**PHƯƠNG PHÁP HỌC MÁY MACHINE LEARNING TRONG CHẤM ĐIỂM THẺ TÍN DỤNG KHÁCH HÀNG**

***Tóm tắt:*** *Chấm điểm tín dụng, quy trình quản lý rủi ro quan trọng của các tổ chức tài chính và ngân hàng, không chỉ là bảng xếp hạng mà còn là chìa khóa quyết định chất lượng cho vay. Trong sự cạnh tranh khốc liệt, chất lượng của quy trình chấm điểm tín dụng quyết định đến sự tồn tại và lợi nhuận của các tổ chức tài chính.. Chấm điểm tín dụng tốt sẽ góp phần làm cho chất lượng cho vay tốt hơn. Chất lượng cho vay là yếu tố quyết định hàng đầu đến sự cạnh tranh, tồn tại và lợi nhuận của các tổ chức tài chính ngân hàng. Các kỹ thuật học máy đã góp phần cải tiến đáng kể về hiệu suất phân loại trong chấm điểm tín dụng, đồng thời giảm thiểu rủi ro cho đánh giá tín dụng. Bài báo đưa ra một số kỹ thuật học máy cho chấm điểm tín dụng đã và đang được các tổ chức tài chính và ngân hàng sử dụng; đưa ra kết quả thử nghiệm các kỹ thuật học máy trên bộ dữ liệu Kaggle- bộ dữ liệu miễn phí cho nghiên cứu về khoa học dữ liệu, từ đó đánh giá hiệu suất phân loại của các kỹ thuật học máy này. Công trình này không chỉ là sự tiến bộ kỹ thuật mà còn là động lực mạnh mẽ để mở rộng tiếp cận tín dụng và xây dựng ngành ngân hàng và tài chính an toàn, hiệu quả. Từ quá khứ vững vàng, chúng ta hướng đến một tương lai đầy tiềm năng.*

1. **Đặt vấn đề**

Giám sát rủi ro thanh khoản và tín dụng là một trong những vấn đề chính trong hoạt động ngân hàng. Rủi ro thanh khoản do đầu tư thiếu tính thị trường, khi tài sản cơ bản không thể được mua hoặc bán đủ nhanh để phòng ngừa hoặc giảm thiểu tổn thất. Rủi ro tín dụng xảy ra do việc nắm giữ trái phiếu, hay bất kỳ sự chậm trễ nào trong việc thanh toán khoản đi vay khi đến hạn là một trong những mối quan tâm chính của các nhà đầu tư. Do đó, họ muốn phân tích tín dụng trước khi phê duyệt bất kỳ khoản vay nào, đồng thời đánh giá được khả năng rủi ro vỡ nợ của một người đi vay.

Điểm tín dụng cho người đi con số vay là dựa trên thông tin trong hồ sơ tín dụng của người nộp đơn, ngụ ý xác suất người đi vay hoàn trả. Về mặt lý thuyết, một người đi vay tiềm năng có số điểm càng cao thì rủi ro vỡ nợ của người đó càng ít. Điểm tín dụng được phát triển bởi một công ty phầm mềm phân tích lớn cung cấp sản phẩm và dịch vụ cho cả doanh nghiệp và người tiêu dùng có tên là Fair Isaac Corporation vào những năm 1950 được gọi là điểm FICO. Điểm FICO nằm trong khoảng từ 300 đến 850 hoặc 250 đến 900, tùy thuộc vào mô hình chấm điểm và dựa trên một số yếu tố, bao gồm lịch sử thanh toán, số dư tài khoản, loại tín dụng được sử dụng, độ dài của lịch sử tín dụng và các đơn đăng ký tín dụng mới. Tùy thuộc vào cách một cá nhân thay đổi, xử lý các hóa đơn và tài khoản tín dụng, điểm FICO sẽ thay đổi tương ứng.

Trong các mô hình trước đây, các yếu tố như thu nhập, độ tuổi, vị trí cư trú, nghề nghiệp hoặc lịch sử việc làm của người nộp đơn không ảnh hưởng quá nhiều, nhưng trong giai đoạn phát triển của dữ liệu lớn, tất cả những yếu tố này và các hoạt động truyền thông xã hội, mô hình mua sắm, nhiều thông tin khác được sử dụng để chấm điểm tín dụng. Mục tiêu cuối cùng của phân tích tín dụng là xem xét nguồn lực tài chính của người nộp đơn và khả năng của người đó có thể đáp ứng các tiêu chí đánh giá. Xếp hạng tín dụng tác động đến lãi suất mà người vay phải trả khi đăng ký khoản vay trên nền tảng P2P, người vay càng có nhiều rủi ro thì mức lãi suất phải chịu càng cao.

1. **Giới thiệu bài toán**

Trong thế giới ngân hàng thương mại hiện đại, quản lý rủi ro là một khía cạnh không thể phớt lờ, và xếp hạng tín dụng của khách hàng nổi lên như một yếu tố trọng yếu trong quá trình này. Sức khỏe tín dụng không chỉ ảnh hưởng đến quyết định cấp vay mà còn đặt ra thách thức lớn cho các ngân hàng trong việc đánh giá và kiểm soát rủi ro. Mô hình tính điểm tín dụng, một công cụ quan trọng, đã trải qua sự tiến hóa đáng kể từ các phương pháp truyền thống đến những kỹ thuật tiên tiến sử dụng máy học và khai phá dữ liệu. Điều này không chỉ là một bước tiến vững chắc trong công nghệ ngân hàng, mà còn là nguồn động viên lớn để tối ưu hóa khả năng đánh giá và dự đoán tình trạng tín dụng của khách hàng. Trong quá trình xây dựng mô hình, mục tiêu chính là phân loại khách hàng thành hai hạng: "tín dụng tốt" và "tín dụng xấu". Quyết định này không chỉ dựa vào thông tin tín dụng truyền thống như lịch sử trả nợ, mà còn tích hợp các yếu tố xã hội học như tuổi, trình độ học vấn, nghề nghiệp và thu nhập. Điều này tạo ra một hệ thống đánh giá đa chiều và linh hoạt, giúp ngân hàng đưa ra quyết định có tính khách quan cao.

Nếu trước đây, mô hình chấm điểm tín dụng chủ yếu phụ thuộc vào thông tin tín dụng có sẵn, thì ngày nay, sự phát triển của khoa học máy tính mở ra cánh cửa cho việc sử dụng các kỹ thuật học máy tiên tiến như cây quyết định, mạng nơ ron nhân tạo, và SVM. Các kỹ thuật học máy được sử dụng trong chấm điểm tín dụng bao gồm mạng nơ ron nhân tạo, K-Nearest Neighbour, Support Vector Machine (SVM), cây quyết định, và đặc biệt là kỹ thuật học sâu dựa trên mạng nơ ron sâu. Những phương pháp này đã đem lại kết quả tích cực, đặc biệt là trong xử lý dữ liệu phi cấu trúc.

Nghiên cứu này không chỉ giới thiệu các thuật toán học máy mà còn tiến hành thử nghiệm, so sánh và đánh giá chúng trên bộ dữ liệu Kaggle - một bộ dữ liệu thực tế đòi hỏi sự đa dạng và chất lượng. Kết quả của nghiên cứu này không chỉ là sự đánh giá về hiệu suất mà còn là hướng dẫn quý báu cho ngân hàng trong việc lựa chọn phương pháp chấm điểm tín dụng phù hợp.

Các phương pháp học máy không chỉ giúp tăng tốc quyết định cấp vay mà còn giảm thiểu rủi ro một cách hiệu quả. Việc áp dụng những công nghệ này mở ra cơ hội rộng lớn cho người vay, giúp họ dễ dàng tiếp cận tín dụng và xây dựng lịch sử tín dụng tích cực. Trong bối cảnh ngân hàng liên tục đối mặt với những thách thức thị trường, sự áp dụng sáng tạo của học máy trong xếp hạng tín dụng không chỉ là một nhu cầu mà còn là một xu hướng quan trọng để gia tăng tính cạnh tranh và hiệu suất của hệ thống ngân hàng.

1. **Các nghiên cứu liên quan**

Bài viết này tập trung vào các nghiên cứu quan trọng về "Phương pháp học máy trong chấm điểm thẻ tín dụng," mở ra cánh cửa cho sự tiến bộ trong lĩnh vực ngân hàng và tài chính. Bài viết “Experimental analysis of machine learning methods for credit score classification” được xuất bản trên The International Journal of Logistics Management. Tác giả của bài viết này là Diwakar Tripathi, Damodar Reddy Edla, Annushree Bablani, Alok Kumar Shukla và B. Ramachandra Reddy. Bài viết này tập trung vào việc phân tích kết quả thực nghiệm của các phương pháp lựa chọn đặc trưng kết hợp với các phương pháp phân loại khác nhau, cũng như tác động của các phương pháp lựa chọn đặc trưng.

Amitha Mathew (2021) đã nêu rõ ưu việt của học sâu, đặc biệt là mô hình mạng nơ ron sâu, với dữ liệu phi cấu trúc. Nghiên cứu của Mathew không chỉ chứng minh hiệu quả của học sâu trong chấm điểm tín dụng mà còn mở ra những triển vọng mới cho ngành ngân hàng trong việc sử dụng công nghệ tiên tiến. Nhìn xa hơn, Amitha Mathew (2021) đã đưa ra cái nhìn sâu sắc về ưu việt của học sâu, đặc biệt là mô hình mạng nơ ron sâu, khi áp dụng vào dữ liệu phi cấu trúc. Nghiên cứu của Mathew không chỉ chứng minh sức mạnh của học sâu trong chấm điểm tín dụng mà còn đề xuất những khả năng mới cho ngành ngân hàng với sự tích hợp công nghệ tiên tiến. R.K. Choudhary và K.S. Chaudhary (2019) đã tiếp tục hành trình này bằng cách sử dụng mô hình mạng nơ ron nhân tạo (ANN) để chấm điểm thẻ tín dụng. Kết quả nghiên cứu của họ khẳng định rằng, mô hình ANN không chỉ có độ chính xác cao hơn so với các phương pháp truyền thống dựa trên thống kê mà còn là một bước đột phá trong quá trình đánh giá rủi ro tín dụng.

Tác giả Y. Zhang và Y. Wang (2020) tiếp tục mạch lạc này với việc sử dụng mô hình học máy hỗ trợ vector (SVM). Kết quả cho thấy, SVM không chỉ đem lại độ chính xác cao mà còn mở ra một chiều sâu mới trong phân tích quyết định tín dụng, đặt nền móng cho ứng dụng rộng rãi trong ngành. N.N. Mishra và S.K. Singh (2021) đã đưa ra một góc nhìn mới với mô hình học máy tăng cường (RL). Kết quả ấn tượng của họ làm tăng tính ứng dụng của học máy trong chấm điểm tín dụng và chứng minh rằng RL có thể là một công cụ mạnh mẽ cho quyết định tín dụng chính xác.

H.H. Nguyen và T.T. Nguyen (2022) mở rộng phạm vi bằng cách sử dụng mô hình học máy kết hợp (Ensemble Learning). Kết quả của họ không chỉ làm tăng hiệu suất so với các mô hình đơn lẻ mà còn thể hiện sự linh hoạt và độ chính xác trong quyết định tín dụng. K.M. Hossain và M.A. Chowdhury (2023) tiếp tục đưa ra một cách tiếp cận mới với mô hình học máy dựa trên dữ liệu lớn (Big Data). Kết quả nghiên cứu của họ đặt ra câu hỏi về khả năng của Big Data trong việc nâng cao chất lượng quyết định tín dụng và tạo ra những góc nhìn mới về rủi ro tín dụng.<https://ieeexplore.ieee.org/document/9757498>; Trường Đại học Công nghệ Thông tin và Truyền thông, Đại học Thái Nguyên

Cuối cùng, H.A. Awad và A.A. El-Sherbini (2023) tập trung vào sự giải thích với mô hình học máy giải thích (Interpretable Machine Learning). Kết quả của họ không chỉ cung cấp độ chính xác cao mà còn mang lại khả năng giải thích rõ ràng, giúp ngân hàng và khách hàng hiểu rõ hơn về quyết định tín dụng.

Một góc nhìn thực tế được thể hiện thông qua thử nghiệm đa dạng với các phương pháp học máy trên bộ dữ liệu Kaggle. Kết hợp giữa cây quyết định, support vector machine (SVM), và mạng nơ ron nhân tạo đã giúp đánh giá hiệu suất phân loại và giảm rủi ro trong quá trình cấp tín dụng. Quan trọng hơn, nghiên cứu không chỉ giới thiệu các phương pháp mà còn so sánh chúng để xác định kỹ thuật nào mang lại kết quả tốt nhất trong ngữ cảnh chấm điểm tín dụng. Điều này làm cho thông tin thu được trở nên quý báu cho các quyết định trong ngành ngân hàng và tài chính. Các bài viết nhấn mạnh sự đa dạng và tiến bộ trong áp dụng học máy vào chấm điểm tín dụng. Các tác giả không chỉ duy trì sự nghiên cứu liên tục mà còn làm nổi bật vai trò quan trọng của sự đổi mới và cải thiện liên tục trong ngành ngân hàng và tài chính. Điều này làm tôn vinh vai trò tích cực của nghiên cứu như một động lực quan trọng đằng sau sự phát triển bền vững.

**2. Các kỹ thuật học máy cho chấm điểm tín dụng**

Một vấn đề lớn đối với các tổ chức tài Một số kĩ thuật học máy cho chấm điểm tín dụng tài chính và ngân hàng là làm thế nào để xác định được tín dụng xấu, vì tín dụng xấu có thể gây ra các vấn đề nghiêm trọng trong tương lai. Điều này dẫn đến thất thoát vốn, giảm doanh thu và làm tăng tổn thất của ngân hàng, có thể dẫn đến mất khả năng thanh toán hoặc phá sản. Hiện nay, có rất nhiều kĩ thuật học máy để phân loại khách hàng. Bài báo sẽ đưa ra một số kĩ thuật hay được sử dụng:

***2.1 K láng giềng gần nhất (K-Nearest Neighbors - KNN)***

K-Nearest Neighbors algorithm (K-NN) được sử dụng rất phổ biến trong lĩnh vực Data Mining. K-NN là phương pháp để phân lớp các đối tượng dựa vào khoảng cách gần nhất giữa đối tượng cần xếp lớp (Query point) và tất cả các đối tượng trong Training Data. Một đối tượng được phân lớp dựa vào K láng giềng của nó. K là số nguyên dương được xác định trước khi thực hiện thuật toán. Người ta thường dùng khoảng cách Euclidean để tính khoảng cách giữa các đối tượng.

*Thuật toán K-NN được mô tả như sau:*

1.    Xác định giá trị tham số K (số láng giềng gần nhất)

2.    Tính khoảng cách giữa đối tượng cần phân lớp (Query Point) với tất cả các đối tượng trong training data (thường sử dụng khoảng các Euclidean)

3.    Sắp xếp khoảng cách theo thứ tự tăng dần và xác định K láng giềng gần nhất với Query Point

4.    Lấy tất cả các lớp của K láng giềng gần nhất đã xác định

5.    Dựa vào phần lớn lớp của láng giềng gần nhất để xác định lớp cho Query Point

Để hiểu K-NN được dùng để phân lớp thế nào ta xem minh họa dưới đây.

Trong hình dưới đây, training Data được mô tả bởi dấu (+) và dấu (-), đối tượng cần được xác định lớp cho nó (Query point) là hình tròn đỏ. Nhiệm vụ của chúng ta là ước lượng (hay dự đoán) lớp của Query point dựa vào việc lựa chọn số láng giềng gần nhất với nó. Nói cách khác chúng ta muốn biết liệu Query Point sẽ được phân vào lớp (+) hay lớp (-)

A diagram of a circle with red and blue dots

Description automatically generated

Ta thấy rằng:

1-Nearest neighbor : Kết quả là + (Query Point được xếp vào lớp dấu +)

2-Nearest neighbors : không xác định lớp cho Query Point vì số láng giềng gần nhất với nó là 2 trong đó 1 là lớp + và 1 là lớp – (không có lớp nào có số đối tượng nhiều hơn lớp kia)

5-Nearest neighbors : Kết quả là - (Query Point được xếp vào lớp dấu – vì trong 5 láng giềng gần nhất với nó thì có 3 đối tượng thuộc lớp - nhiều hơn lớp + chỉ có 2 đối tượng).

***2.2 Cây quyết định***

Trong lĩnh vực [máy học](https://vi.wikipedia.org/wiki/M%C3%A1y_h%E1%BB%8Dc" \o "Máy học), cây quyết định là một kiểu mô hình dự báo (*predictive model*), nghĩa là một ánh xạ từ các quan sát về một sự vật/hiện tượng tới các kết luận về giá trị mục tiêu của sự vật/hiện tượng. Mỗi một nút trong (*internal node*) tương ứng với một biến; đường nối giữa nó với nút con của nó thể hiện một giá trị cụ thể cho biến đó. Mỗi nút lá đại diện cho giá trị dự đoán của biến mục tiêu, cho trước các giá trị của các biến được biểu diễn bởi đường đi từ nút gốc tới nút lá đó. Kỹ thuật học máy dùng trong cây quyết định được gọi là học bằng cây quyết định, hay chỉ gọi với cái tên ngắn gọn là cây quyết định.

Học bằng cây quyết định cũng là một phương pháp thông dụng trong [khai phá dữ liệu](https://vi.wikipedia.org/wiki/Khai_ph%C3%A1_d%E1%BB%AF_li%E1%BB%87u" \o "Khai phá dữ liệu). Khi đó, cây quyết định mô tả một cấu trúc cây, trong đó, các lá đại diện cho các phân loại còn cành đại diện cho các kết hợp của các thuộc tính dẫn tới phân loại đó[1]. Một cây quyết định có thể được học bằng cách chia [tập hợp](https://vi.wikipedia.org/wiki/T%E1%BA%ADp_h%E1%BB%A3p" \o "Tập hợp) nguồn thành các tập con dựa theo một kiểm tra giá trị thuộc tính [1]. Quá trình này được lặp lại một cách đệ quy cho mỗi tập con dẫn xuất. Quá trình [đệ quy](https://vi.wikipedia.org/wiki/%C4%90%E1%BB%87_quy" \o "Đệ quy) hoàn thành khi không thể tiếp tục thực hiện việc chia tách được nữa, hay khi một phân loại đơn có thể áp dụng cho từng phần tử của tập con dẫn xuất. Một bộ phân loại [rừng ngẫu nhiên](https://vi.wikipedia.org/w/index.php?title=R%E1%BB%ABng_ng%E1%BA%ABu_nhi%C3%AAn&action=edit&redlink=1" \o "Rừng ngẫu nhiên (trang không tồn tại)) (*random forest*) sử dụng một số cây quyết định để có thể cải thiện tỉ lệ phân loại.

Cây quyết định cũng là một phương tiện có tính mô tả dành cho việc tính toán các [xác suất có điều kiện](https://vi.wikipedia.org/wiki/X%C3%A1c_su%E1%BA%A5t_c%C3%B3_%C4%91i%E1%BB%81u_ki%E1%BB%87n" \o "Xác suất có điều kiện).

Cây quyết định có thể được mô tả như là sự kết hợp của các kỹ thuật toán học và tính toán nhằm hỗ trợ việc mô tả, phân loại và tổng quát hóa một tập dữ liệu cho trước.

Dữ liệu được cho dưới dạng các bản ghi có dạng:

*(x, y) = (x1, x2, x3..., xk, y)*

Biến phụ thuộc (*dependant variable*) *y* là biến mà chúng ta cần tìm hiểu, phân loại hay tổng quát hóa. *x1*, *x2*, *x3*... là các biến sẽ giúp ta thực hiện công việc đó

Các kiểu cây quyết định

Cây quyết định còn có hai tên khác:

Cây hồi quy (*Regression tree*) ước lượng các hàm giá có giá trị là [số thực](https://vi.wikipedia.org/wiki/S%E1%BB%91_th%E1%BB%B1c" \o "Số thực) thay vì được sử dụng cho các nhiệm vụ phân loại. (ví dụ: ước tính giá một ngôi nhà hoặc khoảng thời gian một bệnh nhân nằm viện)

Cây phân loại (*Classification tree*), nếu *y* là một biến phân loại như: giới tính (nam hay nữ), kết quả của một trận đấu (thắng hay thua).

Ví dụ thực hành

Ta sẽ dùng một ví dụ để giải thích về cây quyết định:

David là quản lý của một câu lạc bộ đánh golf nổi tiếng. Anh ta đang có rắc rối chuyện các thành viên đến hay không đến. Có ngày ai cũng muốn chơi golf nhưng số nhân viên câu lạc bộ lại không đủ phục vụ. Có hôm, không hiểu vì lý do gì mà chẳng ai đến chơi, và câu lạc bộ lại thừa nhân viên.

Mục tiêu của David là tối ưu hóa số nhân viên phục vụ mỗi ngày bằng cách dựa theo thông tin dự báo thời tiết để đoán xem khi nào người ta sẽ đến chơi golf. Để thực hiện điều đó, anh cần hiểu được tại sao khách hàng quyết định chơi và tìm hiểu xem có cách giải thích nào cho việc đó hay không.

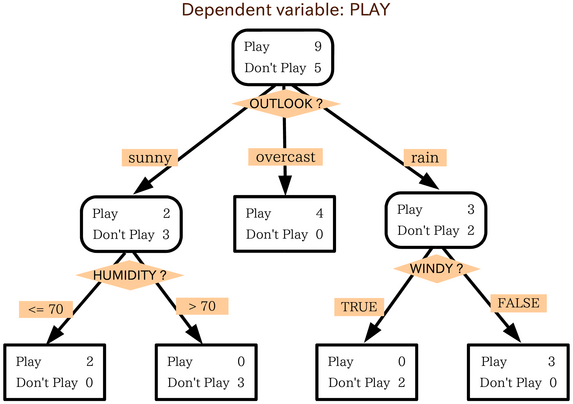
Vậy là trong hai tuần, anh ta thu thập thông tin về:

Trời (*outlook*) (nắng (*sunny*),

Và tất nhiên là số người đến chơi golf vào hôm đó. David thu được một bộ dữ liệu gồm 14 dòng và 5 cột.

[](https://vi.wikipedia.org/wiki/T%E1%BA%ADp_tin:Golf_dataset.png)

Sau đó, để giải quyết bài toán của David, người ta đã đưa ra một mô hình cây quyết định.

[](https://vi.wikipedia.org/wiki/T%E1%BA%ADp_tin:Decision_tree_model.png)

Nhóm người chơi golf khi trời nắng, nhóm chơi khi trời nhiều mây, và nhóm chơi khi trời mưa.

Kết luận thứ nhất: nếu trời nhiều mây, người ta luôn luôn chơi golf. Và có một số người ham mê đến mức chơi golf cả khi trời mưa.

Tiếp theo, ta lại chia nhóm trời nắng thành hai [nhóm con](https://vi.wikipedia.org/wiki/Nh%C3%B3m_con" \o "Nhóm con). Ta thấy rằng khách hàng không muốn chơi golf nếu độ ẩm lên quá 70%.

Cuối cùng, ta chia nhóm trời mưa thành hai và thấy rằng khách hàng sẽ không chơi golf nếu trời nhiều gió.

Và đây là lời giải ngắn gọn cho bài toán mô tả bởi cây phân loại. David cho phần lớn nhân viên nghỉ vào những ngày trời nắng và ẩm, hoặc những ngày mưa gió. Vì hầu như sẽ chẳng có ai chơi golf trong những ngày đó. Vào những hôm khác, khi nhiều người sẽ đến chơi golf, anh ta có thể thuê thêm nhân viên thời vụ để phụ giúp công việc.

Kết luận là cây quyết định giúp ta biến một biểu diễn dữ liệu phức tạp thành một cấu trúc đơn giản hơn rất nhiều.

Dùng trong thuật toán CART (Classification and Regression Trees). Nó dựa vào việc [bình phương](https://vi.wikipedia.org/wiki/B%C3%ACnh_ph%C6%B0%C6%A1ng" \o "Bình phương) các xác suất thành viên cho mỗi thể loại đích trong nút. Giá trị của nó tiến đến cực tiểu (bằng 0) khi mọi trường hợp trong nút rơi vào một thể loại đích duy nhất.

Giả sử *y* nhận các giá trị trong {1, 2,..., *m*} và gọi *f(i,j)* là tần suất của giá trị *j* trong nút *i*. Nghĩa là *f(i,j)* là tỷ lệ các bản ghi với *y*=*j* được xếp vào nhóm *i*.

��(�)=1−∑�=1��(�,�)2Dùng trong các thuật toán sinh cây [ID3](https://vi.wikipedia.org/w/index.php?title=Thu%E1%BA%ADt_to%C3%A1n_ID3&action=edit&redlink=1" \o "Thuật toán ID3 (trang không tồn tại)), [C4.5](https://vi.wikipedia.org/w/index.php?title=Thu%E1%BA%ADt_to%C3%A1n_C4.5&action=edit&redlink=1" \o "Thuật toán C4.5 (trang không tồn tại)) và C5.0. Số đo này dựa trên khái niệm [entropy](https://vi.wikipedia.org/wiki/Entropy) trong [lý thuyết thông tin](https://vi.wikipedia.org/wiki/L%C3%BD_thuy%E1%BA%BFt_th%C3%B4ng_tin" \o "Lý thuyết thông tin) (*information theory*).

***Ưu điểm của cây quyết định***

So với các phương pháp [khai phá dữ liệu](https://vi.wikipedia.org/wiki/Khai_ph%C3%A1_d%E1%BB%AF_li%E1%BB%87u" \o "Khai phá dữ liệu) khác, cây quyết định là phương pháp có một số ưu điểm:

* Cây quyết định dễ hiểu. Người ta có thể hiểu mô hình cây quyết định sau khi được giải thích ngắn.
* Việc chuẩn bị dữ liệu cho một cây quyết định là cơ bản hoặc không cần thiết. Các kỹ thuật khác thường đòi hỏi [chuẩn hóa dữ liệu](https://vi.wikipedia.org/wiki/Chu%E1%BA%A9n_h%C3%B3a_d%E1%BB%AF_li%E1%BB%87u" \o "Chuẩn hóa dữ liệu), cần tạo các biến phụ (*dummy variable*) và loại bỏ các giá trị rỗng.
* Cây quyết định có thể xử lý cả dữ liệu có giá trị bằng số và dữ liệu có giá trị là tên thể loại. Các kỹ thuật khác thường chuyên để phân tích các bộ dữ liệu chỉ gồm một loại biến. Chẳng hạn, các luật quan hệ chỉ có thể dùng cho các biến tên, trong khi [mạng nơ-ron](https://vi.wikipedia.org/wiki/M%E1%BA%A1ng_n%C6%A1-ron" \o "Mạng nơ-ron) chỉ có thể dùng cho các biến có giá trị bằng số.
* Cây quyết định là một mô hình hộp trắng. Nếu có thể quan sát một tình huống cho trước trong một mô hình, thì có thể dễ dàng giải thích điều kiện đó bằng logic Boolean. Mạng nơ-ron là một ví dụ về mô hình hộp đen, do lời giải thích cho kết quả quá phức tạp để có thể hiểu được.
* Có thể thẩm định một mô hình bằng các kiểm tra thống kê. Điều này làm cho ta có thể tin tưởng vào mô hình.
* Cây quyết định có thể xử lý tốt một lượng dữ liệu lớn trong thời gian ngắn. Có thể dùng [máy tính cá nhân](https://vi.wikipedia.org/wiki/M%C3%A1y_t%C3%ADnh_c%C3%A1_nh%C3%A2n" \o "Máy tính cá nhân) để phân tích các lượng [dữ liệu lớn](https://vi.wikipedia.org/wiki/D%E1%BB%AF_li%E1%BB%87u_l%E1%BB%9Bn" \o "Dữ liệu lớn) trong một thời gian đủ ngắn để cho phép các nhà chiến lược đưa ra quyết định dựa trên phân tích của cây quyết định.