D A T A S C I E N C E
and
B U S I N E S S I N T E L L I G E N T

Author: Egi Safitri

Meeting 3

Business Problems and Data Science Solutions

1. Dari Masalah Bisnis ke Tugas Data Mining
2. Metode Supervised vs Unsupervised
3. Data Mining dan Hasilnya
4. Proses Data Mining
5. Implikasi dalam Manajemen Tim Data Science
6. Teknik dan Teknologi Analisis Lainnya

Fundamental Concept

- Prinsip penting dalam Data Science adalah bahwa Data Mining merupakan proses yang terdiri dari beberapa tahap yang telah dipahami dengan baik.
- Beberapa tahap melibatkan penerapan teknologi informasi, seperti penemuan otomatis dan evaluasi pola dalam data, sementara yang lain membutuhkan kreativitas, pemahaman bisnis, dan akal sehat dari seorang analis.

Fundamental Concept

- Karena proses Data Mining membagi tugas pencarian pola dalam data menjadi sejumlah sub-tugas yang terdefinisi dengan baik, proses ini juga berguna dalam menyusun diskusi tentang Data Science.
- Bab ini memperkenalkan proses Data Mining, tetapi sebelum itu, kita akan membahas beberapa jenis tugas Data Mining yang umum.

From Business Problems to Data Mining Tasks

- Setiap masalah bisnis berbasis data memiliki karakteristik unik, termasuk tujuan, keinginan, kendala, dan bahkan faktor manusia.
- Dalam kolaborasi dengan pemangku kepentingan bisnis, **Data Scientist** membagi suatu masalah bisnis menjadi beberapa sub-tugas.
- Solusi dari sub-tugas tersebut dapat digabungkan untuk menyelesaikan keseluruhan masalah bisnis.
- Beberapa sub-tugas ini bersifat unik untuk suatu masalah bisnis tertentu, sementara yang lain merupakan tugas umum dalam **Data Mining**.

From Business Problems to Data Mining Tasks (Contoh)

- Sebagai contoh, masalah **churn pelanggan** dalam industri telekomunikasi di perusahaan **MegaTelCo** bersifat unik.
- Ada aspek khusus dari masalah ini yang berbeda dengan masalah **churn** di perusahaan telekomunikasi lain.
- Namun, salah satu sub-tugas yang kemungkinan besar menjadi bagian dari solusi adalah memperkirakan dari data historis **probabilitas pelanggan akan menghentikan kontraknya setelah masa berlaku habis.**

From Business Problems to Data Mining Tasks

- Meskipun banyak algoritma spesifik telah dikembangkan dalam Data Mining, hanya ada beberapa jenis tugas fundamental yang ditangani oleh algoritma-algoritma tersebut.
- "Individu" dalam konteks ini merujuk pada entitas yang memiliki data, seperti pelanggan atau konsumen, atau bisa juga entitas tak hidup seperti bisnis.

From Business Problems to Data Mining Tasks

- Sering kali, kita ingin menemukan korelasi antara variabel tertentu yang menggambarkan individu dan variabel lainnya.
- Misalnya, dalam data historis, kita mungkin tahu pelanggan mana yang berhenti berlangganan setelah kontraknya habis.
- Kita dapat mencari variabel lain yang berkorelasi dengan pelanggan yang kemungkinan akan meninggalkan layanan di masa mendatang.
- Mencari korelasi semacam ini adalah contoh dasar dari tugas **klasifikasi dan regresi.**

Classification

- **Klasifikasi** dan **estimasi probabilitas kelas** mencoba memprediksi ke kelas mana suatu individu dalam populasi akan termasuk.
- Contoh pertanyaan **klasifikasi**:
  - "Di antara semua pelanggan **MegaTelCo**, siapa yang kemungkinan besar akan merespons penawaran tertentu?"
  - Dua kategori yang bisa digunakan adalah **akan merespons** dan **tidak akan merespons**.

Regression

- **Regresi** (estimasi nilai) berusaha memperkirakan atau memprediksi **nilai numerik** dari suatu variabel untuk individu tertentu.
- Contoh pertanyaan **regresi**:
  - "Berapa banyak layanan yang akan digunakan oleh pelanggan tertentu?"
- **Klasifikasi** memprediksi **apakah sesuatu akan terjadi**, sedangkan **regresi** memprediksi **seberapa banyak sesuatu akan terjadi**.

Similiarity Matching

- **Similarity Matching** bertujuan mengidentifikasi individu yang memiliki kesamaan berdasarkan data yang diketahui tentang mereka.
- Contoh penerapan:
  - **IBM** ingin menemukan perusahaan yang mirip dengan pelanggan bisnis terbaik mereka agar dapat mengoptimalkan strategi pemasaran mereka.
  - Mereka menggunakan **Similarity Matching** berdasarkan data "firmografi" yang menggambarkan karakteristik perusahaan.
- **Similarity Matching** juga menjadi dasar dari metode populer dalam sistem rekomendasi produk (misalnya menemukan orang yang memiliki kesamaan preferensi dengan Anda berdasarkan produk yang mereka beli atau sukai).

Clustering

- **Clustering** mencoba mengelompokkan individu dalam populasi berdasarkan kesamaan mereka, tanpa tujuan tertentu.
- Contoh pertanyaan **clustering**: "*Apakah pelanggan kami membentuk kelompok atau segmen alami?*"
- **Clustering** juga digunakan dalam pengambilan keputusan bisnis, seperti menentukan produk mana yang harus dikembangkan atau bagaimana mengatur tim layanan pelanggan dan penjualan.

Penjelasan Sederhana:

Clustering digunakan untuk menemukan pola dalam data tanpa mengetahui kategori sebelumnya. Misalnya, dalam bisnis, clustering bisa membantu menemukan segmen pelanggan dengan kebiasaan belanja yang mirip.

Co-Occurance Grouping

- Juga dikenal sebagai **frequent itemset mining**, **association rule discovery**, atau **market-basket analysis**.
- Mencari hubungan antara entitas berdasarkan transaksi yang melibatkan mereka.
- Contoh pertanyaan **co-occurrence grouping**: "*Barang apa yang sering dibeli bersamaan?*"
- Contohnya, menganalisis catatan pembelian di supermarket untuk melihat pola pembelian pelanggan.
- Digunakan dalam **sistem rekomendasi**.

Penjelasan Sederhana:

Teknik ini banyak digunakan dalam analisis belanja untuk melihat produk yang sering dibeli bersamaan.

Contoh: pelanggan yang membeli susu juga sering membeli roti, sehingga supermarket bisa menempatkan dua produk ini berdekatan.

Profiling

- **Profiling** mencoba mengidentifikasi pola perilaku khas individu, kelompok, atau populasi.
- Contoh pertanyaan **profiling**: *"Bagaimana pola penggunaan ponsel khas dari segmen pelanggan ini?"*
- **Profiling** sering digunakan untuk mendeteksi anomali dalam sistem keamanan dan deteksi penipuan.
    - Contohnya: mendeteksi penipuan kartu kredit atau pemantauan gangguan dalam sistem komputer.

Penjelasan Sederhana:

Profiling membantu memahami perilaku khas suatu kelompok, misalnya melihat kebiasaan penggunaan internet pelanggan untuk mendeteksi transaksi mencurigakan.

Link Prediction

- **Link prediction** berusaha memprediksi hubungan antara data berdasarkan pola yang ada.
- Contohnya, dalam media sosial:
  - *"Karena Anda dan Karen memiliki 10 teman yang sama, mungkin Anda ingin berteman dengan Karen?"*

Penjelasan Sederhana:

Teknik ini sering digunakan dalam media sosial untuk merekomendasikan teman atau koneksi baru berdasarkan kesamaan jaringan sosial.

Data Reduction

- **Data Reduction** bertujuan untuk mengambil set data yang besar dan menggantinya dengan versi yang lebih kecil tetapi tetap mengandung informasi penting.
- Contoh: dataset besar tentang preferensi menonton film dikurangi menjadi versi lebih kecil yang tetap mencerminkan selera pelanggan.

Penjelasan Sederhana:

Alih-alih menyimpan seluruh data dalam ukuran besar, kita bisa meringkasnya menjadi versi lebih kecil yang tetap berguna. Contoh: menyaring data ribuan transaksi menjadi beberapa pola belanja utama.

Causal Modeling

- **Causal modeling** membantu kita memahami faktor apa yang benar-benar mempengaruhi kejadian lain.
- Contoh:
  - Jika kita menargetkan iklan kepada pelanggan dan mereka kemudian membeli lebih banyak, apakah itu karena iklan tersebut, atau karena pelanggan memang sudah berniat membeli sejak awal?

Penjelasan Sederhana:

Teknik ini digunakan untuk memastikan apakah suatu tindakan benar-benar menyebabkan efek tertentu. Misalnya, jika pelanggan membeli setelah melihat iklan, apakah benar iklan tersebut yang mendorong mereka untuk membeli?

Supervised vs Unsupervised Methods

- **Unsupervised Learning**: Tidak ada target atau tujuan yang spesifik.
    - Contoh: "*Apakah pelanggan kami secara alami terbagi ke dalam kelompok yang berbeda?*"
- **Supervised Learning**: Ada target yang jelas.
    - Contoh: "*Dapatkah kita menemukan kelompok pelanggan yang kemungkinan besar akan membatalkan layanan mereka setelah kontraknya habis?*"

Penjelasan Sederhana:

- **Unsupervised Learning** seperti mencoba mengelompokkan pelanggan berdasarkan kebiasaan mereka tanpa tahu sebelumnya kategori mereka.
- **Supervised Learning** seperti mencoba memprediksi apakah pelanggan akan membeli suatu produk berdasarkan data sebelumnya.

Supervised vs Unsupervised Methods

- Jika target bisa diberikan, masalah bisa dikategorikan sebagai **supervised learning**.
- **Supervised learning** membutuhkan teknik yang berbeda dari **unsupervised learning**, dan hasilnya sering kali lebih berguna.

Penjelasan Sederhana:

Dalam **supervised learning**, kita memiliki data dengan label (misalnya, daftar pelanggan yang membeli produk). Dalam **unsupervised learning**, kita hanya punya data mentah dan mencoba menemukan pola tanpa mengetahui hasil akhirnya.

Supervised vs Unsupervised Methods (Contoh)

- **Supervised Learning**: Harus ada data target.
- Contoh:
  - **Customer ID | Loyal**
  - 1345 | Yes
  - 1234 | No
- Mengumpulkan data tentang target adalah investasi penting dalam **Data Science.**

Penjelasan Sederhana:

Misalnya, kita ingin memprediksi apakah pelanggan akan tetap setia. Kita bisa menggunakan data historis tentang pelanggan yang tetap loyal atau tidak, lalu melatih model berdasarkan data tersebut.

Supervised vs Unsupervised Methods (Label)

- Dalam **supervised learning**, target data harus ada.
- Nilai target untuk individu sering disebut sebagai **label**.
- Contoh: **Customer 1345** memiliki label **"YES"**, yang berarti dia adalah pelanggan setia.

Penjelasan Sederhana:

Label adalah kategori atau nilai yang kita coba prediksi. Jika kita memiliki data pelanggan lama yang setia atau tidak setia, kita bisa melatih model untuk memprediksi apakah pelanggan baru juga akan setia atau tidak.

Supervised vs Unsupervised Methods

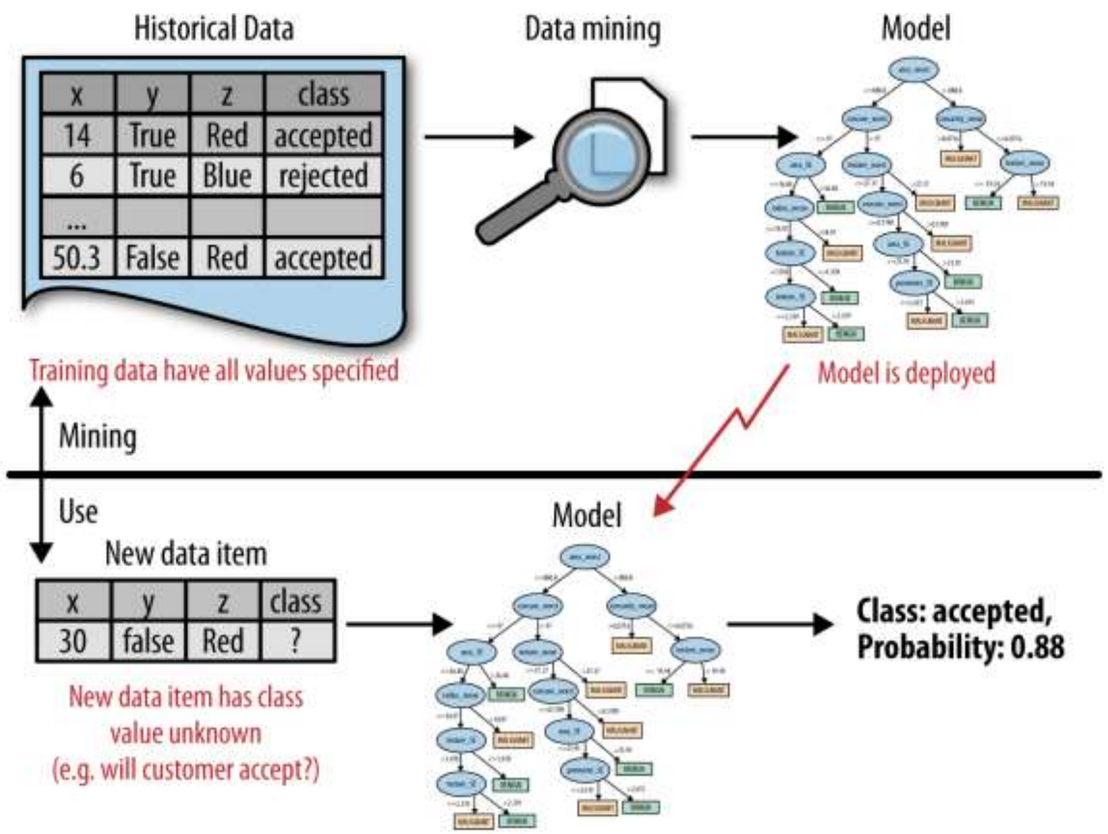
- Dua subkategori utama dalam data mining **supervised** (terawasi):

    1. **Klasifikasi:** Memprediksi apakah pelanggan akan membeli suatu layanan jika diberi insentif.
    2. **Regresi:** Memprediksi seberapa banyak pelanggan akan menggunakan layanan tertentu.

Data Mining and Its Results

- Pada tahap awal data mining, penting untuk menentukan apakah pendekatan akan bersifat **supervised** (terawasi) atau **unsupervised** (tidak terawasi).
- Jika **supervised**, maka perlu mendefinisikan **target yang jelas** yang akan menjadi fokus analisis.
- **Data mining** menggunakan data historis untuk membangun model prediksi.
- Model yang dibuat kemudian diterapkan ke **data baru** untuk memperkirakan **hasil dan probabilitas** tertentu.

Data Mining and Its Results

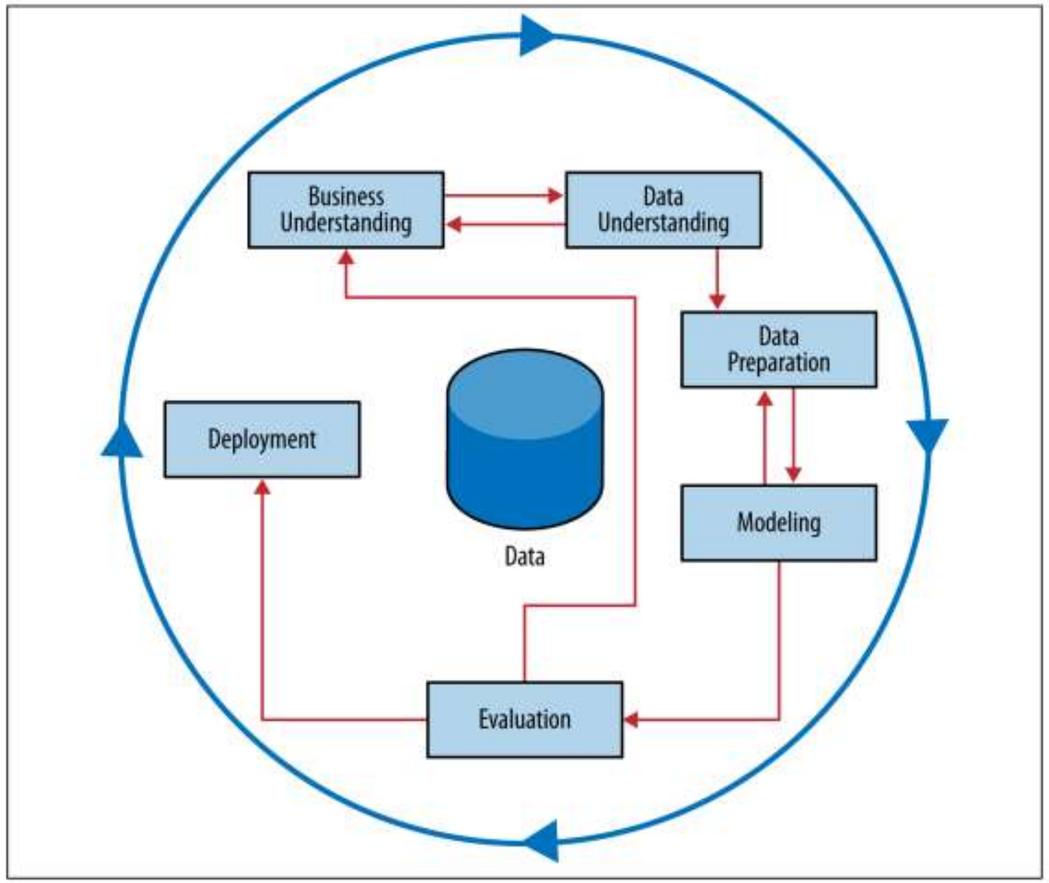


Gambar 1. Perbandingan antara data mining dan penggunaan hasil data mining.

1. **Bagian atas gambar:**
   - Data historis digunakan untuk membangun model prediksi.
   - Data ini sudah memiliki informasi lengkap, termasuk variabel target (class).
2. **Bagian bawah gambar:**
   - Model yang telah dibuat digunakan untuk **menganalisis data baru**.
   - Data baru **tidak memiliki class**, sehingga model akan mencoba menebaknya.
   - Model tidak hanya memberikan hasil prediksi (accepted/tidak), tetapi juga memberikan **probabilitas** bahwa prediksi tersebut benar.

Contoh dalam dunia nyata:
Misalkan kita memiliki data pelanggan sebelumnya yang menunjukkan apakah mereka menerima atau menolak tawaran produk. Model yang dibuat dari data ini bisa digunakan untuk memprediksi apakah pelanggan baru akan menerima tawaran yang sama.Jika model memperkirakan seorang pelanggan memiliki probabilitas 88% untuk menerima tawaran, perusahaan bisa fokus menawarkan produk tersebut kepadanya.

The Data Mining Process



Data mining mengikuti standar CRISP-DM (Cross Industry Standard Process for Data Mining). Proses ini bersifat iteratif, artinya setiap tahapan eksplorasi data akan memperbaiki pemahaman dan akurasi model.

Business Understanding

- **Memahami masalah bisnis** yang ingin diselesaikan sangat penting sebelum menerapkan data mining.
- Tahap ini melibatkan **kreativitas analis** dan perancangan **skenario penggunaan data** agar hasil analisis lebih relevan.

Data Understanding

- **Data adalah bahan baku utama** dalam data mining.
- Sebelum digunakan, kita harus memperhitungkan **biaya dan manfaat** dari setiap sumber data.
- Contoh penerapan:
  - **Deteksi penipuan kartu kredit**
  - **Deteksi penipuan dalam klaim asuransi kesehatan**

Data Preparation

- Proses ini mencakup beberapa langkah:
    1. Mengubah data ke dalam **format tabel.**
    2. Menghapus atau mengisi **nilai yang hilang.**
    3. Mengonversi **tipe data** sesuai kebutuhan analisis.
- **Menghindari kebocoran data (data leaks)**, yaitu situasi di mana data historis mengandung informasi yang tidak tersedia saat pengambilan keputusan.

Modelling

- **Pemodelan (Modeling)** bertujuan untuk **menangkap pola dalam data.**
- Model yang baik mampu **mengenali pola** dan membuat **prediksi yang akurat** berdasarkan data yang ada.

Evaluation

- Mengevaluasi hasil **data mining** secara ketat untuk memastikan validitas dan keandalan sebelum melanjutkan ke tahap berikutnya.
- Evaluasi mencakup **penilaian kuantitatif dan kualitatif**. Pemangku kepentingan harus memastikan bahwa model membawa manfaat lebih banyak daripada risiko.
- Tim **data science** harus mempertimbangkan apakah model tersebut mudah dipahami oleh pemangku kepentingan.

Penjelasan:

Evaluasi adalah tahap penting untuk memastikan model yang dibuat benar-benar bekerja dengan baik sebelum digunakan dalam bisnis.

Deployment

- Model yang telah dibuat harus diterapkan dalam **dunia nyata** untuk menghasilkan keuntungan dari investasi data mining.
- Contoh paling umum adalah **mengintegrasikan model prediksi** ke dalam sistem informasi atau proses bisnis.
- **Contoh penerapan:** Model prediksi churn dapat digunakan dalam strategi **manajemen pelanggan**, misalnya dengan memberikan **penawaran khusus kepada pelanggan yang diprediksi akan berhenti berlangganan**.

Penjelasan:

Setelah model diuji dan dinilai baik, langkah berikutnya adalah menerapkannya dalam bisnis agar menghasilkan manfaat nyata.

Implikasi dalam Manajemen Tim Data Science

- **Kesalahan umum:** Menganggap data mining sama seperti siklus pengembangan perangkat lunak.
- **Perbedaan utama:** Data mining memerlukan keseimbangan antara **keahlian pemrograman** dan **analisis data.**

Penjelasan:

Data science tidak hanya tentang menulis kode, tetapi juga memahami data dan membuat analisis yang berguna bagi bisnis.

Statistik dalam Analisis Bisnis

- Statistik digunakan untuk menghitung nilai numerik penting dalam data.
- Statistik membantu kita memahami **distribusi data**, **menguji hipotesis**, dan **memperkirakan ketidakpastian hasil.**
- **Pengujian hipotesis** membantu menentukan apakah pola dalam data benar-benar signifikan atau hanya kebetulan.

Penjelasan:

Statistik adalah alat penting dalam data science untuk memahami data dan mengambil keputusan berdasarkan bukti yang valid.

Teknik dan Teknologi Analitik Lainnya

- Menyajikan **enam kelompok teknik analitik** yang berhubungan.
- **Membandingkan dan mengkontraskan** teknik ini dengan **data mining.**
- **Data mining** bertujuan untuk **menemukan pola atau informasi tersembunyi dalam data secara otomatis**.
- **Analis bisnis** harus memahami teknik analitik yang paling sesuai untuk suatu masalah.

Penjelasan:

Data mining hanyalah satu dari banyak teknik analisis data. Analis bisnis harus tahu kapan dan bagaimana menggunakan teknik yang berbeda.

Query Data (Data Querying)

- **Query** adalah permintaan spesifik untuk mengambil sebagian data atau statistik tertentu dari database.
- Berbeda dari **data mining**, karena query tidak mencari pola atau model dalam data.
- **Contoh Query SQL:**

```sql
SELECT * FROM CUSTOMERS WHERE AGE > 45 AND SEX='M' AND DOMICILE = 'NE'
```

Penjelasan:

Query digunakan untuk mengambil data tertentu dari database, sedangkan data mining digunakan untuk menemukan pola dalam data.

Data Warehousing

- **Data warehouse** mengumpulkan data dari berbagai sistem dalam suatu perusahaan.
- Data dari sistem **penjualan, keuangan, dan SDM** bisa digabungkan untuk menemukan pola bisnis yang efektif.
- **Contoh:** Menggunakan data dari penjualan dan HR untuk memahami pola karakteristik tenaga penjualan yang sukses.

Penjelasan:

Data warehouse memungkinkan perusahaan menyimpan dan mengelola data dalam skala besar untuk analisis lebih lanjut.

Regression Analysis

- Digunakan untuk **memprediksi atau memperkirakan nilai variabel** berdasarkan data yang dianalisis.

Penjelasan:

Regresi membantu dalam memprediksi angka, misalnya **perkiraan penjualan di masa depan** berdasarkan data historis.

Machine Learning dan Data Mining

- **Machine Learning** adalah kumpulan metode untuk **mengekstrak model prediktif dari data.**
- Dikembangkan dalam berbagai bidang, termasuk **Statistik Terapan dan Pengenalan Pola.**
- **Machine Learning** adalah bagian dari **Kecerdasan Buatan (AI)** yang bertujuan untuk meningkatkan pengetahuan atau kinerja sistem berdasarkan pengalaman.

Penjelasan:

Machine Learning memungkinkan sistem komputer belajar dari data dan membuat prediksi tanpa diprogram secara eksplisit.

Start making small data science projects with the help of Kaggle


THANK
YOU