SHM PUBLISHER
SURYA HIJAU MANFAAT

Tạp chí hệ thống thông tin Thăm dò và nghiên cứu

https://shmpublisher.com/index.php/joiser

p-ISSN 2964-1160 | e-ISSN 2963-6361

Dự đoán phá sản bằng cách sử dụng máy vectơ hỗ trợ thuật toán di truyền (GA-SVM) Lựa chọn và xếp chồng tính năng

Wiena Faqih Abror[1]*, Alamsyah[2], Muhammad Aziz[3]
[1,2]Khoa Khoa học Máy tính, Đại học Negeri Semarang, Indonesia
[3]Khoa Kỹ thuật Điện, Đại học Kookmin, Hàn Quốc

DOI: https://doi.org/10.52465/joiser.v1i2.180

Nhận ngày 08 tháng 7 năm 2023; Được chấp nhận ngày 18 tháng 7 năm 2023; Có sẵn trực tuyến vào ngày 19 tháng 7 năm 2023

Thông tin bài viết

Từ khóa:
Phá sản;
Thuật toán di truyền;
Vectơ hỗ trợ
máy móc;
Xếp chồng

trừu tượng

Phá sản là một tác động gây ra bởi sự thất bại tài chính của một công ty. Phải tránh thất bại tài chính trong công ty để không gây thiệt hại cho công ty. Trong nghiên cứu được thực hiện bằng cách sử dụng tập dữ liệu từ Tạp chí Kinh tế Đài Loan, có tới 6.819 người được đào tạo bằng thuật toán học máy bằng kỹ thuật phân loại. Mục tiêu thu được từ nghiên cứu được thực hiện là đạt được kỹ thuật phân loại có kết quả chính xác tốt nhất. Phương pháp được sử dụng trong nghiên cứu này là tiền xử lý bằng kỹ thuật lấy mẫu thiểu số tổng hợp để xử lý các tập dữ liệu không cân bằng. Sau đó, kết quả của tập dữ liệu cân bằng sẽ được xử lý bằng thuật toán lựa chọn đặc trưng máy vectơ hỗ trợ thuật toán di truyền để rút gọn các thuộc tính của tập dữ liệu. Các tập dữ liệu có thuộc tính bị giảm sẽ được huấn luyện bằng phương pháp xếp chồng với một trình học cơ sở phân loại duy nhất dưới dạng k-láng giềng gần nhất, vịnh ngây thơ, cây quyết định với mô hình cây phân loại và hồi quy, cây quyết định tăng cường độ dốc và tăng cường độ dốc ánh sáng . Trình học meta được sử dụng trong phương pháp xếp chồng là tăng cường độ dốc cực cao. Kết quả về độ chính xác thu được từ nghiên cứu được thực hiện là 99,22%.

1. Giới thiệu

Phá sản tài chính hoặc thất bại tài chính có thể gây tổn hại đến lưu thông kinh tế toàn cầu. Các công ty, tiểu thương, chợ truyền thống và nhà nước có thể cảm nhận được tác động tiêu cực. Các nhà thực hành, nhà đầu tư, chính phủ và các nhà nghiên cứu học thuật cố gắng xác định các biến số gây ra phá sản [1], [2]. Nguyên nhân phá sản có thể thuộc nhiều loại khác nhau, chẳng hạn như nguyên liệu thô tăng, phí nhân viên, cạnh tranh kinh doanh và năng lực quản lý kém. Nguyên nhân phá sản có thể xảy ra ở mỗi doanh nhân. Vì vậy, ở mọi cấp độ, các tác nhân kinh doanh đều có thể có các yếu tố bên trong và bên ngoài [3]. Quản lý tiết kiệm và giảm rủi ro tín dụng kinh tế có thể cải thiện việc đánh giá rủi ro tín dụng [4]. Đề xuất đánh giá phá sản

* ***Đồng tác giả:***

Wiena Faqih Abror,
Khoa Khoa học Máy tính, Đại học Negeri Semarang, Gunungpati, Sekaran, Semarang, Indonesia. Email: wf.abror@gmail.com

thông tin có giá trị cho các nhà đầu tư, ban quản lý, cổ đông và chính phủ để đưa ra quyết định bảo vệ tài chính của họ để phá sản không xảy ra. Nghiên cứu phá sản có thể đưa ra cảnh báo sớm và phát hiện những điểm yếu về tài chính. Dự đoán phân tích phá sản có thể mang lại những lợi ích như giảm chi phí phân tích tín dụng, giám sát tài chính và tỷ lệ thu hồi nợ [5].

Phân tích phá sản cũng tương tự như mô hình phân loại, được lấy từ số liệu thống kê do công ty cung cấp để vạch ra đặc điểm, chỉ số về nguyên nhân phá sản. Các vấn đề phân loại có thể được giải quyết bằng các thuật toán phân loại [6], chẳng hạn như Phân tích phân biệt đa biến (MDA), Hồi quy logistic (LR), Mạng thần kinh (NN), Máy vectơ hỗ trợ (SVM) và Phương pháp tập hợp [7]. Việc sử dụng các thuật toán khác sử dụng các đặc điểm của Mạng thần kinh, chẳng hạn như Mạng thần kinh tái phát (RNN) [8] và Mạng thần kinh chuyển đổi (CNN) [9] đã được phát triển để phân tích các chủ đề tài chính và quản lý [7].

Có thể giảm kích thước chỉ báo cao bằng cách sử dụng lựa chọn tính năng để cải thiện hiệu suất dự đoán [10]. Kohavi và John [11] tuyên bố rằng kích thước cao có thể làm giảm hiệu suất của độ chính xác dự đoán thu được, chẳng hạn như trong Cây quyết định (DT) và Naive-Bayes (NB), vì các thuộc tính trong dữ liệu không liên quan đến thuật toán. Lựa chọn tính năng có thể làm giảm kích thước cao thu được từ sự kết hợp của các chỉ báo FR và CGI [1].

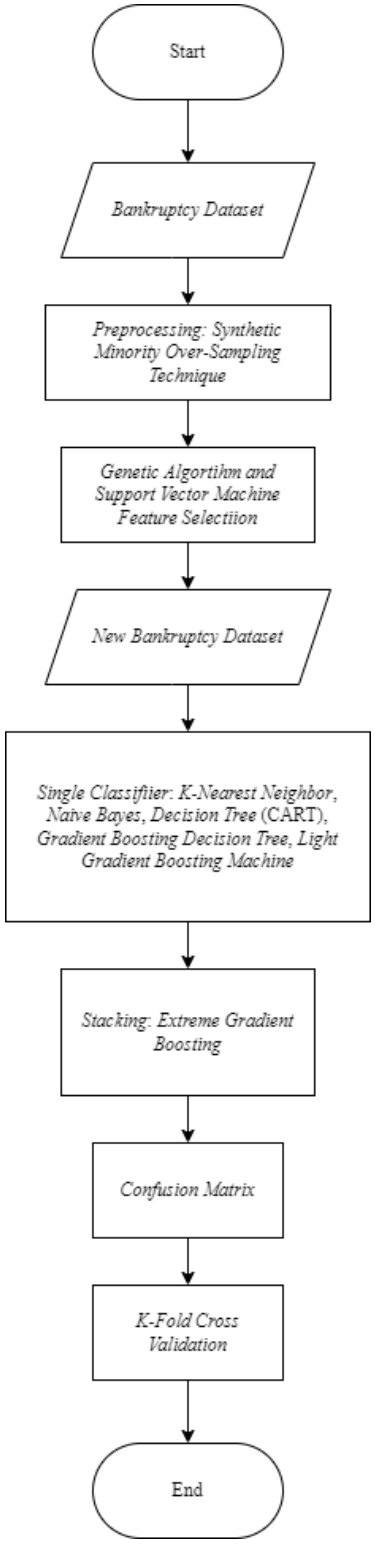
Lựa chọn tính năng được sử dụng tùy theo loại đối tượng dữ liệu được sở hữu. Dữ liệu có nhãn được gọi là dữ liệu được giám sát. Phương pháp bao bọc là một biến thể của phương pháp lựa chọn đối tượng được giám sát có thể được sử dụng để giảm kích thước chiều cao [12].

Thuật toán di truyền (GA) có các đặc điểm chọn lọc tự nhiên bằng cách áp dụng các quy tắc chọn lọc, lai ghép và đột biến [13]. Các quy tắc đặc tính GA có thể được sử dụng để giải quyết các vấn đề tối ưu hóa phi tuyến và hình thành các phân biệt đối xử phân loại. GA có thể được sử dụng làm trình bao bọc trong bước lựa chọn tính năng. Phương thức bao bọc sẽ lặp lại trên mô hình được sử dụng, chẳng hạn như GA, bị giới hạn số lần lặp dựa trên sự hình thành thế hệ mong muốn [12]. Phương pháp trình bao bọc yêu cầu xác thực dữ liệu bằng mô hình đối tượng để đánh giá độ chính xác của các tính năng được chọn thông qua GA [14]. Đối tượng mô hình được sử dụng là Máy vectơ hỗ trợ (SVM). SVM có thể giải quyết các vấn đề phức tạp như dữ liệu có kích thước lớn và phi tuyến [15].

Các kết quả tập hợp con đối tượng và dữ liệu đã đạt đến thế hệ được chỉ định (tiêu chí dừng) sẽ được huấn luyện bằng phương pháp tập hợp. Phương pháp tập hợp sẽ huấn luyện dữ liệu song song và tạo ra các giá trị chính xác [16]. Một trong những thuật toán được áp dụng cho phương pháp tập hợp là xếp chồng. Thuật toán xếp chồng có thể huấn luyện dữ liệu dựa trên thuật toán học máy được sử dụng. Theo tuần tự, thuật toán xếp chồng sẽ chia quá trình đào tạo thành hai phần, đó là phần học cơ bản và phần học meta [17]. Dữ liệu sẽ được đào tạo bằng thuật toán xác định trước ở giai đoạn học cơ sở. Nghiên cứu này sử dụng bốn thuật toán học máy làm người học cơ sở, cụ thể là cây quyết định với tiêu chí chỉ số Gini [18], cây quyết định tăng cường độ dốc [19], k hàng xóm gần nhất [20] và máy tăng cường độ dốc nhẹ [21]. Ở giai đoạn meta-learner, Extreme gradient Boosting (XGBOOST) được sử dụng để huấn luyện dữ liệu kết quả từ trình học cơ sở [22], [23]. Zelenkov và cộng sự, [24] đã sử dụng bộ phân loại tổng hợp mặc dù không sử dụng phương pháp lọc. Kim và Kang [25] tuyên bố rằng việc học tập hợp có thể cải thiện hiệu suất của một bộ phân loại đơn lẻ vì việc học tập hợp có thể cải thiện độ chính xác của bộ phân loại.

2. Phương pháp

Phương pháp nghiên cứu được thực hiện thông qua thiết kế nghiên cứu được thực hiện. Thiết kế nghiên cứu được thực hiện để giải thích dòng nghiên cứu một cách toàn diện. Nhìn chung, quy trình làm việc được chia thành ba giai đoạn, đó là giai đoạn tiền xử lý, giai đoạn lựa chọn tính năng và giai đoạn huấn luyện dữ liệu. Nghiên cứu kết thúc bằng thử nghiệm xác nhận bằng cách sử dụng ma trận nhầm lẫn và xác thực chéo k-fold. Quy trình nghiên cứu được thể hiện trong Hình 1.



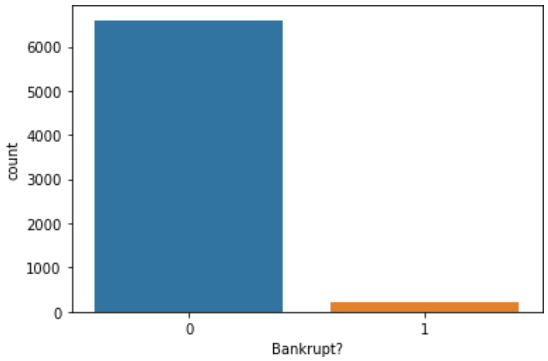
Hình 1. Phương pháp đề xuất

Giải thích chung về quy trình nghiên cứu là giai đoạn tiền xử lý sử dụng phương pháp kỹ thuật lấy mẫu thiểu số tổng hợp, nhằm mục đích xử lý các tập dữ liệu không cân bằng. Giai đoạn lựa chọn tính năng sử dụng phương pháp bao bọc với thuật toán di truyền làm lựa chọn tính năng và sử dụng máy vectơ hỗ trợ để kiểm tra các tính năng đã được tìm kiếm bằng thuật toán di truyền, cũng như giai đoạn huấn luyện dữ liệu được chia thành hai phần bằng cách sử dụng. Đầu tiên là một bộ phân loại duy nhất, chẳng hạn như cây quyết định, vịnh ngây thơ, k hàng xóm gần nhất, cây quyết định tăng cường độ dốc và máy tăng cường độ dốc ánh sáng; cách thứ hai sử dụng tính năng xếp chồng để kết hợp một trình phân loại duy nhất làm trình học cơ bản và trình độ nâng cao

tăng cường độ dốc như một người học meta. Mỗi phần ở giai đoạn đào tạo dữ liệu sẽ được xác thực bằng ma trận nhầm lẫn.

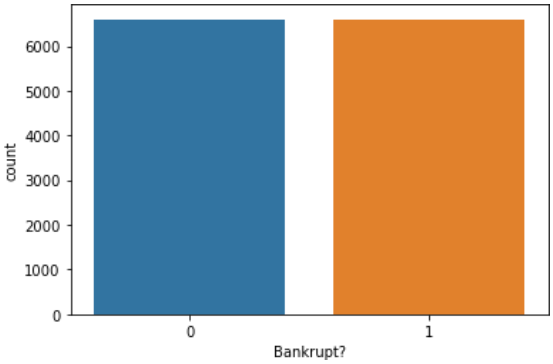
3. Kết quả và thảo luận

Vấn đề được tìm thấy trong nghiên cứu này là sự mất cân bằng dữ liệu. Sự cân bằng của dữ liệu trong tập dữ liệu Mục tiêu Phá sản Đài Loan được hiển thị trong Hình 2 và lớp 0 hiển thị phân bổ của 6.599 tập dữ liệu. Lớp 1 hiển thị trải rộng 220 bộ dữ liệu. Sự khác biệt thể hiện trên biểu đồ cho thấy dữ liệu mất cân bằng trong bộ dữ liệu Phá sản của Đài Loan. Có thể thấy trực quan dưới dạng biểu đồ của tập dữ liệu mục tiêu về Vụ phá sản của Đài Loan trong Hình 2.



Hình 2. Tập dữ liệu không cân bằng

Trực quan hóa tập dữ liệu đích ở dạng biểu đồ được hiển thị trong Hình 3. cho thấy rằng việc phân phối tập dữ liệu đã được cân bằng bằng phương pháp kỹ thuật lấy mẫu quá mức tổng hợp thiểu số.



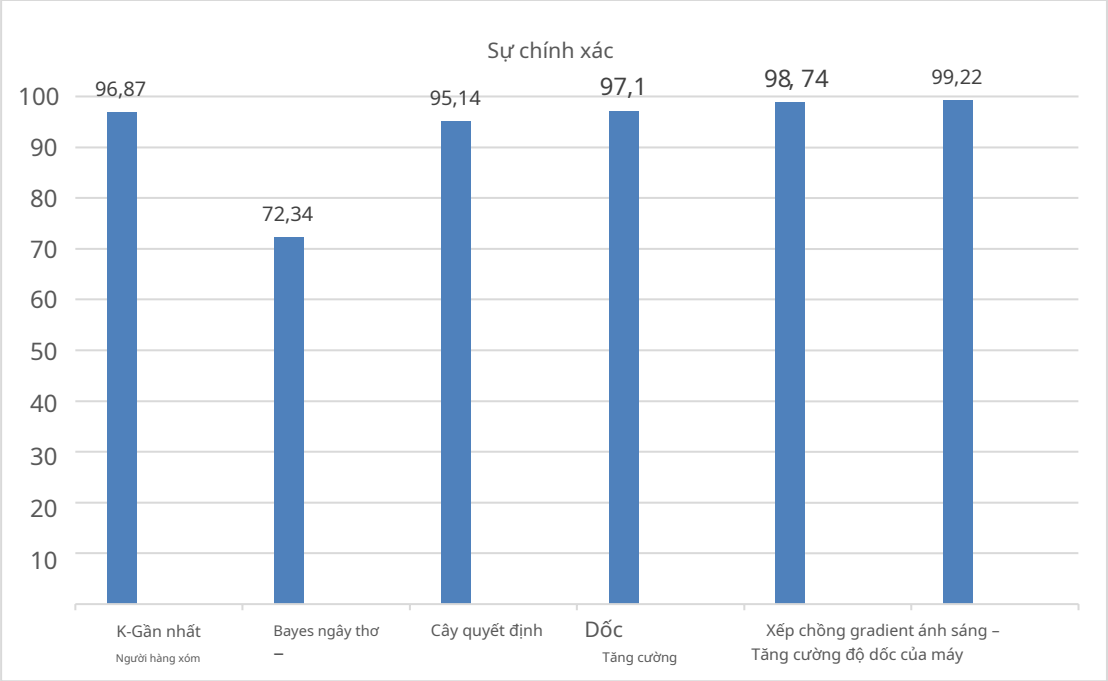
Hình 3. Tập dữ liệu cân bằng

Kết quả cân bằng tập dữ liệu theo thuật ngữ danh nghĩa được thể hiện trong Hình 3. Lớp 0 thể hiện sự phân bố của 6.599 tập dữ liệu và lớp 1 thể hiện sự phân bố của 6.599 tập dữ liệu, nghĩa là sự phân bố của các tập dữ liệu được cân bằng.

Bộ dữ liệu Phá sản của Đài Loan được xử lý bằng cách sử dụng lựa chọn tính năng của máy vectơ hỗ trợ thuật toán di truyền nhằm mục đích giảm các thuộc tính. Thuộc tính mặc định là 96 cùng với cột mục tiêu. Các thuộc tính thu được sau khi được xử lý bằng cách sử dụng lựa chọn tính năng của máy vectơ hỗ trợ thuật toán di truyền là 44 cùng với cột mục tiêu tuyên bố phá sản. Tập dữ liệu mới đã được đào tạo đã được kiểm tra bằng xác thực chéo. Kết quả về độ chính xác được trình bày trong Bảng 1. So sánh trực quan được thể hiện trong Hình 4.

Bảng 1. So sánh hiệu suất của mô hình

Thuật toán	Sự chính xác (%)
K-Hàng xóm gần nhất	96,87
Bayes ngây thơ	72,34
Cây quyết định – Tăng cường độ	95,14
dốc CART Cây quyết định Tăng cường độ	97,13
dốc ánh sáng Xếp chồng máy – Tăng	98,74
cường độ dốc cực cao	99,22



Hình 4. So sánh hiệu suất mô hình

4. Kết luận

Nghiên cứu được thực hiện bằng cách kết hợp một bộ phân loại duy nhất, cụ thể là k-hàng xóm gần nhất, vịnh ngây thơ, cây quyết định sử dụng phương pháp cây phân loại và hồi quy, cây quyết định tăng cường độ dốc và máy tăng cường độ dốc ánh sáng bằng phương pháp xếp chồng. Trình phân loại duy nhất được sử dụng làm trình học cơ sở trong phương pháp xếp chồng. Trình học meta được sử dụng là tăng cường độ dốc cực cao. Kết quả độ chính xác của nghiên cứu được thực hiện là 99,22%.

Người giới thiệu

[1] D. Liang, CC Lu, CF Tsai và GA Shih, "Tỷ lệ tài chính và các chỉ số quản trị doanh nghiệp trong dự đoán phá sản: Một nghiên cứu toàn diện,"*Euro. J. Điều hành. Res.*, tập. 252, không. 2, trang 561–572, 2016, doi: 10.1016/j.ejor.2016.01.012.

[2] MA Muslim, Y. Dasril, H. Javed, WF Abror, DAA Pertiwi và T. Mustaqim, "Thuật toán xếp chồng tập hợp để cải thiện độ chính xác của mô hình trong dự đoán phá sản,"*J. Khoa học dữ liệu. Trí tuệ. Hệ thống.*, tập. 1, không. 1, 2023.

[3] A. Kadim và N. Sunardi, "Phân tích Altman Z-Score Untuk Memprediksi Kebangkrutan Pada Bank Pemerintah (Bumn) Di Indonesia Tahun 2012-2016 Thông tin bài viết Tóm tắt Prodi Manajemen Unpam,"*Keuang. và Investasi)*, tập. 1, không. 3, trang 142–156, 2018.

[4] D. West, S. Dellana và J. Qian, "Chiến lược tổng hợp mạng lưới thần kinh cho các ứng dụng quyết định tài chính,"*Máy tính. Hoạt động. Res.*, tập. 32, không. 10, trang 2543–2559, 2005, doi: 10.1016/j.cor.2004.03.017.

[5] S. Lee và WS Choi, "Mô hình dự đoán phá sản đa ngành sử dụng mạng thần kinh lan truyền ngược và phân tích phân biệt đa biến,"*Hệ thống chuyên gia ứng dụng.*, tập. 40, không. 8, trang 2941–2946, 2013, doi: https://doi.org/10.1016/j.eswa.2012.12.009.

[6] P. Pampouktsi, S. Avdimiotis, M. Maragoudakis, M. Avlonitis, P. Hoogar và G. Ruhago, "Các kỹ thuật học máy ứng dụng được sử dụng cho mục đích lựa chọn và bố trí nguồn nhân lực trong khu vực công,"*J. Inf. Hệ thống. Khám phá.*, tập. 01, không. 01, trang 1–16, 2023.

[7] Y. Qu, P. Quan, M. Lei và Y. Shi, "Đánh giá dự đoán phá sản bằng cách sử dụng kỹ thuật học máy và học sâu,"*Máy tính thủ tục. Khoa học.*, tập. 162, trang 895–899, 2019, doi: https://doi.org/10.1016/j.procs.2019.12.065.

[số 8] E. Chong, C. Han và FC Park, "Mạng lưới học sâu để phân tích thị trường chứng khoán và

dự đoán: Phương pháp luận, trình bày dữ liệu và nghiên cứu trường hợp,"*Hệ thống chuyên gia ứng dụng.*, tập. 83, trang 187–205, 2017, doi: https://doi.org/10.1016/j.eswa.2017.04.030.

[9] T. Hosaka, "Dự đoán phá sản bằng cách sử dụng các tỷ lệ tài chính hình ảnh và mạng lưới thần kinh tích chập,"*Hệ thống chuyên gia ứng dụng.*, tập. 117, trang 287–299, 2019, doi: https://doi.org/10.1016/j.eswa.2018.09.039.

[10] I. Guyon, J. WESTON, S. Barnhill và V. Vadnik, "Lựa chọn gen để phân loại ung thư bằng máy vectơ hỗ trợ,"*Mach. Học hỏi.*, tập. 46, trang 389–422, 2002.

[11] R. Kohavi và GH John, "Trình bao bọc để lựa chọn tập hợp con đối tượng địa lý,"*Nghệ thuật. Trí tuệ.*, tập. 97, không. 1–2, trang 273–324, 1997.

[12] N. El Aboudi và L. Benhlima, "Đánh giá về các phương pháp lựa chọn tính năng của trình bao bọc," trong*2016 Hội nghị quốc tế về kỹ thuật MIS (ICEMIS)*, 2016, trang 1–5. doi: 10.1109/ICEMIS.2016.7745366.

[13] AR Safitri và MA Muslim, "Cải thiện độ chính xác của Bộ phân loại Naive Bayes để xác định việc khách hàng sử dụng thuật toán di truyền và SMOTE,"*J. Máy tính mềm. Khám phá.*, tập. 1, không. 1, trang 70–75, 2020.

[14] WC Lin, YH Lu và CF Tsai, "Lựa chọn tính năng trong các mô hình dự đoán phá sản dựa trên học tập đơn lẻ và tập hợp,"*Hệ thống chuyên gia*, tập. 36, không. 1, trang 1–8, 2019, doi: 10.1111/exsy.12335.

[15] Z. Tao, L. Huiling, W. Wenwen và Y. Xia, "Lựa chọn tính năng dựa trên GA-SVM và tối ưu hóa tham số trong mô hình chi phí nhập viện,"*ứng dụng. Máy tính mềm.*, tập. 75, trang 323–332, 2019, doi: https://doi.org/10.1016/j.asoc.2018.11.001.

[16] X. Dong, Z. Yu, W. Cao, Y. Shi và Q. Ma, "Khảo sát về học tập tổng thể,"*Đằng trước. Máy tính. Khoa học.*, tập. 14, không. 2, trang 241–258, 2020, doi: 10.1007/s11704-019-8208-z.

[17] J. Đậu*et al.*, "Cải thiện khả năng đánh giá lở đất bằng cách sử dụng máy vectơ hỗ trợ với khung máy học tổng hợp đóng bao, tăng cường và xếp chồng ở lưu vực núi, Nhật Bản,"*Sạt lở đất*, tập. 17, không. 3, trang 641–658, 2020, doi: 10.1007/s10346-019-01286-5.

[18] HH Patel và P. Prajapati, "Nghiên cứu và phân tích các thuật toán phân loại dựa trên cây quyết định,"*Int. J. Máy tính. Khoa học. Anh.*, tập. 6, không. 10, trang 74–78, 2018, doi: 10.26438/ijcse/v6i10.7478.

[19] JS Yang, CY Zhao, HT Yu và HY Chen, "Sử dụng GBDT để dự đoán thị trường chứng khoán,"*Máy tính thủ tục. Khoa học.*, tập. 174, không. 2019, trang 161–171, 2020, doi: 10.1016/j.procs.2020.06.071.

[20] W. Xing và Y. Bei, "Phân loại dữ liệu lớn về sức khỏe y tế dựa trên thuật toán phân loại KNN," *Truy cập IEEE*, tập. 8, trang 28808–28819, 2020, doi: 10.1109/ACCESS.2019.2955754.

[21] G. Kế*et al.*, "LightGBM: Cây quyết định tăng cường độ dốc hiệu quả cao,"*Khuyến cáo. Thần kinh Inf. Quá trình. Hệ thống.*, KHÔNG. Tháng 12, trang 3147–3155, 2017.

[22] MA Muslim và Y. Dasril, "Khung dự đoán phá sản của công ty dựa trên hầu hết các các tính năng có ảnh hưởng bằng cách sử dụng XGBoost và xếp chồng việc học tập theo nhóm,"*Int. J. Điện. Máy tính. Anh.*, tập. 11, không. 6, trang 5549–5557, 2021, doi: 10.11591/ijece.v11i6.pp5549-5557.

[23] MA Hồi giáo*et al.*, "Công cụ siêu học kết hợp mô hình mới để cải thiện khả năng dự đoán chính xác P2P cho vay với việc học tập xếp chồng *,"*Trí tuệ. Hệ thống. với ứng dụng.*, tập. 18, không. Tháng 12 năm 2022, tr. 200204, 2023, doi: 10.1016/j.iswa.2023.200204.

[24] Y. Zelenkov, E. Fedorova và D. Chekrizov, "Phương pháp phân loại hai bước dựa trên thuật toán di truyền để dự báo phá sản,"*Hệ thống chuyên gia ứng dụng.*, tập. 88, trang 393–401, 2017, doi: https://doi.org/10.1016/j.eswa.2017.07.025.

[25] MJ Kim và DK Kang, "Lựa chọn bộ phân loại trong quần thể bằng thuật toán di truyền để dự đoán phá sản,"*Hệ thống chuyên gia ứng dụng.*, tập. 39, không. 10, trang 9308–9314, 2012, doi: 10.1016/j.eswa.2012.02.072.