

BAB III

METODOLOGI PENELITIAN

3.1 Gambaran Umum Objek Penelitian

Objek penelitian dalam mengembangkan model prediksi untuk rekomendasi mata kuliah peminatan mahasiswa Sistem Informasi UMN adalah data performa akademik mahasiswa Sistem Informasi UMN Angkatan 2021-2023, yang terdiri dari data nilai mata kuliah yang diperoleh mahasiswa Program Studi Sistem Informasi. Objek penelitian berupa data-data tersebut diperoleh dari Biro Informasi Akademik UMN, yang dibantu melalui dosen pembimbing penelitian skripsi ini. Meski menggunakan data akademik mahasiswa, data ini telah di-*masking* sehingga kerahasiaan data akademik mahasiswa tetap terjaga. Dalam mengembangkan model prediksi peminatan mahasiswa Sistem Informasi UMN, model tersebut akan dilatih berdasar nilai akademik mahasiswa dan akan melakukan prediksi mata kuliah peminatan yang paling cocok dengan mahasiswa yang bersangkutan.

Penelitian ini menguji model terbaik yang mampu menghasilkan akurasi paling tinggi dalam memprediksi peminatan yang paling cocok bagi mahasiswa Program Studi Sistem Informasi UMN yang hendak mengambil peminatan. Penelitian ini berfokus pada pengembangan model dengan menggunakan algoritma Decision Tree, Random Forest dan Extreme Gradient Boosting (XGBoost) untuk melakukan prediksi peminatan mahasiswa. Adapun alasan penggunaan ketiga algoritma dikarenakan pada penelitian terdahulu, algoritma-algoritma ini mampu menghasilkan akurasi tertinggi dalam melakukan klasifikasi [14], [15], [16], [17], [18].

3.2 Metode Penelitian

Dalam pengembangan model *machine learning*, pemilihan metodologi kerja sangatlah penting untuk mencapai keberhasilan. Pada dunia *data mining*, terdapat sebuah *framework* yang paling sering digunakan, yaitu CRISP-DM (*Cross-Industry Standard Process for Data Mining*) yang telah lama menjadi standar dalam proses *data mining*. Namun, seiring berkembangnya teknologi dan kebutuhan industri, muncul *framework* baru yaitu CRISP-ML yang lebih spesifik dikhususkan pada proyek *machine learning*. Dengan kata lain, CRISP-ML merupakan evolusi dari CRISP-DM. Kedua *framework* ini memiliki pendekatan dan fokus yang sedikit berbeda, namun masing-masingnya memberikan struktur yang kokoh untuk mengelola dan menganalisis data secara efektif. Dengan memahami perbedaan dan kelebihan masing-masing *framework*, pengguna *data mining* dapat memilih pendekatan yang paling sesuai dengan kebutuhan dan tujuan penelitian. Berikut ini adalah perbandingan *framework* CRISP-DM, dan CRISP-ML [32]:

Tabel 3.1 Perbandingan *framework* CRISP-DM dan CRISP-ML

Framework	Tahapan	Keunggulan	Kekurangan
CRISP-DM	1. Business Understanding 2. Data Understanding 3. Data Preparation 4. Modeling 5. Evaluation 6. Deployment	Memiliki struktur yang terdefinisi dengan baik, sehingga memudahkan tim dalam mengikuti langkah-langkah proses <i>data mining</i> .	Cenderung berfokus pada pengembangan model, tanpa memerhatikan pemeliharaan dalam jangka panjang.
CRISP-ML	1. Business and Data Understanding 2. Data Engineering (Data Preparation) 3. Machine Learning Model Engineering 4. Evaluation 5. Deployment 6. Monitoring & Maintenance	Mendukung deployment, monitoring, dan retraining model secara otomatis dengan <i>quality assurance</i> sejak awal, sehingga ideal untuk proyek ML skala besar dan berkelanjutan.	Metodologi ini cukup kompleks, membutuhkan pemahaman mendalam serta infrastruktur tambahan, sehingga kurang cocok untuk proyek sederhana.

Berdasarkan perbandingan antara *framework* CRISP-DM dan CRISP-ML pada Tabel 3.1, dapat disimpulkan bahwa *framework* yang paling cocok untuk digunakan dalam penelitian ini adalah CRISP-ML. Meskipun CRISP-DM menjadi salah satu metodologi yang sempat populer dalam pengembangan proyek data mining, namun seiring perkembangan teknologi dan kompleksitas dalam implementasi *machine learning* modern, metodologi ini dinilai kurang memadai karena hanya berfokus pada pengolahan data statis serta pengembangan model tanpa mempertimbangkan aspek penting seperti generalisasi, reusabilitas, dan pemeliharaan model setelah *deployment*. Selain itu, CRISP-DM tidak dirancang untuk menangani dinamika data yang terus berubah yang seringkali terjadi ketika sudah implementasi. Sedangkan pada penelitian ini, sangat penting untuk memelihara model karena akan dipergunakan dalam jangka panjang, terutama dalam memberikan rekomendasi peminatan bagi mahasiswa Sistem Informasi UMN setiap tahun.

Penelitian ini menggunakan metode CRISP-ML, dikarenakan metode ini lebih komprehensif dan adaptif terhadap kebutuhan *machine learning* yang beragam. CRISP-ML dapat membantu memastikan bahwa seluruh proses pengembangan model prediksi dilakukan secara terstruktur, mulai dari pemahaman yang mendalam tentang latar belakang masalah hingga penerapan model yang akurat bahkan hingga pemeliharaan model setelah *deployment*. Hal ini dikarenakan CRISP-ML mengintegrasikan *quality assurance* sejak tahap awal pengembangan model, serta mendukung strategi *retraining* model secara berkelanjutan untuk memastikan hasil yang tetap akurat. Tidak hanya itu, metodologi ini juga menekankan pentingnya *monitoring* performa model setelah *deployment*, sehingga

model yang diterapkan tetap relevan dan akurat terhadap data yang terus berkembang. Dengan menggunakan CRISP-ML, penelitian ini dapat memastikan bahwa setiap langkah dalam proses pengembangan model prediksi yang sistematis dan efisien, sehingga mampu menghasilkan rekomendasi peminatan yang lebih akurat dan relevan bagi mahasiswa.

Menghadapi tantangan dalam pengembangan sistem rekomendasi berbasis machine learning yang akan dipakai secara berkelanjutan, dibutuhkan sebuah pendekatan metodologis yang mampu mencakup seluruh siklus hidup model secara menyeluruh sampai dengan pemeliharaan model supaya tetap relevan digunakan dalam jangka panjang. *Framework* yang dipilih harus mencakup keseluruhan proses, mulai dari pengembangan, *deployment*, hingga monitoring performa model dalam jangka panjang, serta strategi pembaruan (*retraining*) yang berkelanjutan agar sistem dapat tetap relevan terhadap perubahan data dan kebutuhan pengguna dari waktu ke waktu. Dengan begitu, penelitian ini menggunakan *framework* CRISP-ML, sebagai metodologi yang lebih modern dalam melakukan *data mining*. *Framework* ini juga dipilih berdasarkan tujuan pengembangan model prediksi peminatan mahasiswa yang akan diimplementasikan dalam bentuk sebuah situs web. Berdasar pada fokus penelitian ini dalam memberikan rekomendasi peminatan, CRISP-ML menjadi *framework* yang paling sesuai untuk diterapkan. Berikut ini adalah langkah-langkah implementasi proses CRISP-ML pada penelitian ini:

1. *Business Understanding and Data Understanding*

Penelitian ini dilatarbelakangi oleh masalah keterbatasan akses dan pemahaman mahasiswa terhadap informasi mengenai mata kuliah peminatan di Program Studi Sistem Informasi Universitas Multimedia Nusantara (UMN), yang menyebabkan sebagian besar mahasiswa mengalami kesulitan dalam menentukan peminatan yang sesuai dengan minat dan kemampuan akademik mereka. Hasil survei yang dilakukan menunjukkan bahwa 90% mahasiswa angkatan 2024 mengalami kesulitan dalam memilih peminatan, dengan faktor utama berupa kurangnya pemahaman terhadap isi dan prospek mata kuliah peminatan yang tersedia. Kondisi ini diperparah dengan minimnya arahan dari pihak program studi serta kurangnya integrasi teknologi dalam mendukung proses pengambilan keputusan akademik. Oleh karena itu, penelitian ini bertujuan untuk mengembangkan sistem prediksi peminatan berbasis *machine learning* yang dapat memberikan rekomendasi, sehingga dapat membantu mahasiswa dalam menentukan peminatan secara lebih terarah berdasarkan performa akademik mereka.

Tujuan utama penelitian ini adalah untuk membangun model prediksi peminatan mahasiswa Program Studi Sistem Informasi di Universitas Multimedia Nusantara (UMN) berdasarkan data performa akademik. Pengembangan model prediksi pada penelitian ini dapat

bermanfaat bagi mahasiswa yang mengalami kesulitan menemukan mata kuliah peminatan. Model ini diharapkan dapat membantu mahasiswa dalam mengambil keputusan yang tepat terkait pemilihan peminatan, serta memberikan wawasan bagi pihak program studi dalam proses pembinaan akademik.

Dalam penelitian ini, dataset yang digunakan adalah data performa akademik mahasiswa Sistem Informasi UMN angkatan 2021-2023. Dataset ini memuat nilai akademik mahasiswa seperti IPK dan nilai per mata kuliah yang pernah diambil, serta pilihan peminatan sebagai variabel target. Dataset ini terpisah-pisah menjadi beberapa dataset, yaitu dataset angkatan 2023 semester 1, dataset angkatan 2023 semester 2, dataset peminatan angkatan 2023, dan dataset angkatan 2021-2022. Secara keseluruhan, setelah dataset ini digabungkan menjadi 1 dataset angkatan 2021-2023, terdapat 518 jumlah baris dan 11 kolom yang terdiri dari "ID", "PEMINATAN", "PEMINATAN_2", "Algorithms and Programming", "Database Systems", "Discrete Mathematics", "IS Analysis & Design", "Operating System Concept", "Pancasila", "Probability and Statistic", "Seminars on Information System". Pada tahap ini juga dilakukan analisis deskriptif dilakukan untuk memahami distribusi data, mengidentifikasi anomali, dan pengecekan *missing value*.

2. *Data Engineering (Data Preparation)*

Setelah pemahaman data, tahap selanjutnya adalah persiapan data. Hal yang akan dilakukan pada tahap ini seperti pembersihan data, integrasi data dari berbagai sumber jika diperlukan, transformasi data, dan pemilihan data yang relevan. Pada tahap ini perlu dipastikan bahwa data telah siap digunakan untuk proses selanjutnya yaitu pelatihan model. Berikut merupakan tahap persiapan data yang dilewati:

1. Penanganan nilai null atau nilai yang hilang (*missing value*)

Hal pertama yang dapat dilakukan dalam data preparation adalah pengecekan *missing value*. Apabila terdapat jenis data yang hilang atau *missing value*, maka perlu ditangani dengan menghilangkan baris data tersebut atau mengisi kembali baris tersebut dengan data baru yang sesuai, seperti mean, dll.

2. *Feature Selection*

Pemilihan fitur dilakukan untuk menentukan variabel yang akan digunakan dalam pembentukan model prediksi. Fitur-fitur ini dapat mencakup data-data yang memiliki tingkat korelasi tinggi yang didapatkan melalui perhitungan *feature importance* untuk meningkatkan akurasi model prediksi.

3. *Hyperparameter Tuning*

Hyperparameter tuning dilakukan untuk mencari parameter terbaik dari algoritma yang digunakan. Nantinya parameter terbaik akan digunakan untuk menghasilkan akurasi model yang

4. *Data Balancing* dengan SMOTE

Karena data peminatan berpotensi tidak seimbang antar kelas, digunakan metode Synthetic Minority Over-sampling Technique (SMOTE) untuk menyeimbangkan jumlah data antar kelas.

5. *Data Splitting*

Untuk merancang sebuah model, perlu membagi data menjadi dua bagian, yakni data pelatihan dan data pengujian. Maka pada tahap ini akan membagi data menjadi data latih dan uji dengan rasio 80:20.

3. *Machine Learning Model Engineering*

Tahap ini merupakan tahap perancangan model sistem rekomendasi buku dan penggunaan teknik *machine learning*. Pada tahap ini, tiga algoritma yang akan digunakan dalam proses klasifikasi, yaitu Decision Tree, Random Forest, XGBoost. Untuk masing-masing algoritma, akan dilakukan proses:

- *Model training* menggunakan data latih.
- *Hyperparameter tuning* menggunakan metode Randomized Search untuk menemukan konfigurasi paling optimal.
- *Cross-validation* (misalnya k-Fold) untuk memastikan model tidak *overfitting* dan memiliki generalisasi yang baik.

4. *Evaluation*

Setelah model dibuat, tahap selanjutnya adalah tahap evaluasi yang akan mengevaluasi kinerja model prediksi yang telah dirancang sebelumnya, bisa menggunakan metrik evaluasi seperti akurasi, recall, precision, dan f1-score. Evaluasi ini tidak hanya dilakukan untuk membandingkan performa ketiga algoritma, tetapi juga untuk menilai pengaruh dari teknik-teknik peningkatan performa seperti *feature selection*, SMOTE, dan *hyperparameter tuning*.

5. *Deployment*

Setelah model dievaluasi dan memiliki tingkat *error* yang rendah, tahap selanjutnya adalah dengan mengimplementasikan model tersebut. Hal ini dapat dilakukan dengan mengintegrasikan sistem rekomendasi ke dalam

aplikasi berbasis web sehingga dapat digunakan oleh mahasiswa. Aplikasi berbasis web tersebut dirancang menggunakan Streamlit. Penggunaan Streamlit didasarkan pada keunggulannya yang memungkinkan seseorang membuat aplikasi web interaktif untuk penerapan model *machine learning*.

6. *Monitoring & Maintenance*

Model prediksi ini perlu dimonitor secara berkala. Dalam penelitian ini, tahap ini dapat berupa analisis keterbatasan dan potensi pengembangan model di masa depan, serta usulan penelitian lanjutan seperti pengujian model mahasiswa angkatan berikutnya atau integrasi dengan data non-akademik.

3.3 Teknik Pengumpulan Data

Data yang digunakan dalam penelitian ini adalah data performa akademik mahasiswa Sistem Informasi UMN angkatan 2021-2023. Dataset ini memuat nilai akademik mahasiswa seperti IPK, nilai per mata kuliah, dan pilihan peminatan. Data ini terbagi menjadi 3 dataset terpisah berdasarkan angkatan. Informasi yang sifatnya personal atau rahasia seperti nama dan NIM telah di-*masking*. Data ini diperoleh dari Biro Informasi Akademik UMN dan telah disetujui oleh dosen pembimbing skripsi.

3.4 Variabel Penelitian

Variabel penelitian menggunakan data akademik mahasiswa yang diperoleh dari dosen pembimbing skripsi. Pada penelitian ini menggunakan dua buah variabel data yang meliputi independen dan dependen.

1. Variabel Independen

Variabel independen merupakan faktor atau karakteristik yang dianggap sebagai penyebab atau pemicu perubahan pada variabel dependen. Dalam penelitian, variabel independen adalah faktor yang dapat diubah atau dimanipulasi oleh peneliti untuk mengukur dampaknya terhadap variabel dependen. Variabel independen dapat berupa penyebab yang memengaruhi atau memprediksi perubahan pada variabel dependen. Variabel independen pada penelitian ini meliputi nilai mata kuliah dan nilai IPK.

2. Variabel Dependen

Variabel dependen merupakan variabel yang nilainya bergantung pada variabel independen. Dalam penelitian, variabel dependen adalah hasil yang diamati atau diukur untuk menilai efek dari variabel independen. Variabel dependen akan berubah secara otomatis atau berdasarkan perubahan yang terjadi pada variabel independen. Dengan menganalisis hubungan antara variabel independen dan dependen, peneliti dapat mengetahui sejauh mana pengaruh variabel independen terhadap

variabel dependen. Variabel dependen pada penelitian ini adalah mata kuliah peminatan yang diprediksi atau direkomendasikan oleh sistem.

3.5 Teknik Analisis Data

Penelitian ini menggunakan teknik analisis data kuantitatif untuk mengumpulkan dan menganalisis data numerik yang dapat diukur secara sistematis, seperti data penilaian pembaca terhadap buku-buku tertentu. Teknik analisis data ini memungkinkan analisis yang mendalam terhadap efektivitas model prediksi yang dikembangkan dengan bahasa pemrograman Python dan algoritma klasifikasi *machine learning*, yaitu Decision Tree, Random Forest, dan XGBoost. Penelitian ini juga mengadopsi metode data mining CRISP-ML dalam membantu perancangan sistem rekomendasi buku.

Alur kerja teknik analisis pada penelitian ini yang didasarkan pada metode CRISP-ML. Berdasarkan alur tersebut, data diolah dalam tahap persiapan data yang terdiri dari *data cleaning*, *feature selection*, *hyperparameter tuning*, *data balancing*, dan *data splitting*. Setelah data siap, dilakukan pemodelan data dengan algoritma klasifikasi yaitu Decision Tree, Random Forest, dan XGboost. Setelah itu dilakukan evaluasi dengan beberapa metrik evaluasi. Dan setelah melewati semua tahap tersebut, sistem ini akan di-*deploy* dalam bentuk *website* menggunakan Streamlit.

Terdapat beberapa tools yang digunakan dalam penelitian ini untuk mendukung analisis data berdasarkan alur kerja yang telah dirancang menggunakan bahasa pemrograman Python dan tools Visual Studio Code. Pemilihan bahasa pemrograman Python, tools Visual Studio Code, dan Streamlit didasarkan atas keunggulan yang dimiliki oleh berbagai tools ini dalam membantu merancang sistem rekomendasi sebagaimana yang telah dijelaskan pada bagian 2.4. Sehingga, dengan memanfaatkan Python, Visual Studio Code, dan Streamlit, penelitian ini memungkinkan pengembangan model prediksi mata kuliah peminatan dengan baik.