BAB III

METODOLOGI PENELITIAN

3.1 Gambaran Umum Objek Penelitian

Penelitian ini dilakukan untuk merancang dan mengembangkan sistem intervensi peringatan dini dan pemantauan studi mahasiswa Universitas Multimedia Nusantara berbasis *website* dengan memanfaatkan algoritma *clustering* yakni K-Means, K-Medoids, dan DBSCAN. Sejak didirikan pada tahun 2005 oleh Kompas Gramedia, Universitas Multimedia Nusantara (UMN) adalah salah satu perguruan tinggi swasta Indonesia yang menawarkan program pendidikan unggulan di bidang teknologi informasi dan komunikasi (ICT) [58]. Objek yang menjadi fokus dalam penelitian ini adalah data mahasiswa/i Program Studi Sistem Informasi Universitas Multimedia Nusantara (UMN). Lebih tepatnya, data yang digunakan merupakan data yang diperoleh dari Kartu Hasil Studi (KHS) mahasiswa program studi Sistem Informasi UMN yang berasal dari angkatan 2017 hingga 2023. Batas cakupan data mahasiswa yang berasal dari angkatan 2017 hingga 2023 ditetapkan sesuai dengan relevansi penelitian.

Data mahasiswa tersebut diperoleh dari Biro Informasi Akademik (BIA) UMN yang terdiri dari 16 variabel dan 8718 entri data. Namun, data mahasiswa asli yang dipergunakan dalam penelitian ini tidak memuat variabel nama, Nomor Induk Mahasiswa (NIM), serta informasi kontak berupa alamat *e-mail* mahasiswa, orang tua/wali, dan dosen pembimbing karena alasan kerahasiaan *(confidentiality).* Sehingga, akan dilakukan penambahan data *dummy* untuk melengkapi data yang tidak lengkap dan mendukung objektif penelitian karena sistem akhir yang dihasilkan ditujukan untuk mengirimkan notifikasi peringatan melalui *e-mail.* Selain itu, pada tahap *data preparation*, data tersebut juga akan ditambahkan 3 variabel relevan terkait SKS yang dapat memperkaya informasi pada saat pembentukan *cluster*. Sebelum dijadikan data input pada sistem *website*, data mahasiswa UMN akan diproses terlebih dahulu menggunakan algoritma *clustering*

untuk menghasilkan label klaster mahasiswa yang berpotensi mengalami keterlambatan ataupun kegagalan studi.

3.2 Metode Penelitian

3.2.1 Alur Penelitian

Metode yang digunakan dalam penelitian ini tergolong sebagai jenis metode kuantitatif karena data yang menjadi instrumen penelitian berupa angka dan variabel, dapat diukur, serta analisisnya melibatkan teknik statistik [59]. Metode kuantitatif terbukti efektif dalam mengidentifikasi pola dan hubungan dalam data. Dalam upaya ekstraksi informasi dan penemuan pola tersembunyi pada data mahasiswa UMN, keberhasilan penelitian ini bergantung pada metodologi *data mining* sebagai panduan dan alur penelitian yang terstruktur dan sistematis. Pada praktiknya, terdapat tiga jenis metode *data mining* yang populer digunakan yakni: CRISP-DM *(Cross-Industry Standard Process for Data Mining)*, KDD *(Knowledge Discovery in Databases)*, dan SEMMA *(Sample, Explore, Modify, Model, Assess)*, yang masing-masing memiliki pendekatan unik dengan karakteristik, fungsionalitas, dan sistematika tahapan yang berbeda untuk setiap kebutuhan penelitian. Pemilihan kerangka kerja *data mining* yang akan diimplementasikan sebagai metodologi dan alur pada penelitian ini didasarkan pada hasil perbandingan ketiga kerangka kerja tersebut dan kesesuaiannya dengan prasyarat atau kebutuhan penelitian. Berikut merupakan perbandingan dari ketiga metode *data mining* yang dipertimbangkan sebagai kerangka kerja dalam penelitian ini.

Tabel 3.1 Tabel Perbandingan Kerangka Kerja *Data Mining* [35], [60], [61]

Aspek Perbedaan	**Kerangka Kerja *Data Mining***		
	CRISP-DM	**KDD**	**SEMMA**
Fase/Tahapan	CRISP-DM terdiri dari 6 fase: Pemahaman Bisnis, Pemahaman Data, Persiapan Data, Pembuatan Model, Evaluasi Model, dan Penerapan Model	KDD terdiri dari 5 langkah: Seleksi Data, Pra-pemrosesan Data, Transformasi Data, Penambangan Data, Interpretasi dan Evaluasi	SEMMA terdiri dari 5 fase: Sampling, Eksplorasi, Modifikasi, Pemodelan, dan Penilaian
Fokus	CRISP-DM mencakup keseluruhan proses *data mining*, dari tahap	KDD berfokus pada ekstraksi pengetahuan dari data, dengan fokus	SEMMA berfokus pada pengembangan model prediktif, dengan fokus pada

Aspek Perbedaan	Kerangka Kerja *Data Mining*		
	CRISP-DM	**KDD**	**SEMMA**
	pemahaman bisnis hingga penerapan model	pada algoritma dan teknik	aplikasi dan solusi bisnis
Tingkat Detail	CRISP-DM memiliki tingkat detail yang tergolong baik, dengan panduan dan sub-langkah untuk setiap fase	KDD memiliki tingkat detail yang tergolong kurang karena lebih berfokus pada konsep dan terminologi	SEMMA memiliki tingkat detail yang tergolong menengah dengan fokus pada langkah-langkah praktis
Fleksibilitas	CRISP-DM tergolong fleksibel karena dapat disesuaikan dengan proyek dan kebutuhan spesifik	KDD tergolong cukup fleksibel karena memungkinkan iterasi dan eksplorasi data yang lebih bebas	SEMMA tergolong kurang fleksibel karena ebih terstruktur dan fokus pada pengembangan model
Tingkat Adopsi	CRISP-DM sangat populer dan banyak digunakan di berbagai industri	KDD banyak digunakan oleh komunitas akademis dan penelitian	SEMMA populer digunakan oleh SAS Institute dan beberapa perusahaan lain

Berdasarkan hasil perbandingan tiga kerangka kerja *data mining* yang telah dijabarkan dalam Tabel 3.1 di atas, CRISP-DM dipilih sebagai metodologi dan alur pada penelitian ini berdasarkan beberapa faktor. Pertama, CRISP-DM menawarkan enam fase yang mencakup keseluruhan proses data mining, dari awal pemahaman bisnis hingga akhir penerapan model. Kedua, CRISP-DM menyediakan panduan dan sub-langkah yang detail untuk setiap fase. Detail ini membantu peneliti dalam memahami dan melaksanakan setiap langkah dengan lebih terfokus dan jelas, sehingga meminimalisir risiko kesalahan dan meningkatkan kualitas penelitian. Terakhir, CRISP-DM tergolong fleksibel dan dapat diimplementasikan pada berbagai jenis permasalahan di berbagai industri. Hal ini menjadikannya cocok untuk penelitian ini, yang bertujuan untuk menyelesaikan permasalahan spesifik yakni pengembangan sistem peringatan dini intervensi dan pemantauan dengan lingkup mahasiswa UMN. Berikut merupakan penjelasan dari serangkaian fase yang membangun kerangka kerja CRISP-DM yang telah dipilih.

a. *Business Understanding*

Tahap pemahaman bisnis merupakan tahapan pertama dalam CRISP-DM yang bertujuan untuk memahami latar belakang masalah, kebutuhan, dan tujuan pelaksanaan penelitian. Dalam penelitian ini,

permasalahan yang terjadi adalah sistem akademik yang digunakan untuk memantau dan memberikan intervensi pada mahasiswa UMN yang mengalami masalah studi masih terbilang manual. Berdasarkan permasalahan ini, kebutuhan dan tujuan dilangsungkannya penelitian ini adalah untuk mengembangkan sistem intervensi peringatan dini dan pemantauan studi mahasiswa/i UMN dengan memanfaatkan algoritma *clustering* berbasis *website* agar keseluruhan prosesnya dapat berjalan otomatis dan lebih efisien. Tahap ini juga melibatkan pihak-pihak UMN terkait seperti staf BIA maupun teknisi IT universitas dalam proses diskusi pra-penelitian untuk mencegah kesalahpahaman yang mungkin timbul saat perancangan dan pengembangan sistem dengan mendefinisikan lingkup dan objektif proyek secara spesifik.

b. *Data Understanding*

Tahapan pemahaman data merupakan tahapan kedua dalam CRISP-DM yang bertujuan untuk memahami karakteristik dan menilai kualitas data dengan menelaah spesifikasi data yang telah dikumpulkan. Penelitian ini memanfaatkan data mahasiswa UMN sebagai data primer yang bersumber dari Biro Informasi Akademik (BIA). Lebih tepatnya, data yang akan digunakan adalah data hasil studi mahasiswa/i program studi Sistem Informasi angkatan 2017-2023 yang dikompilasi dalam bentuk tabel. Data ini akan digunakan sebagai data pelatihan pada model maupun data pengujian atau *input* pada sistem EIWMS yang akan dikembangkan. Tahapan ini dilakukan untuk memastikan ketepatan format dan kelengkapan variabel data yang dibutuhkan, serta mengeksplorasi distribusi maupun informasi tersembunyi dalam data melalui kegiatan *Exploratory Data Analysis* (EDA).

c. *Data Preparation*

Tahapan persiapan data merupakan tahapan ketiga dalam CRISP-DM yang bertujuan untuk melakukan pra-pemrosesan terhadap data mentah untuk memastikan kesiapan data sebelum pembentukan model. Hasil dan kualitas dari model akhir yang dikembangkan bergantung pada

proses dari tahapan ini. Kegiatan yang akan dilakukan pada tahap ini meliputi pembersihan, transformasi, dan pemilihan fitur data. Pembersihan data dilakukan dengan mengisi *missing value* atau nilai *null* pada data, serta menghilangkan data redundan. Kegiatan transformasi dilakukan untuk menyelaraskan format data dengan mengubah dan mengonversi nama ataupun entri nilai variabel yang perlu disesuaikan ataupun pengubahan struktur bentuk data *(pivot)* pada variabel data agar dapat diproses oleh algoritma *clustering* nantinya. Selain itu, juga dilakukan penambahan kolom atau variabel pada data untuk menunjang kelengkapan informasi dalam pembentukan model dan pengembangan sistem. Pemilihan fitur data dilakukan untuk memilah atribut atau variabel data mahasiswa yang memang relevan dengan objektif penelitian sebelum nantinya diproses dengan algoritma *clustering*.

d. *Modeling*

Tahapan pemodelan merupakan tahapan keempat dalam CRISP-DM yang bertujuan untuk membangun model *machine learning* dari data yang telah diproses sebelumnya untuk mencapai tujuan penelitian. Teknik pemodelan yang digunakan dalam penelitian ini adalah model *clustering* dengan algoritma K-Means, K-Medoids, dan DBSCAN untuk mengidentifikasi *cluster* atau kelompok mahasiswa berdasarkan hasil studi mereka. Tahapan ini dilaksanakan menggunakan *tools* Python sebagai bahasa pemrograman dan Visual Studio Code sebagai perangkat lunaknya. Sebelum dilakukan tahap pembentukan model, data mahasiswa akan melalui proses *splitting data* yakni pembagian data menjadi dua set *training* dan *testing.* Rasio pembagian set data yang akan diterapkan pada penelitian ini adalah 80:20 (80% untuk set pelatihan dan 20% untuk set pengujian) karena umum digunakan dan terbukti dapat memberikan hasil model yang optimal pada penelitian [62]. Set yang akan digunakan pada tahap pemodelan dan evaluasi adalah set data pelatihan *(training)*, sedangkan set data pengujian

(testing) akan digunakan pada tahap *deployment* nanti. Tahapan pemodelan ini dapat dilakukan secara iteratif atau diulang agar dapat menghasilkan model terbaik dengan melakukan kalibrasi atau penyesuaian parameter model.

e. *Evaluation*

Tahapan evaluasi merupakan tahapan kelima dalam CRISP-DM yang bertujuan untuk menilai kualitas dan mengukur performa dari model *clustering* yang dihasilkan pada tahap sebelumnya. Kedua model *clustering* yang dikembangkan menggunakan algoritma K-Means, K-Medoids, dan DBSCAN akan dievaluasi performanya untuk mengidentifikasi model dengan performa terbaik. Beberapa metrik validasi yang digunakan sebagai indikator pembanding performa model *clustering* pada penelitian ini yakni *Silhouette Score* (SH), *Davies-Bouldin Index* (DBI), *Calinski-Harabasz Index* (CHI), dan *Dunn Index* (DI). Metrik evaluasi ini digunakan untuk menilai performa algoritma clustering dalam mengelompokkan data mahasiswa, yaitu dengan memperhatikan pemisahan jarak antar klaster yang terdefinisi dengan baik dan jarak antar anggota klaster yang erat. Kriteria pemilihan model akhir yang akan diimplementasikan pada sistem EIWMS adalah model dengan skor SH yang mendekati 1 positif, indeks DBI yang rendah, serta nilai CHI dan DI yang tinggi. Setelah proses pemilihan model selesai, model yang terpilih kemudian akan diterapkan untuk mengidentifikasi klaster mahasiswa pada set data pengujian.

f. *Deployment*

Tahapan *deployment* merupakan tahapan keenam dan terakhir dalam CRISP-DM yang bertujuan untuk mengimplementasikan hasil model dalam upaya diseminasi (penyebaran) informasi kepada pengguna akhir guna penyelesaian permasalahan yang telah diidentifikasi sebelumnya. Strategi *deployment* yakni penerapan model yang dipilih dalam penelitian ini adalah pengembangan sistem EIWMS berupa antarmuka *website* sederhana yang ditujukan untuk *role* admin.

Sistem EIWMS berbasis *website* yang akan dikembangkan pada penelitian ini memiliki fungsi utama untuk menampilkan data mahasiswa dan informasi label kelompok mahasiswa berdasarkan hasil algoritma *clustering*, serta fungsi pengiriman intervensi peringatan dini berupa notifikasi dalam bentuk pesan elektronik *(e-mail)*. Tahap *deployment* diawali dengan menerapkan hasil model *clustering* terbaik untuk menentukan kelompok *(cluster)* mahasiswa pada set data pengujian *(testing)*. Selanjutnya, label hasil *clustering* mahasiswa dan beberapa data *dummy* pendukung juga akan ditambahkan sebagai kolom variabel baru pada set data *testing* sebelum diekspor dalam bentuk *file* berekstensi .xlsx atau .csv, untuk kemudian diimpor pada basis data sistem yakni MySQL.

Proses pengembangan *website* dilangsungkan dengan memanfaatkan *framework* Laravel dalam pengkodean *website* dan *software* Visual Studio Code. Terakhir, tahapan ini ditutup dengan pelaksanaan *User Acceptance Test* (UAT) untuk memastikan bahwa sistem EIWMS yang telah dikembangkan sudah memenuhi objektif awal penelitian dan memberikan hasil yang diinginkan [63]. Kegiatan UAT dilakukan dengan metode survei berupa kuesioner singkat yang ditujukan kepada para pengguna akhir sistem untuk dapat diisi dalam satuan terukur yakni skala Likert. Skala Likert yang digunakan pada penelitian ini terdiri dari 5 skala poin dan berfungsi untuk mengukur tanggapan pengguna terhadap hasil pengembangan sistem secara kuantitatif [64]. Setelah melewati fase UAT untuk memastikan bahwa sistem dapat berfungsi dengan baik dan tidak memiliki masalah, sistem EIWMS sudah siap untuk digunakan ataupun diintegrasikan pada sistem akademik UMN.

3.2.2 Metode *Data Mining*

Sebelum data dimasukkan sebagai *input* pada aplikasi untuk melaksanakan fungsionalitas peringatan dan pemantauan, data mahasiswa terlebih dahulu diolah untuk menentukan klaster siswa. Pembentukan klaster ini dilakukan

dengan memanfaatkan metode *clustering* yang termasuk dalam algoritma *unsupervised learning*. Penerapan metode *clustering* ini ditujukan untuk mengidentifikasi struktur, pola, ataupun model tersembuyi yang terdapat pada data yang tidak berlabel [39], seperti data mahasiswa UMN. Data yang telah dikelompokkan menjadi beberapa klaster kemudian akan diberikan label *output* untuk membedakan antara mahasiswa yang berpotensi mengalami keterlambatan studi dan memerlukan peringatan dengan mahasiswa yang aman. Hasil dari pemrosesan data ini kemudian menjadi parameter utama pemberian peringatan sistem kepada mahasiswa yang berpotensi mengalami masalah dalam menjalani masa studi.

Penelitian ini mempertimbangkan penggunaan tiga jenis algoritma *clustering* yang populer digunakan karena kesederhanaan modelnya sehingga mudah diinterpretasikan, yakni K-Means, K-Medoids, dan *Density Based Spatial Clustering of Applications with Noise* (DBSCAN). Berikut merupakan tabel perbandingan dari ketiga algoritma *clustering* tersebut.

Tabel 3.2 Tabel Perbandingan Algoritma *Clustering* [41], [65]–[67]

Aspek Perbedaan	**Algoritma *Clustering***		
	K-Means	**K-Medoids**	**DBSCAN**
Jenis *clustering*	*Partitioning-based clustering*	*Partitioning-based clustering*	*Density-based clustering*
Parameter model	Nilai parameter K dalam model K-Means sebagai *centroid* dapat ditentukan secara optimal dengan metode *Elbow*	Nilai parameter K dalam model K-Medoids sebagai *medoids* dapat ditentukan secara optimal dengan berbagai metode, seperti *Elbow Method* maupun *Silhouette Score*	Nilai parameter epsilon (Eps) sebagai jarak maksimum klaster dalam model DBSCAN tidak dapat ditentukan secara optimal tanpa memodifikasi model dasar DBSCAN
Skalabilitas	Algoritma K-Means tergolong efisien sehingga dapat digunakan untuk menangani data berukuran besar dan memiliki skalabilitas sedang	Algoritma K-Medoids tergolong efisien sehingga dapat digunakan untuk menangani data berukuran besar dan memiliki skalabilitas yang tergantung pada metode pencarian	Algoritma DBSCAN dapat digunakan untuk menangani data dalam ukuran apapun, namun kurang efisien untuk menangani data berukuran besar

Aspek Perbedaan	Algoritma *Clustering*		
	K-Means	**K-Medoids**	**DBSCAN**
		medoids yang digunakan	
Sensitivitas terhadap *noise* atau *outlier*	Algoritma K-Means sensitif terhadap *outlier* dalam data	Algoritma K-Medoids sensitif terhadap *noise* dalam data	Algoritma DBSCAN kurang sensitif terhadap *outlier* dalam data
Bentuk	Algoritma K-Means menghasilkan klaster dengan pembagian yang jelas dan cenderung berbentuk lingkaran *(spherical-shaped clusters)*	Algoritma K-Medoids menghasilkan klaster dengan pembagian yang jelas dan cenderung berbentuk sesuai dengan distribusi data	Algoritma DBSCAN dapat menghasilkan klaster dengan pembagian yang acak dan berbentuk garis lengkung ataupun menyebar *(concave / arbitrary-shaped clusters)*
Kompleksitas waktu	Algoritma K-Means tidak membutuhkan waktu yang lama karena kompleksitas waktunya yang rendah yakni O(knt)	Algoritma K-Medoids memiliki kompleksitas waktu yang serupa dengan algoritma K-Means yakni O(knt), sehingga tidak membutuhkan waktu yang lama	Algoritma DBSCAN cukup membutuhkan waktu yang lama karena kompleksitas waktunya yang sedang yakni O(n*logn)

Seperti yang terlihat pada tabel perbandingan kedua algoritma *clustering* di atas, terdapat perbedaan signifikan antara algoritma clustering K-Means, K-Medoids, dan DBSCAN dalam beberapa aspek seperti jenis *clustering*, parameter model, skalabilitas, sensitivitas terhadap *noise* atau *outlier*, bentuk klaster yang dihasilkan, serta kompleksitas waktu. K-Means dan K-Medoids merupakan algoritma *clustering* berbasis partisi yang efektif dalam menangani data dengan pembagian klaster yang jelas. K-Means memungkinkan optimasi parameter model (nilai K) melalui metode *Elbow*, sedangkan K-Medoids menawarkan fleksibilitas dengan berbagai metode optimasi, seperti *Elbow* dan *Silhouette Score*. Kelemahan kedua algoritma ini terletak pada sensitivitasnya terhadap *noise* dan *outlier*. Sementara itu, DBSCAN memiliki kelebihan sebagai algoritma berbasis kepadatan yang tangguh terhadap *noise* atau *outlier*, tetapi tantangannya terletak pada kesulitan menentukan parameter epsilon (Eps) secara optimal tanpa modifikasi pada model konvensionalnya. Ketiga algoritma *clustering* tersebut akan diterapkan dalam fase pemodelan untuk membentuk klaster mahasiswa. Selanjutnya, performa model akan dievaluasi

dengan tujuan memilih model yang menunjukkan kinerja terbaik dan paling sesuai dengan karakteristik data atau tujuan penelitian. Evaluasi kinerja model akan dilakukan menggunakan metrik seperti *Silhouette Score* (SH), *Davies-Bouldin Index* (DBI), *Calinski-Harabasz Index* (CHI), dan *Dunn Index* (DI) guna meneliti keefektifan kedua algoritma dalam pembentukan model *cluster*.

3.3 Teknik Pengumpulan Data

Teknik pengumpulan data yang diterapkan pada penelitian ini tergolong sebagai teknik pengumpulan data primer. Data primer merujuk pada informasi yang diperoleh secara langsung dari sumber melalui metode seperti survei, wawancara, eksperimen, atau observasi yang dilakukan sesuai dengan kebutuhan penelitian tertentu [68]. Teknik pengumpulan data ini digunakan untuk memperoleh data hasil studi mahasiswa Universitas Multimedia Nusantara (UMN) yang bersumber dari Biro Informasi Akademik (BIA) dalam bentuk *file* Excel. BIA merupakan suatu biro atau departemen di UMN yang berperan sebagai penyedia data, pengolah informasi, dan pusat layanan administratif di bidang akademik bagi mahasiswa dan dosen Universitas Multimedia Nusantara. Selain itu, data primer lainnya yang dikumpulkan dengan teknik ini adalah data kepuasan pengguna dari hasil UAT sebagai hasil evaluasi sistem EWS yang telah dikembangkan.

3.3.1 Populasi dan Sampel

Teknik pengambilan sampel melibatkan aktivitas pemilihan *subset* atau bagian dari populasi yang menjadi fokus penelitian. Teknik ini digunakan untuk mengumpulkan data yang dapat merepresentasikan karakteristik suatu populasi secara cepat dan efisien, dibanding melibatkan keseluruhan populasi [69]. Teknik pengambilan sampel data mahasiswa yang digunakan dalam penelitian termasuk ke dalam jenis *non-probability sampling* yakni *purposive sampling*. Dalam penarikan sampel non-probabilitas, elemen-elemen populasi tidak memiliki probabilitas seleksi yang diketahui atau sama. Artinya, tidak semua elemen populasi memiliki peluang yang setara untuk dipilih. *Purposive sampling* digunakan untuk memilih sampel yang cenderung memberikan informasi relevan, dengan fokus pada karakteristik yang diinginkan dalam

penelitian. Meskipun tidak merepresentasikan populasi secara keseluruhan, metode ini bertujuan memasukkan sampel individu-individu yang dapat efektif mengoptimalkan sumber daya penelitian [70].

Teknik pengambilan sampel *purposive sampling* diimplementasikan pada kedua data jenis data primer. Sampel data primer pertama yang dimaksud adalah adalah data hasil studi mahasiswa program studi Sistem Informasi UMN angkatan 2017-2023. Alasan pemilihan program studi dan rentang data dari sampel tersebut didasarkan pada kesesuaian antara lingkup, batasan masalah, dan objek penelitian, serta relevansi kebaruan data yang digunakan dalam penelitian ini. Rentang awal data yakni angkatan 2017 dipilih karena merupakan angkatan terakhir yang dapat lulus pada tahun 2024, sementara rentang akhir yakni angkatan 2023 dipilih karena merupakan angkatan terbaru yang tercatat dalam basis data mahasiswa universitas pada awal tahun 2024. Sampel data primer ini diharapkan dapat merepresentasikan populasi data yakni seluruh mahasiswa/i Universitas Multimedia Nusantara. Sampel data lainnya yang dikumpulkan adalah hasil evaluasi sistem berupa kuesioner yang diisi oleh representatif admin atau staf UMN yang akan menjadi pengguna akhir sistem EIWMS. Populasi yang menjadi fokus sampel data primer ini mencakup seluruh admin atau staf yang tergabung dalam departemen BIA atau *Student Services* UMN sebagai sasaran *end-users* dari penelitian ini.

3.3.2 Periode Pengambilan Data

Periode pengumpulan data dimulai pada bulan Februari 2024 dengan mengirimkan surat permohonan resmi kepada Fakultas Teknik Informatika dan Departemen Biro Informasi Akademik (BIA) sesuai dengan yang tertera dalam Lampiran C. Setelahnya, Departemen BIA merespons permohonan tersebut dengan mengirimkan surat konfirmasi yang memperbolehkan pengambilan data untuk tujuan penelitian, seperti yang terlampir dalam Lampiran D, bersama dengan data yang diminta. Proses tersebut dilanjutkan dengan melakukan pertemuan dengan perwakilan dari BIA, pihak fakultas, dan tim IT UMN. Setelah menyelaraskan tujuan penelitian dengan kebutuhan pihak

universitas, data mahasiswa yang terkompilasi dalam bentuk tabel mulai diolah sebagai sistem EWS. Selain itu, variabel data mahasiswa lainnya yang diperlukan namun tidak disediakan melalui BIA seperti NIM, nama, maupun alamat *e-mail* karena kerahasiaan data, didapatkan melalui pembuatan data *dummy* dengan memanfaatkan pustaka Python sebagai alat pengolah data. Data lainnya yang dikumpulkan adalah data kuesioner mengenai evaluasi hasil akhir sistem EWS, yang disebarkan pada bulan April 2024.

3.4 Variabel Penelitian

Seluruh variabel yang terdapat dalam dataset mahasiswa UMN dapat menjadi parameter yang memengaruhi hasil pengelompokkan *(clustering)* data mahasiswa. Penelitian ini melibatkan 2 (dua) jenis variabel dalam mengelompokkan data mahasiswa UMN, yakni variabel independen sebagai variabel bebas dan dependen sebagai variabel terikat.

3.4.1 Variabel Independen

Variabel independen atau variabel bebas (biasanya dilambangkan dengan X) adalah jenis variabel yang berperan sebagai prediktor dan memengaruhi perubahan nilai variabel lainnya (variabel dependen) [71]. Variabel yang menjadi penentu hasil klasterisasi mahasiswa dalam penelitian ini dipilih berdasarkan indikator yang sebelumnya telah digunakan oleh departemen BIA UMN untuk mengidentifikasi tingkat kemajuan studi siswa dalam sistem semi-manual yang digunakan. Variabel-variabel yang menjadi variabel independen ini berupa variabel numerik, meliputi:

1) Masa studi semester yakni jumlah semester kumulatif yang telah ditempuh;
2) Jenjang semester yang sedang ditempuh saat ini;
3) Besaran IPS yang diperoleh dari semester 1 (satu) hingga saat ini;
4) Besaran IPK yang diperoleh dari semester 1 (satu) hingga saat ini;
5) Jumlah minimal Satuan Kredit Semester (SKS) yang harus diambil pada semester tertentu;

6) Jumlah minimal SKS kumulatif yang harus sudah diambil dan lulus pada semester tertentu;
7) Jumlah prasyarat SKS kelulusan pada program studi tersebut;
8) Jumlah SKS yang diambil pada semester tertentu;
9) Jumlah SKS yang berhasil ditempuh (lulus) pada semester tertentu;
10) Jumlah SKS mata kuliah yang perlu diulang (tidak lulus) pada semester tertentu;
11) Total kumulatif SKS yang diambil oleh masing-masing mahasiswa;
12) Total kumulatif SKS yang diambil dan berhasil ditempuh (lulus) oleh masing-masing mahasiswa;
13) Sisa SKS yang harus diambil sebagai syarat kelulusan oleh masing-masing mahasiswa.

3.4.2 Variabel Dependen

Variabel dependen atau variabel terikat (biasanya dilambangkan dengan Y) merupakan jenis variabel yang berperan sebagai respons dan nilainya akan bergantung atau dipengaruhi oleh variabel lain (variabel independen) [71]. Variabel yang menjadi variabel dependen dalam penelitian ini adalah variabel kategorikal berupa label hasil *cluster* mahasiswa. Label *cluster* mahasiswa yang diperoleh sebagai hasil model *clustering* membagi mahasiswa menjadi beberapa golongan berdasarkan perkembangan studi mereka, yakni mahasiswa dengan kemajuan studi yang sesuai *(on-track)* atau mereka yang berisiko mengalami keterlambatan studi *(at-risk* atau *late)* berdasarkan variabel independen yang telah dijelaskan sebelumnya.

3.5 Teknik Analisis Data

Proses pengolahan dan analisis data pada penelitian ini dilakukan menurut metodologi penelitian yang telah ditentukan sebelumnya, yakni kerangka kerja CRISP-DM dan implementasi algoritma *machine learning*. Algoritma *machine learning* yang digunakan dalam penelitian ini tergolong ke dalam jenis *unsupervised learning* yakni algoritma *clustering* K-Means, K-Medoids, dan DBSCAN. Algoritma tersebut dimanfaatkan pada tahapan pembentukan model dari

data mahasiswa sebelum dapat dijadikan *input* pada sistem EIWMS berupa *website* yang akan dikembangkan. Hasil dari tahapan pembentukan model mahasiswa adalah label *cluster* atau kelompok mahasiswa yang dibagi berdasarkan kemajuan studi mereka. Label kelompok mahasiswa tersebut berfungsi sebagai penanda dalam membedakan antara mahasiswa yang berpotensi mengalami keterlambatan atau kegagalan studi dan yang tidak.

Keseluruhan proses dan teknik dalam menganalisis data membutuhkan bantuan alat *(tools) data mining* dan *data analysis.* Sebagai pertimbangan pemilihan alat pengolahan data, dilakukan suatu komparasi antara 2 (dua) jenis bahasa pemrograman sebagai *tools* yang umum digunakan dalam penerapan *machine learning* karena menyediakan *library* yang sesuai, yakni Python dan R.

Tabel 3.3 Tabel Perbandingan *Tools* Pengolahan Data [72]–[76]

Aspek Perbedaan	***Tools***	
	Python	**R**
Definisi & Tujuan	Python merupakan bahasa pemrograman umum yang digunakan untuk analisis data dan komputasi ilmiah	R merupakan bahasa pemrograman statistik, yang mendukung analisis maupun visualisasi data dengan komputasi dan grafik
Sintaks & Tingkat Kesulitan	Python memiliki sintaks sederhana dan kurva pembelajaran yang linier sehingga mudah dipelajari	R memiliki sintaks yang relatif kompleks dan kurva pembelajaran yang lebih curam dan rumit di awal
Objektif penggunaan	Python cocok digunakan untuk membuat model baru dalam *machine learning* dan *deep learning.* Serta, pengembangan aplikasi GUI, *web,* dan *embedded system*.	R cocok digunakan dalam pembelajaran dan eksplorasi data statistik dengan *library* yang telah tersedia
Cakupan penggunaan	Python sering digunakan dalam proyek *data science* serta pengembangan *web, scripting,* maupun tugas otomasi.	R sering digunakan dalam proses analisis statistik data yang kompleks
Kinerja	Python memiliki kinerja yang umumnya lebih cepat untuk tugas komputasi intensif maupun tugas yang melibatkan *dataset* besar	R memiliki kinerja lebih lambat untuk perhitungan statistik yang kompleks
Integrasi	Python dapat diintegrasikan untuk berjalan *(run)* pada berbagai aplikasi maupun platform *website*	R telah diintegrasikan secara *default* untuk hanya dapat berjalan pada perangkat lokal
Basis Pengguna	Python lebih populer dan memiliki basis pengguna yang luas. Pengguna utama Python meliputi *developer* dan *programmer*	R kurang populer digunakan dan basis pengguna utamanya mencakup para ilmuan dan professional data di bidang R&D yang mengandalkan analisis data

Berdasarkan tabel perbandingan kedua *tools* di atas, bahasa pemrograman yang dipilih untuk menjalankan algoritma *clustering* pada penelitian ini adalah Python.

Python dipilih karena beberapa karakteristiknya yakni: 1) memiliki sintaks yang sederhana, 2) cocok untuk membuat model *machine learning*, 3) cocok diimplementasikan pada proyek *data science* maupun pengembangan *web*, 4) memiliki kinerja yang baik untuk data besar, 5) dapat diintegrasikan dengan berbagai platform khususnya *website*, 6) serta sering digunakan oleh para pengembang dan pemrogram. Sintaks Python yang sederhana dan kinerjanya yang baik dalam mengolah data besar tentunya dapat mempermudah pembuatan model *cluster* mahasiswa dan menghemat waktu pemrosesan data. Python juga sudah sering digunakan dalam berbagai proyek yang melibatkan data ataupun pengembangan sistem karena sifatnya yang mudah diintegrasikan. Bahasa pemrograman ini akan diterapkan pada tahapan pengolahan data, pembentukan model, dan evaluasi model dalam *framework* CRISP-DM.

Hasil akhir dari penelitian ini adalah sistem intervensi peringatan dini dan pemantauan studi mahasiswa yang akan dikembangkan pada fase *deployment* CRISP-DM dan berbasis *website*. Platform *website* dipilih sebagai hasil sistem akhir pada penelitian ini karena sebelumnya BIA UMN sudah memiliki sistem akademik untuk merekap dan mengolah data mahasiswa berbasis *website*. Kesamaan antara platform sistem yang telah ada dan sistem EIWMS yang baru ini diharapkan dapat mempermudah proses integrasi kedua sistem kelak. Selain itu, platform *website* terpilih karena beberapa kelebihannya dibanding platform lain yakni: 1) aksesibilitas pada berbagai perangkat dan tidak memerlukan proses instalasi aplikasi apapun, 2) kemudahan penggunaan dan navigasi, 3) fleksibilitas integrasi dengan beragam sistem dan layanan seperti basis data maupun pengiriman *e-mail*, serta 4) memiliki beragam opsi dari segi pemeliharaan, pembaruan, keamanan, dan skalabilitas yang dapat disesuaikan dengan kebutuhan dan biaya pengembangan.

Framework pengembangan sistem EIWMS berbasis *website* yang digunakan dalam penelitian ini adalah Laravel karena kepraktisan implementasinya dibanding *framework* konvensional seperti HTML, PHP, dan CSS. Penggunaan Laravel juga diharapkan dapat meningkatkan efisiensi dan kemudahan pengembangan sistem.

Beberapa alasan lain pemilihan Laravel yaitu: bersifat *open source*, ketersediaan berbagai *plugin*, ekstensi, dan pustaka pemrograman, kemampuan integrasi dengan layanan basis data yang akan digunakan yakni MySQL, serta skalabilitas implementasinya yang baik. Pada implementasinya, *framework* Laravel dan bahasa pemrograman Python yang telah terpilih sebelumnya akan dioperasikan pada *software* Visual Studio Code karena telah menyediakan ekstensi Python dan Laravel dalam satu perangkat lunak yang sama. Selain itu, VS Code memungkinkan integrasi yang mudah antar bahasa pemrograman dan kerangka kerja, memiliki ukuran yang ringan dan performa yang cepat, serta fleksibel untuk dijalankan pada berbagai platform dengan kebutuhan pengembangan spesifik.