BAB III
METODE PENELITIAN

3.1 Pendekatan Penelitian

Penelitian ini menggunakan pendekatan analisis data berbasis *machine learning*, yaitu metode yang memanfaatkan algoritma pembelajaran mesin untuk mengenali pola dalam data dan membangun model prediktif secara otomatis. Pendekatan ini bersifat data-driven, di mana sistem belajar langsung dari data historis tanpa aturan eksplisit yang ditentukan sebelumnya. Dengan kemampuan adaptif dan efisien dalam mengolah data besar dan kompleks, model yang dihasilkan dapat memprediksi data baru dengan tingkat akurasi tinggi. Oleh karena itu, *machine learning* menjadi pilihan tepat untuk membangun prediksi berbasis data historis (Sariisik & Öğütlü, 2025).

3.2 Sumber Data

Penelitian ini menggunakan data sekunder, yaitu data yang dikumpulkan dan dipublikasikan oleh pihak lain, bukan diperoleh langsung melalui observasi atau wawancara. Sumber data berasal dari platform Satu Data Jakarta, yang menyediakan informasi bulanan mengenai produksi ikan berdasarkan Tempat Pelelangan Ikan (TPI) di DKI Jakarta. Rentang waktu data mencakup Januari 2022 hingga Desember 2023, dengan total sebanyak 120 baris data.

Data terdiri dari lima variabel berikut:

1. periode_data: tahun pengumpulan data (2022 dan 2023)
2. bulan: nomor bulan pengumpulan data (1 = Januari, 2 = Februari, ..., 12 = Desember)
3. tempat_pelelangan_ikan: nama Tempat Pelelangan Ikan (TPI) di DKI Jakarta
4. volume_produksi: volume produksi ikan (dalam kilogram)
5. nilai_produksi: nilai produksi ikan (dalam rupiah)

Data bersifat kuantitatif time series, yang menggambarkan perubahan volume dan nilai produksi ikan dari waktu ke waktu. Data diperoleh melalui pengunduhan file dari platform Satu Data Jakarta dan digunakan sebagai dasar dalam perancangan dan pengembangan model prediksi menggunakan algoritma regresi KNN.

3.3 Peralatan

3.3.1 Kebutuhan Perangkat Keras

Dalam penelitian ini, digunakan perangkat keras dengan spesifikasi berikut untuk mendukung kelancaran proses pengolahan data dan pengembangan model prediksi secara optimal:

1. Laptop dengan prosesor Intel(R) Core(TM) i5-7200U
2. Kecepatan CPU 2.50 GHz
3. RAM sebesar 8 GB (8192 MB)

3.3.2 Kebutuhan Perangkat Lunak

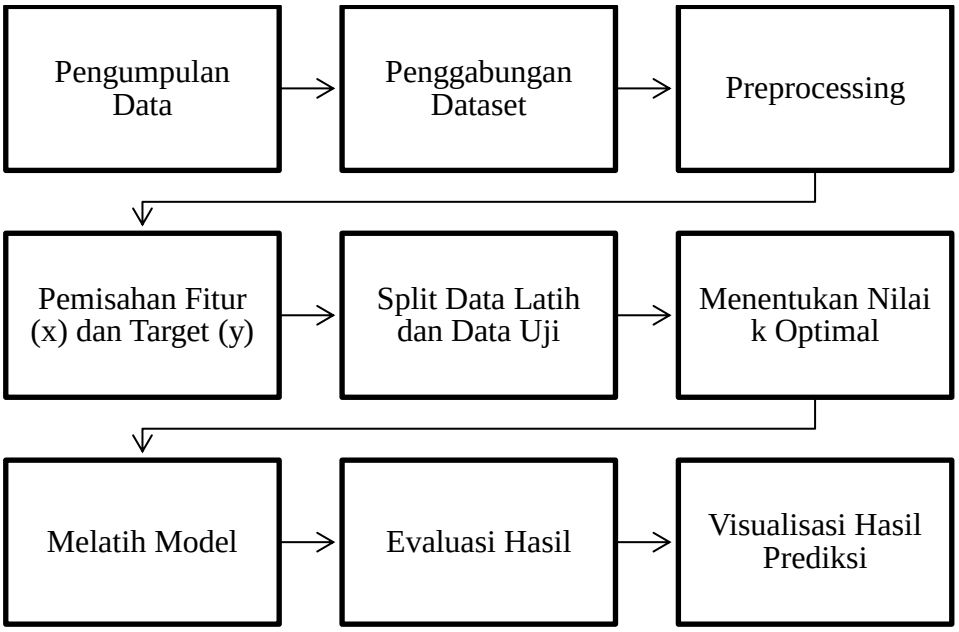
Beberapa perangkat lunak yang digunakan dalam proses pengolahan data dan pembangunan model prediksi adalah:

1. Sistem operasi: Windows 10 Home 64-bit
2. Bahasa pemrograman: Python
3. Platform pengembangan: Google Colaboratory (Colab)

3.4 Analisis Model Prediksi

Model prediksi yang dikembangkan dalam penelitian ini bertujuan untuk memprediksi volume dan nilai produksi ikan di Tempat Pelelangan Ikan (TPI) DKI Jakarta menggunakan algoritma regresi K-Nearest Neighbor (KNN). Model ini memanfaatkan data sekunder berupa data bulanan dari Januari 2022 hingga Desember 2023 yang diperoleh dari platform Satu Data Jakarta. Proses evaluasi performa model dilakukan menggunakan tiga metrik, yaitu Root Mean Squared Error (RMSE), Mean Absolute Error (MAE), dan R-squared (R^2). Model ini diharapkan dapat membantu pihak terkait dalam pengelolaan sumber daya perikanan dengan menyediakan estimasi prediktif yang akurat dan dapat diandalkan.

Gambar berikut menyajikan alur kerja penelitian dalam membangun model prediksi volume dan nilai produksi ikan menggunakan algoritma regresi K-Nearest Neighbor (KNN).



Gambar 3. 1 Alur Penelitian

Seluruh proses dilakukan secara terstruktur dan sistematis untuk memastikan akurasi, validitas, serta keberhasilan dalam membangun model prediksi volume dan nilai produksi perikanan di TPI DKI Jakarta. Gambar 3.1 menyajikan alur utama penelitian, yang dimulai dari tahap pengumpulan data hingga pengujian dan evaluasi model prediksi. Setiap tahapan dalam membangun model prediksi ini dijabarkan sebagai berikut:

3.4.1 Pengumpulan Data

Data yang digunakan dalam penelitian ini merupakan data sekunder yang diperoleh dari sumber resmi, yaitu Satu Data Jakarta. Dataset dengan total 120 baris data yang mencakup informasi periode pengumpulan data, bulan ke berapa, nama TPI tempat data dikumpulkan, volume_produksi (dalam kilogram), dan nilai_produksi (dalam rupiah).

3.4.2 Penggabungan Dataset

Dataset dari tahun 2022 dan 2023 digabungkan menjadi satu dataframe

secara vertikal (baris demi baris). Hasil penggabungan disimpan dalam variabel utama yang kemudian digunakan dalam proses analisis dan pemodelan.

3.4.3 Preprocessing

1. **Pengecekan Nilai Kosong**

   Dilakukan pemeriksaan terhadap nilai kosong (*missing values*) untuk memastikan kelengkapan data. Jika ditemukan, maka dilakukan penanganan yang sesuai, seperti penghapusan atau imputasi.

2. **Penyalinan Data Asli**

   Dataset awal disalin ke dalam variabel baru untuk menjaga keutuhan data asli. Langkah ini bertujuan agar proses validasi ulang atau pengolahan ulang dapat dilakukan tanpa pelu memuat ulang data.

3. **Eksplorasi Awal Distribusi Data**

   Eksplorasi awal terhadap distribusi volume_produksi dan nilai_produksi menunjukkan pola sebaran yang tidak normal (*right-skewed*) dengan rentang nilai yang lebar. Kondisi ini berpotensi menimbulkan bias dalam proses pemodelan, sehingga diperlukan transformasi lebih lanjut untuk menstabilkan distribusi data.

4. **Penanganan Outlier**

   Untuk meminimalkan pengaruh nilai ekstrem yang dapat mengganggu proses pelatihan model, dilakukan penanganan *outlier* pada variabel target volume_produksi dan nilai_produksi menggunakan metode winsorization sebesar 1% di kedua ujung distribusi.

5. **Transformasi Logaritmik**

   Transformasi logaritmik menggunakan np.log1p() diterapkan pada variabel target volume_produksi dan nilai_produksi untuk mengurangi *skewness* dan mendekatkan distribusi terhadap normalitas. Transformasi ini bertujuan untuk meningkatkan kestabilan dan akurasi model regresi.

6. **Encoding Fitur Kategorikal**

   Fitur kategorikal tempat_pelelangan_ikan dikonversi ke dalam format numerik menggunakan metode One-Hot Encoding melalui ColumnTransformer. Proses

ini memastikan hanya fitur kategorikal yang ditransformasi, sementara fitur numerik lainnya tetap utuh. Hasil akhirnya berupa representasi numerik yang siap diproses dalam algoritma *machine learning*.

7. **Normalisasi Fitur Numerik**

   Fitur numerik seperti periode_data dan bulan dinormalisasi menggunakan Min-Max Scaler agar berada dalam rentang [0,1]. Normalisasi ini bertujuan untuk mencegah dominasi fitur berskala besar serta menjaga konsistensi perhitungan jarak pada algortima KNN yang sensitif terhadap perbedaan skala.

3.4.4 Pemisahan Fitur (x) dan Target (y)

Setelah proses preprocessing, dilakukan pemisahan antara fitur (x) dan target (y). Fitur mencakup seluruh variabel *input*, sedangkan target terdiri dari dua variabel utama, yaitu volume_produksi dan nilai_produksi.

3.4.5 Split Data Latih dan Data Uji

Dataset dibagi menjadi data latih (80%) dan data uji (20%) menggunakan fungsi train_test_split, dengan parameter random_state=42 untuk menjaga konsistensi hasil pembagian data, sehingga proses pembagian antara data latih dan data uji menghasilkan hasil yang sama setiap kali kode dijalankan.

3.4.6 Menentukan Nilai k Optimal

Untuk mendapatkan nilai k optimal, dilakukan eksplorasi dengan mencoba berbagai nilai k dalam rentang tertentu. Evaluasi performa model dilakukan pada data uji menggunakan tiga metrik, yaitu metrik RMSE, MAE, R^2. Pemilihan nilai k optimal didasarkan pada hasil terbaik terhadap kedua target.

3.4.7 Melatih Model

Pemodelan dilakukan dengan pendekatan *multi-target regression* menggunakan MultiOutputRegressor dengan KNeighborsRegressor sebagai model dasar. Model dilatih menggunakan data latih, kemudian dievaluasi pada data uji untuk mengukur performa dan mendeteksi kemungkinan *overfitting*.

3.4.8 Prediksi dan Evaluasi Model pada Data Uji

Model yang telah dilatih kemudian dievaluasi menggunakan data uji. Proses evaluasi dimulai dengan melakukan prediksi terhadap data uji, kemudian hasil prediksi dianalisis menggunakan tiga metrik evaluasi, yaitu RMSE, MAE, dan R^2. Evaluasi dilakukan dalam dua skala, yaitu skala logaritmik (hasil langsung dari model) dan skala aktual yang diperoleh melalui transformasi balik menggunakan np.expm1(). Proses evaluasi dilakukan secara terpisah untuk kedua variabel target, yaitu volume_produksi dan nilai_produksi, guna memperoleh gambaran akurasi model secara lebih terhadap masing-masing target.

3.4.9 Visualisasi Hasil Prediksi

Visualisasi dilakukan untuk memberikan gambaran mengenai akurasi dan pola prediksi model. Dua jenis grafik yang digunakan dalam analisis hasil prediksi, yaitu:

1. Line Chart

   Grafik ini menampilkan garis perbandingan antara nilai prediksi dan nilai aktual berdasarkan urutan indeks data uji. Visualisasi ini membantu mengamati kemampuan model dalam mengikuti pola atau tren perubahan nilai target secara keseluruhan.

2. Scatter Plot

   Grafik ini menampilkan sebaran titik antara nilai prediksi dan nilai aktual, dengan garis diagonal sebagai garis referensi ideal (y = x). Grafik ini berguna untuk menilai akurasi prediksi pada setiap data secara individual. Semakin dekat titik terhadap garis, maka semakin tinggi akurasi prediksinya.

Untuk memberikan gambaran mengenai struktur dan format data yang digunakan dalam penelitian ini, Tabel 3.1 menyajikan potongan dari *dataset* produksi perikanan tahun 2022. Kolom tempat_pelelangan_ikan tidak ditampilkan secara eksplisit karena keterbatasan ruang, namun tetap menjadi salah satu variabel utama bersama periode_data, bulan, tempat_pelelangan_ikan, volume_produksi, dan nilai_produksi ikan pada Tempat Pelelangan Ikan (TPI) DKI Jakarta.

Tabel 3. 1 Potongan Dataset Produksi Perikanan Tahun 2022

Periode_data	Bulan	.....	Volume_produksi	Nilai_produksi
2022	1	.....	2386846	94097726700
2022	1	.....	211212	7657302500
2022	1	.....	923485	25462085000
2022	1	.....	272549	10607960000
2022	2	.....	1676641	67752200100
2022	2	.....	173273	5913782500
2022	2	.....	980670	13851485000
2022	2	.....	394082	28056810000
2022	3	.....	1997642	96685395700
2022	3	.....	268097	8460285000
2022	3	.....	895373	23581230000
2022	3	.....	430420	16354815000
2022	4	.....	2938120	137701026400
2022	4	.....	318417	10458452500
2022	4	.....	790862	21275537500
2022	4	.....	451135	17064710000
2022	5	.....	2926818	132472671100
2022	5	.....	269761	9053145000