

2. BAB II

KAJIAN LITERATUR

2.1. Palawija

Palawija adalah istilah yang digunakan dalam pertanian Indonesia untuk menyebut sekelompok tanaman yang ditanam sebagai tanaman sekunder setelah tanaman utama dipanen. Tanaman ini biasanya ditanam untuk meningkatkan kesuburan tanah dan memberikan penghasilan tambahan bagi petani. Beberapa contoh tanaman palawija antara lain jagung, kedelai, kacang tanah, kacang hijau, dan banyak lainnya [8].

Palawija, juga dikenal sebagai tanaman semusim atau tumbuhan biji-bijian, memiliki berbagai manfaat, baik dalam konteks pertanian, gizi, dan kesehatan. Berikut beberapa manfaat palawija [9] :

a. Sumber Nutrisi

Palawija kaya akan nutrisi, seperti protein, serat, vitamin, dan mineral. Palawija dapat menjadi sumber makanan yang penting, terutama di daerah dengan keterbatasan pangan.

b. Protein Nabati

Kacang-kacangan, seperti kacang merah, kacang hitam, dan kacang kedelai, adalah sumber protein nabati yang baik. Palawija dapat menjadi alternatif yang sehat untuk protein hewani, terutama bagi vegetarian dan vegan.

c. Serat

Palawija mengandung serat tinggi, yang membantu meningkatkan pencernaan, mengontrol kadar gula darah, dan mengurangi risiko penyakit jantung.

d. Kaya akan Vitamin dan Mineral

Palawija mengandung berbagai vitamin dan mineral penting, seperti besi, magnesium, folat, dan vitamin B. Palawija membantu memenuhi kebutuhan nutrisi tubuh.

e. Sumber Energi

Karbohidrat yang terdapat dalam palawija, seperti kacang-kacangan dan biji-bijian, merupakan sumber energi yang baik, yang penting bagi aktivitas sehari-hari.

f. Makanan Alternatif

Palawija dapat digunakan sebagai alternatif makanan untuk mengurangi konsumsi daging merah, yang terkait dengan risiko kesehatan tertentu. Palawija mendukung diet yang lebih seimbang dan berkelanjutan.

g. Pertanian Berkelanjutan

Pertumbuhan palawija dalam rotasi tanaman membantu meningkatkan kesehatan tanah dan

mengurangi kebutuhan pupuk kimia. Palawija mendukung pertanian berkelanjutan.

h. Keamanan Pangan

Palawija seringkali dapat disimpan dalam waktu lama dan digunakan saat tidak ada sumber makanan segar yang tersedia. Palawija memberikan keamanan pangan dalam situasi krisis.

Palawija adalah komoditas vital dalam sektor pertanian yang memegang peranan penting dalam memenuhi kebutuhan pangan nasional dan menggerakkan ekspor hasil pertanian ke luar negeri [3]. Akan tetapi, produksi palawija sering mengalami tantangan. Pada tahun 2021, indeks produksi berbagai jenis palawija mencatat berbagai perubahan, di antaranya indeks produksi jagung mengalami penurunan sebesar 0,68%, indeks kedelai turun 29,07%, indeks kacang tanah menurun 7,24%, sementara indeks produksi kacang hijau naik sebesar 11,5%. Begitu juga dengan indeks ubi kayu yang mengalami peningkatan sebesar 2,46%. Tantangan ini menggarisbawahi pentingnya menjaga dan meningkatkan produksi palawija dalam konteks keamanan pangan [4].

2.2. Prediksi (*Forecasting*)

Prediksi (*Forecasting*) adalah proses estimasi nilai masa depan berdasarkan informasi dari data-data sebelumnya. Tujuan utama prediksi adalah membantu dalam pengambilan keputusan dengan mempertimbangkan apa yang mungkin terjadi di masa depan. Prediksi bersifat kualitatif dan kuantitatif. Dalam dunia bisnis, prediksi memiliki peran penting dalam pengambilan keputusan manajerial. Prediksi digunakan sebagai dasar untuk perencanaan dalam jangka pendek, menengah, dan panjang. Dengan prediksi, organisasi dapat mengestimasi penjualan atau penggunaan produk. Sehingga, prediksi berfungsi sebagai alat yang membantu dalam pengambilan keputusan, dengan tujuan membuat keputusan yang baik berdasarkan perkiraan kondisi di masa depan [10].

Prediksi (*Forecasting*) memiliki tujuan sebagai berikut [10]:

a. Mengevaluasi Kebijakan Organisasi

Untuk menilai dampak kebijakan yang sedang diterapkan dan yang pernah diterapkan oleh perusahaan, serta melihat bagaimana dampaknya akan berlangsung di masa depan.

b. Mengatasi Keterlambatan Waktu

Adanya keterlambatan antara pembuatan kebijakan perusahaan dan implementasinya, maka prediksi menjadi penting untuk membantu mengatasi masalah keterlambatan.

c. Dasar Perencanaan Bisnis

Prediksi digunakan sebagai dasar dalam menyusun rencana bisnis suatu perusahaan sehingga dapat meningkatkan efektivitas pelaksanaan rencana.

Berdasarkan periode waktu, prediksi dapat dibagi menjadi tiga jenis [10]:

a. Prediksi Jangka Panjang (Lebih dari 18 bulan)

Peramalan jangka panjang (Lebih dari 18 bulan) adalah prediksi yang melibatkan proyeksi masa depan yang mencakup jangka waktu lebih dari 18 bulan. Misalnya, peramalan terkait dengan investasi besar, perencanaan fasilitas, dan penelitian dan pengembangan.

b. Prediksi Jangka Menengah (3-18 bulan)

Prediksi jangka menengah (3-18 bulan) adalah prediksi yang mencakup periode waktu antara 3 hingga 18 bulan. Contohnya adalah peramalan yang digunakan untuk merencanakan penjualan, produksi, atau kebutuhan tenaga kerja yang bersifat sementara.

c. Prediksi Jangka Pendek (Kurang dari 3 bulan)

Prediksi jangka pendek (Kurang dari 3 bulan) adalah prediksi yang berfokus pada periode waktu kurang dari 3 bulan. Contohnya, peramalan yang berkaitan dengan pembelian bahan baku, jadwal kerja, atau penugasan karyawan.

Prediksi dapat dibagi menjadi tiga jenis berdasarkan fungsinya dalam perencanaan operasi masa depan [10]:

a. Prediksi Ekonomi

Prediksi ekonomi adalah jenis prediksi yang digunakan untuk memahami siklus bisnis dengan memproyeksikan faktor seperti inflasi, ketersediaan uang, serta kebutuhan dana untuk membangun perumahan dan indikator perencanaan ekonomi lainnya.

b. Prediksi Teknologi

Prediksi teknologi adalah prediksi yang berfokus pada tingkat kemajuan teknologi yang dapat menghasilkan produk baru yang menarik.

c. Prediksi Permintaan

Prediksi permintaan adalah prediksi yang melibatkan proyeksi permintaan untuk produk atau layanan suatu perusahaan. Prediksi ini berpengaruh pada pengendalian produksi, kapasitas, sistem penjadwalan, dan menjadi landasan untuk perencanaan dalam bidang keuangan, pemasaran, dan manajemen sumber daya manusia.

Dalam hal prediksi, ada beberapa prinsip penting yang harus diperhatikan untuk memastikan hasil prediksi yang efektif [10]:

a. Ketidakpastian Hasil

Prediksi tidak selalu akurat. Proses prediksi digunakan untuk meramalkan apa yang mungkin terjadi di masa depan dengan menggunakan teknik peramalan, namun hasilnya sering kali tidak tepat dan dapat berbeda dari prediksi awal. Kesalahan dalam prediksi

adalah hal yang wajar dan tidak dapat dihindari.

b. Estimasi Kesalahan

Setiap prediksi seharusnya mencantumkan perkiraan tingkat kesalahan (error). Hal ini mengukur sejauh mana tingkat ketidakpastian dalam prediksi. Jika kesalahan dalam prediksi tidak signifikan, hasil prediksi masih bisa menjadi sumber informasi yang berharga.

c. Akurasi Jangka Waktu

Prediksi umumnya lebih akurat ketika digunakan untuk jangka waktu yang lebih dekat. Jangka waktu prediksi yang lebih jauh ke depan memiliki tingkat ketidakpastian yang lebih tinggi dibandingkan dengan prediksi yang mencakup periode lebih pendek.

Berdasarkan pendekatannya, prediksi dapat dibagi menjadi dua jenis [10]:

a. Teknik Kualitatif

Jenis prediksi ini lebih mengandalkan pendapat dan intuisi manusia sebagai dasar prediksi, sehingga data historis memiliki peran yang lebih kecil.

b. Teknik Kuantitatif

Prediksi jenis ini memperhitungkan data historis yang tersedia sebagai komponen penting dalam proses prediksi. Ada tiga pendekatan utama dalam teknik prediksi kuantitatif, yaitu [10]:

i. Analisis Deret Waktu (*Time Series Analysis*)

Pendekatan ini menggunakan data deret waktu historis sebagai dasar peramalan. Prosesnya melibatkan identifikasi pola dalam data historis dan pemilihan metode peramalan yang sesuai.

ii. Analisis Kausal (*Causal Model*)

Pendekatan ini mempertimbangkan hubungan sebab-akibat antara berbagai variabel. Tujuannya adalah memprediksi keadaan di masa depan dengan mengukur pengaruh variabel independen terhadap variabel dependen yang akan di peramalkan. Dalam pendekatan ini, dua metode yang umum digunakan adalah metode regresi dan korelasi, serta metode input-output.

iii. Persamaan Regresi Simultan

Pendekatan ini sering digunakan untuk perencanaan ekonomi nasional dalam jangka waktu tertentu, baik jangka pendek maupun jangka panjang. Pendekatan ini melibatkan pemodelan simultan dari berbagai variabel yang saling terkait.

2.3. OSEM N

OSEMN merupakan metode dalam data science yang melibatkan lima langkah utama: *Obtain*, *Scrub*, *Explore*, *Model*, dan *iNterpret* [11]. *OSEMN* adalah pendekatan terstruktur dalam analisis data yang membantu memastikan bahwa semua langkah yang diperlukan dilakukan dengan baik, sehingga hasilnya akurat dan dapat diandalkan. Metode *OSEMN* digunakan untuk mengekstrak informasi dan pengetahuan dari data, dan dapat disesuaikan dengan berbagai proyek *data science* [11]. Adapun proses *OSEMN* dapat diuraikan sebagai berikut [12]:

1. *Obtain Data*

Obtain dalam metode *OSEMN* merujuk pada upaya mendapatkan data yang diperlukan untuk suatu proyek. Hal ini dapat dilakukan dengan menggunakan berbagai sumber seperti *database*, *API*, atau *web scraping* [12]. Data yang diperoleh harus relevan dengan proyek dan berkualitas tinggi [13]. Setelah data diperoleh, data tersebut dapat digunakan untuk langkah-langkah selanjutnya dalam metode *OSEMN*, yang meliputi *Scrub*, *Explore*, *Model*, dan *iNterpret*.

2. *Scrub Data*

Scrub dalam metode *OSEMN* mengacu pada langkah pembersihan dan prapemrosesan data untuk memastikan bahwa data tersebut tepat dan siap untuk dianalisis. Tindakan ini melibatkan penghapusan duplikasi, penanganan nilai yang hilang, dan penyesuaian format data agar dapat digunakan [13]. Proses *scrub* ini sangat penting karena membantu memastikan bahwa kualitas data terjaga tinggi dan setiap kesalahan atau inkonsistensi diperbaiki sebelum dilakukan analisis [12].

3. *Explore Data*

Eksplorasi data adalah proses dalam menyelidiki data sebelum beralih ke pemodelan *Machine Learning (ML)* [12]. Tahap ini melibatkan penyelidikan menyeluruh terhadap data, dengan fokus pada semua karakteristiknya. Proses ini dikenal sebagai *Exploratory Data Analysis (EDA)*, di mana kita menganalisis kebersihan data, kesiapan data untuk pemodelan, dan jenis fitur yang perlu diselidiki dengan metode statistik [12]. Pemahaman mendalam terhadap data menjadi kunci utama untuk mengekstrak nilai informasi. Melalui eksplorasi data, kita dapat mengidentifikasi subset data yang akan digunakan dalam langkah-langkah pemodelan selanjutnya [13].

4. *Model Data*

Pemodelan Data merupakan fase menarik dalam proyek *data science*, sering disebut sebagai "tempat keajaiban terjadi" [12]. Proses ini melibatkan penentuan model untuk memprediksi

data. Dalam tahap pemodelan, penting untuk memilih karakteristik yang relevan untuk meramalkan hasil [13].

5. *iNterpret Data*

Interpretasi Data adalah langkah akhir dalam metode *OSEMN*. Interpretasi merujuk pada penyajian data untuk menjawab pertanyaan bisnis dan memberikan wawasan yang dapat diimplementasikan melalui *data science* [12]. Pentingnya interpretasi hasil penelitian terletak pada pemahaman efektivitas penelitian. Dalam hal ini, deskripsi hasil yang jelas penting untuk perbandingan dengan penelitian lain. Hasil perlu diinterpretasikan secara objektif dan kritis sebelum mengevaluasi implikasinya dan menyimpulkan temuan. Meskipun bukan fase teknis, interpretasi berkaitan erat dengan pemahaman data, termasuk cara merangkum hasil model. Proses ini melibatkan penarikan kesimpulan yang bermakna dan rasionalisasi wawasan, membantu menentukan tindakan selanjutnya. Interpretasi melibatkan penyajian data dengan cara yang jelas, memberikan hasil bersama dengan wawasan yang dapat diimplementasikan melalui *data science*, serta memahami analitik prediktif untuk mengulangi hasil positif atau mencegah hasil negatif [13].

Tahapan-tahapan *OSEMN* dapat dilihat pada Gambar 2.1 dibawah ini.



Gambar 2.1 Tahapan *OSEMN*

2.4. *Machine Learning*

Machine learning (ML) adalah pengembangan mesin yang mampu belajar secara mandiri tanpa memerlukan bimbingan dari pengguna. Konsep dasar *machine learning* ini berlandaskan prinsip-prinsip dari berbagai bidang ilmu seperti statistika, matematika, dan data mining, yang memungkinkan mesin untuk belajar dengan cara menganalisis data. Dalam konteks ini, *machine learning* memiliki kapabilitas untuk menggali data sendiri tanpa perlu instruksi khusus, dan juga dapat memahami serta mengekstraksi informasi dari data yang sudah ada maupun data yang diterima, sehingga dapat menjalankan berbagai tugas yang beragam, tergantung pada apa yang telah dipelajarinya [14].

Machine learning memiliki dua teknik, yaitu [14]:

a. *Supervised Learning*

Supervised learning adalah teknik di mana mesin dapat memproses data yang sudah memiliki label tertentu, yang memberikan petunjuk tentang informasi di dalam data tersebut. Dalam hal ini, mesin menggunakan pengalaman belajar dari data yang telah diberi label untuk menentukan target atau hasil yang diinginkan. *Supervised learning* termasuk metode yang menggunakan data yang sudah diberi label untuk memahami pola dan perilaku dalam berbagai aplikasi. *Supervised learning* membantu dalam pembentukan model yang merepresentasikan ciri-ciri distribusi perilaku dari data yang ada. Terbagi menjadi klasifikasi dan regresi, di mana klasifikasi menangani output kategorikal (contohnya warna atau kondisi kesehatan), sementara regresi menangani nilai kontinu (seperti nilai uang atau berat). Berbagai algoritma populer digunakan dalam supervised learning, termasuk *Back-propagation*, *Linear regression*, *Random Forest*, *Support Vector Regression (SVR)*, *Naive Bayesian*, *Metode Rocchio*, *Decision Tree*, *k-Nearest Neighbor*, *Neural Network*, *Logistic Regression*, *Gaussian Mixture models*, *Hidden Markov Models*, *Kernel Regression*, *Deep neural networks*, *Deep belief networks*, *PCA*, dan *Kernel Perceptron*. Ada pula algoritma khusus untuk klasifikasi seperti *Support Vector Machines (SVM)*, *Normal Bayesian Classifier (NBC)*, *K-Nearest Neighbor (KNN)*, *Trees Gradient Boosted (GBT)*, *Random Trees (RT)*, dan *Artificial Neural Networks (ANN)* [14].

b. *Unsupervised Learning*

Unsupervised learning adalah teknik di mana mesin memproses data yang tidak memiliki label atau petunjuk sebelumnya. Berbeda dengan *supervised learning*, *unsupervised learning* tidak memiliki panduan awal. Dalam *unsupervised learning*, sistem menerima sejumlah sampel input tanpa output yang diberikan. Di sini, fokusnya adalah pada mengidentifikasi pola atau struktur yang tersembunyi dalam data yang tidak memiliki label. Jenis pembelajaran ini terbagi menjadi clustering, di mana tujuannya adalah untuk menemukan kelompok dalam data seperti dalam mengelompokkan pelanggan berdasarkan perilaku belanja, dan asosiasi, yang berkaitan dengan menemukan aturan yang menjelaskan pola umum dalam data seperti kecenderungan orang membeli item A juga membeli item B. Terdapat beberapa algoritma populer dalam *unsupervised learning* seperti *k-means*, *Apriori*, *Independent Subspace Analysis (ISA)*, dan *DBSCAN* [14].

Cara kerja machine learning bervariasi tergantung pada teknik atau metode pembelajaran yang digunakan. Namun, prinsip dasarnya tetap sama, termasuk pengumpulan

data, analisis data, pemilihan model atau teknik, pelatihan model yang dipilih, dan evaluasi hasil *machine learning*. Selain algoritma, tingkat akurasi yang diperoleh juga sangat dipengaruhi oleh kualitas dataset yang digunakan. Semakin baik kualitas datasetnya, kemungkinan besar tingkat akurasi hasilnya akan lebih tinggi. Sebaliknya, jika dataset memiliki kualitas rendah, akurasi model cenderung lebih rendah [14].

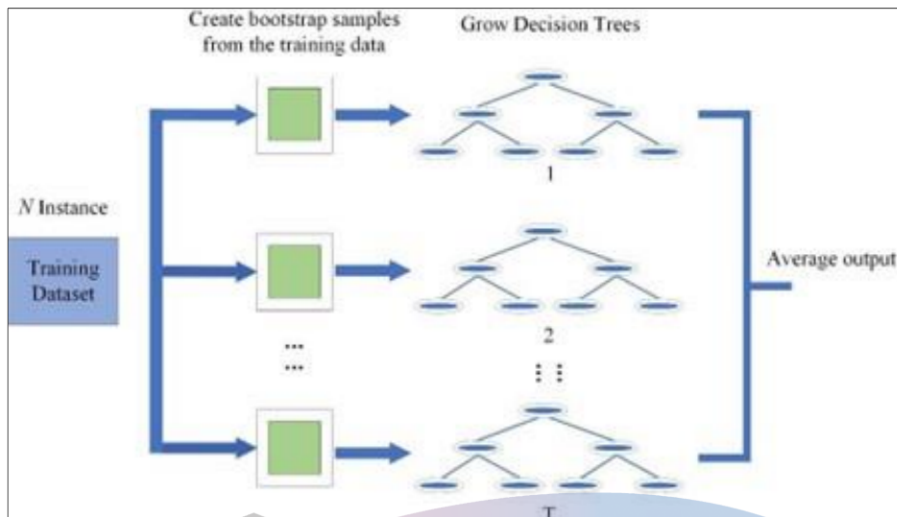
2.5. Random Forest

Random Forest (RF) atau *random decision forests* adalah metode pembelajaran *ensemble* untuk klasifikasi, regresi, dan tugas lainnya. *Random Forest* beroperasi dengan membangun beberapa *decision trees* pada waktu pelatihan yang berbeda, dan mengeluarkan kelas yang mewakili mode kelas (*classification*) atau prediksi (*regression*) dari masing-masing pohon. Algoritma *Random Forest* menggabungkan teknik *classification and regression trees (CARTs)* yang telah berkembang. Setiap *CART* dibangun dengan menggunakan subset acak dari fitur-fitur yang ada. Untuk model pengklasifikasi *Random Forest*, parameter utamanya adalah jumlah pohon keputusan dan jumlah fitur yang digunakan pada setiap simpul saat pohon tumbuh. Selama pelatihan model, jumlah pohon keputusan ditentukan terlebih dahulu. Lebih banyak pohon umumnya lebih baik, tetapi memerlukan waktu lebih lama untuk dihitung. Jumlah fitur yang lebih rendah menyebabkan pengurangan varians yang lebih besar. Jumlah fitur dapat didefinisikan dengan menggunakan rumus empiris $N_f = \sqrt{M}$, dimana M menunjukkan jumlah total fitur [15].

Random Forest dapat diterapkan untuk masalah klasifikasi maupun regresi, tergantung pada jenis pohon yang digunakan apakah itu pohon klasifikasi atau regresi. Skema model regresi *Random Forest* ditunjukkan pada Gambar 2.2. Dengan asumsi bahwa model tersebut terdiri T pohon regresi untuk melakukan prediksi regresi, hasil akhir dari model regresi *Random Forest* adalah sebagai berikut [15]:

$$H(x) = \frac{1}{T} \sum_{i=1}^T h_i(x) \quad (1)$$

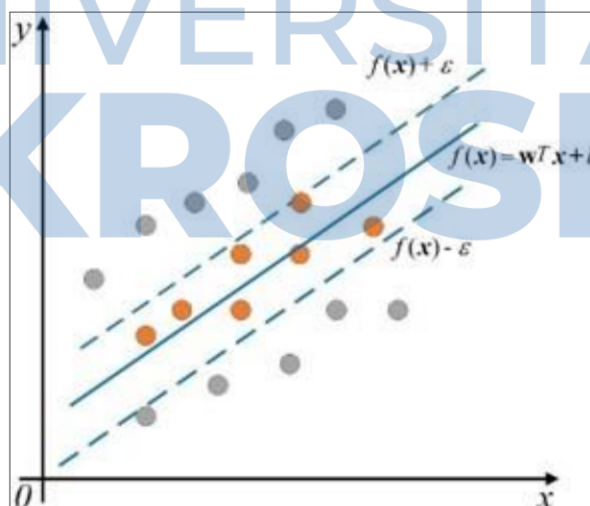
Dalam hal ini, T adalah jumlah pohon regresi, dan $h_i(x)$ adalah *output* dari pohon regresi ke-i pada sampel x. Dengan demikian, prediksi dari *Random Forest* adalah nilai rata-rata dari semua prediksi pohon tersebut.



Gambar 2.2 Diagram Skema *Random Forest Regression (RFR)* [15]

2.6. *Support Vector Regression (SVR)*

Support Vector Regression (SVR) adalah konsep yang berasal dari teori machine learning, yang sebelumnya digunakan untuk menyelesaikan masalah klasifikasi dengan *Support Vector Machine (SVM)*. Dalam konteks yang lebih sederhana, *SVM* digunakan untuk mengkategorikan data, sementara *SVR* digunakan untuk membuat prediksi numerik atau regresi. Tujuan utama *SVR* adalah menemukan fungsi yang sesuai dengan data input dengan sejumlah kesalahan seminimal mungkin [16]. Dengan kata lain, *SVR* bertujuan untuk memberikan prediksi yang sangat akurat dengan tingkat kesalahan sekecil mungkin saat menganalisis data regresi. Diagram skema *SVR* ditunjukkan pada Gambar 2.3.



Gambar 2.3 Diagram Skema *Support Vector Regression (SVR)* [15]

Adapun fungsi regresi dari model *SVR* dapat dituliskan sebagai berikut [17].

$$f(x) = w \varphi(x) + b \quad (2)$$

Keterangan :

w : Vektor pembobot

$\varphi(x)$: Fungsi yang memetakan x dalam suatu dimensi

b : Bias

Dalam menangani masalah yang tidak *linier* menggunakan algoritma *SVR*, kita menggunakan fungsi yang disebut *Kernel*. *Kernel* dalam *data mining* merupakan sebuah fungsi *inner product* (perkalian dalam) yang beroperasi di dalam ruang fitur [18]. Fungsinya adalah untuk memungkinkan penyelesaian masalah yang rumit dalam dimensi yang lebih tinggi tanpa harus secara langsung menghitung pemetaan yang detail. Beberapa jenis fungsi *Kernel* yang umum digunakan, yaitu [17]:

1. *Kernel Radial Basic Function (RBF)*

Kernel Radial Basis Function (RBF) banyak dipilih karena kemampuannya dalam memahami hubungan yang rumit pada data *non-linier*. *Kernel* ini cocok untuk data yang tidak beraturan dan terkenal akan ketepatan prediksinya dalam beragam aplikasi seperti prediksi nilai tukar [19], harga karbon [20], dan stabilitas dinding penahan kantilever [21]. *Kernel RBF* sangat efektif dalam menangkap pola *non-linier* dan telah menunjukkan ketepatan prediksi yang tinggi dalam berbagai penelitian, bahkan mencapai akurasi 95.94% dalam prediksi nilai tukar [19].

$$K(x_i, x_j) = \exp(-\gamma(x_i, x_j)^2) \quad (3)$$

Di sini, γ merupakan *parameter* bebas yang menentukan seberapa besar pengaruh dua titik terhadap satu sama lain. *RBF* berkembang ke arah dimensi yang tak terbatas karena mengalami pertumbuhan eksponensial yang cepat [19].

2. *Kernel Linear*

Kernel linier merupakan yang paling dasar di antara semua jenis *Kernel*. Dalam penggunaannya, tidak ada proyeksi data ke dimensi yang lebih tinggi secara teknis. *Kernel* ini hanya melibatkan hasil perkalian antara x dan y dengan nilai konstan c, tanpa penambahan dimensi baru. Keuntungan dari *Kernel Linear* adalah *Kernel* ini sangat sederhana dan hanya memiliki *parameter* konstan c sebagai satu-satunya komponen [19].

$$K(x_i, x_j) = x_i \cdot x_j + c \quad (4)$$

3. *Kernel Polynomial*

Berbeda dengan *Kernel Linear*, *Kernel* polinomial melibatkan perkalian hasil dari representasi dalam ruang dimensi yang lebih tinggi. Formula *Kernel* polinomial dapat

dinyatakan sebagai

$$K(x_i, x_j) = (\alpha x_i, x_j c)^d \quad (5)$$

di mana terdapat tiga *parameter*: α , c , dan d . Derajat yang paling umum (d) yang digunakan adalah 2 karena derajat yang lebih besar dapat menyebabkan overfitting [19].

4. Kernel Sigmoid

Kernel sigmoid digunakan untuk menangkap hubungan *non-Linear* dan cocok untuk data dengan pola *non-Linear*. *Kernel* ini telah digunakan dalam *SVR* untuk aplikasi seperti prediksi harga saham [22].

$$K(x_i, x_j) = \tanh(y(x_i, x_j) + \theta) \quad (6)$$

Disini, y adalah kemiringan dan θ adalah konstanta intersep [22].

Pemilihan *Kernel* dalam *SVR* tergantung pada sifat data dan kompleksitas hubungan dalam data. *Kernel RBF* sering dipilih karena kemampuannya dalam menangkap pola *non-Linear* dan akurasi prediksi yang tinggi dalam berbagai aplikasi.

2.7. Hyperparameter Tuning

Hyperparameter tuning adalah proses pemilihan nilai optimal untuk *hyperparameter* algoritma *machine learning* untuk meningkatkan kinerjanya. *Hyperparameter* adalah *parameter* yang tidak dipelajari dari data, melainkan ditetapkan oleh pengguna sebelum melatih model. *Hyperparameter tuning* ini sangat penting karena dapat secara signifikan mempengaruhi kemampuan model untuk menggeneralisasi dan membuat prediksi yang akurat. Tujuan penyetelan *hyperparameter* adalah untuk menemukan kombinasi terbaik dari nilai *hyperparameter* yang memaksimalkan kinerja model pada set validasi. Ada tiga metode dalam penyetelan *hyperparameter tuning* [23]:

a. Grid Search

Grid Search merupakan metode penentuan kombinasi model dan *hyperparameter* dengan melakukan uji coba satu persatu kombinasi model dan melakukan validasi pada setiap kombinasi [24]. Dalam prosesnya, *Grid Search* mencoba semua kemungkinan kombinasi nilai *hyperparameter*. Metode ini dapat memakan waktu lama, terutama untuk model dengan banyak *hyperparameter* [23].

b. Random Search

Random Search adalah metode yang mirip dengan *Grid Search*, tetapi nilai *hyperparameter* dipilih secara acak. Metode ini lebih cepat daripada *Grid Search*, tetapi hasil yang diperoleh mungkin tidak optimal. Metode ini bisa efektif jika ruang pencarian besar, tetapi tidak

menjamin bahwa nilai *hyperparameter* optimal akan ditemukan [23].

c. *Bayesian optimization*

Bayesian optimization adalah metode lain yang menggunakan model probabilistik untuk memandu pencarian *hyperparameter* optimal. Metode ini menyeimbangkan eksplorasi dan eksploitasi untuk secara efisien menemukan konfigurasi terbaik [23].

2.8. K-Fold Validation

K-fold cross validation adalah suatu metode yang dipakai untuk mengetahui rata-rata keberhasilan dari suatu sistem dengan cara melakukan iterasi atau perulangan pada atribut masukan yang diacak sehingga didapatkan sistem yang teruji untuk beberapa atribut input yang acak [25]. *K-fold cross validation* merupakan salah satu jenis CV yang berdasarkan pada ukuran dataset. Salah satu contoh k-fold cross-validation yang banyak digunakan ialah 10 fold cross-validation dikarenakan ada sebuah penelitian yang menunjukkan bahwa 10 fold cross-validation dapat memberikan perkiraan akurasi yang kurang bias dibandingkan menggunakan model lainnya.

Kinerja dari K-fold validation yaitu sebagai berikut [25]:

1. Total data dibagi menjadi N bagian.
2. Fold ke-1 pada bagian ke-1 merupakan data uji atau testing dan sisanya merupakan data latih. Kemudian, berdasarkan bagian data tersebut yaitu menghitung nilai akurasi pada kesamaan atau perkiraan hasil pengukuran dengan data sebenarnya berdasarkan banyaknya data.
3. Fold ke-2 pada bagian ke-2 merupakan data uji atau testing dan sisanya merupakan data latih. Kemudian menghitung nilai akurasi berdasarkan data sebenarnya.
4. Demikian seterusnya proses tersebut sampai mencapai fold ke-k. Setelah itu, menghitung rata-rata akurasi dari k buah akurasi di atas. Kemudian rata-rata akurasi tersebut menjadi akurasi terakhir.

2.9. Evaluasi Model

Evaluasi model adalah proses penting dalam *machine learning* yang bertujuan untuk mengukur kinerja dan akurasi dari model yang telah dibangun. Evaluasi model memungkinkan untuk memahami sejauh mana model dapat melakukan tugas yang diinginkan, seperti klasifikasi, regresi, atau tugas lainnya. Terdapat beberapa metrik dan teknik yang dapat digunakan dalam evaluasi model, tergantung pada jenis tugas yang dihadapi. Dalam evaluasi model sering menggunakan metrik seperti *Mean Absolute Error*

(MAE), Mean Squared Error (MSE), dan Root Mean Squared Error (RMSE) untuk mengukur seberapa baik model memprediksi. Evaluasi model yang baik adalah langkah penting dalam mengembangkan model yang akurat dan relevan dengan tujuan yang diinginkan. Selain itu, pemilihan metrik yang sesuai dengan konteks masalah adalah kunci dalam memahami performa model dan membuat keputusan yang tepat berdasarkan hasil prediksi.

Untuk mengukur kinerja algoritma prediksi dengan mengevaluasi perbandingan antara hasil prediksi dan hasil aktualnya dapat dilakukan dengan cara berikut.

1. Mean Absolute Error (MAE)

Mean Absolute Error (MAE) adalah metrik yang mengukur kesalahan rata-rata dalam prediksi dengan menghitung nilai mutlak positif dari selisih antara prediksi dan nilai aktual untuk setiap data [26]. Semakin kecil nilai MAE (Mean Absolute Error), maka semakin baik model tersebut dalam melakukan perkiraan (prediksi) [27]. Persamaan MAE adalah sebagai berikut [28]:

$$MAE = \frac{1}{n} \sum_{i=1}^n |\hat{Y}_i - Y_i| \quad (7)$$

Keterangan:

MAE : Mean Absolute Error

n : Jumlah total data atau observasi

Y_i : Nilai aktual dari observasi ke- i

\hat{Y}_i : Nilai prediksi dari observasi ke- i

Σ : Penjumlahan dari kesalahan absolut pada seluruh data

2. Mean Square Error (MSE)

Mean Square Error (MSE) adalah metrik evaluasi yang mengukur kesalahan peramalan dengan mengkuadratkan selisih antara nilai prediksi dan nilai aktual untuk setiap data [26]. Semakin kecil nilai MSE maka ramalan semakin aktual [29]. Persamaan MSE adalah sebagai berikut [30]:

$$MSE = \frac{1}{n} \sum_{i=1}^n (Y_i - \hat{Y}_i)^2 \quad (8)$$

Keterangan:

MSE : Mean Square Error

n : Jumlah total data atau observasi

Y_i : Nilai aktual dari observasi ke- i

\hat{Y}_i : Nilai prediksi dari observasi ke- i

Σ : Penjumlahan dari kesalahan absolut pada seluruh data

3. Mean Absolute Percentage Error (MAPE)

Mean Absolute Percentage Error (MAPE) adalah metrik evaluasi yang mengukur kesalahan relatif dalam prediksi dengan menghitung presentase penyimpangan antara nilai aktual dan nilai prediksi. MAPE mengukur sejauh mana prediksi benar-benar mendekati nilai aktual dalam bentuk persentase [26]. Persamaan MAPE adalah sebagai berikut [30]:

$$MAPE = \frac{100\%}{n} \sum_{i=1}^n \left| \frac{Y_i - \hat{Y}_i}{Y_i} \right| \quad (9)$$

Keterangan:

MAPE : Mean Absolute Percentage Error

n : Jumlah total data atau observasi

Y_i : Nilai aktual dari observasi ke- i

\hat{Y}_i : Nilai prediksi dari observasi ke- i

Σ : Penjumlahan dari kesalahan absolut pada seluruh data

Nilai MAPE dapat diartikan ke dalam 4 kategori, yaitu [31]:

Tabel 2.1 Rentang Nilai MAPE

Rentang MAPE	Arti Nilai
< 10%	Kemampuan Model Peramalan Sangat Baik
10 - 20%	Kemampuan Model Peramalan Baik
20 - 50%	Kemampuan Model Peramalan Layak
> 50 %	Kemampuan Model Peramalan Buruk

4. Root Mean Squared Error (RMSE)

Root Mean Squared Error (RMSE) adalah metrik evaluasi yang umum digunakan untuk mengukur kesalahan dalam model prediksi data kuantitatif. RMSE digunakan untuk mengukur sebaran simpangan titik data dari garis regresi linier atau seberapa dekat data sekitar garis regresi linier [32]. Persamaan RMSE adalah sebagai berikut [33]:

$$RMSE = \sqrt{\sum_{i=1}^n \frac{(Y_i - \hat{Y}_i)^2}{n}} \quad (10)$$

Keterangan:

RMSE : Root Mean Squared Error

n : Jumlah total data atau observasi

Y_i : Nilai aktual dari observasi ke- i

\hat{Y}_i : Nilai prediksi dari observasi ke- i

Σ : Penjumlahan dari kesalahan absolut pada seluruh data

Nilai *RMSE* yang rendah atau nilainya mendekati angka 0 menunjukkan bahwa hasil peramalan sesuai dengan data aktual. *RMSE* juga dapat berupa angka negatif. Semakin kecil hasil *RMSE* semakin dekat pula prediksinya [31].

2.10. Literature Review

Dalam penyusunan tugas akhir ini, terdapat beberapa studi sebelumnya yang telah dilakukan oleh peneliti lain. Model dan data yang digunakan dalam penelitian sebelumnya, serta hasil-hasil yang diperoleh, akan dijadikan sebagai sumber referensi penting dalam penyusunan tugas akhir ini. Rincian singkat mengenai penelitian yang telah dilakukan dapat dilihat dalam Tabel 2.2.

Tabel 2.2 Literature Review

No.	Referensi	Model	Judul	Area Aplikasi	Hasil
1.	[6]	<i>Random Forest</i> dan <i>Decision Tree</i>	Komparasi Algoritma <i>Random Forest</i> dan <i>Decision Tree</i> untuk Memprediksi Keberhasilan <i>Immunotherapy</i>	Kesehatan	Hasil penelitian menunjukkan bahwa model <i>Decision Tree</i> memiliki tingkat akurasi 84,4 %. Sementara model <i>Random Forest</i> memiliki akurasi 85,5% dan lebih unggul dalam prediksinya dibandingkan dengan <i>Decision Tree</i> .
2.	[34]	<i>Support Vector Regression (SVR)</i>	<i>Prediction And Analysis Of Population Aging In Eight Ethnic Provinces Based On Machine Learning (SVR)</i>	Demografi	Penelitian ini memakai model <i>SVR</i> untuk prediksi pertumbuhan populasi di atas 65 tahun di delapan provinsi etnis Tiongkok selama 9 tahun. Hasilnya menunjukkan kenaikan konsisten di seluruh provinsi, mengkonfirmasi prediksi dan tren pertambahan usia. Evaluasi model memperlihatkan

					akurasi tinggi.
3.	[7]	<i>Support Vector Regression (SVR)</i>	<i>Analysis of Using Support Vector Regression (SVR) Algorithm to Predict The Occurrence of Sea Tides in Tanjung Medang, Riau by Saas Storage in Cloud Computing</i>	Oceanografi	Penelitian ini menggunakan algoritma <i>Support Vector Regression (SVR)</i> untuk memprediksi pasang surut laut di Tanjung Medang, Riau, dengan akurasi tinggi (89%) dan <i>RMSE</i> sekitar 0,45.
4.	[35]	<i>Random Forest</i>	<i>Effective Macrosomia Prediction Using Random Forest Algorithm</i>	Kesehatan	Hasil menunjukkan bahwa model <i>Random Forest</i> lebih baik dengan tingkat sensitif (91,7%) dan spesifik (91,7%) daripada model regresi logistik (sensitivitas: 56,2%, spesifisitas: 82,6%) serta ultrasonografi (sensitivitas: 29,6%, spesifisitas: 97,5%) dalam memprediksi makrosomia. Hal ini mengindikasikan bahwa model <i>Random Forest</i> secara akurat memprediksi makrosomia.
5.	[36]	<i>Support Vector Regression (SVR)</i>	<i>A Support Vector Regression (SVR)-Based Method for Dynamic Load Identification Using Heterogeneous Responses Under Interval Uncertainties</i>	Infrastruktur	Metode <i>Support Vector Regression (SVR)</i> berhasil menghasilkan rekonstruksi beban dinamis yang tidak pasti dengan tingkat akurasi tinggi. Hasilnya menunjukkan metode <i>SVR</i> valid dan efisien dalam mengidentifikasi interval beban dinamis dengan baik.

Hasil tinjauan literatur menunjukkan bahwa penggunaan berbagai model *Machine Learning*, seperti *Random Forest* dan *Support Vector Regression (SVR)*, telah terbukti berhasil dalam berbagai konteks prediksi. *Random Forest* menonjol dalam memprediksi keberhasilan terapi dan makrosomia, dengan tingkat sensitivitas dan spesifisitas yang lebih

tinggi daripada model-model lainnya. Sementara itu, *SVR* mampu memprediksi pertumbuhan populasi, pasang surut laut, dan identifikasi beban dinamis dengan akurasi yang tinggi, bahkan dalam situasi ketidakpastian interval. Dengan hasil evaluasi yang kuat, penggunaan model-model ini dapat dianggap sebagai solusi yang dapat diandalkan dalam upaya prediksi di berbagai konteks.

