**Lab 2**

* **Làm rõ quy trình 7 bước**: Từ Business Understanding đến sử dụng công cụ BI, tập trung vào phân tích sâu, mô hình hóa nâng cao, và trực quan hóa phức tạp.
* **Mỗi nhóm hoàn thiện**: Phân tích chi tiết dữ liệu, triển khai và so sánh 7 mô hình học máy cơ bản (Linear Regression, Decision Tree Regressor, KNN, Ridge, Lasso, Random Forest, SVR), thực hiện kiểm định thống kê (t-test, ANOVA, Wilcoxon), và xây dựng dashboard tương tác.
* **Đánh giá và tối ưu hóa**: So sánh hiệu năng mô hình bằng các chỉ số (MAE, RMSE, R²), kiểm định thống kê, và trực quan hóa kết quả.

**Thông tin nhóm & đề tài**

| **Nhóm** | **Mã SV** | **Họ tên thành viên** | **Đề tài** | **Mục tiêu cụ thể cho Lab 2** |
| --- | --- | --- | --- | --- |
| 5 | SE181933, SE190189, SE190186, SE180683 | Lâm Nguyễn Minh Thanh, Đinh Đại Lộc, Trần Phi Học, Nguyễn Cao Trị | Stock Market | 1. Phân tích ảnh hưởng earnings tới giá; 2. SQL phân cụm phản ứng giá; 3. Chênh lệch giá 3 ngày trước/sau earnings; 4. 7 mô hình dự báo biến động; 5. Line, bar, area, scatter, histogram; 6. Dashboard Power BI. |

**Quy trình 7 bước chi tiết cho Lab 2**

**Bước 1: Business Understanding & Analytic Approach**

**Yêu cầu**:

* Tinh chỉnh bài toán từ Lab 1, xác định mục tiêu cụ thể và người dùng cuối.
* Xem xét lại giả thuyết từ Lab 1 và điều chỉnh dựa trên kết quả.
* Xác định các bên liên quan (stakeholders).
* Định nghĩa lại KPIs chi tiết (RMSE, MAE, R²).
* Đặt các câu hỏi nghiên cứu mở rộng và lập kế hoạch phân tích (so sánh mô hình, kiểm định thống kê, tối ưu dashboard).

**Câu hỏi cho từng nhóm**:

**Nhóm 5 (Stock Market)**:

1. Ai là người dùng cuối của phân tích tác động earnings, và họ cần gì từ kết quả?
2. Kết quả Lab 1 cho thấy giả thuyết nào về tác động earnings cần được xem xét lại?
3. KPIs (RMSE) ở Lab 1 có đạt được không, và mục tiêu mới là gì?
4. Tại sao việc phân tích chênh lệch giá trước/sau earnings lại quan trọng?
5. Câu hỏi nghiên cứu chính của nhóm về tác động earnings là gì?
6. Những yếu tố nào (ngoài earnings) có thể ảnh hưởng đến biến động giá cổ phiếu?
7. Biểu đồ line và bar sẽ giúp người dùng hiểu gì về dữ liệu earnings?
8. Kiểm định thống kê sẽ giúp nhóm rút ra kết luận gì về hiệu năng mô hình?
9. Dashboard Power BI sẽ hỗ trợ nhà đầu tư như thế nào?
10. Những thách thức nào có thể gặp phải khi phân tích tác động earnings?

**Bước 2: Data Collection, Understanding & Preparation**

**Yêu cầu**:

* Mở rộng dữ liệu từ Lab 1 (tối thiểu 3 tháng, nhiều mã cổ phiếu, nhiều địa điểm để so sánh), xử lý chi tiết hơn.
* Thu thập dữ liệu bổ sung từ các nguồn (Yahoo Finance, Alpha Vantage, NOAA, OpenWeather).
* Mô tả chi tiết cấu trúc dữ liệu: số bản ghi, biến, kiểu dữ liệu, chất lượng.
* Tiền xử lý: xử lý missing values (interpolation, KNN imputation), chuẩn hóa/scale dữ liệu, phát hiện và xử lý outliers (IQR, Z-score), feature engineering (rolling mean, volatility, % thay đổi giá).
* Lưu dữ liệu vào cơ sở dữ liệu (SQLite, MySQL) và kiểm tra tính toàn vẹn.

**Câu hỏi cho từng nhóm**:

**Nhóm 5**:

1. Dữ liệu 1 tháng bao gồm những biến nào (bao gồm earnings), và nguồn từ đâu?

* Nguồn: **Yahoo Finance API** (qua yfinance).
* Biến trong DATA\_price.csv: Date, Close, High, Low, Open, Volume.
* Biến trong DATA\_earnings.csv: Earnings Date, EPS Estimate, Reported EPS.
* Biến trong DATA\_sentiment.csv: Earnings Date, compound (tạo giả định).

1. Số bản ghi và kiểu dữ liệu của từng biến trong tập dữ liệu là gì?

Ví dụ với AAPL trong khoảng 1 tháng (tùy tháng):

* DATA\_price.csv:
  + Khoảng 20–22 dòng (ngày giao dịch).
  + Kiểu: Date (datetime), các cột còn lại là float64 hoặc int64.
* DATA\_earnings.csv:
  + ~1–2 bản ghi/tháng.
* DATA\_sentiment.csv:
  + Giống earnings, chỉ khác ở compound.

1. Nhóm đã phát hiện bao nhiêu giá trị thiếu, và cách xử lý là gì?

 **Missing values**:

* Date có thể bị lỗi khi chuyển đổi → dropna(subset=["Date"]).
* Trong giá cổ phiếu: điền bằng interpolation + KNNImputer.

 **Cách xử lý**:

* .interpolate() → lấp khoảng trống tuyến tính.
* KNNImputer(n\_neighbors=3) → dùng lân cận để ước lượng.

1. Phương pháp nào được sử dụng để phát hiện outliers trong giá cổ phiếu?

**Hai bước kết hợp**:

* **Z-score**: loại bỏ giá trị > 3 std dev khỏi trung bình.
* **IQR (Interquartile Range)**: loại bỏ các điểm ngoài 1.5 \* IQR so với Q1/Q3.

1. Đặc trưng mới nào (chênh lệch giá trước/sau earnings,...) đã được tạo, và ý nghĩa là gì?

* price\_change\_pct: phần trăm thay đổi giá trước/sau earnings (1 ngày).
* eps\_result: Beat, Miss, hoặc Meet – phân loại dựa trên so sánh EPS thực tế và ước tính.
* compound: sentiment score (giả định, từ -1 đến 1).
* rolling\_mean\_5, rolling\_std\_5: biến động 5 ngày.
* volatility: chênh lệch High - Low trong ngày.
* Ý nghĩa:
  + Các đặc trưng này giúp mô hình học máy phân tích mối quan hệ giữa tâm lý thị trường, kết quả tài chính, và biến động giá.

1. Dữ liệu đã được chuẩn hóa/scale như thế nào, và tại sao cần thiết?

 **Chuẩn hóa bằng MinMaxScaler** về khoảng [0,1].

 Lý do:

* Các thuật toán như KNN, Linear Regression, Random Forest hoạt động tốt hơn khi dữ liệu trên cùng một thang đo.

1. Nhóm đã lưu dữ liệu vào cơ sở dữ liệu nào, và kiểm tra tính toàn vẹn ra sao?

* **CSDL**: SQLite, file: multi\_stock\_analysis.db.
* **Tính toàn vẹn**:
  + Chuyển toàn bộ định dạng thời gian → %Y-%m-%d.
  + Chỉ insert dữ liệu nếu các trường quan trọng (Date, Earnings Date) hợp lệ.
  + Truy vấn EPS chỉ lấy dòng có close\_before và close\_after.

1. Những thách thức nào gặp phải khi thu thập dữ liệu earnings?

* Một số mã cổ phiếu không có earnings\_dates đầy đủ từ Yahoo.
* Thời gian earnings báo cáo không đều (không theo tháng).
* Phải xử lý timezone và định dạng Earnings Date nhất quán.

1. Dữ liệu có bao nhiêu báo cáo earnings, và điều này ảnh hưởng đến phân tích thế nào?

* Khoảng 4 báo cáo earnings mỗi năm cho mỗi cổ phiếu (~1 mỗi quý).
* Nếu < 10 dòng được merge giữa EPS và sentiment → **bỏ qua bước ML** để tránh lỗi chia nhỏ dữ liệu train/test.
* Do đó, ảnh hưởng lớn đến độ tin cậy của mô hình nếu số lượng điểm nhỏ.

1. Nhóm đã xử lý thế nào nếu dữ liệu có giá trị không hợp lý?

**Bước kiểm tra và làm sạch:**

* Bỏ NaN, sai định dạng ngày → dropna(), to\_datetime(errors='coerce').
* Outliers: loại qua Z-score & IQR.
* Earnings không có EPS Estimate hoặc Reported → tự động loại bỏ trong SQL truy vấn.

**Bước 3: Data Analysis with SQL**

**Yêu cầu**:

* Import dữ liệu từ Lab 1 vào cơ sở dữ liệu và kiểm tra tính nhất quán.
* Viết truy vấn SQL nâng cao (JOIN, SUBQUERY, WINDOW FUNCTIONS).
* Stock Market: Xác định ngày volume bất thường, trend moving average, phân cụm theo ngành.
* Weather: AVG, MAX, MIN theo vùng/thời gian, xác định ngưỡng AQI/nhiệt độ.
* Lưu truy vấn vào thư mục sql/ và tích hợp kết quả vào Python.

**Câu hỏi cho từng nhóm**:

**Nhóm 5**:

1. Nhóm đã import dữ liệu từ Lab 1 vào cơ sở dữ liệu nào, và kiểm tra tính nhất quán thế nào?
2. Truy vấn nào được sử dụng để phân cụm cổ phiếu theo phản ứng giá sau earnings?
3. Làm thế nào để tính chênh lệch giá 3 ngày trước/sau earnings bằng SQL?
4. Truy vấn nào sử dụng window function để phân tích xu hướng giá?
5. Kết quả từ truy vấn phản ứng giá cho thấy điều gì về dữ liệu?
6. Nhóm đã sử dụng JOIN để kết hợp dữ liệu từ các bảng nào?
7. Truy vấn nào xác định cổ phiếu có phản ứng giá mạnh nhất?
8. Làm thế nào để phân loại mức độ phản ứng giá bằng SQL?
9. Kết quả từ các truy vấn SQL đã được tích hợp vào Python như thế nào?
10. Những thách thức nào gặp phải khi viết truy vấn SQL?

**Bước 4: Data Analysis with Python**

**Yêu cầu**:

* Thực hiện phân tích nâng cao bằng Python, mở rộng từ Lab 1.
* Sử dụng thư viện: pandas, numpy, statsmodels, scikit-learn, seaborn.
* Phân tích: thống kê mô tả (skewness, kurtosis), tương quan đa biến (heatmap), phân rã chuỗi thời gian (time-series decomposition).
* Ghi chú rõ ràng phương pháp và lưu kết quả vào file hoặc cơ sở dữ liệu.

**Câu hỏi cho từng nhóm**:

**Nhóm 5**:

1. Nhóm đã sử dụng những thư viện Python nào để phân tích dữ liệu?
2. Thống kê mô tả (mean, std, skewness, kurtosis) của giá cổ phiếu là gì?
3. Tương quan giữa giá và dữ liệu earnings là bao nhiêu, và ý nghĩa của nó?
4. Phân rã chuỗi thời gian của giá cổ phiếu cho thấy xu hướng nào?
5. Kết quả phân tích chênh lệch giá trước/sau earnings có gì đáng chú ý?
6. Nhóm đã thực hiện phân tích nào để xác định tác động của earnings?
7. Dữ liệu giá có phân phối chuẩn không, và điều này ảnh hưởng ra sao?
8. Những phát hiện nào từ phân tích Python sẽ hữu ích cho dự báo giá?
9. Kết quả phân tích đã được lưu ở đâu (file, cơ sở dữ liệu)?
10. Những thách thức nào gặp phải khi phân tích dữ liệu bằng Python?

**Bước 5: Data Visualization**

**Yêu cầu**:

* Tạo ít nhất 10 biểu đồ đa dạng (line, scatter, heatmap, waffle, choropleth, bar, area…).
* Đảm bảo tiêu đề, nhãn trục, chú giải rõ ràng, sử dụng màu sắc phù hợp.
* Lưu biểu đồ vào thư mục figures/ và giải thích ý nghĩa trong báo cáo.

**Câu hỏi cho từng nhóm**:

**Nhóm 5**:

1. Biểu đồ line thể hiện xu hướng giá trước/sau earnings cho thấy điều gì?
2. Biểu đồ bar thể hiện chênh lệch giá theo cổ phiếu cho thấy xu hướng nào?
3. Biểu đồ area thể hiện tích lũy giá theo thời gian cho thấy điều gì?
4. Biểu đồ scatter giữa giá và sentiment có điểm bất thường nào không?
5. Biểu đồ histogram của chênh lệch giá cho thấy phân phối ra sao?
6. Nhóm đã sử dụng màu sắc nào để phân biệt các cổ phiếu?
7. Biểu đồ nào giúp người dùng nhận diện cổ phiếu có phản ứng giá mạnh nhất?
8. Nhóm đã lưu các biểu đồ vào đâu, và chúng được đặt tên ra sao?
9. Ý nghĩa của từng biểu đồ trong báo cáo đã được giải thích thế nào?
10. Những thách thức nào gặp phải khi trực quan hóa dữ liệu?

**Bước 6: Machine Learning Model Implementation**

**Yêu cầu**:

* Triển khai 7 mô hình: Linear Regression, Decision Tree Regressor, KNN, Ridge, Lasso, Random Forest, SVR.
* Chia dữ liệu train/test (80/20 hoặc K-fold cross-validation).
* Đánh giá bằng MAE, RMSE, R² (hoặc Accuracy, Precision, Recall, F1-score nếu classification).
* So sánh hiệu năng: bar chart (MAE, RMSE), line chart (y\_thực tế vs. y\_dự báo), boxplot (phân phối lỗi).
* Tinh chỉnh siêu tham số (grid search, random search) để đạt KPIs.

**Câu hỏi cho từng nhóm**:

**Nhóm 5**:

1. Nhóm đã chia dữ liệu train/test theo tỷ lệ nào, và tại sao?
2. MAE, RMSE, R² của từng mô hình (Linear, Tree, KNN,...) là bao nhiêu?
3. Mô hình nào có RMSE thấp nhất, và nó có đạt KPI (RMSE < 6%) không?
4. Nhóm đã tinh chỉnh siêu tham số của mô hình nào, và kết quả ra sao?
5. Bar chart so sánh RMSE giữa các mô hình cho thấy điều gì?
6. Line chart so sánh y\_thực tế và y\_dự báo có gì đáng chú ý?
7. Boxplot phân phối lỗi của các mô hình cho thấy mô hình nào ổn định nhất?
8. Kiểm định thống kê (t-test, ANOVA) cho thấy sự khác biệt nào giữa các mô hình?
9. Những thách thức nào gặp phải khi huấn luyện 7 mô hình?
10. Nhóm đã cải thiện hiệu năng mô hình như thế nào so với Lab 1?

**Bước 7: Data Analysis with BI Tool**

**Yêu cầu**:

* Sử dụng Power BI (Nhóm 1, 2, 3, 4, 5, 6, 9), Tableau (Nhóm 8), Google Looker (Nhóm 7).
* Thêm bộ lọc tương tác (ngành, thời gian, ngưỡng volume, AQI, vùng).
* Hiển thị biểu đồ chính (line, bar, heatmap, choropleth) và KPIs (MAE, RMSE).
* Thêm tính năng drill-down và tooltip.
* Xuất file dashboard vào thư mục dashboard/.

**Câu hỏi cho từng nhóm**:

**Nhóm 5**:

1. Dashboard Power BI của nhóm có những bộ lọc nào (ngày earnings,...)?
2. Biểu đồ line trong dashboard thể hiện xu hướng giá ra sao?
3. Biểu đồ bar trong dashboard thể hiện chênh lệch giá như thế nào?
4. KPI nào (RMSE,...) được hiển thị trên dashboard?
5. Tính năng drill-down trong dashboard được áp dụng cho biểu đồ nào?
6. Tooltip trong dashboard cung cấp thông tin gì cho người dùng?
7. Dashboard đã giúp nhà đầu tư phân tích tác động earnings như thế nào?
8. Nhóm đã gặp khó khăn gì khi xây dựng dashboard Power BI?
9. File dashboard được lưu ở đâu, và tên file là gì?
10. Những cải tiến nào có thể thực hiện cho dashboard trong tương lai?

**Yêu cầu nộp Lab 2**

* **Lab2\_<tennhom>.ipynb hoặc .py</tennhom>**: Toàn bộ pipeline 7 bước. (1 file chung và 6 file tách từng bước + 1 file dashboard)
* **data/**: Dữ liệu mở rộng (CSV, DB).
* **sql/**: Truy vấn SQL chi tiết.
* **compare\_models.csv**: Bảng so sánh 7 mô hình.
* **statistical\_test.md**: Báo cáo kiểm định thống kê.
* **figures/**: Biểu đồ (bar, boxplot, heatmap, line, map, scartter plot…).
* **dashboard/**: File Power BI, Tableau, hoặc Google Looker.
* **report\_Lab2\_<tennhom>.docx</tennhom>**: Báo cáo chi tiết (mô tả từng bước, trả lời đầy đủ các câu hỏi, kết quả, insight).
* **Deadline**: 13/06/2025