ĐẠI HỌC QUỐC GIA THÀNH PHỐ HỒ CHÍ MINH

ĐẠI HỌC KINH TẾ - LUẬT

FINAL PROJECT

HYBRID ML MODEL FOR FINANCIAL HEALTH CLASSIFICATION

SUBJECT : MACHINE LEARNING AND ARTIFICIAL INTELLIGENCE IN FINANCE

Course Code : 243TC6701

Full name : Phan Thị Thu Huyền

Instructor : Phan Huy Tâm

Tp.HCM, 12/7

1. Introduction

Financial distress and corporate failure can lead to substantial social and economic costs, affecting a wide range of stakeholders including owners, managers, employees, lenders, suppliers, customers, communities, and governments. Consequently, the ability to assess and predict a firm's financial health has become a critical issue in both academic and business contexts.

Traditional methods for predicting bankruptcy or financial difficulties, such as logistic regression and statistical scoring models, have been widely used in credit risk assessment and investment decision-making. However, with the rise of machine learning, recent studies have demonstrated that advanced algorithms – especially hybrid and ensemble approaches – often outperform classical techniques in prediction accuracy and robustness.

In particular, combining unsupervised learning (e.g., clustering) with supervised classification models has gained attention as a promising hybrid strategy. Clustering methods can uncover latent groupings and identify outliers, while classifiers trained on clustered data can generalize well to new observations. Prior research (e.g., Hsieh, 2013) has shown that filtering unrepresentative data using clustering techniques can improve classification performance by reducing noise and overfitting.

Inspired by these advancements, this study aims to develop a robust and scalable framework for evaluating corporate financial health. Specifically, we:

- Apply clustering algorithms to group firms based on financial ratios and identify natural financial health categories (e.g., "Strong", "Moderate", "Weak", "At Risk").
- Assign labels to clusters and train classification models to automate the prediction process for new firms.
- Evaluate model performance using modern metrics such as Silhouette Score, Calinski-Harabasz Index, and classification accuracy.
- Apply the trained models to historical data (2021–2023) to assess financial health trends across time and industries.

This integrated approach not only enhances interpretability but also offers practical implications for stakeholders in investment, policy, and financial monitoring. It aligns with

the ongoing trend of adopting machine learning-based decision-support systems in finance, while contributing to the development of more resilient and interpretable models.

2. Theoretical Background

2.1. Key Financial Concepts

In corporate finance, financial ratios are essential tools for evaluating a company's overall health, operational efficiency, solvency, and profitability. These ratios are derived from core components of financial statements and are commonly grouped based on the specific aspect of performance they measure.

2.1.1. Liquidity Ratios

Liquidity ratios assess a firm's ability to meet its short-term financial obligations using current assets.

Current Ratio: $$\text{Current Ratio} = \frac{\text{Current Assets}}{\text{Current Liabilities}}$$

A ratio above 1 indicates that the company is capable of covering its short-term liabilities with its short-term assets. However, an excessively high value may signal inefficiencies in asset utilization.

Quick Ratio : $$\text{Quick Ratio} = \frac{\text{Current Assets} - \text{Inventory}}{\text{Current Liabilities}}$$

This ratio removes inventory from current assets, offering a more conservative view of liquidity.

Cash Ratio : $$\text{Cash Ratio} = \frac{\text{Cash and Cash Equivalents}}{\text{Current Liabilities}}$$

Measures the company's ability to settle short-term obligations using only its most liquid assets.

2.1.2. Leverage Ratios

Debt-to-Assets Ratio : $$\text{Debt-to-Assets} = \frac{\text{Total Liabilities}}{\text{Total Assets}}$$

Indicates the proportion of a firm's assets that are financed through debt.

Debt-to-Equity Ratio: $$\text{Debt-to-Equity} = \frac{\text{Total Liabilities}}{\text{Shareholders' Equity}}$$

Reflects the company's reliance on borrowed funds compared to shareholder investment.

2.1.3. Efficiancy Ratios :

Efficiency ratios evaluate how well a company manages its short-term assets and operations.

Receivables Turnover : $$\text{Receivables Turnover} = \frac{\text{Net Revenue}}{\text{Average Accounts Receivable}}$$

Shows how many times a firm collects its receivables during a period.

Inventory Turover : $$\text{Inventory Turnover} = \frac{\text{Cost of Goods Sold}}{\text{Average Inventory}}$$

Indicates how effectively inventory is being managed and sold.

2.1.4. Profitability Ratios

These ratios measure a company's ability to generate earnings relative to its resources.

Return on Assets (ROA) : $$\text{ROA} = \frac{\text{Net Income}}{\text{Total Assets}}$$

Assesses the efficiency of asset usage in generating profit.

Return on Equity (ROE) : $$\text{ROE} = \frac{\text{Net Income}}{\text{Shareholders' Equity}}$$

Evaluates the return on shareholder investment.

Net Profit Margin : $$\text{Net Profit Margin} = \frac{\text{Net Income}}{\text{Net Revenue}}$$

Indicates how much of each dollar earned translates into profit.

2.1.5. Cash Flow Ratio

Operating Cash Flow Ratio : $$\text{Operating Cash Flow Ratio} = \frac{\text{Cash Flow from Operating Activities}}{\text{Current Liabilities}}$$

Measures the firm's ability to cover short-term liabilities with cash generated from its core operations.

2.2. Data Standardization Theory

Financial ratios often differ greatly in scale and units (e.g., some are percentages, some are large absolute values), which can distort distance-based models such as KMeans, DBSCAN, or Gaussian Mixture Models.

To address this, all numerical features were standardized using StandardScaler, which transforms the data to have a mean of 0 and standard deviation of 1, as follows:

$$z = \frac{x - \mu}{\sigma}$$

Where:

- xxx is the original value
- μ is the mean of the feature
- σ is the standard deviation.

This transformation ensures that each feature contributes equally to the analysis and prevents variables with larger scales from dominating the clustering results. Standardization is particularly important for distance-based algorithms, as they are sensitive to feature magnitude. Without standardization, features with larger numerical ranges would disproportionately influence cluster formation.

2.3. Principal Component Analysis (PCA)

Principal Component Analysis (PCA) is a linear dimensionality reduction technique that transforms high-dimensional data into a lower-dimensional space while retaining most of the original information. In this study, PCA was applied to:

- Reduce dimensionality from 11 financial ratios to two principal components (PCA1 and PCA2) for easier visualization.
- Reveal the underlying distribution structure of the data, supporting the selection of appropriate clustering models.
- Avoid multicollinearity, which is common among financial indicators.

The PCA results allow the data to be visualized in a 2D scatter plot, making it possible to observe the shape and density of potential clusters.

2.4. Clustering Algorithms

This study evaluates several clustering algorithms to identify groups of firms based on financial indicators. Each algorithm has specific advantages and limitations depending on the data structure:

- **K-Means**: A centroid-based algorithm that partitions data into equally shaped spherical clusters. It performs well when clusters are balanced and well-separated but is sensitive to outliers and not suitable for irregular or overlapping distributions.
- **DBSCAN (Density-Based Spatial Clustering of Applications with Noise)**: Groups points based on density. It handles noise and outliers effectively and does not require specifying the number of clusters in advance. However, it is highly sensitive to the choice of the eps (radius) parameter.
- **DBSCAN + K-Means**: A hybrid approach that first removes outliers using DBSCAN and then applies K-Means on the filtered data. This combination leverages the strengths of both methods but requires careful parameter tuning.
- **GMM (Gaussian Mixture Model)**: A probabilistic model that assumes data is generated from a mixture of several Gaussian distributions. It works well for elliptical and overlapping clusters and provides soft classification (probabilistic assignment), offering more flexibility than K-Means.
- **HDBSCAN (Hierarchical DBSCAN)**: An advanced version of DBSCAN that eliminates the need to specify eps, and better identifies clusters of varying densities. It is more robust for complex datasets and provides a hierarchy of clusters.

Conclusion: Based on the nature of the dataset, four models were selected for comparative analysis: DBSCAN**,** DBSCAN + K-Means**,** GMM, and HDBSCAN**.**

2.5. Classification Algorithms

To predict financial health labels assigned through clustering, supervised learning algorithms are used. The following models were considered:

Random Forest: An ensemble learning method based on decision trees. It is robust, easy to interpret, and performs well on high-dimensional data. However, it can be biased toward majority classes in imbalanced datasets.

XGBoost (Extreme Gradient Boosting): A powerful gradient boosting framework that builds models sequentially to minimize errors. It handles class imbalance more effectively and tends to outperform traditional algorithms in structured data tasks.

XGBoost is widely used in competitions and industry applications due to its high accuracy and efficiency.

3. Data

3.1. Overview of the Original Data

The original dataset was compiled from financial statements, including balance sheets, income statements, cash flow statements, financial notes, and financial plans. The data was collected from 4 Excel files corresponding to the years 2021–2024, covering 1,547 companies, resulting in 1,547 rows.

3.2. Data Preprocessing

3.2.1. Loading and Cleaning Data by Year

Column names were standardized by:

- Removing redundant parts from column names
- Dropping unnecessary columns

3.2.2. Merging Multi-Year Data

After preprocessing, the files were merged into a single consolidated dataset (Vietnam_merged.xlsx), with an additional "Year" column to distinguish between years.

3.2.3. Handling Missing Values

The percentage of missing values was calculated for each column in each year. Any column with over 40% missing values in any year was removed. Additionally, rows with more than 0.5% missing columns were also filtered out.

3.2.4. Selecting Core Financial Columns and Calculating Ratios

The analysis focused on key indicators used in financial health assessment, including: current assets, short-term debt, inventory, receivables, equity, net revenue, net profit after tax, and cash flow from operating activities.

Based on these columns, the following financial ratios were calculated:

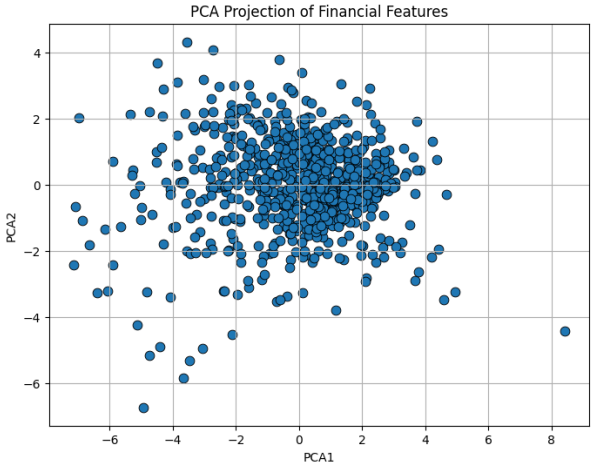
- **Liquidity ratios**: current ratio, quick ratio, cash ratio
- **Leverage ratios**: debt-to-assets, debt-to-equity
- **Efficiency ratios**: receivables turnover, inventory turnover
- **Profitability ratios**: ROA, ROE, net profit margin
- **Cash flow ratio**: operating cash flow ratio

3.2.5. Exporting Cleaned Data

The cleaned and enriched dataset was exported in .csv format for use in the next analysis phase.

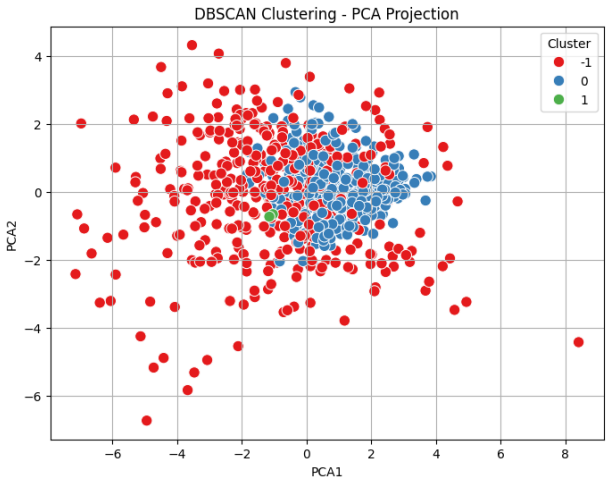
4. Enterprise Clustering

4.1. Identifying the Data Distribution Pattern



The data points exhibit an elliptical distribution, densely concentrated near the center (0,0) and gradually becoming sparser toward the edges. There is no visually distinct separation between clusters, suggesting a dispersed or overlapping structure.

4.2. DBSCAN model



A total of **381 observations**, accounting for nearly half of the dataset, were labeled as outliers. Additionally, **Cluster 1 contains only 3 data points**, indicating an unstable and insignificant cluster formation.

The **Silhouette Score is -0.063**, suggesting poor cluster separation. Data points within the same cluster are not significantly more similar to each other than to those in other clusters. A negative score indicates that the clustering result is of **low quality** and may even distort the analysis.

Conclusion: DBSCAN is not suitable for the current data distribution due to the following reasons:

- The data exhibits a **relatively uniform density**, with no clearly defined cluster boundaries.
- DBSCAN is **highly sensitive to the parameters** eps and min_samples, which limits its effectiveness in this context.
- The algorithm **misclassifies a large number of points as outliers** and produces **unreliable clusters**.

4.3. DBSCAN Combined with K-Means Model

4.3.1. Parameter Combination Analysis

To identify the optimal combination of parameters for outlier removal using DBSCAN prior to clustering with KMeans, multiple configurations of **eps**, **min_samples**, and **number of clusters (k)** were tested. The table below summarizes the top 5 combinations ranked by **Silhouette Score**:

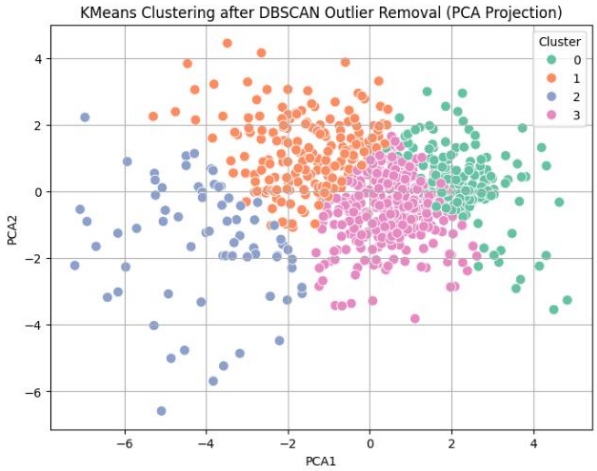
Top 5 tổ hợp có Silhouette Score cao nhất:

eps	min_samples	k	silhouette_score	n_samples
1.00	8	3	0.28	381
1.00	5	3	0.28	411
1.00	3	3	0.25	463
1.50	8	5	0.24	580
1.00	5	4	0.23	411

The best-performing combination was:

- eps = 1.00, min_samples = 5 or 8, and k = 3
- Achieving the **highest Silhouette Score of 0.28**

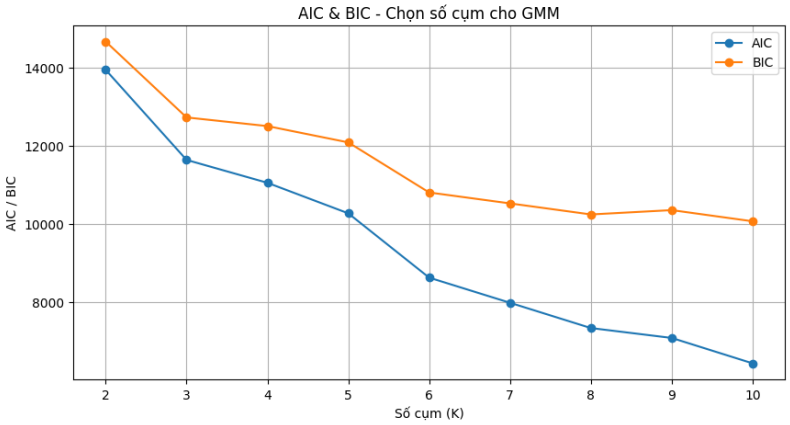
4.3.2. Model



After removing outliers using DBSCAN, the KMeans algorithm was applied to the remaining data, resulting in four clusters with sizes of 345, 195, 180, and 71 observations, respectively. The clustering achieved a **Silhouette Score of 0.173**, indicating a moderate level of cluster separation. Compared to DBSCAN alone, this combined approach shows improved structure and more balanced clusters, with no significant dominance by noise or trivial groups. While the result is not optimal, it suggests that **DBSCAN + KMeans is a more effective strategy** for this dataset, and further refinement of parameters could lead to better clustering performance.

4.4. GMM model

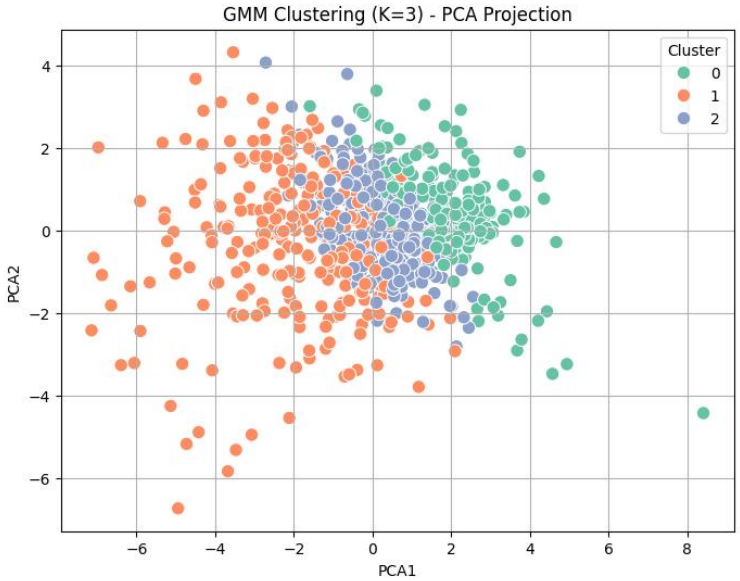
4.4.1. Selecting the Number of Clusters for GMM



Model Selection for GMM Based on AIC & BIC. The AIC and BIC scores were used to determine the optimal number of clusters for the Gaussian Mixture Model (GMM). Both curves show a **consistent decrease** as the number of clusters increases, with a notable drop from **K=2 to K=6**.

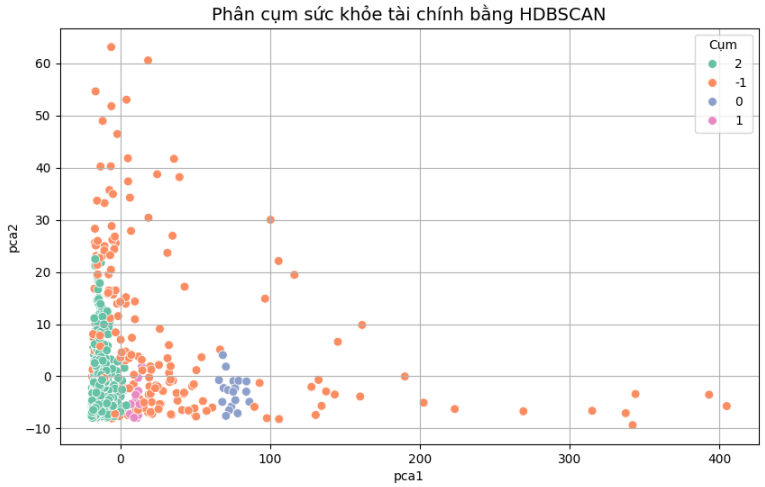
- **AIC** continues to decline steadily, reaching its lowest value at **K=3**, suggesting better model fit with more components.
- **BIC**, however, reaches a plateau around **K=6–8** and fluctuates slightly afterward, indicating that additional clusters beyond this range may offer diminishing returns when considering model complexity.

4.4.2. Model



Gaussian Mixture Model (GMM) with **three clusters (K = 3)** yielded a relatively balanced cluster distribution: 296, 261, and 235 observations per cluster, respectively. However, the **Silhouette Score was 0.096**, indicating **weak clustering performance**. Although the clusters are not dominated by outliers or imbalance, the low score suggests that the separation between clusters is limited, and the model may struggle to capture meaningful group structures in the data. This implies that **GMM with K = 3 may not be sufficient**, and further tuning of the number of components or input features might be necessary to improve clustering quality.

4.5. HDBSCAN model



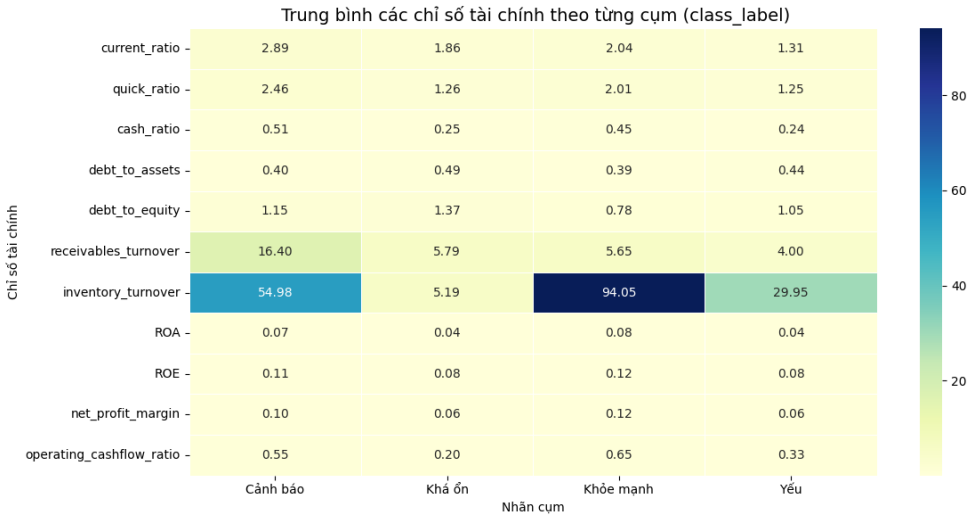
HDBSCAN identified four clusters, including a large cluster (Cluster 2 with 575 samples), two very small clusters (Clusters 0 and 1 with 18 and 17 samples, respectively), and 182 observations labeled as outliers. After excluding the outliers, the model achieved a **Silhouette Score of 0.6431**, indicating **high-quality clustering with well-separated group structures**. The result suggests that HDBSCAN is highly effective at identifying dense, well-defined clusters while isolating less coherent or noisy data points. However, the presence of only one dominant cluster alongside two minor ones reflects an **imbalanced structure**, which may limit interpretability depending on the analytical objective.

Calinski-Harabasz Score = 105.10 indicates that the clusters are compact and well-separated.. **Davies-Bouldin Score = 1.65** suggests a low level of similarity between clusters, which is a sign of good clustering performance.

4.6. Overall Conclusion

HDBSCAN not only demonstrates superior clustering performance from a technical standpoint, but also produces clusters with clear practical significance in assessing corporate financial health. It is the most suitable model for the characteristics of the dataset used in this study.

Interpretation of HDBSCAN Clusters Based on Financial Ratios



The HDBSCAN algorithm identified three main clusters (excluding outliers), each representing a distinct financial profile:

Cluster 0 – Strong financial health:

Companies in Cluster 0 demonstrate strong financial positions. They have the highest profitability ratios, with ROA of 0.08, ROE of 0.12, and net profit margin of 0.12. Liquidity is also solid (current ratio = 2.04; quick ratio = 2.01), while debt levels are relatively low (debt-to-equity = 0.78). Notably, these firms show outstanding operational performance with the highest inventory turnover (94.05) and a healthy operating cash flow ratio (0.65). This cluster likely represents well-managed firms with efficient operations and sustainable financial structures.

Cluster 1 – Weak financial health:

Firms in Cluster 1 display signs of financial weakness. Liquidity is the lowest among all clusters (current ratio = 1.31; quick ratio = 1.25), and leverage is relatively high (debt-to-assets = 0.44; debt-to-equity = 1.05). Profitability metrics are also lower (ROA = 0.04; ROE = 0.08; profit margin = 0.06), and the operating cash flow ratio is modest (0.33). These companies may face financial stress or limited flexibility in meeting short-term obligations.

Cluster 2 – Average performance but high debt:

This is the largest cluster (575 firms), characterized by average financial performance but elevated risk. Liquidity indicators are moderate (current ratio = 1.86), yet the debt-to-equity ratio is the highest across all clusters (1.37), suggesting higher

reliance on external financing. Inventory turnover is low (5.19), and the operating cash flow ratio is the weakest (0.20), indicating potential inefficiencies in operations and weaker internal cash generation.

Outliers (Label = -1):

The outlier group includes firms with unusual financial behavior. They have extremely high efficiency metrics, such as receivables turnover (16.40) and inventory turnover (54.98), along with above-average liquidity (current ratio = 2.89; quick ratio = 2.46). While not forming a coherent cluster, these companies may represent either exceptional performers or entities with atypical financial structures.

5. Classification Model Development

5.1. Model Objective

Following the unsupervised clustering process using HDBSCAN and the labeling of each cluster based on financial health levels (“Healthy”, “Moderate”, “Weak”, “Warning”), the next step involves building a supervised machine learning model to:

- Automatically predict the financial health label of new companies based on the calculated financial indicators.
- Evaluate the separability and practical interpretability of the clustered groups.
- Compare the performance of common machine learning models to determine the most suitable algorithm.

5.2. Model Training Methodology

5.2.1. Preparing Data

The dataset used includes normalized financial ratios derived from the earlier preprocessing phase.

The class_label column, which contains four categories ("Warning", "Weak", "Moderate", "Healthy"), was encoded into numerical format using LabelEncoder to be compatible with machine learning models. The data was then split into training and testing sets in an 80:20 ratio using train_test_split, with stratify applied to ensure balanced class distribution between training and test data.

5.2.2. Random Forest model

```
[Random Forest] Classification Report:
              precision    recall  f1-score   support

           0       0.83      0.81      0.82        37
           1       0.95      0.99      0.97       115
           2       1.00      0.25      0.40         4
           3       0.50      0.33      0.40         3

    accuracy                           0.92       159
   macro avg       0.82      0.60      0.65       159
weighted avg       0.92      0.92      0.91       159

Accuracy RF: 0.9182389937106918
```

Random Forest is an ensemble learning method that combines multiple decision trees and is well known for its robustness to noise and overfitting. When applied to the classification of corporate financial health, the model achieved a relatively high overall accuracy of **91.8%**. However, a closer inspection of the classification report and confusion matrix reveals inconsistencies in its performance across different classes.

The model performs very well on the majority class “Healthy” (Precision = 0.95, Recall = 0.99, F1-score = 0.97), which contains a large number of samples. In contrast, the minority classes such as “Moderate” and “Weak” were poorly predicted, with the “Moderate” group achieving only **25% recall**, and the “Weak” group just **33% recall**. This indicates that Random Forest has a tendency to **favor dominant classes**, making it less effective in recognizing underrepresented yet critical financial conditions. Given that the goal of the model is to detect financially unstable firms as well, this limitation significantly reduces its practical value in risk assessment and early warning systems.

5.2.3. XGBoost model

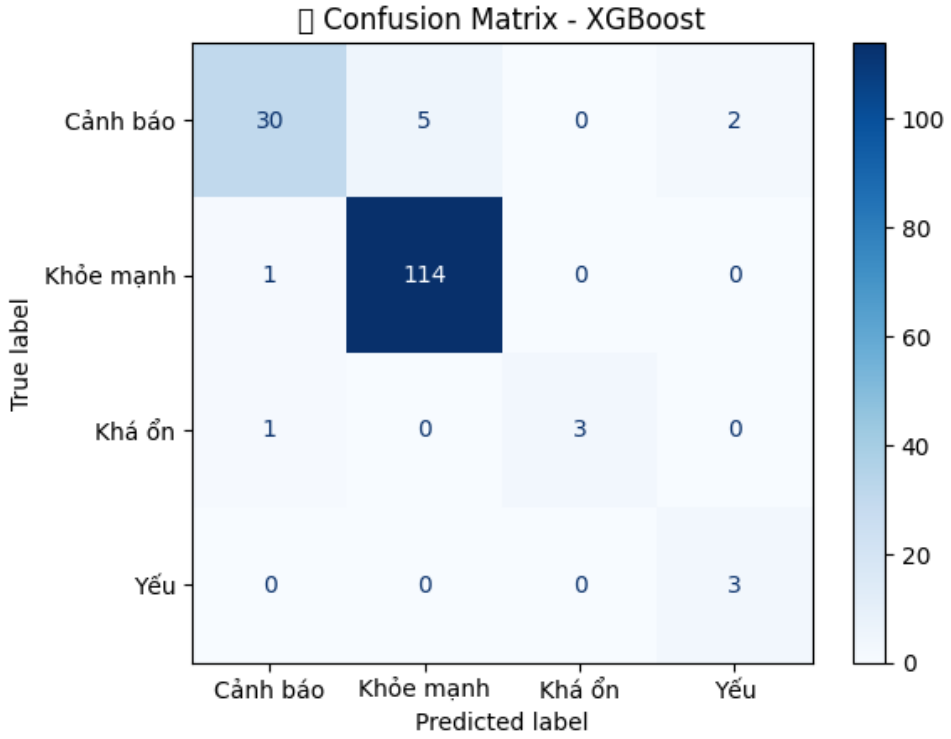
```
[XGBoost] Classification Report:
              precision    recall  f1-score   support

           0       0.94      0.81      0.87        37
           1       0.96      0.99      0.97       115
           2       1.00      0.75      0.86         4
           3       0.60      1.00      0.75         3

    accuracy                           0.94       159
   macro avg       0.87      0.89      0.86       159
weighted avg       0.95      0.94      0.94       159

Accuracy XGB: 0.9433962264150944
```

X GBoost, a state-of-the-art gradient boosting algorithm, is designed to correct the limitations of base learners through iterative training and regularization. Applied to the same dataset, XGBoost significantly outperformed Random Forest, achieving a higher overall accuracy of **94.3%**, along with a **macro F1-score of 0.86** and a **weighted F1-score of 0.94**. Most notably, XGBoost handled the minority classes much better. The "Moderate" class achieved a strong F1-score of **0.86** (compared to 0.40 in Random Forest), and the "Weak" class achieved **0.75**, reflecting the model's capability to capture more nuanced patterns in the data.



The confusion matrix confirms that the model correctly classified almost all "Healthy" and "Warning" cases, with very few misclassifications. The slight overlap between "Warning" and "Healthy" classes is understandable, as companies may exhibit borderline financial indicators. Overall, XGBoost demonstrates **better generalization**, **less class imbalance bias**, and **stronger sensitivity to subtle financial signals**, making it a more suitable model for predicting firm financial health.

5.3. Overall conclusion

Between the two evaluated models, **XGBoost clearly outperformed Random Forest** in both overall accuracy and the ability to correctly classify minority classes such

as “Moderate” and “Weak.” While Random Forest demonstrated solid performance for dominant classes, it struggled with imbalanced data and failed to capture nuances in underrepresented groups. In contrast, XGBoost delivered more balanced classification results, making it a more robust and reliable choice for predicting corporate financial health in this study.

6. Application of the Classification Model to Historical Data (2021–2023)

6.1. Objective of Application

Following the training of the classification model using 2024 data, the next step is to apply the model to historical financial data from 2021 to 2023. This aims to:

- Automatically assign a financial health label to each company based on its financial indicators.
- Examine the distribution of predicted classes over time and across sectors.
- Evaluate the model’s generalizability and the practical consistency of the HDBSCAN-based labeling logic.

The historical dataset was processed using the same cleaning, outlier removal, and standardization procedures as applied to the 2024 data. The financial indicators were then fed into the trained XGBoost model, and predicted labels were assigned accordingly.

6.2. Overall Prediction Results

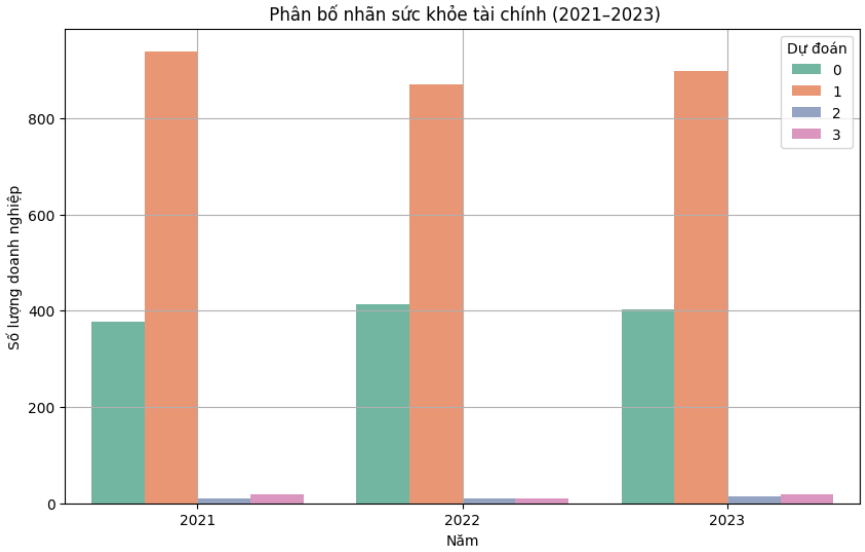
Phân bố theo từng năm:

Năm	predicted_class	
2021	0	378
	1	939
	2	11
	3	19
2022	0	414
	1	871
	2	11
	3	11
2023	0	402
	1	897
	2	15
	3	18

The model produced the following distribution of predicted financial health labels across the 3,986 companies in the 2021–2023 dataset:

The results indicate that approximately **68%** of firms were classified as "Healthy," while only **2.1%** fell into the combined "Weak" and "Warning" categories. This suggests a generally stable financial environment during the evaluated period.

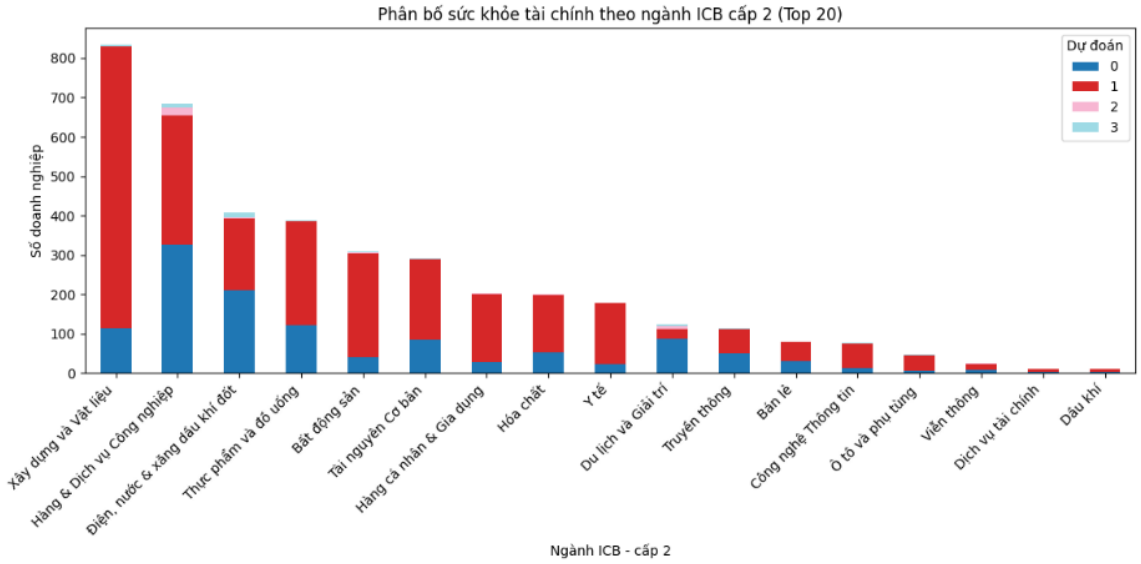
6.3.Temporal analysis



The percentage of firms classified as “Healthy” remained consistent across the three years (approximately 65–70%), with no sharp increase in high-risk categories. This stability suggests that the financial conditions during this period were resilient and that the classification model generalizes well to unseen data from previous years.

6.4. Sector-Based Analysis (ICB – Level 2)

predicted_class Ngành ICB - cấp 2	0	1	2	3	Tổng
Xây dựng và Vật liệu	113	716	0	6	835
Hàng & Dịch vụ Công nghiệp	326	329	19	11	685
Điện, nước & xăng dầu khí đốt	210	183	3	12	408
Thực phẩm và đồ uống	121	264	1	3	389
Bất động sản	40	265	2	3	310
Tài nguyên Cơ bản	85	203	1	2	291
Hàng cá nhân & Gia dụng	28	173	1	0	202
Hóa chất	52	147	1	0	200
Y tế	22	155	0	0	177
Du lịch và Giải trí	88	24	8	5	125



Several insights emerge from this breakdown:

- Defensive sectors such as **Health Care**, **Chemicals**, and **Personal Goods** show strong performance with a high concentration of "Healthy" firms.
- In contrast, **Travel & Leisure** and **Industrial Services** exhibit higher proportions of "Weak" and "Warning" classifications, likely due to operational or external financial pressures.
- **Construction & Materials** contains a large number of firms, mostly "Healthy," but still includes some warning cases, suggesting a degree of heterogeneity within the sector.

6.5. Conclusion of the application

The application of the XGBoost model to historical data from 2021 to 2023 demonstrates strong generalization capability. It effectively reclassifies the financial health status of nearly 4,000 firms in an automated and consistent manner. The distribution of predicted classes shows no signs of abnormal bias and aligns well with the macroeconomic context of the 2021–2023 period. This confirms the model's practical utility and its potential to serve as a reliable tool for sector-level financial monitoring and investment decision support over time.