

BAB III METODE PENELITIAN

3.1 Objek Penelitian

Objek penelitian dalam studi ini adalah data hasil panen produksi buah-buahan di Provinsi Aceh yang diperoleh dari Badan Pusat Statistik (BPS) selama periode tahun 2018 hingga 2023. Dataset yang digunakan terdiri dari 3.542 entri data yang mencerminkan distribusi produksi buah berdasarkan wilayah dan waktu. Setiap entri dalam dataset memuat sejumlah atribut penting, antara lain: kode provinsi, kode kabupaten/kota, tahun produksi, jenis tanaman buah, serta jumlah produksi dalam satuan kuintal.

Jenis tanaman buah yang tercatat dalam *dataset* sangat bervariasi, mencakup komoditas utama seperti alpukat, mangga, pisang, jeruk, durian, rambutan, dan beberapa buah lainnya yang memiliki peran penting dalam sektor pertanian di Aceh. Informasi ini memberikan gambaran yang komprehensif mengenai keberagaman hasil panen serta tren produksi tiap jenis buah selama enam tahun terakhir.

Untuk mempermudah pengolahan dan analisis data, dilakukan serangkaian proses *data preprocessing* yang mencakup teknik *label encoding* untuk mengubah kategori jenis tanaman menjadi format numerik, serta *one-hot encoding* untuk representasi wilayah produksi (kabupaten/kota). Proses ini bertujuan untuk mengoptimalkan kinerja algoritma pembelajaran mesin yang digunakan dalam pemodelan prediksi.

Dalam tahap pemodelan, dua algoritma utama diterapkan, yaitu **Random Forest** dan **Support Vector Machine (SVM)**. Kedua model ini dipilih karena memiliki keunggulan dalam menangani data yang bersifat non-linear dan kompleks. Faktor-faktor yang menjadi variabel input utama dalam prediksi mencakup lokasi geografis (kabupaten/kota), tahun produksi, dan jenis tanaman. Model yang dikembangkan diharapkan mampu mengidentifikasi pola-pola produksi yang tidak terlihat secara eksplisit, sehingga dapat digunakan untuk memprediksi volume produksi buah pada periode waktu mendatang.

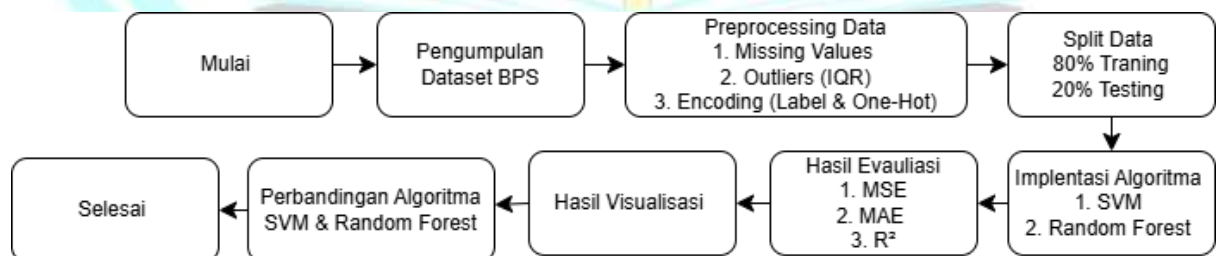
Evaluasi performa model dilakukan dengan menggunakan metrik-metrik akurasi seperti nilai *Mean Absolute Error (MAE)*, *Root Mean Squared Error (RMSE)*, dan *R-squared*

(R^2). Hasil evaluasi ini digunakan untuk membandingkan efektivitas kedua algoritma dalam menangkap pola historis produksi buah. Model dengan performa terbaik nantinya dapat menjadi dasar dalam mendukung pengambilan keputusan berbasis data, khususnya dalam perencanaan strategi pertanian, pengelolaan logistik hasil panen, serta proyeksi ketahanan pangan di wilayah Provinsi Aceh.

Dengan pendekatan ini, penelitian ini tidak hanya memberikan wawasan deskriptif mengenai hasil panen buah, tetapi juga menghadirkan solusi prediktif berbasis kecerdasan buatan (AI) yang bermanfaat bagi pemerintah daerah, petani, dan stakeholder lainnya dalam meningkatkan efisiensi dan produktivitas sektor pertanian lokal.

3.2 Prosedur Penelitian

Gambar 3.1 menggambarkan alur prosedur penelitian yang dilakukan secara terstruktur untuk mencapai tujuan penelitian, yaitu memprediksi produksi buah di Provinsi Aceh menggunakan algoritma machine learning. Penelitian diawali dengan tahap pengumpulan data yang bersumber dari Badan Pusat Statistik (BPS), yang memuat data produksi buah dari berbagai kabupaten/kota dalam rentang waktu tertentu.



Gambar 3.1 Prosedur Penelitian

Selanjutnya, dilakukan tahap preprocessing data untuk memastikan kualitas data yang digunakan dalam pelatihan model. Tahapan ini meliputi penanganan data kosong (missing values), penghapusan outlier menggunakan metode Interquartile Range (IQR), serta proses encoding fitur kategorikal seperti jenis tanaman dan wilayah menggunakan metode label encoding dan one-hot encoding. Langkah ini penting agar data siap digunakan oleh algoritma prediktif. Setelah preprocessing selesai, data dibagi menjadi dua bagian: 80% untuk data latih dan 20% untuk data uji. Selanjutnya, dilakukan implementasi dua algoritma, yaitu Support Vector Machine (SVM) dan Random Forest, untuk membangun model prediksi produksi buah. Masing-masing model kemudian dievaluasi menggunakan tiga metrik utama: Mean Squared Error (MSE), Mean Absolute

Error (MAE), dan R^2 Score, yang menunjukkan tingkat kesalahan dan akurasi model.

Hasil evaluasi kemudian divisualisasikan untuk mempermudah interpretasi dan dibandingkan guna mengetahui algoritma mana yang memberikan performa terbaik. Proses terakhir adalah melakukan perbandingan performa antara SVM dan Random Forest, untuk menentukan model yang paling efektif dalam memprediksi data produksi buah di Aceh secara akurat dan andal.

3.3 Pengumpulan Data

Penelitian ini menggunakan data primer yang diperoleh langsung dari catatan Badan Pusat Statistik (BPS) mengenai produksi buah di Provinsi Aceh selama periode 2018 hingga 2023. *Dataset* terdiri dari 3.542 entri dengan atribut seperti kode provinsi, kode kabupaten/kota, tahun, jenis tanaman, dan jumlah produksi dalam satuan kuintal. Data tersebut kemudian diolah menggunakan *Google Collaboratory* untuk memprediksi hasil produksi berdasarkan pola historis. Setelah data terkumpul, langkah selanjutnya adalah menyusunnya dalam format terstruktur, seperti CSV atau Excel, untuk mempermudah proses pengolahan dan analisis lebih lanjut. Atribut independen yang digunakan meliputi kode provinsi, kode kabupaten/kota, tahun, dan jenis tanaman, sedangkan variabel dependen adalah jumlah produksi. *Dataset* ini menjadi landasan utama dalam pengembangan model *Machine Learning* yang digunakan dalam penelitian ini.

Tabel 3.1 Load Dataset

=== 5 Baris Pertama Dataset ===

	bps_kode_provinsi	bps_nama_provinsi	bps_kode_kabupaten_kota	\
0	11	Aceh	1101	
1	11	Aceh	1102	
2	11	Aceh	1103	
3	11	Aceh	1104	
4	11	Aceh	1105	

	bps_nama_kabupaten_kota	tahun	jenis_tanaman	jumlahproduksi	satuan
0	Kabupaten Simeulue	2018	Alpukat	47.0	Kuintal
1	Kabupaten Aceh Singkil	2018	Alpukat	27.0	Kuintal
2	Kabupaten Aceh Selatan	2018	Alpukat	45.0	Kuintal
3	Kabupaten Aceh Tenggara	2018	Alpukat	197.0	Kuintal
4	Kabupaten Aceh Timur	2018	Alpukat	NaN	Kuintal

3.4 Preprocessing Data

Tahapan *Preprocessing* bertujuan untuk membersihkan data sehingga dapat digunakan dalam proses *Machine Learning* secara maksimal. *Dataset* yang terdiri dari 3.542 entri dengan 8 atribut, yaitu kode provinsi, kode kabupaten/kota, tahun produksi, jenis tanaman, jumlah produksi, dan variabel tambahan hasil encoding, diproses untuk memastikan kualitas data yang optimal. Langkah-langkah utama mencakup penghapusan missing values pada atribut 'jumlah_produksi', deteksi serta penanganan outliers dengan metode IQR untuk menghapus data ekstrem, dan transformasi variabel kategori menjadi format numerik melalui label encoding dan one-hot encoding. Proses ini dirancang untuk mengatasi masalah seperti distribusi data yang tidak normal, noise, atau ketidakseimbangan data, sehingga menghasilkan *Dataset* yang siap digunakan dalam pengembangan model *Machine Learning* yang akurat.

3.4.1 Mengatasi Missing Values

Baris dengan nilai kosong pada kolom “jumlah_produksi” dihapus untuk mencegah bias. Proses ini diawali dengan mengecek jumlah data yang hilang pada kolom-kolom utama, khususnya pada atribut "jumlah_produksi". Setelah identifikasi, baris dengan data yang hilang dihapus, dan jumlah baris sebelum serta sesudah proses ini dicatat. Prosedur ini penting untuk memastikan *Dataset* yang digunakan sepenuhnya valid dan reliabel tanpa adanya celah data kosong yang dapat memengaruhi hasil analisis. Proses ini mengurangi jumlah baris yang tersedia dalam *Dataset* namun meningkatkan validitas data.

3.4.2 Penanganan Outliers

Metode *Interquartile Range (IQR)* digunakan untuk menghapus data ekstrem pada kolom “jumlah_produksi”. Proses ini diawali dengan menghitung nilai kuartil pertama (Q1) dan kuartil ketiga (Q3), kemudian selisih antara Q3 dan Q1 dihitung untuk mendapatkan nilai IQR. Data yang berada di luar rentang $[Q1 - 1.5IQR, Q3 + 1.5IQR]$ dianggap sebagai *outlier* dan dihapus untuk menjaga distribusi data yang lebih normal. Langkah ini dilakukan untuk memastikan data yang digunakan dalam model memiliki distribusi yang tidak bias akibat data ekstrem, sehingga meningkatkan akurasi hasil analisis.

3.4.3 Transformasi Variabel Kategori

Proses transformasi variabel kategori dilakukan dengan tujuan untuk memastikan bahwa data kategori dapat dimanfaatkan dalam algoritma *Machine Learning* yang memerlukan data numerik. Langkah pertama adalah **Label Encoding**, yang diterapkan pada kolom “jenis_tanaman” untuk mengubah kategori seperti 'alpukat', 'mangga', atau 'pisang' menjadi nilai numerik unik yang merepresentasikan setiap kategori. Selanjutnya, dilakukan **One-Hot Encoding** pada kolom “bps_nama_kabupaten_kota” untuk mengubah nama kabupaten/kota menjadi representasi biner. Hal ini menghasilkan kolom tambahan yang mencerminkan keberadaan atau ketiadaan kategori tertentu dalam format 0 atau 1. Kedua metode ini memastikan bahwa data kategori terintegrasi secara optimal dalam model, tanpa mengintroduksi hubungan numerik yang tidak seharusnya antara kategori.

3.5 Split Dataset

Dataset dalam penelitian ini terdiri dari 3.542 entri dengan 8 atribut utama, yaitu kode provinsi, kode kabupaten/kota, tahun produksi, jenis tanaman, jumlah produksi, serta variabel hasil encoding dari *Preprocessing*. *Dataset* ini dibagi menjadi dua bagian untuk memastikan performa model dapat dievaluasi secara objektif. Sebanyak 80% data (2.833 baris) digunakan sebagai data latih (training) untuk membangun dan melatih model *Machine Learning*. Sementara itu, 20% data lainnya (709 baris) digunakan sebagai data uji (testing) untuk mengevaluasi kinerja model dalam memprediksi hasil produksi.

Pembagian ini dilakukan dengan menggunakan fungsi “train_test_split” dari pustaka scikit-learn. Data latih berfungsi untuk mengenalkan pola-pola pada model, sedangkan data uji digunakan untuk menilai kemampuan generalisasi model terhadap data baru. Dengan proporsi pembagian yang umum digunakan ini, penelitian berupaya meminimalkan overfitting dan memastikan hasil prediksi model lebih andal dan akurat dalam konteks data produksi buah di Provinsi Aceh. dibagi menjadi dua bagian:

```

a) Data Training : untuk melatih model
b) Data Testing  : untuk menguji performa model
=====
Ukuran X_train : (1504, 26)
Ukuran X_test  : (377, 26)
Ukuran y_train : (1504,)
Ukuran y_test  : (377,)

```

Gambar 3.2 Hasil Data Training & Data Testing

Proporsi data latih dan uji adalah 80:20. Fitur yang digunakan meliputi kode provinsi, kode kabupaten/kota, tahun, dan hasil encoding variabel kategori.

3.6 Implementasi Model *Machine Learning*

Model yang digunakan dalam penelitian ini adalah *Random Forest* dan *Support Vector Machine (SVM)*. Proses implementasi model dilakukan secara bertahap untuk memastikan setiap algoritma dapat menghasilkan prediksi yang optimal. Berikut adalah penjelasan rinci dari setiap langkah implementasi:

1. **Random Forest**

Model ini menggunakan pendekatan ensemble learning yang menggabungkan banyak pohon keputusan untuk membuat prediksi yang lebih akurat. Model dilatih menggunakan data latih yang terdiri dari 2.833 baris, dengan fitur seperti kode provinsi, kode kabupaten/kota, tahun, jenis tanaman, dan jumlah produksi. Setelah pelatihan, model diuji menggunakan data uji sebanyak 709 baris.

Evaluasi model dilakukan dengan menggunakan metrik seperti Mean Squared Error (MSE), Mean Absolute Error (MAE), dan R^2 Score. Hasil evaluasi menunjukkan bahwa *Random Forest* memberikan MSE sebesar 5.234, MAE sebesar 1.856, dan R^2 Score sebesar 0.92, menandakan kinerja yang sangat baik dalam menangkap pola data.

2. **Support Vector Machine (SVM)**

Algoritma SVM digunakan dengan *Kernel RBF* untuk menangani data yang tidak terpisah secara linear. Proses pelatihan dilakukan pada data latih, di mana algoritma mencari hyperplane optimal yang memisahkan kelas data. Pengaturan

parameter *Kernel* dilakukan secara hati-hati untuk mengoptimalkan performa model.

Setelah pelatihan, model diuji dengan data uji, dan hasil evaluasi menunjukkan bahwa SVM menghasilkan MSE sebesar 6.758, MAE sebesar 2.103, dan R^2 Score sebesar 0.88. Meskipun kinerjanya sedikit di bawah *Random Forest*, algoritma ini tetap memberikan hasil yang cukup andal.

3.7 Evaluasi Model

Evaluasi kinerja model dilakukan dengan menggunakan tiga metrik utama yang umum digunakan dalam penelitian *Machine Learning* yaitu *Mean Squared Error (MSE)*, *Mean Absolute Error (MAE)*, dan R^2 Score.

- a. **Mean Squared Error (MSE)** Mengukur rata-rata kesalahan kuadrat dari prediksi model terhadap nilai aktual. MSE yang lebih kecil menunjukkan model yang lebih baik. Pada penelitian ini, *Random Forest* menghasilkan MSE sebesar 5.234, sedangkan SVM memiliki MSE sebesar 6.758.
- b. **Mean Absolute Error (MAE)** Mengukur rata-rata kesalahan absolut antara nilai prediksi dan aktual, memberikan indikasi langsung tentang selisih rata-rata. Hasil evaluasi menunjukkan bahwa *Random Forest* memiliki MAE sebesar 1.856, sementara SVM mencapai MAE sebesar 2.103.
- c. **R^2 Score** Mengukur seberapa baik model dapat menjelaskan variabilitas data aktual. Nilai R^2 berkisar antara 0 hingga 1, di mana 1 menandakan model sempurna. *Random Forest* mencapai R^2 sebesar 0.92, lebih tinggi dibandingkan SVM yang memiliki R^2 sebesar 0.88.

Evaluasi ini menunjukkan bahwa *Random Forest* lebih unggul dibandingkan SVM dalam hal prediksi data produksi buah di Provinsi Aceh, sebagaimana ditunjukkan oleh nilai Mean Squared Error (MSE) sebesar 5.234 dan Mean Absolute Error (MAE) sebesar 1.856, dengan R^2 Score mencapai 0.92. Sebaliknya, SVM menghasilkan MSE sebesar 6.758, MAE sebesar 2.103, dan R^2 Score sebesar 0.88, yang meskipun sedikit lebih rendah, tetap memberikan hasil prediksi yang andal. Analisis ini memberikan wawasan penting tentang kemampuan masing-masing algoritma dalam menangani *Dataset* yang kompleks, dengan *Random Forest* menunjukkan performa yang lebih

baik dalam menangkap pola data historis. Evaluasi kinerja model dilakukan berdasarkan beberapa metrik utama:

- a. **Mean Squared Error (MSE)** Mengukur rata-rata kesalahan kuadrat prediksi.
- b. **Mean Absolute Error (MAE)** Mengukur rata-rata kesalahan absolut prediksi.
- c. **R² Score** Mengukur seberapa baik model memprediksi data aktual.

3.8 Visualisasi Hasil

Hasil analisis model divisualisasikan untuk mempermudah interpretasi dan memberikan wawasan lebih mendalam terkait kinerja model. Proses visualisasi mencakup beberapa langkah penting:

- a. **Cross Validation** Data dibagi ke dalam beberapa lipatan (folds) untuk memastikan hasil konsisten di berbagai subset data. Pada model *Random Forest*, nilai rata-rata R² dari validasi silang adalah 0.91, menunjukkan kemampuan model yang andal dalam generalisasi data.
- b. **Feature Importance** Menampilkan atribut-atribut yang paling memengaruhi prediksi model. Hasil analisis menunjukkan bahwa jenis tanaman memiliki pengaruh terbesar terhadap hasil prediksi, diikuti oleh lokasi geografis seperti kabupaten/kota.
- c. **Distribusi Residual** Memvisualisasikan perbedaan antara nilai prediksi dan aktual. Grafik distribusi residual untuk *Random Forest* menunjukkan distribusi yang lebih terpusat di sekitar nol dibandingkan dengan SVM, mengindikasikan prediksi yang lebih akurat dan bias yang lebih rendah.
- d. **Residual Plot** Grafik residual menunjukkan penyebaran kesalahan prediksi pada setiap nilai prediksi. Residual dari *Random Forest* cenderung lebih stabil dibandingkan dengan SVM, terutama pada data dengan nilai produksi yang ekstrem.
- e. **Visualisasi Performa Model** Dibuat grafik perbandingan antara nilai prediksi dan nilai aktual untuk kedua model, yang memperlihatkan bahwa *Random Forest* lebih mendekati garis diagonal sempurna, mencerminkan prediksi yang lebih akurat.

Dengan langkah-langkah ini, visualisasi memberikan gambaran yang jelas tentang performa model serta atribut yang paling berpengaruh, sehingga mendukung

pengambilan keputusan yang berbasis data. Hasil analisis model divisualisasikan untuk mempermudah interpretasi:

- a. **Cross Validation** Membagi data ke dalam beberapa lipatan (folds) untuk memastikan hasil konsisten.
- b. **Feature Importance** Menampilkan atribut yang paling memengaruhi prediksi model.
- c. **Distribusi Residual** Memvisualisasikan perbedaan antara nilai prediksi dan aktual.

3.9 Perbandingan Algoritma *Random Forest* dan SVM

Evaluasi model menunjukkan bahwa *Random Forest* memiliki performa lebih baik dibandingkan SVM dalam memprediksi produksi buah di Provinsi Aceh. *Random Forest* mencatat MSE sebesar 1.862.499,60, MAE sebesar 804,43, dan R^2 Score sebesar 0,4799, menunjukkan kesalahan prediksi yang lebih kecil dan kemampuan lebih baik dalam menangkap pola data. Sebaliknya, SVM memiliki MSE sebesar 4.322.770,64, MAE sebesar 1.202,17, dan R^2 Score sebesar -0,2072, yang menandakan bahwa model ini tidak dapat menangkap pola data dengan baik dan memiliki tingkat kesalahan lebih tinggi. Dengan demikian, *Random Forest* lebih direkomendasikan untuk analisis prediksi produksi buah karena memberikan hasil yang lebih akurat dan stabil dibandingkan SVM.

