ĐẠI HỌC QUỐC GIA HÀ NỘI
TRƯỜNG ĐẠI HỌC CÔNG NGHỆ

NINH HOÀI ANH

NGHIÊN CỨU VÀ XÂY DỰNG ỨNG DỤNG PHÂN TÍCH DỮ LIỆU KINH DOANH THIẾT BỊ ĐIỆN TỬ

Ngành: Công nghệ thông tin
Chuyên ngành: Kỹ thuật phần mềm
Mã số: 60480103

LUẬN VĂN THẠC SĨ NGÀNH CÔNG NGHỆ THÔNG TIN

NGƯỜI HƯỚNG DẪN KHOA HỌC: PGS. TS. Nguyễn Hải Châu

Hà Nội - 2017

MỤC LỤC

LỜI CAM ĐOAN

Tôi xin cam đoan luận văn này là do tôi thực hiện, được hoàn thành trên cơ sở tìm kiếm, thu thập, nghiên cứu, tổng hợp phần lý thuyết và các phương pháp kĩ thuật được trình bày trong các tài liệu được công bố trong nước và trên thế giới. Các tài liệu tham khảo đều được nêu ở phần cuối của luận văn. Luận văn này không sao chép nguyên bản từ bất kì một nguồn tài liệu nào khác.

Nếu có gì sai sót, tôi xin chịu mọi trách nhiệm.

Học viên

Ninh Hoài Anh

DANH MỤC CÁC KÝ HIỆU VÀ CHỮ VIẾT TẮT

TT	Ký hiệu	Tiếng Anh	Giải thích theo tiếng Việt
01	ARFF	Attribute - relation file format	Định dạng tập tin thuộc tính - quan hệ
02	CDA	Confirmatory data analysis	Phân tích dữ liệu khẳng định
03	CPU	Central Processing Unit	Bộ vi xử lý trung tâm
04	DOM	Document Object Model	Mô hình đối tượng tài liệu
05	EDA	Exploratory data analysis	Phân tích dữ liệu thăm dò
06	ESS	Explained sum of squares	Tổng bình phương hồi quy
07	HTML	Hypertext markup language	Ngôn ngữ đánh dấu siêu văn bản
08	OLS	Ordinarry least square	Phương pháp bình phương tối thiểu
09	PRF	Popolartion regression function	Hàm hồi quy tổng thể
10	RSS	Residual sum of squares	Tổng bình phương sai số
11	SRF	Sample regression function	Hàm hồi quy mẫu
12	TSS	Total sum of squares	Tổng bình phương toàn phần

DANH MỤC CÁC HÌNH VẼ VÀ ĐỒ THỊ

DANH MỤC CÁC BẢNG

MỞ ĐẦU

Ngày nay, gắn liền với sự phát triển của Internet, mạng xã hội và các thiết bị di động là sự gia tăng dữ liệu không ngừng trên toàn cầu. Dữ liệu được sinh ra từng phút, từng giây, có ở khắp mọi nơi và chúng có thể chỉ cho ta thấy nhiều điều. Tuy nhiên, làm thế nào để dữ liệu trở nên có ý nghĩa lại trở thành một vấn đề không nhỏ đối với những cá nhân, tổ chức sở hữu những khối dữ liệu này. Trên thực tế, nhiều doanh nghiệp chưa được trang bị, ứng dụng hệ thống công nghệ thông tin cần thiết giúp khai thác dữ liệu hiệu quả, từ đó đưa ra những quyết định sáng suốt dựa trên những phân tích có chất lượng thay vì dựa trên trực giác hay kinh nghiệm trong quá khứ.

Với quy mô dữ liệu đa dạng, phong phú, dữ liệu có thể phản ánh thông tin từ nhiều khía cạnh của đời sống xã hội hiện đại. Ví dụ, các vị trí địa lý đều được dữ liệu hóa, đầu tiên là bằng kinh độ, vĩ độ và gần đây là thông qua các hệ thống định vị toàn cầu GPS (Global positioning system). Những cuốn sách, tài liệu giấy đã được số hóa thành ebook, các file tài liệu số với nhiều định dạnh như pdf, txt, rtf. Kể cả những mối quan hệ bạn bè, sự ưa thích "like" cũng được dữ liệu hóa qua các mạng xã hội như Facebook, Zalo,… Những loại dữ liệu này được sử dụng để phân tích nhờ vào sự giúp đỡ của những bộ máy tính với chi phí thấp, những phép toán thông minh, dựa trên những kiến thức toán học được vay mượn từ kiến thức thống kê. Thay vì dạy cho máy tính có thể lái xe hoặc phiên dịch một ngôn ngữ, chúng ta có thể cung cấp đủ dữ liệu để máy tính có thể tính toán ra xác suất của tất cả mọi thứ mà chúng ta muốn tính toán.

Phân tích dữ liệu (Data analysis) là khoa học khám phá dữ liệu thô nhằm rút ra kết luận từ những dữ liệu ấy. Phân tích dữ liệu được sử dụng trong nhiều ngành công nghiệp để hỗ trợ các công ty, tổ chức để đưa ra quyết định kinh doanh tốt hơn hoặc trong các ngành khoa học để xác nhận hay bác bỏ các mô hình, lý thuyết hiện có. Quá trình phân tích dữ liệu bao gồm các bước kiểm định, làm sạch, chuyển đổi, mô hình hóa và phân tích dữ liệu với mục đích tìm thông tin hữu ích, cho thấy kết luận hoặc hỗ trợ ra quyết định dựa trên bộ dữ liệu hiện có.

Vấn đề nghiên cứu và ứng dụng phân tích dữ liệu vào các lĩnh vực rất phổ biến và phát triển trên thế giới. Tuy nhiên, tại Việt Nam, vấn đề này còn chưa được ứng dụng rộng rãi, nhất là trong lĩnh vực kinh doanh thương mại. Trên cơ

sở các nghiên cứu đã có, luận văn tập trung vào các mục tiêu và các vấn đề cần giải quyết sau:

Mục tiêu và phạm vi nghiên cứu:

Luận văn tập trung nghiên cứu về mô hình hồi quy tuyến tính, phương pháp sử dụng mô hình hồi quy tuyến tính trong phân tích dữ liệu, tìm hiểu công cụ hỗ trợ phân tích dữ liệu Weka.

Mục tiêu chính của luận văn là dựa trên công cụ WEKA xây dựng được mô hình hồi quy tuyến tính dự đoán giá của mặt hàng máy tính xách tay trên thị trường Việt Nam thông qua việc phân tích dữ liệu bán hàng của Công ty cổ phần thương mại Nguyễn Kim. Từ đó, hỗ trợ các doanh nghiệp, nhà phân phối máy tính xách tay đưa giá bán cạnh tranh nhất trên thị trường. Bên cạnh đó, cũng giúp người tiêu dùng ước lượng chi phí để mua một chiếc máy tính xách tay phù hợp với nhu cầu của bản thân.

Phương pháp nghiên cứu:

Trong phạm vi luận văn này, tôi đã sử dụng 03 phương pháp nghiên cứu khoa học để tiếp cận và làm rõ những vấn đề của đề tài mà mình đã lựa chọn. Đó là các phương pháp nghiên cứu sau:

- Phương pháp phân tích và tổng hợp lý thuyết: Nghiên cứu các tài liệu khác nhau về mô hình hồi quy tuyến tính, phân tích dữ liệu và công cụ WEKA; phân tích để tìm hiểu sâu sắc đối với mỗi vấn đề và tổng hợp để có cái nhìn tổng quan, đầy đủ về các vấn đề cần tìm hiểu.

- Phương pháp thực nghiệm khoa học: Chủ động tiến hành thu thập, xử lý dữ liệu bán máy tính xách tay; sử dụng công cụ WEKA xây dựng mô hình hồi quy tuyến tính để dự báo giá.

- Phương pháp phân tích, tổng kết kinh nghiệm: Nghiên cứu, phân tích và đánh giá các mô hình đã xây dựng để từng bước xây dựng mô hình phù hợp nhất với độ tin cậy, chính xác cao hơn.

Bố cục của luận văn:

Luận văn được trình bày với bố cục gồm 04 chương với những nội dung chính như sau:

Chương 1 - Đặt vấn đề: Phát biểu bài toán, lựa chọn miền ứng dụng và giới thiệu các phương pháp và công cụ để giải quyết bài toán

Chương 2 - Mô hình hồi quy tuyến tính và công cụ hỗ trợ WEKA: Trình bày cơ sở lý thuyết của mô hình hồi quy, đi vào cụ thể với mô hình hồi quy tuyến tính. Đồng thời, giới thiệu về công cụ WEKA, xây dựng mô hình hồi quy tuyến tính với sự hỗ trợ của WEKA.

Chương 3 - Thực nghiệm và đánh giá kết quả: Sử dụng công cụ WEKA để xây dựng mô hình hồi quy tuyến tính dự báo giá bán máy tính xách tay của Công ty cổ phần thương mại Nguyễn Kim. Tiến hành phân tích, xây dựng mô hình và đánh giá kết quả thu được.

Chương 4 - Kết luận: Trình bày kết quả đạt được của luận văn và định hướng phát triển trong tương lai.

Chương 1
ĐẶT VẤN ĐỀ

1.1. Bài toán phân tích dữ liệu

Dữ liệu được tạo ra từng giây, từng phút trong đời sống xã hội hiện đại. Chúng có thể là dữ liệu web, dữ liệu từ các cảm biến, các tệp lưu nhật ký, dữ liệu cá nhân, dữ liệu từ các thiết bị thông minh,… Từ khối dữ liệu này mà chúng ta có thể tìm kiếm, khai thác và trích xuất ra những thông tin hữu ích. Làm thế nào để có được những thông tin ấy là vấn đề được đặt ra. Phân tích dữ liệu là một trong những chìa khóa giúp chúng ta giải quyết vấn đề nêu trên. Vậy phân tích dữ liệu là gì?

Phân tích dữ liệu là một trong những ứng dụng thực tiễn của kỹ thuật khai phá dữ liệu (Data mining). Phân tích dữ liệu là một quá trình trích xuất thông tin hữu ích từ tập dữ liệu được cung cấp. Các bước cơ bản của quá trình phân tích dữ liệu bao gồm: Kiểm định (Inspecting), làm sạch (Cleaning), chuyển đổi (Transforming), mô hình hóa (Modeling) và phân tích (Analysing) dữ liệu nhằm mục đích tìm kiếm thông tin, cho thấy kết luận, hỗ trợ đưa ra quyết định.

Hình 1.1. Các bước của quá trình phân tích dữ liệu

Trước khi có máy tính, nhiều phương pháp phân tích cho tập dữ liệu nhỏ đã phát triển và tập trung phân tích từng biến riêng lẻ. Ngày nay, khi khả năng tính toán của máy tính đã phát triển vượt bậc, phân tích dữ liệu đã phân tích đồng thời quan hệ của nhiều biến.

Phân tích dữ liệu được chia thành phân tích dữ liệu thăm dò EDA và phân tích dữ liệu khẳng định CDA. Phân tích dữ liệu thăm dò dùng dữ liệu để xác định mối quan hệ giữa các biến độc lập với biến phụ thuộc hay xác định các biến được đưa vào mô hình. Phân tích dữ liệu khẳng định sử dụng dữ liệu để khẳng định giả thiết là đúng hoặc sai. Hai phương pháp này không tách rời nhau mà luôn đi cùng nhau để tìm ra những thông tin hữu ích từ tập dữ liệu đã có. Trước hết, chúng ta sử dụng phương pháp EDA để xây dựng mô hình phù hợp từ tập

dữ liệu đã có. Sau đó, sử dụng phương pháp CDA để khẳng định những thông tin chúng ta nhận định là đúng hoặc sai.

1.2. Lựa chọn miền ứng dụng

Hiện nay, xung quanh chúng ta là một biển dữ liệu khổng lồ và không ngừng mở rộng. Khối dữ liệu này liên quan đến tất cả các khía cạnh của đời sống xã hội. Đáp ứng nhu cầu hiểu biết vô tận của con người, giới học thuật đã nghiên cứu về phân tích dữ liệu nhằm chắt lọc những thông tin cần thiết, bổ ích đối với mỗi cá nhân, mỗi tập thể, mỗi lĩnh vực, mỗi quốc gia,… Giới kinh doanh cũng không thể bỏ qua công cụ hữu ích này để tăng cường lợi nhuận của doanh nghiệp từng ngày, thậm trí từng giờ. Từ đó, nhiều giải pháp công nghệ đã được nghiên cứu với mục đích hỗ trợ các công ty, doanh nghiệp đưa ra các quyết định kinh doanh hợp lý, sáng suốt. Thực tế, các công ty, tập đoàn lớn đã chuyển dần từ chế tạo sản phẩm sang cung cấp dịch vụ phân tích kinh doanh. Trước đây, tập đoàn IBM chế tạo, sản xuất các sản phẩm công nghệ như máy chủ, máy tính xách tay, máy tính để bản, thiết bị cho hạ tầng công nghệ thông tin. Nhưng, ngày nay, tập đoàn IBM đang quan tâm nhiều hơn và cung cấp thêm dịch vụ phân tích kinh doanh.

Trong khuôn khổ luận văn này, tác giả tập trung nghiên cứu, ứng dụng phân tích dữ liệu vào lĩnh vực kinh doanh. Dữ liệu bán hàng của các công ty điện máy là khối dữ liệu đồ sộ với đa dạng các loại mặt hàng của nhiều nhà cung cấp được bày bán với mức giá có thể thay đổi theo thời gian và từng chương trình khuyến mãi khác nhau. Khối dữ liệu này được thể hiện đầy đủ và đáng tin cậy trên website của các công ty điện máy và có thể được thu thập một cách chính xác thông qua các công cụ sẵn có. Tác giả lấy dữ liệu bán hàng của Công ty Cổ phần thương mại Nguyễn Kim là điển hình. Phân tích dữ liệu bán hàng của Công ty cổ phần thương mại Nguyễn Kim để hỗ trợ các công ty điện máy dự đoán và đưa ra giá bán cạnh tranh nhất cho mặt hàng máy tính xách tay trên thị trường Việt Nam.

1.3. Phương pháp và công cụ

1.3.1. Lựa chọn phương pháp

Phân tích dữ liệu khẳng định là lựa chọn không thể bỏ qua để hỗ trợ đưa ra quyết định kinh doanh sáng suốt. Một mô hình dữ liệu được xây dựng dựa trên tập dữ liệu lịch sử. Những thuật toán học máy được sử dụng để xây dựng

mô hình dữ liệu ẩn giấu trong tập dữ liệu này. Sau khi mô hình dữ liệu được xác nhận, nó được coi là tổng quát hóa kiến thức và có thể dự đoán tương lai. Bằng cách này, các doanh nghiệp có thể dự đoán các nguy cơ tiềm ẩn trong tương lai để hoạch định chiến lược kinh doanh phù hợp.

Thống kê cung cấp các phương pháp, kỹ thuật xây dựng mô hình toán học để phân tích dữ liệu. Hai phương pháp thống kê chính được sử dụng trong phân tích dữ liệu là: Thống kê mô tả (Descriptive statistics) và thống kê suy diễn (Inferential statistics). Dữ liệu thống kê thường được thu thập để trả lời các câu hỏi được định trước. Thống kê mô tả tóm tắt dữ liệu từ một mẫu thí nghiệm còn thống kê suy diễn rút ra kết luận từ dữ liệu. Ngày nay, với sự phát triển không ngừng về khả năng tính toán của máy tính, thống kê được sử dụng nhiều trong học máy (Machine learning) nhằm xây dựng các mô hình toán cho các thuật toán học máy. Thống kê suy diễn được sử dụng nhiều trong phân tích dữ liệu khẳng định.

Trong khuôn khổ luận văn này, tác giả tập trung nghiên cứu mô hình hồi quy tuyến tính trong thống kê với mục đích xây dựng mô hình học máy cho bài toán phân tích dữ liệu để dự đoán tương lai.

1.3.2. Lựa chọn công cụ

Hiện tại, các công cụ hỗ trợ phân tích dữ liệu đã xuất hiện nhiều như R, SPSS, WEKA,… Tuy nhiên, tác giả lựa chọn và nghiên cứu phần mềm WEKA. Đây là phần mềm được phát triển bằng Java nhằm phát triển các kỹ thuật học máy và áp dụng chúng vào các bài toán khai phá dữ liệu trong thực tế.

Chương 2
MÔ HÌNH HỒI QUY TUYẾN TÍNH VÀ CÔNG CỤ HỖ TRỢ WEKA

2.1. Mô hình hồi quy tuyến tính

2.1.1. Lý thuyết về mô hình hồi quy

Phân tích hồi quy nghiên cứu sự phụ thuộc của biến phụ thuộc vào một hay nhiều biến độc lập để ước lượng hay dự đoán giá trị trung bình của biến phụ thuộc trên cơ sở các giá trị biết trước của biến độc lập. Phân tích hồi quy được mô hình hóa thông qua dưới dạng:

$$Y = f(X) + \varepsilon \quad (2.1)$$

Trong đó:

- X là biến độc lập

- Y là biến phụ thuộc

- ε là sai số ngẫu nhiên

- f(X) = E(Y|X) là hàm hồi quy tổng thể PRF cho biết giá trị trung bình của biến Y sẽ thay đổi như thế nào khi biến X nhận các giá trị khác nhau

Mô hình (2.1) được gọi là mô hình hồi quy. Để khảo sát mô hình hồi quy người ta tiến hành quan sát các bộ số (Xi, Yi). Ở lần quan sát thứ i, biến X nhận giá trị Xi, biến Y nhận giá trị Yi và sai số ngẫu nhiên là εi. Khi đó, mô hình (2.1) trở thành:

$$Yi = f(Xi) + \varepsilon i = E(Y|Xi) + \varepsilon i \quad (2.2)$$

εi là độ chênh lệch giữa giá trị quan sát Yi của biến phụ thuộc Y với giá trị trung bình của Y khi biến độc lập X nhận giá trị Xi. ε tồn tại bởi nhiều yếu tố tác động. Một yếu tố quan trọng là do ngoài các biến độc lập X đã được đưa vào mô hình có thể còn có các biến khác chưa được xem xét tới cũng ảnh hưởng đến giá trị của biến phụ thuộc Y nên ε đại diện cho phần ảnh hưởng ấy.

Từ (2.2) ta có: $\varepsilon i = Yi - f(Xi)$

$$\Rightarrow \quad \varepsilon i \rightarrow 0 \Leftrightarrow Yi - f(Xi) \rightarrow 0$$

Nếu εi có giá trị càng nhỏ thì biến phụ thuộc Y càng quan hệ mật thiết hay càng phụ thuộc vào biến độc lập X. Vì vậy, ε đóng vai trò quan trọng trong việc

đánh giá chất lượng của mô hình hồi quy. Việc xây dựng mô hình hồi quy tốt thực chất là xác định hàm hồi quy tổng thể f(X) sao cho sai số ngẫu nhiên ε của mô hình nhận giá trị nhỏ nhất. Khi đó, ta có thể ước lượng hay dự đoán giá trị của biến phụ thuộc Y trên cơ sở các giá trị biết trước của biến độc lập X với một độ tin cậy nhất định.

Trong nhiều trường hợp, ta không có điều kiện để xét toàn bộ tổng thể của một vấn đề. Khi đó, ta có thể ước lượng giá trị trung bình của biến phụ thuộc từ tập số liệu mẫu. Thống kê học cung cấp phương pháp điều tra chọn mẫu cho phép lấy tập số liệu tổng thể một số mẫu số liệu để nghiên cứu, phân tích và đưa ra kết quả cho tổng thể với độ tin cậy cho trước. Việc xây dựng hàm hồi quy tổng thể được thực hiện thông qua việc xác định hàm hồi quy mẫu SRF, dùng nó để ước lượng và kiểm định các giả thiết từ đó xây dựng hàm hồi quy tổng thể. Hàm hồi quy mẫu được xây dựng dựa trên tập số liệu mẫu.

Mô hình hồi quy được chia làm 02 loại:

- Mô hình hồi quy đơn với hàm hồi quy tổng thể chỉ có 1 biến độc lập

- Mô hình hồi quy bội với hàm hồi quy tổng thể có từ 2 biến độc lập trở lên

2.1.2. Mô hình hồi quy tuyến tính

Mô hình hồi quy tuyến tính là mô hình hồi quy mà trong đó hàm hồi quy tổng thể có dạng tuyến tính

$$f(Xi) = E(Y|Xi) = \beta_0 + \beta_1 X_1 i + \beta_2 X_2 i + \ldots + \beta_n X_n i \quad (2.3)$$

Trong đó:

- Xi là giá trị của các biến độc lập X ở quan sát thứ i

- E(Y|Xi) là giá trị trung bình của biến phụ thuộc Y khi biến độc lập X nhận các giá trị Xi ở quan sát thứ i

- β_0, β_1, β_2, …, β_n là các tham số hồi quy. Tham số hồi quy β_0 còn được gọi là hệ số tự do, nó cho biết giá trị trung bình của biến phụ thuộc Y là bao nhiêu khi biến độc lập X nhận giá trị "0". Tham số hồi quy β_j còn được gọi là các hệ số góc, nó cho biết giá trị trung bình của biến phụ thuộc Y sẽ thay đổi như thế nào khi giá trị của biến độc lập thứ j $X_j i$ tăng một đơn vị với điều kiện các biến độc lập khác không thay đổi giá trị.

Thật vậy: Giả sử $X_j i^1 = X_j i + 1$

$$\Rightarrow E(Y|Xi)^1 = \beta_0 + \beta_1 X_1 i + \beta_2 X_2 i + \ldots + \beta_j X_j i^1 + \ldots + \beta_n X_n i$$

$$= \beta_0 + \beta_1 X_1 i + \beta_2 X_2 i + \ldots + \beta j(X_j i + 1) + \ldots + \beta_n X_n i$$

$$= \beta_0 + \beta_1 X_1 i + \beta_2 X_2 i + \ldots + \beta_j X_j i + \ldots + \beta_n X_n i + \beta j$$

$$= E(Y|Xi) + \beta j$$

$$\Rightarrow \beta j = E(Y|Xi)^1 - E(Y|Xi)$$

Nếu $\beta j > 0$ thì $E(Y|Xi)^1 > E(Y|Xi)$ tức là giá trị trung bình của Y tăng. Ngược lại, nếu $\beta j < 0$ thì $E(Y|Xi)^1 < E(Y|Xi)$ tức là giá trị trung bình của Y giảm.

Thuật ngữ "tuyến tính" có thể được hiểu theo hai nghĩa: tuyến tính với tham số và tuyến tính đối với biến số. Tuy nhiên, hàm hồi quy tuyến tính luôn được hiểu là với tham số, nó có thể không tuyến tính với biến số.

Như đã trình bày ở phần trước:

- Nếu $f(Xi) = E(Y|Xi) = \beta_0 + \beta_1 Xi$ thì mô hình được gọi là mô hình hồi quy tuyến tính đơn.

- Nếu $f(Xi) = E(Y|Xi) = \beta_0 + \beta_1 X_1 i + \beta_2 X_2 i + \ldots + \beta_n X_n i$ với $n >= 2$ thì mô hình được gọi là mô hình hồi quy tuyến tính bội.

Đối với mô hình hồi quy tuyến tính, hàm hồi quy mẫu có dạng:

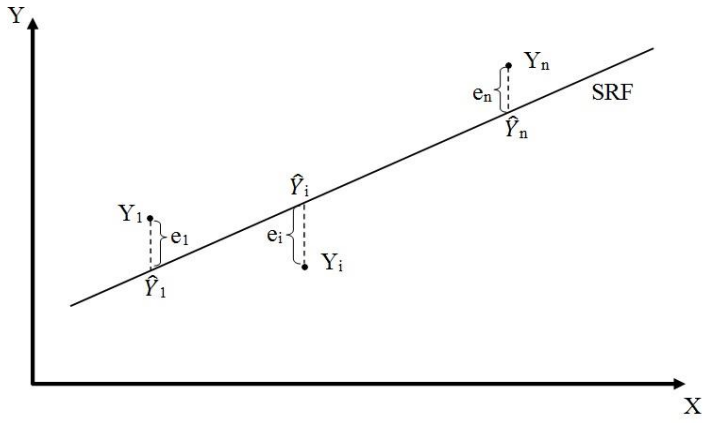
$$\hat{Y}i = \hat{\beta}_0 + \hat{\beta}_1 X_1 i + \hat{\beta}_2 X_2 i + \ldots + \hat{\beta}_n X_n i \qquad (2.3)$$

Trong đó:

- $\hat{\beta}_i$ là ước lượng điểm của β_i

- $\hat{Y}i$ là ước lượng điểm của Yi

Khi đó, sai số ei = Yi - $\hat{Y}i$. Minh họa bằng hình 2.1.



Hình 2.1. Sai số ei giữa Yi và $\hat{Y}i$

Như vậy, việc xây dựng mô hình hồi quy tuyến tính trở thành việc xác định các $\hat{\beta}_i$ sao cho sai số ei nhỏ nhất tức là $\hat{Y}i$ càng gần với giá trị Yi càng tốt.

2.1.3. Phương pháp bình phương tối thiểu để ước lượng các tham số của mô hình hồi quy tuyến tính

Phương pháp bình phương tối thiểu OLS được đưa ra bởi nhà toán học Carl Friedrich Gauss là phương pháp được sử dụng phổ biến nhất trong thống kê để xác định các $\hat{\beta}_i$ sao cho tổng bình phương các sai số ei giữa giá trị quan sát Yi với giá trị $\hat{Y}i$ tính theo hàm hồi quy mẫu là nhỏ nhất. Nội dung phương pháp cụ thể như sau:

Xét trường hợp, hàm hồi quy tổng thể có dạng:

$$f(Xi) = E(Y|Xi) = \beta_0 + \beta_1 Xi$$

và có một mẫu gồm n cặp quan sát (Xi, Yi) với i = 1, 2, …, n.

Ở lần quan sát thứ i, ta có:

- Hàm hồi quy mẫu: $\hat{Y}i = \hat{\beta}_0 + \hat{\beta}_1 Xi$

- Sai số: $e_i = Yi - \hat{Y}i = Yi - \hat{\beta}_0 - \hat{\beta}_1 Xi$

- Tổng bình phương các sai số ei:

$$\sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} \left(Yi - \hat{\beta}_0 - \hat{\beta}_1 Xi\right)^2$$

Việc cần làm là xác định $\hat{\beta}_0$ và $\hat{\beta}_1$ sao cho tổng bình phương các ei là nhỏ nhất. Tức là:

$$f(\hat{\beta}_0, \hat{\beta}_1) = \sum_{i=1}^{n} \left(Yi - \hat{\beta}_0 - \hat{\beta}_1 Xi\right)^2 \Rightarrow \textbf{min}$$

Vì $f(\hat{\beta}_0, \hat{\beta}_1)$ là đa thức bậc 2 của 2 biến $\hat{\beta}_0, \hat{\beta}_1$ nên điều kiện để nó đạt cực tiểu là:

$$\begin{cases} \frac{\partial f(\hat{\beta}_0, \hat{\beta}_1)}{\partial \hat{\beta}_0} = 2\sum_{i=1}^{n} \left(Yi - \hat{\beta}_0 - \hat{\beta}_1 Xi\right)(-1) = 0 \\ \frac{\partial f(\hat{\beta}_0, \hat{\beta}_1)}{\partial \hat{\beta}_1} = 2\sum_{i=1}^{n} \left(Yi - \hat{\beta}_0 - \hat{\beta}_1 Xi\right)(-Xi) = 0 \end{cases} \quad (2.4)$$

Giải hệ phương trình (2.4) ta được:

- $\hat{\beta}_0 = \bar{Y} - \hat{\beta}_1 \bar{X}$

- $\hat{\beta}_1 = \frac{\sum_{i=1}^{n} (Xi - \bar{X})(Yi - \bar{Y})}{\sum_{i=1}^{n} (Xi - \bar{X})^2} = \frac{\sum_{i=1}^{n} YiXi - n\,\bar{X}\,\bar{Y}}{\sum_{i=1}^{n} Xi^2 - n\,(\bar{X})^2}$

Trong đó:

- $\bar{X}$ là giá trị trung bình của X, $\bar{X} = \frac{\sum \text{Xi}}{n}$

- $\bar{Y}$ là giá trị trung bình của Y, $\bar{Y} = \frac{\sum \text{Yi}}{n}$

Các giả thuyết cơ bản của phương pháp bình phương tối thiểu:

- Giả thuyết 1: Quan hệ giữa X và Y là tuyến tính, các giá trị Xi cho trước và không ngẫu nhiên

- Giả thuyết 2: Các sai số e_i là đại lượng ngẫu nhiên có giá trị trung bình bằng 0 tức là $\mathbf{E(e_i|Xi) = 0}$

- Giả thuyết 3: Các sai số e_i là đại lượng ngẫu nhiên có phương sai không đổi tức là $\mathbf{Var(e_i|Xi) = \delta^2 = const}$

- Giả thuyết 4: Không có sự tương quan giữa các e_i tức là

$$\mathbf{Cov}(e_i|e_j) = 0 \text{ với } i \neq j$$

- Giả thuyết 5: Không có sự tương quan giữa e_i và Xi tức là

$$\mathbf{Var(e_i|Xi) = 0}$$

Định lý Gauss – Markov: Khi các giải thuyết 1 đến 5 được đảm bảo thì các ước lượng của phương pháp OLS là các ước lượng tuyến tính, không chệch và có phương sai nhỏ nhất trong lớp các ước lượng tuyến tính không chệch.

Đối với hàm hồi quy 2 biến thì $\hat{\beta}_0, \hat{\beta}_1$ tương ứng là các ước lượng tuyến tính, không chệch và có phương sai nhỏ nhất của β_0, β_1

Hệ số xác định r^2 (coefficient of determination) đo độ phù hợp của hàm hồi quy mẫu:

- Tổng bình phương toàn phần TSS: là tổng bình phương của tất cả các sai lệch giữa giá trị quan sát Yi với giá trị trung bình của chúng

$$\text{TSS} = \sum_{i=1}^{n}(Yi - \bar{Y})^2 = \sum_{i=1}^{n} Yi^2 - n(\bar{Y})^2 \quad (2.5)$$

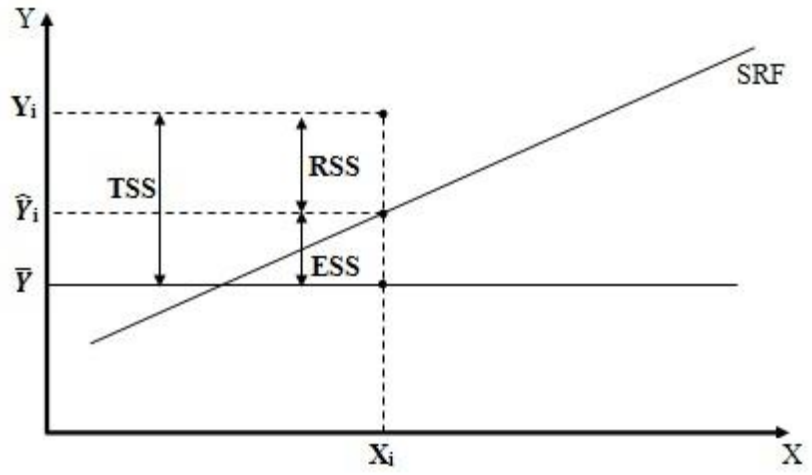
- Tổng bình phương hồi quy ESS: là tổng bình phương tất cả các sai lệch giữa giá trị của Y tính theo hàm hồi quy mẫu và giá trị trung bình

$$\text{ESS} = \sum_{i=1}^{n}\left(\hat{Y}i - \bar{Y}\right)^2 = \left(\hat{\beta}_1\right)^2 \sum_{i=1}^{n} Xi^2 - n(\bar{X})^2 \quad (2.6)$$

- Tổng bình phương sai số RSS: là tổng bình phương tất cả các sai lệch giá trị quan sát Yi với giá trị của Y tính theo hàm hồi quy mẫu

$$RSS = \sum_{i=1}^{n} e^2 = \sum_{i=1}^{n} \left(Yi - \hat{Y}i\right)^2 \quad (2.7)$$

Và TSS = ESS + RSS, minh họa bằng hình 2.2.



Hình 2.2. Mối quan hệ giữa TSS, ESS và RSS

- Hệ số xác định r^2 được xác định bởi công thức:

$$r^2 = 1 - \frac{RSS}{TSS} = \frac{ESS}{TSS} = \left(\hat{\beta}_1\right)^2 \frac{\sum_{i=1}^{n} Xi^2 - n(\bar{X})^2}{\sum_{i=1}^{n} Yi^2 - n(\bar{Y})^2} \quad (2.8)$$

=> $0 \leq r^2 \leq 1$

- Nếu $r^2 \rightarrow 1$: Hàm hồi quy mẫu là phù hợp, tất cả các sai lệch của Yi so với giá trị trung bình $\bar{Y}$ đều được giải thích bằng mô hình hồi quy

- Nếu $r^2 \rightarrow 0$: Hàm hồi quy mẫu là không phù hợp, biến phụ thuộc Y không phụ thuộc vào các biến độc lập X

Hệ số tương quan r (coefficient of correlation) đo độ tương quan giữa biến phụ thuộc Y và biến độc lập X: được xác định bởi công thức:

$$r = \frac{\sum_{i=1}^{n}(Xi - \bar{X})(Yi - \bar{Y})}{\sqrt{\sum_{i=1}^{n}(Xi - \bar{X})^2 \sum_{i=1}^{n}(Yi - \bar{Y})^2}} \quad (2.9)$$

- Có thể chứng minh được: $r = \pm\sqrt{r^2}$

=> Dấu của r trùng với dấu của $\hat{\beta}_1$

- Các tính chất của r:

+ Giá trị của r nằm trong khoảng {-1;1}

+ Nếu r > 0: X và Y có mối tương quan thuận

+ Nếu r < 0: X và Y có mối tương quan nghịch

+ $\hat{\beta}_1 = 0$ thì r = 0 và ngược lại, có thể căn cứ vào dấu của $\hat{\beta}_1$ để xác định tính thuận nghịch của mối tương quan

+ $|r| \to 1$ thì mối tương quan giữa X và Y càng chặt chẽ, nếu $|r| = 1$ thì X và Y có quan hệ hàm số

+ $|r| \to 0$ thì mối tương quan giữa X và Y càng lỏng lẻo, nếu $|r| = 0$ thì X và Y độc lập với nhau

2.1.4. Ứng dụng mô hình hồi quy tuyến tính vào phân tích dữ liệu

Trên thực tế, khi phân tích dữ liệu, chúng ta phải xác định mối quan hệ giữa một biến phụ thuộc vào nhiều biến độc. Ví dụ như, các yếu tố ảnh hưởng đến tốc độc của chiếc xe gắn máy đang chạy trên đường không chỉ phụ thuộc vào phân khối của động cơ mà còn phụ thuộc vào độ ma sát của mặt đường, sức cản của gió, trọng lượng hàng hóa trên xe,… Vì vậy, chúng ta cần xem xét các mô hình hồi quy tuyến tính nhiều hơn 1 biến độc lập.

Khi đó, hàm hồi quy tổng thể với k biến độc lập có dạng:

$$f(Xi) = E(Y|Xi) = \beta_0 + \beta_1 X_1 i + \beta_2 X_2 i + \ldots + \beta_k X_k i$$

Với n quan sát ta có:

$$Y_1 = \beta_0 + \beta_1 X_{11} + \beta_2 X_{21} + \ldots + \beta_k X_{k1} + e_1$$

$$Y_2 = \beta_0 + \beta_1 X_{12} + \beta_2 X_{22} + \ldots + \beta_k X_{k2} + e_2$$

$$\ldots\ldots\ldots\ldots$$

$$Y_n = \beta_0 + \beta_1 X_{1n} + \beta_2 X_{2n} + \ldots + \beta_k X_{kn} + e_n$$

Ký hiệu:

$$Y = \begin{bmatrix} Y_1 \\ Y_2 \\ \ldots \\ Y_n \end{bmatrix};\ \beta = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \ldots \\ \beta_k \end{bmatrix};\ e = \begin{bmatrix} e_1 \\ e_2 \\ \ldots \\ e_n \end{bmatrix} \text{ và } X = \begin{bmatrix} 1 & X_{11} & X_{21} & \ldots & X_{k1} \\ 1 & X_{12} & X_{22} & \ldots & X_{k2} \\ & \ldots\ldots\ldots\ldots\ldots & & & \\ 1 & X_{1n} & X_{2n} & \ldots & X_{kn} \end{bmatrix}$$

Ta có: $Y = X\beta + e$

Hàm hồi quy mẫu có dạng:

$$\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 X_{1i} + \hat{\beta}_2 X_{2i} + \ldots + \hat{\beta}_k X_{ki}$$

Khi đó: $e = \begin{bmatrix} e_1 \\ e_2 \\ \dots \\ e_n \end{bmatrix} = Y - X\hat{\beta}$

Các ước lượng OLS tìm được bằng cách tìm các $\hat{\beta}_i$ sao cho:

$$\sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} \left(Y_i - \hat{\beta}_0 - \hat{\beta}_1 X_{1i} - \hat{\beta}_2 X_{2i} - \dots - \hat{\beta}_k X_{ki}\right)^2 => Min$$

Gọi X^T, Y^T, $\hat{\beta}^T$, e^T lần lượt là ma trận chuyển vị của X, Y, $\hat{\beta}$ và e:

$$X^T = \begin{bmatrix} 1 & 1 & \dots & 1 \\ X_{11} & X_{12} & \dots & X_{1n} \\ X_{21} & X_{22} & \dots & X_{2n} \\ \dots & \dots & \dots & \dots \\ X_{k1} & X_{k2} & \dots & X_{kn} \end{bmatrix}; Y^T = \begin{bmatrix} Y_1 & Y_2 & \dots & Y_n \end{bmatrix}$$

$$\hat{\beta}^T = \begin{bmatrix} \hat{\beta}_1 & \hat{\beta}_2 & \dots & \hat{\beta}_n \end{bmatrix}; e^T = \begin{bmatrix} e_1 & e_2 & \dots & e_n \end{bmatrix}$$

Khi đó:

$$\sum_{i=1}^{n} e_i^2 = e^T e = \left(Y - X\hat{\beta}\right)^T \left(Y - X\hat{\beta}\right)$$

$$= \left(Y^T - \hat{\beta}^T X^T\right)\left(Y - X\hat{\beta}\right)$$

$$= Y^T Y - Y^T X\hat{\beta} - \hat{\beta}^T X^T Y + \hat{\beta}^T X^T X\hat{\beta}$$

$$= Y^T Y - 2\hat{\beta}^T X^T Y + \hat{\beta}^T X^T X\hat{\beta}$$

Hệ phương trình có dạng:

$$\frac{\partial (e^T e)}{\partial \hat{\beta}} = 0 \Rightarrow -2X^T Y + 2X^T X\hat{\beta} = 0$$

$$\Rightarrow X^T Y = X^T X\hat{\beta}$$

$$\Rightarrow \boldsymbol{\hat{\beta} = (X^T X)^{-1} X^T Y} \quad (2.10)$$

Trong đó ma trận $X^T X$ có dạng như sau:

$$X^T X = \begin{bmatrix} n & \sum_{i=1}^{n} X_{1i} & \sum_{i=1}^{n} X_{2i} & \dots & \sum_{i=1}^{n} X_{ki} \\ \sum_{i=1}^{n} X_{1i} & \sum_{i=1}^{n} {X_{1i}}^2 & \sum_{i=1}^{n} X_{1i}X_{2i} & \dots & \sum_{i=1}^{n} X_{1i}X_{ki} \\ \sum_{i=1}^{n} X_{2i} & \sum_{i=1}^{n} X_{2i}X_{1i} & \sum_{i=1}^{n} {X_{2i}}^2 & \dots & \sum_{i=1}^{n} X_{2i}X_{ki} \\ \dots & \dots & \dots & \dots & \dots \\ \sum_{i=1}^{n} X_{ki} & \sum_{i=1}^{n} X_{ki}X_{1i} & \sum_{i=1}^{n} X_{ki}X_{2i} & \dots & \sum_{i=1}^{n} {X_{ki}}^2 \end{bmatrix}$$

Kết quả: Các hệ số hồi quy được ước lượng theo công thức (2.10)

Hệ số xác định r^2 được định nghĩa như là tỷ lệ (%) sự biến động của biến phụ thuộc Y được giải thích bằng các biến độc lập X_k.

$$r^2 = 1 - \frac{RSS}{TSS} = \frac{ESS}{TSS} \quad (0 \leq r^2 \leq 1)$$

Với:

- RSS = $\sum_{i=1}^{n} e^2 = \sum_{i=1}^{n} \left(Yi - \hat{Y}i\right)^2$
- ESS = $\sum_{i=1}^{n} \left(\hat{Y}i - \bar{Y}\right)^2$
- TSS = TSS = $\sum_{i=1}^{n} (Yi - \bar{Y})^2$
- n là số lần quan sát

Hệ số tương quan r nói lên mối tương quan giữa biến phụ thuộc Y và các biến độc lập X_k.

$$r = \sqrt{r^2} \quad (-1 \leq r \leq 1)$$

Hệ số xác định đã điều chỉnh $\overline{r^2}$ để xác định có nên thêm 1 biến độc lập vào mới vào mô hình hay không. Thường thì giá trị của $\overline{r^2}$ có sự khác biệt rất ít so với r^2. Chúng ta có thể quyết định thêm một biến độc lập mới vào mô hình nếu $\overline{r^2}$ tăng lên khi tăng biến đó.

$$\overline{r^2} = 1 - (1 - r^2)\left(\frac{n-1}{n-k}\right)$$

Trong đó: k là số biến độc lập đưa vào mô hình

Ví dụ: Ta có số liệu quan sát của một mẫu được nêu trong Bảng 2.1

i	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**
X_1	8	7	8	8	6	6	5	5	4	3
X_2	2	3	4	4	5	5	6	7	8	8
Y	20	18	19	18	17	17	16	15	13	12

Bảng 2.1. Số liệu theo dõi dữ liệu bán hàng

Trong đó:

- Y là số lượng hàng bán được của một loại hàng (tấn/tháng)

- X1 là thu nhập của người tiêu dùng (triệu đồng/năm)

- X2 là giá bán của loại hàng này (ngàn đồng/kg)

Cần tìm hàm hồi quy: $\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 X_1 + \hat{\beta}_2 X_2$

Lời giải:

Ta tính được:

- $(X^T X)^{-1} = \begin{bmatrix} 10 & 60 & 52 \\ 60 & 388 & 282 \\ 52 & 282 & 308 \end{bmatrix}^{-1} = \frac{1}{1528} \begin{bmatrix} 39980 & -3816 & -3256 \\ -3816 & 376 & 300 \\ -3256 & 300 & 280 \end{bmatrix}$

- $X^T Y = \begin{bmatrix} 165 \\ 1029 \\ 813 \end{bmatrix}$

$\Rightarrow \hat{\beta} = \frac{1}{1528} \begin{bmatrix} 39980 & -3816 & -3256 \\ -3816 & 376 & 300 \\ -3256 & 300 & 280 \end{bmatrix} \begin{bmatrix} 165 \\ 1029 \\ 813 \end{bmatrix} = \begin{bmatrix} 14.99 \\ 0.76 \\ -0.59 \end{bmatrix}$

Vậy hàm hồi quy cần tìm là: $\hat{Y} = 14.99 + 0.76X_1 - 0.59X_2$

Khi đó ta có:

i	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**
Y	20	18	19	18	17	17	16	15	13	12
$\hat{Y}$	19.89	18.54	18.71	18.71	16.6	16.6	15.25	14.66	13.31	12.55

RSS = 2.2886 ESS = 56.1686 TSS = 58.5

r^2 = 0.960147 r = 0.979871 $\overline{r^2}$ = 0.955165

Vậy, với hàm hồi quy tìm được, sự biến động của số lượng hàng bán ra được giải thích theo thu nhập của người dùng và giá bán của sản phẩm với tỷ lệ 96%. Đồng thời, số lượng hàng bán ra có tương quan chặt chẽ với thu nhập của người dùng và giá bán của sản phẩm.

2.2. Công cụ hỗ trợ xây dựng mô hình hồi quy tuyến tính WEKA

2.2.1. Giới thiệu về WEKA

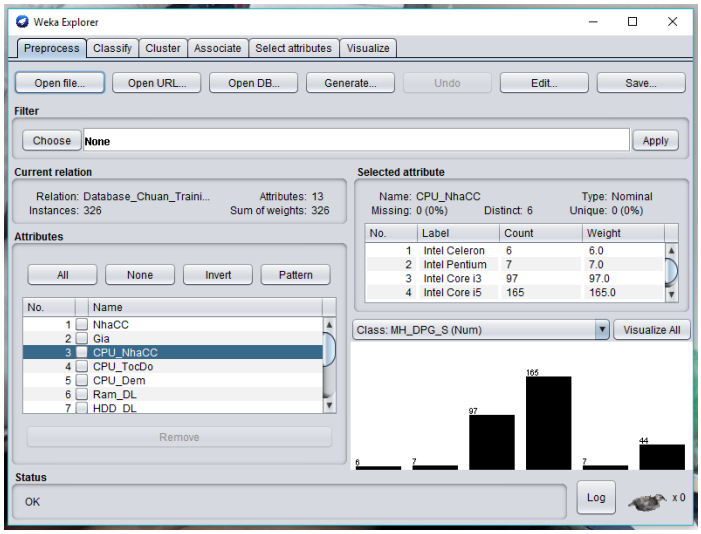
WEKA (**W**aikato **E**nvironment for **K**nowledge **A**nalysis) là một phần mềm khai phá dữ liệu mã nguồn mở được phát triển bởi Đại học Waikato ở New Zealand. WEKA cũng là tên một loài chim chỉ có trên một hòn đảo của New Zealand. WEKA được xây dựng bằng ngôn ngữ Java với mục tiêu xây dựng một công cụ hiện đại phát triển các kỹ thuật học máy và ứng dụng vào các bài toán khai phá dữ liệu trong thực tế.

Năm 1993, Đại học Waikato khởi động dự án nghiên cứu và xây dựng phiên bản đầu tiên của WEKA. Năm 1997, Đại học Waikato quyết định xây dựng lại WEKA từ đầu bằng ngôn ngữ Java và có cài đặt các thuật toán mô hình hóa. Đến năm 2005, WEKA được nhận giải thưởng SIGKDD Data Mining and Knowledge Discovery Service Award.

WEKA được chia sẻ rộng rãi trên website http://www.cs.waikato.ac.nz/~ml/weka/index.html. Hiện tại, phiên bản ổn định mới nhất của Weka là Weka 3.8. Ngoài ra, Đại học Waikato còn cung cấp phiên bản đang phát triển Weka 3.9. Đối với mỗi phiên bản, Weka được cung cấp đầy đủ các phiên bản cho hệ điều hành Windows, Mac OS X, Linux. Lưu ý, máy tính cần phải phiên bản Java cần thiết để để chạy một phiên bản Weka cụ thể. Với bản Weka 3.8 hiện tại, máy tính cần cài đặt phiên bản Java 1.7 trở lên.

WEKA được xây dựng với hơn 600 lớp, tổ chức thành 10 packages, mỗi package thực hiện một nhiệm vụ trong quá trình khai phá dữ liệu. Các lớp, packages này được mô tả một cách chi tiết trong tài liệu hướng dẫn sử dụng của nhà cung cấp. Giao diện đồ họa người sử dụng của WEKA được phát triển theo hướng trực quan và dễ sử dụng.

a. Giao diện chính



b. Giao diện chức năng "Explorer"

Hình 2.3. Một số hình ảnh về giao diện đồ họa người sử dụng của WEKA

2.2.2. Các chức năng chính của WEKA

WEKA cung cấp 5 môi trường làm việc nhằm hỗ trợ người sử dụng hai chức năng chính là khai phá dữ liệu và thực nghiệm, đánh giá các mô hình học máy. Cụ thể:

- *Explorer:* Môi trường cho phép tiến hành khai phá dữ liệu với các tính năng tiền xử lý dữ liệu (Preprocess), phân lớp (Classify), phân cụm (Cluster), khai thác luật kết hợp (Associate). Ngoài ra, nó còn cung cấp thêm tính năng hỗ trợ lựa chọn thuộc tính (Select attributes) và mô hình hóa dữ liệu (Visualize).

- *Experimenter:* Môi trường cho phép thực nghiệm (Setup, Run), so sánh, phân tích (Analyse) các mô hình học máy.

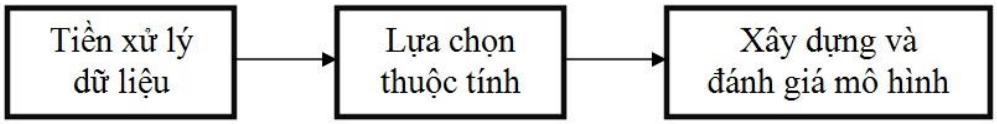
- *KnowledgeFlow:* Môi trường này hỗ trợ các tính năng cơ bản giống như Explorer nhưng với một giao diện kéo thả để hỗ trợ học tập gia tăng.

- *Simple CLI:* Cung cấp một giao diện dòng lệnh đơn giản cho phép thực thi trực tiếp các lệnh của WEKA cho các hệ điều hành không cung cấp giao diện dòng lệnh riêng.

- *Workbench:* Môi trường này là sự kết hợp của 4 môi trường nêu trên, người sử dụng có thể tùy ý chuyển đổi mà không cần phải quay lại cửa sổ "Weka GUI Chooser".

2.2.3. Xây dựng mô hình hồi quy tuyến tính với WEKA

Để xây dựng một mô hình hồi quy tuyến tính với WEKA, cần lựa chọn *Explorer* với các tính năng *Preprocess*, *Classify* và *Select attributes*. Quá trình xây dựng mô hình hồi quy tuyến tính với WEKA được thực hiện theo 03 bước: Tiền xử lý dữ liệu, lựa chọn các thuộc tính, xây dựng và đánh giá mô hình.



Hình 2.4. Các bước xây dựng một mô hình hồi quy tuyến tính với WEKA

Trước tiên, để tiền xử lý dữ liệu, cần chọn tính năng *Preprocess* của *Explorer*. Tính năng *Preprocess* cho phép lựa chọn và chỉnh sửa các tập dữ liệu được sử dụng để khai phá. WEKA có thể tiếp nhận dữ liệu từ các tập dữ liệu, từ các địa chỉ URL và từ các cơ sở dữ liệu SQL (thông qua JDBC).

Dữ liệu đầu vào của WEKA được định dạng chuẩn ARFF với phần mở rộng "*.arff". Tuy nhiên, WEKA cung cấp bộ chuyển đổi dữ liệu từ các định dạng "*.csv", "*.names", ".data", "*.json", "*.libsvm", "*.m", "*.dat", "*.bsi" sang dạng "*.arff". Ngoài ra, cũng có thể bổ sung các định dạng khác bằng cách thêm bộ chuyển đổi tập tin vào package "weka.core.converters". Người sử dụng cần mở tập tin dữ liệu ban đầu, tùy chỉnh dữ liệu rồi lưu lại với định dạng "*.arff".

Một tập tin ARFF là một tập tin văn bản theo bảng mã ASCII mô tả một danh sách các thể hiện (instances) của tập các thuộc tính. Một tập tin ARFF đơn giản có dạng:

```
@relation 1
@attribute name {John,Peter,Marry}
@attribute birthday date "yyyy-MM-dd HH:mm:ss"
@attribute math numeric
@attribute sentence string

@data
John,"2014-07-02 12:00:00",7,'aaa'
Peter,"2014-07-03 12:00:00",8,'aa b'
Marry,"2014-07-04 12:00:00",5,'Acvc aa1'
```

Như đã thấy ở trên, một tập tin ARFF gồm 02 phần riêng biệt. Phần thứ nhất có thể gọi là phần mô tả. Nó chứa thông tin về: tên của quan hệ, các thuộc tính của quan hệ và dạng dữ liệu của mỗi thuộc tính. Phần thứ hai là phần dữ liệu. Nó bắt đầu với từ khóa "@data" trên một dòng riêng biệt. Sau đó là các thể hiện (instance) và mỗi thể hiện trên một dòng. Giá trị các thuộc tính của một thể hiện phân cách nhau bằng một dấu ",". Giá trị các thuộc tính xuất hiện theo thứ tự được khai báo ở phần mô tả. Các giá trị bị thiếu sẽ được thể hiện bằng "?".

Tên của quan hệ được khai báo ở dòng đầu tiên của tập tin với cú pháp:

@relation <Tên quan hệ>

Một thuộc tính được khai báo với cú pháp:

@attribute <Tên thuộc tính> <Kiểu dữ liệu của thuộc tính>

Tên quan hệ và tên của thuộc tính là một chuỗi ký tự không bắt đầu bằng các ký tự đặc biệt "{", "}", ",", "%" và được đặt trong dấu nháy đơn nêu chứa khoảng trắng.

Tập tin ARFF định nghĩa các thuộc tính theo 4 kiểu chính:

- Thuộc tính dạng số: @attribute math **numeric**

- Thuộc tính dạng chuỗi: @attribute sentence **string**

- Thuộc tính định danh: @attribute name {<Giá trị 1>,<Giá trị 2>,<Giá trị 3>}

- Thuộc tính dạng ngày tháng: @attribute birthday date "<Định dạng ngày>", định dạng ngày theo tiêu chuẩn ISO-8601. Ví dụ "yyyy-MM-dd HH:mm:ss".

Đối với dữ liệu đầu vào, cần lưu ý với những thuộc tính có giá trị "0" và những thuộc tính chưa có giá trị (missing value). Thuộc tính có giá trị "0" không phải là thuộc tính chưa có giá trị. Những thuộc tính chưa có giá trị cần được thể hiện một cách rõ ràng bằng dấu "?".

```
@data
John,"2014-07-02 12:00:00",0,'aaa'
Peter,"2014-07-03 12:00:00",8,?
```

Sau khi tiền xử lý dữ liệu, cần lựa chọn các thuộc tính quan trọng, cần thiết để xây dựng mô hình hồi quy tuyến tính. Tập dữ liệu có rất nhiều thuộc tính để mô tả đầy đủ các khía cạnh của dữ liệu, tuy nhiên không phải tất cả các thuộc tính đề phù hợp để xây dựng mô hình hồi quy tuyến tính. Nói cách

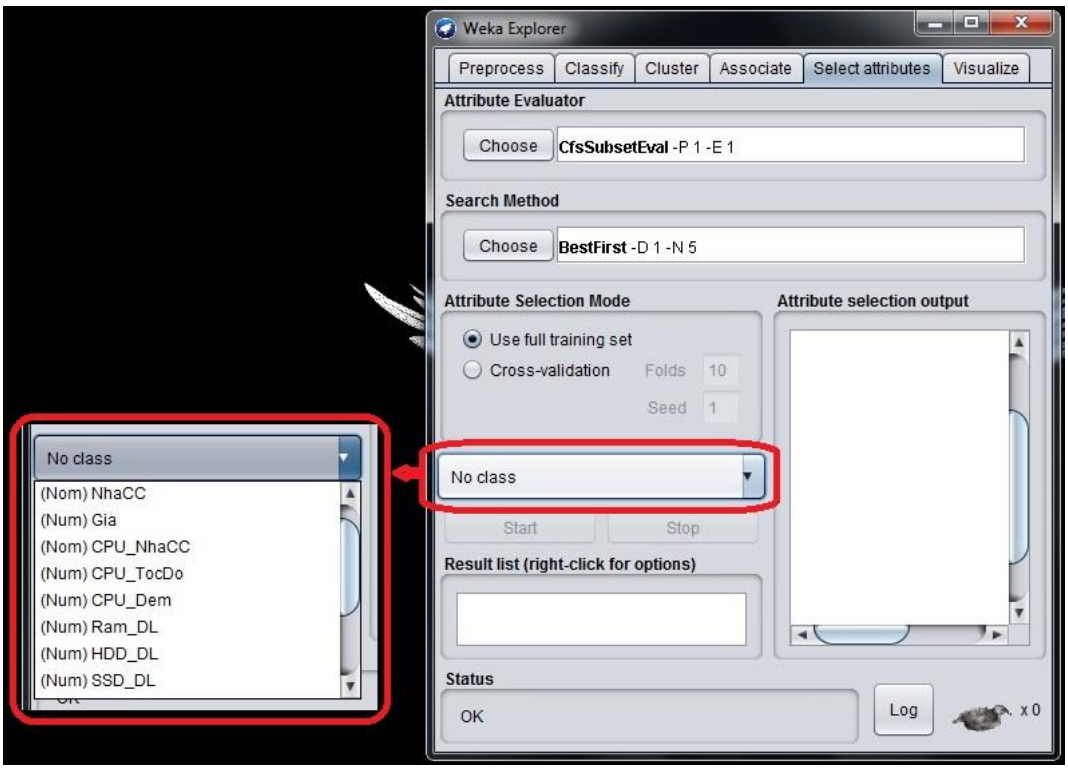
khác, việc lựa chọn thuộc tính chính là lựa chọn tập hợp các biến hồi quy để xây dựng mô hình.

Trên lĩnh vực kinh tế lượng thuần túy, mô hình hồi quy tuyến tính được xây dựng trên tập dữ liệu với các thuộc tính nhận giá trị số. Trên thực tế, tập dữ liệu không chỉ có những thuộc tính nhận giá trị số mà còn những thuộc tính nhận giá trị chuỗi ký tự, ngày tháng,… Đôi khi, những thuộc tính không nhận giá trị số này cũng đóng vai trò không kém phần quan trọng trong việc xây dựng mô hình hồi quy tuyến tính. Chẳng hạn, ngoài thuộc tính về diện tích, chiều dài mặt tiền, thuộc tính về loại nhà (nhà cấp 4, nhà cao tầng, biệt thự), thuộc tính về vị trí của ngôi nhà (trung tâm thành phố hay ở nông thôn) cũng ảnh hưởng đến giá bán của ngôi nhà. Những thuộc tính như vậy được gọi là thuộc tính phân loại.

WEKA cung cấp tính năng *Select attributes* của *Explorer* để hỗ trợ người sử dụng lựa chọn các thuộc tính xây dựng mô hình hồi quy tuyến tính. Tính năng *Select attributes* có nhiệm vụ tìm tập con các thuộc tính của tập dữ liệu để xây dựng được mô hình tin cậy nhất.

Để lựa chọn thuộc tính cần thiết lập bốn đối tượng cụ thể sau:

- Lựa chọn thuộc tính được dự đoán (biến phụ thuộc): Sử dụng dropdown liệt kê tập thuộc tính của tập dữ liệu.



Hình 2.5. Lựa chọn thuộc tính được dự đoán

- Bộ đánh giá thuộc tính (Attribute Evaluator): Để đánh giá tập các thuộc tính của tập dữ liệu. WEKA cung cấp 9 phương pháp đánh giá thuộc tính, gồm:

+ *CfsSubsetEval*: Đánh giá tập thuộc tính bằng cách xem xét khả năng dự đoán của từng thuộc tính riêng lẻ và mức độ dư thừa giữa chúng.

+ *CorrelationAttributeEval:* Đánh giá một thuộc tính dựa trên sự tương quan với lớp.

+ *GainRatioAttributeEval:* Đánh giá một thuộc tính dựa trên tỷ lệ gia tăng.

+ *InfoGainAttributeEval:* Đánh giá một thuộc tính dựa trên thông tin thu được.

+ *OneRAttributeEval:* Đánh giá một thuộc tính bằng cách sử dụng bộ phân loại OneR.

+ *PrincipalComponents:* Thực hiện phân tích thành phần chính và chuyển đổi dữ liệu.

+ *ReliefFAttributeEval*: Đánh giá thuộc tính dựa trên các thể hiện.

+ *SymmetricalUncertAttributeEval:* Đánh giá một thuộc tính dựa trên bất đối xứng.

+ *WrapperSubsetEval:* Đánh giá tập thuộc tính dựa trên một bộ phân loại cùng với xác nhận chéo.

- Phương thức tìm kiếm (Search Method): Để xác định phương pháp tìm kiếm được thực hiện. WEKA cung cấp 3 phương thức tìm kiếm, gồm:

+ *BestFirst:* Tiến hành kỹ thuật leo đồi tham lam kết hợp với quay lui.

+ *GreedyStepwise:* Thực hiện tìm kiếm tham lam về phía trước hoặc phía sau thông qua không gian các tập con thuộc tính.

+ *Ranker:* Xếp hạng các thuộc tính theo đánh giá trọng số của từng thuộc tính. Sử dụng kết hợp với các bộ đánh giá thuộc tính (ReliefF, GainRatio,…).

- Chế độ lựa chọn thuộc tính (Attribute Selection Mode): Xác định chế độ lựa chọn thuộc tính sử dụng tập huấn luyện đầy đủ hoặc tiến hành xác nhận chéo. Để xây dựng mô hình hồi quy tuyến tính, cần lựa chọn sử dụng tập huấn luyện đầy đủ.

Sau cùng, để xây dựng và đánh giá mô hình, WEKA hỗ trợ người sử dụng thông qua tính năng *Classify* của *Explorer*. Người sử dụng cần thiết lập ba đối tượng cụ thể sau:

- Bộ phân lớp (Classifer): Lựa chọn *functions/LinearRegression*. Việc xây dựng mô hình hồi quy tuyến tính được WEKA thực hiện trên cơ sở phương pháp bình phương tối thiểu. Có thể thực hiện lựa chọn thuộc tính bằng phương thức tham lam sử dụng loại bỏ lạc hậu hoặc xây dựng một mô hình đầy đủ từ tất cả các thuộc tính rồi loại bỏ dần các thuộc tính cho đến khi đạt được tiêu chí chấm dứt AIC. Ngoài ra, việc xây dựng mô hình được thực hiện với cơ chế phát hiện các thuộc tính đa cộng tuyến và cơ chế ổn định các trường hợp thoái hóa, giảm tình trạng quá tải thông bằng cách xử phạt các hệ số lớn

Tiêu chuẩn thông tin Akaike (AIC) của Nhà thống kê Akaike Hirotsugu người Nhật chỉ ra sự phù hợp của mô hình. Mô hình có tiêu chuẩn này càng nhỏ thì độ thích hợp của dữ liệu đối với mô hình càng cao. AIC là tiêu chuẩn được sử dụng phổ biến nhất trong các phân tích chuỗi thời gian và được tính theo công thức:

$$AIC = \left(\frac{ESS}{n}\right) e^{(2k/n)}$$

- Các tùy chọn kiểm thử (Test options): Tùy chọn phương pháp kiểm thử. WEKA cung cấp 4 phương pháp, gồm:

+ *Use training set:* Sử dụng tập dữ liệu mà bộ phân loại đã được huấn luyện.

+ *Supplied test set:* Cung cấp tập dữ liệu kiểm thử. Người sử dụng có thể lựa chọn tập dữ liệu kiểm thử bằng cách nháy vào nút "Set…"

+ *Cross-validation:* Tiến hành xác nhận chéo.

+ *Percentage split:* Chia tập dữ liệu thành 2 phần, huấn luyện trên 1 phần và kiểm thử trên phần còn lại. Phân chia tập dữ liệu theo tỷ lệ phần trăm do người sử dụng cài đặt.

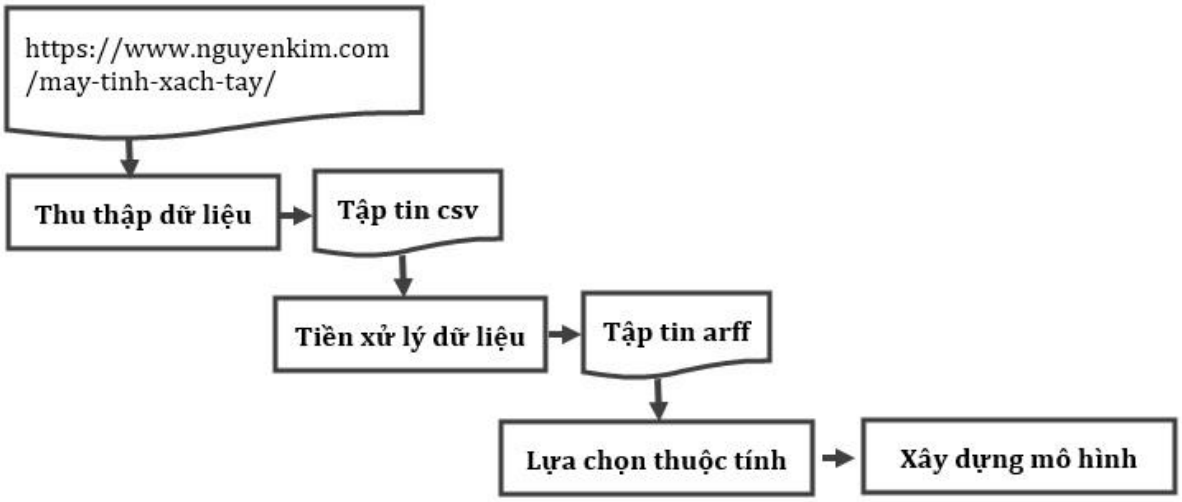
- Lựa chọn thuộc tính được dự đoán (biến phụ thuộc).

Chương 3
THỰC NGHIỆM VÀ ĐÁNH GIÁ KẾT QUẢ

3.1. Phát biểu bài toán thực tế

Công ty Cổ phần thương mại Nguyễn Kim là một thương hiệu hàng đầu trong ngành bán lẻ điện tử tiêu dùng, đạt nhiều giải thưởng trong nước và quốc tế, được người tiêu dùng tin tưởng và đánh giá cao. Với uy tín và thương hiệu của Công ty, Nguyễn Kim thường xuyên giới thiệu hàng nghìn mặt hàng thiết bị điện tử tới người tiêu dùng thông qua trang thông tin điện tử https://www.nguyenkim.com. Đây là nguồn dữ liệu kinh doanh thiết bị điện tử khổng lồ, hoàn toàn có thể khai thác để trích xuất ra những thông tin hữu ích.

Với mục đích nêu trên, tác giả đã tiến hành thu thập dữ liệu kinh doanh mặt hàng máy tính xách tay của Công ty Cổ phần thương mại Nguyễn Kim, xây dựng mô hình hồi quy tuyến tính trên tập dữ liệu thu thập được để dự báo giá bán sản phẩm. Quá trình thực nghiệm được tiến thành gồm 04 bước:



Hình 3.1. Các thực nghiệm xây dựng mô hình hồi quy tuyến tính để dự báo

Trong đó, dữ liệu đầu vào của quá trình thực nghiệm là những thông tin về mặt hàng máy tính xách tay được cung cấp trên trang thông tin điện tử của Công ty Cổ phần thương mại Nguyễn Kim. Dữ liệu đầu ra là một mô hình hồi quy tuyến tính với biến phục thuộc là giá bán mặt hàng máy tính xách tay, các biến độc lập là các thông tin về cấu hình, nhà cung cấp sản phẩm,… Thông qua mô hình hồi quy tuyến tính xây dựng được, người sử dụng có thể tính toán giá bán mặt hàng máy tính xách tay khi có sự thay đổi về cấu hình hay nhà cung cấp sản phẩm.

3.2. Tiến hành xây dựng mô hình

3.2.1. Thu thập dữ liệu

Điều kiện tiên quyết để xây dựng được mô hình hồi quy tuyến tính là cần phải thu thập một tập dữ liệu chính xác, đáng tin cậy và các thuộc tính nhận giá trị số. Do đó, dữ liệu kinh doanh mặt hàng máy tính xác tay của các công ty điện máy là một lựa chọn phù hợp với những ưu điểm như:

- Thuộc tính giá bán (biến phụ thuộc) phụ thuộc khá nhiều vào thông số kỹ thuật của từng dòng sản phẩm (biến độc lập) mà các thuộc tính này đều nhận giá trị số.

- Dữ liệu bán hàng được các nhà phân phối cung cấp đầy đủ trên website thương mại điện tử.

Tuy nhiện, quá trình thu thập dữ liệu cũng gặp phải không ít những khó khăn, điển hình là:

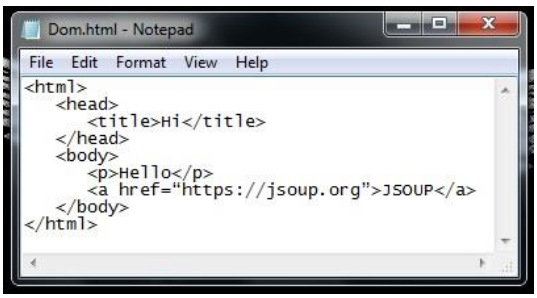
- Thông tin sản phẩm được cung cấp một cách không thống nhất theo một khuôn dạng nhất định.

- Các website thương mại điện tử được thiết kế để hạn chế và gây khó khăn cho việc khai thác dữ liệu.

Sau khi khảo sát thực tế, dữ liệu được thu thập tại website thương mại điện tử của Công ty cổ phần thương mại Nguyễn Kim với địa chỉ http://www.nguyenkim.com/may-tinh-xach-tay. Dữ liệu thu thập với những thông tin chính về tên sản phẩm, bộ vi xử lý CPU, ram, card màn hình, loại màn hình và giá thành sản phẩm.

Website thương mại điện tử của Công ty cổ phần thương mại Nguyễn Kim được thiết kế trên nền tảng HTML. Jsoup được lựa chọn để phân tích và khai thác dữ liệu từ một tài liệu HTML. Nó là một thư viện Java cung cấp các API để phân tích tài liệu HTML thành danh sách các phần tử và khai thác dữ liệu của từng phần tử. Người sử dụng có thể tải trực tiếp bộ thư viện Jsoup dưới dạng tập tin "jar" tại địa chỉ https://jsoup.org/download.

Jsoup phân tích tài liệu HTML thành mô hình DOM. Người sử dụng cần hiểu rõ bố cục của tài liệu HTML để truy cập chính xác đến từng phần tử cụ thể của danh sách.



a. Tập tin Dom.html b. Mô hình DOM của tập tin Dom.html

Hình 3.2. Mô hình DOM của tập tin HTML đơn giản

Jsoup bao gồm nhiều lớp đối tượng, nhưng ba lớp đối tượng chính và quan trọng nhất là *org.jsoup.Jsoup, org.jsoup.nodes.Document* và *org.jsoup.nodes.Element*. Người sử dụng có thể tìm hiểu cụ thể trên website https://jsoup.org.

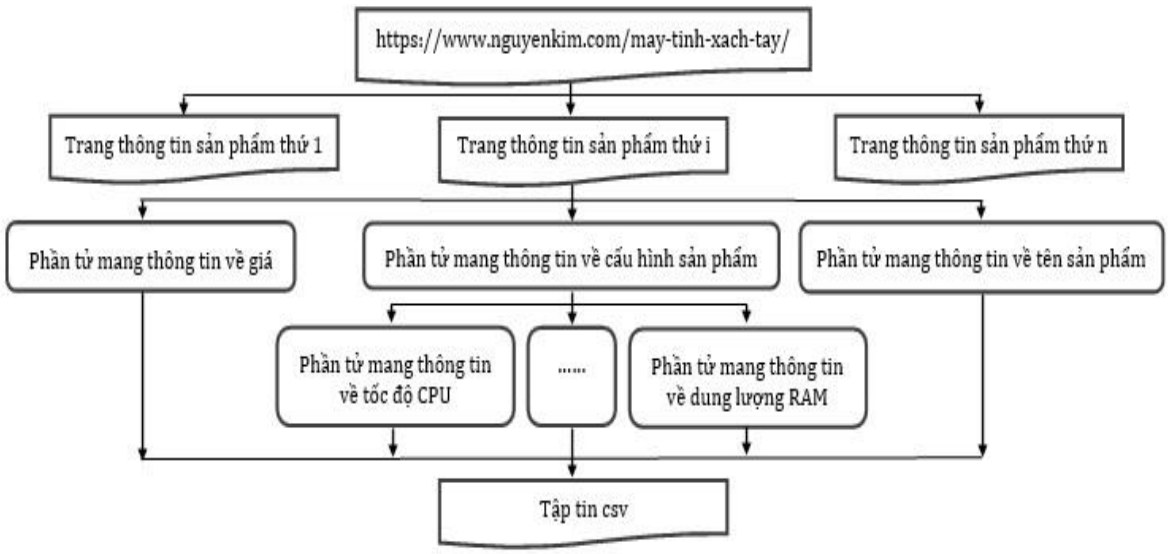
Dữ liệu kinh doanh mặt hàng máy tính xách tay của Công ty cổ phần thương mại Nguyễn Kim được thu thập với những thuộc tính tiêu biểu sau:

TT	Tên thuộc tính	Khuôn dạng dữ liệu	Mô tả
01	NgayTT	4/4/2016	Ngày dữ liệu được thu thập
02	NhaPP	Nguyen Kim	Nhà phân phối sản phẩm
03	NhaCC	Asus, Lenovo,…	Nhà cung cấp sản phẩm
04	TenSP	E402SA WX043D	Tên sản phẩm
05	Gia	5.300.000	Giá bán của sản phẩm
06	CPU_NhaCC	Intel Celeron	Tên nhà cung cấp CPU
07	CPU_TocDo	1.60 GHz, 800 MHz	Tốc độ CPU
08	CPU_Dem	2 MB Cache	Dung lượng bộ nhớ đệm của CPU
09	Ram_Loai	SDRAM DDR3	Loại Ram
10	Ram_DL	2 GB	Dung lượng Ram
11	Ram_Bus	1600 MHz	Tốc độ bus của Ram
12	HDD_Loai	SATA, SDD	Loại ổ cứng
13	HDD_DL	500 GB, 1 TB	Dung lượng ổ cứng
14	Card_Loai	Intel HD, GT 820M	Loại card màn hình
15	Card_DL	2GB, Share	Dung lượng card màn hình

16	MH_Loai	LED HD, Full HD	Loại màn hình
17	MH_KT	14.0 inch, 15.6 inch	Kích thước màn hình
18	MH_DPG	1366 x 768 Pixels	Độ phân giải màn hình

Bảng 3.1. Danh sách thuộc tính của tập dữ liệu thu thập

Qua quá trình khảo sát, chương trình thu thập dữ liệu được cài đặt trên cơ sở ngôn ngữ lập trình Java có sử dụng bộ thư viện Jsoup để khai thác thông tin của từng mặt hàng máy tính xách tay trên thương mại điện tử của Công ty Cổ phần thương mại Nguyễn Kim.



Hình 3.3. Quá trình khai thác thông tin từng sản phẩm

Quá trình thu thập dữ liệu được thực hiện liên tục để theo dõi sự thay đổi về giá bán của các dòng sản phẩm máy tính xác tay theo các thuộc tính được thu thập. Cụ thể, với khoảng thời gian từ 04/4/2016 đến 19/7/2016, tập dữ liệu thu thập ở định dạng ".csv" với 18 thuộc tính có 5.527 dòng dữ liệu với 327 dòng sản phẩm của 06 nhà cung cấp, 16 lần thu thập dữ liệu.

3.2.2. Tiền xử lý dữ liệu

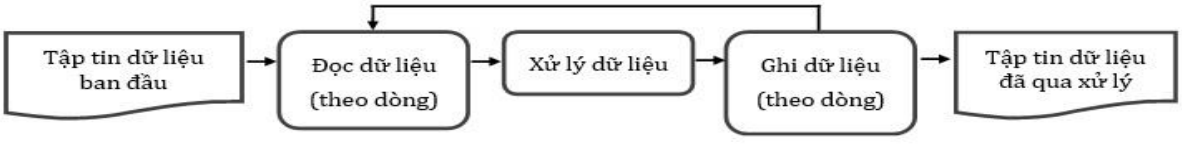
Như đã trình bày ở Chương 2, dữ liệu thu thập cần phải tiền xử lý trước khi đưa vào xây dựng mô hình. Ở đây, việc tiền xử lý dữ liệu được tiến hành theo 02 giai đoạn:

- Giai đoạn 1: Loại bỏ các dữ liệu dư thừa và chuẩn hóa khuôn dạng dữ liệu để WEKA có thể đọc được tập tin dữ liệu ".csv". Cụ thể:

Tên thuộc tính ban đầu	Khuôn dạng dữ liệu ban đầu	Tên thuộc tính mới	Khuôn dạng dữ liệu mới
NgayTT	4/4/2016	NgayTT	2016-04-04 12:00:00
Gia	5.300.000	Gia	5300000
CPU_NhaCC	Intel Celeron	CPU_NhaCC	Intel Celeron
CPU_TocDo	1.60 GHz, 800 MHz	CPU_TocDo	1.60, 0.8
CPU_Dem	2 MB Cache	CPU_Dem	1, 2, 3,…
Ram_DL	2 GB	Ram_DL	2, 4,…
Ram_Bus	1600 MHz	Ram_Bus	1600, 1333,…
HDD_Loai	SATA, SDD, SATA + SDD	HDD	SATA hoặc để trống
		SDD	SDD hoặc để trống
HDD_DL	500 GB, 1 TB, 1 TB + 128 GB	HDD_DL	0, 500, 1024
		SDD_DL	0, 128
Card_DL	2GB, Share	Card_DL	1, 2, 4, “0” với card share
MH_KT	14.0 inch, 15.6 inch	MH_KT	14.0, 15.6
MH_DPG	1366 x 768 Pixels	MH_DPG_W	1280, 1366, …
		MH_DPG_H	768, 800,…
Tất cả các giá trị chưa xác định được biểu diễn bởi “?”			

Bảng 3.2. Danh sách các thuộc tính đã tiền xử lý

Giai đoạn tiền xử lý này, chương trình được cài đặt bằng ngôn ngữ lập trình Java để tự động đọc dữ liệu từ tập tin dữ liệu ban đầu, xử lý các giá trị của từng dòng dữ liệu và ghi dữ liệu vào một tập tin mới.

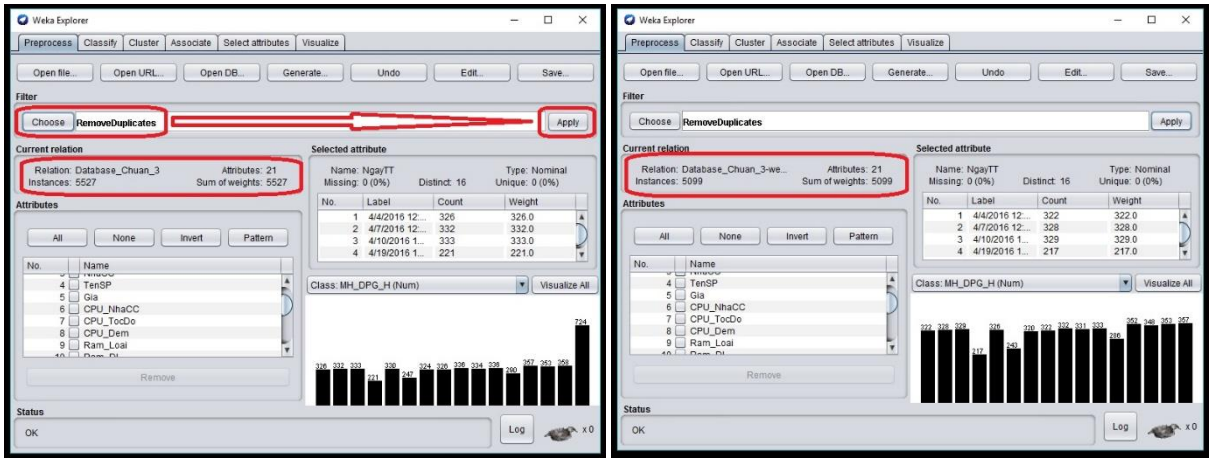


Hình 3.4. Quá trình tiền xử lý dữ liệu giai đoạn 1

Kết thúc giai đoạn 1, tập tin dữ liệu có 21 thuộc tính gồm: *NgayTT, NhaPP, NhaCC, TenSP, Gia, CPU_NhaCC, CPU_TocDo, CPU_Dem, Ram_Loai, Ram_DL, Ram_Bus, HDD, SSD, HDD_DL, SSD_DL, Card_Loai, Card_DL, MH_Loai, MH_KT, MH_DPG_W, MH_DPG_H.*

- Giai đoạn 2: Tiến hành lọc dữ liệu để loại bỏ các bản ghi giống nhau và xử lý các giá trị thiếu (missing value) bằng cách sử dụng các bộ lọc dữ liệu được WEKA cung cấp.

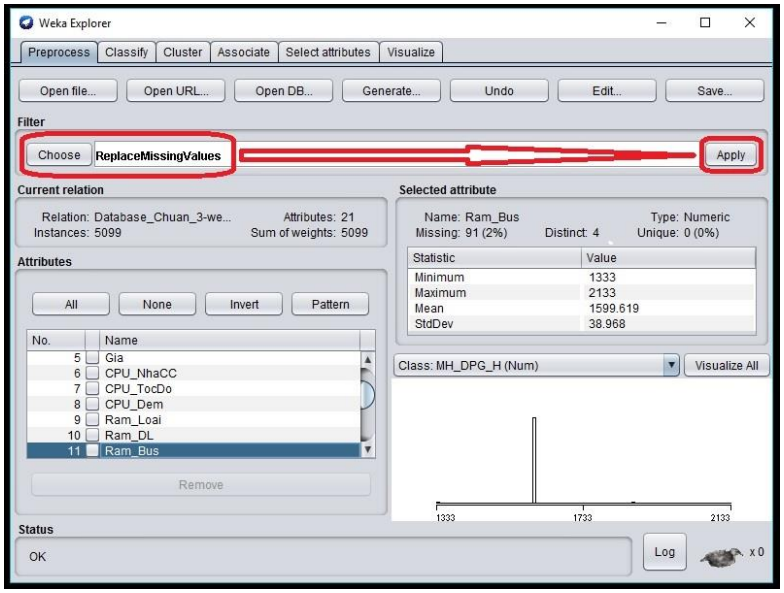
Để loại bỏ các bản ghi trùng lặp, sử dụng bộ lọc *RemoveDuplicates* của WEKA. Các bản ghi được coi là trùng lặp nếu chúng nhận giá trị của các thuộc tính hoàn toàn giống nhau. Sau khi lọc dữ liệu lần thứ nhất, tập tin dữ liệu còn 5.099 dòng dữ liệu.



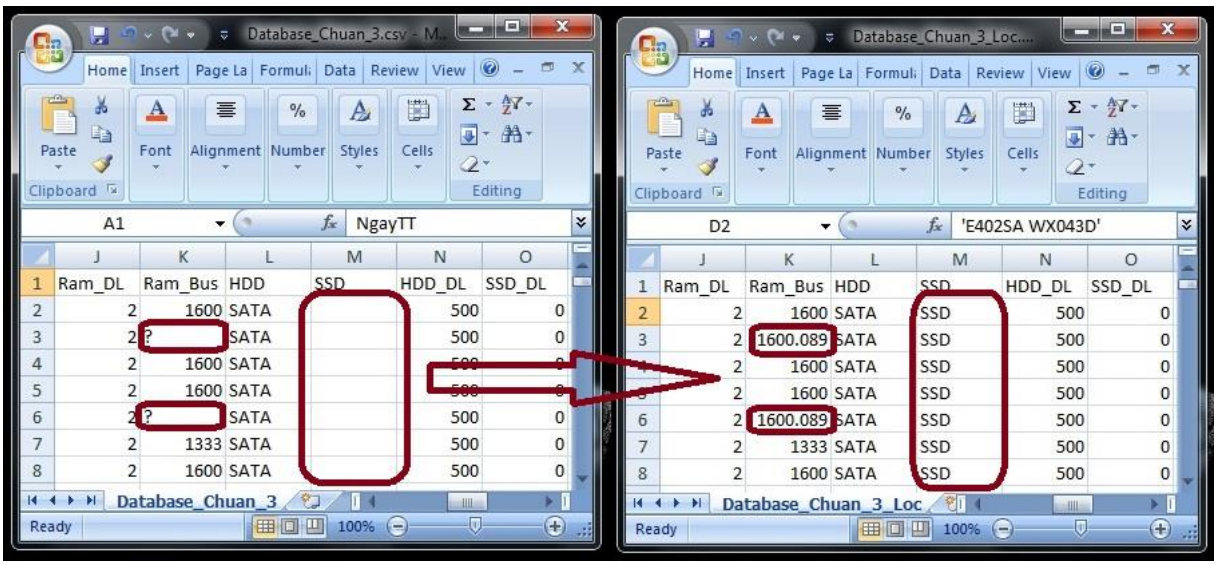
a. Lựa chọn bộ lọc dữ liệu b. Dữ liệu đã qua xử lý của bộ lọc

Hình 3.5. Loại bỏ các bản ghi giống nhau của tập tin dữ liệu

Để xử lý các giá trị thiếu, người sử dụng cần sử dụng bộ lọc *RepalceMissingValue* của WEKA để thay thế các giá trị thiếu dựa trên những giá trị đã biết của tập dữ liệu huấn luyện. Các giá trị thiếu sẽ được thay thế bằng giá trị trung bình cộng những giá trị đã biết của tất cả các bản ghi.



a. Lựa chọn bộ lọc dữ liệu



b. Dữ liệu đã qua xử lý của bộ lọc

Hình 3.6. Xử lý giá trị thiếu trong tập dữ liệu

Cuối cùng, để hoàn tất việc tiền xử lý dữ liệu, người sử dụng cần lưu lại tập dữ liệu với định dạng "*.arff".

3.2.3. Lựa chọn thuộc tính

Lựa chọn thuộc tính là bước đóng vai trò quan trọng trong quá trình xây dựng mô hình. Tập tin dữ liệu thu thập sau khi được tiền xử lý đã có 21 thuộc tính, trong đó thuộc tính "Gia" được xác định là thuộc tính được dự báo hay biến phụ thuộc trong mô hình hồi quy tuyến tính. Người sử dụng cần sử dụng tính năng *Select attributes* của *Explorer* để lựa chọn các thuộc tính độc lập xây dựng mô hình trong số 20 thuộc tính còn lại.

Phương thức tìm kiếm tập con thuộc tính được lựa chọn thông qua phương pháp *BestFirst*. Tập thuộc tính ban đầu chưa có thuộc tính nào được lựa chọn. Tìm kiếm tập con thuộc tính bằng cơ chế leo đồi tham lam kết hợp với cơ chế quay lui.

Phương thức đánh giá thuộc tính *CfsSubsetEval* được lựa chọn để tìm ra tập con thuộc tính có độ tương quan chặt chẽ với thuộc tính "Gia" được dự đoán.

Bốn đối tượng của tính năng *Select attributes* được lựa chọn như sau:

- Thuộc tính được dự đoán: *(Num) Gia*

- Chế độ lựa chọn thuộc tính: Sử dụng tập huấn luyện đầy đủ *Use full training set*

- Phương thức tìm kiếm: *BestFirst*

- Bộ đánh giá thuộc tính: *CfsSubsetEval*

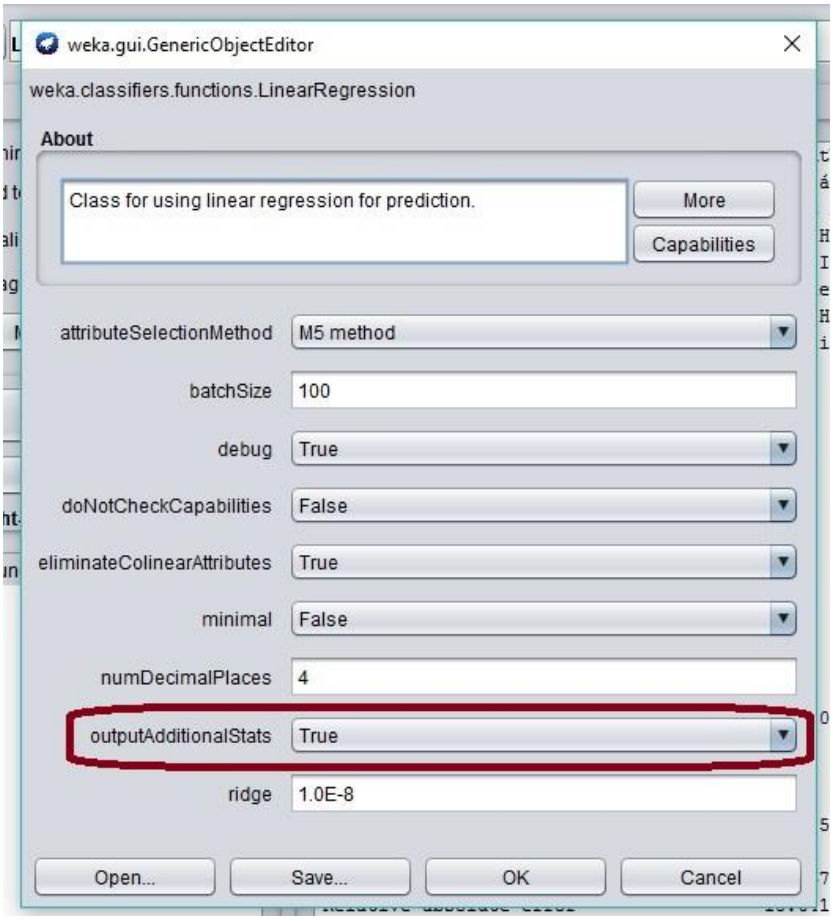
Kết quả 10 thuộc tính được lựa chọn, bao gồm: *NgayTT*, *NhaCC*, *CPU_NhaCC*, *CPU_Dem*, *Ram_DL*, *SSD_DL*, *Card_Loai*, *Card_DL*, *MH_Loai*, *MH_DPG_W*.

Cần loại bỏ các thuộc tính không được lựa chọn của tập tin dữ liệu thu thập và lưu lại tập tin dữ liệu đầu vào sử dụng để xây dựng mô hình hồi quy tuyến tính.

3.2.4. Xây dựng và đánh giá mô hình

Để xây dựng mô hình hồi quy tuyến tính, người sử dụng cần lựa tính năng *Classify* của *Explorer* và thiết lập các đối tượng như sau:

- Bộ phân lớp: Lựa chọn *functions/LinearRegression*. Ngoài ra, người sử dụng cần thiết lập thêm tùy chọn **outputAdditionalStats = True** để có thêm thông tin về kết quả mô hình được xây dựng.



Hình 3.7. Thiết lập bổ sung thông tin dữ liệu đầu ra

- Các tùy chọn kiểm thử: Tiến hành kiểm thử xây dựng mô hình hồi quy tuyến tính 03 lần, mỗi lần lựa chọn một trong 03 tùy chọn *Use training set, Supplied test set và Percentage split*. Trong đó:

+ *Use training set:* Sử dụng tập tin dữ liệu đầu vào.

+ *Supplied test set:* Chia tập tin dữ liệu đầu vào thành 02 phần: Phần 1 gồm dữ liệu 15 lần thu thập đầu tiên để huấn luyện (4742 dòng dữ liệu ≈ 93% dữ liệu), phần 2 gồm dữ liệu của lần thu thập cuối cùng (357 dòng dữ liệu ≈ 7% dữ liệu) để kiểm thử.

+ *Percentage split:* Chia tập tin dữ liệu đầu vào thành 2 phần: Phần 1 có 66% dữ liệu để huấn luyện, phần 2 có 34% dữ liệu còn lại để kiểm thử.

- Lựa chọn thuộc tính được dự đoán: *(Num) Gia*

Kết quả thu được:

	Kiểm thử ***Use training set***	**Kiểm thử** ***Supplied test set***	**Kiểm thử** ***Percentage split***
R^2 value	0.9655	0.9645	0.9655
Adjusted R^2	0.9651	0.96403	0.9651
F-statistic	2238.7534	1955.806	2238.7534
Correlation coefficient	0.9826	0.987	0.9828
Mean absolute error	837,146	879,776	838,952
Root mean squared error	1,138,025	1,158,829	1,125,783
Relative absolute error	20.6454 %	19.2251 %	20.9419 %
Root relative squared error	18.5656 %	16.1924 %	18.5856 %
Total Number of Instances	5,099	357	1,734

Bảng 3.3. Kết quả kiểm thử mô hình

Đánh giá mô hình: Kết quả kiểm thử đối với mô hình hồi quy tuyến tính được xây dựng bằng WEKA trên tập tin dữ liệu đầu vào là chấp nhận được. Cụ thể như sau:

- Hệ số xác định r^2 qua 03 lần kiểm thử đều đạt giá trị lớn hơn 0.96 cho thấy hơn 96% sự thay đổi của biến phụ thuộc “Gia” được giải thích bởi tập các biến độc lập được lựa chọn.

- Từ hệ số xác định r^2 tính được hệ số tương quan qua 03 lần kiểm thử đều đạt giá trị lớn hơn 0.98 cho thấy biến phụ thuộc "Gia" có mối tương quan chặt chẽ với tập các biến độc lập được lựa chọn.

Tuy nhiên, cần phải thêm các biến độc lập chưa được lựa chọn vào mô hình để khảo sát sự phù hợp của mô hình đã được xây dựng. Quá trình thêm các biến độc lập được thực hiện qua 05 lần, cụ thể như sau:

- Lần thứ 1: Thêm biến độc lập "HDD_DL", mô hình có 12 biến gồm *NgayTT, NhaCC, Gia, CPU_NhaCC, CPU_Dem, Ram_DL,* ***HDD_DL****, SSD_DL, Card_Loai, Card_DL, MH_Loai, MH_DPG_W.*

- Lần thứ 2: Thêm biến độc lập "Ram_Bus", mô hình có 13 biến gồm *NgayTT, NhaCC, Gia, CPU_NhaCC, CPU_Dem,* ***Ram_Bus****, Ram_DL,* ***HDD_DL****, SSD_DL, Card_Loai, Card_DL, MH_Loai, MH_DPG_W.*

- Lần thứ 3: Thêm biến độc lập "MH_KT", mô hình có 14 biến gồm *NgayTT, NhaCC, Gia, CPU_NhaCC, CPU_Dem,* ***Ram_Bus****, Ram_DL,* ***HDD_DL****, SSD_DL, Card_Loai, Card_DL, MH_Loai,* ***MH_KT****, MH_DPG_W.*

- Lần thứ 4: Thêm biến độc lập "Ram_Loai", mô hình có 15 biến gồm *NgayTT, NhaCC, Gia, CPU_NhaCC, CPU_Dem,* ***Ram_Bus****,* ***Ram_Loai****, Ram_DL,* ***HDD_DL****, SSD_DL, Card_Loai, Card_DL, MH_Loai,* ***MH_KT****, MH_DPG_W.*

- Lần thứ 5: Thêm biến độc lập "CPU_TocDo", mô hình có 16 biến gồm *NgayTT, NhaCC, Gia, CPU_NhaCC,* ***CPU_TocDo****, CPU_Dem,* ***Ram_Bus****,* ***Ram_Loai****, Ram_DL,* ***HDD_DL****, SSD_DL, Card_Loai, Card_DL, MH_Loai,* ***MH_KT****, MH_DPG_W.*

Kết quả thu được như sau:

	Mô hình ban đầu (11 biến)	Khảo sát lần thứ 1 (12 biến)	Khảo sát lần thứ 2 (13 biến)	Khảo sát lần thứ 3 (14 biến)	Khảo sát lần thứ 4 (15 biến)	Khảo sát lần thứ 5 (16 biến)
R^2 value	0.9655	0.966	0.9661	0.9662	0.9673	0.9673
Adjusted R^2	0.9651	0.96556	0.96567	0.96573	0.96679	0.96679
F-statistic	2238.7534	2306.5965	2241.6488	2211.3551	2034.0351	2034.0351
Correlation coefficient	0.9826	0.9828	0.9829	0.9829	0.9835	0.9835
Mean absolute error	837,146	822,784	819,228	819,648	809,546	809,546
Root mean squared error	1,138,025	1,130,546	1,128,590	1,127,447	1,109,028	1,109,028
Relative absolute error	20.6454 %	20.2912 %	20.2035 %	20.2138 %	19.9647 %	19.9647 %
Root relative squared error	18.5656 %	18.4436 %	18.4117 %	18.3931 %	18.0926 %	18.0926 %
Total Number of Instances	5,099	5,099	5,099	5,099	5,099	5,099

Bảng 3.4. Kết quả quá trình thêm biến độc lập vào mô hình

Qua 05 lần thêm biến độc lập vào mô hình, các hệ số của mô hình đều được cải thiện. Trong đó:

- Hệ số xác định điều chỉnh $\overline{r^2}$ (Adjusted R^2) đều tăng trong 04 lần khảo sát đầu từ 0.9651 lên 0.96679

- Sai số trung bình tuyệt đối (Mean absolute error) đều giảm qua 04 lần khảo sát đầu từ 837,416 xuống 809,546

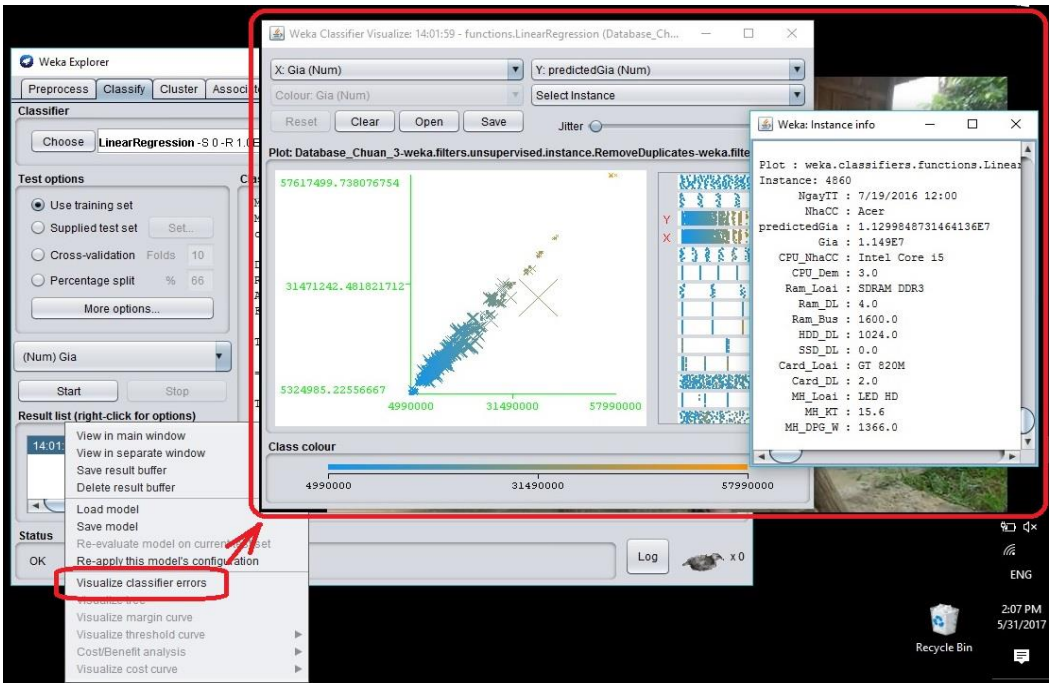
Tổng kết: Thêm các biến độc lập ***Ram_Bus, Ram_Loai, HDD_DL, MH_KT*** vào mô hình là cần thiết. Vậy, mô hình hồi quy tuyến tính được thiết lập với 15 biến, gồm:

- Biến phụ thuộc: *Gia*

- Biến độc lập: *NgayTT, NhaCC, CPU_NhaCC, CPU_Dem, Ram_Bus, Ram_Loai, Ram_DL, HDD_DL, SSD_DL, Card_Loai, Card_DL, MH_Loai, MH_KT, MH_DPG_W*

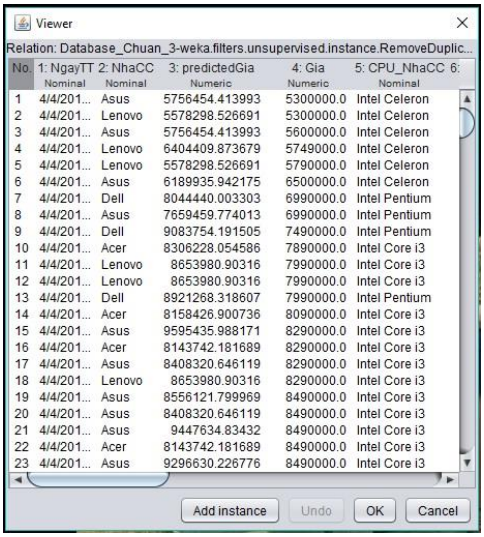
3.3. Tính toán thử nghiệm độ chính xác dự báo

Sau khi xây dựng mô hình hồi quy tuyến tính, WEKA hỗ trợ người sử dụng khảo sát độ chính xác dự báo thông qua việc so sánh giá sản phẩm trên thực tế và giá sản phẩm được dự đoán bằng mô hình hồi quy tuyến tính. Quá trình khảo sát được hỗ trợ thông qua tính năng đồ thị hóa sai số của mô hình hồi quy tuyến tính.

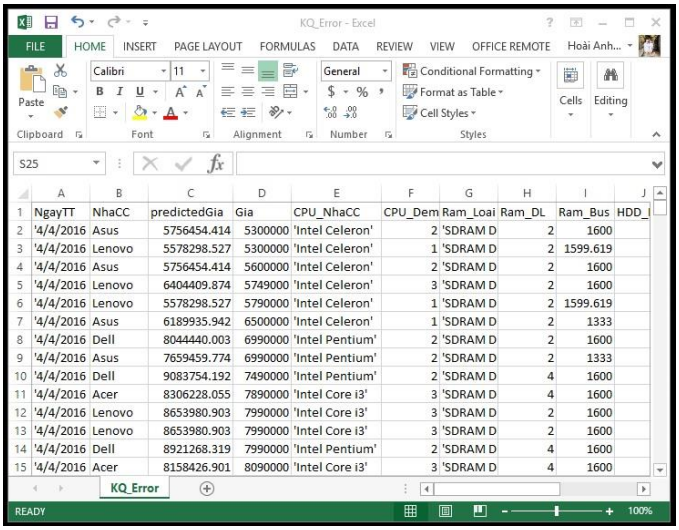


Hình 3.8. Mô hình hóa sai số của mô hình

Ngoài ra, WEKA còn hỗ trợ trích xuất kết quả dự báo của mô hình thành tập tin dữ liệu kết quả với định dạng "arff". Tuy nhiên, để thuận tiện cho việc đánh giá và khảo sát, người sử dụng có thể mở tập tin dữ liệu kết quả dưới định dạng "arff" bằng WEKA và lưu lại dưới định dạng "csv".



a. Tập tin "arff" mở bằng WEKA



b. Tập tin "csv" mở bằng Excel

Hình 3.9. Tập tin dữ liệu kết quả

Qua khảo sát tập tin dữ liệu kết quả, thu được một số thông tin sau:

- Sai số lớn nhất: 9.874.898đ của 01 dòng dữ liệu.

	NhaCC	predictedGia	Gia	Error	CPU_NhaCC
2406	Dell	28,115,101.01	37,990,000.00	9,874,898.99	'Intel Core i7'

- Sai số nhỏ nhất: ≈ 0đ với giá thực tế 02 dòng dữ liệu

	NhaCC	predictedGia	Gia	Error	CPU_NhaCC
4743	HP	42,990,000.00	42,990,000.00	5.99E-06	'Intel Core i7'
5015	Lenovo	15,989,999.99	15,990,000.00	0.005004	'Intel Core i5'
5099	HP	42,990,000.00	42,990,000.00	5.99E-06	'Intel Core i7'

- Sai số nhỏ hơn 500.000đ: có 2130/5099 dòng dữ liệu chiếm ≈ 41,77%.

- Sai số lớn hơn 2.000.000đ: có 344/5099 dòng dữ liệu chiếm ≈ 6,75%.

Chương 4
KẾT LUẬN

Việc ứng dụng phân tích dữ liệu vào công tác dự báo là hướng nghiên cứu có nhiều triển vọng, có thể áp dụng cho nhiều lĩnh vực trong đời sống xã hội. Nó có thể hỗ trợ, chúng ta hoạch đĩnh những chiến lược hay kế hoạch đầu tư phát triển hợp lý. Bên cạnh đó, với sự phát triển không ngừng của Ngành công nghệ thông tin, các công cụ hỗ trợ phân tích dữ liệu ngày càng phong phú và hỗ trợ đắc lực con người trong công tác dự báo.

Thông qua quá trình nghiên cứu về mô hình hồi quy tuyến tính và công cụ hỗ trợ WEKA, luận văn đã tiến hành giải quyết bài toán thực tế về công tác dự báo. Cụ thể, luận văn đã đi sâu nghiên cứu và làm rõ những nội dung sau:

- Đưa ra cơ sở lý thuyết về mô hình hồi quy tuyến tính ứng dụng trong việc phân tích dữ liệu để tiến hành dự báo.

- Tìm hiểu, nghiên cứu công cụ hỗ trợ WEKA trong việc xây dựng mô hình hồi quy tuyến tính để tiến hành dự báo.

- Sử dụng công cụ hỗ trợ WEKA để giải quyết bài toán thực tế về phân tích dữ liệu bán hàng và dự báo giá bán sản phẩm máy tính xách tay của Công ty cổ phần thương mại Nguyễn Kim.

Luận văn đã cho thấy sự hữu ích của việc phân tích dữ liệu để áp dụng, giải quyết các bài toán thực tế. Tuy nhiên, do một số nguyên nhân khách quan và chủ quan, luận văn vẫn còn tồn tại một số hạn chế sau:

- Dữ liệu thu thập của duy nhất một đơn vị dẫn đến công tác dự bảo mới chỉ dừng lại ở phạm vi cục bộ.

- Chưa tìm hiểu hết tất cả các tính năng của công cụ hỗ trợ WEKA để giải quyết các bài toàn thực tế.

Để khắc phục những hạn chế nêu trên, trong thời gian tới, luận văn sẽ tiếp tục nghiên cứu mở rộng phạm vi thu thập dữ liệu, tìm hiểu rõ công cụ hỗ trợ WEKA và các công cụ hỗ trợ khác để tiến hành dự báo có tính khái quát và chính xác hơn.

TÀI LIỆU THAM KHẢO

Tiếng Việt

1. Trần Ngọc Minh (2006), *Kinh tế lượng*, Học viện Công nghệ Bưu chính - Viễn thông, Hà Nội.
2. https://websrv1.ctu.edu.vn/coursewares/kinhte/phantichdulieu/chuong6.htm

Tiếng Anh

3. Ian H. Witten, Eibe Frank, Mark A. Hall (2011), Data Mining Practical Machine Learning Tools and Techniques
4. Ramu Ramanathan (2002), Introductory Econometrics with Applications
5. https://www.ibm.com/developerworks/vn/library/12/ba-predictive-analytics1/