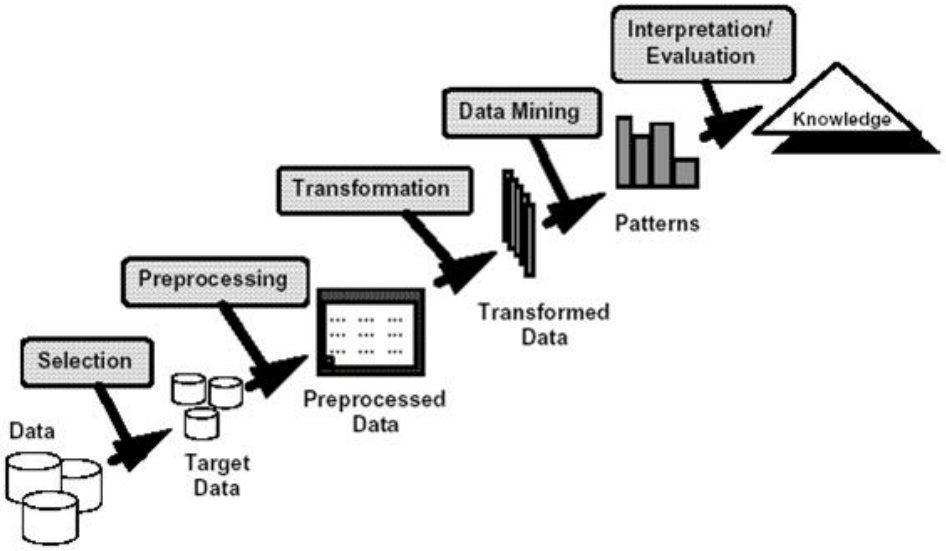
BAB 2
LANDASAN TEORI

2.1 Data Mining

Data mining merupakan proses ekstraksi dari sebuah informasi dan pengetahuan dari data yang sudah ada untuk menemukan pola dan hubungan yang tidak diketahui sebelumnya dalam data. Tujuan *data mining* untuk memperoleh informasi yang berguna untuk membuat keputusan dan memahami fenomena yang terjadi. *Data mining* memiliki banyak istilah lain yang memiliki arti serupa seperti *knowledge mining from data, knowledge extraction, data/pattern analysis, data archaeology, data dredging*, dan *Knowlegde Discovery in Databases*[11]. Proses *data mining* dapat digambarkan seperti berikut :



Gambar 2.1.1 *Data Mining*

Berikut penjelasan dari tahapan-tahapan *data mining* :

1. *Data Selection,* merupakan proses pemilihan data yang akan digunakan untuk penemuan pengetahuan dari *dataset* yang sudah ada. Pemilihan ini didasarkan pada kriteria yang relevan dengan tujuan penelitian. Proses ini sangat penting karena *data mining* belajar dan menemukan hal dari data yang tersedia. Oleh karena itu, proses ini harus dilakukan dengan baik dengan mempertimbangkan sebanyak mungkin atribut yang akan digunakan pada tahap selanjutnya.

2. *Preprocessing,* pada tahap ini terdapat beberapa hal yang harus dilakukan seperti *data cleaning* untuk membersihkan dan menghilangkan tidak konsistennya data. *Data integration* proses penggabungan data dari beberapa sumber menjadi satu.

3. *Data transformation,* proses perubahan data dan dikonsolidasikan ke dalam bentuk yang sesuai untuk *data mining* dengan melakukan operasi ringkasan atau agregasi. Tahap ini juga termasuk tahap untuk melakukan pemilihan dan ekstraksi fitur. Tahapan ini memiliki tujuan untuk memastikan bahwa data dapat dianalisis dengan benar dan efisien. *Data transormation* juga membantu untuk memastikan bahwa data yang digunakan untuk analisis memenuhi kualitas data seperti integritas, akurasi, dan konsistensi.

4. *Data mining,* memilih tugas *data mining* yang sesuai. Memutuskan jenis *data mining* yang akan digunakan. Misalnya *prediction* yang menggunakan jenis *supervised data dining, deskriptif data mining* mencakup aspek *unsupervised* dan visualisasi dari *data mining*. Selanjutnya menentukan metode yang akan digunakan seperti klasifikasi, regresi, atau pengelompokan. Ini sebagian besar tergantung pada tujuan *data mining*, dan juga pada langkah-langkah sebelumnya. Ada dua tujuan utama dalam *data mining*: prediksi dan deskripsi. Setelah memilih metode

spesifik yang akan digunakan untuk pola pencarian maka selanjutnya kita perlu menggunakan algoritma beberapa kali sampai hasil yang diperoleh memuaskan.

5. *Interpretation/Evaluation,* pada tahapan ini merupakan proses untuk memahami pola yang ditemukan dari data, ini melibatkan hasil analisis *data mining* dari klasifikasi. Tujuan dari interpretasi adalah mengubah data mentah menjadi informasi yang dapat diterapkan untuk membuat keputusan yang lebih baik atau memecahkan masalah. Selanjutnya *evaluation,* tahapan untuk mengevaluasi dan menafsirkan pola yang ditambang sehingga mendapatkan pola yang terbaik.

2.2 Classification

Klasifikasi merupakan salah satu metode pada teknik *data mining* dengan jenis *supervised learning*. *Supervised Learning* adalah metode yang digunakan untuk menemukan hubungan antara atribut *input* dan atribut target. Hubungan yang ditemukan direpresentasikan dalam struktur yang disebut sebagai model. Biasanya model menggambarkan dan menjelaskan fenomena yang tersembunyi dalam *dataset* dan dapat digunakan untuk memprediksi nilai atribut target dengan mengetahui nilai atribut *input*. *Supervised learning* dapat diimplementasikan dalam berbagai domain seperti pemasaran, keuangan dan manufaktur.

Klasifikasi adalah bentuk analisis data yang mengekstraksi model yang menggambarkan kelas data penting. Model seperti itu, disebut pengklasifikasi, memprediksi label kelas kategoris (diskrit, tidak terurut). Misalnya, kita dapat membangun model klasifikasi untuk mengkategorikan aplikasi pinjaman bank apakah aman atau berisiko. Analisis semacam itu dapat membantu kita memahami data secara lebih baik. Banyak metode klasifikasi telah diusulkan oleh para peneliti dalam pembelajaran mesin, pengenalan pola, dan statistik. Klasifikasi memiliki banyak algoritma yang berbeda beda, seperti *Decision Trees, Naive Bayes, Rule-Based, KNN, ANN,* dan masih banyak lagi[18].

Tahapan terakhir pada klasifikasi adalah evaluasi, tujuan dari evaluasi ini untuk menghitung seberapa akurat pengklasifikasian dapat memprediksi suatu masalah. Ada beberapa cara untuk mengevaluasi metode klasifikasi ini seperti *Metrics for Evaluating Classifier Performance, Holdout Method and Random Subsampling, Cross-Validation, Bootstrap, Model Selection Using Statistical Tests of Significance, Comparing Classifiers Based on Cost–Benefit and ROC Curves*, dan lain sebagainya.

2.3 Feature Extraction

Feature extraction adalah tahap dalam pemrosesan data dimana karakteristik yang penting dan representatif dari sebuah data diambil dan disimpan untuk digunakan dalam tahap selanjutnya, seperti pengenalan pola atau klasifikasi. Dalam *feature extraction* gambar, karakteristik yang diambil dapat berupa informasi mengenai warna, tekstur, bentuk, dan lokasi objek dalam gambar. *Feature extraction* berguna untuk menemukan representasi data yang baik sehingga menjadi fitur yang berguna. Ada beberapa jenis *feature extraction seperti Standardization, Normalization, Signal enhancement, Extraction of local features, Linear and non-linear space embedding methods, Non-linear expansions*, dan *Feature discretization*. Tahapan ini memiliki 4 aspek penting mencakup *feature construction, feature subset generation (or search strategy), evaluation criterion definition,* dan *evaluation criterion estimation*.

Salah satu metode untuk melakukan *feature extraction* adalah *Extraction of local features* menggunakan algoritma *CNN*. *Feature extraction* menggunakan *Convolutional Neural Network (CNN)* adalah salah satu metode untuk mengambil fitur penting dari suatu gambar atau data citra. Dalam proses ini, beberapa lapisan *convolutional* dan *pooling* dipakai untuk mengidentifikasi dan mengambil fitur yang penting dari citra. Lapisan terakhir dalam jaringan seringkali terdiri dari lapisan *fully connected* yang digunakan untuk memperoleh prediksi akhir. Fitur yang diekstrak dapat digunakan sebagai *input* untuk *task* lain, seperti klasifikasi atau pengenalan objek[19].

2.4 Decision Trees

Decision tree adalah suatu model prediksi yang berbentuk seperti pohon yang menggambarkan hasil dari keputusan dan kemungkinan-kemungkinan yang mungkin terjadi dari setiap keputusan tersebut[18]. *Decision tree* membagi data menjadi beberapa bagian berdasarkan serangkaian keputusan dan kondisi yang telah ditentukan. Setiap cabang dari pohon mewakili suatu keputusan, sedangkan setiap daun dari pohon mewakili suatu hasil atau kelas prediksi. Dalam *decision tree*, keputusan awal disebut sebagai *root node*, sedangkan cabang-cabang yang terbentuk dari *root node* disebut sebagai *internal node*. Setiap *internal node* mengandung kondisi yang harus dipenuhi oleh data sehingga dapat menuju ke salah satu cabang. Setiap cabang kemudian mewakili nilai dari kondisi tersebut dan mengarah ke *node* lain yang lebih kecil atau ke daun[11]. Terdapat beberapa algoritma terapan pada *decision tree* seperti *ID3 (Iterative Dichotomiser 3), C4.5, CART (Chi-square Automatic Interaction Detection)*, dan lain lain[11].

2.5 Algoritma *Decision Tree* C4.5

Algoritma C4.5 adalah algoritma *decision tree* yang dikembangkan dari algoritma *ID3*. C4.5, seperti halnya *ID3*, menggunakan gain informasi sebagai metrik untuk menentukan kriteria pemilihan atribut terbaik pada setiap tahap pemisahan[18]. Algoritma *ID3* menghitung gain informasi untuk setiap atribut dan memilih atribut dengan gain informasi tertinggi sebagai *root node*. Algoritma ini terus membagi data berdasarkan atribut-atribut lain hingga semua data terbagi ke dalam daun yang mewakili kelas prediksi. Kelebihan algoritma C4.5 dibanding *ID3* sebagai berikut:

1. *ID3* hanya cocok untuk data kategorikal dan memiliki kecenderungan *overfitting*.
2. C4.5 dapat mengatasi data numerik dengan menggunakan konsep *threshold*.
3. C4.5 menggunakan gain rasio sebagai pengukur untuk mengatasi masalah *overfitting* pada *ID3*.

Algoritma ini menggunakan nilai *Entropy* untuk menghitung tingkat kesamaan / homogen suatu data, semakin rendah nilai *Entropy* maka semakin homogen sebuah data. Namun dalam pengambilan keputusan *root nodes* terbaik dibutuhkan nilai gain informasi dari setiap atribut, Gain informasi mengukur seberapa banyak informasi baru yang diperoleh setelah membagi data berdasarkan atribut tersebut. Semakin besar gain informasi, semakin baik atribut tersebut untuk menjadi *node*[18].

Berikut adalah tahapan yang akan dilakukan pada algoritma C4.5:

1. Mempersiapkan *dataset* yang akan digunakan
2. Menghitung nilai *root nodes* berdasarkan atribut yang diulang sebanyak x kali hingga menjadi *leaf nodes* dengan menggunakan metode *Entropy* dan dilanjutkan dengan gain informasi. Kedua perhitungan tersebut diperjelesan seperti berikut:
   a. Nilai *Entropy*

$$Entropy(S) = -\sum pi \log2(pi)\ n\ i=1 \qquad (2.1)$$

   Keterangan:

   E = *Entropy*

   S = Himpunan kasus

   n = jumlah data 12

   pi = Proporsi dari Si terhadap S

   b. Gain informasi

$$Gain(S, A) = Entropy(S) - \sum |Si\ |\ |S|\ Entropy(\ n\ i=1\ Si) \qquad (2.2)$$

   Keterangan:

   S = Himpunan kasus

   A = Atribut

   n = Jumlah data pada atribut

   $|Si\ |$ = Jumlah kasus pada partisi ke i

   |S| = jumlah kasus dalam S

3. Kesimpulan hasil perhitungan Membuat kesimpulan berdasarkan hasil perhitungan *decision tree* yang telah dibuat dengan menjelaskan karakteristik model yang telah dibuat.
4. Evaluasi algoritma C4.5 Proses evaluasi matriks dengan melakukan perhitungan akurasi, dan *f1- score*. Namun untuk menghitung akurasi dan *F1-Score* dibutuhkan *confussion matrix* untuk mempermudah proses evaluasi. Ada 4 kondisi saat membuat *confussion matrix* dengan ketentuan sebagai berikut:
   - *True Positve*: kondisi ini merupakan data yang diprediksi positif atau dapat diklasifikasikan secara benar sesuai dengan labelnya oleh model.
   - *True Negative*: kondisi ini merupakan data negatif yang secara benar diklasifikasikan sebagai negatif oleh model.
   - *False Positive*: kondisi di mana data negatif yang salah diklasifikasikan sebagai positif oleh model.
   - *False Negative*: kondisi di mana data positif yang salah diklasifikasikan sebagai negatif oleh model.

Setelah membuat *confussion matrix* diperlukan perhitungan akurasi dan *f1-score*, berikut adalah rumus matematis terkait perhitungan akurasi dan *f1-score* secara detail:

A. Akurasi

Akurasi merupakan nilai metrik yang dapat mengukur pengujian dalam sebuah sistem klasifikasi yang dapat membuat prediksi secara benar. Akurasi diartikan dengan persentase keakuratan sebuah prediksi. Berikut rumus matematis untuk menghitung nilai akurasi:

$$\text{Akurasi} = (TP + TN) / (\text{Total Sampel}) \quad (2.3)$$

Keterangan:

TP = *True Positive*

TN = *True Negative*

Total Sampel = jumlah data yang diuji

B. F1-Score

Nilai *F1-Score* merupakan nilai metrik yang menggabungkan nilai presisi dan *recall* dari sebuah sistem prediksi. Presisi adalah rasio antara *true positive* dengan total prediksi positif. Sedangkan *recall* adalah rasio antara *true positive* dengan contoh aktual positif. Berikut adalah rumus matematis dari *F1-Score*:

$$\text{Presisi} = TP / (TP + FP) \quad (2.4)$$

$$\textit{Recall} = TP / (TP + FN) \quad (2.5)$$

$$\textit{F1-Score} = 2 * (\text{Presisi} * \textit{Recall}) / (\text{Presisi} + \textit{Recall}) \quad (2.6)$$

Keterangan:

TP = *True Positive*

FP = *False Positive*

TN = *True Negative*

2.6 Kaggle

Kaggle adalah platform *online* yang populer di kalangan data *scientist*, ilmuwan data, dan pengembang perangkat lunak untuk berpartisipasi dalam kompetisi data, berbagi *dataset*, dan menjalankan proyek-proyek data *science*. *Kaggle* menyediakan berbagai sumber daya dan alat yang memfasilitasi eksplorasi data, pengembangan model *machine learning*, dan analisis data.

Berikut adalah beberapa komponen dan fitur utama *Kaggle*:

1. **Kompetisi Data (*Data Competitions*)**: *Kaggle* terkenal karena mengadakan berbagai kompetisi data yang menantang peserta untuk memecahkan masalah-masalah dunia nyata dengan menggunakan analisis data dan machine learning. Kompetisi ini sering kali memiliki hadiah uang tunai dan menarik ribuan peserta dari seluruh dunia.
2. ***Dataset* Publik (*Public Datasets*)**: *Kaggle* menyediakan berbagai *dataset* yang dapat diakses secara bebas. Pengguna dapat mencari dan mengunduh *dataset* ini untuk keperluan penelitian, eksperimen, atau pembelajaran.
3. ***Kernels***: *Kaggle* memiliki fitur yang disebut "*kernels*" yang memungkinkan pengguna untuk mengeksekusi kode *Python* (dan beberapa bahasa lain) dalam lingkungan yang siap pakai secara *online*. Ini memungkinkan pengguna untuk berbagi *notebook* interaktif mereka dengan analisis data, visualisasi, dan model *machine learning* dengan komunitas *Kaggle*.
4. **Forum dan Diskusi**: *Kaggle* memiliki forum aktif yang digunakan untuk berdiskusi tentang topik-topik seputar data *science*, *machine learning*, dan kompetisi. Pengguna dapat berbagi pengetahuan, bertanya pertanyaan, dan mencari solusi bersama.
5. ***Course* (*Kaggle Learn*)**: *Kaggle* menyediakan serangkaian kursus *online* gratis yang disebut "*Kaggle Learn*." Kursus ini membantu pengguna untuk memahami konsep-konsep *data science* dan *machine learning* melalui proyek-proyek praktis.

6. **Portofolio**: Pengguna dapat membuat portofolio proyek-proyek data *science* mereka di *Kaggle*, yang dapat digunakan untuk menunjukkan kemampuan dan prestasi mereka kepada calon majikan atau rekan kerja.
7. **Peringkat**: *Kaggle* memiliki sistem peringkat yang memungkinkan pengguna untuk membandingkan kinerja mereka dalam kompetisi data dan kursus dengan pengguna lainnya.
8. **Pekerjaan dan Rekrutmen**: *Kaggle* juga memiliki *sección* "*Jobs*" di mana perusahaan dapat memposting lowongan pekerjaan yang berkaitan dengan data *science* dan ilmu data. Ini memungkinkan para profesional data *science* mencari peluang karier.

Kaggle telah menjadi sumber daya penting bagi komunitas ilmu data karena menyediakan platform yang interaktif, kompetitif, dan mendidik untuk mengembangkan dan menguji keterampilan analisis data dan machine learning mereka. Platform ini telah membantu memajukan lapangan data science dengan cara yang signifikan dengan mengumpulkan para ahli dan mengaktifkan kolaborasi serta persaingan yang sehat dalam memecahkan masalah kompleks.

2.7 Mendeley Data

Mendeley Data adalah platform *online* yang populer di kalangan peneliti, akademisi, dan ilmuwan untuk menyimpan, berbagi, dan mengelola dataset penelitian. *Mendeley Data* menyediakan berbagai sumber daya dan alat yang memfasilitasi pengelolaan data penelitian, aksesibilitas, dan kolaborasi antar peneliti.

Berikut adalah beberapa komponen dan fitur utama *Mendeley Data*:

1. **Penyimpanan *Dataset* (*Data Storage*):** *Mendeley Data* memungkinkan pengguna untuk menyimpan *dataset* penelitian mereka dengan aman. Pengguna dapat mengunggah berbagai jenis *file data* dan mengelola penyimpanannya di satu tempat yang terpusat.

2. **Akses Publik (*Public Access*):** Pengguna dapat memilih untuk membuat *dataset* mereka tersedia secara publik, memungkinkan peneliti lain di

seluruh dunia untuk mengakses, mengunduh, dan menggunakan data tersebut untuk keperluan penelitian lebih lanjut. Ini mendorong keterbukaan dan reproduktibilitas dalam penelitian ilmiah.

3. **Metadata dan Deskripsi (*Metadata and Description*):** *Mendeley Data* menyediakan alat untuk menambahkan metadata yang rinci dan deskripsi dataset. Ini membantu pengguna lain untuk memahami konteks, tujuan, dan metodologi yang digunakan dalam pengumpulan data, serta cara menggunakannya dengan benar.

4. **DOI dan Kutipan (*DOI and Citation*):** Setiap *dataset* yang dipublikasikan di *Mendeley Data* diberikan *DOI (Digital Object Identifier)* unik, yang memungkinkan dataset tersebut untuk diidentifikasi dan dikutip dengan mudah dalam publikasi akademik. Ini membantu meningkatkan visibilitas dan kredibilitas penelitian.

5. **Integrasi dengan Alat Lain (*Integration with Other Tools*):** *Mendeley Data* dapat diintegrasikan dengan berbagai alat dan *platform* lain yang digunakan dalam penelitian, termasuk *Mendeley Reference Manager* dan perangkat lunak analisis data lainnya. Ini mempermudah pengelolaan data dan referensi penelitian secara keseluruhan.

6. **Kolaborasi dan Berbagi (*Collaboration and Sharing*):** *Mendeley Data* memungkinkan peneliti untuk berkolaborasi dengan rekan sejawat mereka dengan berbagi *dataset* secara aman. Pengguna dapat mengatur tingkat akses dan izin, memastikan bahwa data hanya dapat diakses oleh pihak yang berwenang.

7. **Versi Dataset (*Dataset Versions*):** *Mendeley Data* mendukung versi *dataset*, memungkinkan pengguna untuk melacak perubahan dan pembaruan yang dilakukan pada data mereka dari waktu ke waktu. Ini penting untuk menjaga keakuratan dan integritas data penelitian.

8. **Keamanan dan Kepatuhan (*Security and Compliance*):** *Mendeley Data* memastikan keamanan data dengan menerapkan protokol keamanan yang

ketat dan mematuhi standar kepatuhan internasional. Ini memberikan ketenangan pikiran kepada pengguna bahwa data mereka disimpan dengan aman dan sesuai dengan peraturan yang berlaku.

Mendeley Data telah menjadi alat yang penting bagi komunitas penelitian karena menyediakan *platform* yang komprehensif untuk pengelolaan data penelitian. Dengan memfasilitasi penyimpanan, berbagi, dan kolaborasi data, *Mendeley Data* membantu peneliti untuk meningkatkan efisiensi penelitian mereka, meningkatkan transparansi, dan mempromosikan ilmu pengetahuan terbuka.

2.8 Azure Custom Vision

Azure Custom Vision adalah layanan yang disediakan oleh *Microsoft Azure* yang memungkinkan pengguna untuk melatih model *machine learning* khusus untuk pengenalan gambar atau deteksi objek yang sesuai dengan kebutuhan bisnis atau proyek mereka. Layanan ini dirancang untuk membuat implementasi *machine learning* dalam tugas-tugas pengolahan gambar lebih mudah diakses dan diterapkan oleh pengguna dengan berbagai tingkat keahlian.

Berikut adalah beberapa komponen dan fitur utama dari *Azure Custom Vision*:

1. **Pelatihan Model Kustom**: *Azure Custom Vision* memungkinkan pengguna untuk melatih model *machine learning* yang disesuaikan dengan data gambar mereka sendiri. Anda dapat mengunggah gambar, memberikan label kepada objek dalam gambar, dan melatih model untuk mengenali objek atau karakteristik tertentu dalam gambar tersebut.
2. **Pengenalan Gambar**: Setelah model *Custom Vision* Anda dilatih, Anda dapat menggunakannya untuk melakukan pengenalan gambar. Ini bisa digunakan dalam berbagai aplikasi, seperti pengenalan produk dalam gambar *e-commerce*, deteksi penyimpangan dalam produksi, atau pengenalan wajah dalam sistem keamanan.
3. **Deteksi Objek**: Selain pengenalan gambar, *Azure Custom Vision* juga mendukung deteksi objek. Anda dapat melatih model untuk mengenali dan

menandai lokasi objek dalam gambar, yang berguna dalam berbagai aplikasi seperti analisis video, pengawasan industri, atau sistem otomatisasi.

4. **Integrasi dengan *Azure Services***: *Custom Vision* dapat dengan mudah diintegrasikan dengan layanan *Azure* lainnya, seperti *Azure IoT Hub* atau *Azure Stream Analytics*. Ini memungkinkan implementasi model di berbagai solusi *Azure*.
5. **Ekspor Model**: Setelah pelatihan selesai, Anda dapat mengekspor model untuk digunakan di aplikasi atau perangkat yang sesuai dengan kebutuhan Anda. Model ini dapat diintegrasikan dengan perangkat lunak desktop, perangkat mobile, atau aplikasi web.
6. **Optimasi Performa**: Anda dapat mengoptimalkan model Anda untuk kinerja yang lebih baik dengan mengukur dan memeriksa akurasi model, serta melakukan *fine-tuning* jika diperlukan.
7. **Konsol Manajemen**: *Azure* menyediakan konsol manajemen yang mudah digunakan untuk melihat, melacak, dan mengelola model-model yang telah Anda latih.
8. **Keamanan dan Privasi**: *Azure Custom Vision* memperhatikan keamanan dan privasi data Anda. Data pelatihan Anda aman dan dilindungi sesuai dengan standar keamanan *Azure*.