ĐẠI HỌC QUỐC GIA TP.HCM

**TRƯỜNG ĐẠI HỌC KINH TẾ - LUẬT**

**BÁO CÁO TỔNG KẾT**

**ĐỀ TÀI NGHIÊN CỨU KHOA HỌC SINH VIÊN**

**NĂM 2023**

**ỨNG DỤNG MACHINE LEARNING DỰ ĐOÁN KHẢ NĂNG TRẢ CỔ TỨC CỦA DOANH NGHIỆP VIỆT NAM**

**Lĩnh vực khoa học:** Tài chính – Kinh tế

**Chuyên ngành:** Công nghệ - Tài chính

**Nhóm nghiên cứu:**

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **TT** | **Họ tên** | **MSSV** | **Đơn vị** | **Nhiệm vụ** | **Điện thoại** | **Email** |
|  | Nguyễn Trung Trực | K204141936 | Khoa Tài chính – Ngân hàng | Nhóm trưởng | 0336843682 | trucnt20414c@st.uel.edu.vn |
|  | Trần Thu Hà | K204141913 | Khoa Tài chính – Ngân hàng | Tham gia | 0931985145 | hatt20414c@st.uel.edu.vn |
|  | Nguyễn Việt Hùng | K204141918 | Khoa Tài chính – Ngân hàng | Tham gia | 0947135262 | hungnv20414c@st.uel.edu.vn |
|  | Nguyễn Văn Thành | K204140640 | Khoa Tài chính – Ngân hàng | Tham gia | 0343984998 | thanhnv20414c@st.uel.edu.vn |
|  | Nguyễn Ngọc Phương Nhu | K204141924 | Khoa Tài chính – Ngân hàng | Tham gia | 0865849845 | nhunnp20414c@st.uel.edu.vn |

ĐẠI HỌC QUỐC GIA TP.HCM

**TRƯỜNG ĐẠI HỌC KINH TẾ - LUẬT**

**————————**

**BÁO CÁO TỔNG KẾT**

**ĐỀ TÀI NGHIÊN CỨU KHOA HỌC SINH VIÊN NĂM 2023**

**Tên đề tài:**

**ỨNG DỤNG MACHINE LEARNING DỰ ĐOÁN KHẢ NĂNG TRẢ CỔ TỨC CỦA DOANH NGHIỆP VIỆT NAM**

|  |  |  |
| --- | --- | --- |
| **Đại diện nhóm nghiên cứu** | **Giảng viên hướng dẫn** | **Chủ tịch hội đồng** |
| *(Ký, họ tên)* | *(Ký, họ tên)* | *(Ký, họ tên)* |

## 

**Lãnh đạo Khoa/Bộ môn/Trung tâm**

*(Ký, họ tên)*

**LỜI CẢM ƠN**

Với cương vị là sinh viên của trường Đại học Kinh tế - Luật, chúng em nhận thức được rằng bên cạnh việc học tập và trau dồi kiến thức thì hoàn thành Đề tài nghiên cứu khoa học là điều vô cùng quan trọng và cần thiết. Sau bao cố gắng và nỗ lực để hoàn thành đề tài, chúng em biết rằng bất kỳ sự thành công nào cũng đều gắn liền với sự hỗ trợ, sự giúp đỡ của những người xung quanh dù ít hay nhiều, dù trực tiếp hay gián tiếp. Vì thế chúng em xin gửi lời cảm ơn chân thành đến Ban lãnh đạo trường Đại học Kinh tế - Luật, đã tạo điều kiện cho chúng em được thực hiện nghiên cứu khoa học khi vẫn còn là sinh viên.

Em xin chân thành cảm ơn Ban Lãnh đạo Khoa Tài chính – Ngân hàng, đặc biệt là Giáo viên hướng dẫn của nhóm TS Phan Huy Tâm đã hướng dẫn và tận tình chỉ bảo giúp chúng em hoàn thành tốt đề tài nghiên cứu này. Trong thời gian vừa qua, nhóm chúng em không chỉ nhận được những kiến thức bổ ích về chuyên môn mà còn ở rất nhiều lĩnh vực khác. Những trải nghiệm và kiến thức quý báu đó không chỉ giúp chúng em hoàn thành tốt đề tài nghiên cứu mà còn là hành trang theo chúng em trong suốt thời gian học tập và làm việc sau này!

Bởi vì kiến thức và kinh nghiệm còn hạn chế nên đề tài nghiên cứu của nhóm chúng em còn nhiều thiếu sót, kính mong nhận được sự đánh giá, góp ý của quý thầy cô!

**Nhóm nghiên cứu**

**MỤC LỤC**

[**CHƯƠNG 1. TỔNG QUAN NGHIÊN CỨU** 1](#_Toc130803822)

[**1.1. LÝ DO CHỌN ĐỀ TÀI** 1](#_Toc130803823)

[**1.2. MỤC TIÊU NGHIÊN CỨU** 1](#_Toc130803824)

[**1.2.1. Mục tiêu chung** 1](#_Toc130803825)

[**1.2.2. Mục tiêu cụ thể** 2](#_Toc130803826)

[**1.3. ĐỐI TƯỢNG VÀ PHẠM VI CỦA NGHIÊN CỨU** 2](#_Toc130803827)

[**1.3.1. Đối tượng nghiên cứu** 2](#_Toc130803828)

[**1.3.2. Phạm vi nghiên cứu** 2](#_Toc130803829)

[**1.4. CÂU HỎI NGHIÊN CỨU** 3](#_Toc130803830)

[**1.5. PHƯƠNG PHÁP NGHIÊN CỨU** 3](#_Toc130803831)

[**1.6. CẤU TRÚC ĐỀ TÀI** 4](#_Toc130803832)

[**CHƯƠNG 2: CƠ SỞ LÝ THUYẾT** 5](#_Toc130803833)

[**2.1. Cơ sở lý thuyết** 5](#_Toc130803834)

[**2.1.1. Signalling theory** 5](#_Toc130803835)

[**2.1.2. Agency theory** 5](#_Toc130803836)

[**2.1.3. Pecking order theory** 6](#_Toc130803837)

[**2.1.4. Dự đoán cổ tức** 6](#_Toc130803838)

[**2.2. Các bài nghiên cứu trước đây** 6](#_Toc130803839)

[**2.3. Phát triển giả thuyết** 7](#_Toc130803840)

[**2.3.1. Mối quan hệ giữa DIV và tỷ số lợi nhuận trên tài sản (ROA)** 7](#_Toc130803841)

[**2.3.2. Mối quan hệ giữ DIV và lợi nhuận trên vốn (ROE)** 8](#_Toc130803842)

[**2.3.3. Mối quan hệ giữa DIV và dòng tiền tự do (FCF)** 8](#_Toc130803843)

[**2.3.4. Mối quan hệ giữa DIV và đòn bẩy (Leverage)** 9](#_Toc130803844)

[**2.3.5. Mối quan hệ giữa DIV và thanh khoản (Liquidity)** 9](#_Toc130803845)

[**2.3.6. Mối quan hệ giữa DIV và MTB** 10](#_Toc130803846)

[**2.3.7. Mối quan hệ giữa DIV và Capitalization** 10](#_Toc130803847)

[**2.3.8 Mối quan hệ giữa DIV và Company age** 11](#_Toc130803848)

[**2.3.9. Mối quan hệ giữa DIV và Public age** 11](#_Toc130803849)

[**CHƯƠNG 3: PHƯƠNG PHÁP VÀ DỮ LIỆU** 12](#_Toc130803850)

[**3.1. Phương pháp** 12](#_Toc130803851)

[**3.1.1. Logistic Regression** 12](#_Toc130803852)

[**3.1.2. K-Nearest Neighbors** 13](#_Toc130803853)

[**3.1.3. Naive Bayes** 13](#_Toc130803854)

[**3.1.4. Support Vector Machines** 13](#_Toc130803855)

[**3.1.5. Decision Tree** 13](#_Toc130803856)

[**3.1.6. Random Forest** 13](#_Toc130803857)

[**3**.**1.7. Gradient Boost** 14](#_Toc130803858)

[**3.1.8. XGBoost** 14](#_Toc130803859)

[**3.1.9. LightGBM** 14](#_Toc130803860)

[**3.1.10. CatBoost** 14](#_Toc130803861)

[**3.1.11. ANN** 14](#_Toc130803862)

[**3**.**1.12. AdaBoost** 15](#_Toc130803863)

[**3.1.13. Bagging** 15](#_Toc130803864)

[**3.2. Tổng quan dữ liệu** 15](#_Toc130803865)

[**3.3. Thu thập dữ liệu** 15](#_Toc130803866)

[**3.4. Quy trình thực hiện** 17](#_Toc130803867)

[**CHƯƠNG 4. KẾT QUẢ NGHIÊN CỨU VÀ THẢO LUẬN** 18](#_Toc130803868)

[**4.1. Xử lý dữ liệu và thống kê mô tả** 18](#_Toc130803869)

[**4.1.1. Xử lý dữ liệu** 18](#_Toc130803870)

[**4.2. Thống kê mô tả** 21](#_Toc130803871)

[**4.2.1. Tổng quan sau khi xử lý dữ liệu** 21](#_Toc130803872)

[**4.2.2. Categorical data** 23](#_Toc130803873)

[**4.2.3. Numerical data** 25](#_Toc130803874)

[**4.3. Kết quả phân tích dữ liệu** 26](#_Toc130803875)

[**4.3.1. Time consumption** 26](#_Toc130803876)

[**4.3.2. Accuracy** 27](#_Toc130803877)

[**4.3.3. F1 - Score** 28](#_Toc130803878)

[**4.3.4. Precision và Recall** 29](#_Toc130803879)

[**4.4. Đánh giá mô hình machine learning** 30](#_Toc130803880)

[**4.4.1. Các tiêu chí đánh giá mô hình** 30](#_Toc130803881)

[**4.4.2. Lựa chọn mô hình** 30](#_Toc130803882)

[**4.5. Thảo luận về kết quả và so sánh với các nghiên cứu trước đây** 32](#_Toc130803883)

[**4.6. Những giới hạn và hướng phát triển trong tương lai** 33](#_Toc130803884)

[**4.6.1. Hạn chế về phạm vi và nội dung** 33](#_Toc130803885)

[**4.6.2. Hạn chế trong quá trình thực hiện nghiên cứu** 33](#_Toc130803886)

[**4.6.3. Hướng nghiên cứu** 33](#_Toc130803887)

[**CHƯƠNG 5: KẾT LUẬN VÀ KHUYẾN NGHỊ** 35](#_Toc130803888)

[**5.1. Kết luận** 35](#_Toc130803889)

[**5.2. Khuyến nghị** 35](#_Toc130803890)

[***TÀI LIỆU THAM KHẢO*** 37](#_Toc130803891)

**DANH SÁCH BIỂU ĐỒ**

Biểu đồ 4.1. Kiểm tra vị trí dữ liệu ngoại lai...........................................................................18

Biểu đồ 4.2. Phần trăm dữ liệu khuyết trong cột.....................................................................19

Biểu đồ 4.3. Biểu đồ mức độ tương quan giữa các biến..........................................................20

Biểu đồ 4.4. Boxplot và phân phối của ROA,ROE..................................................................21

Biểu đồ 4.5. Boxplot và phân phối của ROA,ROE sau khi loại bỏ outlier..............................21

Biểu đồ 4.6. Khả năng trả cổ tức của các công ty ở 2 sàn........................................................23

Biểu đồ 4.7. So sánh thời gian chạy của từng mô hình............................................................26

Biểu đồ 4.8. So sánh Accuracy của từng mô hình...................................................................27

Biểu đồ 4.9. So sánh F1 - Score của từng mô hình..................................................................28

Biểu đồ 4.10. So sánh Precision và Recall của từng mô hình..................................................29

Biểu đồ 4.11. Mức độ ảnh hưởng của các biến đến kết quả..........................................31

**DANH SÁCH HÌNH**

Hình 4.1. Thông tin dữ liệu......................................................................................................22

Hình 4.2. So sánh khả năng trả cổ tức công ty theo ngành tại Hà Nội và Hồ Chí Minh.........24

Hình 4.3. Các chỉ số thống kê của biến định lượng.................................................................25

**DANH MỤC BẢNG**

Bảng 3.1: Phân loại ngành nghề..............................................................................................16

**DANH SÁCH VIẾT TẮT**

DIV: Dividend Payout Ratio

ROE: Return of Equity

ROA: Return On Assets

KNN: Thuật toán học máy K-Nearest Neighbors

SVM: Thuật toán học máy Support Vector Machines

XGBoost: Extreme Gradient Boosting

LightGBM: Light Gradient Boosting Machine

ANN: Artificial Neural Networks

# **CHƯƠNG 1. TỔNG QUAN NGHIÊN CỨU**

## **1.1. LÝ DO CHỌN ĐỀ TÀI**

Trong kinh doanh, việc trả cổ tức là một trong những phương tiện giúp doanh nghiệp thu hút các nhà đầu tư, đồng thời cũng phản ánh khả năng sinh lời của doanh nghiệp.

Kể từ khi nghiên cứu của Miller và Modigliani (1961) công bố không có sự liên quan giữa cổ tức và giá trị công ty, nhiều nhà kinh tế tài chính đã thực hiện các nghiên cứu khác nhau và đi đến kết luận khác nhau về tác động của chi trả cổ tức đến sự giàu có của các cổ đông. Nhiều nghiên cứu đã chỉ ra rằng hầu hết các nhà quản lý tin rằng cổ tức sẽ ảnh hưởng đến giá cổ phiếu của họ trên thị trường. Vì vậy, các nhà quản lý và các nhà kinh tế đã dành nhiều thời gian để nghiên cứu các yếu tố quyết định đến chính sách chi trả cổ tức và thiết lập một chính sách cho mình (J. Chen & Dhiensiri, 2009)

Với sự phát triển của thuật toán machine learning, việc ứng dụng machine learning để dự đoán khả năng trả cổ tức của doanh nghiệp đã được nhiều nghiên cứu và đưa ra những kết quả khá tích cực. Tuy nhiên, vẫn còn nhiều thách thức trong việc áp dụng machine learning vào dự đoán cổ tức của doanh nghiệp trong thực tế.

Cùng sự mong muốn đóng góp vào lĩnh vực này, đề tài nghiên cứu “**ỨNG DỤNG MACHINE LEARNING DỰ ĐOÁN KHẢ NĂNG TRẢ CỔ TỨC CỦA DOANH NGHIỆP VIỆT NAM”** sẽ tập trung vào xây dựng một mô hình với machine learning để dự đoán khả năng trả cổ tức của các doanh nghiệp tại Việt Nam. Nghiên cứu sẽ sử dụng các phương pháp xử lý dữ liệu và huấn luyện mô hình để đánh giá độ chính xác của mô hình dự đoán..

## **1.2. MỤC TIÊU NGHIÊN CỨU**

### **1.2.1. Mục tiêu chung**

Mục tiêu của nghiên cứu là đưa ra một mô hình với machine learning là đưa ra một kết quả dự báo có kết quả khả quan để dự đoán khả năng trả cổ tức của doanh nghiệp tại Việt Nam. Kết quả của nghiên cứu này sẽ giúp các nhà đầu tư có thể đưa ra quyết định đầu tư hiệu quả và đồng thời cũng có thể hỗ trợ các doanh nghiệp trong việc quản lý tài chính và phát triển kinh doanh.

### **1.2.2. Mục tiêu cụ thể**

Xác định các yếu tố ảnh hưởng đến khả năng trả cổ tức của doanh nghiệp: Nghiên cứu cần phân tích các yếu tố như doanh thu, lợi nhuận, nợ vay, lưu chuyển tiền tệ và các chỉ số tài chính khác để xác định những yếu tố quan trọng nhất ảnh hưởng đến khả năng trả cổ tức của doanh nghiệp.

Tìm kiếm và thu thập dữ liệu: Để xây dựng mô hình với machine learning, nghiên cứu cần thu thập dữ liệu từ các nguồn khác nhau, bao gồm báo cáo tài chính của các doanh nghiệp, dữ liệu thị trường và các chỉ số kinh tế.

Xây dựng mô hình với machine learning: Nghiên cứu cần áp dụng các phương pháp xử lý dữ liệu và huấn luyện mô hình để xây dựng một mô hình với machine learning có kết quả khả quan nhằm dự đoán khả năng trả cổ tức của doanh nghiệp.

Đánh giá độ chính xác của mô hình: Nghiên cứu cần đánh giá độ chính xác của mô hình với machine learning để đảm bảo rằng nó có thể dự đoán khả năng trả cổ tức của doanh nghiệp với độ chính xác cao.

So sánh với các phương pháp dự đoán khác: Nghiên cứu cần so sánh mô hình với machine learning và các phương pháp dự đoán khác để đánh giá hiệu quả của nó, theo sau đó là xác định khả năng ứng dụng thực tiễn của mô hình.

Đề xuất cải tiến cho mô hình: Nghiên cứu cần đề xuất các cải tiến cho mô hình với machine learning để cải thiện độ chính xác và khả năng ứng dụng của nó trong thực tế.

## **1.3. ĐỐI TƯỢNG VÀ PHẠM VI CỦA NGHIÊN CỨU**

### **1.3.1. Đối tượng nghiên cứu**

Đối tượng nghiên cứu: Ứng dụng machine learning để dự đoán khả năng trả cổ tức của doanh nghiệp Việt Nam trên hai sàn chứng khoán là HOSE và HNX.

### **1.3.2. Phạm vi nghiên cứu**

Phạm vi về nội dung: Tác động của Machine Learning đến việc dự đoán khả năng trả cổ tức của các doanh nghiệp tại Việt Nam.

Phạm vi về không gian: Việc nghiên cứu thực hiện trong phạm vi các doanh nghiệp thuộc lãnh thổ Việt Nam, với số lượng là các doanh nghiệp được niêm yết trên hai sàn chứng khoán là HOSE và HNX.

Phạm vi về thời gian: Nghiên cứu được tiến hành dựa trên dữ liệu các báo cáo thường niên và phát triển bền vững trong giai đoạn từ năm 2014 đến năm 2021 của các doanh nghiệp được niêm yết trên hai sàn chứng khoán tại Việt Nam.

## **1.4. CÂU HỎI NGHIÊN CỨU**

Mô hình với thuật toán machine learning nào là phù hợp nhất để dự đoán mức độ trả cổ tức của các doanh nghiệp?

Các yếu tố nào ảnh hưởng đến việc dự đoán trả cổ tức của các doanh nghiệp bằng machine learning?

Phạm vi ứng dụng của machine learning để dự đoán trả cổ tức là như thế nào?

Mức độ chính xác của machine learning trong dự đoán trả cổ tức là bao nhiêu?

Các nhà đầu tư có thể sử dụng các kết quả dự đoán trả cổ tức từ machine learning để đưa ra quyết định đầu tư hợp lý hay không?

## **1.5. PHƯƠNG PHÁP NGHIÊN CỨU**

Dữ liệu nghiên cứu: nhóm lấy dữ liệu trong báo cáo tài chính, và báo cáo thường niên trên hai sàn giao dịch chứng khoán Hà Nội (HNX) và Thành phố Hồ Chí Minh (HOSE). Dữ liệu được thu thập bao gồm: Dividend Payment, ROA, ROE, Free Cash Flow, Leverage, Liquidity, Company Market Cap, Market to book ratio, Company age, Public age, Exchange Name, GICS Industry Name, Type.

Quy trình nghiên cứu:

Bước 1: Thực hiện quá trình tìm và đọc các tài liệu liên quan đến đề tài.

Bước 2: Thu thập và xử lý dữ liệu về các công ty niêm yết trên 2 sàn chứng khoán HOSE và HNX.

Bước 3: Đo lường các yếu tố tác động đến việc chi trả cổ tức và đề xuất mô hình với machine learning phù hợp để dự báo khả năng chi trả cổ tức của doanh nghiệp.

Bước 4: Thử nghiệm và đánh giá, so sánh với các kết quả của các nghiên cứu trước.

Bước 5: Kết luận.

## **1.6. CẤU TRÚC ĐỀ TÀI**

Công trình nghiên cứu gồm 36 trang, 1 bảng, 3 hình và 11 biểu đồ cùng 1 phụ lục. Ngoài phần mở đầu và kết luận, danh mục bảng và biểu đồ, danh sách viết tắt, danh mục tài liệu tham khảo và phụ lục, đề tài được kết cấu thành 5 mục như sau:

CHƯƠNG 1. TỔNG QUAN NGHIÊN CỨU

CHƯƠNG 2: CƠ SỞ LÝ THUYẾT

CHƯƠNG 3: PHƯƠNG PHÁP VÀ DỮ LIỆU

CHƯƠNG 4. KẾT QUẢ NGHIÊN CỨU VÀ THẢO LUẬN

CHƯƠNG 5: KẾT LUẬN VÀ KHUYẾN NGHỊ

# **CHƯƠNG 2: CƠ SỞ LÝ THUYẾT**

## **2.1. Cơ sở lý thuyết**

### **2.1.1. Signalling theory**

Theo (Brigham & Houston, 2021), lý thuyết tín hiệu là hành động mà quản lý hoặc công ty thực hiện để cung cấp các chỉ dẫn/tín hiệu cho nhà đầu tư liên quan đến hiệu suất, điều kiện và triển vọng của công ty trong tương lai. Tín hiệu có thể là tín hiệu tích cực hoặc tín hiệu tiêu cực. Lý thuyết tín hiệu là một lý thuyết mô tả việc thay đổi trong phân phối cổ tức là dấu hiệu cho thấy thu nhập được tạo ra bởi công ty sẽ được phân phối cho cổ đông. Lý thuyết tín hiệu cũng bàn về khía cạnh không đối xứng của thông tin giữa những người trong nội bộ (quản lý) và các bên bên ngoài (cổ đông). Người trong nội bộ có nhiều thông tin hơn về công ty không được công bố và sở hữu bởi các nhà đầu tư. Theo (Chaniago & Ekadjaja, 2022), lý thuyết tín hiệu giải thích rằng thông tin về việc trả cổ tức tiền mặt cho các nhà đầu tư có thể là một tín hiệu tích cực về triển vọng của công ty trong tương lai. Trong khi đó, các công ty có sự giảm trả cổ tức tiền mặt có thể là một tín hiệu tiêu cực về triển vọng của công ty trong tương lai (Affandi, SUNARKO, & YUNANTO, 2019). Lý thuyết tín hiệu này giải thích rằng thông tin về việc trả tiền cổ tức tiền mặt có thể được sử dụng như một hướng dẫn cho nhà đầu tư liên quan đến triển vọng của công ty trong tương lai do sự không đối xứng của thông tin giữa quản lý và nhà đầu tư.

### **2.1.2. Agency theory**

Theo agency theory, đó là một mối quan hệ giữa một hoặc nhiều cổ đông (chủ sở hữu) để ủy quyền quyền ra quyết định cho đại diện. Vấn đề xảy ra khi có mâu thuẫn lợi ích giữa nhà quản lý và nhà đầu tư của họ. Mâu thuẫn lợi ích giữa chủ sở hữu và đại lý có thể xảy ra do hành động/hành vi của đại diện không phù hợp với nguyên tắc, do đó gây ra agency cost (Affandi et al., 2019). Xung đột này có thể được giải quyết bằng cách cân bằng lợi ích giữa nhà quản lý và nhà đầu tư (Amalia & Hermanto, 2018). Sự tồn tại của agency theory được hy vọng có thể giảm thiểu chi phí do bất đối xứng thông tin giữa nhà quản lý và cổ đông. Bất đối xứng thông tin là tình huống một bên biết nhiều thông tin hơn bên kia. Chi phí đại lý có dạng chi phí giám sát, chi phí ràng buộc và tổn thất dư còn lại theo (Harryono, 2020). Chi phí giám sát là chi phí phát sinh và chịu bởi người đại diện để đo lường, quan sát và kiểm soát hành vi của đại lý. Chi phí ràng buộc là chi

phí do đại lý chịu để thiết lập và tuân thủ cơ chế đảm bảo đại lý sẽ hành động trong lợi ích của người đại diện. Tổn thất dư còn lại là sự hy sinh dưới dạng giảm sự thịnh vượng của người đại diện do sự khác biệt trong quyết định của đại lý và người đại diện. Theo (Supriyono & Herdhayinta, 2019), lý thuyết đại lý là mối quan hệ mà đại lý được ủy quyền bởi người đại diện đối với quyết định của người đại diện bằng cách tối đa hóa lợi nhuận của công ty, vì vậy nó có thể giảm thiểu gánh nặng thuế bằng cách tránh thuế.

### **2.1.3. Pecking order theory**

Lý thuyết thứ tự ưu tiên cung cấp các liên kết giữa cấu trúc vốn, chính sách cổ tức và đầu tư của một công ty. Lý thuyết cho rằng các công ty thích sử dụng vốn chủ sở hữu nội bộ để trả cổ tức và tài trợ đầu tư mới. Vốn chủ sở hữu nội bộ xếp ở đầu thứ tự ưu tiên, tiếp theo là nợ, cổ phiếu ưu đãi và vốn chủ sở hữu bên ngoài.

### **2.1.4. Dự đoán cổ tức**

Dự đoán chi trả cổ tức là việc ước tính mức độ khả dụng của tiền thu được từ hoạt động kinh doanh để chi trả cổ tức cho cổ đông trong tương lai. Để dự đoán trả cổ tức, các nhà đầu tư thường sử dụng các phương pháp phân tích tài chính để đánh giá các chỉ số tài chính quan trọng của công ty, như lợi nhuận, doanh thu, cơ cấu tài sản, tính thanh khoản và dòng tiền tự do.

Ngoài các chỉ số tài chính, các nhà đầu tư cũng phân tích các yếu tố khác như xu hướng thị trường, lĩnh vực kinh doanh của công ty, chiến lược kinh doanh và quản lý công ty để đưa ra dự đoán trả cổ tức chính xác hơn.

Tuy nhiên, cần lưu ý rằng dự đoán trả cổ tức là một quá trình phức tạp và không đảm bảo chính xác 100%. Các kết quả dự đoán có thể bị ảnh hưởng bởi các yếu tố khác như biến động thị trường, thay đổi chiến lược kinh doanh của công ty và sự kiện không đoán trước được. Do đó, việc dự đoán trả cổ tức cần được xem là một công cụ hỗ trợ cho quyết định đầu tư, và không thể thay thế cho sự hiểu biết và kinh nghiệm của các nhà đầu tư chuyên nghiệp.

## **2.2. Các bài nghiên cứu trước đây**

Các nghiên cứu về dự đoán trả cổ tức sử dụng phương pháp machine learning là một lĩnh vực đang được quan tâm và phát triển trong lĩnh vực tài chính và đầu tư. Một số nghiên cứu trước đây đã được thực hiện để khảo sát khả năng của các phương pháp machine learning trong việc dự đoán mức độ trả cổ tức của các doanh nghiệp.

Nghiên cứu của Al-Shubiri và Al-Absi (2019) đã sử dụng phương pháp học máy Random Forest và Neural Network để dự đoán trả cổ tức của các công ty Jordan. Kết quả cho thấy hai mô hình này có khả năng dự đoán mức độ trả cổ tức với độ chính xác cao.

Nghiên cứu của (Jain, Shakya, Khatter, & Gupta, 2019) tập trung vào các công ty Ấn Độ và sử dụng phương pháp học máy Support Vector Machine để dự đoán trả cổ tức. Kết quả cho thấy mô hình này có độ chính xác cao trong việc dự đoán mức độ trả cổ tức của các doanh nghiệp.

(Islam et al., 2021) đã sử dụng phương pháp học máy Artificial Neural Network để dự đoán trả cổ tức của các công ty Bangladesh. Kết quả cho thấy mô hình này cũng có khả năng dự đoán mức độ trả cổ tức với độ chính xác cao.

Nghiên cứu của Ramasamy và Taraman (2016) tập trung vào các công ty Malaysia và sử dụng phương pháp học máy Decision Tree để dự đoán trả cổ tức. Kết quả cho thấy mô hình này cũng có độ chính xác cao trong việc dự đoán mức độ trả cổ tức của các doanh nghiệp.

Tuy nhiên, cần lưu ý rằng các nghiên cứu này chỉ mang tính chất tham khảo và việc sử dụng phương pháp machine learning để dự đoán trả cổ tức vẫn đang trong quá trình phát triển và cần được xem xét kỹ lưỡng để đưa ra quyết định đầu tư chính xác. Ngoài ra, việc đánh giá các yếu tố khác như rủi ro, điều kiện thị trường cũng rất quan trọng trong việc đưa ra quyết định đầu tư.

## **2.3. Phát triển giả thuyết**

### **2.3.1. Mối quan hệ giữa DIV và tỷ số lợi nhuận trên tài sản (ROA)**

Khả năng sinh lời của một công ty có ảnh hưởng đáng kể đến khả năng chi trả cổ tức, vì cổ tức là một phần của lợi nhuận ròng của công ty. Các nghiên cứu trước đây, bao gồm (Al-Shubiri, 2011), (Al-Kuwari, 2012), (Amidu & Abor, 2006) đã chỉ ra rằng tỷ suất sinh lời trên tài sản là yếu tố quan trọng đóng vai trò trong khả năng công ty chi trả cổ tức. Nghiên cứu của (Al-Malkawi, 2008) cũng cho thấy rằng khả năng sinh lời là yếu tố quan trọng nhất trong việc quyết định mức độ chi trả cổ tức của một công ty.

***H1: Tỷ số lợi nhuận trên tài sản (ROA) có quan hệ cùng chiều với tỷ lệ chi trả cổ tức.***

**ROA = net income/ total assets**

### **2.3.2. Mối quan hệ giữ DIV và lợi nhuận trên vốn (ROE)**

(ROE) là tỷ lệ cho thấy khả năng của công ty sử dụng các nguồn vốn để kiếm lợi nhuận sau thuế (Simbolon & Sampurno, 2017). Lợi nhuận được đề cập ở đây là lợi nhuận cho cổ đông (lợi nhuận cho cổ đông sở hữu) hoặc lợi nhuận sau thuế (lợi nhuận sau thuế). Khả năng kiếm lợi nhuận của công ty có thể cho thấy khả năng trả cổ tức của công ty. Nếu các công ty có khả năng trả cổ tức cho cổ đông, điều đó có thể tạo ra ấn tượng rằng công ty đó có lợi nhuận. Khi ROE tăng, tỷ lệ trả cổ tức cũng tăng theo. Dựa trên lý thuyết tín hiệu, lợi nhuận tăng trong một công ty có thể cung cấp tín hiệu tích cực cho nhà đầu tư về triển vọng của công ty trong tương lai thông qua việc tăng tỷ lệ trả cổ tức. ROE là tỷ lệ đo lường tính lợi nhuận từ phía cổ đông và là một công cụ thường được sử dụng trong việc ra quyết định. Theo (Affandi et al., 2019), ROE cho thấy bao nhiêu lợi nhuận mà công ty tạo ra từ vốn đầu tư của cổ đông. ROE cũng giúp cổ đông đánh giá tính khả thi của một cổ phiếu khi điều kiện thị trường chung không tốt. Giá trị ROE lớn trong một công ty có thể cho thấy công ty đó hiệu quả trong việc sử dụng vốn chủ sở hữu để tạo ra lợi nhuận. Ngược lại, một ROE nhỏ có thể cho thấy công ty đó kém hiệu quả trong việc quản lý vốn để kiếm lợi nhuận, điều này ảnh hưởng đến việc trả cổ tức cho cổ đông của công ty.

***H2: Tỷ số lợi nhuận trên vốn (ROE) có quan hệ cùng chiều với tỷ lệ chi trả cổ tức.***

**ROE = net income/ total equity**

### **2.3.3. Mối quan hệ giữa DIV và dòng tiền tự do (FCF)**

Dòng tiền tự do (hay còn gọi là "free cash flow" trong tiếng Anh) mô tả mức độ linh hoạt tài chính của một công ty. Dựa trên lý thuyết đại diện (agency theory), nếu công ty có dòng tiền tự do, nhà quản lý công ty sẽ bị áp lực từ các nhà đầu tư để phân phối nó dưới dạng cổ tức. Việc này được thực hiện để ngăn chặn nhà quản lý sử dụng dòng tiền tự do cho những điều không phù hợp với mục tiêu của công ty và có xu hướng gây hại cho nhà đầu tư. Do đó, sự có mặt của các quỹ có thể được sử dụng cho sự thịnh vượng của cổ đông. Cổ đông tin rằng dòng tiền tự do được phân phối dưới dạng cổ tức sẽ có giá trị cao hơn và rủi ro thấp hơn. Các công ty có khả năng tạo ra dòng tiền tự do cao sẽ có thể cung cấp lợi nhuận cao cho cổ đông, so với những công ty tạo ra dòng tiền tự do thấp hàng năm.

***H3: Free cash flow có quan hệ cùng chiều với tỷ lệ chi trả cổ tức.***

**Free cash flow (FCF) = (Operating income + depreciation - taxes - interest expenses – preferred dividends - ordinary dividends)/Total assets**

### **2.3.4. Mối quan hệ giữa DIV và đòn bẩy (Leverage)**

Một số nghiên cứu gần đây cho thấy mức độ đòn bẩy tài chính ảnh hưởng tiêu cực đến chính sách cổ tức ((Jensen, 1987); (Faccio, Lang, & Young, 2001); (Gugler & Yurtoglu, 2003); (Nizar Al‐Malkawi, 2007)). Các nghiên cứu này cho thấy các công ty có đòn bẩy tài chính cao tập trung vào việc duy trì dòng tiền nội bộ để đáp ứng các nghĩa vụ, thay vì phân phối tiền có sẵn cho cổ đông và bảo vệ các chủ nợ. Tuy nhiên, Mollah (2011) đã nghiên cứu một thị trường mới nổi và phát hiện ra mối quan hệ trực tiếp giữa đòn bẩy tài chính và mức độ nợ gây khó khăn cho việc giao dịch. Do đó, các công ty với tỷ lệ đòn bẩy cao có chi phí giao dịch cao hơn và không đủ sức để trả cổ tức cao hơn để tránh chi phí tài chính bên ngoài.

Để phân tích mức độ ảnh hưởng của nợ đến việc trả cổ tức, nghiên cứu này sử dụng tỷ lệ đòn bẩy tài chính, hoặc tỷ lệ của các khoản nợ (tổng nợ ngắn hạn và dài hạn) đối với tổng vốn chủ sở hữu. Đòn bẩy tài chính cho thấy tổng số nợ so với tổng số nợ và vốn chủ sở hữu. Càng sử dụng nhiều nợ, công ty càng bị kiểm soát bởi các chủ nợ và đối mặt với rủi ro tài chính cao hơn. Do đó, nếu các công ty có đòn bẩy tài chính cao hơn, tỷ lệ cổ tức có thể thấp hơn. Điều này đã được nghiên cứu bởi rất nhiều nhà kinh tế như (Jensen, 1987), (Ho, 2003), (Myers & Bacon, 2004), và nhiều người khác. Các công ty phải chi tiêu tiền và tài sản cho các chủ nợ trước khi trả cổ tức cho cổ đông. Ngoài ra, các công ty giữ tỷ lệ nợ cao có thể giảm tỷ lệ cổ tức nếu họ không muốn đối mặt với chi phí cao khi tăng quỹ bên ngoài.

***H4: Đòn bẩy tài chính có quan hệ ngược chiều với tỷ lệ chi trả cổ tức.***

**Financial leverage (LEV) = total liabilities / total equity**

### **2.3.5. Mối quan hệ giữa DIV và thanh khoản (Liquidity)**

Tỷ lệ thanh khoản là một nhóm chỉ số tài chính quan trọng được sử dụng để xác định khả năng của một công ty trả các khoản nợ ngắn hạn mà không cần tăng vốn bên ngoài. Thanh khoản là mức độ mà một công ty có thể trả các khoản nợ ngắn hạn dựa trên tài sản dễ chuyển đổi của nó (Atrill & McLaney, 2002). (Ho, 2003) đã phát hiện ra rằng các công ty của Nhật Bản có tỷ lệ thanh khoản cao hơn thường có tỷ suất cổ tức

cao hơn. Tuy nhiên, (Mehar, 2002) cho rằng có mối quan hệ nghịch đảo giữa tình trạng thanh khoản và việc trả cổ tức. Từ nghiên cứu của Mehar về các công ty trên Sở giao dịch chứng khoán Karachi ở Pakistan, các công ty có vốn lưu động dương sẽ giảm cổ tức. (Myers & Bacon, 2004) nhấn mạnh rằng các công ty có khả năng giảm cổ tức để phân bổ thanh khoản. Tuy nhiên, (Al‐Najjar & Hussainey, 2009) đã chứng minh rằng việc trả cổ tức thấp hay cao không phụ thuộc vào tình trạng thanh khoản tốt hay xấu.

**H5: Liquidity có quan hệ ngược chiều với tỷ lệ chi trả cổ tức.**

**Liquidity (LIQ) = Current ratio = Current assets/Current liabilities**

### **2.3.6. Mối quan hệ giữa DIV và MTB**

Nếu công ty có thể tối đa hóa khả năng của mình thông qua các khoản đầu tư để tạo ra lợi nhuận phù hợp với số tiền vốn ràng buộc, nó có thể tăng giá trị của công ty. Tỷ lệ giá trị thị trường so với giá trị sách của cổ phiếu càng cao, giá trị của công ty càng cao. Nghiên cứu trước đó được thực hiện bởi (Viveronica & Amanah, 2017) và (Panjaitan & Akmalia, 2020) cho thấy rằng tỷ lệ giá trị thị trường so với giá trị sách của cổ phiếu có tác động tích cực đến giá trị công ty.

***H6: MTB có quan hệ cùng chiều với tỷ lệ chi trả cổ tức.***

**Market-to-book ratio = market capitalization / book value**

### **2.3.7. Mối quan hệ giữa DIV và Capitalization**

Đây là mối quan hệ quan trọng giữa việc trả cổ tức và vốn hóa công ty đã được nghiên cứu rất nhiều trong các nghiên cứu tài chính. Theo nghiên cứu của (Baker & Wurgler, 2004), các công ty có vốn hóa lớn thường có xu hướng trả cổ tức cao hơn so với các công ty có vốn hóa nhỏ hơn. Điều này được giải thích bằng cách cho rằng các công ty có vốn hóa lớn thường có nhiều lợi nhuận hơn và do đó có khả năng trả cổ tức cao hơn. Ngoài ra, các nhà đầu tư thường có xu hướng đầu tư vào các công ty có vốn hóa lớn hơn vì chúng được coi là tương đối ổn định và có khả năng sinh lợi cao hơn trong tương lai.

Nghiên cứu của (Fama & French, 2002) cũng chỉ ra rằng các công ty có cổ tức cao hơn thường có vốn hóa lớn hơn và có kết quả kinh doanh tốt hơn. Họ cũng nhận thấy rằng các công ty có cổ tức cao hơn thường được các nhà đầu tư ưa chuộng hơn vì chúng được coi là có khả năng tăng giá cổ phiếu trong tương lai.

***H7: Capitalization có quan hệ cùng chiều với tỷ lệ chi trả cổ tức.***

### **2.3.8 Mối quan hệ giữa DIV và Company age**

Tuổi của doanh nghiệp được tính là số năm kể từ khi thành lập đến năm 2021. Theo lý thuyết, càng lâu càng tốt vì doanh nghiệp được thành lập càng lâu thì sự tin tưởng từ các nhà đầu tư càng mạnh. Điều này được giả định là càng lâu càng có lợi nhuận hơn so với các công ty mới thành lập. Các công ty mới thành lập thường có xu hướng tìm kiếm đầu tư trên thị trường vốn và do đó phụ thuộc vào vốn chủ sở hữu của mình (Zen & Herman, 2012). Lý thuyết điều hành cũng đề cập đến việc càng cao tuổi của công ty thì càng ít bất định của không đối xứng thông tin có thể ảnh hưởng đến việc giá cổ phiếu bị định giá thấp hơn. Lý thuyết tín hiệu cho rằng càng cao tuổi của công ty, càng nhiều kinh nghiệm doanh nghiệp học được để vận hành và quản lý công ty của họ một cách tốt hơn. Những công ty đó có thể cung cấp thông tin và tin tức tốt đẹp cho các nhà chủ chốt của họ, có thể ảnh hưởng tích cực đến giá trị doanh nghiệp.

***H8: CAGE có quan hệ cùng chiều với tỷ lệ chi trả cổ tức.***

### **2.3.9. Mối quan hệ giữa DIV và Public age**

PAGE thể hiện số năm mà công ty đã tồn tại sau khi được niêm yết trên thị trường chứng khoán (Alexander, 2017)

Theo Bassiounyet al., (2016) một công ty đã niêm yết trong thời gian dài ít có khả năng quản lý lợi nhuận so với một công ty mới niêm yết, điều này bởi vì những công ty cũ hơn là các công ty được biết đến với giá trị lớn hơn trên thị trường và vì vậy, họ sẽ muốn tránh danh tiếng xấu. Ngoài ra, các công ty cũ hơn rất nắm bắt các quy tắc và mã đạo đức điều hành hoạt động của họ trong khi các công ty mới sẽ muốn chứng tỏ với thị trường chứng khoán rằng nó đang hoạt động tốt và đã đáp ứng tất cả các yêu cầu của thị trường và công chúng nói chung để thu hút thêm nhà đầu tư. Các công ty cũ có thể đã cải thiện các phương pháp báo cáo tài chính của họ theo thời gian (Alsaeed,2006) và cố gắng nâng cao danh tiếng và hình ảnh của mình trên thị trường (Akhatarruddin, 2005), vì vậy càng cũ thì càng ít khả năng cho công ty gian lận lợi nhuận và càng mới thì càng cao khả năng gian lận lợi nhuận (Sara Wahi, 2016).

***H9: PAGE có quan hệ cùng chiều với tỷ lệ chi trả cổ tức.***

# 

# **CHƯƠNG 3: PHƯƠNG PHÁP VÀ DỮ LIỆU**

## **3.1. Phương pháp**

Dự báo DIV là một bài toán phân loại với kết quả nhị phân Có/Không. Để sử dụng các thuật toán học máy trong Python, các giá trị phản hồi được chuyển đổi thành loại 0/1.

Machine learning là một lĩnh vực rộng trong khoa học máy tính và trí tuệ nhân tạo, trong đó máy tính được lập trình để học từ dữ liệu và tự động cải thiện hiệu suất của chúng theo thời gian. Theo nghiên cứu của các chuyên gia, Machine learning được xem là một phương pháp tiên tiến và có nhiều lợi thế so với phương pháp truyền thống trong việc giải quyết các vấn đề. Thứ nhất, Machine learning có khả năng xử lý các dữ liệu lớn và phức tạp hơn so với phương pháp truyền thống. Các mô hình Machine learning có thể tự động tìm ra các mối quan hệ phức tạp giữa các đặc trưng và kết quả, giúp cho việc dự báo cổ tức trở nên chính xác hơn. Thêm vào đó, nhờ sự linh hoạt của trí tuệ nhân tạo, nó cho phép các mô hình dự đoán được cập nhật và cải tiến dựa trên dữ liệu mới. Điều này giúp cho việc dự báo cổ tức trở nên chính xác hơn theo thời gian. Và cuối cùng, Machine learning có thể tự động hóa các tác vụ và giảm thiểu sự phụ thuộc vào con người. Các mô hình Machine learning có thể tự động học hỏi từ dữ liệu mà không cần phải được lập trình trực tiếp, giúp tiết kiệm thời gian và chi phí.

### **3.1.1. Logistic Regression**

Logistic Regression là một phương pháp học máy được sử dụng để dự đoán kết quả của biến phụ thuộc nhị phân dựa trên một hoặc nhiều biến độc lập. Điểm khác biệt giữa Logistic Regression và Linear Regression là biến phụ thuộc của Logistic Regression là một biến nhị phân (0 hoặc 1), trong khi Linear Regression là biến liên tục.

Logistic Regression sử dụng hàm sigmoid để tính xác suất của biến phụ thuộc nhị phân. Hàm sigmoid sẽ chuyển đổi giá trị đầu vào thành một giá trị xác suất nằm trong khoảng từ 0 đến 1. Với Logistic Regression, chúng ta cần điều chỉnh các tham số sao cho đường sigmoid phù hợp nhất với dữ liệu (Hastie, Tibshirani, Friedman, & Friedman, 2009).

### **3.1.2. K-Nearest Neighbors**

K-Nearest Neighbors (KNN) là một thuật toán học máy sử dụng trong bài toán phân loại và dự đoán. Thuật toán này dựa trên nguyên lý "gần kề nhau thì giống nhau" (the closer, the better) để dự đoán lớp hoặc giá trị của một điểm dữ liệu mới dựa trên k điểm gần nhất trong tập dữ liệu huấn luyện (Cui et al., 2021).

### **3.1.3. Naive Bayes**

Naive Bayes là một thuật toán học máy được sử dụng rộng rãi trong các bài toán phân loại và dự đoán giá trị của các điểm dữ liệu mới. Thuật toán này dựa trên định lý Bayes để tính xác suất của một điểm dữ liệu thuộc về từng lớp, sau đó chọn lớp có xác suất cao nhất cho điểm dữ liệu đó (Nguyen & Do, 2020).

### **3.1.4. Support Vector Machines**

Support Vector Machines (SVM) là một thuật toán học máy được sử dụng trong các bài toán phân loại và hồi quy. Thuật toán SVM tìm cách xác định một siêu mặt phẳng sao cho khoảng cách giữa các điểm dữ liệu thuộc các lớp khác nhau là lớn nhất. Các điểm dữ liệu gần nhất với siêu mặt phẳng được gọi là các vector hỗ trợ (support vectors), từ đó tên của thuật toán được đặt (Cortes & Vapnik, 1995).

### **3.1.5. Decision Tree**

Decision Tree (cây quyết định) là một thuật toán học máy được sử dụng rộng rãi trong các bài toán phân loại và dự đoán giá trị. Thuật toán này dựa trên việc tạo ra một cây phân loại dựa trên tập dữ liệu huấn luyện, trong đó mỗi nút trên cây đại diện cho một thuộc tính của dữ liệu và mỗi nhánh của nút đại diện cho một giá trị của thuộc tính đó. Các quyết định về lớp của một điểm dữ liệu mới được đưa ra bằng cách di chuyển từ gốc của cây đến lá tương ứng với giá trị của các thuộc tính của điểm dữ liệu đó (Han, Kamber, & Pei, 2012).

### **3.1.6. Random Forest**

Random Forest là một thuật toán học máy được sử dụng để giải quyết bài toán phân loại, dự đoán giá trị, và khai thác mối quan hệ giữa các biến đầu vào. Thuật toán này được xây dựng dựa trên một số lượng lớn các cây quyết định độc lập được xây dựng trên các tập con ngẫu nhiên của dữ liệu huấn luyện. Quá trình huấn luyện các cây quyết định trên các tập con khác nhau và kết hợp kết quả của chúng giúp cho Random Forest tránh được overfitting và tăng khả năng tổng quát hóa của mô hình (Breiman, 2001).

### **3**.**1.7. Gradient Boost**

Gradient Boosting là một phương pháp học máy trong đó các mô hình học máy khác nhau được xây dựng theo một cách tăng dần độ khó, mỗi mô hình cố gắng cải thiện dự đoán của mô hình trước đó bằng cách sử dụng đạo hàm của hàm mất mát (loss function). Thuật toán Gradient Boosting thường được sử dụng để giải quyết các bài toán dự đoán giá trị số (regression) và phân loại (classification) (Friedman, 2001).

### **3.1.8. XGBoost**

XGBoost (Extreme Gradient Boosting) là một thuật toán học máy phổ biến trong lĩnh vực dự đoán, được phát triển bởi Dương Dương tại Đại học Washington và có khả năng xử lý dữ liệu lớn và phức tạp. XGBoost kết hợp giữa Gradient Boosting và các kỹ thuật Regularization để cải thiện hiệu suất và tránh overfitting (T. Chen & Guestrin, 2016).

### **3.1.9. LightGBM**

LightGBM (Light Gradient Boosting Machine) là một thuật toán học máy phổ biến được phát triển bởi Microsoft, có khả năng xử lý dữ liệu lớn và phức tạp nhanh hơn so với các thuật toán Gradient Boosting truyền thống như XGBoost. LightGBM sử dụng một kỹ thuật mới gọi là Gradient-based One-Side Sampling (GOSS) để giảm thiểu thời gian huấn luyện mô hình (Ke et al., 2017).

### **3.1.10. CatBoost**

CatBoost là một thuật toán học máy đang được sử dụng phổ biến trong các cuộc thi khoa học dữ liệu và ứng dụng thực tế. Thuật toán này được phát triển bởi công ty Yandex, với khả năng xử lý dữ liệu có cấu trúc và không có cấu trúc.

CatBoost sử dụng một kỹ thuật gọi là Symmetric Decision Tree để tạo ra các cây quyết định. Ngoài ra, nó còn hỗ trợ các tính năng như tự động điều chỉnh tham số và xử lý các giá trị bị khuyết trong dữ liệu (Dorogush, Ershov, & Gulin, 2018).

### **3.1.11. ANN**

ANN là viết tắt của Artificial Neural Networks, hay còn gọi là mạng neural nhân tạo. Đây là một thuật toán học máy được xây dựng dựa trên cách thức hoạt động của các tế bào thần kinh trong não người. Thuật toán ANN là một trong những thuật toán học máy phổ biến nhất hiện nay và được sử dụng rộng rãi trong nhiều lĩnh vực như xử lý

ngôn ngữ tự nhiên, nhận diện hình ảnh, dự đoán tài chính,...(Rumelhart, Hinton, & Williams, 1986)

### **3**.**1.12. AdaBoost**

AdaBoost (Adaptive Boosting) là một thuật toán học máy đa dạng (ensemble) được sử dụng rộng rãi trong các bài toán phân loại và dự đoán. Thuật toán này kết hợp nhiều bộ phân loại yếu để tạo ra một bộ phân loại mạnh hơn (Freund & Schapire, 1997).

### **3.1.13. Bagging**

Bagging (Bootstrap Aggregating) là một thuật toán học máy đa dạng (ensemble) được sử dụng rộng rãi trong các bài toán phân loại và dự đoán. Thuật toán này sử dụng kỹ thuật lấy mẫu ngẫu nhiên có tái lấy mẫu (bootstrap) để tạo ra nhiều bộ dữ liệu con khác nhau, từ đó huấn luyện nhiều bộ phân loại riêng biệt và kết hợp chúng để tạo ra một bộ phân loại mạnh hơn (Breiman, 1996).

## **3.2. Tổng quan dữ liệu**

Nghiên cứu này xem xét các công ty được niêm yết trên thị trường chứng khoán Việt Nam từ năm 2014 đến 2021. Các biến số dữ liệu cấp công ty được lấy từ báo cáo tài chính và bảng cân đối kế toán. Nghiên cứu này là nghiên cứu giải thích với phương pháp định lượng. Dữ liệu được sử dụng trong nghiên cứu này là dữ liệu thứ cấp. Dữ liệu thứ cấp có nghĩa là được xuất bản hoặc sử dụng bởi tổ chức khác. Bảng 1 hiển thị định nghĩa của mỗi biến số.

## **3.3. Thu thập dữ liệu**

Dữ liệu được thu thập từ Phòng Business Intelligence Research Lab & SAP Next-Gen Lab ở tầng 8 của tòa A tại Trường Đại Học Kinh Tế Luật. Phạm vi của dữ liệu bao gồm các ngành nghề đang hoạt động trên Sàn giao dịch chứng khoán TP.HCM và Sàn giao dịch chứng khoán Hà Nội, bao gồm các nhóm ngành nghề như sau: Utilities, Material, Industrials, Consumer discretionary, Real estate, Financials, Energy, Health care, Information technology, Consumer staples, Communication services. Dữ liệu được thu thập từ năm 2014 đến năm 2021.

*Bảng 3.1: Phân loại ngành nghề*

|  |  |
| --- | --- |
| **Category** | **Subcategories** |
| Energy | Energy Equipment & Services, Oil, Gas & Consumable Fuels |
| Material | Chemicals, Construction Materials, Containers & Packaging, Metals & Mining, Paper & Forest Products |
| Industrials | Building Products, Construction & Engineering, Electrical Equipment, Industrial Conglomerates, Machinery, Trading Companies & Distributors, Commercial Services & Supplies, Professional Services, Air Freight & Logistics, Airlines, Marine, Road & Rail, Transportation Infrastructure |
| Consumer Discretionary | Auto Components, Automobiles, Household Durables, Hotels, Restaurants & Leisure, Textiles, Apparel & Luxury Goods, Diversified Consumer Services, Distributors, Specialty Retail, Leisure Products |
| Consumer Staples | Beverages, Food Products, Tobacco, Household Products |
| Health Care | Health Care Equipment & Supplies, Health Care Providers & Services, Pharmaceuticals |
| Financials | Banks, Diversified Financial Services, Capital Markets, Insurance |
| Information Technology | IT Services, Software, Communications Equipment, Electronic Equipment, Instruments & Components |
| Communication Services | Diversified Telecommunication Services, Entertainment, Media |

|  |  |
| --- | --- |
| Utilities | Electric Utilities, Gas Utilities, Water Utilities, Independent Power and Renewable Electricity Producers |
| Real Estate | Real Estate Management & Development |

## **3.4. Quy trình thực hiện**

Bước 1: Thu thập dữ liệu.

Bước 2: Tiền xử lý dữ liệu, bước này bao gồm việc loại bỏ dữ liệu nhiễu và giá trị thiếu, chuẩn hóa dữ liệu, và chuyển đổi dữ liệu sang định dạng phù hợp.

Bước 3: Chọn mô hình Machine learning.

Bước 4: Huấn luyện mô hình, bước này là quá trình huấn luyện mô hình trên dữ liệu đã được tiền xử lý, chia dữ liệu thành các tập huấn luyện và tập kiểm tra.

Bước 5: So sánh và đánh giá kết quả của các mô hình.

Bước 6: Đưa ra kết luận, khuyến nghị và hạn chế trong vấn đề nghiên cứu.

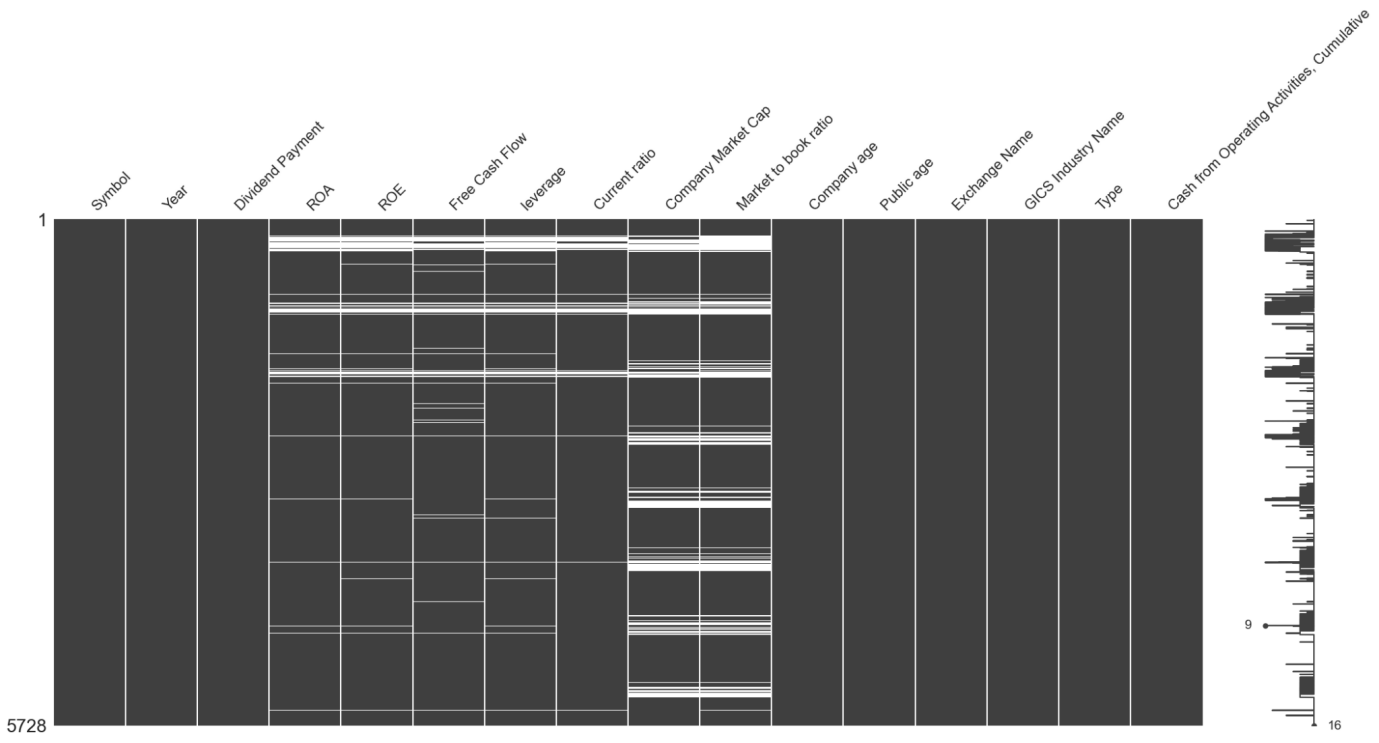
# **CHƯƠNG 4. KẾT QUẢ NGHIÊN CỨU VÀ THẢO LUẬN**

## **4.1. Xử lý dữ liệu và thống kê mô tả**

## **4.1.1. Xử lý dữ liệu**

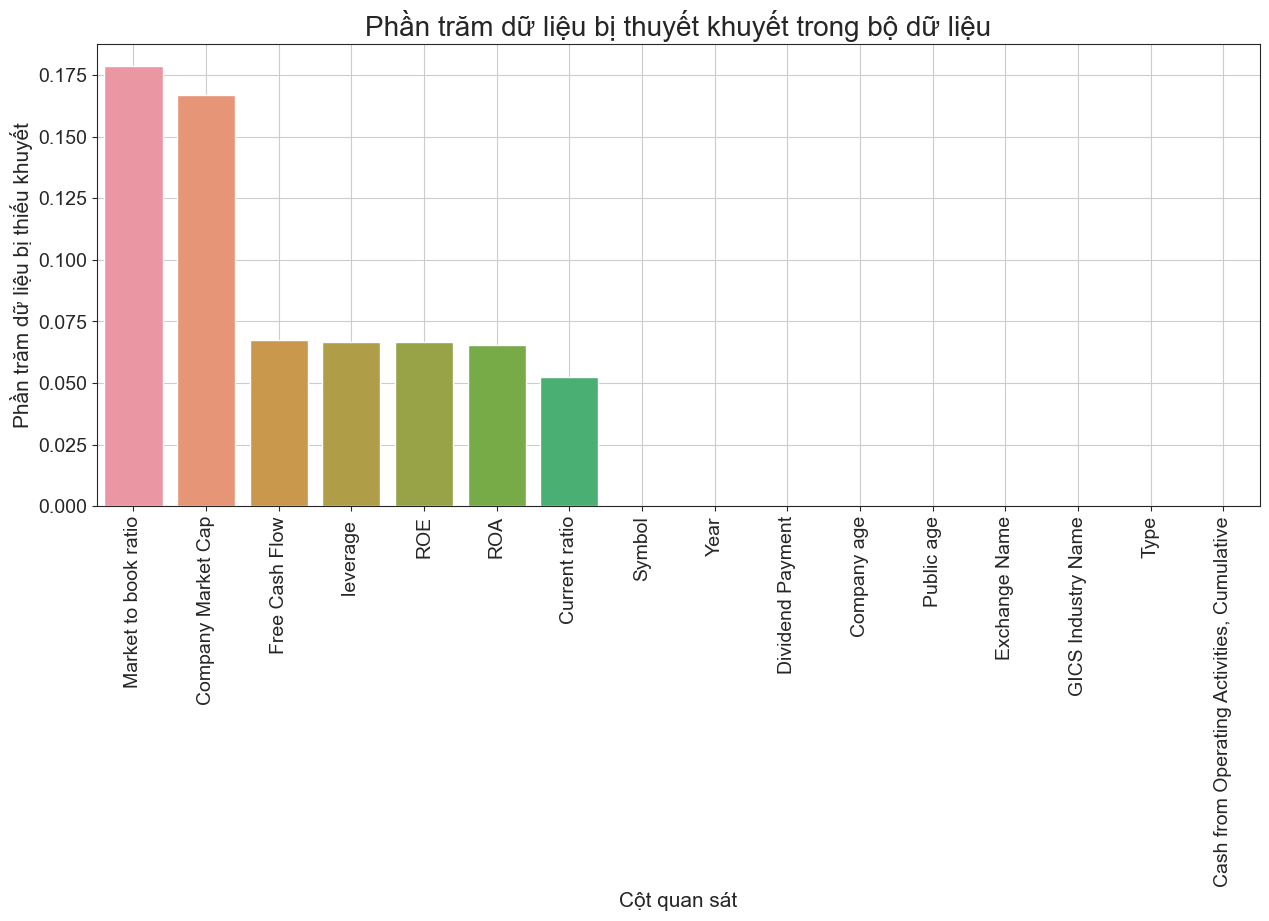
#### **4.1.1.1. Xử lý dữ liệu khuyết**

*Biểu đồ 4.1. Kiểm tra vị trí dữ liệu ngoại lai*

****

**Nhận xét biểu đồ 4.1:** Các cột quan sát bao gồm ROA, ROE, Free Cash Flow, Leverage, Current ratio, Company Market Cap và Market to book ratio chứa dữ liệu ngoại lai, trong khi các cột quan sát khác không có dữ liệu bị thiếu. Tổng số lượng dữ liệu bị thiếu là 4,15% sau khi thực hiện kiểm tra trên toàn bộ tập dữ liệu.

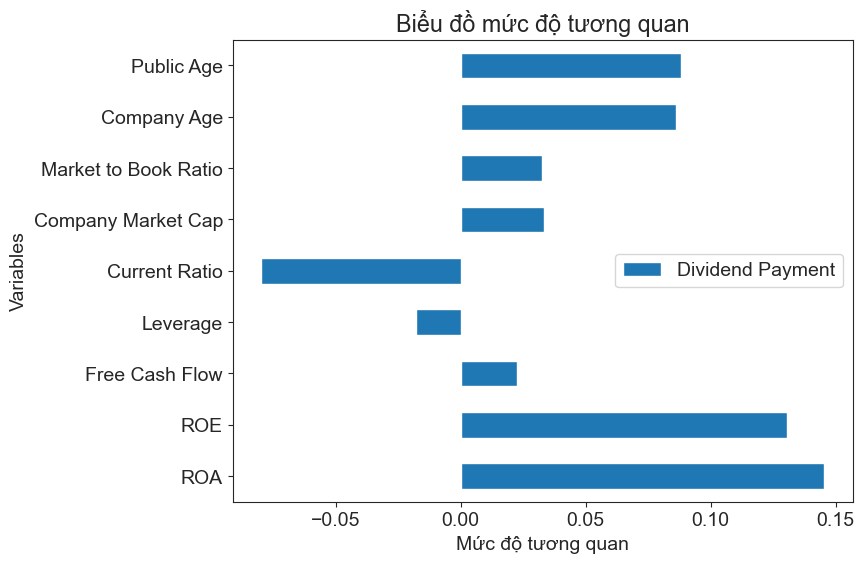
*Biểu đồ 4.2. Phần trăm dữ liệu khuyết trong cột*



**Nhận xét biểu đồ 4.2:** Các cột quan sát Market to book ratio và Company Market Cap có tỷ lệ dữ liệu bị khuyết cao nhất, với khoảng 17,5% tổng số dữ liệu trong mỗi cột. Các cột quan sát khác có tỷ lệ dữ liệu ngoại lai vượt quá 5%. Vì vậy, quyết định đã được đưa ra là loại bỏ toàn bộ các dữ liệu khuyết trong bộ dữ liệu do số lượng dữ liệu khuyến trên tổng bộ dữ liệu thấp.

#### **4.1.1.2. Kiểm tra dữ liệu**

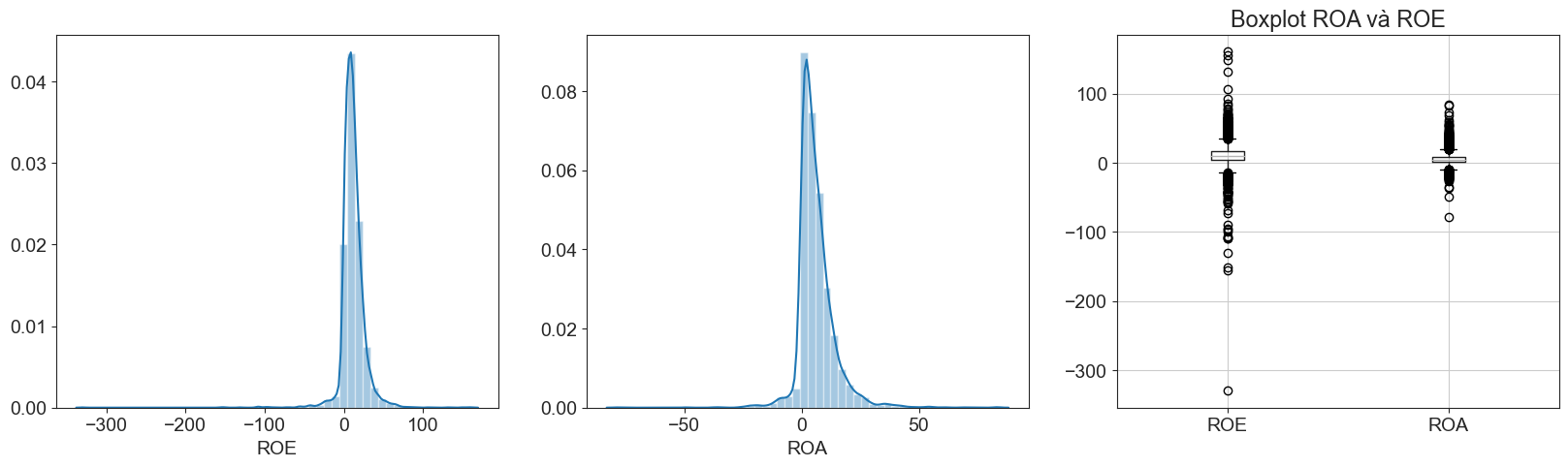
*Biểu đồ 4.3. Biểu đồ mức độ tương quan giữa các biến*

****

**Nhận xét biểu đồ 4.3:** Dựa vào biểu đồ, ta có thể nhận thấy rằng có sự tương quan giữa các biến quan sát và khả năng trả cổ tức của các công ty. Cụ thể, Current ratio, Market to book ratio và Free cash flow tương quan thuận với khả năng trả cổ tức, trong khi đó, Current ratio và leverage tương quan nghịch với khả năng trả cổ tức. Các biến tương quan thấp nhất là Free cash flow, Leverage, Market to book ratio và Company market cap.

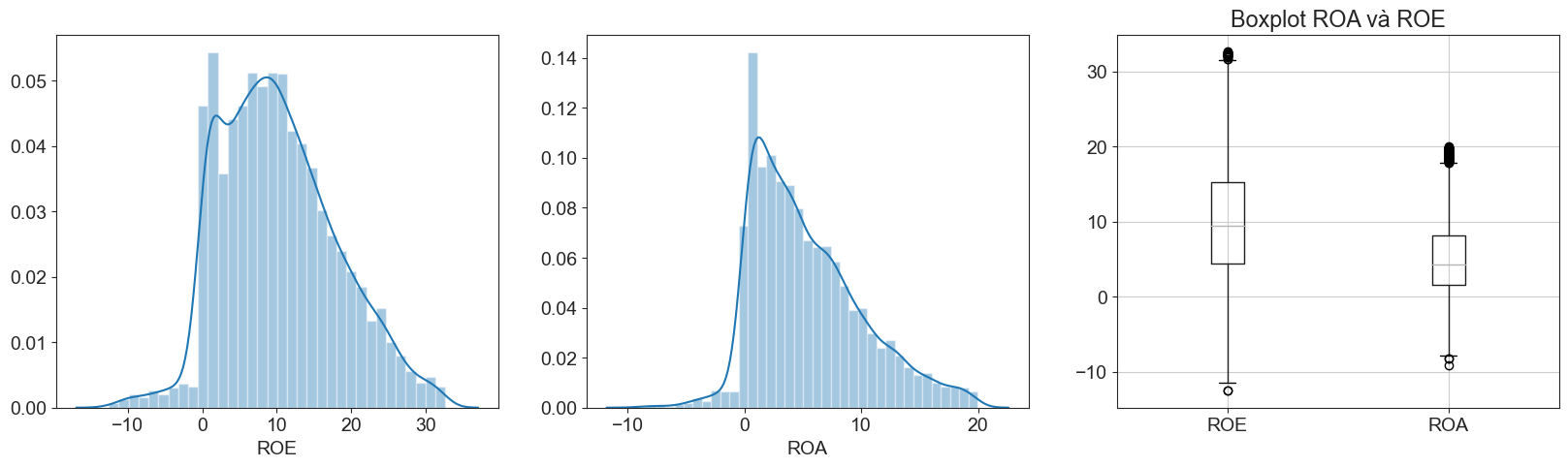
Ngoài ra, ta cũng nhận thấy rằng ROE và ROA là hai biến có mức độ tương quan cao nhất trong các biến quan sát. Vì vậy, để hiểu rõ hơn về sự tương quan của hai biến này với khả năng trả cổ tức, ta cần kiểm tra phân phối dữ liệu và tìm hiểu về các outlier trong hai biến này.

*Biểu đồ 4.4. Boxplot và phân phối của ROA,ROE*

****

**Nhận xét biểu đồ 4.4:** Dữ liệu của ROA và ROE không tuân theo phân phối chuẩn. Boxplot cho thấy dữ liệu không bị lệch, nhưng lại xuất hiện các giá trị ngoại lai. Do đó, quyết định sử dụng phương pháp IQR để loại bỏ các giá trị ngoại lai.

*Biểu đồ 4.5. Boxplot và phân phối của ROA,ROE sau khi loại bỏ outlier*

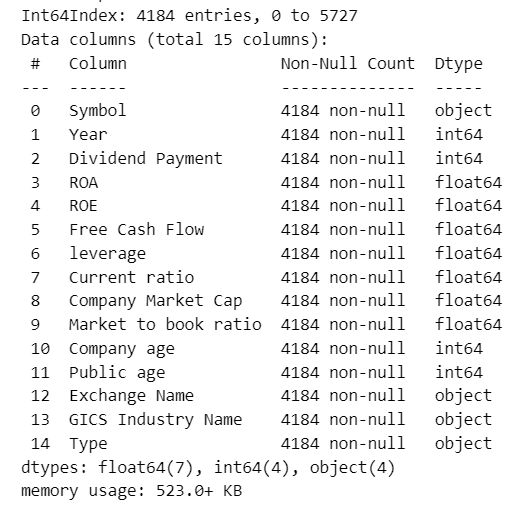


Sau khi loại bỏ outlier bằng IQR thì dữ liệu còn 91,1% so với dữ liệu gốc.

## **4.2. Thống kê mô tả**

### **4.2.1. Tổng quan sau khi xử lý dữ liệu**

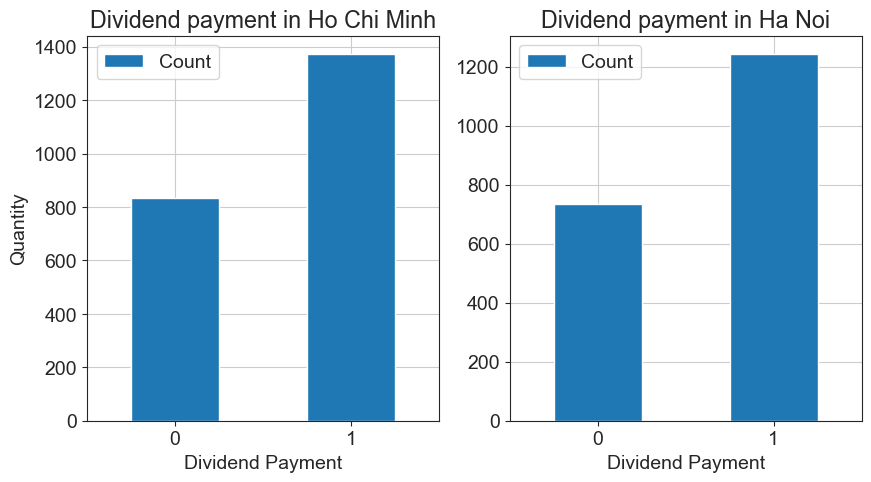
*Hình 4.1. Thông tin dữ liệu*

****

Sau khi xử lý dữ liệu, tập dữ liệu còn lại gồm 4184 hàng và 15 cột quan sát. Trong đó, có 3 cột quan sát là dữ liệu phân loại gồm Symbol, Exchange Name và type. Các cột dữ liệu còn lại đều là dạng số, trong đó có 4 cột dạng số nguyên (Year, Dividend Payment, Company age, Public age) và 7 cột dạng số thực (ROA, ROE, Free cash flow, leverage, current ratio, company market cap và market to book ratio)

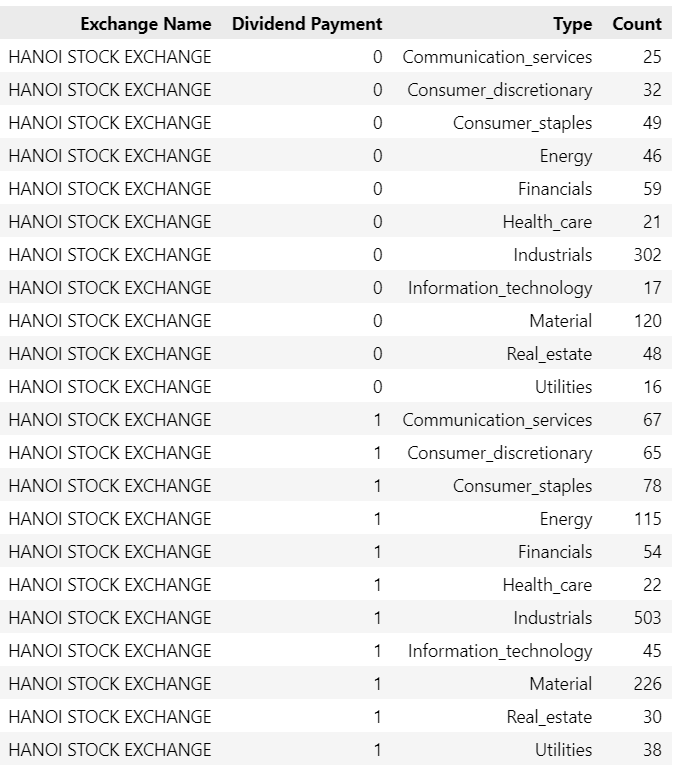
### **4.2.2. Categorical data**

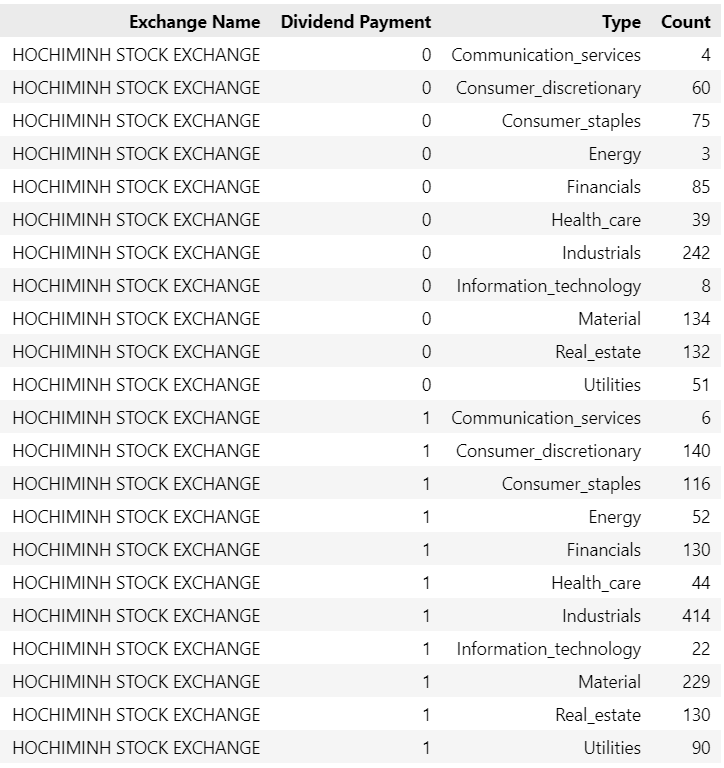
*Biểu đồ 4.6. Khả năng trả cổ tức của các công ty ở 2 sàn*

****

**Nhận xét biểu đồ 4.6:** Nhìn vào 2 biểu đồ có thể thấy khả năng trả cổ tức giữa 2 sàn gần như giống nhau với số công ty có khả năng trả cổ tức cao hơn hẳn số công ty không có khả năng trả cổ tức.

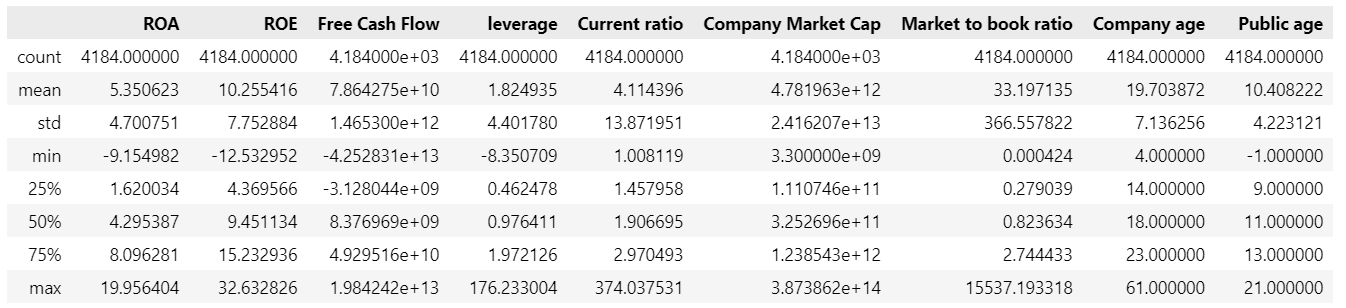
*Hình 4.2. So sánh khả năng trả cổ tức công ty theo ngành tại Hà Nội và Hồ Chí Minh*





### **4.2.3. Numerical data**

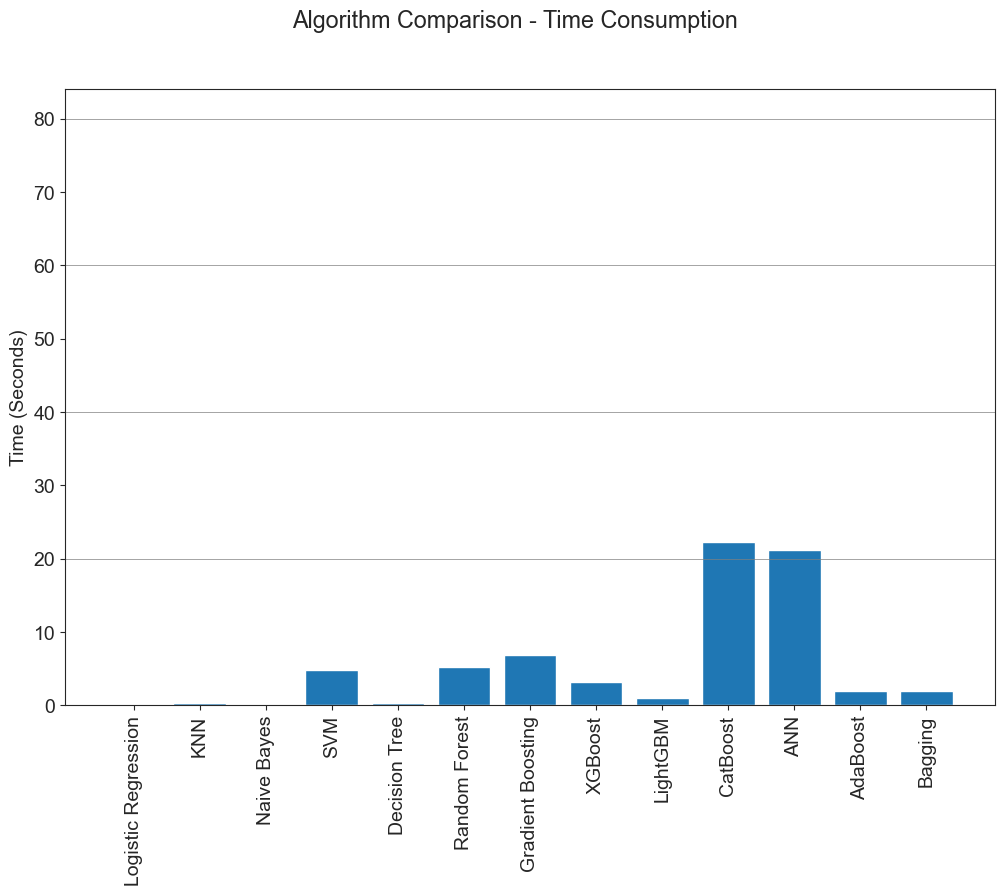
*Hình 4.3. Các chỉ số thống kê của biến định lượng*



## **4.3. Kết quả phân tích dữ liệu**

### **4.3.1. Time consumption**

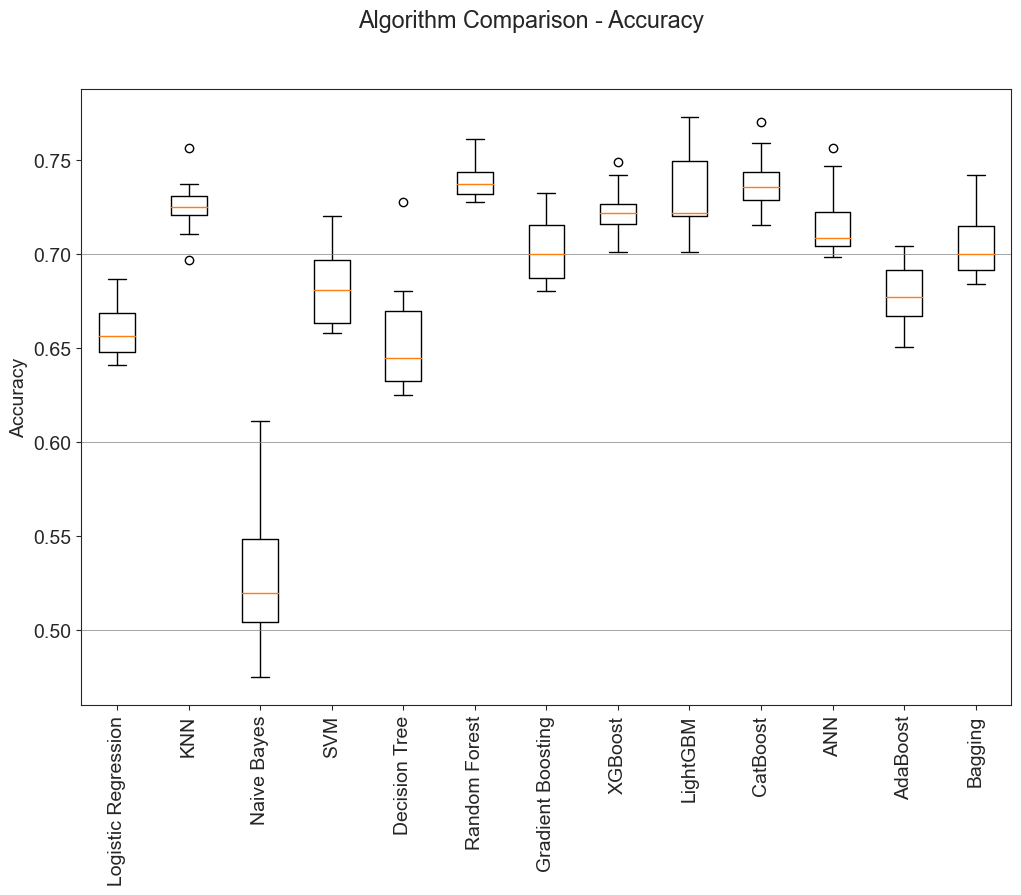
*Biểu đồ 4.7. So sánh thời gian chạy của từng mô hình*

****

**Nhận xét biểu đồ 4.7:** Dựa vào biểu đồ, có thể thấy thời gian chạy nhanh nhất trong các mô hình Machine Learning sử dụng trong nghiên cứu là gần như 0 giây với các thuật toán Logistic Regression, KNN, Naive Bayes và Decision Tree. Thời gian chạy tăng dần từ khoảng 1 đến 5 giây với lần lượt các thuật toán LightGBM, AdaBoost, Bagging, XGBoost, SVM, Random Forest và Gradient Boosting. Hai thuật toán CatBoost và ANN là hai thuật toán có thời gian chạy lâu nhất với hơn 20 giây.

### **4.3.2. Accuracy**

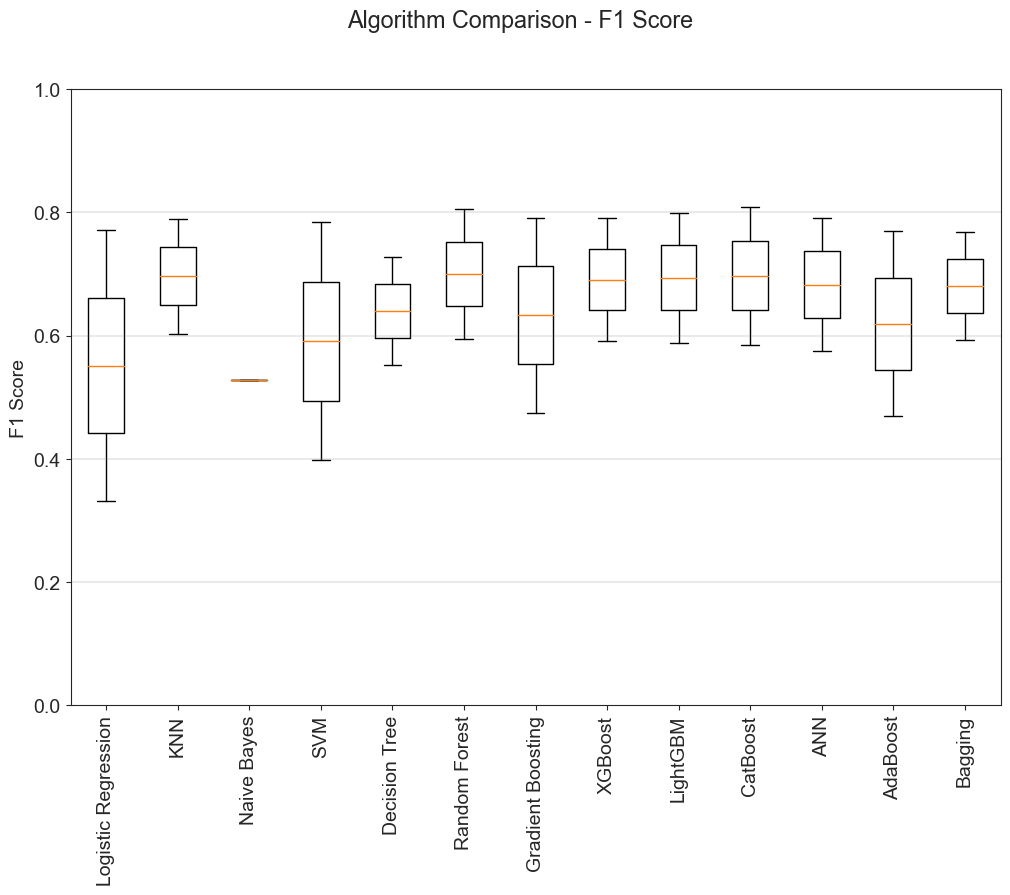
*Biểu đồ 4.8. So sánh Accuracy của từng mô hình*

****

**Nhận xét biểu đồ 4.8:** Dựa trên biểu đồ, có thể nhận thấy mô hình Naive Bayes có độ chính xác thấp nhất (52,8%) và không ổn định. Các mô hình Logistic Regression, SVM, Decision Tree và AdaBoost có phần trăm dự báo trong khoảng từ 65% đến 70%. Gradient Boosting, LightGBM và Bagging dự báo với độ chính xác hơn 70%, nhưng mức độ ổn định không cao. Tiếp theo đó KNN, Random Forest, LightGBM và Catboost dự báo với độ chính xác hơn 72%. Trong số đó, Random Forest (73,6%) và CatBoost (73,9%) có độ chính xác cao nhất và không chênh lệch quá nhiều. Tuy nhiên về mức độ ổn định thì Random Forest ổn định hơn CatBoost.

### **4.3.3. F1 - Score**

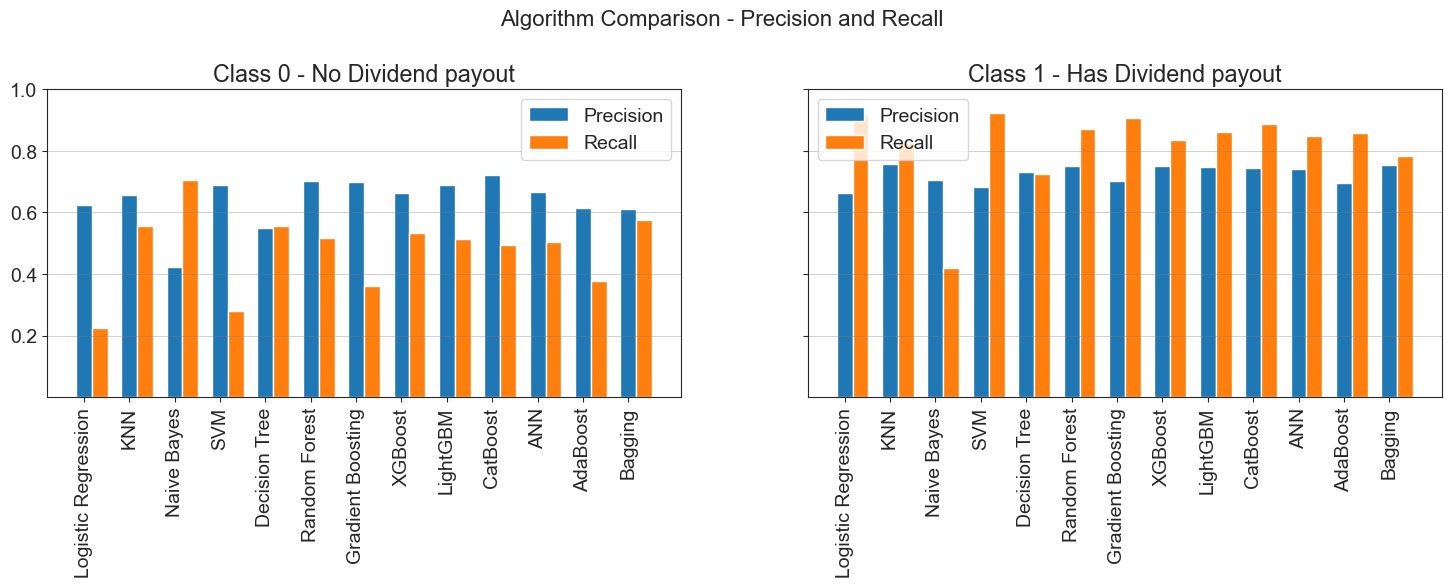
*Biểu đồ 4.9. So sánh F1 - Score của từng mô hình*

****

**Nhận xét biểu đồ 4.9:** Dựa vào biểu đồ trên có thể thấy rằng Naive Bayes có chỉ số F1 thấp nhất ( khoảng 50% ). Tiếp theo đó là mô hình với thuật toán Logictic Regression và SVM ( dưới 60% ). Các mô hình còn lại đều trên 60%, trong đó cao nhất là KNN, Random Forest, XGboost, LightGBM, ANN và Bagging (trên 70%) và mức độ ổn định của chỉ số F1 là gần như ngang nhau.

### **4.3.4. Precision và Recall**

*Biểu đồ 4.10. So sánh Precision và Recall của từng mô hình*

****

**Nhận xét biểu đồ 4.10:**

Lớp 0 (không trả cổ tức):Qua phân tích biểu đồ, ta thấy Precision lớp 0 của Decision Tree, Naive Nayes, AdaBoost và Bagging là thấp nhất (dưới 60%), trong khi các mô hình sử dụng thuật toán khác đều có Precision lớn hơn 60%. Trong đó, SVM, Random Forest, Gradient Boosting và CatBoost có Precision lớp 0 cao nhất (khoảng 70%). Tuy nhiên, Recall lớp 0 của tất cả các thuật toán đều thấp hơn 60%, trong đó thuật toán Logistic Regression có Recall thấp nhất. Ngoại trừ Naive Bayes với Recall lớp 0 cao hơn 70%, các thuật toán khác đều có Recall lớp 0 dưới 55%. Điểm đáng chú ý là mô hình với Decision Tree có tỷ lệ Precision và Recall cân bằng nhất ở lớp 0.

Lớp 1 ( trả cổ tức ):Qua phân tích biểu đồ, ta có thể thấy rằng lớp 1 (trả cổ tức) có Precision cao hơn lớp 0 (không trả cổ tức) với tất cả các thuật toán. Kết quả cao nhất đạt được ở Precision của lớp 1 là với KNN và Random Forest (trên 75%). Tương tự, Recall của các thuật toán cho lớp 1 cũng cao hơn Recall của lớp 0, với SVM và KNN cho kết quả cao nhất (khoảng 90%). Random Forest và Gradient Boosting cũng cho kết quả tương đối cao. Naive Bayes là thuật toán cho kết quả thấp nhất ở Recall của lớp 1. Decision Tree tiếp tục là mô hình cho kết quả cân bằng giữa Precision và Recall của lớp 1.

## **4.4. Đánh giá mô hình machine learning**

### **4.4.1. Các tiêu chí đánh giá mô hình**

Khi đánh giá một mô hình dự báo khả năng trả cổ tức, cần phải cân nhắc các yếu tố khác nhau của các góc nhìn khác nhau, bao gồm: độ chính xác (Precision, Recall và F1-score) và thời gian chạy mô hình. Tùy thuộc vào mục đích sử dụng của mô hình, ta sẽ chọn các chỉ số phù hợp để đánh giá mô hình. Các chỉ số Recall, Precision và F1-score đều liên quan đến độ chính xác của mô hình dự báo khả năng trả cổ tức. Tuy nhiên, các góc nhìn khác nhau có thể có các yêu cầu khác nhau đối với một mô hình dự báo.

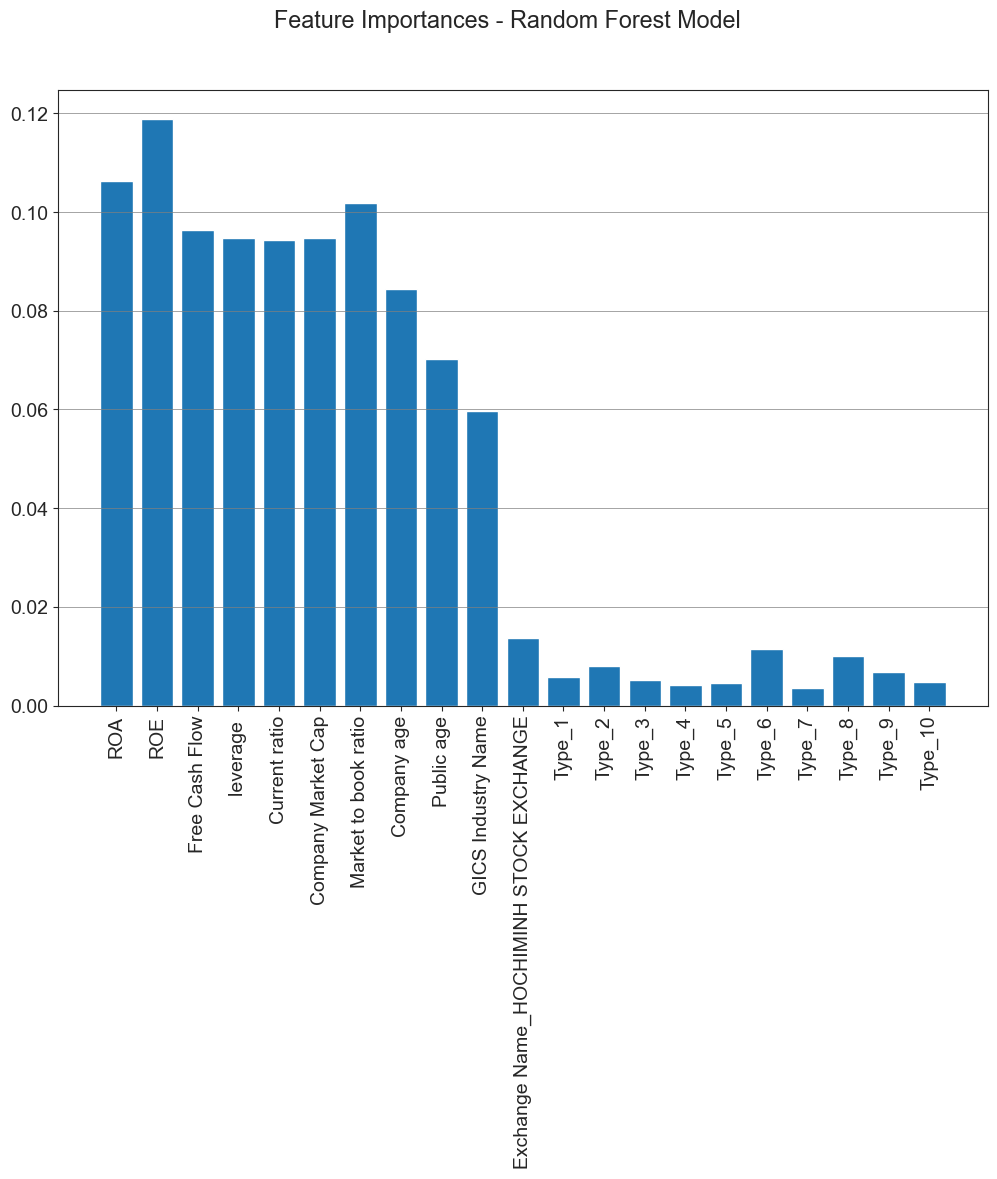
### **4.4.2. Lựa chọn mô hình**

Từ kết quả của các mô hình máy học được áp dụng, mô hình Random Forest cho thấy kết quả tốt nhất trong việc dự báo khả năng trả cổ tức của doanh nghiệp. Với chỉ số Accuracy cao, tỷ lệ Precision và Recall cân bằng tốt ở cả hai lớp 0 và 1, và mức độ ổn định tương đối cao trong các chỉ số F1-score, mô hình Random Forest đáp ứng tốt yêu cầu dự đoán độ chính xác khả năng trả cổ tức của doanh nghiệp.

Một số nguyên nhân cho sự hiệu quả của mô hình Random Forest có thể được giải thích thông qua tính chất của thuật toán. Thứ nhất, trong quá trình xây dựng mô hình, Random Forest là một mô hình Ensemble Learning, nghĩa là nó kết hợp nhiều cây quyết định để tạo thành một mô hình dự đoán. Việc kết hợp này giúp giảm thiểu các hiện tượng overfitting và tăng độ chính xác của mô hình. Nó sử dụng phương pháp bootstrapping để tạo ra các cây quyết định khác nhau cũng giúp tăng độ đa dạng của mô hình và giảm thiểu sự ảnh hưởng của các nhiễu trong dữ liệu (Breiman, 2001). Thứ hai, Random Forest có khả năng xử lý dữ liệu với nhiều features mà không yêu cầu quá nhiều xử lý trước. Các features trong dữ liệu thường có sự tương tác với nhau, và Random Forest có thể xử lý các tương tác này một cách tốt hơn so với các mô hình tuyến tính khác. Thứ ba, Random Forest có khả năng xử lý dữ liệu bị thiếu thông tin (missing data) hoặc nhiễu (noise) một cách hiệu quả. Nó có thể xử lý các giá trị bị thiếu thông tin và nhiễu trong dữ liệu bằng cách sử dụng các cây quyết định khác nhau và kết hợp chúng để đưa ra dự đoán. Cuối cùng, Random Forest là một mô hình linh hoạt và dễ sử dụng. Nó có

thể được sử dụng cho các bài toán dự đoán, phân loại, và cả xác định mức độ quan trọng của các features trong dữ liệu

*Biểu đồ 4.11. Mức độ ảnh hưởng của các biến đến kết quả*

****

**Nhận xét biểu đồ 4.11**: Feature importance là một công cụ quan trọng trong việc đánh giá đóng góp của các biến đầu vào trong quá trình xây dựng mô hình

Random Forest. Phương pháp tính toán feature importance dựa trên việc đo lường mức độ giảm thiểu độ lỗi của các cây quyết định sau khi một biến được loại bỏ khỏi mô hình. Dựa vào biểu đồ 4.5, có thể thấy rằng các biến đều có mức độ đóng góp cao vào kết quả của mô hình. Trong đó cao nhất là ROA, ROE và Market to book ratio. Các biến Free cash flow, leverage, Current ratio, Company maket cap thì có mức độ ảnh hưởng thấp hơn nhưng không đáng kể và gần như bằng nhau. Các biến đầu vào như tên sàn niêm yết và phân loại ngành nghề có mức độ ảnh hưởng thấp hơn hẳn so với các biến nói chung.

## **4.5. Thảo luận về kết quả và so sánh với các nghiên cứu trước đây**

Trong nghiên cứu dự báo khả năng trả cổ tức của doanh nghiệp tại Việt Nam, mô hình máy học (ML) có thể giải quyết được những vấn đề mà mô hình truyền thống không thể, như xử lý dữ liệu lớn và phức tạp, phân tích các yếu tố tác động đến cổ tức một cách hiệu quả. Ngoài ra, mô hình ML còn có khả năng tự động hóa việc phân tích dữ liệu và tạo ra các mô hình dự báo, giúp tiết kiệm thời gian và chi phí so với các phương pháp truyền thống. Mô hình ML cũng rất linh hoạt, có thể học và cập nhật mô hình dự báo dựa trên dữ liệu mới, giúp đưa ra dự báo cổ tức chính xác hơn theo thời gian. Vì vậy, độ chính xác của mô hình ML thường cao hơn so với mô hình truyền thống, nhờ khả năng phân tích và mô hình hóa mối quan hệ giữa các yếu tố ảnh hưởng đến cổ tức.

Để đánh giá hiệu quả của mô hình random forest trong dự đoán khả năng trả cổ tức của các doanh nghiệp Việt Nam, chúng tôi đã so sánh kết quả của nghiên cứu của mình với các nghiên cứu trước đó sử dụng phương pháp máy học tương tự. Kết quả cho thấy mô hình random forest đạt được kết quả gần bằng so với các nghiên cứu trước đó. Trong nghiên cứu của Lien và Rong (2020), họ sử dụng mô hình Support Vector Machines để dự đoán khả năng trả cổ tức và đạt được độ chính xác đạt 74,4%. Trong khi đó, kết quả của mô hình random forest trong bài nghiên cứu trên đạt 73,4%. Ngoài nghiên cứu của Lien và Rong (2020), còn có nhiều nghiên cứu khác cũng sử dụng các phương pháp máy học để dự đoán khả năng trả cổ tức của các doanh nghiệp. Ví dụ, nghiên cứu của Tran và đồng nghiệp (2019) sử dụng mô hình Artificial Neural Network (ANN) và đạt được độ chính xác là 70,7%. Ngoài ra, nghiên cứu của Li và Wang (2017) sử dụng mô hình Logistic Regression và đạt được độ chính xác là 64,3%. So sánh các kết quả này với kết quả của mô hình random forest trong nghiên cứu trên, ta thấy mô hình random forest đạt

được độ chính xác cao hơn so với các mô hình ANN và Logistic Regression. Tóm lại, kết quả nghiên cứu cho thấy mô hình random forest là một phương pháp hiệu quả để dự đoán khả năng trả cổ tức của các doanh nghiệp Việt Nam, và đạt được kết quả tốt hơn so với nhiều phương pháp máy học khác đã được sử dụng trong các nghiên cứu trước đó. Điều này cho thấy rằng, mô hình random forest có khả năng dự đoán khả năng trả cổ tức của doanh nghiệp Việt Nam một cách hiệu quả.

## **4.6. Những giới hạn và hướng phát triển trong tương lai**

### **4.6.1. Hạn chế về phạm vi và nội dung**

Quy mô của cơ sở dữ liệu còn tương đối hẹp, chưa giải thích được tất cả mọi khía cạnh, chủ yếu là các công ty ở Việt Nam. Có một số doanh nghiệp không công bố thông tin đầy đủ về việc trả cổ tức, các chỉ số kinh tế hoặc thông tin được công bố có thể không chính xác hoặc thiếu tính minh bạch. Điều này làm giảm hiệu quả sử dụng Machine Learning để dự đoán các khoản trả cổ tức của doanh nghiệp.

Đồng thời, việc chỉ sử dụng dữ liệu về số liệu ở các báo tài chính mà bỏ qua một số yếu tố phi tài chính, chẳng hạn như thị trường, chính sách chính phủ, hoặc môi trường kinh doanh có thể làm giảm độ chính xác cho dự đoán của Machine Learning. Thuật toán mà bài nghiên cứu nhóm sử dụng chỉ đơn thuần là thuật toán phân loại chưa sử dụng mô hình hồi quy và phân cụm.

### **4.6.2. Hạn chế trong quá trình thực hiện nghiên cứu**

Nhóm nghiên cứu mong muốn đề cập đến yếu tố con người trong đề tài của mình. Các thành viên trong nhóm vẫn còn là sinh viên nên chưa có nhiều kinh nghiệm nghiên cứu. Bên cạnh đó, nhóm còn phải chia sẻ thời gian với các công việc học tập và rèn luyện khác, vì vậy không tránh khỏi việc bài nghiên cứu còn những thiếu xót nhất định. Hơn nữa, đề tài của nhóm vẫn còn tồn đọng nhiều hạn chế, bao gồm việc chưa đào sâu vào các đối tượng doanh nghiệp khác nhau trong xã hội do phạm vi nghiên cứu khá rộng và bị giới hạn về thời gian.

### **4.6.3. Hướng nghiên cứu**

Các bài nghiên cứu trong tương lai có thể mở rộng hơn những yếu tố khác ảnh hưởng như thế nào đến kết quả dự đoán của Machine Learning trong khả năng khả năng trả cổ tức để có cái nhìn bao quát hơn. Trong tương lai, để tăng độ chính xác của mô hình, nhà nghiên cứu có thể kết hợp nhiều loại dữ liệu khác nhau như dữ liệu tài chính,

dữ liệu về ngành công nghiệp, dữ liệu thị trường và dữ liệu xã hội. Việc kết hợp nhiều loại dữ liệu khác nhau sẽ cung cấp cho mô hình thông tin đầy đủ hơn về một doanh nghiệp. Đồng thời, việc tích hợp các phương pháp dự báo kinh tế học và thống kê vào các mô hình dự đoán cổ tức để tăng cường khả năng dự đoán của mô hình.

Sử dụng mô hình học sâu (Deep learning): Mô hình học sâu có thể học được các đặc trưng phức tạp hơn so với các mô hình học máy truyền thống. Chẳng hạn, một mô hình mạng nơ-ron sâu có thể học được mối quan hệ giữa các yếu tố như tỷ lệ lợi nhuận, tỷ lệ nợ vay và tốc độ tăng trưởng doanh thu. Các mô hình học sâu có thể giúp khắc phục vấn đề thiếu dữ liệu đáng tin cậy bằng cách học đại diện tự động của các đặc trưng của doanh nghiệp từ các mô hình khác hoặc dữ liệu bổ sung. Ngoài ra, việc sử dụng các mô hình học sâu cũng có thể giúp mô hình học được các mối quan hệ phức tạp giữa các biến đầu vào và đầu ra, đồng thời có khả năng xử lý dữ liệu không cấu trúc như dữ liệu văn bản hoặc hình ảnh. Sử dụng mô hình học tăng cường (Reinforcement learning): Mô hình học tăng cường có thể được sử dụng để tối ưu hóa chiến lược đầu tư. Bằng cách sử dụng mô hình này, nhà nghiên cứu có thể đưa ra quyết định đầu tư dựa trên khả năng trả cổ tức của một doanh nghiệp, trong đó cổ tức được coi là một phần thưởng hoặc khoản thưởng cho nhà đầu tư.

Tóm lại, những hạn chế kể trên sẽ mở ra những đề tài nghiên cứu mới và chi tiết hơn về ứng dụng Machine Learning trong dự đoán khả năng chi trả cổ tức của các doanh nghiệp Việt Nam.

# 

# **CHƯƠNG 5: KẾT LUẬN VÀ KHUYẾN NGHỊ**

## **5.1. Kết luận**

Trong bài nghiên cứu này, chúng tôi đã áp dụng các mô hình máy học để dự báo cổ tức cho các công ty niêm yết trên thị trường chứng khoán tại Việt Nam. Kết quả cho thấy rằng, trong số các mô hình được đánh giá, mô hình Random Forest đã cho ra kết quả ổn định và chính xác hơn so với các phương pháp truyền thống.

Các biến đầu vào quan trọng nhất trong việc dự báo cổ tức là ROA, ROE và Market to book ratio. Ngoài ra, bài nghiên cứu cũng đã chỉ ra một số hạn chế của mô hình máy học, bao gồm sự phụ thuộc vào dữ liệu đầu vào và thời gian tính toán tăng lên khi sử dụng các mô hình phức tạp hơn. Tuy nhiên, những hạn chế này có thể được giải quyết bằng cách sử dụng các kỹ thuật xử lý dữ liệu và tối ưu hóa mô hình.

Tổng quan, bài nghiên cứu này cho thấy mô hình máy học, đặc biệt là mô hình Random Forest, có khả năng cung cấp kết quả dự báo cổ tức chính xác và ổn định hơn so với các phương pháp truyền thống. Các kết quả này cung cấp một tầm nhìn mới cho việc sử dụng máy học trong dự báo cổ tức tại Việt Nam. Nghiên cứu này cũng đề xuất các hướng nghiên cứu tiếp theo để tối ưu hóa mô hình máy học và giải quyết các hạn chế của nó.

## **5.2. Khuyến nghị**

Dưới góc nhìn của nhà đầu tư, mô hình Random Forest có thể được khuyến nghị như một công cụ hữu ích để đưa ra quyết định đầu tư vào các doanh nghiệp có tiềm năng trả cổ tức cao. Random Forest là một trong những mô hình dự báo cổ tức hiệu quả nhất trong bài nghiên cứu này, với độ chính xác cao và mức độ ổn định tốt. Những thông tin đáng chú ý về Precision và Recall cho cả hai lớp cũng cho thấy mô hình Random Forest có khả năng dự báo đáng tin cậy cả khi doanh nghiệp không trả cổ tức và khi doanh nghiệp trả cổ tức.

Dưới góc nhìn của nhà quản trị, việc sử dụng mô hình dự báo cổ tức có thể giúp quản trị viên đưa ra các quyết định chiến lược dựa trên việc định giá và phân tích các doanh nghiệp theo tiềm năng trả cổ tức. Mô hình Random Forest là một công cụ mạnh mẽ để phân tích và đưa ra dự báo về khả năng trả cổ tức của doanh nghiệp, giúp quản trị viên đưa ra các quyết định đầu tư và quản lý rủi ro một cách chính xác và hiệu quả.

Dưới góc nhìn của chính phủ và các cơ quan liên quan, việc áp dụng mô hình dự báo cổ tức có thể giúp cơ quan chức năng đưa ra các chính sách và quy định hợp lý và hiệu quả về việc quản lý các doanh nghiệp trong lĩnh vực tài chính. Mô hình Random Forest có thể giúp các cơ quan chức năng đưa ra các quyết định hợp lý về định giá và quản lý các doanh nghiệp, cũng như hỗ trợ trong việc phát hiện các rủi ro tiềm ẩn trong hoạt động kinh doanh của các doanh nghiệp. Điều này có thể giúp cơ quan chức năng tăng cường sự giám sát và kiểm soát, đảm bảo sự ổn định và phát triển bền vững của thị trường tài chính.

# ***TÀI LIỆU THAM KHẢO***

Affandi, F., SUNARKO, B., & YUNANTO, A. (2019). The impact of cash ratio, debt to equity ratio, receivables turnover, net profit margin, return on equity, and institutional ownership to dividend payout ratio. *Journal of Research in Management, 1*(4).

Al-Kuwari, D. (2012). Are large shareholders conducting influential monitoring in emerging markets? An investigation into the impact of large shareholders on dividend decisions: The case of Kuwait. *Research in World Economy, 3*(2), 52-67.

Al-Malkawi, H.-A. N. (2008). Factors Influencing Corporate Dividend Decision: Evidence from Jordanian Panel Data. *International journal of business, 13*(2).

Al-Shubiri, F. N. (2011). Determinants of changes dividend behavior policy: Evidence from the Amman stock exchange. *Far East Journal of Psychology and Business, 4*(1), 1-15.

Al‐Najjar, B., & Hussainey, K. (2009). The association between dividend payout and outside directorships. *Journal of Applied Accounting Research*.

Alexander, N. (2017). Factors affecting earnings management in the Indonesian Stock Exchange. *Journal of Finance and Banking Review, 2*(2), 8-14.

Amalia, S., & Hermanto, S. B. (2018). Pengaruh Profitabilitas, Kepemilikan Manajerial, Leverage dan Growth terhadap Kebijakan Dividen. *Jurnal Ilmu dan Riset Akuntansi (JIRA), 7*(8).

Amidu, M., & Abor, J. (2006). Determinants of dividend payout ratios in Ghana. *The journal of risk finance*.

Atrill, P., & McLaney, E. (2002). *Management accounting for non-specialists*: Pearson Higher Education.

Baker, M., & Wurgler, J. (2004). A catering theory of dividends. *The Journal of finance, 59*(3), 1125-1165.

Breiman, L. (1996). Bagging predictors. *Machine learning, 24*, 123-140.

Breiman, L. (2001). Random forests. *Machine learning, 45*, 5-32.

Brigham, E. F., & Houston, J. F. (2021). *Fundamentals of financial management: Concise*: Cengage Learning.

Chaniago, Y. F., & Ekadjaja, A. (2022). Determinant of dividend payout ratiosin consumer goods company. *Jurnal Ekonomi, 27*(03), 100-118.

Chen, J., & Dhiensiri, N. (2009). Determinants of dividend policy: The evidence from New Zealand.

Chen, T., & Guestrin, C. (2016). *Xgboost: A scalable tree boosting system.* Paper presented at the Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining.

Cortes, C., & Vapnik, V. (1995). Support vector machine. *Machine learning, 20*(3), 273-297.

Cui, Y., Chen, R., Chu, W., Chen, L., Tian, D., Li, Y., & Cao, D. (2021). Deep learning for image and point cloud fusion in autonomous driving: A review. *IEEE Transactions on Intelligent Transportation Systems, 23*(2), 722-739.

Dorogush, A. V., Ershov, V., & Gulin, A. (2018). CatBoost: gradient boosting with categorical features support. *arXiv preprint arXiv:1810.11363*.

Faccio, M., Lang, L. H. P., & Young, L. (2001). Dividends and expropriation. *American economic review, 91*(1), 54-78.

Fama, E. F., & French, K. R. (2002). Testing trade-off and pecking order predictions about dividends and debt. *Review of financial studies*, 1-33.

Freund, Y., & Schapire, R. E. (1997). A decision-theoretic generalization of on-line learning and an application to boosting. *Journal of computer and system sciences, 55*(1), 119-139.

Friedman, J. H. (2001). Greedy function approximation: a gradient boosting machine. *Annals of statistics*, 1189-1232.

Gugler, K., & Yurtoglu, B. B. (2003). Corporate governance and dividend pay-out policy in Germany. *European economic review, 47*(4), 731-758.

Han, J., Kamber, M., & Pei, J. (2012). Data mining concepts and techniques third edition. *University of Illinois at Urbana-Champaign Micheline Kamber Jian Pei Simon Fraser University*.

Harryono, S. (2020). The effect analysis of management ownership structure, debt policy and risk on financial performance of Indonesian manufacturing companies. *Journal of Economy, Accounting and Management Science (JEAMS), 1*(2), 24-32.

Hastie, T., Tibshirani, R., Friedman, J. H., & Friedman, J. H. (2009). *The elements of statistical learning: data mining, inference, and prediction* (Vol. 2): Springer.

Ho, H. (2003). Dividend policies in Australia and Japan. *International Advances in Economic Research, 9*(2), 91-100.

Islam, R. M., Khan, M. N., Oldroyd, J. C., Rana, J., Magliago, D. J., Chowdhury, E. K., . . . Hossain, M. B. (2021). Prevalence of diabetes and prediabetes among Bangladeshi adults and associated factors: Evidence from the Demographic and Health Survey, 2017-18. *medRxiv*, 2021.2001. 2026.21250519.

Jain, A., Shakya, A., Khatter, H., & Gupta, A. K. (2019). *A smart system for fake news detection using machine learning.* Paper presented at the 2019 International conference on issues and challenges in intelligent computing techniques (ICICT).

Jensen, D. E. (1987). *The information content of cash flow measures in regard to enterprise dividend policy*: The Pennsylvania State University.

Ke, G., Meng, Q., Finley, T., Wang, T., Chen, W., Ma, W., . . . Liu, T.-Y. (2017). Lightgbm: A highly efficient gradient boosting decision tree. *Advances in neural information processing systems, 30*.

Mehar, A. (2002). Corporate governance and dividend policy.

Myers, M., & Bacon, F. (2004). The determinants of corporate dividend policy. *Academy of Accounting and Financial Studies Journal, 8*(3), 17.

Nguyen, T. T. S., & Do, P. M. T. (2020). Classification optimization for training a large dataset with Naïve Bayes. *Journal of Combinatorial Optimization, 40*(1), 141-169.

Nizar Al‐Malkawi, H. A. (2007). Determinants of corporate dividend policy in Jordan: an application of the Tobit model. *Journal of Economic and Administrative Sciences, 23*(2), 44-70.

Panjaitan, S. A., & Akmalia, A. (2020). Pengaruh Profitabilitas, IOS, Leverage, Kebijakan Dividen Terhadap Nilai Perusahaan Sub-sektor Properti dan Real Estate Yang Terdaftar di BEI 2014-2018. *Prosiding UMY Grace, 1*(1), 543-558.

Rumelhart, D. E., Hinton, G. E., & Williams, R. J. (1986). Learning representations by back-propagating errors. *nature, 323*(6088), 533-536.

Simbolon, K., & Sampurno, R. D. (2017). Analisis Pengaruh Firm Size, DER, Asset Growth, ROE, EPS, Quick Ratio dan Past Dividend terhadap Dividend Payout Ratio (Studi pada Perusahaan Manufaktur yang Terdaftar di BEI Tahun 2011-2015). *Diponegoro Journal of Management, 6*(3), 315-327.

Supriyono, R., & Herdhayinta, H. (2019). Determinants of Bank Profitability: The case of the regional development bank (BPD Bank) in Indonesia. *Journal of Indonesian Economy and Business, 34*(1), 1-17.

Viveronica, D. A., & Amanah, L. (2017). Pengaruh Kebijakan Perusahaan dan Struktur Kepemilikan Terhadap Nilai Perusahaan. *Jurnal Ilmu dan Riset Akuntansi (JIRA), 6*(6).

Zen, S. D., & Herman, M. (2012). Pengaruh Harga Saham Umur Perusahaan dan Rasio Profitabilitas Perusahaan Terhadap Tindakan Perataan Laba yang Dilakukan Oleh Perusahaan Perbankan yang Terdaftar Di Bursa Efek Jakarta. *Jurnal Akuntansi & Manajemen, 2*(2), 57-71.