BAB II
TINJAUAN PUSTAKA

2.1. Penelitian Terkait

Penelitian tentang prediksi prestasi siswa yang telah dilakukan oleh beberapa peneliti sebelumnya dan menjadi latar belakang penelitian ini dijabarkan pada tabel dibawah ini:

Tabel 2.1 . Penelitian Terkait

NO	JUDUL /PENULIS /TAHUN	METODE	HASIL	KEKURANGAN
1	"Data mining Aproach for Predicting Student Performance"/ Edin Osman Begovic, Mirza Suljic/2015(Osman begovi, 2015)	Naïve Bayes. MLP, J48	Naïve Bayes sebesar 76,65%, Multi Layer Percepton (MLP) 71,20%, J48 73,93	Hasil akurasi masih belum optimal algoritma yang menghasilkan nilai tertinggi Naive Bayes dengan tingkat akurasi 76,65 %
2	"Educational Data Mining & Student Performance Prediction Using SVM Techniques/ Mrs. Jamuna dan Mrs.S.A. Shoba(Jamuna & Shoba, 2017)/2017	MLP, NN, NB Tree, REPTree	Multilayer Perceptron (MLP) & Neural Network sebesar 74,80%, ID3 sebesar 73%, NBTree dan REPTree sebesar 71 %, J48, Simple Cart, Decision Table sebesar 68,8%; 69,5%; 68,1%	Hasil akurasi masih belum optimal algoritma yang menghasilkan nilai tertinggi Neural Network dengan tingkat akurasi 74,80 %
3	"Data Mining Prestasi Akademik Dengan Naïve Bayes Berdasarkan *Attribut Importance* (AI)/Ni Komang Sri Julyantari, I Ketut Dedy Suryawan(Komang et al., 2013)/2017	Naïve Bayes, AI	Naïve Bayes = 74,55 % AI	Hasil akurasi masih belum optimal algoritma yang mengahsilkan nilai tertinggi Naive Bayes dengan tingkat akurasi 74,55 %

4	"Analysis of Student Data base Using Classification Techniques/K. Sumathi, Ph.D, S. Kannan, Ph.D, K.Nagarajan(Sumathi et al., 2016)/2017	DT, J48	DT = 77,12 %	Hasil akurasi masih belum optimal algoritma yang mengahsilkan nilai tertinggiDT dengan tingkat akurasi 77,12 %
5	"Student Performance Prediction Using Data Mining Techniques"/Durges Ugale, Jeet Pawar, Sacim Yadav, dan Dr. Chandra Sekar Raut/2018	DT, K-NN, Naive Bayes	Decision Tree sebesar 79% K-NN sebesar 72% Naïve Bayes sebesar 64 % Support Vector Machine sebesar 70%	Hasil akurasi masih belum optimal algoritma yang menghasilkan nilai tertinggi Decision Tree dengan tingkat akurasi 79 %
6	"Student Performance Prediction / Mukul Gharpure, Pushpak Chaudri, Yash Bhole, Sagar Borkar, Aashutosh Awasthi(Shetty et al., 2019)/2020	Machine Learning Alogaritma-Linear Regression	Machine Learning Alogaritma- Linear Regression = 78%	Hasil akurasi masih belum optimal algoritma yang mengahsilkan nilai tertinggi Linear Regression dengan tingkat akurasi 78%
7	"Educational Data Mining in Predicting Student Final Grades/ William Willibrodus Damopoli, Nathan Priyasadie, Amalia Zahra(The & Academy, 2021)/2021	K-NN with rapid miner	k-NN = 77,36%	Hasil akurasi masih belum optimal algoritma yang mengahsilkan nilai tertinggik-NN dengan akurasi 77,36%

Hasil performa dari jurnal yang pertama dengan membandingkan beberapa algoritma yaitu K-NN, neural network, dan naïve bayes, algoritma decision tree memiliki tingkat akurasi yang paling tinggi. Jurnal kedua dengan membandingkan

algoritma naïve bayes, MLP, dan K-NN hasil performa pemodelan algoritma naïve bayes memiliki tingkat akurasi paling tinggi. Hasil performa jurnal ketiga dengan membandingkan algoritma naïve bayes, Regressi Logistik, dan KNN, algoritma KNN memiliki tingkat akurasi yang paling tinggi, Hasil performa jurnal keempat menggunakan satu algoritma yaitu naïve bayes menghasilkan akurasi yang baik yaitu 77,36% untuk menentukan status pasien covid tetapi padajurnal ini tidak ada algoritma pembanding dalam penentuan status covid yang berpotensi untuk mngetahui adanya algoritma yang lebih baik hasil tingkat akurasinya.

Dari hasil review beberapa jurnal baik jurnal nasional dan internasional saya menyimpulkan agar data set dapat terakurasi dengan baik maka data yang dibutuhkan peneliti semakin banyak akan semakin baik akurasinya dengan jumlah data diatas 500 fitur , dan metode yang digunakan untuk beberapa jurnal yang telah direview tingkat akurasi cenderung lebih tinggi dengan menggunakan Decision Tree dengan akurasi hampir mencapai 80 % sesuai pada tabel berikut :

Tabel 2.2 Hasil Akurasi Review Jurnal

No Jurnal	Metode Terbaik dan akurasi
1	Naive Bayes dengan tingkat akurasi 76,65 %
2	Decision Tree dengan tingkat akurasi 79 %
3	Neural Network dengan tingkat akurasi 74,80 %
4	NN dengan akurasi 77,36%
5	Naive Bayes dengan tingkat akurasi 74,55 %
6	DT dengan tingkat akurasi 77,12 %
7	Linear Regression dengan tingkat akurasi 78%

2.2. Prestasi Siswa

Prestasi adalah hasil yang telah dicapai seseorang dalam melakukan kegiatan. Menurut Hamdani (2011:137) prestasi yaitu hasil dari suatu kegiatan yang telah dikerjakan, diciptakan baik secara individual maupun kelompok. Sedangkan menurut Syaiful Bahri Djamarah (2012:21) prestasi yaitu hasil dari suatu kegiatan yang telah dikerjakan, diciptakan, yang menyenangkan hati yang diperoleh dengan jalan keuletan kerja, baik secara individual maupun kelompok dalam bidang kegiatan tertentu. Untuk meraih prestasi ini seorang siswa membutuhkan keuletan dan kegigihan kerja. Prestasi siswa di sekolah menengah secara umum diasosiasikan dengan hasil belajar atau prestasi belajar dimana prestasi belajar adalah sesuatu yang diraih oleh siswa yang dilihat dalam bentuk nilai pengetahuan, nilai sikap, dan nilai keahlian. Gagne (1985:40) menyatakan bahwa prestasi belajar dibedakan menjadi lima aspek, yaitu : kemampuan intelektual, strategi kognitif, informasi verbal, sikap dan keterampilan. Menurut Bloom dalam Suharsimi Arikunto (1990:110) bahwa hasil belajar dibedakan menjadi tiga aspek yaitu *kognitif, afektif dan psikomotorik.* Sementara untuk memahami tentang pengertian belajar kita awali dengan beberapa pendapat para ahli tentang definisi tentang belajar. Cronbach, Harold Spears dan Geoch dalam Sardiman A.M (2005:20) sebagai berikut :

1. Cronbach memberikan definisi :
   "*Learning is shown by a change in behavior as a result ofexperience*"."Belajar adalah memperlihatkan perubahan dalam perilaku sebagai hasil dari pengalaman".
2. Harold Spears memberikan batasan:
   "*Learning is to observe, to read, to initiate, to try something themselves, to listen, to follow direction".*Belajar adalah mengamati, membaca, berinisiasi, mencoba sesuatu sendiri, mendengarkan, mengikuti petunjuk/arahan.
3. Geoch, mengatakan :
   "*Learning is a change in performance as a result of practice".* Adapun "belajar" adalah perubahan dalam penampilan sebagai hasil praktek.

Atau dengan kata lain dari ketiga definisi belajar itu merupakan perubahan tingkah laku atau penampilan, dengan serangkaian kegiatan misalnya dengan membaca, mengamati, mendengarkan, meniru dan lain sebagainya. Juga belajar itu akan lebih baik kalau si subyek belajar itu mengalami atau melakukannya, jadi tidak bersifat verbalistik. Belajar sebagai kegiatan individu sebenarnya merupakan rangsangan-rangsangan individu yang dikirim kepadanya oleh lingkungan. Dengan demikian terjadinya kegiatan belajar yang dilakukan oleh seorang individu dapat dijelaskan dengan rumus antara individu dan lingkungan. Dari pengertian yang telah diuraikan diatas, dapat disimpulkan bahwa prestasi yaitu hasil dari suatu kegiatan yang dilakukan secara sadar yang diciptakan baik secara individu maupun kelompok dan mendapatkan hasil. Prestasi merupakan kecakapan atau hasil kongkrit yang dapat dicapai pada saat atau periode tertentu. Yang merupakan hasil yang telah dicapai siswa dalam proses pembelajaran dan terwujud dalam hasil yang disebut prestasi belajar.

2.3. Data Mining

Secara umum ini penggalian pengetahuan atau informasi dari data base memerlukan suatu proses, dimana proses ini biasanya diistilahkan sebagai Data mining. Data mining bertugas untuk menganalisis sejumlah data untuk mencari pola yang khas atau khusus yang belum diketahui sebelumnya sperti kelompok catatan data atau cluster, anomaly atau catatan yang tidak biasa. Pola ini kemudian yang dijadikan sebagai rujukan untuk menganalisis data lebih lanjut , misal untuk mengklasifikasi dan memprediksi. Berikut ini terdapat beberapa pengertian data mining menurut para ahli, terdiri atas: Pramudiono (2006) Mengemukakan bahwa pengertian data mining adalah adalah serangkaian proses untuk menggali nilai tambah dari suatu kumpulan data berupa pengetahuan yang selama ini tidak diketahui secara manual. Larose ; Data mining adalah bidang yang digabung dari beberapa bidang keilmuan yang menyatukan teknik dari pembelajaran mesin, pengenalan pola, statistik, database, dan visualisasi untuk pengenalan permasalahan pengambilan informasi dari database yang besar(Klinkenberg & Hoffmann, 2014); Data mining merupakan pemilihan atau “menambang” pengetahuan dari jumlah data yang banyak. (Maimon

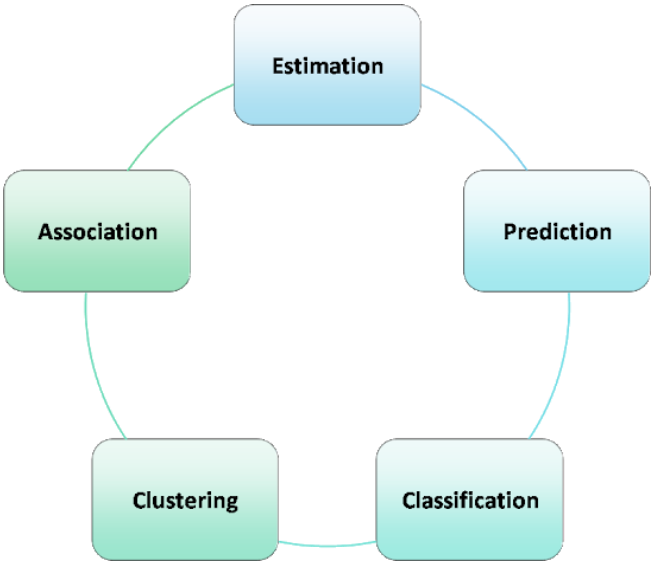
& Rokach, 2011): Data mining adalah istilah yang digunakan untuk menggambarkan proses ekstraksi nilai dari database. Empat hal diperlukan untuk menambang data secara efektif:

a. Data berkualitas tinggi.
b. Data yang "benar".
c. Ukuran sampel yang memadai.
d. Alat yang tepat.

Berikut ini terdapat beberapa metode data mining (Gambar 2.1) terdiri atas:

a. **Estimasi (Estimation): E**stimasi adalah metode data mining dimana kita dapat memperkirakan nilai populasi dengan memakai nilai sampel. Estimasi biasanya diperlukan untuk mendukung keputusan yang baik, menjadwalkan perkejaan, menjadwalkan kedatangan seseorang, menentukan berapa lama projek dapat dilaksanakan dan lainnya.
b. **Prediksi/ Forcasting** : Prediksi adalah metode data mining yang sangat penting. Prediksi hampir sama dengan klasifikasi dan estimasi, kecuali bahwa dalam prediksi nilai dari hasil akan ada di masa mendatang.
c. **Klasifikasi (classification): Classification** adalah metode yang paling umum pada data mining. Persoalan bisnis sperti Churn Analysis, dan Risk Management biasanya melibatkan metode Classification.men-generalisasi struktur yang diketahui untuk diaplikasikan pada data-data baru. Misalkan, klasifikasi penyakit ke dalam sejumlah jenis, klasifikasi email ke dalam spam atau bukan.
d. **Klasterisasi (clustering): Clustering** juga disebut sebagai segmentation. Metoda ini digunakan untuk mengidentifikasi kelompok alami dari sebuah kasus yang di dasarkan pada sebuah kelompok atribut, mengelompokkan data yang memiliki kemiripan atribut.mengelompokkan data, yang tidak diketahui label kelasnya, ke dalam sejumlah kelompok tertentu sesuai dengan ukuran kemiripannya.

e. **Asosiasi (Association): Assosiation** juga disebut sebagai **Market Basket Analysis**. Sebuah problem bisnis yang khas adalah menganalisa tabel transaksi penjualan dang mengidentifikasi produk-produk yang seringkali dibeli bersamaan oleh customer, misalnya apabila orang membeli sambal, biasanya juga dia membeli kecap.



Gambar 2.1 Metode Data Mining

2.4. Prediction

Prediction/ Forecasting merupakan salah satu metode dalam data mining , (Kotu & Deshpande, 2014) : secara umum prediksi dianggap sebagai tindakan yang menjelaskan mengenai masa yang akan dating, akan tetapi prediksi berbeda dengan menebak secara sederhana yang didasari dengan pertimbangan pengalaman, opini, dan informasi lainnya dalam melakukan peramalan. Prediksi menjelaskan sifat dasar kejadian di masa mendatang terhadap peristiwa-peristiwa tertentu berdasarkan apa yang telah terjadi di masa lalu, istilah yang umumnya dikaitkan dengan ‘prediction’ adalah ‘forecasting’. Meskipun banyak orang yang percaya bahwa kedua istilah itu adalah sinonim, tetapi ada perbedaan tipis namun sangat penting diantara keduanya. ’Prediction’ pada umumnya berbasis opini dan pengalaman, ‘forecasting’ berbasis data dan model(Yusuf & Lawan, 2020). Beberapa contoh dari prediksi dalam bisnis dan penelitian adalah :

a. Prediksi harga beras dalam tiga bulan yang akan datang.

b. Seperti apa jadinya nilai saham dari Microsoft Corporation (pada NASDAQ, disimbolkan sebagai MSFT) pada keesokan hari?
c. Sebanyak apa penjualan produk tertentu pada bulan depan?
d. Prediksi persentase kenaikan kecelakaan lalu lintas tahun depan jika batas bawah. kecepatan dinaikan.

Beberapa metode dan teknik yang digunakan dalam klasifikasi dan estimasi dapat pula digunakan untuk keadaan yang tepat) untuk prediksi.

2.5. K- Means

k - Means yang dianggap sebagai salah satu algoritma pengelompokan yang paling banyak digunakan karena kesederhanaannya. Salah satu algoritma yang sangat sering digunakan dalam statistika dan machine learning adalah algoritma K-Means clustering. K-Means clustering adalah salah satu algoritma analisis klaster(clusteranalysis) nonhirarki. Analisis klaster merupakan salah satu alat untuk mengelompokkan data berdasarkan variabel atau feature. Tujuan dari K-Means clustering, seperti metode klaster lainnya, adalah untuk mendapatkan kelompok data dengan memaksimalkan kesamaan karakteristik dalam klaster dan memaksimalkan perbedaan antar klaster (Gambar 2.2). Algoritma K-Means clustering mengelompokkan data berdasarkan jarak antara dataterhadap titik centroid klaster yang didapatkan melalui proses berulang. Analisis perlu menentukan jumlah K sebagai input algoritma. Dalam ranah machine learning, algoritma K-Means clustering termasuk ke dalam jenis unsupervisedlearning. Contoh penggunaan K-Means clustering antara lain :

- Segmentasi pasar (market segmentation)
- Segmentasi Kinerja
- Segmentasi citra
- Kompresi gambar

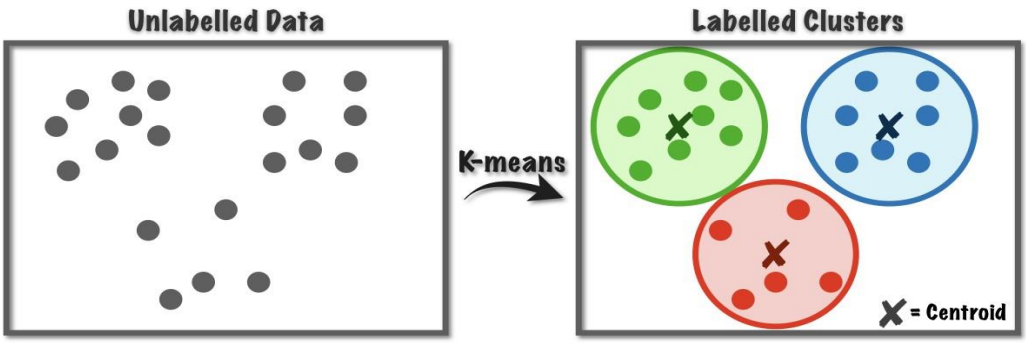
Seperti metode clustering lainnya, K-Means digunakan dalam penelitian eksploratori, confirmatori dan eksplanatori. Untuk tujuan-tujuan eksploratori, K-Means dapat dimanfaatkan untuk melengkapi proses Exploratory Data Analysis (EDA), selain menggunakan analisis statistik deskriptif dan visualisasi data. Sedangkan dalam proses confirmatori dan eksplanatori, K-Means clustering dapat

digunakan untuk melakukan konfirmasi terhadap teori-teori yang sudah ada. Selain itu, algoritma ini juga digunakan untuk melakukan identifikasi jika tiba-tiba terjadi perubahan cluster setelah data baru masuk. Pada tahap terakhir, hasil cluster dapat digunakan sebagai dasar melakukan strategi atau kebijakan terhadap suatu masalah tertentu yang dikaji. Teknik clustering atau segmentasi berbasis prototipe membuat partisi satu tingkat dari objek data. Ada sejumlah teknik seperti itu, tetapi dua di antaranya yang paling menonjol adalah K-Means dan K-medoid. K-Means mendefinisikan prototipe dalam istilah dari centroid, yang biasanya merupakan rata-rata dari sekelompok titik, danbiasanya diterapkan pada objek dalam ruang n-dimensional yang kontinu.

K-Means clustering adalah algoritma deskriptif yang menskalakan dengan baik hingga besar data (Hartigan, 1975). Analisis klaster memiliki aplikasi yang luas, termasuk: segmentasi pelanggan, pengenalan pola, studi biologi, dan web klasifikasi dokumen. K-Means clustering mencoba untuk menemukan k partisi (McQueen, 1967) dalam data, di mana setiap pengamatan milik cluster dengan mean terdekat.

Langkah-langkah dasar untuk K-Means adalah :

1. Pilih k observasi secara sewenang-wenang atau secara acak sebagai centroid cluster awal.
2. Tugaskan setiap pengamatan ke cluster yang memiliki centroid terdekat.
3. Setelah semua pengamatan ditetapkan, hitung ulang posisi dari k centroid.
4. Ulangi langkah 2 dan 3 hingga centroid tidak lagi berubah pengulangan ini membantu meminimalkan variabilitas dalam cluster dan memaksimalkan variabilitas di antara cluster.



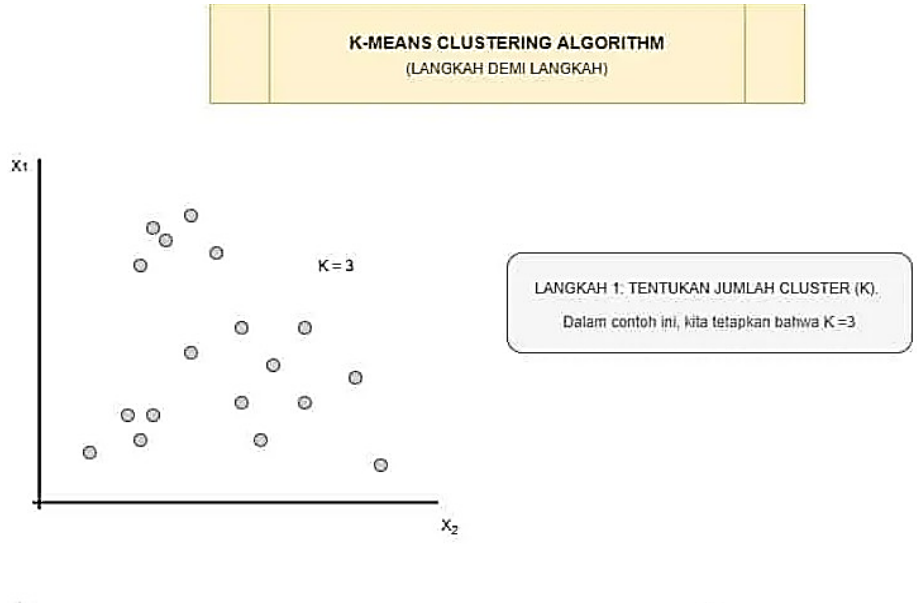
Gambar 2.2 K-Means Clustering

K-Means adalah algoritma yang menemukan pengelompokan ini dalam kumpulan data besar yang tidak memungkinkan untuk dilakukan dengan tangan. Intuisi di balik algoritme sebenarnya cukup lurus ke depan. Untuk memulai, kami memilih nilai untuk k (jumlah cluster) dan secara acak memilih centroid awal (koordinatpusat) untuk setiap cluster. Kami kemudian menerapkan proses dua langkah: Segmentasi clustering adalah salah satu teknik analisis data eksplorasi yang paling umum digunakan untuk mendapatkan intuisi tentang struktur data. Ini dapatdidefinisikan sebagai tugas mengidentifikasi subkelompok dalam data sedemikianrupa sehingga titik data dalam subkelompok (cluster) yang sama sangat miripsedangkan titik data dalam kelompok yang berbeda sangat berbeda. Dengan katalain, kita mencoba untuk menemukan subkelompok yang homogen dalam data sedemikian rupa sehingga titik data di setiap cluster(kelompok)semirip mungkinmenurut ukuran kesamaan seperti jarak berbasis euclidean atau jarak berbasiskorelasi. Keputusan ukuran kesamaan mana yang akan digunakan adalah spesifik aplikasi. Analisis pengelompokan dapat dilakukan berdasarkan fitur di mana kami mencoba menemukan subkelompok sampel berdasarkan fitur atau berdasarkansampel di mana kami mencoba menemukan subkelompok fitur berdasarkan sampel.Kami akan membahas di sini pengelompokan berdasarkan fitur. Clustering digunakan dalam segmentasi pasar; dimana kami mencoba mencari pelanggan yangmirip satu sama lain baik dari segi perilaku atau atribut, segmentasi/kompresi citra;tempat kami mencoba mengelompokkan wilayah yang serupa, pengelompokan dokumen berdasarkan topik, dll. Tidak seperti pembelajaran yang diawasi, pengelompokan dianggap sebagai metode pembelajaran tanpa pengawasan karena kami tidak memiliki kebenaran dasar untuk membandingkan output dari algoritme pengelompokan dengan label yang sebenarnya untuk mengevaluasi kinerjanya.Kami hanya ingin mencoba menyelidiki struktur data dengan mengelompokkantitik data ke dalam subkelompok yang berbeda.

Pada algoritma K-Means proses clustering dilakukan dengan langkah sebagai berikut :

1. **LANGKAH 1: TENTUKAN JUMLAH CLUSTER (K).**

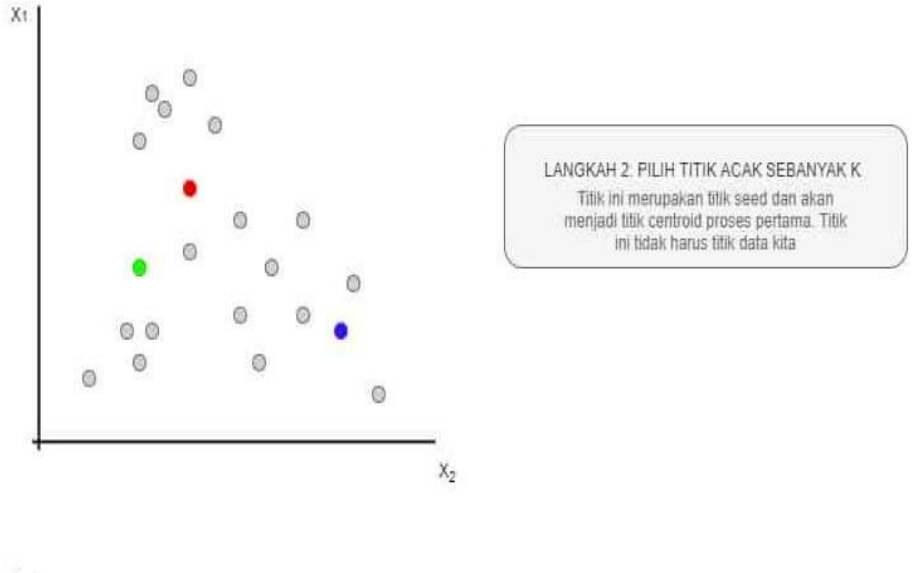
   Langkah awal adalah kita akan menetapkan bahwa akan terdapat 3 kluster maka K = 3 (Gambar 2.3)



Gambar 2.3 Menentukan jumlah kluster

2. **LANGKAH 2: PILIH TITIK ACAK SEBANYAK K.**

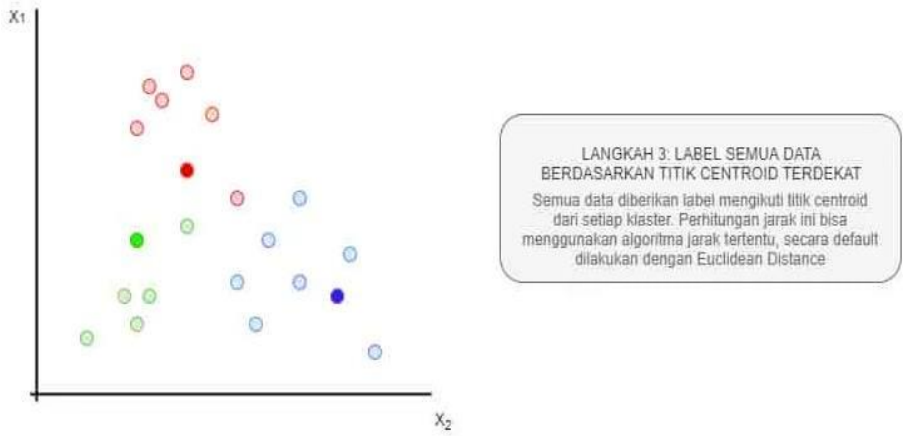
   Memilih titik yang merupakan titik seed dan akan menjadi titik centroid proses pertama (Gambar 2.4). Titik ini tidak harus titik data kita.



Gambar. 2.4 Menentukan titik centroid

3. **LANGKAH 3: LABEL SEMUA DATA BERDASARKAN TITIK CENTROID TERDEKAT.**

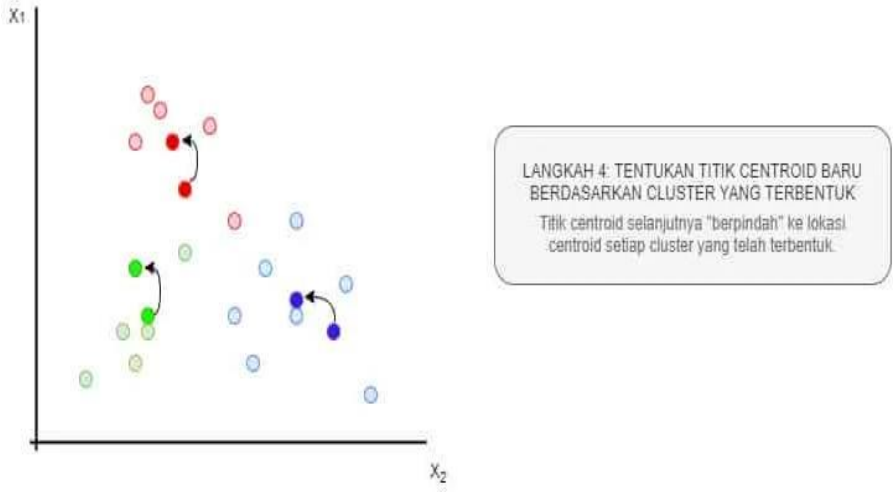
Semua data diberikan label mengikuti titik centroid dari setiap klister (Gambar 2.5). Perhitungan jarak ini bisa menggunakan algoritma jarak tertentu, secara default dilakukan dengan Euclidean Distance



Gambar. 2.5 Memberikan label pada semua titik terdekat centroid

4. **LANGKAH 4: TENTUKAN TITIK CENTROID BARU BERDASARKAN CLUSTER YANG TERBENTUK.**

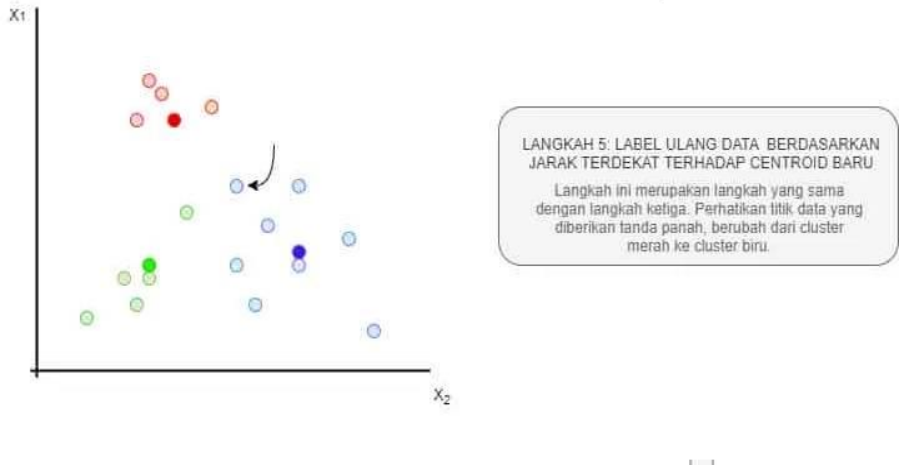
Titik centroid selanjutnya "berpindah" ke lokasi centroid setiap cluster yang telah terbentuk (Gambar 2.6).



Gambar 2.6. Memindahkan centroid ke kluster yang terbentuk

5. **LANGKAH 5: LABEL ULANG DATA BERDASARKAN JARAK TERDEKAT TERHADAP CENTROID BARU.**

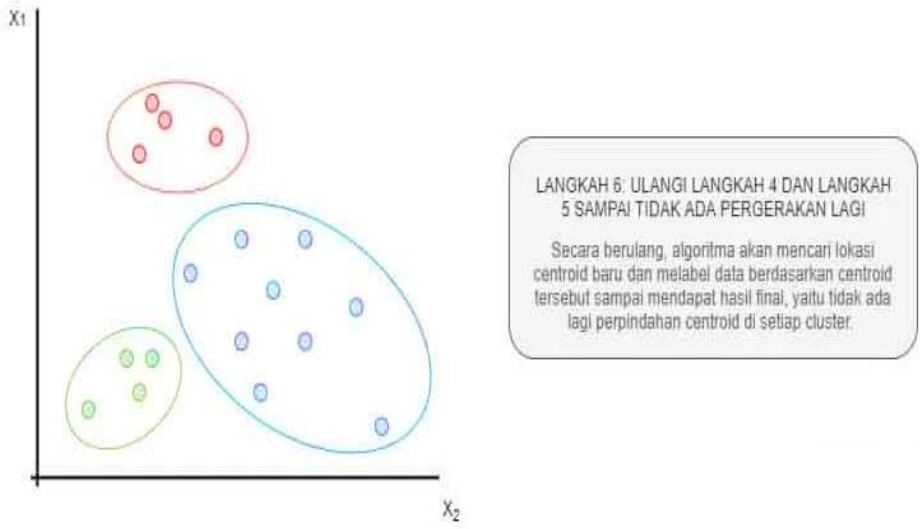
Langkah ini merupakan langkah yang sama dengan langkah ketiga yaitu memberikan label ulang data (Gambar 2.7). Perhatikan titik data yang diberikan tanda panah, berubah dari cluster merah ke cluster biru.



Gambar 2.7. Re-Labeling titik klustering ke centroid baru

6. **LANGKAH 6: ULANGI LANGKAH 4 DAN LANGKAH 5 SAMPAI TIDAK ADA PERGERAKAN LAGI.**

Secara berulang, algoritma akan mencari lokasi centroid baru dan melabel data berdasarkan centroid tersebut sampai mendapat hasil final, yaitu tidak ada lagi perpindahan centroid di setiap cluster(Gambar 2.8).



Gambar. 2.8 Mengulang kembail langkah clustering hingga tercapai kondisi cluster stabil tetap

K-Means adalah algoritma pengelompokan tanpa pengawasan yang dirancang untuk mempartisi data yang tidak berlabel menjadi sejumlah tertentu (itulah "K") dari pengelompokan yang berbeda. Dengan kata lain, K-Means menemukan observasi yang memiliki karakteristik penting dan mengklasifikasikannya bersama ke dalam cluster. Solusi pengelompokan yang baik adalah solusi yang menemukan klaster sedemikian rupa sehingga pengamatan dalam setiap klaster lebih mirip daripada klaster itu sendiri.

2.6. Supervised dan Unsupervised learning

Untuk lebih memahami supervised dan unsupervised. Supervised learning menggunakan data berlabel (labelled data), sedangkan unsupervised learning menggunakan data tanpa label (unlabeled data). Supervised learning digunakan untuk tugas-tugas klasifikasi dan regresi, misal dalam kasus object recognition, predictive analysis dan sentiment analysis. Unsupervised learning digunakan untuk kasus-kasus klastering, asosiasi dan dimensionality reduction.

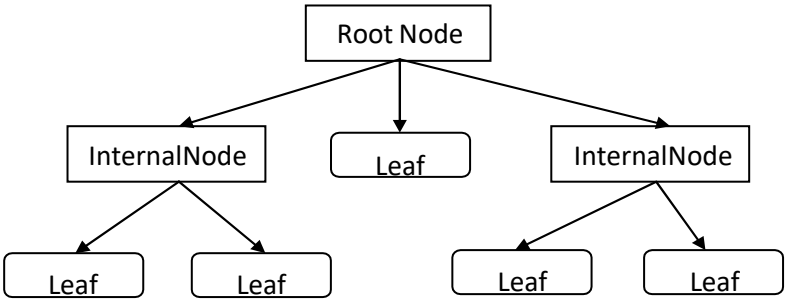
Selanjutnya kita juga perlu memahami jenis data yang dapat dibagi menjadi dua, yaitu data dengan label (labelled data) dan data tanpa label (unlabelled data). Data berlabel (labelled data) adalah data yang memiliki label berupa “tag”, atau kelas yang biasanya dijadikan output model. Sedangkan data yang tidak berlabel (unlabelled data) adalah data yang tidak memiliki label yang digunakan sebagai output pemodelan. Misal kita memiliki dataset berupa kumpulan foto hewan. Jika setiap foto di tag dengan, misalnya nama hewan, maka ini merupakan data berlabel Label ini bukan hanya berupa data kategori tetapi juga data numerik. Lihat ilustrasi (Gambar 2.9) yang kita ambil dari *Grokking Machine Learning* berikut ini.

Gambar 2.9 Illustrasi Grokking Machine Learning

2.7. Decision Tree

Decision Tree adalah struktur flowchart yang menyerupai Tree (pohon), dimana setiap simpul internal menandakan suatu tes pada atribut, setiap cabang merepresentasikan hasil tes, dan simpul daun merepresentasikan kelas atau distribusi kelas. Alur pada Decision Tree di telusuri dari simpul akar ke simpul daun yang memegang prediksi (Kasih, 2019). Selain karena pembangunannya relatif cepat, hasil dari model yang dibangun mudah untuk dipahami. Pada decision tree terdapat 3 jenis node (Gambar 2.10) :

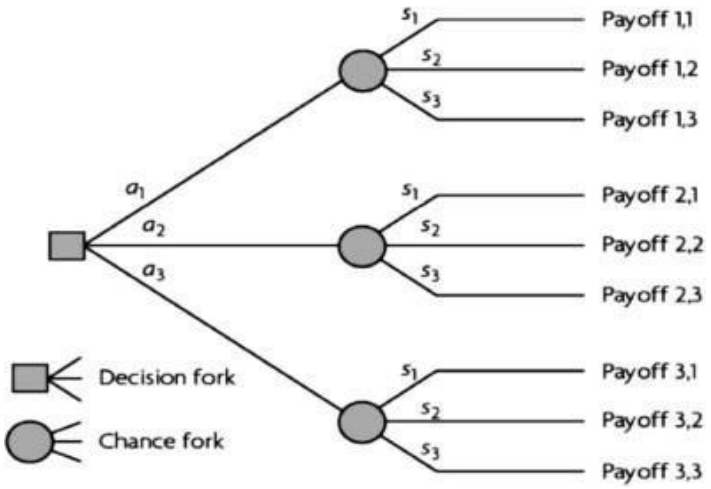
a. Root Node, merupakan node paling atas, pada node ini tidak ada input danbisa tidak mempunyai output atau mempunyai output lebih dari satu.
b. Internal Node, merupakan node percabangan, pada node ini hanya terdapatsatu input dan mempunyai output minimal dua.
c. Leaf node atau terminal node, merupakan node akhir, pada node ini hanyaterdapat satu input dan tidak mempunyai output.



Gambar 2.10. Model Decision Tree

Pohon keputusan merupakan metode klasifikasi dan prediksi yang sangat kuat dan terkenal. Metode pohon keputusan mengubah fakta yang sangat besar menjadi pohon keputusan yang merepresentasikan aturan. Aturan dapat dengan mudah dipahami dengan bahasa alami. Selain itu dapat diekspresikan dalam bentuk bahasa basis data seperti Structure Query Language untuk mencari record pada kategori tertentu (Kusrini dan Emha, 2009). Pohon keputusan juga berguna untuk mengeksplorasi data, menemukan hubungan tersembunyi antara sejumlah calon variabel input dengan variabel target. Sebuah pohon keputusan adalah sebuah

struktur (Gambar 2.11) yang dapat digunakan untuk membagi kumpulan data yang besar menjadi himpunan-himpunan record yang lebih kecil dengan menerapkan serangkaian aturan keputusan, dengan masing-masing rangkaian pembagian, anggota himpunanhasil menjadi mirip satu dengan yang lain.



Gambar 2.11 Bentuk Decision Tree Secara umum

Decision tree memiliki training sample berupa sekumpulan data yang nantinya akan digunakan untuk membangun sebuah tree yang telah diuji kebenarannya. Secara umumDecision Tree adalah untuk membangun pohon keputusan sebagai berikut :

a. Pilih atribut sebagai akar
b. Buat cabang untuk setiap nilai
c. Bagi kasus dalam cabang d. Ulangi proses untuk masing-masing cabang sampai semua kasus pada cabang yang memiliki kelas yang sama. Rumus menghitung nilaientropy menggunakan persamaan :

Entropy (S) $\sum -pi\ log2\ pi\ n\ i$=1 (1)

Keterangan :

S = himpunan kasus
n = jumlah partisi atribut

A Pi = proporsi Si terhadap S
|Si| = jumlah kasus pada partisi ke i
|S| = jumlah kasus dalam S
A= atribut Rumus untuk mencari nilai gain :

Gaint (S,A) $\sum n\ f-1\ \left|si\right|\ \left|s\right|$ Entropy (Si) (2)

2.8. Cross Validation

Cross validation adalah suatu metode tambahan dari teknik data mining yang bertujuan untuk memperoleh hasil akurasi yang maksimal. Metode ini sering juga disebut dengan *k-fold cross validation* dimana percobaan sebanyak *k* kali untuk satu model dengan parameter yang sama (Santosa dan Umam 2018) Dalam bukunya yang berjudul "*Data Mining dan Big Data Analytics.*

Secara umum, kita akan membandingkan *n* model dalam *cross validation ini*, dalam arti lain fungsi dari penggunaan metode *cross validation* adalah

1. Untuk mengetahui performa dari suatu model algoritma dengan melakukan percobaan sebanyak *k* kali.
2. Untuk meningkatkan tingkat performansi dari model tersebut.
3. Untuk mengolah data set dengan kelas yang seimbang.

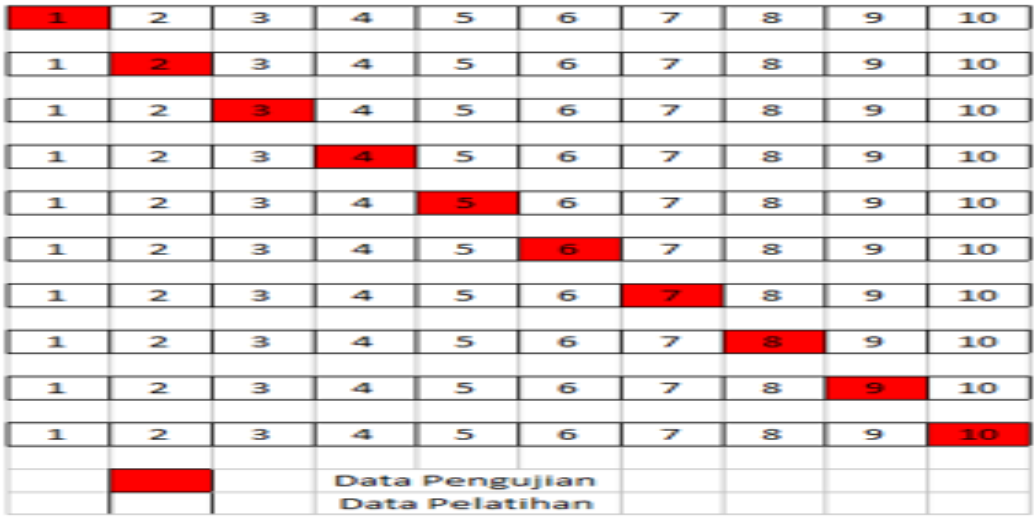
Dalam kasus klasifikasi, ada yang perlu diperhatikan dalam pembagian set data ke sejumlah *k* partisi, yaitu harus melakukan *stratification* yang artinya kita akan mempartisi atau membagi set data tersebut ke *k* partisi dengan komposisi kelas yang seimbang disetiap partisinya. Dengan kata lain, distribusi kelas setiap partisi harus sama antar kelas, yang berarti juga sama dengan distribusi kelas di set data originalnya. Metode Cross-validation (CV) adalah metode statistik yang dapat digunakan untuk mengevaluasi kinerja model atau algoritma dimana data dipisahkan menjadi dua subset yaitu data proses pembelajaran dan data validasi / evaluasi. Biasanya CV K-fold digunakan karena dapat mengurangi waktu komputasi dengan tetap menjaga keakuratan estimasi. Selain penjelasan diatas Cros-validation (CV) biasa disebut sebagai metode hapus-satu, yaitu suatu metode yang bertujuan untuk meminimumkan jumlah kuadrat dari error prediksi untuk variabel respon, dimana prediktor untuk respon tersebut didasarkanpada estimator yang menggunakan seluruh data kecuali data (Putu & Pratiwi, 2017).

Model / algoritma dilatih oleh subset pembelajaran dan divalidasi oleh subset validasi, selanjutnya pemilihan jenis cross-validation dapat didasarkan pada ukuran dataset. Berikut adalah contoh tabel dari cara kerja *cross*

validation dari *5-fold cross validation* yang artinya adalah melakukan percobaan sebanyak 5 kali tahapan. Dari 5 hasil percobaan ini, kita akan catat nilai evaluasi performa dari model tersebut dengan menggunakan *confussionmatrix* kemudian tentukan nilai rata-rata dari setiap percobaan. Maka disitu akan ditemukan percobaan mana yang dapat dijadikan acuan dari penggunaan suatu model algoritma yang terlah dipilih.

Gambar 2.11. Skema 5 Fold Cross – validation

Dalam beberapa penelitian yang sudah dilakukan oleh pakar-pakar data mining, model pengujian atau validasi model dari suatu algoritma klasifikasi, *Cross Validation* lebih sering dipakai ketimbang *Split Validation* karena model validasi dengan menerapkan *10-Cross Validation* sudah merupakan standar dan suatu metode validasi yang canggih atau lebih praktis dan efisien serta mampu meningkatkan sedikit nilai performansinya, oleh karenanya penelitian ini menerapkan 10-Fold Cross-validation dalam perhitungannya. 10-Fold Cross Validation merupakan salah satu dari cross validation yang direkomendasikan untuk pemilihan model terbaik karena cenderung memberikan estimasi akurasi yang lebih baik dalam pengklasifikasian. Dalam 10-Fold Cross Validation, data dibagi menjadi 10 fold yang berukuran sama, sehingga akan memiliki 10 subset data untuk mengevaluasi kinerja model / algoritma (Wong et al.,2019).



Gambar 2.12. Skema 10 Fold Cross – validation

Untuk masing-masing dari 10 subset data tersebut, CV akan menggunakan 9 fold untuk pelatihan dan 1 fold untuk pengujian seperti diilustrasikan pada Gambar 2.12.

Misalkan t_l^k dan y_l^k (k = 1, 2, 3, ..., 10 dan l = 1, 2 ,, 3, ..., Ncv) menjadi nilai target dan nilai prediksi pengamatan data l dalam subset k. Maka, kinerja model atau algoritma dapat diperkirakan dengan mean squared error cross-validation (MSECV) :

$$MSECV = \frac{1}{10}\frac{1}{N_{cv}}\sum_{k=1}^{10}\sum_{l=1}^{N_{cv}}(t_l^k - y_l^k)^2 \qquad (3) \qquad (1).$$

Kita menggunakan CV untuk memilih model yang sesuai dengan membandingkan nilai mean squared error dari cross-validation (MSECV). Model atau algoritma terbaik dipilih jika memiliki nilai MSECV terendah dibanding yang lain. Selanjutnya, misalkan t_m^{val} dan y_m^{val} adalah nilai target dan prediksi nilai untuk data validasi/validasi, dan M adalah jumlah data validasi (m = 1, 2, 3, ..., M). Selanjutnya, kita menggunakan mean square error (MSE)

$$MSE_{val} = \frac{1}{M}\sum_{m=1}^{M}(t_m^{val} - y_m^{val})^2 \qquad (4) \qquad (2).$$

untuk melakukan validasi model terbaik

2.9. Akurasi

Akurasi adalah salah satu metrik untuk mengevaluasi model klasifikasi. Secara informal,akurasi adalah sebagian kecil dari prediksi model kami yang benar. Secaraformal, akurasi memiliki definisi sebagai berikut:

$$\text{Akurasi} = \frac{\text{Number of Corret Prediction}}{\text{Total Number of Prediction}} \quad (5)$$

Untuk klasifikasi biner, akurasi juga dapat dihitung dalam hal positif dan negatif seperti pada persamaan :

$$\text{Akurasi} = \frac{TP+TN}{TP+TN+FP+FN} \quad (6)$$

dimana
TP = True Positif
TN = True Negatif
FP = False Positif
FN = False Negatif

2.10. Kurva ROC dan AUC

Dalam Machine Learning, pengukuran kinerja adalah tugas penting. Jadi dalam masalahklasifikasi, kita dapat mengandalkan Kurva AUC - ROC. Ketika kita perlu memeriksa atau memvisualisasikan kinerja masalah klasifikasi multi-kelas, kita menggunakan kurva AUC (Area Under The Curve) ROC (Receiver Operating Characteristics). Ini adalah salah satu metrik evaluasi terpenting untuk memeriksa kinerja model klasifikasi apa pun. Itu juga ditulissebagai AUROC (Area Di BawahKarakteristik Operasi Penerima) (Sarang Narkhede, 2018).

Seringkali, solusi untuk masalah bisnis atau penelitian tertentu mengarah ke pertanyaan menarik lebih lanjut, yang kemudian dapat diserang menggunakan proses umum yang sama seperti sebelumnya. Pelajaran dari proyek-proyek masa lalu harus selalu dibawa sebagai masukan ke dalam proyek-proyek baru. Berikut ini adalah garis besar dari setiap fase. Meskipun mungkin, masalah yang dihadapi selama fase evaluasi dapat mengirim analis kembali ke salah satu fase sebelumnyauntuk perbaikan, untuk kesederhanaan kami hanya menunjukkan loop yang paling umum,kembali ke tahap pemodelan.(Daniel, n.d.)

2.11. Knowledge Discovery in Databases **(KDD)**

KDD (Knowledge Discovery in Databases) adalah proses yang melibatkan ekstraksi informasi yang berguna, yang sebelumnya tidak diketahui, dan berpotensi berharga dari kumpulan data yang besar. Proses KDD dalam data mining biasanya melibatkan langkah-langkah berikut:

1. Seleksi: Memilih subset data yang relevan untuk dianalisis.
2. Pra-pemrosesan: Membersihkan dan mengubah data agar siap untuk dianalisis. Hal ini dapat mencakup tugas-tugas seperti normalisasi data, penanganan nilai yang hilang, dan integrasi data.
3. Transformasi: Mengubah data ke dalam format yang sesuai untuk penggalian data, seperti matriks atau grafik.
4. Penambangan Data: Menerapkan teknik dan algoritme penggalian data pada data untuk mengekstrak informasi dan wawasan yang berguna. Hal ini dapat mencakup tugas-tugas seperti pengelompokan, klasifikasi, penambangan aturan asosiasi, dan deteksi anomali.
5. Interpretasi: Menafsirkan hasil dan mengekstrak pengetahuan dari data. Hal ini dapat mencakup tugas-tugas seperti memvisualisasikan hasil, mengevaluasi kualitas pola yang ditemukan, dan mengidentifikasi hubungan dan asosiasi di antara data.
6. Evaluasi: Mengevaluasi hasil untuk memastikan bahwa pengetahuan yang diekstrak berguna, akurat, dan bermakna.
7. Penerapan: Menggunakan pengetahuan yang ditemukan untuk memecahkan masalah bisnis dan membuat keputusan.

Proses KDD adalah proses yang berulang dan membutuhkan beberapa kali pengulangan dari langkah-langkah di atas untuk mengekstrak pengetahuan yang akurat dari data.