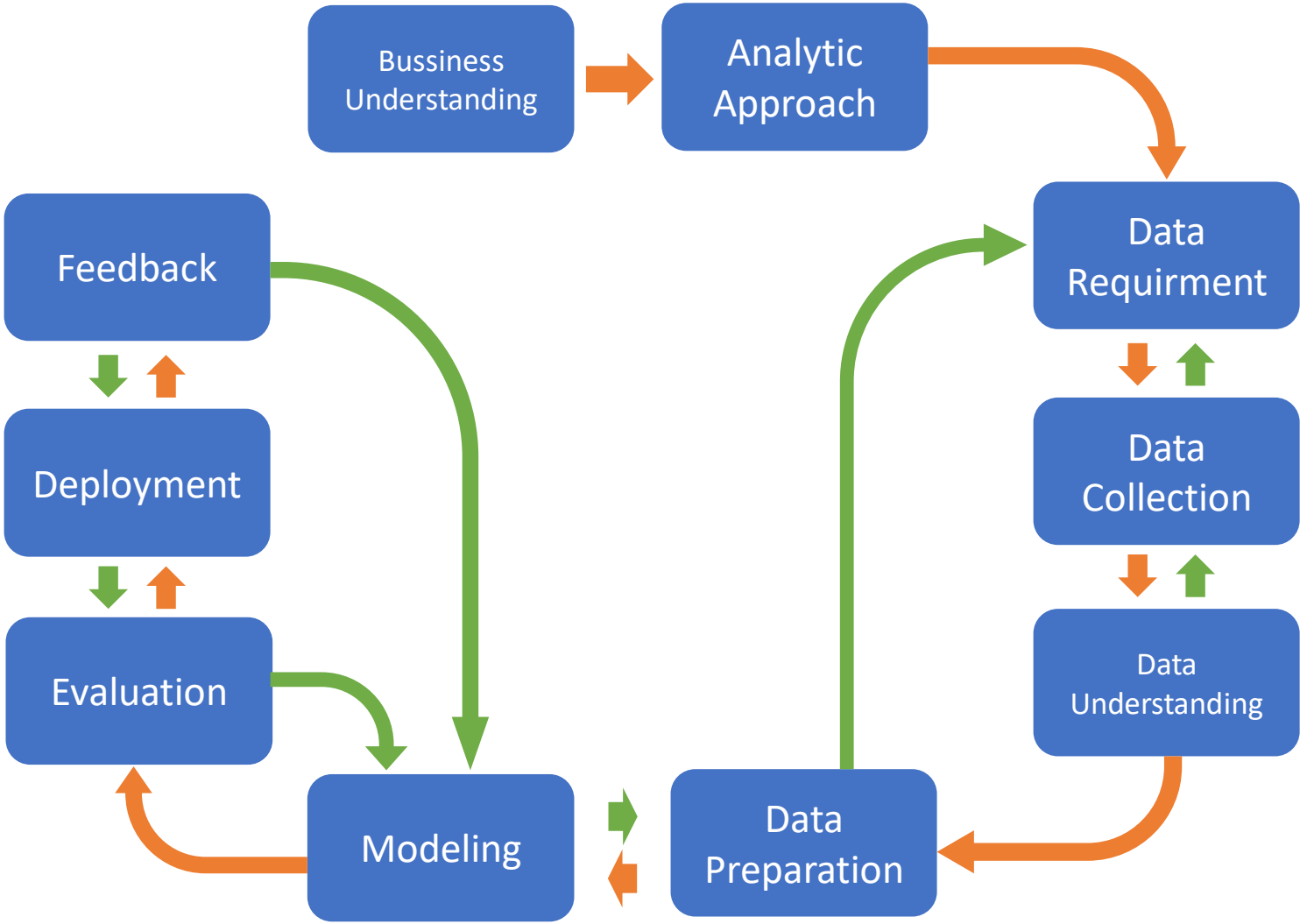
PERTEMUAN 2

Metodologi Data Science

Metodologi data Science

- Metodologi *data science* adalah langkah-langkah digunakan dalam proyek *data science* agar dapat menghasilkan hasil yang optimal yang dapat menjawab pertanyaan dari suatu masalah yang ingin diselesaikan. Metodologi ini tidak bergantung pada teknologi atau *tools* tertentu. Metodologi *data science* yang dibahas disini adalah metode CRISP-DM yang dikemukakan oleh John Rollins yang merupakan seorang senior Data Scientist di IBM

Metodologi Data Science



Tahap Metodologi Data science

- Bussiness Understanding
- Analitic Approach
- Data Requirment
- Data Collection
- Data Understanding
- Data Preparation
- Modeling
- Evaluation
- Deployment
- Feedback

Bussines Understanding

Fase ini mendefinisikan masalah, tujuan, dan persyaratan solusi dari perspektif bisnis.

Misalnya, perusahaan asuransi ingin menggunakan *data science* untuk menyelesaikan masalah, katakanlah pertanyaannya: “Bagaimana cara terbaik untuk mengalokasikan dana kesehatan yang terbatas agar dapat memaksimalkan penggunaannya dalam memberikan layanan yang berkualitas?”

Bussines Understanding (Lanjut..)

Sebelum memulai mengumpulkan data, target dan tujuan dari pertanyaan tersebut perlu didefinisikan terlebih dahulu. Kita memerlukan penjelasan dari si pemberi pertanyaan untuk dalam mengetahui lebih detail target dan tujuannya. Misalnya dalam kasus ini, targetnya adalah menyediakan layanan kesehatan tanpa menaikkan biaya, sedangkan tujuannya adalah meninjau kembali proses yang sudah berjalan untuk mengidentifikasi ketidakefektifan (*inefficiencies*).

Bussines Understanding(Lanjut..)

Setelah target dan tujuan ditentukan, misalnya tim *data scientist* memprioritaskan "perawatan kembali pasien" sebagai area yang efektif untuk ditinjau ulang. Dengan bekal target dan tujuan yang sudah ditentukan, ditemukan bahwa 25-35% pasien yang telah selesai menjalani perawatan akan kembali menjalani perawatan dalam waktu satu tahun, sementara 50% pasien akan kembali menjalani perawatan dalam waktu lima tahun. Dan pasien gagal jantung merupakan pasien terbanyak yang kembali menjalani perawatan.

Bussines Understanding(Lanjut..)

Setelah memahami permasalahan bisnis, selanjutnya adalah mengidentifikasi *business requirement*. Dalam kasus ini *business requirement*-nya disimpulkan sebagai berikut:

- Memprediksi kemungkinan pasien gagal jantung menjalani perawatan kembali.
- Memprediksi kemungkinan pasien (apapun penyakitnya) menjalani perawatan kembali.
- Memahami secara eksplisit hal apa saja yang menyebabkan pasien menjalani perawatan kembali atau tidak.
- Mengaplikasikan prediksi kemungkinan pada pasien baru apakah akan menjalani perawatan kembali atau tidak.

Analitic Approach

Tahap metodologi data science selanjutnya yang dilakukan adalah menentukan pendekatan analitik untuk menyelesaikan masalah. Dalam tahap ini dilakukan pendefinisian masalah dalam konteks statistik atau *machine learning* untuk memperoleh hasil yang diinginkan.

Data Requirment

Tahap Berikutnya metodologi data science adalah data requirements. Pemilihan pendekatan analitik menentukan *data requirements* atau data apa saja yang dibutuhkan agar permasalahan dapat terjawab.

Sama halnya ketika kita ingin memasak jenis makanan tertentu, kita tentu harus menyiapkan bahan-bahan yang dibutuhkan yang sesuai dengan makanan tersebut. Penggunaan bahan-bahan yang tidak sesuai tentunya akan mengakibatkan rasa atau hasil yang kurang memuaskan.

Data Requirment (Lanjutan)

- Dalam tahap ini, asumsikan kita akan **“memasak”** dengan data. Misalnya masalah yang perlu diselesaikan adalah resepnya dan data adalah bahan-bahannya, maka yang perlu kita identifikasi adalah data apa saja yang diperlukan, bagaimana mengumpulkan data tersebut, bagaimana mengolah data tersebut, dan bagaimana menyiapkan data tersebut agar sesuai dengan hasil yang diinginkan.

Data Collection

- Setelah menentukan ***data requirements***, *selanjutya* mengidentifikasi dan mengumpulkan data yang relevan dengan domain masalah. Tahap ini merupakan tahap ke empat di metodologi data science.
- Dalam tahap ini, data yang kita butuhkan tidak tersedia. Maka saat menemukan kendala seperti ini, sangat mungkin bagi kita untuk merevisi kembali *data requirement* dan memutuskan apakah akan mengumpulkan lebih banyak atau lebih sedikit data.

Data Undestanding

- Pada tahap ke lima dari metodologi data science ini, kita akan mengecek apakah ada *missing values*, data yang *imbalanced*, *outlier*, salah format, dan sebagainya yang harus diperbaiki terlebih dahulu.

- Proses *data understanding* yang populer adalah dengan menggunakan statistik deskriptif dan teknik visualisasi. Teknik ini membantu *data scientist* memahami isi data, menilai kualitas data, dan menemukan *insight* awal dari data tersebut.

Data Preparation

- Pada tahapan ini dilakukan pembersihan data, menggabungkan data, dan mengubah data menjadi variabel yang lebih berguna. Agar data dapat diproses secara efektif pada tahap pemodelan, data harus dipersiapkan dengan baik dengan membersihkannya dari *missing values*, *invalid values*, dan data duplikat serta memastikan bahwa seluruh data telah memiliki format yang benar.

Data Preparation(Lanjutan)

Feature engineering

Feature engineering adalah proses transformasi data menjadi fitur-fitur yang lebih representatif dan dalam membantu menyelesaikan masalah dengan lebih baik.

Tahapan ini memakan banyak waktu. Tahap ini bisa menghabiskan sekitar 70% atau bahkan 80% dari keseluruhan proses dalam projek *data science*.

Modeling

Modeling atau pemodelan adalah tahap dalam metodologi *data science* dimana *data scientist* membuat model untuk menjawab permasalahan.

Pemodelan data berfokus pada mengembangkan model, baik itu model deskriptif atau prediktif.

Model ini bergantung pada *analytical approach* yang telah ditentukan sebelumnya, apakah menggunakan pendekatan statistik atau *machine learning*.

Proses pemodelan untuk model prediktif menggunakan data training.

Evaluation

Pengujian terhadap kualitas model apakah model yang dirancang sebelumnya tersebut dapat mengatasi permasalahan bisnis dengan tepat. Evaluasi model memiliki dua fase yaitu:

1. Fase *diagnostic measures*
2. Fase *statistical significance testing*.

Evaluation (Lanjutan..)

Diagnostic measures digunakan untuk memastikan model bekerja dengan baik sesuai yang diharapkan.

Statistical significance testing dapat digunakan untuk memastikan bahwa data yang digunakan telah ditangani dan diinterpretasikan dengan benar di dalam model.

Deployment

Setelah mengasumsikan bahwa model yang dikembangkan menghasilkan hasil yang memuaskan dan disetujui oleh pemangku kepentingan, model tersebut dapat diterapkan atau digunakan dalam lingkungan bisnis.

Feedback

Umpan balik dari stackholder tentang kinerja model yang telah dibangun oleh data scientist, sehingga dapat meningkatkan akurasi dan kegunaan modelnya