

TRƯỜNG ĐẠI HỌC SÀI GÒN
KHOA CÔNG NGHỆ THÔNG TIN



BÁO CÁO
MÔN PHÂN TÍCH DỮ LIỆU

TÊN ĐỀ TÀI
PHÂN TÍCH KHÁM PHÁ DỮ LIỆU

Giảng viên hướng dẫn:

TS. Đỗ Như Tài

Danh sách thành viên:

Nguyễn Trương Hiệp - 3122410110

Nguyễn Văn Minh - 3122410242

Vũ Thị Thanh Ngân - 3122410255

Trương Xuân Hưng - 3122410161

Thành phố Hồ Chí Minh, ngày 14 tháng 10 năm 2025

MỤC LỤC

MỤC LỤC

TÓM TẮT

GIỚI THIỆU

1

1.1. Mục tiêu dự án

1

1.2. Tài liệu và yêu cầu ban đầu

1

1.3. Hình ảnh sản phẩm hoàn thành

1

PHƯƠNG PHÁP THỰC HIỆN VÀ CÁC HIỆU CHỈNH KỸ THUẬT

3

2.1. Phân tích và Lựa chọn Thư viện Thu thập Dữ liệu

3

2.2. Thiết lập Nguồn dữ liệu VN-30

3

2.3. Phân tích và Tái cấu trúc Chức năng Dashboard

4

Tab 1: Tổng quan (Summary)

4

Tab 2: Biểu đồ Kỹ thuật (Chart)

5

Tab 3: Statistics (Thống kê) -> Cổ tức & Chia tách (Dividends & Splits)

6

Tab 4: Báo cáo Tài chính (Financials)

7

Tab 5: Analysis (Phân tích) -> Phân tích & Cổ đông (Holders & Recommendations)

7

Tab 6: Mô phỏng Monte Carlo (Monte Carlo Simulation)

8

Tab 7: Danh mục Đầu tư (Your Portfolio's Trend)

9

2.4. Việt hóa và Cải thiện Giao diện Người dùng (UI/UX)

9

KẾT QUẢ ĐẠT ĐƯỢC

11

KẾT LUẬN

12

TÓM TẮT

Dự án này trình bày quá trình phân tích, hiệu chỉnh và phát triển một ứng dụng Financial Dashboard tương tác, chuyển đổi từ một hệ thống mẫu phân tích S&P 500 sang áp dụng cho chỉ số VN-INDEX 30 của Việt Nam. Trọng tâm của báo cáo là các thách thức kỹ thuật gặp phải và các giải pháp được áp dụng. Vấn đề cốt lõi được xác định là sự không tương thích của thư viện **yahoo_fin** với cấu trúc dữ liệu và tính sẵn có của API cho thị trường Việt Nam.

Giải pháp chính là thống nhất sử dụng thư viện **yfinance** làm nguồn truy cập dữ liệu duy nhất, tận dụng API hướng đối tượng để tăng tính ổn định. Quá trình tái cấu trúc bao gồm việc loại bỏ các tab chức năng không khả thi (Statistics, Analysis) do thiếu hụt dữ liệu đầu vào. Đồng thời, dự án đã phát triển các tính năng thay thế có giá trị tương đương (Cổ tức & Chia tách, Phân tích & Cổ đông) dựa trên các nguồn dữ liệu đáng tin cậy.

Ngoài ra, dự án đã thực hiện các cải tiến về phương pháp luận phân tích (mở rộng mẫu thời gian cho Monte Carlo, chuẩn hóa giá cho danh mục) và hoàn thiện sản phẩm bằng cách Việt hóa toàn bộ giao diện người dùng. Kết quả là một ứng dụng dashboard hoàn chỉnh, ổn định, thân thiện với người dùng Việt Nam và phù hợp với thực tiễn phân tích dữ liệu VN-30.

GIỚI THIỆU

1.1. Mục tiêu dự án

Mục tiêu của dự án là xây dựng một ứng dụng Financial Dashboard tương tác bằng Streamlit. Ứng dụng này tập trung vào việc phân tích và trực quan hóa dữ liệu tài chính của các công ty trong rổ chỉ số VN-INDEX 30 của Việt Nam. Mục tiêu không chỉ dừng lại ở việc hiển thị dữ liệu thô, mà còn nhằm cung cấp một công cụ tổng hợp, mang lại cái nhìn trực quan, đa chiều về sức khỏe tài chính và hiệu suất thị trường của các doanh nghiệp hàng đầu. Trong bối cảnh thông tin tài chính thường bị phân mảnh trên nhiều nền tảng, một dashboard "tất cả trong một" như vậy sẽ hỗ trợ người dùng trong việc tra cứu, so sánh và ra quyết định một cách hiệu quả và nhanh chóng.

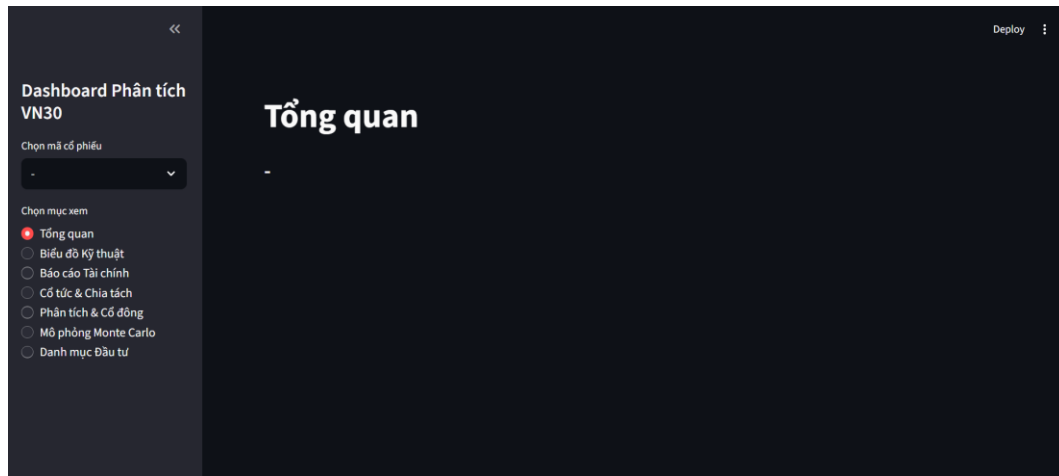
1.2. Tài liệu và yêu cầu ban đầu

Dự án được thực hiện dựa trên một ứng dụng Dashboard mẫu (**findash_app.py**), vốn được thiết kế để phân tích dữ liệu từ thị trường S&P 500 (Mỹ). Yêu cầu cốt lõi được đặt ra là "Việt hóa" ứng dụng này. Đây không chỉ là một bài toán thay đổi nguồn dữ liệu đơn thuần từ S&P 500 sang VN-30, mà còn là một thách thức kỹ thuật thực sự. Thách thức này đòi hỏi việc phân tích sâu về sự tương thích của các thư viện và API đang được sử dụng.

Việc chuyển đổi này không chỉ là thay đổi danh sách mã cổ phiếu. Nó bao hàm việc kiểm tra xem các điểm cuối API (API endpoints) mà thư viện **yahoo_fin** sử dụng có trả về dữ liệu cho các mã **.VN** hay không. Do đó, ứng dụng phải được hiệu chỉnh, thêm hoặc bớt các tính năng để đảm bảo hoạt động ổn định, phù hợp với đặc thù và tính sẵn có (availability) của dữ liệu tài chính tại thị trường Việt Nam.

1.3. Hình ảnh sản phẩm hoàn thành

Dưới đây là hình ảnh giao diện chính của ứng dụng Financial Dashboard VN-30 đã hoàn thành (Tab Tổng quan), minh họa cho kết quả của quá trình chuyển đổi và phát triển:



Hình 1.3: Giao diện chính của ứng dụng Financial Dashboard VN-30

PHƯƠNG PHÁP THỰC HIỆN VÀ CÁC HIỆU CHỈNH KỸ THUẬT

Quá trình hiệu chỉnh ứng dụng gốc đề ra được sản phẩm cuối cùng bao gồm nhiều thay đổi quan trọng về thư viện, nguồn dữ liệu, logic tính năng và giao diện người dùng.

2.1. Phân tích và Lựa chọn Thư viện Thu thập Dữ liệu

Thay đổi mang tính nền tảng và quan trọng nhất của dự án là việc thống nhất thư viện truy cập dữ liệu. Phiên bản gốc sử dụng đồng thời hai thư viện: **yfinance** (viết tắt là **yf**) để lấy dữ liệu giá lịch sử, và **yahoo_fin.stock_info** (viết tắt là **si**) để lấy các dữ liệu thống kê, tài chính và phân tích.

Trong quá trình thử nghiệm ban đầu, thư viện **yahoo_fin** bộc lộ tính không tương thích rõ rệt với các mã cổ phiếu Việt Nam (có hậu tố **.VN**). Cụ thể, các hàm quan trọng được thiết kế chủ yếu cho thị trường Mỹ như **si.get_stats**, **si.get_stats_valuation**, và **si.get_analysts_info** hoạt động không ổn định. Khi được gọi với các mã như 'FPT.VN' hay 'VCB.VN', các hàm này thường xuyên trả về giá trị **None** hoặc DataFrame rỗng.

Việc này dẫn đến hai hệ lụy nghiêm trọng:

1. **Về mặt chức năng:** Các Tab "Statistics" và "Analysis" hoàn toàn không có dữ liệu để hiển thị.
2. **Về mặt kỹ thuật:** Mã nguồn trở nên phức tạp do phải xử lý lỗi (exception handling) cho các giá trị **NoneType** để tránh làm sập đổ (crash) toàn bộ ứng dụng.

Do đó, giải pháp chiến lược được đưa ra là **loại bỏ hoàn toàn thư viện yahoo_fin** và chỉ sử dụng duy nhất **yfinance**. Quyết định này mang lại hai lợi ích lớn:

1. **Tính ổn định:** **yfinance** chứng tỏ khả năng truy xuất dữ liệu cho mã **.VN** tốt hơn và đáng tin cậy hơn.
2. **Tính nhất quán:** **yfinance** cung cấp một API hướng đối tượng (Object-Oriented) mạnh mẽ hơn thông qua **yf.Ticker()**. Bằng cách tạo một đối tượng **stock = yf.Ticker(ticker)**, chúng ta có thể truy cập hầu hết mọi thông tin cần thiết (như **.info**, **.financials**, **.dividends**) thông qua các thuộc tính và phương thức đã được chuẩn hóa.

2.2. Thiết lập Nguồn dữ liệu VN-30

Nguồn dữ liệu mã cổ phiếu (tickers) đã được cập nhật để phản ánh đúng thị trường mục tiêu. Thay vì tự động tải danh sách 500 mã S&P 500 bằng hàm `si.tickers_sp500()` như code gốc, phiên bản mới sử dụng một danh sách Python cố định (list) tên là **VN30_TICKERS**.

Danh sách này chứa 30 mã cổ phiếu VN-30 đã được chuẩn hóa với hậu tố **.VN** (ví dụ: **'ACB.VN'**, **'FPT.VN'**). Lý do của thay đổi này là vì không có hàm API nào cung cấp sẵn danh sách VN-30 cập nhật một cách đáng tin cậy. Việc tạo danh sách cố định đảm bảo tính chính xác tuyệt đối và đúng phạm vi yêu cầu của dự án. Danh sách này được sử dụng cho cả Sidebar chọn mã và Tab "Danh mục Đầu tư", đồng thời giúp ứng dụng khởi động nhanh hơn.

2.3. Phân tích và Tái cấu trúc Chức năng Dashboard

Việc chuyển đổi thư viện và nguồn dữ liệu dẫn đến hàng loạt hiệu chỉnh logic trên các Tab của ứng dụng:

Tab 1: Tổng quan (Summary)

Tab 'Tổng quan' được thiết kế để cung cấp cho người dùng một cái nhìn nhanh và toàn diện về mã cổ phiếu được chọn. Chức năng của tab này là hiển thị các chỉ số tài chính quan trọng (như vốn hóa, P/E, khối lượng giao dịch) và một biểu đồ lịch sử giá tổng thể để nắm bắt xu hướng dài hạn.

Về mặt kỹ thuật, code gốc vốn sử dụng `si.get_quote_table()`. Do hàm này không hoạt động với mã **.VN**, chúng tôi đã chuyển sang sử dụng đối tượng `yf.Ticker(ticker).info`. Đây là một thay đổi lớn về logic, vì `.info` trả về một từ điển (dictionary) lớn. Một hàm mới là `get_summary_data(ticker)` đã được xây dựng để trích xuất thủ công các thông tin quan trọng. Các tên chỉ số cũng được Việt hóa (ví dụ: 'previousClose' thành 'Giá đóng cửa phiên trước'). Đồng thời, một hàm định dạng (`_format_val`) cũng được bổ sung để xử lý các kiểu dữ liệu khác nhau và trình bày chúng dưới dạng văn bản có định dạng rõ ràng.



Hình 2.3.a: Phần tổng quan (dữ liệu của ABC.VN)



Hình 2.3.b: Phần biểu đồ tổng quan (dữ liệu của ABC.VN)

Tab 2: Biểu đồ Kỹ thuật (Chart)

Tab 'Biểu đồ Kỹ thuật' là công cụ phân tích chuyên sâu, cho phép người dùng tùy chỉnh và tương tác với dữ liệu giá. Chức năng của tab này là cung cấp các công cụ trực quan hóa chi tiết như biểu đồ Nến (Candlestick) hoặc Đường (Line), kết hợp với đường Trung bình Động (SMA) và Khối lượng giao dịch (Volume), với khả năng tùy biến khung thời gian linh hoạt.

Chức năng này vốn đã sử dụng thư viện **yfinance** (**yf.download()**) để lấy dữ liệu giá, do đó nó hoàn toàn tương thích với các mã VN-30. Không giống như các Tab 3 và 5 cần loại bỏ, logic cốt lõi của Tab 2 được giữ lại. Quá trình hiệu chỉnh tập trung chủ yếu vào việc cải thiện trải nghiệm người dùng và Việt hóa, như tinh chỉnh hàm **getchartdata**

để đảm bảo việc gộp dữ liệu SMA 50 ngày được thực hiện một cách an toàn và chính xác, sử dụng `pd.merge_asof` để căn chỉnh chuỗi thời gian.



Hình 2.3.c: Phần biểu đồ kỹ thuật (dữ liệu của ABC.VN)

Tab 3: Statistics (Thống kê) -> Cổ tức & Chia tách (Dividends & Splits)

Một thay đổi lớn trong cấu trúc ứng dụng là việc **loại bỏ hoàn toàn Tab "Statistics"** (Tab 3) của bản gốc. Lý do là các hàm `si.get_stats` và `si.get_stats_valuation` mà tab này phụ thuộc hoàn toàn thất bại khi lấy dữ liệu VN-30.

Để thay thế, một Tab 3 mới với tên gọi "Cổ tức & Chia tách" đã được phát triển. Chức năng của tab này là cung cấp thông tin chi tiết về các hành động doanh nghiệp (corporate actions) trong quá khứ, cho phép nhà đầu tư theo dõi lịch sử nhận cổ tức và các lần chia tách cổ phiếu. Tab này sử dụng các hàm `yf.Ticker(ticker).dividends` và `yf.Ticker(ticker).splits` của `yfinance`.

The screenshot shows the 'Cổ tức & Chia tách' (Dividends & Splits) tab for ACB.VN. The sidebar is the same as in the previous image, but the 'Cổ tức & Chia tách' option is now selected. The main area has the title 'Cổ tức & Chia tách' and 'ACB.VN'. Below this are two sections: 'Lịch sử Chi trả Cổ tức' (Dividend History) and 'Lịch sử Chia tách Cổ phiếu' (Stock Split History). Each section contains a table with columns for 'Date' and the respective action (Dividends or Stock Splits).

Lịch sử Chi trả Cổ tức	
Date	Dividends
2025-05-23 00:00:00+07:00	1000
2024-05-31 00:00:00+07:00	1000

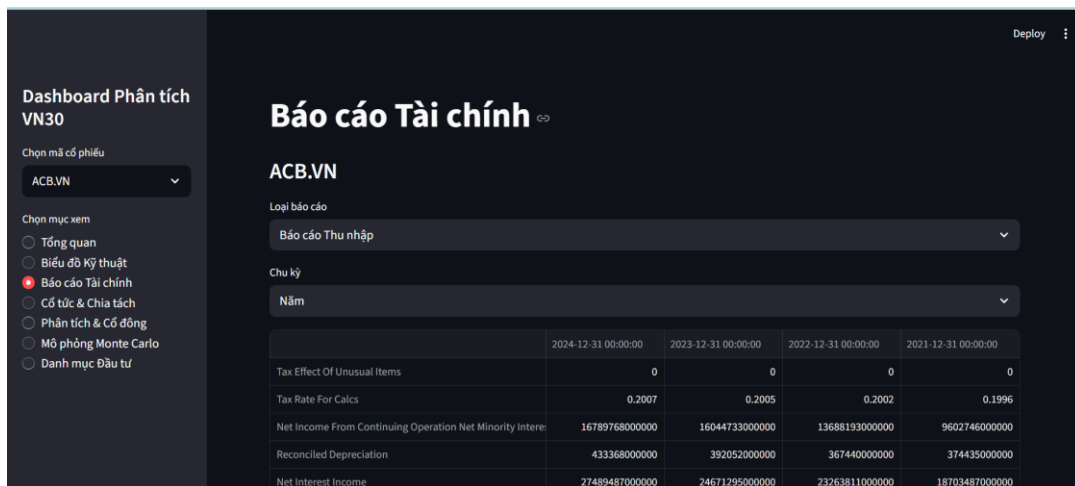
Lịch sử Chia tách Cổ phiếu	
Date	Stock Splits
2025-05-23 00:00:00+07:00	1.15
2024-05-31 00:00:00+07:00	1.15

Hình 2.3.d: Phần Cổ tức & Chia tách (dữ liệu của ABC.VN)

Tab 4: Báo cáo Tài chính (Financials)

Tab 'Báo cáo Tài chính' đóng vai trò cung cấp dữ liệu nền tảng cho việc phân tích cơ bản. Chức năng của tab là cho phép người dùng truy cập và xem chi tiết ba báo cáo tài chính cốt lõi: Báo cáo Thu nhập (Income Statement), Bảng Cân đối Kế toán (Balance Sheet), và Báo cáo Lưu chuyển Tiền tệ (Cash Flow), với tùy chọn xem theo chu kỳ Năm hoặc Quý.

Về mặt kỹ thuật, các hàm cũ của **yahoo_fin** (như **si.get_income_statement()**) đã được thay thế toàn bộ bằng các thuộc tính tương ứng từ đối tượng **yf.Ticker (.financials, .balance_sheet, .cashflow)**, củng cố cho quyết định thống nhất sử dụng **yfinance**.



	2024-12-31 00:00:00	2023-12-31 00:00:00	2022-12-31 00:00:00	2021-12-31 00:00:00
Tax Effect Of Unusual Items	0	0	0	0
Tax Rate For Calcs	0.2007	0.2005	0.2002	0.1996
Net Income From Continuing Operation Net Minority Intere	16789768000000	16044733000000	13688193000000	9602746000000
Reconciled Depreciation	433368000000	392052000000	367440000000	374435000000
Net Interest Income	27489487000000	24671295000000	23263811000000	18703487000000

Hình 2.3.e: Phần báo cáo tài chính (dữ liệu của ABC.VN)

Tab 5: Analysis (Phân tích) -> Phân tích & Cổ đông (Holders & Recommendations)

Tương tự Tab 3, Tab "Analysis" (Tab 5) cũng bị loại bỏ do hàm **si.get_analysts_info()** của **yahoo_fin** không trả về dữ liệu. Các thông tin chi tiết như dự báo doanh thu (Earnings Estimate) là đặc thù của thị trường Mỹ.

Thay vào đó, một tab mới mang tên "Phân tích & Cổ đông" được xây dựng. Chức năng của tab này là cung cấp cái nhìn về các bên liên quan và các đánh giá bên ngoài về cổ phiếu, bao gồm các khuyến nghị từ nhà phân tích chuyên nghiệp và cơ cấu sở hữu của các quỹ đầu tư, tổ chức lớn. Tab này khai thác các dữ liệu có sẵn từ **yfinance** như **stock.recommendations**, **stock.institutional_holders**, và **stock.mutualfund_holders**.

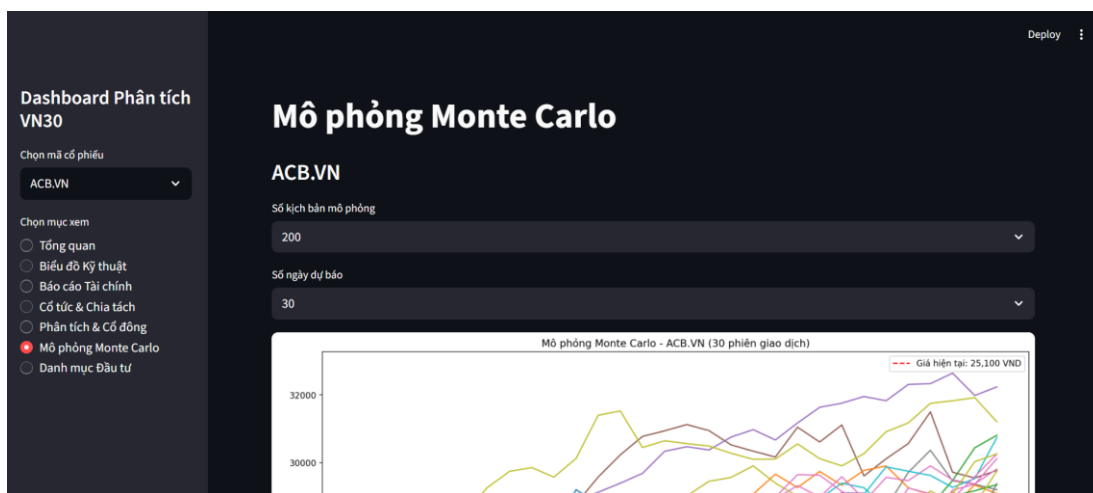


Hình 2.3.f: Phân Phân tích & Cổ động (dữ liệu của ABC.VN)

Tab 6: Mô phỏng Monte Carlo (Monte Carlo Simulation)

Tab 'Mô phỏng Monte Carlo' cung cấp một công cụ phân tích dự báo và quản trị rủi ro. Chức năng của nó là chạy hàng trăm kịch bản giá ngẫu nhiên dựa trên độ biến động lịch sử, từ đó ước tính phạm vi giá có thể xảy ra trong tương lai và tính toán Mức lỗ Tối đa (Value at Risk - VaR).

Đối với tab này, một cải tiến quan trọng về phương pháp luận đã được thực hiện. Code gốc chỉ sử dụng 30 ngày dữ liệu để tính toán độ biến động. Ở phiên bản mới, hàm **yf.download()** được sử dụng để lấy dữ liệu của 365 ngày gần nhất, giúp mô phỏng bao gồm nhiều chu kỳ thị trường hơn, cho ra một độ biến động thực tế hơn, từ đó nâng cao độ tin cậy của mô phỏng và giá trị VaR.



Hình 2.3.g: Phân mô phỏng Monte Carlo (dữ liệu của ABC.VN)

Tab 7: Danh mục Đầu tư (Your Portfolio's Trend)

Tab 'Danh mục Đầu tư' cho phép người dùng so sánh hiệu suất của nhiều mã cổ phiếu cùng lúc. Chức năng chính là trực quan hóa sự tăng trưởng tương đối của các khoản đầu tư khác nhau theo thời gian, giúp người dùng đánh giá hiệu quả của việc phân bổ danh mục.

Code gốc chỉ vẽ biểu đồ giá tuyệt đối (absolute price). Phiên bản mới đã thêm một tính năng quan trọng là **chuẩn hóa (Normalization)** dữ liệu. Bằng cách tạo một DataFrame mới (**normalized_df = (df / df.iloc[0])**), tất cả cổ phiếu được đưa về cùng một mốc 1.0 tại thời điểm bắt đầu. Tính năng chuẩn hóa này cho phép người dùng so sánh trực quan *hiệu suất tăng trưởng (%)* của các mã theo thời gian.



Hình 2.3.h: Phần Danh mục đầu tư (dữ liệu của ABC.VN)

2.4. Việt hóa và Cải thiện Giao diện Người dùng (UI/UX)

Một phần quan trọng trong việc "phát triển thêm" ứng dụng là quá trình bản địa hóa toàn bộ giao diện. Để sản phẩm thực sự "phù hợp với dữ liệu tài chính của Việt Nam", việc giao tiếp với người dùng bằng tiếng Việt là yêu cầu tất yếu.

Quá trình này bao gồm:

Việt hóa Tên Tab: Tất cả các tab điều hướng đã được đổi tên (ví dụ: 'Summary' thành 'Tổng quan', 'Chart' thành 'Biểu đồ Kỹ thuật').

Việt hóa Nhãn và Điều khiển: Toàn bộ các nhãn (**label**) của các widget Streamlit như **st.selectbox**, **st.radio**, **st.date_input** đều được dịch sang tiếng Việt (ví dụ: 'Select a ticker' thành 'Chọn mã cổ phiếu', 'Period' thành 'Chu kỳ').

Việt hóa Nội dung Biểu đồ: Các tiêu đề (**title**) và nhãn trục (**xlabel**, **ylabel**) của biểu đồ Plotly và Matplotlib đều được hiển thị bằng tiếng Việt (ví dụ: 'Giá (VND)', 'Phân phối giá dự báo cuối kỳ').

Việt hóa Thông báo: Các thông báo lỗi (**st.error**), cảnh báo (**st.warning**) và thông tin (**st.info**) đều được viết lại để cung cấp hướng dẫn rõ ràng bằng tiếng Việt.

Những thay đổi này, tuy không ảnh hưởng đến logic xử lý dữ liệu, nhưng có vai trò then chốt trong việc nâng cao trải nghiệm người dùng (UX) và khiến ứng dụng trở nên thân thiện, chuyên nghiệp hơn.

KẾT QUẢ ĐẠT ĐƯỢC

Sản phẩm **findash_app.py** hoàn thiện đã đạt được đầy đủ các mục tiêu đề ra. Dashboard đã chuyển đổi thành công và hoạt động ổn định với 30 mã cổ phiếu VN-INDEX 30. Đây không chỉ là một bản sao chép đơn thuần, mà là một sản phẩm đã được tái cấu trúc (refactored) và cải tiến (enhanced) để phù hợp với môi trường dữ liệu mới.

Việc hợp nhất về thư viện **yfinance** đã loại bỏ các lỗi API tiềm ẩn, mang lại tính ổn định cao. Các tab không tương thích đã được phân tích và thay thế bằng các tính năng phù hợp hơn (như Cổ tức & Chia tách, Phân tích & Cổ đông). Các cải tiến về logic phân tích ở Tab Monte Carlo và Tab Danh mục Đầu tư cũng đã tăng cường đáng kể giá trị sử dụng của công cụ so với phiên bản gốc. Cuối cùng, việc Việt hóa toàn diện giao diện đã hoàn thiện sản phẩm, giúp nó trở nên trực quan và dễ tiếp cận hơn với người dùng Việt Nam.

KẾT LUẬN

Dự án đã thành công trong việc chuyển đổi một dashboard tài chính từ thị trường Mỹ sang Việt Nam. Quá trình này không chỉ đòi hỏi thay đổi nguồn dữ liệu mà còn phải phân tích, đánh giá và thay thế các thư viện, logic nghiệp vụ không còn phù hợp.

Dự án này đã chứng minh tầm quan trọng của việc hiểu rõ giới hạn của công cụ (thư viện) và đặc thù của nguồn dữ liệu (thị trường Việt Nam so với Mỹ). Thông qua việc giải quyết các thách thức về tính tương thích API, cải tiến các phương pháp phân tích và bản địa hóa giao diện, kết quả thu được là một công cụ mạnh mẽ, ổn định và có giá trị phân tích thực tế cho dữ liệu VN-30.