**PHIẾU ĐĂNG KÝ ĐỀ TÀI**

|  |
| --- |
| **1. Tên đề tài:**  Ứng dụng học máy trong phân tích dữ liệu bán lẻ: Dự báo doanh số và Tối ưu hóa chiến lược kinh doanh |
| **2. Thông tin nhóm sinh viên:**  **Sinh viên 1 (Nhóm trưởng):**   * **Họ và tên:** Nguyễn Khang * **Mã sinh viên:** 22174600062 * **Điện thoại:** 0862648906 * **Email:** nkhang.dhkl16a1hn@sv.uneti.edu.vn   **Sinh viên 2:**   * **Họ và tên:** Nguyễn Văn Hoàng * **Mã sinh viên:** 22174600023 * **Điện thoại:** 0365586740 * **Email:** nvhoang.dhkl16a1hn@sn.uneti.edu.vn   **Sinh viên 3:**   * **Họ và tên:** Lê Thị Lan * **Mã sinh viên:** 22174600093 * **Điện thoại:** 0583467602 * **Email:** ltlan.dhkl16a1hn@sv.uneti.edu.vn   **Sinh viên 4:**   * **Họ và tên:** Phùng Thị Linh * **Mã sinh viên:** 22174600001 * **Điện thoại:** 0329869246 * **Email:** ptlinh.dhkl16a1hn@sv.uneti.edu.vn   **Sinh viên 5:**   * **Họ và tên:** Nguyễn Thị Thanh Hoa * **Mã sinh viên:** 22174600052 * **Điện thoại:** 0352307901 * **Email:** ntthoa.dhkl16a1hn@sv.uneti.edu.vn |
| **3. Tóm tắt nội dung đề tài:**  Đề tài “**Ứng dụng học máy trong phân tích dữ liệu bán lẻ: Dự báo doanh số và Tối ưu hóa chiến lược kinh doanh**” tập trung vào việc ứng dụng các kỹ thuật học máy để phân tích dữ liệu bán lẻ từ bộ dữ liệu Retail Sales Dataset (Kaggle) được lấy từ link: [https://www.kaggle.com/datasets/mohammadtalib786/retail-sales-dataset?](https://www.kaggle.com/datasets/mohammadtalib786/retail-sales-dataset?%20), bao gồm 1.000 giao dịch với thông tin về khách hàng (độ tuổi, giới tính), sản phẩm (danh mục, giá), và giao dịch (số lượng, tổng chi tiêu, ngày mua). Mục tiêu chính là:   * Phân tích hành vi tiêu dùng: Khai thác dữ liệu để nhận diện xu hướng mua sắm theo thời gian, độ tuổi, giới tính, và danh mục sản phẩm. * Phân nhóm khách hàng: Sử dụng thuật toán phân cụm (K-Means Clustering) để chia khách hàng thành các phân khúc dựa trên đặc điểm hành vi, từ đó hỗ trợ xây dựng chiến lược tiếp thị và chăm sóc khách hàng cá nhân hóa. * Dự báo doanh số: Áp dụng các mô hình học máy như Hồi quy tuyến tính (Linear Regression), Rừng ngẫu nhiên (Random Forest) giúp doanh nghiệp lập kế hoạch kinh doanh hiệu quả. * Tối ưu hóa chiến lược kinh doanh: Dựa trên kết quả phân tích và dự báo, đề xuất các chiến lược về giá, khuyến mãi, quản lý hàng tồn kho, và tiếp cận khách hàng nhằm tăng doanh thu và giảm lãng phí.   Quá trình thực hiện bao gồm: tiền xử lý dữ liệu (làm sạch, chuẩn hóa), phân tích mô tả, trực quan hóa dữ liệu bằng các biểu đồ (histogram, boxplot, heatmap), và xây dựng các mô hình học máy. Kết quả dự kiến cung cấp các insight hữu ích và mô hình dự báo chính xác, hỗ trợ doanh nghiệp bán lẻ ra quyết định dựa trên dữ liệu. |

|  |  |
| --- | --- |
|  | *Ngày 11 tháng 4 năm 2025* |
|  | **Nhóm trưởng** |
|  | **Nguyễn Khang** |

ĐỀ CƯƠNG CHI TIẾT ĐỀ TÀI

|  |
| --- |
| **1. Tên đề tài:** |
| Ứng dụng học máy trong phân tích dữ liệu bán lẻ: Dự báo doanh số và Tối ưu hóa chiến lược kinh doanh  **2. Mục tiêu đề tài:**  Mục tiêu chính của đề tài là xây dựng quy trình phân tích và khai thác dữ liệu bán lẻ bằng các kỹ thuật học máy hiện đại nhằm hỗ trợ doanh nghiệp trong việc dự báo doanh số và tối ưu hóa chiến lược kinh doanh. Cụ thể, đề tài hướng tới:   * Tiền xử lý và trực quan hóa dữ liệu bán lẻ thực tế từ bộ dữ liệu Retail Sales Dataset (Kaggle), bao gồm việc làm sạch dữ liệu, xử lý giá trị thiếu, chuẩn hóa và mã hóa các biến. * Phân tích hành vi mua hàng của khách hàng dựa trên các đặc trưng như thời gian giao dịch, loại sản phẩm, số lượng mua, và tổng giá trị đơn hàng. * Phân khúc khách hàng bằng các thuật toán phân cụm như K-Means và Spectral Clustering, nhằm xây dựng các nhóm khách hàng có hành vi tiêu dùng tương đồng để phục vụ mục tiêu tiếp thị và chăm sóc khách hàng cá nhân hóa. * Đưa ra các khuyến nghị chiến lược kinh doanh dựa trên kết quả mô hình và phân tích dữ liệu, bao gồm chiến lược giá, chương trình khuyến mãi, và chính sách phân phối nhằm nâng cao hiệu quả hoạt động và lợi nhuận. |
| **3. Tổng quan tình hình nghiên cứu thuộc lĩnh vực đề tài:**  Trong những năm gần đây, việc áp dụng học máy trong lĩnh vực bán lẻ đã trở thành xu hướng nghiên cứu và ứng dụng thực tiễn quan trọng. Nhiều nghiên cứu tập trung vào khai thác dữ liệu bán lẻ để đưa ra các dự báo chính xác và tối ưu hóa các chiến lược kinh doanh. Theo bài báo của Chen et al. (2020), các mô hình học máy như Random Forest và XGBoost đã chứng minh được hiệu quả trong việc phân loại và dự báo doanh số bán hàng nhờ khả năng xử lý dữ liệu phi tuyến và đa chiều.  Ngoài ra, việc phân khúc khách hàng cũng đóng vai trò quan trọng trong tối ưu hóa chiến lược kinh doanh. Các thuật toán phân cụm như K-Means và Spectral Clustering đã được sử dụng để xác định các nhóm khách hàng dựa trên hành vi mua sắm, từ đó giúp doanh nghiệp cá nhân hóa các chiến dịch tiếp thị (Jain, 2010).  Tổng quan các nghiên cứu trên cho thấy việc áp dụng các mô hình học máy trong phân tích dữ liệu bán lẻ là hoàn toàn khả thi và mang lại nhiều giá trị thực tiễn. Tuy nhiên, việc lựa chọn mô hình phù hợp, xử lý dữ liệu hiệu quả và kết hợp nhiều kỹ thuật khác nhau vẫn là thách thức đối với nhà nghiên cứu và doanh nghiệp. |
| **4. Nội dung đề tài:**  Đề tài được triển khai theo các nội dung chính sau:  **Tìm hiểu và phân tích bộ dữ liệu bán lẻ:** Khám phá và tìm hiểu cấu trúc của bộ dữ liệu Retail Sales Dataset từ Kaggle. Mô tả các biến dữ liệu bao gồm: mã đơn hàng, ngày giao dịch, mã khách hàng, mã sản phẩm, loại sản phẩm, số lượng bán, doanh thu,... Thực hiện phân tích các đặc điểm định tính và định lượng để hiểu rõ về nội dung và ý nghĩa của từng thuộc tính trong dữ liệu.  **Tiền xử lý và làm sạch dữ liệu:** Thực hiện các bước tiền xử lý để chuẩn bị dữ liệu cho phân tích và mô hình học máy:   * Xử lý giá trị thiếu, loại bỏ dữ liệu trùng lặp hoặc sai lệch. * Mã hóa các biến phân loại bằng các kỹ thuật như Label Encoding và One-hot Encoding. * Chuẩn hóa dữ liệu để đảm bảo các mô hình học máy hoạt động hiệu quả. * Tạo các đặc trưng mới (feature engineering) như: tổng chi tiêu theo khách hàng, doanh thu theo tuần/tháng, khoảng thời gian giữa các lần mua,...   **Phân tích thăm dò và trực quan hóa dữ liệu (Exploratory Data Analysis - EDA):** Phân tích thống kê mô tả và trực quan hóa dữ liệu nhằm phát hiện xu hướng và mô hình tiềm ẩn:   * Trực quan hóa xu hướng bán hàng theo thời gian, loại sản phẩm, khu vực địa lý (nếu có). * Phân tích mối quan hệ giữa số lượng bán và doanh thu. * Phân tích hành vi mua sắm theo nhóm khách hàng để xác định các yếu tố ảnh hưởng đến doanh số.   **Phân cụm khách hàng (Customer Segmentation)**: Xây dựng đặc trưng khách hàng theo mô hình RFM (Recency - Frequency - Monetary), sau đó áp dụng các thuật toán phân cụm như K-Means, Spectral Clustering để chia khách hàng thành các nhóm có hành vi tương đồng. Đánh giá chất lượng phân cụm bằng các chỉ số như Silhouette Score và trực quan hóa kết quả bằng biểu đồ.  **Xây dựng mô hình dự báo doanh số:**   * Áp dụng các mô hình học máy như Linear Regression, Random Forest, XGBoost để dự báo doanh số. * Thực hiện phân chia tập huấn luyện và kiểm tra, đánh giá mô hình bằng các chỉ số MAE, RMSE và R². * Bên cạnh đó, triển khai các mô hình dự báo chuỗi thời gian như ARIMA, Prophet hoặc LSTM để dự đoán doanh số theo ngày hoặc theo tháng.   **Đề xuất chiến lược kinh doanh**: Dựa trên kết quả phân tích và mô hình, đề xuất các chiến lược nhằm hỗ trợ ra quyết định:   * Đưa ra gợi ý hành động cho từng nhóm khách hàng như chăm sóc khách hàng thân thiết, tiếp thị lại với khách hàng có nguy cơ rời bỏ. * Dự báo tình hình doanh thu trong ngắn và trung hạn để hỗ trợ lập kế hoạch sản xuất và bán hàng. * Xây dựng bảng dashboard trực quan thể hiện các chỉ số kinh doanh quan trọng như tổng doanh thu, tăng trưởng theo tháng, tỉ lệ giữ chân khách hàng,... nhằm phục vụ việc giám sát và ra quyết định của nhà quản lý. |
| **5. Phương pháp thực hiện:**  Trong quá trình thực hiện đề tài, dựa trên kiến thức đã học trong môn Đồ án 2 nhóm dự kiến sử dụng các mô hình và phương pháp trong lĩnh vực khoa học dữ liệu và học máy, phù hợp với từng mục tiêu cụ thể của đề tài.  Phân nhóm khách hàng (Customer Segmentation): Để thực hiện phân nhóm khách hàng dựa trên hành vi mua sắm, nhóm sử dụng các thuật toán phân cụm như K-Means và Spectral Clustering. Thuật toán K-Means sẽ được áp dụng trên các đặc trưng hành vi khách hàng theo mô hình RFM (Recency - tần suất gần nhất mua hàng, Frequency - số lần mua, Monetary - tổng số tiền chi tiêu). Spectral Clustering sẽ giúp đánh giá và cải thiện chất lượng phân cụm trong trường hợp dữ liệu có cấu trúc phức tạp, không thể phân tách tốt bằng K-Means thông thường.  Dự báo doanh số bằng mô hình học máy: Để dự đoán doanh số bán hàng dựa trên các yếu tố như loại sản phẩm, thời gian, khu vực,... nhóm sẽ sử dụng các mô hình hồi quy như Linear Regression, Random Forest và XGBoost. Linear Regression giúp xây dựng mối quan hệ tuyến tính giữa các biến đầu vào và doanh số, trong khi Random Forest và XGBoost là những mô hình mạnh trong xử lý dữ liệu phi tuyến và nhiều chiều sẽ giúp nâng cao độ chính xác trong dự báo. |
| **6. Phân công công việc (dự kiến):**   |  |  |  |  | | --- | --- | --- | --- | | **STT** | **Họ và tên** | **Mã sinh viên** | **Nội dung công việc được phân công** | | **1** | Nguyễn Khang | 22174600062 | Xây dựng các mô hình dự báo doanh số: Linear Regression, Random Forest, XGBoost.  Chia dữ liệu thành tập huấn luyện và kiểm tra, đánh giá mô hình bằng MAE, RMSE, R².  So sánh hiệu quả các mô hình và chọn ra mô hình tốt nhất.  Điều phối và giám sát toàn bộ quá trình thực hiện đề tài đảm bảo tiến độ thực hiện theo đúng kế hoạch. Góp ý và hỗ trợ các thành viên giải quyết các vấn đề kỹ thuật phát sinh và đảm bảo sự phối hợp nhịp nhàng giữa các phần việc. | | **2** | Lê Thị Lan | 22174600093 | Xử lý các giá trị thiếu, dữ liệu trùng lặp hoặc sai lệch.  Mã hóa các biến phân loại bằng Label Encoding hoặc One-hot Encoding.  Chuẩn hóa dữ liệu đầu vào phục vụ huấn luyện mô hình.  Tạo các biến đặc trưng nếu cần: tổng chi tiêu theo khách hàng, doanh thu theo thời gian,...  Tổng hợp báo cáo, thuyết trình. | | **3** | Nguyễn Văn Hoàng | 22174600023 | Xây dựng tập đặc trưng RFM (Recency, Frequency, Monetary) cho từng khách hàng.  Áp dụng thuật toán phân cụm K-Means, Spectral Clustering.  Đánh giá chất lượng phân cụm bằng Silhouette Score.  Trực quan hóa kết quả phân nhóm khách hàng và mô tả đặc điểm từng nhóm.  Hỗ trợ thuyết trình | | **4** | Phùng Thị Linh | 22174600001 | Thực hiện phân tích thăm dò dữ liệu (EDA):  Trực quan hóa xu hướng bán hàng theo thời gian, loại sản phẩm, khu vực.  Phân tích mối quan hệ giữa số lượng bán và doanh thu.  Phân tích hành vi mua sắm của khách hàng theo nhóm (giới tính, độ tuổi nếu có).  Phân tích và đề xuất chiến lược kinh doanh dựa trên kết quả phân tích.  Thiết kế slide và chuẩn bị nội dung thuyết trình, hỗ trợ thuyết trình. | | **5** | Nguyễn Thị Thanh Hoa | 22174600052 | Tìm hiểu bộ dữ liệu “Retail Sales Dataset” từ Kaggle:  Mô tả cấu trúc dữ liệu và ý nghĩa các biến: mã đơn hàng, mã khách hàng, loại sản phẩm, doanh thu,...  Phân tích đặc điểm định tính và định lượng của dữ liệu.  Chuẩn bị báo cáo mô tả sơ bộ dữ liệu và nhận xét ban đầu.  Tổng hợp thông tin từ các insight, bất thường, xu hướng từ dữ liệu trước đó, đưa ra các khuyến nghị.  Duyệt, soát lỗi báo cáo, hỗ trợ thuyết trình. | |
| **7. Dự kiến kết quả đạt được:**  Thông qua việc thực hiện đề tài, nhóm dự kiến đạt được một số kết quả quan trọng. Trước hết, nhóm sẽ hiểu rõ cấu trúc và đặc điểm của bộ dữ liệu bán lẻ, từ đó xác định được các biến quan trọng phục vụ phân tích và xây dựng mô hình. Dữ liệu sau khi được tiền xử lý sẽ đảm bảo sạch, nhất quán và sẵn sàng cho việc áp dụng các thuật toán học máy. Đồng thời, nhóm sẽ trực quan hóa được xu hướng bán hàng, hành vi tiêu dùng và mối quan hệ giữa các yếu tố trong dữ liệu thông qua các biểu đồ trực quan như histogram, boxplot, heatmap và line chart.  Bên cạnh đó, nhóm sẽ áp dụng các kỹ thuật phân cụm như K-Means và Spectral Clustering để phân nhóm khách hàng dựa trên đặc điểm hành vi (RFM), từ đó xây dựng được các chân dung khách hàng cụ thể nhằm hỗ trợ hoạt động marketing hiệu quả hơn. Ngoài ra, nhóm cũng sẽ xây dựng được các mô hình dự báo doanh số sử dụng thuật toán Linear Regression, Random Forest, XGBoost nhằm hỗ trợ doanh nghiệp trong việc dự đoán xu hướng tiêu dùng và lập kế hoạch kinh doanh.  Cuối cùng, nhóm sẽ đề xuất được một số chiến lược kinh doanh cụ thể cho từng nhóm khách hàng như chăm sóc khách hàng trung thành, giữ chân khách hàng có nguy cơ rời bỏ, hay tập trung vào sản phẩm bán chạy theo mùa vụ. Đồng thời, nhóm sẽ xây dựng một bảng dashboard thể hiện các chỉ số kinh doanh quan trọng một cách trực quan nhằm hỗ trợ các nhà quản lý trong việc ra quyết định nhanh chóng và chính xác. |

|  |  |
| --- | --- |
|  | *Ngày 11 tháng 4 năm 2025* |
|  | **Nhóm trưởng**  **Nguyễn Khang** |
|  |  |