

**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

**TÓM TẮT LUẬN VĂN THẠC SĨ  
NGÀNH CÔNG NGHỆ THÔNG TIN**

**Hà Nội - 2017**

## MỤC LỤC

MỞ ĐẦU .....	3
CHƯƠNG 1. ĐẶT VẤN ĐỀ .....	5
1.1. Bài toán phân tích dữ liệu .....	5
1.2. Lựa chọn miền ứng dụng .....	5
1.3. Phương pháp và công cụ .....	5
1.3.1. Lựa chọn phương pháp .....	5
1.3.2. Lựa chọn công cụ .....	6
CHƯƠNG 2. MÔ HÌNH HỒI QUY TUYẾN TÍNH VÀ CÔNG CỤ HỖ TRỢ WEKA .....	7
2.1. Mô hình hồi quy tuyến tính .....	7
2.1.1. Lý thuyết về mô hình hồi quy .....	7
2.1.2. Mô hình hồi quy tuyến tính .....	8
2.1.3. Phương pháp bình phương tối thiểu để ước lượng các tham số của mô hình hồi quy tuyến tính .....	9
2.1.4. Ứng dụng mô hình hồi quy tuyến tính vào phân tích dữ liệu .....	9
2.2. Công cụ hỗ trợ xây dựng mô hình hồi quy tuyến tính WEKA .....	12
2.2.1. Giới thiệu về WEKA .....	12
2.2.2. Các chức năng chính của WEKA .....	13
2.2.3. Xây dựng mô hình hồi quy tuyến tính với WEKA .....	13
CHƯƠNG 3. THỰC NGHIỆM VÀ ĐÁNH GIÁ KẾT QUẢ .....	16
3.1. Phát biểu bài toán thực tế .....	16
3.2. Tiến hành xây dựng mô hình .....	16
3.2.1. Thu thập dữ liệu .....	16
3.2.2. Tiền xử lý dữ liệu .....	17
3.2.3. Lựa chọn thuộc tính .....	19
3.2.4. Xây dựng và đánh giá mô hình .....	20
3.3. Tính toán thử nghiệm độ chính xác dự báo .....	22
CHƯƠNG 4. KẾT LUẬN .....	24
TÀI LIỆU THAM KHẢO .....	25

## MỞ ĐẦU

Ngày nay, 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.

Phân tích dữ liệu 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 tra, làm sạch, chuyển đổi và mô hình hóa 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 và 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. 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, làm sạch, chuyển đổi, mô hình hóa và phân tích dữ liệu nhằm mục đích tìm kiếm thông tin, cho thấy kết luận, hỗ trợ đưa ra quyết định. 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.

#### 1.2. Lựa chọn miền ứng dụng

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,  $\varepsilon$  là sai số ngẫu nhiên và  $f(X) = E(Y|X)$  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  $(X_i, Y_i)$ . Ở lần quan sát thứ  $i$ , biến  $X$  nhận giá trị  $X_i$ , biến  $Y$  nhận giá trị  $Y_i$  và sai số ngẫu nhiên là  $\varepsilon_i$ . Khi đó, mô hình (2.1) trở thành:  $Y_i = f(X_i) + \varepsilon_i = E(Y|X_i) + \varepsilon_i$  (2.2)

$\varepsilon_i$  là độ chênh lệch giữa giá trị quan sát  $Y_i$  của biến phụ thuộc  $Y$  với giá trị trung bình của  $Y$  khi biến độc lập  $X$  nhận giá trị  $X_i$ .  $\varepsilon$  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  $\varepsilon$  đại diện cho phần ảnh hưởng ấy.

Từ (2.2) ta có:  $\varepsilon_i = Y_i - f(X_i) \Rightarrow \varepsilon_i \rightarrow 0 \Leftrightarrow Y_i - f(X_i) \rightarrow 0$

Nếu  $\varepsilon_i$  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,  $\varepsilon$  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  $\varepsilon$  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

$$f(X_i) = E(Y|X_i) = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \dots + \beta_n X_{ni} \quad (2.3)$$

Trong đó:

- $X_i$  là giá trị của các biến độc lập  $X$  ở quan sát thứ  $i$
- $E(Y|X_i)$  là giá trị trung bình của biến phụ thuộc  $Y$  khi biến độc lập  $X$  nhận các giá trị  $X_i$  ở quan sát thứ  $i$
- $\beta_0, \beta_1, \beta_2, \dots, \beta_n$  là các tham số hồi quy. Tham số hồi quy  $\beta_0$  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  $\beta_j$  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  $j$   $X_{ji}$  tăng một đơn vị với điều kiện các biến độc lập khác không thay đổi giá trị.

Thật vậy: Giả sử  $X_{ji}^1 = X_{ji} + 1$

$$\text{Ta có } E(Y|X_i)^1 = E(Y|X_i) + \beta_j \Rightarrow \beta_j = E(Y|X_i)^1 - E(Y|X_i)$$

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

Như đã trình bày ở phần trước:

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

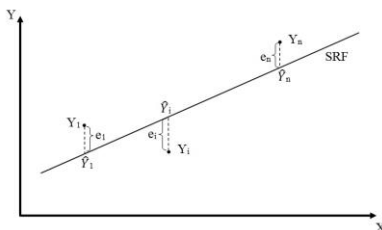
- Nếu  $f(X_i) = E(Y|X_i) = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \dots + \beta_n X_{ni}$  với  $n \geq 2$  thì mô hình được gọi là mô hình hồi quy tuyến tính bội.

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

$$\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 X_{1i} + \hat{\beta}_2 X_{2i} + \dots + \hat{\beta}_n X_{ni} \quad (2.3)$$

Trong đó:  $\hat{\beta}_i$  là ước lượng điểm của  $\beta_i$ ,  $\hat{Y}_i$  là ước lượng điểm của  $Y_i$

Khi đó, sai số  $e_i = Y_i - \hat{Y}_i$ . Minh họa bằng hình 2.1.



Hình 2.1. Sai số  $e_i$  giữa  $Y_i$  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  $e_i$  nhỏ nhất tức là  $\hat{Y}_i$  càng gần với giá trị  $Y_i$  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  $e_i$  giữa giá trị quan sát  $Y_i$  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:

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

và có một mẫu gồm  $n$  cặp quan sát  $(X_i, Y_i)$  với  $i = 1, 2, \dots, n$ .

Ở lần quan sát thứ  $i$ , ta có:

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

- Sai số:  $e_i = Y_i - \hat{Y}_i = Y_i - \hat{\beta}_0 - \hat{\beta}_1 X_i$

- Tổng bình phương các sai số  $e_i$ :

$$\sum_{i=1}^n e_i^2 = \sum_{i=1}^n (Y_i - \hat{\beta}_0 - \hat{\beta}_1 X_i)^2$$

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

$$f(\hat{\beta}_0, \hat{\beta}_1) = \sum_{i=1}^n (Y_i - \hat{\beta}_0 - \hat{\beta}_1 X_i)^2 \Rightarrow \min$$

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:

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

Giải hệ phương trình (2.4) ta được:

$$\hat{\beta}_0 = \bar{Y} - \hat{\beta}_1 \bar{X} \quad \text{và} \quad \hat{\beta}_1 = \frac{\sum_{i=1}^n (X_i - \bar{X})(Y_i - \bar{Y})}{\sum_{i=1}^n (X_i - \bar{X})^2} = \frac{\sum_{i=1}^n Y_i X_i - n \bar{X} \bar{Y}}{\sum_{i=1}^n X_i^2 - n (\bar{X})^2}$$

Trong đó:

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

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

### 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. 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:

$$f(X_i) = E(Y|X_i) = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \dots + \beta_k X_{ik}$$

Với n quan sát ta có:

$$\begin{aligned} Y_1 &= \beta_0 + \beta_1 X_{11} + \beta_2 X_{21} + \dots + \beta_k X_{k1} + e_1 \\ Y_2 &= \beta_0 + \beta_1 X_{12} + \beta_2 X_{22} + \dots + \beta_k X_{k2} + e_2 \\ &\dots\dots\dots \\ Y_n &= \beta_0 + \beta_1 X_{1n} + \beta_2 X_{2n} + \dots + \beta_k X_{kn} + e_n \end{aligned}$$

Ký hiệu:

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

Ta có:  $Y = X\beta + e$

Hàm hồi quy mẫu có dạng:

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

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:

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

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

Khi đó:

$$\sum_{i=1}^n e_i^2 = e^T e = Y^T Y - 2\hat{\beta}^T X^T Y + \hat{\beta}^T X^T X \hat{\beta}$$

Hệ phương trình có dạng:

$$\begin{aligned} \frac{\partial (e^T e)}{\partial \hat{\beta}} &= 0 \Rightarrow -2X^T Y + 2X^T X \hat{\beta} = 0 \\ &\Rightarrow \hat{\beta} = (X^T X)^{-1} X^T Y \end{aligned} \tag{2.10}$$

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  $X_k$ .

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

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  $\bar{r}^2$  để xác định có nên thêm 1 biến độc lập vào mới vào mô hình hay không. Thường thì giá trị của  $\bar{r}^2$  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  $\bar{r}^2$  tăng lên khi tăng biến đó.

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

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)
- $X_1$  là thu nhập của người tiêu dùng (triệu đồng/năm)
- $X_2$  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:

$$(X^T X)^{-1} = \frac{1}{1528} \begin{bmatrix} 39980 & -3816 & -3256 \\ -3816 & 376 & 300 \\ -3256 & 300 & 280 \end{bmatrix} \text{ và } X^T Y = \begin{bmatrix} 165 \\ 1029 \\ 813 \end{bmatrix}$$

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

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.

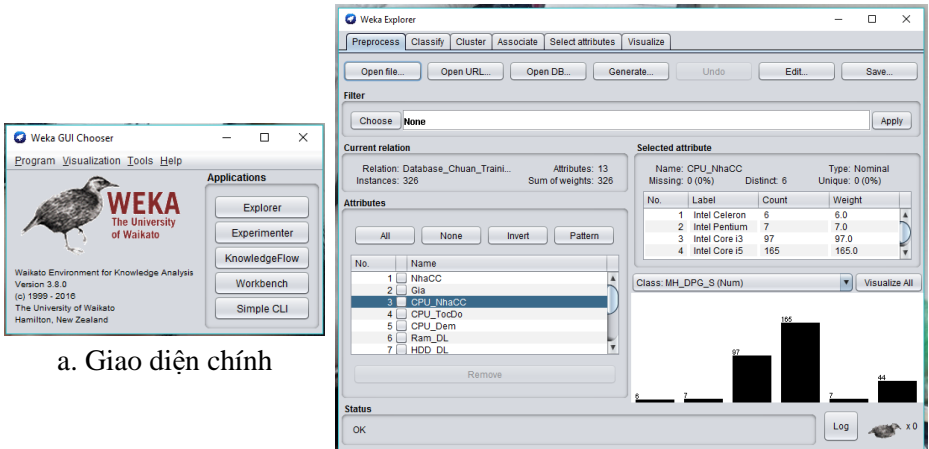
## 2.2. Công cụ hỗ trợ xây dựng mô hình hồi quy tuyến tính WEKA

### 2.2.1. Giới thiệu về WEKA

WEKA (Waikato Environment for Knowledge Analysis) 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.

WEKA được chia sẻ rộng rãi trên website <http://www.cs.waikato.ac.nz/~ml/weka/index.html>. Weka 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. Giao diện đồ họa người sử dụng GUI của WEKA được phát triển theo hướng trực quan và dễ sử dụng.



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

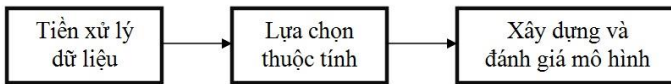
### 2.2.2. Các chức năng chính của WEKA

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.
- *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, người sử dụng cần lựa chọn *Explorer* với các tính năng *Preprocess*, *Classify* và *Select attributes*. Quá trình xây dựng mô hình được 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, người sử dụng cần chọn tính năng *Preprocess* của *Explorer*. Tính năng *Preprocess* cho phép người sử dụng 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, người dùng 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”.

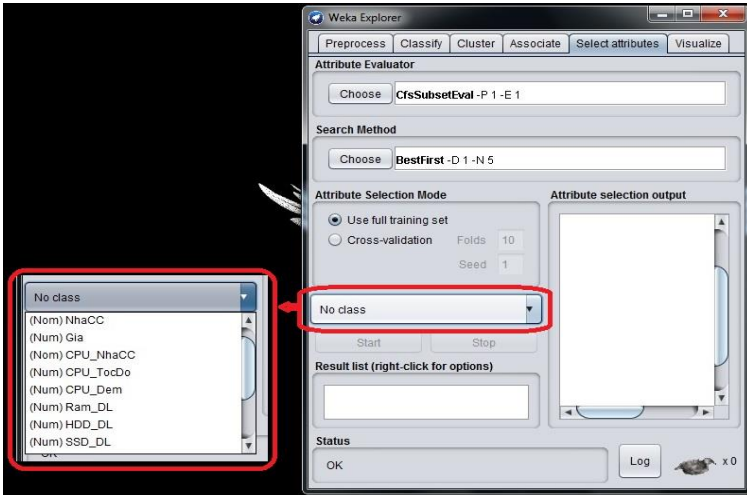
Sau khi tiền xử lý dữ liệu, người sử dụng cần lựa chọn các thuộc tính quan trọng, cần thiết để xây dựng mô hình hồi quy tuyến tính. Tập dữ liệu có rất nhiều thuộc tính để mô tả đầy đủ các khía cạnh của dữ liệu, tuy nhiên không

phải tất cả các thuộc tính để 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 quy để xây dựng mô hình.

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.

Người sử dụng 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

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

- 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, người sử dụng 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 (Classifier): 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 Hirotosugu 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*, *Supplied test set*, *Cross-validation*, *Percentage split*.

- 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

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: Thu thập dữ liệu, tiền xử lý dữ liệu, lựa chọn thuộc tính và xây dựng mô hình.

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,...

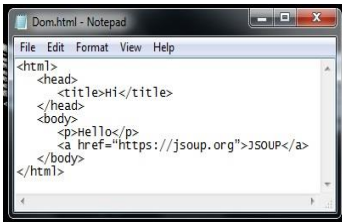
#### 3.2. Tiến hành xây dựng mô hình

##### 3.2.1. Thu thập dữ liệu

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 Central Processing Unit), 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>.

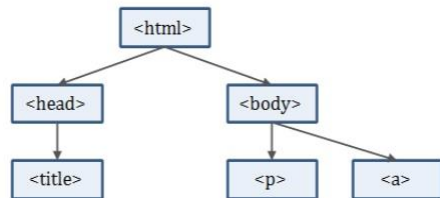
Jsoup phân tích tài liệu HTML thành mô hình DOM. Người sử dụng cần hiểu rõ bộ cục của tài liệu HTML để truy cập chính xác đến từng phần tử cụ thể của danh sách.



```

Dom.html - Notepad
File Edit Format View Help
<html>
<head>
<title>Hi</title>
</head>
<body>
<p>Hello</p>
<a href="https://jsoup.org">JSOUP</a>
</body>
</html>
  
```

a. Tập tin Dom.html



b. Mô hình DOM của tập tin Dom.html

Hình 3.1. 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	NgàyTT	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, SSD	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

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 xách 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” có 5.527 dòng dữ liệu với 327 dòng sản phẩm của 06 nhà cung cấp, 16 lần thu thập dữ liệu và 1.34MB dung lượng 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: Người sử dụng cần 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
NgàyTT	4/4/2016	NgàyTT	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 ổ trống
HDD_DL	500 GB, 1 TB, 1 TB + 128 GB	SDD	SDD hoặc ổ trống
Card_DL	2GB, Share	HDD_DL	0, 500, 1024
MH_KT	14.0 inch, 15.6 inch	SDD_DL	0, 128
MH_DPG	1366 x 768 Pixels	Card_DL	1, 2, 4, "0" với card share
		MH_KT	14.0, 15.6
		MH_DPG_W	1280, 1366, ...
		MH_DPG_H	768, 800,...
		MH_DPG_S	= MH_DPG_W * MH_DPG_H

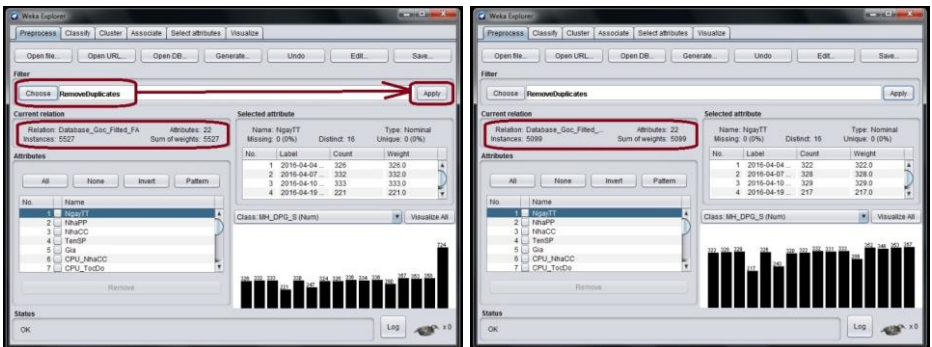
Tất cả các giá trị chưa xác định được biểu diễn bởi "?"

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

Kết thúc giai đoạn 1, tập tin dữ liệu có 862KB dung lượng.

- Giai đoạn 2: Người sử dụng cần mở tập tin dữ liệu bằng WEKA và tiến hành lọc dữ liệu để loại bỏ các bản ghi giống nhau và xử các giá trị thiếu (missing value).

Để loại bỏ các bản ghi giống nhau, người sử dụng cần sử dụng bộ lọc *Unsupervised/Instance/RemoveDuplicates*. Sau khi lọc dữ liệu lần thứ nhất, tập tin dữ liệu còn 5.099 dòng dữ liệu và 869KB dung lượng.



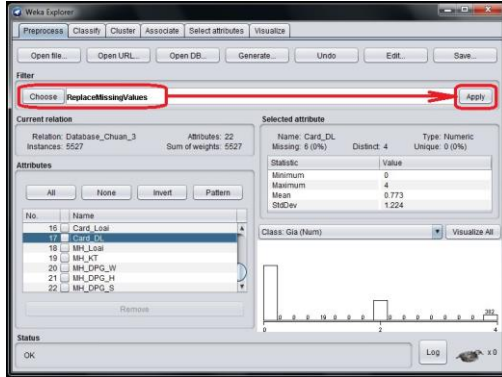
a. Lựa chọn bộ lọc dữ liệu

b. Dữ liệu đã qua xử lý của bộ lọc

Hình 3.2. 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 *Unsupervised/Attribute/RepalceMissingValue* để thay thế các giá trị thiếu dựa

trên những giá trị đã biết của tập dữ liệu huấn luyện. Sau khi lọc dữ liệu lần thứ hai, tập tin dữ liệu có 880KB dung lượng.



a. Lựa chọn bộ lọc dữ liệu

	J	K	L	M	N	O
1	Ram_DL	Ram_Bus	HDD	SSD	HDD_DL	SSD_DL
2	2	1600	SATA		500	0
3	2	?	SATA		500	0
4	2	2	1600	SATA		0
5	2	2	1600	SATA		0
6	2	?	SATA		500	0
7	2	2	1333	SATA		0
8	2	2	1600	SATA		0

	J	K	L	M	N	O
1	Ram_DL	Ram_Bus	HDD	SSD	HDD_DL	SSD_DL
2	2	1600	SATA	SSD	500	0
3	2	1600.089	SATA	SSD	500	0
4	2	2	1600	SATA	SSD	500
5	2	2	1600	SATA	SSD	500
6	2	1600.089	SATA	SSD	500	0
7	2	2	1333	SATA	SSD	500
8	2	2	1600	SATA	SSD	500

b. Dữ liệu đã qua xử lý của bộ lọc

Hình 3.3. 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.

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*.

### 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.

- 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  $\approx 93\%$  dữ liệu), phần 2 gồm dữ liệu của lần thu thập cuối cùng (357 dòng dữ liệu  $\approx 7\%$  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) Giá*

Kết quả thu được:

	Kiểm thử <i>Use training set</i>	Kiểm thử <i>Supplied test set</i>	Kiểm thử <i>Percentage split</i>
Degrees of freedom	5035	4676	5035
R <sup>2</sup> value	0.9655	0.9645	0.9655
Adjusted R <sup>2</sup>	0.9651	0.96403	0.9651
Correlation coefficient	0.9826	0.987	0.9828
Mean absolute error	837.146.4483	879.776.3325	838.952.2489
Root mean squared error	1.138.025.3298	1.158.829.3494	1.125.783.6627
Relative absolute error	20.6454 %	19.2251 %	20.9419 %
Root relative squared error	18.5656 %	16.1924 %	18.5856 %
Total Number of Instances	5099	357	1734

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”
- Lần thứ 2: Thêm biến độc lập “Ram\_Bus”
- Lần thứ 3: Thêm biến độc lập “MH\_KT”
- Lần thứ 4: Thêm biến độc lập “Ram\_Loai”
- Lần thứ 5: Thêm biến độc lập “CPU\_TocDo”

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 <sup>2</sup> value	0.9655	0.966	0.9661	0.9662	0.9673	0.9673
Adjusted R <sup>2</sup>	0.9651	0.9656	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<sup>2</sup>) đều tăng trong 04 lần khảo sát đầu từ 0.9651 lên 0.96679

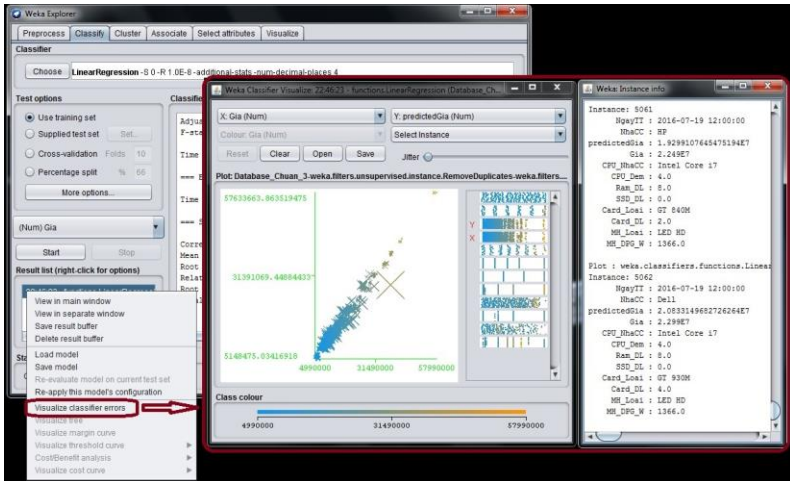
- Sai số trung bình tuyệt đối (Mean absolute error) đều giảm qua 04 lần khảo sát đầu từ 837,416 xuống 809,546

**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 3.5. 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”.

No	1. NgàyTT	2. NhaCC	3. predictedGia	4. Gia	5. CPU_NhaCC	6. 7.
	Nominal	Nominal	Numeric	Numeric	Nominal	CPU_D
1	2016-04-04	Asus	5818731.609781	5300000.0	Intel Celeron	2.0
2	2016-04-04	Lenovo	5632790.062074	5300000.0	Intel Celeron	1.0
3	2016-04-04	Asus	5818731.609781	6990000.0	Intel Celeron	2.0
4	2016-04-04	Lenovo	6431444.62192	5749000.0	Intel Celeron	3.0
5	2016-04-04	Lenovo	5632790.062074	5790000.0	Intel Celeron	1.0
6	2016-04-04	Asus	5419389.329858	6500000.0	Intel Celeron	1.0
7	2016-04-04	Dell	8312909.572931	6990000.0	Intel Pentium	2.0
8	2016-04-04	Asus	6431444.62192	5749000.0	Intel Celeron	3.0
9	2016-04-04	Dell	9354741.523058	7490000.0	Intel Pentium	2.0
10	2016-04-04	Acer	8269510.504871	7890000.0	Intel Core i3	3.0
11	2016-04-04	Lenovo	8546690.750189	7990000.0	Intel Core i3	3.0
12	2016-04-04	Lenovo	8546690.750189	7990000.0	Intel Core i3	3.0
13	2016-04-04	Dell	9354741.523058	7890000.0	Intel Pentium	2.0
14	2016-04-04	Acer	8115179.805481	8090000.0	Intel Core i3	3.0
15	2016-04-04	Asus	9576942.957697	8290000.0	Intel Core i3	3.0
16	2016-04-04	Acer	8269510.504871	8290000.0	Intel Core i3	3.0
17	2016-04-04	Asus	8380780.380878	8290000.0	Intel Core i3	3.0
18	2016-04-04	Lenovo	8546690.750189	8290000.0	Intel Core i3	3.0
19	2016-04-04	Asus	8535111.007469	8490000.0	Intel Core i3	3.0
20	2016-04-04	Asus	8380780.380878	8490000.0	Intel Core i3	3.0
21	2016-04-04	Asus	9422512.259336	8490000.0	Intel Core i3	3.0
22	2016-04-04	Acer	8269510.504871	8490000.0	Intel Core i3	3.0
23	2016-04-04	Asus	9375151.988201	8490000.0	Intel Core i3	3.0
24	2016-04-04	Acer	8269510.504871	8500000.0	Intel Core i3	3.0
25	2016-04-04	Acer	8269510.504871	8590000.0	Intel Core i3	3.0

a. Tập tin “arff” mở bằng WEKA

b. Tập tin “csv” mở bằng Excel

Hình 3.6. 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: 9.623.216đ của 01 dòng dữ liệu.

	<b>NhaCC</b>	<b>predictedGia</b>	<b>Gia</b>	<b>Error</b>	<b>CPU_NhaCC</b>
2406	Dell	28,366,783.67	37,990,000.00	-9,623,216.33	'Intel Core i7'

- Sai số nhỏ nhất: 0đ với giá thực tế 03 dòng dữ liệu

	<b>NhaCC</b>	<b>predictedGia</b>	<b>Gia</b>	<b>Error</b>	<b>CPU_NhaCC</b>
4743	HP	42,990,000.00	42,990,000.00	0.00	'Intel Core i7'
5015	Lenovo	15,990,000.00	15,990,000.00	0.00	'Intel Core i5'
5099	HP	42,990,000.00	42,990,000.00	0.00	'Intel Core i7'

- Sai số nhỏ hơn 500.000đ: có 2060/5099 dòng dữ liệu chiếm 40,4%.

- Sai số lớn hơn 2.000.000đ: có 374/5099 dòng dữ liệu chiếm 7,3%.

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