UBND TỈNH BÌNH DƯƠNG

**TRƯỜNG ĐẠI HỌC THỦ DẦU MỘT**

**TRẦN NGUYỄN BẢO KHANH**

**KHAI PHÁ DỮ LIỆU DỰA TRÊN BÁO CÁO TÌNH HÌNH, KẾT QUẢ GIẢI QUYẾT THỦ TỤC HÀNH CHÍNH THUẾ TẠI CỤC THUẾ BÌNH DƯƠNG**

**CHUYÊN NGÀNH: HỆ THỐNG THÔNG TIN**

**MÃ SỐ: 8480104**

**LUẬN VĂN THẠC SĨ**

**BÌNH DƯƠNG, NĂM 2020**

UBND TỈNH BÌNH DƯƠNG

**TRƯỜNG ĐẠI HỌC THỦ DẦU MỘT**

**TRẦN NGUYỄN BẢO KHANH**

**KHAI PHÁ DỮ LIỆU DỰA TRÊN BÁO CÁO TÌNH HÌNH, KẾT QUẢ GIẢI QUYẾT THỦ TỤC HÀNH CHÍNH THUẾ TẠI CỤC THUẾ BÌNH DƯƠNG**

**CHUYÊN NGÀNH: HỆ THỐNG THÔNG TIN**

**MÃ SỐ: 8480104**

**LUẬN VĂN THẠC SĨ**

**NGƯỜI HƯỚNG DẪN KHOA HỌC: TS NGUYỄN QUỐC HUY**

**BÌNH DƯƠNG, NĂM 2020**

# **LỜI CAM ĐOAN**

Tôi cam đoan luận văn này: “**Ứng dụng học máy vào việc nhận diện tài xế ngủ gật khi lái xe.**” là công trình nghiên cứu của…...

Các tài liệu tham khảo, sản phẩm nghiên cứu sử dụng cho luận văn này được trích dẫn theo đúng quy định.

Bình Dương, ngày 01 tháng 12 năm 2020

Học viên thực hiện luận văn

Nguyễn Hữu Nghĩa

# **LỜI CẢM ƠN**

Lời đầu tiên tôi xin bày tỏ lòng biết ơn chân thành TS Nguyễn Quốc Huy, Thầy đã tận tình chỉ dẫn, định hướng và truyền đạt những kiến thức cho tôi suốt thời gian thực hiện luận văn này.

Tôi xin bày tỏ lòng biết ơn chân thành đến TS. Bùi Thanh Hùng đã hỗ trợ hướng dẫn cùng với quí Thầy Cô giáo trong khoa Kỹ thuật - Công nghệ, Phòng Đào tạo Sau đại học – Trường đại học Thủ Dầu Một đã trang bị cho tôi những kiến thức nền tảng quan trọng và hỗ trợ tận tình trong suốt quá trình tôi theo học.

Tôi xin chân thành cảm ơn Lãnh đạo Cục thuế tỉnh Bình Dương cùng các anh em đồng nghiệp phòng Cục thuế tỉnh Bình Dương đã tạo điều kiện và cung cấp số liệu cho tôi hoàn thiện luận văn này.

Cảm ơn gia đình đã quan tâm, động viên ủng hộ trong suốt thời gian tôi tham gia học tập và thực hiện luận văn này.

Cảm ơn các anh chị đồng nghiệp, bạn bè lớp Cao học Hệ thống thông tin khóa 2017-2019 đã giúp đỡ và đồng hành cùng tôi trong suốt thời gian học tập tại Trường Đại học Thủ Dầu Một.

Mặc dù đã cố gắng song luận văn cũng không tránh khỏi những thiếu sót. Tôi rất mong nhận được những ý kiến đóng góp của quý Thầy Cô để tôi có thể hoàn thiện hơn đề tài của mình.

Xin trân trọng cảm ơn.

# **TÓM TẮT**

Ngày nay Ngành thuế Việt Nam nói chung và Cục thuế Bình Dương nói riêng thực hiện cải cách thủ tục hành chính mang đến sự hài lòng cho người nộp thuế và tăng nguồn thu ngân sách. Dự báo thực hiện cải cách thủ tục hành chính tại Cục thuế Bình Dương từ đó đưa ra giải pháp phù hợp để nâng cao chất lượng dịch vụ hành chính công là nội dung cấp thiết hiện nay. Trong nghiên cứu này chúng tôi đề xuất giải pháp Phân tích dự báo dịch vụ thủ tục hành chính thuế tại Cục thuế tỉnh Bình Dương bằng phương pháp học máy.

Dựa trên số liệu các dịch vụ thủ tục hành chính thuế tại Cục thuế tỉnh Bình Dương trong thời gian 3 năm 2017, 2018, 2019 và nửa đầu năm 2020 đề tài đã tiến hành:

Phân tích so sánh và trực quan hóa số liệu các dịch vụ thủ tục hành chính thuế để thấy được sự khác biệt giữa các quy trình thủ tục thuế và bức tranh tổng quát về các dịch vụ thủ tục hành chính thuế tỉnh Bình Dương.

Từ các yếu tố ảnh hưởng tới thủ tục hành chính thuế và số lượng hồ sơ báo cáo sử dụng dịch vụ thủ tục hành chính thuế, chúng tôi đã tiến hành Dự báo sự phát triển số lượng báo cáo giải quyết TTHC tại Cục thuế tỉnh Bình Dương bằng 4 phương pháp học máy: Random Forest, Linear Regression, K-Nearest Neighbors và Decision Trees. Trên cơ sở thực nghiệm, đánh giá so sánh các kết quả đã đạt được của các mô hình, chúng tôi đã lựa chọn mô hình tối ưu để dự báo sự phát triển số hồ sơ được giải quyết TTHC tại Cục thuế tỉnh Bình Dương. Kết quả thực nghiệm đã chứng minh rằng mô hình dự báo sử dụng phương pháp học máy Random Forest có độ chính xác cao nhất.

# **MỤC LỤC**

[LỜI CAM ĐOAN iii](#_Toc58940593)

[LỜI CẢM ƠN iv](#_Toc58940594)

[TÓM TẮT v](#_Toc58940595)

[DANH MỤC HÌNH vii](#_Toc58940596)

[DANH MỤC BẢNG viii](#_Toc58940597)

[DANH MỤC THUẬT NGỮ VÀ CÁC TỪ VIẾT TẮT ix](#_Toc58940598)

[Chương 1. GIỚI THIỆU CHUNG 1](#_Toc58940599)

[1.1 Tổng quan về công tác thuế và cục thuế tỉnh Bình Dương 1](#_Toc58940600)

[1.2 Khái niệm thủ tục hành chính thuế 2](#_Toc58940601)

[1.3 Tổng quan về phát hiện tri thức và khai phá dữ liệu 3](#_Toc58940602)

[Chương 2. CƠ SỞ LÝ THUYẾT VÀ CÁC NGHIÊN CỨU CÓ LIÊN QUAN 13](#_Toc58940603)

[2.1 Một số phương pháp dự báo 13](#_Toc58940604)

[2.2 Các nghiên cứu liên quan 15](#_Toc58940605)

# Chương 3. MÔ HÌNH ĐỀ XUẤT KHAI PHÁ DỮ LIỆU DỰA TRÊN BÁO CÁO TÌNH HÌNH, KẾT QUẢ GIẢI QUYẾT THỦ TỤC HÀNH CHÍNH THUẾ TẠI CỤC THUẾ TỈNH BÌNH DƯƠNG 17

[3.1 Mô hình đề xuất 17](#_Toc58940606)

3.2 Các mô hình sử dụng trong đề tài 19

[3.2.1 Linear Regression 19](#_Toc58940607)

[3.2.2 K - Nearest Neighbors 21](#_Toc58940608)

[3.2.3 Decision Trees 23](#_Toc58940609)

3.2.4 Random Forest 26

[Chương 4. THỰC NGHIỆM VÀ ĐÁNH GIÁ KẾT QUẢ 30](#_Toc58940610)

[4.1 Dữ liệu 30](#_Toc58940611)

[4.2 Kết quả thực nghiệm 33](#_Toc58940612)

[Chương 5. KẾT LUẬN VÀ HƯỚNG PHÁT TRIỂN 41](#_Toc58940613)

[TÀI LIỆU THAM KHẢO 42](#_Toc58940614)

# **DANH MỤC HÌNH**

Hình 1.1. Quy trình khai phá dữ liệu 7

Hình 3.1: Mô hình dự báo kết quả giải quyết thủ tục hành chính cục Thuế tỉnh Bình Dương 17

Hình 3.2: Mô hình hồi quy tuyến tính 20

Hình 3.3: Mô hình KNN 22

Hình 3.4: Mô hình Cây quyết định 25

Hình 3.5: Mô hình Random forest cho vấn đề hồi quy 28

Hình 4.1: Số lượng giải quyết hồ sơ đúng hạn và quá hạn tháng 12 năm 2018 34

Hình 4.2: Số lượng giải quyết hồ sơ đúng hạn và quá hạn tháng 10 năm 2020 35

Hình 4.3: Biểu đồ thể hiện độ đo trung bình của các mô hình 39

Hình 4.4: Biểu đồ thể hiện độ chính xác trung bình khi dự báo bằng mô hình RF 40

# **Chương 1. GIỚI THIỆU CHUNG**

Trong chương này, chúng tôi đã trình bày một cách tổng quan về tỉnh Bình Dương, tổng quan về cục thuế tỉnh Bình Dương sử dụng khai phá dữ liệu vào xử lý các công việc liên quan. Trình bày tổng quan về khai phá dữ liệu và các bước khai phá dữ liệu dựa trên báo cáo tình hình, kết quả giải quyết thủ tục hành chính thuế tại Cục thuế Bình Dương.

## **1.1 Tổng quan về** công tác thuế và cục thuế tỉnh Bình Dương

Thời đại công nghệ 4.0 đánh dấu kỷ nguyên lên ngôi của trí tuệ nhân tạo. Việc ứng dụng trí tuệ nhân tạo trong công tác quản lý nhà nước và xã hội là tất yếu khách quan và đang diễn ra rộng khắp trên thế giới. Tại Đại hội kế toán thuế công châu Á-Thái Bình Dương (AOTCA) lần thứ 17 tổ chức tại Busan - Hàn Quốc tháng 10/2019 vừa qua, vấn đề ứng dụng AI trong quản lý thuế được nhắc tới như một chủ đề chính và nhận được sự quan tâm sâu sắc của các đại biểu tham dự.

Quản lý thuế và tư vấn thuế là công việc phức tạp, liên quan nhiều đến pháp luật và con người. Các công việc đó cần dựa trên cơ sở từ các số liệu kế toán, tài chính để đưa ra kết quả phân tích, nhận định phục vụ hoạt động kinh doanh, nhằm đảm bảo việc tuân thủ pháp luật thuế với chi phí hợp lý nhất.

Bình Dương là một tỉnh nằm ở vùng trọng điểm kinh tế phía nam, hiện nay có 29 khu công nghiệp và hơn 12 cụm công nghiệp đang hoạt động với trên 32 ngàn doanh nghiệp đăng ký và hơn một triệu lao động đang làm việc. Vì vậy Khai phá dữ liệu dựa trên báo cáo tình hình, kết quả giải quyết thủ tục hành chính thuế tại Cục thuế Bình Dương đã được thực hiện trong những năm gần đây với các công cụ thô, nên tốn nhiều công sức, thời gian và độ chính xác chưa cao. Chính vì thế, để đáp ứng được sự phát triển nhanh của các thủ tục hành chính thuế đòi hỏi cần có một giải pháp hiện đại, hiệu quả. Trong nghiên cứu này chúng tôi đề xuất giải pháp dựa trên học máy vào Phân tích dự báo số lượng các hồ sơ giải quyết tại Cục thuế tỉnh Bình Dương. Kết quả dự đoán nhanh, chính xác sẽ là tiền đề giúp lãnh đạo hoạch định chính sách, chiến lược, đầu tư kinh doanh phù hợp cho tương lai.

Do đó việc nghiên cứu xác định và phân tích dự đoán sự phát triển các thủ tục giải quyết hồ sơ của Ngành thuế Việt Nam nói chung và phân tích dự báo số lượng các hồ sơ được giải quyết tại Cục thuế tỉnh Bình Dương nói riêng nhằm giúp Ban lãnh đạo biết được để xác định, định hướng bản chất sự phát triển các thủ tục thuế, hoàn thuế GTGT, hoàn thuế TNDN, hoàn thuế TNCN, ứng dụng CNTT… định hướng xây dựng quy hoạch, kế hoạch chiến lược phát triển trong những năm tới đó cũng là lý do chúng tôi chọn đề tài này: “Khai phá dữ liệu dựa trên báo cáo tình hình, kết quả giải quyết thủ tục hành chính thuế tại Cục thuế Bình Dương”.

## **1.2 Khái niệm** thủ tục hành chính thuế

TCT tên đầy đủ là Tổng cục thuế là tổ chức được thành lập năm 1945 có chi nhánh rộng khắp trên 63 tỉnh thành trên toàn quốc hiện đang hỗ trợ các thủ tục hành chính thuế như: tài liệu, hoá đơn, chính sách thủ tục thuế từ người nộp thuế.

DVTTHCT: dịch vụ thủ tục hành chính thuế là loại hàng hoá đặc biệt, sản phẩm dịch vụ vô hình, phục vụ chức năng công quyền, loại DVTTHCT do Cục thuế thực hiện, có đặc điểm sau: Quá trình thực hiện và cung cấp dịch vụ hành chính đều bị điều tiết rất chặt chẽ bởi quy định của pháp luật, mỗi người sử dụng dịch vụ đều có quyền ngang nhau trong việc sử dụng dịch vụ.

Phạm vi giải quyết DVTTHCT

Thủ tục “nộp thuế”:

Thủ tục “khai báo thuế”

Thủ tục “hoàn thuế “

Thủ tục “kiểm tra”

Thủ tục “thanh tra”

Ưu điểm của các thủ tục hành chính:

Thủ tục hành chính “khai báo thuế” nhanh, tiết kiệm thời gian, linh hoạt có độ tin cậy và uy tín cao.

Thủ tục hành chính “nộp thuế” thuận lợi, nhanh, tiết kiệm thời gian, chi phí

Thủ tục hành chính “hoàn thuế” nhanh, hỗ trợ kịp thời

Thủ tục hành chính “kiểm tra” nhanh, chuyên nghiệp,

Thủ tục hành chính “thanh tra” nhanh, chuyên nghiệp, phát hiện hành vi gian lận kịp thời, chống thất thoát nguồn ngân sách

Một số nhược điểm của thủ tục hành chính thuế

So với một số thủ tục hành chính khác, việc giải quyết hồ sơ còn chậm hơn

Người nộp thuế phải mang đến tận nơi để giải quyết thủ tục hành chính thuế

Kết quả giải quyết thủ tục hành chính thuế không được công khai minh bạch.

## **1.3 Tổng quan về** phát hiện tri thức và khai phá dữ liệu

Ngày nay, công nghệ thông tin đã trở thành một trong những động lực quan trọng của sự phát triển. Với khả năng số hóa mọi thông tin (số, đồ thị, văn bản, hình ảnh, âm thanh, tiếng nói,…), máy tính đã trở thành một công cụ thông minh, nó được sử dụng để xử lý thông tin với nhiều dạng thông tin thuộc nhiều lĩnh vực khác nhau trong đời sống như: kinh doanh, y học,…

Bên cạnh đó, cùng với sự phát triển của công nghệ lưu trữ dữ liệu phục vụ trong công việc lưu trữ các thông tin liên quan đến nhiều mặt của cuộc sống: kinh doanh, buôn bán, … đã góp phần cải thiện cuộc sống và làm giảm bớt đi việc lưu trữ thông tin dựa trên văn bản.

Đó chính là tiền đề cho sự ra đời của nền kinh tế mới – nền kinh tế số (hay có thể gọi là nền kinh tế tri thức, nền kinh tế dựa trên tri thức). Nền kinh tế đó đã và đang làm cho sự phát triển thông tin lưu trữ ngày càng nhiều, và khả năng linh hoạt của các phần mềm phải đảm đương nhiều công việc trong việc lựa chọn thông tin. Với khối lượng dữ liệu tăng nhanh và khổng lồ như vậy, trong luận văn này dữ liệu của chúng tôi sử dụng được trích từ bộ cơ sở dữ liệu tại Cục thuế tỉnh Bình Dương, với bộ dữ liệu về giải quyết thủ tục hành chính thuế tại Cục thuế tỉnh Bình Dương. Trên cơ sở đó chúng tôi áp dụng các phương pháp khai phá dữ liệu để chuẩn hóa lại cơ sở dữ liệu, giúp chúng tôi có được một bộ dữ liệu có cấu trúc và có thể sử dụng được để đưa vào mô hình xử lý cho bài toán nâng cao chất lượng giải quyết thủ tục hành chính thuế tại Cục thuế tỉnh Bình Dương.

**1.3.1 Khái niệm**

“Khám phá tri thức là quá trình tìm ra những tri thức, đó là những mẫu tìm ẩn, trước đó chưa biết và là thông tin hữu ích đáng tin cậy”. Còn khai phá dữ liệu (KPDL) là một bước quan trọng trong quá trình khám phá tri thức, sử dụng các thuật toán KPDL chuyên dùng với một số qui định về hiệu quả tính toán chấp nhận được để chiết xuất ra các mẫu hoặc các mô hình có ích trong dữ liệu. Nói một cách khác, mục đích của khám phá tri thức và KPDL chính là tìm ra các mẫu hoặc mô hình đang tồn tại trong các cơ sở dữ liệu (CSDL) nhưng vẫn còn bị che khuất bởi hàng núi dữ liệu.

Khám phá tri thức từ CSDL là một quá trình sử dụng các phương pháp và công cụ tin học, trong đó con người là trung tâm của quá trình. Do đó, con người cần phải có kiến thức cơ bản về lĩnh vực cần khám phá để có thể chọn được tập con dữ liệu tốt, từ đó phát hiện các mẫu phù hợp với mục tiêu đề ra. Đó chính là tri thức, được rút ra từ CSDL, thường để phục vụ cho việc giải quyết một loạt nhiệm vụ nhất định trong một lĩnh vực nhất định. Tuy vậy, quá trình khám phá tri thức mang tính chất hướng nhiệm vụ vì không phải là mọi tri thức tìm được đều áp dụng vào thực tế được.

Để có được những thông tin quý báu chúng ta phải tìm ra các mẫu có trong tập CSDL trước. Việc đánh giá các mẫu được tìm thấy cũng là một điều thú vị và tất yếu có tính chất quyết định đến sự sử dụng hay không sử dụng chúng. Đầu ra của một chương trình là khám phá những mẫu có ích được gọi là tri thức. Tri thức được khám phá có các đặc điểm chính:

- Kiến thức cao cấp: Ngày càng có nhiều câu hỏi mang tính chất định tính cần phải trả lời dựa trên một khối lượng dữ liệu khổng lồ đã có. Quá trình để tìm ra kiến thức như vậy không phải từ những phương pháp thống kê cổ điển mà mà nó được được đúc kết từ các kinh nghiệm đã có, được thể hiện trong dữ liệu, những kết quả đó có thể lĩnh hội được.

- Độ chính xác: Dù cho những mẫu khai phá thật sự có trong CSDL hay không thì việc đo lường trị giá của chúng là bắt buộc phải có. Chúng ta sẽ chỉ sử dụng những mẫu nào có độ chính xác càng cao thì hiệu quả công việc đạt được càng lớn, những mẫu có độ chính xác chưa được xác định rõ ràng hoặc không cao thì không nên sử dụng chúng.

- Tính hấp dẫn: Khám phá tri thức được coi là lý thú vì nó có thể vạch ra các xu hướng một cách hoàn thiện. Đó là những điều mới lạ hay những quy trình tìm năng, hữu ích ẩn chứa từ trong dữ liệu trước đó.

- Tính hiệu quả: thời gian chạy của thuật toán khám phá tri thức trên CSDL lớn có thể dự tính và chấp nhận được.

Dữ liệu là tập hợp những bộ thông tin chính xác và quá trình khám phá tri thức được xem là sự lọc bỏ các dư thừa, được rút gọn tới mức tối thiểu chỉ để lại các đặc trưng cơ bản cho dữ liệu. Tri thức được tìm thấy là các thông tin tích hợp, bao gồm các sự kiện và các mối quan hệ trong chúng. Các mối quan hệ này có thể được hiểu ra, có thể được phát hiện, hoặc có thể được học.

Nếu khám phá tri thức là toàn bộ quá trình chiết xuất tri thức từ các CSDL thì KPDL là giai đoạn chủ yếu của quá trình đó. KPDL là một quá trình phát hiện các mẫu mới, thường bao gồm việc thử tìm mô hình phù hợp với tập dữ liệu và tìm kiếm các mẫu từ tập dữ liệu theo mô hình đó. Sử dụng các kỹ thuật và các khái niệm của các lĩnh vực đã được nghiên cứu từ trước như: học máy, nhận dạng, thống kê, hồi quy, xếp loại, phân nhóm, các mô hình đồ thị, các mạng Bayes,… Hầu hết các CSDL đều chứa rất nhiều các mẫu mới và có ích, tuy nhiên mẫu có giá trị với mục tiêu đặt ra phải là những mẫu không tầm thường. Để các mẫu trở nên không tầm thường, hệ thống phải làm nhiều hơn là chỉ mò mẫm thống kê vì kết quả của việc tính toán trực tiếp qua công tác thống kê là đã có đối với người dùng. Một hệ thống tìm kiếm cần phải có khả năng quyết định cần thực hiện tính toán nào và kết quả là có đáng quan tâm để tạo nên tri thức trong ngữ cảnh hiện tại hay không.

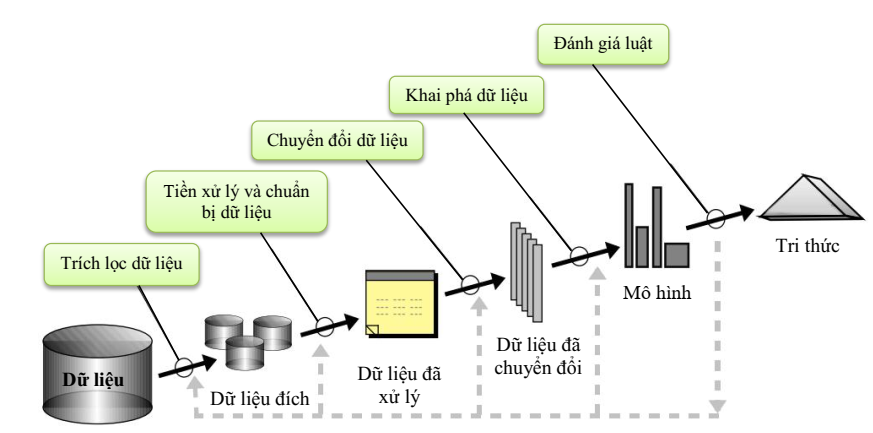
KPDL được sử dụng để tạo ra giả thuyết. Ví dụ như để xác định các yếu tố rủi ro khi cho vay tín dụng, kỹ thuật KPDL phải phát hiện được những người có thu nhập thấp và nợ nhiều là những người sẽ có mức rủi ro cao. Ngoài ra kỹ thuật cũng có thể phát hiện ra những quy luật mà nhà phân tích có thể chưa tìm ra ví dụ như tỷ lệ giữa thu nhập trên nợ và tuổi cũng là các yếu tố xác định mức rủi ro. Để làm được điều này, KPDL sử dụng các thông tin trong quá khứ để học. Nó sẽ tìm kiếm các thông tin này trong các CSDL và sử dụng chúng để tìm ra các mẫu đáng quan tâm.

Nếu xét về mặt ý tưởng và mục đích ứng dụng, KPDL là một nhu cầu tất yếu, một sự nhạy cảm đáp lại sự mong mỏi của giới kinh doanh thì về mặt kỹ thuật, đó thực sự là một khó khăn và là cả sự thách thức đối với những nhà khoa học. KPDL được xây dựng dựa trên việc sử dụng các giải thuật mới, được định hướng theo như cầu kinh doanh để có thể giải quyết tự động các bài toán kinh doanh bằng các kỹ thuật dễ dùng và có thể hiểu được. Các kỹ thuật đang được nghiên cứu và sử dụng hiện nay bao gồm cây quyết định (CART, CHAID, AID), mạng neuron, phương pháp láng giềng gần nhất, các luật suy diễn,…

KPDL không thuộc một ngành công nghiệp nào. Nó sử dụng các kỹ thuật thông minh để khai phá các tri thức tiềm ẩn trong dữ liệu. Có thể coi KPDL ngày nay đang ở trạng thái giống như việc quản trị dữ liệu vào những năm 60, khi mà các ứng dụng quản trị dữ liệu đều không tuân theo một nguyên tắc chung nào cho đến khi mô hình dữ liệu quan hệ ra đời cùng với sức mạnh của ngôn ngữ vấn đáp đã thúc đẩy việc phát triển các ứng dụng quản trị dữ liệu lên nhanh chóng. Tuy vậy, hiện nay trên thế giới đã có rất nhiều ngành công nghiệp sử dụng kỹ thuật KPDL để phục vụ cho hoạt động kinh doanh của mình và đã bước đầu thành công như ngành tài chính, y học, hóa học, bảo hiểm, sản xuất, giao thông, hàng không,… Các kết quả đạt được cho thấy mặc dù kỹ thuật KPDL hiện nay vẫn còn nhiều vấn đề nổi cộm, nhưng với những tri thức mà chuyên gia con người cũng chưa cung cấp được thì KPDL có một tiềm năng to lớn trong việc tạo ra những lợi nhuận đáng kể trong nền kinh tế.

**1.3.2 Quy trình phát hiện tri thức**

Quy trình phát hiện tri thức được mô tả tóm tắt như sau:

  
Hình 1.1: Quy trình khai phá tri thức

Quá trình khám phá tri thức từ CSDL là một quá trình có sử dụng nhiều phương pháp và công cụ tin học nhưng vẫn là một quá trình mà trong đó con người là trung tâm. Do đó, nó không phải là một hệ thống phân tích tự động mà là một hệ thống bao gồm nhiều hoạt động tương tác thường xuyên giữa con người và CSDL, tất nhiên là với sự hỗ trợ của các công cụ tin học. Người sử dụng hệ thống ở đây phải là người có kiến thức cơ bản về lĩnh vực cần phát hiện tri thức để có thể chọn được đúng các tập con dữ liệu, các lớp mẫu phù hợp và đạt tiêu chuẩn quan tâm so với mục đích. Tri thức mà ta nói ở đây là các tri thức rút ra từ các CSDL, thường để phục vụ cho việc giải quyết một loạt nhiệm vụ nhất định trong một lĩnh vực nhất định. Do đó, quá trình phát hiện tri thức cũng mang tính chất hướng nhiệm vụ, không phải là phát hiện mọi tri thức bất kỳ mà là phát hiện tri thức nhằm giải quyết tốt nhiệm vụ đề ra.

Gom dữ liệu (Gathering): Tập hợp dữ liệu là bước đầu tiên trong quá trình khai phá dữ liệu. Đây là bước được khai thác trong một CSDL, một kho dữ liệu và thậm chí các dữ liệu từ các nguồn ứng dụng Web.

Trích lọc dữ liệu (Selection): Ở giai đoạn này lựa chọn những dữ liệu phù hợp với nhiệm vụ phân tích trích rút từ CSDL.

Làm sạch, tiền xử lý và chuẩn bị trước dữ liệu (Cleansing, Pre-processing and Preparation): Giai đoạn thứ ba này là giai đoạn hay bị sao lãng, nhưng thực tế nó là một bước rất quan trọng trong quá trình khai phá dữ liệu. Một số lỗi thường mắc phải trong khi gom dữ liệu là tính không đủ chặt chẻ, logic. Vì vậy, dữ liệu thường chứa các giá trị vô nghĩa và không có khả năng kết nối dữ liệu, ví dụ: điểm = -1. Giai đoạn này sẽ tiến hành xử lý những dạng dữ liệu không chặt chẽ nói trên. Những dữ liệu dạng này được xem như thông tin dư thừa, không có giá trị. Bởi vậy, đây là một quá trình rất quan trọng vì dữ liệu này nếu không được “làm sạch” sẽ gây nên những kết quả sai lệch nghiêm trọng.

Chuyển đổi dữ liệu (Transformation): Tiếp theo là giai đoạn chuyển đổi dữ liệu, dữ liệu được chuyển đổi hay được hợp nhất về dạng thích hợp cho việc khai phá.

Khai phá dữ liệu (Data Mining): Đây là một tiến trình cốt yếu. Ở giai đoạn này nhiều thuật toán khác nhau đã được sử dụng một cách phù hợp để trích xuất thông tin có ích hoặc cá mẫu điển hình trong dữ liệu.

Đánh giá kết quả mẫu (Evaluation of Result): Đây là giai đoạn cuối trong quá trình khai phá dữ liệu. Ở giai đoạn này, các mẫu dữ liệu được chiết xuất, không phải bất cứ mẫu dữ liệu nào cũng đều hữu ích, đôi khi nó còn bị sai lệch. Vì vậy, cần phải ưu tiên những tiêu chuẩn đánh giá để chiết xuất ra các tri thức cần thiết.

Từ quá trình khám phá tri thức trên chúng ta thấy được sự khác biệt giữa khám phá tri thức và khai phá dữ liệu. Trong khi khám phá tri thức là nói đến quá trình tổng thể phát hiện tri thức hữu ích từ dữ liệu. Còn KPDL chỉ là một bước trong quá trình khám phá tri thức, các công việc chủ yếu là xác định được bài toán khai phá, tiến hành lựa chọn phương pháp KPDL phù hợp với dữ liệu có được và tách ra các tri thức cần thiết.

**1.3.3 Các phương pháp khai phá dữ liệu**

Các kỹ thuật KPDL được có thể chia làm 2 nhóm chính:

- Kỹ thuật KPDL mô tả: có nhiệm vụ mô tả về các tính chất hoặc các đặc tính chung của dữ liệu trong CSDL hiện có. Nhóm kỹ thuật này gồm các phương pháp: phân nhóm (Clustering), tổng hợp hóa (Summerization), phát hiện sự biến đổi và độ lệch (Change and deviation detection), phân tích luật kết hợp (Association Rules), ...

- Kỹ thuật KPDL dự đoán: có nhiệm vụ đưa ra các dự đoán dựa vào các suy diễn trên dữ liệu hiện thời. Nhóm kỹ thuật này gồm các phương pháp: phân lớp (Classification), hồi quy (Regression), ...

**Phân lớp và dự đoán (Classification & Prediction)**

Là đặt các mẫu vào các lớp được xác định trước. Nhiệm vụ chính là tìm các hàm ánh xạ các mẫu dữ liệu một cách chính xác vào trong các lớp.Ví dụ một ngân hàng muốn phân loại các khách hành của họ vào trong hai nhóm có nợ hay không nợ, từ đó giúp họ ra quyết định cho vay hay không cho vay. Quá trình phân lớp dữ liệu thường gồm 2 bước: xây dựng mô hình và sử dụng mô hình để phân lớp dữ liệu.

- Bước 1: một mô hình sẽ được xây dựng dựa trên việc phân tích các mẫu dữ liệu sẵn có. Mỗi mẫu tương ứng với một lớp, được quyết định bởi một thuộc tính gọi là thuộc tính lớp. Các mẫu dữ liệu này còn được gọi là tập dữ liệu huấn luyện (training data set). Các nhãn lớp của tập dữ liệu huấn luyện đều phải được xác định trước khi xây dựng mô hình, vì vậy phương pháp này còn được gọi là học có giám sát (supervised learning) khác với phân nhóm dữ liệu là học không có giám sát (unsupervised learning).

- Bước 2: sử dụng mô hình để phân lớp dữ liệu. Trước hết chúng ta phải tính độ chính xác của mô hình. Nếu độ chính xác là chấp nhận được, mô hình sẽ được sử dụng để dự đoán nhãn lớp cho các mẫu dữ liệu khác trong tương lai.

Trong kỹ thuật phân lớp chúng ta có thể sử dụng các phương pháp như: Cây quyết định (Decision Tree), K-Láng giềng gần nhất (k-Nearest Neighbor), Mạng Nơron (Neural networks), Giải thuật di truyền (Genetic algorithms), Mạng Bayesian (Bayesian networks), Tập mờ và tập thô (Rough and Fuzzy Sets).

**Luật kết hợp (Association Rules)**

Luật kết hợp là dạng luật biểu diễn tri thức ở dạng tương đối đơn giản. Mục tiêu của phương pháp này là phát hiện và đưa ra các mối liên hệ giữa các giá trị dữ liệu trong CSDL. Mẫu đầu ra của giải thuật KPDL là tập luật kết hợp tìm được.

Tuy luật kết hợp là một dạng luật khá đơn giản nhưng lại mang rất nhiều ý nghĩa. Thông tin mà dạng luật này đem lại rất có lợi trong các hệ hỗ trợ ra quyết định. Tìm kiếm được những luật kết hợp đặc trưng và mang nhiều thông tin từ CSDL tác nghiệp là một trong những hướng tiếp cận chính của lĩnh vực khai phá dữ liệu.

**Khai thác mẫu tuần tự (Sequential / Temporal patterns)**

Tương tự như khai thác luật kết hợp nhưng có thêm tính thứ tự và tính thời gian. Một luật mô tả mẫu tuần tự có dạng tiêu biểu X -> Y phản ánh sự xuất hiện của biến cố X sẽ dẫn đến việc xuất hiện kế tiếp biến cố Y. Hướng tiếp cận này có tính dự báo cao.

**Phân nhóm- đoạn (Clustering / Segmentation)**

Mục tiêu chính của việc phân nhóm dữ liệu là nhóm các đối tượng tương tự nhau trong tập dữ liệu vào các nhóm sao cho mức độ tương tự giữa các đối tượng trong cùng một nhóm là lớn nhất và mức độ tương tự giữa các đối tượng nằm trong các nhóm khác nhau là nhỏ nhất. Các nhóm có thể tách nhau hoặc phân cấp gối lên nhau và số lượng các nhóm là chưa biết trước. Một đối tượng có thể vừa thuộc nhóm này, nhưng cũng có thể vừa thuộc nhóm khác. Không giống như phân lớp dữ liệu, phân nhóm dữ liệu không đòi hỏi phải định nghĩa trước các mẫu dữ liệu huấn luyện. Vì thế, có thể coi phân nhóm dữ liệu là một cách học bằng quan sát (learning by observation), trong khi phân lớp dữ liệu là học bằng ví dụ (learning by example). Trong phương pháp này chúng ta sẽ không thể biết kết quả các nhóm thu được sẽ như thế nào khi bắt đầu quá trình. Vì vậy, thông thường cần có một chuyên gia về lĩnh vực đó để đánh giá các nhóm thu được. Phân nhóm còn được gọi là học không có giám sát (unsupervised learning). Phân nhóm dữ liệu được sử dụng nhiều trong các ứng dụng về phân đoạn thị trường, phân đoạn khách hàng, nhận dạng mẫu, phân loại trang Web, … Ngoài ra phân nhóm dữ liệu còn có thể được sử dụng như một bước tiền xử lý cho các thuật toán KPDL khác.

**Hồi quy (Regression)**

Hồi quy là việc học một hàm ánh xạ từ một mẫu dữ liệu thành một biến dự đoán có giá trị thực. Nhiệm vụ của hồi quy tương tự như phân lớp, điểm khác nhau chính là ở chỗ thuộc tính để dự báo là liên tục chứ không rời rạc. Việc dự báo các giá trị số thường được làm bởi các phương pháp thống kê cổ điển chẳng hạn như hồi quy tuyến tính. Tuy nhiên phương pháp mô hình hóa cũng có thể được sử dụng như cây quyết định.

**Tổng hợp hóa (Summarization)**

Là công việc liên quan đến các phương pháp tìm kiếm một mô tả tập con dữ liệu. Kỹ thuật mô tả khái niệm và tổng hợp hóa thường áp dụng trong việc phân tích dữ liệu có tính thăm dò và báo cáo tự động. Nhiệm vụ chính là sản sinh ra các mô tả đặc trưng cho một lớp. Mô tả loại này là một kiểu tổng hợp, tóm tắt các đặc tính chung của tất cả hay hầu hết các mục của một lớp. Các mô tả đặc trưng thể hiện theo luật có dạng sau: “Nếu một mục thuộc về lớp đã chỉ trong tiền đề thì mục đó có tất cả các thuộc tính đã nêu trong kết luận”.

**Mô hình hóa sự phụ thuộc (dependency modeling)**

Là việc tìm kiếm một mô hình mô tả sự phụ thuộc giữa các biến, thuộc tính theo hai mức. Mức cấu trúc của mô hình mô tả (thường dưới dạng đồ thị), trong đó, các biến phụ thuộc bộ phận vào các biến khác. Và mức định lượng mô hình mô tả mức độ phụ thuộc. Những phụ thuộc này thường được biểu thị dưới dạng theo luật “nếu - thì” - nếu tiền đề đúng thì kết luận đúng. Về nguyên tắc, cả tiền đề và kết luận đều có thể là sự kết hợp logic của các giá trị thuộc tính. Trên thực tế, tiền đề thường là nhóm các giá trị thuộc tính và kết luận chỉ là một thuộc tính. Hơn nữa, hệ thống có thể phát hiện các luật phân lớp trong đó tất cả các luật cần phải có cùng một thuộc tính do người dùng chỉ ra trong kết luận.

Quan hệ phụ thuộc cũng có thể biểu diễn dưới dạng mạng tin cậy Bayes. Đó là đồ thị có hướng, không chu trình. Các nút biểu diễn thuộc tính và trọng số của liên kết phụ thuộc giữa các nút đó.

**Phát hiện sự biến đổi và độ lệch (Change and deviation detection)**

Tập trung vào khám phá hầu hết sự thay đổi có nghĩa dưới dạng độ đo đã biết trước hoặc giá trị chuẩn, phát hiện độ lệch đáng kể giữa nội dung của tập con dữ liệu thực và nội dung mong đợi. Hai mô hình độ lệch hay dùng là lệch theo thời gian và lệch theo nhóm. Độ lệch theo thời gian là sự thay đổi có ý nghĩa của dữ liệu thời gian. Độ lệch theo nhóm là sự khác nhau của dữ liệu trong hai tập con dữ liệu, ở đây xét cả trường hợp tập con dữ liệu này thuộc tập con kia. Nghĩa xác định dữ liệu trong một nhóm con của đối tượng có khác đáng kể so với toàn bộ đối tượng không? Theo cách này, sai sót dữ liệu hay sai lệch so với giá trị thông thường sẽ được phát hiện.

**1.3.4 Các lĩnh vực liên quan đến phát hiện tri thức và khai phá dữ liệu**

KPDL có nhiều ứng dụng trong thực tế, một số ứng dụng điển hình như:

Bảo hiểm, tài chính và thị trường chứng khoán: phân tích tình hình tài chính và dự báo giá của các loại cổ phiếu trong thị trường chứng khoán. Danh mục vốn và giá, lãi suất, dữ liệu thẻ tín dụng, phát hiện gian lận,…

Điều trị y học và chăm sóc y tế: một số thông tin về chẩn đoán bệnh lưu trong các hệ thống quản lý bệnh viện. Phân tích mối liên hệ giữa triệu chứng bệnh, chẩn đoán và phương pháp điều trị (chế độ dinh dưỡng, thuốc,…).

Sản xuất và chế biến: qui trình, phương pháp chế biến và xử lý sự cố

Text mining & Web mining: phân lớp văn bản và các trang web, tóm tắt văn bản,...

Lĩnh vực khoa học: quan sát thiên văn, dữ liệu gene, dữ liệu sinh vật học, tìm kiếm, so sánh các hệ gene và thông tin di truyền, mối liên hệ gene và các bệnh di truyền,…

Lĩnh vực khác: viễn thông, môi trường, thể thao, âm nhạc, giáo dục,…

**Tổng kết chương 1**

Trong chương 1 của luận văn, chúng tôi đã trình bày một cách tổng quan về tỉnh Bình Dương, tổng quan về cục thuế tỉnh Bình Dương sử dụng khai phá dữ liệu vào xử lý các công việc liên quan. Trình bày tổng quan về khai phá dữ liệu và các bước khai phá dữ liệu dựa trên báo cáo tình hình, kết quả giải quyết thủ tục hành chính thuế tại Cục thuế Bình Dương.

# **Chương 2. CƠ SỞ LÝ THUYẾT VÀ CÁC NGHIÊN CỨU CÓ LIÊN QUAN**

Trong chương 2 này, chúng tôi trình bày một số phương pháp dự báo, phương pháp khai phá dữ liệu và các nghiên cứu liên quan đã được công bố về các thủ tục hành chính thuế như: GTGT, TNDN, TNCN…, kế hoạch phát triển thủ tục hành chính của Ngành thuế Việt Nam nói chung cũng như Cục thuế Bình Dương nói riêng. Chương 2 cũng trình bày một số nghiên cứu liên quan đến mô hình dự báo ngắn hạn, mô hình dự báo dài hạn trên một số nước.

## **2.1 Một số phương pháp dự báo**

Dự báo đã hình thành từ đầu những năm 60 của thế kỉ 20. Khoa học dự báo với tư cách một ngành khoa học độc lập có hệ thống lí luận, phương pháp luận và phương pháp hệ riêng nhằm nâng cao tính hiệu quả của dự báo. Người ta thường nhấn mạnh rằng một phương pháp tiếp cận hiệu quả đối với dự báo là phần quan trọng trong hoạch định.

Như vậy, dự báo là một khoa học và nghệ thuật tiên đoán những sự việc sẽ xảy ra trong tương lai, trên cơ sở phân tích khoa học về các dữ liệu đã thu thập được.

Khi tiến hành dự báo ta căn cứ vào việc thu thập xử lý số liệu trong quá khứ và hiện tại để xác định xu hướng vận động của các hiện tượng trong tương lai nhờ vào một số mô hình toán học.

Dự báo có thể là một dự đoán chủ quan hoặc trực giác về tương lai. Nhưng để cho dự báo được chính xác hơn, người ta cố loại trừ những tính chủ quan của người dự báo.

Ngày nay, dự báo là một nhu cầu không thể thiếu được của mọi hoạt động kinh tế - xác hội, khoa học - kỹ thuật, được tất cả các ngành khoa học quan tâm nghiên cứu.

Có nhiều cách để phân biệt dự báo: thời gian và nội dung…

Về thời gian có thể chia thành 3 loại chính, đó là dự báo ngắn hạn, dự báo trung hạn và dự báo dài hạn.

Dự báo ngắn hạn: Là dự báo có tầm xa dự báo rất ngắn, có thể tuần, tháng... đến dưới một năm. Dự báo loại này thường được dùng cho các quyết định mua sắm, điều độ công việc, phân giao nhiệm vụ, cân đối các mặt trong quản trị tác nghiệp.

Dự báo trung hạn: Khoảng thời gian dự báo thường từ tháng đến 3 năm. Loại dự báo này cần thiết cho việc lập kế hoạch sản xuất, kế hoạch bán hàng, dự thảo ngân sách, kế hoạch tiền mặt, huy động các nguồn lực....

Dự báo dài hạn: Là các dự báo cho khoảng thời gian từ 3 năm trở lên. Loại dự báo này cần cho việc lập các dự án sản xuất sản phẩm mới, lựa chọn các dây chuyền công nghệ, thiết bị mới, mở rộng doanh nghiệp....

Về nội dung có thể chia dự báo thành: Dự báo khoa học, dự báo kinh tế, dự báo xã hội, dự báo tự nhiên, thiên văn học…

- Dự báo khoa học: Là dự kiến, tiên đoán về những sự kiện, hiện tượng, trạng thái nào đó có thể hay nhất định sẽ xảy ra trong tương lai. Theo nghĩa hẹp hơn, đó là sự nghiên cứu khoa học về những triển vọng của một hiện tượng nào đó, chủ yếu là những đánh giá số lượng và chỉ ra khoảng thời gian mà trong đó hiện tượng có thể diễn ra những biến đổi.

- Dự báo kinh tế: Là khoa học dự báo các hiện tượng kinh tế trong tương lai. Dự báo kinh tế được coi là giai đoạn trước của công tác xây dựng chiến lược phát triển kinh tế - xã hội và dự án kế hoạch dài hạn; không đặt ra những nhiệm vụ cụ thể, nhưng chứa đựng những nội dung cần thiết làm căn cứ để xây dựng những nhiệm vụ đó. Dự báo kinh tế bao trùm sự phát triển kinh tế và xã hội của đất nước có tính đến sự phát triển của tình hình thế giới và các quan hệ quốc tế. Thường được thực hiện chủ yếu theo những hướng sau: dân số, nguồn lao động, việc sử dụng và tái sản xuất chúng, năng suất lao động; tái sản xuất xã hội trước hết là vốn sản xuất cố định: sự phát triển của cách mạng khoa học – kĩ thuật và công nghệ và khả năng ứng dụng vào kinh tế; mức sống của nhân dân, sự hình thành các nhu cầu phi sản xuất, động thái và cơ cấu tiêu dùng, thu nhập của nhân dân; động thái kinh tế quốc dân và sự chuyển dịch cơ cấu (nhịp độ, tỉ lệ, hiệu quả); sự phát triển các khu vực và ngành kinh tế (khối lượng động thái, cơ cấu, trình độ kĩ thuật , bộ máy, các mối liên hệ liên ngành); phân vùng sản xuất, khai thác tài nguyên thiên nhiên và phát triển các vùng kinh tế trong nước, các mối liên hệ liên vùng; dự báo sự phát triển kinh tế của thế giới kinh tế. Các kết quả dự báo kinh tế cho phép hiểu rõ đặc điểm của các điều kiện kinh tế - xã hội để đặt chiến lược phát triển kinh tế đúng đắn, xây dựng các chương trình, kế hoạch phát triển một cách chủ động, đạt hiệu quả cao và vững chắc.

- Dự báo xã hội: Dự báo xã hội là khoa học nghiên cứu những triển vọng cụ thể của một hiện tượng, một sự biến đổi, một quá trình xã hội, để đưa ra dự báo hay dự đoán về tình hình diễn biến, phát triển của một xã hội.

- Dự báo tự nhiên, thiên văn học, loại dự báo này thường bao gồm:

+ Dự báo thời tiết: Thông báo thời tiết dự kiến trong một thời gian nhất định trên một vùng nhất định. Trong dự báo thời tiết có dự báo chung, dự báo khu vực, dự báo địa phương, v.v. Về thời gian, có dự báo thời tiết ngắn (1-3 ngày) và dự báo thời tiết dài (tới một năm).

+ Dự báo thuỷ văn: Là loại dự báo nhằm tính để xác định trước sự phát triển các quá trình, hiện tượng thuỷ văn xảy ra ở các sông hồ, dựa trên các tài liệu liên quan tới khí tượng thuỷ văn. Dự báo thuỷ văn dựa trên sự hiểu biết những quy luật phát triển của các quá trình, khí tượng thuỷ văn, dự báo sự xuất hiện của hiện tượng hay yếu tố cần quan tâm…

## **2.2 Các nghiên cứu liên quan**

Các nghiên cứu liên quan đến thực báo các tình hình giải quyết hồ sơ thủ tục hành chính nói chung, đến việc giải quyết hồ sơ thủ tục thuế nói riêng ở Việt Nam đã được công bố ở nhiều nghiên cứu trong và ngoài nước.

Bài phân tích Khắc phục tình trạng trể hẹn hồ sơ trong giải quyết thủ tục hành chính, bài phân tích này nêu rỏ tình hình thực trạng của việc giải quyết hồ sơ. Nhưng chỉ là nhận định trong một thời gian ngắn và dựa vào kết quả tồn đọng hồ sơ. Chưa nêu ra được lý do cụ thể vì sao hồ sơ trể hẹn và cách khắc phục hoặc dự đoán khi nào hồ sơ đó sẽ trể hạn để đưa ra biện pháp.

Đề tài nghiên cứu “Dự báo nhu cầu giao thông và đánh giá khả năng thông hành qua nút giao tại Quận 10 đến 2020”, đề tài nghiên cứu dự báo trên dựa vào một phần mềm dựng sẳn để tính toán dựa trên việc phát sinh hành trình, phân phối hành trình, nên chưa ra được tính khách quan nhất và chưa có sự so sánh giữa các phần mềm xem phần mềm nào là tối ưu nhất.

Đề tài “Sử dụng Data Mining dự báo nhu cầu lao động cho một số ngành nghề trên địa bàn tỉnh Bình Dương” 2019 [7] trình bày cơ bản về các phương pháp dự báo, khai phá dữ liệu, các kỹ thuật khai phá dữ liệu và các ứng dụng của khai phá dữ liệu trong thực tế. Trong đó tập trung vào nghiên cứu các kỹ thuật, phương pháp hồi quy trên các thuật toán và áp dụng vào bài toán dự báo thực tế đó là bài toán dự báo nhu cầu lao động cho một số ngành nghề trên địa bàn tỉnh Bình Dương.

Nhìn chung chưa có nghiên cứu nào về khai phá dữ liệu dựa trên báo cáo tình hình, kết quả giải quyết thủ tục hành chính thuế tỉnh Bình Dương bằng học máy; Tuy nhiên những nghiên cứu trên là nền tảng để chúng tôi nghiên cứu và vận dụng một cách sáng tạo vào việc khai thác dữ liệu dựa vào báo cáo tình hình, kết quả giải quyết thủ tục hành chính thuế nói chung và thủ tục hành chính thuế tỉnh Bình Dương nói riêng.

# **Chương 3. MÔ HÌNH ĐỀ XUẤT KHAI PHÁ DỮ LIỆU DỰA TRÊN BÁO CÁO TÌNH HÌNH, KẾT QUẢ GIẢI QUYẾT THỦ TỤC HÀNH CHÍNH THUẾ TẠI CỤC THUẾ TỈNH BÌNH DƯƠNG**

## **3.1 Mô hình đề xuất**

Mô hình đề xuất về khai phá dữ liệu dựa trên báo cáo tình hình, kết quả giải quyết thủ tục hành chính thuế tại Cục thuế Bình Dương dự báo số lượng hồ sơ được giải quyết được biểu diễn như sau:

5. Trình bày kết quả dự báo

1. Dữ liệu về thủ tục hành chính tại cụng thuế tỉnh Bình Dương

2. Tiền xử lý dữ liệu

3. Lựa chọn mô hình

4. Thực nghiệm, Đánh giá

Mô hình tối ưu

Mô hình chưa tối ưu

6. Theo dõi KQDB

Hình 3.1: Mô hình dự báo kết quả giải quyết thủ tục hành chính cục Thuế

tỉnh Bình Dương

*Bước 1:*

- Thu thập dữ liệu về số lượng hồ sơ và nguồn thu ngân sách các dich vụ thủ tục hành chính thuế tại các đơn vị trực thuộc Cục thuế tỉnh Bình Dương.

- Thu thập dữ liệu các yếu tố ảnh hưởng tới dịch vụ thủ tục hành chính thuế

Bước 2:

- Tiền xử lý dữ liệu và cho ra tập dữ liệu chuẩn để dưa vào dự báo, bao gồm: việc Làm sạch dữ liệu, loại bỏ nhiễu, hiểu chỉnh những thành phần dữ liệu không nhất quán, xử lý các dữ liệu của các biến dư thừa hoặc thiếu thông tin, vector hóa dữ liệu đã được làm sạch.

*Bước 3:*

Lựa chọn mô hình: Gồm bốn mô hình dự báo Linear Regression, K-Nearest Neighbors, Decision Trees và Random Forests với các tham số trên từng mô hình.

*Bước 4:* Thực nghiệm, đánh giá và lựa chọn mô hình phù hợp

- Thực nghiệm và đánh giá, lựa chọn mô hình phù hợp dựa vào các độ đo:

Trong quá trình thực nghiệm, chúng tôi lập lại nhiều lần để thay đổi tỉ lệ phân chia giữa tập huấn luyện và tập kiểm tra với các tập dữ liệu được chương trình chọn ngẫu nhiên và tiến hành kiểm tra chọn độ chính xác của các model tốt nhất

- Thực nghiệm và đánh giá, lựa chọn mô hình phù hợp dựa vào kết quả số liệu dự báo:

Chúng tôi tiến hành cài đặt 04 mô hình dự báo của 04 thuật toán trên với dữ liệu đầu vào, đầu ra như sau:

+ Dữ liệu đầu vào là: Dữ liệu thu thập số lượng các dịch vụ thủ tục hành chính thuế tại Cục thuế tỉnh Bình Dương, dữ liệu về các yếu tố ảnh hưởng tới dịch vụ thủ tục hành chính thuế theo từng tháng.

+ Dữ liệu đầu ra: Số liệu kết quả dự báo số lượng hồ sơ và nguồn thu ngân sách

Để kiểm tra độ chính xác của kết quả dự báo với số liệu gốc hiện có, chúng tôi chọn số dữ liệu làm bộ test đồng thời tiến hành loại bỏ số liệu đó trong bảng dữ liệu và tiến hành chạy thực nghiệm

*Bước 5*: Trình bày kết quả dự báo

Kết quả dự báo phải được trình bày rõ ràng sao cho đễ hiểu các con số được tính toán như thế nào và chỉ ra sự tin cậy trong kết quả dự báo.

Người dự báo phải có khả năng trao đổi các kết quả dự báo theo ngôn ngữ mà các nhà quản lý hiểu được.

Trình bày cả ở dạng viết và dạng nói.

Bảng biểu phải ngắn gọn, rõ ràng.

Chuỗi dữ liệu dài có thể được trình bày dưới dạng đồ thị (cả giá trị thực và dự báo).

Trình bày thuyết trình nên theo cùng hình thức và cùng mức độ với phần trình bày viết.

Bước 6: Theo dõi kết quả dự báo

Lệch giữa giá trị dự báo và giá trị thực phải được thảo luận một cách tích cực, khách quan và cởi mở.

Mục tiêu của việc thảo luận là để hiểu tại sao có các sai số, để xác định độ lớn của sai số.

Trao đổi và hợp tác giữa người sử dụng và người làm dự báo có vai trò rất quan trọng trong việc xây dựng và duy trì quy trình dự báo thành công.

## **3.2 Các mô hình sử dụng trong đề tài**

Có rất nhiều thuật toán để sử dụng cho bài toán dự báo, tuy nhiên trong đề tài này, chúng tôi chọn bốn thuật toán để áp dụng xây dựng mô hình dự báo như sau:

### **3.2.1 Linear Regression**

Thuật ngữ hồi quy được sử dụng đầu tiên năm 1908, bởi Pearson. Mục đích của hồi quy là:

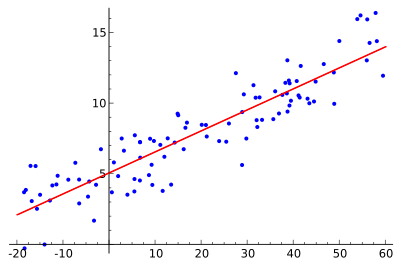
- Vấn đề giao dịch với các ứớc tính của một giá trị sản xuất dựa trên giá trị đầu vào.

- Hồi quy là một kỹ thuật khai thác dữ liệu được sử dụng để phù hợp với một phương trình của tập dữ liệu.

Ngoài ra, mục đích của hồi quy là tìm hiểu thêm về mối quan hệ giữ các biến độc lập (independent) hoặc biến dự đoán (predictor) và một biến phụ thuộc (dependent) hay tiêu chuẩn (criterion). Mô hình hồi quy dựa trên việc xây dựng các đồ thị dựa trên đường thẳng để giải quyết các bài toán có mức độ khó khác nhau. Chính vì vậy, hồi quy còn được biết đến là tất cả những thuật toán liên quan đến dữ liệu số. Hình thức đơn giản nhất của hồi quy là hồi quy tuyến tính, trong đó sử dụng phương trình đại số:

Chúng ta sẽ dự đoán y với biến X cho trước và mục tiêu của thuật toán hồi quy tuyến tính là tìm các giá trị cho các hệ số , ….

Có nhiều loại hồi quy khác nhau được sử dụng trong lĩnh vực thống kê và thường được sử dụng trong lĩnh vực dự đoán, nhưng ý tưởng cơ bản của hồi quy là mô hình được tạo ra mà bản đồ giá trị từ dự đoán có giá trị xảy ra lỗi là thấp nhất trong việc đưa ra một dự đoán.



*Hình 3.2. Mô hình hồi quy tuyến tính*

**Ưu điểm:**

- Trong trường hợp hồi quy tuyến tính, nó xây dựng một mô hình trong đó có mối quan hệ giữa các biến độc lập và phụ thuộc được lên đến nhiệm vụ của nó và cho kết quả tối ưu. Còn đối với hồi quy logic, xây dựng một mô hình dựa trên xác suất mà sự kiện xảy ra trong quẩn thể.

- Cả hồi quy tuyến tính và hồi quy logic đều dựa trên dữ liệu có sẵn để

xây dựng.

- Là một công cụ mạnh trong việc khai thác dữ liệu phân lớp.

- Hồi quy được giới hạn trong việc dự đoán các giá trị số.

**Khuyết điểm:**

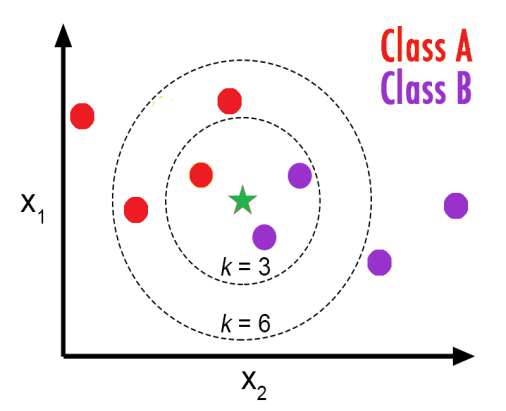
- Hồi quy không được ứng dụng trong việc giải quyết các vấn đề khai thác dữ liệu với mục đích phân tích kết hợp.

- Trong việc xử lý với số lượng dữ liệu lớn, việc lựa chọn hồi quy cho việc khai thác dữ liệu sẽ gặp rất nhiều lỗi và nhiễu trong quá trình khai thác.

### **3.2.2 K - Nearest Neighbors**

Thuật toán K- Nearest neighbour (KNN) gần nhất là thuật toán dùng để phân loại dữ liệu. KNN là một thuật toán học có giám sát mà kết quả của truy vấn hiện mới được phân loại dựa trên đa số các loại KNN. Mục đích KNN được sử dụng là để phân loại các đối tượng mới dựa trên các thuộc tính và các mẫu đào tạo. Việc phân loại không sử dụng bất kỳ mô hình để điều chỉnh, mà chỉ dựa vào bộ nhớ. Với một điểm truy vấn, chúng ta tìm k đối tượng hoặc (điểm đào tạo) gần nhất để thực hiện truy vấn. Việc phân loại được sử dụng đa số trong việc phân loại K đối tượng. Bất kỳ mối quan hệ có thể được chia một cách ngẫu nhiên. KNN sử dụng trong các thuật toán phân loại lân cận là giá trị dự đoán của các ví dụ truy vấn mới.

Thuật toán KNN rất đơn giản. Nó hoạt động dựa trên khoảng cách tối thiểu từ các mẫu truy vấn đến các mẫu đào tạo để xác định K- lân cận gần nhất. Sau khi chúng ta đã tập hợp được được những lân cận gần nhất, chúng ta sẽ sử dụng để dự đoán các mẫu truy vấn.



*Hình 3.3. KNN*

Các bước thực hiện về cách tính toán K lân cận gần nhất, áp dụng thuật toán KNN cho các dữ liệu định lượng:

* Xác định tham số K lân cận gần nhất.
* Tính khoảng cách giữa tất cả các mẫu truy vấn và đào tạo
* Phân loại theo khoảng cách và xác định những lân cận gần nhất dựa vào khoảng cách tối thiểu thứ k.
* Tập hợp các giá trị của Y của những người lân cận gần nhất.

- Sử dụng trung bình của các lân cận gần nhất là giá trị dự đoán của các mẫu truy vấn.

**Ưu điểm:**

* Kết quả dự đoán của dữ liệu mới nhanh chóng. Độ phức tạp của thuật toán bằng 0.
* Dùng để loại bỏ nhiễu trong dữ liệu huấn luyện.

**Khuyết điểm:**

* Rất nhạy cảm với nhiễu.
* Dữ liệu đào tạo nhỏ sẽ không hiệu quả.

### **3.2.3 Decision Trees**

Cây quyết định được sử dụng trong lĩnh vực khai phá dữ liệu và học máy. Cây quyết định thường được sử dụng như là một mô hình dự báo về một đối tượng mục tiêu, để có được kết luận về giá trị của mục tiêu đó. Cây quyết định còn được gọi là cây phân loại hay cây hồi quy.

Cấu trúc của một cây quyết định: trên cây quyết định có 3 loại nút

- Nút gốc: Không có cạnh vào, không có hoặc có nhiều cạnh ra.

- Nút giữa: Có chính xác một cạnh vào, có hai hay nhiều cạnh ra.

- Nút lá: có chính xác một cạnh vào, không có cạnh ra. Nút lá còn là đại diện cho phân loại, ngành đại diện hoặc liên từ của tính năng, từ đó dẫn đến những phân loại.

Trong phân tích quyết định, một cây quyết định có thể được sử dụng để đại diện rõ ràng và trực quan quyết định và ra quyết định. Trong khai phá dữ liệu, cây quyết định mô tả một dữ liệu nhưng không quyết định, các kết quả của cây phân loại dữ liệu có thể là đầu vào cho việc hỗ trợ ra quyết định.

• Việc tạo cây quyết định bao gồm 2 giai đoạn : Tạo cây và tỉa cây .

- Để tạo cây ở thời điểm bắt đầu tất cả những ví dụ huấn luyện là ở gốc sau đó phân chia ví dụ huấn luyện theo cách đệ qui dựa trên thuộc tính được chọn .

- Việc tỉa cây là xác định và xóa những nhánh mà có phần tử hỗn loạn hoặc những phần tử nằm ngoài (những phần tử không thể phân vào một lớp nào đó) .

• Có rất nhiều biến đổi khác nhau về thuật toán xây dựng cây quyết định, mặc dù vậy chúng vẫn tuân theo những bước cơ bản sau :

- Cây được thiết lập từ trên xuống dưới và theo cách thức chia để trị.

- Ở thời điểm bắt đầu, các mẫu huấn luyện nằm ở gốc của cây

- Thuộc tính được phân loại (Rời rạc hóa các thuộc tính dạng phi số)

- Chọn một thuộc tính để phân chia thành các nhánh. Thuộc tính được chọn dựa trên độ đo thống kê hoặc độ đo heuristic.

- Tiếp tục lặp lại việc xây dựng cây quyết định cho các nhánh.

• Điều kiện để dừng việc phân chia:

- Tất cả các mẫu rơi vào một nút thuộc về cùng một lớp (nút lá)

- Không còn thuộc tính nào có thể dùng để phân chia mẫu nữa

- Không còn lại mẫu nào tại nút.

Hiện nay, có nhiều cách để đánh giá cách chia là tốt hay không tốt. Các độ đo dùng để đánh giá và lựa chọn cách chia được định nghĩa trên gốc độ sự phân phối về lớp của các mẫu tin trước và sau khi bị chia. Gọi pi=p(i|t) là tỉ lệ các mẫu tin thuộc vào lớp I của nút t. Trong cách chia đôi, giả sử có hai lớp class=0 và class=1 thì p1=1-p0 (Với p0, p1 là xác suất của class=0 và class =1) . độ đo được phát triển cho việc lựa chọn cách chia tốt nhất dựa trên mức độ không thuần nhất (impurity) của các nút con. độ không thuần nhất càng nhỏ thì phân phối lớp càng lệch. độ không thuần nhất có thể được đo bằng entropy, gini, classification error. Entropy, gini, classification error tại nút t được định nghĩa như sau:

*c*1

*Entropy*   *p*(*i* | *t*) log 2 *p*(*i* | *t*)

*i*0

*c*1

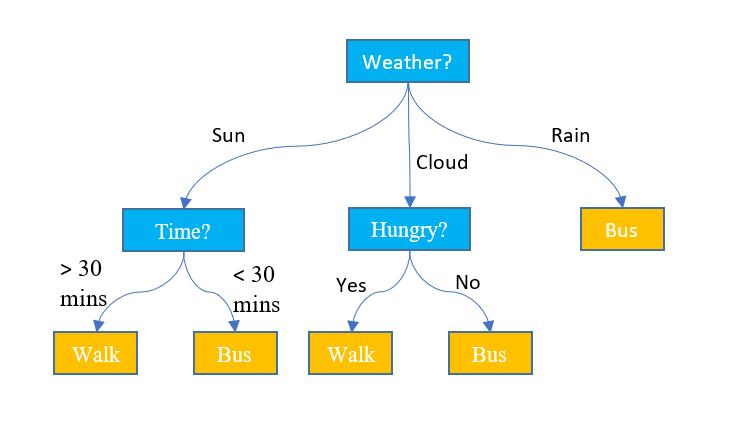


*Gini*(*t*)  1 [ *p*(*i* |*t*)]2

*i*0

*Classification* \_ *error*(*t*)  1  max[ *p*(*i* | *t*)]

Trong đó, c là tổng số lớp, các lớp được đánh số từ 0 đến c-1.



*Hình 3.4. Mô hình Cây quyết định*

**Ưu điểm:**

- Cây quyết định dễ hiểu. Người ta có thể hiểu mô hình cây quyết định sau khi được giải thích ngắn.

- Việc chuẩn bị dữ liệu cho một cây quyết định là cơ bản hoặc không cần thiết. Các kỹ thuật khác thường đòi hỏi chuẩn hóa dữ liệu, cần tạo các biến phụ (dummy variable) và loại bỏ các giá trị rỗng.

- Cây quyết định có thể xử lý cả dữ liệu có giá trị bằng số và dữ liệu có giá trị là tên thể loại. Các kỹ thuật khác thường chuyên để phân tích các bộ dữ liệu chỉ gồm một loại biến. Chẳng hạn, các luật quan hệ chỉ có thể dùng cho các biến tên, trong khi mạng nơ-ron chỉ có thể dùng cho các biến có giá trị bằng số.

- Cây quyết định là một mô hình hộp trắng. Mạng nơ-ron là một ví dụ về mô hình hộp đen, do lời giải thích cho kết quả quá phức tạp để có thể hiểu được.

- Có thể thẩm định một mô hình bằng các kiểm tra thống kê. điều này làm cho ta có thể tin tưởng vào mô hình.

- Cây quyết định có thể xử lý tốt một lượng dữ liệu lớn trong thời gian ngắn. Có thể dùng máy tính cá nhân để phân tích các lượng dữ liệu lớn trong một thời gian đủ ngắn để cho phép các nhà chiến lược đưa ra quyết định dựa trên phân tích của cây quyết định.

**Nhược điểm:**

- Mô hình cây quyết định phụ thuộc rất lớn vào dữ liệu của chúng ta. Thậm chí, với một sự thay đổi nhỏ trong bộ dữ liệu, cấu trúc mô hình cây quyết định có thể thay đổi hoàn toàn.

- Cây quyết định hay gặp vấn đề quá khớp (overfitting)

### **3.2.4 Random Forests**

Random Forest (rừng ngẫu nhiên) là phương pháp học tập thể (ensemble) để phân loại, hồi quy được phát triển bởi Leo Breiman tại đại học California, Berkeley. Breiman cũng đồng thời là đồng tác giả của phương pháp CART. Random Forest (RF) là phương pháp cải tiến của phương pháp tổng hợp bootstrap (bagging). RF sử dụng 2 bước ngẫu nhiên, một là ngẫu nhiên theo mẫu (sample) dùng phương pháp bootstrap có hoàn lại (with replacement), hai là lấy ngẫu nhiên một lượng thuộc tính từ tập thuộc tính ban đầu. Các tập dữ liệu con (sub-dataset) được tạo ra từ 2 lần ngẫu nhiên này có tính đa dạng cao, ít liên quan đến nhau, giúp giảm lỗi phương sai (variance). Các cây CART được xây dựng từ tập các tập dữ liệu con này tạo thành rừng. Khi tổng hợp kết quả, RF dùng phương pháp bỏ phiếu (voting) cho bài toán phân loại và lấy giá trị trung bình (average) cho bài toán hồi quy. Việc kết hợp các mô hình CART này để cho kết quả cuối cùng nên RF được gọi là phương pháp học tập thể.

Đối với bài toán phân loại, cây CART sử dụng công thức Gini như là một hàm điều kiện để tính toán điểm tách nút của cây. Số lượng cây là không hạn chế, các cây trong RF được xây dựng với chiều cao tối đa.

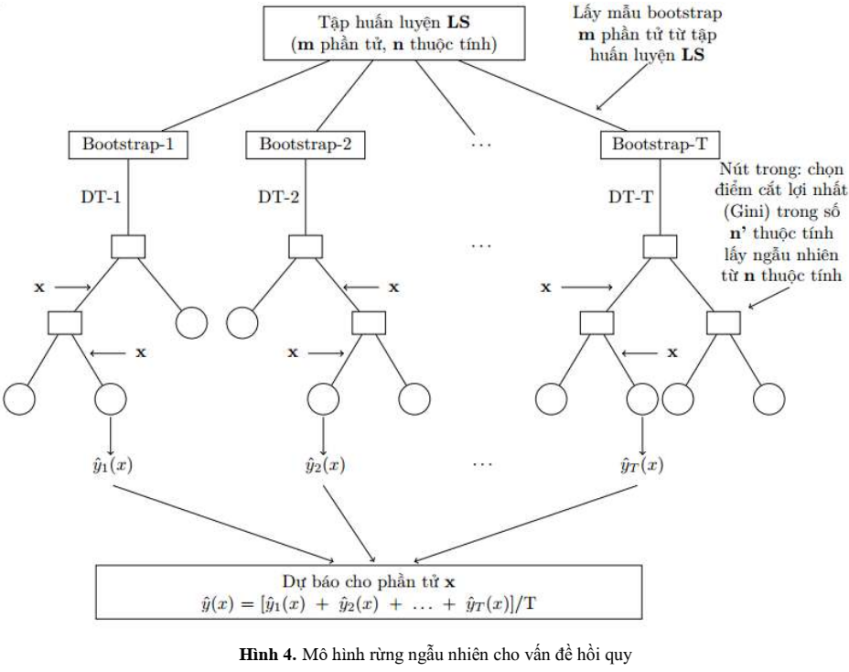
Trong những năm gần đây, RF được sử dụng khá phổ biến bởi những điểm vượt trội của nó so với các thuật toán khác: xử lý được với dữ liệu có số lượng các thuộc tính lớn, có khả năng ước lượng được độ quan trọng của các thuộc tính, thường có độ chính xác cao trong phân loại (hoặc hồi quy), quá trình học nhanh. Trong RF, mỗi cây chỉ chọn một tập nhỏ các thuộc tính trong quá trình xây dựng (bước ngẫu nhiên thứ 2), cơ chế này làm cho RF thực thi với tập dữ liệu có số lượng thuộc tính lớn trong thời gian chấp nhận được khi tính toán. Người dùng có thể đặt mặc định số lượng các thuộc tính để xây dựng cây trong rừng, thông thường giá trị mặc định tối ưu là √𝑝 cho bài toán phân loại và 𝑝⁄3 với các bài toán hồi quy (p là số lượng tất cả các thuộc tính của tập dữ liệu ban đầu). Số lượng các cây trong rừng cần được đặt đủ lớn để đảm bảo tất cả các thuộc tính đều được sử dụng một số lần. Thông thường là 500 cây cho bài toán phân loại, 1000 cây cho bài toán hồi quy. Do sử dụng phương pháp bootstrap lấy mẫu ngẫu nhiên có hoàn lại nên các tập dữ liệu con có khoảng 2/3 các mẫu không trùng nhau dùng để xây dựng cây, các mẫu ngày được gọi là in-bag. Khoảng 1/3 số mẫu còn lại gọi là out-of-bag, do không tham gia vào việc xây dựng cây nên RF dùng luôn các mẫu out-of-bag này để kiểm thử và tính toán độ quan trọng thuộc tính của các cây CART trong rừng.

Tóm tắt thuật toán Random Forest cho phân loại dữ liệu:

Bước 1: Từ tập dữ liệu huấn luyện D, ta tạo dữ liệu ngẫu nhiên (mẫu bootstrap).

Bước 2: Sử dụng các tập con dữ liệu lấy mẫu ngẫu nhiên D1 , D2 ,…, Dk xây dựng nên các cây T1 , T2 ,…, Tk .

Bước 3: Kết hợp các cây: sử dụng chiến lược bình chọn theo số đông với bài toán phân loại hoặc lấy trung bình các giá trị dự đoán từ các cây với bài toán hồi quy.



*Hình 3.5. Mô hình Random forest cho vấn đề hồi quy [6], [14]*

Quá trình học của Random Forest bao gồm việc sử dụng ngẫu nhiên giá trị đầu vào, hoặc kết hợp các giá trị đó tại mỗi node trong quá trình dựng từng cây quyết định. Trong đó Random Forest có một số thuộc tính mạnh như :

(1) Độ chính xác của RF tương đối cao.

(2) Thuật toán giải quyết tốt các bài toán có nhiều dữ liệu nhiễu.

(3) Thuật toán chạy nhanh hơn so với bagging.

(4) Có những sự ước lượng nội tại như độ chính xác của mô hình dự đoán hoặc độ mạnh và liên quan giữa các thuộc tính.

(5) Dễ dàng thực hiện song song.

(6) Tuy nhiên để đạt được các tính chất mạnh trên, thời gian thực thi của thuật toán khá lâu và phải sử dụng nhiều tài nguyên của hệ thống.

Tính chất thứ 4 được quan tâm rất nhiều và là tính chất được sử dụng để giải quyết bài toán trích chọn thuộc tính. Sau khi thực hiện học sẽ thu được một danh sách các thuộc được xếp hạng dựa theo một trong hai tiêu chí. Tiêu chí thứ nhất là thu được sau quá trình kiểm tra độ chính xác sử dụng các mẫu out-of-bag. Tiêu chí thứ hai là mức độ dầy đặc tại các node khi phân chia thuộc thuộc tính, và được tính trung bình trên tất cả các cây.

Qua những tìm hiểu trên về giải thuật RF ta có nhận xét rằng RF là một phương pháp phân loại tốt do:

(1) Trong RF các phương sai (variance) được giảm thiểu do kết quả của RF được tổng hợp thông qua nhiều bộ học (learner).

(2) Việc chọn ngẫu nhiên tại mỗi bước trong RF sẽ làm giảm mối tương quan (correlation) giữa các bộ phận lớp trong việc tổng hợp các kết quả.

Ngoài ra, chúng ta cũng thấy rằng lỗi chung của một rừng các cây phân loại phụ thuộc vào lỗi riêng của từng cây trong rừng cũng như mỗi tương quan giữa các cây.

# **Chương 4. THỰC NGHIỆM VÀ ĐÁNH GIÁ KẾT QUẢ**

Trong chương này chúng tôi tiến hành cài đặt chương trình thực nghiệm các mô hình dự báo, từ kết quả đạt được của các mô hình dự báo, chúng tôi tiến hành đánh giá, so sánh để tìm ra mô hình phù hợp và đề xuất mô hình xử lý cho bài toán khai phá dữ liệu dựa trên báo cáo tình hình, kết quả giải quyết, thủ tục hành chính thuế tại Cục thuế Bình Dương.

## **4.1 Dữ liệu**

Chúng tôi thu thập dữ liệu về các báo cáo tình hình, giải quyết thủ tục hành chính thuế tỉnh Bình Dương được điều tra trong 3 năm (từ năm 2017 đến năm 2019) và 10 tháng năm 2020. Các phòng thực hiện khảo sát: Phòng kê khai kế toán thuế, phòng quản lý nợ, phòng thanh tra kiểm tra, phòng hộ kinh doanh cá nhân, phòng một cửa. Phạm vi các doanh nghiệp hoạt động sản xuất kinh doanh trên địa bàn tỉnh Bình Dương, gồm có: Thuế GTGT, thuế TNCN, thuế TNDN.

Cơ sở dữ liệu đã được chuyển sang file Excel gồm nhiều trường dữ liệu đã được trích lọc về tình hình, giải quyết thủ tục hành chính thuế tỉnh Bình Dương như: Mã hồ sơ, Loại hồ sơ, ngày tháng năm nhận hồ sơ, ngày tháng năm hẹn trả hồ sơ, thời hạn trả hồ sơ, thời gian quá hạn hồ sơ, phòng ban xử lý hồ sơ, …

|  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Mã HS | Loại HS | Ngày nhận | Tháng nhận | Năm nhận | Ngày hẹn | Tháng hẹn | Năm hẹn | Hẹn | Trể hẹn | Phòng ban | … |
| 171229/  ZV01/  0006/  7401/  039336733 | 006 | 29 | 12 | 2017 | 03 | 01 | 2018 | 6 | 1051 | Phòng kê khai và kế toán thuế | … |
| 171229/  ZV01/  0006/  7401/  039336723 | 006 | 29 | 12 | 2017 | 03 | 01 | 2018 | 6 | 7 | Phòng kê khai và kế toán thuế | … |
| 180108/  ZV01/  0006/  7401/  000177836 | 006 | 08 | 01 | 2018 | 10 | 01 | 2018 | 2 | 15 | Phòng kê khai và kế toán thuế |  |
| 180109/  ZV01/  0006/  7401/  000226798 | 006 | 09 | 01 | 2018 | 11 | 01 | 2018 | 2 | 7 | Phòng kê khai và kế toán thuế |  |
| … | … | … | … | … | … | … | … | … | … | … | … |
| 201001/  ZV15/  0041/  7401/  017983123 | 041 | 01 | 10 | 2020 | 07 | 10 | 2020 | 7 | 1 | Phòng Quản Lý Nợ & Cưỡng Chế Nợ Thuế | … |

Bảng 4.1: Hồ sơ và phòng ban xử lý hồ sơ bị quá hạn tại Cục thuế.

Mã hồ sơ bao gồm thông tin ngày, tháng, năm nộp hồ sơ, hồ sơ gồm những thành phần nào, hồ sơ thuộc loại nộp lần đầu hay nộp lại nhiều lần và phần cuối là một mã định danh duy nhất cho hồ sơ đó.

Loại hồ sơ bao gồm những loại như: Trả lời vướng mắc, hồ sơ đăng ký thuế, hồ sơ hoàn thuế, hồ sơ miễn giảm thuế, hồ sơ khiếu nại tố cáo về thuế, hồ sơ mua hóa đơn thuế, hồ sơ đăng ký tự in hóa đơn sử dụng, hồ sơ giải quyết nợ thuế, hồ sơ khác.

Ngày tháng năm nhận và ngày Ngày tháng năm hẹn trả của một hồ sơ đã được quy định thời gian trả cụ thể trong bộ thủ tục hành chính của cơ quan, bộ thủ tục này được quy định trong các quy định của văn bản pháp luật. Nên việc nhận hồ sơ và trả hồ sơ phải nằm trong thời hạn cho phép. Từ đây một vấn đề được đặt ra là vì lý do gì có một số hồ sơ bị trể hẹn.

Phòng ban là nơi xử lý chính hồ sơ do người dân hoặc doanh nghiệp nộp vào, mỗi phòng ban sẽ xử lý một nhiệm vụ riêng, các phòng ban bao gồm: Phòng kê khai và kế toán thuế, phòng kiểm tra nội bộ, phòng kiểm tra thuế, phòng quản lý các khoản thu về đất, phòng thanh tra, phòng quản lý nợ và cưỡng chế, …

Hình thức nộp và hình thức nhận hồ sơ bao gồm hình thức nộp trực tiếp, nộp điện tử hoặc bưu điện. Tùy thuộc vào mỗi loại hồ sơ khác nhau mà có hình thức nộp và hình thức nhận khác nhau để linh hoạt trong quá trình nhận và xử lý.

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| A Tên chỉ tiêu | Tồn đầu kỳ | Mới phát sinh | Bổ sung | Tổng số | Đúng hạn | Quá hạn | Tổng số | Tỷ lệ đúng hạn | Tồn cuối kỳ |
| 1. Trả lời vướng mắc của ĐTNT bằng văn bản | 68 | 218 | 6 | 218 | 216 | 14 | 230 | 93.91 | 56 |
| 2. Hồ sơ đăng ký thuế | 280 | 84 | 0 | 84 | 12 | 287 | 299 | 4.01 | 65 |
| 3. Hồ sơ hoàn thuế | 197 | 296 | 2 | 295 | 103 | 190 | 293 | 35.15 | 199 |
| 3.1. Hoàn thuế GTGT | 56 | 69 | 1 | 68 | 24 | 53 | 77 | 31.17 | 47 |
| - Hồ sơ hoàn thuế đối với dự án ODA | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| - Hồ sơ hoàn thuế GTGT hoàn trước kiểm tra sau | 33 | 57 | 1 | 57 | 17 | 42 | 59 | 28.81 | 31 |
| - Hồ sơ hoàn thuế GTGT kiểm tra trước, hoàn sau | 23 | 12 | 0 | 11 | 7 | 11 | 18 | 38.89 | 16 |
| - Hoàn thuế GTGT tạm | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 3.2. Hoàn thuế TNCN | 8 | 18 | 1 | 18 | 19 | 7 | 26 | 73.08 | 0 |
| 3.3 Hồ sơ hoàn thuế khác | 133 | 209 | 0 | 209 | 60 | 130 | 190 | 31.58 | 152 |
| 4. Hồ sơ miễn giảm thuế | 8 | 10 | 1 | 10 | 5 | 1 | 6 | 83.33 | 12 |
| 4.1 Hồ sơ miễn giảm thuế TNCN | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 4.2 hồ sơ miễn giảm khác | 8 | 10 | 1 | 10 | 5 | 1 | 6 | 83.33 | 12 |
| 5. Hồ sơ khiếu nại, tố cáo về thuế | 0 | 2 | 0 | 2 | 0 | 0 | 0 | 0 | 2 |
| 6. Hồ sơ mua hoá đơn thuế | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 7. Hồ sơ đăng ký tự in hoá đơn, sử dụng hoá đơn tự in thuế | 7 | 30 | 0 | 30 | 13 | 7 | 20 | 65 | 17 |
| 8. Hồ sơ giải quyết nợ thuế (khoanh, xoá) | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 9. Hồ sơ khác | 19 | 85 | 1 | 85 | 75 | 14 | 89 | 84.27 | 15 |

Bảng 4.2: Tổng hợp việc nhận hồ sơ và giải quyết hồ sơ đúng hạn tại Cục thuế trong 1 tháng của từng loại hồ sơ

Chúng tôi cũng tiến hành thu thập dữ liệu các các báo cáo tổng hợp về việc nhận và trả hồ sơ theo từng tháng tính từ năm 2017 đến tháng 10 năm 2020 để có được cái nhìn tổng quan nhất về việc xử lý nhận và trả hồ sơ tại Cục thuế tỉnh Bình Dương. Từ đó đưa vào mô hình để dự đoán tình hình xử lý hồ sơ những tháng tiếp theo như thế nào để báo cáo lãnh đạo đưa ra hướng giải quyết trước, giúp cho hồ sơ được giải quyết nhanh hơn và không còn bị trình trạng bị động trong quá trình xử lý.

## **4.2 Kết quả thực nghiệm**

Để tiến hành phân tích, đánh giá hiệu quả các mô hình khai phá dữ liệu dựa trên báo cáo tình hình, kết quả giải quyết thủ tục hành chính tại Cục thuế tỉnh Bình Dương, chúng tôi sử dụng ngôn ngữ lập trình Python (Python version: 3.6.5); thư viện Numpy, Pandas, Scipy để đọc dữ liệu và xử lý dữ liệu; thư viện Matplotlib, Seaborn để trực quan dữ liệu; thư viện học máy Scikit-Learn, Keras để cung cấp các giải thuật xây dựng các mô hình Linear Discriminant Analysis, K-Nearest Neighbors, Decision trees và Random forests. Cấu hình máy tính chạy thực nghiệm trên laptop với cấu hình CPU Intel Core Intel(R) i5 8265U\_CPU\_@\_1.80GHz, 8 GB RAM, hệ điều hành MS WINDOWS 10. Kết quả thực nghiệm được trình bày trong các phần tiếp theo.

### **4.2.1 Phân tích trực quan hóa số liệu các dịch vụ thủ tục hành chính công**

Kết quả về Phân tích trực quan hóa số liệu được thực hiện trên 18 thủ tục hành chính. Với mỗi đơn vị được biểu diễn trực quan ở 9 trường: tồn đầu kỳ, mới phát sinh, bổ sung, tổng số, đúng hạn, quá hạn, tỉ lệ đúng hạn, tồn cuối kỳ. Mỗi thủ tục hành chính được thể hiện ở một bảng trên mỗi tháng. Chúng tôi lấy từ 2017 đến tháng 10 năm 2020. Như vậy sẽ có 46 hình ảnh trực quan được biểu diễn.

Dưới đây là một số hình biểu diễn trực quan của thủ tục thuế ở các thời điểm khác nhau và trên những biểu đồ khác nhau.

*Hình 4.1. Số lượng giải quyết hồ sơ đúng hạn và quá hạn tháng 12 năm 2018*

*Hình 4.2. Số lượng giải quyết hồ sơ đúng hạn và quá hạn tháng 10 năm 2020*

Dựa vào 2 biểu đồ trực quan chúng ta có nhận định thông qua các năm thì hồ sơ đăng ký thuế là hồ sơ bị trể hạn nhiều nhất với số lượng vượt bậc. Trong luận văn này chúng tôi đã đưa ra mô hình dựa vào báo cáo kết quả thực hiện thủ tục hành chính để đưa ra dự đoán trong giai đoạn tới loại hồ sơ nào sẽ bị trể hạn và phòng nào sẽ bị trể hạn. Từ đó sẽ giúp cho lãnh đạo cục Thuế tỉnh Bình Dương đưa ra những quyết định nhằm nâng cao hiệu quả trong việc xử lý hồ sơ thủ tục hành chính.

### **4.2.2 Thực nghiệm đánh giá mô hình dự đoán, lựa chọn mô hình tối ưu**

**4.2.2.1 Thực nghiệm đánh giá mô hình trên các độ đo.**

Thực nghiệm với tập dữ liệu đã được làm sạch với với các mô hình dự báo đã được nêu ở Chương 3:

- Xây dựng 04 mô hình dự báo Logistic Regression, K-nearest neighbors, Decision trees và Random forests với các tham số trên từng mô hình chúng tôi chọn mặc định, vì dựa vào các độ đo để chọn mô hình tốt nhất trên cùng 01 cơ sở dữ liệu nên không cần điều chỉnh tham số.

- Trong quá trình thực nghiệm, chúng tôi lập lại nhiều lần để thay đổi tỉ lệ phân chia giữa tập huấn luyện và tập kiểm tra với các tập dữ liệu được chương trình chọn ngẫu nhiên và tiến hành kiểm tra chọn độ chính xác của các model tốt nhất bằng nhiều cách khác nhau như: Kiểm tra độ chính xác của kết quả dự báo cho từng thuộc tính (test\_score); Kiểm tra Hệ số tương quan Pearson (Pearson correlation coefficient, kí hiệu r) đo lường mức độ tương quan tuyến tính giữa hai biến; Hệ số tương quan hạng Spearman để kiểm tra mối quan hệ giữa hai biến được xếp hạng hoặc một biến được xếp hạng và một biến đo lường và kết quả tối ưu cho việc phân chia tập dữ liệu là 80% làm tập huấn luyện và 20% còn lại làm tập kiểm tra cho các mô hình dự báo

Sau khi chạy thực nghiệm chúng tôi thu được kết quả của từng độ đo như sau:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Thuật toán | Độ đo | Hồ sơ trể hẹn | | |
| Loại Hồ sơ | Trể hẹn | Phòng ban |
| **Random Forests** | **Spearman** | **1** | **1** | **1** |
| **Pearson** | **1** | **1** | **1** |
| **test\_score** | **0.999984** | **0.999996** | **0.999970** |
| K-nearest Neighbors | Spearman | 0.22 | 0.336 | 0.478 |
| Pearson | 0.43 | 0.623 | 0.467 |
| test\_score | 0.161907 | 0.322153 | 0.205894 |
| Decision Trees | Spearman | 1 | 1 | 1 |
| Pearson | 1 | 0.995 | 1 |
| test\_score | 0.997338 | 0.926353 | 0.999112 |
| Logistic Regression | Spearman | 0.982 | 0.935 | 0.722 |
| Pearson | 1 | 1 | 1 |
| test\_score | 1 | 1 | 1 |

Bảng 4.3: Kết quả độ đo Hồ sơ trể hẹn

Dựa vào bảng số liệu của bảng 4.3 chúng tôi có một số nhận xét như sau: độ đo thuộc thuật toán Random Forests là cao nhất các độ đo tiến gần tới ngưỡng 1, các độ đo thuộc thuật toán Decision Trees cũng ở mức khá cao và tiếp đến là thuật toán Logistic Regression và độ đo của thuật toán K-nearest Neighbors là thấp nhất.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Thuật toán** | **Độ đo** | **Thống kê hồ sơ** | | |
| **Tồn đầu kỳ** | **Tồn cuối kỳ** | **Tỉ lệ trể hạn** |
| Random Forests | **Spearman** | **1** | **1** | **1** |
| **Pearson** | **0.999** | **0.999** | **0.998** |
| **test\_score** | **0.952** | **0.997** | **1** |
| K-nearest Neighbors | Spearman | 0.531 | 0.676 | 0.919 |
| Pearson | 0.093 | 0.658 | 0.861 |
| test\_score | 0.04 | 0.034 | 0.192 |
| Decision Trees | Spearman | 1 | 1 | 1 |
| Pearson | 0.762 | 0.871 | 0.823 |
| test\_score | 0.941 | 0.91 | 0.88 |
| Logistic Regression | Spearman | 0.397 | 0.445 | 0.882 |
| Pearson | 0.293 | 0.709 | 0.708 |
| test\_score | -0.29 | 0.483 | 0.475 |

Bảng 4.4 Kết quả giải quyết hồ sơ theo tháng.

Trong bảng kết quả trên chúng tôi nhận thấy kết quả nghiên hẳn về thuật toán Random Forests là cao nhất, tiếp theo là các độ đo thuộc thuật toán Decision Trees. Nhưng đối với các thuộc tính này thì có sự thay đổi giữa Logistic Regression và K-nearest Neighbors. Các độ đo của thuật toán K-nearest Neighbors vượt lên vị trí thứ 3, trong khi Logistic Regression lại có độ đo chạy về âm.

Từ những số liệu về các độ đo của các mô hình trên, chúng tôi sẽ lấy trung bình các độ đo, có bảng số liệu như sau:

|  |  |  |
| --- | --- | --- |
| Thuật toán | Độ đo | Trung bình |
|  |
| **Random Forests** | **Spearman** | **1** |  |
| **Pearson** | **0.999333333** |  |
| **test\_score** | **0.991491667** |  |
| K-nearest Neighbors | Spearman | 0.526666667 |  |
| Pearson | 0.522 |  |
| test\_score | 0.159325667 |  |
| Decision Trees | Spearman | 1 |  |
| Pearson | 0.9085 |  |
| test\_score | 0.9423005 |  |
| Logistic Regression | Spearman | 0.727166667 |  |
| Pearson | 0.785 |  |
| test\_score | 0.611333333 |  |

Bảng 4.5: Bảng trung bình cộng các độ đo khi đã phân tích

Hình 4.3: Biểu đồ thể hiện độ đo trung bình của các mô hình

Từ biểu đồ của hình 4.3 chúng tôi nhận định độ đo của thuật toán Random Forests là tối ưu nhất. Kết quả này phù hợp với mô hình mà chúng tôi đã đề xuất ở chương 3.

Dựa vào kết quả trên chúng tôi đã thay đổi các tham số của Mô hình Random forests chúng tôi lần lược thay đổi tham số n\_estimators (Số lượng cây trong rừng), tham số này đạt kết quả dự báo tối ưu nhất là từ 500 đến 1.000, tuy nhiên khi chọn tham số tiến gần đến 1.000 thì chương trình chạy tốn nhiều thời gian mà kết quả cũng không thay đổi nhiều, vì vậy chúng tôi chọn tham số n\_estimators =500, các tham số còn lại là chọn mặc định. Sau khi đã tối ưu thuật toán và chạy thử chúng tôi có được bảng số liệu sau:

|  |  |  |
| --- | --- | --- |
| **NĂM  DỰ BÁO** | **ĐỘ LỆCH TRUNG BÌNH** | **ĐỘ CHÍNH XÁC TRUNG BÌNH** |
| 2017 | 23% | 77% |
| 2018 | 19% | 81% |
| 2019 | 15% | 85% |
| 2020 | 6% | 94% |

Bảng 4.6: Dự báo bằng mô hình Random forest trên bộ dữ liệu tự sinh và bộ dữ liệu thật

Hình 4.4: Biểu đồ thể hiện độ chính xác trung bình khi dự báo bằng mô hình Random Forest

Qua thực nghiệm Chúng tôi thu được kết quả như trên, từ kết quả này Chúng tôi thấy khai phá dữ liệu dựa trên báo cáo tình hình kết quả thủ tục hành chính tại cục Thuế tỉnh Bình Dương thì mô hình Random Forest là cho độ chính xác cao nhất. Còn mô hình K-nearest Neighbors cho kết quả không đáng tin cậy nhất. Từ kết quả này chúng tôi chọn mô hình Random Forest để thực hiện trong luận văn này.

# **Chương 5. KẾT LUẬN VÀ HƯỚNG PHÁT TRIỂN**

Chúng tôi vừa trình bày trong bài viết về Khai phá dữ liệu dựa trên báo cáo tình hình kết quả giải quyết thủ tục hành chính thuế tại cục thuế tỉnh Bình dương, trong thời gian 3 năm 2017, 2018, 2019 và đến tháng 10 năm 2020 chúng tôi đã tiến hành:

Phân tích các yếu tố ảnh hưởng đến thủ tục hành chính thuế tại cục thuế tỉnh Bình Dương bằng các độ đo Pearson và Spearman từ đó xác định được các yếu tố then chốt ảnh hưởng tới việc xử lý thủ tục hành chính.

Từ các yếu tố ảnh hưởng tới chất lượng dịch vụ của thủ tục hành chính, chúng tôi đã tiến hành Dự báo sự hiệu quả của việc xử lý thủ tục hành chính tại cục thuế tỉnh Bình Dương bằng 4 phương pháp học máy: Random Forest, Linear Regression, K-Nearest Neighbors và Decision Trees. Trên cơ sở thực nghiệm, đánh giá so sánh các kết quả đã đạt được của các mô hình, chúng tôi đã lựa chọn mô hình tối ưu để dự báo sự hiệu quả của việc xử lý thủ tục hành chính thuế tại cục thuế tỉnh Bình Dương. Quá trình xây dựng các mô hình Linear Regression, K- Nearest Neighbors, Decision Trees và Random Forests cần nhiều thời gian điều chỉnh các bộ tham số, để tạo ra mô hình dự báo tối ưu. Thực nghiệm cho thấy rằng mô hình dự báo sử dụng thuật toán Random Forest cho kết quả chính xác nhất.

Tuy nhiên, vì dữ liệu thu thập còn ít, các yếu tố tác động chưa được thu thập hết để từ đó đưa ra các mô hình phân tích dự báo đạt kết quả tốt hơn. Trong thời gian tới chúng tôi sẽ tiếp tục thu thập dữ liệu, phân tích khai thác dữ liệu trên tất cả các dịch vụ ở từng đơn vị, đánh giá hệ thống trên nhiều khía cạnh để từ đó xây dựng được giải pháp tối ưu khi đưa ứng dụng vào sử dụng trong thực tiễn.

# **TÀI LIỆU THAM KHẢO**

[1]. Niên giám thống kê tỉnh Bình Dương từ năm 2013 đến năm 2017.

[2]. Luật việc làm số 38/2013/QH13, ngày 16 tháng 11 năm 2013.

[3]. Đỗ Văn Chấn (1984) “Dự đoán nhu cầu triển vọng cán bộ chuyên môn của nước ta”, Viện Nghiên cứu Đại học và THCN.

[4]. Phạm Đức Chính (2006). Thị trường lao động, cơ sở lý luận và thực tiễn ở Việt Nam. Hà Nội: Nxb. Chính trị quốc gia.

[5]. Nguyễn Thị Lan Hương, Nguyễn Bá Ngọc, Nguyễn Bích Ngọc (2009). Lao động - việc làm trong thời kỳ hội nhập, Nhà xuất bản Lao động Xã hội.

[6]. Nguyễn Trọng Hoài, Phùng Thanh Bình, Nguyễn Khánh Duy (2009). Dự báo và phân tích dữ liệu trong kinh tế và tài chính, Nhà xuất bản thống kê.

[7]. Tác giả Nguyên Kim Sơn với đề tài “Sử dụng Data Mining dự báo nhu cầu lao động cho một số ngành nghề trên địa bàn tỉnh Bình Dương” 2019.

[8]. Trần Hữu Nam (2000), Đề tài “Nghiên cứu ứng dụng các phương pháp dự báo trong giáo dục - đào tạo”, Viện Nghiên cứu phát triển giáo dục.

[9]. Trần Thị Phương Nam (2014). Cơ sở khoa học của dự báo nhu cầu nhân lực trình độ cao đẳng, đại học ở Việt Nam, Luận án Tiến sĩ Khoa học giáo dục, Viện Khoa học Giáo dục Việt Nam.

[10]. Nguyễn Hà Nam, Nguyễn Trí Thành, Hà Quang Thụy (2013), Giáo Trình Khai Phá Dữ Liệu, Nhà xuất bản đại học quốc gia Hà Nội.

[11]. Đỗ Thanh Nghị, Trần Nguyễn Minh Thư, Bùi Võ Quốc Bảo, Phạm Nguyên Khang, 1984. Xây dựng và cân chỉnh mô hình dự báo mật số rầy nâutrên nền apache spark, Kỷ yếu Hội nghị Khoa học Quốc gia lần thứ IX - Nghiên cứu cơ bản và ứng dụng Công nghệ thông tin (FAIR'9), Trường Đại học Cần Thơ.

[12]. Nghiên cứu “Phát triển dịch vụ bưu chính của Tổng công ty Bưu chính Việt Nam trong giai đoạn hiện nay” của tác giả Trần Thị Hòa, 2011.

[13]. Nguyễn Tiệp (2008), Giáo trình thị trường lao động, nhà xuất bản Lao động – Xã hội, Hà Nội.

[14]. Nguyễn Văn Tuấn, Phân tích dữ liệu với R – Hỏi và đáp, nhà xuất bản Tổng hợp TPHCM, 2018.

[15]. Techtalk via [Viblo](https://viblo.asia/p/10-thuat-toan-hoc-may-ma-cac-ky-su-can-biet-RQqKLn8ml7z) (2018). 10 thuật toán machine learning mà lập trình viên cần biết, <https://techtalk.vn/10-thuat-toan-machine-learning-ma-lap-trinh-vien-can-biet.html>, xem 08/10/2018.

[16]. Machine Learning cơ bản (Dec 28, 2016). Bài 3: Linear Regression <https://machinelearningcoban.com/2016/12/28/linearregression/>, xem 08/10/2018.

[17]. Machine Learning cơ bản (Jan 8, 2017). Bài 6: K-nearest neighbors, <https://machinelearningcoban.com/2017/01/08/knn/>, xem 08/10/2018.

[18]. Machine Learning cơ bản (Jan 14, 2018). Bài 34: Decision Trees (1), <https://machinelearningcoban.com/2018/01/14/id3/>, xem 08/10/2018.

[19]. Leo Breiman, Jerome Friedman, Charles J. Stone & R.A. Olshen, Classification and Regression Trees – Taylor & Francis, 1984.

[20] Breiman, L.: “Bagging predictors”, Machine Learning vol.24(2):123–140, 1996.

[21] Breiman, L.: “Arcing classifiers”, The Annals of Statistics vol.26(3):801-849, 1998.

[22] Breiman, L.: “Random forests”, Machine Learning vol.45(1):5-32, 2001.

[23] Jiawei Han, Micheline Kamber, “Data Mining: Concepts and Techniques”, Second Edition, Morgan Kaufmann Publishers, 2006.

[24] David Hand, Heikki Mannila, Padhraic Smyth, “Principles of Data Mining”, MIT Press, 2001.

[25]. J. Han, M. Kamber, and Jian Pei, Data Mining: Concepts and Techniques (3rd edition). Morgan Kaufmann, 2011.

[26]. John E.Hanke & Dean W.Wichern, Business Forecasting, 8th Edition, 2005.

[27]. Gregory Piatetsky – Shapiro, Data Mining Course (Power Point Version), 2006.

[28]. J.Holton Wilson & Barry Keating, Business Forecasting With Accompanying Excel – Based ForecastXTM Software, 5th Edition, 2007.

[29]. Ross Ihaka. Time Series Analysis, Lecture Notes for 475.726, Statistics Department, University of Auckland, 2005.