BAB II
TINJAUAN PUSTAKA

2.1 Penelitian Terdahulu

Penelitian ini tentunya tidak lepas dari sejumlah penelitian yang telah dilakukan oleh para peneliti sebelumnya, sehingga terdapat kesamaan maupun perbedaan dalam subjek yang diteliti. Dalam menyusun penelitian ini, diperlukan referensi yang memadai untuk jadikan rujukan dalam penelitian ini. Penelitian terdahulu terkait metode *K-Nearest Neighbor* yang digunakan sebagai rujukan yaitu:

1) Penelitian oleh (Seruni et al., 2020) dari Universitas Brawijaya, dengan judul "**Sistem Prediksi Pertumbuhan Jumlah Penduduk Kota Malang menggunakan Metode *K-Nearest Neighbor Regression***". Penelitian ini bertujuan membantu pemerintah dalam perencanaan kota untuk memanfaatkan sumber daya secara optimal. Hasil menunjukkan bahwa *K-NN Regression* dapat memprediksi pertumbuhan penduduk dalam jangka pendek (1-3 tahun).

   Perbedaan dari penelitian yang dilakukan oleh penulis dengan penelitian oleh Seruni adalah pada penelitian oleh Seruni dalam melakukan evaluasi akurasi hasil prediksi *K-Nearest Neighbor* hanya menggunakan satu *metric* saja yaitu MAPE. Sedangkan, penelitian yang dilakukan penulis saat ini dalam melakukan evaluasi akurasi hasil prediksi menggunakan dua *metric* yaitu MAPE dan RMSE.

2) Penelitian oleh (Utomo et al., 2019) dari Universitas Amikom Yogyakarta, dengan judul "**Implementasi Metode *K-Nearest Neighbor* Dan Regresi Linear Dalam Prediksi Harga Emas**". Penelitian ini bertujuan memahami fluktuasi harga emas dan meminimalkan risiko investasi. Hasil menunjukkan bahwa menggabungkan regresi linear dengan K-NN meningkatkan akurasi prediksi, terlihat dari penurunan nilai RMSE.

   Perbedaan dari penelitian yang dilakukan oleh penulisan dengan penelitian oleh Utomo adalah pada penelitian oleh Utomo tidak hanya menggunakan *K-Nearest Neighbor* tetapi juga menggunakan Regresi Linear dalam

prediksi. Tak hanya itu, penelitian oleh Utomo hanya menggunakan satu *metric* dalam mengevaluasi tingkat akurasi hasil prediksi yaitu RMSE. Sedangkan, penulis dalam penelitian ini hanya berfokus pada metode *K-Nearest Neighbor* dan dalam mengukur tingkat akurasi menggunakan dua *metric* yaitu MAPE dan RMSE.

3) Penelitian oleh (Dharmali et al., 2023) dari Universitas Tarumanegara, dengan judul "**Sistem Prediksi Penjualan Triplek Menggunakan Metode Regresi Time Series *K-Nearest Neighbor* (KNN) Pada Toko Makmur Cibubur**". Penelitian ini menggunakan data time series dan mencapai tingkat akurasi rata-rata 62,71% untuk tujuh jenis triplek. Metode ini dapat menjadi solusi untuk membantu toko dalam memprediksi penjualan. Penelitian ini menggunakan bantuan bahasa pemrograman *python* dalam melakukan pengolahan data.

   Perbedaan dari penelitian yang dilakukan oleh penulis dengan penelitian oleh Dharmali adalah pada penelitian oleh Dharmali dalam melakukan evaluasi akurasi hasil prediksi *K-Nearest Neighbor* hanya menggunakan satu metrik saja yaitu MAPE. Sedangkan, penelitian yang dilakukan penulis saat ini dalam melakukan evaluasi akurasi hasil prediksi menggunakan dua *metric* yaitu MAPE dan RMSE.

4) Penelitian oleh (Qirani & Sukarsih, 2024) dari Universitas Islam Bandung, dengan judul "**Penerapan Metode *K-Nearest Neighbor* untuk Prediksi Harga Gas Alam Menggunakan *Python***". Penelitian ini mengeksplorasi pengaruh penghapusan data pencilan terhadap akurasi prediksi. Hasil menunjukkan bahwa tanpa data pencilan, MAPE prediksi harga gas alam lebih baik. Penelitian ini menggunakan bantuan bahasa pemrograman *python* dalam melakukan pengolahan data.

   Perbedaan dari penelitian yang dilakukan oleh penulis dengan penelitian oleh Qirani & Sukarsih adalah pada penelitian oleh Qirani & Sukarsih dalam melakukan peramalan menggunakan *K-Nearest Neighbor*, terdapat perhitungan tiga jarak yaitu *Euclidean*, *Manhattan*, dan *Chebyshev*. Sedangkan, pada penelitian yang dilakukan oleh penulis saat ini hanya menggunakan jarak *Euclidean*.

5) Penelitian oleh (Januar et al., 2023) dari STMIK Primakarya, dengan judul "**Analisis Dan Prediksi Penutupan Harga Saham Pada PT. Adaro Energy Indonesia TBK Menggunakan Algoritma *K-Nearest Neighbor Regression***". Penelitian ini bertujuan membantu investor memprediksi pergerakan harga saham. Hasil menunjukkan bahwa penggunaan 11 atribut lebih efektif dibandingkan 12 atribut, dengan RMSE 35,02 dan R-squared 0,99, yang menekankan pentingnya pemilihan atribut dalam model prediksi. Penelitian ini menggunakan *tools Google Colaboratory* dalam melakukan prediksi atau pengolahan data.

   Perbedaan dari penelitian yang dilakukan oleh penulis dengan penelitian oleh Januar adalah pada penelitian oleh Januar dalam melakukan evaluasi akurasi hasil prediksi K-Nearest Neighbor menggunakan RMSE dan *R-Squared*. Sedangkan, penelitian yang dilakukan oleh penulis menggunakan RMSE dan MAPE dalam mengukur tingkat akurasi hasil prediksi.

6) Penelitian oleh (Susilo et al., 2024) dari Universitas Muhadi Setiabudi, Brebes yang berjudul "**Penerapan Algoritma *K-Nearest Neighbor* untuk Prediksi Penjualan Produk Digital**". Penelitian ini bertujuan untuk menentukan produk terlaris berdasarkan data transaksi dari Agustus 2023 hingga Januari 2024 dengan menerapkan metode *K-Nearest Neighbor*. Hasil dari penelitian ini yaitu penerapan metode *K-Nearest Neighbor* (K-NN) dengan nilai K = 3 berhasil memprediksi jumlah transaksi produk digital untuk dua bulan ke depan. Hasil uji coba menunjukkan bahwa algoritma *K-Nearest Neighbor* (K-NN) efektif dalam mengklasifikasikan data transaksi produk digital di PT. Global Indo Multimedia. Pengujian ini menghasilkan akurasi sebesar 95,24%, yang menandakan bahwa *dataset* tersebut valid untuk digunakan pada tahapan berikutnya. Dengan tingkat akurasi yang cukup tinggi ini, model *K-Nearest Neighbor* (K-NN) dapat menjadi solusi yang handal untuk memprediksi transaksi produk digital berdasarkan data penjualan yang sudah ada. Klasifikasi yang dilakukan dalam penelitian ini menggunakan bantuan *tools* RapidMiner untuk memprediksi penjualan produk digital berdasarkan *data testing*.

Perbedaan penelitian yang dilakukan oleh Susilo dengan penelitian yang dilakukan oleh penulis yaitu prediksi yang dilakukan untuk klasifikasi hasil prediksi penjualan apakah rugi atau tidak. Selain itu, evaluasi akurasi yang digunakan yang terbatas pada *metric* RMSE. Sedangkan, penelitian yang dilakukan oleh penulis yaitu prediksi yang dilakukan untuk memprediksi harga di kondisi tertentu dan tahun tertentu. Selain itu evaluasi akurasi yang digunakan yaitu MAPE dan RMSE.

7) Penelitian oleh (Rahmadini et al., 2023) dari Universitas Muhammadiyah Sumatera Utara yang berjudul "**Penerapan Data *Mining* untuk Memprediksi Harga Bahan Pangan di Indonesia Menggunakan algoritma *K-Nearest Neighbor* (KNN)**". Penelitian ini bertujuan untuk mengimplementasikan algoritma regresi *K- Nearest Neighbor* (KNN) dalam memprediksi harga Komodias beras dan menganalisis algoritma regresi *K-Nearest Neighbor* (K-NN) dengan metode lain untuk memprediksi harga Komodias beras. Dimana hasil penelitian nya, metode *K-Nearest Neighbor* dapat melakukan prediksi harga beras untuk periode Januari 2019 hingga Desember 2021 dilakukan berdasarkan data bulanan. Selain itu, variabel tambahan seperti luas panen (hektar) dan hasil produksi (ton) juga digunakan dalam analisis. Dari berbagai percobaan dengan rentang nilai k antara 2 hingga 10, prediksi terbaik diperoleh dengan nilai k = 2. Pada *data training*, nilai MAE dan RMSE masing-masing sebesar 52,77 dan 96,40, sedangkan pada *data testing* mencapai 55,55 dan 81,64. Nilai k = 2 tersebut telah melalui proses normalisasi agar dapat menghasilkan prediksi yang lebih akurat.

   Perbedaan dari penelitian yang dilakukan oleh penulis dengan penelitian oleh Rahmadini adalah pada penelitian oleh Rahmadini dalam melakukan evaluasi akurasi hasil prediksi *K-Nearest Neighbor* menggunakan satu *metric* saja yaitu MAE dan RMSE. Sedangkan, penelitian yang dilakukan penulis saat ini dalam melakukan evaluasi akurasi hasil prediksi menggunakan dua *metric* yaitu MAPE dan RMSE.

8) Penelitian oleh (Rofiq et al., 2020) dari Unisa Gorontalo yang berjudul "**Penerapan *Data Mining* untuk Menentukan Potensi Hujan Harian**

dengan Menggunakan Algoritma K-*Nearest Neighbor* (K-NN)". Penelitian ini bertujuan untuk menentukan informasi cuaca, sehingga informasi tersebut dapat dimanfaatkan secara maksimal oleh Publik dengan menggunakan algoritma *K-Nearest Neighbor* (K-NN). Hasil dari penelitian ini bahwa Penerapan *Data mining* untuk menentukan potensi hujan harian dengan menggunakan algoritma *K-Nearest Neighbor* (KNN) dapat di klasifikasikan, dengan nilai akurasi 12.493+/-0.000. Sehingga penelitian tersebut dapat membantu dan memudahkan masyarakat dalam mengetahui informasi potensi hujan yang tidak membingungkan. Pada penelitian ini menggunakan RapidMiner untuk membantu proses analisis data.

Perbedaan dari penelitian yang dilakukan oleh penulis dengan penelitian oleh Rofiq adalah pada penelitian oleh Rofiq dalam melakukan evaluasi akurasi hasil prediksi *K-Nearest Neighbor* hanya menggunakan satu *metric* saja yaitu RMSE. Sedangkan, penelitian yang dilakukan penulis saat ini dalam melakukan evaluasi akurasi hasil prediksi menggunakan dua *metric* yaitu MAPE dan RMSE.

Berdasarkan beberapa penelitian terdahulu diatas yang dijadikan rujukan oleh penulis dapat dilihat bahwa terdapat 2 (dua) penelitian yang menggunakan RapidMiner, 2 (dua) penelitian menggunakan *Python*, 1 (satu) menggunakan Google Colaboratory, dan 3 (tiga) tidak menyebutkan *tools* apa yang digunakan dalam melakukan pengolahan data. Sehingga dapat disimpulkan bahwa, RapidMiner dan *Python* umum digunakan dalam melakukan klasifikasi untuk peramalan data. Selain itu, dalam penelitian ini memiliki persamaan dengan penelitian terdahulu yang dijadikan rujukan oleh penulis yaitu keduanya sama sama menggunakan metode *K-Nearest Neighbor* untuk melakukan prediksi atau peramalan pada suatu objek.

2.2 Landasan Teori

Penelitian ini menggunakan landasan teori yang berkaitan dengan peramalan harga komoditi beras premium menggunakan algoritma *K-Nearest Neighbor* (KNN) Regression di Kota Surabaya.

2.2.1 Peramalan

Peramalan secara umum didefinisikan sebagai proses sistematis untuk memprediksi kemungkinan kejadian di masa depan berdasarkan data atau informasi historis dan kondisi saat ini. Tujuan dari peramalan adalah meminimalkan potensi kesalahan dalam pengambilan keputusan atau strategi yang akan diterapkan (Wardana, 2022). Menurut John E. Biegel (1999), peramalan adalah proses analisis untuk memprediksi tingkat permintaan suatu produk atau beberapa produk pada periode waktu tertentu di masa depan. Peramalan bertujuan memberikan gambaran yang akurat terkait kebutuhan pasar, sehingga pelaku bisnis atau pihak terkait dapat mempersiapkan strategi produksi, distribusi, dan pemasaran dengan lebih tepat. Dalam konteks manajemen, peramalan berperan penting dalam pengambilan keputusan jangka pendek maupun jangka panjang agar perusahaan atau organisasi mampu beradaptasi terhadap perubahan permintaan pasar yang dinamis (Hassyddiqy & Hasdiana, 2023). Berdasarkan definisi di atas, peramalan adalah proses sistematis untuk memprediksi kejadian atau permintaan di masa depan guna mendukung pengambilan keputusan yang tepat dan meminimalkan kesalahan berdasarkan data historis dan kondisi terkini.

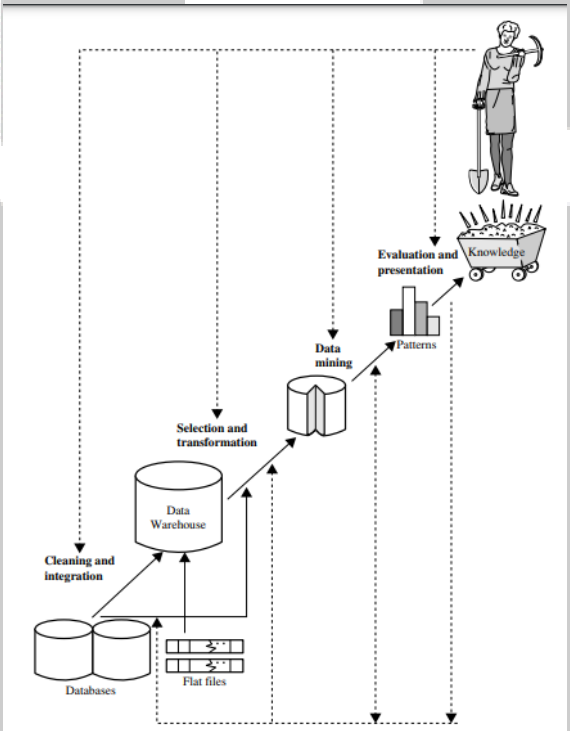
2.2.2 *Data Mining*

Data mining adalah proses menggali pengetahuan atau informasi berharga dari kumpulan data berukuran besar dan kompleks. Proses ini bertujuan untuk menemukan pola, hubungan, atau informasi tersembunyi yang tidak mudah terlihat, sehingga dapat memberikan wawasan yang lebih mendalam dan bermanfaat (Rahayu et al., 2018). Hasil dari *data mining* dapat dimanfaatkan untuk mendukung proses pengambilan keputusan, mengenali tren pasar, meningkatkan efisiensi operasional, dan merumuskan strategi bisnis. Berikut ini beberapa definisi *data mining* menurut para ahli:

1) Menurut Han dan Kamber, *data mining* adalah proses menemukan pola atau informasi berharga dari kumpulan data berukuran besar dengan menerapkan berbagai metode, seperti teknik statistik, matematika, dan kecerdasan buatan.
2) Menurut Berry dan Linoff, *data mining* adalah proses untuk menemukan pola yang berharga atau informasi baru dalam basis data yang besar dengan memanfaatkan algoritma pencarian atau algoritma pembelajaran mesin.

3) Menurut Fayyad, Piatetsky-Shapiro, dan Smyth, *Data mining* adalah proses menggali pengetahuan berharga, pola, atau informasi yang tersembunyi dari sejumlah besar data.

Menurut (Ha et al., 2011) menjelaskan bahwa penggalian data atau informasi terdiri dari beberapa tahapan. Beberapa tahapan dalam proses *data mining* sebagai langkah dalam proses penemuan pengetahuan dapat dilihat pada gambar 2.1 dibawah ini.



Gambar 2. 1 Tahapan *Data Mining*

(Sumber: Jiawei Han, 2011:44)

1) *Data Cleaning*

Data Cleaning atau pembersihan data adalah proses yang digunakan untuk menghilangkan data *noise* dan data yang tidak konsisten dari berbagai basis data yang memiliki format yang berbeda. Tahapan ini bertujuan untuk memastikan bahwa data dari berbagai sumber tersebut dapat diintegrasikan dengan baik dalam satu *database* atau data *warehouse* (Wicaksono & Setiadi, 2023).

2) *Data Integration*

Integrasi data adalah proses mengumpulkan data dari berbagai tabel ke dalam satu tabel baru. Seringkali, data yang dibutuhkan untuk *data mining*

tidak hanya berasal dari satu tabel, tetapi juga dari beberapa tabel atau file teks (Muhammad Syafrullah, 2021).

3) Data Selection

Data Selection adalah tahapan dalam *data mining* yang bertujuan untuk memilih data relevan yang akan digunakan dalam analisis, sementara data yang tidak dibutuhkan atau kurang relevan akan diabaikan. Langkah ini penting untuk meningkatkan efisiensi dan akurasi proses *data mining* dengan hanya memfokuskan pada informasi yang signifikan bagi tujuan analisis, sekaligus mengurangi beban komputasi dan potensi gangguan dari data yang tidak diperlukan (Saputro & Sari, 2020).

4) *Data Transformation*

Data Transformation adalah tahap di mana data yang telah dipilih diubah ke dalam format yang sesuai untuk digunakan dalam prosedur penggalian informasi (*mining procedure*). Proses ini melibatkan langkah-langkah seperti normalisasi, yang bertujuan menyelaraskan data agar memiliki skala yang konsisten, dan agregasi, di mana data diringkas atau digabungkan untuk memudahkan analisis lebih lanjut (Zai, 2022).

5) *Data Mining*

Tahapan utama dalam proses ekstraksi pola data untuk menemukan informasi atau pengetahuan berharga dari data dan mendapatkan pola yang diinginkan (Wicaksono & Setiadi, 2023).

6) *Pattern Evaluation*

Tahapan untuk mengidentifkasi pola yang paling menggambarkan data yang digali informasinya.

7) *Knowledge Presentation*

Tahapan visualisasi dan representasi adalah tahap akhir di mana pengetahuan yang telah ditemukan ditampilkan secara visual kepada pengguna. Tahap ini sangat penting karena memanfaatkan teknik visualisasi untuk membantu pengguna memahami dan menginterpretasikan hasil *data mining* dengan lebih mudah dan intuitif. Visualisasi yang efektif dapat berupa grafik, diagram, atau *dashboard* interaktif yang memungkinkan pengguna

mendapatkan wawasan mendalam dari data secara cepat dan jelas (Zai, 2022).

Data mining bertujuan untuk mengungkap pola dan pengetahuan yang bermanfaat dari data, serta dapat digunakan untuk melakukan prediksi, pengoptimalan, segmentasi pelanggan, deteksi penipuan, dan berbagai aplikasi lainnya. Pada dasarnya, *data mining* menyediakan kerangka kerja teoretis dan alat untuk mengekstraksi informasi dari data (SLN, 2023).

2.2.3 Algoritma

Algoritma adalah prosedur komputasi yang terdefinisi secara jelas, di mana sejumlah nilai atau kumpulan nilai digunakan sebagai masukan dan diproses untuk menghasilkan keluaran berupa nilai atau kumpulan nilai tertentu. Secara sederhana, memahami algoritma dapat diibaratkan seperti merancang rute perjalanan untuk mencapai tujuan dengan cara yang paling efisien dan tepat. Proses ini memastikan setiap langkah diambil secara terstruktur, sehingga hasil akhirnya sesuai dengan yang diharapkan. Seperti mencari jalan terpendek menuju destinasi, algoritma membantu menemukan solusi yang optimal dan efektif dalam berbagai konteks masalah (Ni Nyoman Emang Smrti et al., 2023).

Istilah algoritma berasal dari nama Al-Khawarizmi, merujuk pada karya Abu Ja'far Muhammad Ibnu Musa Al-Khawarizmi, seorang matematikawan dari Persia. Dalam bukunya yang berjudul "Aljabar wal Muqabala," ia memberikan beberapa penjelasan mengenai konsep algoritma. Algoritma didefinisikan sebagai rangkaian langkah-langkah terstruktur yang digunakan untuk menyelesaikan suatu masalah dengan cara yang sistematis dan logis. Dua kata kunci dalam definisi ini adalah "sistematis" dan "logis." "Sistematis" mengacu pada proses yang tersusun secara teratur dan berurutan, sehingga setiap langkah memiliki keterkaitan dan dapat diikuti dengan jelas dari awal hingga akhir. Sementara itu, "logis" berarti setiap langkah dalam algoritma disusun berdasarkan prinsip pemikiran rasional, memastikan bahwa solusi yang dihasilkan masuk akal dan efektif. Kombinasi kedua elemen ini menjamin bahwa algoritma tidak hanya mudah dipahami tetapi juga dapat diterapkan secara konsisten untuk menyelesaikan berbagai masalah (Asih et al., 2020).

2.2.4 Time Series

Time series (deret waktu) adalah kumpulan data yang disusun sesuai dengan urutan waktu saat data tersebut dicatat. Frekuensi waktu yang digunakan bisa berupa tahunan, bulanan, mingguan, harian, atau bahkan per jam. Sebuah time series dapat memiliki satu fitur (*univariate*) atau lebih dari satu fitur *(multivariate*) (Dharmali et al., 2023). Secara umum, time series memiliki tiga pola (Athanasopoulos, 2018), yaitu:

1) *Trend* (Tren)

   Tren muncul ketika terdapat perubahan jangka panjang yang menunjukkan peningkatan atau penurunan dalam data. Tren ini tidak selalu bersifat linier.

2) *Seasonal* (Musim)

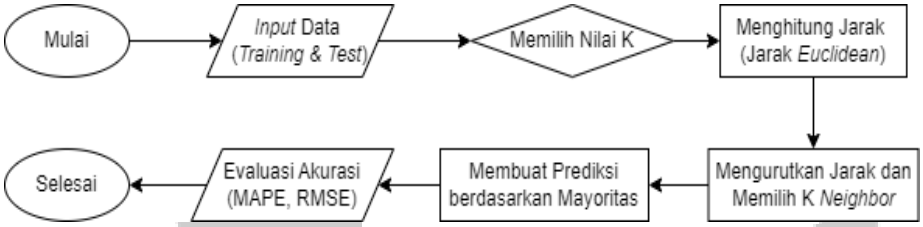
   Pola musiman muncul ketika deret waktu dipengaruhi oleh faktor-faktor musiman, seperti waktu dalam tahun atau hari dalam minggu. Musiman selalu memiliki frekuensi yang tetap dan sudah diketahui.

3) *Cyclic* (Siklus)

   Siklus terjadi ketika data menunjukkan variasi kenaikan dan penurunan yang tidak mengikuti frekuensi tetap. Fluktuasi ini biasanya disebabkan oleh kondisi ekonomi dan seringkali berkaitan dengan "siklus bisnis." Umumnya, durasi fluktuasi ini berlangsung setidaknya 2 tahun.

2.2.5 K-Nearest Neighbor (K-NN)

Algoritma K-Nearest Neighbor merupakan salah satu algoritma yang terdapat dalam teknik klasifikasi. Algoritma *K-Nearest Neighbor* (K-NN) adalah sebuah algoritma yang terkenal karena kemampuannya dalam memprediksi pola berdasarkan data historis (Mustafa & Simpen, 2019). Metode *K-Nearest Neighbor* (KNN) memprediksi kategori dengan memanfaatkan hubungan jarak antara tetangga terdekat. Dalam analisis data, jarak antar tetangga terdekat dapat dihitung menggunakan dua metrik utama, yaitu jarak Manhattan dan jarak Euclidean (Aprihartha et al., 2024). Tahapan dari algoritma *K-Nearest Neighbor* (K-NN) dapat dilihat pada gambar 2.2 dibawah ini.

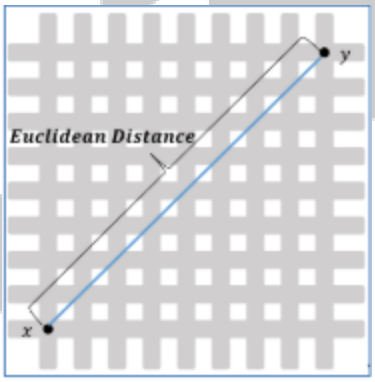


Gambar 2. 2 Tahapan *K-Nearest Neighbor*

Diagram di atas menggambarkan alur proses dari algoritma *K-Nearest Neighbor* (K-NN) untuk peramalan dalam klasifikasi data. Proses dimulai dengan memasukkan data, yang terdiri dari data pelatihan dan data pengujian, ke dalam sistem. Selanjutnya, ditentukan nilai K, yaitu jumlah tetangga terdekat yang akan digunakan dalam perhitungan. Setelah itu, sistem menghitung jarak antara data uji dan seluruh data pelatihan menggunakan metode jarak Euclidean. Hasil perhitungan jarak tersebut kemudian diurutkan untuk memilih K tetangga terdekat. Berdasarkan mayoritas label dari tetangga-tetangga tersebut, sistem membuat prediksi atau klasifikasi. Tahap akhir adalah evaluasi akurasi menggunakan metrik seperti MAPE *(Mean Absolute Percentage Error)* dan RMSE *(Root Mean Square Error)* untuk memastikan kualitas hasil peramalan. Setelah evaluasi selesai, proses pun berakhir dengan menghasilkan informasi tentang performa model.

Jarak Manhattan adalah total panjang jalur yang diproyeksikan ke sumbu, yang harus ditempuh langkah demi langkah untuk mencapai titik lain dalam sistem koordinat, seperti yang diperlihatkan pada Gambar 2.3. Sementara itu, jarak Euclidean mengukur jarak garis lurus antara dua titik, seperti yang ditunjukkan pada Gambar 2.4, dan dapat digunakan untuk mengukur data dalam struktur hiperbola sirkuler di ruang berdimensi tinggi.

Gambar 2. 3 Jarak Manhattan

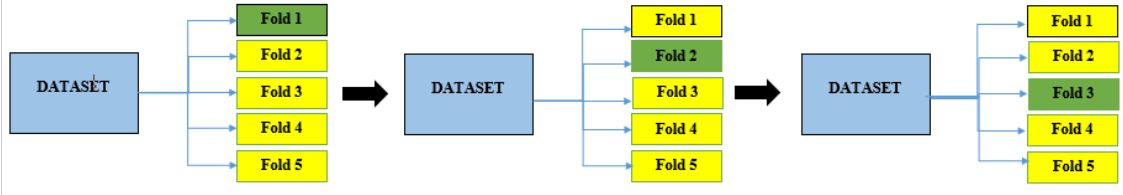


Gambar 2. 4 Jarak Euclidean

(Sumber: Aprihartha, 2024:4)

2.2.6 K-Fold Cross Validation

Salah satu metode yang digunakan dalam melakukan validasi model terbaik adalah *cross validation*. Metode ini menguji keefektifan model yang dibentuk dengan penyusunan ulang pada data untuk dibagi menjadi 2 bagian yaitu *data training* dan *testing* (Widyaningsih et al., 2021). *K-Fold Cross Validation* merupakan salah satu metode yang sering digunakan untuk menghitung akurasi prediksi suatu sistem karena dapat mengurangi waktu komputasi dengan tetap menjaga tingkat keakuratan dari prediksi. Data akan dibagi ke dalam subsets dengan jumlah yang sama dan diulangi sebanyak K subsest dan hasil akurasi klasifikasi yaitu hasil rata-rata dari setiap *data training* dan *data testing*. *K-Fold* yang sering digunakan adalah 3,5,10 dan 20 (Suryati et al., 2023).



Gambar 2. 5 Alur *K-Fold Cross Validation*

Sumber: (Peryanto et al., 2020)

Proses kinerja K-*Fold Cross Validation* yaitu *dataset* akan dibagi menjadi beberapa *fold* / sebuah K bagian di mana setiap *fold* digunakan sebagai set pengujian di beberapa titik. Umumnya pembagian data dibagi menjadi 3, 5, 10, dan 20. Proses pertama, *fold* 1 akan digunakan untuk menguji model dan *fold* sisanya digunakan untuk melatih model. Proses kedua, *fold 2* akan digunakan untuk menguji model dan *fold* sisanya untuk melatih model. Proses ini diulangi hingga setiap *fold* dari nilai *fold* yang ditentukan telah dilakukan pengujian (Peryanto et al., 2020).

2.2.7 Mean Absolute Precentage Error (MAPE)

Mean Absolute Percent Error (MAPE) adalah rata-rata selisih absolut antara nilai prediksi dan nilai aktual, yang disajikan dalam bentuk persentase (Ihzaniah et al., 2023). *Mean Absolute Percent Error* (MAPE) memungkinkan perbandingan akurasi antar model atau antar hasil prediksi dalam konteks yang sama, dengan mengukur persentase rata-rata kesalahan absolut terhadap data aktual.

$$MAPE = \frac{1}{x}\sum \left|\frac{y - y_t}{y}\right| x\ 100\% \qquad (4)$$

y_t = Data hasil peramalan

y = Data Aktual

n = Jumlah Data

Nilai *Mean Absolute Percent Error* (MAPE) berfungsi untuk mengevaluasi tingkat akurasi peramalan (Reba et al., 2021), sebagaimana ditunjukkan dalam tabel 2.1 berikut:

Tabel 2. 1 Akurasi Peramalan MAPE

Nilai (MAPE) *Mean Absolute Percent Error*	**Akurasi Peramalan**
MAPE ≤ 10%	Tinggi
10% ≤ MAPE ≤ 20%	Baik
20% ≤ MAPE ≤ 50%	Layak
MAPE > 50%	Rendah

(Sumber: Felix Reba, 2021:4)

2.2.8 Root Mean Square Error (RMSE)

Root Mean Squared Error (RMSE) adalah akar kuadrat dari rata-rata kuadrat selisih antara nilai aktual dan nilai prediksi, yang digunakan untuk mengukur seberapa besar kesalahan prediksi secara keseluruhan (Ihzaniah et al., 2023). Menurut (Eka et al., 2021), *Root Mean Square Error* (RMSE) atau *Root Mean Square Deviation* adalah metode yang digunakan untuk menghitung tingkat kesalahan (*error*) dalam hasil estimasi. Nilai *error* ini mencerminkan seberapa besar perbedaan atau deviasi antara hasil estimasi dan nilai aktual semakin kecil (mendekati 0) nilai RMSE maka hasil prediksi akan semakin akurat. Tujuan penggunaan RMSE adalah untuk mengukur tingkat kesalahan dalam analisis yang melibatkan metode tertentu, seperti pada *data training* dan *data testing*. *Root Mean Square Error* (RMSE) memberikan bobot lebih pada kesalahan besar dengan menghitung akar rata-rata dari kuadrat kesalahan, sehingga sangat efektif untuk mendeteksi ketidakakuratan yang signifikan dalam hasil prediksi. Berikut ini adalah

persamaan *Root Mean Square Error* (RMSE) yang digunakan untuk menghitung akurasi harga beras premium menggunakan algoritms *K-Nearest Neighbor* (K-NN).

$$RMSE = \sqrt{\frac{\sum(y_t - y_t)^2}{n}} \quad (4)$$

y_t = Nilai Aktual indeks

y_t = Nilai Peramalan

n = Jumlah Sampel

2.2.9 Beras Premium

Beras merupakan salah satu komoditas pangan utama dan sumber karbohidrat bagi lebih dari setengah populasi dunia, terutama di Asia, termasuk Indonesia, memiliki peran sentral dalam ketahanan pangan dan stabilitas ekonomi (Reza et al., 2021). Secara umum, beras di Indonesia dapat dibagi menjadi dua kategori, yaitu beras premium dan beras medium.

Gambar 2. 6 Beras Premium
Sumber: (Kementrian Perdagangan, 2019)

Berdasarkan informasi dari (Hellosehat, 2023), *brand* beras yang termasuk jenis beras premium yaitu, Topi Koki Sentra Ramos, Rojolele, Beras Cap BMW, Beras Cap Bunga, Beras Sumo, Maknyuss Beras Premium, Puregreen Beras Meras, SiPulen, FUKUMI Beras Porang, dan Sundakala Beras Coklat Organik. Selain itu, terdapat merek beras lain yang merupakan jenis beras premium yang diolah dari bahan padi berkualitas jenis bengawan yaitu Beras Pinpin (Jaya Raya, 2025).

Berdasarkan informasi dari (Kementrian Perdagangan, 2019), setiap jenis beras tersebut memiliki ciri-ciri masing-masing yang dapat dilihat pada tabel 2.1 dibawah ini.

Tabel 2. 2 Ciri-ciri Beras Premium dan Medium

Beras Premium	Beras Medium
Warna lebih cerah	Warna lebih gelap
Tidak terdapat butir beras lainnya seperti butir menir atau gabah.	Masih ada butir beras atau gabah
Butir beras patah maksimal 15%	Butir beras patah maksimal 25%

Sumber: (Putra & Sinaga, 2022)

Berdasarkan tabel di atas, dapat disimpulkan bahwa beras premium memiliki mutu yang lebih tinggi dibandingkan beras medium, meskipun kandungan gizi utama dari keduanya relatif serupa. Perbedaan kualitas ini menjadi salah satu faktor yang memengaruhi perbedaan harga antara kedua jenis beras tersebut. Visualisasi dari kedua jenis beras dapat dilihat pada gambar 2.5 dan gambar 2.6.

Gambar 2. 7 Beras Medium
(Sumber: Kementrian Perdagangan, 2019)

2.2.10 R Studio

RStudio adalah lingkungan pengembangan terintegrasi (*Integrated Development Environment*, IDE) untuk bahasa pemrograman R, yang digunakan terutama dalam analisis data dan pemrograman statistik (Nurandi Rachim et al., 2024). R menyimpan data dan fungsi dalam suatu tempat disebut *package* (paket). Ada dua jenis paket R yaitu paket standar yang harus ada dalam setiap perangkat lunak R dan paket yang dikembangkan oleh banyak ahli untuk perluasan komputasi statistik (Amir et al., 2022). Sebagai IDE yang terkenal dalam bidang *data science*, RStudio sangat berguna bagi peneliti, analis data, dan ilmuwan data yang menggunakan R dalam pekerjaan mereka untuk melakukan analisis statistik, visualisasi data, serta pengembangan aplikasi berbasis data (Nur Amalia et al., 2024).