**BỘ GIÁO DỤC VÀ ĐÀO TẠO**

**TRƯỜNG ĐẠI HỌC SƯ PHẠM KỸ THUẬT THÀNH PHỐ HỒ CHÍ MINH**

**KHOA CÔNG NGHỆ THÔNG TIN**

-----🙞🙞🙞🙞🙞-----A logo with a ball and a sphere

Description automatically generated

BÁO CÁO CUỐI KỲ MÔN KHAI PHÁ DỮ LIỆU

**ĐỀ TÀI:**

**XÂY DỰNG CÁC MÔ HÌNH PHÂN TÍCH CẢM XÚC CỦA KHÁCH HÀNG DỰA TRÊN BÌNH LUẬN VÀ CÁC MÔ HÌNH CHO VIỆC DỰ ĐOÁN CHỈ SỐ HẠNH PHÚC CỦA CÁC QUỐC GIA**

**GVHD:** Ths. Trần Trọng Bình

**Lớp:** Thứ 7 (tiết 11 – 14)

**Mã lớp:** DAMI330484\_23\_2\_02

Sinh viên thực hiện: (nhóm 5)

Đặng Nguyễn Quang Huy MSSV: 21133036

Nguyễn Trọng Dũng MSSV: 21133021

Đái Triệu Phi MSSV: 21133068

          Huỳnh Gia Hân         MSSV: 21133031

*Thành phố Hồ Chí Minh, tháng 05, năm 2024*

|  |  |
| --- | --- |
| BỘ GIÁO DỤC VÀ ĐÀO TẠO  **TRƯỜNG ĐẠI HỌC SƯ PHẠM KỸ THUẬT**  **THÀNH PHỐ HỒ CHÍ MINH** | **CỘNG HÒA XÃ HỘI CHỦ NGHĨA VIỆT NAM**  **Độc lập – Tự do – Hạnh phúc** |

**NHẬN XÉT CỦA GIẢNG VIÊN HƯỚNG DẪN**

**Họ và tên sinh viên thực hiện**

* Đặng Nguyễn Quang Huy - 21133036
* Nguyễn Trọng Dũng – 21133021
* Đái Triệu Phi - 21133068
* Huỳnh Gia Hân - 21133031

**Chuyên ngành:** Kỹ thuật dữ liệu (Data Engineering)

**Đề tài:** Xây dựng các mô hình phân tích cảm xúc của khách hàng dựa trên bình luận và các mô hình cho việc dự đoán chỉ số hạnh phúc của các quốc gia.

**Môn học:** Khai phá dữ liệu (Data Mining)

**Nhận xét**

Tp HCM, / / 2024

Giảng viên hướng dẫn

(Tên và chữ ký)

Trần Trọng Bình

**LỜI CẢM ƠN**

Bài báo cáo này là sản phẩm của một quá trình học tập và làm việc nhóm. Để có thể hoàn thành bài báo cáo, chúng em đã nhận được rất nhiều sự hỗ trợ từ thầy và các bạn. Do đó nhóm chúng em xin được phép gửi lời cảm ơn chân thành và sâu sắc nhất đến:

*1) Thầy Trần Trọng Bình – giảng viên bộ môn Khai phá dữ liệu (Data Mining), khoa Công nghệ Thông tin trường Đại học Sư phạm Kỹ thuật thành phố Hồ Chí Minh. Nhóm xin cảm ơn sự quan tâm và giúp đỡ tận tình của thầy trong suốt quá trình giảng dạy. Cảm ơn thầy đã luôn giải đáp những thắc mắc cũng như đưa ra những nhận xét, góp ý giúp nhóm thực hiện cải thiện chất lượng công việc của nhóm.*

*2) Thư viện trường Đại học Sư phạm Kỹ thuật thành phố Hồ Chí Minh, nơi cung cấp môi trường học tập, nghiên cứu và làm việc nhóm để chúng em có thể hoàn thành tốt báo cáo của nhóm mình.*

*3) Các bạn học cùng lớp đã có những nhận xét, đóng góp về mặt kiến thức lẫn tinh thần cho nhóm.*

Đề tài và bài báo cáo được chúng em thực hiện trong khoảng thời gian ngắn, với những hạn chế khác về mặt kiến thức, kỹ thuật và kinh nghiệm trong việc thực hiện báo cáo. Do đó, trong quá trình làm đề tài có những thiếu sót là điều không thể tránh khỏi. Chúng em rất mong nhận được những ý kiến đóng góp quý báu của thầy để kiến thức của chúng em được hoàn thiện hơn và làm tốt hơn nữa trong những lần sau. Chúng em xin chân thành cảm ơn!

**MỤC LỤC**

[I. TỔNG QUAN VỀ DỰ ÁN 1](#_Toc167916232)

[1.1. Giới thiệu về các bài toán 1](#_Toc167916233)

[1.2. Giới thiệu về tập dữ liệu 3](#_Toc167916234)

[II. XÂY DỰNG CÁC MÔ HÌNH PHÂN TÍCH CẢM XÚC CỦA KHÁCH HÀNG VÀ ĐI TÌM KHUÔN MẪU DỮ LIỆU 4](#_Toc167916235)

[2.1. Tiền xử lý dữ liệu dữ liệu bình luận 4](#_Toc167916236)

[2.2. Khám phá và phân tích tập dữ liệu 6](#_Toc167916237)

[2.2.1. Phân tích thống kê mô tả tập dữ liệu 6](#_Toc167916238)

[2.2.2. Phân tích thống kê suy luận 10](#_Toc167916239)

[2.3. Áp dụng mô hình machine learning và deep learning 15](#_Toc167916240)

[2.3.1. Các kĩ thuật biểu diễn từ trong không gian vector 15](#_Toc167916241)

[2.3.2. Xây dựng mô hình đi tìm khuôn mẫu dữ liệu 20](#_Toc167916242)

[2.3.3. Mô hình K-Nearest Neighbors (KNN) cho bài toán phân tích cảm xúc 21](#_Toc167916243)

[2.3.4. Mô hình LSTM cho bài toán phân tích cảm xúc 22](#_Toc167916244)

[2.3.5. Mô hình lai ghép giữa LSTM và CNN cho bài toán phân tích cảm xúc 23](#_Toc167916245)

[2.4. Triển khai mô hình phân tích cảm xúc bằng Flask 25](#_Toc167916246)

[III. XÂY DỰNG CÁC MÔ HÌNH CHO VIỆC DỰ ĐOÁN CHỈ SỐ HẠNH PHÚC CỦA CÁC QUỐC GIA 25](#_Toc167916247)

[3.1. Tiền xử lý dữ liệu 25](#_Toc167916248)

[3.2. Khám phá và phân tích tập dữ liệu 27](#_Toc167916249)

[3.2.1. Phân tích thống kê suy luận 27](#_Toc167916250)

[3.2.2. Phân tích thống kê mô tả tập dữ liệu sau khi đã tiền xử lý 31](#_Toc167916251)

[3.3. Xây dựng các mô hình machine learning và deep learning để dự đoán chỉ số điểm hạnh phúc ở các quốc gia 31](#_Toc167916252)

[3.3.1. Mô hình FNN (Feedforward Neural Network) 32](#_Toc167916253)

[3.3.2. Mô hình Random Forest Regressor 32](#_Toc167916254)

[3.3.3. Mô hình K - Nearest Neighbors (KNN) 33](#_Toc167916255)

[3.4. Chọn ra mô hình tốt nhất cho việc dự đoán chỉ số hạnh phúc 34](#_Toc167916256)

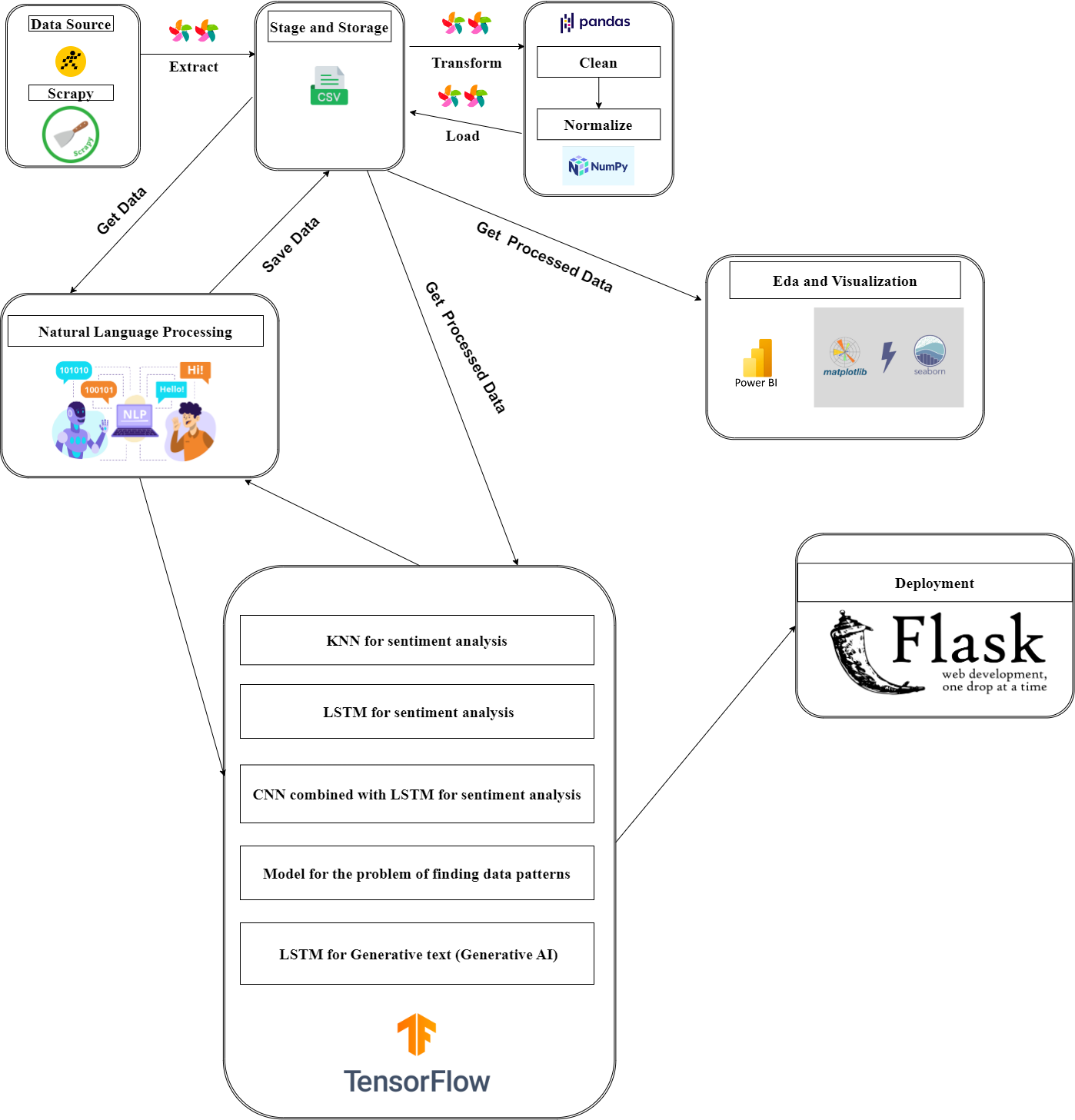
[IV. BẢNG PHÂN CÔNG CÔNG VIỆC 35](#_Toc167916257)

# I. TỔNG QUAN VỀ DỰ ÁN

## 1.1. Giới thiệu về các bài toán

**a) Xây dựng các mô hình phân tích cảm xúc của khách hàng dựa trên những bình luận của họ**

Quy trình công việc



Trong bài toán này sử dụng dữ liệu của một chuỗi cửa hàng bán điện thoại là Thế Giới Di Động. Để theo dõi mức độ yêu thích của khách hàng, chúng em đã thu thập các thông tin đánh giá trên tất cả sản phẩm và ở nhiều dạng khác nhau (xếp hạng, bình luận, …) trong một khoảng thời gian dài.

Mục tiêu của bài toán này chính là tìm được những ưu điểm và nhược điểm của sản phẩm thông qua đánh giá của khách hàng bằng cách xây dựng được các mô hình phân tích cảm xúc dựa trên bình luận của họ.Từ đó cải thiện quá trình kinh doanh và chất lượng sản phẩm theo đúng hướng, thu hút nhiều khách hàng và đem lại doanh thu cao.

**b) Xây dựng các mô hình đi tìm khuôn mẫu dữ liệu**

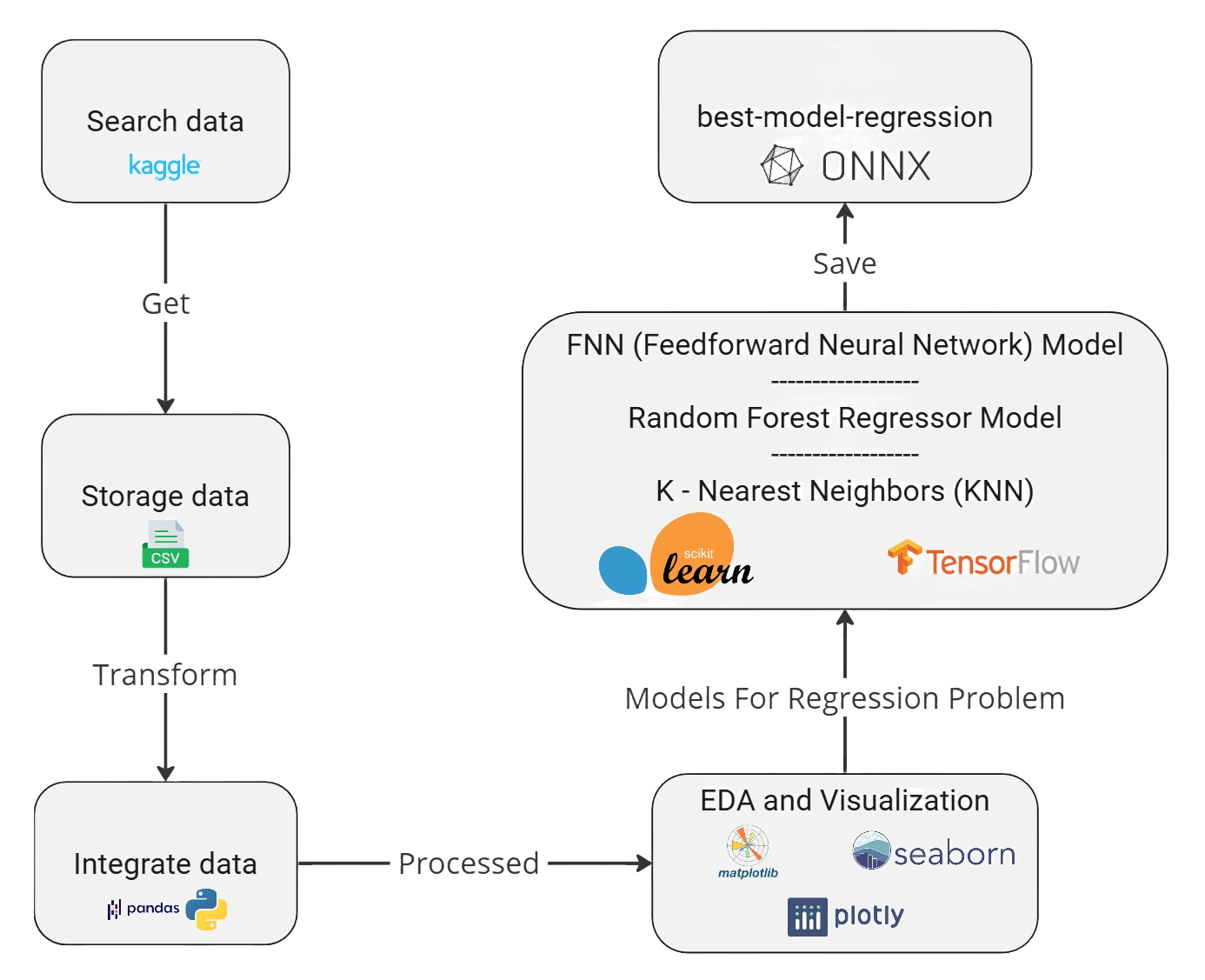
Phân cụm là một phương pháp học máy không giám sát để nhóm các điểm dữ liệu có sự tương đồng cao với nhau. Trong báo cáo này, chúng ta sẽ thảo luận về quá trình phân cụm để gắn nhãn cho tập dữ liệu sentiment, bao gồm các bước tiền xử lý dữ liệu, sử dụng các thuật toán phân cụm khác nhau, và cuối cùng là gắn nhãn và đánh giá kết quả.

**c) Xây dựng các mô hình cho việc dự đoán chỉ số hạnh phúc của các quốc gia**

Trong bài toán này sử dụng dữ liệu là The World Happiness Report được tổng hợp theo từng năm từ 2015 đến 2022. Bộ dữ liệu đã từng được công bố tại Liên Hợp Quốc trong sự kiện kỷ niệm Ngày Quốc tế Hạnh phúc vào ngày 20 tháng 3, có được sự công nhận trên toàn cầu khi các chính phủ, tổ chức và xã hội dân sự ngày càng sử dụng các chỉ số hạnh phúc để làm cơ sở cho các quyết định hoạch định chính sách của họ.

Mục tiêu của bài toán này là tìm hiểu và dự đoán mức độ điểm hạnh phúc của các quốc gia trên thế giới thông qua các thuộc tính liên quan bằng các mô hình học máy (machine learning) và các yếu tố nào có ảnh hưởng đáng kể đến mức độ hạnh phúc này. Từ đó có thể giúp các nhà nghiên cứu và nhà hoạch định chính sách hiểu rõ hơn về những yếu tố ảnh hưởng đến hạnh phúc của người dân và có thể đưa ra các biện pháp cải thiện chất lượng cuộc sống ở quốc gia của mình.

Quy trình thực hiện:



## 1.2. Giới thiệu về tập dữ liệu

**a) Xây dựng các mô hình cho bài toán phân tích cảm xúc của khách hàng và bài toán đi tìm khuôn mẫu dữ liệu**

Tập dữ liệu về bình luận khách hàng gồm 2 tệp csv phục vụ cho việc phân tích và xây dựng các mô hình học máy là Test.csv và Train.csv. Độ lớn của từng tập dữ liệu là:

* Test.csv: 2224 hàng và 5 cột
* Trainfull.csv: 8898 hàng và 5 cột (Train.csv và Dev.csv)

Trong mỗi tập dữ liệu sẽ gồm 5 thuộc tính:

* index: mã số bình luận
* comment: bình luận dạng văn bản tiếng Việt
* n\_star: số sao mà khách hàng đánh giá cho sản phẩm, từ 0 đến 5
* date\_time: thời điểm của mỗi bình luận
* label: đánh giá các chức năng theo trải nghiệm của mỗi khách hàng, gồm 3 mức độ là Positive, Neutral và Negative

**b) Xây dựng các mô hình cho việc dự đoán chỉ số hạnh phúc của các quốc gia**

Tập dữ liệu ban đầu gồm 5 tệp .csv chứa dữ liệu được tổng hợp từ năm 2015 đến năm 2019, tại [World Happiness Report](https://www.kaggle.com/datasets/unsdsn/world-happiness) với các file lần lượt là 2015.csv, 2016.csv, 2017.csv, 2018.csv và 2019.csv. Nhưng sau khi qua tổng hợp lại dữ liệu với các thuộc tính cần thiết, nhận thấy chỉ có được 782 hàng, nên nhóm đã tiến hành đi tìm thêm nguồn dữ liệu để tích hợp vào. Sau khi tìm được tại [World Happiness Report up to 2022](https://www.kaggle.com/datasets/mathurinache/world-happiness-report), nơi có thêm 3 tập dữ liệu nữa từ năm 2020 đến 2022 lần lượt là 2020.csv, 2021.csv và 2022.csv, nhóm đã tổng hợp lại và đạt được kết quả tập dữ liệu là: 1231 hàng và 9 cột.

Các cột trong tập dữ liệu được tích hợp lại bao gồm:

* Country: Tên của đất nước được thực hiện khảo sát
* Happiness Score: Một thước đo được đo lường trên thang điểm từ 0 đến 10, và đây cũng là biến mục tiêu mà ta cần dự đoán
* Economy (GDP per Capita): Mức độ đóng góp của GDP vào việc tính Happiness Score.
* Family: Mức độ mà Family đóng góp vào việc tính Happiness Score
* Health (Life Expectancy): Mức độ mà Health đóng góp vào việc tính Happiness Score
* Freedom: Mức độ mà Freedom đóng góp vào việc tính Happiness Score
* Trust (Government Corruption): Mức độ mà Trust (Government Corruption) đóng góp vào Happiness Score
* Generosity: Mức độ đóng góp của Generosity vào việc tính Happiness Score
* year: Năm thực hiện khảo sát

# II. XÂY DỰNG CÁC MÔ HÌNH PHÂN TÍCH CẢM XÚC CỦA KHÁCH HÀNG VÀ ĐI TÌM KHUÔN MẪU DỮ LIỆU

## 2.1. Tiền xử lý dữ liệu dữ liệu bình luận

Dữ liệu được cung cấp chỉ là dữ liệu dạng thô, không có giá trị hoặc rất khó để phân tích. Do đó, quy trình tiền xử lý dữ liệu để làm sạch trước khi thực hiện khám phá hoặc phân tích chuyên sâu là vô cùng quan trọng

Trong tập dữ liệu này, các kỹ thuật xử lý dữ liệu ngôn ngữ đặc biệt cho tiếng Việt sẽ được áp dụng

* Loại bỏ các chấm câu và icon đặc biệt

Chấm câu và icon là một trong những ký tự không quan trọng trong văn bản và khó phân tích, nên việc loại bỏ nó sẽ làm giảm kích thước văn bản và từ điển

* Thống nhất dạng viết thường cho các ký tự

Một trong những kỹ thuật đặc biệt của dữ liệu văn bản chính là xây dựng kho từ điển, nó sẽ chứa tất cả từ đã xuất hiện trong cả bộ dữ liệu, mỗi từ là khác nhau kể cả in hoa hoặc in thường.

Nếu một từ vừa xuất hiện dạng in hoa và in thường, sẽ xuất hiện 2 lần trong từ điển, nó khiến cho từ điển trở nên lớn và dư thừa. Vì vậy, việc biến tất cả ký tự về một dạng in thường giúp từ điển nhỏ gọn hơn và thuận tiện cho việc phân tích

* Chuẩn hóa số về từ

Trong bài toán này , các số hầu như không có ý nghĩa lớn nó sẽ làm cho cái kho ngữ liệu của mình quá lớn và quá trình huấn luyện sẽ trở nên lâu hơn.Vì thế chúng ta cần biến đổi số thành “number”

* Loại bỏ các từ dừng

Trong ngôn ngữ tự nhiên, từ dừng thường xuyên xuất hiện nhiều trong các câu, chúng không mang ý nghĩa quan trọng trong câu. Vậy nên, việc loại bỏ chúng sẽ giúp câu nhỏ gọn và tập trung được làm nổi bật đặc trưng hơn

Kỹ thuật này sẽ không có tác dụng trong một số bài toán cần ngữ nghĩa hoặc cú pháp của cả câu. Điển hình như các bài toán tạo sinh văn bản hoặc dịch máy, chúng cần học thông tin cả ngữ nghĩa và cú pháp chứ không chỉ là đặc trưng của câu. Cho nên việc loại bỏ từ dừng sẽ giảm khả năng học và độ chính xác của mô hình

* Tokenize cho tiếng Việt

Tokenize là một kỹ thuật cơ bản, nó có nhiệm vụ tách từng từ trong một văn bản/câu thành một danh sách từ. Mỗi từ sẽ được gọi là một token, về sau mỗi từ sẽ được chuyển đổi thành một vector số đặc trưng

Đối với tiếng Việt, khác với tiếng Anh, chúng ta có các từ ghép, nghĩa là hai từ mới tạo nên một nghĩa hoàn chỉnh. Vì vậy, kỹ thuật để xử lý tokenize cho tiếng Việt nói riêng và các ngôn ngữ khác nói chung sẽ khác một phần

Trong dự án này, chúng ta sẽ dùng thư viện có sẵn Pyvi, là một thư viện chuyên xử lý tokenize cho tiếng Việt được chính người Việt thiết kế

* Loại bỏ từ trùng lặp

Trong một câu chúng ta nên loại bỏ từ trùng lặp sẽ giúp cho mô hình dễ đào tạo cũng như dễ nhận dạng hơn

* Chuyển đổi về định dạng datetime

Dữ liệu thời gian hiện tại đang ở dạng chuỗi, để dữ liệu có ý nghĩa về mặt thời gian, chúng ta cần chuyển đổi chúng về đúng định dạng datetime

* Gán nhãn cho mỗi bình luận

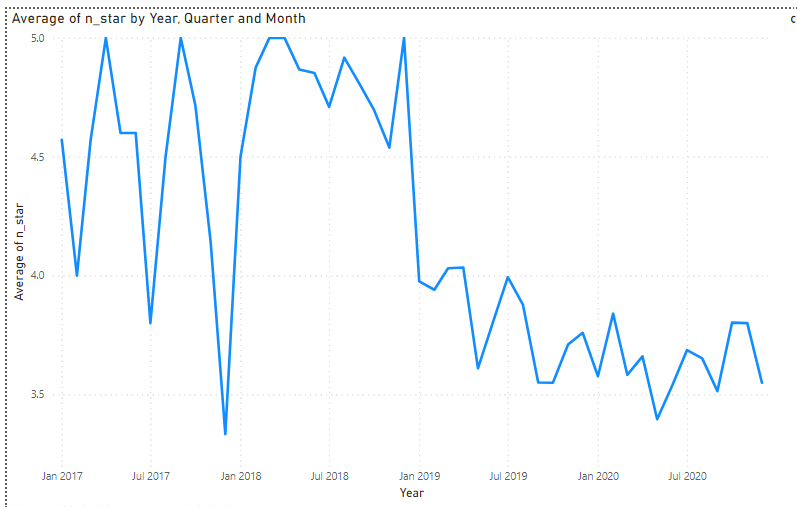
Để phân tích cũng như huấn luyện mô hình học máy, chúng ta cần các cột mục tiêu hoặc gọi là nhãn, trong tập dữ liệu gốc chúng ta chưa có bất kỳ cột nhãn nào nên ta sẽ tiến hành gán nhãn

Chúng ta để ý rằng, tại cột label gốc, dữ liệu là chuỗi các đánh giá tích cực/trung lập/tiêu cực cho từng đặc điểm sản phẩm. Chúng ta sẽ tiến hành đếm số lượng mỗi loại đánh giá, sau đó ước lượng gán nhãn cho bình luận đó là tích cực, trung lập hay tiêu cực.

## 2.2. Khám phá và phân tích tập dữ liệu

### **2.2.1. Phân tích thống kê mô tả tập dữ liệu**

Đầu tiên chúng ta cần biết trong thời gian gần đây, khách hàng đã có những đánh giá gì cho sản phẩm. Áp dụng biểu đồ đường với trục hoành là giá trị thời gian và trục tung là giá trị trung bình số sao



*(Hình 2.2.1.1: Biểu đồ thể hiện khách hàng có những đánh giá các sản phẩm theo thời gian)*

Dựa theo biểu đồ trên, trong những năm đầu 2017 đến đầu năm 2019, đánh giá của khách hàng cho các sản phẩm khá cao, trung bình khoảng 4.5. Nhưng khoảng thời gian từ giữa năm 2019 đến hết năm 2020, đánh giá trung bình đã giảm một cách đáng cảnh báo, chứng tỏ rằng khách hàng rất không hài lòng với chất lượng sản phẩm hoặc dịch vụ của chúng ta. Cần gấp rút kiểm tra lại các khâu quản lý chất lượng sản phẩm hoặc dịch vụ để cải thiện tình hình

Bên cạnh đó, chỉ số sao cũng chỉ là một con số, nó không thực sự chính xác vì người dùng có thể khai khống không đúng với chất lượng sản phẩm/dịch vụ. Chúng ta cần xem xét tiếp một thông số chi tiết hơn chính là đánh giá phân loại Positive/Neutral/Negative của chức năng sản phẩm, sau đó so sánh với chỉ số sao để có cái nhìn chính xác hơn

A graph with blue and orange bars

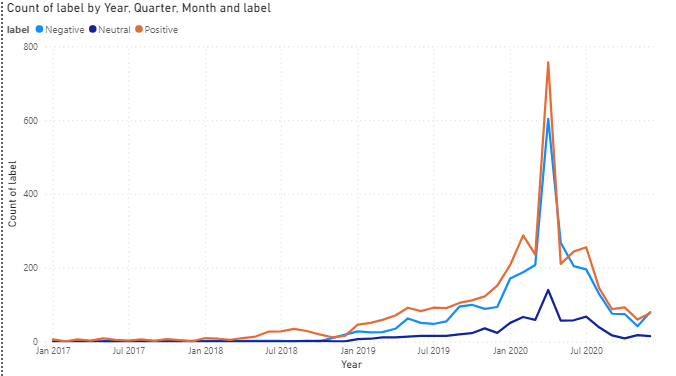
Description automatically generated

*(Hình 2.2.1.2: Biểu đồ thể hiện lượt đánh giá qua các label Positive/Neutral/Negative)*

Dựa vào biểu đồ thanh, chúng ta quan sát tại chỉ số 5 sao thì số lượng phân loại Positive dành cho các chức năng rất cao và áp đảo những loại khác. Tại chỉ số 4 sao, số lượng phân loại Negative và tiêu cực bằng ½ so với Positive.

Tuy nhiên, từ chỉ số 3 đến 1 sao thì phân loại Positive hầu như không có và chủ yếu là loại Negative. Chứng tỏ từ mức 3 đến 1 sao, chất lượng các chức năng của sản phẩm thật sự xuống cấp dẫn đến sự không hài lòng của khách hàng.

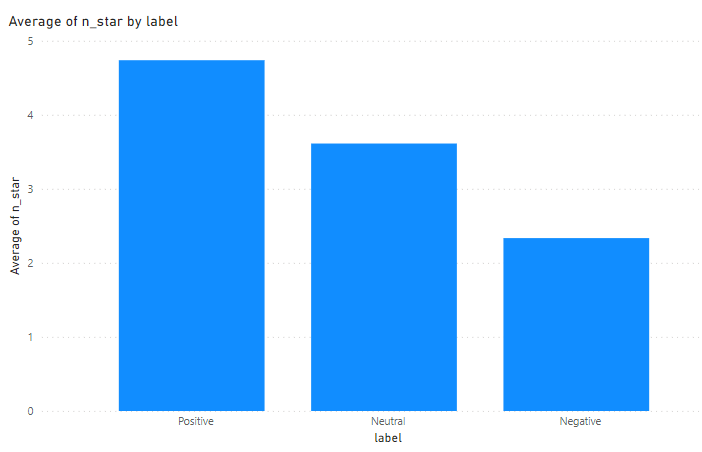
Nhưng chúng ta cần biết thêm rằng, tình trạng chất lượng sản phẩm đi xuống đã tồn tại trong thời gian dài hay không, hay chỉ mới gần đây mới gặp tình trạng như vậy. Ta áp dụng biểu đồ boxplot kết hợp với biểu đồ *Hình 3.1* bên trên để kết luận



*(Hình 2.2.1.3: Biểu đồ thể hiện số lượng label được đánh giá theo thời gian)*

Ta có thể thấy, số lượng đánh giá Positive luôn nằm trên những phân loại khác. Một điểm đáng lưu ý nữa là khoảng đầu năm 2019 đến giữa 2020, số lượng đánh giá phân loại tăng đột biến, đồng nghĩa đến việc số lượng khách hàng tăng vọt trong khoảng thời gian đó. Kết hợp với thông tin từ *Hình 3.1,* chứng tỏ rằng chất lượng sản phẩm nhìn chung khá tốt, có thể chỉ mới sa sút gần đây, trong những cuối năm 2019 và 2020.

Ngoài ra, chúng ta biết thêm rằng với phân loại Positive thì chỉ số sao ở khoảng 4.6, đối với Neutral sẽ là 3.7 và Negative là 2.3



*(Hình 2.2.1.4: Thống kê số lượng n\_star thông qua các label Postive/Neutral/Negative).*

### **2.2.2. Phân tích thống kê suy luận**

Quan sát phân phối của các trạng thái Positive, Negative và Neutral theo Length (Độ dài chuỗi đánh giá), Word\_count (Số lượng từ trong câu đánh giá), mean\_word\_length (Số lượng ký tự trung bình trong một từ) để thấy tương quan giữa trạng thái câu đánh giá và cách người dùng đánh giá.

A comparison of a graph

Description automatically generated with medium confidence

*(Hình 2.2.2.1: Biểu đồ phân phối xác suất theo word\_count - Bên trái, Biểu đồ mật độ xác suất theo word\_count - Bên phải)*

Quan sát biểu đồ thấy được với khoảng 10 đến 40 từ người dùng đã có thể diễn đạt thái độ của họ đối với sản phẩm. Nói cách khác, khoảng 10 đến 40 từ là khoảng thể hiện rõ nhất quan điểm của một câu đánh giá.

A comparison of a graph

Description automatically generated with medium confidence

*(Hình 2.2.2.2: Biểu đồ phân phối xác suất theo Length - Bên trái, Biểu đồ mật độ xác suất theo Length - Bên phải)*

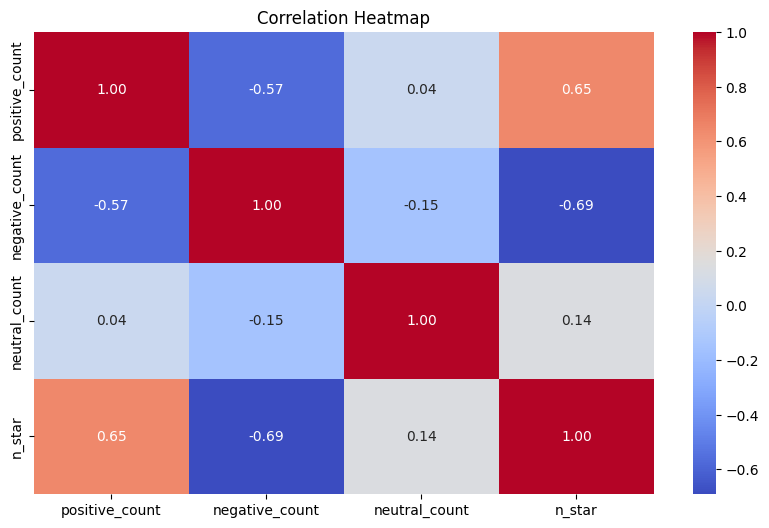
Cũng giống như phân bố theo *word\_count,* Positive, Negative và Neutral tập trung tại giá trị Length 100. Càng xa giá trị này, các câu đánh giá càng không nêu rõ thái độ.

A diagram of a graph

Description automatically generated with medium confidence

*(Hình 2.2.2.3: Biểu đồ phân phối xác suất theo mean\_word\_length - Bên trái, Biểu đồ mật độ xác suất theo mean\_word\_length - Bên phải)*

Ngoài các yếu tố về kích thước của câu đánh giá, số lượng từ ngữ Positive, Negative và Neutral cũng là yếu tố quan trọng thể hiện điểm đánh giá của người dùng (n\_star). Dưới đây là biểu đồ tương quan giữa positive\_count, negative\_count, neutral\_count, n\_star:



*(Hình 2.2.2.3: Ma trận tương quan heatmap cho các cột positive\_count, negative\_count, neutral\_count, n\_star)*

Biểu đồ cho thấy giá trị *positive\_count* tương quan thuận với n\_star (chỉ số tương quan 0.65) nghĩa là trong đánh giá từ người dùng số từ *positive\_count* càng nhiều khả năng càng cao người dùng đó sẽ cho rating cao. Ngược lại, giá trị negative\_counttương quan nghịch với n\_star (chỉ số tương quan -0.69) nghĩa là trong đánh giá từ người dùng số từ *negative\_count* càng nhiều đó sẽ là một rating thấp sao.

A graph of a number of numbers

Description automatically generated

*(Hình 2.2.2.4: Biểu đồ đường thể hiện số lượng đánh giá cho từng trạng thái theo số từ)*

Quan sát thấy được rằng, người dùng có xu hướng dùng khoảng dưới 40 từ để đánh giá. Lượng đánh giá Positive luôn cao hơn Negative, đây là một dấu hiệu tốt đối với lĩnh vực kinh doanh sản phẩm.

Từ các phân tích trên, ta thấy rằng trong phân tích cảm xúc từ bình luận, các yếu tố về số từ, từ *positive\_count, negative\_count* ảnh hưởng rất lớn. Đồng thời, các yếu tố có sự tương quan rõ rệt, do đó, trong quá trình xây dựng mô hình, cần tập trung khai thác các đặc trưng đã đề cập.

Bên cạnh đó, nhóm tiến hành phân tích về các *comment* của người dùng dành cho sản phẩm.

A close up of words

Description automatically generated

*(Hình 2.2.2.5: Biểu đồ đám mây thể hiện các từ khóa liên quan đến trải nghiệm sản phẩm của các người dùng)*

Các từ khóa được sắp xếp theo kích thước, từ lớn nhất đến nhỏ nhất, dựa trên tần suất xuất hiện trong văn bản. Từ khóa lớn nhất trong biểu đồ là "máy", "mới", "tốt", "pin", "chụp", "camera", "game", "mượt", "giá", "thích", "đẹp", "nhân viên". Nhìn chung, biểu đồ này cho thấy người dùng hài lòng với trải nghiệm mua sắm điện thoại của họ. Tuy nhiên, họ vẫn có một số vấn đề cần được giải quyết, chẳng hạn như thời lượng pin và lỗi phần mềm. Có thể những *comment* này sẽ tác động mạnh đến việc đánh giá *positive, negative và neutral*

Nhằm đánh giá số lượng khách hàng đánh giá sản phẩm vào khoảng thời gian nào, và xu hướng đánh giá của khách hàng theo thời gian

A graph with blue lines and numbers

Description automatically generated

*(Hình 2.2.2.6: Biểu đồ thể hiện sự phân bố đánh giá theo thời gian)*

Có thể thấy được từ năm 2019 khách hàng có xu hướng tham gia đánh giá nhiều hơn, và số lượng sản phẩm có lượt đánh giá từ 3 đến 4.5 sao ngày càng nhiều.

## 2.3. Áp dụng mô hình machine learning và deep learning

Sau khi trải qua các phần trên như là thu thập, tiền xử lý dữ liệu , khám phá và phân tích tập dữ liệu.Phần này sẽ tập trung vào việc áp dụng các kỹ thuật machine learning (học máy) và deep learning (học sâu) để phân tích cảm xúc từ các bình luận của khách hàng và đi tìm khuôn mẫu dữ liệu.

### **2.3.1. Các kĩ thuật biểu diễn từ trong không gian vector**

Trước khi vào phần mô hình chúng ta cùng tìm hiểu các kỹ thuật biểu diễn từ trong không gian vector là bước quan trọng trong quá trình xử lý ngôn ngữ tự nhiên (NLP)

Kỹ thuật này chuyển đổi văn bản từ dạng chuỗi ký tự sang dạng số mà các mô hình machine learning và deep learning có thể xử lý. Dưới đây là một số kỹ thuật phổ biến mà Nhóm đã dùng trong dự án này

#### **2.3.1.1. One-hot-Encoding**

- One-Hot Encoding là một trong những kỹ thuật đơn giản nhất để biểu diễn từ. Mỗi từ trong từ điển (vocabulary) được biểu diễn bằng một vector có chiều dài bằng số lượng từ trong từ điển, trong đó chỉ có một phần tử có giá trị 1, các phần tử còn lại đều là 0.

- Ví dụ:

Từ điển: ["apple", "banana", "cherry"]

Biểu diễn như sau:

+"**apple**" -> [1, 0, 0]

+"**banana**" -> [0, 1, 0]

+"**cherry**" -> [0, 0, 1]

- **Ưu điểm:** Dễ hiểu và đơn giản để triển khai.

- **Nhược điểm:**

+ Không hiệu quả với từ điển lớn (vì vector sẽ rất dài và thưa)

+ Không lưu trữ được thông tin về ngữ nghĩa hay mối quan hệ giữa các từ.

#### **2.3.1.2. Bag-of-Words**

Bằng phương pháp túi từ (Bag-of-Words) (Z. S. Harris, 1954), ta sẽ đưa tất cả các từ trong văn bản thành các token sau đó xây dựng bộ từ điển và ghi nhận lại sự xuất hiện của các từ trong từ điển dựa trên các văn bản. Các ví dụ sau sẽ cho thấy cơ chế hoạt động của phương pháp này.

**Ví dụ :Hãy trích xuất đặc trưng của 3 câu sau bằng phương pháp túi từ.**

1. "Hôm qua tôi học lập trình"

2. "Hôm nay tôi cũng học lập trình"

3. "Ngày mai tôi không học lập trình"

**- Thông qua bước tokenization có thể biến đổi 3 câu đã cho thành các câu:**

1. "Hôm\_qua tôi học lập\_trình"

2. "Hôm\_nay tôi cũng học lập\_trình"

3. "Ngày\_mai tôi không học lập\_trình"

-Kho từ điển: [“hôm\_qua”, ”hôm\_nay”, ”ngày\_mai”, “tôi” , ”cũng” , ”không” , ”học”, ”lập trình”]

- Biểu diễn cho câu thứ nhất "Hôm\_qua tôi học lập\_trình" =>[1,0,0,1,0,0,1,1]

- Biểu diễn cho câu thứ hai "Hôm\_nay tôi cũng học lập\_trình” => [0,1,0,1,1,0,1,1]

- Biểu diễn cho câu thứ ba "Ngày\_mai tôi không học lập\_trình" => [0,0,1,1,0,1,1,1]

**- Ưu điểm:**

+ Dễ hiểu và dễ triển khai.

+ Hiệu quả đối với các văn bản ngắn và có từ điển nhỏ.

**- Nhược điểm:**

+ Không lưu trữ được thông tin về ngữ nghĩa hay thứ tự từ.

+ Vector rất dài và thưa khi từ điển lớn, dẫn đến yêu cầu bộ nhớ và tài nguyên tính toán cao.

#### **2.3.1.3. TF-IDF**

-TF-IDF là từ viết tắt của thuật ngữ tiếng Anh "term frequency – inverse document frequency". Như mô tả trong tài liệu “Khai thác những bộ dữ liệu lớn" - Mining of Massive Datasets) (Leskovec et al., 2014), TF-IDF là trọng số của một từ trong văn bản thu được qua thống kê thể hiện mức độ quan trọng của từ này trong một văn bản, mà bản thân văn bản đang xét nằm trong một tập hợp các văn bản.

﻿-TF-IDF gồm hai thành phần TF và IDF, được tính như sau:

+TF (Term Frequency): dùng để ước lượng tần suất xuất hiện của một từ trong một văn bản. Để chuẩn hóa tần suất này, chúng ta sẽ chia nó cho tần suất của từ xuất hiện nhiều nhất trong văn bản đó. Công thức tính TF được biểu diễn dưới đây:

Trong đó:

- f(t,d): là tần số xuất hiện của từ t trong văn bản d

- max {f(w,d):w **∈** d}: là tần suất của từ xuất hiện nhiều nhất trong văn bản d

+ IDF:﻿(Inverse Document Frequency): dùng để ước lượng mức độ quan trọng của một từ trong toàn bộ tập văn bản. Đối với từng văn bản, việc một từ xuất hiện nhiều lần sẽ có mức độ quan trọng cao hơn. Tuy nhiên, đối với toàn bộ tập các văn bản, một từ mà ở văn bản nào cũng xuất hiện sẽ trở nên phổ biến và không còn quan trọng nữa, ví dụ như:

• Từ nối: và, nhưng, tuy nhiên, vì thế, vì vậy,..

• Giới từ: ở, trong, trên,...

• Từ chỉ định: ấy, đó, nhỉ,...

Đó là lí do chúng ta cần phải giảm đi đó bằng cách sử dụng IDF được tính như dưới đây

Trong đó :

+ |D|: là tổng số văn bản trong tập văn bản D

+ |{d **∈** D : t **∈ d**}|: là số lượng văn bản trong tập văn bản D có chứa từ t

﻿

- Một trường hợp ngoại lệ có thể xảy ra ở đây đó chính là khi chúng ta sử dụng TF-IDF, nếu từ t không xuất hiện ở bất kỳ văn bản nào thì |{d **∈** D : t **∈ d**}| của IDF sẽ bằng 0. Do vậy, để tránh phép chia cho 0, chúng ta có thể thay đổi mẫu số của IDF bằng cách cộng một vào cho nó(smoothing).

- Sau cùng, kết hợp cả TF và IDF, ta có được công thức TF-IDF:

**TF-IDF(t,d,D)= TF(t,d) x IDF(t,D)**

- Mỗi tài liệu sẽ được biểu diễn bằng một vector TF-IDF, vector này sẽ có số chiều là K với K độ dài tất cả các từ có trong tất cả các văn bản của tập dữ liệu.

**Ưu điểm:**

- Tập trung vào sự quan trọng của từ.

- Loại bỏ từ không quan trọng.

- Linh hoạt, có thể điều chỉnh.

**Nhược điểm:**

- Không xem xét ngữ cảnh hoặc ý nghĩa của từ.

- Không xử lý được các từ gốc hoặc từ đồng nghĩa.

#### **2.3.1.4. Word2vec**

- Ở các kĩ thuật biểu diễn từ như Bag-of-words và TF-IDF đều không quan tâm đến mối quan hệ giữa các từ và có thể làm cho mô hình học không tốt đạt hiệu suất không cao

- word2vec là một trong những kỹ thuật được sử dụng phổ biến nhất trong lĩnh vực Xử lý ngôn ngữ tự nhiên. word2vec được tạo ra và công bố vào năm 2013 bởi một nhóm các

- Giả sử kho ngữ liệu có V từ phân biệt và ta muốn biến đổi từ thành vector có N chiều thì ma trận trọng số của mạng nơ-ron skip-gram sẽ có kích thước là V x N. Khi đó mỗi từ

(được biểu diễn bằng vector one-hot 1 x V) sẽ được biến đổi thành một vector 1 × N, vector này sẽ là một hàng trong ma trận trọng số.

- Tổng kết lại tầng đầu vào là biểu diễn one hot vector.Sau đó tầng ẩn là học theo đặc trưng để giảm số chiều.Tầng đầu ra là xem biểu diễn của tầng ẩn đó thì tương ứng với từ ngữ cảnh nào nếu theo mô hình skip-gram và ngược lại là mô hình cbow

#### **2.3.1.5. Embedding layer**

Embedding layer trong mạng nơ-ron có chức năng tương tự như Word2Vec trong việc học biểu diễn vector cho các từ. Cụ thể, embedding layer được sử dụng để ánh xạ từng từ trong từ vựng vào các vector có số chiều nhỏ hơn, mà trong đó mỗi chiều đại diện cho một đặc trưng của từ.

### **2.3.2. Xây dựng mô hình đi tìm khuôn mẫu dữ liệu**

#### **2.3.2.1. Giới thiệu về mô hình.**

Phân cụm là một phương pháp học máy không giám sát để nhóm các điểm dữ liệu có sự tương đồng cao với nhau. Trong báo cáo này, chúng ta sẽ thảo luận về quá trình phân cụm để gắn nhãn cho tập dữ liệu sentiment, bao gồm các bước tiền xử lý dữ liệu, sử dụng các thuật toán phân cụm khác nhau, và cuối cùng là gắn nhãn và đánh giá kết quả.

Mô hình sử dụng KMeans là một phương pháp phân cụm dữ liệu dựa trên các điểm giống nhau về mặt đặc

trưng. Quá trình này bao gồm các bước sau:

* Tạo Embedding Cho Dữ Liệu: Sử dụng mô hình embedding để chuyển đổi từng bài viết trong tập dữ liệu thành các vector số.
* Tìm Số Lượng Cụm Tối Ưu: Sử dụng phương pháp Elbow để xác định số lượng cụm tối ưu cho dữ liệu. Phương pháp này dựa trên việc quan sát sự giảm đáng kể trong tổng khoảng cách khi số lượng cụm tăng.
* Áp Dụng KMeans: Tiến hành phân cụm dữ liệu với số lượng cụm đã chọn.
* Giảm Chiều Dữ Liệu và Trực Quan Hóa Kết Quả Phân Cụm: Sử dụng PCA để giảm chiều dữ liệu xuống 2D và trực quan hóa kết quả phân cụm.

#### **2.3.2.2. Ứng dụng vào bài toán đi tìm khuôn mẫu dữ liệu**

Mô hình KMeans được áp dụng để phân loại các đánh giá sản phẩm thành các nhóm ý kiến khác nhau, giúp nhận biết và đánh giá tổng quan về cảm nhận của người dùng. Quá trình này bao gồm:

1. **Phân Cụm Dữ Liệu**: Sử dụng K-Means để phân chia các đánh giá thành các nhóm có ý nghĩa, như "Tích cực", "Tiêu cực" và "Trung lập", dựa trên nội dung và ngữ cảnh.
2. **Gắn Nhãn và Đặt Tên Cụm**: Mỗi đánh giá được gắn nhãn vào một nhóm, và các nhóm được đặt tên để mô tả ý kiến của người dùng.
3. **Đánh Giá Chất Lượng Phân Cụm**: Sử dụng chỉ số Silhouette để đánh giá mức độ tách biệt giữa các nhóm và chất lượng phân cụm.

Ví dụ: Trong ngành bán lẻ, một công ty muốn hiểu ý kiến của khách hàng về sản phẩm mới. Bằng cách sử dụng K-Means để phân tích các đánh giá từ khách hàng, họ có thể nhận biết các cảm nhận tích cực, tiêu cực và trung lập từ người dùng, từ đó cải thiện sản phẩm và dịch vụ của mình.

### **2.3.3. Mô hình K-Nearest Neighbors (KNN) cho bài toán phân tích cảm xúc**

#### **2.3.3.1. Giới thiệu về mô hình**

KNN có thể áp dụng cho 2 dạng bài toán kinh điển trong học máy là :

* Bài toán phân loại :Nếu là bài phân loại sau khi chọn ra số lượng hàng xóm gần gũi nhất thì sẽ dựa trên số lượng nhãn nào chiếm ưu thế thì sẽ gán nhãn đó cho điểm dữ liệu muốn nhận dạng
* Bài toán hồi quy(dự đoán): Nếu là bài toán hồi quy (dự đoán) sau khi chọn ra số lượng hàng xóm gần gũi nhất thì sẽ tính trung bình tất cả hàng xóm với nhau hoặc sẽ nhân thêm với hệ số tương đồng sau đó gán giá trị cho điểm dữ liệu muốn nhận dạng

#### **2.3.3.2. Ứng dụng vào bài toán phân tích cảm xúc**

Ý tưởng áp dụng trong bài toán:

* Ứng với từng câu ta sẽ thực hiện việc tách các từ trong câu.Sau đó sử dụng kĩ thuật TF-IDF để biểu diễn một câu (document).
* Sau đó sử dụng cosine để tính độ tương đồng giữa các câu với nhau.Khi nhận dạng một câu,xem nó là bình luận tích cực,tiêu cực hay là không thích cũng không ghét thì ta xem câu vừa bình luận có độ tương đồng với những câu nào.
* Sau đó chọn k (câu) có độ tương đồng với câu cần nhận dạng.
* Cuối cùng,do bài toán là thuộc bài phân loại trong kỹ thuật khai phá dữ liệu nên sẽ dựa trên số lượng hàng xóm(câu) có nhãn chiếm ưu thế và nhãn của những hàng xóm (câu) sẽ được gán cho câu cần nhận dạng

### **2.3.4. Mô hình LSTM cho bài toán phân tích cảm xúc**

#### **2.3.4.1. Giới thiệu về mô hình LSTM**

Mạng LSTM (Long Short-Term Memory) là một loại mạng nơ-ron hồi quy (RNN) tiên tiến được thiết kế để xử lý các vấn đề liên quan đến chuỗi dữ liệu, như xử lý ngôn ngữ tự nhiên, dịch máy, nhận dạng giọng nói, và nhiều ứng dụng khác. LSTM được thiết kế để giải quyết vấn đề của các mạng RNN truyền thống trong việc xử lý thông tin từ xa và lâu dài, có thể dẫn đến vấn đề biến mất hoặc phân rã đạo hàm.

#### **2.3.4.2. Ứng dụng vào bài toán phân tích cảm xúc**

Ý tưởng:

* Là ứng với từng bình luận ta sẽ thực hiện tách từ trong câu
* Sau đó xây kho ngữ liệu (từ điển).Sau đó đối với từng bình luận ta sẽ thực hiện ánh xạ lên từ điển để chuyển đổi từ thành giá trị.
* Do là mô hình các mạng thần kinh nhân tạo yêu cầu đầu vào số lượng từ trong một câu phải có kích thước bằng nhau cho nên chúng ta cần đệm thêm số 0 cho câu có số từ ít hơn.Vì thế chúng ta cần phải tìm kích thước của câu bình luận có số từ nhiều nhất và đệm thêm số 0 cho những câu khác để làm cho chúng bằng nhau.
* Sau đó sử dụng lớp embedding layer hoạt động của lớp này nhóm đã nêu ở phần 2.3.1.5 lớp này sẽ thực hiện việc nhúng từ để biểu diễn từ trong không gian vector với số chiều do mình lựa chọn để làm tăng hiệu suất mô hình.
* Đầu vào của mô hình sẽ là từng câu bình luận với mỗi câu bình luận sẽ có các từ.Và lớp ẩn là mỗi từ trong một câu bình luận sẽ được nhúng vào lớp embedding với mong muốn lớp LSTM có thể học nhớ được các từ quan trọng trong câu và khi nhận dạng LSTM sẽ đoán được câu đó thuộc vào lớp nào. Do bài toán này thuộc bài phân loại nên output layer sẽ có 3 nơron tương ứng với 3 nhãn positive, negative, neutral

### **2.3.5. Mô hình lai ghép giữa LSTM và CNN cho bài toán phân tích cảm xúc**

#### **2.3.5.1. Giới thiệu về mô hình**

Mạng tích chập (Convolutional Neural Network - CNN) là một loại kiến trúc mạng nơ-ron thường được sử dụng trong các ứng dụng liên quan đến xử lý ảnh và thị giác máy tính. CNN đã đạt được những thành tựu ấn tượng trong việc nhận diện ảnh, nhận dạng vật thể, và các nhiệm vụ khác liên quan đến thị giác máy tính.

CNN bao gồm :

* INPUT: Tầng đầu vào
* CONV: Tầng tích chập
* POOL: Tầng tổng hợp, thông thường là Max pooling hoặc có thể là Average pooling dùng để giảm chiều của ma trận đầu vào.



Như vậy ta có thể thấy một mạng nơ ron tích chập về cơ bản có 2 quá trình khác nhau:

* Quá trình tích chập (convolution): Thông qua các tích chập giữa ma trận đầu vào với bộ lọc để tạo thành các đơn vị trong một tầng mới.
* Quá trình tổng hợp (max pooling): Sau các tầng tích chập, số lượng tham số và khối lượng tính toán trong mạng nơ-ron tăng lên do chiều sâu (số lượng kênh) tăng theo cấp số nhân. Để giảm tải tính toán, quá trình tổng hợp (pooling) được áp dụng để giảm kích thước của các feature maps.

#### **2.3.5.2. Ứng dụng vào bài toán phân tích cảm xúc**

Ý tưởng:

* Là ứng với từng bình luận ta sẽ thực hiện tách từ trong câu.Sau đó sử dụng kỹ thuật biểu diễn từ word2vec để nhúng từ vào không gian vector nhằm mục đích tìm hiểu mối quan hệ giữa các từ cũng như là biểu diễn từ.
* Phương pháp word2vec sẽ vừa biểu diễn từ và vừa tạo một kho ngữ liệu (từ điển) của tất cả các từ trong bình luận.Sau đó chúng ta cần thực hiện việc Padding vì CNN cũng yêu cầu đầu vào số lượng từ trong câu phải bằng nhau.Chúng ta cũng đi tìm câu có số từ nhiều nhất và thực hiện đệm số 0 cho các câu có ít từ hơn
* Lý do nhóm kết hợp cả hai mô hình CNN và LSTM là vì CNN rất mạnh trong việc rút trích những tính chất ẩn tức là hàm ý của câu bình luận.

Ví dụ: Một khách hàng khi mua điện thoại anh ta nói. “Tôi muốn mua điện thoại có giá rẻ chút,nhưng vẫn chơi game ổn” thì tính chất ẩn ở đây là anh ta quan tâm đến cấu hình của máy bao gồm: pin,chip.

Tóm lại CNN sẽ đặt ở layer hidden đầu tiên để rút trích tính chất ẩn của câu bình luận còn LSTM sẽ làm nhiệm vụ nhớ được các tính chất ẩn quan trọng trong câu bình luận đó thì hiệu suất sẽ tốt hơn rất nhiều

## 2.4. Triển khai mô hình phân tích cảm xúc bằng Flask

Nhóm sẽ triển khai trang web để mọi người nhập bình luận đánh giá điện thoại và có kèm theo mô hình gợi ý nhập bình luận.Sau đó nhóm nhận dạng tất cả các bình luận mà mọi người nhập. Có 2 vai trò là user để nhập comment và admin để nhận dạng bình luận của user vừa nhập. Sau đó thống kê số lượng bình luận tích cực, tiêu cực, trung lập cho tất cả các điện thoại trong cơ sở dữ liệu.

# III. XÂY DỰNG CÁC MÔ HÌNH CHO VIỆC DỰ ĐOÁN CHỈ SỐ HẠNH PHÚC CỦA CÁC QUỐC GIA

## 3.1. Tiền xử lý dữ liệu

Ban đầu dữ liệu có được là từ 8 file .csv, với mỗi file có các thuộc tính khác nhau, nên ta cần phải trích lọc ra các thuộc tính cần thiết để thống nhất đưa vào cùng một tập dataframe, đồng thời cũng quy chuẩn về cùng một tên cho mỗi thuộc tính.

Trong tập dataframe này, các kỹ thuật tiền xử lý dữ liệu được áp dụng:

* Xử lý dữ liệu bị thiếu (NULL/NaN): nhằm tránh ảnh hưởng đến tính toàn vẹn và chính xác của phân tích và mô hình hóa dữ liệu, đảm bảo rằng dữ liệu được sử dụng trong quá trình phân tích là đầy đủ và đáng tin cậy.
* Xử lý dữ liệu trùng lặp: nhằm tránh sự chênh lệch trong phân tích và đưa ra kết quả không chính xác. Đảm bảo rằng mỗi mẫu dữ liệu chỉ được tính một lần, giúp giảm bớt khối lượng dữ liệu cần xử lý, từ đó tăng hiệu suất của quá trình tính toán và phân tích.
* Xử lý dữ liệu nhiễu: nhằm cải thiện chất lượng của dữ liệu bằng cách loại bỏ hoặc giảm thiểu ảnh hưởng của các điểm dữ liệu không mong muốn, cải thiện chất lượng của mô hình dự đoán.
* Thu giảm dữ liệu bằng PCA: nhằm giảm số chiều của dữ liệu mà vẫn giữ lại thông tin quan trọng nhất thông qua tinh chỉnh “n\_components” - số lượng chiều mà dữ liệu sẽ được giảm xuống. Lựa chọn “n\_components” sao cho tổng phương sai giải thích đạt tới 90% hoặc 95% của tổng phương sai ban đầu là một tiêu chí phổ biến và hữu ích trong việc giảm chiều dữ liệu bằng PCA. Điều này nhằm bảo toàn thông tin quan trọng nhất, giảm thiểu mất mát thông tin, đơn giản hóa dữ liệu, tránh overfitting và tiết kiệm chi phí tính toán. Trong bài toán này, ta thực hiện thu giảm về còn 2 chiều, được kết quả là 2 cột Happiness Score\_PCA và Economy (GDP per Capita)\_PCA.
* Biến đổi dữ liệu: nhằm thay đổi cấu trúc hoặc đặc điểm của dữ liệu ban đầu để làm cho nó phù hợp hơn cho một mục đích cụ thể.
  + Thực hiện gắn nhãn cho cột Country, dùng để vẽ biểu đồ Correlation để kiểm tra sự tương quan của biến này với các biến khác, và cũng để có thể sử dụng trong các mô hình học máy.
  + Chuẩn hóa các cột dữ liệu ngoại trừ “Country” và “year” theo phương pháp min-max normalization. Theo tập dữ liệu, chỉ có biến mục tiêu là Happniess Score là được tính theo thang điểm từ 0 đến 10, các biến độc lập còn lại là mức độ đóng góp của nó vào việc tính toán Happiness Score, nên ta sẽ thực hiện quy chuẩn các biến đó về cùng 1 thang điểm từ 0 đến 10 cho cùng một miền giá trị.



## 3.2. Khám phá và phân tích tập dữ liệu

### **3.2.1. Phân tích thống kê suy luận**

A graph with a line going up

Description automatically generated

*(Hình 3.2.1.1. Sự dao động của Happiness Score qua từng năm khảo sát)*

Bắt đầu từ 2019, số điểm có xu hướng giảm mạnh (cho đến năm 2022 lúc khảo sát), có thể là tại vì:

* Đại dịch COVID-19: Đại dịch COVID-19 bùng phát vào cuối năm 2019 và lan rộng trên toàn cầu trong năm 2020. Sự lan rộng của virus và các biện pháp phong tỏa và kiểm soát dịch bệnh đã gây ra một loạt các ảnh hưởng tiêu cực đối với nền kinh tế, sức khỏe và tâm lý của người dân. Việc phải sống trong tình hình bất ổn và lo sợ về sức khỏe và tài chính có thể đã làm giảm mức độ hạnh phúc của nhiều người.
* Tác động kinh tế: Đại dịch COVID-19 đã gây ra một số biến động lớn trong nền kinh tế toàn cầu, bao gồm sụp đổ của các ngành công nghiệp quan trọng như du lịch, giải trí và hàng không. Các biện pháp giãn cách xã hội và phong tỏa đã dẫn đến mất việc làm, giảm thu nhập và tăng tỷ lệ thất nghiệp, gây ra lo lắng và lo sợ về tương lai và ảnh hưởng đến tâm trạng và mức độ hạnh phúc của người dân.
* Tăng cường căng thẳng xã hội: Sự căng thẳng xã hội và bất ổn chính trị cũng có thể đã góp phần vào sự giảm mạnh của điểm hạnh phúc từ năm 2019. Các biểu hiện của sự căng thẳng và bất ổn có thể bao gồm các cuộc biểu tình, xung đột và bất đồng quan điểm, tất cả đều gây ra một môi trường không ổn định và lo sợ trong cộng đồng.

A pie chart with different colored circles

Description automatically generated

*(Hình 3.2.1.2. Tỷ lệ các yếu tố ảnh hưởng đến Happiness Score)*

Sự phân phối của các yếu tố ảnh hưởng đến Happiness Score được biểu diễn trong biểu đồ tròn cho thấy sự quan trọng của ba yếu tố chính là Family, Economy và Health đối với mức độ hạnh phúc của một quốc gia. Sự hài lòng trong mối quan hệ gia đình, thu nhập kinh tế và sức khỏe là những yếu tố cơ bản quyết định đến sự hạnh phúc và chất lượng cuộc sống của một quốc gia.

A map of the world

Description automatically generated

*(Hình 3.1.2.3. Bản đồ thể hiện trung bình Happiness Score sau các năm khảo sát)*

* Dựa trên biểu đồ, các quốc gia có chỉ số Hạnh Phúc cao nhất thế giới tập trung ở Bắc Âu, bao gồm: Phần Lan, Đan Mạch, Na Uy, Iceland và Thụy Sĩ
* Các quốc gia có điểm số hạnh phúc thấp tập trung ở Châu Phi, có thể là do điều kiện thời tiết tại Châu Lục này có ảnh hưởng lớn đến sự phát triển của các Quốc gia ở đây. Ngoài ra, Ấn Độ cũng thuộc diện nhóm này, có lẽ là do sự phân hóa, cách biệt giữa tầng lớp giàu nghèo trong xã hội Ấn.
* 2 quốc gia có diện tích lớn nhất là Nga và Trung Quốc lại có chi số Hạnh Phúc trung bình.
* Châu Úc và Châu Mỹ là 2 Châu Lục có số điểm Happiness Score tốt hơn so với các Châu còn lại. Ở Châu Mỹ, các nước có điểm cao đa số đều ở phía Bắc, xung quanh Hoa Kỳ.

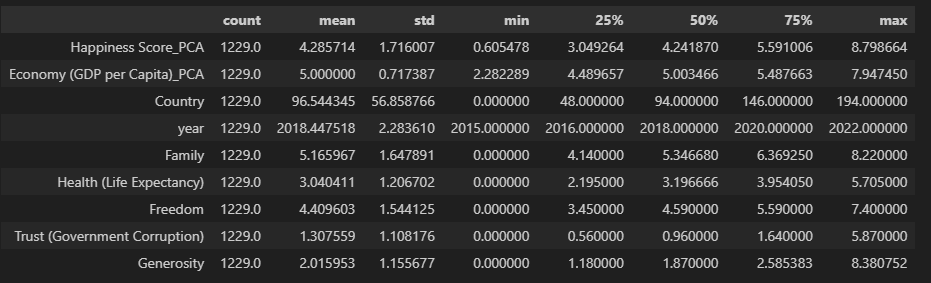
A screen shot of a chart

Description automatically generated

*(Hình 3.2.1.4. Sự tương quan giữa các thuộc tính trước khi thực hiện các kỹ thuật xử lý dữ liệu)*

Happiness Score có mối tương quan tích cực mạnh mẽ với Economy. Điều này có nghĩa là, các quốc gia có GDP cao hơn thường có Happiness Score cao hơn. Đây là điều dễ hiểu vì GDP cao hơn đồng nghĩa với thu nhập cao hơn, mức sống cao hơn và nhiều cơ hội hơn cho người dân. Happiness Score cũng có mối tương quan tích cực mạnh mẽ với Family và Health. Điều này cho thấy rằng các mối quan hệ gia đình tốt đẹp và sức khỏe tốt đóng vai trò quan trọng trong việc nâng cao hạnh phúc của con người. Health và Economy có mối tương quan tích cực cao, thực tế là nền kinh tế càng phát triển thì phúc lợi y tế, xã hội mà người dân có thể sử dụng càng tốt, nên chất lượng sức khỏe cũng được nâng theo. Generosity và Economy có mối tương quan âm nhẹ, có lẽ là ở quốc gia mà người dân có độ hào phóng cao thì sẽ có thu nhập thấp.

### **3.2.2. Phân tích thống kê mô tả tập dữ liệu sau khi đã tiền xử lý**



Trung bình Happiness Score\_PCA là khoảng 4.29, với độ biến động lớn (độ lệch chuẩn là 1.72). Điều này có thể chỉ ra rằng mức độ hạnh phúc của các quốc gia trong dữ liệu có sự phân tán đáng kể. Trung bình GDP trên đầu người là khoảng 5, với độ biến động tương đối thấp (độ lệch chuẩn là 0.72). Điều này có thể cho thấy sự ổn định kinh tế tương đối trong các quốc gia được nghiên cứu.

## 3.3. Xây dựng các mô hình machine learning và deep learning để dự đoán chỉ số điểm hạnh phúc ở các quốc gia

Sau khi đã tiền xử lý dữ liệu xong, xem lại sự tương quan của các thuộc tính, nhận thấy có sự tương quan cao giữa Family và Health, nhóm đã quyết định loại bỏ cột Family để tránh hiện tượng đa cộng tuyến trong lúc xây dựng các mô hình.

Sau khi xử lý xong, tiến hành chia tập dữ liệu thành 2 phần Train và Test với tỉ lệ 8:2.

Kết quả của model sẽ được đánh giá thông qua các chỉ số là:

* MSE (Mean Squared Error): đo lường trung bình của bình phương của sai số giữa dự đoán của mô hình và giá trị thực tế. MSE càng nhỏ, mô hình càng chính xác.
* RMSE (Root Mean Squared Error): căn bậc hai của MSE, càng nhỏ, mô hình càng chính xác.
* MAE (Mean Absolute Error): đo lường trung bình của giá trị tuyệt đối của sai số giữa dự đoán của mô hình và giá trị thực tế, càng nhỏ, mô hình càng chính xác.
* R-squared (R2): đo lường phần trăm phương sai của biến phụ thuộc mà mô hình giải thích được. R2 có giá trị từ 0 đến 1, và giá trị càng gần 1 cho thấy mô hình giải thích được một phần lớn sự biến động của dữ liệu.

### **3.3.1. Mô hình FNN (Feedforward Neural Network)**

#### **3.3.1.1. Giới thiệu về mô hình**

Mô hình Feedforward Neural Network (FNN), còn được biết đến với tên gọi Multilayer Perceptron (MLP), là một kiểu mạng nơ-ron nhân tạo phổ biến. Với cấu trúc gồm các lớp input, các lớp ẩn và lớp output, mô hình này truyền dữ liệu từ lớp input qua các lớp ẩn đến lớp output theo chiều truyền tiến. Các hàm kích hoạt phi tuyến tính được áp dụng tại mỗi nút để tạo ra đầu ra không tuyến tính và học được các đặc trưng phức tạp. Với khả năng học và dự đoán trong nhiều ứng dụng khác nhau như dự đoán, phân loại, nhận diện hình ảnh và xử lý ngôn ngữ tự nhiên, mô hình FNN đóng vai trò quan trọng trong lĩnh vực học máy và trí tuệ nhân tạo.

#### **3.3.1.2. Ứng dụng vào bài toán dự đoán chỉ số hạnh phúc**

Ý tưởng:

* Sử dụng mạng FNN (Feedforward Neural Network) với kiến trúc gồm các lớp Dense - kiểu fully connected và lớp Dropout - một kỹ thuật chính regularization để tránh overfitting.
* Đặt learning rate của tối ưu hóa Adam là 0.001, giúp giảm nguy cơ quá trình huấn luyện bị hỏng do các thay đổi quá lớn trong trọng số, đồng thời giúp mô hình học được các biểu diễn tốt hơn từ dữ liệu.
* Huấn luyện mô hình với 200 epochs, batch size là 32, và sử dụng 20% dữ liệu huấn luyện làm tập validation.

### **3.3.2. Mô hình Random Forest Regressor**

#### **3.3.2.1. Giới thiệu về mô hình**

Mô hình Random Forest Regressor là một thuật toán học máy thuộc nhóm ensemble learning, được sử dụng cho các nhiệm vụ hồi quy. Nó xây dựng một tập hợp các cây quyết định từ các tập con của dữ liệu huấn luyện, sử dụng phương pháp bootstrap sampling và chỉ xem xét một tập hợp con ngẫu nhiên của các đặc trưng tại mỗi nút chia. Dự đoán cuối cùng được lấy bằng cách trung bình các dự đoán từ tất cả các cây. Random Forest Regressor có độ chính xác cao, khả năng xử lý dữ liệu lớn và ít bị ảnh hưởng bởi nhiễu, nhưng đòi hỏi tài nguyên tính toán lớn. Nó được ứng dụng rộng rãi trong dự đoán giá nhà, phân tích tài chính, và khoa học môi trường.

#### **3.3.2.2. Ứng dụng vào bài toán dự đoán chỉ số hạnh phúc**

Ý tưởng:

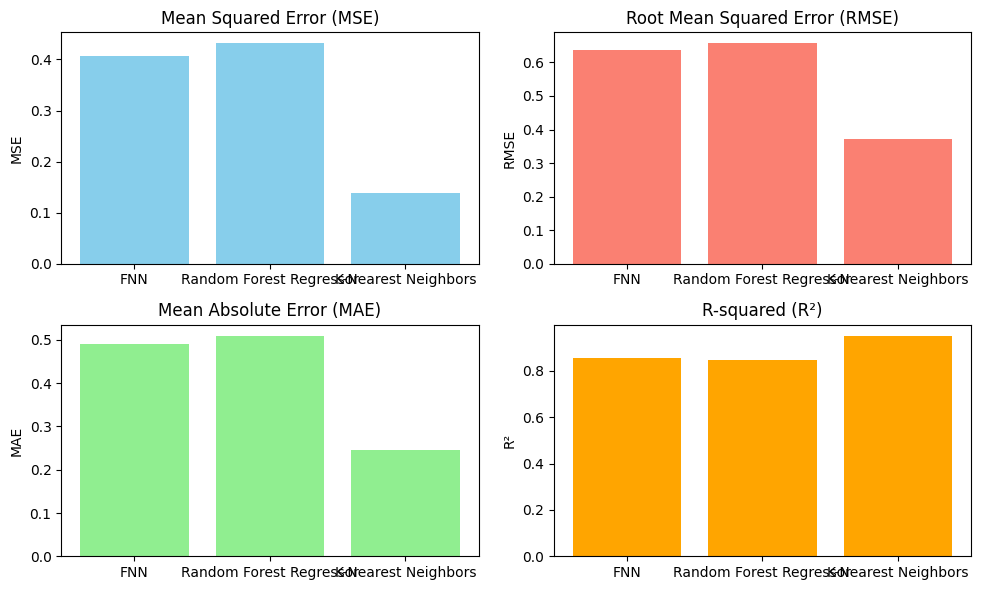
* Khởi tạo mô hình: Sử dụng thư viện scikit-learn để khởi tạo mô hình Random Forest Regressor với tham số random\_state=42 để đảm bảo tính tái lập của kết quả, đảm bảo rằng mỗi khi chạy sẽ nhận được cùng một kết quả.
* Huấn luyện mô hình: Áp dụng dữ liệu huấn luyện để huấn luyện mô hình.
* Sử dụng mô hình đã huấn luyện để dự đoán giá trị trên tập dữ liệu kiểm tra. Sau đó đánh giá hiệu suất mô hình

### **3.3.3. Mô hình K - Nearest Neighbors (KNN)**

Ý tưởng:

* Thiết lập các giá trị siêu tham số: Đặt các giá trị cần thử nghiệm cho các tham số n\_neighbors, weights, algorithm, và metric.
  + n\_neighbors: Số lượng hàng xóm xem xét để đưa ra dự đoán.
  + weights: Cách trọng số được gán cho các điểm lân cận (đều hoặc theo khoảng cách).
  + algorithm: Thuật toán được sử dụng để tìm các điểm lân cận.
  + metric: Hàm khoảng cách được sử dụng để tính toán khoảng cách giữa các điểm.
* Tạo mô hình KNN: Khởi tạo mô hình KNeighborsRegressor
* Tìm kiếm siêu tham số tốt nhất: Sử dụng GridSearchCV - một công cụ trong thư viện scikit-learn dùng để tìm kiếm siêu tham số tốt nhất cho mô hình học máy, với tập các giá trị siêu tham số đã thiết lập để tìm ra cấu hình tốt nhất dựa trên tiêu chí neg\_mean\_squared\_error - dạng phủ định của Mean Squared Error (MSE). Do GridSearchCV mặc định tối đa hóa tiêu chí đánh giá (tức là tìm giá trị lớn nhất), nên nếu muốn tối thiểu hóa MSE, cần sử dụng phủ định của nó. Điều này có nghĩa là giá trị càng lớn (trong ngữ cảnh phủ định, tức là giá trị MSE càng nhỏ), thì mô hình càng tốt.
* Huấn luyện mô hình: Áp dụng Grid Search trên tập huấn luyện
* Dự đoán trên tập kiểm tra: Sử dụng mô hình đã tinh chỉnh để dự đoán trên tập kiểm tra. Sau đó, đánh giá hiệu suất của mô hình bằng cách tính toán Mean Squared Error (MSE).
* Tạo mô hình KNN tốt nhất: Sử dụng các tham số tốt nhất tìm được từ Grid Search.
* Huấn luyện mô hình, dự đoán lại trên tập kiểm tra và tiến hành đánh giá hiệu suất.

## 3.4. Chọn ra mô hình tốt nhất cho việc dự đoán chỉ số hạnh phúc



(Hình 3.4.1. So sánh hiệu suất của các mô hình)

Từ các tiêu chí đánh giá là MSE, RMSE, MAE càng thấp càng tốt, và R-Squared càng gần giá trị 1 càng tốt thì nhóm quyết định chọn mô hình K-Neighbors (KNN) là mô hình tốt nhất cho việc dự đoán chỉ số hạnh phúc từ các thuộc tính liên quan. Nhóm cũng đã tiến hành lưu model ra ngoài thành file “best-model-regression.onnx”.

# IV. BẢNG PHÂN CÔNG CÔNG VIỆC

|  |  |
| --- | --- |
| **HỌ TÊN** | **NHIỆM VỤ** |
| **Đặng Nguyễn Quang Huy** | * Tiền xử lý dữ liệu văn bản(NLP) - mục 2.1 * Sử dụng power BI để phân tích thống kê mô tả tập dữ liệu - mục 2.2.1 * Mô hình K-Nearest Neighbors (KNN) cho bài toán phân tích cảm xúc - mục 2.3.3 * Mô hình LSTM cho bài toán phân tích cảm xúc - mục 2.3.4 * Mô hình LSTM cho bài toán tạo sinh văn bản (làm thêm) * Triển khai mô hình thực tế bằng Flask - mục 2.4 |
| **Đái Triệu Phi** | * Cào dữ liệu từ trang web sử dụng Scrapy <https://www.thegioididong.com/> * Xây dựng mô hình phân cụm để gắn nhãn dữ liệu cào về. Sử dụng Kmeans, DBScan, AgglomerativeClustering. Để phân cụm tập dữ liệu và gắn nhãn. mục – 2.3.5 * Các kĩ thuật biểu diễn từ trong không gian vector mục 2.3.1 |
| **Huỳnh Gia Hân** | * Sử dụng Python để phân tích thống kê suy luận - mục 2.2.2 * Mô hình lai ghép giữa LSTM và CNN cho bài toán phân tích cảm xúc - mục 2.3.5 * Triển khai mô hình thực tế bằng Flask - mục 2.4 |
| **Nguyễn Trọng Dũng** | * Tích hợp và tiền xử lý dữ liệu World Happiness Report - mục 3.1 * Khám phá và phân tích tập dữ liệu bằng Python - mục 3.2 * Xây dựng các mô hình machine learning và deep learning để dự đoán chỉ số điểm hạnh phúc ở các quốc gia - mục 3.3 |