ĐẠI HỌC QUỐC GIA HÀ NỘI
TRƯỜNG ĐẠI HỌC CÔNG NGHỆ

NGUYỄN NGỌC TUÂN

ÁP DỤNG KỸ THUẬT KHAI PHÁ DỮ LIỆU DỰ BÁO THUÊ BAO RỜI MẠNG TRONG MẠNG DI ĐỘNG

LUẬN VĂN THẠC SĨ CÔNG NGHỆ THÔNG TIN

Hà Nội - 2016

ĐẠI HỌC QUỐC GIA HÀ NỘI
TRƯỜNG ĐẠI HỌC CÔNG NGHỆ

NGUYỄN NGỌC TUÂN

ÁP DỤNG KỸ THUẬT KHAI PHÁ DỮ LIỆU DỰ BÁO THUÊ BAO RỜI MẠNG TRONG MẠNG DI ĐỘNG

Ngành: Công nghệ Thông tin

Chuyên ngành: Hệ thống Thông tin

Mã số: 60480104

LUẬN VĂN THẠC SĨ CÔNG NGHỆ THÔNG TIN

NGƯỜI HƯỚNG DẪN KHOA HỌC: PGS.TS. NGUYỄN HÀ NAM

Hà Nội - 2016

MỤC LỤC

LỜI MỞ ĐẦU

Thuê bao rời mạng luôn là vấn đề "đau đầu" của các nhà mạng trong nước cũng như trên thế giới bởi lẽ khách hàng (thuê bao) chính là người mang lại doanh thu và duy trì hoạt động của các nhà mạng. Để duy trì và phát triển hoạt động kinh doanh của mình, các nhà mạng phải tìm mọi cách để phát triển thuê bao mới nhưng đồng thời cũng phải tìm cách để duy trì hoạt động của các thuê bao hiện hữu. Theo các nghiên cứu và thực tế triển khai của các nhà mạng cho thấy, tổng chi phí để phát triển một thuê bao mới cao hơn nhiều so với việc duy trì một thuê bao hiện hữu. Trong khi đó, doanh thu từ các thuê bao hiện hữu (đặc biệt là các thuê bao lâu năm) cao hơn nhiều so với doanh thu của các thuê bao mới (theo thống kê của MobiFone thuê bao lâu năm có doanh thu trung bình cao hơn 48% so thuê bao mới). Chính vì lý do trên, các nhà mạng trên thế giới không ngừng tìm kiếm các giải pháp và nghiên cứu phát triển các ứng dụng để xác định, dự đoán sớm thuê bao có khả năng rời mạng để có biện pháp kịp thời tác động nhằm duy trì thuê bao đó hoạt động..

Xuất phát từ yêu cầu đặt ra đối với đơn vị mình, tôi đã thực hiện đề tài luận văn **"ÁP DỤNG KỸ THUẬT KHAI PHÁ DỮ LIỆU DỰ BÁO THUÊ BAO RỜI MẠNG TRONG MẠNG DI ĐỘNG".** Luận văn đi sâu vào việc áp dụng các kỹ thuật khai phá dữ liệu từ hành vi sử dụng của các thuê bao di động từ đó dự báo thuê bao có khả năng rời mạng

Luận văn gồm có phần mở đầu, kết luận và 04 chương, cụ thể như sau:

Chương 1: Giới thiệu tổng quan về mạng di động và các vấn đề liên quan

Chương này trình bày về tổng quan mạng di động, thuê bao rời mạng. Phát biểu bài toán và các nghiên cứu liên quan

Chương 2: Khai phá dữ liệu và các kỹ thuật phân tích dự báo

Nghiên cứu các cơ sở lý thuyết KPDL, các vấn đề liên quan đến KPDL trong bài toán dự báo thuê bao rời mạng. Tìm hiểu các kỹ thuật khai phá dữ liệu sử dụng trong bài toán phân lớp, dự báo (chuẩn bị dữ liệu, lựa chọn thuộc tính, phân tích đặc trưng, cây quyết định, …) và áp dụng kỹ thuật KPDL trên ứng dụng WEKA.

Chương 3: Giải pháp phát hiện thuê bao di động có khả năng rời mạng

Các phương pháp để phát hiện thuê bao di động có khả năng rời mạng bao gồm các kỹ thuật dựa vào đặc trưng thuê bao, phương pháp ứng dụng kỹ thuật khai phá dữ liệu

Chương 4: Thực nghiệm và đánh giá kết quả

Nội dung chủ yếu là áp dụng các mô hình đã tìm hiểu vào việc dự báo thuê bao rời mạng. Trước tiên áp dụng các kỹ thuật cơ bản trên toàn bộ dữ liệu đã được chuẩn bị với mô hình phân tích đặc trưng, cây quyết định và đánh giá. Tiếp theo sử dụng các kỹ thuật lựa chọn thuộc tính để cải tiến tốc độ xử lý và đánh giá độ chính xác của mô hình dự báo. Cuối cùng tiến hành đánh giá kết quả dự báo của từng mô hình và đưa ra mô hình dự báo khuyến nghị để áp dụng vào bài toán thực tế

Chương 1. Giới thiệu tổng quan về mạng di động và các kiến thức cơ sở liên quan

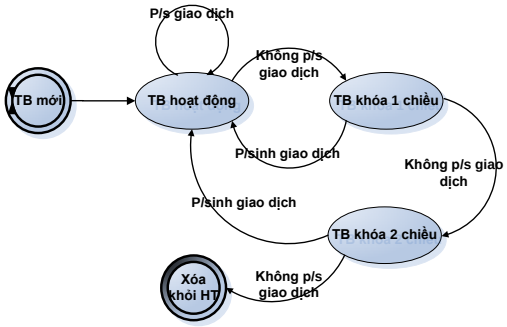
1.1. Giới thiệu về mạng di động

Hai mươi năm qua ghi nhận một thời gian dài phát triển sôi động, vượt bậc của ngành công nghiệp viễn thông nói chung và thông tin di động nói riêng. Được thành lập từ năm 1993, sau 22 năm phát triển, Tổng Công ty Viễn thông MobiFone (tiền thân là Công ty Thông tin di động), từ một doanh nghiệp cỡ vừa và nhỏ, đã phát triển trở thành một doanh nghiệp lớn, hàng đầu Việt Nam, có doanh thu lên tới hơn 40.000 tỷ, lợi nhuận hơn 6.000 tỷ VNĐ hàng năm và cung cấp dịch vụ viễn thông di động cho trên 20 triệu khách hàng sử dụng thường xuyên. Thương hiệu MobiFone cũng trở thành một trong những thương hiệu có uy tín hàng đầu. Bước sang những năm đầu của thập kỷ mới, cùng với sự bão hòa của số lượng thuê bao toàn thị trường, những thách thức đang dần trở nên thực tế hơn, đòi hỏi MobiFone phải có những sự theo dõi sát sao hơn với tính hình kinh doanh. Trong khi ở giai đoạn bùng nổ của thập niên đầu của thế kỷ 21, việc phát triển thuê bao là khá dễ dàng khi gần như cứ bổ sung thêm các trạm mới là có thể có thêm thuê bao, đến những năm đầu thập kỷ thứ hai, khi số lượng thuê bao trên toàn thị trường trở nên bão hòa, mạng lưới cũng đã phủ sóng gần như khắp quốc gia, mỗi khách hàng tiềm năng đều đã có 1 hay nhiều thuê bao, việc phát triển thuê bao mới trở nên khó khăn hơn bao giờ hết. Các nhà cung cấp dịch vụ viễn thông cũng cạnh tranh với nhau gãy gắt, quyết liệt, giành giật nhau từng thuê bao, bao gồm cả các thuê bao đang sử dụng dịch vụ của mạng đối thủ. Chính vì vậy, không chỉ việc phát triển thuê bao mà ngay cả việc giữ thuê bao cũng trở nên cấp bách.

1.2. Sơ lược tình hình nghiên cứu trên thế giới

1.3. Phát biểu bài toán

Nêu vòng đời phát triển thuê bao trả trước và mốc thời gian để .



Mục đích của nghiên cứu: **Phát hiện các thuê bao trả trước lâu năm có khả năng rời mạng để có thể tác động và duy trì thuê bao**

Mục tiêu của nghiên cứu: **Dự báo các thuê bao trả trước lâu năm có khả năng rời mạng khi vẫn còn trong giai đoạn 2 của vòng đời thuê bao tức là không phát sinh cước trong khoảng thời gian 1 tháng**.

1.4. Kết luận chương 1

Chương này giới thiệu về bài toán, những yêu cầu đặt ra cần giải quyết đối với bài toán đồng thời trình bày một số hướng nghiên cứu về thuê bao rời mạng trong mạng di động, hướng tiếp cận của luận văn.

Chương 2 Các kiến thức cơ sở liên quan

2.1. Khai phá dữ liệu

2.1.1. Khái niệm KPDL

KPDL là quá trình khảo sát và phân tích một khối lượng lớn các dữ liệu được lưu trữ trong các cơ sở dữ liệu, kho dữ liệu…để từ đó trích xuất ra các thông tin quan trọng, có giá trị tiềm ẩn bên trong. Do nhu cầu nghiệp vụ cần có cách nhìn thông tin trên quy mô toàn đơn vị.

2.1.2. Những nhóm bài toán của KPDL

2.1.3. Các bước xây dựng một giải pháp về KPDL

Có rất nhiều tác giả đưa ra các bước của một hệ thống KPDL, mọi sự phân chia chỉ mang tính chất tương đối và tư tưởng chủ đạo của nó là như sau:

- Bước 1: Xác định mục tiêu bài toán.
- Bước 2: Thu thập dữ liệu.
- Bước 3: Làm sạch dữ liệu và chuyển đổi dữ liệu.
- Bước 4: Xây dựng mô hình.
- Bước 5: Đánh giá mô hình hay đánh giá mẫu.
- Bước 6: Báo cáo.
- Bước 7: Dự đoán.
- Bước 8: Tích hợp vào ứng dụng
- Bước 9: Quản lý mô hình

2.1.4. Ứng dụng KPDL trong viễn thông

2.2. Một số kỹ thuật KPDL trong phân lớp, dự báo

2.2.1. Cây quyết định

Cây quyết định là một cấu trúc biểu diễn dưới dạng cây. Trong đó, mỗi nút trong (internal node) biểu diễn một thuộc tính, mỗi nhánh (branch) biểu diễn giá trị có thể có của thuộc tính, mỗi lá (leaf node) biểu diễn các lớp quyết định và đỉnh trên cùng của cây gọi là gốc (root).

2.2.2. Phân lớp Naïve Bayes

Định lý Bayes cho phép tính xác suất xảy ra của một sự kiện ngẫu nhiên A khi biết sự kiện liên quan B đã xảy ra. Xác suất này được ký hiệu là P(A|B), và đọc là "xác suất của A nếu có B". Đại lượng này được gọi xác suất có điều kiện vì nó được rút ra từ giá trị được cho của B hoặc phụ thuộc vào giá trị đó.

Phương pháp Naive Bayes phù hợp các bài toán có yêu cầu về chi phí xuất hiện của các giá trị thuộc tính.

- Thiết kế hệ thống phân lớp thường dễ dàng hơn so với các phương pháp khác.

- Các thuộc tính trong tập mẫu học phải độc lập với điều kiện.

- Độ chính xác thuật toán phân lớp phụ thuộc nhiều vào tập dữ liệu học ban đầu.

2.2.3. Mạng nơ ron nhân tạo

Neural nhân tạo là sự mô phỏng đơn giản của neural sinh học. Mỗi neural nhân tạo thực hiện hai chức năng: chức năng tổng hợp đầu vào và chức năng tạo đầu ra. Mỗi neural có một giá trị ngưỡng, chức năng đầu vào chính là tổng có trọng số các tín hiệu vào kết hợp với ngưỡng để tạo ra tín hiều đầu vào. Chức năng tạo đầu ra được thực hiện bằng hàm truyền đạt. Hàm này sẽ nhận tín hiệu đầu vào và tạo tín hiệu đầu ra của neural.

Có ba loại neural trong một mạng Neural được tạo ra với thuật toán Microsoft Neural Network:

Neural đầu vào: Neural đầu vào cung cấp các giá trị thuộc tính đầu vào cho các mô hình khai thác dữ liệu. Đối với thuộc tính đầu vào rời rạc, một neural đầu vào thường đại diện cho một trạng thái đơn nhất từ các thuộc tính đầu vào. Một thuộc tính đầu vào liên tục tạo ra hai neural đầu vào: một neural cho một trạng thái bị thiếu, một neural cho giá trị của chính thuộc tính liên tục đó. Neural đầu vào cung cấp đầu vào cho một hoặc nhiều neural ẩn.

Neural ẩn: neural ẩn nhận đầu vào từ các neural đầu vào và cung cấp đầu ra cho các neural đầu ra.

Neural đầu ra: Neural đầu ra đại diện cho các giá trị của thuộc tính dự đoán của mô hình KPDL. Đối với các thuộc tính đầu ra là rời rạc, một neural đầu ra đại diện duy nhất cho một trạng thái dự đoán của thuộc tính dự đoán, bao gồm cả giá trị thiếu. Nếu các thuộc tính dự đoán liên tục tạo ra hai neural đầu ra: một neural cho một trạng thái bị thiếu, một neural cho chính các giá trị của thuộc tính liên tục đó.

Mỗi đầu vào có một giá trị được gán cho nó có trọng số là w, mô tả sự liên quan giữa đầu vào đến các neural ẩn hoặc các neural đầu ra.

2.2.4. Luật kết hợp

Khai phá luật kết hợp: Là tìm các mẫu phổ biến, sự kết hợp, sự tương quan, hay các cấu trúc nhân quả giữa các tập đối tượng trong các cơ sở dữ liệu giao tác, cơ sở dữ liệu quan hệ, và những kho thông tin khác.

Các ứng dụng: Luật kết hợp có ứng dụng trong nhiều lĩnh vực khác nhau của đời sống như: khoa học, hoạt động kinh doanh, tiếp thị, thương mại, phân tích thị trường chứng khoán, tài chính và đầu tư,...

2.3. Giới thiệu về công cụ weka

2.4. Kết luận chương 2

Chương 2 trình bày một số khái niệm cơ bản và các kiến thức có liên quan về các thuật toán thường được áp dụng trong KPDL ở dạng phân lớp, dự báo đó là: cây quyết định, mạng nơron nhân tạo, phân lớp Naïve Bayes và luật kết hợp.

Chương 3 Giải pháp phát hiện thuê bao di động có khả năng rời mạng

3.1. Giải pháp chung:

Đối với các bài toán về thuê bao rời mạng nói chung thì đều sử dụng mô hình như sau:

Hình 3-1 Mô hình xử lý dữ liệu chung

Trong tất cả các giải pháp trình bày tiếp theo đều sử dụng chung mô hình này và chỉ khác nhau trong khối mô hình phân tích.

Giải thích các chức năng của các khối như sau:

- Thu thập dữ liệu: Có nhiệm vụ thu thập dữ liệu từ các nguồn liên quan đến bài toán, đối với bài toán này thì dữ liệu được thu thập từ các nguồn sẵn có của MobiFone như dữ liệu CDR từ tổng đài, dữ liệu thông tin khách hàng, dữ liệu lịch sử mua hàng và khuyến mại, dữ liệu lịch sử khiếu nại … Các dữ liệu này đều được các hệ thống tác nghiệp của MobiFone xử lý và đã được lưu trữ trong CSDL.
- Xử lý và trích xuất dữ liệu: Bước này thực hiện tiền xử lý và trích xuất dữ liệu để đưa vào mô hình phân tích. Trong khối dữ liệu khổng lồ của mạng di động thì tôi đã dựa vào kinh nghiệm làm việc gần 10 năm tại mạng di động MobiFone về các công việc liên quan đến xử lý, quản lý dữ liệu khách hàng cũng như tham khảo các chuyên gia trong lĩnh vực chăm sóc khách hàng để chọn lọc ra những dữ liệu có liên quan nhất với bài toán.
- Mô hình phân tích: Bước này sẽ thực hiện mô hình hóa các dữ liệu đã được xử lý từ bước trước để xây dựng các mô hình dữ liệu.
- Đánh giá: Đánh giá kết quả thực hiện

3.2. Giải pháp hiện tại của mạng MobiFone

Giải pháp hiện tại của mạng MobiFone như sau:

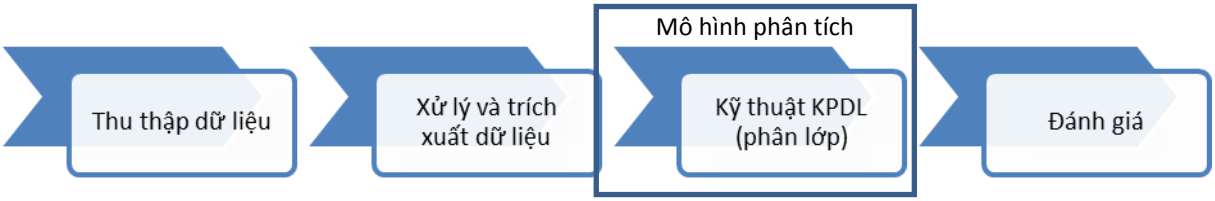
Hình 3-2 Mô hình xử lý hiện tại

Trong giải pháp này thì các chuyên gia MobiFone trong lĩnh vực chăm sóc khác hàng đã sử dụng phương pháp phân tích đặc trưng dữ liệu từ đó tìm ra đặc trưng của các thuê bao có khả năng rời mạng để áp dụng dự báo. Phương pháp này sẽ sử dụng các tập luật để phân lớp dữ liệu. Phương pháp này có ưu điểm là đơn giản rất dễ thực hiện để phân lớp tuy nhiên để xác định chính xác các tập luật đòi hỏi phải rất hiểu được dữ liệu nên không thể tự động hóa được một khi dữ liệu thay đổi đòi hỏi phải có chuyên gia phân

tích lại dữ liệu nên sẽ tốn rất nhiều thời gian để đưa ra được tập luật chính xác. Hơn nữa đối với những loại dữ liệu phức tạp thì phương pháp này cho độ chính xác không cao.

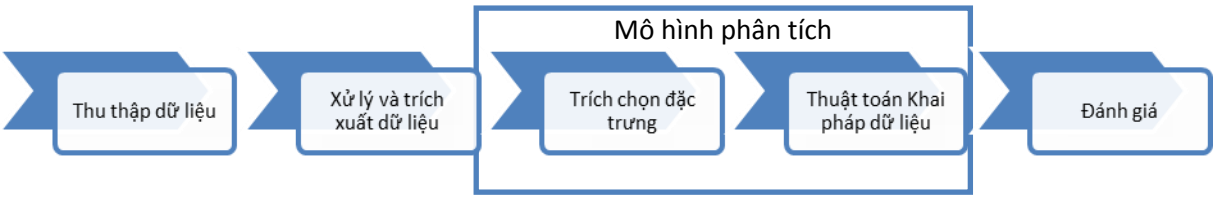
3.3. Giải pháp đề xuất

3.3.1. Giải pháp đề xuất dùng kỹ thuật khai phá dữ liệu



Giải pháp này đề xuất sử dụng để giải quyết bài toán. Trong bước mô hình hóa sẽ sử dụng các thuật toán KPDL cụ thể các thuật toán phân lớp để ứng dụng dự báo thuê bao rời mạng. Phương pháp này có ưu điểm là có thể tự động hóa được tuy nhiên thời gian xử lý dữ liệu thường chậm hơn phương pháp dựa vào đặc trưng.

3.3.2. Giải pháp đề xuất sau khi cải tiến



Giải pháp này là cải tiến của giải pháp dùng thuật toán khai phá dữ liệu. Để làm giảm thời gian xử lý dữ liệu đối thì trước khi đưa số liệu vào xử lý bằng thuật toán khai phá dữ liệu ta sẽ sử dụng các thuật toán trích chọn đặc trưng để tìm ra những thuộc tính phù hợp nhất với các thuật toán. Phương pháp này thời gian xử lý dữ liệu sẽ nhanh hơn giải pháp đề cập mục 3.4 tuy nhiên kết quả dự đoán sẽ không chính xác bằng. Do dữ liệu mạng di động thời rất lớn mà bài toán này thì đòi hỏi thời gian xử lý nhanh do vậy ưu tiên về thời gian xử lý hơn.

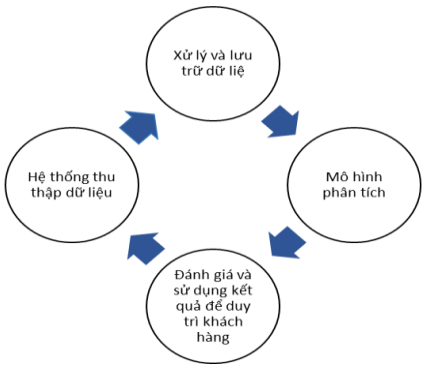
Các thuật toán trích chọn thuộc tính:

- Trích chọn theo Correlation-based: Đánh giá giá trị của một tập hợp các thuộc tính bằng cách xem xét các khả năng tiên đoán riêng rẽ của mỗi thuộc tính cùng với mức độ dư thừa giữa chúng. (M. A. Hall (1998). Correlation-based Feature Subset Selection for Machine Learning.) (Trích chọn dựa vào tương quan)
- Trích chọn theo độ đo GainRatio: Đánh giá giá trị của một thuộc tính bằng cách đo GainRatio của thuộc tính trong quá trình phân lớp:
  GainR(Class, Attribute) = (H(Class) - H(Class | Attribute)) / H(Attribute).
- Trích chọn theo độ đo InfoGain: Đánh giá giá trị của một thuộc tính bằng cách đo InfoGain liên quan đến các phân lớp.

  InfoGain(Class,Attribute) = H(Class) - H(Class | Attribute).

Trích chọn theo phương pháp PCA (principal components analysis – Phân tích thành phần chính)

3.4. Mô hình đề xuất áp dụng thực tế



Trong đó Hệ thống thu thập dữ liệu là hệ thống có chức năng lấy dữ liệu CDR, thông tin khách hàng, log khiếu nại của khách hàng

Hệ thống xử lý và lưu trữ dữ liệu: Đây chính là hệ thống Kho dữ liệu của MobiFone có chức năng xử lý dữ liệu đã thu thập và lưu trữ dưới dạng database.

Mô hình phân tích: Sử dụng kết nối từ WEKA đến DB của hệ thống Kho dữ liệu để lấy thông tin phân tích và đầu ra là kết quả dự báo thuê bao rời mạng

Đánh giá và sử dụng: Đánh giá kết quả dự đoán và đưa số liệu dự đoán để sử dụng thực tế. Qua việc phản hồi kết quả đánh giá sẽ tác động ngược lại quá trình thực hiện ban đầu để tối ưu mô hình dự đoán. Do đặc thù của kinh doanh luôn thay đổi ảnh hưởng hành vi sử dụng của khách hàng do không có mô hình dự đoán nào luôn đúng. Trong quá trình thực hiện đánh giá nhằm cải tiến mô hình để phù hợp nhất với tình hình số liệu thực tế

Chương 4 Thực nghiệm và đánh giá kết quả

Theo các phương pháp dự báo đã trình bày trong chương 3, áp dụng vào số liệu thực tế bài toán: dự báo thuê bao di động rời mạng, luận văn trình bày các kết quả thực nghiệm đạt được trên các phương pháp đã nêu.

4.1 Chuẩn bị dữ liệu

Yêu cầu thu thập dữ liệu: Dữ liệu được sử dụng trong luận văn này là dữ liệu thực tế của mạng di động MobiFone đã được biến đổi khi công bố trong luận văn nhằm đảm bảo tính bảo mật của dữ liệu.
Đối tượng: Thuê bao MobiFone lâu năm (kích hoạt/hòa mạng từ 2 năm trở lên) không phát sinh cước trong tháng 8/2015.
Khoảng thời gian thu thập dữ liệu: thu thập dữ liệu lịch sử của những thuê bao này trong vòng 3 tháng, từ tháng 5 đến tháng 7/2015.
Chi tiết dữ liệu cần thu thập: 86 trường dữ liệu chi tiết, gồm: (chi tiết tại Phụ lục đính kèm):

- Gói cước
- Thời gian kích hoạt
- Loại khách hàng (cá nhân/doanh nghiệp)
- Trạng thái thuê bao tại thời điểm xuất dữ liệu
- Doanh thu tài khoản chính và tài khoản khuyến mại (có chia dịch vụ thoại, sms, data, khác) trong 3 tháng gần nhất (tháng 5, 6 và 7)
- Số lượng, sản lượng cuộc gọi đến/đi, sms đến/đi trong 3 tháng gần nhất (tháng 5, 6 và 7)
- Số ngày cập nhật VLR, số ngày phát sinh cước trong 3 tháng gần nhất (tháng 5, 6 và 7).
- Số tiền nạp/số lần nạp thẻ trong 3 tháng gần nhất.
- Số chương trình khuyến mại tham gia trong 3 tháng gần nhất.
- Và 1 số trường thông tin khác.

Dữ liệu sau khi làm sạch và lấy mẫu gồm gần 2 triệu bản ghi với 86 trường dữ liệu chi tiết.

- Tiền xử lý dữ liệu: Dữ liệu đã được xử lý bởi các hệ thống của MobiFone và đã chuẩn hóa các thông tin chi tiết như sau:

1. Loại bỏ những bản ghi có thông tin thiếu.
2. Loại bỏ những cuộc gọi bất thường
3. Chuẩn hóa các trường dữ liệu theo định dạng quy định

4.2 Mô tả dữ liệu thực nghiệm

Bộ dữ liệu tháng 8/2015, toàn mạng MobiFone có 1.622.229 thuê bao kích hoạt từ 2 năm trở lên và không phát sinh cước. Trong đó, thuê bao khách hàng cá nhân chiếm đa số với 98% (tương đương 1.610.136 thuê bao) và thuê bao khách hàng doanh nghiệp chiến 2% (12.093 thuê bao).

Nhằm phân tích sâu hơn và tìm ra những đặc điểm đặc trưng của thuê bao trước khi rời mạng, nhóm chuyên gia MobiFone đã thực hiện phân tập thuê bao không phát sinh cước trong tháng 8/2015 thành 5 nhóm để phân tích, cụ thể:

- Nhóm I: Thuê bao trả sau.
- Nhóm II: Thuê bao trả trước có tiêu dùng tài khoản chính (TKC) nhưng không đạt 3k3d_vlr trong tháng 7/2015.
- Nhóm III: Thuê bao trả trước chỉ tiêu dùng tài khoản khuyến mại (TKKM) trong tháng 7/2015.
- Nhóm IV: Thuê bao trả trước không phát sinh cước trong tháng 7/2015.
- Nhóm V: Thuê bao trả trước có đạt chuẩn 3k3d_vlr trong tháng 7/2015.

Loại TB	Thuê bao trả sau	Thuê bao trả trước				Tổng
		Không đạt 3k3d_vlr trong tháng trước			Đạt 3k3d_vlr trong tháng trước	
Nhóm	Nhóm I	Nhóm II	Nhóm III	Nhóm IV	Nhóm V	
Thuê bao Fastconnect	17.244	3.096	2.690	14.578	2.154	39.762
Thuê bao thường	34.821	161.527	219.993	1.067.414	105.703	1.589.458
Số lượng TB	52.065	164,623	222.683	1.081.992	107.857	1.629.220
Tỷ lệ	3.20%	10.10%	13.67%	66.41%	6.62%	100.00%

Bảng 4-1 Nhóm tiêu dùng

Tuy nhiên trong mục đích nghiên cứu này ta chỉ tập trung vào thuê bao trả trước đồng thời qua phân tích dữ liệu thì nhóm 2 sẽ gồm 2 nhóm có đặc trưng khác nhau do đó ta chia tiếp nhóm II thành 2 nhóm nhỏ (nhóm 1 – tiêu dùng TKC dưới 3.000đ và nhóm 2 – tiêu dùng TKC từ 3.000đ trở lên)

Từ bộ dữ liệu thực tế trên để ta lấy mẫu dữ liệu để xử lý như sau:

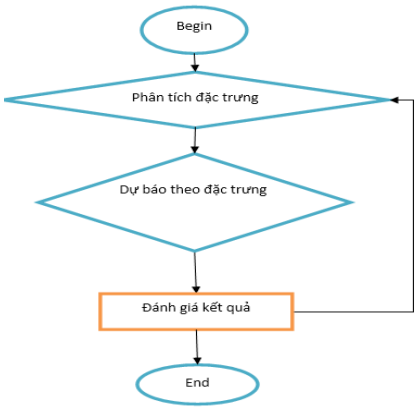
Nhóm	Rời mạng		Không rời mạng	Tổng
	Lấy mẫu chạy	Số liệu đang có		
Nhóm 1	20,000	93,962	136,000	156,000
Nhóm 2	10,000	67,565	68,000	78,000
Nhóm 3 (nhóm III)	20,002	19,993	136,000	156,002
Nhóm 4 (nhóm IV)	25,001	1,061,729	170,000	195,001
Nhóm 5 (nhóm V)	20,000	105,703	136,000	156,000
Tổng	**95,003**	**1,548,952**	**646,000**	**741,003**

Trong đó số liệu không rời mạng là nhóm thuê bao có phát sinh cước trong tháng 8

4.3 Kết quả thực nghiệm theo phương pháp hiện tại

Kết quả phân tích các nhóm dữ liệu đã các chuyên gia MobiFone tìm ra được những đặc điểm của thuê bao trước khi rời mạng. Sau khi áp dụng các đặc trưng này vào bộ dữ liệu kiểm nghiệm để đánh giá thì kết quả như sau. Như đã đề cập ở chương 1, trong phạm vi đề tài này chỉ tập trung vào nhóm thuê bao trả trước mà không tập trung vào nhóm thuê bao trả sau. Do vậy, trong phần này, chỉ tập trung vào việc sử dụng các đặc điểm đã phân tích để dự báo cho các thuê bao trả trước

Cách thức thực hiện như sau:



- Bước 1: Sử dụng tập dữ liệu thuê bao rời mạng trong tháng 8 để thực hiện phân tích đặc trưng của các thuê bao rời mạng theo từng nhóm thuê bao đã mô tả tại mục 4.2.
- Bước 2: Dự báo thuê bao rời mạng trên 1 tập dữ liệu lấy mẫu để đánh giá kết quả dự báo
- Bước 3: Đánh giá kết quả theo phương pháp đánh giá đã nêu tại chương 2, để lựa chọn các đặc trưng tối ưu. Sau khi đánh giá quay lại bước phân tích nếu chưa được kết quả tối ưu.

Việc thực hiện các bước trên nhiều lần cùng với kiến thức chuyên gia về lĩnh vực CSKH của mạng MobiFone đã đưa ra kết quả tối ưu như sau (chi tiết các đặc trưng được nêu tại lục lục 3):

Nhóm 1:

- Kết quả phân lớp

		Kết quả dự đoán		
		Rời mạng	*Không rời mạng*	***Tổng***
Thực tế	*Rời mạng*	16.706	3.294	20.000
	Không rời mạng	8.461	127.539	136.000
	Tổng	25.167	130.833	156.000

Nhóm 2:

- Kết quả phân lớp (dự báo):

		Kết quả dự đoán		
		Rời mạng	*Không rời mạng*	***Tổng***
Thực tế	*Rời mạng*	9.530	470	10.000
	Không rời mạng	2.375	65.625	68.000
	Tổng	11.905	66.095	78.000

Nhóm 3:

- Kết quả phân lớp (dự báo):

		Kết quả dự đoán		
		Rời mạng	*Không rời mạng*	***Tổng***
Thực tế	*Rời mạng*	19,047	955	20,002
	Không rời mạng	913	135,087	136,000
	Tổng	19,960	136,042	156,002

Nhóm 4:

- Kết quả phân lớp (dự báo):

		Kết quả dự đoán		
		Rời mạng	*Không rời mạng*	***Tổng***
Thực tế	*Rời mạng*	22.724	3.953	26.677
	Không rời mạng	2.218	167.782	170.000
	Tổng	24.942	171.735	196.677

Nhóm 5:

- Kết quả phân lớp (dự báo):

		Kết quả dự đoán		
		Rời mạng	*Không rời mạng*	*Tổng*
Thực tế	*Rời mạng*	9.861	10.139	20.000
	Không rời mạng	23.559	112.441	136.000
	Tổng	33.420	122.580	156.000

Tổng hợp kết quả dự báo dữ liệu thực dựa vào các đặc điểm đặc trưng của thuê bao trước khi rời mạng như sau:

Tên độ đo	Nhóm 1	Nhóm 2	Nhóm 3	Nhóm 4	Nhóm 5
Accuracy	92,5%	96,4%	98,7%	96,9%	78,4%
Error_Rate	7,5%	3,6%	1,3%	3,1%	21,6%
Recall	83,5%	80,1%	95,1%	91,1%	49,3%
Precision	66,4%	95,3%	95,1%	85,2%	29,5%

Bảng 4-2 Tổng hợp kết quả của phương pháp hiện tại

4.4 Kết quả thực nghiệm dựa trên khai phá dữ liệu

4.4.1 Kết quả thực nghiệm dựa trên giải pháp khai phá dữ liệu

Trong phần này, ta sẽ sử dụng tập dữ liệu thuê bao chưa được phân lớp (chưa được xác định là rời mạng hay không) làm đầu vào cho mô hình. Sử dụng ứng dụng WEKA và thuật toán cây quyết định để thực nghiệm. Lý do sử dụng thuật toán cây quyết định vì đây thuật toán có thời gian chạy dữ liệu nhanh và đối với dữ liệu này cho kết quả cao hơn so với phương pháp đang thực hiện tại MobiFone nên phù hợp với yêu cầu của MobiFone .

Tập dữ liệu chạy được trích xuất từ bộ dữ liệu thực tế như mô tả tại mục 4.1. Trong tập dữ liệu trên ta chia tập train (luyện tập) và tập test (kiểm chứng) theo nguyên tắc như sau:

Tập luyện tập: Là tập chứa 66% số liệu của bộ dữ liệu đầu vào.

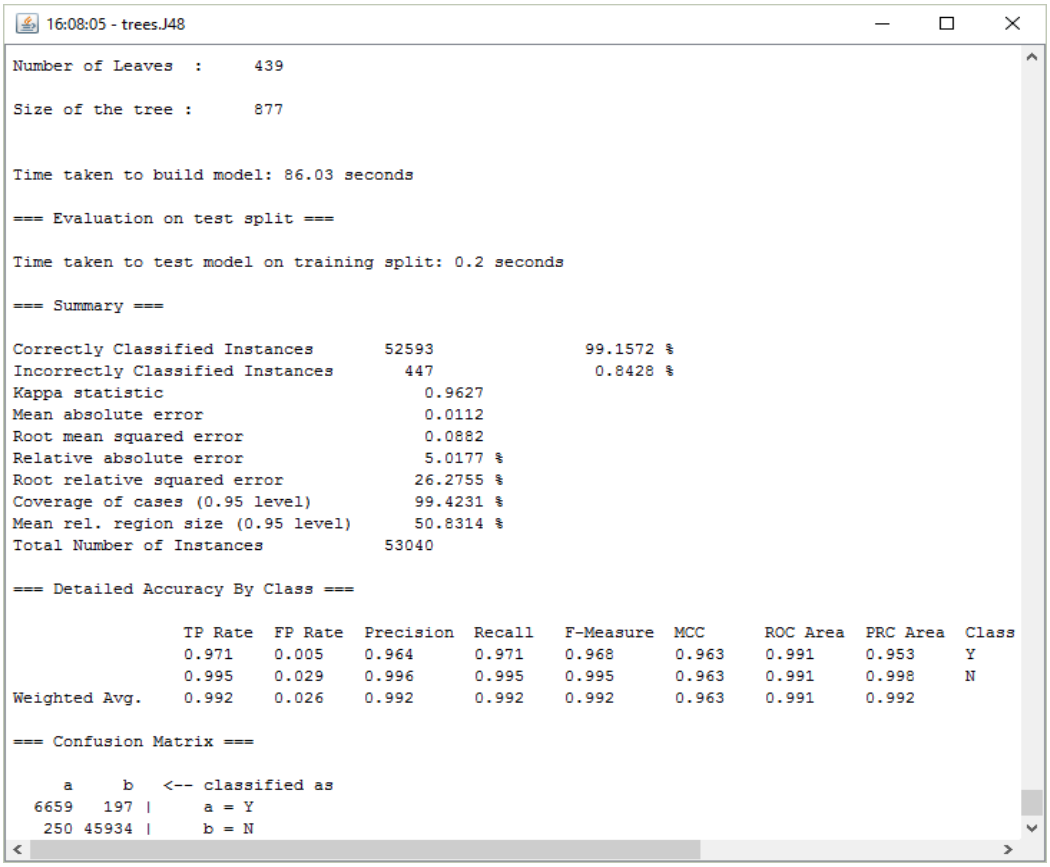
Tập kiểm chứng: Là tập chứa 34% số liệu của bộ dữ liệu đầu vào còn lại.

Thuật toán phân lớp: Cây quyết định theo thuật toán C4.5 được chương trình hóa trong mục phân lớp trên công cụ Weka là J48.

Kết quả chi tiết như sau:

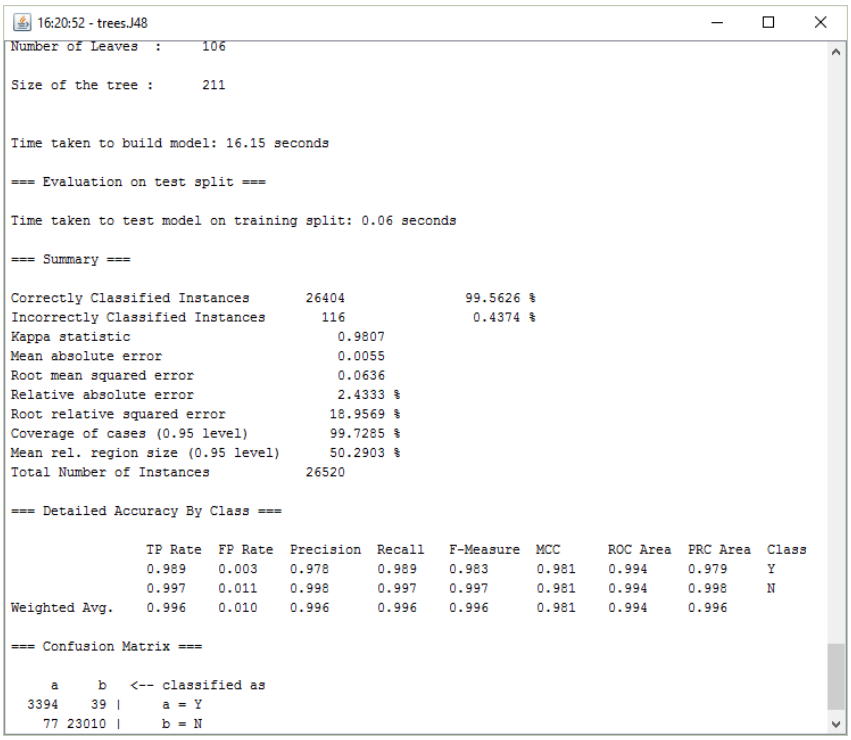
a. Nhóm 1

Tập dữ liệu chạy bao gồm 156.000 TB, trong đó: 20.000 TB rời mạng, 136.000 TB không rời mạng



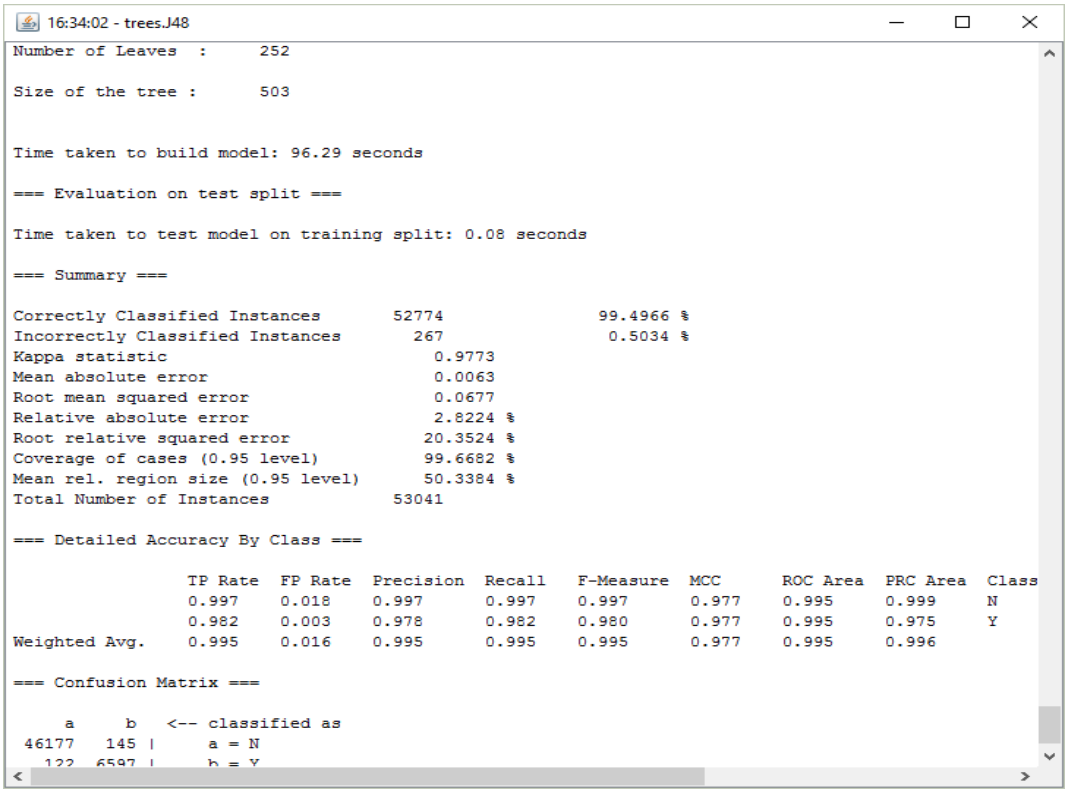
b. Nhóm 2:

Tập dữ liệu chạy bao gồm 156.000 TB, trong đó: 20.000 TB rời mạng, 136.000 TB không rời mạng



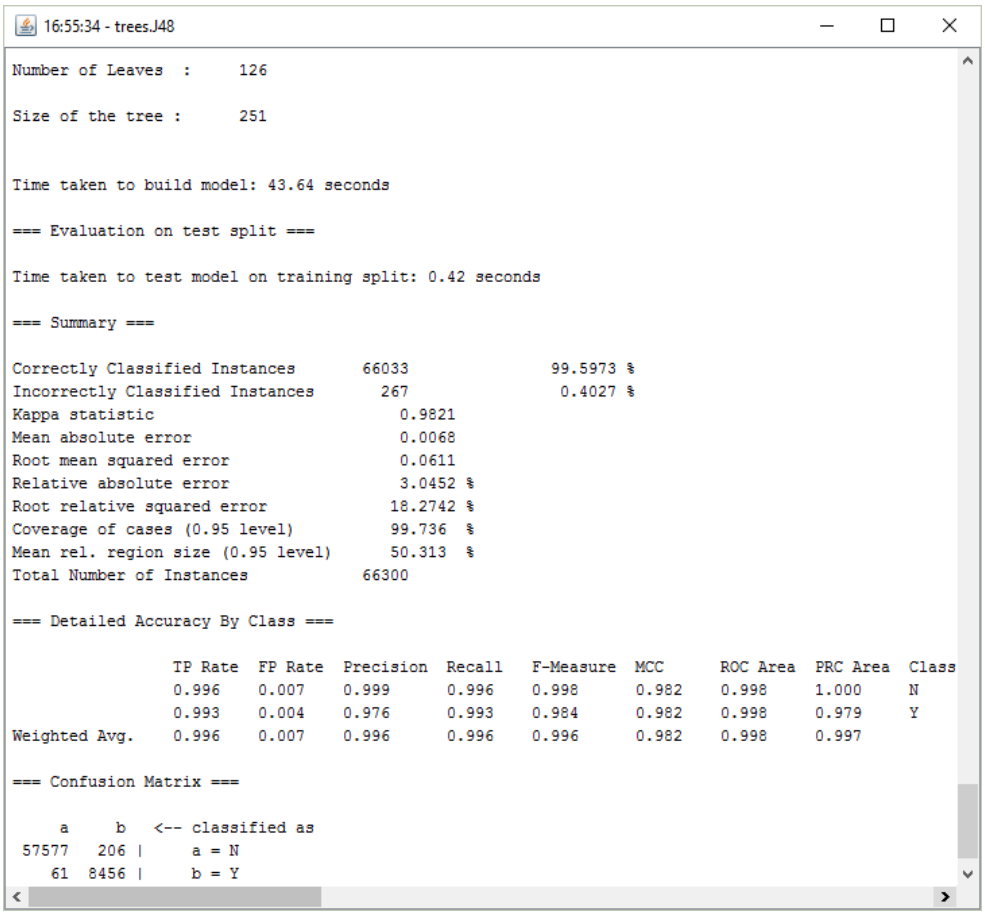
```
16:20:52 - trees.J48
Number of Leaves  :     106

Size of the tree :      211


Time taken to build model: 16.15 seconds

=== Evaluation on test split ===

Time taken to test model on training split: 0.06 seconds

=== Summary ===

Correctly Classified Instances       26404               99.5626 %
Incorrectly Classified Instances       116                0.4374 %
Kappa statistic                          0.9807
Mean absolute error                      0.0055
Root mean squared error                  0.0636
Relative absolute error                  2.4333 %
Root relative squared error             18.9569 %
Coverage of cases (0.95 level)          99.7285 %
Mean rel. region size (0.95 level)      50.2903 %
Total Number of Instances            26520

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0.989    0.003    0.978      0.989    0.983      0.981    0.994     0.979     Y
                 0.997    0.011    0.998      0.997    0.997      0.981    0.994     0.998     N
Weighted Avg.    0.996    0.010    0.996      0.996    0.996      0.981    0.994     0.996

=== Confusion Matrix ===

    a     b   <-- classified as
 3394    39 |     a = Y
   77 23010 |     b = N
```

c. Nhóm 3: TB trả trước chỉ tiêu dùng TKKM trong tháng trước

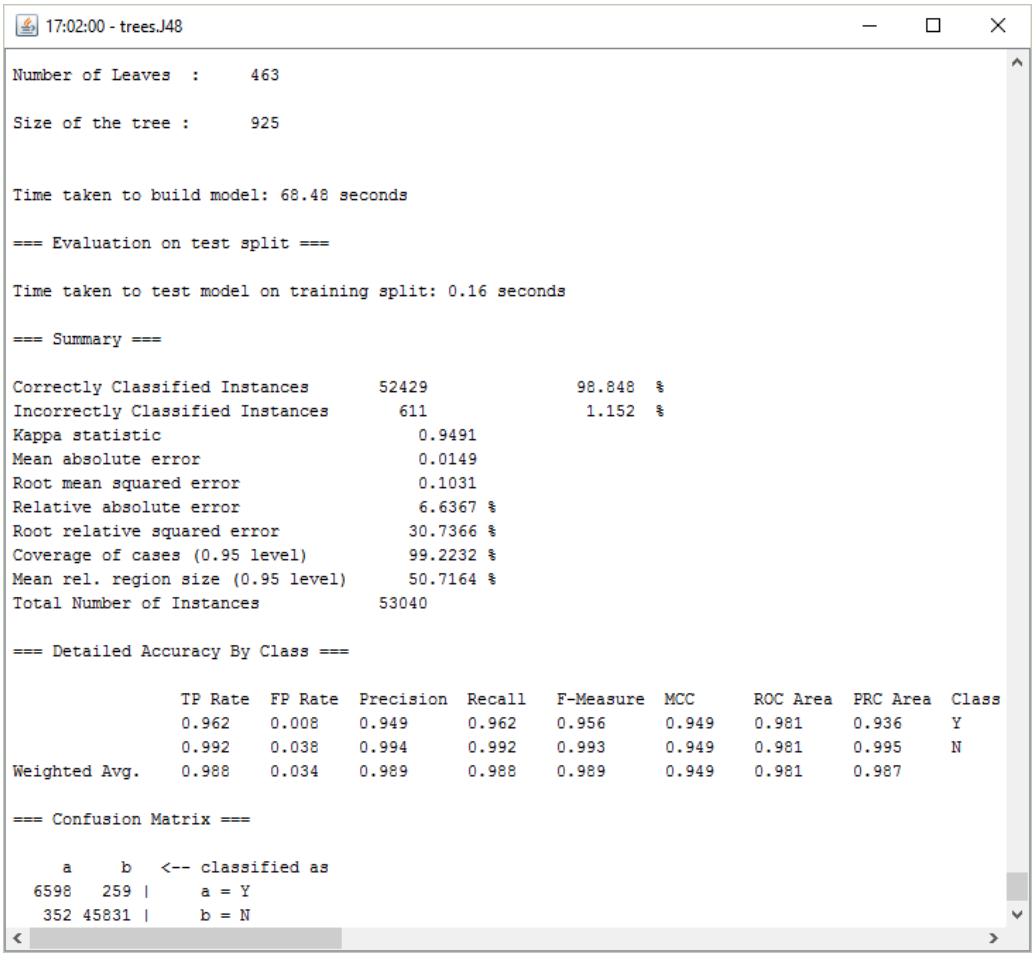


```
16:34:02 - trees.J48
Number of Leaves  :     252

Size of the tree :      503


Time taken to build model: 96.29 seconds

=== Evaluation on test split ===

Time taken to test model on training split: 0.08 seconds

=== Summary ===

Correctly Classified Instances       52774               99.4966 %
Incorrectly Classified Instances       267                0.5034 %
Kappa statistic                          0.9773
Mean absolute error                      0.0063
Root mean squared error                  0.0677
Relative absolute error                  2.8224 %
Root relative squared error             20.3524 %
Coverage of cases (0.95 level)          99.6682 %
Mean rel. region size (0.95 level)      50.3384 %
Total Number of Instances            53041

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0.997    0.018    0.997      0.997    0.997      0.977    0.995     0.999     N
                 0.982    0.003    0.978      0.982    0.980      0.977    0.995     0.975     Y
Weighted Avg.    0.995    0.016    0.995      0.995    0.995      0.977    0.995     0.996

=== Confusion Matrix ===

     a     b   <-- classified as
 46177   145 |     a = N
   122  6597 |     b = Y
```

d. Nhóm 4


16:55:34 - trees.J48
Number of Leaves : 126
Size of the tree : 251
Time taken to build model: 43.64 seconds
=== Evaluation on test split ===
Time taken to test model on training split: 0.42 seconds
=== Summary ===
Correctly Classified Instances 66033 99.5973 %
Incorrectly Classified Instances 267 0.4027 %
Kappa statistic 0.9821
Mean absolute error 0.0068
Root mean squared error 0.0611
Relative absolute error 3.0452 %
Root relative squared error 18.2742 %
Coverage of cases (0.95 level) 99.736 %
Mean rel. region size (0.95 level) 50.313 %
Total Number of Instances 66300
=== Detailed Accuracy By Class ===
TP Rate FP Rate Precision Recall F-Measure MCC ROC Area PRC Area Class
0.996 0.007 0.999 0.996 0.998 0.982 0.998 1.000 N
0.993 0.004 0.976 0.993 0.984 0.982 0.998 0.979 Y
Weighted Avg. 0.996 0.007 0.996 0.996 0.996 0.982 0.998 0.997
=== Confusion Matrix ===
a b <-- classified as
57577 206 | a = N
61 8456 | b = Y

e. Nhóm 5:


17:02:00 - trees.J48
Number of Leaves : 463
Size of the tree : 925
Time taken to build model: 68.48 seconds
=== Evaluation on test split ===
Time taken to test model on training split: 0.16 seconds
=== Summary ===
Correctly Classified Instances 52429 98.848 %
Incorrectly Classified Instances 611 1.152 %
Kappa statistic 0.9491
Mean absolute error 0.0149
Root mean squared error 0.1031
Relative absolute error 6.6367 %
Root relative squared error 30.7366 %
Coverage of cases (0.95 level) 99.2232 %
Mean rel. region size (0.95 level) 50.7164 %
Total Number of Instances 53040
=== Detailed Accuracy By Class ===
TP Rate FP Rate Precision Recall F-Measure MCC ROC Area PRC Area Class
0.962 0.008 0.949 0.962 0.956 0.949 0.981 0.936 Y
0.992 0.038 0.994 0.992 0.993 0.949 0.981 0.995 N
Weighted Avg. 0.988 0.034 0.989 0.988 0.989 0.949 0.981 0.987
=== Confusion Matrix ===
a b <-- classified as
6598 259 | a = Y
352 45831 | b = N

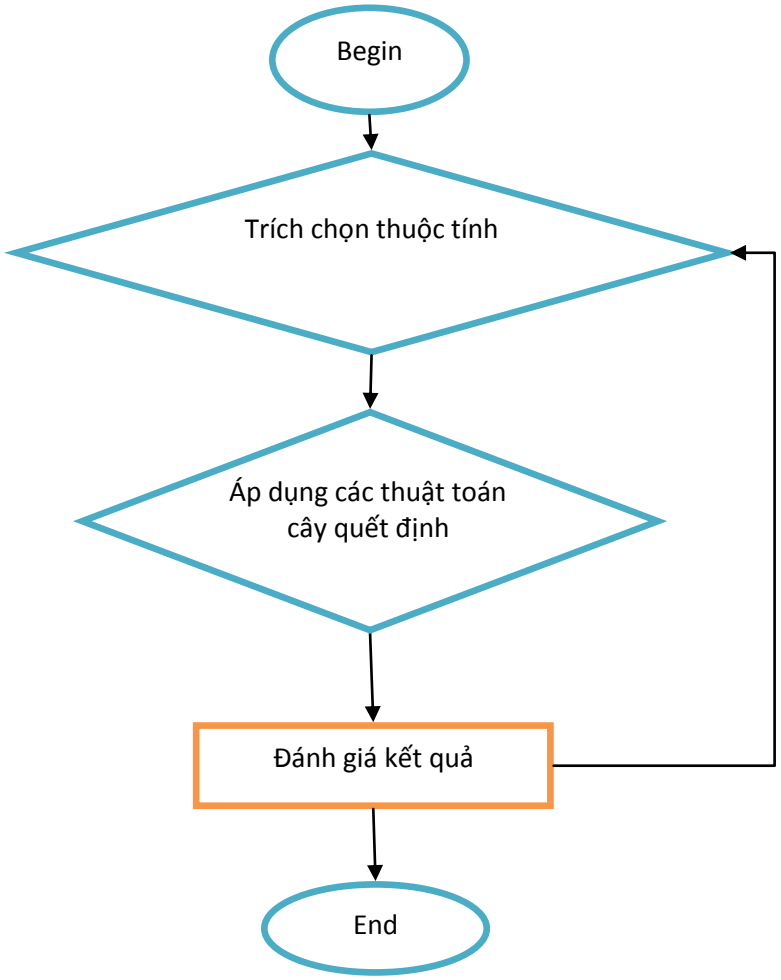
f. Tổng hợp kết quả và đánh giá:
Bảng tổng hợp đánh giá

Tên độ đo	Nhóm 1	Nhóm 2	Nhóm 3	Nhóm 4	Nhóm 5
Accuracy	99.2%	99.6%	99.5%	99.6%	98.8%
Error_Rate	0.8%	0.4%	0.5%	0.4%	1.2%
Recall	97.1%	98.9%	99.7%	99.6%	96.2%
Precision	96.4%	97.8%	99.7%	99.9%	94.9%
Thời gian xây dựng mô hình (giây)	86	16	96	44	68

Bảng 4-3 Bảng tổng hợp kết quả theo giải pháp khai phá dữ liệu

4.4.2 Kết quả thực nghiệm dựa trên khai phá dữ liệu đã cải tiến

Trong phần này, ta sẽ sử dụng tập dữ liệu thuê bao đã được chọn mẫu tại mục 4.1 để thực nghiệm. Các bước của quá trình thực nghiệm như sau:



❖ Bước 1:

- Chạy thuật toán để trích chọn đặc trưng các thuộc tính. Các thuật toán đã thực nghiệm gồm các thuật toán đã nêu tại chương 3. Qua thực nghiệm trên bộ số liệu thực tế thì thuật toán GainRatioAttributeEval phù hợp nhất với các bộ số liệu thực nghiệm. Sau khi thực nghiệm theo thuật toán này thì sẽ ra được danh sách đánh theo thứ tự quan trọng của các thuộc tính và ta tiếp tục phải tìm số lượng thuộc tính tối ưu.

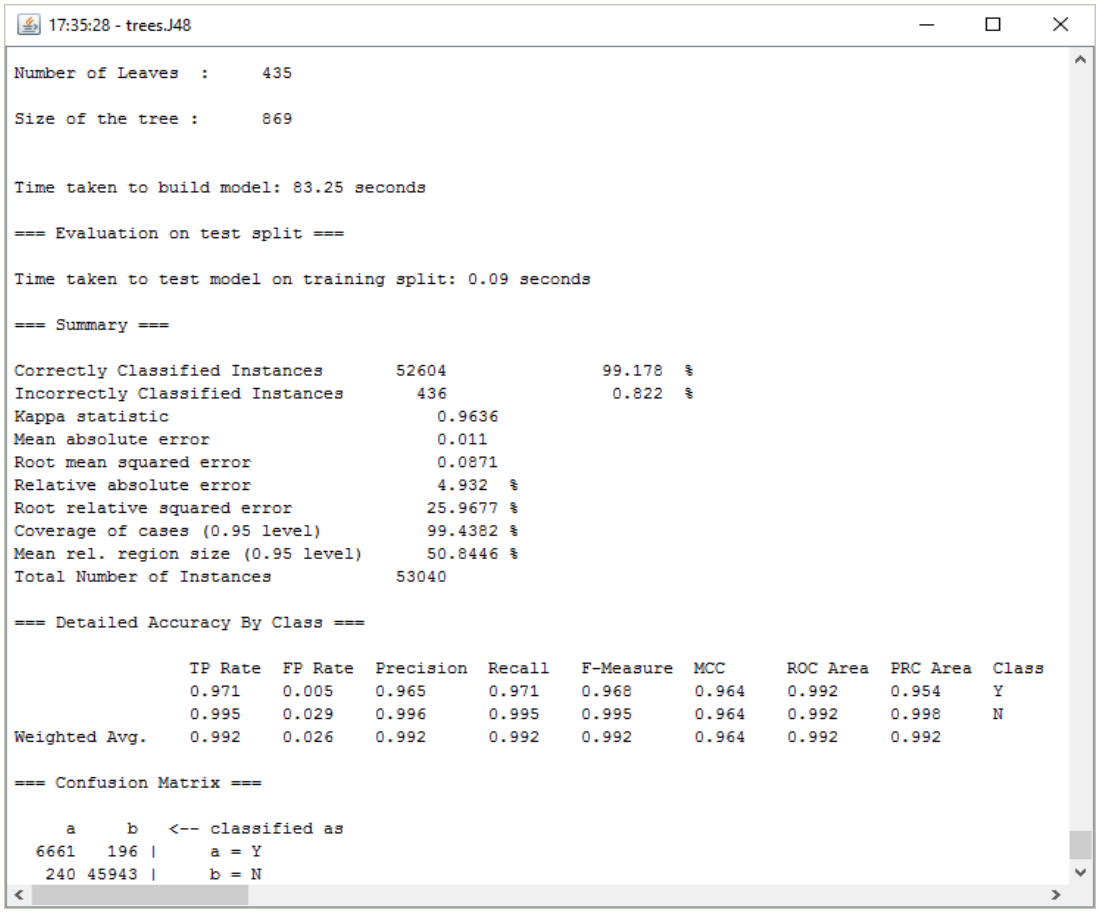
- Lựa chọn số thuộc tính trích chọn: Lựa chọn càng nhiều thuộc tính thì kết quả càng chính xác tuy nhiên thời gian xây dựng mô hình lại càng lâu. Để đảm bảo phù hợp với yêu cầu của MobiFone là xử lý dữ liệu nhanh với độ chính xác chấp nhận được nên ta phải lựa chọn số lượng thuộc tính vừa phải không làm giảm độ chính xác quá nhiều nhưng thời gian xây dựng mô hình nhanh.

❖ Bước 2: Chạy kỹ thuật Cây quyết định (Kỹ thuật J48 trên mô hình WEKA) để xây dựng mô hình phân lớp tập dữ liệu sau khi trích chọn và dự báo đánh giá mô hình đã xây dựng.

Nhóm 1:

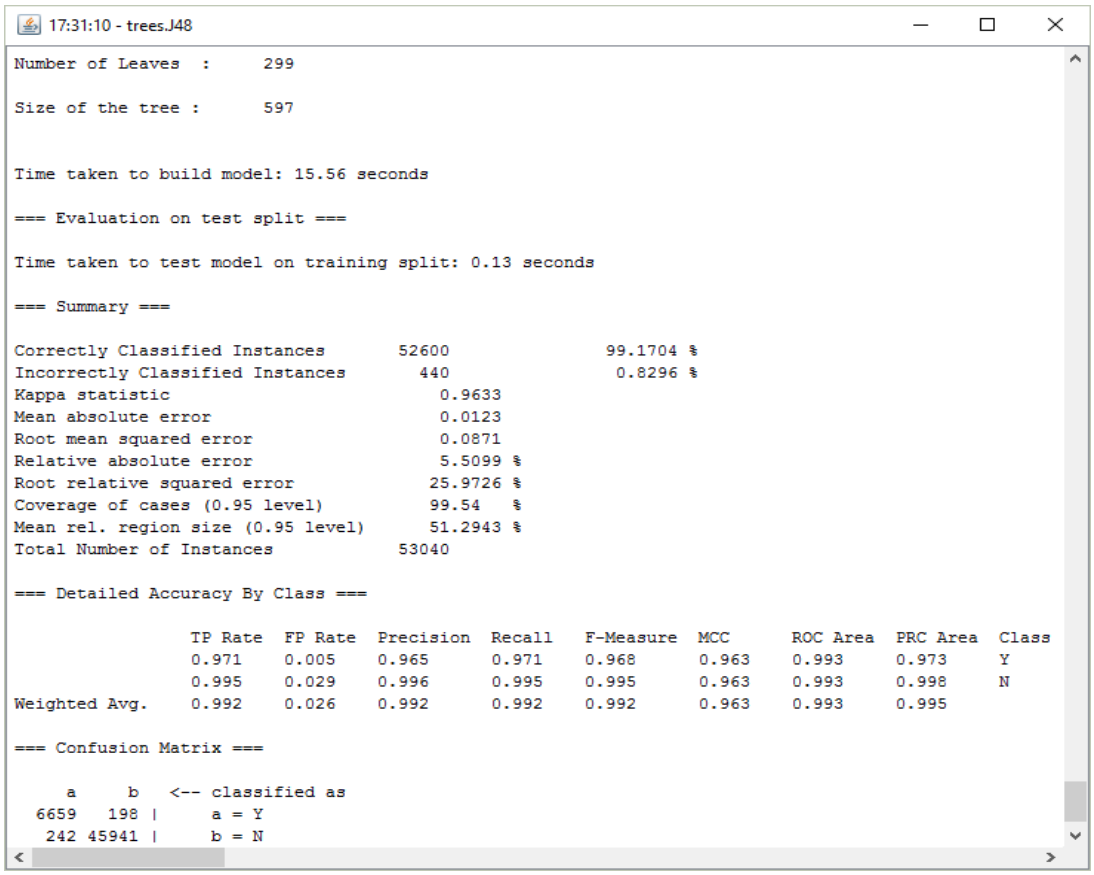
Tập Training: 156.000 TB, trong đó: 20.000 TB rời mạng, 136.000 TB không rời mạng

Kết quả sau khi sắp sếp lại các trường dữ liệu theo xếp hạng của thuật toán GainRatioAttributeEval



Kết quả cho thấy độ chính xác của thuật toán tăng so với trước khi sắp xếp. Thời gian xây dựng mô hình giảm 3s so với bộ dữ liệu chưa xếp hạng. Như vậy có nhận xét thuật toán cây quyết định bị ảnh hưởng bởi thứ tự trường dữ liệu đầu vào.

Lựa chọn số thuộc tính từ xếp hạng trên với 15 thuộc tính được xếp hạng cao nhất kết quả thuật toán C4.5 như sau:

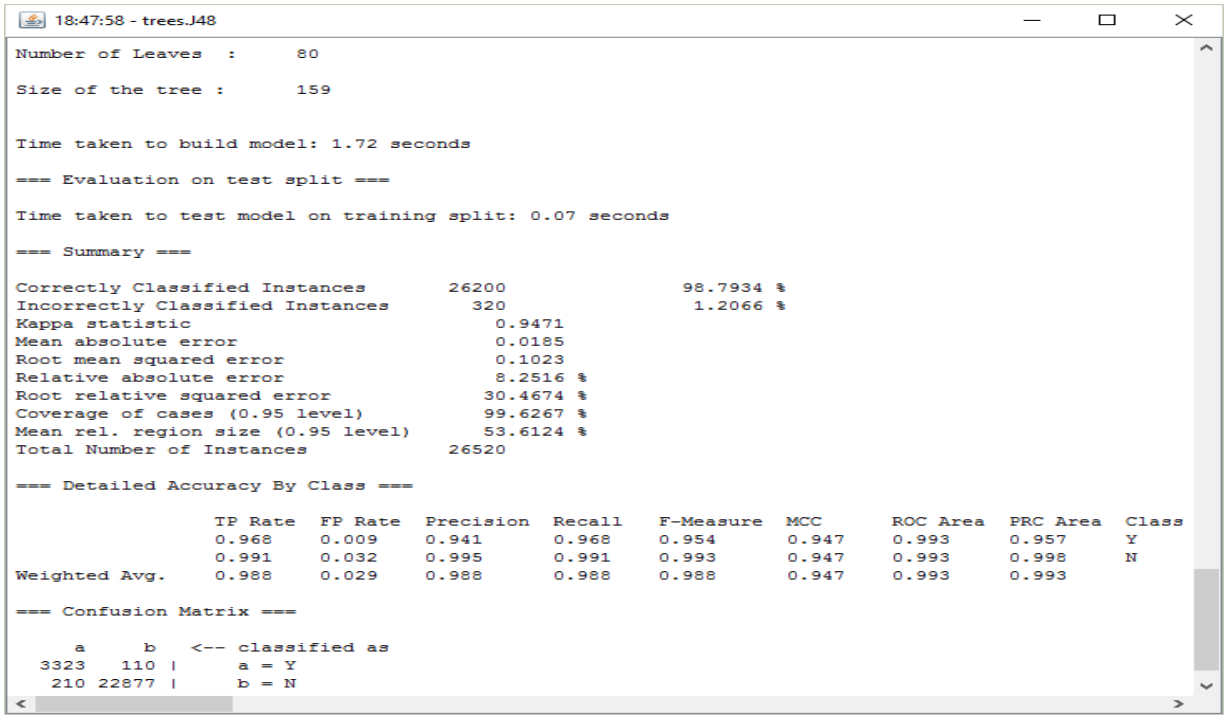


Đánh giá đối với lựa chọn này thì kết quả độ chính xác thuật toán vẫn tương đương so với bộ dữ liệu đã xếp hạng. Thời gian chạy số liệu giảm từ 83s xuống còn 16s

Nhóm 2:

Tập dữ liệu: 78.000 TB, trong đó: 10.000 TB rời mạng, 68.000 TB không rời mạng

Kết quả sau khi thực nghiệm áp dụng thuật toán GainRatioAttributeEval và trích chọn được 10 trường dữ liệu tối ưu như sau:



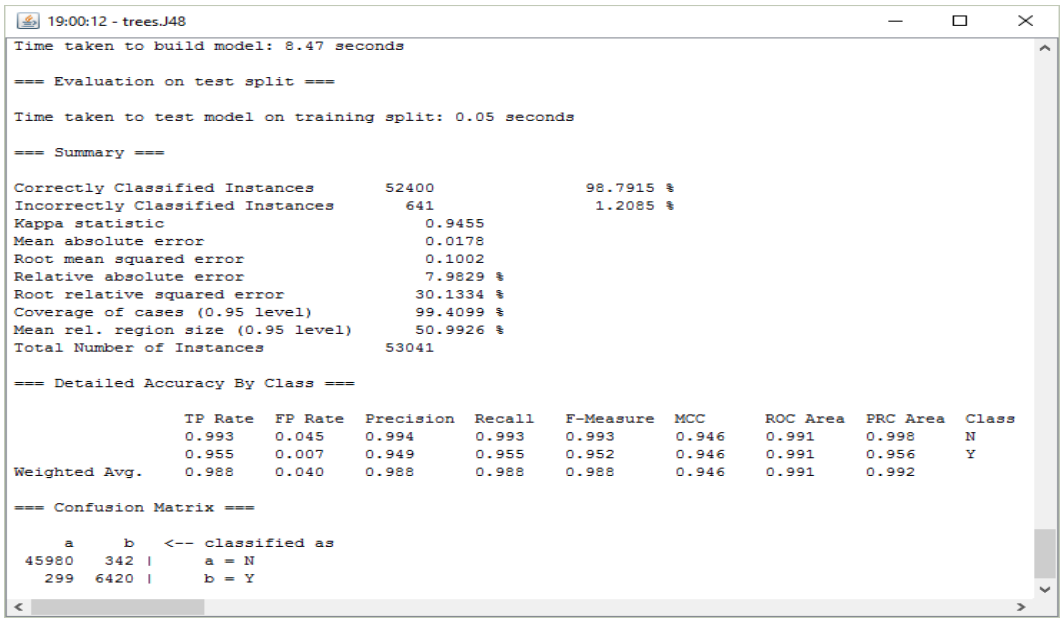
Hình 4-1 Kết quả nhóm 2 của weka

Thời gian dựng mô hình giảm độ chính xác giảm nhưng không đáng kể

Nhóm 3:

Tập dữ liệu chạy: 156.000 TB, trong đó: 20.002 TB rời mạng, 136.000 TB không rời mạng

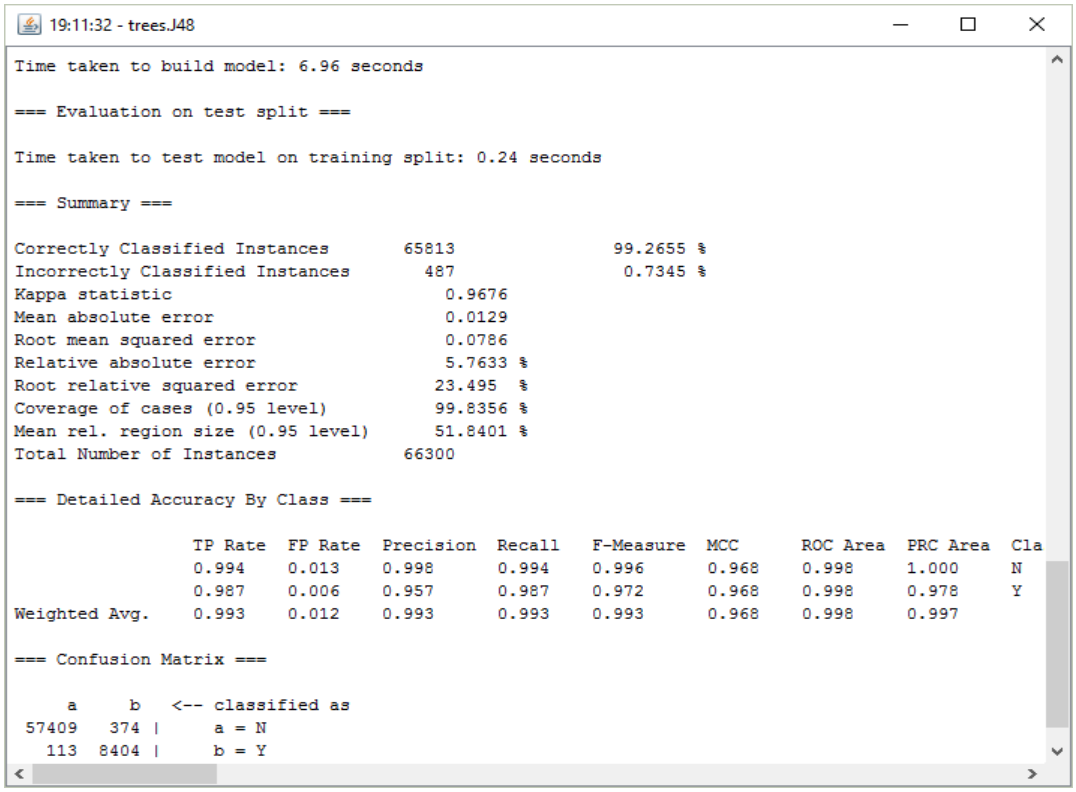
Kết quả sau khi thực nghiệm áp dụng thuật toán trích chọn được 12 thuộc tính:



Nhóm 4:

Tập Training: 195.000 TB, trong đó: 25.001 TB rời mạng, 170.000 TB không rời mạng

Kết quả với 20 thuộc tính lựa chọn

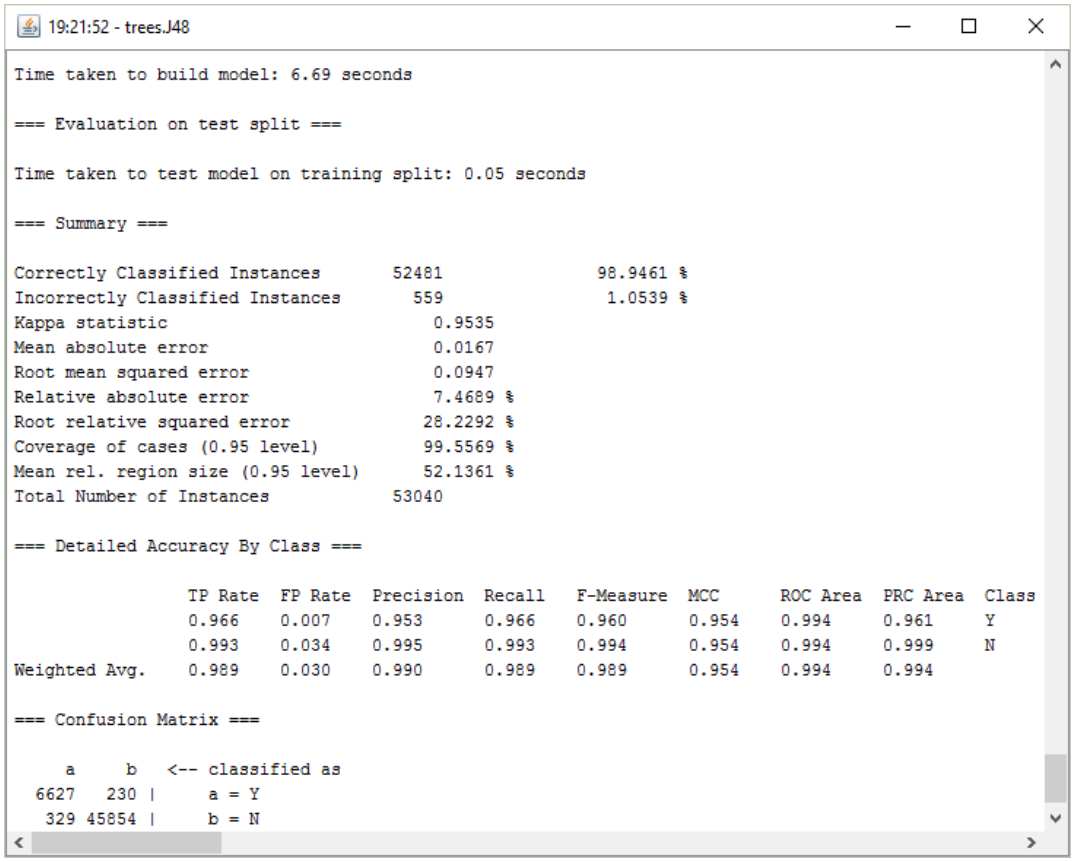


Hình 2 Kết quả nhóm 4 của weka

Nhóm 5:

Tập dữ liệu: 156.000 TB, trong đó: 20.000 TB rời mạng, 136.000 TB không rời mạng

Kết quả với trích chọn 7 thuộc tính.



```
19:21:52 - trees.J48

Time taken to build model: 6.69 seconds

=== Evaluation on test split ===

Time taken to test model on training split: 0.05 seconds

=== Summary ===

Correctly Classified Instances       52481               98.9461 %
Incorrectly Classified Instances       559                1.0539 %
Kappa statistic                          0.9535
Mean absolute error                      0.0167
Root mean squared error                  0.0947
Relative absolute error                  7.4689 %
Root relative squared error             28.2292 %
Coverage of cases (0.95 level)          99.5569 %
Mean rel. region size (0.95 level)      52.1361 %
Total Number of Instances            53040

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0.966    0.007    0.953      0.966    0.960      0.954    0.994     0.961     Y
                 0.993    0.034    0.995      0.993    0.994      0.954    0.994     0.999     N
Weighted Avg.    0.989    0.030    0.990      0.989    0.989      0.954    0.994     0.994

=== Confusion Matrix ===

     a     b   <-- classified as
  6627   230 |     a = Y
   329 45854 |     b = N
```

Hình 3 Kết quả nhóm 5 của weka

Đánh giá mô hình

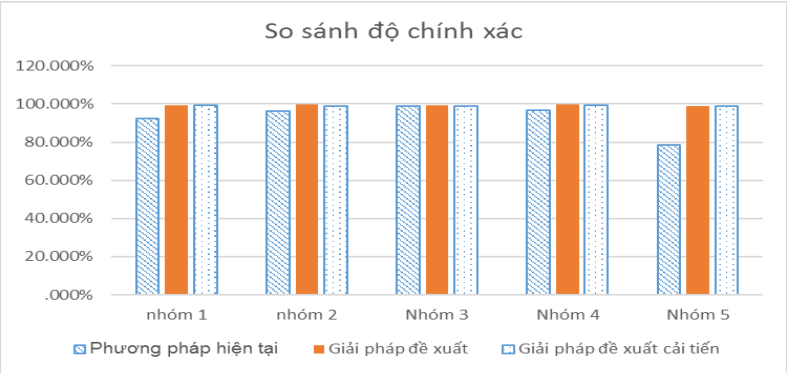
Từ kết quả kiểm nghiệm ở mục 3.3, ta đã tính toán độ chính xác của dự đoán đối với từng nhóm thuê bao dựa vào công thức đã đưa ra như sau:

Tên độ đo	Nhóm 1	Nhóm 2	Nhóm 3	Nhóm 4	Nhóm 5
Accuracy	99.2%	98.8%	98.8%	99.3%	98.9%
Error_Rate	0.8%	1.2%	1.2%	0.7%	1.1%
Recall	96.5%	98.8%	99.3%	99.4%	96.6%
Precision	97.1%	94.1%	99.4%	99.8%	95.3%
Thời gian xây dựng mô hình (giây)	16	2	8	7	7

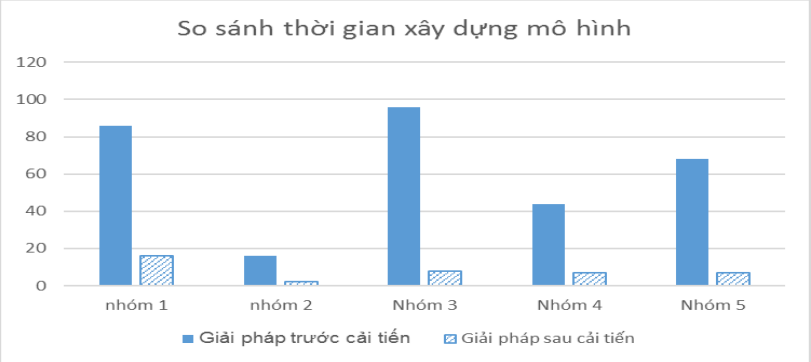
Bảng 4-4 Tổng hợp độ chính xác của giải pháp đề xuất

4.5 So sánh đánh giá kết quả

Để đánh giá kết quả đã thực nghiệm ta so sánh độ chính xác (độ đo Accuracy) của 3 giải pháp đã thực hiện như sau:



Nhận xét: giải pháp hiện tại MobiFone đang sử dụng có độ chính xác thấp nhất, đối với các nhóm 3 và 4 độ chính xác nhỏ hơn nhưng gần tương đương với độ chính xác của các giải pháp dùng thuật toán khai phá dữ liệu, nhưng đối với các nhóm 1,2 và 5 thì độ chính xác thấp hơn đáng kể so với các giải pháp dùng khai phá dữ liệu đặc biệt nhóm 5 thấp hơn rất nhiều (78,4% so với 98,8%). Qua việc so sánh độ chính xác của các giải pháp ta nhận thấy được các đặc trưng của các nhóm thuê bao trong đó các nhóm 3,4 là các nhóm đã được nhận định là các thuê bao rời mạng, nhóm 5 nhận định là nhóm mục tiêu chính để dự báo thì kết quả giải pháp hiện tại của MobiFone chưa cao trong khi đó giải pháp đề xuất cho kết quả rất cao . Độ chính xác của giải pháp đề xuất và giải pháp đề xuất sau khi cải tiến tương đương nhau. Để đánh giá tiếp về giải pháp trước và sau cải tiến ta tiếp tục đánh giá thời gian xây dựng mô hình giữa hai giải pháp này kết quả như biểu đồ dưới:



Qua biểu đồ ta thấy rõ thời gian xây dựng mô hình của giải pháp sau khi cải tiến ít hơn rất nhiều so với giải pháp trước cải tiến. Do vậy giải pháp đề xuất cải tiến đã đạt được mục tiêu đề ra ban đầu là đưa ra giải pháp với độ chính gần tương đương (có thể giảm) tuy nhiên thời gian xây dựng mô hình sẽ nhanh hơn.

4.6 **Kết luận chương 4**

Chương này trình bày kết quả thực nghiệm dựa trên các phương pháp đã thực hiện ở chương 3. Qua thực nghiệm ta có nhận xét như sau: Kết quả dùng thuật toán khai phá dữ liệu cao hơn so với phương pháp hiện tại đang thực hiện

KẾT LUẬN

Từ việc nghiên cứu những yêu cầu cấp thiết đặt ra trong công tác duy trì và phát triển thuê bao của mạng di động, luận văn đã đạt được một số kết quả chính sau đây:

- Xây dựng mô hình dự báo áp dụng kỹ thuật khai phá dữ liệu để phát hiện nhanh chính xác các thuê bao di động có khả năng rời mạng từ đó áp dụng các giải pháp để duy trì thuê bao.

- Triển khai mô hình đề xuất, áp dụng trên dữ liệu thực tế, so sánh với các giải pháp đã sử dụng được áp dụng. Các kết quả đạt được đã cho thấy được tiềm năng áp dụng phương pháp đề xuất vào thực tiễn

Trong thời gian tới chúng tôi sẽ nghiên cứu tích hợp các kỹ thuật này vào các chương trình hỗ trợ kinh doanh của MobiFone đồng thời cải tiến thời gian dự báo cũng như kết quả dự báo. Trong thời gian tới tôi sẽ tiếp tục cập nhật mô hình với dữ liệu của MobiFone để kết quả dự đoán được cải thiện hơn

TÀI LIỆU THAM KHẢO

Tiếng Việt

[1] Phan Xuân Hiếu (2013),*Bài giảng môn học KPDL và kho dữ liệu,* Trường Đại học Công nghệ, Đại học Quốc gia Hà Nội.

[2] Bộ Thông tin và Truyền thông (2014),*Sách Trắng về Công nghệ thông tin và Truyền thông (CNTT-TT) Việt Nam 2014,* Nhà xuất bản Thông tin và Truyền thông, Hà nội.

[3] Hà Quang Thụy (2010),*Bài giảng môn học Kho dữ liệu và KPDL,* Trường Đại học Công nghệ, Đại học Quốc gia Hà Nội.

Tiếng Anh

[4] Leo Breiman, Jerome Friedmen, and Charles J. Stone (1984),*Classification and Regression Trees*, Wadsworth International Group.

[5] M. Chandar, Laha, A., & Krishna, P. (2006),*Modeling churn behavior of bank customers using predictive data mining techniques*, National conference on soft computing techniques for engineering applications.

[6] Jiawei Han and Micheline Kamber (2012),*Data Mining Concepts and Techniques*, Third Edition ed Elsevier Inc.

[7] John Ross Quinlan (1993),*C4.5: Programs for Machine Learning*, Morgan Kaufmann Publishers.

[8] John Ross Quinlan, *Induction of decision trees*, in *Machine Learning 1*. 1986. p. 81-106.

[9] J. Burez, & Van den Poel, D. (2009), *Handling class imbalance in customer churn prediction*, Expert System with Applications,36, 4626-4636.

[10] S. Olafsson, Li, X., & Wu, S. (2008), *Operations research and data mining*, European Journal of Operational Research,187, 2592-1448.

[11] *Weka 3 - Data Mining with Open Source Machine Learning Software in Java*. Available from: http://www.cs.waikato.ac.nz/ml/weka/.