

**BỘ GIÁO DỤC VÀ ĐÀO TẠO
KHOA TOÁN – THỐNG KÊ
TRƯỜNG ĐẠI HỌC KINH TẾ TP. HỒ CHÍ MINH**

KHÓA LUẬN TỐT NGHIỆP

**ỨNG DỤNG THUẬT TOÁN K-MEANS TRONG VIỆC PHÂN
CỤM CỦ PHIẾU NGÀNH NGÂN HÀNG TẠI VIỆT NAM VÀ
DIỄN BIẾN GIÁ TRÊN THỊ TRƯỜNG.**

Giảng viên hướng dẫn : ThS Trần Gia Tùng

Sinh viên : Nguyễn Ngọc Trọng

Khoa : 46

Lớp : FM001

Thành phố Hồ Chí Minh - Năm 2023

**BỘ GIÁO DỤC VÀ ĐÀO TẠO
KHOA TOÁN – THỐNG KÊ**
TRƯỜNG ĐẠI HỌC KINH TẾ TP. HỒ CHÍ MINH

KHÓA LUẬN TỐT NGHIỆP

**ỨNG DỤNG THUẬT TOÁN K-MEANS TRONG VIỆC PHÂN
CỤM CỔ PHIẾU NGÀNH NGÂN HÀNG TẠI VIỆT NAM VÀ
DIỄN BIẾN GIÁ TRÊN THỊ TRƯỜNG.**

Giảng viên hướng dẫn : ThS Trần Gia Tùng

Sinh viên : Nguyễn Ngọc Trọng

Khoa : 46

Lớp : FM001

Thành phố Hồ Chí Minh - Năm 2023

Lời cảm ơn

Đề tài “Ứng dụng thuật toán K-Means trong việc phân cụm cổ phiếu ngành ngân hàng tại Việt Nam và diễn biến giá trên thị trường.” là nội dung tác giả lựa chọn cho bài khoá luận tốt nghiệp sau nhiều năm theo học chương trình Toán tài chính tại trường Đại học Kinh tế TP. Hồ Chí Minh.

Để có thể hoàn thành tốt bài nghiên cứu này, luôn có những lời khuyên chân thành và sự chỉ đạo từ giảng viên hướng dẫn để em có thể thực hiện đúng hướng đề tài. Bên cạnh đó là một môi trường học tập và làm việc tốt mà giảng viên và trường đã tạo ra cho toàn thể sinh viên.

Vì vậy, em xin gửi lời cảm ơn sâu sắc tới toàn thể Ban giám hiệu và đội ngũ giảng viên của trường đã tạo cơ hội cho em được học tập tại đây, cũng như có những kiến thức và kinh nghiệm trong thực tế để hoàn thành tốt bài luận văn. Bên cạnh đó, em xin gửi lời cảm ơn đến giảng viên hướng dẫn vì đã đồng hành cùng em trong suốt quãng đường vừa qua, cảm ơn thầy vì những kiến thức thực tế cũng như kinh nghiệm mà thầy đã chia sẻ với tất cả chúng em.

Em xin chân thành cảm ơn!

Nhận xét của giảng viên

DANH MỤC HÌNH ẢNH

Hình 3.1: Mô phỏng các bước xử lý của bài toán phân cụm.

Hình 4.1: Mô hình phân cụm bằng thuật toán K-Means.

Hình 5.1: Diễn biến giá CTG 1.

Hình 5.2: Diễn biến giá TCB 1.

Hình 5.3: Diễn biến giá TPB 1.

Hình 5.4: Diễn biến giá CTG 2.

Hình 5.5: Diễn biến giá TCB 2.

Hình 5.6: Diễn biến giá TPB 2.

DANH MỤC BẢNG BIỂU

Bảng 3.1: So sánh sự khác nhau giữa phân lớp và phân cụm.

Bảng 4.1: Danh sách các ngân hàng được phân cụm.

Bảng 4.2: Kết quả Shilhoutte Score.

Bảng 4.3: Kết quả phân cụm.

Bảng 5.1: Kết quả đầu tư theo phương án một.

Biểu đồ 5.1: Biểu đồ giá trung bình của ba cụm.

Biểu đồ 5.2: Biểu đồ tỷ suất sinh lợi bình quân phương án hai.

MỤC LỤC

CHƯƠNG 1. GIỚI THIỆU ĐỀ TÀI NGHIÊN CỨU.....	9
1.1 Lý do chọn đề tài	9
1.2 Mục tiêu nghiên cứu và câu hỏi nghiên cứu	11
<i>1.2.1 Mục tiêu nghiên cứu</i>	<i>11</i>
<i>1.2.2 Câu hỏi nghiên cứu.....</i>	<i>11</i>
1.3 Đối tượng và phạm vi nghiên cứu	11
1.4 Ý nghĩa nghiên cứu.....	11
1.5 Kết cấu luận văn	12
CHƯƠNG 2. TỔNG QUAN CÁC NGHIÊN CỨU.....	13
2.1 Các nghiên cứu liên quan trong và ngoài nước	13
<i>2.1.1 Các nghiên cứu về khả năng hoạt động và phá sản của ngân hàng</i>	<i>13</i>
<i>2.1.2 Các nghiên cứu về ứng dụng của thuật toán phân cụm K-Means</i>	<i>14</i>
2.2 Ưu điểm, nhược điểm của các nghiên cứu trước đây.....	15
<i>2.2.1 Ưu điểm</i>	<i>15</i>
<i>2.2.2 Nhược điểm</i>	<i>15</i>
<i>2.2.3 Hướng phát triển</i>	<i>15</i>
CHƯƠNG 3: CƠ SỞ LÝ THUYẾT.....	17
3.1 Tổng quan về học máy.....	17
<i>3.1.1 Học máy.....</i>	<i>17</i>
<i>3.1.2 Học máy trong R.....</i>	<i>19</i>
3.2 Tổng quan về bài toán phân cụm.....	19
<i>3.2.1 Phân cụm trong học máy.....</i>	<i>19</i>
<i>3.2.2 Các bước xử lý của bài toán phân cụm.....</i>	<i>21</i>
3.3 Thuật toán phân cụm K-Means.....	21
<i>3.3.1 Quá trình phát triển.....</i>	<i>21</i>
<i>3.3.2 Các bước của thuật toán</i>	<i>22</i>

3.3.3 <i>Ưu điểm và nhược điểm</i>	23
3.4 Tổng quan về Silhouette Score.....	24
CHƯƠNG 4. XÂY DỰNG MÔ HÌNH PHÂN CỤM VỚI THUẬT TOÁN K-MEANS	26
4.1 Mô hình tổng quát.....	26
4.2 Dữ liệu thực nghiệm	26
4.2.1 <i>Tổng quan về dữ liệu thực nghiệm</i>	26
4.2.2 <i>Thu thập dữ liệu</i>	32
4.2.3 <i>Tiền xử lý dữ liệu</i>	32
4.3 Lựa chọn số lượng cụm phù hợp với dữ liệu thực nghiệm	33
4.4 Phân cụm bằng thuật toán K-Means trên R	34
CHƯƠNG 5. KẾT QUẢ THỰC NGHIỆM VÀ ĐÁNH GIÁ.....	36
5.1 Phát biểu bài toán phân cụm.....	36
5.2 Phương pháp đánh giá kết quả	36
5.3 Đánh giá kết quả và kết luận.....	37
5.4 Ứng dụng phân cụm cổ phiếu bằng thuật toán K-Means	48
5.5 Lợi ích việc phân cụm cổ phiếu bằng thuật toán K-Means	48
CHƯƠNG 6. KẾT LUẬN VÀ KHUYẾN NGHỊ	49
6.1 Kết luận.....	49
6.2 Hạn chế	51
6.3 Kiến nghị và hướng phát triển.....	51
PHỤ LỤC.....	55
TÀI LIỆU THAM KHẢO	55
HÌNH ẢNH.....	57

CHƯƠNG 1. GIỚI THIỆU ĐỀ TÀI NGHIÊN CỨU.

1.1 Lý do chọn đề tài

Trong những năm gần đây, ảnh hưởng của dịch bệnh đã tác động rất nhiều tới đời sống của người dân Việt Nam, nhiều ngành nghề không hoạt động được, người dân không có việc làm. Xuất phát từ mong muốn có thu nhập dựa trên thặng dư vốn trong tình hình dịch bệnh khó khăn ấy nhiều cá nhân, tổ chức luôn cố gắng tìm kiếm kênh đầu tư mới nhằm tạo ra lợi nhuận, điều này đã được chứng minh bằng số lượng tài khoản chứng khoán được đăng ký mới tăng kỷ lục. Theo thống kê, tính tới tháng 7 năm 2023 đã có 7,4 triệu tài khoản chứng khoán được đăng ký, tương đương hơn 7,4% dân số Việt Nam có tài khoản chứng khoán. Tuy nhiên, một lượng lớn các nhà đầu tư cá nhân tham gia vào thị trường chứng khoán khi chưa chuẩn bị nhiều kiến thức rất dễ thất bại trong một thị trường rộng lớn như thế này. Do đó, bên cạnh những kiến thức cơ bản như đọc hiểu báo cáo tài chính trong quá trình định giá, hay sử dụng những chỉ báo kỹ thuật để đưa ra dự đoán giá cổ phiếu trong tương lai... các nhà đầu tư cần tìm hiểu thêm nhiều kiến thức nâng cao, ứng dụng tin học cho việc phân tích và định giá nhất là trong thời kỳ bùng nổ của trí tuệ nhân tạo như hiện nay. Việc ứng dụng các công cụ trong đầu tư tài chính là thật sự cần thiết, thuật toán phân cụm K-Means là một trong số những công cụ đó. Việc phân loại một nhóm cổ phiếu thành nhiều nhóm có những đặc điểm chung khác nhau giúp nhà đầu tư dễ dàng so sánh và đưa ra nhận định của mình trước khi quyết định đầu tư.

Một ngành nghề luôn được các nhà đầu tư quan tâm trên thị trường chứng khoán chính là ngân hàng bởi tầm quan trọng của nó đối với nền kinh tế chung của cả nước. Hệ thống ngân hàng được xem như trái tim của nền kinh tế Việt Nam, chúng luôn chuyển nguồn vốn của cả nền kinh tế. Mặc dù bị ảnh hưởng nhiều từ đại dịch COVID-19 ngành ngân hàng luôn ghi nhận lợi nhuận kỷ lục trong những năm gần đây. Vì thế, những mã cổ phiếu thuộc ngành này được quan

tâm là điều rất dễ hiểu. Nhưng không vì thế mà các nhà đầu tư sẵn sàng đầu tư vào các mã cổ phiếu này kể cả khi kế hoạch lợi nhuận tăng mạnh. Điều này được giải thích bởi tình hình tài chính quốc tế đang có nhiều biến động, những chính sách của Ngân hàng Nhà nước có thể ảnh hưởng đến tâm lý nhà đầu tư khiến giá các cổ phiếu ngân hàng có thể giảm đột ngột. Bên cạnh đó là tâm lý e ngại chất lượng tài sản đảm bảo từ hệ lụy của thị trường trái phiếu doanh nghiệp và bất động sản khiến cho nhà đầu tư băn khoăn khi đầu tư vào các cổ phiếu này. Vì thế việc phân loại các mã cổ phiếu thuộc ngành ngân hàng thành nhiều nhóm là điều cần thiết để giúp nhà đầu tư có cái nhìn tổng quan cũng như chi tiết về từng ngân hàng từ đó đưa ra quyết định đầu tư sao cho hiệu quả và ít rủi ro nhất.

Trong vòng 5 tháng đầu năm 2023, 3 ngân hàng ở Mỹ: Silicon Valley Bank, Signature bank và First Republic Bank đã sụp đổ và có đến 186 ngân hàng đang có nguy cơ sụp đổ theo tờ US Today. Ở Thụy Sĩ, ngân hàng Credit Suisse cũng đã sụp đổ và được mua lại bởi ngân hàng lớn nhất Thụy Sĩ (UBS). Có thể thấy thị trường tài chính thế giới đang khủng hoảng trầm trọng. Điều này là hệ quả tất yếu của hệ thống tài chính hiện hành, cũng như các chính sách hỗ trợ người dân sau đại dịch đi kèm là ảnh hưởng của chiến tranh Nga - Ukraine. Tình hình thế giới là thế, ở Việt Nam cũng không ngoại lệ, trong lịch sử cũng có nhiều ngân hàng hoạt động yếu kém không kiểm soát được những rủi ro trong kinh doanh đã phá sản và gây ra nhiều ảnh hưởng tiêu cực đến thị trường tài chính. Vì thế việc phân loại các ngân hàng theo bất kỳ hình thức nào là vô cùng quan trọng. Một trong số đó là những rủi ro mà các ngân hàng phải đối mặt như là rủi ro thanh khoản, rủi ro tín dụng... Vì thế việc phân cụm theo các chỉ số thể hiện những rủi ro tiềm ẩn của ngân hàng là vô cùng cần thiết khi đưa ra quyết định đầu tư.

Từ những lý do trên, tác giả quyết định lựa chọn đề tài nghiên cứu cho khóa luận tốt nghiệp là “Ứng dụng thuật toán K-Means trong việc phân cụm cổ phiếu ngành ngân hàng tại Việt Nam và diễn biến giá trên thị trường.”

1.2 Mục tiêu nghiên cứu và câu hỏi nghiên cứu

1.2.1 Mục tiêu nghiên cứu

- Phân cụm các mã cổ phiếu thuộc ngành ngân hàng được niêm yết trên thị trường chứng khoán Việt Nam.
- Đánh giá kết quả phân cụm.
- Quan sát diễn biến giá trên thị trường.

1.2.2 Câu hỏi nghiên cứu

- Câu hỏi về lý thuyết: Phân cụm là gì? Thuật toán phân cụm K-Means được hình thành như thế nào?
- Tính thực tế của việc sử kết quả phân cụm K-Means khi đưa ra quyết định đầu tư vào các ngân hàng ở Việt Nam.

1.3 Đối tượng và phạm vi nghiên cứu

Bài nghiên cứu sử dụng đồng thời các chỉ số tài chính chung của các mã cổ phiếu, bên cạnh đó kết hợp những chỉ số riêng dành cho ngân hàng làm đối tượng nghiên cứu. Những chỉ số được sử dụng trong bài nghiên cứu được thu thập theo kết quả trên trang fireant.vn.

1.4 Ý nghĩa nghiên cứu

Với việc ứng dụng thuật toán phân cụm K-Means trong việc phân tích, bài nghiên cứu đưa ra một công cụ mới cho các nhà đầu tư trong việc xây dựng danh mục đầu tư của mình cũng như trong việc quan sát diễn biến giá của các doanh nghiệp trong danh mục đầu tư.

1.5 Kết cấu luận văn

Kết cấu của khoá luận tốt nghiệp gồm 6 chương:

Chương 1: Giới thiệu để tài nghiên cứu trình bày chi tiết về lý do lựa chọn đề tài, mục tiêu nghiên cứu; giới thiệu đối tượng nghiên cứu và phạm vi nghiên cứu cũng như ý nghĩa thực tiễn của đề tài.

Chương 2: Tổng quan các nghiên cứu sẽ tóm tắt về những nghiên cứu trong và ngoài nước có liên quan đến thuật toán phân cụm K-Means cũng như những rủi ro mà ngân hàng có thể gặp phải. Đánh giá những ưu điểm và nhược điểm của những nghiên cứu này và đưa ra các cơ hội còn chưa được khai thác.

Chương 3: Cơ sở lý thuyết sẽ trình bày tổng quan về học máy, giới thiệu bài toán phân cụm, phân lớp. Trình bày chi tiết bài toán phân cụm bằng thuật toán K-Means và ưu nhược điểm của thuật toán.

Chương 4: Xây dựng thuật toán phân cụm K-Means trình bày các bước thực hiện thuật toán; mô tả quá trình để có được kết quả cuối cùng từ các bước thu thập dữ liệu, tiền xử lý dữ liệu, lựa chọn số cụm phù hợp và cuối cùng là thực hiện phân cụm dữ liệu bằng thuật toán K-Means trên phần mềm R.

Chương 5: Kết quả thực nghiệm và đánh giá trình bày kết quả có được sau khi phân cụm. Sau đó dùng những phương pháp đánh giá như so sánh giá trung bình các cụm, sử dụng giá lịch sử và đầu tư theo hai phương án sau đó đưa ra đánh giá về kết quả có được.

Chương 6: Kết luận và khuyến nghị đưa ra kết luận về việc ứng dụng thuật toán K-Means để phân cụm cổ phiếu ngân hàng tại Việt Nam. Tính thực tiễn của kết quả phân cụm và đưa ra các khuyến nghị cho nhà đầu tư khi sử dụng kết quả phân cụm cho việc phân tích và đầu tư.

CHƯƠNG 2. TỔNG QUAN CÁC NGHIÊN CỨU.

2.1 Các nghiên cứu liên quan trong và ngoài nước

2.1.1 Các nghiên cứu về khả năng hoạt động và phá sản của ngân hàng

Kerhar J. Baral (2005) khi nghiên cứu về khả năng hoạt động của các ngân hàng ở Nepal theo mô hình CAMELS đã đưa ra một số kết luận chính: Ngân hàng trong phạm vi nghiên cứu đang cải thiện chất lượng tài sản khi có gắng giảm tài sản xấu trong nhiều năm liền. Các ngân hàng đảm bảo lượng tiền mặt cao, tránh rủi ro về thanh khoản nhưng vì thế đã ảnh hưởng đến lợi nhuận của ngân hàng. Như vậy, ta dễ dàng nhận ra có sự đánh đổi giữa kết quả hoạt động kinh doanh và rủi ro thanh khoản của ngân hàng.

Nil Gunsel (2007) khi nghiên cứu về khả năng thất bại của các ngân hàng khu vực phía Bắc đảo Síp đã cho thấy: khi hoạt động trong một hệ thống quản trị rủi ro kém nhưng lại tăng đòn bẩy sẽ khiến ngân hàng ngày càng gần hơn với khả năng phá sản. Nghiên cứu cũng chỉ ra rằng, các thành phần trong mô hình CAMELS có ý nghĩa quan trọng trong việc dự báo tình trạng kiệt quệ tài chính của các ngân hàng ở đây.

Còn ở Việt Nam, theo Nguyễn Thị Thu Phương (2016) khi nghiên cứu về “Ảnh hưởng của các chỉ số tài chính đến khả năng sinh lợi của hệ thống ngân hàng khu vực Châu Á- Thái Bình Dương”, tác giả kết luận các chi phí ngoài trả lãi trên tổng thu nhập có tác động ngược chiều đến khả năng sinh lợi của ngân hàng. Nghiên cứu cũng chỉ ra được xung đột giữa thu nhập và tính thanh khoản tương tự kết quả của KerHar J. Baral. Bên cạnh đó, trong phạm vi nghiên cứu, tác giả không thấy được tác động của nợ xấu đến khả năng sinh lợi của ngân hàng.

Nguyễn Thị Phương Quyên (2019) khi nghiên cứu “Tác động của các yếu tố rủi ro tài chính đến nguy cơ phá sản ngân hàng thương mại Việt Nam” đã chỉ

ra rằng có mối quan hệ giữa các yếu tố như là: đòn bẩy tài chính, dự phòng rủi ro tín dụng, tỷ lệ thanh khoản, tỷ lệ thu nhập lãi ròng, tỷ lệ chi phí đến khả năng phá sản của các ngân hàng thương mại ở Việt Nam.

Nhìn chung, có nhiều kết luận đã được đưa ra về mối quan hệ giữa các chỉ số tài chính đến kết quả hoạt động kinh doanh cũng như khả năng phá sản của ngân hàng, song vẫn còn sự mâu thuẫn giữa các kết luận. Vì thế việc chỉ dựa vào một hoặc vài kết luận từ các bài nghiên cứu để đưa ra nhận định về kết quả hoạt động hay khả năng phá sản của một ngân hàng là điều rất chủ quan. Điều này khi kết hợp với tâm lý tự tin quá mức sẽ khiến các nhà đầu tư đưa ra quyết định không chính xác khi đầu tư. Do đó việc ứng dụng thuật toán K-Means trong việc gom những mã cổ phiếu ngân hàng có những thuộc tính tương đồng nhau thành một cụm là cần thiết để các nhà đầu tư có cái nhìn khách quan hơn về từng cụm nói chung và từng mã cổ phiếu ngành ngân hàng nói riêng.

2.1.2 Các nghiên cứu về ứng dụng của thuật toán phân cụm K-Means

Thuật toán phân cụm K-Means trong nhiều năm gần đây đã không còn xa lạ với chúng ta, nó được ứng dụng trong nhiều ngành nghề, lĩnh vực cũng như phục vụ nhiều mục đích khác nhau có thể kể đến như: môi trường, y tế, kinh tế, giáo dục... Riêng lĩnh vực ngân hàng, công cụ này cũng đã được ứng dụng để phân loại khách hàng trong thời đại ngân hàng số đang phát triển.

Nguyễn Đức Hiển (2014) với “Mô hình hai giai đoạn dự đoán giá cổ phiếu với K-Means và Fuzzy-SVM” kết luận mô hình mang lại hiệu quả dự đoán cao hơn so với các mô hình đơn như RBN, SVM của một số tác giả trước đó. Việc kết hợp với thuật toán K-Means giúp cải thiện thời gian thực hiện.

Jerzy Korzeniewski (2018) khi tiến hành xây dựng danh mục đầu tư hiệu quả bằng phương pháp phân cụm đã tiến hành thực hiện phân cụm dữ liệu theo

nhiều phương pháp khác nhau. Ông đánh giá kết quả phân cụm của mỗi thuật toán theo kết quả đầu tư trong nhiều khoảng thời gian khác nhau và đã đưa ra kết luận thuật toán phân cụm K-Means có kết quả tốt hơn PAM.

2.2 Ưu điểm, nhược điểm của các nghiên cứu trước đây

2.2.1 Ưu điểm

Đối với những nghiên cứu về các mối quan hệ giữa các chỉ số tài chính đến kết quả hoạt động kinh doanh hay khả năng phá sản của ngân hàng, đã chỉ ra được mối tương quan rõ ràng giữa các yếu tố.

Đối với những nghiên cứu về việc ứng dụng thuật toán phân cụm trong thực tế đều đạt được kết quả tích cực nhất định.

2.2.2 Nhược điểm

Phần lớn nghiên cứu trước đây chưa thể hiện được ứng dụng của thuật toán phân cụm K-Means trong thực tế. Đây là một khoảng trống cần được khai thác và sẽ mang lại lợi ích nếu ứng dụng đúng cách.

2.2.3 Hướng phát triển

Ngày nay với sự phát triển của công nghệ thông tin, việc kết hợp công nghệ với kiến thức thực tế để đưa ra quyết định là việc vô cùng đơn giản và hiệu quả. Nhất là trong việc đầu tư các sản phẩm tài chính nói chung và đầu tư chứng khoán nói riêng. Trong chương trình cử nhân tài chính, kinh tế nói chung, mỗi cá nhân chúng ta đều được trang bị những kiến thức chuyên ngành một cách sâu rộng. Với sự phát triển của Internet, việc tìm hiểu các kiến thức mới trở nên vô cùng dễ dàng. Ta cần tập trung ứng dụng những thuật toán kết hợp với kiến thức đã có để phục vụ cho công việc hàng ngày của chúng ta. Một trong số đó là việc ứng dụng thuật toán phân cụm trong nhiều trường hợp và đánh giá kết quả phân cụm nhằm đưa ra tính khả dụng trong thực tế của thuật toán. Hiện nay việc ứng dụng

thuật toán vào thực tế chưa được phổ biến, nên đây là cơ hội cho những người tiên phong thực hiện.

CHƯƠNG 3: CƠ SỞ LÝ THUYẾT.

3.1 Tổng quan về học máy

3.1.1 Học máy

Trong nhiều năm trở lại đây, AI -Artificial Intelligence (Trí tuệ nhân tạo) hay Machine Learning (Học máy) rất được quan tâm trên thế giới. Trí tuệ nhân tạo đang len lỏi vào mọi lĩnh vực trong cuộc sống một cách âm thầm. Những trợ lý ảo trên xe điện của Tesla, hệ thống nhận diện khuôn mặt của Facebook, hệ thống gợi ý của các nền tảng trực tuyến: Tiktok, Youtube, Netflix, hay máy chơi cờ vây AlphaGo đã chiến thắng người chơi hay nhất thế giới.

Machine learning là một tập con của AI. Theo Wikipedia: “Học máy (Machine Learning) là một lĩnh vực thuộc trí tuệ nhân tạo. Chúng liên quan đến việc nghiên cứu và tạo ra các kỹ thuật cho phép các hệ thống học một cách tự động từ dữ liệu nhằm giải một số vấn đề cụ thể.”

Gần đây, khi mà khả năng tính toán của cá máy tính được nâng cấp một cách đáng kể nhằm tương xứng với lượng dữ liệu khổng lồ thu được bởi các công ty công nghệ lớn trên thế giới, Machine Learning được tiến thêm một bước dài và một lĩnh vực mới được ra đời gọi là Deep Learning (Học sâu). Nhờ có sự ra đời của lĩnh vực này, những việc tưởng chừng không thể vào 20 năm trước đã được thực hiện một cách dễ dàng.

Có hai cách phổ biến để phân loại các thuật toán của học máy: một là dựa trên phương thức học, hai là trên chức năng của mỗi thuật toán.

Nếu phân loại dựa trên phương thức học, học máy được chia làm bốn loại:

- Học có giám sát (Supervised learning): đây là thuật toán dự đoán đầu ra của một dữ liệu mới dựa trên các cặp dữ liệu và kết quả đầu ra đã biết trước. Cặp dữ liệu này còn được gọi là dữ liệu nhãn. Đây cũng là nhóm

phổ biến nhất trong các thuật toán của học máy. Trong nhóm học có giám sát lại được tiếp tục chia thành hai loại chính: phân loại và hồi quy. Trong khi dữ liệu nhãn của phân loại được chia thành một số hữu hạn nhóm thì dữ liệu nhãn của hồi quy là một giá trị thực cụ thể.

- Học không giám sát (Unsupervised learning): đây là thuật toán mà ta không biết được kết quả hay nhãn mà chỉ có dữ liệu đầu vào. Lúc này, thuật toán sẽ dựa vào cấu trúc của dữ liệu để thực hiện một công việc nào đó có thể kể đến như phân nhóm hoặc giảm số chiều của dữ liệu để thuận tiện trong việc lưu trữ và tính toán. Ở nhóm học không giám sát lại được chia nhỏ thành hai loại đó là phân nhóm và kết hợp.
- Học bám giám sát (Semi-Supervised learning): các bài toán sử dụng thuật toán này thường chỉ cần một phần trong chúng được gán nhãn. Những bài toán như vậy là sự kết hợp giữa hai nhóm được kể bên trên.
- Học củng cố (Reinforcement learning): đây là bài toán giúp cho một hệ thống tự động được xác định hành vi dựa trên hoàn cảnh để đạt được lợi ích cao nhất. Thuật toán dạng này được áp dụng chủ yếu trong lý thuyết trò chơi, các nước đi tiếp theo cần được tính toán để đạt hiệu quả cao nhất.

Hiện nay để có thể tìm hiểu về học máy, không thể không kể đến sự hỗ trợ của các ngôn ngữ lập trình có thể kể đến như: C++, Java, Python hay R. Việc sử dụng ngôn ngữ nào để làm công cụ học tập còn phụ thuộc vào cảm nhận của mỗi người và đặc thù của mỗi công việc. Riêng về công việc trong môi trường thống kê, nghiên cứu, phân tích, trình bày dữ liệu thì ngôn ngữ lập trình R tối ưu nhất.

3.1.2 Học máy trong R

Ngôn ngữ lập trình R hay còn gọi là R, được biết đến là một ngôn ngữ lập trình miễn phí với mã nguồn mở được phát triển bởi Ross Ihaka và Robert Gentleman trên ngữ nghĩa khôi từ vựng của ngôn ngữ lập trình Scheme. Với hơn 15000 thư viện trong tất cả lĩnh vực nghiên cứu khác nhau, ngôn ngữ lập trình R được ứng dụng phổ biến trong kinh doanh.

Một số bài nghiên cứu trước đây liên quan tới lĩnh vực tài chính cũng đã sử dụng ngôn ngữ lập trình R có thể kể đến như: “Ứng dụng một số mô hình máy học trong dự báo chiều biến động của thị trường chứng khoán Việt Nam”, “Ứng dụng phần mềm R định giá quyền chọn cho các cổ phiếu trên thị trường chứng khoán Việt Nam” hay “Áp dụng mô hình Garch trên thị trường chứng khoán Việt Nam.” Qua đó thấy được việc sử dụng ngôn ngữ lập trình R nhằm giải quyết những bài toán tài chính là rất phổ biến.

Trong thời đại ngày nay, việc biết thêm ngôn ngữ lập trình sẽ giúp chúng ta rất nhiều trong cả đời sống và công việc. Vẫn còn rất nhiều công việc chưa được tối ưu hoá, nên việc sử dụng ngôn ngữ lập trình, ứng dụng các thuật toán học máy trong công việc là điều cần thiết cho mỗi người chúng ta, đó sẽ là hành trang quý giá khi tham gia vào thị trường lao động ngày nay.

3.2 Tổng quan về bài toán phân cụm

3.2.1 Phân cụm trong học máy

Như đã được liệt kê bên trên, khi phân loại các thuật toán theo phương thức ta có bốn nhóm chính. Bên cạnh đó vẫn còn một cách phân loại khác khi dựa trên chức năng của từng thuật toán. Đặc biệt và được sử dụng nhiều trong học máy là hai thuật toán phân lớp (Classification) và phân cụm (Clustering). Cả hai đều được ứng dụng rất nhiều nhằm giải quyết các vấn đề của con người. Trong bài

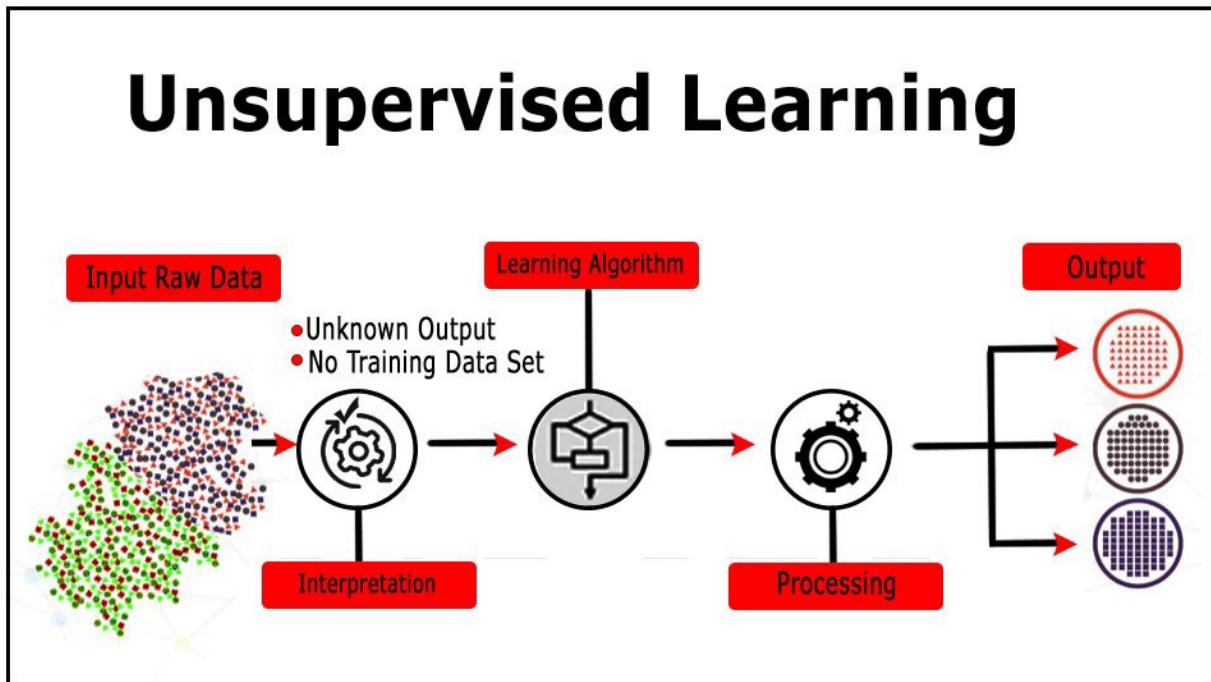
nghiên cứu sẽ sử dụng thuật toán phân cụm dữ liệu trong thực tế. Do đó cùng tìm hiểu đôi nét về thuật toán này.

Đầu tiên ta cần nắm rõ phân cụm là một kỹ thuật trong Data Mining được sử dụng nhằm tìm kiếm, phát hiện các cụm, các dữ liệu tự nhiên tìm ẩn, quan tâm trong tập dữ liệu lớn, từ đó cung cấp thông tin, tri thức hữu ích cho ra quyết định. Mục tiêu của phân cụm dữ liệu là chia các đối tượng thành các cụm thuận nhất và phân biệt với nhau, tức là các nhóm đối tượng thoả mãn: độ tương tự của các đối tượng trong mỗi nhóm cao nhất có thể (tiêu chuẩn liên kết chặt), các đối tượng trong nhóm khác nhau phân biệt nhất có thể (tiêu chuẩn tách rời).

Bảng 3.1: So sánh sự khác nhau giữa phân lớp và phân cụm.

PHÂN CỤM	PHÂN LỚP
Nhóm các đối tượng tương tự, nhờ đó phát hiện cấu trúc ẩn của dữ liệu.	Trích rút đặc trưng từ dữ liệu cho phép phân loại các phần tử mới vào các lớp đã xác định.
Không có nhãn sẵn.	Có nhãn cho một số điểm dữ liệu.
Nhóm các đặc điểm vào cụm dựa vào toạ độ của chúng.	Cần một quy luật cho phép gán nhãn chính xác cho các điểm dữ liệu mới.
Đo kết quả bằng các đặc trưng kiểm chứng độc lập.	Đo kết quả bằng độ chính xác phân lớp.
Học không giám sát.	Học có giám sát.

3.2.2 Các bước xử lý của bài toán phân cụm



Hình 3.1: Mô phỏng các bước xử lý của bài toán phân cụm.

Như vậy, ta đã có được những kiến thức cơ bản của thuật toán phân cụm trong học máy. Nhưng với mỗi mục đích khác nhau, ta có thể tùy chọn những thuật toán phân cụm phù hợp. Một trong số thuật toán phân cụm được biết đến nhiều nhất đó là thuật toán phân cụm K-Means, tác giả cũng đã sử dụng thuật toán này để ứng dụng trong thực tế vì vậy hãy cùng tìm hiểu về thuật toán này.

3.3 Thuật toán phân cụm K-Means

3.3.1 Quá trình phát triển

Trong những năm 50 của thế kỷ trước, một nhà toán học tại Bell Labs là Stuát Lloyd đã công bố một bài báo “ Least Squares Quantization in PCM” mô tả một thuật toán tương tự K-Means, nhưng cũng trong giai đoạn đó ít người quan tâm đến nó. Qua nhiều lần cải tiến, mãi đến năm 1967, John MacQueen đã lần đầu đưa ra khái niệm “K-Means Clustering” trong bài báo “Some Methods for

classification and Analysis of Multivariate Observations". Đồng thời đề xuất sử dụng phương sai để đo lường độ tương đồng giữa các điểm và trung tâm của cụm. Giai đoạn 1980-1990 đã chứng kiến sự phát triển của thuật toán K-Means khi liên tiếp có các phương pháp đánh giá hiệu suất của thuật toán K-Means cũng như mở hướng phát triển mới cho các phương pháp phân cụm. Và đến thời điểm hiện tại, với việc kết hợp với các phương pháp phân cụm khác nhau, sử dụng những độ đo khác nhau cũng như cải thiện việc chọn số cụm ban đầu thuật toán phân cụm K-Means vẫn là một trong những thuật toán phân cụm phổ biến được ứng dụng rộng rãi trong nhiều lĩnh vực từ khoa học dữ liệu, thị giác máy tính cho đến việc phân tích hình ảnh.

3.3.2 Các bước của thuật toán

Các bước thực hiện của thuật toán phân cụm K-Means:

INPUT	Dữ liệu cần được phân cụm và số lượng cụm cần tìm (K).
OUTPUT	Gắn nhãn cụm cho từng điểm dữ liệu của bộ dữ liệu.

- Bước 1: Chọn K điểm bất kỳ làm trung tâm ban đầu của bộ dữ liệu.
- Bước 2: Với mỗi điểm dữ liệu, chúng ta sẽ tính khoảng cách của chúng tới các trung tâm đã có được ở bước một bằng cá độ đo khoảng cách. Sau đó sẽ gán chúng vào cụm có khoảng cách của chúng với trung tâm là gần nhất. Lúc này với cả bộ dữ liệu sẽ có những cụm được hình thành.
- Bước 3: Sau khi hình thành cụm mới, ta cập nhật lại các điểm trung tâm mới tương đương với các giá trị trung bình của các điểm dữ liệu trong cụm đó.

Thuật toán sẽ được thực hiện lặp lại ở bước 2 và 3 cho đến khi không có sự thay đổi trong các cụm.

3.3.3 Ưu điểm và nhược điểm

Như đã biết, phương pháp K-means là một trong những phương pháp phân cụm đơn giản nhất và được sử dụng nhiều trong học không giám sát bởi một số ưu điểm có thể kể đến như:

- Đơn giản và dễ hiểu: Phương pháp này vô cùng đơn giản và dễ hiểu. Phương pháp này làm việc tốt trong nhiều trường hợp, đặc biệt là khi số cụm nhỏ và dữ liệu có kích thước lớn.
- Hiệu quả trong tính toán: Phương pháp có thuật toán đơn giản cho phép xử lý các tập dữ liệu lớn và ít tốn thời gian nhưng vẫn cho kết quả tốt.
- Tính linh hoạt: Khi số lượng cụm được xác định, phương pháp có thể áp dụng cho các tập dữ liệu có kích thước và tập tính khác nhau. Hay nói cách khác phương pháp được sử dụng trong nhiều trường hợp khác nhau.
- Tính nhất quán: Kết quả phân cụm của phương pháp này có tính nhất quán tức là nó sẽ cho cùng một kết quả khi được áp dụng nhiều lần trên cùng một tập dữ liệu và cùng một số lượng cụm.
- Dễ dàng tuỳ chỉnh: phương pháp có thể tuỳ chỉnh để phù hợp với các mục tiêu và đặc tính khác nhau của dữ liệu.

Bên cạnh những ưu điểm vượt trội đó, phương pháp phân cụm K-means cũng còn một số hạn chế như:

- Phụ thuộc vào việc lựa chọn số lượng cụm: phương pháp này yêu cầu nhập số lượng cụm cần phân chia. Việc chọn số lượng cụm sẽ ảnh hưởng đến kết quả phân cụm.
- Nhạy cảm với dữ liệu nhiễu: phương pháp K-Means cũng rất nhạy cảm với dữ liệu nhiễu.

- Không hiệu quả với dữ liệu phi tuyến: tức là các cụm phải có hình dạng cầu hoặc elip với dữ liệu phi tuyến, thuật toán này cũng sẽ cho ra kết quả phân cụm không tốt.
- Khó xử lý với dữ liệu lớn: khi dữ liệu quá lớn, phương pháp K-Means có thể gặp khó khăn trong việc xử lý và tính toán, tốn nhiều tài nguyên và thời gian tính toán.

Tóm lại, phương pháp phân cụm K-Means có những ưu và nhược điểm riêng. Việc sử dụng phương pháp này cần cân nhắc kỹ lưỡng sao cho phù hợp với mục đích nghiên cứu và đặc tính của dữ liệu.

3.4 Tổng quan về Silhouette Score

Ngày nay, có rất nhiều các phương pháp khác nhau phục vụ cho công việc phân tích dữ liệu nói chung và phân cụm dữ liệu nói riêng. Việc đánh giá kết quả cho từng phương pháp là vô cùng quan trọng, nó quyết định việc chúng ta có nên sử dụng phương pháp cho thực tế hay không. Với phương pháp phân cụm K-Means chúng ta có Silhouette Score, đây là một phương pháp đánh giá chất lượng của việc phân cụm bằng cách đánh giá độ tách biệt và tập trung của các cụm.

Cụ thể, điểm Silhouette được tính bằng cách lấy trung bình của các giá trị Silhouette của từng điểm dữ liệu thuộc bộ dữ liệu ban đầu. Mỗi giá trị Silhouette của từng điểm bằng cách lấy trung bình khoảng cách giữa điểm đó và các điểm dữ liệu trong cùng một cụm, và khoảng cách giữa điểm đó và các điểm dữ liệu trong cụm khác. Giá trị này nằm trong đoạn từ -1 đến 1, với -1 cho rằng điểm dữ liệu không phù hợp với cụm và 1 cho thấy điều ngược lại đó là điểm dữ liệu đã được gán vào đúng cụm và khác biệt với các cụm còn lại.

Với việc sử dụng Silhouette Score, chúng ta có thể dễ dàng lựa chọn số lượng cụm phù hợp cho mỗi tập dữ liệu khác nhau dựa trên tính tập trung giữa các điểm dữ liệu trong cụm và tách biệt với các cụm khác. Tuy nhiên, Silhouette

Score còn tồn tại một số hạn chế khi áp dụng cho các bộ dữ liệu phức tạp có cấu trúc không rõ ràng. Trong phạm vi bài nghiên cứu, tác giả sử dụng nhằm lựa chọn số cụm để có được kết quả tốt nhất. Để có thể ứng dụng trong thực tế hay với bộ dữ liệu phức tạp cần áp dụng nhiều phương pháp khác như Elbow Method, Gap Statistic hay phân tích thành phần chính (PCA) để có được kết quả tối ưu và chính xác nhất.

CHƯƠNG 4. XÂY DỰNG MÔ HÌNH PHÂN CỤM VỚI THUẬT TOÁN K-MEANS

4.1 Mô hình tổng quát

Quá trình ứng dụng thuật toán K-Means trong việc phân cụm cổ phiếu thuộc nhóm ngân hàng có mô hình tổng quát như sau:



Hình 4.1: Mô hình phân cụm bằng thuật toán K-Means.

4.2 Dữ liệu thực nghiệm

4.2.1 Tổng quan về dữ liệu thực nghiệm

Với mục tiêu phân cụm các mã cổ phiếu thuộc ngành ngân hàng trên thị trường chứng khoán, tác giả quyết định thu thập dữ liệu về các chỉ số tài chính của các ngân hàng này, bên cạnh đó tác giả thu thập thêm những chỉ số riêng của ngành ngân hàng và chủ động không thu thập các dữ liệu có sự khác nhau quá lớn giữa các ngân hàng như: giá cổ phiếu, vốn hoá,... nhằm có bộ dữ liệu tốt nhằm có được kết quả phân cụm hiệu quả nhất.

- ❖ Tỷ suất sinh lời trên các tài sản có sinh lãi (YOEA - Yield on Earning Assets)

Đây là một chỉ số hay được sử dụng để đánh giá khả năng sinh lời của ngân hàng. Tài sản có sinh lãi hàng năm tạo ra nguồn thu nhập không lồ cho các ngân hàng. YOEA cao thể hiện ngân hàng có mức sinh lời cao từ các tài sản có sinh lãi, bên cạnh đó nếu YOEA thấp có thể hiểu rằng ngân hàng đang sử dụng tài sản sinh lời kém hiệu quả.

Công thức tính:

$$YOEA = \frac{\text{Thu nhập lãi và các khoản thu nhập tương tự}}{\text{Tài sản sinh lãi bình quân}}$$

- ❖ Tỷ lệ thu nhập lãi thuần (NIM – Net Interest Margin)

Chỉ số này thường được đo lường và biểu thị dưới dạng tỷ lệ phần năm. Bên cạnh YOEA thì NIM cũng là một chỉ thể hiện hiệu quả hoạt động của ngân hàng từ khả năng sinh lời từ dòng tiền của ngân hàng đó.

Công thức tính:

$$NIM = \frac{\text{Thu nhập lãi thuần}}{\text{Tài sản sinh lãi bình quân}}$$

- ❖ Tỷ lệ chi phí huy động vốn (COF - Cost of Fund)

Ngân hàng thương mại giữ vai trò quan trọng trong nền kinh tế nói chung và thị trường tài chính nói riêng. Một trong số đó là vai trò trung gian, nghĩa là ngân hàng huy động vốn từ các cá nhân tổ chức đang dư thừa vốn sau đó cho luân chuyển tới các cá nhân, tổ chức đang thiếu hụt vốn và thu lợi nhuận từ sự chênh lệch lãi suất huy động và cho vay. Tỷ lệ chi phí huy động vốn sẽ cho chúng ta biết các ngân hàng đang duy động vốn với mức lãi suất bình quân là bao nhiêu. Chỉ số này càng thấp cho thấy các ngân hàng đang có được nguồn vốn với mức lãi suất càng rẻ, do đó tạo ra lợi thế cạnh tranh trên thị trường tài chính.

Công thức tính:

$$\text{COF} = \frac{\text{Chi phí lãi và các chi phí tương tự}}{\text{Nguồn huy động có tính lãi bình quân}}$$

- ❖ Tỷ lệ cho vay trên tài sản (LAR - Loan to Asset Ratio)

Tỷ lệ cho vay trên tổng tài sản thường được sử dụng để đánh giá rủi ro thanh khoản của ngân hàng. Khi tỷ lệ này càng cao cho thấy ngân hàng đang có tính thanh khoản thấp, lúc này ngân hàng sẽ phải đối mặt với nhiều rủi ro hơn khi tình hình tài chính, kinh tế bất ổn. Mặt khác, với LAR cao cho thấy ngân hàng đang có được lợi nhuận từ các khoản cho vay nhiều hơn do đó thu nhập lãi thuần cũng sẽ tăng lên.

Công thức tính:

$$\text{LAR} = \frac{\text{Dư nợ cho vay}}{\text{Tổng tài sản}}$$

- ❖ Tỷ lệ dư nợ cho vay trên vốn huy động (LDR – Loan to Deposit Ratio)

Đây là một trong những thước đo quan trọng được sử dụng trong hoạt động quản lý và giám sát ngân hàng. Chỉ số này liên quan đến cơ cấu tài sản và nguồn vốn của các ngân hàng, phản ánh tính thanh khoản của ngân hàng. Khá tương đồng với LAR nhưng LDR cụ thể và cho thấy được nhiều ý nghĩa phân tích hơn.

Công thức tính:

$$\text{LDR} = \frac{\text{Dư nợ cho vay}}{\text{Tổng tiền gửi}}$$

❖ Tỷ lệ chi phí trên thu nhập (CIR – Cost to Income Ratio)

Chỉ số này thể hiện mức độ quản trị của ngân. Việc so sánh chi phí hoạt động với tổng doanh thu của ngân hàng đó giúp ta có được cái nhìn khách quan về hiệu quả vận hành của ngân hàng. CIR càng thấp thể hiện ngân hàng đang quản lý chi phí rất hiệu quả. Trước đây, chi phí hoạt động chiếm tỷ lệ lớn nhất đó là chi phí cho nhân viên, nhưng với sự phát triển của công nghệ thông tin, các giao dịch hầu hết được thực hiện trên điện thoại thông minh, một số ngân hàng còn phát hành thẻ trực tuyến mà không cần phải đến quầy giao dịch. Nhưng thay vào đó, ngân hàng cần đầu tư xây dựng hạ tầng công nghệ thông tốt để đảm bảo việc sử dụng dịch vụ của khách hàng là tốt nhất. Có thể trong khoảng thời gian này chưa có sự thay đổi lớn về chi phí hoạt động của ngân hàng, nhưng trong tương lai khi khách hàng thay đổi thói quen sử dụng dịch vụ lúc này sẽ có sự thay đổi đáng kể về chi phí hoạt động của ngân hàng.

Công thức tính:

$$CIR = \frac{\text{Chi phí hoạt động}}{\text{Tổng thu nhập hoạt động}}$$

❖ Tỷ lệ dự phòng rủi ro tín dụng (LLR – Loan Loss Reserves)

Tỷ lệ dự phòng rủi ro tín dụng và tỷ lệ bao phủ nợ xấu có vai trò quan trọng trong việc đánh giá khả năng vỡ nợ của ngân hàng. Qua đó đây cũng là một tỷ lệ được các nhà đầu tư quan tâm khi xem xét đầu tư vào các cổ phiếu này.

Công thức tính:

$$LLR = \frac{\text{Dự phòng rủi ro tín dụng}}{\text{Dự nợ cho vay}}$$

❖ Tỷ lệ bao phủ nợ xấu (LLRNPL)

Đây là một tỷ lệ thể hiện khả năng đối phó trước những rủi ro tín dụng của ngân hàng. Việc có tỷ lệ bao phủ nợ xấu lớn giúp ngân hàng đảm bảo an toàn tài chính, tránh những rủi ro từ thị trường.

Công thức tính:

$$\text{LLRNPL} = \frac{\text{Dự phòng rủi ro tín dụng}}{\text{Tổng nợ xấu}}$$

Ngày nay, khi phân tích tình hình hoạt động của ngân hàng, chúng ta thường sử dụng mô hình CAMEL. Do tính đặc trưng của mỗi ngành nghề mà ta có những khía cạnh cần quan tâm, đối với doanh nghiệp ngân hàng ta quan tâm đến các yếu tố như: mức độ an toàn vốn, chất lượng tài sản, quản trị, thu nhập, thanh khoản. Vì lý do đó mà tác giả đã lựa chọn những yếu tố thuộc các khía cạnh trên để thực hiện phân cụm các mã cổ phiếu thuộc ngành ngân hàng được niêm yết trên thị trường chứng khoán Việt Nam.

Hiện nay tại Việt Nam có tất cả 49 ngân hàng. Trong đó có 31 ngân hàng thương mại cổ phần, 4 ngân hàng 100% vốn nhà nước, 2 ngân hàng chính sách, 2 ngân hàng liên doanh, 9 ngân hàng 100% vốn nước ngoài và 1 ngân hàng hợp tác xã. Nhưng biết đến nhiều nhất là những ngân hàng thương mại cổ phần tại Việt Nam. Để có được dữ liệu một cách chính xác và phổ biến nhất, tác giả quyết định thực hiện thu thập dữ liệu trên các ngân hàng được niêm yết trên thị trường chứng khoán Việt Nam. Dưới đây là danh sách các ngân hàng được thu thập dữ liệu.

Bảng 4.1: Danh sách các ngân hàng được phân cụm.

STT	Tên ngân hàng	Mã cổ phiếu	Sàn niêm yết
1	NH TMCP An Bình	ABB	UPCOM
2	NH TMCP Á Châu	ACB	HOSE
3	NH TMCP Bắc Á	BAB	HNX
4	NH TMCP Đầu tư và Phát triển Việt Nam	BID	HOSE
5	NH TMCP Bản Việt	BVB	UPCOM
6	NH TMCP Công thương Việt Nam	CTG	HOSE
7	NH TMCP Xuất nhập khẩu Việt Nam	EIB	HOSE
8	NH TMCP Phát triển TP Hồ Chí Minh	HDB	HOSE
9	NH TMCP Kiên Long	KLB	UPCOM
10	NH TMCP Bưu điện Liên Việt	LPB	HOSE
11	NH TMCP Quân Đội	MBB	HOSE
12	NH TMCP Hàng hải Việt Nam	MSB	HOSE
13	NH TMCP Nam Á	NAB	UPCOM
14	NH TMCP Phương Đông	OCB	HOSE
15	NH TMCP Xăng dầu Petrolimex	PGB	UPCOM
16	NH TMCP Sài Gòn - Hà Nội	SHB	HOSE
17	NH TMCP Đông Nam Á	SSB	HOSE
18	NH TMCP Sài Gòn Thương Tín	STB	HOSE
19	NH TMCP Kỹ thương Việt Nam	TCB	HOSE
20	NH TMCP Tiên Phong	TPB	HOSE
21	NH TMCP Việt Nam Thương Tín	VBB	UPCOM
22	NH TMCP Ngoại thương Việt Nam	VCB	HOSE
23	NH TMCP Quốc tế Việt Nam	VIB	HOSE
24	NH TMCP Việt Nam Thịnh Vượng	VPB	HOSE

4.2.2 Thu thập dữ liệu

Toàn bộ dữ liệu được tác giả tổng hợp trên trang thông tin cổ phiếu fireant.vn. Khi tìm kiếm các mã cổ phiếu đã được liệt kê bên trên, sẽ xuất hiện các bảng tính thể hiện các dữ liệu phục vụ cho việc phân tích như phân tích cơ bản trên các báo cáo tài chính, hay phân tích kỹ thuật dựa trên dữ liệu quá khứ. Mọi thứ được trực quan hóa do đó rất dễ dàng cho người dùng có thể tìm kiếm và tham khảo.

Tác giả sử dụng tổ hợp các câu lệnh HTML trên trang thông tin cổ phiếu fireant.vn để có thể thu thập được dữ liệu bảng thuộc phần chỉ số tài chính của từng cổ phiếu. Sau đó, sử dụng phần mềm Excel để kết hợp tất cả các bảng dữ liệu thành một bảng thống nhất và thay đổi cấu trúc phù hợp để tiến hành các bước tiếp theo.

4.2.3 Tiền xử lý dữ liệu

Sau khi thu thập dữ liệu, ta có được bảng dữ liệu hoàn chỉnh nhưng chưa thể sử dụng dữ liệu này cho quá trình phân cụm. Do đó ta cần thực hiện bước này để có thể tiến hành phân cụm.

Dữ liệu được thu thập bao gồm nhóm gồm nhiều chỉ số tài chính khác nhau, có những nhóm chỉ số chỉ dao động trong khoảng từ 5 đến 15% nhưng cũng tồn tại những chỉ số ở mức trên dưới 100% vì vậy việc chuẩn hóa dữ liệu là cần thiết để có thể có được bộ dữ liệu phục vụ cho quá trình phân cụm. Ở đây tác giả thực hiện bước co giãn dữ liệu (Scaling data) để chuẩn hóa phạm vi của các đặc trưng dữ liệu và thường được sử dụng trong quá trình tiền xử lý dữ liệu.

4.3 Lựa chọn số lượng cụm phù hợp với dữ liệu thực nghiệm

Sau khi có được bộ dữ liệu đã được chuẩn hoá, ta tiến hành thực hiện lựa chọn số cụm phù hợp cho quá trình phân cụm dữ liệu. Tuỳ thuộc vào từng bộ dữ liệu mà ta có những cách lựa chọn số cụm cho quá trình phân cụm. Trong trường hợp đơn giản nhất, chúng ta quá hiểu về bộ dữ liệu và nắm được số cụm cần thiết, lúc này chỉ cần nhập trực tiếp số cụm mà không cần phải thực hiện bước lựa chọn cụm phù hợp. Tuy nhiên, với độ phức tạp của bộ dữ liệu ta không thể biết trước được số lượng cụm tối ưu do đó cần thực hiện quá trình lựa chọn số cụm cho bộ dữ liệu. Có hai phương pháp chính và thường được sử dụng khi thực hiện thuật toán phân cụm K-Means. Một là phương pháp Elbow hay còn gọi là phương pháp khuỷu tay, ở phương pháp này thuật toán sẽ phân cụm K-Means với một loạt các giá trị k từ 1 đến giới hạn nhất định. Với mỗi giá trị k tương ứng, tính tổng bình phương khoảng cách từ mỗi điểm dữ liệu đến tâm cụm gần nhất sau đó thể hiện tổng này dưới dạng biểu đồ theo mỗi giá trị k. Điểm uốn cong trên biểu đồ thường là giá trị tối ưu cho giá trị k. Hai là phương pháp Silhouette, phương pháp này sẽ tính điểm cho mỗi điểm dữ liệu mỗi điểm dữ liệu trong từng cụm và sau đó tính trung bình điểm Silhouette của tất cả các điểm. Phương pháp này cũng tương tự Elbow khi lặp lại quy trình cho mỗi giá trị k và so sánh điểm Silhouette của từng k, giá trị k cho trung bình Silhouette cao nhất được coi là lựa chọn tối ưu nhất.

Tác giả sử dụng Silhouette Score để lựa chọn số cụm phù hợp cho bộ dữ liệu này với sự hỗ trợ từ phần mềm Orange. Đây là một phần mềm hỗ trợ người dùng trong quá trình phân tích, trực quan hóa dữ liệu.

Bảng 4.2: Kết quả Shilhoutte Score.

Giá trị k	2	3	4	5	6	7	8
Silhouette Score	0.241	0.275	0.251	0.225	0.170	0.167	0.135

Nguồn: Tác giả tự tổng hợp.

Theo kết quả khi tính điểm Sillhoutte trên phần mềm Orange, đã chọn được số lượng cụm phù hợp là 3 với mức điểm Sillhoutte là 0.275, với con số này theo kinh nghiệm từ các chuyên gia phân tích dữ liệu, sau khi có được kết quả phân cụm cần có những kiến thức chuyên môn, thực tế để đánh giá kết quả phân cụm của thuật toán K-Means.

4.4 Phân cụm bằng thuật toán K-Means trên R

Sau khi dữ liệu đã được chuẩn hoá đã có được số cụm tối ưu, tác giả tiến hành thực hiện phân cụm bằng thuật toán K-Means trên ứng dụng R. Với sự phát triển của công nghệ thông tin, hiện nay các thuật toán đã được thiết kế sẵn cho người dùng ứng dụng, bằng một số câu lệnh gọi và sử dụng các thư viện trên R Studio, ta cung cấp các đối số đầu vào cho thuật toán. Sau khi thực hiện phân cụm dữ liệu K-Means trên R, sử dụng câu lệnh để lưu và tải kết quả phân cụm về máy phục vụ cho quá trình phân tích, so sánh.

Các bước thực hiện phân cụm trên R:

```
> library("readxl")  
  
> data <- read_excel("~/Desktop/testn.xlsx")  
  
> data2 <- scale(data)
```

```
> phancum = kmeans( data2, 3, nstart = 25)

> ketqua <- cbind(data, cum = phancum$cluster)

> write.csv(ketqua, file="ketqua.csv")
```

Kết quả có được sau khi phân cụm với số lượng cụm là 3 theo kết quả có được từ điểm Silhouette với số lượng cổ phiếu trong mỗi cụm lần lượt như sau: cụm 1 với ba mã cổ phiếu, cụm 2 với 9 mã và số lượng mã cổ phiếu ở cụm 3 là 12. Kết quả cụ thể như sau:

Bảng 4.3: Kết quả phân cụm.

Cụm	Mã cổ phiếu
1	BID, CTG, VCB
2	ABB, BAB, BVB, EIB, KLB, NAB, PGB, TPB, VBB
3	ACB, HDB, LPB, MBB, MSB, OCB, SHB, SSB, STB, TCB, VIB, VPB

Sau khi có được kết quả phân cụm, tác giả tiến hành thực hiện so sánh, đánh giá sự giống nhau giữa các mã cổ phiếu tổng cùng một cụm và khác nhau giữa các cụm. Đồng thời tiến hành một số phương pháp để đánh giá kết quả phân cụm này.

CHƯƠNG 5. KẾT QUẢ THỰC NGHIỆM VÀ ĐÁNH GIÁ

5.1 Phát biểu bài toán phân cụm

Bài toán phân cụm các mã cổ phiếu ngân hàng được niêm yết trên thị trường chứng khoán Việt Nam:

- Input: Tập hợp các chỉ số tài chính của các mã cổ phiếu ngân hàng.
- Output: Hình thành các cụm dữ liệu.

Sau khi có được kết quả phân cụm, tác giả tiến hành đánh giá hiệu quả của thuật toán phân cụm K-Means.

5.2 Phương pháp đánh giá kết quả

Với đối tượng nghiên cứu là các mã cổ phiếu ngân hàng, để đánh giá hiệu quả của thuật toán phân cụm tác giả quyết định chọn ba danh mục đầu tư tương ứng với ba cụm được hình thành, sau đó sử dụng lịch sử giá của ba danh mục để đánh giá hiệu quả của thuật toán phân cụm K-Means.

Với mỗi cụm đã được phân, tác giả tính giá trung bình của tất cả các cổ phiếu trong cụm, tỷ suất sinh lợi của giá trị trung bình này đại diện cho tỷ suất sinh lợi của cả danh mục đầu tư. Để có cái nhìn tổng quát về các cụm được phân cụm theo thuật toán K-Means, đầu tiên ta xem qua xu hướng giá trung bình của các cụm. Điều này sẽ giúp các nhà đầu tư nhìn nhận được các điểm tương đồng và khác biệt giữa các cụm, kết hợp với những kiến thức vĩ mô, tài chính... để giải thích và kết luận.

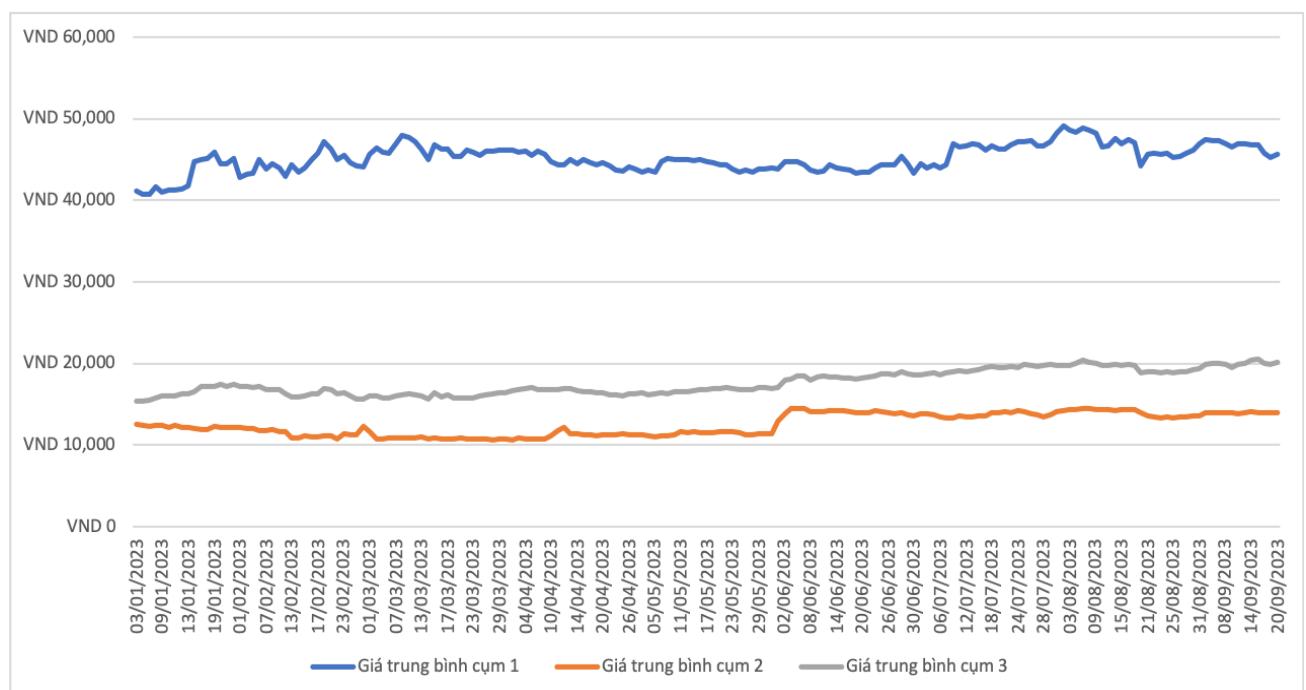
Sau khi nhìn nhận bao quát các cụm, tác giả tiến hành sử dụng giá lịch sử để so sánh tỷ suất sinh lợi bình quân trong trường hợp nhà đầu tư đầu tư theo cả hai hình thức là đầu cơ và đầu tư giá trị. Với hình thức đầu tư đầu cơ, tác giả chọn T+3 là ngày bán cổ phiếu, thực hiện liên tục trong giai đoạn từ đầu năm 2023 cho

đến ngày 17 tháng 9 năm 2023. Đối với nhà đầu tư giá trị, tác giả đánh giá tỷ suất sinh lợi theo 3 thời gian đầu tư khác nhau là 3 tháng, 6 tháng và 9 tháng, qua đó đánh giá được giá trị thực sự của các ngân hàng thuộc các cụm.

Sau khi đánh giá chung kết quả phân cụm theo từng cụm, tác giả tiếp tục quan sát diễn biến của các cổ phiếu thuộc mỗi cụm để có thể đưa ra kết quả đánh giá cụ thể hơn. Với mỗi cụm tác giả lựa chọn các cổ phiếu đại diện và đánh giá diễn tiến trên thị trường qua các giai đoạn.

5.3 Đánh giá kết quả và kết luận

Sau khi có được kết quả phân cụm, tác giả tiến hành tính giá trung bình của ba nhóm cổ phiếu từ đầu năm 2023 đến ngày 20 tháng 9 năm 2023 được kết quả như sau.



Biểu đồ 5.1: Biểu đồ giá trung bình của ba cụm.

Có thể thấy, cụm 1 với ba mã cổ phiếu đầu ngành ngân hàng lần lượt là BID, CTG và VCB có mức giá trung bình lớn nhất trong ba cụm được phân. Ba

mã cổ phiếu này cũng chính là ba cổ phiếu có vốn nhà nước được sử dụng để phân cụm trong bài nghiên cứu này. Đây là một dấu hiệu khả quan cho thấy khả năng phân cụm của thuật toán K-Means tương đối tốt khi tách biệt ba cổ phiếu này thành một cụm so với phần còn lại của bộ dữ liệu. Cụm 1 gồm ba ngân hàng lớn của Việt Nam, trong bối cảnh hình kinh tế, chính trị bất ổn trong những năm gần đây, cụm này sẽ biến động liên tục theo các chính sách của nhà nước cũng như tình hình tài chính thế giới. Điều này có thể thấy qua biểu đồ giá trung bình của cụm 1 trong giai đoạn từ đầu năm đến nay cuối tháng 9 năm 2023 này. Dễ dàng thấy được sự khác biệt so với hai cụm còn lại, khi đường giá trung bình của cụm 1 thay đổi liên tục trong giai đoạn 3 tháng đầu năm và 3 tháng gần đây. Ngược lại với cụm 1, cụm 2 và 3 có đường giá trung bình ít biến động hơn. Điều này có thể được giải thích bởi số lượng các mã cổ phiếu trong hai cụm này là lớn hơn nhiều so với cụm 1 khi cụm 2 gồm 9 mã cổ phiếu và con số này là 12 với cụm 3. Việc có nhiều cổ phiếu trong một cụm làm cho giá trung bình của cả cụm ít thay đổi bởi sự bù trừ cho nhau khi có sự thay đổi giá của từng cổ phiếu trong cụm.

Tuy khác biệt nhau về sự biến động trong giai đoạn này, nhưng tất cả các mã cổ phiếu thuộc nhóm ngân hàng nói riêng và các cụm nói chung đều có sự tăng trưởng nhất định. Theo kết quả của Tạp chí Tài chính, trong hơn nay đầu năm 2023 cổ phiếu ngân hàng có sự tăng mạnh nhờ sự hỗ trợ của các chính sách. Nhiều cổ phiếu thuộc nhóm ngân hàng tăng mạnh trong 7 tháng đầu năm 2023 có thể kể đến: PGB tăng hơn 77%, NAB tăng 44% và VBB tăng 50% đều thuộc cụm 2, bên cạnh đó STB tăng 23% cùng với VPB tăng 20% thuộc cụm 3 và cuối cùng là hai ông lớn của cụm 1 là BID và VCB tăng lần lượt 19% và 11%. Sự tăng trưởng mạnh mẽ này diễn ra vào giai đoạn tháng 5 khi các dòng vốn bắt đầu gia nhập thị trường, kết hợp với các biện pháp mà Ngân hàng Nhà nước thực hiện với mục đích thúc đẩy hỗ trợ doanh nghiệp, tăng trưởng kinh tế. Sau giai đoạn tăng lãi suất tiền tiết kiệm vào giai đoạn cuối năm 2022 đầu năm 2023 nhằm

siết chặt nguồn tiền đã được bơm ra ngoài thị trường trong giai đoạn đại dịch Covid-19, chính sách tiền tệ nói lỏng được thực hiện và ngày càng rõ ràng trong năm 2023. Theo đó, ta có thể quan sát được trên biểu đồ giá trung bình của cụm 2 và 3 có sự thay đổi rõ ràng vào giai đoạn cuối tháng 5.

Nhìn chung, theo biểu đồ giá trung bình của mỗi cụm, ta thấy có sự phân biệt rõ ràng giữa ba cụm về giá. Khi cụm 1 bao gồm 3 cổ phiếu đầu ngành và có giá cao nhất, trong khi cụm 2 chứa tất cả các cổ phiếu đang được niêm yết trên sàn UPCOM và có giá trung bình thấp nhất trong ba cụm. Cụm 3 gồm những cổ phiếu còn lại là những ngân hàng đang hoạt động tốt và có nhiều cơ hội tăng trưởng trong tương lai. Bên cạnh đó, tuy không sử dụng các dữ liệu thể hiện quy mô, cơ cấu vốn... để phân cụm dữ liệu nhưng kết quả có được khá đúng với thực tế về quy mô của của từng cụm. Khi cụm 1 bao gồm ba ngân hàng được xem là “big 4” ngân hàng góp mặt, cụm 2 bao gồm những ngân hàng với quy mô tài sản nhỏ so với thị trường, cụm 3 bao gồm những ngân hàng theo sau cụm 1 với quy mô tài sản lớn thuộc top 10 ngân hàng có tổng tài sản lớn như: MBB, TCB, VPB, ACB, SHB, HDB, STB.

Tiếp theo, tác giả sử dụng giá trung bình lịch sử của các cụm và đưa ra hai quyết định đầu tư. Giả định, giá trung bình của mỗi cụm đại diện cho một mã cổ phiếu, tác giả thực hiện mua 3 mã cổ phiếu tương ứng với 3 cụm này vào ngày T+0 và bán ra ở ngày T+3, thực hiện liên tục trong khoảng thời gian từ đầu năm 2023 đến ngày mua cuối cùng là ngày 17 tháng 9 năm 2023. Ở quyết định đầu tư thứ 2, mua các mã chứng khoán này và nắm giữ trong vòng 1 năm từ tháng 9 năm 2022 đến tháng 9 năm 2023.

Ở cả hai lựa chọn đầu tư, tác giả đã bỏ qua các khoản chi phí giao dịch cũng như thuế thu nhập cá nhân đối với giao dịch chứng khoán. Ở quyết định đầu tiên khi thực hiện liên tục giao dịch mua bán trong vòng 9 tháng đầu năm 2023 ta được kết quả như sau.

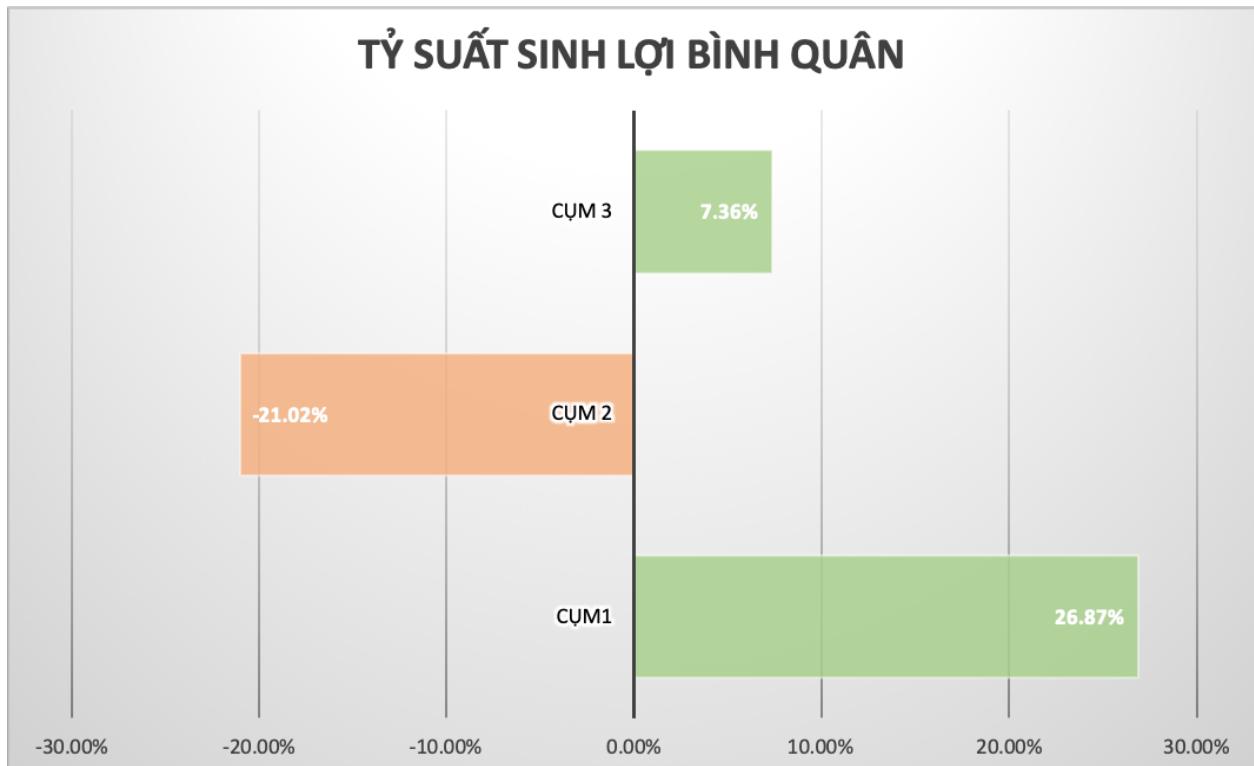
Bảng 5.1: Kết quả đầu tư theo phương án một.

Cụm	Sinh lợi trung bình	Thua lỗ trung bình	Tỷ lệ sinh lợi	Tỷ lệ thua lỗ	Tỷ suất sinh lợi bình quân
1	2.21%	-1.65%	48.52%	51.48%	0.22%
2	3.02%	-2.35%	46.58%	53.42%	0.16%
3	1.86%	-1.68%	61.05%	38.95%	0.48%

Nguồn: Tác giả tự tổng hợp

Có thể thấy, với quyết định đầu tư đầu tiên cụm 3 có tỷ suất sinh lợi bình quân cùng với xác suất lời cao hơn hẳn so với hai cụm còn lại. Cụm 3 luôn có khối lượng giao dịch được duy trì cao. Trong tháng 9 năm 2023, cụm 3 luôn duy trì ở khối lượng giao dịch trung bình hơn 9 triệu trong khi con số này dao động ở mức 1 triệu với cụm 1 và gần 400 ngàn cổ phiếu ở cụm 2. Do đó, dễ dàng nhận ra các mã cổ phiếu trong cụm 3 thường có tính thanh khoản cao hơn, vì thế cũng là sự lựa chọn ưa thích của các nhà đầu tư nhỏ lẻ. Cụm 1 với mức giá trung bình khá cao khi bao gồm các mã CTG, BID và đặc biệt là VCB nên khối lượng giao dịch của cụm này khá ít, nhưng mức sinh lợi trung bình cụm này khá cao, tỷ suất sinh lợi ổn định. Còn với cụm 2, ta thấy sinh lợi trung bình cao nhất khi cụm này có mức giá trung bình thấp, bên cạnh đó mức thua lỗ trung bình cũng cao. Lợi nhuận đi kèm rủi ro, vì thế ta thấy tỷ suất sinh lợi bình quân cụm này thấp nhất. Các mã cổ phiếu ở cụm 3 lại đáp ứng được những hạn chế của hai cụm còn lại. Với mức giá trung bình ổn định, không cao cũng không quá thấp giúp các nhà đầu tư dễ dàng tham gia đầu tư. Kết hợp với kết quả bên trên, cụm 3 bao gồm những mã cổ phiếu của những ngân hàng lớn chỉ sau “Big 4” ngân hàng ở Việt Nam, và đều là những cái tên có kết quả hoạt động sôi nổi trên thị trường tài chính. Vì thế mà khối lượng giao dịch trung bình của cụm này tương đối lớn và cho kết quả đầu tư theo phương án đầu tiên là tốt nhất.

Ngược lại với phương án đầu tiên thứ nhất, phương án thứ hai thường được sử dụng để đầu tư vào giá trị cốt lõi của công ty, những cổ phiếu được bơm giá và thiếu minh bạch trong báo cáo kết quả hoạt động kinh doanh thường không được sử dụng để đầu tư theo phương án này. Với phương án đầu tư giá trị, tác giả thực hiện mua các mua và nắm giữ cổ phiếu với mức giá trung bình đại diện cho mỗi cụm và nắm giữ trong vòng một năm từ tháng 9 năm 2022 đến tháng 9 năm



2023.

Biểu đồ 5.2: Biểu đồ tỷ suất sinh lợi bình quân phương án hai.

Như vậy, với ba mã cổ phiếu được xem là những ngân hàng đầu ngành thì cụm 1 cho thấy được kết quả đầu tư giá trị khả quan nhất với tỷ suất sinh lợi vượt trội so với hai cụm còn lại. Tiếp theo đó là cụm 3 với tỷ suất sinh lợi bình quân là 7.36%, con số này nằm ở mức ngang bằng so với lãi suất tiết kiệm kỳ hạn 12 tháng vào khoảng thời gian này năm trước. Riêng cụm 2 có kết quả đầu tư âm nếu đầu tư theo phương án hai. Có thể thấy các mã cổ phiếu trong cụm này phần lớn không phù hợp với những nhà đầu tư giá trị.

Sau khi có được kết quả tổng quan của mỗi cụm, để có thể kiểm chứng được những kết luận đã đưa ra, tác giả tiến hành quan sát diễn biến giá trên thị trường của một số cổ phiếu trong từng cụm cổ phiếu nhằm củng cố các kết luận đã nêu trên.

Trước tiên để có thể kiểm chứng kết luận về kết quả đầu tư theo phương án hai, tức người mua sẽ mua và nắm giữ cổ phiếu trong vòng một năm. Đối với cụm 1 tác giả sử dụng biểu đồ giá của mã CTG, mã TCB với cụm 2 và TPB với cụm thứ 3. Kết quả có được như sau.



Hình 5.1: Diễn biến giá CTG 1.



Hình 5.2: Diễn biến giá TCB 1.



Hình 5.3: Diễn biến giá TPB 1.

Có thể thấy, nhìn chung trong vòng một năm qua đường xu hướng chính của cả 3 cụm đều có xu hướng tăng. Nhưng có sự khác nhau về tỷ lệ tăng, với lần lượt 2 mã CTG và TCB đều có mức tăng từ 40-45% trong khi con số này ở mức

30% với VPB. Đây cũng là cơ sở để các nhà đầu tư dài hạn tham khảo khi đưa ra quyết định đầu tư. Điều này tương ứng với kết quả đánh giá chung cho mỗi cụm khi đầu tư theo phương án hai. Theo kết quả đó, CTG đại diện cụm 1 và TCB đại diện cụm 3 đều cho kết quả sinh lợi dương, kết quả sinh lợi của VPB thì ngược lại với kết quả chung của cả cụm 2. Điều này cho thấy, kết quả chung của cả cụm chỉ mang tính chất tổng quát, tham khảo ở mức tổng thể, chưa thể kết luận các cổ phiếu trong cụm đó có kết quả đầu tư như cả cụm. Vì thế, khi có được kết quả phân cụm cần kết hợp nhiều kiến thức thực tế, vĩ mô để đánh giá lại từng kết quả trong cụm để có được kết quả tốt nhất.

Có thể nói năm 2021 là một dấu mốc lớn của thị trường chứng khoán Việt Nam khi có rất nhiều kỷ lục được thiết lập trong năm. Dòng tiền đổ vào thị trường trong bối cảnh dịch bệnh phức tạp, số lượng tài khoản đăng ký mới kỷ lục cùng với đó là nhiều yếu tố thuận lợi để mang lại cho thị trường một năm 2021 vô cùng tươi đẹp. Tuy nhiên, qua đến năm 2022 ta chứng kiến sự sụt giảm trầm trọng của toàn thị trường bởi sự chuyển dịch dòng tiền bởi tình hình chính trị nhiều biến động, bên cạnh đó là những chính sách của các nước nhằm kìm hãm lạm phát đang tăng cao. Thị trường chứng khoán Việt Nam nói chung giữ được đà tăng trưởng từ năm 2021 trong quý đầu tiên của năm 2022, sau đó bước vào giai đoạn điều chỉnh giảm mạnh bắt đầu từ tháng 4, và có những bước phục hồi trong giai đoạn giữa đến cuối năm. Năm 2022 được xem như liều thuốc thử đối với các cổ phiếu ngành ngân hàng, vì thế ta có thể đánh giá các mã cổ phiếu này dựa trên diễn biến trên thị trường trong năm 2022.



Hình 5.4: Diễn biến giá CTG 2.

Với mã CTG, đúng như kết quả chung của thị trường chứng khoán năm 2022, đầu năm 2022 chứng kiến sự gia tăng theo đà từ năm 2021, sau đó điều chỉnh tăng vào tháng 5 và tháng 8. Đến cuối năm có sự sụt giảm đáng kể nhưng từ đó kết quả đã khả quan hơn tính đến thời điểm hiện tại như trên ta đã đề cập. Bên cạnh đó, biên độ giảm của mã cổ phiếu này là ít nhất trong cả 3 cổ phiếu được so sánh và có các nhịp hồi liên tục, cho thấy khả năng phục hồi tốt theo thị trường. Điều này cho thấy nếu có những chuyển biến trên thị trường chứng khoán thì mã chứng khoán này vẫn được các nhà đầu tư kỳ vọng sẽ phục hồi trong tương lai.



Hình 5.5: Diễn biến giá TCB 2.



Hình 5.6: Diễn biến giá TPB 2.

Cả hai mã TCB và TPB nhìn chung đều có xu hướng như kết quả chung của thị trường. Nhưng khác với CTG, có thể thấy hai mã cổ phiếu này có sự sụt giảm lớn trong giai đoạn này. TCB và TPB lần lượt giảm 45% và 50%, một tỷ lệ điều chỉnh đáng kể. Tuy vậy khả năng phục hồi của TCB vẫn tốt hơn như kết quả

có được ở trên. Qua đó, ta dễ dàng thấy được, dẫu trong giai đoạn khó khăn chung của cả thị trường chứng khoán nhưng cổ phiếu CTG vẫn đảm bảo mức giảm không đáng kể và phục hồi tốt. Trong khi hai cổ phiếu còn lại chưa quay lại được đỉnh cũ trong giai đoạn đầu năm 2021.

Tuy chỉ xem xét qua ba cổ phiếu đại diện cho ba cụm đã được phân, nhưng kết quả có vẫn khá khả quan khi có sự khác biệt giữa các mã cổ phiếu của mỗi cụm. Nhưng bên cạnh đó vẫn còn một số lưu ý, các cổ phiếu trong mỗi cụm có thể cho ra kết quả đầu tư trái ngược nhau, ví dụ cụ thể là mã cổ phiếu TPB ở cụm 2. Như kết quả đầu tư theo phương án hai, nếu theo giá trung bình của cả cụm ta thu được lợi nhuận âm, tuy nhiên thực tế lợi nhuận trong năm của mã này là dương.

Bên cạnh đó, thị trường chứng khoán Việt Nam được cho là thị trường cận biên, hay thị trường kém hiệu quả theo lý thuyết thị trường hiệu quả. Bằng chứng trong nhiều năm gần đây, đã xảy ra nhiều trường hợp sai phạm trên thị trường chứng khoán Việt Nam. Vì thế chỉ dựa vào phân tích cơ bản theo những chỉ số tài chính, hay kết hợp với dữ liệu quá khứ phục vụ cho phân tích kỹ thuật chỉ mang tính chủ quan. Do đó tác giả khi quan sát diễn biến giá trên thị trường của các mã cổ phiếu đại diện chỉ mang tính chất tham khảo kỳ vọng của các nhà đầu tư vào các mã cổ phiếu này. Những mà cổ phiếu được kỳ vọng là những mã cổ phiếu có giá trị cốt lõi cao, được mong đợi sẽ có kết quả hoạt động kinh doanh tốt. Song những cổ phiếu có diễn biến giá lịch sử tốt cũng chưa chắc chắn tăng trưởng vì sự tăng trưởng trong quá khứ là kết quả của các chiêu trò nhằm thổi giá và bán hàng loạt của các nhà đầu cơ.

Với các phương pháp đánh giá như trên, có thể nói kết quả của thuật toán phân cụm K-Means mang lại nhiều giá trị cho các nhà đầu tư trong việc phân tích các mã cổ phiếu ngành ngân hàng cũng như đưa ra quyết định đầu tư. Các mã sau khi được phân cụm có mối liên hệ chặt chẽ trong cùng một cụm cũng như có sự

khác nhau lớn giữa các cụm với nhau. Song vẫn còn một số ngoại lệ tuỳ thuộc vào phương thức đánh giá kết quả phân cụm.

5.4 Ứng dụng phân cụm cổ phiếu bằng thuật toán K-Means

Trong phạm vi bài khoá luận này, tác giả sử dụng các mã *cổ phiếu* thuộc ngành ngân hàng để phân cụm dữ liệu và có được kết quả khả quan, vì vậy việc ứng dụng thuật toán phân cụm K-Means cho việc phân tách dữ liệu của nhiều ngành nghề khác nhau là hoàn toàn khả dụng. Với việc những thuật toán này ngày càng được cải tiến và có được kết quả tốt hơn, việc ứng dụng thuật toán này trong thực tế sẽ mang lại lợi ích nhất định cho các nhà đầu tư.

5.5 Lợi ích việc phân cụm cổ phiếu bằng thuật toán K-Means

Đối với các nhà đầu tư, đây được xem là một phương thức để chọn lọc những cổ phiếu thuộc bất kỳ ngành nghề nào mà mình mong muốn. Các nhà đầu tư có thể chủ động trong việc đưa ra quyết định đầu tư mà không cần sự giới thiệu từ các nhà môi giới.

Đối với các công ty chứng khoán, việc phân cụm cổ phiếu bằng thuật toán K-Means kết hợp với lượng thông tin và kiến thức có sẵn sẽ tạo ra những phương án đầu tư thích hợp, từ đó đưa ra những gợi ý phù hợp với từng nhu cầu đầu tư của khách hàng hơn. Bên cạnh đó các công ty chứng khoán cũng sẽ có lợi khi sử dụng kết quả phân cụm để đưa ra quyết định đầu tư vào các cổ phiếu trên thị trường, tạo niềm tin lớn với khách hàng của mình.

Cuối cùng là các nhà phát triển, việc tích hợp các công cụ phân tích vào các phần mềm giao dịch chứng khoán của các công ty là vô cùng lớn. Các nhà phát triển phần mềm có thể thêm vào những công cụ như thế này giúp cải thiện chất lượng phần mềm ngày càng tốt hơn. Không những trong lĩnh vực chứng khoán, thuật toán này cũng có thể được sử dụng cho nhiều ngành nghề khác nhau mà các công ty cần xem xét sử dụng.

CHƯƠNG 6. KẾT LUẬN VÀ KHUYÊN NGHỊ

6.1 Kết luận

Với những gì có được trong phạm vi bài khoá luận, tác giả đưa ra những kết luận, bên cạnh đó là những hạn chế của thuật toán trong việc ứng dụng thực tế cũng như đưa ra những hướng phát triển mới cho đề tài.

Đầu tiên, bài khoá luận cung cấp thông tin giúp chúng ta hiểu rõ hơn về thuật toán phân cụm K-Means và ứng dụng của thuật toán trong thực tế. Hiện nay, có rất nhiều các công cụ hỗ trợ khác nhau phục vụ nhu cầu phân tích, đầu tư của con người, phương pháp phân cụm K-Means là một trong số đó. Với việc hiểu rõ được nội dung của thuật toán và ứng dụng vào thực tế đã mang lại nhiều lợi ích cho các nhà đầu tư. Cũng thông qua bài toán này, chúng ta còn thấy được tiềm năng của phương pháp phân cụm K-Means vì tính đa dụng và thân thiện với nhiều loại dữ liệu khác nhau tùy theo từng lĩnh vực ứng dụng.

Khi được ứng dụng cụ thể vào việc phân cụm cổ phiếu ngành ngân hàng tại Việt Nam, phương pháp phân cụm K-Means đã cho được kết quả tích cực. Kết quả đã cho thấy được điều đó, khi có sự khác nhau về quy mô cũng như giá trung bình giữa các cụm mặc dù tác giả không sử dụng những chỉ số liên quan đến các giá trị này. Trong mỗi cụm, các cổ phiếu có sự tương đồng mạnh mẽ. Điều này đáp ứng được hai tiêu chí khi đánh giá kết quả phân cụm là đánh giá trong và đánh giá ngoài. Cho thấy kết quả có được sau khi phân cụm là khả quan. Tiếp theo đó, khi tiến hành hình thành danh mục đầu tư đại diện cho các cụm và đầu tư theo hai phương án khác nhau cũng đã cho thấy được kết quả khác biệt rõ ràng và hợp lý. Tất cả cho thấy tính khả dụng của thuật toán khi áp dụng vào thực tế và cụ thể là ứng dụng vào việc phân cụm cổ phiếu phục vụ cho phân tích và đầu tư.

Cuối cùng khi tiến hành quan sát diễn biến giá trên thị trường của một số cổ phiếu thuộc các cụm cũng mang lại một số kết quả nhất định. Trong giai đoạn

khó khăn nhất thị trường gần như có xu hướng giảm đáng kể. Khi VN-INDEX giảm 500 điểm, kinh tế khó khăn vì lạm phát và chiến tranh, cổ phiếu CTG thuộc nhóm 1 vẫn cho thấy sức mạnh của mình khi chỉ sụt giảm gần 20. Đây là con số nhỏ khi so sánh với hai cổ phiếu còn lại được quan sát. Do đó có thể khẳng định có sự khác biệt giữa các cụm, cụ thể là các cổ phiếu thuộc mỗi cụm khi quan sát diễn biến giá trên thị trường. Bên cạnh đó, khi kiểm chứng lại kết quả đầu tư dài hạn khi nắm giữ ba cổ phiếu được quan sát từ tháng 9 năm 2022 đến tháng 9 năm 2023 cũng cho kết quả gần giống với kết quả đầu tư theo phương án hai. Ngoại trừ trường hợp của mã cổ phiếu TPB, khi vẫn mang lại lợi nhuận trong khi kết quả đầu tư của cả cụm là ngược lại. Điều này có thể được giải thích bởi kết quả phân cụm dựa trên những chỉ số tài chính của các ngân hàng. Việc các cổ phiếu được gom vào một cụm cho thấy các ngân hàng này có bộ chỉ số ở mức tương đồng với nhau, do đó việc khác nhau về kết quả đầu tư hay diễn biến giá trên thị trường của mỗi mã cổ phiếu trong cụm khác nhau là điều có thể xảy ra. Từ đó suy ra, kết quả phân cụm là một cụm tương đồng nhau về các dữ liệu đầu vào. Do đó, các nhà đầu tư khi sử dụng kết quả phân cụm cần chú ý mục đích của mình. Trách nhiệm của họ là đưa ra quyết định đầu tư cẩn kêt hợp nhiều thông tin và kiến thức để có thể có được kết quả đầu tư tốt nhất.

Bên cạnh những kết quả có được trong bài nghiên cứu này, thuật toán K-Means còn được sử dụng để đa dạng hóa danh mục đầu tư. Một danh mục đầu tư nếu không được đa dạng hóa sẽ có rủi ro lớn khi có sự biến động trên thị trường qua đó không đảm bảo được tỷ suất sinh lợi kỳ vọng cho nhà đầu tư. Do đó việc đa dạng hóa danh mục đầu tư là việc vô cùng quan trọng trong quản lý rủi ro tài chính. Mục tiêu của đa dạng hóa danh mục đầu tư là giảm thiểu rủi ro. Thuật toán phân cụm K-Means sẽ đóng góp phần quan trọng trong việc xây dựng danh mục đầu tư đa dạng. Đa dạng hóa danh mục đầu tư là phân chia tỷ trọng đầu tư vào

nhiều loại tài sản khác nhau nhằm giảm thiểu rủi ro. Bằng cách kết hợp nhiều phương pháp phân tích khác nhau để ta sẽ tìm được các ngành nghề có mối tương quan phù hợp với mục tiêu đa dạng hóa danh mục đầu tư, khi có biến động xấu trên thị trường ngành nghề này sẽ hỗ trợ ngành nghề bị ảnh hưởng, sau đó sử dụng thuật toán phân cụm K-Means lựa chọn những cổ phiếu tốt những thuộc những nhóm ngành đã có được từ đó xây dựng danh mục đầu như mục tiêu đã đề ra.

6.2 Hạn chế

Tuy có nhiều kết quả tích cực, nhưng bài khoá luận vẫn còn tồn đọng nhiều thiếu sót và hạn chế.

Dữ liệu được sử dụng trong bài khoá luận được thu thập từ một trang thứ ba, chưa được trực tiếp tính toán từ kết quả báo cáo tài chính của những ngân hàng ở Việt Nam. Quy trình thu thập dữ liệu chưa được tự động hoá mặc dù có sử dụng ngôn ngữ lập trình hỗ trợ, các bước tổng hợp vẫn còn thủ công và chưa tối ưu.

Kết quả khi thực hiện đầu tư theo các phương án là kết quả chủ quan với số quan sát nhỏ, do đó chưa đủ ý nghĩa để đưa ra quyết định đầu tư.

6.3 Kiến nghị và hướng phát triển

Từ những kết quả khả quan thu được bên cạnh những hạn chế của khoá luận, tác giả đề xuất những hướng phát triển mới nhằm tăng tính ứng dụng của thuật toán phân cụm K-Means vào thực tế.

- Khi tiến hành sử dụng thuật toán phân cụm K-Means vào thực tế, cần sử dụng dữ liệu tốt nhất, đa dạng hóa các dữ liệu đầu vào để có được kết quả tối ưu.

- Ứng dụng thuật toán vào nhiều lĩnh vực khác nhau, không chỉ phân cụm cổ phiếu mà còn nhiều ngành nghề có thể sử dụng thuật toán này. Tuỳ theo mục tiêu phân cụm mà có thể thay đổi dữ liệu phù hợp để có được kết quả tốt nhất.

Hiện nay, bên cạnh phương pháp phân cụm K-Means vẫn có một số phương pháp phân cụm được sử dụng phổ biến có thể kể đến như phân tích phân biệt hay mô hình Merton. Trong khi phương pháp phân tích phân biệt được sử dụng để phân biệt các quan sát trong bộ dữ liệu dựa trên các biến độc lập tương tự như thuật toán K-Means, thì mô hình Merton đi sâu hơn vào phân tích khi được dùng để đánh giá rủi ro tín dụng của một khoản nợ công ty. Cụ thể nếu áp dụng trực tiếp vào bài nghiên cứu, việc phân tích phân biệt có thể phân tách bộ dữ liệu thành 2 nhóm khác biệt tuỳ theo mục tiêu của người sử dụng, còn mô hình Merton sẽ được áp dụng để đánh giá mức độ rủi ro tín dụng ngân hàng. Nhưng khi thực hiện riêng những phương pháp này đều có những hạn chế nhất định.

Đối với phân tích phân biệt, phương pháp này có một số giả định như dữ liệu đầu vào phải tuân theo phân phối chuẩn, hay phương sai của các biến độc lập bằng nhau. Do đó nếu không đảm bảo được những giả định này có khả năng sẽ cho ra kết quả không tốt. Phương pháp này còn giới hạn số lượng nhóm, phân tích phân biệt thông thường chỉ áp dụng cho việc phân biệt giữa hai nhóm, việc sử dụng phân tích phân biệt cho nhiều nhóm sẽ trở nên phức tạp và không hiệu quả. Bên cạnh đó phương pháp này không xử lý được dữ liệu dạng chuỗi hoặc thời gian, điều này hạn chế đối với một số ứng dụng.

Mô hình Merton cũng không ngoại lệ khi tồn tại một số nhược điểm. Mô hình này vẫn tồn tại giả định về phân phối chuẩn. Mô hình Merton dựa trên giả định rằng giá trị tài sản của công ty tuân theo phân phối chuẩn. Tuy nhiên điều này là hoàn toàn không đúng với thị trường, ngày ngày một thuật ngữ mới được thay thế và rất thường được sử dụng là “không chắc chắn”, việc giả định tài sản

tuân theo phân phối chuẩn là nhược điểm lớn của mô hình. Mô hình còn bỏ qua rủi ro của thị trường, giả định cơ cấu nợ của công ty không thay đổi. Mô hình này chỉ dùng lại ở việc đánh giá mức độ rủi ro tín dụng dựa trên xác suất chứ không dự đoán sự xảy ra của sự vỡ nợ.

Có thể thấy, bất kỳ phương pháp hay mô hình nào được sử dụng hiện nay đều có những ưu và nhược điểm nhất định. Do đó, xu hướng ngày nay các nhà đầu tư thường kết hợp nhiều phương pháp, mô hình phục vụ cho nhiều mục đích khác nhau. Để có thể đưa ra được quyết định đầu tư, hay xây dựng một danh mục đầu tư đa dạng như đã đề cập ở trên, ngoài phương pháp phân cụm K-Means cần phối hợp với phương pháp phân tích phân biệt hay mô hình Merton trước khi thực hiện phân cụm để có được kết quả phân cụm khách quan và dễ dàng so sánh. Cụ thể, có thể sử dụng phương pháp phân tích phân biệt để lựa chọn những cổ phiếu phù hợp và loại bỏ những cổ phiếu không phù hợp trước khi phân cụm, hay sử dụng mô hình Merton loại những cổ phiếu có mức độ rủi ro tín dụng cao ra khỏi mô hình trước khi phân cụm theo thuật toán K-Means. Việc sử dụng nhiều mô hình như thế sẽ làm cho kết quả phân cụm trở nên khách quan và dễ dàng so sánh hơn. Vì những dữ liệu đầu vào của bài khoá luận có chứa nhiều chỉ số đánh giá rủi ro, nguy cơ phá sản của ngân hàng nên sau khi phân cụm theo thuật toán K-Means ta cũng có thể áp dụng một số phương pháp đánh giá khả năng phá sản của ngân hàng ngoài mô hình Merton. Có thể kể đến như mô hình Z-Score của Edward Altman được ông nghiên cứu và áp dụng cho không chỉ trên các công ty sản xuất mà còn được sử dụng với những công ty dịch vụ với nhiều biến thể khác nhau. Hay mô hình Z-Score của Roy được đề xuất đầu tiên năm 1952. Bên cạnh đó, có thể sử dụng mô hình CAMELS để có cái nhìn cụ thể hơn về mỗi ngân hàng mà nhà đầu tư quan tâm khi đưa ra quyết định đầu tư. Có thể thấy có rất nhiều phương pháp, mô hình có thể được kết hợp để bù trừ những khuyết điểm cho nhau, vì thế mỗi phương pháp hay mô hình chỉ dừng lại ở mức độ tham khảo

nhưng để có thể đưa ra quyết định đầu tư cuối cùng cần có sự phối hợp chặt chẽ và hợp lý giữa các phương pháp với nhau từ đó cho ra kết quả đầu tư tối ưu nhất.

PHỤ LỤC

TÀI LIỆU THAM KHẢO

Keshavarz Haddadha, H., Alipour, H., & Kheradyar, S. (2023). Designing an Optimal Model Regarding Early Warning System of Bankruptcy of Banks in Iran Application of Grounded Theory and Econometric Models. *Environmental Energy and Economic Research*, 7(2), 1-20.

Nanda, S. R., Mahanty, B., & Tiwari, M. K. (2010). Clustering Indian stock market data for portfolio management. *Expert Systems with Applications*, 37(12), 8793-8798.

Shin, H. W., & Sohn, S. Y. (2004). Segmentation of stock trading customers according to potential value. *Expert systems with applications*, 27(1), 27-33.

Aslam, B., Bhuiyan, R. A., & Zhang, C. (2023). Portfolio Construction with K-Means Clustering Algorithm Based on Three Factors. In *MATEC Web of Conferences* (Vol. 377, p. 02006). EDP Sciences.

Gunsel, N. (2005). Financial ratios and the probabilistic prediction of bank failure in North Cyprus. *Editorial Advisory Board e*, 18(2), 191-200.

Khaddafi, M., Heikal, M., & Nandari, A. (2017). Analysis Z-score to predict bankruptcy in banks listed in Indonesia stock exchange. *International Journal of Economics and Financial Issues*, 7(3), 326-330.

Grönholm, R. (2023). Performance of clustering-based stock portfolios: case: Exploratory study of k-means clustering for S&P500 in 2010-2022 data using combinations of selected key figures.

Korzeniewski, J. (2018). Efficient stock portfolio construction by means of clustering.

Dũng, N. V., & Chi, H. L. (2022). Phân tích cơ bản, chỉ số tài chính và lợi suất cổ phiếu: nghiên cứu thực nghiệm trên thị trường chứng khoán Việt Nam. *Tạp chí Quản lý Kinh tế Quốc tế (Journal of International Economics and Management)*, (147), 1-16.

Nguyễn, Đ. H. (2014). MÔ HÌNH HAI GIAI ĐOẠN DỰ ĐOÁN GIÁ CỔ PHIẾU VỚI K-MEANS VÀ FUZZY-SVM.

Phương, N. T. T. (2016). Ảnh hưởng của các chỉ số tài chính đến khả năng sinh lợi của hệ thống ngân hàng khu vực Châu Á-Thái Bình Dương.

Quyên, N. T. P. (2019). Tác động của các yếu tố rủi ro tài chính đến nguy cơ phá sản ngân hàng thương mại Việt Nam.

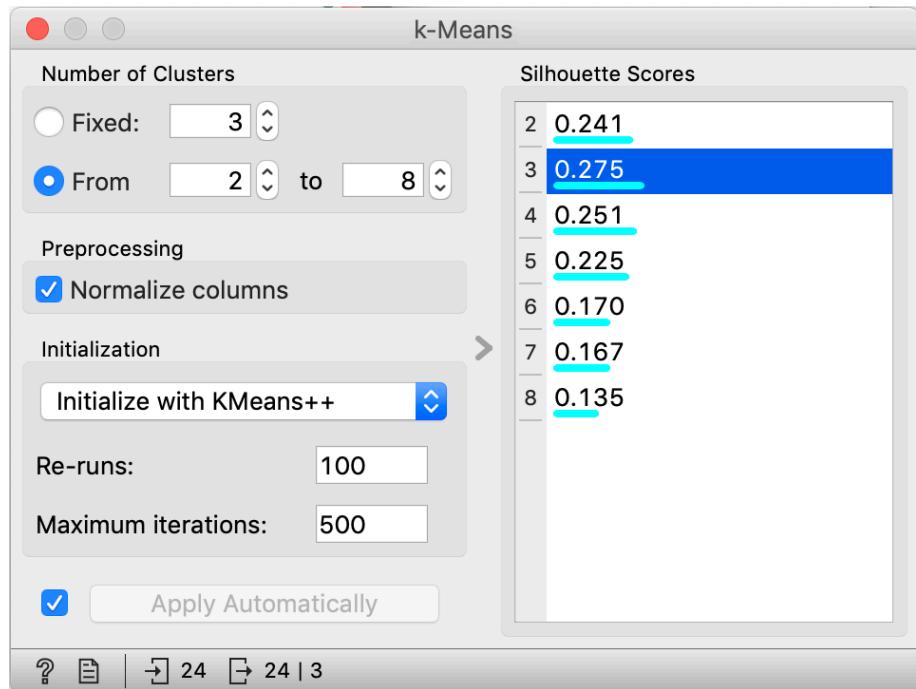
HÌNH ẢNH

	A	B	C	D	E	F	G	H	I	J
1	STT	Mã	YOEA (%)	NIM (%)	COF (%)	LAR (%)	LDR (%)	CIR (%)	LLRL (%)	LLRNPL (%)
2	1	ABB	7.17	2.58	7.83	54.4	96.04	51.29	1.53	43.39
3	2	ACB	8.64	4.4	4.72	68.8	100.37	37.41	1.15	159.27
4	3	BAB	9.62	2.17	7.94	71.41	91.68	63.19	1.11	203.77
5	4	BID	7.14	2.83	4.54	76.67	105.4	34.61	2.43	216.92
6	5	BVB	8.34	1.96	10.59	65.85	106.98	76.7	1.56	52.27
7	6	CTG	7.36	3.03	4.52	73.08	103.76	29.22	2.15	188.37
8	7	EIB	8.07	3.02	5.55	69.28	85.46	50.73	1.2	55.8
9	8	HDB	10.6	4.88	5.89	59.38	92.8	38.01	1.32	70.42
10	9	KLB	9.69	2.98	10.75	55.27	78.11	57.88	1.21	68.62
11	10	LPB	9.32	3.61	6.09	72.35	113.06	42.83	1.75	142.11
12	11	MBB	9.17	5.57	4.04	64.26	108.97	32.58	2.25	238.03
13	12	MSB	8.32	4.35	4.31	57.44	108.19	38.44	1.63	69.17
14	13	NAB	9.34	3.31	6.38	64.54	88.89	46.69	1.03	77.22
15	14	OCB	9.05	3.85	5.88	60.38	115.5	32.54	1.51	59.24
16	15	PGB	8.1	3.33	8.46	64.38	96.87	50.18	1	37.99
17	16	SHB	10.21	3.82	6.49	69.67	99.52	22.35	1.77	65.03
18	17	SSB	8.09	3.13	5.38	66.02	131.37	40.3	1.71	98.9
19	18	STB	9.43	4.3	5.35	74.01	91.8	42.95	1.38	130.97
20	19	TCB	8.45	4.49	4.41	63.69	122.15	33.9	1.24	157.34
21	20	TPB	8.51	3.69	4.94	51.58	88.94	43.02	1.35	135
22	21	VBB	7.84	1.66	6.41	59.23	84.72	63.34	1.06	26.93
23	22	VCB	6.57	3.56	3.25	69.1	88.76	30.64	3.21	316.86
24	23	VIB	9.54	4.7	8.02	61.98	114.31	32.43	1.42	53.89
25	24	VPB	11.21	6.3	5.47	66.05	126.07	28.61	2.8	54.4

Hình 1: Dữ liệu thu thập.

	A	B	C	D	E	F	G	H	I	J
1	STT	Mã	YOEA (%)	NIM (%)	COF (%)	LAR (%)	LDR (%)	CIR (%)	LLRL (%)	LLRNPL (%)
2	1	ABB	-1.382241	-0.964558	0.87621629	-1.6045838	-0.4034928	0.67473638	-0.1485759	-0.9249757
3	2	ACB	-0.0887274	0.68121911	-0.7302879	0.58538602	-0.0923399	-0.3899102	-0.8095576	0.60575434
4	3	BAB	0.77361496	-1.33531	0.93303798	0.98231805	-0.7168016	1.58750972	-0.8791346	1.19358214
5	4	BID	-1.4086392	-0.7384897	-0.8232688	1.78226535	0.269115	-0.6046804	1.41690697	1.36728856
6	5	BVB	-0.3527098	-1.5252074	2.30192416	0.13674638	0.3826535	2.62377593	-0.0963932	-0.8076743
7	6	CTG	-1.2150521	-0.5576351	-0.8336	1.23629371	0.15126491	-1.018113	0.92986785	0.9901541
8	7	EIB	-0.5902939	-0.5666778	-0.3015424	0.65838502	-1.1637697	0.63178234	-0.7225863	-0.7610444
9	8	HDB	1.63595733	1.11527022	-0.1259117	-0.8472192	-0.6363186	-0.343888	-0.5138553	-0.5679198
10	9	KLB	0.83521085	-0.6028488	2.38457389	-1.4722731	-1.6919393	1.18021338	-0.7051921	-0.5916971
11	10	LPB	0.50963261	-0.0331567	-0.0225995	1.12527441	0.81956116	0.02582356	0.23409767	0.37907738
12	11	MBB	0.37764143	1.7392187	-1.0815492	-0.1050628	0.52565453	-0.7603888	1.10381039	1.6461435
13	12	MSB	-0.3703086	0.63600545	-0.9420778	-1.1422568	0.46960388	-0.3109054	0.02536662	-0.5844318
14	13	NAB	0.52723143	-0.3044386	0.1272031	-0.06248	-0.9172905	0.32189962	-1.0182886	-0.4780944
15	14	OCB	0.27204848	0.18386887	-0.1310773	-0.695138	0.99489911	-0.7634569	-0.1833644	-0.7156033
16	15	PGB	-0.5638956	-0.2863532	1.20164961	-0.086813	-0.3438492	0.58959533	-1.0704714	-0.9963076
17	16	SHB	1.29278027	0.15674068	0.18402479	0.7176967	-0.1534207	-1.545067	0.26888618	-0.6391196
18	17	SSB	-0.5726951	-0.4672078	-0.3893577	0.16260019	2.13531434	-0.1682367	0.16452066	-0.19171
19	18	STB	0.60642614	0.59079179	-0.4048545	1.37772926	-0.7081785	0.035028	-0.4094897	0.23192228
20	19	TCB	-0.2559162	0.76260369	-0.8904217	-0.1917491	1.47276686	-0.65914	-0.6530093	0.58025979
21	20	TPB	-0.2031198	0.03918517	-0.6166445	-0.20334529	-0.9136975	0.04039725	-0.4616725	0.28515703
22	21	VBB	-0.7926804	-1.7964893	0.14269993	-0.8700314	-1.2169459	1.59901527	-0.9661059	-1.1424059
23	22	VCB	-1.9102057	-0.0783703	-1.4896323	0.6310104	-0.9266323	-0.9091938	2.77365881	2.68745733
24	23	VIB	0.70321967	0.95250106	0.97436285	-0.451808	0.90938593	-0.7718943	-0.3399127	-0.7862747
25	24	VPB	2.17272146	2.3993381	-0.3428672	0.16716263	1.75445733	-1.0649022	2.06049438	-0.7795378

Hình 2: Dữ liệu sau khi được chuẩn hóa.



Hình 3: Kết quả lựa chọn số cụm.

	A	B	C	D	E	F	G	H	I	J
Enter		YOEA (%)	NIM (%)	COF (%)	LAR (%)	LDR (%)	CIR (%)	LLRL (%)	LLRNPL (%)	cum
2	1	-1.382241	-0.964558	0.87621629	-1.6045838	-0.4034928	0.67473638	-0.1485759	-0.9249757	2
3	2	-0.0887274	0.68121911	-0.7302879	0.58538602	-0.0923399	-0.3899102	-0.8095576	0.60575434	3
4	3	0.77361496	-1.33531	0.93303798	0.98231805	-0.7168016	1.58750972	-0.8791346	1.19358214	2
5	4	-1.4086392	-0.7384897	-0.8232688	1.78226535	0.269115	-0.6046804	1.41690697	1.36728856	1
6	5	-0.3527098	-1.5252074	2.30192416	0.13674638	0.3826535	2.62377593	-0.0963932	-0.8076743	2
7	6	-1.2150521	-0.5576351	-0.8336	1.23629372	0.15126491	-1.018113	0.92986785	0.9901541	1
8	7	-0.5902939	-0.5666778	-0.3015424	0.65838502	-1.1637697	0.63178234	-0.7225863	-0.7610444	2
9	8	1.63595733	1.11527022	-0.1259117	-0.8472192	-0.6363186	-0.343888	-0.5138553	-0.5679198	3
10	9	0.83521085	-0.6028488	2.38457389	-1.4722731	-1.6919393	1.18021338	-0.7051921	-0.5916971	2
11	10	0.50963261	-0.0331567	-0.0225995	1.12527441	0.81956117	0.02582356	0.23409767	0.37907738	3
12	11	0.37764143	1.7392187	-1.0815492	-0.1050628	0.52565453	-0.7603888	1.10381039	1.6461435	3
13	12	-0.3703086	0.63600545	-0.9420778	-1.1422568	0.46960388	-0.3109054	0.02536662	-0.5844318	3
14	13	0.52723143	-0.3044386	0.1272031	-0.06248	-0.9172905	0.32189962	-1.0182886	-0.4780944	2
15	14	0.27204849	0.18386888	-0.1310773	-0.695138	0.99489911	-0.7634569	-0.1833644	-0.7156033	3
16	15	-0.5638956	-0.2863532	1.20164961	-0.086813	-0.3438492	0.58959533	-1.0704714	-0.9963076	2
17	16	1.29278027	0.15674068	0.18402479	0.7176967	-0.1534207	-1.545067	0.26888618	-0.6391196	3
18	17	-0.5726951	-0.4672078	-0.3893577	0.16260019	2.13531434	-0.1682367	0.16452066	-0.19171	3
19	18	0.60642614	0.59079179	-0.4048545	1.37772926	-0.7081785	0.035028	-0.4094897	0.23192228	3
20	19	-0.2559162	0.76260369	-0.8904217	-0.1917491	1.47276686	-0.65914	-0.6530093	0.58025979	3
21	20	-0.2031198	0.03918517	-0.6166445	-2.0334529	-0.9136975	0.04039725	-0.4616725	0.28515703	2
22	21	-0.7926804	-1.7964893	0.14269993	-0.8700314	-1.2169459	1.59901527	-0.9661059	-1.1424059	2
23	22	-1.9102057	-0.0783703	-1.4896323	0.6310104	-0.9266323	-0.9091938	2.77365882	2.68745733	1
24	23	0.70321967	0.95250106	0.97436285	-0.451808	0.90938593	-0.7718943	-0.3399127	-0.7862747	3
25	24	2.17272146	2.3993381	-0.3428672	0.16716263	1.75445733	-1.0649022	2.06049439	-0.7795378	3

Hình 4: Kết quả sau khi phân cụm.