Báo cáo
Đồ án Thực hành

Môn học: Nhập môn Khoa học dữ liệu

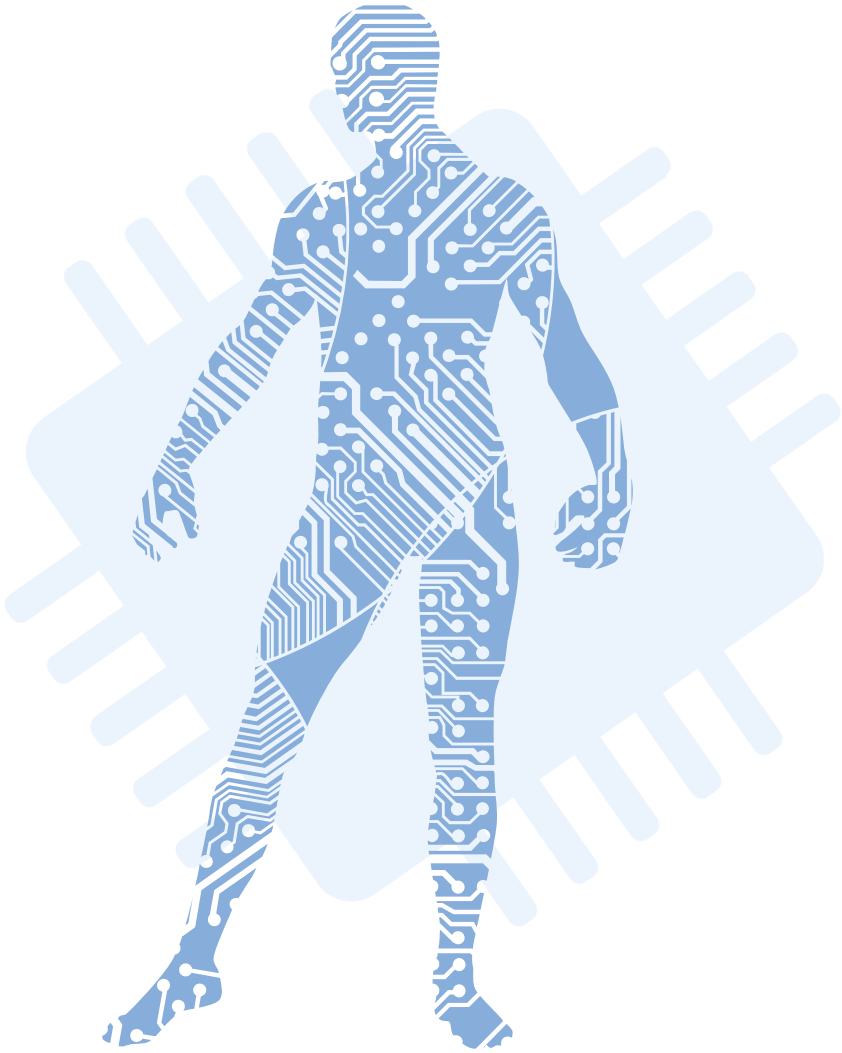
GVHD:

- Trần Đại Chí
- Nguyễn Bảo Long
- Lê Nhựt Nam
- Nguyễn Thái Vũ

Dân số Thế giới

(Thực hiện quy trình khoa học dữ liệu và mô hình hóa dữ liệu với bộ dữ liệu liên quan đến dân số thế giới)

Nội dung

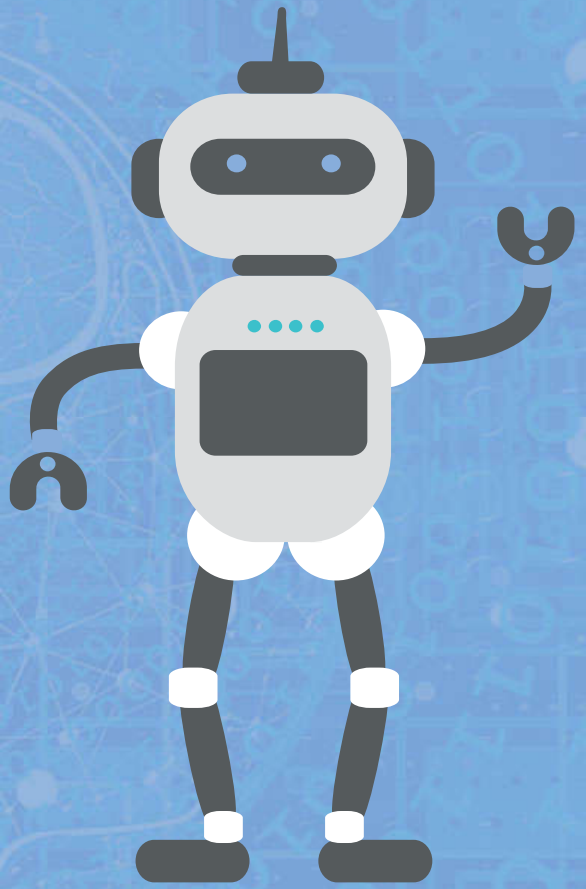


NHÓM 5

Các thành viên

- Nguyễn Việt Khoa - 20120120
- Nguyễn Đức Minh Quân - 20120357
- Nguyễn Hoàng Việt - 20120402
- Nguyễn Tấn Sơn - 20120565

Giới thiệu

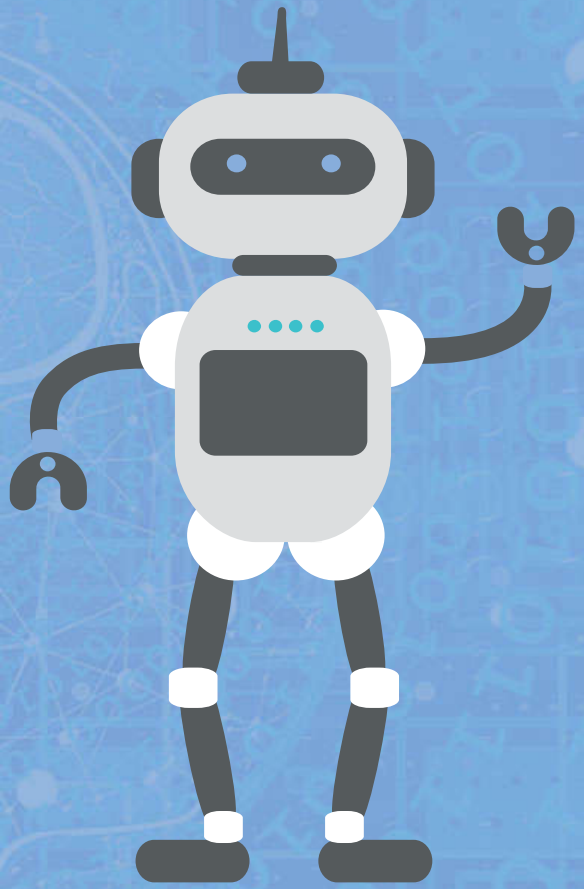


Giới thiệu

Đồ án lần này được thực hiện để tổng hợp lại tất cả kiến thức và kĩ năng đã được học ở môn Nhập môn Khoa học dữ liệu giúp cho sinh viên có thể ôn tập và cũng cố kiến thức cũng như nâng cao khả năng của bản thân.

Tập dữ liệu của đồ án lần này được lấy từ trang web https://www.worldometers.info/world-population/population-by-country.

Tập dữ liệu được thu thập bằng cách sử dụng thư viện scrapy để cào dữ liệu tại trang web trên sau đó lưu vào một file csv.



Khám phá và tiền xử lý dữ liệu

Khám phá và tiền xử lý

Khám phá dữ liệu

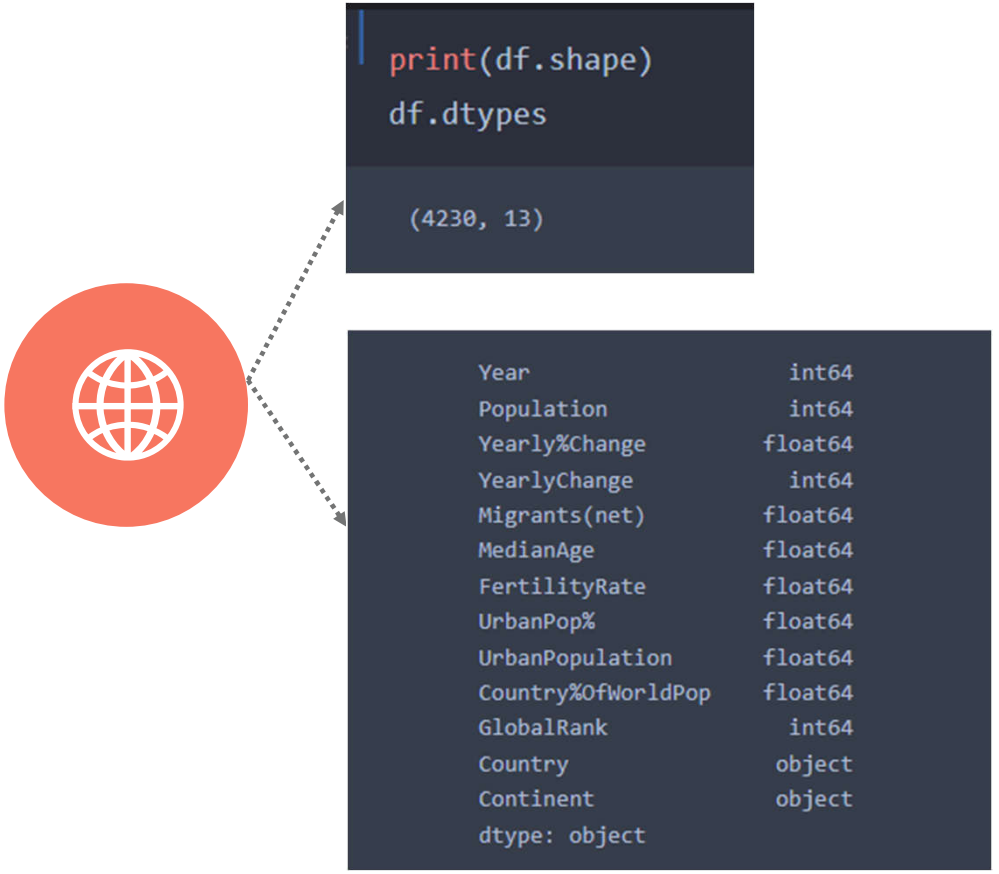
Dữ liệu có 4230 dòng và 13 cột.

Mỗi dòng ở đây thể hiện cho các đặc điểm liên quan đến dân số của một quốc gia ở một năm cụ thể.

Các cột thể hiện những mô tả về đặc điểm liên quan đến dân số:

- Country: tên quốc gia
- Year: năm thực hiện thống kê
- Population: dân số
- Yearly%Change: tỉ lệ gia tăng dân số
- YearlyChange: số dân thay đổi hằng năm
- Migrants(net): số người di cư đến
- MedianAge: độ tuổi trung vị
- FertilityRate: tỉ suất sinh
- UrbanPop%: tỉ lệ dân thành thị
- UrbanPopulation: số dân thành thị
- %OfWorldPop: tỉ lệ % so với dân số thế giới
- GlobalRank: xếp hạng về dân số
- Continent: châu lục mà quốc gia đó trực thuộc

Các kiểu dữ liệu trong tập này vừa có dạng số vừa có dạng object (thực chất là dạng chuỗi), một số chỗ là NaN nên sẽ được gán là kiểu float trong pandas.



```
print(df.shape)
df.dtypes
```

```
(4230, 13)
```

```
Year                    int64
Population              int64
Yearly%Change         float64
YearlyChange            int64
Migrants(net)         float64
MedianAge             float64
FertilityRate         float64
UrbanPop%             float64
UrbanPopulation       float64
Country%OfWorldPop    float64
GlobalRank              int64
Country                object
Continent              object
dtype: object
```

Khám phá và tiền xử lý

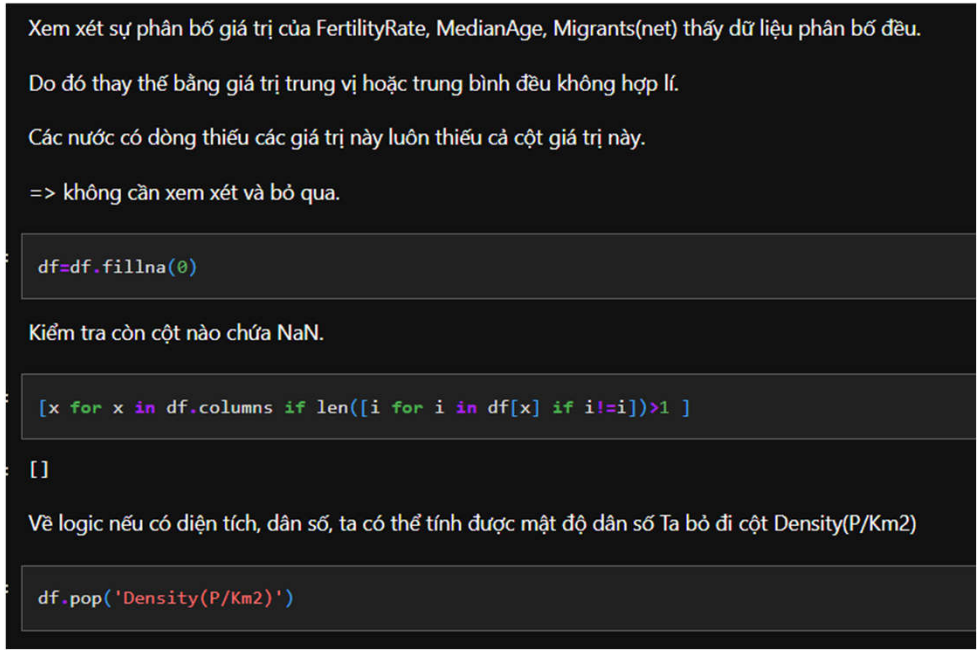
Tiền xử lý dữ liệu

Chúng ta có 2 vấn đề cần xử lí với kiểu dữ liệu này:

- Tìm cách thay thế các NaN hoặc giá trị thiếu ở các cột dạng số.
- Bỏ đi một số cột có thể suy ra được từ dữ liệu của các cột khác.

Dùng hồi quy để thay thế NaN, dựa trên cột UrbanPop% (vì những dữ liệu còn lại của FertilityRate, MedianAge, Migrants(net) phân bố khá đều và độ chênh lệch không cao, do đó thay thế bằng giá trị trung vị hoặc trung bình đều không hợp lí). Bên cạnh đó, các nước có dòng thiếu các giá trị ở FertilityRate, MedianAge, Migrants(net) luôn thiếu cả cột giá trị này (tức là năm nào cũng thiếu) nên ta sẽ không thực hiện hồi quy.

Về logic nếu có diện tích, dân số, ta có thể tính được mật độ dân số nên ta hoàn toàn có thể bỏ đi cột Density(P/Km2). Sau đó lưu dữ liệu đã được xử lí lại vào file csv.



Đặt câu hỏi
và trả lời câu hỏi

Đặt câu hỏi và trả lời

Câu hỏi 1

Làm thế nào để dễ dàng tra cứu thông tin dân số theo từng nước/lục địa qua các năm, tỉ lệ phần trăm dân số của từng nước/lục địa so với thế giới (qua các năm)?

Tiền xử lý để trả lời câu hỏi: Do thông tin cần trình bày có rất nhiều đặc điểm cần thể hiện (dân số của các nước/lục địa, sự thay đổi theo thời gian,…) nên nhóm đã trực quan dữ liệu bằng interactive plot trong thư viện plotly (biểu đồ tương tác, có thể thay đổi dữ liệu hiển thị theo yêu cầu của người dùng).

Trả lời: Bằng cách vẽ các biểu đồ.

Ý nghĩa của câu hỏi: Giúp việc tra cứu các thông tin về dân số một cách dễ dàng hơn.

Đặt câu hỏi và trả lời

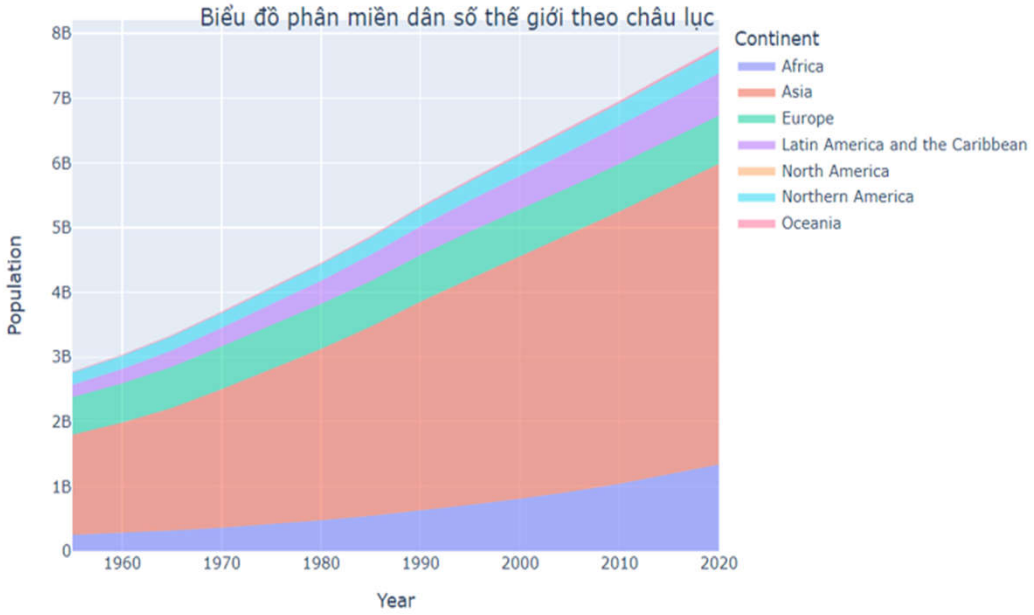
Câu hỏi 1

Làm thế nào để dễ dàng tra cứu thông tin dân số theo từng nước/lục địa qua các năm, tỉ lệ phần trăm dân số của từng nước/lục địa so với thế giới (qua các năm)?

Tiền xử lý để trả lời câu hỏi: Do thông tin cần trình bày có rất nhiều đặc điểm cần thể hiện (dân số của các nước/lục địa, sự thay đổi theo thời gian,…) nên nhóm đã trực quan dữ liệu bằng interactive plot trong thư viện plotly (biểu đồ tương tác, có thể thay đổi dữ liệu hiển thị theo yêu cầu của người dùng).

Trả lời: Bằng cách vẽ các biểu đồ.

Ý nghĩa của câu hỏi: Giúp việc tra cứu các thông tin về dân số một cách dễ dàng hơn.



Đặt câu hỏi và trả lời

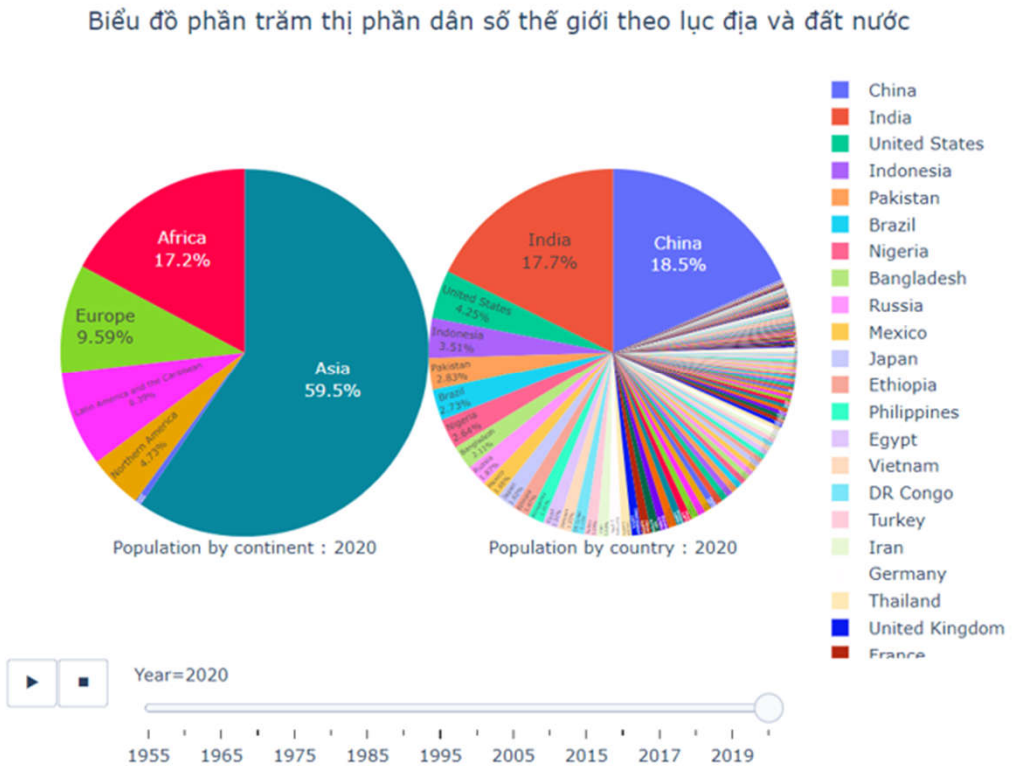
Câu hỏi 1

Làm thế nào để dễ dàng tra cứu thông tin dân số theo từng nước/lục địa qua các năm, tỉ lệ phần trăm dân số của từng nước/lục địa so với thế giới (qua các năm)?

Tiền xử lý để trả lời câu hỏi: Do thông tin cần trình bày có rất nhiều đặc điểm cần thể hiện (dân số của các nước/lục địa, sự thay đổi theo thời gian,…) nên nhóm đã trực quan dữ liệu bằng interactive plot trong thư viện plotly (biểu đồ tương tác, có thể thay đổi dữ liệu hiển thị theo yêu cầu của người dùng).

Trả lời: Bằng cách vẽ các biểu đồ

Ý nghĩa của câu hỏi: Giúp việc tra cứu các thông tin về dân số một cách dễ dàng hơn.



Đặt câu hỏi và trả lời

Câu hỏi 2

Có nước nào có tỉ lệ dân thành thị giảm gần đây không? Điều đó mang ý nghĩa như thế nào?

Tiền xử lý để trả lời câu hỏi: Chọn ra dữ liệu của 3 năm mới nhất trong df để so sánh giá trị cột UrbanPop%, nếu quốc gia nào có giá trị UrbanPop% giảm dần từ 2018 đến 2020 thì đó là nước có tỉ lệ dân thành thị giảm gần đây.

Trả lời: Vẽ biểu đồ, các quốc gia được hiển thị trên biểu đồ là đáp án của câu hỏi này.

Ý nghĩa của câu hỏi: Từ thông tin này ta có thể tìm hiểu thêm về cách thức và mục đích của việc giảm tỉ lệ dân thành thị ở các quốc gia đó. Có thể việc giảm tỉ lệ dân thành thị cũng liên quan đến việc giảm ùn tắc giao thông, tái cơ cấu nền kinh tế ở một số vùng miền để giảm sự chênh lệch về kinh tế giữa các vùng trong nước.

Đặt câu hỏi và trả lời

Câu hỏi 3

Mối tương quan giữa độ tuổi trung vị với tỉ lệ tăng dân số?

**Trả lời:* Vẽ biểu đồ, các quốc gia được hiển thị trên biểu đồ là đáp án của câu hỏi này.
Giải thích: Mỗi điểm thể hiện cho một quốc gia, độ lớn của điểm thể hiện tổng dân số, cột x và y thể hiện lần lượt tỉ lệ gia tăng dân số và độ tuổi trung vị của quốc gia đó.

**Ý nghĩa của câu hỏi:* Từ thông tin này ta có thể biết thêm mối tương quan giữa độ tuổi trung vị với tỉ lệ tăng dân số: độ tuổi trung vị càng thấp (tức dân số càng trẻ) thì tỉ lệ gia tăng dân số càng cao. Trong khi đó, các nước dân số già lại đang có xu hướng thực hiện lối sống "hưởng thụ", không quá áp đặt việc sinh con để có thể sống tự do hơn, điều đó khiến dân số của các nước này có khả năng ngày càng "già" hơn nữa.

Mô hình hóa dữ liệu

ARTIFICIAL INTELLIGENCE

Mô hình hóa dữ liệu

a. Đánh giá BestToLive (quốc gia đáng sống)

Ban đầu nhóm request html để lấy thông tin về tên của các quốc gia đáng sống (top 50 quốc gia đáng sống trên thế giới, theo thống kê từ tờ báo US News với đường dẫn https://www.usnews.com/news/best-countries/rankings/quality-of-life).
Sau đó tiến hành gán nhãn cho các quốc gia trong dataframe df.

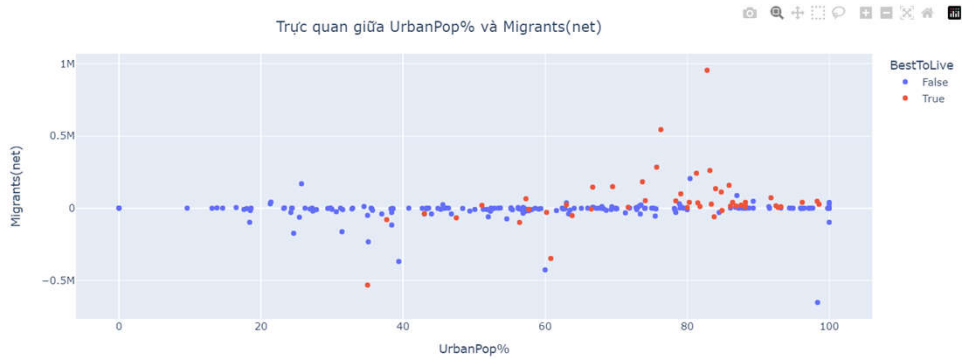
	Country	Population	Yearly%Change	MedianAge	UrbanPop%	Migrants(net)	FertilityRate	BestToLive
0	Afghanistan	38928346	2.33	18.4	25.4	-62920.0	4.56	False
1	Albania	2877797	-0.11	36.4	63.5	-14000.0	1.62	False
2	Algeria	43851044	1.85	28.5	72.9	-10000.0	3.05	False
3	American Samoa	55191	-0.22	0.0	88.1	0.0	0.00	False
4	Andorra	77265	-0.19	0.0	87.8	0.0	0.00	False

Mô hình hóa dữ liệu

a. Đánh giá BestToLive (quốc gia đáng sống)

Chọn ra dữ liệu thuộc năm 2020 để phân lớp. Nhóm lựa chọn việc phân lớp theo hai thuộc tính.

- Phân lớp dựa vào UrbanPop% và Migrants(net): Vẽ đồ thị trực quan với trục hoành là trục UrbanPop% và trục tung là Migrants(net), và điểm chấm màu đỏ là các nước được cho là “Đáng sống”.



Với tập train có cột label được xem là có đáng sống hay không? (True hoặc False), ta xây dựng mô hình **Hồi quy logistic** để đưa ra dự đoán dựa trên 2 yếu tố UrbanPop% và Migrants.

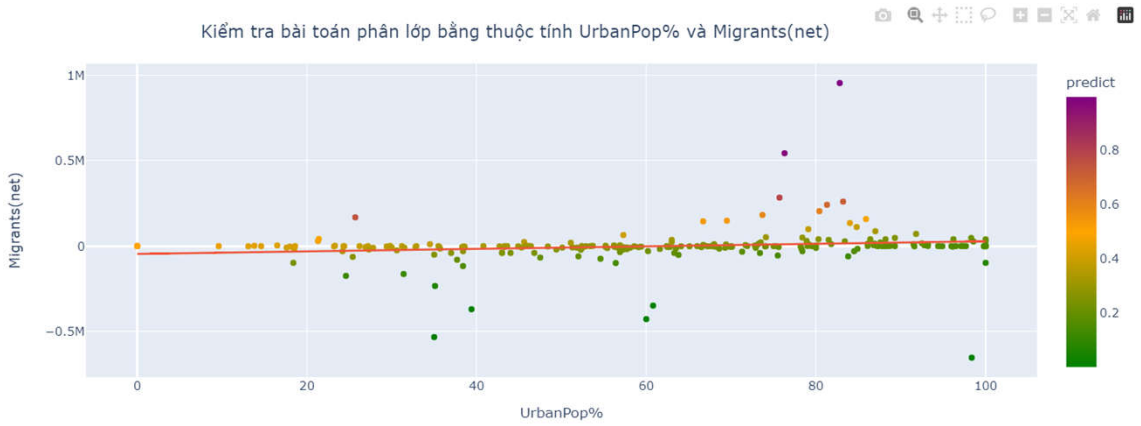
```
Binary cross entropy loss là :  6.466844936886641
Độ chính xác của mô hình :  0.8127659574468085
```

Mô hình hóa dữ liệu

a. Đánh giá BestToLive (quốc gia đáng sống)

Chọn ra dữ liệu thuộc năm 2020 để phân lớp. Nhóm lựa chọn việc phân lớp theo hai thuộc tính.

Kiểm tra bài toán:



Đánh giá: Như vậy, theo mô hình thì dân số thành thị cao nhưng số dân nhập cư tới thấp thì dự đoán sẽ rất thấp.

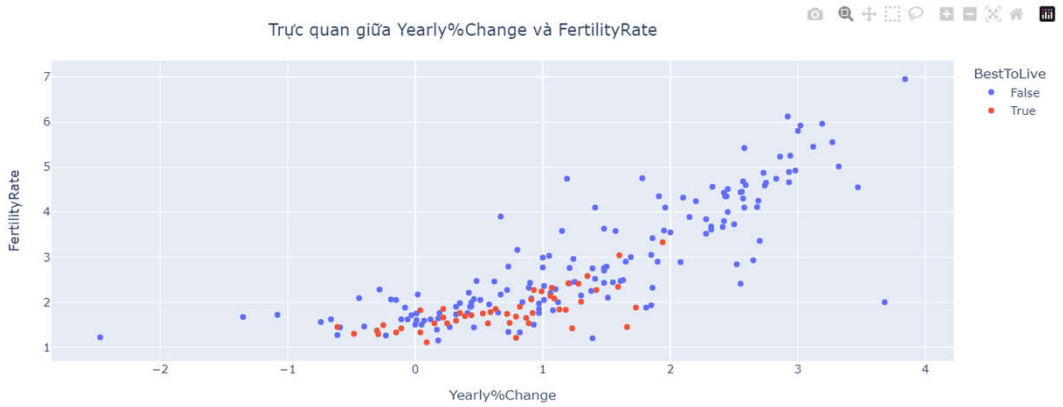
Có vẻ hợp lý khi kết hợp 2 yếu tố này ngoài thực tế thì dự đoán lại rất cao, còn không dự đoán sẽ lệch về phía nữa trên đồ thị và có xu hướng lệch nhẹ về bên trái. Với độ chính xác trên tập train là khoảng 0.81 thì lượng dân nhập cư lớn sẽ phản ánh phần nào mức độ đáng sống của 1 quốc gia. Vì đơn giản "*Ai cũng muốn đến đây sống*".

Mô hình hóa dữ liệu

a. Đánh giá BestToLive (quốc gia đáng sống)

Chọn ra dữ liệu thuộc năm 2020 để phân lớp. Nhóm lựa chọn việc phân lớp theo hai thuộc tính.

- Phân lớp dựa vào FertilityRate và Yearly%Change (bỏ đi các giá trị bằng 0 trong FertilityRate):



Mô hình Hồi quy Logistic cho 2 thuộc tính Yearly%Change và FertilityRate.

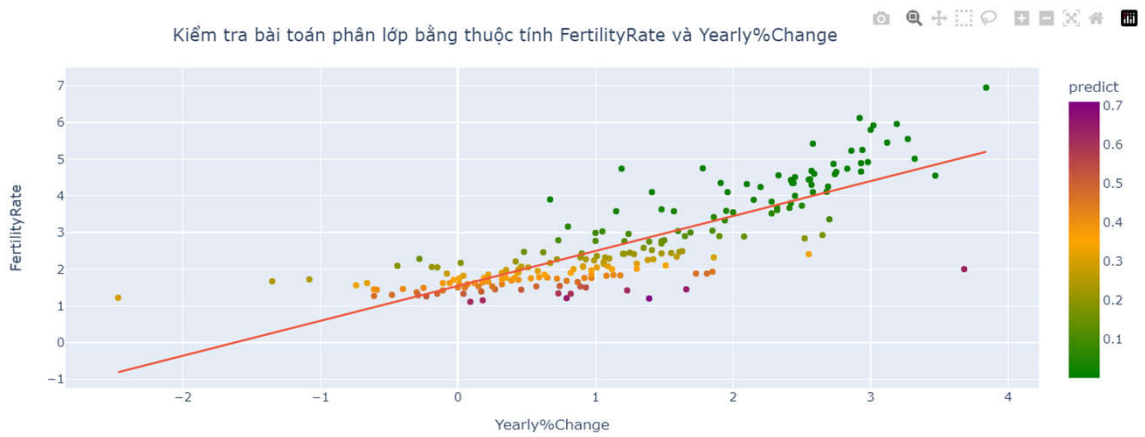
```
Binary cross entropy loss là :  8.763605932921974
Độ chính xác của mô hình :  0.746268656716418
```

Mô hình hóa dữ liệu

a. Đánh giá BestToLive (quốc gia đáng sống)

Chọn ra dữ liệu thuộc năm 2020 để phân lớp. Nhóm lựa chọn việc phân lớp theo hai thuộc tính.

Kiểm tra bài toán:



Đánh giá: Có vẻ như tỉ lệ sinh và số lượng dân số không thay đổi không quyết định nhiều lắm về chất lượng quốc gia đáng sống.

Mô hình hóa dữ liệu

b. Dự đoán dân số trong tương lai

Ở bài này ta sẽ sử dụng mô hình Hồi quy tuyến tính (Lineaer Regression) và Hồi quy đa thức (Polynomial Regression) bậc 2. Trong đó Hồi quy tuyến tính có thể xem là Hồi quy đa thức bậc 1.

Tạo ra 2 hàm linear_regression_1 và linear_regression_2 để biểu diễn hồi quy tuyến tính (linear bậc 1) và hồi quy đa thức bậc 2 (linear bậc 2) để so sánh sự khác biệt. Ngoài ra, phải xử lý trường hợp mô hình dự đoán số dân của một số nước giảm về 0:

```
#Có những nước với dụ đoán GIẢM có nguy cơ diệt vong (0 dân)
dict_pred['Population']=[i if i>0 else 1000 for i in dict_pred.pop('Population')]
dict_pred = [dict(zip(dict_pred,t)) for t in zip(*dict_pred.values())]
for row in dict_pred :
    yield row
```

Mô hình hóa dữ liệu

b. Dự đoán dân số trong tương lai

Ta sẽ tiến hành kiểm tra hàm xây dựng được qua tình hình dân số nước Afghanistan.

Có vẻ như hàm xây dựng đã đúng, ta tiến hành áp dụng mô hình tuyến tính cho tệp dữ liệu của 275 quốc gia. Và sau đó, đưa ra dự đoán dân số vào các năm 2022, 2025, 2030, 2035, 2040, 2050.

Sau khi đưa ra dự đoán, ta sẽ dựa vào hàm mất mát xây dựng được để đánh giá mô hình.

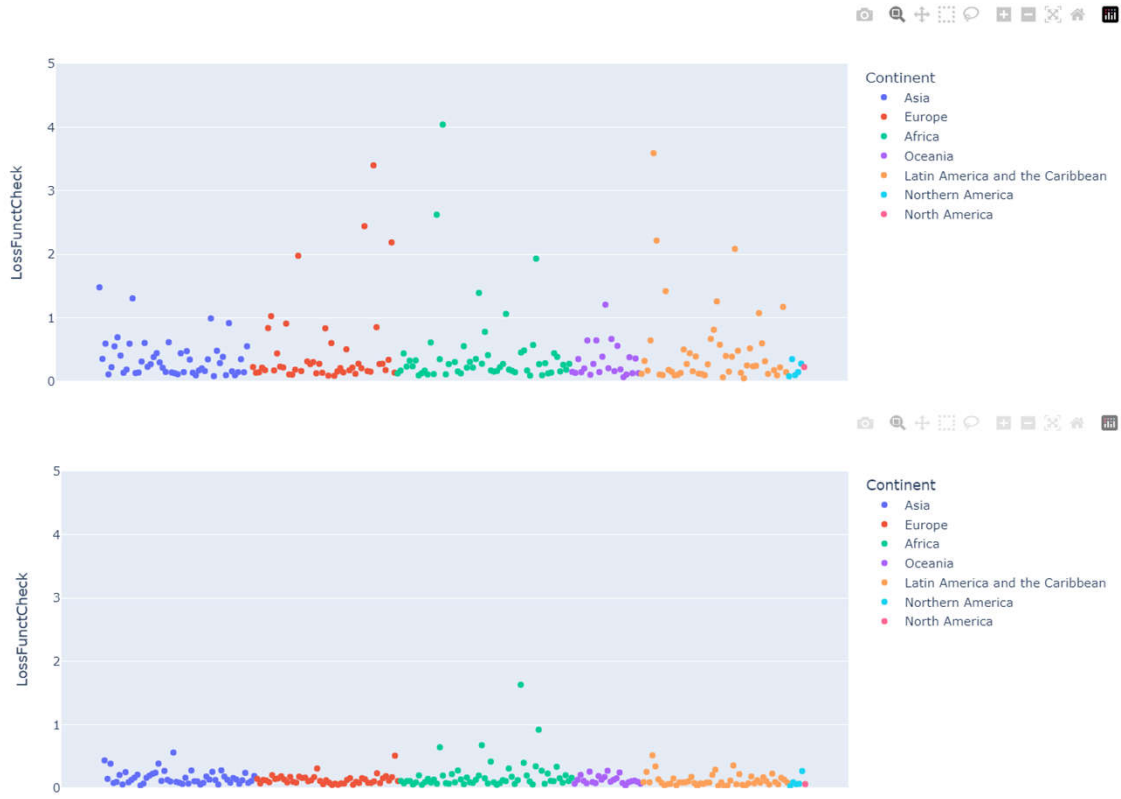
Mô hình hóa dữ liệu

b. Dự đoán dân số trong tương lai

Vậy là ta đã có những dữ đoán cho tình hình dân số tương lai, liệu mô hình đã hiệu quả chưa? Ta cùng đánh giá qua hành mất mát.

Nhận xét:

- Điểm trừ của Linear bậc 1 : biên độ mất mát bị lệch do ảnh hướng từ các số liệu lệch đột biến. Ví dụ giả sử Afghanistan có dân số xu hướng tăng theo thời gian nhưng dân số năm 2022 và 2025 được dự đoán sẽ nhỏ hơn dân số năm 2020.
- Điểm trừ của LD2 : Overfit, không phù hợp để chuẩn đoán các mốc thời gian xa hơn, vì ta dự đoán tới năm 2050. Tuy nhiên rất tốt để chuẩn đoán các giá trị liền kề, nhất là năm 2022 (Hiện tại) từ dữ liệu tới năm 2020.



Mô hình hóa dữ liệu

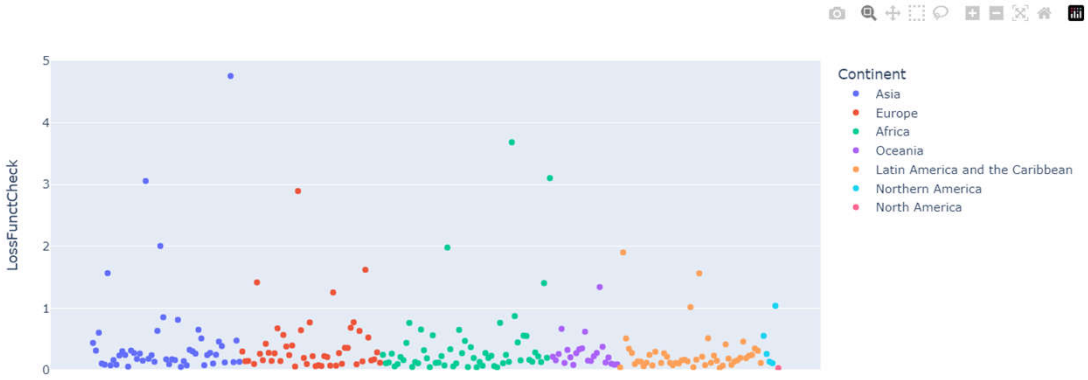
b. Dự đoán dân số trong tương lai

Giải pháp:

Giả thuyết về độ mất mát lớn: Dữ liệu được thu thập vào năm 1950, là năm hậu chiến tranh thế giới thứ 2 và vẫn còn diễn ra xung đột vũ trang ở một số quốc gia và vùng lãnh thổ, dẫn đến biến động về dân số.

Vì thế ta đặt mốc ổn định của thế giới là năm 1990 (tình hình thế giới có thể được coi là ít xung độ vũ trang) để cải tiến dự đoán của mô hình Hồi quy Tuyến tính.

Sau khi thay đổi, có vẻ các dự đoán không có vấn đề gì nữa. Tiếp theo ta sử dụng hàm Hồi quy bậc 2 để dự đoán tương lai gần là năm 2022 so với năm 2020.



	Year	YearlyChange	Migrants(net)	MedianAge	FertilityRate	UrbanPopulation	Yearly%Change	UrbanPop%	Country	Continent	LossFunctCheck	Population
0	2022	3.346121e+06	-411126.352852	40.369599	1.550087	9.384110e+08	0.229571	64.382646	China	Asia	0.091138	1.457553e+09
1	2022	1.296889e+07	-700805.569106	28.761390	1.933698	5.040785e+08	0.901463	35.038327	India	Asia	0.040062	1.438649e+09
2	2022	2.172761e+06	982945.785393	39.282261	2.014710	2.809522e+08	0.641556	82.957421	United States	Northern America	0.268629	3.386704e+08
3	2022	2.806483e+06	-158829.810498	30.976352	2.100638	1.657408e+08	0.999069	59.001419	Indonesia	Asia	0.071430	2.809099e+08
4	2022	4.412810e+06	-294709.132706	23.015787	3.110085	8.112318e+07	1.904715	35.015452	Pakistan	Asia	0.055317	2.316782e+08
...	...	...	...	...	...	...	...	...	...	...	...	...
230	2022	3.853897e+01	0.000000	0.000000	0.000000	3.472436e+02	1.002488	9.032608	Montserrat	Latin America and the Caribbean	0.219118	3.844333e+03
231	2022	1.252865e+02	0.000000	0.000000	0.000000	2.475632e+03	3.603189	71.198181	Falkland Islands	Latin America and the Caribbean	0.085769	3.477100e+03
232	2022	3.929776e+01	0.000000	0.000000	0.000000	7.014603e+02	2.764501	49.346007	Niue	Oceania	0.100013	1.421514e+03
233	2022	2.705358e+01	0.000000	0.000000	0.000000	0.000000e+00	2.197878	0.000000	Tokelau	Oceania	0.091223	1.230895e+03
234	2022	-2.240483e-01	0.000000	0.000000	0.000000	8.360805e+02	-0.026765	99.880537	Holy See	Europe	0.122860	8.370805e+02

Mô hình hóa dữ liệu

b. Dự đoán dân số trong tương lai

Đưa ra dự đoán và kết luận:

- Tổng dân số thế giới 2022:

```
sum_2022 = LD['Population'].sum()
sum_2022
```

✓ 0.6s

```
8001369602.050169
```

- Dự đoán trong tương lai:

```
2030: 8648797081.485008
2031: 8734648873.204184
2032: 8820645342.809118
2033: 8906786490.299911
2034: 8993072315.676558
2035: 9079502818.93888
```

Dự đoán năm **2035** dân số sẽ đạt **9 tỷ** người

Kết luận: Khá sát với thực tế khi dân số 2022 đã đạt mốc **8 tỷ** người.


Thank You
Cảm ơn Thầy đã lắng nghe
và kính mời Thầy đặt câu hỏi
cho chúng em