

**BỘ GIÁO DỤC VÀ ĐÀO TẠO**

**TRƯỜNG ĐẠI HỌC MỎ - ĐỊA CHẤT**

**PHAN THANH TÚ (2221050845)**

**NGUYỄN ĐÌNH TRÁNG**

**TRẦN ĐĂNG QUANG**

TÊN ĐỀ TÀI:

DỰ ĐOÁN ĐIỂM HỌC SINH

**BÀI TẬP LỚN**

**MÔN: KỸ NGHỆ TRI THỨC VÀ HỌC MÁY**

**Hà Nội, Năm 2025**

**LỜI CẢM ƠN**

Trong quá trình thực hiện đồ án tốt nghiệp, chúng em đã nhận được nhiều sự giúp đỡ, đóng góp ý kiến và sự hướng dẫn tận tình từ thầy cô và bạn bè.

Chúng em xin gửi lời cảm ơn chân thành đến thầy Thắng và cô Dung – giảng viên Bộ môn Kỹ nghệ tri thức và học máy, Trường Đại học Mỏ Địa Chất – những người đã tận tình hướng dẫn, chỉ bảo chúng em trong suốt quá trình thực hiện khoá luận.

Bên cạnh đó, chúng em cũng xin bày tỏ lòng biết ơn sâu sắc đến các thầy cô trong Trường Đại học Mỏ Địa Chất, đặc biệt là các thầy cô trong Bộ môn Kỹ nghệ trí thức và học máy, những người đã truyền đạt cho chúng em những kiến thức đại cương và chuyên ngành quý báu, giúp chúng em có nền tảng lý thuyết vững vàng, đồng thời tạo điều kiện thuận lợi cho chúng em trong quá trình học tập và thực hiện bài tập lớn.

Cuối cùng, chúng em xin chân thành cảm ơn gia đình và bạn bè – những người luôn ở bên cạnh, luôn quan tâm, động viên và giúp đỡ chúng em trong suốt quá trình học tập và hoàn thành đồ án tốt nghiệp.

Do thời gian thực hiện có hạn và phạm vi nghiên cứu rộng, cùng với kinh nghiệm còn hạn chế, đồ án này khó tránh khỏi những thiếu sót. Chúng em rất mong nhận được sự góp ý, chỉ bảo từ thầy cô để có thể bổ sung, hoàn thiện kiến thức của mình, nhằm phục vụ tốt hơn cho công tác thực tế sau này.

Chúng em xin chân thành cảm ơn!

**MỤC LỤC**

1. **Giới thiệu đề tài**  
   1.1. Lý do nghiên cứu  
   1.2. Mục tiêu của dự án  
   1.3. Tầm quan trọng của dự án  
   1.4. Phạm vi nghiên cứu
2. **Tổng quan lý thuyết**  
   2.1. Hồi quy tuyến tính đa biến (Multiple Linear Regression)  
   2.2. Gradient Descent  
   2.3. Các thuật toán học máy khác
3. **Thiết kế hệ thống**  
   3.1. Kiến trúc hệ thống tổng quan  
   3.2. Sơ đồ hệ thống  
   3.3. Cấu trúc dữ liệu  
   3.4. Mô hình dữ liệu
4. **Quá trình huấn luyện mô hình**  
   4.1. Tiền xử lý dữ liệu  
   4.2. Phân chia dữ liệu  
   4.3. Huấn luyện mô hình hồi quy tuyến tính  
   4.4. Đánh giá mô hình
5. **Triển khai hệ thống**  
   5.1. Backend với Flask  
   5.2. Frontend với React  
   5.3. Kiểm thử hệ thống
6. **Đánh giá kết quả**  
   6.1. So sánh với giá trị trung bình  
   6.2. Phân tích lỗi mô hình  
   6.3. Đánh giá trải nghiệm người dùng
7. **Định hướng phát triển trong tương lai**  
   7.1. Cải thiện mô hình  
   7.2. Mở rộng ứng dụng  
   7.3. Tối ưu hóa hiệu suất
8. **Kết luận**
9. **Tài liệu tham khảo**
10. **Phụ lục**  
    10.1. Code nguồn  
    10.2. Kết quả huấn luyện mô hình  
    10.3. Ảnh chụp màn hình của ứng dụng

**1. Giới thiệu đề tài**

**1.1 Lý do nghiên cứu**

**Trong bối cảnh giáo dục ngày nay, việc dự đoán kết quả học tập của học sinh là một vấn đề quan trọng đối với không chỉ các giáo viên mà còn đối với phụ huynh và các nhà quản lý giáo dục. Một trong những vấn đề mà nhiều phụ huynh gặp phải là không thể biết trước kết quả học tập của con em mình, điều này khiến cho việc định hướng học tập và hỗ trợ học sinh gặp khó khăn. Vì vậy, việc có thể dự đoán điểm số của học sinh dựa trên các yếu tố đầu vào đơn giản sẽ giúp phụ huynh và giáo viên có được những thông tin quan trọng để đưa ra những quyết định phù hợp.**

**Dự án của chúng tôi nhằm phát triển một hệ thống có thể dự đoán điểm số học sinh dựa trên một số thuộc tính cơ bản như thời gian học, tình trạng gia đình, sức khỏe, và các điểm số trước đó. Hệ thống này không chỉ giúp phụ huynh có cái nhìn rõ ràng hơn về khả năng học tập của con em mình mà còn giúp các giáo viên đưa ra những biện pháp hỗ trợ kịp thời.**

**1.2 Mục tiêu của dự án**

**Mục tiêu chính của dự án là xây dựng một mô hình dự đoán điểm số học sinh, sử dụng thuật toán hồi quy tuyến tính đa biến. Với các dữ liệu đầu vào như thời gian học (studytime), số lần thất bại (failures), mức độ quan hệ gia đình (famrel), sức khỏe (health), việc có Internet (internet), số ngày vắng mặt (absences), tình trạng gia đình hỗ trợ (famsup), sự hỗ trợ của trường học (schoolsup), cùng với điểm số các kỳ học trước (G1 và G2), hệ thống sẽ có thể tính toán và dự đoán điểm số cuối kỳ (G3) cho học sinh.**

**Mục tiêu cụ thể của dự án là:**

* **Phát triển một mô hình học máy có thể dự đoán điểm số học sinh một cách nhanh chóng và chính xác.**
* **Cung cấp một công cụ đơn giản và dễ sử dụng cho phụ huynh và giáo viên để theo dõi và hỗ trợ học sinh.**

**1.3 Tầm quan trọng của dự án**

**Dự án này có tầm quan trọng lớn đối với giáo dục, đặc biệt là trong việc cải thiện cách thức hỗ trợ học sinh. Việc dự đoán điểm số học sinh có thể giúp phụ huynh đưa ra những quyết định tốt hơn trong việc hỗ trợ con em mình học tập. Các giáo viên cũng có thể sử dụng công cụ này để xác định những học sinh có thể cần sự hỗ trợ đặc biệt, từ đó đưa ra các biện pháp can thiệp kịp thời.**

**Ngoài ra, dự án này cũng đóng góp vào việc áp dụng các công nghệ học máy trong giáo dục, tạo ra các giải pháp tiên tiến giúp nâng cao chất lượng giảng dạy và học tập. Hệ thống dự đoán điểm số cũng có thể trở thành một công cụ hữu ích trong việc tối ưu hóa các chiến lược giảng dạy và phát triển chương trình học.**

**1.4 Phạm vi nghiên cứu**

**Phạm vi nghiên cứu của dự án này tập trung vào việc sử dụng các dữ liệu có sẵn từ bộ dữ liệu điểm số học sinh, được thu thập từ các nguồn tài liệu trực tuyến và các nghiên cứu trước đó. Dữ liệu đầu vào sẽ bao gồm các yếu tố đơn giản như thời gian học, tình trạng gia đình, và các điểm số học kỳ trước của học sinh, và mục tiêu là dự đoán điểm số cuối kỳ (G3).**

**Phạm vi nghiên cứu của dự án chủ yếu bao gồm:**

* **Dữ liệu có sẵn: Bộ dữ liệu học sinh sẽ được sử dụng để huấn luyện mô hình. Dữ liệu này đã được thu thập và có sẵn từ các nghiên cứu trước đây.**
* **Kiến thức học máy cơ bản: Dự án này sẽ áp dụng các kiến thức học máy cơ bản, đặc biệt là về thuật toán hồi quy tuyến tính, để xây dựng mô hình dự đoán.**
* **Tự học và nghiên cứu: Phần lớn kiến thức về học máy và kỹ thuật mô hình hóa sẽ được tự học thông qua các tài liệu nghiên cứu và khóa học trực tuyến.**

**2. Tổng quan lý thuyết**

**2.1 Hồi quy tuyến tính đa biến (Multiple Linear Regression)**

Hồi quy tuyến tính đa biến (Multiple Linear Regression - MLR) là một mô hình học máy được sử dụng để dự đoán một biến phụ thuộc (dependent variable) từ một hoặc nhiều biến độc lập (independent variables). Đây là một mở rộng của hồi quy tuyến tính đơn biến (Simple Linear Regression), nơi mối quan hệ giữa biến phụ thuộc và các biến độc lập được mô tả bằng một hàm tuyến tính.

Công thức tổng quát của mô hình hồi quy tuyến tính đa biến là:

A black screen with white text

AI-generated content may be incorrect.

**Các giả định cơ bản** của hồi quy tuyến tính đa biến:

1. **Mối quan hệ tuyến tính**: Giả định rằng mối quan hệ giữa biến phụ thuộc và các biến độc lập là tuyến tính.
2. **Độc lập giữa các sai số**: Các sai số (residuals) giữa các quan sát phải độc lập với nhau.
3. **Phân phối chuẩn của sai số**: Sai số của mô hình nên tuân theo phân phối chuẩn.
4. **Đồng nhất phương sai (Homoscedasticity)**: Phương sai của sai số phải là hằng số, tức là không phụ thuộc vào giá trị của các biến độc lập.

**Ví dụ minh họa**:  
Giả sử chúng ta muốn dự đoán điểm số cuối kỳ G3của học sinh dựa trên các yếu tố như thời gian học (studytime), số lần thất bại (failuresfailuresfailures), và điểm các kỳ học trước (G1, G2). Mô hình hồi quy tuyến tính đa biến có thể được biểu diễn như sau:



Chúng ta sẽ huấn luyện mô hình để tìm ra các trọng số w1, w2, w3, w4 và bias b sao cho độ chính xác của mô hình là cao nhất.

**2.2 Gradient Descent**

**Gradient Descent** là một thuật toán tối ưu hóa được sử dụng để tìm giá trị tối ưu của các tham số (trọng số và hệ số chệch) trong các mô hình học máy, bao gồm hồi quy tuyến tính. Mục tiêu của thuật toán là tối thiểu hóa hàm chi phí (cost function) bằng cách điều chỉnh các tham số theo hướng giảm dần của gradient (đạo hàm).

**Khái niệm lý thuyết**:  
Thuật toán Gradient Descent cập nhật các trọng số www và hệ số chệch bbb qua từng bước lặp, sao cho hàm chi phí (cost function) đạt giá trị nhỏ nhất. Hàm chi phí trong trường hợp của hồi quy tuyến tính là **Mean Squared Error (MSE)**, được tính như sau:

A black background with white text

AI-generated content may be incorrect.

**Cập nhật trọng số**:  
Để tối thiểu hóa hàm chi phí, Gradient Descent sẽ tính toán gradient của hàm chi phí theo các trọng số và hệ số chệch, sau đó cập nhật các tham số theo công thức sau:

A black background with white text

AI-generated content may be incorrect.

Thuật toán sẽ tiếp tục cập nhật các trọng số và hệ số chệch qua các bước lặp cho đến khi hàm chi phí hội tụ (không thay đổi nhiều giữa các lần lặp).

**2.3 Lí do chọn thuật toán**

Hồi quy tuyến tính đa biến (Multiple Linear Regression) là lựa chọn tối ưu cho bài toán dự đoán điểm số học sinh vì một số lý do sau:

1. **Mô hình đơn giản và dễ hiểu**: Hồi quy tuyến tính đa biến cung cấp một mô hình dễ hiểu với các trọng số rõ ràng, giúp phụ huynh và giáo viên dễ dàng theo dõi các yếu tố ảnh hưởng đến điểm số của học sinh.
2. **Dữ liệu đầu vào đơn giản**: Dữ liệu trong bài toán của bạn chủ yếu bao gồm các yếu tố như thời gian học (studytime), số lần thất bại (failure), mức độ quan hệ gia đình (famrel), sức khỏe (health), việc có Internet (internet), số ngày vắng mặt (absences), tình trạng gia đình hỗ trợ (famsup), sự hỗ trợ của trường học (schoolsup), và điểm số trước đó (G1, G2), các yếu tố này có thể được mô tả tốt bằng mô hình hồi quy tuyến tính.
3. **Hiệu suất tốt với dữ liệu tuyến tính**: Hồi quy tuyến tính đa biến rất hiệu quả khi có mối quan hệ tuyến tính giữa các biến độc lập và biến phụ thuộc. Trong trường hợp này, các yếu tố như thời gian học, điểm số trước đó có thể có ảnh hưởng tuyến tính đến điểm số cuối kỳ của học sinh.
4. **Tính khả thi**: Hồi quy tuyến tính đa biến yêu cầu ít tài nguyên tính toán và dễ dàng triển khai, giúp tiết kiệm thời gian và tài nguyên cho việc huấn luyện và triển khai hệ thống.
5. **Thiết kế hệ thống**  
   3.1. Kiến trúc hệ thống tổng quan  
   3.2. Sơ đồ hệ thống  
   3.3. Cấu trúc dữ liệu  
   3.4. Mô hình dữ liệu
6. **Quá trình huấn luyện mô hình**  
   4.1. Tiền xử lý dữ liệu  
   4.2. Phân chia dữ liệu  
   4.3. Huấn luyện mô hình hồi quy tuyến tính  
   4.4. Đánh giá mô hình
7. **Triển khai hệ thống**  
   5.1. Backend với Flask  
   5.2. Frontend với React  
   5.3. Kiểm thử hệ thống
8. **Đánh giá kết quả**  
   6.1. So sánh với giá trị trung bình  
   6.2. Phân tích lỗi mô hình  
   6.3. Đánh giá trải nghiệm người dùng
9. **Định hướng phát triển trong tương lai**  
   7.1. Cải thiện mô hình  
   7.2. Mở rộng ứng dụng  
   7.3. Tối ưu hóa hiệu suất
10. **Kết luận**
11. **Tài liệu tham khảo**
12. **Phụ lục**  
    10.1. Code nguồn  
    10.2. Kết quả huấn luyện mô hình  
    10.3. Ảnh chụp màn hình của ứng dụng