

BAB II

TINJAUAN PUSTAKA

2.1 Landasan Teori

2.1.1 Perguruan Tinggi

Perguruan tinggi adalah lembaga pendidikan tinggi yang memiliki tanggung jawab besar dalam mengembangkan ilmu pengetahuan, membentuk individu yang berkeahlian, dan mempersiapkan sumber daya manusia yang kompeten dan terampil. Institusi ini tidak hanya berfungsi sebagai tempat belajar, tetapi juga sebagai pusat inovasi yang berkontribusi pada pembangunan masyarakat (Rahayu, 2019). Sebagai lembaga pengembangan ilmu, perguruan tinggi menyediakan platform untuk penelitian, pengembangan keahlian, dan pembentukan karakter generasi muda. Tujuannya adalah menciptakan masyarakat yang berpengetahuan luas, adaptif terhadap perubahan, dan mampu berkontribusi pada berbagai sektor kehidupan, baik secara nasional maupun global.

2.1.2 Kelulusan Mahasiswa

Kelulusan mahasiswa merupakan pencapaian penting dalam pendidikan tinggi dan mencerminkan keberhasilan penyelesaian suatu program studi yang memenuhi persyaratan yang ditetapkan oleh universitas. Menurut Putri et al., bahwa ketepatan waktu kelulusan merupakan sejauh mana mahasiswa menyelesaikan studinya dalam batas waktu yang ditentukan dan itu memegang peranan penting dalam dunia pendidikan (Putri et al., 2024). Hal ini tercermin dalam regulasi seperti Peraturan Badan Akreditasi Nasional Perguruan Tinggi (BAN-PT), yang menetapkan bahwa persentase lulusan tepat waktu menjadi salah satu parameter penilaian akreditasi. Mahasiswa dianggap lulus tepat waktu jika menyelesaikan beban studi sesuai ketentuan, yaitu kurang dari atau sama dengan empat tahun untuk program strata-1 (S1). Pencapaian kelulusan tepat waktu berpengaruh langsung terhadap kualitas akreditasi program studi dan reputasi universitas. Selain itu, bagi mahasiswa, kelulusan tepat

waktu berdampak positif pada efisiensi biaya pendidikan dan kesiapan memasuki dunia kerja. Sebaliknya, kelulusan yang terlambat dapat meningkatkan beban finansial, psikologis, dan sosial.

2.1.3 Data Science

Data science merupakan bidang yang menggabungkan ilmu komputer, statistika, dan pengetahuan domain bisnis untuk menganalisis dan mengekstrak wawasan dari data berukuran besar. Dengan memanfaatkan teknik komputasi, data science mampu mengolah data mentah menjadi informasi yang memiliki nilai strategis bagi organisasi atau individu. Proses ini melibatkan pengumpulan, pembersihan, dan analisis data untuk menemukan pola-pola tersembunyi yang dapat membantu pengambilan keputusan (Hairani & Amrullah, 2020). Data science tidak hanya berfokus pada aspek teknis seperti pengkodean dan analisis statistik, tetapi juga pada penerapan hasil analisis untuk mendukung pengambilan keputusan yang lebih baik.

2.1.4 CRISP-DM (Cross-Industry Standard Process Model for Data Mining)

CRISP-DM adalah metodologi yang digunakan untuk menerapkan data mining dalam berbagai industri, yang terdiri dari enam tahap utama (Suhanda et al., 2020). Tahap pertama adalah Business Understanding, yang berfokus pada pemahaman masalah bisnis yang ingin dipecahkan melalui data mining. Tujuan dari tahap ini adalah untuk menentukan tujuan proyek dan mendefinisikan masalah yang akan dianalisis, serta mengidentifikasi potensi solusi berbasis data. Setelah itu, pada tahap Data Understanding, dilakukan eksplorasi data awal untuk memahami karakteristik data yang tersedia, mengidentifikasi masalah kualitas data, serta menemukan wawasan awal yang relevan untuk analisis lebih lanjut. Tahap berikutnya adalah Data Preparation, yang melibatkan pembersihan, transformasi, dan pemrosesan data agar siap digunakan untuk pemodelan. Pada tahap Modeling, berbagai algoritma data mining diterapkan untuk membangun model prediktif atau klasifikasi berdasarkan data yang telah dipersiapkan. Kemudian tahap Evaluation dilakukan untuk menilai kinerja model dan memastikan bahwa

model yang dibangun memenuhi tujuan bisnis yang telah ditetapkan. Terakhir, pada tahap Deployment, model yang telah berhasil dievaluasi diterapkan dalam dunia nyata untuk memberikan solusi yang bermanfaat bagi bisnis, baik dalam bentuk laporan, sistem otomatisasi, atau aplikasi yang mendukung pengambilan keputusan.

2.1.5 Machine Learning

Machine Learning adalah bidang yang berkembang pesat yang berfokus pada pembuatan algoritma yang memungkinkan komputer untuk menganalisis data untuk mengidentifikasi pola dan menghasilkan prediksi atau keputusan yang lebih akurat (Louridas & Ebert, 2016). Proses ini melibatkan tiga jenis pembelajaran utama: supervised learning, unsupervised learning, dan reinforcement learning. Hanya saja, dalam penelitian yang dilakukan, seperti prediksi kelulusan mahasiswa tepat waktu, jenis machine learning yang digunakan adalah supervised learning. Metode ini memerlukan data yang sudah diberi label (labeling) sebagai input, sehingga algoritma dapat belajar dari data historis untuk memprediksi hasil.

2.1.6 Supervised learning

Supervised learning adalah jenis machine learning di mana model dilatih pada kumpulan data berlabel untuk memprediksi hasil atau mengklasifikasikan data (Sharma, 2024). Hal ini bertujuan untuk melatih model menggunakan input dan output yang diharapkan, dan kemudian menggunakan model tersebut untuk mengklasifikasikan atau memperkirakan output dengan data yang tidak terlihat. Supervised learning melibatkan algoritma yang mempelajari pemetaan input-output dari data yang sudah dilabeli sebelumnya, yang diterapkan di berbagai bidang.

2.1.7 Algoritma XGBoost

Extreme Gradient Boosting (XGBoost) adalah teknik pembelajaran mesin yang digunakan untuk menyelesaikan masalah klasifikasi. Algoritma ini menggunakan metode Gradient Boosting Decision Tree, yang menggabungkan beberapa model kecil, biasanya Decision Tree, untuk membuat model yang lebih kuat dan memberikan hasil yang lebih baik dalam akurasi (Kollongei, 2024). Algoritma ini menggunakan Decision Tree yang dibangun sekuensial dengan kedalaman yang sama, sehingga basis utamanya adalah level atau kedalaman. Menurut (Muslim Karo Karo, 2020), XGBoost memasukkan teknik seperti regularization dan loss function untuk meningkatkan efisiensi algoritma peningkatan gradient. Keunggulan utama XGBoost terletak pada kemampuan regulasinya, yaitu L1 (Lasso) dan L2 (Ridge), yang mencegah overfitting. Algoritma ini juga memiliki fitur seperti paralelisasi komputasi, pengelompokan sparsitas, dan pruning pohon untuk meningkatkan kinerja pada dataset besar dan kompleks. Cara pembuatan pohon XGBoost dilakukan sebagai berikut:

1. Menentukan nilai probability awal dari data target, class value yaitu jumlah data yang akan diolah.

$$Probability(p) = \frac{\Sigma(Class\ Value)}{\Sigma(Class)} \quad (2.1)$$

2. Menentukan nilai probability awal dari data target, class value yaitu jumlah data yang akan diolah.

$$Residual\ (Y) = Class\ Value - Probability \quad (2.2)$$

3. Membuat root awal dari classification tree dengan residual yang telah ditentukan dengan menjumlahkan semua residual tersebut. Selanjutnya, membuat leaf atau dengan mengklasifikasikan berdasarkan feature yang ada
4. Menghitung similarity atau kesamaan antara data

$$Similarity\ score = \frac{((Residual))^2}{\Sigma(px(1-p)) + \lambda} \quad (2.3)$$

Keseluruhan nilai residual dimasukkan ke dalam satu leaf yang sama dan dihitung nilai similarity score dari leaf tersebut.

5. Langkah selanjutnya adalah menghitung nilai gain dari left similarity dan right similarity

$$Gain = \frac{1}{2} \left[\frac{G_L^2}{H_L + \lambda} + \frac{G_R^2}{H_R + \lambda} - \frac{(G_L + G_R)^2}{(H_L + H_R) + \lambda} \right] - \gamma \quad (2.4)$$

Setelah perhitungan dilakukan, nilai gain tertinggi akan dipilih menjadi cabang yang memisahkan residual.

6. Kemudian dilakukan iterasi atau percabangan menggunakan pruning untuk menghitung selisih antara gain dari cabang paling bawah dari tree dengan nilai gamma yang sudah ditetapkan

$$L(\theta) = \sum_i^n (\gamma_i - \hat{\gamma})^2 \quad (2.5)$$

Apabila dalam operasi kondisi tersebut mendapatkan hasil diatas <0, maka leaf tersebut dipangkas dan data tersebut tidak digunakan lagi. Namun, jika bernilai >0 artinya leaf tidak bisa dipangkas.

7. Selanjutnya menghitung output value dari setiap leaf.

$$Output\ Value = \frac{((\sum Residual))}{\sum (P \times (1-p) + \lambda)} \quad (2.7)$$

8. Setelah diketahui output dari setiap leaf perlu dilakukan scale dengan mengalikan dengan learning rate.

$$Scale\ data(P) = \log \left(\frac{p}{1} - p \right) + (Learning\ rate \times output\ value) \quad (2.8)$$

9. Nilai learning rate biasanya dengan range 0-1.

$$Probability\ baru = \frac{e^P}{1 + e^P} \quad (2.9)$$

Selanjutnya membuat probability baru dengan menggabungkan nilai-nilai tersebut.

2.1.8 Confusion Matrix

Confusion matrix adalah alat evaluasi yang penting dalam klasifikasi model untuk menggambarkan kinerja model dengan membandingkan hasil prediksi dengan nilai aktual (Görtler et al., 2022). Setiap baris dalam Matrix Confusion menunjukkan kelas data asli dan setiap kolom menunjukkan prediksi model.

Tabel 2.1 Confusion matrix untuk prediksi model benar dan salah

	Predictive Positive	Predicted Negative
Actual Positive	True Positive (TP)	False Negative (FN)
Actual Negative	False Positive (FP)	True Negative (TN)

Jenis prediksi confusion matrix adalah sebagai berikut:

1. True Positive (TP): Prediksi benar untuk kelas positif.
2. True Negative (TN): Prediksi benar untuk kelas negatif.
3. False Positive (FP): Prediksi salah untuk kelas positif.
4. False Negative (FN): Prediksi salah untuk kelas negatif.

Kinerja model algoritma dapat dihitung sebagai berikut:

- Akurasi adalah angka yang menunjukkan seberapa sering model memprediksi kelas positif dan negatif dengan benar terhadap keseluruhan data.

$$Akurasi = \frac{TP+TN}{TP+TN+FP+FN} \quad (2.10)$$

- Presisi adalah rasio prediksi benar positif dibandingkan dengan total hasil prediksi yang positif.

$$Presisi = \frac{TP}{TP+FP} \quad (2.11)$$

- Recall adalah rasio dari prediksi benar positif dibandingkan dengan total data benar positif.

$$Recall = \frac{TP}{TP+FN} \quad (2.12)$$

- F1 Skor adalah perbandingan presisi dan recall rata-rata yang dibobotkan.

$$F1 - Score = 2x \frac{Precision \times Recall}{Precision + Recall} \quad (2.13)$$

Untuk mengevaluasi model klasifikasi dan membandingkan berbagai model untuk memilih yang terbaik, confusion matrix sangat berguna. Namun, penting untuk diingat bahwa ia hanya berfungsi jika jumlah data yang tersedia cukup besar dan mewakili populasi. Oleh karena itu, agar model dapat diuji dengan benar, pembagian data yang tersedia menjadi training set dan testing set sangat penting sebelum membuat model klasifikasi.

2.1.9 SHAP (SHapley Additive exPlanations)

SHAP (SHapley Additive exPlanations) adalah metode interpretasi model machine learning yang berbasis pada teori Shapley Values. SHAP digunakan untuk menjelaskan kontribusi masing-masing fitur input terhadap prediksi model, baik pada level global (seluruh data) maupun lokal (per individu). Metode ini memberikan nilai kontribusi (SHAP value) bagi setiap fitur, yang menggambarkan seberapa besar pengaruh fitur tersebut dalam menaikkan atau menurunkan output prediksi. Dengan pendekatan ini, SHAP bisa memahami tidak hanya fitur mana yang penting, tetapi juga bagaimana fitur tersebut mempengaruhi hasil prediksi secara positif atau negatif. Berdasarkan penelitian yang dilakukan oleh (I.U. Ekanayake, D.P.P. Meddage, n.d.), penggunaan machine learning yang konvensional memiliki kelemahan dalam menginterpretasikan variable yang dipakai oleh model, dengan adanya SHAP ini hasil prediksi model dapat dianalisis secara kuat dengan membandingkan variable pada feature importance model yang diujikan.

$$h(z') = \phi_o + \sum_{i=1}^N \phi_i z'_i \quad (2.14)$$

$$\begin{aligned}\varnothing_i &= \sum_{K \subseteq M(i)} \frac{|K|!(N - |K| - 1)!}{N!} [g_x(K \cup \{i\}) - g_x(K)] \\ g_x(K) &= E[g(x)|x_K]\end{aligned}\tag{2.15}$$

2.1.10 Python

Python adalah bahasa pemrograman tingkat yang mendukung pengembangan perangkat lunak dengan sintaksis yang jelas dan mudah dibaca. Keunggulan Python terletak pada fleksibilitasnya, memungkinkan digunakan dalam berbagai bidang termasuk pengembangan web, analisis data, kecerdasan buatan, dan lainnya yang berkaitan dengan pemrograman. Python bersifat open source, memungkinkan pengguna untuk mengakses dan memodifikasinya sesuai kebutuhan (Akbar et al., 2023).

2.1.11 Scikit-Learn (Sklearn)

Scikit-Learn adalah pustaka Python yang dirancang khusus untuk machine learning yang mencakup algoritma untuk tugas-tugas seperti klasifikasi, regresi, klustering, dan prediksi. Pustaka ini dibangun untuk kebutuhan membangun model machine learning menggunakan Python, seperti NumPy, SciPy, dan Matplotlib, sehingga mudah diintegrasikan dengan alat-alat analisis data lainnya (Ajibode et al., 2023). Scikit-Learn juga dapat melakukan fitur pre-processing data, seperti cleaning data, encoding, pelabelan data, splitting, evaluasi dan seleksi fitur, sehingga mempermudah pengembang dalam membangun model machine learning dari awal hingga akhir.

2.1.12 Google Colaboratory

Google Colaboratory, sering disebut Google Colab, adalah layanan berbasis cloud yang disediakan oleh Google untuk menjalankan kode Python langsung di browser (Putri & Nur, 2023). Colab dirancang khusus untuk kebutuhan pengembangan dan analisis data, termasuk machine learning, analisis data, dan penelitian. Dengan Colab, pengembang dapat membuat, berbagi, dan menjalankan notebook Jupyter tanpa perlu instal perangkat lunak tambahan.

2.2 Penelitian Terkait

2.2.1 State of the art

Tabel 2.2 menunjukkan hasil kajian literatur terhadap beberapa penelitian sebelumnya yang terkait dengan evaluasi kinerja model algoritma Xgboost dan model lainnya pada prediksi dan variabel input yang mempengaruhi ketepatan waktu kelulusan mahasiswa.

Tabel 2.2 State of the art

No	Judul, penulis, tahun	Metode/solusi	Hasil
1	Prediction of Student Graduation Time Using the Best Algorithm (Riyanto et al., 2019)	Neural Network (NN), support vector machin (SVM) and Decision Tree (DT) algorithms	Penelitian ini menggunakan variabel jenis kelamin, jurusan SMA, asal SMA, IPK semester 1, IPK semester 2, IPK semester 3, IPK semester 3, IPK semester 4, IPK semester 5, dan IPK semester 6. Penelitian ini berturut-turut memperoleh tingkat akurasi 85,18%, 84,96% dan 85,18% untuk SVM, DT dan NN.
2	Analisis Prediksi Kelulusan Mahasiswa Tepat Waktu Menggunakan Metode Data Mining Naïve Bayes: Systematic Review (Setiyani et al., 2020)	Metode Naïve Bayes	Penelitian ini menggunakan variabel jenis kelamin, asal dan status sekolah, pendapatan ayah, dan IPK. Penelilian ini memperoleh Tingkat akurasi sebesar 69,33%
3	Analysis of graduation prediction on time based on student academic performance using the Naïve Bayes Algorithm with	Naïve Bayes Algorithm	Penelitian ini menggunakan variabel jenis kelamin, asal daerah, asal sekolah, dan IPK. Penelitian memperoleh tingkat akurasi 70,83%.

	data mining implementation (Sembiring & Tambunan, 2021)		
4	Predicting Students Graduate on Time Using C4.5 Algorithm (Yuliansyah et al., 2021)	C4.5 Algorithm	Penelitian ini menggunakan variabel program studi, usia masuk, IPK, dan TOEPL. Penelitian memperoleh tingkat akurasi sebesar 90%
	Playing Smart with Numbers: Predicting Student Graduation Using the Magic of NaiveBayes (Mehta, 2023)	Naïve Bayes algorithm	Penelitian menggunakan variabel jenis kelamin, IPS1 – IPS8 dan IPK. Penelitian ini memperoleh tingkat akurasi sebesar 85%
5	Prediksi Ketepatan Waktu Studi Mahasiswa Bidik Misi Menggunakan K-Nearest Neighbour (Priyambodo et al., 2022)	K-Nearest Neighbour	Penelitian ini menggunakan atribut variabel jenis kelamin, nilai TOEFL, indeks prestasi semester 1 hingga semester 4, rata-rata nilai ijazah SMA, rata-rata nilai ujian nasional, program studi, nilai prestasi keaktifan organisasi. Penelitian ini memperoleh Tingkat akurasi sebesar 93.93%.
6	Prediction of Undergraduate Student's Study Completion Status Using MissForest Imputation in Random Forest and XGBoost Models (Nirmala et al., 2022)	Random Forest and XGBoost Models	Penelitian ini menggunakan variabel jumlah SKS per semester, usia masuk universitas, IPK, dan IPS. Penelitian ini memperoleh tingkat akurasi 92,33% dan 92,18%.

7	Machine Learning Algorithms with Parameter Tuning to Predict Students' Graduation-on-time: A Case Study in Higher Education (Bakri et al., 2022)	Random Forest, Support Vector Machine (Linear Kernel), Support Vector Machine (Polynomial Kernel), K-Nearest Neighbors, and Naïve Bayes.	Penelitian ini menggunakan variabel usia masuk universitas, IPS1-4, program studi, dan pendapatan orangtua. Penelitian ini memperoleh tingkat akurasi 86,7%, 85,7%, 85,9%, 83,2% dan 85,6%.
8	Student Graduation Time Prediction Using Logistic Regression, Decision Tree, Support Vector Machine, And Adaboost Ensemble Learning (Desfiandi & Soewito, 2023)	Logistic Regression (LR), Decision Tree (DT), Support Vector Machine (SVM)	Penelitian ini menggunakan variabel pendapatan orangtua, IPS2, IPS3, IPS4, IPS5. Penelitian ini berturut-turut memperoleh tingkat akurasi 75%, 70% dan 76% untuk LR, DT dan SVM
9	Prediksi Kelulusan Mahasiswa Tepat Waktu Menggunakan Algoritma C4.5 Pada Uin Syarif Hidayatullah Jakarta (Hasibuan & Mahdiana, 2023)	Decision Tree Classifier (C4.5)	Penelitian ini menggunakan variabel jurusan, jenis seleksi, tahun angkatan, kelamin, provinsi, asal sekolah, jurusan sekolah, ipsmt1, ipsmt2, dan ipsmt3. Penelitian memperoleh tingkat akurasi kelulusan mahasiswa sebesar 75,52%.
10	Evaluating Machine Learning Models for Predicting Graduation Timelines in	Support Vector Machines, Decision	Penelitian ini menggunakan variabel usia, status perkawinan, jenis kelamin, asal daerah, pendapatan keluarga, jenis sekolah. Penelitian memperoleh tingkat

	Moroccan Universities (Sadqui et al., 2023)	Tree, Naive Bayes, Logistic Regression, and Random Forest	akurasi sebesar 69%, 80%, 74%, 78% dan 82% untuk SVM, decision tree, naïve bayes, logistic regression, dan random forest.
11	Prediction Of Student Graduation Using The K-Nearest Neighbor Method Case Study in Politeknik Negeri Tanah Laut (Sari et al., 2023)	K-Nearest Neighbor Method	Penelitian ini hanya menggunakan variabel IPS1 – IPS5 dan memperoleh tingkat akurasi sebesar 91.67%
12	Comparison of C4.5 and Naive Bayes Algorithm Methods in Prediction of Student Graduation on Time (Gerhana et al., 2019)	C4.5 and Naive Bayes Algorithm	Penelitian ini hanya menggunakan variabel IPK, dan TOEFL dan memperoleh tingkat akurasi sebesar 88,7% dan 87,6%.
13	Profiles of University Students Who Graduate on Time: A Cohort Study from the Chilean Context (Moraga-Pumarino et al., 2023)	Classification Tree (CART)	Penelitian ini menggunakan nilai indeks prestasi SMA, score uji matematika dan Bahasa, jenis SMA, jenis kelamin, pendidikan orangtua, IPS1, IPS2, score matematika S1 dan S2, status kerja, jarak tempat tinggal. Penelitian ini memperoleh 79,5%
14	A neuro-fuzzy model for predicting and analyzing student graduation performance in computing programs (Mehdi & Nachouki, 2023)	neuro-fuzzy model	Penelitian hanya menggunakan IPS1-IPS6 dan memperoleh tingkat akurasi sebesar 77%

15	Predicting graduation grades using Machine Learning: A case study of Can Tho University students (Nguyen et al., 2023)	K- nearest neighbor, Neural network, Decision tree, Random forest, and Gradient boosting	Penelitian hanya menggunakan jenis kelamin, usia, program studi, IPK, dan jumlah SKS dan memperoleh tingkat akurasi sebesar 80,2%, 77,8%, 85%, 87,2% dan 86,1% untuk KNN, Neural Network, decision tree, random forest, dan gradient boosting
16	Student Graduation Prediction Using Decision Tree Method with C4.5 Algorithm (Sarbaini & Ulfa, 2024)	C4.5 Algorithm	Penelitian ini menggunakan variabel jenis kelamin, asal SMA, pendapatan orangtua, IP dan jumlah SKS persemester. Penelitian memperoleh tingkat akurasi kelulusan mahasiswa sebesar 78,7%.
17	Klasifikasi Ketepatan Waktu Lulus Mahasiswa Jurusan Matematika Universitas Pattimura Menggunakan Metode Naïve Bayes Classifier (Tomu et al., 2024)	Naïve Bayes Classifier	Penelitian ini hanya menggunakan atribut variabel usia, jenis kelamin, jalur penerimaan masuk universitas dan IPK dan memperoleh tingkat akurasi sebesar 70%.
18	Prediksi Kelulusan Tepat Waktu Berdasarkan Riwayat Akademik Menggunakan Metode Naïve Bayes (Imam Riadi et al., 2024)	Metode Naïve Bayes	Penelitian ini menggunakan atribut variabel jenis sekolah, provinsi, program studi, rata-rata matematika, IPK dan nilai TOEFL dan memperoleh Tingkat akurasi sebesar 72%

19	Data Mining Model Klasifikasi Untuk Ketepatan Waktu Kelulusan Mahasiswa (Rafelin et al., 2024)	Algoritma Naïve Bayes	Penelitian ini menggunakan variabel asal sekolah, jenis kelamin, jurusan sekolah, umur masuk universitas, jalur masuk universitas, pendapatan orangtua, IPK. Penelitian memperoleh Tingkat akurasi 75%
20	Predicting Time to Graduation of Open University Students: An Educational Data Mining Study (Santoso et al., 2024)	Random forest and neural networks algorithms	Penelitian ini menggunakan variabel usia mahasiswa, jenis kelamin, asal daerah, IPK dan penelitian memperoleh tingkat akurasi sebesar 76%
21	Algorithmic Prediction of Students On-Time Graduation from the University (Alfahid, 2024)	SVM and Random Forest	Penelitian ini menggunakan variabel jenis kelamin, IPK SMA, hasil uji masuk universitas, IPK, IPS dan program studi. Penelitian ini memperoleh tingkat akurasi sebesar 83% dan 83,1% untuk SVM and Random Forest
22	Predicting time to graduation at a large enrollment American university (Aiken et al., 2020)	Logistic regression dan Gradient boosted trees	Penelitian ini menggunakan variabel jenis kelamin, IPK SMA, pendapatan orangtua, jumlah SKS/semester, IPK, IPS dan program studi. Penelitian ini memperoleh tingkat akurasi sebesar 84.2% dan 86,2% untuk LR and Gradient boosted
23	Implementation of Data Mining for Predicting Student Graduation Using the K-Nearest Neighbor Algorithm at Jambi Muhammadiyah University (Amandha et al., 2024)	K-Nearest Neighbor Algorithm	Penelitian ini menggunakan variabel jenis kelamin, IPS1 – IPS6. Penelitian ini memperoleh tingkat akurasi sebesar 83.33%

24	Implementation of Random Forest Algorithm for Graduation Prediction (Riskiyono & Mahdiana, 2024)	Random Forest Algorithm.	84.03%
25	Predicting student's dropout in university classes using two-layer ensemble machine learning approach: A novel stacked generalization (Niyogisubizo et al., 2022)	Random Forest, Gradient Boosting, dan XGBoost	Penelitian ini menghasilkan masing-masing persentase akurasi sebesar 91.66%, 86.66%, dan 91.66%.

2.2.2 Matriks Penelitian

Tabel 2.3 Matriks penelitian

Judul	Permasalahan	Indikator
Analisis Feature Importance Pada Prediksi Kelulusan Mahasiswa Menggunakan Algoritma Extreme Gradient Boosting (Xgboost)	1. Seberapa akurat model XGBoost dalam memprediksi kelulusan mahasiswa? 2. Apa saja faktor-faktor yang mempengaruhi ketepatan waktu kelulusan mahasiswa?	a. Evaluasi model berdasarkan metrik seperti akurasi, precision, recall, dan F1-score. b. Data akademik terdiri dari penghasilan orang tua, pendidikan orang tua, tanggungan, jenis kelamin, program studi, fakultas, dan IP semester 1-8.
Sumber Data	Metode Penelitian	
UPT TIK Universitas Siliwangi	Pendekatan kuantitatif dalam penelitian ini menggunakan algoritma XGBoost untuk menganalisis faktor-faktor yang memengaruhi kelulusan mahasiswa tepat waktu. Algoritma XGBoost dipilih karena kemampuannya dalam menangani data yang kompleks dan mencegah overfitting melalui pengoptimalan parameter yang mempengaruhi.	

2.3 Kebaharuan Penelitian

Penelitian ini memiliki beberapa kebaruan signifikan. Berdasarkan hasil investigasi dalam beberapa penelitian sebelumnya pada Tabel 2.2 yang menerapkan model XGBoost dan lainnya, atribut input dan jumlah data memiliki pengaruh besar dalam prediksi kelulusan kuliah tepat waktu mahasiswa. Oleh karena itu dalam penelitian ini, pertama mengusulkan penggunaan pendekatan fitur lebih umum (Universal) dalam prediksi kelulusan kuliah tepat waktu mahasiswa untuk mencakup area probabilitas yang lebih luas, dengan memanfaatkan metode machine learning XGBoost. Kedua mengusulkan penggunaan *feature importance* dalam mengoptimalkan pemilihan fitur untuk menyederhanakan model sehingga fitur mana saja yang berpengaruh terhadap proses analisis. Matriks perbedaan penelitian terdahulu dengan penelitian yang dilakukan, yang dapat dilihat pada tabel 2.4 sebagai berikut.

Tabel 2.4 Perbandingan Penelitian

No	Jurnal	Perbandingan penelitian
1.	Prediction of Student Graduation Time Using the Best Algorithm (Riyanto et al., 2019)	Terdahulu: Memprediksi menggunakan 3 algoritma yaitu NN, DT dan SVM dengan menggunakan variabel jenis kelamin, jurusan SMA, asal SMA, IPK semester 1, IPK semester 2, IPK semester 3, IPK semester 3, IPK semester 4, IPK semester 5, dan IPK semester 6. Baru: Menggunakan XGBoost yang menganalisis kontribusi setiap fitur secara detail melalui feature importance, meningkatkan interpretabilitas model.
2.	Analisis Prediksi Kelulusan Mahasiswa Tepat Waktu Menggunakan Metode Data	Terdahulu: Penelitian ini menggunakan variabel jenis kelamin, asal dan status sekolah, pendapatan

	Mining Naïve Bayes: Systematic Review (Setiyani et al., 2020)	ayah, dan IPK. Penelitian ini memperoleh Tingkat akurasi sebesar 69,33% Baru: Pendekatan XGBoost menghasilkan prediksi lebih akurat dan mendalam melalui analisis pengaruh tiap fitur secara menyeluruh.
3.	Analysis of graduation prediction on time based on student academic performance using the Naïve Bayes Algorithm with data mining implementation (Sembiring & Tambunan, 2021)	Terdahulu: Penelitian ini menggunakan variabel jenis kelamin, asal daerah, asal sekolah, dan IPK. Penelitian memperoleh tingkat akurasi 70,83%. Baru: XGBoost memanfaatkan kekuatan pemrosesan fitur kompleks dan disertai analisis SHAP untuk pemahaman individu terhadap hasil prediksi.
4.	Predicting Students Graduate on Time Using C4.5 Algorithm (Sarbaini & Ulfa, 2024)	Terdahulu: Penelitian ini menggunakan variabel program studi, usia masuk, IPK, dan TOEPL. Penelitian memperoleh tingkat akurasi sebesar 90% Baru: Meskipun akurasi tinggi, model XGBoost unggul dalam interpretasi fitur menggunakan SHAP untuk analisis penyebab kelulusan.
5.	Playing Smart with Numbers: Predicting Student Graduation Using the Magic of NaiveBayes (Mehta, 2023)	Terdahulu: Penelitian menggunakan variabel jenis kelamin, IPS1 – IPS8 dan IPK. Penelitian ini memperoleh tingkat akurasi sebesar 85%. Baru: XGBoost digunakan untuk menyajikan visualisasi dampak tiap fitur terhadap prediksi dengan SHAP, meningkatkan transparansi.

6.	Prediksi Ketepatan Waktu Studi Mahasiswa Bidik Misi Menggunakan K-Nearest Neighbour (Priyambodo et al., 2022)	Terdahulu: Penelitian ini menggunakan atribut variabel jenis kelamin, nilai TOEFL, indeks prestasi semester 1 hingga semester 4, rata-rata nilai ijazah SMA, rata-rata nilai ujian nasional, program studi, nilai prestasi keaktifan organisasi. Penelitian ini memperoleh Tingkat akurasi sebesar 93.93%. Baru: XGBoost dikombinasikan dengan SHAP dan feature importance, menghasilkan model yang adaptif dan interpretatif terhadap latar belakang mahasiswa Bidik Misi.
7.	Prediction of Undergraduate Student's Study Completion Status Using MissForest Imputation in Random Forest and XGBoost Models (Priyambodo et al., 2022)	Terdahulu: Penelitian ini menggunakan variabel jumlah SKS per semester, usia masuk universitas, IPK, dan IPS. Penelitian ini memperoleh tingkat akurasi 92,33% dan 92,18%. Baru: Penelitian baru memanfaatkan XGBoost sepenuhnya dengan analisis fitur untuk validasi interpretasi dan efisiensi model.
8.	Machine Learning Algorithms with Parameter Tuning to Predict Students' Graduation-on-time: A Case Study in Higher Education (Desfiandi & Soewito, 2023)	Terdahulu: Penelitian ini menggunakan variabel usia masuk universitas, IPS1-4, program studi, dan pendapatan orangtua. Penelitian ini memperoleh tingkat akurasi 86,7%, 85,7%, 85,9%, 83,2% dan 85,6%. Baru: XGBoost dilengkapi SHAP menjelaskan pola akademik-ekonomi mahasiswa secara rinci dan strategis bagi institusi.

9.	Prediksi Kelulusan Mahasiswa Tepat Waktu Menggunakan Algoritma C4.5 Pada Uin Syarif Hidayatullah Jakarta (Hasibuan & Mahdiana, 2023)	Terdahulu: Penelitian ini menggunakan variabel jurusan, jenis seleksi, tahun angkatan, kelamin, provinsi, asal sekolah, jurusan sekolah, ipsmt1, ipsmt2, dan ipsmt3. Penelitian memperoleh tingkat akurasi kelulusan mahasiswa sebesar 75,52%. Baru: XGBoost memungkinkan penelusuran kontribusi fitur seperti jalur seleksi dan prestasi akademik secara lebih spesifik.
10.	Evaluating Machine Learning Models for Predicting Graduation Timelines in Moroccan Universities (Sadqui et al., 2023)	Terdahulu: Penelitian ini menggunakan variabel usia, status perkawinan, jenis kelamin, asal daerah, pendapatan keluarga, jenis sekolah. Penelitian memperoleh tingkat akurasi sebesar 69%, 80%, 74%, 78% dan 82% untuk SVM, decision tree, naïve bayes, logistic regression, dan random forest. Baru: XGBoost menawarkan prediksi yang tidak hanya akurat, tetapi juga transparan dan dapat dijelaskan menggunakan SHAP.

Adapun hasil dari penelitian ini, memberikan kontribusi sebagai berikut:

1. Kontribusi Teoritis:

Penelitian ini tidak hanya memfokuskan pada prediksi semata, tetapi juga pada interpretabilitas model melalui analisis feature importance. Dengan memanfaatkan algoritma XGBoost yang memiliki kemampuan menangani data kompleks dan menghasilkan skor pentingnya tiap fitur, penelitian ini memberikan pendekatan terintegrasi antara predictive modeling dan explainable AI (XAI). Penemuan bahwa fitur sosial ekonomi seperti penghasilan orang tua dan status pekerjaan berpengaruh

besar terhadap ketepatan waktu kelulusan memberikan kontribusi baru dalam literatur akademik, terutama dalam penggunaan machine learning untuk konteks pendidikan tinggi.

2. Kontribusi Praktis:

Model XGBoost yang digunakan bersifat high-performance dan mampu memberikan visualisasi pentingnya fitur menggunakan teknik seperti SHAP. Hal ini memungkinkan pihak lembaga untuk lebih memahami faktor-faktor risiko dari setiap mahasiswa dan melakukan intervensi secara lebih terarah. Misalnya, mahasiswa dengan latar belakang ekonomi rentan dapat diberikan dukungan tambahan untuk meningkatkan kemungkinan kelulusan tepat waktu.

3. Kontribusi Institusional.

Dengan pendekatan berbasis data yang bisa diterapkan ulang (replicable), penelitian ini bisa dijadikan rujukan oleh institusi lain dalam membangun sistem pemantauan akademik berbasis machine learning. Ini menjadikan hasil penelitian tidak hanya bermanfaat untuk satu institusi, tetapi bisa diadaptasi secara luas dengan penyesuaian data yang minimal, memperkuat fungsi decision support system di ranah pendidikan tinggi.