BAB II
TINJAUAN PUSTAKA

2.1 *Knowledge Discovery in Database* (KDD)

Pola dan hubungan big data dapat ditemukan melalui penggunaan data historis, teknik yang dikenal sebagai penemuan pengetahuan melalui *data mining* (KDD). KDD sering digunakan untuk menggambarkan proses penyaringan melalui database besar untuk informasi yang relevan. KDD menurut (Mardianti and Fauzi, 2020) merupakan kegiatan yang digunakan dalam pengumpulan data, dan mengolah data untuk menemukan pola hubungan yang teratur dalam basis data yang besar. Saat ini, KDD semakin populer karena penurunan kebutuhan akan pengenalan pola yang telah terjadi sebagai akibat dari perkembangan ini. Hasil penambangan data dapat digunakan di masa depan untuk membuat keputusan yang lebih tepat tentang masa depan. Menurut (Buulolo, 2020) *Knowledge Discovery in Database* (KDD) mencakup proses lengkap mengekstraksi atau menemukan pola, pengetahuan, dan informasi yang berpotensi berguna dari kumpulan data dalam jumlah besar, dimana pengetahuan dan informasi yang dihasilkan dari KDD bersifat sah, baru, mudah di mengerti serta bermanfaat.

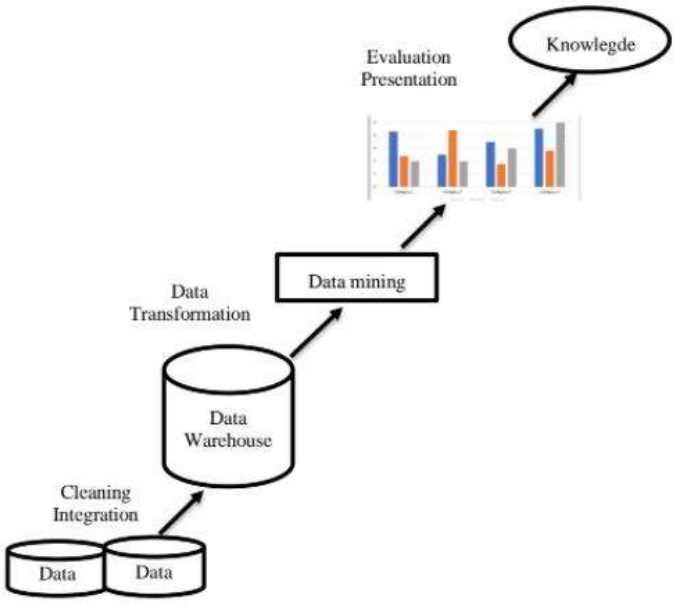
Menurut Tomar, Agarwal dalam (Widaningsih, 2019) KDD adalah proses menemukan informasi baru yang berharga yang lebih mudah dipahami daripada sistem penyimpanan data yang besar dan kompleks. Dalam proses KDD, hasil dari kumpulan data ditafsirkan dengan menyatukan kumpulan data dengan informasi

dari bidang lain. Penting untuk dicatat bahwa proses KDD dimulai dan diakhiri dengan penetapan dan evaluasi tujuan.

Atas dasar hal di atas, dapat dinyatakan bahwa KDD adalah teknik untuk menemukan atau mengekstraksi informasi dari toko data besar melalui pengumpulan dan pemrosesan data untuk menghasilkan data atau informasi baru yang mudah dipahami dan berguna.

Tahapan Proses *Knowledge Discovery in Database* (KDD)

(Buulolo, 2020) menyatakan bahwa proses KDD dapat dipecah menjadi beberapa langkah, seperti yang ditunjukkan pada Gambar 1.1.



Gambar 1. 1 Tahapan Proses KDD
Sumber : (Buulolo, 2020)

1. Data

   Hal pertama yang harus dipersiapkan dalam proses KDD yaitu data. Data yang di gunakan merupakan data yang sudah terpisah dengan data operasional.

2. *Selection*

   Perlu dilakukan pemilihan data karena tidak semua data yang ada dapat dipergunakan. Membuat kumpulan data target, menentukan variabel, memilih sampel data, dan menyimpan data dalam file adalah contoh kegiatan pemilihan data.

3. *Pre-processing (Cleaning)*

   Pada titik ini, data yang dipilih akan dibersihkan. Setelah menghapus duplikat dan menyelesaikan inkonsistensi informasi, proses pembersihan selesai. Data juga dapat diperkaya dengan menambahkan informasi tambahan yang relevan, yang dikenal sebagai enrichment, selama tahap praproses ini.

4. *Transformation*

   Ada berbagai algoritma dan metode yang dapat digunakan dalam *data mining*. Seperti berdiri, format data yang dibutuhkan oleh setiap algoritma atau metode berbeda. Ketika proses KDD digunakan, data yang telah disiapkan sebelumnya dimasukkan ke dalamnya, dan setiap perubahan yang diperlukan dilakukan terlebih dahulu.

5. *Data Mining*

Tahap utama KDD adalah *data mining*. Pengetahuan dan informasi dapat ditambang dari data menggunakan algoritma atau metode tertentu berdasarkan pengetahuan atau informasi yang sedang dicari.

6. *Interpretation (Evaluation)*

Proses *data mining* menghasilkan pengetahuan atau informasi yang dapat dengan mudah dipahami oleh pihak terkait, seperti informasi yang ditampilkan dalam bentuk grafik, pohon keputusan, atau aturan. Periksa untuk melihat apakah proses *data mining* menghasilkan pengetahuan atau informasi yang berlainan dengan keyakinan atau fakta yang dimiliki sebelumnya.

7. *Knowledge*

Tujuan utama dari proses KDD adalah untuk mengumpulkan pengetahuan atau data yang berguna dan dapat dimengerti. Sesuai dengan manfaat atau kegunaan dari pengetahuan atau informasi yang dihasilkan, itu dilaksanakan.

2.2 *Data Mining*

Data mining ialah metode yang dipakai buat menciptakan pengetahuan terkini dalam sejumlah besar data. Terdapat beberapa tahap yang wajib diiringi buat melaksanakan prosedur ini (Sitepu and Buulolo, 2017). Tujuan dari *data mining* adalah untuk membuat database dari data yang sebelumnya tidak tersedia. Menganalisis data didefinisikan sebagai proses mencari dan mengidentifikasi koneksi di antara kumpulan data yang berbeda. Data dalam database dapat disaring

secara efisien menggunakan hubungan ini. Ekstraksi informasi terkini dari sejumlah besar data, yang menunjang dalam pengambilan ketentuan, diketahui sebagai *data mining.* Istilah "*knowledge discovery*" kadang-kadang digunakan untuk menggambarkan penambangan data.

Sedangkan menurut (Arhami and Nasir, 2020) *data mining* adalah proses "menggali" melalui data untuk menemukan informasi atau pengetahuan baru yang bermanfaat bagi pengguna. Mengumpulkan, mengekstraksi, menganalisis, dan melaporkan data semuanya termasuk dalam proses *data mining*. Analisis pola data tersembunyi juga dikenal sebagai *data mining*. Analisis gudang data, algoritma *data mining*, dan memfasilitasi pengambilan keputusan bisnis dan informasi lainnya dikumpulkan di area umum dan digunakan untuk tujuan ini.

2.2.1 Pengelompokan *Data Mining*

Menurut (Buulolo, 2020) *data mining* dibagi menjadi berbagai kategori tergantung pada tugas yang dapat diselesaikan, termasuk:

1. Deskripsi: Guna mempromosikan kegiatan, perlu untuk mengidentifikasi atau menemukan pola yang terjadi secara teratur. Ketika pelanggan sering membeli produk A dan B pada saat yang sama, manajemen supermarket memindahkan A ke lokasi yang berdekatan di katalog. Untuk membuatnya sesederhana mungkin bagi pelanggan untuk membeli produk.

2. Klasifikasi: Berdasarkan korelasi antara kondisi dan variabel target. Bencana alam, misalnya, dipecah menjadi tiga kategori: bencana, sedang, dan non-bencana.
3. Prediksi: Prediksi dan klasifikasi sangat mirip. Peramalan adalah fungsi penambangan data yang umum. Nilai hasil prediksi akan digunakan di masa depan berdasarkan data sebelumnya.
4. Estimasi: Menurut definisi istilah "estimasi," itu adalah prediksi; Namun, berbeda dari klasifikasi dalam bahwa estimasi adalah angka daripada pengelompokan abjad.
5. Pengklasteran: Pengklasteran adalah sekelompok data yang memiliki karakteristik umum (tipe yang sama). Pengamatan, catatan data, atau kelas dan objek serupa adalah contoh data yang dapat dikelompokkan ke dalam Pengklasteran.
6. Asosiasi: Asosiasi adalah pengelompokan, serikat pekerja, atau jenis pengelompokan lainnya. Contoh umum dari *data mining* adalah mengejar atribut yang muncul atau selalu muncul bersamaan dengan pembelian lebih dari satu item, seperti ketika Anda membeli produk A, Anda juga membeli B, Anda juga membeli B, Anda juga membeli C dan sebagainya.

2.2.2 Prediksi

Peramalan adalah kombinasi dari seni dan sains. Prediksi Adalah mungkin untuk memprediksi nilai masa depan dengan berfokus pada data dan informasi yang relevan, apakah itu informasi serta data dari periode lalu ataupun data dari periode saat ini. Prediksi dapat dilakukan dengan berbagai cara, termasuk keduanya:

1. Metode Kualitatif

Model matematika tidak diperlukan untuk menggunakan metode ini. Akibatnya, tidak mungkin untuk memprediksi masa depan hanya berdasarkan data yang dikumpulkan (peramalan jangka panjang). Adalah umum bagi perkiraan kualitatif untuk mengandalkan pendapat para ahli di bidangnya.

Selain itu, metode ini hemat biaya karena tidak memerlukan banyak data dan dapat diperoleh dalam hitungan menit. Namun, kelemahan metode ini adalah sifat subjektif dari data yang dihasilkannya.

2. Metode Kuantitatif

Untuk memprediksi masa depan, metode ini bergantung pada data mentah dan aturan matematis yang menyertainya. Metode kuantitatif memiliki banyak model yang berbeda untuk membuat prediksi, seperti:

a. Model-model regresi

Memprediksi variabel yang mempunyai ikatan linier dengan variabel bebas yang dikenal serta bisa diharapkan ialah tujuan dari model ini.

b. Model Ekonometrik

Segmen ekonomi dirangsang oleh variabel independen dalam persamaan regresi yang digunakan dalam model.

c. Model *Time Series Analysis* (Deret Waktu)

Ini adalah model yang menggunakan data dari masa lalu untuk memprediksi tren masa depan berdasarkan data saat ini.

2.2.3 Time Series Analisys

Istilah "*Time Series*" mengacu pada kumpulan data yang mencakup periode waktu tertentu. Prediksi data masa depan menggunakan persamaan matematika dan statistik juga merupakan bagian dari peramalan deret waktu. Data *Time Series* tersedia dalam berbagai bentuk, termasuk: (Octavia, Tanti and Yulia, Yulia dan Lydia, 2015):

a. Siklus

Pola seperti ini dapat dilihat dengan cara pola ini naik dan turun. Pola periodik dapat ditemukan dengan menghapus pola musiman dari kumpulan data jika digunakan mingguan atau bulanan.

b. Random

Pola acak yang tidak dapat digambarkan dengan menggambar. Sebagai akibat dari keadaan darurat, tidak ada cara untuk memprediksi atau menjelaskan ketidakteraturan pola ini.

c. *Trend*

Peningkatan atau penurunan komponen jangka panjang pola data dapat dilihat sebagai tren karena pola tren berubah dari waktu ke waktu.

d. Musiman

Dalam sebuah pola, gerakan terbukti diulang dari waktu ke waktu. Data plakat yang dikumpulkan mingguan dan bulanan dapat menunjukkan tren ini. Metode *moving averages, smoothing expenential* musim dingin klasik adalah metode umum untuk menentukan nilai pola musiman.

Teknik peramalan *Time series* terdiri atas:

1. "Statisika"
    a. *Moving Average*
    b. *Exponential Smoothing*
    c. Regresi
    d. *ARIMA (Box Jenkins*)
2. Kecerdasan buatan
    a. *Neural Network*
    b. Algoritma Genetika
    c. *Simulated Annealing*
    d. *Genetic Programming*
    e. Klasifikasi
    f. *Hybrid"*

2.2 Metode *Data Mining*

Metode klasifikasi digunakan untuk menemukan model yang menggambarkan dan membedakan kelas konseptual data oleh penulis dalam riset ini. Data yang dipakai guna membangun model ini diperoleh lewat pemeriksaan *data set* praktik. Dengan menggunakan model ini, bisa memperkirakan label kelas dari objek yang tak dikenal (Nikmatun and Waspada, 2019).

Menurut (Muslim *et al.*, 2019) Klasifikasi adalah strategi untuk memprediksi kelas entitas yang labelnya tidak diketahui dengan menemukan model yang menjelaskan atau membedakan gagasan atau kelas fakta. Sedangkan menurut (Iriadi & Nuraeni, 2016) dalam (Wijaya and Fauzi, 2020), klasifikasi data ialah proses dalam menemukan suatu properti-properti sama pada himpunan objek pada database, serta mengklasifikasikan ke kelas-kelas berbeda. Tujuannya yaitu menemukan model pada *training set* yang membedakan antara atribut ke dalam kategori atau kelas yang sesuai.

Dari sekian banyak teknik dalam pengklasifikasian, diantaranya yaitu *K-Nearest Neighbour, Naïve Bayes, Decision Tree, Neural Networks-Support Vector Machine* dan lain-lain.

2.3 Algoritma *K-Nearest Neighbour* (K-NN)

Sejak tahun 1970, teknik *K-Nearest Neighbour* (K-NN) telah digunakan dalam estimasi statistik dan pengenalan pola (Arhami and Nasir, 2020). K-NN adalah salah satu pendekatan non-parametrik pada saat itu. K-NN adalah algoritma klasifikasi atau pendekatan yang banyak digunakan dalam *data mining*. K-NN juga

termasuk kedalam kategori regresi yang juga dapat digunakan untuk memprediksi seperti hal nya regresi.

Secara umum, ada 2 komponen yang paling penting dalam K-NN, yaitu :

1. K sebagai parameter yang akan melingkupi sejauh mana atau sejumlah data mana yang akan menjadi ukuran untuk pertimbangan penentuan label atau kelas dari objek latih.
2. Jarak, jarak antara item data yang akan diuji dan semua objek data pelatihan yang diketahui harus diketahui sehingga penempatan objek pelatihan yang lebih dekat dengan tetangganya dapat ditentukan. Untuk menentukan jarak terdekat tersebut maka dapat dipilih fungsi atau metode jarak mana pun yang sesuai dengan kasus yang akan diselesaikan. Jarak yang berbeda akan menentukan posisi atau letak yang berbeda.

Tahapan metode *K-Nearest Neighbor*, antara lain:

1. Parameter k (jumlah tetangga terdekat) ditentukan.
2. Menghitung jarak antara data yang akan dinilai dan data yang akan dievaluasi dengan pelatihan penuh;
3. Mengurutkan jarak yang dihasilkan;
4. Menentukan jarak yang paling dekat dengan urutan;
5. Menempatkan kelas yang sebanding bersama-sama;
6. Carilah jumlah kelas yang dibagikan oleh tetangga yang berdekatan dan tetapkan kelas-kelas tersebut untuk dinilai sebagai kelas data.

Rumus Euclidean Distance persamaan 2.1 dapat digunakan untuk menghitung jarak antara dua titik, seperti titik data *training* dan titik data *testing*.

Persamaan 2.1 adalah rumus untuk jarak komputasi menggunakan algoritma K-Nearest Neighbor (K-NN) dan *Euclidean Distance*:

$$d\ (P,Q) = \sqrt{\sum_{i=1}^{n}(Pi - Qi)^2}$$ **Rumus 2. 1** Perhitungan Jarak *Euclidean*

Dimana

$d(P,Q)$	: jarak *euclidien*
n	: jumlah data *training*
P	: inputan data ke -1 dari data *training*
Q	: inputan data ke -1 dari data *testing*"

Ada sejumlah teknik pembelajaran berbasis kasus yang termasuk dalam pendekatan K-Nearest Neighbor (K-NN) untuk pembelajaran mesin. Algoritma juga merupakan metode pembelajaran yang tidak efisien. Di K-NN, kelompok objek terdekat (serupa) dengan objek ditemukan menggunakan data baru atau data uji yang berasal dari data pelatihan (Islami, 2018). Data pembelajaran yang paling dekat dengan objek digunakan untuk mengklasifikasikannya menggunakan algoritma *K-Nearest Neighbor*. Kasus-kasus baru ditemukan dengan menghitung jarak antara kasus baru dan kasus lama yang cocok dengan sejumlah fungsi yang ada berdasarkan berat kasus baru. Hal ini dimungkinkan untuk menggunakan rumus *Euclidean* untuk mengetahui jarak antara dua titik, yaitu x dalam data *training* dan y dalam data *testing*.

D adalah jarak antara titik pada data pelatihan x dan titik pengujian data y yang akan dikategorikan, di mana x=x1,x2,…,xi dan y=y1,y2,…,yi dan I menunjukkan nilai atribut dan n menunjukkan dimensi atribut.

Tahapan menghitung metode Algoritma *K-Nearest Neighbor*, yakni:

1. Parameter K (Jumlah tetangga terdekat) sedang ditentukan.
2. Menghitung kuadrat setiap objek dari jarak *euclid* (*instance query*) menggunakan contoh data yang diberikan;
3. Setelah itu, atur item ke dalam kelompok dengan jarak Euclidean terpendek;
4. Mengumpulkan kategori Y (Klasifikasi *Nearest Neighbor*);
5. Sebagian besar nilai *instans query* yang telah ditentukan dapat diantisipasi dengan menggunakan kategori *Nearest Neighbor*.

2.4 *Software* Pendukung

Gambar 2. 1 Logo *Rapidminer*
Sumber : (*www.rapidminer.com*, 2021)

Pada penelitian ini peneliti menggunakan *software RapidMinner*. *RapidMinner* merupakan *software data mining* gratis untuk keperluan akademik dimana bertujuan untuk memproses *data mining* dengan cakupan analisis teks, mengestrak pola yang ada dari kumpulan data besar dengan menggunakan metode statistik, kecerdasan buatan, dan database untuk digabungkan. Tujuan analisis ini yakni guna mendapat informasi dengan nilai kriteria tertinggi dari data yang diolah.

2.5 Penelitian Terdahulu

Penelitian bergantung pada penelitian terdahulu. Berikut penelitian yang dapat berfungsi sebagai panduan untuk penelitian ini, termasuk:

1. Peneltian oleh (Sitepu and Buulolo, 2017) Nomor 1 Vol.7 dengan judul **"IMPLEMENTASI ALGORITMA *NEAREST NEIGHBOR* PADA PENERIMAAN PEGAWAI BARU PADA MTS IKHWANUTS TSALITS TALUN KENAS"**, penelitian ini menyatakan bahwa K-NN efektif dalam sistem penerimaan pegawai dengan cara penghitungan nilai kedekatan diantara kasus lama dengan kasus baru. Dikarenakan setiap nilainya dapat ditentukan oleh pengguna (*user*) dengan persyaratan yang di tentukan yaitu nilai IPK, nilai TOEFL, prestasi, pengalaman, umur, dan status. Maka dari itu penelitian ini dapat dijadikan rujukan.
2. Penelitian (Winarso and Arribe, 2017) Nomor 2 Vol.8 dengan judul **"SELEKSI PEGAWAI DAN DOSEN UMRI BERBASIS *E-RECRUITMENT* MENGGUNAKAN METODE *K-NEAREST NEIGHBOR*"**, menyatakan bahwa Dalam seleksi administrasi, metode K-NN dapat digunakan untuk memilih calon karyawan dan dosen dengan

menghitung kesamaan antara persyaratan dan data pelamar. Universitas Muhammadiyah Riau juga telah berhasil mengembangkan sistem e-recruitment untuk mengevaluasi calon karyawan dan dosen. Penelitian ini dapat berfungsi sebagai panduan.

3. Penelitian oleh (Khasanah, Harjoko and Candradewi, 2016) Nomor 2 Vol. 6 yang berjudul "**KLASIFIKASI SEL DARAH PUTIH BERDASARKAN CIRI WARNA DAN BENTUK DENGAN METODE *K-NEAREST NEIGHBOR* (K-NN)"** menyatakan Mengklasifikasikan sel darah berdasarkan jenisnya menggunakan mikroskop telah lama menjadi praktik standar di laboratorium hematologi. Laboratorium hematologi menggunakannya sebagai alat diagnostik dan pemantauan utama. Ada juga prosedur manual yang memakan waktu untuk lulus serangkaian tes laboratorium. Akibatnya, penelitian ini berfokus pada tahap awal klasifikasi otomatis jenis sel darah putih di bidang medis. Morfologi sel darah dapat digunakan dalam teknik pengolahan gambar untuk mengatasi masalah waktu dan diagnosis dini. Peneliti menggunakan K-Nearest Neighbor untuk mengklasifikasikan sel darah putih berdasarkan morfologi sel (K-NN). Hough circle, threshold, dan feature extraction adalah algoritma pemrosesan gambar yang digunakan. Akhirnya, klasifikasi K-Nearest Neighbor (K-NN) digunakan. Untuk mengidentifikasi jenis gambar, 100 gambar dianalisis. Hasil dari tes segmentasi dan klasifikasi menunjukkan tingkat akurasi masing-masing 78% dan 64%.

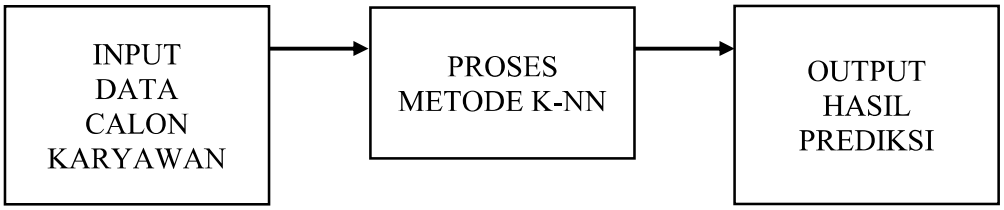
4. Penelitian (Lizarti and Ulfah, 2019) Nomor 1 Vol.4 berjudul **"PENERAPAN ALGORITMA *K-NEAREST NEIGHBOR* UNTUK PENENTUAN PEMINATAN STUDI STMIK AMIK RIAU"** menyatakan Mahasiswa di STMIK Amik Riau didorong untuk mengejar minat akademik mereka berdasarkan bakat dan minat mereka yang unik. Mahasiswa dalam Program Studi Teknik Informatika STMIK Amik Riau mengkhususkan diri dalam bisnis dan jaringan. Kemampuan dan minat siswa harus dipertimbangkan ketika memilih minat. Nilai mahasiswa dan tingkat kelulusan sangat dipengaruhi oleh tingkat minat mereka dalam kursus. Seperti berdiri, pemilihan minat belajar mahasiswa dibuat semata-mata atas dasar persahabatan dan bukan kemampuan akademik. Karena diyakini memberikan saran yang baik dan tepat untuk belajar, itu menjadi jawaban atas masalah memilih jurusan atau konsentrasi studi. Pengelompokan data dapat dilakukan dengan menggunakan algoritma klasifikasi *K-Nearest Neighbor* (K-NN). Data nilai kursus prasyarat dari semester pertama hingga kelima digunakan dalam penelitian ini. Algoritma K-NN dapat diimplementasikan menggunakan aplikasi berbasis PHP dan *MySQL*. Jika dibandingkan dengan hasil yang dihitung secara manual, output sistem 100% akurat. Untuk menguji kinerja algoritma, gunakan alat *RapidMiner*. Algoritma K-NN bekerja dengan baik pada 183 data pelatihan dan 100 data pengujian, dengan *acuracy*, *Recall*, *Precision*, F *Measure*, dan *Classification Error* hasil masing-masing 98%, 100%, 100%, 91,67%, dan 2%. Mahasiswa Teknik Informatika STMIK Amik

Riau dapat memperoleh manfaat dari penelitian ini dengan menerima saran minat studi.

5. Penelitian (Imron and Kusumah, 2018) Nomor 1 Vol. 1 dengan judul **"*APPLICATION OF DATA MINING CLASSIFICATION METHOD FOR STUDENT GRADUATION PREDICTION USING K-NEAREST NEIGHBOR* (K-NN) *ALGORITHM*"** menyatakan bahwa tingkat kelulusan mahasiswa merupakan salah satu indikator untuk meningkatkan akreditasi suatu mata kuliah. Diperlukan untuk memantau dan mengevaluasi kelulusan siswa kecenderungan, tepat waktu atau tidak. Salah satunya adalah memprediksi tingkat kelulusan dengan memanfaatkan teknik *data mining*. Metode Klasifikasi *Data mining* yang digunakan adalah algoritma *K-Nearest Neighbor* (K-NN). Data yang digunakan berasal dari data siswa, data nilai siswa, dan data kelulusan siswa untuk tahun 2010-2012 dengan total 2.189 catatan. Atribut yang digunakan adalah jenis kelamin, asal sekolah, program studi IP Semester 1-6. Hasil menunjukkan bahwa metode K-NN menghasilkan akurasi yang tinggi sebesar 89,04%.4.

2.6 Kerangka Pemikiran

Perlu menggunakan metode K-NN untuk memprediksi perekrutan karyawan di PT. DSAW. Perangkat lunak *RapidMiner* digunakan dalam penelitian ini untuk menganalisis data sebelumnya dan memprediksi efisiensi rekrutmen. Kerangka penelitian didasarkan pada latar belakang dan metode yang digunakan:



Gambar 2.1 Kerangka pemikiran

Sumber : Data Penelitian (2021)