**Ảnh có chứa quảng trường

Mô tả được tạo tự độngĐẠI HỌC QUỐC GIA THÀNH PHỐ HỒ CHÍ MINH**

TRƯỜNG ĐẠI HỌC CÔNG NGHỆ THÔNG TIN

🙞🕮🙝



**BÁO CÁO ĐỒ ÁN CUỐI KỲ**

MÔN HỌC : DỮ LIỆU LỚN

Đề tài : DỰ ĐOÁN LỢI NHUẬN DỰA TRÊN CHI PHÍ BỎ

RA ĐỂ NGHIÊN CỨU VÀ PHÁT TRIỂN SẢN PHẨM

Giảng viên hướng dẫn:  **ThS. Nguyễn Hồ Duy Tri**

Sinh viên:

Lê Tuấn Khanh MSSV: 19521681

Hồ Trọng Khang MSSV: 19521661

Nguyễn Cao Khoa MSSV: 19521694

Lê Tiến Vinh MSSV: 19522521

**TP.Hồ Chí Minh, tháng 12 2022**

**ĐẠI HỌC QUỐC GIA THÀNH PHỐ HỒ CHÍ MINH**

TRƯỜNG ĐẠI HỌC CÔNG NGHỆ THÔNG TIN

🙞🕮🙝



**BÁO CÁO ĐỒ ÁN CUỐI KỲ**

MÔN HỌC : DỮ LIỆU LỚN

Đề tài : DỰ ĐOÁN LỢI NHUẬN DỰA TRÊN CHI PHÍ BỎ

RA ĐỂ NGHIÊN CỨU VÀ PHÁT TRIỂN SẢN PHẨM

Giảng viên hướng dẫn:  **ThS. Nguyễn Hồ Duy Tri**

Sinh viên:

Lê Tuấn Khanh MSSV: 19521681

Hồ Trọng Khang MSSV: 19521661

Nguyễn Cao Khoa MSSV: 19521694

Lê Tiến Vinh MSSV: 19522521

**TP.Hồ Chí Minh, tháng 12 2022**

# **LỜI CẢM ƠN**

Nhóm xin gửi lời cảm ơn và lòng biết ơn sâu sắc tới ThS. Nguyễn Hồ Duy Tri (giảng viên môn Dữ liệu Lớn). Thầy đã tận tình hướng dẫn, giúp đỡ, cung cấp tài liệu và tạo điều kiện cho nhóm tác giả hoàn thành đồ án này.

Trong thời gian thực hiện đề tài, nhóm đã cố gắng vận dụng những kiến thức nền tảng đã tích lũy đồng thời kết hợp với việc học hỏi và nghiên cứu những kiến thức mới để ứng dụng vào thực hiện đề tài này. Mặc dù đã hết sức cố gắng, nhưng chắc chắn đồ án vẫn còn những thiếu sót. Chính vì vậy, nhóm rất mong nhận được những sự góp ý quý báu từ Thầy nhằm hoàn thiện những kiến thức của và là hành trang để nhóm thực hiện tiếp các đề tài khác trong tương lai.

Xin chân thành cảm ơn Thầy!

Nhóm sinh viên thực hiện

# **NHẬN XÉT CỦA GIÁO VIÊN HƯỚNG DẪN**

*……., ngày……...tháng……năm 2022…*

**Người nhận xét**

*(Ký tên và ghi rõ họ tên)*

**MỤC LỤC**

[LỜI CẢM ƠN 3](#_Toc122458550)

[NHẬN XÉT CỦA GIÁO VIÊN HƯỚNG DẪN 4](#_Toc122458551)

[MỤC LỤC 5](#_Toc122458552)

[DANH MỤC ẢNH 7](#_Toc122458553)

[CHƯƠNG 1: MỤC TIÊU VÀ HƯỚNG LÀM 10](#_Toc122458554)

[**1.1** **Tổng quan đề tài** 10](#_Toc122458555)

[**1.2** **Mục tiêu đề tài** 10](#_Toc122458556)

[CHƯƠNG 2: NGUỒN VÀ CHI TIẾT TẬP DỮ LIỆU 11](#_Toc122458557)

[**2.1** **Tổng quan dữ liệu** 11](#_Toc122458558)

[**2.2** **Chi tiết dữ liệu** 11](#_Toc122458559)

[**2.2.1** **Thêm các thư viện cần thiết.** 12](#_Toc122458560)

[**2.2.2** **Đọc dữ liệu.** 12](#_Toc122458561)

[**2.2.3** **Check các dòng dữ liệu** 12](#_Toc122458562)

[CHƯƠNG 3: XỬ LÝ DỮ LIỆU 13](#_Toc122458563)

[**3.1** **Kiểm tra Kiểu dữ liệu các cột** 13](#_Toc122458564)

[**3.2** **Kiểm tra thông tin các cột** 13](#_Toc122458565)

[**3.3** **Kiểm tra dữ liệu bị thiếu** 14](#_Toc122458566)

[**3.4** **Tạo view để truy vấn dữ liệu** 14](#_Toc122458567)

[**3.5** **Phân loại các thuộc tính có trong tập dữ liệu** 14](#_Toc122458568)

[**3.6** **Kiểm tra giá trị cột RD Spend** 15](#_Toc122458569)

[**3.7** **Kiểm tra giá trị cột Profit** 16](#_Toc122458570)

[CHƯƠNG 4: CHECK OUTLINERS, LÀM SẠCH VÀ KIỂM SOÁT NGOẠI LỆ DỮ LIỆU 17](#_Toc122458571)

[**4.1** **Checkoutliners, làm sạch dữ liệu cho cột RD Spend** 17](#_Toc122458572)

[**4.2** **Checkoutliners, làm sạch dữ liệu cho cột Profit** 18](#_Toc122458573)

[CHƯƠNG 5: PHÂN TÍCH DỮ LIỆU 20](#_Toc122458574)

[**5.1** **Trực quan hóa các chi phí bỏ ra để nghiên cứu và phát triển sản phẩm ( RDSpend)** 20](#_Toc122458575)

[**5.2** **Trực quan hóa giữa Chi phí (RDSpend) và Lợi nhuận (Profit)** 22](#_Toc122458576)

[CHƯƠNG 6: CHUẨN BỊ DỮ LIỆU 23](#_Toc122458577)

[**6.1** **Phân chia dữ liệu thành 2 phần Feature (X) và Target (Y)** 23](#_Toc122458578)

[**6.1.1** **Show dữ liệu X ( RDSpend)** 24](#_Toc122458579)

[**6.1.2** **Show dữ liệu Y ( Profit)** 24](#_Toc122458580)

[**6.2** **Chia dữ liệu Test Train theo theo tỉ lệ Test 30% và Train 70%** 25](#_Toc122458581)

[CHƯƠNG 7: XÂY DỰNG THUẬT TOÁN LINEAR REGRESSION 27](#_Toc122458582)

[**7.1** **Xây dựng thuật toán Linear Regression ( Hướng làm 1)** 27](#_Toc122458583)

[**7.1.1** **Tạo mô hình hồi quy tuyến tính ( Linear Regression )** 27](#_Toc122458584)

[**7.1.2** **Ứng dụng mô hình hồi quy tuyến tính trên tập dữ liệu Test** 32](#_Toc122458585)

[**7.1.3** **Tạo và ứng dựng chỉ số đánh giá RMSE - Root Mean Square Error** 34](#_Toc122458586)

[**7.1.4** **Trực quan hóa và so sánh kết quả dự đoán với thực tế** 35](#_Toc122458587)

[**7.2** **Xây dựng thuật toán Linear Regression ( Hướng làm 2)** 37](#_Toc122458588)

[**7.2.1** **Tạo mô hình hồi quy tuyến tính ( Linear Regression )** 37](#_Toc122458589)

[**7.2.2** **Ứng dụng mô hình dự đoán trên tập Test** 39](#_Toc122458590)

[**7.2.3** **Tạo và ứng dựng chỉ số đánh giá RMSE - Root Mean Square Error** 40](#_Toc122458591)

[**7.2.4** **Trực quan hóa và so sánh kết quả dự đoán với thực tế** 41](#_Toc122458592)

[**7.3** **Xây dựng thuật toán Linear Regression ( Hướng làm 3 sử dụng thư viện Mllib)** 42](#_Toc122458593)

[**7.3.1** **Chuẩn bị dữ liệu** 42](#_Toc122458594)

[**7.3.2** **Ứng dụng mô hình hồi quy tuyến tính trên tập dữ liệu test** 44](#_Toc122458595)

[**7.3.3** **Tạo và ứng dựng chỉ số đánh giá RMSE - Root Mean Square Error** 45](#_Toc122458596)

[**7.3.4** **Trực quan hóa và so sánh kết quả dự đoán với thực tế** 46](#_Toc122458597)

[CHƯƠNG 8: SO SÁNH THUẬT TOÁN 47](#_Toc122458598)

[**8.1** **So sánh, đánh giá các hướng làm** 47](#_Toc122458599)

[**8.1.1** **Thời gian thuật toán thực hiện** 47](#_Toc122458600)

[**8.1.2** **Độ chính xác thông qua chỉ số RMSE** 48](#_Toc122458601)

[**8.1.3** **Độ chính xác thông qua trực quan dữ liệu** 49](#_Toc122458602)

[CHƯƠNG 9: KẾT LUẬN 50](#_Toc122458603)

[**9.1** **Ưu và nhược điểm giữa các hướng làm** 50](#_Toc122458604)

[**9.2** **Hướng phát triển** 50](#_Toc122458605)

[CHƯƠNG 10: TÀI LIỆU THAM KHẢO 51](#_Toc122458606)

# **DANH MỤC ẢNH**

[**Hình 1: Thêm thư viện 11**](#_Toc122457200)

[**Hình 2: Đọc dữ liệu vào 12**](#_Toc122457201)

[**Hình 3: Kiểm tra số dòng dữ liệu 12**](#_Toc122457202)

[**Hình 4: Kiểm tra dữ liệu các cột 13**](#_Toc122457203)

[**Hình 5: Kiểm tra thông tin các cột 13**](#_Toc122457204)

[**Hình 6: Kiểm tra dữ liệu bị thiếu 14**](#_Toc122457205)

[**Hình 7: Tạo view để truy vấn dữ liệu 14**](#_Toc122457206)

[**Hình 8: Tạo dataframe để dùng cho việc dự đoán và đổi tên cột 15**](#_Toc122457207)

[**Hình 9: Kiểm tra giá trị cột RDSpend 15**](#_Toc122457208)

[**Hình 10: Kiểm tra giá trị cột Profit 16**](#_Toc122457209)

[**Hình 11: Lấy giá trị của RDSpend 17**](#_Toc122457210)

[**Hình 12: Đưa giá trị vào df 17**](#_Toc122457211)

[**Hình 13: Biểu đồ giá trị trực quan và ngoại lệ của RDSpend 17**](#_Toc122457212)

[**Hình 14: Lấy giá trị của Profit 18**](#_Toc122457213)

[**Hình 15: Đưa giá trị vào df 18**](#_Toc122457214)

[**Hình 16: Biểu đồ giá trị trực quan và ngoại lệ của Profit 18**](#_Toc122457215)

[**Hình 17: Kiểm tra giá trị ngoại lệ của Profit 19**](#_Toc122457216)

[**Hình 18: Lấy các giá trị của RDSpend và Profit 19**](#_Toc122457217)

[**Hình 19: Đưa các giá trị vào df 19**](#_Toc122457218)

[**Hình 20: Trực quan hóa các giá trị chi phí 20**](#_Toc122457219)

[**Hình 21: Trực quan hóa giữa chi phí và lợi nhuận 21**](#_Toc122457220)

[**Hình 22: Chia dữ liệu thành Feature (X) và Target (y) 22**](#_Toc122457221)

[**Hình 23: Dữ liệu của X 23**](#_Toc122457222)

[**Hình 24: Dữ liệu của y 23**](#_Toc122457223)

[**Hình 25: Chia dữ liệu thành Test và Train 24**](#_Toc122457224)

[**Hình 26: Các thông tin của X train 24**](#_Toc122457225)

[**Hình 27: Các thông tin của X test 24**](#_Toc122457226)

[**Hình 28: Các thông tin của y train 25**](#_Toc122457227)

[**Hình 29: Các thông tin của y test 25**](#_Toc122457228)

[**Hình 30: Đưa giá trị của 2 tập train test lên RDD theo dạng list 26**](#_Toc122457229)

[**Hình 31: Số lượng dữ liệu trong train test 26**](#_Toc122457230)

[**Hình 32: Hàm tính giá trị trung bình (Mean) 27**](#_Toc122457231)

[**Hình 33: Hàm tính mẫu số B1 (Slope) 27**](#_Toc122457232)

[**Hình 34: Tính mẫu số và mean của x train 28**](#_Toc122457233)

[**Hình 35: Tính mẫu số và mean của y train 28**](#_Toc122457234)

[**Hình 36: Giá trị mẫu số và mean tính được 29**](#_Toc122457235)

[**Hình 37: Hàm tính tử số B1 (Slope) 29**](#_Toc122457236)

[**Hình 38: Giá trị tử số tính được 29**](#_Toc122457237)

[**Hình 39: Hàm tính hệ số B0 và B1 30**](#_Toc122457238)

[**Hình 40: Hệ số B0 và B1 tính được 30**](#_Toc122457239)

[**Hình 41: Hàm tính toán dựa trên công thức hồi quy tuyến tính 31**](#_Toc122457240)

[**Hình 42: Chuyển rdd\_X\_test thành danh sách các giá trị 31**](#_Toc122457241)

[**Hình 43: Dự đoán kết quả và tính toán thời gian chạy 31**](#_Toc122457242)

[**Hình 44: Kết quả dự đoán 32**](#_Toc122457243)

[**Hình 45: Chuyển rdd\_y\_test thành danh sách các giá trị 32**](#_Toc122457244)

[**Hình 46: Hàm tính chỉ số RMSE 33**](#_Toc122457245)

[**Hình 47: Hàm đánh giá thuật toán trên tập test 34**](#_Toc122457246)

[**Hình 48: Đánh giá thuật toán 34**](#_Toc122457247)

[**Hình 49: Giá trị gốc 34**](#_Toc122457248)

[**Hình 50: So sánh giá trị dự đoán và giá trị gốc 35**](#_Toc122457249)

[**Hình 51: Trực quan giá trị dự đoán và giá trị thực tế 35**](#_Toc122457250)

[**Hình 52: Tạo tập dữ liệu train test 36**](#_Toc122457251)

[**Hình 53: Hàm tính toán hệ số B0 và B1 37**](#_Toc122457252)

[**Hình 54: Hệ số B0 và B1 tính được 38**](#_Toc122457253)

[**Hình 55: Hàm tính toán dựa trên công thức hồi quy tuyến tính 38**](#_Toc122457254)

[**Hình 56: Dự đoán kết quả và tính toán thời gian chạy 38**](#_Toc122457255)

[**Hình 57: Hàm tính chỉ số RMSE 39**](#_Toc122457256)

[**Hình 58: Hàm đánh giá thuật toán hồi quy trên tập test 39**](#_Toc122457257)

[**Hình 59: Đánh giá thuật toán hồi quy 40**](#_Toc122457258)

[**Hình 60: Lấy giá trị gốc và dự đoán 40**](#_Toc122457259)

[**Hình 61: So sánh giá trị dự đoán với giá trị gốc 40**](#_Toc122457260)

[**Hình 62: Trực quan giá trị dự đoán với giá trị gốc 41**](#_Toc122457261)

[**Hình 63: Chuyển dữ liệu RDD sang DF 41**](#_Toc122457262)

[**Hình 64: Chuyển giá trị cột RDSpend vào cột Features 42**](#_Toc122457263)

[**Hình 65: Thông tin của dữ liệu 43**](#_Toc122457264)

[**Hình 66: Thực hiện mô hình hồi quy tuyến tính trên tập test 43**](#_Toc122457265)

[**Hình 67: Thêm giá trị dự đoán vào DF 44**](#_Toc122457266)

[**Hình 68: Đánh giá mô hình bằng RMSE 44**](#_Toc122457267)

[**Hình 69: So sánh giá trị dự đoán và giá trị thực tế 45**](#_Toc122457268)

[**Hình 70: Trực quan giá trị dự đoán và giá trị thực tế 45**](#_Toc122457269)

[**Hình 71: So sánh thời gian thực hiện 46**](#_Toc122457270)

[**Hình 72: So sánh độ chính xác bằng RMSE 47**](#_Toc122457271)

[**Hình 73: So sánh độ chính xác qua việc dự đoán 48**](#_Toc122457272)

# **CHƯƠNG 1: MỤC TIÊU ĐỀ TÀI**

* 1. **Tổng quan đề tài**

Hiện nay, chúng ta đang sống trong thời kỳ của cuộc Cách mạng công nghiệp lần thứ 4. Đây là thời kỳ gắn với những đột phá về công nghệ cũng như chuyển đổi số trong các doanh nghiệp. Công nghệ thông tin ngày càng phát triển thì đời sống của con người càng cần nhiều ứng dụng để hỗ trợ, phát triển hơn. Điều đó ngày càng thu hút nhiều Startup với các giải pháp công nghệ của mình để có thể bắt kịp với xu hướng cộng đồng hiện tại.

Tuy nhiên, với 1 Startup hay bất cứ công ty nào, thì lợi nhuận luôn được xem như là 1 thước đo để đánh giá sự thành công của công ty đó. Và đã có không ít công ty phải đổ vỡ, phá sản vì những tính toán không hợp lí về chi phí bỏ ra để có thể đầu tư và phát triển sản phẩm của mình. Cũng chính vì những sự thay đổi chóng mặt, sự sôi động trong thị trường, mà trong đồ án môn học Môn Dữ Liệu Lớn lần này, nhóm chúng em tiến hành nghiên cứu dự đoán lợi nhuận dựa trên chi phí bỏ ra để có thể phát triển các sản phẩm.

Qua đó, có thể hỗ trợ doanh nghiệp có 1 cái nhìn tổng quan về thị trường, chi phí và nhiều thứ khác bỏ ra để có thể thành công trong kinh doanh.

* 1. **Mục tiêu đề tài**
* Hiểu và nắm được dữ liệu bài toán.
* Có thể xử lý, làm sạch dữ liệu ban đầu thành nguồn dữ liệu có giá trị phục vụ giải quyết vấn đề đặt ra.
* Khám phá, phân tích và trực quan hóa tập dữ liệu.
* Tự xây dựng được mô hình Linear Regression để dự đoán lợi nhuận và chỉ số đánh giá mức độ chính xác của mô hình.
* So sánh, đánh giá được sự hiệu quả, tính chính xác, ưu, nhược điểm giữa các mô hình Linear Regression tự xây dựng với thư viện MMLib.

# **CHƯƠNG 2: NGUỒN VÀ CHI TIẾT TẬP DỮ LIỆU**

* 1. **Tổng quan dữ liệu**

Để có thể thực hiện dự án thì nhóm chúng em chọn 1 tập dữ liệu Dataset thuộc Kaggle được lấy từ cuộc khảo sát của 1000 Startup.

* Tên Dataset: 1000\_companies\_profit
* Kích thước: 51.2 kB
* Loại tệp tin khi tải xuống từ Kaggle: .csv
* Loại tệp tin sau khi giải nén: .csv
* Tần suất cập nhật: Hằng tháng.
* Link Dataset: [Link](https://www.kaggle.com/datasets/rupakroy/1000-companies-profit?fbclid=IwAR0Y7e_asSIPuiJUml0xnuPyiGLcTngXvYSUqY6D9FTeXaNI-WX8LPnld2Y)   
  1. **Chi tiết dữ liệu**

Tập dữ liệu 1000\_companies\_profit có 1000 dòng thể hiện cho 1000 Startup tham gia khảo sát và 5 thuộc tính dữ liệu bao gồm:

R&D Spend : Chi phí nghiên cứu và phát triển ( Research and Development ) được thể hiện dưới dạng Double ( ví dụ: 91749.16 )

2. Administration Spend: Chi phí quản lý ( là chi phí hoạt động của công ty ) được thể hiện dưới dạng Double ( ví dụ: 114175.79 )

3. Marketing Spend: Chi phí quảng cáo được thể hiện dưới dạng Double ( ví dụ: 294919.57 )

4. State: địa điểm của công ty Startup đó ( ví dụ: Florida )

5. Profit: Lợi nhuận của công ty được thể hiện dưới dạng Double ( ví dụ: 124266.9)

* + 1. **Thêm các thư viện cần thiết.**

Chart

Description automatically generated

Hình : Thêm thư viện

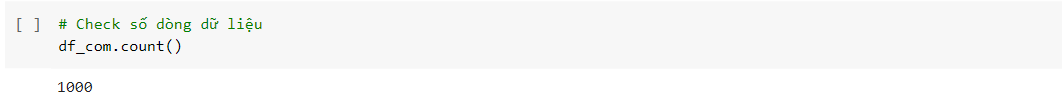
* + 1. **Đọc dữ liệu.**

Table

Description automatically generated

Hình : Đọc dữ liệu vào

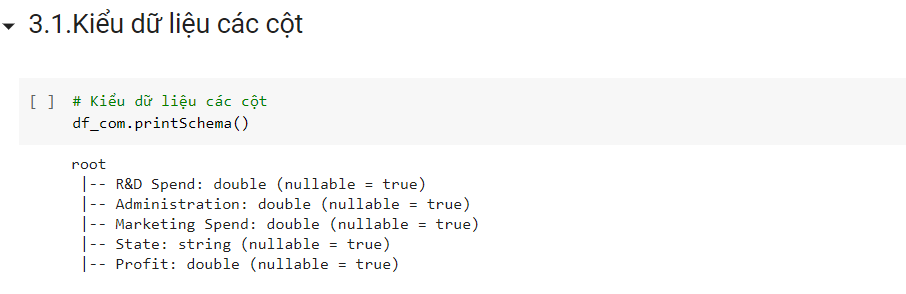
* + 1. **Check các dòng dữ liệu**



Hình : Kiểm tra số dòng dữ liệu

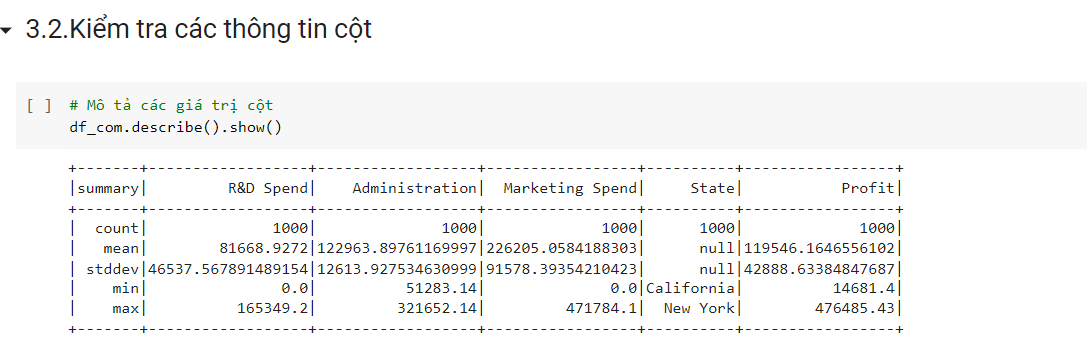
# **CHƯƠNG 3: XỬ LÝ DỮ LIỆU**

* 1. **Kiểm tra Kiểu dữ liệu các cột**



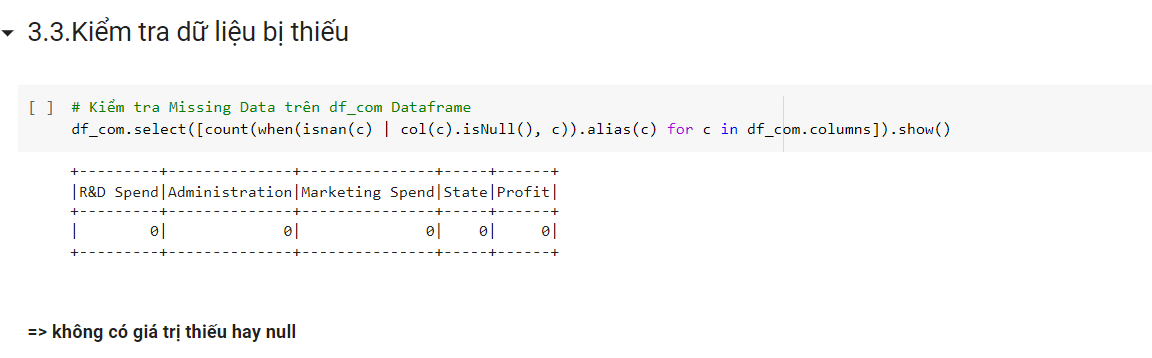
Hình : Kiểm tra dữ liệu các cột

* Kiểu dữ liệu của các cột, trùng với kiểu dữ liệu ban đầu từ Nguồn.
  1. **Kiểm tra thông tin các cột**



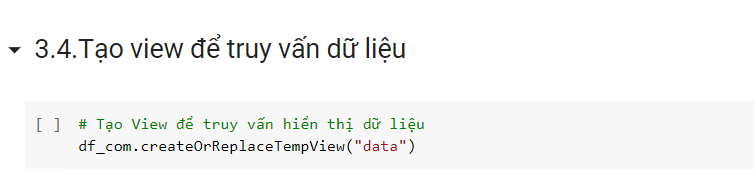
Hình : Kiểm tra thông tin các cột

* 1. **Kiểm tra dữ liệu bị thiếu**



Hình : Kiểm tra dữ liệu bị thiếu

* 1. **Tạo view để truy vấn dữ liệu**



Hình : Tạo view để truy vấn dữ liệu

* 1. **Phân loại các thuộc tính có trong tập dữ liệu**

Vì trọng tâm bài toán là dự đoán lợi nhuận dựa trên chi phí nghiên cứu và phát triển thuật toán bằng Hồi quy tuyến tính nên nhóm em sẽ tiến hành loại bỏ các trường dữ liệu không cần thiết chỉ giữ lại những trường có giá trị. Cụ thể các trường được giữ lại:

* RD Spend: Chi phí nghiên cứu và phát triển
* Profit: Lợi nhuận

Graphical user interface, text, application

Description automatically generated

Hình : Tạo dataframe để dùng cho việc dự đoán và đổi tên cột

* 1. **Kiểm tra giá trị cột RD Spend**



Hình : Kiểm tra giá trị cột RDSpend

* Không có giá trị khác thường
  1. **Kiểm tra giá trị cột Profit**

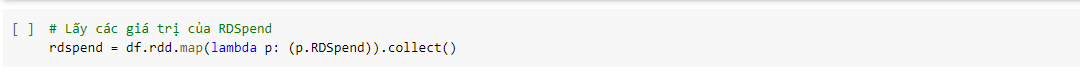


Hình : Kiểm tra giá trị cột Profit

**CHƯƠNG 4: CHECK OUTLINERS, LÀM SẠCH VÀ KIỂM SOÁT NGOẠI LỆ DỮ LIỆU**

## **4.1 Checkoutliners, làm sạch dữ liệu cho cột RD Spend**

**Bước 1:** Lấy các giá trị của RDSpend.



Hình : Lấy giá trị của RDSpend

**Bước 2:** Đưa các giá trị vừa lấy vào DataFrame.

Graphical user interface, text, application

Description automatically generated

Hình : Đưa giá trị vào df

**Bước 3:** Trực quan hóa giá trị tổng quan và ngoại lệ của thuộc tính RDSpend.

Chart, box and whisker chart

Description automatically generated

Hình : Biểu đồ giá trị trực quan và ngoại lệ của RDSpend

* Sau khi trực quan hóa dữ liệu, nhóm nhận thấy dữ liệu RDSpend là 1 tập dữ liệu đẹp, các giá trị phân bổ đồng đều và không có giá trị nào là bất thường.

## **4.2 Checkoutliners, làm sạch dữ liệu cho cột Profit**

**Bước 1:** Lấy các giá trị của Profit.

Text

Description automatically generated with medium confidence

Hình : Lấy giá trị của Profit

**Bước 2:** Đưa các giá trị vừa lấy vào DataFrame.

Graphical user interface, text, application

Description automatically generated

Hình : Đưa giá trị vào df

**Bước 3:** Trực quan hóa giá trị tổng quan và ngoại lệ của Profit.

Chart, box and whisker chart

Description automatically generated

Hình : Biểu đồ giá trị trực quan và ngoại lệ của Profit

* Sau khi trực quan hóa dữ liệu của Profit, ta có thể thấy dữ liệu phân bố chủ yếu ở mức từ 0 đến 200000, tuy nhiên, có 3 dữ liệu nằm ngoài 300000, nhóm sẽ tiến hành phân tích 3 dữ liệu này ở bước tiếp theo.

**Bước 4:** Sử dụng DataFrame để kiểm tra các giá trị Profit nằm ngoài 300000.

Graphical user interface, text, application

Description automatically generated

Hình : Kiểm tra giá trị ngoại lệ của Profit

* Sử dụng đối chiếu trên 2 DataFrame, ta có thể thấy, khi lợi nhuận càng cao ( >300000 ) thì chi phí bỏ ra để nghiên cứu và phát triển cũng rất cao, điều này là hoàn toàn đúng và hợp logic và có thể xảy ra trong thực tế. Nhóm sẽ xem đây là đặc tính của dữ liệu và không tiến hành xử lí.

# **CHƯƠNG 5: PHÂN TÍCH DỮ LIỆU**

**Bước 1:** Lấy ra các giá trị của RDSpend và Profit.

Graphical user interface, text, application

Description automatically generated

Hình : Lấy các giá trị của RDSpend và Profit

**Bước 2:** Đưa các giá trị trên vào DataFrame.

Graphical user interface, text, application

Description automatically generated

Hình : Đưa các giá trị vào df

* 1. **Trực quan hóa các chi phí bỏ ra để nghiên cứu và phát triển sản phẩm ( RDSpend)**

Chart, bar chart

Description automatically generated

Hình : Trực quan hóa các giá trị chi phí

* Với kết quả trực quan hóa, ta có thể thấy mức chi phí nằm chủ yếu ở (120000 cho đến 150000).
* Mức chi phí thấp nhất nằm ở trên 150000.
  1. **Trực quan hóa giữa Chi phí (RDSpend) và Lợi nhuận (Profit)**

Chart, bar chart

Description automatically generated

Hình : Trực quan hóa giữa chi phí và lợi nhuận

* Dựa vào sơ đồ trực quan hóa, ta có thể thấy, khi chi phí bỏ ra càng cao thì lợi nhuận thu lại càng lớn.

# **CHƯƠNG 6: CHUẨN BỊ DỮ LIỆU**

## **6.1 Phân chia dữ liệu thành 2 phần Feature (X) và Target (Y)**

Table

Description automatically generated

Hình : Chia dữ liệu thành Feature (X) và Target (y)

### **6.1.1 Show dữ liệu X ( RDSpend)**

Table

Description automatically generated

Hình : Dữ liệu của X

### **6.1.2 Show dữ liệu Y ( Profit)**

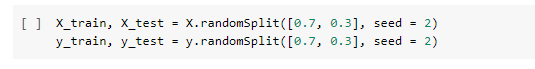
Table

Description automatically generated

Hình : Dữ liệu của y

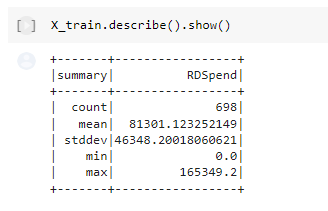
## **6.2 Chia dữ liệu Test Train theo theo tỉ lệ Test 30% và Train 70%**

**Bước 1:** Chia tập dữ liệu Test và Train.



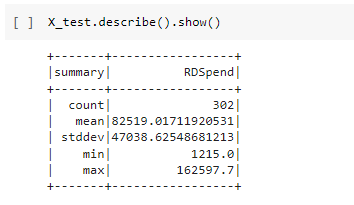
Hình : Chia dữ liệu thành Test và Train

**Bước 2:** Show các giá trị X Train.



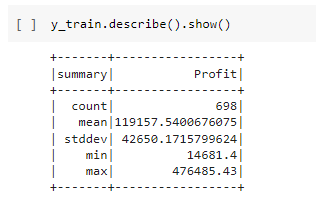
Hình : Các thông tin của X train

**Bước 3:** Show các giá trị X Test.



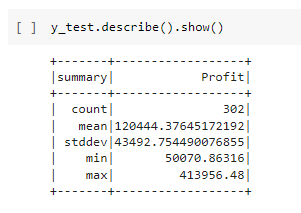
Hình : Các thông tin của X test

**Bước 4:** Show các giá trị Y Train.



Hình : Các thông tin của y train

**Bước 5:** Show các giá trị Y Test.



Hình : Các thông tin của y test

# **CHƯƠNG 7: XÂY DỰNG THUẬT TOÁN LINEAR REGRESSION**

## **7.1 Xây dựng thuật toán Linear Regression ( Hướng làm 1)**

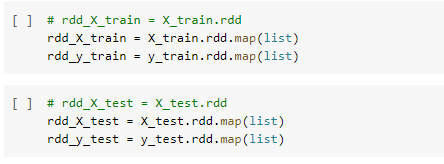
### **7.1.1 Tạo mô hình hồi quy tuyến tính ( Linear Regression )**

Mô hình hồi quy tuyến tính đơn giản là một đường được xác định bởi các hệ số được ước tính từ dữ liệu huấn luyện. Sau khi các hệ số được ước tính, chúng ta có thể sử dụng chúng để đưa ra dự đoán.

Phương trình đưa ra dự đoán với mô hình hồi quy tuyến tính đơn giản như sau:

Nhóm sẽ bắt đầu tính các hệ số B0 và B1 cho mô hình hồi quy tuyến tính.

**Bước 1:** Sau khi đã phân tách dữ liệu ra làm 2 tập training và testing, nhóm tiến hành đưa dữ liệu lên trên RDD theo dạng list để tiến hành xử lý.



Hình : Đưa giá trị của 2 tập train test lên RDD theo dạng list

**Bước 2:** Kiểm tra giá trị X train và Y train.

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

Description automatically generated

Hình : Số lượng dữ liệu trong train test

**Bước 3:** Xây dựng hàm tính giá trị trung bình và mẫu số B1 của cả biến đầu vào và đầu ra từ dữ liệu huấn luyện.

Giá trị trung bình là trung bình của một danh sách các số

Công thức:

Text

Description automatically generated with medium confidence

Dưới đây là một hàm có tên là ***mean()*** thực hiện tính giá trị trung bình cho danh sách các số.

Graphical user interface, text

Description automatically generated

Hình : Hàm tính giá trị trung bình (Mean)

Công thức tổng tính hệ số B1 (slope):

Text

Description automatically generated

Công thức tính mẫu B1:

Text

Description automatically generated

Graphical user interface, text, application

Description automatically generated

Hình : Hàm tính mẫu số B1 (Slope)

**Bước 4:** Tính mẫu số và Mean(x).

Graphical user interface, text, application

Description automatically generated

Hình : Tính mẫu số và mean của x train

**Bước 5:** Tính mẫu số và Mean(y)

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

Description automatically generated

Hình : Tính mẫu số và mean của y train

**Bước 6:** Thể hiện các giá trị vừa tính được

A picture containing diagram

Description automatically generated

Hình : Giá trị mẫu số và mean tính được

**Bước 7:** Tiến hành thực hiện công thức tính tử số B1 ở Bước 3.

Text

Description automatically generated with medium confidence

Text

Description automatically generated

Hình : Hàm tính tử số B1 (Slope)

**Bước 8:** Thể hiện giá trị tính được.

A picture containing text

Description automatically generated

Hình : Giá trị tử số tính được

**Bước 9:** Tính toán hệ số B0 và B1.

Text

Description automatically generated

Hình : Hàm tính hệ số B0 và B1

**Bước 10:** Hiển thị hệ số của mô hình.

Graphical user interface, text, application

Description automatically generated

Hình : Hệ số B0 và B1 tính được

**Bước 12:** Tiến hành xây dựng hàm dự đoán lợi nhuận dựa trên mô hình Hồi quy tuyến tính đơn giản.

Dưới đây là hàm có tên ***simple\_linear\_regression()*** thực hiện phương trình dự đoán để đưa ra dự đoán trên tập dữ liệu thử nghiệm. Nó cũng liên kết với nhau ước tính các hệ số trên dữ liệu đào tạo từ các bước trên.

Text

Description automatically generated

Hình : Hàm tính toán dựa trên công thức hồi quy tuyến tính

### **7.1.2 Ứng dụng mô hình hồi quy tuyến tính trên tập dữ liệu Test**

**Bước 1:** Chuyển RDD Test thành danh sách các giá trị (dạng List).

Graphical user interface, application, Word

Description automatically generated

Hình : Chuyển rdd\_X\_test thành danh sách các giá trị

**Bước 2:** Hiển thị thời gian dự đoán kết quả lợi nhuận.

Text

Description automatically generated

Hình : Dự đoán kết quả và tính toán thời gian chạy

**Bước 3:** Xuất kết quả dự đoán lợi nhuận trên mô hình hồi quy tuyến tính.

Table

Description automatically generated with medium confidence

Hình : Kết quả dự đoán

**Bước 4:** Chuyển RDD y\_test thành danh sách các giá trị.

Graphical user interface, text, application

Description automatically generated

Hình : Chuyển rdd\_y\_test thành danh sách các giá trị

**7.1.3** **Tạo và ứng dựng chỉ số đánh giá RMSE - Root Mean Square Error**

Để đánh giá được độ chinh xác của mô hình cũng sẽ xây dựng thêm một hàm để quản lý việc đánh giá các dự đoán được gọi là ***evaluate\_algorithm()*** và một hàm khác để ước tính *Sai số toàn phương trung bình căn – Root Mean Square Error* của các dự đoán được gọi là ***rmse\_metric().***

Công thức:

Text

Description automatically generated

* là giá trị ước lượng
* là biến độc lập
* n = (N – k – 1)
* N: số tổng lượng quan sát
* K: tổng lượng biến

Text, letter

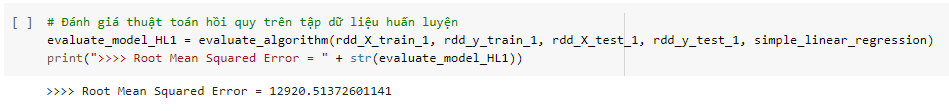
Description automatically generated

Hình : Hàm tính chỉ số RMSE

Text, letter

Description automatically generated

Hình : Hàm đánh giá thuật toán trên tập train



Hình : Đánh giá thuật toán

### **7.1.4 Trực quan hóa và so sánh kết quả dự đoán với thực tế**

**Bước 1:** Hiển thị giá trị gốc

Table

Description automatically generated

Hình : Giá trị gốc

**Bước 2:** So sánh giá trị Profit dự đoán và giá trị gốc

Table

Description automatically generated with low confidence

Hình : So sánh giá trị dự đoán và giá trị gốc

**Bước 3:** Trực quan hóa so sánh kết quả dự đoán và thực tế

Chart, line chart

Description automatically generated

Hình : Trực quan giá trị dự đoán và giá trị thực tế

## **7.2 Xây dựng thuật toán Linear Regression ( Hướng làm 2)**

### **7.2.1 Tạo mô hình hồi quy tuyến tính ( Linear Regression )**

**Bước 1:** Tạo 2 tập dữ liệu training data và testing data.

Tạo ra 1 tập dữ liệu rdd\_train và rdd\_test bằng hàm zip() bởi 2 tập dữ liệu x (rdd\_X\_train\_1) và tập dữ liệu y (rdd\_y\_train\_1).

**rdd\_train** và **rdd\_test** sẽ có từng dòng dữ liệu sẽ là các cặp khóa giá trị (key – value).

Với số lượng dòng dữ liệu của **rdd\_train** và **rdd\_test** lần lượt là 698 và 302.

Text

Description automatically generated

Hình : Tạo tập dữ liệu train test

**Bước 2:** Định nghĩa hàm Tính hệ số ước lượng - estimate\_coef()

Đây là hàm tính toán và trả về 2 hệ số ước tính: b\_0 và b\_1.

Với Phương trình của **đường hồi quy** được biểu diễn như sau:

(1)

Chúng ta cần phải tính:

**|**

và

**|**

**Trong đó:**

* SS\_xy: là tổng các độ lệch chéo (cross-deviations) của y và x:
* SS\_xx: là tổng các độ lệch bình phương của x:
* ̅y: là giá trị trung bình của y.
* ̅x: là giá trị trung bình của x.
* n: là số lượng các giá trị x hoặc y.

Text

Description automatically generated

Hình : Hàm tính toán hệ số B0 và B1

**Bước 3:** Tính hệ số ước lượng B1 và B0

Đầu vào sẽ là tập rdd\_train trước đó. Đầu ra sẽ là 2 chỉ số ước lượng của công thức tính phương trình hồi quy.

Text

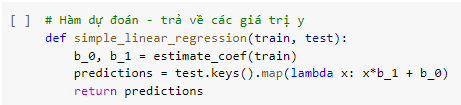
Description automatically generated

Hình : Hệ số B0 và B1 tính được

**Bước 4:** Định nghĩa hàm dự đoán các giá trị y - simple\_linear\_regression()

Hàm **simple\_linear\_regression()** có đối số đầu vào là 2 tập dữ liệu RDD đại diện 2 tập dữ liệu training data và testing data.

Hàm này sẽ gọi tới hàm **estimate\_coef()** để trả về 2 giá trị **b\_0** và **b\_1**. Sau đó sẽ thực hiện tính toán với các dữ liệu x của testing data và trả về các giá trị y được dự đoán (tạm gọi là **y\_predict**).



Hình : Hàm tính toán dựa trên công thức hồi quy tuyến tính

### **7.2.2 Ứng dụng mô hình dự đoán trên tập Test**

Thực hiện việc tính toán tập dữ liệu y\_predict

Tại đây, chúng ta sẽ chạy hàm **simple\_linear\_regression()** và trả về tập dữ liệu y\_predict cũng như sẽ tính toán thời gian chạy thuật toán.

Text

Description automatically generated

Hình : Dự đoán kết quả và tính toán thời gian chạy

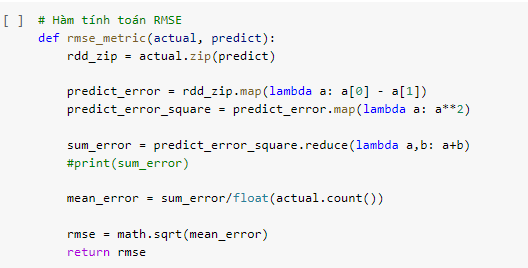
**7.2.3** **Tạo và ứng dựng chỉ số đánh giá RMSE - Root Mean Square Error**

**Bước 1:** Định nghĩa hàm tính RMSE – rmse\_metric()

Hàm có đối số đầu vào là 2 tập dữ liệu **y\_actual** (y thực tế - là các giá trị y trong tập test data) và và **y\_predict** (các giá trị y được dự đoán từ phương trình hồi quy).

Hàm sẽ trả về độ sai số giữa y\_actual và y\_predict.

Công thức được sử dụng để tính RMSE:



Hình : Hàm tính chỉ số RMSE

**Bước 2:** Định nghĩa hàm đánh giá thuật toán - evaluate\_algorithm()

Hàm có đối số đầu vào là 2 tập RDD của 2 tập dữ liệu train data và test data.

Trong thân hàm sẽ gọi tới hàm rmse\_metric() để lấy và trả về giá trị sai số RMSE.

Text

Description automatically generated

Hình : Hàm đánh giá thuật toán hồi quy trên tập train

**Bước 3:** Tính toán chỉ số RMSE

Graphical user interface, text

Description automatically generated

Hình : Đánh giá thuật toán hồi quy

**7.2.4 Trực quan hóa và so sánh kết quả dự đoán với thực tế**

**Bước 1:** Lấy giá trị gốc và dự đoán

Graphical user interface, application

Description automatically generated

Hình : Lấy giá trị gốc và dự đoán

**Bước 2:** So sánh giá trị Profit dự đoán và giá trị gốc

Table

Description automatically generated

Hình : So sánh giá trị dự đoán với giá trị gốc

**Bước 3:** Trực quan hóa so sánh kết quả dự đoán và thực tế.

Chart, line chart

Description automatically generated

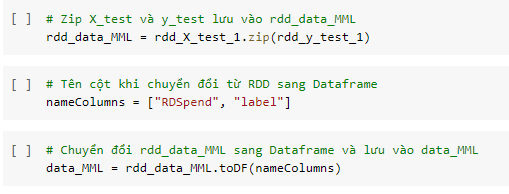
Hình : Trực quan giá trị dự đoán với giá trị gốc

## **7.3 Xây dựng thuật toán Linear Regression ( Hướng làm 3 sử dụng thư viện Mllib)**

Để khách quan về tính chính xác của thuật toán Hồi quy tuyến tính mà nhóm đã xây dựng, nhóm tiến hành xử dụng tập dữ liệu test áp dụng với thuật toán Hồi quy tuyến tính trên thư viện MLlib.

### **7.3.1 Chuẩn bị dữ liệu**

**Bước 1:** Chuyển đổi RDD sang DataFrame



Hình : Chuyển dữ liệu RDD sang DF

**Bước 2:** Chuyển đổi giá trị cột Feature thông qua VectorAssembler

Table

Description automatically generated

Hình : Chuyển giá trị cột RDSpend vào cột Features

**Bước 3:** Hiển thị dữ liệu

Table

Description automatically generated

Hình : Thông tin của dữ liệu

**7.3.2** **Ứng dụng mô hình hồi quy tuyến tính trên tập dữ liệu test**

Graphical user interface, text, application

Description automatically generated

Hình : Thực hiện mô hình hồi quy tuyến tính trên tập test

Table

Description automatically generated

Hình : Thêm giá trị dự đoán vào DF

**7.3.3 Tạo và ứng dựng chỉ số đánh giá RMSE - Root Mean Square Error**

Text

Description automatically generated

Hình : Đánh giá mô hình bằng RMSE

### **7.3.4 Trực quan hóa và so sánh kết quả dự đoán với thực tế**

**Bước 1:** Đưa giá trị test và dự đoán vào DataFrame để tiến hành so sánh và trực quan hóa.

Graphical user interface, text, application

Description automatically generated

Hình : So sánh giá trị dự đoán và giá trị thực tế

**Bước 2:** Trực quan hóa so sánh kết quả trên tập test và kết quả dự đoán

Chart, line chart

Description automatically generated

Hình : Trực quan giá trị dự đoán và giá trị thực tế

# **CHƯƠNG 8: SO SÁNH THUẬT TOÁN**

* 1. **So sánh, đánh giá các hướng làm**

Nhóm sẽ tiến hành so sánh và đánh giá 2 hướng mô hình làm khác nhau mà nhóm đã thực hiện so với mô hình sử dụng thuật toán Linear Regression do thư viện MLlib cung cấp.

* + 1. **Thời gian thuật toán thực hiện**

Chart, bar chart

Description automatically generated

Hình : So sánh thời gian thực hiện

* + 1. **Độ chính xác thông qua chỉ số RMSE**

Chart, bar chart

Description automatically generated

Hình : So sánh độ chính xác bằng RMSE

* + 1. **Độ chính xác thông qua trực quan dữ liệu**

Chart

Description automatically generated with low confidence

Hình : So sánh độ chính xác qua việc dự đoán

**CHƯƠNG 9: KẾT LUẬN**

* 1. **Ưu và nhược điểm giữa các hướng làm**
* Thời gian thực thi dự đoán giữa hướng làm 1 và 2 nhanh hơn so với hướng làm 3.
* Độ chính xác, đúng đắn giữa giá trị dự đoán và giá trị kiểm thử trên 3 hướng làm gần như tương đồng với nhau.
  1. **Hướng phát triển**
* Nhóm sẽ tiến hành mở rộng mô hình Hồi quy tuyến tính đơn giản (Simple Linear Regression) với 2 hệ số thành mô hình Hồiquy tuyến tính phức tạp hơn với nhiều đối số truyền vào (Multiple Linear Regression) để có thể đưa ra được những dự đoán khách quan hơn.
* Ngoài ra, nhóm cũng sẽ xây dựng thêm một số mô hình dự đoán đơn (Single Model) chẳng hạn như Decision Tree,.. và mô hình tập hợp (Embedded Model) chẳng hạn như Random Forest. XGBoots để so sánh tính chính xác, khả năng dự đoán của các mô hình trên cùng một tập dữ liệu.

**CHƯƠNG 10: TÀI LIỆU THAM KHẢO**

**[1]** – spark.apache.org, "**RDD Programming Guide**", “[*https://spark.apache.org/docs/latest/rdd-programming-guide.html*](https://spark.apache.org/docs/latest/rdd-programming-guide.html)”

**[2]** – machinelearningmastery.com, "**How To Implement Simple Linear Regression From Scratch With Python**", “[*https://machinelearningmastery.com/implement-simple-linear-regression-scratch-python*/](https://machinelearningmastery.com/implement-simple-linear-regression-scratch-python/)”

**[3]** – medium.com, " **Linear Regression Full Tutorial Guide**", “<https://medium.com/@rupeshthetech/linear-regression-full-tutorial-guide-f32ac17caa9f>”

**[4] –** geeksforgeeks.org**, “Linear Regression (Python Implementation)”, “**[*https://www.geeksforgeeks.org/linear-regression-python-implementation*/](https://www.geeksforgeeks.org/linear-regression-python-implementation)”