BAB II
TINJAUAN PUSTAKA

2.1 Teori Dasar

2.1.1 Basis Data

"Basis data (*database*) adalah sistem terkomputerisasi yang tujuan utamanya adalah memelihara data yang sudah diolah atau informasi dan membuat informasi tersedia saat dibutuhkan. Pada intinya basis data adalah media untuk menyimpan data agar dapat diakses dengan mudah dan cepat"(Sukamto dan Shalahuddin, 2013).

Berdasarkan dari pengertian basis data di atas, dapat dikatakan pengertian dari basis data merupakan kumpulan dari data-data yang tersimpan di media yang berfungsi sebagai penyimpanan data, dan penyimpanan tersebut saling berhubungan.

2.1.2 *Unified Modeling Language* (UML)

UML adalah suatu metode dalam permodelan secara visual yang digunakan sebagai sarana perancangan sistem berorientasi objek. UML yang digunakan pada penelitian ini adalah sebagai berikut:

1. *Use Case* Diagram

   Use Case Diagram adalah salah satu jenis diagram dalam UML, diagram ini menggambarkan hubungan interaksi antara sistem dengan aktor. Adapun komponen di dalam *use case diagram* adalah sebagai berikut:

   a. Sistem

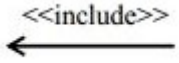
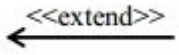
      Sebuah sistem digambarkan ke dalam bentuk persegi. Fungsinya untuk membatasi use case dengan interaksi dari luar sistem. Sistem pada umumnya diberikan label yang sesuai. Namun, umumnya sistem ini tidaklah diberi gambar karena tidak terlalu memberikan makna pada sebuah diagram.

   b. Aktor

      Aktor tidak memberikan kontrol terhadap sistem, namun hanya memberikan gambaran mengenai hubungannya dengan sistem.

Berikut adalah komponen-komponen pada use case Diagram:

Tabel 2. 1 Komponen-Komponen Use Case Diagram

Simbol	Keterangan
	Aktor: Mewakili peran orang, sistem yang lain, atau alat ketika berkomunikasi dengan *use case*.
	*Use Case:* Abstraksi dan interaksi antara sistem dan aktor.
	Association: Abstraksi dari penghubung antara aktor dengan *use case*.
	*Generalisasi*: Menunjukkan spesialisasi aktor untuk dapat berpartisipasi dengan *use case*.
	Menunjukan bahwa suatu *use case* seluruhnya merupakan fungsionalitas dari *use case* lainnya.
	Menunjukan bahwa suatu *use case* merupakan tambahan fungsional dari *use case* lainnya jika suatu kondisi terpenuhi.

2. *Class Diagram*

Class Diagram merupakan salah satu jenis struktur diagram pada UML yang berfungsi untuk menggambarkan struktur serta deskripsi *class*, atribut, metode, dan hubungan dari setiap objek. *Class diagram* memiliki 3 (tiga) komponen penyusun, berikut adalah

komponen- komponennya:

a. Komponen Atas

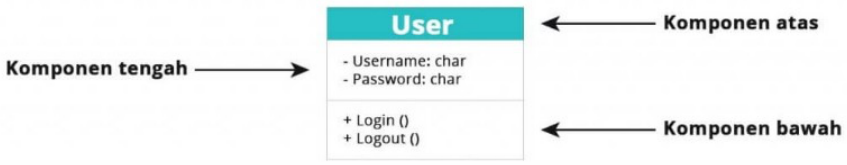
Komponen ini berisikan nama *class*. Setiap *class* pasti memiliki nama yang berbeda-beda, sebutan lain untuk nama ini adalah *simplename* (nama sederhana).

b. Komponen Tengah

Komponen ini berisikan atribut dari *class*, komponen digunakan untuk menjelaskan kulitas dari suatu kelas. Atribut ini dapat ditulis lebih detail, dengan cara memasukkan tipe nilai.

c. Komponen Bawah

Komponen ini menyertakan operasi yang ditampilkan dalam bentukdaftar. Operasi ini dapat menggambarkan bagaimana suatu *class* dapat berinteraksi dengan data.

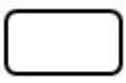


Gambar 2. 1 Komponen Penyusun Class Diagram

3. *Activity Diagram*

Activity Diagram adalah salah satu permodelan yang dapat menggambarkan proses-proses yang terjadi pada sebuah sistem, selain itu permodelan ini menggambarkan proses-proses tersebut secara vertical. Berikut adalah komponen-komponen pada *activity diagram*:

Tabel 2. 2 Komponen-Komponen Activity Diagram

Simbol	Nama	Keterangan
	Status Awal	Sebuah diagram aktivitas memiliki sebuah status awal.
	Aktivitas	Aktivitas yang dilakukan sistem, aktivitas biasanya diawali dengan

		kata kerja.
	Percabangan/*Decision*	Percabangan dimana ada pilihan aktivitas yang lebih dari satu.
	Penggabungan/*Join*	Penggabungan dimana yang mana lebih dari satu aktivitas lalu digabungkan jadi satu.
	Status Akhir	Status akhir yang dilakukan sistem, sebuah diagram aktivitas memiliki sebuah status akhir.
	*Swimlane*	*Swimlane* memisahkan organisasi bisnis yang bertanggung jawab terhadap aktivitas yang terjadi.

4. *Sequence* Diagram

Sequence Diagram adalah sebuah diagram yang digunakan untuk menjelaskan dan menampilkan interaksi antar objek-objek dalam sebuah sistem secara terperinci. Selain itu *sequence diagram* juga akan menampilkan pesan atau perintah yang dikirim, beserta waktu pelaksanaannya. Objek-objek yang berhubungan dengan berjalannya proses operasi biasanya diurutkan dari kiri ke kanan. Berikut adalah beberapa komponen-komponen dalam *sequence diagram*:

a. Aktor

Komponen yang pertama adalah aktor. Komponen ini menggambarkan seorang pengguna (*user*) yang berada diluar sistem dan sedang berinteraksi dengan sistem. Dalam *sequence diagram*, aktor biasanya digambarkan dengan simbol *stick figure*.

b. Activation Box

Selanjutnya ada *activation box*. Komponen *activation box* ini merepresentasikan waktu yang dibutuhkan suatu objek untuk menyelesaikan tugasnya. Semakin lama waktu yang diperlukan, maka secara otomatis *activation box*-nya juga akan menjadi lebih panjang. Komponen ini digambarkan dengan bentuk persegipanjang.

c. Lifeline

Berikutnya adalah *lifeline*. Komponen ini digambarkan dengan bentuk garis putus-putus. *Lifeline* ini biasanya memiliki kotak yang berisi objek yang memiliki fungsi untuk menggambarkan aktifitas dari objek.

d. Objek

Komponen berikutnya adalah objek. Komponen objek ini digambarkan memiliki bentuk kotak yang berisikan nama dari objek dengan garis bawah. Biasanya objek berfungsi untuk mendokumentasikan perilaku sebuah objek pada sebuah sistem.

e. Messages

Terakhir ada messages atau pesan. Komponen ini untuk menggambarkan komunikasi antar objek. *Messages* biasanya muncul secara berurutan pada *lifeline*. Komponen messages ini direpresentasikan dengan anak panah. Inti dari sebuah diagram urutan terdapat pada komponen *lifeline* dan *messages* ini.

2.1.3 Data Warehouse

"Data *Warehouse* adalah suatu konsep dan kombinasi teknologi yang memfasilitasi organisasi untuk mengelola dan memelihara data historis yang diperoleh dari system atau aplikasi operasional" (Purba, 2020).

Dari definisi tersebut dapat disimpulkan data warehouse merupakan tempat penyimpanan data tunggal yang lengkap dan konsisten dengan karakterisitik berorientasi subjek, terintegrasi, tidak volatile, dan bervariasi waktu yang dapat digunakan untuk mendukung keputusan.

2.1.4 Data Mart

"Data mart merupakan bagian dari data warehouse yang berada pada tingkatan yang lebih kecil seperti level departemen pada suatu organisasi atau perusahaan" (Setiawan, 2017).

Dari penjelasan diatas data mart dapat digunakan untuk mengoptimalkan pemanfaatan data histori transaksi.

2.1.5 Data Mining

"Data Mining adalah proses yang memanfaatkan teknik-teknik statistik, matematika, dan kecerdasan buatan untuk mengekstrak dan mengidentifikasi informasi dan knowledge selanjutnya atau pola-pola yangberasal dari sekumpulan data yang sangat besar" (Sano, 2019).

Berdasarkan definisi data maing diatas, maka dapat dijelaskan data mining adalah suatu metode yang mampu melakukan pengolahan data berskala besar, oleh karena itu data mining ini memiliki peranan penting dalam bidang industri, keuangan, cuaca, ilmu dan teknologi.

2.1.6 Extraction, Transformation, dan Load (ETL)

Dharayani, Laksitowening dan Yanuarfiani (2015) menyatakan bahwa ETL adalah salah satu proses pada data warehouse. Proses dari ETLadalah mengumpulkan data dari berbagai macam sumber. ETL juga berfungsi untuk mengolah data menjadi data yang bersih sesuai ketentuan data warehouse. Proses ETL pada umumnya terdiri dari berbagai macam aktvitas dan membutuhkan waktu serta memori yang cukup besar.

Terdapat 3 (tiga) langkah yang dapat dilakukan dalam Menyusun proses ETL dan membuat data terintegrasi dari sumber ke tujuan, yaitu sebagai berikut:

1. Ektraksi Data

   Pada langkah pertama proses ETL ini, data terstruktur dan tidak terstruktur diimpor dan dikonsolidasikan ke dalam suatu wadah penyimpanan. Data mentah dapat diekstraksi dari berbagai sumber berikut ini:

   a. Database yang ada dan legacy system.
   b. Cloud, Hybrid, dan on-premises environments.
   c. Aplikasi penjualan dan pemasaran.
   d. Mobile devices dan apps.
   e. CRM systems.
   f. Data storage platforms.
   g. Data Warehouse.
   h. Analytics Tools.

2. Transformasi

   Transformasi ETL merupakan pembersihan dan mempersiapkan agregasi untuk analisis. Langkah ini sangat penting dalam proses ETL karena membantu memastikan data yang akan diolah sepenuhnya siap dan kompatibel. Proses transformasi ETL terbagi menjadi beberapa proses yaitu sebagai berikut:

   a. Pembersihan, data yang tidak konsisten dihilangkan.
   b. Standardisasi, memasang aturan pemformatan ke kumpulan data.
   c. Deduplikasi, data yang sama dibuang atau dikecualikan.
   d. Verifikasi, data yang tidak dapat digunakan dihapus dan anomaly ditandai.
   e. Pengurutan, data diatur menurut jenisnya.
   f. Tugas lainnya, aturan tambahan yang dapat meningkatkan kualitas data.

3. *Loading*/Memuat Data

   Loading adalah proses terakhir dari ETL, yaitu memuat data yang sudah diubah ke tujuan baru. Data tersebut dapat dimuat sekaligus (*full load*) atau interval terjadwal (*incremental load*).

2.1.7 Operasi Data Mining

Menurut sifatnya operasi data mining dibedakan menjadi 2 bagian, yaitu sebagai berikut:

1. Prediksi

   Untuk menjawab pertanyaan apa dan sesuatu yang bersifat abstrak atautransparan. Operasi prediksi digunakan untuk validasi *hipotesis*, *querying*, dan pelaporan.

2. Penemuan

   "Bersifat transparan dan untuk menjawab pertanyaan "mengapa?". Operasi penemuan digunakan untuk analisis data eksplorasi, pemodelan prediktif, segmentasi database, analisis keterkaitan (*link analisys*) dan deteksi deviasi" (Hermawati, 2013).

2.1.8 Teknik Data Mining

Beberapa teknik dan sifat data mining adalah sebagai berikut:

1. Klasterisasi

   Klasterisasi adalah mempartisi data-set menjadi beberapa sub-net ataukelompok sedemikian rupa sehingga elemen-elemen dari suatu kelompok tertentu memiliki set property yang di share bersama, dengan tingkat similaritas yang tinggi dalam suatu kelompok yang rendah. Disebut juga dengan "*unsupervised learnig*".

2. Regresi

   Regresi adalah mempredikisi nilai dari suatu variabel kontinyu yang diberikan berdasarkan nilai dari variabel yang lain, dengan mengasumsikan sebuah model ketergantungan linier atau non linier.

3. Klasifikasi

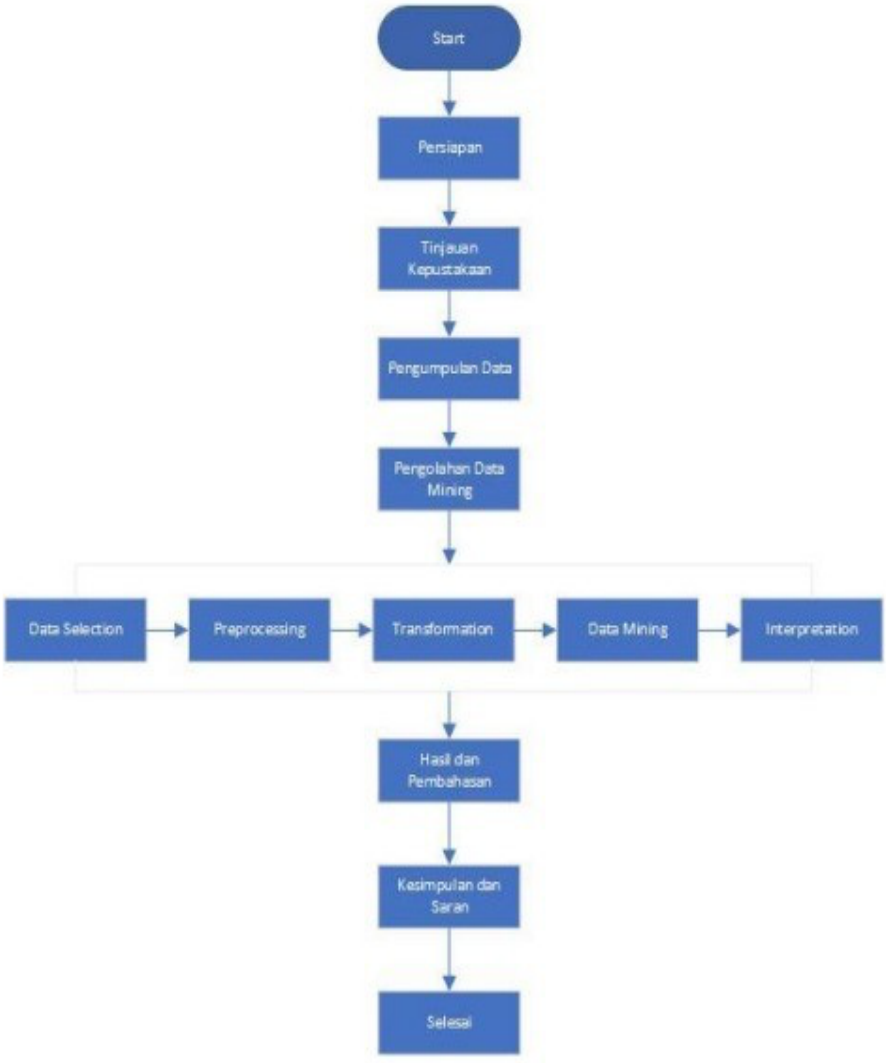
   Klasifikasi adalah menentukan sebuah record data baru ke salah satu dari beberapa kategori (kelas) yang telah didefenisikan sebelumnya dan disebut juga dengan "*supervised learning*".

4. Kaidah Asosiasi (*Association Rule)*

   "Kaidah Asosiasi adalah mendeteksi kumpulan atribut-atribut yang muncul bersamaan (*co-occur*) dalam frekuensi yang sering dan membentuk sejumlah kaidah dari kumpulan-kumpulan tersebut" (Hermawati, 2013).

2.1.9 Knowledge Discovery in Database (KDD)

Proses dan teknik penyaringan data menentuka mutu pengetahuan dan informasi yang akan diperoleh. Istilah lain untuk data mining adalah Knowledge Discovery in Databases (KDD). KDD merupakan sebuah proses yang terdiri dari serangkaian proses interasi yang terurut, dan data mining merupakan salah satu langkah dalam proses KDD. Urutan langkahdalam KDD adalah sebagai berikut:



Gambar 2. 2 Langkah-langkah dalam Knowledge Discovery in Databases (KDD)

1. Data *Selection*

   Pemilihan (seleksi) data dari sekumpulan data operasi perlu dilakukan sebelum tahap penggalian informasi dalam KDD dimulai.

2. *Pre-processing/Cleaning*

   Sebelum proses data mining dapat dilaksanakan, perlu dilakukan proses cleaning pada data yang menjadi fokus KDD. Proses cleaning mencakup antara lain membangun duplikasi data, memeriksa data yang inkonsisten dan memperbaiki kesalahan pada data, seperti kesalahan cetak (tipografi).

3. *Transformation*

   Coding adalah proses transformasi pada data yang telah dipilih, sehingga data tersebut sesuai untuk proses data mining. Proses coding dalam 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, metode atau algoritma dalam data mining sangat bervariasi.

5. *Interpretation/Evaluation*

   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.

2.1.10 *K-Nearest Neighbor* (KNN)

Gorunescu (2013) menyatakan algoritma KNN adalah merupakan sebuah metode untuk melakukan klasifikasi terhadap obyek baru berdasarkan (K) tetangga terdekatnya KNN termasuk algoritma *supervised learning*, yang mana hasil dari query instance baru, diklasifikasikan berdasarkan mayoritas dari kategori pada KNN. Kelas yang paling banyak muncul, yang akan menjadi kelas hasil.

Perbedaan antara *supervised learning* dengan *unsupervised learning* adalah pada *supervised learning* bertujuan untuk menemukan polabaru dalam data dengan menghubungkan pola data yang sudah ada dengan data yang baru. Sedangkan pada *unsupervised learning*, data belum memiliki pola apapun, dan tujuan *unsupervised learning* untuk menemukanpola dalam sebuah data. Tujuan dari algoritma KNN adalah untuk mengklasifikasi objek baru berdasarkan atribut dan *training samples*.

Menurut Prasetyo (2014) Pada algoritma ini terdapat beberapa cara, yang berguna mencari nilai tetangga terdekat yaitu sebagai berikut:

1. Manhattan
2. Cosine
3. Correlation
4. Hamming
5. Euclidean.

Menurut (Sayad, 2014) rumus perhitungan jarak dengan Euclidean dapat dilihat pada gambar 2.3 di bawah ini:

$$\sqrt{\sum_{i=1}^{K}(Xi - Yi)^2}$$

Gambar 2. 3 Rumus Perhitungan Jarak Euclidean

Sedangkan untuk menghitung akurasi data menggunakan rumusseperti di bawah ini:

$$\text{Akurasi} = \frac{\text{Jumlah pengujian yang diperiksa benar}}{\text{Jumlah data yang diuji}} \times 100$$

2.1.11 Bahasa Pemrograman Python

Edwardo (2018) menyatakan *python* adalah *scripting language* yang berorientasi objek. Bahasa pemrograman ini dapat digunakan untuk pengembangan perangkat lunak dan bisa dijalankan melalui berbagai sistem operasi. Saat ini, python juga merupakan bahasa yang populer bagi bidang data *science* dan analisis. Hal ini dikarenakan oleh dukungan bahasa *python* terhadap *library-library* yang didalamnya

menyediakan fungsi analisis data dan fungsi *machine learning*, data *preprocessing tools*, serta visualisasi data.

Dalam membuat sebuah proyek data mining dengan *python*, dapat menggunakan anaconda, dimana anaconda telah menyediakan berbagai kelengkapan python yang lebih dikhususkan untuk kebutuhan analisis data. IDE (*Integrated Development Enviroment*) yang dapat digunakan antara lain *Jupyter Notebook*, *Google Collabs* dengan *extension* .ipynb yang sudah merupakan bawaan dari anaconda Navigator.

Berikut ini adalah beberapa alasan Python menjadi bahasa yang populer, khususnya dalam ranah analisis data dan data science:

1. Ketersediaan akan *open-source library*, *frameworks*, *tools* untuk data mining, contohnya adalah *SciKit Learn*, *TensorFlow*.
2. Relatif lebih mudah dipahami. penulisan code di *python* relatif lebih singkat dibandingkan bahasa pemrograman yang lain.
3. Multifungsi, tidak hanya untuk data *processing*, namun juga bisa untuk tugas lain seperti membuat website dan tampilan GUI (*Graphical UserInterface*).

2.2 Tinjauan Studi

Berikut beberapa referensi yang mendukung penelitian ini:

1. Jurnal penelitian yang dilakukan oleh Ike Yolanda dan Hasanul Fahmi dengan judul "Penerapan Data Mining Untuk Prediksi Penjualan Produk Roti Terlaris Pada PT.Nippon Indosari Corpindo Tbk Menggunakan Metode *K-Nearest Neighbor*". Penelitian ini membahas mengenai penerapan data mining pada perusahaan PT. Nippon Indosari Corpindo yang membutuhkan sebuah prediksi penjualan produk roti terlaris, agar mempermudah pihak perusahaan dalam memproduksi roti mana yang paling banyak diproduksi. Maka untuk mengetahui penjualan produk roti terlaris dibutuhkan sebuah data mining untuk memecahkan sebuah masalah dengan mengunakan metode k- nearest neighbor.
2. Jurnal penelitian yang dilakukan oleh Yulia Rizki Amalia dengan judul "Penerapan Data Mining Uuntuk Prediksi Penjualan Produk Elektronik Terlaris Menggunakan Metode K- Nearest Neighbor (Studi Kasus : PT.

Bintang Multi Sarana Palembang)". Penelitian ini membahas mengenai pemanfaatan data mining untuk memprediksi penjualan produk elektronik dan furniture agar pihak perusahaan bisa mengetahui penjualan produk elektronik yang paling banyak diminati oleh konsumen. Selain itu prediksi ini bertujuan untuk mempermudah bagian penyedia stok barang pada perusahaan dalam melakukan perencanaan dan penyedian stok barang.

3. Jurnal penelitian yang dilakukan oleh Dedi Handoko, Heru Satria Tambunan, Jaya Tata Hardinata dengan judul "Analisis Penjualan Produk Paket Kuota Internet Dengan Metode K- Narest Neighbor". Penelitian ini membahas tentang penjualan produk paket kuota internet, dimana perusahaan tersebut belum mampu memprediksi penjualan paket kuota internet sehingga produk hanya terjual beberapa saja. Dari permasalahan tersebut maka dibutuhkan prediksi untuk penjualan kedepannya guna untuk mempermudah pihak perusahaan dalam perencanaan penyediaan stok dan mengetahui penjualan produk paket kuota internet yang terlaris. Metode yang digunakan dalam penelitian tersebut adalah Metode *K-Nearest Neighbor*.
4. Jurnal penelitian yang dilakukan oleh Ayu Azlina Putri dengan judul "Penerapan Data Mining Untuk Memprediksi Penjualan Buah Dan Sayur Menggunakan Metode *K-Nearest Neighbor* (Studi Kasus : PT. Central Brastagi Utama)". Penelitian ini membahas mengenai penjualan produk buah dan sayur pada PT. Central Brastagi Utama yang terdapat permasalahan belum adanya sistem yang mengatur prediksi atau peramalan untuk penjualan buah dan sayur, sehingga terjadi penumpukan produk dan selain itu juga banyak produk yang rusak dan busuk. Dari permasalahan tersebut, maka dilakukan penerapan data mining untuk memprediksi penjualan buah dan sayur menggunakan metode *K-Nearest Neighbor*. Tujuannya agar dapat meminimalisir penumpukan atau kerusakan pada produk dan juga dapat meminimalisir kerugian pada perusahaan.
5. Jurnal penelitian yang dilakukan oleh Katon Wijana dengan judul "Program Bantu Prediksi Penjualan Barang Menggunakan Metode KNN (Studi Kasus: U.D. Anang)". Penelitian ini membahas tentang toko U.D. Anang yang bergerak dibidang penjualan bahan bangunan. Permasalahan yang terdapat

pada toko ini adalah proses transaksi ditoko ini masih dilakukan secara konvensional. Hal ini menjadi kurang maksimal oleh karena sistem pengolahan data hanya dilakukan oleh karyawan dan tanpa sistem komputerisasi sehingga pengelola toko kesulitan dalam memikirkan peluang usaha untuk meningkatkan transaksi penjualan berdasarkan riwayattransaksi. Dari permasalahan tersebut, maka diterapkan metode KNN yangdiharapkan dapat membantu perencanaan dalam setiap transaksi penjualan.

6. Jurnal penelitian yang dilakukan oleh Aisha Alfani W.P.R, Fahrur Rozi, Farid Sukmana dengan judul "Prediksi Penjualan Produk Unilever Menggunakan Metode *K-Nearest Neighbor*". Penelitian ini membahas tentang sebuah toko retail semi grosir, dimana terdapat permasalahan yaitu karena banyaknya permintaan konsumen akan produk Unilever berdasarkan data 3 tahun terakhir, maka dibutuhkan prediksi untuk penjualan produk Unilever terlaris. Hal ini berguna untuk mempermudah pihak pemilik toko dalam perencanaan penyediaan stok. Untuk mengetahui penjualan produk Unilever terlaris digunakan teknik klasifikasi *data mining* dan algoritma *K-Nearest Neighbor*.
7. Jurnal penelitian yang dilakukan oleh Bagus Hardiyanto, Fahrur Rozi berjudul "Prediksi Penjualan Sepatu Menggunakan Metode *K-Nearest Neighbor*". Penelitian ini membahas tentang prediksi penjualan sepatuuntuk klasifikasi potensi pelanggan baru di toko Obral Murah dengan menggunakan metode K-*Nearest Neighbor*. Agar toko Obral Murah tetap menjadi toko favorit dan tidak kalah dengan pesaing pesaing baru, untuk menghindari hal tersebut maka perlu adanya prediksi penjualan untuk melihat potensi dari para pelanggan dan barang yang disukai pelanggan.