BAB II
TINJAUAN PUSTAKA DAN DASAR TEORI

2.1. Tinjauan Pustaka

Pemanfaatan teknik ML untuk menganalisis data oseanografi CTD sebagai bahan kajian prediksi sumber daya kelautan relatif baru. Oleh karena itu perkembangan ilmu ML dari topik penelitian sebelumnya mengalami perkembangan yang lebih variatif seiring dengan bertambahnya wawasan ML yang dipelajari dengan menggabungkan ilmu oseanografi. Faktor kesamaan dengan penelitian terdahulu menjadi bahan inspiratif sebagai hasil pemikiran yang layak untuk dikutip dan disertai kutipan sebagai pengembangan keilmuan yang melandasi penelitian ini. Berikut ini adalah tinjauan literatur yang disajikan pada Tabel 2.1, yang menguraikan ringkasan berbagai bidang penelitian sebelumnya.

Tabel 2.1 Ringkasan tinjauan literatur penelitian terdahulu

Area Riset	Metode	Permasalahan	Hasil Riset	Referesi
Earth Science	*CNN, LSTM, Con LSTM*	AI secara progresif diterapkan pada penelitian maritim, melengkapi model dan pengamatan peramalan laut tradisional. AI model algoritmik yang mereferensikan uji coba aplikasi, dan secara metodis menguraikan tren penelitian dari mesin pembelajar.	Penelitian ini menyoroti penggunaan algoritme AI untuk mendeteksi dan mengenali gelombang panas, fenomena El Niño., dan es laut untuk meramalkan komponen laut.	(Song, dkk, 2023)
Applied Computing and Geosciences	*KNN, LR, DT, SVM, NB, RF, MLP*	Memodelkan akuifer heterogen yang akurat, menjadi tantangan besar dalam hidrogeologi. Ada kebutuhan mendesak untuk mengembangkan metode baru yang mengubah data beresolusi tinggi.	Prediksi unit hidrostratigrafi menunjukkan skor Fl pelatihan 97% dan akurasi pengujian 91%, meningkat menjadi 100% dan 95% setelah penyetelan.	(Tilahun & Korus, 2023)

Tabel 2.1 Ringkasan tinjauan literatur penelitian terdahulu (lanjutan)

Area Riset	Metode	Permasalahan	Hasil Riset	Referesi
Knowledge-Based Systems	*SVR, RF, ANN*	ML digunakan untuk memprediksi suhu permukaaan jalan dalam rangka Pemeliharaan Jalan Musim Dingin (WRM) di negara Nordik yang hemat biaya dan dapat membuat rencana awal yang tepat untuk mengembangkan sistem pendukung keputusan dalam meningkatkan keselamatan lalu lintas di jalan sekaligus mengurangi biaya dan dampak negatif lingkungan.	Empat metode berbeda digunakan untuk memprediksi suhu permukaan jalan yang tidak efisien. Salah satunya adalah dengan regresi linier, merupakan teknik regresi statistik klasik; tiga lainnya adalah metode teknik pembelajaran mesin, termasuk mendukung regresi vektor, multilayer perceptron jaringan saraf tiruan, dan regresi hutan acak.	(Hatamzad, dkk, 2022)
Marine Science	*A_CNN*	Dikarenakan kurangnya pengetahuan tentang wilayah laut dalam dan wilayah kutub, manusia belum dapat menguraikan fenomena dan pola spesifik di lautan. AI dilatih pada model matematika dengan struktur menggunakan sejumlah besar data statistik untuk mendapatkan fitter yang berisi fitur statistik melekat pada data pelatihan. Ini dapat diterapkan untuk memecahkan masalah optimasi.	Pertama, data kelautan memasuki era data besar yang kaya informasi. Kedua, infrastruktur pengelolaan data kelautan yang ada harus digunakan untuk meningkatkan pemrosesan dan pengelolaan data agar data dapat dibaca oleh mesin. Ketiga, dengan integrasi ilmu kelautan dan teknologi data besar.	(Jiang & Zhu, 2022)
Environment Management	*ANN & SVM*	ML digunakan untuk memberi peringatan adanya insiden potensi pemekaran alga berbahaya (HAB) juga untuk pemodelan dan peramalan masalah kualitas air di pelabuhan yang terkena dampak HAB.	Penelitian ini menggunakan dua model ML yaitu ANN & SVM. Kinerja model SVM hasilnya lebih baik daripada model ANN dalam memprediksi kualitas air.	(Deng, dkk, 2021)

Tabel 2.1 itu adalah ringkasan tinjauan literatur penelitian terdahulu yang merupakan referensi perkembangan keilmuan ML pada penelitian ini terutama sitasi tahun 2021 – 2022, hal yang menjadi pertimbangan:

- Penggunaan ML untuk pemodelan dan prediksi (Deng, dkk, 2021).
- AI dilatih pada model matematika dengan menggunakan data statistik dalam jumlah besar (Jiang & Zhu, 2022).
- Penggunaan ML untuk prediksi (Hatamzad, dkk, 2022).

Sedangkan sitasi tahun 2023 menjadi referensi pembanding, alasannya:

- AI diterapkan secara progresif pada penelitian kelautan, untuk melengkapi model yang sudah ada dan pengganti monitoring peramalan laut yang masih bersifat tradisional (Song, dkk, 2023).
- AI digunakan sebagai pengembangan metode baru yang mengubah data beresolusi tinggi menjadi perwakilan parameter hidrogeologi akuifer (Tilahun & Korus, 2023).

2.2. Dasar Teori

2.2.1. *Machine Learning* (ML)

ML adalah komputer memodifikasi data dengan cara membuat prediksi, sehingga menjadi lebih akurat karena akurasi diukur dari seberapa benar data yang dimodifikasi atau diprediksi sesuai dengan keinginan (Marsland, 2011). Definisi ML lainnya, adalah ilmu komputer yang memberikan kemampuan sistem komputer untuk belajar, semakin meningkatkan kinerja dengan data maka tanpa diprogram akan dengan sendirinya dieksplisit oleh ML (Mathur, 2019). Pengetahuan mengenai ML secara luas, bahwa ML memiliki 6 tahapan metodologi yang ditunjukan pada Gambar 2.1 (Samudrala, 2018):

- *Business Problem*

Permasalahan sistem organisasi atau sistem aktivitas merupakan hal yang penting dalam ML untuk menjadi penggerak utama, agar dapat diatasi dengan algoritma yang sesuai dengan ML.

- *Data Discovery*

Proses mengidentifikasi dan mengoleksi sumber data internal maupun eksternal yang dapat membantu mengatasi permasalahan bisnis. Kebanyakan model ML tidak dapat menggunakan sebagian besar atribut pada data mentah karena akan menghasilkan kesalahan jika digunakan secara langsung (Boehm, dkk, 2019). Contoh, atribut yang digunakan pada data mentah adalah nama dan alamat. Jika atribut digunakan langsung sebagai fitur pada ML, maka atribut akan diperlakukan sebagai fitur kategorikal dengan domain yang sangat besar dikarenakan nama atau alamat memiliki artian yang sangat luas. Oleh karena itu data yang digunakan harus diolah terlebih dahulu dari data mentah menjadi sebuah vektor fitur yang tepat dengan cara transformasi data, hal itu disebut *feature engineering/feature extraction.*

- *Model Selection*

Pemilihan model dengan algoritma ML, tujuannya untuk mencapai bias dan varian yang rendah. Bias, merupakan asumsi penyederhanaan yangdibuat oleh model untuk membuat fungsi target agar lebih mudah belajar. Varian, merupakan jumlah target yang akan berubah jika data pelatihan yang digunakan berbeda. Pemilihan model meliputi dua proses:

✓ Pemilihan algoritma, proses penentuan hipotesis mana yang berfungiuntuk aplikasi. Praktiknya, pemilihan algoritma tidak hanya dipengaruhi oleh tingkat akurasi prediksi saja, melainkan kombinasi kompleks dari faktor; teknis dan non teknis, biaya sumber daya, ketersediaan alat pelatihan, dan penilaian khusus penggunaan fungsi prediksi.

✓ Penyetelan *hyper parameter,* proses untuk menentukan nilai *hyper parameter* yang digunakan oleh model ML.

- *Training the Model*

Model yang terlatih akan didapat bila memenuhi unsur-unsur seperti ini:

✓ *Training Data*

Biasanya, 60 - 70% dari data histori untuk membangun model ML.

✓ *Validation Data*

Biasanya, 20 - 30% dari data historis digunakan sebagai data validasi.

Validasi digunakan untuk memverifikasi keakuratan model yang terlatih dan melakukan penyesuaian terhadap *hyper parameter* untuk mencapai tingkat akurasi yang diinginkan.

✓ *Test Data*

Biasanya, 10% - 20% dari data histori yang dijadikan sebagai data uji. Hasil prediksi model dibandingkan dengan hasil historis untuk menilai kebenaran model.

- *Roll out the model*

Model yang telah dilatih diintegrasikan ke dalam antarmuka pengguna dengan aplikasi, seperti aplikasi seluler untuk prediksi berdasarkan *input*.

- *Review model*

Model yang digunakan perlu dilatih ulang agar tetap layak, karena:

✓ Model yang digunakan untuk menghasilkan prediksi perlu dicatat pada catatan baru, sehingga ada kebutuhan untuk memasukan data baru dalam rangka memperbaiki pelatihan.

✓ Belajar dari kesalahan, agar ada umpan balik terus menerus tentang kesimpulan yang diberikan benar atau tidak. Hal ini diperlukan untuk memasukkan catatan baru ke pelatihan.

✓ Prediksi menggunakan kumpulan data yang sama selama kurun waktu tertentu oleh banyak pihak akan memiliki hasil yang semakin berkurang dan akan mengakibatkan terjadinya *overfitting* atau *underfitting*. *Overfitting* adalah performansi data pada model yang digunakan untuk memprediksi sangat bagus, tetapi kenyataannya buruk akibat terlalu banyak data yang digunakan saat *training* (Shalev-Shwartz & Ben-David, 2014). *Underfitting* adalah model yang tidak bisa melakukan generalisasi data baru, sehingga tidak bisa dijadikan sebagai model data pelatihan (Brownlee, 2016).

 Evaluasi model harus dilakukan, tujuan untuk mengetahui ketepatan model yang digunakan dan harus diperhatikan permasalahan apa yang ingin diselesaikan, apakah masalah regresi atau klasifikasi. Metrik evaluasi yang umumnya digunakan dalam analisis regresi, ditunjukandengan persamaan 2.1, 2.2, 2.3, 2.4 (Yilmazer & Kocaman, 2020):

- MAE (*Mean Absolute Error*), yaitu perbedaan antara nilai sebenarnya dan nilai prediksi yang diekstraksi dengan rata-rata perbedaan absolut atas kumpulan data.

$$MAE = \frac{1}{n}\sum_{i=1}^{n} (Yi - Yi) \quad (2.1)$$

Y ' = Nilai Prediksi
Y = Nilai Sebenarnya
n = Jumlah Data

- MSE (*Mean Squared Error*), yaitu perbedaan antara nilai sebenarnya dan nilai prediksi yang diekstraksi dengan kuadrat perbedaan rata-rata atas kumpulan data.

$$MSE = \frac{1}{n}\sum_{i=1}^{n} (Yi - Yi)^2 \quad (2.2)$$

Y ' = Nilai Prediksi
Y = Nilai Sebenarnya
n = Jumlah Data

- RMSE (Root Mean Square Error), yaitu tingkat kesalahan oleh akar kuadrat MSE.

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n} (Yi - Yi)^2} \quad (2.3)$$

Y ' = Nilai Prediksi
Y = Nilai Sebenarnya
n = Jumlah Data

- R2 (Koefisien Determinasi), yaitu seberapa baik nilai koefisien dibandingkan dengan nilai yang sebenarnya. Ukuran nilai dari 0 – 1 yang ditafsirkan sebagai presentasi.

$$R2 = \frac{SSregression}{SStotal} \quad (2.4)$$

SSregression adalah jumlah kuadrat akibat regresi (jumlah kuadrat)
SStotal adalah jumlah total kuadrat

Algoritma ML dibagi menjadi 3 jenis, yaitu; *supervised learning*, *unsupervised learning*, dan *reinforcement learning* (Ray, 2019):

Supervised Learning, algoritma yang dilatih agar dapat memilih fungsi yang paling menggambarkan *input*. Fungsi algoritma SL tergantung pada asumsi yang digunakan, jika asumsi tidak terpenuhi maka hasil pengolahan data akan

menjadi bias. Oleh karena itu, algoritma SL membutuhkan data latih yang benar sehingga sistem dapat mempelajari model data. SL menggunakan data latih tujuannya untuk melakukan pembelajaran, yaitu data latih yang sudah berlabel. Data yang telah berlabel akan dilatih pada mesin untuk selanjutnya dipahami dan disimpan oleh mesin, prosesnya terdapat 3 tahapan:

- ✓ *Input data*, yaitu data yang di *input* telah berlabel.
- ✓ *Model training*, yaitu data yang diproses oleh supervisor untuk dilatih dan menghasilkan *output* yang diinginkan dengan menggunakan algoritma tertentu yang sesuai dengan karakteristik datanya.
- ✓ *Output*, yaitu hasil pemodelan data yang telah dilatih.

Jenis model SL; *Classificasi, K-NN, Naive Bayes, Decision Tree, Linier Regression, Random Forest, Support Vector Machine, Neural Network*. SL memiliki kelebihan, yaitu prosesnya sederhana dan mudah dipahami. Algoritma SL bersifat *powerfull* untuk klasifikasi dan data yang digunakan bukanlah data *real time*, sehingga memerlukan data baru untuk memprediksi hasil. Kekurangan algoritma SL, memerlukan waktu komputasi yang cukup panjang untuk pelatihan dan menggunakan algoritma yang lebih kompleks. Penggunaan SL untuk menyelesaikan suatu permasalahan, dibagi menjadi dua jenis (Marsland, 2011), yaitu:

- ➢ *Regression*, metode pemodelan prediktif yang digunakan untuk mempelajari hubungan antara variabel dependen (target) dengan satu atau lebih variabel independen (prediktor). Variabel target, adalah variabel yang akan di prediksi atau dipelajari. Variabel prediktor, adalah variabel yang menjelaskan nilai target pada variabel dependen. Variabel independen dinotasikan dengan (X), variabel dependen dinotasikan dengan (Y). Pembahasan regresi, fokus pada variabel (Y) yang harus berupa nilai numerik. Sedangkan variabel (X) dapat berupa nilai numerik atau katergorikal. Regresi bertujuan untuk menemukan suatu fungsi yang memodelkan data dengan cara meminimalkan selisih antara nilai prediksi dengan nilai sebenarnya. Regresi digunakan untuk memprediksi nilai kontinu. Metode SL biasanya digunakan untuk *forecasting*, prediksi *time*

series analysis dan mencari hubungan sebab akibat diantara variable (Y) dan variabel (X). Tipe Analisis Regresi dapat dibedakan berdasarkan; jumlah variable (X), tipe variable (X), dan bentuk kurva regresi. Jumlah variabel (X), terdiri dari; Bila hanya ada 1 varibel (X) maka harus menggunakan Regresi Linier Sederhana, bila variable (X) >1 maka harus menggunakan Regresi Berganda. Tipe regresi yang akan digunakan dalam penelitian ini adalah Regresi Linier Berganda (Marsland, 2011), yang mana ciri-cirinya sebagai berikut:

- Hubungan antara variable (Y) dengan lebih dari satu variable (X) menggunakan garis lurus yang sesuai.
- Variabel (X) kontinu atau diskrit, variable (Y) kontinu.

Tipe analisis regresi dapat dibedakan berdasarkan: jumlah variabel independen, tipe variabel independen, dan bentuk-bentuk kurva. Berikut ini gambar 2.1 yang akan menjelaskan hal tersebut:

Jumlah variable independen	Tipe variable independen	Bentuk Kurva
• 1 variable (Regresi linier tunggal) • > 1 variable (Regresi linier berganda)	• Kontinu • Diskrit	• Linier • Logistik • Polinomial

Gambar 2.1: Jumlah dan tipe variabel serta bentuk kurva regresi

Gambar 2.1 menunjukkan uraian mengenai; jumlah variabel independen, tipe variable independen, dan bentuk kurva:

- Jumlah variabel independen; jenis 1 variabel disebut regresi linier tunggal, dan jenis lebih dari 1 variabel disebut linier berganda. Jenis jumlah variabel independen lebih dari 1 digunakan pada kajian ini.
- Tipe variable independen, ada variabel diskrit dan kontinu. Kajian ini menggunakan tipe variabel kontinu keseluruhannya.
- Bentuk kurva regresi; kurva linier, logistik, dan polinomial. Kajian ini menggunakan kurva linier untuk melihat korelasi antar data.

Algoritma yang dapat digunakan untuk menyelesaikan permasalahan hubungan antara variabel target dengan variabel prediktor yang lebih dari satu, diantaranya:

- *Decision Tree*, adalah model prediktif bentuk pohon keputusan yang terdiri dari simpul (*node)* yang mewakili keputusan atau tes terhadap fitur. Cabang merepresentasikan hasil keputusan atau tes, dan daun mewakili hasil prediksi. Pada setiap simpul, algoritma memilih fitur terbaik untuk membagi data berdasarkan kriteria seperti; g*ini impurity* atau *gain information*. Metode DT ada dua jenis, yaitu *Classification and Regression Trees* (CART). Penerapannya harus menggunakan algoritma $C_{4.5}$, yaitu algoritma yang umumnya menggunakan data numerik dan kategorikal. Metodologinya, menyiapkan data latih untuk memilih atribut yang dihitung, kemudian menggunakan rumus *Entropy, Split information, dan Gain Rasio,* ditunjukan pada persamaan 2.5, 2.6, dan 2.7 (Rokach & Maimon, 2010).

$$Entropy(S) = \sum_i^c - Pi\ log2\ Pi \quad (2.5)$$

c = jumlah kelas ; Pi = data objek ; i = jumlah sampel kumpulan data

$$Split\ Information = \sum_i - \frac{Si}{S} log_2 \frac{Si}{S} \quad (2.6)$$

S = kumpulan sampel data.

S1 sampai Sc = himpunan bagian sampel data yang dibagi berdasarkan banyaknya variasi nilai atribut A.

$$Gain\ Ratio(S,A) = \frac{Gain(S,A)}{Split\ Information(S,A)} \quad (2.7)$$

- *Linear Regression,* adalah persamaan matematis yang menggambarkan hubungan antara variabel bebas dan terikat, sering disebut persamaan regresi. Tujuannya adalah menemukan garis linear yang menggambarkan hubungan. Hubungan variabel itu ditunjukan dalam persamaan 2.8 (Marsland, 2011):

$$Y = \alpha + \beta 1\ X2 + \beta 2\ X2 + \beta n\ Xn + e \quad (2.8)$$

Y = Variabel terikat atau variabel respon.

X = Variabel bebas atau variabel prediktor.

α = Konstanta.

β = Kemiringan atau estimasi koefisien

Sedangkan penggunaan metode statistik *Residual Sum of Squares* (RSS) untuk mengidentifikasi tingkat perbedaan dataset yang tidak diprediksi oleh model regresi, ditunjukan pada persamaan 2.9 (Maulud & Abdulazeez, 2020).

$$RSS = \sum_{i=1}^{n} (y_i - y)^2 \qquad = \sum_{i=1}^{n} (y_i - (a + \beta x_i))^2 \qquad (2.9)$$

Y i adalah nilai variabel yang diamati

Y adalah nilai yang diperkirakan dengan regresi linier

X i adalah nilai independen

α dan β bersifat konstan

- *Random Forest,* disebut dengan e*nsemble machine learning* adalah teknik yang menggunakan beberapa algoritme pohon keputusan untuk membuat satu algoritme yang kuat (Breiman, 2001). RF bergantung pada banyak pohon, yang membantu memiliki daya prediksi yang paling efisien dengan lebih sedikit ketidakpastian (Tilahun & Korus, 2023). RF digunakan sebagai pembanding karena tidak ditemukan pada metode lain (Speiser, dkk, 2019). RF juga dapat menangani kumpulan data yang mengandung variabel kontinu seperti pada kasus regresi dalam kajian ini. Hutan Acak terdiri dari klasifikasi berbentuk pohon {h(x, θ k), k = 1, .} di mana θk adalah vektor acak yang terdistribusi secara independen dan setiap pohon akan memilih kelas paling populer pada input X. Berikut adalah karakteristik akurasi dari hutan acak: ada pengklasifikasi h1(x), h2(x), hk (x) dan dengan kumpulan data latih yang di distribusi vektor acak Y, X (Han, dkk, 2018). Penerapan algoritma RF dengan tahapan; menentukan berapa jumlah pohon keputusan, mengambil data sampel secara random untuk membentuk pohon keputusan, data sampel dihitung dengan Indeks Gini untuk menentukan simpul teratas, ditunjukan pada persamaan 2.10 (Speiser, dkk, 2019).

$$Gini = 1 - \sum (pi)2\ n \qquad (2.10)$$

i = 1 ; pi = probabilitas objek yang diklasifikasikan dalam fitur tertentu.

- *Clasificasion,* model klasifikasi digunakan untuk memprediksi keluaran yang bernilai diskrit dan menjadi ukuran model ML dari hasil prediksi sebagai ukuran keakuratan model. Poin paling penting mengenai klasifikasi, untuk mendapatkan keluarannya adalah diskrit (Samudrala, 2018). Pengertian klasfikasi secara matematis, adalah mendekati fungsi pemetaan (f) dari variabel *input* ke variabel *output.* Penerapan klasifikasi, caranya harus melatih data yang diklasifikasikan. Misalnya data nomor telepon yang "spam" dan "bukan spam" digunakan sebagai data pelatihan, selanjutnya data yang telah dilatih dapat digunakan untuk mendeteksi nomor telepon yang tidak dikenal. Ada 2 cara penerapan klasifikasi, pertama yang dinamakan *Learner Classification* yaitu menunggu data pengujian, selanjutnya menyimpan data pelatihan klasifikasi yang dilakukan hanya setelah mendapatkan data pengujian. Dalam hal ini, waktu yang dibutuhkan untuk pelatihan lebih sedikit, tetapi lebih banyak waktu untuk memprediksi. Kedua yang dinamakan *Eager Classification*, yaitu tanpa harus menunggu data pengujian setelah menyimpan data pelatihan. Dalam hal ini, waktu yang dibutuhkan untuk pelatihan lebih banyak, tetapi lebih sedikit waktu yang dibutuhkan untuk memprediksi.

  Unsupervised Learning, algoritma yang diajari untuk mengenali pola dan struktur data tanpa menggunakan label, memungkinkan model untuk mengeksplorasi dan menemukan informasi yang belum diketahui sebelumnya. Model diberikan *dataset* yang tidak berlabel dan diharapkan dapat menemukan pola atau hubungan sendiri. *unsupervised learning* juga didefinisikan sebagai seperangkat pelatihan dimana hanya ada *input* data dan tidak ada variabel keluaran yang sesuai (Brownlee, 2016). Tujuan dari *unsupervised learning*, untuk memodelkan distribusi data agar dapat mempelajari lebih lanjut tentang data. *Unsupervised Learning* memiliki beberapa jenis, diantaranya:

  - *Clustering,* pengelompokan data ke dalam beberapa *cluster* berdasarkan kesamaan karakteristik internal, contoh algoritma:
    - ✓ *K-Means,* memisahkan data ke kelompok berdasarkan pusat *cluster*.

- ✓ *Hierarchical Clustering,* membangun hirarki kelompok dengan menggabungkan kelompok berdasarkan kedekatan spasial.
- ✓ *Density-Based Spatial Clustering of Applications with Noise* (DBSCAN), mengidentifikasi kelompok kepadatan data.

- *Dimensionality Reduction*, mengurangi jumlah fitur dalam data tanpa kehilangan informasi signifikan, contoh algoritma:
  - ✓ *Principal Component Analysis* (PCA), mereduksi dimensi data dengan mengidentifikasi arah utama variasi.
  - ✓ *t-Distributed Stochastic Neighbor Embedding* (t-SNE), mereduksi dimensi data untuk visualisasi pada ruang yang lebih rendah.
  - ✓ *Autoencoders,* jaringan saraf tiruan yang mempelajari representasi data melalui peta dari dimensi tinggi ke rendah.
- *Association Rule Learning,* mencari asosiasi yang kuat antara variabel dalam data, contoh algoritma:

  Apriori, menemukan asosiasi antar item dalam dataset transaksional, seperti dalam analisis keranjang belanja.
- *Generative Modeling,* model belajar untuk memahami dan mereplikasi distribusi data untuk pembuatan data baru yang mirip, contoh algoritma:
  - ✓ *Variational Autoencoders*, hasil data baru dari distribusi data.
  - ✓ *Generative Adversarial Networks,* membangun generator yang dapat membuat data sulit dibedakan dari data asli.
- *Anomaly Detection,* mengidentifikasi *instance* yang tidak umum dalam data, contoh algoritma:
  - ✓ *One-Class SVM,* mengidentifikasi pola normal dan mendeteksi anomali berdasarkan deviasi dari pola normal.
  - ✓ *Isolation Forest*, membangun pohon isolasi untuk mengidentifikasi *instance* anomali.
- *Self-Organizing Maps*, metode pengelompokan dan reduksi dimensi yang menggunakan jaringan saraf untuk menghasilkan peta dua dimensi dari data, contoh algoritma:

Self-Organizing Maps, mengorganisir data dalam *grid* dua dimensi di mana kelompok yang serupa berdekatan satu sama lain.

Setiap jenis *unsupervised learning* memiliki aplikasi dan kegunaan tersendiri, tergantung pada karakteristik data dan tujuan analisis. Kombinasi dari beberapa teknik dapat memberikan wawasan yang lebih mendalam ke dalam struktur dan pola yang mungkin tersembunyi dalam data tanpa adanya label. *Unsupervised Learning* yang digunakan saat kita ingin mengeksplorasi data dan menemukan pola atau struktur tanpa memiliki label khusus untuk pelatihan model yang diinginkan. *Unsupervised Learning* juga memiliki aplikasi yang luas, termasuk analisis data eksploratif. Adanya teknologi yang terus berkembang, *unsupervised learning* terus menjadi area penelitian yang menarik untuk menjelajahi dan memahami data secara lebih mendalam.

Orange3 memiliki beberapa widget yang dirancang khusus untuk tugas regresi dan berikut ini adalah beberapa *widget orange3* yang dapat digunakan:
*Linear Regression, m*embangun model regresi linear.
Dukungan *Orange3, Widget* "*Linear Regression*" dapat digunakan untuk menerapkan regresi linear pada data.
Regression Tree, membangun model regresi berbasis pohon keputusan.
Dukungan *Orange3, Widget* "*Regression Tree*" memungkinkan penerapan regresi berbasis pohon.
Random Forest Regressor, menggunakan algoritma *Random Forest* dalam tugas regresi.
Dukungan *Orange3, Widget* "*Random Forest*" dapat digunakan untuk membangun model regresi dengan menggunakan *Random Forest*.
Support Vector Regression (SVR), menerapkan regresi dengan SVM.
Dukungan *Orange3, Widget* "SVR" memungkinkan penerapan regresi menggunakan SVR.
Gradient Boosting Regressor (GBR), membangun model regresi dengan menggunakan teknik GBR.
Dukungan *Orange3, Widget* "*Gradient Boosting*" dapat digunakan untuk menerapkan regresi dengan teknik *gradient boosting*.

Widget orange3 dapat digunakan untuk membangun model regresi, mengevaluasi performanya, dan memvisualisasikan hasilnya secara intuitif dalam alur kerja visual *orange3*. Adapun langkah umumnya adalah memasukan data, memilih model regresi yang sesuai, melatih model, dan mengevaluasi kinerja menggunakan metrik yang relevan.

- *Reinforcement Learning*, pelatihan yang menggabungkan konsep *supervised learning* dan *unsupervised learning* yaitu suatu agen belajar untuk membuat keputusan dengan berinteraksi dalam suatu lingkungan untuk mencapai tujuan tertentu. Tujuannya adalah untuk memaksimalkan nilai tertentu selama agen berinteraksi dengan lingkungan yang dinamis. Konsep dasar *reinforcement learning* secara analogik, untuk menggambarkan pembelajaran dan penyesuaian terus-menerus dalam hubungan.

2.2.2. Orange3 untuk *Unsupervised Learning*

Orange3 adalah perangkat lunak sumber terbuka yang digunakan untuk analisis data dan ML. Penggunaan *orange3* dapat digunakan untuk membangun model *unsupervised learning* dengan mudah melalui antarmuka pengguna grafis yang intuitif. Berikut ini adalah poin-poin penting tentang penggunaan *orange3* untuk membangun model *unsupervised learning*:

- Antarmuka Pengguna Visual, o*range3* menyediakan antarmuka pengguna visual berbasis alur kerja yang memungkinkan pengguna untuk membangun model dengan menggabungkan *widget-widget* pada *canvas*. *Widget* mewakili berbagai langkah dalam proses analisis data atau pembuatan model, dan dapat disusun secara visual untuk membentuk alur kerja sesuai dengan kebutuhan. Visual berbasis alur kerja, mencakup langkah pemrosesan data hingga pembuatan model dan evaluasi hasilnya. Hal ini membuatnya lebih mudah untuk memahami aliran pekerjaan dan mengidentifikasi setiap langkah dalam analisis data.
- Manipulasi dan Eksplorasi Data, sebelum membangun model, *orange3* dapat digunaian untuk melakukan manipulasi dan eksplorasi data. Sehingga distribusi data, identifikasi nilai yang hilang dapat diketahui, dan dapat

melakukan berbagai operasi praproses data melalui *widget* yang disediakan. Evaluasi Model, setelah membangun model, *orange3* menyediakan *widget* untuk mengevaluasi kinerja model dengan menggunakan metrik evaluasi yang relevan untuk tugas *unsupervised learning*, seperti metode evaluasi klaster atau reduksi dimensi.

- Visualisasi Hasil, *orange3* dapat memvisualisasikan data, hasil reduksi dimensi, dan hasil klaster untuk membantu memahami struktur data.

Dengan kombinasi antarmuka pengguna visual, widget-widget khusus unsupervised learning, dan fungsionalitas eksplorasi data yang kuat, Orange3 merupakan alat yang berguna untuk eksplorasi dan pemodelan *unsupervised learning*. Ini sangat berguna bagi pengguna yang ingin membangun model tanpa pengetahuan mendalam tentang pemrograman atau statistika.

2.2.3. Oseanografi *Conductivity, Temperature, Depth* (CTD)

Oseanografi adalah cabang ilmu bumi yang mempelajari samudra dan laut, termasuk segala aspek yang terkait seperti fisika laut, kimia laut, biologi laut, dan geologi laut (Kennish, 2019):

- ✓ Fisika Laut, mempelajari sifat fisik air laut termasuk; suhu, tekanan, gelombang laut, dan arus laut. Pakar fisika laut akan memahami bagaimana terjadinya perubahan yang mempengaruhi ekosistem laut.
- ✓ Kimia Laut, meneliti komposisi kimia air laut termasuk; distribusi garam, unsur kimia, dan polutan. Studi tentang hal ini akan membantu dalam pemahaman mengenai siklus biogeokimia di dalam laut.
- ✓ Biologi Laut, fokus pada kehidupan laut, baik itu mikroorganisme hingga makrofauna. Pakar biologi laut memahami perilaku, evolusi, dan interaksi antarorganisme di dalam ekosistem laut.
- ✓ Geologi Laut, mempelajari formasi dan struktur dasar laut, termasuk pegunungan bawah laut, lembah-lembah, dan dasar samudra. Ini mencakup pemahaman tentang tektonika lempeng, aktivitas vulkanik, dan pembentukan kontur dasar laut.

Fungsi utama perangkat CTD adalah untuk mendeteksi bagaimana konduktivitas dan suhu berubah secara relatif terhadap kedalaman (Halverson, dkk, 2017). Konduktivitas adalah ukuran seberapa baik suatu larutan air laut menghantarkan listrik dan ini berhubungan langsung dengan salinitas. Dengan mengukur konduktivitas air laut, salinitas dapat diturunkan dari suhu dan tekanan air yang sama. Kedalaman kemudian diturunkan dari pengukuran tekanan dengan menghitung massa jenis air dari suhu dan salinitas. CTD adalah parameter data penting yang digunakan dalam semua disiplin ilmu oseanografi, karena memberikan informasi penting tentang sifat fisik, kimia, dan biologi.

Pada kapal eksplorasi, perangkat CTD sering dipasang pada rangkaian pengambilan sampel air laut yang dikenal sebagai *roset* yang diturunkan ke dalam air melalui kabel. Beberapa botol pengambilan sampel air laut disebut *niskin* yang dilekatkan pada *roset*. Botol-botol ini terbuka saat *roset* disebarkan dan dapat dipicu untuk menutup, mengumpulkan sampel air laut pada kedalaman tertentu untuk analisis selanjutnya. Perangkat CTD sering menyertakan instrumen tambahan, seperti; sensor untuk mengukur oksigen, pH air, kadar nitrat dan klorofil, kekeruhan, dan kecepatan arus air. Semua pengukuran ini dapat dilihat berdampingan dalam kaitannya dengan kedalaman. Perangkat CTD dapat mengirimkan data kembali ke kapal secara *real time*, sementara yang lain menyimpan data hingga instrumen dipulihkan dan data diunduh untuk ditinjau. Data yang dikumpulkan dari perangkat CTD digunakan untuk menghasilkan profil karakteristik air laut secara relatif terhadap kedalaman. Dengan membandingkan data CTD pada setiap kedalaman, maka karakteristik fisik air dapat dianalisis. Dari data ini, para ilmuwan dapat mendeteksi perubahan air laut yang memerlukan penelitian lebih lanjut. Dengan demikian, data CTD memiliki peran penting dalam membantu para ilmuwan untuk membuat keputusan mengenai area lokasi dalam kegiatan untuk mengeksplorasi selanjutnya.

2.2.4. Sumber Daya Laut

Sumber daya yang meliputi ruang lingkup kehidupan laut (flora dan fauna, seperti; organisme mikroskopis dan habitat laut) mulai dari perairan dalam sampai ke daerah pasang surut di pantai dataran tinggi dan daerah muara yang luas (Mukhophadya, dkk, 2020). Sumber daya laut terdiri dari:

- ✓ Sumber daya yang dapat dipulihkan, seperti; ikan, garam, rumput laut, terumbu karang, fosfat, ombak laut, mutiara, plankton, alga.
- ✓ Sumber daya yang tidak dapat dipulihkan, seperti; minyak dan gas bumi.

Sumber daya laut memiliki aspek potensi ekonomi bilamana dikelola menjadi sumber daya yang mendukung kegiatan ekonomi manusia. Berikut adalah aspek potensi ekonomi dari sumber daya laut:

- ✓ Perikanan adalah salah satu aspek utama eksploitasi sumber daya laut memiliki potensi ekonomi yang melibatkan analisis keberlanjutan penangkapan ikan dan manajemen perikanan.
- ✓ Energi terbarukan, seperti; energi gelombang, pasang surut, dan angin laut memiliki potensi ekonomi yang melibatkan penilaian efisiensi teknologi dan integrasi sumber daya energi terbarukan dalam infrastruktur energi.
- ✓ Dasar laut mengandung potensi tambang mineral dan sumber daya alam lainnya memiliki potensi ekonomi yang melibatkan penelitian tentang keberlanjutan eksploitasi, teknologi pertambangan.
- ✓ Pariwisata laut memiliki potensi ekonomi yang signifikan, termasuk kegiatan seperti selam, snorkeling, perjalanan kapal pesiar, dan lainnya. Potensi ekonomi di sektor pariwisata yang melibatkan penilaian daya tarik destinasi laut, infrastruktur pariwisata, dan keberlanjutan.
- ✓ Lautan menyediakan jalur transportasi internasional yang penting memiliki potensi ekonomi yang melibatkan analisis terhadap hubungan perdagangan global, efisiensi transportasi laut, dan pengembangan pelabuhan internasional.
- ✓ Sumber daya minyak dan gas bumi di laut memiliki dampak ekonomi yang signifikan memilik potensi ekonomi di sektor ini melibatkan penilaian cadangan, teknologi ekstraksi, dan dampak sosial-ekonomi.

- ✓ Sumber daya laut juga dapat digunakan untuk pengembangan produk, seperti; kosmetik, obat-obatan, dan bahan kimia memiliki potensi ekonomi melibatkan penelitian dalam inovasi dan pemanfaatan berbagai jenis organisme laut.

Pemahaman yang holistik tentang potensi ekonomi sumber daya laut melibatkan pengintegrasian berbagai disiplin ilmu dan pendekatan untuk memastikan pemanfaatan yang berkelanjutan (Bax, dkk, 2021).