# **LINEAR REGRESSION**

# TITLE & ABSTRACT

## What question were the investigators asking?

Nhóm tác giả dự đoán kết quả học tập của sinh viên bằng giải thuật Hồi quy tuyến tính.

## What techniques did they use?

Họ triển khai các ứng dụng mới của máy học trong dạy và học dựa trên nền tảng của sinh viên, điểm số học tập trong quá khứ của sinh viên và xem xét các thuộc tính khác.

Ngoài ra, họ xác định triển khai giải thuật Hồi quy tuyến tính (Linear Regression) để dự đoán kết quả học tập của sinh viên.

## What were their major results?

Kết quả chính của họ là những dự đoán kết quả học tập của sinh viên có ảnh hưởng bởi những thuộc tính điều kiện nào.

## What conclusions did they draw?

Khi quy mô lớp học càng lớn, giáo viên sẽ khó hỗ trợ từng học viên trong mỗi khóa học mở, điều này có thể làm tăng tỷ lệ bỏ học khi kết thúc khóa học. Do đó, việc sử dụng và triển khai các ứng dụng Học máy (cụ thể là Hồi quy tuyến tính – Linear Regression) sẽ giúp dự đoán được kết quả học tập của sinh viên. Từ đó, nhóm nghiên cứu có thể giúp cho giáo viên dễ dàng hỗ trợ cho sinh viên tốt hơn.

# INTRODUCTION

## Why did they start?

Nhóm tác giả cho rằng: “Khi việc sử dụng máy tính và internet ở khắp mọi nơi, sự sẵn có của dữ liệu có thể được phân tích tăng lên nhanh chóng. Dữ liệu có thể là bất cứ thứ gì liên quan đến dân số, dữ liệu học tập của sinh viên và sở thích của mọi người. Chúng ta có thể thấy rằng dữ liệu mới xuất hiện theo thời gian. Phân tích dữ liệu là nhiệm vụ khó khăn đối với con người.”

Ngoài ra, họ cũng cho rằng: “…máy tính có thể phân tích dữ liệu hiệu quả hơn con người vì nó lưu trữ dữ liệu kỹ thuật số theo cách được định dạng tốt.”

Mặt khác, họ e sợ về việc giáo viên sẽ khó hỗ trợ sinh viên nếu quy mô lớp hơn ngày càng lớn.

Do đó, họ đã thực hiện bài nghiên cứu này vừa để chứng minh nhận định trên, vừa đưa ra dự đoán học tập của sinh viên để cải thiện chất lượng giáo dục tốt hơn.

## Why is this research important?

Dự đoán kết quả học tập của sinh viên là rất quan trọng vì nó có thể cho giáo viên biết về những sinh viên có khả năng bỏ học và dự đoán có thể hỗ trợ thêm cho những sinh viên cần cải thiện kết quả học tập.

## Is it justified (review of literature?)

Nhóm nghiên cứu đã khẳng định:

* “Học máy là một nhánh của Trí tuệ nhân tạo cung cấp khả năng tự động học hỏi từ những kinh nghiệm trong quá khứ.”
* “Các máy được lập trình rõ ràng.”
* “Nó mang lại khả năng cho máy tính khiến con người và máy móc trông giống nhau ở khía cạnh học tập.”

## What is the research question/s of interest (purpose/objectives) or hypothesis?

Trong bài báo nghiên cứu này, nhóm tập giả tập trung vào cách thực hiện hồi quy tuyến tính đối với kết quả học tập của sinh viên khi xem xét tập dữ liệu của sinh viên.

## Terms defined?

Thuật ngữ được xác định trong bài báo là "Linear Regression", nghĩa là "Hồi quy tuyến tính".

# METHOD

## What did they do?

Nhóm tác giả đã thực hiện các bước như sau:

* Thu thập bộ dữ liệu cần thiết: Đây là bước mà họ xem là quan trong nhất cho công việc nghiên cứu.
* Tiền xử lý dữ liệu: Bước này được thực hiện sau khi thu thập dữ liệu, dữ liệu sẽ được chuyển thành dạng mà họ mong muốn.
* Làm sạch dữ liệu: Tìm ra những dữ liệu không đầy đủ, không liên quan trong tập dữ liệu và loại bỏ chúng để thu được kết quả chính xác của công việc.
* Phân loại dữ liệu: Có thể chọn bất kỳ một trong các thuật toán như Hồi quy tuyến tính, Máy vectơ hỗ trợ, Phân loại tiêu chuẩn Naive Bayes, thuật toán Cây quyết định… nhưng nhóm nghiên cứu đã xác định chọn thuật toán Hồi quy tuyến tính.
* Ngoài ra, họ cũng thực hiện Trực quan hóa dữ liệu (Data Visualization) và biểu diễn trên đồ thị Boxplot.

## Flow execution

Nhóm nghiên cứu đã thực hiện theo mô hình như sau:

Diagram

Description automatically generated

# RESULTS

## What did they find?

Họ chia tập dữ liệu của mình thành hai phần là tập dữ liệu huấn luyện và kiểm tra. Sau đó, chúng ta sẽ áp dụng hàm lm() trên dữ liệu “huấn luyện” và hàm predict() trên dữ liệu “kiểm tra”, đồng thời tạo trực quan hóa mô hình hồi quy của chúng ta với đường hồi quy và khoảng tin cậy 95%:

Dưới đây là **bảng Giá trị của các biến**:

Table

Description automatically generated

**Bảng Giá trị phù hợp của dữ liệu:**

**Table

Description automatically generated**

Họ **trực quan hóa điểm dự đoán:**

**Chart, scatter chart

Description automatically generated**

Ngoài ra, ở phần Trực quan hóa dữ liệu, họ xác định mục đính chính là **tìm các mẫu trực quan**. Họ dựng biểu đồ điểm học tập so với giới tính, tuổi tác, tình trạng phụ huynh bằng cách sử dụng gói GGPLOT. GGPLOT yêu cầu ba thành phần chính:

• Định nghĩa dữ liệu ở dạng khung dữ liệu.

• Mô tả tính thẩm mỹ cho hình ảnh hoặc cách ánh xạ các thuộc tính.

• Xác định hình học hoặc loại đồ họa sẽ được sử dụng.

Dưới đây là biểu đồ **Trực quan hóa dữ liệu giữa Điểm học tập và giới tính của sinh viên:**

Chart, bar chart

Description automatically generated

Họ quan sát tỷ lệ phần trăm điểm học tập liên quan đến giới tính của sinh viên, và dưới đây là biểu đồ **Trực quan hóa dữ liệu giữa Điểm học tập và tuổi của sinh viên:**

Chart, histogram

Description automatically generated

Với phần biểu diễn trên đồ thị Boxplot, họ thực hiện 2 biểu đồ ứng với 2 Trực quan hóa boxplot về:

* Học sinh hoàn thành lớp chuẩn bị (pre-class) có điểm học tập tốt hơn.
* Học sinh nữ đạt điểm cao hơn học sinh nam.

Dưới đây là 2 biểu diễn Boxplot mà họ thực hiện:

Chart, box and whisker chart

Description automatically generated

Chart, histogram

Description automatically generated

## Were the number of subjects in each group or subgroup used in the analysis specified?

Có, số lượng đối tượng trong mỗi nhóm hoặc phân nhóm được sử dụng trong phân tích được chỉ định, tham số là sinh viên.

## Did the results relate to the specified objective/hypotheses?

Các kết quả thu được (chủ yếu là các kết quả từ phần Trực quan hóa dữ liệu) có liên quan đến mục tiêu ban đầu là dự đoán kết quả học tập của sinh viên.

## Do the tables and figures “speak for themselves?”

Không hẳn, các bảng và số liệu được thể hiện trong bài báo chỉ nếu lên 1 phần hoặc chỉ khảo sát về dữ liệu chứ không tự đưa ra được kết luận cho mục tiêu ban đầu.

## Are the tables adequately titled, labeled?

Các bảng có đủ tiêu đề và được dán nhãn.

# DISCUSSION

## What do the results mean?

Các nghiên cứu hiện tại cho thấy rằng kết quả học tập của học sinh cũng phụ thuộc vào nền tảng của học sinh và các thuộc tính khác. Nhiều công trình nghiên cứu xác nhận rằng ngoài thành tích học tập trong quá khứ, nền tảng của học sinh và các thuộc tính khác thực sự có ảnh hưởng đáng kể đến kết quả học tập của học sinh.

## Did authors discuss the results in relations to the objectives/hypotheses?

Các tác giả có thảo luận về kết quả trong mối quan hệ với các mục tiêu nêu lên ban đầu.

## Were the results discussed in relation to those from similar studies?

Các kết quả được thảo luận liên quan đến những kết quả từ các nghiên cứu tương tự. Nhóm tác giả có đề cập và trình bày cụ thể tại mục II. LITERATURE REVIEW về những nghiên cứu liên quan khác mà họ có nhắc tới (cụ thể là 10 bài).

## Are the authors justified in the strength of the statements they make in the study?

Các tác giả có hợp lý về sức mạnh của các tuyên bố mà họ đưa ra trong nghiên cứu.

## Did they offer alternative explanations for results?

Nhóm nghiên cứu có đưa ra lời giải thích, nhưng là giải thích về sự hiểu quả và độ chính xác của nghiên cứu, không phải cho mục tiêu mà nhóm tác giả đã nêu từ đầu.

Hiệu quả của việc sử dụng máy học trong lĩnh vực giáo dục **phụ thuộc vào thuật toán và cách sử dụng dữ liệu**. Việc **lựa chọn thuật toán để thực hiện** dự đoán kết quả học tập của học **sinh là rất quan trọng**. Độ chính xác của kết quả phụ thuộc vào thuật toán Máy học.