BAB I

PENDAHULUAN

1.1 Latar Belakang

Transformasi digital dalam ranah pendidikan modern telah mendorong pemanfaatan *Artificial Intelligence* dan *Machine Learning*, khususnya melalui pendekatan *Data Mining*. Dari perspektif pendidikan tinggi, kelulusan tepat waktu merupakan salah satu indikator utama keberhasilan akademik. Bagi mahasiswa, kelulusan tepat waktu dapat mengurangi beban finansial dan mempercepat transisi ke dunia profesional. Sementara bagi institusi perguruan tinggi, tingkat kelulusan secara langsung berdampak pada penilaian akreditasi dan reputasi akademis[1]–[3]. Oleh karena itu, diperlukan suatu pendekatan analitik prediktif yang mampu mengidentifikasi risiko keterlambatan kelulusan sekaligus memetakan akar faktor penghambatnya secara jelas dan terukur.

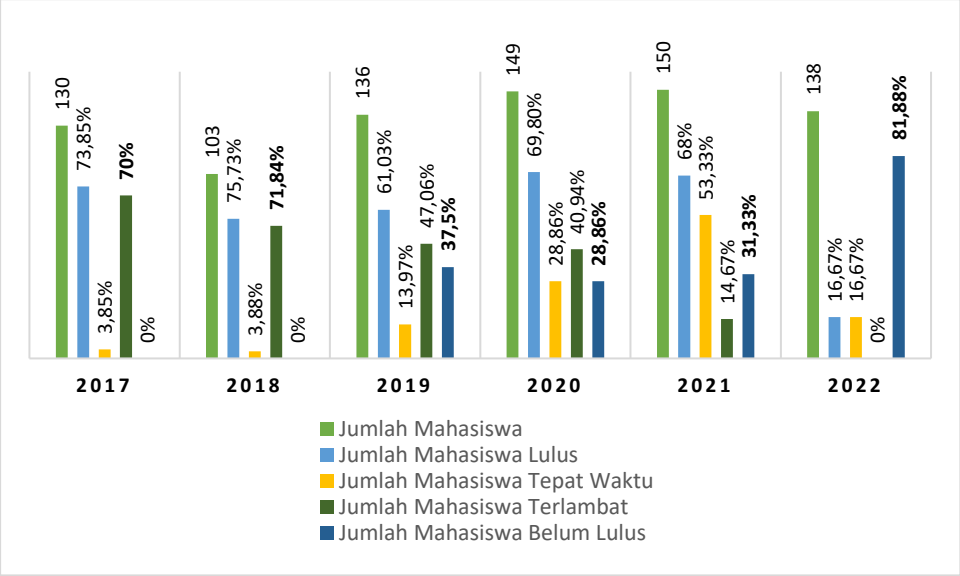
Kebutuhan untuk menerapkan pemodelan analitik ini sangat relevan dengan kondisi di Program Studi Teknik Informatika, Fakultas Sains dan Teknologi, Universitas Islam Negeri (UIN) Sunan Gunung Djati Bandung. Program studi ini telah berhasil meraih status akreditasi dengan status "Unggul"[4], [5]. Dalam upaya menjaga dan meningkatkan kualitas tersebut, pemenuhan indikator kelulusan tepat waktu (8 semester) berdasarkan pedoman akademik universitas menjadi tolok ukur evaluasi yang sangat penting[6].

Tabel 1. 1 Rekapitulasi Status Kelulusan Mahasiswa Teknik Informatika Angkatan 2017-2022

Angkatan	Jumlah Mahasiswa	Jumlah Mahasiswa Lulus	Jumlah Mahasiswa Tepat Waktu	Jumlah Mahasiswa Terlambat	Jumlah Mahasiswa Mengundurkan Diri	Lainnya	Jumlah Mahasiswa Belum Lulus
2017	130	96	5	91	32	2	0
2018	103	78	4	74	22	3	0
2019	136	83	19	64	0	2	51
2020	149	104	43	61	2	0	43
2021	150	102	80	22	0	1	47

Angkatan	Jumlah Mahasiswa	Jumlah Mahasiswa Lulus	Jumlah Mahasiswa Tepat Waktu	Jumlah Mahasiswa Terlambat	Jumlah Mahasiswa Mengundurkan Diri	Lainnya	Jumlah Mahasiswa Belum Lulus
2022	138	23	23	0	2	0	113
TOTAL	**806**	**486**	**174**	**312**	**58**	**8**	**254**

Tabel 1.1 memaparkan rekapitulasi data status mahasiswa angkatan tahun 2017 hingga 2022 pada Program Studi Teknik Informatika. Data tersebut merincikan sebaran status akademik mahasiswa secara keseluruhan, mulai dari jumlah mahasiswa yang lulus tepat waktu, lulus terlambat, mengundurkan diri, hingga jumlah mahasiswa yang belum lulus. Untuk melihat proporsi dan tren dari fenomena keterlambatan kelulusan ini secara lebih jelas, perbandingan persentasenya divisualisasikan pada Gambar 1.1.



Gambar 1. 1 Persentase Status Kelulusan Mahasiswa

Gambar 1.1 merupakan visualisasi persentase status kelulusan mahasiswa yang diolah berdasarkan data pada Tabel 1.1. Pengamatan terhadap data kelulusan dari angkatan tahun 2017 hingga 2022 menunjukkan adanya rasio keterlambatan lulus yang cukup tinggi serta penumpukan mahasiswa yang belum menyelesaikan pendidikannya. Meskipun angkatan tahun 2017 dan 2018 mencatatkan tingkat kelulusan secara penuh (tidak ada mahasiswa yang belum lulus), tingginya angka keterlambatan kelulusan pada kedua angkatan tersebut sangat mengkhawatirkan, dengan 70% pada angkatan 2017 dan 71,84% pada angkatan 2018. Disisi lain,

untuk angkatan yang masih aktif menyelesaikan studi, proporsi mahasiswa yang belum lulus mengalami lonjakan signifikan, dimulai dari angkatan 2019 (37,5%), dan terus meningkat dengan angka 28,86% hingga 31,33% pada angkatan 2020 dan 2021, hingga mencapai titik puncaknya di angkatan 2022 sebesar 81,88%. Jika fenomena ini tidak diberi perhatian atau tindakan, maka akan berdampak langsung pada penurunan kualitas lulusan dan berpotensi mengganggu status akreditasi "Unggul" yang sudah dicapai oleh program studi. Fenomena ini menunjukkan bahwa intervensi berbasis komputasional sangat diperlukan untuk mendeteksi potensi keterlambatan sejak dini.

Berdasarkan tinjauan pustaka, keterlambatan dalam menyelesaikan studi dipengaruhi oleh faktor internal (motivasi, kesehatan, kemampuan akademis, keterlibatan organisasi, dan tantangan tugas akhir) dan faktor eksternal (pengaruh sosial, peluang kerja, fasilitas pembelajaran, dan komunikasi akademik)[7]–[18]. Untuk menganalisis interaksi antar faktor tersebut, pendekatan klasifikasi dalam *data mining* terbukti efektif untuk memprediksi pola tersembunyi dari data historis. Sebagai bentuk penyempurnaan dari algoritma *ID3* dan *C4.5* gagasan *Ross Quinlan*, algoritma klasifikasi *C5.0* dalam bentuk *Decision Tree* dipilih karena dikenal memiliki kemampuan komputasi yang sangat handal berkat efisiensi penggunaan memorinya, kecepatan eksekusi, serta kemampuannya dalam menghasilkan representasi aturan keputusan yang mudah dipahami[19]. Di ranah akademik, keandalan *C5.0* telah dibuktikan oleh beberapa studi terdahulu, seperti riset oleh Puteri dkk. yang berhasil memprediksi risiko *drop out* mahasiswa dengan akurasi mencapai 95%, mengungguli model *Deep Learning* berbasis *GRU*[20]. Pemanfaatan *C5.0* juga berhasil diterapkan dalam memprediksi kelulusan mata kuliah tertentu hingga mengevaluasi tingkat pemahaman belajar daring mahasiswa[21], [22].

Meskipun algoritma *C5.0* menunjukkan kinerja yang baik, terdapat kesenjangan penelitian yang nyata pada beberapa penelitian terkait kelulusan mahasiswa terdahulu. Mayoritas penelitian dalam ranah pendidikan tersebut umumnya hanya menguji *C5.0* pada *dataset* yang sudah bersih, berukuran kecil, dan dijalankan menggunakan pengaturan parameter bawaan. Sebagai contoh, penelitian kelulusan mata kuliah oleh Tanjung dkk. memiliki keterbatasan karena hanya menguji sampel

data sangat kecil sebanyak 15 mahasiswa dengan transformasi kategori manual[23]. Keterbatasan yang sama juga ditemukan dalam penelitian tentang pemahaman belajar daring oleh Priyatna & Sanwani yang modelnya berjalan menggunakan setelan bawaan tanpa adanya optimasi parameter otomatis, sehingga akurasi yang dihasilkan masih terbatas[22]. Selain itu, penelitian terkait kelulusan mahasiswa oleh Zebua dkk. serta Nusari dkk. juga mengabaikan masalah ketidakseimbangan data dan membiarkan parameter *C5.0* berjalan apa adanya[24], [25]. Padahal, data dari dunia nyata sering kali bersifat mentah, tidak seimbang, dan belum memiliki label target, yang dapat memicu terjadinya bias keputusan serta *overfitting* pada model *C5.0*.

Untuk mengatasi ketiadaan label target pada data mentah tersebut, proses pelabelan secara manual sering kali tidak efisien dan rentan terhadap penilaian yang subjektif. Sebagai solusinya, penelitian ini mengusulkan pendekatan *pseudo-labeling* melalui algoritma *unsupervised learning K-Means Clustering*. Algoritma ini mampu memetakan pola tersembunyi dari sampel *dataset* secara matematis untuk membentuk klaster kelas secara objektif. Hasil pengelompokan inilah yang kemudian dijadikan label target sebelum data diproses oleh algoritma klasifikasi[26].

Setelah data memiliki label target, tantangan selanjutnya adalah menangani proporsi kelas yang tidak seimbang serta keterbatasan parameter bawaan. Jika diabaikan, algoritma *C5.0* cenderung bias pada kelas mayoritas dan menghasilkan struktur pohon yang terlalu rumit. Untuk mencegah hal tersebut, penelitian ini mengadopsi teknik penyeimbangan data dan penyetelan parameter yang telah terbukti berhasil di luar bidang pendidikan[27]. Secara spesifik, teknik *Synthetic Minority Over-sampling Technique (SMOTE)* diterapkan guna menyeimbangkan jumlah data dengan cara menghasilkan data buatan yang polanya mengikuti data minoritas, sehingga model dapat belajar secara proporsional. Sementara itu, optimasi algoritma dilakukan menggunakan metode *GridSearchCV* yang didukung pengujian *Cross-Validation* untuk menyeleksi parameter terbaik secara otomatis. Dalam ranah medis dan keamanan siber, kombinasi *SMOTE* dan *GridSearchCV* telah terbukti andal meningkatkan performa *Decision Tree* secara signifikan[28],

[29]. Namun, pemanfaatannya untuk mengoptimalkan performa algoritma *C5.0* dalam memecahkan masalah keterlambatan masa studi masih sangat terbatas.

Tantangan utama dari penerapan model klasifikasi yang telah dioptimalkan secara kompleks adalah karakteristiknya yang menyerupai kotak hitam *(black box)*, sehingga sulit bagi manusia untuk menilai pengaruh matematis dari setiap atribut prediktor. Pihak pengambil keputusan di perguruan tinggi tidak hanya membutuhkan persentase akurasi prediksi kelulusan, tetapi juga menuntut transparansi mengenai alasan logis mengapa seorang mahasiswa diklasifikasikan berisiko terlambat. Untuk memecahkan masalah transparansi ini, penelitian terkait kecerdasan buatan mulai mengintegrasikan konsep *Explainable AI (XAI)* melalui pendekatan *Shapley Additive exPlanations (SHAP).* Berdasarkan *Game Theory*, *SHAP* mampu memvisualisasikan logika keputusan model, sehingga besarnya pengaruh dari masing-masing faktor penghambat kelulusan dapat dipetakan secara transparan[30].

Berdasarkan permasalahan yang telah dipaparkan, penelitian ini diusulkan untuk membangun sebuah pemodelan prediktif yang dapat mengklasifikasikan tingkat risiko keterlambatan masa studi mahasiswa berdasarkan faktor penghambat. Pada tahap prapemrosesan, penanganan data yang tidak ideal dilakukan menggunakan teknik *pseudo-labeling K-Means* untuk membentuk kelas target dan metode *SMOTE* untuk menyeimbangkan distribusi data. Selanjutnya, pemodelan prediktif ini menggunakan algoritma *Decision Tree C5.0* yang dioptimasi dengan *GridSearchCV* untuk mencapai kinerja klasifikasi yang akurat, sekaligus memanfaatkan metode *Explainable AI (XAI*) melalui pendekatan *Shapley Additive exPlanations (SHAP)* guna menganalisis kontribusi setiap faktor penghambat secara transparan. Penelitian ini menggunakan 9 atribut faktor penghambat kelulusan mahasiswa yang relevan dengan literatur terdahulu sebagai variabel utama dalam *dataset*, yaitu: motivasi, kesehatan, kemampuan akademis, keterlibatan organisasi, tantangan tugas akhir, pengaruh sosial, peluang kerja, fasilitas pembelajaran, dan komunikasi akademik. Penelitian ini memanfaatkan data sekunder yang bersumber dari rekapitulasi kuesioner terkait faktor penghambat studi mahasiswa. *Dataset* sekunder ini diperoleh melalui teknik studi dokumentasi terhadap rekapitulasi jawaban kuesioner yang sebelumnya telah disebarkan secara

internal oleh pihak Program Studi Teknik Informatika UIN Sunan Gunung Djati Bandung.

1.2 Rumusan Masalah

Berdasarkan latar belakang yang telah diuraikan, permasalahan komputasional yang menjadi fokus utama dalam penelitian ini dirumuskan sebagai berikut:

1. Bagaimana penerapan teknik *pseudo-labeling K-Means* dan *SMOTE* pada tahap prapemrosesan data, serta hasil optimasi parameter *GridSearchCV* pada model algoritma *C5.0* dalam memprediksi risiko keterlambatan kelulusan?
2. Bagaimana tingkat akurasi dan kinerja model algoritma *C5.0* teroptimasi dalam memprediksi risiko keterlambatan kelulusan?
3. Bagaimana interpretasi metode *SHAP* dalam memetakan kontribusi dan pengaruh masing-masing faktor penghambat terhadap prediksi risiko kelulusan mahasiswa?

1.3 Tujuan Penelitian

Penelitian ini dilaksanakan dengan tujuan komputasional sebagai berikut:

1. Mengimplementasikan teknik *pseudo-labeling K-Means* dan *SMOTE* pada tahap prapemrosesan data, serta mendapatkan kombinasi parameter optimal algoritma *C5.0* menggunakan metode *GridSearchCV*.
2. Mengukur tingkat akurasi dan kinerja model algoritma *C5.0* teroptimasi dalam memprediksi risiko keterlambatan kelulusan mahasiswa.
3. Menganalisis kontribusi faktor penghambat kelulusan mahasiswa secara visual dan terukur menggunakan metode *SHAP*.

1.4 Manfaat Penelitian

Terdapat manfaat yang diperoleh dari penelitian ini, di antaranya:

1. Memberikan kontribusi ilmiah dalam bidang *Data Science*, khususnya sebagai acuan metodologi dalam menerapkan optimasi parameter otomatis *(GridSearchCV)* pada algoritma *C5.0* untuk menghasilkan model klasifikasi yang optimal.
2. Menjadi referensi akademis dalam penerapan konsep *Explainable AI (XAI)* melalui pendekatan *SHAP*, guna mengubah model prediksi yang rumit

menjadi grafik visual yang mudah dipahami dan diinterpretasikan oleh manusia.

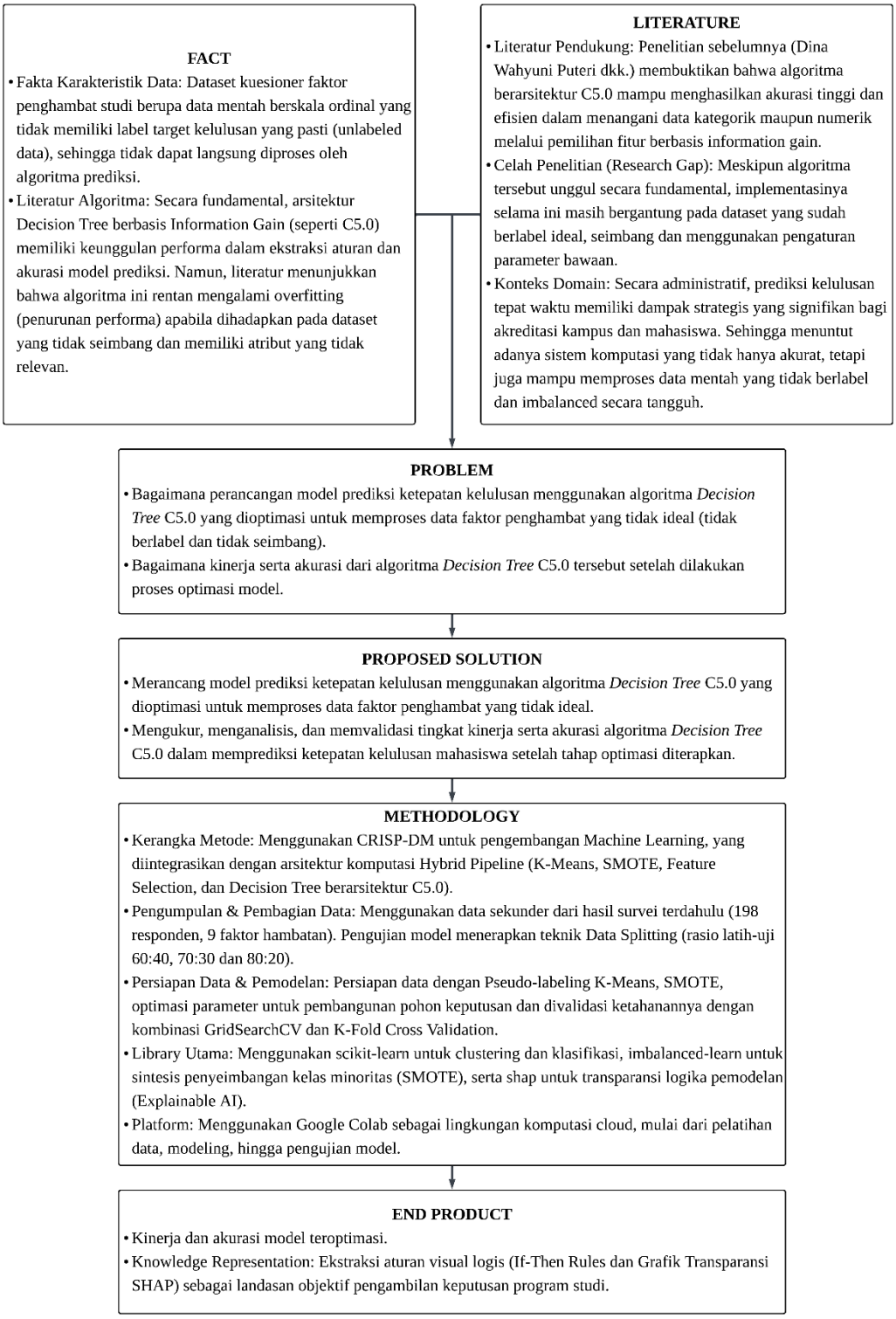
1.5 Batasan Masalah

Terdapat batasan masalah dalam penelitian ini, di antaranya:

1. Data penelitian dibatasi pada *dataset* sekunder hasil survei terdahulu "Pendataan Faktor Penghambat Ketepatan Waktu Kelulusan Studi" tahun 2023 terhadap mahasiswa Program Studi Teknik Informatika UIN Sunan Gunung Djati Bandung, berjumlah (198 responden). Rincian dari *dataset* kuesioner tersebut dilampirkan pada Lampiran A.
2. Atribut prediktor dibatasi pada 9 atribut faktor penghambat studi, mencakup faktor internal (motivasi, kesehatan, kemampuan akademis, keterlibatan organisasi, dan tantangan tugas akhir) serta faktor eksternal (pengaruh sosial, peluang kerja, fasilitas pembelajaran, dan komunikasi akademik).
3. Atribut target dibatasi pada dua kelas berupa kelas "Risiko Tinggi" dan "Risiko Rendah" yang dilabeli berdasarkan pendekatan prapemrosesan data.
4. Algoritma *machine learning* yang digunakan untuk membangun model prediksi dibatasi hanya pada algoritma *Decision Tree C5.0*.
5. Implementasi sistem secara keseluruhan dibangun menggunakan bahasa pemrograman *Python*.
6. Hasil akhir penelitian dibatasi pada evaluasi kinerja model, tanpa melakukan implementasi ke dalam bentuk perangkat lunak.

1.6 Kerangka Pemikiran

Kerangka pemikiran yang digunakan dalam penelitian ini ditampilkan pada gambar 1.2 berikut.



Gambar 1. 2 Kerangka Pemikiran

Gambar 1.2 menyajikan kerangka pemikiran yang menjadi alur utama dalam pelaksanaan penelitian ini. Penelitian ini dilatarbelakangi oleh ketiadaan perancangan model prediksi terkait kelulusan mahasiswa menggunakan algoritma *Decision Tree C5.0* teroptimasi untuk memproses *dataset* yang kurang ideal.

Dataset kuesioner yang digunakan masih berupa data mentah yang belum memiliki target kelas risiko, jumlah kelasnya tidak seimbang, dan memuat atribut yang tidak relevan. Oleh karena itu, diperlukan tahapan prapemrosesan data dan pemodelan yang tepat untuk memperbaiki masalah data tersebut, sekaligus membuktikan peningkatan akurasi model setelah dioptimasi.

Untuk memecahkan masalah komputasional tersebut sekaligus mencapai tujuan penelitian , penelitian ini mengusulkan sebuah rangkaian optimasi model prediktif. Pendekatan komputasional ini diawali dengan tahap pembagian data latih dan data uji dengan rasio pembagian 60:40, 70:30 dan 80:20. Tahap selanjutnya *Pseudo-Labeling* menggunakan algoritma *K-Means Clustering* untuk memetakan data secara objektif ke dalam kelas target, yaitu "Risiko Tinggi" dan "Risiko Rendah". Selanjutnya mengimplementasikan algoritma *SMOTE* di dalam rangkaian optimasi guna menyeimbangkan distribusi kelas melalui sintesis data buatan. Proses ini dikombinasikan dengan *Feature Selection (SelectKBest)* untuk mengurangi dimensi data dan menghapus atribut-atribut yang tidak memberikan pengaruh informasi yang berarti.

Data yang telah tervalidasi dan seimbang tersebut diproses menggunakan algoritma klasifikasi *Decision Tree* berarsitektur *C5.0*. Untuk memastikan kestabilan dan keandalan dalam komputasi, proses pelatihan model diatur secara otomatis dengan penggunaan optimasi parameter dengan *GridSearchCV*, serta divalidasi menggunakan metode *K-Fold Cross Validation*. Serta tahap evaluasi performa setiap variasi rasio model diukur menggunakan *Confusion Matrix*.

Sebagai inovasi dalam aspek transparansi dari kecerdasan buatan, penelitian ini mengintegrasikan pendekatan metode *Explainable AI (XAI)* melalui pustaka *Shapley Additive exPlanations (SHAP)*. Pendekatan ini bertujuan untuk mengurai logika di balik pemodelan, sehingga sistem dapat memaparkan kontribusi masing-masing atribut dengan jelas. Hasil akhir dari penelitian ini tidak hanya menghasilkan model prediksi *C5.0* yang akurat, tetapi juga mengekstraksi representasi pengetahuan melalui visualisasi interpretasi *SHAP*. Pengetahuan ini dapat menjadi landasan objektif bagi pihak program studi dalam menyusun kebijakan strategis untuk mencegah keterlambatan masa studi mahasiswa.

1.7 Sistematika Penulisan

Urutan penulisan pada penelitian ini adalah sebagai berikut.

BAB I **Pendahuluan**

Bab ini memaparkan secara komprehensif mengenai latar belakang masalah, rumusan masalah, tujuan dan manfaat penelitian, batasan masalah, kerangka pemikiran, serta diakhiri dengan sistematika penulisan.

BAB II **Tinjauan Pustaka**

Bab ini menguraikan landasan teori untuk melakukan penelitian dan tinjauan literatur yang berhubungan dengan penelitian yang sedang dilakukan agar membantu dalam menyelesaikan penelitian.

BAB III **Metodologi Penelitian**

Bab ini berisikan metode yang dibutuhkan dan menjelaskan rencana untuk menyelesaikan segala permasalahan dalam penelitian ini.

BAB IV **Hasil dan Pembahasan**

Bab ini memaparkan hasil yang didapat dalam melakukan penelitian dan pembahasan dari metodologi yang digunakan dalam memecahkan masalah di penelitian ini

BAB V **Kesimpulan dan Saran**

Bab ini memaparkan kesimpulan dan saran berdasarkan penelitian yang dilakukan.