|  |
| --- |
| **BỘ GIÁO DỤC VÀ ĐÀO TẠO**  **TRƯỜNG ĐẠI HỌC NGUYỄN TẤT THÀNH**  **KHOA CÔNG NGHỆ THÔNG TIN**  Description: C:\Documents and Settings\Administrator\Desktop\logo dai hoc_khong nen.png  **TIỂU LUẬN MÔN HỌC**  **DEEP LEARNING TRONG KHOA HỌC DỮ LIỆU**  **DỰ ĐOÁN PHÂN LỚP LỢI NHUẬN CỦA APPLE TRÊN GIÁ CỔ PHIẾU**  Giảng viên giảng dạy: THS.HỒ KHÔI  Sinh viên thực hiện: BÙI TIẾN SANG  MSSV : 2000004684  Chuyên ngành : KHOA HỌC DỮ LIỆU  Môn học : DEEP LEARNING TRONG KHOA HỌC DỮ LIỆU  Khóa : 2020-2024    **Tp.HCM, ngày 28 tháng 12 Năm 2022** |
| **BỘ GIÁO DỤC VÀ ĐÀO TẠO**  **TRƯỜNG ĐẠI HỌC NGUYỄN TẤT THÀNH**  **KHOA CÔNG NGHỆ THÔNG TIN**  Description: C:\Documents and Settings\Administrator\Desktop\logo dai hoc_khong nen.png  **TIỂU LUẬN MÔN HỌC**  **DEEP LEARNING TRONG KHOA HỌC DỮ LIỆU**  **DỰ ĐOÁN PHÂN LỚP LỢI NHUẬN CỦA APPLE TRÊN GIÁ CỔ PHIẾU**  Giảng viên giảng dạy: THS.HỒ KHÔI  Sinh viên thực hiện: BÙI TIẾN SANG  MSSV : 2000004684  Chuyên ngành : KHOA HỌC DỮ LIỆU  Môn học : DEEP LEARNING TRONG KHOA HỌC DỮ LIỆU  Khóa : 2020-2024    **Tp.HCM, ngày 28 tháng 12 Năm 2022** |

|  |  |
| --- | --- |
| Trường Đại học Nguyễn Tất Thành  **Khoa Công Nghệ Thông Tin**  🙜 🙜 🙝 🙝 | CỘNG HÒA XÃ HỘI CHỦ NGHĨA VIỆT NAM  **Độc lập – Tự do – Hạnh phúc**  🙜 🙜 🙝 🙝 |

**NHIỆM VỤ ĐỒ ÁN**

Họ và tên: BÙI TIẾN SANG MSSV: 2000004684

Chuyên ngành:KHOA HỌC DỮ LIỆU Lớp: 20DTH2A

Email:buitiensang191@gmail.com SĐT : 0369273641

Tên đề tài: DỰ ĐOÁN PHÂN LỚP LỢI NHUẬN CỦA APPLE TRÊN GIÁ CỔ PHIẾU

Giảng viên giảng dạy: Ths.HỒ KHÔI

Thời gian thực hiện:13/12/2022 đến 26/12/2022.

Nhiệm vụ/nội dung (mô tả chi tiết nội dung, yêu cầu, phương pháp… ):

* Thực hiện công việc A: Nắm được cơ bản các thuật toán.
* Thực hiện công việc B: Áp dụng việc phân tích, xây dựng dữ liệu và sử dụng các thuật toán cây quyết định,hồi quy logistic, hồi quy tuyến tính, RNN vào bài toán
* Thực hiện công việc C: Tiến hành dự đoán phân lớp lợi nhuận của Apple trên giá cổ phiếu.

Nội dung và yêu cầu đã được thông qua bộ môn.

*TP.HCM, ngày 28 tháng 12 năm 2022*

|  |  |
| --- | --- |
| **TRƯỞNG BỘ MÔN**  *(Ký và ghi rõ họ tên)* | **GIÁO VIÊN HƯỚNG DẪN**  *(Ký và ghi rõ họ tên)* |

**LỜI MỞ ĐẦU**

Trong thời đại ngày nay, khoa học và kỹ thuật không ngừng phát triển, việc dữ liệu tạo ngày càng được hoàn thiện và đóng góp không nhỏ trong quá trình nâng cao, cải thiện cuộc sống của nhân loại. Các bài toán thống kê, dự báo thực tế và làm việc với dữ liệu chính là nhiệm vụ cơ bản của ngành Khoa học dữ liệu. Thông qua quá trình huấn luyện với các bộ dữ liệu đầu vào, kết quả thu được với độ chính xác cao có thể giúp con người nắm bắt rõ hơn về số liệu cũng như mối tương quan giữa chúng, dự báo trước được dữ liệu tương lai để đưa ra những giải pháp phù hợp. Chính vì vậy, việc khoa học dữ liệu vào quá trình phân tích, tìm hiểu quan hệ giữa các dữ liệu đầu vào này sẽ giúp ích rất nhiều trong việc ra quyết định, thực hiện các chính sách thay đổi thích hợp góp phần vào quá trình công nghiệp hóa, hiện đại hóa đất nước. Với bộ dữ liệu đầu vào, chúng em sẽ tiến hành phân tích lớp lợi nhuận của một ông công ty lớn đó là Apple và đưa ra các dự đoán và biện pháp chủ quan nhằm cải thiện các vấn đề cốt lõi đã đề ra.

**LỜI CẢM ƠN**

Với lòng biết ơn sâu sắc nhất, em xin gửi đến quý Thầy Cô ở Khoa Công Nghệ Thông Tin Trường Đại Học Nguyễn Tất Thành đã truyền đạt vốn kiến thức quý báu của quý thầy cô cho chúng em trong suốt thời gian học tập tại trường. Nhờ có những lời hướng dẫn, dạy bảo của thầy cô nên đề tài dự đoán phân lớp lợi nhuận của apple dựa trên giá cổ phiếu thành công tốt đẹp

Một lần nữa, em xin chân thành cảm ơn thầy Ths.Hồ Khôi– người đã trực tiếp giúp đỡ, quan tâm, hướng dẫn em hoàn thành tốt sản phẩm và bài báo cáo này trong thời gian qua.

Bài báo cáo của  em còn hạn chế và còn nhiều bỡ ngỡ nên không tránh khỏi những thiếu sót, em rất mong nhận được những ý kiến đóng góp quý báu của quý Thầy để kiến thức của em trong lĩnh vực này được hoàn thiện hơn đồng thời có điều kiện bổ sung, nâng cao ý thức và trình độ của mình của mình.

Em xin chân thành cảm ơn các quý thầy cô rất nhiều!

Sinh viên thực hiện

BÙI TIẾN SANG

|  |  |
| --- | --- |
| TRƯỜNG ĐẠI HỌC NGUYỄN TẤT THÀNH  **TRUNG TÂM KHẢO THÍ** | **KỲ THI KẾT THÚC HỌC PHẦN**  **HỌC KỲ I NĂM HỌC 2022 - 2023** |

**PHIẾU CHẤM THI TIỂU LUẬN/ĐỒ ÁN**

BM-ChT-11

Môn thi: Deep learning trong khoa học dữ liệu Lớp học phần: 20DTH2A

Nhóm sinh viên thực hiện :Bùi Tiến Sang + Châu Nhật Tín

1. Tham gia đóng góp:

2. Tham gia đóng góp:

3. Tham gia đóng góp:

4. Tham gia đóng góp:

5. Tham gia đóng góp:

6. Tham gia đóng góp:

7. Tham gia đóng góp:

8. Tham gia đóng góp:

Ngày thi: Phòng thi:

Đề tài tiểu luận/báo cáo của sinh viên : DỰ ĐOÁN PHÂN LỚP LỢI NHUẬN CỦA APPLE TRÊN GIÁ CỔ PHIẾU

Phần đánh giá của giảng viên (căn cứ trên thang rubrics của môn học):

|  |  |  |  |
| --- | --- | --- | --- |
| **Tiêu chí (theo CĐR HP)** | **Đánh giá của GV** | **Điểm tối đa** | **Điểm đạt được** |
| Cấu trúc của báo cáo |  |  |  |
| Nội dung |  |  |  |
| * Các nội dung thành phần |  |  |  |
| * Lập luận |  |  |  |
| * Kết luận |  |  |  |
| Trình bày |  |  |  |
| **TỔNG ĐIỂM** |  | **10** |  |

|  |  |
| --- | --- |
|  | **Giảng viên chấm thi**  *(ký, ghi rõ họ tên)* |

**NHẬN XÉT CỦA GIẢNG VIÊN GIẢNG DẠY**

*Tp.HCM, Ngày 28 tháng 12 năm 2022*

**Giảng viên giảng dạy**

(Ký tên và ghi rõ họ tên)

**MỤC LỤC**

[CHƯƠNG 1: GIỚI THIỆU 1](#_Toc122968187)

[1.Giới thiệu đề tài. 1](#_Toc122968188)

[1.2.Lý do chọn đề tài 2](#_Toc122968189)

[1.3.Mục tiêu đề tài. 2](#_Toc122968190)

[1.4.Phương pháp đề tài. 2](#_Toc122968191)

[1.4.1.Kiến trúc mạng nơ ron nhân tạo. 3](#_Toc122968192)

[1.4.2.Mạng nơ ron hồi quy. 4](#_Toc122968193)

[1.5.Đối tượng và phạm vi nghiên cứu. 8](#_Toc122968194)

[CHƯƠNG 2.ỨNG DỤNG VÀ THUẬT TOÁN 9](#_Toc122968195)

[2.Giới thiệu về ngôn ngữ lập trình python. 9](#_Toc122968196)

[2.1.Python là gì ? 9](#_Toc122968197)

[2.2. Xếp hạng ngôn ngữ. 10](#_Toc122968198)

[2.3. Lịch sử của Python. 11](#_Toc122968199)

[2.4 Mô tả thuật toán. 12](#_Toc122968200)

[2.4.1. Cây quyết định. 12](#_Toc122968201)

[2.4.2. Mô hình cây quyết định 12](#_Toc122968202)

[2.4.3. Hồi quy Logistic. 12](#_Toc122968203)

[2.4.4. Hồi quy tuyến tính. 13](#_Toc122968204)

[2.4.5. RNN. 15](#_Toc122968205)

[2.5. Xây dựng bộ dữ liệu. 15](#_Toc122968206)

[2.6. Áp dụng Python vào bài toán. 16](#_Toc122968207)

[2.6.1. Cây quyết định. 16](#_Toc122968208)

[2.6.2. Hồi quy Logistic. 16](#_Toc122968209)

[2.6.3. Hồi quy tuyến tính. 17](#_Toc122968210)

[2.6.4. RNN. 17](#_Toc122968211)

[CHƯƠNG 3.XÂY DỰNG ỨNG DỤNG NGÔN NGỮ PYTHON 18](#_Toc122968212)

[3.Xây dựng ứng dụng và giải thích 18](#_Toc122968213)

[3.1 Tiến hành xử lý tiền dữ liệu 18](#_Toc122968214)

[3.2 Tiến hành áp dụng thuật toán và chạy chương trình 23](#_Toc122968215)

[KẾT LUẬN 33](#_Toc122968216)

[Kết quả đạt được. 33](#_Toc122968217)

[Hạn chế và hướng phát triển . 33](#_Toc122968218)

[TÀI LIỆU THAM KHẢO 34](#_Toc122968219)

**DANH MỤC HÌNH**

[Hình 1 Tỷ lệ doanh thu của Apple theo theo danh mục sản phẩm trong Q4/2022 2](#_17dp8vu)

[Hình 2 Cấu trúc mạng nơ-ron nhân tạo(RNN) 3](#_1ksv4uv)

[Hình 3 Quy trình xử lý của mạng nơ ron nhân tạo 4](#_44sinio)

[Hình 4 Mạng nơ ron hồi quy 4](#_z337ya)

[Hình 5 Xử lý bên trong mạng nơ ron nhân tạo 5](#_3j2qqm3)

[Hình 6 Các phép tính toán trên một chuỗi 5](#_1y810tw)

[Hình 7 Bảng xếp hạng ngôn ngữ 10](#_3as4poj)

[Hình 8 : Guido van Rossum 11](#_2p2csry)

[Hình 9 Ví dụ cây quyết định về việc có nên mua laptop hay không 12](#_ihv636)

[Hình 10 Ví dụ hồi quy Logistic 13](#_1hmsyys)

[Hình 11 Ví dụ hồi quy tuyến tính 14](#_2grqrue)

[Hình 12 Mô hình RNN 15](#_3fwokq0)

[Hình 13 Thống kê doanh số bán hàng 16](#_4f1mdlm)

**KÍ HIỆU CÁC CỤM TỪ VIẾT TẮT**

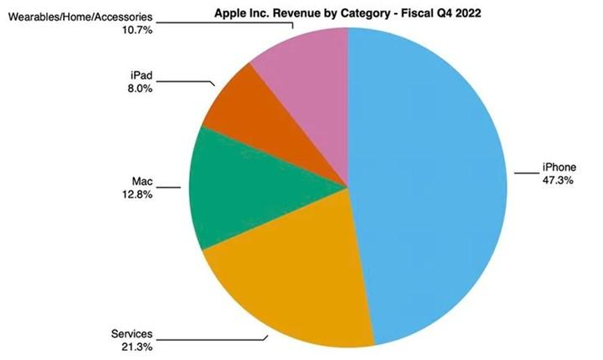
|  |  |
| --- | --- |
| Chữ viết tắt | Ý nghĩa |
| ANN | Artificial Neural Network |
| RNN | Recurrent Neural Network |

# CHƯƠNG 1: GIỚI THIỆU

## 1.Giới thiệu đề tài.

Trong thời buổi công nghệ hiện nay thì không ai lại không biết đến cổ phiếu mà đặt biệt là cổ phiếu của Apple. Cổ phiếu của Apple là một trong số những cổ phiếu thành công nhất trên thị trường chứng khoán . Vào ngày 03/01/2022, cổ phiếu AAPL stock đã đạt vốn hóa thị trường gần 3.000 tỷ USD. Với một năm đầy sự bất ổn trên thị trường, Apple tiếp tục công bố lợi nhuận vượt dự báo bất chấp các vấn đề chuỗi cung ứng toàn cầu. Năm nay, gã khổng lồ công nghệ thậm chí còn có hiệu suất hoạt động vượt trội so với chỉ số S&P 500. Có thể nói hãng công nghệ Apple là một trong những hãng công nghệ có doanh thu khủng mới đây Apple đã công bố báo cáo doanh thu và lợi nhuận cho quý tài chính thứ 4 năm 2022. Tất cả các báo cáo thu nhập này được đưa ra trong bối cảnh lo ngại về suy thoái kinh tế ở Hoa Kỳ và mọi con mắt đang đổ dồn Apple để xem liệu công ty có vượt qua cơn bão tốt hơn các đối thủ cạnh tranh hay không. Và kết quả đúng như mọi người trông đợi. Trong quý này, Apple đã báo cáo doanh thu kỷ lục là 90.1 tỷ USD (hơn 2.2 triệu tỷ đồng) và lợi nhuận lên đến 20.7 tỷ USD (hơn 513 nghìn tỷ đồng). Apple không còn báo cáo doanh số bán hàng theo đơn vị cho bất kỳ sản phẩm nào của mình, mà thay vào đó báo cáo phân tích doanh thu theo danh mục sản phẩm. Dưới đây là danh sách phân tích đầy đủ cho quý tài chính thứ 4 của Apple:

* iPhone: 42.63 tỷ USD (hơn 1 triệu tỷ đồng - tăng 9.8% so với cùng kỳ).
* Mac: 11.51 tỷ USD (hơn 285 nghìn tỷ đồng - tăng 25.4% so với cùng kỳ).
* iPad: 7.17 tỷ USD (hơn 178 nghìn tỷ đồng - giảm 13.1% so với cùng kỳ).
* Thiết bị đeo, nhà và phụ kiện: 9.65 tỷ USD (hơn 239 nghìn tỷ đồng - tăng 9.8% so với cùng kỳ).
* Dịch vụ: 19.19 tỷ USD (hơn 476 nghìn tỷ đồng - tăng 5% so với cùng kỳ).



Hình 1 Tỷ lệ doanh thu của Apple theo theo danh mục sản phẩm trong Q4/2022

## 1.2.Lý do chọn đề tài

Cùng với sự phát triển của các công ty do nhu cầu phân lớp lợi nhuận của các công ty ngày càng cao để có thể phát triển công ty một cách vượt bậc hơn cho các năm sau đó của công ty. Thì chúng em quyết chọn đề tài dự đoán phân lớp lợi nhuận của Apple là một trong những hãng công nghệ có lợi nhuận khủng nhất thế giới để tìm ra những biểu đồ và giải pháp cụ thể nhất .

## 1.3.Mục tiêu đề tài.

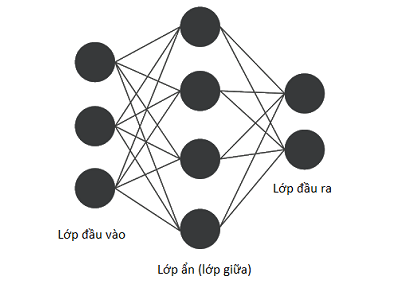
Dự đoán phân lớp lợi nhuận của Apple thông qua cổ phiếu để đưa ra các phương hướng để nâng cao chất lượng và đưa ra các phương pháp tối ưu nhất để phát triển được lợi nhuận của công ty .

## 1.4.Phương pháp đề tài.

Nghiên cứu trên lý thuyết gồm kiến trúc mạng nơ ron nhân tạo và mạng nơron hồi quy. bằng cách sử dụng các thuật toán như là cây quyết định , hồi quy logistics, hồi quy tuyến tính và xây dựng mô hình RNN

### **1.4.1.Kiến trúc mạng nơ ron nhân tạo.**

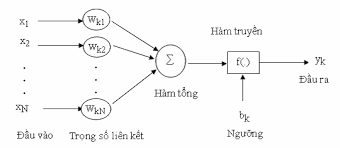
Mạng nơ ron nhân tạo (Artificial Neural Network-ANN) là một mô hình xử lý thông tin được mô phỏng dựa trên hoạt động của hệ thống thần kinh của sinh vật bao gồm số lượng lớn các nơ ron được gắn kết để xử lý thông tin. ANN hoạt động giống như bộ não con người, được học bởi kinh nghiệm ( thông tin qua huấn luyện), có khả năng lưu giữ các tri thức và sử dụng các tri thức đó trong việc dự đoán các dữ liệu chưa biết (unseen data). Một mạng nơ ron là một nhóm các nút nối với nhau, mô phỏng mạng nơ ron thần kinh của não người. Mạng nơ ron nhân tạo được thể hiện thông qua ba thành phần cơ bản: mô hình của nơ ron, cấu trúc và sự liên kết giữa các nơ ron. .



Hình 2 Cấu trúc mạng nơ-ron nhân tạo(RNN)

Cấu trúc cơ bản của mạng nơ-ron nhân tạo trông như sau: Mỗi vòng tròn được gọi là một “nút” (node) và nó mô phỏng một nơ-ron duy nhất. Ở bên trái là các nút đầu vào (input), ở giữa là các nút ẩn (hidden) và bên phải là các nút đầu ra (output).

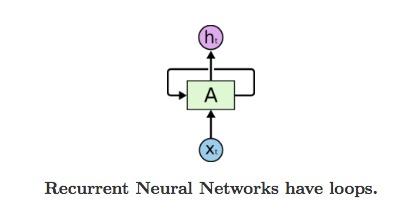
Trong đó, lớp ẩn (Hidden Layer) gồm các nơ ron, nhận dữ liệu input từ các nơ ron ở lớp (Layer) trước đó và chuyển đổi các input này cho các lớp xử lý tiếp theo. Trong một mạng ANN có thể có nhiều Hidden Layer. Lợi thế lớn nhất của mạng ANN là khả năng được sử dụng như một cơ thể xấp xỉ hàm tùy ý mà học được từ các dữ liệu quan sát. Tuy nhiên, sử dụng chúng không đơn giản như vậy, một số các đặc tính và kinh nghiệm khi thiết kế một mạng nơ ron ANN .



*Hình 3 Quy trình xử lý của mạng nơ ron nhân tạo*

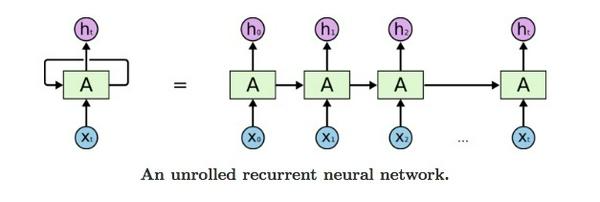
### **1.4.2.Mạng nơ ron hồi quy.**

Mạng neural hồi quy là một mạng neural chứa một vòng lặp bên trong nó.



Hình 4 Mạng nơ ron hồi quy

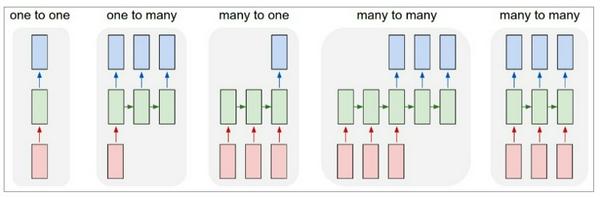
Trong hình trên, Alà mạng neural hồi quy. Nó nhận một đầu vào xt, tiến hành xử lý và đưa ra đầu ra ht. Điểm đặc biệt của Alà nó sẽ lưu lại giá trị của htđể sử dụng cho đầu vào tiếp theo. Có thể coi một mạng neural hồi quy là một chuỗi những mạng con giống hệt nhau, mỗi mạng sẽ truyền thông tin nó vừa xử lý cho mạng phía sau nó. Nếu ta tách từng vòng lặp xử lý trong Ara thành từng mạng con theo cách suy nghĩ như trên thì ta sẽ có một mạng có kiến trúc như sau:



*Hình 5 Xử lý bên trong mạng nơ ron nhân tạo*

Chuỗi các đầu vào x0, x1, ... , xtlà những sự kiện xảy ra theo thứ tự thời gian. Những sự kiện này đều có mối liên hệ về thông tin với nhau và thông tin của chúng sẽ được giữ lại để xử lý sự kiện tiếp theo trong mạng neural hồi quy. Vì tính chất này, mạng neural hồi quy phù hợp cho những bài toán với dữ liệu đầu vào dưới dạng chuỗi với các sự kiện trong chuỗi có mối liên hệ với nhau. Vì vậy, mạng neural hồi quy có ứng dụng quan trọng trong các bài toán xử lý ngôn ngữ tự nhiên như: Dịch máy - Neural Machine Translation, Phân loại ngữ nghĩa - Semantic classification, Nhận dạng giọng nói: Speech Recognition.

Một trong những điểm mạng của mạng neural hồi quy là cho phép tính toán trên một chuỗi các vector. Dưới đây là các kiểu hoạt động của mạng neural hồi quy:



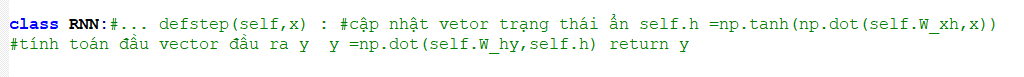
*Hình 6 Các phép tính toán trên một chuỗi*

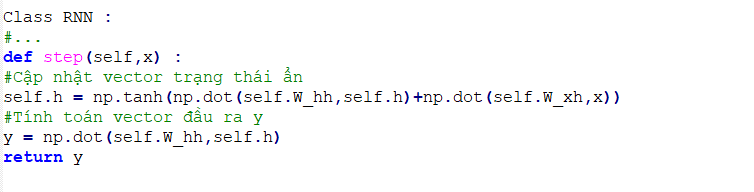
Mỗi hình chữ nhật là 1 vector và các mũi tên thể hiện các hàm biến đổi. Vector đầu vào có màu đỏ, vector đầu ra có màu xanh biển và vector trạng thái thông tin trao đổi giữa các mạng con có màu xanh lá. Từ trái sang phải ta có:

* Mạng neural kiểu Vanilla: Đầu vào và đầu ra có kích thước cố định ( Bài toán nhận diện ảnh - Image Classification)
* Đầu ra có dạng chuỗi: Đầu vào cố định và đầu ra là một chuỗi các vector ( Bài toán tạo tiêu đề cho ảnh - Image Captioning)
* Đầu vào có dạng chuỗi: Đầu vào là một chuỗi vector và đầu ra cố định ( Bài toán phân loại ngữ nghĩa - Sentiment Classification)
* Đầu vào và đầu ra có dạng chuỗi: Bài toán Dịch máy - Neural Machine Translation
* Đầu vào và đầu ra có dạng chuỗi đồng bộ: Đầu vào và đầu ra là một chuỗi vector có độ dài bằng nhau ( Bài toán phân loại video và gắn nhãn từng frame - Video Classification)

Có thể nhận thấy rằng độ dài các chuỗi đầu vào hay đầu ra tại mỗi trường hợp không bắt buộc phải cố định vì kích thước vector trạng thái thông tin trao đổi trong mạng neural hồi quy là cố định. Giờ chúng ta sẽ đi sâu hơn vào phương thức hoạt động của mạng neural hồi quy.

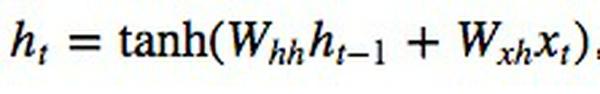
Mạng neural hồi quy nhận một vector đầu vào x và đưa ra vector đầu ra y. Để có thể lưu trữ được thông của các sự kiện trong quá khứ, mạng neural hồi quy lưu trữ trong chính nó một vector trạng thái ẩn h. Vector trạng thái này sẽ lưu giữ những thông tin của những sự kiện đã được xử lý bằng cách cập nhật lại giá trị mỗi khi một sự kiện mới được xử lý. Dưới đây là đoạn code minh họa quá trình này.





* W\_hh: Ma trận weights cho vector trạng thái ẩn.
* W\_xh: Ma trận weights cho vector đầu vào x.
* W\_hy: Ma trận weights dùng để tính vector đầu ra y

Vector trạng thái ẩn self.h được khởi tạo là một vector không. Hàm np.tanh là một hàm phi tuyến hyperbolic đưa giá trị của từng phần tử về khoảng [-1, 1]. Công thức toán học dùng để tính vector trạng thái h:



Các ma trận weights này sẽ được cập nhật trong quá trình huấn luyện để điều chỉnh hành vi của mạng neural hồi quy. Với một đầu vào x là một chuỗi các vector có độ dài 10, mạng neural hồi quy hoạt động như sau:



Đó là cách hoạt động của mạng neural hồi quy dạng many-to-one có đầu vào là chuỗi các vector và đầu ra là một vector có kích thước cố định. Tuy nhiên, kiến trúc này có một nhược điểm: Nếu kích thước chuỗi vector đầu vào xlà rất lớn thì việc tính toán vector trạng thái ẩn h sẽ phải đi qua nhiều lớp tính toán. Trong quá trình backpropagate để cập nhật các tham số weights, vì việc đạo hàm phải đi qua nhiều lớp tính toán của vector hnên các giá trị cập nhật(gradient) cũng sẽ lớn dần lên, việc cập nhật các weights không theo ý muốn và khiến mạng không ổn định. Vì vậy, các biến thể nâng cấp của mạng neural hồi quy đã ra đời để khắc phục vấn đề này: LSTM - Long Short Term Memoryvà GRU - Gated Recurrent Unit

## 1.5.Đối tượng và phạm vi nghiên cứu.

Doanh thu theo quý bán của Apple, chủ yếu là giá bán và số lượng bán ra theo từng quý.

Phạm vi nghiên cứu là lợi nhuận của Apple trên giá cổ phiếu và đưa ra các phương hướng trong tương tai để phát triển thêm và tăng thêm lợi nhuận của công ty.

# CHƯƠNG 2.ỨNG DỤNG VÀ THUẬT TOÁN

## 2.Giới thiệu về ngôn ngữ lập trình python.

### **2.1.Python là gì ?**

Python là một ngôn ngữ lập trình thông dịch (interpreted), hướng đối tượng (object-oriented), và là một ngôn ngữ bậc cao (high-level) ngữ nghĩa động (dynamic semantics). Python hỗ trợ các module và gói (packages), khuyến khích chương trình module hóa và tái sử dụng mã. Trình thông dịch Python và thư viện chuẩn mở rộng có sẵn dưới dạng mã nguồn hoặc dạng nhị phân miễn phí cho tất cả các nền tảng chính và có thể được phân phối tự do.

Sau đây là các đặc điểm của Python:

* Ngữ pháp đơn giản, dễ đọc.
* Vừa hướng thủ tục (procedural-oriented), vừa hướng đối tượng (object-oriented)
* Hỗ trợ module và hỗ trợ gói (package)
* Xử lý lỗi bằng ngoại lệ (Exception)
* Kiểu dữ liệu động ở mức cao.
* Có các bộ thư viện chuẩn và các module ngoài, đáp ứng tất cả các nhu cầu lập trình.
* Có khả năng tương tác với các module khác viết trên C/C++ (Hoặc Java cho Jython, hoặc .Net cho IronPython).
* Có thể nhúng vào ứng dụng như một giao tiếp kịch bản (scripting interface).

Python dễ dàng kết nối với các thành phần khác:

* Python có thể kết nối với các đối tượng COM, .NET (Ironpython, Python for .net), và CORBA, Java… Python cũng được hỗ trợ bởi Internet Communications Engine (ICE) và nhiều công nghệ kết nối khác.
* Có thể viết các thư viện trên C/C++ để nhúng vào Python và ngược lại.

Python là ngôn ngữ có khả năng chạy trên nhiều nền tảng.

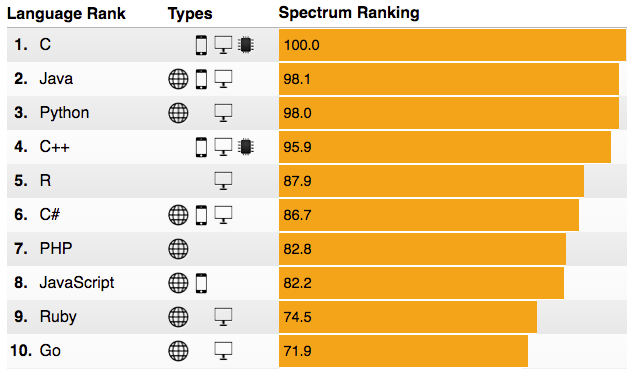
* Python có cho mọi hệ điều hành: Windows, Linux/Unix, OS/2, Mac, Amiga, và những hệ điều hành khác. Thậm chí có cả những phiên bản chạy trên .NET, máy ảo Java, và điện thoại di động (Nokia Series 60). Với cùng một mã nguồn sẽ chạy giống nhau trên mọi nền tảng.

Python rất đơn giản và dễ học

* Python có cộng đồng lập trình rất lớn, hệ thống thư viện chuẩn, và cả các thư viện mã nguồn mở được chia sẻ trên mạng. nó là một mã nguồn mở.
* Cài đặt Python dùng giấy phép nguồn mở nên được sử dụng và phân tối tự do, ngay cả trong việc thương mại. Giấy phép Python được quản lý bởi Python Software Foundation.

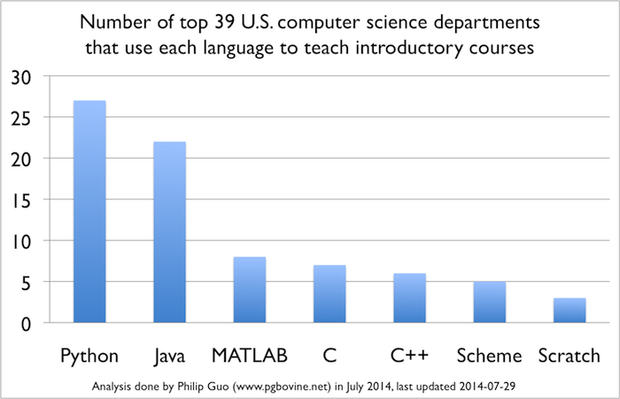
### **2.2. Xếp hạng ngôn ngữ.**

Hiện nay ngôn ngữ Python được xếp hạng thứ 3 trong Top 10 các ngôn ngữ lập trình phổ biến nhất đang được thế giới sử dụng. (Bảng xếp hạng năm 2016):



*Hình 7* Bảng xếp hạng ngôn ngữ

Theo thống kê từ top 39 trường giảng dạy khoa học máy tính thì đa phần các trường sử dụng ngôn ngữ Python để giảng dạy:



Hình 2.1 Thống kê sử dụng ngôn ngữ trong các trường đại học

### **2.3. Lịch sử của Python.**

Python đã được hình thành vào cuối những năm 1980, và việc thực hiện nó vào tháng 12 năm 1989 bởi Guido van Rossum tại Centrum Wiskunde & Informatica (CWI) ở Hà Lan như là một kế thừa cho ngôn ngữ ABC (tự lấy cảm hứng từ SETL) có khả năng xử lý ngoại lệ và giao tiếp với Hệ điều hành Amoeba. Van Rossum là tác giả chính của Python, và vai trò trung tâm của ông trong việc quyết định hướng phát triển của Python.



*Hình 8 :* Guido van Rossum

Hơn sáu năm trước, vào tháng 12 năm 1989, tôi đã tìm kiếm một dự án lập trình "sở thích" mà nó đã chiếm đóng tâm trí tôi trong suốt tuần lễ Giáng sinh. Văn phòng của tôi ... sẽ đóng cửa, nhưng tôi đã có một máy tính ở nhà, và không có nhiều thứ khác trên tay. Tôi quyết định viết một bộ thông dịch (interprester) cho ngôn ngữ kịch bản mới mà tôi đã từng nghĩ đến: một hậu duệ của ABC có thể hấp dẫn các hacker Unix/C. Tôi đã chọn Python như là một tiêu đề làm việc cho dự án.

Python 2.0 đã được phát hành vào ngày 16 tháng 10 năm 2000 và có nhiều tính năng mới, bao gồm bộ thu gom rác theo chu kỳ (cycle-detecting garbage) và hỗ trợ Unicode. Với việc phát hành này quá trình phát triển đã được thay đổi và trở nên minh bạch hơn và cộng đồng hậu thuẫn

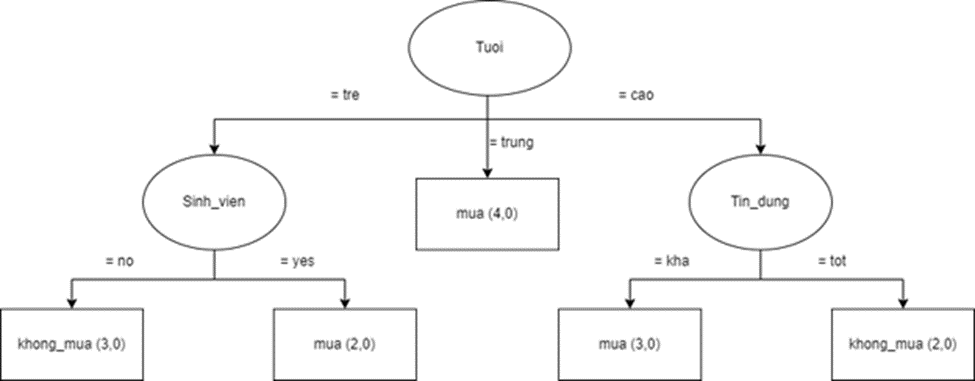
## 2.4 Mô tả thuật toán.

### **2.4.1. Cây quyết định.**

Cây quyết định (decision tree) là cấu trúc dữ liệu dạng cây biểu diễn quá trình ra quyết định dựa vào các điều kiện được xác lập từ các giá trị dữ liệu cụ thể được biểu diễn tại các nút và nhánh của cây.

### **2.4.2. Mô hình cây quyết định**

Xuất phát từ nút gốc (root) tới nút trung gian (nút nội – internal node) và kết thúc ở nút lá (leaf node). Trong đó: Nút gốc và nút nội chứa các biểu thức điều kiện rẽ nhánh .

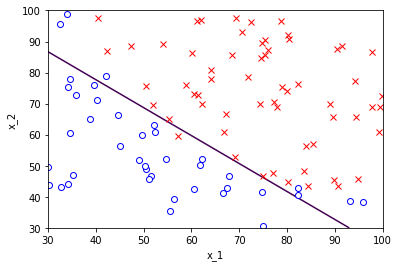


*Hình 9* Ví dụ cây quyết định về việc có nên mua laptop hay không

### **2.4.3. Hồi quy Logistic.**

Phương pháp hồi quy logistic là một mô hình hồi quy nhằm dự đoán giá trị đầu ra rời rạc (discrete target variable) y ứng với một vectơ đầu vào x. Việc này tương đương với chuyện phân loại các đầu vào x vào các nhóm y tương ứng.

xem một bức ảnh có chứa một con mèo hay không. Thì ở đây ta coi đầu ra y=1 y=1 nếu bức ảnh có một con mèo và y=0y=0 nếu bức ảnh không có con mèo nào. Đầu vào \mathbf{x}x ở đây sẽ là các pixel một bức ảnh đầu vào.



*Hình 10* Ví dụ hồi quy Logistic

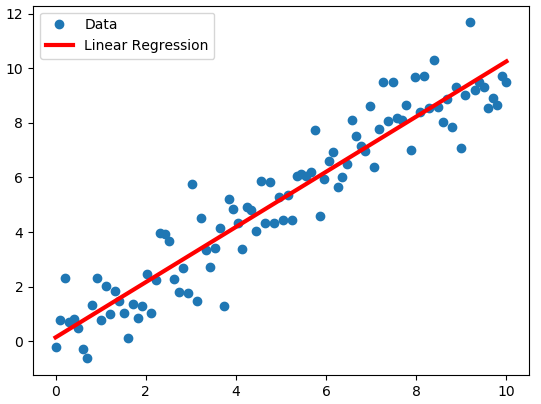
### **2.4.4. Hồi quy tuyến tính.**

“*Phân tích* ***hồi quy tuyến tính (linear regression)*** *là một phương pháp phân tích quan hệ giữa biến phụ thuộc Y với một hay nhiều biến độc lập X.”*

Bao gồm 2 mô hình: *hồi quy tuyến tính quần thể* và *hồi quy tuyến tính mẫu*.

**Tuyến tính (*linear*)** hiểu một cách đơn giản là thẳng, phẳng. Trong không gian hai chiều, một hàm số được gọi là tuyến tính nếu đồ thị của nó có dạng một đường thẳng. Trong không gian ba chiều, một hàm số được gọi là tuyến tính nếu đồ thị của nó có dạng một mặt phẳng. Trong không gian nhiều hơn ba chiều, khái niệm mặt phẳng không còn phù hợp nữa, thay vào đó, một khái niệm khác ra đời được gọi là siêu mặt phẳng (hyperplane)

**Hồi quy tuyến tính (*linear regression*)** được hiểu là một phương pháp dùng để *dự báo một biến (biến phụ thuộc/biến kết quả) theo giá trị của biến còn lại (biến độc lập/biến giải thích)*. Mục tiêu của **hồi quy tuyến tính**: là để làm sao một *đường thẳng có thể tạo được sự phân bố gần nhất với hầu hết các điểm*. Do đó làm giảm khoảng cách (sai số) của các điểm dữ liệu cho đến đường đó.



Hình 11 Ví dụ hồi quy tuyến tính

Nếu chúng ta có từ 2 biến độc lập trở lên khi đó hồi quy tuyến tính giản đơn trở thành hồi quy đa biến.

**\*Phương trình đường thẳng:**

**y = w0 + w1\*X**

**Trong đó:**

x là biến độc lập

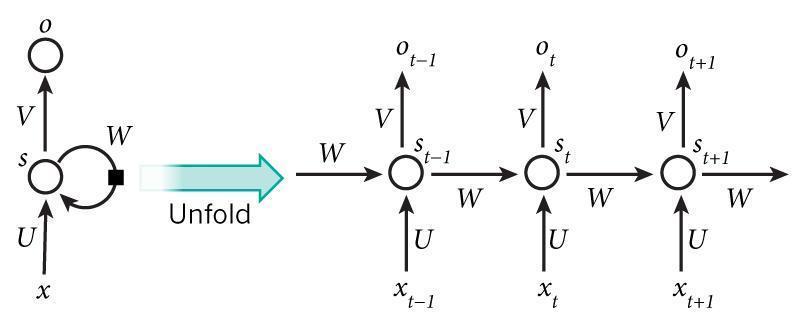
y là biến phụ thuộc

w0 là Hằng số

w1 là Hệ số mối quan hệ giữa X và Y

### **2.4.5. RNN.**

RNN (Recurrent Neural Network) là sử dụng chuỗi các thông tin. Trong các mạng nơ-ron truyền thống tất cả các đầu vào và cả đầu ra là độc lập với nhau. Tức là chúng không liên kết thành chuỗi với nhau.



Hình 12 Mô hình RNN

Mô hình trên mô tả phép triển khai nội dung của một RNN. Triển khai ở đây có thể hiểu đơn giản là ta vẽ ra một mạng nơ-ron chuỗi tuần tự.

## 2.5. Xây dựng bộ dữ liệu.

Dữ liệu được lấy từ thống kê doanh số bán hàng theo quý mới nhất của Apple. Thông thường thì người ta sẽ dùng thống kê theo quý chớ không theo năm nên vậy chúng em cũng thống kê theo quý.

Table

Description automatically generated

Hình 13 Thống kê doanh số bán hàng

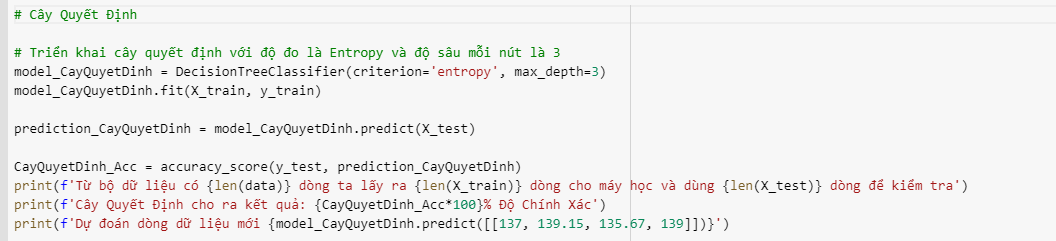
Dữ liệu được đưa ra trên hình 13 thể hiện NgayThang, Gia\_MoCua, Gia\_Tran, Gia\_San , Gia\_DongCua, KhoiLuongGD, LoiNhuan\_Hay không.

Từ bảng dữ liệu ta sử dụng các Gia\_MoCua, Gia\_Tran, Gia\_San , Gia\_DongCua, KhoiLuongGD Ta sẽ lấy 1 cột làm biến phụ thuộc vì đang làm về lợi nhuận nên sẽ lấy biến [‘LoiNhuan\_Hay\_Khong’] để làm biến phụ thuộc.Tiến hành xem data có bao nhiêu hàng bao nhiêu cột , phát hiện các giá trị còn thiếu của tập dữ liệu, tiến hành lọc ra Gia\_MoCua, Gia\_Tran, Gia\_San , Gia\_DongCua, KhoiLuongGD, LoiNhuan.

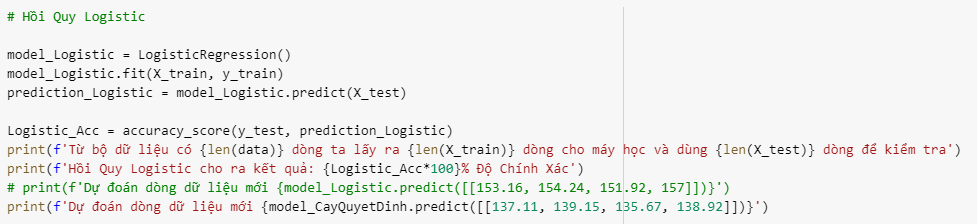
Kiểm tra các giá trị trong dữ liệu có bị null hay không sau đó tiến hành xem các mối tương quan của các cột với nhau.Sau đó tiến hành thống kê mô tả tập dữ liệu và vẽ biểu đồ tương quan của đồ thị các dữ liệu với nhau .

## 2.6. Áp dụng Python vào bài toán.

### **2.6.1. Cây quyết định.**



### **2.6.2. Hồi quy Logistic.**

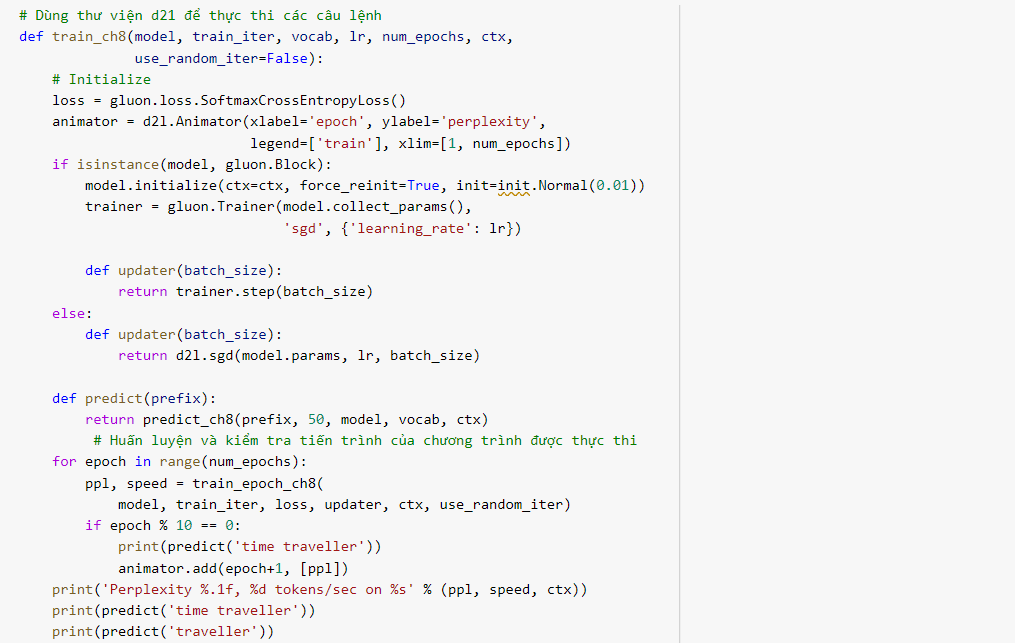


### **2.6.3. Hồi quy tuyến tính.**

Text

Description automatically generated with medium confidence

### **2.6.4. RNN.**

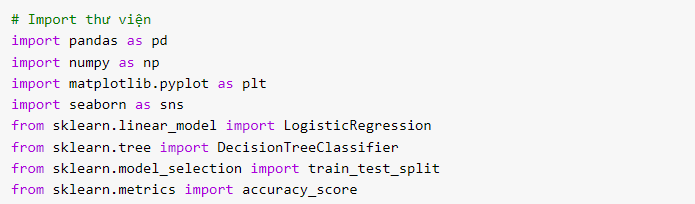


Các mô hình xây dựng dữ liệu trên chỉ có 1 phần để diễn tả được các thuật toán khi sử dụng ngôn ngữ lập trình python. Tiếp theo thì chúng em sẽ xây dựng ứng dụng bằng ngôn ngữ python một cách cụ thể , chân thật thông qua chương 3 của bài báo cáo.

# CHƯƠNG 3.XÂY DỰNG ỨNG DỤNG NGÔN NGỮ PYTHON

## 3.Xây dựng ứng dụng và giải thích

Tiến hành import các thư viện cần thiết để sử dụng cho bài toán.



Tiếp tục liên kết dữ liệu cho bài toán

Table

Description automatically generated

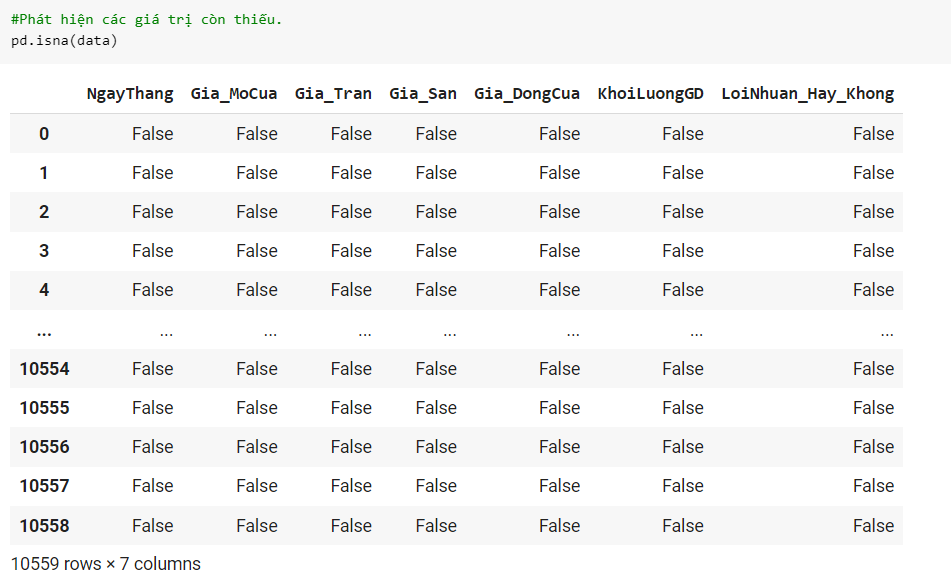
## 3.1 Tiến hành xử lý tiền dữ liệu

Sau khi dùng read.csv để đọc file csv lên thì tiếp theo đó là dùng data.shape để kiểm tra xem dữ liệu gồm bao nhiêu cột và dòng.

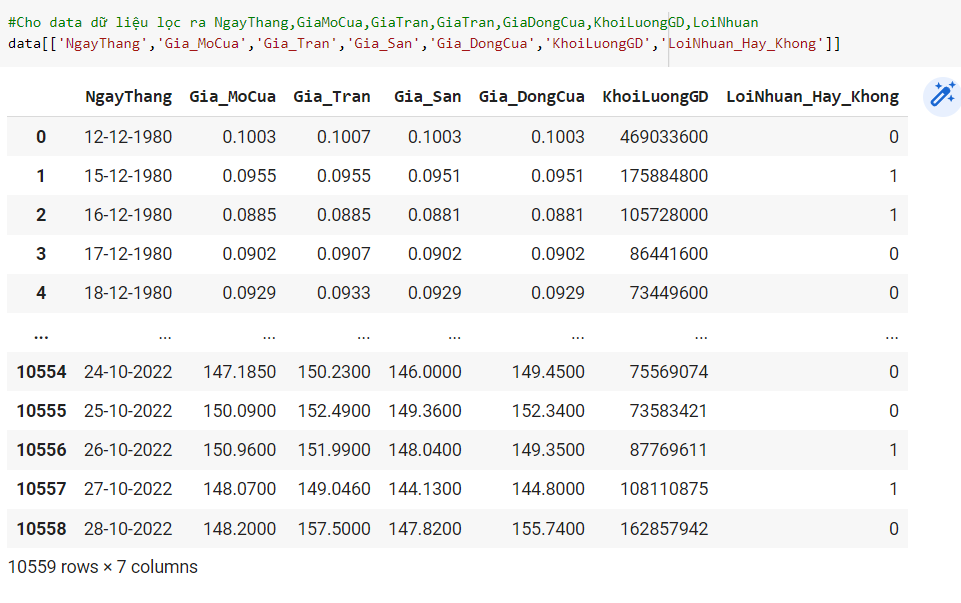
A picture containing text

Description automatically generated

Và tiếp theo đó sử dụng hàm isna để kiểu tra phát hiện các giá trị còn thiếu.



Tiến hành xử lý dữ liệu để lọc ra những dữ liệu cần thiết.



Theo đó là khái quát về data

Table

Description automatically generated

Sau khi khái quát, tiến hành xem thử trong data có giá trị nào bị null hay không .

Table

Description automatically generated

Sau khi tiến hành kiểm tra các giá trị bị null trong tập dữ liệu thì tiến hành xây dựng thêm mối tương quan giữa các cột với nhau .

Table

Description automatically generated

Dùng lệnh data để thể hiện đầy đủ bảng dữ liệu.

Table

Description automatically generated

Tiến hành vẽ biểu đồ ma trận với tương quan giữa các cột với nhau.

A picture containing table

Description automatically generated

Ta thấy sự tương quan của biểu đồ ma trận.

Sau khi thể hiện được sự tương quan của dữ liệu bắt đầu thống kê mô tả

Table

Description automatically generated

Sau đó vẽ biểu đồ tương quan giữa các cột bằng thư viện matplotlib.pyplot  bằng cách sử dụng plt.subplots() và sử dụng hàm for lòng vào nhau để thể hiện sự tương quan giữa các cột và in ra biểu đồ.

A picture containing text

Description automatically generated

Chart, line chart, scatter chart

Description automatically generated

Chart, line chart

Description automatically generated

Chart, line chart

Description automatically generated

Phân chia dataset làm 4 biến độc lập:'Gia\_MoCua','Gia\_Tran','Gia\_San','Gia\_DongCua'.

Sau đó lấy cột LoiNhuan\_Hay\_Khong làm biến phụ thuộc .

Graphical user interface, text, application

Description automatically generated

## 3.2 Tiến hành áp dụng thuật toán và chạy chương trình

Tiếp theo thì phân chia data và train thử như thế nào , ở bước này chúng ta chỉ đơn giản là chia dữ liệu thành 80% và 20% và sau khi train ra có cộng lại giống như ban đầu hay không thôi . Và sau đó là chúng em xây dựng cây quyết định để triển khai cây quyết định

Graphical user interface, text

Description automatically generated with medium confidence

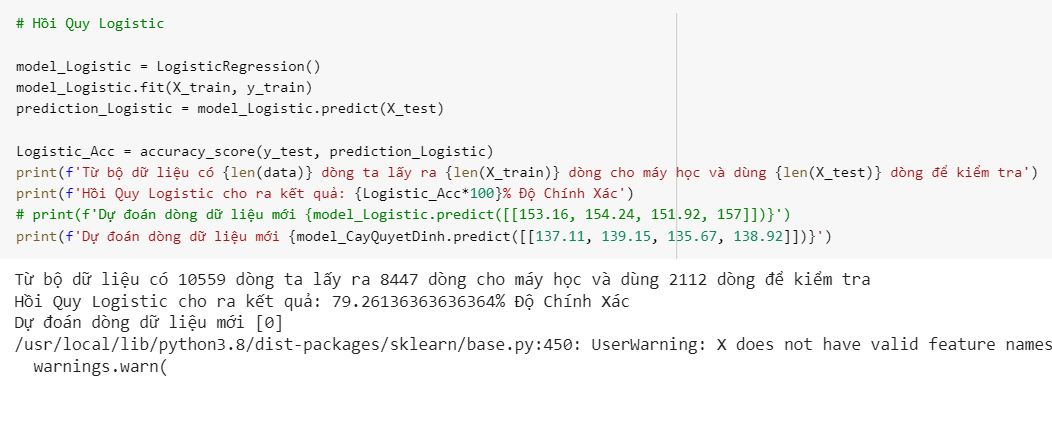
Và sau khi triển khai xong cây quyết định thì đưa dữ liệu train ở trên vào để cho cây quyết định dự đoán và độ chính xác và cho ra kết quả rất khả quan.

Graphical user interface, text, application

Description automatically generated

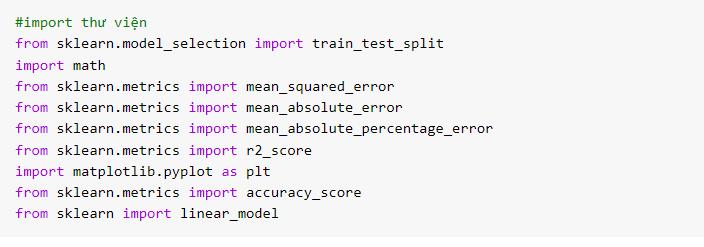
Cây Quyết Định cho ra kết quả độ chính xác : 83.57007575757575%

Trong Deep Learning thi không thể bỏ qua thuật toán xây dựng hồi quy Logistic và dự đoán bằng hồi quy tuyến tính .Để xây dựng hồi quy logistic thi ban đầu tạo 1 biến = Logistic Regression() và fit() nó vào xong theo đó thì vẫn dùng dữ liệu đã train là 80% và 20% để cho ra kết quả hồi quy và cho 1 giá trị bất kì để dự đoán.



Hồi Quy Logistic cho ra kết quả độ chính xác là : 79.26136363636364**%**

Như đã nói ở trên đã có hồi quy logistics thì cũng phải có dự đoán bằng hồi quy tuyến tính.Ban đầu vẫn như bao bài toán khác thì cũng phải khai báo (import) thư viện .



Sau khi Import thư viện xong thì sẽ lấy 1 cột làm biến phụ thuộc vì đang làm về lợi nhuận nên sẽ lấy biến [‘LoiNhuan\_Hay\_Khong’] để làm biến phụ thuộc.Sau đó là train dữ liệu để bước tiếp đó mình có thể dự đoán bằng hồi quy tuyến tính.

Graphical user interface, application

Description automatically generated with medium confidence

Cũng giống như hồi quy logistic ban đầu cho model = linear\_model.LinearRegression() và fit nó và cho pred\_y =model.predict(X\_train) đã train ở trên và sử dụng hàm scatter và plot để vẽ biểu đồ hồi quy tuyến tính , và kèm theo đó là hiển thị sai số của biểu đồ hồi quy.

Chart

Description automatically generated with low confidence

Sau đó chúng ta tiến hành dự đoán.

Graphical user interface, text, application

Description automatically generated

A picture containing chart

Description automatically generated

Kết quả thu được : Hồi Quy Tuyến Tính cho ra kết quả: 79.26136363636364%

Mean Squared Error: 0.22726313235020973

Root Mean Squared Error: 0.4767212312769484 và 0.0879157589945537

Và cuối cùng là RNN là mạng nơ ron để xây dựng mạng nơ ron thi trước tiên phải cài mxnet.

A picture containing diagram

Description automatically generated

Và cài cả thư viện d21 và thư viện matplotlib\_inline.

Graphical user interface, text, application, chat or text message, website

Description automatically generated

Để xây dụng một mạng nơ ron thì chúng ta sử dụng hàm one\_hot và hàm init\_rnn\_state và kể cả sử dụng thư viện d21 để sử dụng cho các câu lệnh.

Graphical user interface, text, application

Description automatically generated

Tiếp theo thì chúng ta biểu diễn one hot với 0 và 2.

Calendar

Description automatically generated with medium confidence

Sử dụng Hàm one\_hot biến đổi một minibatch như vậy thành một tensor 3 chiều với kích thước chiều cuối cùng bằng kích thước bộ từ vựng.

Text

Description automatically generated with low confidence

Kết quả in ra là : **35,32,28**

Tiếp theo chúng ta khởi tạo các tham số cho mô hình RNN.

Text

Description automatically generated with low confidence

Từ đây chúng ta đưa vào các thông số lớp ẩn , thông số lớp đầu ra và đính kèm cho chúng chúng những giải màu tương thích

Graphical user interface, text, application

Description automatically generated

Bắt đầu khởi tạo  các trạng thái ẩn bằng hàm init\_rnn\_state,Hàm này trả về tuple gồm một ndarray chứa giá trị 0 và có kích thước là (kích thước batch, số nút ẩn).Và định nghĩa cách tính toán trạng thái ẩn và đầu ra tại một bước thời gian. Hàm kích hoạt ở đây là tanh.

Text

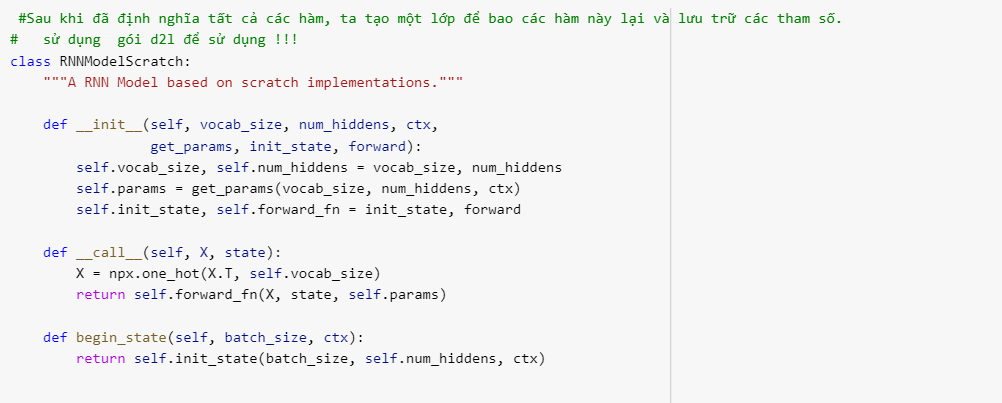
Description automatically generated

Định dạng đầu vào

Text

Description automatically generated

Sau khi đã định nghĩa tất cả các hàm, ta tạo một lớp để báo các hàm này lại và lưu trữ các tham số thì chúng ta sử dụng  gói d2l để sử dụng !!! .

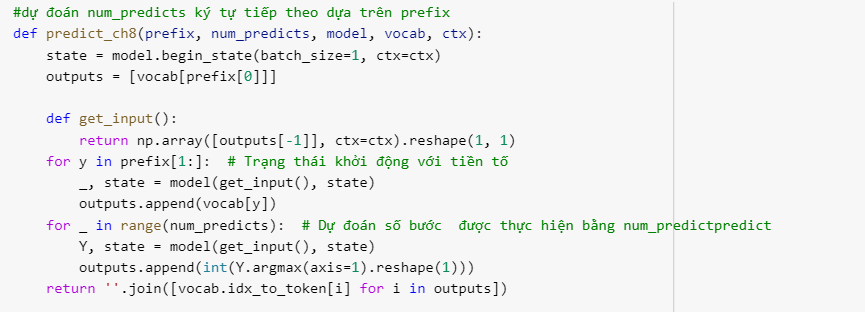


Kiểm tra nhanh chiều của đầu vào và đầu ra, và xem chiều của trạng thái ẩn có thay đổi hay không

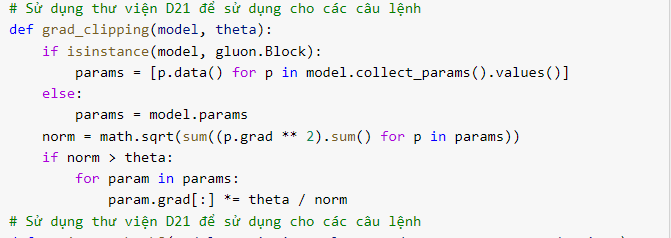
Text, letter

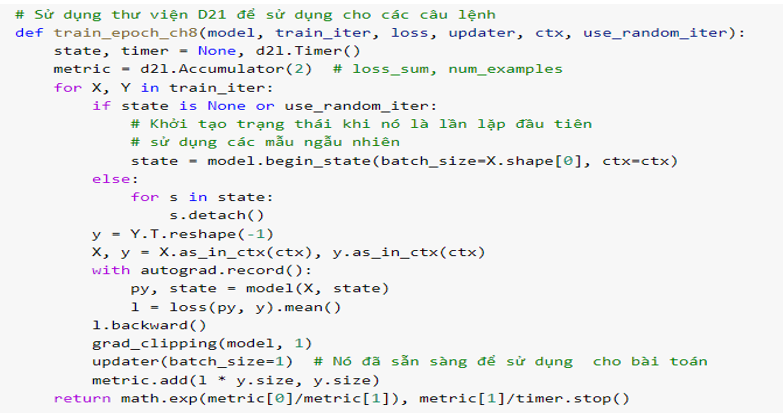
Description automatically generated

Dự đoán num\_predicts ký tự tiếp theo dựa trên prefix. thể hiện được Trạng thái khởi động với tiền tố và dự đoán số bước  được thực hiện bằng num\_predict.

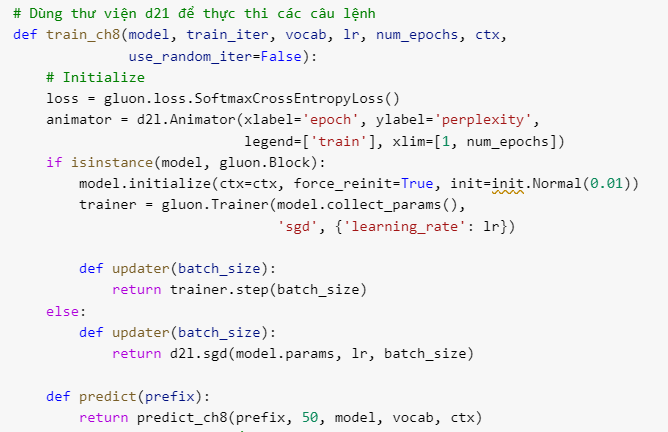


Sau đó chúng em tiếp tục sử dụng thư viện D21  để chạy các câu lệnh và khởi tạo lại các trạng thái khi nó là lần lặp đầu tiên và sử dụng các mẫu ngẫu nhiên để sử dụng cho bài toán

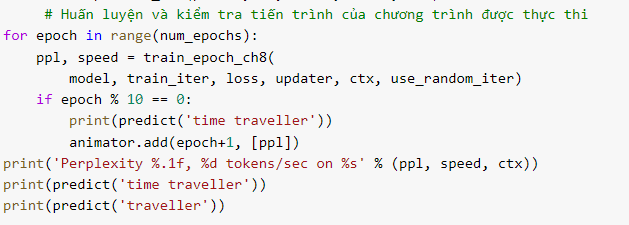




Bắt đầu  huấn luyện  nó và kiểm tra tiến trình thực  thi của chương trình.



Tiếp tục huấn luyện.



Tiến hành train dữ liệu bằng  câu lệnh num\_epochs,

Đầu tiên thử cho chạy 1000 lần trong 1 lần lặp.Ta thu được kết quả như sau :   
Chart

Description automatically generated

Thử train lại thêm 1 lần nữa với 1000 dòng với 2 bước lặp.

A picture containing text

Description automatically generated

Ta sẽ thu được kết quả như sau :

Chart, histogram

Description automatically generated

Chúng ta có thể thấy được sự thay đổi rõ rệt sau 2 lần train và đưa ra 1 kết quả rất là khả quan.

# KẾT LUẬN

## Kết quả đạt được.

Hiểu rõ hơn về việc phân tích dữ liệu, xử lý và đưa ra các phương pháp để xây dựng. Thực thi được hầu hết các mô hình hồi quy và dự đoán và cả cây quyết định và đặc biệt là RNN mạng nơ ron để dự đoán phân lớp lợi nhuận từ giá trị cổ phiếu.

Link sản phẩm :

<https://drive.google.com/drive/folders/1YZaR65SQyaQaxs4UPBCrPi2Viv_7qf0-?usp=share_link>

## Hạn chế và hướng phát triển .

Song theo những mặt tích cực thì trong phần code thì cũng gặp rất nhiều lỗi và sự cố phát sinh trong quá trình xây dựng thuật toán cũng như chưa lưu ý chi tiết những phần code thật sự dễ đọc và dễ hiểu, nhưng song theo đó thì chúng em sẽ cố gắng phát triển hơn , chi tiết hơn trong quá trình xây dựng thuật toán cũng như ứng dụng để đáp ứng được yêu cầu cũng như là nâng cao trình độ để sau này có thể thực chiến tại các công ty một cách thật chuyên nghiệp.

# TÀI LIỆU THAM KHẢO

[1] Nguyễn Thanh Tuấn(2009), Deeplearning cơ bản The Legrand Orange Book Template by Mathias Legrand is used : <https://nttuan8.com/sach-deep-learning-co-ban/>

[2] Ths.Hồ Khôi(2022),Lecturer in deep learning at the Faculty of Information Technology, Nguyen Tat Thanh University.

[3] Từ [DeepLearning.AI](http://deeplearning.ai) (12-23/12/2022) : Start or Advance Your Career in AI.

[4] Đắm chìm vào học sâu (12-23/12/2022) :<https://d2l.aivivn.com/>

[5]Tham khảo sửa chữa bài (12-23/12/2022) : <https://stackoverflow.com/>

[6]File csv Apple revenue from 1980 to 2022 | [Apple revenue from 1980 to 2022 | Kaggle](https://www.kaggle.com/datasets/meerashareef/apple-revenue-from-1980-to-2022)

[7] He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep residual learning for image recognition. Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 770–778).

[8] Bishop, C. M. (2006). Pattern recognition and machine learning. springer.