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: DƯƠNG VĂN QUÝ

  Lớp: K57-KMT MSSV : K215480106118

Giáo viên GIẢNG DẠY: NGUYỄN VĂN HUY

Link GitHub: …………………………………………………………

**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: DƯƠNG VĂN QUÝ*

*Lớp*: K57-KMT        *Ngành: Kĩ Thuật Máy Tính*

*Giáo viên hướng dẫn : NGUYỄN VĂN HUY*

*Ngày giao đề : 20/5/2025*  *Ngày hoàn thành : 30/5/2025*

*Tên đề tài : Phần tích và dự báo giá cổ phiếu*

*Yêu cầu : Phân tích và dự báo giá cổ phiếu*

***Đầu bài:*** *Web app dự báo giá cổ phiếu.*

***Đầu vào:***

* [*Stock Market Data - Kaggle*](https://www.kaggle.com/datasets/paultimothymooney/stock-market-data)

***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)***

**MỤC LỤC**

**I . Giới thiệu đầu bài......................................................................Trang 5**

**II . Cơ sở lý thuyết..........................................................................Trang 6**

**2.1. Xử lý dữ liệu thời gian với Pandas**

**2.2. Mô hình LSTM**

**2.3. Chuẩn hoá và chia sẻ dữ liệu**

**III . Thiết kế và xây dựng chương trình.....................................Trang 7**

**3.1. Sơ đồ khối hệ thống**

**3.2. Sơ đồ khối thuật toán chính**

**3.3. Cấu trúc dữ liệu**

**3.4. Mã nguồn chương trình**

**IV . Thực nghiệm và kết luận.......................................................Trang 11**

**4.1. Thực nghiệm**

**4.2. Kết luận**

**GIỚI THIỆU ĐẦU BÀI**

Trong thời đại công nghệ số phát triển mạnh mẽ, việc phân tích và dự báo giá cổ phiếu trở thành một trong những ứng dụng quan trọng và thiết thực của khoa học dữ liệu trong lĩnh vực tài chính. Thị trường chứng khoán vốn dĩ biến động mạnh và chịu ảnh hưởng từ nhiều yếu tố như kinh tế vĩ mô, tâm lý nhà đầu tư, tin tức, chính sách, v.v... Việc sử dụng các mô hình học sâu, đặc biệt là mạng nơ-ron LSTM (Long Short-Term Memory), mở ra hướng tiếp cận mới giúp cải thiện độ chính xác trong dự đoán chuỗi thời gian giá cổ phiếu.

Đề tài "Phân tích và dự báo giá cổ phiếu" được thực hiện với mục tiêu:

* Áp dụng quy trình khoa học dữ liệu để xử lý và phân tích dữ liệu giá cổ phiếu.
* Xây dựng mô hình dự báo giá cổ phiếu trong tương lai dựa vào dữ liệu lịch sử.
* Đánh giá độ chính xác của mô hình và rút ra nhận xét, định hướng phát triển.

Trong khuôn khổ bài tập lớn này, sinh viên vận dụng các kiến thức đã học trong môn Khoa học Dữ liệu như: tiền xử lý dữ liệu, trực quan hóa, xử lý chuỗi thời gian, chuẩn hóa dữ liệu, tách tập huấn luyện và kiểm tra, và đặc biệt là áp dụng mô hình học sâu LSTM để dự đoán giá đóng cửa của cổ phiếu.

Ngoài ra, sản phẩm còn được kiểm thử bằng dữ liệu thực tế từ thị trường, thông qua biểu đồ trực quan và các chỉ số sai số như MSE, RMSE. Từ đó rút ra kết luận về hiệu quả và khả năng ứng dụng của mô hình trong thực tiễn.

**CƠ SỞ LÝ THUYẾT**

2.1. Xử lý dữ liệu thời gian với Pandas

Pandas là một thư viện mạnh mẽ trong Python để xử lý dữ liệu dạng bảng và chuỗi thời gian. Một số thao tác cơ bản như đọc dữ liệu CSV, chuyển đổi kiểu dữ liệu thời gian, thiết lập chỉ mục thời gian và vẽ biểu đồ được sử dụng để làm sạch và chuẩn bị dữ liệu.

2.2. Mô hình LSTM

LSTM (Long Short-Term Memory) là một dạng mạng nơ-ron hồi tiếp (RNN) có khả năng học các phụ thuộc dài hạn. Mỗi đơn vị LSTM bao gồm các cổng: cổng vào, cổng quên và cổng đầu ra để điều khiển dòng thông tin.

Ưu điểm:

* Giảm thiểu hiện tượng mất thông tin ở các chuỗi dài.
* Hiệu quả trong các bài toán chuỗi thời gian như dự báo giá, dịch máy, nhận dạng giọng nói.

2.3. Chuẩn hóa và chia dữ liệu

* Chuẩn hóa dữ liệu về [0, 1] giúp tăng hiệu quả học của mô hình.
* Dữ liệu được chia thành tập huấn luyện (80%) và kiểm tra (20%).
* Dữ liệu được cắt thành các chuỗi có độ dài 60 để huấn luyện.

**THIẾT KẾ VÀ XÂY DỰNG CHƯƠNG TRÌNH**

3.1. Sơ đồ khối hệ thống

Dữ liệu đầu vào (CSV)

↓

Tiền xử lý và chuẩn hóa

↓

Tạo chuỗi dữ liệu với look\_back

↓

Xây dựng mô hình LSTM

↓

Huấn luyện mô hình

↓

Dự đoán và đánh giá kết quả

3.2. Sơ đồ khối thuật toán chính

a) Đọc và xử lý dữ liệu

b) Tạo tập train/test

c) Huấn luyện mô hình

d) Dự đoán

e) Trực quan hóa kết quả

3.3. Cấu trúc dữ liệu

* Dữ liệu đầu vào: extended\_stock\_data.csv
* Các cột chính: Date (ngày), Close (giá đóng cửa)
* Sau chuẩn hóa, tạo chuỗi 60 ngày để dự đoán giá ngày tiếp theo

3.4. Mã nguồn chương trình:

|  |
| --- |
| import streamlit as st import pandas as pd import numpy as np import matplotlib.pyplot as plt from sklearn.preprocessing import MinMaxScaler from tensorflow.keras.models import Sequential from tensorflow.keras.layers import LSTM, Dense  st.title("📈 Dự báo giá cổ phiếu bằng LSTM") st.write("Tải lên một file CSV chứa dữ liệu giá cổ phiếu để phân tích và dự báo.")  uploaded\_file = st.file\_uploader("📤 Vui lòng tải lên file CSV có cột 'Date' và 'Close'.", type="csv")  def create\_dataset(dataset, look\_back=1):  X, Y = [], []  for i in range(len(dataset) - look\_back):  a = dataset[i:(i + look\_back), 0]  X.append(a)  Y.append(dataset[i + look\_back, 0])  return np.array(X), np.array(Y)  if uploaded\_file is not None:  df = pd.read\_csv(uploaded\_file)   # Kiểm tra cột cần thiết  if 'Date' not in df.columns or 'Close' not in df.columns:  st.error("❌ File CSV phải có cột 'Date' và 'Close'.")  else:  df['Date'] = pd.to\_datetime(df['Date'])  df = df.sort\_values('Date')   st.success("✅ File hợp lệ. Hiển thị dữ liệu gần nhất:")  st.dataframe(df.tail(10))   # Chuẩn bị dữ liệu cho LSTM  data = df[['Close']].values.astype('float32')   scaler = MinMaxScaler(feature\_range=(0, 1))  data\_scaled = scaler.fit\_transform(data)   look\_back = 10 # số bước thời gian dựa vào để dự đoán   X, Y = create\_dataset(data\_scaled, look\_back)   # Chia tập train/test: 80% train  train\_size = int(len(X) \* 0.8)  X\_train, X\_test = X[:train\_size], X[train\_size:]  Y\_train, Y\_test = Y[:train\_size], Y[train\_size:]   # Reshape input cho LSTM [samples, time steps, features]  X\_train = np.reshape(X\_train, (X\_train.shape[0], X\_train.shape[1], 1))  X\_test = np.reshape(X\_test, (X\_test.shape[0], X\_test.shape[1], 1))   # Xây dựng mô hình LSTM  model = Sequential()  model.add(LSTM(50, input\_shape=(look\_back, 1)))  model.add(Dense(1))  model.compile(loss='mean\_squared\_error', optimizer='adam')   # Huấn luyện mô hình  st.write("⏳ Đang huấn luyện mô hình LSTM...")  model.fit(X\_train, Y\_train, epochs=20, batch\_size=16, verbose=0)  st.success("✅ Đã huấn luyện xong mô hình!")   # Dự báo  train\_predict = model.predict(X\_train)  test\_predict = model.predict(X\_test)   # Chuyển ngược về giá trị thực  train\_predict = scaler.inverse\_transform(train\_predict)  Y\_train\_real = scaler.inverse\_transform(Y\_train.reshape(-1,1))  test\_predict = scaler.inverse\_transform(test\_predict)  Y\_test\_real = scaler.inverse\_transform(Y\_test.reshape(-1,1))   # Vẽ biểu đồ so sánh dự báo  st.subheader("Biểu đồ dự báo giá cổ phiếu")   fig, ax = plt.subplots(figsize=(12,6))  ax.plot(df['Date'][look\_back:look\_back+len(train\_predict)], train\_predict, label='Dự báo train')  ax.plot(df['Date'][look\_back+len(train\_predict):look\_back+len(train\_predict)+len(test\_predict)], test\_predict, label='Dự báo test')  ax.plot(df['Date'], df['Close'], label='Giá thực tế', color='black', alpha=0.6)  ax.set\_xlabel('Ngày')  ax.set\_ylabel('Giá đóng cửa')  ax.legend()  st.pyplot(fig) else:  st.info("📥 Vui lòng tải lên file CSV để bắt đầu dự báo.") |

**THỰC NGHIỆM VÀ KẾT LUẬN**

4.1. Thực nghiệm

* Dataset sử dụng có khoảng 2000 dòng dữ liệu lịch sử giá cổ phiếu
* Mô hình LSTM gồm 2 lớp, được huấn luyện trong 10 epoch, batch size 32
* Mô hình dự báo sát với thực tế và thể hiện tốt xu hướng
* Biểu đồ đầu ra hiển thị rõ ràng sự chênh lệch giữa giá dự đoán và thực tế

4.2. Kết luận

* Mô hình đã dự báo được xu hướng giá cổ phiếu tương đối tốt
* Quá trình thực hiện giúp sinh viên hiểu rõ hơn về chuỗi thời gian, LSTM và xử lý dữ liệu thực tế
* Hướng phát triển: thêm đặc trưng như khối lượng, giá mở cửa, giá cao/thấp, hoặc áp dụng mô hình CNN-LSTM