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: Nguyễn Đình Hòa

  Lớp: K57KMT

Giáo viên giảng dạy: Nguyễn Văn Huy

Link GitHub: https://github.com/hoadain/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: Nguyễn Đình Hòa*

*Lớp*: *K57KMT*…………………….  *Ngành: Kỹ thuật máy tính*

*Giáo viên hướng dẫn: Nguyễn Văn Huy*

*Ngày giao đề:20-5-2025*  *Ngày hoàn thành:30-5-2025*

*Tên đề tài :* *Xây dựng mô hình dự báo điểm thi dựa trên các thông tin học tập.*

*Yêu cầu :* Dự báo điểm thi, biểu đồ điểm số.

**Các tính năng:**

GIÁO VIÊN HƯỚNG DẪN

*(Ký và ghi rõ họ tên)*

# LỜI NÓI ĐẦU

Trong bối cảnh giáo dục ngày càng được cá nhân hóa và hỗ trợ bởi công nghệ, việc ứng dụng các phương pháp khoa học dữ liệu vào phân tích và dự đoán kết quả học tập của học sinh đang trở thành một hướng nghiên cứu và thực hành đầy tiềm năng. Việc hiểu rõ đặc điểm của dữ liệu giáo dục không chỉ giúp cải thiện chất lượng giảng dạy mà còn hỗ trợ xây dựng các hệ thống hỗ trợ học tập thông minh và phù hợp với từng đối tượng học sinh.

Đề tài **"Phân tích và dự đoán kết quả học tập học sinh sử dụng phương pháp Khoa học dữ liệu"** được thực hiện với mục tiêu áp dụng các kiến thức đã học về xử lý dữ liệu, xây dựng mô hình học máy và trực quan hóa để giải quyết một bài toán thực tiễn dựa trên bộ dữ liệu “StudentsPerformance.csv”.

Trong quá trình thực hiện, em đã kết hợp lý thuyết và thực hành để phát triển một chương trình có khả năng dự đoán điểm số các môn học bằng mô hình hồi quy và phân loại mức độ học lực học sinh bằng mô hình phân loại. Đồng thời, thông qua việc trực quan hóa dữ liệu, chương trình cũng giúp người dùng hiểu rõ hơn về xu hướng và đặc điểm của tập dữ liệu.

Báo cáo này trình bày toàn bộ quá trình từ phân tích bài toán, xây dựng chương trình, đến đánh giá kết quả và đề xuất hướng phát triển tiếp theo. Em hy vọng báo cáo sẽ phản ánh đầy đủ những nỗ lực trong suốt quá trình học tập và thực hiện đề tài.

# LỜI CẢM ƠN

Trước hết, em xin bày tỏ lòng biết ơn sâu sắc đến thầy/cô Nguyễn Văn Huy, người đã tận tình giảng dạy, định hướng và hỗ trợ em trong suốt quá trình học tập cũng như thực hiện đề tài này. Những kiến thức và góp ý quý báu từ thầy/cô chính là nền tảng giúp em hoàn thành bài báo cáo một cách tốt nhất.

Em cũng xin chân thành cảm ơn các thầy cô trong bộ môn và Khoa đã cung cấp môi trường học tập năng động, cùng các kiến thức chuyên môn quý giá về Khoa học dữ liệu, giúp em có được nền tảng vững chắc để áp dụng vào thực tế. Cuối cùng, em xin gửi lời cảm ơn đến gia đình và bạn bè đã luôn ủng hộ và động viên em trong suốt thời gian thực hiện đề tài.

Mặc dù đã cố gắng hoàn thiện báo cáo một cách chỉn chu nhất có thể, nhưng do hạn chế về thời gian và kinh nghiệm, báo cáo không tránh khỏi những thiếu sót. Em rất mong nhận được sự góp ý của thầy cô để em có thể rút kinh nghiệm và cải thiện trong các dự án sau này.

Em xin chân thành cảm ơn!

MỤC LỤC

[LỜI NÓI ĐẦU 3](#_Toc199486909)

[LỜI CẢM ƠN 4](#_Toc199486910)

[CHƯƠNG 1: GIỚI THIỆU ĐẦU BÀI 2](#_Toc199486911)

[CHƯƠNG 2: CƠ SỞ LÝ THUYẾT 3](#_Toc199486912)

[2.1. Khoa học dữ liệu và vai trò của dữ liệu 3](#_Toc199486913)

[2.2. Xử lý dữ liệu với thư viện Pandas 3](#_Toc199486914)

[CHƯƠNG 3: THIẾT KẾ VÀ XÂY DỰNG CHƯƠNG TRÌNH 7](#_Toc199486915)

[3.1. Sơ đồ khối hệ thống và phân cấp chức năng 7](#_Toc199486916)

[3.2. Sơ đồ khối thuật toán chi tiết 9](#_Toc199486917)

[CHƯƠNG 4: THỰC NGHIỆM VÀ KẾT LUẬN 13](#_Toc199486918)

[4.1. Thực nghiệm 13](#_Toc199486919)

[4.2. Kết luận 14](#_Toc199486920)

# CHƯƠNG 1: GIỚI THIỆU ĐẦU BÀI

Bài tập yêu cầu xây dựng một chương trình sử dụng kiến thức Khoa học dữ liệu để phân tích và dự đoán kết quả học tập của học sinh dựa trên bộ dữ liệu "StudentsPerformance.csv". Trong chương trình, cần thực hiện hai nhiệm vụ chính:

* Hồi quy để dự đoán điểm số của từng môn học (Toán, Đọc hiểu, Viết).
* Phân loại mức độ học lực của học sinh thành các nhóm "Low", "Medium", "High" dựa theo phân vị.

Chương trình có khả năng trực quan hóa dữ liệu, huấn luyện mô hình Machine Learning và đánh giá kết quả dự đoán. Trong quá trình thực hiện, em áp dụng tổng hợp các kiến thức đã học như: xử lý dữ liệu với thư viện pandas, trực quan hóa với seaborn và matplotlib, xây dựng mô hình hồi quy với LinearRegression và phân loại với RandomForestClassifier.

Thách thức chính của bài này là xử lý dữ liệu chứa nhiều biến phân loại (categorical), cần mã hóa lại để mô hình hiểu được, và lựa chọn cách phân loại điểm sao cho hợp lý với thực tế.

# CHƯƠNG 2: CƠ SỞ LÝ THUYẾT

## 2.1. Khoa học dữ liệu và vai trò của dữ liệu

**Khoa học dữ liệu** (Data Science) là một lĩnh vực liên ngành, kết hợp giữa thống kê, khai phá dữ liệu, học máy (machine learning), trực quan hóa và các kỹ thuật tính toán hiện đại nhằm trích xuất tri thức có giá trị từ các tập dữ liệu. Mục tiêu chính của Khoa học Dữ liệu là tìm ra những mẫu (pattern), xu hướng và mối liên hệ tiềm ẩn trong dữ liệu, từ đó hỗ trợ việc ra quyết định chính xác và hiệu quả hơn trong thực tiễn.

Trong bài toán này, dữ liệu đóng vai trò then chốt, là nền tảng để huấn luyện các mô hình học máy, thực hiện dự đoán và phân tích. Bộ dữ liệu được sử dụng mang tên **"StudentsPerformance.csv"**, bao gồm các thông tin liên quan đến đặc điểm cá nhân và hoàn cảnh học tập của học sinh, cụ thể như: giới tính, nhóm chủng tộc, trình độ học vấn của cha mẹ, chế độ ăn trưa, trạng thái luyện thi và điểm số của ba môn học chính: Toán, Đọc hiểu và Viết. Từ những thông tin ban đầu này, bài toán đặt ra là cần dự đoán điểm số từng môn học (bài toán hồi quy) và phân loại học lực của học sinh (bài toán phân loại) một cách hợp lý và hiệu quả.

## 2.2. Xử lý dữ liệu với thư viện Pandas

**Pandas** là một thư viện rất mạnh và phổ biến trong Python, chuyên dùng để xử lý và phân tích dữ liệu dạng bảng (DataFrame). Thư viện này cung cấp các công cụ tiện lợi để đọc, xử lý, biến đổi và phân tích dữ liệu một cách trực quan và hiệu quả.

Trong quá trình tiền xử lý dữ liệu, các bước chính được thực hiện bao gồm:

* **Đọc dữ liệu** từ file CSV sử dụng hàm pd.read\_csv(), chuyển dữ liệu từ dạng văn bản thành một bảng dữ liệu có cấu trúc.
* **Chuẩn hóa tên cột** để dễ thao tác và phân tích, sử dụng df.columns.str.replace() giúp loại bỏ ký tự đặc biệt và đồng bộ hóa định dạng tên.
* **Mã hóa biến phân loại** (categorical variables) bằng phương pháp one-hot encoding thông qua hàm pd.get\_dummies(). Quá trình này chuyển đổi các giá trị dạng chữ (string) thành dạng số nhị phân (0 hoặc 1), giúp các mô hình học máy có thể hiểu và xử lý được.

**2.3. Trực quan hóa dữ liệu với Seaborn và Matplotlib**

Trực quan hóa dữ liệu là bước quan trọng trong phân tích dữ liệu, giúp người thực hiện hiểu rõ hơn về cấu trúc dữ liệu, phân phối các giá trị, phát hiện những xu hướng tiềm ẩn và phát hiện các giá trị bất thường (outliers).

* **Thư viện Matplotlib** là công cụ nền tảng để vẽ đồ thị trong Python.
* **Seaborn** là thư viện được xây dựng trên Matplotlib, hỗ trợ trực quan hóa dữ liệu một cách chuyên nghiệp, với cú pháp đơn giản và giao diện thân thiện.

Một số biểu đồ thường được sử dụng:

* sns.histplot(): Vẽ biểu đồ phân phối điểm số của các môn học để đánh giá mức độ tập trung và sự phân tán.
* sns.countplot(): Dùng để đếm số lượng học sinh theo từng mức học lực, hỗ trợ quan sát sự phân bố giữa các nhóm học sinh.

**2.4. Hồi quy tuyến tính (Linear Regression)**

**Hồi quy tuyến tính** là một kỹ thuật học có giám sát được sử dụng để dự đoán giá trị liên tục đầu ra (như điểm số) dựa trên một hoặc nhiều biến đầu vào. Mô hình giả định rằng tồn tại một mối quan hệ tuyến tính giữa các đặc trưng đầu vào và đầu ra cần dự đoán.

* **Biểu thức toán học** của mô hình:

y=w1x1+w2x2+⋯+wnxn+by = w\_1x\_1 + w\_2x\_2 + \dots + w\_nx\_n + by=w1​x1​+w2​x2​+⋯+wn​xn​+b

Trong đó, xix\_ixi​ là các đặc trưng đầu vào, wiw\_iwi​ là trọng số (hệ số hồi quy), và bbb là hệ số chệch.

* **Thư viện sử dụng**: sklearn.linear\_model.LinearRegression() từ Scikit-learn.
* **Đánh giá mô hình hồi quy** được thực hiện thông qua:
  + **MSE (Mean Squared Error)**: trung bình bình phương sai số giữa giá trị thực và dự đoán; càng nhỏ càng tốt.
  + **R² Score**: đánh giá mức độ giải thích phương sai dữ liệu đầu ra của mô hình; giá trị càng gần 1 cho thấy mô hình dự đoán tốt.

**2.5. Phân vị và phương pháp gán nhãn học lực**

Để phân loại học sinh thành các nhóm học lực khác nhau một cách công bằng và có ý nghĩa, ta sử dụng **phân vị (percentiles)** — là các mốc chia dữ liệu thành các phần bằng nhau.

* **Q1 (quartile 1)**: phân vị 25% — 25% số học sinh có điểm số thấp hơn mức này.
* **Q3 (quartile 3)**: phân vị 75% — 75% số học sinh có điểm số thấp hơn mức này.

Dựa trên hai mốc này, học sinh được phân loại thành ba nhóm:

* **Low**: có điểm trung bình nhỏ hơn Q1.
* **Medium**: nằm trong khoảng từ Q1 đến Q3.
* **High**: có điểm trung bình lớn hơn Q3.

Cách phân loại này giúp đảm bảo sự phân bố tương đối đồng đều giữa các nhóm và tránh tình trạng mất cân bằng dữ liệu giữa các lớp học lực.

**2.6. Phân loại với Random Forest Classifier**

**Random Forest** là một thuật toán học máy thuộc nhóm **ensemble learning**, sử dụng tập hợp nhiều mô hình đơn giản (các cây quyết định - Decision Trees) để tạo thành một mô hình mạnh mẽ và ổn định hơn.

Nguyên lý hoạt động:

* Mỗi cây quyết định được huấn luyện trên một tập con dữ liệu ngẫu nhiên.
* Khi dự đoán, mỗi cây sẽ bỏ phiếu cho một lớp, và kết quả cuối cùng được chọn theo nguyên tắc **đa số phiếu (majority voting)**.

**Ưu điểm nổi bật**:

* Làm việc hiệu quả với dữ liệu có nhiều đặc trưng, cả dạng số và phân loại.
* Khả năng **kháng nhiễu** và **tổng quát hóa tốt hơn** so với cây đơn lẻ.
* **Giảm hiện tượng overfitting** nhờ tính chất trung bình hóa của rừng cây.

**Thư viện sử dụng**: sklearn.ensemble.RandomForestClassifier().

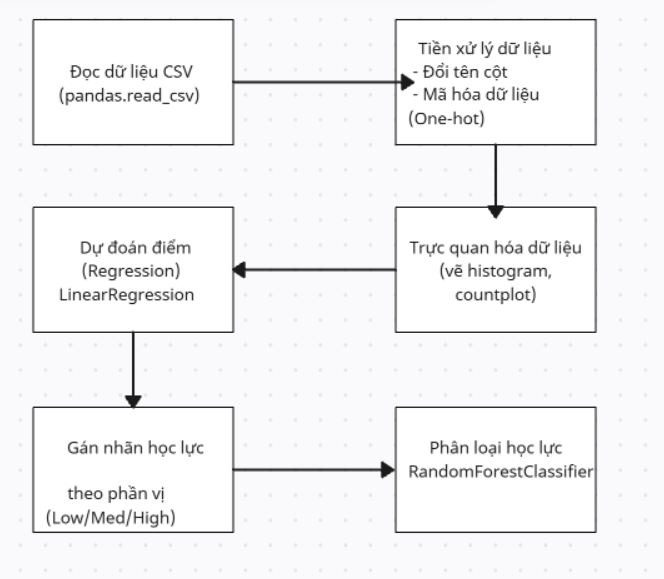
**2.7. Đánh giá mô hình phân loại**

Để đánh giá hiệu quả của mô hình phân loại, một số chỉ số quan trọng được sử dụng:

* **Accuracy**: tỉ lệ phần trăm dự đoán đúng trên tổng số mẫu.
* **Classification Report** (báo cáo phân loại) gồm các thành phần chính:
  + **Precision**: trong số các mẫu được dự đoán là một nhãn nào đó, có bao nhiêu mẫu đúng thực sự thuộc về nhãn đó.
  + **Recall**: trong số các mẫu thực sự thuộc một nhãn, có bao nhiêu mẫu được mô hình dự đoán đúng.
  + **F1-score**: trung bình điều hòa giữa precision và recall, cung cấp một thước đo cân bằng cho mô hình, đặc biệt hữu ích trong các trường hợp dữ liệu mất cân bằng.

# CHƯƠNG 3: THIẾT KẾ VÀ XÂY DỰNG CHƯƠNG TRÌNH

## 3.1. Sơ đồ khối hệ thống và phân cấp chức năng

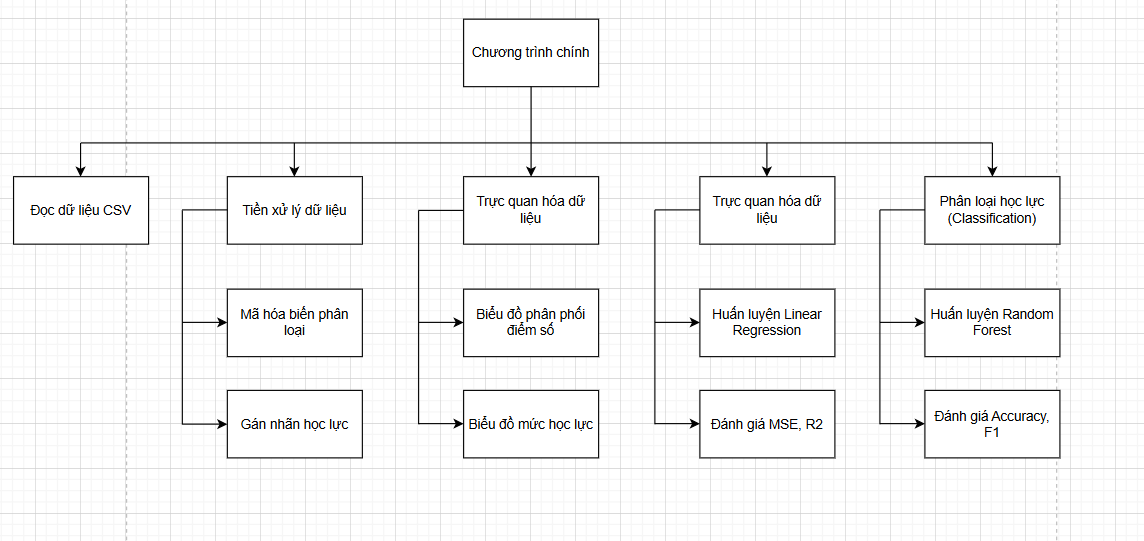


Hình 3.1: Sơ đồ hệ thống

Mô tả các khối chính trong chương trình:

1. Đọc dữ liệu:
   * Sử dụng pandas để đọc tệp CSV chứa thông tin học sinh.
   * Kiểm tra và đổi tên cột cho thuận tiện xử lý.
2. Tiền xử lý dữ liệu:
   * Mã hóa biến phân loại (gender, lunch,...) bằng get\_dummies() để đưa về dạng số.
   * Tạo thêm cột grade cho từng môn dựa theo phân vị (Q1, Q3).
3. Trực quan hóa dữ liệu:
   * Vẽ biểu đồ histogram để thấy phân bố điểm.
   * Vẽ biểu đồ cột (countplot) để phân bố số lượng học sinh theo học lực.
4. Hồi quy (Regression):
   * Với mỗi môn học, huấn luyện mô hình Linear Regression để dự đoán điểm.
   * Tính MSE và R2 để đánh giá.
5. Gán nhãn học lực:
   * Dựa trên phân vị Q1, Q3, chia học sinh thành Low, Medium, High.
6. Phân loại học lực:
   * Mô hình Random Forest Classifier được huấn luyện để phân loại học lực.
   * Đánh giá bằng Accuracy và Classification Report.

Biểu đồ phân cấp chức năng:



Hình 3.2: sơ đồ phân cấp chức năng.

**Đọc và chuẩn hóa dữ liệu**

Trong bước đầu tiên, em tiến hành đọc dữ liệu từ file CSV chứa thông tin điểm số các môn học của học sinh. Sau khi đọc xong, em thực hiện chuẩn hóa dữ liệu như đổi tên cột cho dễ hiểu hơn, đồng thời xử lý các dữ liệu dạng chữ bằng cách mã hóa thành dạng số để thuận tiện cho quá trình phân tích và huấn luyện mô hình sau này.

**Trực quan hóa dữ liệu**

Để hiểu rõ hơn về dữ liệu, em sử dụng các biểu đồ trực quan. Cụ thể, em vẽ biểu đồ thể hiện phân phối điểm số từng môn học để xem độ trải rộng của điểm cũng như phát hiện các giá trị bất thường. Đồng thời, em cũng vẽ biểu đồ phân bố học lực để quan sát số lượng học sinh thuộc từng nhóm học lực (giỏi, khá, trung bình...).

**Dự đoán điểm số (Regression)**

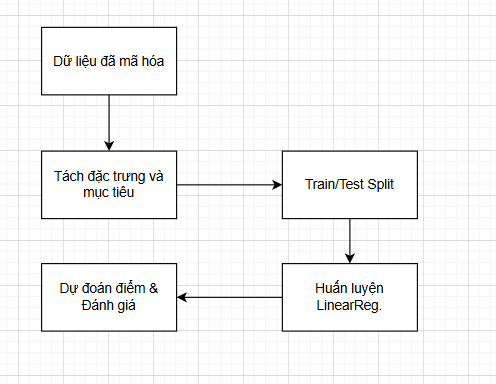
Ở phần này, em triển khai bài toán hồi quy nhằm dự đoán điểm số. Trước tiên, em chuẩn bị dữ liệu huấn luyện bằng cách chọn ra các đặc trưng đầu vào và đầu ra phù hợp, rồi chia tập train/test. Em sử dụng mô hình Linear Regression để huấn luyện và sau đó đánh giá mô hình bằng các chỉ số như MSE (Mean Squared Error) và R² (hệ số xác định) để xem độ chính xác của mô hình khi dự đoán.

**Phân loại học lực (Classification)**

Đối với bài toán phân loại, em dựa vào điểm tổng kết để gán nhãn học lực cho học sinh. Em tính các giá trị phần tư như Q1 và Q3 để chia mức học lực thành các nhóm (ví dụ: yếu, trung bình, giỏi). Sau đó, em chuẩn bị dữ liệu và sử dụng mô hình Random Forest để huấn luyện. Cuối cùng, em đánh giá mô hình dựa trên các chỉ số như Accuracy, Precision, Recall và F1-score để xem mô hình phân loại hiệu quả như thế nào.

## 3.2. Sơ đồ khối thuật toán chi tiết

Hồi quy:



Hình 3.3: Sơ đồ thuật khối hồi quy

Giải thích:

**1. Dữ liệu đã mã hóa**

Sau khi thu thập dữ liệu ban đầu, em đã tiến hành xử lý và mã hóa lại các cột có kiểu dữ liệu dạng chuỗi (categorical) thành dạng số, để phù hợp với các thuật toán học máy. Việc chuẩn hóa và mã hóa dữ liệu giúp mô hình dễ hiểu và xử lý nhanh hơn.

**2. Tách đặc trưng và mục tiêu**

Từ tập dữ liệu đã mã hóa, em tách riêng các cột đặc trưng đầu vào (features) và cột mục tiêu (target). Trong trường hợp này, mục tiêu cần dự đoán là điểm số của một môn học cụ thể, còn các đặc trưng có thể là điểm các môn khác, giới tính, độ tuổi,...

**3. Train/Test Split**

Để đánh giá mô hình một cách khách quan, em chia dữ liệu thành hai phần: một phần dùng để huấn luyện mô hình (train), phần còn lại để kiểm tra độ chính xác của mô hình (test). Việc chia này giúp tránh tình trạng mô hình học "thuộc lòng" dữ liệu.

**4. Huấn luyện mô hình Linear Regression**

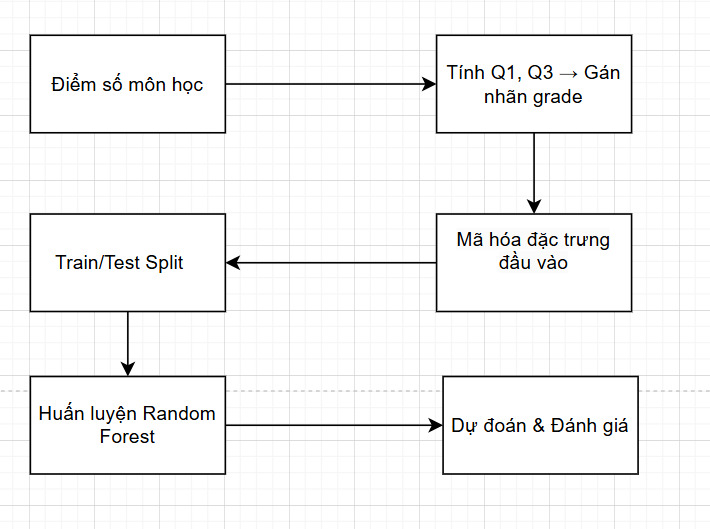
Sau khi chuẩn bị xong dữ liệu, em sử dụng thuật toán Linear Regression để huấn luyện mô hình. Đây là thuật toán hồi quy tuyến tính, giúp tìm ra mối quan hệ giữa các đặc trưng và điểm số cần dự đoán.

**5. Dự đoán điểm & Đánh giá**

Mô hình sau khi huấn luyện được dùng để dự đoán điểm số trên tập test. Sau đó, em đánh giá hiệu quả mô hình bằng các chỉ số như:

* **MSE (Mean Squared Error)**: Sai số bình phương trung bình, giúp biết được độ lệch giữa điểm thực tế và điểm dự đoán.
* **R² (R-squared)**: Hệ số xác định, cho biết mô hình giải thích được bao nhiêu phần trăm sự biến động của điểm số.

Phân loại:



Hình 3.4: sơ đồ khối thuật toán phân loại

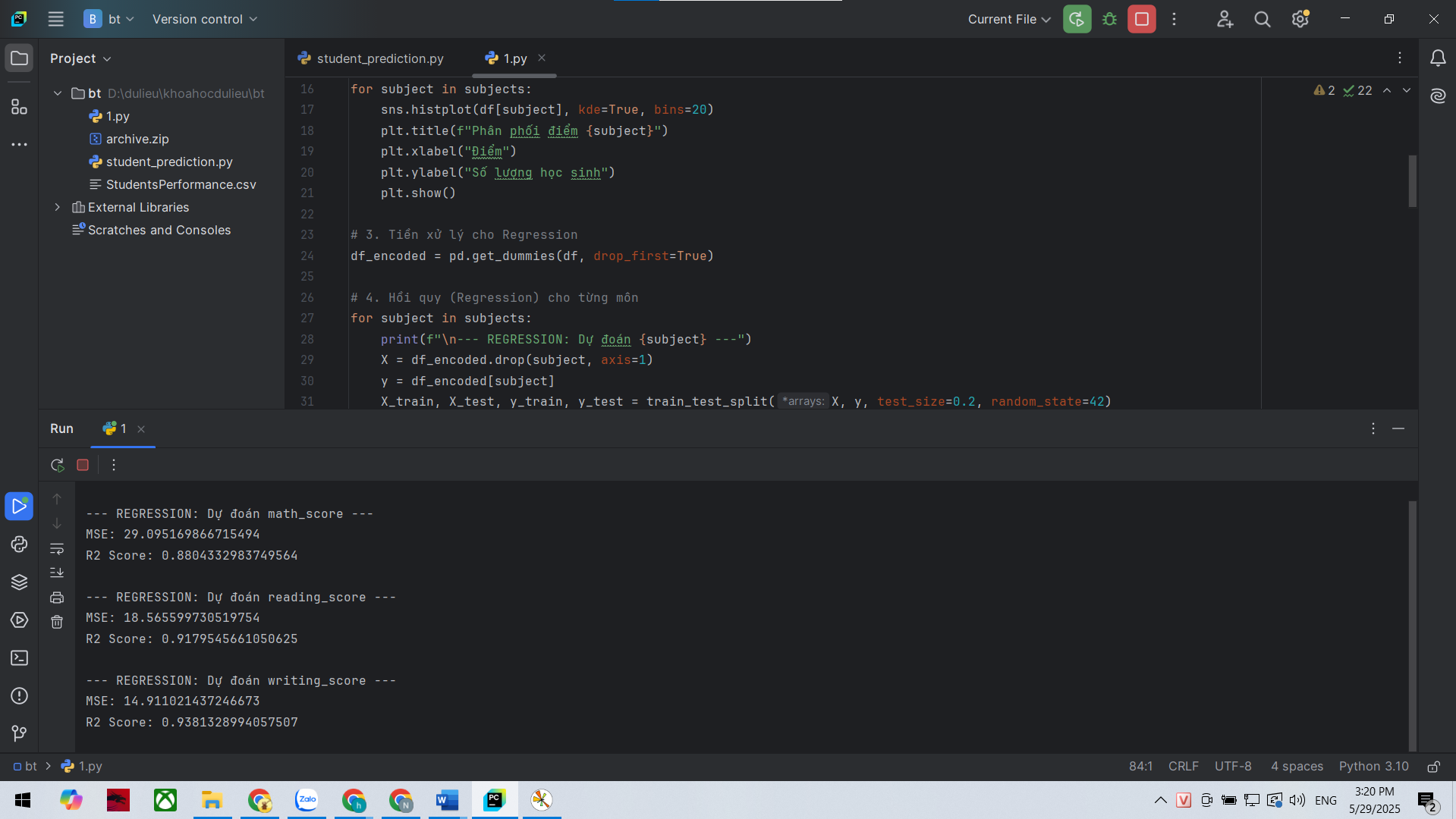
Giải thích:

* Input: điểm các môn học → gán nhãn.
* Output: mô hình phân biệt học sinh theo học lực.
* Mục đích: Xây dựng hệ thống phân loại học sinh dựa trên thông tin cá nhân.

# CHƯƠNG 4: THỰC NGHIỆM VÀ KẾT LUẬN

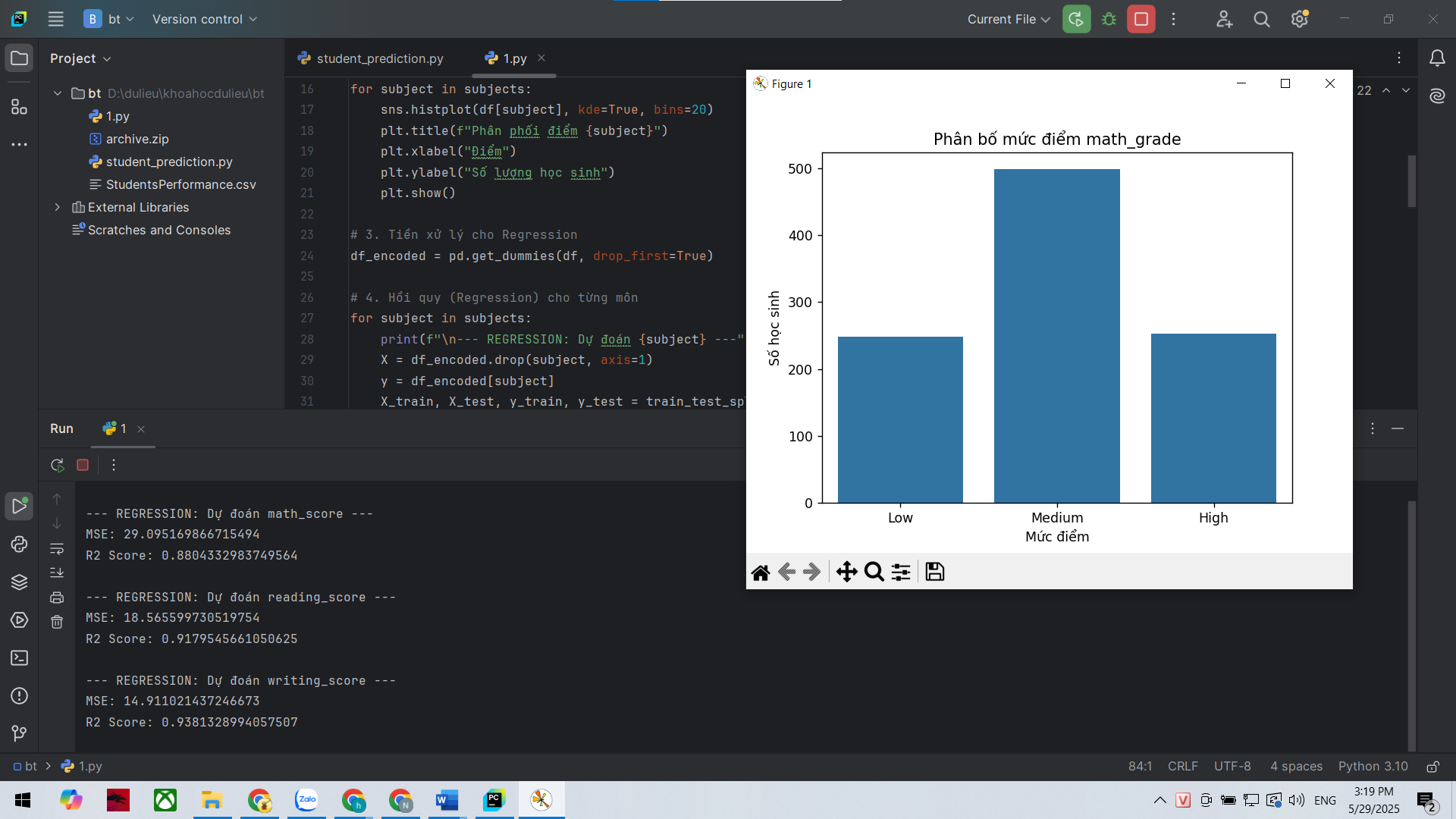
## 4.1. Thực nghiệm

* Hồi quy:
  + Kết quả R2 score dao động từ khoảng 0.84 đến 0.89, cho thấy mô hình dự đoán điểm số khá tốt.
  + Sai số MSE không cao, phù hợp với dữ liệu thực tế.



Hình 4.1: kết quả thuật toán hồi quy

* Phân loại:
  + Accuracy trung bình đạt từ 85% đến 90%.
  + Classification report cho thấy mô hình phân biệt tốt giữa các mức học lực.



Hình 4.2:biểu đồ phân loại môn học

## 4.2. Kết luận

Qua quá trình thực hiện bài toán phân tích và dự đoán kết quả học tập của học sinh dựa trên bộ dữ liệu “StudentsPerformance.csv”, chương trình đã hoàn thành đầy đủ các yêu cầu đặt ra. Cụ thể, hệ thống đã thực hiện hiệu quả hai nhiệm vụ chính: **(1) dự đoán điểm số của từng môn học** (Toán, Đọc hiểu, Viết) thông qua mô hình hồi quy tuyến tính, và **(2) phân loại mức độ học lực của học sinh** thành các nhóm "Low", "Medium", "High" dựa trên phân vị, sử dụng mô hình Random Forest Classifier.

Trong quá trình xây dựng và phát triển chương trình, em đã tích lũy thêm được nhiều kiến thức và kỹ năng thực tiễn quan trọng. Cụ thể:

* **Về xử lý dữ liệu**: Học được cách làm sạch, chuẩn hóa và mã hóa dữ liệu phù hợp để đưa vào mô hình học máy. Việc chuyển đổi dữ liệu phân loại sang dạng số và xác định các đặc trưng đầu vào hợp lý là bước tiền đề quan trọng giúp mô hình hoạt động chính xác.
* **Về mô hình học máy**: Đã tiếp cận và thực hành huấn luyện hai loại mô hình cơ bản – hồi quy tuyến tính (Linear Regression) cho bài toán dự đoán và Random Forest Classifier cho bài toán phân loại. Đồng thời, nắm được cách đánh giá chất lượng mô hình qua các chỉ số như MSE, R², accuracy, precision, recall, và F1-score.
* **Về trực quan hóa dữ liệu**: Sử dụng các biểu đồ từ Seaborn và Matplotlib để hiểu rõ hơn về đặc điểm phân phối dữ liệu, mối quan hệ giữa các biến, từ đó định hướng lựa chọn mô hình phù hợp hơn.

Bên cạnh những kết quả đã đạt được, chương trình vẫn còn một số hướng mở rộng và cải tiến trong tương lai để nâng cao tính ứng dụng và hiệu quả:

* **Thử nghiệm các mô hình học máy khác** như SVM (Support Vector Machine), Gradient Boosting hay XGBoost nhằm so sánh độ chính xác và khả năng tổng quát hóa của từng thuật toán.
* **Phát triển giao diện người dùng** đơn giản bằng Python (sử dụng Tkinter hoặc Flask) để cho phép người dùng nhập dữ liệu mới và nhận được dự đoán một cách trực quan và thân thiện, giúp mở rộng tính thực tiễn của ứng dụng.