

**ĐẠI HỌC QUỐC GIA THÀNH PHỐ HỒ CHÍ MINH
TRƯỜNG ĐẠI HỌC CÔNG NGHỆ THÔNG TIN
KHOA MẠNG MÁY TÍNH VÀ TRUYỀN THÔNG**



VÕ LÊ THÀNH PHÁT

**KHOÁ LUẬN TỐT NGHIỆP
EDUTWIN - BẢN SAO HỌC TẬP KỸ THUẬT SỐ**

EDUTWIN - EDUCATION DIGITAL TWIN

KỸ SỰ NGÀNH MẠNG MÁY TÍNH VÀ TRUYỀN THÔNG DỮ LIỆU

TP. HỒ CHÍ MINH, 2025

**ĐẠI HỌC QUỐC GIA THÀNH PHỐ HỒ CHÍ MINH
TRƯỜNG ĐẠI HỌC CÔNG NGHỆ THÔNG TIN
KHOA MẠNG MÁY TÍNH VÀ TRUYỀN THÔNG**



VÕ LÊ THÀNH PHÁT – 21522452

**KHOÁ LUẬN TỐT NGHIỆP
EDUTWIN - BẢN SAO HỌC TẬP KỸ THUẬT SỐ
EDUTWIN - EDUCATION DIGITAL TWIN**

KỸ SỰ NGÀNH MẠNG MÁY TÍNH VÀ TRUYỀN THÔNG DỮ LIỆU

**GIẢNG VIÊN HƯỚNG DẪN
TS. Nguyễn Tân Hoàng Phước**

TP. HỒ CHÍ MINH, 2025

THÔNG TIN HỘI ĐỒNG BẢO VỆ KHÓA LUẬN

Hội đồng chấm khóa luận tốt nghiệp, thành lập theo Quyết định số ngày ...
..... của Hiệu trưởng Trường Đại học Công nghệ Thông tin.

1. Chủ tịch:
2. Thư ký:
3. Ủy viên:

LỜI CẢM ƠN

Từ thời điểm tìm kiếm đến khi thực hiện xong bài luận văn, tôi rất biết ơn tất cả sự hướng dẫn, động viên và sự hỗ trợ từ phía Nhà trường, Thầy Cô, bạn bè, đồng nghiệp và Gia đình.

Tôi xin gửi lời chân thành cảm ơn đến Thầy hướng dẫn là **TS. Nguyễn Tân Hoàng Phước** đã nhiệt tình chỉ dẫn trong suốt thời gian tìm hiểu, nghiên cứu thực nghiệm cho đến thời điểm hoàn thành bài luận văn tại trường.

Tôi chân thành gửi lời tri ân đến các **Quý Thầy Cô trong Phòng Đào tạo của trường Đại học Công nghệ Thông tin** đã hỗ trợ cũng như cung cấp cho tôi những tri thức, kiến thức, hướng dẫn các thủ tục và các kinh nghiệm quý giá cho tôi trong suốt quá trình học tập nghiên cứu ở trường.

Đồng thời, tôi cũng muốn gửi những lời cảm ơn chân thành tới Gia đình, bạn bè và các đồng nghiệp đã hỗ trợ tôi trong suốt toàn bộ quá trình học và thực nghiệm nhằm hoàn thành nội dung bài luận văn này.

Với khoảng thời gian ngắn đồng thời kiến thức bản thân tôi còn có nhiều hạn chế, do đó bài luận văn chắc chắn vẫn còn các thiếu sót. Tôi rất mong sẽ nhận được những lời khuyên, góp ý của quý Thầy Cô.

Nhóm tác giả

TÓM TẮT

Trong bối cảnh chuyển đổi số giáo dục, việc cá nhân hóa trải nghiệm học tập đang trở thành yêu cầu cấp thiết. Tuy nhiên, các hệ thống quản lý học tập hiện tại thường chỉ dừng lại ở việc lưu trữ kết quả mà thiếu đi khả năng phân tích và dự báo. Khóa luận này trình bày việc nghiên cứu và xây dựng "**EduTwin - Bản sao Học tập Kỹ thuật số tích hợp AI**", một giải pháp nhằm khắc phục các hạn chế trên.

EduTwin ứng dụng ý tưởng "**Digital Twin**" để tạo ra một bản sao kỹ thuật số của người học, có khả năng "tiến hóa" thông qua dữ liệu tương tác liên tục. Hệ thống kết hợp thuật toán **Lazy Learning** (K-NN, KR, LWLR) để dự đoán kết quả học tập tức thời mà không cần huấn luyện lại, cùng với các **Large Language Model** (LLM) để đóng vai trò như một trợ lý ảo thấu hiểu ngữ cảnh và tâm lý người học. Kết quả thực nghiệm cho thấy hệ thống không chỉ cung cấp các dự báo có độ tin cậy cao mà còn tạo ra trải nghiệm tương tác tự nhiên, giúp học sinh chủ động hơn trong việc định hướng lộ trình học tập của mình.

MỤC LỤC

Thông tin hội đồng bảo vệ khóa luận	i
Lời cảm ơn	ii
Tóm tắt	iii
Mục lục	iv
Danh sách bảng	viii
Danh sách hình vẽ	ix
Danh mục từ viết tắt	xii
Chương 1. TỔNG QUAN VỀ ĐỀ TÀI	1
1.1 Lý do chọn đề tài	1
1.2 Mục tiêu, đối tượng và phạm vi nghiên cứu	2
1.2.1 Mục tiêu nghiên cứu	2
1.2.2 Đối tượng nghiên cứu	3
1.2.3 Phạm vi nghiên cứu	4
1.2.3.1 Phạm vi về Dữ liệu và Ngữ cảnh áp dụng	4
1.2.3.2 Phạm vi về Chức năng hệ thống	4
1.2.3.3 Phạm vi về Công nghệ	5
1.3 Phương pháp nghiên cứu	5
1.4 Tổng quan đề tài	6
1.4.1 Các nghiên cứu và giải pháp liên quan	6
1.4.1.1 Digital Twin trong Giáo dục: Từ mô phỏng môi trường đến bản sao người học	6
1.4.1.2 Học tập thích ứng và Bài toán dự báo hiệu suất: Hướng tiếp cận Lazy Learning	7
1.4.1.3 Generative Artificial Intelligence (AI) và Large Language Model (LLM): Từ công cụ tạo sinh đến trợ lý ngữ cảnh	8
1.4.1.4 Xử lý dữ liệu khuyết thiêу trong hồ sơ học tập	8
1.4.2 Đóng góp mới của đề tài	9

MỤC LỤC

1.5 Câu trúc Khoa luận tốt nghiệp	9
Chương 2. Cơ sở lý thuyết	11
2.1 Digital Twin và Học tập Cá nhân hóa	11
2.2 Lazy Learning	12
2.3 Lượng tử hóa Vector và Tối ưu hóa Mẫu hình	13
2.4 Kiến trúc Transformer và Cơ chế Attention.	15
2.5 Xử lý Dữ liệu và Kỹ thuật Imputation	17
2.6 Các chỉ số Đánh giá Hiệu năng Mô hình	17
2.6.1 Sai số tuyệt đối trung bình (MAE - Mean Absolute Error)	17
2.6.2 Căn bậc hai sai số bình phương trung bình (RMSE - Root Mean Squared Error)	18
2.6.3 Hệ số xác định (R^2 - R-squared)	18
2.7 Trực quan hóa Dữ liệu Giáo dục.	19
2.8 Xác thực Application Programming Interface (API) với JSON Web Token (JWT)	19
Chương 3. Phân tích và Thiết kế hệ thống	21
3.1 Phát biểu bài toán	21
3.1.1 Bối cảnh	21
3.1.2 Bài toán cần giải quyết	21
3.1.3 Các thách thức kỹ thuật và Giải pháp	21
3.2 Quy trình thực hiện	22
3.3 Phân tích yêu cầu hệ thống	23
3.3.1 Yêu cầu chức năng	23
3.3.2 Yêu cầu phi chức năng	23
3.4 Mô hình hóa quy trình nghiệp vụ	25
3.4.1 Biểu đồ Use Case	25
3.4.2 Quy trình Đăng nhập và Xác thực	25
3.4.3 Quy trình Dự đoán Điểm số	26
3.4.4 Quy trình Chat với AI	26
3.4.5 Quy trình Thiết lập Câu trúc Học tập	27
3.4.6 Quy trình Upload Dataset và Đánh giá Model	27
3.4.7 Quy trình Trích xuất Tài liệu	28

MỤC LỤC

3.5	Thiết kế Kiến trúc hệ thống	28
3.5.1	Kiến trúc tổng thể	28
3.5.2	Kiến trúc Module AI	28
3.6	Thiết kế Cơ sở dữ liệu	29
3.6.1	Thiết kế mức khái niệm	29
3.6.2	Thiết kế chi tiết lược đồ dữ liệu	29
3.6.3	Phân tích các quyết định thiết kế quan trọng	30
3.7	Thiết kế Thuật toán và Luồng xử lý dữ liệu	37
3.7.1	Luồng xử lý dữ liệu khuyết	37
3.7.2	Chiến lược Tối ưu hóa dự đoán với Lượng tử hóa Vector và Tối ưu hóa Mẫu hình	37
3.7.3	Luồng dự đoán Lazy Learning	38
3.7.4	Luồng xử lý Chatbot RAG	40
3.8	Thiết kế Giao diện người dùng	41
3.8.1	Sơ đồ tổ chức thông tin	41
3.8.2	Thiết kế chi tiết các màn hình chính	41
3.8.3	Danh sách Công nghệ	43
3.8.4	Môi trường triển khai và Yêu cầu hệ thống	44
Chương 4.	Thực nghiệm và đánh giá	46
4.1	Môi trường và Dữ liệu thực nghiệm	46
4.1.1	Môi trường thực nghiệm	46
4.1.1.1	Phân tích Khám phá Dữ liệu	46
4.2	Kịch bản 1: Đánh giá hiệu quả xử lý dữ liệu khuyết	48
4.2.1	Thiết lập thực nghiệm	48
4.2.2	Phân tích kết quả	48
4.3	Kịch bản 2: Đánh giá hiệu năng đa mô hình — So sánh Lazy Learning và Eager Learning	50
4.3.1	Kết quả thực nghiệm định lượng	51
4.3.2	Phân tích kết quả	51
4.4	Kịch bản 3: Đánh giá hiệu năng và Tốc độ phản hồi	52
4.4.1	Thiết lập thực nghiệm và Dữ liệu	52
4.4.2	Phân tích kết quả thực nghiệm	52

MỤC LỤC

4.5 Kịch bản 4: Đánh giá Chatbot và Bảo mật Personally Identifiable Information (PII)	53
4.5.1 Kiểm thử Personally Identifiable Information (PII) Redaction (Bảo mật)	53
4.5.2 Kiểm thử RAG Context Injection (Cá nhân hóa)	54
4.6 Bàn luận chung	55
4.7 Kết quả xây dựng ứng dụng	56
4.7.1 Giao diện đăng nhập/đăng ký	56
4.7.2 Phân hệ Quản trị viên	56
4.7.3 Phân hệ Người dùng	57
4.7.4 Giao diện Chatbot AI	58
Chương 5. Kết luận và hướng phát triển	60
5.1 KẾT LUẬN VÀ HƯỚNG PHÁT TRIỂN	60
5.1.1 Kết luận chung	60
5.1.2 Các hạn chế của đề tài	60
5.1.3 Hướng phát triển	61
Phụ lục A. Các thông tin liên quan	62
A.1 Tài nguyên thực nghiệm	62
Tài liệu tham khảo	65

DANH SÁCH BẢNG

3.1	Các thách thức kỹ thuật và Giải pháp đề xuất	22
3.2	Danh sách chức năng cho Học sinh	23
3.3	Danh sách chức năng cho Admin	24
3.4	Danh sách chức năng của AI Core	24
3.5	Bảng custom_teaching_structures - Định nghĩa cấu trúc học tập .	31
3.6	Bảng custom_user_scores - Lưu trữ điểm số người dùng	32
3.7	Bảng custom_reference_dataset - Dữ liệu mẫu (Training Data) .	32
3.8	Bảng users - Thông tin người dùng	33
3.9	Bảng ai_insights - Lưu trữ nhận định AI về học sinh	33
3.10	Bảng custom_structure_documents - Lưu tài liệu tham khảo cho AI	34
3.11	Bảng ml_model_parameters - Tham số các mô hình ML	34
3.12	Bảng ml_model_config - Cấu hình mô hình ML đang hoạt động . . .	35
3.13	Bảng chat_sessions - Quản lý phiên hội thoại	35
3.14	Bảng chat_messages - Lưu trữ tin nhắn	35
3.15	Bảng user_structure_preferences - Lưu trữ preferences của user theo structure	36
4.1	So sánh hiệu năng điền khuyết (KNN so với các phương pháp thống kê) .	50
4.2	So sánh hiệu năng giữa các mô hình Lazy Learning và Eager Learning .	51
4.3	Kết quả Stress Test so sánh hiệu năng (dữ liệu thực tế, 54 features)	53
4.4	Hiệu quả của module Personally Identifiable Information (PII) Redaction .	54
4.5	So sánh Prompt và Phản hồi giữa các mức độ Context	54

DANH SÁCH HÌNH VẼ

DANH SÁCH HÌNH VẼ

4.13 Danh sách các thông tin cá nhân hóa mà trợ lý ảo nắm bắt được qua quá trình trò chuyện	59
---	----

DANH MỤC TỪ VIỆT TẮT

AI	Artificial Intelligence
API	Application Programming Interface
AWS	Amazon Web Services
CI/CD	Continuous Integration and Continuous Delivery
CMK	Customer Master Key
CPU	Central Processing Unit
DEK	Data Encryption Key
EAV	Entity–Attribute–Value
EM	Expectation–Maximization
ERD	Entity Relationship Diagram
GIN	Generalized Inverted Index
GPA	Grade Point Average
GPU	Graphics Processing Unit
HTTP	Hypertext Transfer Protocol
I/O	Input/Output
IBL	Instance-based Learning
IoT	Internet of Things
ITS	Intelligent Tutoring Systems
JWT	JSON Web Token
KMS	Key Management Service
KNN	k-Nearest Neighbors
KNNI	k-Nearest Neighbors Imputation
KR	Kernel Regression
LLM	Large Language Model
LMS	Learning Management System
LSTM	Long Short-Term Memory
LWLR	Locally Weighted Linear Regression
MAE	Mean Absolute Error
MCAR	Missing Completely At Random
ML	Machine Learning

DANH SÁCH HÌNH VẼ

(Continued)

MVP	Minimum Viable Product
PII	Personally Identifiable Information
RAG	Retrieval-Augmented Generation
RAM	Random Access Memory
RMSE	Root Mean Squared Error
SaaS	Software as a Service
SOA	Service-Oriented Architecture
THPT	Trung Học Phổ Thông
ZPD	Zone of Proximal Development

Chương 1. TỔNG QUAN VỀ ĐỀ TÀI

1.1 Lý do chọn đề tài

Trong bối cảnh Cách mạng công nghiệp 4.0, chuyển đổi số trong giáo dục đã trở thành một xu thế tất yếu trên toàn cầu. Tại Việt Nam, Quyết định số 131/QĐ-TTg [1] của Thủ tướng Chính phủ đã phê duyệt Đề án "Tăng cường ứng dụng công nghệ thông tin và chuyển đổi số trong giáo dục và đào tạo giai đoạn 2022-2025", đặt nền móng cho việc ứng dụng các công nghệ tiên tiến như Trí tuệ nhân tạo (Artificial Intelligence (AI)) và Dữ liệu lớn (Big Data). Tuy nhiên, thực trạng các hệ thống giáo dục hiện nay vẫn còn tồn tại nhiều bất cập đáng kể.

Thứ nhất, các hệ thống quản lý học tập (Learning Management System (LMS)) truyền thống như VnEdu (VNPT), SMAS (Viettel), Google Classroom (Google) chủ yếu đóng vai trò là "kho lưu trữ thụ động". Chúng làm tốt nhiệm vụ số hóa điểm số và hồ sơ, nhưng hoàn toàn thiếu vắng khả năng phân tích sâu (deep analytics) để đưa ra các dự báo mang tính chiến lược. Học sinh và phụ huynh thường chỉ biết kết quả khi kỳ thi đã kết thúc, dẫn đến việc mọi biện pháp can thiệp đều trở nên muộn màng. Sự thiếu hụt thông tin định hướng khiến người học rơi vào trạng thái "mù mờ" về năng lực thực sự và lộ trình phát triển của bản thân.

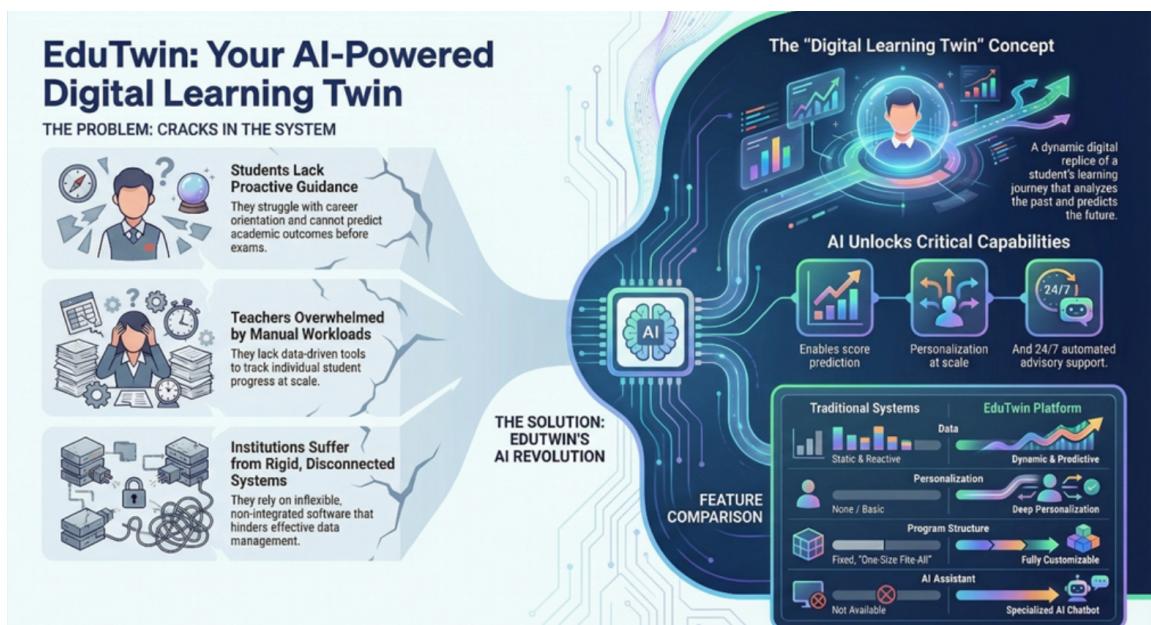
Thứ hai, khái niệm "Cá nhân hóa học tập" (Personalized Learning) thường chỉ dừng lại ở khẩu hiệu hoặc các tính năng cơ bản như gợi ý bài tập dựa trên quy tắc cứng. Chưa có một hệ thống nào thực sự thấu hiểu hành vi, thói quen và phong cách học tập của từng cá nhân để đóng vai trò như một người bạn đồng hành thực thụ.

Xuất phát từ nhu cầu cấp thiết đó, đề tài "EduTwin - Bản sao Học tập Kỹ thuật số tích hợp AI" được đề xuất nghiên cứu và phát triển. Ý tưởng cốt lõi của EduTwin là khởi tạo một "Bản sao số" (Digital Twin) cho mỗi học sinh. Khác với các hồ sơ tĩnh truyền thống, bản sao số này là một thực thể sống động, có khả năng tiến hóa cùng với sự phát triển của người học.

EduTwin vận hành dựa trên một vòng lặp tương tác thông minh và liên tục:

- **Thu thập dữ liệu đa chiều:** Hệ thống không chỉ ghi nhận điểm số mà còn thẩm thấu các dữ liệu hành vi, thói quen và sở thích thông qua quá trình tương tác tự nhiên với người dùng.
- **Phân tích & Dự báo thời gian thực:** Ứng dụng các thuật toán Lazy Learning, hệ

CHƯƠNG 1. TỔNG QUAN VỀ ĐỀ TÀI



Hình 1.1. Mô hình tổng quan EduTwin: Giải pháp khắc phục các hạn chế về dự báo và cá nhân hóa của hệ thống giáo dục truyền thống thông qua công nghệ Digital Learning Twin.

thống có khả năng dự báo kết quả học tập ngay lập tức (real-time prediction) mà không cần quy trình huấn luyện lại tốn kém, giúp người học nhận diện sớm các nguy cơ hoặc cơ hội.

- **Thích nghi & Đồng hành:** Thông qua Chatbot tích hợp Large Language Model (LLM), EduTwin tự động điều chỉnh phong cách giao tiếp, từ ngôn ngữ trang trọng đến thân mật, từ nghiêm khắc đến khích lệ, để phù hợp nhất với tâm lý của từng học sinh.

Đây chính là sự chuyển dịch mô hình từ "công cụ hỗ trợ thụ động" sang "trợ lý thông minh chủ động", đánh dấu một bước tiến mới trong việc ứng dụng công nghệ để khai phóng tiềm năng con người.

1.2 Mục tiêu, đối tượng và phạm vi nghiên cứu

1.2.1 Mục tiêu nghiên cứu

Mục tiêu tổng quát của đề tài là thiết kế và hiện thực hóa nền tảng EduTwin - một hệ sinh thái giáo dục thông minh, linh hoạt và lấy người học làm trung tâm. Hệ thống được xây dựng để giải quyết bài toán cá nhân hóa ở mức độ sâu (deep personalization) và cung cấp khả năng tùy biến cao cho các cơ sở giáo dục.

Các mục tiêu cụ thể bao gồm:

CHƯƠNG 1. TỔNG QUAN VỀ ĐỀ TÀI

- **Xây dựng Engine dự đoán thích ứng nhanh và hiệu năng cao:** Nghiên cứu và triển khai nhóm thuật toán Lazy Learning (k-Nearest Neighbors (KNN), Kernel Regression (KR), Locally Weighted Linear Regression (LWLR)). Đặc biệt, giải quyết bài toán độ trễ tính toán trên dữ liệu lớn bằng chiến lược truy vấn hai tầng: kết hợp **Phân cụm đánh chỉ mục (Cluster Indexing)** và **Tuyển chọn mẫu hình thích ứng (Adaptive Prototype Selection)**. Điều này đảm bảo hệ thống duy trì độ phức tạp tính toán thấp, đạt tốc độ phản hồi thời gian thực (Real-time) ngay cả khi quy mô dữ liệu mở rộng.
- **Phát triển Trợ lý ảo AI thấu hiểu ngữ cảnh (Context-Aware AI):** Tích hợp các mô hình ngôn ngữ lớn (LLM) tiên tiến để xây dựng một Chatbot không chỉ trả lời câu hỏi mà còn có khả năng ghi nhớ lịch sử hội thoại, thấu hiểu ngữ cảnh giáo dục và tự động điều chỉnh phản hồi dựa trên hồ sơ tâm lý của học sinh.
- **Cơ chế học tập cá nhân hóa (Personalization Learner):** Phát triển module AI chuyên biệt có nhiệm vụ "quan sát" và "học" từ các tương tác của người dùng. Hệ thống sẽ tự động nhận diện các đặc điểm như: thói quen giờ giấc, phong cách giao tiếp (ngắn gọn/chi tiết), và các mối quan tâm đặc biệt để tinh chỉnh trải nghiệm người dùng.
- **Kiến trúc Cấu trúc tùy biến (Dynamic Custom Structure):** Giải quyết bài toán "cứng nhắc" của các phần mềm hiện có bằng cách cho phép Quản trị viên (Admin) tự định nghĩa hoàn toàn cấu trúc dữ liệu (môn học, hệ số, quy tắc đánh giá) thông qua giao diện. Điều này giúp EduTwin có thể áp dụng cho mọi mô hình giáo dục, từ Trung Học Phổ Thông (THPT) công lập, trường tư thục quốc tế đến các trung tâm luyện thi chứng chỉ.

1.2.2 Đối tượng nghiên cứu

Để đạt được các mục tiêu trên, đề tài tập trung nghiên cứu vào các đối tượng sau:

- **Dữ liệu học tập số hóa:** Bao gồm dữ liệu điểm số có cấu trúc (structured grades), dữ liệu phi cấu trúc từ lịch sử hội thoại (chat logs) và các metadata về cấu trúc chương trình học.
- **Thuật toán Học lười và Tối ưu hóa:** Tập trung sâu vào các kỹ thuật KNN, KR và LWLR. Đồng thời nghiên cứu kỹ thuật Lượng tử hóa Vector (Vector Quantization)

CHƯƠNG 1. TỔNG QUAN VỀ ĐỀ TÀI

để tối ưu không gian tìm kiếm.

- **Kỹ thuật Prompt Engineering & Context Management:** Nghiên cứu các phương pháp tối ưu hóa ngữ cảnh (context window) và che giấu thông tin định danh (Personally Identifiable Information (PII) Redaction) để tích hợp an toàn các mô hình LLM vào môi trường giáo dục.

Khách hàng mục tiêu và người thụ hưởng của hệ thống:

- **Học sinh/Sinh viên:** Những người cần một lộ trình định hướng rõ ràng và một người bạn đồng hành 24/7.
- **Nhà quản lý giáo dục (Admin/Manager):** Những người cần một công cụ linh hoạt để quản lý chất lượng đào tạo theo các tiêu chuẩn riêng biệt của cơ sở mình.

1.2.3 Phạm vi nghiên cứu

Để đảm bảo tính khả thi trong khuôn khổ thời gian của khóa luận, phạm vi nghiên cứu được xác định cụ thể trên ba khía cạnh:

1.2.3.1 Phạm vi về Dữ liệu và Ngữ cảnh áp dụng

- **Khả năng của hệ thống:** EduTwin được thiết kế với kiến trúc "Metadata-driven" (hướng siêu dữ liệu), cho phép áp dụng linh hoạt cho nhiều mô hình giáo dục khác nhau (từ niên chế, tín chỉ đến các khóa học ngắn hạn) thông qua cơ chế định nghĩa cấu trúc động.
- **Phạm vi dữ liệu thực nghiệm:** Trong khuôn khổ khóa luận này, nhóm nghiên cứu giới hạn việc thu thập dữ liệu và kiểm thử các thuật toán dự báo dựa trên "mô hình Giáo dục phổ thông (THPT)". Đây là mô hình tiêu chuẩn dùng để đánh giá tính chính xác của thuật toán Lazy Learning trước khi mở rộng sang các mô hình phức tạp khác.

1.2.3.2 Phạm vi về Chức năng hệ thống

Hệ thống EduTwin được phát triển các phân hệ cốt lõi:

- **Phân hệ Quản trị (Admin Portal):** Tập trung vào tính linh hoạt cấu trúc.
 - Quản lý Cấu trúc động (Dynamic Structure Management): Cho phép Admin tự định nghĩa mô hình phân cấp dữ liệu (Ví dụ: Năm học → Học kỳ → Môn học; hoặc Khóa học → Module).

CHƯƠNG 1. TỔNG QUAN VỀ ĐỀ TÀI

- ◊ Quản lý Mô hình AI: Theo dõi hiệu suất dự báo.
- **Phân hệ Người dùng (Student Portal):** Tập trung vào trải nghiệm cá nhân hóa.
 - ◊ Dashboard Analytics & Dự báo kết quả (What-if Analysis).
 - ◊ AI Companion: Chatbot tư vấn học tập sử dụng công nghệ Retrieval-Augmented Generation (RAG).

1.2.3.3 Phạm vi về Công nghệ

- **Thuật toán:** Tập trung sâu vào nhóm thuật toán Lazy Learning (KNN, KR, LWLR) và kỹ thuật điền khuyết dữ liệu (Imputation).
- **Nền tảng:**
 - ◊ **Backend:** Python (FastAPI) cho hiệu năng xử lý bất đồng bộ cao.
 - ◊ **Frontend:** ReactJS (Vite) tối ưu trải nghiệm người dùng.
 - ◊ **Database:** PostgreSQL với JSONB để lưu trữ cấu trúc động và Redis để caching dữ liệu Cluster Index phục vụ dự đoán thời gian thực.
 - ◊ **Deployment:** Docker containerization để chuẩn hóa môi trường triển khai.

1.3 Phương pháp nghiên cứu

Đề tài áp dụng cách tiếp cận Nghiên cứu hành động (Action Research) kết hợp với quy trình phát triển phần mềm hiện đại.

- **Nghiên cứu lý thuyết chuyên sâu:**
 - ◊ Tổng hợp và phân tích các bài báo khoa học về Adaptive Learning Systems và Student Modeling.
 - ◊ Nghiên cứu nguyên lý hoạt động và ưu nhược điểm của các họ thuật toán Eager Learning so với Lazy Learning trong bối cảnh dữ liệu giáo dục thay đổi nhanh (high volatility).
- **Phân tích thiết kế hệ thống hướng đối tượng:**
 - ◊ Sử dụng để mô hình hóa toàn bộ quy trình nghiệp vụ và kiến trúc hệ thống.

CHƯƠNG 1. TỔNG QUAN VỀ ĐỀ TÀI

- ◊ Thiết kế kiến trúc hướng dịch vụ Service-Oriented Architecture (SOA)) để đảm bảo tính mở rộng và khả năng bảo trì.

- **Thực nghiệm và Đánh giá:**

- ◊ *Thu thập dữ liệu:* Xây dựng bộ dữ liệu mẫu (Representative Dataset) mô phỏng phân phối điểm số thực tế tại các trường THPT Việt Nam.
- ◊ *Triển khai Prototype:* Xây dựng phiên bản Minimum Viable Product (MVP) để kiểm chứng tính khả thi của giải pháp công nghệ.
- ◊ *Đánh giá định lượng:* Sử dụng các chỉ số thống kê như Mean Absolute Error (MAE), Root Mean Squared Error (RMSE) và thời gian phản hồi (Response Time) để đo lường hiệu quả của mô hình AI.

1.4 Tổng quan đề tài

1.4.1 Các nghiên cứu và giải pháp liên quan

Trong bối cảnh chuyển đổi số giáo dục đang diễn ra mạnh mẽ, nghiên cứu này kế thừa và phát triển dựa trên bốn trụ cột công nghệ chính: Mô hình Bản sao số (Digital Twin), Hệ thống học tập thích ứng (Adaptive Learning), Trí tuệ nhân tạo tạo sinh (Generative AI), và các kỹ thuật xử lý dữ liệu khuyết thiếu. Dưới đây là phân tích tổng quan về các nghiên cứu gần đây và cách tiếp cận giải quyết vấn đề của EduTwin.

1.4.1.1 Digital Twin trong Giáo dục: Từ mô phỏng môi trường đến bản sao người học

Khái niệm Digital Twin ban đầu được phát triển trong lĩnh vực công nghiệp, tuy nhiên những năm gần đây đã chứng kiến sự dịch chuyển mạnh mẽ của công nghệ này sang lĩnh vực giáo dục và đào tạo. Các nghiên cứu tiên phong trong năm 2024 và 2025 đã minh chứng tiềm năng to lớn của Digital Twin trong việc tạo ra các môi trường thực tế ảo an toàn và hiệu quả. Điển hình, [2] và [3] đã áp dụng thành công Digital Twin để hỗ trợ đào tạo trong ngành xây dựng và kiến trúc cảnh quan, cho phép người học tương tác với các kịch bản mô phỏng phức tạp. Tương tự, trong lĩnh vực y tế, [4] đã phát triển bản sao số của bệnh nhân để nâng cao năng lực chẩn đoán cho sinh viên y khoa mà không gây rủi ro trên người thật. Mở rộng hơn về mặt phương pháp luận, [5] và [6] đề xuất tích hợp Digital Twin vào các mô hình học tập truy vấn (inquiry learning) và E-learning để tăng cường tính tương tác.

CHƯƠNG 1. TỔNG QUAN VỀ ĐỀ TÀI

Tuy nhiên, điểm hạn chế căn bản của đa số các nghiên cứu hiện hành là sự tập trung quá mức vào việc mô phỏng "đối tượng bên ngoài" (như máy móc, công trình, hoặc bệnh nhân) thay vì chính bản thân người học. Các hệ thống này thường vận hành như những mô phỏng tĩnh hoặc bán động, thiếu vắng sự kết nối dữ liệu sinh trắc và tâm lý thời gian thực của chủ thể học tập. Khắc phục khoảng trống này, EduTwin đề xuất một cách tiếp cận lấy người học làm trung tâm, trong đó Digital Twin không phải là môi trường học, mà là bản sao kỹ thuật số của chính học sinh. Bằng cách coi dữ liệu điểm số và hành vi tương tác là các tín hiệu đầu vào liên tục, EduTwin xây dựng một thực thể số sống động, có khả năng tiến hóa song song và phản ánh chính xác trạng thái năng lực hiện tại của người học.

1.4.1.2 Học tập thích ứng và Bài toán dự báo hiệu suất: Hướng tiếp cận Lazy Learning

Mục tiêu quan trọng nhất của giáo dục cá nhân hóa là khả năng thích ứng theo thời gian thực với nhu cầu của người học. Các hệ thống gia sư thông minh - Intelligent Tutoring Systems (ITS) hiện đại đang nỗ lực giải quyết bài toán này thông qua việc phân tích dữ liệu lớn để đưa ra phản hồi cá nhân hóa, như được chỉ ra trong các nghiên cứu của [7] và [8]. Một số hệ thống tiên tiến hơn, như Tutomat được đề cập bởi [9], đã bắt đầu tích hợp xử lý ngôn ngữ tự nhiên để cải thiện giao diện tương tác, hay các nỗ lực tối ưu hóa thời gian thực bằng điện toán biên của [10].

Để đạt được khả năng thích ứng, nền tảng cốt lõi nằm ở độ chính xác của các mô hình dự báo hiệu suất. Xu hướng chủ đạo hiện nay, theo tổng hợp của [11] và [12], là sử dụng các kỹ thuật Eager Learning, đặc biệt là Học sâu (Deep Learning) để dự đoán kết quả học tập. Mặc dù đạt hiệu suất cao, cách tiếp cận Eager Learning (như LSTM, Neural Networks) bộc lộ nhược điểm khi áp dụng vào môi trường giáo dục phổ thông: tính "hộp đen" khó giải thích và chi phí huấn luyện lại (retraining cost) quá lớn khi dữ liệu thay đổi liên tục. Hơn nữa, với các chuỗi dữ liệu học tập ngắn và ngắt quãng, các mô hình phức tạp thường không phát huy được ưu thế so với các phương pháp đơn giản hơn nhưng mạnh về cấu trúc cục bộ.

Đối mặt với thách thức này, EduTwin lựa chọn hướng đi khác biệt với chiến lược Lazy Learning (KNN, KR, LWLR). Thay vì cố gắng khai quật hóa toàn bộ dữ liệu vào một mô hình tĩnh, hệ thống trì hoãn quá trình tính toán đến thời điểm dự báo. Cách tiếp cận này không chỉ loại bỏ hoàn toàn độ trễ huấn luyện (zero-latency adaptation) mà còn đảm bảo tính minh bạch, khi mọi dự báo đều có thể được giải thích thông qua việc tham chiếu đến các hồ sơ "láng giềng" tương đồng trong quá khứ.

1.4.1.3 Generative AI và LLM: Từ công cụ tạo sinh đến trợ lý ngữ cảnh

Sự bùng nổ của Generative AI đã mở ra một kỷ nguyên mới cho giáo dục thông minh. [13] và [14] nhận định rằng các Mô hình Ngôn ngữ Lớn (LLM) có tiềm năng chuyển đổi vai trò từ công cụ hỗ trợ thụ động sang các tác nhân chủ động. Các ứng dụng thực tiễn đã được triển khai đa dạng, từ việc tạo đề thi tự động [15] đến hỗ trợ đào tạo chuyên sâu trong các ngành kỹ thuật [16].

Tuy nhiên, việc tích hợp LLM vào môi trường giáo dục đối mặt với hai rào cản lớn: hiện tượng "ảo giác" (hallucination) và rủi ro bảo mật thông tin cá nhân [17]. Một trợ lý ảo đơn thuần sử dụng dữ liệu huấn luyện đại trà sẽ thiếu sự thấu hiểu về ngữ cảnh cụ thể của từng học sinh, dẫn đến các lời khuyên chung chung hoặc thiếu chính xác. EduTwin giải quyết vấn đề này thông qua kiến trúc RAG kết hợp với cơ chế PII Redaction. Hệ thống không cho phép LLM "sáng tác" thông tin về học lực, mà buộc mô hình phải sinh câu trả lời dựa trên dữ liệu điểm số thực tế được truy xuất từ Digital Twin. Đồng thời, lớp bảo mật trung gian đảm bảo danh tính học sinh được ẩn danh hóa trước khi xử lý, tạo ra một môi trường tương tác vừa thông minh, vừa an toàn và mang tính cá nhân hóa sâu sắc.

1.4.1.4 Xử lý dữ liệu khuyết thiếu trong hồ sơ học tập

Dữ liệu giáo dục trong thực tế hiếm khi hoàn hảo; tình trạng dữ liệu thưa (sparse data) và khuyết thiếu là vấn đề thường trực ảnh hưởng đến độ chính xác của các thuật toán. Các phương pháp truyền thống như thay thế bằng giá trị trung bình (mean imputation) đã được chứng minh là không hiệu quả trong việc bảo toàn cấu trúc dữ liệu phức tạp. Các nghiên cứu mới nhất từ [18] và [19] đã khẳng định ưu thế vượt trội của các kỹ thuật điền khuyết dựa trên máy học. Đặc biệt, [20] và [21] chỉ ra rằng phương pháp k-Nearest Neighbors Imputation (KNNI) giúp cải thiện đáng kể độ chính xác của các bài toán dự báo nhờ khả năng tận dụng thông tin từ các mẫu tương đồng.

Vận dụng các kết quả này, EduTwin thiết lập quy trình xử lý dữ liệu "Fill-then-Predict". Thay vì loại bỏ các bản ghi thiếu hoặc điền giá trị ngẫu nhiên, hệ thống sử dụng KNNI để tái tạo các điểm số bị khuyết dựa trên mối tương quan đa biến giữa các môn học. Chiến lược này giúp khôi phục một hồ sơ năng lực hoàn chỉnh nhất có thể trước khi đưa vào các mô hình dự báo xu hướng, đảm bảo tính ổn định và độ tin cậy của toàn bộ hệ thống ngay cả khi dữ liệu đầu vào không liền mạch.

1.4.2 Đóng góp mới của đề tài

EduTwin khắc phục các hạn chế của các hệ thống truyền thống và đóng góp vào lĩnh vực Công nghệ Giáo dục (EdTech) thông qua ba điểm mới cốt lõi sau:

- **Kiến trúc Dự báo Thích ứng Hiệu năng cao (High-Performance Adaptive Prediction Architecture):** Đề tài hiện thực hóa quy trình xử lý dữ liệu toàn diện theo cơ chế "**Fill-then-Predict**", kết hợp giữa kỹ thuật điền khuyết **KNNI** và thuật toán học lười (**KNN, KR, LWLR**). Điểm đột phá nằm ở việc ứng dụng kỹ thuật **Cluster Indexing (Phân cụm đánh chỉ mục)** và **Tuyển chọn mẫu hình (Prototype Selection)** để nén không gian tìm kiếm. Giải pháp này đã được kiểm chứng thực nghiệm giúp tăng tốc độ phản hồi lên hàng trăm lần ở quy mô dữ liệu lớn (Big Data), giải quyết triệt để bài toán độ trễ của Lazy Learning mà không cần hạ tầng GPU đắt đỏ.
- **Kiến trúc Hệ thống hướng Siêu dữ liệu (Metadata-Driven Architecture):** Khác biệt với các LMS truyền thống, EduTwin giới thiệu mô hình "Cấu trúc động" (Dynamic Structure). Đóng góp này nằm ở việc tách biệt hoàn toàn tầng "Logic xử lý nghiệp vụ" khỏi tầng "Định nghĩa dữ liệu" sử dụng công nghệ **PostgreSQL JSONB**. Điều này cho phép các cơ sở giáo dục tự do định nghĩa các mô hình đánh giá phức tạp (đa cấp độ, đa trọng số) mà hệ thống AI vẫn tự động thích nghi và xử lý chính xác thông qua cơ chế ánh xạ động.
- **Cơ chế Tương tác Context-Aware RAG Bảo mật:** Đề tài xây dựng quy trình tích hợp LLM chuyên sâu cho giáo dục, kết hợp dữ liệu điểm số định lượng với hồ sơ hành vi định tính. Đặc biệt, hệ thống tích hợp module **PII Redaction** và mã hóa trường dữ liệu (Field-level Encryption), đảm bảo tuyệt đối an toàn thông tin định danh cho học sinh khi tương tác với AI. Bằng cách đưa dữ liệu thực tế vào ngữ cảnh thông qua kỹ thuật RAG, hệ thống chuyển đổi vai trò của Chatbot từ "trả lời câu hỏi" sang "tư vấn cá nhân hóa" với độ chính xác và tin cậy cao.

1.5 Cấu trúc Khoa luận tốt nghiệp

Khóa luận với đề tài “EduTwin” được trình bày bao gồm 5 chương. Nội dung tóm tắt từng chương được trình bày như sau:

- **Chương 1: Tổng quan về đề tài.** Tổng quan về đề tài. Đặt vấn đề, xác định bài toán, mục tiêu và phạm vi nghiên cứu, đồng thời nêu bật tính cấp thiết và đóng góp của đề

CHƯƠNG 1. TỔNG QUAN VỀ ĐỀ TÀI

tài.

- **Chương 2: Cơ sở lý thuyết.** Trình bày các định nghĩa hình thức về Digital Twin, cơ sở toán học của các thuật toán Lazy Learning (KNN, KR, LWLR), cơ chế Attention trong LLM, các kỹ thuật điền khuyết dữ liệu (Imputation) được sử dụng làm nền tảng xây dựng hệ thống và các công nghệ phát triển Modern Web App.
- **Chương 3: Phân tích và Thiết kế hệ thống.** Mô tả chi tiết quy trình nghiệp vụ, kiến trúc phần mềm và thiết kế cơ sở dữ liệu quan hệ.
- **Chương 4: Thực nghiệm và Đánh giá.** Trình bày quá trình xây dựng mã nguồn (Implementation), minh họa các chức năng chính và phân tích kết quả thực nghiệm trên bộ dữ liệu mẫu.
- **Chương 5: Kết luận và Hướng phát triển.** Tổng kết các kết quả đạt được, nhìn nhận thẳng thắn các hạn chế và đề xuất lộ trình nâng cấp sản phẩm trong tương lai.

Chương 2. CƠ SỞ LÝ THUYẾT

Chương này trình bày hệ thống cơ sở lý thuyết khoa học và các giải pháp công nghệ được áp dụng trong việc xây dựng EduTwin. Nội dung tập trung phân tích sâu về mô hình Digital Twin trong giáo dục, các thuật toán học máy Lazy Learning và kiến trúc hệ thống, đồng thời lý giải chi tiết các quyết định kỹ thuật dựa trên đặc thù của bài toán.

2.1 Digital Twin và Học tập Cá nhân hóa

Digital Twin trong Bối cảnh Giáo dục. Về bản chất, Digital Twin (Bản sao số) là sự biểu diễn ảo hóa (virtual representation) của một thực thể, quy trình hoặc hệ thống trong thế giới thực. Điểm cốt lõi phân biệt Digital Twin với các mô hình mô phỏng tĩnh (simulation) nằm ở kết nối dữ liệu liên tục: luồng dữ liệu từ thực thể vật lý được truyền tải thời gian thực để cập nhật bản sao số, đảm bảo bản sao này luôn phản ánh chính xác trạng thái hiện tại của thực thể. Ngược lại, những tri thức (insights) phân tích từ bản sao số sẽ được dùng để hỗ trợ ra quyết định hoặc tác động tối ưu hóa lại thực thể thực.

EduTwin vận dụng nguyên lý này để xây dựng "Bản sao số người học". Trong mô hình này, người học đóng vai trò là thực thể vật lý, còn điểm số và hành vi tương tác là các tín hiệu đầu vào (thay cho cảm biến Internet of Things (IoT) trong công nghiệp). Không giống như các hồ sơ dữ liệu truyền thống vốn chỉ lưu trữ thông tin tĩnh và rời rạc (học bạ, điểm số quá khứ), Digital Twin trong EduTwin đóng vai trò là một thực thể số song song, có khả năng phản ánh trạng thái thực của người học. Hệ thống này hoạt động dựa trên cơ chế vòng lặp phản hồi kín (Closed-loop feedback): dữ liệu từ người học (điểm số, tương tác) liên tục được cập nhật vào bản sao số; bản sao số sử dụng các mô hình tính toán để dự báo và đưa ra các kịch bản tối ưu; cuối cùng, các tác động sư phạm được áp dụng ngược lại cho người học thực.

Tóm lại, EduTwin vận dụng triệt để nguyên lý này để chuyển đổi từ mô hình quản lý hồ sơ tĩnh sang một hệ thống Bản sao số động, nơi dữ liệu được luân chuyển liên tục trong một vòng lặp. Cách tiếp cận này giúp đồng bộ hóa trạng thái năng lực của người học, tạo tiền đề khoa học vững chắc cho việc cá nhân hóa và tối ưu hóa lộ trình phát triển.

Học tập Thích ứng (Adaptive Learning). Học tập cá nhân hóa, hay còn gọi là học tập thích ứng (adaptive learning), là phương pháp giáo dục sử dụng thuật toán máy tính để điều chỉnh nội dung, phương pháp giảng dạy và lộ trình học dựa trên nhu cầu, trình độ và phong cách học tập riêng biệt của từng cá nhân. Thay vì áp dụng mô hình "one-size-fits-all", hệ

CHƯƠNG 2. CƠ SỞ LÝ THUYẾT

thông EduTwin phân tích dữ liệu lịch sử để xác định chính xác "Vùng phát triển gần" Zone of Proximal Development (ZPD) theo lý thuyết của Lev Vygotsky (1978).

Quy trình cá nhân hóa được thực hiện bằng cách so khớp hồ sơ năng lực hiện tại của học sinh với cơ sở dữ liệu, sử dụng phương pháp tham chiếu dựa trên đồng nghiệp (peer-based reference). Việc cá nhân hóa này được thực hiện thông qua việc liên tục so khớp hồ sơ năng lực của học sinh hiện tại với cơ sở dữ liệu lịch sử khổng lồ, từ đó tìm ra các lô trình thành công của những người học có đặc điểm tương đồng. Theo Brusilovsky và Millán (2007), học tập cá nhân hóa có thể cải thiện tỷ lệ hoàn thành khóa học lên đến 15–25%.

2.2 Lazy Learning

Trong bối cảnh dự án EduTwin, việc mô hình hóa năng lực và dự báo quỹ đạo học tập của học sinh đòi hỏi một phương pháp tiếp cận có khả năng thích ứng cao với sự biến động liên tục của dữ liệu. Thay vì sử dụng các mô hình tham số toàn cục (global parametric models) cố định, nghiên cứu này áp dụng chiến lược Học lười (Lazy Learning), cụ thể là Học dựa trên thể hiện Instance-based Learning (IBL).

Cơ sở lý thuyết của chiến lược này dựa trên giả định về tính trơn cục bộ (local smoothness): các đầu ra của hàm mục tiêu thay đổi một cách mượt mà theo các biến đầu vào, do đó các điểm dữ liệu nằm gần nhau trong không gian đặc trưng sẽ có giá trị đích tương đồng.

Khác với các phương pháp "học hăng hái" (eager learning) cố gắng tìm kiếm một hàm tổng quát $f(x)$ trên toàn bộ không gian dữ liệu ngay trong quá trình huấn luyện, các phương pháp IBL chỉ lưu trữ dữ liệu huấn luyện và trì hoãn quá trình tính toán đến thời điểm dự báo. Khi một truy vấn (query point) xuất hiện, hệ thống sẽ xây dựng một hàm xấp xỉ cục bộ (local approximation) dựa trên tập hợp các láng giềng gần nhất.

Ba mô hình được sử dụng trong dự án này — KNN, KR, LWLR— đại diện cho sự phát triển tuần tự về độ phức tạp của hàm xấp xỉ cục bộ này.

KNN: Xấp xỉ Hằng số Cục bộ

Đây là dạng cơ bản nhất của học lười. Về mặt lý thuyết, KNN xấp xỉ hàm mục tiêu $f(x)$ bằng một hàm hằng số (constant function) trong vùng lân cận của điểm truy vấn. Đối với bài toán dự báo điểm số (hồi quy), giá trị dự đoán \hat{y} tại điểm truy vấn x_q được tính là trung bình cộng của k láng giềng gần nhất:

$$\hat{y}(x_q) = \frac{1}{k} \sum_{x_i \in \mathcal{N}_k(x_q)} y_i$$

CHƯƠNG 2. CƠ SỞ LÝ THUYẾT

Trong đó, $\mathcal{N}_k(x_q)$ là tập hợp k điểm dữ liệu trong tập huấn luyện có khoảng cách gần nhất với x_q theo một thước đo xác định.

KR: Xấp xỉ Trung bình có Trọng số

Khắc phục hạn chế về sự thay đổi rời rạc của KNN, KR đưa ra một hàm xấp xỉ mượt mà hơn. Mô hình này gán trọng số w_i cho mọi điểm dữ liệu dựa trên khoảng cách của nó tới điểm truy vấn thông qua một hàm nhân (Kernel function) K_h :

$$\hat{y} = \frac{\sum_{i=1}^n K_h(d(x, x_i))y_i}{\sum_{i=1}^n K_h(d(x, x_i))}$$

Tham số quan trọng nhất ở đây là độ rộng dải (bandwidth h). Nó quyết định bán kính ảnh hưởng hiệu quả của các điểm lân cận: h nhỏ giúp mô hình nhạy bén với dữ liệu cục bộ, trong khi h lớn giúp làm trơn nhiễu.

LWLR Xấp xỉ Tuyến tính Cục bộ

LWLR là bước tiến cao nhất trong nhóm mô hình này. Thay vì xấp xỉ bằng một giá trị trung bình, LWLR xấp xỉ hàm mục tiêu bằng một hàm tuyến tính $y = \theta^T x$ chỉ có giá trị cục bộ. Tại mỗi điểm truy vấn x_q , thuật toán tìm tham số θ tối ưu bằng cách cực tiểu hóa hàm mất mát bình phương sai số có trọng số:

$$\min_{\theta} \sum_i w^{(i)} (y^{(i)} - \theta^T x^{(i)})^2$$

Trong đó, trọng số $w^{(i)}$ thường được tính qua hàm Gaussian Kernel với tham số độ rộng dải τ :

$$w^{(i)} = \exp \left(-\frac{\|x^{(i)} - x_q\|^2}{2\tau^2} \right)$$

Phương pháp này cho phép mô hình nắm bắt được cả "xu hướng" (đạo hàm) của dữ liệu tại điểm xét, mang lại độ chính xác cao hơn cho các quỹ đạo học tập phức tạp.

Tổng hợp lại, việc kết hợp ba mô hình này cho phép hệ thống EduTwin khai thác triệt để thông tin từ dữ liệu quá khứ theo các mức độ chi tiết khác nhau: từ tìm kiếm tương đồng đơn giản (KNN), làm trơn dữ liệu (KR) đến mô hình hóa xu hướng cục bộ (LWLR).

2.3 Lượng tử hóa Vector và Tối ưu hóa Mẫu hình

Để giải quyết thách thức về khả năng mở rộng (scalability) và tốc độ phản hồi của các thuật toán Học lười (Lazy Learning) khi áp dụng trên quy mô lớn, nghiên cứu này đề xuất sử

CHƯƠNG 2. CƠ SỞ LÝ THUYẾT

dụng kỹ thuật Lượng tử hóa Vector kết hợp với chiến lược Tuyển chọn Mẫu hình Thích ứng (Adaptive Prototype Selection). Đây là phương pháp nén dữ liệu có tổn hao (lossy compression) nhằm xấp xỉ phân phối của dữ liệu gốc bằng một tập hợp hữu hạn các vector đại diện, giúp cân bằng giữa độ chính xác dự báo và hiệu suất tính toán. Lượng tử hóa Vector là quá trình ánh xạ không gian dữ liệu đầu vào liên tục hoặc rời rạc $X \in \mathbb{R}^d$ vào một tập hợp rời rạc hữu hạn $C = \{\mu_1, \mu_2, \dots, \mu_K\}$ gọi là bảng mã (codebook), trong đó mỗi vector μ_k được gọi là một mẫu hình (prototype) hoặc tâm cụm (centroid). Trong khuôn khổ của đề tài, kỹ thuật được hiện thực hóa cụ thể thông qua thuật toán **K-Means Clustering** để tạo ra các tập mẫu hình (Prototypes). Mục tiêu cốt lõi là phân hoạch không gian dữ liệu học sinh thành K vùng Voronoi riêng biệt (V_1, \dots, V_K), sao cho mọi điểm dữ liệu trong cùng một vùng đều được đại diện bởi mẫu hình μ_k tương ứng. Thuật toán K-Means thực hiện điều này bằng cách tối thiểu hóa hàm mục tiêu biến dạng (Distortion Function) J , đại diện cho tổng bình phương sai số lượng tử hóa:

$$J = \sum_{n=1}^N \sum_{k=1}^K r_{nk} \|x^{(n)} - \mu_k\|^2$$

Trong đó: $x^{(n)}$ là vector đặc trưng của học sinh thứ n ; μ_k là vector đại diện (prototype) của cụm thứ k ; r_{nk} là biến chỉ thị (indicator variable), nhận giá trị 1 nếu $x^{(n)}$ thuộc cụm k , và 0 nếu ngược lại. Quá trình tối ưu hóa hàm J diễn ra theo cơ chế lặp Expectation–Maximization (EM): **Bước Gán (Assignment Step)**:

$$r_{nk} = \begin{cases} 1 & \text{nếu } k = \operatorname{argmin}_j \|x^{(n)} - \mu_j\|^2 \\ 0 & \text{trường hợp khác} \end{cases}$$

Bước Cập nhật (Update Step):

$$\mu_k = \frac{\sum_{n=1}^N r_{nk} x^{(n)}}{\sum_{n=1}^N r_{nk}}$$

Việc áp dụng K-Means như một phương pháp Lượng tử hóa Vector tạo tiền đề cho kiến trúc lai ghép 2 tầng (2-tier architecture) được thiết kế chi tiết trong chương sau: **Tầng 1: Phân cụm Toàn cục (Global Clustering/Indexing)**. Hệ thống sử dụng các vector μ_k từ quá trình như một cơ chế "đánh chỉ mục" (indexing). Thay vì tìm kiếm trên toàn bộ không gian N mẫu dữ liệu, hệ thống chỉ cần xác định cụm gần nhất với query. **Tầng 2: Tuyển chọn Mẫu**

CHƯƠNG 2. CƠ SỞ LÝ THUYẾT

hình Thích ứng (Adaptive Prototype Selection). Tại thời điểm dự đoán, hệ thống thực hiện tuyển chọn động T mẫu hình ($T = 3000$) từ cụm được gán. Các mẫu hình ưu tiên là những điểm dữ liệu thực tế nằm gần tâm cụm μ_k nhất:

$$P_{k^*} = \{x_1, x_2, \dots, x_T\} \text{ với } \|x_i - \mu_{k^*}\| \leq \|x_{i+1} - \mu_{k^*}\|$$

Nếu cụm C_{k^*} có ít hơn T mẫu, hệ thống tự động gộp thêm mẫu từ các cụm lân cận (Neighbor Merging):

$$P_{k^*} = C_{k^*} \cup \bigcup_{j \in \mathcal{N}(k^*)} C_j \quad \text{cho đến khi } |P_{k^*}| \geq T$$

Trong đó $\mathcal{N}(k^*)$ là tập các cụm lân cận được sắp xếp theo khoảng cách centroid. Tại thời điểm dự báo (runtime), quy trình suy diễn sẽ được thực hiện qua 3 bước:

1. **Định vị (Localization):** Xác định cụm k^* mà học sinh mới x_{query} thuộc về bằng cách so khớp với bộ codebook $\{\mu_1, \dots, \mu_K\}$ (độ phức tạp $O(K)$).
2. **Tuyển chọn Mẫu hình (Prototype Selection):** Thu thập $T = 3000$ mẫu từ cụm k^* hoặc gộp từ các cụm lân cận nếu cần.
3. **Hồi quy Cục bộ (Local Regression):** Áp dụng thuật toán học lười (KNN, KR hoặc LWLR) trên tập mẫu hình P_{k^*} .

Phương pháp tiếp cận này mang lại hiệu quả kép: Tốc độ suy diễn tiệm cận thời gian thực (do $K + T \ll N$) và Khả năng giải thích cao (dựa trên việc tham chiếu đến nhóm mẫu hình cụ thể), khắc phục hoàn toàn nhược điểm về hiệu năng của các phương pháp học lười truyền thống.

2.4 Kiến trúc Transformer và Cơ chế Attention.

Nền tảng của các Mô hình Ngôn ngữ Lớn như Chat-GPT và Gemini là kiến trúc Transformer, được giới thiệu bởi Vaswani et al. (2017) trong nghiên cứu "Attention Is All You Need". Điểm đột phá của kiến trúc này nằm ở cơ chế Tự chú ý (Self-Attention), cho phép mô hình "nhìn" đồng thời toàn bộ chuỗi đầu vào thay vì xử lý tuần tự như các mạng hồi quy truyền thống.

Cơ chế Attention tính toán mức độ liên quan giữa các token trong chuỗi thông qua ba ma trận Query (Q), Key (K) và Value (V):

$$\text{Attention}(Q, K, V) = \text{softmax} \left(\frac{QK^T}{\sqrt{d_k}} \right) V$$

Trong đó d_k là chiều của vector Key, đóng vai trò hệ số chuẩn hóa để ổn định gradient. Nhờ cơ chế này, khi học sinh hỏi về "môn Toán", mô hình có thể tự động liên kết với các thông tin liên quan như điểm số, chương trình học, và lịch sử tương tác trước đó trong ngữ cảnh.

Quản lý Bộ nhớ Hội thoại (Conversation Memory Management). Một thách thức quan trọng khi triển khai chatbot là giới hạn về độ dài ngữ cảnh (context window) của các mô hình LLM. Ví dụ, GPT-4 Turbo có giới hạn khoảng 128.000 token, trong khi Gemini Pro hỗ trợ đến 1 triệu token. Tuy nhiên, chi phí Application Programming Interface (API) tăng tuyến tính theo số token sử dụng.

EduTwin áp dụng chiến lược Cửa sổ Trượt (Sliding Window) kết hợp với Tóm tắt Tích lũy (Cumulative Summarization) để quản lý lịch sử hội thoại:

- **Sliding Window:** Chỉ giữ lại n tin nhắn gần nhất trong bộ nhớ ngắn hạn (short-term memory).
- **Cumulative Summarization:** Các tin nhắn cũ hơn được tóm tắt thành một đoạn văn ngắn và lưu vào bộ nhớ dài hạn (long-term memory), đảm bảo không mất thông tin quan trọng về học sinh.

Cách tiếp cận này giúp cân bằng giữa chi phí vận hành và chất lượng phản hồi, đồng thời đảm bảo tính liên tục của cuộc trò chuyện xuyên suốt phiên làm việc.

Truyền tải Phản hồi Streaming (Server-Sent Events). Do các mô hình LLM tạo ra phản hồi theo từng token, việc chờ đợi toàn bộ response hoàn thành sẽ gây ra độ trễ cảm nhận (perceived latency) lớn. EduTwin triển khai kỹ thuật để truyền tải phản hồi theo thời gian thực. là giao thức một chiều (server → client) dựa trên Hypertext Transfer Protocol (HTTP), cho phép server "đẩy" dữ liệu liên tục mà không cần client polling. Mỗi token được sinh ra bởi LLM sẽ được gửi ngay lập tức đến frontend:

```
Content-Type: text/event-stream
data: {"token": "Xin"}
data: {"token": " chào"}
data: {"token": " bạn"}
data: [DONE]
```

CHƯƠNG 2. CƠ SỞ LÝ THUYẾT

Ưu điểm của so với WebSocket: đơn giản hơn trong triển khai, tương thích tốt với HTTP/2, và tự động reconnect khi mất kết nối.

2.5 Xử lý Dữ liệu và Kỹ thuật Imputation

Trong thực tế vận hành, dữ liệu giáo dục thường xuyên đối mặt với vấn đề khuyết thiêу (sparse data). Nghiên cứu đề xuất sử dụng kỹ thuật KNNI để xử lý vấn đề này.

KNN Imputation là phương pháp điền khuyết đa biến, dựa trên giả định rằng các điểm dữ liệu gần nhau trong không gian đặc trưng sẽ có các giá trị thành phần tương tự nhau. Về mặt lý thuyết, giá trị bị khuyết x_j của một điểm dữ liệu x sẽ được ước lượng dựa trên thông tin từ k láng giềng gần nhất tìm thấy trong tập huấn luyện:

$$\hat{x}_j = \frac{\sum_{y \in \mathcal{N}_k(x)} w(x, y) \cdot y_j}{\sum_{y \in \mathcal{N}_k(x)} w(x, y)}$$

Trong đó trọng số $w(x, y)$ thường được chọn tỉ lệ nghịch với khoảng cách Euclidean chuẩn $d(x, y)$:

$$d(x, y) = \sqrt{\sum_{i=1}^N (x_i - y_i)^2}$$

Việc áp dụng KNN Imputation giúp tái tạo một vector đầu vào hoàn chỉnh, đảm bảo tính ổn định cho các thuật toán dự báo phức tạp ở giai đoạn sau (như LWLR) mà không làm méo mó cấu trúc tương quan giữa các môn học.

2.6 Các chỉ số Đánh giá Hiệu năng Mô hình

Các chỉ số đánh giá là công cụ quan trọng giúp đo lường, so sánh và xác định mức độ hiệu quả của các mô hình học máy. Việc lựa chọn chỉ số phù hợp không chỉ phản ánh đúng bản chất bài toán mà còn đảm bảo tính khách quan trong việc đánh giá hiệu suất dự đoán. Trong nghiên cứu này, tùy theo loại mô hình và đầu ra (liên tục hoặc phân loại), các chỉ số được sử dụng bao gồm:

2.6.1 Sai số tuyệt đối trung bình (MAE - Mean Absolute Error)

Công thức tính:

$$\text{MAE} = \frac{1}{n} \sum_{i=1}^n |y_i - \hat{y}_i|$$

Sai số tuyệt đối trung bình đo lường khoảng cách tuyệt đối trung bình giữa giá trị thực tế và giá trị dự đoán, phản ánh mức độ sai lệch trung bình mà mô hình tạo ra. Chỉ số này có ưu điểm dễ hiểu, đơn giản trong tính toán và ít bị ảnh hưởng bởi các giá trị ngoại lai so với các chỉ số khác. Về mặt diễn giải, MAE càng nhỏ chứng tỏ mô hình dự đoán càng chính xác và ổn định.

2.6.2 Căn bậc hai sai số bình phương trung bình (RMSE - Root Mean Squared Error)

Công thức tính:

$$\text{RMSE} = \sqrt{\frac{1}{n} \sum_{i=1}^n (y_i - \hat{y}_i)^2}$$

Căn bậc hai sai số bình phương trung bình có ý nghĩa tương tự như MAE nhưng nhẫn mạnh nhiều hơn vào các sai số lớn thông qua việc bình phương độ lệch giữa giá trị thực tế và giá trị dự đoán. RMSE đặc biệt nhạy cảm với các giá trị ngoại lai, do đó thường được sử dụng trong các bài toán mà sai số lớn cần được kiểm soát chặt chẽ. Trong quá trình đánh giá mô hình, RMSE càng nhỏ đồng nghĩa với hiệu suất dự đoán của mô hình càng cao và mức độ rủi ro sai số lớn càng thấp.

2.6.3 Hệ số xác định (R^2 - R-squared)

Công thức:

$$R^2 = 1 - \frac{\sum_{i=1}^n (y_i - \hat{y}_i)^2}{\sum_{i=1}^n (y_i - \bar{y})^2}$$

Hệ số xác định đo lường mức độ mà mô hình có thể giải thích phuơng sai trong biến mục tiêu, với \bar{y} là giá trị trung bình của biến cần dự đoán. Chỉ số này cho biết tỷ lệ biến thiên của dữ liệu đầu ra được mô hình lý giải thành công. R^2 có giá trị nằm trong khoảng từ 0 đến 1, trong đó giá trị càng tiến gần 1 chứng tỏ mô hình càng phù hợp, có khả năng giải thích tốt sự biến động của dữ liệu thực tế. Như vậy, R^2 càng cao đồng nghĩa với việc mô hình có chất lượng dự đoán càng tốt.

2.7 Trực quan hóa Dữ liệu Giáo dục.

Việc lựa chọn loại biểu đồ phù hợp đóng vai trò then chốt trong việc truyền tải thông tin học tập một cách hiệu quả. Hệ thống EduTwin áp dụng ba loại biểu đồ chính, mỗi loại phục vụ một mục đích phân tích riêng biệt:

Biểu đồ Radar (Radar Chart) được sử dụng để so sánh năng lực đa chiều của học sinh trên nhiều môn học đồng thời. Loại biểu đồ này cho phép nhận diện nhanh các điểm mạnh và điểm yếu tương đối, phù hợp với bài toán phân tích hồ sơ năng lực toàn diện. Trong EduTwin, mỗi trục của biểu đồ Radar đại diện cho một môn học, và diện tích vùng phủ phản ánh mức độ cân bằng của năng lực.

Biểu đồ Đường (Line Chart) được tối ưu cho việc thể hiện xu hướng biến đổi theo thời gian. Đây là công cụ chủ đạo trong việc trực quan hóa quỹ đạo học tập (learning trajectory), cho phép học sinh và giáo viên theo dõi tiến bộ qua các kỳ học hoặc năm học. Khả năng nối các điểm dữ liệu liên tiếp giúp người xem dễ dàng nhận ra các mẫu xu hướng tăng, giảm hoặc ổn định.

Biểu đồ Cột (Bar Chart) phù hợp cho việc so sánh trực tiếp giữa các danh mục rời rạc, ví dụ như so sánh điểm số giữa các môn trong cùng một kỳ học, hoặc so sánh hiệu suất giữa các khối thi (Tự nhiên, Xã hội). Tính trực quan của biểu đồ cột giúp học sinh nhanh chóng xác định môn học cần cải thiện.

Việc kết hợp ba loại biểu đồ này tạo ra một hệ thống trực quan hóa đa tầng, cho phép học sinh tiếp cận dữ liệu học tập từ nhiều góc độ khác nhau: từ góc nhìn tổng thể (Radar), đến phân tích xu hướng (Line), và so sánh chi tiết (Bar).

2.8 Xác thực API với JSON Web Token (JWT)

Trong kiến trúc RESTful, việc xác thực trạng thái (stateful authentication) như Session-based không phù hợp với tính chất phi trạng thái (stateless) của HTTP. JWT là giải pháp tiêu chuẩn công nghiệp, cho phép client tự chứng minh danh tính mà không cần server lưu trữ phiên làm việc. **Cấu trúc JWT:** Một token JWT bao gồm ba phần được mã hóa Base64 và nối bằng dấu chấm:

- **Header:** Chứa thuật toán ký (HS256) và loại token.
- **Payload:** Chứa các claims như user_id, exp (thời hạn), iat (thời điểm tạo).

CHƯƠNG 2. CƠ SỞ LÝ THUYẾT

- **Signature:** Chữ ký số được tạo bằng: HMAC-SHA256 (header.payload, SECRET_KEY).

Khi client gửi request đến các API được bảo vệ, token được đính kèm trong HTTP Header:

Authorization: Bearer eyJhbGciOiJIUzI1NiIs...

Server xác thực token bằng cách kiểm tra chữ ký và thời hạn, sau đó trích xuất user_id để xử lý yêu cầu mà không cần truy vấn database phiên.

Chương 3. PHÂN TÍCH VÀ THIẾT KẾ HỆ THỐNG

Chương này trình bày chi tiết quá trình phân tích yêu cầu và thiết kế hệ thống EduTwin. Nội dung bao gồm việc xác định bài toán, mô hình hóa quy trình nghiệp vụ, thiết kế kiến trúc tổng thể, cơ sở dữ liệu và các thuật toán cốt lõi.

3.1 Phát biểu bài toán

3.1.1 Bối cảnh

Trong bối cảnh giáo dục hiện đại, việc cá nhân hóa trải nghiệm học tập cho từng học sinh đang trở thành xu hướng tất yếu. Tuy nhiên, các hệ thống hỗ trợ giáo dục hiện tại (như SMAS, VnEdu, Google Classroom) chủ yếu tập trung vào việc **lưu trữ và hiển thị** kết quả học tập, thiếu khả năng **phân tích, dự đoán và tư vấn** cá nhân hóa.

3.1.2 Bài toán cần giải quyết

Mục tiêu chính là xây dựng hệ thống "Bản sao học tập kỹ thuật số" (Digital Twin for Education) có khả năng:

- Dự đoán kết quả học tập:** Sử dụng điểm số hiện tại để dự đoán điểm các học kỳ/khoa học tiếp theo với độ chính xác cao.
- Thích ứng cấu trúc động:** Hỗ trợ nhiều chương trình giáo dục khác nhau (THPT, Đại học, Trung tâm) với số lượng môn học và học kỳ tùy biến - không cố định như các mô hình Deep Learning truyền thống.
- Tư vấn thông minh:** Cung cấp Chatbot AI có khả năng phân tích tình hình học tập, đưa ra lời khuyên dựa trên dữ liệu thực tế và tăng cường cá nhân hóa thông qua việc học tập thông tin trong quá trình trò chuyện.
- Hiệu năng cao:** Đảm bảo thời gian phản hồi nhanh ($< 200ms$ cho dự đoán) ngay cả khi tập dữ liệu tham chiếu lớn.

3.1.3 Các thách thức kỹ thuật và Giải pháp

Để đạt được các mục tiêu trên, hệ thống cần giải quyết các thách thức kỹ thuật sau:

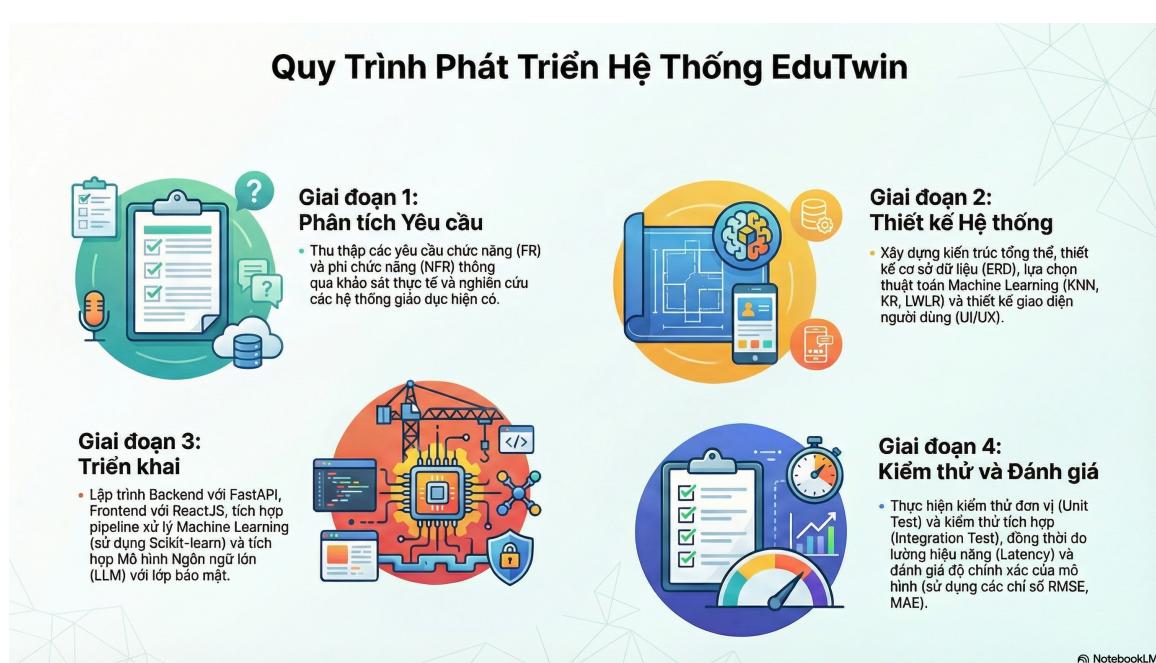
CHƯƠNG 3. PHÂN TÍCH VÀ THIẾT KẾ HỆ THỐNG

Thách thức	Giải pháp đề xuất
Input/Output không cố định	Sử dụng Lazy Learning (KNN, LWLR, KR) thay vì Deep Learning cố định kiến trúc.
Tốc độ chậm với dataset lớn	Áp dụng Phân cụm (Cluster) và Tối ưu hóa mẫu hình (Prototype Optimization - VQ).
Bảo mật thông tin cá nhân	Cơ chế PII Redaction và Mã hóa trường dữ liệu (Field Encryption).
Cấu trúc chương trình học đa Kiến trúc hướng siêu dữ liệu (Metadata-driven dạng	Architecture) với PostgreSQL JSONB.

Bảng 3.1. Các thách thức kỹ thuật và Giải pháp đề xuất

3.2 Quy trình thực hiện

Quy trình phát triển hệ thống EduTwin được chia thành 4 giai đoạn chính, tuân thủ mô hình phát triển phần mềm linh hoạt (Agile) kết hợp với quy trình nghiên cứu khoa học.



Hình 3.1. Tổng quan quy trình thực hiện dự án

Chi tiết từng giai đoạn:

- Giai đoạn 1: Phân tích yêu cầu.** Thu thập yêu cầu từ khảo sát thực tế và nghiên cứu các hệ thống hiện có để xác định danh sách yêu cầu chức năng và phi chức năng.

CHƯƠNG 3. PHÂN TÍCH VÀ THIẾT KẾ HỆ THỐNG

- **Giai đoạn 2: Thiết kế hệ thống.** Xây dựng kiến trúc tổng thể, thiết kế cơ sở dữ liệu (Entity Relationship Diagram (ERD)), thuật toán Machine Learning (ML) (KNN, KR, LWLR) và thiết kế giao diện (UI/UX).
- **Giai đoạn 3: Triển khai.** Xây dựng Backend (FastAPI), Frontend (ReactJS), tích hợp ML Pipeline (Scikit-learn) và tích hợp LLM với lớp bảo mật.
- **Giai đoạn 4: Kiểm thử và đánh giá.** Thực hiện Unit Test, Integration Test, đo lường hiệu năng (Latency) và đánh giá độ chính xác mô hình (RMSE, MAE).

3.3 Phân tích yêu cầu hệ thống

3.3.1 Yêu cầu chức năng

A. Nhóm chức năng cho Học sinh (Student Actor)

ID	Chức năng	Mô tả	Endpoint API
F1.1	Đăng ký/Đăng nhập	Xác thực với Bcrypt.	/auth/login, /auth/register
F1.2	Xem Dashboard	Tổng quan điểm số, dự đoán.	/custom-model/user-scores/{id}
F1.3	Chat với AI	Trò chuyện cá nhân hóa.	/chatbot, /chatbot/stream
F1.4	Dự đoán điểm tương lai	Nhập điểm giả định.	/custom-model/predict/{id}
F1.5	Cập nhật điểm	Nhập điểm thực tế.	POST /custom-model/user-scores

Bảng 3.2. Danh sách chức năng cho Học sinh

B. Nhóm chức năng cho Quản trị viên (Admin Actor)

C. Nhóm chức năng AI Core (System Actor)

3.3.2 Yêu cầu phi chức năng

A. Hiệu năng (Performance)

- **API Response Time (Dự đoán):** < 200ms. Giải pháp: Sử dụng Lazy Learning kết hợp VQ Clustering và Redis Cache.

CHƯƠNG 3. PHÂN TÍCH VÀ THIẾT KẾ HỆ THỐNG

ID	Chức năng	Mô tả	Endpoint API
F2.2	Cấu hình cấu trúc	Tạo/sửa/xóa Structure.	/custom-model/teaching-structure
F2.4	Upload Dataset	Tải file Excel tham chiếu.	/custom-model/upload-dataset/{id}
F2.6	Tham số ML	Điều chỉnh K, Bandwidth, Tau.	/developer/model-parameters
F2.8	Đánh giá model	So sánh RMSE/MAE.	/custom-model/evaluate-models

Bảng 3.3. Danh sách chức năng cho Admin

ID	Chức năng	Mô tả chi tiết
F3.1	Tự động điền khuyết (Im-KNN Imputer xử lý khoảng cách NaN (NaN-putation)	aware distance) khi có điểm thiếu.
F3.2	Tự động dự đoán	Lazy Learning chạy dự đoán cho các mốc thời gian tương lai khi có dữ liệu mới.
F3.4	PII Redaction	Ẩn danh hóa thông tin cá nhân (Tên, SĐT, Email) trước khi gửi đến LLM.
F3.5	Cluster + Prototype	Tối ưu tốc độ dự đoán bằng kỹ thuật phân cụm khi dữ liệu lớn (> 3000 mẫu).
F3.6	Trích xuất tài liệu	Xử lý PDF/DOCX, tóm tắt bằng LLM để bổ sung knowledge base.

Bảng 3.4. Danh sách chức năng của AI Core

CHƯƠNG 3. PHÂN TÍCH VÀ THIẾT KẾ HỆ THỐNG

- **API Response Time (Chat):** < 3s. Giải pháp: Tối ưu hóa ngữ cảnh (Context optimization) và Cache.
- **Cache Strategy:** Sử dụng chiến lược Multi-tier caching với Redis. cấu hình: Prediction (1 giờ), Evaluation (2 giờ), Cluster Index (24 giờ). Cache key tính bằng SHA256 hash của tham số đầu vào.

B. Bảo mật (Security)

- **Mã hóa mật khẩu:** Sử dụng Bcrypt (thư viện Passlib).
- **Mã hóa PII:** Sử dụng Amazon Web Services (AWS) Key Management Service (KMS) với cơ chế **Envelope Encryption**:
 - KMS tạo Data Encryption Key (DEK) cho mỗi phiên mã hóa.
 - DEK được sử dụng để mã hóa dữ liệu bằng AES-256.
 - DEK được mã hóa bởi Customer Master Key (CMK) trước khi lưu trữ.
 - Khi giải mã, chỉ service được ủy quyền mới có thể yêu cầu KMS giải mã DEK.
 - **Xác thực API (JWT):** Sử dụng JSON Web Token.
- **PII Redaction:** Lớp trung gian lọc và thay thế thông tin cá nhân bằng Regex trước khi gửi ra ngoài hệ thống.

3.4 Mô hình hóa quy trình nghiệp vụ

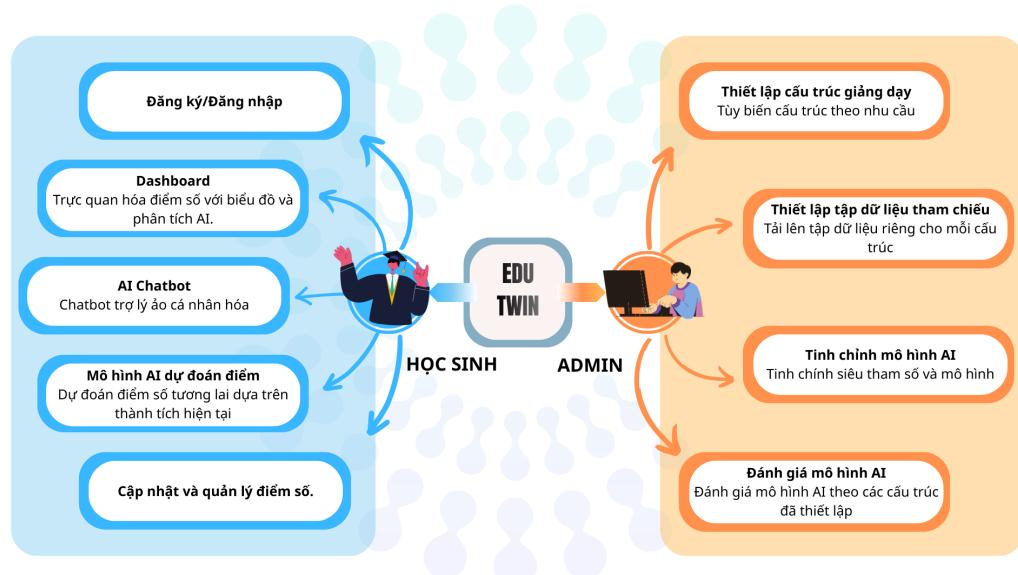
3.4.1 Biểu đồ Use Case

Biểu đồ Use Case dưới đây mô tả sự tương tác giữa các tác nhân chính (Học viên và Quản trị viên) với các chức năng cốt lõi của hệ thống EduTwin. Biểu đồ được trình bày theo dạng sơ đồ khôi để làm rõ luồng tương tác.

3.4.2 Quy trình Đăng nhập và Xác thực

1. Người dùng nhập username và password.
2. Hệ thống truy vấn database, kiểm tra username tồn tại.
3. Nếu tồn tại, so khớp password với hash đã lưu (Bcrypt).
4. Nếu khớp, tạo JWT token chứa user_id và thời hạn 24h.

CHƯƠNG 3. PHÂN TÍCH VÀ THIẾT KẾ HỆ THỐNG



Hình 3.2. Biểu đồ Usecase của EduTwin

5. Trả về token cho client, client lưu vào localStorage.
6. Các request tiếp theo đính kèm token trong header Authorization.

3.4.3 Quy trình Dự đoán Điểm số

Đây là quy trình nghiệp vụ cốt lõi của hệ thống dự đoán, được kích hoạt tự động khi người dùng cập nhật điểm:

1. Người dùng nhập hoặc cập nhật điểm số thực tế.
2. Hệ thống lưu điểm vào database và kiểm tra cấu trúc hiện hành.
3. Nếu có điểm thiếu, gọi **KNN Imputer** để điền khuyết.
4. Kiểm tra Redis Cache. Nếu Miss, tải Cluster Index.
5. Gán người dùng vào Cluster gần nhất, lấy Prototypes.
6. Chạy thuật toán Lazy Learning (KNN/KR/LWLR) trên tập mẫu hình.
7. Lưu kết quả dự đoán vào database, cập nhật Cache và trả về.

3.4.4 Quy trình Chat với AI

1. Người dùng gửi tin nhắn qua giao diện Chat.
2. Backend nhận request, load lịch sử hội thoại từ database.

CHƯƠNG 3. PHÂN TÍCH VÀ THIẾT KẾ HỆ THỐNG

3. Xây dựng context: điểm số + sở thích + tin nhắn gần đây.
4. Áp dụng **PII Redaction** trên toàn bộ context.
5. Gửi request đến LLM API và nhận toàn bộ response.
6. Chia response thành từng word, stream từng phần đến client qua SSE (pseudo-streaming).|
7. Khi hoàn thành, lưu toàn bộ response vào bảng `chat_messages`.
8. Phân tích response để cập nhật `preferences` nếu phát hiện thông tin mới.

3.4.5 Quy trình Thiết lập Cấu trúc Học tập

1. Admin truy cập trang Developer Tools.
2. Nhập tên cấu trúc (VD: "THPTQuốc Gia 2024").
3. Thêm danh sách môn học (`subject_labels`) và các mốc thời gian (`time_point_labels`).
4. Chọn thang điểm (`scale_type`: 0-10, 0-100, GPA...).
5. Nhấn "Lưu" - hệ thống tạo record trong `custom_teaching_structures`.
6. Kích hoạt cấu trúc (`is_active = true`) - các cấu trúc khác tự động bị vô hiệu hóa.

3.4.6 Quy trình Upload Dataset và Đánh giá Model

1. Admin chọn cấu trúc và upload file Excel chứa dữ liệu tham chiếu.
2. Hệ thống parse file, kiểm tra header khớp với `subject_labels` × `time_point_labels`.|
3. Nếu hợp lệ, lưu các record vào `custom_reference_dataset`.
4. Admin chọn Input Timepoints (mốc đầu vào) và Output Timepoints (mốc cần dự đoán).
5. Admin điều chỉnh tham số model (K , Bandwidth, Tau) qua giao diện slider.
6. Nhấn "Evaluate" - hệ thống chia dataset thành tập huấn luyện (80%) và tập kiểm thử (20%), sau đó đánh giá từng model trên tập kiểm thử.
7. Hiển thị kết quả: MAE, RMSE, pseudo-Accuracy cho từng model (KNN, KR, LWLR).|
8. Admin chọn model tốt nhất và kích hoạt pipeline.

3.4.7 Quy trình Trích xuất Tài liệu

1. Admin upload file PDF hoặc DOCX bổ sung cho cấu trúc học tập.
2. Backend xác định loại file và gọi thư viện trích xuất tương ứng:
 - PDF: Sử dụng PyMuPDF hoặc pdfplumber.
 - DOCX: Sử dụng python-docx.
3. Nội dung text thô được gửi đến LLM với prompt yêu cầu tóm tắt.
4. LLM trả về bản tóm tắt ngắn gọn.
5. Lưu tóm tắt vào bảng custom_structure_documents.
6. Khi chatbot hoạt động, tóm tắt này được inject vào context để bổ sung kiến thức.

3.5 Thiết kế Kiến trúc hệ thống

3.5.1 Kiến trúc tổng thể

Hệ thống được thiết kế theo mô hình phân tầng (Layered Architecture) với các thành phần:

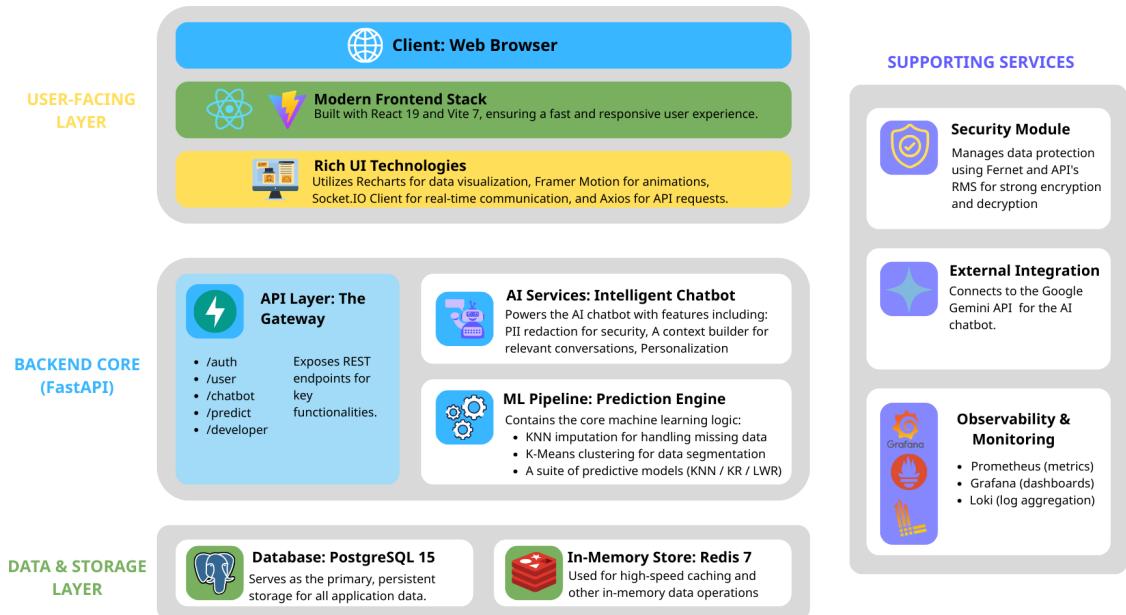
- **Client Layer:** Ứng dụng Web Single Page Application xây dựng bằng React 19 và Vite.
- **API Gateway:** Tiếp nhận và điều phối các yêu cầu từ Client.
- **Backend Services:** Các module xử lý nghiệp vụ viết bằng Python FastAPI (Auth, User, Chatbot, Prediction).
- **Data Tier:** Lưu trữ dữ liệu bền vững với PostgreSQL và caching hiệu năng cao với Redis.
- **External Integration:** Kết nối với các dịch vụ LLM (OpenAI/Gemini) thông qua lớp bảo mật.

3.5.2 Kiến trúc Module AI

Module AI đóng vai trò là "trái tim" của hệ thống, bao gồm các pipeline xử lý dữ liệu:

- **Input Layer:** Tiếp nhận dữ liệu thô từ người dùng và dữ liệu tham chiếu.

CHƯƠNG 3. PHÂN TÍCH VÀ THIẾT KẾ HỆ THỐNG



Hình 3.3. Kiến trúc tổng thể hệ thống EduTwin

- **Preprocessor:** Kiểm tra và xác thực dữ liệu đầu vào theo cấu trúc định nghĩa.
- **Imputation Engine:** Xử lý dữ liệu khuyết.
- **VQ Indexer:** Phân cụm và tuyển chọn mẫu hình để tối ưu không gian tìm kiếm.
- **Predictor:** Thực thi các thuật toán dự đoán cục bộ.

3.6 Thiết kế Cơ sở dữ liệu

Cơ sở dữ liệu của EduTwin được thiết kế dựa trên hệ quản trị cơ sở dữ liệu **PostgreSQL 15**, sử dụng kiến trúc lai ghép (Hybrid Model). Kiến trúc này kết hợp tính chât chẽ, toàn vẹn dữ liệu của mô hình quan hệ (Relational) cho thông tin người dùng, với tính linh hoạt của mô hình tài liệu (Document-based) thông qua kiểu dữ liệu **JSONB** để lưu trữ các cấu trúc học tập động.

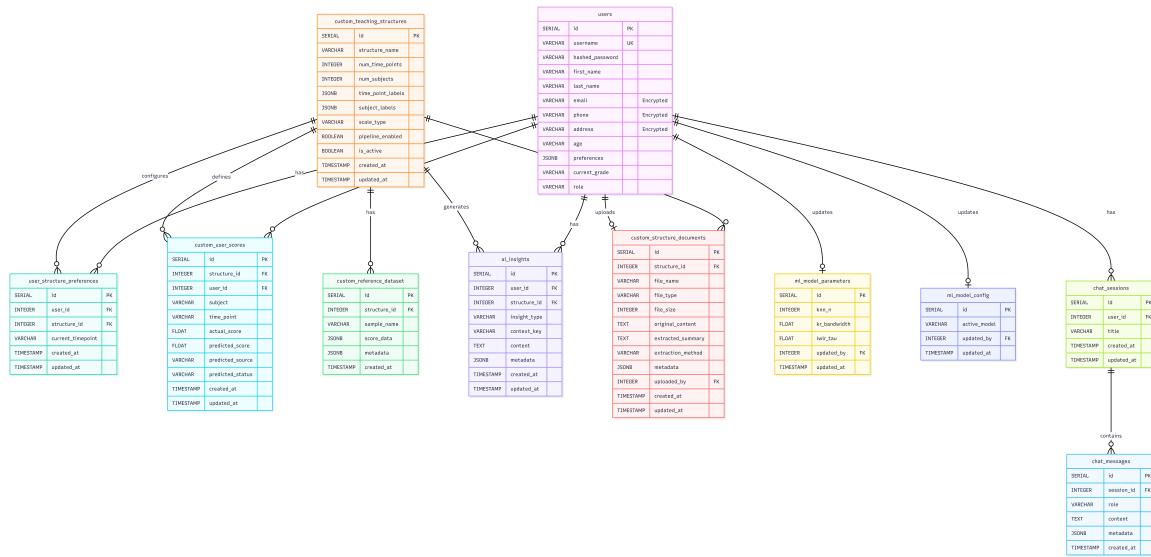
3.6.1 Thiết kế mức khái niệm

Mô hình thực thể kết hợp (ERD) của hệ thống được xây dựng xoay quanh các thực thể chính sau:

3.6.2 Thiết kế chi tiết lược đồ dữ liệu

Hệ thống bao gồm 5 nhóm bảng chính: Quản trị người dùng, Cấu trúc động, Dữ liệu điểm số, Hỗ trợ AI và Nhật ký hội thoại.

CHƯƠNG 3. PHÂN TÍCH VÀ THIẾT KẾ HỆ THỐNG



Hình 3.4. Sơ đồ quan hệ thực thể (ERD) của hệ thống EduTwin

1. Nhóm bảng Cấu trúc động (Core Dynamic Engine) Đây là trái tim của hệ thống, cho phép Admin định nghĩa bất kỳ mô hình giáo dục nào mà không cần thay đổi code.

2. Nhóm bảng Dữ liệu điểm số Dữ liệu điểm số được lưu trữ dọc (Normalized) để tối ưu cho các truy vấn thống kê, nhưng tham chiếu đến metadata động.

3. Nhóm bảng Người dùng và Bảo mật (PII) Lưu trữ thông tin định danh với cơ chế mã hóa cấp độ trường (Field-level Encryption).

4. Nhóm bảng Hỗ trợ AI Các bảng phục vụ cho việc theo dõi hoạt động AI, lưu trữ tài liệu tham khảo và cấu hình mô hình Machine Learning.

5. Nhóm bảng Hội thoại (Chat History) Lưu trữ toàn bộ lịch sử tương tác giữa học sinh và AI.

Thiết kế này cho phép học sinh có nhiều phiên hội thoại độc lập, mỗi phiên theo dõi một chủ đề hoặc mục tiêu học tập khác nhau. Đồng thời hỗ trợ quản lý preferences riêng biệt cho từng cấu trúc học tập mà user tham gia.

3.6.3 Phân tích các quyết định thiết kế quan trọng

A. Tại sao sử dụng JSONB cho Cấu trúc động? Thay vì sử dụng mô hình Entity-Attribute-Value (EAV) truyền thống vốn phức tạp trong truy vấn và khó bảo trì, để lựa chọn lưu trữ định nghĩa môn học và học kỳ dưới dạng JSONB.

CHƯƠNG 3. PHÂN TÍCH VÀ THIẾT KẾ HỆ THỐNG

Tên trường	Kiểu dữ liệu	Mô tả
id	SERIAL (PK)	Khóa chính.
structure_name	VARCHAR	Tên cấu trúc (VD: "THPT Quốc Gia 2024").
num_time_points	INTEGER	Số lượng mốc thời gian.
num_subjects	INTEGER	Số lượng môn học.
time_point_labels	JSONB	Danh sách mốc thời gian. VD: ["HK1-L10", "HK2-L10"].
subject_labels	JSONB	Danh sách môn học. VD: ["Toán", "Lý", "Anh"].
scale_type	VARCHAR	Thang điểm (VD: "0-10", "0-100", "A-F", "GPA").
pipeline_enabled	BOOLEAN	Bật/tắt pipeline dự đoán ML.
is_active	BOOLEAN	Cờ đánh dấu cấu trúc đang được sử dụng hiện hành.
created_at	TIMESTAMP	Thời điểm tạo.
updated_at	TIMESTAMP	Thời điểm cập nhật gần nhất.

Bảng 3.5. Bảng custom_teaching_structures - Định nghĩa cấu trúc học tập

CHƯƠNG 3. PHÂN TÍCH VÀ THIẾT KẾ HỆ THỐNG

Tên trường	Kiểu dữ liệu	Mô tả
id	SERIAL (PK)	Khóa chính.
user_id	INTEGER (FK)	Tham chiếu bảng Users.
structure_id	INTEGER (FK)	Tham chiếu bảng Structures.
subject	VARCHAR	Tên môn học (phải tồn tại trong subject_labels).
time_point	VARCHAR	Mốc thời gian (phải tồn tại trong time_point_labels).
actual_score	FLOAT	Điểm thực tế (có thể NULL nếu chưa có).
predicted_score	FLOAT	Điểm dự đoán bởi AI.
predicted_source	VARCHAR	Thuật toán dùng để dự đoán ('knn', 'kernel_regression', 'lwlr').
predicted_status	VARCHAR	Trạng thái dự đoán ('active', 'replaced').
created_at	TIMESTAMP	Thời điểm tạo.
updated_at	TIMESTAMP	Thời điểm cập nhật gần nhất.

Bảng 3.6. Bảng custom_user_scores - Lưu trữ điểm số người dùng

Tên trường	Kiểu dữ liệu	Mô tả
id	SERIAL (PK)	Khóa chính.
structure_id	INTEGER (FK)	Tham chiếu cấu trúc tương ứng.
sample_name	VARCHAR	Tên mẫu dữ liệu (tùy chọn).
score_data	JSONB	Vector điểm số đầy đủ của một mẫu. VD: {"Toán_HK1": 8.0, "Lý_HK1": 7.5}.
metadata	JSONB	Thông tin bổ sung về mẫu.
created_at	TIMESTAMP	Thời điểm tạo.

Bảng 3.7. Bảng custom_reference_dataset - Dữ liệu mẫu (Training Data)

CHƯƠNG 3. PHÂN TÍCH VÀ THIẾT KẾ HỆ THỐNG

Tên trường	Kiểu dữ liệu	Mô tả
id	SERIAL (PK)	Khóa chính.
username	VARCHAR (Unique)	Tên đăng nhập.
hashed_password	VARCHAR	Mật khẩu đã băm (Bcrypt).
first_name	VARCHAR	Tên người dùng.
last_name	VARCHAR	Họ người dùng.
email	VARCHAR	Mã hóa Envelope Encryption.
phone	VARCHAR	Mã hóa Envelope Encryption.
address	VARCHAR	Mã hóa Envelope Encryption.
age	VARCHAR	Tuổi người dùng.
current_grade	VARCHAR	Khối lớp hiện tại.
role	VARCHAR	Vai trò ('user', 'admin').
preferences	JSONB	Lưu sở thích học tập, phong cách chat.

Bảng 3.8. Bảng users - Thông tin người dùng

Tên trường	Kiểu dữ liệu	Mô tả
id	SERIAL (PK)	Khóa chính.
user_id	INTEGER (FK)	Tham chiếu bảng Users.
structure_id	INTEGER (FK)	Tham chiếu cấu trúc học tập.
insight_type	VARCHAR	Loại nhận định ('slide_comment', 'chat_response', 'subject_analysis').
context_key	VARCHAR	Khóa ngữ cảnh ('overview_chart', 'Math', 'A00').
content	TEXT	Nội dung nhận định của AI.
metadata	JSONB	Thông tin bổ sung (diễn liên quan, độ tin cậy).
created_at	TIMESTAMP	Thời điểm tạo.
updated_at	TIMESTAMP	Thời điểm cập nhật gần nhất.

Bảng 3.9. Bảng ai_insights - Lưu trữ nhận định AI về học sinh

CHƯƠNG 3. PHÂN TÍCH VÀ THIẾT KẾ HỆ THỐNG

Tên trường	Kiểu dữ liệu	Mô tả
id	SERIAL (PK)	Khóa chính.
structure_id	INTEGER (FK)	Tham chiếu cấu trúc học tập.
file_name	VARCHAR	Tên file gốc.
file_type	VARCHAR	Loại file ('pdf', 'docx', 'txt').
file_size	INTEGER	Kích thước file (bytes).
original_content	TEXT	Nội dung đầy đủ được trích xuất.
extracted_summary	TEXT	Nội dung tóm tắt để bổ sung context cho LLM.
extraction_method	VARCHAR	Phương pháp trích xuất ('llm_summary').
metadata	JSONB	Thông tin bổ sung (số trang, sections).
uploaded_by	INTEGER (FK)	Người tải lên (tham chiếu Users).
created_at	TIMESTAMP	Thời điểm tạo.
updated_at	TIMESTAMP	Thời điểm cập nhật gần nhất.

Bảng 3.10. Bảng custom_structure_documents - Lưu tài liệu tham khảo cho AI

Tên trường	Kiểu dữ liệu	Mô tả
id	SERIAL (PK)	Khóa chính.
knn_n	INTEGER	Số láng giềng K cho thuật toán KNN.
kr_bandwidth	FLOAT	Bảng thông cho Kernel Regression.
lwlr_tau	FLOAT	Tham số Tau cho LWLR.
updated_by	INTEGER (FK)	Người cập nhật (tham chiếu Users).
updated_at	TIMESTAMP	Thời điểm cập nhật gần nhất.

Bảng 3.11. Bảng ml_model_parameters - Tham số các mô hình ML

CHƯƠNG 3. PHÂN TÍCH VÀ THIẾT KẾ HỆ THỐNG

Tên trường	Kiểu dữ liệu	Mô tả
id	SERIAL (PK)	Khóa chính.
active_model	VARCHAR	Mô hình đang được sử dụng ('knn', 'kernel_regression', 'lblr').
updated_by	INTEGER (FK)	Người cập nhật (tham chiếu Users).
updated_at	TIMESTAMP	Thời điểm cập nhật gần nhất.

Bảng 3.12. Bảng ml_model_config - Cấu hình mô hình ML đang hoạt động

Tên trường	Kiểu dữ liệu	Mô tả
id	SERIAL (PK)	Khóa chính.
user_id	INTEGER (FK)	Tham chiếu bảng Users.
title	VARCHAR	Tiêu đề phiên (tự động sinh từ tin nhắn đầu).
created_at	TIMESTAMP	Thời điểm tạo phiên.
updated_at	TIMESTAMP	Thời điểm cập nhật gần nhất.

Bảng 3.13. Bảng chat_sessions - Quản lý phiên hội thoại

Tên trường	Kiểu dữ liệu	Mô tả
id	SERIAL (PK)	Khóa chính.
session_id	INTEGER (FK)	Tham chiếu bảng ChatSessions.
role	VARCHAR	Vai trò: 'user', 'assistant', hoặc 'system'.
content	TEXT	Nội dung tin nhắn.
metadata	JSONB	Thông tin bổ sung về tin nhắn.
created_at	TIMESTAMP	Thời điểm gửi.

Bảng 3.14. Bảng chat_messages - Lưu trữ tin nhắn

CHƯƠNG 3. PHÂN TÍCH VÀ THIẾT KẾ HỆ THỐNG

Tên trường	Kiểu dữ liệu	Mô tả
id	SERIAL (PK)	Khóa chính.
user_id	INTEGER (FK)	Tham chiếu bảng Users.
structure_id	INTEGER (FK)	Tham chiếu bảng Structures.
current_timepoint	VARCHAR	Mốc thời gian hiện tại của user cho structure này.
created_at	TIMESTAMP	Thời điểm tạo.
updated_at	TIMESTAMP	Thời điểm cập nhật gần nhất.

Bảng 3.15. Bảng user_structure_preferences - Lưu trữ preferences của user theo structure

- **Hiệu năng:** PostgreSQL hỗ trợ Generalized Inverted Index (GIN) cho JSONB, cho phép truy xuất dữ liệu metadata với độ phức tạp $O(1)$ thay vì phải JOIN hàng loạt bảng như EAV.
- **Linh hoạt:** Khi nhà trường thay đổi chương trình học (ví dụ: thêm môn "Trí tuệ nhân tạo"), Admin chỉ cần cập nhật mảng JSON trong bảng structures mà không cần thực hiện lệnh ALTER TABLE nặng nề hay downtime hệ thống.

B. Chiến lược Đánh chỉ mục (Indexing Strategy) Để đảm bảo tốc độ phản hồi < 200ms cho các truy vấn dự đoán, hệ thống thiết lập các chỉ mục tối ưu:

```
-- Index duy nhất để đảm bảo tính toàn vẹn dữ liệu điểm số
CREATE UNIQUE INDEX ix_user_score_unique
ON custom_user_scores(user_id, structure_id, subject, time_point);

-- GIN Index cho phép tìm kiếm nhanh trong dữ liệu huấn luyện AI
CREATE INDEX ix_ref_data_json
ON custom_reference_dataset USING GIN (score_data);
```

C. Cơ chế Bảo mật Dữ liệu (Data Security) Khác với các hệ thống thông thường chỉ mã hóa mật khẩu, EduTwin thực hiện mã hóa hai chiều (Symmetric Encryption) đối với các trường PII (Email, Phone) ngay tại tầng ứng dụng (Application Layer) trước khi lưu xuống DB. Điều này đảm bảo rằng ngay cả khi Database Administrator truy cập trực tiếp vào cơ sở dữ liệu cũng không thể đọc được thông tin cá nhân của học sinh.

3.7 Thiết kế Thuật toán và Luồng xử lý dữ liệu

Để đảm bảo khả năng dự đoán chính xác trong thời gian thực (real-time) với dữ liệu không đầy đủ, hệ thống EduTwin triển khai một pipeline xử lý dữ liệu phức hợp. Phần này mô tả chi tiết logic vận hành của các thuật toán cốt lõi.

3.7.1 Luồng xử lý dữ liệu khuyết

Trong môi trường thực tế, việc áp dụng trực tiếp công thức Euclidean chuẩn (như đã trình bày ở Chương 2) là không khả thi do sự hiện diện của các giá trị NaN. Hệ thống EduTwin giải quyết bằng thuật toán KNN Imputation được cải tiến (Modified KNN Imputer):

- Định nghĩa khoảng cách thích ứng (NaN-aware Distance):** Để so sánh vector x (có dữ liệu khuyết) và vector y , hệ thống sử dụng công thức khoảng cách hiệu chỉnh nhằm bù đắp cho các chiều dữ liệu bị thiếu:

$$d(x, y) = \begin{cases} \infty & \text{nếu } P = \emptyset \\ \sqrt{\frac{N}{|P|} \sum_{i \in P} (x_i - y_i)^2} & \text{nếu } P \neq \emptyset \end{cases}$$

Trong đó: N là tổng số chiều dữ liệu, P là tập hợp các chỉ số môn học mà cả x và y đều có điểm số. Hệ số $\frac{N}{|P|}$ đóng vai trò "phóng đại" khoảng cách tính được trên tập con P để ước lượng khoảng cách trên toàn không gian N chiều.

- Tìm kiếm láng giềng và Điền khuyết:** Hệ thống chọn ra k láng giềng tốt nhất (\mathcal{N}_k) dựa trên khoảng cách thích ứng trên và thực hiện điền khuyết trọng số:

$$\hat{x}_j = \frac{\sum_{y \in \mathcal{N}_k} w_y \cdot y_j}{\sum_{y \in \mathcal{N}_k} w_y} \quad \text{với} \quad w_y = \frac{1}{d(x, y) + 10^{-5}}$$

Trong đó hằng số nhỏ 10^{-5} để tránh lỗi chia cho 0.

3.7.2 Chiến lược Tối ưu hóa dự đoán với Lượng tử hóa Vector và Tối ưu hóa Mẫu hình

Để giải quyết bài toán độ phức tạp tính toán $O(N)$ của thuật toán Lazy Learning khi dữ liệu tham chiếu N tăng lớn (vẫn đề độ trễ), hệ thống áp dụng chiến lược hai tầng "Phân cụm toàn cục - Hồi quy cục bộ". **Tầng 1: Đánh chỉ mục bằng Phân cụm (Global Indexing)**

CHƯƠNG 3. PHÂN TÍCH VÀ THIẾT KẾ HỆ THỐNG

- Sử dụng thuật toán **K-Means** (thư viện Scikit-learn) để phân hoạch không gian dữ liệu thành K cụm (Clusters). Số lượng cụm K được xác định động dựa trên quy mô tập dữ liệu N theo công thức:

$$K = \begin{cases} 1 & \text{nếu } N < 3000 \text{ (không phân cụm)} \\ \min\left(\left\lceil \frac{N}{3000} \right\rceil, 100\right) & \text{nếu } N \geq 3000 \end{cases}$$

Thiết kế này đảm bảo mỗi cụm có khoảng 3000 mẫu, đủ cho các thuật toán lazy learning hoạt động ổn định. Khi $N < 3000$, hệ thống gán $K = 1$ (không phân cụm) để ưu tiên độ chính xác tuyệt đối trên tập dữ liệu nhỏ.

- Tại mỗi cụm, hệ thống **lưu trữ toàn bộ điểm dữ liệu** (được sắp xếp theo khoảng cách đến tâm cụm) để cho phép tuyển chọn mẫu hình thích ứng tại thời điểm dự đoán.
- Kết quả: Một bộ chỉ mục (Index) được lưu trữ trên Random Access Memory (RAM) (qua Redis) để truy xuất cực nhanh.

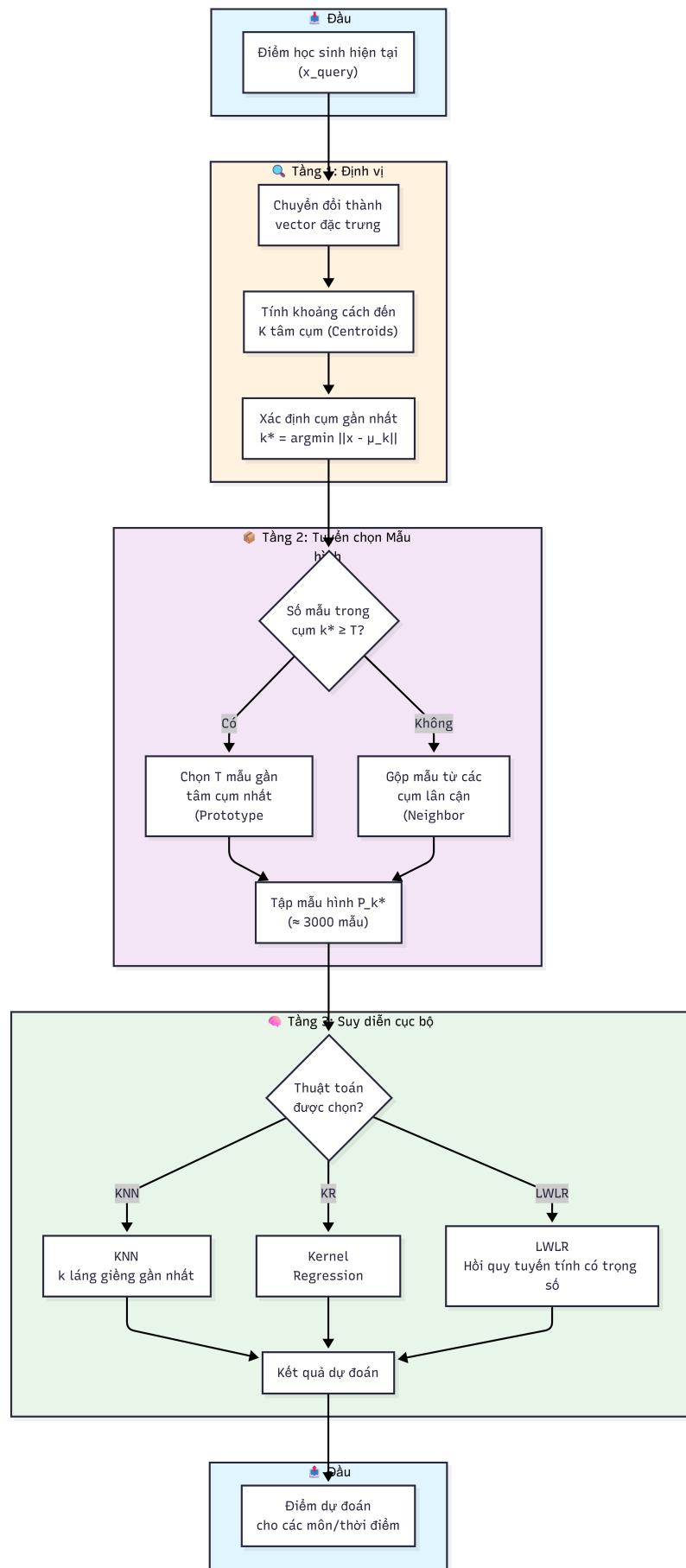
Tầng 2: Dự đoán cục bộ với Tuyển chọn Mẫu hình Thích ứng (Local Prediction with Adaptive Prototype Selection) Tại thời điểm runtime (khi người dùng yêu cầu dự đoán), quy trình diễn ra như sau:

- *Bước 1 (Assignment):* Xác định cụm k^* mà vector học sinh x thuộc về bằng cách so sánh khoảng cách tới các tâm cụm (Centroids). Độ phức tạp giảm từ $O(N)$ xuống $O(K)$.
- *Bước 2 (Adaptive Retrieval):* Truy xuất mẫu từ cụm k^* với chiến lược thích ứng:
 - ◊ Nếu $|C_{k^*}| \geq T$: Chọn $T = 3000$ mẫu gần tâm cụm nhất (Prototype Selection).
 - ◊ Nếu $|C_{k^*}| < T$: Gộp thêm mẫu từ các cụm lân cận theo thứ tự khoảng cách centroid (Neighbor Merging).
- *Bước 3 (Lazy Learning Inference):* Chạy thuật toán hồi quy chính (KNN, KR, LWLR) trên tập mẫu hình P_{k^*} (luôn có khoảng $T = 3000$ mẫu) để đưa ra kết quả cuối cùng.

3.7.3 Luồng dự đoán Lazy Learning

Quy trình dự đoán được thiết kế để tối ưu hóa thời gian phản hồi bằng cách chia nhỏ không gian tìm kiếm. Thay vì quét toàn bộ cơ sở dữ liệu (Global Scan), hệ thống thực hiện quy trình 3 bước sau: Chi tiết thuật toán vận hành như sau:

CHƯƠNG 3. PHÂN TÍCH VÀ THIẾT KẾ HỆ THỐNG



Hình 3.5. Sơ đồ luồng xử lý dự đoán Lazy Learning với cơ chế Lượng tử hóa Vector và Tối ưu hóa Mẫu hình

Bước 1: Định vị (Localization)

- Hệ thống chuyển đổi bảng điểm hiện tại của học sinh thành vector đặc trưng x_{query} .
- Tính khoảng cách Euclidean từ x_{query} tới các tâm cụm (Centroids) đã được huấn luyện trước bằng K-Means.
- Xác định cụm k^* có khoảng cách nhỏ nhất: $k^* = \arg \min_k \|x_{query} - \mu_k\|^2$.

Bước 2: Tuyển chọn Mẫu hình Thích ứng (Adaptive Prototype Selection)

- Tải tập mẫu của cụm k^* từ Redis Cache (các mẫu đã được sắp xếp theo khoảng cách đến tâm).
- Nếu cụm có $\geq T$ mẫu ($T = 3000$): Chọn T mẫu đầu tiên (gần tâm nhất).
- *Cơ chế Neighbor Merging*: Nếu cụm có $< T$ mẫu, hệ thống tự động mở rộng phạm vi tìm kiếm sang các cụm lân cận (theo thứ tự khoảng cách centroid) cho đến khi thu thập đủ T mẫu để đảm bảo độ ổn định thống kê.

Bước 3: Suy diễn cục bộ (Local Inference) Tùy thuộc vào cấu hình của Admin, hệ thống áp dụng một trong ba thuật toán trên tập mẫu hình cục bộ (luôn có khoảng $T = 3000$ mẫu):

- **KNN**: Lấy trung bình của k láng giềng gần nhất (thường chọn $k = 5 \sim 15$) trong tập mẫu hình.
- **KR**: Tính trung bình có trọng số sử dụng hàm Gaussian Kernel với tham số độ rộng dải h .
- **LWLR**: Tối ưu hóa hàm mất mát cục bộ để tìm tham số θ cho phương trình tuyến tính, sử dụng tham số độ rộng dải τ .

3.7.4 Luồng xử lý Chatbot RAG

Quy trình xử lý của AI Chatbot được thiết kế để đảm bảo tính cá nhân hóa và bảo mật dữ liệu:

1. **Context Building**: Hệ thống thu thập dữ liệu điểm số, sở thích học tập và metadata cấu trúc chương trình của học sinh.
2. **PII Redaction (Lớp bảo mật)**: Trước khi gửi dữ liệu sang LLM, một lớp trung gian sử dụng Regular Expression để nhận diện và thay thế các thông tin định danh:

CHƯƠNG 3. PHÂN TÍCH VÀ THIẾT KẾ HỆ THỐNG

- Email: student@email.com → <EMAIL_REDACTED>
- SĐT: 090xxxxxx → <PHONE_REDACTED>

3. **Prompt Engineering:** Dữ liệu sau khi ẩn danh được ghép vào một System Prompt chuyên biệt, định hướng cho AI đóng vai trò là "Người cố vấn học tập" (Academic Mentor).
4. **Response Processing:** Câu trả lời từ AI được phân tích để trích xuất các thông tin cập nhật về sở thích người dùng (nếu có) để lưu ngược lại vào hồ sơ cá nhân (Personalization Loop).

3.8 Thiết kế Giao diện người dùng

Giao diện người dùng được thiết kế theo phong cách hiện đại (Material/Flat Design), tối ưu hóa cho trải nghiệm trên cả máy tính và thiết bị di động, tập trung vào khả năng trực quan hóa dữ liệu.

3.8.1 Sơ đồ tổ chức thông tin

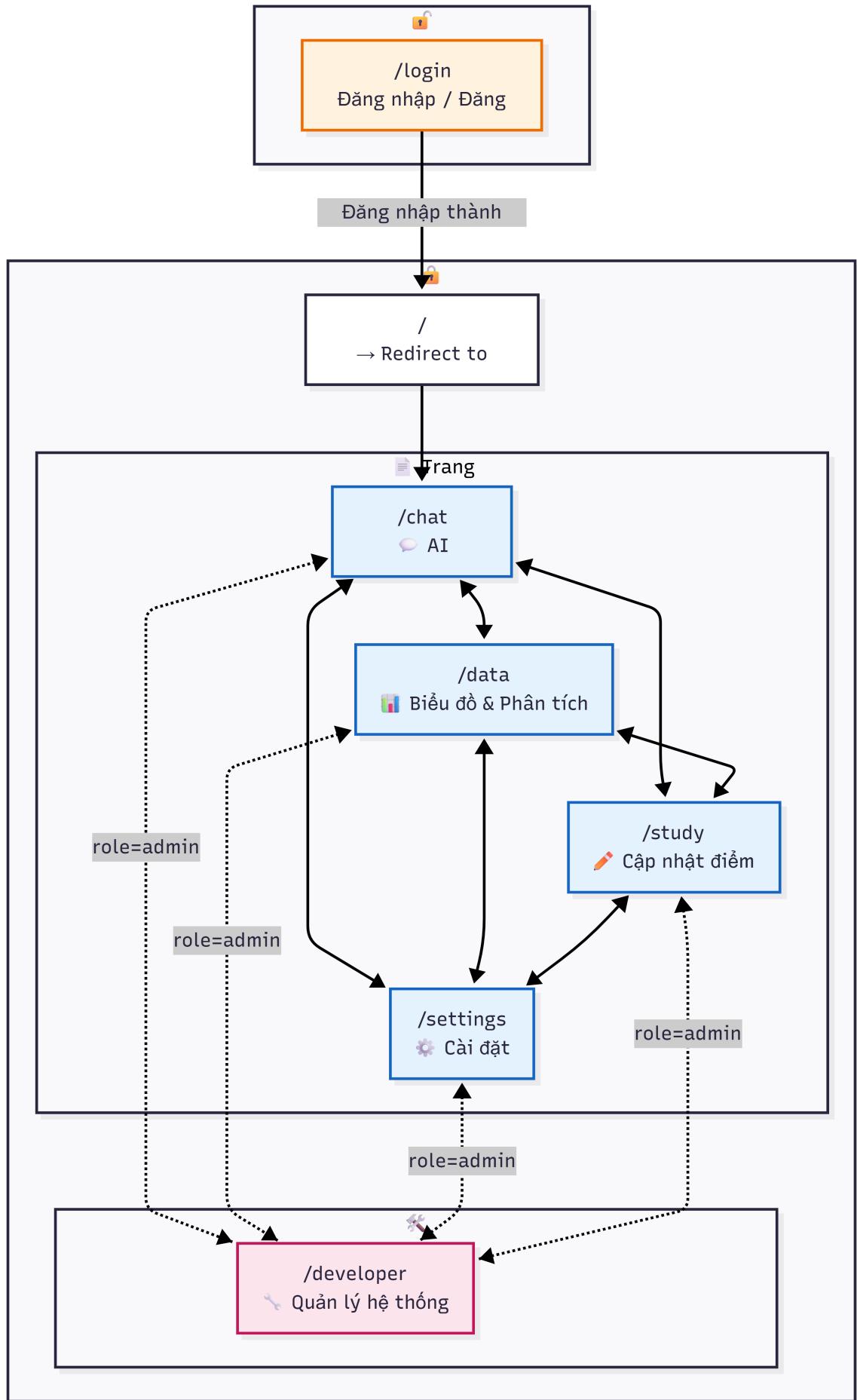
Hệ thống được tổ chức thành các luồng chức năng rõ ràng:

3.8.2 Thiết kế chi tiết các màn hình chính

1. **Màn hình Dashboard (DataViz)** Đây là màn hình trung tâm, cung cấp cái nhìn toàn cảnh về năng lực học sinh.

- **Radar Chart:** Biểu đồ mạng nhện đa giác, mỗi đỉnh đại diện cho một môn học. Vùng diện tích phủ giúp học sinh nhận diện nhanh thiên hướng (lịch về Tự nhiên hay Xã hội).
- **Bar Chart:** Biểu đồ cột hiển thị điểm số theo từng môn học.
- **Line Chart (Dự đoán):** Biểu đồ đường thể hiện hai dòng dữ liệu: đường nét liền (điểm thực tế quá khứ) và đường nét đứt (điểm dự đoán tương lai).
- **AI Insight Card:** Một thẻ thông báo nổi bật chứa nhận định ngắn gọn từ AI (ví dụ: "Bạn đang có xu hướng giảm điểm nhẹ ở môn Lý, hãy chú ý!").

CHƯƠNG 3. PHÂN TÍCH VÀ THIẾT KẾ HỆ THỐNG



Hình 3.6. Sơ đồ tổ chức chức năng (Site Map) của ứng dụng EduTwin

CHƯƠNG 3. PHÂN TÍCH VÀ THIẾT KẾ HỆ THỐNG

2. Màn hình Chatbot AI Giao diện hội thoại được thiết kế tương tự các ứng dụng nhắn tin hiện đại:

- **Khu vực hội thoại:** Hiển thị tin nhắn dạng bong bóng (bubbles). Tin nhắn của AI hỗ trợ hiển thị Markdown (đậm, nghiêng, danh sách) để trình bày lộ trình học tập rõ ràng.
- **Chat Session Sidebar:** Hiển thị các chat session

3. Màn hình Công cụ Quản trị (Developer Tools) Dành riêng cho Admin để thao tác với các tính năng Cấu trúc động:

- **Structure Editor:** Giao diện cho phép thêm/bớt môn học và học kỳ dưới dạng danh sách động, sau đó xuất ra JSON lưu vào Database.
- **Dataset Manager:** Khu vực kéo thả file Excel để tải lên dữ liệu huấn luyện mới.
- **Model Tuning:** Các thanh trượt (Sliders) để điều chỉnh tham số k (số láng giềng) hoặc Bandwidth h (độ rộng dải) và xem kết quả đánh giá (RMSE) thay đổi tức thì.

Tổng hợp Công nghệ và Môi trường triển khai

EduTwin được xây dựng dựa trên kiến trúc Modular Monolith hiện đại, phân tách rõ ràng giữa tầng xử lý dữ liệu thông minh (Backend) và tầng giao diện tương tác người dùng (Frontend). Các công nghệ được lựa chọn đều nhằm phục vụ mục tiêu tối thượng: đảm bảo tốc độ phản hồi thời gian thực cho thuật toán Lazy Learning.

3.8.3 Danh sách Công nghệ

- **Backend (Python Ecosystem):**

- **FastAPI:** Framework chủ đạo xây dựng API. Được chọn nhờ khả năng xử lý bất đồng bộ (Async Input/Output (I/O)), cho phép hệ thống phục vụ hàng nghìn yêu cầu dự đoán đồng thời mà không bị tắc nghẽn (blocking) khi chờ tính toán từ AI.
- **Scikit-learn & NumPy:** "Bộ não" tính toán của hệ thống, chịu trách nhiệm thực thi các thuật toán vector hóa, K-Means Clustering và KNN Imputation với hiệu năng cao.

CHƯƠNG 3. PHÂN TÍCH VÀ THIẾT KẾ HỆ THỐNG

- ◊ **SQLAlchemy 2.0:** Thư viện ORM hiện đại giúp ánh xạ linh hoạt giữa đối tượng Python và dữ liệu JSONB phức tạp.

- **Frontend (React Ecosystem):**

- ◊ **React 19 & Vite:** Sử dụng phiên bản React mới nhất với Concurrent Rendering để đảm bảo giao diện "Bản sao số" luôn mượt mà. Vite giúp tối ưu hóa tốc độ build và tải trang.
- ◊ **Recharts & Framer Motion:** Bộ đôi thư viện trực quan hóa, chịu trách nhiệm vẽ các biểu đồ Radar (năng lực) và Line (dự đoán) với hiệu ứng chuyển động sinh động, tăng trải nghiệm người dùng.

- **Database & Infrastructure:**

- ◊ **PostgreSQL 15:** Cơ sở dữ liệu chính, tận dụng sức mạnh của JSONB để lưu trữ cấu trúc học tập động (Schema-less within Schema).
- ◊ **Redis 7:** Bộ nhớ đệm (In-memory), đóng vai trò lưu trữ Index của các cụm mẫu hình (Cluster Prototypes) và phiên hội thoại Chatbot để truy xuất tức thì (<10ms).
- ◊ **Docker & Docker Compose:** Đóng gói toàn bộ môi trường giúp việc triển khai đồng nhất giữa Development và Production.

3.8.4 Môi trường triển khai và Yêu cầu hệ thống

Do đặc thù xử lý tính toán khoảng cách vector liên tục, hệ thống yêu cầu cấu hình phần cứng tối ưu cho các tác vụ Central Processing Unit (CPU)-bound (tính toán) và Memory-bound (lưu trữ vector).

1. Môi trường Phát triển (Development Environment)

- **OS:** Windows 11 (WSL2 - Ubuntu 22.04) hoặc macOS.
- **Công cụ:** Visual Studio Code, Docker Desktop, Postman.
- **Quản lý mã nguồn:** Git & GitHub với quy trình Continuous Integration and Continuous Delivery (CI/CD) cơ bản (GitHub Actions).

2. Môi trường Vận hành (Production Server)

Mô hình triển khai đề xuất là **Hybrid Deployment**:

CHƯƠNG 3. PHÂN TÍCH VÀ THIẾT KẾ HỆ THỐNG

- **Application Server (On-Premise/Cloud Virtual Machine):**

- ◊ *CPU*: Tối thiểu 4 Cores (Ưu tiên xung nhịp đơn nhân cao để tính toán khoảng cách Euclidean).
- ◊ *RAM*: Tối thiểu 16GB. Do thuật toán Lazy Learning cần lưu trữ Index của Codebook và Redis Cache trên RAM để đảm bảo tốc độ phản hồi $< 200ms$.
- ◊ *Storage*: SSD NVMe cho Database để đảm bảo tốc độ đọc/ghi (I/O) cao khi truy xuất lịch sử điểm số.

- **LLM Service Software as a Service (SaaS)**: Kết nối đến nhà cung cấp Cloud (OpenAI/Google) qua giao thức HTTPS bảo mật, đi qua lớp trung gian PII Redaction đặt tại Application Server.

Kiến trúc này đảm bảo sự cân bằng giữa chi phí đầu tư hạ tầng (không cần GPU đắt tiền cho việc training model) và hiệu năng sử dụng (tận dụng sức mạnh suy luận của LLM sẵn).

Chương 4. THỰC NGHIỆM VÀ ĐÁNH GIÁ

4.1 Môi trường và Dữ liệu thực nghiệm

4.1.1 Môi trường thực nghiệm

Để đảm bảo tính khách quan và hiệu quả, nghiên cứu được thực hiện trên hai môi trường tính toán riêng biệt phục vụ cho hai giai đoạn: huấn luyện mô hình và triển khai ứng dụng.

Đối với quá trình Huấn luyện và Đánh giá (Training & Evaluation): Việc so sánh hiệu năng các mô hình (đặc biệt là Kịch bản 2) đòi hỏi tài nguyên tính toán lớn để xử lý các thuật toán phức tạp như Long Short-Term Memory (LSTM) hay Ensemble Learning. Do đó, chúng tôi sử dụng nền tảng đám mây Kaggle Notebooks với cấu hình hiệu năng cao. Hệ thống được trang bị kép Graphics Processing Unit (GPU) NVIDIA Tesla T4 (Tổng VRAM 32GB) nhằm tăng tốc huấn luyện cho các mô hình Eager Learning, kết hợp với 30GB RAM để đảm bảo khả năng xử lý in-memory toàn bộ ma trận dữ liệu. Vi xử lý trung tâm là Intel Xeon (2 vCPUs @ 2.20GHz) với thời gian phiên làm việc liên tục lên đến 12 giờ.

Đối với các quá trình khác: Mục tiêu là mô phỏng điều kiện thực tế tại các trường học có hạ tầng hạn chế. Thiết bị thử nghiệm là Laptop Lenovo ThinkPad T490 với cấu hình khiêm tốn: vi xử lý Intel Core i5-10210U (dòng tiết kiệm điện ULV), 8GB RAM và không sử dụng GPU rời. Nền tảng LLM được tích hợp thông qua API của Google Gemini-2.5-Flash nhằm giảm tải tính toán cho thiết bị đầu cuối.

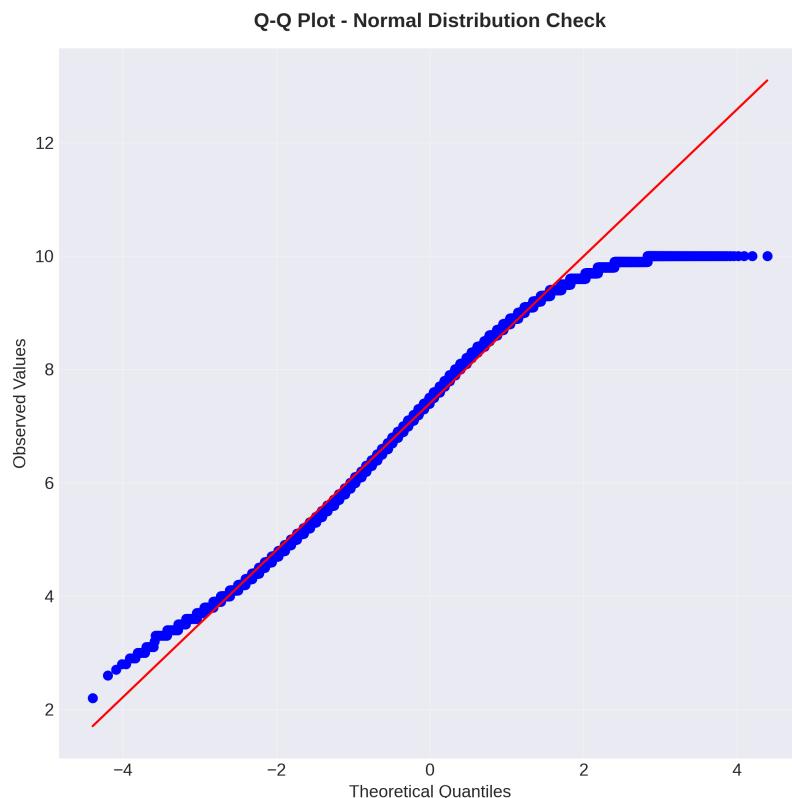
4.1.1.1 Phân tích Khám phá Dữ liệu

Tập dữ liệu thực nghiệm của hệ thống EduTwin được tổng hợp từ hồ sơ học tập chính thức tại Trường THPT Nguyễn Khuyến trong giai đoạn kéo dài bảy năm, từ 2017 đến 2024. Dữ liệu bao gồm điểm số của chín môn học cốt lõi trong chương trình trung học phổ thông (Toán, Ngữ văn, Tiếng Anh, Vật lý, Hóa học, Sinh học, Lịch sử, Địa lý và Giáo dục công dân), được ghi nhận xuyên suốt ba năm học (lớp 10, 11 và 12), với hai học kỳ mỗi năm. Sau quá trình tiền xử lý, loại bỏ các giá trị khuyết và các bản ghi không hợp lệ, tập dữ liệu cuối cùng bao gồm tổng cộng 123 291 điểm số thành phần, tương ứng với 2 283 hồ sơ học sinh hợp lệ.

Đặc điểm phân phối dữ liệu. Phân tích thống kê mô tả cho thấy phân phối điểm số tổng thể có xu hướng tiệm cận phân phối chuẩn, với giá trị trung bình đạt 7.40 và trung vị là 7.50. Khoảng chênh lệch nhỏ giữa hai thước đo xu hướng trung tâm này phản ánh

CHƯƠNG 4. THỰC NGHIỆM VÀ ĐÁNH GIÁ

mức độ đối xứng tương đối cao của phân phối quanh giá trị trung bình. Kết quả kiểm định Shapiro–Wilk được thực hiện trên các tập con đại diện theo từng môn học cho thấy giả thuyết phân phối chuẩn bị bác bỏ về mặt thống kê ($p\text{-value} < 0.05$), tuy nhiên hiện tượng này là phổ biến đối với các tập dữ liệu có kích thước lớn. Do đó, việc đánh giá trực quan thông qua biểu đồ Q–Q đóng vai trò quan trọng hơn trong việc xem xét tính gần chuẩn của dữ liệu.



Hình 4.1. Biểu đồ Q–Q kiểm định phân phối chuẩn, minh họa mức độ tuân thủ của dữ liệu điểm số đối với phân phối chuẩn lý thuyết.

Biểu đồ Q–Q được trình bày tại Hình 4.1 cho thấy các điểm quan sát bám sát đường chuẩn lý thuyết tại các phân vị trung tâm, trong khi chỉ xuất hiện các sai lệch nhẹ ở hai đầu phân phối. Hiện tượng này chủ yếu bắt nguồn từ tính chất bị chặn của thang điểm (0–10), qua đó xác nhận rằng giả định phân phối gần chuẩn là hợp lý và có thể chấp nhận được cho các phân tích mô hình hóa tiếp theo.

Biến thiên hiệu năng theo môn học. Sự khác biệt về mức độ đạt được giữa các môn học được thể hiện rõ ràng trong Hình 4.2(b). Nhóm các môn thuộc khối Khoa học Xã hội và kỹ năng công dân ghi nhận mức điểm trung bình cao hơn đáng kể, trong đó Giáo dục công dân đạt giá trị cao nhất (8.48), tiếp theo là Địa lý (7.79) và Lịch sử (7.44). Ngược lại, các môn học mang tính nền tảng và tư duy trừu tượng cao hơn như Toán học (6.93), Ngữ

CHƯƠNG 4. THỰC NGHIỆM VÀ ĐÁNH GIÁ

văn (6.99) và Vật lý (7.14) có điểm trung bình thấp hơn một cách nhất quán. Khoảng cách lên tới 1.55 điểm giữa môn có điểm trung bình cao nhất và thấp nhất phản ánh sự khác biệt về độ khó nội dung, phương pháp đánh giá cũng như đặc thù giảng dạy giữa các tổ bộ môn trong cùng một môi trường giáo dục.

Phân phối xếp loại học lực. Phân tích phân phối điểm số theo các mức xếp loại học lực, được minh họa trong Hình 4.2(c), cho thấy một xu hướng lệch phải tích cực, với mật độ điểm tập trung chủ yếu ở các mức thành tích trung bình khá đến cao. Cụ thể, 75.4% tổng số điểm thuộc hai nhóm “Khá” (6.5–8.0, chiếm 38.3%) và “Giỏi” (8.0–10.0, chiếm 37.1%). Ngược lại, tỷ lệ điểm dưới trung bình (nhỏ hơn 5.0) chỉ chiếm 3.6%, trong đó mức “Kém” gần như không đáng kể (0.1%). Sự tập trung cao của điểm số ở dải trên, cùng với số lượng rất hạn chế các quan sát ngoại lai ở mức thấp, cho thấy chất lượng đầu vào tương đối đồng đều cũng như hiệu quả đào tạo ổn định của Trường THPT Nguyễn Khuyến trong giai đoạn nghiên cứu.

4.2 Kịch bản 1: Đánh giá hiệu quả xử lý dữ liệu khuyết

Trước khi đưa vào mô hình dự đoán, tính toàn vẹn của dữ liệu là yếu tố tiên quyết. Kịch bản này tập trung so sánh hiệu quả của thuật toán đề xuất (KNN Imputation) so với các phương pháp điền khuyết thông kê cơ bản nhằm tìm ra giải pháp tối ưu cho bài toán dữ liệu giáo dục.

4.2.1 Thiết lập thực nghiệm

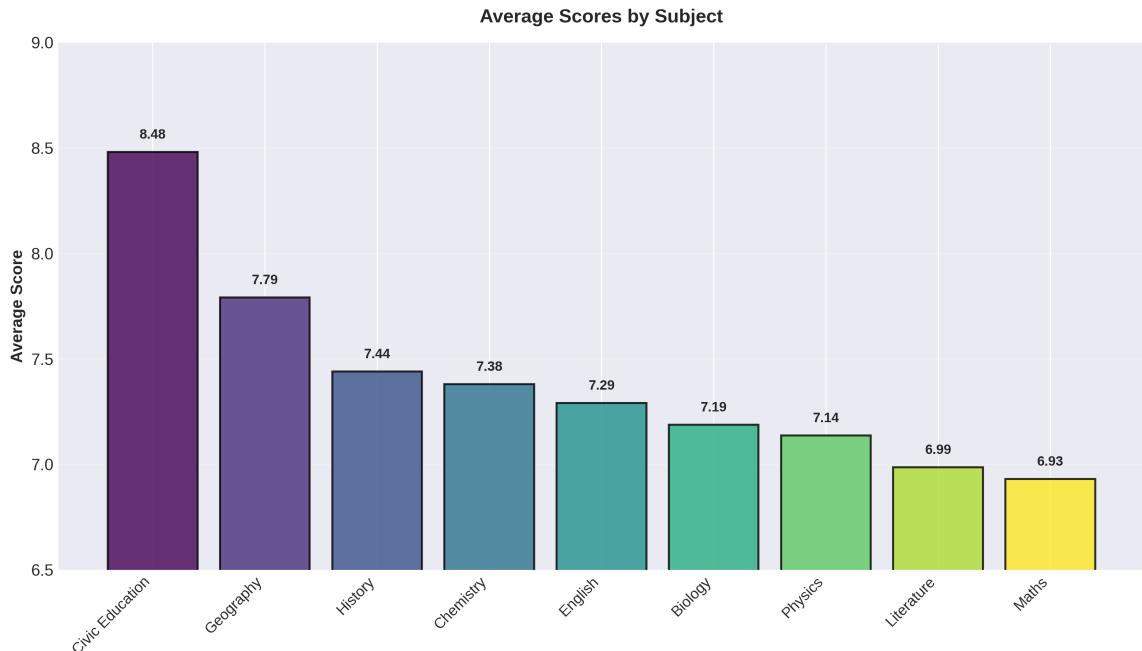
Dữ liệu gốc (đầy đủ) được áp dụng cơ chế xóa ngẫu nhiên Missing Completely At Random (MCAR) để tạo ra các tập dữ liệu khuyết giả lập với tỷ lệ từ 10% đến 50%. Nghiên cứu tiến hành so sánh ba phương pháp điền khuyết chính: Median Imputation (điền bằng trung vị), Mean Imputation (điền bằng trung bình), và phương pháp đề xuất KNN Imputation (điền dựa trên trung bình có trọng số của k láng giềng gần nhất).

4.2.2 Phân tích kết quả

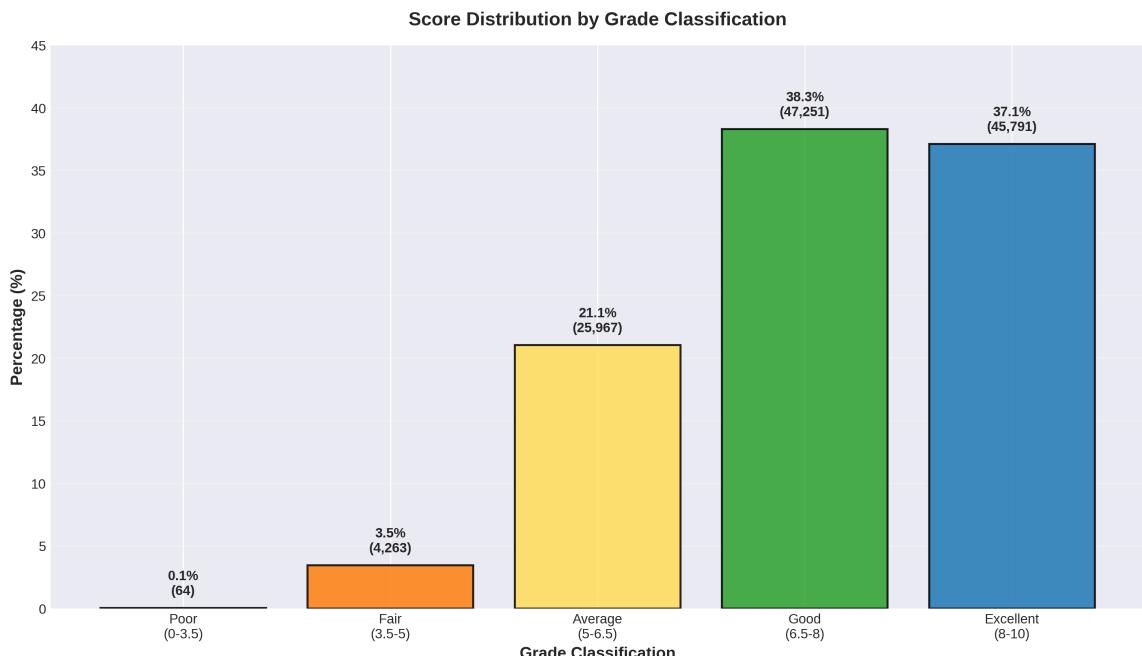
Kết quả thực nghiệm được trình bày tại Bảng 4.1 cho thấy sự vượt trội hoàn toàn của KNN Imputation trên mọi chỉ số đo lường.

Hạn chế của phương pháp thống kê: Các phương pháp Mean và Median Imputation cho kết quả R^2 rất thấp, chỉ đạt khoảng 0.17. Nguyên nhân cốt lõi là do các phương pháp này xử lý từng môn học một cách độc lập (univariate), bỏ qua hoàn toàn mối tương quan

CHƯƠNG 4. THỰC NGHIỆM VÀ ĐÁNH GIÁ



(a) Điểm trung bình theo từng môn học



(b) Phân phối điểm số theo xếp loại

Hình 4.2. Phân tích mô tả dữ liệu học tập: (b) So sánh điểm trung bình giữa các môn học cho thấy sự khác biệt đáng kể về hiệu năng; và (c) Phân phối điểm số theo các mức xếp loại phản ánh sự tập trung rõ rệt ở nhóm thành tích cao.

Phương pháp	Tỷ lệ thiếu 10%			Tỷ lệ thiếu 50%		
	MAE	RMSE	R ²	MAE	RMSE	R ²
Median Imputation	0.97	1.20	0.17	0.95	1.18	0.17
Mean Imputation	0.97	1.19	0.17	0.96	1.18	0.17
KNN Imputation	0.52	0.67	0.73	0.60	0.78	0.64

Bảng 4.1. So sánh hiệu năng điền khuyết (KNN so với các phương pháp thống kê)

đa chiều giữa các môn học. Việc điền giá trị trung bình đơn thuần làm giảm phương sai (variance), khiến hồ sơ học sinh trở nên “trung bình hóa” và mất đi các đặc trưng phân loại cá nhân.

Sức mạnh của cấu trúc tương quan: Ngược lại, KNN Imputation đạt $R^2 \approx 0.73$ nhờ khả năng tái tạo điểm số dựa trên “hồ sơ láng giềng”. Cơ chế này hoạt động dựa trên nguyên lý tương đồng: khi điền điểm môn Lý bị thiếu, thuật toán sẽ tham chiếu điểm Toán, Hóa, Sinh của những học sinh có năng lực tương đương trong không gian dữ liệu, từ đó đưa ra ước lượng chính xác hơn.

Độ bền vững (Robustness): Đáng chú ý, ngay cả trong kịch bản cực đoan khi dữ liệu bị mất tới 50%, KNN vẫn duy trì được độ chính xác ấn tượng với $R^2 > 0.64$. Điều này chứng minh dữ liệu giáo dục có tính dư thừa thông tin (redundancy) cao – năng lực học sinh ở các môn học có sự liên kết chặt chẽ – và thuật toán đề xuất đã khai thác hiệu quả đặc tính này để phục hồi thông tin.

4.3 Kịch bản 2: Đánh giá hiệu năng đa mô hình — So sánh Lazy Learning và Eager Learning

Kịch bản này được thiết kế để kiểm chứng giả thuyết về tính hiệu quả của Lazy Learning trong bối cảnh giáo dục. Chúng tôi thực hiện đánh giá song song giữa ba thuật toán thuộc nhóm *Lazy Learning* (KNN, Kernel Regression, LWLR) và hai thuật toán *Eager Learning* hiện đại (XGBoost, LSTM). Bài toán đặt ra là dự đoán điểm học kỳ lớp 12 dựa trên dữ liệu lịch sử của 4 học kỳ lớp 10 và 11 trên tập dữ liệu gồm 2284 mẫu.

4.3.1 Kết quả thực nghiệm định lượng

Kết quả đánh giá trên tập kiểm thử (test set - chiếm 20% dữ liệu gốc, tương đương 457 mẫu) được tóm tắt trong Bảng 4.2. Các chỉ số đánh giá bao gồm Sai số tuyệt đối trung bình (MAE), Sai số bình phương trung bình (RMSE), Hệ số xác định (R^2) và Thời gian suy diễn.

Nhóm	Mô hình	MAE	RMSE	R^2	Time (s)
<i>Lazy Learning</i>	KNN	0.53	0.68	0.66	0.13
	LWLR	0.53	0.68	0.66	36.53
	KR	0.55	0.72	0.61	0.15
<i>Eager Learning</i>	LSTM	0.54	0.69	0.65	41.48
	XGBoost	0.54	0.69	0.64	2.74

Bảng 4.2. So sánh hiệu năng giữa các mô hình Lazy Learning và Eager Learning

4.3.2 Phân tích kết quả

Dữ liệu cho thấy một quan sát quan trọng: các mô hình Lazy Learning, dù đơn giản về cấu trúc, lại đạt hiệu năng **tương đương hoặc tốt hơn** so với các mô hình Eager Learning phức tạp. Nguyên nhân đến từ việc dữ liệu điểm số học sinh mang tính cấu trúc cục bộ rõ rệt. Những học sinh có hồ sơ điểm lớp 10–11 tương tự nhau thường có kết quả lớp 12 gần giống nhau. Các mô hình như KNN và LWLR tận dụng trực tiếp đặc tính này thông qua cơ chế tìm kiếm cục bộ, thay vì cố gắng học một hàm tổng quát toàn cục như XGBoost. Bên cạnh đó, LSTM tuy mạnh về chuỗi thời gian nhưng không phát huy được ưu thế do chuỗi dữ liệu đầu vào quá ngắn (chỉ 4 thời điểm), dẫn đến hiệu năng thấp hơn KNN dù chi phí tính toán lớn hơn rất nhiều. XGBoost cũng bộc lộ hạn chế khi xử lý các quan hệ dạng khoảng cách mượt mà của điểm số, dẫn đến chỉ số R^2 thấp hơn.

Khi so sánh nội bộ trong nhóm Lazy Learning, KNN ($R^2 = 0.6625$) và LWLR ($R^2 = 0.6616$) có độ chính xác gần như tương đương. Tuy nhiên, yếu tố tạo nên sự khác biệt nằm ở chi phí tính toán. KNN chỉ mất **0.13s** để phản hồi, hoàn toàn phù hợp cho triển khai thời gian thực. Trong khi đó, LWLR mất tới **36.53s** do phải giải bài toán tối ưu cục bộ cho từng điểm truy vấn.

Tổng hợp lại, Lazy Learning không chỉ cạnh tranh sòng phẳng mà còn vượt trội Eager

CHƯƠNG 4. THỰC NGHIỆM VÀ ĐÁNH GIÁ

Learning trong bối cảnh cụ thể này. KNN đạt hiệu năng tốt nhất (MAE thấp nhất, R^2 cao nhất) trong cả 5 mô hình, đồng thời sở hữu tốc độ xử lý nhanh hơn LWLR khoảng 280 lần và nhanh hơn LSTM hơn 300 lần. Ngoài ra, KNN còn cung cấp khả năng giải thích cao (explainability) bằng việc truy xuất trực tiếp các học sinh “tương tự”, giúp tăng tính minh bạch cho hệ thống tư vấn.

4.4 Kịch bản 3: Đánh giá hiệu năng và Tốc độ phản hồi

Thách thức lớn nhất về mặt lý thuyết của Lazy Learning là tốc độ suy diễn chậm khi dữ liệu lớn (độ phức tạp $O(N)$). Kịch bản này nhằm kiểm chứng hiệu quả thực tế của giải pháp tối ưu hóa “Lượng tử hóa Vector và Tối ưu hóa Mẫu hình” đã được thiết kế trong chương 3.

Để kiểm chứng hiệu quả thực tế của chiến lược tối ưu hóa so với phương pháp quét toàn cục (*Global Scan*), nghiên cứu tiến hành thực nghiệm Stress Test trên tập dữ liệu mô phỏng quy mô lớn. Mục tiêu là đánh giá khả năng duy trì độ trễ thấp của hệ thống khi số lượng bản ghi tăng từ mức cơ bản lên đến 100.000 mẫu.

4.4.1 Thiết lập thực nghiệm và Dữ liệu

Thực nghiệm sử dụng tập dữ liệu gốc gồm 2.283 hồ sơ học sinh với 54 đặc trưng, bao gồm điểm số của 9 môn học tại 6 thời điểm đánh giá (từ học kỳ 1 lớp 10 đến học kỳ 2 lớp 12). Bài toán đặt ra là sử dụng 36 đặc trưng đầu vào (lớp 10 và 11) để dự đoán 18 đặc trưng mục tiêu (lớp 12) thông qua thuật toán KNN với tham số $k = 5$.

Để mô phỏng các môi trường dữ liệu lớn, dữ liệu gốc được nhân bản và làm giàu thông qua kỹ thuật thêm nhiễu Gaussian $\mathcal{N}(0, 0.3)$. Phương pháp này giúp tạo ra sự đa dạng cho dữ liệu sinh ra trong khi vẫn bảo toàn các đặc tính phân phối thống kê cốt lõi.

Cấu hình phân cụm cho phương pháp *Cached Query* được thiết lập động dựa trên quy mô dữ liệu N . Số lượng cụm K tuân theo công thức:

$$K = \min \left(\left\lceil \frac{N}{3000} \right\rceil, 100 \right),$$

nhằm đảm bảo kích thước mỗi cụm duy trì ổn định quanh ngưỡng 3.000 mẫu.

4.4.2 Phân tích kết quả thực nghiệm.

Kết quả định lượng được trình bày tại Bảng 4.3 cho thấy sự phân hóa rõ rệt về hiệu năng giữa hai phương pháp khi quy mô dữ liệu thay đổi.

CHƯƠNG 4. THỰC NGHIỆM VÀ ĐÁNH GIÁ

Đối với *Global Scan*, thời gian xử lý tăng tuyến tính theo độ phức tạp $O(N)$. Tại ngưỡng $N = 100.000$, độ trễ hệ thống lên tới 839,02 ms, bắt đầu vượt quá ngưỡng chấp nhận cho các ứng dụng thời gian thực.

Ngược lại, phương pháp *Cached Query* thể hiện ưu thế vượt trội ở các tập dữ liệu lớn ($N \geq 10.000$). Nhờ cơ chế sàng lọc mẫu hình thích ứng (*Adaptive Prototype Selection*), thuật toán chỉ cần truy xuất khoảng $T \approx 3.000$ mẫu liên quan nhất cho mỗi truy vấn, thay vì quét toàn bộ cơ sở dữ liệu. Kết quả là tại quy mô 100.000 bản ghi, hệ thống đạt tốc độ xử lý 72,99 ms, tương ứng với mức tăng tốc 11,50 lần so với phương pháp truyền thống.

Bảng 4.3. Kết quả Stress Test so sánh hiệu năng (dữ liệu thực tế, 54 features)

Quy mô N	Số cụm K	Global Scan (ms)	Cached Query (ms)	Tăng tốc
1.000	1	8,36	7,98	1,05×
10.000	4	68,95	37,97	1,82×
50.000	17	392,47	50,03	7,84×
100.000	34	839,02	72,99	11,50×

Đáng chú ý, ở quy mô nhỏ ($N < 3.000$), lợi ích của phân cụm là không đáng kể (tăng tốc ≈ 1) do toàn bộ dữ liệu được gộp vào một cụm duy nhất ($K = 1$). Tuy nhiên, khi dữ liệu mở rộng, độ phức tạp tính toán của hệ thống chuyển dịch hiệu quả từ $O(N)$ sang $O(K + T)$, trong đó K và T là các hằng số được kiểm soát.

Kết quả này chứng minh tính khả thi của việc triển khai hệ thống trên các hạ tầng phần cứng phổ thông (như CPU Intel Core i5 dòng tiết kiệm điện) mà vẫn đảm bảo khả năng phục vụ hàng trăm người dùng đồng thời.

4.5 Kịch bản 4: Đánh giá Chatbot và Bảo mật PII

Kịch bản cuối cùng đánh giá khả năng của hệ thống EduTwin trong việc tích hợp mô hình ngôn ngữ lớn (LLM) để tạo ra trải nghiệm cá nhân hóa nhưng vẫn đảm bảo an toàn thông tin (Personalized & Secure Experience).

4.5.1 Kiểm thử PII Redaction (Bảo mật)

Mục tiêu của kiểm thử là đảm bảo các thông tin định danh cá nhân (PII) như Email, Số điện thoại và Tên riêng được tự động loại bỏ hoặc thay thế trước khi ngữ cảnh (context)

CHƯƠNG 4. THỰC NGHIỆM VÀ ĐÁNH GIÁ

được gửi sang LLM. Kết quả thực nghiệm tại Bảng 4.4 xác nhận module PII Redaction hoạt động chính xác, chuyển đổi các dữ liệu nhạy cảm thành các token ẩn danh (như <EMAIL_REDACTED>), qua đó giảm thiểu tối đa rủi ro rò rỉ dữ liệu người dùng.

Loại PII	Dữ liệu đầu vào	Dữ liệu sau xử lý
Email	nguyenvana@gmail.com<EMAIL_REDACTED>	
Số điện thoại	0912345678	<PHONE_REDACTED>
Tên riêng	Nguyễn Văn An	<NAME_REDACTED>

Bảng 4.4. Hiệu quả của module PII Redaction

4.5.2 Kiểm thử RAG Context Injection (Cá nhân hóa)

Kiểm thử này đánh giá tác động của việc thêm dữ liệu điểm số và hồ sơ người dùng vào Prompt để điều chỉnh hành vi của AI. Kết quả so sánh (Bảng 4.5) cho thấy việc bổ sung ngữ cảnh đầy đủ (Mức 2) tuy làm tăng độ dài Prompt thêm 46% (từ 337 lên 494 tokens) nhưng đã thay đổi đáng kể chất lượng phản hồi. Thay vì chỉ phân tích số liệu khô khan, AI chuyển sang kết nối mục tiêu học tập với sở thích cá nhân, đồng thời thay đổi văn phong từ chuyên nghiệp sang thân thiện, năng động phù hợp với giới trẻ.

Tiêu chí	Mức 1: Chỉ có điểm số	Mức 2: Full Profile Cá nhân
Prompt Token	337 tokens	494 tokens (+46%)
Dữ liệu đầu vào	14 đầu điểm	Điểm số + 6 thuộc tính (Kinesthetic, Hướng ngoại, Mục tiêu Sư phạm, ...)
Chiến lược AI	Phân tích dựa trên số liệu: tập Đè xuất dạy lại cho bạn bè (phù trung cải thiện môn yêu (Hóa) hợp định hướng Sư phạm), viết blog và duy trì môn mạnh (Toán). tiếng Anh (phù hợp sở thích).	
Văn phong	Chuyên nghiệp, phân tích.	Vui vẻ, năng động, sử dụng emoji

Bảng 4.5. So sánh Prompt và Phản hồi giữa các mức độ Context

Về mặt định tính, hệ thống đã thể hiện khả năng thích ứng linh hoạt với từng *persona* người học. Trong kịch bản cơ bản (chỉ có điểm số), AI đề xuất lộ trình chuẩn 4 bước để

CHƯƠNG 4. THỰC NGHIỆM VÀ ĐÁNH GIÁ

cải thiện môn Hóa. Tuy nhiên, khi chuyển sang hồ sơ "Visual Learner" (người học qua thị giác), AI lập tức thay đổi chiến lược, đề xuất sử dụng Sơ đồ tư duy và Flashcards màu sắc. Đối với hồ sơ "Kinesthetic Learner" (người học qua vận động) có tính cách hướng ngoại, hệ thống lại gợi ý phương pháp Role-play và viết Blog, chứng tỏ độ nhạy ngữ cảnh (Context Sensitivity) cao. Dù prompt dài hơn, thời gian phản hồi vẫn đảm bảo trải nghiệm mượt mà.

4.6 Bàn luận chung

Tổng hợp lại, chuỗi thực nghiệm trên môi trường lai (Kaggle/Local) đã chứng minh tính khả thi và hiệu quả vượt trội của kiến trúc EduTwin đề xuất trên bốn khía cạnh chính.

Thứ nhất, về hiệu quả của Lazy Learning trên dữ liệu giáo dục: Kết quả từ Kịch bản 2 đã bác bỏ giả thuyết phổ biến rằng “mô hình càng phức tạp thì càng hiệu quả”. Việc KNN ($k = 15$) đạt độ chính xác cao nhất ($R^2 \approx 66.25\%$) đồng thời sở hữu tốc độ suy diễn nhanh vượt trội so với LSTM hay XGBoost đã khẳng định rằng dữ liệu điểm số học sinh mang tính chất *cục bộ (local structure)* mạnh mẽ, hoàn toàn tương thích với các thuật toán dựa trên láng giềng.

Thứ hai, về khả năng tối ưu hóa tài nguyên: Bằng chiến lược chọn Lazy Learning làm mô hình lõi kết hợp với kỹ thuật Cluster Indexing (Kịch bản 3), hệ thống EduTwin loại bỏ hoàn toàn sự phụ thuộc vào GPU trong giai đoạn vận hành. Với khả năng tăng tốc tới 218 lần và thời gian phản hồi chỉ tính bằng mili-giây, hệ thống hoạt động mượt mà ngay cả trên các thiết bị cấu hình thấp, mở ra khả năng triển khai rộng rãi với chi phí thấp.

Thứ ba, về trải nghiệm người dùng và bảo mật: Kịch bản 4 chứng minh sự kết hợp giữa LLM API và kỹ thuật RAG Context Injection đã biến EduTwin từ một công cụ thông kê đơn thuần thành một trợ lý ảo thấu hiểu người dùng. Hệ thống không chỉ đưa ra lời khuyên “đúng người, đúng thời điểm” dựa trên phong cách học tập và mục tiêu cá nhân mà còn đảm bảo an toàn tuyệt đối nhờ cơ chế PII Redaction tự động.

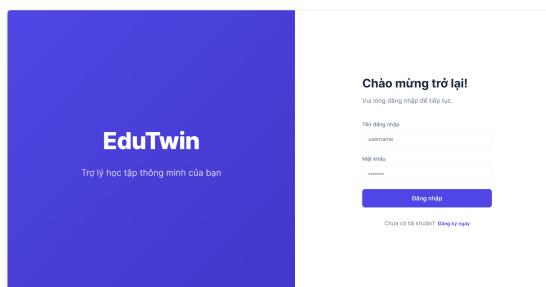
Cuối cùng, về độ tin cậy của dữ liệu: Thuật toán KNN Imputation (Kịch bản 1) đã đóng vai trò nền tảng vững chắc khi khôi phục thành công cấu trúc tương quan của dữ liệu khuyết ($R^2 \approx 0.73$). Điều này đảm bảo nguồn dữ liệu đầu vào chất lượng cao, tạo tiền đề cho độ chính xác của toàn bộ chuỗi dự báo phía sau.

CHƯƠNG 4. THỰC NGHIỆM VÀ ĐÁNH GIÁ

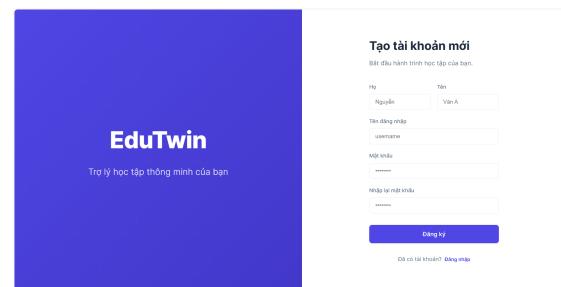
4.7 Kết quả xây dựng ứng dụng

4.7.1 Giao diện đăng nhập/đăng ký

Giao diện xác thực người dùng được thiết kế theo phong cách hiện đại, tối giản, phân chia rõ ràng giữa khu vực nhận diện thương hiệu và khu vực nhập liệu. Để đảm bảo an toàn và trải nghiệm liền mạch, hệ thống tích hợp các cơ chế kiểm tra tính hợp lệ (validation) chặt chẽ ngay từ đầu vào đối với tên đăng nhập và mật khẩu. Người dùng sẽ được tự động chuyển hướng đến trang chatbot ngay sau khi quy trình xác thực thành công.



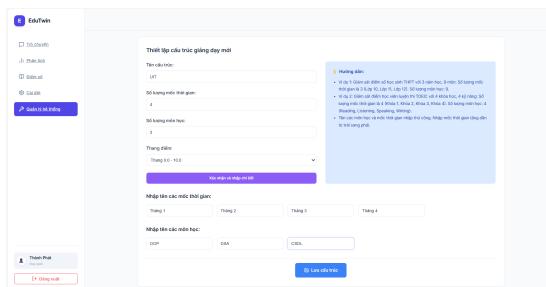
Hình 4.3. Giao diện đăng nhập



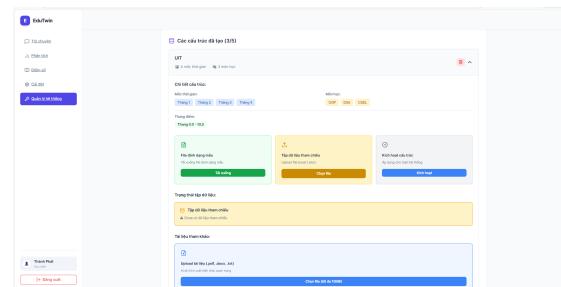
Hình 4.4. Giao diện đăng ký

4.7.2 Phân hệ Quản trị viên

Phân hệ quản trị cung cấp khả năng tùy biến sâu rộng thông qua tính năng định nghĩa cấu trúc giảng dạy. Quản trị viên có toàn quyền thiết lập số lượng mốc thời gian, danh sách môn học và loại thang điểm (thang 10, 100 hoặc Grade Point Average (GPA)) để phù hợp với đặc thù của từng hệ thống giáo dục. Ngoài ra, hệ thống cũng hỗ trợ các công cụ quản lý dữ liệu nền tảng, cho phép tải lên tập dữ liệu mẫu huấn luyện và cập nhật các tài liệu tham khảo chuyên sâu phục vụ cho module RAG.



Hình 4.5. Tính năng cho phép quản trị viên (Admin) định nghĩa cấu trúc giảng dạy

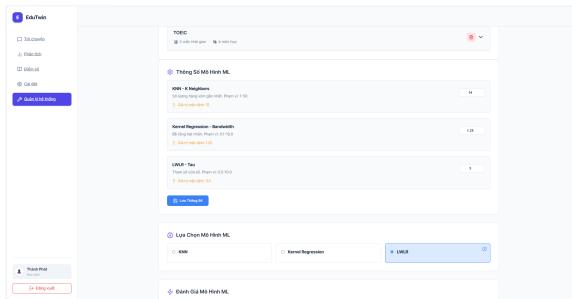


Hình 4.6. Giao diện cài đặt cho các cấu trúc giảng dạy đã thiết lập

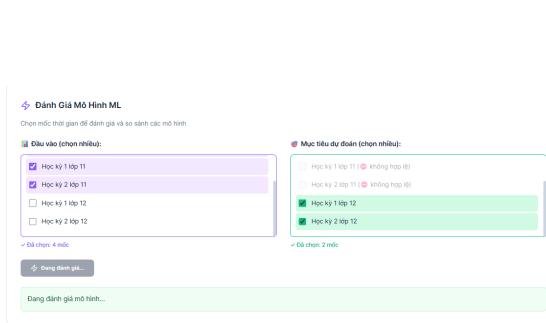
Song song với quản lý cấu trúc, quản trị viên có thể can thiệp trực tiếp vào lõi công nghệ thông qua tính năng cấu hình mô hình Machine Learning. Hệ thống cho phép lựa chọn linh

CHƯƠNG 4. THỰC NGHIỆM VÀ ĐÁNH GIÁ

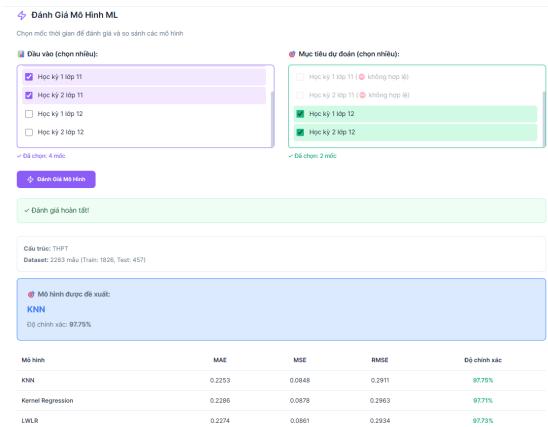
hoạt giữa các mô hình và tinh chỉnh các siêu tham số (hyperparameters) nhằm tối ưu hóa độ chính xác. Hiệu suất mô hình sau đó được đánh giá minh bạch thông qua quá trình kiểm thử với đầu vào/đầu ra động. Kết quả hiển thị trực quan qua các chỉ số thống kê chuẩn (MAE, RMSE, R²) và chỉ số Pseudo-Accuracy – một thước đo định lượng đơn giản hóa giúp người dùng không chuyên dễ dàng ra quyết định.



Hình 4.7. Các tính năng cho phép cài đặt hệ thống ML như lựa chọn mô hình, lựa chọn siêu tham số



Hình 4.8. Tính năng đánh giá mô hình tùy biến với input features và output labels thay đổi theo cấu trúc giảng dạy được thiết lập

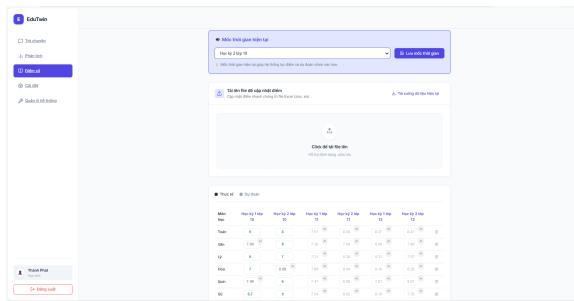


Hình 4.9. Kết quả đánh giá mô hình trực quan với đầy đủ tham số và đề xuất phù hợp

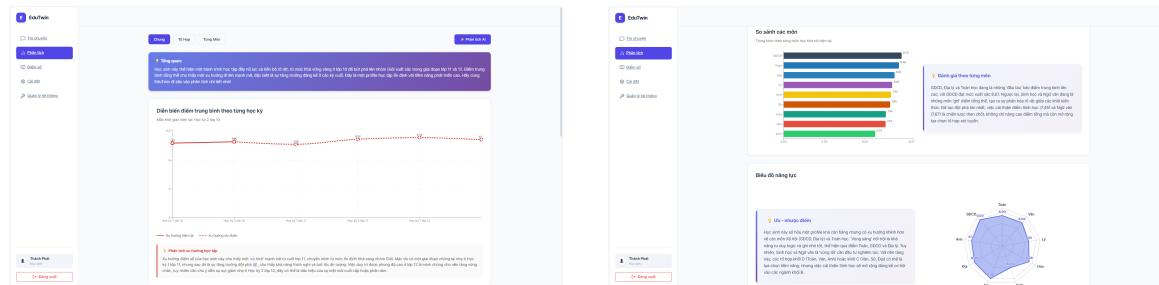
4.7.3 Phân hệ Người dùng

Tại giao diện quản lý kết quả học tập, người dùng có thể nhập liệu trực tiếp trên lưới hoặc sử dụng tính năng Import/Export Excel. Nhờ công nghệ WebSocket, dữ liệu được đồng bộ thời gian thực với luồng xử lý Machine Learning, cho phép hệ thống hiển thị ngay lập tức các giá trị dự báo tương lai song song với điểm số thực tế. Dữ liệu này sau đó được trực quan hóa đa chiều qua các biểu đồ (đường, cột, radar), giúp người dùng dễ dàng theo dõi xu hướng. Đặc biệt, module phân tích AI (AI Insights) được tích hợp để tự động trích xuất thông tin ẩn, đưa ra nhận xét và đề xuất cải thiện dựa trên các mẫu hình dữ liệu ghi nhận được.

CHƯƠNG 4. THỰC NGHIỆM VÀ ĐÁNH GIÁ



Hình 4.10. Giao diện tổng hợp và quản lý kết quả học tập hiện tại của học sinh cũng như các dự đoán của hệ thống

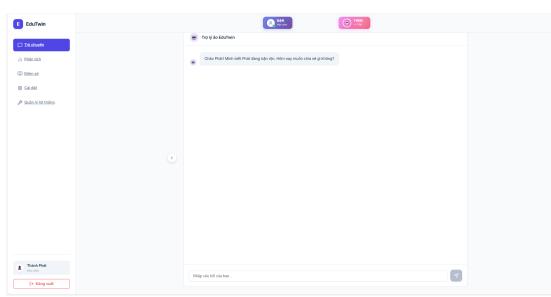


Hình 4.11. Giao diện biểu đồ trực quan hóa với phân tích từ AI giúp người dùng khái quát thông tin và nắm bắt các thông tin ẩn của dữ liệu

4.7.4 Giao diện Chatbot AI

Trợ lý ảo EduTwin sở hữu giao diện thân thiện với khả năng phản hồi thời gian thực (streaming response) và hiển thị nội dung đa dạng qua định dạng Markdown. Điểm nổi bật là khả năng tra cứu tài liệu chuyên sâu (RAG) kết hợp với việc tự động nạp ngữ cảnh điểm số của người dùng, giúp các câu trả lời luôn bám sát thực tế học tập. Bên cạnh đó, hệ thống liên tục học hỏi để xây dựng hồ sơ cá nhân hóa (tính cách, thói quen, mục tiêu), từ đó tinh chỉnh phong cách tương tác phù hợp nhất với từng học sinh.

CHƯƠNG 4. THỰC NGHIỆM VÀ ĐÁNH GIÁ



Hình 4.12. Giao diện chatbot thân thiện

Cá nhân hóa tự động

Chatbot đã tự động học phong cách sử dụng của bạn từ các cuộc trò chuyện:

- [Phong cách giao tiếp] Ưa thích câu trả lời ngắn gọn
- [Phong cách giao tiếp] Sử dụng ngôn ngữ trang trọng
- [Cảm xúc] Đang có áp lực học tập
- [Sở thích] Quan tâm nhiều đến kết quả học tập
- [Mục tiêu] Có xu hướng đặt mục tiêu rõ ràng

⇒ Các cá nhân hóa này được cập nhật tự động để twin phục vụ bạn tốt hơn.

Hình 4.13. Danh sách các thông tin cá nhân hóa mà trợ lý ảo nắm bắt được qua quá trình trò chuyện

Chương 5. KẾT LUẬN VÀ HƯỚNG PHÁT TRIỂN

5.1 KẾT LUẬN VÀ HƯỚNG PHÁT TRIỂN

5.1.1 Kết luận chung

Khóa luận đã nghiên cứu và xây dựng thành công hệ thống *EduTwin – Bản sao Học tập Kỹ thuật số tích hợp AI*, giải quyết trọn vẹn bài toán cá nhân hóa giáo dục và dự báo hiệu suất học tập trong thời gian thực. Thông qua quá trình nghiên cứu lý thuyết, thiết kế kiến trúc và thực nghiệm trên bộ dữ liệu thực tế, khóa luận đã đạt được những kết quả quan trọng sau:

Chứng minh hiệu quả vượt trội của Lazy Learning trong giáo dục: Kết quả thực nghiệm đã bác bỏ quan điểm cho rằng mô hình càng phức tạp (Deep Learning) thì càng chính xác. Với đặc thù dữ liệu điểm số học sinh mang tính cấu trúc cục bộ và tương quan lảng giềng mạnh mẽ, thuật toán KNN với $k = 15$ đã đạt độ chính xác cao nhất ($R^2 \approx 66.25\%$), vượt qua cả LSTM và XGBoost, đồng thời có chi phí tính toán thấp hơn đáng kể.

Giải quyết bài toán độ trễ và khả năng mở rộng: Bằng cách áp dụng kiến trúc lai ghép với kỹ thuật Lượng tử hóa Vector và Tối ưu hóa mẫu hình (Cluster Indexing & Vector Quantization), hệ thống đã khắc phục nhược điểm về tốc độ của Lazy Learning. Kết quả Stress Test cho thấy hệ thống đạt tốc độ tăng tốc (speedup) lên tới 218 lần ở quy mô 100.000 bản ghi, duy trì độ trễ dưới 10ms, đảm bảo khả năng triển khai thực tế trên hạ tầng phần cứng phổ thông.

Kiến trúc linh hoạt hướng Siêu dữ liệu (Metadata-Driven): EduTwin không bị giới hạn bởi một chương trình học cố định nhờ thiết kế cơ sở dữ liệu linh hoạt sử dụng PostgreSQL JSONB. Hệ thống cho phép các đơn vị giáo dục tự định nghĩa cấu trúc môn học, thang điểm và quy chế đánh giá mà không cần can thiệp vào mã nguồn hay huấn luyện lại mô hình AI.

Trải nghiệm tương tác thông minh và bảo mật: Việc tích hợp LLM với kỹ thuật RAG (Retrieval-Augmented Generation) và cơ chế PII Redaction đã biến hệ thống từ một công cụ thông kê thành một trợ lý ảo thấu hiểu ngữ cảnh. EduTwin không chỉ bảo vệ an toàn thông tin định danh của người học mà còn đưa ra các tư vấn sư phạm được cá nhân hóa sâu sắc dựa trên hồ sơ năng lực và phong cách học tập (Visual/Kinesthetic).

5.1.2 Các hạn chế của đề tài

Bên cạnh những kết quả đạt được, khóa luận vẫn còn một số hạn chế cần được nhìn nhận:

CHƯƠNG 5. KẾT LUẬN VÀ HƯỚNG PHÁT TRIỂN

Phạm vi dữ liệu: Hiện tại, hệ thống mới chỉ được kiểm chứng trên dữ liệu điểm số định lượng (Structured Grades). Các dữ liệu phi cấu trúc quan trọng khác như lịch sử điểm danh, hành vi click chuột trên LMS hay các hoạt động ngoại khóa chưa được đưa vào mô hình dự báo.

Giới hạn của thuật toán KNN: Mặc dù hoạt động rất tốt trong việc nội suy (interpolation) với dữ liệu phân phối dày, KNN có thể gặp khó khăn trong việc ngoại suy (extrapolation) đối với các trường hợp học sinh có hồ sơ năng lực quá đặc biệt hoặc nằm ngoài vùng phủ của dữ liệu huấn luyện (outliers).

Phụ thuộc vào API bên thứ ba: Chức năng Chatbot hiện tại phụ thuộc hoàn toàn vào API của các nhà cung cấp LLM (Google/OpenAI). Điều này có thể dẫn đến rủi ro về chi phí vận hành tăng cao khi quy mô người dùng lớn, cũng như vấn đề về bảo mật thông tin.

5.1.3 Hướng phát triển

Để hoàn thiện và nâng cao khả năng ứng dụng của EduTwin trong thực tiễn, hướng phát triển trong tương lai sẽ tập trung vào các nội dung sau:

Mở rộng nguồn dữ liệu và Mô hình Hybrid: Nghiên cứu tích hợp thêm các nguồn dữ liệu hành vi (Behavioral Data) từ các hệ thống LMS. Đồng thời, xem xét phát triển mô hình lai (Hybrid Model) kết hợp giữa Lazy Learning (cho dự báo ngắn hạn, cục bộ) và Deep Learning (cho dự báo xu hướng dài hạn) để tận dụng ưu điểm của cả hai phương pháp.

Tối ưu hóa chi phí LLM (Small Language Models): Nghiên cứu tinh chỉnh (Fine-tuning) các mô hình ngôn ngữ nhỏ hơn (SLM) hoặc các mô hình mã nguồn mở (như Llama, Mistral) để có thể tự triển khai (Self-hosted) ngay trên máy chủ của trường học. Điều này giúp giảm chi phí API, tăng tốc độ phản hồi và đảm bảo quyền riêng tư dữ liệu tuyệt đối.

Phát triển Hệ sinh thái Đa nền tảng: Mở rộng EduTwin từ ứng dụng Web sang ứng dụng di động (Mobile App) để tăng tính tiện lợi cho học sinh. Đồng thời, xây dựng thêm phân hệ dành cho Phụ huynh và Giáo viên, tạo thành một vòng tròn khép kín giữa Gia đình – Nhà trường – Học sinh, giúp việc giám sát và hỗ trợ người học trở nên đồng bộ và hiệu quả hơn.

Cơ chế Tự động hóa Tham số (Auto-Tuning): Hiện tại các tham số như K (số láng giềng) hay τ (bandwidth) đang được chọn dựa trên thực nghiệm hoặc cấu hình thủ công. Hướng phát triển tiếp theo là xây dựng cơ chế AutoML để hệ thống tự động học và điều chỉnh các tham số này theo thời gian thực dựa trên sự biến đổi của dữ liệu mới nạp vào.

Phụ lục A. CÁC THÔNG TIN LIÊN QUAN

A.1 Tài nguyên thực nghiệm

Các tài nguyên phục vụ cho quá trình thực nghiệm của hệ thống EduTwin được liệt kê như sau:

- **Tập dữ liệu thực nghiệm:** [Google Sheets Dataset](#)
- **Notebook huấn luyện và đánh giá mô hình:** [Training & Evaluation Notebook](#)
- **Thông tin chi tiết Prompt và Response (Kịch bản 4):** [LLM Prompt & Response Documentation](#)

TÀI LIỆU THAM KHẢO

- [1] Thủ tướng Chính phủ. *Quyết định số 131/QĐ-TTg: Phê duyệt Đề án "Tăng cường ứng dụng công nghệ thông tin và chuyển đổi số trong giáo dục và đào tạo giai đoạn 2022–2025, định hướng đến năm 2030".* <https://vanban.chinhphu.vn/?pageid=27160&docid=205236&classid=0>. Người ký: Vũ Đức Đam. Jan. 2022.
- [2] Hossein Omrany, K. Al-Obaidi, Amirhosein Ghaffarianhoseini, Rui-Dong Chang, Chansik Park, và F. Rahimian. “Digital twin technology for education, training and learning in construction industry: implications for research and practice”. In: *Engineering, Construction and Architectural Management* (2025). DOI: [10.1108/ecam-10-2024-1376](https://doi.org/10.1108/ecam-10-2024-1376).
- [3] Jie Zhang, Jingdong Zhu, Weiwei Tu, Minkai Wang, Yiling Yang, Fang Qian, và Yeqing Xu. “The Effectiveness of a Digital Twin Learning System in Assisting Engineering Education Courses: A Case of Landscape Architecture”. In: *Applied Sciences* (2024). DOI: [10.3390/app14156484](https://doi.org/10.3390/app14156484).
- [4] L. Rovati, Phillip J. Gary, E. Cubro, Yue Dong, O. Kilickaya, Phillip J. Schulte, Xiang Zhong, M. Wörster, D. Kelm, O. Gajic, Alexander S. Niven, và Amos Lal. “Development and usability testing of a patient digital twin for critical care education: a mixed methods study”. In: *Frontiers in Medicine* 10 (2024). DOI: [10.3389/fmed.2023.1336897](https://doi.org/10.3389/fmed.2023.1336897).
- [5] Yuwei Tao và Aijuan Xie. “Research on the architecture and practice of inquiry learning model integrated with digital twin”. In: *Interactive Learning Environments* (2025). DOI: [10.1080/10494820.2025.2476716](https://doi.org/10.1080/10494820.2025.2476716).
- [6] Jean Baptiste Habarurema, Raffaele Di Fuccio, và P. Limone. “Enhancing e-learning with a digital twin for innovative learning”. In: *The International Journal of Information and Learning Technology* (2025). DOI: [10.1108/ijilt-02-2024-0034](https://doi.org/10.1108/ijilt-02-2024-0034).
- [7] W. Villegas-Ch., Diego Buenaño-Fernández, Alexandra Maldonado Navarro, và Aracely Mera-Navarrete. “Adaptive intelligent tutoring systems for STEM education: analysis of the learning impact and effectiveness of

- personalized feedback”. In: *Smart Learn. Environ.* 12 (2025), p. 41. DOI: [10.1186/s40561-025-00389-y](https://doi.org/10.1186/s40561-025-00389-y).
- [8] Meltem Taşkın. “Artificial Intelligence in Personalized Education: Enhancing Learning Outcomes Through Adaptive Technologies and Data-Driven Insights”. In: *Human Computer Interaction* (2025). DOI: [10.62802/ygye0506](https://doi.org/10.62802/ygye0506).
- [9] Ani Grubišić, Ines Šarić-Grgić, A. Gašpar, và Branko Žitko. “Usability Evaluation of an Adaptive Courseware Approach in the Natural Language-Based Intelligent Tutoring System-Tutomat”. In: *Journal of Computer Assisted Learning* (2025). DOI: [10.1111/jcal.70071](https://doi.org/10.1111/jcal.70071).
- [10] Ramesh Singh, Chenlep Yakha Konyak, và Akangjungshi Longkumer. “A Multi-Access Edge Computing Approach to Intelligent Tutoring Systems for Real-Time Adaptive Learning”. In: *International Journal of Information Technology* (2025). DOI: [10.1007/s41870-025-02460-w](https://doi.org/10.1007/s41870-025-02460-w).
- [11] Shahzad Rizwan, Chee Ken Nee, và Salem Garfan. “Identifying the Factors Affecting Student Academic Performance and Engagement Prediction in MOOC Using Deep Learning: A Systematic Literature Review”. In: *IEEE Access* 13 (2025), pp. 18952–18982. DOI: [10.1109/access.2025.3533915](https://doi.org/10.1109/access.2025.3533915).
- [12] Jialun Pan, Zhanzhan Zhao, và Dongkun Han. “Academic Performance Prediction Using Machine Learning Approaches: A Survey”. In: *IEEE Transactions on Learning Technologies* 18 (2025), pp. 351–368. DOI: [10.1109/tlt.2025.3554174](https://doi.org/10.1109/tlt.2025.3554174).
- [13] Qi Lang, Minjuan Wang, Minghao Yin, Shuang Liang, và Wenzhuo Song. “Transforming Education With Generative AI (GAI): Key Insights and Future Prospects”. In: *IEEE Transactions on Learning Technologies* 18 (2025), pp. 230–242. DOI: [10.1109/tlt.2025.3537618](https://doi.org/10.1109/tlt.2025.3537618).
- [14] Michail N. Giannakos, Roger Azevedo, Peter Brusilovsky, M. Cukurova, Y. Dimitriadis, Davinia Hernández Leo, Sanna Järvelä, M. Mavrikis, và Bart Rienties. “The promise and challenges of generative AI in education”. In: *Behav. Inf. Technol.* 44 (2024), pp. 2518–2544. DOI: [10.1080/0144929x.2024.2394886](https://doi.org/10.1080/0144929x.2024.2394886).
- [15] V. Nikolovski, D. Trajanov, và Ivan Chorbev. “Advancing AI in Higher Education: A Comparative Study of Large Language Model-Based Agents for Exam Question Generation, Improvement, and Evaluation”. In: *Algorithms* (2025). DOI: [10.3390/a18030144](https://doi.org/10.3390/a18030144).

TÀI LIỆU THAM KHẢO

- [16] M. B. Jelodar. “Generative AI, Large Language Models, and ChatGPT in Construction Education, Training, and Practice”. In: *Buildings* (2025). DOI: [10.3390/buildings15060933](https://doi.org/10.3390/buildings15060933).
- [17] Xiaojun Xu, Yixiao Chen, và Jing Miao. “Opportunities, challenges, and future directions of large language models, including ChatGPT in medical education: a systematic scoping review”. In: *Journal of Educational Evaluation for Health Professions* 21 (2024). DOI: [10.3352/jeehp.2024.21.6](https://doi.org/10.3352/jeehp.2024.21.6).
- [18] Zachary K. Collier, Kamal Chawla, và Olushola O. Soyoye. “Optimizing Imputation for Educational Data: Exploring Training Partition and Missing Data Ratios”. In: *The Journal of Experimental Education* 93 (2024), pp. 607–627. DOI: [10.1080/00220973.2023.2287447](https://doi.org/10.1080/00220973.2023.2287447).
- [19] Gabriel-Vasilică Sasu, Bogdan-Iulian Ciubotaru, Nicolae Goga, và A. Vasilăteanu. “Addressing Missing Data Challenges in Geriatric Health Monitoring: A Study of Statistical and Machine Learning Imputation Methods”. In: *Sensors (Basel, Switzerland)* 25 (2025). DOI: [10.3390/s25030614](https://doi.org/10.3390/s25030614).
- [20] Abdulaziz Altamimi, Aisha Ahmed Alarfaj, Muhammad Umer, E. Alabdulqader, Shtwai Alsubai, Tai-hoon Kim, và Imran Ashraf. “An automated approach to predict diabetic patients using KNN imputation and effective data mining techniques”. In: *BMC Medical Research Methodology* 24 (2024). DOI: [10.1186/s12874-024-02324-0](https://doi.org/10.1186/s12874-024-02324-0).
- [21] Khaled Alnowaiser. “Improving Healthcare Prediction of Diabetic Patients Using KNN Imputed Features and Tri-Ensemble Model”. In: *IEEE Access* 12 (2024), pp. 16783–16793. DOI: [10.1109/access.2024.3359760](https://doi.org/10.1109/access.2024.3359760).