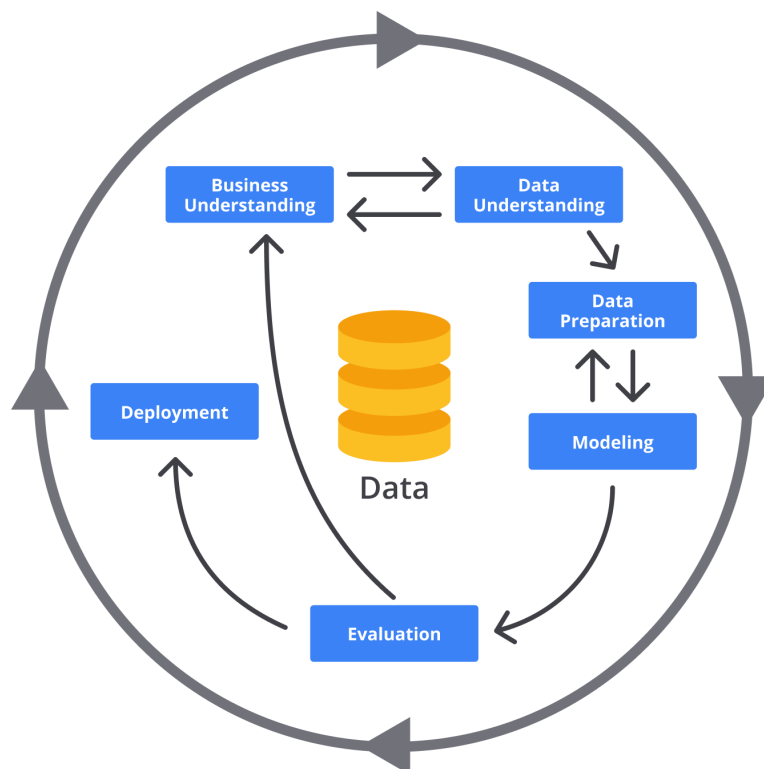


Nama Dosen : Teguh Iman Hermanto, M.Kom
 Mata Kuliah : Machine Learning 2
 Pembahasan : Metode CRISP-DM
 Pokok Pemb : - Business Understanding
 - Data Understanding
 - Data Preparation
 - Modeling
 - Evaluation
 - Deployment



1. Business Understanding

Business Understanding adalah fase pertama dalam metodologi CRISP-DM. Fase ini bertujuan untuk memahami konteks bisnis dari proyek machine learning atau data mining yang akan dikerjakan. Fase ini sangat penting karena menjadi dasar untuk menentukan arah dan tujuan proyek. Tanpa pemahaman yang jelas tentang kebutuhan bisnis, proyek bisa kehilangan fokus dan tidak memberikan nilai yang diharapkan.

Tujuan Business Understanding:

- 1) Memahami Masalah Bisnis: Mengidentifikasi masalah atau peluang bisnis yang ingin diselesaikan.
- 2) Menentukan Tujuan Proyek: Menetapkan tujuan yang jelas dan terukur untuk proyek machine learning.
- 3) Mengidentifikasi Stakeholder: Mengetahui siapa saja yang terlibat dan berkepentingan dengan hasil proyek.
- 4) Membuat Rencana Proyek: Menyusun rencana kerja yang mencakup sumber daya, timeline, dan metrik keberhasilan.

Aktivitas Utama dalam Business Understanding:

- a. **Project Domain** : memetakan latar belakang dari proyek ML
- b. **Problem Statements** : mendefinisikan masalah yang akan dipecahkan
- c. **Goals** : menentukan target dan tujuan yang ingin dicapai dalam proyek ML
- d. **Solution Statements** : membuat rangkaian solusi untuk memecahkan masalah dengan tujuan yang sudah ditargetkan

2. Data Understanding

Data Understanding adalah fase kedua dalam metodologi CRISP-DM. Fase ini bertujuan untuk mengenal dan memahami data yang akan digunakan dalam proyek machine learning. Tujuannya adalah untuk mengumpulkan data, mengeksplorasi karakteristiknya, dan mengidentifikasi masalah atau keanehan dalam data (seperti missing values, outlier, atau ketidakkonsistenan). Fase ini sangat penting karena kualitas data sangat memengaruhi performa model machine learning.

Tujuan Data Understanding

- 1) Mengumpulkan Data: Mengidentifikasi dan mengumpulkan data yang relevan dari berbagai sumber.
- 2) Menjelajahi Data: Memahami struktur, format, dan karakteristik data.
- 3) Mengevaluasi Kualitas Data: Mengidentifikasi masalah seperti missing values, duplikat, atau outlier.
- 4) Mengidentifikasi Pola atau Insight Awal: Menemukan hubungan atau tren yang mungkin berguna untuk analisis lebih lanjut.

Aktivitas Utama dalam Data Understanding

1. **Pengumpulan Data:**
 - Mengumpulkan data dari database, file CSV, API, atau sumber lainnya.
 - Memastikan data yang dikumpulkan relevan dengan tujuan proyek.
2. **Deskripsi Data:**
 - Menganalisis struktur data (jumlah baris, kolom, tipe data, dll.).
 - Membuat ringkasan statistik (mean, median, standar deviasi, dll.).
3. **Eksplorasi Data:**
 - Membuat visualisasi data (histogram, scatter plot, box plot, dll.).
 - Mengidentifikasi hubungan antara variabel (korelasi, distribusi, dll.).
4. **Evaluasi Kualitas Data:**
 - Mengecek missing values, duplikat, atau outlier.
 - Memverifikasi konsistensi dan akurasi data.

3. Data Preparation

Data Preparation adalah fase ketiga dalam metodologi CRISP-DM. Fase ini bertujuan untuk mempersiapkan data yang telah dikumpulkan dan dipahami agar siap digunakan untuk pemodelan *machine learning*. Fase ini seringkali memakan waktu paling banyak dalam proyek *machine learning* karena data mentah biasanya tidak langsung bisa digunakan. Data perlu dibersihkan, diubah, dan diformat ulang agar sesuai dengan kebutuhan model.

Tujuan Data Preparation:

1. Membersihkan Data: Menangani masalah seperti missing values, outlier, dan data yang tidak konsisten.
2. Mengubah Data: Mentransformasi data agar sesuai dengan kebutuhan model (misalnya, normalisasi, encoding, dll.).
3. Menggabungkan Data: Mengintegrasikan data dari berbagai sumber jika diperlukan.
4. Membuat Dataset Final: Menyiapkan dataset yang siap untuk digunakan dalam pemodelan.

Aktivitas Utama dalam Data Preparation:

1. **Pembersihan Data:**
 - Menangani missing values (menghapus, mengisi, atau menginterpolasi).
 - Menangani outlier (menghapus atau mengubah).
 - Menghapus duplikat data.
2. **Transformasi Data:**
 - Normalisasi atau standardisasi data numerik.
 - Encoding data kategorikal (misalnya, one-hot encoding atau label encoding).
 - Membuat fitur baru (*feature engineering*) berdasarkan domain knowledge.
3. **Penggabungan Data:**
 - Menggabungkan data dari berbagai sumber (misalnya, menggabungkan data pelanggan dengan data transaksi).
 - Melakukan join atau merge pada tabel-tabel yang relevan.
4. **Pembuatan Dataset Final:**
 - Membagi data menjadi training set dan test set.
 - Menyimpan dataset dalam format yang siap digunakan (misalnya, CSV, database, dll.).

4. Modeling

Modeling adalah fase keempat dalam metodologi CRISP-DM. Fase ini bertujuan untuk membangun model *machine learning* yang dapat memprediksi atau mengklasifikasikan data berdasarkan tujuan proyek. Pada fase ini, data yang sudah dipersiapkan (dari fase Data Preparation) digunakan untuk melatih model dan mengevaluasi performanya. Fase ini melibatkan pemilihan algoritma, pelatihan model, dan tuning parameter untuk mencapai performa terbaik.

Tujuan Modeling:

1. Membangun Model: Membuat model *machine learning* yang sesuai dengan masalah yang ingin diselesaikan.
2. Mengevaluasi Performa Model: Mengukur seberapa baik model bekerja menggunakan metrik yang relevan.
3. Meningkatkan Performa Model: Menyesuaikan parameter atau memilih algoritma yang lebih baik untuk meningkatkan akurasi atau performa model.

Aktivitas Utama dalam Modeling:

1. **Pemilihan Algoritma:**
 - Memilih algoritma yang sesuai dengan jenis masalah (klasifikasi, regresi, clustering, dll.).
 - Contoh algoritma: Decision Tree, Random Forest, Logistic Regression, SVM, Neural Networks, dll.

2. **Pelatihan Model:**
 - Membagi data menjadi training set dan test set.
 - Melatih model menggunakan training set.
3. **Evaluasi Model:**
 - Menguji model menggunakan test set.
 - Menggunakan metrik evaluasi seperti akurasi, presisi, recall, F1-score, RMSE, dll.
4. **Tuning Parameter:**
 - Menyesuaikan hyperparameter model untuk meningkatkan performa.
 - Menggunakan teknik seperti Grid Search atau Random Search.
5. **Validasi Model:**
 - Memvalidasi model menggunakan teknik seperti cross-validation untuk memastikan model tidak overfitting.

5. Evaluation

Evaluation adalah fase kelima dalam metodologi CRISP-DM. Fase ini bertujuan untuk mengevaluasi model *machine learning* yang telah dibangun dan memastikan bahwa model tersebut memenuhi tujuan bisnis serta memiliki performa yang baik. Evaluasi tidak hanya fokus pada metrik teknis (seperti akurasi atau presisi), tetapi juga mempertimbangkan dampak bisnis dari model tersebut. Fase ini juga mencakup peninjauan ulang terhadap proses yang telah dilakukan untuk memastikan tidak ada langkah yang terlewat atau kesalahan yang terjadi.

Tujuan Evaluation:

1. Menilai Performa Model: Mengevaluasi seberapa baik model bekerja menggunakan metrik yang relevan.
2. Memvalidasi Kesesuaian dengan Tujuan Bisnis: Memastikan model memberikan nilai bisnis yang diharapkan.
3. Mengidentifikasi Masalah atau Kelemahan: Menemukan area di mana model mungkin kurang baik atau perlu ditingkatkan.
4. Membuat Rekomendasi: Menentukan apakah model siap untuk diimplementasikan atau perlu disempurnakan lebih lanjut.

6. Deployment

Deployment adalah fase terakhir dalam metodologi CRISP-DM. Fase ini bertujuan untuk menerapkan model *machine learning* yang telah dibangun dan dievaluasi ke dalam lingkungan produksi, sehingga model dapat digunakan oleh pengguna akhir atau sistem bisnis. Deployment melibatkan integrasi model dengan infrastruktur yang ada, pengujian lebih lanjut, dan pemantauan performa model setelah diimplementasikan. Fase ini juga mencakup pembuatan dokumentasi dan rencana pemeliharaan untuk memastikan model tetap berfungsi dengan baik seiring waktu.