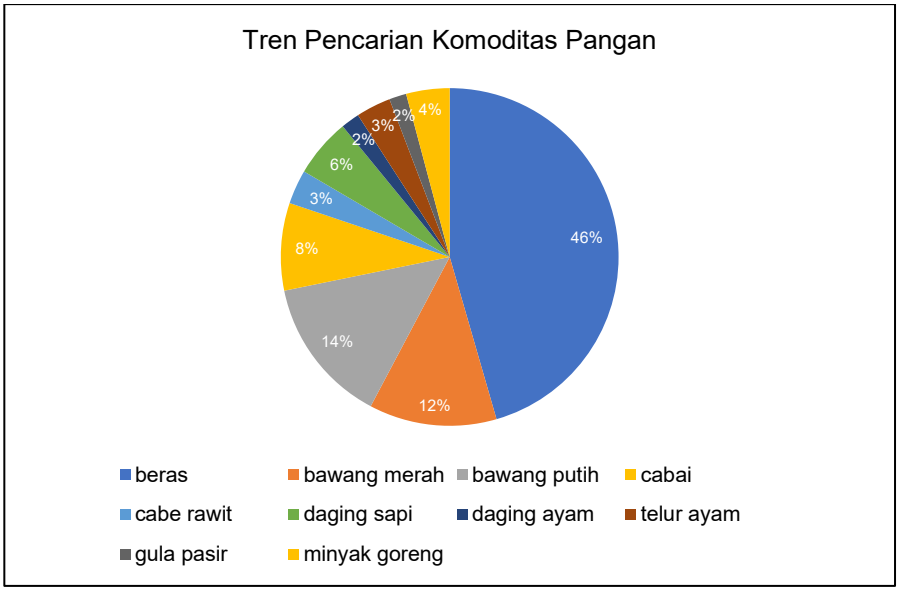
BAB I
PENDAHULUAN

1.1 Latar Belakang

Ketahanan pangan merupakan salah satu pilar utama dalam mewujudkan stabilitas sosial, politik, dan ekonomi suatu negara. Di Indonesia, sektor pangan memiliki peranan strategis karena sebagian besar penduduknya masih sangat bergantung pada komoditas pangan pokok sebagai sumber konsumsi utama. Fluktuasi harga yang tajam pada komoditas pangan dapat menyebabkan tekanan terhadap daya beli masyarakat, terutama kelompok ekonomi rentan, serta memperburuk ketimpangan sosial. Terlebih lagi, ketidakstabilan harga pangan juga berkontribusi terhadap laju inflasi nasional dan dapat memperlemah efektivitas kebijakan moneter dan fiskal. Badan Pangan Nasional menunjukkan bahwa pada April 2025, dari total inflasi tahunan sebesar 1,17%, komponen pangan menyumbang andil sebesar 0,64%. Oleh karena itu, pengelolaan harga pangan secara sistematis dan berbasis data merupakan kebutuhan mendesak dalam memperkuat fondasi ketahanan nasional. Terdapat beberapa komoditas pangan yang memiliki kontribusi signifikan dalam pembentukan angka inflasi (strategis) menurut PIHPS (Pusat Informasi Harga Pangan Strategis) Nasional yaitu beras, bawang merah, bawang putih, cabai merah, cabai rawit, daging sapi, daging ayam ras, telur ayam ras, gula pasir, dan minyak goreng (Utomo, 2020).

Fenomena ini bahkan lebih terasa di tingkat daerah. Salah satunya di Provinsi Sulawesi Selatan, kelompok "Makanan, Minuman, dan Tembakau" menjadi pendorong utama inflasi pada Juli 2024 dengan andil mencapai 0,96% terhadap total inflasi provinsi sebesar 1,74%. Komoditas beras secara spesifik diidentifikasi sebagai pemicu utama inflasi di wilayah tersebut. Sebagai wilayah agraris sekaligus pusat distribusi pangan di Kawasan Timur Indonesia, Sulawesi Selatan menghadapi tantangan serius dalam menjaga stabilitas harga pangan, khususnya pada masa-masa tertentu seperti hari besar keagamaan, musim panen, dan gangguan distribusi. Dalam konteks ini, perhatian masyarakat terhadap harga pangan tertentu dapat diamati melalui data pencarian daring. Perhatian masyarakat terhadap harga pangan tertentu juga menunjukkan tingkat sensitivitas yang tinggi. Berdasarkan data dari Google Trends, lima komoditas pangan strategis yang paling sering dicari oleh masyarakat Sulawesi Selatan adalah beras, bawang putih, bawang merah, cabai merah, dan daging sapi. Tingginya frekuensi pencarian ini menunjukkan urgensi komoditas tersebut dalam kehidupan sehari-hari serta tingginya sensitivitas harga di mata konsumen.



Gambar 1. Tren Pencarian Komoditas Pangan pada Google Trends
Sumber: Pengolahan Data (2025)

Fluktuasi harga pada komoditas pangan tidak hanya berdampak pada konsumen akhir, tetapi juga menyulitkan pelaku usaha, petani, dan pengambil kebijakan dalam mengambil keputusan berbasis data. Ketidakstabilan harga dapat menurunkan daya beli masyarakat, meningkatkan laju inflasi, serta mengganggu kelancaran distribusi dan rantai pasok. Tingginya tingkat fluktuasi harga ini dapat dilihat secara detail pada Tabel 3 (hal 40). Sebagai contoh, harga cabai merah melonjak lebih dari 88% dalam lima bulan, dari Rp 26.974 pada Februari 2022 menjadi Rp 50.690 pada Juli 2022. Lonjakan tajam ini langsung membebani pengeluaran rumah tangga karena cabai merupakan kebutuhan harian. Sementara itu, harga bawang merah sempat naik lebih dari 130%, dari Rp 25.498 pada Januari 2022 ke Rp 60.429 pada Juli 2022, yang memicu keresahan konsumen dan menekan daya beli. Di sisi lain, harga beras menunjukkan tren kenaikan bertahap, dari Rp 9.933 pada Januari 2022 menjadi Rp 14.148 pada Maret 2024 (naik sekitar 42%). Fluktuasi harga, baik kenaikan yang berisiko meningkatkan inflasi maupun penurunan tajam yang merugikan petani dan peternak, menciptakan ketidakpastian pasar. Ketidakpastian ini menyulitkan pemerintah dalam merancang kebijakan stabilisasi harga yang tepat waktu dan berbasis data. Oleh karena itu, diperlukan pendekatan prediktif yang dapat menghasilkan estimasi harga yang akurat dan berkelanjutan. (Komansilan dkk., 2024).

Kebutuhan akan alat prediksi yang akurat semakin mendesak tercermin dari intensitas intervensi pasar yang dilakukan oleh Pemerintah Provinsi Sulawesi Selatan. Pemprov Sulsel secara rutin meluncurkan Gerakan Pangan Murah (GPM) dengan tujuan utama untuk stabilisasi pasokan dan harga sekaligus mengendalikan inflasi. Sepanjang tahun 2024 saja, GPM telah dilaksanakan sebanyak 622 kali, dan upaya ini dilanjutkan pada tahun 2025. Frekuensi intervensi yang sangat tinggi ini

menunjukkan bahwa mekanisme pengendalian dan peramalan harga yang digunakan saat ini belum cukup presisi atau cenderung bersifat reaktif. Hal ini menciptakan kebutuhan mendesak untuk mengadopsi model peramalan yang lebih akurat dan adaptif guna mendukung kebijakan intervensi pasar yang *prediktif* dan efisien.

Namun demikian, berbagai metode konvensional yang banyak digunakan dalam studi sebelumnya menunjukkan keterbatasan dalam merespons kompleksitas data harga pangan. Pendekatan sederhana seperti Regresi Linear, yang diterapkan oleh (Arinal & Azhari, 2023) untuk memprediksi harga beras, menghasilkan nilai RMSE sebesar 337,996, yang mengindikasikan adanya galat signifikan karena ketidakmampuannya menangkap pola non-linear. Bahkan model yang dirancang untuk data musiman seperti *Triple Exponential Smoothing Holt-Winters* juga menunjukkan keterbatasan serius, di mana sebuah studi peramalan persediaan beras mencatat nilai MAPE yang sangat tinggi mencapai 26,4% (Amri dkk., 2025). Sementara itu, model ARIMA yang populer untuk data deret waktu juga menghadapi tantangan serupa. Sebuah studi pada peramalan harga bawang merah dengan ARIMA menghasilkan MAPE sebesar 19,81%, dengan penulis menyimpulkan bahwa model ini kesulitan menangani data non-linear sehingga menghasilkan galat yang besar. Kegagalan-kegagalan ini, mulai dari asumsi linearitas yang kaku hingga ketidakmampuan menangani pola musiman yang kompleks, mengindikasikan pentingnya penggunaan pendekatan yang lebih adaptif seperti *machine learning*, yang mampu mengenali pola-pola dinamis dari berbagai sumber data (Rahayu dkk., 2023).

Komponen selanjutnya dari konsep peramalan adalah pemilihan metode yang mampu menghasilkan prediksi dengan akurasi tinggi dan waktu komputasi yang efisien. Dalam era data saat ini, metode berbasis *machine learning* semakin luas digunakan karena kemampuannya dalam menangani volume data besar dan pola hubungan non-linear yang kompleks. Penelitian (Erdianto, 2023) menunjukkan bahwa model *Random Forest* dan XGBoost secara konsisten mengungguli regresi linier dan *regression tree* dalam memprediksi harga sepuluh komoditas pangan strategis di Indonesia, dengan lima komoditas terbaik dihasilkan oleh *Random Forest* dan tiga lainnya oleh XGBoost. Misalnya, prediksi harga telur ayam ras menggunakan *Random Forest* menghasilkan MAPE hanya sebesar 2,46%, sementara prediksi cabai merah dengan XGBoost mencapai MAPE 16,63%, lebih baik dibandingkan metode lainnya pada komoditas tersebut (Yang dkk., 2023).

Kinerja unggul XGBoost juga dibuktikan dalam studi (Salsabila & Widhastika, 2024), di mana XGBoost diterapkan untuk memprediksi harga daging sapi murni di Papua Barat dan menunjukkan akurasi tinggi dengan nilai MAPE hanya 1,59%, jauh lebih unggul dari ARIMA yang kesulitan menangani data non-stasioner dan volatil. Selain itu, dalam konteks klasifikasi prediktif pada kasus gagal panen gabah, *Random Forest* dan XGBoost menunjukkan akurasi 77% dan 70% secara berurutan, menjadi dua metode terbaik dari delapan model yang diuji (Nizami dkk., 2023). Hal ini mengonfirmasi kelebihan utama kedua metode tersebut dalam menangani ketidakseimbangan data, *noise*, dan hubungan non-*linear* antar variabel. *Random*

Forest dikenal karena stabilitasnya dalam menghindari *overfitting* melalui mekanisme *bagging*, sementara XGBoost unggul dalam mengoptimalkan akurasi melalui *boosting* bertahap dan regularisasi yang efektif. Oleh karena itu, penelitian ini memilih *Random Forest* dan XGBoost sebagai dua pendekatan utama dalam pemodelan harga pangan, seiring dengan keberhasilan dan keandalannya yang terbukti dalam berbagai studi serupa (Yang dkk., 2023).

Dengan menggunakan pendekatan berbasis algoritma *Machine Learning*, penelitian ini bertujuan untuk optimasi sistem prediksi harga lima komoditas pangan strategis di Sulawesi Selatan. Berdasarkan hal tersebut peneliti melakukan penelitian berjudul "**Optimasi Peramalan Harga Pangan Di Sulawesi Selatan Menggunakan *Machine Learning* (Studi Perbandingan *Random Forest* Dan *Extreme Gradient Boosting*)**". Penelitian ini tidak hanya memberikan kontribusi akademik dalam bidang *forecasting* berbasis *machine learning*, tetapi juga diharapkan dapat memberikan dampak praktis sebagai rekomendasi kebijakan bagi lembaga pemerintah, pelaku pasar serta masyarakat luas dalam mengantisipasi gejolak harga pangan secara lebih dini dan efisien.

1.2 Rumusan Masalah

Berdasarkan permasalahan yang telah dijabarkan, berikut merupakan rumusan masalah yang akan diteliti dalam penelitian ini:

a. Bagaimana mengembangkan model matematis berbasis algoritma *Machine Learning* untuk memprediksi harga komoditas pangan strategis di Sulawesi Selatan?
b. Apa perbandingan performa antara berbagai algoritma *Machine Learning* dan Konvensional dalam melakukan peramalan harga berdasarkan data historis komoditas pangan?
c. Metode Peramalan apa yang paling optimal dalam melakukan peramalan harga komoditas pangan Sulawesi Selatan berdasarkan hasil evaluasi model (RMSE, MAPE, dan R2)?
d. Bagaimana hasil peramalan harga pangan ini dapat digunakan sebagai dasar dalam mendukung pengambilan keputusan terkait dalam pengambilan kebijakan oleh pemerintah dan pelaku industri pangan?

1.3 Tujuan Penelitian

Berdasarkan rumusan masalah di atas, adapun tujuan penelitian berikut ini:

a. Mengembangkan model matematis berbasis algoritma *Machine Learning* untuk memprediksi harga komoditas pangan strategis di Sulawesi Selatan.
b. Membandingkan performa berbagai algoritma *Machine Learning* dan model konvensional dalam melakukan peramalan harga berdasarkan data historis harga pangan.
c. Mengidentifikasi model peramalan yang paling optimal dalam melakukan prediksi harga untuk masing-masing komoditas pangan strategis berdasarkan hasil evaluasi model (RMSE, MAPE, dan R2).
d. Menganalisis potensi pemanfaatan hasil peramalan harga pangan sebagai dasar dalam pengambilan keputusan oleh pemerintah dan pelaku industri pangan.

1.4 Manfaat Penelitian

Penelitian ini diharapkan dapat memberikan manfaat bagi berbagai pihak, diantaranya:

a. Bagi Petani dan Pelaku Usaha
   Memberikan prediksi harga yang akurat sebagai dasar pengambilan keputusan dalam perencanaan produksi, distribusi, dan strategi penjualan untuk mengurangi risiko kerugian akibat fluktuasi harga.
b. Bagi Akademisi
   Menambah literatur ilmiah dalam bidang *forecasting* dan *Machine Learning*, serta menjadi referensi bagi penelitian lanjutan terkait prediksi harga komoditas pangan di Sulawesi Selatan.
c. Bagi Pemerintah dan Instansi Terkait
   Memberikan informasi berbasis data sebagai dasar pengambilan kebijakan harga dan distribusi pangan strategis.
d. Bagi Penulis
   Menjadi sarana pengembangan kompetensi dalam analisis data dan penerapan *Machine Learning* untuk menyelesaikan permasalahan nyata di bidang ketahanan pangan.

1.5 Batasan Masalah

Untuk mencapai tujuan penelitian dengan baik, penulis memberikan batasan ruang lingkup permasalahan. Adapun batasan masalah pada penelitian ini adalah:

a. Komoditas yang dianalisis dibatasi pada 5 komoditas pangan strategis berdasarkan data dari Pusat Informasi Harga Pangan Strategis (PIHPS) Nasional
b. Data yang digunakan dibatasi pada data harga historis harian atau, yang diperoleh dari sumber terbuka seperti PIHPS, BPS, atau lembaga resmi lainnya, dalam rentang waktu tahun 2022-2024.
c. Metode peramalan dibatasi pada metode *Random Forest* dan *Extreme Gradient Boosting* (XGBoost), dan *Linear Regression*.
d. Faktor eksternal yang dimasukkan dalam pemodelan dibatasi pada variabel yang tersedia secara kuantitatif, seperti waktu, tren musiman, dan data historis, tanpa memasukkan faktor kualitatif seperti kebijakan pemerintah, iklim sosial-politik, atau spekulasi pasar.

1.6 Tinjauan Pustaka

1.6.1 Komoditas Pangan

Pangan merupakan kebutuhan dasar yang esensial bagi keberlangsungan hidup manusia dan menjadi elemen penting dalam pembangunan nasional. Ketahanan pangan suatu negara ditentukan oleh kemampuan negara tersebut dalam menjamin ketersediaan, keterjangkauan, dan keamanan pangan bagi seluruh penduduknya. Komoditas pangan tidak hanya dipandang sebagai barang konsumsi, tetapi juga sebagai bagian dari sistem sosial, budaya, dan ekonomi masyarakat. Di Indonesia, yang dikenal sebagai negara agraris, komoditas pangan memiliki peran strategis dalam menopang stabilitas ekonomi dan kesejahteraan masyarakat. Komoditas pangan juga menjadi salah satu komponen utama dalam pengendalian inflasi dan pengukuran indeks harga konsumen (IHK). Oleh karena itu, pemerintah secara aktif memantau dan mengelola pergerakan harga komoditas pangan guna menjaga kestabilan pasar dan daya beli masyarakat.

Menurut Pusat Informasi Harga Pangan Strategis (PIHPS), kriteria Barang Kebutuhan Pokok (BAPOK) yang ditetapkan pemerintah, terdapat sejumlah komoditas pangan yang dikategorikan sebagai komoditas strategis. Komoditas-komoditas ini memiliki tingkat konsumsi tinggi, peran penting dalam struktur pengeluaran rumah tangga, dan sensitivitas harga yang tinggi terhadap faktor musiman maupun non-musiman. Komoditas utama yang tergolong strategis antara lain adalah: beras, gula pasir, daging ayam ras, telur ayam ras, bawang merah, bawang putih, cabai merah, cabai rawit, daging sapi, dan minyak goreng.

Komoditas-komoditas tersebut memiliki karakteristik yang berbeda-beda dalam hal produksi, distribusi, dan pola konsumsi, sehingga pengelolaannya memerlukan perhatian khusus dari pemerintah maupun pelaku pasar. Selain sebagai kebutuhan domestik, sebagian dari komoditas ini juga berperan dalam kegiatan ekspor-impor dan menjadi indikator penting dalam ketahanan pangan nasional.
(Ilmandira dkk., 2022)

1.6.2 *Time Series Prediction*

Deret waktu adalah serangkaian pengamatan yang diukur secara berurutan sepanjang waktu. Pengukuran ini dapat dilakukan secara terus-menerus sepanjang waktu atau dilakukan pada serangkaian titik waktu yang terpisah. Berdasarkan konvensi, kedua jenis deret ini masing-masing disebut deret waktu berkelanjutan dan diskrit, meskipun variabel yang diukur dapat berupa diskrit atau kontinu dalam kedua kasus tersebut. Secara efektif, setiap pengamatan pada variabel yang diukur adalah pengamatan *bivariat* dengan waktu sebagai variabel kedua. Menurut (Chatfield, 2000) Tujuan utama analisis deret waktu adalah:

a. Deskripsi
   Untuk menggambarkan data menggunakan statistik ringkasan dan/atau metode grafis. Plot waktu dari data sangat berharga.
b. Pemodelan
   Untuk menemukan model statistik yang sesuai untuk menggambarkan proses pembuatan data. Model *univariat* untuk variabel tertentu hanya didasarkan pada nilai-nilai masa lalu dari variabel tersebut, sedangkan model *multivariat* untuk

variabel tertentu mungkin didasarkan tidak hanya pada nilai-nilai masa lalu dari variabel tersebut, tetapi juga pada nilai-nilai masa kini dan masa lalu dari variabel (prediktor) lainnya. Dalam kasus terakhir, variasi dalam satu deret dapat membantu menjelaskan variasi dalam deret lainnya. Tentu saja, semua model adalah perkiraan dan pembuatan model merupakan seni sekaligus sains.

c. Peramalan

Untuk memperkirakan nilai seri di masa mendatang. Sebagian besar penulis menggunakan istilah ‘peramalan’ dan ‘prediksi’ secara bergantian. Ada perbedaan yang jelas antara peramalan kondisi mapan, di mana memperkirakan masa depan akan sangat mirip dengan masa lalu, dan peramalan *What-if* di mana model *multivariat* digunakan untuk mengeksplorasi dampak perubahan variabel kebijakan.

d. Kontrol

Peramalan yang baik memungkinkan analis untuk mengambil tindakan guna mengendalikan proses tertentu, baik itu proses industri, ekonomi, bidang lainnya.

Sebelum memulai pemodelan dan peramalan data deret waktu, penting untuk terlebih dahulu mengeksplorasi data untuk mendapatkan pemahaman awal tentang karakteristiknya. Langkah ini sangat bermanfaat untuk proses pemodelan di kemudian hari. Plot waktu adalah alat utama yang digunakan, tetapi grafik lainnya serta statistik ringkasan juga bisa membantu dalam memahami pola data. Proses ini juga memungkinkan analis untuk membersihkan data dengan menghapus atau memperbaiki kesalahan. Pendekatan ini sering disebut sebagai Analisis Data Awal. Bagian ini juga memperkenalkan beberapa metode deskriptif sederhana yang khusus dirancang untuk menganalisis data deret waktu. Dua sumber utama variasi dalam banyak deret waktu adalah tren dan variasi musiman. Pengetahuan dasar tentang model deret waktu akan sangat membantu dalam memahami konsep-konsep ini, dan gagasan yang dijelaskan di sini akan memperdalam pemahaman tentang perilaku deret waktu.

Sebelum menjelaskan berbagai model, berikut gambaran umum analisis deret waktu, yang bertujuan memisahkan variasi dalam deret waktu menjadi beberapa komponen penting yaitu :

a. Seasonality (S)

Jenis variasi ini umumnya terjadi secara tahunan dan muncul pada banyak deret waktu, baik yang diukur mingguan, bulanan, atau triwulanan, ketika pola perilaku yang serupa terlihat pada waktu-waktu tertentu dalam setahun. Contohnya adalah pola penjualan es krim yang selalu tinggi di musim panas. Perlu diperhatikan bahwa jika deret waktu hanya diukur setiap tahun (sekali dalam setahun), maka tidak mungkin mengetahui apakah terdapat variasi musiman.

b. Trend (T)

Variasi ini muncul ketika suatu deret waktu menunjukkan pertumbuhan yang stabil atau penurunan yang berkelanjutan selama beberapa periode waktu berturut-turut. Contohnya adalah perilaku indeks harga ritel di Inggris yang meningkat setiap tahun selama bertahun-tahun. Tren dapat didefinisikan secara umum sebagai "perubahan jangka panjang dalam rata-rata", namun tidak ada definisi matematis yang sepenuhnya memadai. Persepsi tentang tren akan tergantung pada panjang deret waktu yang diamati. Selain itu, pemisahan variasi menjadi komponen musiman dan tren tidak selalu unik.

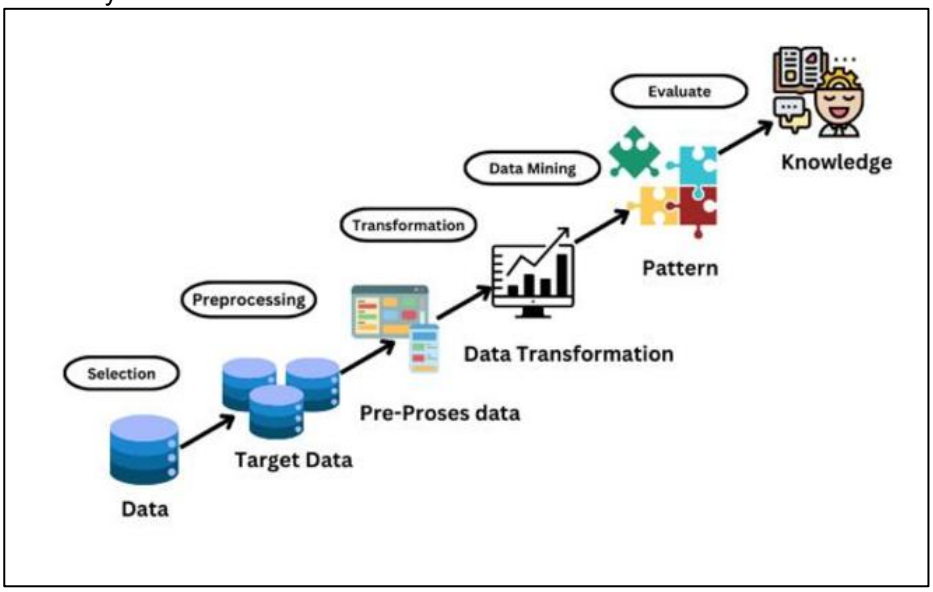
c. *Cycles (C)*
   Variasi Ini mencakup variasi siklis yang teratur dengan periode selain satu tahun. Contohnya termasuk siklus bisnis yang terjadi dalam jangka waktu lima tahun dan ritme harian (disebut variasi diurnal) dalam perilaku biologis makhluk hidup.
d. Fluktuasi tak teratur
   Istilah 'fluktuasi tak teratur' sering digunakan untuk menggambarkan variasi yang tersisa setelah tren, musiman, dan efek sistematis lainnya dihilangkan. Variasi ini bisa saja benar-benar acak sehingga tidak dapat diprediksi. Namun, mereka juga mungkin menunjukkan korelasi jangka pendek atau termasuk perubahan tiba-tiba yang terjadi satu kali.

1.6.3 KDD (*Knowledge Discovery In Database*)

Knowledge Discovery In Database (KDD) merupakan metode untuk memperoleh pengetahuan dari data base yang ada. Dalam data base terdapat tabel - tabel yang saling berhubungan / berelasi. Hasil pengetahuan yang diperoleh dalam proses tersebut dapat digunakan sebagai basis pengetahuan (*knowledge base*) untuk keperluan pengambilan keputusan. *Knowledge Discovery in Database* (KDD) dan *data mining* sering kali digunakan secara bergantian untuk menjelaskan proses penggalian informasi tersembunyi dalam suatu basis data yang besar. Sebenarnya kedua istilah tersebut memiliki konsep yang berbeda, tetapi berkaitan satu sama lain, dan salah satu tahapan dalam keseluruhan proses KDD adalah *data mining*. Proses KDD secara garis besar dapat dijelaskan sebagai berikut. (Mardi, 2017)

1. *Data Selection*
   Pemilihan (seleksi) data dari sekumpulan data operasional perlu dilakukan sebelum tahap penggalian informasi dalam *Knowledge Discovery in Database* (KDD) dimulai. Data hasil seleksi yang akan digunakan untuk proses *data mining*, disimpan dalam suatu berkas terpisah dari basis data operasional.
2. *Pre-processing / Cleaning*
   Sebelum proses *data mining* dapat dilaksanakan, perlu dilakukan proses *cleaning* pada data yang menjadi fokus *Knowledge Discovery in Database* (KDD). Proses *cleaning* mencakup antara lain membuang duplikasi data, memeriksa data yang inkonsisten, dan memperbaiki kesalahan pada data, seperti kesalahan cetak. Juga dilakukan proses *enrichment*, yaitu proses “memperkaya” data yang sudah ada dengan data atau informasi lain yang relevan dan diperlukan untuk *Knowledge Discovery in Database* (KDD), seperti data atau informasi eksternal lainnya yang diperlukan.
3. *Transformation*
   Coding adalah proses transformasi pada data yang telah dipilih, sehingga data tersebut sesuai untuk proses *data mining*. Proses *coding* dalam *Knowledge Discovery in Database* (KDD) merupakan proses kreatif dan sangat tergantung pada jenis atau pola informasi yang akan dicari dalam basis data.
4. *Data mining*
   Data mining adalah proses mencari pola atau informasi menarik dalam data terpilih dengan menggunakan teknik atau metode tertentu. Teknik-teknik, metode-metode, atau algoritma dalam *data mining* sangat bervariasi. Pemilihan metode atau algoritma yang tepat sangat bergantung pada tujuan dan proses *Knowledge Discovery in Database* (KDD) secara keseluruhan.
5. *Interpretation / Evaluation*
   Pola informasi yang dihasilkan dari proses *data mining* perlu ditampilkan dalam bentuk yang mudah dimengerti oleh pihak yang berkepentingan. Tahap ini

merupakan bagian dari proses *Knowledge Discovery in Database* (KDD) yang disebut *interpretation*. Tahap ini mencakup pemeriksaan apakah pola atau informasi yang ditemukan bertentangan dengan fakta atau hipotesis yang ada sebelumnya.



Gambar 2. Proses *Knowledge Discovery in Database Process* (KDD)
Sumber: (Buani, 2024)

1.6.4 *Data mining*

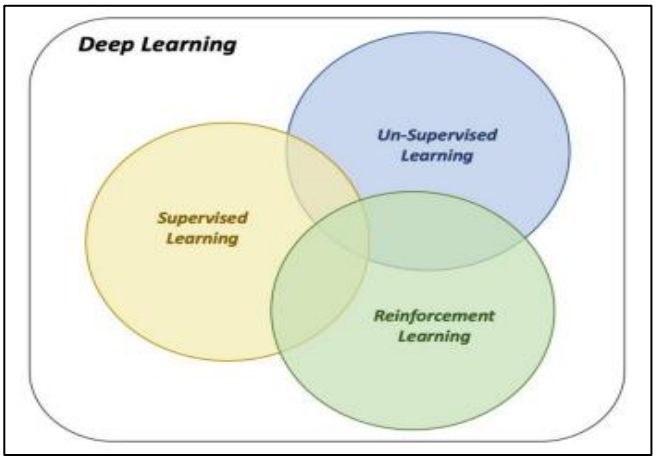
Data mining adalah proses menemukan hubungan baru yang mempunyai arti, pola dan kebiasaan dengan memilah-milah sebagian besar data yang disimpan dalam media penyimpanan dengan menggunakan teknologi pengenalan pola seperti teknik statistik dan matematika. *Data mining* merupakan gabungan dari beberapa disiplin ilmu yang menyatukan teknik dari pembelajaran mesin, pengenalan pola, statistik, *database*, dan visualisasi untuk penanganan permasalahan pengambilan informasi dari data base yang besar. *Data mining* menggunakan teknik statistik, matematika, kecerdasan buatan, dan *Machine Learning* untuk mengekstraksi dan mengidentifikasi informasi yang bermanfaat dan pengetahuan yang terkait dari berbagai *database* besar. Salah satu kesulitan untuk mendefinisikan *data mining* adalah kenyataan bahwa *data mining* mewarisi banyak aspek dan teknik dari bidang-bidang ilmu yang dulu sudah mapan terlebih dulu (Mardi, 2017).

1.6.5 *Machine Learning*

Machine Learning (ML) merupakan bidang studi yang fokus kepada desain dan analisis algoritma sehingga memungkinkan komputer untuk dapat belajar. Menurut (Zhang dkk., 2022), *Machine Learning* berisi sebuah algoritma yang bersifat *generic* (umum) di mana algoritma tersebut dapat menghasilkan sesuatu yang menarik atau bermanfaat dari sejumlah data tanpa harus menulis kode yang spesifik. Pada intinya, algoritma yang generik tersebut ketika diberikan sejumlah data maka ia dapat membangun sebuah aturan atau model atau inferensi dari data tersebut. Sebagai contoh sebuah algoritma untuk mengenali tulisan tangan dapat digunakan untuk mendeteksi email yang berisi spam dan bukan spam tanpa mengganti kode. Algoritma yang sama ketika diberikan data pelatihan yang berbeda menghasilkan logika klasifikasi yang berbeda.

1.6.6 *Deep Learning*

(Haris dkk., 2021) menyatakan bahwa *Deep learning* (DL) yang merupakan sebuah teknik berbasis jaringan saraf tiruan telah banyak digunakan dalam beberapa tahun terakhir sebagai salah satu metode implementasi *Machine Learning* (ML). Pada beberapa artikel disebutkan bahwa DL tidak hanya spesifik untuk bidang tertentu, tetapi telah didefinisikan sebagai bentuk pembelajaran umum yang dapat menyelesaikan hampir berbagai macam masalah di berbagai bidang. DL menggunakan mekanisme *deep architecture of learning* atau pendekatan *hierarchical learning*. *Learning* atau pembelajaran dalam hal ini adalah sebuah prosedur yang berisi proses estimasi parameter-parameter suatu model sehingga model yang dikembangkan (*algoritme*) dapat menyelesaikan suatu tugas atau permasalahan tertentu. DL menggunakan beberapa lapisan (*layers*) di antara lapisan masukan (*input layer*) dan lapisan keluaran (*output layer*). Arsitektur tersebut dapat digunakan untuk melakukan pemrosesan non linier dengan beberapa tahap yang hasilnya dapat digunakan untuk *feature learning* dan klasifikasi pola (*pattern classification*).



Gambar 3. Kategori dalam *Deep Learning* (DL)
Sumber: (Haris dkk., 2021)

Beberapa teknik dalam *deep learning* dapat dikategorikan menjadi *supervised*, *semi-supervised*, dan *unsupervised*. Kategori lain, seperti *Reinforcement Learning* (RL) atau *Deep* RL (DRL), sering kali dikategorikan menjadi *semi-supervised* atau *unsupervised*. Pengategorian tersebut diperlihatkan pada Gambar. 2

1. *Deep Supervised Learning*: Teknik *learning* yang digunakan dalam kategori ini menggunakan data yang telah diberi label sebelumnya (*labeled* data). Contoh yang populer dalam kategori ini adalah *Deep Neural Networks* (DNN), CNN, (RNN), termasuk juga LSTM, dan *Gated Recurrent Unit* (GRU).
2. *Deep Semi-Supervised Learning*
   Deep Semi-supervised learning menggunakan teknik *learning* yang menggunakan sebagian data yang telah diberi label sebelumnya (*partially labeled*

data). Pada beberapa kasus, DRL, *Generative Adversarial Networks* (GAN), serta RNN, termasuk LSTM, dan GRU juga menggunakan teknik *learning* ini.

3. *Deep Unsupervised Learning*
   Teknik *learning* ini menggunakan data yang tidak diberi label sebelumnya (*unlabeled data*). *Auto Encoders* (AE), *Restricted Boltzmann Machines* (RBM), dan generasi terbaru dari GAN menggunakan teknik *learning* ini pada implementasinya.
4. *Deep Reinforcement Learning*
   Teknik *learning* ini digunakan pada lingkungan atau *environments* yang tidak diketahui (*unknown environments*).

1.6.7 *Random Forest*

Random Forest adalah metode pembelajaran *ensemble* berbasis pohon keputusan yang secara sistematis membangun banyak pohon keputusan acak, kemudian menggabungkan hasil prediksi dari seluruh pohon tersebut dengan cara merata-ratakannya. Pendekatan ini bertujuan untuk meningkatkan akurasi prediksi sekaligus mengurangi kemungkinan *overfitting*. Dalam implementasinya, *Random Forest* digunakan sebagai metode klasifikasi ketika variabel target bersifat kategorikal, dan sebagai metode regresi ketika variabel target bersifat numerik kontinu. Salah satu keunggulan *Random Forest* adalah kemampuannya dalam mengevaluasi kontribusi masing-masing fitur terhadap performa model melalui perhitungan nilai penting (*feature importance*). Nilai ini membantu peneliti mengidentifikasi fitur-fitur yang paling berpengaruh terhadap hasil prediksi.

Random Forest mengombinasikan prinsip dari algoritma *bagging* dan pohon keputusan CART (*Classification and Regression Tree*). Proses pembentukannya mencakup beberapa tahapan sebagai berikut:

Pertama, sejumlah sampel dipilih secara acak dari *dataset* pelatihan untuk membentuk beberapa *subset* data (*bootstrap sampling*). Masing-masing *subset* digunakan untuk melatih satu pohon keputusan. Dalam proses pelatihan ini, sejumlah k fitur dipilih secara acak dari keseluruhan fitur untuk menentukan titik pemisahan terbaik (*split point*) pada setiap *node*.

Kedua, proses pemisahan dilakukan dengan membagi ruang input menjadi N wilayah (R1, R2, …, Rn) berdasarkan minimisasi galat kuadrat. Asumsikan x(j) sebagai fitur ke-j dan s sebagai nilai ambang pemisahan, maka dua wilayah dapat didefinisikan sebagai:

$$R_{1(j,s)} = x|x^{(j)} \leq \text{s}, R_{2(j,s)} = x|x^{(j)} > s \qquad (1)$$

Keterangan :

$R_{1(j,s)}$: Region pertama tempat data dengan fitur ke- j bernilai kurang dari atau sama dengan ambang s

$R_{2(j,s)}$: Region kedua tempat data dengan fitur ke-j bernilai lebih dari ambang s

$x^{(j)}$: Nilai fitur ke- j

s : Nilai ambang batas (*threshold*)

j : Indeks fitur yang digunakan untuk *split*

Pemisahan ini diulang secara rekursif sampai kondisi penghentian terpenuhi. Pohon regresi kemudian dihasilkan dengan menghitung nilai rata-rata dari setiap wilayah partisi, sebagaimana dirumuskan dalam:

$$f(x) = \sum_{n=1}^{N} C_n I \qquad (2)$$

Keterangan :

$f(x)$: Nilai prediksi dari pohon regresi untuk input x

N : Jumlah region atau daun (*leaves*) dalam pohon

C_n : Nilai prediksi (rata-rata target) untuk region R_n

$I(x \in R_n)$: Fungsi indikator, bernilai 1 jika x berada dalam region R_n, dan 0 jika tidak

Ketiga, proses pelatihan diulangi pada semua *subset* data, dengan fitur dan titik pemisahan yang dipilih secara acak pada setiap iterasi, untuk membentuk pohon-pohon tambahan. Setelah itu hasil akhir dari *Random Forest* diperoleh dengan menghitung rata-rata prediksi dari seluruh pohon yang telah dibangun.
(Yang dkk., 2023)

1.6.8 *Model XGBoost*

XGBoost (*Extreme Gradient Boosting*) adalah algoritma pembelajaran terawasi yang termasuk dalam keluarga metode *ensemble* dan dikenal karena efisiensinya dalam memproses data dalam jumlah besar serta kemampuannya menghasilkan prediksi yang sangat akurat. XGBoost bekerja dengan membangun pohon keputusan secara bertahap, di mana setiap pohon baru dilatih untuk memperbaiki kesalahan (residual) dari pohon sebelumnya melalui optimasi berbasis gradien.

Fungsi objektif dalam XGBoost terdiri dari dua komponen utama: fungsi kerugian (*loss function*) yang mengukur perbedaan antara prediksi dan nilai aktual, serta istilah regularisasi yang bertujuan mengontrol kompleksitas model dan mencegah *overfitting*. Rumus umum fungsi objektifnya adalah:

$$L^{(t)} = \sum_{i=1}^{n} l(y_i, \hat{y}_i) + \sum_{t=1}^{T} \Omega(f_t) \qquad (3)$$

Keterangan :

$L^{(t)}$: Fungsi objektif total pada iterasi ke-t

$l(y_i, \hat{y}_i)$: Fungsi loss antara nilai aktual y_i dan hasil prediksi $\hat{y}_i$

$\Omega(f_t)$: Fungsi regularisasi yang menghukum kompleksitas model f_t

n : Jumlah total sampel pada data latih

T : Jumlah total pohon yang dibangun dalam model

f_t : Pohon keputusan ke-t

Output akhir dari XGBoost dihitung sebagai jumlah output dari seluruh pohon yang telah dibangun:

$$\hat{y} = \sum_{t=1}^{T} f_t(x_i) \qquad (4)$$

Keterangan :

$\hat{y}$: Hasil prediksi akhir dari model XGBoost

$f_t(x_i)$: *Output* dari pohon ke-t untuk sampel x_i

x_i : Sampel ke-i dari data latih

T : Jumlah total pohon dalam model di mana $f_t(x_i)$ adalah *output* dari pohon ke-t untuk sampel x_i

XGBoost juga menyediakan metode interpretasi fitur melalui tiga metrik utama: *Gain*, *Cover*, dan *Frequency*. *Gain* menunjukkan kontribusi fitur terhadap pengurangan galat, *Cover* mencerminkan jumlah observasi yang terpengaruh oleh fitur tersebut, dan *Frequency* mengindikasikan seberapa sering fitur muncul dalam proses pemisahan *node* pada pohon-pohon dalam model. Di antara ketiganya, *Gain* umumnya dianggap paling representatif dalam menilai pentingnya fitur. Dengan demikian, XGBoost tidak hanya menawarkan akurasi prediksi yang tinggi, tetapi juga memberikan wawasan tentang fitur mana yang paling relevan terhadap target prediksi, menjadikannya alat yang kuat untuk pemodelan dan analisis data kompleks**.** (Yang dkk., 2023)

1.7 Penelitian Terdahulu

Beberapa penelitian telah dilakukan sebelumnya dengan topik yang relevan dengan penelitian ini. Penelitian terdahulu ini digunakan sebagai referensi yang memiliki metode ataupun permasalahan sejenis yang diangkat dalam penelitian ini. Berikut beberapa penelitian terdahulu yang terkait dengan Peramalan harga komoditas dengan menggunakan *machine learning:*

Tabel 1. Penelitian terdahulu

No	Peneliti	Judul	Metode	Hasil	Objek
1.	Yang dkk.,. (2023)	*Forecasting Crude Oil Futures Prices using Extreme Gradient Boosting*	XGBoost, *Random Forest*	XGBoost menghasilkan performa prediksi terbaik dalam memproyeksikan harga minyak mentah berjangka dibandingkan *benchmark* dan *Random Forest*. Penelitian juga menunjukkan bahwa penggunaan data dengan jeda waktu antara pencarian dan pemrosesan informasi oleh investor menghasilkan model prediksi yang lebih akurat. Evaluasi dilakukan menggunakan MAPE, MAE, dan RMSE.	Harga minyak mentah berjangka di China
2.	Harjanto dkk., (2022)	*Export Commodity Price Forecasting in Indonesia Using Decision Tree, Random Forest, and*	*Decision Tree*, *Random Forest*, LSTM	LSTM memiliki kinerja terbaik dalam memprediksi harga komoditas ekspor Indonesia berdasarkan nilai MAPE terendah, yaitu 0.121 untuk batubara, 0.494 untuk kopi, dan 0.282 untuk minyak sawit. *Random Forest* dan	Harga ekspor minyak sawit, kopi, dan batu bara (Indonesia)

No	Peneliti	Judul	Metode	Hasil	Objek
		Long Short-Term Memory		*Decision Tree* digunakan sebagai pembanding, namun performanya tidak sebaik LSTM. Penelitian ini juga menggunakan *feature scaling* dan validasi silang.	
3.	Zhuoran Li (2023)	*Comparison of XGBoost and LSTM Models for Stock Price Prediction*	XGBoost, LSTM	XGBoost lebih unggul dalam menangani *dataset* besar dengan performa stabil terhadap fluktuasi jangka panjang, sedangkan LSTM memberikan hasil yang lebih baik untuk *dataset* kecil dengan pola jangka pendek. Evaluasi dilakukan dengan RMSE, visualisasi kurva aktual dan prediksi, serta analisis residual.	Harga saham Chevron (CVX)
4.	Oukhouya ddk., (2023)	*Forecasting International Stock Market Trends: XGBoost, LSTM, LSTM-XGBoost*	XGBoost, LSTM, LSTM-XGBoost	Model *hybrid* LSTM-XGBoost menunjukkan kinerja terbaik dalam memprediksi arah tren indeks saham global. Kombinasi model LSTM dan XGBoost mampu menangkap pola deret waktu sekaligus memperbaiki prediksi melalui *boosting*. Optimasi parameter dilakukan dengan *grid search* dan evaluasi melalui akurasi arah (*directional accuracy*).	Indeks saham internasional: MASI, CAC 40, DAX, FTSE 250, NASDAQ, HKEX
5.	Murugesan dkk., (2023)	*Deep Learning Based Models to Forecast Agricultural Commodities Prices*	Vanilla LSTM, Bi-LSTM, Stacked LSTM, CNN-LSTM, Conv-LSTM	Model Bi-LSTM dan CNN-LSTM memberikan hasil paling stabil dan akurat dalam memprediksi harga komoditas pertanian berdasarkan evaluasi metrik MAE, MSE, RMSE, dan MAPE. Studi ini membandingkan lima varian LSTM dan menunjukkan pentingnya pemilihan	Beras, gandum, pisang, kacang gram, kacang tanah (India)

No	Peneliti	Judul	Metode	Hasil	Objek
				arsitektur untuk efisiensi dan akurasi prediksi.	
6.	Zakiya Yahaya Shehu ddk., (2024)	*Improving the Accuracy of Food Commodity Price Prediction Model Using Deep Learning Algorithm*	Hybrid LSTM-CNN, Whale Optimization Algorithm (WOA)	Model LSTM-WOA memberikan hasil terbaik dengan RMSE 0.0071–0.0073, MSE 0.0061, MAE 0.0082–0.0083, dan R^2 0.972–0.975, lebih unggul dibanding CNN-LSTM dan CNN standar. Selain itu, penelitian menunjukkan bahwa integrasi WOA mampu mengatasi kelemahan LSTM-CNN terkait *overfitting* dan biaya komputasi, sehingga model lebih stabil pada data besar.	Harga komoditas pangan (jagung, gula, beras, kacang-kacangan) di Nigeria
7.	Sunoto & Mangapul Siahaan (2025)	*Analisis Prediksi Hasil Padi dengan Model MLPRegressor dan Long Short-Term Memory*	MLPRegressor, LSTM	LSTM arsitektur 2-42-42-42-1 memberikan akurasi tertinggi 94,12% (MSE 0.00305660), sedangkan MLPRegressor 2-22-1 memberikan akurasi 91,18% (MSE 0.00471975). Selain itu, hasil penelitian menekankan bahwa LSTM lebih unggul dalam menangkap pola musiman, sedangkan MLPRegressor lebih sederhana namun kurang adaptif pada data temporal.	Produktivitas padi 34 provinsi di Indonesia (data BPS 2018–2023)
8.	Anil Kumar Mahto dkk., (2021)	*Short-Term Forecasting of Agriculture Commodities in Context of Indian Market*	Artificial Neural Network (ANN), ARIMA	ANN mengungguli ARIMA dengan MAPE dan RMSPE lebih rendah, sehingga lebih efektif untuk data non-linear dalam prediksi harga jangka pendek. Selain itu, penelitian membuktikan bahwa ANN dapat mengatasi ketidakstabilan harga yang dipengaruhi faktor musiman	Harga kedelai (2014–2018) dan bunga matahari (2011–2016) di India

No	Peneliti	Judul	Metode	Hasil	Objek
				dan non-linearitas, yang tidak bisa ditangani ARIMA dengan baik.	
9.	Yeong Hyeon Gu dkk., (2022)	*Forecasting Agricultural Commodity Prices Using Dual Input Attention LSTM*	Dual Input Attention LSTM (DIA-LSTM)	DIA-LSTM menggunakan data harga, volume perdagangan, dan data meteorologi, menghasilkan MAPE lebih rendah (1.41%–4.26%) dibanding *benchmark* model. Menggunakan *dynamic main production area* meningkatkan akurasi 2.8%–5.5%. Selain itu, penelitian menegaskan bahwa penggunaan *dynamic main production area* memberikan gambaran lebih realistis terhadap pengaruh iklim dibanding pendekatan statis.	Harga bulanan kubis dan lobak di Korea Selatan
10	S. Hossain dkk., (2023)	*A tree based eXtreme Gradient Boosting (XGBoost) machine learning model to forecast the annual rice production in Bangladesh*	XGBoost, *Decision Tree*, *Random Forest*	XGBoost menghasilkan prediksi paling akurat dibandingkan model lain, dengan *error* terkecil dan performa stabil. Selain itu, penelitian menekankan pentingnya pemilihan fitur agrikultur (curah hujan, suhu, area tanam) dalam meningkatkan akurasi model.	Produksi beras tahunan di Bangladesh
11	Veena K dkk., (2025)	*Price Prediction of Agriculture Commodities Using Machine Learning and SARIMAX*	SARIMAX, *Random Forest*, SVR, XGBoost	Model hibrida SARIMAX-ML lebih akurat dibanding model tunggal, mampu menangkap pola musiman dan faktor eksternal (cuaca, permintaan). Selain itu, integrasi *hybrid* meningkatkan keandalan prediksi baik jangka pendek maupun panjang.	Harga komoditas pertanian di India

No	Peneliti	Judul	Metode	Hasil	Objek
13	Ranjit K. Paul dkk., (2022)	*Machine learning techniques for forecasting agricultural prices: A case of brinjal in Odisha, India*	GRNN, SVR, *Random Forest*, GBM, ARIMA	GRNN menunjukkan performa terbaik di sebagian besar pasar, disusul RF dan GBM. Hasil menunjukkan model ML non-linear lebih unggul daripada ARIMA dalam memprediksi harga sayuran yang fluktuatif.	Harga brinjal (terong) di 17 pasar utama Odisha, India
14	Eric Bauer & Ron Kohavi (1999)	*An empirical comparison of voting classification algorithms: Bagging, Boosting, and Variants*	*Bagging*, AdaBoost, Arc-x4, *Decision Tree*, Naive Bayes	*Bagging* menurunkan variansi model tidak stabil (misalnya *Decision Tree*), sedangkan AdaBoost mampu menurunkan bias dan variansi sekaligus. Namun, pada model stabil seperti Naive Bayes, *boosting* justru meningkatkan *error*. Ini menunjukkan efektivitas *ensemble* tergantung pada model dasar.	Data simulasi & *dataset* benchmark (UCI *Machine Learning Repository*)
15	R. Ragunath & R. Rathipriya (2025)	*Forecasting Agriculture Commodity Price Trend using Novel Competitive Ensemble Regression Model*	Competitive *Ensemble* Regression (CER) dengan 5 model dasar: SVR, DTR, GPR, RFR, XGBoost	CER menunjukkan akurasi lebih tinggi dibanding model tunggal dan *ensemble* klasik. RMSE menurun: 48.3% (*Barley*), 36.8% (Ragi), 32.9% (*Wheat*). Prediksi harga komoditas Juli 2024–Juni 2025 juga dihasilkan, misalnya harga padi diprediksi Rp3.158–Rp3.146/bulan.	Harga komoditas pertanian di India (Padi, Jagung, Barley, Gandum, Ragi)

BAB II
METODE PENELITIAN

2.1 Jenis dan Sumber Data

Penelitian ini menggunakan data sekunder berupa data historis harga lima komoditas pangan strategis (beras, bawang putih, bawang merah, cabai merah, dan daging sapi). Data diperoleh dari situs resmi Pusat Informasi Harga Pangan Strategis (PIHPS) Nasional untuk wilayah Sulawesi Selatan.

2.2 Periode dan Jumlah Data

Data yang digunakan mencakup harga harian dari tahun 2022 hingga 2024, dengan total sebanyak 782 observasi per komoditas. Rentang waktu selama tiga tahun dipilih karena mampu menangkap pola tren jangka panjang, fluktuasi musiman, anomali yang mungkin terjadi dalam siklus pasar serta memastikan model dilatih pada kondisi pasar yang masih relevan dengan situasi saat ini.

2.3 Teknik Pengumpulan Data

Pengumpulan data dilakukan dengan cara:

1. Identifikasi komoditas
   Pemilihan objek penelitian melalui analisis Google Trends untuk menentukan lima komoditas pangan yang paling banyak dicari oleh masyarakat Sulawesi Selatan.
2. Literature Review
   Peneliti melakukan kajian terhadap berbagai penelitian terdahulu yang relevan guna memperoleh landasan teori serta pendekatan metodologis yang tepat untuk digunakan dalam penelitian ini.
3. Pengumpulan Data Historis
   Tahap ini merupakan proses teknis pengambilan data harga harian dari situs PIHPS Nasional (https://www.bi.go.id/hargapangan) untuk wilayah Sulawesi Selatan, mencakup lima komoditas yang telah ditentukan dalam 3 tahun.

2.4 Teknik Pengolahan Data dan Analisis

Dalam penelitian ini, tahapan pemodelan dilakukan untuk membangun dan membandingkan performa algoritma *Machine Learning*, yaitu *Random Forest* dan *Extreme Gradient Boosting* (XGBoost), dengan metode konvensional *Linear Regression* dalam memprediksi harga lima komoditas pangan strategis di Sulawesi Selatan. *Linear Regression* secara spesifik ditetapkan sebagai metode konvensional (*baseline*) setelah melalui tahapan tinjauan literatur dan analisis karakteristik data yang menunjukkan volatilitas tinggi serta pola non-linear. Berdasarkan perbandingan teoritis, metode *time series* klasik seperti ARIMA atau SARIMA dinilai kurang relevan untuk studi kasus ini karena mensyaratkan asumsi stasioneritas data yang ketat, yang sulit dipenuhi oleh komoditas pangan dengan fluktuasi harga ekstrem akibat guncangan eksternal. Oleh karena itu, *Linear Regression* dipilih untuk merepresentasikan pendekatan parametrik sederhana, dengan tujuan memvalidasi secara empiris bahwa kompleksitas data harga pangan di Sulawesi Selatan memang memerlukan penanganan algoritma non-linear, sekaligus mengukur signifikansi peningkatan akurasi yang ditawarkan oleh model *Machine Learning*. Keseluruhan proses pemodelan ini dilakukan secara sistematis melalui beberapa tahapan, mulai

dari pengumpulan data, pra-pemrosesan (*preprocessing*), pembangunan model, pelatihan model (*training*), evaluasi kinerja, hingga analisis hasil prediksi.

Setiap algoritma diuji menggunakan data historis harga pangan yang telah dibersihkan dan disesuaikan formatnya, dengan mempertimbangkan aspek-aspek seperti tren waktu, musiman, serta fluktuasi harga. Kinerja kedua model kemudian dibandingkan menggunakan metrik evaluasi yang relevan seperti MAPE, *Root Mean Squared Error* (RMSE), dan R-squared (R^2) untuk menentukan model mana yang paling optimal dalam melakukan peramalan harga pangan.

2.2.1 Pengumpulan Data

Tahap pertama adalah mengumpulkan data yang digunakan sebagai *input* dalam pembuatan dan pelatihan model. Data mencakup berbagai informasi penting, seperti harga masing masing komoditas yang akan diprediksi. Contoh data yang di ambil dari *website* PIHPS.

Tabel 2. Contoh Data Mentah

Komoditas (Rp)	03/ 01/ 2022	04/ 01/ 2022	05/ 01/ 2022	06/ 01/ 2022	07/ 01/ 2022
Beras	9,900	9,900	9,900	9,900	9,900
Daging Sapi	116,000	116,000	116,000	116,000	116,000
Bawang Merah	25,200	25,350	25,300	25,300	25,150
Bawang Putih	26,850	27,000	26,900	26,800	26,800
Cabai Merah	34,550	34,200	33,450	33,200	32,550

Sumber: Data Sekunder (2025)

Informasi historis harga komoditas sangat penting untuk membangun pola yang dapat digunakan dalam pemodelan prediksi harga pangan. Data yang digunakan sebagai *input* dalam model meliputi tanggal sebagai indeks waktu serta harga komoditas harian sebagai variabel utama. Dalam penelitian ini, variabel *input* disusun berdasarkan data harga harian dari lima komoditas strategis, yaitu beras, bawang putih, bawang merah, cabai merah, dan daging sapi.

2.2.2 Pra-Pemrosesan Data

Setelah data harga komoditas berhasil dikumpulkan, tahap berikutnya adalah proses pra-pemrosesan untuk memastikan data siap digunakan sebagai masukan dalam model Peramalan. Proses ini penting dilakukan untuk membersihkan data dari kesalahan, menyesuaikan format, sehingga model dapat mengenali pola historis secara efektif dan menghasilkan prediksi yang akurat.

a. Membersihkan Data
   Pada tahap ini, data diperiksa untuk mengidentifikasi nilai yang hilang (*missing values*), duplikasi, atau nilai yang tidak valid. Nilai kosong dapat diatasi dengan metode interpolasi, pengisian rata-rata, atau penghapusan baris data tertentu, tergantung pada jumlah dan dampaknya terhadap *dataset* secara keseluruhan. Proses ini dilakukan agar model hanya bekerja dengan data yang konsisten dan berkualitas tinggi.
b. Membagi Data ke dalam *Training* dan *Testing Set*
   Setelah data dibersihkan dan disiapkan, data kemudian dibagi menjadi dua *subset*: data pelatihan (*training set*) dan data pengujian (*testing set*). Data

pelatihan digunakan untuk membangun model, sedangkan data pengujian digunakan untuk mengevaluasi kinerja model terhadap data yang belum pernah dilihat sebelumnya. Dalam penelitian ini, proporsi pembagian data adalah 80% untuk *training* set dan 20% untuk *testing* set. Rasio ini memberikan keseimbangan optimal antara jumlah data yang cukup untuk melatih model agar mengenali pola historis, dan data pengujian yang memadai untuk mengevaluasi kemampuan generalisasi model terhadap data baru. Dengan jumlah data harian yang besar dan rentang waktu tiga tahun (2022–2024), rasio 80:20 dianggap paling sesuai karena mampu menjaga stabilitas model dan menghindari *overfitting*, sekaligus memberikan hasil evaluasi yang representatif (Adinugroho, 2022).

2.2.3 Mendesain dan Membuat Model

Pada tahap desain dan pembuatan model, langkah pertama yang dilakukan adalah mempersiapkan data historis harga pangan dari lima komoditas strategis, yaitu beras, bawang putih, bawang merah, cabai merah, dan daging sapi. Data ini diperoleh dari Pusat Informasi Harga Pangan Strategis (PIHPS) dan telah melalui proses pra-pemrosesan, yang meliputi pembersihan data, penanganan data hilang, serta penyesuaian format agar sesuai dengan kebutuhan pemodelan.

Setelah data diproses, langkah selanjutnya adalah membagi data menjadi data latih (*training set*) dan data uji (*testing set*) dengan rasio 80:20. Tujuannya adalah untuk memastikan model dapat dilatih dengan data historis yang cukup, serta diuji menggunakan data baru untuk mengukur akurasi prediksinya secara objektif.

Dalam penelitian ini, dua algoritma *Machine Learning* digunakan, yaitu *Random Forest* dan *Extreme Gradient Boosting* (XGBoost). Selain itu, metode konvensional yang akan digunakan sebagai pembanding dari metode *machine learning* yaitu *Linear Regression.*

Random Forest bekerja dengan membangun sejumlah pohon keputusan (*decision tree*) secara acak dan menggabungkan hasilnya untuk menghasilkan prediksi yang lebih stabil dan akurat. Algoritma ini cocok untuk menangani data non-linear dan memiliki toleransi yang baik terhadap *noise* serta *overfitting*. XGBoost adalah algoritma berbasis *boosting* yang meningkatkan akurasi prediksi dengan membangun model secara bertahap, di mana setiap model baru fokus untuk memperbaiki kesalahan dari model sebelumnya. XGBoost dikenal karena efisiensinya dalam komputasi dan performanya yang unggul dalam berbagai kompetisi data *science*.

2.2.4 Melatih Model

Tahap pelatihan model dilakukan setelah proses pra-pemrosesan dan pembagian data selesai dilakukan. Tujuan dari tahap ini adalah untuk melatih model agar dapat mengenali pola historis dalam data harga komoditas pangan dan menghasilkan prediksi yang akurat.

a. Menentukan Fungsi Evaluasi

   Fungsi evaluasi yang digunakan dalam penelitian ini meliputi *Root Mean Squared Error* (RMSE), *Mean Absolute Percentage Error* (MAPE), dan koefisien determinasi (R^2). RMSE digunakan untuk mengukur seberapa besar kesalahan prediksi dalam satuan yang sama dengan data asli, sehingga memberikan

interpretasi yang lebih intuitif terhadap deviasi prediksi. MAPE memberikan gambaran persentase kesalahan rata-rata terhadap nilai aktual, yang berguna untuk memahami akurasi model dalam konteks relatif. Sementara itu, R^2 digunakan untuk menilai seberapa baik model menjelaskan variabilitas data, dengan nilai mendekati 1 menunjukkan tingkat kecocokan yang tinggi antara hasil prediksi dan data aktual.

b. Penentuan *Hyperparameter* (Parameter Tuning)
   Model *Random Forest* dan XGBoost memiliki sejumlah parameter yang perlu diatur agar mencapai performa terbaik. Penyesuaian parameter seperti jumlah pohon (*n_estimators*), kedalaman maksimum pohon (*max_depth*), dan tingkat pembelajaran (*learning_rate*) dilakukan menggunakan pendekatan *grid search*, *random search* ataupun optimasi parameter secara manual . Penyetelan ini dilakukan secara eksperimen agar model mampu beradaptasi dengan karakteristik data dan menghasilkan prediksi yang optimal.
c. Proses *Training*
   Data historis harga komoditas pangan yang telah diproses dibagi menjadi dua bagian, yaitu 80% untuk data pelatihan dan 20% untuk data pengujian. Model dilatih menggunakan data pelatihan untuk mempelajari pola, kemudian diuji menggunakan data pengujian untuk mengukur akurasi prediksi. Selama proses pelatihan, model akan melalui iterasi pembelajaran untuk menyesuaikan parameter internal hingga mencapai hasil prediksi yang stabil dan akurat.
   (Qiu dkk., 2020)

2.2.5 Analisis dan Perbandingan

Dalam mengevaluasi kinerja model prediksi harga komoditas pangan, terdapat tiga metrik utama yang digunakan untuk memberikan penilaian yang menyeluruh. Pertama, *Root Mean Squared Error* (RMSE) digunakan untuk mengukur seberapa besar kesalahan prediksi dalam satuan harga, dengan memberikan penalti lebih besar terhadap kesalahan yang ekstrem. Kedua, *Mean Absolute Percentage Error* (MAPE) berguna untuk membandingkan performa model antar komoditas dengan harga yang berbeda, karena menampilkan kesalahan dalam bentuk persentase yang mudah dipahami. R^2 *Score* digunakan untuk menilai sejauh mana model mampu menjelaskan variasi dalam data historis dan menangkap pola-pola yang ada. Gabungan ketiga metrik ini memberikan wawasan lengkap mengenai akurasi prediksi model, baik dari sisi kesalahan absolut, kesalahan relatif, maupun kemampuan dalam merepresentasikan karakteristik data. Metrik evaluasi untuk setiap model secara umum diformulasikan seperti persamaan sebagai berikut. (Arwansyah dkk., 2024)

$$RMSE = \sqrt{\frac{1}{n}\sum_{t=1}^{n}(A_t - F_t)^2} \quad (5)$$

$$MAPE = \frac{1}{n}\sum_{t=1}^{n}\left|\frac{A_t - F_t}{A_t}\right| \times 100\% \quad (6)$$

$$R^2 = 1 - \frac{\sum_{t=1}^{n}(A_t - F_t)^2}{\sum_{t=1}^{n}(\bar{A} - A_t)^2} \quad (7)$$

Keterangan :

A_t : nilai aktual *(true value)*

F_t : nilai prediksi *(predicted value)*

n : jumlah data

$\bar{A}$: *rata-rata nilai aktual*

2.2.6 Validasi Model

Untuk memastikan bahwa model terbaik yang diperoleh tidak hanya unggul berdasarkan metrik evaluasi seperti RMSE, MAPE, dan R^2, tetapi juga kuat secara statistik, penelitian ini menggunakan dua metode validasi tambahan, yaitu Uji *Diebold–Mariano* (DM *Test*) dan Uji *Durbin–Watson* (DW *Test*). Kedua uji ini berperan penting dalam mengevaluasi kualitas model peramalan pada konteks deret waktu.

a. Uji *Diebold–Mariano* (DM Test)

Uji *Diebold–Mariano* (DM Test) digunakan untuk menguji apakah terdapat perbedaan yang signifikan antara performa dua model peramalan yang berbeda (Diebold & Mariano, 1995). Uji ini membandingkan rata-rata *loss differential* (perbedaan kesalahan) antara dua model berdasarkan kriteria kesalahan tertentu seperti *Mean Absolute Error* (MAE) atau *Mean Squared Error* (MSE). Secara matematis, statistik uji DM dirumuskan sebagai berikut:

$$DM = \frac{\bar{d}}{\sqrt{\frac{\hat{\sigma}_d^2}{n}}} \quad (8)$$

dengan:

$$\bar{d} = \frac{1}{n}\sum_{t=1}^{n} d_t \quad (9)$$

$$d_t = L(e_{1t}) - L(e_{2t}) \quad (10)$$

di mana:

d_t : selisih *loss function* antara dua model pada periode ke-t,

$L(e_{it})$: fungsi *loss* model ke-i (misalnya e_{it}^2untuk MSE atau $|e_{it}|$untuk MAE),

$\hat{\sigma}_d^2$: varians dari deret d_t,

n : jumlah observasi.

Statistik DMmengikuti distribusi t dengan derajat kebebasan $n-1$. Hipotesis yang diuji adalah:

H_0 : $E(d_t) = 0$(tidak ada perbedaan signifikan antara dua model)

H_1 : $E(d_t) \neq 0$(terdapat perbedaan signifikan antara dua model)

Apabila nilai p-value < 0,05, maka hipotesis nol ditolak, yang berarti terdapat perbedaan performa yang signifikan antara kedua model. Dalam penelitian ini, DM *Test* digunakan untuk membandingkan model *Linear Regression, Random Forest*, dan XGBoost pada masing-masing komoditas. Hasil uji ini memberikan dasar statistik yang kuat untuk menentukan model terbaik secara objektif, bukan hanya berdasarkan nilai metrik evaluasi (Coroneo & Iacone, 2024).

b. Uji *Durbin–Watson* (DW Test)

Uji *Durbin–Watson* (DW Test) digunakan untuk mendeteksi adanya autokorelasi residual pada model regresi atau peramalan deret waktu. Autokorelasi menunjukkan sejauh mana error pada suatu periode dipengaruhi oleh error pada periode sebelumnya. Jika residual tidak acak (autokorelasi tinggi), maka model belum sepenuhnya menangkap pola dalam data dan hasil prediksi menjadi kurang reliabel.

Rumus statistik *Durbin–Watson* adalah sebagai berikut:

$$DW = \frac{\sum_{t=2}^{n}(e_t - e_{t-1})^2}{\sum_{t=1}^{n} e_t^2} \quad (11)$$

dengan:

e_t : residual pada periode ke-t,

n : jumlah observasi.

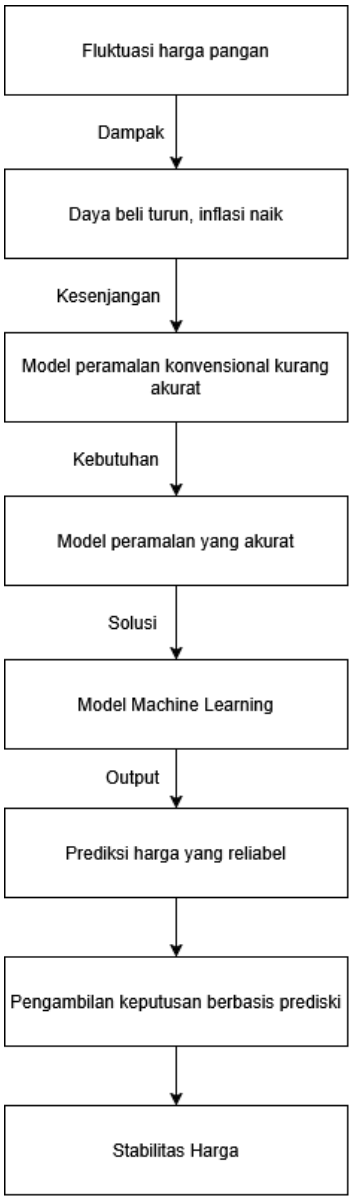
Nilai DW berada dalam rentang 0 hingga 4, dengan interpretasi umum sebagai berikut:

Nilai DW	Interpretasi
< 1.5	:Terdapat autokorelasi positif
1.5 – 2.5	:Tidak ada autokorelasi (residual acak)
> 2.5	:Terdapat autokorelasi negatif

(Ashley, 2025).

2.5 Kerangka Pikir

Kerangka pemikiran menjelaskan hubungan antara variabel dan memudahkan penulis untuk menyelesaikan karya tulis dengan lebih terarah dan mudah diselesaikan. Adapun kerangka berpikir pada penelitian ini ialah sebagai berikut:



2.6 Diagram Alir Penelitian

Diagram alir penelitian merupakan salah satu cara bagi penulis dalam melakukan penelitian secara sistematis. Berikut adalah alur penelitian yang dilakukan oleh peneliti, antara lain: