Trong bài báo cáo này thì em sẽ trình bày tổng quát về các dạng của ML thông qua các bài code đặc trưng về từng loại:

* Supervised learning (Học có giám sát)

+ Regression (Hồi quy)

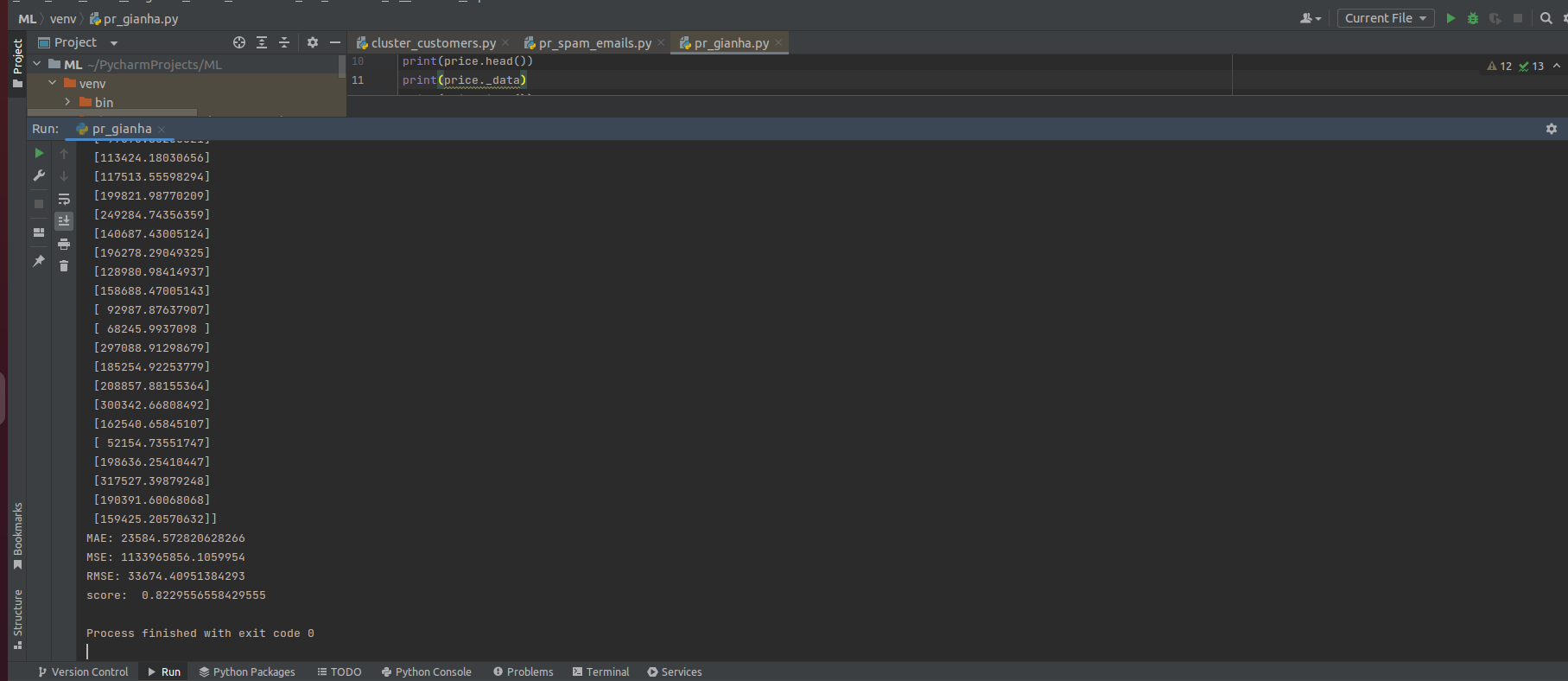
+ Classified (Phân loại)

* Unsupervised learning (Học không giám sát)

***Dự đoán giá nhà (Regression):***

Chart, histogram

Description automatically generated



Score khoảng 0.82, điểm vẫn còn có thể cao hơn nếu fix xong missing\_value, trong bài này thì em chưa tiền xử lí dữ liệu, em chỉ đang sử dụng các columns không bị missing\_value

Trong bài này em sử dụng LinearRegression để dự đoán giá nhà, sử dụng seaborn để vẽ biểu đồ để xem 1 cách tổng quát.

* Saleprice dùng để thể hiện giá nhà
* Count là số lượng nhà có mức giá như thế

***Phân loại emails (Phân loại)***

Text

Description automatically generated with medium confidence

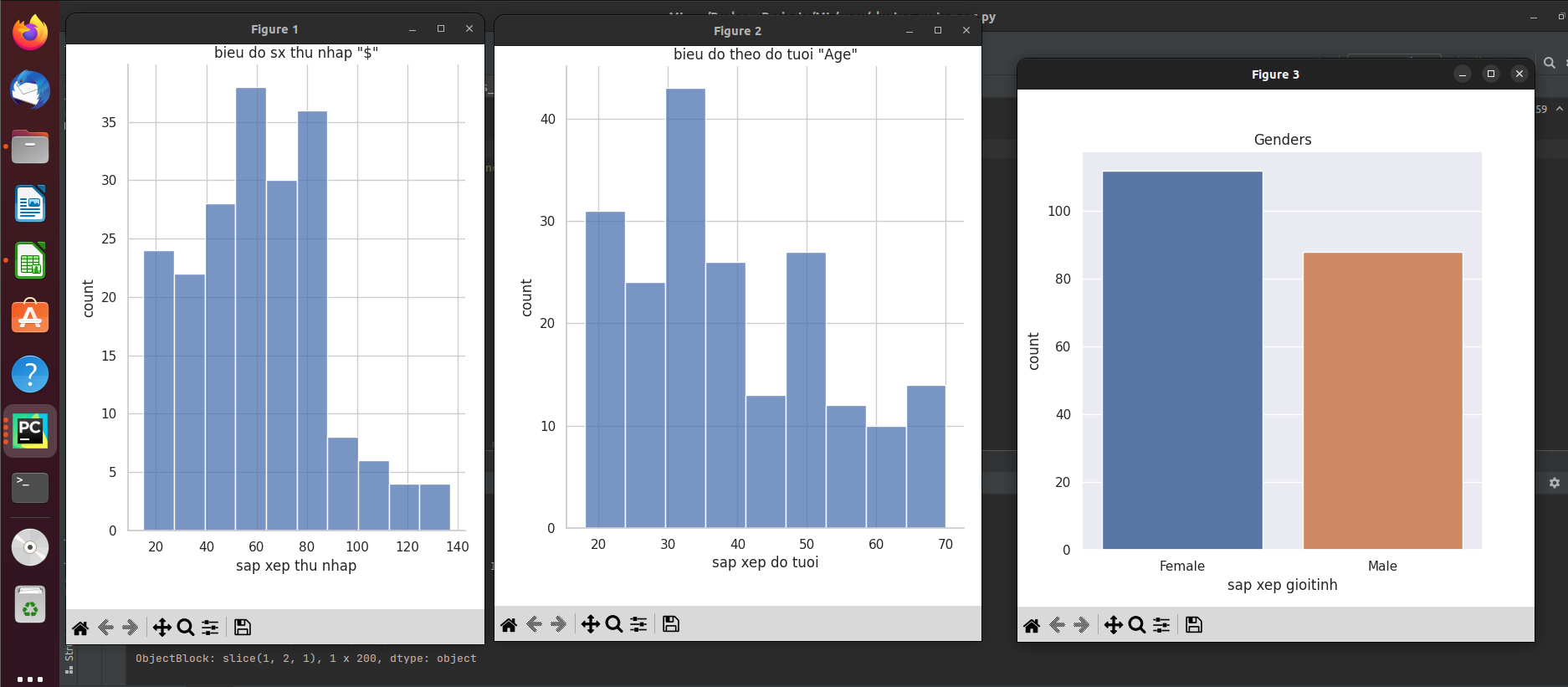
***Ý tưởng bài toán:*** đầu tiên em sẽ mở file lên để xem các title của các cột tiện cho việc phân loại, text sẽ là đầu vào còn label\_num sẽ là đầu ra, sau đó sẽ đến bước chuẩn hóa dữ liệu

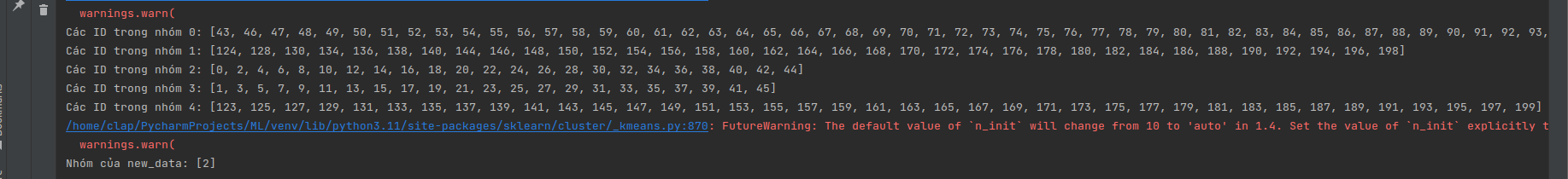
Em sử dụng CountVectorizer, 1 hàm khá phổ biến để chuẩn hóa ngôn ngữ. sau đó cũng như các bài toán học có giám sát khác thì em cũng sẽ chia dữ liệu thành train và test, 80% train và 20% test.

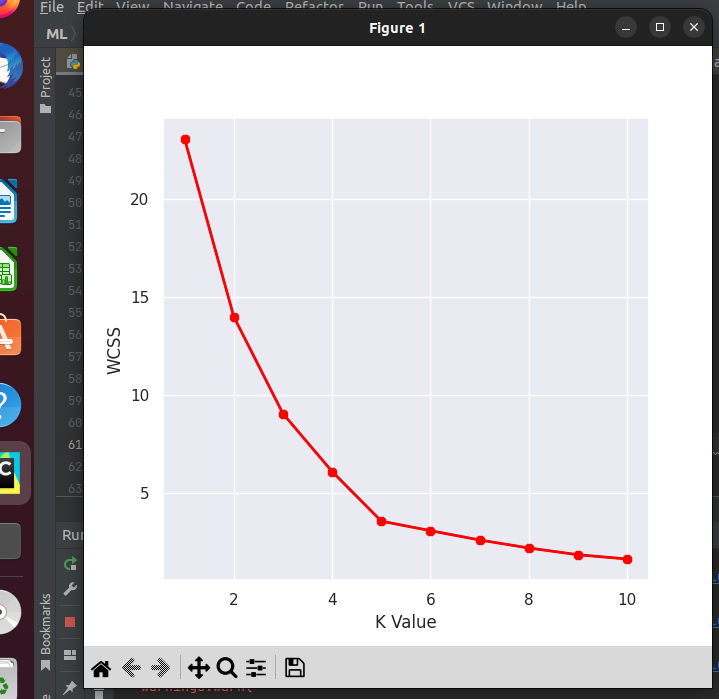
Tiếp theo thì em import các models vào để train và so sánh độ chính xác của các models

Trong mô hình này thì em sử dụng 1 số mô hình như SVC, navie\_bayes,… để kiểm tra xem độ chính xác giữa các mô hình, thì có thể thấy rõ rằng mô hình navies\_bayes hoạt động rất tốt trong bài toán này, cụ thể là MultinomialNB , và thứ tự độ chính xác của từng mô hình được sắp theo hình trên và thứ tự này khá chính xác bởi vì em sử dụng phương pháp cross\_val\_score để đánh giá mô hình và em chia thành 5 fold sau đó lấy trung bình của 5 fold đó lại thì nó sẽ ra list như trên, Tạo 1 cái Hàm để gọi ra các models khi chạy.

***Phân loại khách hàng dựa vào thu nhập (Học không giám sát)***

******





Chart, scatter chart

Description automatically generated

***Ý tưởng bài toán:*** vì đây là bài toán phân cụm nên cần tập trung phân cụm dữ liệu dựa trên biểu đồ, như mọi bài toán thì việc đầu tiên cần làm là tiền xử lý dữ liệu và chuẩn hóa dữ liệu.chuyển genders thành 0 và 1, sau đó dùng MinMaxScaler() để chuẩn hóa.

Vì bài toán này dùng kmeans để phân cụm nên trước khi phân cụm được thì ta sẽ sử dụng biểu đồ elbow để chọn k, nhìn biểu đồ elbow(line màu đỏ) ta thấy dữ liệu bắt đầu giảm ngang ngang nhau (giảm không sâu) tại k = 5, vì thế nên ta chọn k =5

Khi phân cụm thì có thểthấy dữ liệu đã gom khá tốt về 5 cụm tương ứng, tuy nhiên chỉ với biểu đồ này thì vẫn chưa biết được ID nào thuộc về cụm nào, vì vậy chúng ta phải in ID trong từng cụm

Tiếp theo em khởi tạo 1 new data và tiến hành dự đoán xem data này thuộc cụm nào?

silhouette\_score\_prediction là 1 trong những hàm phổ biến để đánh giá mô hình trong kmeans, giá trị của loại score này từ -1 đến 1, và càng gần 1 thì càng tốt. Em chỉ sử dụng 2/5 cột để train thì kết quả đạt được là khoảng 0.56 , có thể khi sử dụng 4/5 cột thì kết quả sẽ khác :”<