DATA SCIENCE

MOH DZAKY IRHAB

DATA SCIENCE

Data Science adalah cara kita menjawab pertanyaan dan membuat keputusan dengan menggunakan data

KOMPONEN UTAMA

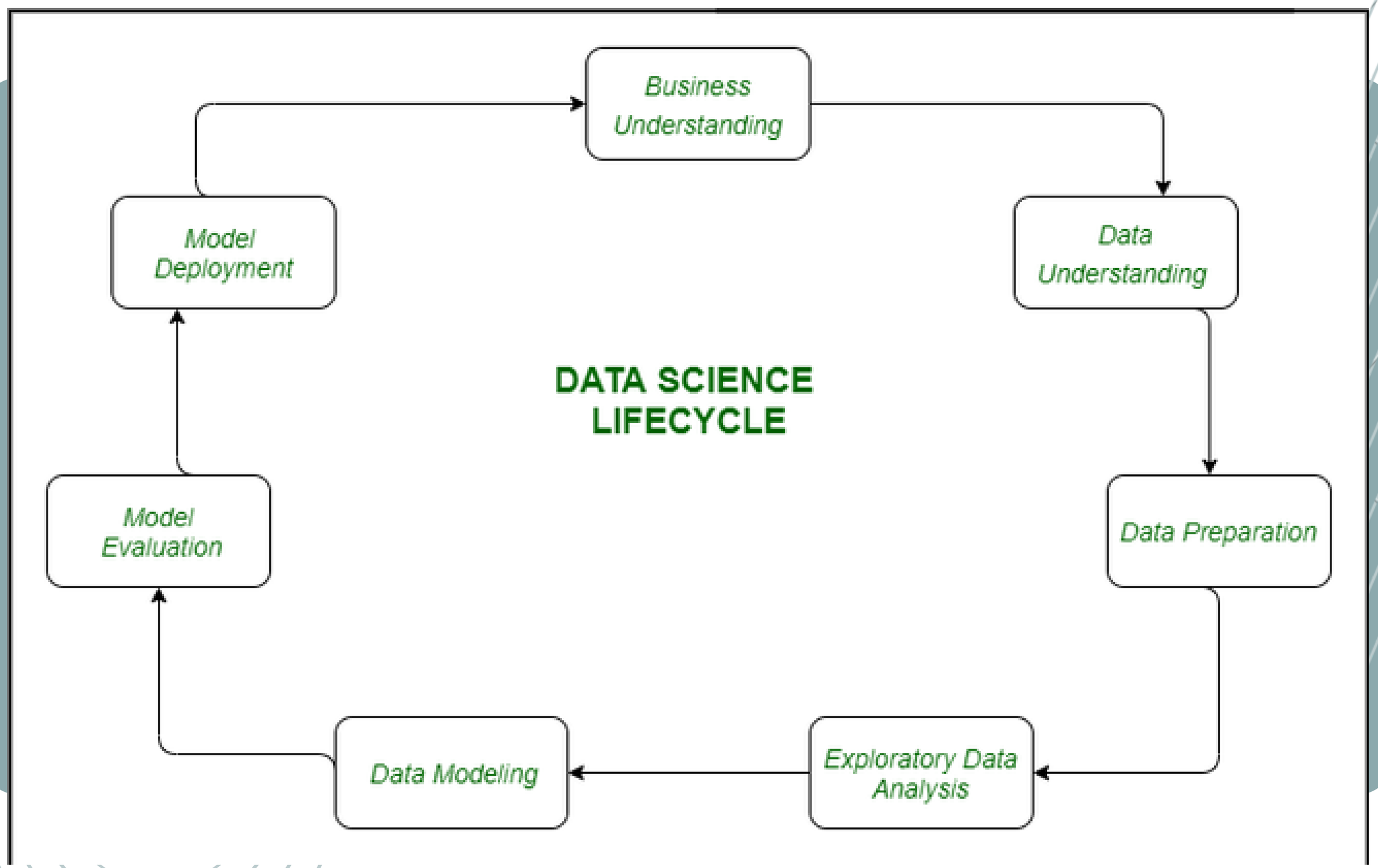
1. Data:
   - Data adalah fakta-fakta atau informasi yang kita kumpulkan, seperti angka, teks, atau gambar.
2. Pertanyaan atau Masalah:
   - Data Science dimulai dengan pertanyaan atau masalah yang ingin dipecahkan.
3. Analisis Data:
   - Proses menggali dan menganalisis data untuk menemukan pola atau jawaban.
4. Model dan Prediksi:
   - Membuat model (seperti mesin prediktif) untuk memprediksi hal-hal berdasarkan data yang telah ada.
5. Pengambilan Keputusan:
   - Menggunakan hasil analisis dan prediksi untuk membuat keputusan yang lebih baik.

MASALAH

Bayangkan kamu punya data tentang berapa lama kamu belajar setiap hari dan nilai ujian kamu. Pertanyaannya adalah, "Apakah ada hubungan antara waktu belajar dan nilai ujian?"

LANGKAH

1. **Pengumpulan Data: Catat waktu belajar dan nilai ujian setiap hari.**
2. **Analisis Data: Periksa data untuk melihat apakah ada pola atau hubungan.**
3. **Model Prediksi: Buat model yang dapat memprediksi nilai ujian berdasarkan waktu belajar.**
4. **Pengambilan Keputusan: Dengan model tersebut, kamu bisa membuat keputusan tentang berapa lama sebaiknya kamu belajar untuk mendapatkan nilai ujian yang baik.**


Business Understanding
Data Understanding
Data Preparation
Exploratory Data Analysis
Data Modeling
Model Evaluation
Model Deployment
DATA SCIENCE LIFECYCLE

DATA MINING

Data mining adalah proses pengumpulan dan pengolahan data yang bertujuan untuk mengekstrak informasi penting pada data.

DATA SCIENCE

Data Science is about data gathering, analysis and decision-making.
Data Science is about finding patterns in data, through analysis, and make future predictions.

DATA UNDERSTANDING

TIPE DATA NUMERIK

Tipe data numerik membantu dalam perhitungan matematis, statistik, dan pemodelan. Mereka menyediakan informasi kuantitatif yang dapat diolah untuk mendapatkan wawasan.

TIPE DATA KATEGORI

Tipe data kategorikal membantu dalam mengelompokkan dan mengidentifikasi pola berdasarkan kategori atau kelompok. Mereka dapat digunakan dalam analisis frekuensi, membuat grafik kategorikal, dan pemahaman struktur kualitatif dari data.

KONSEP DASAR STATISTIK - DATA UNDERSTANDING

1. **Mean (Rata-rata):**
   - **Rata-rata adalah jumlah semua nilai dibagi dengan jumlah total nilai.**
   - **Contoh: Rata-rata dari 2, 4, dan 6 adalah (2 + 4 + 6) / 3 = 4.**
2. **Median:**
   - **Median adalah nilai tengah saat data diurutkan dari terkecil hingga terbesar.**
   - **Contoh: Median dari 3, 1, dan 5 adalah 3 karena saat diurutkan, menjadi 1, 3, 5.**
3. **Mode:**
   - **Mode adalah nilai yang paling sering muncul dalam suatu set data.**
   - **Contoh: Mode dari 2, 3, 4, 2, dan 5 adalah 2 karena 2 muncul lebih sering.**
4. **Deviasi Standar:**
   - **Deviasi standar mengukur seberapa tersebar data dari nilai rata-ratanya.**
   - **Contoh: Deviasi standar dari 1, 2, dan 3 adalah sekitar 0.82.**

DATA PREPARATION

1. Handling Missing Values:
   - Definisi:
     - Nilai-nilai yang hilang dalam data bisa mengganggu analisis dan pemodelan.
2. Encoding Variabel Kategorikal:
   - Definisi:
     - Model-machine learning memerlukan variabel kategorikal untuk diubah ke dalam format numerik.
3. Scaling dan Normalisasi:
   - Definisi:
     - Menskalakan dan menormalkan data membantu model-machine learning bekerja dengan lebih baik.

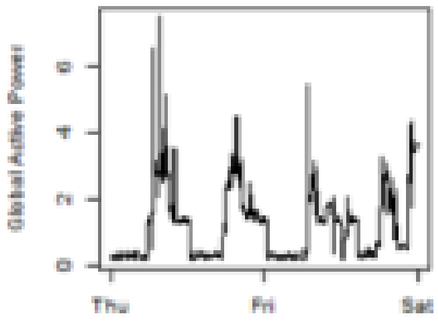
EDA

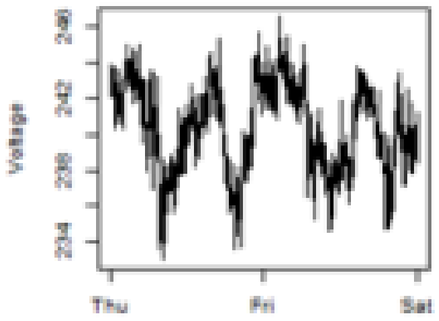
Raw Data

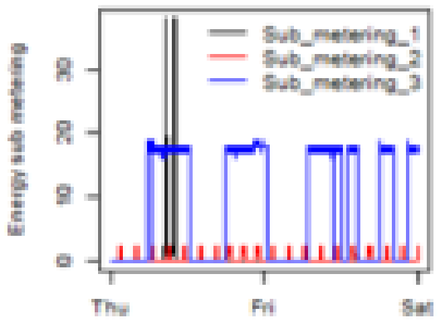
Data Preprocessing

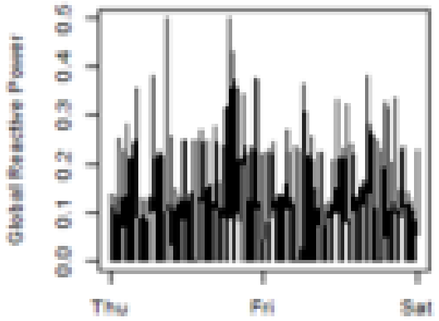
Clean, Structured Data

Exploratory Data Analysis (EDA)









Models & Algorithms

Data-driven Products

- Insights
- Reports
- Visual Graphs

MODELLING

Modelling:
Definisi:

- **Modelling merujuk pada proses membuat representasi matematis atau statistik dari suatu sistem atau fenomena dengan tujuan untuk memahami, menjelaskan, atau meramalkan perilaku sistem tersebut.**

Pentingnya:

- **Pemahaman yang Lebih Baik: Modelling membantu kita memahami hubungan antar variabel dalam data.**
- **Peramalan dan Prediksi: Model dapat digunakan untuk meramalkan atau memprediksi hasil berdasarkan data yang ada.**
- **Optimisasi Keputusan: Dengan model, kita dapat mengoptimalkan keputusan dan strategi berdasarkan informasi yang diberikan oleh data.**

MACHINE LEARNING VS DEEP LEARNING

Machine Learning:
Definisi:

- **Machine Learning (ML) adalah cabang dari kecerdasan buatan yang menggunakan algoritma untuk memberikan sistem kemampuan untuk belajar dari data dan membuat keputusan atau tindakan tanpa secara eksplisit diprogram**

Deep Learning:
Definisi:

- **Deep Learning adalah cabang dari Machine Learning yang menggunakan neural networks berlapis (deep neural networks) untuk menggali struktur fitur dalam data dan membuat keputusan.**

PREDIKSI VS KLASIFIKASI VS KLUSTERING

Modelling untuk Prediksi:
Definisi:

- **Modelling prediksi bertujuan untuk membangun suatu model yang dapat memprediksi nilai-nilai di masa depan berdasarkan pola atau tren yang teridentifikasi dalam data historis.**

Modelling untuk Klasifikasi:
Definisi:

- **Modelling klasifikasi digunakan untuk mengidentifikasi kategori atau label tertentu dari suatu data berdasarkan pola atau karakteristik tertentu yang ada dalam data tersebut.**

PREDIKSI VS KLASIFIKASI VS KLUSTERING

Modelling untuk Klustering:
Definisi:

- **Modelling klustering digunakan untuk mengelompokkan data ke dalam kelompok-kelompok yang memiliki kesamaan karakteristik atau pola tertentu.**

MODEL EVALUATION

Evaluasi model adalah tahap kritis dalam pembangunan model untuk memastikan bahwa model yang dikembangkan dapat memberikan kinerja yang baik dan dapat diandalkan dalam mengatasi tugas tertentu

Model Evaluation untuk Model Prediksi:

1. **Mean Squared Error (MSE):**
   - Mengukur seberapa dekat prediksi model dengan nilai sebenarnya.
   - Semakin kecil MSE, semakin baik modelnya.
2. **R-Squared (R^2):**
   - Memberikan indikasi seberapa baik variabilitas dalam data dapat dijelaskan oleh model.
   - Rentang nilai R^2 antara 0 dan 1, dan semakin mendekati 1, semakin baik modelnya.
3. **MAE (Mean Absolute Error):**
   - Mengukur rata-rata nilai absolut dari perbedaan antara prediksi dan nilai sebenarnya.

Model Evaluation untuk Model Prediksi:

4. MAPE (Mean Absolute Percentage Error):
   - Mengukur rata-rata persentase kesalahan prediksi terhadap nilai sebenarnya.
5. Analisis Residual:
   - Memeriksa residu (selisih antara prediksi dan nilai sebenarnya) untuk memastikan tidak ada pola yang tersisa.

Model Evaluation untuk Model Klasifikasi:

1. **Accuracy:**
   - Seberapa akurat model dalam memprediksi label kelas.
   - (Jumlah prediksi benar) / (Jumlah total prediksi).
2. **Precision:**
   - Seberapa banyak dari yang diprediksi sebagai positif yang sebenarnya positif.
   - TP / (TP + FP).
3. **Recall (Sensitivitas):**
   - Seberapa banyak dari yang sebenarnya positif yang terdeteksi oleh model.

TP / (TP + FN).

Model Evaluation untuk Model Klasifikasi:

4. F1-Score:

- **Kombinasi dari precision dan recall.**
- **2 * (Precision * Recall) / (Precision + Recall).**

5. ROC-AUC (Receiver Operating Characteristic - Area Under the Curve):

- **Mengukur kinerja model di berbagai tingkat ambang batas.**
- **Semakin tinggi AUC, semakin baik kinerja modelnya.**

Model Evaluation untuk Model Klustering:

1. **Silhouette Score:**
   - **Mengukur seberapa baik objek dalam satu kluster dibandingkan dengan kluster lainnya.**
   - **Rentang nilai -1 hingga 1, dan nilai yang lebih tinggi menunjukkan kluster yang lebih baik.**
2. **Davies-Bouldin Index:**
   - **Mengukur seberapa baik kluster terbentuk dengan memperhitungkan jarak antara kluster.**
   - **Semakin rendah nilainya, semakin baik kluster.**

Model Evaluation untuk Model Klustering:

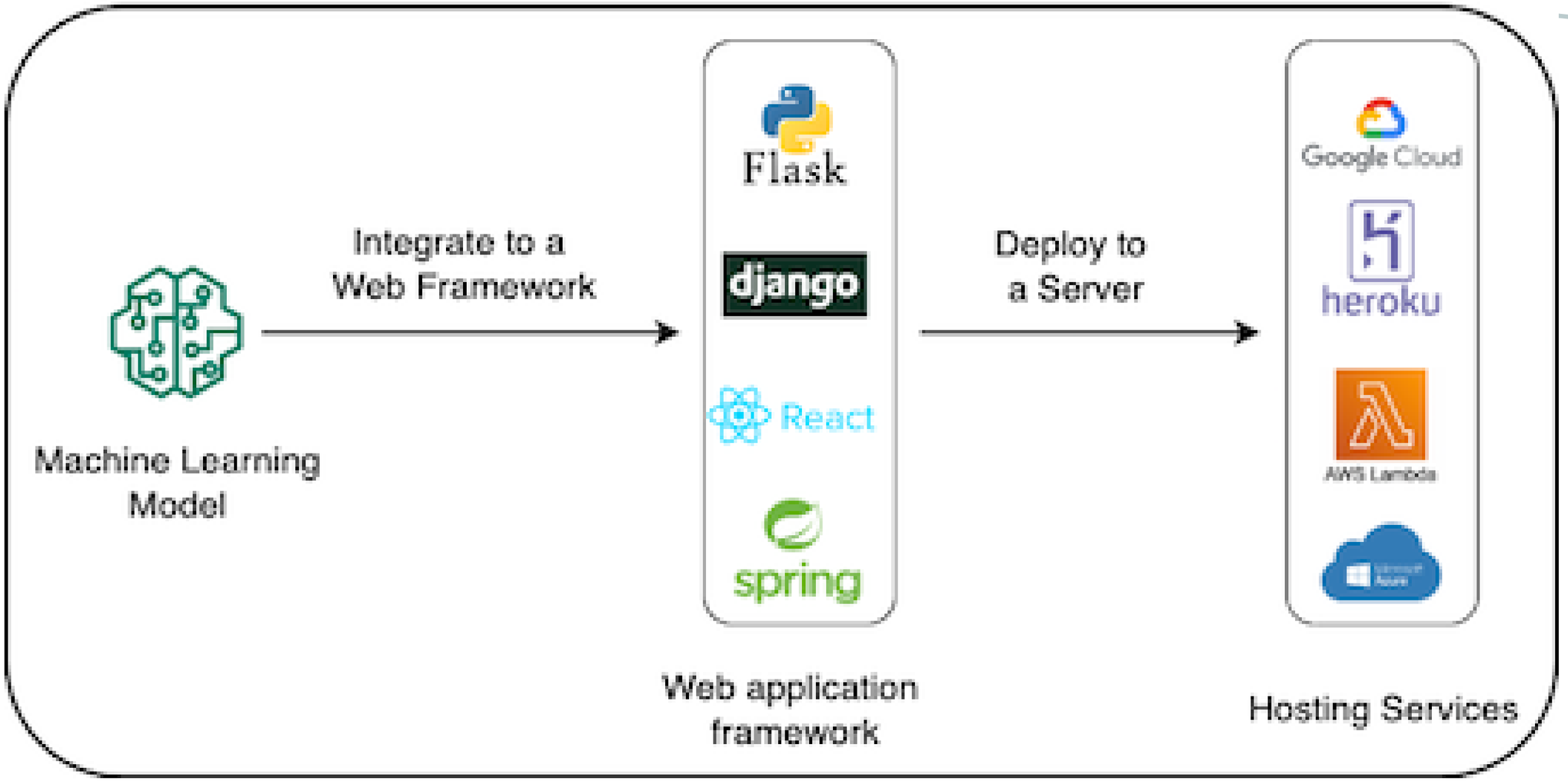
3. Adjusted Rand Index (ARI):

- **Mengukur seberapa serupa hasil klustering dengan ground truth jika ada.**
- **Rentang nilai -1 hingga 1, dan nilai yang lebih tinggi menunjukkan kesamaan yang lebih baik.**

4. Homogeneity, Completeness, dan V-Measure:

- **Menyediakan metrik lain untuk mengevaluasi kualitas klustering.**
- **Homogeneity mengukur sejauh mana setiap kluster hanya berisi anggota dari satu kelas.**
- **Completeness mengukur sejauh mana semua anggota dari satu kelas ada dalam satu kluster.**
- **V-Measure adalah harmonisasi dari keduanya.**

DEPLOYING


Model
Deployment
Integrate to a
Web Framework
Machine Learning
Model
Flask
django
React
spring
Web application
framework
Deploy to
a Server
Google Cloud
heroku
AWS Lambda
Hosting Services

PRAKTEK

THANK YOU

Have any question?