|  |  |
| --- | --- |
|  |  |
| **TÔN QUANG THUẬN** | **BỘ CÔNG THƯƠNG**  **TRƯỜNG ĐẠI HỌC CÔNG NGHIỆP HÀ NỘI**  **---------------------------------------** |
|  |
| **ĐỒ ÁN TỐT NGHIỆP ĐẠI HỌC** |
| **NGÀNH KHOA HỌC MÁY TÍNH** |
| **XÂY DỰNG ỨNG DỤNG DỰ ĐOÁN SỰ BIẾN THIÊN CỦA CÁC MÃ CHỨNG KHOÁN TRONG THỊ TRƯỜNG CHỨNG KHOÁN BẰNG PYTHON** |
|  |
|  |
| **CBHD:Th.s Nguyễn Lan Anh** |
| **KHOA HỌC MÁY TÍNH** | **Sinh viên: Tôn Quang Thuận** |
| **Mã số sinh viên: 2020607280** |
|  |
|  |
|  |
| **Hà Nội – Năm 2024** |
|  |

**MỤC LỤC**

[MỤC LỤC ii](#_Toc168192753)

[DANH MỤC CHỮ VIẾT TẮT v](#_Toc168192754)

[DANH MỤC HÌNH ẢNH vi](#_Toc168192755)

[DANH MỤC BẢNG BIỂU viii](#_Toc168192756)

[LỜI CẢM ƠN ix](#_Toc168192757)

[LỜI MỞ ĐẦU x](#_Toc168192758)

[CHƯƠNG 1: TÌM HIỂU VÀ PHÁT BIỂU BÀI TOÁN 1](#_Toc168192759)

[**1.1.** **Chứng khoán là gì?** 1](#_Toc168192760)

[**1.1.1.** **Cổ phiếu là gì?** 1](#_Toc168192761)

[**1.1.2.** **Trái phiếu là gì?** 3](#_Toc168192762)

[**1.1.3.** **Chứng quyền là gì?** 4](#_Toc168192763)

[**1.1.4.** **Đặc điểm của chứng khoán** 5](#_Toc168192764)

[**1.2.** **Thị trường chứng khoán là gì?** 6](#_Toc168192765)

[**1.3.** **Vai trò, chức năng của thị trường chứng khoán** 7](#_Toc168192766)

[**1.3.1.** **Huy động vốn cho nên kinh tế** 7](#_Toc168192767)

[**1.3.2.** **Cung cấp môi trường đầu tư cho công chúng** 8](#_Toc168192768)

[**1.3.3.** **Tạo tính thanh khoản cho các chứng khoán** 8](#_Toc168192769)

[**1.3.4.** **Đánh giá hoạt động của doanh nghiệp** 8](#_Toc168192770)

[**1.3.5.** **Tạo môi trường giúp chính phủ thực hiện các chính sách vĩ mô** 8](#_Toc168192771)

[**1.4.** **Phát biểu bài toán** 9](#_Toc168192772)

[**1.4.1.** **Input** 9](#_Toc168192773)

[**1.4.2.** **Output** 9](#_Toc168192774)

[**1.4.3.** **Mục tiêu** 9](#_Toc168192775)

[**1.4.4.** **Ràng buộc** 10](#_Toc168192776)

[**1.4.5.** **Cách tiếp cận** 10](#_Toc168192777)

[**1.4.6.** **Các bước thực hiện** 10](#_Toc168192778)

[**1.5. Khảo sát các phương pháp dự đoán** 11](#_Toc168192779)

[**1.5.1. Phương pháp thống kê (Statistical Methods)** 11](#_Toc168192780)

[**1.5.2. Các mô hình học máy (Machine Learning)** 11](#_Toc168192781)

[**1.5.3. Các mô hình học sâu (Deep Learning)** 12](#_Toc168192782)

[CHƯƠNG 2: TÌM HIỂU MỘT SỐ THUẬT TOÁN AI VÀ MỘT SỐ THƯ VIỆN PYTHON CẦN DÙNG 14](#_Toc168192783)

[**2.1. Thuận toán LSTM** 14](#_Toc168192784)

[**2.2.1. Giới thiệu về LSTM** 14](#_Toc168192785)

[**2.2.2. Mạng trí nhớ ngắn hạn định hướng dài hạn (LSTM – Long Short-Term Memory)** 14](#_Toc168192786)

[**2.2.2. Ý tưởng đằng sau LSTM** 16](#_Toc168192787)

[**2.2.3. Thứ tự các bước của LSTM** 17](#_Toc168192788)

[**2.2.4. Các biển thể của LSTM** 20](#_Toc168192789)

[**2.2.5. Ưu điểm** 22](#_Toc168192790)

[**2.2.6. Nhược điểm** 22](#_Toc168192791)

[**2.2. Thuật toán RNN** 23](#_Toc168192792)

[**2.2.1. Giới thiệu về RNN** 23](#_Toc168192793)

[**2.2.2. Dạng của Linear Regression** 23](#_Toc168192794)

[**2.2.3. Sai số dự đoán** 24](#_Toc168192795)

[**2.2.4. Hàm mất mát** 24](#_Toc168192796)

[**2.2.5. Nghiệm cho bài toàn Liearn Regression** 25](#_Toc168192797)

[**2.2.6. Phân loại bài toán RNN** 26](#_Toc168192798)

[**2.2.7. Ứng dụng bài toán RNN** 27](#_Toc168192799)

[**2.2.8. Ưu điểm** 27](#_Toc168192800)

[**2.2.9. Nhược điểm** 28](#_Toc168192801)

[**2.3. Thuật toán KNN** 28](#_Toc168192802)

[**2.3.1. Giới thiệu về thuật toán KNN** 28](#_Toc168192803)

[**2.3.2. Cách thức hoạt động** 29](#_Toc168192804)

[**2.3.3. Ứng dụng của thuật toán** 30](#_Toc168192805)

[**2.3.4. Ưu điểm** 31](#_Toc168192806)

[**2.3.5. Nhược điểm** 32](#_Toc168192807)

[**2.4. Giới thiệu về Python** 32](#_Toc168192808)

[**2.4.1. Ưu điểm** 33](#_Toc168192809)

[**2.4.2. Nhược điểm** 33](#_Toc168192810)

[**2.5. Một số thư viện cần dùng** 34](#_Toc168192811)

[**2.5.1. Numpy** 34](#_Toc168192812)

[**2.3.2. Keras** 35](#_Toc168192813)

[**2.3.3. Pandas** 37](#_Toc168192814)

[**2.2.4. Scikit-learn** 38](#_Toc168192815)

[CHƯƠNG 3: SỬ DỤNG LSTM ĐỂ THỰC NGHIỆM BÀI TOÁN 40](#_Toc168192816)

[**3.1. Tổng quan về dữ liệu** 40](#_Toc168192817)

[**3.2. Xử lý dữ liệu** 42](#_Toc168192818)

[**3.3. Huấn luyện mô hình** 44](#_Toc168192819)

[**3.4. Kết quả thực nghiệm** 48](#_Toc168192820)

[**3.5. So sánh độ chính xác của mô hình** 51](#_Toc168192821)

[CHƯƠNG 4: XÂY DỰNG VÀ KIỂM THỬ 52](#_Toc168192822)

[**4.1. Xây dựng ứng dụng** 52](#_Toc168192823)

[**4.2. Kiểm thử mô hình** 56](#_Toc168192824)

[**4.2.1. Kế hoạch kiểm thử** 56](#_Toc168192825)

[**4.2.2. Kiểm thử mô hình** 57](#_Toc168192826)

[**4.2.3. Kiểm thử giao diện đăng nhập** 59](#_Toc168192827)

[**4.2.4. Kết quả kiểm thử** 60](#_Toc168192828)

[KẾT LUẬN 61](#_Toc168192829)

[TÀI LIỆU THAM KHẢO 62](#_Toc168192830)

# **DANH MỤC CHỮ VIẾT TẮT**

|  |  |
| --- | --- |
| **Ký hiệu** | **Giải thích** |
| LSTM | Mô hình Long Short – Term Memory |
| RNN | Recurrent Neural Network |
| KNN | K-Nearest Neighbors |

# **DANH MỤC HÌNH ẢNH**

[Hình 1. 1. Hình ảnh về chứng khoán 1](#_Toc168167475)

[Hình 1. 2. Hình ảnh thị trường chứng khoán 7](#_Toc168167476)

[Hình 2. 1. Sự lặp lại kiến trúc module trong mạng RNN chứa một tầng ẩn 15](#_Toc168190869)

[Hình 2. 2. Sự lặp lại kiến trúc module trong mạng LSTM chứa 4 tầng ẩn (3 sigmoid và 1 tanh) tương tác 15](#_Toc168190870)

[Hình 2. 3. Diễn giải các kí hiệu trong đồ thị mạng nơ ron 16](#_Toc168190871)

[Hình 2. 4. Đường đi của ô trạng thái (cell state) trong mạng LSTM 16](#_Toc168190872)

[Hình 2. 5. Một cổng của hàm sigmoid trong LSTM 17](#_Toc168190873)

[Hình 2. 6. Tầng cổng quên (forget gate layer) 18](#_Toc168190874)

[Hình 2. 7. Cập nhật giá trị cho ô trạng thái bằng cách kết hợp 2 kết quả từ tầng cổng vào và tầng ẩn hàm tanh 19](#_Toc168190875)

[Hình 2. 8. Ô trạng thái mới 19](#_Toc168190876)

[Hình 2. 9. Điều chỉnh thông tin ở đầu ra thông qua hàm tanh 20](#_Toc168190877)

[Hình 2. 10. Kết nối ổng tiểu (peehole) liên kết trực tiếp ô trạng thái với các cổng 21](#_Toc168190878)

[Hình 2. 11. Cấu trúc điều chỉnh thêm mới và bỏ vào thông tin đồng thời 21](#_Toc168190879)

[Hình 2. 12. Cấu trúc cổng truy hồi đơn vị (GRU – Gated Recurrent Unit) 22](#_Toc168190880)

[Hình 2. 13. Các dạng bài toán RNN 26](#_Toc168190881)

[Hình 2. 14. Thuật toán KNN 30](#_Toc168190882)

[Hình 3. 1. Hình ảnh dữ liệu được lấy từ Yahoo Finance 40](#_Toc168253674)

[Hình 3. 2. Dữ liệu mã chứng khoán GOOG 42](#_Toc168253675)

[Hình 3. 3. Kết quả thông tin của dữ liệu được lấy 43](#_Toc168253676)

[Hình 3. 4. Hình ảnh kết quả dự đoán giá của mã chứng khoán 7 ngày tới 49](#_Toc168253677)

[Hình 3. 5. Biểu đồ mã chứng khoán trước khi dự đoán 50](#_Toc168253678)

[Hình 3. 6. Biểu đồ mã chứng khoán khi thêm phần dự đoán 50](#_Toc168253679)

[Hình 3. 7. Độ chính xác của mô hình LSTM 51](#_Toc168253680)

[Hình 3. 8. Độ chính xác của mô hình KNN 51](#_Toc168253681)

[Hình 4. 1. Giao diện đăng nhập 52](#_Toc168191790)

[Hình 4. 2. Giao diện đăng kí 53](#_Toc168191791)

[Hình 4. 3. Giao diện lấy lại mật khẩu 54](#_Toc168191792)

[Hình 4. 4. Giao diện ứng dụng dự đoán mã chứng khoán 55](#_Toc168191793)

[Hình 4. 5. Giao diện khi đã nhập mã chứng khoán 56](#_Toc168191794)

# **DANH MỤC BẢNG BIỂU**

[Bảng 4. 1. Bảng kiểm thử mô hình 57](#_Toc168191829)

[Bảng 4. 2. Bảng kiểm thử giao diện đăng nhập 59](#_Toc168191830)

# **LỜI CẢM ƠN**

Lời đầu tiên cho em xin gửi lời cảm ơn đến quý Thầy Cô đang công tác tại khoa Công nghệ thông tin trường Đại Học Công Nghiệp Hà Nội đã tận tình hướng dẫn và truyền đạt cho em những kiến thức, những kinh nghiệm nghề nghiệp để em có hành trang vững chắc bước vào đời. Không những vậy, Thầy Cô đã tạo mọi điều kiện thuận lợi nhất cho em thực hiện đồ án tốt nghiệp của mình trong suốt thời gian vừa qua.

Em xin gửi lời cảm ơn chân thành đến cô Ths Nguyễn Lan Anh đã giúp đỡ, chỉ bảo và hỗ trợ hướng dẫn em trong suốt thời gian làm đồ án tốt nghiệp. Cô đã bỏ rất nhiều thời gian và công sức để giúp cho em có được một đồ án tốt nghiệp để em có thể hoàn thành khóa học của mình một cách tốt nhất.

Trong quá trình làm đồ án tốt nghiệp, cũng như là trong quá trình làm bài báo cáo đồ án tốt nghiệp do trình độ cũng như kinh nghiệm thực tiễn còn hạn chế nên đồ án không thể tránh khỏi những thiếu sót, em rất mong nhận được ý kiến đóng góp của Thầy Cô để em học thêm được nhiều kinh nghiệm và sẽ hoàn chỉnh hơn bài báo cáo đồ án tốt nghiệp của em.

Cuối cùng em kính chúc tất cả các Thầy Cô khoa Công nghệ thông tin của trường Đại Học Công Nghiệp Hà Nội dồi dào sức khỏe, đạt được nhiều thành công trong công việc.

Em xin chân thành cảm ơn!

Sinh viên thực hiện

*Tôn Quang Thuận*

# **LỜI MỞ ĐẦU**

Như mọi người đã biết, thị trường chứng khoán ngày nay được xem là thời kì phát triển mạnh mẽ nhất với hàng trăm công ty được niêm yết trên sàn và giá thay đổi theo từng giây từng phút. Khi dự báo giá chứng khoán, các nhà phân tích thường phải nghiên cứu các loại báo cáo tài chính, nguồn dữ liệu tổng hợp từ nhiều năm trước. Tuy nhiên việc đối mặt với các yếu tố khách quan như là dịch bệnh, thiên tai có thể ảnh hưởng rất lớn tới giá của một nhóm cổ phiếu củ thể.

Thị trường chứng khoán không chỉ dành riêng cho những người giàu kinh nghiệm, những chủ công ty, mà nó còn dành cho những sinh viên đang ngồi trên ghế nhà trường. Nếu có thể cập nhật thông tin, dự đoán chính xác thì bộ môn này có thế đem lại nguồn thu nhập thụ động cho mỗi cá nhân.

Thị trường chứng khoán tuy được nhiều người biết đến nhưng cách khai thác và ra thu nhập của bộ môn này còn nhiều khó khăn. Với mục đích để cải thiện một số khó khăn như là phân tích hướng đi của giá chứng khoán hay dự đoán giá chứng khoán trong tương lai, em đã quyết định lựa chọn bài toán “ Xây dựng ứng dụng dự đoán sự biến thiên của các mã chứng khoán trong thị trường chứng khoán bằng Python”.

Với việc ứng dụng mô hình sử dụng trí tuệ nhân tạo AI vào dự đoán giá chứng khoán có thể giúp người chơi có cái nhìn khách quan và chủ động hơn, với phương pháp này chúng ta có thể dự đoán được giá chứng khoán và xu hướng phát triển của nó như thế nào trong tương lai.

Để thực hiện đề tài này, trước tiên tôi phải tìm hiểu chứng khoán và thị trường chứng khoán hiện nay như thế nào. Sau đó, tôi tiến hành thu thập và tiền xử lý bộ dữ liệu huấn luyện cho các module trí tuệ nhân tạo. Tiếp theo, tôi tập trung vào việc nghiên cứu, phát biểu bài toán dự đoán giá chứng khoán bằng mô hình LSTM; nghiên cứu chuyên sâu về mô hình LSTM; thực nghiệm kết quả và xây dựng chương trình demo.

*Nội dung quyển báo cáo đồ án tốt nghiệp sẽ bao gồm các chương như sau:*

**Chương 1: Tìm hiểu và phát biểu bài toán**

Trong chương 1, tôi tiến hành tìm hiểu về chứng khoán và thị trường chứng khoán là gì? Sau phần khảo sát, tôi trình bày về bài toán được đặt ra là dự đoán sự biến thiên của các mã chứng khoán.

**Chương 2: Tìm hiểu một số thuật toán AI và một số thư viện Python cần dùng**

Sau khi đưa ra một số phương pháp và thuật toán để giải quyết bài toán ở chương 1, ở chương 2 này sẽ tìm hiểu về một số thuật toán đã nhắc đến ở chương 1. Đồng thời tìm hiểu thêm về một số thứ viện và công nghệ cần dùng để thực hiện việc xây dựng chương trình.

**Chương 3: Sử dụng LSTM để thực nghiệm bài toán**

Tại chương 3, tập trung trình bày về quá trình thực nghiệm cũng như các kết quả đạt được với kỹ thuật giải quyết bài toán được đề xuất. Đồng thời tiến hành so sánh kết quả thực nghiệm thu được từ phương pháp đề xuất giải quyết bài với một số phương pháp phổ biến hiện nay và đưa ra nhận xét.

**Chương 4: Xây dựng ứng dụng và kiểm thử**

Để tận dụng kết quả thực nghiệm đã thu được, tiến hành xây dựng sản phẩm demo là một ứng dụng có chức năng chính là dự đoán sự biên thiên của các mã chứng khoán. Rồi kiểm thử mô hình cũng như các chức năng của ứng dụng.

**Phần kết luận:**

Cuối cùng trong phần kết luật, tôi tổng hợp các kết quả đạt được, các hướng phát triển và mở rộng đề tài nghiên cứu trong tương lai.

# **CHƯƠNG 1: TÌM HIỂU VÀ PHÁT BIỂU BÀI TOÁN**

* 1. **Chứng khoán là gì?**

Chứng khoán là một loại tài sản gồm cổ phiếu, chứng chỉ quỹ, chứng khoán phái sinh, trái phiếu, chứng quyền có bảo đảm...

Theo định nghĩa trong Luật Chứng khoán, chứng khoán là tài sản bao gồm cổ phiếu, chứng chỉ quỹ, chứng khoán phái sinh, trái phiếu, chứng quyền, chứng quyền có bảo đảm, quyền mua cổ phần, chứng chỉ lưu ký. Những loại tài sản này có điểm chung là một bằng chứng xác nhận sở hữu hợp pháp của người sở hữu (gọi chung là nhà đầu tư) với tài sản của doanh nghiệp hoặc tổ chức phát hành.



Hình 1. 1. Hình ảnh về chứng khoán

* + 1. ***Cổ phiếu là gì?***

Cổ phiếu được xem là chứng khoán vốn bởi nó xác nhận quyền và lợi ích như như dự họp đại hội đồng cổ đông, biểu quyết, đề cử Hội đồng quản trị, hưởng cổ tức... của người sở hữu với một phần vốn của doanh nghiệp. Đây là loại chứng khoán phổ biến và được giao dịch nhiều nhất trên thị trường.

Tương tự cổ phiếu, chứng chỉ quỹ cũng là chứng khoán vốn nhưng để khẳng định quyền sở hữu của nhà đầu tư với một phần vốn góp của quỹ đầu tư chứng khoán. Bản chất đầu tư vào chứng chỉ quỹ là ủy thác vốn cho chuyên gia của các tổ chức quản lý quỹ.

* Các đặc điểm chính:
* Người phát hành: Có thể là chính phủ, các địa phương hoặc các công ty.
* Giá trị, mệnh giá: Là giá trị gốc của trái phiếu, được trả lại cho nhà đầu tư khi trái phiếu đáo hạn.
* Lãi suất cố định hoặc biến động: Trái phiếu có thể có lãi suất cố định (được xác định và không đổi suốt thời gian tồn tại của trái phiếu) hoặc lãi suất biến động (có thể thay đổi theo một chỉ số tham chiếu nhất định).
* Kỳ hạn: Độ dài thời gian từ khi phát hành trái phiếu cho đến khi trả hết nợ gốc, có thể từ vài năm đến vài chục năm.
* Lợi Ích và Rủi Ro
* Lợi ích: Trái phiếu thường an toàn hơn so với cổ phiếu, đặc biệt là trái phiếu chính phủ, và cung cấp lãi suất thu nhập cố định.
* Rủi ro: Trái phiếu có rủi ro lãi suất (giá trị của trái phiếu giảm khi lãi suất tăng) và rủi ro tín dụng (rủi ro mất vốn nếu người phát hành không thể trả nợ).
* Các Loại Trái Phiếu
* Trái phiếu chính phủ: Phát hành bởi chính phủ quốc gia (như trái phiếu kho bạc Mỹ) hoặc chính quyền địa phương (như trái phiếu thành phố hoặc quận).
* Trái phiếu doanh nghiệp: Phát hành bởi các công ty nhằm mục đích tăng vốn cho các hoạt động kinh doanh.
* Trái phiếu chuyển đổi: Có thể chuyển đổi thành một số lượng cổ phiếu xác định của công ty phát hành trong tương lai.
  + 1. ***Trái phiếu là gì?***

Trái phiếu là loại chứng khoán để xác nhận nghĩa vụ nợ của doanh nghiệp hoặc chính phủ với người nắm giữ (còn gọi là trái chủ). Vì trái phiếu là chứng khoán nợ nên khi doanh nghiệp phá sản hoặc giải thể thì trái chủ được ưu tiên thanh toán trước người nắm giữ cổ phiếu (còn gọi là cổ đông).

* Các Đặc Điểm Chính
* Người phát hành: Có thể là chính phủ, các địa phương hoặc các công ty.
* Giá trị Mệnh giá: Là giá trị gốc của trái phiếu, được trả lại cho nhà đầu tư khi trái phiếu đáo hạn.
* Lãi suất cố định hoặc biến động: Trái phiếu có thể có lãi suất cố định (được xác định và không đổi suốt thời gian tồn tại của trái phiếu) hoặc lãi suất biến động (có thể thay đổi theo một chỉ số tham chiếu nhất định).
* Kỳ hạn: Độ dài thời gian từ khi phát hành trái phiếu cho đến khi trả hết nợ gốc, có thể từ vài năm đến vài chục năm.
* Lợi Ích và Rủi Ro
* Lợi ích: Trái phiếu thường an toàn hơn so với cổ phiếu, đặc biệt là trái phiếu chính phủ, và cung cấp lãi suất thu nhập cố định.
* Rủi ro: Trái phiếu có rủi ro lãi suất (giá trị của trái phiếu giảm khi lãi suất tăng) và rủi ro tín dụng (rủi ro mất vốn nếu người phát hành không thể trả nợ).
* Các Loại Trái Phiếu
* Trái phiếu chính phủ: Phát hành bởi chính phủ quốc gia (như trái phiếu kho bạc Mỹ) hoặc chính quyền địa phương (như trái phiếu thành phố hoặc quận).
* Trái phiếu doanh nghiệp: Phát hành bởi các công ty nhằm mục đích tăng vốn cho các hoạt động kinh doanh.
* Trái phiếu chuyển đổi: Có thể chuyển đổi thành một số lượng cổ phiếu xác định của công ty phát hành trong tương lai.
  + 1. ***Chứng quyền là gì?***

Chứng quyền có bảo đảm là loại chứng khoán do công ty chứng khoán phát hành và được niêm yết với mã giao dịch riêng. Người giữ chứng quyền nhận được quyền mua hoặc quyền bán chứng khoán cơ sở (như cổ phiếu, ETF, chỉ số...) tại mức giá và thời điểm đã được xác định hoặc hưởng khoản chênh lệch giá tại ngày đáo hạn. Giá chứng quyền thường thấp hơn nhiều so với chứng khoán cơ sở nhưng có tính đòn bẩy cao và biết được khoản lỗ tối đa, do đó phù hợp với những nhà đầu tư ưa mạo hiểm.

* Đặc điểm chính của Chứng quyền:
* Tài sản cơ sở: Chứng quyền thường được liên kết với cổ phiếu của một công ty nhưng cũng có thể được liên kết với các loại tài sản khác.
* Giá thực hiện: Đây là giá mà người sở hữu chứng quyền có thể mua hoặc bán tài sản cơ sở.
* Ngày đáo hạn: Ngày cuối cùng mà chứng quyền có thể được thực hiện.
* Các loại Chứng quyền:
* Chứng quyền mua (Call Warrants): Cho phép người nắm giữ mua tài sản cơ sở với giá thực hiện trước khi chứng quyền đáo hạn.
* Chứng quyền bán (Put Warrants): Cho phép người nắm giữ bán tài sản cơ sở với giá thực hiện trước khi chứng quyền đáo hạn.
* Lợi ích và Rủi ro:
* Lợi ích:

+ Đòn bẩy: Chứng quyền cho phép nhà đầu tư sử dụng đòn bẩy để tăng cường tiềm năng sinh lời từ các biến động nhỏ của giá cổ phiếu cơ sở.

+ Lựa chọn: Người mua chứng quyền có quyền nhưng không có nghĩa vụ phải thực hiện quyền mua hoặc bán, cho phép họ quyết định dựa trên điều kiện thị trường.

* Rủi ro:

+ Mất vốn: Nếu chứng quyền không được thực hiện trước ngày đáo hạn, người mua chứng quyền sẽ mất toàn bộ số tiền đã thanh toán để mua chứng quyền.

+ Biến động giá cao: Giá của chứng quyền có thể rất nhạy cảm với các thay đổi về giá của tài sản cơ sở, làm cho nó trở thành một công cụ đầu tư rủi ro cao.

* + 1. ***Đặc điểm của chứng khoán***

Đặc điểm hay thuộc tính của chứng khoán cũng rất đa dạng tùy thuộc vào các loại hình chứng khoán khác nhau. Trong các thuộc tính của chứng khoán thì có 3 thuộc tính chung và nội trội nhất đó là:

* Chứng khoán có tính thanh khoản cao

Tính thanh khoản hay tính lỏng của tài sản là khả năng chuyển đổi tài sản đó thành tiền mặt. Tùy thuộc vào khoảng thời gian và phí cần thiết cho việc chuyển đổi cũng như rủi ro giảm sút giá trị của tài sản đó do chuyển đổi, ta sẽ đánh giá được tính thanh khoản của tài sản đó cao hay thấp. So với các tài sản khác thì chứng khoán có tính thanh khoản cao hơn, thể hiện qua khả năng chuyển đổi cao trên thị trường. Các loại hình chứng khoán khác nhau sẽ có khả năng chuyển đổi là khác nhau.

* Tính rủi ro của chứng khoán: Các loại chứng khoán đều tồn tại những rủi ro tác động đến giá trị của chứng khoán. Có 2 loại rủi ro chính ảnh hưởng đến chứng khoán đó là:
* Rủi ro có hệ thống hay rủi ro thị trường có tác động tới hầu hết hoặc toàn bộ các loại tài sản. Loại rủi ro này chịu sự ảnh hưởng đến từ các điều kiện kinh tế chung như: lạm phát, biến động tỷ giá hối đoái, lãi suất v.v.
* Rủi ro không có hệ thống là loại rủi ro có tác động đến chỉ một hoặc một nhóm nhỏ các tài sản. Loại rủi ro này thông thường sẽ liên quan tới điều kiện của tổ chức phát hành.
* Mức độ rủi ro của các loại chứng khoán được sắp xếp theo thứ tự tăng dần là: Chứng khoán nợ (Trái phiếu), Chứng khoán vốn (Cổ phiếu) và cuối cùng là Chứng khoán phái sinh.
* Tình sinh lời của chứng khoán

Một trong những thuộc tính của chứng khoán được quan tâm nhất khi tìm hiểu về chứng khoán đó là khả năng sinh lời. Tính sinh lời trong chứng khoán là khả năng đem lại lợi nhuận lớn cho nhà đầu tư trong tương lai. Lợi nhuận này có thể thu được từ lợi tức được phân chia hàng năm và việc tăng giá của chứng khoán trên thị trường.

* 1. **Thị trường chứng khoán là gì?**

Thị trường chứng khoán là nơi diễn ra các giao dịch mua bán trao đổi các loại chứng khoán Các quan hệ mua bán trao đổi này làm thay đổi chủ sở hữu của chứng khoán, và như vậy, thực chất đây là quá trình vận động của tư bản chuyển từ tư bản sở hữu sang tư bản kinh doanh.

Thị trường chứng khoán không giống với các thị trường các hàng hóa thông thường khác vì hàng hóa của thị trường chứng khoán là loại hàng hóa đặc biệt, là quyền sở hữu về tư bản. Loại hàng hóa này cũng có giá trị và giá trị sử dụng. Như vậy, có thể nói, bản chất của thị trường chứng khoán là thị trường thể hiện mối quan hệ giữa cung và cầu của vốn đầu tư nào đó, giá cả của chứng khoán chứa đựng thông tin về chi phí vốn hay giá cả của vốn đầu tư. Thị trường chứng khoán là hình thức phát triển bậc cao của nền sản xuất và lưu thông hàng hóa.



Hình 1. 2. Hình ảnh thị trường chứng khoán

* 1. **Vai trò, chức năng của thị trường chứng khoán**

Thị trường chứng khoán có vai trò rất lớn cho nền kinh tế đất nước.

* + 1. ***Huy động vốn cho nên kinh tế***

Khi các nhà đầu tư mua chứng khoán do các công ty phát hành, số tiền nhàn rỗi của họ được đưa vào hoạt động sản xuất kinh doanh và qua đó góp phần mở rộng sản xuất xã hội. Bằng cách hỗ trợ các hoạt động đầu tư của công ty, thị trường chứng khoán đã có những tác động quan trọng đối với sự phát triển của nền kinh tế quốc dân. Thông qua thị trường chứng khoán, chính phủ và chính quyền ở các địa phương cũng huy động được các nguồn vốn cho mục đích sử dụng và đầu tư phát triển hạ tầng kinh tế, phục vụ các nhu cầu chung của xã hội.

* + 1. ***Cung cấp môi trường đầu tư cho công chúng***

Thị trường chứng khoán cung cấp cho công chúng một môi trường đầu tư lành mạnh với các cơ hội lựa chọn phong phú. Các loại chứng khoán trên thị trường rất khác nhau về tính chất, thời hạn và độ rủi ro, cho phép các nhà đầu tư có thể lựa chọn cho loại hàng hóa phù hợp với khả năng, mục tiêu và sở thích của mình. Chính vì vậy, thị trường chứng khoán góp phần đáng kể làm tăng mức tiết kiệm quốc gia.

* + 1. ***Tạo tính thanh khoản cho các chứng khoán***

Nhờ có thị trường chứng khoán các nhà đầu tư có thể chuyển đổi các chứng khoán họ sở hữu thành tiền mặt hoặc các loại chứng khoán khác khi họ muốn. Khả năng thanh khoản (khả năng chuyển đổi thành tiền mặt) là một trong những đặc tính hấp dẫn của chứng khoán đối với người đầu tư. Đây là yếu tố cho thấy tính linh hoạt, an toàn của vốn đầu tư. Thị trường chứng khoán hoạt động càng năng động và có hiệu quả thì càng có khả năng nâng cao tính thanh khoản của các chứng khoán giao dịch trên thị trường.

* + 1. ***Đánh giá hoạt động của doanh nghiệp***

Thông qua giá chứng khoán, hoạt động của các doanh nghiệp được phản ảnh một cách tổng hợp và chính xác, giúp cho việc đánh giá và so sánh hoạt động của các doanh nghiệp được nhanh chóng và thuận tiện, từ đó cũng tạo ra một môi trường cạnh tranh lành mạnh nhằm nâng cao hiệu quả sử dụng vốn, kích thích áp dụng công nghệ mới, cải tiến sản phẩm.

* + 1. ***Tạo môi trường giúp chính phủ thực hiện các chính sách vĩ mô***

Các chỉ báo của thị trường chứng khoán phản ánh động thái của nền kinh tế một cách nhạy bén và chính xác. Giá các chứng khoán tăng lên cho thấy đầu tư đang mở rộng, nền kinh tế tăng trưởng; và ngược lại giá chứng khoán giảm sẽ cho thấy các dấu hiệu tiêu cực của nền kinh tế.

Vì thế, thị trường chứng khoán được gọi là phong vũ biểu của nền kinh tế và là một công cụ quan trọng giúp chính phủ thực hiện các chính sách kinh tế vĩ mô. Thông qua thị trường chứng khoán, chính phủ có thể mua và bán trái phiếu chính phủ để tạo ra nguồn thu bù đắp thâm hụt ngân sách và quản lý lạm phát.

Ngoài ra, chính phủ cũng có thể sử dụng một số chính sách, biện pháp tác động vào thị trường chứng khoán nhằm định hướng đầu tư đảm bảo cho sự phát triển cân đối của nền kinh tế.

* 1. **Phát biểu bài toán**
     1. ***Input***

Dữ liệu lịch sử của các mã chứng khoán bao gồm các thông tin sau:

* Giá mở cửa (Open)
* Giá cao nhất (High)
* Giá thấp nhất (Low)
* Giá đóng cửa (Close)
* Khối lượng giao dịch (Volume)
* Các chỉ số kỹ thuật khác (nếu có)
  + 1. ***Output***

Dự đoán giá đóng cửa của một mã chứng khoán trong một khoảng thời gian nhất định trong tương lai (ví dụ: ngày kế tiếp, tuần kế tiếp, tháng kế tiếp).

* + 1. ***Mục tiêu***
* Xây dựng mô hình dự đoán có thể sử dụng các thông tin đầu vào để dự đoán chính xác giá đóng cửa của mã chứng khoán.
* Mô hình cần phải được tối ưu hóa để giảm thiểu sai số dự đoán và nâng cao độ chính xác.
  + 1. ***Ràng buộc***
* Dữ liệu lịch sử của các mã chứng khoán phải đủ lớn và đủ chi tiết để huấn luyện mô hình.
* Mô hình dự đoán phải có khả năng xử lý các biến động mạnh và bất thường của thị trường chứng khoán.
* Thời gian tính toán và dự đoán của mô hình phải nhanh chóng để có thể áp dụng trong các tình huống thực tế.
  + 1. ***Cách tiếp cận***
* Thu thập và tiền xử lý dữ liệu lịch sử của các mã chứng khoán từ các nguồn đáng tin cậy.
* Áp dụng các kỹ thuật học máy và học sâu để xây dựng mô hình dự đoán.
* Huấn luyện và đánh giá mô hình trên tập dữ liệu kiểm tra.
* Triển khai mô hình dự đoán vào một ứng dụng demo để thử nghiệm và đánh giá hiệu suất.
  + 1. ***Các bước thực hiện***
* **Thu thập dữ liệu:** Lấy dữ liệu lịch sử của các mã chứng khoán từ các nguồn như Yahoo Finance, Alpha Vantage, Quandl.
* **Tiền xử lý dữ liệu:** Làm sạch dữ liệu, xử lý các giá trị thiếu, biến đổi dữ liệu, tạo các đặc trưng mới.
* **Lựa chọn mô hình:** Thử nghiệm với nhiều mô hình khác nhau như Hồi quy tuyến tính, Cây quyết định, Rừng ngẫu nhiên, Mạng nơ-ron hồi quy, Mạng nơ-ron tích chập, LSTM.
* **Huấn luyện và đánh giá:** Huấn luyện các mô hình trên tập dữ liệu huấn luyện và đánh giá hiệu suất trên tập dữ liệu kiểm tra.
* **Triển khai ứng dụng:** Xây dựng ứng dụng web để người dùng có thể nhập mã chứng khoán và xem dự đoán giá.

## **1.5. Khảo sát các phương pháp dự đoán**

### ***1.5.1. Phương pháp thống kê (Statistical Methods)***

Phương pháp thống kê trong dự đoán (statistical methods for prediction) bao gồm nhiều kỹ thuật được sử dụng để dự báo các giá trị tương lai dựa trên dữ liệu lịch sử. Dưới đây là một số phương pháp thông dụng:

* Hồi quy tuyến tính (Linear Regression):

Dự đoán giá trị của một biến phụ thuộc dựa trên giá trị của một hoặc nhiều biến độc lập. Mô hình hồi quy tuyến tính đơn giản sử dụng phương trình , trong đó là biến độc phụ thuộc, là biến độc lập, và là các hệ số hồi quy.

* Hồi quy bội (Multiple Regression):

Mở rộng hồi quy tuyến tính đơn để dự đoán giá trị của một biến phụ thuộc dựa trên nhiều biến độc lập. Mô hình này có dạng

* Phân tích chuỗi thời gian (Time Series Analysis):

Dự đoán các giá trị tương lai dựa trên các giá trị trong quá khứ của cùng một biến. Một số phương pháp phổ biến bao gồm:

+ ARIMA (AutoRegressive Integrated Moving Average): Kết hợp ba thành phần chính là tự hồi quy (AR), tích phân (I), và trung bình động (MA) để dự báo.

+ Exponential Smoothing: Sử dụng trọng số giảm dần theo thời gian để dự báo các giá trị tương lai.

### ***1.5.2. Các mô hình học máy (Machine Learning)***

Các mô hình học máy (machine learning) được sử dụng rộng rãi trong dự đoán vì chúng có khả năng học từ dữ liệu và cải thiện độ chính xác theo thời gian. Dưới đây là một số mô hình học máy phổ biến trong dự đoán:

* Hồi quy tuyến tính (Linear Regression):

Dự đoán một biến liên tục dựa trên giá trị của một hoặc nhiều biến độc lập. Đây là mô hình đơn giản và dễ hiểu.

* Hồi quy logistic (Logistic Regression):

Dự đoán xác suất của một biến phụ thuộc nhị phân (binary dependent variable). Thường được sử dụng trong các bài toán phân loại (classification).

* Cây quyết định (Decision Trees):

Sử dụng cấu trúc cây để phân chia dữ liệu dựa trên các điều kiện, giúp dễ dàng hiểu và diễn giải kết quả.

* Rừng ngẫu nhiên (Random Forest):

Kết hợp nhiều cây quyết định để cải thiện độ chính xác và giảm nguy cơ overfitting. Rừng ngẫu nhiên sử dụng phương pháp bagging để lấy mẫu và xây dựng nhiều cây quyết định.

* Máy vector hỗ trợ (Support Vector Machines - SVM):

Dùng để phân loại và hồi quy, SVM tìm kiếm một siêu phẳng tốt nhất để phân chia dữ liệu trong không gian cao chiều.

* K láng giềng gần nhất (K-Nearest Neighbors - KNN):

Dự đoán giá trị của một điểm dựa trên các giá trị của các điểm láng giềng gần nhất. KNN có thể được sử dụng cho cả phân loại và hồi quy.

* Mạng nơ-ron nhân tạo (Artificial Neural Networks - ANN):

Mô phỏng hoạt động của não người, mạng nơ-ron gồm nhiều tầng (layers) và có thể học các mô hình phức tạp từ dữ liệu lớn.

### ***1.5.3. Các mô hình học sâu (Deep Learning)***

Học sâu (deep learning) là một nhánh của học máy tập trung vào việc sử dụng các mạng nơ-ron nhân tạo với nhiều tầng (layers) để học từ dữ liệu phức tạp. Dưới đây là một số mô hình học sâu phổ biến được sử dụng trong dự đoán:

* Mạng Nơ-ron Dày đặc (Dense Neural Networks - DNN):

Còn được gọi là Mạng nơ-ron feedforward, DNN là loại mạng nơ-ron đơn giản nhất, trong đó các nơ-ron được kết nối đầy đủ giữa các tầng. DNN thường được sử dụng cho các bài toán phân loại và hồi quy cơ bản.

* Mạng Nơ-ron Tích chập (Convolutional Neural Networks - CNN):

CNN được thiết kế đặc biệt để xử lý dữ liệu dạng hình ảnh. Chúng sử dụng các lớp tích chập để tự động học các đặc trưng từ hình ảnh, và sau đó sử dụng các lớp dày đặc để dự đoán.

* Mạng Nơ-ron Hồi quy (Recurrent Neural Networks - RNN):

RNN thích hợp cho dữ liệu tuần tự hoặc chuỗi thời gian, chẳng hạn như văn bản, âm thanh hoặc dữ liệu tài chính. RNN sử dụng các nơ-ron có trạng thái ẩn để nhớ thông tin từ các bước trước đó.

* Long Short-Term Memory (LSTM):

LSTM là một biến thể của RNN được thiết kế để xử lý vấn đề vanishing gradient và có khả năng nhớ thông tin trong thời gian dài. LSTM rất hiệu quả cho các bài toán liên quan đến chuỗi thời gian dài như dịch máy, dự đoán tài chính và nhận dạng giọng nói.

# **CHƯƠNG 2: TÌM HIỂU MỘT SỐ THUẬT TOÁN AI VÀ MỘT SỐ THƯ VIỆN PYTHON CẦN DÙNG**

## **2.1. Thuận toán LSTM**

### ***2.2.1. Giới thiệu về LSTM***

Long Short – Term Memory (LSTM) là một cải tiến của mạng RNN cho phép học các mối tương quan dài mà mạng nơ-ron hồi quy bị hạn chế. Mối tương quan dài được hiều là sự ảnh hưởng của các giá trị bỏ vào ở những thời điểm bắt đầu và các giá trị bỏ ra ở những thời điểm sau.

RNN (Recurrent Neural Network) có thể xử lý thông tin dạng chuỗi (Sequence/ Time-series). RNN có thể mang thông tin của frame(ảnh) từ state trước tới các state sau, rồi ở state cuối là sự kết hợp của tất cả các ảnh để dự đoán hành động.

Ví dụ bài toán là dự đoán từ tiếp theo trong đoạn văn. Đoạn đầu tiên “Mặt trời mọc ở hướng …”, ta có thể chỉ sử dụng các từ trước trong câu để đoán là đông. Tuy nhiên, với đoạn, “Tôi là người Việt Nam. Tôi đang sống ở nước ngoài. Tôi có thể nói trôi chảy tiếng …” thì rõ ràng là chỉ sử dụng từ trong câu đấy hoặc câu trước là không thể dự đoán được từ cần điền là Việt. Ta cần các thông tin từ state ở trước đó rất xa => cần long term memory điều mà RNN không làm được => Cần một mô hình mới để giải quyết vấn đề này => Long short term memory (LSTM) ra đời.

### ***2.2.2. Mạng trí nhớ ngắn hạn định hướng dài hạn (LSTM – Long Short-Term Memory)***

Mạng trí nhớ ngắn hạn định hướng dài hạn còn được viết tắt là LSTM làm một kiến trúc đặc biệt của RNN(Recurrent Neural Networ) có khả năng học được sự phụ thuộc trong dài hạn (long-term dependencies). Kiến trúc này đã được phổ biến và sử dụng rộng rãi cho tới ngày nay. LSTM đã tỏ ra khắc phục được rất nhiều những hạn chế của RNN trước đây về triệt tiêu đạo hàm. Tuy nhiên cấu trúc của chúng có phần phức tạp hơn mặc dù vẫn giữ được tư tưởng chính của RNN là sự sao chép các kiến trúc theo dạng chuỗi. Một mạng RNN tiêu chuẩn sẽ có kiến trúc rất đơn giản chẳng hạn như đối với kiến trúc gồm có một tầng ẩn là hàm tanh như bên dưới.

A screenshot of a computer

Description automatically generated

Hình 2. 1. Sự lặp lại kiến trúc module trong mạng RNN chứa một tầng ẩn

LSTM cũng có một chuỗi dạng như thế nhưng phần kiến trúc lặp lại có cấu trúc khác biệt hơn. Thay vì chỉ có một tầng đơn, chúng có tới 4 tầng ẩn (3 tầng sigmoid và 1 tanh) tương tác với nhau theo cấu trúc đặc biệt.

A screenshot of a computer

Description automatically generated

Hình 2. 2. Sự lặp lại kiến trúc module trong mạng LSTM chứa 4 tầng ẩn (3 sigmoid và 1 tanh) tương tác

Các kí hiệu có thể diễn giải như sau:

A black arrow pointing to the right

Description automatically generated

Hình 2. 3. Diễn giải các kí hiệu trong đồ thị mạng nơ ron

Trong sơ đồ tính toán trên, mỗi một phép tính sẽ triển khai trên một véc tơ. Trong đó hình tròn màu hồng biển diễn một toán tử đối với véc tơ như phép cộng véc tơ, phép nhân vô hướng các véc tơ. Màu vàng thể hiện hàm activation mà mạng nơ ron sử dụng để học trong tầng ẩn, thông thường là các hàm phi tuyến sigmoid và tanh. Kí hiệu 2 đường thẳng nhập vào thể hiện phép chập kết quả trong khi kí hiệu 2 đường thẳng rẽ nhánh thể hiện cho nội dung véc tơ trước đó được sao chép để đi tới một phần khác của mạng nơ ron.

### ***2.2.2. Ý tưởng đằng sau LSTM***

Ý tưởng chính của LSTM là thành phần ô trạng thái (cell state) được thể hiện qua đường chạy ngang qua đỉnh đồ thị như hình vẽ bên dưới:

A diagram of a diagram

Description automatically generated

Hình 2. 4. Đường đi của ô trạng thái (cell state) trong mạng LSTM

Ô trạng thái là một dạng băng chuyền chạy thẳng xuyên suốt toàn bộ chuỗi với chỉ một vài tương tác tuyến tính nhỏ giúp cho thông tin có thể truyền dọc theo đồ thị mạng nơ ron ổn định.

LSTM có khả năng xóa và thêm thông tin vào ô trạng thái và điều chỉnh các luồng thông tin này thông qua các cấu trúc gọi là cổng.

Cổng là cơ chế đặc biệt để điều chỉnh luồng thông tin đi qua. Chúng được tổng hợp bởi một tầng ẩn của hàm activation sigmoid và với một toán tử nhân như đồ thị.

A black background with a yellow and pink rectangle with a black letter

Description automatically generated

Hình 2. 5. Một cổng của hàm sigmoid trong LSTM

Hàm sigmoid sẽ cho đầu ra là một giá trị xác xuất nằm trong khoảng 0 đến 1, thể hiện rằng có bao nhiêu phần thông tin sẽ đi qua cổng. Giá trị bằng 0 ngụ ý rằng không cho phép thông tin nào đi qua, giá trị bằng 1 sẽ cho toàn bộ thông tin đi qua.

Một mạng LSTM sẽ có 3 cổng có kiến trúc dạng này để bảo vệ và kiếm soát các ô trạng thái.

### ***2.2.3. Thứ tự các bước của LSTM***

Bước đầu tiên trong LSTM sẽ quyết định xem thông tin nào chúng ta sẽ cho phép đi qua ô trạng thái (cell state). Nó được kiểm soát bởi hàm sigmoid trong một tầng gọi là tầng quên (forget gate layer). Đầu tiên nó nhận đầu vào là 2 giá trị và và trả về một giá trị nằm trong khoảng 0 và 1 cho mỗi giá trị của ô trạng thái . Nếu giá trị bằng 1 thể hiện ‘giữ toàn bộ thông tin’ và bằng 0 thể hiện ‘bỏ qua toàn bộ chúng’.

Trở lại ví dụ về ngôn ngữ, chúng ta đang cố gắng dự báo từ tiếp theo dựa trên toàn bộ những từ trước đó. Trong những bài toán như vậy, ô trạng thái có thể bao gồm loại của chủ ngữ hiện tại, để cho đại từ ở câu tiếp theo được sử dụng chính xác. Chẳng hạn như chúng đang mô tả về một người bạn là con trai thì các đại từ nhân xưng ở tiếp theo phải là anh, thằng, hắn thay vì cô, con ấy. Tuy nhiên chủ ngữ không phải khi nào cũng cố định. Khi chúng ta nhìn thấy một chủ ngữ mới, chúng ta muốn quên đi loại của một chủ ngữ cũ. Do đó tầng quên cho phép cập nhật thông tin mới và lưu trữ giá trị của nó khi có thay đổi theo thời gian.

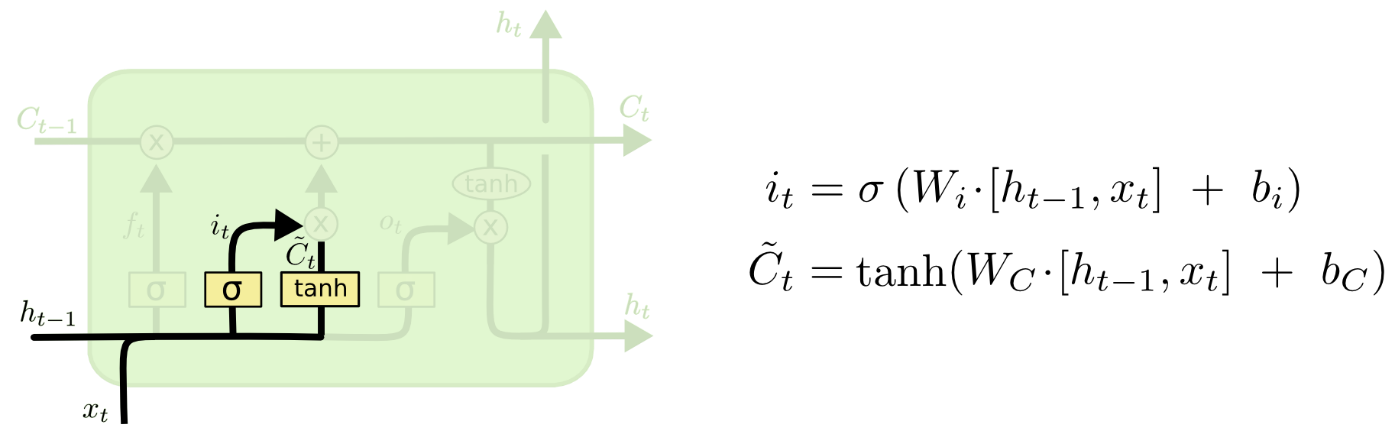
A diagram of a rectangular object with arrows pointing to the side

Description automatically generated

Hình 2. 6. Tầng cổng quên (forget gate layer)

Bước tiếp theo chúng sẽ quyết định loại thông tin nào sẽ được lưu trữ trong ô trạng thái. Bước này bao gồm 2 phần. Phần đầu tiên là một tầng ẩn của hàm sigmoid được gọi là tầng cổng vào (input gate layer) quyết định giá trị bao nhiêu sẽ được cập nhật. Tiếp theo, tầng ẩn hàm tanh sẽ tạo ra véc tơ của một giá trị trạng thái mới mà có thể được thêm vào trạng thái. Tiếp theo kết hợp kết quả của 2 tầng này để tạo thành một cập nhật cho trạng thái.

Trong ví dụ của mô hình ngôn ngữ, chúng ta muốn thêm loại của một chủ ngữ mới vào ô trạng thái để thay thế phần trạng thái cũ muốn quên đi.



Hình 2. 7. Cập nhật giá trị cho ô trạng thái bằng cách kết hợp 2 kết quả từ tầng cổng vào và tầng ẩn hàm tanh

Đây là thời điểm để cập nhật một ô trạng thái cũ, sang một trạng thái mới . Những bước trước đó đã quyết định làm cái gì, và tại bước này chỉ cần thực hiện nó.

Chúng ta nhận trạng thái cũ với tương ứng với việc quên những thứ quyết định được phép quên sớm. Phần tử đề cử là một giá trị mới được tính toán tương ứng với bao nhiêu được cập nhật vào mỗi giá trị trạng thái.

A diagram of a rectangular object with arrows pointing up

Description automatically generated

Hình 2. 8. Ô trạng thái mới

Cuối cùng cần quyết định xem đầu ra sẽ trả về bao nhiêu. Kết quả ở đầu ra sẽ dựa trên ô trạng thái, nhưng sẽ là một phiên bản được lọc. Đầu tiên, chúng ta chạy qua một tầng sigmoid nơi quyết định phần nào của ô trạng thái sẽ ở đầu ra. Sau đó, ô trạng thái được đưa qua hàm tanh (để chuyển giá trị về khoảng -1 và 1) và nhân nó với đầu ra của một cổng sigmoid, do đó chỉ trả ra phần mà chúng ta quyết định.

A diagram of a graphing diagram

Description automatically generated with medium confidence

Hình 2. 9. Điều chỉnh thông tin ở đầu ra thông qua hàm tanh

### ***2.2.4. Các biển thể của LSTM***

Những gì mà chúng ta vừa mổ tả cho đến giờ là một mạng LSTM rất thông thường. Nhưng không phải toàn bộ LSTM đều tương tự như trên. Trên thực tế, có vẻ như hầu hết mọi bài báo liên quan đến LSTM đều sử dụng những version khác nhau đôi chút. Sự khác biệt là rất nhỏ nhưng rất đáng để đề cập một ít trong số nhứng kiến trúc này.

Một trong những biến thể nối tiếng nhất của LSTM được giới thiệu bởi Gers & Schmidhuber (2000) thêm một kết nối ống tiểu (peehole connection) để các cổng có thể kết nối trực tiếp đến các ô trạng thái.

A rectangular object with arrows pointing to the center

Description automatically generated

Hình 2. 10. Kết nối ổng tiểu (peehole) liên kết trực tiếp ô trạng thái với các cổng

Một biến thể khác là sử dụng cặp đôi cổng vào và cổng ra. Thay vì quyết định riêng rẽ bỏ qua thông tin nào và thêm mới thông tin nào, chúng ta sẽ quyết định chúng đồng thời. Các thông tin chỉ bị quên khi chúng ta muốn cập nhập vào một vài thông tin mới.

A diagram of a rectangular object with arrows

Description automatically generated

Hình 2. 11. Cấu trúc điều chỉnh thêm mới và bỏ vào thông tin đồng thời

Một dạng biến thể khá mạnh khác của LSTM là cổng truy hồi đơn vị (Gated Recurrent Unit - GRU) được giới thiệu bởi Cho, et al. (2014). Nó kết hợp cổng quên và cổng vào thành một cổng đơn gọi là cập nhật (update gate). Nó cũng nhập các ô trạng thái và trạng thái ẩn và thực hiện một số thay đổi khác. Kết quả của mô hình đơn giản hơn nhiều so với mô hình LSTM chuẩn, và đã trở nên khá phổ biến.

A close-up of a sign

Description automatically generated

Hình 2. 12. Cấu trúc cổng truy hồi đơn vị (GRU – Gated Recurrent Unit)

### ***2.2.5. Ưu điểm***

* Giữ thông tin lâu dài: LSTM có khả năng lưu trữ thông tin trong thời gian dài mà không bị biến mất, nhờ vào cơ chế của các cổng (gate) như cổng quên (forget gate), cổng vào (input gate) và cổng ra (output gate).
* Khả năng xử lý chuỗi dài: So với RNN thông thường, LSTM có khả năng xử lý chuỗi dài mà không gặp phải vấn đề gradient biến mất (vanishing gradient problem).
* Hiệu suất trong việc học dài hạn: Nhờ vào khả năng giữ thông tin qua nhiều bước thời gian, LSTM thích hợp cho các nhiệm vụ cần xử lý dữ liệu dài hạn và nhận dạng mẫu phức tạp.
* Đa dạng hóa cấu trúc: LSTM cho phép cấu hình nhiều lớp, các cổng và kích thước của mỗi lớp, từ đó có thể tạo ra các kiến trúc mô hình phức tạp và phù hợp với nhiều loại dữ liệu.

### ***2.2.6. Nhược điểm***

* Tính tính toán phức tạp: LSTM có thể đòi hỏi nhiều tài nguyên tính toán hơn so với các mô hình nơ-ron truyền thống, đặc biệt là khi mạng trở nên rất sâu và phức tạp.
* Khó điều chỉnh tham số: Với các mô hình LSTM có kích thước lớn và phức tạp, việc điều chỉnh các tham số trở nên khó khăn và tốn kém.
* Có thể dễ dàng bị overfitting: Như các mô hình nơ-ron khác, LSTM cũng dễ bị overfitting nếu không được điều chỉnh và kiểm soát đúng cách, đặc biệt là khi dữ liệu huấn luyện có kích thước nhỏ.
* Khó hiểu: Với những người mới bắt đầu, LSTM có thể khó hiểu do có nhiều cơ chế và cổng cần phải hiểu rõ để áp dụng hiệu quả.

## **2.2. Thuật toán RNN**

### ***2.2.1. Giới thiệu về RNN***

Recurrent Neural Networks (RNN) là một loại mạng nơ-ron nhân tạo đặc biệt, có khả năng xử lý dữ liệu tuần tự và duy trì thông tin từ các bước trước đó thông qua việc sử dụng các vòng lặp trong cấu trúc mạng. Khác với các mạng nơ-ron truyền thống (feedforward neural networks), RNN có các kết nối ngược cho phép thông tin lưu giữ từ trạng thái trước đó để ảnh hưởng đến trạng thái hiện tại. Điều này làm cho RNN trở nên lý tưởng cho các nhiệm vụ liên quan đến dữ liệu tuần tự như xử lý ngôn ngữ tự nhiên, dịch máy, nhận dạng giọng nói và phân tích chuỗi thời gian.

### ***2.2.2. Dạng của Linear Regression***

Một hàm số đơn giản (1)

Trong đó, là các hằng số, còn được gọi là bias. Mối quan hệ bên trên là một mối quan hệ tuyến tính (linear).

Trong phương trình (1), nếu chúng ta đặt = là vector (cột) hệ số cần phải tối ưu và (đọc là x *bar* trong Tiếng Anh) là vector (hàng) dữ liệu đầu vào mở rộng. Số 1 ở đầu được thêm vào để phép tính đơn giản và thuận tiện cho việc tính toán. Khi đó, phương trình (1) có thể viết lại dưới dạng:

Chú ý rằng là một vector hàng.

### ***2.2.3. Sai số dự đoán***

Chúng ta mong muốn rằng sự sai khác e giữa giá trị dự đoán (đọc là y hat trong tiếng Anh) là nhỏ nhất. Nói cách khác, chúng ta muốn giá trị sau đây càng nhỏ càng tốt:

Trong đó hệ số (lại) là để thuận tiện cho việc tính toán (khi tính đạo hàm thì số sẽ bị triệt tiêu). Chúng ta cần vì có thể là một số âm, việc nói e nhỏ nhất sẽ không đúng vì khi là rất nhỏ nhưng sự sai lệch là rất lớn.

### ***2.2.4. Hàm mất mát***

Điều tương tự xảy ra với tất cả các cặp *(input, outcome)* (, với N là số lượng dữ liệu quan sát được. Điều chúng ta muốn, tổng sai số là nhỏ nhất, tương đương với việc tìm **w** để hàm số sau đạt giá trị nhỏ nhất:

(2)

Hàm số được gọi là hàm mất mát (loss function) của bài toán Linear Regression. Chúng ta mong muốn rằng sự mất mát (sai số) là nhỏ nhất, điều đó đồng nghĩa với việc tìm vector hệ số **w** sao cho giá trị của hàm mất mát này càng nhỏ càng tốt. Giá trị của **w** làm cho hàm mất mát đạt giá trị nhỏ nhất được gọi là điểm tối ưu (optimal point), ký hiệu:

Trước khi đi tìm lời giải, chúng ta đỏn giản hóa phép toán trong phương trình hàm mất mát (2). Đặt là vector cột chứa tất cả các output của training data; là ma trận dữ liệu đầu vào (mở rộng) mà mỗi hàng của nỏ là một điểm dữ liệu. Khi đó hàm số mất mát được viết dưới dạng ma trận đơn giản hơn:

= (3)

Với là Euclidean norm (chuẩn Eulid, hay khoảng cách Euclid), nói cách khác là tổng của bình phương mỗi phần tử của vector **z**. Tới đây, ta đã có một dạng đơn giản của hàm mất mát được viết như phương trình (3).

### ***2.2.5. Nghiệm cho bài toàn Liearn Regression***

Cách phổ biến nhất để tìm nghiệm cho một bài toàn tối ưu (chúng ta đã biết từ khi học cấp 3) là giải phương trình đạo hàm (gradient) bằng 0! Tất nhiên đó là khi việc tính đạo hàm và việc giải phương trình đạo hàm bằng 0 không quá phức tạp. Thật may mắn, với các mô hình tuyến tính, hai việc này là khả thi.

Đạo hàm theo **w** của hàm mất mát là:

Phương trình đạo hàm bằng 0 tương đương với:

*(4)*

(Kí hiệu nghĩa là đặt ).

Nếu ma trận vuông khả nghịch (non-singular hay invertible) thì phương trình (4) có nghiệm duy nhất: .

Vậy nếu ma trận A không khả nghịch (có định thức bằng 0) thì sao? Trong trường hợp này thì hoặc phương trình (4) vô nghiệm, hoặc là nó có vô số nghiệm. Khi đó, chúng ta sử dụng khái niệm giả nghịch đảo (đọc là A dagger trong tiếng Anh). Với khái niệm giả nghịch đảo, điểm tối ưu của bài toán Linear Regression có dạng:

### ***2.2.6. Phân loại bài toán RNN***

A diagram of a number of rectangular objects

Description automatically generated with medium confidence

Hình 2. 13. Các dạng bài toán RNN

* **One to one**: mẫu bài toán cho Neural Network (NN) và Convolutional Neural Network (CNN), 1input và 1 output, ví dụ với CNN input là ảnh và uotput là ảnh dduocj segment.
* **One to many**: bài toán có 1 input nhưng nhiều output, ví dụ: bài toán caption cho ảnh, input là 1 ảnh nhưng output là nhiều chữ mô tả cho ảnh đấy, dưới dạng một câu.
* **Many to one**: bài toán có nhiều input nhưng chỉ có 1 output, ví dụ bài toán phân loại hành động trong video, input là nhiều ảnh (frame) tách ra từ video, ouptut là hành động trong video
* **Many to many**: bài toán có nhiều input và nhiều output, ví dụ bài toán dịch từ tiếng anh sang tiếng việt, input là 1 câu gồm nhiều chữ: “I love Vietnam” và output cũng là 1 câu gồm nhiều chữ “Tôi yêu Việt Nam”.

### ***2.2.7. Ứng dụng bài toán RNN***

Về cơ bản nếu bạn thấy sequence data hay time-series data và bạn muốn áp dụng deep learning thì bạn nghĩ ngay đến RNN. Dưới đây là một số ứng dụng của RNN:

* Speech to text: Chuyển giọng nói sang text.
* Sentiment classification: phân loại số sao cho các bình luận, ví dụ: input: “ứng dụng tốt”, output: 4 sao.
* Machine translation: Bài toán dịch tự động giữa các ngôn ngữ.
* Video recognition: Nhận diện hành động trong video.
* Heart attack: Dự đoán đột quỵ tim.
* ….

### ***2.2.8. Ưu điểm***

* Xử Lý Dữ Liệu Tuần Tự: RNN có khả năng xử lý dữ liệu tuần tự, như chuỗi thời gian hoặc văn bản, nhờ vào cấu trúc có các kết nối ngược, cho phép thông tin từ các bước trước ảnh hưởng đến các bước sau.
* Bộ Nhớ Ngắn Hạn: RNN có khả năng duy trì bộ nhớ ngắn hạn về thông tin trước đó, giúp ích trong các ứng dụng cần nhận dạng các mẫu liên quan trong chuỗi dữ liệu.
* Ứng Dụng Rộng Rãi: RNN được ứng dụng rộng rãi trong nhiều lĩnh vực như xử lý ngôn ngữ tự nhiên, nhận dạng giọng nói, phân tích chuỗi thời gian, và nhiều ứng dụng khác.

### ***2.2.9. Nhược điểm***

* Vanishing/Exploding Gradient Problem: Khi độ dài của chuỗi dữ liệu tăng lên, RNN có thể gặp phải vấn đề gradient biến mất hoặc gradient bùng nổ trong quá trình học, làm giảm hiệu quả của việc học và khiến mô hình khó cập nhật trọng số một cách chính xác.
* Bộ Nhớ Hạn Chế: RNN thông thường có khả năng lưu trữ và truy xuất thông tin từ các bước thời gian trước đó chỉ trong khoảng ngắn, hạn chế trong việc xử lý các chuỗi dài hoặc thông tin phụ thuộc xa.
* Tính Toán Phức Tạp: Việc tính toán gradient cho từng bước thời gian thông qua quá trình lan truyền ngược qua thời gian (BPTT) có thể rất phức tạp và tốn nhiều tài nguyên tính toán.
* Hiệu Suất Không Cao Với Dữ Liệu Dài Hạn: Đối với các tác vụ yêu cầu duy trì thông tin dài hạn, RNN thông thường không hoạt động hiệu quả, và cần sử dụng các biến thể như LSTM hoặc GRU để cải thiện.

## **2.3. Thuật toán KNN**

### ***2.3.1. Giới thiệu về thuật toán KNN***

(K-NN) là một thuật toán học máy linh hoạt và được sử dụng rộng rãi, chủ yếu được sử dụng vì tính đơn giản và dễ thực hiện của nó. Nó không yêu cầu bất kỳ giả định nào về phân phối dữ liệu cơ bản. Nó cũng có thể xử lý cả dữ liệu số và dữ liệu phân loại, khiến nó trở thành lựa chọn linh hoạt cho nhiều loại tập dữ liệu khác nhau trong các nhiệm vụ phân loại và hồi quy. Đây là một phương pháp phi tham số đưa ra dự đoán dựa trên sự giống nhau của các điểm dữ liệu trong một tập dữ liệu nhất định. K-NN ít nhạy cảm hơn với các ngoại lệ so với các thuật toán khác.

Thuật toán K-NN hoạt động bằng cách tìm K lân cận gần nhất với một điểm dữ liệu nhất định dựa trên thước đo khoảng cách, chẳng hạn như khoảng cách Euclide. Lớp hoặc giá trị của điểm dữ liệu sau đó được xác định bằng đa số phiếu bầu hoặc mức trung bình của K hàng xóm. Cách tiếp cận này cho phép thuật toán thích ứng với các mẫu khác nhau và đưa ra dự đoán dựa trên cấu trúc cục bộ của dữ liệu.

### ***2.3.2. Cách thức hoạt động***

Thuật toán K-Nearest Neighbors (KNN) hoạt động theo nguyên tắc tương tự, trong đó nó dự đoán nhãn hoặc giá trị của điểm dữ liệu mới bằng cách xem xét nhãn hoặc giá trị của K lân cận gần nhất của nó trong tập dữ liệu huấn luyện.

Để đưa ra dự đoán, thuật toán tính toán khoảng cách giữa mỗi điểm dữ liệu mới trong tập dữ liệu thử nghiệm và tất cả các điểm dữ liệu trong tập dữ liệu huấn luyện. Khoảng cách Euclide là thước đo khoảng cách thường được sử dụng trong K-NN, nhưng các thước đo khoảng cách khác, chẳng hạn như khoảng cách Manhattan hoặc khoảng cách Minkowski, cũng có thể được sử dụng tùy thuộc vào vấn đề và dữ liệu. Khi khoảng cách giữa điểm dữ liệu mới và tất cả các điểm dữ liệu trong tập dữ liệu huấn luyện được tính toán, thuật toán sẽ tiến hành tìm K hàng xóm gần nhất dựa trên những khoảng cách này. Phương pháp cụ thể để chọn các lân cận gần nhất có thể khác nhau, nhưng cách tiếp cận phổ biến là sắp xếp khoảng cách theo thứ tự tăng dần và chọn K điểm dữ liệu có khoảng cách ngắn nhất.

Sau khi xác định K hàng xóm gần nhất, thuật toán đưa ra dự đoán dựa trên nhãn hoặc giá trị liên quan đến những hàng xóm này. Đối với các nhiệm vụ phân loại, lớp đa số trong số K hàng xóm được gán làm nhãn dự đoán cho điểm dữ liệu mới. Đối với các tác vụ hồi quy, giá trị trung bình hoặc trung bình có trọng số của các giá trị K lân cận được gán làm giá trị dự đoán.

Giả sử X là tập dữ liệu huấn luyện có n điểm dữ liệu, trong đó mỗi điểm dữ liệu được biểu thị bằng vectơ đặc trưng d chiều X\_i và Y là nhãn hoặc giá trị tương ứng cho từng điểm dữ liệu trong X. Cho một điểm dữ liệu mới x, thuật toán sẽ tính toán khoảng cách giữa x và mỗi điểm dữ liệu X\_i trong X bằng cách sử dụng thước đo khoảng cách, chẳng hạn như khoảng cách Euclide:



Thuật toán chọn K điểm dữ liệu từ X có khoảng cách ngắn nhất đến x. Đối với các nhiệm vụ phân loại, thuật toán gán nhãn y thường xuyên nhất trong số K lân cận gần nhất với x. Đối với các tác vụ hồi quy, thuật toán tính trung bình hoặc trung bình có trọng số của các giá trị y của K lân cận gần nhất và gán nó làm giá trị dự đoán cho x.

A black and white image of a circle and a circle with small black triangles

Description automatically generated with medium confidence

Hình 2. 14. Thuật toán KNN

### ***2.3.3. Ứng dụng của thuật toán***

Xử lý trước dữ liệu - Trong khi xử lý bất kỳ vấn đề nào về Machine Learning, trước tiên chúng tôi thực hiện phần EDA , trong đó nếu chúng tôi thấy rằng dữ liệu chứa các giá trị bị thiếu thì cũng có nhiều phương pháp tính toán khả dụng. Một trong những phương pháp như vậy là KNN Imputer , một phương pháp quảng cáo khá hiệu quả thường được sử dụng cho các phương pháp tính toán phức tạp.

Nhận dạng mẫu – Thuật toán KNN hoạt động rất tốt nếu bạn đã huấn luyện thuật toán KNN bằng cách sử dụng bộ dữ liệu MNIST và sau đó thực hiện quy trình đánh giá thì chắc hẳn bạn đã thấy rằng độ chính xác quá cao.

Công cụ đề xuất - Nhiệm vụ chính được thực hiện bởi thuật toán KNN là gán một điểm truy vấn mới cho một nhóm tồn tại trước đã được tạo bằng cách sử dụng một kho dữ liệu khổng lồ. Đây chính xác là những gì được yêu cầu trong hệ thống gợi ý để chỉ định mỗi người dùng vào một nhóm cụ thể và sau đó cung cấp cho họ các đề xuất dựa trên sở thích của nhóm đó.

### ***2.3.4. Ưu điểm***

* Đơn giản và dễ hiểu: Thuật toán KNN rất dễ hiểu và triển khai. Nó không yêu cầu các giả định phức tạp về dữ liệu.
* Không yêu cầu huấn luyện mô hình: KNN là thuật toán không cần huấn luyện trước (lazy learning). Điều này có nghĩa là không cần giai đoạn huấn luyện phức tạp như các thuật toán khác.
* Linh hoạt: KNN có thể được sử dụng cho cả bài toán phân loại và hồi quy. Nó cũng hoạt động tốt trên cả dữ liệu số và dữ liệu phân loại.
* Không yêu cầu giả thiết về phân phối dữ liệu: KNN không yêu cầu dữ liệu phải tuân theo bất kỳ phân phối nào, do đó nó linh hoạt trong nhiều loại dữ liệu khác nhau.
* Khả năng mở rộng: KNN có thể dễ dàng điều chỉnh độ phức tạp của mô hình thông qua tham số K. Điều này cho phép tìm được cân bằng giữa độ chính xác và độ phức tạp.

### ***2.3.5. Nhược điểm***

* Tốn kém về tính toán: KNN yêu cầu tính toán khoảng cách từ điểm cần dự đoán tới tất cả các điểm trong tập dữ liệu huấn luyện, điều này có thể rất tốn kém về mặt tính toán đối với các tập dữ liệu lớn.
* Nhạy cảm với dữ liệu nhiễu và ngoại lai: KNN có thể bị ảnh hưởng mạnh bởi các điểm dữ liệu ngoại lai hoặc nhiễu trong tập dữ liệu, điều này có thể dẫn đến dự đoán không chính xác.
* Hiệu suất kém trên dữ liệu nhiều chiều: Khi số chiều của dữ liệu tăng cao, khoảng cách giữa các điểm dữ liệu trở nên ít phân biệt hơn, điều này có thể làm giảm hiệu suất của KNN (curse of dimensionality).
* Yêu cầu bộ nhớ lớn: Vì KNN cần lưu trữ toàn bộ tập dữ liệu huấn luyện để tính toán khoảng cách, nó có thể yêu cầu một lượng lớn bộ nhớ, đặc biệt với các tập dữ liệu lớn.
* Khó khăn trong việc chọn giá trị K: Việc chọn giá trị K phù hợp là một vấn đề quan trọng. Giá trị K quá nhỏ có thể dẫn đến mô hình bị overfitting, trong khi giá trị K quá lớn có thể dẫn đến mô hình bị underfitting.
* Không cung cấp thông tin về tầm quan trọng của đặc trưng: KNN không có cơ chế để đánh giá tầm quan trọng của các đặc trưng khác nhau trong dữ liệu. Điều này có thể là hạn chế khi làm việc với dữ liệu có nhiều đặc trưng.

## **2.4. Giới thiệu về Python**

Python là một ngôn ngữ lập trình bậc cao, thông dịch và đa mục đích, được phát triển bởi Guido van Rossum và ra mắt lần đầu vào năm 1991. Python nổi tiếng với cú pháp rõ ràng và dễ đọc, điều này giúp các lập trình viên dễ dàng học và sử dụng.

### ***2.4.1. Ưu điểm***

* Cú pháp đơn giản và dễ đọc: Python có cú pháp gần gũi với ngôn ngữ tự nhiên, giúp các lập trình viên dễ dàng học và viết mã. Điều này cũng giúp giảm thời gian phát triển và bảo trì mã nguồn.
* Thư viện và framework phong phú: Python có một thư viện tiêu chuẩn rất mạnh mẽ và phong phú. Ngoài ra, có rất nhiều thư viện bên thứ ba hỗ trợ các lĩnh vực khác nhau như khoa học dữ liệu (NumPy, Pandas), học máy (TensorFlow, Scikit-learn), và phát triển web (Django, Flask).
* Đa nền tảng: Python có thể chạy trên nhiều hệ điều hành khác nhau như Windows, macOS, Linux, và Unix. Giúp các lập trình viên phát triển ứng dụng một cách linh hoạt và dễ dàng triển khai trên nhiều môi trường.
* Hỗ trợ nhiều phong cách lập trình: Python hỗ trợ lập trình thủ tục, lập trình hướng đối tượng, và lập trình hàm, cung cấp sự linh hoạt trong việc thiết kế và triển khai ứng dụng.
* Cộng đồng lớn và tích cực: Python có một cộng đồng người dùng rất lớn, cung cấp rất nhiều tài liệu, hướng dẫn, và hỗ trợ trực tuyến. Các hội nghị và sự kiện như PyCon tạo cơ hội cho các lập trình viên trao đổi kiến thức và kinh nghiệm.
* Tương tác với các ngôn ngữ khác: Python có thể dễ dàng tích hợp và tương tác với các ngôn ngữ lập trình khác như C, C++, và Java, giúp mở rộng khả năng sử dụng.

### ***2.4.2. Nhược điểm***

* Tốc độ thực thi: Python là ngôn ngữ thông dịch, do đó, tốc độ thực thi có thể chậm hơn so với các ngôn ngữ biên dịch như C++ hay Java. Điều này có thể trở thành vấn đề đối với các ứng dụng yêu cầu hiệu suất cao.
* Sử dụng bộ nhớ: Python sử dụng nhiều bộ nhớ hơn do bản chất của nó là ngôn ngữ cấp cao và thông dịch. Điều này có thể trở thành hạn chế khi phát triển các ứng dụng lớn hoặc trong các môi trường hạn chế tài nguyên.
* Thiếu kiểm tra kiểu tĩnh: Python là ngôn ngữ động, không kiểm tra kiểu tĩnh, điều này có thể dẫn đến lỗi thời gian chạy. Việc thiếu kiểm tra kiểu tĩnh cũng có thể làm cho việc duy trì mã nguồn lớn trở nên khó khăn hơn.
* Ứng dụng trong các tác vụ đa luồng: GIL (Global Interpreter Lock) của Python làm hạn chế khả năng thực hiện các tác vụ đa luồng hiệu quả trong Python. Điều này có thể là một vấn đề đối với các ứng dụng cần thực hiện nhiều tác vụ đồng thời.
* Không thích hợp cho một số loại ứng dụng: Python không phải là lựa chọn tốt nhất cho các ứng dụng di động hoặc các ứng dụng có yêu cầu thời gian thực cao do hạn chế về hiệu suất và quản lý bộ nhớ.

## **2.5. Một số thư viện cần dùng**

### ***2.5.1. Numpy***

NumPy (Numerical Python) là một thư viện cơ bản cho tính toán khoa học trong Python. Nó cung cấp hỗ trợ cho các mảng đa chiều (arrays), cùng với một tập hợp lớn các hàm toán học và logic để thao tác với các mảng này một cách hiệu quả.

* Ưu điểm:
* NumPy được viết bằng C, nên các phép tính trên mảng NumPy thường nhanh hơn nhiều so với các phép tính tương tự trên các danh sách (lists) Python.
* NumPy cung cấp hỗ trợ mạnh mẽ cho các mảng đa chiều, cho phép bạn thao tác với dữ liệu dưới dạng ma trận, tensor, v.v.
* NumPy tích hợp tốt với các thư viện khác như SciPy, Pandas, Matplotlib, và scikit-learn, giúp bạn xây dựng các ứng dụng phân tích dữ liệu và học máy dễ dàng hơn.
* NumPy cung cấp nhiều hàm toán học phức tạp như Fourier Transform, Linear Algebra, và Random Number Generation.
* NumPy mảng tiêu thụ ít bộ nhớ hơn và nhanh hơn để truy cập và thao tác với các phần tử của mảng so với danh sách Python.
* Nhược điểm:
* Với những người mới bắt đầu học lập trình Python, việc làm quen với các khái niệm của NumPy và cách sử dụng các hàm của nó có thể khá khó khăn.
* Mặc dù NumPy mảng mạnh mẽ và hiệu quả, nhưng chúng ít linh hoạt hơn so với danh sách Python khi cần lưu trữ các kiểu dữ liệu hỗn hợp hoặc thực hiện các thao tác phức tạp không được hỗ trợ trực tiếp.
* Đối với các phép tính rất phức tạp hoặc không tiêu chuẩn, bạn có thể phải viết thêm mã C hoặc sử dụng các thư viện khác để mở rộng khả năng của NumPy.

### ***2.3.2. Keras***

Keras là một thư viện mã nguồn mở dành cho việc xây dựng và huấn luyện các mô hình học sâu (deep learning) một cách dễ dàng và hiệu quả. Nó được thiết kế để giúp người dùng phát triển nhanh chóng các mô hình học sâu mà không cần phải xử lý nhiều chi tiết phức tạp của các framework học sâu cơ bản. Keras có thể chạy trên các backend khác nhau như TensorFlow, Theano, và Microsoft Cognitive Toolkit .

* Ưu điểm:
* Keras được thiết kế với trọng tâm là khả năng sử dụng và khả năng mở rộng, với một API trực quan và đơn giản giúp người dùng dễ dàng bắt đầu và xây dựng các mô hình phức tạp.
* Mặc dù đơn giản, Keras vẫn rất mạnh mẽ và có thể mở rộng để xây dựng các mô hình học sâu phức tạp. Người dùng có thể dễ dàng thêm các lớp mới và các hàm mất mát tùy chỉnh.
* Keras có thể chạy trên nhiều backend khác nhau như TensorFlow, Theano, và CNTK, giúp linh hoạt trong việc lựa chọn framework phù hợp với nhu cầu của bạn.
* Keras là một phần của TensorFlow và được tích hợp chặt chẽ, cho phép người dùng tận dụng các tính năng tiên tiến của TensorFlow như phân tán và tối ưu hóa hiệu suất.
* Keras có một cộng đồng lớn và nhiều tài liệu, hướng dẫn, và ví dụ thực tế, giúp người dùng học hỏi và giải quyết các vấn đề một cách nhanh chóng.
* Nhược điểm:
* Mặc dù Keras rất tiện dụng, nhưng đôi khi nó có thể chậm hơn so với việc sử dụng trực tiếp các framework backend như TensorFlow, vì nó có thể tạo ra thêm lớp trừu tượng và do đó tăng thêm chi phí tính toán.
* Keras thiết kế để đơn giản hóa quá trình phát triển mô hình, điều này cũng có nghĩa là người dùng sẽ có ít kiểm soát chi tiết hơn so với việc sử dụng trực tiếp các framework backend. Điều này có thể là một hạn chế đối với các mô hình yêu cầu tùy chỉnh cao.
* Do Keras là một lớp trừu tượng trên các framework backend, việc khắc phục sự cố có thể trở nên phức tạp hơn khi bạn phải tìm hiểu qua nhiều lớp mã để tìm ra nguyên nhân của vấn đề.

### ***2.3.3. Pandas***

Pandas là một thư viện mã nguồn mở trong Python, được thiết kế để hỗ trợ thao tác và phân tích dữ liệu. Pandas cung cấp các cấu trúc dữ liệu và hàm thao tác với bảng số liệu và chuỗi thời gian (time series) một cách dễ dàng và trực quan. Hai cấu trúc dữ liệu chính của Pandas là DataFrame (bảng dữ liệu 2 chiều) và Series (dữ liệu 1 chiều).

* Ưu điểm
* Pandas cung cấp các hàm thao tác dữ liệu rất mạnh mẽ và dễ sử dụng, giúp bạn dễ dàng làm sạch, biến đổi và phân tích dữ liệu.
* Pandas được xây dựng trên nền tảng NumPy, do đó thừa hưởng hiệu suất cao và khả năng xử lý hiệu quả của NumPy.
* Pandas có thể xử lý nhiều loại dữ liệu khác nhau, bao gồm dữ liệu từ CSV, Excel, SQL, và nhiều định dạng khác.
* Pandas cung cấp nhiều công cụ và hàm mạnh mẽ để phân tích dữ liệu, bao gồm lọc dữ liệu, nhóm dữ liệu (grouping), ghép dữ liệu (merging), và xử lý dữ liệu chuỗi thời gian.
* Nhược điểm
* Khi xử lý các bộ dữ liệu rất lớn, Pandas có thể tiêu tốn nhiều bộ nhớ, gây khó khăn cho việc làm việc với các bộ dữ liệu không vừa bộ nhớ (memory).
* Mặc dù Pandas rất nhanh cho các bộ dữ liệu vừa và nhỏ, nhưng khi làm việc với các bộ dữ liệu rất lớn, nó có thể trở nên chậm chạp và kém hiệu quả.
* Đối với những người mới bắt đầu, cú pháp và các hàm của Pandas có thể hơi phức tạp và khó nắm bắt.

### ***2.2.4. Scikit-learn***

Scikit-learn là một thư viện Python được sử dụng rộng rãi cho học máy và phân tích dữ liệu. Đây là một công cụ mạnh mẽ và linh hoạt, bao gồm nhiều thuật toán học máy, từ học giám sát (như phân loại và hồi quy) đến học không giám sát (như phân cụm và giảm chiều). Scikit-learn được thiết kế để tích hợp dễ dàng với các thư viện khoa học số liệu của Python như NumPy và SciPy.

* Ưu điểm:
* Dễ sử dụng: Scikit-learn có API đơn giản và nhất quán, làm cho quá trình triển khai và thử nghiệm các mô hình học máy trở nên dễ dàng hơn.
* Tài liệu hướng dẫn tốt: Thư viện có một lượng tài liệu rất lớn và hữu ích, bao gồm nhiều ví dụ và hướng dẫn cụ thể, giúp người mới bắt đầu có thể dễ dàng tiếp cận.
* Cộng đồng mạnh mẽ và hỗ trợ rộng rãi: Với cộng đồng lớn và năng động, các vấn đề và cập nhật thường được giải quyết nhanh chóng.
* Hiệu suất tối ưu: Scikit-learn được tối ưu hóa cho hiệu suất, với một số thành phần được viết bằng Cython để tăng tốc.
* Hỗ trợ nhiều thuật toán: Bao gồm một loạt các thuật toán cho cả học có giám sát và không giám sát, cũng như các công cụ cho tiền xử lý dữ liệu, chọn mô hình và đánh giá mô hình.
* Nhược điểm:
* Không tốt cho quy mô lớn: Scikit-learn không được thiết kế cho dữ liệu cực kỳ lớn hoặc tính toán phân tán, điều này có thể hạn chế khi làm việc với bộ dữ liệu cỡ lớn hay trong môi trường phân tán.
* Không hỗ trợ mô hình học sâu: Không có hỗ trợ bản địa cho các mô hình học sâu như mạng nơ-ron, người dùng cần sử dụng các thư viện khác như TensorFlow hoặc PyTorch cho những nhu cầu đó.
* Thiếu tính linh hoạt: Mặc dù Scikit-learn cung cấp nhiều thuật toán, nhưng nó không cho phép tùy chỉnh sâu các thuật toán đó. Điều này có thể không đáp ứng được cho những nhu cầu rất cụ thể hoặc nâng cao.
* Tối ưu hóa tham số có hạn: Mặc dù có các công cụ như GridSearchCV và RandomizedSearchCV cho tìm kiếm tham số, nhưng quá trình tìm kiếm này có thể không hiệu quả về mặt tính toán đối với các không gian tham số lớn.

# **CHƯƠNG 3: SỬ DỤNG LSTM ĐỂ THỰC NGHIỆM BÀI TOÁN**

Sau khi tìm hiểu về các thuật toán AI, ưu và nhược điểm của từng thuật toán. Thuật toán LSTM có ưu điểm nội trội và phù hợp với bài toàn đặt ra.

## **3.1. Tổng quan về dữ liệu**

Dữ liệu tài chính được lấy trực tiếp từ Yahoo Finance. Yahoo Finance là một trong những nguồn cung cấp dữ liệu tài chính lớn và uy tín, bao gồm thông tin về giá cổ phiếu, các chỉ số thị trường, tỷ giá hối đoái, hàng hóa, và các dữ liệu tài chính khác. Chính vì thế tôi đã sử dụng thư viện ‘yfinance’, một thư viện của Python phổ biến để lấy dữ liệu.

Dữ liệu gồm các thông tin sau:

A table with numbers and letters

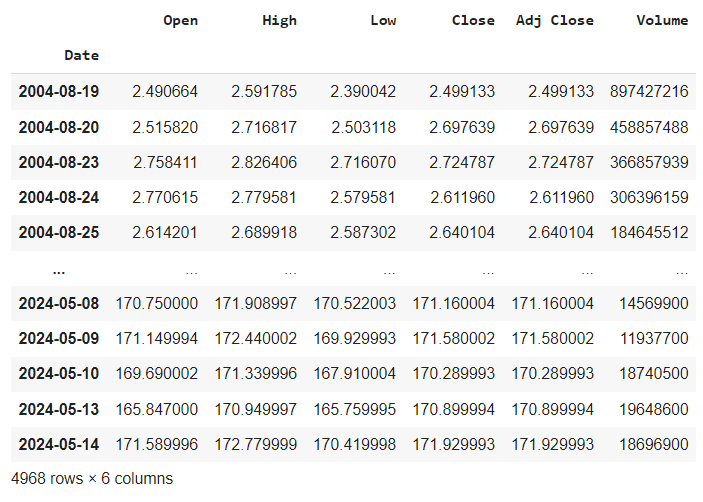
Description automatically generated

Hình 3. 1. Hình ảnh dữ liệu được lấy từ Yahoo Finance

* Date: Ngày xác định dữ liệu mà nó được ghi nhận.
* Open: Đại diện cho giá mở cửa của chứng khoán vào đầu phiên giao dịch của ngày củ thể đó. Đây là giá mà chứng khoán bắt đầu được giao dịch khi thị trường mở cửa vào buổi sáng. Giá mở cửa là một chỉ số quan trọng vì nó có thể phản ánh phản ứng của thị trường đối với các tin tức hoặc sự kiện xảy ra ngoài giờ giao dịch.
* High: đại diện cho giá cao nhất mà cổ phiếu hoặc tài sản đạt được trong phiên giao dịch của ngày cụ thể đó. Đây là mức giá cao nhất mà các nhà giao dịch đã sẵn sàng trả cho cổ phiếu trong suốt khoảng thời gian giao dịch của ngày đó.
* Low: đại diện cho giá thấp nhất mà cổ phiếu hoặc tài sản đạt được trong phiên giao dịch của ngày cụ thể đó. Đây là mức giá thấp nhất mà các nhà giao dịch đã sẵn sàng bán cổ phiếu trong suốt khoảng thời gian giao dịch của ngày đó.
* Close: đại diện cho giá đóng cửa của cổ phiếu hoặc tài sản vào cuối phiên giao dịch của ngày cụ thể đó. Đây là mức giá cuối cùng mà cổ phiếu được giao dịch trước khi thị trường đóng cửa. Giá đóng cửa là một trong những chỉ số quan trọng nhất vì nó thường được sử dụng để đánh giá hiệu suất hàng ngày của cổ phiếu và là cơ sở cho nhiều phân tích kỹ thuật và tài chính.
* Adj Close: đại diện cho giá đóng cửa được điều chỉnh để phản ánh các sự kiện như chia cổ tức, chia tách cổ phiếu (stock split), và các thay đổi về vốn khác. Giá này cung cấp một cái nhìn chính xác hơn về giá trị thực sự của cổ phiếu sau khi các điều chỉnh này được tính đến.
* Volume: đại diện cho khối lượng giao dịch, tức là tổng số lượng cổ phiếu hoặc tài sản được mua và bán trong một phiên giao dịch cụ thể. Khối lượng giao dịch là một chỉ số quan trọng vì nó cung cấp thông tin về mức độ hoạt động và sự quan tâm của thị trường đối với một cổ phiếu cụ thể.

## **3.2. Xử lý dữ liệu**

Lấy dữ liệu 20 năm của mã chứng khoán ‘GOOG’ (mã chứng khoán của Tập đoàn Alphabet), ta được một bảng dữ liệu gồm 4968 hàng và 6 cột như hình dưới:



Hình 3. 2. Dữ liệu mã chứng khoán GOOG

Tiến hành phân tích dữ liệu thấy dữ liệu sạch không có ô nào khuyết.

A screenshot of a computer code

Description automatically generated

Hình 3. 3. Kết quả thông tin của dữ liệu được lấy

## **3.3. Huấn luyện mô hình**

* Lấy tập dữ liệu:

Adj\_close\_price = google\_data[['Adj Close']]

* Tiến hành lấy cột dữ liệu ‘Adj Close’ của tập dữ liệu đã lấy ở trên để chuẩn bị chuẩn hóa và chia tập dữ liệu thành train model và test model train.
* Chuẩn hóa dữ liệu:

from sklearn.preprocessing import MinMaxScaler

scaler = MinMaxScaler(feature\_range=(0,1))

scaled\_data = scaler.fit\_transform(Adj\_close\_price)

scaled\_data

* MinMaxScaler: Được sử dụng để chuẩn hóa dữ liệu sao cho tất cả các giá trị nằm trong khoảng từ 0 đến 1.
* fit\_transform: Hàm này sẽ học các giá trị cực đại và cực tiểu từ dữ liệu đầu vào Adj\_close\_price và sau đó chuyển đổi dữ liệu này thành khoảng từ 0 đến 1.
* Tạo tập dữ liệu huấn luyện và kiểm tra

x\_data = []

y\_data = []

for i in range(60, len(scaled\_data)):

    x\_data.append(scaled\_data[i-60:i])

    y\_data.append(scaled\_data[i])

import numpy as np

x\_data, y\_data = np.array(x\_data), np.array(y\_data)

* x\_data: Chứa các tập con của dữ liệu với độ dài là 60 (sử dụng 60 bước thời gian để dự đoán bước thời gian tiếp theo).
* y\_data: Chứa giá trị tiếp theo cần dự đoán từ x\_data.
* np.array: Chuyển đổi danh sách x\_data và y\_data thành mảng numpy để dễ dàng xử lý trong mô hình.
* Chia tập dữ liệu thành huấn luyện và kiểm tra

splitting\_len = int(len(x\_data)\*0.7)

x\_train = x\_data[:splitting\_len]

y\_train = y\_data[:splitting\_len]

x\_test = x\_data[splitting\_len:]

y\_test = y\_data[splitting\_len:]

* splitting\_len: Tính toán số lượng mẫu dữ liệu thuộc tập huấn luyện (70% của tổng dữ liệu).
* x\_train, y\_train: Tập huấn luyện chứa các mẫu dữ liệu và nhãn tương ứng.
* x\_test, y\_test: Tập kiểm tra chứa các mẫu dữ liệu và nhãn tương ứng.
* Khởi tạo lớp LSTM:

class LSTM:

    def \_\_init\_\_(self, input\_dim, hidden\_dim):

        self.input\_dim = input\_dim

        self.hidden\_dim = hidden\_dim

        # Initialize weights

        self.Wf = np.random.randn(hidden\_dim, hidden\_dim + input\_dim)

        self.Wi = np.random.randn(hidden\_dim, hidden\_dim + input\_dim)

        self.Wo = np.random.randn(hidden\_dim, hidden\_dim + input\_dim)

        self.Wc = np.random.randn(hidden\_dim, hidden\_dim + input\_dim)

        self.bf = np.zeros((hidden\_dim, 1))

        self.bi = np.zeros((hidden\_dim, 1))

        self.bo = np.zeros((hidden\_dim, 1))

        self.bc = np.zeros((hidden\_dim, 1))

* \_\_init\_\_ method: Phương thức khởi tạo cho lớp LSTM.
* input\_dim: Kích thước của đầu vào tại mỗi bước thời gian.
* hidden\_dim: Kích thước của trạng thái ẩn.
* Wf, Wi, Wo, Wc: Ma trận trọng số tương ứng cho cổng quên, cổng đầu vào, cổng đầu ra và trạng thái ứng cử viên.
* bf, bi, bo, bc: Vector bias tương ứng cho các cổng và trạng thái ứng cử viên.
* Các hàm kích hoạt:

  def sigmoid(self, x):

        return 1 / (1 + np.exp(-x))

    def tanh(self, x):

        return np.tanh(x)

* Hàm sigmoid: Được sử dụng để tính toán đầu ra của các cổng quên, cổng đầu vào và cổng đầu ra. Hàm sigmoid sẽ nén đầu vào vào khoảng (0,1).
* Hàm tanh: Được sử dụng để tính toán trạng thái ứng cử viên và trạng thái ẩn. Hàm tanh sẽ nén đầu vào vào khoảng (-1, 1).
* Phương thức forward:

def forward(self, x):

        h, c = np.zeros((self.hidden\_dim, 1)), np.zeros((self.hidden\_dim, 1))

        for t in range(x.shape[0]):

            x\_t = x[t].reshape(-1, 1)

            concat = np.vstack((h, x\_t))

            f = self.sigmoid(np.dot(self.Wf, concat) + self.bf)

            i = self.sigmoid(np.dot(self.Wi, concat) + self.bi)

            o = self.sigmoid(np.dot(self.Wo, concat) + self.bo)

            c\_hat = self.tanh(np.dot(self.Wc, concat) + self.bc)

            c = f \* c + i \* c\_hat

            h = o \* self.tanh(c)

        return h, c

* forward method: Phương thức tính toán trạng thái ẩn và trạng thái ô qua các bước thời gian.
* h, c: Trạng thái ẩn và trạng thái ô được khởi tạo là các vector không.
* for t in range(x.shape[0]): Vòng lặp qua các bước thời gian của đầu vào x.
* x\_t = x[t].reshape(-1, 1): Lấy đầu vào tại thời điểm t và chuyển thành vector cột.
* concat = np.vstack((h, x\_t)): Kết hợp trạng thái ẩn hiện tại và đầu vào tại thời điểm t.
* f, i, o, c\_hat: Tính toán đầu ra của các cổng quên, đầu vào, đầu ra và trạng thái ứng cử viên sử dụng các trọng số và bias tương ứng.
* c = f \* c + i \* c\_hat: Cập nhật trạng thái ô.
* h = o \* self.tanh(c): Cập nhật trạng thái ẩn.
* Định nghĩa và huấn luyện mô hình:

from keras.models import Sequential

from keras.layers import Dense

+ Nhập các lớp cần thiết từ Keras:

* Sequential: Lớp để tạo mô hình tuần tự.
* Dense, LSTM: Các lớp Dense và LSTM để xây dựng mô hình

model = Sequential()

model.add(LSTM(128, return\_sequences=True, input\_shape=(x\_train.shape[1],1)))

model.add(LSTM(64,return\_sequences=False))

model.add(Dense(25))

model.add(Dense(1))

+ Tạo mô hình tuần tự và thêm các lớp vào mô hình:

* Lớp LSTM đầu tiên với 128 đơn vị và trả về toàn bộ chuỗi đầu ra (return\_sequences=True). Kích thước đầu vào là (x\_train.shape[1], 1).
* Lớp LSTM thứ hai với 64 đơn vị và không trả về toàn bộ chuỗi đầu ra (return\_sequences=False).
* Lớp Dense với 25 đơn vị.
* Lớp Dense đầu ra với 1 đơn vị để dự đoán giá trị cuối cùng.

model.compile(optimizer='adam', loss='mean\_squared\_error')

* Sử dụng bộ tối ưu hóa Adam và hàm mất mát mean\_squared\_error.

model.fit(x\_train, y\_train, batch\_size=1, epochs = 2)

* Huấn luyện mô hình với dữ liệu x\_train và y\_train, kích thước batch là 1 và số epoch là 2.

model.summary()

* Hiển thị thông tin chi tiết về các lớp và số lượng tham số trong mô hình.

## **3.4. Kết quả thực nghiệm**

Dự đoán giá trong tương lai dựa trên dữ liệu 60 ngày gần nhất bằng cách sử dụng một mô hình đã huấn luyện (chẳng hạn như LSTM) và bộ điều chỉnh tỷ lệ (scaler).

import numpy as np

def predict\_future\_days(model, scaler, last\_60days\_sc, days\_to\_predict=7):

    future\_predictions = []

    current\_input = last\_60days\_sc.copy()  # Sao chép dữ liệu 60 ngày cuối cùng

    for \_ in range(days\_to\_predict):

        # Chuẩn bị dữ liệu để dự đoán

        X\_test = []

        X\_test.append(current\_input)

        X\_test = np.array(X\_test)

        X\_test = np.reshape(X\_test, (X\_test.shape[0], X\_test.shape[1], 1))

        # Dự đoán giá

        pred\_price = model.predict(X\_test)

        pred\_price\_inverse = scaler.inverse\_transform(pred\_price)  # Chuyển đổi về giá trị gốc

        future\_predictions.append(pred\_price\_inverse[0][0])

        # Cập nhật đầu vào cho lần dự đoán tiếp theo

        new\_input = np.append(current\_input[1:], pred\_price)  # Thêm giá dự đoán vào chuỗi dữ liệu, loại bỏ giá trị đầu tiên

        current\_input = new\_input.reshape(-1, 1)

    return future\_predictions

# Sử dụng hàm để dự đoán 7 ngày tới

predicted\_prices = predict\_future\_days(model, scaler, last\_60days\_sc, days\_to\_predict=7)

print("Dự đoán giá đóng cửa điều chỉnh 7 ngày tới: ", predicted\_prices)

Kết quả:

A close-up of a number

Description automatically generated

Hình 3. 4. Hình ảnh kết quả dự đoán giá của mã chứng khoán 7 ngày tới

Biểu đồ được vẽ trước khi dự đoán:

A graph showing a line

Description automatically generated with medium confidence

Hình 3. 5. Biểu đồ mã chứng khoán trước khi dự đoán

Biểu đồ sau khi dự đoán:

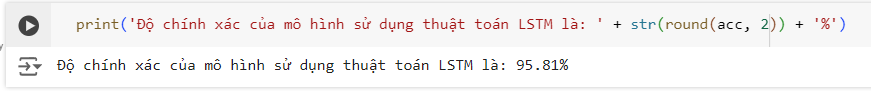
**A graph with blue lines

Description automatically generated**

Hình 3. 6. Biểu đồ mã chứng khoán khi thêm phần dự đoán

## **3.5. So sánh độ chính xác của mô hình**

Độ chính xác của mô hình sử dụng thuật toán LSTM:



Hình 3. 7. Độ chính xác của mô hình LSTM

Độ chính xác của mô hình sử dụng thuật toán KNN:

A close up of text

Description automatically generated

Hình 3. 8. Độ chính xác của mô hình KNN

Từ kết quả độ chính xác của 2 mô hình ta thấy mô hình LSTM có độ chính xác cao hơn mô hình KNN.

# **CHƯƠNG 4: XÂY DỰNG ỨNG DỤNG VÀ KIỂM THỬ**

## **4.1. Xây dựng ứng dụng**

Giao diện đăng nhập của ứng dụng:

**A screenshot of a login page

Description automatically generated**

Hình 4. 1. Giao diện đăng nhập

Giao diện đăng kí của ứng dụng:

A screenshot of a computer

Description automatically generated

Hình 4. 2. Giao diện đăng kí

Giao diện lấy lại mật khẩu:

A screenshot of a computer screen

Description automatically generated

Hình 4. 3. Giao diện lấy lại mật khẩu

Sau khi đăng nhập thành công, giao diện ứng dụng xuất hiện gồm có:

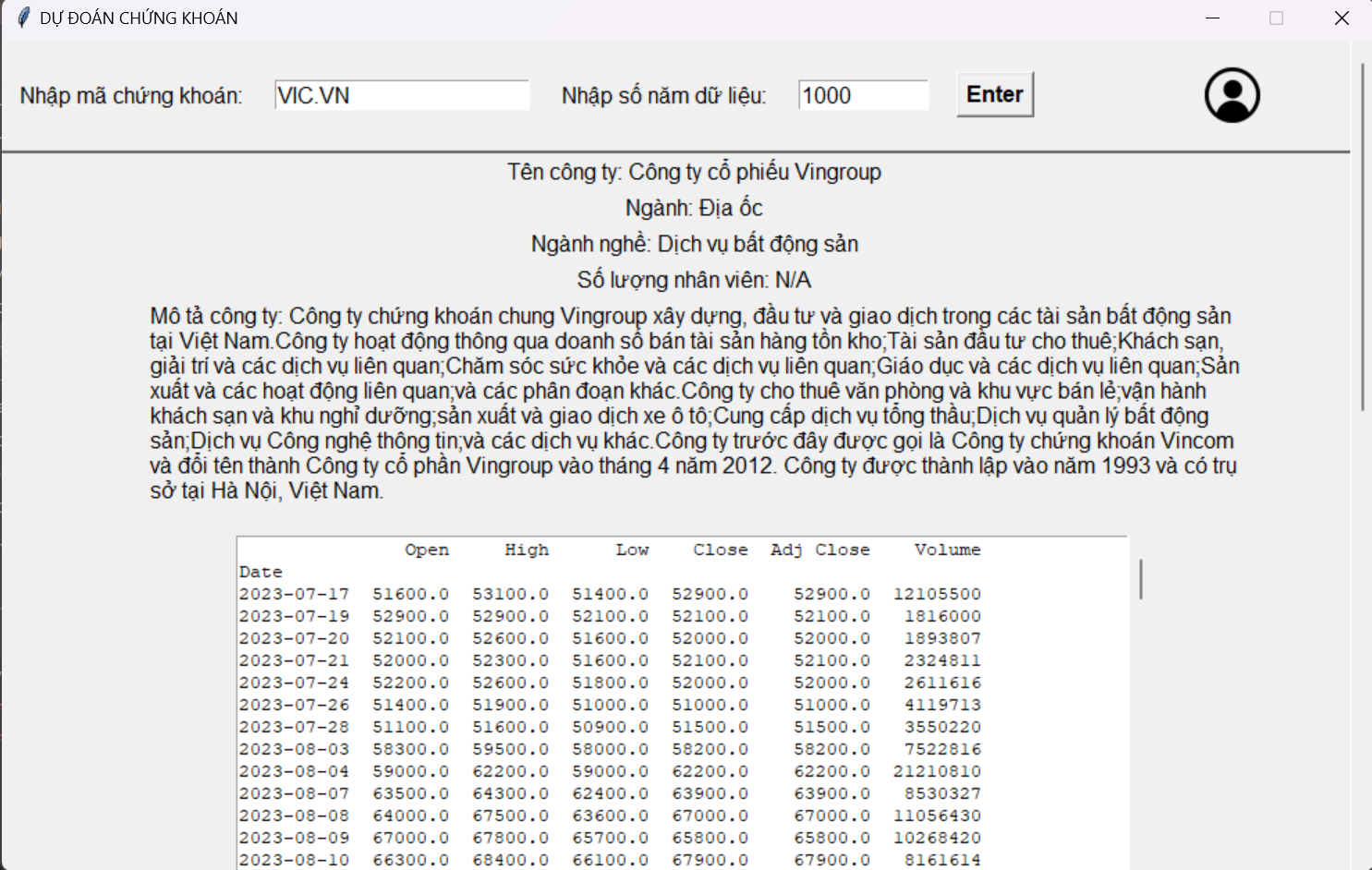
* Nhập mã chứng khoán: Nơi nhập mã chứng khoán mà mình muốn dự đoán
* Nhập số năm dữ liệu: Nơi nhập số năm dữ liệu mà mình muốn hiện thị ra (tính từ ngày hiện tại mình nhập đến ngày cách đó số năm mình đã nhập)

A screenshot of a computer

Description automatically generated

Hình 4. 4. Giao diện ứng dụng dự đoán mã chứng khoán

Sau khi nhập đủ các thông tin như: mã chứng khoán và số năm, nhấn ‘Enter’ ứng dụng sẽ xuất hiện các thông tin về mã chứng khoán, biểu đồ, giá dự đoán và biểu đồ dự đoán



Hình 4. 5. Giao diện khi đã nhập mã chứng khoán

## **4.2. Kiểm thử mô hình**

### ***4.2.1. Kế hoạch kiểm thử***

* Kiểm thử mô hình:
* Kiểm thử tải dữ liệu: trả về DataFrame chứa dữ liệu tải về từ yf.download.
* Kiểm thử chuẩn bị dữ liệu: trả về các giá trị đúng cho new\_df, last\_60days\_sc, và scaler.
* Kiểm thử dự đoán dữ liệu: trả về các giá trị dự đoán chính xác dựa trên model và dữ liệu đầu vào.
* Kiểm thử vẽ đồ thị: đảm bảo không gây ra lỗi khi vẽ đồ thị.
* Kiểm thử giao diện đăng nhập:
* Kiểm thử giao diện: hiện thị đầy đủ các thành phần của giao diện
* Kiểm thử chức năng đăng nhập: đăng nhập được vào ứng dụng
* Kiểm thử chức năng đăng kí: đăng kí được tài khoản mới
* Kiểm thử chức năng quên mật khẩu: hiện thị cửa sổ quên mật khẩu
* Kiểm thử chức năng ẩn và hiện mật khẩu: ẩn và hiện mật khẩu hoạt động đúng

### ***4.2.2. Kiểm thử mô hình***

Bảng 4. 1. Bảng kiểm thử mô hình

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| STT | Case | Đầu vào | Đầu ra mong muốn | Kết quả |
| 1 | Kiểm thử hàm load\_stock\_data | * stock: Mã cổ phiếu (string). * start: Ngày bắt đầu (datetime). * end: Ngày kết thúc (datetime). | DataFrame chứa dữ liệu về giá cổ phiếu, được tải về từ yfinance. | Pass |
| 2 | Kiểm thử hàm prepare\_data | * data: DataFrame chứa dữ liệu về giá cổ phiếu. * days: Số ngày cuối cùng được sử dụng để chuẩn bị dữ liệu (int). | * new\_df: DataFrame chứa dữ liệu giá cổ phiếu đã lọc. * last\_60days\_sc: Mảng NumPy chứa giá cổ phiếu chuẩn hóa của 60 ngày cuối cùng. * scaler: MinMaxScaler đã được fit với dữ liệu. | Pass |
| 3 | Kiểm thử hàm predict\_future\_days | * model: Mô hình dự đoán giá cổ phiếu. * scaler: MinMaxScaler đã được fit với dữ liệu. * last\_60days\_sc: Mảng NumPy chứa giá cổ phiếu chuẩn hóa của 60 ngày cuối cùng. * days\_to\_predict: Số ngày dự đoán tiếp theo (int). | Mảng NumPy chứa giá cổ phiếu dự đoán cho số ngày tiếp theo. | Pass |
| 4 | Kiểm thử hàm plot\_predictions | * original\_data: DataFrame chứa dữ liệu giá cổ phiếu gốc. * predicted\_data: Mảng NumPy chứa giá cổ phiếu dự đoán cho các ngày tiếp theo.   - stock: Mã cổ phiếu (string). | Biểu đồ matplotlib hiển thị dữ liệu giá cổ phiếu gốc và dự đoán. | Pass |

### ***4.2.3. Kiểm thử giao diện đăng nhập***

Bảng 4. 2. Bảng kiểm thử giao diện đăng nhập

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| STT | Case | Đầu vào | Đầu ra mong muốn | Kết quả |
| 1 | Giao diện người dùng | Kiểm tra xem các thành phần giao diện như cửa sổ đăng nhập, các label, các entry field, nút đăng nhập, nút đăng ký, nút quên mật khẩu có hiển thị đúng không. | Giao diện người dùng được hiển thị đúng và đầy đủ, không có lỗi hiển thị. | Pass |
| 2 | Đăng nhập | Người dùng nhập tên người dùng và mật khẩu đúng. | Hiển thị thông báo "Đăng nhập thành công" và chuyển đến trang chính của ứng dụng. | Pass |
| Người dùng nhập tên người dùng hoặc mật khẩu không chính xác. | Hiển thị thông báo lỗi "Tài khoản không tồn tại hoặc mật khẩu không đúng". | Pass |
| 3 | Đăng kí | Người dùng nhập đầy đủ thông tin và mật khẩu | Hiển thị thông báo "Đăng ký thành công". | Pass |
| Người dùng nhập tên người dùng đã tồn tại, email đã được sử dụng | Hiển thị thông báo lỗi tương ứng. | Pass |
| 4 | Quên mật khẩu | Người dùng click vào nút quên mật khẩu. | Hiển thị cửa sổ quên mật khẩu và đóng cửa sổ đăng nhập. | Pass |
| 5 | Ẩn mật khẩu | Người dùng click vào nút ẩn mật khẩu. | Mật khẩu được ẩn và nút chuyển thành nút hiển thị mật khẩu. | Pass |
| 6 | Hiện thị mật khẩu | Người dùng click vào nút hiển thị mật khẩu. | Mật khẩu được hiển thị và nút chuyển thành nút ẩn mật khẩu. | Pass |

### ***4.2.4. Kết quả kiểm thử***

* Tỉ lệ test case đạt (Passed): 100%

# **KẾT LUẬN**

Đề tài “Xây dựng ứng dụng dự đoán sự biến thiên của các mã chứng khoán trong thị trường chứng khoán bằng Python” xuất phát từ những nhu cầu thực tế mà ngày nay trong lĩnh vực tài chính – chứng khoán cần có. Qua quá trình thực hiện đề tài, em đã tìm hiểu, tích lũy và vận dụng được những kiến thức công nghệ được học từ trường, từ nơi thực tập và đã hoàn thành được những phần sau đây.

* Tiền xử lý được dữ liệu
* Huấn luyện mô hình để dự đoán
* Có kế hoạch và đánh giá kiểm thử
* Xây dựng được giao diện ứng dụng

**Hướng phát triển**

* Tiếp tục hoàn thiện các tính năng mở rộng phần mềm nhằm cải tiến và nâng cấp chương trình.
* Từng bước xây dựng ứng dụng đa nền tảng
* Hoàn thiện chương trình, sửa chữa các lỗi, cải thiện tính bảo mật.
* Tích hợp đăng nhập qua Facebook, Google.

# **TÀI LIỆU THAM KHẢO**

**Tiếng việt**

[1] Trần Hùng Cường, Trần Thanh Hùng, *Giáo trình khai phá dữ liệu*, NXB Đại học Công nghiệp Hà Nội.

[2] Nguyễn Phương Nga, Trần Hùng Cường, *Giáo trình trí tuệ nhân tạo*

**Tiếng Anh**

[1] Machine Learning with Python for Everyone (Addison-Wesley Data & Analytics Series) – Softcover

[2] L. F. Sanchez, “Automatic Number Plate Recognition System Using Machine Learning Techniques,” no. August, pp. 2017–2018, 2018.

[3] Python Machine Learning: Machine Learning and Deep Learning with Python, scikit-learn, and TensorFlow 2

**Website**

[1] Trang tài liệu về chứng khoán

<https://chung-khoan-la-gi-phan-biet-cac-loai-chung-khoan.html>

[2] Trang tài liệu về mô hình mạng LSTM

<https://Ly_thuyet_ve_mang_LSTM.html>