# **I.MỤC LỤC**

[**I.MỤC LỤC** 1](#_Toc199418834)

[**II.LỜI CẢM ƠN** 3](#_Toc199418835)

[**II.NỘI DUNG BÀI TẬP LỚN** 4](#_Toc199418836)

[**Chương 1.Giới thiệu đầu bài** 4](#_Toc199418837)

[***1.1.Thông tin đề tài*** 4](#_Toc199418838)

[***1.2.Thông tin đề tài*** 4](#_Toc199418839)

[**Chương 2.Cơ sở lý thuyết** 6](#_Toc199418840)

[***2.1.Giới thiệu thư viện pandas*** 6](#_Toc199418841)

[***2.2.Giới thiệu mô hình hồi quy tuyến tính*** 6](#_Toc199418842)

[***2.3.Giới thiệu thư viện Matplotlib*** 7](#_Toc199418843)

[**Chương 3.Thiết kế và xây dựng chương trình** 8](#_Toc199418844)

[***3.1.Sơ đồ khối hệ thống*** 8](#_Toc199418845)

[***3.2.Sơ đồ khối thuật toán chính*** 9](#_Toc199418846)

[***3.2.1.Sơ đồ khối*** 10](#_Toc199418847)

[***3.2.2.Phân tích các khối*** 10](#_Toc199418848)

[***3.2.3.Quan hệ đầu vào-ra giữa các khối*** 10](#_Toc199418849)

[***3.3.Cấu trúc dữ liệu*** 10](#_Toc199418850)

[***3.1.1.Cấu trúc dữ liệu đầu vào*** 10](#_Toc199418851)

[***3.1.2.Cấu trúc dữ liệu sau tiền xử lý*** 11](#_Toc199418852)

[***3.1.2.Cấu trúc dữ liệu cho mô hình*** 11](#_Toc199418853)

[***3.1.3.Cấu trúc dữ liệu đầu ra*** 11](#_Toc199418854)

[***3.4.Chương trình*** 11](#_Toc199418855)

[***3.4.1.Quá trình huấn luyện*** 11](#_Toc199418856)

[**Chương 4.Thực nghiệm và kết luận** 12](#_Toc199418857)

[***4.1.Thực nghiệm*** 12](#_Toc199418858)

[***4.2.Kết luận*** 12](#_Toc199418859)

[**III.TÀI LIỆU THAM KHẢO** 13](#_Toc199418860)

# **II.LỜI CẢM ƠN**

Trước hết, em xin gửi lời cảm ơn chân thành đến thầy **Nguyễn Văn Huy**, người đã tận tình hướng dẫn, chỉ bảo và hỗ trợ em trong suốt quá trình thực hiện đồ án môn học này. Những ý kiến đóng góp quý báu cùng sự định hướng của thầy đã giúp em hoàn thiện sản phẩm một cách tốt nhất

Em cũng xin cảm ơn các thầy cô trong Khoa Điện Tử, Bộ môn Công Nghệ Thông Tin đã tạo điều kiện thuận lợi để em có cơ hội học tập, nghiên cứu và thực hành, giúp em tích lũy thêm nhiều kiến thức và kinh nghiệm quan trọng trong lĩnh vực tin học công nghiệp

Cuối cùng, em xin gửi lời cảm ơn sâu sắc đến gia đình và bạn bè, những người đã luôn động viên, khích lệ và ủng hộ em trong suốt quá trình thực hiện đồ án. Chính sự hỗ trợ và tin tưởng này đã giúp em vượt qua những khó khăn và hoàn thành tốt đề tài

Em hy vọng sản phẩm này sẽ thể hiện được những nỗ lực và tâm huyết của bản thân, đồng thời góp phần vào việc ứng dụng công nghệ trong lĩnh vực tin học công nghiệp

Thái Nguyên, ngày….tháng…..năm 2025

**SINH VIÊN**

*(Ký ghi rõ họ tên)*

# **II.NỘI DUNG BÀI TẬP LỚN**

## **Chương 1.Giới thiệu đầu bài**

## ***1.1.Thông tin đề tài***

-Tên đề tài: Dự báo điểm thi sinh viên

## ***1.2.Thông tin đề tài***

-Dữ liệu đầu vào: Dữ liệu students-performance-in-exams

<https://www.kaggle.com/datasets/spscientist/students-performance-in-exams>

-Mục tiêu và phạm vi đề tài:

Đề tài nhằm xây dựng một hệ thống dự báo điểm thi cho sinh viên dựa trên việc phân tích các thông tin học tập từ dữ liệu lịch sử. Hệ thống chủ yếu tập trung vào việc:

+Thu thập dữ liệu liên quan đến quá trình học tập và kết quả hoặc điểm thi của sinh viên (thông qua file CSV có chứa các thông tin như số giờ học, số bài tập hoàn thành, điểm kiểm tra giữa kỳ – cuối kỳ, …)

+Tiền xử lý dữ liệu học tập để làm sạch, chuẩn hóa thông tin và loại bỏ giá trị thiếu hoặc không hợp lệ

+Huấn luyện các mô hình dự báo (ví dụ: mô hình hồi quy tuyến tính, Logistic Regression hay các mô hình học máy khác) nhằm tìm ra mối liên hệ giữa các biến số như hoạt động học tập, kỷ luật, chỉ số chuyên cần và điểm thi của sinh viên

+Đánh giá hiệu quả của mô hình bằng các chỉ số thống kê như RMSE, MAE,...

+Trực quan hóa kết quả dự báo qua biểu đồ và bảng dữ liệu trên giao diện web, giúp sinh viên và giáo viên dễ dàng nhận diện xu hướng và có kế hoạch cải thiện chất lượng học tập

-Yêu cầu đề tài:

+Dữ liệu đầu vào: Tập hợp thông tin học tập và kết quả thi của sinh viên, được lưu ở dạng file CSV. Các cột dữ liệu có thể bao gồm: Mã sinh viên, tên, môn học, điểm số từng kỳ thi, số buổi học tham gia, số bài tập nộp,…

+Dữ liệu đầu ra: Bảng dự báo điểm thi cho các kỳ thi tương lai và các biểu đồ trực quan thể hiện xu hướng tiến bộ học tập, từ đó đưa ra dự báo khả năng đạt (vượt đạt/không đạt) của từng sinh viên

+Phương pháp xử lý: Sử dụng thư viện Pandas để tiền xử lý dữ liệu; áp dụng các mô hình dự báo hồi quy (và/hoặc mô hình phân loại đối với bài toán đánh giá vượt đạt hay không) sau đó trực quan hóa kết quả bằng Matplotlib

+Giao diện người dùng: Xây dựng ứng dụng web cho phép người dùng upload file dữ liệu, lựa chọn tham số và xem kết quả dự báo cùng với biểu đồ trực quan

## **Chương 2.Cơ sở lý thuyết**

## ***2.1.Giới thiệu thư viện pandas***

Pandas là một thư viện mã nguồn mở trong Python, được sử dụng rộng rãi trong phân tích dữ liệu và khoa học dữ liệu. Được phát triển bởi Wes McKinney vào năm 2008, Pandas cung cấp các công cụ mạnh mẽ để thao tác và xử lý dữ liệu một cách hiệu quả

Các tính năng chính của Pandas:

+Hỗ trợ xử lý dữ liệu dạng bảng với cấu trúc DataFrame

+Hỗ trợ nhiều định dạng dữ liệu đầu vào như CSV, JSON, SQL, Excel

+Cho phép thao tác lọc, nhóm, nối dữ liệu, tính toán thống kê nhanh chóng

+Hỗ trợ xử lý giá trị thiếu (NaN), làm sạch dữ liệu hiệu quả

## ***2.2.Giới thiệu mô hình hồi quy tuyến tính***

Hồi quy tuyến tính (Linear Regression) là một phương pháp dự báo phổ biến trong học máy, được sử dụng để mô hình hóa mối quan hệ giữa một hoặc nhiều biến độc lập (đặc trưng học tập) và biến phụ thuộc (điểm số dự báo)

Phương trình hồi quy tuyến tính: y=w0+w1x1+w2x2+...+wnxn+ϵ

+y: Giá trị điểm thi dự báo

+xi: Các đặc trưng ảnh hưởng đến điểm thi (số giờ học, điểm trung bình học kỳ trước, số bài tập hoàn thành…)

+wi: Hệ số hồi quy, biểu diễn mức độ ảnh hưởng của từng đặc trưng

+ϵ - epsilon: Sai số của mô hình

## ***2.3.Giới thiệu thư viện Matplotlib***

Matplotlib là một thư viện mạnh mẽ trong Python dùng để trực quan hóa dữ liệu, hỗ trợ biểu diễn kết quả dự báo một cách trực quan

Ứng dụng Matplotlib trong dự báo điểm thi:

+Vẽ biểu đồ điểm số thực tế và dự báo

+Biểu đồ phân phối điểm thi

+Biểu đồ xu hướng học tập theo thời gian

## **Chương 3.Thiết kế và xây dựng chương trình**

## ***3.1.Sơ đồ khối hệ thống***

Hệ thống dự báo điểm thi sinh viên có thể được chia thành các module chính như sau:

+ Thu thập dữ liệu: Nhận dữ liệu từ nguồn CSV (các tập dữ liệu học tập của sinh viên)

+ Tiền xử lý dữ liệu: Làm sạch, xử lý giá trị thiếu, chuyển đổi định dạng

+ Phân tích dữ liệu: Thống kê thông tin học tập và điểm số

+ Huấn luyện mô hình: Áp dụng các thuật toán như Hồi quy tuyến tính hoặc Logistic Regression để dự báo

+ Dự báo điểm thi: Sử dụng mô hình để đưa ra kết quả dự đoán điểm số hoặc phân loại sinh viên đạt hay không đạt

+ Trực quan hóa kết quả: Biểu diễn kết quả bằng bảng dữ liệu và biểu đồ

+ Giao diện web: Người dùng nhập dữ liệu, chọn mô hình, xem kết quả

A diagram with black text and white text

AI-generated content may be incorrect.

## ***3.2.Sơ đồ khối thuật toán chính***

Sơ đồ khối thuật toán chính sẽ được chia thành các bước:

+ Bước 1: Nhập dữ liệu CSV từ sinh viên

+ Bước 2: Tiền xử lý dữ liệu (loại bỏ giá trị thiếu, chuẩn hóa dữ liệu)

+ Bước 3: Chọn mô hình dự báo (hồi quy tuyến tính hoặc phân loại)

+ Bước 4: Huấn luyện mô hình với dữ liệu đã xử lý

+ Bước 5: Dự báo điểm thi hoặc phân loại kết quả

+ Bước 6: Hiển thị kết quả dự báo và biểu đồ lên giao diện web

## ***3.2.1.Sơ đồ khối***

## ***3.2.2.Phân tích các khối***

+ Nhập dữ liệu: Đọc file CSV chứa thông tin học tập của sinh viên

+ Tiền xử lý dữ liệu: Chuẩn hóa thông tin điểm số, loại bỏ dữ liệu thiếu

+ Huấn luyện mô hình: Áp dụng mô hình hồi quy tuyến tính hoặc phân loại Logistic Regression

+ Dự báo điểm thi: Sử dụng mô hình đã huấn luyện để đưa ra kết quả dự báo

+ Hiển thị kết quả: Xuất ra bảng dữ liệu và biểu đồ trực quan

## ***3.2.3.Quan hệ đầu vào-ra giữa các khối***

## ***3.3.Cấu trúc dữ liệu***

## ***3.1.1.Cấu trúc dữ liệu đầu vào***

Dữ liệu đầu vào được lưu dưới dạng bảng DataFrame của Pandas, chứa các thông tin:

* gender: Giới tính sinh viên
* race/ethnicity: Dân tộc
* parental level of education: Trình độ học vấn của phụ huynh
* lunch: Loại suất ăn
* test preparation course: Tình trạng tham gia ôn luyện
* math score: Điểm thi môn toán (hoặc các môn học khác)

## ***3.1.2.Cấu trúc dữ liệu sau tiền xử lý***

Sau khi tiền xử lý, dữ liệu được chuẩn hóa và sẵn sàng đưa vào mô hình:

* Loại bỏ giá trị thiếu
* Chuyển đổi biến categorical thành dạng số (OneHot Encoding)
* Chuẩn hóa điểm số để phù hợp với mô hình hồi quy

## ***3.1.2.Cấu trúc dữ liệu cho mô hình***

Dữ liệu đầu vào sẽ gồm:

* Tập huấn luyện (X\_train, y\_train): Các đặc trưng học tập và điểm số để huấn luyện mô hình
* Tập kiểm thử (X\_test, y\_test): Dữ liệu dùng để đánh giá mô hình

## ***3.1.3.Cấu trúc dữ liệu đầu ra***

Sau khi chạy mô hình, hệ thống sẽ xuất kết quả dự báo:

* predicted\_score: Điểm thi dự báo
* pass\_fail: Phân loại sinh viên đạt hay không đạt

## ***3.4.Chương trình***

## ***3.4.1.Quá trình huấn luyện***

* Bước 1: Đọc dữ liệu từ file CSV
* Bước 2: Tiền xử lý dữ liệu (chuẩn hóa, xử lý giá trị thiếu)
* Bước 3: Chia tập dữ liệu thành tập huấn luyện và kiểm thử
* Bước 4: Huấn luyện mô hình hồi quy tuyến tính hoặc Logistic Regression
* Bước 5: Kiểm thử mô hình và đánh giá độ chính xác
* Bước 6: Lưu mô hình và xuất kết quả dự báo

## **Chương 4.Thực nghiệm và kết luận**

## ***4.1.Thực nghiệm***

## ***4.2.Kết luận***

- Những điểm đã đạt được:

+ Xây dựng hệ thống webapp hoạt động ổn định: Cho phép người dùng tải lên dữ liệu học tập, lựa chọn mô hình dự báo và xem kết quả phân tích

+ Xử lý dữ liệu học tập hiệu quả: Sử dụng thư viện Pandas để làm sạch dữ liệu, xử lý giá trị thiếu và chuẩn hóa thông tin

+ Triển khai các mô hình dự báo: Áp dụng hồi quy tuyến tính để dự đoán điểm thi và Logistic Regression để phân loại sinh viên đạt hay không đạt

+ Trực quan hóa kết quả: Sử dụng Matplotlib để hiển thị biểu đồ giúp phân tích xu hướng điểm thi và sự ảnh hưởng của các yếu tố học tập

- Những điểm hạn chế:

+ Dữ liệu mẫu vẫn còn nhỏ, có thể ảnh hưởng đến độ chính xác của mô hình

+ Chưa thử nghiệm với các mô hình học sâu (Deep Learning) hoặc phương pháp dự báo tiên tiến hơn

+ Giao diện web có thể mở rộng thêm các tính năng tùy chỉnh trực quan hơn

- Hướng cải tiến:

+ Mở rộng tập dữ liệu: Tích hợp thêm nhiều nguồn dữ liệu học tập thực tế để cải thiện hiệu quả dự báo

+ Thử nghiệm với các mô hình nâng cao: Áp dụng Random Forest, XGBoost hoặc Neural Network để tăng độ chính xác

+ Phát triển giao diện web: Bổ sung bảng phân tích chi tiết và biểu đồ động giúp tương tác tốt hơn với người dùng

+ Tích hợp API: Cho phép nhận dữ liệu học tập từ các hệ thống quản lý đào tạo trong trường học

# **III.TÀI LIỆU THAM KHẢO**

*- McKinney, W. (2017). Python for Data Analysis. O'Reilly Media. → Giới thiệu về thư viện Pandas và cách xử lý dữ liệu hiệu quả.*

*- Pedregosa, F., et al. (2011). Scikit-learn: Machine Learning in Python. Journal of Machine Learning Research. → Nghiên cứu về các thuật toán dự báo như Linear Regression và Logistic Regression*

*- Kaggle Dataset - Students Performance in Exams → Tập dữ liệu mẫu được sử dụng để thử nghiệm mô hình dự báo điểm thi*