BAB II
TINJAUAN PUSTAKA DAN LANDASAN TEORI

2.1 Tinjauan Pustaka

Bagian ini merupakan bagian analisa dari penelitian-penelitian sebelumnya yang relevan dengan topik penelitian ini. Tujuan dari Bab ini adalah untuk menunjukkan pengetahuan dan pemahaman yang mendalam tentang kerangka konseptual dan teoritis yang mendukung penelitian ini. Selain itu pada bagian ini juga disampaikan teori-teori yang digunakan untuk mendukung penelitian.

2.1.1. Penelitian Terkait

Amrieh dkk (2016) mengusulkan model baru untuk prediksi performa siswa berdasarkan teknik data mining dengan atribut/fitur data baru, yang disebut fitur perilaku siswa. Jenis fitur ini terkait interaksi pelajar dengan sistem manajemen e-*learning* yang diperoleh dari data mahasiswa semester satu dan dua. Metode yang digunakan adalah *ensemble* yaitu menggunakan sekumpulan model, kemudian menggabungkannya untuk mengambil yang terbaik. Algoritma klasifikasi yang digunakan adalah (ANN, NB and DT) dan menggunakan metode *ensemble* (Bagging, Boosting dan RF). Pada penelitian ini disampaikan bahwa fitur perilaku sumber yang diakses merupakan fitur yang paling efektif dalam memprediksi performa siswa. Metode *ensembel* DT dengan Boosting merupakan metode terbaik dengan tingkat akurasi sebesar 82,2%.

Aluko dkk. (2018) mengusulkan prediksi performa akademik mahasiswa dengan memanfaatkan informasi yang terkandung dalam nilai akademik. IPK digunakan sebagai metrik untuk mengukur performa akademik mahasiswa. IPK dibagi menjadi 2 (lulus/gagal) IPK 2,4 lulus dan IPK di bawahnya gagal. Algoritma yang digunakan adalah LR dan SVM. Model SVM mengungguli model LR dalam hal akurasi. Hasil penelitian juga menunjukkan bahwa nilai akademik merupakan prediktor yang baik untuk performa akademik mahasiswa.

Hellas dkk (2018) mengusulkan model klasifikasi yang berbeda untuk memprediksi performa siswa, menggunakan data yang dikumpulkan dari data

pendaftaran mahasiswa dan data aktivitas yang dihasilkan dari sistem manajemen pembelajaran universitas (LMS) periode 2011 sampai dengan 2013. Data pendaftaran berisi informasi siswa seperti fitur sosiodemografi, dasar masuk universitas dan jenis kehadiran. Data LMS merekam keterlibatan siswa dengan aktivitas pembelajaran online mereka. Kontribusi penting dari penelitian ini adalah pertimbangan heterogenitas siswa dalam membangun model prediksi. Hal ini didasarkan pada pengamatan bahwa siswa dengan fitur sosiodemografi atau model belajar yang berbeda dapat menunjukkan motivasi belajar yang berbeda-beda. Lebih lanjut, eksperimen mengungkapkan bahwa mempertimbangkan fitur pendaftaran dan aktivitas perkuliahan dapat membantu mengidentifikasi siswa yang rentan secara lebih tepat. Empat algoritma digunakan untuk memprediksi performa akademik mahasiswa yaitu NB, J48, SMO, JRip. Eksperimen mengungkapkan bahwa tidak ada metode tunggal yang menunjukkan kinerja unggul dalam semua aspek untuk memprediksi performa siswa. Kombinasi J48 dan JRip, berkontribusi secara signifikan dengan menghasilkan output yang dapat dipahami masing-masing dalam bentuk pohon dan aturan. Selanjutnya, dengan mempertimbangkan fitur gabungan, diperoleh hasil prediksi yang lebih unggul dalam mengidentifikasi siswa yang tidak berhasil dibandingkan dengan mempertimbangkan fitur secara terpisah.

Ramaswami dkk (2019) mengumpulkan data interaksi siswa dengan LMS (Xorro-Q) selama satu semester dan satu mata kuliah, data dibagi dua yaitu partisipasi di dalam kelas dan diluar kelas. Selanjutnya mereka membandingkan algoritma NB, LR, KNN dan RF untuk menemukan algoritma yang paling baik dalam hal akurasi untuk memprediksi performa akademik mahasiswa. Hasil penelitian memperlihatkan bahwa fitur partisipasi diluar kelas memberikan dampak yang baik namun tidak signifikan, selain itu dari hasil penelitian ini algoritma Random Forest adalah algoritma terbaik dari algoritma lainnya.

Algoritma untuk memprediksi performa akademik mahasiswa berbeda-beda antar peneliti oleh karena itu (Arifin dkk. 2021) membandingkan beberapa algoritma yang sering digunakan dalam memprediksi performa akademik mahasiswa. Mereka membandingkan 5 algoritma GLM, RF, DT, GBT, DL, dan SVM. Data yang dipergunakan merupakan kumpulan data yang terdiri dari data

keuangan, data sosial, dan data akademik dari 12.411 siswa. Hasil percobaan menunjukkan bahwa GBT adalah algoritma yang paling efektif dalam memprediksi performa akademik mahasiswa dengan nilai RMSE paling rendah. Algoritma terbaik diperlukan dalam memprediksi performa akademik. Sebuah perbandingan diperlukan sebelum menentukan algoritma yang akan digunakan. Beberapa peneliti menyatakan bahwa algoritma terbaik dari penelitian mereka bervariasi. Tidak ada satu algoritma yang terbaik dalam memprediksi performa akademik siswa. Hal ini dipengaruhi oleh data yang diperoleh dan proses prapengolahan. Oleh karena itu, peneliti perlu membandingkan beberapa algoritma sebelum dipergunakan dalam memprediksi performa akademik.

2.1.2. Fitur-fitur Untuk Prediksi Performa Akademik Mahasiswa

Fitur-fitur untuk memprediksi performa akademik mahasiswa antara peneliti satu dan peneliti lainnya tidak memiliki kesamaan, namun terdapat beberapa irisan dalam pemilihan fitur yang digunakan. Proses pencarian fitur yang memiliki pengaruh terhadap prediksi performa akademik merupakan salah satu pembeda dan akan menjadi temuan baru dari peneliti. Ramaswami dkk (2019) mengumpulkan data interaksi siswa dari LMS (Xorro-Q) selama satu semester dan satu mata kuliah, data dibagi dua yaitu partisipasi di dalam kelas dan diluar kelas. Fitur perilaku siswa yang merupakan fitur terkait interaksi mahasiswa dengan LMS telah diusulkan (Amrieh dkk. 2016). Yağcı (2022) menggunakan fitur ujian tengah semester dan ujian akhir semester dari mahasiswa yang mengambil mata kuliah bahasa. Arun dkk (2018) menggunakan fitur dari Nilai dan Status Quiz1, FinalTest1, FinalTest2. Helal dkk (2018) mengusulkan model klasifikasi yang berbeda untuk memprediksi performa mahasiswa, menggunakan data yang dikumpulkan dari data pendaftaran mahasiswa dan data aktivitas yang dihasilkan dari sistem manajemen pembelajaran universitas (LMS). Aluko dkk (2018) mengusulkan prediksi performa akademik mahasiswa dengan variabel masukannya adalah nilai yang diperoleh pada ujian O-level (11 mata pelajaran), cara masuk (JAMB/Entri Langsung) dan skor nilai ujian tersier terpadu (JAMB).

2.1.3. Keaslian Penelitian

Dari tinjauan pustaka yang disampaikan, dapat ditunjukkan penelitian utama yang mendasari penelitian ini yang telah dilakukan oleh beberapa penelitian lain, dapat dilihat pada Tabel 2.1.

Tabel 2. 1 Penelitian yang melandasi keaslian dari penelitian

No	Judul, Penulis, dan Tahun	Tujuan	Metode	Hasil	Catatan
1	"***Mining Educational Data to Predict Student's academic Performance using Ensemble Methods***", Amrieh dkk., 2016 (Q3)	Mengusulkan **model baru** untuk prediksi performa siswa berdasarkan teknik data mining dengan **atribut/fitur data baru**, yang disebut fitur perilaku siswa. Jenis fitur ini terkait interaksi pelajar dengan sistem manajemen e-learning	Mengusulkan model performa siswa dengan menggunakan metode ensembel. Algoritma klasifikasi yang digunakan adalah (ANN, NB and DT) dan menggunakan metode ***ensemble*** (Bagging, Boosting dan RF) **Tiga fitur kategori utama**: (1) fitur demografis. (2) Fitur latar belakang akademik. (3) Fitur perilaku.	Hasil yang diperoleh mengungkapkan bahwa ada **hubungan yang kuat antara perilaku peserta didik dan performa akademik mereka**. Fitur **sumber yang akses** adalah fitur perilaku yang paling efektif pada model performa siswa. Setelah proses pelatihan selesai, model prediktif diuji menggunakan siswa pendatang baru yang tidak berlabel, akurasi yang dicapai lebih dari 80%.	Penelitian selanjutnya akan lebih fokus dalam **menganalisa fitur-fitur** lainnya. Metode ensembel lebih akurat dalam memprediksi performa siswa namun metode ini tidak cocok jika diaplikasikan secara realtime karena membutuhkan sumberdaya yang besar.
2	"***Towards reliable prediction of academic performance of architecture students using data mining techniques***", Aluka dkk., 2018 (Q2)	1. Menyelidiki efisiensi teknik data mining untuk memprediksi performa akademik mahasiswa berdasarkan informasi yang terkandung dalam **nilai akademik** 2. Mencari pengaruh nilai akademik terhadap performa mahasiswa	Mengambil data dari data akademik selanjutnya menggunakan algoritma data mining Logistic Regression dan Support Vector Machine (SVM) Data dari mahasiswa tingkat awal, skor ujian masuk yang diperoleh dari data akademik. IPK digunakan sebagai metrik untuk mengukur performa mahasiswa. IPK **dibagi menjadi 2 (lulus/gagal) Ipk 2,4 Lulus dan ipk dibawahnya gagal** dengan skala 0-5 Tool *R* dan *Rminer*	Model SVM mengungguli model regresi logistik dalam hal akurasi. Secara bersama-sama, terbukti bahwa nilai akademik merupakan prediktor yang baik untuk performa akademik mahasiswa **Nilai akademik merupakan prediktor yang baik untuk performa akademik mahasiswa**.	Penelitian ini hanya memanfaatkan data nilai akademik siswa serta skor ujian masuk untuk memprediksi performa akademik mahasiswa. Metode SVM lebih unggul dibandingkan Logistic Regression **Nilai akademik sangat berpengaruh kepada prediksi performa siswa**

No	Judul, Penulis, dan Tahun	Tujuan	Metode	Hasil	Catatan
4	"***Predicting academic performance by considering student heterogeneity***", Helal dkk. 2018 (Q1)	Mengusulkan model klasifikasi yang berbeda untuk memprediksi performa siswa. Dengan menggunakan data yang diperoleh dari LMS-Moodle. Mengusulkan **sub populasi siswa berdasarkan fitur demografis dan akademik untuk membangun sub-model siswa**, dan mengevaluasi kegunaannya dalam mengidentifikasi siswa yang rentan.	Melakukan eksperimen dengan empat metode klasifikasi (Naïve Bayes, J48, SMO, Jrip) yang berbeda untuk memvalidasi efektivitas pendekatan yang diusulkan. (Menggunakan tool WEKA) Membagi data kedalam sub-sub dan menguji kedalam sub-sub model klasifikasi Dengan mempertimbangkan fitur **demografis siswa, akademik, dan aktivitas pembelajaran** secara terpisah serta bersama-sama untuk membantu identifikasi siswa "berisiko".	Didapatkan hasil: Efektivitas penggunaan sub-populasi siswa dalam memprediksi performa akademik siswa. Hasil ini menunjukkan bahwa, meskipun tidak semua sub-model tingkat kedua mencapai prediksi yang unggul, beberapa masih dapat memberikan wawasan tentang performa siswa, dan dengan demikian membantu dalam desain dukungan siswa yang lebih bertarget. **Fitur pendaftaran dan aktivitas pembelajaran** membantu dalam mengidentifikasi siswa yang rentan secara lebih tepat.	Penelitian ke depan sebaiknya mengidentifikasi indikator risiko siswa internasional, karena mereka mungkin memiliki beberapa fitur yang berbeda dari siswa domestik, seperti asal etnis yang beragam, peluang pendanaan, bahasa asli, dan faktor lainnya. Kedua, akan sangat berguna untuk **mempertimbangkan fitur gabungan** untuk modul tertentu dan fitur kategorisasi (misalnya sosial dan informasi) **dalam hal partisipasi siswa dalam kegiatan LMS.**
5	"***Supervised data mining approach for predicting student performance***", Wan dkk, 2019 (Q3)	Mengembangkan model prediksi menggunakan algoritma klasifikasi untuk memprediksi performa mahasiswa di universitas. Model prediksi yang dikembangkan dapat digunakan untuk mengidentifikasi **atribut terpenting dalam data**	Teknik pemodelan prediktif model K-Nearest Neighbor, Naïve Bayes, Decision Tree dan Logistic Regression Model digunakan untuk memprediksi performa siswa apakah sangat baik atau tidak baik. Validasi model dilakukan dengan 10-fold cross-validation untuk memvalidasi setiap pengklasifikasi. Ketika tes selesai, performa rata-rata pada tes dihitung untuk menentukan keakuratan model yang dikembangkan	**Akurasi Naïve Bayes** mengungguli algoritma klasifikasi lainnya. Naïve Bayes mengungkapkan bahwa faktor yang paling signifikan berkontribusi terhadap prediksi siswa yang sangat baik adalah ketika nilai siswa A+ dan A dalam Analisis Multivariat; A+, A dan A- di Pemrograman SAS dan A, A- dan B+ di ITS 472.	Beberapa atribut pengaruh tinggi untuk memprediksi performa mahasiswa dapat dipertimbangkan oleh universitas untuk merencanakan tindakan lebih lanjut untuk perbaikan. Studi ini dapat diperluas lebih lanjut untuk memprediksi **performa mahasiswa dengan menggunakan atribut lain.**

No	Judul, Penulis, dan Tahun	Tujuan	Metode	Hasil	Catatan
6	***"Predicting students' final degree classification using an extended profile"***, Al-Sudani & Palaniappan, 2019 (Q1)	**Mengkombinasi faktor institusional, akademik, demografi, dan ekonomi** untuk memprediksi performa siswa dengan mengklasifikasikan gelar siswa ke dalam kelas baik atau dasar.	Model NN juga dibandingkan dengan pengklasifikasi lain khususnya k-Nearest Neighbour, Decision Tree dan Support Vector Machine pada dataset yang sama menggunakan fitur yang sama.	Hasilnya menunjukkan bahwa NN mengungguli semua pengklasifikasi lain dalam hal akurasi klasifikasi secara keseluruhan dan menunjukkan metode yang akan digunakan dalam usaha memprediksi performa mahasiswa di universitas secara otomatis.	Seperangkat faktor kelembagaan tambahan seperti program studi dan jenis biaya dan faktor lain seperti alamat dan jenis kelamin adalah digunakan pada tahap awal analisis, tetapi kemudian **dihapus karena redundansi atau korelasi yang lemah dengan model prediksi.**
7	***"Predicting student performance in higher education using multi-regression models",*** Leo dan Yulia. 2020 (Q3)	Membangun sistem yang dirancang dapat mengelompokkan siswa ke dalam kelompok-kelompok yang model prediksinya relatif sama. Dengan menjelajahi kelompok siswa ini, pengetahuan tentang faktor-faktor yang menentukan performa siswa.	Pengambilan data dari LMS Moodle dibagi kedalam 3 kategori data siswa, aktivitas dan interaksi dengan LMS. Metode prediksi membandingkan antara single linear regression dengan multi-regression	Berdasarkan hasil pengujian, model regresi berganda lebih baik dalam menjelaskan variabel dependen daripada regresi linier tunggal. Apalagi dengan bertambahnya jumlah model regresi linier, RMSE cenderung menurun secara bertahap.	**Interaksi dengan LMS dapat meningkatkan akurasi prediksi performa siswa** Penerapan multi regression lebih akurat dibandingkan dengan single linear regression. Data yang digunakan merupakan data data siswa dan aktivitas interaksi dengan LMS
8	***"Predicting academic success in higher education: literature review and best practices",*** Alyahyan, E., & Düştegör, D. 2020 (Q1)	Studi ini menyajikan serangkaian pedoman yang jelas untuk diikuti dalam menggunakan EDM dalam memprediksi keberhasilan siswa	Merumuskan EDM Framework dalam memprediksi siswa sukses berdasarkan literatur, terdapat 6 langkah yaitu: 1. Data Collection 2. Initial Preparation 3. Statistical Analysis 4. Data Preprocessing 5. Data Mining 6. Evaluation	Menyajikan model pengukuran/evaluasi dalam melakukan prediksi akademik siswa Keberhasilan akademik siswa diukur dari prestasi akademik, hasil penelitian menunjukkan **bahwa prestasi akademik sebelumnya**, demografi siswa, aktivitas e-learning, atribut psikologis, adalah faktor yang paling umum dilaporkan	Sementara temuan yang dilaporkan didasarkan pada literatur (misalnya definisi potensi keberhasilan akademik, fitur untuk mengukurnya, faktor penting), data tambahan yang tersedia dapat dengan mudah dimasukkan dalam analisis, termasuk data fakultas (misalnya kompetensi, kriteria rekrutmen, kualifikasi akademik) mungkin untuk menemukan determinan baru.

No	Judul, Penulis, dan Tahun	Tujuan	Metode	Hasil	Catatan
9	***"Comparative Analysis on Educational Data Mining Algorithm to Predict Academic Performance"*** Arifin dkk., 2021 (Prosiding IEEE)	**Menemukan algoritma data mining terbaik** dari beberapa algoritma yang sering digunakan oleh peneliti-peneliti sebelumnya dalam memprediksi performa akademik. Menemukan fitur yang paling berpengaruh terhadap performa akademik mahasiswa	Proses pengambilan data dari data repository, selanjutnya mereview data untuk dilakukan pemilihan fitur sebelum diaplikasikan pada algoritma prediksi. Membandingkan algoritma *Generalized linear model (GLM), Deep learning (DL), Decision tree (DT), Random Forest (RF), Gradient boosted trees (GBT), Support vector machine (SVM)* Tool yang digunakan adalah Rapid Miner	GBT merupakan algoritma yang paling efektif dalam memprediksi performa siswa dibandingkan dengan algoritma lainnya. Selain itu, menggunakan beberapa algoritma dalam perangkat lunak untuk mengidentifikasi akurasi prediksi sangat penting untuk membandingkan algoritma dan menentukan teknik yang paling tepat untuk digunakan dalam pembuatan aplikasi.	Algoritma terbaik diperlukan dalam memprediksi performa akademik. Sebuah perbandingan diperlukan sebelum menentukan algoritma apa yang akan digunakan. Beberapa peneliti menyatakan bahwa algoritma terbaik dari penelitian mereka bervariasi. Tidak ada satu algoritma yang terbaik dalam memprediksi. Hal ini dipengaruhi oleh data yang diperoleh dan proses preprocessing.
10	***"Predicting students at risk of academic failure using ensemble model during pandemic in a distance learning system"*** Halit Karalar dkk, 2021 (Q1)	Tujuan dari penelitian ini adalah menemukan teknik data mining dan machine learning untuk mengidentifikasi siswa yang beresiko gagal akademik selama pandemi dengan menggunakan data pembelajaran secara sinkron dan asinkron	Penelitian ini mengusulkan model ensemble optimal yang memprediksi siswa berisiko menggunakan kombinasi algoritma mesin pembelajaran yang relevan. Data yang digunakan merupakan data dari **LMS (Moodle) dan Conference Management Software (Adobe Connect) selama satu semester**. Variabel atau atribut yang digunakan diantaranya; jenis kelamin, tingkat, jumlah catatan kuliah dan materi kuliah yang diunduh, total waktu yang dihabiskan dalam sesi online, jumlah kehadiran, dan skor kuis. Memprediksi hasil **skor IPK (lulus/gagal)**, mata kuliah dipilih sebagai kolom target sebagai preferensi umum untuk menentukan performa siswa.	Penelitian ini menyimpulkan bahwa pendekatan model pembelajaran ensemble yang terdiri dari kombinasi algoritma klasifikasi Extra Trees (ET), Random Forest (RF) dan Logistic Regression (LR), paling baik untuk memprediksi siswa yang berisiko gagal akademik. Hasil dari penelitian ini adalah bahwa **skor kuis, gelar, jumlah beban catatan kuliah, jumlah unduhan materi pelajaran lainnya, dan total waktu yang dihabiskan untuk menonton video kursus yang direkam efektif dalam memprediksi** siswa yang berisiko.	Pada penelitian selanjutnya disarankan untuk menguji model pembelajaran ensemble.

No	Judul, Penulis, dan Tahun	Tujuan	Metode	Hasil	Catatan
11	***Model Educational Data Mining Berbasis Gradient Boosted Trees Untuk Prediksi Performa Akademik Mahasiswa*** **(Keaslian Penelitian)**	1. Mengembangkan model EDM dalam memprediksi performa akademik mahasiswa 2. Menggabungkan fitur akademik dan non-akademik untuk memprediksi performa akademik 3. Menemukan fitur yang paling penting dalam memprediksi performa akademik mahasiswa 4. Meningkatkan akurasi prediksi dengan menggunakan optimasi hyperparameter 5. Memberikan rekomendasi metode dalam memprediksi performa akademik mahasiswa di perguruan tinggi	1. Mengusulkan gabungan dari fitur akademik dan non-akademik untuk prediksi performa akademik mahasiswa dengan menggunakan model mesin pembelajaran GBT. 2. Mengumpulkan dan memproses data akademik dan non-akademik sebagai data untuk prediksi performa akademik mahasiswa 3. Menghitung pengaruh setiap fitur dalam memprediksi performa akademik mahasiswa 4. Melakukan optimasi metode prediksi dengan cara tuning hyperparameter 5. Membangun model untuk sistem informasi prediksi performa akademik mahasiswa	1. Model GBT merupakan model yang memiliki kesalahan terkecil dibandingkan dengan model-model lainnya. 2. Mendapatkan data performa akademik yang siap digunakan dari data akademik dan non-akademik. 3. Menemukan fitur-fitur yang mempunyai pengaruh besar dan pengaruh terkecil dari semua fitur yang diusulkan. 4. Menemukan metode untuk meningkatkan akurasi prediksi performa akademik mahasiswa. 5. Membangun model untuk sistem informasi prediksi performa akademik mahaiswa	1. Penggunaan data akademik berasal dari catatan LMS dan nilai IPS, kedepan sebaiknya ditambahkan nilai ulangan harian dan nilai tugas-tugas serta nilai ujian 2. Terdapat banyak fitur non-akademik namun karena keterbatasan peneliti maka pada penelitian ini hanya menggunakan fitur demografi, ekonomi dan organisasi kampus. 3. Banyaknya metode optimasi model mesin pembelajaran, namun pada penelitian ini hanya menggunakan satu metode, kedepan bisa dicoba berbagai metode optimasi lainnya.

2.2 Landasan Teori

Membangun model EDM untuk sistem informasi prediksi performa akademik mahasiswa membutuhkan berbagai teori yang harus dipahami. Beberapa teori dasar yang harus dipahami adalah tentang Sistem Informasi, Performa (Kinerja), EDM, CRISP-DM, Organisasi Kemahasiswaan, Moodle, Seleksi Fitur, Algoritma *Gradient Boosted Trees*, *Tuning Hyperparameters*, *Cross Validation*, MAE, MSE, RMSE dan R^2. Adapun masing-masing pentingnya teori-teori tersebut didalam penelitiaan ini dijelaskan dibagian awal dari masing-masing penjelasan setiap teori.

2.2.1. Sistem Informasi

Pemahaman teori tentang Sistem Informasi harus diketahui secara mendalam, hal ini terkait dengan inti dari penulisan disertasi, dimana topik utama dari penulisan ini adalah sistem informasi berbasis model EDM untuk memprediksi performa akademik mahasiswa. Sistem informasi, menurut Laudon dan Laudon (2010), adalah sekelompok komponen yang saling terkait yang mengumpulkan, memproses, menyimpan, dan mendistribusikan informasi untuk mendukung pengambilan keputusan dan pengontrolan dalam sebuah organisasi. O'Brien dan Marakas (2011) mendefinisikan sistem informasi sebagai kombinasi orang, perangkat keras, perangkat lunak, jaringan komunikasi, dan sumber daya data yang mengumpulkan, mengubah, dan menyebarkan informasi di sebuah organisasi. Turban (2009) menggambarkan sistem informasi sebagai proses mengumpulkan, memproses, menyimpan, menganalisis, dan menyebarkan informasi untuk tujuan tertentu, umumnya dikomputerisasi.

Dengan merinci definisi-definisi tersebut, dapat disimpulkan bahwa sistem informasi merupakan kombinasi terorganisir antara orang-orang, informasi, jaringan komunikasi, perangkat keras, perangkat lunak, serta aturan dan prosedur. Semua ini berinteraksi untuk memproses dan menyebarluaskan informasi yang diperlukan untuk pengambilan keputusan di sebuah organisasi.

Fungsi sistem informasi, menurut O'Brien dan Marakas (2011), meliputi mendukung fungsi bisnis seperti keuangan, akuntansi, operasional, pemasaran, dan sumber daya manusia. Selain itu, sistem informasi digunakan untuk meningkatkan

efisiensi proses produksi, produktivitas pekerja, pelayanan pelanggan, sebagai sumber utama informasi untuk pengambilan keputusan, pengembangan produk dan jasa yang kompetitif, serta sebagai komponen utama dalam infrastruktur bisnis dan jaringan kehandalan.

Sistem informasi memiliki tiga peran utama dalam aplikasi bisnis, yakni mendukung proses dan operasi bisnis, membantu pembuatan keputusan oleh pekerja dan manajer, serta mendukung strategi untuk keunggulan kompetitif perusahaan. Hal ini mencakup berbagai aktivitas seperti mencatat pembelian, melacak persediaan, menggaji pekerja, membuat keputusan tentang produk dan layanan, serta strategi penjualan online.

Komponen utama sistem informasi, menurut Stair dan Reynolds (2010), mencakup masukan (pengumpulan data mentah), proses (konversi data menjadi output yang berguna), keluaran (menghasilkan informasi), dan umpan balik (output untuk membuat perubahan dalam aktivitas input atau proses).

Sebuah sistem informasi yang baik harus melalui proses evaluasi, terdapat banyak model untuk mengevaluasi sistem informasi. DeLone and McLean (1992) menciptakan D&M IS Success Model untuk mengukur tingkat keberhasilan suatu sistem informasi. Model ini menjelaskan bahwa kualitas sistem dan informasi, secara independen maupun bersama-sama, mempengaruhi penggunaan dan kepuasan pemakai. Tingkat penggunaan dapat memengaruhi kepuasan pemakai secara positif atau negatif, yang selanjutnya berdampak pada individu dan organisasi (Jogianto 2007). Kepuasan pemakai terhadap sistem informasi dipahami sebagai pandangan nyata pemakai terhadap sistem, bukan hanya aspek teknisnya (Guimaraes dkk. 2003). Davis (1989) menyatakan bahwa kepuasan pemakai berkaitan dengan respons penerima terhadap output sistem informasi. Doll dan Torkzadeh (1988) menggambarkan kepuasan pemakai sebagai sikap afektif terhadap suatu aplikasi komputer tertentu yang diinteraksinya secara langsung. Model evaluasi Doll & Torkzadeh, yang disebut model *End User Computing* (EUC) Satisfaction, menyoroti kepuasan pengguna akhir terhadap aspek teknologi dengan mempertimbangkan isi, keakuratan, format, kemudahan penggunaan, dan ketepatan waktu. Pengukuran kepuasan pemakai menjadi metode yang umum digunakan

untuk menilai keberhasilan sistem informasi karena kepuasan pemakai dianggap mencerminkan keberhasilan sistem tersebut (Jogianto 2007).

2.2.2. Performa (Kinerja) Akademik

Teori tentang performa dibutuhkan untuk mengetahui apa yang dimaksud dengan performa dan faktor apa saja yang mempengaruhi performa seseorang. Victor H. Vroom (1964) dikenal dengan kontribusinya dalam pengembangan teori motivasi di tempat kerja. Salah satu teori yang paling terkenal yang dia perkenalkan adalah Teori *Expectancy* atau *Expectancy Theory*. Teori ini pertama kali dijelaskan oleh Vroom dalam bukunya yang berjudul "*Work and Motivation*" yang diterbitkan pada tahun 1964. Teori *Expectancy* Vroom terdiri dari tiga poin yaitu: *Expectancy* (Harapan), *Instrumentality* (Instrumentalitas), dan *Valence* (Nilai). *Expectancy* mengacu pada keyakinan individu bahwa usahanya akan menghasilkan tingkat performa tertentu. Dalam konteks ini, individu pertama-tama menilai sejauh mana mereka percaya bahwa mereka dapat mencapai tujuan tertentu. Sedangkan *Instrumentality* berkaitan dengan keyakinan individu bahwa mencapai tingkat performa tertentu akan menghasilkan hasil yang diinginkan atau hadiah. Ini melibatkan keyakinan bahwa mencapai tujuan akan mengarah pada konsekuensi positif. Sementara itu *Valence* mencerminkan seberapa berharga atau diinginkannya hadiah tersebut bagi individu. Ini melibatkan penilaian subjektif tentang sejauh mana individu menginginkan hasil atau hadiah tertentu. Teori *Expectancy* Vroom sering dijelaskan melalui model EIV, di mana motivasi individu (M) dijelaskan sebagai hasil dari perkalian tiga faktor: *Expectancy* (E), *Instrumentality* (I), dan *Valence* (V). Vroom menekankan bahwa hubungan antara tiga elemen ini bersifat multipikatif, yang berarti bahwa jika salah satu faktor mendekati nol, motivasi keseluruhan akan menurun. Menurut teori ini, individu akan memutuskan untuk bertindak atau tidak berdasarkan penilaian mereka terhadap E, I, dan V. Jika mereka percaya bahwa usaha mereka akan menghasilkan tingkat performa yang tinggi, bahwa tingkat performa tersebut akan mengarah pada hasil yang diinginkan, dan bahwa mereka menginginkan hasil tersebut, mereka cenderung termotivasi untuk bertindak. Teori *Expectancy* Vroom memberikan

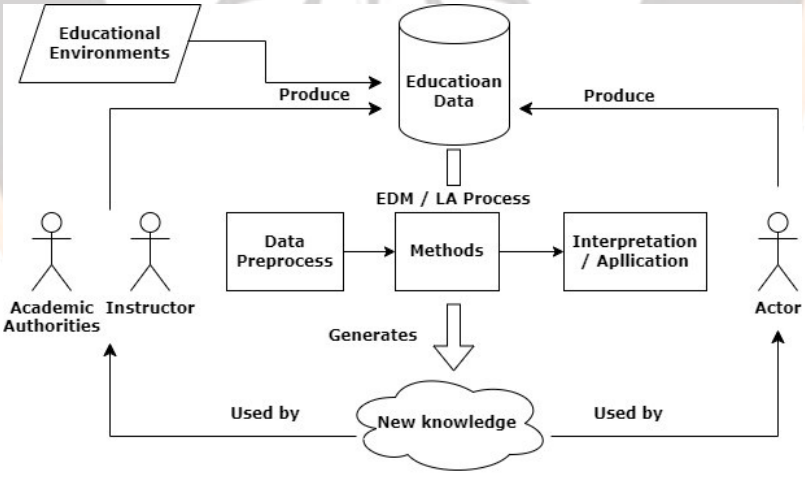
pemahaman tentang bagaimana individu membuat keputusan motivasi berdasarkan ekspektasi mereka tentang hubungan antara usaha, performa, dan hasil. Beberapa yang dapat menghambat performa adalah kurang memiliki keterampilan, sumber daya, dan kemampuan teknis yang tepat untuk mencapai tujuan. Dengan memahami hambatan tersebut dapat berupaya menerapkan pelatihan, pembinaan atau pengawasan yang sesuai, atau berdiskusi terhadap pencapaian tujuan. Teori ekspektasi merupakan salah satu teori motivasi yang paling sentral, dan mendukung pandangan bahwa teori ekspektasi dapat memprediksi usaha dan performa. Teori ini telah menjadi landasan bagi penelitian lebih lanjut tentang motivasi dan pengambilan keputusan di berbagai konteks.

Kasmir (2016) menyatakan bahwa "kinerja adalah hasil kerja dan perilaku kerja yang telah dicapai dalam menyelesaikan tugas-tugas dan tanggung jawab yang diberikan dalam suatu periode tertentu". Emron Edison (2016) menyatakan bahwa "kinerja adalah hasil yang diperoleh oleh suatu organisasi baik organisasi tersebut bersifat *profit oriented* dan *non profit oriented* yang dihasilkan selama satu periode waktu". Sementara itu, Lamas (2015) menyampaikan bahwa terdapat perbedaan antara prestasi akademik dan performa akademik mahasiswa, prestasi mengacu pada penyelesaian dan pencapaian tingkat tertentu yang dapat dicapai oleh seorang siswa setelah serangkaian pendidikan atau pelatihan, sedangkan performa mengacu pada hasil ujian suatu mata pelajaran atau keseluruhan mata kuliah (IPS). Metrik seperti nilai ujian, nilai tes dan nilai rata-rata (IPK) merupakan salah satu indikator yang digunakan untuk menentukan performa akademik mahasiswa (Caro dkk. 2014). Hal ini juga selaras dengan yang disampaikan oleh Cai dan Cao (2018) yang menyatakankan bahwa performa akademik merupakan prestasi siswa pada tahap tertentu. Berdasarkan hal tersebut maka pada penelitian ini penulis tidak menggunakan fitur ujian masuk dan prestasi sebelumnya. Penelitian ini berfokus pada prediksi performa akademik mahasiswa semester dua dan empat dengan menggunakan fitur akademik (IPS dan LMS) dan fitur nonakademik (demografi, ekonomi, organisasi kampus).

2.2.3. EDM (Educational Data Mining)

Pemahaman teori EDM dibutuhkan untuk mengetahui sumber data, metode dan hasil dari EDM tersebut. EDM merupakan proses penemuan pengetahuan yang iteratif. Gambar-1 menunjukkan bahwa proses dimulai dan berlangsung di lingkungan pendidikan, seperti ruang kelas tradisional, sistem *e-learning*, dan sistem pendidikan berbasis web yang adaptif dan cerdas. Lingkungan pendidikan ini menghasilkan data mentah, termasuk data penggunaan dan interaksi mahasiswa, informasi kursus, dan data akademik (Mais Haj Qasem, Raneem Qaddoura 2017).

Proses EDM terdiri dari formulasi hipotesis, pengujian, dan penyempurnaan. Hipotesis dikembangkan dari berbagai lingkungan pendidikan yang menciptakan data dalam volume besar. Proses utama EDM dimulai dengan validasi data (yaitu, menemukan hubungan antara variabel, parameter, dan item data) yang juga dikenal sebagai prapengolahan data. Setelah prapengolahan selanjutnya diolah menggunakan data mining, dan hasil/interpretasi diberikan kepada berbagai pengguna pendidikan. Rekomendasi lebih lanjut disarankan untuk mengatasi masalah ataupun penyempurnaan tugas pendidikan (Romero dan Ventura 2013).



Gambar 2. 1 Desain EDM (Romero dkk. 2020)

Tujuan EDM di Perguruan Tinggi adalah; 1) Memprediksi perilaku belajar siswa di masa depan dengan penggunaan model siswa, tujuan ini dapat dicapai dengan menciptakan model siswa yang menggabungkan karakteristik siswa, termasuk informasi terperinci seperti pengetahuan, perilaku, dan motivasi mereka

untuk belajar. 2) Menemukan atau meningkatkan domain model melalui berbagai metode dan aplikasi EDM, dimungkinkan untuk menemukan model baru dan meningkatkan model yang ada. 3) Mempelajari efek dukungan pendidikan terhadap capaian sistem pembelajaran. 4) Mendapatkan pengetahuan ilmiah tentang pembelajaran dengan membangun dan menggabungkan model siswa, teknologi dan perangkat lunak yang digunakan (Suthar, Patel, dan Patel 2019).

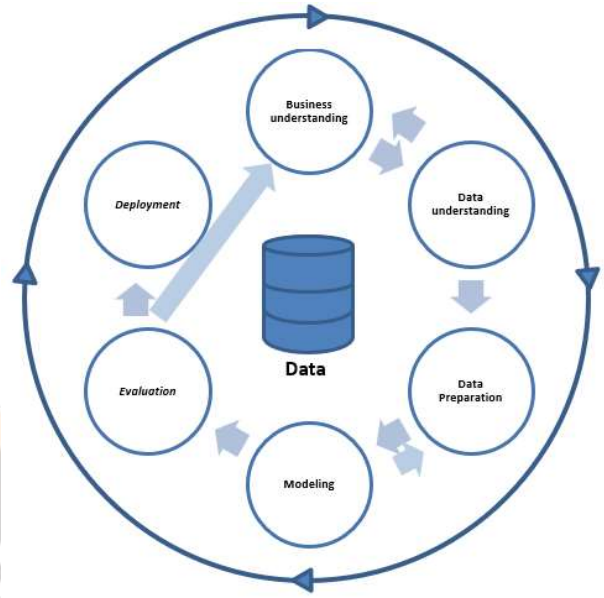
Penerapan EDM merupakan siklus berulang dari pembentukan hipotesis, pengujian, dan penyempurnaan dalam proses pendidikan. Komponen utama EDM adalah sistem pendidikan, teknik data mining dan pengguna. Sistem pendidikan terdiri dari ruang kelas tradisional, sistem *e-learning*, LMS, sistem adaptif berbasis web, *intelligent tutoring systems*, kuesioner dan kuis. Teknik data mining yang digunakan dalam EDM yaitu; *statistics, visualization, clustering, classification, association rule mining* dan *sequence mining, text mining.* Pengguna EDM adalah mahasiswa atau pelajar, instruktur, guru atau dosen, administrasi, peneliti dan pejabat sekolah (Romero dkk. 2010).

2.2.4. CRoss-Industry Standard Process for Data Mining

CRoss-Industry Standard Process for Data Mining (CRISP-DM) merupakan metodologi penggunaan data mining dalam menyelesaikan masalah bisnis suatu organisasi. Dengan merujuk pada praktik terbaik, para praktisi dan peneliti *Data Mining* mengusulkan beberapa proses, seperti alur kerja atau pendekatan dengan tahapan-tahapan yang mudah, untuk meningkatkan peluang keberhasilan dalam pelaksanaan berbagai proyek *Data Mining*. Upaya ini menghasilkan beberapa proses yang dijadikan standar, yang dikenal sebagai CRISP-DM. Penerapan metodologi CRISP-DM pada dunia pendidikan telah dilakukan banyak peneliti (Buenaño Fernández dan Luján-Mora 2016; Castro R., Espitia P., dan Montilla 2018; Layth Almahadeen, Murat Akkaya 2017; Oreski, Pihir, dan Konecki 2017; Schäfer, Zeiselmair, dan Becker 2020), berdasarkan hal tersebut maka pada penelitian ini menggunakan metodologi CRISP-DM. CRISP-DM menyediakan standar proses *Data Mining* sebagai strategi pemecahan masalah secara umum dari bisnis atau unit penelitian (Larose 2006).

Proses *Data Mining* berdasarkan CRISP-DM terdiri dari enam fase (Larose 2006), adalah sebagai berikut:

1. *Business Understanding Phase* (Fase Pemahaman Bisnis)
   a. Menetapkan tujuan dan kebutuhan proyek secara rinci dalam konteks bisnis atau unit penelitian secara menyeluruh.
   b. Mengartikan tujuan dan batasan menjadi rumusan masalah data mining.
   c. Membuat strategi awal untuk mencapai tujuan yang telah ditetapkan.
2. *Data Understanding Phase* (Fase Pemahaman Data)
   a. Koleksi data.
   b. Gunakan analisis penyelidikan data untuk mendalami data dan mencari pemahaman awal.
   c. Evaluasi kualitas data.
   d. Jika perlu, pilih subset kecil data yang mungkin mengungkap pola dari masalah tersebut.
3. *Data Preparation Phase* (Fase Persiapan Data)
   a. Persiapkan data awal dengan mengumpulkan informasi yang diperlukan untuk seluruh fase berikutnya, yang merupakan tugas yang memerlukan upaya intensif.
   b. Pilih kasus dan variabel yang akan dianalisis, disesuaikan dengan jenis analisis yang akan dilakukan.
   c. Modifikasi variabel-variabel tertentu jika diperlukan.
   d. Siapkan data awal agar siap digunakan dalam proses pemodelan.
4. *Modelling Phase* (Fase Pemodelan)
   a. Pilih dan terapkan teknik pemodelan yang tepat.
   b. Kalibrasi aturan model untuk mencapai hasil optimal.
   c. Penting untuk dicatat bahwa beberapa teknik dapat digunakan untuk permasalahan data mining yang serupa.
   d. Jika diperlukan, proses dapat kembali ke tahap pengolahan data agar sesuai dengan spesifikasi teknik data mining tertentu.



Gambar 2. 2 Proses CRISP-DM (Larose 2006)

5. *Evaluation Phase* (Fase Evaluasi)
   a. Evaluasi satu atau lebih model yang diterapkan pada fase pemodelan untuk menilai kualitas dan efektivitas sebelum diimplementasikan.
   b. Tentukan apakah model-model tersebut mencapai tujuan yang ditetapkan pada awal.
   c. Identifikasi apakah terdapat permasalahan bisnis atau penelitian yang belum ditangani secara memadai.
   d. Buat keputusan terkait pemanfaatan hasil dari proses data mining.
6. *Deployment Phase* (Fase penyebaran)
   a. Menerapkan model yang telah dibuat. Pembentukan model tidak menunjukkan penyelesaian proyek.
   b. Sebagai contoh cara menyebarkan: Pembuatan laporan.

2.2.5. Organisasi Kemahasiswaan

Organisasi kemahasiswaan perlu dipahami lebih lanjut, hal ini disebabkan bahwa organisasi kemahasiswaan merupakan salah satu fitur yang diusulkan pada penelitian ini. Usulan fitur ini didasari dari temuan para peneliti sebelumnya dimana terdapat beberapa pengaruh organisasi kemahasiswaan terhadap mahasiswa diantaranya; Arola Anderson dkk. (2021) menyampaikan bahwa aktivitas

organisasi kemahasiswaan pada semester pertama memiliki pengaruh positif terhadap optimis mahasiswa, membantu mengatasi perubahan (Bohnert dkk. 2010), berpengaruh positif terhadap perkembangan dan perilaku anak (Morris 2015), memiliki pengaruh terhadap masa tunggu lulusan dalam mendapatkan pekerjaan (Arifin dkk. 2022). Sebelum menggunakan fitur ini untuk prediksi performa akademik, penulis telah menguji pengaruh dari fitur ini terhadap kesuksesan alumni dalam mendapatkan pekerjaan setelah lulus, hasilnya menunjukkan bahwa mahasiswa yang memiliki nilai IPK tinggi dan mengikuti organisasi kampus berpeluang lebih besar dalam mendapatkan pekerjaan jika dibandingkan dengan mahasiswa yang tidak mengikuti organisasi kampus.

2.2.6. Moodle

Modular Object-Oriented Developmental Learning Environment (Moodle) merupakan salah satu *learning management system* (LMS) yang paling sering digunakan oleh para peneliti EDM dalam memprediksi performa akademik mahasiswa. Sistem Moodle menyimpan banyak informasi rinci tentang konten pembelajaran, pengguna, dan penggunaan dalam database relasional. Conijn dkk., (2017) menganalisis 17 mata kuliah campuran dengan 4.989 siswa menggunakan log LMS Moodle. Tujuan mereka adalah untuk memprediksi nilai akhir mahasiswa dari variabel prediktor LMS dan nilai tugas, menggunakan model logistik (lulus/gagal) dan regresi standar (nilai akhir). Mereka memprediksi performa mahasiswa selama 10 minggu pertama. Akurasi sedikit meningkat saat minggu pertama, dengan peningkatan yang mencolok setelah minggu ke-5, saat nilai tugas tersedia. Pada minggu ke-5, model regresi menunjukkan 0,43 penyesuaian R2, dan pengklasifikasi biner lulus/gagal memperoleh akurasi 67% pada minggu ke-3. Sementara itu, Gerritsen menggunakan file log Moodle dari 17 mata pelajaran untuk meramalkan apakah seorang mahasiswa akan lulus atau gagal dalam suatu mata pelajaran (klasifikasi biner) (Gerritsen 2017). Dari tujuh model, *multilayer perceptron* mengungguli pengklasifikasi lainnya, berhasil mendeteksi siswa berisiko dengan akurasi 66,1%.

Moodle terdiri dari kuis, tugas, dan forum. Kuis adalah perangkat yang bermanfaat untuk mahasiswa dalam menguji pengetahuan mereka dan meninjau setiap mata pelajaran yang dipelajari. Moodle dapat digunakan untuk memberi mahasiswa umpan balik yang cepat tentang kinerja mereka dan untuk mengukur pemahaman mereka tentang materi. Tugas adalah alat untuk mengumpulkan pekerjaan siswa. Mahasiswa tidak harus login pada saat yang sama untuk berkomunikasi dengan guru atau teman sekelasnya. Dengan demikian, siswa dapat meluangkan waktu untuk menyusun balasan, yang dapat mengarah pada diskusi yang lebih bijaksana. Banyak penelitian menunjukkan bahwa lebih banyak siswa yang bersedia berpartisipasi dalam forum asinkron daripada yang bersedia berbicara di kelas. Forum menciptakan banyak peluang untuk mencontoh percakapan yang lakukan di kelas, merumuskan diskusi tugas antara kelompok, atau mengumpulkan ide dan pertanyaan terbaik dari forum (Romero dkk. 2010).

Bravo-Agapito dkk (2021) mengekstrak variabel-variabel yang digunakan untuk memprediksi performa akademik mahasiswa dari LMS Moodle, dijabarkan dalam Tabel 2.3.

Tabel 2. 2 Ekstrak variabel Moodle untuk prediksi performa akademik siswa

No	Nama	Deskripsi
1	*Total_logins*	Jumlah total login ke platform Moodle
2	*N_access_forum*	Frekuensi akses mahasiswa ke semua forum
3	*N_added_messages_forum*	Jumlah total pesan yang ditambahkan oleh mahasiswa di forum
4	*N_access_didactic_units*	Frekuensi akses siswa terhadap bahan ajar
5	*N_access_glossaries*	Frekuensi akses siswa ke semua glosarium
6	*Total_assignments*	Jumlah tugas kursus
7	*N_assignments_consulted*	Frekuensi siswa berkonsultasi tugas
8	*N_assignments_submitted*	Jumlah tugas yang diserahkan
9	*N_access_questionnaires*	Frekuensi siswa mengakses semua kuesioner
10	*N_attempts_questionnaires*	Frekuensi siswa mencoba untuk memecahkan kuesioner
11	*N_answered_questions*	Jumlah pertanyaan yang dijawab di semua kuesioner
12	*N_questionnaire_views*	Frekuensi siswa mengamati angket
13	*N_questionnaires_submitted*	Frekuensi siswa menyerahkan angket
14	*N_reviews_questionnaire*	Frekuensi siswa upaya untuk merevisi kuesioner
15	*Days_first_access*	Jumlah hari hingga akses pertama ke kelas LMS
16	*N_entries course*	Jumlah total entri ke kursus

2.2.7. Seleksi Fitur

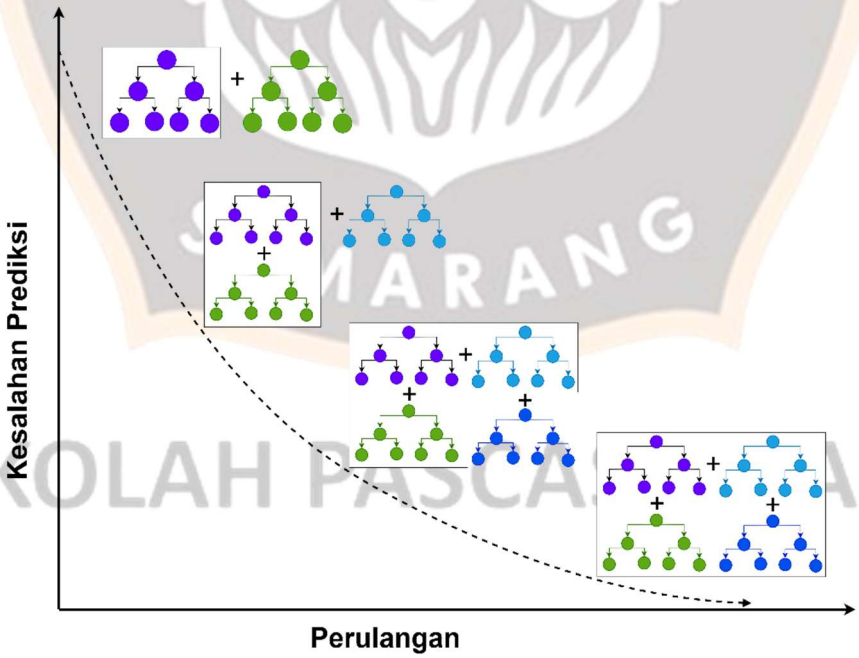
Seleksi fitur merupakan usaha pada tahap prepengolahan data untuk mengurangi fitur yang tidak memiliki pengaruh besar dan tidak relevan dalam meningkatkan kinerja model. Dalam proses pengumpulan data terdapat beberapa fitur yang tidak relevan, bahkan fitur-fitur tersebut cenderung memperlambat kinerja model jika tetap digunakan. Sementara itu menurut (Hall 1999), tujuan seleksi fitur adalah mengidentifikasi fitur-fitur yang memiliki kepentingan serupa dalam dataset, sambil mengeliminasi fitur yang memberikan informasi tidak relevan dan berlebihan. Proses seleksi fitur mengurangi dimensi data, memungkinkan algoritma beroperasi lebih efisien, meningkatkan akurasi prediksi, dan menghindari atribut yang kurang bermanfaat dalam proses belajar (Blum dan Langley 1997). Untuk mengatasi masalah ini, langkah pemrosesan awal dapat digunakan untuk menghilangkan atribut yang tidak relevan sebelum menerapkan algoritma *Data Mining*.

Terdapat dua metode fitur seleksi yaitu *filter*, *wrapper*, dan *embedded*. Metode filter menggunakan ukuran statistik untuk memberikan skor pada setiap fitur, dan fitur diberi peringkat berdasarkan skor tersebut untuk dipertahankan atau dihapus dari dataset. Kriteria penilaian metode *filter* melibatkan jarak, informasi, ketergantungan, dan konsistensi. Penggunaan metode *filter* umumnya sebagai tahap prapengolahan data, tidak bergantung pada algoritma *Machine Learning* tertentu, dan berfokus pada korelasi fitur dengan variabel hasil. Di sisi lain, metode *wrapper* memerlukan satu jenis algoritma *Machine Learning*, menggunakan kinerjanya sebagai kriteria evaluasi, dan berusaha meningkatkan kinerja algoritma dengan mencari fitur yang paling sesuai. Metode *wrapper* melakukan evaluasi terhadap semua kombinasi fitur, dengan risiko kemahalan komputasi jika himpunan fiturnya besar. Sedangkan metode *embedded* selector menggabungkan keunggulan metode filter dan *wrapper* dengan mengimplementasikan algoritma yang memiliki pemilihan fitur bawaan.

2.2.8. Algoritma *Gradient Boosted Trees*

Penggunaan algoritma GBT pada penelitian ini didasari dari penemuan diawal penelitian, dimana peneliti telah membandingkan GBT dengan lima algoritma prediksi lain. Dari perbandingan tersebut ditemukan bahwa GBT merupakan algoritma yang memiliki akurasi paling tinggi dibanding algoritma pembandingnya (Arifin dkk. 2021). Algoritma GBT dipilih sebagai metode prediksi yang dapat mengatasi data numerik, nominal, dan missing data dengan handal, menjadi pilihan unggul dalam membangun model prediksi dibandingkan dengan metode prediksi lainnya (Friedman 2002). GBT merupakan metode *ensemble* yang membangun sejumlah pohon keputusan. Setiap pohon yang dibangun mengacu pada kelemahan pohon sebelumnya. Dasar pemikiran GBT adalah model terbaik berikutnya akan meminimalkan kesalahan prediksi keseluruhan jika dikombinasikan dengan model sebelumnya.

GBT bekerja dengan mengombinasikan beberapa model lemah menjadi model yang lebih kuat melalui pendekatan iteratif. Setiap iterasi bertujuan untuk meningkatkan model sebelumnya dengan menambahkan model baru, dilakukan secara berulang-ulang hingga model mencapai kriteria tertentu, seperti nilai *loss function* yang memadai.



Gambar 2. 3 Visualisasi iterasi model boosting

GBT memiliki kelebihan yaitu memiliki tingkat akurasi yang tinggi terutama pada data kompleks, fleksibilitas pada berbagai jenis data tanpa persyaratan yang ketat, dan kecepatan komputasi yang memadai. Namun, terdapat beberapa kelemahan, seperti kebutuhan tuning parameter yang cermat untuk optimalitas, risiko *overfitting* jika parameter tidak diatur dengan baik, dan ketergantungan pada jumlah data yang besar untuk mendapatkan model yang akurat dan stabil.

Tahapan pengujian algoritma GBT yang dilakukan dengan menggunakan prosedur berikut:

1. ***Input***. Data $\{(X_i, Y_i)\}_{i=1}^{n}$, dan …(1)
   perbedaan ***Loss Function*** $L(Y_i, F(x))$ …(2)
2. **Langkah pertama**: Inisialisasi model dengan nilai konstan
   $F_0(X) = \frac{argmin}{\gamma} \sum_{i=1}^{n} L(Y_i, \gamma)$ …(3)
3. **Langkah kedua**: Untuk m = 1 ke M:
   (A) Hitung $r_{im} = -\left[\frac{\partial L(y_i, F(X_i))}{\partial F(X_i)}\right]_{F(x)=F_{m-1}(x)}$ untuk i= 1,…,n …(4)
   (B) Paskan pohon regresi dengan nilai r_{im} dan buat daerah terminal R_{jm} untuk j =1… J_m …(5)
   (C) Untuk j = 1...J_m
   Hitung $\gamma_{jm} = \frac{argmin}{\gamma} \sum_{x_i \in R_{ij}} L(y_i, F_{m-1}(x_i) + \gamma)$ …(6)
   (D) Perbarui $F_m(x) = F_{m-1}(x) + \upsilon \sum_{j=1}^{J_m} \gamma_{jm} I(x \epsilon R_{jm})$ …(7)
4. **Langkah ketiga**: Keluaran $F_M(x)$

2.2.9. *Tuning hyperparameters*

Penggunaan penyetelan *hyperparameter* ini bertujuan untuk mendukung tujuan penelitian poin ketiga, yaitu untuk meningkatkan akurasi prediksi dari algoritma GBT. Hal yang mendasari penggunaan metode ini berasal dari berbagai sumber bahwa performa model tergantung pada pemilihan *hyperparameter* yang tepat (Shekhar dkk. 2021). Pilihan konfigurasi *hyperparameter* diketahui memengaruhi performa model mesin pembelajaran secara signifikan (Zhang dkk. 2022). Terdapat dua metode untuk penyetelan *hyperparameter* yaitu manual dan

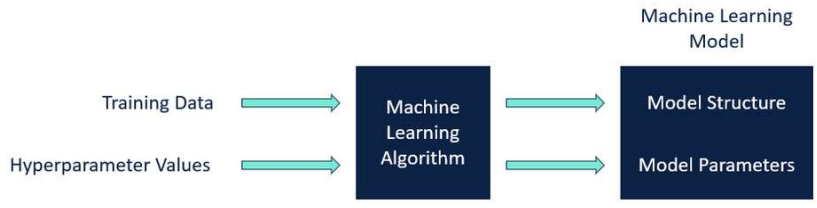
otomatis. Pencarian manual membutuhkan banyak waktu dalam upaya menemukan nilai terbaiknya (Putatunda dan Rama 2019), sementara itu metode otomatis membutuhkan algoritma pencarian yang tepat untuk mendapatkan nilai terbaiknya.

Algoritma pencarian untuk penyetelan *hyperparameter* yang paling sering digunakan adalah *Grid Search, Random Search, Optuna, dan HyperOpt.* (Shekhar dkk. 2021) membandingkan *HyperOpt, Optuna, Optunity,* dan SMAC. *Grid Search* dan *Random Search*, atau perbandingan di antara mereka, tampaknya menjadi dua topik yang dibahas di sebagian besar karya sebelumnya tentang penyetelan *hyperparameter* (Andonie dan Florea 2020; Duarte dan Wainer 2017; Florea dan Andonie 2019; Liashchynskyi dan Liashchynskyi 2019; MacKay dkk. 2019; Shekar dan Dagnew 2019; Villalobos-Arias dkk. 2020; Wong dkk. 2019). Sementara itu, (Shekhar dkk. 2021) membandingkan *Optuna, HyperOpt*, dan metode lainnya. (Putatunda dan Rama 2019) bandingkan teknik penyetelan *hyperparameter*: *Random Search, Grid Search, Hyperopt*, dan *Randomized-Hyperopt*. (Putatunda dan Rama 2018) perbandingan *Hyperopt, Random Search, dan Grid Search* untuk menyetel *hyperparameter* algoritma XGBoost. (Joy dan Selvan 2022) bandingkan *Random Search, Grid Search, Optuna*. Berdasarkan referensi-referensi diatas maka pada penelitian ini akan membandingkan algoritma pencarian yang paling sering digunakan yaitu *Random Search, Grid Search, Optuna* dan *Hyperopt* dalam proses pembuatan model prediksi.

Pembuatan model pada mesin pembelajaran memerlukan tiga komponen yaitu data pelatihan, parameter, dan *hyperparameter*. Model memperoleh nilai parameter selama pelatihan dengan belajar dari data yang diberikan, nilai ini tidak dapat ditetapkan secara manual. Data pelatihan (input) adalah kumpulan data yang dipergunakan untuk melatih model. Data pelatihan adalah apa yang dimanfaatkan algoritma (instruksi untuk membangun model) untuk mengidentifikasi pola dalam data dan mengeksploitasi pola untuk membuat prediksi. Data pelatihan (biasanya) tidak menjadi bagian dari model secara langsung, mereka hanya digunakan untuk mengajarkan model seperti apa data itu, dan bagaimana menangani data untuk membuat perkiraan yang berarti dari variabel target.

Komponen kedua adalah parameter, parameter dipelajari dari data pelatihan, yaitu bobot, ambang batas, atau koefisien yang memungkinkan model membuat prediksi. Ketika algoritma "mempelajari" data pelatihan yang disediakan untuk membuat model, algoritma benar-benar menyesuaikan parameter model akhir untuk membuat prediksi terbaik dari variabel target berdasarkan data pelatihan. Parameter pada akhirnya disimpan sebagai bagian dari model yang dipelajari. Parameter adalah bagian dari model yang secara otomatis disesuaikan agar sesuai dengan data spesifik dan kasus penggunaannya.

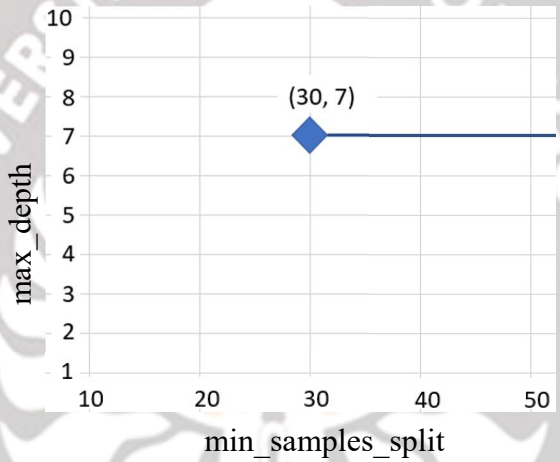
Komponen ketiga dari model terlatih adalah *hyperparameter*, dimana komponen inilah yang mengatur proses pelatihan model. *Hyperparameter* biasanya menentukan kapan model selesai dilatih, atau berapa banyak rekaman yang dipertimbangkan algoritma pada satu waktu selama pelatihan. Tidak seperti parameter, *hyperparameter* bersifat konstan selama proses pelatihan, ditetapkan sebelum pelatihan model dan tidak disesuaikan selama proses pelatihan. *Hyperparameter* berdampak langsung pada nilai akhir parameter model, dan beberapa *hyperparameter* juga berdampak langsung pada struktur model pembelajaran mesin yang dilatih. Karena *hyperparameter* menentukan struktur sebenarnya dari algoritma pembelajaran mesin dan proses pelatihan model, tidak ada cara untuk "mempelajari" nilai-nilai ini menggunakan fungsi kerugian dan data pelatihan.



Gambar 2. 4 Parameter mesin pembelajaran

Parameter model meliputi koefisien model dalam model regresi linier (Ma dkk. 2018). Sementara itu *hyperparameter* tidak memperoleh nilainya dari data. Oleh karena itu, peneliti harus mengaturnya secara manual. Pada pembangunan model tertentu, peneliti selalu menetapkan nilai pada *hyperparameter* (misalnya, sebelum proses pelatihan) (Ma dkk. 2018). Parameter model dapat dikendalikan oleh

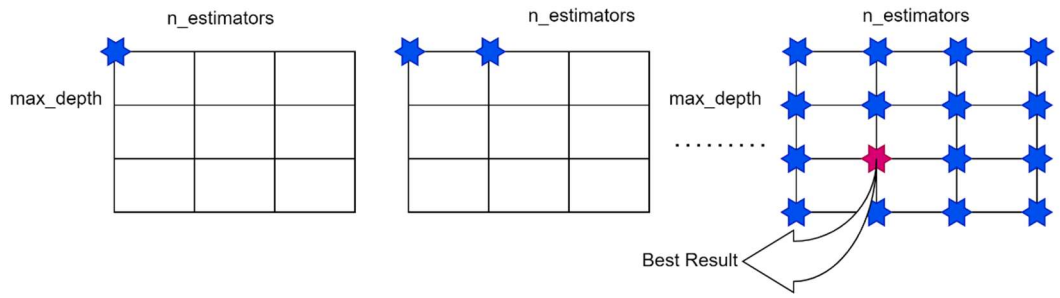
hyperparameter model. Dengan kata lain, performa model dapat dipengaruhi oleh *hyperparameter* model. Oleh karena itu, kita harus memodifikasi nilai *hyperparameter* model untuk menghasilkan output model yang optimal atau terbaik. Ruang pencarian *hyperparameter* sangat penting untuk proses penyetelan *hyperparameter*. Semua kemungkinan kombinasi nilai *hyperparameter* yang ditentukan pengguna dapat ditemukan di ruang pencarian. Ruang pencarian 2 dimensi untuk dua *hyperparameter* terpisah *max_depth* dan *min_samples_split* ditunjukkan pada Gambar 2.5.



Gambar 2. 5 Ruang *hyperparameter* dua dimensi

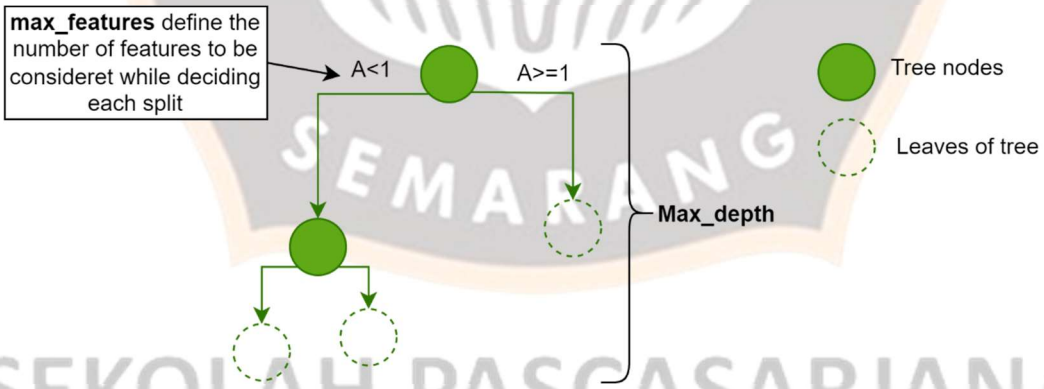
Berdasarkan hasil dari perbandingan algoritma pencarian *hyperparameter* yang telah dilakukan oleh penulis ditemukan bahwa Grid Search (GS) merupakan algoritma terbaik dalam menemukan kombinasi nilai dari parameter yang diusulkan. GS adalah metode sederhana dan mudah digunakan (Bergstra dan Bengio 2012). Metode ini paling sering digunakan oleh peneliti untuk melakukan optimasi dan terbukti meningkatkan akurasi (Andonie dan Florea 2020; Duarte dan Wainer 2017; Florea dan Andonie 2019; Joy dan Selvan 2022; Liashchynskyi dan Liashchynskyi 2019; MacKay dkk. 2019; Putatunda dan Rama 2018; Shekar dan Dagnew 2019; Villalobos-Arias dkk. 2020; Villalobos-Arias dan Quesada-López 2021; Wong dkk. 2019). GS adalah teknik untuk mencari ruang pencarian untuk semua kemungkinan kombinasi *hyperparameter* yang diberikan oleh pengguna. GS adalah metode yang secara tradisional telah digunakan dengan melihat melalui

semua kombinasi parameter potensial. Seluruh ruang parameter dipertimbangkan dan dipartisi menjadi kotak saat melakukan pencarian kisi. Titik-titik grid kemudian masing-masing dinilai sebagai *hyperparameter*. GS mencoba setiap kombinasi dari daftar nilai-nilai tersebut dari *hyperparameter* dan mengevaluasi model untuk setiap kombinasi. Pola yang diikuti mirip dengan kisi-kisi, di mana semua nilai ditempatkan dalam bentuk matriks. Setiap set parameter dipertimbangkan dan akurasi dicatat. Setelah semua kombinasi dievaluasi, model dengan kumpulan parameter yang memberikan akurasi tertinggi dianggap yang terbaik, representasi visual dari GS ditunjukkan pada Gambar 2.6.



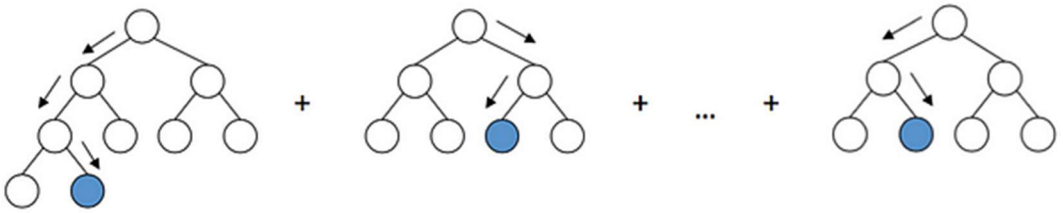
Gambar 2. 6 Representasi visual GS

(Sumber: https://www.numpyninja.com/post/hyper-parameter-tuning-using-grid-search-and-random-search)



Gambar 2. 7 Parameter GBT

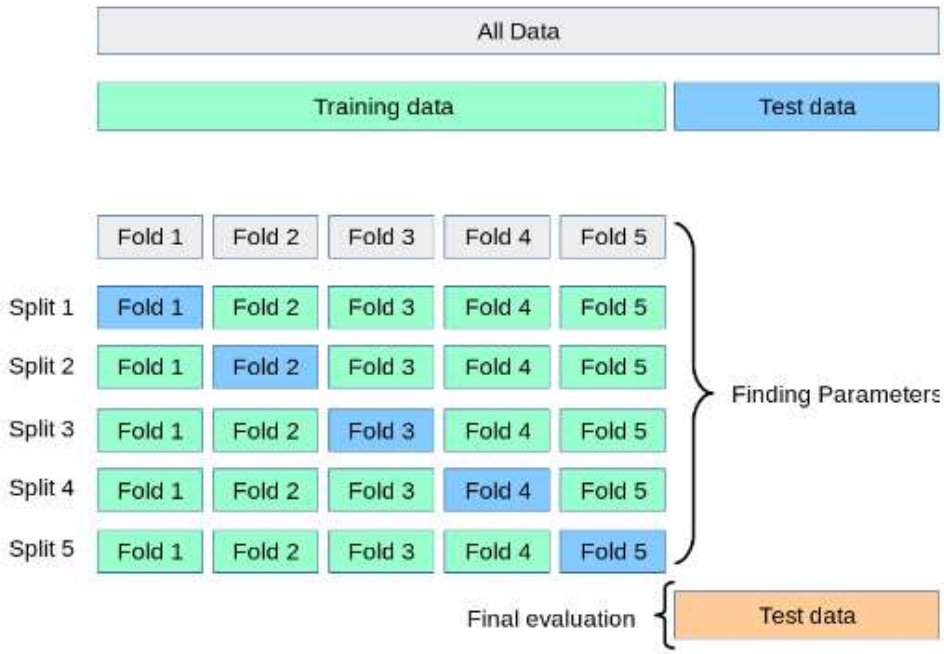
(Sumber: https://www.kaggle.com/hyperparameter/gbt)



Gambar 2. 8 Visualisasi pencarian parameter

2.2.10. Cross Validation

Cross validation adalah teknik untuk pengamatan dari kumpulan data pengujian dan pelatihan. Pengujian hasil prediksi dapat dilakukan menggunakan metode x-validation, yang melibatkan 5 langkah (5-fold cross-validation). Dalam x-validation, data dibagi menjadi 5 subset yang memiliki jumlah yang sama. Setiap subset secara bergantian diambil sebagai data uji, sementara 4 subset lainnya digunakan sebagai data pelatihan. Pendekatan ini meningkatkan akurasi pengukuran hasil, mengurangi kemungkinan inkonsistensi data selama tahap prediksi (Witten, Frank, dan Hall 2011).



Gambar 2. 9 Visualisasi pembagian data untuk validasi silang

(Sumber: https://scikit-learn.org/stable/modules/cross_validation.html)

2.2.11. MAE, MSE, RMSE dan R^2

MAE, MSE, RMSE dan R^2 merupakan matrik yang akan dipergunakan untuk mengevaluasi model yang dibangun. Langkah penting dalam setiap model pembelajaran mesin adalah mengevaluasi keakuratan model. *Mean Absolute Error* (MAE), *Mean Squared Error* (MSE) dan *Root Mean Squared Error* (RMSE) digunakan untuk mengevaluasi kinerja model dalam analisis regresi.

MAE mewakili rata-rata perbedaan absolut antara nilai aktual dan prediksi dalam kumpulan data. Ini mengukur rata-rata residu dalam kumpulan data.

$$MAE = \sum_{i=1}^{n} \frac{|y-\hat{y}|}{n} \quad \ldots(8)$$

Dimana:

$\sum$ = Jumlah keseluruhan nilai

y = Nilai aktual

$\hat{y}$ = Nilai hasil prediksi

i = Urutan data pada database

n = Banyaknya data

Tabel 2. 3 Contoh soal untuk menghitung MAE, MSE dan RMSE

Tinggi (m)	Berat (kg)	Prediksi Berat (kg)	Kesalahan	$\lvert y - \hat{y} \rvert$
1.6	88	74.2	13.8	13.8
1.6	76	74.2	1.8	1.8
1.5	56	71.6	-15.6	15.6
			Jumlah	31.2

$$MAE = \frac{31.3}{3} = 10.4$$

Sedangkan MSE mewakili rata-rata perbedaan kuadrat antara nilai asli dan prediksi dalam kumpulan data. Ini mengukur varian dari residual.

$$MSE = \sum_{i=1}^{N} \frac{(y-\hat{y})^2}{n} \quad \ldots(9)$$

Tabel 2. 4 Contoh perhitungan MSE

Tinggi (m)	Berat (kg)	Prediksi Berat (kg)	$(y - \hat{y})$	$(y - \hat{y})^2$
1.6	88	74.2	13.8	190.44
1.6	76	74.2	1.8	3.24
1.5	56	71.6	-15.6	243.36
			Jumlah	437.04

$$MSE = \frac{437.04}{3} = 145.68$$

Adapun RMSE merupakan sebuah metode evaluasi yang digunakan untuk mengukur kinerja model dengan cara menghitung seberapa besar kesalahan antara nilai observasi dan nilai prediksi. Semakin rendah nilai RMSE, maka tingkat akurasi model akan semakin tinggi (Chai dan Draxler 2014). Metode ini juga dikenal sebagai indikator akurasi dalam sistem rekomendasi, di mana RMSE memberikan penalti yang lebih besar untuk perbedaan yang signifikan antara hasil aktual dan prediksi (Gunawan, Tania, dan Suhartono 2016) yang dapat disesuaikan dengan persamaan sebagai berikut:

$$RMSE = \sqrt{MSE} = \sqrt{\sum_{i=1}^{N} \frac{(y-\hat{y})^2}{n}} \quad \ldots(10)$$

MSE dan RMSE menghukum kesalahan prediksi yang besar dibandingkan dengan MAE. Namun, RMSE lebih banyak digunakan daripada MSE untuk mengevaluasi kinerja model regresi dengan model acak lainnya karena memiliki unit yang sama dengan variabel dependen (sumbu Y). MSE adalah fungsi yang dapat dibedakan yang memudahkan untuk melakukan operasi matematika dibandingkan dengan fungsi yang tidak dapat dibedakan seperti MAE. Oleh karena itu, dalam banyak model, RMSE digunakan sebagai metrik default untuk menghitung Loss Function meskipun lebih sulit untuk diinterpretasikan daripada MAE. Nilai MAE, MSE, dan RMSE yang paling rendah menandakan akurasi model regresi tinggi. RMSE menunjukkan seberapa baik suatu model regresi dapat memprediksi nilai suatu variabel respon secara absolut. Berikut ini contoh evaluasi keakuratan prediksi.

Tabel 2. 5 Contoh Perhitungan RMSE

Tinggi (m)	Berat (kg)	Prediksi Berat (kg)	$(y - \hat{y})$	$(y - \hat{y})^2$
1.6	88	74.2	13.8	190.44
1.6	76	74.2	1.8	3.24
1.5	56	71.6	-15.6	243.36
			Jumlah	437.04

$$MSE = \frac{437.04}{3} = 145.68$$

$$RMSE = \sqrt{MSE}$$

$$RMSE = \sqrt{145.68} = 12.07$$

R-squared (R^2) adalah sebuah metrik yang menunjukkan sejauh mana variabel independen (eksogen) memengaruhi variabel dependen (endogen). R^2 memiliki rentang nilai antara 0 hingga 1, mencerminkan sejauh mana kombinasi variabel independen secara bersama-sama mempengaruhi nilai variabel dependen. Penggunaan nilai R^2 membantu dalam mengevaluasi sejauh mana suatu variabel laten independen tertentu berkontribusi terhadap variabel laten dependen (Hair dkk. 2011). Semakin tinggi nilai R^2 menunjukkan bahwa model prediksi sangat baik untuk digunakan (Ghozali 2016).