BAB II
TINJAUAN PUSTAKA DAN DASAR TEORI

2.1 Tinjauan Pustaka

Rahmadini et al. (2023) melakukan penelitian terkait prediksi harga beras di Indonesia menggunakan dataset dari situs Pusat Informasi Harga Pangan Strategis (PIHPS) Nasional periode Januari 2019 hingga Desember 2021. Algoritma K-Nearest Neighbor (KNN) diterapkan pada data yang telah dinormaliasi dan menghasilkan performa terbaik pada nilai k = 2. Model menunjukkan nilai Mean Absolute Error (MAE) dan Root Mean Squared Error (RMSE) sebesar 52,77 dan 96,40 pada data latih, serta 55,55 dan 81,64 pada data uji. Hasil ini menunjukkan bahwa algoritma KNN mampu menghasilkan prediksi yang cukup akurat, dan mengindikasikan kestabilan harga beras selama periode tersebut.

Sudriyanto et al. (2023) membandingkan kinerja algoritma KNN, Neural Network (NN), dan Support Vector Machine (SVM) dalam memprediksi harga saham PT Astra Internasional Tbk (ASII.JK). Hasil evaluasi menunjukkan bahwa model KNN memiliki nilai RMSE terendah 0,037, dibandingkan dengan NN sebesar 0,048, dan SVM sebesar 0,075. Hal ini menunjukkan bahwa KNN memberikan performa terbaik dalam prediksi harga saham ASII.JK.

Husdi et al. (2023) melakukan prediksi persediaan telur ayam pada UD Unggas Karya Mandiri menggunakan algoritma KNN. Model KNN berhasil memprediksi jumlah persediaan sebanyak 2.519 rak telur dengan tingkat akurasi mencapai 96,98% berdasarkan pengujian menggunakan Mean Absolute Percentage Error (MAPE).

Wijaya et al. (2023) menerapkan algortima KNN untuk memprediksi harga jagung di Kabupaten Tulungagung berdasarkan data sejak Januari 2020. Hasil prediksi harga jagung selama lima minggu berturut-turut pada Mei 2023 adalah 6.400; 6.416; 6.333; 6.333; dan 0, dengan nilai RMSE sebesar 185,15 serta akurasi model sebesar 95%.

Huda & Ula (2024) membandingkan performa lima akurasi algoritma Naive Bayes, Regresi Logistik (LR), Random Forest, Support Vector Machine (SVM), dan K-Nearest Neighbor (KNN) dalam memprediksi diabetes berdasarkan data 768 pasien dengan sembilan fitur diagnosis. Random Forest menghasilkan akurasi tertinggi (83%), diikuti oleh KNN dan SVM (masing-masing 80%), LR (78%), dan Naive Bayes (76%). Hasil ini menunjukkan bahwa Random Forest dan KNN merupakan metode prediktif yang potensial.

Qirani & Sukarsih (2024) melakukan penelitian untuk memprediksi harga gas alam menggunakan algoritma KNN, dengan evaluasi menggunakan MAPE. Hasil prediksi untuk bulan Juni menunjukkan nilai sebesar $2,8215, dengan harga aktual sebesar $2,839 dan MAPE sebesar 0,6164%. Nilai kesalahan yang rendah ini mengindikasikan bahwa KNN mampu menghasilkan prediksi yang sangat besar mendekati nilai aktual.

Fadhilah & Nudin (2024) menggunakan algoritma KNN untuk memprediksi harga mobil bekas. Evaluasi model menghasilkan nilai RMSE sebesar 1.324.544, yang menunjukkan bahwa algortima KNN mampu memodelkan harga mobil bekas dengan cukup akurat dan optimal, tergantung pada nilai k yang digunakan.

Tabel 2. 1 Perbandingan Penelitian

Peneliti	**Topik**	**Metode**	**Hasil**
Rahmadini et al. (2023)	Prediksi harga beras	KNN	Kinerja terbaik pada k = 2, MAE 55,55 harga relatif stabil.
Sudriyanto et al. (2023)	Prediksi harga saham ASII.JK	KNN, NN, SVM	KNN unggul dengan RMSE 0,037 dibanding model lain.
Husdi et al. (2023)	Prediksi persediaan telur ayam	KNN	Akurasi tinggi MAPE 96,98%, prediksi

Tabel 2.1 Lanjutan

Peneliti	Topik	Metode	Hasil
			mencapai 95%.
Wijaya et al. (2024)	Prediksi harga jagung	KNN	RMSE 185,15 dengan akurasi prediksi mencapai 95%.
Huda & Ula (2024)	Prediksi diabetes	KNN, RF, SVM, LR, NB	RF dan KNN unggul dengan akurasi 83% dan 80%.
Qirani & Sukarsih (2024)	Prediksi harga gas alam	KNN	Prediksi akurat dengan MAPE 0,62%.
Fadhilah & Nudin (2024)	Prediksi harga mobil bekas	KNN	RMSE 1,32 juta, prediksi cukup akurat dengan KNN.
Diajukan (2025)	Prediksi volume dan nilai produksi ikan	KNN	MAE 371.453 kg & Rp14,4 miliar, RMSE 674.739 kg & 25,2 miliar, dan R^2 92% & 83%. Model akurat menangkap pola data.

2.2 Dasar Teori

2.2.1 K-Nearest Neighbor (KNN)

KNN pertama kali diperkenalkan oleh Evelyn Fix dan Joseph Hodges pada tahun 1951, kemudian dikembangkan lebih lanjut oleh Thomas Cover. KNN

merupakan metode dalam supervised learning, yaitu metode pembelajaran yang menggunakan data *training* untuk membimbing proses pembelajaran dalam membentuk model prediksi yang optimal. Konsep dasar dari KNN adalah mencari k tetangga terdekat berdasarkan kemiripan biasanya dihitung dengan jarak, seperti Euclidean *distance*, yang kemudian digunakan untuk menentukan nilai *output*.

Langkah-langkah Algoritma KNN adalah:

1. Membuat data *training* dengan label kelas yang diketahui.
2. Membuat data *testing* dengan label kelas yang tidak diketahui.
3. Menentukan nilai k.
4. Mengklasifikasikan data *testing* dengan menetapkan kelas berdasarkan mayoritas kelas dari k tetangga terdekat dalam data *training*.

Langkah-langkah tersebut jika digabungkan dengan proses analisis menggunakan algoritma KNN, sebagai berikut (1) membagi data menjadi data *training* dan data *testing*, (2) melakukan transformasi data (jika diperlukan), (3) menghitung jarak antar data, (4) menentukan nilai k, (5) menentukan output berdasarkan tetangga terdekat baik untuk data *training* maupun data *testing*, (6) evaluasi performa model.

Penentuan banyaknya data *training* dan data *testing* dapat menggunakan dua skema meliputi: (1) pengambilan sampel acak tanpa pengembalian dengan mengabaikan proporsi kelas, atau (2) pengambilan sampel secara acak tanpa pengembalian dengan mempertimbangkan proporsi kelas (Sa'adah et al. 2021).

Pada algoritma KNN, kedekatan antara data *training* dan data *testing* menjadi faktor utama dalam proses klasifikasi maupun prediksi. Ukuran kedekatan ini umumnya dihitung menggunakan matrik jarak, di mana jarak Euclidean adalah yang paling umum digunakan. Rumus adalah sebagai berikut (Situju et al. 2023):

$$D_{xy} = \sqrt{\sum_{i=1}^{n} (xi - yi)^2}$$

Keterangan:

D = Jarak kedekatan

x = Data *testing*

y = Data *training*

n = Jumlah atribut individu

i = Atribut individu antara 1 sampai dengan n

Selain digunakan dalam klasifikasi, algoritma KNN juga dapat diterapkan untuk permasalahan regresi, yaitu memprediksi nilai kontinu. Penjelasan lebih lanjut mengenai pendekatan regresi KNN dibahas pada subbab berikut.

2.2.2 Regresi K-Nearest Neighbor (KNN Regression)

Regresi K-Nearest Neighbor (KNN Regression) merupakan metode non-parametrik yang digunakan untuk memprediksi nilai kontinu dari suatu data baru. Prediksi didasarkan pada rata-rata nilai target dari sejumlah k tetangga terdekat dalam data pelatihan. Metode ini tidak membentuk model eksplisit pada saat pelatihan, melainkan menyimpan seluruh data pelatihan dan melakukan prediksi langsung saat data baru diberikan (*instance-based learning* atau *lazy learning*).

Berbeda dengan KNN untuk klasifikasi yang menetapkan kelas berdasarkan mayoritas tetangga, regresi KNN menghitung rata-rata nilai target dari k tetangga terdekat untuk menghasilkan prediksi. Prosedur algoritma regresi KNN secara umum adalah sebagai berikut:

1. Menentukan nilai k (jumlah tetangga terdekat).
2. Menghitung jarak antara data uji dan seluruh data pelatihan, menggunakan metrik seperti Euclidean distance.
3. Mengurutkan hasil jarak secara naik dan memilih k data dengan jarak terdekat.
4. Mengambil rata-rata nilai target dari k tetangga terdekat tersebut.
5. Menetapkan rata-rata tersebut sebagai hasil prediksi.

Prediksi akhir dihitung menggunakan rumus:

$$\hat{y} = \frac{1}{k}\sum_{i=1}^{k} yi$$

Keterangan:

$\hat{y}$ = nilai hasil prediksi

yi = nilai target dari tetangga ke-i

k = jumlah tetangga terdekat

Pemilihan nilai k sangat memengaruhi performa model. Nilai k yang terlalu kecil

dapat menyebabkan prediksi terlalu sensitif terhadap noise, sedangkan k yang terlalu besar dapat menyebabkan prediksi menjadi terlalu rata sehingga kehilangan pola penting dalam data. Oleh karena itu, perlu dilakukan uji coba untuk menentukan nilai k yang optimal.

Selain itu, regresi KNN dapat diperluas untuk menangani multi-target regression, yaitu prediksi lebih dari satu variabel target secara bersamaan. Hal ini dimungkinkan dengan membungkus regresor dasar menggunakan struktur seperti *MultiOutputRegressor*, yang memungkinkan model mempelajari hubungan antara fitur input dengan lebih dari satu target sekaligus (Srisuradetchai & Suksrikran, 2024).

2.2.3 Volume dan Nilai Produksi Ikan

Produksi ikan merupakan salah satu faktor utama yang menentukan nilai ekonomi sektor perikanan. Volume produksi ikan dapat bervariasi tergantung pada lokasi penangkapan dan jenis ikan yang ditangkap. Selain itu, nilai produksi yang diperoleh dari hasil penjualan ikan mencerminkan aspek ekonomi dari kegiatan penangkapan. Volume dan nilai produksi ikan menjadi indikator keberhasilan usaha penangkapan. Kedua indikator ini digunakan untuk mengevaluasi kelayakan usaha perikanan serta tingkat pendapatan nelayan. Suatu usaha penangkapan ikan dikatakan optimal apabila terjadi peningkatan baik dari sisi volume maupun nilai produksi. Faktor seperti harga ikan dan jumlah hasil tangkapan ikan memiliki pengaruh signifikan terhadap pendapatan nelayan (Khoerunnisa et al. 2025).

2.2.4 Machine Learning dan Data Mining

Sejak 1950-an, para peneliti telah mengembangkan algoritma yang memungkinkan mesin belajar dari data. Arthur Samuel, salah satu pelopor dalam bidang ini, mendefinisikan *machine learning* sebagai “bidang studi yang memberikan komputer kemampuan untuk belajar tanpa diprogram secara eksplisit”. Definisi ini menekankan bagaimana sistem dapat beradaptasi dan meningkatkan kinerjanya melalui pembelajaran dari data historis. Di era digital saat ini, perkembangan *machine learning* semakin pesat dan menjadi landasan bagi berbagai

inovasi teknologi, seperti sistem pengenalan wajah, kendaraan otonom, asisten virtual, dan sistem rekomendasi. Dengan terus bertambahnya volume dan kompleksitas data, *machine learning* memainkan peran penting dalam pengolahan informasi, mendukung proses pengambilan keputusan, serta memberikan prediksi terhadap tren di masa depan dengan lebih akurat (Bintoro et al. 2024).

Sementara itu, *Data mining* merupakan proses ekstraksi informasi berharga dari kumpulan data yang besar dan kompleks. Tujuannya adalah untuk mengidentifikasi pola, hubungan, atau wawasan tersembunyi yang tidak dapat diperoleh melalui analisis biasa. *Data mining* memungkinkan analisis data yang lebih mendalam guna mendukung pengambilan keputusan yang lebih tepat. Beberapa manfaat utama *data mining* meliputi peningkatan kualitas pengambilan keputusan, penemuan pola dan tren, prediksi terhadap kondisi di masa depan, segmentasi pelanggan, efisiensi operasional, pengelolaan risiko, serta peningkatan layanan terhadap pelanggan.

Untuk menjalankan proses *data mining* secara sistematis, digunakan tahapan yang dikenal sebagai *Knowledge Discovery in Databases* (KDD). KDD merupakan proses menyeluruh yang mencakup berbagai tahapan mulai dari pengumpulan data hingga interpretasi hasil. Adapun tahapan-tahapan dalam proses KDD menurut Rahayu et al. (2024) meliputi:

1. **Seleksi dan Pemahaman Data (*Selection and Preprocessing*)**
   Memilih data yang relevan untuk analisis data dan memahami karakteristik awal data. Proses ini melibatkan pembersihan data, pengelolaan nilai yang hilang dan praproses data untuk tahap analisis berikutnya.
2. **Pemahaman Data (*Data Cleaning and Understanding*)**
   Data yang dipilih dilakukan analisis lebih lanjut yang melibatkan pemahaman statistik sederhana, visualisasi data dan analisis deskriptif untuk mengidentifikasi pola yang perlu ditangani.
3. **Transformasi Data (*Data Transformation*)**
   Proses ini melibatkan konversi dan transformasi data, mencakup normalisasi data, konversi format atau pembentukan fitur baru yang lebih relevan untuk analisis.

4. **Pembentukan Model atau Pemodelan (*Data Mining*)**

   Menerapkan berbagai teknik seperti klasifikasi, klastering, regresi atau asosiasi untuk mengekstrak pola dari data. Model prediktif atau deskriptif dibangun selama tahap ini.

6. **Evaluasi Model (*Model Evaluation*)**

   Setelah model dibangun, evaluasi dilakukan untuk menilai sejauh mana model efektif dan dapat diandalkan.

7. **Interpretasi dan Visualisasi Hasil (*Interpretation and Visualization*)**

   Hasil analisis *data mining* diinterpretasikan dalam konteks bisnis atau ilmiah. Visualisasi digunakan untuk membantu pemahaman dan komunikasi hasil dengan pemangku kepentingan.

8. **Penggunaan Pengetahuan yang Ditemukan (*Knowledge Utilization*)**

   Pola yang diidentifikasi selama proses KDD digunakan untuk mendukung keputusan atau tindakan yang lebih baik.

2.2.5 Python

Python merupakan bahasa pemrograman fungsional, prosedural, dan berorientasi objek yang dikembangkan oleh Guido van Rossum pada akhir 1980-an. Python banyak diminati karena sintaksnya sederhana dan fleksibel. Saat ini, Python banyak digunakan untuk pengembangan perangkat lunak, pengembangan web, serta *Graphical User Interface* (GUI). Keunggulan dari bahasa pemrograman Python adalah bersifat serbaguna, gratis (*open source*), dan memiliki sintaks yang mudah dipahami. Python juga merupakan salah satu bahasa pemrograman dengan pertumbuhan tercepat. Sejak tahun 2003, Python konsisten menempati peringkat sepuluh besar bahasa pemrograman terpopuler (Anam et al. 2023).

Pengembangan Python terus dilakukan oleh komunitas yang tergabung dalam Python Software Foundation, sebuah organisasi nirlaba yang memegang hak cipta atas Python sejak versi awal. Python merupakan bahasa pemrograman tingkat tinggi yang mendukung eksekusi multi-guna secara interpretatif dengan pendekatan *Object-Oriented Programming* (OOP). Sebagai bahasa pemrograman tingkat tinggi, Python mudah dipelajari dan memiliki manajemen memori otomatis. Python

memungkinkan pengembangan aplikasi yang cepat (*Rapid Application Development*) karena didukung oleh berbagai pustaka, baik standar maupun eksternal, seperti NumPy, Pandas, Seaborn, dan lainnya (Rahcmadani et al. 2020).

2.2.6 Root Mean Squared Error (RMSE)

Root Mean Squared Error (RMSE) adalah metrik evaluasi untuk mengukur perbedaan antara nilai prediksi yang dihasilkan oleh suatu model dengan nilai aktual. RMSE berfungsi untuk mengukur secara kuantitatif besarnya kesalahan prediksi dalam satuan yang sama dengan data aktual, sehingga memudahkan interpretasi tingkat akurasi model. RMSE memiliki akurasi yang tinggi dalam melakukan pengukuran dengan probabilitas hingga 50% pada algoritma yang digunakan untuk memprediksi data, berikut persamaan rumus (Tri Wijaya et al. 2023):

$$RMSE = \sqrt{\sum \frac{(y' - y)^2}{n}}$$

Keterangan:

Y' = Nilai prediksi

Y = Nilai sebenarnya

n = Jumlah data

2.2.7 Mean Absolute Error (MAE)

Mean Absolute Error (MAE) adalah metrik evaluasi yang digunakan untuk mengukur rata-rata kesalahan absolut antara nilai aktual dan nilai prediksi oleh model. Semakin kecil nilai MAE, semakin akurat model tersebut dalam melakukan prediksi. MAE memberikan gambaran yang sederhana namun efektif tentang seberapa besar kesalahan prediksi dalam satuan yang sama dengan data aslinya.
Berikut persamaan untuk menghitung nilai MAE (Yoshua et al. 2023):

$$MAE = \frac{1}{n} \sum | yi - y'i|$$

Keterangan:

n = Jumlah data yang diprediksi

y_i = Nilai aktual data ke-i

y'_i = Nilai prediksi data ke-i

2.2.8 R-squared (R^2)

R-squared (R^2) adalah ukuran proporsi variabilitas data asli yang dapat dijelaskan oleh model. Nilai R^2 digunakan untuk menilai seberapa baik model regresi menyesuaikan dengan data, yaitu seberapa dekat garis regresi dengan data asli. Semakin tinggi nilai R^2, semakin baik model dalam menjelaskan variasi data. Persamaan menghitung nilai R^2 (Aji Saputra et al. 2024):

$$R^2 = \frac{\sum\left(Y_i - \hat{Y}_1\right)^2}{\sum(Y_i - \bar{Y}_i)^2}$$

Keterangan:

R^2 = R-squared

Yi = Nilai pengamatan

$\hat{Y}_1$ = Nilai Y

$\bar{Y}_i$ = Nilai rata-rata pengamatan

2.2.9 TPI DKI Jakarta

Pelabuhan Perikanan Samudera Nizam Zachman (PPSNZ) merupakan pelabuhan perikanan tipe A terbesar di Indonesia yang berlokasi di DKI Jakarta. PPSNZ merupakan pelabuhan ikan berskala internasional, dengan jumlah kapal yang berpangkalan sebanyak 1.545 unit pada tahun 2023. Jenis kapal yang paling dominan adalah Kapal Purse Seine, dengan jumlah armada mencapai 339 unit. PPSNZ memiliki fasilitas untuk melayani kapal berukuran lebih dari 60 GT (*Gross Tonnage*) dan kapal-kapal ikan yang beroperasi di perairan lepas pantai, Zona Ekonomi Eksklusif (ZEE), serta laut lepas (perairan internasional). Jumlah ikan yang didaratkan mencapai sekitar 40.000 ton per tahun. Selain itu, pelabuhan ini juga menyediakan layanan ekspor dan area industri untuk mendukung kegiatan pengolahan hasil perikanan (Suherman et al. 2022; Wicaksono et al. 2025).

Muara Angke merupakan kawasan perikanan yang terletak di Kecamatan Penjaringan, Jakarta Utara. Kawasan ini dibangun pada tahun 2004 dengan luas 3,4 hektare dan memiliki kapasitas tampung hingga 50 unit kapal. Muara Angke

dikembangkan sebagai kawasan nelayan karena lokasinya yang strategis di Pantai Utara Jawa. Pada tahun 2023, jumlah kapal yang keluar-masuk tercatat mencapai 1.500 unit, dengan 80% di antaranya berukuran 3 – 30 GT dan 20% berukuran lebih dari 30 GT. Jenis kapal yang paling dominan adalah kapal pengangkut ikan, yaitu sebanyak 1.599 unit, sedangkan kapal penangkap ikan tercatat sebanyak 112 unit. Mayoritas kapal di Muara Angke memiliki 11 – 20 ABK (Anak Buah Kapal), yaitu sebanyak 1.010 unit. Selanjutnya, terdapat 678 unit kapal dengan jumlah ≤ 10 ABK, serta 13 unit kapal dalam kategori 21 – 33 ABK, yang merupakan jumlah paling sedikit (Ramadhani et al. 2023; Hutagalung et al. 2025).

Kamal Muara merupakan salah satu basis perikanan tangkap di Jakarta Utara. Saat ini, terdapat 7 perahu yang dioperasikan oleh nelayan pengguna jaring tongkol, serta 70 orang nelayan bagan tancap, 7 orang bagan apung, dan 100 orang pengguna alat tangkap sero. Selain itu, telah terbentuk Kelompok Usaha Bersama (KUB) yang beranggotakan 15 – 20 nelayan per kelompok (Anggraini et al. 2021).

Kabupaten Administrasi Kepulauan Seribu memiliki dua kecamatan, salah satunya adalah Kecamatan Kepulauan Seribu Selatan yang merupakan daerah dengan potensi tinggi di bidang perikanan dan berperan penting sebagai pemasok bahan baku ikan. Rata-rata nelayan menggunakan kapal berukuran 2 – 5 GT dengan jenis alat pancing seperti kotrekan, tonda, dan gala setan. Penangkapan ikan umumnya dilakukan pada siang hari dengan durasi 10 – 12 jam, dan jenis ikan yang diperoleh antara lain ikan kurisi, kembung, dan selar (Wiryati et al. 2021).

Kecamatan Cilincing merupakan salah satu kecamatan penghasil perikanan tangkap di Kota Administrasi Jakarta Utara. Kecamatan ini berbatasan langsung dengan pesisir DKI Jakarta dan didominasi oleh aktivitas industri, kepelabuhan, serta perikanan. Produksi ikan dari TPI Cilincing pada tahun 2017 tercatat sebesar 5.892.185 kg, menurun pada tahun 2019 menjadi 4.942.117 kg, kemudian meningkat kembali pada tahun 2021 menjadi 13.949.640 kg (Wahyuni et al., 2022).