Pemodelan Tekanan Hidrostatik Berbasis Machine Learning CatBoost dengan Data Oseanografi Multivariabel

Azril Fahmiardi[1], Ariel James[2]

[1]Departemen Teknik Elektro dan Informatika, Universitas Gadjah Mada
[2]Departemen Ilmu Komputer, Universitas Negeri Semarang

Abstract

Tekanan hidrostatik merupakan parameter fundamental dalam oseanografi yang memengaruhi berbagai proses fisik dan biologis di lingkungan laut. Penelitian ini menyajikan pendekatan berbasis machine learning untuk memprediksi tekanan hidrostatik dengan memanfaatkan data oseanografi multivariabel. Model CatBoostRegressor digunakan sebagai algoritma utama karena kemampuannya dalam menangani data kategorikal, missing values, dan relasi nonlinier. Metodologi yang dikembangkan mencakup preprocessing data komprehensif, termasuk penanganan missing values berbasis KNN, normalisasi multistage, dan transformasi data untuk mengurangi skewness. Feature engineering dilakukan dengan mempertimbangkan aspek fisis oseanografi, meliputi ekstraksi fitur temporal dengan encoding siklus Fourier, rekonstruksi fitur cahaya, dan pembentukan fitur interaksi berdasarkan prinsip hidrodinamika. Feature selection diterapkan secara sistematis menggunakan pendekatan dua tahap untuk memilih subset fitur paling informatif. Model akhir mencapai performa tinggi dengan $R^2 = 0.9568$, RMSE = 0.0004, dan MAE = 0.0002 pada validasi silang 5-fold. Learning curve menunjukkan stabilitas model dengan gap yang mengecil antara skor training dan validasi seiring penambahan data pelatihan. Hasil ini mendemonstrasikan potensi pendekatan machine learning dalam prediksi parameter oseanografi dan aplikasinya untuk monitoring lingkungan laut.

Kata Kunci: *tekanan hidrostatik, machine learning, CatBoost, oseanografi, feature engineering*

Pendahuluan

Lautan merupakan komponen vital dalam sistem iklim global yang mencakup lebih dari 70% permukaan Bumi dan berperan penting dalam mengatur pola cuaca, iklim, serta kehidupan di planet ini [1]. Pemahaman mendalam tentang dinamika laut menjadi krusial dalam konteks perubahan iklim global dan dampaknya terhadap ekosistem laut [2]. Salah satu parameter fisik yang fundamental dalam oseanografi adalah tekanan hidrostatis, yang tidak hanya mencerminkan kondisi fisik kolom air tetapi juga memengaruhi berbagai proses biogeokimia dan ekologi di lingkungan laut [3].

Tekanan hidrostatis di laut merupakan hasil dari kombinasi kompleks berbagai faktor fisik dan biologis, termasuk kedalaman, densitas air laut, temperatur, salinitas, serta distribusi massa air pada berbagai lapisan vertikal [4]. Para oseanografer secara rutin memantau tekanan hidrostatis untuk memahami dinamika arus laut, stratifikasi massa air, perubahan habitat bawah laut, serta untuk mendukung prediksi kondisi laut dalam berbagai aplikasi operasional [5]. Pengukuran dan prediksi tekanan hidrostatis yang akurat memiliki implikasi luas, mulai dari navigasi maritim, eksplorasi sumber daya laut, hingga pemahaman tentang respons ekosistem laut terhadap perubahan lingkungan.

Metode tradisional untuk memprediksi tekanan hidrostatis umumnya mengandalkan persamaan hidrostatik sederhana yang mengasumsikan densitas air laut konstan atau menggunakan model empiris berbasis parameter fisik tunggal. Namun, pendekatan ini memiliki keterbatasan signifikan dalam menangkap kompleksitas interaksi multivariabel yang terjadi di lingkungan laut nyata. Variabilitas spasial dan temporal yang tinggi dalam parameter oseanografi, seperti fluktuasi suhu, salinitas, kandungan oksigen, serta aktivitas biologis, menciptakan pola nonlinear yang sulit diprediksi menggunakan model konvensional [6].

Dalam dekade terakhir, penerapan teknik *machine learning* dalam ilmu kelautan telah mengalami perkembangan pesat dan menunjukkan potensi besar dalam mengatasi tantangan prediksi parameter oseanografi. Algoritma pembelajaran mesin, khususnya yang berbasis *ensemble methods* seperti Random Forest, Gradient Boosting, dan CatBoost, telah terbukti efektif dalam menangani dataset multivariabel dengan hubungan nonlinear kompleks [7]. CatBoostRegressor, sebagai salah satu implementasi *advanced gradient boosting*, memiliki keunggulan khusus dalam menangani data kategorikal, *missing values*, dan mampu menangkap pola interaksi yang rumit antar variabel prediktor [8].

Keberhasilan penerapan *machine learning* dalam

prediksi parameter oseanografi telah dilaporkan dalam berbagai studi. Zhang et al. [9] menunjukkan bahwa *ensemble methods* dapat meningkatkan akurasi prediksi suhu permukaan laut hingga 15% dibandingkan model statistik tradisional. Wang et al. [10] berhasil menggunakan *gradient boosting* untuk memprediksi konsentrasi klorofil-a dengan tingkat akurasi yang tinggi berdasarkan data satelit multispektral. Namun, aplikasi khusus untuk prediksi tekanan hidrostatis menggunakan *snapshot* profil data oseanografi yang komprehensif masih terbatas dan memerlukan pengembangan lebih lanjut.

Tantangan utama dalam prediksi tekanan hidrostatis menggunakan *machine learning* meliputi heterogenitas data oseanografi, keberadaan *missing values* yang signifikan, distribusi data yang tidak normal (*skewed*), serta perlunya mempertahankan interpretabilitas fisik dari model yang dibangun. Data oseanografi umumnya memiliki karakteristik kompleks dengan variabel yang berasal dari berbagai domain (fisik, kimia, biologis, dan temporal), masing-masing dengan skala dan distribusi yang berbeda. *Preprocessing* data yang tepat menjadi kunci keberhasilan dalam aplikasi *machine learning* untuk data oseanografi. Teknik seperti imputasi berbasis K-Nearest Neighbors (KNN), normalisasi *multistage*, dan transformasi untuk mengatasi *skewness* telah terbukti efektif dalam meningkatkan kualitas data input [11].

Feature engineering yang berbasis pengetahuan domain oseanografi, termasuk ekstraksi fitur temporal, rekonstruksi variabel cahaya, dan pembentukan fitur interaksi antar parameter fisik, dapat secara signifikan meningkatkan performa model prediktif [12]. Pendekatan ini memungkinkan model untuk menangkap pola kompleks yang mungkin tidak terdeteksi dari data mentah, sekaligus mempertahankan interpretabilitas fisik yang penting dalam konteks oseanografi.

Penelitian ini bertujuan mengembangkan model prediksi tekanan hidrostatis yang akurat dan robust menggunakan machine learning berbasis *CatBoostRegressor*. Dengan dataset komprehensif yang meliputi 52 variabel oseanografi dari domain fisik, kimia, biologis, dan temporal, dikembangkan pipeline preprocessing dan feature engineering sistematis untuk optimasi performa. Pendekatan meliputi penanganan missing values berdasarkan karakteristik kelompok fitur, normalisasi dan transformasi multistage, serta feature selection kombinasi importance-based dan recursive feature elimination.

Kontribusi utama penelitian ini adalah:

1. Pengembangan metodologi preprocessing komprehensif untuk data oseanografi multivariabel,
2. Implementasi feature engineering berbasis pengetahuan domain kelautan untuk meningkatkan relevansi prediktif,
3. Evaluasi sistematis performa model *CatBoostRegressor* dalam prediksi tekanan hidrostatis.

Metode Analisis

Sumber Data

Dataset yang digunakan dalam penelitian ini diperoleh dari kompetisi Regression Rumble — Neurontara Data Clash 2025 yang berfokus pada prediksi tekanan hidrostatik menggunakan variabel-variabel oseanografi. Data mentah bersumber dari arsip observasi dan reanalisis terbuka, termasuk dataset ERA5 yang dikembangkan oleh European Centre for Medium-Range Weather Forecasts (ECMWF) serta program-program pemantauan laut terbuka lainnya yang berlisensi Creative Commons. Seluruh variabel dalam dataset telah melalui proses kurasi dan penamaan ulang yang dilakukan oleh Edgar V (Kolombia) dan panitia Neurontara Data Clash untuk keperluan edukatif, dengan catatan bahwa tidak terdapat afiliasi resmi dengan penyedia data asli.

Data Cleaning

Pembersihan data merupakan tahapan krusial dalam proses analisis prediktif, khususnya pada pemodelan regresi. Tahapan ini bertujuan untuk memastikan kualitas dan konsistensi data agar tidak mengganggu hasil eksplorasi dan performa model. Beberapa langkah utama yang diterapkan dalam proses ini mencakup penanganan data temporal, standarisasi format numerik, serta penanganan nilai pencilan (outlier).

1. **Penanganan Data Temporal**
   Fitur temporal depth_reading_time pada awalnya terbaca sebagai tipe objek (string), yang menjadi kendala dalam eksplorasi pola waktu seperti distribusi pengukuran per jam atau tren musiman. Oleh karena itu, dilakukan konversi ke format datetime menggunakan pendekatan toleran kesalahan dengan errors='coerce', yang mengubah entri tidak valid menjadi NaT (Not a Time) untuk mencegah kegagalan pemrosesan.
2. **Standarisasi Format Numerik**
   Beberapa fitur numerik dalam dataset memiliki format desimal lokal yang menggunakan tanda koma (,) sebagai pemisah. Format ini menyebabkan nilai-nilai tersebut tidak dikenali sebagai numerik oleh pustaka analisis data seperti *pandas*, melainkan terbaca sebagai string. Hal ini menghambat analisis statistik maupun visualisasi. Untuk mengatasi permasalahan tersebut, dilakukan konversi sistematis ke format standar menggunakan fungsi pemetaan untuk mengganti koma menjadi titik. Selanjutnya, nilai-nilai dikonversi

ke tipe numerik menggunakan `pd.to_numeric()` dengan parameter `errors='coerce'`.

Feature	Sample Values
`oxygen_saturation_50m`	21,8; 22,6
`perceived_water_density`	28,2; 33,8
`sediment_deposition`	0,1; 1,4
`dissolved_gas_pressure`	0,21; 0,28
`current_turbulence`	7,9; 2,5
`sediment_porosity_0_to_10cm`	0,204; 0,293
`sediment_porosity_10_to_30cm`	0,162; 0,268
`sediment_porosity_30_to_100cm`	0,257; 0,286
`sediment_porosity_100_to_250cm`	0,321; 0,311
`perpendicular_light_intensity`	151,9; 586,8
`thermal_emissions`	596,9; 1386,7

Table 1: *Fitur Numerik dengan Format Desimal Koma*

3. **Penanganan Outlier**
   Nilai pencilan (outlier) dapat mendistorsi hasil eksplorasi data dan mempengaruhi ketepatan model regresi. Untuk mengatasi hal ini, diterapkan metode *capping* berbasis interkuartil (Interquartile Range, IQR). Teknik ini mempertahankan semua observasi tetapi membatasi nilai-nilai ekstrem agar tidak melebihi batas bawah dan atas yang ditentukan, yaitu:

$$\text{Batas Bawah} = Q1 - 1.5 \times \text{IQR} \quad (1)$$

$$\text{Batas Atas} = Q3 + 1.5 \times \text{IQR} \quad (2)$$

   Seluruh fitur numerik dalam data dikenali dan diproses dengan metode ini secara iteratif. Pendekatan *capping* dipilih alih-alih eliminasi data untuk menjaga volume data tetap utuh, namun tetap mengurangi potensi bias dari pengaruh nilai ekstrem.

Exploratory Data Analysis (EDA)

Dataset terdiri dari dua komponen utama: data train dengan 15.321 sampel dan 52 fitur termasuk target, serta data test dengan 6.567 sampel dan 52 fitur tanpa target. Analisis struktur data mengungkapkan adanya perbedaan signifikan antara kedua dataset tersebut. Fitur target `hydrostatic_pressure` hanya terdapat pada data train dikarenakan untuk pengujian, sementara fitur `total_light_exposure` hanya terdapat pada data test.

1. **Karakteristik Dataset**
   Dataset terdiri dari data *train* dengan 15.321 sampel dan 52 fitur termasuk target, serta data *test* dengan 6.567 sampel dan 52 fitur tanpa target. Analisis struktur data mengungkapkan adanya perbedaan signifikan antara kedua dataset tersebut. Fitur target `hydrostatic_pressure` hanya terdapat pada data *train* dikarenakan digunakan untuk pelatihan, sementara fitur `total_light_exposure` hanya terdapat pada data *test*.

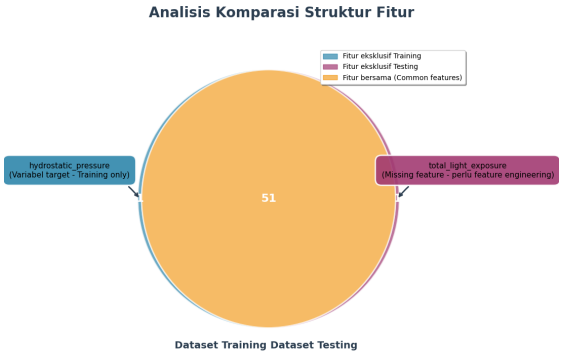

Figure 1: *Perbandingan struktur fitur pada data train dan test*

Perbedaan struktur ini menimbulkan tantangan khusus dalam pengembangan model, karena fitur `total_light_exposure` yang tersedia pada data *test* tidak dapat dimanfaatkan langsung saat melatih model. Hal ini mengindikasikan kebutuhan akan pendekatan rekonstruksi fitur pada tahap *feature engineering*.

2. **Analisis Missing Values**
   Identifikasi *missing values* menunjukkan pola yang bervariasi pada berbagai kelompok fitur. Visualisasi pada Gambar 2 menunjukkan distribusi *missing values* secara keseluruhan.

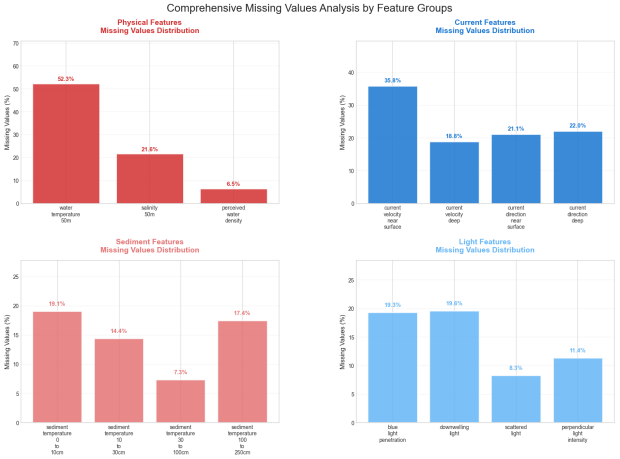

Figure 2: *Distribusi missing values berdasarkan kategori fitur*

Analisis kuantitatif berdasarkan kelompok fitur mengungkapkan bahwa fitur-fitur sedimen memiliki persentase missing values tertinggi (rata-rata 16,44%), diikuti oleh fitur arus (12,82%), fitur fisik (11,27%), dan fitur cahaya (8,65%). Perbedaan signifikan ini menunjukkan pendekatan imputasi harus mempertimbangkan karakteristik kelompok fitur. Oleh karena itu, pembahasan selanjutnya

difokuskan pada tiap kelompok fitur.

Kelompok Fitur	Rata-rata Missing Values (%)
Sediment Features	16.44
Current Features	12.82
Physical Features	11.27
Light Features	8.65

Table 2: *Persentase Missing Values per Kelompok Fitur*

3. **Analisis Distribusi**
   Beberapa fitur prediktor menunjukkan pola distribusi yang tidak simetris , yang dapat memengaruhi stabilitas model prediktif jika tidak ditangani dengan tepat. Tabel berikut menampilkan 10 fitur dengan skewness tertinggi:

Fitur	Skewness
downwelling_light	1.27
mesoplankton_density	1.18
bioluminescence_intensity	1.17
microplankton_density	1.13
scattered_light	1.12
perpendicular_light_intensity	1.10
dissolved_gas_pressure	1.08
blue_light_penetration	1.07
sediment_deposition	0.91
salinity_50m	0.90

Table 3: *Daftar fitur dengan skewness tertinggi*

4. **Analisis Temporal**
   Fitur temporal `depth_reading_time` dianalisis dengan mengekstrak komponen waktu seperti jam, hari, dan bulan. Pola siklik tekanan hidrostatis berdasarkan komponen waktu menunjukkan keteraturan konsisten, terutama pada siklus harian. Hasil ini menunjukkan bahwa waktu pengambilan data berperan penting dalam karakteristik tekanan hidrostatis.

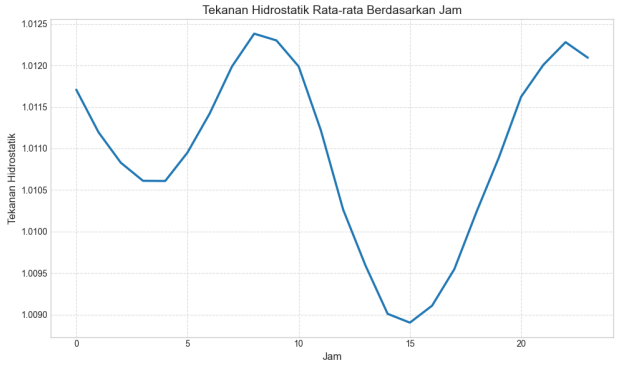

Figure 3: *Tekanan Hidrostatis Rata-rata berdasarkan Jam*

Gambar 3 memperlihatkan fluktuasi tekanan hidrostatis dalam siklus 24 jam, dengan pola jelas antara siang dan malam. Pola siklik ini menjadi dasar pengembangan fitur trigonometri berbasis waktu pada tahap feature engineering.

Preprocessing

Setelah melakukan eksplorasi data analisis, tahap selanjutnya adalah preprocessing data untuk mempersiapkan dataset yang optimal bagi proses pemodelan. Tahap preprocessing meliputi serangkaian transformasi data untuk mengatasi permasalahan yang teridentifikasi pada tahap EDA.

1. **Penanganan Missing Values**
   Penanganan *missing values* dilakukan dengan pendekatan yang disesuaikan berdasarkan karakteristik kelompok fitur, mengingat pola ketidaklengkapan data yang bervariasi antar kelompok fitur seperti yang telah diidentifikasi pada tahap EDA. Untuk imputasi data, kami mengimplementasikan metode KNN (*K-Nearest Neighbors*) dengan $k = 5$ pada setiap kelompok fitur. Pendekatan ini memungkinkan pemanfaatan korelasi antar fitur dalam kelompok yang sama untuk menghasilkan estimasi yang lebih akurat bagi nilai-nilai yang hilang. Secara khusus, algoritma KNN memanfaatkan kemiripan antar sampel berdasarkan fitur-fitur yang tersedia untuk mengisi nilai yang hilang, sehingga mempertahankan pola dan hubungan antar variabel yang relevan secara oseanografi.
   Untuk kasus di mana imputasi KNN tidak dapat diterapkan (misalnya jika hanya terdapat satu fitur dalam kelompok), kami menggunakan imputasi dengan nilai *median* untuk fitur numerik dan nilai *modus* untuk fitur kategorikal.
2. **Normalisasi dan Transformasi Data**
   Berdasarkan analisis komprehensif terhadap distribusi data dan karakteristik oseanografi, kami mengidentifikasi enam fitur utama yang memerlukan normalisasi khusus, yaitu: `water_temperature_50m`, `salinity_50m`, `oxygen_saturation_50m`, `seafloor_pressure`, `plankton_density`, dan `perceived_water_density`. Pemilihan fitur-fitur ini tidak semata-mata didasarkan pada nilai *skewness*, melainkan merupakan hasil integrasi antara temuan statistik dan pertimbangan ilmiah. Meskipun `water_temperature_50m` memiliki nilai skewness 0,73 dan tidak termasuk dalam sepuluh fitur dengan skewness tertinggi, variabel ini tetap dimasukkan karena merupakan komponen utama dalam persamaan keadaan air laut yang mengaitkan suhu (T), salinitas (S), dan tekanan (p) dalam membentuk densitas (ρ) melalui fungsi:

$$\rho = \rho(T, S, p) \tag{3}$$

   Ketiga parameter tersebut membentuk satu kesatuan triadik yang berperan sentral dalam dinamika

oseanografi, sehingga diperlukan normalisasi yang konsisten untuk menjaga integritas hubungan non-linear antar variabel tersebut.
Dari sisi teknis, perbedaan rentang nilai antar fitur yang signifikan—seperti suhu (4,2–18,8°C) dibandingkan densitas (1022–1029 kg/m³)—dapat menimbulkan ketidakseimbangan kontribusi dalam proses pembelajaran model. Untuk itu, dilakukan dua tahap transformasi:

(a) **Min-Max Scaling**: Menyeragamkan rentang fitur dalam `COLUMNS_TO_NORMALIZE` ke skala 0–1.
(b) **Transformasi Yeo-Johnson**: Mengurangi skewness fitur yang sangat menceng (`skewness` > 1.0).

Pendekatan ini terbukti efektif berdasarkan uji eksperimental karena mampu mengurangi bias distribusi sekaligus mempertahankan relasi fisis yang relevan. Selain aspek statistik, normalisasi ini juga mempertimbangkan keterkaitan langsung fitur terhadap tekanan hidrostatik sebagai target model.
Fitur seperti `seafloor_pressure` dan `perceived_water_density` berkontribusi langsung dalam persamaan tekanan hidrostatik:

$$P = \rho \times g \times h \tag{4}$$

Sementara itu, `oxygen_saturation_50m` dan `plankton_density` mencerminkan faktor biologis yang memengaruhi stratifikasi kolom air dan secara tidak langsung berdampak terhadap fluktuasi tekanan.

Feature Engineering dan Feature Selection

Feature engineering dilakukan untuk mengekstraksi informasi yang lebih bermakna dari data mentah dan domain oceanografi. Proses ini melibatkan transformasi dan penciptaan fitur baru untuk meningkatkan kinerja model prediktif. Selanjutnya, feature selection diterapkan untuk mengidentifikasi subset fitur yang paling informatif, mengurangi dimensionalitas, dan menghindari overfitting.

Feature Engineering Feature engineering dilakukan melalui beberapa tahapan sistematis berikut:

1. **Ekstraksi Fitur Temporal**
   Waktu pengukuran (depth_reading_time) ditransformasikan menjadi fitur-fitur temporal eksplisit, seperti year, month, day, hour, dan day_of_year, yang memungkinkan model untuk menangkap tren musiman dan harian. Untuk mengakomodasi sifat periodik waktu, dilakukan encoding menggunakan fungsi Fourier berbasis sinus dan kosinus [13], seperti:

$$\text{month\_sin} = \sin\left(\frac{2\pi \cdot \text{month}}{12}\right) \tag{5}$$

$$\text{hour\_cos} = \cos\left(\frac{2\pi \cdot \text{hour}}{24}\right) \tag{6}$$

   Selain itu, pembagian waktu menjadi beberapa periode siang–malam (day_period) digunakan untuk menangkap perbedaan aktivitas ekologis berdasarkan waktu.

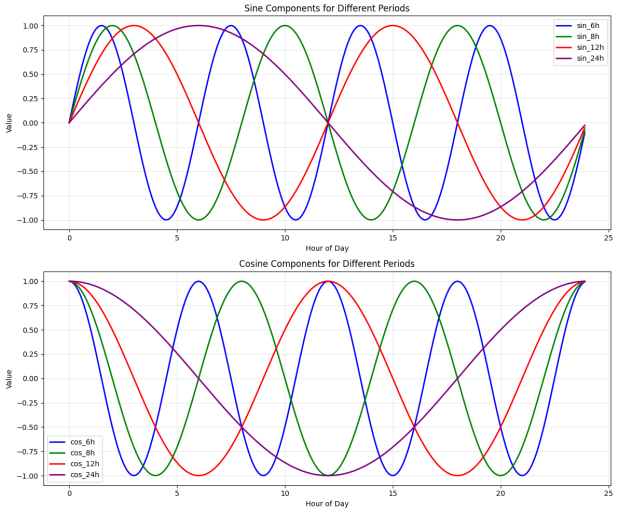

Figure 4: *Komponen Fourier untuk berbagai periode: (a) Komponen sinus untuk periode 6, 8, 12, dan 24 jam; (b) Komponen kosinus untuk periode yang sama. Encoding dengan fungsi trigonometri ini memungkinkan model menangkap pola siklik dari data temporal.*

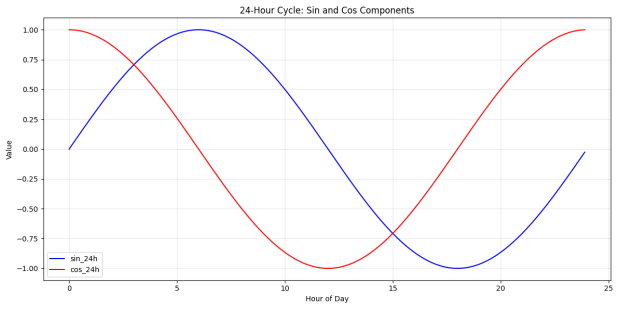

Figure 5: *Visualisasi komponen sinus dan kosinus untuk siklus 24 jam. Kombinasi kedua komponen ini memungkinkan model untuk mengenali posisi tepat dalam siklus harian tanpa diskontinuitas di batas periode.*

2. **Rekonstruksi dan Ekspansi Fitur Cahaya**
   Karena variabel total_light_exposure tidak tersedia secara langsung, fitur ini direkonstruksi menggunakan kombinasi fisik dari spektrum cahaya [14]:

$$\begin{aligned}\text{total\_light} = 0.35 \cdot \text{blue} + 0.40 \cdot \text{downwelling} \\ + 0.15 \cdot \text{scattered} \\ + 0.10 \cdot \text{perpendicular}\end{aligned} \tag{7}$$

Nilai tersebut kemudian dikoreksi terhadap faktor kekeruhan dengan eksponensial negatif dari turbidity, serta dimodifikasi berdasarkan jam menggunakan fungsi parabola simetris terhadap pukul 12:

$$f(h) = -\frac{(h-12)^2}{40} + 1 \tag{8}$$

3. **Fitur Domain Oceanografi Lanjutan**
   Dibuat berbagai rasio, interaksi, dan transformasi non-linear berdasarkan prinsip fisik dan biogeokimia [15, 16].

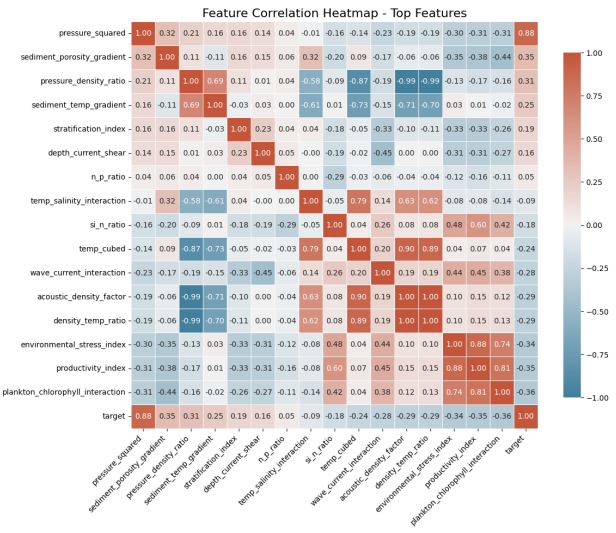

Figure 6: *Peta korelasi (heatmap) antar fitur domain oceanografi dan variabel target. Visualisasi ini menunjukkan hubungan antar fitur utama dengan kode warna, di mana warna merah menunjukkan korelasi positif kuat dan warna biru menunjukkan korelasi negatif kuat.*

Table 4: *Korelasi fitur domain oceanografi dengan variabel target*

Korelasi Positif	
Fitur	**Nilai**
pressure_squared	0.885
sediment_porosity_gradient	0.350
pressure_density_ratio	0.315
sediment_temp_gradient	0.252
stratification_index	0.186
depth_current_shear	0.158
n_p_ratio	0.050
temp_salinity_interaction	-0.088
Korelasi Negatif	
si_n_ratio	-0.180
temp_cubed	-0.239
wave_current_interaction	-0.281
acoustic_density_factor	-0.287
density_temp_ratio	-0.288
environmental_stress_index	-0.339
productivity_index	-0.351
plankton_chlorophyll_interaction	-0.356

Tabel 4 menunjukkan korelasi antara fitur-fitur domain oceanografi dengan variabel target. Fitur dengan korelasi positif tertinggi adalah pressure_squared (0.885), yang mengindikasikan hubungan kuadratik yang kuat antara tekanan dan variabel target. Sebaliknya, plankton_chlorophyll_interaction memiliki korelasi negatif terkuat (-0.356), menunjukkan bahwa peningkatan interaksi antara plankton dan klorofil cenderung menurunkan nilai target.

4. **Fitur Polinomial dan Interaksi Tingkat Tinggi**
   Digunakan PolynomialFeatures (derajat 2, interaksi saja) dari 20 fitur terpilih untuk menghasilkan fitur interaksi, misalnya temp_salinity_interaction dan temp_salinity_quadratic, yang menangkap relasi non-linier antar variabel penting.

Feature Selection Untuk menghindari overfitting dan mengurangi kompleksitas model, dilakukan dua tahap seleksi fitur secara berurutan. Proses ini dirancang agar efisien dan tetap mempertahankan informasi prediktif yang penting bagi target hydrostatic pressure.

1. **Seleksi Awal Berdasarkan Feature Importance (CatBoost)**
   Langkah pertama menggunakan model CatBoostRegressor untuk menghitung feature importance berdasarkan pohon boosting. Semua fitur awal (tanpa transformasi polinomial) dilatih, dan diurutkan berdasarkan bobot kontribusinya. Dua

puluh fitur teratas dipilih untuk proses transformasi nonlinier berikutnya. Hasil seleksi ini tidak hanya mengurangi noise fitur yang kurang relevan, tetapi juga memastikan bahwa fitur interaksi hanya dibentuk dari variabel-variabel dominan.

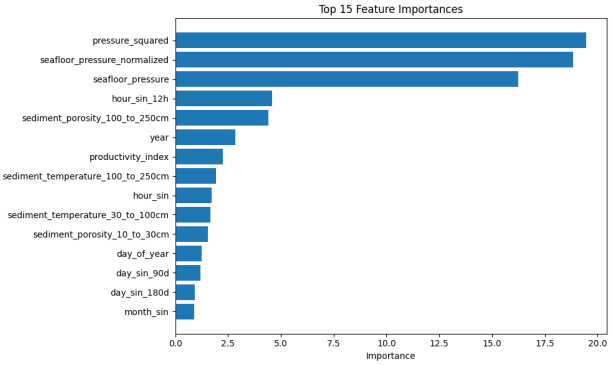

Figure 7: *Feature importance dari model CatBoost awal*

2. **Seleksi Lanjutan Setelah Transformasi Polinomial**
   Selanjutnya, fitur-fitur interaktif dibentuk dari 20 fitur penting tersebut menggunakan PolynomialFeatures (derajat 2, interaction_only). Dataset hasil transformasi kemudian diseleksi kembali menggunakan pendekatan SelectFromModel, di mana digunakan XGBRegressor hanya sebagai alat bantu estimator untuk mengukur kekuatan kontribusi tiap fitur baru.
   Seleksi akhir mempertahankan hanya fitur-fitur dengan nilai importance di atas threshold (≥ 0.001). Untuk dataset berdimensi lebih rendah (jumlah fitur < 50), digunakan pula Recursive Feature Elimination with Cross Validation (RFECV) sebagai validasi akhir pemilihan fitur.

Pendekatan dua tahap ini memastikan bahwa model akhir hanya dilatih pada fitur yang informatif dan tidak redundant, mengurangi risiko overfitting serta mempercepat inferensi.

Hasil dan Pembahasan

Kinerja Model

Evaluasi kinerja model dilakukan menggunakan pendekatan validasi silang (cross-validation) sebanyak 5 lipatan (5-fold CV) untuk mengukur generalisasi model terhadap data tak terlihat. Model utama yang digunakan adalah CatBoostRegressor, sebuah algoritma boosting berbasis pohon keputusan yang mampu menangani data kategorikal, missing value, serta relasi nonlinier secara efisien. Penggunaan boosting sendiri mengacu pada pendekatan additive model optimization yang dijelaskan oleh Friedman [17].

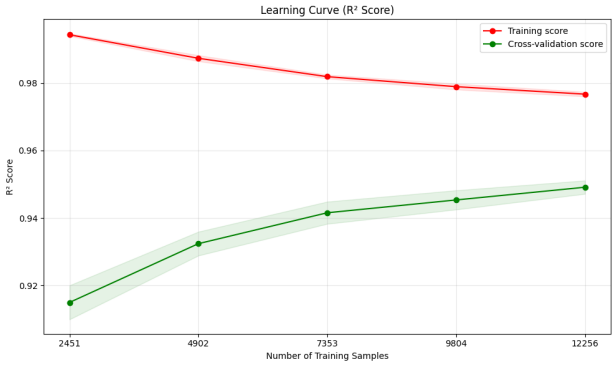

Figure 8: *Learning curve model CatBoostRegressor menunjukkan skor R^2 pada data training dan validasi silang terhadap peningkatan ukuran sampel pelatihan. Garis merah menunjukkan performa pada data latih, sedangkan garis hijau menunjukkan performa pada data validasi.*

Analisis Learning Curve Analisis learning curve pada Gambar 8 memberikan insight penting tentang perilaku model saat jumlah data pelatihan bertambah:

Table 5: *Perubahan skor R^2 berdasarkan jumlah sampel pelatihan*

Training Samples	Training R²	CV R² (Mean)
2451	0.9943	0.9150
4902	0.9873	0.9324
7353	0.9819	0.9416
9804	0.9789	0.9454
12256	0.9767	0.9491

Berdasarkan learning curve, dapat diidentifikasi beberapa karakteristik model:

- **Penurunan Skor Training**: Skor R^2 pada data latih sedikit menurun dari 0.994 menjadi 0.977 seiring bertambahnya data. Fenomena ini normal karena model belajar dari lebih banyak variasi data dan tidak lagi "menghafal" pola spesifik.
- **Peningkatan Skor Validasi**: Skor validasi silang meningkat dari 0.915 menjadi 0.949, yang mengindikasikan bahwa kemampuan generalisasi model terhadap data baru semakin baik.
- **Konvergensi Gap**: Selisih antara skor training dan validasi mengecil hingga mencapai sekitar 0.0276 pada titik akhir. Meskipun masih ada sedikit overfitting, gap ini relatif kecil dan menunjukkan trend penurunan yang mengindikasikan model yang seimbang.

Berdasarkan analisis ini, model CatBoostRegressor menunjukkan keseimbangan yang baik antara bias dan varians, dengan potensi peningkatan performa marginal jika ditambahkan lebih banyak data pelatihan.

Hasil Evaluasi Cross-Validation Model terbaik yang diperoleh melalui penalaan hiperparameter menunjukkan performa prediktif yang tinggi terhadap target hydrostatic_pressure. Berikut ringkasan metrik evaluasi:

- Root Mean Square Error (RMSE): 0.0004
- Mean Absolute Error (MAE): 0.0002
- Coefficient of Determination (R^2): 0.9568

Nilai R^2 yang tinggi menunjukkan bahwa 95.6% variasi dalam target dapat dijelaskan oleh fitur-fitur prediktor, sedangkan nilai kesalahan absolut yang rendah menandakan bahwa prediksi sangat dekat dengan nilai aktual.

Visualisasi Evaluasi Model Model divisualisasikan melalui dua pendekatan utama:

- Plot Aktual vs Prediksi menunjukkan bahwa sebagian besar prediksi mengikuti garis identitas ($y = x$), menandakan estimasi yang presisi.
- Plot Residual menunjukkan distribusi error yang menyebar secara acak di sekitar nol, tanpa pola sistematik, yang mengindikasikan ketidakhadiran bias signifikan.

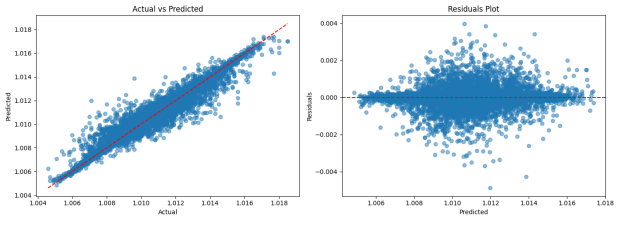

Figure 9: *Evaluasi Model: (a) Hubungan antara nilai aktual dan prediksi. (b) Plot residual.*

Kesimpulan

Penelitian ini berhasil membangun model prediksi tekanan hidrostatik berbasis algoritma machine learning CatBoost dengan memanfaatkan data oseanografi multivariabel yang kompleks dan tinggi dimensi. Proses pemodelan dimulai dari pembersihan dan transformasi data, analisis eksploratif, hingga teknik feature engineering dan selection yang mempertimbangkan aspek fisis, statistik, dan domain ilmu kelautan.

Hasil eksplorasi menunjukkan bahwa tekanan hidrostatik sangat dipengaruhi oleh fitur-fitur seperti densitas air, suhu, salinitas, dan tekanan dasar laut. Selain itu, pola siklik temporal serta kondisi biologis seperti densitas plankton juga memberikan kontribusi signifikan. Normalisasi dan transformasi data yang dilakukan berhasil meningkatkan kestabilan distribusi dan relevansi fisis antar variabel.

Dengan pendekatan preprocessing yang cermat dan rekayasa fitur yang mempertimbangkan struktur data serta dinamika oseanografi, model CatBoost mampu memanfaatkan interaksi nonlinier antar fitur secara efektif. Pendekatan ini membuka peluang lebih luas untuk penerapan machine learning dalam memahami dinamika laut secara presisi, khususnya dalam prediksi tekanan hidrostatik yang esensial bagi aplikasi oseanografi operasional, rekayasa laut, dan monitoring lingkungan laut.

References

[1] T. P. Boyer et al. *World Ocean Database 2018*. Vol. 87. NOAA, 2018.

[2] S. Levitus et al. “World ocean heat content and thermosteric sea level change (0–2000 m), 1955–2010”. In: *Geophysical Research Letters* 39.10 (2012).

[3] R. A. Feely et al. “Impact of anthropogenic CO2 on the CaCO3 system in the oceans”. In: *Science* 305.5682 (2004), pp. 362–366.

[4] L. D. Talley et al. *Descriptive Physical Oceanography: An Introduction*. 6th ed. Academic Press, 2011.

[5] S. A. Thorpe. *An Introduction to Ocean Turbulence*. Cambridge University Press, 2007.

[6] J. A. Barth et al. “Delayed-mode calibration of autonomous CTD profile data by -S climatology”. In: *Journal of Atmospheric and Oceanic Technology* 16.6 (1999), pp. 188–202.

[7] T. Chen and C. Guestrin. “XGBoost: A Scalable Tree Boosting System”. In: *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. 2016, pp. 785–794.

[8] L. Prokhorenkova et al. “CatBoost: Unbiased Boosting with Categorical Features”. In: *Advances in Neural Information Processing Systems*. Vol. 31. 2018, pp. 6638–6648.

[9] Y. Zhang et al. “Machine learning applications in ocean temperature prediction: A comprehensive review”. In: *Deep Sea Research Part I* 168 (2021), pp. 103–118.

[10] L. Wang et al. “Satellite-derived chlorophyll-a prediction using gradient boosting regression”. In: *Remote Sensing of Environment* 246 (2020), pp. 111–125.

[11] S. García et al. *Data Preprocessing in Data Mining*. Springer, 2015.

[12] I. Guyon and A. Elisseeff. “An introduction to variable and feature selection”. In: *Journal of Machine Learning Research* 3 (2003), pp. 1157–1182.

[13] B. Wehbe and F. Kirchner. “3D-DUO: 3D detection of underwater objects in low-resolution multibeam echosounder maps”. In: *Ocean Engineering* (2025).

[14] F. F. Mojtahedi et al. “Forecasting of Turbidity Currents using DL-CFD Hybrid Models”. In: *Ocean Engineering* (2025).

[15] S. Xu et al. “Rotation-Invariant Sonar Image Segmentation”. In: *IEEE Journal of Oceanic Engineering* (2025).

[16] H. F. Tolie et al. “Underwater Image Quality Assessment Using Dispersion Metrics”. In: *IEEE Transactions* (2025).

[17] J. H. Friedman. “Greedy Function Approximation: A Gradient Boosting Machine”. In: *Annals of Statistics* 29.5 (2001), pp. 1189–1232.