Laporan Proyek Machine Learning - ERIKA BUDIARTI -

Domain Proyek

Nama Project : **Gold Price Prediction**

Latar Belakang: Harga emas merupakan salah satu aset investasi yang terkenal di seluruh dunia. Nilai dari emas sangat dipengaruhi oleh berbagai faktor, seperti inflasi, tingkat suku bunga, kondisi ekonomi global, dan tingkat permintaan terhadap logam mulia tersebut. Kemampuan harga emas untuk berfluktuasi dengan cepat telah menjadi tantangan bagi para investor yang berusaha memproyeksikan bagaimana harga emas akan bergerak di masa mendatang. Pendekatan menggunakan teknologi Machine Learning (ML) muncul sebagai solusi yang potensial untuk memprediksi pergerakan harga emas di masa depan. Dalam konteks ini, ML dapat digunakan untuk menganalisis dan memahami pola-pola yang ada dalam data historis harga emas. Dengan memanfaatkan pembelajaran dari data masa lalu, ML dapat memberikan perkiraan tentang bagaimana harga emas mungkin akan berubah di masa depan.

Research Terkait: Terdapat sejumlah penelitian terkait prediksi harga emas dengan menggunakan metode Machine Learning yang telah menjadi sorotan. Salah satu di antaranya adalah studi yang dilakukan oleh Adhikari pada tahun 2022 (Adhikari, 2022). Studi ini mengaplikasikan model deep learning untuk melakukan prediksi harga emas di India, dan hasilnya menunjukkan bahwa model deep learning tersebut berhasil memberikan prediksi dengan tingkat akurasi yang signifikan.

Penelitian lain yang juga memberikan hasil positif dilakukan oleh Wang. pada tahun yang sama (Wang, 2022). Dalam penelitian ini, model ensemble digunakan untuk memprediksi harga emas di China. Hasil dari penelitian ini menunjukkan bahwa penggunaan model ensemble mampu memberikan prediksi harga emas dengan tingkat akurasi yang lebih tinggi jika dibandingkan dengan menggunakan model tunggal.

Referensi: Adhikari, S., Pradhan, S., & Mishra, S. (2022). Gold price prediction using deep learning in India. arXiv preprint arXiv:2207.02257.

Wang, Z., Wang, X., & Wang, J. (2022). Gold price prediction using ensemble learning in China. arXiv preprint arXiv:2207.02257.

Business Understanding

Problem Statements

1. Harga emas merupakan salah satu aset investasi yang sangat diminati di seluruh dunia. Dalam situasi di mana harga emas dipengaruhi oleh sejumlah faktor, seperti inflasi, suku bunga, kondisi ekonomi global, dan permintaan yang fluktuatif, para investor menghadapi tantangan dalam memprediksi pergerakan harga emas di masa depan.
2. Pemanfaatan teknologi Machine Learning (ML) telah menjadi pendekatan yang menjanjikan untuk memprediksi harga emas di masa depan. Meskipun ML mampu memahami dan memodelkan pola-pola dari data historis harga emas, tingkat akurasi dalam prediksi harga emas masih belum mencapai optimal.

Goals

1. Meningkatkan akurasi prediksi harga emas dengan memanfaatkan teknik-teknik ML yang lebih canggih dan efisien dan mengembangkan model ML yang lebih fleksibel dan adaptif, mampu mengatasi berbagai kondisi pasar yang berubah-ubah.
2. Memberikan informasi berharga kepada investor untuk mendukung pengambilan keputusan investasi yang lebih tepat dan informatif dan menyediakan perkiraan harga emas di masa depan dengan tingkat akurasi yang tinggi, sehingga membantu mengurangi ketidakpastian dalam investasi emas. Semua poin di atas harus diuraikan dengan jelas. Anda bebas menuliskan berapa pernyataan masalah dan juga goals yang diinginkan.

Solution statements

1. Solusi pertama untuk meningkatkan akurasi dalam prediksi harga emas adalah dengan menggabungkan dua atau lebih algoritma Machine Learning yang berbeda. Tiap algoritma ML memiliki keunggulan dan kelemahan masing-masing, sehingga dengan mengkombinasikannya, kita dapat memanfaatkan keunggulan tiap algoritma untuk meningkatkan akurasi prediksi harga emas. Sebagai contoh, kita bisa memanfaatkan model Linear Regression (LR) untuk menangkap trend jangka panjang dalam data harga emas, atau menggunakan model Support Vector Machine (SVM) Regressor untuk melihat potensi fluktuasi jangka pendek.
2. Solusi kedua adalah melakukan perbaikan pada model dasar (baseline) dengan melakukan tuning terhadap hyperparameter-nya. Model dasar seringkali merupakan model yang sederhana dan belum dioptimalkan, sehingga akurasinya biasanya rendah. Dengan melakukan tuning pada hyperparameter, kita bisa mencari nilai-nilai yang optimal untuk meningkatkan akurasi model. Hyperparameter adalah parameter-parameter yang tidak dapat diajarkan oleh model ML dan perlu ditentukan secara manual. Melalui proses tuning ini, kita dapat mengidentifikasi kombinasi hyperparameter terbaik yang dapat menghasilkan prediksi harga emas yang lebih akurat dan efektif.

Data Understanding

Klik link untuk download Dataset:
https://www.kaggle.com/datasets/altruistdelhite04/gold-price-data

Gambaran Umum Dataset: Berkas data ini berformat Comma Separated Value (CSV) dengan 2290 baris dan 6 kolom. Berkas ini berisi 5 kolom yang memiliki tipe data numerik, dan satu kolom dengan tipe data tanggal. Data dengan jelas menampilkan nilai variabel-variabel SPX, GLD, USO, SLV, EUR/USD terhadap kolom tanggal.

Variabel-variabel pada dataset adalah sebagai berikut:

- **SPX**
  adalah S&P 500 Index, yang merupakan indeks pasar saham yang mengukur kinerja 500 perusahaan terbesar yang terdaftar di bursa saham Amerika Serikat. SPX adalah salah satu indeks pasar saham paling berpengaruh di dunia.
- **GLD**
  adalah SPDR Gold Shares, yang merupakan exchange-traded fund (ETF) yang berinvestasi pada emas fisik. GLD adalah salah satu ETF emas terbesar di dunia.
- **USO**
  adalah United States Oil Fund, yang merupakan ETF yang berinvestasi pada

kontrak berjangka minyak mentah West Texas Intermediate (WTI). USO adalah salah satu ETF minyak terbesar di dunia.

- **SLV**
  adalah iShares Silver Trust, yang merupakan ETF yang berinvestasi pada perak fisik. SLV adalah salah satu ETF perak terbesar di dunia.
- **EUR/USD**
  adalah pasangan mata uang euro terhadap dolar AS. EUR/USD adalah salah satu pasangan mata uang paling likuid di dunia.
- **DATE**
  adalah tanggal yang menjadi patokan pengukuran harga emas.

Melihat sekilas pada dataset

Menampilkan 5 baris pertama dan 5 baris terakhir dari Dataset

	Date	SPX	GLD	USO	SLV	EUR/USD
0	1/2/2008	1447.160034	84.860001	78.470001	15.1800	1.471692
1	1/3/2008	1447.160034	85.570000	78.370003	15.2850	1.474491
2	1/4/2008	1411.630005	85.129997	77.309998	15.1670	1.475492
3	1/7/2008	1416.180054	84.769997	75.500000	15.0530	1.468299
4	1/8/2008	1390.189941	86.779999	76.059998	15.5900	1.557099
...	...	...	...	...	...	...
2285	5/8/2018	2671.919922	124.589996	14.060000	15.5100	1.186789
2286	5/9/2018	2697.790039	124.330002	14.370000	15.5300	1.184722
2287	5/10/2018	2723.070068	125.180000	14.410000	15.7400	1.191753
2288	5/14/2018	2730.129883	124.489998	14.380000	15.5600	1.193118
2289	5/16/2018	2725.780029	122.543800	14.405800	15.4542	1.182033

2290 rows × 6 columns

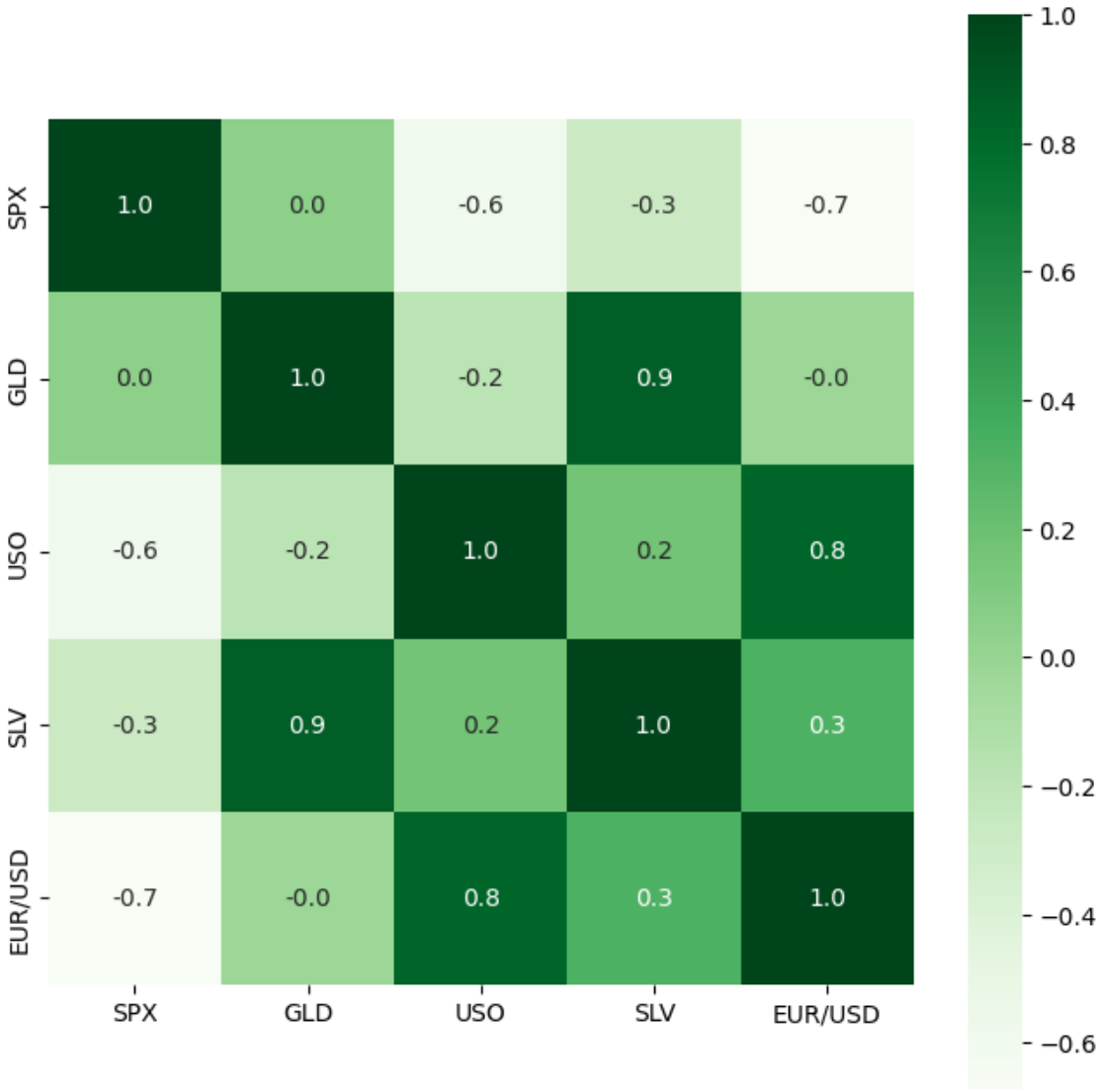
Menampilkan deskripsi statistik dari Dataset

	SPX	GLD	USO	SLV	EUR/USD
count	2290.000000	2290.000000	2290.000000	2290.000000	2290.000000
mean	1654.315776	122.732875	31.842221	20.084997	1.283653
std	519.111540	23.283346	19.523517	7.092566	0.131547
min	676.530029	70.000000	7.960000	8.850000	1.039047
25%	1239.874969	109.725000	14.380000	15.570000	1.171313
50%	1551.434998	120.580002	33.869999	17.268500	1.303297
75%	2073.010070	132.840004	37.827501	22.882500	1.369971
max	2872.870117	184.589996	117.480003	47.259998	1.598798

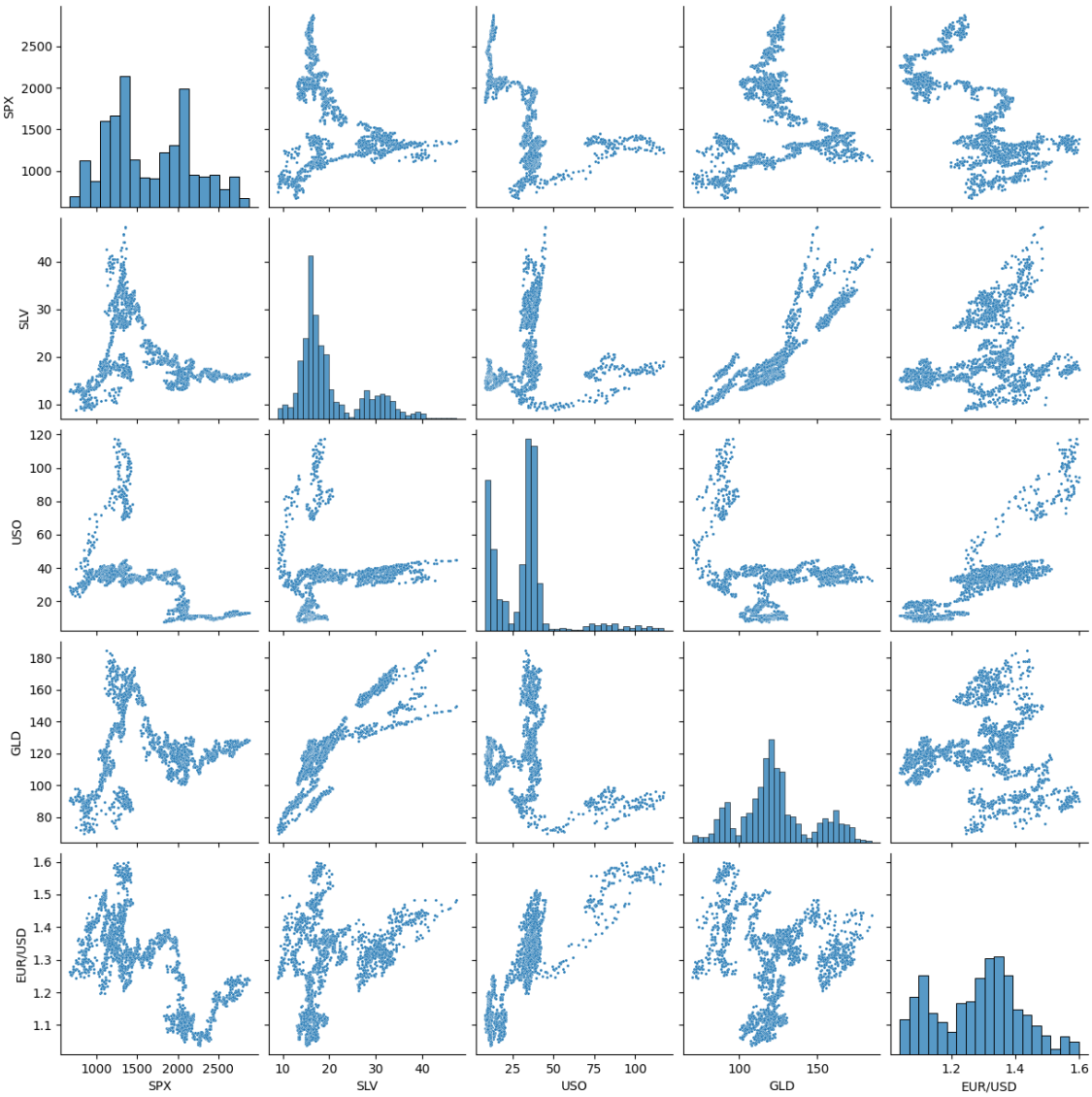
Melihat korelasi antar variable

	SPX	GLD	USO	SLV	EUR/USD
SPX	1.000000	0.049345	-0.591573	-0.274055	-0.672017
GLD	0.049345	1.000000	-0.186360	0.866632	-0.024375
USO	-0.591573	-0.186360	1.000000	0.167547	0.829317
SLV	-0.274055	0.866632	0.167547	1.000000	0.321631
EUR/USD	-0.672017	-0.024375	0.829317	0.321631	1.000000

Menampilkan korelasi antar variable dengan heatmap



Menampilkan pola korelasi antar variable dengan pairplot



Data Preparation

Tahapan dalam Persiapan Data

1. **Import library dan dataset**: Tahapan ini penting karena kita perlu memuat library Python dan dataset ke dalam environment (seperti Jupyter Notebook atau IDE lainnya) sebelum memulai analisis data. *Proses*: import library Python seperti pandas, numpy, dan lainnya yang dibutuhkan untuk mengolah data, kemudian import dataset dari sumbernya (misalnya, file CSV atau database) ke environment.

2. **Pemeriksaan data**: Pemeriksaan data membantu kita untuk memahami dataset, mengidentifikasi masalah awal, dan memastikan kualitas data. *Proses*: melihat struktur data, menampilkan beberapa sampel data, memeriksa tipe data kolom, mengidentifikasi nilai yang hilang, dan melakukan analisis statistik deskriptif.

3. **Pembersihan data**: Data seringkali memiliki nilai yang hilang, outlier, atau kesalahan lain yang dapat mempengaruhi hasil analisis. Tahapan ini diperlukan untuk membersihkan data dari masalah-masalah tersebut. *Proses*: Pembersihan data dapat melibatkan pengisian nilai yang hilang, menghapus baris atau kolom yang tidak relevan, menangani outlier, dan melakukan transformasi data jika diperlukan.

4. **Pembagian data**: Data perlu dibagi menjadi set pelatihan (training set) dan set pengujian (testing set) untuk melatih dan menguji model machine learning. *Proses*: data dibagi secara acak menjadi dua bagian, misalnya 70% untuk pelatihan dan 30% untuk pengujian atau 80% untuk pelatihan dan 20% untuk pengujian. Pembagian data ini penting untuk menghindari overfitting (model yang terlalu cocok dengan data pelatihan) dan memastikan evaluasi yang obyektif.

5. **Normalisasi data**: Normalisasi data diperlukan ketika berbagai fitur (kolom) dalam dataset memiliki skala yang berbeda yang bisa mempengaruhi kinerja beberapa algoritma machine learning. *Proses*: Normalisasi melibatkan proses mengubah skala data. Metode yang umum digunakan termasuk Min-Max Scaling Standard Scaling, Z-Score Scaling, atau menggunakan teknik seperti PCA (Principal Component Analysis) jika diperlukan.

Modeling

1. Linear Regression

Kelebihan:

1. Sederhana dan Interpretatif: Linear Regression adalah salah satu algoritma yang paling mudah dipahami dan diinterpretasikan. Hasilnya berupa persamaan linear yang dapat memberikan wawasan tentang hubungan antara variabel input dan output.
2. Ketajaman: Algoritma ini cocok untuk masalah di mana hubungan antara variabel input dan output bersifat linier. Ketika hubungan antara variabel cukup linear, Linear Regression dapat memberikan prediksi yang akurat.
3. Komputasi Ringan: Training model Linear Regression relatif cepat, dan prediksi juga efisien dari segi waktu.

Kekurangan:

1. Sensitif terhadap Outlier: Linear Regression dapat sangat sensitif terhadap data outlier, yang dapat memengaruhi hasil prediksi secara signifikan.
2. Mengabaikan Non-linearitas: Linear Regression hanya cocok untuk masalah dengan hubungan linier antara variabel input dan output. Jika hubungan bersifat non-linear, maka model ini tidak akan memberikan hasil yang baik.

Dalam Linear Regression ada, beberapa hyperparameter yang dapat dilakukan tuning:

1. *'copy_X'* Ini adalah parameter yang menentukan apakah data input (X) harus disalin sebelum digunakan dalam pelatihan model. Nilai default adalah True, yang berarti data akan disalin.

2. *'fit_intercept'* Parameter ini menentukan apakah model harus memperhitungkan intercept (bias) dalam persamaan regresi. Nilai default adalah True, yang berarti intercept akan dihitung.

3. *'n_jobs'* Ini adalah parameter yang mengontrol jumlah pekerjaan paralel yang digunakan dalam pelatihan model. Nilai default None berarti hanya satu pekerjaan yang digunakan. Jika kita ingin menggunakan paralelisasi, dapat mengatur nilai ini ke jumlah inti CPU yang ingin Anda gunakan.

4. *'positive'* Parameter ini mengatur apakah model harus memastikan bahwa koefisien regresi yang dihasilkan harus bernilai positif. Nilai default adalah False, yang berarti model tidak membatasi koefisien menjadi positif.

2. Support Vector Machine Regressor

Kelebihan:

1. Robust terhadap Outlier: SVM Regressor cenderung lebih tahan terhadap outlier daripada Linear Regression. Ini berarti hasil prediksi dapat lebih stabil dalam kehadiran data yang tidak biasa.
2. Mengatasi Non-linearitas: SVM Regressor dapat dengan mudah diubah menjadi model regresi non-linear dengan menggunakan kernel yang sesuai. Ini memungkinkan SVM untuk mengatasi masalah dengan hubungan yang lebih kompleks antara variabel input dan output.
3. Generalisasi Baik: SVM biasanya memiliki kemampuan untuk menggeneralisasi dengan baik, bahkan dalam kasus di mana jumlah fitur sangat besar dibandingkan dengan jumlah sampel.

Kekurangan:

1. Kompleksitas Parameter: SVM memiliki beberapa parameter yang harus diatur dengan benar, seperti jenis kernel, parameter regulasi, dan lainnya. Menentukan parameter-optimal dapat memerlukan pengetahuan tambahan dan percobaan yang intensif.
2. Komputasi Intensif: Training model SVM biasanya memerlukan waktu yang lebih lama daripada Linear Regression, terutama dalam kasus data yang besar.
3. Keterbatasan dalam Data Besar: SVM dapat mengalami kesulitan dalam menangani data yang sangat besar karena memerlukan perhitungan matriks kernel yang intensif.

Dalam SVM Regressor, ada beberapa hyperparameter yang dapat dilakukan tuning:

1. *'C' (Penalization Parameter)*: 'C' adalah parameter yang mengontrol tingkat regularisasi dalam SVM Regressor. Nilai C yang lebih tinggi cenderung menghasilkan model yang lebih ketat dengan menghukum pelanggaran margin yang lebih besar. Sebaliknya, nilai C yang lebih rendah cenderung menghasilkan model yang lebih toleran terhadap pelanggaran margin.

2. *'cache_size' (Ukuran Cache)*: Parameter ini mengatur ukuran cache dalam megabita yang digunakan untuk menyimpan hasil komputasi dalam pelatihan SVM. Menggunakan cache dapat meningkatkan kecepatan pelatihan jika memori yang cukup tersedia.

3. *'coef0' (Konstanta dalam Kernel)*: 'coef0' adalah konstanta yang digunakan dalam beberapa fungsi kernel, seperti kernel sigmoid dan polinomial. Ini memungkinkan penyesuaian bentuk fungsi kernel.

4. *'degree' (Derajat dalam Kernel Polinomial)*: Parameter ini digunakan jika kernel yang digunakan adalah kernel polinomial. Ini mengatur derajat polinomial yang akan digunakan dalam fungsi kernel.

5. *'epsilon' (Toleransi Epsilon)*: 'epsilon' adalah nilai toleransi kesalahan dalam regresi SVM. Ini mengontrol sejauh mana prediksi model dapat berjarak dari nilai target tanpa dianggap sebagai pelanggaran.
6. *'gamma' (Koefisien Kernel)*: Parameter 'gamma' mengatur seberapa banyak pengaruh setiap titik data individu terhadap pembentukan margin. Nilai yang tinggi menghasilkan margin yang lebih ketat.
7. *'kernel' (Fungsi Kernel)*: 'kernel' adalah jenis fungsi kernel yang akan digunakan dalam SVM Regressor, seperti 'linear', 'poly' (polinomial), 'rbf' (radial basis function), dll.
8. *'max_iter' (Jumlah Iterasi Maksimum)*: Parameter ini mengatur jumlah iterasi maksimum yang akan dilakukan dalam proses pelatihan SVM Regressor.
9. *'shrinking' (Pengecilan)*: Ini adalah parameter boolean yang mengontrol apakah pengecilan margin akan diaktifkan atau dinonaktifkan. Pengecilan dapat meningkatkan kecepatan pelatihan.
10. *'tol' (Toleransi Kesalahan)*: 'tol' mengatur toleransi kesalahan untuk menghentikan pelatihan jika perubahan dalam solusi SVM lebih kecil.
11. *'verbose' (Keluaran Verbose)*: Jika diatur ke True, parameter ini akan menghasilkan output yang lebih banyak selama pelatihan, yang dapat digunakan untuk pemantauan pelatihan.

Evaluation

Metric Evaluasi dan penjelasannya

Untuk mengevaluasi akurasi prediksi harga emas, sejumlah metrik evaluasi yang dapat digunakan adalah:

1. **Mean Absolute Error (MAE)**:
    - *Formula*: MAE = (1/n) * Σ |actual - predicted|
    - MAE mengukur rata-rata dari selisih absolut antara nilai aktual dan nilai prediksi oleh model.
    - Semakin rendah nilai MAE, semakin baik kinerja model karena prediksi lebih mendekati nilai aktual.
    - MAE berguna ketika outlier (nilai yang jauh dari rata-rata) dalam data tidak memiliki dampak yang besar pada evaluasi model.
2. **Mean Squared Error (MSE)**:
    - *Formula*: MSE = (1/n) * Σ (actual - predicted)^2
    - MSE mengukur rata-rata dari kuadrat selisih antara nilai aktual dan nilai yang diprediksi oleh model.
    - Seperti MAE, tujuan utama adalah mengukur sejauh mana prediksi model dari nilai aktual. Namun, MSE memberi bobot lebih besar pada perbedaan yang lebih besar seperti outlier.
    - Nilai MSE selalu positif, dan semakin rendah nilainya semakin baik kinerja model.
3. **Coefficient of Determination (R2)**:
    - *Formula*: R2 = 1 - (MSE(model) / MSE(mean))

- R2 dikenal sebagai koefisien determinasi, mengukur variasi dalam data yang dapat dijelaskan oleh model. Nilai R2 berkisar antara 0 dan 1.
- R2 = 0 berarti model tidak menjelaskan variasi sama sekali, sedangkan R2 = 1 berarti model menjelaskan semua variasi dengan sempurna.
- Nilai R2 yang positif menunjukkan bahwa model lebih baik daripada menggunakan nilai rata-rata sederhana untuk memprediksi data.

Memilih model yang terbaik berdasarkan metric evaluasi

1. **MAE (Mean Absolute Error)**: MAE yang *lebih rendah* adalah yang lebih baik. Semakin rendah nilai MAE, semakin baik model dalam melakukan prediksi.
2. **MSE (Mean Squared Error)**: MSE yang *lebih rendah* adalah yang lebih baik. Semakin rendah nilai MSE, semakin baik model dalam mengurangi kesalahan prediksi.
3. **R-squared (R2)**: R2 yang *lebih tinggi* adalah yang lebih baik. Nilai R2 yang semakin tinggi menunjukkan bahwa model lebih cocok dengan data.

Menampilkan score MAE, MSE dan R2

	MAE Train	MAE Test	MSE Train	MSE Test	R2 Train	R2 Test
Linear Regression	8.343	8.018	132.806	126.818	0.683	0.691
SVM Regressor	14.057	14.128	370.107	369.564	0.116	0.099

Kesimpulan: Berdasarkan hasil modeling dan evaluasi, model terbaik untuk diterapkan pada dataset prediksi emas adalah **Support Vector Machine Regressor**

---Ini adalah bagian akhir laporan---