

Pergeseran Feature Importance pada Prediksi Pasar Saham Teknologi Menggunakan Machine Learning: Studi Komparatif Pra dan Pasca Pandemi COVID-19

DZAKY RAYSSA BUNTORO

Program Studi Informatika, Fakultas Ilmu Komunikasi dan Informatika, Universitas Muhammadiyah Surakarta, Indonesia¹

dzakyrayssa7@gmail.com¹

Correspondence: dzakyrayssa7@gmail.com¹

ABSTRAK

Tujuan: Penelitian ini bertujuan untuk menganalisis apakah model machine learning (XGBoost dan Random Forest) mengalami pergeseran dalam menentukan fitur terpenting untuk memprediksi arah pergerakan saham teknologi sebelum dan sesudah pandemi COVID-19.

Metode: Data yang digunakan adalah time-series harian dari Invesco QQQ Trust (QQQ) sebagai representasi sektor teknologi Amerika Serikat, serta variabel makroekonomi dan volatilitas. Periode penelitian dibagi menjadi dua rezim: pra-pandemi (2018–2019) dan pasca-pandemi (2021–2022). Model dilatih secara terpisah untuk masing-masing rezim, kemudian dilakukan analisis komparatif terhadap feature importance. Evaluasi model menggunakan metrik Accuracy dan F1 Score.

Hasil: Hasil menunjukkan adanya peningkatan prediktabilitas pasar pada periode pasca-pandemi, dengan F1 Score XGBoost meningkat signifikan sekitar 61% (dari 0,346 menjadi 0,556) dan Random Forest meningkat drastis lebih dari 230% (dari 0,164 menjadi 0,544). Analisis feature importance menunjukkan pergeseran dominasi faktor: pra-pandemi dipengaruhi secara merata oleh harga, teknikal, dan makroekonomi, sedangkan pasca-pandemi lebih didominasi faktor makroekonomi (FedFundsRate) dan volatilitas (ATR, VIX).

Kesimpulan: Penelitian ini menyimpulkan bahwa pandemi COVID-19 menyebabkan perubahan rezim prediktif di pasar saham teknologi, dengan meningkatnya peran faktor makroekonomi dan volatilitas. Temuan ini menegaskan pentingnya adaptasi model prediksi serta memberikan wawasan praktis bagi investor dalam memahami dinamika pasar pasca-pandemi.

Kata Kunci

machine learning; feature importance; pasar saham; pandemi COVID-19; XGBoost; Random Forest

ABSTRACT

Objective: This study aims to examine whether machine learning models (XGBoost and Random Forest) experience shifts in determining the most important features for predicting the direction of technology stock movements before and after the COVID-19 pandemic.

Methods: The dataset consists of daily time-series data from the Invesco QQQ Trust (QQQ) as a proxy for the U.S. technology sector, along with macroeconomic and volatility indicators. The research period is divided into two regimes: pre-pandemic (2018–2019) and post-pandemic (2021–2022). Models were trained separately for each regime, and feature importance was compared. Model performance was evaluated using Accuracy and F1 Score.

Results: The findings revealed an increase in market predictability during the post-pandemic period, with XGBoost's F1 Score rising from 0.346 to 0.556 and Random Forest's from 0.164 to 0.544. Feature importance analysis showed a clear shift: pre-pandemic models relied more evenly on price, technical, and macroeconomic variables, while post-pandemic models were dominated by macroeconomic factors (FedFundsRate) and volatility indicators (ATR, VIX).

Conclusion: The study concluded that the COVID-19 pandemic triggered a predictive regime shift in the technology stock market, with macroeconomic and volatility factors playing a more dominant role. These findings highlighted the need for adaptive predictive models and provide practical insights for investors in navigating the post-pandemic market dynamic

Key Words

machine learning; feature importance; stock market; COVID-19 pandemic; XGBoost; Random Forest

I. PENDAHULUAN

Memprediksi pergerakan pasar saham merupakan salah satu tantangan terbesar dalam analisis keuangan modern, di mana kompleksitas dan dinamika pasar sering membuat metode statistik tradisional tidak memadai (Tsay, 2005). Dalam beberapa tahun terakhir, *machine learning* (ML) muncul sebagai pendekatan yang lebih adaptif karena mampu menangkap pola non-linear yang rumit dari data historis dalam jumlah besar. Model seperti XGBoost dan Random Forest terbukti efektif dalam tugas ini, bahkan mampu mempelajari pola yang mungkin tidak terlihat oleh model statistik biasa (Gu et al., 2020).

Meski demikian, tantangan terbesar dalam menggunakan model ML adalah sifatnya yang sering dipandang sebagai "kotak hitam" (*black box*), di mana alasan di balik keputusan prediksi tidak selalu jelas. Oleh karena itu, teknik analisis *feature importance* menjadi krusial untuk membuka mekanisme internal model. Teknik ini memungkinkan peneliti untuk memahami variabel mana seperti suku bunga, volume perdagangan, atau indikator teknikal yang dianggap paling berpengaruh oleh model dalam membuat prediksi (Fischer & Krauss, 2018).

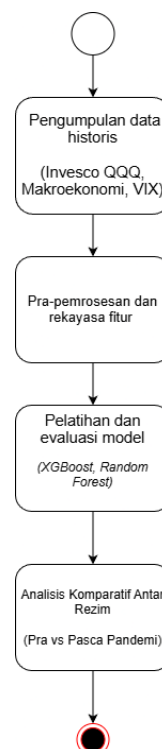
Konteks interpretabilitas ini menjadi sangat relevan ketika pasar mengalami guncangan struktural besar. Pandemi COVID-19 yang terjadi pada tahun 2020 adalah peristiwa luar biasa yang secara fundamental mengubah "aturan main" di pasar keuangan global (Baker et al., 2020). Sebelum pandemi, investor cenderung memperhatikan indikator teknikal dan berita perusahaan, sedangkan setelah pandemi fokus disinyalir beralih ke faktor makroekonomi seperti inflasi dan kebijakan bank sentral (Thorbecke, 2020). Perubahan drastis ini menimbulkan pertanyaan penelitian penting: apakah hierarki informasi yang dianggap "penting" oleh model prediktif juga ikut bergeser?

Berdasarkan konteks tersebut, penelitian ini memberikan kontribusi dengan membandingkan secara empiris hierarki *feature importance* dari model XGBoost dan Random Forest pada dua periode berbeda: pra-pandemi (2018–2019) dan pasca-pandemi (2021–2022). Tujuannya adalah untuk membuktikan adanya pergeseran rezim prediktif akibat guncangan struktural serta memberikan wawasan praktis bagi investor mengenai faktor-faktor dominan yang menggerakkan pasar saham teknologi di era pasca-pandemi.

II. MATERIAL DAN METODE

Desain Penelitian

Penelitian ini menggunakan pendekatan kuantitatif komparatif untuk menganalisis pergeseran hierarki *feature importance* dari model machine learning pada dua rezim pasar yang berbeda: pra-pandemi dan pasca-pandemi. Alur penelitian meliputi lima tahap utama: (1) pengumpulan data historis, (2) pra-pemrosesan dan rekayasa fitur (*feature engineering*), (3) pelatihan dan evaluasi model (*XGBoost, Random Forest*), (4) ekstraksi *feature importance*, dan (5) analisis komparatif antarrezim.



Gambar 1. Diagram Alur Penelitian (Sumber: Pengolahan Data Penulis, 2025)

Data dan Sumber Data

Objek penelitian adalah data *time-series* harian dari Invesco QQQ Trust (ticker: QQQ), sebuah Exchange-Traded Fund (ETF) yang melacak indeks Nasdaq-100 dan dianggap sebagai representasi sektor teknologi Amerika Serikat. Data harga QQQ dan Indeks Volatilitas CBOE (VIX) diperoleh dari Yahoo Finance. Data makroekonomi (FedFundsRate, CPI, DXY) diperoleh dari Federal Reserve Economic Data (FRED). Untuk menjamin replikabilitas penelitian, seluruh proses komputasi dilakukan menggunakan bahasa pemrograman Python versi 3.10. Manipulasi dan pembersihan data dilakukan menggunakan pustaka Pandas (v1.5.3) dan NumPy (v1.24). Implementasi model *machine learning* dibangun menggunakan pustaka Scikit-learn (v1.3.0) untuk algoritma Random Forest dan metrik evaluasi, serta pustaka XGBoost (v1.7.6) untuk model *gradient boosting*. Visualisasi data dan *feature importance* dihasilkan menggunakan Matplotlib (v3.7) dan Seaborn (v0.12)

Periode penelitian dibagi menjadi dua rezim non-overlapping. Tahun 2020 sengaja dikecualikan untuk menghindari anomali transisi pandemi, sebagaimana ringkasan pada Tabel 1.

Tabel 1. Deskripsi Dataset per Rezim (Sumber: Data diolah Penulis, 2025)

Rezim	Deskripsi	Periode Mulai	Periode Selesai	Jumlah Observasi (Hari)
Rezim 1	Pra-Pandemi	2018-01-01	2019-12-31	~502
Rezim 2	Pasca-Pandemi	2021-01-01	2022-12-31	~503

Variabel Penelitian

Variabel target adalah variabel biner: return harian pada hari berikutnya ($t+1$) bernilai 1 jika positif (harga naik) dan 0 jika negatif (harga turun). Formulasi ini mengikuti pendekatan umum dalam literatur prediksi arah pasar (Henrique et al., 2019). Variabel independen mencakup harga, volume, indikator teknikal, makroekonomi, dan volatilitas pasar sebagaimana dirangkum pada Tabel 2.

Tabel 2. Variabel Penelitian (Sumber: Tinjauan literatur dan pengolahan Penulis, 2025)

Kategori	Nama Variabel	Deskripsi Singkat
Harga & Volume	Open, High, Low, Close, Volume	Data harga & volume harian
Indikator Teknikal	SMA_20/50, RSI_14, MACD, ATR_14	Tren, momentum, volatilitas
Makroekonomi	FedFundsRate, TreasuryYield10Y, CPI, DXY	Kondisi ekonomi makro
Volatilitas Pasar	VIX	Indeks ketakutan pasar

Pra-pemrosesan Data

Data dari berbagai sumber disinkronkan berdasarkan tanggal. Teknik *forward fill* diterapkan pada data makroekonomi untuk mengisi hari tanpa rilis data dengan nilai terakhir yang diketahui. Baris awal yang mengandung nilai NaN akibat perhitungan indikator teknikal dihapus. Dataset setiap rezim dipisahkan menjadi 80% data latih dan 20% data uji secara kronologis untuk mencegah *lookahead bias*.

Model Machine Learning

Dua model ensemble digunakan dalam penelitian ini:

1. **XGBoost (Extreme Gradient Boosting):** model berbasis boosting yang membangun pohon keputusan secara sekuensial untuk memperbaiki kesalahan dari pohon sebelumnya. XGBoost dikenal unggul dalam berbagai kompetisi prediksi finansial.

2. **Random Forest:** model berbasis *bagging* yang membangun pohon keputusan secara paralel dari sub-sampel data berbeda. Model ini digunakan sebagai pembandingan dan untuk uji robustnes.

Evaluasi Model

Kinerja model klasifikasi dievaluasi menggunakan metrik yang diturunkan dari *confusion matrix* (James et al., 2013):

- **Accuracy:** proporsi prediksi benar terhadap total data.
- **Precision:** tingkat akurasi prediksi positif.
- **Recall (Sensitivity):** kemampuan model menangkap semua kasus positif.
- **F1 Score:** rerata harmonik Precision dan Recall, dipilih sebagai metrik utama karena lebih seimbang untuk data pasar yang volatil.

III. HASIL

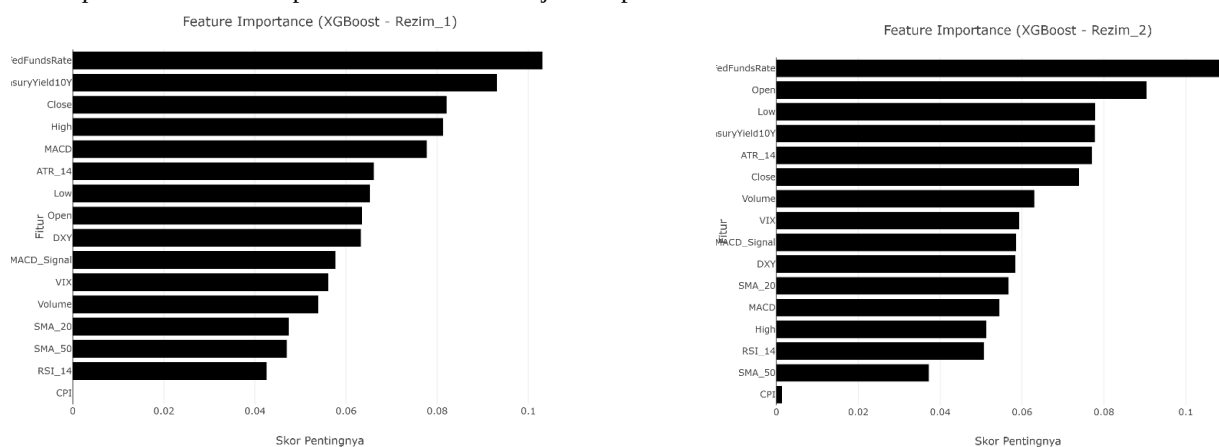
Kinerja model dievaluasi pada dua rezim pasar dengan metrik Accuracy dan F1 Score. Hasil ditunjukkan pada Tabel 3.

Tabel 3. Perbandingan Kinerja Model Klasifikasi pada Set Pengujian (Sumber: Hasil pengolahan data Penulis, 2025)

Model	Metrik	Rezim 1 (Pra-Pandemi)	Rezim 2 (Pasca-Pandemi)
XGBoost	Accuracy	0.418	0.473
	F1 Score	0.346	0.556
Random Forest	Accuracy	0.440	0.484
	F1 Score	0.164	0.544

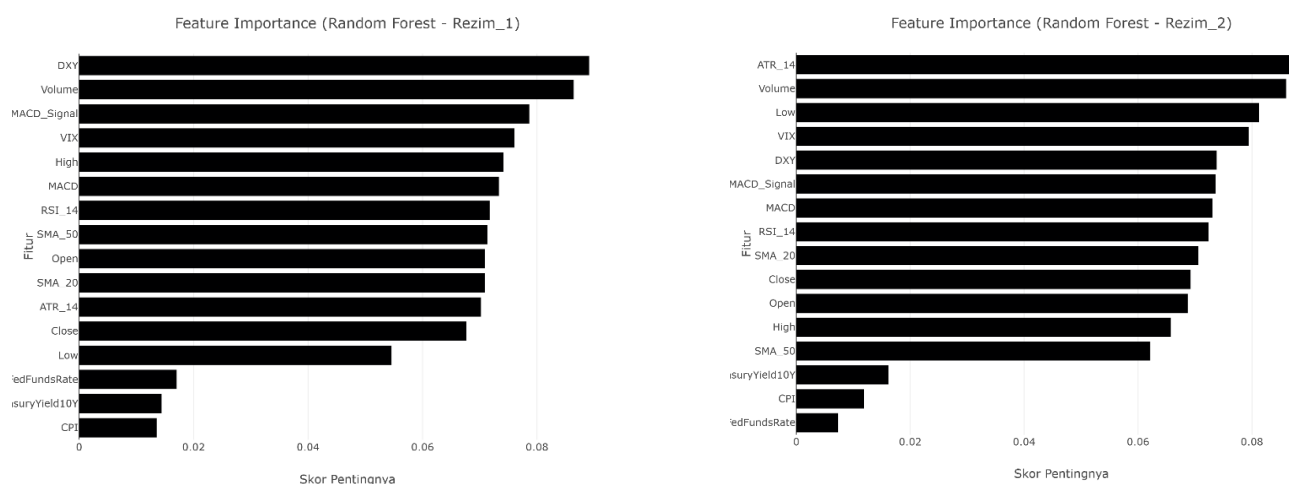
Hasil pada Tabel 3 menunjukkan adanya peningkatan kinerja prediktif di rezim pasca-pandemi. F1 Score XGBoost meningkat dari 0.346 menjadi 0.556, sementara Random Forest naik dari 0.164 menjadi 0.544. Hal ini mengindikasikan bahwa model machine learning lebih mampu menangkap pola pasar setelah pandemi. Hasil pada Tabel 3 menunjukkan adanya peningkatan kinerja prediktif yang signifikan di rezim pasca-pandemi. F1 Score XGBoost meningkat dari 0,346 menjadi 0,556, sementara Random Forest melonjak dari 0,164 menjadi 0,544. Rendahnya skor akurasi dan F1 pada periode pra-pandemi mengindikasikan tingkat kesalahan yang tinggi, di mana model cenderung mengalami *underfitting* atau kesulitan melakukan generalisasi akibat sinyal pasar yang lemah dan penuh *noise*. Sebaliknya, peningkatan performa yang konsisten pada data uji (*test set*) di periode pasca-pandemi menunjukkan bahwa model berhasil menangkap pola fundamental baru tanpa indikasi *overfitting* yang signifikan. Hal ini mengonfirmasi bahwa struktur pasar pasca-pandemi memiliki pola deterministik yang lebih kuat dibandingkan periode sebelumnya

Analisis *feature importance* dilakukan untuk mengidentifikasi pergeseran faktor dominan. Perbandingan hierarki *feature importance* XGBoost pada kedua rezim ditunjukkan pada Gambar 1.



Gambar 2. Perbandingan Feature Importance Model XGBoost pada Rezim Pra dan Pasca Pandemi Sumber: Hasil Pengolahan Data Penulis (2025)

Sebagai uji robustnes, analisis serupa dilakukan menggunakan Random Forest, sebagaimana ditampilkan pada Gambar 2.



Gambar 3. Perbandingan Feature Importance Model Random Forest pada Rezim Pra dan Pasca Pandemi Sumber: Hasil Pengolahan Data Penulis (2025)

Diskusi

Penelitian ini menunjukkan adanya pergeseran fundamental dalam faktor prediktif pasar saham teknologi akibat pandemi COVID-19, yakni transisi dari rezim pasar multi-faktor ke rezim yang didominasi oleh narasi makroekonomi tunggal.

Pertama, dominasi variabel makroekonomi pada periode pasca-pandemi memiliki landasan teoritis yang kuat. Pada periode pra-pandemi, *feature importance* tersebar merata karena pasar digerakkan oleh sentimen idiosinkratik perusahaan. Namun, pasca-pandemi, XGBoost menempatkan *FedFundsRate* sebagai fitur paling krusial. Hal ini sejalan dengan teori valuasi aset, di mana saham teknologi dikategorikan sebagai saham pertumbuhan (*growth stocks*) dengan durasi ekuitas yang panjang. Valuasi saham jenis ini sangat sensitif terhadap perubahan tingkat diskonto yang direpresentasikan oleh suku bunga The Fed. Oleh karena itu, kebijakan moneter menjadi sinyal deterministik utama yang "mematikan" sinyal-sinyal teknikal minor lainnya.

Kedua, peningkatan nilai F1 Score yang signifikan pada rezim pasca-pandemi (XGBoost: 0.556; Random Forest: 0.544) mengonfirmasi hipotesis (Thorbecke, 2020) bahwa krisis dapat meningkatkan prediktabilitas pasar. Meskipun volatilitas pasar meningkat, arah pergerakan harga menjadi lebih mudah dibaca oleh mesin karena didorong oleh satu faktor dominan (kebijakan suku bunga dan inflasi) dibandingkan periode normal yang penuh dengan *noise* acak.

Ketiga, perbedaan hasil antar model memberikan perspektif kausalitas yang saling melengkapi tanpa perlu justifikasi berulang. XGBoost berhasil menangkap "penyebab" pergerakan pasar melalui suku bunga (*FedFundsRate*), sedangkan Random Forest lebih efektif menangkap "dampak" psikologis pasar melalui indikator volatilitas (*ATR* dan *VIX*). Kombinasi ini memperlihatkan mekanisme pasar pasca-pandemi yang bekerja dalam siklus *Risk-On/Risk-Off* yang ketat, di mana setiap kebijakan bank sentral langsung direspon dengan lonjakan volatilitas yang terukur.

Secara praktis, temuan ini mengimplikasikan bahwa strategi investasi pasca-pandemi tidak lagi bisa mengandalkan analisis teknikal klasik semata. Investor perlu memprioritaskan pemantauan terhadap kalender ekonomi makro dan kebijakan The Fed sebagai input utama dalam model pengambilan keputusan mereka.

IV. KESIMPULAN

Penelitian ini membuktikan bahwa pandemi COVID-19 telah memicu perubahan struktural dalam faktor-faktor prediktif pasar saham teknologi Amerika Serikat. Berdasarkan hasil komparasi model XGBoost dan Random Forest antara periode pra-pandemi (2018–2019) dan pasca-pandemi (2021–2022), dapat ditarik tiga kesimpulan utama:

1. **Peningkatan Prediktabilitas Pasar:** Terjadi lonjakan kinerja prediktif yang signifikan pada rezim pasca-pandemi, di mana model mampu menangkap pola pasar dengan lebih akurat. Hal ini mengindikasikan bahwa pasar yang didominasi oleh narasi makroekonomi tunggal cenderung lebih deterministik dibandingkan pasar pra-pandemi yang dipenuhi *noise* idiosinkratik.

2. **Pergeseran Fundamental Fitur:** Telah terjadi transisi dominasi fitur dari indikator teknikal dan harga menuju faktor makroekonomi dan volatilitas. Secara spesifik, *FedFundsRate* menjadi prediktor utama, yang mengonfirmasi sensitivitas valuasi saham teknologi terhadap kebijakan suku bunga di era baru ini.
3. **Mekanisme Kausalitas:** Perbedaan preferensi fitur antar model memperlihatkan dinamika pasar yang utuh: XGBoost menangkap sinyal kebijakan moneter sebagai penyebab utama pergerakan, sementara Random Forest menangkap respons volatilitas pasar sebagai dampak langsungnya.

Implikasi praktis dari temuan ini menuntut investor untuk mengevaluasi ulang strategi yang hanya mengandalkan analisis teknikal historis. Dalam rezim pasar saat ini, pemantauan terhadap kebijakan The Fed dan indikator volatilitas (VIX) menjadi instrumen navigasi yang jauh lebih krusial daripada pola grafik harga masa lalu.

Keterbatasan dan Saran

Penelitian ini memiliki keterbatasan pada pemilihan fitur, yang belum mencakup data sentimen berbasis teks atau indikator global lain. Selain itu, lingkup penelitian hanya terbatas pada ETF QQQ. Studi selanjutnya dapat memperluas cakupan ke sektor lain atau pasar internasional, serta menguji model deep learning seperti LSTM dengan analisis interpretatif berbasis SHAP (Bussmann et al., 2021).

V. DAFTAR PUSTAKA

- Baker, S. R., Bloom, N., Davis, S. J., Kost, K., Sammon, M., & Viratyosin, T. (2020). The unprecedented stock market impact of COVID-19. *The Review of Asset Pricing Studies*, 10(4), 742–758. <https://doi.org/10.1093/rapstu/raaa008>
- Bussmann, N., Giudici, P., Marinelli, D., & Papenbrock, J. (2021). Explainable machine learning in credit risk management. *Computational Economics*, 57(1), 203–228. <https://doi.org/10.1007/s10614-020-10042-0>
- Fischer, T., & Krauss, C. (2018). Deep learning with long short-term memory networks for financial market predictions. *European Journal of Operational Research*, 270(2), 654–669. <https://doi.org/10.1016/j.ejor.2017.11.054>
- Gu, S., Kelly, B., & Xiu, D. (2020). Empirical asset pricing via machine learning. *The Review of Financial Studies*, 33(5), 2223–2273. <https://doi.org/10.1093/rfs/hhz137>
- Henrique, B. M., Sobreiro, V. A., & Kimura, H. (2019). Literature review: Machine learning techniques applied to financial market prediction. *Expert Systems with Applications*, 124, 226–251. <https://doi.org/10.1016/j.eswa.2019.01.012>
- James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). *An introduction to statistical learning: With applications in R*. Springer. <https://doi.org/10.1007/978-1-4614-7138-7>
- Murphy, J. J. (1999). *Technical analysis of the financial markets: A comprehensive guide to trading methods and applications*. New York Institute of Finance.
- Shiller, R. J. (2003). From efficient markets theory to behavioral finance. *Journal of Economic Perspectives*, 17(1), 83–104. <https://doi.org/10.1257/089533003321164967>
- Thorbecke, W. (2020). The impact of the COVID-19 pandemic on the U.S. economy: Evidence from the stock market. *Journal of Risk and Financial Management*, 13(10), 233. <https://doi.org/10.3390/jrfm13100233>
- Tsay, R. S. (2005). *Analysis of financial time series* (2nd ed.). John Wiley & Sons.