BAB III
METODE PENELITIAN

A. Jenis Penelitian

Penelitian ini merupakan penelitian kuantitatif prediktif berbasis data mining yang bertujuan membangun dan mengevaluasi model prediksi prestasi siswa menggunakan data historis siswa. Pendekatan kuantitatif digunakan karena seluruh variabel penelitian dinyatakan dalam bentuk numerik, yaitu nilai prestasi siswa, tingkat kedisiplinan, status ekonomi dan motivasi belajar yang diolah menggunakan algoritma regresi linier berganda dan *support vector regression*(SVR) untuk menghasilkan model prediksi.

B. Setting Penelitian

1. Lokasi Penelitian

   Penelitian ini dilakukan di SMKN 1 Sinunukan, Kabupaten Mandailing Natal, Provinsi Sumatera Utara.

2. Waktu Penelitian

   Penelitian ini direncanakan akan dilaksanakan selama kurang lebih Enam bulan, terhitung sejak Desember 2025 hingga Mei 2026.

C. Bahan Penelitian

1. Sumber Dan Jumlah Data

   a. Data primer: diperoleh melalui penyebaran kuesioner motivasi belajar kepada siswa(motivasi belajar)

   b. Data sekunder: diperoleh dari administrasi sekolah (nilai rapor, kehadiran, dan data ekonomi siswa)

   Berdasarkan data resmi SMKN 1 Sinunukan tahun ajaran 2024/2025, jumlah populasi sebanyak 374 siswa. Data penelitian diperoleh dari beberapa sumber, yaitu data Status Ekonomi, data Kedisiplinan, data Prestasi Siswa, dan data Motivasi Belajar. Karena data tersebut berasal dari sumber yang berbeda, dilakukan proses pencocokan dan penggabungan data berdasarkan identitas siswa untuk memastikan bahwa setiap data berasal dari siswa yang sama dan sesuai dengan variabel penelitian. Pada saat proses pengumpulan data dari sekolah,

UIN IMAM BONJOL PADANG

sebagian siswa sedang melaksanakan Praktik Lapangan (PL), sedangkan sebagian lainnya telah menyelesaikan masa pendidikan. Kondisi tersebut menyebabkan tidak seluruh siswa memiliki data yang lengkap pada seluruh variabel penelitian. Berdasarkan hasil pencocokan dan pemeriksaan kelengkapan data, diperoleh sebanyak 198 siswa yang memiliki data lengkap pada seluruh variabel penelitian. Selanjutnya, 198 data siswa tersebut digunakan sebagai dataset dalam penelitian.

2. Proses Labeling Data

   Proses labelling dilakukan untuk memastikan setiap data memiliki representasi yang sesuai untuk kebutuhan pemodelan.

   a. Label target (Y): Prestasi siswa direpresentasikan dalam bentuk nilai rapor (numerik), sehingga tidak memerlukan transformasi tambahan.
   b. Label variabel ekonomi $(X1)$: Data status ekonomi berupa penghasilan orang tua dalam satuan rupiah.
   c. Label variabel kedisiplinan $(X2)$: Diukur berdasarkan persentase kehadiran siswa, yaitu perbandingan antara jumlah hari siswa hadir dengan jumlah hari efektif pembelajaran. Hasil persentase tersebut kemudian dikonversi ke bentuk desimal dan digunakan sebagai data numerik dalam proses analisis.
   d. Label variabel motivasi $(X3)$: Data hasil kuesioner Likert dikonversi menjadi skor numerik melalui proses:
      1) Penjumlahan skor tiap item
      2) Perhitungan rata-rata
      3) Digunakan sebagai nilai akhir motivasi

   Proses labelling ini bertujuan agar seluruh variabel dapat diproses.

3. Normalisasi Data

   Normalisasi data dilakukan untuk menyeragamkan skala atau rentang nilai antarvariabel numerik, sehingga perbedaan satuan dan besarnya nilai antarvariabel tidak menyebabkan salah satu variabel

UIN IMAM BONJOL PADANG

lebih dominan dalam proses pemodelan. Hal ini penting terutama pada algoritma *Support Vector Regression* (SVR) yang menggunakan perhitungan berdasarkan jarak dan skala nilai data, sehingga perbedaan rentang antarvariabel dapat memengaruhi proses pembentukan model.

Pada penelitian ini digunakan metode Min–Max Normalization untuk mengubah nilai setiap variabel numerik ke dalam rentang 0 sampai 1. Dengan demikian, setiap variabel memiliki skala yang seragam sebelum digunakan dalam proses pemodelan. Transformasi Min–Max dilakukan menggunakan rumus yang dikemukakan oleh Pradasyah & Baita, (2025):

$$X' = \frac{X - X_{min}}{X_{max} - X_{min}} \quad (6)$$

Metode ini mengubah nilai setiap variabel ke dalam rentang 0 hingga 1, sehingga:

a. mempercepat proses pelatihan model,
b. menghindari dominasi variabel dengan skala besar,
c. meningkatkan stabilitas dan akurasi model SVR.

Variabel yang dinormalisasi pada penelitian ini adalah variabel numerik yang digunakan sebagai input model, yaitu status ekonomi, kedisiplinan dan motivasi belajar siswa agar seluruh variabel berada pada skala yang sebanding sebelum proses pemodelan dilakukan.

4. Contoh Dataset

Tabel ini hanya menampilkan contoh dataset untuk menggambarkan struktur dan format data yang digunakan dalam penelitian.

Tabel 3. 1 Contoh dataset

ID	Nama	Total Penghasilan	Total_ Hadir	Persentase Kehadiran	Nilai Rapor
1	Siswa 1	3.500.000	47	0.783	80
2	Siswa 2	2.800.000	40	0.667	75

UIN IMAM BONJOL PADANG

3	Siswa 3	1.800.000	47	0.783	88
4	Siswa 4	4.000.000	40	0.667	82
5	Siswa 5	5.200.000	42	0.700	85

5. Variabel Penelitian

Variabel yang digunakan dalam penelitian ini terdiri dari variabel independen (variabel bebas) dan variabel dependen (variabel terikat). Setiap variabel didefinisikan secara operasional agar dapat digunakan dalam proses pemodelan prediksi menggunakan metode regresi linier berganda dan *support vector regression* (SVR).

a. Status Ekonomi (X_1)

Status ekonomi merupakan variabel independen yang merepresentasikan kondisi ekonomi keluarga siswa. Variabel ini diperoleh dari data administratif sekolah berupa penghasilan orang tua, yaitu hasil penjumlahan antara penghasilan ayah dan penghasilan ibu.

Jika salah satu orang tua tidak memiliki penghasilan tetap atau tidak bekerja, maka nilai penghasilannya dianggap nol agar perhitungan total pendapatan tetap konsisten.

Variabel ini dinyatakan dalam bentuk numerik (rupiah) dan digunakan secara langsung dalam proses pemodelan tanpa dilakukan kategorisasi, sehingga tetap mempertahankan informasi kontinu yang lebih detail mengenai kondisi ekonomi siswa.

b. Kedisiplinan (X_2)

Kedisiplinan merupakan variabel independen yang mencerminkan tingkat kepatuhan siswa dalam mengikuti kegiatan pembelajaran. Variabel ini diukur menggunakan persentase kehadiran siswa selama satu periode pembelajaran.

Nilai persentase ini kemudian dikonversi ke dalam bentuk desimal (0–1) untuk memudahkan proses pengolahan data dalam model.

UIN IMAM BONJOL PADANG

Semakin tinggi persentase kehadiran, maka semakin tinggi tingkat kedisiplinan siswa dalam mengikuti proses belajar mengajar. Variabel ini digunakan secara langsung dalam bentuk numerik dalam proses pemodelan.

c. Motivasi Belajar (X_3)

merupakan variabel independen yang menggambarkan dorongan siswa dalam mengikuti proses pembelajaran. Data motivasi diperoleh melalui kuesioner menggunakan skala Likert dengan rentang nilai 1 sampai 5. Penggunaan kuesioner untuk mengukur motivasi belajar dengan skala Likert juga diterapkan dalam penelitian Utami dkk (2026). Setiap item pernyataan mencerminkan aspek motivasi belajar, seperti semangat belajar, ketekunan, minat, dan keinginan untuk berprestasi. Pengukuran motivasi belajar melalui instrumen yang disusun berdasarkan indikator dan diuji validitas serta reliabilitasnya juga diterapkan pada penelitian siswa SMK oleh Ningsih dkk (2025).

Setiap item pernyataan dalam kuesioner mencerminkan aspek motivasi belajar, seperti:

1) semangat dalam belajar
2) ketekunan dalam menyelesaikan tugas
3) minat terhadap pelajaran
4) keinginan untuk berprestasi

Skor motivasi dihitung dengan cara:

1) Menjumlahkan seluruh skor item kuesioner
2) Menghitung rata-rata skor

Nilai rata-rata tersebut digunakan sebagai representasi numerik dari motivasi belajar siswa dan digunakan secara langsung dalam proses pemodelan.

d. Prestasi Siswa (Y)

Prestasi siswa (Y) dalam penelitian ini direpresentasikan oleh nilai rapor siswa, sehingga nilai rapor digunakan sebagai variabel target

UIN IMAM BONJOL PADANG

dalam proses pemodelan prediksi. Nilai rapor dipilih karena mencerminkan prestasi siswa secara menyeluruh, memiliki bentuk data numerik, dan tersedia secara konsisten dalam data sekolah, sehingga sesuai digunakan sebagai variabel target dalam proses pemodelan prediksi menggunakan algoritma regresi linier berganda dan *support vector regression*.

6. Pembagian Data Training dan Testing

Setelah melalui tahap *data preparation*, dataset yang telah siap digunakan untuk pemodelan terdiri dari 198 data siswa. Dataset tersebut dibagi menjadi data *training* dan data *testing* dengan proporsi 80:20. Menurut Joseph dkk (2022), pembagian data dengan proporsi 80:20 merupakan salah satu rasio yang umum digunakan, yaitu 80% data digunakan sebagai data *training* dan 20% data digunakan sebagai data *testing*. Pembagian tersebut memungkinkan model dibangun menggunakan sebagian besar data, sementara sebagian data lainnya tetap dipisahkan untuk mengevaluasi kemampuan prediksi model terhadap data yang tidak digunakan dalam proses pelatihan.

Berdasarkan proporsi tersebut, dari 198 data siswa diperoleh 158 data sebagai data *training* dan 40 data sebagai data *testing*. Pembagian data dilakukan menggunakan fungsi train_test_split dari *library* Scikit-learn dengan test_size=0.20 dan random_state=42. Penggunaan random_state=42 bertujuan agar proses pembagian data dapat direproduksi sehingga menghasilkan pembagian data yang sama ketika proses dijalankan kembali.

Data *training* digunakan untuk membangun model Regresi Linier Berganda (RLB) dan Support Vector Regression (SVR). Pada SVR, proses pencarian kombinasi parameter dilakukan menggunakan *GridSearchCV* terhadap data *training* dengan metode *cross-validation*. Setelah diperoleh kombinasi parameter terbaik, model SVR dibentuk menggunakan data *training* dan selanjutnya digunakan untuk melakukan prediksi terhadap data *testing*. Data *testing* tidak digunakan dalam proses

UIN IMAM BONJOL PADANG

pelatihan maupun dalam penentuan parameter model sehingga dapat digunakan untuk mengukur kemampuan model dalam melakukan prediksi terhadap data yang belum digunakan sebelumnya.

Untuk menjaga konsistensi proses pemodelan, normalisasi Min-Max pada variabel prediktor dilakukan dengan membentuk parameter normalisasi berdasarkan data *training*. Parameter normalisasi tersebut kemudian diterapkan pada data *testing*. Dengan demikian, informasi dari data *testing* tidak digunakan dalam proses pembentukan model.

Selanjutnya, performa model dievaluasi menggunakan data *testing* dengan tiga metrik, yaitu *Mean Squared Error* (MSE), *Root Mean Squared Error* (RMSE), dan *Coefficient of Determination* (R^2). MSE dan RMSE digunakan untuk mengukur tingkat kesalahan hasil prediksi, sedangkan R^2 digunakan untuk mengetahui proporsi variasi nilai prestasi siswa yang dapat dijelaskan oleh model.

D. Alat Penelitian

Alat penelitian terdiri dari perangkat lunak dan perangkat keras yang digunakan untuk mengolah data dan membangun model prediksi.

1. Perangkat Lunak *(Software)*

   Perangkat lunak yang digunakan dalam penelitian ini adalah:

   a. Sistem Operasi Windows 11
   b. Bahasa pemrograman Python (menggunakan library seperti pandas, numpy, sklearn, dan matplotlib)
   c. *Google Colab* untuk menjalankan pemodelan
   d. *Microsoft Excel* untuk pemeriksaan dan persiapan awal dataset
   e. Browser sebagai akses pendukung literatur dan dokumentasi

2. Perangkat Keras *(Hardware)*

   Perangkat keras yang digunakan dalam penelitian ini meliputi:

   a. Laptop Lenovo LOQ 15IAX9 dengan prosesor Intel Core i5-12450HX
   b. Memori RAM 12 GB DDR5
   c. Penyimpanan 512 GB SSD

UIN IMAM BONJOL PADANG

d. Kartu grafis NVIDIA GeForce RTX 3050 6GB

E. Prosedur Penelitian

Prosedur penelitian ini disusun secara sistematis berdasarkan tahapan *Cross-Industry Standard Process for Data Mining* (CRISP-DM) untuk memastikan proses pengolahan dan analisis data hingga penerapan model dapat dilakukan secara terstruktur. Tahapan penelitian terdiri atas *Business Understanding*, *Data Understanding*, *Data Preparation*, *Modeling*, *Evaluation*, dan *Deployment*.

1. Business Understanding

   Tahap pertama adalah *business understanding*, yaitu memahami permasalahan dan kebutuhan penelitian pada SMKN 1 Sinunukan. Pada tahap ini ditentukan tujuan penelitian, yaitu membangun model prediksi prestasi siswa berdasarkan status ekonomi, kedisiplinan, dan motivasi belajar. Hasil yang diharapkan adalah diperolehnya model yang mampu memberikan prediksi prestasi siswa serta dapat dibandingkan performanya menggunakan metode Regresi Linier Berganda dan *Support Vector Regression* (SVR).

2. Data Understanding

   Tahap kedua adalah *data understanding*, yaitu memahami data yang diperoleh dari pihak sekolah dan hasil penyebaran kuesioner kepada siswa. Data sekunder yang digunakan meliputi nilai rapor, data kehadiran siswa, dan data penghasilan orang tua, sedangkan data primer diperoleh melalui kuesioner untuk mengukur motivasi belajar siswa. Pada tahap ini dilakukan pemeriksaan terhadap struktur dataset, jumlah data, jenis data, kelengkapan data, serta identifikasi kemungkinan adanya data yang tidak lengkap atau tidak konsisten.

   Selain itu, pada tahap ini dilakukan pengujian validitas dan reliabilitas terhadap instrumen kuesioner motivasi belajar untuk memastikan bahwa instrumen yang digunakan mampu mengukur variabel motivasi belajar secara tepat dan konsisten. Uji validitas dilakukan terhadap setiap butir pernyataan kuesioner, sedangkan uji reliabilitas dilakukan

UIN IMAM BONJOL PADANG

untuk mengetahui konsistensi instrumen sebelum data hasil kuesioner digunakan dalam tahap pengolahan dan pemodelan.

3. Data Preparation

   Tahap ketiga adalah *data preparation*, yaitu mempersiapkan dataset agar dapat digunakan dalam proses pemodelan. Tahapan ini meliputi pembersihan dan penyesuaian format data, penggabungan data yang diperlukan, pengolahan variabel penelitian, Selain itu, dilakukan penentuan variabel independen yang terdiri dari status ekonomi, kedisiplinan, dan motivasi belajar, serta variabel dependen berupa prestasi siswa. Data yang telah dipersiapkan kemudian digunakan sebagai input dalam proses pemodelan.

4. Modeling

   Tahap keempat adalah *modeling*, yaitu membangun model prediksi menggunakan dua metode, yaitu Regresi Linier Berganda dan *Support Vector Regression* (SVR). Dataset yang telah melalui tahap *data preparation* dibagi menjadi data *training* dan data *testing* dengan proporsi 80:20. Data *training* digunakan dalam proses pembangunan model, sedangkan data *testing* digunakan pada tahap evaluasi untuk mengukur kemampuan model dalam melakukan prediksi terhadap data yang belum digunakan dalam proses pelatihan.

   Regresi Linier Berganda digunakan untuk memodelkan hubungan linier antara status ekonomi, kedisiplinan, dan motivasi belajar terhadap prestasi siswa. Sementara itu, SVR digunakan untuk membangun model regresi yang mampu menangkap hubungan yang lebih kompleks. Pada model SVR dilakukan proses pencarian kombinasi parameter menggunakan *GridSearchCV* pada data *training* untuk memperoleh parameter yang digunakan dalam membangun model SVR.

5. Evaluation

   Tahap kelima adalah *evaluation*, yaitu mengevaluasi dan membandingkan kinerja model Regresi Linier Berganda dan SVR. Evaluasi dilakukan menggunakan data *testing* yang telah dipisahkan

UIN IMAM BONJOL PADANG

pada tahap *modeling* dengan proporsi 20% dari keseluruhan data. Kinerja masing-masing model diukur menggunakan *Mean Squared Error* (MSE), *Root Mean Squared Error* (RMSE), dan *Coefficient of Determination* (R^2). MSE dan RMSE digunakan untuk mengukur tingkat kesalahan prediksi, sedangkan R^2 digunakan untuk mengetahui kemampuan model dalam menjelaskan variasi nilai Prestasi Siswa. Hasil evaluasi digunakan untuk mengetahui dan membandingkan kinerja Regresi Linier Berganda dan SVR berdasarkan nilai MSE, RMSE, dan R^2.

6. Deployment

   Tahap keenam adalah *deployment*, yaitu menerapkan model yang telah melalui proses pemodelan dan evaluasi ke dalam bentuk aplikasi prediksi menggunakan Streamlit. Aplikasi digunakan sebagai implementasi hasil penelitian sehingga pengguna dapat memasukkan data yang diperlukan dan memperoleh hasil prediksi Prestasi Siswa berdasarkan model yang telah dibangun dan dievaluasi.

   Selesai.

UIN IMAM BONJOL
PADANG