BAB II
LANDASAN TEORI

2.1 Tinjauan Pustaka

Penelitian ini menggunakan pendekatan analisis klaster untuk mengidentifikasi pola karakter santri dalam konteks lingkungan pesantren dan gaya belajar yang beragam. Pemilihan metode ini dilatarbelakangi oleh keberhasilan berbagai studi sebelumnya yang membuktikan efektivitas teknik klasterisasi, seperti *K-Means Clustering* dan *Hierarchical Clustering*, dalam mengelompokkan individu berdasarkan gaya belajar dan karakteristik perilaku. Dengan mempertimbangkan kompleksitas data karakter santri, metode klasterisasi dipandang mampu memberikan gambaran segmentasi yang lebih komprehensif dan bermakna. Oleh karena itu, kajian pustaka dalam bagian ini menyajikan berbagai penelitian yang telah menerapkan metode klaster sebagai pijakan teoritis sekaligus pembanding metodologis untuk mendukung relevansi pendekatan yang digunakan dalam penelitian ini.

Tabel 2.1 Penelitian terdahulu terkait *clustering*

No	Peneliti (Tahun)	Judul	Metode	Tujuan/Goals	Hasil
1	(Deak dkk., 2025)	*Clustering* Undergraduate Students Based on Physical Fitness Levels	K-Means *Clustering*	Mengelompokkan mahasiswa berdasarkan kebugaran fisik	Perbedaan signifikan ditemukan dengan 2 klaster utama pada laki-laki dan 1 pada perempuan
2	(Ma dkk., 2024)	Application of Cluster Analysis to Identify Different Reader Groups	Gaussian Mixture Models, K-Means, K-Medoids, *Hierarchical Clustering*	Mengidentifikasi kelompok pembaca berdasarkan keterlibatan dalam aplikasi literasi digital	K-Means memberikan hasil terbaik, mengidentifikasi 9 kelompok pembaca dengan pola keterlibatan berbeda dalam

SEKOLAH PASCASARJANA

No	Peneliti (Tahun)	Judul	Metode	Tujuan/Goals	Hasil
					pembelajaran literasi
3	(Meng dkk., 2024)	Nano-Integrating Green and Low-Carbon Concepts into Ideological and Political Education	K-Means *Clustering*	Mengidentifikasi pola penerapan konsep hijau dalam pendidikan politik dan ideologis	Cluster 3 memiliki integrasi praktik hijau paling kuat
4	(Siswipraptini dkk., 2023)	IT Job Profile Using *Hierarchical Clustering*	Average-Linkage *Hierarchical Clustering*	Mengklasifikasikan profil pekerjaan IT berdasarkan kompetensi	10 klaster pekerjaan IT dengan keterampilan berbeda ditemukan
5	(Richelle dkk., 2024)	Cluster Analysis of Medical Students' Attitudes regarding People who Use Drugs	*Hierarchical Clustering* & K-Means	Mengelompokkan mahasiswa kedokteran berdasarkan sikap terhadap pengguna narkoba	4 klaster mahasiswa ditemukan
6	(Bueno dkk., 2023)	Hierarchical and K-means *Clustering* to Assess Thermal Dissatisfaction and Productivity in University Classrooms	Hierarchical & K-Means *Clustering*	Mengelompokkan mahasiswa berdasarkan tingkat ketidaknyamanan termal dan produktivitas	4 klaster terbentuk; klaster dengan produktivitas tertinggi memiliki tingkat ketidaknyamanan termal terendah. K-Means validasi klaster dengan tingkat kesuksesan klasifikasi 95,4%.

SEKOLAH PASCASARJANA

No	Peneliti (Tahun)	Judul	Metode	Tujuan/Goals	Hasil
7	(Liu dkk., 2022)	Use Latent Profile Analysis and K-Means *Clustering* to Identify Student Anxiety Profiles	Latent Profile Analysis (LPA), K-Means *Clustering*	Mengelompokkan siswa berdasarkan tingkat kecemasan mereka	3 profil kecemasan siswa ditemukan
8	(Araka dkk., 2022)	Using Educational Data Mining Techniques to Identify Profiles in Self Regulated Learning	Expectation - Maximization, Agglomerative Hierarchical, K-Means *Clustering*	Mengidentifikasi profil pembelajaran mandiri	Agglomerative *Hierarchical Clustering* menjadi metode terbaik untuk deteksi SRL
9	(Ramírez-Correa dkk., 2021)	Identifying Engineering Undergraduates' Learning Style Profiles	K-Means *Clustering*, Decision Tree C4.5	Mengembangkan model prediksi untuk identifikasi gaya belajar mahasiswa teknik	Model prediksi memiliki akurasi 82,93%
10	(Mohd Talib dkk., 2023)	Identification of Student Behavioral Patterns in Higher Education Using K-Means Clustering and Support Vector Machine	K-Means *Clustering*, Support Vector Machine	Mengidentifikasi fitur pembentuk klaster berkarakteristik holistik dan mengembangkan model prediksi performa mahasiswa untuk tiap klaster	Ditemukan 3 klaster mahasiswa (performa rendah, sedang, tinggi); model prediksi berbasis label klaster menghasilkan akurasi lebih tinggi dibandingkan label nilai semester

SEKOLAH PASCASARJANA

No	Peneliti (Tahun)	Judul	Metode	Tujuan/Goals	Hasil
11	(Batool dkk., 2020)	*Clustering with the Average Silhouette Width*	Average Silhouette Width dan metode clustering	Mengkaji penggunaan Average Silhouette Width dalam evaluasi dan pemilihan jumlah klaster.	Average Silhouette Width dapat digunakan sebagai ukuran evaluasi clustering, tetapi memiliki keterbatasan sehingga tidak selalu tepat dijadikan satu-satunya dasar dalam menentukan jumlah klaster.
12	(IFAMR, 2026)	*From insight to impact: segmenting consumers by their regenerative agriculture priorities*	Silhouette, Calinski-Harabasz, Elbow, Gap Statistic, dan interpretasi hasil	Membandingkan beberapa pendekatan untuk menentukan jumlah klaster dalam proses segmentasi.	K=2 memperoleh Silhouette sebesar 0,203, sedangkan K=3 sebesar 0,182. Meskipun demikian, K=3 dipilih berdasarkan kombinasi beberapa bukti evaluasi serta pertimbangan *theoretical interpretability* dan *managerial actionability*.

SEKOLAH PASCASARJANA

No	Peneliti (Tahun)	Judul	Metode	Tujuan/Goals	Hasil
13	(Han dkk., 2011)	*Data Mining: Concepts and Techniques*	Berbagai metode clustering dan validasi	Menjelaskan konsep, metode, dan evaluasi dalam proses data mining, termasuk clustering.	Evaluasi clustering perlu disesuaikan dengan karakteristik data dan tujuan analisis, sehingga hasil pengelompokan tidak hanya dinilai berdasarkan satu ukuran evaluasi.

Dalam Tabel 2.1 terdapat 10 jurnal yang membahas mengenai teknik *clustering.* Penelitian oleh Ma dkk. (2024) menerapkan analisis klaster menggunakan Gaussian Mixture Models, K-Means, K-Medoids, dan *Hierarchical Clustering* untuk mengidentifikasi kelompok pembaca dalam aplikasi literasi digital. Hasil analisis menunjukkan bahwa K-Means *Clustering* memberikan hasil terbaik dengan 9 klaster pembaca, yang masing-masing menunjukkan pola keterlibatan berbeda dalam pembelajaran literasi. Pembaca dengan kemampuan awal rendah lebih banyak berinteraksi dengan permainan berbasis fonologi dan decoding, sementara pembaca dengan kemampuan lebih tinggi lebih fokus pada pemahaman dan kosa kata. Studi ini menyoroti pentingnya pemetaan keterlibatan pembaca untuk mengembangkan strategi personalisasi dalam pembelajaran digital, guna meningkatkan efektivitas dan aksesibilitas literasi berbasis teknologi.

Penelitian oleh Meng dkk. (2024) menggunakan K-Means *Clustering* untuk mengidentifikasi pola penerapan konsep hijau dalam pendidikan politik dan ideologis di perguruan tinggi. Hasil analisis menemukan 3 klaster utama, dengan Cluster 3 memiliki integrasi konsep hijau paling kuat dibandingkan dengan dua

UNIVERSITAS DIPONEGORO SEMARANG SEKOLAH PASCASARJANA

klaster lainnya. Klaster ini menunjukkan bahwa mahasiswa yang masuk dalam kelompok tersebut lebih aktif dalam mendukung praktik keberlanjutan dan lebih memahami pentingnya konsep hijau dalam kebijakan politik. Sementara itu, klaster lainnya menunjukkan tingkat adopsi konsep hijau yang lebih rendah. Studi ini menyoroti bahwa pendidikan berbasis keberlanjutan dapat membantu meningkatkan kesadaran lingkungan mahasiswa serta mendorong implementasi kebijakan hijau di institusi pendidikan.

Penelitian oleh Siswipraptini dkk. (2023) menerapkan Average-Linkage *Hierarchical Clustering* untuk mengelompokkan profil pekerjaan IT berdasarkan kompetensi yang dibutuhkan di industri. Dengan menganalisis 1.065 iklan lowongan kerja IT dari Tech in Asia dan Jobstreet Indonesia, penelitian ini menemukan 10 klaster utama pekerjaan IT berdasarkan keterampilan yang sering muncul, seperti pengembangan perangkat lunak, keamanan siber, analisis data, jaringan komputer, dan kecerdasan buatan. Setiap klaster dianalisis menggunakan teknik text mining dan semantic analysis, yang memungkinkan pemetaan keterampilan spesifik untuk setiap peran pekerjaan. Hasil penelitian ini divalidasi oleh 10 profesional IT melalui diskusi kelompok terfokus (FGD), menunjukkan bahwa metode *clustering* dapat menjadi alat penting dalam menyesuaikan kurikulum pendidikan dengan kebutuhan industri.

Penelitian oleh Richelle dkk. (2024) menggunakan kombinasi *Hierarchical Clustering* dan K-Means *Clustering* untuk mengelompokkan mahasiswa kedokteran berdasarkan sikap mereka terhadap pengguna narkoba (Substance Use Disorder - SUD). Studi ini menganalisis 923 mahasiswa kedokteran tahun terakhir di Belgia, menggunakan kuesioner beSAAS dengan 23 item terkait sikap terhadap pengguna narkoba. Hasilnya mengidentifikasi 4 klaster utama, yaitu *The Inclusives* (27,9%) yang memiliki sikap paling positif, *The Centrists* (23,6%) yang memiliki pengalaman pribadi dengan penggunaan zat, *The Moralists* (27,6%) yang memiliki stereotip kuat terhadap pengguna narkoba, serta *The Specialist Care-Oriented* (20,8%) yang lebih mendukung perawatan spesialis. Studi ini menunjukkan bahwa faktor seperti pengalaman pribadi, latar belakang budaya, dan spesialisasi yang diminati berpengaruh terhadap sikap mahasiswa terhadap

SEKOLAH PASCASARJANA

pengguna narkoba, sehingga perlu adanya pendekatan edukasi yang lebih spesifik untuk mengurangi stigma dalam lingkungan medis.

Penelitian oleh Bueno dkk. (2023) menggunakan metode *Hierarchical Clustering* dan *K-Means Clustering* untuk mengevaluasi hubungan antara ketidaknyamanan termal dan produktivitas mahasiswa di ruang kelas perguruan tinggi. Studi ini dilakukan terhadap 519 mahasiswa dari berbagai program studi teknik di Brasil, dengan mengumpulkan data lingkungan termal (suhu udara, kelembapan, kecepatan angin) dan data persepsi pribadi menggunakan kuisioner. Dua metode pengukuran digunakan untuk menilai produktivitas, yaitu produktivitas persepsi diri (PROD1) dan yang berbasis model termal PMV (PROD2). Hasil analisis klaster mengelompokkan mahasiswa ke dalam empat klaster yang menunjukkan tingkat produktivitas dan ketidaknyamanan termal yang berbeda. Klaster dengan tingkat produktivitas tertinggi menunjukkan persepsi termal netral, sementara klaster dengan persepsi termal dingin cenderung memiliki produktivitas lebih rendah. Validasi hasil dengan analisis diskriminan menunjukkan tingkat keberhasilan klasifikasi sebesar 95,4%. Penelitian ini menunjukkan adanya hubungan yang signifikan antara kenyamanan termal dan produktivitas persepsi diri, serta menunjukkan bahwa waktu pelaksanaan kuliah juga memengaruhi persepsi termal dan hasil belajar mahasiswa. Dengan demikian, pemahaman terhadap kondisi termal dalam ruang kelas sangat penting untuk meningkatkan kenyamanan dan performa akademik mahasiswa.

Penelitian oleh Deak dkk. (2025) menerapkan K-Means *Clustering* untuk mengelompokkan mahasiswa berdasarkan tingkat kebugaran fisik. Studi ini menganalisis 324 mahasiswa dari jurusan pendidikan jasmani di Romania dan Ukraina, dengan menggunakan beberapa indikator kebugaran seperti VO2max, tes sit-up, push-up, dan lompatan jauh. Hasil analisis mengidentifikasi dua klaster utama pada laki-laki dan satu klaster pada perempuan. Laki-laki dalam Klaster 1 menunjukkan daya tahan kardiorespirasi dan kekuatan otot yang lebih tinggi, sementara Klaster 3 memiliki indeks massa tubuh lebih tinggi tetapi menunjukkan koordinasi dan kecepatan ekstremitas atas yang lebih baik. Klaster 2 yang terdiri dari perempuan menunjukkan performa lebih baik dalam koordinasi tetapi memiliki

SEKOLAH PASCASARJANA

kekuatan otot lebih rendah dibandingkan laki-laki. Dengan indeks Davies-Bouldin sebesar 0,017, hasil *clustering* ini memiliki pemisahan klaster yang kuat. Studi ini menyoroti perlunya program kebugaran yang lebih dipersonalisasi berdasarkan karakteristik fisik dan gender untuk meningkatkan kesehatan mahasiswa.

Penelitian oleh Liu dkk. (2022) menggunakan kombinasi Latent Profile Analysis (LPA) dan K-Means *Clustering* untuk mengidentifikasi profil kecemasan siswa. Studi ini melibatkan 9.738 siswa sekolah menengah di Tiongkok dan mengkategorikan mereka ke dalam tiga klaster utama, yaitu *Low-Risk Group* (29,5%), *Mild-Risk Group* (38,1%), dan *High-Risk Group* (32,4%). Analisis menggunakan LPA dan K-Means *Clustering* menunjukkan hasil yang konsisten dalam pengelompokan siswa, dengan tingkat kecemasan yang lebih tinggi ditemukan pada siswa perempuan, mereka yang memiliki prestasi akademik lebih rendah, dan siswa dengan tekanan akademik tinggi. Metode regresi ordinal multinomial menunjukkan bahwa siswa dalam kelompok risiko tinggi lebih mungkin mengalami kecemasan yang signifikan dibandingkan kelompok lainnya. Studi ini menyoroti pentingnya intervensi psikologis berbasis data untuk mengurangi kecemasan siswa dan meningkatkan kesejahteraan mental mereka.

Penelitian oleh Araka dkk. (2022) menggunakan Expectation-Maximization, Agglomerative *Hierarchical Clustering*, dan K-Means *Clustering* untuk mengidentifikasi profil pembelajaran mandiri (Self Regulated Learning - SRL) pada mahasiswa yang menggunakan sistem pembelajaran daring. Studi ini menganalisis data dari platform Learning Management System (LMS) di Open University, UK, dengan tujuan mengevaluasi algoritma mana yang paling efektif dalam mendeteksi pola SRL. Hasilnya menunjukkan bahwa Agglomerative *Hierarchical Clustering* menghasilkan klaster dengan pemisahan paling jelas dibandingkan metode lainnya. Penelitian ini mengidentifikasi tiga kelompok utama, yaitu *High-Level SRL Engagers*, *Moderate-Level Engagers*, dan *Low-Level Engagers*. Klaster pertama menunjukkan keterlibatan tinggi dalam strategi pembelajaran seperti manajemen waktu dan pencarian bantuan, sedangkan klaster terakhir memiliki keterlibatan yang lebih pasif. Temuan ini menekankan pentingnya desain sistem pembelajaran yang dapat mengakomodasi berbagai

SEKOLAH PASCASARJANA

tingkat regulasi diri mahasiswa untuk meningkatkan efektivitas pembelajaran daring.

Penelitian oleh Ramírez-Correa dkk. (2021) menerapkan K-Means *Clustering* dan Decision Tree C4.5 untuk mengembangkan model prediksi gaya belajar mahasiswa teknik. Studi ini menganalisis data dari mahasiswa teknik di Chili, dengan tujuan mengidentifikasi pola belajar yang dapat membantu pengajar menyesuaikan metode pembelajaran. Hasil analisis menunjukkan dua klaster utama, dengan mayoritas mahasiswa menunjukkan preferensi terhadap gaya belajar sensing (84%), visual (76%), dan aktif (70%). Model prediksi berbasis Decision Tree C4.5 memiliki akurasi sebesar 82,93%, yang menunjukkan efektivitas tinggi dalam mendeteksi pola belajar mahasiswa dengan hanya menggunakan sejumlah kecil pertanyaan awal. Studi ini menekankan pentingnya penyesuaian metode pengajaran berbasis data untuk meningkatkan hasil akademik mahasiswa teknik, terutama dalam lingkungan pembelajaran hibrida.

Berdasarkan penelitian terdahulu, penentuan jumlah klaster perlu mempertimbangkan lebih dari satu ukuran evaluasi. Silhouette Score memberikan informasi mengenai kohesi dan separasi berdasarkan jarak antarobjek, tetapi penggunaannya memiliki keterbatasan apabila dijadikan satu-satunya dasar penentuan jumlah klaster (Batool & Hennig, 2021). Penelitian IFAMR (2024) juga menunjukkan bahwa konfigurasi dengan nilai Silhouette lebih tinggi tidak selalu ditetapkan sebagai solusi akhir. Pada penelitian tersebut, K=2 memperoleh nilai Silhouette sebesar 0,203, sedangkan K=3 sebesar 0,182, tetapi K=3 dipilih berdasarkan kombinasi beberapa ukuran evaluasi serta pertimbangan interpretabilitas. Temuan tersebut menjadi dasar bahwa penentuan jumlah klaster dalam penelitian ini dilakukan dengan mempertimbangkan bukti struktural, statistik, dan substantif secara bersama-sama.

Dari berbagai hasil penelitian yang telah dikaji, dapat disimpulkan bahwa metode *clustering*, khususnya K-Means dan *Hierarchical Clustering*, terbukti efektif dalam mengelompokkan individu berdasarkan pola perilaku, preferensi belajar, hingga karakter psikososial. Pendekatan ini mampu mengungkap heterogenitas dalam populasi peserta didik dan menyediakan dasar untuk intervensi

SEKOLAH PASCASARJANA

yang lebih personal dan tepat sasaran. Di samping itu, literatur juga menegaskan pentingnya pemetaan karakter dan gaya belajar dalam konteks pendidikan untuk meningkatkan efektivitas proses pembelajaran, keterlibatan siswa, dan penguatan nilai-nilai kepribadian.

Relevansi pendekatan *clustering* semakin kuat ketika diaplikasikan dalam konteks pesantren yang memiliki karakteristik unik, baik dari segi lingkungan sosial, nilai spiritual, maupun pola hidup kolektif. Oleh karena itu, pemilihan atribut-atribut penelitian yang mencakup aspek karakter seperti kedisiplinan, tanggung jawab, kejujuran, empati, hingga preferensi gaya belajar VAK dilakukan secara cermat dengan mempertimbangkan kontribusi teoritis dan temuan empiris dari studi-studi sebelumnya. Atribut-atribut tersebut menjadi dasar dalam pemodelan klasterisasi untuk memahami profil santri secara lebih menyeluruh, sebagaimana akan dijelaskan pada subbab berikutnya.

2.2 Dasar Teori

Bab ini menyajikan landasan teori yang digunakan sebagai dasar dalam penelitian. Teori-teori yang dibahas mencakup konsep-konsep utama yang berkaitan dengan topik penelitian, sehingga dapat mendukung pemahaman terhadap permasalahan yang dikaji. Selain itu, disajikan pula hasil penelitian terdahulu yang relevan sebagai bahan pendukung dalam proses analisis. Landasan teori ini diharapkan dapat memberikan dasar yang kuat sebelum memasuki pembahasan yang lebih rinci pada subbab berikutnya.

SEKOLAH PASCASARJANA

2.2.1 Pondok Pesantren

Pesantren adalah lembaga pendidikan Islam yang memiliki peran strategis dalam pembentukan identitas dan karakter keislaman masyarakat Indonesia. Sebagai institusi pendidikan yang unik, pesantren menawarkan pendekatan holistik yang mengintegrasikan pendidikan agama, ilmu pengetahuan, serta pembinaan karakter. Lingkungan pesantren yang sarat nilai-nilai spiritual dan moral menjadikan institusi ini sebagai basis utama dalam melestarikan tradisi Islam di Indonesia (Fatmawati, 2024).Pesantren tidak hanya mencetak santri yang berilmu,

tetapi juga menanamkan nilai-nilai seperti kedisiplinan, kemandirian, dan keikhlasan yang menjadi bekal utama dalam kehidupan bermasyarakat.

Secara yuridis formal, pesantren didefinisikan berdasarkan Undang-Undang Nomor 18 Tahun 2019 tentang Pesantren sebagai lembaga yang berbasis masyarakat dan didirikan oleh perseorangan, yayasan, organisasi masyarakat Islam, dan/atau masyarakat yang menanamkan keimanan dan ketakwaan kepada Allah SWT, menyemaikan akhlak mulia serta memegang teguh ajaran Islam rahmatan lil alamin yang tercermin dari sikap religius, moderat, cinta tanah air, dan nilai-nilai luhur kebangsaan dalam rangka mewujudkan tujuan pendidikan nasional (Undang-Undang Republik Indonesia, 2019). Definisi ini menegaskan bahwa pesantren bukan sekadar lembaga pendidikan agama, melainkan juga lembaga pembentukan karakter bangsa yang memiliki akar kuat dalam masyarakat.

Secara tipologis, pesantren di Indonesia dapat diklasifikasikan ke dalam tiga kategori utama berdasarkan kurikulum dan orientasi pendidikannya. Pertama, pesantren salafi (tradisional), yaitu pesantren yang mempertahankan pengajaran kitab kuning sebagai inti kurikulum dengan menggunakan metode bandongan dan sorogan tanpa mengintegrasikan kurikulum pendidikan formal dari pemerintah. Kedua, pesantren khalafi (modern), yaitu pesantren yang telah mengintegrasikan kurikulum pendidikan formal (seperti MI, MTs, MA, atau SMP/SMA) ke dalam sistem pendidikannya, sehingga santri mendapatkan ijazah pendidikan nasional sekaligus pendidikan keagamaan. Ketiga, pesantren terpadu, yaitu kombinasi dari kedua model sebelumnya yang memadukan tradisi pengajaran kitab kuning dengan sistem pendidikan modern secara berimbang (Departemen Agama Republik Indonesia, 2003). Tipologi ini penting untuk dipahami karena karakteristik santri dan pola pembinaan di masing-masing jenis pesantren dapat berbeda secara signifikan.

Dalam kajian oleh Fatmawati (2024), pesantren juga berfungsi sebagai ruang transformasi sosial, khususnya dalam peran perempuan. Penelitian ini menyoroti peran penting pesantren dalam mengakomodasi perubahan peran kepemimpinan perempuan, meskipun masih menghadapi kendala budaya patriarkal. Hal ini menunjukkan bahwa pesantren memiliki fleksibilitas untuk

SEKOLAH PASCASARJANA

berkembang seiring dengan perubahan sosial tanpa kehilangan esensi nilai-nilai Islam.

Unsur pokok pesantren umumnya meliputi kyai, santri, asrama, masjid atau musholla, sistem pengajian, ustadz/asatidz, serta kegiatan harian santri. Kyai berperan sebagai pusat otoritas moral dan spiritual, sedangkan ustadz menjadi pendamping proses pembelajaran dan pembinaan. Asrama menjadi ruang pembiasaan, masjid menjadi pusat aktivitas ibadah, dan kegiatan harian menjadi sarana pembentukan kedisiplinan, tanggung jawab, serta kemampuan sosial santri. Dengan demikian, pesantren memiliki konteks pendidikan yang berbeda dari sekolah reguler karena proses pembinaan tidak hanya berbasis akademik, tetapi juga berbasis nilai, perilaku, dan rutinitas kehidupan sehari-hari.

Dalam konteks penelitian ini, pondok pesantren dipahami sebagai lingkungan pendidikan yang menghasilkan berbagai data santri. Data tersebut tidak hanya berupa penilaian santri, tetapi juga informasi mengenai karakter, kecenderungan belajar, dan performa pembinaan. Oleh sebab itu, pesantren menjadi konteks yang relevan untuk penerapan analisis profil berbasis data, khususnya dengan pendekatan clustering yang bertujuan menemukan kelompok santri berdasarkan kemiripan karakteristik.

2.2.2 Karakter Santri

Karakter dalam konteks pesantren merujuk pada seperangkat nilai, sikap, dan perilaku yang tertanam melalui proses pendidikan dan pembiasaan yang berlangsung secara konsisten. (Lickona, 1991), dalam bukunya *Educating for Character*, mendefinisikan karakter sebagai sekumpulan kebiasaan baik yang mencakup dimensi pengetahuan moral (*moral knowing*), perasaan moral (*moral feeling*), dan tindakan moral (*moral action*). Dalam konteks pesantren, karakter tidak hanya diajarkan secara verbal, tetapi juga dibentuk melalui pembiasaan, keteladanan, aturan kehidupan bersama, dan rutinitas keagamaan.

Karakter santri dalam penelitian ini diposisikan sebagai dimensi yang menggambarkan perilaku dan sikap santri dalam kehidupan pesantren. Dimensi ini mencakup lima atribut kuesioner, yaitu kedisiplinan, tanggung jawab, kejujuran,

SEKOLAH PASCASARJANA

empati, dan interaksi sosial. Kelima atribut tersebut dipilih karena mewakili karakter personal dan karakter sosial. Kedisiplinan, tanggung jawab, dan kejujuran menggambarkan karakter personal, sedangkan empati dan interaksi sosial menggambarkan karakter sosial santri dalam kehidupan kolektif di pesantren.

Kedisiplinan berkaitan dengan kemampuan santri mematuhi aturan, mengatur waktu, dan menjalankan kewajiban secara konsisten. Dalam lingkungan pesantren, kedisiplinan terlihat pada kepatuhan terhadap jadwal ibadah, pembelajaran, kegiatan asrama, serta tata tertib pondok. Penelitian mengenai academic self-discipline menunjukkan bahwa disiplin belajar memiliki hubungan dengan regulasi diri dan pencapaian belajar, sehingga atribut ini relevan untuk menggambarkan kesiapan santri dalam mengikuti sistem kehidupan pesantren.

Tanggung jawab menunjukkan kesadaran santri dalam menjalankan peran dan kewajiban, baik sebagai pelajar, anggota kamar, anggota kelas, maupun bagian dari komunitas pesantren. (Ward, 2005) mengembangkan instrumen tanggung jawab siswa yang menekankan aspek sikap dan perilaku dalam menjalankan kewajiban. Dalam konteks pesantren, tanggung jawab dapat terlihat melalui kesediaan santri menyelesaikan tugas, menjaga kebersihan, mengikuti kegiatan, dan mematuhi amanah yang diberikan.

Kejujuran merupakan nilai moral yang menunjukkan kesesuaian antara perkataan, tindakan, dan kenyataan. Dalam pendidikan karakter, kejujuran menjadi nilai dasar yang penting karena berhubungan dengan integritas dan kepercayaan. (Kosgolla dkk., 2023) menunjukkan bahwa kejujuran dapat dikaji melalui pelaporan diri dalam penelitian sosial dan perilaku. Pada kehidupan pesantren, kejujuran tercermin dalam perilaku santri ketika mengerjakan tugas, mengikuti aturan, melaporkan keadaan, dan berinteraksi dengan sesama. Atribut ini relevan karena pesantren tidak hanya membina kemampuan akademik, tetapi juga membentuk akhlak dan integritas.

Empati menggambarkan kemampuan santri memahami perasaan, kondisi, dan kebutuhan orang lain. Dalam lingkungan asrama yang bersifat kolektif, empati menjadi penting karena santri hidup bersama dengan teman yang memiliki latar belakang, kebiasaan, dan kemampuan berbeda. Empati mendukung terbentuknya

SEKOLAH PASCASARJANA

kepedulian, kerja sama, dan hubungan sosial yang sehat. Oleh karena itu, empati dapat diposisikan sebagai bagian dari karakter sosial santri.

Interaksi sosial menggambarkan kemampuan santri membangun hubungan dengan teman, pengurus, ustadz, dan lingkungan pesantren. (Gao dkk., 2024) menunjukkan bahwa interaksi sosial dalam pembelajaran dapat berhubungan dengan efisiensi belajar melalui kehadiran sosial dan keterlibatan belajar. Meskipun interaksi sosial dapat menjadi variabel tersendiri dalam beberapa penelitian, dalam penelitian ini interaksi sosial diposisikan sebagai indikator karakter sosial. Alasannya, fokus penelitian bukan mengukur hubungan sosial secara terpisah, tetapi menggambarkan perilaku sosial santri sebagai bagian dari profil santri. Dengan memasukkan interaksi sosial, karakter santri tidak hanya dipahami dari sisi pribadi, tetapi juga dari cara santri berhubungan dengan lingkungan sosialnya.

2.2.3 Gaya Belajar VAK (Visual, Auditori, Kinestetik)

Gaya belajar adalah cara yang cenderung dipilih oleh seseorang untuk menerima, memproses, dan mengingat informasi baru. Setiap individu memiliki preferensi gaya belajar yang berbeda-beda, dan pemahaman terhadap preferensi ini dapat membantu perancangan metode pembelajaran yang lebih efektif (Fleming & Mills, 1992). Salah satu model gaya belajar yang paling banyak digunakan dalam penelitian pendidikan adalah model VAK (Visual, Auditory, Kinesthetic) yang dikembangkan oleh Walter Burke Barbe, Raymond Swassing, dan Michael Milone pada tahun 1979 (Barbe dkk., 1979).

Model VAK membagi preferensi gaya belajar ke dalam tiga modalitas sensorik utama sebagai berikut:

1. Gaya Belajar Visual : Individu dengan gaya belajar visual cenderung menyerap informasi yang efektif melalui indera penglihatan. Mereka lebih mudah memahami materi yang disajikan dalam bentuk diagram, grafik, peta konsep, warna, dan representasi visual lainnya. Dalam konteks pesantren, santri dengan kecenderungan visual umumnya lebih mudah memahami materi yang ditulis di papan tulis, buku, atau ketika mereka membaca kitab secara mandiri.

SEKOLAH PASCASARJANA

2. Gaya Belajar Auditori : Individu dengan gaya belajar auditori menyerap informasi yang efektif melalui indera pendengaran. Mereka lebih mudah belajar melalui ceramah, diskusi, penjelasan lisan, dan pengulangan secara verbal. Dalam konteks pesantren, metode bandongan di mana kyai membacakan dan menjelaskan kitab secara lisan sementara santri menyimak sangat sesuai dengan gaya belajar ini.
3. Gaya Belajar Kinestetik : Individu dengan gaya belajar kinestetik menyerap informasi yang efektif melalui pengalaman langsung, gerakan, dan praktik. Mereka lebih mudah belajar melalui kegiatan yang melibatkan aktivitas fisik, simulasi, atau eksperimen. Dalam konteks pesantren, kegiatan praktik ibadah, latihan ceramah, olahraga, dan kegiatan keterampilan sangat relevan bagi santri dengan kecenderungan kinestetik.

Identifikasi gaya belajar santri dalam penelitian ini dilakukan menggunakan instrumen kuesioner berbasis skala Likert yang mengukur kecenderungan masing-masing modalitas. Pendekatan ini konsisten dengan penelitian-penelitian sebelumnya seperti Ramírez-Correa dkk., 2021 yang menggunakan K-Means Clustering untuk mengidentifikasi profil gaya belajar mahasiswa berdasarkan respons kuesioner, serta (Araka dkk., 2022) yang menunjukkan efektivitas metode clustering dalam mengidentifikasi pola belajar mandiri berdasarkan data perilaku. Dengan memahami preferensi gaya belajar santri secara kolektif per klaster, pengelola pesantren dapat merancang pendekatan pembelajaran yang lebih sesuai dan efektif.

2.2.4 Penilaian Santri

Penilaian santri merupakan dimensi yang menggambarkan capaian santri berdasarkan data yang tercatat di pesantren. Dalam penelitian ini, istilah penilaian santri digunakan sebagai istilah operasional karena indikator yang digunakan tidak hanya mencakup penilaian santri, tetapi juga partisipasi pembelajaran dan capaian pembinaan keagamaan. Oleh karena itu, istilah penilaian santri dipandang lebih sesuai daripada penilaian santri semata.

Pemilihan dimensi penilaian santri didasarkan pada pendekatan learning analytics dan student performance yang memanfaatkan data pembelajaran untuk

SEKOLAH PASCASARJANA

memahami kondisi peserta didik. Dalam learning analytics, data akademik, aktivitas belajar, keterlibatan peserta didik, dan catatan pembelajaran dapat digunakan untuk membangun profil peserta didik serta mendukung pengambilan keputusan pendidikan. (Gonzalez-Nucamendi dkk., 2022) menjelaskan bahwa learning analytics dapat digunakan untuk menentukan dimensi profil peserta didik dan mendukung prediksi performa akademik. (Johar dkk., 2023) juga menunjukkan bahwa analisis keterlibatan belajar dapat digunakan untuk memahami dan meningkatkan performa pembelajaran peserta didik.

Dalam penelitian ini, penilaian santri terdiri atas tiga atribut, yaitu nilai diskusi, nilai rata-rata rapor, dan nilai rutinitas ibadah. Nilai diskusi menggambarkan partisipasi santri dalam proses pembelajaran, kemampuan mengemukakan pendapat, keterlibatan dalam musyawarah atau diskusi kelas, serta respons terhadap materi. Penggunaan nilai diskusi sebagai bagian dari penilaian santri sejalan dengan konsep student engagement, yaitu keterlibatan peserta didik dalam proses belajar. Bergdahl dkk. (2024) menjelaskan bahwa keterlibatan peserta didik merupakan aspek penting dalam proses pendidikan karena berkaitan dengan pengalaman belajar dan hasil pendidikan.

Nilai rata-rata rapor digunakan untuk menggambarkan capaian akademik formal santri berdasarkan penilaian yang tercatat di pesantren. Dalam kajian educational data mining dan learning analytics, data akademik sering digunakan sebagai indikator utama untuk memahami performa peserta didik. (Fazil dkk., 2024) menggunakan data akademik, aktivitas pembelajaran, dan penilaian untuk membangun representasi perilaku peserta didik dalam analisis performa. Dengan demikian, nilai rata-rata rapor dalam penelitian ini digunakan sebagai indikator capaian akademik formal santri.

Nilai rutinitas ibadah digunakan untuk menggambarkan capaian pembinaan keagamaan santri. Atribut ini dipilih karena konteks penelitian berada di lingkungan pesantren, yaitu lembaga pendidikan yang tidak hanya menekankan capaian akademik, tetapi juga pembiasaan ibadah, pembentukan akhlak, dan penguatan karakter religius. (Solihah dkk., 2024) menunjukkan bahwa pendidikan karakter religius dapat dibentuk melalui kegiatan pembiasaan seperti pengulangan aktivitas

SEKOLAH PASCASARJANA

keagamaan dan pendampingan ibadah. Rahmawaty & Supriyadi, 2025 juga menjelaskan bahwa pembiasaan kegiatan ibadah rutin dapat menjadi sarana internalisasi nilai karakter religius peserta didik.

Dimensi penilaian santri menjadi pembeda penting dalam penelitian ini karena datanya berasal dari catatan pesantren, bukan dari persepsi responden. Penggabungan data kuesioner dan data institusional memungkinkan profil santri yang dihasilkan tidak hanya menggambarkan persepsi diri santri, tetapi juga mempertimbangkan capaian yang tercatat secara administratif oleh pesantren. Dengan demikian, data kuesioner digunakan untuk menggambarkan karakter dan gaya belajar, sedangkan data penilaian santri digunakan untuk menggambarkan capaian pembelajaran dan pembinaan santri.

Tabel 2.2 Atribut Penilaian Santri dalam Penelitian

Atribut Penilaian Santri	Makna dalam Penelitian	Jenis Data
Nilai Diskusi	Menggambarkan partisipasi dan keterlibatan santri dalam proses pembelajaran.	Data pesantren
Nilai Rata-rata Rapor	Menggambarkan capaian akademik formal santri.	Data pesantren
Nilai Rutinitas Ibadah	Menggambarkan capaian pembinaan keagamaan dan konsistensi ibadah santri.	Data pesantren

2.2.5 Instrumen Penelitian

Instrumen penelitian merupakan alat yang digunakan untuk memperoleh data sesuai dengan variabel penelitian. Dalam penelitian ini, instrumen digunakan terutama untuk mengukur dimensi karakter santri dan gaya belajar. Sementara itu, data penilaian santri diperoleh dari dokumen atau catatan resmi pesantren. Penempatan instrumen penelitian sebelum machine learning menjadi penting karena data yang dianalisis menggunakan clustering berasal dari proses pengukuran tersebut.

2.2.5.1 Kuesioner

Kuesioner merupakan teknik pengumpulan data yang dilakukan dengan menyajikan sejumlah pernyataan kepada responden untuk memperoleh informasi mengenai sikap, persepsi, kebiasaan, atau kecenderungan perilaku. Dalam penelitian ini, kuesioner digunakan untuk mengukur delapan atribut, yaitu lima

SEKOLAH PASCASARJANA

atribut karakter santri dan tiga atribut gaya belajar. Setiap pernyataan disusun agar responden dapat menilai kesesuaian pernyataan dengan kondisi dirinya.

Penggunaan kuesioner relevan karena beberapa atribut yang diteliti, seperti kejujuran, empati, kedisiplinan, tanggung jawab, interaksi sosial, dan gaya belajar, tidak seluruhnya dapat diambil dari data administratif pesantren. Atribut tersebut membutuhkan pengukuran berdasarkan respons santri. Oleh karena itu, kuesioner menjadi sumber data utama untuk dimensi karakter dan gaya belajar.

2.2.5.2 Skala Likert

Skala Likert adalah salah satu jenis skala pengukuran psikometri yang paling umum digunakan dalam penelitian ilmu sosial dan pendidikan. Dikembangkan oleh Rensis Likert pada tahun 1932, skala ini dirancang untuk mengukur sikap, persepsi, atau perilaku responden dengan menyajikan serangkaian pernyataan yang kemudian dinilai oleh responden menggunakan kategori respons yang terurut (Likert, 1932). Setiap kategori respons diberi nilai numerik yang mencerminkan tingkat persetujuan atau frekuensi tertentu.

Dalam penelitian ini, digunakan skala Likert dengan empat titik (1–4) tanpa opsi netral di tengah. Penggunaan skala genap ini disengaja untuk menghindari kecenderungan responden memilih kategori tengah (central tendency bias) yang dapat mengurangi variabilitas data. Keempat kategori respons yang digunakan adalah: (1) Sangat Tidak Sesuai, (2) Tidak Sesuai, (3) Sesuai, dan (4) Sangat Sesuai. Nilai yang lebih tinggi mencerminkan kesesuaian yang lebih besar antara pernyataan dan kondisi aktual responden. Skor akhir per konstruk diperoleh dengan menghitung rata-rata dari butir-butir yang membentuk konstruk tersebut, sehingga menghasilkan nilai kontinu dalam rentang 1 hingga 4 yang dapat digunakan langsung dalam analisis klasterisasi.

2.2.5.3 Validitas dan Reliabilitas Instrumen

Validitas instrumen mengacu pada sejauh mana suatu alat ukur benar-benar mengukur konstruk yang dimaksudkan. Dalam penelitian ini, validitas dinilai melalui dua pendekatan. Pertama, validitas isi *(content validity)* dinilai oleh panel ahli menggunakan Indeks Aiken's V sebagaimana telah dijelaskan sebelumnya (Aiken, 1985). Kedua, validitas butir (item validity) dinilai melalui korelasi item-

SEKOLAH PASCASARJANA

total (corrected item-total correlation), yaitu korelasi antara skor masing-masing butir dengan total skor konstruk. Butir dinyatakan valid apabila nilai korelasi item-total (r hitung) melebihi nilai r tabel pada taraf signifikansi tertentu (Azwar, 2012).

Reliabilitas mengacu pada tingkat konsistensi dan kestabilan suatu instrumen dalam mengukur konstruk yang sama. Salah satu ukuran reliabilitas yang paling umum digunakan adalah Cronbach's Alpha (α), yang digunakan untuk menilai konsistensi internal berdasarkan rata-rata korelasi antar butir dalam suatu konstruk (Cronbach, 1951). Secara matematis, cronchbach's Alpha dirumuskan sebagai berikut:

$$a = \frac{k}{k-1}\left(1 - \frac{\sum \sigma_i^2}{\sigma_t^2}\right) \tag{2.11}$$

di mana *k* menyatakan jumlah butir dalam suatu konstruk, σ_i^2 adalah varians dari butir ke-*i*, dan σ_t^2 merupakan varians total dari keseluruhan skor konstruk.

Nilai α berada pada rentang 0 hingga 1, di mana semakin tinggi nilainya menunjukkan tingkat konsistensi internal yang semakin baik. Secara umum, nilai α≥0,70 dianggap reliabel untuk penelitian di bidang ilmu sosial, sedangkan nilai α≥0,80 menunjukkan reliabilitas yang sangat baik (Hair dkk., 2019). Namun, dalam penelitian eksploratori, nilai α≥0,60 masih dapat diterima sebagai batas minimum reliabilitas (DeVellis, 2016).

Dengan demikian, nilai Cronbach's Alpha yang memenuhi ambang batas tersebut menunjukkan bahwa butir-butir dalam suatu konstruk memiliki konsistensi internal yang baik dan mampu mengukur aspek yang sama secara stabil. Dalam penelitian ini, pengujian reliabilitas menggunakan Cronbach's Alpha diterapkan pada tahap uji coba instrumen sebagaimana dijelaskan pada sub-bab 3.2.2.4, dengan hasil perhitungan untuk masing-masing konstruk dirangkum pada Tabel 4.1 (Bab IV).

2.2.6 Machine Learning

Machine Learning (ML) merupakan cabang dari kecerdasan buatan *(Artificial Intelligence)* yang berfokus pada pengembangan sistem yang mampu belajar dari data dan meningkatkan kinerjanya secara otomatis tanpa harus diprogram secara eksplisit. Melalui pendekatan ini, komputer dapat mengidentifikasi pola dan

SEKOLAH PASCASARJANA

hubungan dalam data sehingga dapat digunakan untuk mendukung proses pengambilan keputusan secara lebih adaptif dan berbasis data (Gheibi dkk., 2021).

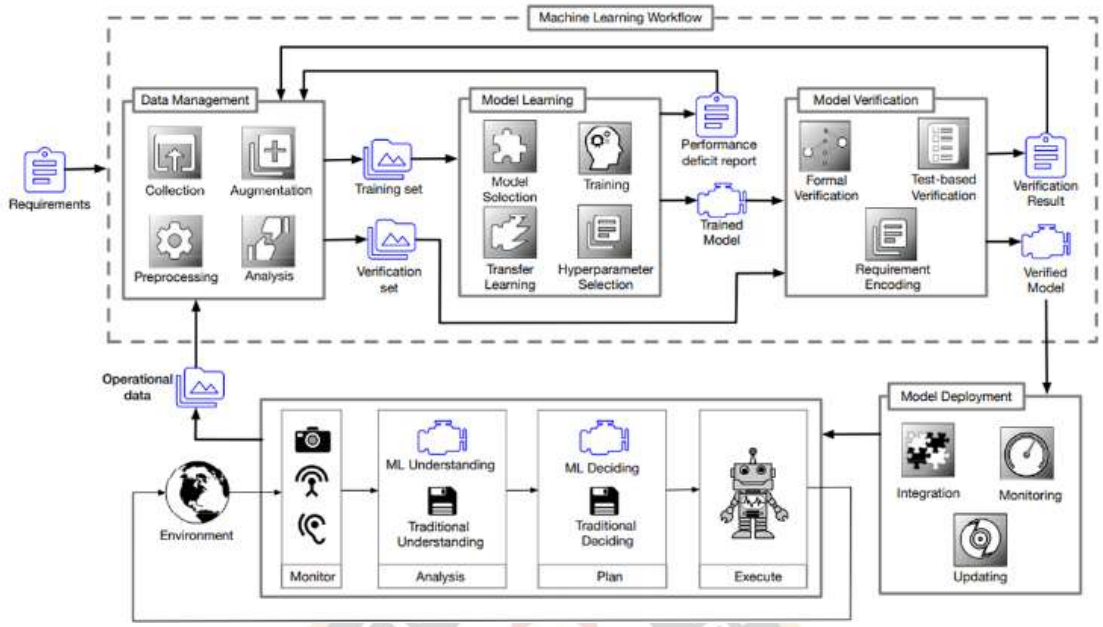
Dalam implementasinya, pengembangan model machine learning dilakukan melalui suatu siklus yang sistematis. Siklus ini umumnya terdiri dari beberapa tahapan utama, yaitu data management, model learning, model evaluation *(verification)*, dan *deployment* (Ashmore dkk., 2021). Pada tahap data management, dilakukan proses pengumpulan data, pembersihan data *(data cleaning)*, transformasi, serta eksplorasi awal untuk memastikan kualitas data yang digunakan. Tahap ini menjadi sangat penting karena kualitas data akan sangat memengaruhi performa model yang dihasilkan.

Tahap berikutnya adalah model learning, yaitu proses pelatihan model menggunakan algoritma tertentu berdasarkan data yang telah dipersiapkan. Pada tahap ini dilakukan pemilihan metode yang sesuai dengan karakteristik data serta tujuan analisis. Selain itu, proses pelatihan juga mencakup penyesuaian parameter agar model dapat menghasilkan performa yang optimal.

Selanjutnya adalah tahap model *evaluation* atau *verification*, di mana model diuji menggunakan data yang tidak digunakan saat proses pelatihan. Tujuan dari tahap ini adalah untuk mengetahui kemampuan model dalam melakukan generalisasi terhadap data baru. Evaluasi dilakukan menggunakan metrik tertentu yang disesuaikan dengan jenis permasalahan, seperti akurasi untuk klasifikasi atau metrik validitas klaster untuk clustering (Ashmore dkk., 2021).

Tahap terakhir adalah deployment, yaitu proses penerapan model ke dalam lingkungan nyata. Pada tahap ini, model tidak hanya digunakan tetapi juga dipantau secara berkala untuk memastikan performanya tetap stabil seiring dengan perubahan data di dunia nyata. Tahapan siklus hidup machine learning tersebut dapat dilihat pada Gambar 2.1.

SEKOLAH PASCASARJANA



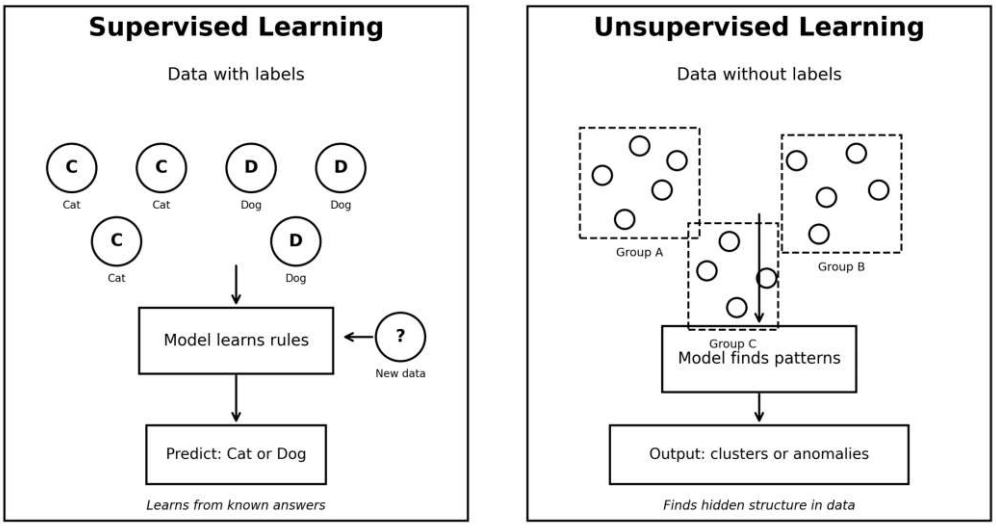
Gambar 2.1 Machine Learning *Lifecycle*

Berdasarkan Gambar 2.1, dapat dilihat bahwa proses pengembangan machine learning bersifat iteratif, di mana setiap tahap saling berkaitan dan dapat diulang kembali apabila diperlukan untuk meningkatkan kualitas model yang dihasilkan (Ashmore dkk., 2021).

Berdasarkan pendekatan pembelajarannya, machine learning secara umum dibagi menjadi beberapa jenis, di antaranya supervised learning dan unsupervised learning (Badillo dkk., 2020). Supervised learning merupakan metode pembelajaran yang menggunakan data berlabel, di mana setiap data memiliki target atau output yang telah diketahui. Model dilatih untuk mempelajari hubungan antara input dan output tersebut sehingga dapat digunakan untuk melakukan prediksi terhadap data baru.

SEKOLAH PASCASARJANA

Sebaliknya, unsupervised learning merupakan metode pembelajaran yang tidak menggunakan label pada data. Dalam pendekatan ini, model berusaha menemukan pola atau struktur tersembunyi dalam data secara mandiri. Salah satu teknik utama dalam pendekatan ini adalah clustering, yaitu proses pengelompokan data berdasarkan tingkat kemiripan karakteristik (Badillo dkk., 2020).



Gambar 2.2 Supervised dan Unsupervised Learning
(Fauzi, 2021)

Gambar 2.2 menunjukkan perbedaan mendasar antara supervised learning dan unsupervised learning. Pada supervised learning, model dilatih menggunakan data yang telah memiliki label, sedangkan pada unsupervised learning, model bekerja tanpa label dan berfokus pada pencarian pola dalam data (Badillo dkk., 2020; Fauzi, 2021).

Selain kedua pendekatan tersebut, terdapat pula metode lain seperti semi-supervised learning dan reinforcement learning. Semi-supervised learning menggabungkan data berlabel dan tidak berlabel dalam proses pembelajaran, sedangkan reinforcement learning berfokus pada pembelajaran berbasis interaksi dengan lingkungan melalui mekanisme reward dan punishment (Goodfellow dkk., 2016). Namun, pendekatan tersebut tidak digunakan dalam penelitian ini karena tidak sesuai dengan karakteristik data yang digunakan.

SEKOLAH PASCASARJANA

Dalam konteks penelitian ini, pendekatan yang digunakan adalah unsupervised learning, khususnya teknik clustering. Hal ini dikarenakan data yang digunakan tidak memiliki label kategori tertentu, sehingga diperlukan metode yang mampu mengelompokkan data secara otomatis berdasarkan kemiripan karakteristik.

Keunggulan machine learning terletak pada kemampuannya dalam mengolah data dalam jumlah besar serta mengidentifikasi pola yang kompleks yang sulit ditemukan menggunakan metode konvensional. Selain itu, machine learning juga

memungkinkan analisis yang lebih fleksibel dan adaptif terhadap berbagai jenis data (Gheibi dkk., 2021). Namun demikian, penerapannya juga memiliki beberapa tantangan, seperti ketergantungan pada kualitas data, kebutuhan komputasi yang cukup tinggi, serta kompleksitas dalam interpretasi hasil model (Ashmore dkk., 2021).

Dengan demikian, machine learning menjadi pendekatan yang relevan dalam penelitian ini, terutama dalam mendukung proses analisis berbasis data untuk mengidentifikasi pola karakter santri melalui teknik clustering seperti *Hierarchical Clustering* dan K-Means.

2.2.7 Clustering

Clustering merupakan teknik dalam unsupervised learning yang bertujuan mengelompokkan data ke dalam beberapa klaster berdasarkan tingkat kemiripan antarobjek. Objek yang memiliki karakteristik mirip ditempatkan pada klaster yang sama, sedangkan objek yang berbeda ditempatkan pada klaster yang berbeda. (Jain dkk., 1999) menjelaskan bahwa clustering merupakan salah satu teknik penting dalam analisis data karena dapat menemukan struktur tersembunyi pada data tanpa memerlukan label kelas.

Tujuan utama clustering adalah membentuk kelompok yang memiliki homogenitas tinggi di dalam klaster dan heterogenitas tinggi antar klaster. Dengan kata lain, data dalam satu klaster diharapkan memiliki kesamaan yang kuat, sedangkan data pada klaster berbeda diharapkan memiliki perbedaan yang cukup jelas. Dalam konteks pendidikan, clustering dapat digunakan untuk mengidentifikasi kelompok peserta didik berdasarkan pola belajar, karakteristik perilaku, performa akademik, atau keterlibatan pembelajaran.

Terdapat beberapa jenis clustering, di antaranya *partitioning clustering*, *hierarchical clustering*, *density-based clustering*, dan *model-based clustering*. *Partitioning clustering* membagi data ke dalam sejumlah klaster tertentu, contohnya K-Means. *Hierarchical clustering* membentuk struktur bertingkat yang dapat divisualisasikan melalui dendrogram. *Density-based clustering* mengelompokkan data berdasarkan kepadatan titik, sedangkan model-based

SEKOLAH PASCASARJANA

clustering membangun klaster berdasarkan asumsi model probabilistik tertentu. Penelitian ini menggunakan *Hierarchical Clustering* dan K-Means karena keduanya sesuai untuk data numerik dan dapat saling melengkapi dalam kerangka hybrid.

2.2.7.1 K-Means *Clustering*

Menurut (Suyanto, 2019), metode K-Means adalah algoritma klasterisasi yang paling tua dan paling banyak digunakan dalam berbagai aplikasi kecil hingga menengah karena kemudahan implementasinya. Algoritma ini bekerja dengan meminimalkan jarak rata-rata antara data dengan pusat klaster (*centroid*), sehingga data dengan kesamaan fitur lebih mungkin berada dalam grup yang sama. Algoritma ini pertama kali diperkenalkan oleh Mac Queen pada tahun 1967 dan hingga kini merupakan salah satu metode klasterisasi yang paling populer. Berikut adalah tahapan Algoritma K-Means :

1. Input:
   a. Data berjumlah n.
   b. Fungsi jarak d (misalnya Euclidean, Manhattan, atau Minkowski).
   c. Jumlah klaster k yang akan dicari.
2. Proses:
   a. Inisialisasi dengan memilih k titik pusat awal secara acak.
   b. Menghitung jarak setiap data (x) terhadap pusat klaster (c).
   c. Mengelompokkan data ke klaster dengan jarak terdekat (menghitung fungsi keanggotaan data).
   d. Memperbarui pusat klaster (c) sebagai rata-rata posisi semua data di klaster tersebut.
   e. Ulangi langkah ini hingga tidak ada perubahan klaster (konvergensi) atau jumlah iterasi maksimum tercapai.
3. Output:
   a. Posisi akhir dari pusat klaster (c).
   b. Pembagian data ke dalam *k* klaster.

Persamaan yang digunakan untuk menghitung pusat klaster diberikan dengan rumus:

SEKOLAH PASCASARJANA

$$C_i = \frac{1}{M} \sum_{j=1}^{M} X_j \quad (2.4)$$

Dengan x_j adalah data dalam klaster tersebut, dan M adalah jumlah data dalam klaster. Berikut adalah penjelasan detail :

1. Fungsi Jarak : Algoritma K-Means bergantung pada cara menghitung jarak antara data dan pusat klaster.
    a. *Euclidean distance :* Mengukur jarak lurus antara dua titik dalam ruang multidimensi.
    b. *Manhattan distance :* Mengukur jarak menggunakan pendekatan "blok kota" (sum of absolute differences).
    c. *Minkowski distance :* Merupakan generalisasi dari Euclidean dan Manhattan.
2. Kriteria Konvergensi, proses akan berhenti jika:
    a. Tidak ada data yang mengubah keanggotaan klasternya.
    b. Perubahan posisi pusat klaster menjadi sangat kecil.
    c. Jumlah iterasi maksimum tercapai.

2.2.7.2 Hierarchical *Clustering*

Menurut Stephen C. Johnson dalam penelitiannya yang berjudul *Hierarchical Clustering schemes*, *Hierarchical Clustering* adalah metode pengelompokan data yang membentuk struktur hierarki (bertingkat) berupa pohon (dendrogram), di mana setiap objek awalnya dianggap sebagai klaster tersendiri, kemudian secara bertahap digabungkan berdasarkan kedekatan (similaritas) hingga membentuk satu klaster besar atau sebaliknya (Johnson, 1967).

Secara umum, pendekatan dalam *Hierarchical Clustering* terbagi menjadi dua, yaitu *Divisive Hierarchical Clustering* (*Top Down*) dan *Agglomerative Hierarchical Clustering* (*Bottom Up*).

A. *Divisive Hierarchical Clustering*

Divisive Hierarchical Clustering merupakan pendekatan top-down dalam *Hierarchical Clustering*. Metode ini memulai proses dari satu klaster besar yang mencakup seluruh data, kemudian membaginya secara bertahap menjadi klaster-klaster yang lebih kecil berdasarkan tingkat ketidaksamaan antaranggota. Proses pembagian terus dilakukan hingga setiap objek menjadi klaster tunggal (*singleton*

cluster), atau hingga kriteria penghentian tertentu terpenuhi, misalnya jumlah klaster yang diinginkan telah tercapai (Shetty & Singh, 2021).

Menurut (Shetty & Singh, 2021), langkah-langkah *divisive Hierarchical Clustering* adalah sebagai berikut:

1. Mulailah dengan semua data dalam satu klaster besar.
2. Pada setiap iterasi, keluarkan "outsider" (data yang paling tidak cocok) dari klaster yang paling tidak kohesif.
3. Ulangi proses ini sampai setiap data berada di klaster tunggal (singleton cluster), atau hingga kriteria penghentian tertentu terpenuhi.

Pendekatan ini memiliki konsep yang berlawanan dengan *agglomerative clustering*. Meskipun lebih intuitif secara konsep, *divisive clustering* umumnya membutuhkan biaya komputasi yang lebih tinggi dan sangat sensitif terhadap keputusan awal karena pemisahan pada tahap awal berdampak signifikan terhadap hasil akhir.

B. *Agglomerative Hierarchical Clustering*

Agglomerative Hierarchical Clustering merupakan pendekatan *bottom-up* dalam *Hierarchical Clustering*. Pada tahap awal, setiap objek dianggap sebagai satu klaster tunggal (*singleton cluster*). Selanjutnya, dua klaster yang memiliki tingkat kemiripan paling tinggi atau jarak paling dekat akan digabungkan secara bertahap. Proses ini terus berlangsung hingga seluruh data tergabung menjadi satu klaster besar, atau hingga jumlah klaster yang diinginkan tercapai (Everitt dkk., 2011).

Langkah-langkah *agglomerative Hierarchical Clustering* menurut (Rokach & Maimon, 2005) adalah sebagai berikut:

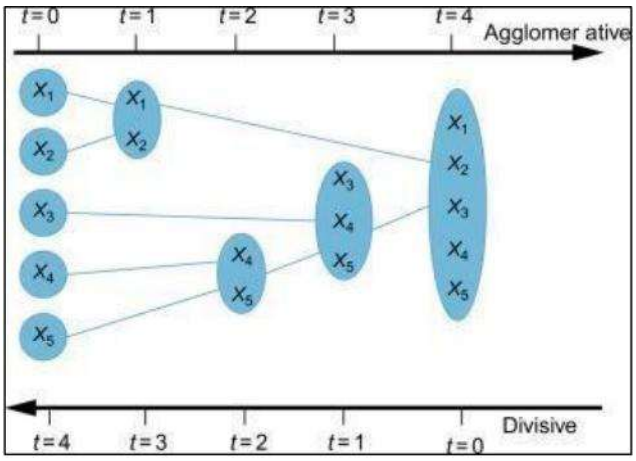
1. Anggap setiap data sebagai klaster tunggal *(singleton cluster).*
2. Hitung jarak antar klaster menggunakan ukuran jarak tertentu, seperti jarak Euclidean, kemudian gabungkan dua klaster dengan jarak minimum.
3. Ulangi proses ini hingga hanya ada satu klaster besar yang mencakup semua data, atau hingga jumlah klaster yang diinginkan tercapai.

Dalam proses *agglomerative clustering*, penggabungan klaster ditentukan oleh dua komponen utama, yaitu ukuran jarak (*distance measure*) dan metode linkage. Ukuran jarak digunakan untuk menghitung tingkat perbedaan antarobjek,

SEKOLAH PASCASARJANA

sedangkan metode linkage menentukan cara menghitung jarak antar klaster yang sudah terbentuk. Beberapa metode linkage yang umum digunakan antara lain *single linkage, complete linkage, average linkage, centroid linkage, dan Ward Linkage.*

Gambar 2.3 menunjukkan ilustrasi antara pendekatan *agglomerative* dan *divisive* dalam metode *Hierarchical Clustering*.



Gambar 2.3 *Hierarchical Clustering*
(Shetty & Singh, 2021)

Dalam penelitian ini, metode linkage yang digunakan adalah Ward Linkage (juga dikenal sebagai Ward's Minimum Variance Method), yang pertama kali diusulkan oleh Ward pada tahun 1963. Berbeda dengan metode linkage lain yang mengukur jarak antar titik data, Ward Linkage bekerja dengan meminimalkan peningkatan total within-cluster sum of squares (WCSS) atau variansi internal klaster setiap kali dua klaster digabungkan. Prinsip ini menjadikan Ward Linkage sebagai metode yang cenderung menghasilkan klaster yang lebih kompak dan berukuran relatif seimbang (Murtagh & Legendre, 2014).

SEKOLAH PASCASARJANA

Secara matematis, jarak Ward antara dua klaster A dan B dihitung sebagai:

$$d_{Ward}(A, B) = \frac{n_A n_B}{n_A + n_B}(\mu A - \mu B)^2 \qquad (2.5)$$

Dengan :

- n_A *dan* n_B adalah jumlah anggota pada klaster *A* dan *B*.
- μ_A dan μ_B adalah nilai rata-rata (centroid) dari masing-masing klaster,

Penggabungan dua klaster terjadi ketika selisih WCSS yang dihasilkan paling kecil, artinya penggabungan tersebut paling sedikit meningkatkan variabilitas internal. Metode ini dipilih karena terbukti menghasilkan klaster yang lebih kohesif

dan interpretatif dibandingkan metode single linkage maupun complete linkage, terutama pada data yang tidak memiliki bentuk klaster yang ekstrem (Everitt dkk., 2011).

2.2.8 Dendrogram dan Penentuan Jumlah Klaster

Dendrogram merupakan visualisasi berbentuk pohon yang menggambarkan proses penggabungan klaster secara hierarkis dalam *Hierarchical Clustering*. Pada dendrogram, sumbu vertikal menunjukkan jarak penggabungan atau linkage distance, sedangkan sumbu horizontal menunjukkan objek atau observasi yang dikelompokkan. Semakin rendah posisi penggabungan dua cabang pada sumbu vertikal, semakin tinggi tingkat kemiripan antara objek atau klaster yang digabungkan. Sebaliknya, semakin tinggi posisi penggabungan, semakin besar perbedaan antara klaster yang disatukan (Murtagh & Legendre, 2014).

Penentuan jumlah klaster yang optimal dari dendrogram dapat dilakukan dengan teknik pemotongan dendrogram *(dendrogram cutting)*. Pemotongan dilakukan pada tingkat linkage distance tertentu dengan memperhatikan perubahan jarak penggabungan antar tahap. Apabila terdapat lonjakan linkage distance yang besar pada suatu tahap, kondisi tersebut mengindikasikan bahwa klaster yang akan digabungkan sudah memiliki perbedaan karakteristik yang cukup besar. Oleh karena itu, titik pemotongan umumnya dipilih sebelum lonjakan terbesar terjadi, sehingga jumlah cabang utama yang terbentuk setelah pemotongan dapat digunakan sebagai pertimbangan dalam menentukan jumlah klaster yang optimal (Everitt dkk., 2011).

SEKOLAH PASCASARJANA

Pemilihan titik cut-off tidak dilakukan hanya berdasarkan keberadaan lonjakan semata, melainkan berdasarkan perubahan linkage distance yang paling signifikan selama keseluruhan proses penggabungan klaster. Dengan demikian, interpretasi dendrogram bersifat kontekstual dan harus mempertimbangkan pola perubahan jarak secara menyeluruh, bukan hanya satu titik lonjakan tunggal (Jain & Dubes, 1988).

2.2.9 Metode Gabungan *Hierarchical Clustering* dan K-Means *Clustering*

Penggabungan metode *Hierarchical Clustering* dengan *K-Means Clustering* bertujuan untuk menghasilkan hasil pengelompokan (*clustering*) yang lebih optimal, terutama dalam menangani data dengan bentuk dan ukuran yang bervariasi. Pendekatan ini menggabungkan keunggulan visualisasi dan pemetaan awal dari *Hierarchical Clustering* dengan efisiensi iteratif dari *K-Means Clustering* dalam menyempurnakan posisi pusat klaster (*centroid*) (Ruswanti dkk., 2024)

Langkah awal dalam metode gabungan ini adalah menerapkan *Hierarchical Clustering* terhadap dataset yang telah melalui tahap pra-pemrosesan. Teknik ini membentuk struktur pohon atau *dendrogram* yang menggambarkan bagaimana data dikelompokkan secara bertingkat berdasarkan kemiripan, menggunakan metode seperti *single linkage* atau *Ward linkage* (Bueno dkk., 2023). Keunggulan utama dari *Hierarchical Clustering* adalah kemampuannya dalam memetakan struktur data tanpa memerlukan jumlah klaster secara eksplisit di awal serta kemampuannya mengenali bentuk dan ukuran klaster yang kompleks (Ruswanti dkk., 2024).

Namun, meskipun kuat dalam hal eksplorasi data dan visualisasi awal, *Hierarchical Clustering* memiliki beberapa keterbatasan teknis yang menyebabkan metode ini sering tidak dilanjutkan hingga tahap akhir analisis. Salah satunya adalah keterbatasan dalam mengoptimalkan ulang struktur klaster setelah proses penggabungan selesai, karena sifatnya yang *greedy* dan tidak dapat membatalkan keputusan penggabungan sebelumnya. Selain itu, metode ini memiliki sensitivitas terhadap outlier dan hasilnya bisa sangat bergantung pada pilihan metode linkage (Richelle dkk., 2024b).

SEKOLAH PASCASARJANA

Untuk mengatasi keterbatasan tersebut, hasil awal dari *Hierarchical Clustering* digunakan untuk menentukan *initial centroids* bagi K-Means *Clustering*. Dengan cara ini, kelemahan K-Means yang biasanya sensitif terhadap pemilihan titik awal dapat dihindari. Selanjutnya, K-Means bekerja dengan mengoptimalkan posisi pusat klaster melalui proses iteratif, yaitu mengelompokkan data berdasarkan kedekatan dengan centroid, lalu menghitung ulang centroid hingga nilai tersebut stabil atau batas iterasi tercapai (Ruswanti dkk., 2024).

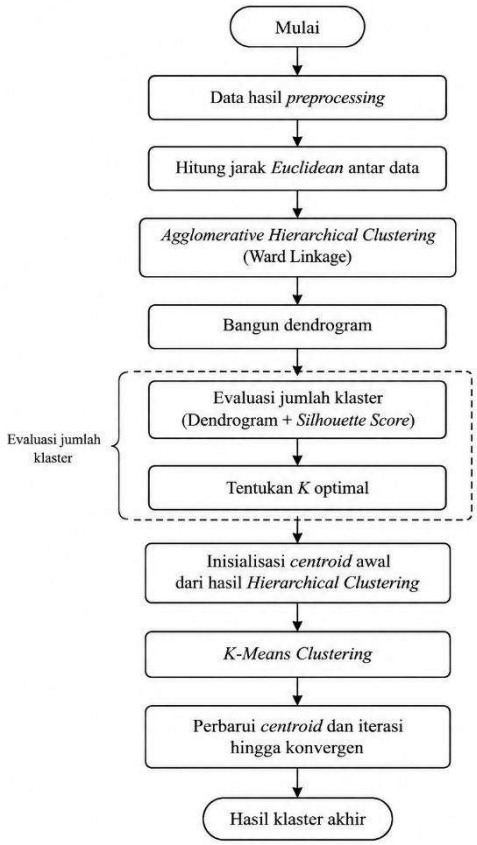
Dengan demikian, meskipun keduanya merupakan metode *clustering*, kombinasi keduanya dibutuhkan karena *Hierarchical Clustering* unggul dalam eksplorasi struktur data awal, sedangkan K-Means efektif dalam optimasi akhir pengelompokan (Bueno dkk., 2023). Metode gabungan ini telah terbukti menghasilkan pengelompokan yang lebih baik dalam berbagai konteks, seperti studi mengenai produktivitas mahasiswa dalam ruang kelas (Bueno dkk., 2023)dan segmentasi sikap mahasiswa kedokteran (Richelle dkk., 2024).

Untuk menentukan jumlah klaster optimal dari dendrogram, digunakan pendekatan visual dengan menetapkan nilai threshold pemotongan pada tingkat ketinggian *(height)* dendrogram yang menunjukkan lonjakan jarak penggabungan (linkage distance) yang signifikan antar klaster. Nilai threshold ini ditentukan berdasarkan analisis visual terhadap perubahan jarak saat dua klaster terakhir digabung. Semakin besar lonjakan jarak tersebut, semakin besar kemungkinan bahwa dua klaster yang digabung sudah tidak terlalu mirip, sehingga pemotongan dilakukan sebelum lonjakan itu terjadi. Pemotongan ini menghasilkan sejumlah cabang besar yang mewakili klaster yang paling alami. Untuk mendukung pemotongan tersebut, juga digunakan metrik Silhouette Score untuk membandingkan beberapa jumlah klaster dan memilih nilai K dengan kualitas pengelompokan terbaik.

Dalam penelitian ini, digunakan Euclidean Distance sebagai fungsi pengukuran jarak karena kesederhanaannya dan kesesuaiannya dengan data interval yang berskala seragam (telah distandarisasi), seperti skor kuesioner skala Likert. Euclidean Distance merupakan pilihan umum dalam *clustering* karena menghitung jarak garis lurus antar titik dalam ruang multidimensi, yang intuitif dan konsisten untuk data kontinu. Meskipun terdapat ukuran jarak lain seperti Manhattan, Mahalanobis, dan Jaccard, ukuran tersebut tidak digunakan karena karakteristik data penelitian ini berupa data numerik kontinu yang telah distandarisasi, maka Euclidean Distance dipandang sebagai pendekatan yang paling tepat, efisien, dan sejalan dengan karakteristik data.

SEKOLAH PASCASARJANA

Gambar 2.4 menggambarkan proses *clustering* menggunakan kombinasi antara metode *Hierarchical Clustering* dengan *K-means clustering*. Ilustrasi kombinasi kedua metode tersebut dapat dilihat pada Gambar 2.4.



Gambar 2.4 Gabungan Metode *Hierarchical Clustering* dan *K-Means* diadaptasi dari Bueno dkk. (2023) dan Ruswanti dkk. (2024)

Tujuan dari gabungan metode ini adalah memperoleh struktur klaster yang representatif dan optimal, baik secara visual maupun secara numerik. Berikut ini adalah uraian teknis dari tahapan-tahapan proses yang dilakukan :

1. Input dan Persiapan Data

Proses dimulai dengan pengumpulan dan input data berupa hasil tanggapan responden (santri) melalui instrumen kuesioner yang telah dikembangkan dan divalidasi. Data ini kemudian masuk ke tahap pre-processing, yang meliputi :

SEKOLAH PASCASARJANA

a. Normalisasi/standardisasi data untuk menyetarakan skala antar variabel,
b. Penanganan nilai kosong atau tidak valid (missing value),
c. Identifikasi dan penghapusan outlier berdasarkan batas IQR atau Z-score.

Langkah ini bertujuan untuk menghindari bias jarak akibat perbedaan skala antar fitur saat dilakukan proses *clustering* (Suyanto, 2019).

2. *Agglomerative Hierarchical Clustering (AHC)*

Agglomerative Hierarchical Clustering (AHC) adalah suatu metode *hierarhical clustering* yang memiliki sifat bottom-up yaitu menggabungkan n ke dalam satu kelompok tunggal (Bouguettaya dkk., 2015). Dalam proses agglomerative clustering, penggabungan dua klaster pada setiap tahap ditentukan oleh metode linkage yang mengukur kedekatan antar klaster. Metode linkage yang digunakan dalam penelitian ini adalah *Ward Linkage*, yaitu metode yang menggabungkan klaster berdasarkan peningkatan variansi internal (*within-cluster sum of squares*) yang paling minimal, sehingga menghasilkan struktur klaster yang lebih kompak dan seimbang. Penjelasan lengkap mengenai *Ward Linkage* beserta rumus perhitungannya disajikan pada sub-bab 2.2.8.

Langkah - langkah perhitungan *Agglomerative Hierarchical Clustering* adalah sebagai berikut :

1. Pada penelitian ini perhitungan jarak menggunakan *Euclidean Distance.* Seperti pada rumus di bawah ini :

$$X_{rs} = \sqrt{\sum_{k=1}^{p}(y_{rk} - y_{sk})^2} \quad (2.6)$$

SEKOLAH PASCASARJANA

Keterangan :

x_{rs} = Jarak antara objek r dengan s

y_{rk} = Nilai objek r pada variabel ke k

y_{sk} = Nilai objek s pada variabel ke k

P = Banyaknya variabel yang diamati

2. Penggabungan dua objek terdekat :

Apabila nilai jarak dari objek a dan b yang terkecil dibanding jarak antar objek yang lain, maka gabungan dua objek yang pertama yaitu x_ab.

3. Perbarui matriks jarak :

   Apabila x_ab merupakan jarak paling kecil dalam perhitungan jarak yang menggunakan *Euclidian Distance,* maka rumus untuk metode *agglomerative* yang digunakan, yaitu Ward Linkage, adalah:

   Rumus perhitungan jarak Ward Linkage beserta uraian variabelnya disajikan pada persamaan (2.5) sub-bab 2.2.8.

4. Selanjutnya lakukan langkah ke 2 dan langkah ke 3 sampai tersisa satu *cluster.*

5. Penentuan Nilai K Optimal :

   Nilai jumlah klaster optimal (K) tidak diambil saat proses mencapai K = 1, melainkan diperoleh dengan memotong dendrogram pada titik lonjakan jarak (linkage distance) yang signifikan. Evaluasi dibantu dengan metrik Silhouette Score (Bueno dkk., 2023; Ruswanti dkk., 2024), yaitu :

$$s(i) = \frac{b(i)-a(i)}{\max\{a(i),b(i)\}} \qquad (2.7)$$

   Dengan *a(i)* = rata-rata jarak intra-klaster, dan *b(i)* = rata-rata jarak ke klaster tetangga terdekat.

3. *K-Means Clustering*

Hasil dari proses Agglomerative *Hierarchical Clustering* digunakan sebagai dasar untuk inisialisasi centroid pada algoritma K-Means. Titik pusat klaster (centroid) yang terbentuk dari pemotongan dendrogram pada jumlah klaster optimal (K) akan dijadikan sebagai centroid awal dalam proses K-Means. Pendekatan ini dimaksudkan untuk mengatasi kelemahan K-Means yang sensitif terhadap pemilihan centroid secara acak.

Selanjutnya dilakukan proses pengelompokan menggunakan algoritma K-Means *Clustering*. Dengan titik awal centroid yang telah ditentukan sebelumnya, K-Means mengelompokkan data berdasarkan jarak minimum ke centroid, kemudian melakukan pembaruan posisi centroid secara iteratif hingga proses konvergen atau jumlah iterasi maksimum tercapai.

SEKOLAH PASCASARJANA

2.2.10 Evaluasi Klaster

Evaluasi klaster dilakukan untuk menilai kualitas hasil clustering dan mendukung pemilihan jumlah klaster yang sesuai. Dalam penelitian ini, evaluasi klaster dilakukan menggunakan Silhouette Score dan Elbow Method. Kedua metode tersebut digunakan sebagai pendukung interpretasi dendrogram sehingga keputusan jumlah klaster tidak hanya bergantung pada visualisasi.

2.2.10.1 Silhoutte Score

Silhouette Score merupakan salah satu metode evaluasi clustering yang paling populer dan banyak digunakan karena kemampuannya mengukur keseimbangan antara kohesi dan separasi secara simultan (Rousseeuw, 1987). Metode ini memberikan penilaian untuk setiap data individu, sehingga tidak hanya memberikan gambaran kualitas model secara keseluruhan, tetapi juga memungkinkan identifikasi data mana yang mungkin salah dikelompokkan.

Nilai Silhouette untuk setiap data dihitung dengan membandingkan dua komponen utama. Komponen pertama, $a(i)$, adalah rata-rata jarak antara data ke-i*i* dengan seluruh data lain dalam cluster yang sama. Komponen ini merepresentasikan tingkat kohesi; semakin kecil nilai $a(i)$, semakin rapat dan padat cluster tempat data tersebut berada. Komponen kedua, $b(i)$, adalah rata-rata jarak minimum antara data ke-*i* dengan seluruh data pada cluster lain yang paling dekat (nearest neighboring cluster). Komponen ini merepresentasikan tingkat separasi, semakin besar nilai $b(i)$, semakin jauh cluster tersebut dari cluster lainnya. Secara matematis, nilai Silhouette Score untuk setiap data ke-i*i* dirumuskan sebagai berikut:

$$s(i) = \frac{b(i)-a(i)}{\max\{a(i),b(i)\}} \tag{2.8}$$

Dari rumus tersebut, dapat dianalisis beberapa skenario:

a. Jika $s(i)$ mendekati 1 ($b(i) \gg a(i)$), artinya data berada sangat dekat dengan anggota cluster-nya sendiri dan sangat jauh dari cluster lain. Ini adalah kondisi ideal.

SEKOLAH PASCASARJANA

b. Jika $s(i)$ mendekati 0 ($b(i) \approx a(i)$), artinya data berada di perbatasan antara dua cluster. Penempatan data tersebut bersifat ambigu.
c. Jika $s(i)$ bernilai negatif ($b(i) < a(i)$), artinya data secara rata-rata lebih dekat ke cluster lain dibandingkan ke cluster-nya sendiri. Ini merupakan indikasi kuat bahwa data tersebut salah penempatan.

Nilai Silhouette Score untuk keseluruhan model adalah rata-rata dari $s(i)$ untuk seluruh data. Semakin tinggi nilai rata-rata ini (mendekati 1), semakin baik kualitas clustering yang dihasilkan. Dalam penelitian ini, Silhouette Score dipilih sebagai metrik evaluasi utama karena kemampuannya yang intuitif dalam menilai keseimbangan kohesi-separasi dan kemudahannya dalam visualisasi melalui *silhouette plot*.

Meskipun Silhouette Score banyak digunakan dalam evaluasi clustering, nilai tersebut tidak selalu dapat dijadikan satu-satunya dasar dalam menentukan jumlah klaster. Batool dkk. (2020) menunjukkan bahwa Average Silhouette Width memiliki keterbatasan ketika digunakan untuk memperkirakan jumlah klaster, sehingga penggunaannya perlu dipertimbangkan bersama informasi atau metode evaluasi lainnya. Dengan demikian, nilai Silhouette yang lebih tinggi menunjukkan kualitas pemisahan yang lebih baik menurut ukuran tersebut, tetapi tidak secara otomatis menunjukkan bahwa konfigurasi tersebut merupakan struktur klaster yang paling sesuai dengan tujuan analisis.

2.2.10.2 Interpretasi dan Keterbatasan Silhouette Score

Nilai Silhouette berada pada rentang −1 sampai 1. Nilai yang semakin mendekati 1 menunjukkan bahwa suatu objek memiliki kedekatan yang lebih tinggi dengan klasternya sendiri dibandingkan dengan klaster lain, sedangkan nilai mendekati 0 menunjukkan adanya tumpang tindih antarklaster. Nilai negatif dapat menunjukkan bahwa suatu objek lebih dekat dengan klaster lain dibandingkan dengan klaster tempat objek tersebut ditempatkan (Kaufman & Rousseeuw, 1990).

Interpretasi nilai Silhouette perlu mempertimbangkan karakteristik data dan tujuan analisis. Pada data yang merepresentasikan karakteristik individu, batas antarklaster tidak selalu bersifat tegas karena karakteristik antarindividu dapat memiliki kemiripan dan tumpang tindih. Oleh karena itu, nilai Silhouette yang tidak

SEKOLAH PASCASARJANA

tinggi tidak secara langsung menunjukkan bahwa hasil clustering tidak dapat digunakan. Nilai tersebut perlu dibaca bersama dengan struktur klaster, ukuran evaluasi lainnya, serta interpretasi karakteristik kelompok yang terbentuk.

2.2.10.3 Elbow Method

Elbow Method merupakan pendekatan heuristik yang digunakan untuk menentukan jumlah cluster (*k*) yang optimal, khususnya dalam konteks algoritma *centroid-based* seperti K-Means. Metode ini bekerja dengan menganalisis hubungan antara jumlah cluster (*k*) dan nilai *Within-Cluster Sum of Squares* (WCSS), yang juga dikenal sebagai inersia (*inertia*).

Nilai WCSS menggambarkan total variasi dalam cluster, yaitu jumlah kuadrat jarak antara setiap data dengan pusat cluster (*centroid*) tempatnya bernaung. Secara matematis, WCSS untuk *k* cluster dirumuskan sebagai berikut:

$$\mathrm{WCSS} = \sum_{i=1}^{k} \sum_{x \in C_i} \|x - \mu_i\|^2 \qquad (2.9)$$

Di mana *Ci* adalah himpunan data pada cluster ke-*i* dan μi adalah centroid dari cluster *Ci*. Konsekuensi logis dari rumus ini adalah: semakin banyak cluster yang dibuat, maka ukuran setiap cluster akan semakin kecil dan anggota-anggotanya akan semakin dekat dengan centroid-nya, sehingga nilai WCSS akan semakin menurun. Pada titik ekstrem, jika *k* sama dengan jumlah data, maka setiap data menjadi centroid-nya sendiri dan WCSS akan bernilai nol yang secara praktis tidak berguna.

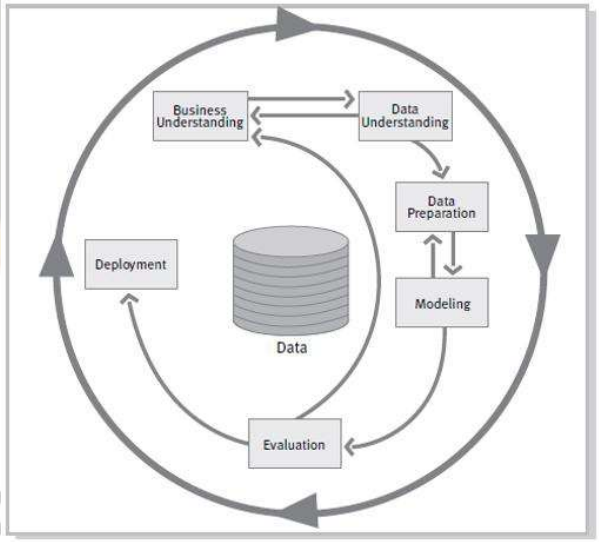
Prinsip Elbow Method adalah mencari titik di mana penambahan jumlah cluster tidak lagi memberikan penurunan WCSS yang signifikan. Titik ini divisualisasikan sebagai "siku" (*elbow*) pada grafik garis antara *k* (sumbu-X) dan WCSS (sumbu-Y). Sebelum titik siku, kurva menurun tajam; setelah titik siku, kurva melandai. Titik transisi inilah yang diusulkan sebagai jumlah cluster optimal. Keunggulan Elbow Method terletak pada kesederhanaannya. Namun, metode ini memiliki kelemahan yang cukup signifikan, yaitu interpretasi titik "siku" seringkali bersifat subjektif, terutama jika kurva penurunan WCSS bersifat halus dan tidak menunjukkan perubahan yang tajam. Oleh karena itu, dalam penelitian ini, Elbow Method tidak digunakan secara tunggal, melainkan sebagai pelengkap yang

SEMARANG
SEKOLAH PASCASARJANA

dikombinasikan dengan metode lain seperti Silhouette Score dan analisis dendrogram untuk menentukan jumlah cluster optimal secara triangulasi.

2.2.11 CRISP-DM

CRISP-DM (Cross-Industry Standard Process for Data Mining) menyediakan standar proses umum untuk pemecahan masalah dari bisnis atau unit penelitian dengan strategi penambangan data. Penambangan data (data mining) merupakan proses untuk menemukan pola dan tren yang berguna berdasarkan dataset yang besar. Beberapa data analis sering mengulang-ulang tahap pemahaman data dan pengolahan data, sehingga sulit maju ketahap selanjutnya. CRISP-DM dapat berfungsi sebagai peta jalan yang bermanfaat untuk menjaga penambangan data terorganisir dan bergerak menuju pemenuhan tujuan penelitian (Larose & Larose, 2014). Siklus hidup CRISP-DM dapat dilihat pada Gambar 2.5.



Gambar 2.5 Siklus Hidup CRISP-DM (Chapman dkk. 2000)

Siklus hidup CRISP-DM dibagi kedalam 6 fase, dimana setiap fase bisa bergerak dinamis. Hasil dari suatu fase akan menentukan langkah selanjutnya. Panah menunjukkan ketergantungan antar fase dan lingkaran terluar menunjukkan proses penanambangan data itu sendiri. Untuk memahami lebih dalam mengenai siklus hidup CRISP-DM, berikut ini penjabarannya (Chapman & others, 2000):

1. Fase Pemahaman Bisnis *(Business Understanding)*

Fase ini merupakan fase awal yang berfokus untuk memahami tujuan dan persyaratan proyek dari perspektif bisnis, kemudian mengubah pengetahuan ini menjadi definisi masalah penambangan data dan merancang rencana awal untuk mencapai tujuan. Di dalam fase ini terdapat *determine business objective*, *Assess Situation*, *Determine Data Mining Goals* dan *Produce Project Plan*.

2. Fase Pemahaman Data (*Data Understanding Phase*)

   Fase ini tentang pemahaman data yang dimulai dengan pengumpulan data awal, dilanjutkan dengan eksplorasi data, identifikasi masalah kualitas data dan akhirnya dihasilkan pengetahuan awal mengenai data untuk membuat hipotesa langkah selanjutnya. Di dalam fase ini terdapat *collect initial data*, deskripsi data, eksplor data dan *verify data quality* 18.

3. Fase Pengolahan Data (*Data Preparation Phase*)

   Fase ini tentang pengolahan data mencakup semua kegiatan yang diperlukan untuk membangun dataset awal menjadi dataset akhir (data yang akan dimasukkan ke dalam pemodelan). Fase ini cenderung dilakukan beberapa kali dan tidak dalam urutan yang ditentukan. Dalam fase ini akan banyak tugas yang harus dilakukan, diantaranya : *select/collect* data, clean data, construct data, integrate data, dan format data.

4. Fase Pemodelan (*Modelling Phase*)

   Dalam fase ini, berbagai teknik pemodelan dipilih dan diterapkan, kemudian kalibrasi parameter model ke nilai optimal. Biasanya, ada beberapa teknik untuk jenis masalah penambangan data yang sama, sehingga kembali ke fase pengolahan data sering diperlukan. Di dalam fase ini terdapat *select modelling technique*, *generate test design* dan *Assess Model*.

5. Fase Evaluasi (*Evaluation Phase*)

   Fase ini telah dibangun model yang tampaknya memiliki kualitas tinggi dari perspektif analisis data. Sebelum melanjutkan ke penyebaran/deployment akhir model, penting untuk mengevaluasinya secara meyeluruh dan meninjau langkah-langkah sebelumnya untuk memastikan model tersebut benar mencapai tujuan bisnis. Tujuan utama fase ini adalah untuk memastikan apakah

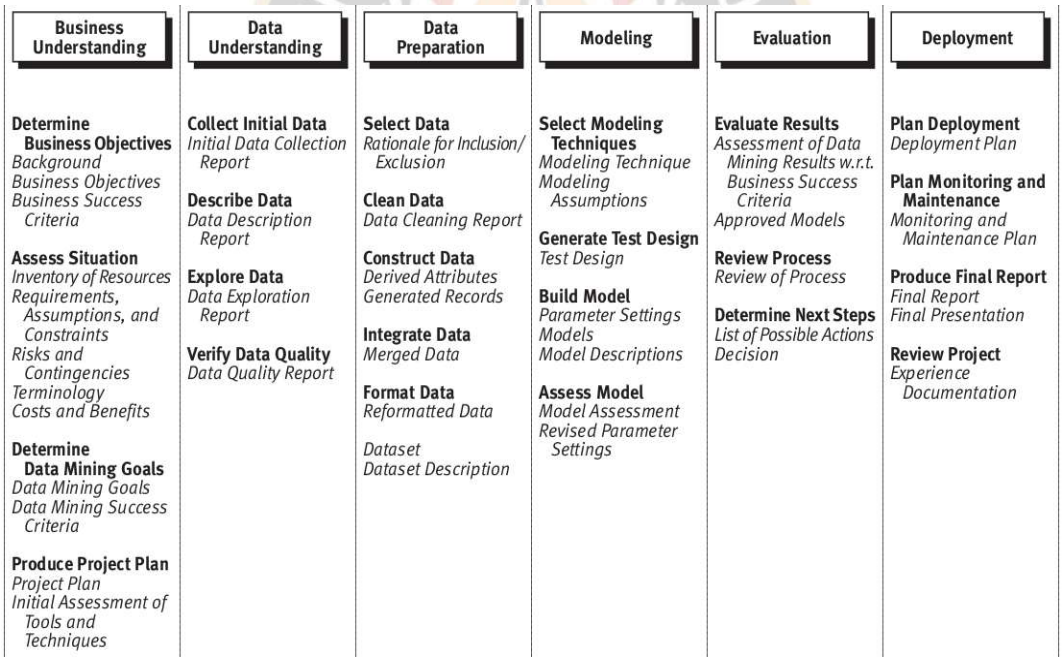
SEKOLAH PASCASARJANA

ada beberapa masalah bisnis penting yang belum ditangani. Pada akhir fase ini, hasil penambangan data harus sesuai dengan tujuan penambangan data yang telah dijabarkan di fase awal. Dalam fase ini ada *evaluate results*, *review process* dan *determine next steps*.

6. Fase Penyebaran (*Deployment Phase*)

   Fase ini tentang pemanfaatan model yang telah dibuat dimana perlu diatur dan disajikan sedemikian rupa sehingga pelanggan dapat menggunakannya, misalnya: personalisasi halaman web secara real-time. Dalam banyak kasus, pelanggan bukanlah seorang data analist sehingga perlu tindakan tambahan agar pelanggan benar-benar memanfaatkan model yang dibuat. Di dalam fase ini terdapat plan, monitoring & maintenance, produce final report dan review project.

Agar lebih jelas penjabaran tiap fase CRISP-DM bisa dilihat pada Gambar 2.6.



Gambar 2.6 Penjabaran Elemen Tiap Fase CRISP-DM (Chapman dkk. 2000)