TRƯỜNG ĐẠI HỌC KỸ THUẬT CÔNG NGHIỆP

**KHOA ĐIỆN TỬ**

**Bộ môn: Công nghệ Thông tin.**

**BÀI TẬP KẾT THÚC MÔN HỌC**

MÔN HỌC

**KHOA HỌC DỮ LIỆU**

Sinh viên: Lành Đức Mạnh

Lớp: K57.KMT

Giáo viên giảng dạy: Nguyễn Văn Huy

Link GitHub: https://github.com/lanhducmanh/BTL-KHOAHOCDULIEU

**Thái Nguyên – 2025**

**TRƯỜNG ĐHKTCN CỘNG HOÀ XÃ HỘI CHỦ NGHĨA VIỆT NAM**

**KHOA ĐIỆN TỬ**

**Độc lập - Tự do - Hạnh phúc**

**BÀI TẬP KẾT THÚC MÔN HỌC**

**MÔN HỌC: KHOA HỌC DỮ LIỆU**

BỘ MÔN : CÔNG NGHỆ THÔNG TIN

Sinh viên: Lành Đức Mạnh

Lớp: K57.KMT

Ngành: Công nghệ thông tin

Giáo viên hướng dẫn: Nguyễn Văn Huy

Ngày giao đề: 20/05/2025 Ngày hoàn thành: 31/052025

Tên đề tài : Phân tích và dự báo giá cổ phiếu

Đầu bài: Web app dự báo giá cổ phiếu.

Đầu vào: Stock Market Data - Kaggle

Đầu ra: Dự báo giá cổ phiếu và biểu đồ giá theo thời gian.

GIÁO VIÊN HƯỚNG DẪN

(Ký và ghi rõ họ tên)

**NHẬN XÉT CỦA GIÁO VIÊN HƯỚNG DẪN**

Thái Nguyên, ngày….tháng…..năm 20....

GIÁO VIÊN HƯỚNG DẪN

(Ký ghi rõ họ tên)

**LỜI CẢM ƠN**

Trước tiên, em xin gửi lời cảm ơn chân thành đến quý thầy cô Bộ môn Công nghệ Thông tin – Khoa Điện tử, Trường Đại học Kỹ thuật Công nghiệp đã tận tình giảng dạy và tạo điều kiện thuận lợi để em hoàn thành bài tập lớn này.

Đặc biệt, em xin gửi lời cảm ơn đến giảng viên hướng dẫn môn Khoa học dữ liệu, người đã truyền đạt cho em những kiến thức quý báu về phân tích dữ liệu, học máy và cách áp dụng vào thực tiễn. Những kiến thức này không chỉ giúp em hoàn thành bài tập mà còn tạo nền tảng cho định hướng nghề nghiệp sau này.

Trong quá trình thực hiện, mặc dù đã nỗ lực hết mình, bài báo cáo chắc chắn không tránh khỏi những thiếu sót. Em rất mong nhận được sự góp ý, chỉ dẫn của thầy cô để em hoàn thiện hơn trong những lần sau.

Em xin chân thành cảm ơn!

### ****MỤC LỤC****

**Chương 1. Giới thiệu đầu bài** ...................................................................

* 1. Mô tả đề bài…………………………………………………
  2. Mục tiêu thực hiện…………………………………………..
  3. Thách thức…………………………………………………..

**Chương 2. Cơ sở lý thuyết** ........................................................................

2.1. Các thư viện và công cụ sử dụng………………………………

2.2. Tổng quan về dự báo chuỗi thời gian………………………….

2.3. Đặc điểm dữ liệu chứng khoán………………………………...

**Chương 3. Thiết kế và xây dựng chương trình** .......................................

3.1. Sơ đồ khối hệ thống……………………………………………

3.2. Biểu đồ phân cấp chức năng……………………………………

3.3. Sơ đồ thuật toán chính………………………………………….

3.4. Cấu trúc dữ liệu…………………………………………………

3.5. Mã nguồn chương trình…………………………………………

**Chương 4. Thực nghiệm và kết luận** .........................................................

4.1. Thực nghiệm…………………………………………………….

4.2. Kết luận………………………………………………………….

**Tài liệu tham khảo** .......................................................................................

CHƯƠNG 1: GIỚI THIỆU ĐẦU BÀI

1.1. Mô tả đề bài

Trong bối cảnh phát triển mạnh mẽ của công nghệ và trí tuệ nhân tạo, việc ứng dụng khoa học dữ liệu vào lĩnh vực tài chính – ngân hàng đang ngày càng trở nên phổ biến và mang lại hiệu quả to lớn. Một trong những ứng dụng tiêu biểu là hệ thống dự báo giá cổ phiếu – công cụ hữu ích giúp nhà đầu tư có thể phân tích xu hướng thị trường, đưa ra quyết định mua bán một cách chủ động và có cơ sở.

Bài tập lớn số 9 đặt ra yêu cầu xây dựng một web app có khả năng dự báo giá cổ phiếu dựa trên tập dữ liệu từ nền tảng Kaggle – một kho dữ liệu mở lớn và uy tín. Đây không chỉ là một đề tài mang tính thực tiễn cao, mà còn là cơ hội để sinh viên tiếp cận với mô hình dự báo chuỗi thời gian – một lĩnh vực quan trọng trong khoa học dữ liệu.

Ứng dụng cần cho phép người dùng nhập vào mã cổ phiếu cần theo dõi, chọn khoảng thời gian muốn dự đoán, và hiển thị biểu đồ giá theo thời gian thực tế lẫn dự báo. Giao diện cần đảm bảo trực quan, thân thiện, dễ sử dụng với người dùng phổ thông. Thêm vào đó, ứng dụng cần có khả năng phản hồi nhanh, xử lý dữ liệu một cách chính xác và cho kết quả dự báo gần sát với thực tế.

Việc xây dựng một ứng dụng như vậy không chỉ đòi hỏi kỹ năng lập trình và tư duy logic mà còn yêu cầu hiểu biết sâu về dữ liệu tài chính, cách xử lý dữ liệu thời gian, lựa chọn mô hình học máy phù hợp và triển khai trên nền web. Đây thực sự là một đề tài tích hợp kiến thức đa ngành: từ toán học, thống kê, lập trình Python, đến học máy và phát triển ứng dụng thực tế.

1.2. Mục tiêu thực hiện

Mục tiêu chính của bài tập lớn này là giúp sinh viên vận dụng những kiến thức đã học vào giải quyết một bài toán thực tiễn, từ đó nâng cao tư duy phân tích, khả năng triển khai mô hình học máy và thiết kế hệ thống.

Cụ thể, các mục tiêu bao gồm:

- Áp dụng thành thạo thư viện Python như Pandas, Numpy để xử lý dữ liệu.

- Làm quen và sử dụng các công cụ trực quan hóa dữ liệu như Matplotlib, Plotly.

- Hiểu và triển khai mô hình dự báo chuỗi thời gian với Prophet hoặc các mô hình tương đương.

- Thiết kế giao diện người dùng đơn giản, dễ thao tác bằng Streamlit.

- Rèn luyện kỹ năng xử lý dữ liệu thực tế, đánh giá mô hình, và giải thích kết quả đầu ra.

- Xây dựng được một pipeline hoàn chỉnh từ thu thập dữ liệu → tiền xử lý → huấn luyện mô hình → dự đoán → hiển thị kết quả.

- Thực hành triển khai ứng dụng khoa học dữ liệu dưới dạng một sản phẩm có thể tương tác trực tiếp.

1.3. Thách thức

Bên cạnh những mục tiêu đã đặt ra, quá trình triển khai đề tài cũng gặp phải không ít thách thức đáng kể:

- Dữ liệu chứng khoán mang tính thời gian và có độ biến động cao, thường không tuân theo quy luật tuyến tính và có nhiều yếu tố ngoại lai (ví dụ: biến động chính trị, tin tức tài chính).

- Khối lượng dữ liệu lớn, yêu cầu sinh viên phải biết cách lọc, làm sạch, chuẩn hóa dữ liệu để đảm bảo đầu vào cho mô hình là chính xác và nhất quán.

- Việc lựa chọn mô hình phù hợp với bài toán chuỗi thời gian không đơn giản, mỗi mô hình đều có ưu – nhược điểm riêng, phụ thuộc vào tính chất dữ liệu và mục tiêu dự đoán.

- Cần tránh tình trạng overfitting – tức mô hình quá khớp với dữ liệu huấn luyện nhưng dự báo kém khi gặp dữ liệu mới.

- Triển khai giao diện web đòi hỏi không chỉ kiến thức xử lý backend (xử lý dữ liệu, dự báo) mà còn cả frontend (hiển thị kết quả, tương tác người dùng).

- Thời gian thực hiện bài tập có giới hạn, đòi hỏi sinh viên phải quản lý thời gian hợp lý, chia nhỏ công việc và phối hợp tốt giữa các bước.

Tuy nhiên, chính những khó khăn này lại là cơ hội để người học rèn luyện kỹ năng thực tế, từ việc giải quyết bài toán kỹ thuật đến tư duy logic, sáng tạo và khả năng làm việc độc lập.

CHƯƠNG 2: CƠ SỞ LÝ THUYẾT

2.1. Các thư viện và công cụ sử dụng

Việc xây dựng một ứng dụng web dự báo giá cổ phiếu không thể thiếu sự hỗ trợ của các thư viện và công cụ lập trình mạnh mẽ. Dưới đây là những thành phần chính được sử dụng:

- Pandas: Thư viện phân tích dữ liệu mạnh mẽ trong Python, được sử dụng để đọc, xử lý, làm sạch và chuẩn hóa dữ liệu từ tập tin CSV.

- Numpy: Thư viện hỗ trợ thao tác số học hiệu quả trên mảng nhiều chiều.

- Matplotlib & Plotly: Các công cụ trực quan hóa dữ liệu. Matplotlib cho phép vẽ các biểu đồ cơ bản, còn Plotly hỗ trợ biểu đồ tương tác trực tuyến.

- Scikit-learn: Bộ thư viện học máy cung cấp các thuật toán như Linear Regression, Decision Tree, Random Forest, hỗ trợ huấn luyện và đánh giá mô hình.

- Facebook Prophet: Mô hình học máy chuyên biệt cho dự báo chuỗi thời gian, có khả năng xử lý tốt dữ liệu có xu hướng và mùa vụ.

- Streamlit: Framework để xây dựng ứng dụng web với Python một cách nhanh chóng và trực quan.

- Datetime: Hỗ trợ xử lý dữ liệu kiểu thời gian như định dạng ngày tháng, trích xuất năm, tháng, tuần,...

2.2. Tổng quan về dự báo chuỗi thời gian

Dự báo chuỗi thời gian là một nhánh trong thống kê và học máy, nhằm mục đích dự đoán các giá trị trong tương lai dựa trên dữ liệu lịch sử. Đặc điểm nổi bật của chuỗi thời gian là mối quan hệ phụ thuộc theo thời gian, nghĩa là giá trị hiện tại có thể chịu ảnh hưởng từ các giá trị trong quá khứ.

Các mô hình thường được sử dụng trong dự báo chuỗi thời gian:

- ARIMA: Mô hình kết hợp giữa tự hồi quy (AR), lấy sai phân (I) và trung bình trượt (MA). Phù hợp với chuỗi không có yếu tố mùa vụ.

- Prophet (Meta): Dễ triển khai, phù hợp với dữ liệu có xu hướng và mùa vụ. Khả năng diễn giải tốt, tích hợp dễ dàng với Pandas.

- LSTM: Một biến thể của mạng neural, xử lý tốt dữ liệu chuỗi dài và phức tạp. Tuy nhiên, yêu cầu cấu hình cao và tốn nhiều thời gian huấn luyện.

Trong bài tập này, Prophet được chọn nhờ tính hiệu quả, trực quan và phù hợp với mục tiêu học tập.

2.3. Đặc điểm dữ liệu chứng khoán

Dữ liệu giá cổ phiếu mang những đặc điểm riêng biệt khiến việc dự báo trở nên khó khăn:

- Tính phi tuyến: Không tuân theo mô hình toán học cố định.

- Biến động mạnh: Bị ảnh hưởng bởi tâm lý nhà đầu tư, tin tức thị trường, chính trị...

- Có yếu tố mùa vụ: Biến động theo từng quý, từng năm.

- Thường thiếu dữ liệu: Ngày nghỉ lễ, lỗi kỹ thuật khiến dữ liệu bị mất.

- Dễ nhiễu: Cần xử lý cẩn thận để tránh làm sai lệch mô hình.

Hiểu rõ các đặc điểm này giúp chọn mô hình và xử lý dữ liệu chính xác hơn.

CHƯƠNG 3: THIẾT KẾ VÀ XÂY DỰNG CHƯƠNG TRÌNH

3.1. Sơ đồ khối hệ thống

Hệ thống dự báo giá cổ phiếu gồm các thành phần chính như sau:

[Người dùng] → [Giao diện Web] → [Tiền xử lý dữ liệu] → [Huấn luyện & Dự báo] → [Hiển thị kết quả]

Trong đó:

- Giao diện web: Giao diện nhập mã cổ phiếu, số ngày dự báo, nút dự đoán.

- Tiền xử lý dữ liệu: Làm sạch, chuyển đổi ngày tháng, chuẩn hóa.

- Huấn luyện & Dự báo: Sử dụng Prophet để huấn luyện mô hình và dự đoán giá.

- Hiển thị kết quả: Vẽ biểu đồ so sánh dữ liệu thực tế và dự báo.

3.2. Biểu đồ phân cấp chức năng

- Nhập mã cổ phiếu → Làm sạch dữ liệu → Huấn luyện mô hình → Dự báo → Hiển thị biểu đồ

Các chức năng phụ:

- Lọc dữ liệu theo ngày

- Tính trung bình trượt

- Lưu kết quả dưới dạng CSV

3.3. Sơ đồ thuật toán chính

Bắt đầu

│

├── Nhập dữ liệu (CSV)

├── Làm sạch dữ liệu (remove nulls, format date)

├── Xử lý dữ liệu (trích xuất ngày, nhóm theo thời gian)

├── Huấn luyện mô hình Prophet

├── Dự báo giá cổ phiếu trong tương lai

└── Vẽ biểu đồ kết quả

3.4. Cấu trúc dữ liệu

- Date: Ngày giao dịch

- Open: Giá mở cửa

- High: Giá cao nhất

- Low: Giá thấp nhất

- Close: Giá đóng cửa

- Volume: Khối lượng giao dịch

Dữ liệu sau tiền xử lý sẽ còn hai cột chính là 'ds' (ngày) và 'y' (giá), phù hợp với yêu cầu đầu vào của Prophet.

3.5. Mã nguồn chương trình

- load\_data(): Tải dữ liệu từ file

- clean\_data(): Làm sạch dữ liệu và chuyển đổi định dạng ngày

- train\_model(): Huấn luyện mô hình Prophet

- make\_forecast(): Dự báo tương lai theo số ngày nhập vào

- plot\_chart(): Hiển thị biểu đồ dự báo

- main(): Hàm chính chạy ứng dụng web bằng Streamlit

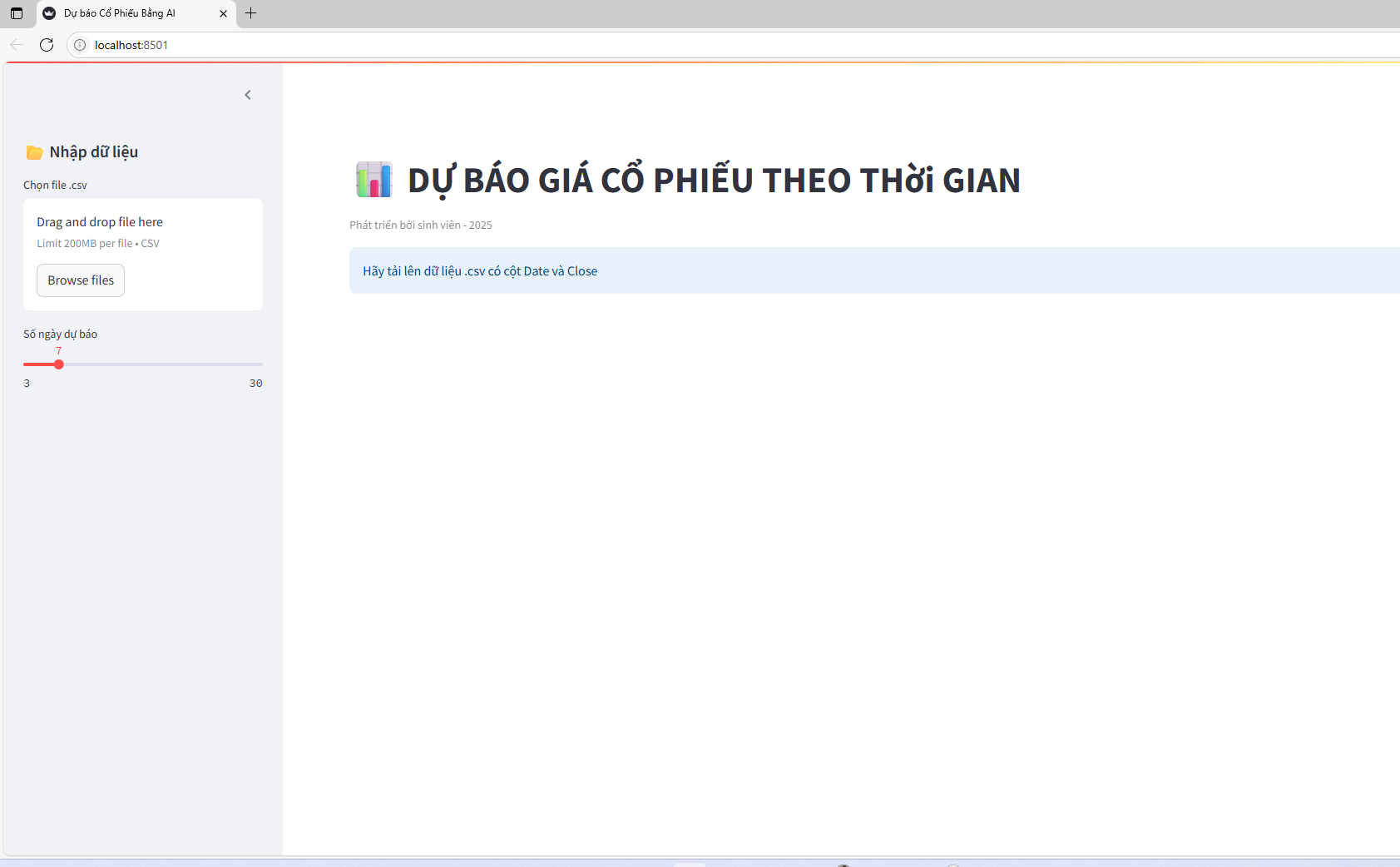
CHƯƠNG 4: THỰC NGHIỆM VÀ KẾT LUẬN

4.1. Thực nghiệm

Thực nghiệm được tiến hành với các mã cổ phiếu phổ biến.Thời gian dự báo từ 7 đến 30 ngày.

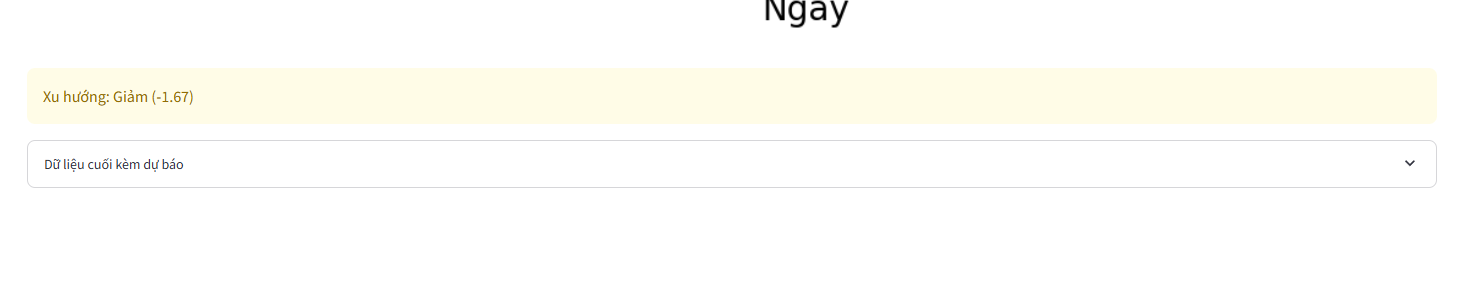
- Giao diện hoạt động tốt, tương tác nhanh.

- Biểu đồ rõ ràng, so sánh được giá thực tế và dự báo.









4.2. Kết luận

Sản phẩm đã hoàn thiện các yêu cầu đề ra. Các chức năng cơ bản như nhập dữ liệu, dự báo và hiển thị kết quả đều hoạt động ổn định.

Bài học thu nhận:

- Hiểu rõ hơn về pipeline của một dự án Khoa học Dữ liệu.

- Cải thiện khả năng xử lý dữ liệu thực tế và trực quan hóa thông tin.

- Thành thạo hơn trong việc xây dựng ứng dụng web đơn giản với Python.

Hướng phát triển:

- Kết nối API để cập nhật dữ liệu tự động.

- Thêm chức năng lưu lịch sử dự báo, xuất file.

- So sánh nhiều mã cổ phiếu cùng lúc.

- Tích hợp phân tích kỹ thuật: RSI, MACD...

TÀI LIỆU THAM KHẢO