BAB II
TINJAUAN PUSTAKA

2.1 Penelitian Yang Relevan

Pada bab tinjauan Pustaka ini akan ditambahkan beberapa penelitian terdahulu yang berkaitan dengan masalah tersebut untuk dibahas guna membandingkan dan membimbing untuk memahami dan merancang suatu metode yang akan digunakan:

Penelitian pertama, **"Perbandingan Akurasi Algoritma C4.5 Dan KNN Untuk Prediksi Kelulusan Mahasiswa Penerima Beasiswa"** (*Andrian Fakih et al., 2025*) Hasil dari Penelitian yang telah dilakukan untuk mengembangkan skema penerimaan beasiswa yang optimal dan pemerataan bagi universitas. Metode yang digunakan bisa diterapkan dengan harapan dapat membantu administrator universitas agar dapat menyeleksi dengan mudah. Metode tradisional yang digunakan sebelumnya memakan banyak waktu. Berdasarkan penelitian yang telah dilakukan dapat disimpulkan bahwasanya algoritma KNN memiliki performansi yang lebih baik dibanding algoritma *Decision Tree* c4.5 dikarenakan hasil yang didapatkan oleh algoritma KNN dengan tingkat akurasi 74.95%, presisi sebesar 20.9%, *recall* sebesar 77.10% dan AUC 0.752, sedangkan jika menggunakan algoritma C4.5 tingkat akurasi hanya mendapatkan 71.79%, presisi sebesar 39.0%, *recall* sebesar 45.30% dengan hasil AUC 0.734.

Penelitian Kedua, **"Algoritma KNN, Naive Bayes, SVM Dan *Decision Tree* Dalam Menentukan Kelulusan Mahasiswa"** (*Teuku Afriliansyah et al., 2023*) Hasil temuan kami dengan memprediksi tingkat kelulusan mahasiswa dengan menggunakan lima model algoritma didapatkan hasil terbaik atau hasil yang paling akurat dalam melakukan prediksi yaitu dengan model algoritma KNN dengan melakukan optimasi pada data tingkat akurasi yang didapatkan sebesar 96,95% termasuk kedalam kategori sangat baik. Selain itu hasil prediksi yang didapatkan lebih dominan kepada mahasiswa yang terlambat dibandingkan dengan mahasiswa yang lulus tepat waktu, Adapun faktor

keterlambatan mahasiswa dipengaruhi oleh Indeks Prestasi Komulatif (IPK). Oleh karena itu dengan mengetahui hasil prediksi tingkat kelulusan mahasiswa dengan menggunakan model algoritma terbaik diharapkan mampu memberikan nilai maksimal dalam melakuan evaluasi akademik.

Penelitian Ketiga, **"Perbandingan Metode Data Mining Modelklasifikasi Naive Bayes, Decission Tree Dan K-Nearest Neighbour Dalam Memprediksi Ketepatan Kelulusan Mahasiswa Prodi Teknik Informatika Di Universitas Pamulang"** (*Ichsan Ramhdani et al., 2023*) Hasil dari penelitian Data mining dengan menggunakan model klasifikasi naive bayes adalah cara yang paling baik karena memiliki nilai akurasi yang lebih baik daripada decission tree dan alogoritma k-NN. Model yang dihasilkan oleh naive bayes cukup baik (AUC=0,726) digunakan untuk memprediksi kelulusan mahasiswa prodi teknik informatika di Universitas Pamulang. Nilai akurasinya 75,78% dalam memprediksi ketepatan waktu lulus. Atribut kelompok nilai IPK lebih berpengaruh dalam memprediksi kelulusan tepat waktu daripada atribut jenis kelamin.

Penelitian Keempat, **"Penerapan Algoritma *Decision Tree* Dalam Klasifikasi Data Prediksi Kelulusan Mahasiswa"** (*M Riski Qisthiano et al., 2023*) Hasil akhir penelitian yang sudah dijalankan oleh peneliti pada penerapan algoritma *Decision Tree* terhadap data klasifikasi terhadap penggunaan data prediksi mahasiswa yang dapat lulus tepat waktu atau tidak dengan memakai data alumni sebagai landasan tolak ukur kelulusan setiap mahaiswa apakah tepat waktu atau sebaliknya, sampel data berasal dari data alumni yang telah dikumpulkan dari beberapa perguruan tinggi di Kota Palembang. Berdasarkan pengujian didapat hasil akurasi tertinggi yaitu sebesar 0.8739, atau sebesar 87.93% dengan pembagian data 90% data *training*, dan 10% data *test*ing yang digunakan dari 1739 record data dengan menggunakan metode *Decision Tree* menghasilkan akurasi sebesar 87.93%.

Penelitian Kelima, **"Analisis Perbandingan *Decision Tree* C4.5 Dan KNN Dalam Perizinan Bongkar Muatan Kapal"** (*Naurah Nafizah et al., 2023*) Hasil akhir

penelitian yang sudah dijalankan oleh peneliti pada penerapan algoritma *Decision Tree* terhadap data klasifikasi terhadap penggunaan data prediksi mahasiswa yang dapat lulus tepat waktu atau tidak dengan memakai data alumni sebagai landasan tolak ukur kelulusan setiap mahaiswa apakah tepat waktu atau sebaliknya, sampel data berasal dari data alumni yang telah dikumpulkan dari beberapa perguruan tinggi di Kota Palembang. Berdasarkan pengujian didapat hasil akurasi tertinggi yaitu sebesar 0.8739, atau sebesar 87.93% dengan pembagian data 90% data *training*, dan 10% data *test*ing yang digunakan dari 1739 record data dengan menggunakan metode *Decision Tree* menghasilkan akurasi sebesar 87.93%.

Penelitian Keenam, **"Prediksi Kelulusan Mahasiswa Tepat Waktu Menggunakan Metode Naïve Bayes Dan *Decision Tree* Pada Universitas Stella Maris Sumba"** (*Julianti Suwartini et al., 2024*) Hasil dari penelitian ini dengan Atribut yang digunakan untuk Klasifikasi Data Mining terdiri atas 13 atribut yaitu NIM, Jenis Kelamin, Status Mahasiswa, Umur, Indeks Prestasi Semester 1, Indeks Prestasi Semester 2, Indeks Prestasi Semester 3, Indeks Prestasi Semester 4, Indeks PrestasiSemester 5, Indeks Prestasi Semester 6, Indeks Prestasi Semester 7, Indeks Prestasi Semester 8, Indek Prestasi Komulatif, Keterangan Sebagai atribut hasil. Dari hasil Atribut yang digunakan untuk Klasifikasi Data Mining Prestasi Semester 6, Indeks Prestasi Semester 7, Indeks Prestasi Semester 8, Indek Prestasi Komulatif, Keterangan Sebagai atribut hasil. Dari hasil proses pengujian dengan tools RapidMiner Menggunakan dua metode yang telah dilakukan. Decision Tree(C4.5) memperoleh hasil akurasi sebesar 70.18% dan Metode Naïve Bayes memperoleh hasil akurasi tertinggi sebesar 71.24%.

Penelitian Ketujuh, **"Klasifikasi Algoritma K-Nearest Neighbor, Naive Bayes, *Decision Tree* Untuk Prediksi Status Kelulusan Mahasiswa S1"** (*Enggar Novianto et al., 2023*) Pada penelitian ini untuk memprediksi kelulusan mahasiswa dengan menggunakan 3 metode yaitu K-NN, Naïve Bayes dan Decision Tree. Atribut yang digunakan untuk klasifikasi data mining terdiri dari NIM, Jenis Kelamin, IPS1, IPS2,

IPS3, IPS4, IPS5, IPS6, IPS7, IPK dan status. Dari hasil proses pengujian dan prediksi dengan aplikasi RapidMiner menggunakan tiga metode yang telah dilakukan. Metode KNerest Neighbor (KNN) memperoleh hasil akurasi sebesar 96.67%, pada pengujian prediksi dengan metode Naïve Bayes memperoleh nilai akurasi sebesar 77.33%, sedangkan metode *Decision Tree* memperoleh hasil akurasi sebesar 94.00%. Sehingga metode K-NN menjadi metode terbaik dalam klasifikasi komparasi dalam memprediksi kelulusan mahasiswa tepat waktu dengan prediksi nilai akurasi 96.67%. Oleh karena itu, disarankan pada penelitian selanjutnya untuk membuat perbandingan antara metode klasifikasi dan metode clustering untuk mengetahui nilai akurasi dari penggunaan kedua metode algoritma tersebut.

Penelitian Kedelapan **"PREDIKSI PREDIKAT KELULUSAN MAHASISWA MENGGUNAKAN NAIVE BAYES DAN *DECISION TREE* PADA UNIVERSITAS XYZ"** (*Agung Wibowo et al., 2022*) Pada penelitian ini untuk memprediksi kelulusan mahasiswa dengan menggunakan perbandingan metode yaitu Naïve Bayes dan Decission Tree, data yang dipakai disini diambil dari mahasiswa yang lulus dengan prdikat indeks prestasi kurang sejumlah 5, data dari predikat indeks prestasi memuaskan sejumlah 42, data dari predikat indeks prestasi sangat memuaskan sejumlah 150, dan data dari predikat dengan pujian sejumlah 42. *Training* data yang digunakan sebanyak 228 wisudawan dan 5 wisudawan merupakan *test*ing data yang digunakan untuk mengetahui tingkat akurasi algoritma NBC dan C.4.5. Atribut yang digunakan sebagai mahasiswa di setiap periodenya. parameter adalah 11 atribut, dimana 10 adalah predictor dan 1 adalah hasil. Hasil penelitian ini perlu dikembangkan lagi dengan peningkatan jumlah atribut dan data, dan perlu dibuat sebuah sistem penentu predikat kelulusan mahasiswa dari pola yang telah dihasilkan supaya dapat membantu perguruan tinggi untuk meningkatkan predikat kelulusan.

Penelitian Kesembilan **"Prediksi Kelulusan Mahasiswa Tepat Waktu Berdasarkan Dengan Riwayat Akademik Menggunakan Metode *K-Nearest***

***Neighbor*"** (*Imam Riadi et al., 2024*) Pada penelitian ini untuk memprediksi kelulusan mahasiswa dengan menggunakan metode K-Nearest Neighboar dan Hasil penelitian sebagai pengembangan ilmu pengetahuan dan membuktikan apakah algoritma *KNearest Neighbor* bisa digunakan sebagai metode klasifikasi kelulusan mahasiswa tepat waktu dan kelulusan mahasiswa tidak tepat waktu. Setelah dilakukan tahapan-tahapan riset, dapat ditarik kesimpulan bahwa klasifikasi kelulusan mahasiswa melalui penerapan algoritma *K-Nearest Neighbors* membuktikan bisa digunakan untuk klasifikasi kelulusan mahasiswa tepat waktu dan lulus tidak tepat waktu. Proses klasifikasi kelulusan mahasiwa menjadi semakin akurat apabila kapasitas *dataset* penelitiam yang digunakan dengan kuantitas banyak dan atribut yang dipakai lebih *variatif*. Klasifikasi pada penelitian ini menggunakan *dataset* mahasiwa FTI UAD sebanyak 93 data mahasiswa terdiri 78 *data training* dan sebanyak 12 mahasiswa *data testing.* Kaidah penelitian dengan *K-Nearest Neighbors* untuk mengukur akurasi penelitian dengan metode *Confusion Matrix*. Penelitian menggunakan variabel nama, asal SMA/SMK, wilayah asal SMA/SMK, dan nilai rata-rata matematika. Pengujian klasifikasi diuji dengan *cluster data* mulai dari k=2, k=3, k=4, k=5, k=6 dan k=7. Dari pengujian *cluster* menunjukkan *cluster* dengan nilai k=3 adalah nilai paling tinggi dari nilai yang lainnya, dengan hasil pengujian akurasi 78% dan presisi 100%.

Penelitian Kesepuluh **"Perbandingan Metode Algoritma *K-Nearest Neighbors* Dan Random Forest Dalam Prediksi Kelulusan Mahasiswa Mahasiswa Universitas Muhammadiyah Jakarta"** (*Rizal Hadi Hidayatullah et al., 2025*) Pada penelitian ini untuk memprediksi kelulusan mahasiswa Dengan data akademik yang mencakup Indeks Prestasi Kumulatif (IPK), jumlah SKS yang telah ditempuh, tingkat kehadiran, nilai mata kuliah inti, dan durasi penyelesaian tugas akhir, model menunjukkan akurasi tertinggi sebesar 86% saat menggunakan parameter K=5 dengan metrik jarak Euclidean. Hasil penelitian mengungkapkan bahwa variabel IPK, jumlah SKS, dan durasi penyelesaian tugas akhir adalah faktor dominan yang memengaruhi kelulusan mahasiswa. Model prediksi ini memungkinkan identifikasi mahasiswa yang berisiko terlambat lulus

sehingga pihak universitas dapat memberikan intervensi akademik yang lebih dini dan terarah. Pengembangan model dengan memasukkan variabel non-akademik, seperti keikutsertaan dalam organisasi, status pekerjaan, dan motivasi belajar, diharapkan dapat meningkatkan akurasi dan fleksibilitas model dalam memprediksi kelulusan mahasiswa. Dengan demikian, hasil penelitian ini memberikan kontribusi nyata dalam mendukung pengelolaan pendidikan yang lebih baik di Universitas Muhammadiyah Jakarta.

2.2 Matriks Penelitian

Table 2.1 Matriks Penelitian

No	Nama Peneliti	Judul Penelitian	Sumber Data	Metode yang digunakan	Kesimpulan
1	*Andrian Fakih et al., 2025*	Perbandingan Akurasi Algoritma C4.5 Dan KNN Untuk Prediksi Kelulusan Mahasiswa Penerima Beasiswa	Kaggle, (1000-dataset-beasiswa)	Decision Tree (C4.5) dan K-Nearest Neighbor (K-NN)	Algoritma C4.5 memiliki performa lebih baik dibandingkan K-NN dalam klasifikasi penerima beasiswa berdasarkan: akurasi sebesar 74,95% dan nilai AUC lebih baik. **Kelebihanpenelitian**: Menggunakan evaluasi lengkap (Accuracy, Precision, Recall, AUC) dan Menggunakan teknik cross-validation untuk mengurangi overfitting **Kekurangan Penelitian:** Precision C4.5 relatif rendah (20,9%) dan Tidak dijelaskan optimasi parameter (misalnya tuning nilai K pada K-NN)
2	*Teuku Afriliansyah et al., 2023*	Algoritma Ann, Knn, Naive Bayes, Svm Dan Decision Tree Dalam Menentukan Kelulusan Mahasiswa	Sistem Informasi Akademik (IOSYS) Fakultas Teknik	Artificial Neural Network (ANN), K-Nearest Neighbor (KNN), Naïve Bayes, Support	Algoritma KNN menjadi model terbaik dalam memprediksi kelulusan mahasiswa dengan akurasi sebesar 96,95%, dan Optimasi data menggunakan SMOTEENN terbukti meningkatkan akurasi dibanding penelitian

				Vector Machine (SVM), Decision Tree	sebelumnya yang tidak menggunakan optimasi. **Kelebihan Penelitian:** Membandingkan 5 algoritma sekaligus dan Menggunakan optimasi data (SMOTEENN) dengan menggunakan data asli **Kekurangan Penelitian:** Dataset hanya berasal dari satu fakultas (kurang generalisasi) dan Tidak dilakukan cross-validation (hanya split data)
3	*Ichsan Ramhdani et al., 2023*	Perbandingan Metode Data Mining Model klasifikasi Naive Bayes, Decission Tree Dan K-Nearest Neighbour Dalam Memprediksi Ketepatan Kelulusan Mahasiswa Prodi Teknik Informatika Di Universitas Pamulang	Data Akademik Mahasiswa Program Studi Teknik Informatika , Universitas Pamulang	Naïve Bayes, Decision Tree, K-Nearest Neighbour (KNN)	Decision Tree memiliki tingkat akurasi tertinggi dibandingkan metode lainnya. Kemudian Naïve Bayes dan KNN juga mampu melakukan klasifikasi dengan baik, namun performanya sedikit di bawah Decision Tree. **Kelebihan Penelitian:** Memberikan insight bagi prodi untuk evaluasi akademik dan Menggunakan evaluasi performa klasifikasi yang standar **Kekurangan Penelitian:** Tidak membandingkan dengan algoritma berbasis ensemble dan Tidak dijelaskan analisis feature importance secara mendalam.

4	*M Riski Qisthiano et al., 2023*	Penerapan Algoritma *Decision Tree* Dalam Klasifikasi Data Prediksi Kelulusan Mahasiswa	Data alumni mahasiswa perguruan tinggi kota palembang	Decision Tree (C4.5)	Algoritma Decision Tree efektif digunakan untuk memprediksi kelulusan mahasiswa tepat waktu, Dengan 1.739 data, model menghasilkan akurasi terbaik sebesar **87,93%**. **Kelebihan Penelitian:** Menggunakan data nyata dari beberapa perguruan tinggi (lebih variatif) dan Menggunakan validasi sebanyak 5 kali **Kekurangan Penelitian:** Tidak menggunakan cross-validation murni (menggunakan variasi split data) dan Tidak dibahas feature importance atau rule yang terbentuk secara detail
5	*Naurah Nafizah et al., 2023*	Analisis Perbandingan *Decision Tree* C4.5 Dan KNN Dalam Perizinan Bongkar Muatan Kapal	Instansi pelabuhan / sistem perizinan terkait	Decision Tree (C4.5), K-Nearest Neighbor (KNN)	Algoritma C4.5 lebih unggul dibandingkan KNN pada studi kasus ini. Model dapat membantu instansi dalam mempercepat proses evaluasi permohonan izin. **Kelebihan Penelitian:** Studi kasus nyata di bidang logistik/Pelabuhan dan Hasil model dapat digunakan sebagai decision support system **Kekurangan Penelitian:** Dataset terbatas pada satu instansi dan Tidak dilakukan optimasi parameter KNN (misalnya tuning nilai K)

6	*Julianti Suwartini et al., 2024*	Prediksi Kelulusan Mahasiswa Tepat Waktu Menggunakan Metode Naïve Bayes Dan *Decision Tree* Pada Universitas Stella Maris Sumba	Data primer (data akademik mahasiswa) Universitas Stella Maris Sumba	Decision Tree (C4.5), Naïve Bayes	Naïve Bayes menghasilkan akurasi terbaik (71,24%). Kedua metode memiliki performa yang cukup baik dan dapat membantu pihak kampus dalam evaluasi akademik. **KelebihanPenelitian:** Menghasilkan model yang dapat membantu pengambilan keputusan akademik dan Decision Tree menghasilkan rule/pohon keputusan yang interpretatif **Kekurangan Penelitian:** Tidak menggunakan optimasi parameter lanjutan dan Tidak membandingkan dengan algoritma lain seperti KNN, SVM, atau ANN.
7	*Enggar Novianto et al., 2023*	Klasifikasi Algoritma K-Nearest Neighbor, Naive Bayes, *Decision Tree* Untuk Prediksi Status Kelulusan Mahasiswa S1	Akademik Fakultas Hukum Universitas Sebelas Maret	K-Nearest Neighbor (K-NN), Naïve Bayes, Decision Tree	K-NN menjadi metode paling optimal dalam penelitian ini, Decision Tree juga menunjukkan performa sangat baik dan Naïve Bayes memiliki akurasi lebih rendah dibandingkan dua metode lainnya. **KelebihanPenelitian:** Membandingkan 3 algoritma popular dan Dataset nyata dari institusi pendidikan **Kekurangan Penelitian:** Dataset hanya dari satu program studi (kurang generalisasi) dan Tidak dijelaskan tuning parameter (misalnya nilai K pada K-NN)

8	*Agung Wibowo et al., 2022*	Prediksi Predikat Kelulusan Mahasiswa Menggunakan Naive Bayes Dan *Decision Tree* Pada Universitas Xyz	Data akademik mahasiswa Universitas XYZ (perguruan tinggi swasta di Semarang)	Naive Bayes Classifier (NBC), Decision Tree (C4.5)	Predikat kelulusan mahasiswa dapat diprediksi menggunakan metode data mining. Naive Bayes memiliki performa lebih baik dengan akurasi sebesar 86% dibandingkan Decision Tree dalam penelitian ini. **KelebihanPenelitian:** Menggunakan metode CRISP-DM yang sistematis dan Evaluasi lengkap (accuracy, kappa, waktu komputasi) **Kekurangan Penelitian:** Dataset hanya dari satu program studi, Jumlah data relatif kecil (239 data) sehingga Data testing sangat terbatas (5 data)
9	*Imam Riadi et al., 2024*	Prediksi Kelulusan Mahasiswa Tepat Waktu Berdasarkan Dengan Riwayat Akademik Menggunakan Metode *K-Nearest Neighbor*	Data akademik mahasiswa dari bagian akademik perguruan tinggi	K-Nearest Neighbor (K-NN)	Nilai K tertentu menghasilkan performa terbaik dibandingkan nilai K lainnya dan Riwayat akademik (IPS/IPK semester awal) sangat berpengaruh terhadap ketepatan kelulusan. **Kelebihan Penelitian:** Fokus pada satu algoritma sehingga analisis lebih mendalam dan Menggunakan data akademik nyata **Kekurangan Penelitian:** Performa dapat menurun jika data tidak seimbang dan Tidak dibahas optimasi parameter secara mendalam.

10	*Rizal Hadi Hidayatullah et al., 2025*	Perbandingan Metode Algoritma *K-Nearest Neighbors* Dan Random Forest Dalam Prediksi Kelulusan Mahasiswa Mahasiswa Universitas Muhammadiyah Jakarta	Data akademik mahasiswa (perguruan tinggi Universitas Muhammadiyah Jakarta)	*K-Nearest Neighbors, Random Forest*	Prediksi lama studi dapat membantu program studi dalam meningkatkan mutu dan akreditasi, dan performa random forest lebih baik dengan memiliki tingkat akurasi tinggi. **Kelebihan Penelitian:** Menggunakan validasi 10-fold cross validation (cukup kuat secara evaluasi) dan Membandingkan dua metode populer **KekuranganPenelitian:** Performa dapat menurun jika data tidak seimbang dan Tidak dibahas optimasi parameter secara mendalam dan Dataset relatif terbatas pada satu program studi

Untuk Penelitian yang relevan pada di atas dapat disimpulkan sebagai berikut: Dari 10 jurnal di atas yang terdiri dari 10 jurnal internasional terdapat jurnal dengan topik sama yaitu melakukan prediksi kelulusan mahasiswa, terdapat jurnal dengan topik sama namun metodenya berbeda dan terdapat juga jurnal dengan topik berbeda namun metodenya sama. Perbedaan penelitian ini dengan penelitian sebelumnya adalah:

1. Pada beberapa penelitian terdahulu, menggunakan Algoritma *Naïve Bayes*, *Random Forest* dan SVM.
2. Untuk Objek Penelitian tidak hanya berada di universitas melainkan ada beberapa penelitian terdahulu yang menggunakan objek muatan kapal.
3. Untuk penggunaan metode tidak hanya menggunakan *CRISP – DM,* melainkan ada beberapa penelitian terdahulu menggunakan metode *Waterfall, Rapid Miner dll.*

4. Jurnal Penelitian terkait tidak hanya dilakukan di perguruan tinggi atau universitas, tetapi juga ada beberapa penelitian terdahulu yang melakukan penelitian di SMA/SMK.
5. Data yang digunakan adalah data primer, sedangkan beberapa penelitian terdahulu menggunakan data sekunder atau data yang diambil dari sumber tidak langsung, seperti *website, google dll.*
6. Fokus penelitian ini adalah: Melakukan Analisa perbandingan dua algoritma antara algoritma *K-Nearest Neighboars* (KNN) dan *Decission Tree* untuk model prediksi kelulusan mahasiswa dan mencari tingkat akurasi tertinggi diantara kedua algoritma tersebut menggunakan dataset indeks predikat semester (IPS) dengan variable status kelulusan mahasiswa, yaitu apakah mahasiswa diprediksi lulus tepat waktu atau tidak lulus tepat waktu.

2.3 Kesenjangan Penelitian (*GAP Research*)

Berdasarkan tinjauan pustaka terhadap penelitian-penelitian yang relevan, teridentifikasi adanya beberapa kesenjangan (*research gap*) yang menjadi landasan bagi penelitian ini. Walaupun prediksi kelulusan mahasiswa telah banyak diteliti, masih terdapat beberapa aspek penting yang belum dikaji secara mendalam, di antaranya:

1. **Kesenjangan Metodologi**: Sejumlah penelitian terdahulu telah menggunakan berbagai algoritma *Machine Learning* modern seperti, *Naïve Bayes, Support Vector Regression (SVR),* Random Forest untuk melakukan prediksi. Di sisi lain, beberapa penelitian memang menggunakan KNN dan *Decission Tree,* namun seringkali diterapkan pada objek masalah yang berbeda seperti prediksi kasus Data penerimaan Beasiswa, Muatan Kapal dan melakukan prediksi suatu penyakit. Kesenjangan yang ada adalah kurangnya analisis perbandingan secara langsung antara algoritma KNN dan *Decission Tree,* terutama dalam konteks kelulusan mahasiswa.
2. **Kesenjangan Kontekstual (Lokasi):** Penelitian-penelitian terdahulu ada berfokus di lokasi yang berbeda, penelitian sebelumnya dilakukan di, universitas pamulang,

universitas Muhammadiyah Jakarta, universitas XYZ, Universitas stella maris sumba, perguruan tinggi di Palembang, dan belum ada yang melakukan secara spesifik di perguruan tinggi Institut Teknologi PLN

3. **Kesenjangan Data:** data yang digunakan adalah data primer yaitu data yang diambil langsung di histori kampus dengan melakukan wawancara langsung ke pihak yang terkait sedangkan penelitian terdahulu seringkali menggunakan data sekunder atau data yang diambil secara tidak langsung seperti data akademik dari sistem akademik, *Google* atau *website,* namun ada perbandingan singkat antara data primer dan sata sekunder dimana data primer lebih kontekstual dan diperoleh langsung oleh sumber, sedangkan data sekunder tidak selalu sesuai dengan konteks penelitian dan validasi bisa tidak terbaru.
4. **Kesenjangan Fokus Analisis**: Penelitian ini secara khusus bertujuan untuk memvalidasi dan membandingkan secara langsung algoritma *K-Nearest Neighbors* dan *Decission Tree.* Fokus ini memberikan nilai kebaruan yang praktis, yaitu mencari model yang lebih akurat dengan menggunakan dataset dan variabel independen (Lulus tepat waktu, dan lulus tidak tepat waktu).

2.4 Landasan Teori

Landasan teori membahas teori-teori yang berhubungan dengan penelitian yang penulis kerjakan. Berikut adalah teori-teori tersebut:

2.4.1 Teori Prediksi

Prediksi adalah proses memperkirakan secara sistematis apa yang mungkin terjadi di masa depan berdasarkan informasi masa lalu dan sekarang yang kita miliki, sambil meminimalkan kesalahan (perbedaan antara apa yang telah terjadi dan hasil yang diprediksi). Prediksi tidak membutuhkan jawaban yang jelas atas suatu peristiwa yang akan terjadi, tetapi berusaha mencari jawaban yang sedekat mungkin dengan peristiwa tersebut.(*Prasatya et al., 2020)*

2.4.2 Teori Kelulusan

Menurut Kamus Besar Bahasa Indonesia (KBBI), arti kata kelulusan adalah keguguran (melahirkan anak sebelum masanya). Arti lainnya dari kelulusan adalah hal (keadaan) lulus (ujian dan sebagainya).

Menurut Buku Bimbingan Akademik dan Komunikasi Mahasiswa Program Studi S1 Teknik Informatika IT-PLN, kelulusan adalah peralihan dari status mahasiswa menjadi sarjana untuk program Strata-1 dengan menyelesaikan 144-160 SKS setelah menyelesaikan 8 (delapan) semester kegiatan pendidikan.

2.4.3 Machine Learning

Machine Learning (ML) adalah mesin yang dikembangkan untuk dapat belajar sendiri tanpa bimbingan pengguna. Pembelajaran mesin dibangun di bidang lain seperti statistik, matematika, dan penambangan data sehingga mesin dapat belajar dengan menganalisis data tanpa memprogram ulang atau memberikan perintah. Dalam hal ini, pembelajaran mesin mampu mengambil data yang ada dengan instruksinya sendiri. *Machine Learning* juga dapat mempelajari data yang ada dan data yang diperolehnya sehingga dapat melakukan tugas-tugas tertentu. Tugas yang bisa dilakukan *Machine Learning* juga beragam, tergantung apa yang dipelajarinya, mungkin termasuk orang pertama yang menggunakan istilah *Machine Learning* untuk menyebut program komputer yang dirancang dengan prinsip: *"Programming computers to learn from experience should eventually eliminate the need for much of this detailed programming effort."* (*Rizka Yudana et al., 2023)*

Shalev-Shwartz dan Ben-David dalam bukunya yang berjudul *Understanding Machine Learning* menjelaskan istilah "pembelajaran" didalam *Machine Learning* sebagai sebuah proses yang bertujuan untuk mengkonversi pengalaman yang diperoleh sebuah algoritma menjadi pengetahuan yang dimiliki oleh algoritma tersebut. Sebuah algoritma *Machine Learning* memperoleh "pengalaman" pada saat dilakukan proses

pembelajaran (*training*) menggunakan input data (data *training*). “Pengetahuan” sebagai hasil perolehan dari proses pembelajaran disimpan didalam sebuah struktur internal (misalnya: parameter model *Machine Learning*) sebagai hasil generalisasi dari sejumlah pola didalam data *training*. Pengetahuan hasil *training* selanjutnya digunakan terhadap data baru atau data *test*ing (lihat gambar diatas). Sebuah algoritma *Machine Learning* dikatakan memiliki kinerja pembelajaran sangat baik apabila mampu mengenali pola dari data baru secara akurat. Metrik kinerja pembelajaran *Machine Learning* merupakan alat untuk mengukur atau menilai sejauh mana keakuratan hasil prediksi yang dihasilkan oleh pengetahuan hasil pembelajaran *Machine Learning* terhadap data aktual yang dijadikan target pengujian.(*Putri et al., 2023)*

2.4.4 *K-Nearest Neighbors* (KNN)

Algoritma *K-Nearest Neighbors* (KNN) merupakan algoritma yang berfungsi untuk mengklasifikasikan suatu data berdasarkan data pembelajaran (train dataset), yang diambil dari K tetangga terdekatnya (*nearest neighbor*), dengan K adalah banyaknya tetangga terdekat, Dengan kata lain tujuan dari algoritma ini yaitu mengklasifikasikan objek baru berdasarkan atribut dan contoh dari data *training*. Contoh kasus, misalkan untuk memprediksi lama studi mahasiswa digunakan data mahasiswa yang telah menyelesaikan masa studinya. KNN adalah algoritma *supervised learning* yang maksudnya algoritma ini menggunakan data yang telah ada dan outputnya telah diketahui. KNN banyak dipergunakan pada aplikasi data mining, *pattern recognition, image processing*, dll. Algoritma *K- Nearest Neighbors* menggunakan klasifikasi ketetanggaan (*neighbors*) sebagai nilai prediksi dari *query instance* yang baru. Algoritma ini sederhana, bekerja berdasarkan jarak terpendek dari *query instance* ke *training sample* untuk menentukan ketetanggaannya. (*Etriyanti Sistem Informasi et al., 2021*)

Adapun langkah-langkah untuk menghitung Metode *K-Nearest Neighbors* adalah sebagai berikut:

1. Tentukan parameter K.
2. Hitung jarak antara penilaian dan setiap sesi pelatihan.
3. Urutkan jarak yang sudah terbentuk.
4. Tentukan jarak terdekat sesuai urutan K.
5. Tentukan kelas yang sesuai.
6. Temukan jumlah kelas tetangga terdekat dan tetapkan kelas tersebut sebagai kelas data untuk evaluasi.

Perhitungan jarak antara data baru dan lama dapat dilakukan sesuai dengan rumus berikut:

$$d_i = \sqrt{\sum_{i=1}^{p} (x_{2i} - x_{1i})^2}$$

Keterangan:

x1 = Data sampel atau data *training*

x2 = Data *test* atau data uji

i = Data Variabel

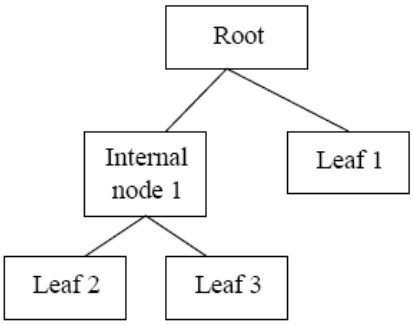
d = Jarak

p = Dimensi Data *Decision Tree*

2.4.5 Decission Tree

Decision Tree merupakan salah satu metode klasifikasi yang menggunakan representasi struktur pohon yang berisi alternatif-alternatif pemecahan dari suatu permasalahan *Decision Tree* juga dapat menghasilkan pohon keputusan yang dapat akan digunakan untuk mengidentifikasi, menganalisis dan melihat hubungan antara atribut-atribut yang digunakan, bahwa *Decision Tree* merupakan salah satu metode klasifikasi yang direpresentasikan dengan struktur pohon (*Tree*) atribut direpresentasikan oleh *node*,

nilai dari atribut direpresentasikan oleh cabang, kelas direpresentasikan oleh daun. (*Cahyaningtyas et al., 2020)*



Gambar 2. 1 *Decission Tree (Ramdhani, 2023)*

2.4.6 Pyhton

Python adalah bahasa pemrograman yang ditafsirkan untuk tujuan umum dengan filosofi desain yang berpusat pada keterbacaan kode. *Python* bertujuan untuk menjadi bahasa dengan perpustakaan fungsi standar yang besar dan komprehensif, menggabungkan keterampilan dan kemampuan dengan sintaks kode yang sangat jelas. *Python* juga didukung oleh komunitas yang besar. *Python* mendukung banyak paradigma pemrograman yang berbeda, terutama pemrograman berorientasi objek, imperatif, dan fungsional. Secara khusus, *Python* tersedia sebagai bahasa pemrograman dinamis dengan manajemen memori otomatis. *Python* merupakan sebuah bahasa pemrograman tingkat tinggi yang dibuat oleh *Guido Van Rossum* dan dirilis pada tahun 1991 *Python* juga merupakan bahasa yanng sangat populer belakangan ini. Selain itu *Python* juga merupakan bahasa pemrograman yang multi fungsi salah satunya pada bidang *Machine Learning* dan *Deep Learning*. Kode *Python* sekarang dapat berjalan di beberapa platform sistem operasi.(*Riziq sirfatullah Alfarizi et al., 2023)*

2.4.7 Google Colab

Colaboratory adalah proyek penelitian *Google*, dan dibuat untuk membantu menyebarkan pendidikan dan penelitian pembelajaran mesin. Ini adalah lingkungan

notebook Jupyter, yang dapat digunakan tanpa pengaturan apa pun dan berjalan sepenuhnya di *cloud*. *Google* menawarkan penggunaan GPU secara gratis dan ini merupakan fitur yang menarik bagi para pengembang. Alasan untuk membuatnya tersedia untuk umum adalah untuk membuat perangkat lunaknya menjadi standar di bidang akademisi untuk mengajar pembelajaran mesin dan ilmu data. Mereka mungkin juga memiliki jangka Panjang berencana membangun basis klien untuk Google *Cloud* API yang dijual per penggunaan.

Dengan menggunakan *Colab*, programmer dapat menulis, mengedit, dan mengeksekusi kode dengan *Python*. Selain itu, pustaka *Python* populer seperti *NumPy* dan Matplotlib dapat digunakan untuk menganalisis dan memvisualisasikan data. Hal ini juga memungkinkan kita untuk mengintegrasikan perpustakaan open-source bernama *PyTorch, TensorFlow*, dan *OpenCV*. (*Ray et al., 2021)*

2.4.8 Holdout validation

Hold-Out adalah salah satu dari jenis pengujian *Cross Validation* yang berfungsi untuk menilai kinerja proses sebuah metode algoritma dengan membagi dataset menjadi dua bagian yaitu data pelatihan dan data pengujian. Kelebihan *hold-out* dibandingkan dengan *cross validation* lainnya adalah sederhana dengan waktu komputasi yang cepat. Pada penelitian yang dilakukan oleh Saat Saat mengacak data yang akan dipecah menjadi data pelatihan dan pengujian, kemungkinan besar satu atau lebih klasifikasi akan merepresentasikannya secara berlebihan. Data latih dan uji yang dihasilkan tidak representatif dalam arti klasifikasi tersebut lebih dominan dibandingkan dengan klasifikasi lainnya. Oleh karena itu, diperlukan metode *holdout*. Metode ini memungkinkan data pelatihan dan pengujian yang dihasilkan untuk mewakili setiap klasifikasi secara proporsional. *(Manhitu et al., 2025)*

2.2.9 *Confusion Matrix*

Untuk menentukan *Confusion matrix* adalah sebuah tabel yang digunakan untuk menampilkan jumlah data uji yang diklasifikasikan dengan benar dan salah, sehingga

memudahkan dalam mengevaluasi akurasi suatu sistem klasifikasi. Dengan menggunakan *confusion matrix*, kita dapat melihat secara detail kinerja suatu sistem klasifikasi dan mengidentifikasi di mana terjadi kesalahan klasifikasi. *Confusion matrix* merupakan sebuah teknik yang mudah dan efektif dalam mengukur kinerja sistem klasifikasi. Tujuan utama dari *confusion matrix* adalah untuk menilai performa atau akurasi dari suatu sistem klasifikasi dalam mengklasifikasikan data uji *(Nurhidayat & Dewi, 2023)*

		True Class	
		Positive	Negative
Predicted Class	Positive	true positives count (TP)	false negatives count (FP)
	Negative	false positives count (FN)	true negatives count (TN)

Gambar 2. 2 *Confusion Matriks (Riadi et al., 2024)*

Berdasarkan gambar 2.4 diatas terdapat beberapa nilai didalam matriks yaitu (TP) "*True Positive*", (TN) "*True Negative*", (FP) "*False Positive*", dan (FN) "*False Negative*", semua kemungkinan kejadian sebenarnya *positif* (P) dan semua kemungkinan kejadian sebenarnya *negatif* (N). Nilai yang dimaksud dapat digunakan untuk menghitung akurasi dengan persamaan dibawah ini.

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} x\ 100\%$$

Akurasi digunakan sebagai ukuran untuk menilai keakuratan model yang melakukan klasifikasi. Sementara untuk menghitung level presesi prediksi peristiwa, dapat menggunakan rumus berikut:

$$Precision = \frac{TP}{TP + FP}$$

Presisi menunjukkan seberapa akurat suatu model memprediksi kejadian positif dalam serangkaian kegiatan prediksi. Perhitungan presisi biasanya digunakan untuk pengembangan model prediksi hujan di suatu daerah. Selain presisi dan akurasi, untuk dapat melihat lebih detail kinerja suatu sistem, *recall* atau sensitifitas sistem terhadap suatu kelas juga dapat dilihat. *Recall* dapat dihitung dengan menggunakan persamaan.

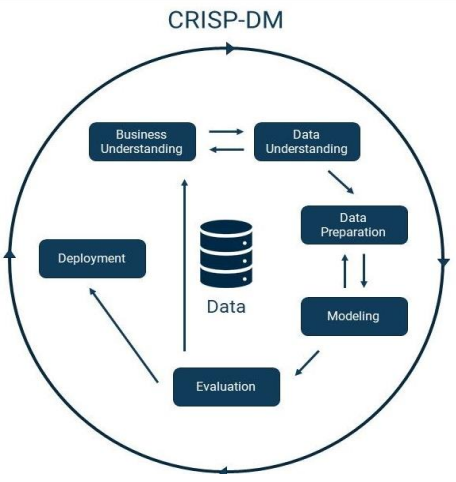
$$Recall = \frac{TP}{TP + FN}$$

F1-Score adalah harmonic mean dari precision dan *recall*. Yang secara matematika dapat ditulis seperti persamaan berikut

$$F1 - score = 2\,x\,\frac{Precision\ x\ Recall}{Precision + Recall}$$

Nilai terbaik *F1-Score* adalah 1.0 dan nilai terburuknya adalah 0. Secara representasi, jika *F1-Score* punya skor yang 1.0 atau mendekati 1.0 mengindikasikan bahwa model klasifikasi kita punya precision dan *recall* yang baik.

2.2.10 *CRISP - DM*



Gambar 2. 3 *CRISP – DM (Schröer et al., 2021)*

1. ***BusinessUnderstanding***

   Tahap ini adalah memahami kebutuhan pelanggan secara mendalam. Kegiatan yang dilakukan pada tahap ini adalah menentukan tujuan bisnis, menilai situasi ketersediaan sumber daya, menentukan tujuan pengumpulan data, dan menghasilkan rencana proyek

2. ***Data Understanding***

   Selanjutnya adalah tahap pemahaman data, yaitu mengidentifikasi, mengumpulkan, dan menganalisis kumpulan data yang dapat membantu untuk mencapai tujuan proyek. Kegiatan pada tahap ini adalah mengumpulkan data awal, menjelaskan data, menjelajahi data, dan memverifikasi kualitas data.

3. ***DataPreparation***

   Fase ini sering disebut "*data mining*", yaitu menyiapkan kumpulan data akhir untuk pemodelan. Kegiatan pada fase ini adalah memperbaiki kualitas data agar sesuai dengan proses *modelling* yang akan dilakukan berikutnya.

4. ***Modeling***

   Membuat dan menilai berbagai model berdasarkan beberapa teknik pemodelan yang berbeda. Pada tahap ini, ada empat tugas, yaitu memilih teknik pemodelan, menghasilkan desain pengujian, membangun model, dan yang terakhir menilai model.

5. ***Evaluation***

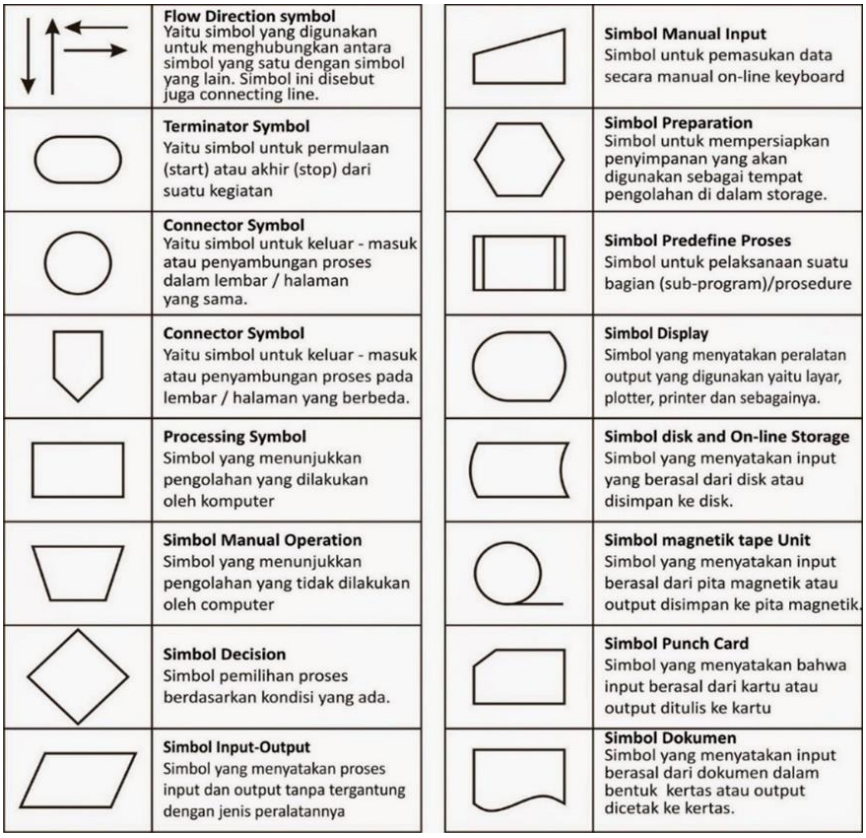
   Fase evaluasi ini melihat lebih luas model yang paling sesuai dengan bisnis dan yang harus dilakukan selanjutnya. Ada tiga kegiatan yang mewakili fase evaluasi, yaitu evaluasi hasil, proses peninjauan, dan penentuan langkah selanjutnya.

2.2.11 Flowchart

Flowchart adalah grafik yang secara logis menampilkan aliran (flows) dalam suatu program atau prosedur sistem. Flowchart adalah metode yang menggambarkan langkah-langkah pemecahan masalah dengan mewakili simbol-simbol yang berbeda yang mudah dipahami, digunakan, dan standar. Tujuan penggunaan flowchart adalah untuk

menjelaskan fase pemecahan masalah secara sederhana, jelas, teratur, dan dapat dipahami dengan menggunakan notasi standar. Langkah-langkah yang disajikan untuk memecahkan masalah harus jelas, sederhana, dan tepat.(Zamrodah, 2016)

Beberapa simbol yang digunakan dalam menggambar suatu flowchart adalah sebagai berikut:



Gambar 2. 4 Simbol – Simbol Flowchart

Flowchart memiliki banyak jenis, Sebagian diantaranya adalah:

1. Flowchart Sistem

   Flowchart sistem adalah diagram yang mewakili aliran fungsional dari keseluruhan sistem. Bagan alir sistem menunjukkan bagaimana sistem bekerja atau apa yang dilakukan sistem.

2. Flowchart Dokumen

Flowchart dokumeen didefinisikan sebagai bagan alir yang menampilkan alur pesan dan formulir yang berisi duplikat yang mirip dengan bagan alur dokumen, atau bisa juga disebut formulir bagan alur (flowchart).

3. Flowchart Skematik

   Flowchart skematik dapat didefinisikan sebagai flowchart yang mirip dengan flowchart sistem untuk menggambarkan aliran dalam suatu sistem. Perbedaan antara flowchart adalah penggunaan simbol flowchart sistem serta penggunaan gambar komputer dan perangkat lain yang digunakan. Tujuan menggunakan gambar-gambar ini adalah untuk membuatnya lebih mudah dipahami bagi mereka yang tidak terbiasa dengan simbol diagram alur, tetapi gambar-gambar ini membuat gambar-gambar ini berat dan memakan waktu untuk menggambar.

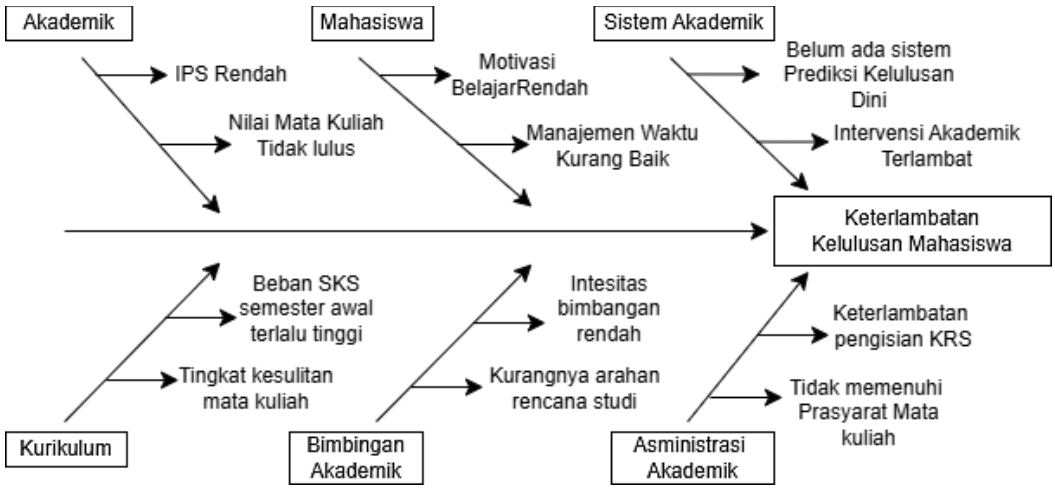
4. Flowchart Program

   Program flowchart (bagan alir flowchart) adalah bagan yang menjelaskan secara terperinci langkah-langkah dari proses program. Flowchart program didasarkan pada asal mula flowchart sistem. Ada dua jenis diagram alur program, diagram alur logika program dan diagram alur program komputer rinci (*detailed computer program flowchart*). *Logic* flowchart Ini digunakan untuk menggambarkan secara logis diagram dalam pemrograman komputer. Diagram logis dari program dibuat menggunakan analisis sistem. Diagram terperinci dari logika program komputer (*detailed computer program flowchart*), bagan alir ini di buat oleh pemrogram.

5. Flowchart Proses

   Flowchart Proses adalah diagram alur proses adalah diagram alur yang menggambarkan proses dalam proses yang benar-benar dapat membantu analis yang memanfaatkannya dalam teknik industri.

2.5 Diagram Fishbone Keterlambatan Kelulusan Mahasiswa



Gambar 2. 6 Diagram *Fishbone*

Diagram fishbone pada Gambar diatas digunakan untuk menggambarkan hubungan sebab–akibat dari permasalahan keterlambatan kelulusan mahasiswa**.** Diagram ini bertujuan untuk mengidentifikasi dan mengelompokkan faktor-faktor utama yang berpotensi memengaruhi masa studi mahasiswa. Berdasarkan hasil analisis, faktor-faktor penyebab keterlambatan kelulusan mahasiswa dikelompokkan ke dalam enam kategori utama, yaitu faktor akademik, mahasiswa, kurikulum, bimbingan akademik, administrasi akademik, dan sistem akademik.

2.5.1 Faktor Akademik

Faktor ini meliputi Indeks Prestasi Semester (IPS) yang rendah, nilai mata kuliah dasar yang tidak lulus, serta pengulangan mata kuliah. Faktor akademik dianggap sebagai faktor yang paling dominan dan objektif karena dapat diukur secara kuantitatif dan tersedia dalam data akademik historis

2.5.2 Faktor Mahasiswa

Faktor ini mencakup motivasi belajar, manajemen waktu, kedisiplinan, serta kemampuan adaptasi mahasiswa terhadap lingkungan perkuliahan. Meskipun faktor ini

memiliki pengaruh terhadap kelulusan mahasiswa, namun bersifat subjektif dan sulit diukur secara kuantitatif sehingga tidak dimasukkan sebagai variabel dalam pemodelan penelitian ini.

2.5.3 Faktor Kurikulum

Faktor ini meliputi beban SKS pada semester awal, tingkat kesulitan mata kuliah dasar, serta urutan pengambilan mata kuliah. Faktor kurikulum dapat memengaruhi capaian akademik mahasiswa, khususnya pada semester awal, yang pada akhirnya berdampak pada ketepatan waktu kelulusan.

2.5.4 Faktor Bimbingan Akademik

Faktor ini mencakup intensitas bimbingan, monitoring perkembangan akademik mahasiswa, serta arahan dalam penyusunan rencana studi. Kurangnya bimbingan akademik yang optimal dapat menyebabkan mahasiswa mengalami kesulitan dalam pengambilan keputusan akademik yang berdampak pada keterlambatan kelulusan.

2.5.5 Faktor Administrasi Akademik

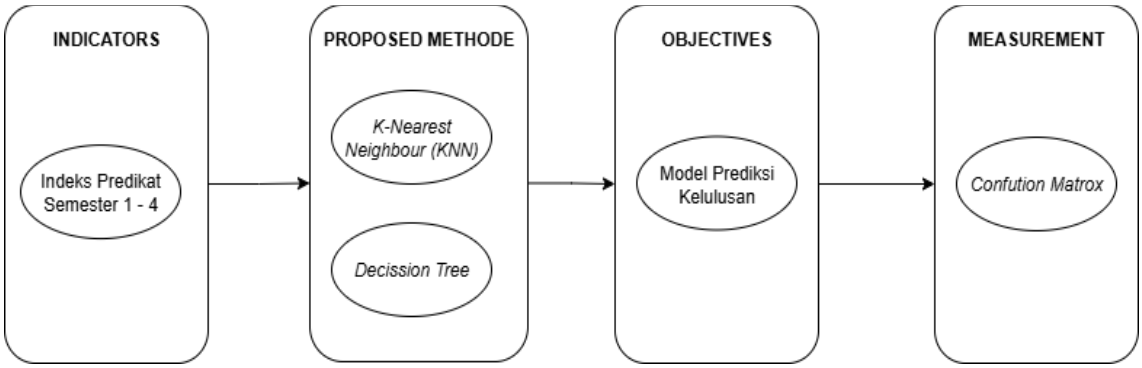
Faktor ini meliputi keterlambatan pengisian KRS, kesalahan pengambilan mata kuliah, serta ketidaksesuaian pemenuhan prasyarat mata kuliah. Permasalahan pada aspek administrasi akademik dapat menyebabkan mahasiswa tidak dapat mengikuti perkuliahan sesuai rencana studi.

2.5.6 Faktor Sistem Akademik

Faktor ini meliputi belum tersedianya sistem prediksi kelulusan secara dini, keterlambatan intervensi akademik, serta monitoring akademik yang belum berbasis data secara optimal. Faktor ini menjadi dasar perlunya penerapan pendekatan data mining dalam penelitian ini sebagai upaya mendukung pengambilan keputusan akademik yang lebih tepat dan proaktif.

2.6 Kerangka Pemikiran

Kerangka pemikiran memberikan gambaran tahapan demi tahap mulai dari proses awal analisis data hingga menghasilkan kesimpulan dari data yang dianalisis. Adapun kerangka pemikiran yang dilakukan sebagai berikut:



Gambar 2. 6 Kerangka Pemikiran

Pada Gambar diatas menjelaskan kerangka pemikiran bagaimana penelitian ini dilakukan. Kerangka pemikiran menunjukkan ketertarikan antara satu komponen dengan komponen lainnya. Berikut penjelasan dari tiap komponen:

1. Indicators

Indicators adalah sebuah parameter yang akan diuji dalam suatu penelitian. *Indicators* juga didefinisikan dengan masukan (input) yang nantinya akan diolah pada tahap selanjutnya. Pada penelitian ini, yang menjadi *indicators* adalah data nilai Indeks Prestasi Semester 1 sampai 4 yang kemudian data ini akan diolah pada tahap selanjutnya.

2. Proposed Method

Proposed Method berisi metode yang ditetapkan sebagai solusi untuk memecahkan masalah berdasarkan indikator yang telah diperoleh. Dalam penelitian ini, penulis menggunakan algoritma *K-Nearest Neighbors (KNN)*, yaitu metode klasifikasi yang didasarkan pada kedekatan jarak dengan data lain, serta *Decision Tree*, yaitu metode klasifikasi yang membentuk pohon keputusan dengan memetakan aturan dari data untuk

memprediksi hasil. Kedua metode tersebut digunakan sebagai proposed method dalam menganalisis indicator yang telah disebutkan sebelumnya.

3. Objectives

Objectives merupakan suatu tujuan atau sasaran yang ingin dicapai dalam suatu penelitian. Adapun tujuan yang ingin dicapai pada penelitian ini ialah untuk mendapatkan kedua model algoritma yaitu *K-Nearest Neigbors* dan *Decission Tree* untuk prediksi kelulusan mahasiswa berdasarkan indeks predikat semester awal.

4. Measurement

Measurement adalah proses untuk mengukur hasil dari suatu penelitian. Pada penelitian ini, hasil dari nilai prediksi yang didapatkan dilakukan proses untuk melakukan pengecekan akurasi model. Pengujian dilakukan menggunakan *Confusion Matrix* untuk mengukur nilai *accuracy, precision, recall, F1 score* sebagai ketepatan dari hasil model yang telah terbentuk. Kedua metode (*K-Nearest Neigbors dan* dan *Decision Tree*) dibandingkan performanya berdasarkan metrik tersebut.