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: ……Vi Tuấn Đạt……

Lớp: ……K57KMT……. Ngành: …Kĩ Thuật máy tính……

Giáo viên hướng dẫn: …Nguyễn Văn Huy……

Ngày giao đề: *Ngày hoàn thành:*

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

Yêu cầu :

- Web app dự báo giá cổ phiếu.

- 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…..

GIÁO VIÊN HƯỚNG DẪN

(Ký ghi rõ họ tên)

MỤC LỤC

LỜI CAM ĐOAN

Tôi xin cam đoan bài tập lớn "Khoa học dữ liệu: Phân tích và dự báo giá cổ phiếu." này là công trình nghiên cứu của riêng tôi. Các số liệu sử dụng trong luận văn là trung thực. Các kết quả nghiên cứu được trình bày trong đồ án chưa từng được công bố tại bất kỳ công trình nào khác.

Tên sinh viên

Vi Tuấn Đạt

DANH MỤC HÌNH VẼ

LỜI NÓI ĐẦU

Trong thời đại công nghệ số hiện nay, thị trường chứng khoán ngày càng trở nên sôi động và đóng vai trò quan trọng trong nền kinh tế của mỗi quốc gia. Các nhà đầu tư cá nhân và tổ chức đều phải đối mặt với thách thức trong việc phân tích và dự báo biến động giá cổ phiếu để đưa ra các quyết định đầu tư chính xác. Trong bối cảnh đó, việc ứng dụng các kỹ thuật khoa học dữ liệu, học máy và đặc biệt là học sâu (deep learning) vào bài toán dự báo giá cổ phiếu đang trở thành xu hướng tất yếu.

Trong quá trình thực hiện, đồ án sử dụng dữ liệu lịch sử giá cổ phiếu từ các sàn giao dịch lớn như NASDAQ, NYSE, SP500 và Forbes2000, được thu thập từ nguồn dữ liệu công khai trên nền tảng Kaggle. Ứng dụng được xây dựng trên nền tảng Python với framework Flask, kết hợp cùng các thư viện như Pandas, NumPy, scikit-learn, TensorFlow và Plotly để xử lý, huấn luyện mô hình và trực quan hóa kết quả.

Bên cạnh mục tiêu học thuật, đồ án còn giúp người học rèn luyện tư duy phân tích dữ liệu, kỹ năng lập trình, triển khai mô hình học máy và phát triển hệ thống ứng dụng thực tế từ đầu đến cuối. Đây là cơ hội quý báu để áp dụng kiến thức đã học vào một bài toán có ý nghĩa thực tiễn cao.

Em xin chân thành cảm ơn sự quan tâm giúp đỡ của thầy Nguyễn Văn Huy cùng toàn thể các thầy cô giáo và các bạn đã giúp đỡ em hoàn thành đề tài này.

CHƯƠNG 1: TỔNG QUAN VỀ ĐỀ TÀI

1.1. Bối cảnh và lý do chọn đề tài

Trong thời đại số hóa, dữ liệu tài chính, đặc biệt là dữ liệu giá cổ phiếu, đang ngày càng được thu thập và phân tích rộng rãi để hỗ trợ các quyết định đầu tư. Tuy nhiên, việc phân tích và dự báo giá cổ phiếu không phải là một bài toán đơn giản, bởi biến động thị trường thường chịu ảnh hưởng của nhiều yếu tố khó kiểm soát. Chính vì vậy, việc ứng dụng các mô hình học máy, đặc biệt là mô hình học sâu, đang ngày càng được chú trọng để tăng cường độ chính xác trong việc dự báo xu hướng thị trường.

Hình 1: bối cảnh

Với mong muốn vận dụng kiến thức đã học vào một bài toán thực tiễn và có ý nghĩa, nhóm đã chọn đề tài "Phân tích và dự báo giá cổ phiếu" làm đồ án môn học Khoa học Dữ liệu.

1.2. Mục tiêu đề tài

Mục tiêu chính của đề tài là xây dựng một ứng dụng web cho phép người dùng lựa chọn sàn giao dịch, mã cổ phiếu và số ngày cần dự báo, từ đó hệ thống sẽ:

- Trích xuất dữ liệu giá cổ phiếu lịch sử từ kho dữ liệu có sẵn.

- Huấn luyện mô hình dự báo theo từng mã cổ phiếu cụ thể.

- Trả kết quả dự báo giá cổ phiếu trong tương lai.

- Hiển thị biểu đồ tương tác, kết hợp giữa dữ liệu thực tế và dự báo.

1.3. Yêu cầu và tính năng của chương trình

Chương trình có các chức năng chính như sau:

Lấy dữ liệu tự động từ thư mục đã cho (chứa dữ liệu từ các sàn: NASDAQ, NYSE, SP500, Forbes2000).

Tự động nhận diện mã cổ phiếu theo từng sàn, không cần nhập tay.

Huấn luyện mô hình LSTM cho từng mã nếu chưa có mô hình, tránh huấn luyện lại không cần thiết.

Cho phép người dùng chọn số ngày cần dự báo.

Hiển thị biểu đồ dự báo giá cổ phiếu bằng biểu đồ tương tác (Plotly), có thể zoom, pan, hover rõ từng điểm.

Tự động lưu mô hình và kết quả để tái sử dụng.

1.4. Thách thức khi thực hiện

Trong quá trình thực hiện đề tài, nhóm gặp phải một số thách thức như:

Tiền xử lý dữ liệu từ nhiều nguồn và định dạng khác nhau, không đồng bộ về thời gian hoặc định dạng ngày tháng.

Huấn luyện mô hình LSTM hiệu quả cho mỗi mã cổ phiếu có dữ liệu dài/ngắn khác nhau.

Tối ưu hóa tốc độ phản hồi trên web, đặc biệt khi người dùng chọn các mã chưa có mô hình.

Trực quan hóa dữ liệu chuyên nghiệp, giúp người dùng dễ hiểu và dễ phân tích.

Tương thích đa trình duyệt và tối ưu giao diện người dùng (UI/UX).

1.5. Kiến thức đã vận dụng

Trong suốt quá trình thực hiện đồ án, nhóm đã vận dụng tổng hợp nhiều kiến thức đã học, bao gồm:

Xử lý dữ liệu với Pandas, NumPy.

Tiền xử lý và chuẩn hóa dữ liệu bằng MinMaxScaler.

Xây dựng mô hình học sâu (LSTM) bằng TensorFlow/Keras.

Trực quan hóa dữ liệu bằng thư viện Plotly.

Phát triển ứng dụng web với Flask, sử dụng Blueprint, template engine (Jinja2).

Tổ chức project khoa học dữ liệu theo hướng module hóa và dễ bảo trì.

Tóm tắt chương: *trình bày bối cảnh thực tiễn và lý do lựa chọn đề tài dự báo giá cổ phiếu, đồng thời nêu rõ mục tiêu của hệ thống là xây dựng một ứng dụng web có khả năng tự động lấy dữ liệu, huấn luyện mô hình và hiển thị dự báo giá cổ phiếu tương tác. Chương cũng liệt kê các yêu cầu chức năng, thách thức gặp phải trong quá trình triển khai và những kiến thức đã được áp dụng như xử lý dữ liệu, xây dựng mô hình học sâu và phát triển ứng dụng web.*

CHƯƠNG 2: CƠ SỞ LÝ THUYẾT

2.1. Dữ liệu chuỗi thời gian (Time Series)

Dữ liệu chuỗi thời gian là tập hợp các giá trị dữ liệu được ghi nhận theo thời gian với khoảng cách đều nhau, ví dụ: giá cổ phiếu theo ngày, tháng, quý. Việc phân tích chuỗi thời gian cho phép nhận biết xu hướng (trend), mùa vụ (seasonality), và nhiễu (noise) để dự báo giá trị tương lai dựa trên các giá trị quá khứ.

Trong bài toán này, **dữ liệu giá đóng cửa (Close price)** của cổ phiếu được coi là một chuỗi thời gian đầu vào cho mô hình dự báo.

2.2. Mô hình học sâu LSTM (Long Short-Term Memory)

LSTM là một loại mạng nơ-ron hồi tiếp (Recurrent Neural Network – RNN), được thiết kế đặc biệt để xử lý các chuỗi dữ liệu dài và có phụ thuộc thời gian. Khác với RNN truyền thống dễ gặp vấn đề **mất thông tin dài hạn** (vanishing gradient), LSTM sử dụng **các cổng (gates)** để lưu giữ và điều tiết thông tin quan trọng trong quá trình học.

Cấu trúc cơ bản của một lớp LSTM gồm 3 thành phần chính:

Cổng vào (Input Gate): Quyết định thông tin nào sẽ được thêm vào trạng thái nhớ.

Cổng quên (Forget Gate): Quyết định thông tin nào cần loại bỏ khỏi trạng thái nhớ.

Cổng đầu ra (Output Gate): Quyết định thông tin nào sẽ được sử dụng làm đầu ra tại thời điểm hiện tại.

LSTM rất phù hợp để dự báo chuỗi thời gian như giá cổ phiếu, vì nó có khả năng ghi nhớ xu hướng trong quá khứ và học được các mô hình ẩn trong dữ liệu.

2.3. Chuẩn hóa dữ liệu (Normalization)

Để tăng hiệu quả học của mô hình, dữ liệu đầu vào thường được chuẩn hóa về cùng một khoảng giá trị. Trong đồ án này, nhóm sử dụng phương pháp **Min-Max Scaling**, đưa giá trị về khoảng [0, 1] theo công thức:

$$X_{scaled} = \frac{X - X_{min}}{X_{max} - X_{min}}$$

Sau khi dự báo, kết quả sẽ được **giải chuẩn hóa** (inverse transform) để đưa về giá trị thực.

2.4. Kiến trúc ứng dụng Flask

Flask là một framework web nhẹ và linh hoạt trong Python. Ứng dụng Flask trong đồ án được tổ chức theo mô hình **modular structure** gồm các thành phần:

Blueprint: phân chia các chức năng thành modules (trang chủ, dự báo, API).

Templates (Jinja2): hỗ trợ tạo giao diện động HTML.

Static: chứa CSS, JS và hình ảnh phục vụ frontend.

Flask cho phép kết hợp chặt chẽ giữa backend Python và frontend HTML/JS để xây dựng một hệ thống dự báo hoàn chỉnh.

2.5. Trực quan hóa dữ liệu với Plotly

Plotly là một thư viện vẽ biểu đồ tương tác cho web. Ưu điểm:

Biểu đồ đẹp, hỗ trợ zoom, pan, hover chuyên nghiệp.

Tương thích tốt với Flask thông qua Jinja2.

Hiển thị đồng thời **giá thực tế và giá dự báo** rõ ràng.

Nhờ Plotly, người dùng có thể dễ dàng phân tích xu hướng cổ phiếu trực tiếp từ trình duyệt mà không cần dùng phần mềm chuyên dụng.

2.6. Một số công nghệ khác

Pandas & NumPy: xử lý, phân tích dữ liệu dạng bảng và mảng.

Scikit-learn: chuẩn hóa dữ liệu, tiền xử lý.

TensorFlow/Keras: xây dựng, huấn luyện mô hình LSTM.

Matplotlib (giai đoạn đầu): vẽ biểu đồ thử nghiệm trước khi chuyển sang Plotly.

Tóm tắt chương: *Cung cấp các kiến thức lý thuyết làm nền tảng cho đồ án. Trước hết là khái niệm về dữ liệu chuỗi thời gian – loại dữ liệu đặc trưng trong dự báo tài chính. Tiếp đến là mô hình học sâu LSTM, một biến thể của mạng nơ-ron hồi tiếp, có khả năng học được các quan hệ dài hạn trong chuỗi dữ liệu. Chương cũng giới thiệu kỹ thuật chuẩn hóa dữ liệu bằng MinMaxScaler nhằm nâng cao hiệu quả huấn luyện mô hình. Về mặt triển khai, chương trình sử dụng Flask làm framework backend cho web, và Plotly để trực quan hóa biểu đồ dự báo. Một số thư viện hỗ trợ khác cũng được đề cập như Pandas, NumPy, scikit-learn, TensorFlow và Matplotlib.*

CHƯƠNG 3: THIẾT KẾ VÀ XÂY DỰNG CHƯƠNG TRÌNH

3.1. Sơ đồ khối hệ thống

Hệ thống dự báo giá cổ phiếu được chia thành 4 module chính, mỗi module đảm nhiệm một nhóm chức năng cụ thể:

1 - Giao diện người dùng (Frontend):

Giao diện web được xây dựng bằng HTML + CSS + JavaScript, sử dụng template engine Jinja2 của Flask.

Cho phép người dùng chọn sàn giao dịch, mã cổ phiếu và số ngày dự đoán.

2 - Xử lý dữ liệu (Data Processing):

Tự động tìm kiếm và đọc dữ liệu CSV từ các thư mục theo sàn.

Tiền xử lý: đọc dữ liệu, chuẩn hóa bằng MinMaxScaler, kiểm tra định dạng ngày tháng.

3 - Huấn luyện & dự báo (Model Training & Prediction):

Nếu chưa có mô hình, hệ thống sẽ huấn luyện mô hình LSTM theo từng mã cổ phiếu.

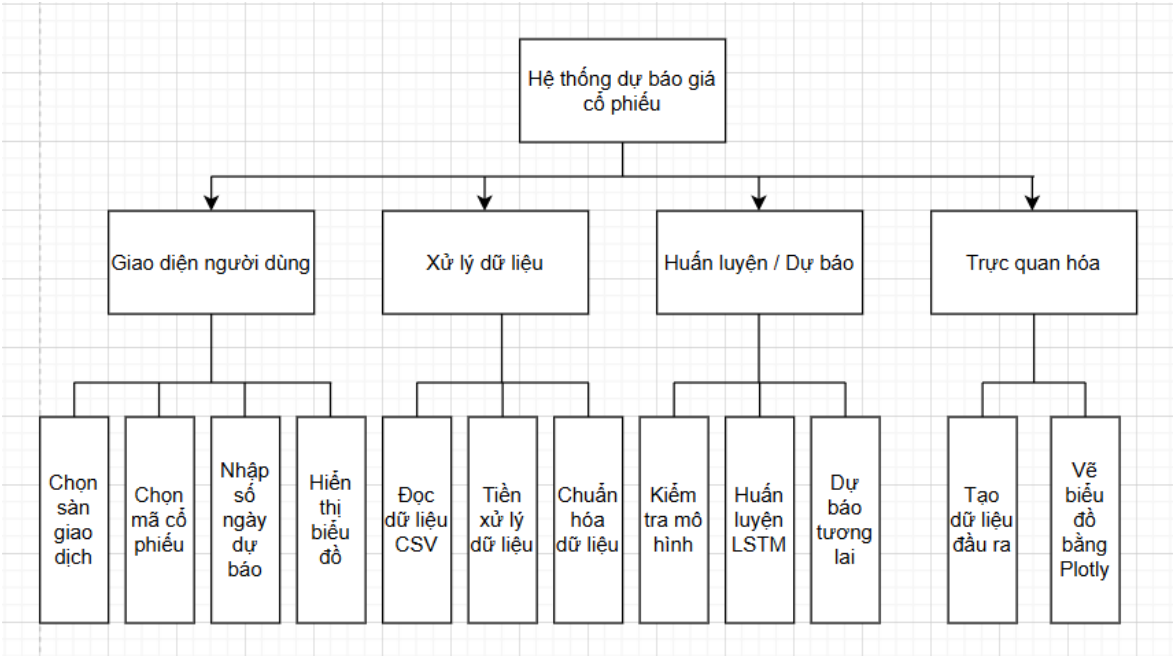
Nếu đã có mô hình, hệ thống chỉ dự báo giá theo số ngày yêu cầu.

4 - Trực quan hóa (Visualization):

Hiển thị kết quả dự báo bằng biểu đồ tương tác Plotly.

Phân biệt rõ phần dữ liệu thực tế và dữ liệu dự báo.

3.2. Biểu đồ phân cấp chức năng



Hình 2: Biểu đồ phân cấp chức năng

3.3. Sơ đồ khối thuật toán chính

- **Khối đọc và chuẩn hóa dữ liệu**

  Đầu vào: file CSV

  Chức năng: đọc dữ liệu cột `Date` và `Close`, chuyển định dạng ngày, chuẩn hóa dữ liệu với MinMaxScaler

- **Khối kiểm tra mô hình**

  Đầu vào: tên sàn, mã cổ phiếu

  Chức năng: kiểm tra xem mô hình đã được huấn luyện chưa. Nếu chưa, huấn luyện và lưu lại

- **Khối huấn luyện mô hình LSTM**

  Đầu vào: dữ liệu đã chuẩn hóa

  Chức năng: tạo tập dữ liệu chuỗi, huấn luyện mô hình LSTM với 2 lớp, lưu model `.h5` và scaler `.pkl`

- **Khối dự báo**

  Đầu vào: 60 ngày gần nhất

  Chức năng: dự báo liên tiếp `n` ngày tiếp theo, mỗi lần lặp dùng đầu ra trước đó làm đầu vào mới

- **Khối hiển thị kết quả**

  Đầu ra: danh sách ngày dự báo + giá tương ứng

  Chức năng: truyền dữ liệu sang giao diện và hiển thị bằng biểu đồ Plotly

3.4. Cấu trúc dữ liệu

3.4.1. Dữ liệu đầu vào

Mỗi sàn có thư mục `csv/` chứa nhiều file CSV.

Mỗi file có cấu trúc giống nhau:

Trường	Ý nghĩa
Date	Ngày giao dịch
Open	Giá mở cửa
High	Giá cao nhất
Low	Giá thấp nhất
Close	Giá đóng cửa (dùng để dự báo)
Volume	Khối lượng giao dịch

3.4.2. Dữ liệu mô hình

Model được lưu tại: `mo_hinh/lstm_models/<san>_<ma>.h5`

Scaler tương ứng: `..._scaler.pkl`

3.5. Các hàm chính trong chương trình.

Hàm	Mô tả
doc_du_lieu_csv()	Đọc và xử lý file CSV từ thư mục theo sàn
tao_du_lieu_lstm()	Tạo chuỗi dữ liệu đầu vào cho LSTM
huan_luyen_model(ma)	Huấn luyện mô hình LSTM nếu chưa có
du_bao_gia(ma, so_ngay)	Dự báo giá n ngày tiếp theo
render_template("du_bao.html", ...)	Truyền kết quả sang giao diện HTML
ve_bieu_do_plotly()	Vẽ biểu đồ dữ liệu bằng thư viện Plotly

Tóm tắt chương: *trình bày chi tiết thiết kế và quá trình xây dựng hệ thống. Hệ thống được chia thành bốn module chính: giao diện người dùng, xử lý dữ liệu, huấn luyện/dự báo và trực quan hóa. Mỗi module đảm nhận một nhóm chức năng rõ ràng, từ việc đọc và chuẩn hóa dữ liệu từ file CSV, đến huấn luyện mô hình LSTM, dự báo giá cổ phiếu và hiển thị biểu đồ dự báo. Chương này cũng trình bày sơ đồ khối thuật toán, mô tả từng bước trong quy trình hoạt động của hệ thống. Ngoài ra, cấu trúc dữ liệu đầu vào và đầu ra cũng được phân tích rõ ràng. Cuối cùng là phần giới thiệu các hàm chính trong chương trình – những thành phần cốt lõi đảm bảo hệ thống hoạt động chính xác và hiệu quả.*

CHƯƠNG 4: THỰC NGHIỆM VÀ KẾT LUẬN

4.1. Thực nghiệm

4.1.1. Môi trường thực thi

Ngôn ngữ: Python 3.10

Framework: Flask

Thư viện sử dụng: Pandas, NumPy, scikit-learn, TensorFlow (Keras), Plotly

Giao diện: HTML, CSS, JavaScript, Jinja2

Trình duyệt kiểm thử: Google Chrome

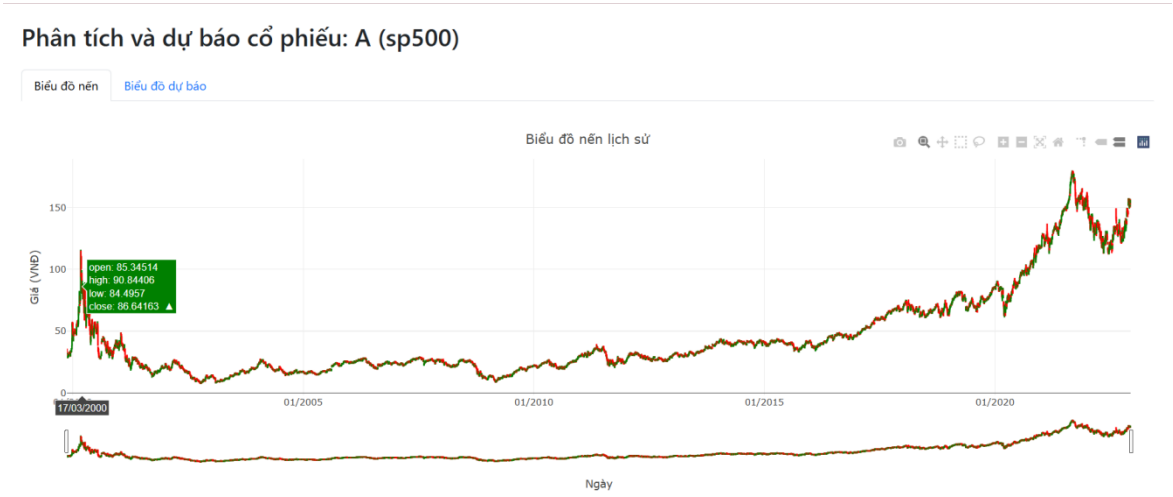
Hệ điều hành: Windows 10

4.1.2. Kịch bản kiểm thử và kết quả

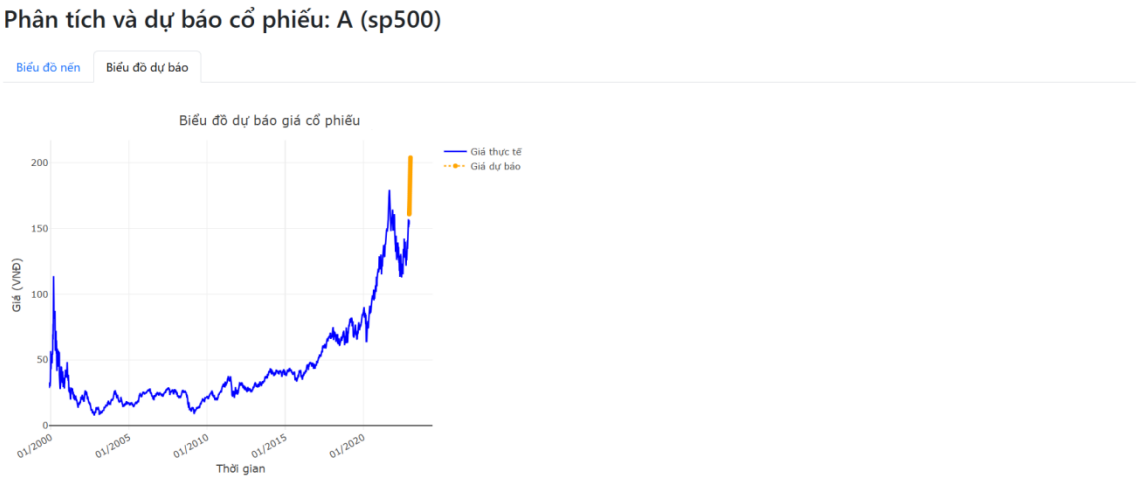
STT	Chức năng kiểm thử	Mô tả thao tác kiểm tra	Kết quả
1	Truy cập trang chủ	Người dùng vào localhost:5000	Đạt
2	Chọn sàn và hiển thị mã cổ phiếu	Chọn NASDAQ → danh sách mã tự động hiện ra	Đạt
3	Nhập số ngày dự báo và chạy dự báo	Chọn 30 ngày, nhấn nút “Dự báo”	Đạt
4	Tự động huấn luyện model nếu chưa có	Chọn mã chưa từng chạy → hệ thống huấn luyện mới	Đạt
5	Tái sử dụng model đã huấn luyện	Chọn lại mã cổ phiếu đã có model → không huấn luyện lại	Đạt
6	Hiển thị biểu đồ giá cổ phiếu	Biểu đồ hiện rõ giá thực tế và dự báo, có zoom/hover	Đạt
7	Biểu đồ theo thời gian chuẩn (tháng/năm)	Trục X hiển thị tháng/năm, hover hiển thị ngày/tháng/năm	Đạt

4.1.3. Hình ảnh minh họa

Hình 3:Giao diện chọn sàn và mã cổ phiếu



Hình 4: Biểu Đồ dữ liệu phân tích



Hình 5:Biểu đồ dự đoán

4.2. Kết luận

4.2.1. Những gì đồ án đã thực hiện được

Xây dựng thành công ứng dụng web Flask phục vụ dự báo giá cổ phiếu.

Tự động quét và hiển thị danh sách mã cổ phiếu theo từng sàn giao dịch.

Áp dụng mô hình học sâu LSTM để huấn luyện và dự báo giá cổ phiếu.

Tự động huấn luyện mô hình khi cần và tái sử dụng khi đã có sẵn.

Hiển thị biểu đồ tương tác bằng Plotly, chuyên nghiệp và dễ sử dụng.

Giao diện người dùng thân thiện, tùy chỉnh được số ngày dự báo.

4.2.2. Kiến thức học được

Vận dụng kiến thức về xử lý chuỗi thời gian trong thực tế.

Hiểu rõ cách xây dựng mô hình LSTM và áp dụng cho dữ liệu tài chính.

Biết cách tổ chức, triển khai và bảo trì một ứng dụng Flask hoàn chỉnh.

Thành thạo việc kết hợp frontend và backend trong dự án web thực tế.

Kỹ năng làm việc theo quy trình và viết tài liệu kỹ thuật.

4.2.3. Hướng cải tiến trong tương lai

Cho phép so sánh nhiều mã cổ phiếu trong cùng một biểu đồ.

Bổ sung tùy chọn mô hình dự báo khác như ARIMA, GRU, Transformer.

Thêm các chỉ báo kỹ thuật như RSI, MACD, Đường trung bình (MA).

Lưu lại lịch sử dự báo của người dùng theo từng phiên.

Tích hợp API RESTful để mở rộng ra ứng dụng di động hoặc dịch vụ web.

Thêm hệ thống gợi ý cổ phiếu tiềm năng dựa trên phân tích dữ liệu lớn.

TÀI LIỆU THAM KHẢO

https://chatgpt.com/
https://www.kaggle.com/datasets/paultimothymooney/stock-market-data
https://scikit-learn.org
https://pandas.pydata.org
https://www.tensorflow.org
https://machinelearningmastery.com