

PHÂN TÍCH VÀ DỰ BÁO HOẠT ĐỘNG ĐẦU TƯ TẠI KHU CÔNG NGHỆ CAO THÀNH PHỐ HỒ CHÍ MINH

Huỳnh Trọng Thừa*, Tân Hạnh*, Lưu Nguyễn Kỳ Thư*,
Huỳnh Trung Trụ*, Lê Bích Loan*, Huỳnh Lê Minh*, Lê Thị Kim Vân*

*Học viện Công nghệ Bưu chính Viễn thông Cơ sở tại Thành phố Hồ Chí Minh

*Ban Quản lý Khu Công nghệ cao Thành phố Hồ Chí Minh

Tóm tắt: Khu Công nghệ cao TP.HCM (KCNC) là một trong ba KCNC quốc gia, trải qua 18 năm KCNC đã có hơn 160 dự án đầu tư và là KCNC thành công nhất trong cả nước. Để công tác quản lý và phân tích đánh giá thực trạng hoạt động của các dự án đầu tư tại KCNC theo đúng quy định của pháp luật về đầu tư cũng như Ban Quản lý KCNC (BQL) thực hiện nhiệm vụ tham mưu cho Ủy Ban Nhân Dân TPHCM (UBND) những số liệu cần thiết phục vụ cho công tác dự báo, định hướng quy hoạch phát triển kinh tế xã hội của Thành phố. Đây là thách thức không hề nhỏ đối với BQL, do đó mục tiêu của nghiên cứu là tập trung vào việc thiết kế và xây dựng hệ thống tích hợp dữ liệu về hoạt động của các dự án đầu tư để tăng cường công tác quản lý, phân tích đánh giá và khai thác dữ liệu dựa trên kỹ thuật dự báo chuỗi thời gian Facebook Prophet nhằm tạo ra các thông tin mang tính dự báo, hỗ trợ ra quyết định và góp phần định hướng xây dựng Thành phố thông minh.

Từ khóa: Phân tích dự báo, chuỗi thời gian, Prophet, đầu tư, KCNC.

I. ĐẶT VẤN ĐỀ

KCNC là một trong ba KCNC quốc gia, sau hơn 18 năm hình thành và phát triển, KCNC là địa điểm đầu tư lý tưởng với nhiều tập đoàn công nghệ cao trên thế giới quan tâm và muốn đặt dự án tại đây. Theo báo cáo số 1363/KCNC-QLKHCN&HTQT của KCNC [1], tính đến hết tháng 12/2020, KCNC có 161 dự án được cấp Giấy chứng nhận đăng ký đầu tư còn hiệu lực với tổng vốn là 44.563,258 tỷ đồng/110 dự án trong nước và 5.698,561 triệu USD/51 dự án nước ngoài. KCNC có 85 dự án đang hoạt động chiếm 52,8% và 76 dự án đang triển khai hoạt động chiếm 47,2%, trong đó có 33 dự án đang xây dựng và 43 dự án đang thực hiện các thủ tục chuẩn bị đầu tư. Ước giá trị sản xuất sản phẩm công nghệ cao năm 2020 của Khu Công nghệ cao đạt 20,69 tỷ USD tăng 23,3% so với cùng kỳ và tăng 3,45% so với kế hoạch đề ra, trong đó giá trị xuất khẩu đạt 19,08 tỷ USD tăng 20,15% và giá trị nhập khẩu đạt 17,49 tỷ USD tăng 21,2 so với cùng kỳ. Lũy kế đến cuối năm 2020, giá trị sản xuất sản phẩm của KCNC đạt 85,41 tỷ USD, trong đó giá trị xuất khẩu đạt 80,41 tỷ USD và giá trị nhập khẩu đạt 72,7 tỷ USD. Ước giá trị nộp ngân sách nhà nước năm 2020 của các dự án trong KCNC đạt khoảng từ 200 triệu USD trở lên. Giải quyết việc làm cho 45,096 lao động,

trong đó lao động trong nước là 44,576 lao động và lao động người nước ngoài là 520 lao động.

KCNC đang vào giai đoạn tăng tốc, đổi mới sáng tạo đáp ứng yêu cầu là đầu tàu cho đô thị sáng tạo của Thành phố Thủ Đức. Do đó yêu cầu quản lý số lượng lớn các doanh nghiệp với đa dạng lĩnh vực hoạt động và nguồn nhân lực phong phú nhiều quốc gia là một thách thức đối với BQL. Khó khăn lớn nhất hiện nay của BQL là xử lý nguồn dữ liệu về doanh nghiệp, các dự án đầu tư, dữ liệu về chuyên gia... tại KCNC đang vừa thiếu lại vừa thừa, vừa trùng lặp lại vừa mâu thuẫn dữ liệu do sự phân tán dữ liệu, sự quản lý thiếu tính nhất quán và sự chia sẻ dữ liệu chưa đầy đủ của các phòng/ban/đơn vị quản lý và các doanh nghiệp đang hoạt động trong KCNC. Điều này gây ra những trở ngại chính sau đây:

- Công tác quản lý, hỗ trợ các dự án đầu tư và tương tác giữa BQL với các doanh nghiệp chưa thật sự hiệu quả.
- Tiếp cận, cập nhật và phổ biến thông tin gặp nhiều khó khăn và chưa kịp thời.
- Công tác hoạch định chiến lược phát triển của KCNC và tham mưu cho UBND TP.HCM cũng như Chính phủ trong việc phát triển công nghệ cao chưa toàn diện.

Theo báo cáo đánh giá kiến trúc tổng thể KCNC giai đoạn 2019-2025 [2] do Công ty cổ phần Công nghệ DTT cập nhật ngày 23/04/2019 cho biết hiện tại KCNC chưa đầu tư nền tảng dịch vụ chia sẻ, tích hợp cho riêng mình, do vậy hầu hết các ứng dụng, đặc biệt là các ứng dụng dùng chung chưa được chia sẻ và tích hợp với nhau. Ngoài ra, các dữ liệu về doanh nghiệp, dự án đầu tư, dữ liệu về chuyên gia hoạt động trong các lĩnh vực khoa học công nghệ tại KCNC chưa được tổ chức khoa học, đang được quản lý rải rác dưới dạng excel và các dạng lưu trữ phi cấu trúc khác. Vì vậy, việc nghiên cứu thiết kế xây dựng một hệ thống tích hợp dữ liệu về hoạt động của các dự án đầu tư, phân tích để tạo ra các thông tin mang tính dự báo, chia sẻ dữ liệu doanh nghiệp dùng chung tại KCNC để giải quyết các khó khăn nêu trên cũng như để đáp ứng yêu cầu báo cáo thống kê và hỗ trợ ra quyết định là một nhu cầu cấp thiết đối với BQL. Điều này giúp Lãnh đạo của KCNC nắm bắt kịp thời tình hình hoạt động của doanh nghiệp và nhanh chóng đưa ra những điều chỉnh, giải pháp kịp thời cũng như các định hướng thích hợp về mặt chiến lược cho KCNC.

Bài báo này có 5 phần, các phần còn lại được trình bày

Tác giả liên hệ: Huỳnh Trọng Thừa,

Email: htthua@ptithcm.edu.vn

Đến tòa soạn: 22/2/2021, chỉnh sửa: 12/7/2021, chấp nhận đăng: 21/7/2021

như sau. Phần II trình bày các nghiên cứu liên quan. Phần III trình bày thiết kế hệ thống. Phần IV trình bày kết quả đạt được. Phần V sẽ kết luận cho bài báo này.

II. CÁC NGHIÊN CỨU LIÊN QUAN

Theo tìm hiểu của chúng tôi thì chưa có nghiên cứu nào công bố kết quả liên quan trực tiếp đến dữ liệu đầu tư trong các Khu Công nghiệp đặc biệt là các Khu Công nghệ cao. Vì vậy, trong phần này, chúng tôi trình bày các nghiên cứu liên quan đến các kỹ thuật tích hợp, phân tích và khai phá dữ liệu cho một số lĩnh vực kinh tế tương tự. Một vài nghiên cứu như thế được mô tả dưới đây.

Tân Hạnh và cộng sự đã thực hiện nghiên cứu [3] nhằm đề xuất giải pháp kỹ thuật tích hợp cơ sở dữ liệu doanh nghiệp phục vụ từ nhiều nguồn dữ liệu khác nhau cho cơ quan quản lý nhà nước tỉnh Bình Dương đảm bảo việc chia sẻ thông tin số phù hợp với quy định của pháp luật và đáp ứng các yêu cầu thống kê báo cáo thông tin doanh nghiệp cho các cơ quan quản lý Nhà nước của tỉnh Bình Dương. Đề tài chỉ dừng lại ở mức tích hợp dữ liệu từ nhiều nguồn mà chưa triển khai các giải pháp phân tích dữ liệu.

Một số nghiên cứu dựa trên các mô hình dự báo kinh tế lượng. Nguyễn Văn Huân và Lê Anh Tú thực hiện nghiên cứu [4] nhằm xây dựng hệ thống phân tích dữ liệu và dự báo doanh thu cho doanh nghiệp vừa và nhỏ trên cơ sở thu thập, phân tích và kiểm định dữ liệu nhằm đưa ra những giải pháp xây dựng kế hoạch hoạt động sản xuất kinh doanh sao cho đạt hiệu quả cao nhất. Các phương pháp sử dụng trong nghiên cứu gồm phương pháp hồi quy tuyến tính, điều hòa mũ, trung bình động giản đơn và trung bình động có trọng số. Nghiên cứu [5] do Đỗ Văn Đại thực hiện nhằm áp dụng một số phương pháp khai phá và dự báo như phương pháp chuyên gia, phương trình hồi quy và chuỗi thời gian để hoạch định và điều hành chính sách tài chính, xây dựng mô hình dự báo chỉ số thống kê xã hội chủ yếu, dự báo biến động giá chứng khoán, dự báo sự tác động của vốn đầu tư từ nước ngoài, dự báo giá một số mặt hàng tư liệu sản xuất.

Một số khác tập trung nghiên cứu xây dựng mô hình dự báo dựa trên phương pháp phân lớp và cây quyết định. Phạm Trung Kiên thực hiện nghiên cứu [6] nhằm đề xuất và xây dựng mô hình quản lý rủi ro bằng phương pháp khai phá dữ liệu. Mô hình này cung cấp cơ sở tri thức để hỗ trợ công tác quản lý rủi ro trên cả hai phương diện phân loại và dự đoán rủi ro. Dựa trên tính chất của dữ liệu cho vay, tác giả đã nghiên cứu và áp dụng thuật toán phân lớp Navie Bayes và cây quyết định để phân tích và dự báo rủi ro trong giao dịch tín dụng ngân hàng. Nghiên cứu [7] do Nguyễn Ngọc Tuấn thực hiện nhằm giải quyết bài toán duy trì khách hàng thuê bao di động. Theo đó tác giả dựa vào hành vi sử dụng của các thuê bao di động và nghiên cứu áp dụng kỹ thuật phân lớp gồm lựa chọn thuộc tính, phân tích đặc trưng và cây quyết định để xây dựng mô hình dự báo. Priyanga Chandrasekar và cộng sự thực hiện nghiên cứu [8] nhằm chứng minh cây quyết định là công cụ có giá trị để phân loại, mô tả và khái quát hóa dữ liệu. J48 là một thuật toán cây quyết định được sử dụng để tạo ra mô hình phân loại. Nghiên cứu này trình bày phương pháp cải thiện độ chính xác để khai thác cây quyết định bằng việc tiền xử lý dữ liệu. Nghiên cứu đã áp dụng phân tách bộ lọc được giám sát trên thuật toán J48 để xây dựng cây quyết định. Kết quả thu được từ các thực nghiệm cho thấy độ chính xác của J48 sau khi phân tách tốt hơn J48 trước khi phân tách. Nghiên cứu [9] do Niketa Gandhi và cộng sự thực

hiện dựa trên mã nguồn mở WEKA để áp dụng các kỹ thuật khai thác dữ liệu cho tập dữ liệu nông nghiệp. Độ nhạy, tính đặc trưng và độ chính xác đã được tính toán để đánh giá kết quả thực nghiệm. Nghiên cứu cho thấy các bộ phân loại J48 và LADTree cung cấp hiệu suất tốt nhất trong số các bộ phân loại được sử dụng cho vùng khí hậu bán khô của tập dữ liệu thu thập tại Ấn Độ.

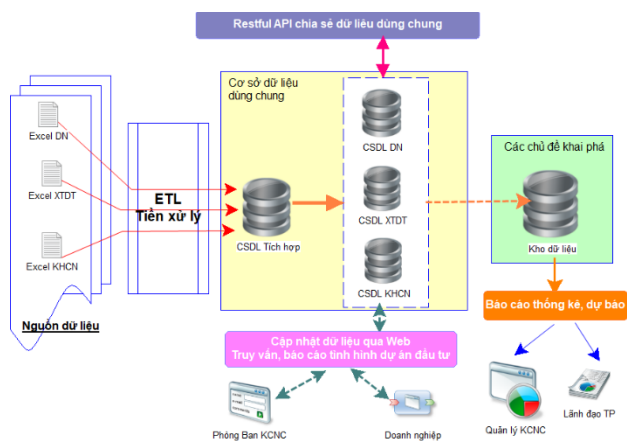
Một số nghiên cứu giải quyết bài toán dự báo dựa trên sự kết hợp của nhiều kỹ thuật khai phá và phương pháp khác nhau. Tác giả Bùi Quang Thành thực hiện nghiên cứu [10] bằng cách dựa vào các thuật toán tối ưu kết hợp phương pháp trí tuệ nhân tạo như mạng nơ-ron nhân tạo, vec-tơ hỗ trợ, hồi quy logic kết hợp mô hình tối ưu hóa Simulated Annealing. Dựa trên GIS, tác giả sử dụng các dữ liệu về vị trí mẫu bệnh sốt rét và đặc điểm dịch tễ học để xây dựng cơ sở dữ liệu các yếu tố ảnh hưởng đến nguy cơ bệnh sốt rét. Từ đó, tác giả kết hợp mô hình ANFIS (Adaptive Neuro-Inference System) và Simulated Annealing để đưa ra bản đồ nguy cơ sốt rét tại tỉnh Đắk Nông. Mukesh Kumar và A.J. Singh thực hiện nghiên cứu [11] nhằm phân tích dữ liệu sinh viên dựa trên nhiều thuật toán khác nhau như Cây quyết định, Naive Bayes, Random Forest, PART và Bayes Network. Sau khi thực hiện phân tích trên các số liệu khác nhau bằng các thuật toán khai thác dữ liệu khác nhau, nghiên cứu đã tìm thấy thuật toán Random Forest cho kết quả tốt nhất so với thuật toán khác. Việc phân tích các thuật toán khai phá dữ liệu khác nhau đã rút ra được một nhận thức đáng giá về cách các thuật toán này dự đoán hiệu suất học tập của sinh viên. Lưu Quốc Đạt cùng cộng sự đã trình bày mô hình quyết định đa tiêu chuẩn (MCDM) tích hợp để lựa chọn và phân nhóm nhà cung cấp xanh của Công ty TN2019 HH Canon Việt Nam trong nghiên cứu [12]. Trong đó, phương pháp phân tích thứ bậc (AHP) được sử dụng để xác định trọng số và phương pháp điểm lý tưởng (TOPSIS) được sử dụng để xếp hạng và phân nhóm các nhà cung cấp xanh. Lý thuyết tập mờ đã được ứng dụng nhằm lượng hóa giá trị tỷ lệ của các lựa chọn cũng như trọng số của các tiêu chuẩn đánh giá.

Gần đây, khi đại dịch Covid xuất hiện trên toàn thế giới, rất nhiều nhà khoa học đã sử dụng các mô hình dự báo chuỗi thời gian để hỗ trợ công tác phòng chống dịch, một vài trong đó là nghiên cứu [13] do Naresh Kumar và Seba Susan thực hiện nhằm dự báo số ca nhiễm và số ca tử vong do nhiễm Covid dựa trên hai mô hình ARIMA [14] và Facebook Prophet [15]. Parikshit N. Mahalle và cộng sự thực hiện nghiên cứu [16] để dự báo tình hình diễn biến của dịch Covid bằng mô hình Facebook Prophet. Nghiên cứu này chỉ ra rằng mô hình này cho kết quả dự báo chính xác hơn một số thuật toán máy học khác. Nguyen Quoc Duong và cộng sự thực hiện nghiên cứu [17] để tìm ra mô hình ARIMA phù hợp nhất trong cảnh báo diễn biến tiếp theo của đại dịch Covid-19.

III. THIẾT KẾ HỆ THỐNG

A. Mô hình hệ thống

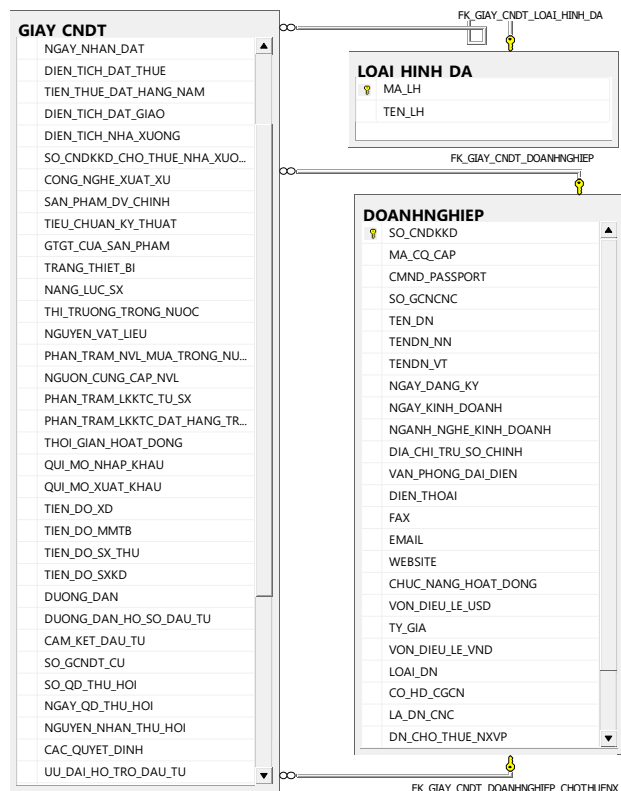
Kiến trúc tổng thể mô hình hệ thống phân tích dữ liệu doanh nghiệp tại KCNC được thể hiện như hình 1. Theo mô hình kiến trúc đề xuất, hệ thống chia làm 3 phần chính: thu thập dữ liệu từ nhiều nguồn (nhiều phòng ban quản lý của KCNC, doanh nghiệp và đặc biệt là nhiều định dạng dữ liệu khác nhau), xử lý và lưu trữ vào cơ sở dữ liệu tích hợp, phân tích và khai phá dữ liệu phục vụ công tác dự báo và điều hành hoạt động quản lý tại KCNC.



Hình 1. Kiến trúc hệ thống

B. Thiết kế cơ sở dữ liệu cho hệ thống

Cơ sở dữ liệu được thiết kế riêng cho từng phòng nghiệp vụ. Theo yêu cầu khảo sát và phân tích, chúng tôi thiết kế cơ sở dữ liệu phục vụ mục đích dùng chung cho 3 phòng (phòng doanh nghiệp, phòng xúc tiến đầu tư, và phòng quản lý khoa học công nghệ), sau đó tổng hợp đưa vào kho dữ liệu theo từng nghiệp vụ như hoạt động đầu tư, hoạt động khoa học công nghệ, nghiên cứu ươm tạo... Trong phạm vi nghiên cứu này, chúng tôi trình bày một số bảng dữ liệu chính liên quan trực tiếp đến mục tiêu phân tích hoạt động đầu tư tại KCNC. Một trong số đó thể hiện ở hình 2.



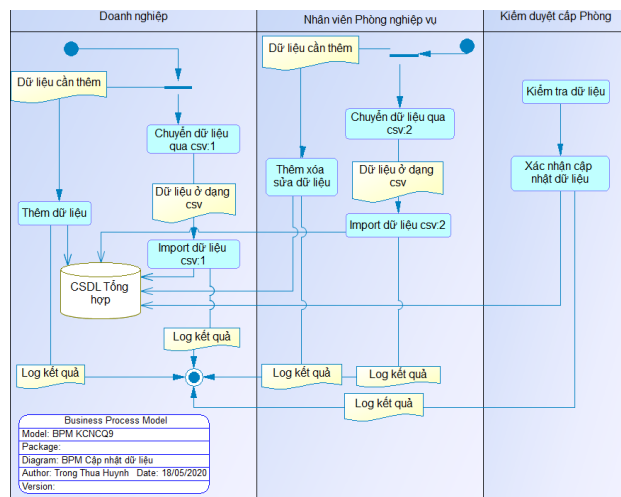
Hình 2. Các bảng dữ liệu chính của kho dữ liệu

Theo đó, doanh nghiệp muốn đăng ký dự án đầu tư thì phải được cấp phép thông qua Giấy chứng nhận đầu tư (bảng GIAY_CNDT) có các thông tin: số GCNDT, tên dự án, ngày đăng ký, qui mô dự án, mục tiêu hoạt động, vốn đầu tư dự án tính theo USD, vốn đầu tư dự án tính theo VND, tỷ giá, diện tích đất sử dụng, ngày nhận đất, dự án sử dụng công nghệ gì, xuất xứ từ đâu, sản phẩm, dịch vụ

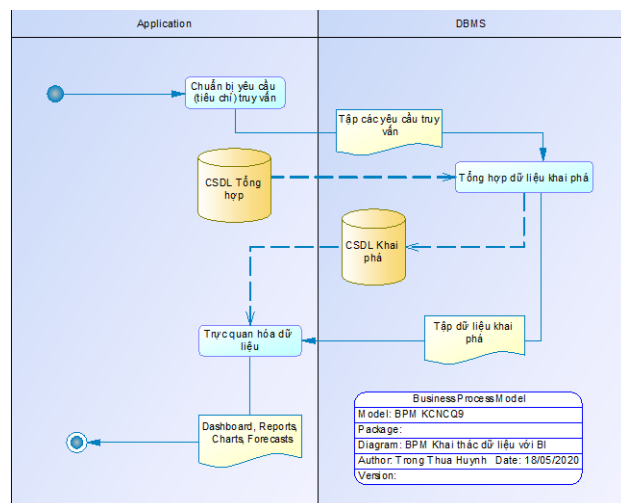
chính của dự án, loại hình đầu tư... Doanh nghiệp khi thành lập có các thông tin (chưa trong bảng DOANHNGHIEP) như số chứng nhận đăng ký kinh doanh, tên doanh nghiệp, tên doanh nghiệp, ngày đăng ký, ngày kinh doanh... Hình thức đầu tư chi tiết (bảng HINH_THUC_DAU_TU_CHI_TIET) chứa các loại và tên hình thức đầu tư như: Sản xuất, Dịch vụ, Đào tạo, Ươm tạo...

C. Thiết kế luồng chức năng hệ thống

Phần này chúng tôi trình bày 2 luồng chính: luồng cập nhật dữ liệu (hình 3) và luồng phân tích và khai thác dữ liệu (hình 4).



Hình 3. Luồng cập nhật dữ liệu cho hệ thống



Hình 4. Luồng phân tích và khai thác dữ liệu

D. Thiết kế giải pháp dự báo

Để đưa ra giải pháp dự báo, chúng tôi tiến hành thu thập yêu cầu và đánh giá dữ liệu có được tại KCNC, tổng hợp được một số nhu cầu dự báo và tập trung vào thiết kế giải pháp dự báo cho các dữ liệu hệ thống thu thập được dựa trên thuật toán phân tích dữ liệu chuỗi thời gian Facebook Prophet [15]. Prophet dùng thuật toán mã nguồn mở và thực hiện dự báo tốt hơn mà không yêu cầu tập dữ liệu huấn luyện trong các phương pháp dự báo dựa trên chuỗi thời gian. Đặc điểm chính của thuật toán này là nó xử lý chính xác hơn đối với dữ liệu chuỗi thời gian và được sử dụng chính trong dự báo và lập kế hoạch [16].

1) Phân tích chuỗi thời gian

Dữ liệu chuỗi thời gian là một chuỗi các quan sát được lưu trữ theo thứ tự thời gian. Dữ liệu chuỗi thời gian thường nổi bật khi theo dõi các số liệu kinh doanh, giám sát các quy trình công nghiệp... Phân tích chuỗi thời gian là một cách tiếp cận để phân tích dữ liệu có dạng chuỗi thời gian để rút ra các đặc điểm có ý nghĩa của dữ liệu và tạo ra những hiểu biết hữu ích khác được áp dụng trong tình hình kinh doanh.

Kỹ thuật dự báo chuỗi thời gian có thể trả lời các câu hỏi kinh doanh như lượng hàng tồn kho cần duy trì, lưu lượng truy cập trang web bán hàng trực tuyến hay bao nhiêu sản phẩm sẽ được bán trong tháng tới... Tất cả đều là những vấn đề quan trọng của chuỗi thời gian cần giải quyết. Mục tiêu cơ bản của phân tích chuỗi thời gian thường là xác định một mô hình mô tả mẫu của chuỗi thời gian và có thể được sử dụng để dự báo.

Các kỹ thuật dự báo chuỗi thời gian cổ điển (phổ biến nhất là ARIMA) xây dựng dựa trên các mô hình thống kê đòi hỏi nhiều nỗ lực để điều chỉnh mô hình, xử lý dữ liệu theo ngành. Chúng ta phải điều chỉnh các tham số của phương pháp liên quan đến vấn đề cụ thể khi một mô hình dự báo không hoạt động như mong đợi. Việc điều chỉnh các tham số này đòi hỏi sự hiểu biết kỹ lưỡng về cách hoạt động của các mô hình chuỗi thời gian cơ bản. Nhiều doanh nghiệp sẽ khó khăn trong vấn đề dự báo nếu không có nhóm chuyên viên về khoa học dữ liệu. Ngoài ra, việc thiếu một nền tảng dự báo đầy đủ hoặc các dịch vụ hỗ trợ khác sẽ không đáp ứng được các kỳ vọng của lãnh đạo.

2) Giải pháp dự báo cho dữ liệu chuỗi thời gian

Chúng tôi thiết kế giải pháp dự báo cho KCNC dựa trên nền tảng mã nguồn mở có tên là Prophet do Facebook phát triển, nó hỗ trợ cho cả Python và R bên cạnh việc cung cấp các thông số trực quan, dễ điều chỉnh. Ngay cả những người thiếu chuyên môn sâu về các mô hình dự báo chuỗi thời gian cũng có thể sử dụng Prophet để tạo ra các dự báo có ý nghĩa cho nhiều vấn đề trong các tình huống kinh doanh.

Việc đưa ra các dự báo chất lượng cao không phải là một vấn đề dễ dàng ngay cả đối với máy móc hoặc hầu hết các nhà phân tích. Nhóm chuyên gia khoa học dữ liệu của Facebook đã quan sát và nhận diện hai vấn đề chính trong quá trình tạo ra nhiều loại dự báo kinh doanh, đó là:

- Các kỹ thuật dự báo hoàn toàn tự động là không dễ dàng và chúng thường quá cứng nhắc để kết hợp các giả định hữu ích hoặc phân tích dựa vào kinh nghiệm.
- Các nhà phân tích có thể đưa ra các dự báo chất lượng cao là khá hiếm vì dự báo là một kỹ năng khoa học dữ liệu chuyên biệt đòi hỏi nhiều kinh nghiệm.

Dưới đây là một số ưu điểm của Prophet mà theo nhóm chuyên gia khoa học dữ liệu của Facebook chỉ ra:

- Rất nhanh, vì nó được xây dựng bằng Stan, một ngôn ngữ lập trình để suy luận thống kê được viết bằng C++.
- Một mô hình hồi quy cộng hưởng trong đó các xu hướng phi tuyến phù hợp với tính thời vụ hàng năm, hàng tuần và hàng ngày, cộng với các hiệu ứng ngày lễ:
 - Xu hướng đường cong tăng trưởng tuyến tính

hoặc logistic từng đoạn. Prophet tự động phát hiện những thay đổi trong xu hướng bằng cách chọn điểm thay đổi từ dữ liệu;

- Thành phần theo mùa hàng năm được lập mô hình bằng chuỗi Fourier;
 - Thành phần theo mùa hàng tuần sử dụng biến giả;
 - Danh sách các ngày lễ quan trọng do người dùng cung cấp.
- Mạnh mẽ ngay cả khi thiếu dữ liệu và thay đổi theo xu hướng, có khả năng xử lý tốt cho các trường hợp ngoại lệ.
 - Dễ dàng thay đổi quy trình và điều chỉnh dự báo cũng như bổ sung thông tin chi tiết về dữ liệu doanh nghiệp.

3) Mô hình dự báo dựa trên thuật toán Prophet

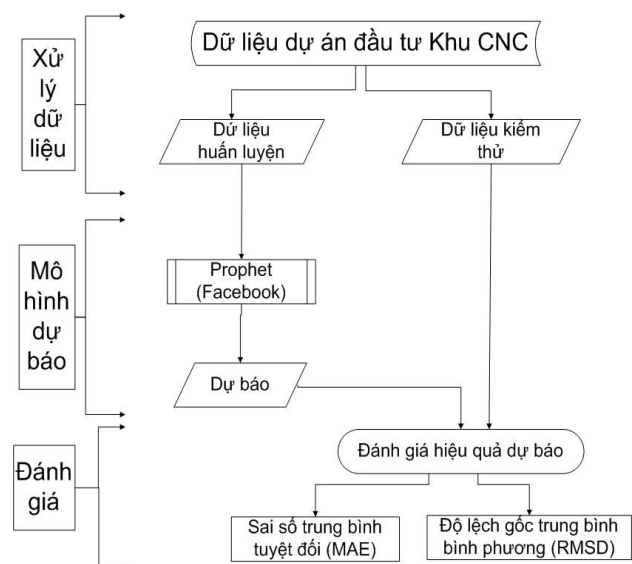
Prophet sử dụng mô hình chuỗi thời gian có thể phân tách với ba thành phần mô hình chính: xu hướng, tính thời vụ và ngày lễ. Chúng được kết hợp trong phương trình sau:

$$y(t) = g(t) + s(t) + h(t) + \epsilon_t \quad (1)$$

Trong đó:

- $g(t)$: đường cong tăng trưởng tuyến tính hoặc logistic từng đoạn để lập mô hình những thay đổi không theo chu kỳ trong chuỗi thời gian;
- $s(t)$: thay đổi định kỳ (hàng tuần, hàng tháng, hàng năm);
- $h(t)$: ảnh hưởng của ngày nghỉ (do người dùng cung cấp) với lịch trình không thường xuyên;
- ϵ_t : mô tả lỗi cho bất kỳ thay đổi bất thường nào không có trong mô hình.

Sử dụng thời gian như một bộ hồi quy, Prophet thực hiện điều chỉnh một số hàm tuyến tính và phi tuyến tính của thời gian như các thành phần của phương trình dự báo (1). Prophet làm đơn giản vấn đề dự báo như một dạng bài toán chỉnh sửa đường cong hơn là xem xét rõ ràng sự phụ thuộc dựa trên thời gian của mỗi quan sát trong một chuỗi thời gian.



Hình 5. Khung giải pháp dự báo

Sau cùng, tại giai đoạn đánh giá, thông tin dự báo được đem đi đánh giá với tập dữ liệu kiểm thử thông qua các chỉ số đo lường như sai số tuyệt đối trung bình MAE (Mean Absolute Error), sai số bình phương trung bình MSE (Mean Squared Error), phần trăm sai số tuyệt đối trung bình MAPE (Mean Absolute Percent Error), sai số bình phương gốc RMSE (Root Mean Squared Error).

IV. HIỆN THỰC HỆ THỐNG

A. Thu thập và xử lý dữ liệu

Để có được dữ liệu cần phân tích dự báo, chúng tôi thực hiện thu thập dữ liệu từ 2 cách chính: nhập liệu từ web (cho phép nhân viên của KCNC nhập dưới sự phê duyệt của Lãnh đạo Phòng và/hoặc đại diện doanh nghiệp với sự cho phép của BQL KCNC) và công cụ nhập dữ liệu hàng loạt từ các file dữ liệu nguồn excel.

Column Name	Data Type	Column Name	Data Type
SO_CNDKKD	nvarchar(20)	MA_CTHTDT	nchar(20)
TENDA_VT	nvarchar(100)	CHI_TIET_HINH_THUC_DAU_TU	nvarchar(100)
TEN_DN	nvarchar(400)	MA_LVCNC	nchar(20)
SO_GCNDT_CU	nvarchar(20)	LINH_VUC_CNC	nvarchar(100)
SO_GCNDT_MOI	nvarchar(20)	MUC_TIEU_HOAT_DONG	nchar(10)
NGAY_DANG_KY	datetime	PHAN_KHU	nvarchar(50)
MA_LH	nchar(10)	DIA_DIEM	nvarchar(1000)
TEN_LH	nvarchar(50)	DIEN_TICH_DAT_THUE	float
QUOC_GIA_FDI	nvarchar(100)	SO_CNDKKD_CHO_THUE_NHA_XU...	nchar(20)
ID_TTHS	int	TEN_DN_CHO_THUE	nvarchar(100)
TINH_TRANG	nvarchar(50)	DIEN_TICH_NHA_XUONG	float
NAM_CAP	int	SO_QD_THU_HOI	nvarchar(400)
VON_DAU_TU_USD	money	NGAY_QD_THU_HOI	date
VON_DAU_TU_VND	money	NGUYEN_NHAN_THU_HOI	nvarchar(500)
TY_GIA	float		
TEN_DA	nvarchar(400)		
MA_HTDT	nchar(20)		
HINH_THUC_DAU_TU	nvarchar(100)		

Cấu trúc Bảng dữ liệu dùng để lưu trữ dữ liệu tạm thời trong quá trình xử lý nhập dữ liệu hàng loạt

Hình 6. Bảng dữ liệu tạm phục vụ tiền xử lý dữ liệu

Để phục vụ việc thu thập dữ liệu qua giao diện web, chúng tôi thiết kế và hiện thực website cho giúp người sử dụng có thể nhập trực tiếp dữ liệu vào hệ thống để đưa vào kho dữ liệu, đồng thời website cũng cho phép Lãnh đạo Phòng/BQL KCNC có thể giám sát, xem thông tin báo cáo thông kê các thông tin liên quan đến hoạt động đầu tư của các doanh nghiệp tại KCNC. Chúng tôi hiện thực website bằng mô hình .NET MVC, Microsoft SQL Sever, Entity Framework và Bootstrap hỗ trợ responsive phù hợp với mọi thiết bị và độ phân giải màn hình.

Để nhập dữ liệu hàng loạt từ các file dữ liệu excel, chúng tôi thiết kế và cài đặt bảng dữ liệu như hình 6 để lưu dữ liệu tạm thời, xử lý dữ liệu trùng lặp, thiếu dữ liệu khóa chính trong các bảng dữ liệu trước khi đưa vào kho dữ liệu.

Trong phần hiện thực, chúng tôi đã thu thập được dữ liệu về 197 dự án đầu tư (22 dự án bị thu hồi), 202 doanh nghiệp tham gia. Trong đó 63 dự án đầu tư trực tiếp nước ngoài (FDI), 123 dự án đầu tư trong nước với tổng vốn đầu tư trên 70 nghìn tỷ đồng trong khoảng thời gian từ 2003 đến 2020.

B. Phân tích dự báo

Dựa trên số liệu thu thập như đề cập ở phần A, để thực hiện chức năng dự báo, chúng tôi sử dụng ngôn ngữ Python kết hợp với thuật toán dự báo dữ liệu chuỗi thời gian Prophet với cấu hình chuẩn như sau:

```
m = Prophet(interval_width=0.95, growth='linear',
weekly_seasonality=False, daily_seasonality=False,
holidays=None)
forecast = m.predict(future)
```

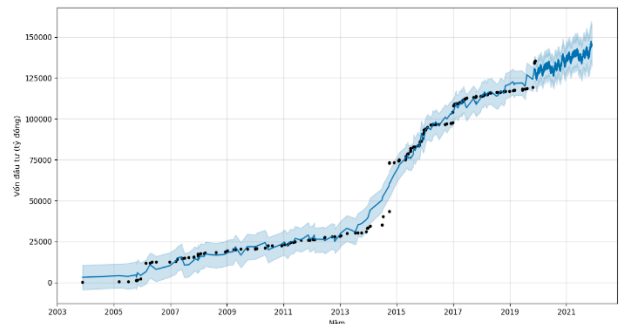
Mô hình dự báo đề xuất trong hình 5 được hiện thực trên nền Anaconda3 sử dụng các thư viện phân tích dữ liệu như pandas, pyodbc, fbprophet, matplotlib, numpy. Tất cả thư viện, nền tảng và tham số cấu hình chuẩn sử dụng ở trên được tham khảo từ [18].

Trong mô hình dự báo này, tham số khoảng tin cậy *interval_width* được thiết lập là 95% thể hiện mức độ không chắc chắn trong phương pháp lấy mẫu của thuật toán Prophet. Giá trị này thường dùng trong các mô hình dự báo dựa trên thông tin như ARIMA [17] hay các mô hình dự báo dữ liệu chuỗi thời gian tương tự [14].

Rất nhiều kết quả dự báo thu được trong quá trình phân tích dữ liệu, trong khuôn khổ bài báo này, chúng tôi chỉ trình bày một số kết quả chính dưới đây liên quan đến dòng vốn đầu tư vào KCNC.

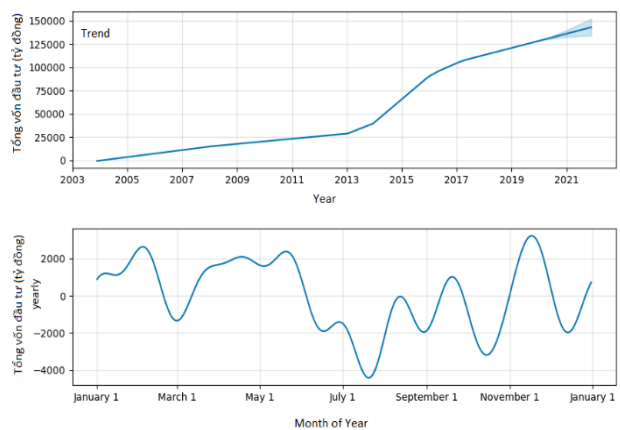
1) Dự báo vốn đầu tư vào KCNC

Hình 7 là kết quả dự báo 2 năm tới về tổng số vốn lũy kế bằng VND sẽ được đầu tư vào KCNC. Kết quả đường dự báo rất sát với đường dữ liệu thực tế đến năm 2020 (các chấm đen). Phần màu xanh ở cuối đường biến thiên thể hiện thông tin dự báo tổng số vốn đầu tư vào KCNC đến năm 2022.



Hình 7. Dự báo vốn đầu tư vào KCNC theo VND

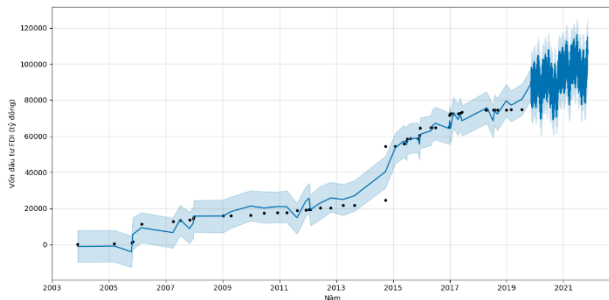
Hình 8 thể hiện xu hướng dòng vốn đầu tư vào KCNC theo các năm với khoảng tin cậy 95%. Cùng với đó là đường biểu diễn biến động số vốn đầu tư theo các tháng trong năm.



Hình 8. Xu hướng đầu tư vốn vào KCNC

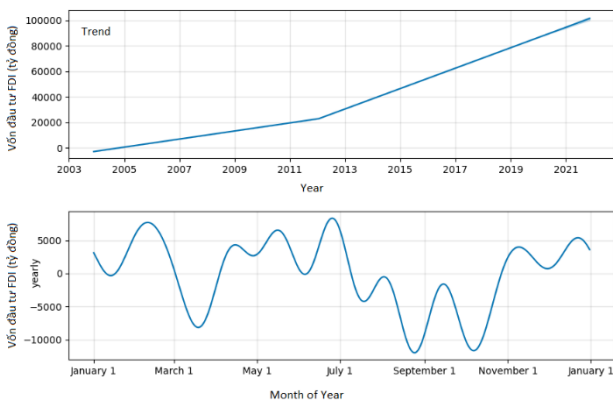
2) Dự báo vốn đầu tư FDI vào KCNC

Hình 9 là kết quả dự báo 2 năm tới về tổng số vốn đầu tư FDI lũy kế bằng VND sẽ được đầu tư vào KCNC. Kết quả đường dự báo rất sát với đường dữ liệu thực tế đến năm 2020 (các chấm đen). Phần màu xanh ở cuối đường biến thiên thể hiện thông tin dự báo số vốn đầu tư FDI vào KCNC đến năm 2022.



Hình 9. Dự báo vốn FDI đầu tư vào KCNC

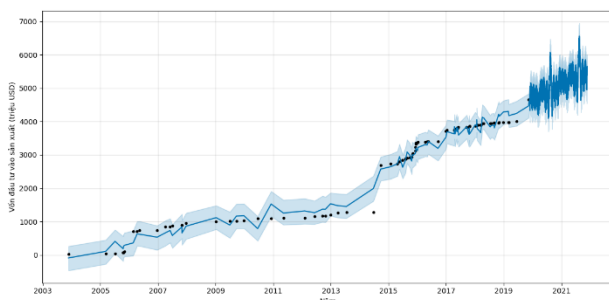
Hình 10 thể hiện xu hướng dòng vốn FDI đầu tư vào KCNC theo các năm với khoảng tin cậy 95%. Cùng với đó là đường biểu diễn biến động số vốn đầu tư FDI theo các tháng trong năm.



Hình 10. Xu hướng vốn FDI đầu tư vào KCNC

3) Dự báo vốn đầu tư vào sản xuất vào KCNC

Hình 11 là kết quả dự báo 2 năm tới về tổng số vốn đầu tư bằng USD sẽ được đầu tư vào KCNC để hoạt động sản xuất (những hoạt động khác trong KCNC ngoài sản xuất gồm có: dịch vụ công nghệ cao, nghiên cứu phát triển ương tạo, đào tạo...). Kết quả đường dự báo rất sát với đường dữ liệu thực tế đến năm 2020 (các chấm đen). Phần màu xanh ở cuối đường biến thiên thể hiện thông tin dự báo số vốn đầu tư vào lĩnh vực sản xuất đến năm 2022.



Hình 11. Dự báo vốn USD đầu tư lĩnh vực sản xuất

Hình 12 thể hiện xu hướng dòng vốn đầu tư vào hoạt

động sản xuất tại KCNC với khoảng tin cậy 95%. Cùng với đó là đường biểu diễn biến động số vốn đầu tư FDI theo các tháng trong năm.

Nhìn chung, biểu đồ dự báo của dòng vốn đầu tư FDI và đầu tư vào lĩnh vực sản xuất là khá tương đồng nhau. Do vậy, có thể thấy rằng đa phần vốn đầu tư FDI vào KCNC chủ yếu là để hoạt động sản xuất mà không tham gia vào các lĩnh vực khác như dịch vụ hay nghiên cứu ương tạo...

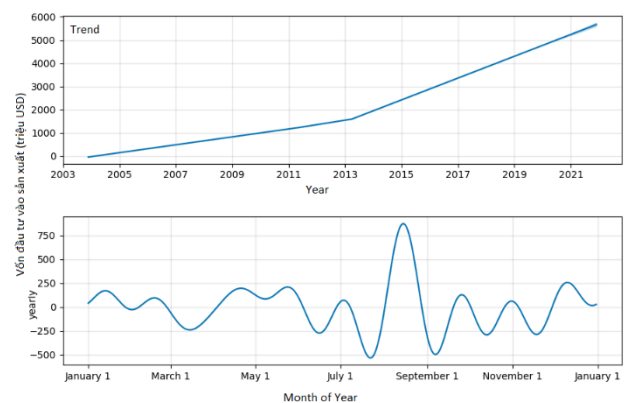
4) Đánh giá sai số của kết quả dự báo

Có nhiều tiêu chí đánh giá chất lượng hay sự phù hợp của mô hình như sai số tuyệt đối trung bình MAE, phần trăm sai số tuyệt đối trung bình MAPE, sai số bình phương gốc RMSE như mô tả trong mô hình dự báo ở hình 5. Chúng tôi thực hiện cross-validation [19] để đánh giá hiệu suất dự đoán trong 365 ngày cho ba tập dữ liệu huấn luyện (tổng vốn đầu tư, vốn FDI, và vốn đầu tư vào sản xuất). Kết quả các chỉ số thu được như bảng I là chúng tôi lấy giá trị tốt nhất (min) với hàm cross-validation [18] và performance_metrics [18] với khoảng horizon [18] là 365 ngày.

Bảng I. Hiệu suất của mô hình dự báo qua các tiêu chí đánh giá sai số

	RMSE	MAE	MAPE
Tổng vốn đầu tư	16997,06089	12249,64	0,297411
Vốn FDI	24506,28	20360,43	0,64090621
Vốn đầu tư vào sản xuất	707,5835	509,0966	0,3213

Theo bảng I, có sự khác biệt khác lớn giữa 3 tập dữ liệu dùng trong dự báo cho thấy sự thay đổi bất thường trong sai số ϵ_t trong phương trình (1). Ngoài ra, mục đích là đánh giá sai số dự báo với những bộ số liệu khác nhau chứ không đánh giá thuật toán dự báo Prophet nên chỉ tiêu MAPE được sử dụng trong trường hợp này thay vì dựa vào hai tiêu chí RMSE và MAE. Rõ ràng, theo bảng I, giá trị MAPE tốt nhất cho các tập dữ liệu tổng vốn đầu tư, vốn FDI, và vốn đầu tư vào sản xuất lần lượt là 0.29, 0.64 và 0.32 cho thấy kết quả dự báo có khả năng chấp nhận khá cao.



Hình 12. Xu hướng vốn đầu tư vào lĩnh vực sản xuất

V. KẾT LUẬN

Bài báo đã trình bày những thách thức mà BQL KCNC đối diện trong công tác quản lý hoạt động của các dự án

đầu tư cũng như công tác hoạch định chính sách và tham mưu cho cấp lãnh đạo cao hơn. Kết quả nghiên cứu này cũng là cơ sở để xây dựng các mô hình dự báo thông minh cho các Khu công nghệ cao khác trong cả nước. Trong nghiên cứu này, chúng tôi đã thiết kế và xây dựng hệ thống tích hợp dữ liệu về hoạt động của các dự án đầu tư gồm các chức năng quan trọng như công cụ thu thập dữ liệu, mô hình phân tích và dự báo hoạt động đầu tư tại KCNC dựa trên thuật toán Prophet của Facebook. Bên cạnh đó, nghiên cứu cũng thực hiện đánh giá mô hình dự báo đề xuất thông qua dữ liệu thực tế tại KCNC và nhận được các kết quả khá chính xác về các thông tin dự báo với các sai số đánh giá mô hình trong ngưỡng chấp nhận được. Hướng phát triển tiếp theo là đưa vào mô hình một số thuật toán khai phá dữ liệu khác để đánh giá hiệu suất của từng thuật toán ứng với từng tập dữ liệu huấn luyện.

LỜI CẢM ƠN

Nghiên cứu này được thực hiện bởi sự hỗ trợ nguồn dữ liệu từ BQL KCNC để thực hiện nhiệm vụ “Phân tích dữ liệu doanh nghiệp cho Khu Công nghệ cao TP.HCM” được phê duyệt bởi Sở Khoa học và Công nghệ TP.HCM. Cảm ơn BQL KCNC đã hỗ trợ nhóm thực hiện nghiên cứu này.

TÀI LIỆU THAM KHẢO

- [1] B. KCNC, “Báo cáo kết quả hoạt động của Khu Công nghệ cao TP.HCM,” Ban Quản lý Khu Công nghệ cao TP.HCM, 2020.
- [2] DTT, “Kiến trúc tổng thể CNTT Khu Công nghệ cao TP.HCM Giai đoạn 2019 - 2025,” Ban Quản lý Khu Công nghệ cao TP.HCM, 2019.
- [3] Tân Hạnh, Lưu Nguyễn Kỳ Thư, Huỳnh Trọng Thừa, “Giải pháp tích hợp dữ liệu doanh nghiệp phục vụ cho các cơ quan quản lý nhà nước tỉnh Bình Dương,” Sở Thông tin và Truyền thông Tỉnh Bình Dương, 2013.
- [4] Nguyen Van Huan, Le Anh Tu, “Giải pháp xây dựng hệ thống phân tích dữ liệu và dự báo doanh thu cho doanh nghiệp vừa và nhỏ tại Việt Nam,” *Tạp chí KHOA HỌC & CÔNG NGHỆ*, vol. 135, no. 5, pp. 191-198, 2015.
- [5] D. V. Dai, “Hệ thống phân tích dữ liệu và dự báo kinh tế,” Đại học Quốc gia Hà Nội, 2016.
- [6] P. T. Kien, “Khai thác và phân tích dữ liệu nhằm quản lý rủi ro trong giao dịch tín dụng,” Đại học Quốc gia Hà Nội, 2019.
- [7] N. N. Tuan, “Áp dụng kỹ thuật khai phá dữ liệu dự báo thuê bao rời mạng trong mạng di động,” Đại học Quốc gia Hà Nội, 2016.
- [8] Priyanga Chandrasekar et al., “Improving the Prediction Accuracy of Decision Tree Mining with Data Preprocessing,” in *IEEE 41st Annual Computer Software and Applications Conference (COMPSAC)*, 2017.
- [9] Niketa Gandhi et al., “Application of data mining techniques for predicting rice crop yield in semi-arid climatic zone of India,” in *2017 IEEE Technological Innovations in ICT for Agriculture and Rural Development (TIAR)*, 2017.
- [10] B. Q. Thanh, “Nghiên cứu tích hợp viễn thám, hệ thống tin địa lý và mô hình khai phá dữ liệu trong đánh giá nguy cơ mắc bệnh sốt rét tại tỉnh Đắk Nông,” Đại học Quốc gia Hà Nội, 2019.
- [11] Mukesh Kumar, A J Singh, “Evaluation of Data Mining Techniques for Predicting Student's Performance,” *International Journal of Modern Education and Computer Science*, vol. 8, no. 8, pp. 25-31, 2017.
- [12] Luu Quoc Dat et al., “Xây dựng mô hình ra quyết định đa tiêu chuẩn tích hợp để lựa chọn và phân nhóm nhà cung cấp xanh,” *VNU Journal of Science Economics and Business*, vol. 33, no. 1, pp. 43-54, 2017.
- [13] Naresh Kumar, Seba Susan, “COVID-19 Pandemic Prediction using Time Series,” in *11th International Conference on Computing, Communication and Networking Technologies*, 2020.
- [14] S. S. Wulff, “Time series analysis: Forecasting and control,” *Journal of Quality Technology*, vol. 49, no. 4, 2017.
- [15] Sean J. Taylor, Benjamin Letham, “Forecasting at Scale,” *THE AMERICAN STATISTICIAN*, vol. 72, no. 1, pp. 37-45, 2018.
- [16] Parikshit N. Mahalle et al., “Data Analytics: COVID-19 Prediction Using Multimodal Data,” *Intelligent Systems and Methods to Combat Covid-19*, pp. 1-10, 2020.
- [17] Nguyen Quoc Duong et al., “Predicting the Pandemic COVID-19 Using ARIMA Model,” *VNU Journal of Science: Mathematics - Physics*, vol. 36, no. 4, 2020.
- [18] F. O. Source, “Diagnostics,” Facebook, [Online]. Available: <https://facebook.github.io/prophet/docs/diagnostics.html>. [Accessed 14 July 2021].
- [19] P. Gupta, “Cross-Validation in Machine Learning,” Towards Data Science, 05 June 2017. [Online]. Available: <https://towardsdatascience.com/cross-validation-in-machine-learning-72924a69872f>. [Accessed 02 August 2020].

ANALYSIS AND FORECAST OF INVESTMENT ACTIVITIES IN SAIGON HIGH TECHNOLOGY PARK

Abstract: Saigon Hi-Tech Park (SHTP) is one of the three national hi-tech parks. Over the past 18 years, SHTP has had more than 160 investment projects and is the most successful Hi-Tech Park in our country. In order for the management and analysis of the current status of investment projects in the SHTP in accordance with the law on investment as well as the Management Board of SHTP (BOM) to perform the task of advising Ho Chi Minh People's Committee of necessary data to serve the forecasting, orientation and planning of socio-economic development of the City. This is not a small challenge for the BOM, so the aim of this study is to focus on designing and developing an integrated data system of the operation of investment projects to enhance management, analysis, evaluation and exploiting data based on Facebook Prophet, a time series forecasting technique, to create predictive information which supports decision making and contributes to the direction of developing the smart City.

Keywords: predictive analytics, time series, Prophet, investment, SHTP.



Huỳnh Trọng Thừa, tốt nghiệp Cử nhân Công nghệ Thông tin tại Đại học Khoa học Tự nhiên TP.HCM, Thạc sĩ Kỹ thuật Máy tính tại Đại học Kyung Hee (Hàn Quốc), và Tiến sĩ Khoa học Máy tính tại Đại học Bách Khoa - Đại học Quốc gia TP.HCM. Lĩnh vực nghiên cứu: An toàn phần mềm, mật mã học, điều tra số.

Email: htthua@ptithcm.edu.vn

Tân Hạnh, hiện là Phó Giám đốc Học viện Công nghệ Bưu chính Viễn thông. Lĩnh vực nghiên cứu chính: Máy học, thu thập thông tin, và khai phá dữ liệu.

Email: tanhanh@ptithcm.edu.vn