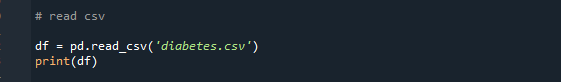
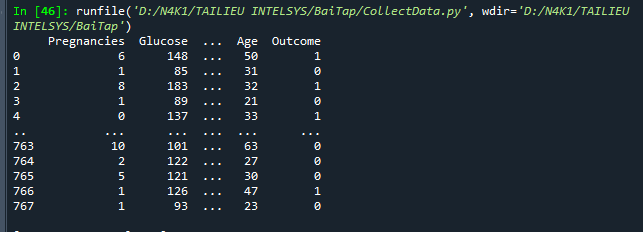
**DATA COLLECTION**

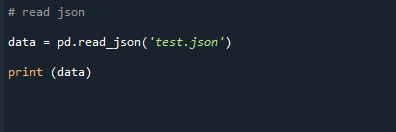
* File csv

****

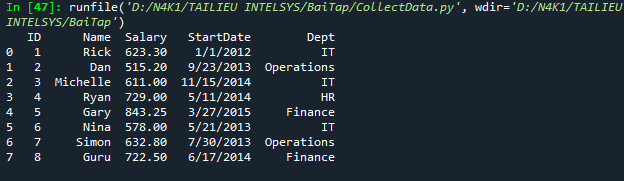
Kết quả:



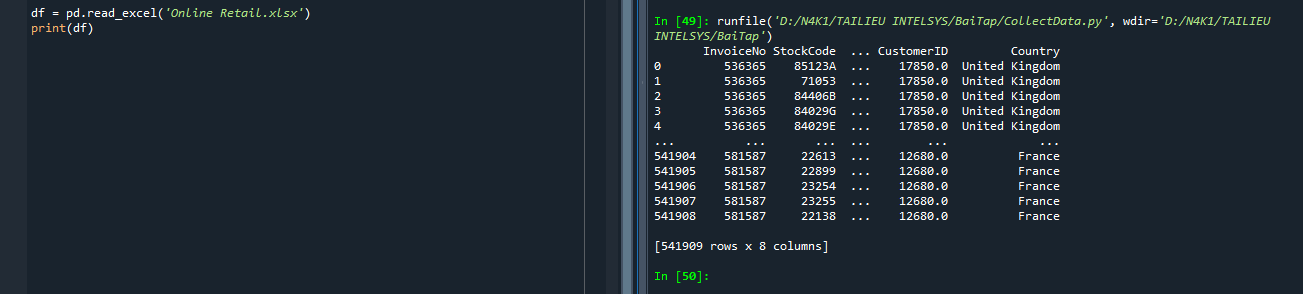
* File json



Kết quả :



* File exel

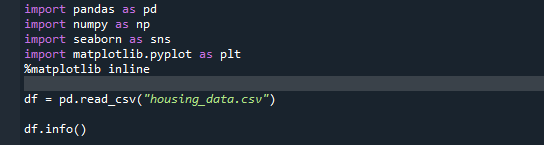


**DATA CLEANING**

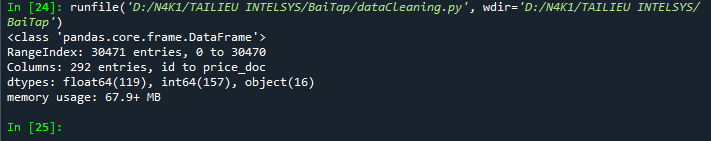
* Data Cleaning (Làm sạch dữ liệu) là quá trình thay đổi hoặc loại bỏ dữ liệu không chính xác, trùng lặp, bị hỏng hoặc không đầy đủ bên trong dataset.
* Data Cleaning không chỉ đơn thuần là loại bỏ dữ liệu, mà còn bao gồm sửa lỗi cú pháp và chính tả, sửa đổi các lỗi như thiếu mã, trường trống, xác định các điểm dữ liệu trùng lặp và chuẩn hóa tập dữ liệu.

1. Các bước cleaning data

* Để tiến hành cleaning data trước hết ta sẽ load dữ liệu từ file csv vào DataFrame sử dụng hàm read\_csv()

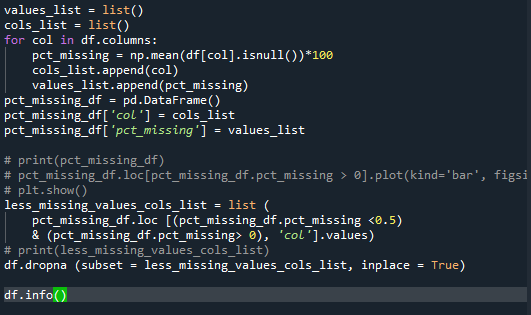


Kết quả

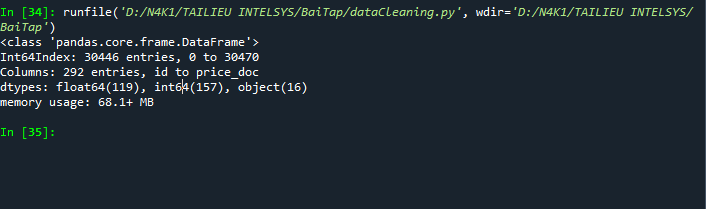


1. Xử lí dữ liệu bị thiếu

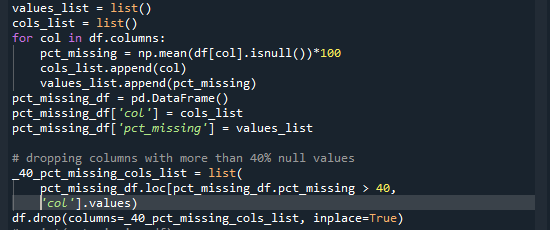
* Xoá hàng
* Một trong các phương pháp phổ biến để xử lí các dữ liệu bị thiếu là xoá bỏ hàng chứa dữ liệu đó
* Phương pháp này thường được sử dụng với tập dữ liệu lớn và việc loại bỏ một vài hàng sẽ không ảnh hưởng lớn đến kết quả.
* Nếu tập dữ liệu quá nhỏ sẽ dẫn đến mất mát dữ liệu ở các cột khác



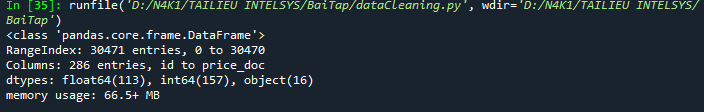
Kết quả :



* Xoá cột
* Trong trường hợp một cột chứa quá nhiều dữ liệu bị thiếu thì ta có thể tính đên việc loại bỏ cột đó gia khỏi dữ liệu nếu nó không quá ảnh hưởng đến kết quả.



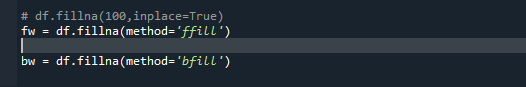
Kết quả:



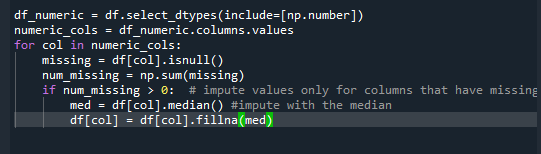
* Thay thế các giá trị bị thiếu
* Ngoài việc xoá bỏ các giá trị thiếu ta còn có thể thay thế nó bằng một giá trị khác
* Cách này giúp ta không phải xoá bỏ các hàng chỉ vì một số giá trị thiếu.
* Một số kiểu thay thế
* Thay thế bằng một giá trị ngẫu nhiên



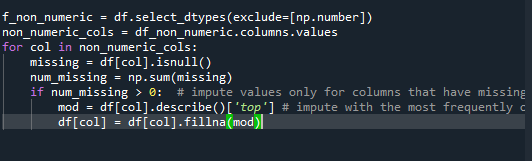
* Các giá trị bị thiếu có thể được thay thế bằng các giá trị trước (pad/ffill) hoặc sau (bfill/backfill) chúng



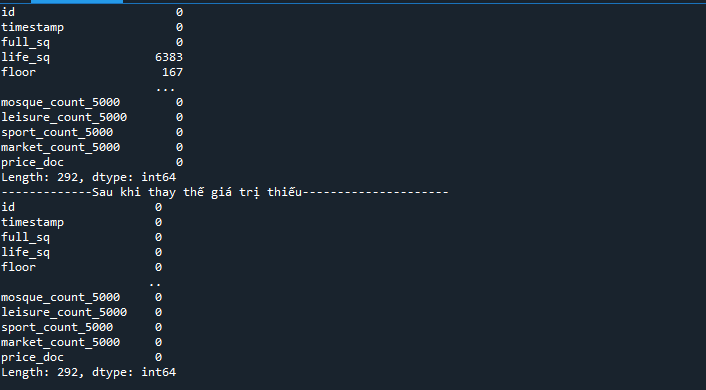
* Thay thế bằng giá trị trung vị của cột



* Thay thế bằng giá trị xuất hiện nhiều nhất trong cột



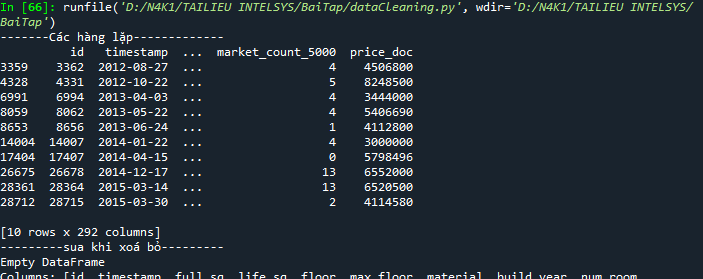
Kết quả :



1. Xoá bỏ các hàng trùng lặp

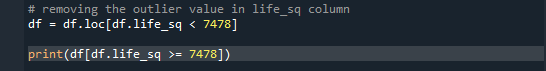
* Các hàng trùng lặp là các hàng chứa các giá trị giống nhau. Chúng thường không cung cấp thêm thông tin gì do đó ta cần loại bỏ chúng khỏi tập dứ liệu.  
  

Kết quả :

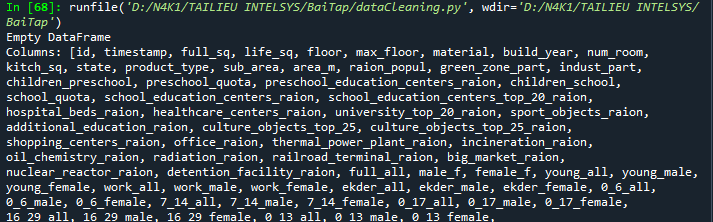


1. Outliers (dữ liệu ngoại lai)

* Là những điểm dữ liệu khác xa so với phần còn lại của dữ liệu.
* Các yếu tố ngoại lai có thể ảnh hưởng đáng kể đến hiệu suất của mô hình Học máy. Do đó, điều quan trọng là xác định những điểm bất thường và xử lí chúng.

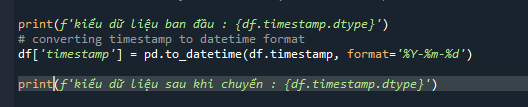


Kết quả :

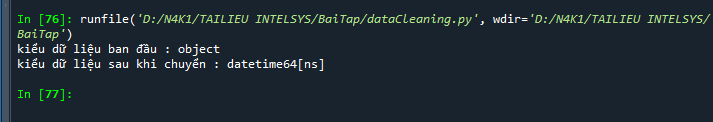


1. Xử lí kiểu dữ liệu

* Thường trong tập dữ liệu, các giá trị không được lưu trữ theo kiểu dữ liệu chính xác. Điều này có thể tạo ra sự cố trong các giai đoạn sau và có thể cho ra kết quả không mong muốn hoặc có thể gặp lỗi trong khi thực hiện. Do đó ta cần phải chuyển các kiểu dữ liệu về đúng kiểu dữ liệu mong muốn.

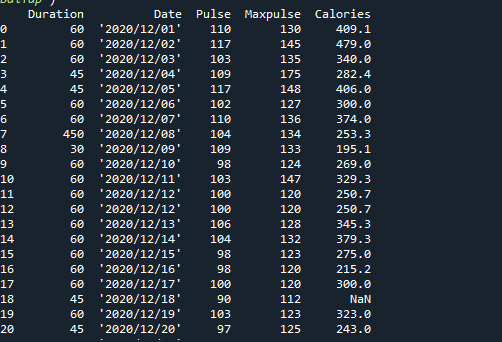


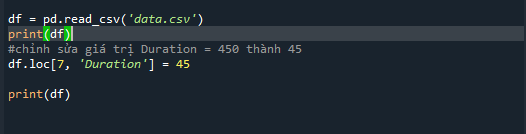
Kết quả :



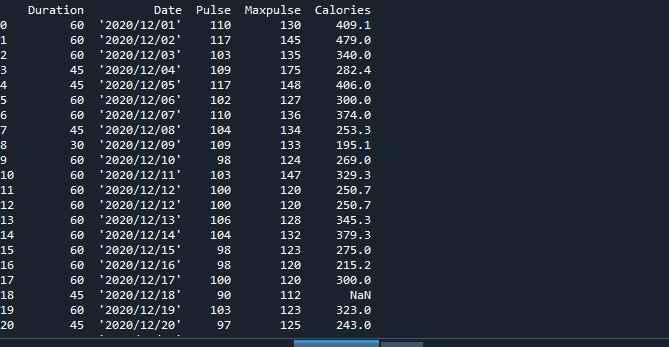
1. Xử lí dữ liệu sai

* Dữ liệu sai là dữ liệu được lưu trữ không chính xác. Vd tuổi một người là có giá trị là “1.22”.



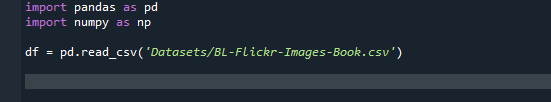


Kết quả :

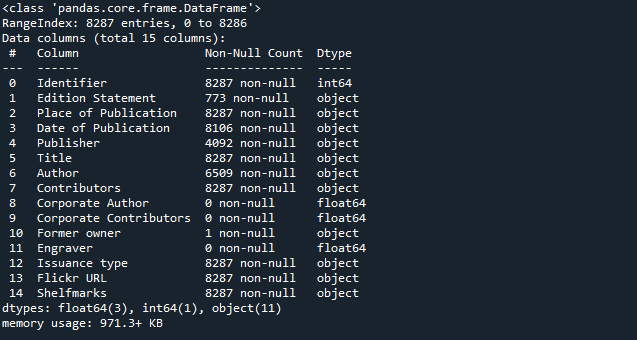


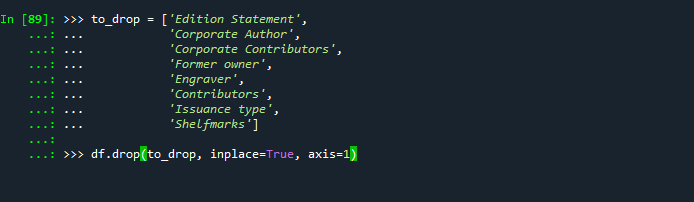
1. Bài Tập

* Load dữ liệu

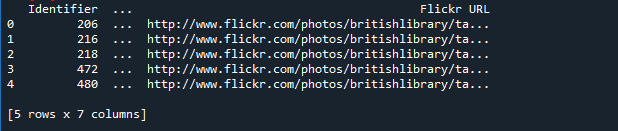


Kết quả :

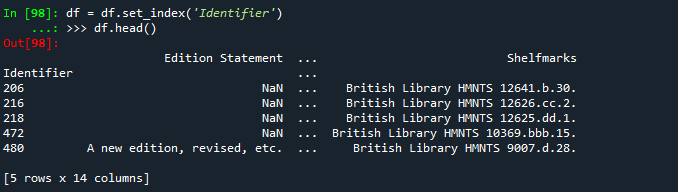


* Xoá cột

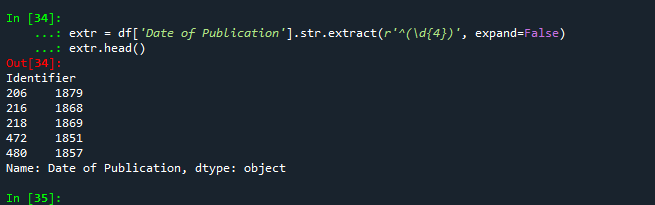
Kết quả:



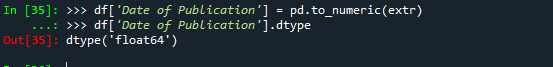
* Thay đổi chỉ mục

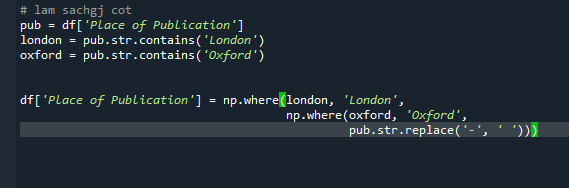


* Xử lí trường dữ liệu

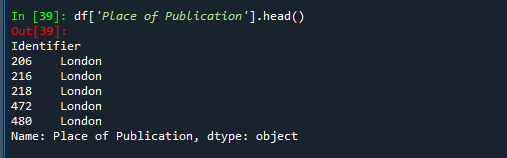


Kết quả





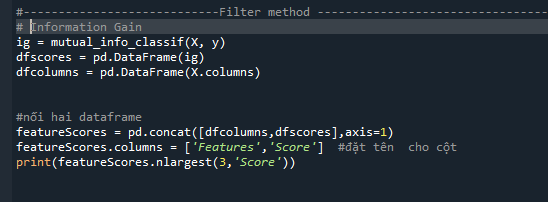
Kết quả :



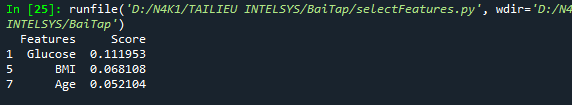
**SELECT FEATURES**

1. Present techniques for selecting features

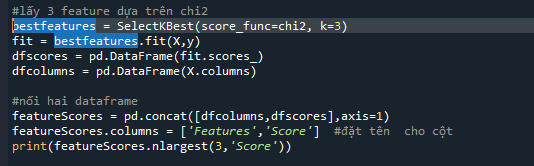
* Filter methods
* Đây là nhóm các phương pháp lựa chọn feature dựa trên các chỉ số thống kê, tương quan… giữa các feature với label.
* Các phương pháp này thường được thực hiện trong quá trình xử lí dữ liệu và gần như không phụ thuộc vào thuật toán học.
* Các phương pháp này cần thời gian tính toán thấp và ít tốn kém.
* Các phương pháp thuộc nhóm Filter methods:
* Information Gain: IG được tính dựa trên sự giảm của hàm Entropy khi tập dữ liệu được phân chia trên một thuộc tính. Phương pháp này lựa chọn các thuộc tính dựa trên giá trị của IG



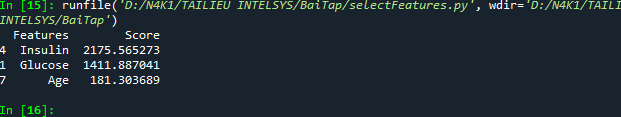
Kết quả :



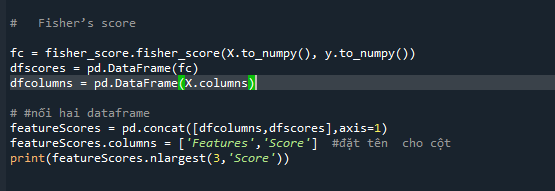
* Chi-square Test: Phương pháp này lựa chọn dựa trên các feature có điểm chi-square tốt nhất



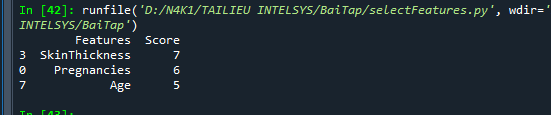
Kết quả:



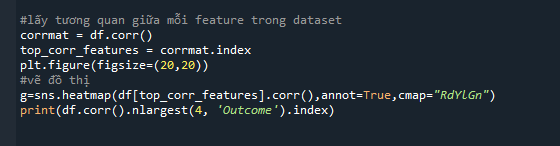
* Fisher’s Score : phương pháp này lựa chọn các tính năng dựa trên điểm được tính theo tiêu chí Fisher.



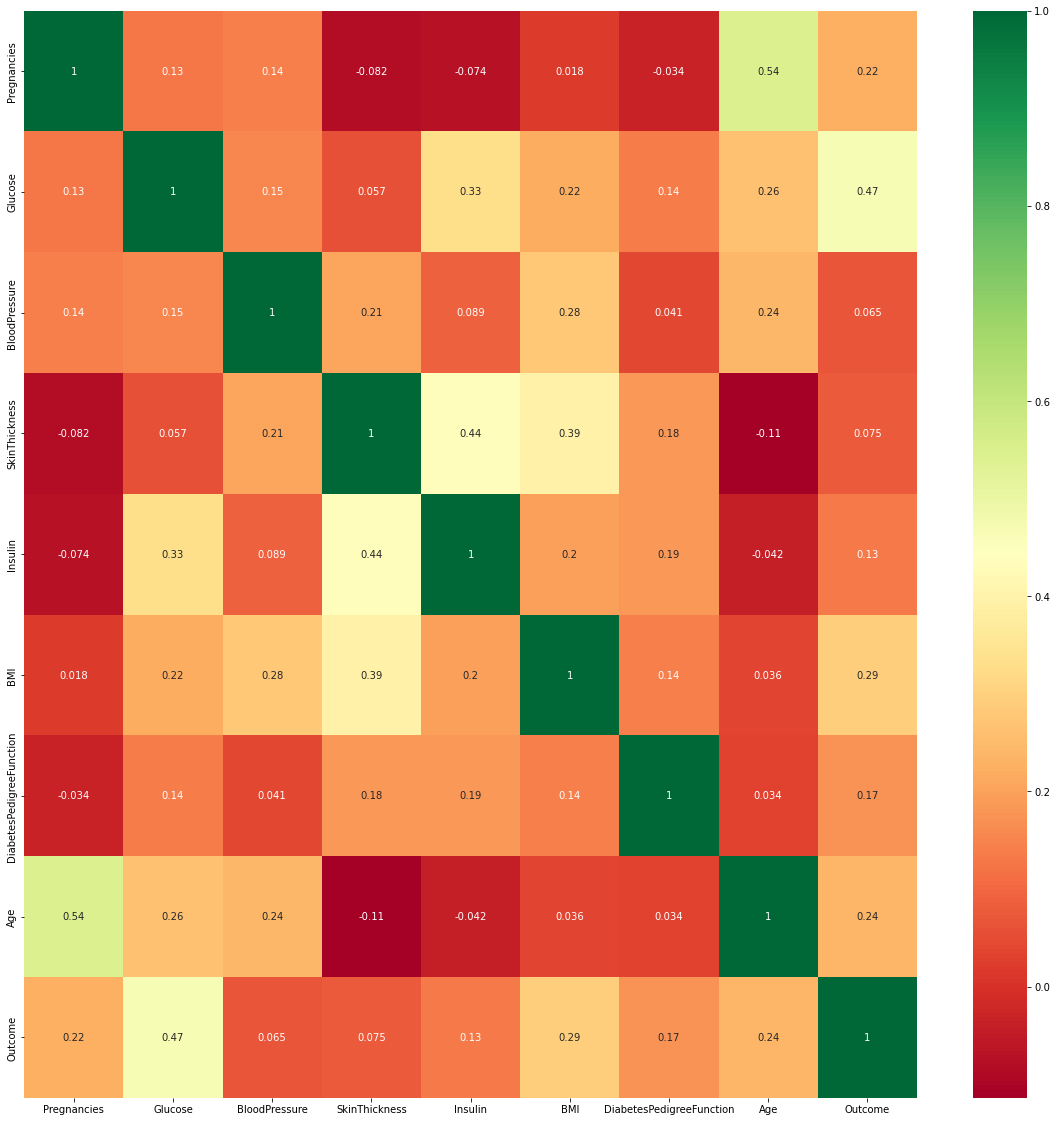
Kết quả:

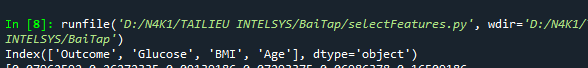


* Correlation Coefficient: Tương quan là thước đo mối quan hệ tuyến tính của 2 hoặc nhiều biến. Thông qua mối tương quan, chúng ta có thể dự đoán biến này so với biến khác. Logic đằng sau việc sử dụng tương quan để lựa chọn feature là các feature tốt có tương quan cao với label.

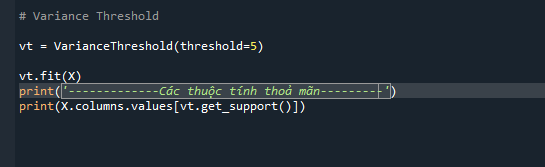


Kết quả :

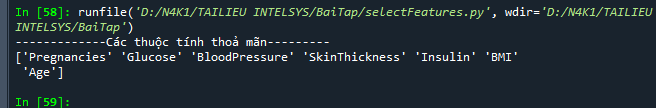




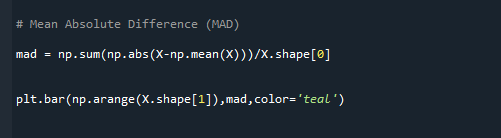
* Variance Threshold: Ngưỡng phương sai là một cách tiếp cận cơ sở đơn giản để lựa chọn feature. Nó loại bỏ tất cả các feature mà phương sai không đáp ứng một số ngưỡng. Theo mặc định, nó loại bỏ tất cả các đối tượng không có phương sai, tức là các đối tượng có cùng giá trị trong tất cả các mẫu.



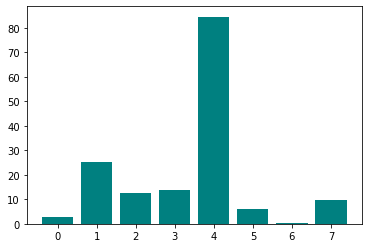
Kết quả :



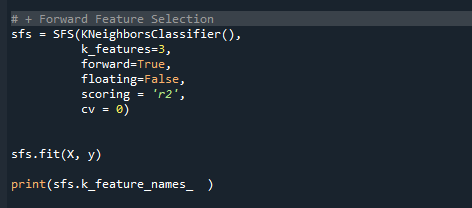
* Mean Absolute Difference (MAD) : phương pháp này tư tương tự như phương pháp Variance Threshold nhưng sự khác biệt là không có bình phương trong MAD. Phương pháp này tính toán chênh lệch tuyệt đối trung bình từ giá trị trung bình.



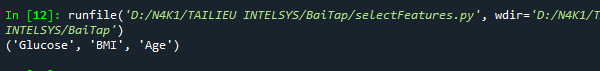
Kết quả :



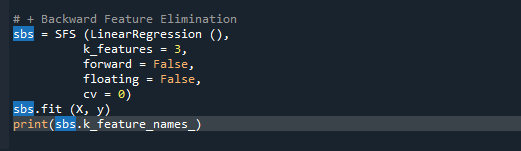
* Wrapper Methods
* Đây là các phương pháp lựa chọn feature dựa trên một thuật toán học máy cụ thể.
* Nó tuân theo một cách tiếp cận tìm kiếm tham lam bằng cách đánh giá tất cả các kết hợp có thể có của các feature so với tiêu chí đánh giá
* Trong phương pháp này ta sẽ lựa chọn một tập con feature và sử dụng nó để huấn luyện cho mô hình. Sau đó dựa vào kết quả của mô hình ta sẽ thêm hoặc loại bỏ các feature sau đó tiếp tục huấn luyện bằng tập feature mới.
* Quá trính này lập đi lập lại liên tục cho đến khi thoả mãn các tiêu chí định ra từ trước. Các tiêu chí này phụ thuộc vào việc theo dõi xem hiệu suất không tăng hoặc giảm vượt quá một ngưỡng nhất định hay không, tùy thuộc vào phương pháp đang sử dụng.
* Một số phương pháp tiêu biểu :
* Forward Feature Selection : ở phương pháp này ban đầu chúng ta bắt đầu với một tập hợp các tính năng trống và tiếp tục thêm một tính năng để cải thiện tốt nhất mô hình sau mỗi lần lặp lại. Tiêu chí dừng là cho đến khi việc bổ sung một biến mới không cải thiện hiệu suất của mô hình



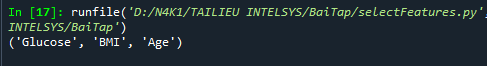
Kết quả



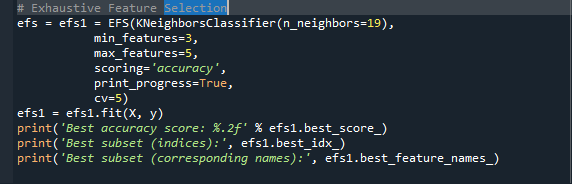
* Backward Feature Elimination : phương pháp này tương tự với phương pháp Forward Feature Selection nhương khác ở điểm ta sẽ bắt đầu với tập còn là tập tất cả các feature ban đầu. Sau đó sẽ loại bỏ dần các feature không quan trọng dựa trên kết quả huấn luyện. Tiêu chí dừng là cho đến khi không quan sát thấy sự cải thiện nào về hiệu suất của mô hình sau khi tính năng bị loại bỏthiện thêm nữa.



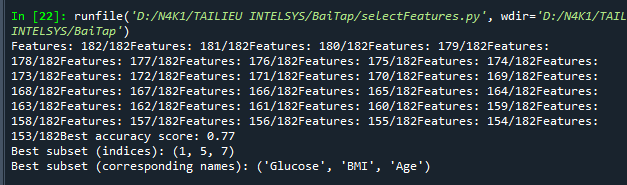
Kết quả :



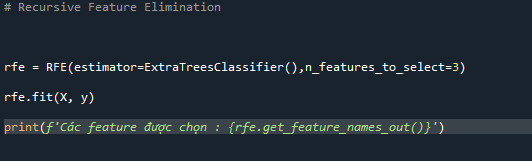
* Exhaustive Feature Selection: Đây là phương pháp lựa chọn tính năng mạnh mẽ nhất được đề cập cho đến nay. Nó thử mọi sự kết hợp có thể có của các biến và trả về tập hợp con hoạt động tốt nhất.



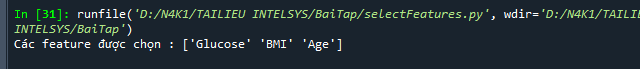
Kết quả :



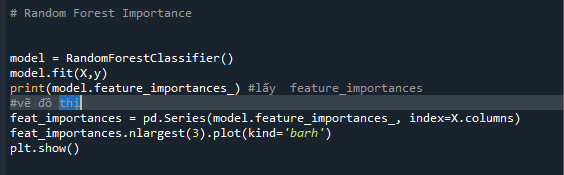
* RFE xếp hạng các tính năng dựa trên mức độ quan trọng của chúng và trả về các tính năng hàng đầu sau khi loại bỏ các tính năng ít quan trọng nhất, trong đó n là do người dùng đưa ra.



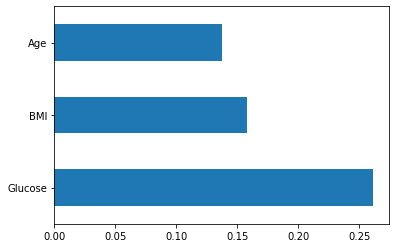
Kết quả :



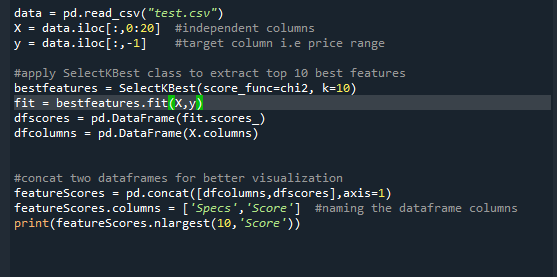
* Embedded Methods :
* Trong phương pháp này thuật toán lựa chọn feature được kết hợp như một phần của thuật toán học. Các kỹ thuật này tận dụng chính các mô hình Học máy để xếp hạng và cho điểm các feature dựa trên tầm quan trọng của chúng. Các phương pháp này nhanh hơn và chính xác hơn các Filter methods và cũng có tính đến sự kết hợp của các tính năng.
* Các phương pháp phổ biến :
* Tree-based methods: Các phương pháp này như Random Forest, Gradient Boosting cung cấp cho ta tầm quan trọng của tính năng cũng như một cách để chọn các tính năng. Mức độ quan trọng của tính năng cho chúng ta biết tính năng nào quan trọng hơn trong việc tạo tác động đến tính năng mục tiêu.



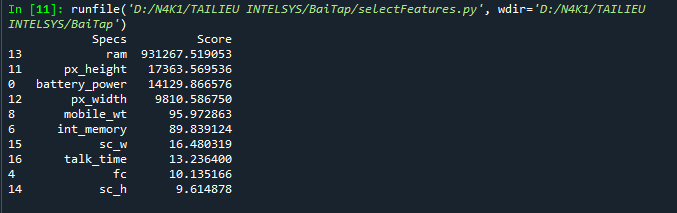
Kết quả:



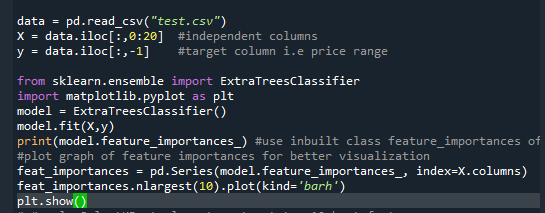
1. Execute 3 techniques you present
2. Univariate Selection



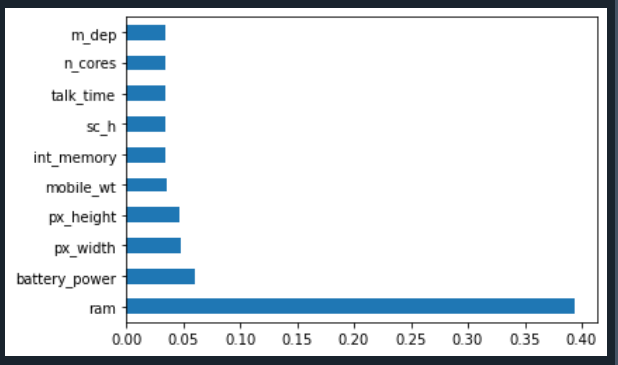
Kết quả:



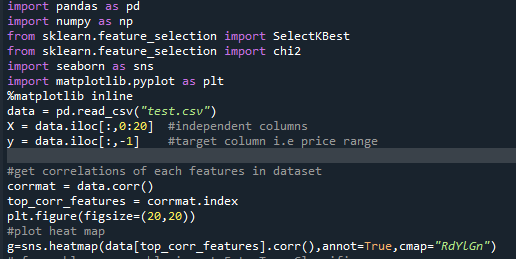
1. Feature Importance



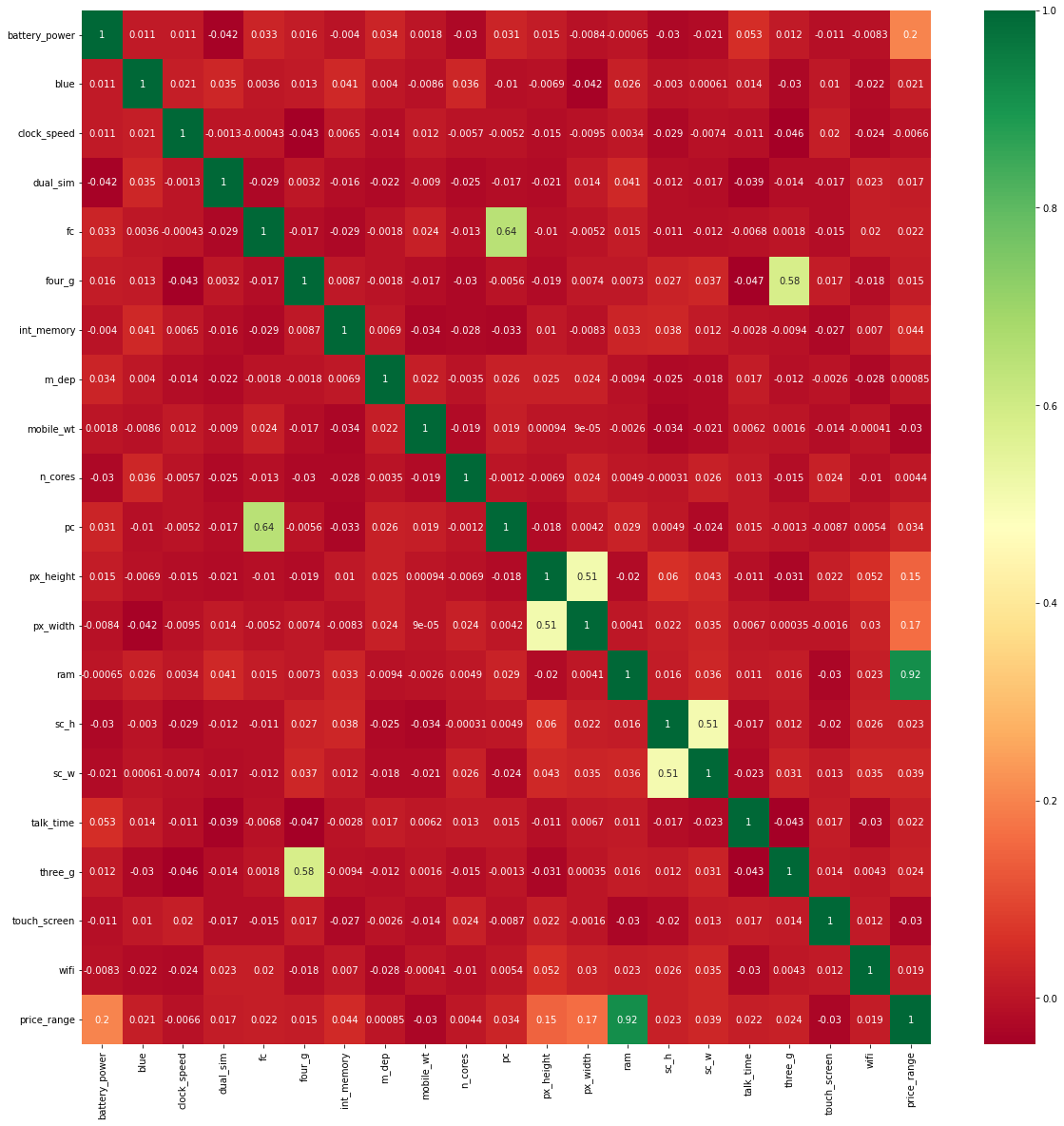
Kết quả :



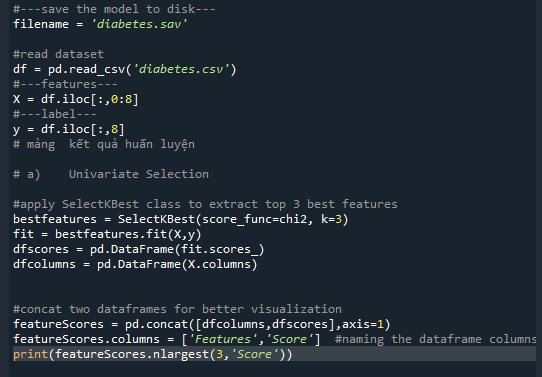
1. Correlation Matrix with Heatmap



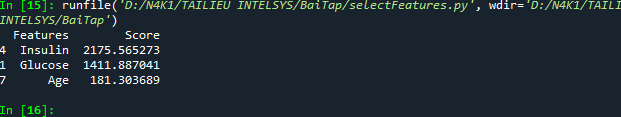
Kết quả



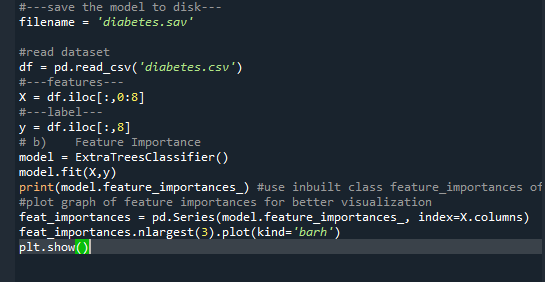
1. Present 3 techniques with data from Chap 12
2. Univariate Selection



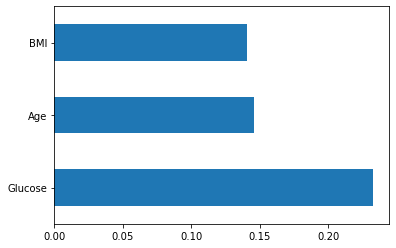
Kết quả:



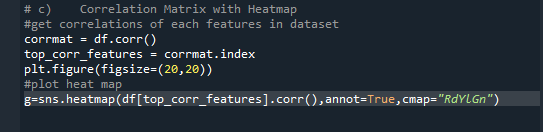
1. Feature Importance



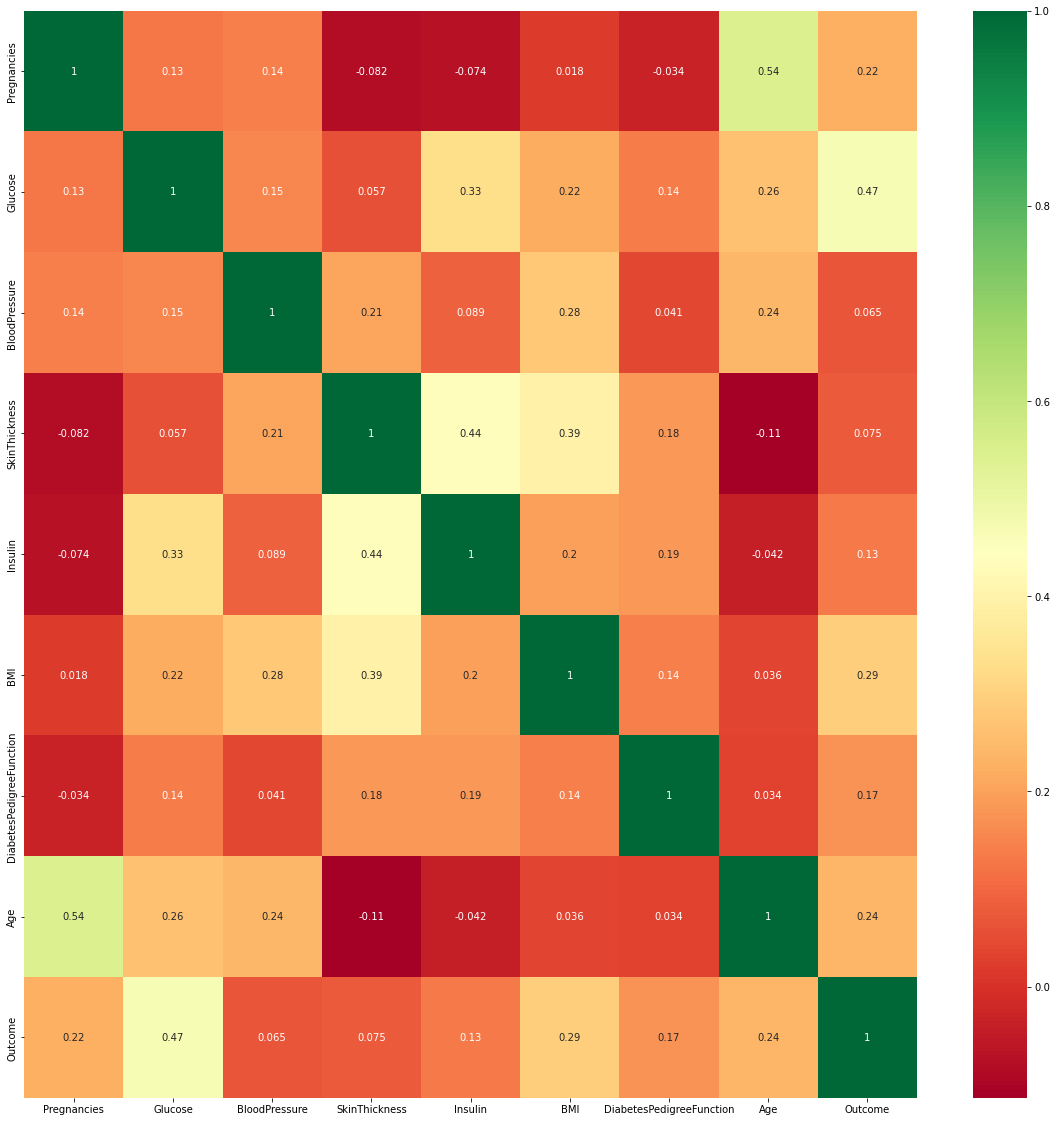
Kết quả :



1. Correlation Matrix with Heatmap



Kết quả



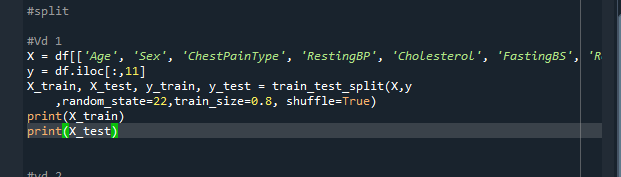
**SPLIT DATA**

* Train - Test Split: là một kỹ thuật để đánh giá hiệu suất của một thuật toán học máy. Nó có thể được sử dụng cho các vấn đề phân loại hoặc hồi quy và có thể được sử dụng cho bất kỳ thuật toán học có giám sát nào.
* Quy trình này liên quan đến việc lấy một tập dữ liệu và chia nó thành hai tập hợp con. Tập hợp con đầu tiên được sử dụng để phù hợp với mô hình và được gọi là tập dữ liệu đào tạo. Tập hợp con thứ hai không được sử dụng để đào tạo mô hình; thay vào đó, phần tử đầu vào của tập dữ liệu được cung cấp cho mô hình, sau đó các dự đoán được đưa ra và so sánh với các giá trị dự kiến. Tập dữ liệu thứ hai này được gọi là tập dữ liệu thử nghiệm.

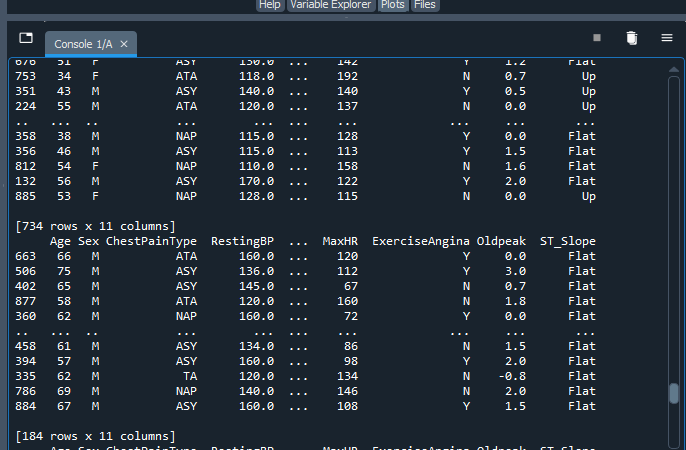
+ Train Dataset: Được sử dụng để huấn luyện mô hình machine learning.

+ Test Dataset: Dùng để đánh giá mô hiệu suất của mô hình.

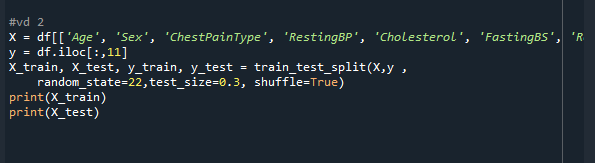
* Tỷ lệ giữa Train-Test là ko có tối ưu.
* Bạn phải chọn tỷ lệ phần trăm phân tách đáp ứng các mục tiêu của dự án với những cân nhắc bao gồm:
* Chi phí tính toán trong đào tạo mô hình.
* Chi phí tính toán trong việc đánh giá mô hình.
* Mức độ chính xác dựa trên mẫu training.
  + Mức độ chính xác dựa trên mẫu test.
* Tuy nhiên, tỷ lệ phần trăm phân chia phổ biến bao gồm:
  + Train: 80%, Test: 20%
  + Train: 67%, Test: 33%
  + Train: 50%, Test: 50%

****

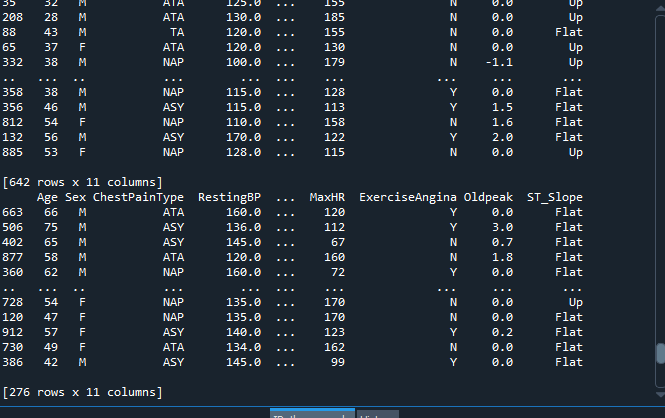
**Kếtquả :**

****

**Vd2 :**

****

**Kết quả :**

****