BAB II
TINJAUAN PUSTAKA

2.1. Prediksi

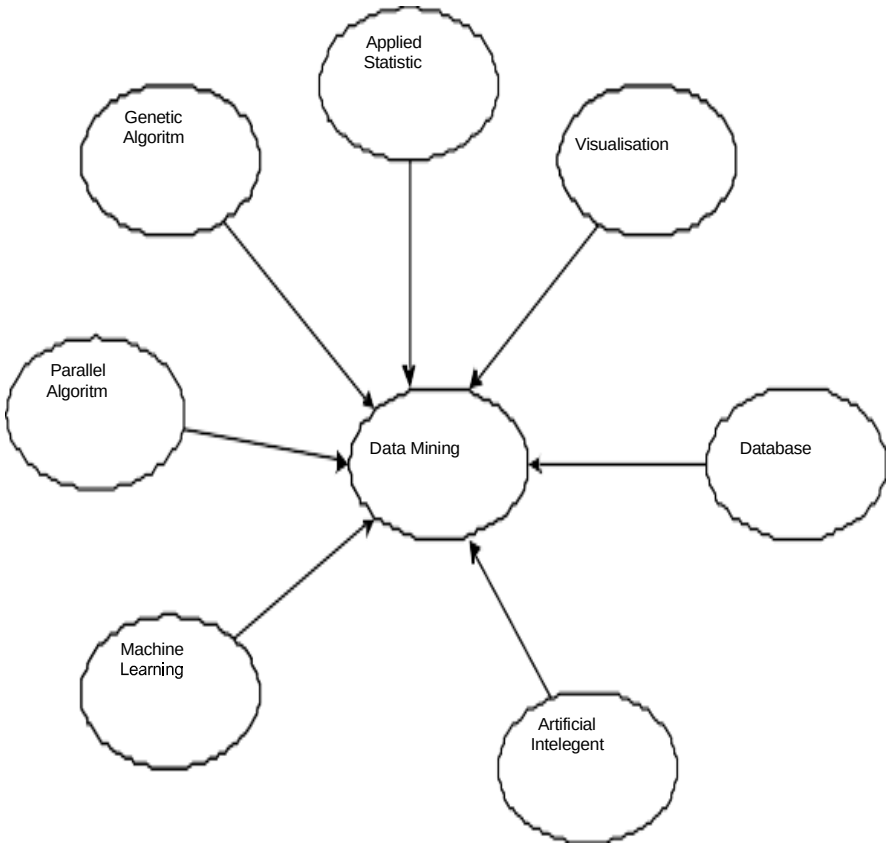
Prediksi merupakan bentuk seni dan ilmu untuk memperkirakan kejadian di masa depan. Hal ini dapat dilakukan dengan melibatkan pengambilan data historis dan memproyeksikannya ke masa mendatang dengan suatu bentuk model matematis. Selain itu, bisa juga merupakan prediksi intuisi yang bersifat subjektif. Atau dapat juga dilakukan dengan menggunakan kombinasi model matematis yang disesuaikan dengan pertimbangan yang baik dari seorang manajer. Prediksi berperan sangat penting dalam bisnis (Mustafa and Simpen, 2018).

2.2. Data Mining

Menurut Pramudiono (2017) perkembangan *data mining* yang pesat tidak dapat lepas dari perkembangan teknologi informasi yang memungkinkan data dalam jumlah yang besar terakumulasi. Tetapi pertumbuhan yang pesat dari akumulasi data telah menciptakan suatu kondisi yang disebut dengan "*rich of data but poor of information"* karena data yang terkumpul itu tidak dapat digunakan dalam suatu aplikasi yang berguna. Bahkan tidak jarang kumpulan data tersebut dibiarkan begitu saja sehingga tercipta "*data tombs"* (kuburan data).

Dalam jurnal ilmiah, *data mining* juga dikenal dengan nama KDD (*Knowledge Discovery in Database*). Namun pada tahun 1995, telah diadakan *International KDD Conference* di Montreal yang berhasil mendefinisikan bahwa KDD merupakan suatu proses dalam mengenali informasi atau suatu kebenaran baru dan benar-benar berguna serta mengenali pola yang dapat dimengerti dari data. Tujuan utama dari proses KDD adalah memprediksikan nilai-nilai yang berguna dari variabel-variabel yang ada atau menemukan pola-pola dari sebuah gugusan data yang dapat diinterpretasikan oleh manusia. Sesuai dengan tujuan tersebut, maka proses dalam mengenali informasi baru dan penemuan pola tersebut perlu diaplikasikan dengan *data mining*. Sehingga sebenarnya *data mining* merupakan suatu bagian yang tidak dapat dilepaskan dari proses KDD (Pramudiono, 2017).

Perlu diketahui bahwa *data mining* merupakan salah satu bidang yang cukup banyak didukung oleh cabang ilmu lain di dalam teknologi informasi yaitu statistik, teknologi basis data, *machine learning*, sistem pakar, algoritma paralel, algoritma genetika, pengenalan pola, visualisasi data, dan lain-lain (Pramudiono, 2017).



Gambar 2. 1 *Data mining* merupakan bidang multidisipliner

Menurut Pramudiono (2017) Ada beberapa faktor yang menjadi alasan utama mengapa menggunakan *data mining*:

1. Banyaknya data yang terkumpul sehingga memerlukan waktu yang sangat lama dan tenaga ahli yang cukup banyak untuk menganalisisnya.
2. Komputer menjadi salah satu pilihan utama karena kemampuannya dalam kecepatan, ketepatan, tidak pernah lelah dan mudah dioperasikan.
3. Tekanan dari kompetisi bisnis yang terus menguat sehingga menjadikan informasi menjadi sangat penting dan harus segera dimiliki.
4. Mampu menemukan suatu pola yang tidak terpikirkan sama sekali.

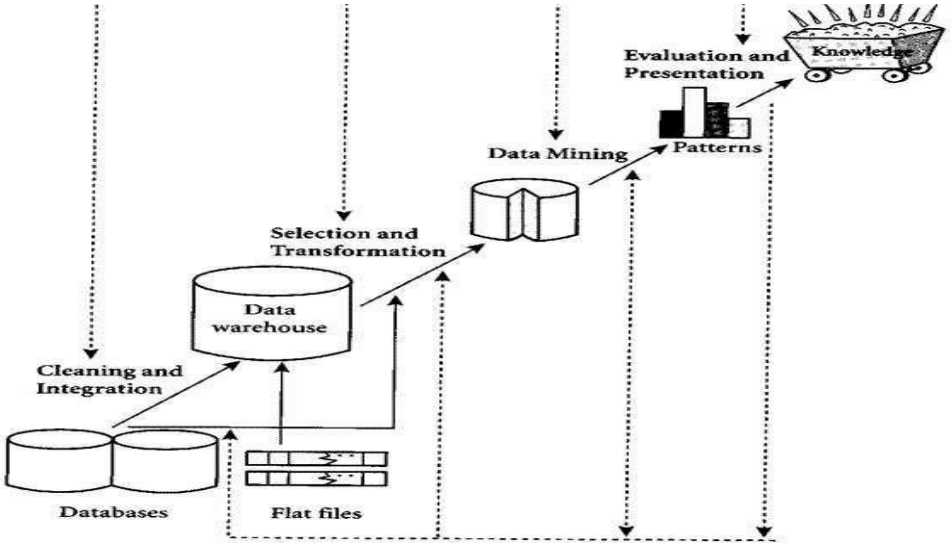
Menurut Sucahyo (2016) *data mining* merupakan salah satu aktifitas dibidang perangkat lunak yang dapat memberikan ROI *(Return of Investment)* yang tinggi. Hal yang perlu diperhatikan adalah bahwa *data mining* berbeda dengan *query tools. Query* dan *data mining* merupakan dua hal yang saling melengkapi. Keberadaan *data mining* bukan untuk menggantikan *query* tetapi menambahkan beberapa tambahan yang berarti. Jika menggunakan *query* sederhana maka informasi yang dapat diakses sekitar 80% dari data yang ada dalam basis data sedangkan 20% lagi akan menjadi informasi tersembunyi yang memerlukan teknik-teknik khusus dalam mengaksesnya .

2.2.1. Tahap-Tahap *Data Mining*

Menurut Sucahyo (2016) karena *data mining* adalah suatu rangkaian proses maka dibagi menjadi beberapa tahap antara lain :

a. Pembersihan data: untuk membuang data yang tidak konsisten dan noise.
b. Integrasi data: untuk menggabungkan data dari beberapa sumber.
c. Transformasi data : untuk mengubah data menjadi bentuk yang sesuai untuk di-*mining.*
d. Aplikasi teknik *data mining.*
e. Evaluasi pola yang ditemukan : untuk menemukan informasi yang menarik ataupun bernilai.
f. Presentasi pengetahuan dengan teknik visualisasi.

Tahap-tahap diatas dapat digambarkan sebagai berikut :



Gambar 2. 2 Tahapan-Tahapan dalam data mining

2.2.2. Teknik *Data Mining*

Menurut Sucahyo (2016) berdasarkan proses, yaitu :

1. *Supervised Learning*

   Dalam *supervised learning* disyaratkan agar data analis telah mengidentifikasi atribut tujuan. Sebagai contoh, bila ada suatu pertanyaan tentang siapakah pelanggan yang baru-baru ini membeli mobil baru, untuk itu dapat dibuat target atribut 1 untuk "YA" dan 0 untuk "TIDAK". Teknik-teknik yang termasuk dalam bagian ini antar lain *Clasification, Regression,* dan lain-lain (Sucahyo, 2016).

2. *Unsupervised Learning*

   Berbeda dengan *supervised learning*, dalam *unsupervised learning* data analis tidak perlu mengidentifikasi atribut target. Teknik-teknik *data mining* yang termasuk ke dalam bagian ini adalah *Clustering, Association Rule,* dan lain-lain (Sucahyo, 2016).

Berikut ini adalah gambaran tentang teknik *data mining* yang paling populer dari teknik-teknik *data mining* yang ada:

1. *Classification*

   Classification adalah proses untuk menemukan model atau fungsi yang menjelaskan atau membedakan konsep atau kelas data, dengan tujuan untuk dapat memperkirakan kelas dari suatu objek yang labelnya tidak diketahui. Model itu sendiri bisa berupa aturan "jika maka", *decision tree* ataupun formula matematis (Pramudiono, 2017). Metode-metode classification yang lain adalah *Bayesian, Neural Network, Genetic Algorithm, Fuzzy, Case-based Reasoning dan K-Nearest Neigboor* .

2.3. *K-Nearest Neighbor* (K-NN)

Algoritma *K-Nearest Neighbor* (K-NN) adalah suatu metode yang menggunakan algoritma supervised. K-NN termasuk kelompok *instance- based learning*.Algoritma ini juga merupakan salah satu teknik lazy learning.K-NN dilakukan dengan mencari kelompok k objek dalam data training yang paling dekat (mirip) dengan objek pada data baru atau data testing (Sumarlin, 2015) .

Secara umum untuk mendefinisikan jarak antara dua objek x dan y, digunakan rumus jarak Euclidean pada persamaan 2.1.

$$dyx = \sqrt{\sum_{f=1}^{n} (x_1 - y_1)^2} \quad \text{.....(2.1)}$$

Keterangan :

Dyx = Jarak *Euclidean*

X_1 = record ke- i

Y_1 = record ke- Y

2.4. *RapidMiner*

RapidMiner merupakan perangakat lunak yang bersifat terbuka (*open source*). RapidMiner adalah sebuah solusi untuk melakukan analisis terhadap *data mining*, *text mining* dan analisis prediksi. Berbagai teknik deskriptif dan prediksi digunakan RapidMiner untuk memberikan kepada kepada pengguna sehingga dapat membuat keputusan yang paling baik. Terdapat kurang lebih 500 operator *data mining* yang dimiliki RapidMiner termasuk operator untuk *input*, *output*, *datapreprocessing* dan *visualisasi*. RapidMiner merupakan *software* yang berdiri sendiri untuk analisis data dan sebagai mesin *data mining* yang dapat diintegrasikan pada produknya sendiri. RapidMiner ditulis dengan munggunakan bahasa java sehingga dapat bekerja di semua sistem operasi (Aprilla Dennis, 2018). RapidMiner memiliki beberapa sifat sebagai berikut:

1. Ditulis dengan bahasa pemrograman Java sehinggadapat dijalankan di berbagai sistem operasi.
2. Konsep multi-layer untuk menjamin tampilan data yang efisien dan menjamin penanganan data.
3. Memiliki GUI, command line mode, dan Java APIyang dapat dipanggil dari program lain.

Beberapa Fitur dari RapidMiner, antara lain:

1. Banyaknya algoritma *data mining*, seperti *decision tree* dan *self-organization map*.

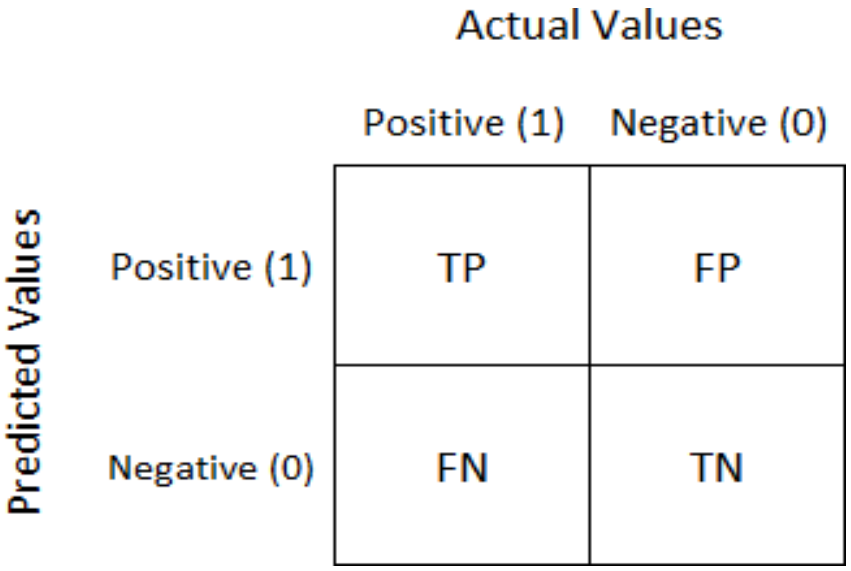
2. Bentuk grafis yang canggih, seperti tumpang tindih diagram histogram, *tree chart* dan 3D *Scatter plots*.
3. Banyaknya variasi *plugin*, seperti *text plugin* untuk melakukan analisis teks.
4. Menyediakan prosedur *data mining* dan *machine learning* termasuk: ETL (*extraction, transformation, loading*), data *preprocessing*, *visualisasi*, *modelling* dan evaluasi.
5. Proses *data mining* tersusun atas operator-operator yang *nestable*, dideskripsikan dengan XML, dan dibuat dengan GUI
6. Mengintegrasikan proyek data mining Weka dan statistika R.

2.5. *Cross Validation*

Validasi dan pengujian adalah Pengujian yang dilakukan untuk mengetahui semua fungsi bekerja dengan baik atau tidak.Validasi dilakukan dengan *Ten-fold CrossValidation*. *Ten-fold Cross Validation* adalah validasi yang dilakukan dengan cara membagi suatu set data menjadi sepuluh segmen yang berukuran sama besar dengan cara melakukan pengacakan data. Validasi dan pengujian dilakukan untuk mengetahui tingkat akurasi, presisi, dan recall dari hasil prediksi klasifikasi.Akurasi adalah persentase dari catatan yang diklasifikasikan dengan benar dalam pengujian dataset.Presisi adalah persentase data yang diklasifikasikan sebagai model baik yang sebenarnya juga baik.*Recall* adalah pengukuran tingkat pengenalan positif sebenarnya (Altujjar *et al.*, 2016).

2.6. *Confusion matrix*

Confusion matrix merupakan salah satu metode yang digunakan untuk menilai akurasi dan mengukur kemampuan suatu metode klasifikasi. *Confusion matrix* menyimpan informasi yang membandingkan hasil klasifikasi yang dilakukan oleh sistem dengan hasil klasifikasi yang sebenarnya. *Confusion matrix* merupakan suatu *tools* penting dalam metode visualisasi yang digunakan di dalam mesin pembelajaran yang biasanya berisi dua kategori atau lebih *Invalid source specified. Confusion matrix* prediksi dua kelas dapat dilihat pada gambar berikut:



Gambar 2.3 Tabel Confusion Matrix

Sumber : (towardsdatascience.com 2020)

Matriks tersebut memiliki empat nilai yang dijadikan acuan dalam perhitungan, dimana:

True Positive (TP) = ketika kelas yang diprediksi positif dan faktanya positif.

True Negative (TN) = ketika kelas yang diprediksi negatif dan faktanya negatif.

False Positive (FP) = ketika kelas yang diprediksi positif dan faktanya negatif.

False Negatif (FN) = ketika kelas yang diprediksi megatif dan faktanya positif.

Berdasarkan nilai TP, TN, FP dan FN dapat diperoleh nilai akurasi. Nilai akurasi menggambarkan seberapa akurat system dapat mengklasifikasi data secara benar. Nilai akurasi menggambarkan seberapa akurat system dapat mengklasifikasi data secara benar. Dari nilai akurasi, presisi dan *recall* diperoleh persamaan sebagai berikut:

Akurasi = $\frac{TP+TN}{Total\ /(TP+TN+FP+FN)}$ ..... (2.2)

Presisi = $\frac{TP}{(FP+TP)}$ ..... (2.3)

Recall = $\frac{TP}{(FN+TP)}$ .....(2.4)

$$F1\text{-}Score = \frac{Presisi\ X\ Recall}{Presisi + Recall} \quad (2.4)$$

2.7. Penelitian Terdahulu

Untuk mendukung penelitian ini diperlukan tinjauan pustaka yang diambil dari beberapa jurnal penelitian yang berkaitan dengan judul penelitian ini dan pokok bahasan berbagai penelitian terdahulu terkait dengan penelitian ini dapat dilihat pada Tabel 2.1 sebagai betikut:

Tabel 2. 1 Penelitian Terdahulu

No	Nama (Tahun)	Judul	Metode	Hasil
1	Suhartini and Bahtiar (2019)	Klasifikasi Algoritma K-Nearest Neighbor Berbasis Particle Swarm Optimization Untuk Kelayakan Bantuan Rehabilitasi Rumah Tidak Layak Huni Pada Desa Lenek Duren Kecamatan Aikmel Kabupaten Lombok Timur Suhartini1,Hariman.	K-Nearest Neighbor	Hasil terbaik yang didapat adalah algoritma K-NN menghasilkan nilai akurasi sebesar 89,29% dan nilai AUC 0,786. Kemudian algoritma K-NN berbasis Particle Swarm Optimization menghasilkan nilai akurasi sebesar 95,33% dan nilai AUC 0.970. Setelah melakukan pengujian terhadap kedua model tersebut memiliki perbedaan tingkat akurasi sebesar 6,04% dan perbedaan nilai AUC 0.184.
2	Yulianti and Nurdin (2018)	Sistem Pendukung Keputusan Penerimaan Bantuan Siswa Miskin (Bsm) Berbasis Online	K-Nearest Neighbor (KNN)	Aplikasi ini dibuat berbasis online dengan menggunakan pemograman

No	Nama (Tahun)	Judul	Metode	Hasil
		Dengan Metode Knn (K-Nearest Neighbor) (Studi kasus : SMPN 1 Koto XI Tarusan) Teknologi informasi adalah suatu teknologi yang digunakan untuk mengolah pribadi		PHP dan MySQL sebagai tempat penyimpanan. Proses dalam menentukan penerimaan BSM dengan menggunakan aplikasi yang dibuat ini dapat membantu panitia beasiswa dalam menginput data dan mendapatkan hasil dari seleksi BSM
3	Wahyuningsih and Utari (2018)	Perbandingan Metode K-Nearest Neighbor, Naïve Bayes dan Decision Tree untuk Prediksi Kelayakan Pemberian Kredit Data akan diuji dengan menggunakan k-folds cross validation (k=10)	K-Nearest Neighbor, Naïve Bayes dan Decision Tree	hasil perbandingan tersebut didapat hasil akurasi metode Decision Tree (J-48) yang lebih unggul dibandingkan dengan metode K-NN dan Naïve Bayes. Hasil yang didapat dari perbandingan ketiga algoritma tersebut adalah, algoritma Decision Tree (J-48) dengan akurasi sebesar 92,21%, algoritma K-Nearest Neighbor memiliki tingkat akurasi sebesar 81,82% dan algoritma Naïve Bayes memiliki tingkat akurasi

No	Nama (Tahun)	Judul	Metode	Hasil
				sebesar 81,83%.
4	Kaesmetan (2016)	Penentuan Penerima Beras Raskin Di Kelurahan Oesapa Barat Menggunakan Metode K-Nearest Neighbor (KNN)	K-Nearest Neighbor (KNN)	Hasil penelitian yaitu penerapan metode K-Nearest Neighbor dalam pengklasifikasian status gizi dengan menggunakan formulasi perhitungan jarak eucledian memiliki kinerja yang baik.
5	Supriana and Astuti (2019)	Implementasi K-Nearest Neighbor Pada Penentuan Keluarga Miskin Bagi Dinas Sosial Kabupaten Tabanan	K-Nearest Neighbor	Sistem yang dibangun akan mengidentifikasi sebuah keluarga berdasarkan 5 katagori kesejahteraan sehingga akan memberikan kemudahan penilaian untuk petugas pendata program kemiskinan. Model pengembangan sistem menggunakan algoritma K-Nearest Nighbor dalam memodelkan dan mengklasifikasi rumah tangga. Hasil penelitian menunjukkan sistem memiliki tingkat akurasi penilaian sebesar 83%.

No	Nama (Tahun)	Judul	Metode	Hasil
6	Arkhiansyah and Rasikun (2018)	Aplikasi Perhitungan Key Performance Indicators (Kpi) Jurusan Berbasis Website Pada Institut Informatika Dan Bisnis	Key Performance Indicators	Menghasilkan Aplikasi perhitungan key performance indicators jurusan berbasis website pada Institut Informatika dan Bisnis Informatics Darmajaya Bandar Lampung. Penggunaan aplikasi ini dapat memberikan kemudahan dalam perhitungan key performance indicators (KPI) jurusan khususnya kepada ketua jurusan sebagai user.