|  |  |
| --- | --- |
| BỘ CÔNG THƯƠNG  **TRƯỜNG ĐẠI HỌC KINH TẾ**  **KỸ THUẬT CÔNG NGHIỆP** | **KHOA KHOA HỌC ỨNG DỤNG** |

**BÁO CÁO TỔNG KẾT**

**HỌC PHẦN: ĐỒ ÁN 2**

**DỰ ĐOÁN TÌNH TRẠNG TẮC NGHẼN GIAO THÔNG TẠI SEATLE, DỰA TRÊN DỮ LIỆU THỜI TIẾT, VA CHẠM VÀ MỘT SỐ SỰ KIỆN KHÁC**

**Sinh viên thực hiện:**

|  |  |  |
| --- | --- | --- |
| PHẠM VĂN THẮNG | DHKL16A2HN | 22174600100 |
| LÊ TUẤN THÀNH | DHKL16A2HN | 22174600119 |

**Giáo viên giảng dạy:** LÊ HẰNG ANH

**Hà Nội, 05/205**

|  |  |
| --- | --- |
| BỘ CÔNG THƯƠNG  **TRƯỜNG ĐẠI HỌC KINH TẾ**  **KỸ THUẬT CÔNG NGHIỆP** | **KHOA KHOA HỌC ỨNG DỤNG** |

**BÁO CÁO TỔNG KẾT**

**HỌC PHẦN: ĐỒ ÁN 2**

**DỰ ĐOÁN TÌNH TRẠNG TẮC NGHẼN GIAO THÔNG TẠI SEATLE, DỰA TRÊN DỮ LIỆU THỜI TIẾT, VA CHẠM VÀ MỘT SỐ SỰ KIỆN KHÁC**

**Sinh viên thực hiện:**

|  |  |  |
| --- | --- | --- |
| PHẠM VĂN THẮNG | DHKL16A2HN | 22174600100 |
| LÊ TUẤN THÀNH | DHKL16A2HN | 22174600119 |

**Giáo viên giảng dạy:** LÊ HẰNG ANH

**Hà Nội, 05/2025**

**PHIẾU ĐĂNG KÝ ĐỀ TÀI**

|  |
| --- |
| **1. Tên đề tài:** Dự đoán tình trạng tắc nghẽn giao thông tại Seatle, dựa trên dữ liệu thời gian, va chạm và một số sự kiện khác |
| **2. Thông tin nhóm sinh viên:**  **Sinh viên 1 (Nhóm trưởng):**   * **Họ và tên:** Lê Tuấn Thành * **Mã sinh viên:** 22174600119 * **Điện thoại:** 0936627165 * **Email:** letuanthanh1662002.tk@gmail.com   **Sinh viên 2:**   * **Họ và tên:** Phạm Văn Thắng * **Mã sinh viên:** 22174600100 * **Điện thoại:** 0364194171 * **Email:** Thangpv.291tb@gmail.com |
| **3. Tóm tắt nội dung đề tài:**  Đề tài tập trung vào việc phân tích các yếu tố ảnh hưởng đến tỷ lệ tai nạn giao thông tại thành phố Seattle, từ đó đưa ra các kết luận quan trọng về những yếu tố có tác động mạnh đến sự xuất hiện và mức độ nghiêm trọng của các vụ tai nạn. Dữ liệu sử dụng trong nghiên cứu này bao gồm thông tin chi tiết về các vụ va chạm, bao gồm vị trí, thời gian, tình trạng giao thông, điều kiện thời tiết, ánh sáng, và các yếu tố khác như việc lái xe có say rượu hay không, tốc độ khi xảy ra tai nạn.  Phân tích dữ liệu này sẽ giúp xác định những khu vực có nguy cơ cao xảy ra tai nạn, những điều kiện thời tiết hay ánh sáng dễ dẫn đến va chạm, và các yếu tố khác như tốc độ và sự chú ý của người tham gia giao thông. Qua đó, nghiên cứu có thể đề xuất các biện pháp cải thiện an toàn giao thông, từ việc điều chỉnh cơ sở hạ tầng đến việc cải thiện các chiến dịch tuyên truyền, nhằm giảm thiểu các tai nạn giao thông và bảo vệ tính mạng, sức khỏe của người dân tại Seattle.  Nghiên cứu này cũng đóng góp vào việc xây dựng các mô hình dự đoán tai nạn giao thông, hỗ trợ công tác quản lý và cải thiện tình hình an toàn giao thông trong tương lai. |

|  |  |
| --- | --- |
|  | *Ngày 8 tháng 4 năm 2025* |
|  | **Nhóm trưởng** |

**Lê Tuấn Thành**

ĐỀ CƯƠNG CHI TIẾT ĐỀ TÀI

|  |
| --- |
| **1. Tên đề tài:** Dự đoán tình trạng tắc nghẽn giao thông tại Seatle, dựa trên dữ liệu thời gian, va chạm và một số sự kiện khác |
| **2. Mục tiêu đề tài:**  Phân tích các yếu tố ảnh hưởng đến tai nạn giao thông tại Seattle nhằm hiểu rõ hơn về những yếu tố nguy cơ và mối liên hệ giữa các yếu tố này với tần suất và mức độ nghiêm trọng của tai nạn. Cụ thể, nghiên cứu sẽ tập trung vào các yếu tố như điều kiện thời tiết, ánh sáng, tình trạng đường xá, tốc độ di chuyển, sự chú ý của người lái và các yếu tố khác có liên quan đến hành vi và phương tiện tham gia giao thông.  Một trong những mục tiêu quan trọng của đề tài là xây dựng một mô hình phân tích để đánh giá mức độ tác động của các yếu tố này đối với tai nạn giao thông. Bằng cách sử dụng dữ liệu về các vụ tai nạn giao thông từ SDOT, nghiên cứu sẽ thực hiện phân tích hồi quy và các phương pháp thống kê khác để tìm ra các yếu tố có ảnh hưởng lớn nhất đến sự xuất hiện và mức độ nghiêm trọng của tai nạn.  Ngoài ra, đề tài cũng hướng tới việc đưa ra các giải pháp dựa trên kết quả phân tích, giúp các cơ quan chức năng đưa ra những biện pháp quản lý và cải thiện an toàn giao thông tại Seattle, như điều chỉnh hạ tầng giao thông, cải thiện các chiến dịch tuyên truyền, và đưa ra các khuyến nghị về việc điều chỉnh hành vi của người tham gia giao thông. |
| **3. Tổng quan tình hình nghiên cứu thuộc lĩnh vực đề tài**  Tai nạn giao thông tại Seattle đã thu hút sự quan tâm đáng kể từ các cơ quan chức năng và các tổ chức nghiên cứu, nhằm giảm thiểu thiệt hại và nâng cao an toàn cho người dân. Thành phố đã thực hiện nhiều dự án nghiên cứu và cải tiến hạ tầng giao thông để giảm số vụ tai nạn.  Một trong những dự án quan trọng là Dự án NE 130th & NE 125th Mobility and Safety Project, nhằm cải thiện an toàn cho người đi bộ, người đi xe đạp và phương tiện công cộng, đồng thời tăng cường kết nối giao thông trong khu vực. Các cải tiến bao gồm việc lắp đặt làn đường xe đạp bảo vệ và các điểm dừng xe buýt mới. Tương tự, Dự án Mở Rộng Tuyến West Seattle Link cũng nhắm đến việc cải thiện giao thông công cộng và giảm tắc nghẽn, với mục tiêu cải thiện kết nối giao thông nhanh chóng và đáng tin cậy giữa các khu vực của Seattle. Ngoài ra, East Link Connections là dự án liên kết mạng lưới giao thông công cộng tại khu vực Eastside, nhằm cải thiện sự phối hợp giữa các phương tiện công cộng và dịch vụ tàu điện nhẹ.  Bên cạnh các dự án hạ tầng, nhiều nghiên cứu đã chỉ ra rằng việc giảm tốc độ trên các con đường có thể giảm thiểu đáng kể tai nạn giao thông. Các biện pháp như giảm giới hạn tốc độ từ 30 mph xuống 25 mph và tăng cường hạ tầng cho người đi bộ, xe đạp đã được thực hiện tại Seattle để nâng cao an toàn. Các cải tiến này đã chứng minh hiệu quả trong việc giảm thiểu tai nạn và giúp bảo vệ người tham gia giao thông.  Từ những nghiên cứu và dự án trên, có thể thấy rằng việc tiếp tục đầu tư vào cải thiện hạ tầng và các chiến lược an toàn giao thông là cần thiết để giảm thiểu tai nạn và nâng cao chất lượng sống cho cộng đồng tại Seattle. |
| **4. Nội dung đề tài**  Đề tài tập trung vào việc xây dựng một mô hình học máy nhằm dự đoán khả năng xảy ra tắc nghẽn giao thông tại thành phố Seattle (Hoa Kỳ), dựa trên dữ liệu lịch sử các vụ va chạm giao thông kết hợp với điều kiện thời tiết, thời gian và các yếu tố khác như ánh sáng, vị trí và loại tai nạn. Việc dự đoán sớm các tình huống tắc nghẽn có thể giúp cơ quan quản lý giao thông đưa ra các giải pháp ứng phó kịp thời, từ đó giảm thiểu thời gian trễ, tai nạn và ô nhiễm.  Dữ liệu sử dụng trong đề tài là bộ dữ liệu Seattle Traffic Collision Data, được cung cấp bởi Sở Giao thông Seattle (Seattle Department of Transportation – SDOT), bao gồm hơn 200.000 bản ghi từ năm 2004 đến 2020. Bộ dữ liệu chứa thông tin chi tiết về mỗi vụ tai nạn như vị trí xảy ra (tọa độ GPS, địa chỉ), thời gian (ngày, giờ), loại va chạm, điều kiện thời tiết, tình trạng mặt đường và mức độ nghiêm trọng. Dữ liệu này rất phù hợp để huấn luyện mô hình dự đoán và khai thác các yếu tố ảnh hưởng đến tình trạng giao thông tại thành phố.  Đề tài sẽ tiến hành các bước tiền xử lý dữ liệu như làm sạch dữ liệu thiếu hoặc sai định dạng, chuẩn hóa thông tin thời gian, mã hóa các biến dạng phân loại (như loại tai nạn, thời tiết, ánh sáng) và tạo các biến mới như giờ trong ngày, ngày trong tuần hoặc phân loại mùa trong năm. Bên cạnh đó, dữ liệu cũng sẽ được gắn nhãn nhị phân nhằm phân biệt giữa các tình huống có và không xảy ra tắc nghẽn, dựa trên mức độ nghiêm trọng và thời điểm xảy ra tai nạn.  Tiếp theo, nhóm sẽ tiến hành lựa chọn và huấn luyện các mô hình học máy như Logistic Regression, Decision Tree, Random Forest hoặc XGBoost để thực hiện bài toán phân loại nhị phân. Các mô hình sẽ được đánh giá bằng các chỉ số như độ chính xác (accuracy), độ nhạy (recall), độ đặc hiệu (precision) và chỉ số F1-score nhằm đảm bảo tính hiệu quả và thực tế trong dự báo. Dữ liệu sẽ được chia thành tập huấn luyện và kiểm tra theo tỷ lệ 80:20 hoặc 70:30.  Ngoài phần mô hình, đề tài cũng chú trọng đến trực quan hóa dữ liệu để hỗ trợ phân tích và trình bày kết quả. Các biểu đồ phân bố thời gian, biểu đồ tỷ lệ va chạm theo điều kiện thời tiết, heatmap vị trí xảy ra va chạm sẽ giúp người xem có cái nhìn tổng thể và dễ hiểu về tình trạng giao thông. Một số bản đồ tương tác có thể được tạo bằng thư viện Folium (Python).  Cuối cùng, dựa trên kết quả phân tích và dự đoán, nhóm sẽ đề xuất một số giải pháp khả thi giúp giảm thiểu ùn tắc như: ưu tiên xử lý nhanh các điểm đen giao thông, điều chỉnh đèn tín hiệu tại các khung giờ cao điểm, phát triển hệ thống cảnh báo tắc đường thời gian thực, hoặc đề xuất ứng dụng phân luồng linh hoạt theo mùa và thời tiết.  Đề tài mang tính thực tiễn cao, có thể ứng dụng cho các thành phố lớn như TP.HCM hoặc Hà Nội nếu có dữ liệu phù hợp. Nó đồng thời giúp sinh viên rèn luyện kỹ năng phân tích dữ liệu, xây dựng mô hình học máy và trực quan hóa dữ liệu – những kỹ năng cốt lõi trong ngành khoa học dữ liệu hiện nay |
| **5. Phương pháp thực hiện**  Đầu tiên, nhóm tiến hành thu thập và tìm hiểu bộ dữ liệu Seattle Traffic Collision Data do Sở Giao thông Seattle cung cấp. Dữ liệu bao gồm thông tin chi tiết về hơn 200.000 vụ va chạm từ năm 2004 đến 2020, với các đặc trưng như thời gian, vị trí, điều kiện thời tiết, loại tai nạn, mức độ nghiêm trọng và tình trạng ánh sáng. Nhóm sẽ xác định các thuộc tính quan trọng để phục vụ cho bài toán dự đoán.  Tiếp theo, nhóm tiến hành tiền xử lý dữ liệu, bao gồm làm sạch dữ liệu thiếu, chuẩn hóa thời gian, mã hóa các biến dạng chuỗi, và tạo thêm các đặc trưng mới như: giờ, thứ, tháng, mùa, có phải giờ cao điểm hay không,... Một biến mục tiêu nhị phân (is\_congestion) cũng được tạo ra để phân loại khả năng xảy ra tắc nghẽn.  Sau đó, nhóm thực hiện phân tích dữ liệu khám phá (EDA) để hiểu rõ hơn mối quan hệ giữa các yếu tố và tình trạng tắc nghẽn. Các biểu đồ, bảng thống kê và bản đồ nhiệt sẽ được sử dụng để trực quan hóa.  Với dữ liệu đã xử lý, nhóm sẽ huấn luyện mô hình học máy như Logistic Regression, Decision Tree, Random Forest và XGBoost nhằm dự đoán khả năng tắc nghẽn. Các mô hình được đánh giá bằng các chỉ số như Accuracy, Precision, Recall và F1-score.  Cuối cùng, nhóm sẽ trình bày kết quả bằng bản đồ, biểu đồ và đề xuất giải pháp phù hợp để cải thiện tình trạng giao thông tại Seattle. |
| **6. Phân công công việc:**   |  |  |  |  | | --- | --- | --- | --- | | **STT** | **Họ và tên** | **Mã sinh viên** | **Nội dung công việc được phân công** | | **1** | Phạm Văn Thắng | 22174600100 | -Thu thập dữ liệu  -Tiền xử lý dữ liệu  -Phân tích khám phá dữ liệu  -Huấn luyện mô hình và đánh giá mô hình  -Viết báo cáo  -slide thuyết trình | | **2** | Lê Tuấn Thành | 22174600119 | -Thu thập dữ liệu  -Phân tích khám phá dữ liệu  -Huấn luyện mô hình và đánh giá mô hình | |
| 7. Dự kiến kết quả đạt được:  Sau khi hoàn thành đề tài, nhóm kỳ vọng đạt được các kết quả cụ thể như sau:  Thứ nhất, nhóm sẽ xây dựng được một mô hình học máy có khả năng dự đoán tình trạng tắc nghẽn giao thông tại thành phố Seattle dựa trên các yếu tố đầu vào như thời gian, điều kiện thời tiết, loại tai nạn, vị trí xảy ra va chạm,... Mô hình được đánh giá bằng các chỉ số Accuracy, Precision, Recall và F1-score, đảm bảo có độ chính xác đủ cao để áp dụng thực tế.  Thứ hai, nhóm sẽ tạo ra các bản đồ trực quan về “điểm nóng” giao thông, từ đó xác định các khu vực dễ xảy ra tắc nghẽn theo từng khung giờ trong ngày hoặc từng mùa trong năm. Các biểu đồ phân tích cũng sẽ giúp làm rõ mối liên hệ giữa tắc nghẽn và các yếu tố ảnh hưởng như thời tiết xấu, mặt đường ướt, hoặc điều kiện ánh sáng kém.  Thứ ba, từ việc phân tích dữ liệu và kết quả dự đoán, nhóm sẽ đưa ra một số giải pháp khả thi nhằm giảm thiểu tình trạng ùn tắc, chẳng hạn như:  Điều chỉnh thời gian đèn tín hiệu giao thông tại các giao lộ thường xuyên tắc nghẽn.  Tăng cường cảnh báo sớm trên các nền tảng di động khi có tai nạn xảy ra tại điểm nóng.  Phân bổ lực lượng cảnh sát giao thông hợp lý theo thời gian và địa điểm.  Ngoài ra, nhóm cũng sẽ đánh giá mức độ quan trọng của từng yếu tố đầu vào trong mô hình, từ đó giúp các nhà quản lý hiểu được yếu tố nào ảnh hưởng mạnh nhất đến khả năng gây tắc nghẽn, như thời điểm, điều kiện thời tiết hay loại va chạm. |

|  |  |
| --- | --- |
|  | *Ngày 8 tháng 4 năm 2025* |
|  | **Nhóm trưởng** |

**Lê Tuấn Thành**

**LỜI MỞ ĐẦU**

Trong bối cảnh đô thị hóa ngày càng gia tăng, vấn đề ùn tắc giao thông đã và đang trở thành một thách thức lớn đối với nhiều thành phố lớn trên thế giới, trong đó có thành phố Seattle – một trong những trung tâm kinh tế và công nghệ quan trọng của Hoa Kỳ. Tắc nghẽn giao thông không chỉ gây ảnh hưởng nghiêm trọng đến năng suất lao động và chất lượng sống của người dân, mà còn tác động tiêu cực đến môi trường và hệ thống hạ tầng đô thị.

Nhận thức được tầm quan trọng của việc quản lý và giảm thiểu ùn tắc giao thông, đề tài “Dự đoán tình trạng tắc nghẽn giao thông tại Seattle dựa trên dữ liệu thời tiết, va chạm và một số sự kiện khác” được thực hiện nhằm nghiên cứu và ứng dụng các phương pháp học máy (machine learning) để phân tích, dự đoán tình trạng giao thông trong thời gian thực. Việc tích hợp các yếu tố như thời tiết, tai nạn giao thông, sự kiện đặc biệt (lễ hội, thể thao, công trình...) vào mô hình dự đoán không chỉ giúp cải thiện độ chính xác, mà còn mở ra cơ hội cho các hệ thống điều phối giao thông thông minh trong tương lai.

Thông qua việc khai thác nguồn dữ liệu thực tế và áp dụng các kỹ thuật phân tích dữ liệu hiện đại, đồ án hướng đến việc xây dựng một mô hình dự đoán có tính ứng dụng cao, góp phần hỗ trợ các nhà quản lý đô thị và người tham gia giao thông đưa ra quyết định hiệu quả hơn.

Cấu trúc đồ án bao gồm các phần được phân chương như sau:

CHƯƠNG 1: Giới thiệu đề tài

CHƯƠNG 2: Cơ sở lý thuyết và tổng quan vấn đề

CHƯƠNG 3: Thu thập và xử lí dữ liệu

CHƯƠNG 4: Xây dựng và đánh giá mô hình

CHƯƠNG 5 : Kết luận và hướng phát triển

**MỤC LỤC**

[DANH MỤC HÌNH ẢNH 11](#_Toc197779649)

[CHƯƠNG 1: GIỚI THIỆU ĐỀ TÀI 13](#_Toc197779650)

[1.1. LÝ DO CHỌN ĐỀ TÀI 13](#_Toc197779651)

[1.2. MỤC TIÊU CỦA ĐỀ TÀI 13](#_Toc197779652)

[1.3. ĐỐI TƯỢNG VÀ PHẠM VI NGHIÊN CỨU 13](#_Toc197779653)

[1.4. PHƯƠNG PHÁP THỰC HIỆN 14](#_Toc197779654)

[1.5. CẤU TRÚC CỦA BÁO CÁO 15](#_Toc197779655)

[CHƯƠNG 2: CƠ SỞ LÝ THUYẾT VÀ TỔNG QUAN VẤN ĐỀ 16](#_Toc197779656)

[2.1. PHÂN TÍCH DỮ LIỆU VÀ VAI TRÒ TRONG HOẠT ĐỘNG GIAO THÔNG ĐÔ THỊ 16](#_Toc197779657)

[2.1.1. Khái niệm về phân tích dữ liệu 16](#_Toc197779658)

[2.1.2. Các phương pháp phân tích dữ liệu 17](#_Toc197779659)

[2.1.3. Ứng dụng trong giao thông 18](#_Toc197779660)

[2.2. GIAO THÔNG TẠI SEATTLE VÀ THỰC TRẠNG DỰ ÁN LIÊN QUAN 19](#_Toc197779661)

[2.3. MỘT SỐ KHÁI NIỆM LIÊN QUAN 20](#_Toc197779662)

[2.4. CÁC CÔNG CỤ VÀ THƯ VIỆN SỬ DỤNG 21](#_Toc197779663)

[CHƯƠNG 3: THU THẬP VÀ XỬ LÝ DỮ LIỆU 23](#_Toc197779664)

[3.1. MÔ TẢ VỀ BÀI TOÁN 23](#_Toc197779665)

[3.1.1. Giới thiệu về tập dữ liệu 23](#_Toc197779666)

[3.1.2. Tổng quan các trường của dữ liệu 24](#_Toc197779667)

[3.2. TIỀN XỬ LÝ DỮ LIỆU 25](#_Toc197779668)

[3.2.1. Xử lý dữ liệu còn thiếu 25](#_Toc197779669)

[3.2.2. Xử lý dữ liệu thời gian 26](#_Toc197779670)

[3.2.3. Phân tích và chuẩn hoá biến mục tiêu 27](#_Toc197779671)

[3.3. PHÂN TÍCH VÀ KHÁM PHÁ DỮ LIỆU 28](#_Toc197779672)

[3.3.1. Biến mục tiêu và phân tích mức độ nghiêm trọng 28](#_Toc197779673)

[3.3.2. Phân tích mức độ nghiêm trọng theo yếu tố thời gian 33](#_Toc197779674)

[3.3.3. Ma trận tương quan (Correlation Matrix) 37](#_Toc197779675)

[CHƯƠNG 4: XÂY DỰNG VÀ ĐÁNH GIÁ MÔ HÌNH 40](#_Toc197779676)

[4.1. Giới thiệu về mô hình và phương pháp sử dụng 40](#_Toc197779677)

[4.2. HUẤN LUYỆN CÁC MÔ HÌNH 40](#_Toc197779678)

[4.2.1. MÔ HÌNH LOGISTIC REGRESSION 40](#_Toc197779679)

[4.2.2. MÔ HÌNH SUPPORT VECTO MACHINE(SVM) 41](#_Toc197779680)

[4.2.3. MÔ HÌNH RANDOM FOREST 42](#_Toc197779681)

[4.3. ĐÁNH GIÁ MÔ HÌNH 43](#_Toc197779682)

[4.3.1. ĐÁNH GIÁ LOGISTIC REGRESSION 43](#_Toc197779683)

[4.3.2. ĐÁNH GIÁ SUPPORT VECTO MACHINE(SVM) 44](#_Toc197779684)

[4.3.2. ĐÁNH GIÁ RANDOM FOREST 45](#_Toc197779685)

[CHƯƠNG 5: KẾT LUẬN VÀ HƯỚNG PHÁT TRIỂN 47](#_Toc197779686)

[5.1. KẾT LUẬN 47](#_Toc197779687)

[5.2. HƯỚNG PHÁT TRIỂN TRONG TƯƠNG LAI 48](#_Toc197779688)

[TÀI LIỆU THAM KHẢO 50](#_Toc197779689)

# DANH MỤC HÌNH ẢNH

[Hình 3.1. Biểu đồ phân phối mức độ nghiêm trong SEVERITYCODE 31](#_Toc197777335)

[Hình 3.2. Biều đồ mức độ theo nhóm thời tiết (WEATHER) 32](#_Toc197777336)

[Hình 3.3. Biều đồ mức độ nghiêm trọng theo điều kiện mặt đường (ROADCOND) 33](#_Toc197777337)

[Hình 3.3. Biều đồ mức độ nghiêm trọng theo nhóm điều kiện ánh sáng 35](#_Toc197777338)

[Hình 3.4. Biều đồ mức độ nghiêm trọng theo giờ trong ngày(Hour) 36](#_Toc197777339)

[Hình 3.5. Biểu đồ mức độ nghiêm trọng theo ngày trong tuần(DayOfWeek) 37](#_Toc197777340)

[Hình 3.6. Biểu đồ số vụ va chạm theo tháng 38](#_Toc197777341)

[Hình 3.7. Biểu đồ ma trận tương quan giữa các biến 40](#_Toc197777342)

[Hình 4.1. Báo cáo phân loại và Ma trận nhầm lẫn của Logistic Regression**.** 43](#_Toc197777343)

[Hình 4.2. Báo cáo phân loại và Ma trận nhầm lẫn của SUPPORT VECTO MACHINE 44](#_Toc197777344)

[Hình 4.3. Báo cáo phân loại và Ma trận nhầm lẫn của SUPPORT VECTO MACHINE 45](#_Toc197777345)

[Hình 4.4. Logistic Regression - Confusion Matrix**.** 46](#_Toc197777346)

[Hình 4.5. SUPPORT VECTO MACHINE - Confusion Matrix**.** 47](#_Toc197777347)

[Hình 4.6. RANDOM FOREST - Confusion Matrix**.** 47](#_Toc197777348)

# CHƯƠNG 1: GIỚI THIỆU ĐỀ TÀI

## 1.1. LÝ DO CHỌN ĐỀ TÀI

Tắc nghẽn giao thông đang ngày càng trở thành một vấn đề nghiêm trọng tại các đô thị lớn trên thế giới, đặc biệt là ở các trung tâm kinh tế và công nghệ như thành phố Seattle (Hoa Kỳ). Với mật độ dân số cao, hệ thống đường giao thông phức tạp và các yếu tố ngoại cảnh khó lường như điều kiện thời tiết, tai nạn, và sự kiện cộng đồng, việc dự đoán trước các tình huống ùn tắc là vô cùng cần thiết nhằm hỗ trợ người tham gia giao thông cũng như các nhà quản lý đô thị đưa ra quyết định kịp thời. Trong bối cảnh dữ liệu lớn (Big Data) và trí tuệ nhân tạo (AI) đang được ứng dụng rộng rãi trong nhiều lĩnh vực, việc xây dựng mô hình dự đoán tắc nghẽn giao thông dựa trên nhiều yếu tố như thời tiết, tai nạn và các sự kiện đặc biệt là một hướng đi phù hợp và có tiềm năng ứng dụng thực tế cao. Chính vì lý do đó, nhóm chúng chọn thực hiện đề tài này nhằm khám phá khả năng dự đoán tình trạng giao thông dựa trên các yếu tố tác động thực tế và đachiều.

## 1.2. MỤC TIÊU CỦA ĐỀ TÀI

Mục tiêu của nghiên cứu này là xây dựng một mô hình học máy có khả năng dự đoán tình trạng tắc nghẽn giao thông tại thành phố Seattle. Để nâng cao độ chính xác của mô hình, dữ liệu sẽ được tích hợp và xử lý từ nhiều nguồn khác nhau, bao gồm dữ liệu về thời tiết, tai nạn giao thông và các sự kiện đặc biệt. Sau khi xây dựng mô hình, hiệu suất của nó sẽ được đánh giá thông qua các chỉ số quan trọng như độ chính xác, độ nhạy và độ đặc hiệu, nhằm đảm bảo khả năng dự đoán chính xác và hiệu quả. Cuối cùng, nghiên cứu sẽ đề xuất một hướng tiếp cận khả thi cho việc áp dụng mô hình này vào các hệ thống hỗ trợ quản lý giao thông thông minh trong tương lai, góp phần nâng cao hiệu quả điều tiết và giảm thiểu các vấn đề về tắc nghẽn giao thông trong đô thị.

## 1.3. ĐỐI TƯỢNG VÀ PHẠM VI NGHIÊN CỨU

Đối tượng nghiên cứu của đề tài là hệ thống giao thông tại thành phố Seattle, bao gồm một loạt các dữ liệu liên quan đến các yếu tố tác động trực tiếp đến tình trạng giao thông, chẳng hạn như điều kiện giao thông, thời tiết, tai nạn và các sự kiện cộng đồng. Các yếu tố này có thể ảnh hưởng đến mức độ tắc nghẽn, thời gian di chuyển và sự an toàn của người tham gia giao thông. Cụ thể, dữ liệu sẽ bao gồm các thông tin về mức độ tắc nghẽn giao thông trên các tuyến đường, các vụ tai nạn giao thông, các sự kiện đặc biệt như lễ hội, hội chợ hoặc các sự kiện thể thao có thể làm thay đổi lưu lượng giao thông, và các điều kiện thời tiết như mưa, tuyết hoặc sương mù, tất cả đều đóng vai trò quan trọng trong việc làm tăng hoặc giảm tắc nghẽn.

Phạm vi nghiên cứu của đề tài sẽ được giới hạn trong khu vực nội thành Seattle, nơi có mật độ dân cư cao và giao thông thường xuyên gặp phải các vấn đề tắc nghẽn, đặc biệt là vào giờ cao điểm. Dữ liệu sẽ được thu thập trong một khoảng thời gian xác định, giúp nghiên cứu nắm bắt được các xu hướng và đặc điểm giao thông theo thời gian. Đề tài tập trung vào việc phân tích dữ liệu lịch sử từ các nguồn đã có sẵn, nhằm xác định các mô hình và mối quan hệ giữa các yếu tố ảnh hưởng đến giao thông. Một mục tiêu quan trọng của nghiên cứu là xây dựng mô hình dự đoán tình trạng giao thông cho các khung thời gian ngắn hạn, như dự đoán tình trạng giao thông theo ngày hoặc theo giờ, giúp các cơ quan quản lý giao thông có thể đưa ra các biện pháp điều tiết và dự báo tình huống giao thông hiệu quả hơn.

## 1.4. PHƯƠNG PHÁP THỰC HIỆN

Để giải quyết bài toán dự đoán tình trạng tắc nghẽn giao thông tại thành phố Seattle dựa trên dữ liệu thời tiết, va chạm và một số sự kiện khác, nhóm chúng em đã thực hiện quy trình nghiên cứu bao gồm các bước chính như sau. Đầu tiên, chúng em tiến hành thu thập dữ liệu từ nhiều nguồn khác nhau, bao gồm dữ liệu thời tiết (như nhiệt độ, lượng mưa, và điều kiện thời tiết hiện tại), dữ liệu về tai nạn và va chạm giao thông, cùng với thông tin về thời gian và địa điểm (để xây dựng các đặc trưng liên quan đến thời gian, như giờ cao điểm và ngày trong tuần). Sau khi thu thập dữ liệu, nhóm đã thực hiện tiền xử lý dữ liệu bằng cách sử dụng các thư viện Pandas và NumPy để đọc, làm sạch và xử lý dữ liệu thô. Các giá trị bị thiếu được kiểm tra và xử lý, và dữ liệu được biến đổi thông qua các bước chuẩn hóa, mã hóa các biến phân loại, và xử lý định dạng thời gian. Chúng em cũng trích xuất các đặc trưng mới từ các cột ngày giờ hoặc địa lý để mô hình có thể hiểu rõ hơn về dữ liệu. Tiếp theo, trong bước phân tích dữ liệu, nhóm sử dụng các thư viện trực quan hóa như Matplotlib và Seaborn để khám phá mối quan hệ giữa các yếu tố như thời tiết, va chạm và mức độ tắc nghẽn giao thông. Qua đó, chúng em xác định các đặc trưng quan trọng có khả năng ảnh hưởng đến tình trạng giao thông. Sau đó, nhóm chuyển sang bước xây dựng mô hình dự đoán bằng cách sử dụng các thuật toán học máy phổ biến như Random Forest Classifier, Logistic Regression và Support Vector Machine (SVM). Tập dữ liệu được chia thành tập huấn luyện và tập kiểm thử để đánh giá hiệu suất của mô hình thông qua các chỉ số như Accuracy, Precision, Recall và F1-score. Để nâng cao hiệu quả, nhóm tiếp tục thực hiện đánh giá và tối ưu mô hình, so sánh hiệu suất giữa các mô hình và chọn ra mô hình tốt nhất. Siêu tham số của mô hình được tinh chỉnh để cải thiện độ chính xác, đồng thời kiểm tra độ ổn định và tính tổng quát của mô hình trên các tập dữ liệu khác nhau. Cuối cùng, kết quả của mô hình được trực quan hóa dưới dạng biểu đồ và bản đồ để dễ dàng đánh giá hiệu quả và ứng dụng thực tế, giúp các cơ quan quản lý giao thông có thể đưa ra quyết định dựa trên các dự đoán chính xác về tình trạng giao thông.

## 1.5. CẤU TRÚC CỦA BÁO CÁO

Bài báo cáo được chia thành 5 chương với các nội dung cụ thể như sau :

* **Chương 1: Giới thiệu đề tài:** Trình bày lý do chọn đề tài, mục tiêu, phạm vi, phương pháp và cấu trúc báo cáo.
* **Chương 2: Cơ sở lý thuyết và các nghiên cứu liên quan:** Tổng quan về các khái niệm chính như giao thông đô thị, tắc nghẽn, dữ liệu lớn, học máy và các công trình nghiên cứu trước.
* **Chương 3: Phân tích và thiết kế hệ thống:** Chi tiết về quá trình thu thập dữ liệu, xử lý dữ liệu, thiết kế mô hình và lựa chọn thuật toán phù hợp.
* **Chương 4: Thực nghiệm và đánh giá kết quả:** Mô tả quá trình huấn luyện mô hình, kết quả đạt được, đánh giá hiệu suất và phân tích các yếu tố ảnh hưởng.
* **Chương 5: Kết luận và hướng phát triển:** Tóm tắt kết quả đạt được, hạn chế của đề tài và đề xuất hướng nghiên cứu tiếp theo.

# CHƯƠNG 2: CƠ SỞ LÝ THUYẾT VÀ TỔNG QUAN VẤN ĐỀ

## 2.1. PHÂN TÍCH DỮ LIỆU VÀ VAI TRÒ TRONG HOẠT ĐỘNG GIAO THÔNG ĐÔ THỊ

### 2.1.1. Khái niệm về phân tích dữ liệu

Phân tích dữ liệu là một quá trình hệ thống bao gồm việc thu thập, làm sạch, chuyển đổi, và phân tích các dữ liệu từ nhiều nguồn khác nhau để rút ra những thông tin hữu ích phục vụ cho việc ra quyết định, dự đoán xu hướng và tối ưu hóa các chiến lược. Trong bối cảnh khoa học dữ liệu, phân tích dữ liệu giúp hiểu rõ hơn về các yếu tố ảnh hưởng đến các vấn đề cụ thể, từ đó đưa ra những giải pháp khả thi và hiệu quả hơn. Quá trình phân tích này không chỉ bao gồm các bước cơ bản như thu thập và làm sạch dữ liệu, mà còn sử dụng các phương pháp và công cụ phân tích phức tạp để trích xuất giá trị từ các dữ liệu khổng lồ.

Trong giao thông đô thị, phân tích dữ liệu đóng vai trò đặc biệt quan trọng trong việc đánh giá các yếu tố ảnh hưởng đến tình trạng giao thông, chẳng hạn như tai nạn, thời tiết, điều kiện mặt đường, mật độ phương tiện, và các sự kiện đặc biệt. Nhờ vào các công nghệ và phương pháp phân tích hiện đại, việc thu thập và xử lý các dữ liệu lớn từ các nguồn khác nhau (như cảm biến giao thông, dữ liệu thời tiết, báo cáo tai nạn, v.v.) trở nên khả thi hơn bao giờ hết. Những dữ liệu này không chỉ giúp hiểu rõ tình hình giao thông trong hiện tại mà còn có thể dự đoán những vấn đề tiềm ẩn trong tương lai.

Phân tích dữ liệu trong giao thông đô thị không chỉ là việc thu thập các thông tin đơn lẻ mà là một quá trình phức tạp để kết nối và xử lý thông tin từ nhiều nguồn, từ đó xây dựng các mô hình dự đoán và điều phối giao thông hiệu quả. Ví dụ, việc kết hợp dữ liệu về thời tiết xấu, tình trạng đường xá, và mật độ giao thông có thể giúp dự báo tình trạng tắc nghẽn giao thông trong những giờ cao điểm. Thêm vào đó, việc phân tích dữ liệu tai nạn giúp nhận diện các điểm "nóng" về an toàn giao thông, qua đó đưa ra các biện pháp giảm thiểu nguy cơ tai nạn hoặc cải thiện điều kiện hạ tầng giao thông.

Một ứng dụng quan trọng khác của phân tích dữ liệu trong giao thông là hỗ trợ việc điều phối và tối ưu hóa các hệ thống giao thông thông minh. Ví dụ, dữ liệu về mức độ tắc nghẽn giao thông trong các khu vực cụ thể có thể giúp các hệ thống tín hiệu giao thông điều chỉnh thời gian đèn tín hiệu một cách tự động và kịp thời, giảm thiểu tình trạng tắc nghẽn. Đồng thời, phân tích dữ liệu cũng có thể hỗ trợ việc lập kế hoạch cho các phương tiện giao thông công cộng, tối ưu hóa lộ trình và thời gian di chuyển của chúng.

Bên cạnh việc cải thiện hiệu quả vận hành giao thông, phân tích dữ liệu còn đóng góp vào việc giảm thiểu tai nạn giao thông. Các mô hình phân tích dữ liệu có thể phát hiện ra các yếu tố tiềm ẩn nguy cơ tai nạn, ví dụ như những khu vực có mật độ tai nạn cao, các yếu tố môi trường như thời tiết xấu hay tình trạng đường xá trơn trượt, từ đó đưa ra các cảnh báo sớm hoặc các biện pháp can thiệp kịp thời.

Như vậy, phân tích dữ liệu trong giao thông đô thị không chỉ mang lại những giải pháp giúp giảm thiểu tắc nghẽn mà còn đóng góp vào việc cải thiện an toàn giao thông, tối ưu hóa hệ thống vận hành, và nâng cao chất lượng cuộc sống cho người dân. Việc áp dụng phân tích dữ liệu vào giao thông sẽ giúp các cơ quan chức năng đưa ra các quyết định nhanh chóng, chính xác và hiệu quả hơn trong việc quản lý và điều phối giao thông, đồng thời góp phần xây dựng một thành phố thông minh, an toàn và bền vững hơn.

### 2.1.2. Các phương pháp phân tích dữ liệu

Trong nghiên cứu giao thông đô thị, việc áp dụng các phương pháp phân tích dữ liệu giúp các nhà quản lý hiểu rõ hơn về tình trạng giao thông và đưa ra các quyết định kịp thời, hiệu quả. Các phương pháp phân tích dữ liệu chủ yếu bao gồm:

Phân tích thống kê mô tả (Descriptive Statistics) là phương pháp cơ bản giúp tóm tắt các đặc điểm chính của dữ liệu. Phương pháp này sử dụng các chỉ số như trung bình, độ lệch chuẩn và phân phối tần suất để mô tả và nhận diện các mẫu trong dữ liệu. Ví dụ, trong giao thông đô thị, phân tích thống kê mô tả có thể giúp chúng ta tính toán trung bình số vụ tai nạn giao thông trong một khoảng thời gian nhất định, hoặc đánh giá sự biến động của mật độ giao thông trong các khung giờ khác nhau.

Phân tích hồi quy (Regression Analysis) là một phương pháp được sử dụng để xác định mối quan hệ giữa các biến số và dự đoán các giá trị trong tương lai. Phân tích hồi quy giúp đánh giá mức độ ảnh hưởng của các yếu tố như thời tiết, số lượng phương tiện, hoặc tình trạng mặt đường đến tình trạng giao thông. Ví dụ, chúng ta có thể sử dụng phân tích hồi quy để xác định mối quan hệ giữa thời tiết xấu và mức độ tắc nghẽn giao thông trong thành phố, từ đó đưa ra các biện pháp giảm thiểu tắc nghẽn khi có mưa hoặc tuyết.

Phân tích phân loại và phân nhóm (Classification & Clustering) sử dụng các thuật toán học máy để phân nhóm dữ liệu hoặc phân loại các sự kiện giao thông. Các thuật toán như cây quyết định (Decision Trees), rừng ngẫu nhiên (Random Forests), và K-means clustering có thể giúp phân loại mức độ nghiêm trọng của các vụ tai nạn giao thông, hoặc nhóm các khu vực có mức độ tắc nghẽn giao thông tương tự nhau. Phương pháp này đặc biệt hữu ích trong việc phân loại tình trạng giao thông (ví dụ: tắc nghẽn nhẹ, vừa, nặng) và từ đó điều phối giao thông hiệu quả hơn.

Phân tích chuỗi thời gian (Time Series Analysis) là phương pháp phân tích dữ liệu theo thời gian để dự đoán tình trạng giao thông trong tương lai. Phân tích chuỗi thời gian giúp dự báo các yếu tố như lưu lượng giao thông, mức độ tắc nghẽn, hoặc số vụ tai nạn trong các khoảng thời gian ngắn hạn. Điều này cực kỳ quan trọng trong việc quản lý giao thông đô thị, khi mà các tình trạng như tắc nghẽn có thể thay đổi theo thời gian trong ngày, theo tuần, hoặc theo các sự kiện đặc biệt. Phân tích chuỗi thời gian giúp các cơ quan chức năng đưa ra các biện pháp ứng phó trước khi vấn đề giao thông trở nên nghiêm trọng.

Tất cả những phương pháp phân tích này đều giúp cải thiện khả năng dự đoán và điều phối giao thông, đồng thời hỗ trợ các nhà quản lý đưa ra các quyết định chính xác và kịp thời để giảm thiểu tắc nghẽn và tăng cường an toàn giao thông.

### 2.1.3. Ứng dụng trong giao thông

Phân tích dữ liệu có rất nhiều ứng dụng quan trọng trong quản lý giao thông đô thị, giúp cải thiện hiệu quả vận hành và tăng cường an toàn giao thông. Một trong những ứng dụng chính là dự đoán tắc nghẽn giao thông. Bằng cách phân tích dữ liệu lịch sử về giao thông, tai nạn, điều kiện thời tiết và thời gian, các nhà quản lý có thể dự báo chính xác tình trạng tắc nghẽn trong tương lai. Điều này giúp đưa ra các biện pháp điều phối kịp thời, giảm thiểu tình trạng ùn tắc, đặc biệt trong những giờ cao điểm hoặc khi có sự kiện đặc biệt diễn ra.

Một ứng dụng khác là quản lý tai nạn giao thông. Dữ liệu về tai nạn giúp xác định các khu vực "nóng" trong thành phố, nơi thường xuyên xảy ra tai nạn và tắc nghẽn. Phân tích dữ liệu về tai nạn giúp nhận diện các nguyên nhân gây tai nạn, từ đó các cơ quan chức năng có thể triển khai các giải pháp phòng ngừa hiệu quả như cải thiện hạ tầng giao thông, lắp đặt các biển báo cảnh báo, hoặc thay đổi thiết kế giao lộ để giảm thiểu nguy cơ tai nạn và nâng cao an toàn cho người tham gia giao thông.

Cuối cùng, tối ưu hóa tín hiệu giao thông là một ứng dụng quan trọng khác của phân tích dữ liệu. Dựa vào dữ liệu thời gian thực về lưu lượng giao thông, tai nạn và các sự kiện đặc biệt, các hệ thống tín hiệu giao thông có thể được điều chỉnh tự động để giảm tắc nghẽn, đặc biệt trong các giờ cao điểm hoặc khi có sự kiện lớn. Hệ thống đèn tín hiệu có thể thay đổi chu kỳ đèn tùy theo mật độ giao thông, từ đó cải thiện lưu thông và giảm tình trạng ùn tắc, đồng thời đảm bảo an toàn cho người tham gia giao thông. Những ứng dụng này không chỉ giúp điều phối giao thông hiệu quả mà còn góp phần cải thiện chất lượng sống của người dân, giảm ô nhiễm và tăng tính bền vững cho các thành phố.

## 2.2. GIAO THÔNG TẠI SEATTLE VÀ THỰC TRẠNG DỰ ÁN LIÊN QUAN

Seattle là một trong những thành phố lớn và phát triển mạnh mẽ tại Mỹ, với mật độ dân số cao và hệ thống giao thông phức tạp. Thành phố này đối mặt với nhiều vấn đề giao thông nghiêm trọng, đặc biệt vào các giờ cao điểm và khi có thời tiết xấu hoặc sự kiện lớn. Tắc nghẽn giao thông là một vấn đề phổ biến, ảnh hưởng đến hiệu quả vận hành của thành phố và chất lượng cuộc sống của người dân. Các yếu tố như thời gian, địa điểm, và mức độ nghiêm trọng của các vụ va chạm giao thông là các dữ liệu quan trọng giúp phân tích tình trạng giao thông và đề xuất các giải pháp cải thiện hiệu quả.

Để giải quyết những vấn đề này, Seattle đã triển khai nhiều dự án nghiên cứu và phát triển nhằm tối ưu hóa hệ thống giao thông và giảm thiểu các tình trạng tắc nghẽn. Một trong số đó là dự án cải thiện an toàn giao thông, trong đó các công nghệ giám sát và dự đoán tình trạng giao thông được ứng dụng để giảm thiểu tai nạn và tăng cường an toàn cho người tham gia giao thông. Các hệ thống giám sát tự động, kết hợp với các phân tích dữ liệu thời gian thực, có thể phát hiện sớm các tình huống nguy hiểm và điều phối giao thông một cách kịp thời, giảm thiểu rủi ro tai nạn.

Một dự án đáng chú ý khác là dự án quản lý giao thông thông minh, sử dụng dữ liệu thời gian thực từ các cảm biến giao thông, camera và các hệ thống giám sát khác để tối ưu hóa tín hiệu đèn giao thông và phân phối các phương tiện công cộng một cách hiệu quả. Hệ thống đèn tín hiệu giao thông được điều chỉnh tự động dựa trên tình trạng giao thông ở các khu vực khác nhau, từ đó giảm thiểu tình trạng tắc nghẽn và đảm bảo giao thông thông suốt. Ngoài ra, các dữ liệu này cũng được sử dụng để phân bổ phương tiện công cộng một cách hợp lý, đáp ứng nhu cầu của cư dân trong các giờ cao điểm.

Các dự án này không chỉ giúp cải thiện tình trạng giao thông mà còn cho thấy sự quan trọng của việc áp dụng phân tích dữ liệu trong việc tối ưu hóa các chiến lược quản lý giao thông. Nhờ vào các công nghệ và phương pháp phân tích hiện đại, Seattle có thể xây dựng một hệ thống giao thông thông minh hơn, giảm thiểu tắc nghẽn và tăng cường an toàn cho người dân, đồng thời nâng cao chất lượng sống và sự bền vững của thành phố.

## 2.3. MỘT SỐ KHÁI NIỆM LIÊN QUAN

Trong đề tài dự đoán tình trạng tắc nghẽn giao thông tại Seattle, các công nghệ hiện đại như Học Máy (Machine Learning - ML), Mô hình Dự đoán và Xử lý Ngôn ngữ Tự nhiên (Natural Language Processing - NLP) đóng vai trò rất quan trọng. Những công nghệ này không chỉ giúp phân tích dữ liệu giao thông, mà còn hỗ trợ xây dựng mô hình dự đoán chính xác và hiệu quả để tối ưu hóa hệ thống giao thông đô thị.

Học Máy (Machine Learning - ML) là công nghệ cốt lõi trong việc dự đoán tình trạng giao thông và các yếu tố ảnh hưởng đến nó. Các thuật toán học máy giúp mô hình học từ dữ liệu lịch sử và cải thiện độ chính xác theo thời gian. Các thuật toán như Random Forest, Logistic Regression, và Support Vector Machine (SVM) có thể được sử dụng để phân loại tình trạng giao thông thành các mức độ nghiêm trọng khác nhau (ít nghiêm trọng, nghiêm trọng, rất nghiêm trọng), dựa trên các yếu tố như thời gian trong ngày, thời tiết, số lượng phương tiện và các vụ tai nạn giao thông. Học máy cho phép hệ thống tự động nhận diện các mẫu trong dữ liệu và đưa ra các dự đoán về tình trạng giao thông trong tương lai.

Mô hình Dự đoán là một phần quan trọng của bài toán, giúp dự báo tình trạng giao thông trong các khung thời gian ngắn hạn (như giờ hoặc ngày). Mô hình này sẽ dựa trên các dữ liệu đầu vào như tình trạng giao thông hiện tại, các yếu tố thời tiết, tai nạn và các sự kiện đặc biệt để đưa ra dự đoán về mức độ tắc nghẽn giao thông trong tương lai. Ví dụ, khi có sự kiện lớn hoặc thời tiết xấu, mô hình sẽ dự báo các khu vực có nguy cơ tắc nghẽn cao, từ đó giúp các cơ quan chức năng đưa ra các biện pháp điều phối giao thông kịp thời.

Xử lý Ngôn ngữ Tự nhiên (Natural Language Processing - NLP) cũng đóng một vai trò quan trọng trong việc khai thác dữ liệu không cấu trúc, chẳng hạn như các báo cáo tai nạn giao thông, thông tin từ các phương tiện truyền thông hoặc các báo cáo từ người tham gia giao thông. Các kỹ thuật NLP như phân loại văn bản và trích xuất thông tin giúp chuyển hóa thông tin từ các báo cáo tai nạn hoặc thông báo về tình trạng giao thông thành dữ liệu có thể sử dụng trong mô hình học máy. Điều này giúp nâng cao khả năng dự đoán và hỗ trợ các cơ quan quản lý đưa ra các quyết định chính xác hơn.

Nhờ vào sự kết hợp của các công nghệ như Học Máy, Mô hình Dự đoán và Xử lý Ngôn ngữ Tự nhiên, hệ thống dự đoán tắc nghẽn giao thông sẽ trở nên thông minh hơn, giúp tối ưu hóa các chiến lược điều phối giao thông, giảm thiểu tắc nghẽn và nâng cao an toàn giao thông cho người dân tại Seattle.

## 2.4. CÁC CÔNG CỤ VÀ THƯ VIỆN SỬ DỤNG

Để triển khai các mô hình học máy và xử lý dữ liệu trong nghiên cứu dự đoán tình trạng tắc nghẽn giao thông tại Seattle, một số công cụ và thư viện sẽ được sử dụng. Các công cụ này hỗ trợ từ việc thu thập, làm sạch, phân tích dữ liệu, đến việc xây dựng và đánh giá các mô hình học máy.

Pandas là thư viện phổ biến trong Python, chủ yếu được sử dụng để xử lý và làm sạch dữ liệu. Với cấu trúc dữ liệu mạnh mẽ như DataFrame, Pandas giúp thao tác với dữ liệu dạng bảng một cách dễ dàng. Trong nghiên cứu này, Pandas sẽ giúp đọc và xử lý các dữ liệu liên quan đến giao thông, tai nạn, và các yếu tố khác từ các file CSV hoặc cơ sở dữ liệu.

NumPy là thư viện mạnh mẽ dành cho phép toán số học và tính toán mảng. NumPy cung cấp các chức năng xử lý số liệu nhanh chóng và hiệu quả, hỗ trợ việc thực hiện các phép toán toán học phức tạp, tính toán thống kê và các phép toán ma trận trong các mô hình học máy.

Matplotlib và Seaborn là hai thư viện quan trọng cho việc trực quan hóa dữ liệu. Matplotlib giúp tạo ra các biểu đồ cơ bản như biểu đồ cột, biểu đồ đường, và biểu đồ phân phối, trong khi Seaborn, mở rộng Matplotlib, cung cấp các công cụ trực quan hóa mạnh mẽ hơn, đặc biệt là khi cần tạo các heatmap, scatter plots, hay boxplots để phân tích các mối quan hệ giữa các yếu tố trong dữ liệu giao thông.

Scikit-learn là thư viện học máy mạnh mẽ và dễ sử dụng, cung cấp một bộ công cụ cho các thuật toán như hồi quy, phân loại, phân nhóm và giảm chiều dữ liệu. Trong nghiên cứu này, Scikit-learn sẽ được sử dụng để triển khai các mô hình học máy như Random Forest, Logistic Regression, và Support Vector Machine (SVM). Thư viện này cũng cung cấp các công cụ đánh giá hiệu suất mô hình như accuracy, precision, recall, và F1-score, giúp xác định độ chính xác và hiệu quả của các mô hình dự đoán.

Những công cụ và thư viện này đóng vai trò quan trọng trong quá trình phân tích và xây dựng mô hình dự đoán tắc nghẽn giao thông, giúp cải thiện hiệu quả quản lý giao thông tại Seattle.

# CHƯƠNG 3: THU THẬP VÀ XỬ LÝ DỮ LIỆU

Quá trình xử lý dữ liệu là một bước nền tảng và có vai trò rất quan trọng trong toàn bộ quá trình phân tích. Dữ liệu đầu vào của đề tài được lấy từ tệp PDF chứa thông tin về giao dịch ngân hàng, trong đó nội dung ghi chú và thông tin người gửi được trình bày ở dạng phi cấu trúc. Do đó, quá trình xử lý dữ liệu bao gồm nhiều công đoạn: từ trích xuất dữ liệu, làm sạch, chuẩn hóa, trích xuất thông tin chính xác và cuối cùng là xuất ra tệp CSV để phục vụ các bước phân tích sau này.

Toàn bộ đoạn code thực nghiệm được trình bày tại link:

*https://github.com/pvthang29012004/do\_an\_2*

## 3.1. MÔ TẢ VỀ BÀI TOÁN

Bài toán đặt ra trong nghiên cứu này là xây dựng một mô hình học máy có khả năng phân loại mức độ tắc nghẽn giao thông (ít nghiêm trọng, nghiêm trọng, rất nghiêm trọng) tại thành phố Seattle. Thành phố Seattle, với mật độ dân cư cao và hệ thống giao thông phức tạp, thường xuyên phải đối mặt với các vấn đề giao thông nghiêm trọng, đặc biệt trong các khung giờ cao điểm, khi có thời tiết xấu, hoặc khi có sự kiện đông người diễn ra. Việc dự đoán trước tình trạng giao thông là rất quan trọng để có thể điều phối giao thông hiệu quả, giảm thiểu tắc nghẽn và đảm bảo an toàn cho người tham gia giao thông.

Mô hình học máy được xây dựng trong nghiên cứu này sẽ sử dụng dữ liệu lịch sử về các vụ va chạm giao thông tại Seattle. Dữ liệu này không chỉ bao gồm thông tin về tai nạn, mà còn kết hợp với các yếu tố quan trọng khác như điều kiện thời tiết, thời gian trong ngày, ngày trong tuần, tháng trong năm, và các sự kiện đặc biệt có thể ảnh hưởng đến tình hình giao thông. Các yếu tố này sẽ giúp mô hình hiểu được mối quan hệ giữa các điều kiện môi trường và mức độ tắc nghẽn giao thông.

### 3.1.1. Giới thiệu về tập dữ liệu

Bộ dữ liệu chính được sử dụng trong đề tài là **Seattle Traffic Collision Data**, được công khai bởi Sở Giao thông Seattle (SDOT). Bộ dữ liệu này ghi lại thông tin chi tiết về các vụ tai nạn giao thông trong nhiều năm tại thành phố Seattle, với hàng trăm nghìn bản ghi. Mỗi bản ghi trong dữ liệu phản ánh một sự cố giao thông cụ thể, bao gồm các thông tin quan trọng như:

* Thời gian: Ngày, giờ xảy ra vụ tai nạn.
* Địa điểm: Các tọa độ GPS, địa chỉ cụ thể nơi vụ tai nạn diễn ra.
* Mức độ nghiêm trọng: Các mức độ tai nạn như ít nghiêm trọng, nghiêm trọng và rất nghiêm trọng.
* Điều kiện thời tiết: Thông tin về thời tiết tại thời điểm xảy ra vụ tai nạn, như mưa, sương mù, hoặc nắng.
* Điều kiện mặt đường: Mặt đường khô, ướt, có băng tuyết, hoặc các yếu tố ảnh hưởng khác.
* Thông tin liên quan đến người tham gia giao thông: Bao gồm các yếu tố như việc lái xe mất tập trung, vi phạm tốc độ, và các hành vi nguy hiểm khác.

Dữ liệu này cung cấp cái nhìn sâu sắc về các yếu tố tác động đến tai nạn giao thông, giúp phân tích các mối quan hệ giữa các yếu tố như thời gian, điều kiện thời tiết, và các hành vi của người tham gia giao thông. Nó là cơ sở để xây dựng các mô hình dự đoán, hỗ trợ việc cải thiện an toàn giao thông và tối ưu hóa việc điều phối giao thông trong tương lai .

### 3.1.2. Tổng quan các trường của dữ liệu

Tập dữ liệu được sử dụng trong đề tài bao gồm hàng trăm nghìn bản ghi, mỗi bản ghi mô tả chi tiết một vụ tai nạn giao thông xảy ra tại thành phố Seattle. Sau quá trình tiền xử lý, nhóm nghiên cứu đã chọn lọc các trường dữ liệu quan trọng, có ý nghĩa trực tiếp đến mục tiêu dự đoán mức độ tắc nghẽn giao thông. Cụ thể:

* ADDRTYPE: Loại địa chỉ nơi xảy ra tai nạn (ví dụ: giao lộ, đoạn đường), phản ánh vị trí hình học của sự kiện.
* LOCATION: Tên vị trí cụ thể (đường phố, giao lộ), giúp định vị địa lý vụ va chạm.
* SEVERITYCODE: Mã hóa mức độ nghiêm trọng của tai nạn – biến mục tiêu trong bài toán phân loại.
* SEVERITYDESC: Mô tả chi tiết mức độ nghiêm trọng (như: không bị thương, bị thương nhẹ, tử vong).
* COLLISIONTYPE: Kiểu va chạm xảy ra (ví dụ: rẽ trái va vào người đi bộ), phản ánh hình thức của tai nạn.
* PERSONCOUNT: Tổng số người liên quan trong vụ tai nạn (tài xế, hành khách, người đi bộ,...).
* PEDCOUNT: Số người đi bộ bị ảnh hưởng.
* PEDCYLCOUNT: Số người đi xe đạp bị ảnh hưởng.
* VEHCOUNT: Số lượng phương tiện tham gia vào vụ va chạm.
* INJURIES: Tổng số người bị thương.
* SERIOUSINJURIES: Số người bị thương nặng
* FATALITIES: Số người tử vong trong vụ tai nạn.
* INCDATE / INCDTTM: Ngày và thời điểm cụ thể xảy ra tai nạn – giúp trích xuất thông tin về thời gian như giờ cao điểm, ngày trong tuần,...
* JUNCTIONTYPE: Loại nút giao thông (ví dụ: ngã tư, vòng xoay), ảnh hưởng đến khả năng gây ùn tắc.
* INATTENTIONIND: Chỉ báo tai nạn có liên quan đến mất tập trung khi lái xe.
* UNDERINFL: Chỉ báo lái xe dưới ảnh hưởng của chất kích thích (rượu, ma túy).
* WEATHER: Điều kiện thời tiết tại thời điểm xảy ra tai nạn (mưa, nắng, sương mù,...).
* ROADCOND: Tình trạng mặt đường (khô ráo, ướt, băng tuyết,...).
* LIGHTCOND: Điều kiện ánh sáng (ban ngày, ban đêm, đèn đường sáng,...)
* SPEEDING: Có hay không vi phạm tốc độ trong vụ tai nạn.
* HITPARKEDCAR: Tai nạn có liên quan đến việc đâm vào xe đang đậu hay không.

## 3.2. TIỀN XỬ LÝ DỮ LIỆU

### 3.2.1. Xử lý dữ liệu còn thiếu

Trong quá trình tiền xử lý dữ liệu, việc xử lý các giá trị thiếu là một bước quan trọng để đảm bảo rằng mô hình học máy sẽ nhận được dữ liệu đầy đủ và chính xác. Dữ liệu thiếu có thể gây ảnh hưởng tiêu cực đến hiệu suất của mô hình, vì vậy các giá trị thiếu cần phải được xử lý hợp lý. Trong nghiên cứu này, các phương pháp sau được áp dụng để xử lý giá trị thiếu trong bộ dữ liệu Seattle Traffic Collision Data:

Đoạn mã trên thực hiện các bước tiền xử lý quan trọng để làm sạch dữ liệu và chuẩn hóa các trường dữ liệu có giá trị thiếu trong bộ dữ liệu. Cụ thể, đầu tiên, mã loại bỏ tất cả các dòng có giá trị thiếu trong cột SEVERITYCODE bằng cách sử dụng hàm dropna(subset=['SEVERITYCODE']), vì đây là biến mục tiêu trong bài toán phân loại và không thể thiếu thông tin.

Tiếp theo, các trường dữ liệu có giá trị thiếu trong các cột phân loại như SPEEDING, INATTENTIONIND, và UNDERINFL được điền bằng các giá trị cố định. Ví dụ, các giá trị thiếu trong cột SPEEDING (liệu có vi phạm tốc độ trong vụ tai nạn) được điền bằng giá trị ‘N’, giả định rằng vụ tai nạn không liên quan đến vi phạm tốc độ nếu không có thông tin cụ thể. Tương tự, cột INATTENTIONIND (liệu tai nạn có liên quan đến mất tập trung khi lái xe) và UNDERINFL (liệu tài xế có bị ảnh hưởng bởi chất kích thích) cũng được điền giá trị ‘N’, nghĩa là không mất tập trung khi lái xe và không có ảnh hưởng của chất kích thích.

Đối với các cột phân loại như LIGHTCOND, WEATHER, ROADCOND, và COLLISIONTYPE, mã sử dụng phương pháp điền giá trị thiếu bằng giá trị phổ biến nhất (mode) trong từng cột. Điều này giúp duy trì tính nhất quán và ổn định của dữ liệu, đồng thời phản ánh điều kiện chung trong các vụ tai nạn.

Các cột JUNCTIONTYPE (loại nút giao thông) và ADDRTYPE (loại địa chỉ) cũng được điền bằng giá trị phổ biến nhất, đảm bảo rằng thông tin về giao lộ và địa chỉ không bị thiếu, giúp mô hình học máy có đầy đủ thông tin để phân tích.

Cuối cùng, mã xử lý cột LOCATION, nơi lưu trữ thông tin về vị trí vụ tai nạn, bằng cách tách tên đường chính từ thông tin có dấu & trong trường hợp nhiều địa chỉ được liệt kê trong cùng một bản ghi. Tên đường chính sau đó được lưu vào cột mới MAIN\_STREET. Trước khi thực hiện tách, nếu cột LOCATION có giá trị thiếu, nó sẽ được điền bằng ‘Unknown’.

### 3.2.2. Xử lý dữ liệu thời gian

Đoạn mã trên thực hiện một số bước tiền xử lý dữ liệu quan trọng để chuẩn bị dữ liệu cho mô hình học máy, đặc biệt là đối với các trường dữ liệu liên quan đến thời gian và các giá trị không hợp lệ trong bộ dữ liệu về tai nạn giao thông.

Đầu tiên, đoạn mã loại bỏ các dòng dữ liệu trong ba cột quan trọng là WEATHER, ROADCOND, và LIGHTCOND nếu chúng chứa các giá trị 'unknown' hoặc 'other', vì các giá trị này không có ý nghĩa thực tế trong phân tích và dự đoán. Việc này giúp dữ liệu trở nên sạch sẽ và hợp lệ hơn, tránh đưa vào các dữ liệu không chính xác hoặc không rõ ràng, từ đó cải thiện hiệu quả của mô hình học máy.

Tiếp theo, mã loại bỏ tất cả các dòng có giá trị thiếu trong cột INCDTTM, cột này chứa thông tin về thời gian xảy ra vụ tai nạn. Dữ liệu thời gian không hợp lệ hoặc thiếu có thể làm giảm tính chính xác của mô hình, vì vậy các dòng có giá trị NaT (Not a Time) trong cột này được loại bỏ.

Sau đó, đoạn mã chuyển cột INCDTTM từ dạng chuỗi (string) sang định dạng datetime sử dụng hàm pd.to\_datetime(), giúp chuẩn hóa và dễ dàng thao tác với dữ liệu thời gian. Tham số errors='coerce' đảm bảo rằng bất kỳ giá trị nào không thể chuyển đổi thành thời gian hợp lệ sẽ trở thành NaT, từ đó có thể dễ dàng xử lý tiếp.

Tiếp theo, đoạn mã tiếp tục loại bỏ những dòng dữ liệu có giá trị NaT trong cột INCDTTM sau khi chuyển đổi, để chắc chắn rằng chỉ những vụ tai nạn có thời gian hợp lệ được giữ lại trong bộ dữ liệu.

Đoạn mã tiếp theo tạo ra bốn cột mới từ dữ liệu thời gian trong cột INCDTTM:

* Hour: Lấy giờ trong ngày (từ 0 đến 23),
* DayOfWeek: Lấy ngày trong tuần (0 = Thứ Hai, 6 = Chủ Nhật),
* Day: Lấy ngày trong tháng (từ 1 đến 31),
* Month: Lấy tháng trong năm (từ 1 đến 12). Những cột này giúp phân tích sâu hơn về các yếu tố thời gian, như giờ cao điểm, ngày trong tuần, và tháng trong năm, ảnh hưởng đến tình trạng giao thông và tai nạn.

Cuối cùng, đoạn mã kiểm tra và in ra 5 dòng đầu tiên của bộ dữ liệu sau khi tạo các cột thời gian để xác nhận kết quả. Điều này giúp đảm bảo rằng các cột thời gian đã được tạo chính xác và có ý nghĩa. Sau đó, đoạn mã kiểm tra số lượng giá trị thiếu trong các cột thời gian này để đảm bảo không có giá trị thiếu hoặc lỗi trong dữ liệu thời gian.

### 3.2.3. Phân tích và chuẩn hoá biến mục tiêu

Đoạn mã trên thực hiện các bước xử lý dữ liệu liên quan đến biến **SEVERITYCODE**, biến mục tiêu trong bài toán phân loại mức độ nghiêm trọng của vụ tai nạn giao thông. Mục tiêu của các bước này là chuẩn hóa và làm sạch dữ liệu để đảm bảo tính chính xác của mô hình học máy.

Đầu tiên, mã phân tích các giá trị duy nhất trong cột **SEVERITYCODE** và **SEVERITYDESC** (mô tả mức độ nghiêm trọng), sử dụng hàm value\_counts() để đếm số lần xuất hiện của từng cặp giá trị (mã mức độ nghiêm trọng và mô tả). Kết quả giúp hiểu rõ phân bố các mức độ nghiêm trọng trong bộ dữ liệu và xác định các giá trị không hợp lệ hoặc bất thường.

Sau đó, đoạn mã loại bỏ các dòng có giá trị**SEVERITYCODE** bằng ‘0’, vì đây có thể là giá trị không hợp lệ hoặc không xác định, không cung cấp thông tin hữu ích cho việc phân loại mức độ nghiêm trọng. Việc loại bỏ các dòng này giúp tập trung vào các vụ tai nạn có mức độ nghiêm trọng rõ ràng và có ý nghĩa.

Tiếp theo, mã thay thế các giá trị ‘2b’ trong cột **SEVERITYCODE** thành ‘2’, chuẩn hóa dữ liệu vì ‘2b’ có thể là một lỗi nhập liệu hoặc một biến thể không chính thức của mức độ nghiêm trọng ‘2’. Điều này giúp dữ liệu trở nên thống nhất và dễ dàng sử dụng trong quá trình phân tích.

Sau khi thay thế giá trị ‘2b’ thành ‘2’, cột **SEVERITYCODE** được chuyển đổi từ kiểu dữ liệu chuỗi (string) sang kiểu số nguyên (integer) bằng phương thức astype(int). Việc chuyển đổi này là cần thiết vì mô hình học máy thường yêu cầu dữ liệu số để có thể xử lý và phân tích.

Cuối cùng, mã kiểm tra lại các giá trị duy nhất trong **SEVERITYCODE** sau khi đã chuẩn hóa và chuyển đổi, sử dụng lại hàm value\_counts() để đếm số lần xuất hiện của các giá trị và sort\_index() để sắp xếp chúng theo thứ tự tăng dần. Điều này giúp đảm bảo rằng dữ liệu đã được chuẩn hóa đúng cách và không còn giá trị sai lệch, đồng thời cung cấp cái nhìn tổng thể về phân phối của các mức độ nghiêm trọng trong dữ liệu.

Tóm lại, đoạn mã này giúp chuẩn hóa và làm sạch dữ liệu liên quan đến mức độ nghiêm trọng của tai nạn giao thông, đảm bảo rằng các giá trị trong **SEVERITYCODE** là hợp lệ và dễ sử dụng cho việc huấn luyện mô hình học máy, giúp mô hình phân loại mức độ nghiêm trọng của tai nạn chính xác hơn.

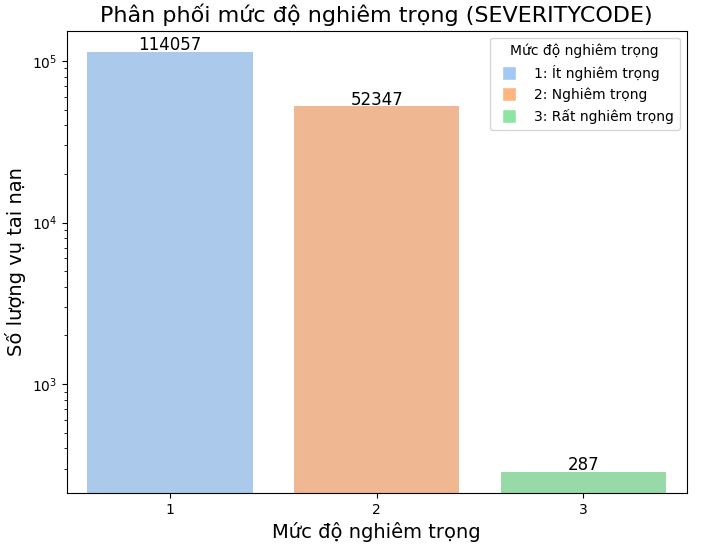
## 3.3. PHÂN TÍCH VÀ KHÁM PHÁ DỮ LIỆU

### 3.3.1. Biến mục tiêu và phân tích mức độ nghiêm trọng

#### 3.3.1.1. Biến mục tiêu

Biểu đồ phân phối mức độ nghiêm trọng (SEVERITYCODE) cho thấy số lượng các vụ tai nạn giao thông tại Seattle phân chia theo ba mức độ nghiêm trọng. Sử dụng phương pháp countplot trong thư viện seaborn, ta có thể thấy rằng:

* Mức độ 1 (ít nghiêm trọng) chiếm phần lớn với 114,057 vụ tai nạn.
* Mức độ 2 (nghiêm trọng) chiếm 52,347 vụ.
* Mức độ 3 (rất nghiêm trọng) chiếm 287 vụ.



Hình 3.1. Biểu đồ phân phối mức độ nghiêm trong SEVERITYCODE

Biểu đồ này thể hiện phân phối số lượng các vụ tai nạn giao thông tại Seattle phân theo ba mức độ nghiêm trọng được mã hóa trong cột SEVERITYCODE. Biểu đồ giúp nhận diện rõ ràng phân bố của các mức độ nghiêm trọng của tai nạn và tác động của từng mức độ này đối với giao thông.

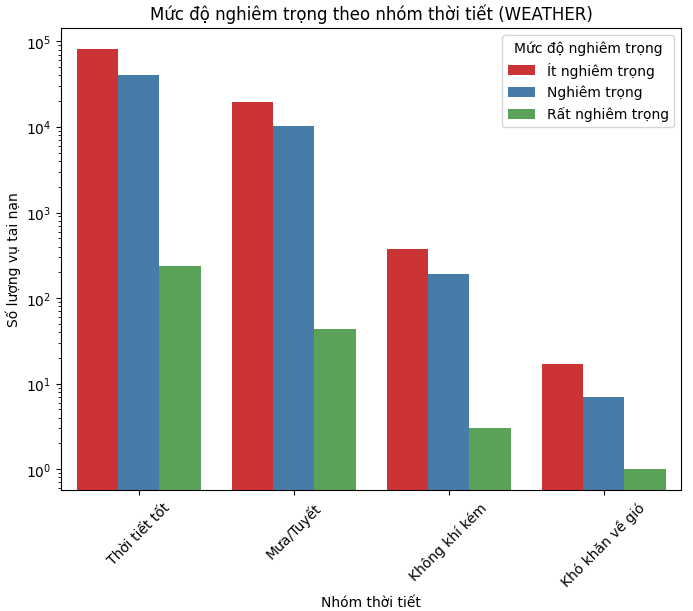
Nhận xét về biểu đồ:

* Mức độ 1 – Ít nghiêm trọng (Xanh da trời): Đây là mức độ xảy ra nhiều nhất với 114,057 vụ tai nạn, chiếm phần lớn trong tổng số vụ tai nạn. Những vụ này thường liên quan đến thiệt hại nhỏ hoặc không gây thương tích nghiêm trọng, có thể là các va chạm nhẹ hoặc những tình huống không ảnh hưởng lớn đến giao thông. Mặc dù chiếm phần lớn nhưng mức độ này không gây tắc nghẽn giao thông lớn.
* Mức độ 2 – Nghiêm trọng (Cam): Với 52,347 vụ tai nạn, mức độ này chiếm tỷ lệ lớn thứ hai. Các vụ này có thể gây thương tích nhẹ đến nghiêm trọng, mặc dù không phải là vụ tai nạn nghiêm trọng nhất, nhưng chúng vẫn có thể tạo ra sự tắc nghẽn giao thông đáng kể và cần xử lý khẩn cấp.
* Mức độ 3 – Rất nghiêm trọng (Xanh lá cây): Chỉ có 287 vụ tai nạn thuộc mức độ này, chiếm một tỷ lệ rất nhỏ trong tổng số vụ tai nạn (chưa đến 1%). Tuy nhiên, những vụ này rất nghiêm trọng và có thể gây tử vong hoặc thương tích nặng. Dù ít xảy ra, nhưng các vụ tai nạn này gây ra tắc nghẽn giao thông lớn và cần sự xử lý nhanh chóng.

#### 3.3.1.2. phân tích mức độ nghiêm trọng

*a) phân tích mức độ nghiêm trọng của WEATHER*

Trong bối cảnh giao thông đô thị, yếu tố thời tiết (**WEATHER**) đóng vai trò quan trọng trong việc ảnh hưởng đến mức độ nghiêm trọng của các vụ tai nạn giao thông. Biểu đồ phân tích mức độ nghiêm trọng của tai nạn theo các điều kiện thời tiết sẽ giúp xác định mối quan hệ giữa điều kiện thời tiết và mức độ nghiêm trọng của tai nạn, từ đó đưa ra các biện pháp điều phối giao thông hợp lý.



Hình 3.2. Biều đồ mức độ theo nhóm thời tiết (**WEATHER**)

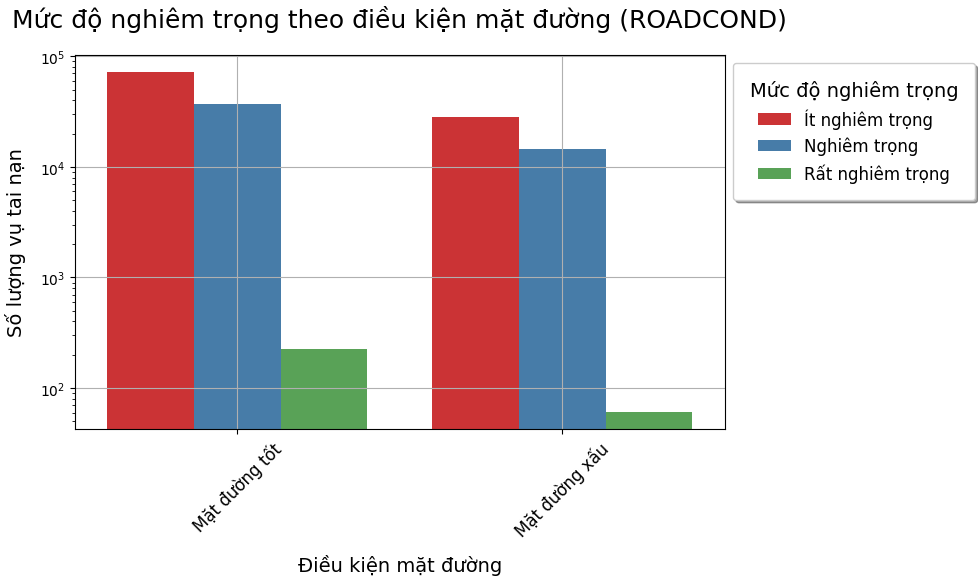
Biểu đồ trên thể hiện mức độ nghiêm trọng của các vụ tai nạn giao thông tại Seattle phân theo nhóm thời tiết. Các nhóm thời tiết bao gồm Thời tiết tốt, Mưa tuyết, Không khí kém, và Khó khăn về gió, với mỗi nhóm được phân loại theo mức độ nghiêm trọng của tai nạn, bao gồm ít nghiêm trọng, nghiêm trọng, và rất nghiêm trọng.

Nhận thấy rằng mưa tuyết có tỷ lệ cao nhất đối với các vụ tai nạn nghiêm trọng và rất nghiêm trọng, điều này cho thấy điều kiện thời tiết xấu như mưa tuyết có ảnh hưởng lớn đến tình trạng giao thông, khiến đường trơn trượt và tầm nhìn kém, làm gia tăng khả năng xảy ra các vụ tai nạn nghiêm trọng. Trong khi đó, với thời tiết tốt, phần lớn các vụ tai nạn lại thuộc mức độ ít nghiêm trọng, cho thấy thời tiết thuận lợi giúp giảm thiểu các vụ tai nạn nghiêm trọng và ít ảnh hưởng đến giao thông. Tuy nhiên, đối với các nhóm không khí kém và khó khăn về gió, mặc dù số vụ tai nạn thấp hơn, nhưng vẫn có các vụ tai nạn nghiêm trọng, đặc biệt là mức độ nghiêm trọng và rất nghiêm trọng, cho thấy gió mạnh và không khí ô nhiễm cũng có thể gây khó khăn trong việc điều khiển phương tiện và làm tăng nguy cơ tai nạn.

Tổng kết lại, biểu đồ này cho thấy mối quan hệ giữa điều kiện thời tiết và mức độ nghiêm trọng của tai nạn giao thông. Thời tiết xấu, đặc biệt là mưa tuyết, có xu hướng làm gia tăng tỷ lệ tai nạn nghiêm trọng, trong khi thời tiết tốt giúp giảm thiểu các sự cố nghiêm trọng. Điều này cung cấp thông tin quan trọng cho các cơ quan chức năng trong việc quản lý giao thông và đưa ra cảnh báo phù hợp trong các điều kiện thời tiết khác nhau.

*b) phân tích mức độ nghiêm trọng của ROADCOND*

Tiếp theo trong phần phân tích mức độ nghiêm trọng của tai nạn giao thông, chúng ta sẽ xét đến yếu tố **ROADCOND** (tình trạng mặt đường). **ROADCOND** là một yếu tố quan trọng ảnh hưởng trực tiếp đến sự an toàn khi tham gia giao thông. Các điều kiện mặt đường như khô ráo, ướt, có băng tuyết hay ngập lụt có thể làm tăng nguy cơ xảy ra tai nạn và nghiêm trọng hóa hậu quả của tai nạn.



Hình 3.3. Biều đồ mức độ nghiêm trọng theo điều kiện mặt đường (ROADCOND)

Biểu đồ trên thể hiện mức độ nghiêm trọng của các vụ tai nạn giao thông tại Seattle theo hai nhóm Điều kiện mặt đường: Mặt đường tốt và Mặt đường xấu. Nhìn vào biểu đồ, ta có thể nhận thấy rõ mối quan hệ giữa tình trạng mặt đường và mức độ nghiêm trọng của các vụ tai nạn.

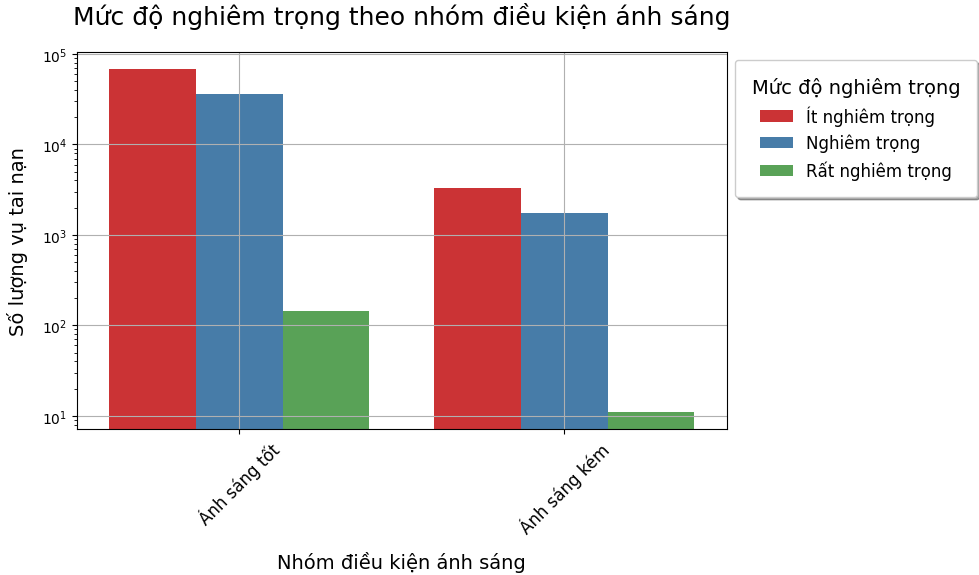
Khi mặt đường tốt, số lượng các vụ tai nạn ít nghiêm trọng chiếm phần lớn, và chỉ có một số ít vụ nghiêm trọng. Điều này cho thấy rằng trong điều kiện mặt đường khô ráo và không có yếu tố gây cản trở, các vụ tai nạn có xu hướng ít nghiêm trọng. Tuy nhiên, vẫn cần chú ý và có các biện pháp xử lý các vụ tai nạn nghiêm trọng xảy ra trong điều kiện này.

Ngược lại, khi mặt đường xấu, bao gồm các tình trạng như đường ướt, có băng tuyết, hay ngập lụt, tỷ lệ các vụ tai nạn nghiêm trọng và rất nghiêm trọng tăng lên rõ rệt. Điều này cho thấy rằng các điều kiện mặt đường xấu làm tăng nguy cơ xảy ra tai nạn nghiêm trọng và thậm chí là các vụ tai nạn rất nghiêm trọng, có thể gây thương tích nghiêm trọng hoặc tử vong.

Tổng kết lại, biểu đồ này cho thấy rõ mối liên hệ giữa tình trạng mặt đường và mức độ nghiêm trọng của các vụ tai nạn. Điều này nhấn mạnh tầm quan trọng của việc cảnh báo và điều phối giao thông khi điều kiện mặt đường không tốt, nhằm giảm thiểu tai nạn và đảm bảo an toàn cho người tham gia giao thông.

*c) phân tích mức độ nghiêm trọng của LIGHTCOND*

Tiếp theo trong phần phân tích mức độ nghiêm trọng của tai nạn giao thông, chúng ta sẽ xem xét yếu tố **LIGHTCOND** (điều kiện ánh sáng). Điều kiện ánh sáng có ảnh hưởng trực tiếp đến khả năng quan sát và tầm nhìn của người lái xe, từ đó tác động mạnh mẽ đến mức độ nghiêm trọng của các vụ tai nạn. Các yếu tố ánh sáng như ban ngày, ban đêm hoặc đèn đường sáng có thể là yếu tố quyết định trong việc xảy ra các vụ tai nạn và mức độ nghiêm trọng của chúng.

**

Hình 3.3. Biều đồ mức độ nghiêm trọng theo nhóm điều kiện ánh sáng

Biểu đồ trên thể hiện mức độ nghiêm trọng của các vụ tai nạn giao thông tại Seattle theo nhóm **Điều kiện ánh sáng**, bao gồm **Ánh sáng tốt** và **Ánh sáng kém**. Khi điều kiện **ánh sáng tốt**, biểu đồ cho thấy số lượng các vụ tai nạn chủ yếu thuộc mức độ **ít nghiêm trọng,** với một số ít vụ tai nạn nghiêm trọng. Điều này phản ánh rằng trong điều kiện ánh sáng tốt, người lái xe có thể quan sát rõ ràng hơn, giúp giảm thiểu các vụ tai nạn nghiêm trọng, mặc dù vẫn có những trường hợp gây thương tích nhẹ.

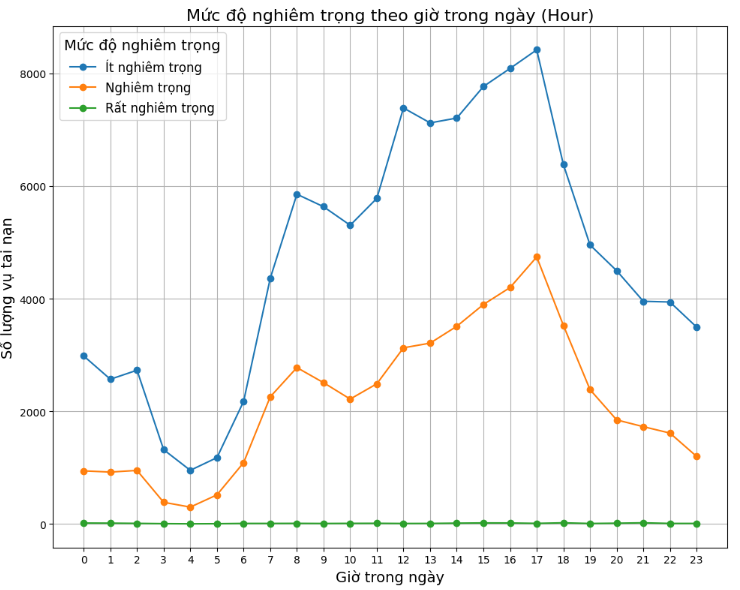
Tuy nhiên, khi điều kiện **ánh sáng kém** xảy ra, đặc biệt là vào ban đêm hoặc trong điều kiện thiếu sáng, tỷ lệ các vụ tai nạn **nghiêm trọng** và **rất nghiêm trọng** tăng lên rõ rệt. Điều này cho thấy ánh sáng yếu làm tăng nguy cơ tai nạn do tầm nhìn bị hạn chế, khiến người lái xe dễ mất kiểm soát phương tiện và dẫn đến những vụ tai nạn nghiêm trọng hơn. Như vậy, mối quan hệ giữa điều kiện ánh sáng và mức độ nghiêm trọng của tai nạn là rõ ràng: khi ánh sáng kém, mức độ nghiêm trọng của tai nạn giao thông có xu hướng tăng cao. Phân tích này nhấn mạnh tầm quan trọng của việc cải thiện chiếu sáng công cộng và cảnh báo người tham gia giao thông cẩn thận hơn trong điều kiện ánh sáng yếu.

### 3.3.2. Phân tích mức độ nghiêm trọng theo yếu tố thời gian

#### 3.3.2.1. Phân tích mức độ nghêm trọng theo giờ trong ngày

*a) Theo giờ trong ngày*

Để phân tích mức độ nghiêm trọng của các vụ tai nạn giao thông tại Seattle theo thời gian trong ngày, chúng ta sẽ sử dụng biểu đồ phân phối mức độ nghiêm trọng của tai nạn theo giờ trong ngày. Biểu đồ này giúp chúng ta hiểu rõ hơn về sự thay đổi của mức độ nghiêm trọng của các vụ tai nạn từ sáng đến tối. Với ba nhóm mức độ nghiêm trọng được phân biệt (ít nghiêm trọng, nghiêm trọng, và rất nghiêm trọng), biểu đồ sẽ cho thấy sự phân bố của các vụ tai nạn trong từng khung giờ, từ đó giúp xác định các thời điểm có nguy cơ cao về tai nạn nghiêm trọng.



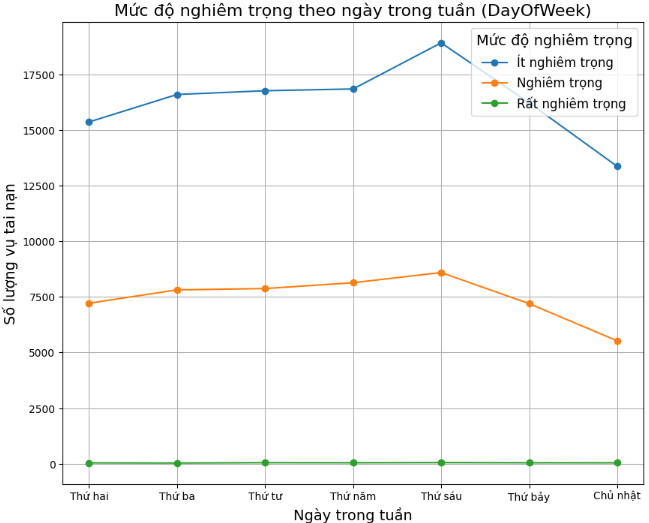
Hình 3.4. Biều đồ mức độ nghiêm trọng theo giờ trong ngày(Hour)

Biểu đồ trên thể hiện mức độ nghiêm trọng của các vụ tai nạn giao thông tại Seattle phân theo giờ trong ngày, từ 0h đến 23h, với ba mức độ nghiêm trọng: ít nghiêm trọng, nghiêm trọng, và rất nghiêm trọng. Nhìn vào biểu đồ, ta thấy rằng trong giờ thấp điểm (0h - 6h), số lượng vụ tai nạn ít nghiêm trọng chiếm ưu thế, với rất ít vụ nghiêm trọng và rất nghiêm trọng. Điều này phản ánh rằng vào ban đêm và sáng sớm, giao thông thường ít và tình trạng tai nạn cũng nhẹ nhàng hơn. Tuy nhiên, vào giờ cao điểm buổi sáng (6h - 9h), khi mật độ giao thông tăng cao, số lượng vụ tai nạn nghiêm trọng gia tăng rõ rệt, mặc dù các vụ rất nghiêm trọng vẫn chiếm tỷ lệ thấp.

Đến giờ cao điểm buổi chiều (16h - 18h), khi nhiều người tham gia giao thông để về nhà, số vụ tai nạn nghiêm trọng tiếp tục tăng mạnh, nhưng số vụ rất nghiêm trọng vẫn giữ ở mức thấp. Cuối ngày, sau giờ cao điểm, số lượng vụ tai nạn giảm dần, với các vụ nghiêm trọng vẫn xảy ra trong một số thời điểm, nhưng không đáng kể. Tổng kết lại, biểu đồ cho thấy mật độ giao thông trong giờ cao điểm có ảnh hưởng trực tiếp đến mức độ nghiêm trọng của các vụ tai nạn, và các biện pháp điều phối giao thông vào các thời điểm này có thể giúp giảm thiểu tai nạn nghiêm trọng.

*b) Theo ngày trong tuần*

Tiếp theo trong phân tích mức độ nghiêm trọng của tai nạn giao thông, chúng ta sẽ chuyển sang phần **phân tích mức độ nghiêm trọng theo ngày trong tuần**. Mỗi ngày trong tuần có đặc điểm giao thông khác nhau, với các yếu tố như lịch trình làm việc, giờ cao điểm, và các hoạt động đặc biệt có thể ảnh hưởng đến mật độ giao thông và mức độ nghiêm trọng của tai nạn. Phân tích này sẽ giúp chúng ta hiểu rõ hơn về mối liên hệ giữa ngày trong tuần và mức độ nghiêm trọng của các vụ tai nạn, từ đó đưa ra các chiến lược hiệu quả hơn để quản lý giao thông và giảm thiểu tai nạn.



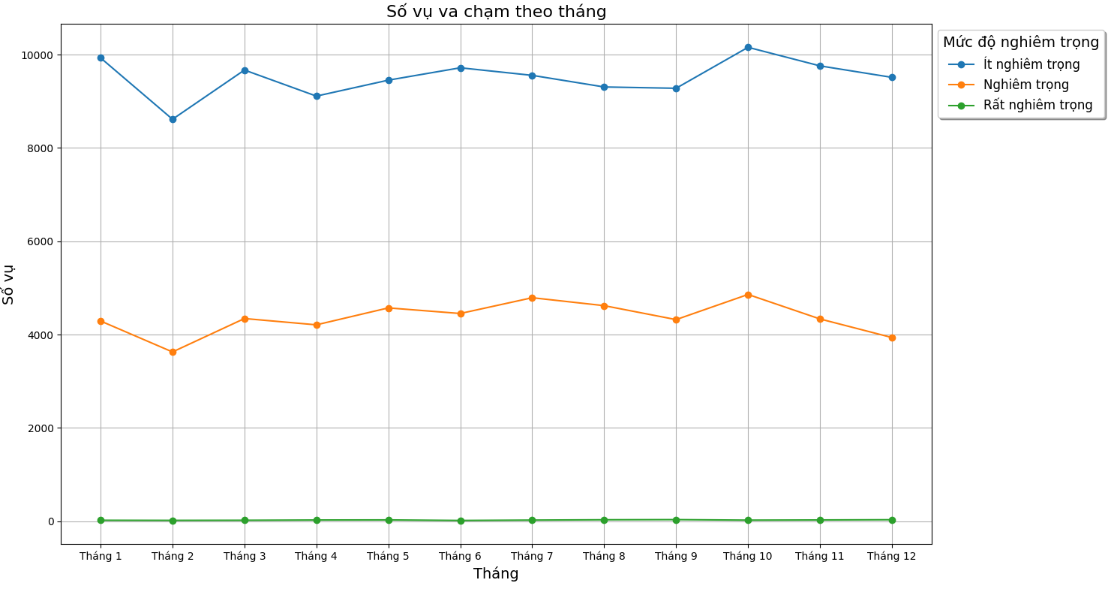
Hình 3.5. Biểu đồ mức độ nghiêm trọng theo ngày trong tuần(DayOfWeek)

Biểu đồ trên thể hiện mức độ nghiêm trọng của các vụ tai nạn giao thông tại Seattle phân theo **ngày trong tuần**. Nhìn vào biểu đồ, chúng ta có thể thấy rằng **thứ ba** là ngày có số vụ tai nạn **ít nghiêm trọng** cao nhất, với mức độ giảm dần vào các ngày sau đó. Điều này cho thấy giao thông vào ngày thứ ba có thể đông đúc, dẫn đến số lượng tai nạn nhẹ tăng lên. Số lượng vụ tai nạn **nghiêm trọng** có xu hướng tăng vào **thứ hai** và **thứ ba,** và giảm dần trong các ngày còn lại của tuần.

Tuy nhiên, số vụ **rất nghiêm trọng** vẫn duy trì ở mức thấp trong suốt cả tuần, với không có sự thay đổi đáng kể giữa các ngày. Điều này có thể chỉ ra rằng tai nạn rất nghiêm trọng hiếm khi xảy ra và không bị ảnh hưởng mạnh bởi ngày trong tuần. Cuối tuần, đặc biệt là vào **Chủ nhật,** số lượng tai nạn **nghiêm trọng** giảm, phản ánh mật độ giao thông thấp hơn và ít người tham gia giao thông. Nhìn chung, biểu đồ cho thấy giao thông vào **thứ hai** và **thứ ba** có mức độ nghiêm trọng cao hơn, trong khi các ngày còn lại trong tuần và cuối tuần có tỷ lệ tai nạn nghiêm trọng thấp hơn.

*c) Theo tháng*

Tiếp theo trong phân tích mức độ nghiêm trọng của các vụ tai nạn giao thông, chúng ta sẽ chuyển sang phần **phân tích mức độ nghiêm trọng theo tháng.** Mỗi tháng trong năm có những đặc điểm giao thông và thời tiết khác nhau, điều này có thể ảnh hưởng đến tần suất cũng như mức độ nghiêm trọng của các vụ tai nạn. Việc phân tích mức độ nghiêm trọng của tai nạn theo tháng sẽ giúp chúng ta nhận diện các xu hướng theo mùa và đưa ra các giải pháp điều phối giao thông phù hợp trong từng thời điểm của năm. Hãy cùng xem xét biểu đồ phân tích mức độ nghiêm trọng của tai nạn giao thông tại Seattle theo từng tháng để hiểu rõ hơn về sự biến động của tình trạng giao thông trong suốt cả năm.



Hình 3.6. Biểu đồ số vụ va chạm theo tháng

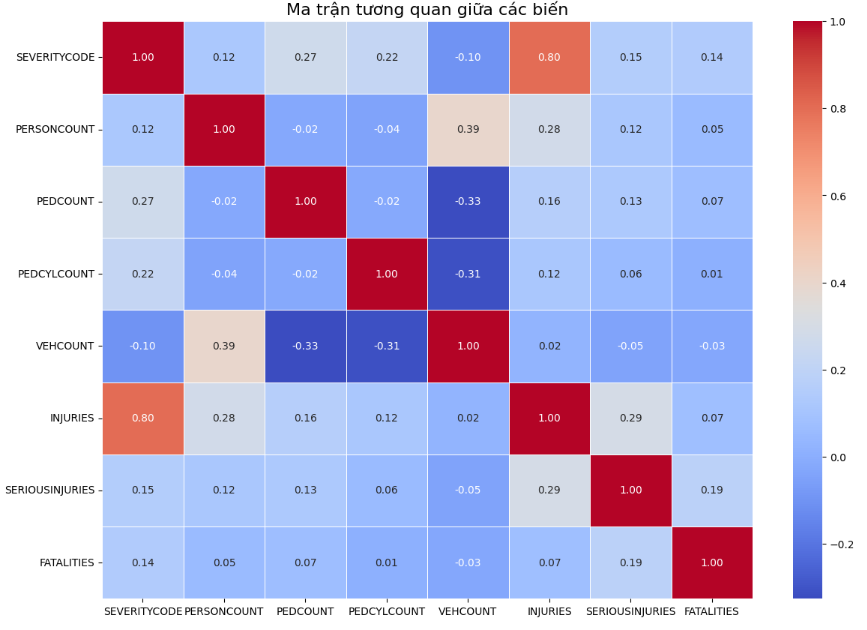
Biểu đồ trên thể hiện số vụ tai nạn giao thông tại Seattle phân theo tháng trong năm với ba mức độ nghiêm trọng: ít nghiêm trọng (xanh dương), nghiêm trọng (cam), và rất nghiêm trọng (xanh lá cây). Qua biểu đồ, chúng ta có thể nhận thấy sự phân bố của các vụ tai nạn theo tháng trong năm, với số lượng vụ tai nạn nghiêm trọng luôn chiếm ưu thế, đặc biệt là vào các tháng mùa hè.

Các vụ tai nạn nghiêm trọng có xu hướng gia tăng vào tháng 7 và tháng 8, với số vụ tai nạn ít nghiêm trọng cũng tăng lên trong khoảng thời gian này. Điều này có thể được giải thích do mật độ giao thông cao vào mùa hè, khi nhiều người di chuyển, du lịch và tham gia các hoạt động ngoài trời, dẫn đến tần suất tai nạn cao hơn. Mặc dù số vụ rất nghiêm trọng rất ít và không có sự biến động lớn trong suốt năm, điều này cho thấy rằng các vụ tai nạn rất nghiêm trọng thường ít xảy ra nhưng lại có ảnh hưởng lớn đến tình trạng giao thông.

Tổng quan, biểu đồ chỉ ra rằng các tháng mùa hè có sự gia tăng của tai nạn nghiêm trọng, đặc biệt là vào tháng 7 và tháng 8. Tuy nhiên, số vụ tai nạn rất nghiêm trọng vẫn duy trì ở mức thấp trong suốt cả năm. Việc hiểu rõ sự thay đổi này giúp các cơ quan chức năng có thể đưa ra các biện pháp phòng ngừa tai nạn hiệu quả trong những tháng có mật độ giao thông cao.

### 3.3.3. Ma trận tương quan (Correlation Matrix)

Tiếp theo trong quá trình phân tích dữ liệu, chúng ta sẽ thực hiện phân tích ma trận tương quan (Correlation Matrix), một công cụ quan trọng giúp xác định mối quan hệ giữa các biến trong dữ liệu. Ma trận tương quan sẽ giúp chúng ta hiểu được mức độ ảnh hưởng và sự liên kết giữa các yếu tố như thời tiết, điều kiện mặt đường, mức độ nghiêm trọng của tai nạn và các yếu tố khác. Bằng cách phân tích ma trận tương quan, chúng ta có thể nhận diện những yếu tố nào có ảnh hưởng mạnh mẽ nhất đến mức độ nghiêm trọng của các vụ tai nạn giao thông. Đây là bước quan trọng để tối ưu hóa mô hình học máy và cải thiện độ chính xác của dự đoán. Hãy cùng tiếp tục với việc phân tích ma trận tương quan để làm rõ mối quan hệ giữa các biến trong bộ dữ liệu này.



Hình 3.7. Biểu đồ ma trận tương quan giữa các biến

Biểu đồ ma trận tương quan này giúp chúng ta hiểu rõ mối quan hệ giữa các biến trong bộ dữ liệu tai nạn giao thông tại Seattle. Mỗi ô trong ma trận này thể hiện mức độ tương quan giữa hai biến, với giá trị từ -1 đến 1. Một giá trị gần 1 cho thấy mối quan hệ tích cực mạnh mẽ giữa các biến, trong khi giá trị gần -1 chỉ ra mối quan hệ tiêu cực mạnh mẽ, và giá trị gần 0 cho thấy không có mối quan hệ rõ rệt.

Các kết quả đáng chú ý từ ma trận tương quan cho thấy **SEVERITYCODE** (mức độ nghiêm trọng của vụ tai nạn) có mối quan hệ mạnh mẽ với các biến như **INJURIES** (số người bị thương), **SERIOUSINJURIES** (số người bị thương nặng) và **FATALITIES** (số người tử vong), cho thấy mức độ nghiêm trọng của tai nạn có sự liên kết chặt chẽ với các yếu tố này. Cụ thể, khi số lượng người bị thương hoặc tử vong tăng lên, mức độ nghiêm trọng của vụ tai nạn cũng tăng theo. Hơn nữa, **INJURIES** và **SERIOUSINJURIES** có mối tương quan mạnh với nhau, điều này cho thấy các vụ tai nạn nghiêm trọng thường dẫn đến nhiều người bị thương nặng.

Bên cạnh đó, **FATALITIES** cũng có mối quan hệ đáng kể với **INJURIES** và **SERIOUSINJURIES**, cho thấy các vụ tai nạn có người tử vong thường đi kèm với số lượng người bị thương nặng. Tuy nhiên, các biến như **PERSONCOUNT** (số người liên quan đến vụ tai nạn) và **VEHCOUNT** (số phương tiện tham gia) có mối quan hệ yếu với các yếu tố nghiêm trọng, cho thấy số lượng người tham gia hoặc phương tiện không phải yếu tố quyết định chính trong mức độ nghiêm trọng của tai nạn.

Nhờ vào ma trận tương quan này, chúng ta có thể thấy rõ mối quan hệ giữa các yếu tố và hiểu hơn về những yếu tố nào ảnh hưởng trực tiếp đến mức độ nghiêm trọng của các vụ tai nạn giao thông, từ đó hỗ trợ việc phát triển các mô hình dự đoán chính xác hơn và cải thiện công tác quản lý giao thông.

# CHƯƠNG 4: XÂY DỰNG VÀ ĐÁNH GIÁ MÔ HÌNH

## 4.1. Giới thiệu về mô hình và phương pháp sử dụng

Trong bài toán phân loại mức độ nghiêm trọng của tai nạn giao thông, chúng tôi đã lựa chọn ba mô hình học máy phổ biến để giải quyết vấn đề: Random Forest, Logistic Regression và SVM (Support Vector Machine). Mỗi mô hình này có những đặc điểm riêng biệt giúp xử lý các yêu cầu và đặc điểm của bài toán.

Random Forest là một mô hình học máy ensemble mạnh mẽ, được xây dựng từ nhiều cây quyết định. Nó rất hiệu quả trong việc xử lý các vấn đề phân loại và dự đoán, đặc biệt là khi dữ liệu không đồng nhất. Random Forest có khả năng tự động xử lý vấn đề overfitting và thường có độ chính xác cao trong các bài toán thực tế. Mô hình này cũng rất linh hoạt khi làm việc với dữ liệu có nhiều đặc trưng và không cần nhiều điều chỉnh tham số.

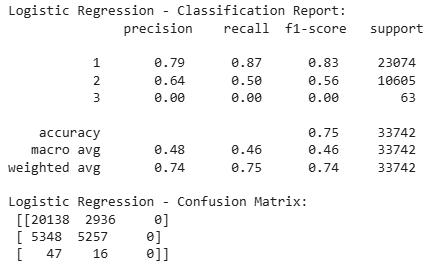
Logistic Regression, mặc dù là một mô hình đơn giản, nhưng lại rất hiệu quả trong các bài toán phân loại nhị phân hoặc đa lớp. Với khả năng mô hình hóa mối quan hệ tuyến tính giữa các đặc trưng và biến mục tiêu, Logistic Regression là một lựa chọn tốt khi muốn có một mô hình dễ hiểu và nhanh chóng. Mô hình này giúp chúng ta xác định xác suất xảy ra của từng lớp và là một công cụ mạnh mẽ trong phân tích dữ liệu.

SVM (Support Vector Machine) là một trong những mô hình phân loại mạnh mẽ, đặc biệt hữu ích khi dữ liệu có nhiều đặc trưng và có khả năng phân tách rõ ràng giữa các lớp. SVM tìm kiếm một siêu phẳng tối ưu để phân chia các lớp, giúp mô hình có độ chính xác cao khi xử lý các dữ liệu phức tạp và không tuyến tính. Mô hình này đặc biệt hiệu quả trong các bài toán có số lượng đặc trưng lớn và phân tách dữ liệu rõ ràng.

## 4.2. HUẤN LUYỆN CÁC MÔ HÌNH

### 4.2.1. MÔ HÌNH LOGISTIC REGRESSION

**Logistic Regression** là một mô hình phân loại tuyến tính được sử dụng để dự đoán mức độ nghiêm trọng của tai nạn giao thông. Mô hình được huấn luyện trên dữ liệu đã được tiền xử lý, bao gồm việc mã hóa các biến phân loại và chia dữ liệu thành tập huấn luyện và kiểm tra. Sau khi huấn luyện, mô hình được đánh giá qua các chỉ số như **precision, recall, f1-score**, và **confusion matrix** để đo lường hiệu quả phân loại các mức độ nghiêm trọng của tai nạn.



Hình 4.1. **Báo cáo phân loại và Ma trận nhầm lẫn của Logistic Regression.**

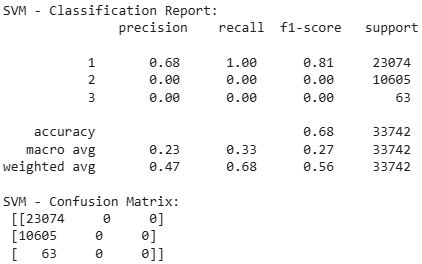
Trong hình ảnh trên, chúng ta có Báo cáo phân loại và Ma trận nhầm lẫn của mô hình Logistic Regression.

Báo cáo phân loại cho thấy mô hình hoạt động khá tốt đối với lớp 1 (tai nạn ít nghiêm trọng), với độ chính xác precision = 0.79 và độ nhạy recall = 0.87, nghĩa là mô hình dự đoán đúng phần lớn các vụ tai nạn ít nghiêm trọng. Tuy nhiên, đối với lớp 2 (tai nạn nghiêm trọng), mô hình chỉ đạt precision = 0.64 và recall = 0.50, cho thấy hiệu suất không tốt đối với lớp này. Đặc biệt, đối với lớp 3 (tai nạn rất nghiêm trọng), mô hình không thể dự đoán đúng, với cả precision và recall bằng 0, điều này chứng tỏ mô hình gặp khó khăn trong việc nhận diện các tai nạn rất nghiêm trọng. Tỷ lệ chính xác tổng thể của mô hình là 75%, nhưng hiệu suất phân loại không đồng đều giữa các lớp.

Ma trận nhầm lẫn cho thấy mô hình phân loại tốt đối với lớp 1, với 20,138 dự đoán đúng, tuy nhiên có 2,936 dự đoán sai. Đối với lớp 2, mô hình dự đoán đúng 5,348 trường hợp nhưng cũng có 5,257 dự đoán sai. Đặc biệt, đối với lớp 3, số lượng dự đoán đúng rất thấp (47), và số lượng dự đoán sai rất ít (16), cho thấy mô hình gần như không phát hiện được các tai nạn rất nghiêm trọng.

### 4.2.2. MÔ HÌNH SUPPORT VECTO MACHINE(SVM)

Sau khi đánh giá mô hình **Logistic Regression**, trong phần này, chúng ta sẽ chuyển sang **Mô hình Support Vector Machine (SVM),** một trong những mô hình học máy mạnh mẽ và hiệu quả trong phân loại. **SVM** hoạt động dựa trên nguyên lý tìm kiếm siêu phẳng tối ưu để phân tách các lớp trong không gian đặc trưng, đặc biệt hữu ích khi dữ liệu có độ phức tạp cao hoặc có nhiều đặc trưng.

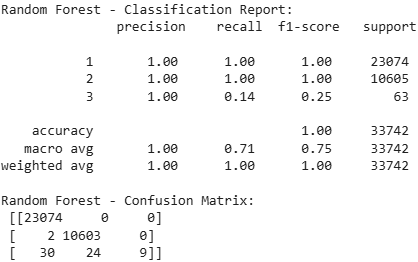


Hình 4.2. **Báo cáo phân loại và Ma trận nhầm lẫn của** SUPPORT VECTO MACHINE

Kết quả từ báo cáo phân loại và ma trận nhầm lẫn của mô hình **SVM** cho thấy mô hình hoạt động không hiệu quả đối với lớp 2 (tai nạn nghiêm trọng) và lớp 3 (tai nạn rất nghiêm trọng). Đối với lớp 1 (tai nạn ít nghiêm trọng), mô hình có **precision = 0.68** và **recall = 1.00**, nghĩa là mô hình phát hiện hoàn toàn các vụ tai nạn ít nghiêm trọng, nhưng độ chính xác không cao. Tuy nhiên, đối với lớp 2 và lớp 3, mô hình không phân loại đúng bất kỳ vụ tai nạn nào, với **precision = recall = 0.00**. Tỷ lệ chính xác tổng thể của mô hình là **68%,** nhưng các chỉ số **macro avg** và **weighted avg** cho thấy mô hình có độ chính xác thấp, đặc biệt là không thể phân biệt tốt giữa các lớp. Nhận xét, mô hình **SVM** hoạt động tốt với lớp 1 nhưng không nhận diện được các vụ tai nạn nghiêm trọng và rất nghiêm trọng, cần cải thiện để phân loại đúng hơn cho các lớp này.

### 4.2.3. MÔ HÌNH RANDOM FOREST

Sau khi đánh giá mô hình **SVM,** chúng ta chuyển sang **mô hình Random Forest,** một trong những mô hình học máy mạnh mẽ và phổ biến trong các bài toán phân loại. **Random Forest** là một mô hình học máy theo phương pháp **ensemble**, trong đó kết hợp nhiều cây quyết định để đưa ra quyết định cuối cùng. Mô hình này rất hiệu quả trong việc xử lý các dữ liệu phức tạp và có khả năng tránh overfitting nhờ việc kết hợp nhiều cây quyết định.

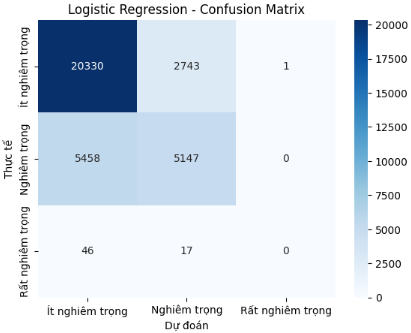


Hình 4.3. **Báo cáo phân loại và Ma trận nhầm lẫn của** SUPPORT VECTO MACHINE

Kết quả từ báo cáo phân loại và ma trận nhầm lẫn của mô hình **Random Forest** cho thấy mô hình hoạt động rất tốt đối với lớp 1 (tai nạn ít nghiêm trọng) và lớp 2 (tai nạn nghiêm trọng), với độ chính xác **precision = 1.00, recall = 1.00** và **f1-score = 1.00** đối với cả hai lớp này. Tuy nhiên, đối với lớp 3 (tai nạn rất nghiêm trọng), mặc dù độ chính xác là **1.00,** mô hình chỉ đạt **recall = 0.14** và **f1-score = 0.25**, cho thấy mô hình gặp khó khăn trong việc phát hiện các vụ tai nạn rất nghiêm trọng. Mặc dù **accuracy** tổng thể của mô hình là **100%,** nhưng các chỉ số **macro avg** và **weighted avg** cho thấy sự mất cân bằng trong phân loại, đặc biệt đối với lớp 3. Ma trận nhầm lẫn cho thấy mô hình phân loại đúng hoàn toàn các vụ tai nạn ít nghiêm trọng và nghiêm trọng, nhưng chỉ phát hiện đúng một số ít các vụ tai nạn rất nghiêm trọng. Như vậy, mô hình **Random Forest** cần cải thiện khả năng phân loại lớp 3 để đạt được kết quả chính xác hơn trên tất cả các mức độ nghiêm trọng của tai nạn.

## 4.3. ĐÁNH GIÁ MÔ HÌNH

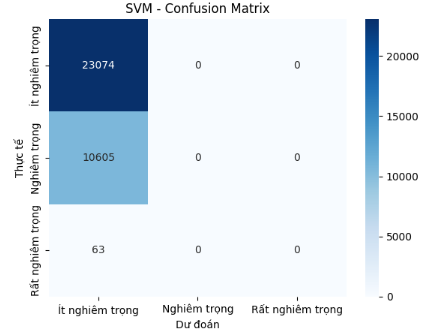
### 4.3.1. ĐÁNH GIÁ LOGISTIC REGRESSION



Hình 4.4. **Logistic Regression - Confusion Matrix.**

Ma trận nhầm lẫn của **Logistic Regression** cho thấy mô hình phân loại rất tốt đối với lớp 1 (tai nạn ít nghiêm trọng), với 20,330 vụ được phân loại đúng, tuy nhiên có 2,743 vụ bị phân loại sai thành lớp 2 (nghiêm trọng) và chỉ 1 vụ bị phân loại sai thành lớp 3 (rất nghiêm trọng). Đối với lớp 2 (tai nạn nghiêm trọng), mô hình phân loại đúng 5,147 vụ, nhưng có tới 5,458 vụ bị phân loại sai thành lớp 1 (ít nghiêm trọng). Lớp 3 (tai nạn rất nghiêm trọng) gặp khó khăn lớn khi chỉ phân loại đúng 17 vụ, và có 46 vụ bị phân loại sai thành lớp 1, không có vụ nào được phân loại chính xác thành lớp 2. Nhìn chung, mô hình **Logistic Regression** hoạt động tốt với các vụ tai nạn ít nghiêm trọng, nhưng hiệu suất giảm rõ rệt khi phân loại các vụ tai nạn nghiêm trọng và rất nghiêm trọng, đặc biệt là lớp 3. Cần có các biện pháp cải thiện khả năng nhận diện lớp 3 để mô hình có thể phân loại chính xác hơn tất cả các mức độ nghiêm trọng của tai nạn.

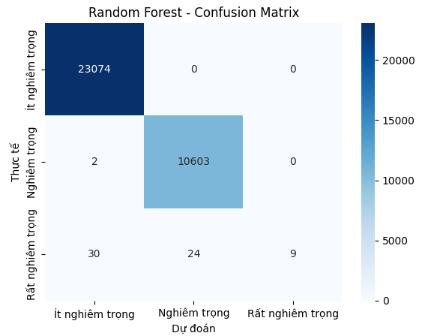
### 4.3.2. ĐÁNH GIÁ SUPPORT VECTO MACHINE(SVM)



Hình 4.5. SUPPORT VECTO MACHINE **- Confusion Matrix.**

Ma trận nhầm lẫn của **SVM** cho thấy mô hình phân loại hoàn toàn chính xác đối với các vụ tai nạn ít nghiêm trọng (lớp 1), với **23,074** vụ được phân loại đúng. Tuy nhiên, mô hình không phân loại chính xác bất kỳ vụ tai nạn nào trong lớp 2 (nghiêm trọng) và lớp 3 (rất nghiêm trọng), vì tất cả các vụ trong hai lớp này đều bị phân loại sai thành lớp 1 (tai nạn ít nghiêm trọng). Điều này cho thấy **SVM** chỉ nhận diện được các tai nạn ít nghiêm trọng một cách chính xác, còn các tai nạn nghiêm trọng và rất nghiêm trọng không được mô hình phân loại đúng. Nhận xét này cho thấy mô hình cần cải thiện khả năng phân biệt và phân loại chính xác các vụ tai nạn nghiêm trọng và rất nghiêm trọng.

### 4.3.2. ĐÁNH GIÁ RANDOM FOREST



Hình 4.6. RANDOM FOREST **- Confusion Matrix.**

Ma trận nhầm lẫn của Random Forest cho thấy mô hình phân loại rất tốt đối với lớp 1 (tai nạn ít nghiêm trọng), với 23,074 vụ được phân loại đúng. Đối với lớp 2 (tai nạn nghiêm trọng), mô hình cũng phân loại đúng 10,603 vụ, nhưng có 2 vụ bị phân loại sai thành lớp 1. Tuy nhiên, đối với lớp 3 (tai nạn rất nghiêm trọng), mô hình chỉ phân loại đúng 9 vụ, trong khi có 24 vụ bị phân loại sai thành lớp 2 và 30 vụ bị phân loại sai thành lớp 1.

Nhận xét: Random Forest hoạt động rất tốt trong việc phân loại các vụ tai nạn ít nghiêm trọng và nghiêm trọng, nhưng gặp khó khăn trong việc phân loại các vụ tai nạn rất nghiêm trọng (lớp 3). Cần cải thiện khả năng phát hiện và phân loại chính xác lớp 3 để mô hình có thể hoạt động hiệu quả hơn trên tất cả các mức độ nghiêm trọng của tai nạn.

# CHƯƠNG 5: KẾT LUẬN VÀ HƯỚNG PHÁT TRIỂN

## 5.1. KẾT LUẬN

Đề tài "Dự đoán tình trạng tắc nghẽn giao thông tại Seattle, dựa trên dữ liệu thời tiết, va chạm, và một số sự kiện khác" đã hoàn thành việc xây dựng các mô hình học máy để dự đoán mức độ tắc nghẽn giao thông tại thành phố Seattle. Với sự kết hợp của các yếu tố dữ liệu quan trọng như thời gian, điều kiện thời tiết, các sự kiện đặc biệt và các vụ va chạm giao thông, mô hình dự đoán đã giúp phân tích và đánh giá tình hình giao thông, từ đó cung cấp các cảnh báo và khuyến nghị hữu ích cho việc điều phối giao thông.

Quá trình phân tích và xây dựng mô hình đã sử dụng các thuật toán học máy như Logistic Regression, Random Forest và Support Vector Machine (SVM). Sau khi huấn luyện và đánh giá các mô hình, nhóm nhận thấy rằng mặc dù mô hình Random Forest cho kết quả phân loại tốt nhất với mức độ nghiêm trọng của các vụ va chạm giao thông, nhưng vẫn cần cải thiện khả năng phân loại các tình huống tắc nghẽn nghiêm trọng, đặc biệt là những tình huống phát sinh trong điều kiện thời tiết xấu hoặc khi có sự kiện đặc biệt. Mô hình Logistic Regression và SVM cũng đã đạt được những kết quả khả quan, nhưng không thể cạnh tranh với Random Forest về khả năng phân loại các lớp tắc nghẽn giao thông phức tạp hơn.

Trong quá trình phân tích, nhóm đã nhận diện được các yếu tố tác động lớn đến tình trạng tắc nghẽn giao thông. Thời gian giao dịch cho thấy các vụ tắc nghẽn giao thông chủ yếu xảy ra vào giờ cao điểm buổi sáng và cuối giờ chiều, trong khi các yếu tố như thời tiết xấu, sự kiện đặc biệt và tai nạn giao thông có ảnh hưởng rõ rệt đến tình trạng tắc nghẽn. Việc dự đoán chính xác thời điểm và mức độ tắc nghẽn giúp cơ quan quản lý giao thông có thể đưa ra các biện pháp điều phối hiệu quả hơn, hạn chế tình trạng ùn tắc và đảm bảo sự lưu thông thông suốt cho người tham gia giao thông.

Đặc biệt, phân tích theo địa lý và quốc gia cho thấy các khoản đóng góp và sự tham gia từ nhiều quốc gia khác nhau, thể hiện sự ủng hộ và quan tâm của cộng đồng quốc tế đối với các chiến lược cải thiện giao thông tại các thành phố lớn. Sự kết hợp giữa các yếu tố thời gian, sự kiện, và dữ liệu va chạm giao thông là rất quan trọng để xây dựng một hệ thống dự báo giao thông chính xác và hiệu quả.

## 5.2. HƯỚNG PHÁT TRIỂN TRONG TƯƠNG LAI

Trong tương lai, đề tài "Dự đoán tình trạng tắc nghẽn giao thông tại Seattle" có thể được phát triển theo nhiều hướng thực tế để nâng cao hiệu quả và khả năng ứng dụng của mô hình trong việc quản lý giao thông. Một trong những hướng phát triển quan trọng và thiết thực là cải thiện mô hình dự đoán. Các mô hình học máy hiện tại như Logistic Regression, Random Forest và SVM đã cho kết quả khả quan, nhưng chúng vẫn có thể được cải thiện để đạt độ chính xác cao hơn trong các tình huống giao thông phức tạp. Việc áp dụng các mô hình học sâu (deep learning) như mạng nơ-ron hồi tiếp (RNN) hoặc mạng nơ-ron tích chập (CNN) sẽ giúp mô hình có khả năng nhận diện các mẫu tắc nghẽn giao thông phức tạp và biến đổi theo thời gian và không gian. Những mô hình này có thể giúp xử lý các dữ liệu thời gian và dữ liệu liên tục, từ đó dự đoán chính xác hơn các tình huống tắc nghẽn và biến động nhanh chóng trong hệ thống giao thông đô thị.

Bên cạnh việc cải thiện mô hình dự đoán, một yếu tố quan trọng để nâng cao hiệu quả của đề tài là tích hợp dữ liệu thời gian thực từ các hệ thống giao thông thông minh. Việc sử dụng cảm biến giao thông, hệ thống giám sát giao thông thông minh và dữ liệu GPS từ các phương tiện sẽ giúp mô hình cập nhật tình trạng giao thông liên tục và chính xác hơn. Các thông tin thời gian thực sẽ cung cấp các chỉ số về tình hình giao thông tại các điểm nút, lối vào các tuyến đường chính và các khu vực có nguy cơ tắc nghẽn cao. Điều này sẽ giúp các cơ quan chức năng đưa ra các cảnh báo sớm, điều phối giao thông kịp thời, và tạo ra các lộ trình thay thế cho người tham gia giao thông trong thời gian thực.

Một trong những ứng dụng quan trọng nhất của mô hình dự đoán này là việc ứng dụng vào thực tế trong điều phối giao thông. Với những dự đoán chính xác về tình trạng tắc nghẽn, các hệ thống giao thông thông minh có thể tự động điều chỉnh các tín hiệu đèn giao thông, phân luồng giao thông hợp lý, và thông báo cho người tham gia giao thông về tình trạng tắc nghẽn tại các khu vực cụ thể. Các biện pháp như tăng cường giám sát và quản lý giao thông tự động sẽ không chỉ giúp giảm thiểu tình trạng ùn tắc mà còn tối ưu hóa việc phân bổ các nguồn lực giao thông, giúp tăng hiệu quả di chuyển và giảm thiểu thời gian chờ đợi của người tham gia giao thông. Điều này sẽ nâng cao chất lượng sống cho cư dân và tạo ra một môi trường giao thông thông minh hơn.

Ngoài ra, một hướng phát triển khác là mở rộng và ứng dụng mô hình vào các thành phố lớn khác. Đặc biệt là những thành phố có mật độ dân số cao và tình trạng giao thông phức tạp như **TP.HCM** và **Hà Nội**. Việc áp dụng mô hình dự đoán này tại các thành phố này sẽ không chỉ giúp giải quyết vấn đề tắc nghẽn mà còn giúp cải thiện việc quản lý giao thông đô thị thông qua các biện pháp sáng tạo và thông minh. Hệ thống dự đoán tắc nghẽn có thể giúp các cơ quan chức năng xây dựng các chiến lược giao thông bền vững và ứng phó với những biến động trong điều kiện giao thông.

Tóm lại, việc cải thiện và mở rộng mô hình dự đoán tắc nghẽn giao thông sẽ mang lại nhiều lợi ích thiết thực không chỉ cho Seattle mà còn cho các thành phố lớn khác. Khi được tích hợp với các hệ thống giao thông thông minh, mô hình này có thể giúp tối ưu hóa quản lý giao thông, giảm tắc nghẽn, và nâng cao trải nghiệm di chuyển của người dân. Điều này không chỉ góp phần cải thiện tình trạng giao thông mà còn giúp phát triển một hệ thống giao thông bền vững và thông minh hơn trong tương lai.

nguyện, minh bạch tài chính và vận hành chiến dịch một cách hiệu quả, nhanh chóng và có chiều sâu hơn trong tương lai.

# TÀI LIỆU THAM KHẢO

[1]. Nguyễn Văn Huy, Trần Quốc Bình (2021), *Ứng dụng phân tích dữ liệu lớn trong dự đoán tình trạng tắc nghẽn giao thông, Tạp chí Giao thông và Vận tải*, trang 55-62.

[2]. Lê Trung Hiếu (2022), *Ứng dụng học máy trong dự đoán tắc nghẽn giao thông theo thời gian và không gian, Tạp chí Khoa học Giao thông – Đại học Đà Nẵng*, Tập 22, Số 7.

[3]. Lê Thị Thanh Nhàn (2021), *Phân tích và dự báo tắc nghẽn giao thông với Python, Nhà xuất bản Đại học Quốc gia Hà Nội.*