Konsep Data Mining

Sevi Nurafni
Data Sains, Fakultas Sains dan Teknologi
Universitas Koperasi Indonesia
Github.com/sevinurafni

Contents:

- Data Simulasi
- Simulasi Monte Carlo
- Data Training Data Testing
- Metodologi, Model, dan Algoritma Data mining

Data Simulasi

Data Simulasi

- Simulasi dapat dianggap sebagai tiruan dari proses dunia nyata dari waktu ke waktu
- Simulasi data adalah proses pengambilan data dalam jumlah besar dan menggunakannya untuk meniru skenario atau kondisi dunia nyata.
- Namun, simulasi hanya seakurat model yang menjadi dasarnya, sehingga penting untuk memiliki pemahaman yang baik mengenai model tersebut sebelum menggunakan pendekatan ini.

https://psyteachr.github.io/data-skills-v1/simulation.html

Kelebihan Data Simulasi

- Memungkinkan pembuatan model yang komprehensif dari sistem yang kompleks dan dinamis;
- Memberdayakan pengambilan keputusan berbasis data dan perencanaan strategis;
- Membantu menguji hipotesis, memahami hubungan, dan meningkatkan prediksi;
- Memungkinkan studi tentang fenomena yang sulit atau tidak mungkin untuk diselidiki secara langsung; dan
- Menghasilkan data sintetis yang mewakili populasi atau kondisi tertentu yang kemudian dapat digunakan untuk pengembangan ML dan AI.

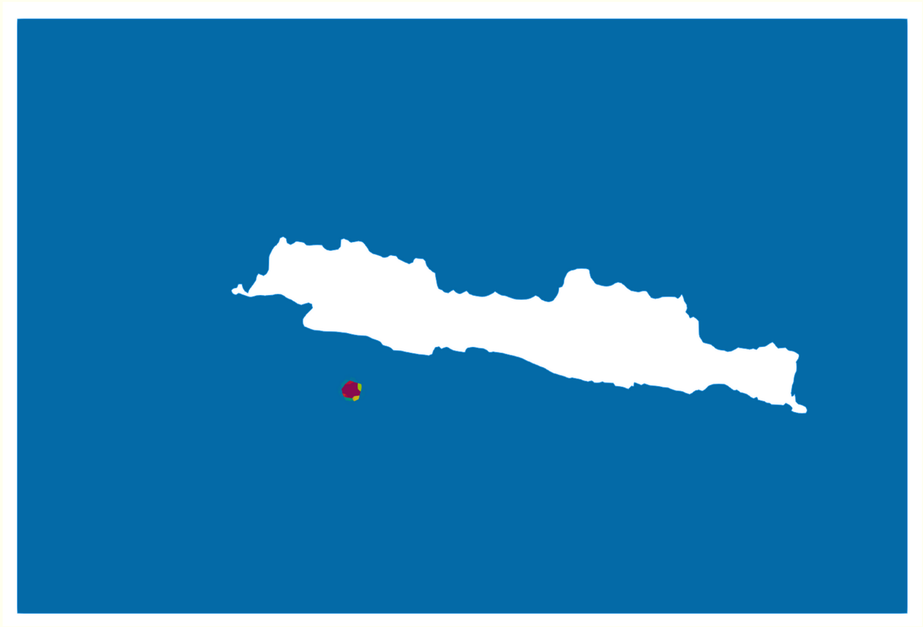
https://psyteachr.github.io/data-skills-v1/simulation.html

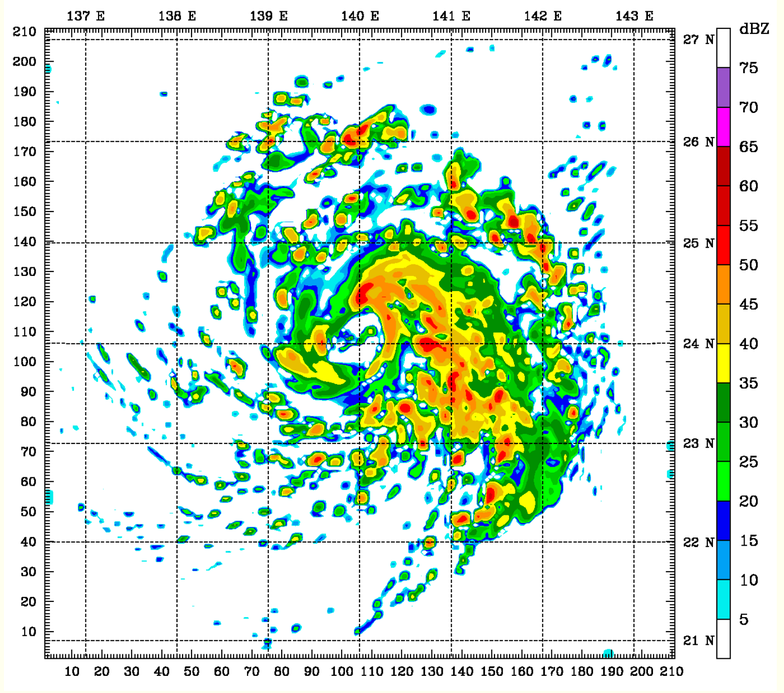
Analisis Hubungan Kompleks Antar Data

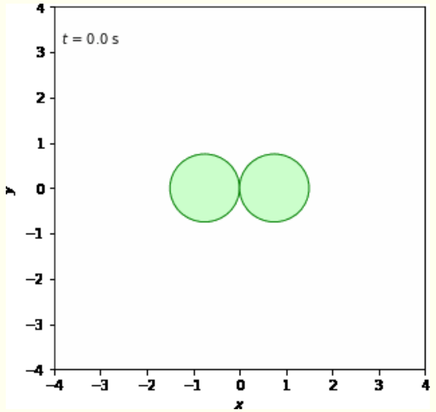
- Simulasi data memungkinkan kita melihat hubungan kompleks antara berbagai variabel dalam sistem:
  - Interaksi Multivariabel: Bagaimana perubahan satu variabel mempengaruhi variabel lain.
  - Dampak Kebijakan: Memprediksi dampak dari kebijakan tertentu pada sistem.
  - Keterkaitan Sistem: Memahami bagaimana komponen-komponen dalam sistem saling mempengaruhi satu sama lain.

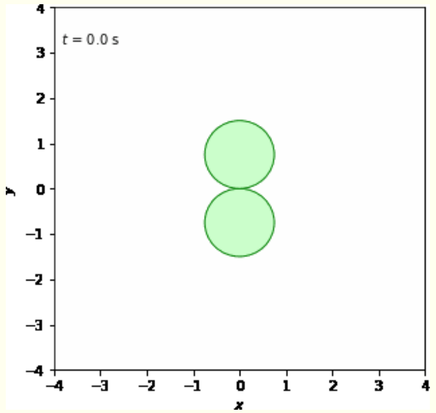
https://mostly.ai/blog/data-simulation

Studi Kasus Simulasi









Studi Kasus Simulasi

$$\vec{R} = \frac{m_1\vec{r_1} + m_2\vec{r_2}}{m_1 + m_2} \quad (1)$$

$$\vec{r} = \vec{r_2} - \vec{r_1} \quad (2)$$

distance-vector length $\vec{r}$,

$$\vec{r} = \vec{r_{b1}} + \vec{r_{b2}} \quad (3)$$

$$\vec{r_1} = \vec{R} - \frac{m_2}{m_1 + m_2}\vec{r} \quad (4)$$

$$\vec{r_2} = \vec{R} - \frac{m_1}{m_1 + m_2}\vec{r} \quad (5)$$

Tesis

2.3 Statement of the Problem

Let $\Omega \subset \mathbb{R}$ be bounded domain and T a positive constant. We consider the problem to find $(\phi, u) : \overline{\Omega}$ x $[0,T] \to \mathbf{R}$ x $\mathbb{R}^2$ such that

$$\begin{cases} \frac{\partial \phi}{\partial t} + \nabla \cdot (u\phi) = 0 & \text{in} \quad \Omega \times (0,T), \\ \rho\phi \left[\frac{\partial u}{\partial t} + (u \cdot \nabla)u\right] - 2\mu\nabla \cdot (\phi D(u)) + \rho g \phi \nabla \eta = 0 & \text{in} \quad \Omega \times (0,T), \\ \phi = \eta + \zeta & \text{in} \quad \Omega \times (0,T), \end{cases} \quad (2.20)$$

where ∇ is the nabla operator on R^2, t represents time and T the final time, with boundary conditions

$$u = 0 \quad \text{on} \quad \Gamma_D \times (0,T), \quad (2.21)$$

$$(D(u)n) \times n = 0, u \cdot n = 0 \quad \text{on} \quad \Gamma_S \times (0,T), \quad (2.22)$$

$$u = c\frac{\eta}{\phi}n \quad \text{on} \quad \Gamma_T \times (0,T) \quad (2.23)$$

and initial conditions

$$u = u^0, \quad \eta = \eta^0 \quad \text{in} \quad \Omega, \quad \text{at} \quad t = 0 \quad (2.24)$$

Monte Carlo

Pendahuluan

Simulasi Metode Monte Carlo (MC) adalah bagian dari algoritme komputasi yang menggunakan proses pengambilan sampel acak berulang untuk membuat estimasi numerik dari parameter yang tidak diketahui.

Metode ini sangat berguna dalam berbagai bidang, seperti keuangan, fisika, teknik, dan ilmu komputer, karena mampu menangani ketidakpastian dan variabilitas.

Jenis Distribusi

- Distribusi Uniform: Semua nilai dalam rentang tertentu memiliki peluang yang sama.
- Distribusi Normal: Nilai berkumpul di sekitar mean dengan probabilitas yang menurun secara simetris.
- Distribusi Eksponensial: Peluang menurun secara eksponensial, sering digunakan untuk model kejadian jarang.
- Distribusi Binomial: Menggambarkan jumlah keberhasilan dalam sejumlah percobaan tertentu.

Cara Kerja

- Siapkan model prediktif, mengidentifikasi variabel dependen yang akan diprediksi dan variabel independen (juga dikenal sebagai variabel input, risiko, atau prediktor) yang akan mendorong prediksi.
- Tentukan distribusi probabilitas dari variabel independen. Gunakan data historis dan/atau penilaian subjektif analis untuk menentukan rentang nilai yang mungkin terjadi dan menetapkan bobot probabilitas untuk masing-masing nilai.
- Jalankan simulasi berulang kali, menghasilkan nilai acak dari variabel independen. Lakukan hal ini hingga diperoleh hasil yang cukup untuk membuat sampel yang representatif dari jumlah kombinasi yang hampir tak terbatas.

https://www.ibm.com/topics/monte-carlo-simulation

Contoh Kasus - Simulasi Monte Carlo untuk Harga Saham

Inisialisasi:

- Tentukan parameter model seperti harga awal saham (S0), tingkat pengembalian (mu), volatilitas (sigma), waktu total (T), dan jumlah langkah waktu (N).
- Tentukan jumlah simulasi (num_simulations).

Generate Data:

- Untuk setiap simulasi:
  - Inisialisasi harga saham pada waktu t=0 dengan harga awal (S0).
  - Untuk setiap langkah waktu:
    - Hasilkan nilai acak normal standar (Z).

Rule:

- Setiap nilai acak dimasukan ke model BGM untuk memperbarui harga saham

Simpan Hasil:

- Simpan harga saham untuk setiap simulasi pada setiap langkah waktu.

Tampilkan Hasil:

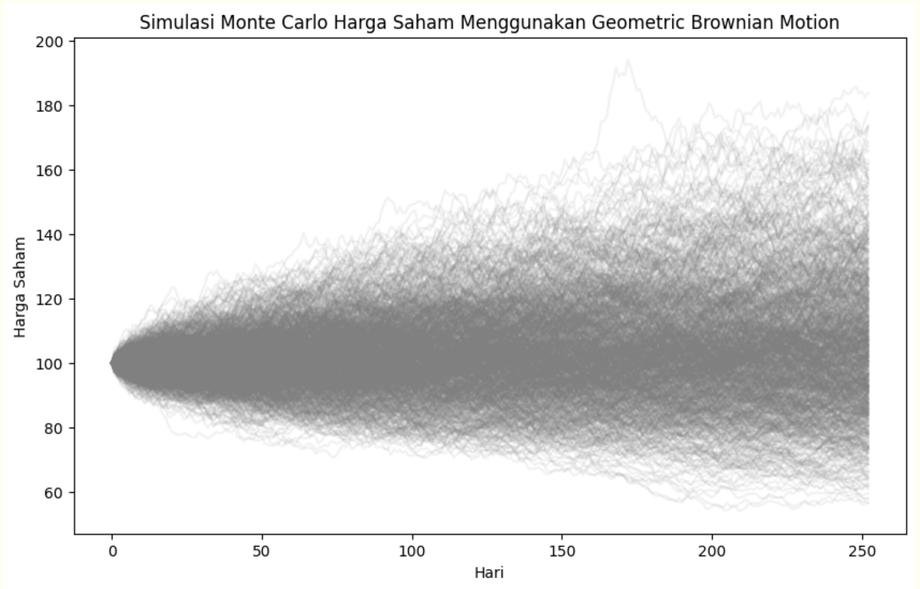
- Visualisasikan data yang dihasilkan.

Contoh Kasus - Simulasi Monte Carlo untuk Harga Saham

```
import numpy as np
import matplotlib.pyplot as plt

def simulate_gbm(S0, mu, sigma, T, N, num_simulations):
    dt = T / N
    prices = np.zeros((num_simulations, N + 1))
    prices[:, 0] = S0

    for t in range(1, N + 1):
        Z = np.random.standard_normal(num_simulations)
        prices[:, t] = prices[:, t - 1] * np.exp((mu -
     0.5 * sigma**2) * dt + sigma * np.sqrt(dt) * Z)

    return prices

# Parameter model
S0 = 100  # Harga awal saham
mu = 0.05  # Tingkat pengembalian
sigma = 0.2  # Volatilitas
T = 1.0  # Waktu total (misalnya 1 tahun)
N = 252  # Jumlah langkah waktu (misalnya 252 hari
perdagangan dalam setahun)
num_simulations = 1000  # Jumlah simulasi
```

https://colab.research.google.com/drive/1Cl1M8XffACV383pSlH5PhBuR3Lm4vDPz?usp=sharing

```
# Generate data menggunakan simulasi Monte Carlo
prices = simulate_gbm(S0, mu, sigma, T, N, num_simulations)

# Visualisasikan hasil
plt.figure(figsize=(10, 6))
plt.plot(prices.T, color='grey', alpha=0.1)
plt.title('Simulasi Monte Carlo Harga Saham Menggunakan
Geometric Brownian Motion')
plt.xlabel('Hari')
plt.ylabel('Harga Saham')
plt.show()
```



Data Training
Data Testing

Training VS Testing

- Data Training: Kumpulan data yang digunakan untuk membangun model. Data ini digunakan untuk 'melatih' algoritma sehingga dapat memahami pola dan hubungan dalam data.
- Data Testing: Kumpulan data yang digunakan untuk menguji model yang telah dibangun. Data ini tidak digunakan dalam proses pelatihan sehingga bisa digunakan untuk mengevaluasi kinerja model secara objektif.

As Andrew Ng said, “The training data is the food for the model, and the testing data is the dessert.”

https://www.geeksforgeeks.org/training-data-vs-testing-data/

Hastie, T., Tibshirani, R., & Friedman, J. (2009). The Elements of Statistical Learning: Data Mining, Inference, and Prediction. Springer.

Metodologi, Model, dan Algoritma pada Data Mining

Metodologi

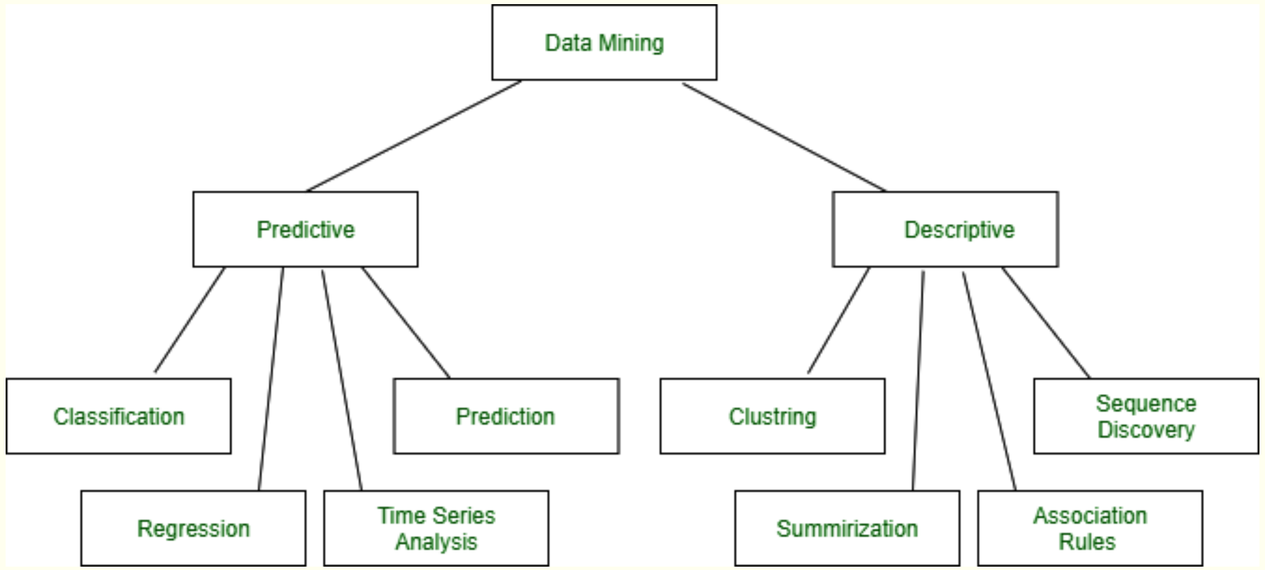
- Metodologi data mining mengacu pada proses dan langkah-langkah yang diikuti untuk melakukan data mining. Salah satu metodologi yang paling populer adalah CRISP-DM (Cross Industry Standard Process for Data Mining). CRISP-DM terdiri dari enam fase utama:
    1. Business Understanding: Memahami tujuan proyek dan kebutuhan bisnis.
    2. Data Understanding: Mengumpulkan dan mengeksplorasi data untuk mendapatkan wawasan awal.
    3. Data Preparation: Membersihkan dan memformat data agar siap untuk dianalisis.
    4. Modeling: Memilih dan menerapkan berbagai teknik pemodelan untuk data.
    5. Evaluation: Mengevaluasi model untuk memastikan memenuhi kebutuhan bisnis.
    6. Deployment: Menerapkan model ke dalam proses bisnis sehari-hari.

https://www.geeksforgeeks.org/training-data-vs-testing-data/

Hastie, T., Tibshirani, R., & Friedman, J. (2009). The Elements of Statistical Learning: Data Mining, Inference, and Prediction. Springer.

Model

Model mengacu pada metode yang biasanya digunakan untuk menyajikan informasi dan berbagai cara yang dapat digunakan untuk menerapkan informasi pada pertanyaan dan masalah tertentu.



https://www.google.com/search?client=safari&sca_esv=2012520a95282aaf&rls=en&q=models+in+data+mining+definition&uds=ADvngMhV_RxKzGxLaOkkOkvk2ezDGF0Kt9YcdDa3rNln0bFj6ov3SbPWCb4FHcwDYQePpd_sZrmT0H4ZpFUA-OUkJZL7GQQGpqt9gjrQ8qFV8uTwydXpaSpsilCr--73Vm7G0JAMxzcJ4iJEWgkpxtdFT_PvhZQ1rQeAaJ6rjnOUnjyuYp-QVZsR7vtSb8AgQoct3NDto4ak5hglTx3w470Ve_Nbc0svi8SvLGNhUvz6Bvz8T-Txqab_rDMQ6gLt0fFF6uauTJ1uxgr4XmoSQX75T2LANMd3odc4d8WiHiNfOckllvqnwqmUFMm7xG0viSZDHnyxEHJ2-ed99_qRChS-qg033g0A3A&udm=2&prmd=ivsnbmtz&sa=X&ved=2ahUKEwjridiBgZGGAxUPWGwGHYVCDPMQtKgLegQIDRAB&biw=1440&bih=820&dpr=2#vhid=s8naieCU9MB8FM&vssid=mosaic

Algoritma

Algoritma adalah langkah-langkah atau prosedur yang digunakan untuk membangun model data mining.

- Regresi Linear: Digunakan untuk memprediksi nilai kontinu berdasarkan hubungan linear antara variabel independen dan dependen.
- Decision Tree: Digunakan untuk klasifikasi dan prediksi dengan cara membagi data ke dalam subset berdasarkan aturan keputusan.
- K-means: Digunakan untuk mengelompokkan data menjadi beberapa cluster berdasarkan kemiripan karakteristik.

https://www.google.com/search?client=safari&sca_esv=2012520a95282aaf&rls=en&q=models+in+data+mining+definition&uds=ADvngMhV_RxKzGxLaOkkOkvk2ezDGF0Kt9YcdDa3rNln0bFj6ov3SbPWCb4FHcwDYQePpd_sZrmT0H4ZpFUA-OUkJZL7GQQGpqt9gjrQ8qFV8uTwydXpaSpsilCr--73Vm7G0JAMxzcJ4iJEWgkpxtdFT_PvhZQ1rQeAaJ6rjnOUnjyuYp-QVZsR7vtSb8AgQoct3NDto4ak5hglTx3w470Ve_Nbc0svi8SvLGNhUvz6Bvz8T-Txqab_rDMQ6gLt0fFF6uauTJ1uxgr4XmoSQX75T2LANMd3odc4d8WiHiNfOckllvqnwqmUFMm7xG0viSZDHnyxEHJ2-ed99_qRChS-qg033g0A3A&udm=2&prmd=ivsnbmtz&sa=X&ved=2ahUKEwjridiBgZGGAxUPWGwGHYVCDPMQtKgLegQIDRAB&biw=1440&bih=820&dpr=2#vhid=s8naieCU9MB8FM&vssid=mosaic

Tugas

1. Jalankan ulang program Monte Carlo untuk harga saham sebanyak 5 kali dengan jumlah simulasi yang berbeda-beda. Apa yang anda pahami dari output tersebut

2. Dalam konteks Data Mining, jelaskan perbedaan antara metodologi, model,

dan algoritma. Bagaimana ketiga konsep ini saling berhubungan dalam proses

analisis data?

https://www.google.com/search?client=safari&sca_esv=2012520a95282aaf&rls=en&q=models+in+data+mining+definition&uds=ADvngMhV_RxKzGxLaOkkOkvk2ezDGF0Kt9YcdDa3rNln0bFj6ov3SbPWCb4FHcwDYQePpd_sZrmT0H4ZpFUA-OUkJZL7GQQGpqt9gjrQ8qFV8uTwydXpaSpsilCr--73Vm7G0JAMxzcJ4iJEWgkpxtdFT_PvhZQ1rQeAaJ6rjnOUnjyuYp-QVZsR7vtSb8AgQoct3NDto4ak5hglTx3w470Ve_Nbc0svi8SvLGNhUvz6Bvz8T-Txqab_rDMQ6gLt0fFF6uauTJ1uxgr4XmoSQX75T2LANMd3odc4d8WiHiNfOcklIvqnwqmUFMm7xG0viSZDHnyxEHJ2-ed99_qRChS-qg033g0A3A&udm=2&prmd=ivsnbmtz&sa=X&ved=2ahUKEwjridiBgZGGAxUPWGwGHYVCDPMQtKgLegQIDRAB&biw=1440&bih=820&dpr=2#vhid=s8naieCU9MB8FM&vssid=mosaic

Terima Kasih

Jangan ragu untuk mengirim pesan kepada saya untuk mengajukan pertanyaan dan diskusi