

**BỘ GIÁO DỤC VÀ ĐÀO TẠO**  
**TRƯỜNG ĐẠI HỌC NGUYỄN TẤT THÀNH**  
**KHOA CÔNG NGHỆ THÔNG TIN**

---



**TIỂU LUẬN MÔN HỌC**  
**CHUYÊN ĐỀ CHUYÊN SÂU KHDL 2**

**ỨNG DỤNG THUẬT TOÁN ARIMA VÀ TRỰC QUAN HOÁ**  
**BÁO CÁO TRÊN POWER BI VỚI TẬP DỮ LIỆU VỀ NGUỒN VỐN**  
**ĐẦU TƯ TRỰC TIẾP NƯỚC NGOÀI VÀO VIỆT NAM**

<b>Giảng viên hướng dẫn:</b>	<b>ThS. VƯƠNG XUÂN CHÍ</b>
<b>Sinh viên thực hiện:</b>	<b>NGÔ CÔNG HUÂN</b>
<b>MSSV:</b>	<b>2000002680</b>
<b>Môn học:</b>	<b>Chuyên đề chuyên sâu KHDL 2</b>
<b>Lớp:</b>	<b>21DTH2C</b>
<b>Khoá:</b>	<b>2021</b>

**Tp.HCM, tháng 5 năm 2024**

**BỘ GIÁO DỤC VÀ ĐÀO TẠO**  
**TRƯỜNG ĐẠI HỌC NGUYỄN TẤT THÀNH**  
**KHOA CÔNG NGHỆ THÔNG TIN**

---



**TIỂU LUẬN MÔN HỌC**  
**CHUYÊN ĐỀ CHUYÊN SÂU KHDL 2**

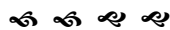
**ỨNG DỤNG THUẬT TOÁN ARIMA VÀ TRỰC QUAN HOÁ**  
**BÁO CÁO TRÊN POWER BI VỚI TẬP DỮ LIỆU VỀ NGUỒN VỐN**  
**ĐẦU TƯ TRỰC TIẾP NƯỚC NGOÀI VÀO VIỆT NAM**

**Giảng viên hướng dẫn:** ThS. VƯƠNG XUÂN CHÍ  
**Sinh viên thực hiện:** NGÔ CÔNG HUÂN  
**MSSV:** 2000002680  
**Môn học:** Chuyên đề chuyên sâu KHDL 2  
**Lớp:** 21DTH2C  
**Khoá:** 2021

**Tp.HCM, tháng 5 năm 2024**

TRƯỜNG ĐẠI HỌC NGUYỄN TẤT THÀNH  
TRUNG TÂM KHẢO THÍ

KÌ THI KẾT THÚC HỌC PHẦN  
HỌC KÌ II NĂM HỌC 2023 - 2024



## PHIẾU CHẤM THI TIỂU LUẬN

Môn thi: **Chuyên đề chuyên sâu KHDL 2** .....Mã lớp học phần: **21DTH2C**.....

Nhóm sinh viên thực hiện:

Sinh viên 1: **Ngô Công Huân** .....Tham gia đóng góp: **100%** .....

Sinh viên 2: **Phan Quốc Điền** .....Tham gia đóng góp: **100%** .....

Ngày thi: **16/05/2024** .....Phòng thi: **L.903** .....

Đề tài tiểu luận của sinh viên : **Ứng dụng thuật toán ARIMA và trực quan hoá báo cáo trên Power BI với tập dữ liệu về Nguồn vốn đầu tư trực tiếp Nước ngoài vào Việt Nam**

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

Tiêu chí (theo CĐR HP)	Đánh giá của giáo viên	Điểm tối đa	Điểm đạt được
Cấu trúc của báo cáo			
Nội dung			
- Các nội dung thành phần			
- Lập luận			
- Kết luận			
Trình bày			
<b>TỔNG ĐIỂM</b>		10	

**Giảng viên chấm thi**

(Ký và ghi rõ họ tên)

ThS. Vương Xuân Chí

## **LỜI CẢM ƠN**

Đầu tiên, em xin gửi lời cảm ơn chân thành đến Trường Đại học Nguyễn Tất Thành vì đã tạo điều kiện thuận lợi cho em trong quá trình học tập và hoàn thành Tiểu luận môn học Chuyên đề chuyên sâu KHDL 2. Đặc biệt, em xin bày tỏ lòng biết ơn sâu sắc đến Thạc sĩ Vương Xuân Chí - người đã trực tiếp hướng dẫn em trong tiểu luận môn học này.

Thầy Vương Xuân Chí đã dành thời gian và công sức để truyền đạt cho em kiến thức cơ bản về trực quan hoá dữ liệu và đưa ra báo cáo phân tích với Power BI. Đó là những kiến thức hết sức quý báu không chỉ trong quá trình thực hiện tiểu luận mà còn là hành trang tiếp bước cho em trong quá trình học tập và lập nghiệp sau này.

Trong quá trình thực hiện tiểu luận, tuy rằng đã rất cố gắng nhưng với kiến thức và kinh nghiệm còn hạn chế nên sẽ khó tránh khỏi những sai sót. Em kính mong Thầy chỉ bảo thêm để em có thể hoàn thành tốt tiểu luận môn học này và có thêm kinh nghiệm cho những công việc trong tương lai. Em xin kính chúc Thầy luôn mạnh khỏe, hạnh phúc và luôn gặp niềm vui trong cuộc sống.

## LỜI MỞ ĐẦU

Trong thời đại của sự kết nối toàn cầu, Việt Nam đã trở thành điểm đến hấp dẫn cho các nhà đầu tư nước ngoài, với việc thu hút vốn đầu tư trực tiếp từ nước ngoài (FDI) đóng vai trò quan trọng trong việc thúc đẩy sự phát triển kinh tế của đất nước. Trong bối cảnh này, việc dự đoán và đánh giá xu hướng FDI có thể cung cấp thông tin quý giá cho các quyết định chiến lược của chính phủ và các nhà đầu tư.

Trong bài viết này, chúng tôi tiếp cận vấn đề này bằng cách áp dụng một phương pháp phân tích dự báo mạnh mẽ là thuật toán học máy ARIMA (AutoRegressive Integrated Moving Average). Chúng tôi tập trung vào việc áp dụng ARIMA để dự đoán xu hướng FDI của Việt Nam trong tương lai, dựa trên dữ liệu lịch sử về FDI và các yếu tố kinh tế, xã hội có liên quan.

Bằng cách kết hợp sức mạnh của ARIMA và khả năng trực quan hóa dữ liệu của công cụ Power BI, chúng tôi không chỉ cung cấp các dự đoán chính xác về FDI mà còn giúp hiểu rõ hơn về các yếu tố ảnh hưởng đến xu hướng này. Bài viết cũng nhấn mạnh sự quan trọng của việc áp dụng công nghệ và phương pháp phân tích dữ liệu hiện đại để hỗ trợ quyết định chiến lược và quản lý rủi ro trong môi trường kinh doanh ngày càng biến động của Việt Nam.

Mục tiêu của chúng tôi là cung cấp một cơ sở lý luận và thực hành vững chắc cho việc sử dụng ARIMA và Power BI trong việc dự đoán FDI của Việt Nam, đồng thời khám phá tiềm năng của việc tích hợp các công nghệ và phương pháp này vào quy trình ra quyết định kinh doanh và chính sách kinh tế.

# MỤC LỤC

<b>LỜI CẢM ƠN .....</b>	<b>i</b>
<b>LỜI MỞ ĐẦU .....</b>	<b>ii</b>
<b>DANH MỤC HÌNH ẢNH .....</b>	<b>iv</b>
<b>CHƯƠNG 1: GIỚI THIỆU .....</b>	<b>1</b>
1.1. Lý do chọn đề tài .....	1
1.2. Mục tiêu nghiên cứu .....	1
1.3. Phương pháp nghiên cứu .....	1
1.4. Ý nghĩa nghiên cứu .....	1
<b>CHƯƠNG 2: CƠ SỞ LÝ THUYẾT .....</b>	<b>2</b>
2.1. Power BI .....	2
2.1.1. Tổng quan về Power BI .....	2
2.1.2. Giới thiệu Power BI .....	2
2.1.3. Giao diện chính của Power BI .....	2
2.1.4. Lợi ích của Power BI .....	3
2.2. Mô hình ARIMA .....	4
2.2.1. Tổng quan về mô hình .....	4
2.2.2. Xây dựng mô hình .....	5
2.2.3. Đánh giá mô hình .....	6
<b>CHƯƠNG 3: THỰC NGHIỆM VÀ KẾT QUẢ .....</b>	<b>7</b>
3.1. Xử lý dữ liệu với Python .....	7
3.1.1. Giới thiệu về tập dữ liệu .....	7
3.1.2. Tiền xử lý dữ liệu .....	7
3.1.3. Dự đoán với mô hình học máy .....	9
3.2. Trực quan hoá báo cáo trên Power BI .....	11
3.2.1. Summary Dashboard .....	11
3.2.2. Partner Dashboard .....	12
3.2.3. Provinces Dashboard .....	13
3.2.4. Industries Dashboard .....	15
<b>CHƯƠNG 4: KẾT LUẬN .....</b>	<b>17</b>
4.1. Kết luận .....	17
4.2. Hạn chế và hướng phát triển của đề tài .....	17
<b>TÀI LIỆU THAM KHẢO .....</b>	<b>19</b>

## DANH MỤC HÌNH ẢNH

Hình 3.1. Đổi tên các thuộc tính trong dữ liệu.....	8
Hình 3.2. Xử lý giá trị null trong tập dữ liệu .....	8
Hình 3.3. Loại bỏ kí tự dư thừa và đổi kiểu dữ liệu cho các giá trị .....	9
Hình 3.4. Hàm thêm và xoá dữ liệu để chuẩn bị dữ liệu cho mô hình .....	10
Hình 3.5. Dự đoán với mô hình học máy.....	10
Hình 3.6. Summary Dashboard.....	11
Hình 3.7. Partner Dashboard 2022.....	12
Hình 3.8. Partner Dashboard 2025.....	13
Hình 3.9. Provinces Dashboard 2022.....	13
Hình 3.10. Provinces Dashboard 2025.....	14
Hình 3.11. Industries Dashboard 2022.....	15
Hình 3.12. Industries Dashboard 2025.....	16

# CHƯƠNG 1: GIỚI THIỆU

Trong thời đại số hóa ngày nay, dữ liệu trở thành nguồn tài nguyên vô cùng quan trọng và đa dạng. Khai thác dữ liệu là một lĩnh vực ngày càng phát triển, nơi các phương pháp và thuật toán được áp dụng để tìm ra thông tin ẩn sau các tập dữ liệu lớn. Môn học "Chuyên đề chuyên sâu KHDL 2" chính là hành trình chinh phục những khối dữ liệu phức tạp để trích xuất tri thức hữu ích và ứng dụng chúng trong thực tế.

## 1.1. Lý do chọn đề tài

Việc lựa chọn đề tài này xuất phát từ nhận thức về sự quan trọng của FDI đối với phát triển kinh tế của Việt Nam. Sự gia tăng đáng kể trong lưu lượng FDI đến Việt Nam trong những năm gần đây đã làm tăng sự quan tâm về việc dự đoán và quản lý xu hướng này. Đồng thời, sự phát triển của các công nghệ phân tích dữ liệu mở ra cơ hội mới để tiếp cận và hiểu rõ hơn về dòng vốn này.

## 1.2. Mục tiêu nghiên cứu

Mục tiêu của nghiên cứu là áp dụng thuật toán học máy ARIMA để dự đoán xu hướng FDI của Việt Nam và trực quan hoá kết quả trên Power BI. Bằng cách này, chúng tôi mong muốn cung cấp một công cụ mạnh mẽ cho các nhà quản lý và nhà đầu tư để hỗ trợ quyết định chiến lược và quản lý rủi ro.

## 1.3. Phương pháp nghiên cứu

Phương pháp nghiên cứu của chúng tôi bao gồm ba bước chính: thu thập dữ liệu lịch sử, xử lý dữ liệu và xây dựng mô hình ARIMA, và trực quan hoá kết quả trên Power BI. Bằng cách kết hợp giữa các kỹ thuật phân tích dữ liệu và công nghệ trực quan hoá, chúng tôi hi vọng tạo ra một phương pháp toàn diện và hiệu quả.

## 1.4. Ý nghĩa nghiên cứu

Nghiên cứu này hứa hẹn mang lại giá trị lớn trong việc hiểu rõ về cách Random Forest có thể được áp dụng trong lĩnh vực bất động sản và phát triển đô thị. Việc xây dựng một mô hình phân loại chính xác có thể giúp đưa ra quyết định đầu tư thông minh, cũng như định hình hướng phát triển của thị trường bất động sản trong khu vực Hà Nội. Đồng thời, nghiên cứu cũng mở ra cơ hội cho sự ứng dụng của các phương pháp khai thác dữ liệu và machine learning trong các lĩnh vực liên quan đến quản lý và phát triển đô thị.



## **CHƯƠNG 2: CƠ SỞ LÝ THUYẾT**

### **2.1. Power BI**

#### **2.1.1. Tổng quan về Power BI**

Power BI là một bộ công cụ “phân tích kinh doanh” để phân tích dữ liệu và chia sẻ thông tin chi tiết của Microsoft. Kết nối để dễ dàng truy cập vào dữ liệu trên Dashboard (bảng điều khiển), Reports (báo cáo) tập dữ liệu trong Power BI.

Công cụ BI (Business Intelligence) tự phục vụ(self-service) phân tích dữ liệu cho nhân viên, được sử dụng bởi các nhà phân tích dữ liệu và các chuyên gia phân tích thông tin kinh doanh.

Power BI được sử dụng bởi cả đại diện bộ phận và ban quản lý, với các báo cáo và dự báo được tạo ra để hỗ trợ các đại diện tiếp thị và bán hàng, đồng thời cung cấp dữ liệu cho việc quản lý về cách bộ phận hoặc cá nhân ...

#### **2.1.2. Giới thiệu Power BI**

Power BI là một công cụ phân tích giúp hình ảnh hóa, trực quan hóa bằng cách kết nối các nguồn dữ liệu rời rạc và phân tích thông qua các báo cáo và bảng điều khiển (dashboard).

Tích hợp của các dịch vụ phần mềm, ứng dụng và bộ kết nối khác nhau nhằm chuyển các nguồn dữ liệu rời rạc thành những thông tin hữu ích.

Công cụ BI thông minh nhất được thiết kế cho người dùng tự điều chỉnh mà Microsoft cho rời đời từ năm 2015. Được nhiều người lựa chọn khi xét đến khía cạnh khả năng lưu trữ dữ liệu.

Có thể nhìn vào dữ liệu từ nhiều góc độ khác nhau và các báo cáo, bảng điều khiển có thể dễ dàng được tạo ra bởi bất kỳ ai trong tổ chức mà không cần có sự hỗ trợ từ đội ngũ IT và quản trị viên.

#### **2.1.3. Giao diện chính của Power BI**

Giao diện của Power BI được chia thành ba phần chính:

- Report: Phần này là nơi người dùng có thể tạo và tương tác với các báo cáo và biểu đồ dựa trên dữ liệu đã được nhập vào Power BI. Trong phần Report, người dùng có thể thêm các biểu đồ, đồ thị, bảng, và các thành phần trực quan khác để hiển thị thông tin dưới dạng các báo cáo trực quan và dễ hiểu.

- **Data:** Phần Data là nơi người dùng có thể xem và quản lý dữ liệu được sử dụng trong báo cáo. Từ phần này, người dùng có thể xem cấu trúc dữ liệu, thực hiện các thao tác biến đổi và làm sạch dữ liệu bằng cách sử dụng Power Query, và kiểm tra các thông số về dữ liệu như số lượng hàng và cột.
- **Relationship:** Phần Relationship cho phép người dùng xác định và quản lý mối quan hệ giữa các bảng dữ liệu. Điều này là quan trọng để kết hợp dữ liệu từ nhiều nguồn khác nhau và tạo ra các báo cáo phức tạp có chứa thông tin từ các bảng dữ liệu khác nhau. Trong phần này, người dùng có thể xác định các mối quan hệ khóa ngoại giữa các bảng và kiểm tra mối quan hệ này thông qua biểu đồ quan hệ.

#### **2.1.4. Lợi ích của Power BI**

Power BI mang lại nhiều lợi ích cho người dùng, bao gồm:

- **Trực quan hoá dữ liệu:** Power BI cho phép người dùng trực quan hoá dữ liệu từ nhiều nguồn khác nhau thành các biểu đồ, đồ thị và bảng điều khiển (dashboard) dễ hiểu và thú vị. Việc này giúp người dùng hiểu rõ hơn về dữ liệu và phân tích các xu hướng, mối quan hệ và biến động.
- **Tích hợp dữ liệu đa nguồn:** Power BI có khả năng kết nối với nhiều nguồn dữ liệu khác nhau, từ cơ sở dữ liệu cục bộ đến dịch vụ đám mây và các ứng dụng trực tuyến khác. Điều này giúp người dùng tổ chức và kết hợp dữ liệu từ nhiều nguồn khác nhau để tạo ra cái nhìn toàn diện và đa chiều về dữ liệu.
- **Phân tích dữ liệu mạnh mẽ:** Power BI cung cấp các công cụ và tính năng phân tích dữ liệu mạnh mẽ, cho phép người dùng thực hiện các phép biến đổi, tính toán và phân tích dữ liệu phức tạp một cách dễ dàng. Điều này giúp người dùng hiểu rõ hơn về dữ liệu và đưa ra các quyết định dựa trên thông tin đáng tin cậy.
- **Tạo báo cáo tùy chỉnh:** Power BI cho phép người dùng tạo ra các báo cáo và bảng điều khiển tùy chỉnh theo nhu cầu của họ. Từ việc chọn các biểu đồ và đồ thị phù hợp đến việc thêm các tính năng tương tác và bộ lọc, người dùng có thể tạo ra các báo cáo linh hoạt và mạnh mẽ để trình

bày thông tin một cách rõ ràng và hiệu quả.

- Chia sẻ và hợp tác dễ dàng: Power BI cho phép người dùng chia sẻ và làm việc cùng nhau trên các báo cáo và bảng điều khiển một cách dễ dàng. Tính năng chia sẻ và hợp tác này giúp tăng cường sự hợp tác và trao đổi thông tin giữa các thành viên trong tổ chức, từ đó giúp cải thiện quyết định kinh doanh và hiệu suất làm việc.

## **2.2. Mô hình ARIMA**

### **2.2.1. Tổng quan về mô hình**

Mô hình ARIMA (AutoRegressive Integrated Moving Average) là một phương pháp thống kê được sử dụng để phân tích và dự đoán chuỗi dữ liệu thời gian. Mô hình ARIMA kết hợp ba thành phần chính: AR (AutoRegressive), I (Integrated), và MA (Moving Average).

- Thành phần AutoRegressive (AR): Thành phần này mô tả mối quan hệ giữa giá trị hiện tại của chuỗi dữ liệu và các giá trị trước đó trong chuỗi. Trong mô hình ARIMA, thành phần AR thường được ký hiệu là  $p$ , và được biểu diễn bởi  $AR(p)$ . Giá trị của  $p$  xác định số lượng các giá trị trước đó được sử dụng để dự đoán giá trị hiện tại.
- Thành phần Integrated (I): Thành phần này liên quan đến việc loại bỏ xu hướng tăng/giảm từ chuỗi dữ liệu bằng cách thực hiện phép toán chuyển đổi (difference). Thành phần này giúp làm cho chuỗi dữ liệu ổn định và phù hợp với mô hình ARIMA. Thành phần I thường được ký hiệu là  $d$ , và mô hình ARIMA được biểu diễn bởi  $ARIMA(p, d, q)$ , trong đó  $d$  là một số nguyên không âm.
- Thành phần Moving Average (MA): Thành phần này mô tả mối quan hệ giữa giá trị hiện tại và các giá trị nhiễu trong chuỗi dữ liệu. Thành phần MA thường được ký hiệu là  $q$ , và mô hình ARIMA được biểu diễn bởi  $ARIMA(p, d, q)$ . Giá trị của  $q$  xác định số lượng giá trị nhiễu trước đó được sử dụng để dự đoán giá trị hiện tại.

Kết hợp ba thành phần AR, I, và MA, mô hình ARIMA có thể được sử dụng để dự đoán và phân tích các chuỗi dữ liệu thời gian, từ dự báo xu hướng tương lai

đến đánh giá các biến động và chu kỳ trong dữ liệu. Mô hình này đã được áp dụng rộng rãi trong nhiều lĩnh vực như tài chính, kinh tế, y học, và thời tiết.

### 2.2.2. Xây dựng mô hình

Quá trình xây dựng mô hình ARIMA (AutoRegressive Integrated Moving Average) thường gồm các bước sau:

- Xác định dữ liệu thích hợp: Thu thập dữ liệu và kiểm tra tính ổn định của chuỗi thời gian. Chuỗi thời gian cần phải không có xu hướng hoặc chu kỳ và có tính ổn định.
- Chuẩn bị dữ liệu: Đảm bảo dữ liệu đầu vào đã được xử lý và là chuỗi thời gian đơn nhất. Nếu cần, thực hiện các biến đổi như chia tỷ lệ hoặc logarit để giảm phương sai và giúp dữ liệu trở nên ổn định hơn.
- Xác định các tham số ARIMA: Sử dụng các biểu đồ ACF (Autocorrelation Function) và PACF (Partial Autocorrelation Function) để xác định các tham số của mô hình ARIMA. Thông qua quan sát các đỉnh trên biểu đồ ACF và PACF, bạn có thể xác định các tham số  $p$ ,  $d$ ,  $q$  cho mô hình ARIMA.
- Xây dựng mô hình ARIMA: Dựa trên các tham số đã xác định ở bước trước, xây dựng mô hình ARIMA bằng cách sử dụng hàm ARIMA trong các thư viện phù hợp (như statsmodels trong Python).
- Kiểm định mô hình: Sử dụng các phương pháp kiểm định như kiểm tra ACF và PACF của dãy thời gian còn sót lại (residuals) để đảm bảo rằng mô hình không có sự tương quan còn lại.
- Dự đoán và đánh giá mô hình: Sử dụng mô hình đã xây dựng để dự đoán giá trị trong tương lai. Đánh giá hiệu suất của mô hình bằng cách so sánh dự đoán với dữ liệu thực tế, sử dụng các độ đo như RMSE (Root Mean Square Error), MAE (Mean Absolute Error) hoặc SMAPE (Symmetric Mean Absolute Percentage Error).
- Tinh chỉnh mô hình (nếu cần): Nếu mô hình không đạt được hiệu suất mong muốn, bạn có thể điều chỉnh các tham số ARIMA hoặc thử các mô hình khác nhau như SARIMA (Seasonal ARIMA).

- Triển khai mô hình: Khi mô hình đã đạt được hiệu suất mong muốn, bạn có thể triển khai nó để dự đoán giá trị trong thực tế.

### 2.2.3. Đánh giá mô hình

Kiểm tra độ chính xác của dự đoán: Đánh giá mức độ chính xác của mô hình bằng cách so sánh giữa giá trị dự đoán và giá trị thực tế của chuỗi thời gian. Các độ đo thường được sử dụng bao gồm:

RMSE (Root Mean Square Error): Đo lường sự chênh lệch giữa các dự đoán và giá trị thực tế, với mục tiêu là giảm bớt sai số.

MAE (Mean Absolute Error): Đo lường trung bình của giá trị tuyệt đối của sự chênh lệch giữa dự đoán và giá trị thực tế.

SMAPE (Symmetric Mean Absolute Percentage Error): Đo lường tỷ lệ phần trăm của sự chênh lệch giữa dự đoán và giá trị thực tế, phù hợp cho dữ liệu có phạm vi biến động lớn.

Kiểm tra tính ổn định của residuals: Đảm bảo rằng residuals của mô hình không có sự tương quan còn lại. Điều này có thể được thực hiện bằng cách kiểm tra ACF và PACF của residuals hoặc sử dụng các kiểm định thống kê như Ljung-Box test.

So sánh với các mô hình khác: Đôi khi, để đánh giá mô hình ARIMA, bạn cũng cần so sánh nó với các mô hình khác như mô hình hồi quy tuyến tính, mô hình Holt-Winters, hoặc các mô hình dự báo khác phù hợp với dữ liệu.

Kiểm tra tính linh hoạt của mô hình: Một mô hình ARIMA linh hoạt sẽ có khả năng dự đoán tốt trên nhiều loại dữ liệu thời gian khác nhau và có thể điều chỉnh tốt với các điều kiện biến đổi.

Kiểm tra ổn định và tính chất của dữ liệu dự báo: Xác định xem dự báo từ mô hình ARIMA có ổn định qua thời gian hay không. Điều này giúp đảm bảo rằng mô hình vẫn hoạt động tốt khi áp dụng cho dữ liệu mới.

## CHƯƠNG 3: THỰC NGHIỆM VÀ KẾT QUẢ

### 3.1. Xử lý dữ liệu với Python

#### 3.1.1. Giới thiệu về tập dữ liệu

Tên bộ dữ liệu: Đầu tư trực tiếp nước ngoài vào Việt Nam từ 2015 - 2022

Được xuất bản bởi: Open Development Vietnam

Tăng trưởng kinh tế có mối tương quan tích cực với sự gia tăng hàng năm của việc thu hút FDI vào Việt Nam. Vốn FDI chiếm tỷ trọng đáng kể trong tổng vốn đầu tư của toàn xã hội. Sự gia tăng FDI giải ngân sẽ mở rộng quy mô sản xuất của các ngành kinh tế, từ đó tạo điều kiện thúc đẩy tăng trưởng kinh tế. FDI cũng giúp thúc đẩy xuất khẩu, góp phần vào thặng dư cán cân thương mại của Việt Nam, từ đó thúc đẩy tăng trưởng GDP. Dữ liệu về đầu tư trực tiếp nước ngoài từ năm 2015 đến cuối năm 2022 cung cấp cho người đọc số lượng đầu tư mới và số vốn đầu tư của các nhà đầu tư nước ngoài.

Bộ dữ liệu gồm có 3 tệp riêng được phân loại theo:

- Đối tác đầu tư (fdi\_country\_partners\_en.csv)
- Các tỉnh đầu tư (fdi\_provinces\_vi.csv)
- Các ngành đầu tư (fdi\_industry\_vi.csv)

Dữ liệu bao gồm các thuộc tính như đối tác đầu tư, số lượng dự án mới được cấp phép, vốn đăng ký (triệu đô la Mỹ), số lượng dự án điều chỉnh, vốn điều chỉnh, số lượng góp vốn, giá trị góp vốn (triệu đô la Mỹ), tên tỉnh được đầu tư, tên ngành đầu tư.

#### 3.1.2. Tiền xử lý dữ liệu

Nhìn chung, cả 3 tập dữ liệu đều có cấu trúc giống nhau nên ta sẽ áp dụng cách làm sạch dữ liệu tương tự nhau cho tất cả tập dữ liệu.

Đầu tiên, khi đọc tập dữ liệu và in ra ta thấy được rằng tên các cột có chứa các kí tự đặc biệt và những khoảng trắng dư thừa nên ta sẽ tiến hành sửa lại tên của chúng.

Ngành	Số dự án cấp mới	Vốn đăng ký cấp mới (triệu USD)	Số lượt dự án điều chỉnh	Vốn đăng ký điều chỉnh (triệu USD)	Số lượt góp vốn mua cổ phần	Giá trị góp vốn, mua cổ phần (triệu USD)	Năm
1 Công nghiệp chế biến, chế tạo	955	8927.8	517	6305.4	NaN	NaN	2015
2 Sản xuất phân phối điện, khí, nước, điều hòa	9	2795.3	8	14.0	NaN	NaN	2015
3 Hoạt động kinh doanh bất động sản	34	2145.4	12	248.3	NaN	NaN	2015
4 Xây dựng	112	573.6	26	162.9	NaN	NaN	2015
5 Bán buôn và bán lẻ; sửa chữa ô tô, mô tô, xe máy	306	375.2	89	166.8	NaN	NaN	2015

Hình 3.1. Đổi tên các thuộc tính trong dữ liệu

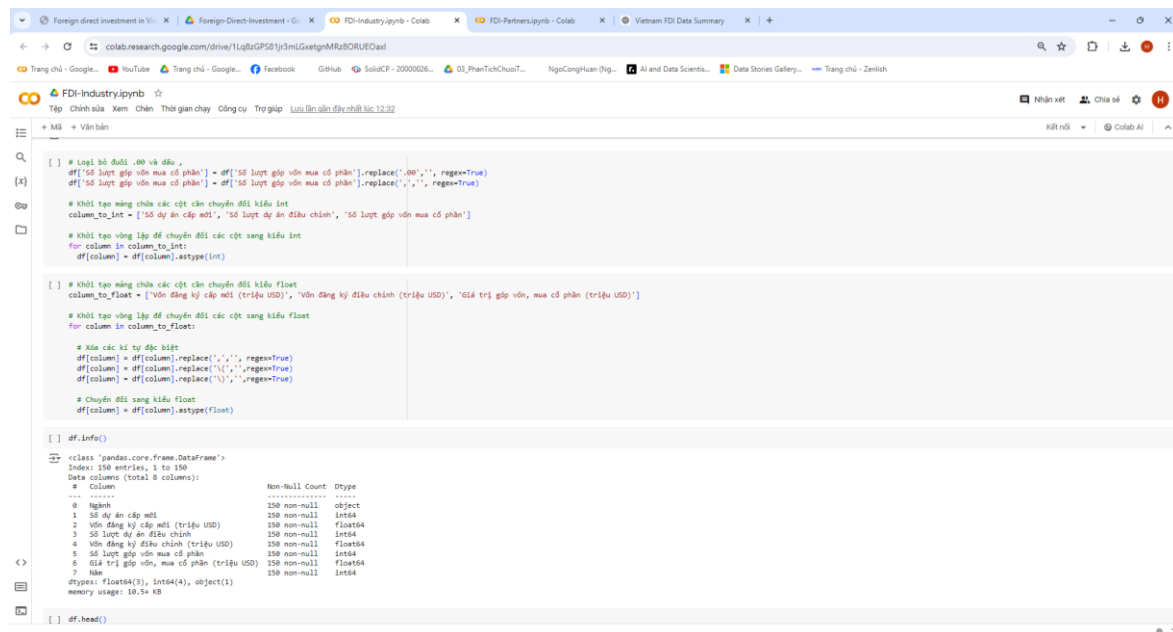
Tiếp theo, ta sẽ thống kê số lượng các giá trị null trong mỗi cột. Ta thấy rằng số lượng giá trị null chiếm khá nhiều trong tập dữ liệu. Tuy nhiên, ở trong ngữ cảnh của bài toán này ta có thể hiểu rằng giá trị null đồng nghĩa với việc không có nguồn vốn đầu tư vào ở thuộc tính đó nên ta sẽ tiến hành thay thế các giá trị null thành số 0.

Bối tác	Số dự án cấp mới	Vốn đăng ký cấp mới (triệu USD)	Số lượt dự án điều chỉnh	Vốn đăng ký điều chỉnh (triệu USD)	Số lượt góp vốn mua cổ phần	Giá trị góp vốn, mua cổ phần (triệu USD)	Năm
0	150	150	290	268	114	113	0
1	150	150	290	268	114	113	0
2	150	150	290	268	114	113	0
3	150	150	290	268	114	113	0
4	150	150	290	268	114	113	0

Hình 3.2. Xử lý giá trị null trong tập dữ liệu

Sau đó ta kiểm tra các giá trị trong tập dữ liệu thì thấy rằng các giá trị này có một vài lỗi như số thập phân có chứa dấu phẩy hay có các kí tự đặc biệt dư thừa trong

các giá trị. Ta sẽ tiến hành làm sạch lại các giá trị và chuyển đổi chúng sang kiểu float và int ứng với từng thuộc tính.



```
[ ] # Loại bỏ đuôi .00 và dấu .
df['Số lượt góp vốn mua cổ phần'] = df['Số lượt góp vốn mua cổ phần'].replace('.', '', regex=True)
df['Số lượt góp vốn mua cổ phần'] = df['Số lượt góp vốn mua cổ phần'].replace(',', '', regex=True)

# Khởi tạo mảng chứa các cột cần chuyển đổi kiểu int
column_to_int = ['Số dự án cấp mới', 'Số lượt dự án điều chỉnh', 'Số lượt góp vốn mua cổ phần']

# Khởi tạo vòng lặp để chuyển đổi các cột sang kiểu int
for column in column_to_int:
    df[column] = df[column].astype(int)

[ ] # Khởi tạo mảng chứa các cột cần chuyển đổi kiểu float
column_to_float = ['Vốn đăng ký cấp mới (triệu USD)', 'Vốn đăng ký điều chỉnh (triệu USD)', 'Giá trị góp vốn, mua cổ phần (triệu USD)']

# Khởi tạo vòng lặp để chuyển đổi các cột sang kiểu float
for column in column_to_float:
    # Xóa các ký tự đặc biệt
    df[column] = df[column].replace(' ', '', regex=True)
    df[column] = df[column].replace('\n', '', regex=True)
    df[column] = df[column].replace('\r', '', regex=True)

    # Chuyển đổi sang kiểu float
    df[column] = df[column].astype(float)

[ ] df.info()

<class 'pandas.core.frame.DataFrame'>
Int64Index: 150 entries, 1 to 150
Data columns (total 8 columns):
 #   Column                                     Non-Null Count  Dtype  Dtype
---  --
 0   Ngành                                     150 non-null    object
 1   Số dự án cấp mới                         150 non-null    int64
 2   Vốn đăng ký cấp mới (triệu USD)          150 non-null    float64
 3   Số lượt dự án điều chỉnh                 150 non-null    int64
 4   Vốn đăng ký điều chỉnh (triệu USD)        150 non-null    float64
 5   Số lượt góp vốn mua cổ phần               150 non-null    int64
 6   Giá trị góp vốn, mua cổ phần (triệu USD)  150 non-null    float64
 7   Năm                                       150 non-null    int64
dtypes: float64(3), int64(4), object(1)
memory usage: 10.5+ KB

[ ] df.head()
```

Hình 3.3. Loại bỏ ký tự dư thừa và đổi kiểu dữ liệu cho các giá trị

### 3.1.3. Dự đoán với mô hình học máy

Chúng ta sẽ áp dụng mô hình ARIMA để dự đoán toàn bộ các đặc trưng thuộc từng đối tác/ngành/tỉnh vào các năm 2023, 2024, 2025 bằng cách sử dụng dữ liệu từ năm 2015 đến năm 2022.

Tuy nhiên, vấn đề gặp phải ở đây là mô hình ARIMA đòi hỏi chuỗi thời gian phải là liên tục và bộ dữ liệu của chúng ta ở mỗi năm khác nhau thì số lượng các đối tác/ngành/tỉnh lại khác nhau. Hướng xử lý cho vấn đề này đó là ta sẽ tiến hành thêm dữ liệu vào các năm bị thiếu bằng cách cho tất cả các đặc trưng bằng 0 (trong ngữ cảnh của bài toán nghĩa là năm đó không có đối tượng này có nguồn vốn đầu tư)

Sau khi thêm dữ liệu vào, bộ dữ liệu mới đã có đầy đủ thông tin từ năm 2015 đến năm 2022 theo từng đối tượng. Ta sẽ tiến hành áp dụng mô hình với thư viện pmdarima trong Python và cụ thể là mô hình auto\_arima. Mô hình này sẽ tự động tìm ra bộ tham số p,q,d tốt nhất cho từng loại đặc trưng.

Khi đã dự đoán hết tất cả đặc trưng thuộc tất cả đối tượng từ năm 2023 đến năm 2025, ta sẽ kết hợp bộ dữ liệu gốc và dự liệu dự đoán lại với nhau và lưu thành tệp csv mới.



```

[ ] ##### Hàm Thêm dữ liệu #####
def fill_data(df):
    # Thống kê các năm và đối tượng
    objects = df.iloc[:,0].unique()
    years = df.iloc[:,1].unique()

    # Duyệt qua từng đối tượng
    for obj in objects:
        # DataFrame tạm thời chứa thông tin nguồn vốn của đối tượng qua các năm
        df_temp = df[df.iloc[:,0] == obj]

        # Thống kê các năm mà đối tượng này không có dữ liệu
        years_of_obj = [y for y in years if not in df_temp.iloc[:,1].values]

        # Thêm dữ liệu vào các năm mà đối tượng không tồn tại
        if years_of_obj:
            for year in years_of_obj:
                data = {}
                data[df.columns[0]] = obj
                data[df.columns[-1]] = year

                for col in df.columns[1:-1]:
                    data[col] = 0

                df_new = pd.DataFrame([data])
                df = pd.concat([df, df_new], ignore_index=True)

    return df

##### Hàm xoá dữ liệu đã thêm #####
def remove_fill_data(df):
    df = df[df.columns[1:-1].sum(axis=1) != 0]

```

Hình 3.4. Hàm thêm và xoá dữ liệu để chuẩn bị dữ liệu cho mô hình

```

[ ] ##### Hàm dự đoán tổng trực đầu tư nước ngoài #####
def predict_fdi(df):
    df = df.sort_values(by="năm")
    data = df.iloc[:,1].values

    # Khởi tạo mô hình
    model = sm.tsa.ARIMA(data, seasonal=False, trace=True)
    n_periods = 3
    forecasts = model.predict(n_periods=n_periods)
    results = forecasts

    return results

##### Hàm dự đoán tất cả các trường #####
def fill_data(df):
    df = df.fillna(0)
    for obj in df.iloc[:,0].unique():
        df_temp = df[df.iloc[:,0] == obj]
        data_predict = {}

        for col in df.columns[1:-1]:
            results = predict_fdi(df_temp[df_temp.columns[0] == col, df_temp.columns[-1]])
            data_predict[col] = results

        data_predict[df_temp.columns[-1]] = [y for y in range(2023, 2026)]
        data_predict[df_temp.columns[0]] = [obj for i in range(2023, 2026)]

        df_predict = pd.DataFrame(data_predict)
        df = pd.concat([df, df_predict], ignore_index=True)

    df = remove_fill_data(df)
    df = df.reset_index(drop=True)

    return df

df = main(df)

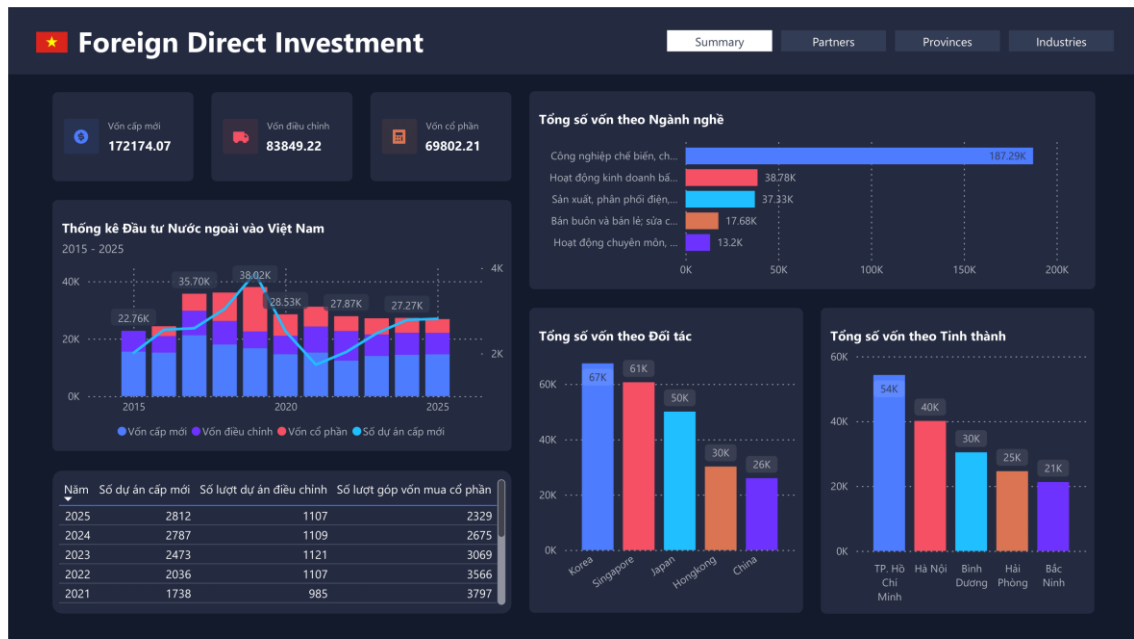
```

Hình 3.5. Dự đoán với mô hình học máy

Cuối cùng, ta thu được 3 tập dữ liệu hoàn chỉnh từ năm 2015 đến năm 2025 và ta sẽ bắt đầu tiến hành trực quan hoá bộ dữ liệu với Power BI.

## 3.2. Trực quan hoá báo cáo trên Power BI

### 3.2.1. Summary Dashboard



Hình 3.6. Summary Dashboard

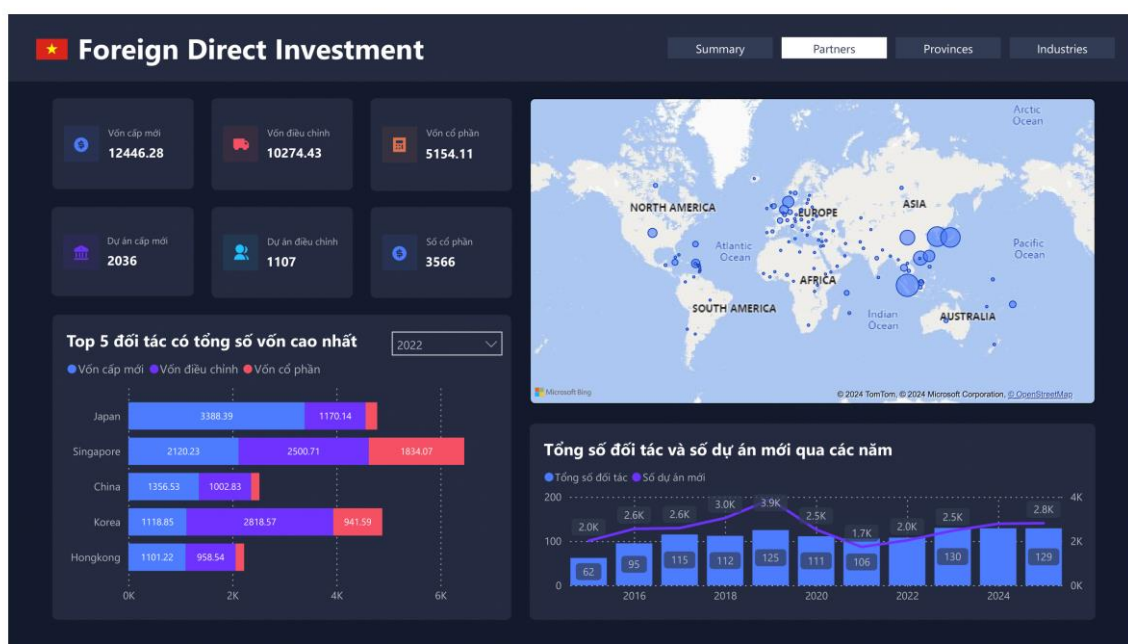
Từ bảng báo cáo cho thấy tổng giá trị từ các nguồn vốn trong năm 2015 – 2025 cao nhất thuộc về vốn cấp mới, cụ thể như sau:

- Tổng số vốn cấp mới: 172174.07 triệu USD
- Tổng số vốn điều chỉnh: 83849.22 triệu USD
- Tổng số vốn cổ phần: 69802.21 triệu USD

Dựa vào biểu đồ “Thống kê Đầu tư Nước ngoài vào Việt Nam”, ta thấy rằng tổng số vốn và số dự án cấp mới có xu hướng tăng mạnh từ năm 2015 đến năm 2019. Tuy nhiên vào cuối năm 2019, do đại dịch Covid 19 bùng phát nên số dự án mới đã giảm đáng kể đến năm 2021. Và từ năm 2022 trở đi nguồn vốn đã dần dần hồi phục mạnh mẽ hơn.

Qua 3 biểu đồ cột theo từng loại đối tượng, ngành công nghiệp chế biến chiếm tỷ trọng cao nhất, các quốc gia đầu tư vào Việt Nam nhiều nhất lần lượt là Korea, Singapore, Japan,... và Thành phố Hồ Chí Minh với Thủ đô Hà Nội là hai khu vực có nhiều tiềm năng thu hút nguồn vốn nhất nước ta.

### 3.2.2. Partner Dashboard



Hình 3.7. Partner Dashboard 2022

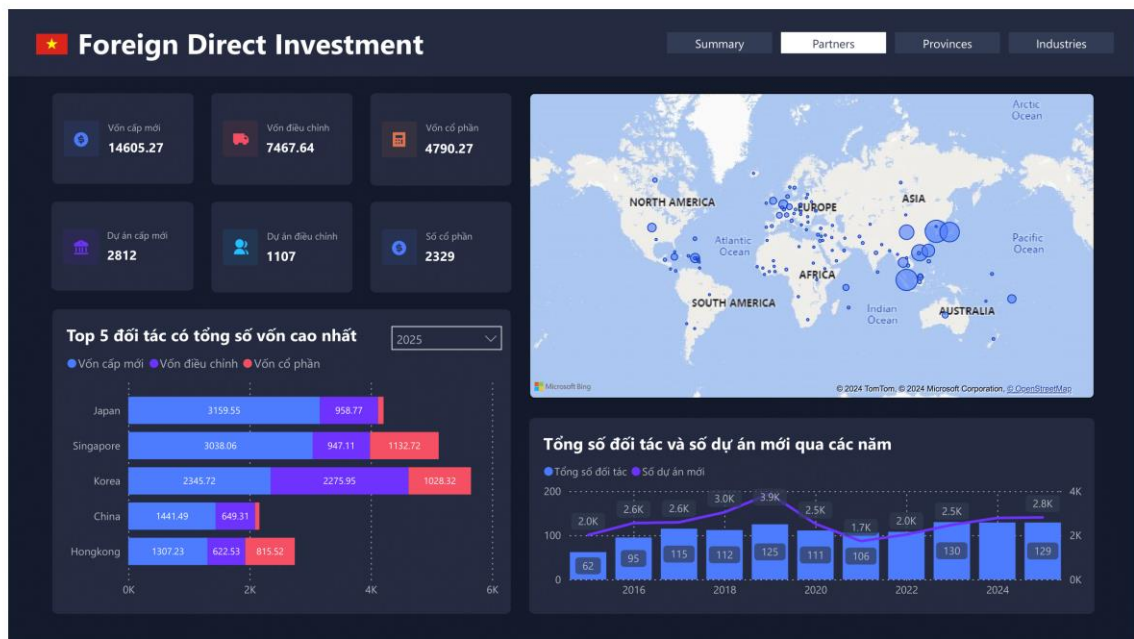
#### Năm 2022:

Có 108 quốc gia và vùng lãnh thổ đầu tư vào Việt Nam trong năm nay. Singapore vẫn là nguồn đầu tư nước ngoài hàng đầu của Việt Nam với 6,46 tỷ USD, chiếm 23,3% tổng vốn FDI đăng ký vào nước này (giảm 39,7% so với cùng kỳ năm ngoái). Hàn Quốc đứng thứ hai với khoảng 4,88 tỷ USD, giảm 1,5% so với cùng kỳ năm ngoái. Nhật Bản đứng thứ ba với tổng vốn đầu tư đăng ký hơn 4,78 tỷ USD, chiếm 17,3% và tăng 22,7% so với cùng kỳ năm trước. Tiếp theo là Trung Quốc, Hồng Kông (Trung Quốc), Đài Loan (Trung Quốc) v.v.

Về số lượng dự án, Hàn Quốc đứng đầu danh sách nhà đầu tư về số dự án đăng ký cấp mới và điều chỉnh vốn (chiếm 20,4% số dự án cấp mới, 32,6% số dự án điều chỉnh và 34,1% số vốn góp, mua cổ phần).

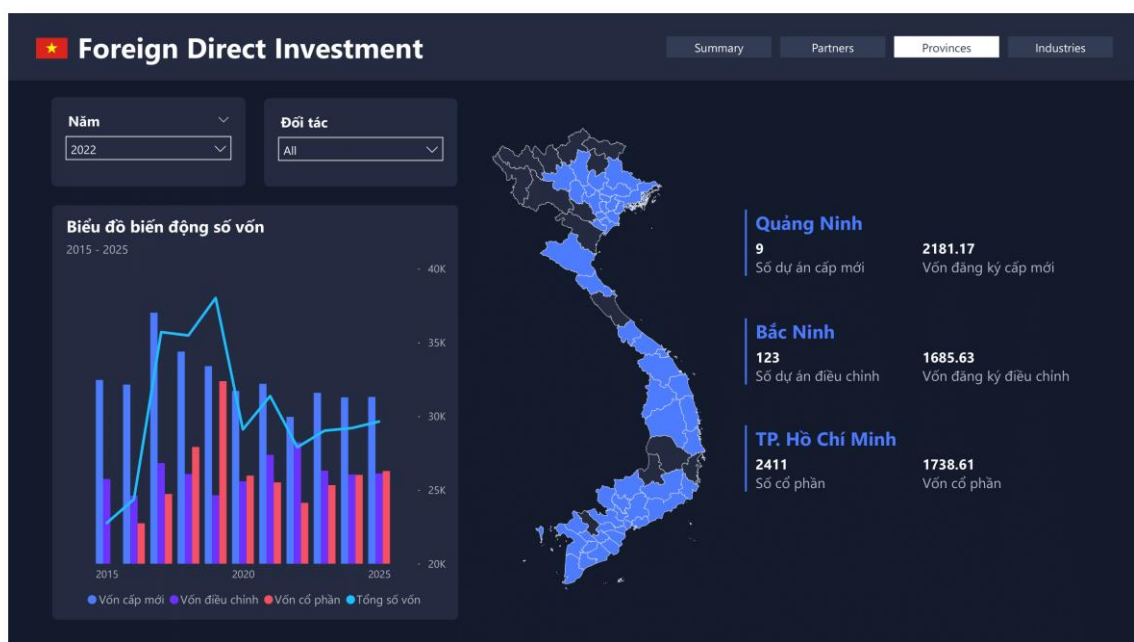
#### Năm 2025:

Theo kết quả dự đoán với mô hình học máy, số đối tác và số dự án mới có xu hướng tăng nhẹ và Korea là đối tác có tổng nguồn vốn cao nhất trong năm.



Hình 3.8. Partner Dashboard 2025

### 3.2.3. Provinces Dashboard



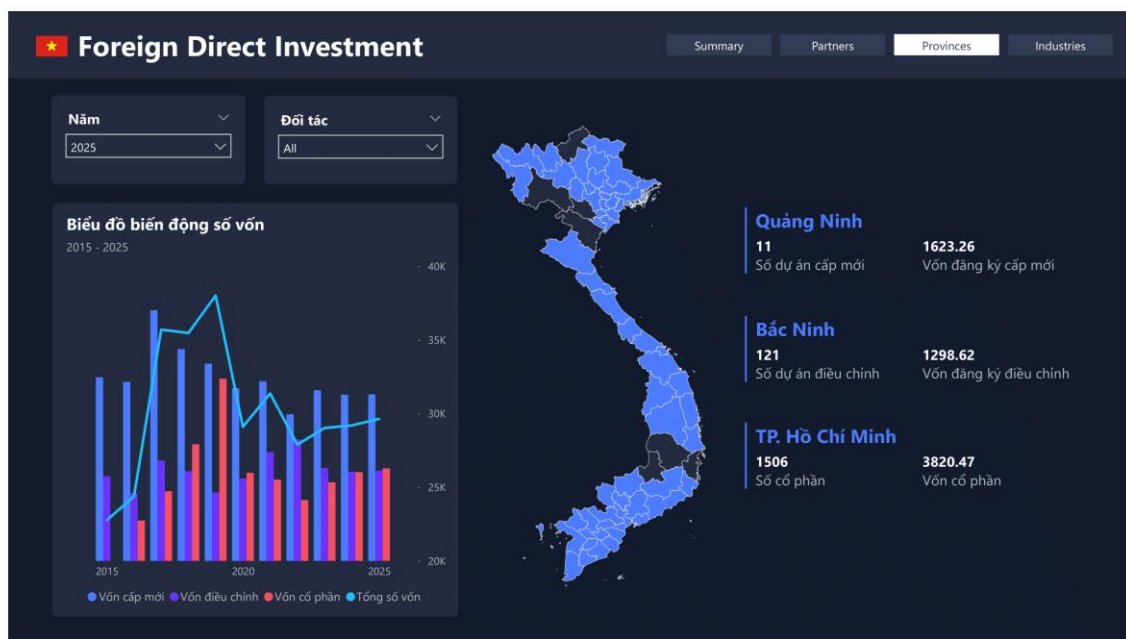
Hình 3.9. Provinces Dashboard 2022

#### Năm 2022:

Các nhà đầu tư nước ngoài đã đầu tư vào 54 tỉnh, thành phố trong cả nước trong năm 2022. Hồ Chí Minh dẫn đầu với hơn 3,94 tỷ USD, chiếm 14,2% tổng vốn và tăng 5,4% so với cùng kỳ năm 2021. Bình Dương đứng thứ hai với tổng vốn đầu tư hơn 3,14 tỷ USD, chiếm 47,3% so cùng kỳ. Quảng Ninh đứng thứ ba với tổng vốn

đầu tư đăng ký 2,37 tỷ USD, chiếm 8,5% và tăng gấp 2 lần so với cùng kỳ năm 2021. Tiếp theo là Bắc Ninh, Hải Phòng, Hà Nội,...

Về số lượng dự án mới, nhà đầu tư nước ngoài vẫn tập trung vào các thành phố lớn có hạ tầng thuận tiện như TP.HCM, Hà Nội. Trong đó, TP.HCM dẫn đầu cả về số dự án cấp mới (43,9%), góp vốn, mua cổ phần (67,6%) và đứng thứ hai về số dự án đăng ký điều chỉnh vốn (17,3%, sau Hà Nội là 18,6%).

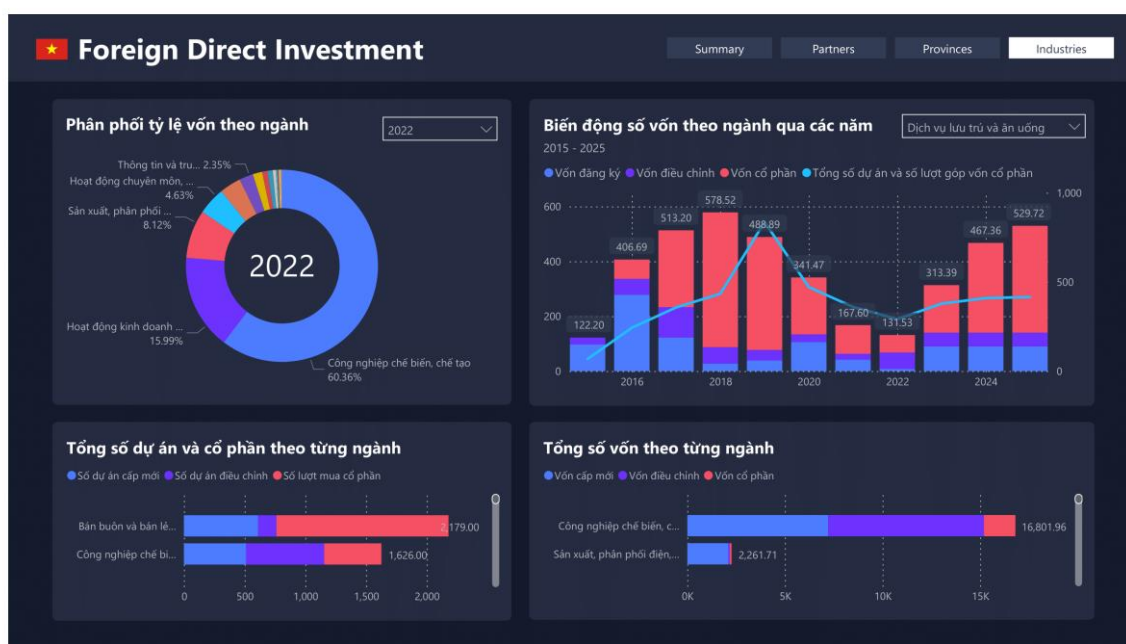


Hình 3.10. Provinces Dashboard 2025

### Năm 2025:

Qua kết quả dự đoán, dẫn đầu các loại vốn là Quảng Ninh, Bắc Ninh và TP. Hồ Chí Minh. Vốn đăng ký cấp mới ở Quảng Ninh giảm mạnh và Số cổ phần ở TP. Hồ Chí Minh tăng lên, còn nguồn vốn Bắc Ninh chỉ thay đổi nhẹ so với năm 2022,

### 3.2.4. Industries Dashboard

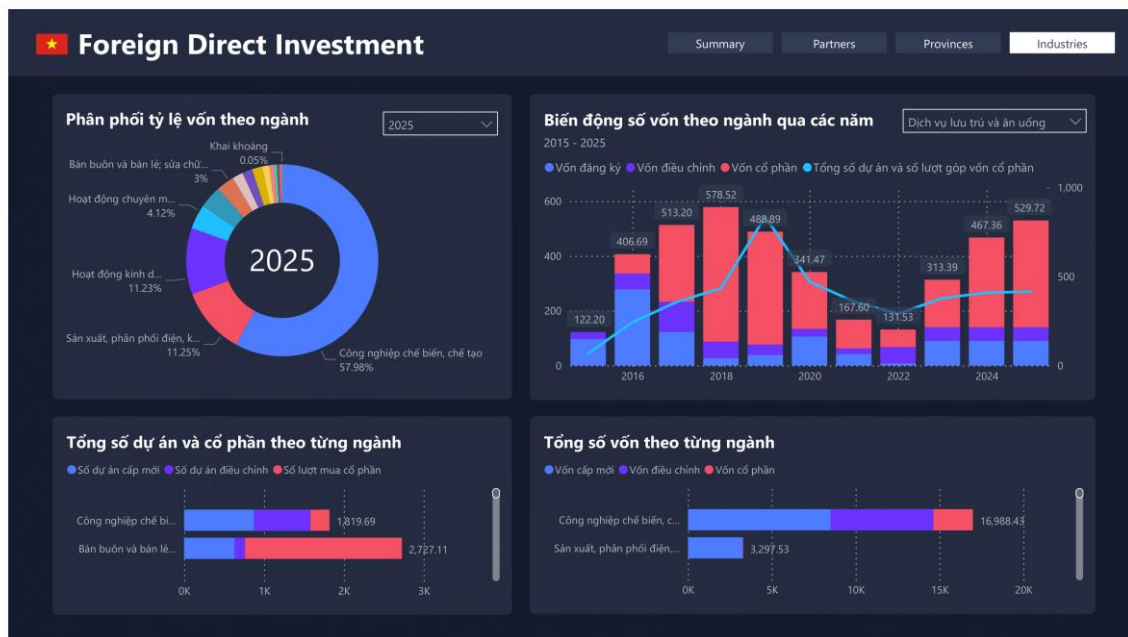


Hình 3.11. Industries Dashboard 2022

#### Năm 2022:

Các nhà đầu tư nước ngoài rót vốn vào 19/21 ngành trong hệ thống phân loại kinh tế quốc gia, trong đó công nghiệp chế biến, chế tạo dẫn đầu với tổng vốn đầu tư hơn 16,8 tỷ USD, chiếm 60,6% tổng vốn cả nước. Tiếp theo là bất động sản với tổng vốn đầu tư 4,45 tỷ USD, chiếm 16,1% tổng vốn đầu tư đăng ký. Tiếp theo là sản xuất, phân phối điện với trên 2,26 tỷ USD và hoạt động khoa học công nghệ với gần 1,29 tỷ USD. Phần còn lại là các lĩnh vực khác.

Điều đáng chú ý là bán buôn và bán lẻ, chế biến, sản xuất và hoạt động khoa học công nghệ là những ngành có số lượng dự án đăng ký mới nhiều nhất, lần lượt chiếm 30%, 25,1% và 16,3% tổng số dự án.



Hình 3.12. Industries Dashboard 2025

### Năm 2025:

Từ năm 2022 đến năm 2025, xu hướng của các ngành chiếm tỷ lệ nguồn vốn cao vẫn tăng lên. Đứng đầu về tổng nguồn vốn vẫn là ngành Công nghệ chế biến chế tạo và tổng số dự án, cổ phần là ngành Bán buôn, chế tạo.



## CHƯƠNG 4: KẾT LUẬN

### 4.1. Kết luận

Báo cáo này cung cấp cái nhìn tổng quan về tình hình đầu tư trực tiếp nước ngoài vào Việt Nam từ năm 2015 đến 2022. Báo cáo nhấn mạnh vai trò quan trọng của FDI trong việc thúc đẩy tăng trưởng kinh tế, mở rộng quy mô sản xuất, và cải thiện cán cân thương mại của Việt Nam. Dữ liệu được phân loại chi tiết theo các đối tác đầu tư, các tỉnh nhận đầu tư, và các ngành nghề đầu tư, cho phép người đọc có cái nhìn sâu rộng về bức tranh FDI tại Việt Nam trong giai đoạn này.

Ứng dụng thuật toán ARIMA để dự đoán vốn đầu tư trực tiếp FDI từ năm 2023 đến 2025 đã cho thấy những kết quả khả quan. Thuật toán ARIMA, với khả năng xử lý tốt các dữ liệu thời gian và dự đoán xu hướng tương lai dựa trên quá khứ, đã cung cấp các dự báo chi tiết về xu hướng FDI trong những năm tới. Những dự báo này có thể hỗ trợ các nhà hoạch định chính sách và các nhà đầu tư trong việc lập kế hoạch và đưa ra các quyết định chiến lược nhằm tối ưu hóa lợi ích từ dòng vốn FDI.

Kết quả dự báo từ mô hình ARIMA cho thấy sự tiếp tục tăng trưởng của dòng vốn FDI vào Việt Nam, mặc dù có thể gặp phải một số biến động do các yếu tố kinh tế và chính trị toàn cầu. Những phân tích này cũng nhấn mạnh tầm quan trọng của việc duy trì môi trường đầu tư ổn định và cải thiện các chính sách thu hút đầu tư để giữ vững đà tăng trưởng.

Tóm lại, việc sử dụng thuật toán ARIMA không chỉ giúp dự đoán chính xác xu hướng FDI mà còn cung cấp cơ sở dữ liệu quan trọng để các bên liên quan có thể đưa ra các quyết định chiến lược. Việc dự báo FDI từ 2023 đến 2025 là bước đi quan trọng để đảm bảo Việt Nam tiếp tục là điểm đến hấp dẫn cho các nhà đầu tư nước ngoài, từ đó thúc đẩy sự phát triển kinh tế bền vững.

### 4.2. Hạn chế và hướng phát triển của đề tài

#### Hạn chế:

Độ chính xác của dự báo: Mặc dù thuật toán ARIMA là công cụ mạnh mẽ trong việc dự đoán dữ liệu thời gian, độ chính xác của dự báo vẫn phụ thuộc vào chất lượng và độ chi tiết của dữ liệu lịch sử. Nếu dữ liệu bị thiếu hoặc không đầy đủ, mô hình có thể đưa ra các dự báo không chính xác.

Biến động kinh tế và chính trị: Các yếu tố kinh tế và chính trị toàn cầu không



thể dự đoán trước có thể ảnh hưởng mạnh mẽ đến dòng vốn FDI. ARIMA chủ yếu dựa vào các xu hướng lịch sử, do đó, không thể dự đoán chính xác các sự kiện bất ngờ như khủng hoảng kinh tế, xung đột chính trị hay đại dịch.

Giới hạn của mô hình: ARIMA tập trung vào các yếu tố thời gian và có thể bỏ qua những yếu tố phi thời gian quan trọng khác như thay đổi chính sách, sự phát triển công nghệ, và thay đổi trong môi trường kinh doanh toàn cầu.

Phức tạp trong điều chỉnh mô hình: Việc điều chỉnh các tham số của mô hình ARIMA để phù hợp với dữ liệu cụ thể có thể phức tạp và đòi hỏi sự hiểu biết sâu sắc về thống kê và phân tích dữ liệu. Sai sót trong việc điều chỉnh tham số có thể dẫn đến kết quả dự báo không chính xác.

### **Hướng phát triển:**

Kết hợp các mô hình khác: Để cải thiện độ chính xác của dự báo, có thể kết hợp ARIMA với các mô hình dự đoán khác như mô hình hồi quy, mạng nơ-ron nhân tạo, hoặc mô hình học sâu. Điều này có thể giúp khai thác tối đa các ưu điểm của từng mô hình và giảm bớt các hạn chế riêng lẻ.

Cải thiện chất lượng dữ liệu: Để tăng cường độ chính xác của dự báo, cần đầu tư vào việc thu thập và chuẩn bị dữ liệu chất lượng cao. Điều này bao gồm việc thu thập dữ liệu chi tiết hơn về các yếu tố ảnh hưởng đến FDI và đảm bảo tính toàn vẹn của dữ liệu.

Phân tích tác động của các yếu tố phi thời gian: Nghiên cứu thêm về các yếu tố phi thời gian như chính sách đầu tư, môi trường kinh doanh, và tình hình chính trị có thể cung cấp một cái nhìn toàn diện hơn và cải thiện độ chính xác của dự báo.

Ứng dụng công nghệ tiên tiến: Sử dụng các công nghệ tiên tiến như phân tích dữ liệu lớn (big data analytics) và trí tuệ nhân tạo (AI) có thể giúp xử lý và phân tích một lượng lớn dữ liệu phức tạp, từ đó cung cấp các dự báo chính xác và kịp thời hơn.

Tăng cường hợp tác quốc tế: Hợp tác với các tổ chức quốc tế và các chuyên gia trong lĩnh vực đầu tư và dự báo kinh tế có thể mang lại những hiểu biết mới và cải thiện phương pháp luận, từ đó nâng cao chất lượng và độ chính xác của dự báo.

Bằng cách khắc phục các hạn chế hiện tại và theo đuổi các hướng phát triển mới, nghiên cứu về dự đoán FDI sẽ ngày càng chính xác và hữu ích, đóng góp quan trọng vào việc hoạch định chính sách và chiến lược đầu tư tại Việt Nam.

## **TÀI LIỆU THAM KHẢO**

- [1] Hà Minh Tân, Phân tích dữ liệu chuỗi thời gian và ứng dụng, Hồ Chí Minh: Đại học Nguyễn Tất Thành, 2024.
- [2] Vương Xuân Chí, Chuyên đề chuyên sâu Khoa học dữ liệu 2, Hồ Chí Minh: Đại học Nguyễn Tất Thành, 2024.