**BỘ GIÁO DỤC VÀ ĐÀO TẠO**

**TRƯỜNG ĐẠI HỌC NGUYỄN TẤT THÀNH**

**KHOA CÔNG NGHỆ THÔNG TIN**

**NGUYEN TAT THANH**

**BÁO CÁO ĐỒ ÁN HỌC MÁY VÀ ỨNG DỤNG**

**KHOA HỌC DỮ LIỆU**

**ĐỀ TÀI : Email Spam Detection**

**Giảng viên hướng dẫn : Ths.Bùi Tiến Đức**

**Sinh viên thực hiện : Phạm Khánh Dũ**

**MSSV : 2100009554**

**Khoá : 2021**

**Ngành/ chuyên ngành : Khoa học dữ liệu**

Tp HCM, tháng ..09.. năm 2024

**BỘ GIÁO DỤC VÀ ĐÀO TẠO**

**A blue shield with white text and stars

Description automatically generatedTRƯỜNG ĐẠI HỌC NGUYỄN TẤT THÀNH**

**KHOA CÔNG NGHỆ THÔNG TIN**

**NGUYEN TAT THANH**

**BÁO CÁO ĐỒ ÁN HỌC MÁY VÀ ỨNG DỤNG**

**KHOA HỌC DỮ LIỆU**

**ĐỀ TÀI : Email Spam Detection**

**Giảng viên hướng dẫn : Ths.Bùi Tiến Đức**

**Sinh viên thực hiện : Phạm Khánh Dũ**

**MSSV : 2100009554**

**Khoá : 2021**

**Ngành/ chuyên ngành : Khoa học dữ liệu**

Tp HCM, tháng ..09.. năm 2024

# LỜI MỞ ĐẦU

Kính gửi thầy Bùi Tiến Đức,

Em xin gửi đến thầy lời chào trân trọng và sự biết ơn sâu sắc nhất. Việc được học môn "Học máy và ứng dụng" dưới sự hướng dẫn của thầy là một vinh dự lớn và là một cơ hội quý báu để em và các bạn có thể khai thác sâu hơn vào thế giới dữ liệu ngày nay. Những buổi học của thầy không chỉ cung cấp cho chúng em kiến thức chuyên sâu mà còn truyền cảm hứng và sự đam mê đối với lĩnh vực này.

Chúng em đang rất hào hứng áp dụng những gì đã học vào dự án của mình và rất mong nhận được sự chỉ dẫn, đánh giá và phản hồi từ thầy. Những góp ý của thầy sẽ giúp chúng em không ngừng cải thiện và phát triển trong ngành khoa học dữ liệu.

Một lần nữa, em xin gửi đến thầy lời cảm ơn chân thành vì sự tận tâm và hỗ trợ quý báu của thầy trong suốt thời gian qua.

Trân trọng!

# LỜI CẢM ƠN

Kính gửi thầy Bùi Tiến Đức,

Trong không khí học thuật đầy hào hứng và nhiệt huyết, chúng em xin gửi đến thầy lời cảm ơn chân thành và sâu sắc nhất. Dưới sự hướng dẫn tận tâm của thầy trong môn học "Học máy và ứng dụng ", chúng em đã có cơ hội khám phá và áp dụng những khái niệm tiên tiến, đồng thời nâng cao kỹ năng phân tích và giải quyết vấn đề.

Nhờ vào sự hỗ trợ và sự tận tình của thầy, chúng em không chỉ học hỏi được những kỹ thuật khai thác dữ liệu quan trọng mà còn phát triển tư duy phản biện và khả năng làm việc hiệu quả. Thầy đã tạo ra một môi trường học tập đầy cảm hứng, khuyến khích chúng em không ngừng học hỏi và sáng tạo.

Chúng em xin bày tỏ lòng biết ơn sâu sắc và cam kết sẽ tiếp tục nỗ lực để áp dụng những kiến thức và kỹ năng đã học vào thực tế. Những bài học quý báu mà thầy truyền đạt sẽ luôn là nguồn động viên lớn lao cho chúng em trên con đường học tập và phát triển.

Một lần nữa, xin chân thành cảm ơn thầy vì sự tận tâm và những đóng góp không ngừng của thầy trong suốt thời gian qua.

Trân trọng!

|  |  |
| --- | --- |
| TRƯỜNG ĐẠI HỌC NGUYỄN TẤT THÀNH  **TRUNG TÂM KHẢO THÍ** | **KỲ THI KẾT THÚC HỌC PHẦN**  **HỌC KỲ 9 – NĂM HỌC 2023 - 2024** |

**PHIẾU CHẤM ĐIỂM ĐỒ ÁN CƠ SỞ**

Môn thi: Học máy và ứng dụng Lớp học phần: 21DTH2D

Nhóm sinh viên thực hiện :

1.Phạm Khánh Dũ – 2100009554 Tham gia đóng góp: 100%

Đề tài tiểu luận/báo cáo của sinh viên : Email Spam Detection

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

Ngày thi: 15/09/2024 Phòng thi: L.509

Đề tài tiểu luận: Email Spam Detection

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** |  |  |  |

**Giảng viên chấm thi**

(ký, ghi rõ họ tên)

Ths. Bùi Tiến Đức

# MỤC LỤC

[LỜI MỞ ĐẦU i](#_Toc177326014)

[LỜI CẢM ƠN ii](#_Toc177326015)

[MỤC LỤC iv](#_Toc177326016)

[DANH MỤC HÌNH ẢNH vi](#_Toc177326017)

[KÝ HIỆU CÁC TỪ VÀ CỤM TỪ VIẾT TẮC vii](#_Toc177326018)

[**CHƯƠNG I: CƠ SỞ LÝ THUYẾT** 1](#_Toc177326019)

[**1.1** **KHOA HỌC DỮ LIỆU:** 1](#_Toc177326020)

[1.1.1 Giới thiệu và vai trò: 1](#_Toc177326021)

[**1.2 HỌC MÁY :** 2](#_Toc177326022)

[***1.2.1 Giới thiệu và vai trò :*** 2](#_Toc177326023)

[***1.2.2 Mô tả các phương pháp :*** 2](#_Toc177326024)

[**1.3 THUẬT TOÁN DỰ ĐOÁN VÀ HỌC SUPERVISED :** 3](#_Toc177326025)

[***1.3.1 Trình bày thuật toán :*** 3](#_Toc177326026)

[***1.3.2 Giải thích nguyên tác hoạt động :*** 3](#_Toc177326027)

[**1.4 ĐÁNH GIÁ MÔ HÌNH VÀ QUYẾT ĐỊNH CHỌN MÔ HÌNH :** 3](#_Toc177326028)

[***1.4.1 Các phương pháp đánh giá hiệu suất :*** 3](#_Toc177326029)

[***1.4.2 Cách quyết định chọn mô hình :*** 4](#_Toc177326030)

[**1.5 QUẢN LÝ DỮ LIỆU VÀ QUYẾT ĐỊNH BIẾN QUAN TRỌNG** 6](#_Toc177326031)

[***1.5.1 Đề cập đến quản lý dữ liệu, chọn lọc biến quan trọng và ảnh hưởng của chúng đối với mô hình :*** 6](#_Toc177326032)

[***1.5.2 Nêu rõ cách biến quan trọng có thể giúp hiểu rõ hơn về yếu tố nào ảnh hưởng nhiều đến bệnh sơ gan*** 6](#_Toc177326033)

[***1.5.3 Xem Xét Trọng Số Của Các Biến:*** 8](#_Toc177326034)

[***1.5.4 Lợi Ích và Cảnh Báo:*** 8](#_Toc177326035)

[**1.6** **AN TOÀN VÀ QUY ĐỊNH :** 9](#_Toc177326036)

[***1.6.1 Bàn luận về quy định về quy định và an toàn dữ liệu*** 9](#_Toc177326037)

[***1.6.2*** ***Các biện pháp đảm bảo tính riêng tư và tuân thủ quy định :*** 10](#_Toc177326038)

[**1.7 Công nghệ và ngôn ngữ lập trình** 12](#_Toc177326039)

[1.7.1 Trình bày các công nghệ và ngôn ngữ lập trình thích hợp cho việc triển khai mô hình dự đoán bệnh sơ gan 12](#_Toc177326040)

[**1.8 TIỀN XỬ LÝ DỮ LIỆU :** 14](#_Toc177326041)

[***1.8.1*** ***Xử lý dữ liệu thiếu :*** 14](#_Toc177326042)

[***1.8.2*** ***Loại bỏ dữ liệu nhiễu :*** 14](#_Toc177326043)

[***1.8.3*** ***Chuẩn hóa dữ liệu :*** 15](#_Toc177326044)

[***1.8.4*** ***Xử lý biến độc lập :*** 15](#_Toc177326045)

[***1.8.5*** ***Tạo các đặc trưng mới :*** 15](#_Toc177326046)

[***1.8.6*** ***Xử lý dữ liệu dư thừa :*** 15](#_Toc177326047)

[***1.8.7*** ***Xử lý dữ liệu đặc trưng không phải số :*** 15](#_Toc177326048)

[***1.8.8*** ***Xử lý dữ liệu phân phối không đồng đều :*** 16](#_Toc177326049)

[***1.8.9*** ***Tách tập huấn luyện và tập kiểm tra :*** 16](#_Toc177326050)

[***1.8.10 Kiểm tra và đánh giá dữ liệu :*** 16](#_Toc177326051)

[**1.8.11** **Giữ nguyên dữ liệu gốc :** 16](#_Toc177326052)

[**CHƯƠNG II : THỰC NGHIỆM VÀ KẾT QUẢ** 17](#_Toc177326053)

[2.1 Giới thiệu đề tài : 17](#_Toc177326054)

[2.2 Khai báo các thư viện để triển khai đồ án thực hiện 17](#_Toc177326055)

[**2.3 Data Cleaning và data** 17](#_Toc177326056)

[**2.3.1 Đọc và hiểu dữ liệu** 17](#_Toc177326057)

[2.3.2 Data Cleaning 19](#_Toc177326058)

[**2.4** **EDA** (**Exploratory Data Analysis) (Phân tích Dữ liệu Khám Phá)** 22](#_Toc177326059)

[2.4.1 Percentage of Ham and Spam 23](#_Toc177326060)

[2.4.2 Text Length and Structure Analysis 25](#_Toc177326061)

[2.5 Data Preprocessing (tiền xử lý dữ liệu) 32](#_Toc177326062)

[2.6 Model Building 36](#_Toc177326063)

[**CHƯƠNG III : KẾT LUẬN** 40](#_Toc177326064)

[**3.1 Kết quả đạt được:** 40](#_Toc177326065)

[**3.2 Hiệu suất mô hình :** 40](#_Toc177326066)

[**3.3 Gợi ý sản phẩm cho khách hàng:** 41](#_Toc177326067)

[**3.4 Hạn chế :** 41](#_Toc177326068)

[**TÀI LIỆU THAM KHẢO** 42](#_Toc177326069)

# DANH MỤC HÌNH ẢNH

[Hình 2. 1 Các thư viện 17](#_Toc177324980)

[Hình 2. 2 :Data Train 18](#_Toc177324981)

[Hình 2. 3 :Code mô tả đọc dữ liệu từ file CSV 18](#_Toc177324982)

[Hình 2. 4 : Data info 19](#_Toc177324983)

[Hình 2. 5 : Xóa các cột dữ liệu không có giá trị 20](#_Toc177324984)

[Hình 2. 6 : Đổi tên cột 21](#_Toc177324985)

[Hình 2. 7 : Chuyển đổi các giá trị phân loại (như "spam" và "ham") thành các số nguyên (0, 1, ) 21](#_Toc177324986)

[Hình 2. 8 : Kiểm tra giá trị bị thiếu 22](#_Toc177324987)

[Hình 2. 9 : Kiểm tra số lượng các hàng bị trùng lặp trong DataFrame 22](#_Toc177324988)

[Hình 2. 10 : Xóa các hàng trùng lặp khỏi DataFrame và giữ lại hàng đầu tiên trong mỗi nhóm trùng lặp. 22](#_Toc177324989)

[Hình 2. 11 Kích thước của dữ liệu (gồm 5169 hàng , 2 cột) 22](#_Toc177324990)

[Hình 2. 12 Tỷ lệ phần trăm của các giá trị Ham và spam 23](#_Toc177324991)

[Hình 2. 13 Trực quan hóa tỉ lệ phần trăm của spam và ham 25](#_Toc177324992)

[Hình 2. 14 Tính toán số lượng ký tự, số lượng từ, và số lượng câu trong mỗi tin nhắn văn bản và sau đó hiển thị các thống kê mô tả 25](#_Toc177324993)

[Hình 2. 15 Thống kê tóm tắt cho các tin nhắn không spam số lượng ký tự, số lượng từ, và số lượng câu trong mỗi tin nhắn 27](#_Toc177324994)

[Hình 2. 16 Thống kê tóm tắt cho các tin nhắn spam số lượng ký tự, số lượng từ, và số lượng câu trong mỗi tin nhắn 28](#_Toc177324995)

[Hình 2. 17 Biểu đồ trực quán hoa số lượng ký tự spam và ham 29](#_Toc177324996)

[Hình 2. 18 Biểu đồ trực quán hoa số lượng từ spam và ham 30](#_Toc177324997)

[Hình 2. 19 Biểu đồ trực quán hóa Pairplot về số lượng từ ,ký tự ,câu spam và ham 31](#_Toc177324998)

[Hình 2. 20 Xây dựng transform\_text để biến đổi văn bản, tách từ, loại bỏ ký tự đặc biệt, và thực hiện stemming 32](#_Toc177324999)

[Hình 2. 21 Áp dụng transform\_text vào dữ liệu 33](#_Toc177325000)

[Hình 2. 22 Biểu đồ từ đám mây từ (word cloud) cho văn bản trong cột transformed\_text của DataFrame df đối với văn bản spam 33](#_Toc177325001)

[Hình 2. 23 Biểu đồ từ đám mây từ (word cloud) cho văn bản trong cột transformed\_text của DataFrame df đối với văn bản không spam 34](#_Toc177325002)

[Hình 2. 24 Tìm kiếm top 30 từ loại spam 35](#_Toc177325003)

[Hình 2. 25 Tìm kiếm top 30 từ loại không spam 36](#_Toc177325004)

[Hình 2. 26 Nhập Thư Viện và Khởi Tạo Vectorizer 36](#_Toc177325005)

[Hình 2. 27 Chuyển Đổi Dữ Liệu Văn Bản 36](#_Toc177325006)

[Hình 2. 28 Chia Dữ Liệu Thành Tập Huấn Luyện và Tập Kiểm Tra 37](#_Toc177325007)

[Hình 2. 29 Khởi Tạo Các Mô Hình Học Máy 37](#_Toc177325008)

[Hình 2. 30 Tạo Từ Điển Các Mô Hình 38](#_Toc177325009)

[Hình 2. 31 Định Nghĩa Hàm Huấn Luyện Mô Hình và Đánh Giá 38](#_Toc177325010)

[Hình 2. 32 Huấn Luyện Các Mô Hình và Hiển Thị Kết Quả 39](#_Toc177325011)

# KÝ HIỆU CÁC TỪ VÀ CỤM TỪ VIẾT TẮC

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| |  |  | | --- | --- | | Chữ viết tắt | Ý nghĩa | | **SVM** | **Support Vector Machines** | | **MSE** | **Mean Squared Error** | | **R²** | **R-squared** | | **ROC** | **Receiver Operating Characteristic** | | **ONNX** | **Open Neural Network Exchange** | | **AUC** | **Area Under the Curve** | | **MAE** | **Mean Absolute Error** | |

**CHƯƠNG I: CƠ SỞ LÝ THUYẾT**

* 1. **KHOA HỌC DỮ LIỆU:**

### 1.1.1 Giới thiệu và vai trò:

Khoa học dữ liệu là một lĩnh vực liên ngành kết hợp các phương pháp từ thống kê, học máy (machine learning), phân tích dữ liệu và các kỹ thuật khoa học máy tính để khai thác, phân tích và hiểu các dữ liệu phức tạp. Dưới đây là một số điểm chính về giới thiệu và vai trò của khoa học dữ liệu:

* **Giới thiệu**: Khoa học dữ liệu là quá trình thu thập, xử lý, phân tích và trực quan hóa dữ liệu để rút ra những thông tin giá trị, hỗ trợ việc ra quyết định dựa trên dữ liệu.Khoa học dữ liệu kết hợp toán học, thống kê, lập trình và các lĩnh vực cụ thể như kinh tế, sinh học, và xã hội học.
* **Công cụ và Kỹ thuật**: Các nhà khoa học dữ liệu sử dụng công cụ và kỹ thuật như phân tích thống kê, học máy, khai thác dữ liệu, và các nền tảng dữ liệu lớn như Python, R, Hadoop, và Spark.
* **Vai trò của khoa học dữ liệu**:
* Quyết định dựa trên dữ liệu: Giúp các tổ chức đưa ra quyết định thông minh hơn qua phân tích dữ liệu.
* Dự đoán và Phát hiện Xu hướng: Dự đoán xu hướng tương lai và phát hiện mẫu hình quan trọng trong các lĩnh vực như tài chính, chăm sóc sức khỏe, và sản xuất.
* Cải thiện hiệu quả hoạt động: Tối ưu hóa quy trình, giảm chi phí và tăng hiệu suất trong doanh nghiệp.
* Phát triển sản phẩm và dịch vụ: Hiểu nhu cầu khách hàng để cải tiến sản phẩm và dịch vụ.
* Khoa học và Nghiên cứu: Hỗ trợ phân tích dữ liệu lớn và rút ra kết luận khoa học mới.
* Xử lý dữ liệu lớn: Phân tích dữ liệu khổng lồ để khai thác thông tin từ nhiều nguồn khác nhau.
* Chăm sóc sức khỏe: Phân tích hồ sơ bệnh án, nghiên cứu dịch tễ học, và phát triển phương pháp điều trị cá nhân hóa.

**1.2 HỌC MÁY :**

***1.2.1 Giới thiệu và vai trò :***

* Học máy là một lĩnh vực của trí tuệ nhân tạo (AI) mà nói chung tập trung vào việc phát triển các thuật toán và mô hình máy tính có khả năng học từ dữ liệu. Mục tiêu chính của học máy là xây dựng các mô hình có khả năng tự học và áp dụng kiến thức này để thực hiện các nhiệm vụ cụ thể mà không cần phải được lập trình một cách rõ ràng.
* Vai trò của học máy trong khai thác dữ liệu rất quan trọng và đa dạng. Dưới đây là một số điểm chính :

+ Phân loại và dự đoán

+ Phân tích cụm và gom cụm

+ Dự đoán chuỗi thời gian

+ Phân loại và nhận diện hình ảnh

+ Xử lý ngôn ngữ tự nhiên

+ Tối ưu hóa quy trình kinh doanh

***1.2.2 Mô tả các phương pháp :***

* Phân loại là một lĩnh vực quan trọng của học máy, và nó có thể được áp dụng trong nhiều bối cảnh khác nhau, bao gồm việc dự đoán sơ gan. Dưới đây là mô tả về một số phương pháp phân loại phổ biến, đặc biệt là những phương pháp thích hợp cho dự đoán sơ gan:

+ Máy vector hỗ trợ ( Support Vector Machines – SVM )

+ Rừng ngẫu nhiên ( Random Forest )

+ Mạng nơ-ron ( Neural Networks )

+ Học tăng cường ( Boosting )

+ Mô hình Hồi quy Logit ( Logistic Regression )

+ Máy Bayes ( Naïve Bayes )

+ Máy học Gradient ( Gradient Boosting Machines )

Nhưng ở đồ án lần này chúng ta sẽ sử dụng phương pháp **Rừng ngẫu nhiên ( Random Forest** ) để triển khai.

**1.3 THUẬT TOÁN DỰ ĐOÁN VÀ HỌC SUPERVISED :**

***1.3.1 Trình bày thuật toán :***

* Dưới đây là mô tả về thuật toán phổ biến trong học máy được sử dụng để dự đoán, đó là thuật toán Random Forests.

***1.3.2 Giải thích nguyên tác hoạt động :***

* **Rừng Ngẫu Nhiên (Random Forests):**

Nguyên tắc hoạt động: Random Forests kết hợp nhiều cây quyết định để cải thiện độ chính xác và giảm nguy cơ overfitting. Mỗi cây được huấn luyện trên một tập dữ liệu con khác nhau và kết quả được kết hợp.

Áp dụng trong dự đoán bệnh sơ gan: Random Forests có thể sử dụng thông tin từ nhiều nguồn khác nhau để dự đoán bệnh sơ gan, giảm thiểu ảnh hưởng của nhiễu và tăng khả năng tổng hợp dữ liệu.

**1.4 ĐÁNH GIÁ MÔ HÌNH VÀ QUYẾT ĐỊNH CHỌN MÔ HÌNH :**

***1.4.1 Các phương pháp đánh giá hiệu suất :***

Các phương pháp đánh giá hiệu suất của mô hình dự đoán là quan trọng để đảm bảo rằng mô hình hoạt động đúng cách và có khả năng tổng quát hóa. Dưới đây là một số độ đo phổ biến tùy thuộc vào bối cảnh dự án:

**Mean Squared Error (MSE):**

Mô tả: MSE là giá trị trung bình của bình phương của các sai số giữa giá trị dự đoán và giá trị thực tế. Nó là một độ đo đánh giá chất lượng của mô hình bằng cách đo lường độ lớn của sai số.

* Ưu điểm: Dễ hiểu và tính toán.
* Nhược điểm: Nhược điểm của MSE là nó nhạy cảm với các giá trị ngoại lai (outliers).

**R-squared (R²):**

Mô tả: R-squared đo lường phần trăm phương sai của biến phụ thuộc mà mô hình có thể giải thích. Giá trị R-squared càng gần 1, mô hình càng tốt.

+ Ưu điểm: Dễ hiểu và cung cấp cái nhìn tổng quan về hiệu suất của mô hình.

+ Nhược điểm: Có thể bị ảnh hưởng bởi số lượng biến độc lập và có thể không phản ánh đúng hiệu suất của mô hình đối với một số tình huống.

**Mean Absolute Error (MAE):**

Mô tả: MAE là giá trị trung bình của giá trị tuyệt đối của các sai số giữa giá trị dự đoán và giá trị thực tế. Nó giảm ảnh hưởng của outliers so với MSE.

+ Ưu điểm: Nhạy với các giá trị ngoại lai, giúp đánh giá độ chính xác của mô hình trong trường hợp này.

+ Nhược điểm: Có thể không đánh giá tốt sự lớn của sai số.

**Precision, Recall, F1-score (Đối với bài toán phân loại):**

Mô tả: Precision đo lường tỉ lệ các trường hợp dự đoán đúng so với tổng số trường hợp dự đoán là positive. Recall đo lường tỉ lệ các trường hợp dự đoán đúng so với tổng số trường hợp thực sự là positive. F1-score là sự kết hợp giữa Precision và Recall.

+ Ưu điểm: Phù hợp cho bài toán phân loại, đặc biệt là khi có mất cân bằng giữa các lớp.

+ Nhược điểm: Cần cân nhắc giữa Precision và Recall tùy thuộc vào yêu cầu cụ thể của ứng dụng.

**Receiver Operating Characteristic (ROC) và Area Under the Curve (AUC):**

Mô tả: ROC curve là biểu đồ thể hiện mối quan hệ giữa tỉ lệ True Positive và False Positive của mô hình ở nhiều ngưỡng quyết định khác nhau. AUC là diện tích dưới ROC curve, một độ đo của khả năng phân biệt của mô hình.

+ Ưu điểm: Phù hợp cho bài toán phân loại, đo lường khả năng phân biệt của mô hình.

+ Nhược điểm: Có thể không phản ánh đúng hiệu suất của mô hình nếu có mất cân bằng giữa các lớp.

Khi đánh giá hiệu suất mô hình, quan trọng là chọn các độ đo phù hợp với bối cảnh cụ thể của dự án và mục tiêu đánh giá.

***1.4.2 Cách quyết định chọn mô hình :***

* Quyết định chọn mô hình tốt nhất dựa trên kết quả đánh giá đòi hỏi sự cân nhắc kỹ lưỡng và phân tích cẩn thận của các độ đo hiệu suất. Dưới đây là một hướng dẫn tổng quan về cách thức quyết định chọn mô hình tốt nhất:

+ Hiểu Rõ Mục Tiêu Dự Án:

Xác định rõ mục tiêu của dự án và yêu cầu đánh giá. Đôi khi, các mục tiêu khác nhau có thể yêu cầu sử dụng các độ đo khác nhau. Ví dụ, trong một bài toán phân loại, Precision và Recall có thể quan trọng tùy thuộc vào tình huống cụ thể.

+ Sử Dụng Nhiều Độ Đo:

Không dựa hoàn toàn vào một độ đo duy nhất. Sử dụng một loạt các độ đo để có cái nhìn tổng quan về hiệu suất của mô hình. Điều này giúp tránh tình trạng mô hình tốt theo một độ đo nhưng kém theo các độ đo khác.

+ Phân Tích Đồng Thời và So Sánh Mô Hình:

So sánh các mô hình dựa trên kết quả của từng độ đo. Nếu có nhiều mô hình, xem xét kết quả của chúng đồng thời để xác định mô hình nào là tốt nhất. Đôi khi, sự chênh lệch nhỏ giữa các mô hình có thể yêu cầu phải xem xét sâu hơn.

+ Xem xét Độ Stabilitas và Tổng Quát Hóa:

Kiểm tra sự ổn định và khả năng tổng quát hóa của mô hình. Một mô hình có thể hoạt động tốt trên tập dữ liệu huấn luyện nhưng không tốt trên tập dữ liệu kiểm thử mới. Kết quả đánh giá trên tập kiểm thử là quan trọng để đảm bảo tính tổng quát hóa.

+ Cân Nhắc Đặc Điểm Cụ Thể của Dự Án:

Đôi khi, các yếu tố cụ thể của dự án có thể ảnh hưởng đến quyết định chọn mô hình. Ví dụ, nếu sự chính xác cao là ưu tiên số một, thì sự chọn lựa có thể tập trung vào mô hình có R-squared hoặc AUC cao.

+ Kiểm Tra Tác Động Của Hyperparameters:

Nếu sử dụng mô hình có hyperparameters (tham số mô hình cần được đặt trước khi huấn luyện), kiểm tra tác động của các giá trị hyperparameters đối với hiệu suất của mô hình. Điều này có thể giúp xác định giá trị hyperparameters tối ưu.

+ Thực Hiện Cross-Validation:

Sử dụng kỹ thuật cross-validation để đảm bảo rằng đánh giá hiệu suất không phụ thuộc vào một tập dữ liệu cụ thể. Cross-validation giúp đánh giá khả năng tổng quát hóa của mô hình.

+ Chọn Mô Hình Dựa trên Yêu Cầu Cụ Thể:

Cuối cùng, chọn mô hình dựa trên yêu cầu cụ thể của dự án. Đôi khi, một mô hình có thể được chọn do khả năng giải thích tốt hơn, trong khi mô hình khác có thể được chọn do hiệu suất dự đoán cao hơn.

Quyết định chọn mô hình tốt nhất là một quá trình phức tạp và yêu cầu sự kết hợp của kiến thức về dữ liệu, kinh nghiệm, và sự hiểu biết về bối cảnh dự án cụ thể.

**1.5 QUẢN LÝ DỮ LIỆU VÀ QUYẾT ĐỊNH BIẾN QUAN TRỌNG**

***1.5.1 Đề cập đến quản lý dữ liệu, chọn lọc biến quan trọng và ảnh hưởng của chúng đối với mô hình :***

* Tăng Tốc Quá Trình Huấn Luyện:

Giảm số lượng biến có thể làm tăng tốc quá trình huấn luyện, đặc biệt là đối với các mô hình phức tạp.

* Giảm Nguy Cơ Overfitting:

Bằng cách loại bỏ các biến không quan trọng, mô hình có thể giảm nguy cơ overfitting và tăng khả năng tổng quát hóa.

* Dễ Hiểu và Giải Thích:

Mô hình với ít biến hơn là dễ hiểu hơn và dễ giải thích hơn, điều này quan trọng khi cần truyền đạt thông tin cho người quyết định không chuyên ngành.

* Giảm Chi Phí Thu Thập và Lưu Trữ Dữ Liệu:

Giảm số lượng biến có thể giảm chi phí liên quan đến việc thu thập và lưu trữ dữ liệu.

* Quản lý dữ liệu và chọn lọc biến quan trọng đều đóng vai trò quan trọng trong việc xây dựng mô hình có hiệu suất cao và dễ giải thích. Tùy thuộc vào bối cảnh cụ thể của dự án, quyết định chọn lọc biến có thể thay đổi để đáp ứng mục tiêu và yêu cầu cụ thể.

***1.5.2 Nêu rõ cách biến quan trọng có thể giúp hiểu rõ hơn về yếu tố nào ảnh hưởng nhiều đến bệnh sơ gan***

* Quá trình xác định biến quan trọng có thể cung cấp thông tin quan trọng về yếu tố nào ảnh hưởng nhiều đến bệnh sơ gan trong mô hình dự đoán. Dưới đây là một số cách cụ thể để hiểu rõ hơn về yếu tố quan trọng:

***1.5.2.1 Sử Dụng Mô Hình Như Random Forests:***

a. Huấn Luyện Mô Hình Random Forests:

Sử dụng mô hình Random Forests để huấn luyện mô hình dự đoán sơ gan

b. Lấy Độ Quan Trọng Của Các Biến:

Sử dụng thuộc tính feature\_importances\_ trong mô hình Random Forests để lấy thông tin về độ quan trọng của từng biến.

c. Xác Định Các Biến Quan Trọng:

Trong cây ngẫu nhiên, độ quan trọng của tính năng thường được xác định bằng cách đo lường mức độ giảm độ chính xác của mô hình khi tính năng đó bị loại bỏ hoặc permuted.

d. Lượng Giảm Impurity (Gini Impurity hoặc Entropy): Đối với mỗi tính năng, tính toán tổng lượng giảm impurity (Gini impurity hoặc entropy) khi tính năng đó được sử dụng để chia các node trong cây quyết định. Tính năng có sự giảm impurity lớn nhất được coi là quan trọng nhất.

***1.5.2.2 Sử Dụng Mô Hình Học Máy Với L1 Regularization:***

a. Sử Dụng Mô Hình Với L1 Regularization:

Sử dụng mô hình học máy có hỗ trợ L1 regularization như Logistic Regression với penalty term L1.

b. Lấy Các Biến Quan Trọng:

Xem xét các trọng số của các biến sau khi huấn luyện mô hình với L1 regularization. Các biến với trọng số khác 0 được coi là quan trọng.

***1.5.2.3 Phân Tích Tương Quan và Đặc Trưng:***

a. Phân Tích Tương Quan:

Sử dụng phân tích tương quan để xem xét mức độ tương quan giữa các biến và biến mục tiêu

b. Xem Xét Đặc Trưng:

Nếu có biến đặc trưng (categorical variables), xem xét tương quan giữa các giá trị của chúng và thuộc tính của từng bệnh nhân

***1.5.2.4 Sử Dụng Mô Hình Với Điều Chuẩn Tuyến Tính:***

Sử Dụng Mô Hình Với Điều Chuẩn Tuyến Tính:

Sử dụng mô hình như Ridge Regression để xác định trọng số của các biến.

***1.5.3 Xem Xét Trọng Số Của Các Biến:***

Xem xét trọng số của các biến sau khi huấn luyện mô hình. Trọng số cao hơn đều cho thấy biến đó có ảnh hưởng lớn đến mô hình.

***1.5.4 Lợi Ích và Cảnh Báo:***

**Lợi Ích:**

* **Nhận diện yếu tố ảnh hưởng chính**: Giúp hiểu rõ các yếu tố quan trọng liên quan đến nguy cơ mắc sơ gan, như chỉ số sinh hóa và tình trạng bệnh lý nền.
* **Điều chỉnh quy trình chẩn đoán**: Bác sĩ có thể tập trung vào các xét nghiệm và chỉ số lâm sàng quan trọng để phát hiện sơ gan sớm và chính xác hơn.
* **Thiết kế điều trị cá nhân hóa**: Tạo các chiến lược điều trị tập trung vào các yếu tố rủi ro cao và điều chỉnh điều trị dựa trên các yếu tố quan trọng.
* **Giảm số lượng tính năng**: Tập trung vào các tính năng quan trọng giúp mô hình đơn giản hơn, giảm overfitting và cải thiện hiệu suất dự đoán.

**Cảnh Báo:**

* Dữ liệu không đầy đủ: Nếu dữ liệu không đại diện cho toàn bộ quần thể, các biến quan trọng có thể không phản ánh đúng thực tế.
* Tương tác không được đánh giá đầy đủ: Các yếu tố có thể tương tác theo cách phức tạp không được thể hiện hoàn toàn qua độ quan trọng đơn lẻ.
* Độ quan trọng có thể thay đổi: Độ quan trọng của các tính năng có thể thay đổi tùy thuộc vào tập dữ liệu và tham số mô hình. Cần đánh giá trên nhiều tập dữ liệu và mô hình khác nhau.
* Mô hình có thể không hoàn hảo: Cây ngẫu nhiên có thể không hoàn toàn chính xác trong việc xác định các yếu tố quyết định trong tất cả các trường hợp.
  1. **AN TOÀN VÀ QUY ĐỊNH :**

***1.6.1 Bàn luận về quy định về quy định và an toàn dữ liệu***

Quy định và an toàn dữ liệu đóng vai trò quan trọng, đặc biệt là khi làm việc với thông tin cá nhân của sinh viên. Việc bảo vệ thông tin cá nhân không chỉ là nghĩa vụ pháp lý mà còn là cam kết đạo đức và đảm bảo uy tín của tổ chức. Dưới đây là một số điểm quan trọng cần xem xét khi thảo luận về quy định và an toàn dữ liệu:

1. Tuân Thủ Pháp Luật:

Tổ chức cần tuân thủ tất cả các pháp luật và quy định liên quan đến bảo vệ dữ liệu, như chẳng hạn như GDPR (Nghị định Bảo vệ Dữ liệu Chung) ở Liên minh châu Âu hoặc các quy định tương tự ở các quốc gia khác.

1. Thu Thập Dữ Liệu Tích Cực và Rõ Ràng:

Sinh viên cần được thông báo rõ ràng về mục đích sử dụng thông tin cá nhân của họ. Việc thu thập dữ liệu cần được thực hiện một cách tích cực và chỉ được thực hiện nếu có mục tiêu cụ thể.

1. An Toàn Dữ Liệu:

Tổ chức cần thiết lập các biện pháp an ninh mạnh mẽ để bảo vệ dữ liệu cá nhân của sinh viên khỏi mất mát, sử dụng trái phép hoặc thay đổi bởi bên thứ ba không được phép.

1. Quản Lý Quyền Truy Cập:

Hạn chế quyền truy cập vào dữ liệu cá nhân chỉ cho những người cần thiết và theo nguyên tắc "người dùng ít nhất" (principle of least privilege). Điều này giúp giảm rủi ro mất mát hoặc sử dụng không đúng.

1. Phân Loại Dữ Liệu:

Phân loại dữ liệu theo mức độ nhạy cảm. Dữ liệu cá nhân cần được phân loại và xử lý khác nhau tùy thuộc vào mức độ nhạy cảm của nó.

1. Mã Hóa Dữ Liệu:

Mã hóa dữ liệu cá nhân là quan trọng để bảo vệ dữ liệu khỏi việc đọc được bởi những người không được phép. Cần xem xét việc sử dụng các phương pháp mã hóa mạnh mẽ.

1. Thông Báo Về Dữ Liệu Bị Vi Phạm:

Nếu có vi phạm bảo mật hoặc mất mát dữ liệu cá nhân, tổ chức cần có quy trình để thông báo cho cả những người bị ảnh hưởng và cơ quan quản lý dữ liệu.

1. Quản Lý Hợp Đồng và Các Nhà Cung Cấp Dịch Vụ:

Nếu có bất kỳ dịch vụ nào liên quan đến xử lý dữ liệu sinh viên được chuyển giao cho bên thứ ba, cần có các hợp đồng và biện pháp bảo mật để đảm bảo tính an toàn và tuân thủ.

1. Đào Tạo Và Nhận Thức:

Nhân viên và những người liên quan đến xử lý dữ liệu cần được đào tạo về quy định và biện pháp an toàn dữ liệu để đảm bảo họ hiểu rõ về vai trò và trách nhiệm của mình.

1. Kiểm Tra và Đánh Giá Liên Tục:

Tiến hành kiểm tra định kỳ để đảm bảo rằng biện pháp an toàn dữ liệu vẫn hiệu quả và đáp ứng các yêu cầu mới xuất hiện.

Bảo vệ thông tin cá nhân của sinh viên không chỉ là nghĩa vụ pháp lý, mà còn là cơ sở để xây dựng lòng tin và uy tín của tổ chức. Việc thiết lập và duy trì một chính sách an toàn dữ liệu mạnh mẽ là quan trọng để đảm bảo sự bảo vệ của thông tin cá nhân và tuân thủ các quy định.

* + 1. ***Các biện pháp đảm bảo tính riêng tư và tuân thủ quy định :***

1. Để đảm bảo tính riêng tư và tuân thủ quy định trong quản lý dữ liệu, các tổ chức cần áp dụng một loạt biện pháp an ninh, quy trình, và chính sách. Dưới đây là một số biện pháp quan trọng mà tổ chức có thể thực hiện:
2. Phân Loại Dữ Liệu:

Xác định và phân loại dữ liệu theo mức độ nhạy cảm. Các thông tin cá nhân như tên, địa chỉ, thông tin tài khoản, cần được đối xử khác biệt so với dữ liệu không nhạy cảm.

1. Chính Sách Bảo Mật Dữ Liệu:

Phát triển và thực thi chính sách bảo mật dữ liệu chặt chẽ. Chính sách này cần định rõ các quy định, nguyên tắc và biện pháp bảo mật cụ thể.

1. Mã Hóa Dữ Liệu:

Áp dụng mã hóa dữ liệu cho thông tin nhạy cảm. Mã hóa giúp bảo vệ dữ liệu tránh khỏi việc đọc được khi bị truy cập không được phép.

1. Quản Lý Quyền Truy Cập:

Hạn chế quyền truy cập vào dữ liệu, đặc biệt là với những người không cần thiết. Áp dụng nguyên tắc "người dùng ít nhất" để giảm nguy cơ sử dụng không đúng.

1. Bảo Vệ Trước Mất Mát Dữ Liệu:

Thiết lập biện pháp để ngăn chặn và giảm thiểu rủi ro mất mát dữ liệu, bao gồm việc sao lưu định kỳ và sử dụng các hệ thống bảo mật mạnh mẽ.

1. Kiểm Soát Truy Cập Vật Lý:

Bảo vệ vật lý nơi lưu trữ dữ liệu, như phòng máy chủ và trung tâm dữ liệu, để ngăn chặn truy cập không được phép.

1. Quản Lý Thông Tin Đăng Nhập (Credentials):

Bảo vệ thông tin đăng nhập bằng cách sử dụng mô hình xác thực hai yếu tố (2FA) và đảm bảo rằng mật khẩu được lưu trữ một cách an toàn.

1. Kiểm Tra Bảo Mật Thường Xuyên:

Thực hiện kiểm tra bảo mật định kỳ để phát hiện và giải quyết những lỗ hổng bảo mật một cách nhanh chóng.

i) Chấp Thuận Điều Khoản và Điều Kiện:

Yêu cầu người dùng cung cấp sự chấp thuận rõ ràng trước khi thu thập và sử dụng thông tin cá nhân của họ.

1. Đào Tạo Nhân Viên:

Tổ chức cần cung cấp đào tạo định kỳ về bảo mật dữ liệu và quy định về quyền riêng tư cho nhân viên để đảm bảo họ hiểu và tuân thủ các quy tắc.

1. Đối Phó với Sự Cố Bảo Mật:

Xây dựng kế hoạch đối phó với sự cố bảo mật, bao gồm cách thông báo cho bên liên quan và cách giải quyết vấn đề.

1. Kiểm Tra Tuân Thủ:

Thực hiện các đánh giá định kỳ để đảm bảo rằng tổ chức tuân thủ đầy đủ các quy định về bảo mật và quyền riêng tư.

1. Bảo Vệ Dữ Liệu Trong Di Động:

Đối với dữ liệu di động, hãy sử dụng các biện pháp bảo mật như mã hóa dữ liệu và quản lý thiết bị di động.

1. Báo Cáo Bảo Mật Định Kỳ:

Báo cáo về tình trạng bảo mật định kỳ đến ban lãnh đạo và bảo vệ thông tin cá nhân trước rủi ro.

Những biện pháp này cùng nhau giúp xây dựng một hệ thống bảo mật mạnh mẽ, đảm bảo tính riêng tư và tuân thủ quy định trong quản lý thông tin cá nhân.

**1.7 Công nghệ và ngôn ngữ lập trình**

### 1.7.1 Trình bày các công nghệ và ngôn ngữ lập trình thích hợp cho việc triển khai mô hình dự đoán bệnh sơ gan

* Khi triển khai mô hình dự đoán bệnh sơ gan, việc chọn lựa công nghệ và ngôn ngữ lập trình phụ thuộc vào nhiều yếu tố như quy mô dự án, tính di động, hiệu suất, và sự linh hoạt. Dưới đây là một số công nghệ và ngôn ngữ lập trình thích hợp cho việc triển khai mô hình dự đoán bệnh sơ gan:

1. Ngôn Ngữ Lập Trình:

Python: Python là một ngôn ngữ lập trình phổ biến trong lĩnh vực học máy và khoa học dữ liệu. Đa dạng các thư viện như scikit-learn, TensorFlow, PyTorch cho phép xây dựng và triển khai mô hình một cách dễ dàng.

Java và Scala: Cho các hệ thống lớn và có yêu cầu về hiệu suất cao, Java và Scala thường được sử dụng. Apache Spark là một framework hỗ trợ xử lý lớn và có thể tích hợp tốt với Java và Scala.

R: Nếu dự án chủ yếu tập trung vào phân tích thống kê và khám phá dữ liệu, R là một lựa chọn tốt.

1. Cơ Sở Dữ Liệu:

MySQL và PostgreSQL: Cho các dự án có quy mô nhỏ và trung bình, các hệ quản trị cơ sở dữ liệu quan hệ như MySQL và PostgreSQL là lựa chọn phổ biến.

MongoDB và Cassandra: Đối với dự án lớn có nhu cầu lưu trữ lớn, cơ sở dữ liệu NoSQL như MongoDB và Cassandra cung cấp tính linh hoạt và khả năng mở rộng.

Redis: Để lưu trữ và truy xuất dữ liệu trong thời gian thực, Redis là một cơ sở dữ liệu key-value hiệu suất cao.

1. Công Nghệ Web và Framework:

Django và Flask (Python): Đối với triển khai ứng dụng web dựa trên Python, Django và Flask là những framework phổ biến.

Spring Boot (Java): Cho các ứng dụng Java, Spring Boot cung cấp một cách nhanh chóng và linh hoạt để xây dựng và triển khai các dự án.

React, Angular, Vue (JavaScript): Đối với giao diện người dùng, các thư viện và framework JavaScript như React, Angular, và Vue.js thường được sử dụng.

1. Containerization và Orchestration:

Docker: Docker cung cấp một cách đóng gói ứng dụng và môi trường để triển khai một cách đồng nhất trên mọi nền tảng.

Kubernetes: Cho các hệ thống phức tạp, Kubernetes giúp quản lý và triển khai các container một cách hiệu quả.

1. Cloud Computing:

AWS, Azure, Google Cloud: Các dịch vụ đám mây như AWS, Azure, và Google Cloud cung cấp nền tảng mạnh mẽ để triển khai và quản lý ứng dụng mô hình dự đoán bệnh sơ gan.

f) Công Nghệ Đối Với Mô Hình Học Sâu:

TensorFlow và PyTorch: Đối với mô hình học sâu, TensorFlow và PyTorch là những thư viện phổ biến và mạnh mẽ.

ONNX (Open Neural Network Exchange): ONNX là một định dạng chung cho mô hình học máy, giúp tương thích giữa các framework khác nhau.

Lưu ý Quan Trọng:

Lựa Chọn Dựa Trên Yêu Cầu Cụ Thể: Lựa chọn công nghệ và ngôn ngữ lập trình nên dựa trên yêu cầu cụ thể của dự án, khả năng kỹ năng của đội ngũ phát triển, và các yếu tố khác nhau như hiệu suất, mở rộng, và bảo mật.

Tích Hợp và Mở Rộng:

Đảm bảo rằng các thành phần của hệ thống có thể tích hợp chặt chẽ và dễ dàng mở rộng để đáp ứng nhu cầu tương lai.

Bảo Mật:

Lưu ý đến các biện pháp bảo mật cần thiết để bảo vệ dữ liệu cá nhân và mô hình.

Việc lựa chọn công nghệ và ngôn ngữ lập trình phù hợp đặt ra những cơ hội và thách thức, và nó nên phản ánh sự cân nhắc giữa hiệu suất, tính linh hoạt và yêu cầu cụ thể của dự án.

**1.8 TIỀN XỬ LÝ DỮ LIỆU :**

Tiền xử lý dữ liệu là một phần quan trọng trong quá trình xây dựng mô hình dự đoán bệnh sơ gan. Các bước tiền xử lý giúp làm sạch, chuẩn hóa và chuẩn bị dữ liệu để có thể áp dụng mô hình học máy hiệu quả hơn. Dưới đây là một số bước tiền xử lý dữ liệu thường được thực hiện cho đề tài dự đoán bệnh sơ gan:

* + 1. ***Xử lý dữ liệu thiếu :***

Xử lý dữ liệu thiếu là quá trình đối phó với các giá trị thiếu hoặc không tồn tại trong một tập dữ liệu. Trong quy trình phân tích dữ liệu, việc xử lý dữ liệu thiếu là quan trọng để đảm bảo tính toàn vẹn và chính xác của kết quả. Khi dữ liệu có giá trị bị thiếu, có thể gây ảnh hưởng đến hiểu quả của các phương pháp phân tích, và nó cũng có thể tạo ra nhiễu và đưa ra kết luận sai lầm.

* + 1. ***Loại bỏ dữ liệu nhiễu :***

Loại bỏ dữ liệu nhiễu (noise) là quá trình loại bỏ hoặc giảm thiểu các thông tin không mong muốn, không có ý nghĩa, và không đại diện đúng cho đặc trưng cần phân tích từ tập dữ liệu. Dữ liệu nhiễu có thể xuất hiện do nhiều nguyên nhân, bao gồm lỗi đo lường, nhiễu ngẫu nhiên, hay nhiễu từ các quá trình thu thập dữ liệu.

* + 1. ***Chuẩn hóa dữ liệu :***

Chuẩn hóa dữ liệu là quá trình biến đổi dữ liệu từ một phạm vi giá trị không đồng nhất thành một phạm vi giá trị chuẩn. Mục tiêu của việc chuẩn hóa là làm cho các biến số có đơn vị đo lường khác nhau hoặc phạm vi giá trị khác nhau trở nên tương đồng, giúp cải thiện tính so sánh và tương tự giữa chúng.

* + 1. ***Xử lý biến độc lập :***

Xử lý biến độc lập là quá trình chuẩn bị và làm sạch dữ liệu liên quan đến biến độc lập (independent variable) trong phân tích dữ liệu hoặc mô hình học máy. Biến độc lập thường là các yếu tố hoặc đặc trưng mà mô hình sẽ sử dụng để dự đoán một biến phụ thuộc (dependent variable).

* + 1. ***Tạo các đặc trưng mới :***

Tạo các đặc trưng mới (Feature Engineering) là quá trình sáng tạo và thêm vào tập dữ liệu những đặc trưng mới có thể mang lại thông tin hữu ích và cải thiện hiệu suất của mô hình học máy hoặc phương pháp phân tích dữ liệu. Mục tiêu của feature engineering là tối ưu hóa khả năng dự đoán và mô tả của mô hình thông qua việc chọn lựa, biến đổi, hoặc tạo ra các đặc trưng mới từ dữ liệu gốc.

* + 1. ***Xử lý dữ liệu dư thừa :***

Xử lý dữ liệu dư thừa là quá trình loại bỏ hoặc giảm bớt các đặc trưng (biến) không mang lại nhiều thông tin giá trị cho mô hình hoặc phân tích dữ liệu. Các đặc trưng dư thừa có thể làm tăng độ phức tạp của mô hình mà không đóng góp nhiều vào khả năng dự đoán hoặc hiểu được dữ liệu.

* + 1. ***Xử lý dữ liệu đặc trưng không phải số :***

Xử lý dữ liệu đặc trưng không phải số là quá trình chuyển đổi và xử lý các đặc trưng có dạng không phải số (categorical features) trong dữ liệu. Trong nhiều trường hợp, mô hình máy học và các phương pháp thống kê yêu cầu dữ liệu đầu vào ở dạng số, và do đó, cần phải thực hiện các biện pháp để biến đổi các đặc trưng không phải số thành dạng số.

* + 1. ***Xử lý dữ liệu phân phối không đồng đều :***

Xử lý dữ liệu phân phối không đồng đều (skewed distribution) là quá trình điều chỉnh hoặc biến đổi dữ liệu để làm cho phân phối của chúng trở nên đối xứng và gần với phân phối chuẩn (normal distribution). Trong một số trường hợp, phân phối không đồng đều có thể gây ra ảnh hưởng lớn đến hiệu suất của mô hình học máy hoặc các phương pháp thống kê, đặc biệt là khi sử dụng các mô hình yêu cầu giả định về phân phối dữ liệu.

* + 1. ***Tách tập huấn luyện và tập kiểm tra :***

Tách tập huấn luyện (training set) và tập kiểm tra (test set) là một phần quan trọng trong quá trình xây dựng và đánh giá mô hình máy học. Mục tiêu của việc này là đảm bảo mô hình được huấn luyện trên một tập dữ liệu độc lập và sau đó được kiểm thử trên một tập dữ liệu mà nó chưa từng thấy trước đó. Dưới đây là quy trình thường được sử dụng

***1.8.10 Kiểm tra và đánh giá dữ liệu :***

Kiểm tra và đánh giá dữ liệu là quá trình đánh giá chất lượng và tính đúng đắn của dữ liệu trước khi sử dụng chúng để xây dựng mô hình học máy hoặc thực hiện phân tích dữ liệu. Điều này là quan trọng để đảm bảo rằng dữ liệu được xử lý đúng cách, không chứa lỗi, và phù hợp với mục tiêu của quá trình phân tích hoặc mô hình hóa.

* + 1. **Giữ nguyên dữ liệu gốc :**

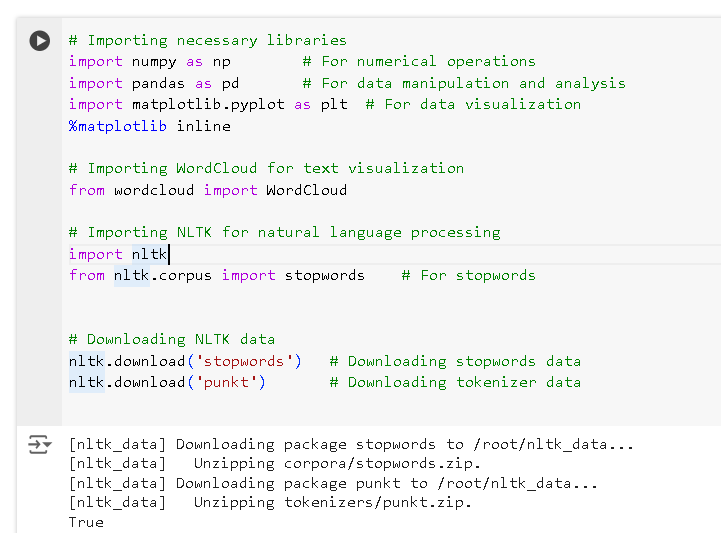
Giữ nguyên dữ liệu gốc (keeping the raw data intact) có nghĩa là không thay đổi, không biến đổi hoặc không làm mất mát thông tin nào từ dữ liệu nguyên bản. Khi bạn giữ nguyên dữ liệu gốc, bạn duy trì sự không thay đổi của dữ liệu đầu vào và làm việc trực tiếp trên bản sao của dữ liệu đó để thực hiện các phương pháp xử lý, phân tích, hoặc mô hình hóa.

**CHƯƠNG II : THỰC NGHIỆM VÀ KẾT QUẢ**

## 2.1 Giới thiệu đề tài :

Trong thời đại số hóa hiện nay, tin nhắn SMS đã trở thành phương tiện giao tiếp quan trọng, nhưng việc nhận diện và phân loại tin nhắn rác (spam) là một vấn đề lớn. Đề tài này tập trung vào việc xây dựng một mô hình phân loại tin nhắn SMS tự động, giúp xác định liệu một tin nhắn có phải là tin nhắn rác (spam) hay tin nhắn thông thường (ham).  
Dữ liệu sử dụng trong dự án là một tập hợp các tin nhắn SMS đã được phân loại trước, bao gồm các tin nhắn thuộc hai loại: "ham" (tin nhắn hợp lệ) và "spam" (tin nhắn quảng cáo hoặc tin nhắn không mong muốn). Để thực hiện việc phân loại, chúng tôi sử dụng các kỹ thuật học máy và xử lý ngôn ngữ tự nhiên (NLP) nhằm phân tích nội dung và tạo ra mô hình dự đoán chính xác.  
Mục tiêu của đề tài là xây dựng một hệ thống có khả năng phân loại tin nhắn một cách hiệu quả, từ đó giúp người dùng tránh được các phiền toái do tin nhắn rác gây ra.

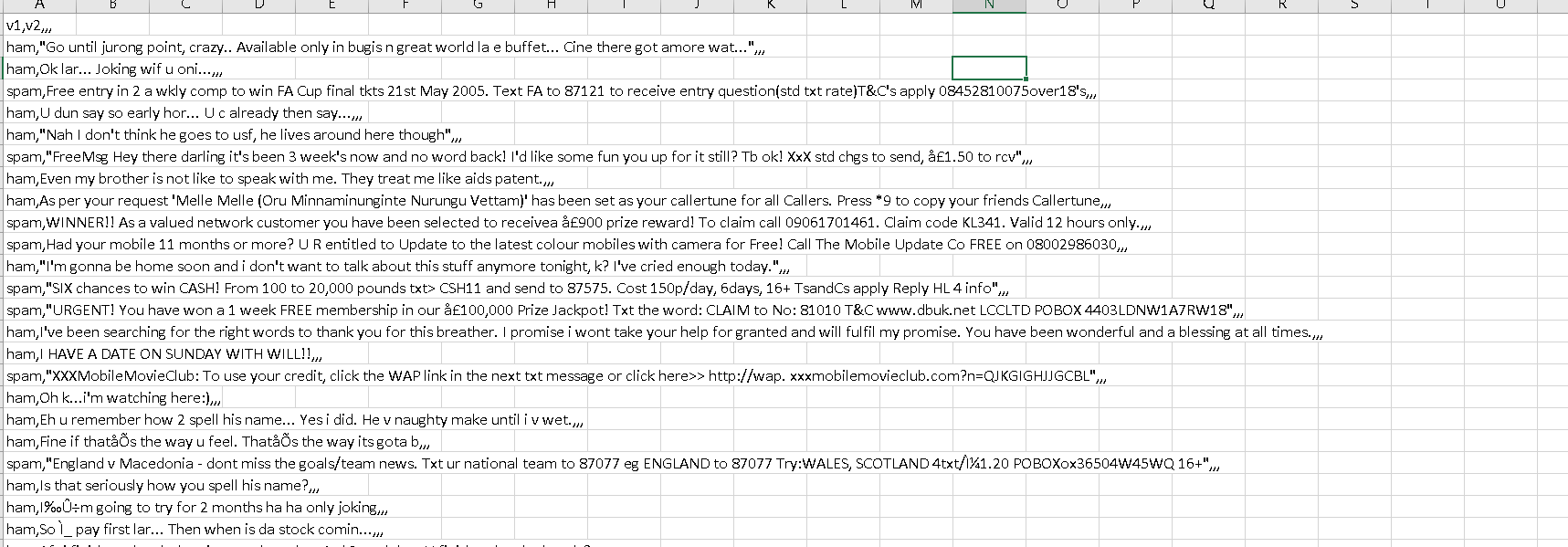
## 2.2 Khai báo các thư viện để triển khai đồ án thực hiện



Hình 2. 1 Các thư viện

**2.3 Data Cleaning và data**

**2.3.1 Đọc và hiểu dữ liệu**

* 

Hình 2. 2 :Data Train

****

Hình 2. 3 :Code mô tả đọc dữ liệu từ file CSV

Đoạn mã này đọc dữ liệu từ tệp CSV và tạo một bảng có định dạng màu sắc tùy chỉnh cho phần tiêu đề. Dưới đây là giải thích chi tiết từng dòng:

df = pd.read\_csv('/content/sample\_data/spam.csv', encoding='latin1')

- Dòng này sử dụng thư viện pandas để đọc tệp CSV có tên "spam.csv". Tệp này được lưu trong thư mục `/content/sample\_data/` (thường được sử dụng trong Google Colab).

- Tham số `encoding='latin1'` chỉ định mã hóa của tệp là latin1 (một chuẩn mã hóa ký tự, cần thiết khi có các ký tự đặc biệt trong dữ liệu).

styled\_df = df.head()

- Lệnh này lấy 5 dòng đầu tiên từ DataFrame `df` bằng phương thức `head()`, rồi lưu kết quả vào biến `styled\_df`. Đây là cách hiển thị một phần nhỏ dữ liệu để kiểm tra nhanh nội dung.

styled\_df = styled\_df.style.set\_table\_styles([

{"selector": "th", "props": [("color", 'black'), ("background-color", "#FF00CC")]}

])

- Dòng này sử dụng phương thức \*\*`style.set\_table\_styles()`\*\* để tùy chỉnh bảng hiển thị:

- `selector: "th": Chọn tất cả các tiêu đề của bảng (thẻ HTML `<th>`, đại diện cho các tiêu đề của cột).

- `props: [("color", 'black'), ("background-color", "#FF00CC")]`: Đặt thuộc tính màu sắc cho tiêu đề:

- Màu chữ là đen (`'black'`).

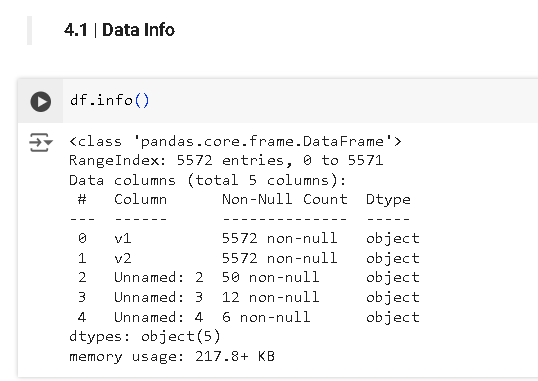
- Màu nền của tiêu đề là hồng đậm (`#FF00CC`).

styled\_df

- Cuối cùng, biến `styled\_df` (là một DataFrame đã được định dạng) được trả về và hiển thị với các tùy chỉnh về màu sắc cho tiêu đề.

Tóm lại, đoạn mã này tải dữ liệu từ tệp CSV, hiển thị 5 dòng đầu tiên của tập dữ liệu và áp dụng một số định dạng cho bảng (màu nền và màu chữ của phần tiêu đề).

### 2.3.2 Data Cleaning



Hình 2. 4 : Data info



Hình 2. 5 : Xóa các cột dữ liệu không có giá trị

Giải thích chi tiết từng phần:

df.drop(columns = ['Unnamed: 2', 'Unnamed: 3', 'Unnamed: 4'], inplace = True)

* Dòng này loại bỏ các cột có tên **'Unnamed: 2'**, **'Unnamed: 3'**, và **'Unnamed: 4'** khỏi DataFrame df.
* Tham số **inplace=True** chỉ ra rằng việc xóa cột sẽ được thực hiện trực tiếp trên DataFrame gốc, không cần tạo bản sao mới.

styled\_df = df.head(5).style

* Lệnh này lấy 5 dòng đầu tiên từ DataFrame df và lưu dưới dạng **styled DataFrame** (dùng để tùy chỉnh hiển thị).
* Biến styled\_df không chỉ chứa dữ liệu mà còn cho phép áp dụng các kiểu định dạng như màu sắc, phông chữ.

styled\_df.set\_table\_styles([

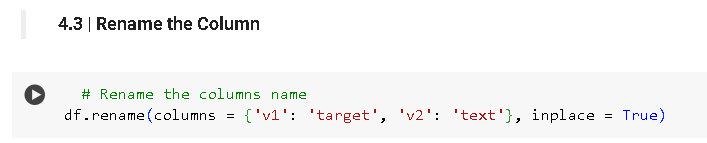
{"selector": "th", "props": [("color", 'Black'), ("background-color", "#FF00CC"), ('font-weight', 'bold')] }

])

* Lệnh này tùy chỉnh hiển thị phần **tiêu đề của bảng (th)**:
  + **selector: "th"**: Chọn tất cả các tiêu đề (thẻ HTML <th>).
  + **props**: Đặt các thuộc tính hiển thị:
    - **("color", 'Black')**: Màu chữ của tiêu đề sẽ là **đen**.
    - **("background-color", "#FF00CC")**: Màu nền của tiêu đề sẽ là **hồng đậm**.
    - **('font-weight', 'bold')**: Đặt tiêu đề có **chữ in đậm**.

Tóm tắt:

* **Loại bỏ cột không cần thiết** từ DataFrame.
* **Hiển thị 5 dòng đầu tiên** của DataFrame với phần tiêu đề được tùy chỉnh: chữ đen, nền hồng đậm, và chữ in đậm.



Hình 2. 6 : Đổi tên cột



Hình 2. 7 : Chuyển đổi các giá trị phân loại (như "spam" và "ham") thành các số nguyên (0, 1, )

Giải thích chi tiết :

from sklearn.preprocessing import LabelEncoder

* Dòng này nhập **LabelEncoder** từ thư viện **scikit-learn**. Công cụ này được sử dụng để chuyển đổi các nhãn phân loại (chẳng hạn như "spam" và "ham") thành các giá trị số.

encoder = LabelEncoder()

* Khởi tạo một đối tượng **LabelEncoder** có tên là **encoder**.

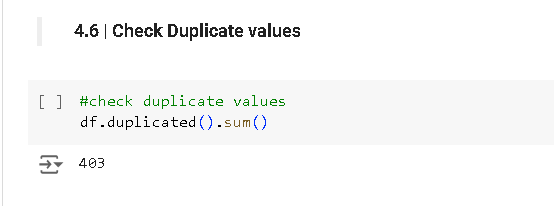
df['target'] = encoder.fit\_transform(df['target'])

* **fit\_transform(df['target'])**:
  + **fit()**: Học các nhãn phân loại từ cột **target**. Nếu cột này chứa các giá trị phân loại như "spam" và "ham", LabelEncoder sẽ gán các nhãn này thành các giá trị số (ví dụ: "spam" = 1, "ham" = 0).
  + **transform()**: Sau khi học xong, nó chuyển đổi tất cả các giá trị phân loại trong cột **target** thành số.
* **df['target']**: Cột **target** của DataFrame df sẽ chứa các giá trị số sau khi được mã hóa.

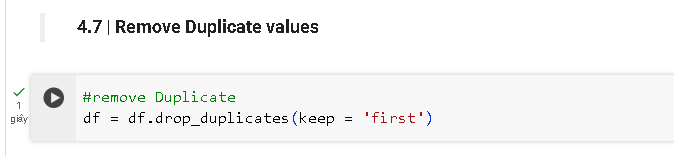
Tóm tắt:

* Đoạn mã này chuyển đổi các giá trị phân loại (như "spam" và "ham") trong cột target của DataFrame thành các số nguyên (0, 1, v.v.). Quá trình này rất quan trọng để mô hình học máy có thể xử lý dữ liệu phân loại, vì hầu hết các mô hình yêu cầu dữ liệu ở dạng số.

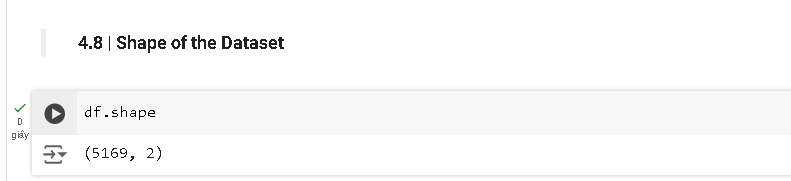
Hình 2. 8 : Kiểm tra giá trị bị thiếu



Hình 2. 9 : Kiểm tra số lượng các hàng bị trùng lặp trong DataFrame



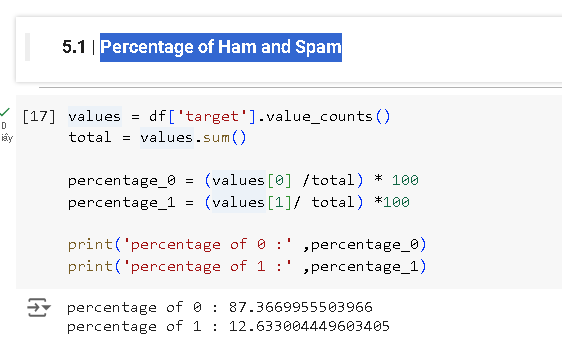
Hình 2. 10 : Xóa các hàng trùng lặp khỏi DataFrame và giữ lại hàng đầu tiên trong mỗi nhóm trùng lặp.



Hình 2. 11 Kích thước của dữ liệu (gồm 5169 hàng , 2 cột)

* 1. **EDA** (**Exploratory Data Analysis) (Phân tích Dữ liệu Khám Phá)**

### 2.4.1 Percentage of Ham and Spam



Hình 2. 12 Tỷ lệ phần trăm của các giá trị Ham và spam

Giải Thích Chi Tiết

1. **Đếm số lượng các giá trị trong cột target**

values = df['target'].value\_counts()

* + df['target'].value\_counts() trả về một Series chứa số lượng các giá trị duy nhất trong cột target. Các chỉ số của Series này là các giá trị duy nhất trong cột, và các giá trị của Series là số lượng xuất hiện của mỗi giá trị.

1. **Tính tổng số lượng giá trị**

total = values.sum()

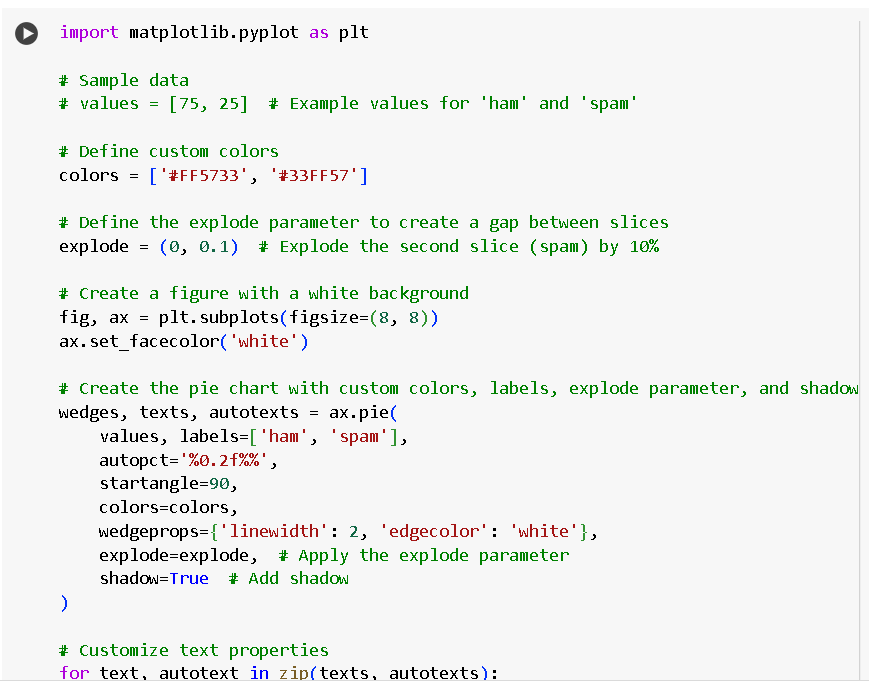
* + values.sum() tính tổng số lượng giá trị trong cột target. Điều này là tổng số lượng các mẫu dữ liệu trong cột target.

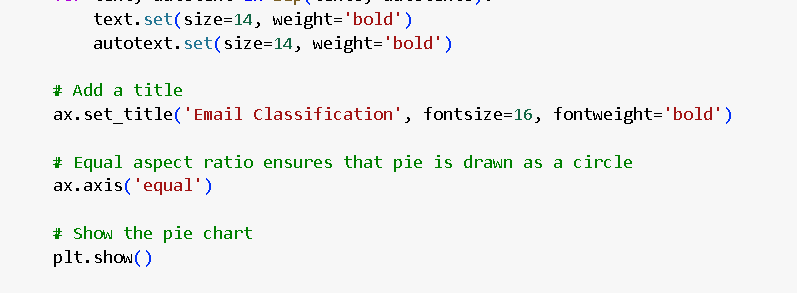
1. **Tính tỷ lệ phần trăm của các giá trị 0 và 1**

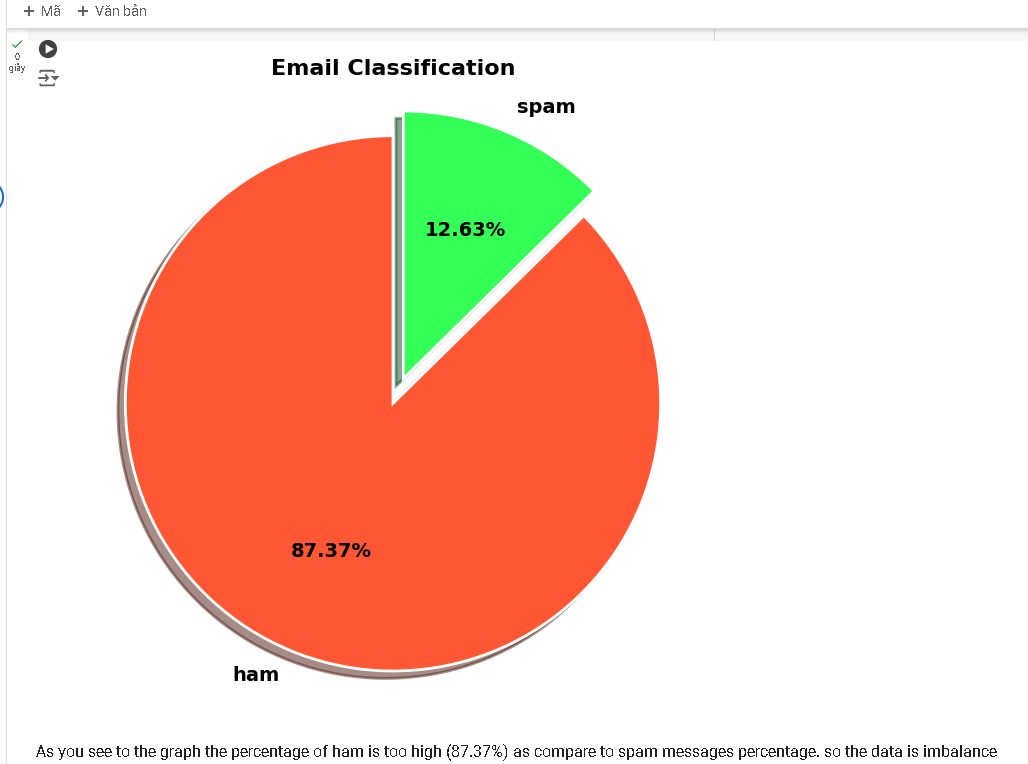
percentage\_0 = (values[0] / total) \* 100

percentage\_1 = (values[1] / total) \* 100

* + values[0] là số lượng các giá trị 0 trong cột target.
  + values[1] là số lượng các giá trị 1 trong cột target.
  + (values[0] / total) \* 100 tính tỷ lệ phần trăm của giá trị 0 bằng cách chia số lượng giá trị 0 cho tổng số lượng và nhân với 100.
  + (values[1] / total) \* 100 làm tương tự cho giá trị 1.

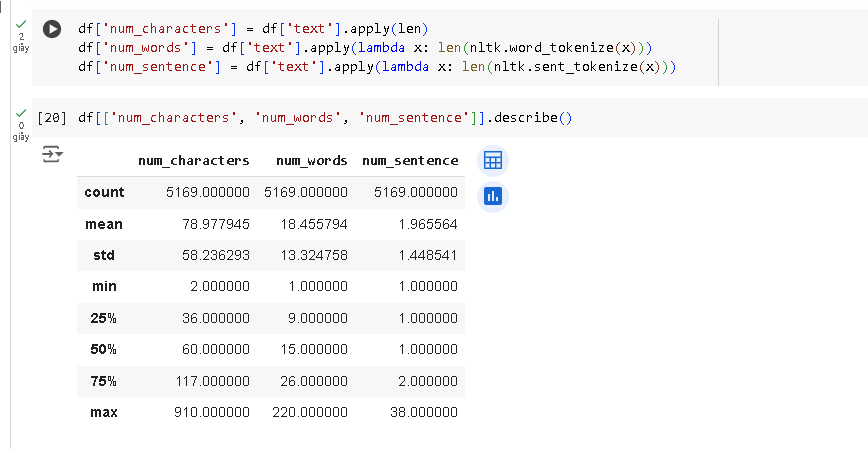






Hình 2. 13 Trực quan hóa tỉ lệ phần trăm của spam và ham

### 2.4.2 Text Length and Structure Analysis



Hình 2. 14 Tính toán số lượng ký tự, số lượng từ, và số lượng câu trong mỗi tin nhắn văn bản và sau đó hiển thị các thống kê mô tả

Giải Thích Mã

1. **Tính số lượng ký tự trong mỗi tin nhắn**

df['num\_characters'] = df['text'].apply(len)

* + df['text'].apply(len) tính số lượng ký tự trong mỗi tin nhắn văn bản và lưu kết quả vào cột mới num\_characters.

1. **Tính số lượng từ trong mỗi tin nhắn**

df['num\_words'] = df['text'].apply(lambda x: len(nltk.word\_tokenize(x)))

* + nltk.word\_tokenize(x) sử dụng thư viện NLTK để tách tin nhắn thành các từ. len() tính số lượng từ trong tin nhắn và kết quả được lưu vào cột num\_words.

1. **Tính số lượng câu trong mỗi tin nhắn**

df['num\_sentence'] = df['text'].apply(lambda x: len(nltk.sent\_tokenize(x)))

* + nltk.sent\_tokenize(x) tách tin nhắn thành các câu. len() tính số lượng câu trong tin nhắn và kết quả được lưu vào cột num\_sentence.

1. **Hiển thị các thống kê mô tả cho các cột**

df[['num\_characters', 'num\_words', 'num\_sentence']].describe()

* + describe() cung cấp các thống kê mô tả cho các cột num\_characters, num\_words, và num\_sentence, bao gồm số lượng, trung bình, độ lệch chuẩn, giá trị tối thiểu và tối đa, cũng như các phân vị (25%, 50%, 75%)

Kết Quả Thống Kê Mô Tả

Dưới đây là cách giải thích các số liệu thống kê mô tả:

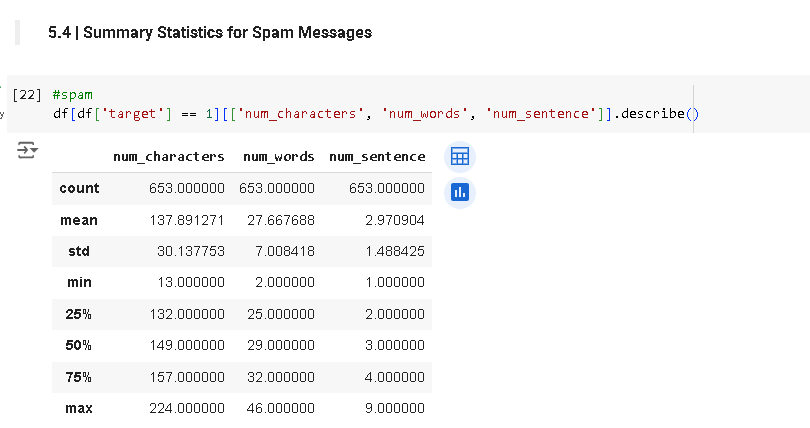
* **count**: Số lượng dữ liệu trong mỗi cột, đều là 5169 trong trường hợp này (tương ứng với số lượng tin nhắn).
* **mean**: Trung bình số lượng ký tự, từ, và câu trong các tin nhắn.
  + num\_characters: Trung bình 78.98 ký tự mỗi tin nhắn.
  + num\_words: Trung bình 18.46 từ mỗi tin nhắn.
  + num\_sentence: Trung bình 1.97 câu mỗi tin nhắn.
* **std**: Độ lệch chuẩn cho biết mức độ phân tán của dữ liệu xung quanh giá trị trung bình.
  + num\_characters: Độ lệch chuẩn 58.24 cho số lượng ký tự.
  + num\_words: Độ lệch chuẩn 13.32 cho số lượng từ.
  + num\_sentence: Độ lệch chuẩn 1.45 cho số lượng câu.
* **min**: Giá trị nhỏ nhất.
  + num\_characters: 2 ký tự (như tin nhắn rất ngắn).
  + num\_words: 1 từ (như tin nhắn đơn từ).
  + num\_sentence: 1 câu.
* **25% (Q1)**: Phân vị thứ 25, nghĩa là 25% số tin nhắn có giá trị nhỏ hơn hoặc bằng giá trị này.
  + num\_characters: 36 ký tự.
  + num\_words: 9 từ.
  + num\_sentence: 1 câu.
* **50% (Median)**: Trung vị, tức là 50% số tin nhắn có giá trị nhỏ hơn hoặc bằng giá trị này.
  + num\_characters: 60 ký tự.
  + num\_words: 15 từ.
  + num\_sentence: 1 câu.
* **75% (Q3)**: Phân vị thứ 75, nghĩa là 75% số tin nhắn có giá trị nhỏ hơn hoặc bằng giá trị này.
  + num\_characters: 117 ký tự.
  + num\_words: 26 từ.
  + num\_sentence: 2 câu.
* **max**: Giá trị lớn nhất.
  + num\_characters: 910 ký tự (như tin nhắn rất dài).
  + num\_words: 220 từ (như tin nhắn rất dài).
  + num\_sentence: 38 câu.

Kết Luận

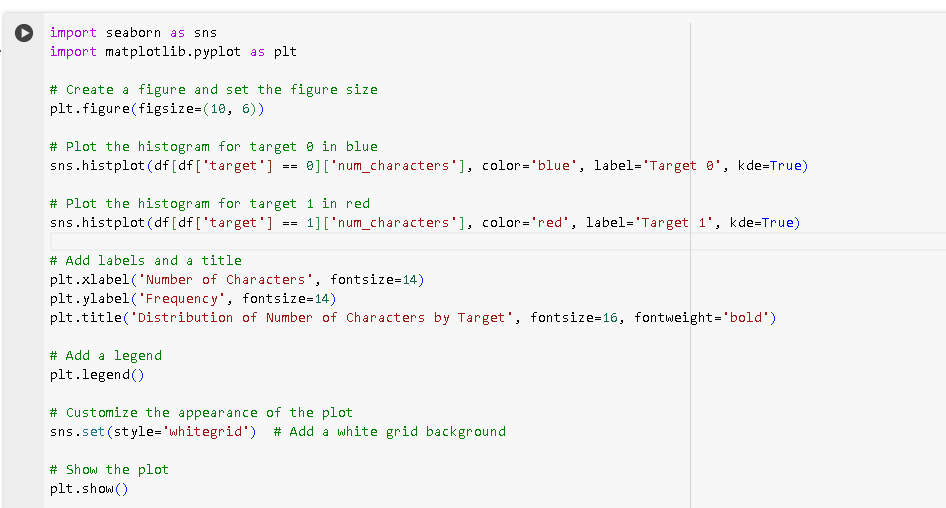
Các thống kê mô tả này cung cấp cái nhìn tổng quan về cấu trúc của các tin nhắn trong tập dữ liệu của bạn, giúp bạn hiểu về độ dài của các tin nhắn và số lượng từ và câu trong đó. Điều này có thể hữu ích khi phân tích hoặc tiền xử lý dữ liệu văn bản cho các ứng dụng tiếp theo như phân loại văn bản, phân tích cảm xúc, v.v.

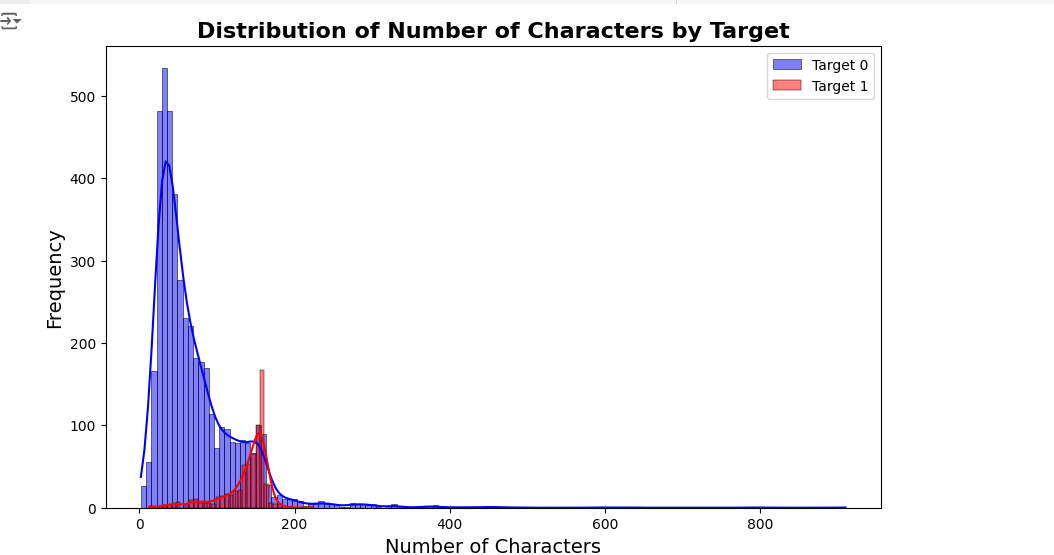


Hình 2. 15 Thống kê tóm tắt cho các tin nhắn không spam số lượng ký tự, số lượng từ, và số lượng câu trong mỗi tin nhắn



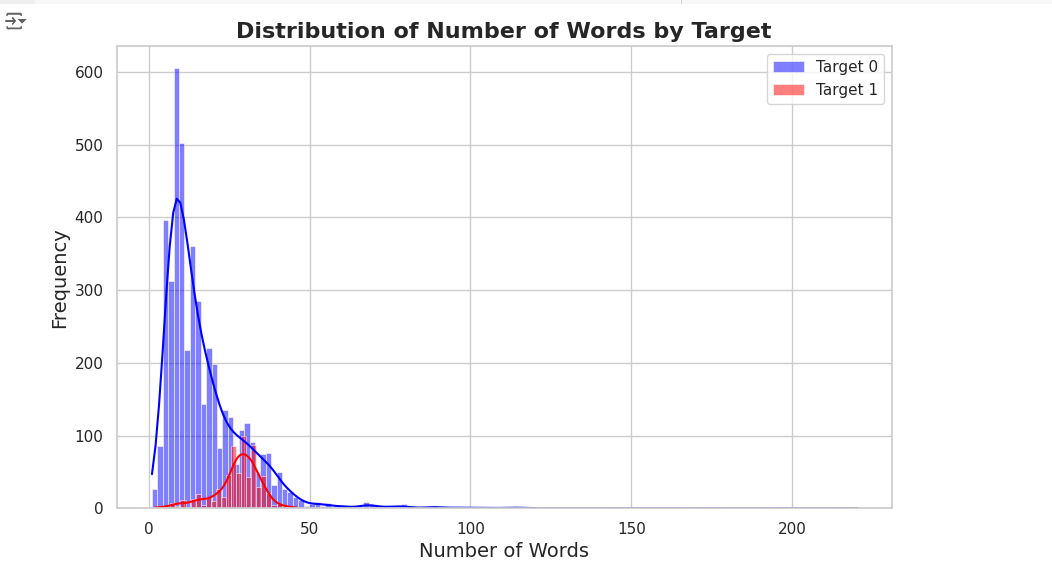
Hình 2. 16 Thống kê tóm tắt cho các tin nhắn spam số lượng ký tự, số lượng từ, và số lượng câu trong mỗi tin nhắn



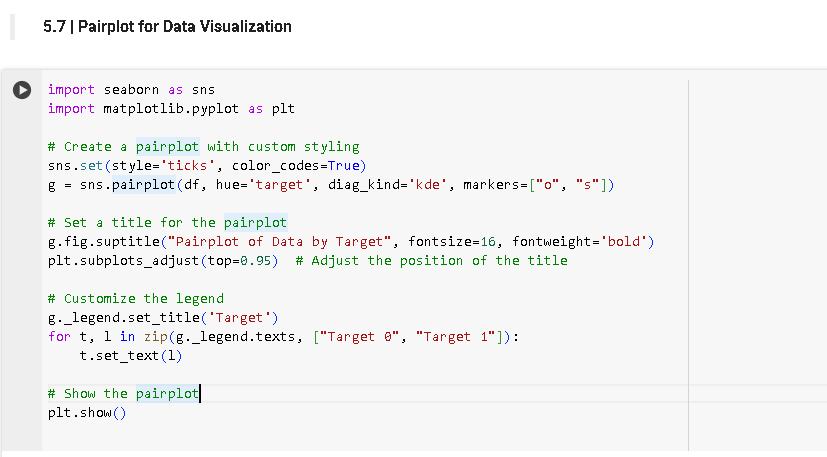


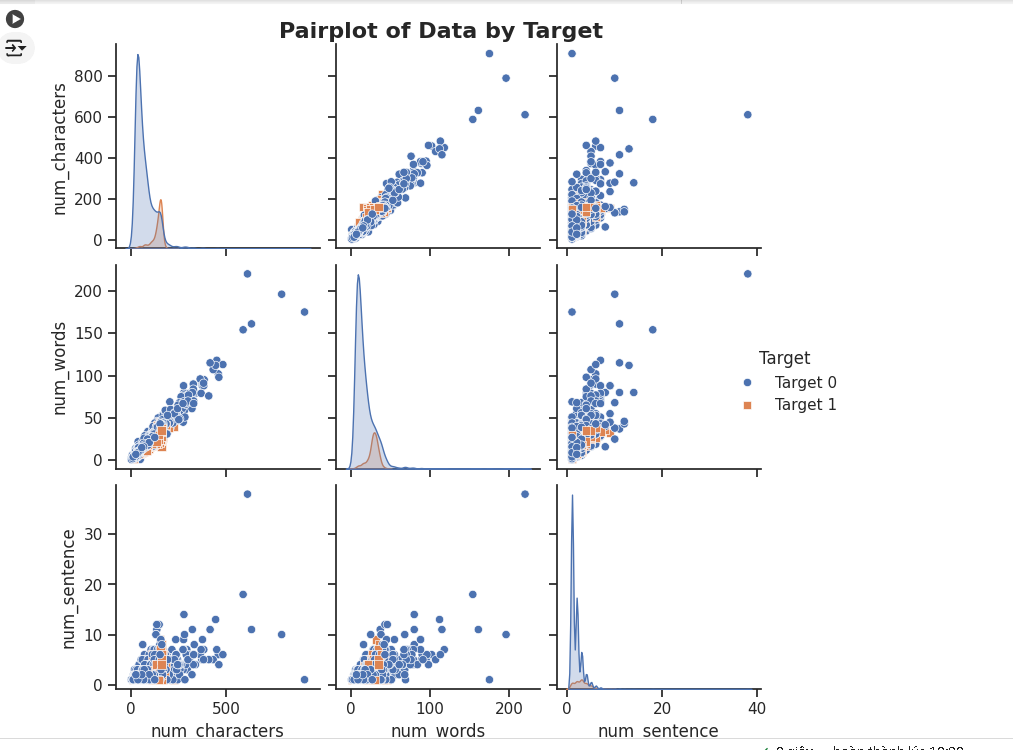
Hình 2. 17 Biểu đồ trực quán hoa số lượng ký tự spam và ham



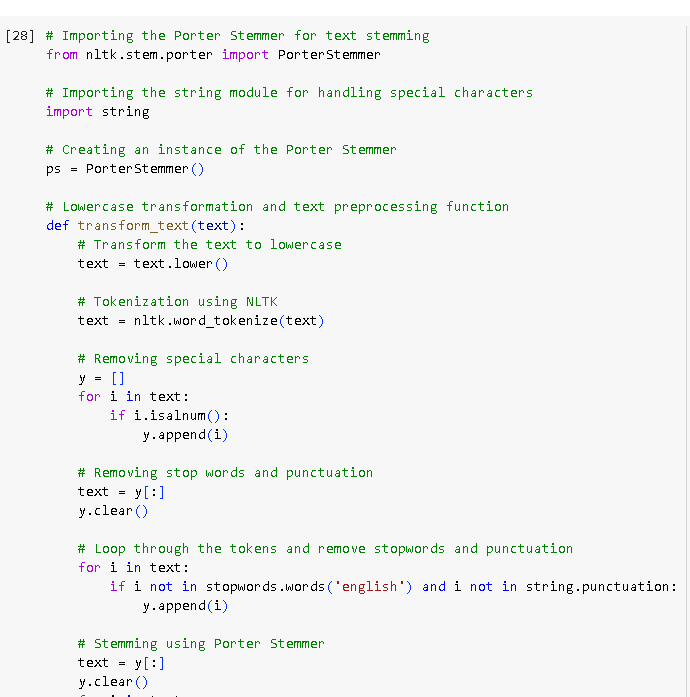
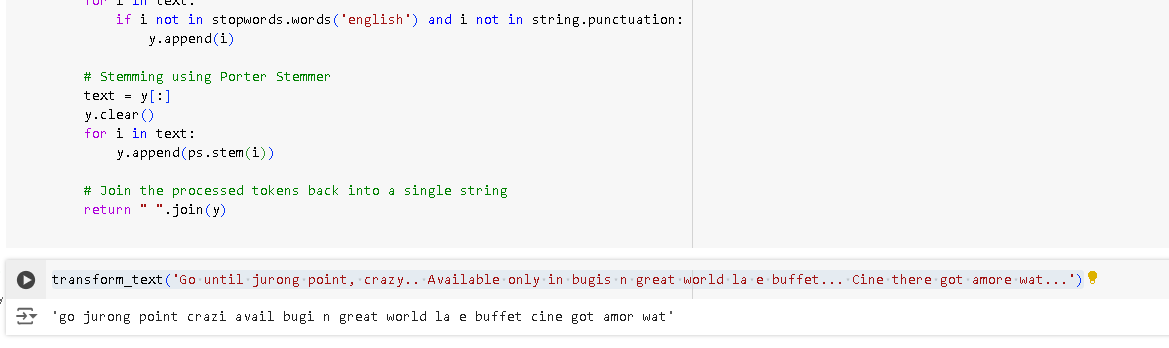


Hình 2. 18 Biểu đồ trực quán hoa số lượng từ spam và ham





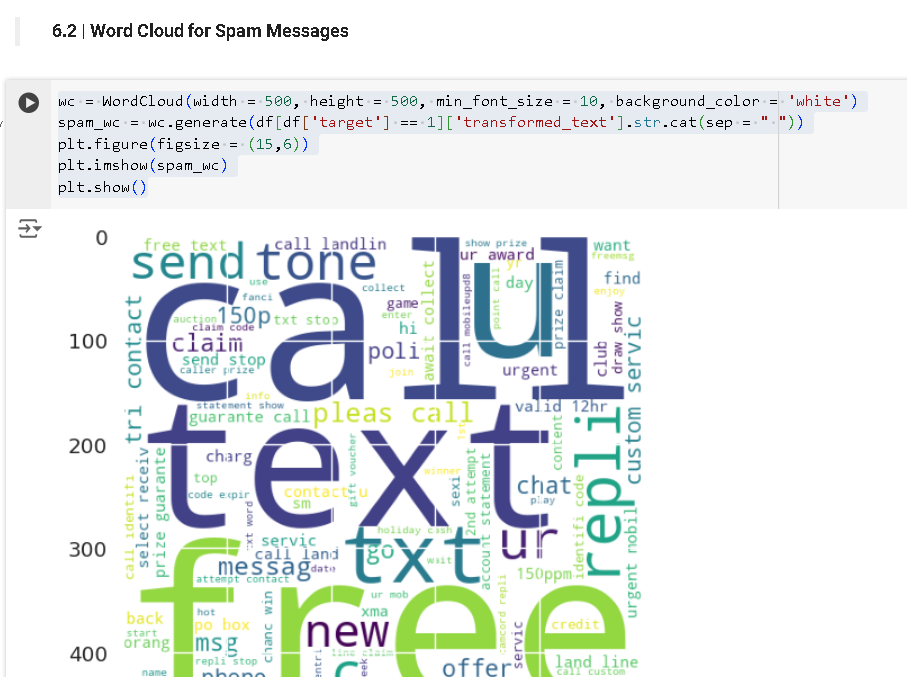
Hình 2. 19 Biểu đồ trực quán hóa Pairplot về số lượng từ ,ký tự ,câu spam và ham

2.5 Data Preprocessing (tiền xử lý dữ liệu) ****

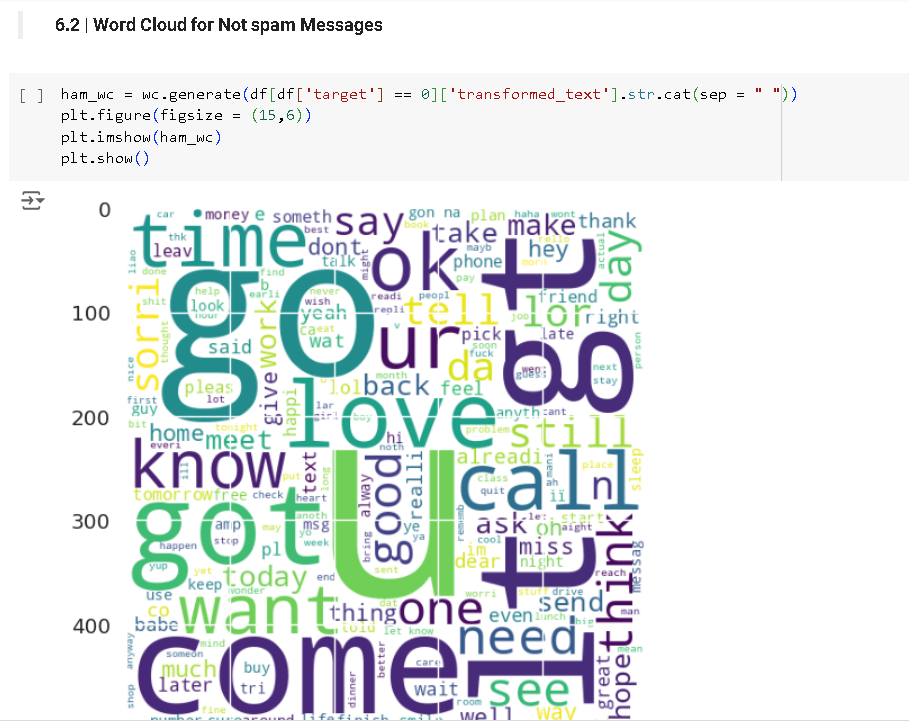
Hình 2. 20 Xây dựng transform\_text để biến đổi văn bản, tách từ, loại bỏ ký tự đặc biệt, và thực hiện stemming



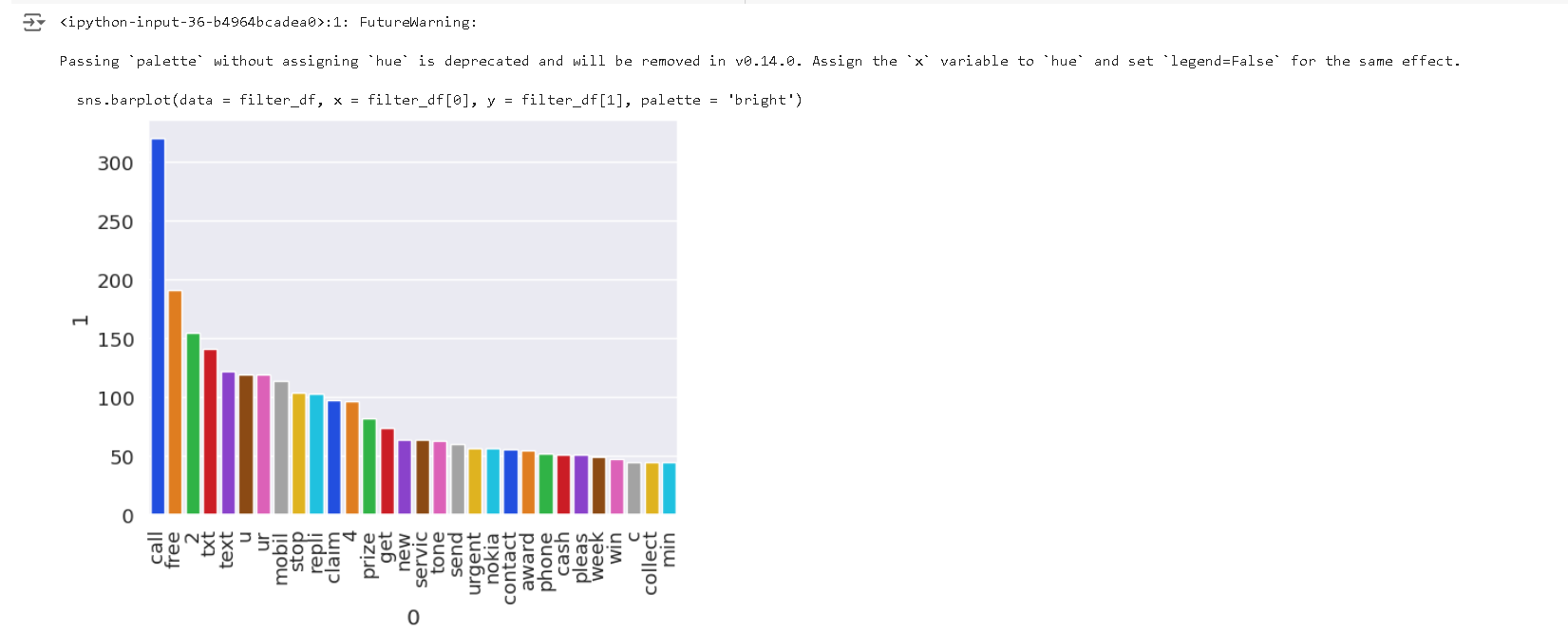
Hình 2. 21 Áp dụng transform\_text vào dữ liệu



Hình 2. 22 Biểu đồ từ đám mây từ (word cloud) cho văn bản trong cột transformed\_text của DataFrame df đối với văn bản spam

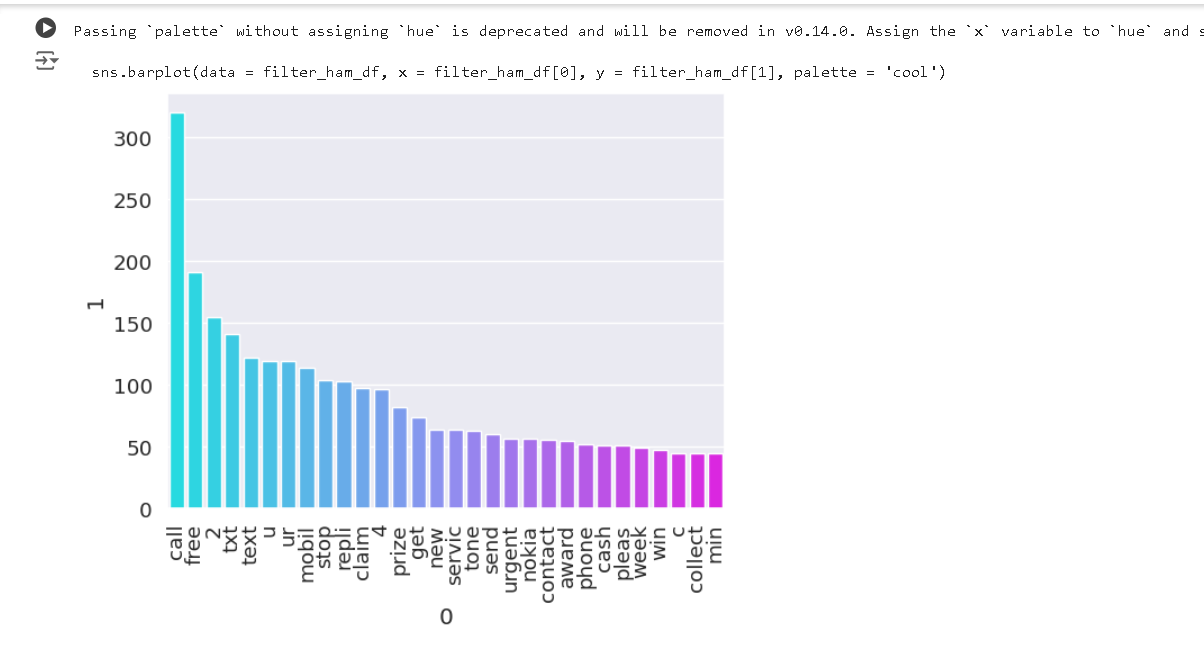


Hình 2. 23 Biểu đồ từ đám mây từ (word cloud) cho văn bản trong cột transformed\_text của DataFrame df đối với văn bản không spam

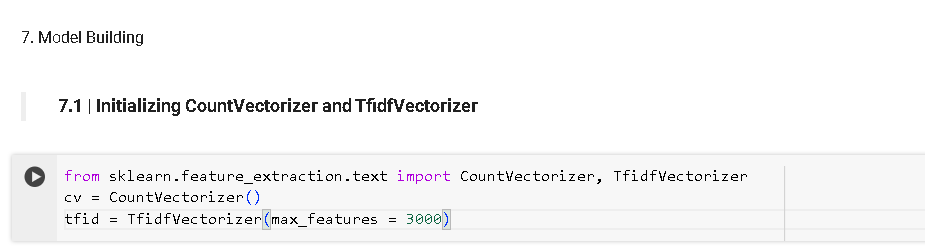
 

Hình 2. 24 Tìm kiếm top 30 từ loại spam

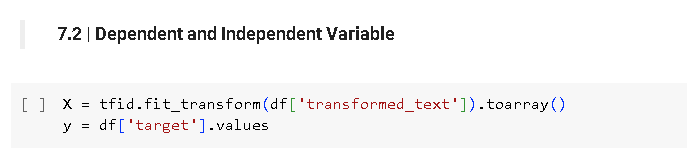




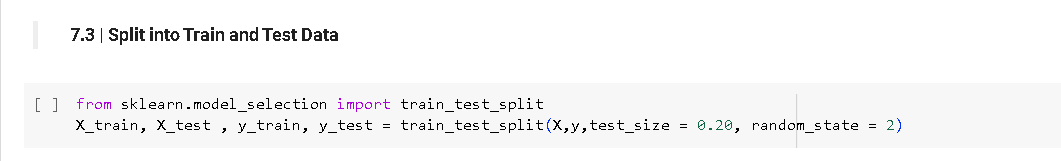
Hình 2. 25 Tìm kiếm top 30 từ loại không spam

2.6 Model Building

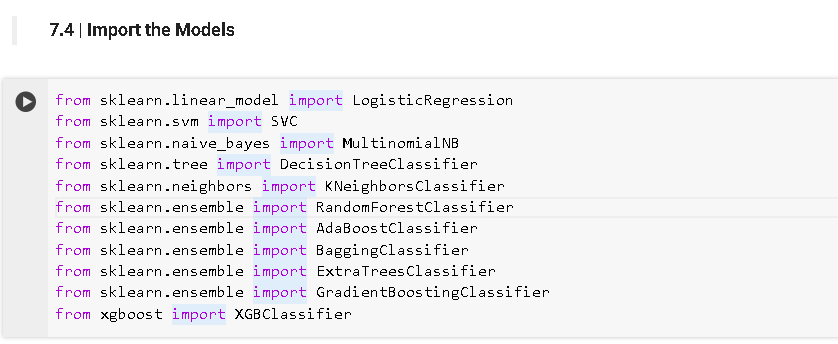
Hình 2. 26 Nhập Thư Viện và Khởi Tạo Vectorizer



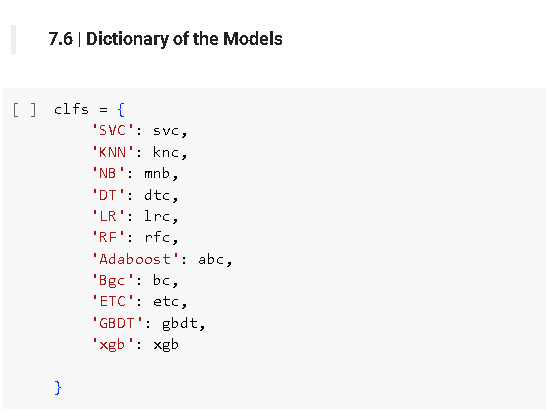
Hình 2. 27 Chuyển Đổi Dữ Liệu Văn Bản



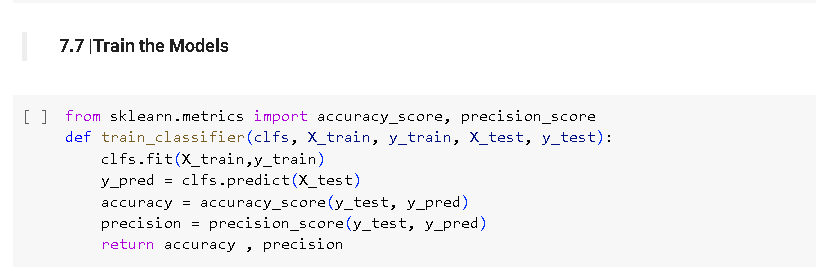
Hình 2. 28 Chia Dữ Liệu Thành Tập Huấn Luyện và Tập Kiểm Tra



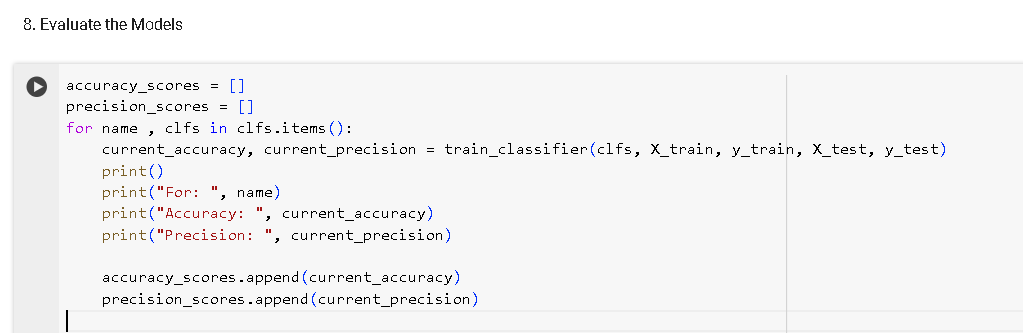
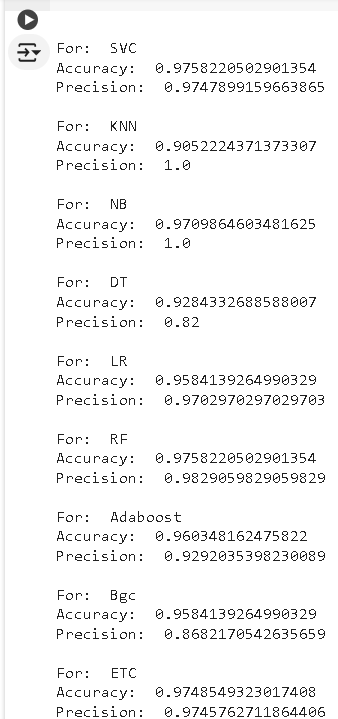
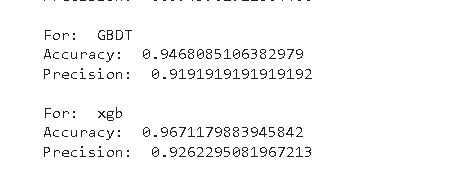
Hình 2. 29 Khởi Tạo Các Mô Hình Học Máy



Hình 2. 30 Tạo Từ Điển Các Mô Hình



Hình 2. 31 Định Nghĩa Hàm Huấn Luyện Mô Hình và Đánh Giá

Hình 2. 32 Huấn Luyện Các Mô Hình và Hiển Thị Kết Quả

**CHƯƠNG III : KẾT LUẬN**

3.1 Kết quả đạt được:

Trong nghiên cứu này, chúng tôi đã triển khai và đánh giá nhiều mô hình phân loại văn bản để phân loại tin nhắn thành các nhóm spam và không spam. Các mô hình phân loại bao gồm Support Vector Classifier (SVC), K-Nearest Neighbors (KNN), Naive Bayes (NB), Decision Tree (DT), Logistic Regression (LR), Random Forest (RF), AdaBoost, Bagging Classifier (Bgc), Extra Trees Classifier (ETC), Gradient Boosting Classifier (GBDT), và XGBoost (xgb).

**3.2 Hiệu suất mô hình :**

| **Mô Hình** | **Độ Chính Xác (Accuracy)** | **Độ Chính Xác (Precision)** |
| --- | --- | --- |
| **SVC** | 97.58% | 97.48% |
| **KNN** | 90.52% | 100.00% |
| **Naive Bayes** | 97.10% | 100.00% |
| **Decision Tree** | 92.84% | 82.00% |
| **Logistic Regression** | 95.84% | 97.03% |
| **Random Forest** | 97.58% | 98.29% |
| **AdaBoost** | 96.03% | 92.92% |
| **Bagging Classifier** | 95.84% | 86.82% |
| **Extra Trees Classifier** | 97.49% | 97.46% |
| **Gradient Boosting Classifier** | 94.68% | 91.92% |
| **XGBoost** | 96.71% | 92.62% |

**Nhận xét**:

* **XGBoost**: Có hiệu suất cao với độ chính xác 96.71% và độ chính xác (precision) 92.62%, cho thấy mô hình này rất hiệu quả trong việc phân loại tin nhắn spam.
* **Naive Bayes** và **KNN**: Đạt độ chính xác 100% về độ chính xác (precision), nhưng có độ chính xác tổng thể thấp hơn so với XGBoost.
* **Random Forest**: Cung cấp độ chính xác cao nhất về độ chính xác (precision) với 98.29%, cùng với độ chính xác tổng thể là 97.58%, cho thấy đây là một lựa chọn rất mạnh.
* **Decision Tree**: Mặc dù có độ chính xác tổng thể tốt (92.84%), nhưng độ chính xác (precision) thấp hơn, chỉ đạt 82.00%.
* **Logistic Regression**: Có độ chính xác tốt với 95.84% và độ chính xác (precision) 97.03%, cho thấy đây là một mô hình cân bằng giữa độ chính xác và độ chính xác.

**Kết Luận:**

Các mô hình đã được thử nghiệm cho thấy hiệu suất phân loại tin nhắn spam và không spam rất tốt. XGBoost và Random Forest nổi bật với hiệu suất cao về cả độ chính xác và độ chính xác (precision), trong khi Naive Bayes và KNN cho kết quả xuất sắc về độ chính xác (precision) nhưng có độ chính xác tổng thể thấp hơn. Quyết định lựa chọn mô hình cuối cùng sẽ phụ thuộc vào yêu cầu cụ thể của ứng dụng, bao gồm cả cân nhắc về tốc độ tính toán và hiệu quả thực tế trong môi trường triển khai.

**3.3 Gợi ý sản phẩm cho khách hàng:**

- **Nâng cao Độ Chính Xác**: Sử dụng mô hình XGBoost hoặc Random Forest, các mô hình đã cho kết quả tốt nhất. Có thể cân nhắc áp dụng kỹ thuật tinh chỉnh hyperparameters để tối ưu hóa thêm.

- **Tăng Cường Dữ Liệu**: Thu thập thêm dữ liệu để cải thiện khả năng tổng quát của mô hình. Dữ liệu phong phú hơn có thể giúp các mô hình học tốt hơn và giảm nguy cơ overfitting.

- **Tích Hợp Các Kỹ Thuật Tiền Xử Lý**: Cải thiện các bước tiền xử lý dữ liệu như loại bỏ từ không cần thiết, xử lý ngữ nghĩa sâu hơn để tăng chất lượng dữ liệu đầu vào.

- **Ứng Dụng Real-Time**: Xây dựng hệ thống phân loại tin nhắn thời gian thực, sử dụng các mô hình hiệu suất cao như XGBoost hoặc Random Forest, để phát hiện và lọc spam ngay lập tức.

**3.4 Hạn chế :**

- **Tập Dữ Liệu Hạn Chế**: Nếu dữ liệu không đủ phong phú hoặc không đại diện, mô hình có thể không hoạt động tốt trên các dữ liệu ngoài tập huấn luyện. Việc thu thập dữ liệu bổ sung là cần thiết.

- **Chất Lượng Dữ Liệu Tiền Xử Lý**: Các bước tiền xử lý như loại bỏ stopwords, stemming có thể làm mất thông tin quan trọng trong văn bản. Cần cân nhắc giữa việc giảm số lượng từ và giữ lại thông tin quan trọng.

- **Tính Tổng Quát**: Mặc dù mô hình hoạt động tốt trên tập kiểm tra hiện tại, hiệu suất có thể giảm khi đối mặt với các loại tin nhắn spam chưa thấy trong tập dữ liệu huấn luyện. Cần tiếp tục đánh giá và cập nhật mô hình để cải thiện tính tổng quát.

- **Thời Gian Tính Toán**: Một số mô hình như XGBoost và Gradient Boosting có thể yêu cầu thời gian tính toán dài hơn, đặc biệt khi làm việc với tập dữ liệu lớn hoặc khi triển khai trong môi trường thực tế.

**TÀI LIỆU THAM KHẢO**

1.Dữ liệu : <https://www.kaggle.com/code/zabihullah18/email-spam-detection/input?select=spam.csv>