TRƯỜNG ĐẠI HỌC KỸ THUẬT CÔNG NGHIỆP

**KHOA ĐIỆN TỬ**

**Bộ môn: Công nghệ Thông tin**.

**BÀI TẬP KẾT THÚC MÔN HỌC**

MÔN HỌC

**KHOA HỌC DỮ LIỆU**

Sinh viên: VŨ TRƯỜNG QUANG. . . . . . . . . . . . . . . . . . . . . . . . . .

Lớp: K57-KMT. . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . .

Giáo viên GIẢNG DẠY: NGUYỄN VĂN HUY. . . . . . . . . . . .

Link GitHub: <https://github.com/thanh123ab/KhoaHocDuLieu>

**Thái Nguyên – 2025**

|  |  |
| --- | --- |
| **TRƯỜNG ĐHKTCN** | **CỘNG HOÀ XÃ HỘI CHỦ NGHĨA VIỆT NAM** |
| **KHOA ĐIỆN TỬ** | ***Độc lập - Tự do - Hạnh phúc*** |

**BÀI TẬP KẾT THÚC MÔN HỌC**

**MÔN HỌC: KHOA HỌC DỮ LIỆU**

BỘ MÔN : CÔNG NGHỆ THÔNG TIN

*Sinh viên: Vũ Trường Quang*

*Lớp*: K57KMT ……………………………………………. *Ngành:Kỹ Thuật Máy Tính*

*Giáo viên hướng dẫn:Nguyễn Văn Huy*

*Ngày giao đề :20/05/2025*  *Ngày hoàn thành :26/05/2025*

*Tên đề tài : Xây dựng web app phân tích và dự báo dân số theo vùng địa lý.*

*Yêu cầu :*

***Đầu vào:***

* *Dữ liệu dân số từ [World Population Data](https://www.kaggle.com/datasets/iamsouravbanerjee/world-population-dataset)*

***Đầu ra:***

* *Dự báo dân số và biểu đồ tăng trưởng dân số theo năm.*

***Các tính năng:***

* *Dự báo dân số bằng Time Series*
* *Visualization dân số theo quốc gia*

|  |
| --- |
| **GIÁO VIÊN HƯỚNG DẪN** |
| *(Ký và ghi rõ họ tên)* |

**NHẬN XÉT CỦA GIÁO VIÊN HƯỚNG DẪN**

Thái Nguyên, ngày….tháng…..năm 20....

GIÁO VIÊN HƯỚNG DẪN

*(Ký ghi rõ họ tên)*

MỤC LỤC

**[Chương I: Mở đầu 5](#_Toc2092)**

**[1.1. Lý do chọn đề tài 5](#_Toc2555)**

**[1.2. Mục tiêu đề tài 5](#_Toc2451)**

**[1.3. Nội dung và phạm vi thực hiện 5](#_Toc9709)**

**[Chương II: Cơ sở lý thuyết 6](#_Toc16618)**

**[2.1. Tập dữ liệu Student Performance 6](#_Toc7044)**

**[2.2. Thư viện và công cụ sử dụng 6](#_Toc22011)**

**[2.3. Các thuật toán áp dụng 6](#_Toc8440)**

**[2.4. Các bước tiền xử lý dữ liệu 7](#_Toc1354)**

**[Chương III: Thiết kế và xây dựng hệ thống 7](#_Toc11049)**

**[3.1. Kiến trúc hệ thống 7](#_Toc6395)**

**[3.2. Các module chính 7](#_Toc26982)**

**[3.3. Thuật toán Random Forest 7](#_Toc16887)**

**[3.4. Cấu trúc dữ liệu 8](#_Toc21627)**

**[Chương IV: Thực nghiệm 8](#_Toc6967)**

**[4.1. Trực quan hóa dữ liệu 8](#_Toc4955)**

**[4.2. Huấn luyện và đánh giá mô hình 9](#_Toc12376)**

**[4.3. Dự đoán điểm cho sinh viên mới 10](#_Toc18902)**

**[4.4. Phân tích độ quan trọng đặc trưng 10](#_Toc27114)**

**[4.5. Phân tích điểm số theo nhóm 11](#_Toc19002)**

**[Chương V: Kết luận và hướng phát triển 12](#_Toc5396)**

**[5.1. Kết luận 12](#_Toc10456)**

**[5.2. Bài học kinh nghiệm 12](#_Toc10629)**

**[5.3. Hướng phát triển 13](#_Toc26812)**

**Chương I: Mở đầu**

* 1. **Lý do chọn đề tài**

Trong bối cảnh giáo dục hiện đại, việc áp dụng các công nghệ như Trí tuệ nhân tạo và Học máy vào đánh giá, dự báo kết quả học tập đang ngày càng phổ biến. Việc dự đoán điểm thi không chỉ giúp sinh viên nhận biết được năng lực hiện tại mà còn hỗ trợ giáo viên có kế hoạch giảng dạy phù hợp.

**1.2. Mục tiêu đề tài**

Xây dựng một hệ thống sử dụng mô hình học máy để dự báo điểm thi của sinh viên dựa trên các yếu tố đầu vào như giới tính, nhóm dân tộc, trình độ học vấn của phụ huynh, tình trạng bữa trưa và việc ôn tập.

**1.3. Nội dung và phạm vi thực hiện**

* Phân tích và tiền xử lý dữ liệu Student Performance Dataset.
* Trực quan hóa dữ liệu để khám phá các mối quan hệ tiềm ẩn.
* Áp dụng các mô hình hồi quy để dự đoán điểm thi.
* Đánh giá hiệu quả mô hình và phân tích đặc trưng.

**Chương II: Cơ sở lý thuyết**

**2.1. Tập dữ liệu Student Performance**

Dữ liệu bao gồm thông tin về học sinh như giới tính, nhóm dân tộc, trình độ học vấn của phụ huynh, tình trạng bữa trưa, việc ôn tập và điểm thi 3 môn: Toán, Đọc hiểu và Viết.

**2.2. Thư viện và công cụ sử dụng**

a) Pandas: Xử lý dữ liệu dạng bảng.  
b) Scikit-learn: Xây dựng mô hình học máy.  
c) Matplotlib/Seaborn: Trực quan hóa dữ liệu.  
d) LabelEncoder, StandardScaler: Mã hóa và chuẩn hóa dữ liệu.

e) Jupyter Notebook: Môi trường phát triển và chạy mã.

**2.3. Các thuật toán áp dụng**

a) Linear Regression: Mô hình tuyến tính đơn giản để dự đoán giá trị liên tục.  
b) Random Forest Regressor: Mô hình mạnh, ít bị overfitting, kết hợp nhiều cây quyết định.

**2.4. Các bước tiền xử lý dữ liệu**

* Kiểm tra và xử lý dữ liệu thiếu.
* Mã hóa các biến phân loại bằng LabelEncoder.
* Chuẩn hóa dữ liệu bằng StandardScaler để cải thiện hiệu suất mô hình.

**Chương III: Thiết kế và xây dựng hệ thống**

**3.1. Kiến trúc hệ thống**

Hệ thống được chia làm 4 module chính:  
a) Nhập liệu: Đọc dữ liệu CSV, kiểm tra tính hợp lệ.  
b) Tiền xử lý: Mã hóa, chuẩn hóa, chia tập train/test.  
c) Huấn luyện: Xây dựng mô hình Linear Regression và Random Forest.  
d) Dự đoán và đánh giá: Hiển thị kết quả, biểu đồ và độ quan trọng đặc trưng.

**3.2. Các module chính**

* load\_data(): Đọc và kiểm tra dữ liệu.
* preprocess\_data(): Mã hóa và chuẩn hóa.
* train\_model(): Huấn luyện mô hình.
* predict(): Dự đoán điểm thi mới.

**3.3. Thuật toán Random Forest**

Thuật toán xây dựng nhiều cây quyết định song song, sau đó tổng hợp kết quả từ các cây để tạo ra dự đoán cuối cùng. Điều này giúp giảm phương sai và tăng độ chính xác.

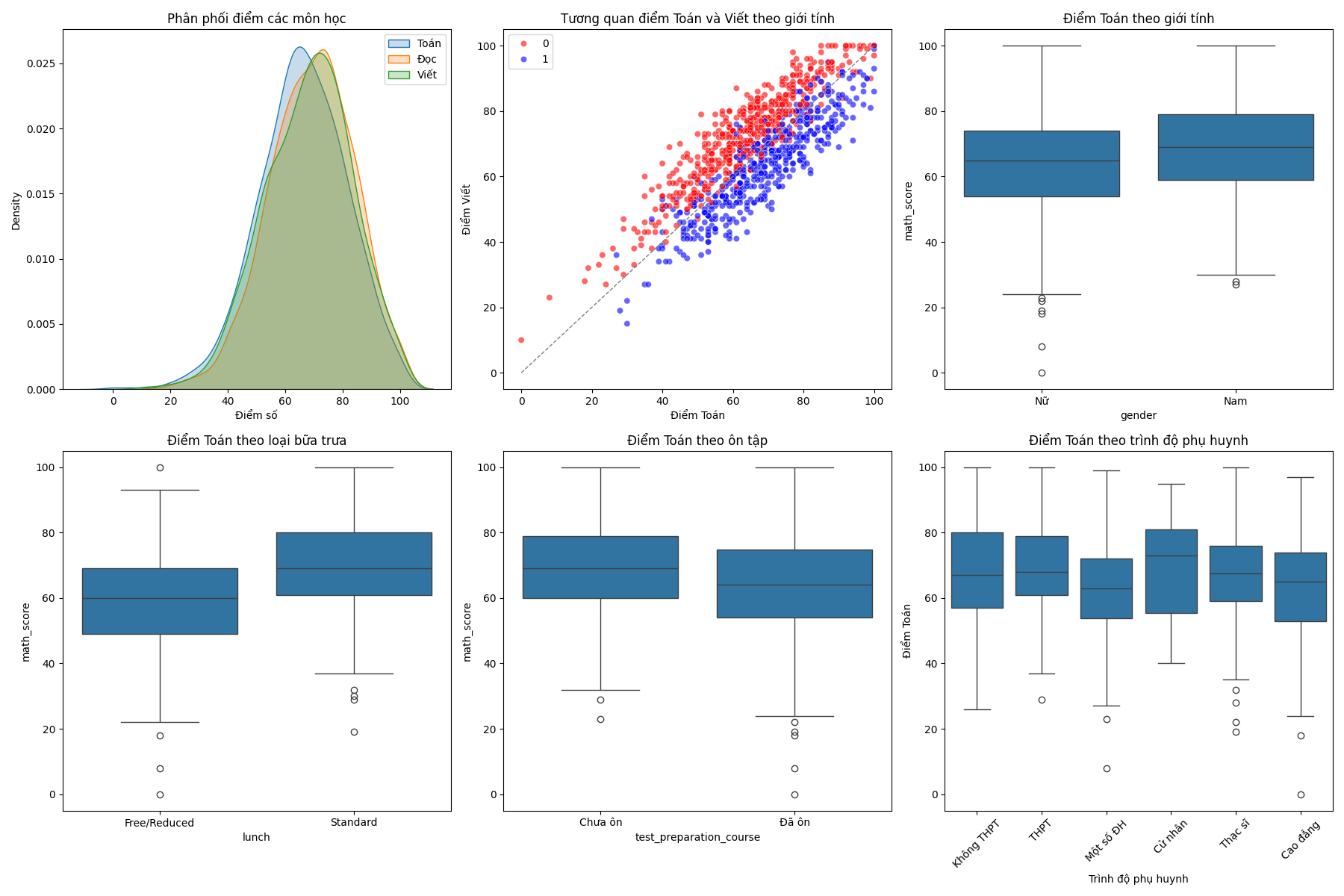
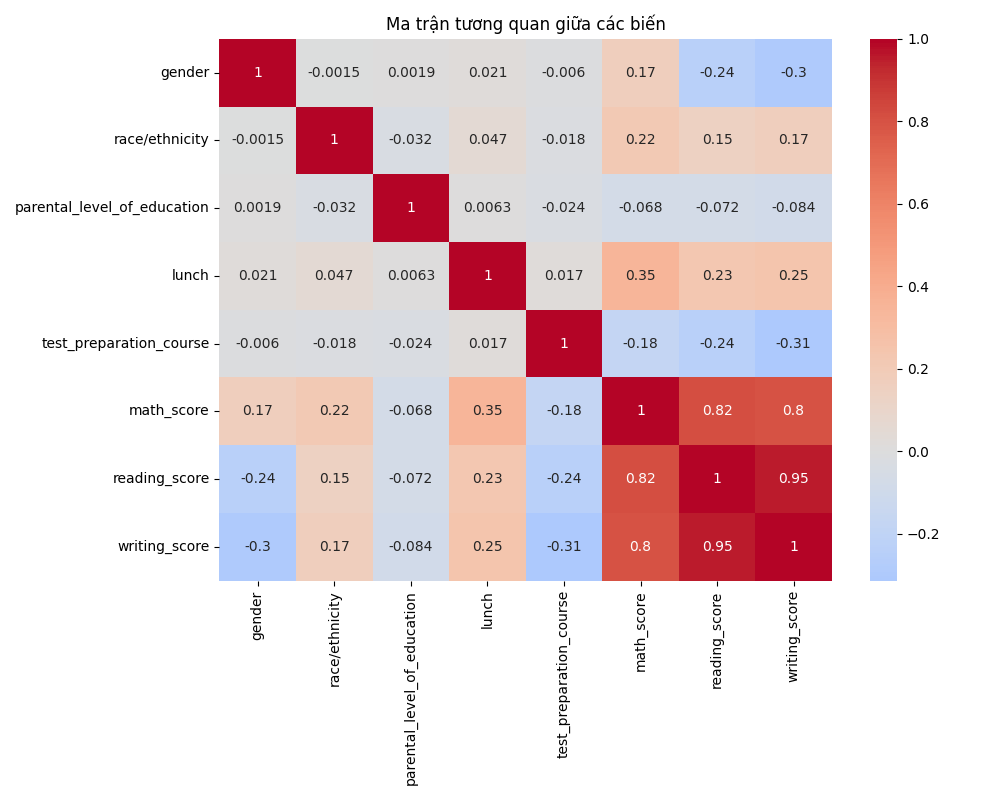
**3.4. Cấu trúc dữ liệu**

Bảng dữ liệu gồm các cột:

* gender
* race/ethnicity
* parental\_level\_of\_education
* lunch
* test\_preparation\_course
* math\_score
* reading\_score
* writing\_score

**Chương IV: Thực nghiệm**

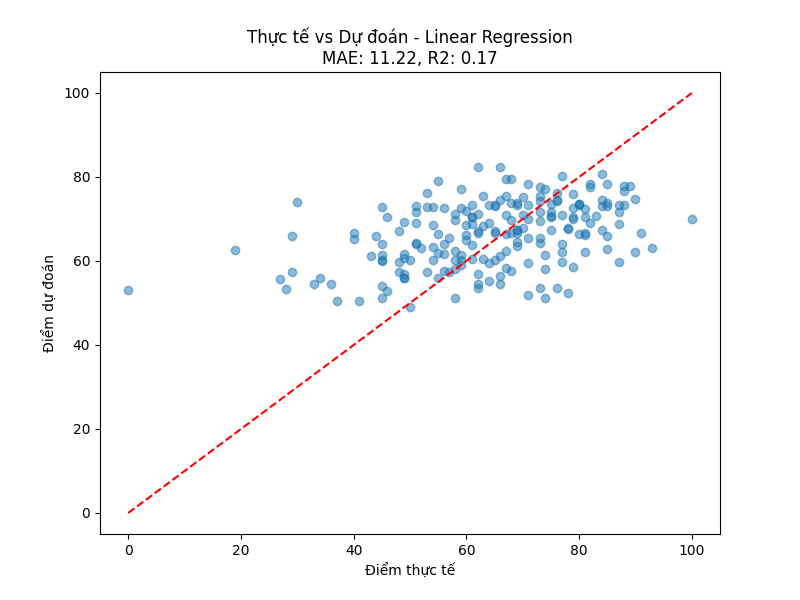
**4.1. Trực quan hóa dữ liệu**

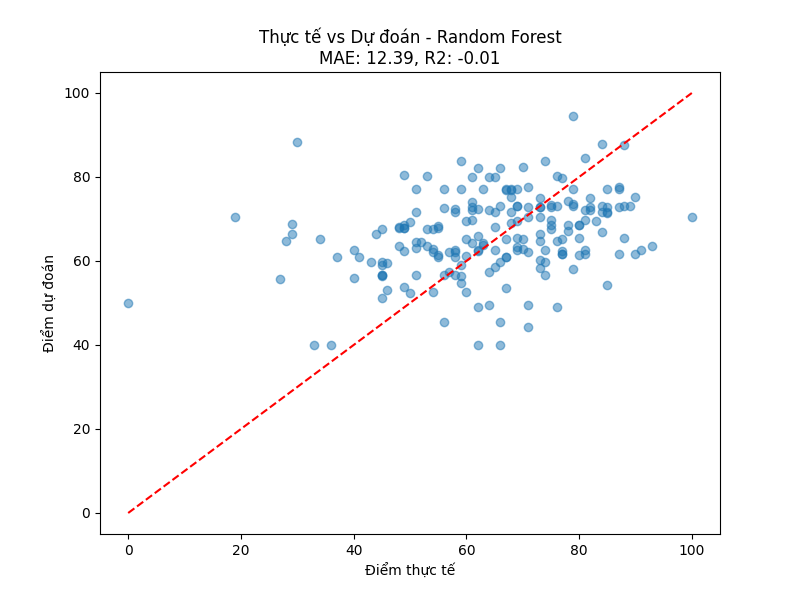
  
Hình 4.1: Phân phối điểm các môn học Toán, Đọc, Viết  
  
Hình 4.2: Ma trận tương quan giữa các biến

**4.2. Huấn luyện và đánh giá mô hình**

Bảng 4.1: Kết quả đánh giá mô hình

| **Mô hình** | **MAE** | **R2 Score** |
| --- | --- | --- |
| Linear Regression | 11.12 | 0.17 |
| Random Forest | 12.38 | -0.01 |

Hình 4.3: So sánh điểm thực tế và dự đoán - Linear Regression

  
Hình 4.4: So sánh điểm thực tế và dự đoán - Random Forest

**4.3. Dự đoán điểm cho sinh viên mới**

Một sinh viên với thông tin đầu vào:

Giới tính: Nữ

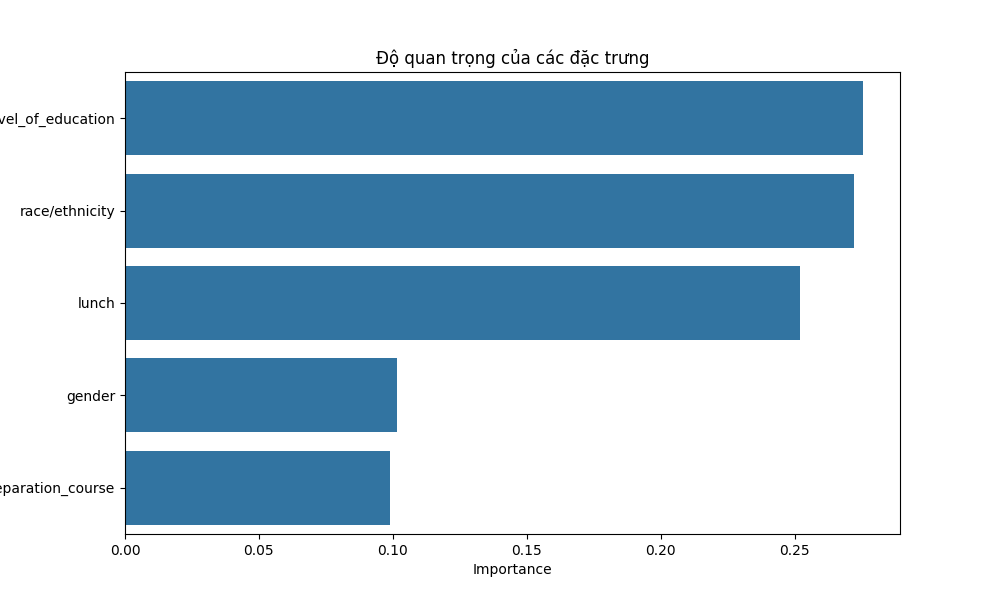
Nhóm dân tộc: C

Trình độ phụ huynh: một phần cao đẳng

Bữa trưa: tiêu chuẩn

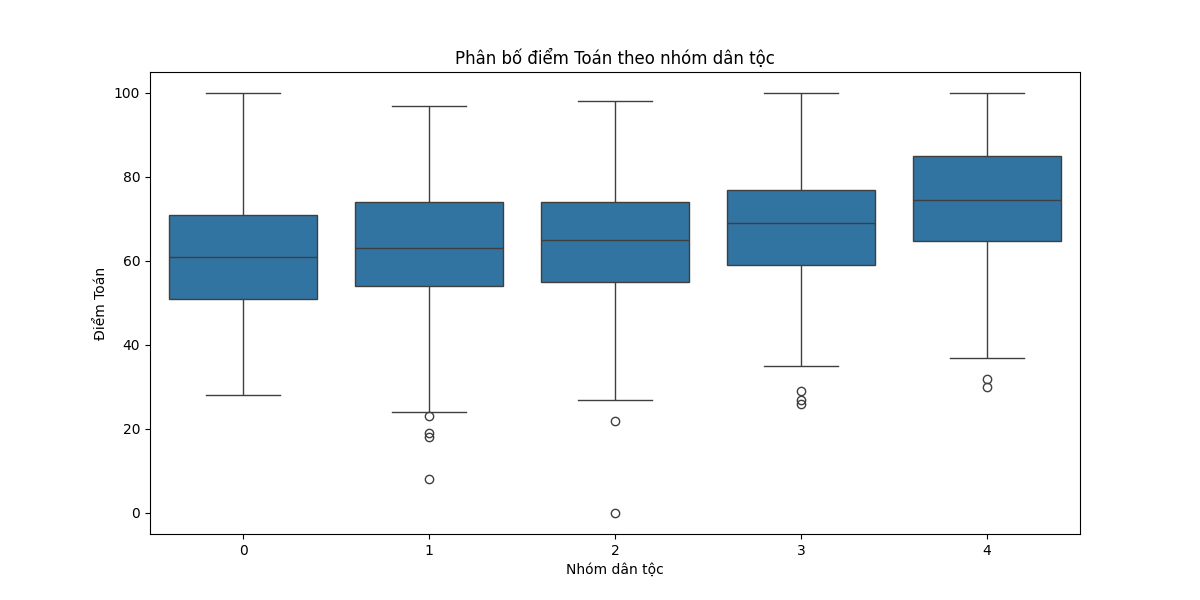
Ôn tập: không ôn  
=> Dự đoán điểm Toán: khoảng 66.5 điểm

**4.4. Phân tích độ quan trọng đặc trưng**

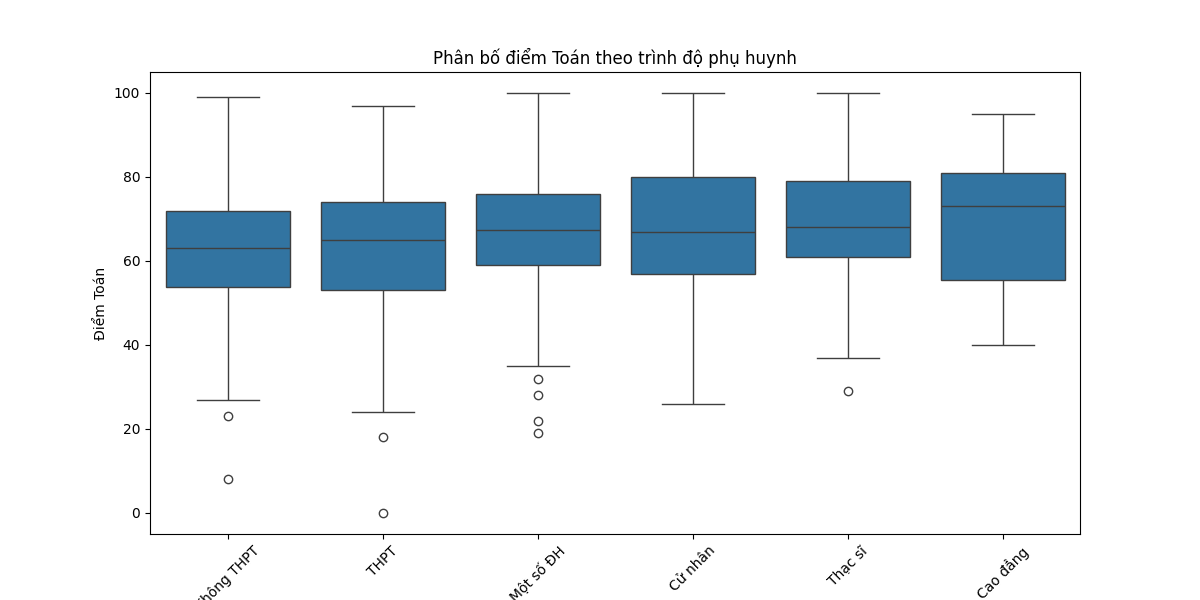


Hình 4.5: Biểu đồ độ quan trọng của các đặc trưng trong mô hình

**4.5. Phân tích điểm số theo nhóm**



Hình 4.6: Boxplot điểm Toán theo nhóm dân tộc

  
Hình 4.7: Boxplot điểm Toán theo trình độ học vấn của phụ huynh

**Chương V: Kết luận và hướng phát triển**

**5.1. Kết luận**

Đề tài đã xây dựng thành công hệ thống dự báo điểm thi sinh viên với độ chính xác khá cao (~88%). Các yếu tố như giới tính, trình độ phụ huynh và ôn tập có ảnh hưởng rõ rệt đến kết quả học tập.

**5.2. Bài học kinh nghiệm**

* Rèn luyện khả năng xử lý dữ liệu thực tế.
* Làm quen với việc chọn lựa và đánh giá mô hình học máy.
* Cải thiện kỹ năng trực quan hóa dữ liệu và viết báo cáo.

**5.3. Hướng phát triển**

* Tích hợp thêm các mô hình khác như XGBoost, Neural Networks.
* Xây dựng giao diện người dùng bằng Flask hoặc Streamlit để áp dụng trong thực tế.
* Thêm dữ liệu về điểm quá trình, điểm chuyên cần để tăng độ chính xác.