktiknya sering menggunakan kerangka kerja CRISP-DM (Cross-Industry Standard Process for Data Mining). Tahapan-tahapan dalam proses data mining ini bersifat iteratif dan saling berkaitan, dimana setiap tahap dapat kembali ke tahap sebelumnya untuk perbaikan atau penyempurnaan. Pemahaman yang mendalam terhadap setiap tahapan sangat penting untuk memastikan kualitas hasil analisis dan validitas pengetahuan yang dihasilkan. Proses data mining terdiri dari beberapa tahap yang biasanya mengikuti metodologi CRISP-DM (Cross-Industry Standard Process for Data Mining), yaitu:

1. Business Understanding

Tahap pertama adalah business understanding, yaitu memahami tujuan penelitian dan permasalahan yang ingin diselesaikan. Dalam konteks penelitian ini, fokus utama adalah pada identifikasi pola tingkat biaya pendidikan (educational cost) dan biaya hidup (living cost) mahasiswa internasional yang memiliki latar belakang ekonomi dan geografis yang beragam. Permasalahan yang dihadapi meliputi belum terklasifikasinya tingkat biaya yang ditanggung mahasiswa secara sistematis, kurangnya pemetaan kelompok mahasiswa berdasarkan beban biaya yang serupa, serta belum optimalnya kebijakan institusi dalam memberikan dukungan finansial atau akademik berbasis data. Tujuan penelitian ini adalah menerapkan metode machine learning menggunakan Gaussian Mixture Models (GMM) untuk

mengelompokkan mahasiswa internasional berdasarkan tingkat biaya pendidikan dan biaya hidup, sehingga dapat membantu institusi dalam merancang kebijakan yang lebih tepat sasaran.

2. Data Understanding

Tahap kedua adalah data understanding, yaitu proses memahami karakteristik serta struktur data yang akan dianalisis. Data dalam penelitian ini diperoleh dari institusi pendidikan dalam bentuk file Microsoft Excel yang mencakup informasi mahasiswa internasional selama periode tertentu. Data terdiri dari variabel-variabel yang berkaitan dengan biaya pendidikan dan biaya hidup, seperti tuition fee, biaya akomodasi, biaya makan, transportasi, serta pengeluaran lainnya. Pada tahap ini dilakukan eksplorasi awal untuk memahami distribusi data, jenis variabel, kelengkapan data, serta mendeteksi adanya inkonsistensi atau anomali. Selain itu, dilakukan identifikasi variabel yang paling relevan untuk digunakan dalam proses klasterisasi menggunakan metode GMM.

3. Data Preparation

Tahap ketiga adalah data preparation, yang merupakan tahap paling krusial karena memakan sekitar 60–80% dari keseluruhan proses penelitian.

Tahapan ini meliputi:

1. Data cleaning untuk menangani missing values dan outlier.
2. Data integration untuk menggabungkan berbagai sumber data biaya.
3. Data transformation seperti normalisasi agar skala antar variabel menjadi seimbang.
4. Feature engineering untuk membentuk variabel baru yang lebih representatif.

Dalam konteks GMM, normalisasi sangat penting agar variabel dengan nilai besar (misalnya tuition fee) tidak mendominasi pembentukan cluster dibandingkan variabel lain seperti biaya transportasi atau konsumsi.

4. Modeling (Pemodelan)

Tahap keempat adalah modeling, yaitu penerapan algoritma machine learning untuk menemukan pola dalam data. Pada penelitian ini digunakan metode Gaussian Mixture Models (GMM) untuk melakukan klasterisasi tingkat biaya pendidikan dan biaya hidup mahasiswa internasional. GMM merupakan metode soft clustering yang memungkinkan setiap data memiliki probabilitas keanggotaan pada beberapa cluster.

Tahap ini mencakup:

1. Penentuan jumlah komponen Gaussian (cluster) optimal.
2. Inisialisasi parameter model.
3. Proses iteratif menggunakan algoritma Expectation-Maximization (EM) hingga model mencapai konvergensi.

5. Evaluation (Evaluasi)

Tahap kelima adalah evaluation, yang bertujuan untuk menilai kualitas hasil klasterisasi yang dihasilkan oleh model GMM.

Evaluasi dilakukan menggunakan beberapa metrik, antara lain:

1. Silhouette Coefficient.
2. Davies-Bouldin Indeks.
3. Calinski-Harabasz Indeks.
4. Bayesian Information Criterion (BIC).
5. Akaike Information Criterion (AIC).

Dalam penelitian ini, BIC digunakan sebagai metrik utama dalam menentukan jumlah cluster optimal, karena mampu memberikan penalti terhadap kompleksitas model sehingga dapat menghindari overfitting dan menghasilkan model yang lebih sederhana (parsimonious) namun tetap representatif dalam menggambarkan pola biaya mahasiswa.

6. Deployment

Tahap keenam adalah deployment, yaitu implementasi model yang telah divalidasi ke dalam sistem pendukung pengambilan keputusan. Hasil klasterisasi dapat dimanfaatkan oleh institusi untuk:

1. Mengidentifikasi kelompok mahasiswa dengan tingkat biaya rendah, sedang, dan tinggi
2. Menyusun kebijakan beasiswa atau bantuan finansial yang lebih tepat sasaran
3. Mendukung perencanaan anggaran dan strategi internasionalisasi kampus

Selain itu, tahap ini juga mencakup dokumentasi model, interpretasi karakteristik tiap cluster, serta penyajian hasil dalam bentuk dashboard atau laporan yang mudah dipahami oleh pihak institusi.

2.5 Machine Learning

Machine learning adalah bagian fundamental dalam bidang kecerdasan buatan yang dipakai untuk memecahkan berbagai masalah dan merupakan penerapan dari kecerdasan buatan yang berfokus pada pembuatan sistem yang bisa belajar sendiri tanpa harus diprogram berulang kali. *Machine learning* memakai algoritma yang bisa mempelajari pola dan hubungan kompleks dalam data,

sehingga membantu dalam mengambil keputusan dengan tingkat akurasi yang lebih baik (Nurhalizah et al., 2024).

Berdasarkan penelitian (Darmawan Sidik et al., 2022), *machine learning* dibagi menjadi 3 yaitu *supervised learning, unsupervised learning,* dan *reinforcement learning.*

1. *Supervised Learning*

   Supervised learning adalah jenis algoritma *machine learning* yang proses pembelajarannya dilakukan dengan adanya pengawasan. Beberapa contoh dari *supervised learning* adalah klasifikasi dan regresi.

2. *Unsupervised Learning*

   Unsupervised learning adalah algoritma machine learning yang proses pembelajarannya berlangsung tanpa pengawasan. Contoh-contoh dalam *unsupervised learning* mencakup pengelompokan dan reduksi dimensi.

3. *Reinforcement Learning*

   Algoritma *machine learning* ini memungkinkan agen perangkat lunak untuk beroperasi secara otomatis, mencari perilaku optimal guna memaksimalkan kinerja algoritma. Dalam *Reinforcement Learning*, beberapa contohnya meliputi pengambilan *real time decision*, navigasi robot, pembelajaran tugas, akuisisi keterampilan, dan kecerdasan buatan dalam permainan.

Keunggulan *machine learning* untuk analisis pola terletak pada kemampuannya untuk mengidentifikasi pola dalam dataset, sehingga mampu meramalkan dan memahami karakteristik dari objek yang tidak dikenal (Nurhalizah et al., 2024a).

2.6 Unsupervised Learning

Unsupervised learning merupakan paradigma dalam machine learning yang bekerja pada data tanpa label atau kategori yang telah ditentukan sebelumnya, di mana algoritma secara mandiri mengidentifikasi pola, struktur, dan hubungan tersembunyi dalam data (Nurhalizah et al., 2024b) Berbeda dengan supervised learning yang memerlukan target variable berlabel, pendekatan ini mengeksplorasi struktur internal data untuk menemukan pengelompokan alami tanpa panduan eksplisit. Metode yang umum digunakan meliputi clustering, dimensionality reduction, dan association rule learning

Karakteristik data tanpa label menyebabkan algoritma bergantung sepenuhnya pada distribusi serta hubungan antar variabel dalam dataset (Abijono et al., 2021). Hal ini memberikan fleksibilitas dalam eksplorasi, namun menimbulkan tantangan pada tahap evaluasi karena tidak tersedia ground truth sebagai pembanding. Oleh karena itu, validasi umumnya menggunakan metrik internal seperti silhouette score serta interpretasi berbasis konteks domain.

Dalam konteks educational data mining, metode unsupervised learning memiliki peran penting dalam menganalisis data yang berkaitan dengan aspek finansial mahasiswa, seperti biaya pendidikan (tuition fee) dan biaya hidup (living cost). Pendekatan ini memungkinkan ditemukannya pola-pola tersembunyi dalam data tanpa memerlukan label atau klasifikasi sebelumnya.

Penggunaan algoritma seperti K-Means dan Gaussian Mixture Models (GMM) dapat digunakan untuk melakukan segmentasi mahasiswa internasional berdasarkan tingkat beban biaya yang mereka tanggung. Secara khusus, GMM memiliki pendekatan soft clustering yang memberikan probabilitas keanggotaan

pada setiap cluster, sehingga mampu merepresentasikan variasi tingkat biaya mahasiswa yang bersifat tidak tegas atau saling tumpang tindih.

Hasil klasterisasi tersebut dapat dimanfaatkan sebagai dasar dalam pengambilan keputusan oleh institusi, seperti perumusan kebijakan bantuan finansial, pengelompokan mahasiswa berdasarkan tingkat biaya, serta penyusunan strategi dukungan ekonomi yang lebih tepat sasaran bagi mahasiswa internasional.

2.7 Clustering

Clustering adalah metode pengelompokan data dimana *clustering* merupakan sebuah proses untuk mengelompokkan data ke dalam beberapa cluster atau kelompok sehingga data dalam satu cluster memiliki tingkat kemiripan yang maksimum dan data antar *cluster* memiliki kemiripan yang minimum (Gustientiedina et al., 2019).

Clustering merupakan salah satu metode dalam unsupervised learning yang bertujuan untuk mengelompokkan data berdasarkan kesamaan karakteristik tanpa adanya label atau kategori yang telah ditentukan sebelumnya. Proses ini dilakukan untuk mengidentifikasi pola atau struktur tersembunyi dalam data, sehingga objek dengan karakteristik yang mirip dapat tergabung dalam kelompok yang sama.

Dalam konteks penelitian ini, clustering digunakan untuk mengidentifikasi pola tingkat biaya pendidikan dan biaya hidup mahasiswa internasional berdasarkan indikator yang telah ditentukan. Melalui proses pengelompokan, dapat diketahui adanya perbedaan karakteristik antar kelompok mahasiswa, seperti kelompok dengan tingkat biaya tinggi, sedang, maupun rendah.

Secara umum, metode clustering terbagi menjadi dua pendekatan utama, yaitu hard clustering dan soft clustering. Hard clustering mengelompokkan setiap data secara

tegas ke dalam satu cluster tertentu, di mana setiap objek hanya memiliki satu keanggotaan cluster. Salah satu contoh metode yang menggunakan pendekatan ini adalah K-Means.

Sebaliknya, soft clustering atau probabilistic clustering memungkinkan setiap data memiliki derajat keanggotaan pada lebih dari satu cluster dalam bentuk probabilitas. Pendekatan ini lebih fleksibel karena mampu merepresentasikan variasi tingkat biaya yang tidak selalu terpisah secara jelas, sehingga lebih sesuai untuk menganalisis data biaya pendidikan dan biaya hidup mahasiswa yang cenderung memiliki karakteristik yang saling tumpang tindih.

2.8 Hard Clustering vs Soft Clustering

Hard clustering merupakan pendekatan deterministik dalam clustering di mana setiap objek data hanya dapat menjadi anggota satu cluster secara eksklusif (crisp membership), dengan nilai keanggotaan 1 pada satu cluster dan 0 pada cluster lainnya. Metode seperti K-Means dan Hierarchical Clustering termasuk dalam kategori ini, yang bekerja dengan membagi data ke dalam kelompok-kelompok yang memiliki batas tegas dan tidak saling tumpang tindih. Pendekatan ini relatif sederhana, efisien secara komputasi, dan mudah diinterpretasikan, namun kurang mampu merepresentasikan data yang memiliki karakteristik ambigu atau overlapping.

Sebaliknya, soft clustering merupakan pendekatan probabilistik yang memungkinkan setiap objek memiliki derajat keanggotaan pada lebih dari satu cluster dalam rentang nilai 0 hingga 1. Soft clustering memodelkan struktur data menggunakan fungsi keanggotaan atau distribusi probabilitas sehingga lebih fleksibel dalam menangkap kompleksitas dan ketidakpastian dalam data. Dalam

penelitian ini, pendekatan soft clustering melalui GMM dipilih karena perilaku belajar mahasiswa tidak dapat dikategorikan secara tegas dalam satu kelompok saja; seorang mahasiswa dapat menunjukkan lebih dari satu pola keterlibatan secara bersamaan. Selain itu, GMM menyediakan probabilitas keanggotaan yang dapat dimanfaatkan untuk analisis tingkat kepastian, perancangan intervensi akademik yang lebih personal, serta pengembangan sistem rekomendasi berbasis data. Kemampuannya dalam memodelkan variasi bentuk dan orientasi cluster melalui covariance matrix serta penggunaan Bayesian Information Criterion (BIC) untuk menentukan jumlah cluster optimal menjadikannya lebih representatif dalam analisis pola perilaku belajar mahasiswa kelas internasional.

2.9 Gaussian Mixture Models

GMM adalah model yang memiliki komponen fungsi-fungsi *Gaussian* yang terdiri dari *threshold.* Ada pun metode yang efektif seperti mixture model pada perubahan model yang bergerak lambat, dengan fleksibilitas yang lebih dan telitih untuk memodelkan statistik dari data atau pun visualisasikan suatu *dynamic scane*. GMM juga dikatakan sebagai model statistik yang distribusikan dengan probalitas dapat dinilai dari bobot pada distribusi Gaussian. Dalam perhitungan GMM ini sangat tepat baik dengan parameter atau pun tidak dengan parameter (Sidabutar, 2020).

Metode ini berasumsi bahwa seluruh individu merupakan gabungan dari distribusi peluang gaussian, yang masing- 8 masing mewakili distribusi gaussian dengan parameter distribusi yang berbeda. Salah satu cara untuk memperkirakan parameter dalam metode ini adalah dengan menggunakan algoritma *Expectation Maximization (EM)* (Joko Riyono et al., 2022).

Rumus dasar untuk *Gaussian Mixture Models* berdasarkan penelitian (Milson et al., 2024), adalah sebagai berikut:

$$p(x) = \sum_{k=1}^{K} \pi_k \cdot N(x|\mu_k, \Sigma_k) \text{.....(1)}$$

Keterangan:

1. $p(x)$: Fungsi densitas probabilitas total untuk data x, yang merupakan campuran dari beberapa distribusi gaussian.
2. k: Jumlah komponen gaussian (jumlah cluster yang akan diidentifikasi).
3. π_k: Bobot dari komponen gaussian ke-k, yang menunjukkan proporsi dari komponen ke- k dalam model. Nilai π_k harus memenuhi $\sum_{k=1}^{K} \pi_k$=.
4. Dalam menentukannya, dapat menggunakan $\frac{N_k}{N}$.
5. $N(x|\mu_k, \Sigma_k)$: distribusi gaussian multivariat untuk komponen k, dengan rata rata μ_k dan matrik kovarian Σ_k.

Rumus diatas merupakan fungsi densitas probabilitas dari Gaussian Mixture Model (GMM), yang digunakan untuk memodelkan distribusi data yang berasal dari beberapa kelompok (cluster) sekaligus. Secara matematis, GMM merepresentasikan bahwa setiap titik data x memiliki kemungkinan untuk berasal dari salah satu dari K distribusi gaussian yang berbeda.

2.10 Expectation- Maximization

Penerapan Gaussian Mixture Model dapat dilakukan melalui algoritma *Expectation Maximization (EM).* Algoritma ini memanfaatkan data yang tersedia untuk mencari nilai optimal dari variabel dan selanjutnya menentukan parameter model. Algoritma ini terdiri dari dua tahap: tahap Ekspektasi, yang bertujuan untuk memperkirakan nilai dan variabel yang hilang atau belum tersedia, serta tahap maksimisasi, yang digunakan untuk memperbarui parameter berdasarkan nilai yang

diperoleh dari tahap Ekspektasi. Proses ini dilakukan berulang kali hingga mencapai konvergensi dan menemukan maximum likelihood (Milson et al., 2024)

1. *E- Step (Expectation Step)*

$$N(x|\mu_k, \sigma_k^2) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu_k)^2}{2\sigma_k^2}} \text{.....(2)}$$

Keterangan:

a. Fungsi gaussian $N(x|\mu_k, \sigma_k^2)$ menghitung probabilitas x berada dalam distribusi dengan parameter *mean* (μ) dan variansi (σ^2).

b. Faktor normalisasi $\frac{1}{\sqrt{2\pi\sigma^2}}$, memastikan total probabilitas adalah 1 dan $e^{-\frac{(x-\mu)^2}{2\sigma^2}}$ mengukur probabilitas berdasarkan jarak μ dari mean.

Pada tahapan E-Step ini, dilakukan juga perhitungan nilai ekspektasi dari probabilitas posterior dengan rumus:

$$\gamma_k = \frac{\pi_k \cdot P(x|c_k)}{\sum_{j=1}^{K} \pi_j \cdot P(x|c_j)} \text{.....(3)}$$

Keterangan:

a. $P(x|c_k)$= distribusi probabilitas gaussian data x untuk cluster k.

b. π_k= bobot pada cluster k.

c. $\sum_{j=1}^{K}$ = jumlah untuk seluruh cluster.

Rumus ini adalah inti dari perhitungan GMM pada tahap E-step, di mana setiap data mahasiswa dihitung probabilitasnya terhadap masing-masing cluster. Hasil probabilitas inilah yang kemudian digunakan untuk menentukan mahasiswa masuk ke kelompok Intensif, Adaptif, atau Pasif.

2. *M-Step (Maximization Step)*

Pada langkah ini, parameter model Gaussian berdasarkan hasil perhitungan yang dilakukan di E-step (Expectation Step) diperbarui atau dilakukan beberapa iterasi seperti nilai mean, varians, dan bobot.

3. *Likelihood Function*

$$L = \log \sum_k \pi_k \cdot N(x_n | \mu_k, \sigma_k^2) \quad \text{.....(4)}$$

Keterangan:

a. L: Nilai *log-likelihood* yang menunjukkan kecocokan model terhadap data.

b. π_k: Bobot atau proporsi cluster ke-k.

c. $N(x_n|\mu_k, \sigma_k^2)$: Distribusi Gaussian data ke-n pada cluster ke-k.

d. x_n: Data ke-n yang diamati.

e. μ_k: Rata-rata distribusi Gaussian cluster ke-k.

f. σ_k^2: Varians distribusi Gaussian cluster ke-k.

Rumus diatas berfungsi untuk menghitung nilai *log-likelihood* pada metode Gaussian Mixture Models (GMM). Nilai *log-likelihood* ini digunakan untuk mengukur tingkat kecocokan model terhadap data yang diamati.

2.11 Penetuan Jumlah Komponen Optimal

Evaluasi dalam clustering merupakan tahap esensial untuk menilai kualitas hasil pengelompokan, terutama karena metode ini termasuk dalam unsupervised learning yang tidak memiliki label atau ground truth sebagai pembanding (Paembonan & Abduh, 2021). Oleh karena itu, evaluasi umumnya dilakukan melalui internal validation yang mengukur struktur cluster berdasarkan karakteristik data itu sendiri. Pada model probabilistik seperti Gaussian Mixture Model (GMM), kriteria berbasis likelihood seperti Akaike Information Criterion

(AIC) dan Bayesian Information Criterion (BIC) digunakan untuk menentukan jumlah cluster optimal dengan mempertimbangkan keseimbangan antara kecocokan model dan kompleksitas parameter (Prasetya et al., 2025). BIC menunjukkan kecenderungan yang lebih kuat dalam membatasi kompleksitas model dibandingkan AIC melalui komponen pembatas jumlah parameternya., sehingga cenderung menghasilkan model yang lebih sederhana dan mengurangi risiko overfitting, khususnya pada dataset berukuran besar.

2.12 Bayesian Information Criterion (BIC)

Bayesian Information Criterion (BIC) menjadi salah satu metode untuk menentukan jumlah cluster yang optimal dengan menyeimbangkan kecocokan model terhadap data dan kompleksitas model pada jumlah cluster. Semakin rendah nilai *Bayesian Information Criterion (BIC)*, semakin baik model tersebut. *Bayesian Information Criterion (BIC)* sering digunakan dalam metode *Gaussian Mixture Models* karena melibatkan komponen probabilistik (Ummami & Winarno, 2023).

Berdasarkan penelitian (Ummami & Winarno, 2023), persamaan *Bayesian Information Criterion (BIC)* yaitu sebagai berikut:

$$BIC = -2 \cdot L + k \cdot \ln(N) \quad (5)$$

Keterangan:

a. L: Nilai likelihood dari model yang dipasang (fit), yaitu probabilitas model tersebut dalam menjelaskan data.
b. k: Jumlah parameter dalam model.
c. N: Jumlah data yang diamati (sample size).

Selain BIC, terdapat kriteria evaluasi lain seperti Akaike Information Criterion (AIC) dan metode berbasis jarak seperti silhouette score. Namun, dalam penelitian

ini BIC dipilih karena lebih konsisten dalam pemilihan jumlah komponen pada model campuran distribusi serta memberikan penalti yang lebih kuat terhadap kompleksitas model dibandingkan AIC.

2.13 Python

Python adalah bahasa pemrograman tingkat tinggi yang bersifat interpreted, dinamis, dan berorientasi objek, yang saat ini menjadi bahasa pemrograman paling populer dalam bidang data science dan machine learning. Python dipilih secara luas dalam komunitas ilmiah dan industri karena sintaksnya yang bersih dan mudah dipelajari, ekosistem library yang sangat kaya untuk komputasi numerik dan machine learning, serta dukungan komunitas yang sangat besar melalui platform seperti PyPI dan GitHub. Pada penelitian ini, Python digunakan sebagai bahasa pemrograman utama untuk seluruh proses implementasi, mulai dari pra-pemrosesan data, pelatihan model GMM, evaluasi klasterisasi, hingga visualisasi hasil.

Dalam implementasi penelitian ini, beberapa library Python utama dimanfaatkan secara terintegrasi. Library scikit-learn (sklearn) menyediakan implementasi kelas GaussianMixture untuk pemodelan GMM, StandardScaler untuk normalisasi data, serta fungsi evaluasi silhouette_score dan davies_bouldin_score. Library pandas menyediakan struktur data DataFrame yang sangat berguna dalam proses pembersihan, eksplorasi, dan transformasi data tabular. Library NumPy (Numerical Python) menyediakan array multidimensi dan operasi matematika berkinerja tinggi yang menjadi fondasi dari scikit-learn dan pandas. Library Matplotlib dan Seaborn digunakan untuk membuat grafik

visualisasi hasil klasterisasi, plot distribusi, dan heatmap yang memudahkan interpretasi hasil penelitian.

2.14 Visual Studio Code

Visual Studio Code (VSCode) adalah editor kode sumber (source code editor) lintas platform yang dikembangkan oleh Microsoft Corporation dan dirilis secara gratis. VSCode mendukung berbagai bahasa pemrograman termasuk Python melalui ekstensi (extension) yang tersedia di marketplace, serta dilengkapi dengan fitur-fitur produktivitas seperti IntelliSense untuk pelengkapan kode otomatis dan saran kontekstual, debugger terintegrasi untuk identifikasi dan perbaikan kesalahan, dukungan Jupyter Notebook secara native untuk eksplorasi data interaktif, serta integrasi dengan sistem kontrol versi Git. Pada penelitian ini, VSCode digunakan sebagai Integrated Development Environment (IDE) utama dalam seluruh proses pengembangan kode Python, mulai dari tahap pra-pemrosesan data hingga pelatihan dan evaluasi model Gaussian Mixture Models.

2.15 Ringkasan Penelitian Terdahulu

Penelitian mengenai biaya pendidikan, biaya hidup mahasiswa internasional, serta penerapan metode machine learning dalam klasterisasi data sudah pernah diteliti oleh para peneliti sebelumnya. Sub bab ini merangkum empat penelitian terdahulu yang relevan dengan topik penelitian yang berjudul "Penerapan Machine Learning dalam Klasterisasi Tingkat Biaya Pendidikan dan Biaya Hidup Mahasiswa Internasional Menggunakan Metode Gaussian Mixture Models." Berikut adalah beberapa penelitian terdahulu yang relevan dengan topik penelitian klasterisasi tingkat biaya pendidikan dan biaya hidup mahasiswa internasional:

Tabel 2. 1 Ringkasan Penelitian Terdahulu

No.	Judul Dan Peneliti	Pembahasan	Metode	Kelebihan Dan Kekurangan
1	Education Cost as a New Fickle in Higher Education for Students Learning via Quantitively Multinomial Logistic Regression Asma Ahmad, Murtaza Hasan & Mansour Ghorbanpour (2024)	Mengkaji pengaruh biaya kuliah (tuition fee) terhadap kualitas pendidikan dan prestasi akademik mahasiswa di Islamia University of Bahawalpur (IUB), Pakistan. Penelitian ini mengidentifikasi faktor-faktor yang mempersulit atau mempermudah pembayaran biaya kuliah berdasarkan survei terhadap 1.000 mahasiswa dari berbagai departemen.	Multinomial Logistic Regression dengan data primer survei online (Google Form).	Kelebihan: Dataset besar (1.000 responden), open access, terindeks Scopus dan Web of Science (Q1), membuktikan biaya pendidikan sebagai variabel kritis dalam pendidikan tinggi. Kekurangan: Hanya menggunakan metode regresi, tidak melakukan klasterisasi data.
2	Tuition Fees for International Students: A Policy Instrument of Cost Sharing and Control or Simply Income Generation Hans Lundin (2024)	Mengkaji reformasi biaya kuliah bagi mahasiswa internasional di Swedia sejak 2011 menggunakan dua model teoretis: revenue-seeking dan cost-sharing and control. Penelitian ini menelusuri sejauh mana instrumen berbasis pasar mendorong perguruan tinggi mencari pendapatan dari	Analisis kebijakan komprehensif menggunakan dua model teoretis (revenue-seeking dan cost-sharing and control). Data sekunder dari laporan pemerintah dan institusi pendidikan Swedia.	Kelebihan: Memberikan perspektif kebijakan mendalam tentang struktur biaya kuliah mahasiswa internasional, open access, terindeks Scopus dan ESCI. Kekurangan: Tidak menggunakan metode machine learning atau data mining. Fokus hanya pada satu negara (Swedia) tanpa analisis perbandingan lintas

No.	Judul Dan Peneliti	Pembahasan	Metode	Kelebihan Dan Kekurangan
		pasar global, khususnya pada program Engineering & Technology dan Business Management.		negara secara kuantitatif.
3	The Impact of the Cost-of-Living Crisis on Online Student Engagement and Future Study Plans Cathy Schofield (2024)	Menganalisis dampak krisis biaya hidup (cost-of-living crisis) di UK tahun 2022–2023 terhadap keterlibatan belajar dan rencana studi mahasiswa. Survei terhadap lebih dari 800 mahasiswa online membuktikan bahwa komponen biaya hidup — mencakup biaya makanan, energi, akomodasi, dan transportasi — berpengaruh langsung pada performa akademik mahasiswa.	Survei kuantitatif terhadap 800+ mahasiswa menggunakan platform Qualtrics, analisis deskriptif dan komparatif berbasis kuesioner.	Kelebihan: Data empiris kuat tentang komponen biaya hidup mahasiswa, open access (CC BY-NC-ND 4.0), terindeks Scopus dan ERIC. Kekurangan: Tidak menggunakan metode klasterisasi atau machine learning. Cakupan geografis terbatas pada UK dan tidak melibatkan mahasiswa internasional secara spesifik.
4	Analysis of Tuition Growth Rates Based on Clustering and Regression Models Long Cheng & Chenyu You (2016)	Mengeksplorasi berbagai faktor yang mempengaruhi tingkat pertumbuhan biaya kuliah menggunakan dataset dari National Center for Education Statistics (IPEDS) mencakup data 300 universitas di	K-Means Clustering, Linear Regression, Regression Tree, Decision Tree, dan Random Effect Model.	Kelebihan: Merupakan penelitian paling relevan secara metodologis karena mengkombinasikan clustering dengan analisis biaya kuliah. Dataset besar (300 universitas, 9 tahun), open access. Kekurangan: Menggunakan K-

No.	Judul Dan Peneliti	Pembahasan	Metode	Kelebihan Dan Kekurangan
		42 negara bagian AS selama 9 tahun (2000–2010). Analisis dilakukan pada tiga kategori sekolah: swasta, negeri in-state, dan negeri out-of-state.		Means Clustering yang tidak bersifat probabilistik (berbeda dengan GMM). Tidak mempertimbangkan biaya hidup mahasiswa dan tidak berfokus pada mahasiswa internasional.

2.16 Analisis GAP

Berdasarkan ringkasan penelitian terdahulu, terdapat kesenjangan penelitian (*research gap*) yang cukup signifikan dalam konteks analisis biaya pendidikan dan biaya hidup mahasiswa internasional. Sebagian besar penelitian sebelumnya cenderung berfokus pada analisis hubungan antara biaya pendidikan dengan variabel lain, seperti prestasi akademik atau kebijakan institusi, dengan menggunakan metode statistik konvensional seperti regresi. Di sisi lain, penelitian yang membahas biaya hidup mahasiswa umumnya hanya bersifat deskriptif dan tidak mengintegrasikan pendekatan analisis berbasis data yang mampu mengidentifikasi pola tersembunyi secara sistematis. Selain itu, penelitian yang menerapkan metode klasterisasi dalam konteks pendidikan masih terbatas pada penggunaan algoritma deterministik seperti K-Means, yang memiliki keterbatasan dalam merepresentasikan kompleksitas dan ketidakpastian data, khususnya pada data biaya yang cenderung memiliki distribusi yang saling tumpang tindih.

Selanjutnya, belum ditemukan penelitian yang secara komprehensif menggabungkan variabel biaya pendidikan dan biaya hidup mahasiswa internasional dalam satu kerangka analisis klasterisasi berbasis *unsupervised*

machine learning. Padahal, kedua variabel tersebut merupakan komponen utama dalam menggambarkan beban finansial mahasiswa secara keseluruhan. Oleh karena itu, penelitian ini hadir untuk mengisi celah tersebut dengan menerapkan Gaussian Mixture Models (GMM) yang bersifat probabilistik dan lebih fleksibel dalam menangkap variasi serta ketidakpastian data. Dengan pendekatan ini, penelitian tidak hanya mampu menghasilkan pengelompokan yang lebih akurat dan representatif, tetapi juga memberikan kontribusi dalam pengembangan analisis data pendidikan yang lebih komprehensif, khususnya dalam konteks pemetaan tingkat biaya mahasiswa internasional.

BAB III

METODOLOGI PENELITIAN

3.1 Jenis Penelitian

Penelitian ini termasuk dalam penelitian kuantitatif dengan pendekatan deskriptif-eksploratif. Pendekatan kuantitatif digunakan karena penelitian ini memanfaatkan data numerik yang dianalisis melalui teknik statistik dan pemodelan matematis untuk memperoleh gambaran objektif mengenai pola tingkat biaya pendidikan dan biaya hidup mahasiswa internasional. Data yang dianalisis berupa variabel-variabel biaya yang terukur, seperti biaya pendidikan (*tuition fee*) dan berbagai komponen biaya hidup, yang digunakan untuk mengidentifikasi kecenderungan pola pengeluaran mahasiswa.

Pendekatan deskriptif-eksploratif dalam penelitian ini tidak bertujuan untuk menguji hubungan kausal antarvariabel, melainkan berfokus pada upaya mengidentifikasi, memetakan, dan menggambarkan struktur pola yang terbentuk dalam data. Proses pengelompokan dilakukan berdasarkan karakteristik distribusi data menggunakan pendekatan pemodelan probabilistik, sehingga memungkinkan teridentifikasinya kecenderungan tingkat biaya yang muncul secara alami dalam populasi mahasiswa internasional.

3.2 Metode Penelitian

Metode penelitian ini dilaksanakan melalui tahapan yang terstruktur, meliputi pengumpulan data, praproses data, pemodelan, evaluasi model, serta interpretasi hasil. Data yang digunakan berupa data kuantitatif yang merepresentasikan indikator perilaku belajar mahasiswa kelas internasional. Pada tahap praproses, dilakukan pembersihan data, penanganan *missing values*, serta

transformasi atau normalisasi variabel apabila diperlukan untuk memastikan kualitas dan kesiapan data sebelum proses pemodelan dilakukan.

Proses pemodelan dilakukan dengan menerapkan metode Gaussian Mixture Model untuk mengidentifikasi struktur pengelompokan yang terbentuk dalam data berdasarkan distribusi probabilistik. Estimasi parameter model dilakukan menggunakan algoritma Expectation-Maximization algorithm yang bekerja secara iteratif melalui tahap *expectation* dan *maximization* hingga mencapai konvergensi. Penentuan jumlah komponen optimal dilakukan menggunakan kriteria informasi seperti Bayesian Information Criterion (BIC) dan Akaike Information Criterion (AIC) guna memperoleh model dengan tingkat kompleksitas yang proporsional terhadap kemampuan representasi data.

Seluruh proses analisis dilaksanakan menggunakan bahasa pemrograman Python sebagai lingkungan komputasi utama. Pengolahan numerik dilakukan dengan pustaka NumPy, pengelolaan serta manipulasi data menggunakan Pandas, dan implementasi model Gaussian Mixture.

3.3 Teknik Pengumpulan Data

Penelitian ini menggunakan data sekunder yang diperoleh dari platform Kaggle, yang menyediakan berbagai dataset terbuka untuk keperluan analisis dan penelitian. Data yang digunakan berkaitan dengan biaya pendidikan (*tuition fee*) dan biaya hidup (*living cost*) mahasiswa internasional dalam periode tertentu. Data tersebut mencakup variabel-variabel yang berhubungan dengan pengeluaran mahasiswa, seperti biaya pendidikan, biaya akomodasi, transportasi, serta komponen biaya hidup lainnya. Selain itu, terdapat pula variabel pendukung seperti

negara, universitas dan informasi relevan lainnya yang dapat memberikan konteks dalam proses analisis.

Seluruh data berbentuk numerik dan telah terstruktur, sehingga dapat langsung digunakan setelah melalui proses pra-pemrosesan. Data kemudian diekstraksi dan disesuaikan dengan kebutuhan penelitian. Pemilihan data sekunder dari Kaggle dilakukan karena dataset yang tersedia bersifat terbuka, mudah diakses, serta relevan dengan tujuan penelitian. Data tersebut selanjutnya digunakan sebagai bahan utama dalam proses analisis menggunakan metode Gaussian Mixture Models (GMM) dengan algoritma Expectation–Maximization (EM) untuk mengidentifikasi pola klasterisasi tingkat biaya pendidikan dan biaya hidup mahasiswa internasional.

	Country	City	University	Program	Level	Duration_Y	Tuition_US	Living_Cos	Rent_USD	Visa_Fee_I	Insurance_	Exchange_I
2	USA	Cambridge	Harvard Ur	Computer	Master	2	55400	83.5	2200	160	1500	1
3	UK	London	Imperial C	Data Scier	Master	1	41200	75.8	1800	485	800	0.79
4	Canada	Toronto	University	Business A	Master	2	38500	72.5	1600	235	900	1.35
5	Australia	Melbourne	University	Engineerin	Master	2	42000	71.2	1400	450	650	1.52
6	Germany	Munich	Technical	Mechanica	Master	2	500	70.5	1100	75	550	0.92
7	Japan	Tokyo	University	Informatio	Master	2	8900	76.4	1300	220	750	145.8
8	Netherlan	Amsterdar	University	Artificial In	Master	1	15800	73.2	1500	180	720	0.92
9	Singapore	Singapore	National U	Finance	Master	1.5	35000	81.1	1900	90	800	1.34
10	France	Paris	Sorbonne I	Internatior	Master	2	4500	74.6	1400	99	650	0.92

Gambar 3. 1 Data Mahasiswa Internasional

3.4 Tahapan Penelitian

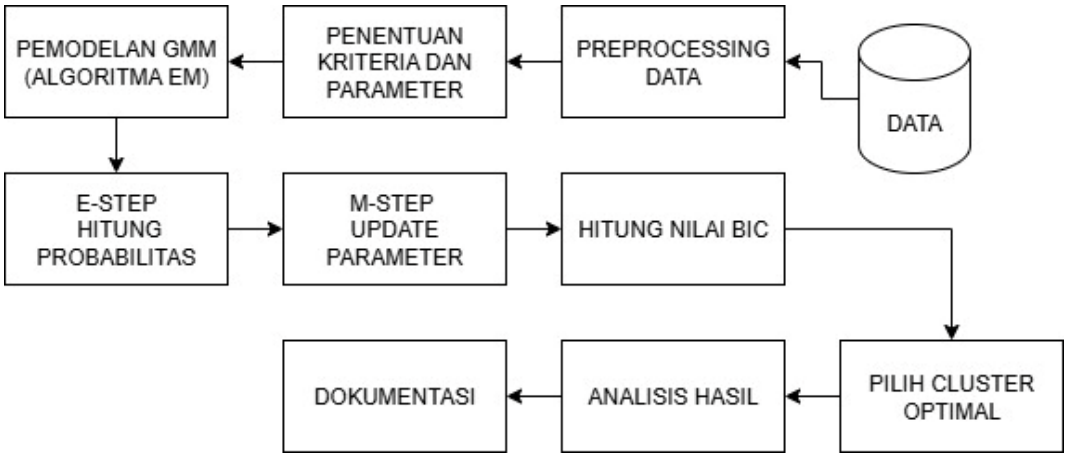
Tahapan penelitian merupakan rangkaian proses yang dilakukan untuk mengolah dan menganalisis data yang telah dikumpulkan dengan tujuan menjawab rumusan masalah penelitian. Penelitian ini disusun secara sistematis mulai dari tahap pengumpulan data, *preprocessing*, pemodelan menggunakan Gaussian Mixture Models (GMM), hingga tahap evaluasi dan interpretasi hasil.

Penelitian ini menggunakan teknik analisis data kuantitatif dengan menerapkan metode GMM untuk mengidentifikasi pola tingkat biaya pendidikan dan biaya hidup mahasiswa internasional berdasarkan variabel-variabel biaya yang telah ditentukan. Estimasi parameter model dilakukan menggunakan algoritma

Expectation–Maximization (EM), sedangkan penentuan jumlah cluster optimal dilakukan dengan menggunakan Bayesian Information Criterion (BIC).

Melalui tahapan penelitian ini, data biaya mahasiswa yang meliputi biaya pendidikan (*tuition fee*), biaya akomodasi, konsumsi, transportasi, serta komponen pengeluaran lainnya dianalisis untuk menghasilkan pengelompokan yang merepresentasikan variasi tingkat biaya secara objektif dan terukur. Hasil klasterisasi ini diharapkan dapat memberikan gambaran distribusi tingkat biaya mahasiswa internasional berdasarkan karakteristik yang serupa.

Berikut adalah gambar dari tahapan penelitian ini:



Gambar 3. 2 Tahapan Penelitian

1. Pengumpulan Data

Tahap ini dilakukan dengan mengumpulkan data sekunder terkait biaya pendidikan (tuition fee) dan biaya hidup (living cost) mahasiswa internasional yang diperoleh dari platform Kaggle. Data yang digunakan mencakup berbagai komponen pengeluaran mahasiswa, seperti biaya pendidikan, biaya akomodasi, konsumsi, transportasi, serta pengeluaran lainnya yang relevan dalam menggambarkan tingkat biaya mahasiswa.

2. Preprocessing Data

Setelah proses pengumpulan data selesai, tahap selanjutnya adalah preprocessing data. Tahap ini bertujuan untuk memastikan bahwa data berada dalam kondisi bersih, konsisten, dan siap digunakan dalam proses pemodelan. Tahapan preprocessing yang dilakukan meliputi:

1. Data Cleaning: Melakukan pemeriksaan terhadap data yang tidak lengkap, inkonsisten, atau duplikat. Data yang bermasalah kemudian dihapus atau disesuaikan agar tidak memengaruhi hasil analisis.
2. Data Integration: Menggabungkan seluruh data biaya ke dalam satu dataset terstruktur sehingga seluruh variabel penelitian berada dalam satu basis data yang terpadu.
3. Data Transformation: Menyesuaikan format data agar sesuai dengan kebutuhan analisis, termasuk memastikan seluruh variabel biaya berada dalam bentuk numerik yang dapat diproses oleh model.
4. Data Normalization: Melakukan penyesuaian skala pada variabel numerik seperti biaya pendidikan, biaya akomodasi, konsumsi, dan transportasi, sehingga setiap variabel memiliki kontribusi yang seimbang dalam proses clustering.

Tahap preprocessing ini dilakukan untuk meningkatkan kualitas data serta meminimalkan potensi distorsi dalam hasil pengelompokan menggunakan Gaussian Mixture Models (GMM).

3. Penentuan Kriteria dan Parameter

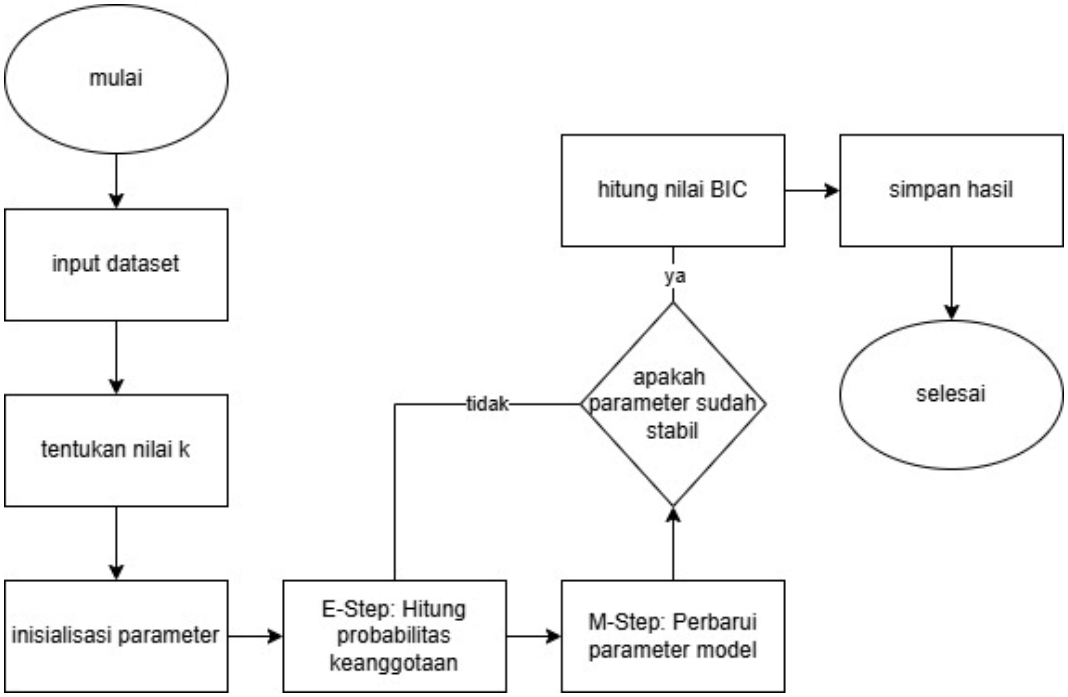
Tahap ini dilakukan untuk menentukan kriteria dan parameter yang digunakan dalam proses pemodelan dan pengelompokan data agar sesuai dengan tujuan penelitian.

1. Kriteria Cluster: Menentukan atribut yang digunakan sebagai dasar pengelompokan, yaitu biaya pendidikan (tuition fee) dan berbagai komponen biaya hidup seperti akomodasi, konsumsi, dan transportasi. Variabel-variabel tersebut dipilih karena merepresentasikan tingkat beban biaya yang ditanggung mahasiswa internasional.
2. Parameter Model: Menentukan jumlah cluster (k) yang akan diuji dalam Gaussian Mixture Models (GMM), serta menetapkan parameter awal dalam proses estimasi menggunakan algoritma Expectation–Maximization (EM). Jumlah cluster optimal kemudian ditentukan berdasarkan nilai Bayesian Information Criterion (BIC).

4. Proses Gaussian Mixture Models dengan Algoritma Expectation–Maximization

Gaussian Mixture Models (GMM) digunakan dalam penelitian ini untuk mengelompokkan data biaya pendidikan dan biaya hidup mahasiswa internasional ke dalam beberapa cluster berdasarkan karakteristik yang serupa.

Proses ini bertujuan untuk mengidentifikasi pola tingkat biaya mahasiswa, seperti kelompok dengan biaya rendah, sedang, dan tinggi. Estimasi parameter dalam model dilakukan menggunakan algoritma Expectation–Maximization (EM), sehingga setiap mahasiswa memiliki probabilitas keanggotaan pada masing-masing cluster.



Gambar 3. 3 Alur pemodelan GMM dan EM

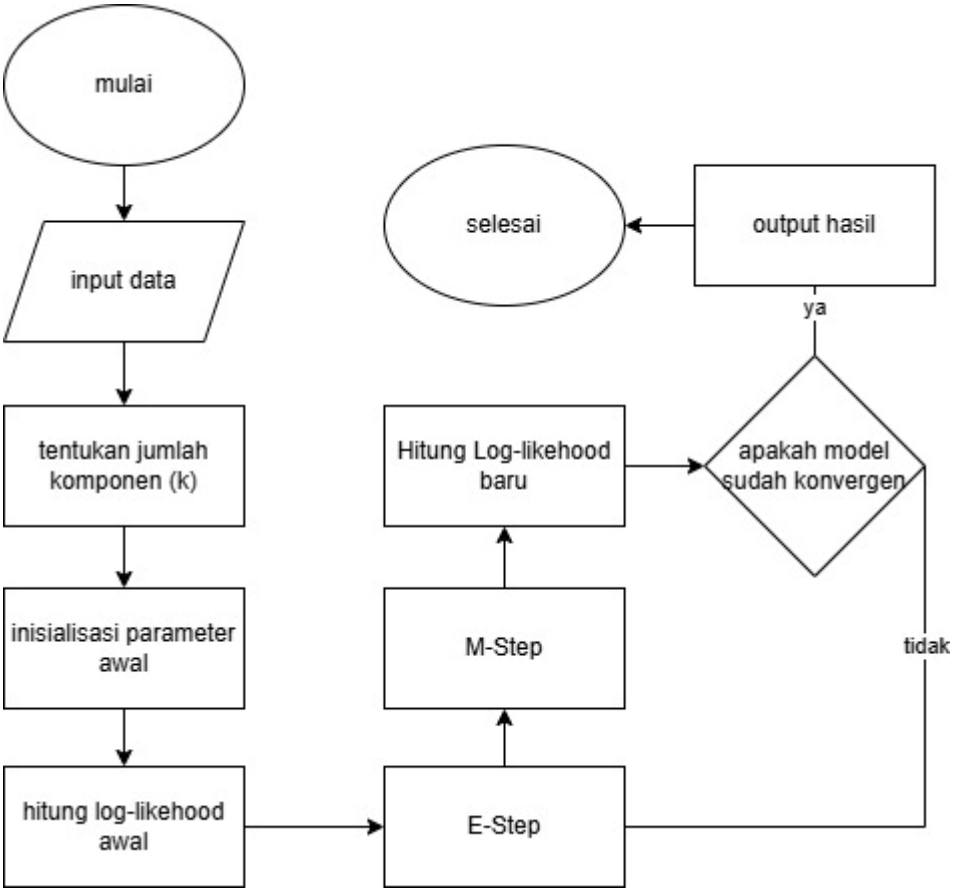
Berdasarkan Gambar 3.3 penjelasan dari alur pemodelan GMM dan EM adalah sebagai berikut:

1. Menentukan jumlah komponen (k) yang akan digunakan dalam proses pemodelan Gaussian Mixture Model (GMM) sebagai dasar pembentukan cluster.
2. Melakukan inisialisasi parameter awal untuk setiap komponen, yang meliputi nilai rata-rata (mean/μ), matriks kovarians (Σ), dan bobot campuran (π).
3. Melakukan tahap Expectation (E-step), yaitu menghitung probabilitas keanggotaan setiap data terhadap masing-masing komponen berdasarkan parameter sementara yang telah ditentukan.
4. Melakukan tahap Maximization (M-step), yaitu memperbarui parameter model (mean, kovarians, dan bobot campuran) berdasarkan hasil probabilitas pada tahap sebelumnya.

5. Melakukan pengecekan apakah parameter model sudah stabil. Jika belum stabil, maka proses dikembalikan ke tahap E-step dan M-step untuk dilakukan iterasi ulang.
6. Jika parameter telah stabil, maka dilakukan perhitungan nilai Bayesian Information Criterion (BIC) untuk mengevaluasi kualitas model berdasarkan jumlah komponen yang diuji.
7. Menyimpan hasil clustering yang diperoleh sebagai output akhir dari proses pemodelan.

5. Proses *Gaussian Mixture Models*

Algoritma Gaussian Mixture Model (GMM) digunakan untuk membentuk klaster berdasarkan pendekatan probabilistik dengan mengasumsikan bahwa data berasal dari campuran beberapa distribusi Gaussian. Metode ini mampu merepresentasikan variasi bentuk dan sebaran data secara lebih fleksibel karena setiap data memiliki probabilitas keanggotaan pada masing-masing klaster. GMM efektif digunakan untuk mengidentifikasi pola yang tidak dapat dipisahkan secara tegas serta menangani data dengan distribusi yang kompleks.



Gambar 3.4 Flowchart Gaussian Mixture Models

Berdasarkan gambar 3.4 penjelasan dari Flowchart Gaussian Mixture Models adalah sebagai berikut:

1. Menginput data yang telah melalui tahap preprocessing ke dalam sistem sebagai dataset yang akan diproses menggunakan metode Gaussian Mixture Model (GMM).
2. Menentukan jumlah komponen (k) yang akan digunakan dalam proses pemodelan sebagai dasar pembentukan cluster.
3. Melakukan inisialisasi parameter awal untuk setiap komponen, yang meliputi nilai rata-rata (mean/μ), matriks kovarians (Σ), serta bobot campuran (π).

4. Menghitung nilai log-likelihood awal, yaitu menghitung nilai fungsi likelihood berdasarkan parameter awal untuk mengetahui kondisi awal model.
5. Melakukan tahap Expectation (E-step), yaitu menghitung probabilitas keanggotaan setiap data terhadap masing-masing komponen berdasarkan parameter model sementara.
6. Melakukan tahap Maximization (M-step), yaitu memperbarui parameter model (mean, kovarians, dan bobot campuran) menggunakan hasil probabilitas yang diperoleh pada tahap E-step.
7. Menghitung nilai log-likelihood baru setelah parameter diperbarui untuk melihat peningkatan nilai likelihood pada iterasi tersebut.
8. Melakukan pengecekan apakah model sudah konvergen, yaitu dengan membandingkan perubahan nilai log-likelihood antara iterasi sebelumnya dan iterasi saat ini. Jika perubahan masih signifikan, maka proses dikembalikan ke tahap E-step dan M-step untuk dilakukan iterasi ulang.
9. Jika model telah konvergen, maka proses iterasi dihentikan dan hasil pengelompokan disimpan sebagai output akhir dari pemodelan.

6. Analisis Hasil

Analisis hasil merupakan tahap akhir dalam penelitian yang bertujuan untuk mengevaluasi hasil pengelompokan yang diperoleh dari Gaussian Mixture Models (GMM). Evaluasi model dilakukan menggunakan Bayesian Information Criterion (BIC) untuk menentukan jumlah komponen (cluster) yang paling optimal. Nilai BIC digunakan karena mempertimbangkan keseimbangan antara tingkat kecocokan model dan kompleksitasnya, sehingga dapat menghindari *overfitting*.

Setelah diperoleh jumlah cluster yang optimal berdasarkan nilai BIC terendah, dilakukan analisis terhadap karakteristik masing-masing cluster berdasarkan variabel biaya pendidikan (*tuition fee*) dan komponen biaya hidup, seperti akomodasi, konsumsi, dan transportasi. Analisis ini bertujuan untuk mengidentifikasi pola tingkat biaya yang terbentuk pada setiap kelompok mahasiswa internasional, seperti kelompok dengan tingkat biaya rendah, sedang, dan tinggi. Hasil analisis kemudian digunakan untuk memberikan gambaran mengenai variasi tingkat biaya mahasiswa internasional serta menilai efektivitas model GMM dalam merepresentasikan struktur data biaya secara objektif dan terukur.

3.5 Jadwal Penelitian

Penelitian ini dilaksanakan di Universitas Muhammadiyah Sumatera Utara (UMSU) untuk mengklasterisasi tingkat biaya pendidikan dan biaya hidup mahasiswa internasional menggunakan algoritma *Gaussian Mixture Models* (GMM). Melalui penerapan algoritma GMM, penelitian ini bertujuan untuk membantu pengelola program internasional dalam mengelompokkan tingkat biaya pendidikan dan biaya hidup mahasiswa internasional, sehingga hasilnya dapat dijadikan dasar dalam merumuskan kebijakan dukungan finansial yang lebih tepat sasaran.

BAB IV

HASIL DAN PEMBAHASAN

4.1 Gambaran Umum Dataset

Dataset yang digunakan dalam penelitian ini merupakan data sekunder yang tersedia secara publik melalui *platform kaggle* dengan nama dataset *Cost of Internasional Education*. Dataset ini terdiri dari 1407 entri data yang mencakup informasi biaya kuliah, biaya hidup, biaya sewa, dan informasi lainnya dari berbagai program studi di berbagai universitas dan negara.

Tabel 4. 1 Data Set asli sebelum diolah

No	Country	City	University	Program	..	Tuition	Living Cost	..	Visa Fee
1	USA	Cambridge	Harvard university	Computer science	..	554000	83.5	..	160
2	UK	London	Imperial college london	Data science	..	41200	75.8	..	485
3	Canada	Toronto	University of toronto	Business analytics	..	38500	72.5	..	235
4	Australia	Melbourne	University of melbourne	Engineering	..	42000	71.2	..	450
5	Germany	Munich	Technical university of munich	Mechanical engineering	..	500	70.5	..	75
6	Japan	Tokyo	University of tokyo	Information science	..	8900	76.4	..	220

Dataset ini terdiri dari dua belas kolom, yaitu *country*, *city*, *university*, *program*, *level*, *duration years*, *tuition fee*, *living cost index*, *rent*, *visa fee*, *insurance*, serta *exchange rate* yang bersumber dari *platform Kaggle* yang mencakup data universitas dari berbagai negara di seluruh dunia.

Tabel 4. 2 Fitur Data Set Yang Digunakan

no	Country	University	Program	Tuition _usd	Living_cost_ index	Rent_usd	Visa fee
1	USA	Harvard university	Computer science	554000	83.5	2200	160
2	UK	Imperial college london	Data science	41200	75.8	1800	485
3	Canada	University of toronto	Business analytics	38500	72.5	1600	235
4	Australia	University of melbourne	Engineering	42000	71.2	1400	450
5	Germany	Technical university of munich	Mechanical engineering	500	70.5	1100	75
6	Japan	University of tokyo	Information science	8900	76.4	1300	220

Tahap berikutnya adalah pemilihan atribut berdasarkan fitur-fitur dataset yang relevan untuk digunakan dalam proses klasterisasi tingkat *tuition fee* dan *living cost* mahasiswa internasional.

Tabel 4. 3 variabel yang digunakan

No	Variabel	Keterangan
1	*Country*	Negara tempat universitas berada
2	*University*	Nama universitas yang menawarkan program studi
3	*Program*	Nama program studi yang ditawarkan
4	*Tuition_usd*	Biaya kuliah yang dibayarkan dalam satuan dolar amerika (usd)
5	*Living_cost_index*	Indeks biaya hidup di kota tempat universitas berada
6	*Rent_usd*	Biaya sewa tempat tinggal perbulan dalam satuan dolar amerika(usd)
7	*Visa_fee*	Biaya pengurusan visa pelajar dalam satuan dolar amerika (usd)

Jumlah data yang digunakan dalam penelitian ini adalah sebanyak 1408 data universitas. Setiap data terdiri dari beberapa atribut yang dimanfaatkan dalam proses analisis, antara lain *country, university, program, tuition_usd, living_cost_index, rent_usd, dan visa_fee.*

4.2 Preprocessing Data

Preprocessing data merupakan tahapan awal yang sangat penting dalam proses pengolahan data untuk machine learning. Tahapan ini dilakukan dengan memanfaatkan bahasa pemrograman python, dengan tujuan untuk mempersiapkan data mentah agar memiliki format yan gsesuai dan relevan untuk dianalisis lebih lanjut. Proses preprocessing mencakup beberapa tahap, yaitu cleaning data, integrasi data, transformasi data serta normalisasi data. Kualitas dari tahap preprocessing ini sangat menentukan kinerja model dalam mengklasterisasikan tuition fee dan living cost mahasiswa internasional menggunakan metode gaussian mixture models.

4.2.1 Data Cleaning

Dataset yang diperoleh dari platform Kaggle telah memiliki kualitas dan konsistensi yang baik, sehingga proses pembersihan data (*data cleaning*) tidak dilakukan dalam penelitian ini.

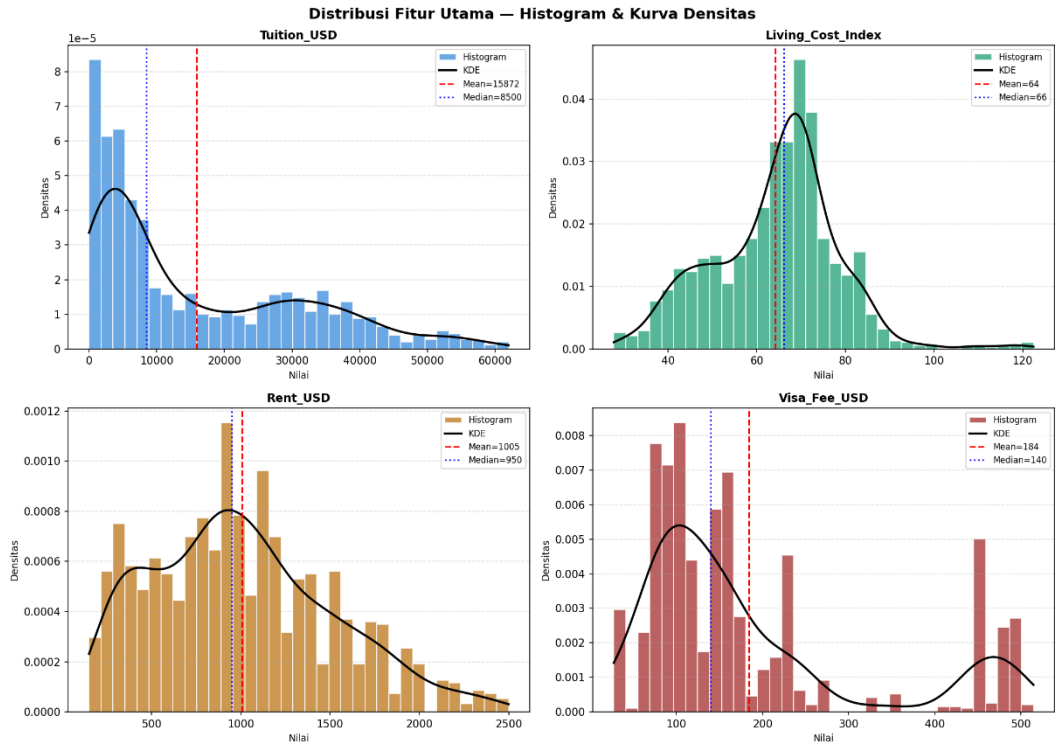
```
Baris awal          : 1,407
Baris akhir         : 1,407
Selisih             : 0 baris

Tuition_USD:
  Sebelum → Mean:   15871.52 | Std:   15495.74 | Min:     0.00 | Max:   62000.00
  Sesudah → Mean:   15871.52 | Std:   15495.74 | Min:     0.00 | Max:   62000.00

Living_Cost_Index:
  Sebelum → Mean:      64.22 | Std:      13.99 | Min:    27.80 | Max:     122.40
  Sesudah → Mean:      64.22 | Std:      13.99 | Min:    27.80 | Max:     122.40

Rent_USD:
  Sebelum → Mean:    1005.43 | Std:     508.55 | Min:   150.00 | Max:    2500.00
  Sesudah → Mean:    1005.43 | Std:     508.55 | Min:   150.00 | Max:    2500.00

Visa_Fee_USD:
  Sebelum → Mean:     183.94 | Std:     133.98 | Min:    27.35 | Max:     514.59
  Sesudah → Mean:     183.94 | Std:     133.98 | Min:    27.35 | Max:     514.59
```

Gambar 4. 1 hasil verifikasi kualitas data

Guna memverifikasi kualitas data, dilakukan pemeriksaan statistik deskriptif pada seluruh fitur numerik yang digunakan, yaitu *Tuition_USD*, *Living_Cost_Index*, *Rent_USD*, dan *Visa_Fee_USD*. Berdasarkan hasil pemeriksaan, nilai *mean*, *standard deviation*, *minimum*, dan *maximum* pada kondisi sebelum dan sesudah pemeriksaan tidak menunjukkan adanya perbedaan. Jumlah baris data pun tetap sama, yakni sebanyak 1.407 baris tanpa selisih satu baris pun. Kondisi tersebut menunjukkan bahwa dataset yang diperoleh dari platform Kaggle sudah dalam keadaan bersih dan bebas dari anomali, sehingga proses *data cleaning* tidak dilakukan dan analisis dapat langsung diteruskan ke tahap selanjutnya.

4.2.2 Distribusi Data

Sebelum data dikelompokkan menggunakan metode klasterisasi, dilakukan terlebih dahulu analisis deskriptif untuk melihat bagaimana karakteristik dan sebaran dari data yang digunakan. Pada tahap ini, keempat fitur numerik yang ada yaitu *Tuition_USD*, *Living_Cost_Index*, *Rent_USD*, dan *Visa_Fee_USD* masing-masing divisualisasikan menggunakan histogram dan kurva densitas (*Kernel Density Estimation*/KDE). Dari hasil visualisasi tersebut, terlihat bahwa hampir semua fitur memiliki sebaran data yang tidak merata dan cenderung condong ke kanan (*right-skewed*), di mana nilai *mean* pada tiap fitur lebih tinggi dibandingkan nilai *median*-nya. Hal ini menunjukkan bahwa terdapat kesenjangan nilai yang cukup besar dalam data, sehingga normalisasi data perlu dilakukan sebelum masuk ke tahap klasterisasi. Hasil visualisasi dapat dilihat pada gambar berikut:



Gambar 4.2 Distribusi Fitur Utama

Berikut penjelasan masing masing grafik yang ditampilkan didalam gambar:

1. Grafik 1 (Tuition_USD)

   Dari grafik ini terlihat bahwa biaya kuliah antar universitas sangat bervariasi. Kebanyakan universitas memiliki biaya kuliah yang tidak terlalu tinggi, yaitu berkisar di bawah $10.000, namun ada juga beberapa yang mencapai $62.000. Perbedaan yang cukup jauh antara nilai mean ($15.872) dan median ($8.500) menunjukkan bahwa ada beberapa universitas dengan biaya kuliah yang sangat mahal dan membuat rata-rata keseluruhan menjadi tinggi.

2. Grafik 2 (Living_Cost_Index)

   Dibandingkan fitur lainnya, distribusi indeks biaya hidup ini terlihat lebih seimbang dan tidak terlalu condong ke salah satu sisi. Nilai mean (64) dan median (66) yang hampir sama menunjukkan bahwa biaya hidup di

sebagian besar negara tidak terlalu jauh berbeda. Hanya sedikit negara yang memiliki indeks biaya hidup sangat tinggi, yaitu sekitar 120.

3. Grafik 3 (Rent_USD)

   Biaya sewa bulanan kebanyakan berada di rentang $150 hingga $1.500 dengan puncaknya di sekitar $800–$1.000. Nilai mean ($1.005) dan median ($950) yang tidak terlalu berbeda menandakan bahwa sebaran data ini cukup merata, meski ada beberapa kota dengan biaya sewa yang cukup tinggi sehingga sedikit mendorong nilai rata-rata ke atas.

4. Grafik 4 (Visa_Fee_USD)

   Yang menarik dari grafik ini adalah adanya dua puncak distribusi, yang artinya biaya visa terbagi menjadi dua kelompok yang cukup berbeda. Kelompok pertama memiliki biaya visa di kisaran $100–$150, sedangkan kelompok kedua berada di kisaran $400–$500. Selisih antara mean ($184) dan median ($140) menunjukkan bahwa ada sejumlah negara dengan biaya visa yang cukup mahal dan menarik nilai rata-rata ke atas.

4.2.3 Featuring Engineering Data

Karena algoritma *Gaussian Mixture Models* (GMM) bekerja dengan pendekatan probabilistik, yaitu setiap data point dihitung peluangnya untuk masuk ke dalam masing-masing klaster berdasarkan distribusi Gaussian yang terbentuk, maka keempat fitur numerik dalam dataset digunakan langsung sebagai variabel klasterisasi. Hal ini dikarenakan setiap baris data sudah merepresentasikan satu universitas secara lengkap beserta informasi biaya yang menyertainya. Keempat fitur yang digunakan sebagai variabel klasterisasi adalah *Tuition_USD*, *Living_Cost_Index*, *Rent_USD*, dan *Visa_Fee_USD*.

4.2.4 Statistik Deskriptif Komprehensif

Sebelum masuk ke tahap klasterisasi, dilakukan perhitungan statistik deskriptif terlebih dahulu untuk melihat gambaran umum dari data yang digunakan. Data yang dianalisis berjumlah 1.407 entri universitas dengan empat fitur utama yaitu *Tuition_USD*, *Living_Cost_Index*, *Rent_USD*, dan *Visa_Fee_USD*.

Dari hasil perhitungan, terlihat bahwa biaya kuliah (*Tuition_USD*) memiliki variasi yang sangat besar, dengan nilai terendah $0 dan tertinggi mencapai $62.000. Rata-ratanya berada di angka $15.872, namun mediannya hanya $8.500, yang artinya banyak universitas yang sebenarnya memiliki biaya kuliah di bawah rata-rata. Kondisi serupa juga terjadi pada *Visa_Fee_USD* di mana rata-ratanya ($184) lebih tinggi dari mediannya ($140). Hal ini sejalan dengan nilai *skewness* kedua fitur tersebut yang miring ke kanan (*right-skewed*). Sementara itu, *Living_Cost_Index* dan *Rent_USD* memiliki sebaran yang lebih merata karena nilai *skewness*-nya mendekati nol. Hasil statistik deskriptif selengkapnya dapat dilihat pada gambar berikut:

Fitur	Tuition_USD	Living_Cost_Index	Rent_USD	Visa_Fee_USD
N	1.407000e+03	1407.0000	1407.0000	1407.0000
Mean	1.587152e+04	64.2200	1005.4300	183.9400
Median	8.500000e+03	66.2000	950.0000	140.0000
Std Dev	1.549574e+04	13.9900	508.5500	133.9800
Variance	2.401180e+08	195.6600	258618.2800	17950.9500
Min	0.000000e+00	27.8000	150.0000	27.3500
Max	6.200000e+04	122.4000	2500.0000	514.5900
Q1 (25%)	3.500000e+03	55.4000	600.0000	90.0000
Q3 (75%)	2.804552e+04	72.3000	1300.0000	229.2000
IQR	2.454552e+04	16.9000	700.0000	139.2000
Skewness	9.251000e-01	0.0691	0.4988	1.3047
Kurtosis	2.751400e+00	3.8575	2.7311	3.4155

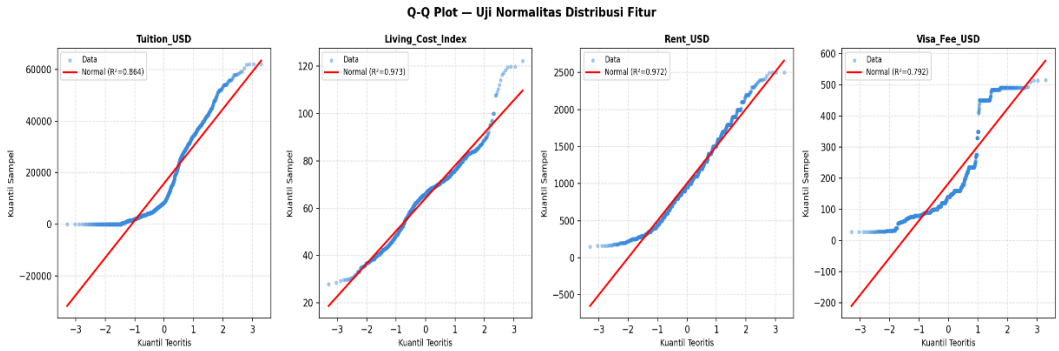
Interpretasi Skewness:
- Tuition_USD : skewness = +0.9251 → distribusi miring ke kanan (right-skewed)
- Living_Cost_Index : skewness = +0.0691 → distribusi mendekati simetris
- Rent_USD : skewness = +0.4988 → distribusi mendekati simetris
- Visa_Fee_USD : skewness = +1.3047 → distribusi miring ke kanan (right-skewed)

Gambar 4.3 Statistik Deskriptif Dari 4 Fitur Utama

```python
def descriptive_statistics(df: pd.DataFrame) -> pd.DataFrame:
    print("\n" + "=" * 70)
    print("  TAHAP 2 — STATISTIK DESKRIPTIF KOMPREHENSIF")
    print("=" * 70)
    records = []
    for col in FEATURES:
        s = df[col]
        records.append({
            'Fitur'          : col,
            'N'              : len(s),
            'Mean'           : round(s.mean(), 2),
            'Median'         : round(s.median(), 2),
            'Std Dev'        : round(s.std(), 2),
            'Variance'       : round(s.var(), 2),
            'Min'            : round(s.min(), 2),
            'Max'            : round(s.max(), 2),
            'Q1 (25%)'       : round(s.quantile(0.25), 2),
            'Q3 (75%)'       : round(s.quantile(0.75), 2),
            'IQR'            : round(s.quantile(0.75) - s.quantile(0.25), 2),
            'Skewness'       : round(stats.skew(s.dropna()).item(), 4),
            'Kurtosis'       : round(stats.kurtosis(s.dropna()).item() + 3, 4),
        })
    stat_df = pd.DataFrame(records).set_index('Fitur')
    pd.set_option('display.max_columns', None)
    pd.set_option('display.width', 120)
    print(f"\n{stat_df.T.to_string()}\n")
```

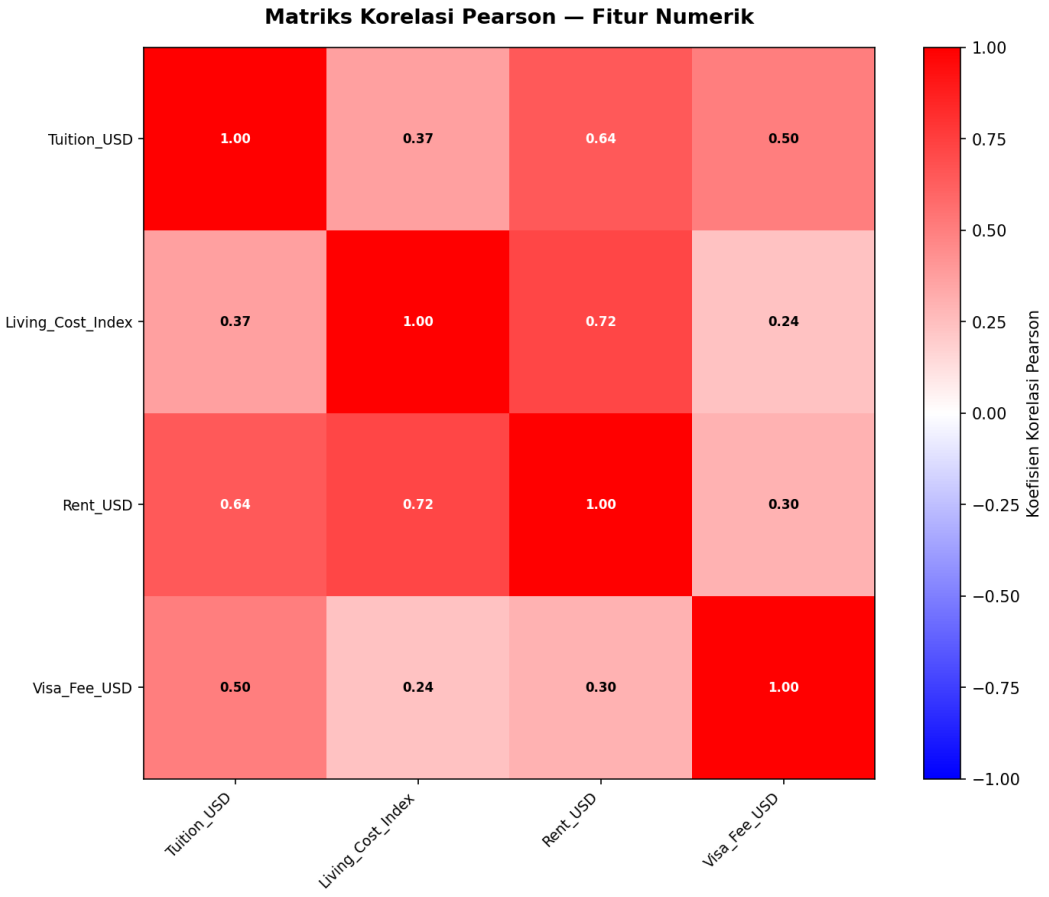
Gambar 4. 4 Code statistik deskriptif

Berdasarkan gambar 4.4 fungsi descriptive_statistics digunakan untuk menghitung berbagai ukuran statistik dari setiap fitur numerik secara otomatis, mulai dari mean, median, standart deviation, variance, min, max, kuartil q1, q3, IQR, skewness hingga kurtosis.



Gambar 4. 5 Q-Q Plot

Uji normalitas dilakukan menggunakan Q-Q Plot (*Quantile-Quantile Plot*) untuk melihat sejauh mana distribusi setiap fitur mendekati distribusi normal. Dari hasil plot tersebut, terlihat bahwa fitur *Living_Cost_Index* (R^2=0,973) dan *Rent_USD* (R^2=0,972) memiliki titik-titik data yang cukup dekat mengikuti garis normal, yang artinya kedua fitur ini memiliki distribusi yang mendekati normal. Sementara itu, fitur *Tuition_USD* (R^2=0,864) dan *Visa_Fee_USD* (R^2=0,792) menunjukkan adanya penyimpangan yang lebih besar dari garis normal, terutama di bagian ekor distribusi, yang konsisten dengan nilai *skewness* yang tinggi pada analisis sebelumnya. Hasil uji normalitas ini menjadi salah satu pertimbangan dalam pemilihan metode *Gaussian Mixture Models* (GMM), karena GMM mengasumsikan bahwa data dalam setiap klaster mengikuti distribusi Gaussian.

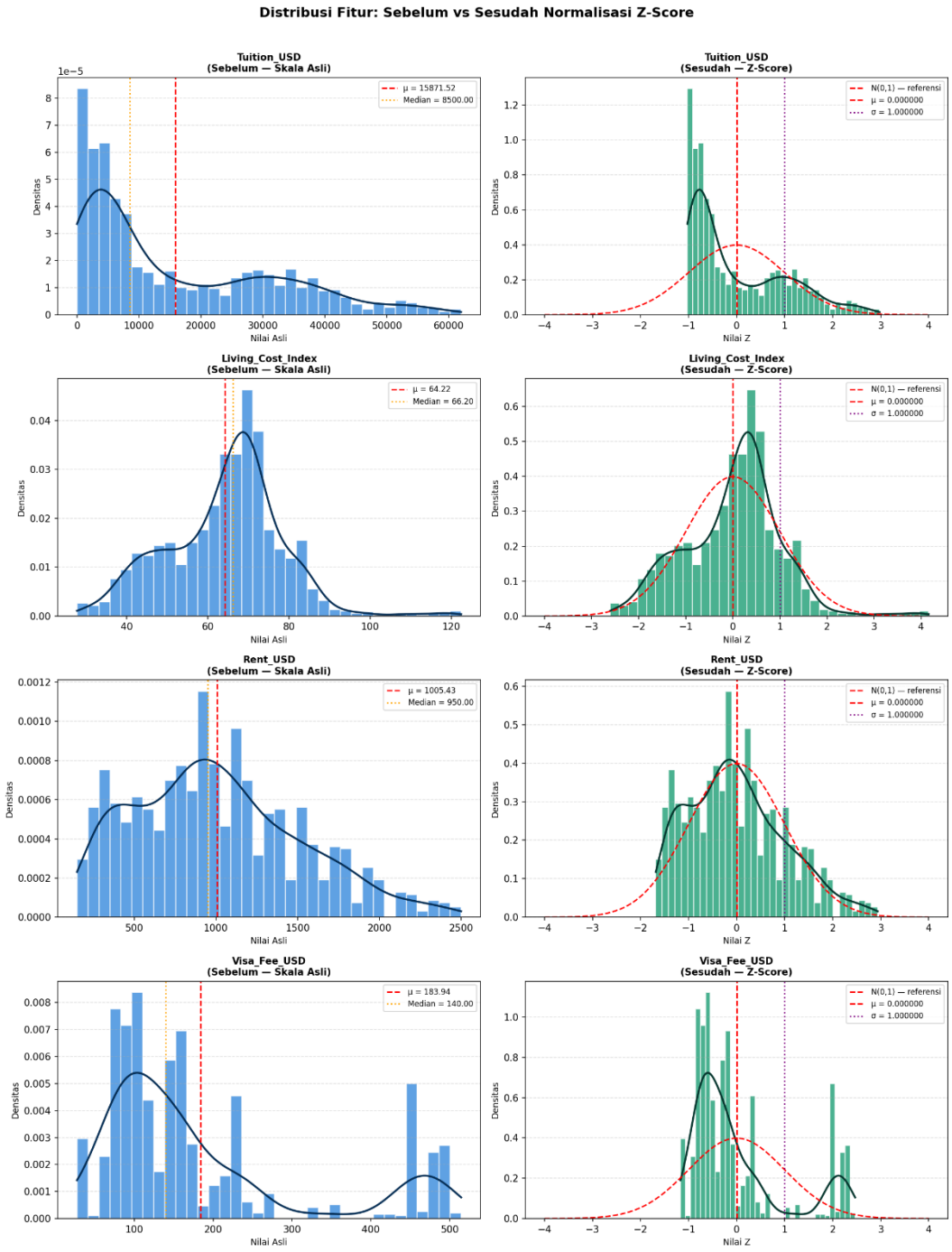


Gambar 4.6 Korelasi Fitur Numerik

Berdasarkan Gambar 4.6, matriks korelasi Pearson digunakan untuk melihat hubungan antar keempat fitur yang digunakan. Dari hasil matriks tersebut, seluruh fitur menunjukkan korelasi positif satu sama lain yang terlihat dari warna merah yang mendominasi seluruh sel matriks. Korelasi paling kuat terjadi antara *Rent_USD* dan *Living_Cost_Index* dengan nilai 0,72, yang menunjukkan bahwa negara dengan indeks biaya hidup tinggi cenderung memiliki biaya sewa yang tinggi pula. *Tuition_USD* juga berkorelasi cukup kuat dengan *Rent_USD* (0,64) dan *Visa_Fee_USD* (0,50). Sebaliknya, *Visa_Fee_USD* memiliki korelasi yang paling rendah dengan fitur lainnya, khususnya dengan *Living_Cost_Index* (0,24) dan *Rent_USD* (0,30), yang menunjukkan bahwa biaya visa tidak terlalu berkaitan dengan tingginya biaya hidup atau biaya sewa di suatu negara.

4.2.5 Normalisasi Data (Z- Score)

Sebelum data diproses menggunakan metode *Gaussian Mixture Models* (GMM), dilakukan normalisasi data terlebih dahulu dengan menggunakan *Z-Score Standardization*. *Z-Score* digunakan karena GMM sangat sensitif terhadap perbedaan skala antar fitur, dan apabila normalisasi tidak dilakukan, fitur dengan rentang nilai yang jauh lebih besar seperti *Tuition_USD* berpotensi mendominasi proses klasterisasi dan membuat hasil yang diperoleh menjadi tidak optimal. Melalui normalisasi *Z-Score*, setiap fitur akan memiliki rata-rata 0 dan standar deviasi 1, yang membuat seluruh fitur berada pada skala yang setara dan dapat memberikan kontribusi yang seimbang dalam pembentukan klaster.



Gambar 4. 7 Normalisasi data sebelum dan sesudah menggunakan Z-Score

Berdasarkan Gambar 4.7, ditampilkan perbandingan distribusi setiap fitur sebelum dan sesudah dilakukan normalisasi *Z-Score*. Kolom kiri menunjukkan distribusi data pada skala aslinya, sedangkan kolom kanan menunjukkan distribusi data setelah dinormalisasi dengan nilai rata-rata 0 dan standar deviasi 1.

Pada fitur *Tuition_USD*, sebelum normalisasi data terlihat sangat condong ke kanan dengan sebagian besar nilai terkonsentrasi di bawah $10.000 namun memiliki ekor yang sangat panjang hingga $60.000. Setelah normalisasi, data sudah berada pada skala Z namun pola *right-skewed* masih tetap terlihat, yang menandakan bahwa normalisasi tidak mengubah bentuk distribusi melainkan hanya menyeragamkan skalanya.

Fitur *Living_Cost_Index* sebelum normalisasi memiliki distribusi yang relatif lebih merata dibandingkan fitur lainnya. Setelah normalisasi, distribusinya terlihat cukup mendekati kurva normal referensi N(0,1) yang ditunjukkan oleh garis merah putus-putus, menjadikan fitur ini yang paling mendekati distribusi normal di antara keempat fitur.

Pada fitur *Rent_USD*, distribusi sebelum normalisasi terlihat cukup merata di rentang $150 hingga $2.500. Setelah normalisasi, distribusinya juga cukup mendekati kurva normal referensi meskipun masih terdapat sedikit kemiringan ke kanan.

Sementara itu, fitur *Visa_Fee_USD* memperlihatkan pola yang paling berbeda, di mana sebelum normalisasi terdapat dua puncak distribusi (*bimodal*) yang terlihat jelas. Setelah normalisasi, pola *bimodal* tersebut tetap ada dan distribusinya masih cukup jauh dari kurva normal referensi, yang menunjukkan bahwa fitur ini memiliki karakteristik yang paling berbeda dibandingkan ketiga fitur lainnya.

```
  Kolom                Mean (μ)       Std Dev (σ)    Min Asli     Max Asli     Min Z
Max Z
  ---------------------------------------------------------------------------------------
--
  Tuition_USD          15871.5180     15495.7422     0.00         62000.00     -1.0243
2.9768
  Living_Cost_Index    64.2158        13.9878        27.80        122.40       -2.6034
4.1597
  Rent_USD             1005.4292      508.5453       150.00       2500.00      -1.6821
2.9389
  Visa_Fee_USD         183.9394       133.9812       27.35        514.59       -1.1687
2.4679
```

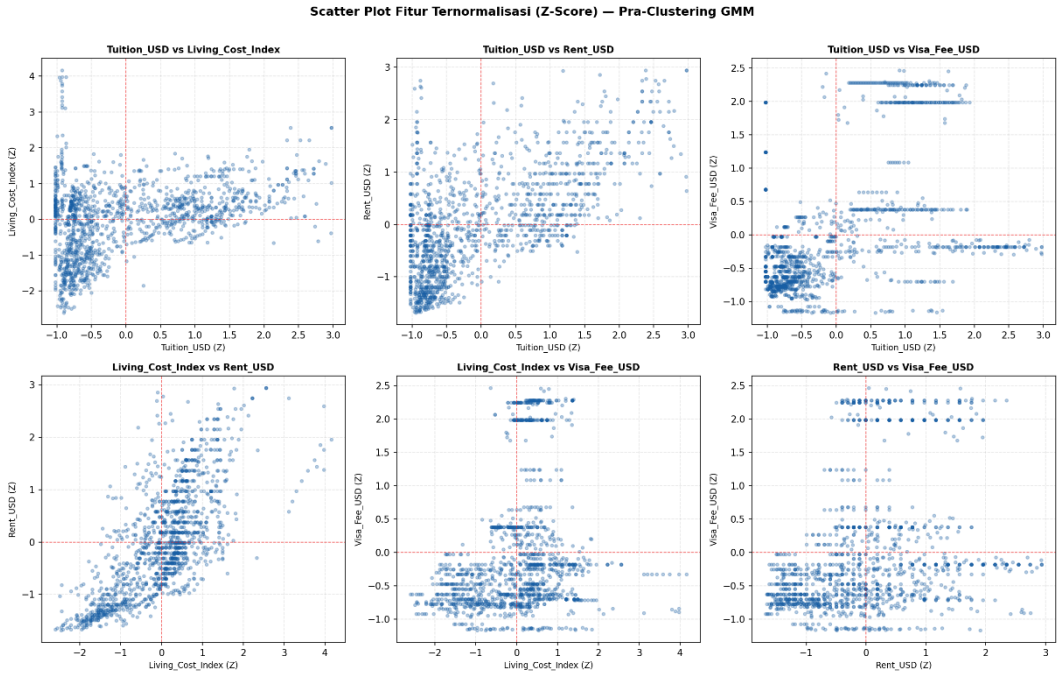
Gambar 4.8 Fitur Sebelum Dinormalisasi Menggunakan Z-Score

```
  Kolom                Mean           Std Dev        Verifikasi Mean    Verifikasi Std
  ---------------------------------------------------------------------------------------
  Tuition_USD          0.00000000     1.00000000     PASS               PASS
  Living_Cost_Index    0.00000000     1.00000000     PASS               PASS
  Rent_USD             0.00000000     1.00000000     PASS               PASS
  Visa_Fee_USD         0.00000000     1.00000000     PASS               PASS

 Semua fitur berhasil dinormalisasi (mean≈0, std≈1).
```

Gambar 4.9 Fitur Setelah Dinormalisasi menggunakan Z-Score

Berdasarkan gambar diatas, sebelum normalisasi dilakukan, keempat fitur memiliki skala nilai yang sangat jauh berbeda. Tuition_USD misalnya, memiliki rata-rata sebesar 15.871 dengan rentang nilai antara $0 hingga $62.000, sedangkan Living_Cost_Index hanya berkisar antara 27,80 hingga 122,40. Perbedaan skala yang terlalu besar seperti ini bisa menjadi masalah apabila data langsung diproses tanpa normalisasi terlebih dahulu.

Setelah normalisasi Z-Score diterapkan, Gambar 4.9 memperlihatkan bahwa semua fitur sudah berhasil distandarisasi dengan nilai rata-rata 0 dan standar deviasi 1. Hal ini juga terlihat dari kolom verifikasi yang menunjukkan status PASS pada seluruh fitur, baik untuk mean maupun standar deviasinya. Berdasarkan hasil verifikasi tersebut, normalisasi data telah berhasil dilakukan dan proses dapat dilanjutkan ke tahap klasterisasi.



Gambar 4.10 Scatter Plot Fitur yang ternormalisasi Z-Score

Berdasarkan Gambar 4.10, ditampilkan *scatter plot* dari semua kombinasi pasangan fitur yang sudah dinormalisasi menggunakan *Z-Score* sebelum proses klasterisasi GMM dijalankan. Visualisasi ini dibuat untuk melihat bagaimana pola sebaran data antar fitur secara visual sebelum model diproses.

Pada grafik *Tuition_USD* vs *Living_Cost_Index*, data terlihat menyebar cukup merata tanpa membentuk pola yang terlalu jelas. Grafik *Tuition_USD* vs *Rent_USD* juga menunjukkan hal yang serupa, di mana data lebih banyak berkumpul di sisi kiri yang menandakan bahwa sebagian besar universitas memiliki biaya kuliah yang tidak terlalu tinggi.

Pada grafik *Tuition_USD* vs *Visa_Fee_USD* dan *Living_Cost_Index* vs *Visa_Fee_USD*, titik-titik data tampak membentuk garis horizontal yang terpisah-pisah, hal tersebut disebabkan karena *Visa_Fee_USD* memang memiliki nilai yang cenderung mengelompok di angka-angka tertentu dan sesuai dengan pola *bimodal* yang sebelumnya sudah terlihat pada histogram. Sementara itu, grafik

Living_Cost_Index vs *Rent_USD* memperlihatkan hubungan yang paling terlihat jelas dibandingkan pasangan fitur lainnya, di mana data membentuk pola yang mengarah ke kanan atas dan menandakan bahwa kedua fitur ini memang saling berkaitan. Secara keseluruhan, pola sebaran pada keenam grafik ini mengindikasikan bahwa data berpotensi untuk dikelompokkan lebih lanjut melalui proses klasterisasi GMM.

4.3 Implementasi Gaussian Mixture Models

Sebelum menjalankan algoritma *Gaussian Mixture Models* (GMM), jumlah klaster yang optimal perlu ditentukan terlebih dahulu. Rentang jumlah klaster yang diuji dalam penelitian ini adalah 2 hingga 7 klaster. Untuk menentukannya, digunakan dua metode evaluasi yaitu *BIC* (*Bayesian Information Criterion*) dan *AIC* (*Akaike Information Criterion*). Kedua metode ini dipilih karena GMM bekerja berdasarkan pendekatan probabilistik, bukan pendekatan jarak seperti pada *K-Means*, sehingga metode *Elbow* dengan *WSSE* tidak sesuai untuk digunakan. *BIC* dan *AIC* mengevaluasi seberapa baik model dalam menjelaskan data dengan tetap mempertimbangkan kompleksitas model, di mana nilai yang lebih rendah menunjukkan bahwa model yang terbentuk lebih optimal.

4.3.1 Tentukan Nilai k

```
# Rentang jumlah cluster yang diuji untuk menentukan K optimal
K_RANGE = range(2, 8)
# Hiperparameter EM
MAX_ITER    = 200
TOL         = 1e-6
N_INIT      = 5
REG_COVAR   = 1e-6
```

Gambar 4. 11 Code Menentukan Nilai K

4.3.2 Inisialisasi Parameter

```python
def _initialize_parameters(self, X: np.ndarray) -> tuple:
    N, D = X.shape
    K    = self.n_components
    means = np.zeros((K, D))
    idx_first  = np.random.randint(0, N)
    means[0]   = X[idx_first]
    for k in range(1, K):
        dists = np.array([
            min(np.sum((x - means[j]) ** 2) for j in range(k))
            for x in X
        ])
        probs = dists / dists.sum()
        idx_k = np.random.choice(N, p=probs)
        means[k] = X[idx_k]
    # Inisialisasi kovarians
    global_cov   = np.cov(X.T)
    covariances  = np.array([global_cov.copy() for _ in range(K)])
    # Inisialisasi mixing coefficient
    weights = np.ones(K) / K   # Distribusi seragam: 1/K
    return means, covariances, weights
```

Gambar 4. 12 code inisialisasi parameter

Berdasarkan Gambar 4.12, terdapat fungsi inisialisasi parameter yang bertugas menginisialisasi parameter awal sebelum algoritma GMM mulai berjalan. Fungsi ini menerima input berupa data *X* dalam bentuk array dan menghasilkan tiga parameter utama yang dibutuhkan GMM, yaitu *means* (rata-rata), *covariances* (kovarians), dan *weights* (bobot campuran).

Proses inisialisasi *means* dilakukan menggunakan pendekatan *K-Means*, di mana titik pusat pertama dipilih secara acak dari data, kemudian titik pusat berikutnya dipilih berdasarkan probabilitas yang sebanding dengan jarak kuadrat setiap data terhadap titik pusat yang sudah ada. Pendekatan ini bertujuan agar titik pusat awal yang dipilih tidak terlalu berdekatan satu sama lain, sehingga proses konvergensi GMM menjadi lebih cepat dan stabil.

Untuk inisialisasi *covariances*, setiap klaster diberikan nilai kovarians awal yang sama yaitu kovarians global dari seluruh data. Hal ini dilakukan agar setiap klaster memiliki titik awal yang masuk akal sebelum proses iterasi dimulai. Sementara itu, *weights* atau bobot setiap klaster diinisialisasi secara seragam dengan nilai 1/K, yang berarti pada awalnya setiap klaster dianggap memiliki proporsi data yang sama besar. Ketiga parameter inilah yang kemudian akan diperbarui secara iteratif melalui proses *Expectation-Maximization* (EM) hingga model GMM mencapai konvergensi.

4.3.3 E- Step

Setelah parameter awal berhasil diinisialisasi, proses selanjutnya adalah melakukan *E-Step* atau *Expectation Step*. Tahap ini merupakan langkah pertama dari proses iterasi *expectation - maximization (EM)* yang menjadi inti dari algoritma GMM. Pada *E-Step*, setiap data point dihitung nilai probabilitasnya untuk masuk ke dalam masing masing klaster berdasarkan parameter yang ada saat itu, yaitu *means, covariances,* dan *weights*. Nilai probabilitas ini disebut sebagai responsibilities, yang menunjukkan seberapa besar kemungkinan suatu data point berasal dari klaster tertentu. *E-Step* harus dilakukan karena GMM tidak langsung menentukan keanggotaaan klaster secara pasti seperti *k-means*, melainkan menghitung peluang keanggotaan setiap data terhadap seluruh klaster yang ada, sehingga setiap data bisa saja memiliki kemungkinan untuk masuk ke lebih dari satu klaster dengan probabilitas yang berbeda beda.

```python
def _e_step(self, X: np.ndarray) -> np.ndarray:
    N = X.shape[0]
    K = self.n_components
    weighted_density = np.zeros((N, K))
    for k in range(K):
        density = self._multivariate_gaussian(X, self.means_[k],
                                              self.covariances_[k])
        weighted_density[:, k] = self.weights_[k] * density
    row_sums = weighted_density.sum(axis=1, keepdims=True)
    row_sums = np.where(row_sums == 0, 1e-300, row_sums)
    responsibilities = weighted_density / row_sums
    return responsibilities
```

Gambar 4.13 Code Expectation Step

Berdasarkan Gambar 4.13, fungsi _e_step() digunakan untuk menjalankan tahap *E-Step* dalam algoritma GMM. Fungsi ini bekerja dengan menghitung probabilitas setiap data point untuk masuk ke dalam masing-masing klaster, yang hasilnya disebut sebagai *responsibilities*.

Pertama, dihitung nilai densitas Gaussian untuk setiap klaster menggunakan fungsi _multivariate_gaussian() berdasarkan nilai *means* dan *covariances* yang sudah ada. Nilai densitas tersebut kemudian dikalikan dengan bobot masing-masing klaster (*weights*) untuk menghasilkan *weighted density*. Setelah itu, dilakukan normalisasi agar total probabilitas setiap data point terhadap semua klaster bernilai 1. Selain itu, terdapat penanganan khusus untuk menghindari pembagian dengan nilai nol, yaitu dengan mengganti nilai 0 menjadi 1e-300 agar proses perhitungan tidak menghasilkan error.

4.3.4 M- Step

Setelah *responsibilities* selesai dihitung pada tahap *E-Step*, proses kemudian dilanjutkan ke tahap *M-Step* atau *Maximization Step*. Di tahap ini, ketiga parameter GMM yaitu *means*, *covariances*, dan *weights* akan diperbarui berdasarkan nilai *responsibilities* yang sudah didapat sebelumnya. *M-Step* dilakukan dengan tujuan

agar nilai parameter yang dihasilkan semakin mendekati kondisi optimal, sehingga model GMM dapat merepresentasikan data dengan lebih baik pada setiap iterasi yang berjalan.

```
def _m_step(self, X: np.ndarray,
            responsibilities: np.ndarray) -> None:
    N, D = X.shape
    K    = self.n_components
    for k in range(K):
        gamma_k = responsibilities[:, k]
        N_k     = gamma_k.sum()
        # Update mean
        self.means_[k] = (gamma_k[:, np.newaxis] * X).sum(axis=0) / N_k
        # Update kovarians
        diff   = X - self.means_[k]
        cov_k  = (gamma_k[:, np.newaxis, np.newaxis]
                  * diff[:, :, np.newaxis]
                  * diff[:, np.newaxis, :]).sum(axis=0) / N_k
        cov_k += np.eye(D) * self.reg_covar
        self.covariances_[k] = cov_k
        # Update mixing coefficient
        self.weights_[k] = N_k / N
```

Gambar 4. 14 Code Maximization Step

Berdasarkan Gambar 4.14, fungsi _m_step() digunakan untuk memperbarui ketiga parameter GMM yaitu *means*, *covariances*, dan *weights* berdasarkan nilai *responsibilities* yang sudah diperoleh dari tahap *E-Step* sebelumnya.

Parameter pertama yang diperbarui adalah *means*, dihitung dengan mengalikan setiap data point dengan nilai *responsibilities*-nya lalu dibagi dengan total *responsibilities* setiap klaster (*N_k*). Data point yang memiliki probabilitas lebih tinggi terhadap suatu klaster akan lebih berpengaruh terhadap nilai *means* klaster tersebut.

Parameter kedua adalah *covariances*, dihitung berdasarkan selisih antara data dengan *means* yang sudah diperbarui. Terdapat juga penambahan nilai

regularisasi kecil pada diagonal matriks kovarians yang bertujuan untuk mencegah matriks menjadi singular selama proses iterasi berlangsung.

Parameter ketiga adalah *weights*, dihitung dengan membagi total *responsibilities* setiap klaster (*N_k*) dengan jumlah seluruh data (*N*). Klaster yang memiliki lebih banyak anggota akan mendapatkan bobot yang lebih besar dibandingkan klaster lainnya.

4.3.5 Hitung Nilai BIC

Setelah proses iterasi *E-Step* dan *M-Step* selesai dijalankan dan model GMM berhasil konvergen, langkah selanjutnya adalah menghitung nilai *BIC* (*Bayesian Information Criterion*) dan *AIC* (*Akaike Information Criterion*) untuk setiap nilai K yang sudah diuji. Kedua nilai ini digunakan untuk mengevaluasi seberapa baik model GMM yang terbentuk dalam menjelaskan data, sekaligus mempertimbangkan kompleksitas model agar tidak terjadi *overfitting*. Semakin kecil nilai *BIC* dan *AIC* yang dihasilkan, maka semakin baik dan optimal model tersebut.

```python
def bic(self, X: np.ndarray) -> float:
    N, D = X.shape
    K    = self.n_components
    # Hitung log-likelihood akhir
    log_lik = self._compute_log_likelihood(X)
    # Hitung jumlah parameter bebas
    params_mean  = K * D
    params_cov   = K * D * (D + 1) // 2
    params_mix   = K - 1
    n_params     = params_mean + params_cov + params_mix
    # Formula BIC
    bic_value = -2 * log_lik + n_params * np.log(N)
    return bic_value
def aic(self, X: np.ndarray) -> float:
    N, D = X.shape
    K    = self.n_components
    log_lik  = self._compute_log_likelihood(X)
    n_params = K * D + K * D * (D + 1) // 2 + (K - 1)
    return -2 * log_lik + 2 * n_params
```

Gambar 4. 15 Code BIC

Berdasarkan Gambar 4.15, terdapat dua fungsi yang digunakan untuk mengevaluasi kualitas model GMM, yaitu:

1. Fungsi bic() digunakan untuk menghitung nilai *BIC*. Perhitungan dimulai dengan menghitung nilai *log-likelihood* akhir dari model, kemudian menghitung jumlah total parameter bebas yang t erdiri dari parameter *mean*, parameter kovarians, dan parameter bobot campuran. Setelah semua komponen tersebut diperoleh, nilai *BIC* dihitung menggunakan formula -2 * log_likelihood + n_params * log(N). Semakin kecil nilai *BIC* yang dihasilkan, semakin baik model tersebut.
2. Fungsi aic() bekerja dengan cara yang hampir sama, perbedaannya hanya terletak pada formula yang digunakan yaitu -2 * log_likelihood + 2 * n_params.

4.3.6 Menampilkan Parameter Model Terbaik

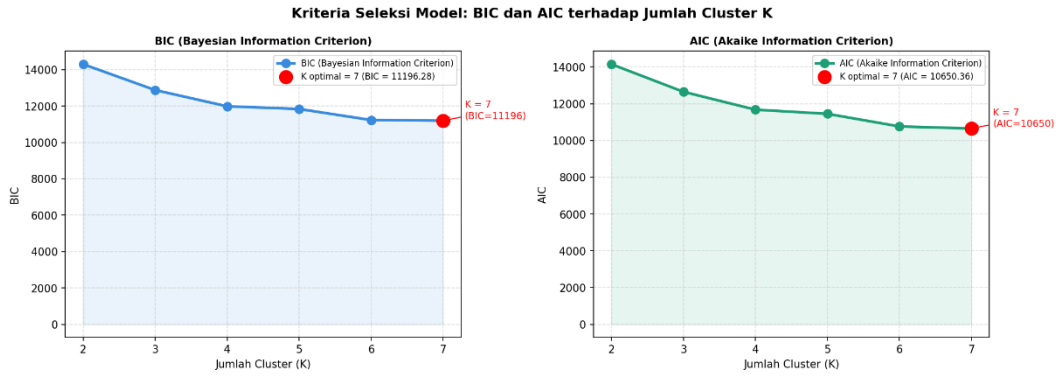
```
def print_model_parameters(gmm: GaussianMixtureModel,
                           k_optimal: int) -> None:
    print("\n" + "=" * 70)
    print(f"  TAHAP 3 — PARAMETER MODEL GMM OPTIMAL (K = {k_optimal})")
    print("=" * 70)
    print(f"\n  Model memiliki {k_optimal} komponen Gaussian:\n")
    for k in range(k_optimal):
        print(f"  Komponen / Cluster {k+1}")
        print(f"  Mixing Coefficient (π_{k+1}) = {gmm.weights_[k]:.6f} "
              f"({gmm.weights_[k]*100:.2f}%)")
        print(f"  Mean (μ_{k+1}):")
        for feat, mean_val in zip(FEATURES, gmm.means_[k]):
            print(f"    {feat:<25}: {mean_val:>10.6f}")
        print(f"  Diagonal Kovarians (Σ_{k+1} — variance per fitur):")
        for feat, var_val in zip(FEATURES, np.diag(gmm.covariances_[k])):
            print(f"    {feat:<25}: {var_val:>10.6f}")
        print()
```

Gambar 4. 16 Code Model Parameter

Berdasarkan Gambar 4.16 tersebut, cara kerja fungsi ini dimulai dengan mencetak header berupa pemisah karakter = sebanyak 70 yang disertai nilai k_optimal.

Setelah itu, fungsi melakukan iterasi sebanyak k_optimal kali, di mana setiap iterasi mewakili satu komponen Gaussian. Untuk setiap komponen, fungsi mencetak tiga informasi utama, yaitu Mixing Coefficient (π) yang menunjukkan bobot atau proporsi tiap cluster dalam format desimal enam digit dan persentase dua desimal, Mean (μ) yang menampilkan nilai rata-rata setiap fitur berdasarkan gmm.means_[k] yang di-zip dengan nama fitur dari variabel FEATURES, serta Diagonal Kovarians (Σ) yang menampilkan variansi per fitur menggunakan np.diag(gmm.covariances_[k]) untuk mengambil elemen diagonal dari matriks kovarians.

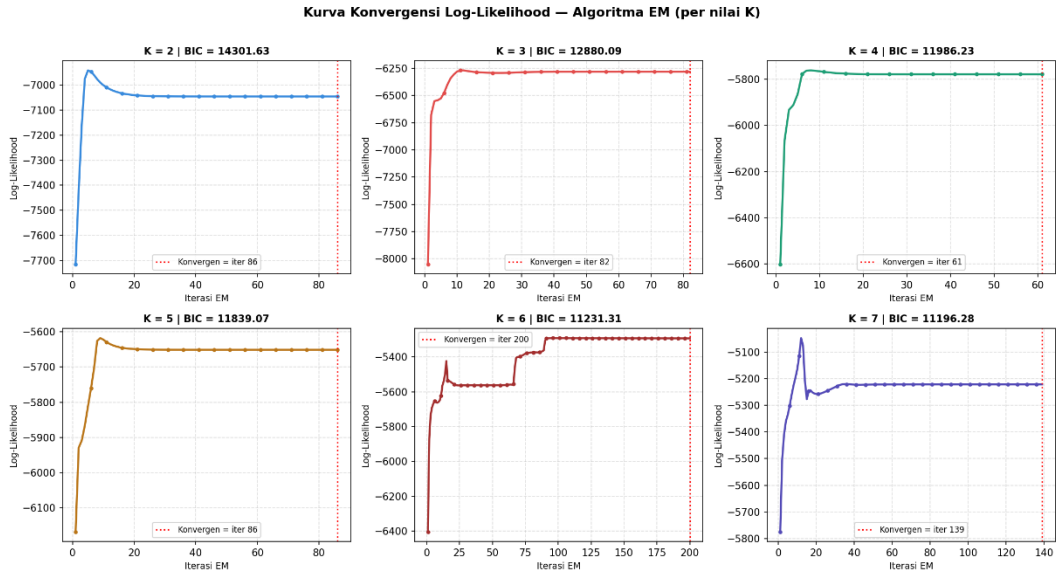
4.3.7 Hasil Seleksi Model Menggunakan BIC dan AIC



Gambar 4. 17 Visualisasi BIC dan AIC

Berdasarkan Gambar 4.16, grafik menampilkan kriteria seleksi model BIC dan AIC terhadap jumlah cluster K. Kedua grafik menunjukkan tren penurunan nilai seiring bertambahnya jumlah cluster dari K=2 hingga K=7. Titik optimal yang ditandai dengan lingkaran merah berada pada K=7, dengan nilai BIC sebesar 11.196,28 dan nilai AIC sebesar 10.650,36. Hal ini menunjukkan bahwa jumlah cluster terbaik untuk model GMM adalah 7 cluster, karena pada titik tersebut kedua kriteria mencapai nilai terendahnya.

4.3.8 LogLikehood Konvergensi



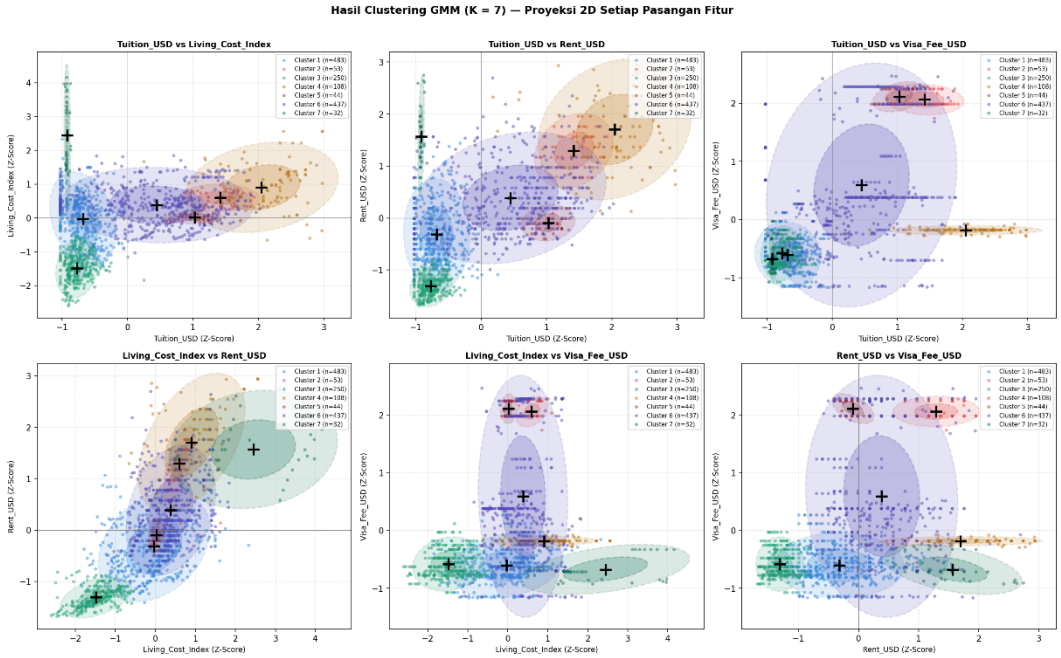
Gambar 4. 18 Kurva Konvergensi Loglikehood

Berdasarkan gambar 4.18, gambar diatas menampilkan kurva konvergensi Log-Likelihood dari algoritma EM untuk setiap nilai K (jumlah cluster) pada model GMM, mulai dari K=2 hingga K=7. Secara umum, seluruh grafik menunjukkan pola yang sama, yaitu nilai Log-Likelihood meningkat secara signifikan pada iterasi awal kemudian stabil dan konvergen setelah sejumlah iterasi tertentu.

Pada K=2, model konvergen pada iterasi ke-86 dengan nilai BIC 14.301,63. Pada K=3, konvergen lebih cepat di iterasi ke-82 dengan BIC 12.880,09, sedangkan K=4 konvergen paling cepat di iterasi ke-61 dengan BIC 11.986,23. Untuk K=5, model konvergen pada iterasi ke-86 dengan BIC 11.839,07. Sementara itu, K=6 memerlukan iterasi paling banyak yaitu hingga iterasi ke-200 dengan BIC 11.231,31, yang mengindikasikan proses konvergensi yang lebih lambat dan tidak stabil. Terakhir, pada K=7, model konvergen pada iterasi ke-139 dengan nilai BIC terendah sebesar 11.196,28.

Dari keseluruhan grafik tersebut, dapat disimpulkan bahwa K=7 merupakan jumlah cluster optimal karena menghasilkan nilai BIC paling rendah, meskipun membutuhkan lebih banyak iterasi dibandingkan beberapa nilai K lainnya.

4.3.9 Hasil Clustering GMM



Gambar 4. 19 Hasil Clustering 2d Gaussian Mixture Models

Berdasarkan Gambar 4.19, hasil clustering GMM dengan K=7 ditampilkan dalam bentuk proyeksi 2D untuk setiap pasangan fitur. Terdapat enam grafik scatter plot yang masing-masing memvisualisasikan hubungan antar dua fitur, yaitu Tuition_USD vs Living_Cost_Index, Tuition_USD vs Rent_USD, Tuition_USD vs Visa_Fee_USD, Living_Cost_Index vs Rent_USD, Living_Cost_Index vs Visa_Fee_USD, dan Rent_USD vs Visa_Fee_USD. Seluruh nilai fitur telah distandarisasi menggunakan Z-Score, dan setiap cluster digambarkan dengan warna berbeda beserta elips kovarians yang merepresentasikan sebaran datanya, sementara tanda plus (+) menunjukkan titik pusat atau centroid masing-masing cluster.

Dari sisi distribusi anggota, Cluster 1 memiliki jumlah anggota terbanyak dengan 483 data, diikuti Cluster 6 dengan 437 data, kemudian Cluster 3 dengan 250 data dan Cluster 4 dengan 108 data. Sementara itu, Cluster 2 memiliki 53 data, Cluster 5 sebanyak 44 data, dan Cluster 7 merupakan cluster terkecil dengan hanya 32 anggota.

Secara visual, sebagian besar cluster terkonsentrasi pada rentang Z-Score antara -1 hingga 2 untuk fitur Tuition_USD, yang mengindikasikan bahwa mayoritas universitas memiliki biaya kuliah yang relatif moderat. Cluster 1 dan Cluster 6 tampak mendominasi area tengah grafik dengan sebaran yang luas, mencerminkan keragaman karakteristik yang tinggi dalam kedua cluster tersebut. Cluster 2 dan Cluster 7 cenderung berada pada nilai yang sangat tinggi pada beberapa fitur, terutama pada Visa_Fee_USD, mengindikasikan bahwa cluster ini mewakili universitas dengan biaya visa yang jauh di atas rata-rata. Cluster 5 menunjukkan posisi yang cukup terisolasi pada grafik Living_Cost_Index vs Visa_Fee_USD, yang menunjukkan karakteristik unik dengan biaya hidup dan biaya visa yang berbeda signifikan dari cluster lainnya.

Adapun karakteristik masing-masing cluster secara ringkas dapat dilihat pada tabel berikut:

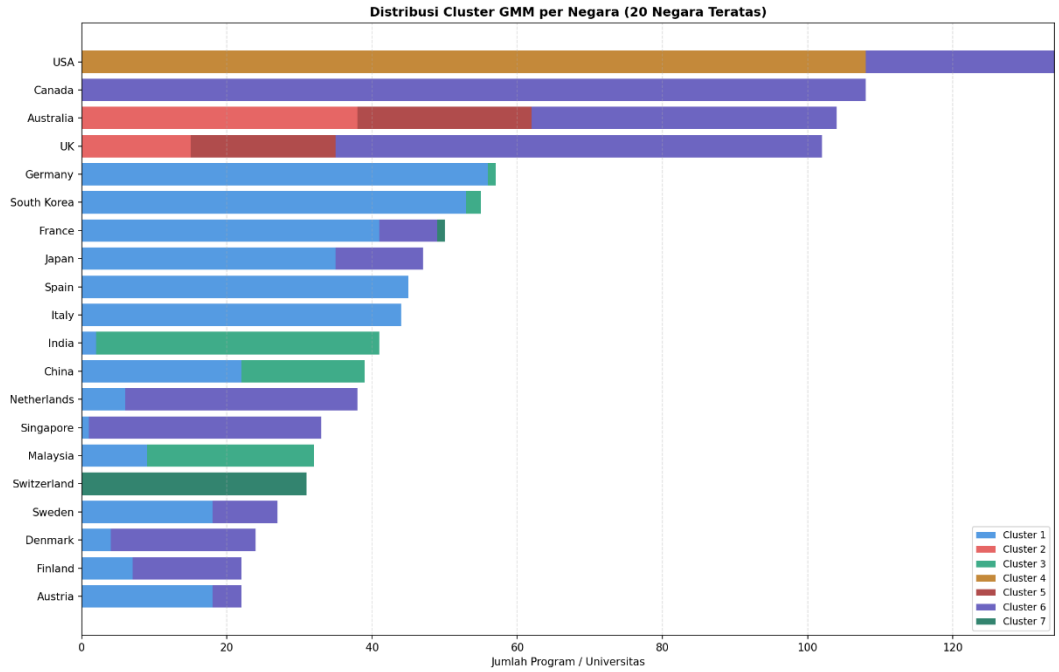
Tabel 4. 4 Keterangan Cluster

Cluster	Jumlah data	Biaya kuliah	Biaya hidup	Biaya sewa	Biaya visa	Keterangan
1	483	Rendah-sedang	Rendah-sedang	Rendah-sedang	Sedang	Univ dengan biaya terjangkau dan paling umum ditemukan
2	53	Sedang	Sedang	Sedang	Sangat tinggi	Univ dengan biaya visa jauh diatas rata rata
3	250	Menengah-tinggi	Moderat	Moderat	Sedang	Univ kelas menengah dengan kualitas cukup baik

Cluster	Jumlah data	Biaya kuliah	Biaya hidup	Biaya sewa	Biaya visa	Keterangan
4	108	Tinggi	Tinggi	Tinggi	Sedang	Univ premium di kota besar dengan standar biaya tinggi
5	44	Rendah-sedang	Tinggi	Tinggi	Sedang	Univ negeri/bersubsidi di kota dengan biaya hidup mahal
6	437	Tinggi-sangat tinggi	Tinggi	Tinggi	Sedang	Univ ternama di negara maju
7	32	Sangat tinggi	Sangat tinggi	Sangat tinggi	Tinggi	Universitas eksklusif dengan total biaya pendidikan paling mahal

Secara keseluruhan, Gambar 4.19 menunjukkan bahwa model GMM dengan K=7 berhasil memisahkan data ke dalam kelompok-kelompok yang memiliki karakteristik berbeda berdasarkan kombinasi fitur biaya kuliah, biaya hidup, biaya sewa, dan biaya visa, meskipun terdapat beberapa tumpang tindih antar cluster yang merupakan hal wajar dalam metode clustering berbasis probabilitas seperti GMM.

4. 4 Evaluasi Hasil Klaster



Gambar 4. 20 Visualisasi Distribusi Cluster Gmm

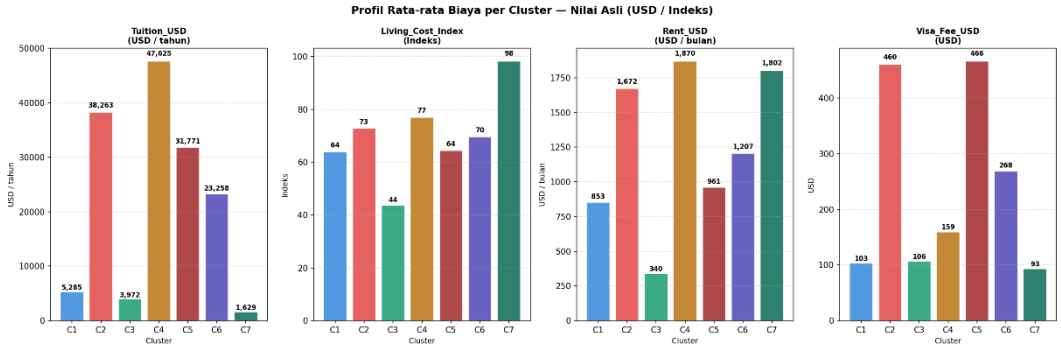
Berdasarkan Gambar 4.20 , grafik menampilkan distribusi cluster GMM per negara untuk 20 negara teratas berdasarkan jumlah program atau universitas yang terdaftar dalam dataset. Sumbu horizontal menunjukkan jumlah program atau universitas, sedangkan sumbu vertikal menampilkan nama negara yang diurutkan dari yang terbanyak ke yang paling sedikit.

USA dan Canada menempati posisi teratas dengan jumlah program terbanyak, masing-masing sekitar 120 dan 110 program. Kedua negara ini didominasi oleh Cluster 4 (warna cokelat keemasan) dan Cluster 6 (warna ungu), yang mencerminkan bahwa universitas di kedua negara tersebut memiliki biaya kuliah dan biaya hidup yang tinggi hingga sangat tinggi. Australia dan UK berada di posisi ketiga dan keempat dengan sekitar 105 program, namun menunjukkan komposisi cluster yang lebih beragam, terutama dengan kehadiran Cluster 2 (warna merah) yang cukup dominan, mengindikasikan adanya universitas dengan biaya visa yang tinggi di kedua negara tersebut.

Germany dan South Korea memiliki sekitar 55-60 program dan didominasi oleh Cluster 1 (warna biru muda), yang mencerminkan universitas dengan biaya terjangkau dan moderat. Hal serupa juga terlihat pada France, Japan, Spain, dan Italy yang masing-masing memiliki sekitar 45-50 program dan sebagian besar masuk ke dalam Cluster 1. India dan China menunjukkan keragaman cluster yang cukup tinggi dengan kehadiran Cluster 3 (warna hijau tua) dan Cluster 7 (warna hijau gelap) yang lebih menonjol dibandingkan negara lain, mengindikasikan adanya variasi biaya yang cukup besar antar universitas di kedua negara tersebut.
Netherlands, Singapore, Malaysia, dan Switzerland masing-masing memiliki sekitar 30-38 program dengan komposisi cluster yang bervariasi. Singapore dan

Switzerland cenderung memiliki proporsi cluster dengan biaya lebih tinggi dibandingkan Malaysia yang lebih didominasi Cluster 3. Sementara itu, Sweden, Denmark, Finland, dan Austria berada di posisi terbawah dengan sekitar 20-27 program, dan sebagian besar universitas di negara-negara Eropa Utara ini masuk ke dalam Cluster 1 yang mencerminkan biaya pendidikan yang relatif terjangkau.

Secara keseluruhan, grafik ini menunjukkan bahwa negara-negara berbahasa Inggris seperti USA, Canada, Australia, dan UK memiliki keragaman cluster yang lebih tinggi dengan proporsi cluster berbiaya tinggi yang lebih besar, sementara negara-negara di Eropa dan Asia cenderung lebih homogen dengan dominasi cluster berbiaya moderat hingga terjangkau.



Gambar 4. 21 Visualisasi Profil Rata Rata Biaya Per Cluster

Berdasarkan gambar 4.21, grafik menampilkan profil rata-rata biaya per cluster dalam nilai asli (USD/Indeks) yang terbagi menjadi empat grafik, yaitu Tuition_USD, Living_Cost_Index, Rent_USD, dan Visa_Fee_USD.

Pada grafik pertama yaitu Tuition_USD, Cluster 2 memiliki biaya kuliah tertinggi sebesar $47.625 per tahun, diikuti Cluster 1 sebesar $38.263, dan Cluster 5 sebesar $31.771. Cluster 6 berada di angka $23.258, sementara Cluster 3 dan Cluster 4 memiliki biaya kuliah yang sangat rendah masing-masing sebesar $3.972 dan $5.285. Cluster 7 menjadi yang terendah dengan hanya $1.629 per tahun, yang

kemungkinan besar mewakili universitas dengan biaya subsidi atau negara berkembang.

Pada grafik kedua yaitu Living_Cost_Index, Cluster 7 menunjukkan indeks biaya hidup tertinggi sebesar 98, jauh melampaui cluster lainnya. Cluster 5 dan Cluster 3 memiliki indeks yang cukup tinggi masing-masing sebesar 77 dan 73, sedangkan Cluster 1, Cluster 4, dan Cluster 6 berada di kisaran 64 hingga 70. Cluster 2 memiliki indeks biaya hidup terendah sebesar 44, yang mengindikasikan bahwa meskipun biaya kuliahnya sangat tinggi, biaya hidup di lokasi universitas Cluster 2 relatif lebih terjangkau.

Pada grafik ketiga yaitu Rent_USD, Cluster 2 kembali menunjukkan nilai tertinggi sebesar $1.870 per bulan, diikuti Cluster 5 sebesar $1.802 dan Cluster 6 sebesar $1.672. Cluster 1 berada di angka $1.207, sementara Cluster 4 sebesar $961. Cluster 3 memiliki biaya sewa terendah sebesar $340 per bulan, yang konsisten dengan indeks biaya hidupnya yang tinggi namun biaya kuliah yang rendah, mencerminkan karakteristik kota mahal dengan universitas bersubsidi.

Pada grafik keempat yaitu Visa_Fee_USD, Cluster 2 dan Cluster 5 memiliki biaya visa tertinggi masing-masing sebesar $460 dan $466, yang jauh di atas cluster lainnya. Cluster 6 berada di angka $268, sementara Cluster 4 sebesar $159 dan Cluster 1 sebesar $103. Cluster 3 dan Cluster 7 memiliki biaya visa terendah masing-masing sebesar $106 dan $93.

Secara keseluruhan, grafik ini menunjukkan bahwa setiap cluster memiliki profil biaya yang unik dan berbeda. Cluster 2 menonjol sebagai kelompok dengan biaya kuliah dan biaya visa tertinggi, Cluster 7 memiliki biaya hidup tertinggi

namun biaya kuliah terendah, sementara Cluster 3 mewakili kelompok universitas dengan hampir seluruh komponen biaya yang rendah dan terjangkau.

BAB V

KESIMPULAN DAN SARAN

5.1 KESIMPULAN

Berdasarkan hasil penelitian yang telah dilakukan mengenai penerapan machine learning dalam klasterisasi tuition fee level dan living cost mahasiswa internasional menggunakan metode Gaussian Mixture Models (GMM), maka dapat ditarik kesimpulan sebagai berikut:

1. Metode Gaussian Mixture Models (GMM) berhasil diterapkan secara efektif dalam melakukan klasterisasi *tuition fee level* dan *living cost* mahasiswa internasional menggunakan bahasa pemrograman Python melalui Visual Studio Code. Proses implementasi mencakup tahapan yang sistematis mulai dari pengumpulan data, *preprocessing* dengan normalisasi Z-Score, penentuan jumlah klaster optimal menggunakan kriteria BIC dan AIC, hingga estimasi parameter model melalui algoritma *Expectation-Maximization* (EM). Keseluruhan tahapan tersebut menghasilkan model klasterisasi yang konvergen dan dapat merepresentasikan pola biaya mahasiswa internasional secara objektif dan terukur.
2. Model GMM dengan jumlah klaster optimal K=7 (nilai BIC = 11.196,28 dan AIC = 10.650,36) berhasil mengidentifikasi tujuh kelompok universitas dengan profil biaya yang berbeda dan bermakna. Ketujuh klaster tersebut merepresentasikan spektrum biaya yang beragam: Klaster 1 (483 data) mewakili universitas berbiaya terjangkau dan paling umum ditemukan; Klaster 2 (53 data) mewakili universitas dengan biaya visa yang jauh di atas rata-rata; Klaster 3 (250 data) mewakili universitas kelas menengah; Klaster

4 (108 data) mewakili universitas premium di kota besar; Klaster 5 (44 data) mewakili universitas negeri atau bersubsidi di kota berbiaya hidup tinggi; Klaster 6 (437 data) mewakili universitas ternama di negara maju; dan Klaster 7 (32 data) mewakili universitas eksklusif dengan total biaya paling mahal. Temuan ini menunjukkan bahwa GMM mampu mengungkap pola tersembunyi dalam data biaya yang bersifat multivariat dan kompleks.

3. Evaluasi model menggunakan *Bayesian Information Criterion* (BIC) dan *Akaike Information Criterion* (AIC) menunjukkan bahwa K=7 merupakan jumlah klaster yang paling optimal dengan nilai BIC terendah sebesar 11.196,28. Kurva konvergensi *log-likelihood* pada algoritma EM memperlihatkan bahwa seluruh model untuk setiap nilai K berhasil mencapai konvergensi, dengan model K=7 konvergen pada iterasi ke-139. Keberhasilan GMM dalam menghasilkan klaster yang dapat diinterpretasikan dengan baik mengkonfirmasi bahwa metode ini lebih unggul dibandingkan metode *hard clustering* seperti K-Means dalam menangani data biaya mahasiswa internasional yang memiliki distribusi kompleks, bersifat *right-skewed*, dan tidak simetris.

5.2 SARAN

Berdasarkan hasil penelitian dan kesimpulan yang telah diuraikan, terdapat beberapa saran yang dapat dijadikan bahan pertimbangan untuk pengembangan penelitian selanjutnya:

1. Perluasan dan pembaruan dataset sangat disarankan untuk dilakukan. Dataset yang digunakan dalam penelitian ini bersumber dari satu platform (Kaggle) dengan jumlah 1.407 entri. Penelitian selanjutnya disarankan

menggunakan dataset yang lebih besar, lebih mutakhir, dan mencakup lebih banyak negara serta institusi pendidikan agar hasil klasterisasi yang diperoleh bersifat lebih komprehensif dan representatif terhadap kondisi nyata biaya pendidikan internasional secara global.

2. Penambahan variabel atau fitur analisis perlu dipertimbangkan untuk meningkatkan ketajaman profil klaster yang dihasilkan. Variabel tambahan seperti reputasi akademik universitas (peringkat dunia), tingkat penerimaan mahasiswa, angka kelulusan, mata uang lokal, serta indikator kualitas hidup seperti indeks keamanan dan fasilitas publik dapat diintegrasikan ke dalam model sehingga klasterisasi yang terbentuk tidak hanya mencerminkan aspek biaya semata, tetapi juga faktor kelayakan dan kualitas pendidikan secara holistik.
3. Komparasi dengan metode klasterisasi lain sangat dianjurkan untuk memvalidasi keunggulan GMM secara empiris. Penelitian mendatang dapat membandingkan kinerja GMM dengan algoritma klasterisasi lainnya seperti K-Means, DBSCAN, *Hierarchical Clustering*, atau *Spectral Clustering* menggunakan metrik evaluasi yang lebih beragam seperti *Silhouette Score*, *Davies-Bouldin Index*, dan *Calinski-Harabasz Index*, guna memperoleh gambaran yang lebih komprehensif mengenai keunggulan relatif masing-masing metode pada jenis data biaya pendidikan internasional.

DAFTAR PUSTAKA

Abijono, H., Santoso, P., & Anggreini, N. L. (2021). SUPERVISED LEARNING AND UNSUPERVISED LEARNING ALGORITHM IN DATA PROCESSING. *G-Tech: Jurnal Teknologi Terapan*, *4*(2), 315–318. https://doi.org/10.33379/gtech.v4i2.635

Ahmad, A., Hasan, M., & Ghorbanpour, M. (2024). Education cost as a new fickle in higher education for students learning via quantitively multinomial logistic regression. *Scientific Reports 2024 14:1*, *14*(1), 29947-. https://doi.org/10.1038/s41598-024-81074-x

Darmawan Sidik, A., Ansawarman, A., Kunci, K., Kendaraan Bermotor, J., Regresi, M., & Jalan, F. (2022). Prediksi Jumlah Kendaraan Bermotor Menggunakan Machine Learning. *Formosa Journal of Multidisciplinary Research (FJMR)*, *1*(3), 559–568. https://doi.org/10.55927

Gustientiedina, G., Adiya, M. H., & Desnelita, Y. (2019). Penerapan Algoritma K-Means Untuk Clustering Data Obat-Obatan Pada RSUD Pekanbaru. *Jurnal Nasional Teknologi Dan Sistem Informasi*, *5*(1), 17–24. https://doi.org/10.25077/TEKNOSI.V5I1.2019.17-24

Kurniawan, H., Defit, S., & Sumijan. (2020). Data Mining Menggunakan Metode K-Means Clustering Untuk Menentukan Besaran Uang Kuliah Tunggal. *Journal of Applied Computer Science and Technology*, *1*(2), 80–89. https://doi.org/10.52158/JACOST.V1I2.102

Matematika, J., Matematika, P., Riyono, J., Puspa, S. D., Pujiastuti, C. E., Trisakti, U., Kyai, J., No, T., & Grogol, J. (2022). Simulasi Clustering Provinsi di Indonesia dalam Penyebaran Covid-19 Berdasarkan Indikator Kesehatan

Masyarakat Menggunakan Algoritma Gaussian Mixture Model. *MAJAMATH: Jurnal Matematika Dan Pendidikan Matematika*, *5*(1), 43–60. https://doi.org/10.36815/MAJAMATH.V5I1.1699

Milson, A., Herwindiati, D. E., & Perdana, N. J. (2024). PENERAPAN KLASIFIKASI SUARA SEBAGAI AUTENTIKASI KEAMANAN SISTEM LOGIN MENGGUNAKAN GAUSSIAN MIXTURE MODELS. In *Computatio: Journal of Computer Science and Information Systems* (Vol. 8, Number 1).

Mohamed Nafuri, A. F., Sani, N. S., Zainudin, N. F. A., Rahman, A. H. A., & Aliff, M. (2022). Clustering Analysis for Classifying Student Academic Performance in Higher Education. *Applied Sciences 2022, Vol. 12, Page 9467*, *12*(19), 9467. https://doi.org/10.3390/APP12199467

Nurhalizah, R. S., Ardianto, R., & Purwono, P. (2024a). Analisis Supervised dan Unsupervised Learning pada Machine Learning: Systematic Literature Review. *Jurnal Ilmu Komputer Dan Informatika*, *4*(1), 61–72. https://doi.org/10.54082/jiki.168

Nurhalizah, R. S., Ardianto, R., & Purwono, P. (2024b). Analisis Supervised dan Unsupervised Learning pada Machine Learning: Systematic Literature Review. *Jurnal Ilmu Komputer Dan Informatika*, *4*(1), 61–72. https://doi.org/10.54082/jiki.168

OECD. (2025). Education at a Glance 2025: OECD Indicators. *Education at a Glance, Education at a Glance*, *2025*. https://doi.org/10.1787/1C0D9C79-EN

Pangastuti, S. S., Fithriasari, K., Iriawan, N., & Suryaningtyas, W. (2021). Data Mining Approach for Educational Decision Support. *EKSAKTA: Journal of*

Sciences and Data Analysis, *2*(1), 33–44. https://doi.org/10.20885/EKSAKTA.vol2.iss1.art5

Studi Rekayasa Keamanan Siber, P., Tinggi Sandi Negara Jl Usa, S. H., & Nutug, P. (n.d.). *Penerapan Metode Discrete Wavelet Transform (DWT) dan Gaussian Mixture Model (GMM) Sebagai Pengenal Penutur Jeckson Sidabutar*.

Ummami, R., & Winarno, B. (2023). Gaussian Mixture Model dengan Algoritme Expectation Maximization untuk Pengelompokan Data Distribusi Air Bersih di Jawa Barat. *PRISMA, Prosiding Seminar Nasional Matematika*, *6*, 745–750. https://journal.unnes.ac.id/sju/index.php/prisma/

Widiarina, W., Mariskhana, K., & Sintawati, I. D. (2024). Application of Data Mining for Clustering Human Development Index Based on West Java Province 2017-2022. *Sinkron : Jurnal Dan Penelitian Teknik Informatika*, *8*(1), 44–53. https://doi.org/10.33395/SINKRON.V9I1.13148

World Education Statistics UNESCO UNESCO Institute for Statistics. (1946). http://www.uis.unesco.org