TỔNG LIÊN ĐOÀN LAO ĐỘNG VIỆT NAM

**TRƯỜNG ĐẠI HỌC TÔN ĐỨC THẮNG**

**KHOA CÔNG NGHỆ THÔNG TIN**



**ĐỒ ÁN CUỐI KỲ**

**Business Intelligence**

Người hướng dẫn: **ThS DƯƠNG HỮU PHÚC**

Người thực hiện: **Trần Chí Huy - 52200213**

**Nguyễn Tuấn Khanh - 52200201**

**Nguyễn Lê Bảo Ngọc - 52200073**

Khoá  **: 26**

**THÀNH PHỐ HỒ CHÍ MINH, NĂM 2024**

TỔNG LIÊN ĐOÀN LAO ĐỘNG VIỆT NAM

**TRƯỜNG ĐẠI HỌC TÔN ĐỨC THẮNG**

**KHOA CÔNG NGHỆ THÔNG TIN**



**ĐỒ ÁN CUỐI KỲ**

**Business Intelligence**

Người hướng dẫn: **ThS DƯƠNG HỮU PHÚC**

Người thực hiện: **Trần Chí Huy - 52200213**

**Nguyễn Tuấn Khanh - 52200201**

**Nguyễn Lê Bảo Ngọc - 52200073**

Khoá  **: 26**

**THÀNH PHỐ HỒ CHÍ MINH, NĂM 2024**

LỜI CẢM ƠN

Em muốn gửi lời cảm ơn chân thành đến thầy Dương Hữu Phúc - người đã giúp đỡ, dạy dỗ và hướng dẫn em trong môn Business Intelligence. Những kiến thức và kinh nghiệm mà thầy chia sẻ đã giúp em hiểu và tiếp thu môn học một cách tốt hơn. Em cảm thấy rất biết ơn và sẽ luôn tận dụng những gì đã được học để phát triển bản thân và đóng góp cho xã hội.

Xin chân thành cảm ơn!

**CÔNG TRÌNH ĐƯỢC HOÀN THÀNH**

**TẠI TRƯỜNG ĐẠI HỌC TÔN ĐỨC THẮNG**

Tôi xin cam đoan đây là công trình nghiên cứu của riêng tôi và được sự hướng dẫn khoa học của Th.S Dương Hữu Phúc; Các nội dung nghiên cứu, kết quả trong đề tài này là trung thực và chưa công bố dưới bất kỳ hình thức nào trước đây. Những số liệu trong các bảng biểu phục vụ cho việc phân tích, nhận xét, đánh giá được chính tác giả thu thập từ các nguồn khác nhau có ghi rõ trong phần tài liệu tham khảo.

Ngoài ra, trong luận văn còn sử dụng một số nhận xét, đánh giá cũng như số liệu của các tác giả khác, cơ quan tổ chức khác đều có trích dẫn và chú thích nguồn gốc.

**Nếu phát hiện có bất kỳ sự gian lận nào tôi xin hoàn toàn chịu trách nhiệm về nội dung luận văn của mình.** Trường đại học Tôn Đức Thắng không liên quan đến những vi phạm tác quyền, bản quyền do tôi gây ra trong quá trình thực hiện (nếu có).

TP. Hồ Chí Minh, ngày 01 tháng 12 năm 2024

Tác giả

(ký tên và ghi rõ họ tên)

**Mục Lục**

[CHƯƠNG 1 – TỔNG QUAN 3](#_Toc184091540)

[1.1 Giới thiệu đề tài 3](#_Toc184091541)

[1.2 Đặc tả hệ thống 3](#_Toc184091542)

[1. Phạm vi hệ thống 4](#_Toc184091543)

[2. Chức năng hệ thống 4](#_Toc184091544)

[3. Giao diện người dùng 4](#_Toc184091545)

[4. Luồng hoạt động của hệ thống 5](#_Toc184091546)

[5. Yêu cầu kỹ thuật 5](#_Toc184091547)

[6. Yêu cầu phi chức năng 6](#_Toc184091548)

[7. Kết quả đầu ra 6](#_Toc184091549)

[CHƯƠNG 2 - PHÂN TÍCH HỆ THỐNG 6](#_Toc184091550)

[2.1 Cơ sở dữ liệu 6](#_Toc184091551)

[2.2 Huấn Luận mô hình học Máy 9](#_Toc184091552)

[2.3 Dự Báo Kết Quả 10](#_Toc184091553)

[1. Inflation (Lạm phát) 10](#_Toc184091554)

[2. GDP Growth (Tăng trưởng GDP) 10](#_Toc184091555)

[3. Marketing Effect (Hiệu quả từ chi phí marketing) 11](#_Toc184091556)

[4. Innovation Effect (Hiệu quả từ đổi mới) 11](#_Toc184091557)

[5. Exchange Rate (Tỷ giá hối đoái) 11](#_Toc184091558)

[6. Interest Rate (Lãi suất) 11](#_Toc184091559)

[7. Consumer Trend (Xu hướng người tiêu dùng) 12](#_Toc184091560)

[8. Competition Index (Chỉ số cạnh tranh) 12](#_Toc184091561)

[Giá trị giả định: CompetitionIndex = 5. 12](#_Toc184091562)

[9. Tax Rate (Thuế suất) 12](#_Toc184091563)

[10. Marketing Expense (Chi phí marketing) 12](#_Toc184091564)

[11. Innovation Index (Chỉ số đổi mới) 12](#_Toc184091565)

[CHƯƠNG 3: SƠ ĐỒ TỔNG QUAN 13](#_Toc184091566)

[3.1 Use case và đặc tả use case: 13](#_Toc184091567)

[3.2 ERD và đặc tả ERD: 15](#_Toc184091568)

[3.3 DFD và đặc tả DFD 17](#_Toc184091569)

[CHƯƠNG 4 - THIẾT KẾ 20](#_Toc184091570)

[4.1 Nguyên tắc thiết kế giao diện 20](#_Toc184091571)

[4.2 Chi tiết giao diện 20](#_Toc184091572)

[4.2.1 Màn hình Nhập Dữ liệu 20](#_Toc184091573)

[4.2.2 Màn hình Tiền xử lý Dữ liệu 21](#_Toc184091574)

[4.2.3 Màn hình Dự báo 22](#_Toc184091575)

[4.2.4 Màn hình Kết quả và Trực quan hóa 22](#_Toc184091576)

[4.3 Nguyên tắc thiết kế màu sắc 23](#_Toc184091577)

[4.4 Mô tả luồng giao diện 23](#_Toc184091578)

[4.5 Mô hình giao diện 23](#_Toc184091579)

[CHƯƠNG 5 – KIỂM THỬ VÀ ĐÁNH GIÁ 24](#_Toc184091580)

[5.1 Mục đích kiểm thử 24](#_Toc184091581)

[5.2 Kiểm thử chức năng 24](#_Toc184091582)

[5.3 Kiểm thử hiệu suất 25](#_Toc184091583)

[5.4 Đánh giá mô hình học máy 25](#_Toc184091584)

[5.5 Kết quả kiểm thử 26](#_Toc184091585)

[5.6 Kết luận 27](#_Toc184091586)

[TÀI LIỆU THAM KHẢO 28](#_Toc184091587)

# CHƯƠNG 1 – TỔNG QUAN

* 1. Giới thiệu đề tài

Trong bối cảnh nền kinh tế ngày càng cạnh tranh và phát triển, việc dự đoán chính xác doanh thu tương lai đóng vai trò quan trọng đối với các doanh nghiệp. Dự đoán này không chỉ hỗ trợ trong việc đưa ra các quyết định chiến lược mà còn giúp tối ưu hóa nguồn lực và nâng cao hiệu quả hoạt động kinh doanh.

Dự án "Ứng dụng WinForm dự báo tình hình kinh doanh của doanh nghiệp" được phát triển nhằm cung cấp một công cụ hỗ trợ phân tích và dự báo doanh thu dựa trên dữ liệu kinh doanh lịch sử. Ứng dụng tích hợp các chức năng chính như tiền xử lý dữ liệu cơ bản, chọn lọc dữ liệu, áp dụng các thuật toán học máy, và hiển thị biểu đồ trực quan về xu hướng doanh thu trong tương lai.

Mục tiêu của đề tài là xây dựng một ứng dụng dễ sử dụng, đáp ứng yêu cầu của các nhà quản lý doanh nghiệp trong việc đưa ra các dự báo chính xác và hiệu quả, đồng thời áp dụng các công nghệ hiện đại để xử lý và phân tích dữ liệu.

* 1. Đặc tả hệ thống

1. Phạm vi hệ thống

**Mục tiêu**:  
Hệ thống được thiết kế để hỗ trợ dự báo tình hình kinh doanh của cửa hàng thông qua việc phân tích dữ liệu lịch sử và hiển thị kết quả dự báo dưới dạng biểu đồ trực quan.

**Đối tượng sử dụng**:  
Bất kỳ ai có nhu cầu dự báo tình hình kinh doanh cho cửa hàng của mình, từ chủ cửa hàng nhỏ lẻ đến các nhà quản lý doanh nghiệp.

**2. Chức năng hệ thống**

1. **Tiền xử lý dữ liệu**:
   * Nhập dữ liệu từ file Excel (.csv, .xlsx).
   * Kiểm tra và làm sạch dữ liệu (xử lý giá trị null, dữ liệu sai định dạng).
   * Chuyển đổi dữ liệu về định dạng phù hợp cho mô hình học máy.
2. **Chọn lọc dữ liệu**:
   * Cho phép người dùng chọn các cột dữ liệu cần thiết cho việc dự báo (doanh thu, thời gian, sản phẩm, v.v.).
3. **Dự đoán bằng Machine learning**
   * Sử dụng mô hình FastTree Regression
   * Tự động huấn luyện mô hình dựa trên dữ liệu đã chọn.
   * Đánh giá độ chính xác của mô hình (R² score).
4. **Hiển thị kết quả**:
   * Hiển thị biểu đồ dự báo doanh thu tương lai dựa trên mô hình học máy.
   * Cho phép người dùng chọn mốc thời gian để dự báo.

**3. Giao diện người dùng**

Hệ thống bao gồm 3 form chính:

* **Form 1**: Tiền xử lý dữ liệu
  + Giao diện cho phép người dùng đăng tải file Excel và thực hiện các bước làm sạch dữ liệu.
* **Form 2**: Chọn lọc dữ liệu
  + Giao diện cho phép người dùng chọn các cột dữ liệu cần thiết cho quá trình học máy.
* **Form 3**: Học máy và trực quan hóa dữ liệu
  + Giao diện cho phép người dùng chỉnh sửa thời gian tới tương lai để có thể trực quan được biểu đồ chart.

**4. Luồng hoạt động của hệ thống**

1. Người dùng tải file Excel lên từ Form 1.
2. Hệ thống tiến hành tiền xử lý dữ liệu (xóa dữ liệu lỗi, chuyển đổi định dạng).
3. Chuyển sang Form 2, người dùng chọn cột dữ liệu cần xử lý.
4. Người dùng nhấn nút "Học máy" trên Form 3 để hệ thống tự động huấn luyện mô hình dựa trên dữ liệu đã chọn.
5. Người dùng chọn mốc thời gian cần dự báo, nhấn nút "Trực quan" để xem biểu đồ dự báo doanh thu.

**5. Yêu cầu kỹ thuật**

* **Ngôn ngữ lập trình**: C#, sử dụng Windows Forms.
* **Thư viện và công cụ**:
  + **OfficeOpenXml**: Đọc và ghi file Excel (.xlsx).
  + **System.Data** và **System.Linq**: Xử lý dữ liệu và truy vấn LINQ.
  + **Microsoft.VisualBasic.FileIO**: Xử lý dữ liệu từ file CSV.
  + **System.Globalization**: Chuyển đổi định dạng thời gian và xử lý định dạng dữ liệu theo ngôn ngữ.
  + **Microsoft.ML**: Huấn luyện mô hình học máy.
  + **Microsoft.ML.Data** và **Microsoft.ML.Trainers.FastTree**: Cung cấp các thuật toán và pipeline học máy.
  + **System.Windows.Forms** và **System.Windows.Forms.DataVisualization.Charting**: Xây dựng giao diện và hiển thị biểu đồ trực quan.
  + **System.Text**: Xử lý chuỗi văn bản.
  + **System.Collections.Generic**: Quản lý danh sách dữ liệu.
* **Yêu cầu môi trường chạy**:
  + Windows với .NET Framework hoặc .NET Core 8.0 trở lên.

1. **Yêu cầu phi chức năng**

* **Hiệu năng**: Thời gian xử lý dữ liệu và huấn luyện mô hình không vượt quá 2 phút với tập dữ liệu nhỏ (~10.000 dòng).
* **Giao diện người dùng**: Đơn giản, dễ sử dụng cho người không có kỹ năng chuyên môn về dữ liệu.
* **Khả năng mở rộng**: Có thể tích hợp thêm thuật toán học máy hoặc các biểu đồ nâng cao trong tương lai.

1. **Kết quả đầu ra**

Dự báo doanh thu cho khoảng thời gian tương lai.

Hiển thị kết quả dưới dạng biểu đồ

CHƯƠNG 2 - PHÂN TÍCH HỆ THỐNG

2.1 Cơ sở dữ liệu

**Nguồn dữ liệu**

Hệ thống không sử dụng cơ sở dữ liệu truyền thống để lưu trữ thông tin. Dữ liệu đầu vào được lấy từ các tệp Excel có định dạng .csv hoặc .xlsx. Mỗi lần xử lý dữ liệu là một lần hoạt động độc lập, không yêu cầu lưu trữ lâu dài.

**Các thực thể và thuộc tính**

Hệ thống được thiết kế để làm việc với dữ liệu lịch sử bán hàng của các cửa hàng, trong đó:

* **Thực thể chính**:
  + Cửa hàng: Thông tin về các cửa hàng có nhu cầu dự báo tình hình kinh doanh của mình.
* \*Thuộc tính của thực thể:

|  |  |  |
| --- | --- | --- |
| **Tên thuộc tính** | **Mô tả** | **Loại dữ liệu** |
| **Time** | Thời gian bán hàng | Date/Time |
| **Revenue** | Doanh thu | Numeric |
| **Cate** | Danh mục sản phẩm | String |
| **Price** | Giá sản phẩm | Numeric |
| **Quantity** | Số lượng sản phẩm bán được | Numeric |

**Quy trình xử lý dữ liệu**

* 1. **Xóa dữ liệu cũ:** Dọn dẹp dữ liệu đã xử lý trước đó.
  2. **Sao chép dữ liệu:** Chuyển dữ liệu từ DataGridView sang DataTable để xử lý.
  3. **Xóa hàng trống:** Loại bỏ các hàng có ô dữ liệu trống.
  4. **Xóa hàng trùng lặp:** Loại bỏ các hàng có giá trị giống nhau.
  5. **Chuẩn hóa dữ liệu:** Chuyển tất cả dữ liệu sang chữ thường để đảm bảo tính nhất quán.

**Đặc điểm hệ thống**

* Hệ thống chỉ hoạt động dựa trên tệp dữ liệu được nhập vào từ người dùng trong mỗi phiên làm việc.
* Dữ liệu không được lưu trữ sau khi xử lý và kết quả dự báo không được lưu trong cơ sở dữ liệu.
* Không yêu cầu tính bảo mật cao vì dữ liệu không nhạy cảm và chỉ được xử lý tạm thời.

**Đặc tả và trình bày đặc tính của dữ liệu Mẫu**

Tệp dữ liệu  xe đạp ở châu Âu chứa thông tin chi tiết về hoạt động bán hàng của một doanh nghiệp, bao gồm các thông tin liên quan đến sản phẩm, khách hàng, và tài chính.

Dữ liệu được đính kèm file.

* 1. **Nội dung chính của dữ liệu**
* **Thời gian bán hàng:**  
  Dữ liệu bao gồm các thông tin về ngày, tháng, và năm bán hàng, cho phép phân tích xu hướng doanh số theo thời gian.
* **Thông tin khách hàng:**  
  Các thông tin như tuổi, nhóm tuổi, giới tính, quốc gia, và bang của khách hàng giúp đánh giá hành vi mua sắm theo đặc điểm nhân khẩu học.
* **Thông tin sản phẩm:**  
  Dữ liệu về danh mục sản phẩm (Product\_Category), danh mục phụ (Sub\_Category), và từng sản phẩm cụ thể (Product) hỗ trợ phân tích mức độ phổ biến của sản phẩm và hiệu quả kinh doanh của từng danh mục.
* **Thông tin bán hàng:**  
  Các số liệu như số lượng đơn hàng (Order\_Quantity), giá vốn (Unit\_Cost), giá bán (Unit\_Price), chi phí (Cost), lợi nhuận (Profit), và doanh thu (Revenue) phản ánh hiệu suất tài chính của doanh nghiệp.
  1. **Đặc tả dữ liệu**
* **Tổng quan:**
  + Tổng số dòng: 113,036.
  + Tổng số cột: 18.
* **Phân loại cột dữ liệu:**
  + **Thời gian:**
    - Date (ngày đầy đủ), Day (ngày trong tháng), Month, Year.
  + **Khách hàng:**
    - Customer\_Age (tuổi), Age\_Group (nhóm tuổi), Customer\_Gender (giới tính), Country (quốc gia), State (bang/tỉnh).
  + **Sản phẩm:**
    - Product\_Category, Sub\_Category, Product.
  + **Bán hàng:**
    - Order\_Quantity (số lượng đặt hàng), Unit\_Cost (giá đơn vị), Unit\_Price (giá bán đơn vị).
  + **Tài chính:**
    - Profit (lợi nhuận), Cost (chi phí), Revenue (doanh thu).
* **Kiểu dữ liệu:**
  + 9 cột dạng số (int64).
  + 9 cột dạng chuỗi (object).
* **Không có giá trị thiếu:** Toàn bộ các cột đều đầy đủ dữ liệu.

**3.  Trình bày đặc tính của dữ liệu**

**a. Thống kê mô tả dữ liệu số**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Nhóm tuổi | Số lượng khách hàng (Count) | Doanh thu trung bình (Average Revenue) | Lợi nhuận trung bình (Average Profit) | Số lượng đơn trung bình (Average Order Quantity) |
| Adults (35-64) | 55,824 | 762.83 | 292.38 | 12.05 |
| Seniors (64+) | 730 | 422.00 | 189.27 | 13.53 |
| Young Adults (25-34) | 38,654 | 793.08 | 294.58 | 11.56 |
| Youth (<25) | 17,828 | 657.57 | 245.38 | 12.12 |

**2.2 Huấn Luận mô hình học Máy**

**FastTree Regression** là một thuật toán máy học sử dụng mô hình cây quyết định gradient (Gradient Boosted Decision Trees - GBDT) để giải quyết các bài toán hồi quy. Nó đặc biệt tối ưu hóa hiệu suất và tốc độ tính toán, phù hợp với các tập dữ liệu lớn.

* **Tối thiểu hóa hàm mất mát (Loss Function):**  
  FastTree Regression thường sử dụng các hàm mất mát như Mean Squared Error (MSE) để đo lường sai số giữa giá trị thực tế và dự đoán.
* **Xây dựng cây quyết định từng bước:**  
  Từng cây mới được thêm vào sẽ học từ lỗi của mô hình hiện tại, giúp cải thiện dự đoán tổng thể.
* **Gradient Boosting:**  
  FastTree tối ưu hóa cây bằng cách tính gradient của hàm mất mát để điều chỉnh các nút chia (splits) trong cây.

**2.3 Dự Báo Kết Quả**

Các giá trị giả định kinh tế vĩ mô vào mô hình dự báo

**1. Inflation (Lạm phát)**

Giá trị giả định: inflationRate = 0.05f (5% mỗi năm).

Ý nghĩa:

Phản ánh mức độ tăng giá chung của hàng hóa và dịch vụ trong nền kinh tế.

Tác động đến "Price" (giá sản phẩm), "Quantity" (số lượng bán), và "Profit" (lợi nhuận).

Tính toán:

Giá trị được cập nhật hàng năm bằng cách nhân với (1 + inflationRate). Ví dụ: nếu lạm phát là 5%/năm, giá trị của năm tiếp theo sẽ tăng 5% so với năm trước.

**2. GDP Growth (Tăng trưởng GDP)**

Giá trị giả định: gdpGrowthRate = 0.04f (4% mỗi năm).

Ý nghĩa:

Thể hiện tốc độ tăng trưởng kinh tế của quốc gia, ảnh hưởng đến sức mua của người tiêu dùng và doanh thu của doanh nghiệp.

Tính toán:

Hệ số điều chỉnh GDP (gdpAdjustment) được nhân với (1 + gdpGrowthRate) mỗi năm để phản ánh tăng trưởng GDP liên tục.

**3. Marketing Effect (Hiệu quả từ chi phí marketing)**

Giá trị giả định: marketingEffect = 0.02f (2% tăng trưởng mỗi năm từ marketing).

Ý nghĩa:

Phản ánh tác động của các chiến dịch marketing đến việc thúc đẩy doanh thu và số lượng bán hàng.

Ứng dụng:

Không được trực tiếp sử dụng trong mã tính toán giá trị, nhưng có thể được tích hợp để điều chỉnh "Quantity" hoặc "Profit".

**4. Innovation Effect (Hiệu quả từ đổi mới)**

Giá trị giả định: innovationEffect = 0.03f (3% tăng trưởng mỗi năm từ đổi mới).

Ý nghĩa:

Đo lường tác động của việc đầu tư vào đổi mới (sản phẩm, quy trình, công nghệ) đối với sự phát triển dài hạn.

Ứng dụng:

Chỉ số "InnovationIndex" được tăng thêm giá trị (int)(innovationEffect \* 10) mỗi năm, cho thấy mức độ cải tiến liên tục.

**5. Exchange Rate (Tỷ giá hối đoái)**

Giá trị giả định: ExchangeRate = 1.02f.

Ý nghĩa:

Tỷ giá được giả định cố định trong suốt giai đoạn dự báo.

Ảnh hưởng đến giá cả xuất nhập khẩu và các sản phẩm liên quan đến thị trường quốc tế.

**6. Interest Rate (Lãi suất)**

Giá trị giả định: InterestRate = 6.5f.

Ý nghĩa:

Lãi suất giả định cố định trong giai đoạn dự báo.

Tác động đến chi phí vay vốn và sức mua của người tiêu dùng.

**7. Consumer Trend (Xu hướng người tiêu dùng)**

Giá trị giả định: ConsumerTrend = 7 + (int)(gdpAdjustment \* 2).

Ý nghĩa:

Phản ánh sự thay đổi hành vi và thói quen tiêu dùng của người mua sắm.

Được điều chỉnh tăng theo "GDP Growth" để mô phỏng sự gia tăng nhu cầu tiêu dùng khi nền kinh tế phát triển.

1. **Competition Index (Chỉ số cạnh tranh)**

Giá trị giả định: CompetitionIndex = 5.

Ý nghĩa:

Được giả định là cố định trong suốt giai đoạn dự báo.

Biểu thị mức độ cạnh tranh trong ngành, có thể điều chỉnh để mô phỏng sự gia tăng hoặc giảm cạnh tranh.

**9. Tax Rate (Thuế suất)**

Giá trị giả định: TaxRate = 20.0f.

Ý nghĩa:

Thuế suất giả định cố định ở mức 20%.

Ảnh hưởng đến lợi nhuận ròng của doanh nghiệp.

**10. Marketing Expense (Chi phí marketing)**

Giá trị giả định: MarketingExpense = 10000f.

Ý nghĩa:

Giả định chi phí marketing cố định, có thể được thay đổi để đánh giá tác động của marketing lên doanh số.

**11. Innovation Index (Chỉ số đổi mới)**

Giá trị giả định: InnovationIndex = 8 + (int)(innovationEffect \* 10).

Ý nghĩa:

Tăng trưởng hàng năm dựa trên "innovationEffect".

Được sử dụng để phản ánh mức độ phát triển sản phẩm mới và công nghệ trong doanh nghiệp.

CHƯƠNG 3: SƠ ĐỒ TỔNG QUAN

3.1 Use case và đặc tả use case:

Use case diagram:

A diagram of a sales forecasting

Description automatically generated

**Đặc tả use-case:**

**Actor:**

* **Người dùng (User)**: Bất kỳ người dùng nào có dataset và mong muốn dự báo tình hình kinh doanh.

**Các Use-case chính:**

1. Nhập dữ liệu.
2. Chọn lọc dữ liệu.
3. Huấn luyện mô hình học máy.
4. Hiển thị kết quả dự báo bằng biểu đồ trực quan.

**Đặc tả Use-case chi tiết**

1. **Use-case: Nhập dữ liệu**

**Tác nhân liên quan:** Người dùng  
**Mô tả:** Người dùng tải lên tệp dữ liệu đầu vào từ máy tính.  
**Luồng cơ bản:**

1. Người dùng mở Form 1.
2. Nhấn nút "Chọn file".
3. Hệ thống hiển thị hộp thoại để người dùng chọn tệp Excel.
4. Người dùng chọn tệp và nhấn "Mở".
5. Hệ thống kiểm tra dữ liệu, xử lý giá trị null hoặc thiếu cột, và hiển thị dữ liệu trong bảng.
6. Tiền xử lí dữ liệu cơ bản.
7. Lưu file đã được xử lí (nếu muốn)
8. **Use-case: Chọn lọc dữ liệu**

**Tác nhân liên quan:** Người dùng  
**Mô tả:** Người dùng chọn các cột dữ liệu cần thiết cho quá trình dự báo.  
**Luồng cơ bản:**

1. Người dùng mở Form 2.
2. Hệ thống hiển thị danh sách các cột dữ liệu có sẵn.
3. Người dùng sử dụng hộp kiểm để chọn các cột muốn sử dụng.
4. Hệ thống lưu thông tin các cột được chọn để sử dụng trong quá trình học máy.
5. **Use-case: Huấn luyện mô hình học máy**

**Tác nhân liên quan:** Người dùng  
**Mô tả:** Hệ thống sử dụng dữ liệu đã chọn để huấn luyện mô hình học máy.  
**Luồng cơ bản:**

1. Người dùng mở Form 3.
2. Nhấn nút "Huấn luyện mô hình".
3. Đánh giá độ chính xác mô hình (R² score).
4. Hiển thị thông báo huấn luyện hoàn tất và độ chính xác của mô hình.

**Use-case: Hiển thị kết quả dự báo**

**Tác nhân liên quan:** Người dùng  
**Mô tả:** Hệ thống hiển thị biểu đồ dự báo doanh thu trong tương lai.  
**Luồng cơ bản:**

1. Người dùng sử dụng Form 3 sau khi mô hình đã được huấn luyện.
2. Hệ thống hiển thị biểu đồ dự báo theo mặc định.
3. Người dùng có thể chọn khoảng thời gian để thay đổi dữ liệu hiển thị.
4. Hệ thống cập nhật biểu đồ theo khoảng thời gian đã chọn.

3.2 DFD và đặc tả DFD

DF Diagram:

*A screen shot of a computer

Description automatically generated*

Đặc tả DFD:

External Entities:

1. Người dùng:
   * Nhập dữ liệu (file Excel .csv hoặc .xlsx) vào hệ thống.
   * Lựa chọn các cột dữ liệu, thiết lập các điều kiện lọc.
   * Nhận kết quả dự báo và biểu đồ doanh thu.

Processes:

1. Nhập và Tiền xử lý Dữ liệu:
   * Tải dữ liệu từ file người dùng cung cấp.
   * Xử lý dữ liệu bao gồm kiểm tra định dạng, điền giá trị thiếu, xử lý ngoại lệ, và chuyển đổi định dạng.
2. Chọn Lọc Dữ liệu:
   * Hiển thị cột dữ liệu, chọn cột cần thiết, và lọc dữ liệu theo điều kiện người dùng đặt ra.
3. Xây dựng và Đánh giá Mô hình:
   * Sử dụng dữ liệu đã chọn để huấn luyện mô hình dự báo.
   * Đánh giá độ chính xác của mô hình bằng các chỉ số (ví dụ: R-squared).
4. Dự báo và Trực quan:
   * Dự báo doanh thu cho khoảng thời gian tương lai.
   * Hiển thị kết quả dưới dạng biểu đồ.

Luồng Dữ liệu (Data Flows):

1. Từ Người dùng → Hệ thống:
   * File dữ liệu đầu vào.
   * Các yêu cầu lọc, chọn cột.
2. Trong Hệ thống:
   * Dữ liệu được truyền qua các bước: Tiền xử lý → Lựa chọn cột → Huấn luyện → Dự báo.
   * Kết quả dự báo được truyền tới quy trình hiển thị biểu đồ.
3. Từ Hệ thống → Người dùng:
   * Kết quả biểu đồ dự báo.
   * Tùy chọn lưu file kết quả.

Không có Lưu trữ Tạm thời:

* Hệ thống xử lý dữ liệu từ khi nhập cho đến khi dự báo mà không lưu trữ dữ liệu.

CHƯƠNG 4 - THIẾT KẾ

4.1 Nguyên tắc thiết kế giao diện

Ứng dụng dự báo doanh thu được thiết kế với các tiêu chí sau:

* Đơn giản, trực quan: Giao diện dễ sử dụng, không yêu cầu người dùng phải có kiến thức chuyên sâu về kỹ thuật hay nghiệp vụ.
* Tương tác dễ dàng: Cung cấp các nút thao tác rõ ràng như "Tải dữ liệu", "Xử lý", "Dự báo", "Hiển thị kết quả".
* Thống nhất: Màu sắc, phông chữ, và cách bố trí được áp dụng nhất quán trên toàn bộ ứng dụng.
* Hỗ trợ hướng dẫn người dùng: Các thông báo hoặc chỉ dẫn khi người dùng tải dữ liệu hoặc gặp lỗi.

4.2 Chi tiết giao diện

Ứng dụng bao gồm các màn hình chính sau:

4.2.1 Màn hình Nhập Dữ liệu

* Mục đích: Cho phép người dùng tải lên file dữ liệu dạng .csv hoặc .xlsx.
* Thành phần:
  + Nút Browse: Mở cửa sổ chọn file từ máy tính.
  + Nhãn hiển thị thông báo tệp đã được tải thành công.
  + A screenshot of a computer

    Description automatically generatedNút Tiền xử lý dữ liệu để chuyển đến bước tiếp theo.

4.2.2 Màn hình Tiền xử lý Dữ liệu

* Mục đích: Hiển thị thông tin dữ liệu đã tải lên, cho phép người dùng tùy chỉnh cách xử lý dữ liệu.
* Thành phần:
  + Bảng hiển thị dữ liệu với các tùy chọn như:
    - Xử lý giá trị thiếu (Điền 0 hoặc loại bỏ).
    - Lựa chọn các cột cần thiết.
  + Nút Xử lý và Lưu.

A screenshot of a computer

Description automatically generated

A screenshot of a computer

Description automatically generated4.2.3 Màn hình Dự báo

* Mục đích: Hiển thị các tùy chọn huấn luyện mô hình và dự báo.
* Nút Huấn luyện mô hình.
* Nút Thực hiện dự báo.

4.2.4 Màn hình Kết quả và Trực quan hóa

* Mục đích: Hiển thị kết quả dự báo và trực quan hóa bằng biểu đồ.
* Thành phần:
  + Biểu đồ dạng cột (bar chart) thể hiện doanh thu trong tương lai.
  + Tùy chọn lưu kết quả dự báo dưới dạng file.

A screenshot of a graph

Description automatically generated

4.3 Nguyên tắc thiết kế màu sắc

* Màu chủ đạo: Tông màu trắng xám nhạt, tạo cảm giác nhẹ nhàng, tối giản, thân thiện.
* Phông chữ: Sử dụng phông chữ dễ đọc như Segoe UI.

4.4 Mô tả luồng giao diện

1. Bước 1: Người dùng mở ứng dụng, tải file dữ liệu từ máy tính.
2. Bước 2: Dữ liệu được hiển thị, người dùng thực hiện các bước tiền xử lý.
3. Bước 3: Sau khi xử lý, người dùng chọn mô hình và thực hiện dự báo.
4. Bước 4: Kết quả được hiển thị trên biểu đồ, người dùng có thể lưu hoặc xem lại.

4.5 Mô hình giao diện

* Sơ đồ bố trí giao diện (Wireframe):
  + Trang Nhập dữ liệu: Khu vực chọn file và nút tiếp tục.
  + Trang Tiền xử lý: Bảng dữ liệu và các tùy chọn.
  + Trang Dự báo: Dropdown menu và nút thực thi.
  + Trang Kết quả: Biểu đồ và tùy chọn lưu.

# CHƯƠNG 5 – KIỂM THỬ VÀ ĐÁNH GIÁ

## 5.1 Mục đích kiểm thử

Kiểm thử và đánh giá hệ thống là một bước quan trọng trong quá trình phát triển phần mềm, giúp đảm bảo tính chính xác, hiệu suất và độ tin cậy của ứng dụng. Mục đích của kiểm thử trong đề án này là:

* Đảm bảo rằng ứng dụng hoạt động đúng chức năng, xử lý dữ liệu chính xác.
* Đánh giá hiệu suất của hệ thống, đảm bảo rằng ứng dụng có thể xử lý một lượng lớn dữ liệu mà không gặp phải sự cố.
* Kiểm tra độ chính xác của mô hình học máy, đặc biệt là khả năng dự báo doanh thu trong tương lai.

## 5.2 Kiểm thử chức năng

Kiểm thử chức năng tập trung vào việc xác nhận rằng tất cả các chức năng trong hệ thống hoạt động đúng như yêu cầu đã đặt ra.

1. **Nhập và tiền xử lý dữ liệu**:
   * Kiểm thử việc tải tệp dữ liệu (.csv hoặc .xlsx) vào hệ thống.
   * Đảm bảo rằng hệ thống xử lý các giá trị null và thiếu cột đúng cách.
   * Xác nhận rằng dữ liệu sau khi được xử lý hiển thị chính xác trong bảng và có thể tiếp tục sang các bước tiếp theo.
2. **Chọn lọc dữ liệu**:
   * Kiểm thử tính năng cho phép người dùng chọn các cột cần thiết cho quá trình học máy.
   * Đảm bảo rằng hệ thống chỉ sử dụng các cột đã chọn để huấn luyện mô hình.
3. **Huấn luyện mô hình học máy**:
   * Kiểm tra việc huấn luyện mô hình fasttree regression
   * Đảm bảo rằng hệ thống tự động huấn luyện mô hình và đánh giá độ chính xác của mô hình thông qua các chỉ số phù hợp.
4. **Hiển thị kết quả**:
   * Kiểm tra việc hiển thị biểu đồ dự báo doanh thu chính xác theo mốc thời gian đã chọn.
   * Đảm bảo rằng biểu đồ có thể được tùy chỉnh theo yêu cầu của người dùng (chọn khoảng thời gian, v.v.).

## 5.3 Kiểm thử hiệu suất

Kiểm thử hiệu suất đảm bảo rằng hệ thống có thể hoạt động ổn định và hiệu quả với lượng dữ liệu lớn và trong thời gian xử lý hợp lý.

1. **Thời gian xử lý**:
   * Kiểm thử thời gian tải dữ liệu và quá trình huấn luyện mô hình với các tập dữ liệu có kích thước khác nhau (từ vài nghìn dòng đến ~10.000 dòng).
   * Đảm bảo rằng thời gian xử lý không vượt quá 2 phút với tập dữ liệu nhỏ (khoảng 10.000 dòng).
2. **Tải dữ liệu và tiền xử lý**:
   * Kiểm tra hiệu suất khi tải và tiền xử lý dữ liệu từ các tệp Excel với nhiều dữ liệu và nhiều cột.
   * Đảm bảo rằng hệ thống có thể xử lý dữ liệu mà không gặp phải sự cố về bộ nhớ hoặc tốc độ.
3. **Khả năng mở rộng**:
   * Kiểm thử tính mở rộng của hệ thống bằng cách thử nghiệm với tập dữ liệu lớn hơn (vượt quá 10.000 dòng) và xác nhận rằng hệ thống vẫn hoạt động ổn định.

## 5.4 Đánh giá mô hình học máy

Đánh giá mô hình học máy là một phần quan trọng để xác định hiệu quả dự báo của hệ thống. Dưới đây là các chỉ số và phương pháp đánh giá sẽ được sử dụng:

1. **Đánh giá độ chính xác của mô hình**:
   * **R² (R-squared) score**: Được sử dụng để đo lường mức độ phù hợp của mô hình với dữ liệu thực tế. Giá trị R² gần 1 cho thấy mô hình dự báo rất chính xác, trong khi giá trị gần 0 cho thấy mô hình không phù hợp.
     + Công thức tính R²:

A mathematical equation with numbers and symbols

Description automatically generated

* + - Trong đó:

A white background with black text

Description automatically generated

1. **Đánh giá mô hình với các thuật toán khác nhau**:
   * So sánh hiệu quả dự báo của mỗi thuật toán dựa trên R² score và thời gian huấn luyện mô hình.
   * Chọn thuật toán có hiệu quả nhất dựa trên các chỉ số đánh giá và yêu cầu của người dùng.
2. **Kiểm thử mô hình với dữ liệu mới**:
   * Để đánh giá tính tổng quát của mô hình, bạn có thể kiểm thử mô hình dự báo với bộ dữ liệu mới mà mô hình chưa được huấn luyện trước đó.
   * Kiểm tra khả năng của mô hình trong việc dự báo chính xác cho các trường hợp chưa thấy trước.

## 5.5 Kết quả kiểm thử

Sau khi thực hiện các bước kiểm thử và đánh giá, bạn sẽ tổng hợp lại các kết quả để xác định mức độ chính xác và hiệu quả của mô hình. Các kết quả này có thể bao gồm:

* **R² score** của mô hình học máy cho từng tập dữ liệu thử nghiệm.
* Thời gian huấn luyện mô hình và thời gian xử lý dữ liệu.
* Các chỉ số hiệu suất khi xử lý tập dữ liệu lớn và nhỏ.

## 5.6 Kết luận

Dựa trên kết quả kiểm thử và đánh giá, bạn có thể đưa ra kết luận về hiệu quả của hệ thống:

* Nếu mô hình đạt được R² score cao có thể kết luận rằng mô hình học máy hoạt động chính xác và có khả năng dự báo doanh thu tốt.
* Nếu hệ thống đáp ứng các yêu cầu về hiệu suất và xử lý dữ liệu trong thời gian hợp lý, hệ thống có thể được triển khai và sử dụng trong môi trường thực tế.

# TÀI LIỆU THAM KHẢO

[Tutorial: Forecast bike rental demand - time series - ML.NET | Microsoft Learn](https://learn.microsoft.com/en-us/dotnet/machine-learning/tutorials/time-series-demand-forecasting)

[Introduction to Machine Learning in C# with ML.NET](https://gilberttanner.com/blog/introduction-to-machine-learning-in-c-with-ml-net/)

[Machine Learning with ML.NET - Guide to Decision Trees](https://rubikscode.net/2021/02/22/machine-learning-with-ml-net-guide-to-decision-trees/)

[Jon Wood - YouTube](https://www.youtube.com/@JonWood)

<https://www.kaggle.com/datasets/sadiqshah/bike-sales-in-europe> : Nguồn tài liệu của dữ liệu mẫu bike sales in eu