**BỘ GIÁO DỤC VÀ ĐÀO TẠO**

**TRƯỜNG ĐẠI HỌC CẦN THƠ**

**TRƯỜNG CÔNG NGHỆ THÔNG TIN & TRUYỀN THÔNG**

**Logo

Description automatically generated**

**LUẬN VĂN**

**NGÀNH CÔNG NGHỆ THÔNG TIN**

**Đề tài**

**PHÂN TÍCH NHÂN TỐ DỰ ĐOÁN HIỆU SUẤT**

**HỌC TẬP CỦA SINH VIÊN BẰNG THUẬT**

**TOÁN HỌC MÁY**

**FACTOR ANALYSIS OF STUDENT ACADEMIC**

**PERFORMANCE PREDICTION USING CLASSICAL**

**MACHINE LEARNING ALGORITHMS**

**Sinh viên: Trương Huỳnh Tú Như – B2111893**

**Khóa: Khóa 47**

**Cần Thơ, 03/2025**

**BỘ GIÁO DỤC VÀ ĐÀO TẠO**

**TRƯỜNG ĐẠI HỌC CẦN THƠ**

**TRƯỜNG CÔNG NGHỆ THÔNG TIN & TRUYỀN THÔNG**

**Logo

Description automatically generated**

**LUẬN VĂN TỐT NGHIỆP ĐẠI HỌC**

**NGÀNH CÔNG NGHỆ THÔNG TIN**

**Đề tài**

**PHÂN TÍCH NHÂN TỐ DỰ ĐOÁN HIỆU SUẤT**

**HỌC TẬP CỦA SINH VIÊN BẰNG THUẬT TOÁN HỌC MÁY**

**FACTOR ANALYSIS OF STUDENT ACADEMIC**

**PERFORMANCE PREDICTION USING CLASSICAL**

**MACHINE LEARNING ALGORITHMS**

|  |  |
| --- | --- |
| **Người hướng dẫn** | **Sinh viên thực hiện** |
| **PGS.Ts. Nguyễn Thanh Hải**  **Ths. Sử Kim Anh** | **Trương Huỳnh Tú Như**  **Mã số: B2111893**  **Khóa 47** |

***Cần Thơ, 03/2025***

**LỜI CẢM ƠN**

Trong suốt bốn năm học đại học cũng như hơn ba tháng thực hiện đề tài, em đã nhận được sự quan tâm, hỗ trợ tận tình từ quý thầy cô, gia đình và bạn bè. Đây là nguồn động lực to lớn, giúp em có cơ hội học tập, trau dồi và trưởng thành như hôm nay.

Em xin bày tỏ lòng biết ơn chân thành đến Quý Thầy Cô tại Trường Công nghệ Thông tin và Truyền thông, Trường Đại học Cần Thơ, những người không chỉ trang bị cho em kiến thức chuyên môn vững chắc mà còn truyền cảm hứng về tinh thần trách nhiệm và niềm đam mê trong học tập cũng như nghiên cứu.

Em muốn bày tỏ lòng biết ơn chân thành đến cô Sử Kim Anh và thầy Nguyễn Thanh Hải, người đã nhiệt tình hướng dẫn và chỉ bảo em trong suốt quá trình làm luận văn. Nhờ những kiến thức và kinh nghiệm thực tế mà thầy cô chia sẻ, em đã có thể hoàn thành tốt các nhiệm vụ trong quá trình nghiên cứu. Sự hỗ trợ quý báu từ thầy cô đã giúp em vượt qua khó khăn và hoàn thiện bài báo cáo này một cách tốt nhất.

Bên cạnh đó, em muốn gửi lời cảm ơn đến các bạn bè, những người đã luôn đồng hành, động viên và chia sẻ kinh nghiệm quý báu trong suốt quá trình thực hiện niên luận.

Trong quá trình nghiên cứu và thực hiện bài báo cáo, mặc dù đã cố gắng hết sức, nhưng không tránh khỏi những sai sót và thiếu sót. Em rất mong nhận được sự thông cảm và ý kiến đóng góp từ Quý Thầy Cô và các anh chị để em có thể rút kinh nghiệm và hoàn thiện hơn.

Trân trọng!

|  |  |
| --- | --- |
|  | Cần Thơ, ngày 7 tháng 03 năm 2025  Sinh viên thực hiện |

**NHẬN XÉT CỦA GIẢNG VIÊN HƯỚNG DẪN**

.......................................................................................................................................

.......................................................................................................................................

.......................................................................................................................................

.......................................................................................................................................

.......................................................................................................................................

.......................................................................................................................................

.......................................................................................................................................

.......................................................................................................................................

.......................................................................................................................................

.......................................................................................................................................

.......................................................................................................................................

.......................................................................................................................................

.......................................................................................................................................

.......................................................................................................................................

.......................................................................................................................................

.......................................................................................................................................

.......................................................................................................................................

.......................................................................................................................................

.......................................................................................................................................

.......................................................................................................................................

.......................................................................................................................................

.......................................................................................................................................

.......................................................................................................................................

|  |  |
| --- | --- |
| Giảng viên hướng dẫn | Cần Thơ, ngày 7 tháng 12 năm 2024  Giảng viên hướng dẫn |

**TÓM TẮT**

Các mô hình học máy hoạt động trên dữ liệu gene family có thể yêu cầu tài nguyên tính toán đáng kể do số lượng đặc trưng rất lớn, dẫn đến thời gian huấn luyện kéo dài và đòi hỏi rất nhiều thời gian để tính toán. Việc chọn lọc đặc trưng không chỉ giúp giảm thiểu chi phí tính toán, đẩy nhanh quá trình xử lý mà còn hạn chế tối đa việc loại bỏ những đặc trưng quan trọng có ảnh hưởng đến kết quả dự đoán. Trong nghiên cứu này, chúng tôi đánh giá ảnh hưởng của phương pháp Lime trong việc chọn lọc đặc trưng kết hợp với mô hình Random Forest (RF). Kết quả thực nghiệm cho thấy phương pháp chọn lọc đặc trưng dựa trên Lime đạt độ chính xác tốt hơn so với phương pháp chọn lọc truyền thống của RF. Đặc biệt, trong hầu hết các trường hợp Lime giúp cải thiện độ chính xác so với dữ liệu gốc và đặc biệt trên các tập dữ liệu T2D, COL, OBE và WT2D. Bên cạnh đó, khi so sánh với các phương pháp trước đây như PCA và RD\_PRO, kết hợp RF + Lime cho thấy hiệu suất phân loại tốt hơn. Những kết quả này cho thấy Lime là một phương pháp hiệu quả trong việc lựa chọn đặc trưng, giúp nâng cao độ chính xác của mô hình mà vẫn giữ lại các yếu tố quan trọng.

***Từ khóa:*** *Gene Family; Lime; Học máy; Lựa chọn đặc trưng*

**ABTRACT**

Machine learning models operating on gene family data can require significant computational resources due to the large number of features, resulting in long training times and requiring a lot of computational time. Feature selection not only helps to reduce computational costs and speed up the processing but also minimizes the removal of important features that affect the prediction results. In this study, we evaluated the effect of the Lime method on feature selection combined with the Random Forest (RF) model. Experimental results show that the Lime-based feature selection method achieves better accuracy than the traditional RF selection method. In particular, in most cases Lime improves accuracy compared to the original data and especially on the T2D, COL, OBE, and WT2D datasets. In addition, compared to previous methods such as PCA and RD \textunderscore PRO, the combination of RF + Lime shows better classification performance. These results show that Lime is an effective method for feature selection, which improves the accuracy of the model while retaining important factors.

***Keywords****: Gene family; Lime; Machine Learning; Feature selection*

**MỤC LỤC**

[CHƯƠNG 1. GIỚI THIỆU 12](#_Toc184413157)

[1.1. Đặt vấn đề 12](#_Toc184413158)

[1.2. Những nghiên cứu liên quan 13](#_Toc184413159)

[1.3. Mục tiêu đề tài 15](#_Toc184413160)

[1.4. Đối tượng và phạm vi nghiên cứu 15](#_Toc184413161)

[1.4.1. Đối tượng nghiên cứu 15](#_Toc184413162)

[1.4.2. Phạm vi nghiên cứu 15](#_Toc184413163)

[1.5. Phương pháp nghiên cứu 16](#_Toc184413164)

[CHƯƠNG 2. CƠ SỞ LÝ THUYẾT 17](#_Toc184413165)

[2.1. Python 17](#_Toc184413167)

[2.2. Machine learning 17](#_Toc184413168)

[2.2.1. SVM 17](#_Toc184413169)

[2.2.2. Naive Bayes 17](#_Toc184413170)

[2.2.3. Random Forest 18](#_Toc184413171)

[2.2.4. Decision Tree 18](#_Toc184413172)

[2.2.5. Logistic Regression 18](#_Toc184413173)

[CHƯƠNG 3. PHƯƠNG PHÁP THỰC HIỆN 19](#_Toc184413174)

[3.1. Tập dữ liệu 19](#_Toc184413178)

[3.2. Trực quan hoá dữ liệu 20](#_Toc184413181)

[3.3. Tiền xử lý dữ liệu và xếp hạng các đặc trưng bằng thuật toán Gain Ratio 22](#_Toc184413183)

[CHƯƠNG 4. KẾT QUẢ THỰC NGHIỆM 24](#_Toc184413190)

[4.1. Thiết lập môi trường 24](#_Toc184413195)

[4.2. Cơ sở đánh giá 24](#_Toc184413196)

[4.3. Điều chỉnh siêu tham số và đào tạo mô hình 25](#_Toc184413197)

[4.3.1. Mô hình SVM 25](#_Toc184413198)

[4.3.2. Mô hình Navie Bayes 26](#_Toc184413204)

[4.3.3. Mô hình Random Forest 27](#_Toc184413213)

[4.3.4. Mô hình Decision Tree 28](#_Toc184413214)

[4.3.5. Mô hình Logistic Regression 29](#_Toc184413215)

[4.3.6. So sánh các mô hình 29](#_Toc184413216)

[CHƯƠNG 5. KẾT LUẬN 33](#_Toc184413217)

[5.1. Kết quả đạt được 33](#_Toc184413223)

[5.2. Hạn chế 33](#_Toc184413224)

[5.3. Hướng phát triển 33](#_Toc184413225)

DANH MỤC HÌNH

[Hình 1: Biểu đồ phân phối các giá trị số 21](#_Toc184413226)

[Hình 2: Biểu đồ phân phối các giá trị phân loại 21](#_Toc184413227)

[Hình 3: Biểu đồ sắp xếp thuộc tính theo tỷ lệ trung bình của các tính năng 23](#_Toc184413228)

[Hình 4: Kết quả của mô hình SVM 25](#_Toc184413229)

[Hình 5: Kết quả của mô hình Naive Bayes 26](#_Toc184413230)

[Hình 6: Kết quả của mô hình Random Forest 27](#_Toc184413231)

[Hình 7: Kết quả của mô hình Decision Tree 28](#_Toc184413232)

[Hình 8: Kết quả của mô hình Logistic Regression 29](#_Toc184413233)

[Hình 9: So sánh Accuracy giữa các thuật toán khác nhau 30](#_Toc184413234)

[Hình 10: So sánh F1-score giữa các thuật toán khác nhau 31](#_Toc184413235)

**DANH MỤC BẢNG**

[Bảng 1: Mô tả các thuộc tính có trong dữ liệu 19](#_Toc184413237)

[Bảng 2: So sánh mô hình đề xuất của chúng tôi với nghiên cứu gần đây 31](#_Toc184413238)

**DANH MỤC TỪ VIẾT TẮT**

|  |  |  |
| --- | --- | --- |
| **Viết tắt** | **Diễn giải** | **Nghĩa tiếng việt** |
| XAI | Explainable Artificial Intelligence | Trí tuệ nhân tạo có thể giải thích |
| LIME | Local Interpretable Model-Agnostic Explanations | Giải thích mô hình cục bộ có thể diễn giải được - không phụ thuộc |
| RF | Random Forest | Rừng ngẫu nhiên |
| RNA | Ribonucleic Acid | Axit Ribonucleic |
| DNA  DT  SVM  ML  KNN  DL  NN  ACC  PCA  PcoA  CCA  NMF  CNN  SFS  PDPs  AS  AUC  MCC | Deoxyribonucleic Acid  Decision tree  Support vector machine  Machine learning  K -Nearest Neighbors  Deep Learning  Neural Network  Accuracy  Principal Component Analysis  Principal Coordinates Analysis  Correspondence Analysis  Non-negative Matrix Factorization  Convolutional neural network  Sequential forward selection  Partial dependence plots  Attribute selection  Area Under Curve  Matthews Correlation Coefficient | Axit Deoxyribonucleic  Cây quyết định  Máy vector hỗ trợ  Học máy  K láng giềng  Học sâu  Mạng lưới nơ-ron  Độ chính xác  Phân tích thành phần chính  Phân tích tọa độ chính  Phân tích thư từ  Phân tích ma trận không âm  Mạng nơ-ron tích chập  Lựa chọn tuần tự chuyển tiếp  Biểu đồ phụ thuộc một phần  Lựa chọn thuộc tính  Diện tích dưới đường cong  Hệ số tương quan Matthews |

# CHƯƠNG 1. GIỚI THIỆU

## Đặt vấn đề

Hệ vi sinh vật, với sự đa dạng và phức tạp của nó đã tạo nên một hệ sinh thái độc đáo. Việc khám phá và hiểu biết về sự đa dạng này là chìa khóa cho cả nghiên cứu sức khỏe lẫn môi trường [1]. Trong hệ vi sinh vật, có khoảng hơn 100 nghìn tỷ vi sinh vật sống cộng sinh, chúng không chỉ đóng vai trò quan trọng trong quá trình trao đổi chất mà còn ảnh hưởng đến nhiều khía cạnh sức khỏe của con người. Chính vì thế, việc khám phá và hiểu rõ tác động của hệ vi sinh vật đối với sức khỏe cũng như các bệnh lý đã trở thành một thách thức lớn, đồng thời là cơ hội đầy tiềm năng trong nghiên cứu y sinh học hiện đại [2].

Khái niệm "Y học cá nhân hóa" đang là 1 xu hướng nghiên cứu mới trong lĩnh vực y tế. Phương pháp tiếp cận này dựa trên việc nhận biết sự khác biệt giữa các bệnh nhân mắc cùng một bệnh, đồng thời thay đổi quan niệm "một giải pháp phù hợp cho tất cả" [3]. Thông thường, khi điều trị bệnh cho bệnh nhân, các bác sĩ sẽ dùng chung phương pháp, điều này dẫn đến khả năng khỏi bệnh không cao do sự biểu hiện gen của mỗi cá nhân là không giống nhau. Từ đó, phương pháp điều trị bằng phương pháp "y học cá nhân hoá" ra đời để đáp ứng mục tiêu nâng cao hiệu quả chữa bệnh [4].

Từ xưa đã có nhiều phương pháp nghiên cứu về vi sinh vật bằng hình thức nuôi cấy, tuy nhiên nó có những hạn chế [5]. Dù có rất nhiều bộ gen vi khuẩn, nhưng chỉ có một tỷ lệ nhỏ trong số chúng có thể nuôi cấy thành công. Cộng thêm việc ở trong tự nhiên các vi khuẩn cùng loài thường không tương tác trực tiếp với nhau. Vì vậy, nuôi cấy vô tính không phản ánh đầy đủ mối quan hệ giữa các vi khuẩn cũng như những biến đổi sinh học của chúng. Trong những năm gần đây, một phương pháp tiên tiến với tên gọi metagenomics đã được phát triển để nghiên cứu vi sinh vật trong một môi trường nhất định bằng cách sử dụng kỹ thuật sàng lọc gen chức năng hoặc phân tích trình tự [6]. Trong y học, 1 trong những mục tiêu của nó là nhằm nâng cao sức khoẻ của con người. Metagenomics có thể được áp dụng để nhận biết được các gen kháng thuốc trong mầm bệnh, cũng như giám sát các đợt bùng phát của bệnh. Sử dụng phương pháp metagenomics mang đến một công cụ mạnh mẽ để xác định các loài vi sinh vật liên quan đến bệnh, đồng thời giúp hiểu rõ quá trình đồng chuyển hóa giữa vật chủ và hệ vi khuẩn với độ chính xác cao hơn trong phân loại. Một ví dụ cụ thể như Các loại virus về hô hấp RNA bao gồm Coronaviridae, Orthomyxoviridae,.. đã được phát hiện trong nghiên cứu [7] trong thời gian sớm. Nhờ vào công nghệ giải trình tự thế hệ mới được trích dẫn bởi [8].

Gene family là tập hợp các gen có nguồn gốc từ một tổ tiên chung, chúng được hình thành qua quá trình nhân đôi gen và tiến hóa dần theo thời gian, những gen này có cùng nguồn gốc và có chức năng sinh học tương tự nhau, các gen trong cùng một họ thường có trình tự DNA tương đồng. Các gen được sắp xếp vào cùng một họ dựa trên sự tương đồng về trình tự nucleotide hoặc protein [9]. Cụ thể sự hình thành của họ gen là khi một gen bị sao chép, và các bản sao này có thể thay đổi theo thời gian, sau đó tạo ra những gen với chức năng mới hoặc tương tự thông qua 3 cơ chế nhân đôi gen, chuyển gen và phân hoá gen. Hiện tại, số lượng gene family vẫn chưa thể xác định chính xác vì nó phụ thuộc vào cách phân loại và dữ liệu sử dụng. Một số ứng dụng của nó phải kể đến như việc giúp hiểu sâu hơn về cơ chế di truyền và sự liên quan của chúng với các bệnh di truyền, ngoài ra tối ưu hóa các phương pháp điều trị và dự đoán nguy cơ mắc bệnh. Có thể giúp xác định các vùng gene có liên quan đến bệnh tật từ đó nghiên cứu ra các mục tiêu phát triển thuốc [10]. Dữ liệu về các gene family có thể mở rộng hiểu biết của chúng ta về sự đa dạng và chức năng của các cộng đồng vi sinh vật, để từ đó ứng dụng vào nghiên cứu môi trường, sức khỏe con người, và nhiều lĩnh vực khác. Cụ thể trong y học, khi xác định được các họ gen liên quan đến bệnh di truyền sẽ có thể tìm ra các phương pháp điều trị mới, đặc biệt các bệnh hiểm nghèo hay bệnh hiếm gặp khi mà hiện nay nó là nổi sợ hãi của nhiều người. Ngoài ra, đối với vi sinh vật, khi phân tích sâu hơn về họ gen của nó có thể giúp phát hiện các cơ chế kháng thuốc để từ đó có nhu cầu phát triển các loại kháng sinh mới [11].

Dù các mô hình dự đoán sử dụng (Machine learning) ML hứa hẹn mang lại những hiểu biết về sinh học của hệ vi sinh vật, nhưng vẫn còn một số thách thức và hạn chế cần được khắc phục. Với khối lượng dữ liệu khổng lồ từ metagenomics, đồng nghĩa với việc tính đa chiều cao có thể gây ra hiện tượng quá khớp và làm suy giảm khả năng tổng quát hóa, ảnh hưởng đến hiệu suất [12]. Việc dữ liệu có nhiều chiều sẽ gây khó khăn trong việc giải thích và tính toán số liệu thống kê. Thêm vào đó, nhiều tính năng có thể bị nhiễu hoặc dư thừa, và số lượng giả thuyết để giải thích dữ liệu thường vượt xa số lượng quan sát và tính năng. Điều này tạo ra nhu cầu về một kỹ thuật giảm chiều để chuyển đổi dữ liệu có số chiều lớn thành dạng dữ liệu với số chiều nhỏ hơn. Trong quá trình xử lý, việc đảm bảo tính toàn vẹn những đặc điểm của bộ gen phải được ưu tiên hàng đầu. Đây là một bước thiết yếu, giúp tạo ra các hình ảnh dữ liệu dễ hiểu cho con người, giúp giảm độ phức tạp của mô hình, tăng tốc độ xử lý, và đôi khi cải thiện độ chính xác bằng cách loại bỏ các đặc trưng không liên quan, đồng thời cung cấp cơ sở cho các phân tích sâu hơn.

Trong nghiên cứu này chúng tôi sử dụng 1 công cụ giúp phân tích và giảm số chiều của bộ dữ liệu gene family, được gọi là LIME (Local Interpretable Model-Agnostic Explanations) [13]. Với bất kỳ một nghiên cứu, cụ thể như gene family thì vấn đề minh bạch luôn là yếu tố không thể thiếu. Và vai trò của LIME giúp mô tả cách thức hoạt động của các thuật toán trong việc phân tích, xử lý dữ liệu và hiểu rõ các đặc trưng cụ thể mà mô hình máy học tập trung vào khi đưa ra dự đoán. LIME có khả năng xử lý các đặc trưng liên tục và cung cấp các ước tính mang tính cục bộ nhằm giải thích từng dự đoán riêng lẻ đồng thời đánh giá mức độ tác động của từng đặc trưng đến kết quả dự đoán [14]. Thông qua việc áp dụng LIME, chúng tôi có thể xác định được những đặc trưng nào có ảnh hưởng quan trọng đến quyết định của mô hình. Dựa trên kết quả phân tích này, chúng tôi sẽ chọn lọc các đặc trưng có tác động lớn để tái huấn luyện mô hình, nhằm cải thiện độ chính xác và hiệu quả cao hơn.

Đóng góp của nghiên cứu:

* Nghiên cứu này đánh giá hiệu quả của Lime trong lựa chọn đặc trưng trên dữ liệu gene family và so sánh hiệu suất với phương pháp chọn lọc đặc trưng truyền thống của Random Forest (RF).
* Thực hiện so sánh Lime với các phương pháp chọn đặc trưng khác và cho thấy Lime hoạt động tốt, độ chính xác cao hơn so với PCA và RD\_PRO trên các tập dữ liệu tương đương.
* Nghiên cứu phân tích tác động của việc giảm chiều dữ liệu đến mô hình RF, cho thấy Lime giúp giảm số đặc trưng đầu vào mà vẫn duy trì độ chính xác cao, thậm chí vượt trội so với mô hình RF sử dụng toàn bộ đặc trưng gốc.

## Những nghiên cứu liên quan

Với những tiến bộ hiện nay, các phương pháp học máy (ML) và học sâu (DL) đã nổi lên như những cách tiếp cận hàng đầu để giải quyết nhiều thách thức khác nhau trong các ứng dụng chăm sóc sức khỏe thông minh, bao gồm phân tích hình ảnh y tế, dự đoán bệnh tật, và khám phá thuốc [15]. Nó đóng vai trò là bộ não cũng như kiến thức của bác sĩ, việc kết hợp máy móc và con người hứa hẹn mang lại những bước tiến đột phá trong y học, từ chẩn đoán chính xác hơn đến điều trị cá nhân hóa và quản lý sức khỏe hiệu quả hơn. Nhiều nhà nghiên cứu đã và đang nỗ lực trong lĩnh vực này để đưa ra những cải tiến và phát triển các tính năng mới [16]. Đã có rất nhiều công trình nghiên cứu liên quan đến đề tài này sẽ được giới thiệu.

Nghiên cứu [17] đã dùng 5 thuật toán học máy có giám sát khác nhau như rừng ngẫu nhiên (RF), lưới đàn hồi (EN), cây quyết định (DT), máy vectơ hỗ trợ (SVM) và mạng nơ-ron (NN). Trong số đó, kết quả RF hoạt động tốt nhất trong các mô hình để dự đoán các bệnh. 12 kỹ thuật phân loại được trích dẫn trong [18], đã chỉ ra lộ trình các phương pháp dựa trên máy học để dự đoán gen bệnh. Bao gồm các phương pháp dựa trên nhị phân, đơn phân, bán giám sát, tổng hợp và học sâu. Cuối cùng, các phương pháp sử dụng cây và mạng nơ-ron đạt hiệu suất tương đối tốt, thể hiện qua các phép đo độ chính xác (ACC).

Trong nghiên cứu khác, các nhà nghiên cứu đã áp dụng phương pháp giảm chiều dựa trên Recursive Feature Elimination kết hợp với mô hình Random Forest nhằm cải thiện độ chính xác của các tác vụ dự đoán trên dữ liệu metagenomic [19]. Kết quả cho thấy, phương pháp này mang lại độ chính xác cao, đạt 0,927% với ba mươi đặc trưng được lựa chọn, tạo ra sự cải thiện rõ rệt so với các kỹ thuật lựa chọn đặc trưng ngẫu nhiên. Một phương pháp mới sử dụng autoencoders được trích dẫn trong [20], đây là 1 kỹ thuật học sâu giúp hạn chế những trở ngại của các phương pháp giảm chiều hiện có. Quá trình này được thực hiện bằng cách chuyển đổi các vector tần suất nucleotide thành một không gian chiều thấp hơn, từ đó giúp giảm tải tính toán và cải thiện khả năng nhận dạng các cụm dữ liệu. Kết quả cho thấy việc sử dụng Autoencoders không chỉ giúp giảm chiều dữ liệu hiệu quả mà còn duy trì được các đặc tính quan trọng của dữ liệu ban đầu. Với kỹ thuật Autoencoders tương tự như trên, nghiên cứu [21] giảm chiều dữ liệu bằng cách mã hóa dữ liệu metagenomics thành không gian chiều thấp hơn trong quá trình huấn luyện mạng nơ-ron, và sau đó dữ liệu được chuyển đổi trở lại không gian ban đầu, nhằm so sánh với dữ liệu gốc và điều chỉnh mạng neural để học cách biểu diễn tối ưu. Kết quả của phương pháp này cho thấy nó vượt trội hơn so với các phương pháp giảm chiều khác như Phân tích thành phần chính (PCA) và t-SNE. Nhiều kỹ thuật giảm chiều dữ liệu được so sánh trong [22]. Điển hình như phương pháp Principal Coordinates Analysis (PCoA) hoạt động tốt trên dữ liệu không tuyến tính và hỗ trợ phân tách các mẫu thuộc các môi trường sống khác nhau. Principal Component Analysis (PCA) tập trung vào giữ lại các thành phần chính có phương sai lớn nhất. Tuy nhiên, nó có hạn chế trong việc xử lý dữ liệu phân loại của microbiome. Bên cạnh đó, Canonical Correspondence Analysis (CCA) là 1 phương pháp cho phép kết hợp các biến giải thích trong quá trình giảm chiều, hiệu quả trong việc phân biệt các mẫu dựa trên các yếu tố môi trường, ... Nghiên cứu [23] đã sử dụng nhiều phương pháp lọc đặc trưng như principal component analysis (PCA), nó giúp tìm các chiều mới, để biểu diễn dữ liệu một cách tối ưu nhất dựa trên sự biến thiên trong dữ liệu ban đầu. Ngoài ra còn có phương pháp Non negative Matrix Factorization (NMF) giúp giảm chiều dữ liệu bằng cách phân tách ma trận dữ liệu ban đầu thành hai ma trận nhỏ hơn, với các giá trị không âm. Thay vì làm việc với ma trận gốc có nhiều biến số, NMF giảm số chiều xuống còn ít hơn, giúp giữ lại thông tin quan trọng.

Ở nhiều bài báo khác nhau, các nhà khoa học đã sử dụng Lime để giải thích và giảm chiều cho dữ liệu phức tạp, đặc biệt trong các lĩnh vực như sinh học, y học và xử lý ngôn ngữ tự nhiên. Bài báo [24] đề xuất phương pháp XAI-reduct để giảm chiều dữ liệu trong phân loại bệnh tim trong khi vẫn duy trì độ chính xác tương đương với dữ liệu gốc. Các mô hình học máy được dùng như Random Forest, SVM, Neural Networks và áp dụng XAI (bao gồm SHAP, LIME, SHAPASH, PDP và DALEX) để chọn đặc trưng quan trọng nhất. Kết quả cho thấy XAI-reduct giúp giảm số đặc trưng mà không ảnh hưởng đến độ chính xác, đồng thời cải thiện tính minh bạch và độ tin cậy của mô hình trong chẩn đoán bệnh tim. Nghiên cứu của tác giả Jaeyoung Shin xuất bản vào năm 2023 đã đánh giá tính hiệu quả của thuật toán LIME trong việc chọn lọc đặc trưng cho dữ liệu fNIRS được trích dẫn trong [25]. Nghiên cứu so sánh hiệu suất của LIME với các phương pháp chọn lọc đặc trưng khác như minimum-redundancy maximum-relevance (mRMR), t-test và sequential forward selection (SFS) trên nhiều bộ dữ liệu fNIRS công khai. Kết quả cho thấy LIME vượt trội hơn các phương pháp khác về độ chính xác sau khi phân loại, chứng minh tính hiệu quả của LIME trong việc chọn lọc đặc trưng cho dữ liệu fNIRS. Một mô hình học máy được sử dụng nhằm phát hiện sớm bệnh Parkinson được trích dẫn trong [26]. Nghiên cứu sử dụng mô hình mạng nơ-ron tích chập (CNN) VGG16 kết hợp với kỹ thuật học chuyển giao để phân loại hình ảnh DaTSCAN. Với độ chính xác 95,20%, độ nhạy 97,50% và độ đặc hiệu 90,90%, mô hình cho thấy hiệu suất cao trong việc nhận diện bệnh Parkinson. Để đảm bảo tính minh bạch, nghiên cứu áp dụng phương pháp LIME để cung cấp các giải thích trực quan cho từng dự đoán, giúp phân biệt giữa bệnh nhân Parkinson và người không mắc bệnh. Kết quả cho thấy hệ thống đề xuất có thể hỗ trợ hiệu quả các nhân viên y tế trong việc chẩn đoán sớm bệnh Parkinson. Nghiên cứu [27] trình bày phương pháp Specific-Input LIME nhằm cải thiện giải thích mô hình học sâu trên dữ liệu dạng bảng. Phương pháp này kết hợp kỹ thuật LIME truyền thống với tầm quan trọng của đặc trưng (feature importance) và biểu đồ phụ thuộc từng phần (PDPs) để thể hiện rõ ảnh hưởng của các đặc trưng đến dự đoán. Kết quả thực nghiệm cho thấy phương pháp giúp tăng tính ổn định của LIME và cân bằng giữa giải thích cục bộ và toàn cục. Bài báo [28] nghiên cứu việc giảm chiều dữ liệu trong bộ dữ liệu biểu hiện gene, sử dụng phương pháp lựa chọn thuộc tính (AS) và phân tích thành phần chính (PCA). Dữ liệu được sử dụng là các bộ gene expression, với mục tiêu cải thiện độ chính xác phân loại. Phương pháp đề xuất kết hợp đánh giá tính nhất quán (CSE), độ dư thừa tối thiểu và độ liên quan tối đa mRMR) cho thấy lựa chọn thuộc tính vượt trội hơn PCA, kết quả gia tăng đáng kể so với khi dùng toàn bộ tập dữ liệu.

## Mục tiêu đề tài

Mục tiêu chính của nghiên cứu trong bài viết này là giúp chọn ra các đặc trưng quan trọng trong dữ liệu gene family để giảm thiểu tính phức tạp của dữ liệu đồng thời cải thiện hiệu suất tốt hơn dữ liệu ban đầu. Mô hình học máy được sử dụng trong nghiên cứu này là mô hình RF. Quá trình cụ thể như sau:

* Dùng phương pháp RF để lựa chọn ra 1000 đặc trưng quan trọng hàng đầu trong dữ liệu gốc tương ứng với hàng triệu đặc trưng.
* Sử dụng phương pháp giải thích mô hình LIME để xác định 100 đặc trưng quan trọng nhất trong dữ liệu vừa được chọn.
* Sử dụng các đặc trưng quan trọng từ các phương pháp LIME và RF để huấn luyện mô hình học máy: RF để phân loại bệnh hoặc không bệnh.
* Tiếp theo độ chính xác của mô hình RF được đánh giá dựa trên 2 phương pháp lựa chọn đặc trưng bằng LIME và RF. Các phương pháp sau đó sẽ được so sánh để xác định phương pháp tối ưu nhất cho dữ liệu gene family.
* Cuối cùng, chúng tôi so sánh kết quả dự đoán với kết quả thực tế từ các nghiên cứu trước để đánh giá tính hiệu quả của các mô hình.

## Đối tượng và phạm vi đề tài

Với đề tài này, đối tượng nghiên cứu của chúng tôi là dữ liệu gene family, bao gồm các đặc trưng sinh học của các loài vi sinh vật có trong mẫu bệnh thu thập từ cơ thể người, … Bộ dữ liệu này tổng cộng có 6 tập, trong đó gồm các bệnh xơ gan (CIR), ung thư đại trực tràng (COL), viêm ruột (IBD), béo phì (OBE), đái tháo đường (T2D, WT2D).

Mục tiêu phân loại học lực:

- Đề tài này nghiên cứu và cải tiến các phương pháp mô hình máy học để thực hiện phân loại các học lực của học sinh.

- Xử lý làm sạch dữ liệu trước khi đưa vào huấn luyện mô hình.

- So sánh các mô hình máy học chưa tinh chỉnh và đã tinh chỉnh.

## Phương pháp nghiên cứu

- Nghiên cứu các tài liệu tham khảo, bài báo có liên quan đến máy học trên các tạp chí uy tín.

- Nghiên cứu tài liệu liên quan đến các bộ dataset đã dùng.

- Nghiên cứu các thuật toán khác nhau để đánh giá kết quả đạt được khi chạy với dataset đã dùng.

## Bố cục quyển luận văn

- Chương 1: Giới thiệu.

- Chương 2: Cơ sở lý thuyết.

- Chương 3: Thiết kế và cài đặt giải pháp.

- Chương 4: Kiểm thử và đánh giá.

- Chương 5: Kết luận.

# CHƯƠNG 2. CƠ SỞ LÝ THUYẾT



## Python

Python là một ngôn ngữ lập trình được ưa chuộng và sử dụng nhiều trong phát triển web, phần mềm, khoa học dữ liệu và học máy. Với cú pháp ngắn gọn, đơn giản và chạy được trên nhiều nền tảng khác nhau, vì thế đã thu hút được đông đảo lập trình viên ở mọi châu lục. Đây là phần mềm mã nguồn mở, miễn phí, có thể tích hợp dễ dàng với các loại hệ thống bất kỳ, giúp nâng cao hiệu suất và giảm thiểu thời gian chạy và phát triển ứng dụng.

Ưu điểm:

* Đơn giản và dễ đọc: Cú pháp của ngôn ngữ này rõ ràng và trực quan, mã của nó cũng ngắn gọn, dễ hiểu. Đối với những người mới bắt đầu vào lập trình sẽ thấy dễ hơn.
* Tính linh hoạt: nó có khả năng ứng dụng rộng rãi trong nhiều lĩnh vực như web, khoa học dữ liệu, học máy, và tự động hóa. Với nguồn thư viện phong phú, có thể sử dụng theo nhiều mục đích tuỳ theo nhu cầu cá nhân.
* Tính cộng đồng: là ngôn ngữ mã nguồn mở nên Python có một cộng đồng lớn mạnh, cung cấp nhiều tài liệu và công cụ miễn phí. Ta có thể dễ dàng tìm kiếm sự hỗ trợ, học hỏi qua các diễn đàn, khóa học.

## Machine learning

Máy học (ML), là lĩnh vực nghiên cứu cho phép máy tính tự cải thiện hiệu suất dựa trên dữ liệu mẫu (training data) hoặc kinh nghiệm tích lũy từ những gì đã học được. Nhờ đó, máy học có khả năng tự dự đoán hoặc đưa ra quyết định mà không yêu cầu phải lập trình cụ thể trước.



### RF

Rừng ngẫu nhiên là một thuật toán học máy có giám sát có thể giải quyết cả bài toán phân loại, hồi quy. Rừng Ngẫu Nhiên được hình thành từ nhiều cây quyết định khác nhau, mỗi cây sẽ được huấn luyện trên một tập con của bộ dữ liệu và sau đó dự đoán kết quả. Các tập con này được chọn một cách ngẫu nhiên để đảm bảo cho mô hình đa dạng, không bị trùng lặp. Khi quá trình học kết thúc, kết quả dự đoán từ từng cây sẽ được tổng hợp lại (thông qua giá trị dự đoán trung bình hoặc phương pháp biểu quyết) để đưa ra kết quả cuối cùng, từ đó giúp cải thiện độ chính xác và khả năng tổng quát hóa của mô hình.

## Local Interpretable Model-Agnostic Explanations

Sự phát triển của trí tuệ nhân tạo phụ thuộc vào mức độ tin cậy, vì vậy các mô hình không đưa ra lời giải thích rõ ràng cho quyết định của chúng sẽ khó được chấp nhận. Để tăng cường tính minh bạch và nâng cao sự tin tưởng, phương pháp LIME đã được đề xuất nhằm làm sáng tỏ cách thức mô hình đưa ra dự đoán.

Bằng cách thay đổi ngẫu nhiên các giá trị đặc trưng và quan sát sự thay đổi trong kết quả dự đoán, LIME xây dựng một tập dữ liệu mới nhằm xác định mức độ ảnh hưởng của từng đặc trưng đến quyết định của mô hình. Hiểu một cách đơn giản, LIME là phương pháp giúp phân tích mức độ ảnh hưởng của từng đặc trưng đối với mô hình. LIME tạo mẫu mới bằng cách thay đổi đặc trưng dữ liệu, tính khoảng cách với dữ liệu gốc và gán trọng số. Dựa trên trọng số này, LIME xác định các đặc trưng quan trọng nhất và huấn luyện một mô hình tuyến tính bất kỳ để giải thích ảnh hưởng của chúng đến dự đoán của mô hình.



Về mặt toán học, Lime có thể được thể hiện như sau:

 (2.1)

Công thức trên thể hiện đầy đủ mô hình toán học của phương pháp LIME do tác giả Ribeiro đề xuất trong [13]. Cụ thể:

+ : xác định giá trị g trong tập G sao cho hàm bên trong đạt giá trị nhỏ nhất.

+ *f*: mô hình phức tạp ban đầu.

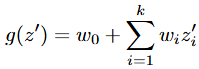
+ *g*: mô hình đơn giản đang được tìm kiếm để giải thích *f.*

+ *x*: phân bố xác suất xung quanh điểm , giúp tạo dữ liệu mới.

*+ L(f,g,x):* là một hàm mất mát để đo độ sai lệch giữa mô hình gốc *f* và mô hình đơn giản hóa *g*.

*+ Ω(g):* là một hàm phạt áp dụng lên *g* để đảm bảo nó đơn giản và dễ hiểu*.*

Kết quả của công thức trên là một mô hình tuyến tính 𝑔, 𝑔 sẽ có công thức sau:

 (2.2)

Trong đó:

+ *w0*: giá trị chặn.

+ *wi*: trọng số (Contribution) của từng đặc trưng *i*.

+ *zi’*: giá trị của đặc trưng *i* trong dạng đơn giản hóa.

Các trọng số *wi* chính là giá trị đóng góp của từng đặc trưng vào quyết định của mô hình. Chúng tôi sẽ sử dụng nó để sắp xếp và chọn các tính năng quan trọng nhất được mô tả trong công thức () dưới đây:

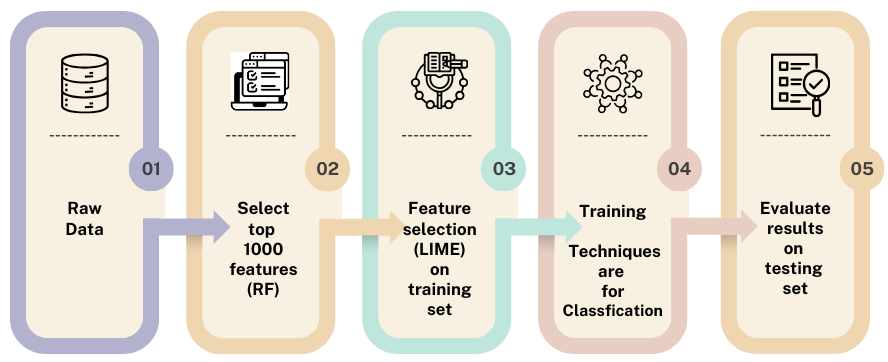
*Trọng số i = wi*  (2.3)

# CHƯƠNG 3. THIẾT KẾ VÀ CÀI ĐẶT GIẢI PHÁP



## Kiến trúc tổng quát

Trong đề tài này, chúng tôi đề xuất một quy trình giảm chiều dữ liệu dựa trên sự kết hợp giữa thuật toán Random Forest và phương pháp giải thích mô hình LIME nhằm lựa chọn các đặc trưng quan trọng trong tập dữ liệu gene family gốc. Kiến trúc tổng quát được thể hiện ở hình.



Ban đầu, tập dữ liệu gốc bao gồm hai tệp: một tệp chứa các đặc trưng của từng mẫu và một tệp chứa nhãn tương ứng. Chúng tôi sẽ thực hiện việc loại bỏ cột đầu và chuyển vị tệp đầu tiên, ở tệp thứ 2 chỉ lấy cột cuối nhằm đảm bảo tính tương thích của 2 tệp. Sau đó, thuật toán Random Forest được sử dụng để đánh giá mức độ quan trọng của từng đặc trưng thông qua chỉ số quan trọng đặc trưng (việc này chỉ thực hiện trên tập train). Với các thông số như n\_estimators = 100, max\_depth = None, min\_samples\_split = 2, min\_samples\_leaf = 1. Dựa trên kết quả này, chúng tôi chọn ra 1000 đặc trưng có đóng góp cao nhất vào quá trình phân loại. Lý do chúng tôi dùng Random Forest để lọc ra 1.000 đặc trưng tốt nhất trước khi sử dụng Lime là để giảm bớt yêu cầu về bộ nhớ. Vì kích thước tập dữ liệu gốc quá lớn, nếu chạy Lime trực tiếp trên toàn bộ dữ liệu, máy tính của chúng tôi sẽ không đủ khả năng xử lý.

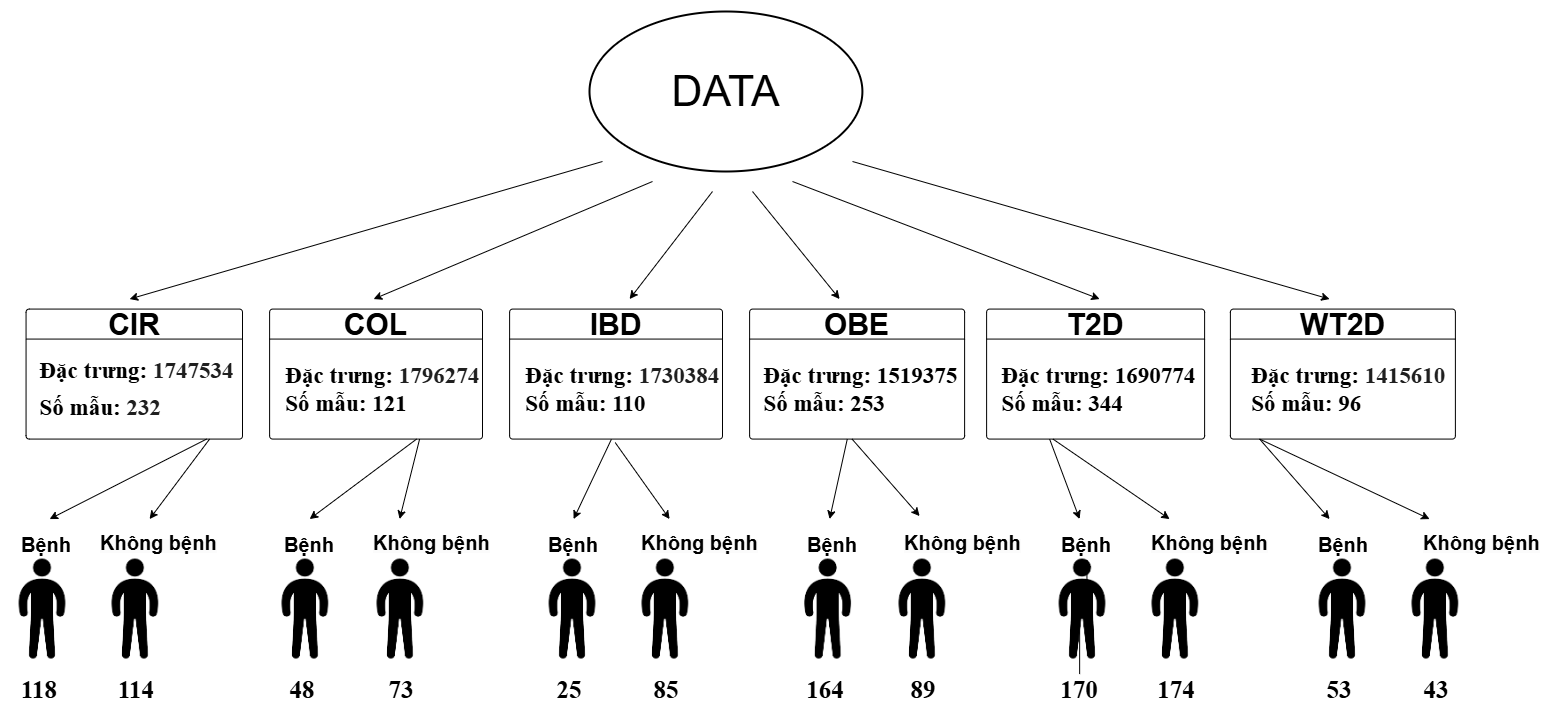
Bước tiếp theo, phương pháp LIME (Local Interpretable Model-Agnostic Explanations) được áp dụng để giải thích ảnh hưởng của từng đặc trưng trong tập 1000 đặc trưng đã chọn. Hiểu một cách đơn giản, LIME là phương pháp giúp phân tích mức độ ảnh hưởng của từng đặc trưng đối với mô hình. Thay vì chỉ giải thích một vài trường hợp riêng lẻ, trong nghiên cứu này, chúng tôi áp dụng LIME để phân tích toàn bộ mẫu trong tập dữ liệu huấn luyện. Sau đó, chúng tôi tính trung bình giá trị đóng góp (contribution) của mỗi đặc trưng trên tất cả các mẫu, lấy trị tuyệt đối và chọn ra 100 đặc trưng có mức đóng góp cao nhất. Những đặc trưng này sau đó được sử dụng để chạy lại mô hình, giúp tối ưu hiệu suất trong khi vẫn giữ được thông tin quan trọng nhất từ dữ liệu ban đầu

Dựa trên danh sách đặc trưng được rút gọn này, chúng tôi tiến hành chạy lại mô hình Random Forest với các tập con đặc trưng có kích thước lần lượt là 10, 20, ..., 100. Mỗi mô hình được đánh giá dựa trên các chỉ số như độ chính xác (Accuracy), độ chính xác (Precision), độ nhạy (Recall) và điểm F1 (F1-score) để xác định mức độ ảnh hưởng của việc giảm chiều dữ liệu đối với hiệu suất phân loại. Ngoài việc lựa chọn đặc trưng bằng LIME, chúng tôi cũng thực hiện thí nghiệm với danh sách đặc trưng được chọn trực tiếp từ chỉ số quan trọng đặc trưng (Feature Importance) của Random Forest. Cụ thể, chúng tôi sử dụng top 10, 20, ..., 100 đặc trưng quan trọng nhất từ Random Forest để huấn luyện lại mô hình và so sánh với kết quả của phương pháp LIME.

Ttrong hầu hết các trường hợp, kết quả cho thấy các đặc trưng được chọn bởi LIME mang lại độ chính xác cao hơn so với đặc trưng chọn từ Random Forest, điều này chứng minh rằng việc xem xét ảnh hưởng của đặc trưng lên từng mẫu dữ liệu giúp tối ưu hóa quá trình giảm chiều dữ liệu. Do đó, phương pháp kết hợp giữa LIME và Random Forest không chỉ giúp giảm số lượng đặc trưng mà còn có thể cải thiện hiệu suất phân loại so với việc chỉ sử dụng Random Forest đơn thuần. Bằng cách so sánh kết quả của các mô hình này, chúng tôi có thể xác định tập đặc trưng tối ưu giúp cân bằng giữa độ chính xác và tính giải thích của mô hình

## Tập dữ liệu

Trong nghiên cứu này, chúng tôi sử dụng sáu tập dữ liệu chứa thông tin về sự phong phú của họ gen mang đến những hiểu biết quan trọng về các yếu tố di truyền có liên quan đến sự phát triển của nhiều bệnh lý. Cụ thể, nghiên cứu tập trung vào các bệnh liên quan đến rối loạn chuyển hóa và viêm nhiễm, bao gồm viêm ruột (IBD), xơ gan (CIR), ung thư đại trực tràng (COL), béo phì (OBE) và tiểu đường loại 2 (T2D) và WT2D được trích xuất từ gói CuratedMetagenomicData, như đã mô tả trong nghiên cứu [29]. Mỗi bệnh lý này đều có sự liên quan đến các biến thể gene khác nhau, đóng vai trò quan trọng trong cơ chế hình thành và phát triển bệnh.



Trong các nhóm bệnh, CIR có số lượng đặc trưng là 1,747,534 với 118 mẫu bệnh và 114 mẫu không bệnh. Nhóm COL bao gồm 1,796,274 đặc trưng, 48 mẫu bệnh và 73 mẫu không bệnh. Đối với IBD, số lượng đặc trưng đạt 1,730,384, với 25 mẫu bệnh và 85 mẫu không bệnh. Nhóm OBE có số lượng đặc trưng thấp hơn, đạt 1,519,375, với 164 mẫu bệnh và 89 mẫu không bệnh. Trong khi đó, T2D có 170 mẫu bệnh và 174 mẫu không bệnh với 1,690,774 đặc trưng. Cuối cùng, WT2 có số lượng đặc trưng thấp nhất là 1,415,610, với 53 mẫu bệnh và 43 mẫu không bệnh.

Để có được cái nhìn tổng quát về dữ liệu, chúng tôi đã tiến hành trực quan hóa, đây là một bước quan trọng không chỉ giúp minh họa mà còn hỗ trợ phân tích, giải thích kết quả nghiên cứu và đưa ra quyết định chính xác. Bên cạnh đó, việc trực quan hóa còn giúp xác định dữ liệu bị thiếu hoặc mất cân bằng, tạo cơ sở cho các bước xử lý tiếp theo.

## Mô hình dự đoán

Quá trình chọn mô hình thử nghiệm đóng vai trò hết sức quan trọng. Trong nghiên cứu này, chúng tôi sử dụng một khuôn khổ tính toán có sẵn là Deepmg được tác giả đề xuất trong [30]. DeepMG giúp hỗ trợ trực quan hóa, huấn luyện và dự đoán dữ liệu bằng các thuật toán học máy và học sâu, cung cấp nhiều bộ phân loại như Mạng Nơ-ron Tích Chập (CNN), Hồi Quy Tuyến Tính, Máy Hỗ Trợ Vectơ (SVM), K-láng giềng gần nhất (KNN), Tăng cường độ dốc (Gradient Boosting) và Rừng Ngẫu Nhiên (Random Forest). Trong nghiên cứu này, chúng tôi sử dụng DeepMG để huấn luyện mô hình phân loại dữ liệu bằng thuật toán Rừng Ngẫu Nhiên (Random Forest).

Mô hình Random Forest đã được thiết lập với các tham số như sau: n\_estimators là 500 (tương ứng số lượng cây trong rừng), max\_depth là -1 (không giới hạn độ sâu của cây), min\_samples\_split là 2 (quy định số lượng mẫu tối thiểu để một nút có thể phân chia), max\_features là -2 (số lượng đặc trưng được chọn ngẫu nhiên tại mỗi lần phân chia), và random\_state là None (duy trì tính ngẫu nhiên của mô hình).



# CHƯƠNG 4. KIỂM THỬ VÀ ĐÁNH GIÁ



## Thiết lập môi trường

Để bắt đầu nghiên cứu, các công cụ cần thiết được chuẩn bị. Quá trình nghiên cứu được thực hiện trong hai môi trường ảo riêng biệt. Môi trường thứ nhất được cài đặt trên máy tính cá nhân, sử dụng Python 3.6 trong môi trường ảo Anaconda. Môi trường thứ hai được triển khai trên máy chủ của trường với phiên bản Python tương tự. Nghiên cứu sử dụng nhiều thư viện quan trọng như: Numpy, deepmg, scikit-learn, keras, pandas, … đã được cài đặt vào máy. Mã nguồn được thực thi trên máy tính chạy hệ điều hành Ubuntu 22.04.5 LTS, CPU Intel(R) Core(TM) i5-6200U CPU @ 2.30GHz, tổng dung lượng bộ nhớ: 12GB.

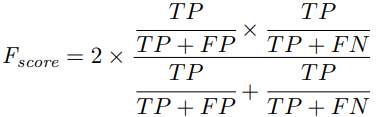
## Các độ đo đánh giá hiệu quả

Trong gói Deepmg, tác giả sử dụng các phương pháp đánh giá mô hình như accuracy (ACC), Area Under Curve (AUC), Matthews Correlation Coefficient (MCC), f1-score, Recall, ... để đánh giá hiệu suất của mô hình. Để đánh giá các độ đo trên, 4 tình huống dự đoán dưới đây là không thể thiếu. Bao gồm: True Positive (TP), False Positive (FP), True Negative (TN), False Negative (FN).

Trong bài toán này, chỉ số ACC được sử dụng làm thước đo chính để đánh giá hiệu suất của mô hình phân loại sau khi quá trình trích chọn đặc trưng được thực hiện. Về mặt khái niệm, độ chính xác thể hiện tỷ lệ giữa số lượng dự đoán đúng và tổng số dự đoán (bao gồm cả đúng và sai). Công thức tính được trình bày trong Phương trình (4.1):

 (4.1)

Điểm F1 được xác định bằng cách kết hợp đồng thời độ chính xác và độ nhạy. Việc sử dụng TP và FN giúp đánh giá độ nhạy, trong khi TP và FP được dùng để đo lường độ chính xác. Công thức tính điểm F1 được trình bày trong Phương trình (4.2):

 (4.2)

Bên cạnh đó còn độ đo AUC, một mô hình được coi là hoàn hảo khi giá trị AUC đạt 1. Giá trị AUC càng lớn thì hiệu suất phân loại của mô hình càng cao. Công thức tính AUC được thể hiện trong phương trình (4.3) dưới đây:

 (4.3)

Trong đó:

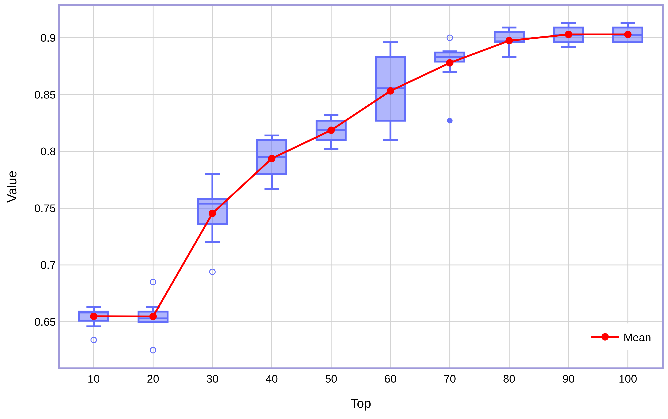
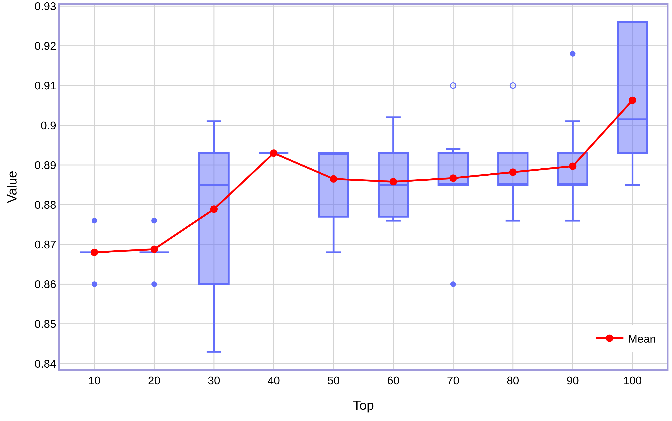
* TP: đây là số trường hợp mà mô hình dự đoán đúng là dương tính.
* TN: đây là số trường hợp mà mô hình dự đoán đúng là âm tính.
* FP: đây là số trường hợp mà mô hình dự đoán nhầm từ âm tính thành dương tính.
* FN: đây là số trường hợp mà mô hình dự đoán nhầm từ dương tính thành âm tính.
* TPR (True Positive Rate): tỉ lệ mô hình dự đoán là dương tính và kết quả là dương tính.
* FPR (False Positive Rate): tỉ lệ mô hình dự đoán là dương tính nhưng kết quả lại là âm tính.

## Kết quả kiểm thử

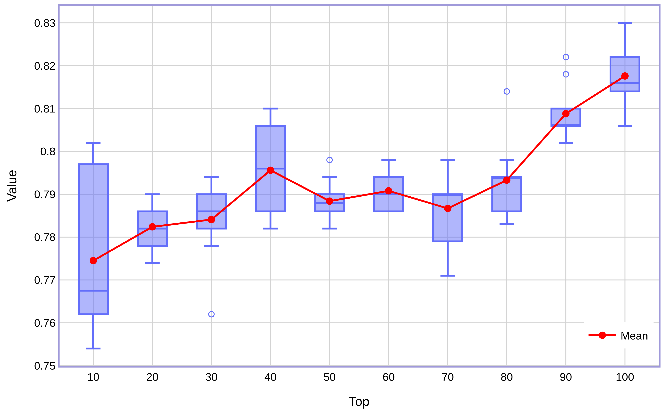
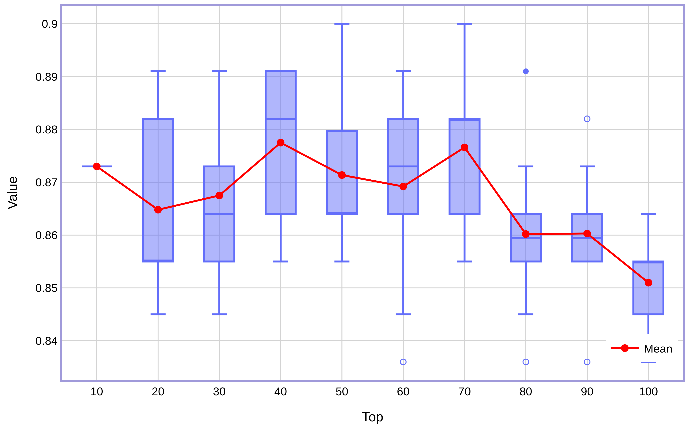


### LIME

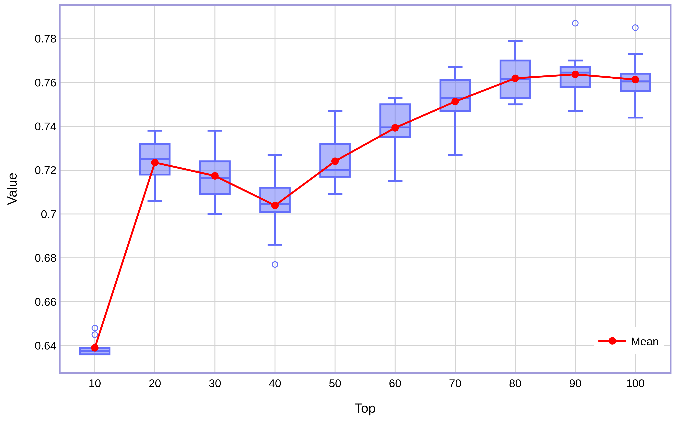
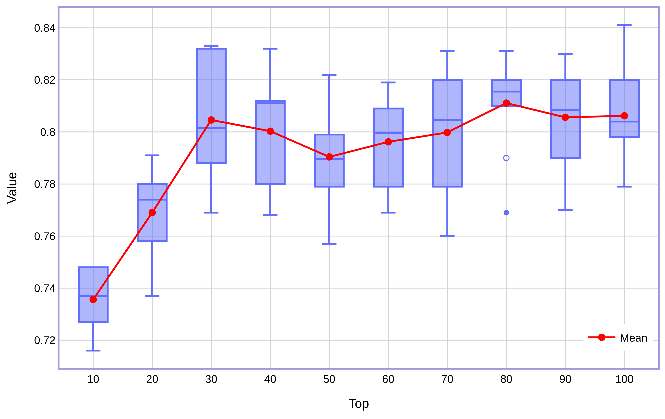
Sau khi tiến hành biểu diễn các kết quả được thu thập như hình A với 6 tập dữ liệu CIR, COL, IBD, OBE, T2D, WT2D em thấy rằng, hiệu suất hầu hết các tập có xu hướng tăng khi số lượng đặc trưng tăng. Trong đó, một số tập có sự thay đổi rõ rệt, trong khi một số khác chỉ biến đổi nhẹ.



1. CIR (b) COL

(c) IBD (d) OBE



(e) T2D (f) WT2D

Với các tập dữ liệu CIR, OBE và T2D, em quan sát thấy sự thay đổi đáng kể trong phân bố dữ liệu. Cụ thể, trong hình A-a với kết quả của CIR, giá trị tăng dần khi số lượng đặc trưng (Top) tăng, từ khoảng 0.65 ở Top 10 lên xấp xỉ 0.9 ở Top 100. Đạt kết quả cao nhất tại top 90 và 100 với giá trị 0.9029. Độ phân tán ban đầu khá nhỏ, sau đó tăng lên ở mức khoảng Top 40 - 60 và giảm dần khi đạt khoảng mức Top 80 - 100, cho thấy mô hình hoạt động ổn định hơn khi số lượng đặc trưng nhiều hơn. Với tập dữ liệu OBE trong hình A, cho thấy giá trị trung bình có xu hướng tăng dần theo số lượng đặc trưng (Top). Cụ thể, ở mức Top 10 đến Top 30, giá trị trung bình có xu hướng tăng nhẹ nhưng không đáng kể. Tuy nhiên, từ Top 40 trở đi, giá trị trung bình bắt đầu tăng rõ rệt, đạt đỉnh ở Top 100 với kết quả 0.8176. Về độ phân tán, ban đầu, độ phân tán khá lớn ở các mức Top 10 và Top 40. Tuy nhiên, từ Top 50 trở đi thì có xu hướng giảm dần, cho thấy mô hình trở nên ổn định hơn. Bên cạnh đó, sự xuất hiện của outlier cho thấy có một số giá trị bất thường so với phần lớn dữ liệu. Tuy nhiên, số lượng outlier không nhiều, cho thấy mô hình không bị ảnh hưởng quá lớn bởi nhiễu. Tương tự, với kết quả được thu thập như hình A của tập dữ liệu T2D. Hiệu suất có xu hướng tăng dần đồng biến theo số lượng đặc trưng (Top). Cụ thể, từ Top 10 đến Top 20: giá trị trung bình tăng mạnh từ 0.64 lên khoảng 0.72. Từ Top 20 đến Top 40: giá trị trung bình giảm nhẹ từ 0.72 xuống 0.7. Điều này có thể do một số đặc trưng mới được thêm vào không đóng góp nhiều hoặc có tác động không ổn định. Từ Top 50 trở đi: Giá trị trung bình có xu hướng tăng ổn định, đạt đỉnh tại top 90 với giá trị 0.7637. Điều này cho thấy khi số lượng đặc trưng đủ lớn, mô hình đạt hiệu suất tối ưu. Độ phân tán nhìn chung không quá lớn và giảm dần khi tăng số đặc trưng, chứng tỏ việc chọn nhiều đặc trưng giúp mô hình Lime ổn định hơn.

Ngược lại, các tập COL, IBD và WT2D cho thấy sự thay đổi ít rõ rệt hơn khi áp dụng phương pháp LIME. Với tập dữ liệu COL trong hình A, giá trị trung bình có xu hướng tăng dần theo số lượng đặc trưng (Top). Ở Top 10 và Top 20, giá trị trung bình gần như không thay đổi đáng kể. Từ Top 30 trở đi, giá trị trung bình bắt đầu tăng, có sự giảm nhẹ ở top 50 và đạt đỉnh ở Top 100 với kết quả ACC là 0.9063. Về độ phân tán, ở top 30 ta thấy độ phân tán khá lớn và giảm dần khi số lượng đặc trưng càng cao, tuy nhiên ở top 100 lại tăng trở lại. Ngoài ra, các điểm outlier xuất hiện nhiều ở Top 10, 20 và 70. Tuy nhiên, số lượng outlier không quá nhiều, và kết quả trung bình của mô hình vẫn tăng đều, cho thấy sự ảnh hưởng của outlier là không đáng kể. Đối với tập IBD, có thể nhận thấy rằng giá trị trung bình giảm dần khi số lượng đặc trưng tăng. Từ Top 30 đến Top 50, giá trị trung bình tăng mạnh, đạt đỉnh tại Top 40 với giá trị 0.8775. Từ Top 50 trở đi, giá trị trung bình tiếp tục dao động nhưng có xu hướng giảm rõ rệt từ Top 80 trở đi, đặc biệt tại Top 100, giá trị giảm xuống mức thấp nhất (0.851). Về độ phân tán, ta thấy độ phân tán tổng thế khá cao và xuất hiện nhiều outlier. Điều này cho thấy, có thể mô hình bị ảnh hưởng của một số đặc trưng không ổn định. Tại Top 80 đến 100, phương sai thu hẹp hơn, nhưng giá trị trung bình lại giảm xuống, cho thấy mô hình có thể đã mất đi một số đặc trưng quan trọng, dẫn đến hiệu suất thấp hơn. Cuối cùng là tập WT2D, ở hình A chúng tôi thấy giá trị trung bình có xu hướng tăng dần theo số lượng đặc trưng (Top). Cụ thể, tăng mạnh từ top 10 đến 30. Sau đó, giá trị trung bình dao động không đáng kể, không tăng mạnh như giai đoạn đầu. Kết quả trung bình cao nhất là 0.8111 tại vị trí top 80. Độ phân tán của phần này cũng hơi cao và giảm dần khi số đặc trưng tăng. Có xuất hiện Outlier ở Top 80, nhưng không quá ảnh hưởng đến kết quả trung bình.

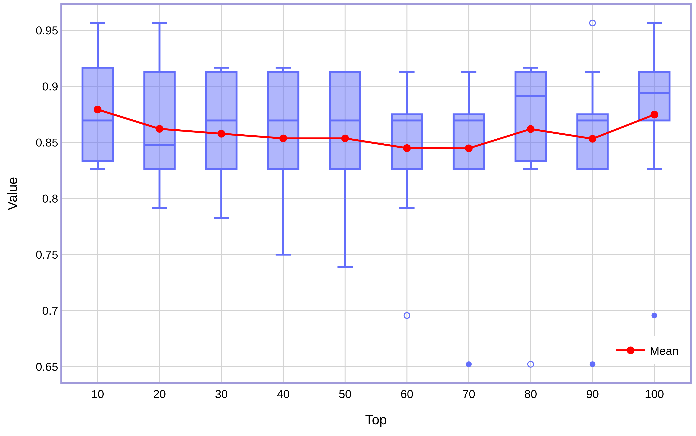
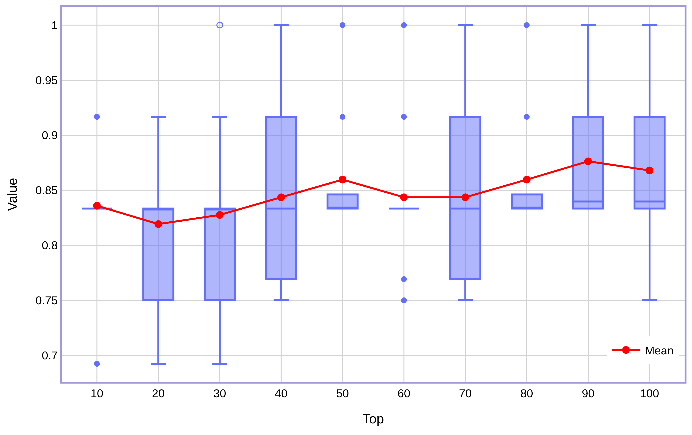
|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | Accuracy | AUC | Precision | F1-score |
| CIR | 0.8102 ±0.0917 | 0.8800 ±0.0937 | 0.8633±0.0519 | 0.7892±0.1271 |
| COL | 0.8852 ±0.0104 | 0.9467 ±0.0340 | 0.9276±0.0151 | 0.8312±0.0203 |
| IBD | 0.8671 ±0.0078 | 0.9360±0.0145 | 0.6652±0.0466 | 0.5072±0.0496 |
| OBE | 0.7922 ±0.0121 | 0.8333 ±0.0251 | 0.8127±0.0171 | 0.8461±0.0082 |
| T2D | 0.7285 ±0.0358 | 0.8055 ±0.0408 | 0.7354±0.0531 | 0.7279±0.0204 |
| WT2D | 0.7920 ±0.0218 | 0.8673 ±0.0297 | 0.8185±0.0273 | 0.6340±0.0147 |

Bảng 2 trình bày kết quả đánh giá mô hình trên các tập dữ liệu sử dụng đặc trưng được chọn bằng Lime. Nhìn chung, mô hình đạt độ chính xác (Accuracy) và AUC cao trên hầu hết các tập dữ liệu, với COL có hiệu suất tốt nhất (Accuracy = 0.8852, AUC = 0.9467). Độ chính xác trên T2D thấp nhất (0.7285), nhưng vẫn duy trì AUC tương đối tốt (0.8055). Precision và F1-score cũng cho thấy sự ổn định trên các tập dữ liệu, với COL và WT2D có giá trị cao nhất, trong khi IBD có F1-score thấp nhất (0.5072), thể hiện sự mất cân bằng trong phân loại. Kết quả này cho thấy phương pháp chọn lọc đặc trưng bằng Lime giúp mô hình đạt hiệu suất tốt trên nhiều tập dữ liệu, nhưng vẫn có sự khác biệt về hiệu suất giữa các loại bệnh

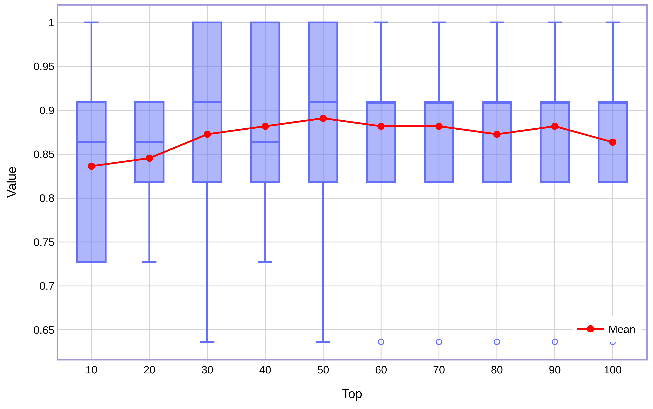
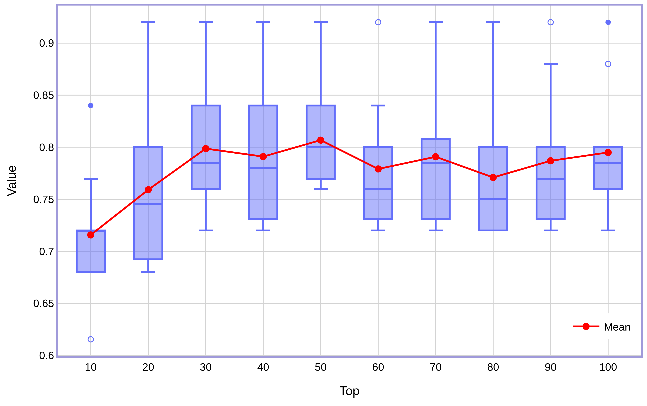


### RF

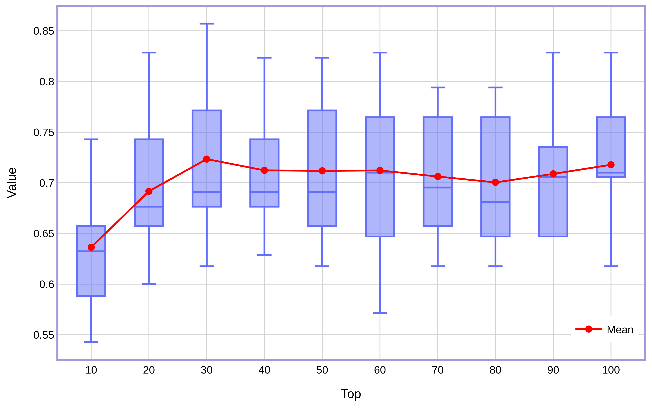
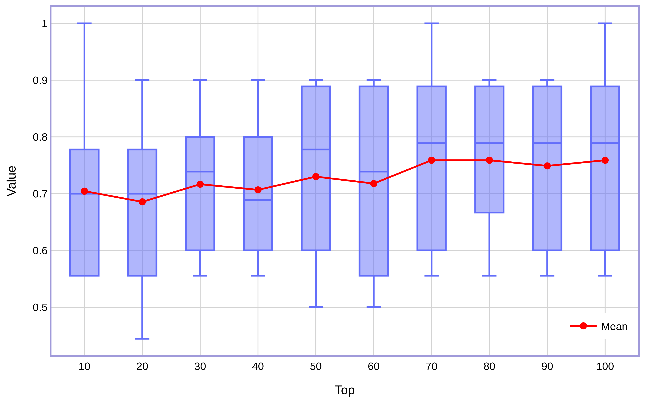
Khi áp dụng Random forest, kết quả có phần thấp hơn khi dùng Lime. Tuy nhiên, ở các tập T2D, OBE, IBD vẫn có sự gia tăng đáng kể về hiệu suất.



1. CIR (b) COL



(c) IBD (d) OBE



(e) T2D (f) WT2D

Cụ thể ở tập T2D trong hình A-b, cho kết quả tổng quan có xu hướng tăng nhưng không có sự tăng mạnh như Lime. Giá trị ACC tăng nhiều nhất từ top 10 đến 30 và sau đó không có nhiều thay đổi quá lớn cho tới top 100. Độ chính xác đạt cao nhất tại top 30 với giá trị 0.723. Về độ phân tán, hầu hết ở các top cho thấy độ phân tán khá rộng. Mặc dù giá trị trung bình không thay đổi nhiều, nhưng mức độ phân tán của dữ liệu vẫn lớn. Tiếp theo là tập OBE được thể hiện trong hình B-b cho thấy sự dao động không ổn định. Từ Top 10 đến 30, giá trị trung bình có xu hướng tăng, nhưng sau đó lại giảm nhẹ ở các mức Top 50 đến 70 và tiếp tục dao động đến Top 100. Đạt kết quả cao nhất ở top 50 với ACC là 0.807. Điều đó cho thấy mô hình Random Forest có thể không đảm bảo được tính ổn định khi số lượng đặc trưng thay đổi. Ngoài ra, độ phân tán của nó cao hơn đáng kể so với Lime. Đặc biệt, ở các mức Top 20, 30, 40, độ phân tán đạt mức cao nhất, cho thấy các giá trị dự đoán có sự phân tán lớn hơn. Đáng chú ý hơn là các outlier trong mô hình tương đối nhiều, cho thấy mô hình có thể bị ảnh hưởng nhiều bởi các giá trị ngoại lai hoặc các điểm dữ liệu nhiễu. Kế tiếp là tập IBD, biểu đồ của cho thấy một xu hướng khá ổn định. Giá trị trung bình tăng dần từ Top 10 đến Top 50, đạt mức cao nhất là 0.891 tại Top 50, sau đó giữ ổn định với sự giảm nhẹ khi tới top 100, chứng tỏ Random Forest có thể xử lý tốt dữ liệu ngay cả khi số lượng đặc trưng tăng lên. Tuy nhiên, độ phân tán của RF lại vô cùng lớn khoảng top 10 đến top 50. Từ Top 60 trở đi, phương sai thu hẹp lại đáng kể, cho thấy mô hình dần ổn định hơn khi có nhiều đặc trưng hơn.

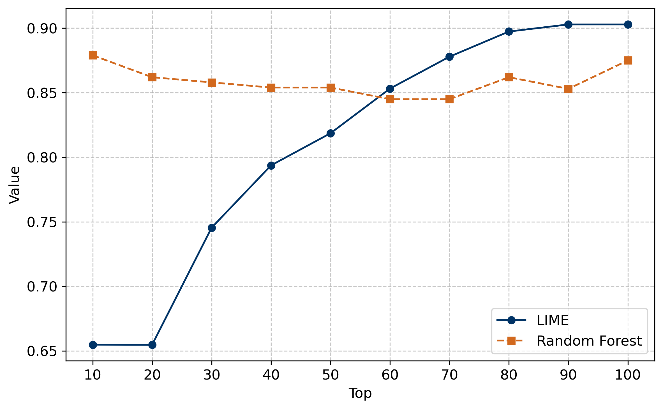
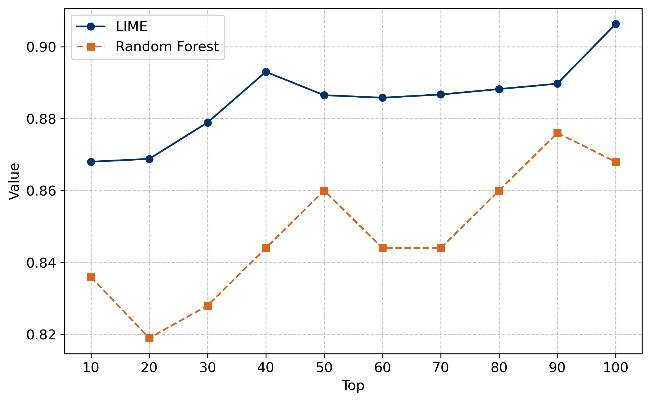
Trong khi đó, ở các tập WT2D, COL, CIR không có sự gia tăng đáng kể. Với tập WT2D trong hình A, giá trị trung bình tăng nhẹ từ 0.704 (Top 10) lên khoảng 0.759 (Top 100). Đạt kết quả cao nhất ở các top 70, 80 và 100 với kết quả acc trung bình là 0.759. Về độ phân tán, ở tất cả các top độ phân tán tương đối lớn, không thu hẹp nhiều khi tăng số lượng đặc trưng. Nhìn chung, khi dùng RF hiệu suất không thay đổi rõ rệt khi số lượng đặc trưng tăng. Tiếp theo là tập COL, giá trị trung bình ban đầu có xu hướng giảm nhẹ từ Top 10 đến Top 20, sau đó tăng dần đến Top 50, rồi tiếp tục dao động đến Top 100 được trình bày trong hình. Kết quả cao nhất đạt 0.876 tại top 90. Ngoài ra, độ phân tán của Random Forest tương đối lớn. Với các giá trị dự đoán có sự phân tán lớn, điều đó làm cho mô hình không ổn định. Ngoài ra, số lượng outlier khá nhiều, gần như xuất hiện ở hầu hết các mức cho thấy mô hình Random Forest có thể bị ảnh hưởng nhiều bởi các giá trị ngoại lai hoặc dữ liệu nhiễu, làm giảm tính ổn định của dự đoán. Và cuối cùng là tập CIR được hiển thị trong hình A, đường trung bình có xu hướng giảm nhẹ từ Top 10 (0.879)- top có ACC cao nhất xuống mức thấp hơn (0.845) trước khi tăng nhẹ ở Top 100. Độ phân tán luôn lớn ở tất cả các mức Top. Các outliers rải rác nhiều hơn, cho thấy mô hình có thể bị ảnh hưởng bởi một yếu tố bất kỳ nào đó hoặc không phản ứng tốt khi thay đổi số lượng đặc trưng.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | Accuracy | AUC | Precision | F1-score |
| CIR | 0.8587 ±0.0107 | 0.9063 ±0.0065 | 0.8855±0.0125 | 0.8564±0.0110 |
| COL | 0.8479 ±0.0170 | 0.9085 ±0.0243 | 0.8306±0.0191 | 0.8058±0.0250 |
| IBD | 0.8710 ±0.0169 | 0.9298 ±0.0094 | 0.7316±0.0696 | 0.5814±0.0405 |
| OBE | 0.7795 ±0.0249 | 0.8150 ±0.0496 | 0.7962±0.0183 | 0.8399±0.0186 |
| T2D | 0.7020 ±0.0235 | 0.7755 ±0.0239 | 0.7100±0.0268 | 0.6920±0.0254 |
| WT2D | 0.7288 ±0.0251 | 0.8253 ±0.0163 | 0.7522±0.0257 | 0.7611±0.0214 |

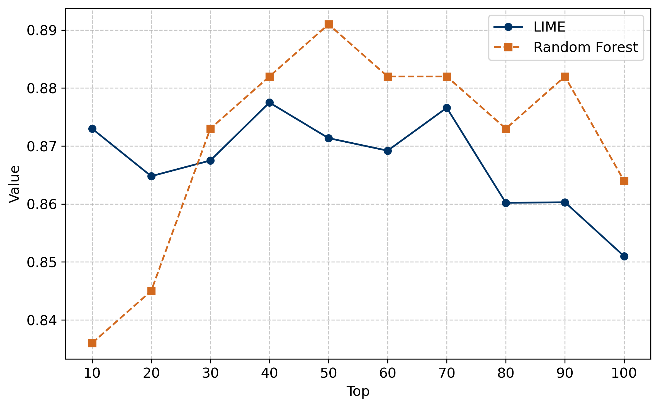
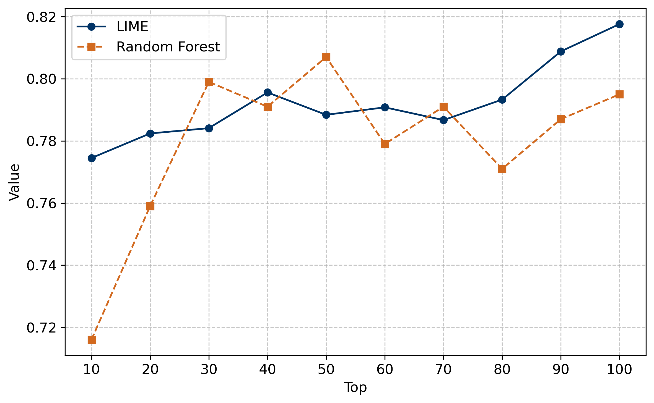
Bảng 3 trình bày kết quả mô hình khi sử dụng đặc trưng được chọn lọc bằng phương pháp RF. Tổng quan, mô hình đạt hiệu suất khá ổn định trên các tập dữ liệu, với độ chính xác (Accuracy) dao động từ 0.7020 (T2D) đến 0.8710 (IBD). AUC cho thấy khả năng phân biệt lớp tốt, đặc biệt trên IBD (0.9298) và COL (0.9085). Precision và F1-score cũng duy trì ở mức cao, với CIR đạt F1-score cao nhất (0.8564) và IBD thấp nhất (0.5814), cho thấy sự mất cân bằng trong phân loại. So với kết quả từ Lime (Bảng 2), RF đạt F1-score cao hơn trên một số tập dữ liệu như CIR và OBE, nhưng kém hơn trên T2D và WT2D. Điều này cho thấy Lime có thể giúp cải thiện hiệu suất trên một số tập dữ liệu nhất định, đặc biệt là với các trường hợp khó phân loại.

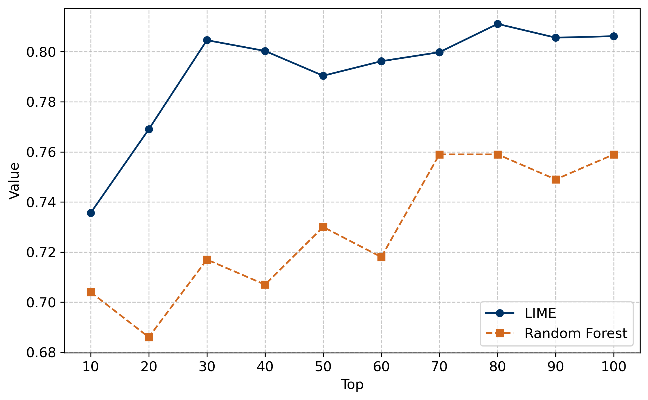


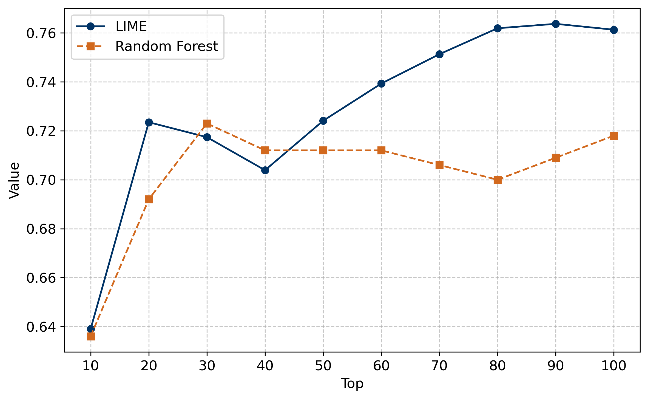
### So sánh



1. CIR (b) COL



 (c) IBD (d) OBE



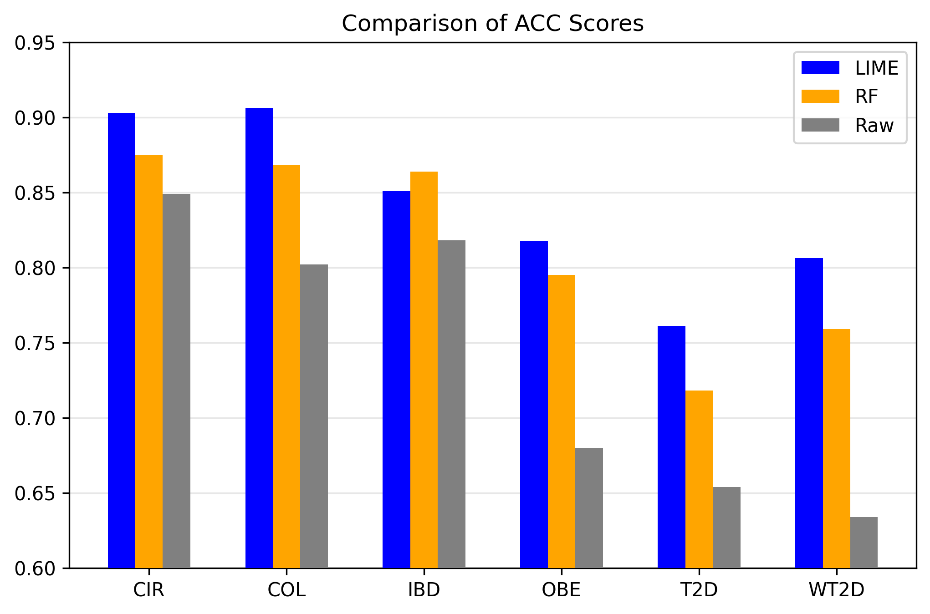
(e) T2D (f) WT2D

Từ hình A có thể thấy được phương pháp Lime đạt kết quả tương đối cao hơn khi dùng RF. Với tập dữ liệu CIR, mặc dù từ top 10 đến top 30, phương pháp RF hoạt động hiệu quả hơn LIME. Tuy nhiên, khi tăng dần số lượng đặc trưng, LIME bắt đầu phát huy hiệu quả, có khả năng tận dụng thông tin từ nhiều đặc trưng hơn để cải thiện hiệu suất, và cuối cùng vượt trội hơn RF. Tiếp theo là tập COL và tâp WT2D, cả hai đều có đặc điểm chung là hai đường xanh và cam nằm cách biệt nhau khá rõ ràng và không có điểm nào giao nhau. Phương pháp Lime duy trì mức Valid Accuracy cao hơn nhiều so với RF. Trong khi RF có một số điểm tụt giảm thất thường, cho thấy phương pháp này không ổn định trên cả 2 tập dữ liệu. Với tập IBD, đây là trường hợp khác biệt khi mà RF phần lớn có các điểm đạt Valid Accuracy cao hơn Lime. Quan sát hình cho thấy, tuy RF và Lime có sự dao động tương đối tương đồng, nhưng tổng quan kết quả RF lại tốt hơn Lime. Trong biểu đồ OBE, LIME luôn có độ chính xác cao hơn RF, đặc biệt khi số lượng đặc trưng tăng dần. RF cải thiện nhanh chóng trong giai đoạn đầu, nhưng sau đó dao động khá mạnh. Ngược lại, LIME duy trì sự ổn định và tiếp tục cải thiện hiệu suất khi số đặc trưng lớn hơn. Điều này cho thấy LIME có khả năng tận dụng thông tin từ nhiều đặc trưng hơn so với RF. Và cuối cùng là tập dữ liệu T2D, RF có sự tăng trưởng nhanh ban đầu nhưng càng về sau hiệu suất gần như không thay đổi. Trong khi đó, LIME tiếp tục cải thiện và đạt độ chính xác cao hơn RF khi số lượng đặc trưng lớn hơn. Điều đó cho thấy càng về sau các đặc trưng dần có đóng góp nhiều hơn.

Trong tất cả các bộ dữ liệu, độ chính xác (ACC) của phương pháp LIME thường tăng khi số lượng đặc trưng quan trọng tăng lên. Tuy nhiên, độ ổn định của ACC lại không đồng nhất giữa các bộ dữ liệu. Trong khi RF cũng có xu hướng tăng nhưng hiệu suất tổng thể lại không cao hơn Lime

### Thảo luận kết quả và so sánh với các nghiên cứu trước

Trong phần này, em so sánh độ chính xác của các phương pháp giảm chiều đã đề xuất khi giữ lại 100 đặc trưng quan trọng nhất trong bộ dữ liệu ban đầu. Số lượng đặc trưng này chỉ chiếm khoảng 0.006% so với toàn bộ dữ liệu gốc. Từ bảng 2 và hình 5, cho thấy được phương pháp giảm chiều bằng Lime hầu như đạt giá trị cao nhất trên ACC với mô hình RF.



|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | Accuracy | AUC | Precision | F1-score |
| CIR | 0.9029 ±0.0065 | 0.8750 ±0.0689 | 0.8960±0.0443 | 0.8490±0.0671 |
| COL | 0.9063 ±0.0173 | 0.8660 ±0.0661 | 0.8760±0.0677 | 0.8020±0.0919 |
| IBD | 0.8510 ±0.0076 | 0.8640 ±0.0932 | 0.8550±0.0728 | 0.8180±0.0576 |
| OBE | 0.8176 ±0.0070 | 0.7950 ±0.0594 | 0.8020±0.0710 | 0.6800±0.0592 |
| T2D | 0.7613 ±0.0110 | 0.7180 ±0.0601 | 0.7240±0.0452 | 0.6540±0.0651 |
| WT2D | 0.8062 ±0.0169 | 0.7590 ±0.0411 | 0.7690±0.1327 | 0.6340±0.1398 |

Với hiệu suất cao nhất đến 0.9063 trong tập dữ liệu COL. So sánh với phương pháp RF, mặc dù RF cho kết quả tốt nhất trên tập IBD (0.891), nhưng LIME vẫn chiếm ưu thế trên các tập còn lại, đặc biệt là WT2D và T2D. Với 2 tập trên, phương pháp LIME mang lại sự cải thiện đáng kể, với độ chính xác tăng từ 0.654 lên 0.7285 cho T2D và từ 0.634 lên 0.7919 cho WT2D. Ngoài ra, trên tập dữ liệu OBE, LIME đạt độ chính xác 0.7922, cao hơn đáng kể so với RF (0.7795) và dữ liệu gốc (0.680). Điều này cho thấy, ngay cả với các tập dữ liệu có độ phức tạp cao, phương pháp LIME vẫn có thể duy trì và cải thiện độ chính xác so với RF thông thường. So với tập gốc, cả hai phương pháp đều cho ra kết quả cao hơn đáng kể. Tuy nhiên, khi so sánh với tập 1000 đặc trưng, Lime vẫn duy trì hiệu suất cao hơn trên hầu hết các tập, ngoại trừ tập IBD thấp hơn 1 xíu. RF thì ngược lại, chỉ có tập IBD cho kết quả cao hơn, còn các tập khác đều cho kết quả thấp hơn.

Kết quả so sánh cho thấy RF hoạt động ổn định và có xu hướng cải thiện độ chính xác so với dữ liệu gốc, nhưng khi áp dụng LIME, độ chính xác tiếp tục được nâng cao. Điều này cho thấy LIME không chỉ giúp giảm chiều dữ liệu mà còn có khả năng giữ lại những đặc trưng quan trọng, từ đó cải thiện hiệu suất mô hình. Hơn nữa, biểu đồ trong Hình 5 minh họa rõ ràng sự khác biệt này. Các cột màu xanh (LIME) phần lớn có xu hướng cao hơn so với các cột màu cam (RF) và xám (Raw), đặc biệt trên các tập dữ liệu COL, OBE, T2D và WT2D. Từ đó có thể chứng minh, việc sử dụng LIME không chỉ mang lại hiệu suất cao hơn trên từng tập dữ liệu riêng lẻ mà còn giúp tổng thể mô hình hoạt động tốt hơn trên nhiều loại dữ liệu khác nhau.

Bảng 4 so sánh các giá trị ACC cao nhất trong các top đặc trưng với kết quả từ các nghiên cứu trước. Kết quả cho thấy hiệu suất của nghiên cứu này đã được cải thiện đáng kể.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | Accuracy | AUC | Precision | F1-score |
| CIR | 0.9029 ±0.0065 | 0.8750 ±0.0689 | 0.8960±0.0443 | 0.8490±0.0671 |
| COL | 0.9063 ±0.0173 | 0.8660 ±0.0661 | 0.8760±0.0677 | 0.8020±0.0919 |
| IBD | 0.8510 ±0.0076 | 0.8640 ±0.0932 | 0.8550±0.0728 | 0.8180±0.0576 |
| OBE | 0.8176 ±0.0070 | 0.7950 ±0.0594 | 0.8020±0.0710 | 0.6800±0.0592 |
| T2D | 0.7613 ±0.0110 | 0.7180 ±0.0601 | 0.7240±0.0452 | 0.6540±0.0651 |
| WT2D | 0.8062 ±0.0169 | 0.7590 ±0.0411 | 0.7690±0.1327 | 0.6340±0.1398 |

# 

# CHƯƠNG 5. KẾT LUẬN

Trong nghiên cứu này, em đã khảo sát ảnh hưởng của phương pháp giảm chiều dữ liệu bằng Lime đối với độ chính xác của mô hình Random Forest (RF) khi áp dụng trên dữ liệu gene family. Kết quả thực nghiệm cho thấy phương pháp lựa chọn đặc trưng dựa trên Lime đạt hiệu suất cạnh tranh so với phương pháp lựa chọn đặc trưng truyền thống của RF, đặc biệt là trên các tập dữ liệu CIR, T2D, COL, OBE và WT2D, nơi Lime đạt độ chính xác vượt trội hơn. Cụ thể, đạt độ chính xác cao nhất trên tập COL là 0.9063. Đáng chú ý, Lime cho kết quả tốt hơn so với dữ liệu gốc trong hầu hết tất cả các trường hợp, cho thấy khả năng giữ lại các tính năng quan trọng của phương pháp này trong khi đã lựa chọn đặc trưng.

Kết quả này cho thấy Lime là một công cụ hữu ích giúp giảm số lượng đặc trưng mà vẫn giữ được độ chính xác cao của mô hình. Tuy nhiên, Lime đòi hỏi thời gian tính toán khá cao, vì thế cần được xem xét khi áp dụng trên các tập dữ liệu lớn hơn. Trong các ứng dụng thực tế, việc kết hợp Lime với các phương pháp chọn lọc đặc trưng khác có thể giúp tối ưu hóa cả về hiệu suất lẫn tốc độ xử lý, từ đó mở ra hướng nghiên cứu mới cho các mô hình phân tích dữ liệu lớn và học máy trong lĩnh vực sinh học.



## Kết quả đạt được

- Bài nghiên cứu đã xây dựng một hệ thống dự đoán hiệu suất học tập của sinh viên dựa trên các thuật toán máy học truyền thống như SVM, Random Forest, Naive Bayes, Decision Tree, và Logistic Regression.

- Việc trực quan hóa các yếu tố quan trọng thông qua Gain Ratio đã giúp xác định rõ các yếu tố như ngày vắng mặt và mức độ tương tác học tập, … làm cơ sở cho các phân tích sâu hơn.

- Việc sử dụng GridSearchCV đã cải thiện đáng kể hiệu suất mô hình, đặc biệt với các thuật toán SVM và Random Forest, giúp tăng độ chính xác và F1-score lên đến mức tối ưu.

- Các yếu tố quan trọng ảnh hưởng đến kết quả học tập như số ngày vắng mặt, số lần truy cập khóa học, và khảo sát phụ huynh đã được xác định và phân tích.

## Hạn chế

* Tập dữ liệu được sử dụng quy mô vẫn còn nhỏ, chưa đa dạng.
* Kết quả của các mô hình sau khi đào tạo vẫn còn thấp.
* Nghiên cứu chủ yếu thực hiện trên các mô hình máy học truyền thống và chưa mở rộng ra với các mô hình lớn hơn.

## Hướng phát triển

* Thu thập thêm nhiều nguồn dữ liệu.
* Mở rộng dữ liệu: Tích hợp thêm các bộ dữ liệu từ nhiều nguồn khác nhau để tăng tính tổng quát của mô hình.
* Cải thiện mô hình: Nghiên cứu và áp dụng các thuật toán tiên tiến hơn để tăng cường độ chính xác.
* Ứng dụng thực tế: Phát triển hệ thống thành công cụ hỗ trợ trực tuyến cho các trường đại học nhằm giám sát và can thiệp kịp thời với các sinh viên có nguy cơ gặp khó khăn trong học tập.

**TÀI LIỆU THAM KHẢO**

[1] J. Chen *et al.*, “Sedentary lifestyle, physical activity, and gastrointestinal diseases: evidence from mendelian randomization analysis,” *eBioMedicine*, vol. 103, p. 105110, May 2024, doi: 10.1016/j.ebiom.2024.105110.

[2] Dabhade, P., Agarwal, R., Alameen, K., Fathima, A., Sridharan, R., Gopakumar,

G.: Educational data mining for predicting students’ academic performance using

machine learning algorithms. Materials Today: Proceedings 47, 5260–5267 (2021),

http://dx.doi.org/10.1016/j.matpr.2021.05.646

[3] Bhutto, E.S., Siddiqui, I.F., Arain, Q.A., Anwar, M.: Predicting students’ academic

performance through supervised machine learning. In: 2020 International Conference

on Information Science and Communication Technology (ICISCT). IEEE (Feb2020),

http://dx.doi.org/10.1109/ICISCT49550.2020.9080033

[4] Ya˘gcı, M.: Educational data mining: prediction of students’ academic performance

using machine learning algorithms. Smart Learning Environments 9(1) (Mar 2022),

http://dx.doi.org/10.1186/s40561-022-00192-z

[5] Ojajuni, O., Ayeni, F., Akodu, O., Ekanoye, F., Adewole, S., Ayo, T., Misra, S., Mbarika, V.: Predicting Student Academic Performance Using Machine Learning, p. 481–491. Springer International Publishing (2021), http://dx.doi.org/10.1007/978-3-030-87013-3 36

[6] Hussain, S., Khan, M.Q.: Student-performulator: Predicting students’ academic performance at secondary and intermediate level using machine learning. Annals of Data Science 10(3), 637–655 (Jun 2021), http://dx.doi.org/10.1007/s40745-021-00341-0

[7] Sarwat, S., Ullah, N., Sadiq, S., Saleem, R., Umer, M., Eshmawi, A.A., Mohamed, A., Ashraf, I.: Predicting students’ academic performance with conditional generative adversarial network and deep svm. Sensors 22(13), 4834 (Jun 2022), http://dx.doi.org/10.3390/s22134834

[8] Jayaprakash, S., Krishnan, S., Jaiganesh, V.: Predicting students academic perfor-

mance using an improved random forest classifier. In: 2020 International Confer-

ence on Emerging Smart Computing and Informatics (ESCI). IEEE (Mar 2020),

http://dx.doi.org/10.1109/ESCI48226.2020.9167547

[9] Almutairi, S., Shaiba, H., Bezbradica, M.: Predicting Students’ Academic Performance and Main Behavioral Features Using Data Mining Techniques, p. 245–259. Springer International Publishing (2019), http://dx.doi.org/10.1007/978-3-030-36365-9 21

[10] Rajendran, S., Chamundeswari, S., Sinha, A.A.: Predicting the academic performance of middle- and high-school students using machine learning algorithms. Social Sciences and amp; Humanities Open 6(1), 100357 (2022), http://dx.doi.org/10.1016/j.ssaho.2022.100357

[11] Alsharari, F., Teodorescu, D.: Stress factors impacting the academic performance of saudi students studying in the us: A study of student perceptions. International Journal of Social Science and Business 4(1), 13–26 (2019)

[12] Yassein, N.A., M Helali, R.G., Mohomad, S.B.: Predicting student academic performance in ksa using data mining techniques. Journal of Information Technology and amp; Software Engineering 07(05) (2017), http://dx.doi.org/10.4172/2165-7866.1000213

[13] Contreras, D., Delgadillo, J., Riveros, G.: Is home overcrowding a significant factor in children’s academic performance? evidence from latin america. International Journal of Educational Development 67, 1–17 (May 2019), http://dx.doi.org/10.1016/j.ijedudev.2019.01.006

[14] Ding, W., Lehrer, S.F., Rosenquist, J., Audrain-McGovern, J.: The impact of poor health onacademic performance: New evidence using genetic markers. Journal of Health Economics

28(3), 578–597 (May 2009), http://dx.doi.org/10.1016/j.jhealeco.2008.11.006

[15] Krukowski, R.A., Smith West, D., Philyaw Perez, A., Bursac, Z., Phillips, M.M., Raczyn-ski, J.M.: Overweight children, weight-based teasing and academic performance. International

Journal of Pediatric Obesity 4(4), 274–280 (2009)

[16] Nayak, J.K.: Relationship among smartphone usage, addiction, academic performance and the moderating role of gender: A study of higher education students in india. Computers and amp; Education 123, 164–173 (Aug 2018), http://dx.doi.org/10.1016/j.compedu.2018.05.007

[17] Sabia, J.J.: The effect of body weight on adolescent academic performance. South-

ern Economic Journal 73(4), 871–900 (Apr 2007), http://dx.doi.org/10.1002/j.2325-

8012.2007.tb00809.x

[18] Johnson, S.R., Stage, F.K.: Academic engagement and student success: Do high- impact practices mean higher graduation rates? The Journal of Higher Education 89(5), 753–781 (Apr2018), http://dx.doi.org/10.1080/00221546.2018.1441107

[19] Alnasyan, B., Basheri, M., Alassafi, M.: The power of deep learning techniques for

predicting student performance in virtual learning environments: A systematic litera-

ture review. Computers and Education: Artificial Intelligence 6, 100231 (Jun 2024),

http://dx.doi.org/10.1016/j.caeai.2024.100231

[20] Zhang, X., Xue, R., Liu, B., Lu, W., Zhang, Y.: Grade prediction of student academic performance with multiple classification models. In: 2018 14th International Conference on Natural Computation, Fuzzy Systems and Knowledge Discovery (ICNC-FSKD). IEEE (Jul 2018), http://dx.doi.org/10.1109/FSKD.2018.8687286

[21] Pandey, M., Kumar Sharma, V.: A decision tree algorithm pertaining to the student performance analysis and prediction. International Journal of Computer Applications 61(13), 1–5(Jan 2013), http://dx.doi.org/10.5120/9985-4822

[22] Ren, Z., Ning, X., Rangwala, H.: Grade prediction with temporal course-wise influence (2017), https://arxiv.org/abs/1709.05433

[23] Namoun, A., Alshanqiti, A.: Predicting student performance using data mining and learning analytics techniques: A systematic literature review. Applied Sciences 11(1), 237 (Dec 2020), http://dx.doi.org/10.3390/app11010237

[24] Chen, W., Brinton, C.G., Cao, D., Mason-Singh, A., Lu, C., Chiang, M.: Early

detection prediction of learning outcomes in online short-courses via learning be-

haviors. IEEE Transactions on Learning Technologies 12(1), 44–58 (Jan 2019),

http://dx.doi.org/10.1109/TLT.2018.2793193

[25] Xu, X., Wang, J., Peng, H., Wu, R.: Prediction of academic performance associated with in-ternet usage behaviors using machine learning algorithms. Computers in Human Behavior 98, 166–173 (Sep 2019), http://dx.doi.org/10.1016/j.chb.2019.04.015

[26] Baashar, Y., Alkawsi, G., Mustafa, A., Alkahtani, A.A., Alsariera, Y.A., Ali, A.Q.,

Hashim, W., Tiong, S.K.: Toward predicting student’s academic performance us-

ing artificial neural networks (anns). Applied Sciences 12(3), 1289 (Jan 2022),

http://dx.doi.org/10.3390/app12031289

[27] Hasan, H.R., Rabby, A.S.A., Islam, M.T., Hossain, S.A.: Machine learning algorithm

for student’s performance prediction. In: 2019 10th International Conference on Com-

puting, Communication and Networking Technologies (ICCCNT). IEEE (Jul 2019),

http://dx.doi.org/10.1109/ICCCNT45670.2019.8944629

[28] Louppe, G.: Understanding random forests: From theory to practice (2014),

https://arxiv.org/abs/1407.7502

[29] He, Z., Lin, D., Lau, T., Wu, M.: Gradient boosting machine: A survey (2019),

https://arxiv.org/abs/1908.06951

[30] Izza, Y., Ignatiev, A., Marques-Silva, J.: On explaining decision trees (2020),

https://arxiv.org/abs/2010.11034

[31] Korosi, G., Esztelecki, P., Farkas, R., Toth, K.: Clickstream-based outcome prediction in short video moocs. In: 2018 International Conference on Computer, Information and Telecommunication Systems (CITS). IEEE (Jul 2018), http://dx.doi.org/10.1109/CITS.2018.8440182

[32] Al-Shabandar, R., Hussain, A., Laws, A., Keight, R., Lunn, J., Radi, N.: Machine

learning approaches to predict learning outcomes in massive open online courses. In:

2017 International Joint Conference on Neural Networks (IJCNN). IEEE (May 2017),

http://dx.doi.org/10.1109/IJCNN.2017.7965922

[33] Liu, Y., Hui, Y., Hou, D., Liu, X.: A novel student achievement prediction method

based on deep learning and attention mechanism. IEEE Access 11, 87245–87255 (2023), http://dx.doi.org/10.1109/ACCESS.2023.3305248

[34] Alqatow, I., Rattrout, A., Jayousi, R.: Prediction of Student Performance with Machine Learning Algorithms Based on Ensemble Learning Methods, p. 520–529. Springer Nature Singapore (2023), http://dx.doi.org/10.1007/978-981-99-7254-8 40

[35] Nabil, A., Seyam, M., Abou-Elfetouh, A.: Deep neural networks for predicting students’ performance. In: Proceedings of the 52nd ACM Technical Symposium on Computer Science Education. SIGCSE ’21, ACM (Mar 2021), http://dx.doi.org/10.1145/3408877.3439685

[36] Amrieh, E.A., Hamtini, T., Aljarah, I.: Mining educational data to predict student’s academic performance using ensemble methods. International Journal of Database Theory and Application 9(8), 119–136 (Aug 2016), http://dx.doi.org/10.14257/ijdta.2016.9.8.13