ỦY BAN NHÂN DÂN TP. HỒ CHÍ MINH

TRƯỜNG ĐẠI HỌC SÀI GÒN

-

**SỬ DỤNG KHAI PHÁ DỮ LIỆU ĐỂ DỰ ĐOÁN HỌC LỰC CỦA HỌC SINH TRUNG HỌC**

BÁO CÁO MÔN KHAI PHÁ DỮ LIỆU

NGUYỄN THÀNH LONG

3121410297

DCT1212

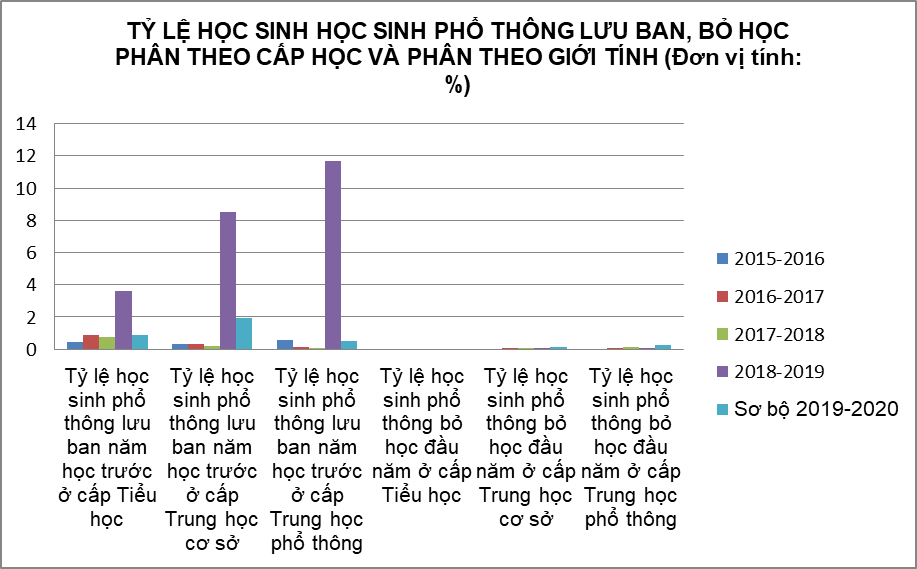
TP. HỒ CHÍ MINH, THÁNG 5 NĂM 2022

**Chương 1: Giới thiệu**

##### 1.1. Bối cảnh của bài toán

Trẻ em chính là chủ nhân tương lai của đất nước. Mai sau dân có giàu, nước có mạnh hay không đều phải dựa vào từng đứa trẻ một. Giáo dục đóng vai trò then chốt trong sự phát triển của mỗi quốc gia, là nền tảng cho nguồn nhân lực tương lai. Đảm bảo chất lượng giáo dục và nâng cao kết quả học tập của học sinh luôn là mục tiêu hàng đầu. Tuy nhiên, thực tế cho thấy vẫn còn tồn tại một bộ phận học sinh có kết quả học tập chưa cao, tiềm ẩn nguy cơ bỏ học hoặc không phát triển hết năng lực bản thân.

Theo thống kê của Bộ Giáo dục và Đào tạo, tỷ lệ học sinh tốt nghiệp trung học cơ sở đạt trên 90%. Tuy nhiên, vấn đề học sinh yếu kém, chán học và bỏ học vẫn còn hiện hữu, đặc biệt là ở khu vực nông thôn và miền núi.



Hình 1.1 Tỷ lệ học sinh phổ thông lưu ban, bỏ học phân theo cấp học

và phân theo giới tính

Trước thực trạng đó, việc chủ động tìm kiếm giải pháp để nâng cao chất lượng giáo dục và hỗ trợ kịp thời cho học sinh gặp khó khăn trong học tập là vô cùng cần thiết. Khai phá dữ liệu (Data Mining) với khả năng phân tích dữ liệu lớn và dự đoán xu hướng nổi lên như một công cụ hữu ích, hứa hẹn mang đến giải pháp hiệu quả cho vấn đề này.



Hình 1.1 Cô giáo và các em học sinh đang vui đùa dưới sân

##### 1.2. Định nghĩa bài toán: dữ liệu đầu vào, dữ liệu đầu ra, thuật toán

Bài toán đặt mục tiêu xây dựng mô hình dự đoán học lực của học sinh và ước lượng các yếu tố ảnh hưởng đến kết quả học tập, cũng như ước lượng lượng rượu tiêu thụ giữa các học sinh. Bộ dữ liệu gốc có hai phần: học sinh học lớp toán và học sinh học lớp tiếng Bồ Đào Nha. Trong đề tài này, chúng ta chỉ quan tâm tới học sinh học lớp toán.

Dữ liệu đầu vào bao gồm:

* Thông tin cá nhân của học sinh (giới tính, tuổi tác, địa chỉ,...).
* Hoàn cảnh gia đình (nghề nghiệp của bố mẹ, học vấn của bố mẹ, tình trạng hôn nhân của bố mẹ,...).
* Thói quen học tập (thời gian tự học, sự giúp đỡ của gia đình/nhà trường,...).
* Các yếu tố khác (sức khoẻ, hoạt động ngoại khóa, sử dụng internet,...).

Dữ liệu đầu ra sẽ là:

* Cột G1 là cột điểm kỳ 1 của các em học sinh.

Bài toán được giải quyết bằng cách sử dụng các thuật toán học máy như Linear Regression, Random Forest để lựa chọn ra các đặc trưng quan trọng ảnh hưởng đến cột điểm G1, sau đó sử dụng Xgboost, Lightgbm, Catboost cho việc huấn luyện mô hình dự đoán giá trị đầu ra rồi dùng hàm đánh giá để so sánh hiệu suất giữa các mô hình để chọn ra mô hình cho kết quả tốt nhất.

##### 1.3. Các giải pháp hiện tại/thủ công để giải quyết bài toán

Để tìm ra xem học sinh cần giúp đỡ hay không có rất nhiều cách thủ công không cần áp dụng máy móc nhưng đặc biệt cần phần lớn sự quan tâm đến từ thầy, cô giáo cũng như phụ huynh của các em. Các cá nhân này cần để tâm hết sức đến kết quả học tập của các em thường xuyên mỗi học kỳ để nhận biết sớm nhất sự chênh lệch điểm số giữa các kỳ. Nếu các kỳ có sự chênh lệch lớn thì nên bắt đầu tìm hiểu để đưa ra kết luận xem sự chênh lệch này là tốt hay xấu cho học sinh. Từ đó khẩn trương tìm ra giải pháp để giúp đỡ các em kịp thời.

Để làm được như vậy, các biện pháp thủ công là không thể thiếu, như là:

* Theo dõi sổ liên lạc: Giáo viên ghi chép điểm số, nhận xét về từng học sinh vào sổ liên lạc.
* Gặp gỡ phụ huynh định kỳ: Trao đổi thông tin về tình hình học tập, rèn luyện của học sinh giữa giáo viên và phụ huynh.
* Tổ chức học nhóm: Giúp đỡ học sinh yếu kém bằng cách ghép nhóm với học sinh học khá giỏi.

Tuy nhiên, các phương pháp này còn mang tính chủ quan vì cần phải dựa vào nhiều đánh giá của giáo viên và phụ huynh, khó khăn trong việc xử lý thông tin, thiếu cơ sở dữ liệu để dự đoán.

##### 1.4. Các giải pháp hiện tại/thủ công để giải quyết bài toán

Việc khai phá dữ liệu đóng vai trò chủ đạo, là trái tim của bài toán. Khai phá dữ liệu với khả năng xử lý lượng lớn dữ liệu, tự động hóa quá trình phân tích và đưa ra dự đoán khách quan mang đến nhiều ưu thế vượt trội trong việc giải quyết bài toán. Nếu không có khai phá dữ liệu, ta không thể tìm ra điều gì đã truyền cảm hứng, động lực cho các em học tốt hơn cũng như dữ liệu nào sẽ làm ảnh hưởng tiêu cực đến kết quả học tập của các em. Ta có thể dự đoán được học sinh nào sẽ có nguy cơ điểm thấp, nhưng sau đó làm sao để giúp đỡ các em thì đều phải nhờ vào các thông tin bổ ích có được từ việc khai phá dữ liệu.

##### 1.5. Kết quả dự đoán/ứng dụng của khai phá dữ liệu sau khi giải quyết bài toán trên

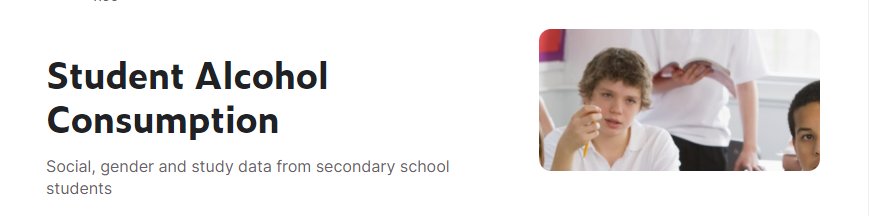
Sau khi có kết quả dự đoán, ta sẽ khoanh vùng các em có kết quả dự đoán điểm sẽ thấp. Dùng các đặc trưng có ảnh hưởng mạnh đến kết quả cuối kỳ làm chuẩn, rồi xem xét từng em dựa trên từng đặc trưng để có thể lên kế hoạch phù hợp với các em. Mỗi cây mỗi hoa, mỗi nhà mỗi cảnh, các em cũng vậy, cũng cần có phương pháp riêng phù hợp với hoàn cảnh của mỗi cá nhân. Là cơ sở để ta có thể xây dựng một hệ thống cảnh báo sớm các em học sinh có nguy cơ học tập sa sút cho giáo viên, phụ huynh. Cá nhân hoá phương pháp giảng dạy của giáo viên, giúp giáo viên biết rõ điểm mạnh, điểm yếu của từng học sinh để thiết kế bài giảng phù hợp, hỗ trợ học sinh yếu kém và phát huy hết khả năng của học sinh khá giỏi. Cùng lúc cung cấp thông tin để ban lãnh đạo dựa vào đó đưa ra các giải pháp mang tính vĩ mô nhằm nâng cao chất lượng giáo dục.

**Chương 2: Mô tả dữ liệu**

##### 1.5. Kích thước dữ liệu, chiều dữ liệu, trích dẫn nguồn dữ liệu

Dữ liệu ở định dạng có đuôi “.csv”, nặng khoảng 42 kB, chứa 33 cột (thuộc tính, mỗi cột đại diện cho một thông tin về học sinh) và 395 dòng (mẫu dữ liệu, mỗi dòng đại diện cho một học sinh). Nguồn dữ liệu mà đồ án này sử dụng đến từ Kaggle, một trang web nổi tiếng mà bất kỳ ai có làm việc với dữ liệu phần lớn đều đã từng nghe qua. Dữ liệu có tên “Student Alcohol Consumption”, được thu thập từ các cuộc khảo sát của học sinh lớp toán ở trong trường.

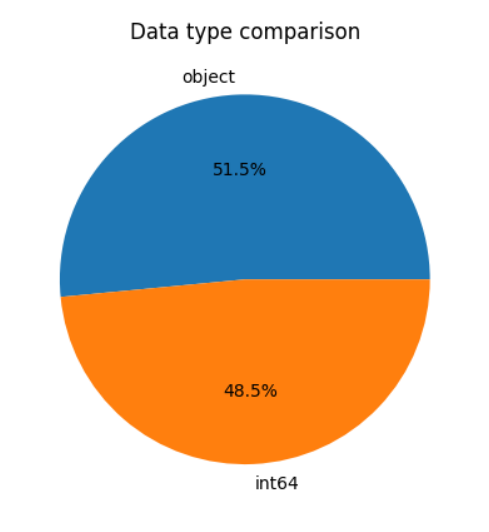
Nguồn dữ liệu: <https://www.kaggle.com/datasets/uciml/student-alcohol-consumption>



Hình 2.1 Trang bìa của dữ liệu ở trên web Kaggle

##### 1.5. Các kiểu dữ liệu

Bộ dữ liệu chứa hai kiểu dữ liệu chính là “object (51.5%)” và “int64 (48.5%)”. Object đại diện cho dữ liệu dạng chuỗi ký tự, là thông tin định tính như giới tính, địa chỉ, nghề nghiệp của bố mẹ,…int64 đại diện cho dữ liệu dạng số nguyên, là thông tin định lượng như tuổi, số lượng thành viên trong gia đình, điểm số, số lượng rượu tiêu thụ,…Một số biến trông có vẻ là số, nhưng trên thực tế, chúng mang tính phân loại. Ví dụ: Medu - trình độ học vấn của mẹ (0 - không có, 1 - trình độ tiểu học, 2 – lớp 5 đến lớp 9, 3 – trình độ trung học cơ sở hoặc 4 – trình độ đại học). Chỉ có 5 đặc điểm số nguyên thực sự là tuổi, sự vắng mặt và điểm số (G1, G2, G3). Và không có dữ liệu bị thiếu.

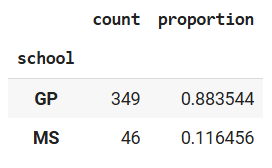


Hình 2.2 Biểu đồ tròn thể hiện sự phân phối của các kiểu dữ liệu trong bộ dữ liệu

* [Thống kê mô tả về dữ liệu]

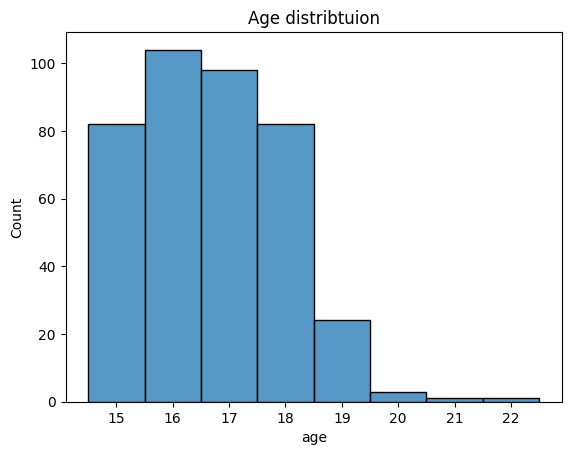
Thứ nhất, mục tiêu của chúng ta là hiểu, nắm bắt được các tác nhân có ảnh hưởng đến học lực của học sinh. Thứ hai, đào sâu vào sự ảnh hưởng của lượng rượu tiêu thụ ở mỗi bậc học lực.

Trong đoạn này, chúng ta sẽ xem xét biến mục tiêu (điểm số) và một số thuộc tính cơ bản (tuổi tác, giới tính, v.v.) cùng với mức độ tiêu thụ rượu bia để hiểu rõ hơn về tập dữ liệu. Chúng ta sẽ xem xét một số thuộc tính có thể ảnh hưởng đến biến mục tiêu và mức độ tiêu thụ rượu bia dựa trên quan điểm chủ quan và ảnh hưởng của chúng đến biến mục tiêu và mức độ tiêu thụ rượu bia.



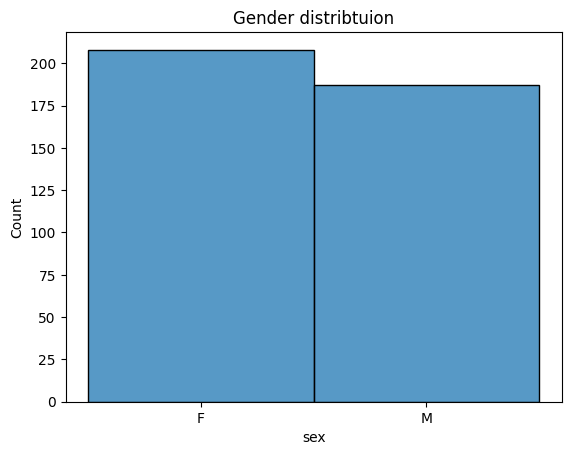
Hình 2.3. Thống kê số lượng học sinh của mỗi trường

Trong bộ dữ liệu chúng ta có hai ngôi trường khác nhau. Trong đó 88% học sinh đến từ trường Gabriel Pereira, phần còn lại thuộc về Mousinho da Silveira.



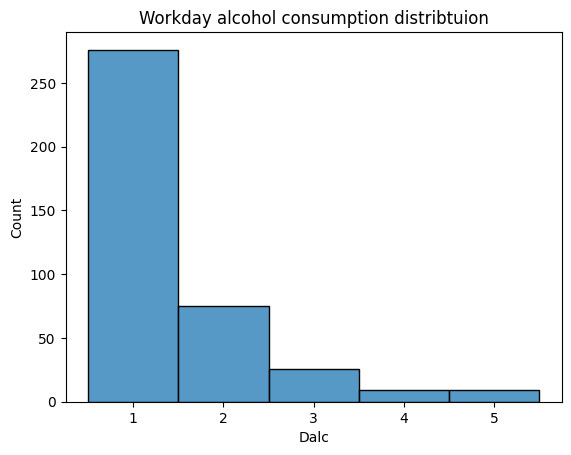
Hình 2.4. Biểu đồ phân phối độ tuổi của học sinh

Số lượng lớn học sinh nằm trong độ tuổi từ 15 đến 18. Học sinh có số tuổi lớn nhất là 22 tuổi.



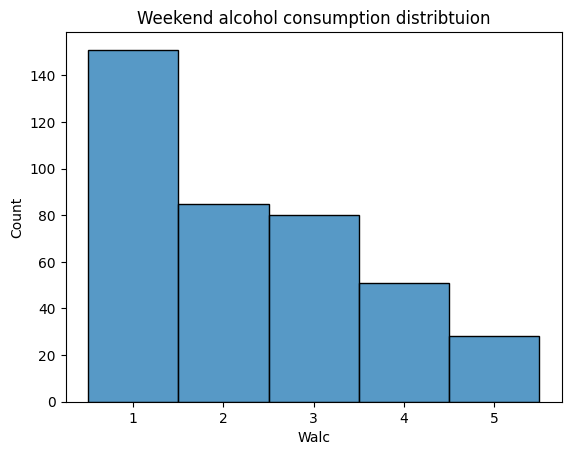
Hình 2.5. Biểu đồ phân phối giới tính của học sinh

Trong bộ dữ liệu này, số học sinh nam nhiều hơn số học sinh nữ nhưng sự cách biệt không đáng kể.



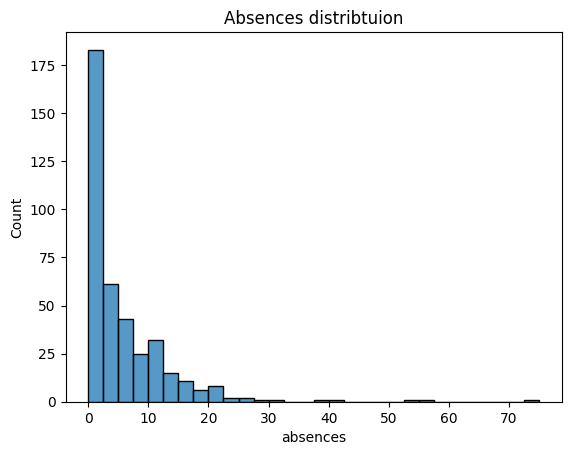
Hình 2.6. Biểu đồ phân phối lượng rượu bia tiêu thụ trong ngày làm việc

Từ biểu đồ 2.6. ta có thể kết luận rằng phần lớn học sinh (gần 70%) không uống rượu bia trong ngày thường ( 1 trong Dalc có nghĩa là mức tiêu thụ rất thấp). Vì trong dữ liệu không có lựa chọn 0 nên ta có thể hiểu được rằng sẽ có những trường hợp các bạn không uống rượu bia nhưng phải chọn 1 vì không có sự lựa chọn nào khác. Vì vậy, ta ngầm khẳng định rằng “mức tiêu thụ rất thấp” đồng nghĩa với “không tiêu thụ rượu bia”. 19% uống nhiều hơn bên “không tiêu thụ rượu bia” và 2.5& uống xuyên suốt các ngày.



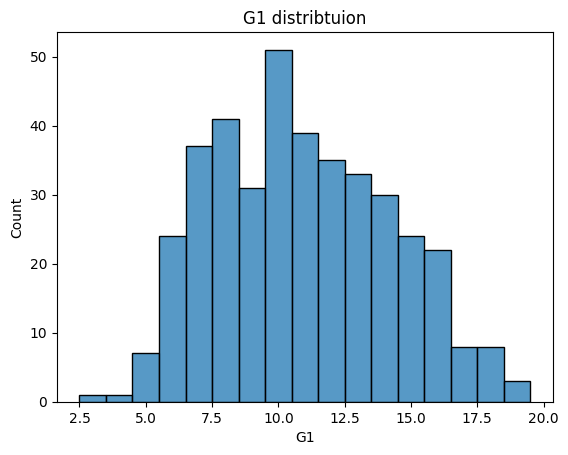
Hình 2.7. Biểu đồ phân phối mức tiêu thụ rượu bia ở các ngày cuối tuần

Có vẻ như học sinh thích uống vào thời điểm cuối tuần. Hợp lý, vì các ngày này học sinh không đến trường nên sẽ có nhiều thời gian rảnh hơn để hẹn hò, đi chơi. Tuy nhiên thì vẫn có rất nhiều học sinh không uống kể cả cuối tuần lẫn ngày thường, con số này đạt tới hơn 54%!

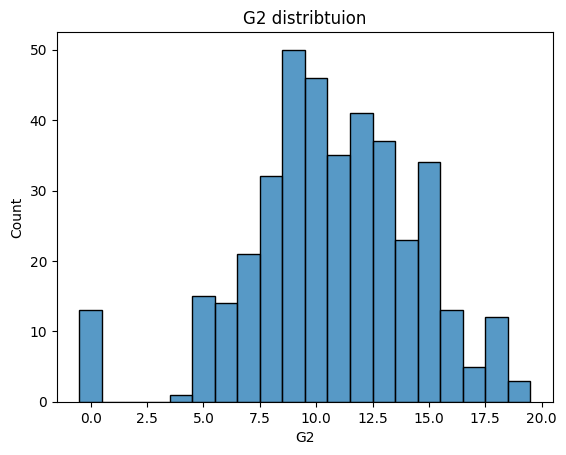


Hình 2.8. Biểu đồ phân phối số ngày vắng lớp của học sinh

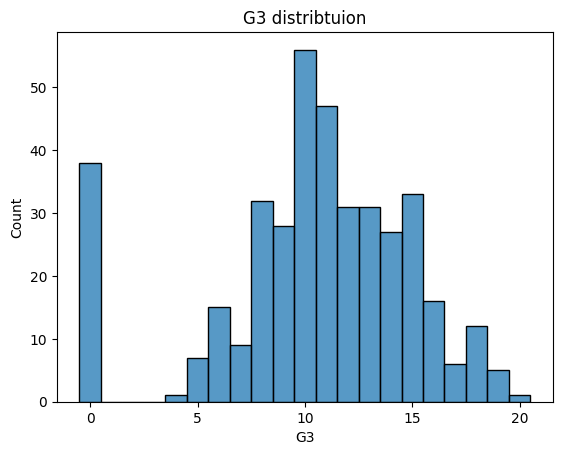
Phần lớn học sinh rất ít khi vắng lớp hoặc chưa bao giờ vắng. Nhưng cũng có một vài cá biệt vắng hơn 30 buổi.



Hình 2.9. Biểu đồ phân phối số điểm của kỳ 1



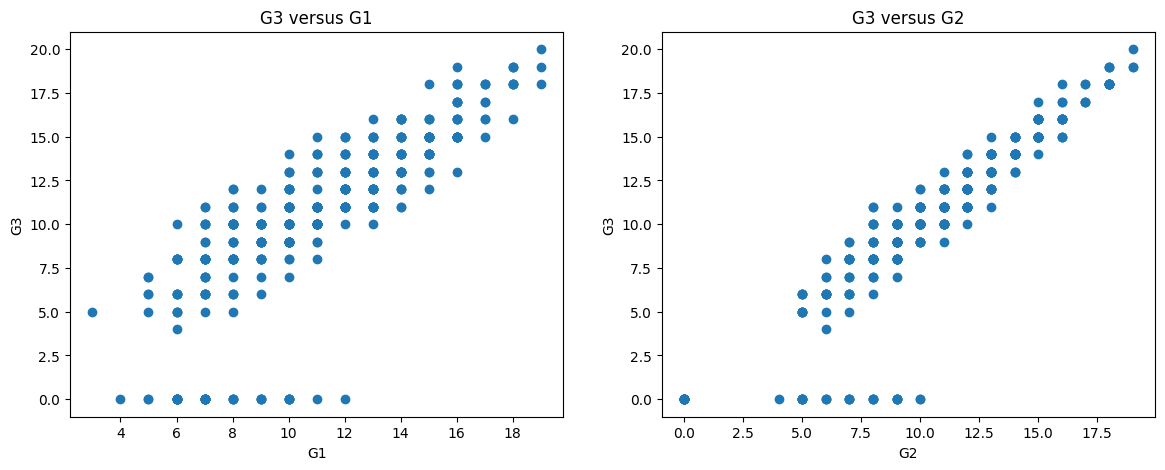
Hình 2.10. Biểu đồ phân phối số điểm của kỳ 2



Hình 2.11. Biểu đồ phân phối số điểm của kỳ 3

Biểu đồ phân phối số điểm của ba kỳ nhìn qua rất giống với phân phối chuẩn. G2 và G3 có các dữ liệu ngoại lai ở vị trí 0 điểm. Vì theo như [Portugal Grading System (scholaro.com)](https://www.scholaro.com/db/Countries/Portugal/Grading-System), số đỉểm thấp nhất có thể có được ở Bồ Đào Nha là 1. Sau khi tìm hiểu, ta có hai giả thuyết sau:

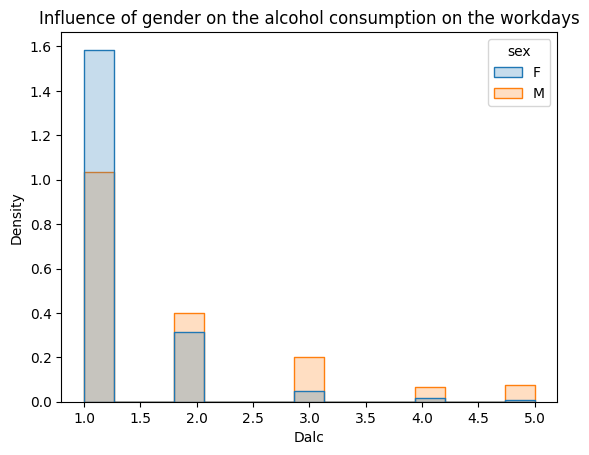
1. Học sinh của nhóm đầu tiên có G2, nhưng G3 bằng 0 có thể do bị thiếu giá trị hoặc vào thời điểm tạo tập dữ liệu, học sinh không có điểm do một số trường hợp.
2. Học sinh của nhóm thứ hai có điểm 0 ở G2 và cả G3. Chúng ta có thể giả định rằng những sinh viên này đã rời trường sau tiết học đầu tiên.



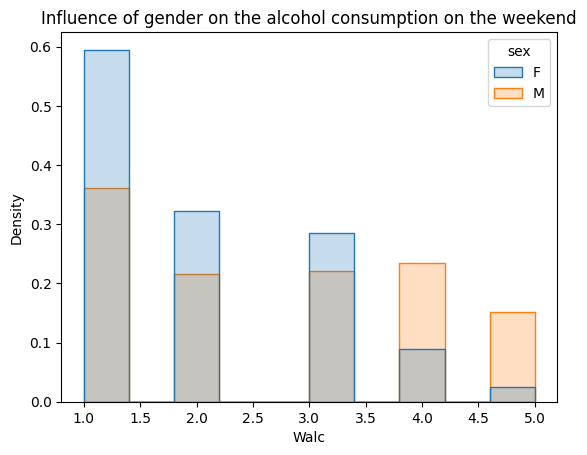
Hình 2.12. Biểu đồ Scatter giữa G3 và G1 (bên trái), giữa G3 và G2 (bên phải)

Dựa trên các điểm được scatter tập trung quanh line y=x, ta có giả thuyết các học sinh thường có điểm số gần như nhau qua các kỳ học. Vì vậy, để giảm chiều dữ liệu và giúp mọi thứ đơn giản hơn, ta chỉ lấy G1, cột điểm không có giá trị ngoại lai làm mục tiêu (sau này có thể dùng kết quả này áp dụng lên các kỳ sau).

Tiếp theo, ta sẽ tìm hiểu về lượng rượu bia tiêu thụ giữa các em học sinh. Đầu tiên chính là ảnh hưởng của giới tính lên mức tiêu thụ rượu bia.



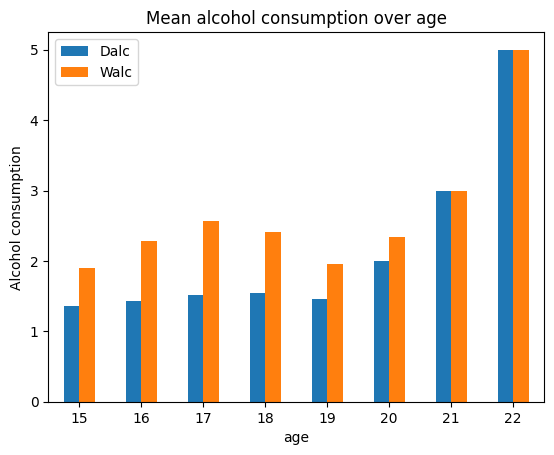
Hình 2.13. Lượng tiêu thụ rượu bia (dựa trên giới tính) ở các ngày trong tuần



Hình 2.13. Lượng tiêu thụ rượu bia (dựa trên giới tính) ở các ngày cuối tuần

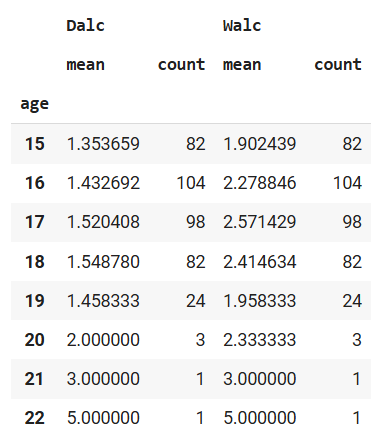
Chúng ta có thể thấy rõ rằng nam tiêu thụ rượu bia nhiều hơn nữ. Và cả hai nhóm này đều uống vào ngày cuối tuần nhiều hơn là trong tuần.

Bây giờ hãy tìm hiểu sự ảnh hưởng của tuổi tác trên lượng rượu bia tiêu thụ.



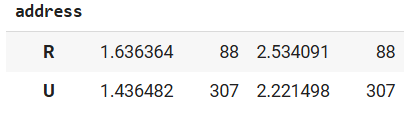
Hình 2.14. Lượng tiêu thụ rượu bia trung bình(dựa trên tuổi tác)

Khi tuổi càng tăng, lượng rượu tiêu thụ vào cuối tuần cũng tăng và đạt đỉnh điểm ở độ tuổi 17, sau đó giảm đi một chút và lại tăng đáng kể ở độ tuổi 20-22. Tuy nhiên, như chúng ta nhớ ở phần trước, chúng ta chỉ có một số học sinh trên 19 tuổi nên dữ liệu này không mang tính đại diện cho họ. Chúng ta có thể tóm tắt trên một bảng mức tiêu thụ rượu trung bình ở mọi lứa tuổi và số lượng học sinh ở độ tuổi này.

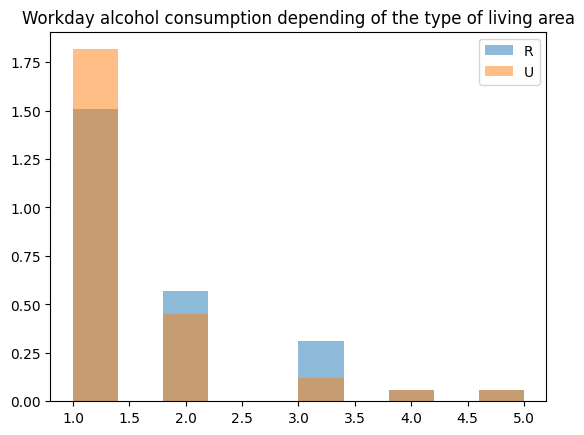


Hình 2.15. Bảng thống kê tuổi tác và lượng rượu bia tiêu thụ trung bình

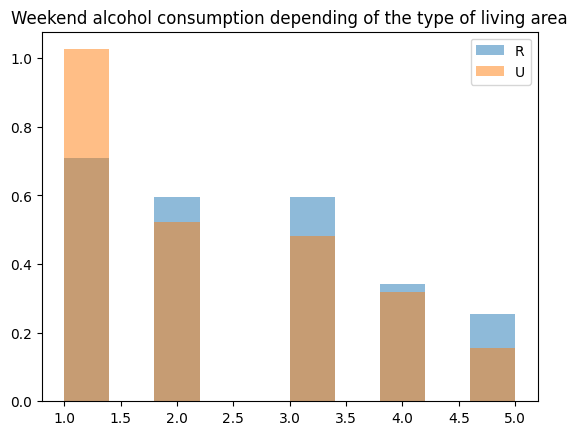
Các đặc điểm khác có thể ảnh hưởng đến việc tiêu thụ rượu bia là khu vực nơi sinh sống (thành thị hoặc nông thôn) và tình trạng chung sống của cha mẹ (sống chung hoặc ly thân).



Hình 2.16. Bảng thống kê số lượng học sinh từng khu vực

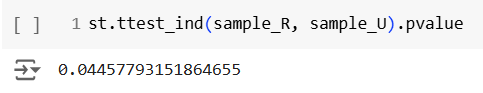


Hình 2.17. Biểu đồ phân phối lượng rượu bia tiêu thụ trong tuần theo từng khu vực



Hình 2.18. Biểu đồ phân phối lượng rượu bia tiêu thụ cuối tuần theo từng khu vực

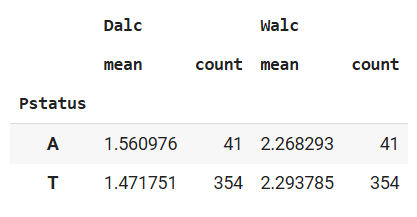
Có vẻ như học sinh sống ở nông thôn tiêu thụ nhiều rượu hơn, nhưng số lượng học sinh này không quá cao. Để chắc chắn chúng ta có thể tiến hành kiểm tra thống kê. Chúng ta sử dụng t-test, đặt mức ý nghĩa (significance level) bằng 0.05.

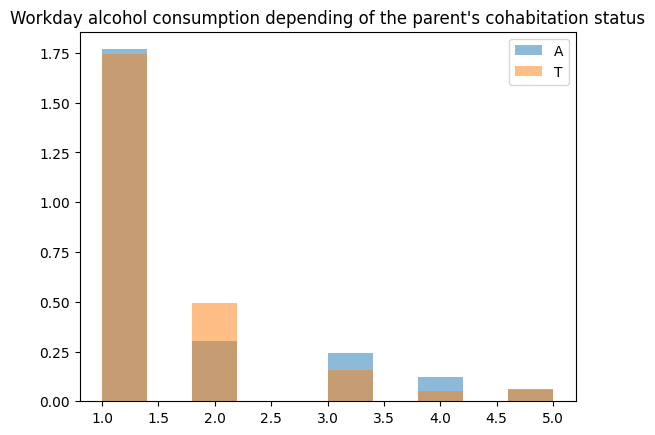


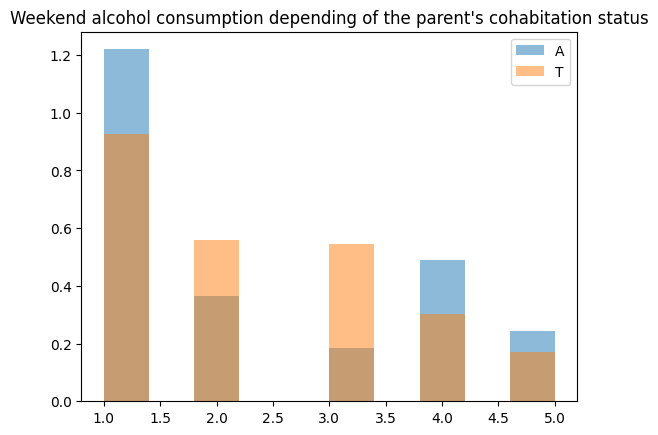
Hình 2.19. Kết quả chạy t-test

Giá trị p (p-value) nhỏ hơn mức ý nghĩa (significance level) cho thấy xác suất thu được mẫu dữ liệu này một cách ngẫu nhiên là rất thấp (trong điều kiện không có sự khác biệt về mức độ tiêu thụ rượu bia). Do đó, chúng ta có thể kết luận rằng có sự khác biệt đáng kể về mức độ tiêu thụ rượu bia giữa các nhóm. Cụ thể, học sinh sống ở khu vực nông thôn có xu hướng uống rượu bia nhiều hơn học sinh sống ở khu vực thành thị.

Tiếp theo là đến tình trạng hôn nhân của cha mẹ.



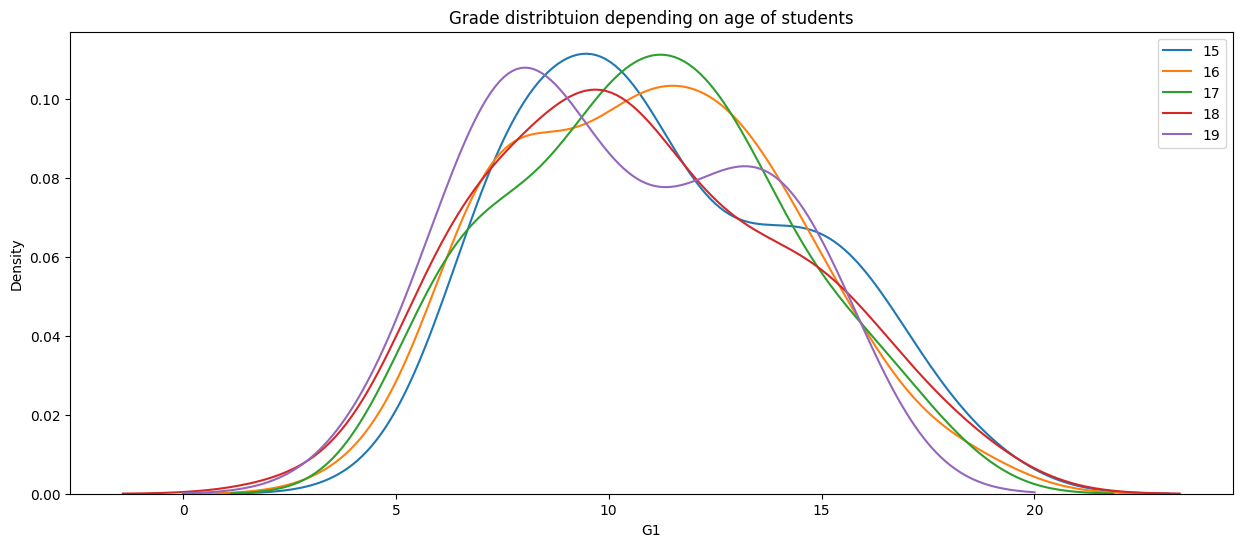




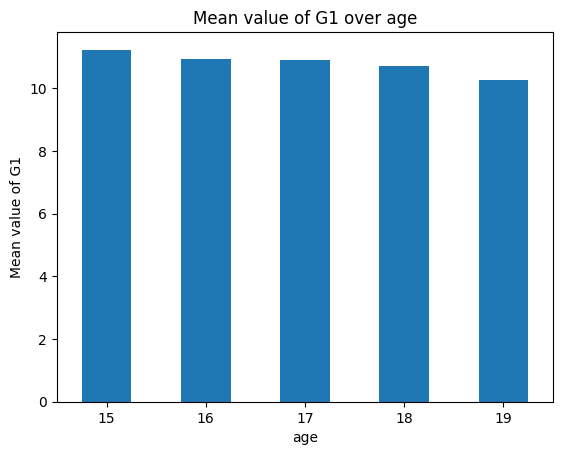
Học sinh sống xa cha mẹ uống nhiều hơn một chút vào các ngày trong tuần và ít hơn một chút vào cuối tuần so với những học sinh sống cùng cha mẹ. Nếu tính đến mẫu nhỏ học sinh mà cha mẹ sống xa nhau, chúng ta không thể nói rằng sự khác biệt là đáng kể. Vì vậy, tình trạng chung sống của cha mẹ không ảnh hưởng đến việc uống rượu của học sinh.

Các đặc trưng ảnh hưởng đến điểm số:

Chúng ta sẽ tìm hiểu làm sao mà các đặc trưng được kể trên sẽ ảnh hưởng đến biến mục tiêu (G1). Bắt đầu với tuổi tác, nó có ảnh hưởng đến điểm số ở trường không? Học sinh có đạt điểm cao hơn khi lớn lên không hay ngược lại? Chúng ta sẽ chỉ tính đến những học sinh dưới 20 tuổi vì chúng ta có đủ dữ liệu về họ.

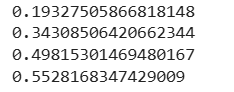


Hình 2.x. Phân bố điểm cho từng lứa tuổi.

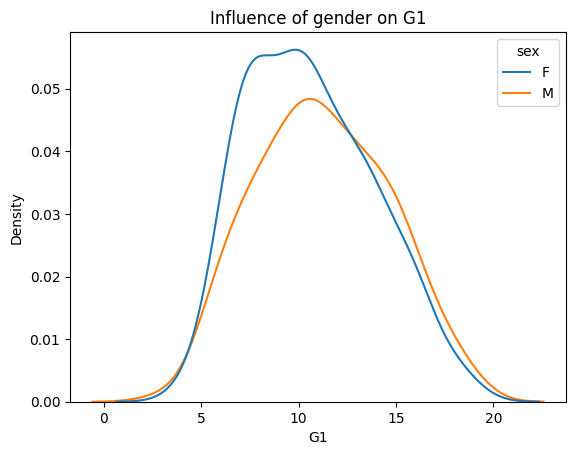


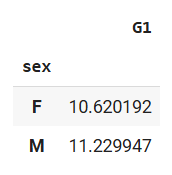
Hình 2.x. Điểm trung bình của từng lứa tuối

Sự phân bố của G1 qua các lứa tuổi không chênh lệch quá nhiều. Giá trị trung bình giảm nhẹ. Hãy kiểm tra xem sự khác biệt giữa các lứa tuổi có đáng kể hay không bằng cách áp dụng phép kiểm thống kê t-test với mức có nghĩa bằng 0.05.

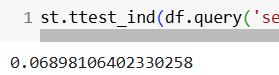


Không có sự khác biệt đáng kể, từ đây ta có thể kết luận tuổi tác không ảnh hưởng đến điểm số.

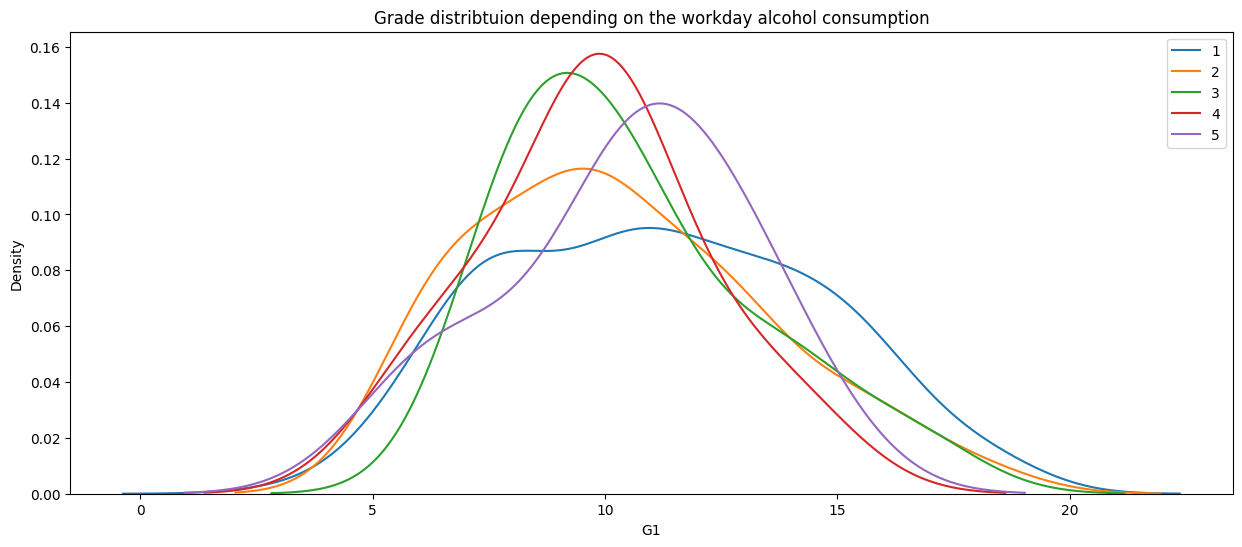


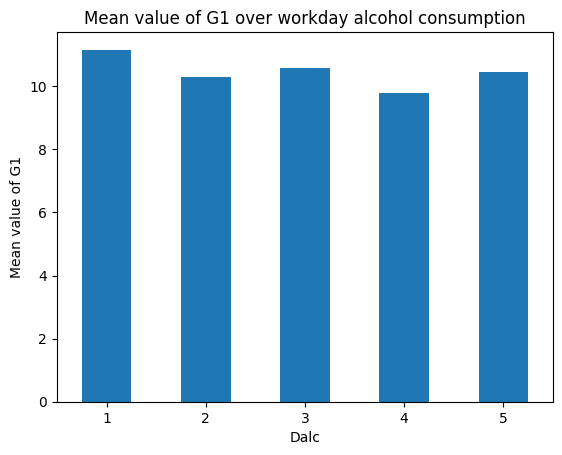


Sự phân bố hơi khác nhau và giá trị trung bình của nam cao hơn nữ. Sự khác biệt không đủ lớn để tin cậy nên chúng ta sẽ lại áp dụng kiểm định thống kê, cụ thể là t-test với mức ý nghĩa 0,05.



Giá trị p cao hơn mức ý nghĩa nên chúng ta không có đủ lý do để bác bỏ giả thuyết khống. Dựa trên điều này, chúng ta có thể kết luận rằng giá trị trung bình G1 của nam và nữ là như nhau. Và có thể nói rằng giới tính không ảnh hưởng đến điểm số.





Phân bố của các nhóm là khác nhau. Đối với những học sinh ít hoặc không tiêu thụ rượu

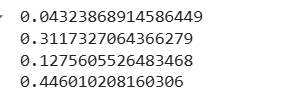
bia vào ngày thường, phân bố điểm số có một đỉnh rộng. Ngược lại, đỉnh phân bố hẹp hơn đối với những học sinh tiêu thụ nhiều rượu bia hơn.

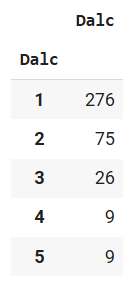
Giá trị trung bình của các nhóm có sự khác biệt nhỏ, và nhóm học sinh ít uống rượu bia

nhất có điểm trung bình cao nhất. Tuy nhiên, do kích thước mẫu tương đối nhỏ và số lượng học sinh ở mỗi nhóm không đồng đều nên chúng ta chưa thể đưa ra kết luận chắc chắn.

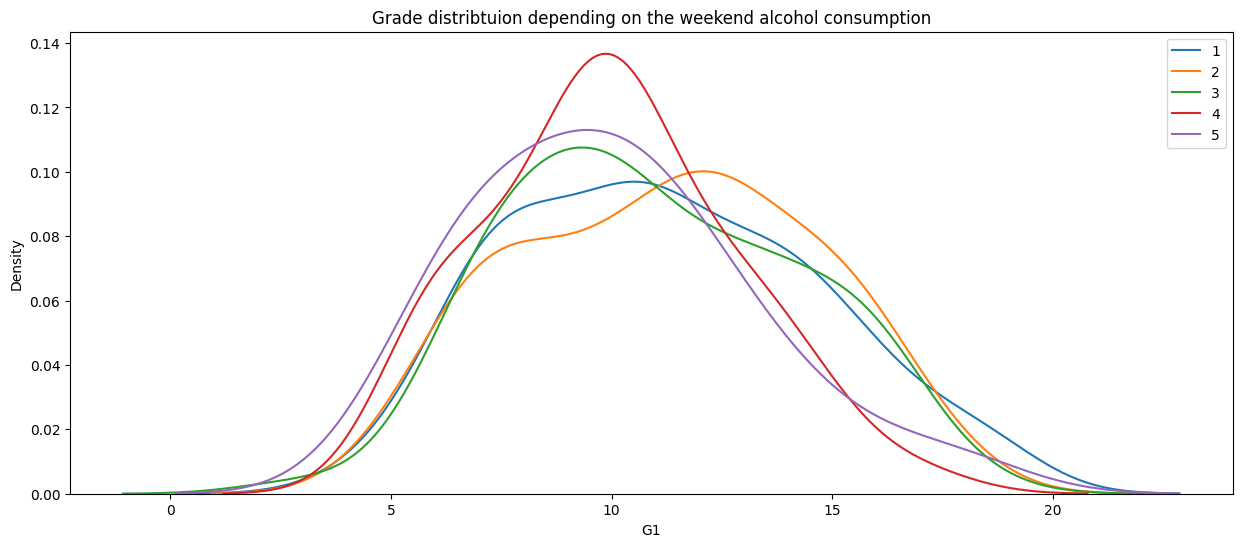
Để kiểm chứng sự khác biệt về điểm số giữa các nhóm có ý nghĩa thống kê hay không, ta

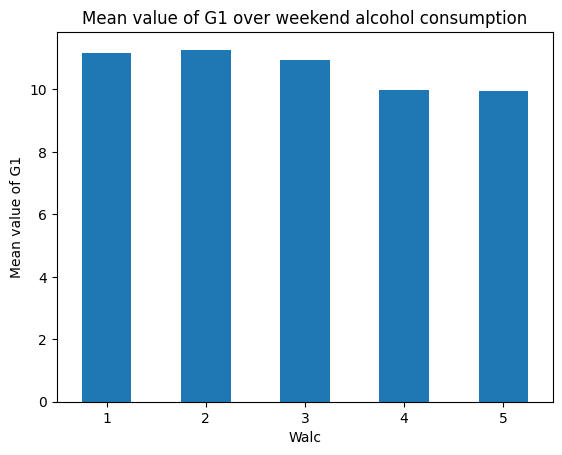
có thể sử dụng kiểm định t-test với mức ý nghĩa 0.05.



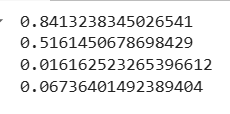


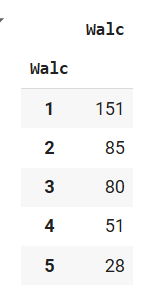
Kết quả kiểm định cho thấy có sự khác biệt có ý nghĩa thống kê về điểm số giữa nhóm ít uống rượu nhất và nhóm uống nhiều hơn một chút. Sự khác biệt về điểm số giữa các nhóm còn lại là không có ý nghĩa thống kê. Tuy nhiên, có thể là do bị ảnh hưởng bởi kích thước mẫu quá nhỏ. Nên ta nghiêng về giả thuyết rằng việc uống rượu bia vào ngày thường có ảnh hưởng đến điểm số, và uống ít rượu bia hơn sẽ dẫn đến điểm cao hơn.





Vì nhóm 2 có giá trị trung bình cao nhất, nên nó được dùng để so sánh với các nhóm khác. Vẫn như cũ chúng ta dùng t-test và mức ý nghĩa ở 0.05.





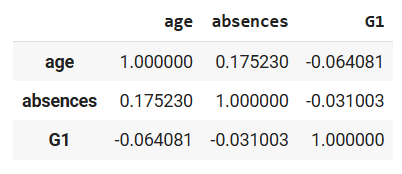
Dù cho giá trị p ở trường hợp cuối có phần cao hơn mức ý nghĩa, ta có giả thuyết rằng tiêu thụ nhiều rượu bia vào cuối tuần khiến cho điểm thấp đi. Học sinh có lượng rượu bia tiêu thụ thấp và vừa cuối tuần sẽ có điểm bằng nhau.

Dựa theo các luận điểm trên, ta rút ra được một vài giả thuyết như sau:

1. Độ tuổi của phần lớn học sinh trong tập dữ liệu nằm trong khoảng từ 15 đến 18. Học sinh lớn tuổi nhất là 22 tuổi.
2. Phần lớn học sinh (gần 70%) không uống rượu bia trong ngày làm việc, 19% uống nhiều hơn một ít, 2,5% uống nhiều trong ngày làm việc.
3. Tình hình tiêu thụ rượu bia thay đổi vào cuối tuần. Chỉ có 38% học sinh không uống rượu bia, 22% uống ít, 20% uống vừa phải, 13% - cao hơn mức trung bình, 7% - uống nhiều nhất.
4. Số điểm có phân phối tương tự như phân phối chuẩn. Có mối tương quan tuyến tính chặt chẽ giữa điểm trong kỳ đầu tiên, điểm trong kỳ thứ hai và điểm kỳ cuối cùng.
5. Học sinh nam uống rượu bia nhiều hơn học sinh nữ.
6. Với sự gia tăng tuổi tác, mức tiêu thụ rượu vào cuối tuần cũng tăng lên và đạt đỉnh điểm ở tuổi 17.
7. Học sinh sống ở nông thôn uống nhiều rượu bia hơn học sinh sống ở thành thị.
8. Tuổi tác và giới tính không ảnh hưởng đến điểm số.
9. Tiêu thụ rượu bia ảnh hưởng đến điểm số. Học sinh mà có mức tiêu thụ rượu bia cao sẽ có mức điểm thấp hơn.

Điều rất quan trọng là phải hiểu những đặc trưng nào ảnh hưởng đến mục tiêu và những đặc trưng nào không. Trong phần trước, chúng ta đã thực hiện việc này một cách ngẫu nhiên: chỉ xem xét ngẫu nhiên một số tính năng có thể ảnh hưởng đến mục tiêu dựa trên kiến thức và kinh nghiệm. Nhưng chúng ta có tới 30 đặc trưng và sẽ khá khó khăn nếu xem xét chính xác tất cả chúng. Chính vì vậy, có một cách hữu ích đó là chúng ta sẽ ước tính tầm quan trọng của đặc trưng và sau đó chỉ những đặc trưng quan trọng nhất mới được xem xét chính xác. Có nhiều cách khác nhau để ước tính tầm quan trọng của đặc trưng: phân tích tương quan, mô hình học máy,…

**Phân tích tương quan** có nghĩa là tìm mối tương quan tuyến tính giữa các biến số. Thật không may, chúng tôi chỉ có 2 biến số (tuổi và sự vắng mặt) và trong trường hợp của chúng tôi, việc áp dụng phân tích tương quan là vô nghĩa.

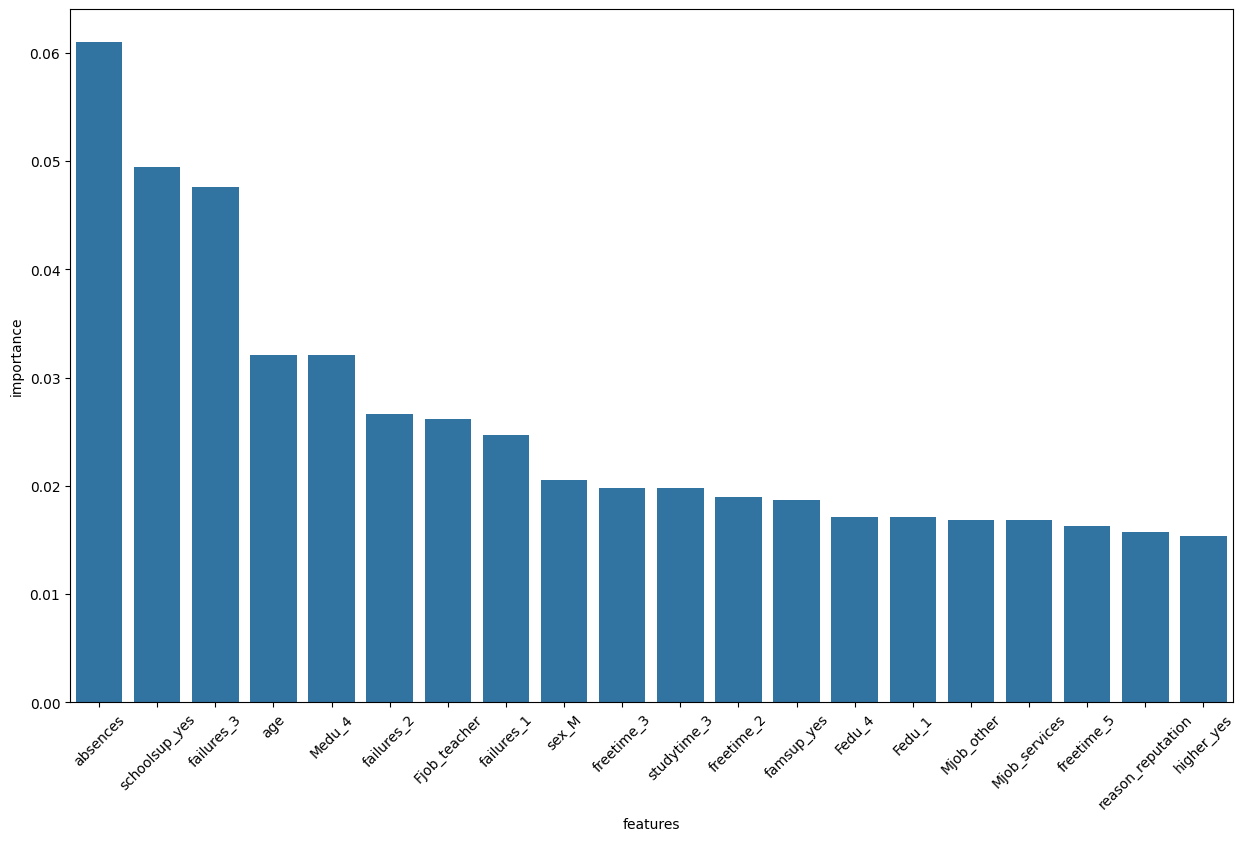


Sự tương quan giữa các đặc trưng rất yếu.

**Phân tích theo trọng số hồi quy tuyến tính**, nếu chúng ta đưa tất cả các tính năng về cùng một tỷ lệ, chúng ta có thể đánh giá tầm quan trọng của tính năng theo trọng số của mô hình hồi quy tuyến tính. Và sau khi thực hiện thuật toán ta rút ra được các đặc trưng quan trọng nhất trong số 30 đặc trưng:

1. failures - số lần rớt trước đây.
2. freetime - thời gian rảnh sau giờ học.
3. Medu – trình độ học vấn của mẹ.
4. studytime – thời gian tự học mỗi tuần.
5. schoolsup – các buổi dạy thêm ở trường.
6. famrel – chất lượng tình cảm gia đình.
7. Fjob – nghề nghiệp của cha.
8. Mjob - nghề nghiệp của mẹ.
9. traveltime – thời gian di chuyển tới trường.
10. higher – muốn học cao hơn.
11. health – tình trạng sức khoẻ hiện tại.
12. famsup – gia đình hỗ trợ việc học.

**Phân tích dựa theo** **Random Forest**, phương pháp này dựa trên thuộc tính có sẵn cho RandomForestRegressor. Quy trình rất đơn giản: fit RandomForestRegressor rồi gọi thuộc tính.

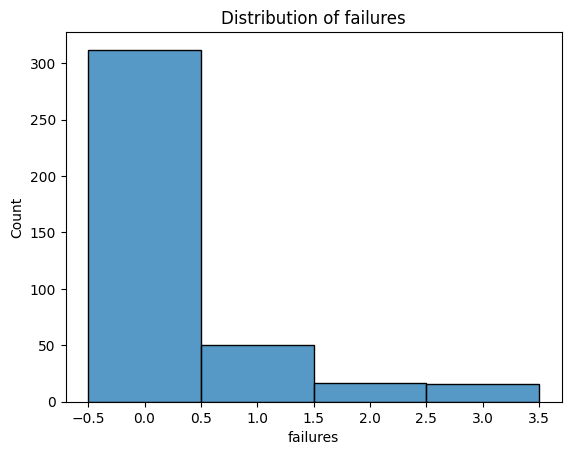


Các tính năng quan trọng nhất dựa trên Random Forest:

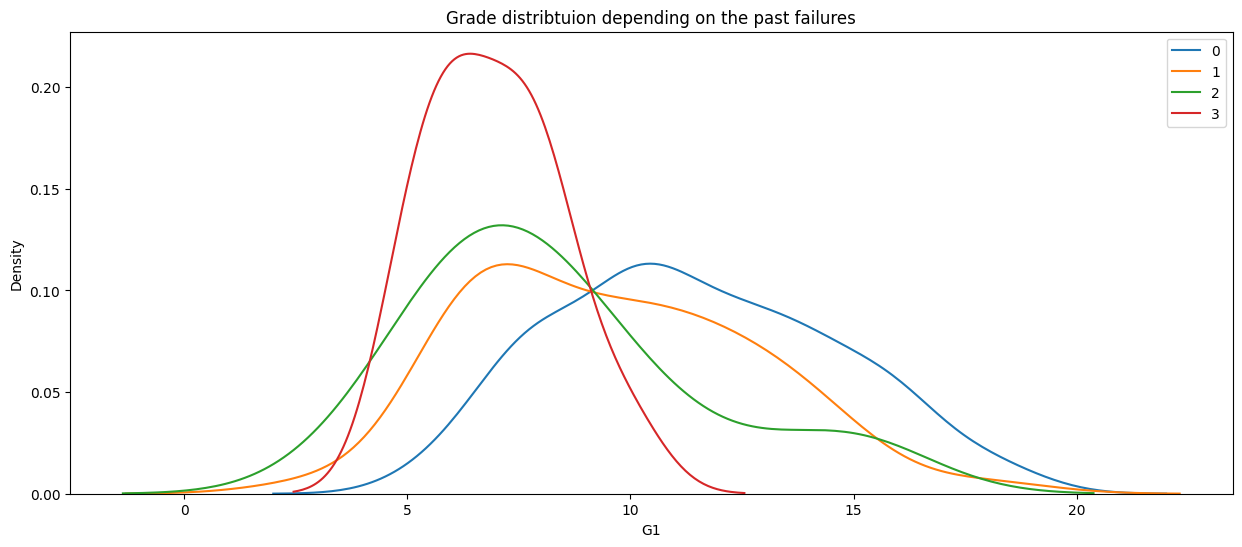
1. absences - số lần vắng mặt ở trường.
2. schoolsup - hỗ trợ giáo dục thêm.
3. failures - số lần thất bại của lớp trước đây.
4. tuổi.
5. Medu - trình độ học vấn của mẹ.
6. Fjob - nghề của cha.
7. Sex – giới tính.
8. freetime - thời gian rảnh sau giờ học.
9. studytime - thời gian tự học hàng tuần.
10. famsup - hỗ trợ giáo dục từ gia đình.
11. Fedu – trình độ học vấn của cha.
12. Mjob - công việc của mẹ.
13. reason - lý do để chọn trường này.
14. higher - muốn học cao hơn.

Từ 3 nhóm phân tích trên, ta kết luận được rằng:

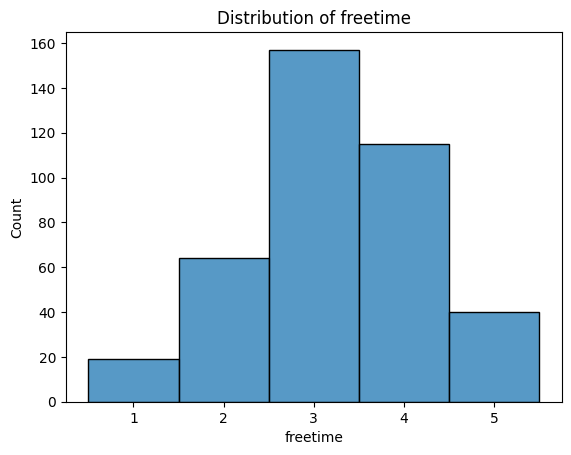
* Phân tích tương quan không được áp dụng vì phần lớn các đặc điểm mang tính phân loại.
* Dựa trên phân tích trọng số hồi quy tuyến tính, 5 đặc điểm hàng đầu có ảnh hưởng lớn nhất đến mục tiêu là: failures, freetime, Medu, studytime, schoolsup.
* Dựa trên Rừng ngẫu nhiên, 5 tính năng hàng đầu có ảnh hưởng lớn nhất đến mục tiêu là: absences, schoolsup, failures, age, Medu.
* Cả hai thuật toán đều đồng ý tôn trọng các đặc trưng như failures, freetime, Medu and schoolsup, nhưng có một số mâu thuẫn, trong khi Random Forest coi age, absences and sex là các đặc trưng quan trọng, thì trọng số của các đặc trưng này trong hồi quy tuyến tính lại rất nhỏ. Vì thế nó cần được làm rõ trong bước tiếp theo. Các tính năng quan trọng của cả hai thuật toán cần được nghiên cứu chi tiết.



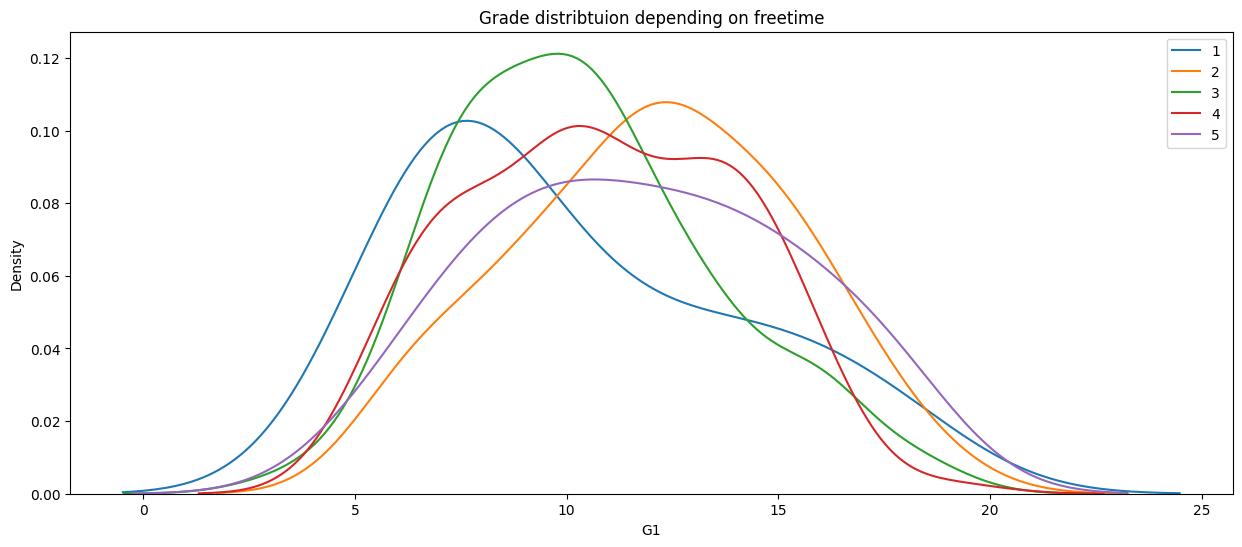
79% học sinh không có rớt lần nào trước đây. Chúng ta hãy xem thực tế này ảnh hưởng đến điểm số như thế nào.

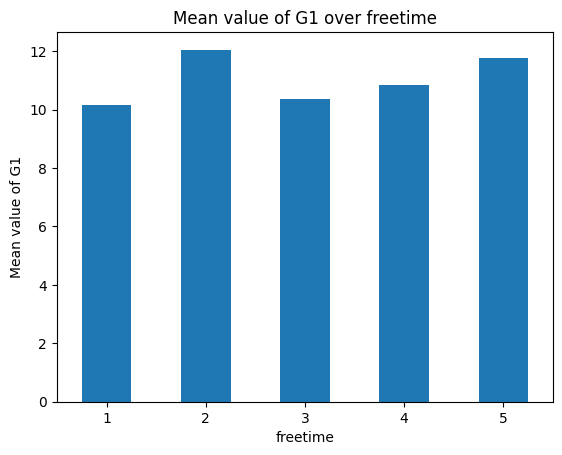


Dựa trên sự phân bổ, chúng ta có thể thấy rõ rằng những học sinh không rớt lần nào trước đây sẽ có điểm cao hơn. Điều đó có nghĩa là ở bước trước, cả hai thuật toán đều đúng về tầm quan trọng của tính năng này.

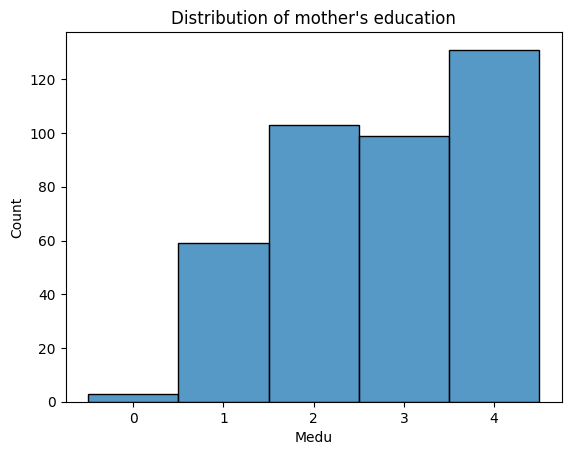


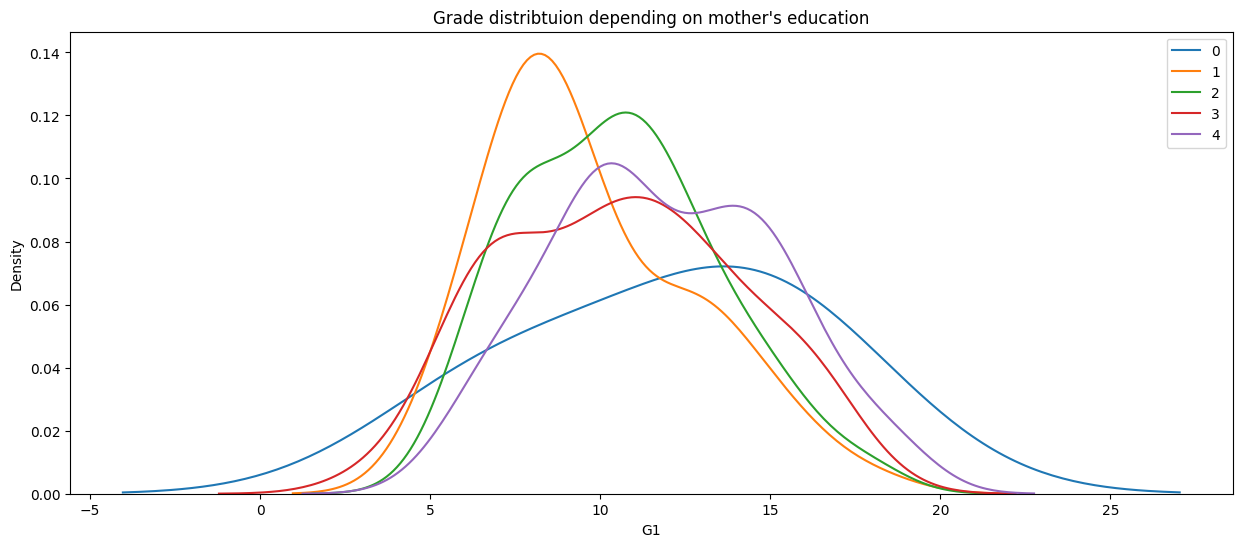
Đây là một phân phối bình thường. 5% học sinh có thời gian rảnh rất ít, 10% học sinh có nhiều thời gian rảnh. Chúng ta hãy xem nó ảnh hưởng đến mục tiêu như thế nào.

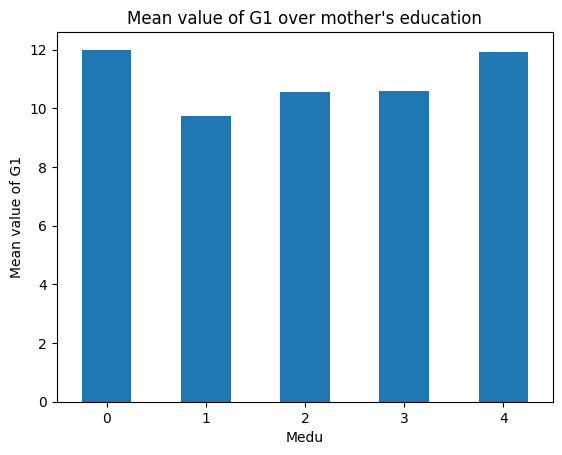




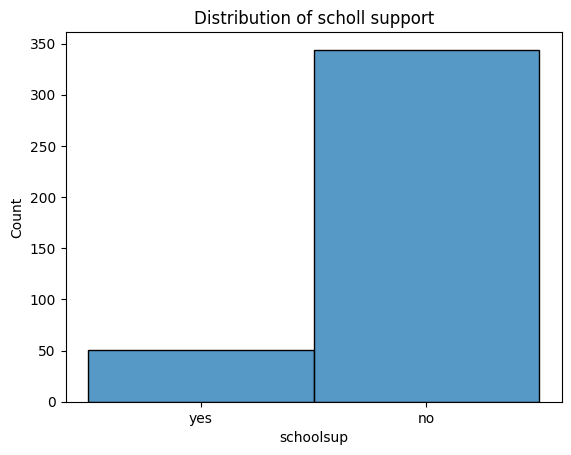
Từ việc phân phối và so sánh các giá trị trung bình, chúng ta có thể kết luận rằng những học sinh có rất ít thời gian rảnh sau giờ học có điểm kém hơn, điều này hợp lý vì các em có thể bận rộn với việc tập luyện thể thao hoặc trường âm nhạc (hoặc bất cứ điều gì) và họ không có đủ thời gian cho học tập.



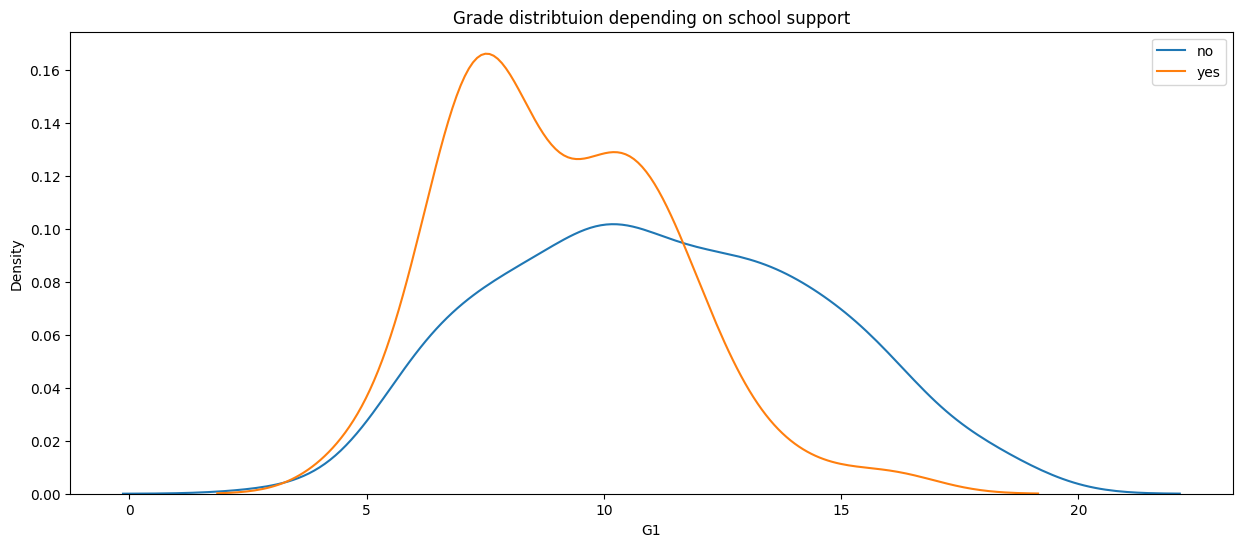


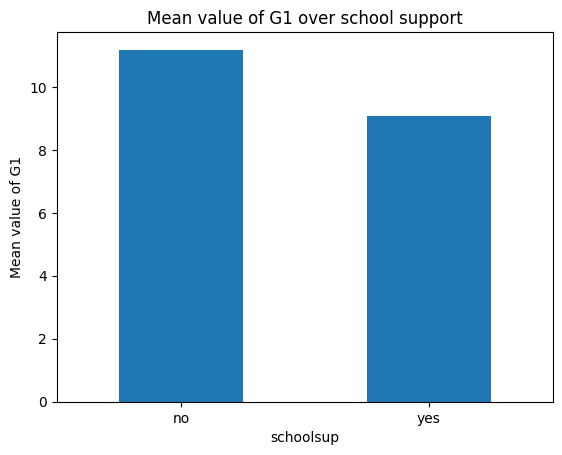


Mẫu dành cho học sinh, những người mẹ không có trình độ học vấn là quá nhỏ. Vì vậy, chúng ta có thể không cân nhắc đến nó. Dựa trên những quan sát khác, chúng ta có thể kết luận rằng trình độ học vấn của người mẹ càng cao thì điểm càng cao.

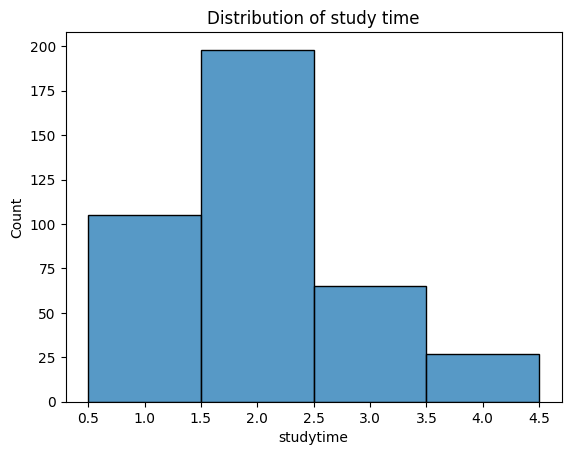


Chỉ 13% học sinh được hỗ trợ giáo dục thêm. Chúng ta có thể cho rằng họ khá kém môn toán nên phải học thêm, và có lẽ điểm của họ sẽ thấp hơn.

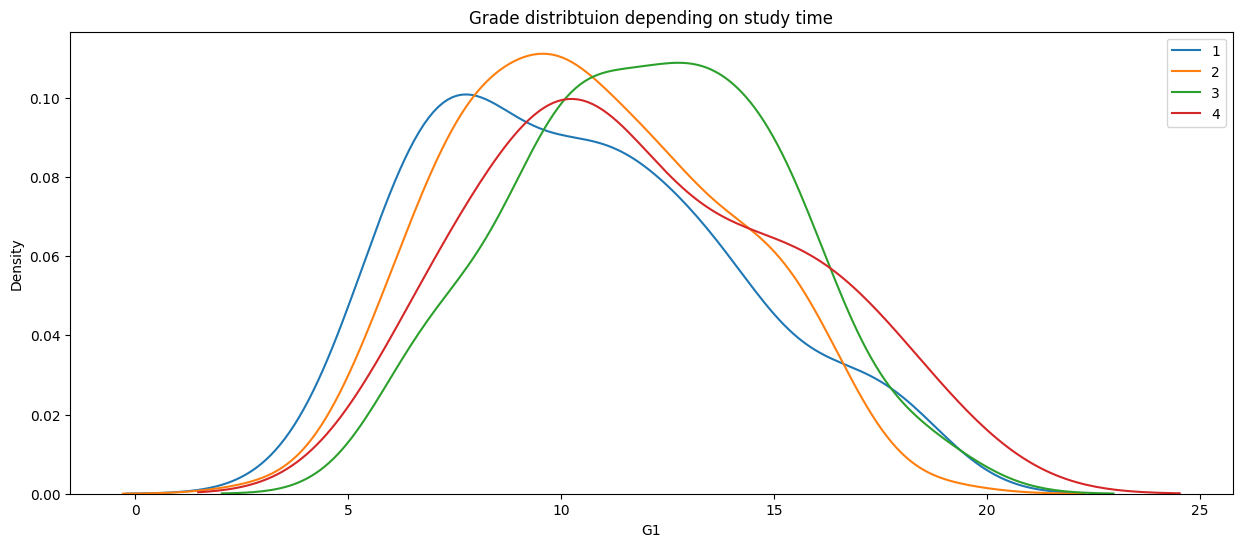


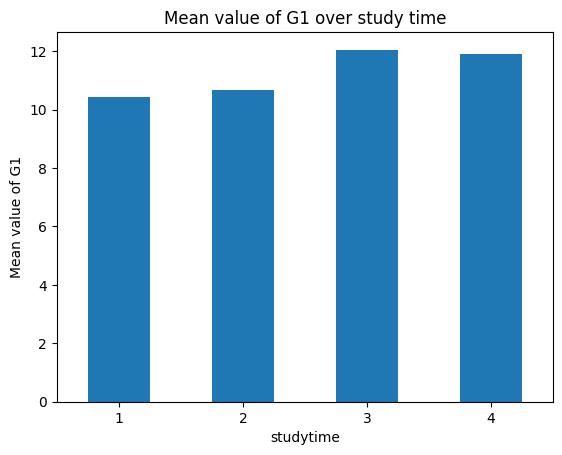


Vâng, giả định của chúng ta đã đúng. Những học sinh không được hỗ trợ giáo dục thêm sẽ có điểm trung bình cao hơn.



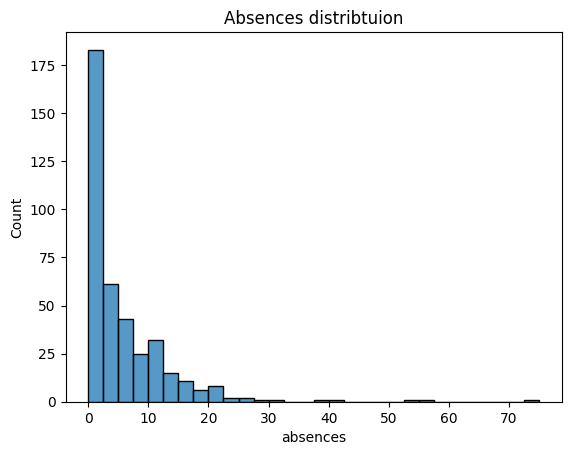
50% sinh viên học từ 2 đến 5 giờ mỗi tuần, 27% - dưới 2 giờ, 7% - hơn 10 giờ. Có vẻ như sinh viên không thích dành thời gian cho việc học. Bây giờ chúng ta có thể kiểm tra xem nó ảnh hưởng như thế nào đến điểm số của họ. Chúng ta có thể cho rằng, càng học nhiều giờ thì điểm càng cao.



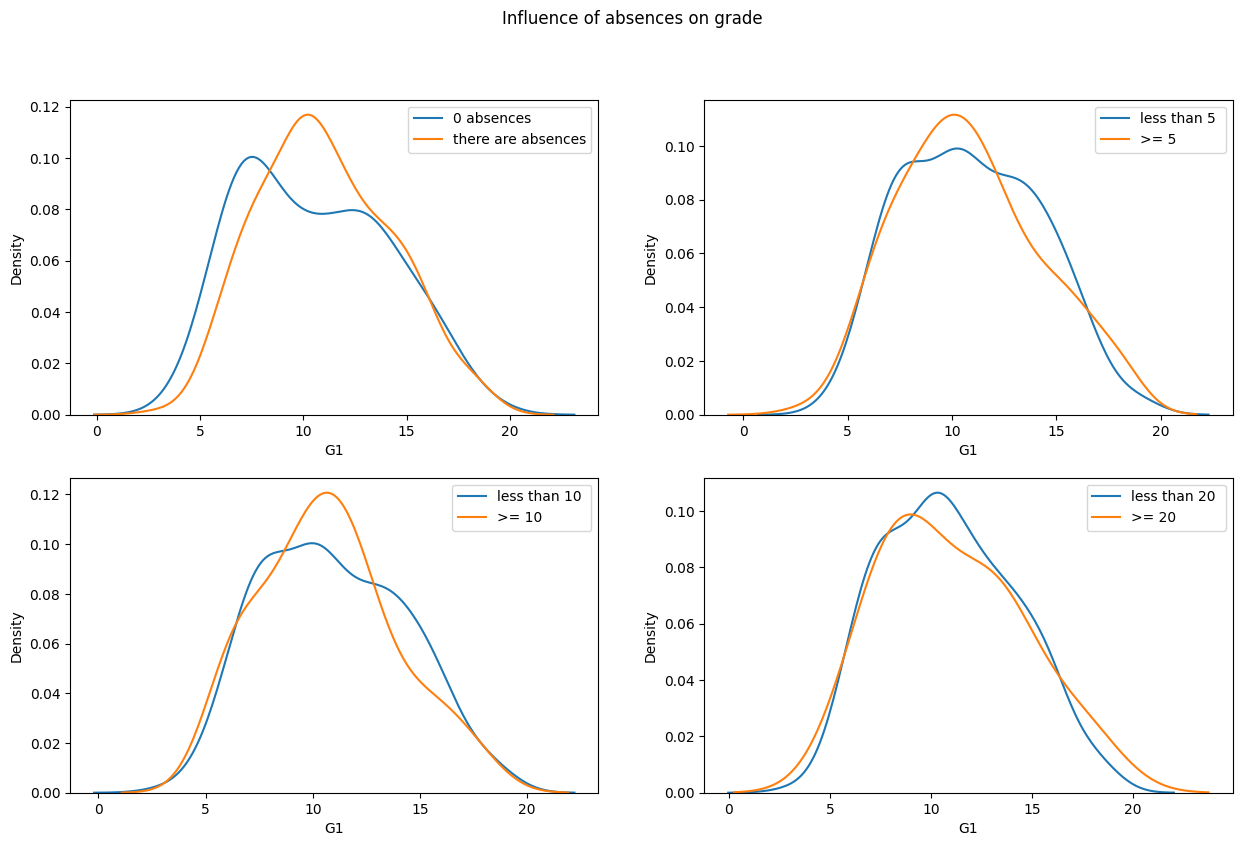


Những học sinh học hơn 5 giờ mỗi tuần sẽ đạt điểm cao hơn những học sinh học ít hơn 5 giờ.

Cho đến bây giờ chúng ta đã xem xét các đặc trưng được gọi là "quan trọng" theo cả hai phương pháp được sử dụng. Hãy kiểm tra đặc trưng absences. Dựa trên Random Forest, đặc trưng này có tầm quan trọng cao nhất, nhưng trọng lượng của nó trong Hồi quy tuyến tính lại thấp.



Dựa trên lẽ thường, những học sinh vắng mặt nhiều sẽ có điểm thấp hơn. Hãy chọn 4 ngưỡng threshold và so sánh điểm của những học sinh vắng nhiều và ít hơn ngưỡng đó: không vắng, ít hơn 5, ít hơn 10 và ít hơn 20.



Từ những phân bổ này, chúng ta có thể kết luận rằng số lần vắng mặt không ảnh hưởng đến điểm số. Thật kỳ lạ khi Random Forest lại coi tính năng này là quan trọng. Ngoài ra, Random Forest coi các đặc trưng quan trọng là giới tính và độ tuổi, nhưng trong phân tích của chúng ta phát hiện ra rằng không phải vậy. Dựa trên điều này, chúng ta có thể kết luận rằng việc tính toán tầm quan trọng của đối tượng dựa trên hệ số hồi quy tuyến tính sẽ đáng tin cậy hơn.

* [Những giả thiết khi thu thập dữ liệu]

1. Những học sinh không rớt bất kỳ lần nào trong lớp học trước đây sẽ có điểm cao hơn.
2. Những học sinh có rất ít thời gian rảnh sau giờ học sẽ có điểm thấp hơn.
3. Việc học của mẹ ảnh hưởng đến điểm số. Trình độ học vấn của mẹ càng cao thì điểm càng cao.
4. Những học sinh phải học thêm sẽ có điểm thấp hơn.
5. Những học sinh học hơn 5 giờ mỗi tuần sẽ đạt điểm cao hơn những học sinh học ít hơn 5 giờ.
6. Số lần vắng mặt không ảnh hưởng đến điểm số.
7. Vì sự vắng mặt, tuổi tác và giới tính được Random Forest coi là các đặc trưng quan trọng, nhưng thực tế không phải vậy, chúng ta có thể kết luận rằng việc tính toán tầm quan trọng của các đặc trưng dựa trên hệ số hồi quy tuyến tính là đáng tin cậy hơn cho tập dữ liệu của chúng ta, bởi vì các đặc trưng quan trọng từ Hồi quy tuyến tính cho thấy rằng họ ảnh hưởng đến điểm số.

**Chương 3: Phương pháp khai phá dữ liệu**

* [Quy trình khai phá dữ liệu]

Tiền xử lý dữ liệu

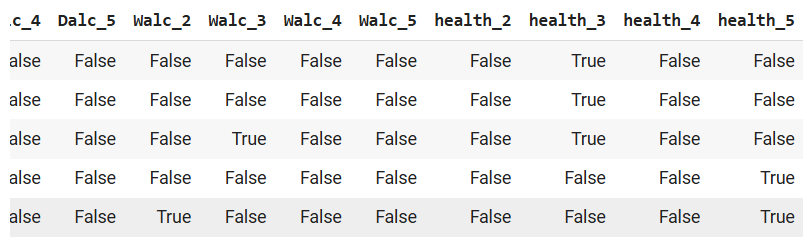
Chuẩn bị dữ liệu

Chúng ta sẽ chuẩn bị dữ liệu cho việc huấn luyện mô hình máy học. Như đã đề cập từ trước, các cột điểm G1, G2, G3 có mối tương quan tuyến tính sâu sắc vì vậy chúng ta sẽ chỉ tập trung vào cột điểm G1 và chỉ dùng nó cho phần xây dựng mô hình. Ta có 30 đặc trưng, tất cả cột trừ index và điểm số. Mục tiêu của chúng ta là G1.

features = df.copy().drop(['G1', 'G2', 'G3'], axis=1)  
target = df.copy()['G1']

Mã hóa dữ liệu dạng phân loại

Chúng ta sử dụng one hot encoding để mã hoá

 Chia tập train – test Để ước lượng hiệu suất của mô hình trong tương lai, chúng ta sẽ chia ra tập train và tập test với tỷ lệ 80 – 20.

features\_train, features\_test, target\_train, target\_test = train\_test\_split(features, target, test\_size=0.2, random\_state=5)

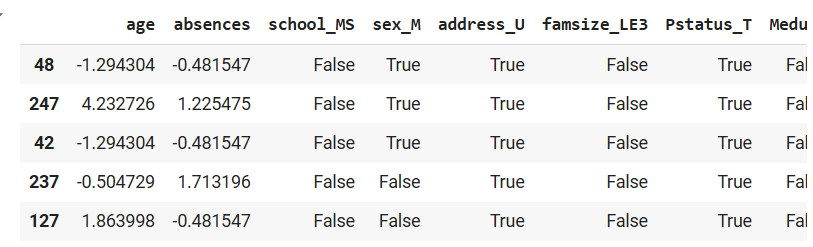
Scale các đặc trưng

Scale 2 đặc trưng bằng phương pháp standardized.

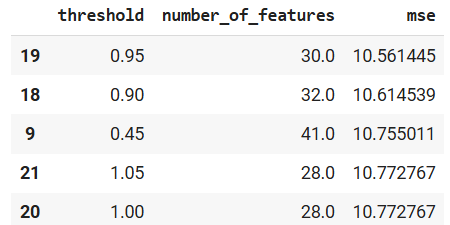
scaler = StandardScaler()

features\_train[['age', 'absences']] = scaler.fit\_transform(features\_train[['age', 'absences']])

features\_test[['age', 'absences']] = scaler.transform(features\_test[['age', 'absences']])



Loại bỏ các đặc trưng không ảnh hưởng đến cột G1 của chúng ta, giúp giảm chiều dữ liệu. Chúng ta sẽ làm như cách mà ta tìm ra các đặc trưng quan trọng như đã làm ở trên nhưng lần này chỉ làm việc với tập train. Đặt một ngưỡng threshold cho các đặc trưng quan trọng, các đặc trưng không quan trọng sẽ được loại bỏ.



Mean\_squared\_error tốt nhất ở ngưỡng 0,9 với 32 đặc trưng. Loại bỏ các đặc trưng không quan trọng (bao gồm cả dummy). Từ 69 đặc trưng, chúng ta đã giảm chiêù dữ liệu xuống còn 32 đặc trưng.

* [Nguyên lý hoạt động của mô hình/thuật toán]

Đích đến của bài toán là dự đoán ra số điểm của học sinh. Để làm được như vậy chúng ta sẽ cần hoàn thành 3 bước sau:

1. Huấn luyện mô hình.
2. Đánh giá hiệu suất của mô hình trên tập thử nghiệm.
3. Kiểm tra tính phù hợp của mô hình (so sánh với mô hình hằng số).

* [Cách cài đặt mô hình, các tham số (nếu có) của mô hình]

Để đào tạo, chúng ta sẽ sử dụng XGBRegressor từ thư viện xgboost. Sự lựa chọn dựa trên thực tế là mô hình này thường mang lại kết quả tốt nhất trên dữ liệu bảng.

GridSearchCV sẽ được sử dụng để điều chỉnh các siêu tham số của mô hình.

parameters = {'max\_depth' : [2, 3, 4],

              'n\_estimators' : [20, 30, 50, 75, 100],

              'eta' : [0.5, 0.3, 0.1, 0.05]}

grid\_search\_xgb = GridSearchCV(estimator=XGBRegressor(random\_state=42),

                               param\_grid=parameters,

                               scoring='neg\_mean\_squared\_error',

                               cv=5,

                               verbose=1)

grid\_search\_xgb.fit(features\_train\_truncated, target\_train)

Tham số tốt nhất: {'eta': 0.05, 'max\_depth': 3, 'n\_estimators': 100}

regressor = XGBRegressor(max\_depth=2,

                         n\_estimators=20,

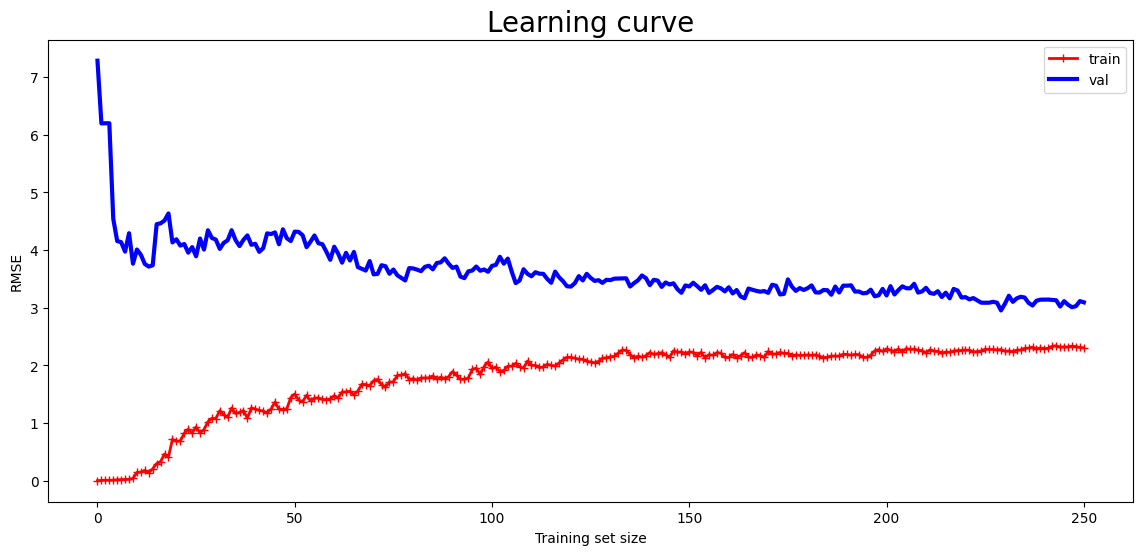
                         eta=0.5,

                         random\_state=42)

regressor.fit(features\_train\_truncated, target\_train)

* [Tiêu chí đánh giá mô hình]

RMSE (lỗi bình phương trung bình gốc) sẽ được sử dụng để đánh giá.



* Với bài toán phân loại (Classification): Accuracy, Precision, Recall, F1-score, AUC
* Với bài toán hồi quy (Regression): mean square error (MSE), root mean square error (RMSE), mean absolute error (MAE).
* Phân cụm (Clustering): Silhouette Coefficient, Dunn’s Index

**Chương 4: Thực nghiệm**

* [Yêu cầu về chương trình]
* Ngôn ngữ lập trình, phiên bản

Ngôn ngữ lập trình được sử dụng là Python, phiên bản 3.10.12, tích hợp sẵn trên Google Colab. Khuyến khích chạy chương trình trên Google Colab để thuận tiện theo dõi cũng như đồng bộ các phiên bản của chương trình, cho ra trải nghiệm tốt nhất.

* Thư viện, phiên bản

Các thư viện được sử dụng trong chương trình gồm có:

1. Numpy, 1.25.2
2. Pandas, 2.0.3
3. Seaborn, 0.13.1
4. import matplotlib.pyplot as plt
5. from scipy import stats as st
6. from sklearn.preprocessing import StandardScaler
7. from sklearn.linear\_model import LinearRegression
8. from sklearn.ensemble import RandomForestRegressor
9. from sklearn.model\_selection import GridSearchCV
10. from sklearn.model\_selection import train\_test\_split
11. from sklearn.neighbors import KNeighborsRegressor
12. from sklearn.metrics import mean\_squared\_error
13. from sklearn.model\_selection import cross\_val\_predict
14. from xgboost import XGBRegressor
15. from sklearn.model\_selection import RepeatedKFold
16. from sklearn.dummy import DummyRegressor

* Cấu hình máy tính

Intel Core i5-1135g7 (phần cứng thật của laptop), ram 12 GB (của Google Colab), ổ đĩa 100 GB (của Google Colab)

* [Mô tả chi tiết các bước khai phá dữ liệu]
* Input, output từng bước

**Chương 5: Kết quả**

* [Kết quả khai phá dữ liệu]
* Dựa trên [Tiêu chí đánh giá mô hình]. Mô hình có chính xác hay không?
* Dựa trên tính giải thích được của mô hình (interpretability). Có thể giải thích tại sao mô hình ra kết quả chính xác được hay không?
* [So sánh kết quả thực tế với kết quả dự đoán]
* Chất lượng của dữ liệu
* Nguyên lý hoạt động của thuật toán có giải quyết được vấn đề hay không?

**Chương 6: Thảo luận và kết luận**

Khả năng ứng dụng của giải pháp/mô hình

Mô hình dự đoán học lực của học sinh dựa trên khai phá dữ liệu mang nhiều tiềm năng ứng dụng trong thực tế:

* Hỗ trợ giáo viên: Giúp giáo viên xác định học sinh có nguy cơ học tập sa sút sớm, từ đó có kế hoạch hỗ trợ cá nhân hóa hiệu quả.
* Hỗ trợ phụ huynh: Cung cấp thông tin về tình hình học tập của con em mình, giúp phụ huynh nắm bắt kịp thời và cùng giáo viên hỗ trợ con em mình.
* Phát triển phương pháp giảng dạy: Giúp giáo viên hiểu rõ điểm mạnh, điểm yếu của từng học sinh, từ đó thiết kế bài giảng phù hợp và nâng cao hiệu quả giảng dạy.
* Xây dựng chính sách giáo dục: Cung cấp dữ liệu phân tích cho ban lãnh đạo trường học và cơ quan quản lý giáo dục, hỗ trợ việc đưa ra các giải pháp mang tính vĩ mô nhằm nâng cao chất lượng giáo dục.

Ưu điểm – nhược điểm của giải pháp/mô hình

Ưu điểm:

* Khách quan: Mô hình dựa trên phân tích dữ liệu khách quan, giảm thiểu yếu tố chủ quan của con người.
* Hiệu quả: Nâng cao hiệu quả quản lý giáo dục, phát hiện sớm vấn đề và đưa ra giải pháp kịp thời.
* Dễ dàng áp dụng: Mô hình có thể dễ dàng áp dụng cho nhiều trường học, nhiều đối tượng học sinh.

Nhược điểm:

* Phụ thuộc vào chất lượng dữ liệu: Mô hình sẽ cho kết quả chính xác khi dữ liệu thu thập đầy đủ, chính xác và đại diện.
* Tính giải thích: Việc giải thích chi tiết kết quả dự đoán của mô hình có thể phức tạp, đòi hỏi chuyên môn về khai phá dữ liệu.
* Thiếu dữ liệu: Mô hình có thể không chính xác khi thiếu dữ liệu về một số đặc trưng quan trọng.

Đề xuất

* Nâng cao chất lượng dữ liệu: Cần thu thập dữ liệu đầy đủ, chính xác và đại diện để mô hình dự đoán chính xác hơn.
* Xây dựng hệ thống theo dõi dữ liệu: Thường xuyên cập nhật dữ liệu và đánh giá hiệu quả của mô hình.
* Phát triển công cụ trực quan: Tạo giao diện trực quan, dễ sử dụng để giáo viên và phụ huynh có thể dễ dàng tiếp cận và sử dụng thông tin từ mô hình.
* Kết hợp với các giải pháp khác: Kết hợp mô hình với các giải pháp hỗ trợ học tập khác như học liệu trực tuyến, học nhóm, để nâng cao hiệu quả hỗ trợ học sinh.

Kết luận:

Mô hình dự đoán học lực học sinh dựa trên khai phá dữ liệu là một giải pháp tiềm năng, hứa hẹn mang đến nhiều lợi ích cho việc nâng cao chất lượng giáo dục. Tuy nhiên, việc phát triển và ứng dụng mô hình cần được thực hiện một cách khoa học, dựa trên sự kết hợp giữa công nghệ và chuyên môn giáo dục.

add\_circle

**Tài liệu tham khảo**