

BAB II

TINJAUAN PUSTAKA

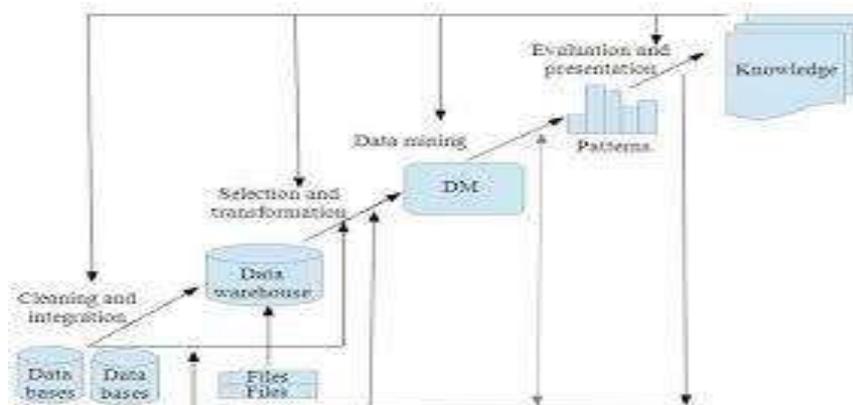
2.1 *Knowledge Discovery in Database (KDD)*

Knowledge Discovery in Database (KDD) merupakan kegiatan yang meliputi pengumpulan data, pemakaian data, historis untuk menemukan keteraturan, pola atau hubungan di dalam *set* data yang berukuran besar. Proses *Knowledge Discovery in Database* ini melibatkan hasil proses pada *Data mining* (proses pengekstrak kecenderungan suatu pola data), yang kemudian mengubah hasilnya secara akurat menjadi informasi yang mudah dipahami. *Knowledge Discovery in Databases (KDD)* untuk menunjukkan pada keseluruhan proses pencarian pengetahuan didalam kumpulan data dalam jumlah yang besar dan menggolongkan *Data mining* ini sebagai salah satu langkah didalam proses KDD dikarenakan dengan penerapannya terhadap satu algoritma spesifik didalam mencari pola-pola (model) dalam data (Jaya et al., 2019).

Pengumpulan *Data mining* bukan hanya sekedar bahwa data hanya terkumpul melainkan mencakup analisis dan prediksi dari informasi yang akan di tampilkan. Data yang telah terkumpul akan di simpan pada database dan kemudian di proses sehingga data tersebut dapat dijadikan pengambilan keputusan dalam melihat informasi yang akan digunakan (Saputri et al., 2021)

2.1.1 Tahapan proses KDD

Pada tahapan proses KDD berdasarkan jurnal penelitian (Syahputra et al., 2020). dari sebagai berikut:



Gambar 2. 1 Tahapan proses KDD
Sumber : (Syahputra et al., 2020).

1. *Data Selection*, pada proses ini dilakukan dengan pemilihan himpunan data, menciptakan himpunan target, atau fokus terhadap subset variabel (sampel data) dimana penemuan (*discovery*) akan dilakukan dan hasil seleksi disimpan pada suatu berkas yang terpisah dari basis data operasional.
2. *Pre-processing*, pada proses ini dilakukan dengan membuang data yang tidak konsisten atau *noise*, duplikasi data dan memperbaiki kesalahan data.
3. *Transformation*, pada proses ini merupakan *mentransformasikan* atau menggabungkan data yang telah ada ke tempat melakukan proses mining dengan cara melakukan peringkasan (*agregasi*)
4. *Data mining*, pada proses ini ialah mencari pola tau informasi menarik dalam data terpilih dengan menggunakan teknik, metode atau algoritma tertentu sesuai dengan proses KDD secara keseluruhan.
5. *Interpretation*, pada proses ini ialah menerjemahkan pola-pola yang telah dihasilkan dari data mining, mengevaluasi (menguji) apakah pola atau informasi yang di temukan sesuai atau bertentangan dengan fakta atau hipotesa sebelumnya.

2.2 Data Mining

2.2.1 Pengertian Data Mining

Menurut (Nainel et al., 2020) *Data mining* adalah suatu langkah analisis terhadap proses penemuan di dalam basis data atau knowledge discovery in database atau KDD. Pengetahuan tersebut bisa dapat berupa pola data atau relasi antar data yang valid(yang belum dei ketahui sebelumnya). *Data mining* di tujukan untuk mengestrak pengetahuan dari sekumpulan data sehingga diperoleh strktu yang dapat dimengerti manusia serta meliputi basis data dan managemen data.

Menuut (Singingi et al., 2020) *Data mining* merupakan proses menemukan sebuah hubungan dengan data yang tidak diketahui oleh pengguna dan menyajikan data dengan cara yang dapat dipahami sehingga hubungan tersebut dapat menjadi dasar pengambilan keputusan. Dalam *Data mining* terdapat beberapa cara untuk mendapatkan pola informasi yang tersembunyi yaitu: Klasterisasi (*clustering*), Regresi (*Regresesion*), Asosiasi (*asosiation*), dan klasifikasi (*clasification*).

2.2.2 Teknik Data Mining

Beberapa teknik *Data mining* dan sifat *Data mining* sebagai berikut berdasarkan jurnal penelitian(Yolanda & Fahmi, 2021):

1. Klasterisasi merupakan pempratisi *data set* menjadi beberapa *sub net* atau kelompok dengan sedemikian rupa sehingga elemen-elemen dari suatu kelompok tertentu menjadi *set property* yang di bagikan secara bersamaan.

2. Regresi merupakan suatu prediksi nilai dari suatu variabel *continiue* yang diberikan berdasarkan nilai dari variabel yang lain, dengan memberikan sebuah asumsi dengan sebuah model ketergantungan linear dan non linear
3. Klasifikasi merupakan penentu sebuah *record* data baru ke salah satu dari beberapa kategori (kelas) yang telah didefenisikan sebelumnya dan disebut juga dengan “*supervised learning*”
4. Kaidah asosiasi (*assosiation rule*) adalah mendeteksi kumpulan atribut-atribut yang muncul secara bersamaan (*co occur*) dalam frekuensi yang sering dan membentuk sejumlah kaidah dari kumpulan-kumpulan tersebut

2.2.3 Keuntungan dan Kelemahan Data mining

2.2.3.1 Keuntungan data mining

Analisis data menggunakan *Data mining* dapat memberikan keuntungan bagi seorang pengusaha untuk mengoptimalkan manajemen dan waktu. Berikut beberapa keuntungan dari data mining:

1. Managemen hubungan pelanggan yang lebih baik
2. Perkiraan tren pasar
3. Membantu dalam persaingan
4. Menarik dan mempertahankan pelanggan
5. Analisis data yang lebih akurat.

2.2.3.2 Kelemahan data mining

Selain keuntungan yang dapat di peroleh *Data mining* juga memiliki kelemahan yang perlu diketahui sebagai berikut:

1. Biaya terbilang cukup mahal

2. Keamanan data
3. *Data mining* melanggar privasi pengguna

2.2.4 Klasifikasi Data Mining

Klasifikasi adalah suatu cara untuk menganalisa data histori yang telah di simpan pada database dan secara otomatis menghasilkan suatu model yang dapat memprediksi suatu perilaku dimasa akan datang. Model intruksi ini terdiri dari generalisasi pada baris-baris data yang digunakan untuk pelatihan yang akan membantu membedakan kelas-kelas standar dengan harapan dapat memprediksi kelas tersebut secara akurat peristiwa-pristiwa yang akan mendatang (Diki & Nurina, 2022).

Data mining klasifikasi juga mempunyai beberapa model algoritma yang dapat digunakan untuk memprediksi yaitu metode naive bayes, yang digunakan sebagai membagikan masalah ke dalam bentuk kelas-kelas dan ada juga metode KNN (*K-Nearest Neighbor*). KNN digunakan sebagai metode yang melakukan klasifikasi terhadap objek yang akan diteliti berdasarkan data pembelajaran dengan data yang paling dekat terhadap objek tersebut. Pada KNN memiliki keunggulan yaitu dapat memprediksi secara tegas terhadap data training dan efektif terhadap data yang besar (Neighbor, 2018).

2.2.5 Pengelompokan Data Mining

Data mining terbagi atas beberapa pengelompokan berdasarkan tugas dan peran yang dapat dilakukan, yaitu:(Saputri et al., 2021)

1. Deskripsi

Deskripsi merupakan gambaran dari pola yang cenderung berada di dalam data dan memungkinkan memberikan penjelasan dari suatu pola tersebut.

2. Estimasi

Merupakan suatu gambaran yang hampir serupa dengan klasifikasi namun pada estimasi dapat menentukan model dan fungsi dengan target estimasi lebih kearah numerik dari pada kearah kategori. Model di bangun menggunakan *record* lengkap dengan menyediakan nilai variabel target sebagai nilai prediksi.

3. Prediksi

Prediksi umumnya hampir sama dengan klasifikasi dan estimasi, namun pada prediksi nilai dari periode sebelumnya dijadikan bahan prediksi sebuah hasil yang akan ada pada masa mendatang.

4. Klasifikasi

Klasifikasi merupakan suatu proses dalam menemukan model atau fungsi dengan gambaran yang membedakan kelas data atau konsep dengan tujuan agar dapat memprediksikan kelas untuk data yang tidak diketahui kelasnya.

5. Pengklusteran

Merupakan pengelompokan *record*, pengamatan yang membentuk kelas objek yang memiliki kemiripan.

6. Asosiasi

Merupakan suatu metode untuk menemukan produk yang sering dibeli bersamaan dengan pelanggan atau bisa juga disebut atribut yang sering muncul dalam satu waktu.

2.2.6 Operasi Data Mining

Menurut sifatnya operasi *Data mining* dibedakan menjadi dua bagian yaitu 1 prediksi (*prediction driven*) digunakan sebagai penjawab pertanyaan tentang sesuatu yang bersifat abstrak. Operasi prediksi digunakan untuk validasi hipotesis, querying dan laporan. Sifat ke 2 ialah penemuan (*discovery driven*) bersifat transparan untuk menjawab pertanyaan “mengapa”. Operasi penemuan digunakan untuk analisis data eksplorasi, pemodelan prediktif, segmentasi *database* (Syahputra, 2020).

2.3 Produksi

Produksi merupakan suatu proses atau kegiatan untuk menciptakan dan menambah kegunaan suatu barang atau jasa untuk mengubah sumber-sumber tenaga kerja, mesin, bahan dan dana menjadi suatu hasil keuntungan. Sifat produksi yaitu mengolah bahan baku dan bahan pembantu secara manual dengan menggunakan peralatan sehingga menghasilkan suatu produk yang memiliki nilai jual (Singingi et al., 2020).

2.3.1 Faktor-faktor yang mempengaruhi Produksi

Faktor produksi merupakan hal yang paling penting pada setiap proses produksi karena jika tidak tersedia salah satunya maka bisa dipastikan produksi tidak akan berjalan dengan lancar dengan resiko tidak ada hasil produksi yang akan dihasilkan, proses produksi macet sehingga usaha akan mengalami kerugian. Faktor-faktor yang mempengaruhi produksi sebagai berikut(Singingi et al., 2020) :

1. Tenaga Kerja merupakan faktor produksi asli karena memiliki unsur fisik, pikiran serta kemampuan yang dimiliki tenaga kerja.

2. Modal merupakan suatu yang bisa dimanfaatkan menunjang proses produksi lain dalam bentuk uang, peralatan dan sebagainya.
3. Keahlian merupakan keterampilan seseorang dalam memanfaatkan faktor produksi untuk menghasilkan barang
4. Jumlah yang dihasilkan merupakan banyaknya barang yang di produksi dalam menentukan keberhasilan produksi.
5. Ketepatan waktu adalah yang penting dalam produksi semakin tepat waktu maka proses produksi dapat dikatakan berhasil efektif dan efisien.

2.3.2 Manfaat Produksi

Manfaat-manfaat produksi yang mendukung adalah sebagai berikut:

1. Melancarkan produksi produk
2. Memberi keuntungan pada pengusaha sehingga bisa dijadikan sebagai modal untuk dapat meningkatkan bisnis termasuk mengembangkan faktor produksi selanjutnya
3. Produk sesuai harapan konsumen baik dari segi kualitas dan kuantitasnya

2.3.3 Penjualan

Penjualan merupakan faktor utama bagi setiap perusahaan karena dengan adanya penjualan maka suatu perusahaan akan memperoleh keuntungan yang lebih supaya usaha yang di lakukan dapat berjalan dan dapat berkembang. Namun bisa dilihat bahwa persaingan bisnis di era jaman modren sekarang begitu ketat, setiap perusahaan diwajibkan untuk membuat suatu persiapan diri secara profesional dan fleksibel sehingga perusahaan mampu bertahan dan mampu bertumbuh. Perusahaan yang baik ialah perusahaan yang memiliki strategi agar dapat memenuhi kebutuhan

pemasaran. Maka dari itu persaingan terletak pada bagaimana sebuah perusahaan dapat menghadirkan dan menyediakan hasil produksi yang berkualitas dan dengan stok barang yang selalu terpenuhi (Putri, 2021).

2.4 Prediksi / *Forecasting*

Prediksi (*forecasting*) merupakan suatu bentuk perhitungan dengan sistem peramalan keadaan di masa akan datang terhadap pengujian keadaan masa lalu. Peramalan produksi di masa akan datang merupakan suatu cara dalam menentukan besarnya perkiraan jumlah produksi, dengan ketentuan potensi yang lebih jauh di masa akan datang. Prediksi berfungsi sebagai suatu sistem yang dapat membantu pengusaha atau pemilik usaha dalam mengambil keputusan dan menentukan produksi atau barang yang harus tersedia oleh sebuah perusahaan atau toko penjualan. Selain demikian prediksi atau peramalan dapat membantu pihak pengusaha atau toko dalam perencanaan penyediaan stok karena prediksi memberikan hasil keluaran terbaik sehingga dapat meminimalisir resiko kesalahan dari perencanaan. Prediksi digunakan sebagai penentu atau menemukan informasi dari sejumlah data besar sehingga memerlukan *Data mining* (Yolanda & Fahmi, 2021)

2.5 Algoritma *K-Nearest Neighbor*

K-Nearest Neighbor merupakan suatu metode yang menggunakan algoritma *supervised learning* dimana hasil dari *Sample* uji yang baru di klafikasikan berdasarkan mayoritas dari kategori pada K-NN sehingga menemukan tingkat akurasi yang tinggi (Utami & Handoko, 2020).

Algoritma KNN juga dapat digunakan untuk menghitung jarak antara dua titik. Untuk melakukan ini algoritma menggunakan rumus jarak Euclidean, yang menghitung jarak antara dua titik dengan menggunakan persamaan matematika yaitu: $\sqrt{(x_2-x_1)^2+(y_2-y_1)^2}$.

Algoritma ini juga dapat digunakan untuk mengklasifikasikan data berdasarkan jarak terdekat dengan data *training*. Dengan menggunakan Algoritma *K-Nearest Neighbor*, nilai dari K dalam Algoritma KNN ditentukan oleh jumlah tetangga terdekat yang ingin kita cari. Jika K bernilai 1, maka hanya satu tetangga terdekat yang akan dicari, jika K bernilai 2, maka dua tetangga terdekat (Yolanda & Fahmi, 2021). Terdapat Langkah-langkah Algoritma KNN (Singingi et al., 2020) Pilih jumlah k yang diinginkan, Menghitung jarak antara setiap data dengan *data testing*, Pilih k tetangga terdekat, Mengambil keputusan berdasarkan keterbacaan , Berikan label data.

2.5.1 Jarak Euclidean

Namun pada penelitian ini hanya menggunakan jarak *Euclidean* maka adapun langkah-langkah *K-Nearest Neighbour* sebagai berikut (Yolanda & Fahmi, 2021):

1. Menentukan parameter K = jumlah tetangga terdekat.
2. Hitung jarak antara data baru dengan semua data *training*. Berikut rumus perhitungan jarak dengan *Euclidean* seperti di bawah ini:

$$d_i = \sqrt{\sum_{k=1}^i (x_i - x_i)^2}$$

Dimana :

X_1 : *Sampel data / data training*

X_2 : Data uji / *testing*

d : Jarak

k : dimensi atribut / data

3. Urutkan jarak tersebut dan tetapkan tetangga terdekat berdasarkan jarak minimum ke-K.
4. Periksa kelas dari tetangga terdekat.
5. Gunakan mayoritas sederhana dari kelas tetangga terdekat sebagai nilai prediksi data baru.

2.5.2 Kelebihan dan Kekurangan Algoritma *K-Nearest Neighbour*

2.5.2.1 Kelebihan Algoritma *K-Nearest Neighbour*

1. *Robust* terhadap data yang *noise*
2. Efektif jika *training* data berjumlah banyak

2.5.2.2 Kekurangan Algoritma *K-Nearest Neighbour*

1. Perlu menunjukkan parameter K(jumlah tetangga terdekat)
2. Berdasarkan perhitungan nilai jarak(*Distance Based Learning*) tidak jelas perhitungan jarak mana yang sebaiknya digunakan dan *atribut* mana yang memberikan hasil yang baik
3. Nilai komputasinya tinggi karena perlu menghitung jarak dan nilai baru ke semua data yang ada di data *training*.

2.5 *Software* Pendukung

Software pendukung ialah perangkat tambahan yang digunakan oleh peneliti untuk mendukung penelitiannya. Berikut *software* pendukung yang digunakan oleh peneliti:

2.5.1 *Rapid miner*

Rapid miner atau yang sering di sebut dengan nama lain *YALE* merupakan *software* yang berdiri sendiri untuk dapat menganalisis data dan sebagai sebuah solusi untuk melakukan analisis terhadap *Data mining*, *text mining* dan analisis prediksi. *Rapid miner* juga merupakan perangkat lunak bersifat terbuka (*open source*). *Rapid miner* memiliki kurang lebih 500 operator *Data mining* termasuk *operator input*, *output*, *data preprocessing* dan visualisasi (Sawit et al., 2019).

Menurut Kdnuggets berdasarkan jurnal penelitian (Putri, 2021) *RapidMiner* juga menyediakan berbagai jenis alat *data mining* seperti algoritme klasifikasi, regresi, kluster, pengelompokan, asosiasi, dan *data mining*. Alat ini dapat digunakan untuk mengoptimalkan proses *data mining* dan membuat analisis data lebih mudah. *Rapidminer* juga dapat menyediakan berbagai jenis perencanaan *data mining*, seperti pemilihan metode dan pengaturan parameter. Ini membuat lebih mudah bagi pengguna untuk mengimplementasikan proses *data mining* yang lebih efektif. Selain itu, *Rapidminer* juga menyediakan berbagai alat visual untuk menganalisis data dan mengevaluasi hasil. Alat ini membuat lebih mudah untuk mengambil kesimpulan tentang data y.

Terdapat sifat-sifat pada rapid minner:

1. Intuitif - *Rapidminer* memungkinkan pengguna untuk dengan cepat memahami bagaimana alat ini beroperasi dengan menyediakan antarmuka yang mudah digunakan dan intuitif.
2. Fleksibel - *Rapidminer* memungkinkan pengguna untuk membuat proses ML yang disesuaikan dengan kebutuhan individu.

3. Mudah Diintegrasikan - *Rapidminer* dapat dengan mudah diintegrasikan dengan berbagai sistem dan basis data, sehingga cocok untuk berbagai aplikasi.
4. *Scalable* - *Rapidminer* dapat diperluas untuk menangani tugas berukuran besar.
5. *Open Source* - *Rapidminer* merupakan perangkat lunak *open source*, sehingga dapat dengan mudah dikembangkan dan ditingkatkan.



Gambar 2. 2 *Rapid miner*
Sumber:(Syahputra et al., 2020)

2.5.2 *Microsoft excel*

Microsoft excel merupakan sebuah aplikasi yang berfungsi untuk dapat mengolah data secara otomatis menggunakan berbagai bentuk rumus perhitungan dasar, pengolahan data, pembuatan tabel, pembuatan grafik hingga manajemen data. *Excel* juga menyediakan formula yang dapat membantu memudahkan untuk melakukan perhitungan dalam pengolahan data tersebut. Aplikasi *microsoft excel* berupa lembaran kerja atau *spreadsheet* yang memiliki fitur-fitur kalkulasi yang baik(Yolanda & Fahmi, 2021).

2.6 Penelitian Terdahulu

Pada penelitian ini, penulis mengacu pada penelitian lain sebagai referensi.

Berikut penelitian terdahulu yang digunakan oleh peneliti:

1. Berdasarkan penelitian (Laksamana, 2021) dengan judul “Penerapan *Data mining* untuk memprediksi jumlah total produksi hcl pada perusahaan PT.Lontar Papyrus menggunakan algoritma regresi linear berganda” Memberi penjelasan bahwa *Data mining* merupakan proses menemukan sebuah hubungan dengan data yang tidak diketahui oleh pengguna dan menyajikan data dengan cara yang dapat dipahami sehingga hubungan tersebut dapat menjadi dasar pengambilan keputusan. *Data mining* juga merupakan kegiatan yang mencakup pengumpulan data, pemakaian data historis yang menjelaskan keteraturan, pola dalam hubungan di dalam set data berukuran besar.
2. Berdasarkan penelitian (Putri, 2021) dengan judul “Penerapan *Data mining* untuk memprediksi penjualan buah dan sayur menggunakan metode *K-Nearest Neighbor*” memberi penjelasan bahwa Terdapat Langkah-langkah Algoritma KNN (Singingi et al., 2020) Pilih jumlah k yang diinginkan, Menghitung jarak antara setiap data dengan data test, Pilih k tetangga terdekat, Mengambil keputusan berdasarkan keterbacaan , Berikan label data.
3. Berdasarkan penelitian (Syahputra, 2020) dengan judul “*Data mining* menentukan prediksi stok barang pada PT.Siantar Top tbk Medan dengan menggunakan metode regresi linear berganda” memberikan penjelasan bahwa Menurut sifatnya operasi *Data mining* dibedakan menjadi dua bagian yaitu 1

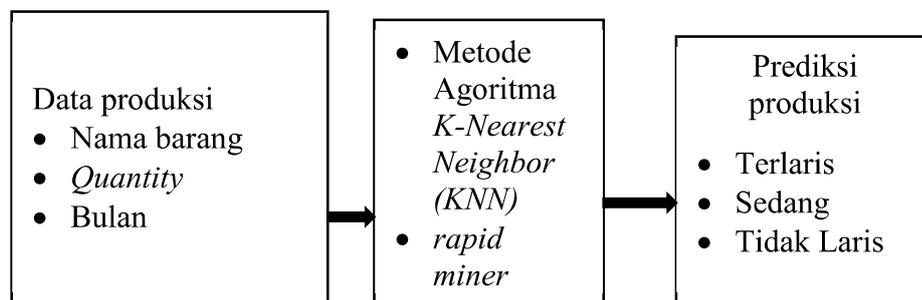
prediksi (*prediction driven*) digunakan sebagai penjawab pertanyaan tentang sesuatu yang bersifat abstrak. Operasi prediksi digunakan untuk validasi hipotesis, querying dan laporan. Sifat ke 2 ialah penemuan (*discovery driven*) bersifat transparan untuk menjawab pertanyaan “mengapa”. Operasi penemuan digunakan untuk analisis data eksplorasi, pemodelan prediktif, segmentasi *database*.

4. Berdasarkan penelitian (Sawit, 2019) dengan judul “ *Data mining* untuk memprediksi hasil produksi buah sawit pada PT. Bumi Sawit Sukses (BSS) menggunakan metode *K-Nearest Neighbor*” memberi penjelasan bahwa Klusterisasi merupakan pempratisi *data set* menjadi beberapa *sub net* atau kelompok dengan sedemikian rupa sehingga elemen-elemen dari suatu kelompok tertentu menjadi *set property* yang di bagikan secara bersamaan.
5. Berdasarkan penelitian (Singingi, 2020) dengan judul “ Penerapan algoritma *K-Nearest Neighbor* dalam klasifikasi data hasil produksi kelapa sawit pada KUD Tirta Kencana “Memberi penjelasan bahwa Klasifikasi merupakan penentu sebuah *record* data baru ke salah satu dari beberapa kategori (kelas) yang telah didefinisikan sebelumnya dan disebut juga dengan “*supervised learning*”.
6. Berdasarkan penelitian (Yolanda, 2021) dengan judul “ Penerapan *Data mining* untuk memprediksi penjualan produk roti terlaris pada PT. Nippon Indosari Corpindo Tbk menggunakan metode *K-Nearest Neighbor*” Menjelaskan bahwa *Rapid miner* atau yang sering di sebut dengan nama lain *YALE* merupakan *software* yang berdiri sendiri untuk dapat menganalisis data

dan sebagai sebuah solusi untuk melakukan analisis terhadap *data mining*, *text mining* dan analisis prediksi.

2.7 Kerangka Pemikiran

Kerangka pemikiran merupakan suatu ide yang dimiliki oleh peneliti tentang suatu pokok permasalahan yang akan dicari dan diselesaikan mulai dari tahap awal hingga tahap akhir dari suatu penelitian.



Gambar 2. 3 Kerangka Pemikiran

Sumber: Data Penelitian 2022

Keterangan:

1. Data produksi merupakan database yang ada di Takadeli dan di jadikan sebagai data training untuk untuk dapat di olah. Data penjualan takadeli Nama barang, *quantity*, bulan.
2. Algoritma KNN, Algoritma ini digunakan untuk melakukan klasifikasi terhadap objek berdasarkan data pembelajaran yang jaraknya paling dekat dengan objek tersebut dan di kelola dengan menggunakan *software Rapid minner* dan perhitungan manual dengan *microsoft excel*.
3. Prediksi produksi merupakan sebagai sebuah solusi untuk melakukan analisis semple uji terhadap data uji atau data training dengan hasil yang akan di capai kategori grade terlaris=A, sedang=B, tidak laris=C.