BAB II

TINJAUAN PUSTAKA

2.1 Data Mining

Di era digital yang dipenuhi data, data mining muncul sebagai kekuatan transformatif yang mengubah informasi mentah menjadi wawasan berharga. Data mining adalah suatu proses pencarian data secara otomatis dapat mendapatkan sebuah model dari database yang besar. [6] Prosesnya melibatkan serangkaian langkah, mulai dari pengumpulan dan pembersihan data, hingga penerapan algoritma dan teknik analisis yang canggih. Tujuannya adalah untuk mengungkap pola tersembunyi, tren yang tak terlihat, dan korelasi yang signifikan yang mungkin terlewatkan jika hanya mengandalkan metode analisis tradisional. Manfaat data mining sangat luas, mencakup peningkatan efisiensi operasional, personalisasi pengalaman pelanggan, pengembangan produk yang lebih baik, dan pengambilan keputusan yang lebih akurat. Dalam dunia bisnis, data mining digunakan untuk memperoleh informasi yang berguna dan berdampak pada keputusan bisnis yang lebih baik. [7] Di bidanglain, sektor kesehatan sendiri telah menjadi salah satu penerima manfaat utama dari potensi besar yang dimiliki oleh *big data.* Dengan informasi yang lebih akurat dan cepat, pelayanan kesehatan dapat disesuaikan secara personal untuk memenuhi kebutuhan pasien secara lebih efektif. [8]

Data mining merupakan proses menemukan korelasi baru yang bermanfaat, pola dan trend dengan menambang sejumlah *repository* data dalam jumlah besar, menggunakan teknologi pengenalan pola seperti statistik dan teknik matematika. [9] Istilah data mining kadang disebut juga *Knowledge Discovery* Salah satu teknik yang dibuat dalam data mining adalah bagaimana menelusuri data yang ada untuk membangun sebuah model, kemudian menggunakan model tersebut agar dapat mengenali pola data yang lain yang tidak berada dalam basis data yang tersimpan. Kebutuhan untuk prediksi juga dapat memanfaatkan teknik ini. Dalam data mining, pengelompokan data juga bisa dilakukan. [10] Data Mining dapat dibagi menjadi empat kelompok, yaitu model *Prediction Modelling* (prediksi), *Cluster Analysis* (analisis kelompok), *Association Analysis* (analisis asosiasi) dan *Anomaly*

Detection (deteksi anomali). Data mining adalah proses otomatis yang memanfaatkan teknik statistik, matematika, kecerdasan buatan, dan *machine learning* untuk menemukan informasi berharga dan berguna yang tersembunyi dalam database besar.

2.1.1 Fungsi Data Mining

Data mining sangat penting dalam proses penggalian informasi berharga dari data. Fungsinya yang utama adalah mengekstrak pola, tren, dan informasi penting yang tersembunyi di dalam data dalam data berukuran besar. Wawasan ini sangat berharga untuk berbagai keperluan, termasuk meningkatkan pengetahuan tentang pelanggan, menyempurnakan alur kerja bisnis, menunjang pengambilan keputusan yang tepat, dan merancang strategi yang lebih jitu. Sebagai contoh, dalam bisnis ritel, data mining memungkinkan analisis data penjualan untuk mengetahui produk terlaris, mengidentifikasi pelanggan loyal, dan mencari cara untuk meningkatkan omzet penjualan.

a. Prediction

Prediction atau fungsi prediksi merupakan salah satu fungsi data mining. Maksudnya yaitu dari proses nanti akan menemukan pola tertentu dari suatu data. Pola tersebut dapat diketahui dari variabel-variabel yang ada pada data. Pola yang didapat bisa digunakan untuk memprediksi variabel lain yang belum diketahui nilai ataupun jenisnya. Karena itulah fungsi satu ini dikatakan sebagai fungsi prediksi. Nantinya bisa digunakan untuk memprediksi variabel tertentu yang tidak ada dalam suatu data. Hal ini tentunya memudahkan dan menguntungkan bagi mereka pemilik kepentingan yang memerlukan prediksi akurat untuk membuat hal penting tersebut menjadi lebih baik.

b. Description

Description atau fungsi deskripsi merupakan proses penting dalam analisis data yang bertujuan untuk mengidentifikasi ciri-ciri krusial yang terdapat dalam *database*. Melalui proses ini, peneliti dapat menemukan pola-pola yang relevan, yang dapat memberikan wawasan berharga tentang data yang dianalisis. Hasil dari fungsi deskripsi membantu dalam membuat prediksi yang lebih akurat dan mendukung pengambilan keputusan yang lebih baik.

c. Classification

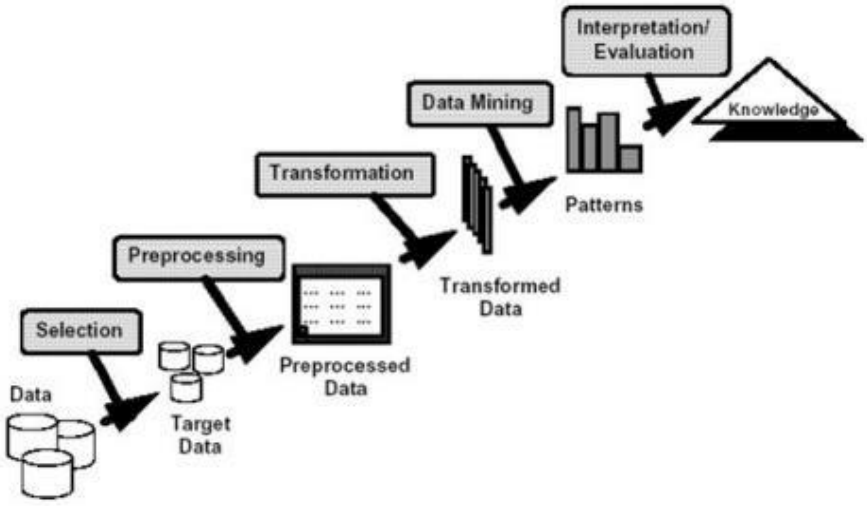
Classification atau klasifikasi adalah proses menemukan pola atau ciri untuk mendeskripsikan kelompok atau konsep dari data. Proses yang digunakan untuk mendeskripsikan data penting dan dapat memprediksi tren data di masa depan. Proses yang digunakan untuk menggambarkan data tersebut adalah hal yang terdapat pada masa mendatang. Contoh: Pelanggan suatu perusahaan telah berpindah ke pesaing perusahaan lainnya.

d. Assosiation

Association atau asosiasi adalah proses yang dipakai untuk menemukan suatu hubungan yang terdapat pada nilai atribut daripada sekumpulan data. Mengidentifikasikan hubungan antara kejadian-kejadian yang terjadi pada suatu waktu.

2.1.2 Tahapan dalam Data Mining

Tahapan yang dilakukan pada proses data mining diawali dari seleksi data dari data sumber ke data target, tahap *pre-processing* untuk memperbaiki kualitas data, transformasi, data mining serta tahap interpretasi dan evaluasi yang menghasilkan output berupa pengetahuan baru yang diharapkan memberikan kontribusi yang lebih baik. Secara detail ditunjukkan pada gambar berikut. Tahap-tahap data mining adalah sebagai berikut:



Gambar 2. 1 Tahapan Data Mining

a. Data Selection

Pemilihan (seleksi) data dari sekumpulan data operasional perlu dilakukan sebelum tahap penggalian informasi dalam KDD dimulai. Data hasil seleksi yang digunakan untuk proses data mining, disimpan dalam suatu berkas, terpisah dari basis data operasional. Data-data yang tidak relevan itu akan dibuang karena keberadaannya bisa mengurangi mutu atau akurasi dari hasil data mining nantinya.

b. Pre-Processing/Cleaning

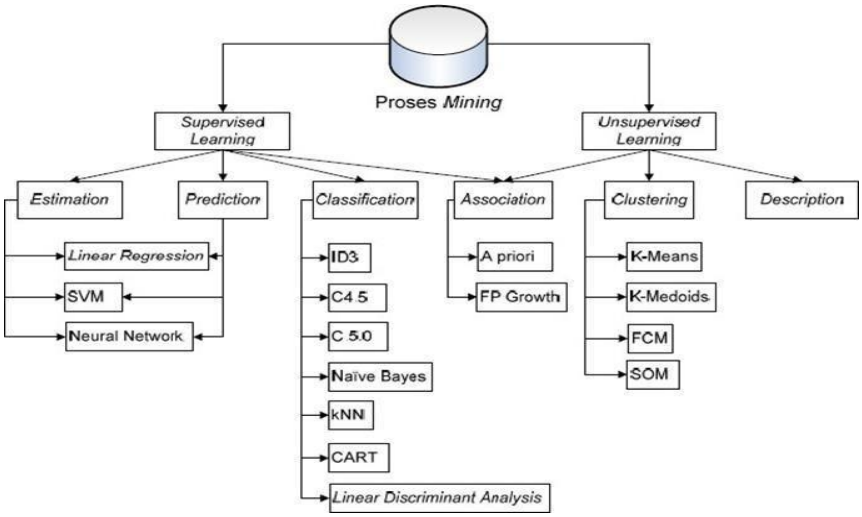
Sebelum melakukan proses data mining, dilakukan proses pembersihan pada data yang difokuskan pada KDD. Proses pembersihan tersebut antara lain menghapus data duplikat, memeriksa konsistensi data, dan memperbaiki kesalahan data. Dari data yang diambil, pembersihan data dilakukan ketika data hilang, data duplikat, atau dicetak berlebih.

c. Transformation

Transformasi adalah sekumpulan instruksi untuk mengubah suatu masukan menjadi keluaran yang diinginkan (input-pemrosesan-output). Proses yang dilakukan adalah mentransformasikan data ke dalam bentuk yang diinginkan sesuai dengan algoritma pengurutan yang digunakan. Untuk data numerik, normalisasi dilakukan untuk menyeimbangkan perbedaan skala yang mempengaruhi hasil. Proses transformasi didasarkan pada data yang dipilih, sehingga data tersebut sesuai untuk proses data mining. Proses pengkodean pada KDD bersifat kreatif dan sangat bergantung pada jenis atau pola informasi yang dicari dalam *database.*

d. Data Mining

Proses mencari pola atau informasi menarik dalam data terpilih dengan menggunakan teknik atau metode tertentu. Teknik metode atau algoritma dalam data mining sangat bervariasi. Pemilihan metode atau algoritma yang tepat sangat bergantung pada tujuan dan proses KDD secara keseluruhan. Teknik, metode algoritma dalam data mining sangat bervariasi. Pemilihan metode atau algoritma yang tepat sangat bergantung pada tujuan dan proses KDD secara keseluruhan.



Gambar 2. 2 Metode Data Mining

e. Interpretation/Evaluation

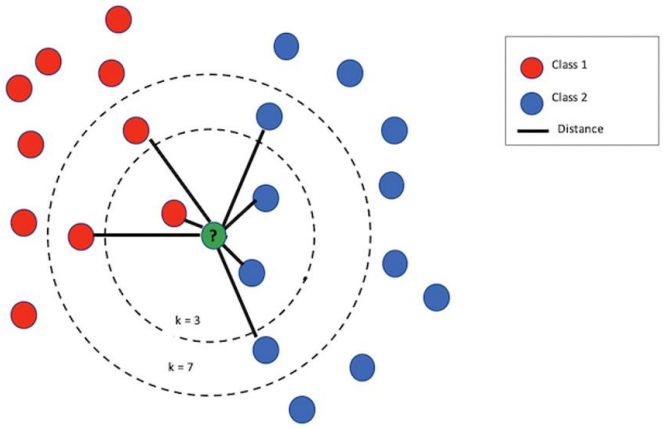
Pola informasi yang dihasilkan dari proses data mining perlu ditampilkan dalam bentuk yang mudah dimengerti oleh pihak yang berkepentingan. Tahap ini merupakan bagian dari proses KDD yang disebut *interpretation*. Tahap ini mencakup pemeriksaan apakah pola atau informasi yang ditemukan bertentangan dengan fakta atau hipotesis yang ada sebelumnya. Pola informasi oleh proses data mining wajib diperlihatkan ke bentuk yang mudah difahami pihak yang mempunyai kepentingan. Fase ini menjadi bagian KDD yang dinamakan interpretasi. Tahapan ini meliputi pemeriksaan apakah siklus ataupun informasi yang diputuskan berlawanan dengan fakta ataupun asumsi yang sudah ada**.**

2.2 KNN (K-Nearest Neighbor)

Machine learning dapat didefinisikan sebagai aplikasi komputer dan algoritma matematika yang diadopsi dengan cara pembelajaran yang berasal dari data dan menghasilkan prediksi di masa yang akan datang. [11] Istilah *machine learning* pertama kali dikemukakan oleh beberapa ilmuwan matematika seperti Adrien Marie Legendre, Thomas Bayes dan Andrey Markov pada tahun 1920-an dengan mengemukakan dasar- dasar machine learning dan konsepnya. [12] Salah satu contoh dari penerapan *machine learning* yang cukup terkenal adalah *Deep Blue* yang dibuat oleh IBM pada tahun 1996. *Deep Blue* merupakan *machine learning* yang dikembangkan agar bisa belajar dan bermain catur. Deep Blue juga telah diuji

coba dengan bermain catur melawan juara catur profesional dan *Deep Blue* berhasil memenangkan pertandingan tersebut. [13]

Algoritma Nearest Neighbor merupakan salah satu metode klasifikasi yang digunakan untuk pemecahan masalah pada bidang data mining. Algoritma KNN (K-Nearest Neighbor) merupakan teknik klasifikasi yang sering digunakan, yang dikenalkan oleh Fix dan Hodges pada tahun 1951, dan telah diakui sebagai algoritma sederhana yang terbaik. [14] Sama halnya dengan beberapa metode klasifikasi lainnya, algoritma ini memiliki ciri yaitu dengan pendekatan untuk mencari kasus dengan menghitung kedekatan kasus yang baru dengan kasus yang lama. Adapun teknik yang digunakan yaitu berdasarkan bobot dari sejumlah objek kasus yang ada. Metode KNN dikenal juga dengan *lazy learner* (pembelajar malas) karena tidak ada proses belajar (dari data) melainkan belajar dari data (tetangga) terdekat secara langsung pada saat klasifikasi.



Gambar 2. 3 Klasifikasi Berdasarkan Tetangga Terdekat

Terdapat beberapa cara untuk mengukur kedekatan jarak antara data baru (*testing* data) dan data lama (*training* data), diantaranya yaitu *euclidean distance* dan *manhattan distance (city block distance),* yang paling sering digunakan adalah *euclidean distance.*

KNN dapat digunakan untuk melakukan klasifikasi atau regresi. KNN bekerja dengan cara membandingkan titik data baru dengan sekumpulan data yang telah

dilatih dan dihafal. Algoritma ini berasumsi bahwa titik-titik yang serupa akan berada di dekat satu sama lain.

2.3 Depression

Depression atau depresi merupakan salah satu indikator penting atas rendahnya kesehatan mental seseorang dan telah menjadi masalah kesehatan utama di seluruh dunia karena jumlah penderitanya yang meningkat setiap tahun. [15] Salah satu gangguan kesehatan mental yang paling umum dialami oleh orang di seluruh dunia, depresi terjadi pada anak-anak, remaja, dewasa, bahkan orang tua. Menurut *World Health Organization* (WHO), depresi berada pada peringkat ke-4 sebagai penyakit yang paling umum di seluruh dunia. [16] Pikiran yang dipenuhi dengan pesimisme dapat menyebabkan seseorang mengalami depresi. Hal ini ditandai dengan kecenderungan untuk menyalahkan diri sendiri, orang lain, dan lingkungan sekitar atas setiap masalah yang timbul.

2.4 Kaggle

Kaggle adalah sebuah platform daring yang sangat dikenal di kalangan ilmuwan data dan mereka yang tertarik dengan *machine learning.* Platform ini menyediakan beragam fasilitas, seperti koleksi data yang melimpah, ajang kompetisi *machine learning*, serta forum komunitas yang ramai. Para pengguna dapat mengikuti kompetisi untuk menguji keahlian mereka, belajar dari sesama ilmuwan data, dan bahkan meraih hadiah. Tak hanya itu, Kaggle juga menyajikan aneka kursus dan tutorial tentang *machine learning*, menjadikannya tempat yang ideal bagi pemula maupun pakar.

2.5 Google Colab

Google Colab (Google Colaboratory) adalah platform berbasis cloud yang dikembangkan oleh Google, digunakan untuk menjalankan kode Python langsung melalui browser tanpa menginstal perangkat lunak tambahan. Berbasis pada Jupyter Notebook, layanan ini menyajikan platform online yang interaktif dan mudah digunakan, di mana dapat digunakan untuk menulis, mengeksekusi, serta mendokumentasikan kode. Dengan Google Colab, berbagai tugas pemrograman seperti *machine learning,* analisis data, pemrosesan gambar, pemodelan statistik,

dan pengujian algoritma dapat dilakukan dengan efisien tanpa konfigurasi perangkat keras atau perangkat lunak.

2.6 Penelitian Terdahulu

Penelitian yang dilakukan oleh Muhamad Septa Utama SP, Handoyo Widi Nugroho pada tahun 2023 dengan judul penelitian "Kajian Algoritma C4.5 dan K-NN Untuk Memprediksi Penduduk Miskin" dan menggunakan algoritma C4.5 dan KNN (K-Nearest Neighbor), didapatkan bahwa C4.5 memiliki keuntungan dalam interpretabilitas hasil, penghapusan fitur yang tidak relevan, dan efisiensi yang tinggi. Di sisi lain, metode KNN memiliki kelebihan dalam menangani data latih yang memiliki banyak noise dan efektivitas pada data latih yang besar masing-masing algoritma memiliki tingkat akurasi yang berbeda. [17]

Penelitian juga dilakukan dengan judul "Komparasi Algoritma Random Forest, Naïve Bayes dan K-Nearest Neighbor Dalam klasifikasi Data Penyakit Jantung" dan ditulis oleh Amril Samosir, Ms Hasibuan, Wahyu Eko Justino, Tri Hariyono pada tahun 2021 memberikan perbandingan uji coba, performance measure algoritma Naïve Bayes memiliki hasil yang lebih baik dibanding dengan algoritma, K-Nearest Neighbor dan Random Forest dengan metode K-Fold Cross Validation. Algoritma Naïve Bayes dapat memberikan rerata hasil akurasi sebesar 0,91 AUC, 0,84 CA, 0,84 F1, 0,839 Precision dan 0,84 Recall. [18]

Penelitian lain dengan judul "Analisis Perbandingan Algoritma Decision Tree, kNN, dan Naive Bayes untuk Prediksi Kesuksesan Start-up" yang diteliti oleh Adhitya Prayoga Permana, Kurniyatul Ainiyah, Khadijah Fahmi Hayati Holle pada tahun 2021 menunjukkan hasil perbandingan antara algoritma Decision Tree, kNN, dan Naive Bayes, untuk melakukan klasifikasi terhadap 923 data start-up, menunjukkan algoritma Decision Tree merupakan algoritma yang paling cocok untuk digunakan di antara algoritma KNN dan Naive Bayes. [19]

Penelitian yang dilakukan oleh Muhammad Fahrul Aditya, Andri Pramuntadi, Dhina Puspasari Wijaya, Yanuar Wicaksono pada tahun 2024 dengan judul "Penerapan Metode Clustering Untuk Prediksi Produksi Bawang Merah (Ensemble K-Nearest Neighbors)" menunjukkan bahwa prediksi harga bawang merah tidak berbeda jauh dengan harga bawang merah sebenarnya. [20]

Penelitian terkait algoritma prediksi yang dilakukan oleh Rahmadini, Enjel Erika Lorencis Lubis, Aji Priansyah, Yolanda R.W.N., Tuti Meutia pada tahun 2023 dengan judul "Penerapan Data Mining Untuk Memprediksi Harga Bahan Pangan Di Indonesia Menggunakan Algoritma K-Nearest Neighbor" menghasilkan bahwa algoritma KNN (K-Nearest Neighbor) diperoleh kesimpulan dari total percobaan dengan menggunakan range k=(2,10) maka hasil prediksi terbaik dengan menggunakan k = 2 dengan MAE dan RMSE untuk data training 52.77 dan 96.40 dan untuk data testing 55.55 dan 81.64 dan parameter K = 2 yang sudah dinormalisasi, dan pada visualisasi model, diperoleh kesimpulan grafik harga beras tidak terlalu signifikan mengalami penurunan serta kenaikan. [21]

Penelitian lain yang dilakukan oleh Dewi Cahyantia, Alifah Rahmayania, Syafira Ainy Husniar dengan judul "Analisis Performa metode KNN (K-Nearest Neighbor) Pada Dataset Pasien Pengidap Kanker Payudara" pada tahun 2020 mendapatkan hasil perhitungan 569 data yang di bagi menjadi 20% training dan 80% testing dengan K = 3,4 dan 5 mendapat nilai tertinggi untuk akurasi adalah 0,93 pada 20% keempat(K3), 20% Pertama(K4) dan 20% pertama(K5), untuk Presisi dengan nilai 0,97 pada 20% keempat(K3), untuk Recall dengan nilai 0,98 pada 20% ketiga(K3) dan F-measure dengan nilai 0,94 pada 20% keempat(K3) dan 20% ketiga(K5). [22]

Terakhir ada penelitian dengan judul "Penerapan Data Mining Untuk Menentukan Potensi Hujan Harian Dengan Menggunakan Algoritma K Nearest Neighbor (KNN)" yang dilakukan oleh Rofiq Harun, Kartika Chandra Pelangi, Yuliyanti Lasena pada tahun 2020 menunjukkan bahwa Penerapan Data mining untuk menentukan potensi hujan harian dengan menggunakan algoritma K-Nearest Neighbor(KNN) dapat di klasifikasikan, Sehingga membantu dan memudahkan masyarakat dalam mengetahui informasi potensi hujan yang tidak membingungkan. [23]

Tabel 2. 1 Penelitian Terdahulu

No.	Judul Penelitian	Penulis	Metode	Hasil
1.	Kajian Algoritma C4.5 dan K-NN Untuk Memprediksi Penduduk Miskin	Muhamad Septa Utama SP, Handoyo Widi Nugroho (2023)	C4.5 dan KNN (K-Nearest Neighbor*)*	Metode C4.5 akan menggunakan *decision tree* untuk mengklasifikasikan data, sehingga hasilnya dapat memberikan wawasan yang lebih jelas tentang faktor-faktor yang mempengaruhi kemiskinan. Sementara itu, KNN akan memanfaatkan jarak terdekat untuk mengelompokkan wilayah-wilayah dengan karakteristik serupa, yang dapat membantu dalam perencanaan program bantuan yang lebih tepat sasaran.
2.	Komparasi Algoritma Random Forest, Naïve Bayes dan K-Nearest Neighbor Dalam klasifikasi Data Penyakit Jantung	Amril Samosir, Ms Hasibuan, Wahyu Eko Justino, Tri Hariyono (2021)	Random Forest, Naïve Bayes dan KNN (K-Nearest Neighbor)	Hasil dari prediksi ketiga algoritma yang memiliki nilai paling benar yaitu pada Algoritma Naïve Bayes, dengan nilai 0,0,0,0,0,0,1,1,0,0, artinya sangat sesuai dengan nilai data testing yaitu : 0,0,0,0,0,0,1,1,0,0 utntuk Algoritma K-NN dengan nilai 1,0,0,1,0,0,1,1,0,0 memiliki 2 nilai kesalahan, yaitu nilai 1 pada baris pertama dan nilai 1 pada baris ke empat dan Random Forest dengan nilai 1,0,0,1,0,0,1,1,0,0 memiliki 2 nilai kesalahan, yaitu nilai 1 pada baris pertama dan nilai 1 pada baris ke empat.
3.	Analisis Perbandingan Algoritma Decision Tree, KNN, dan	Adhitya Prayoga Permana, Kurniyatul Ainiyah,	Decision Tree, KNN, dan Naive Bayes	Hasil akurasi Decision Tree adalah sebesar 79,29%, sedangkan algoritma kNN dengan

	Naive Bayes untuk Prediksi Kesuksesan Start-up	Khadijah Fahmi Hayati Holle (2021)		66,69%, dan Naive Bayes dengan 64,21%. Selanjutnya untuk nilai presisinya, Decision Tree masil lebih unggul dengan nilai 78,99%, diikuti algoritma kNN dengan 55,13%, dan Naive Bayes 51,32%. Dari hasil performa recall, ternyata algoritma Naive Bayes menunjukkan hasil paling baik dengan 79,16%, sedangkan Decision Tree 56,27% dan kNN dengan 40,14%.
4.	Penerapan Metode Clustering Untuk Prediksi Produksi Bawang Merah (Ensemble K-Nearest Neighbors)	Moh. Khoiru Alfin, Aang Alim Murtopo, Nurul Fadilah (2022)	Ensemble KNN (K-Nearest Neighbors)	Dapat dilihat bahwa prediksi harga bawang merah tidak berbeda jauh dengan harga bawang merah sebenarnya. Rentang prediksi harga Bawang Merah 2022 dari Rp 23.400 menjadi Rp 23.700 per kilogram. Kisaran ini mirip dengan harga Bawang Merah sebenarnya yaitu sekitar Rp 23.100 hingga Rp 23.800 per kilogram.
5.	Penerapan Data Mining Untuk Memprediksi Harga Bahan Pangan Di Indonesia Menggunakan Algoritma K-Nearest Neighbor	Rahmadini, Enjel Erika Lorencis Lubis, Aji Priansyah, Yolanda R.W.N., Tuti Meutia (2023)	KNN (K-Nearest Neighbor)	Dari total percobaan dengan menggunakan *range* k (2,10) maka hasil prediksi terbaik dengan menggunakan k = 2 dengan MAE dan RMSE untuk data training 52.77 dan 96.40 dan untuk data testing 55.55 dan 81.64 dan parameter K = 2 yang sudah dinormalisasi, dan pada visualisasi model, diperoleh kesimpulan grafik harga beras tidak terlalu signifikan mengalami penurunan serta kenaikan.

6.	Analisis Performa metode KNN (K-Nearest Neighbor) Pada Dataset Pasien Pengidap Kanker Payudara	Dewi Cahyantia, Alifah Rahmayania, Syafira Ainy Husniar (2020)	KNN (K-Nearest Neighbor)	Dari perhitungan 569 data yang di bagi menjadi 20% training dan 80% testing dengan K = 3,4 dan 5 mendapat nilai tertinggi untuk akurasi adalah 0,93 pada 20% keempat(K3), 20% Pertama(K4) dan 20% pertama(K5), untuk Presisi dengan nilai 0,97 pada 20% keempat(K3), untuk Recall dengan nilai 0,98 pada 20% ketiga(K3) dan F-measure dengan nilai 0,94 pada 20% keempat(K3) dan 20% ketiga(K5).
7.	Penerapan Data Mining Untuk Menentukan Potensi Hujan Harian Dengan Menggunakan Algoritma K Nearest Neighbor (KNN)	Rofiq Harun, Kartika Chandra Pelangi, Yuliyanti Lasena (2020)	KNN (K-Nearest Neighbor)	Hasil pengujian menunjukan bahwa prediksi penentuan cuaca harian dengan algoritma K-Nearest Neighbor mendapatkan Nilai RMSE 9.899 +/- 0.000.