BỘ GIÁO DỤC VÀ ĐÀO TẠO

TRƯỜNG ĐẠI HỌC MỞ THÀNH PHỐ HỒ CHÍ MINH

...ଓ ...

BÁO CÁO PHÂN TÍCH KẾT QUẢ MỨC ĐỘ HẠNH PHÚC TRÊN THẾ GIỚI TỪ NĂM 2019 TỚI NĂM 2023

Danh sách thành viên:

2151010220 - Nguyễn Đặng Phú Mẫn

2151013053 - Hoàng Quốc Minh

2151010230 - Lê Thị Yến My

Lớp: CS2101

Giảng viên: Nguyễn Văn Bảy

TP. Hồ Chí Minh, ngày 16 tháng 1 năm 2023

MỤC LỤC

DỮ LIỆU MỨC ĐỘ HẠNH PHÚC TRÊN THẾ GIỚI TỪ NĂM 2019 TỚI NĂM 2023

1. Mô tả dữ liệu

1.1. Giới thiệu về bộ dữ liệu

- Báo cáo Hạnh phúc Thế giới (World Happiness Report) là tập hợp những bài báo cáo về thứ hạng mức độ hạnh phúc ở các quốc gia vào mỗi năm, dựa trên một số yếu tố trong cuộc sống hằng ngày để đánh giá. Tính đến 2023 thì đã có 11 bài báo cáo (Bài báo cáo đầu tiên là năm 2012, bài báo cáo mới nhất là năm 2023) [1].

- **Cách thức thu thập dữ liệu:**

  - Điểm hạnh phúc của từng nước được dựa trên những bài khảo sát của một công ty chuyên về khảo sát - bỏ phiếu ở Mỹ tên Gallup, Inc.
  - Những người đại diện từng quốc gia sẽ được yêu cầu đánh giá cuộc sống hiện tại của họ dựa trên một thang đo, với 10 điểm là cuộc sống hạnh phúc nhất có thể và 0 điểm là cuộc sống không hạnh phúc nhất có thể [2].
  - Việc thực hiện đánh giá thông qua những phép đo mức sống có phần chủ quan này là một cách tiếp cận từ dưới lên (bottom-up) để giúp những người trả lời khảo sát có thể tự đánh giá mức sống của chính bản thân [3]. Trong bối cảnh này, những giá trị trên Thang Cantril là minh chứng rõ ràng cho việc những người trả lời khảo sát biết mức sống của mình ở đâu dựa vào quan điểm của riêng họ [4].
  - Chính vì cách thu thập này nên không phải quốc gia nào cũng sẽ có mặt trong các bài báo cáo.

- **Các thuộc tính được dùng để đánh giá trong các bài Báo cáo Hạnh phúc Thế giới:**

  - Các quốc gia tham gia vào báo cáo sẽ được xếp hạng dựa trên Điểm hạnh phúc, và điểm đó sẽ được tính theo 6 thuộc tính sau: GDP đầu người (Logged GDP per capita), Hỗ trợ xã hội (Social support), Tuổi thọ khỏe mạnh kỳ vọng (Healthy life expectancy), Tự do khi lựa chọn trong cuộc sống (Freedom to make life choices), Độ rộng lượng (Generosity) và Mức độ nhận thức về tham nhũng (Perceptions of corruption) [2].
  - Trong những nghiên cứu quy mô toàn cầu trước đây về sự khác biệt mức sống của các quốc gia với nhau thì 6 thuộc tính trên nằm trong những thuộc tính được sử dụng để nghiên cứu nhiều nhất. Những thuộc tính khác như Sự bất công hay Thất nghiệp, tuy cũng quan trọng nhưng sẽ không được sử dụng trong bài báo cáo, lý do là vì chưa có đủ dữ liệu quốc tế để có thể so sánh được với tất cả các quốc gia [2].

- **Cách thức so sánh mức độ hạnh phúc:**

  - Tất cả các nước có trong các bài báo cáo sẽ được so sánh với một quốc gia giả tưởng tên Dystopia.
  - Dystopia là quốc gia gồm những người ít hạnh phúc nhất trên thế giới. Mục đích của việc xây dựng lên Dystopia là để dùng nó như một chuẩn so sánh với các quốc

gia còn lại trên thế giới, tất cả quốc gia đều có lợi thế hơn khi được so sánh với Dystopia (không có quốc gia nào có điểm hạnh phúc thấp hơn Dystopia) [1].

- Điểm số thấp nhất ở 6 thuộc tính chính là đặc điểm của Dystopia. Và vì cuộc sống sẽ vô cùng khó khăn khi một người sống trong một quốc gia với GDP đầu người thấp nhất, hỗ trợ xã hội thấp nhất, tuổi thọ khỏe mạnh kỳ vọng thấp nhất, tự do khi lựa chọn trong cuộc sống thấp nhất, độ rộng lượng thấp nhất và mức độ nhận thức về tham nhũng thấp nhất, nên quốc gia này mới được gọi là "Dystopia" (phản địa đàng), ngược lại với "Utopia" (một nơi hoàn hảo nhất về mọi mặt) [1].

- **Những chỉ trích mà báo cáo này đã nhận:**

- Những bản Báo cáo Hạnh phúc thế giới cũng đã gặp phải những chỉ trích từ các cá nhân, trong đó bao gồm cả những nhà phê bình và giới chuyên môn.
- Cách thức thu thập dữ liệu có phần chủ quan này được cho rằng là không thể phản ánh hoàn toàn được sự hạnh phúc của một quốc gia. [5]
- Ý tưởng cho rằng có thể đánh giá hạnh phúc quốc gia chỉ qua khảo sát cũng đã gặp phải nhiều sự tranh cãi từ các nhà kinh tế. Họ chỉ ra rằng, đánh giá về hạnh phúc của mỗi người có thể bị ảnh hưởng bởi nhiều yếu tố như là hệ thống giáo dục, sức khỏe hay những yếu tố khác nữa. [5]

- **Giải thích chi tiết các bộ dữ liệu nhóm em sẽ sử dụng trong bài báo cáo.**

- Nhóm chúng em sẽ sử dụng 5 bộ dữ liệu từ năm 2019 đến năm 2023 để phân tích trong bài báo cáo này.
- Với mỗi bộ dữ liệu, chúng em sẽ sử dụng 8 cột, bao gồm Tên quốc gia, Điểm hạnh phúc, GDP đầu người, Hỗ trợ xã hội, Tuổi thọ trung bình, Mức độ tự do các lựa chọn trong cuộc sống, Mức độ rộng lượng và Mức độ nhận thức về tham nhũng.
- Bộ dữ liệu 2019 gồm 156 instances, 9 attributes:

	Overall rank	Country or region	Score	GDP per capita	Social support	Healthy life expectancy	Freedom to make life choices	Generosity	Perceptions of corruption
0	1	Finland	7.769	1.340	1.587	0.986	0.596	0.153	0.393
1	2	Denmark	7.600	1.383	1.573	0.996	0.592	0.252	0.410
2	3	Norway	7.554	1.488	1.582	1.028	0.603	0.271	0.341
3	4	Iceland	7.494	1.380	1.624	1.026	0.591	0.354	0.118
4	5	Netherlands	7.488	1.396	1.522	0.999	0.557	0.322	0.298
...	...	...	...	...	...	...	...	...	...
151	152	Rwanda	3.334	0.359	0.711	0.614	0.555	0.217	0.411
152	153	Tanzania	3.231	0.476	0.885	0.499	0.417	0.276	0.147
153	154	Afghanistan	3.203	0.350	0.517	0.361	0.000	0.158	0.025
154	155	Central African Republic	3.083	0.026	0.000	0.105	0.225	0.235	0.035
155	156	South Sudan	2.853	0.306	0.575	0.295	0.010	0.202	0.091

156 rows × 9 columns

- Bộ dữ liệu 2020 gồm 153 instances, 9 attributes:

	Country name	Regional indicator	Ladder score	Standard error of ladder score	upperwhisker	lowerwhisker	Logged GDP per capita	Social support	Healthy life expectancy	Freedom to make life choices	Generosity	Perceptions of corruption	Ladder score in Dystopia	Explained by: Log GDP per capita	Explained by: Social support	Explained by: Healthy life expectancy	Explained by: Freedom to make life choices	Explained by: Generosity	Explained by: Perceptions of corruption	Dystopia + residual
0	Finland	Western Europe	7.8087	0.031156	7.869766	7.747634	10.639267	0.954330	71.900825	0.949172	-0.059482	0.195445	1.972317	1.285190	1.499526	0.961271	0.662317	0.159670	0.477857	2.762835
1	Denmark	Western Europe	7.6456	0.033492	7.711245	7.579955	10.774001	0.955991	72.402504	0.951444	0.066202	0.168489	1.972317	1.326949	1.503449	0.979333	0.665040	0.242793	0.495260	2.432741
2	Switzerland	Western Europe	7.5599	0.035014	7.628528	7.491272	10.979933	0.942847	74.102448	0.921337	0.105911	0.303728	1.972317	1.390774	1.472403	1.040533	0.628954	0.269056	0.407946	2.350267
3	Iceland	Western Europe	7.5045	0.059616	7.621347	7.387653	10.772559	0.974670	73.000000	0.948892	0.246944	0.711710	1.972317	1.326502	1.547567	1.000843	0.661981	0.362330	0.144541	2.460688
4	Norway	Western Europe	7.4880	0.034837	7.556281	7.419719	11.087804	0.952487	73.200783	0.955750	0.134533	0.263218	1.972317	1.424207	1.495173	1.008072	0.670201	0.287985	0.434101	2.168266
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
148	Central African Republic	Sub-Saharan Africa	3.4759	0.115183	3.701658	3.250141	6.625160	0.319460	45.200001	0.640881	0.082410	0.891807	1.972317	0.041072	0.000000	0.000000	0.292814	0.253513	0.028265	2.860198
149	Rwanda	Sub-Saharan Africa	3.3123	0.052425	3.415053	3.209547	7.600104	0.540835	61.098846	0.900589	0.055484	0.183541	1.972317	0.343243	0.522876	0.572383	0.604088	0.235705	0.485542	0.548445
150	Zimbabwe	Sub-Saharan Africa	3.2992	0.058674	3.414202	3.184198	7.865712	0.763093	55.617260	0.711458	-0.072064	0.810237	1.972317	0.425564	1.047835	0.375038	0.377405	0.151349	0.080929	0.841031
151	South Sudan	Sub-Saharan Africa	2.8166	0.107610	3.027516	2.605684	7.425360	0.553707	51.000000	0.451314	0.016519	0.763417	1.972317	0.289083	0.553279	0.208809	0.065609	0.209935	0.111157	1.378751
152	Afghanistan	South Asia	2.5669	0.031311	2.628270	2.505530	7.462861	0.470367	52.590000	0.396573	-0.096429	0.933687	1.972317	0.300706	0.356434	0.266052	0.000000	0.135235	0.001226	1.507236

153 rows × 20 columns

- Bộ dữ liệu 2021 gồm 149 instances, 9 attributes:

	Country name	Regional indicator	Ladder score	Standard error of ladder score	upperwhisker	lowerwhisker	Logged GDP per capita	Social support	Healthy life expectancy	Freedom to make life choices	Generosity	Perceptions of corruption	Ladder score in Dystopia	Explained by: Log GDP per capita	Explained by: Social support	Explained by: Healthy life expectancy	Explained by: Freedom to make life choices	Explained by: Generosity	Explained by: Perceptions of corruption	Dystopia + residual
0	Finland	Western Europe	7.842	0.032	7.904	7.780	10.775	0.954	72.000	0.949	-0.098	0.186	2.43	1.446	1.106	0.741	0.691	0.124	0.481	3.253
1	Denmark	Western Europe	7.620	0.035	7.687	7.552	10.933	0.954	72.700	0.946	0.030	0.179	2.43	1.502	1.108	0.763	0.686	0.208	0.485	2.868
2	Switzerland	Western Europe	7.571	0.036	7.643	7.500	11.117	0.942	74.400	0.919	0.025	0.292	2.43	1.566	1.079	0.816	0.653	0.204	0.413	2.839
3	Iceland	Western Europe	7.554	0.059	7.670	7.438	10.878	0.983	73.000	0.955	0.160	0.673	2.43	1.482	1.172	0.772	0.698	0.293	0.170	2.967
4	Netherlands	Western Europe	7.464	0.027	7.518	7.410	10.932	0.942	72.400	0.913	0.175	0.338	2.43	1.501	1.079	0.753	0.647	0.302	0.384	2.798
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
144	Lesotho	Sub-Saharan Africa	3.512	0.120	3.748	3.276	7.926	0.787	48.700	0.715	-0.131	0.915	2.43	0.451	0.731	0.007	0.405	0.103	0.015	1.800
145	Botswana	Sub-Saharan Africa	3.467	0.074	3.611	3.322	9.782	0.784	59.269	0.824	-0.246	0.801	2.43	1.099	0.724	0.340	0.539	0.027	0.088	0.648
146	Rwanda	Sub-Saharan Africa	3.415	0.068	3.548	3.282	7.676	0.552	61.400	0.897	0.061	0.167	2.43	0.364	0.202	0.407	0.627	0.227	0.493	1.095
147	Zimbabwe	Sub-Saharan Africa	3.145	0.058	3.259	3.030	7.943	0.750	56.201	0.677	-0.047	0.821	2.43	0.457	0.649	0.243	0.359	0.157	0.075	1.205
148	Afghanistan	South Asia	2.523	0.038	2.596	2.449	7.695	0.463	52.493	0.382	-0.102	0.924	2.43	0.370	0.000	0.126	0.000	0.122	0.010	1.895

149 rows × 20 columns

- Bộ dữ liệu 2022 gồm 147 instances, 9 attributes:

	RANK	Country	Happiness score	Whisker-high	Whisker-low	Dystopia (1.83) + residual	Explained by: GDP per capita	Explained by: Social support	Explained by: Healthy life expectancy	Explained by: Freedom to make life choices	Explained by: Generosity	Explained by: Perceptions of corruption
0	1	Finland	7,821	7,886	7,756	2,518	1,892	1,258	0,775	0,736	0,109	0,534
1	2	Denmark	7,636	7,710	7,563	2,226	1,953	1,243	0,777	0,719	0,188	0,532
2	3	Iceland	7,557	7,651	7,464	2,320	1,936	1,320	0,803	0,718	0,270	0,191
3	4	Switzerland	7,512	7,586	7,437	2,153	2,026	1,226	0,822	0,677	0,147	0,461
4	5	Netherlands	7,415	7,471	7,359	2,137	1,945	1,206	0,787	0,651	0,271	0,419
...	...	...	...	...	...	...	...	...	...	...	...	...
141	142	Botswana*	3,471	3,667	3,275	0,187	1,503	0,815	0,280	0,571	0,012	0,102
142	143	Rwanda*	3,268	3,462	3,074	0,536	0,785	0,133	0,462	0,621	0,187	0,544
143	144	Zimbabwe	2,995	3,110	2,880	0,548	0,947	0,690	0,270	0,329	0,106	0,105
144	145	Lebanon	2,955	3,049	2,862	0,216	1,392	0,498	0,631	0,103	0,082	0,034
145	146	Afghanistan	2,404	2,469	2,339	1,263	0,758	0,000	0,289	0,000	0,089	0,005

146 rows × 12 columns

- Bộ dữ liệu 2023 gồm 137 instances, 19 attributes:

	Country name	Ladder score	Standard error of ladder score	upperwhisker	lowerwhisker	Logged GDP per capita	Social support	Healthy life expectancy	Freedom to make life choices	Generosity	Perceptions of corruption	Ladder score in Dystopia	Explained by: Log GDP per capita	Explained by: Social support	Explained by: Healthy life expectancy	Explained by: Freedom to make life choices	Explained by: Generosity	Explained by: Perceptions of corruption	Dystopia + residual
0	Finland	7.804	0.036	7.875	7.733	10.792	0.969	71.150	0.961	-0.019	0.182	1.778	1.888	1.585	0.535	0.772	0.126	0.535	2.363
1	Denmark	7.586	0.041	7.667	7.506	10.962	0.954	71.250	0.934	0.134	0.196	1.778	1.949	1.548	0.537	0.734	0.208	0.525	2.084
2	Iceland	7.530	0.049	7.625	7.434	10.896	0.983	72.050	0.936	0.211	0.668	1.778	1.926	1.620	0.559	0.738	0.250	0.187	2.250
3	Israel	7.473	0.032	7.535	7.411	10.639	0.943	72.697	0.809	-0.023	0.708	1.778	1.833	1.521	0.577	0.569	0.124	0.158	2.691
4	Netherlands	7.403	0.029	7.460	7.346	10.942	0.930	71.550	0.887	0.213	0.379	1.778	1.942	1.488	0.545	0.672	0.251	0.394	2.110
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
132	Congo (Kinshasa)	3.207	0.095	3.394	3.020	7.007	0.652	55.375	0.664	0.086	0.834	1.778	0.531	0.784	0.105	0.375	0.183	0.068	1.162
133	Zimbabwe	3.204	0.061	3.323	3.084	7.641	0.690	54.050	0.654	-0.046	0.766	1.778	0.758	0.881	0.069	0.363	0.112	0.117	0.905
134	Sierra Leone	3.138	0.082	3.299	2.976	7.394	0.555	54.900	0.660	0.105	0.858	1.778	0.670	0.540	0.092	0.371	0.193	0.051	1.221
135	Lebanon	2.392	0.044	2.479	2.305	9.478	0.530	66.149	0.474	-0.141	0.891	1.778	1.417	0.476	0.398	0.123	0.061	0.027	-0.110
136	Afghanistan	1.859	0.033	1.923	1.795	7.324	0.341	54.712	0.382	-0.081	0.847	1.778	0.645	0.000	0.087	0.000	0.093	0.059	0.976

136 rows × 19 columns

1.2. Giới thiệu về phương pháp thu thập dữ liệu

- Tìm kiếm và thu thập dữ liệu từ trang web https://www.kaggle.com/
- Kaggle là một trang web cạnh tranh khoa học dữ liệu phổ biến cung cấp các bộ dữ liệu công khai miễn phí mà chúng ta có thể sử dụng để tìm hiểu về Data Science (Khoa học dữ liệu) và Machine Learning.
- Nhóm đã sử dụng ngôn ngữ Python để thực hiện việc khai phá các bộ dữ liệu.
- Bộ dữ liệu từ năm 2019 đến năm 2022 được lấy ở trang web: https://www.kaggle.com/datasets/mathurinache/world-happiness-report
- Bộ dữ liệu năm 2023 được lấy ở trang web: https://worldhappiness.report/ed/2023/

2. Tiền xử lý dữ liệu

Làm sạch dữ liệu (Data Cleaning): Là quá trình loại bỏ các lỗi, nhiễu trong dữ liệu, bao gồm:

- Lấy dataframe của năm 2019 làm chuẩn vì trong Dataset World Happiness 2019 có đầy đủ các giá trị cần thiết và không dư thừa để tính toán mức độ hạnh phúc của các quốc gia trên thế giới.
- Dataframe 2019 khi đã xóa các cột dữ liệu không cần thiết:

	Overall rank	Country or region	Score	GDP per capita	Social support	Healthy life expectancy	Freedom to make life choices	Generosity	Perceptions of corruption
0	1	Finland	7.769	1.340	1.587	0.986	0.596	0.153	0.393
1	2	Denmark	7.600	1.383	1.573	0.996	0.592	0.252	0.410
2	3	Norway	7.554	1.488	1.582	1.028	0.603	0.271	0.341
3	4	Iceland	7.494	1.380	1.624	1.026	0.591	0.354	0.118
4	5	Netherlands	7.488	1.396	1.522	0.999	0.557	0.322	0.298
...	...	...	...	...	...	...	...	...	...
151	152	Rwanda	3.334	0.359	0.711	0.614	0.555	0.217	0.411
152	153	Tanzania	3.231	0.476	0.885	0.499	0.417	0.276	0.147
153	154	Afghanistan	3.203	0.350	0.517	0.361	0.000	0.158	0.025
154	155	Central African Republic	3.083	0.026	0.000	0.105	0.225	0.235	0.035
155	156	South Sudan	2.853	0.306	0.575	0.295	0.010	0.202	0.091

156 rows × 9 columns

- Dataframe 2020 khi đã xóa các cột dữ liệu không cần thiết:

	Country	Regional indicator	Happy score	GDP	Explained by: Social support	Explained by: Healthy life expectancy	Explained by: Freedom to make life choices	Explained by: Generosity	Explained by: Perceptions of corruption
0	Finland	Western Europe	7.8087	1.285190	1.499526	0.961271	0.662317	0.159670	0.477857
1	Denmark	Western Europe	7.6456	1.326949	1.503449	0.979333	0.665040	0.242793	0.495260
2	Switzerland	Western Europe	7.5599	1.390774	1.472403	1.040533	0.628954	0.269056	0.407946
3	Iceland	Western Europe	7.5045	1.326502	1.547567	1.000843	0.661981	0.362330	0.144541
4	Norway	Western Europe	7.4880	1.424207	1.495173	1.008072	0.670201	0.287985	0.434101
...	...	...	...	...	...	...	...	...	...
148	Central African Republic	Sub-Saharan Africa	3.4759	0.041072	0.000000	0.000000	0.292814	0.253513	0.028265
149	Rwanda	Sub-Saharan Africa	3.3123	0.343243	0.522876	0.572383	0.604088	0.235705	0.485542
150	Zimbabwe	Sub-Saharan Africa	3.2992	0.425564	1.047835	0.375038	0.377405	0.151349	0.080929
151	South Sudan	Sub-Saharan Africa	2.8166	0.289083	0.553279	0.208809	0.065609	0.209935	0.111157
152	Afghanistan	South Asia	2.5669	0.300706	0.356434	0.266052	0.000000	0.135235	0.001226

153 rows × 9 columns

- Dataframe 2021 khi đã xóa cột dữ liệu không cần thiết:

	Country	Regional indicator	Happy score	GDP	Social support	Healthy life expectancy	Freedom to make life choices	Generosity	Perceptions of corruption
0	Finland	Western Europe	7.842	1.446	1.106	0.741	0.691	0.124	0.481
1	Denmark	Western Europe	7.620	1.502	1.108	0.763	0.686	0.208	0.485
2	Switzerland	Western Europe	7.571	1.566	1.079	0.816	0.653	0.204	0.413
3	Iceland	Western Europe	7.554	1.482	1.172	0.772	0.698	0.293	0.170
4	Netherlands	Western Europe	7.464	1.501	1.079	0.753	0.647	0.302	0.384
...	...	...	...	...	...	...	...	...	...
144	Lesotho	Sub-Saharan Africa	3.512	0.451	0.731	0.007	0.405	0.103	0.015
145	Botswana	Sub-Saharan Africa	3.467	1.099	0.724	0.340	0.539	0.027	0.088
146	Rwanda	Sub-Saharan Africa	3.415	0.364	0.202	0.407	0.627	0.227	0.493
147	Zimbabwe	Sub-Saharan Africa	3.145	0.457	0.649	0.243	0.359	0.157	0.075
148	Afghanistan	South Asia	2.523	0.370	0.000	0.126	0.000	0.122	0.010

149 rows × 9 columns

- Dataframe 2022 khi đã xóa cột dữ liệu không cần thiết, chuyển dấu “,” về lại dấu “.” trong số thực và đã chuẩn hóa các giá trị số từ object về float hoặc int:

	RANK	Country	Happy score	GDP	Social support	Healthy life expectancy	Freedom to make life choices	Generosity	Perceptions of corruption
0	1	Finland	7.821	1.892	1.258	0.775	0.736	0.109	0.534
1	2	Denmark	7.636	1.953	1.243	0.777	0.719	0.188	0.532
2	3	Iceland	7.557	1.936	1.320	0.803	0.718	0.270	0.191
3	4	Switzerland	7.512	2.026	1.226	0.822	0.677	0.147	0.461
4	5	Netherlands	7.415	1.945	1.206	0.787	0.651	0.271	0.419
...	...	...	...	...	...	...	...	...	...
142	143	Rwanda*	3.268	0.785	0.133	0.462	0.621	0.187	0.544
143	144	Zimbabwe	2.995	0.947	0.690	0.270	0.329	0.106	0.105
144	145	Lebanon	2.955	1.392	0.498	0.631	0.103	0.082	0.034
145	146	Afghanistan	2.404	0.758	0.000	0.289	0.000	0.089	0.005
146	147	xx	NaN	NaN	NaN	NaN	NaN	NaN	NaN

147 rows × 9 columns

- Dataframe 2023 khi đã xóa các cột dữ liệu không cần thiết:

	Country	Happy score	GDP	Social support	Healthy life expectancy	Freedom to make life choices	Generosity	Perceptions of corruption
0	Finland	7.804	1.888	1.585	0.535	0.772	0.126	0.535
1	Denmark	7.586	1.949	1.548	0.537	0.734	0.208	0.525
2	Iceland	7.530	1.926	1.620	0.559	0.738	0.250	0.187
3	Israel	7.473	1.833	1.521	0.577	0.569	0.124	0.158
4	Netherlands	7.403	1.942	1.488	0.545	0.672	0.251	0.394
...	...	...	...	...	...	...	...	...
132	Congo (Kinshasa)	3.207	0.531	0.784	0.105	0.375	0.183	0.068
133	Zimbabwe	3.204	0.758	0.881	0.069	0.363	0.112	0.117
134	Sierra Leone	3.138	0.670	0.540	0.092	0.371	0.193	0.051
135	Lebanon	2.392	1.417	0.476	0.398	0.123	0.061	0.027
136	Afghanistan	1.859	0.645	0.000	0.087	0.000	0.093	0.059

136 rows × 8 columns

3. Mô tả tập dữ liệu sau xử lý

Tổng quan các bộ dữ liệu sau khi tiền xử lý:

- Không có hàng dữ liệu nào có giá trị null ở cả 5 bộ dữ liệu.
- Đã rút lại các bộ dữ liệu chỉ còn 8 thuộc tính quan trọng nhất để đưa vào phân tích:
  - Tên quốc gia – Country name (kiểu dữ liệu: string): Đây là tên của các quốc gia
  - Điểm hạnh phúc – Happiness score (kiểu dữ liệu: float): Đây là thang đo điểm hạnh phúc chủ quan từ 0-10. Thang điểm này được gọi là Thang đo cuộc sống Cantril, hay đơn giản hơn là Thang đo cuộc sống [2].
  - GDP đầu người - Logged GDP per capita (kiểu dữ liệu: float): Những dữ liệu GDP về sức mua tương đương (PPP) được lấy từ cơ sở dữ liệu World Development Indicators. Dữ liệu GDP của các nước Đài Loan, Syria, Palestine, Venezuela, Djibouti và Yemen thì được lấy từ Penn World Table 10.01 [2].
  - Hỗ trợ xã hội - Social support (kiểu dữ liệu: float): (có thể hiểu đơn giản là khi bạn gặp khó thì có ai sẽ giúp bạn không) là trung bình của câu trả lời có/không đối với câu hỏi: "Khi bạn gặp khó khăn, thì bạn có người thân hay bạn bè mà bạn tin tưởng rằng họ có thể giúp bạn, vào bất cứ khi nào bạn cần, hay không?" [2].
  - Tuổi thọ khỏe mạnh kỳ vọng - Healthy life expectancy (kiểu dữ liệu: float): Dữ liệu được lấy từ kho lưu trữ của Global Health Observatory của Tổ chức Y tế Thế giới (WHO) [2].
  - Tự do khi lựa chọn trong cuộc sống - Freedom to make life choices (kiểu dữ liệu: float): là trung bình của câu trả lời có/không đối với câu hỏi: "Bạn hài lòng, hay không hài lòng với sự tự do mà bạn có khi đưa ra các quyết định cho cuộc đời bạn?" [2].
  - Độ rộng lượng - Generosity (kiểu dữ liệu: float): là trung bình của câu trả lời có/không đối với câu hỏi: "Bạn có hỗ trợ tiền đến các tổ chức từ thiện trong tháng vừa qua không?" [2].
  - Mức độ nhận thức về tham nhũng - Perceptions of corruption (kiểu dữ liệu: float): được đo dựa vào các câu trả lời của hai câu hỏi: "Tham nhũng có lan rộng trong chính phủ hay không?" và "Tham nhũng có lan rộng trong các doanh nghiệp hay không?". Mức độ nhận thức chung lại được tính bằng trung bình của hai câu trả lời có/không của hai câu hỏi đó. Nếu như bị thiếu mức độ nhận thức ở chính phủ, thì mức độ nhận thức chung sẽ là mức độ nhận thức ở các doanh nghiệp. Mức độ nhận thức ở tầm quốc gia cũng chỉ là trung bình mức độ nhận thức ở mức cá nhân mà thôi [2].

4. Mô hình khai phá dữ liệu

4.1. Trực quan hóa dữ liệu:

Các giá trị trung bình, trung vị và lớn nhất của các thuộc tính của 5 bộ dữ liệu qua các năm:

- Đầu tiên, tính toán các giá trị trung bình, giá trị trung vị và giá trị lớn nhất của thuộc tính điểm hạnh phúc và 3 trong số 6 thuộc tính chính của bộ dữ liệu là GDP đầu người, hỗ trợ xã hội và tuổi thọ khỏe mạnh kỳ vọng. Lý do lựa chọn 3 thuộc tính trên là vì nhóm đã cho rằng 3 thuộc tính này sẽ có ảnh hưởng lớn nhất đến hạnh phúc của một người, từ đó dẫn đến hạnh phúc của một quốc gia.
- Điểm hạnh phúc: Ta có thể thấy rằng giá trị trung bình và giá trị lớn nhất của thuộc tính điểm hạnh phúc tan dần theo năm, và đến năm 2022 (giá trị lớn nhất giảm) và 2023 (giá trị trung bình giảm) thì có suy giảm một chút. Tuy nhiên, giá trị trung vị qua các năm tăng mạnh, và không hề có dấu hiệu suy giảm kể từ năm 2019.

	2019	2020	2021	2022	2023
Trung bình	5.41	5.47	5.53	5.55	5.54
Trung vị	5.38	5.52	5.53	5.57	5.69
Lớn nhất	7.77	7.81	7.84	7.82	7.8

- GDP đầu người: Ta có thể thấy rằng cả 3 giá trị đều bị giảm vào năm 2020, nguyên nhân là do ngay lúc này đại dịch Covid-19 bùng phát gây ảnh hưởng không nhỏ đến nền kinh tế thế giới. Nhưng ngay sau đó thì cả 3 giá trị đều tăng mạnh vào năm 2021 và 2022, nguyên nhân có thể do vào lúc này thế giới đã thích nghi được với dịch bệnh và nền kinh tế cũng vì thế mà phát triển trở lại. Năm 2023 không có sự thay đổi lớn nào so với năm 2022.

	2019	2020	2021	2022	2023
Trung bình	0.91	0.87	0.98	1.41	1.41
Trung vị	0.96	0.92	1.03	1.45	1.45
Lớn nhất	1.68	1.54	1.75	2.21	2.2

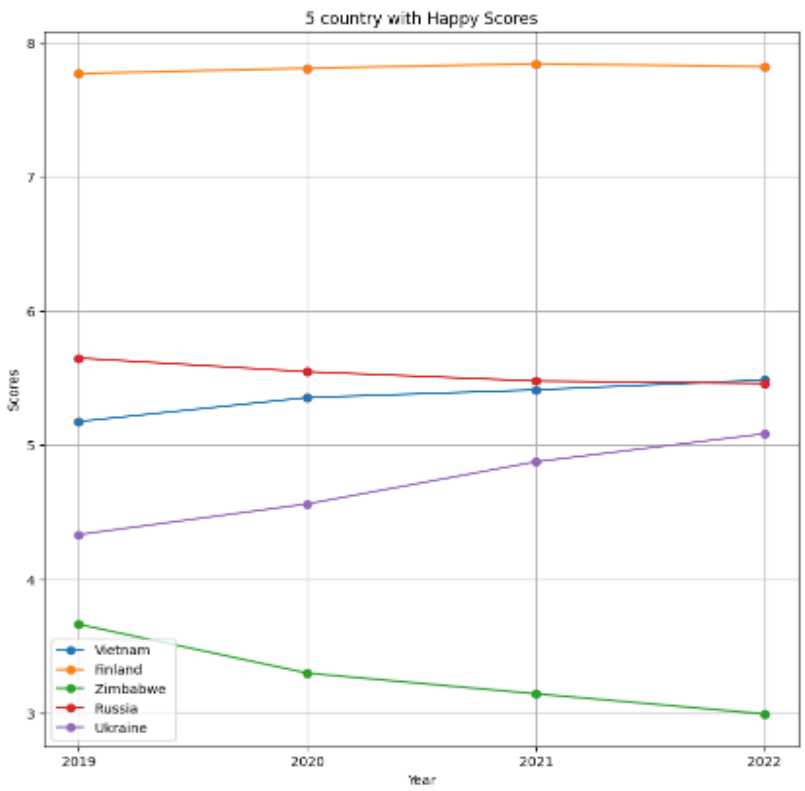
- Hỗ trợ xã hội: Ta có thể thấy rằng cả 3 giá trị đều bị giảm vào năm 2020, nguyên nhân là do ngay lúc này đại dịch Covid-19 bùng phát. Khác với GDP, cả 3 giá trị trên tiếp tục giảm vào năm 2021, nguyên nhân là vì khác với nền kinh tế, thú có thể thích nghi được với tình hình dịch bệnh và dễ dàng phục hồi, thì vào giai đoạn này, cách ly xã hội đang ở mức cao trào nhất, điều này dễ đến sự tiếp xúc giữa người với người giảm đi, vì thế điểm hỗ trợ xã hội cũng giảm theo. Phải cho đến năm 2022 thì cả 3 giá trị mới tăng lại, và đến năm 2023 thì đã quay lại được mức trước dịch.

	2019	2020	2021	2022	2023
Trung bình	1.21	1.16	0.79	0.91	1.16
Trung vị	1.27	1.2	0.8	0.96	1.45
Lớn nhất	1.6	1.55	1.17	1.32	1.62

- Tuổi thọ khỏe mạnh kỳ vọng: Ta có thể thấy rằng cả 3 giá trị đều bị giảm dần qua các năm 2020 và 2021, nguyên nhân là do ngay lúc này đại dịch Covid-19 bùng phát, số lượng người chết vì dịch bệnh là rất nhiều, vì thế dẫn đến tuổi thọ khỏe mạnh trung bình bị suy giảm. Phải cho đến năm 2022 khi tình hình dịch bệnh đã ổn định hơn thì cả 3 giá trị mới tăng trở lại. Năm 2023 ta có thể thấy rằng chỉ số Tuổi thọ khỏe mạnh kỳ vọng giảm là do vì trên thế giới hiện nay diễn ra các cuộc xung đột giữa Nga và Ukraine, Israel và Palestine, giữa các nước ở khu vực trung đông. Điều này ảnh hưởng đến việc suy thoái nền kinh tế, gây tác động mạnh lên chỉ số tuổi thọ khỏe mạnh kỳ vọng.

	2019	2020	2021	2022	2023
Trung bình	0.73	0.69	0.52	0.59	0.37
Trung vị	0.79	0.76	0.57	0.62	0.39
Lớn nhất	1.14	1.14	0.9	0.94	0.7

Biểu đồ đường thể hiện số điểm hạnh phúc của 5 nước: Việt Nam, Zimbabwe, Findland, Ukraine, Nga.



Nhóm chọn 5 nước trên vì Việt Nam hiện đang là quốc gia phát triển, Finland là nước có các chính sách đãi ngộ giáo dục, nhân tài tốt, Zimbabwe là nước đang có lạm phát cao, nước Nga và Ukraine đang có chiến tranh.

- Việt Nam là một nước đang phát triển chúng ta có thể thấy được các chỉ số GDP, Hỗ trợ xã hội, Tự do lựa chọn trong cuộc sống tăng theo từng năm dẫn tới chỉ số Điểm hạnh phúc cũng tăng theo. Tuy nhiên vào các năm 2020 tới 2022 do dịch Covid-19 nên chỉ số Tuổi thọ khỏe mạnh kỳ vọng có phần giảm nhẹ nhưng không nhiều.

Bảng dữ liệu 4 chỉ số có sự tương quan cao với happy score của Việt Nam

	2019	2020	2021	2022	2023
Social support	1.346	0.392	0.873	0.932	1.212
GDP	0.741	0.518	0.817	1.252	1.349
Healthy life expectancy	0.851	0.307	0.616	0.611	0.381
Freedom to make life choice	0.543	0.381	0.679	0.707	0.741

- Finland là một nước phát triển, là một nước luôn có mặt trong top 10 xếp hạng về chỉ số happy score chúng ta có thể thấy được các chỉ số GDP, Hỗ trợ xã hội, Tự do lựa chọn trong cuộc sống, Tuổi thọ khỏe mạnh kỳ vọng đều ổn định và tăng theo từng năm. Tuy nhiên vào các năm 2020 tới 2022 do dịch Covid-19 nên tỷ lệ sức khỏe có phần giảm nhẹ nhưng không nhiều.

Bảng dữ liệu 4 chỉ số có sự tương quan cao với happy score của Finland

	2019	2020	2021	2022	2023
Social support	1.587	1.106	1.106	1.258	1.585
GDP	1.34	1.446	1.446	1.892	1.888
Healthy life expectancy	0.986	0.741	0.741	0.775	0.535
Freedom to make life choice	0.596	0.691	0.691	0.736	0.772

- Zimbabwe là một nước đang phát triển nhưng bên trong nội tình đất nước không ổn định, lạm phát cao. Ta có thể thấy với các chỉ số đã nêu trên thì Zimbabwe luôn ở mức thấp. Riêng chỉ số GDP ở mức cao là do đồng tiền Zim lạm phát quá nhiều. Từ đó dẫn tới chỉ số Điểm hạnh phúc của Zimbabwe lúc nào cũng ở mức thấp.

Bảng dữ liệu 4 chỉ số có sự tương quan cao với happy score của Zimbabwe

	2019	2020	2021	2022	2023
Social support	1.114	1.04783	0.649	0.69	0.881
GDP	0.366	0.42556	0.457	0.947	0.758
Healthy life expectancy	0.433	0.37503	0.243	0.27	0.069
Freedom to make life choice	0.361	0.37740	0.359	0.329	0.363

- Nga và Ukraine là một nước đang phát triển chúng ta có thể thấy rằng là Điểm hạnh phúc của hai nước trên tăng theo theo từng năm nhưng đến năm 2022 và 2023 Điểm hạnh phúc có phần bị giảm do chiến tranh giữa hai nước xảy ra tác động mạnh tới nền kinh tế khiến cho các chỉ số như số GDP, Hỗ trợ xã hội, Tự do lựa chọn trong cuộc sống giảm.

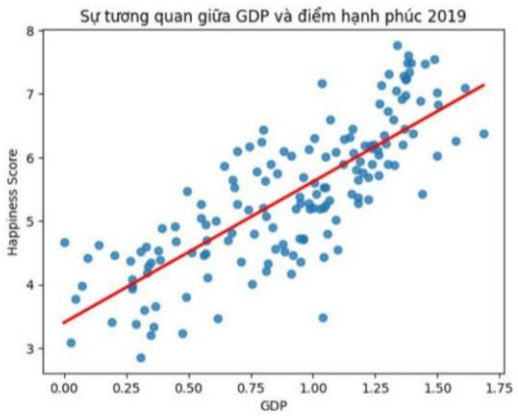
Bảng dữ liệu 4 chỉ số có sự tương quan cao với happy score của Ukraine

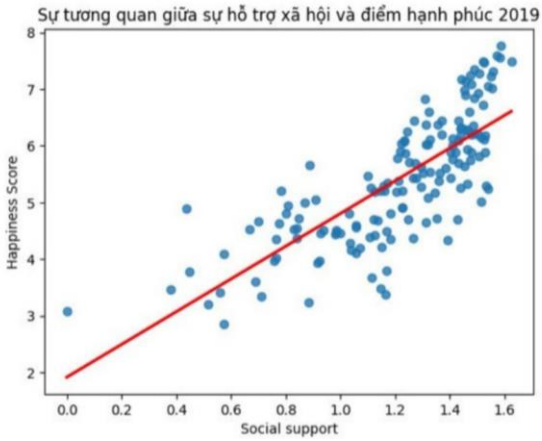
	2019	2020	2021	2022	2023
Social support	1.39	1.321	0.958	1.081	1.354
GDP	0.82	0.7804	0.979	1.411	1.358
Healthy life expectancy	0.739	0.698	0.517	0.583	0.355
Freedom to make life choice	0.178	0.31942	0.417	0.473	0.551

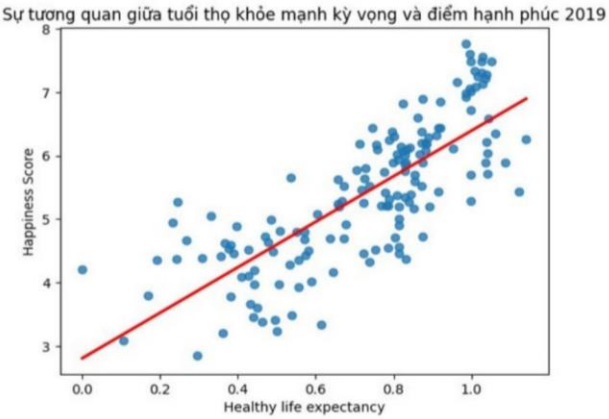
Bảng dữ liệu 4 chỉ số có sự tương quan cao với happy score của Nga

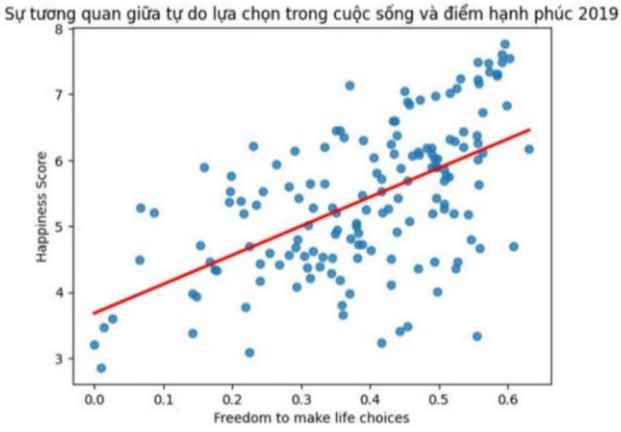
	2019	2020	2021	2022	2023
Social support	1.452	1.37	0.992	1.095	1.383
GDP	1.183	1.12	1.241	1.685	1.68
Healthy life expectancy	0.726	0.68	0.511	0.586	0.366
Freedom to make life choice	0.334	0.39	0.409	0.401	0.449

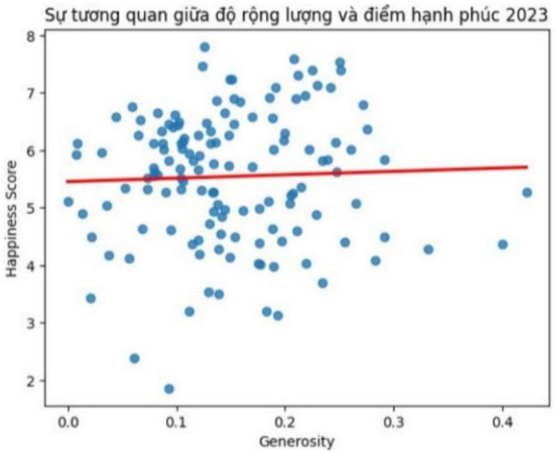
Sự tương quan dữ liệu giữa các thuộc tính và điểm hạnh phúc năm 2019

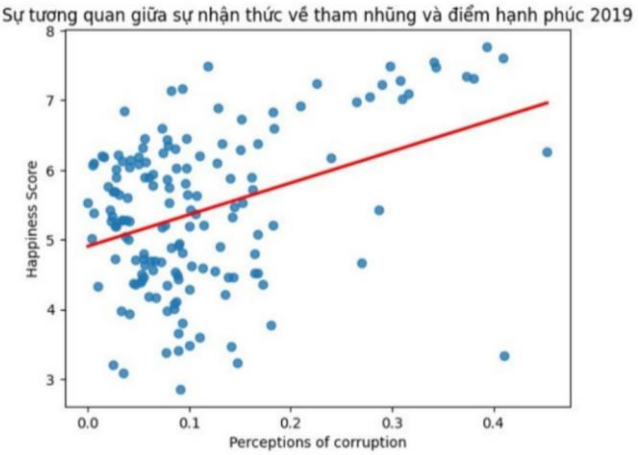




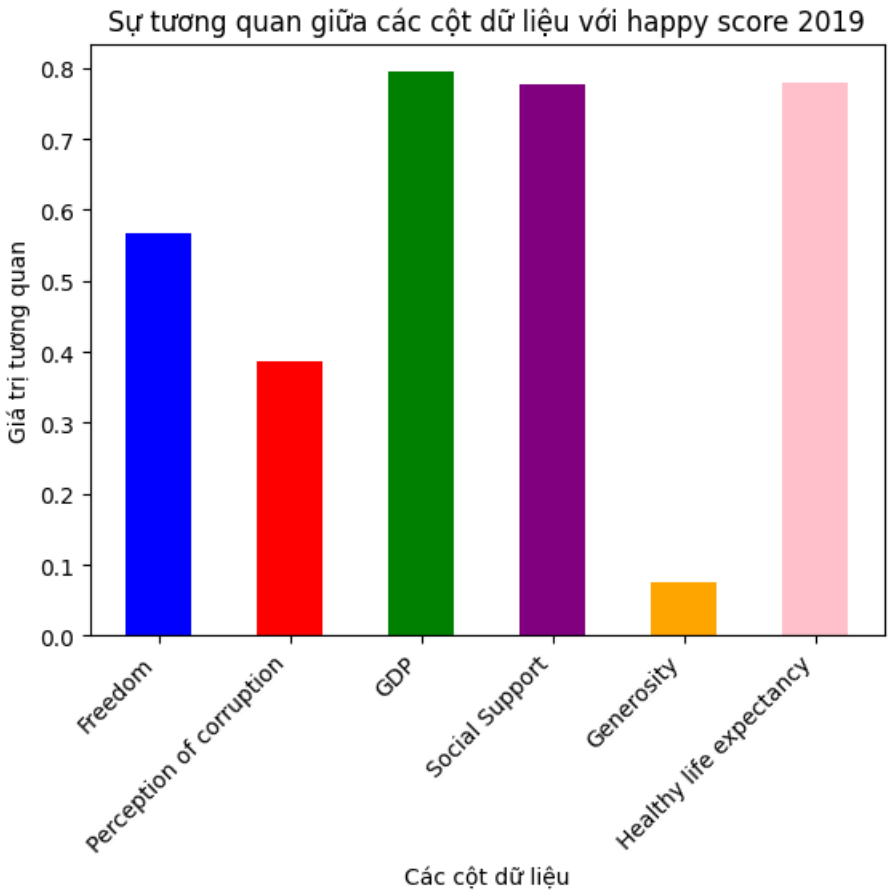






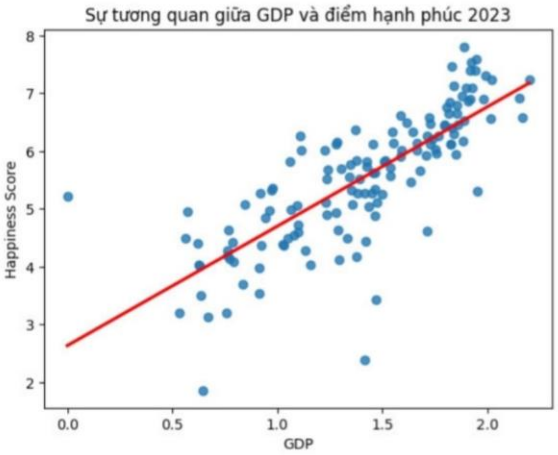


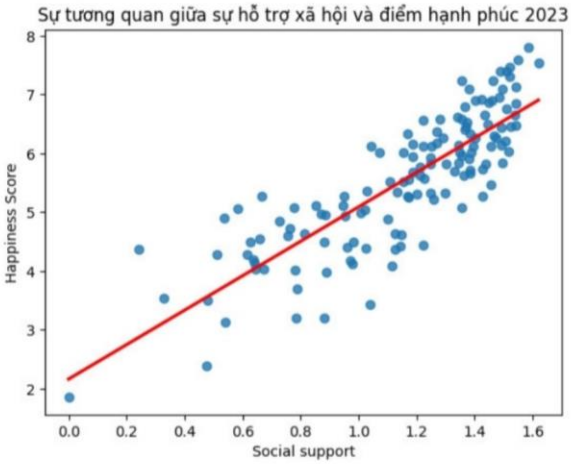
Biểu đồ thể hiện sự tương quan (correalation) giữa các cột dữ liệu với cột điểm hạnh phúc (Happy score) trong năm 2019

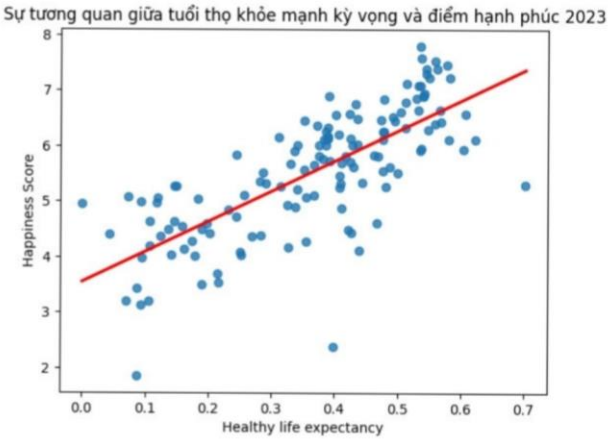


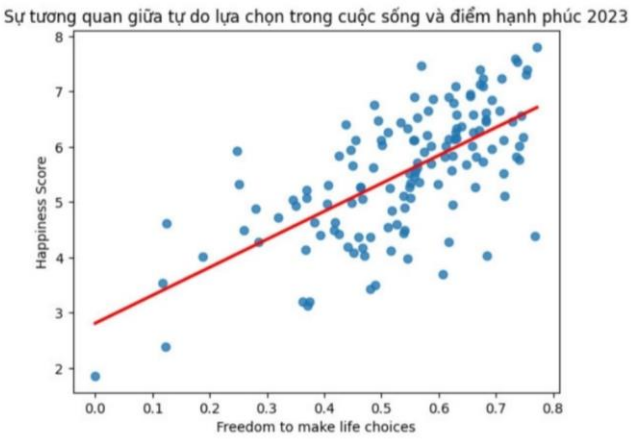
***Nhận xét:** Biểu đồ cột trên cho ta thấy được rằng trong năm 2019 các thuộc tính ảnh hưởng lên Điểm hạnh phúc theo thứ tự từ cao xuống thấp là GDP - Tuổi thọ khỏe mạnh kỳ vọng - Hỗ trợ xã hội - Tự do lựa chọn trong cuộc sống - Mức độ nhận thức về tham nhũng - Độ rộng lượng. Từ đây, thấy được GDP, Tuổi thọ khỏe mạnh kỳ vọng, Hỗ trợ xã hội là những thuộc tính tác động nhiều vào điểm hạnh phúc nhất.

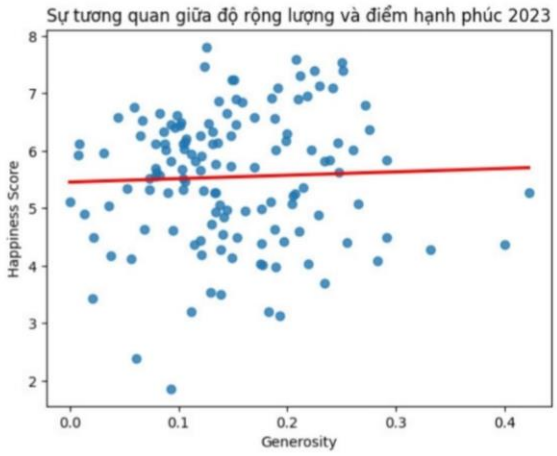
Sự tương quan dữ liệu giữa các thuộc tính và điểm hạnh phúc năm 2023

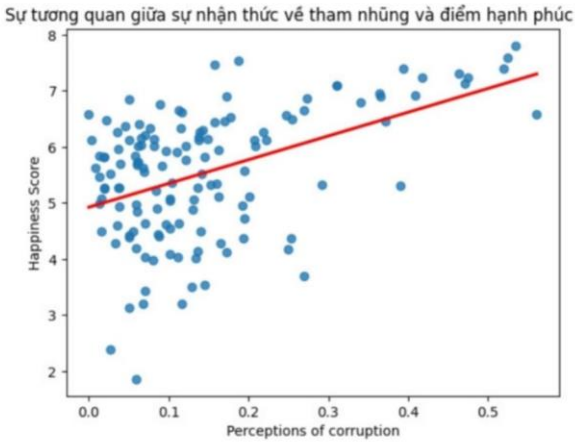




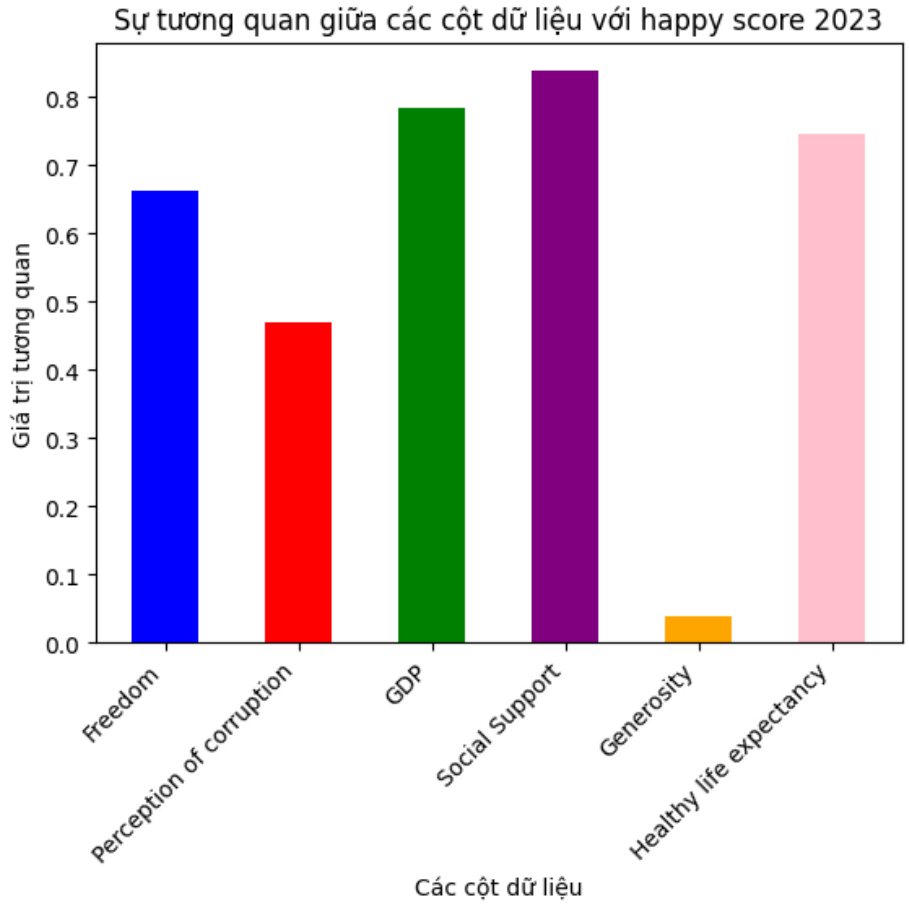








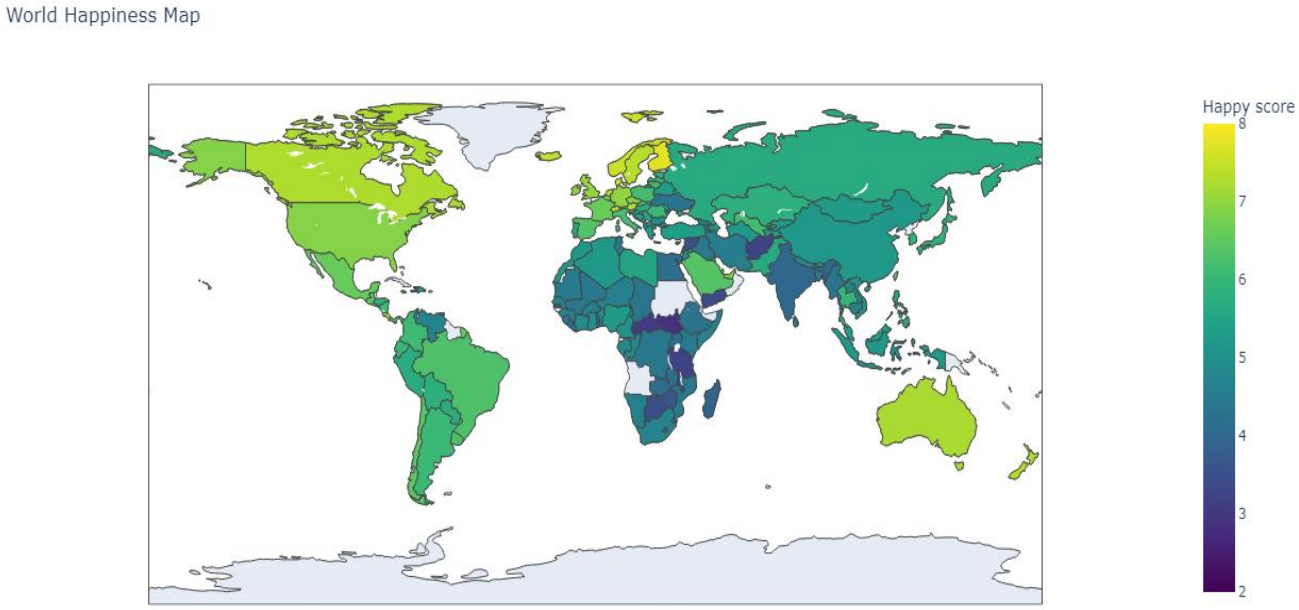
Biểu đồ thể hiện sự tương quan (correalation) giữa các cột dữ liệu với cột điểm hạnh phúc (Happy score) trong năm 2023



***Nhận xét:** Biểu đồ cột trên cho ta thấy được rằng trong năm 2023 các thuộc tính ảnh hưởng lên Điểm hạnh phúc theo thứ tự từ cao xuống thấp là Hỗ trợ xã hội - GDP - Tuổi thọ khỏe mạnh kỳ vọng - Tự do lựa chọn trong cuộc sống - Mức độ nhận thức về tham nhũng - Độ rộng lượng. Từ đây, thấy được Hỗ trợ xã hội, GDP, Tuổi thọ khỏe mạnh kỳ vọng, là những thuộc tính tác động nhiều vào điểm hạnh phúc nhất.

=> Từ 2 biểu đồ thể hiện sự tương quan giữa các cột dữ liệu với cột điểm hạnh phúc trong năm 2019 và năm 2023 và cùng nhận xét trên, ta thấy được sau 3 năm kể từ năm 2019 thì điểm Hỗ trợ xã hội, GDP, Tuổi thọ khỏe mạnh kỳ vọng vẫn sẽ là các thuộc tính tác động vào điểm số hạnh phúc nhiều nhất. Nhưng năm 2019 thì GDP tác động vào Điểm hạnh phúc nhiều nhất và năm 2023 Hỗ trợ xã hội tác động vào Điểm hạnh phúc nhiều nhất.

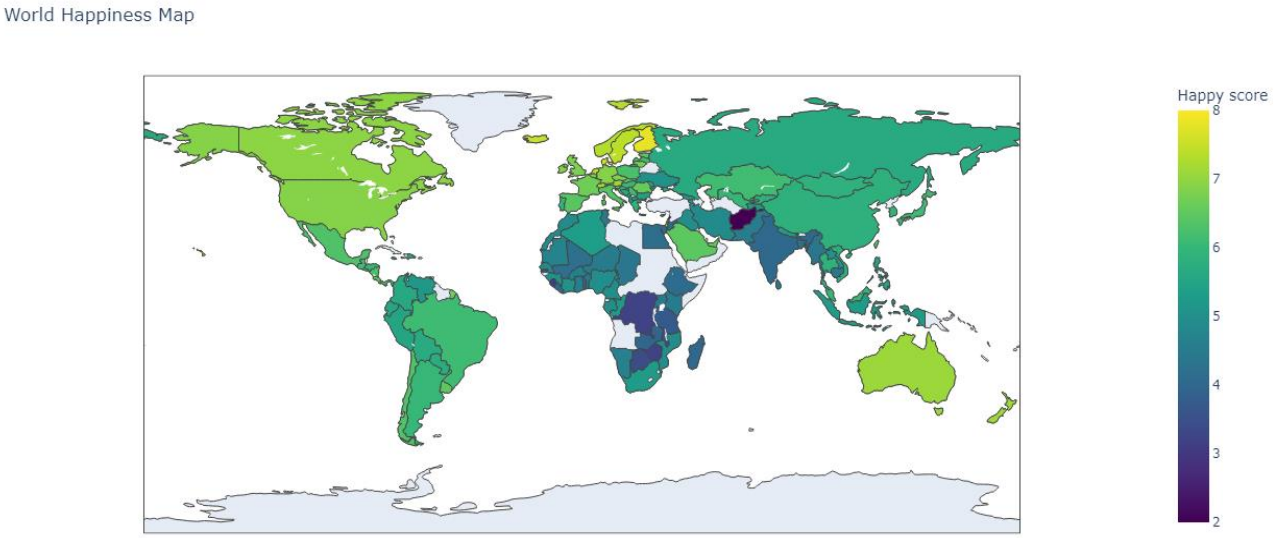
Bản đồ màu Điểm hạnh phúc trên thế giới vào năm 2019



- Qua bản đồ này, ta có thể rút ra một vài điểm nổi bật như sau:

- Những nước có điểm hạnh phúc cao (từ 6.5 trở lên) tập trung đa số ở hai khu vực lớn là Tây Âu, Bắc Mỹ và Châu Đại Dương. Điều này có thể thấy được do nền kinh tế các nước ở các khu vực này là vô cùng phát triển, nên đời sống người dân cũng do đó được đảm bảo hơn. Ngoài ra thì các mảng liên quan đến xã hội như giáo dục, y tế cũng được đảm bảo (điển hình là nền giáo dục ở Phần Lan đã được đánh giá là nền giáo dục hàng đầu thế giới). Từ đó dẫn đến việc mức sống trung bình cao hơn, và người dân hạnh phúc hơn.
- Những nước có điểm hạnh phúc từ trung bình đến thấp (từ 5 điểm đổ xuống) tập trung đa số ở các nước Tây Á và Châu Phi. Điều này có thể thấy là do nền kinh tế các nước các khu vực này tương đối khó khăn, yếu kém, không thể đảm bảo được mức sống cho đa phần người dân. Ngoài ra cũng còn các yếu tố tự nhiên, ví dụ như Châu Phi có vị trí nằm gần như cân xứng so với Xích đạo, phần lớn lãnh thổ nằm giữa hai đường chí tuyến, lục địa có dạng hình khối rõ rệt, ảnh hưởng của biển vào đất liền bị hạn chế, tất cả những yếu tố đó làm cho châu Phi có khí hậu nóng nhất thế giới. Hơn nữa, yếu tố xã hội là xung đột, khủng bố, cũng góp phần ảnh hưởng đến đời sống không được hạnh phúc của người dân các nước này.
- Những nước còn lại (điểm nằm trong khoảng từ 5-6.5) đa phần đều là các nước đang phát triển. Tuy nhiên cũng có sự hiện diện của những nước phát triển như Trung Quốc và Nhật Bản, nguyên nhân dẫn đến chuyện này có thể là do tình hình văn hóa - xã hội ở từng quốc gia (điển hình như văn hóa làm việc ở Nhật Bản cũng đã góp phần vào việc số người tự tử ở Nhật tương đối cao).

Bản đồ màu Điểm hạnh phúc trên thế giới vào năm 2023



- Qua bản đồ này, ta có thể rút ra một vài điểm nổi bật như sau:

- So với năm 2019, phân bố các nước hạnh phúc, không hạnh phúc và các nước còn lại gần như không thay đổi mặc cho thế giới đã trải qua nhiều sự kiện lớn như đại dịch Covid 19, lạm phát trên quy mô toàn cầu, chiến tranh ở Ukraine và nhiều vấn đề khác về môi trường.
- Với đại dịch Covid, tuy đại dịch đã gây ảnh hưởng khá lớn lên hai thuộc tính của bộ dữ liệu là tuổi thọ khỏe mạnh và GDP đầu người, nhưng theo những khảo sát được thực hiện vào năm 2022 để so sánh mức sống trung bình giữa thời kỳ trước và sau dịch, mức sống xã hội trung bình trên toàn thế giới lại tăng. Đúng là ở các nước phương Tây, mức sống trung bình có xu hướng giảm nhưng giảm không nhiều, ngược lại thì ở những nước còn lại trên thế giới thì mức sống trung bình lại tăng, và khi bù trừ lại với nhau thì mức sống trung bình trên toàn thế giới tổng quan là tăng. Ta có thể thấy rằng tuy GDP đầu người và tuổi thọ khỏe mạnh dự đoán có xu hướng giảm, nhưng thuộc tính hỗ trợ xã hội lại tăng, dựa vào sự tương quan của các thuộc tính năm 2023, có thể thấy được hỗ trợ xã hội có ảnh hưởng lớn nhất lên điểm hạnh phúc, vì thế lượng tăng và giảm đã gần như cân bằng nhau. Từ những điều trên ta có thể rút ra rằng Covid gần như không có tác động lớn đến điểm hạnh phúc của các quốc gia [2].
- Chiến tranh giữa Nga và Ukraine, tuy đây là một vấn đề với quy mô khá lớn, nhưng chung quy chỉ ảnh hưởng đến những nước có liên quan trực tiếp đến cuộc chiến. Các nước còn lại hầu như không bị ảnh hưởng gì hoặc chỉ ảnh hưởng tương đối ít, và nếu có thì chỉ ảnh hưởng về mặt kinh tế (ta có thể thấy rõ qua việc giá xăng ở Việt Nam năm 2023 đã có nhiều lần tăng giảm). Tù những điều trên ta có thể rút ra được là chiến tranh của Nga và Ukraine cũng không gây ảnh hưởng mạnh gì đến mức sống và mức độ hạnh phúc ở đa phần các nước trên thế giới.

=> Từ 2 bản đồ trên ta thấy rằng sự phân bố các nước hạnh phúc trên thế giới gần như không thay đổi.

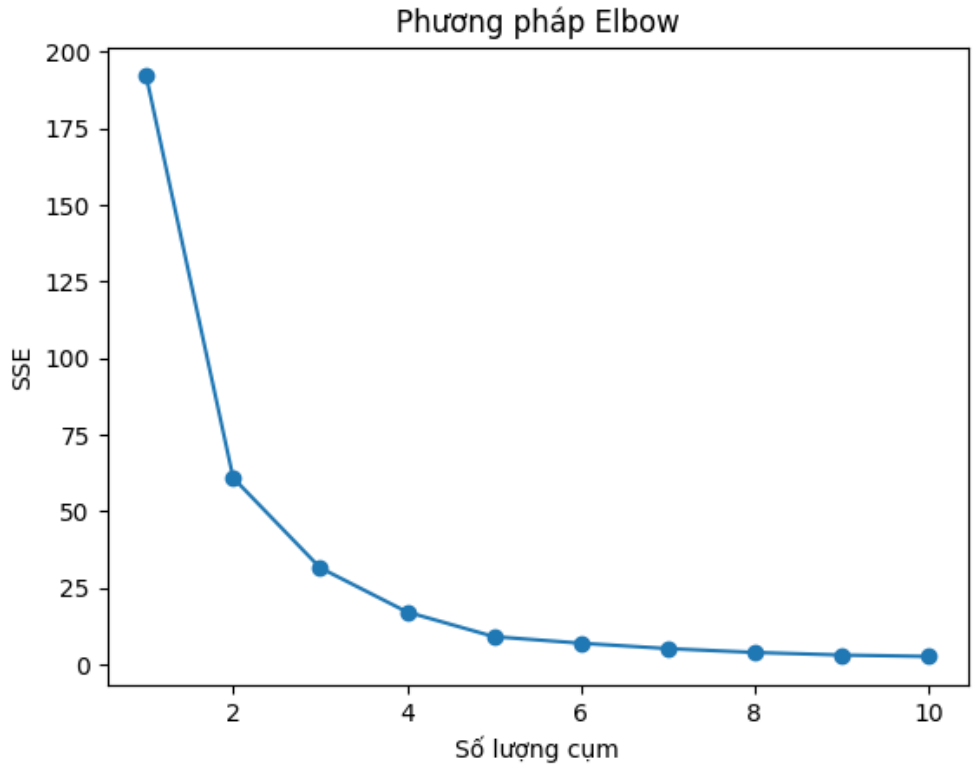
4.2. Gom cụm theo K Mean

K Mean là thuật toán gom cụm, chia tập dữ liệu thành các cụm, các dòng dữ liệu được gán nhãn theo tên của cụm. Thuật toán sẽ dựa vào chỉ số SSE theo phương pháp Elbow để chọn ra số cụm tối ưu của tập dữ liệu. Chúng ta sẽ dựa vào Clustroid (đối với các kiểu dữ liệu) hoặc Centroid (đối với dữ liệu số) đây là cách để so khoảng cách giữa các điểm dữ liệu sau mỗi lần phân cụm. Khi nào Clustroid hoặc Centroid không đổi thì ta sẽ tìm ra được các cụm chính xác.

- **Đối với dataset 2019**

Để lựa chọn k cụm chính xác nhất, nhóm tiến hành sử dụng phương pháp Elbow:

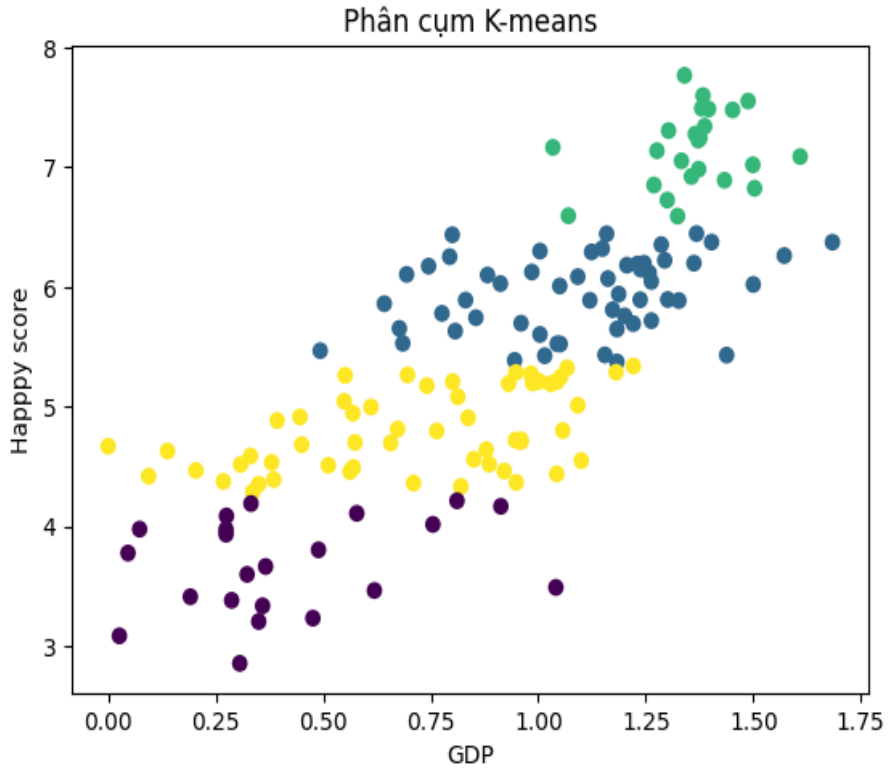
- Do ở Dataset 2019 nhóm nhận thấy GDP tác động vào Điểm hạnh phúc nhiều nhất nên nhóm quyết định gom cụm theo 2 thuộc tính trên và áp dụng hai thuộc tính trên cho các dataset còn lại vì theo GDP qua hầu hết các năm đều tác động mạnh vào Điểm hạnh phúc.



=> Dựa vào kết quả trên, nhóm chọn k = 4.

- Kết quả sau khi gom cụm như sau:

-



- Quan sát biểu đồ trên chúng ta thấy có 4 cụm:
  - **Cụm màu tím (cluster = 0)**: bao gồm các nước có điểm GDP ở mức thấp và điểm hạnh phúc cũng ở mức thấp:

	Country	Happy score	GDP	Social support	Healthy life expectancy	Freedom to make life choices	Generosity	Perceptions of corruption	Cluster
140	Liberia	3.975	0.073	0.922	0.443	0.370	0.233	0.033	0
141	Comoros	3.973	0.274	0.757	0.505	0.142	0.275	0.078	0
142	Madagascar	3.933	0.274	0.916	0.555	0.148	0.169	0.041	0
143	Lesotho	3.802	0.489	1.169	0.168	0.359	0.107	0.093	0
144	Burundi	3.775	0.046	0.447	0.380	0.220	0.176	0.180	0
145	Zimbabwe	3.663	0.366	1.114	0.433	0.361	0.151	0.089	0
146	Haiti	3.597	0.323	0.688	0.449	0.026	0.419	0.110	0
148	Syria	3.462	0.619	0.378	0.440	0.013	0.331	0.141	0
149	Malawi	3.410	0.191	0.560	0.495	0.443	0.218	0.089	0
150	Yemen	3.380	0.287	1.163	0.463	0.143	0.108	0.077	0
151	Rwanda	3.334	0.359	0.711	0.614	0.555	0.217	0.411	0
152	Tanzania	3.231	0.476	0.885	0.499	0.417	0.276	0.147	0
153	Afghanistan	3.203	0.350	0.517	0.361	0.000	0.158	0.025	0
154	Central African Republic	3.083	0.026	0.000	0.105	0.225	0.235	0.035	0

- **Cụm màu xanh dương(cluster = 1):** bao gồm các nước có điểm GDP và Điểm hạnh phúc ở mức trung bình đến mức khá:

	Country	Happy score	GDP	Social support	Healthy life expectancy	Freedom to make life choices	Generosity	Perceptions of corruption	Cluster
24	Taiwan	6.446	1.368	1.430	0.914	0.351	0.242	0.097	1
25	Chile	6.444	1.159	1.369	0.920	0.357	0.187	0.056	1
26	Guatemala	6.436	0.800	1.269	0.746	0.535	0.175	0.078	1
27	Saudi Arabia	6.375	1.403	1.357	0.795	0.439	0.080	0.132	1
28	Qatar	6.374	1.684	1.313	0.871	0.555	0.220	0.167	1
29	Spain	6.354	1.286	1.484	1.062	0.362	0.153	0.079	1
30	Panama	6.321	1.149	1.442	0.910	0.516	0.109	0.054	1
31	Brazil	6.300	1.004	1.439	0.802	0.390	0.099	0.086	1
32	Uruguay	6.293	1.124	1.465	0.891	0.523	0.127	0.150	1
33	Singapore	6.262	1.572	1.463	1.141	0.556	0.271	0.453	1
34	El Salvador	6.253	0.794	1.242	0.789	0.430	0.093	0.074	1
35	Italy	6.223	1.294	1.488	1.039	0.231	0.158	0.030	1
36	Bahrain	6.199	1.362	1.368	0.871	0.536	0.255	0.110	1
37	Slovakia	6.198	1.246	1.504	0.881	0.334	0.121	0.014	1
38	Trinidad & Tobago	6.192	1.231	1.477	0.713	0.489	0.185	0.016	1
39	Poland	6.182	1.206	1.438	0.884	0.483	0.117	0.050	1
40	Uzbekistan	6.174	0.745	1.529	0.756	0.631	0.322	0.240	1
41	Lithuania	6.149	1.238	1.515	0.818	0.291	0.043	0.042	1
42	Colombia	6.125	0.985	1.410	0.841	0.470	0.099	0.034	1
43	Slovenia	6.118	1.258	1.523	0.953	0.564	0.144	0.057	1
44	Nicaragua	6.105	0.694	1.325	0.835	0.435	0.200	0.127	1
45	Kosovo	6.100	0.882	1.232	0.758	0.489	0.262	0.006	1
46	Argentina	6.086	1.092	1.432	0.881	0.471	0.066	0.050	1
47	Romania	6.070	1.162	1.232	0.825	0.462	0.083	0.005	1
48	Cyprus	6.046	1.263	1.223	1.042	0.406	0.190	0.041	1
49	Ecuador	6.028	0.912	1.312	0.868	0.498	0.126	0.087	1
50	Kuwait	6.021	1.500	1.319	0.808	0.493	0.142	0.097	1
51	Thailand	6.008	1.050	1.409	0.828	0.557	0.359	0.028	1

- **Cụm màu xanh lá(cluster = 2):** bao gồm các nước có điểm GDP và Điểm hạnh phúc ở mức khá đến mức cao:

	Country	Happy score	GDP	Social support	Healthy life expectancy	Freedom to make life choices	Generosity	Perceptions of corruption	Cluster
0	Finland	7.769	1.340	1.587	0.986	0.596	0.153	0.393	2
1	Denmark	7.600	1.383	1.573	0.996	0.592	0.252	0.410	2
2	Norway	7.554	1.488	1.582	1.028	0.603	0.271	0.341	2
3	Iceland	7.494	1.380	1.624	1.026	0.591	0.354	0.118	2
4	Netherlands	7.488	1.396	1.522	0.999	0.557	0.322	0.298	2
5	Switzerland	7.480	1.452	1.526	1.052	0.572	0.263	0.343	2
6	Sweden	7.343	1.387	1.487	1.009	0.574	0.267	0.373	2
7	New Zealand	7.307	1.303	1.557	1.026	0.585	0.330	0.380	2
8	Canada	7.278	1.365	1.505	1.039	0.584	0.285	0.308	2
9	Austria	7.246	1.376	1.475	1.016	0.532	0.244	0.226	2
10	Australia	7.228	1.372	1.548	1.036	0.557	0.332	0.290	2
11	Costa Rica	7.167	1.034	1.441	0.963	0.558	0.144	0.093	2
12	Israel	7.139	1.276	1.455	1.029	0.371	0.261	0.082	2
13	Luxembourg	7.090	1.609	1.479	1.012	0.526	0.194	0.316	2
14	United Kingdom	7.054	1.333	1.538	0.996	0.450	0.348	0.278	2
15	Ireland	7.021	1.499	1.553	0.999	0.516	0.298	0.310	2
16	Germany	6.985	1.373	1.454	0.987	0.495	0.261	0.265	2

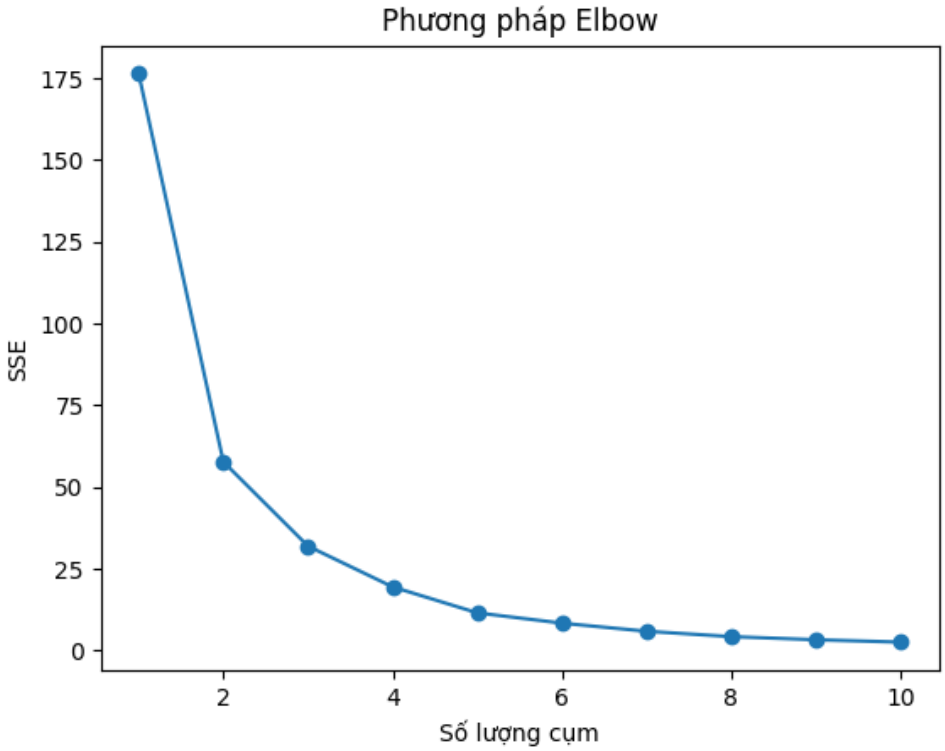
17	Belgium	6.923	1.356	1.504	0.986	0.473	0.160	0.210	2
18	United States	6.892	1.433	1.457	0.874	0.454	0.280	0.128	2
19	Czech Republic	6.852	1.269	1.487	0.920	0.457	0.046	0.036	2
20	United Arab Emirates	6.825	1.503	1.310	0.825	0.598	0.262	0.182	2
21	Malta	6.726	1.300	1.520	0.999	0.564	0.375	0.151	2
22	Mexico	6.595	1.070	1.323	0.861	0.433	0.074	0.073	2
23	France	6.592	1.324	1.472	1.045	0.436	0.111	0.183	2

- **Cụm màu vàng(cluster = 3):** bao gồm các nước có GDP và Điểm hạnh phúc ở mức thấp đến trung bình:

	Country	Happy score	GDP	Social support	Healthy life expectancy	Freedom to make life choices	Generosity	Perceptions of corruption	Cluster
79	Malaysia	5.339	1.221	1.171	0.828	0.508	0.260	0.024	3
80	Belarus	5.323	1.067	1.465	0.789	0.235	0.094	0.142	3
81	Greece	5.287	1.181	1.156	0.999	0.067	0.000	0.034	3
82	Mongolia	5.285	0.948	1.531	0.667	0.317	0.235	0.038	3
83	North Macedonia	5.274	0.983	1.294	0.838	0.345	0.185	0.034	3
84	Nigeria	5.265	0.696	1.111	0.245	0.426	0.215	0.041	3
85	Kyrgyzstan	5.261	0.551	1.438	0.723	0.508	0.300	0.023	3
86	Turkmenistan	5.247	1.052	1.538	0.657	0.394	0.244	0.028	3
87	Algeria	5.211	1.002	1.160	0.785	0.086	0.073	0.114	3
88	Morocco	5.208	0.801	0.782	0.782	0.418	0.036	0.076	3
89	Azerbaijan	5.208	1.043	1.147	0.769	0.351	0.035	0.182	3
90	Lebanon	5.197	0.987	1.224	0.815	0.216	0.166	0.027	3
91	Indonesia	5.192	0.931	1.203	0.660	0.491	0.498	0.028	3
92	China	5.191	1.029	1.125	0.893	0.521	0.058	0.100	3
93	Vietnam	5.175	0.741	1.346	0.851	0.543	0.147	0.073	3
106	Albania	4.719	0.947	0.848	0.874	0.383	0.178	0.027	3
107	Venezuela	4.707	0.960	1.427	0.805	0.154	0.064	0.047	3
108	Cambodia	4.700	0.574	1.122	0.637	0.609	0.232	0.062	3
109	Palestinian Territories	4.696	0.657	1.247	0.672	0.225	0.103	0.066	3
110	Senegal	4.681	0.450	1.134	0.571	0.292	0.153	0.072	3
111	Somalia	4.668	0.000	0.698	0.268	0.559	0.243	0.270	3
112	Namibia	4.639	0.879	1.313	0.477	0.401	0.070	0.056	3
113	Niger	4.628	0.138	0.774	0.366	0.318	0.188	0.102	3
114	Burkina Faso	4.587	0.331	1.056	0.380	0.255	0.177	0.113	3
115	Armenia	4.559	0.850	1.055	0.815	0.283	0.095	0.064	3
116	Iran	4.548	1.100	0.842	0.785	0.305	0.270	0.125	3
117	Guinea	4.534	0.380	0.829	0.375	0.332	0.207	0.086	3
118	Georgia	4.519	0.886	0.666	0.752	0.346	0.043	0.164	3
119	Gambia	4.516	0.308	0.939	0.428	0.382	0.269	0.167	3
120	Kenya	4.509	0.512	0.983	0.581	0.431	0.372	0.053	3
121	Mauritania	4.490	0.570	1.167	0.489	0.066	0.106	0.088	3
122	Mozambique	4.466	0.204	0.986	0.390	0.494	0.197	0.138	3
123	Tunisia	4.461	0.921	1.000	0.815	0.167	0.059	0.055	3
124	Bangladesh	4.456	0.562	0.928	0.723	0.527	0.166	0.143	3
125	Iraq	4.437	1.043	0.980	0.574	0.241	0.148	0.089	3
126	Congo (Kinshasa)	4.418	0.094	1.125	0.357	0.269	0.212	0.053	3
127	Mali	4.390	0.385	1.105	0.308	0.327	0.153	0.052	3
128	Sierra Leone	4.374	0.268	0.841	0.242	0.309	0.252	0.045	3
129	Sri Lanka	4.366	0.949	1.265	0.831	0.470	0.244	0.047	3

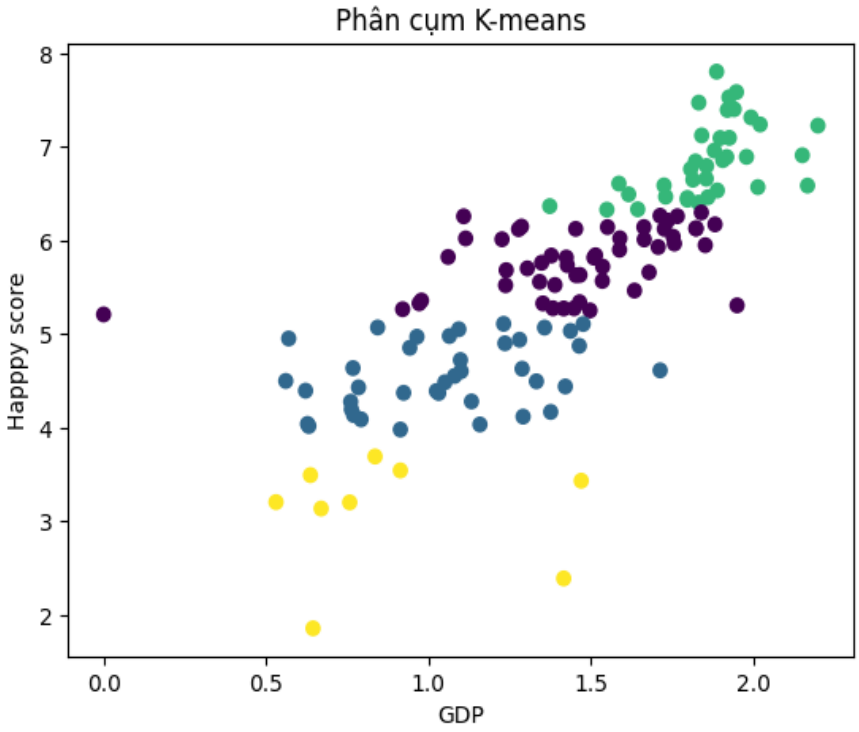
- **Đối với dataset 2023**

Để lựa chọn k cụm chính xác nhất, nhóm tiến hành sử dụng phương pháp Elbow:



=> Dựa vào kết quả trên, nhóm chọn k = 4.

- Kết quả sau khi gom cụm như sau:



- Quan sát biểu đồ trên ta thấy các nước được chia theo 4 cụm dựa theo GDP và Happy score.

- **Cụm màu vàng(cluster = 3)**: bao gồm các nước có GDP, Điểm hạnh phúc ở mức thấp nhất (Điểm số GDP dao động từ 0.531 đến 1.471, điểm số hạnh phúc dao động từ 1.859 đến 3.694).
    - Tình hình lạm phát, nội tình bên trong quốc gia không bao giờ ổn định. Không những chỉ số GDP thấp mà ngoài ra đó còn có các chỉ số khác cũng mở mức thấp nhất như là:
    - Tự do lựa chọn trong cuộc sống, Hỗ trợ xã hội, Trợ cấp của chính phủ, xã hội),... Đời sống người dân vẫn chưa được cải thiện, được quan tâm đúng mực

	Country	Happy score	GDP	Social support	Healthy life expectancy	Freedom to make life choices	Generosity	Perceptions of corruption	Cluster
128	Tanzania	3.694	0.836	0.787	0.214	0.607	0.234	0.269	3
129	Comoros	3.545	0.914	0.327	0.215	0.117	0.129	0.145	3
130	Malawi	3.495	0.637	0.479	0.189	0.490	0.139	0.129	3
131	Botswana	3.435	1.471	1.041	0.087	0.480	0.021	0.071	3
132	Congo (Kinshasa)	3.207	0.531	0.784	0.105	0.375	0.183	0.068	3
133	Zimbabwe	3.204	0.758	0.881	0.069	0.363	0.112	0.117	3
134	Sierra Leone	3.138	0.670	0.540	0.092	0.371	0.193	0.051	3
135	Lebanon	2.392	1.417	0.476	0.398	0.123	0.061	0.027	3
136	Afghanistan	1.859	0.645	0.000	0.087	0.000	0.093	0.059	3

- **Cụm màu xanh dương (cluster = 1):** bao gồm các nước có GDP ở mức trung bình, Điểm hạnh phúc ở mức thấp đến trung bình (Điểm số GDP dao động từ 0.561 đến 1.714, điểm số hạnh phúc dao động từ 3.982 đến 5.111).
    - Chúng ta có thể thấy rằng những nước này là các nước đang có chiến tranh diễn ra đa số ở Châu Phi (Ethiopia, Nigeria, Gabon…) nhưng tiêu biểu phải kể đến đó là Iraq, Ukraine. Nhưng cũng có một số quốc gia ở đông Nam Á như: Cambodia, Laos, Myanmar. Các quốc gia trên có mức GDP trung bình do là các nước đang phát triển hoặc là nội tình bên trong quốc gia không ổn định khiến đời sống người dân còn bấp bênh.

	Country	Happy score	GDP	Social support	Healthy life expectancy	Freedom to make life choices	Generosity	Perceptions of corruption	Cluster
88	Laos	5.111	1.232	0.853	0.257	0.715	0.185	0.162	1
89	Georgia	5.109	1.477	0.947	0.366	0.539	0.000	0.201	1
90	Guinea	5.072	0.844	0.776	0.072	0.369	0.204	0.102	1
91	Ukraine	5.071	1.358	1.354	0.355	0.551	0.265	0.016	1
92	Ivory Coast	5.053	1.094	0.584	0.120	0.467	0.138	0.131	1
93	Gabon	5.035	1.438	1.021	0.183	0.346	0.036	0.102	1
94	Nigeria	4.981	1.065	1.007	0.092	0.448	0.176	0.013	1
95	Cameroon	4.973	0.965	0.871	0.118	0.405	0.144	0.059	1
96	Mozambique	4.954	0.570	0.885	0.000	0.625	0.161	0.192	1
97	Iraq	4.941	1.281	0.953	0.324	0.351	0.134	0.038	1
99	Morocco	4.903	1.236	0.535	0.337	0.540	0.013	0.085	1
100	Iran	4.876	1.465	1.102	0.411	0.281	0.229	0.130	1
101	Senegal	4.855	0.943	0.727	0.231	0.519	0.142	0.060	1
102	Mauritania	4.724	1.099	0.764	0.244	0.320	0.130	0.195	1
103	Burkina Faso	4.638	0.768	0.814	0.107	0.419	0.188	0.113	1
104	Namibia	4.631	1.289	1.126	0.145	0.383	0.069	0.071	1
105	Turkiye	4.614	1.714	1.148	0.467	0.125	0.095	0.096	1
106	Ghana	4.605	1.101	0.756	0.197	0.526	0.211	0.035	1
107	Pakistan	4.555	1.081	0.657	0.158	0.511	0.141	0.102	1
108	Niger	4.501	0.561	0.628	0.137	0.540	0.154	0.140	1
109	Tunisia	4.497	1.333	0.981	0.422	0.259	0.022	0.016	1
110	Kenya	4.487	1.051	0.881	0.190	0.418	0.291	0.055	1
111	Sri Lanka	4.442	1.422	1.224	0.426	0.539	0.120	0.086	1
112	Uganda	4.432	0.785	1.144	0.201	0.425	0.197	0.051	1
113	Chad	4.397	0.622	0.962	0.043	0.393	0.255	0.088	1
114	Cambodia	4.393	1.025	1.024	0.283	0.768	0.176	0.051	1
115	Benin	4.374	0.924	0.242	0.124	0.481	0.114	0.253	1
116	Myanmar	4.372	1.032	1.125	0.269	0.460	0.400	0.194	1
117	Bangladesh	4.282	1.133	0.513	0.355	0.617	0.139	0.165	1
118	Gambia	4.279	0.761	0.614	0.174	0.286	0.332	0.033	1
119	Mali	4.198	0.763	0.637	0.106	0.441	0.121	0.059	1
120	Egypt	4.170	1.377	0.972	0.326	0.467	0.038	0.250	1
121	Togo	4.137	0.770	0.642	0.161	0.367	0.149	0.136	1
122	Jordan	4.120	1.292	0.980	0.438	0.517	0.056	0.173	1
123	Ethiopia	4.091	0.793	1.114	0.250	0.451	0.283	0.101	1
124	Liberia	4.042	0.628	0.644	0.141	0.471	0.219	0.071	1
125	India	4.036	1.159	0.674	0.252	0.685	0.175	0.111	1
126	Madagascar	4.019	0.632	0.779	0.178	0.187	0.177	0.134	1
127	Zambia	3.982	0.914	0.890	0.095	0.545	0.189	0.080	1

- **Cụm màu tím (cluster = 0):** bao gồm các nước có GDP ở mức khá cao, Điểm hạnh phúc ở mức trung bình đến mức khá (Điểm số GDP dao động từ 0.0 đến 1.951, điểm hạnh phúc dao động từ 5.211 đến 6.3).
  - Chúng ta có thể thấy rằng những nước này là các nước đang phát triển tiêu biểu kể đến đó là Việt Nam và các quốc gia phát triển có thể kể đến như là Hàn Quốc, Trung Quốc, Hy Lạp (Greece), Thái Lan,… Đời sống của nhân dân ở các nước này đa số là ổn định, chính phủ khá quan tâm và chăm lo cho đời sống nhân dân, quyền tự do dân chủ được đảm bảo. Nhưng vẫn có một nước có GDP ở mức 0 nhưng điểm hạnh phúc vẫn ở mức khá cao

đó là Venezuela. Tình hình bên trong Venezuela chưa bao giờ ổn định, chiến tranh liên miên, lạm phát thì cao, đời sống nhân dân khổ cực. Nhưng do Venezuela có khá nhiều hoa hậu thắng các giải thưởng quốc tế ở mức thành tích cao nên mới có chuyện là điểm hạnh phúc cao.

81	Hong Kong S.A.R. of China	5.308	1.951	1.201	0.702	0.407	0.123	0.390	0
82	Albania	5.277	1.449	0.951	0.480	0.549	0.133	0.037	0
83	Indonesia	5.277	1.384	1.169	0.314	0.663	0.422	0.038	0
84	South Africa	5.275	1.417	1.428	0.149	0.464	0.090	0.019	0
85	Congo (Brazzaville)	5.267	0.921	0.665	0.145	0.464	0.134	0.136	0
86	North Macedonia	5.254	1.498	1.171	0.408	0.515	0.207	0.020	0
87	Venezuela	5.211	0.000	1.257	0.341	0.369	0.205	0.084	0
40	Latvia	6.213	1.737	1.505	0.405	0.580	0.107	0.071	0
41	Bahrain	6.173	1.883	1.269	0.389	0.748	0.199	0.138	0
42	Guatemala	6.150	1.287	1.188	0.310	0.631	0.106	0.066	0
43	Kazakhstan	6.144	1.664	1.491	0.389	0.628	0.136	0.149	0
44	Serbia	6.144	1.552	1.343	0.424	0.617	0.246	0.081	0
45	Cyprus	6.130	1.824	1.224	0.580	0.455	0.104	0.050	0
46	Japan	6.129	1.825	1.396	0.622	0.556	0.009	0.207	0
47	Croatia	6.125	1.727	1.455	0.475	0.500	0.087	0.003	0
48	Brazil	6.125	1.454	1.250	0.387	0.558	0.131	0.137	0
49	El Salvador	6.122	1.278	1.044	0.383	0.713	0.079	0.222	0
50	Hungary	6.041	1.754	1.519	0.435	0.501	0.105	0.065	0
51	Argentina	6.024	1.590	1.388	0.427	0.587	0.088	0.082	0
52	Honduras	6.023	1.115	1.072	0.341	0.613	0.189	0.062	0
53	Uzbekistan	6.014	1.227	1.347	0.375	0.740	0.260	0.208	0
54	Malaysia	6.012	1.665	1.155	0.385	0.659	0.222	0.122	0
55	Portugal	5.968	1.758	1.356	0.537	0.693	0.031	0.037	0
56	South Korea	5.951	1.853	1.188	0.603	0.446	0.112	0.163	0
57	Greece	5.931	1.708	1.247	0.535	0.248	0.008	0.097	0
58	Mauritius	5.902	1.589	1.382	0.336	0.574	0.121	0.110	0
59	Thailand	5.843	1.515	1.344	0.461	0.624	0.291	0.013	0
60	Mongolia	5.840	1.379	1.494	0.244	0.425	0.239	0.058	0
61	Kyrgyzstan	5.825	1.061	1.439	0.417	0.735	0.234	0.018	0
62	Moldova	5.819	1.425	1.302	0.375	0.610	0.093	0.020	0
63	China	5.818	1.510	1.249	0.468	0.666	0.115	0.145	0
64	Vietnam	5.763	1.349	1.212	0.381	0.741	0.134	0.122	0
65	Paraguay	5.738	1.428	1.427	0.392	0.678	0.148	0.062	0
65	Paraguay	5.738	1.428	1.427	0.392	0.678	0.148	0.062	0
66	Montenegro	5.722	1.537	1.385	0.424	0.563	0.170	0.061	0
67	Jamaica	5.703	1.305	1.329	0.411	0.587	0.079	0.039	0
68	Bolivia	5.684	1.240	1.187	0.329	0.648	0.103	0.060	0
69	Russia	5.661	1.680	1.383	0.366	0.449	0.120	0.091	0
70	Bosnia and Herzegovina	5.633	1.467	1.361	0.429	0.485	0.247	0.008	0
71	Colombia	5.630	1.455	1.213	0.486	0.562	0.080	0.068	0
72	Dominican Republic	5.569	1.536	1.227	0.351	0.623	0.083	0.195	0
73	Ecuador	5.559	1.343	1.173	0.476	0.560	0.079	0.069	0
74	Peru	5.526	1.390	1.153	0.499	0.549	0.073	0.027	0
75	Philippines	5.523	1.238	1.108	0.286	0.714	0.104	0.141	0
76	Bulgaria	5.466	1.635	1.457	0.408	0.557	0.106	0.013	0
77	Nepal	5.360	0.979	1.027	0.281	0.567	0.215	0.104	0
78	Armenia	5.342	1.466	1.134	0.443	0.551	0.053	0.160	0
79	Tajikistan	5.330	0.972	1.248	0.291	0.599	0.104	0.292	0
80	Algeria	5.329	1.353	1.298	0.409	0.252	0.073	0.152	0

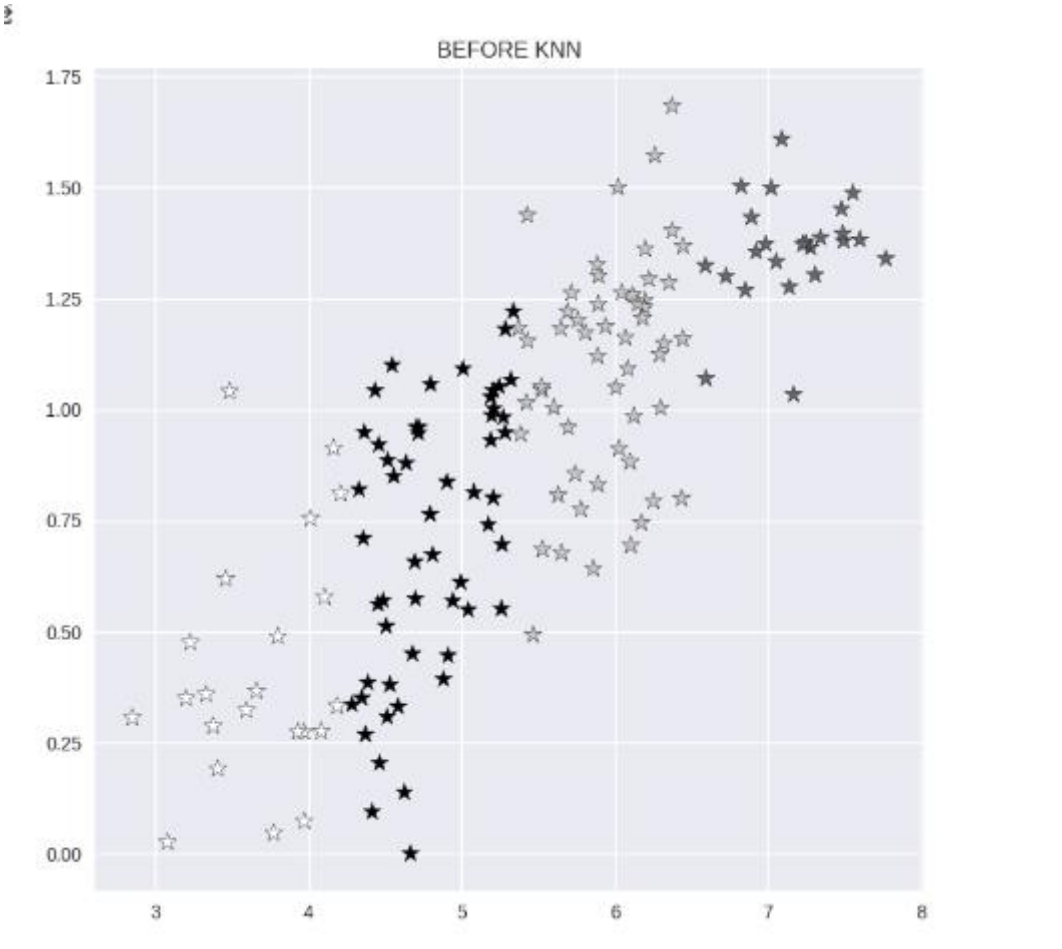
- **Cụm màu xanh lá (cluster = 2):** bao gồm các nước có GDP ở mức cao, Điểm hạnh phúc ở mức cao (Điểm số GDP dao động từ 1.374 đến 2.2, điểm hạnh phúc dao động từ 6.33 đến 7.804).
    - Các nước ở cụm xanh lá là các nước phát triển có thể kể đến như là Anh, Pháp, Mỹ, Finland,... Đời sống ở nhân dân các nước trên ở mức cao, chính phủ quan tâm chăm sóc tốt cho nhân dân, nội tình đất nước ổn định không có nhiều biến động.

	Country	Happy score	GDP	Social support	Healthy life expectancy	Freedom to make life choices	Generosity	Perceptions of corruption	Cluster
0	Finland	7.804	1.888	1.585	0.535	0.772	0.126	0.535	2
1	Denmark	7.586	1.949	1.548	0.537	0.734	0.208	0.525	2
2	Iceland	7.530	1.926	1.620	0.559	0.738	0.250	0.187	2
3	Israel	7.473	1.833	1.521	0.577	0.569	0.124	0.158	2
4	Netherlands	7.403	1.942	1.488	0.545	0.672	0.251	0.394	2
5	Sweden	7.395	1.921	1.510	0.562	0.754	0.225	0.520	2
6	Norway	7.315	1.994	1.521	0.544	0.752	0.212	0.463	2
7	Switzerland	7.240	2.022	1.463	0.582	0.678	0.151	0.475	2
8	Luxembourg	7.228	2.200	1.357	0.549	0.710	0.149	0.418	2
9	New Zealand	7.123	1.842	1.544	0.513	0.672	0.230	0.471	2
10	Austria	7.097	1.927	1.382	0.535	0.630	0.191	0.310	2
11	Australia	7.095	1.899	1.497	0.532	0.677	0.242	0.310	2
12	Canada	6.961	1.881	1.484	0.541	0.656	0.218	0.364	2
13	Ireland	6.911	2.152	1.425	0.539	0.656	0.186	0.409	2
14	United States	6.894	1.980	1.460	0.390	0.557	0.210	0.172	2
15	Germany	6.892	1.919	1.401	0.539	0.618	0.153	0.365	2
16	Belgium	6.859	1.907	1.449	0.528	0.590	0.137	0.273	2
17	Czechia	6.845	1.823	1.544	0.477	0.693	0.158	0.050	2
18	United Kingdom	6.796	1.857	1.366	0.511	0.626	0.272	0.340	2
19	Lithuania	6.763	1.808	1.511	0.432	0.487	0.059	0.089	2
20	France	6.661	1.856	1.433	0.566	0.582	0.083	0.270	2
21	Slovenia	6.650	1.815	1.539	0.532	0.707	0.144	0.113	2
22	Costa Rica	6.609	1.587	1.340	0.503	0.683	0.099	0.116	2
23	Romania	6.589	1.726	1.280	0.423	0.631	0.044	0.000	2
24	Singapore	6.587	2.168	1.354	0.607	0.660	0.170	0.561	2
25	United Arab Emirates	6.571	2.015	1.223	0.401	0.745	0.188	0.247	2
26	Taiwan Province of China	6.535	1.890	1.372	0.492	0.562	0.067	0.178	2
27	Uruguay	6.494	1.617	1.445	0.435	0.683	0.102	0.254	2
28	Slovakia	6.469	1.731	1.544	0.472	0.494	0.128	0.022	2
29	Saudi Arabia	6.463	1.861	1.370	0.351	0.682	0.093	0.170	2
30	Estonia	6.455	1.798	1.526	0.494	0.728	0.153	0.372	2
31	Spain	6.436	1.798	1.491	0.567	0.533	0.101	0.157	2
32	Italy	6.405	1.832	1.365	0.559	0.438	0.097	0.063	2
33	Kosovo	6.368	1.374	1.269	0.372	0.639	0.275	0.045	2
34	Chile	6.334	1.645	1.384	0.511	0.546	0.131	0.076	2
35	Mexico	6.330	1.550	1.169	0.389	0.632	0.086	0.115	2

4.3. Phân lớp theo KNN

KNN là thuật toán phân lớp dựa vào các nhãn dữ liệu. Thuật toán sẽ chia tập dữ liệu thành hai bộ train và test (với điều kiện là train và test phải có đầy đủ nhãn). Thuật toán sẽ đưa ra các dữ liệu dự đoán dựa trên tập dữ liệu train và các chỉ số k để xem mức độ chính xác. Mức độ chính xác càng cao thì sai số giữa tập dữ liệu dự đoán và tập dữ liệu test là rất ít

- **Đối với dataset 2019**
  - Nhóm quyết định sử dụng lại nhãn Cluster (Cụm) đã được dán khi thực hiện thuật toán gom cụm K Mean đối với bộ dataset.
  - Trực quan hóa bộ dữ liệu trước khi chia ra bộ dữ liệu train và test.



  - Có thể thấy bộ dữ liệu ban đầu được chia theo 4 nhãn Cluster. Ở biểu đồ trên ta có thể thấy rằng có một vài data point (điểm dữ liệu) tuy ở nhãn này nhưng vị trí của data point đó bị lạc qua nhãn khác.
  - Nhóm chia dataset 2019 thành hai bộ dữ liệu train và test:

+ Bộ train:

Happy scc	GDP	Social sup	Healthy li	Freedom	Generosit	Perceptio	Cluster
5.247	1.052	1.538	0.657	0.394	0.244	0.028	3
3.231	0.476	0.885	0.499	0.417	0.276	0.147	0
4.49	0.57	1.167	0.489	0.066	0.106	0.088	3
7.6	1.383	1.573	0.996	0.592	0.252	0.41	2
5.211	1.002	1.16	0.785	0.086	0.073	0.114	3
4.799	1.057	1.183	0.571	0.295	0.043	0.055	3
4.418	0.094	1.125	0.357	0.269	0.212	0.053	3
5.265	0.696	1.111	0.245	0.426	0.215	0.041	3
4.107	0.578	1.058	0.426	0.431	0.247	0.087	0
7.769	1.34	1.587	0.986	0.596	0.153	0.393	2
4.461	0.921	1	0.815	0.167	0.059	0.055	3
6.125	0.985	1.41	0.841	0.47	0.099	0.034	1
5.467	0.493	1.098	0.718	0.389	0.23	0.144	1
4.719	0.947	0.848	0.874	0.383	0.178	0.027	3
4.085	0.275	0.572	0.41	0.293	0.177	0.085	0
7.554	1.488	1.582	1.028	0.603	0.271	0.341	2
6.192	1.231	1.477	0.713	0.489	0.185	0.016	1
3.802	0.489	1.169	0.168	0.359	0.107	0.093	0
4.189	0.332	1.069	0.443	0.356	0.252	0.06	0
5.287	1.181	1.156	0.999	0.067	0	0.034	3
5.809	1.173	1.508	0.729	0.41	0.146	0.096	1
6.223	1.294	1.488	1.039	0.231	0.158	0.03	1
4.639	0.879	1.313	0.477	0.401	0.07	0.056	3
4.906	0.837	1.225	0.815	0.383	0.11	0.13	3
4.707	0.96	1.427	0.805	0.154	0.064	0.047	3
4.681	0.45	1.134	0.571	0.292	0.153	0.072	3
4.166	0.913	1.039	0.644	0.241	0.076	0.067	0
4.286	0.336	1.033	0.532	0.344	0.209	0.1	3
3.488	1.041	1.145	0.538	0.455	0.025	0.1	0
4.332	0.82	1.39	0.739	0.178	0.187	0.01	3
7.054	1.333	1.538	0.996	0.45	0.348	0.278	2
4.559	0.85	1.055	0.815	0.283	0.095	0.064	3
3.38	0.287	1.163	0.463	0.143	0.108	0.077	0
5.718	1.263	1.252	1.042	0.417	0.191	0.162	1

+ Bộ test:

Happy scc	GDP	Social sup	Healthy li	Freedom t	Generosit	Perceptio	Cluster
6.021	1.5	1.319	0.808	0.493	0.142	0.097	1
5.323	1.067	1.465	0.789	0.235	0.094	0.142	3
5.529	0.685	1.328	0.739	0.245	0.181	0	1
7.278	1.365	1.505	1.039	0.584	0.285	0.308	2
5.648	1.183	1.452	0.726	0.334	0.082	0.031	1
3.597	0.323	0.688	0.449	0.026	0.419	0.11	0
7.48	1.452	1.526	1.052	0.572	0.263	0.343	2
4.913	0.446	1.226	0.677	0.439	0.285	0.089	3
6.446	1.368	1.43	0.914	0.351	0.242	0.097	1
7.494	1.38	1.624	1.026	0.591	0.354	0.118	2
5.386	0.945	1.212	0.845	0.212	0.263	0.006	1
4.668	0	0.698	0.268	0.559	0.243	0.27	3
6.293	1.124	1.465	0.891	0.523	0.127	0.15	1
6.825	1.503	1.31	0.825	0.598	0.262	0.182	2
4.7	0.574	1.122	0.637	0.609	0.232	0.062	3
5.208	0.801	0.782	0.782	0.418	0.036	0.076	3

- Bộ dữ liệu train và test đảm bảo đủ các nhãn đã được gán.
- Nhóm khởi tạo mô hình KNN với giá trị k = 1 (Đang huấn luyện mô hình với k =1), và sau đó thử test lại với bộ dữ liệu test. Nhóm thu được kết quả như sau:

  + Chúng ta sẽ tập trung chính vào chỉ số accuracy ở mỗi lần test.

```
KNN Model Evaluation with k =1:
[[1 0 0 0]
 [0 5 0 1]
 [0 0 4 0]
 [0 1 0 4]]
              precision    recall  f1-score   support

           0       1.00      1.00      1.00         1
           1       0.83      0.83      0.83         6
           2       1.00      1.00      1.00         4
           3       0.80      0.80      0.80         5

    accuracy                           0.88        16
   macro avg       0.91      0.91      0.91        16
weighted avg       0.88      0.88      0.88        16
```

=> k = 1 thì accuracy là 0.88

+ Với k = 3:

```
KNN Model Evaluation with k = 3 :
[[1 0 0 0]
 [0 5 0 1]
 [0 0 4 0]
 [0 1 0 4]]
              precision    recall  f1-score   support

           0       1.00      1.00      1.00         1
           1       0.83      0.83      0.83         6
           2       1.00      1.00      1.00         4
           3       0.80      0.80      0.80         5

    accuracy                           0.88        16
   macro avg       0.91      0.91      0.91        16
weighted avg       0.88      0.88      0.88        16
```

=> k = 3 thì accuracy là 0.88

+ Với k = 5:

```
KNN Model Evaluation with k = 5 :
[[1 0 0 0]
 [0 6 0 0]
 [0 0 4 0]
 [0 1 0 4]]
              precision    recall  f1-score   support

           0       1.00      1.00      1.00         1
           1       0.86      1.00      0.92         6
           2       1.00      1.00      1.00         4
           3       1.00      0.80      0.89         5

    accuracy                           0.94        16
   macro avg       0.96      0.95      0.95        16
weighted avg       0.95      0.94      0.94        16
```

=> k = 5 thì accuracy là 0.94

+ Với k = 9:

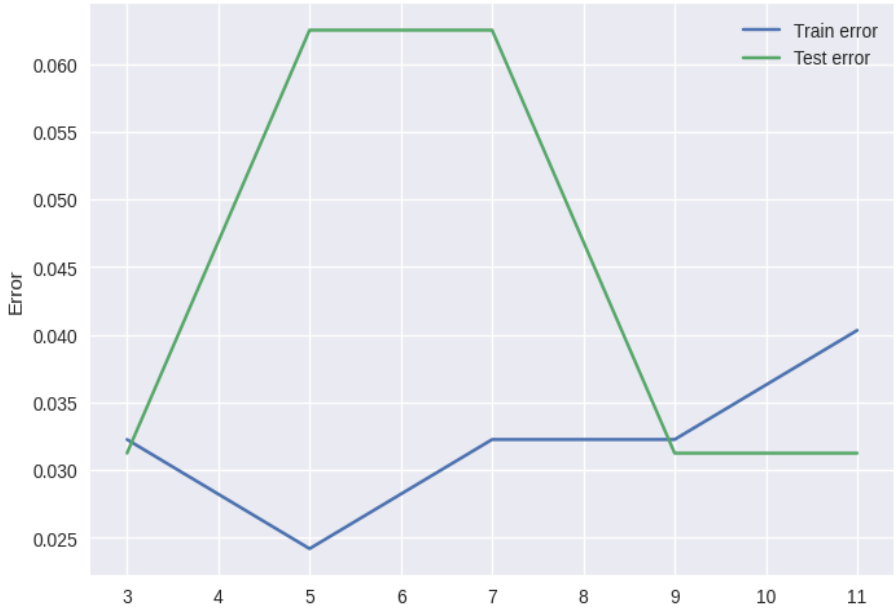
```
KNN Model Evaluation with k = 9 :
[[ 3  0  0  0]
 [ 0 11  0  0]
 [ 0  0  5  0]
 [ 0  1  0 12]]
              precision    recall  f1-score   support

           0       1.00      1.00      1.00         3
           1       0.92      1.00      0.96        11
           2       1.00      1.00      1.00         5
           3       1.00      0.92      0.96        13

    accuracy                           0.97        32
   macro avg       0.98      0.98      0.98        32
weighted avg       0.97      0.97      0.97        32
```

=> k = 9 thì accuracy là 0.97

- Đồ thị sai số và bảng test error, train error:



	K	Train Error	Test Error
0	3	0.032258	0.03125
1	5	0.024194	0.06250
2	7	0.032258	0.06250
3	9	0.032258	0.03125
4	11	0.040323	0.03125

- Để mô hình trên có tính hiệu quả cao thì thông qua những chỉ số k và accuracy cũng như là đồ thị sai số được đưa ra nhóm quyết định chọn k = 9 vì accuracy cao, quan sát đồ thị ta có thể thấy rằng k =9 trở đi thì test error trở nên rất ít. Chọn k = 9 để mô hình KNN trở nên hiệu quả.
- Kết quả gắn nhãn dựa trên dự đoán của thuật toán:

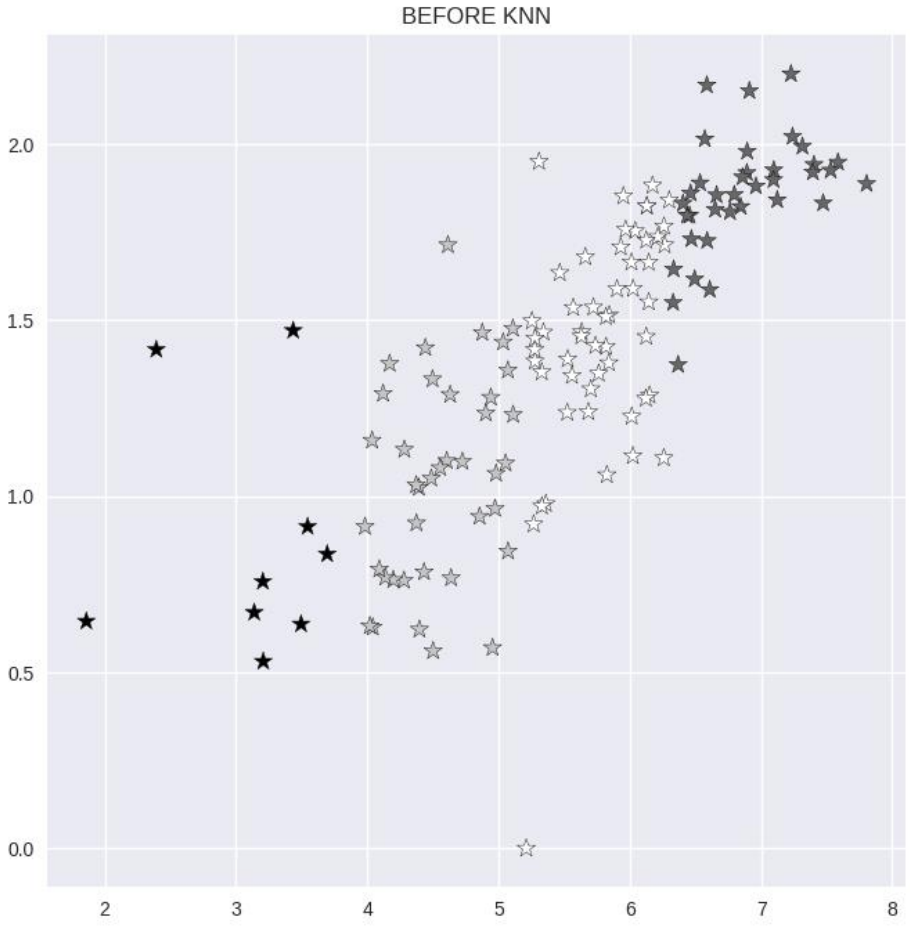
```
KNN Classification Results on New Data:
[3 1 1 1 3 1 3 3 3 0 2 3 2 3 1 1 3 3 0 1 3 2 2 2 1 3 1 1 1 3 1 0]
```

- Kết quả gán nhãn sau khi kiểm tra (dựa trên bộ dữ liệu test để kiểm tra lại):

96	3
69	1
82	3
76	1
114	3
29	1
94	3
132	3
93	3
139	0
19	2
90	3
15	2
125	3
24	1
30	1
119	3
101	3
152	0
78	1
98	3
18	2
12	2
9	2
31	1
104	3
68	1
55	1
51	1
97	3
45	1
147	0

- Như chúng ta có thể thấy với giá trị k = 9 ta có thể huấn luyện mô hình KNN hiệu quả vì sự sai khác giữa dự đoán cho nhãn gán (predict) và kiểm tra nhãn dựa trên bộ dữ liệu test là không quá nhiều (chỉ sai 1 vài chỗ).

- **Đối với dataset 2023**
  - Nhóm quyết định sử dụng lại nhãn Cluster (Cụm) đã được dán khi thực hiện thuật toán gom cụm K Mean đối với bộ dataset.



- Có thể thấy bộ dữ liệu ban đầu được chia theo 4 nhãn Cluster. Ở biểu đồ trên ta có thể thấy rằng có một vài data point (điểm dữ liệu) tuy ở nhãn này nhưng vị trí của data point đó bị lạc qua nhãn khác.
- Nhóm chia dataset 2023 thành hai bộ dữ liệu train và test:

+ Bộ train:

Happy sco	GDP	Social sup	Healthy li	Freedom t	Generosit	Perceptio	Cluster
5.277	1.449	0.951	0.48	0.549	0.133	0.037	0
4.091	0.793	1.114	0.25	0.451	0.283	0.101	1
5.703	1.305	1.329	0.411	0.587	0.079	0.039	0
5.277	1.384	1.169	0.314	0.663	0.422	0.038	0
5.211	0	1.257	0.341	0.369	0.205	0.084	0
4.555	1.081	0.657	0.158	0.511	0.141	0.102	1
4.036	1.159	0.674	0.252	0.685	0.175	0.111	1
7.473	1.833	1.521	0.577	0.569	0.124	0.158	2
6.405	1.832	1.365	0.559	0.438	0.097	0.063	2
4.374	0.924	0.242	0.124	0.481	0.114	0.253	1
5.053	1.094	0.584	0.12	0.467	0.138	0.131	1
4.605	1.101	0.756	0.197	0.526	0.211	0.035	1
6.894	1.98	1.46	0.39	0.557	0.21	0.172	2
5.308	1.951	1.201	0.702	0.407	0.123	0.39	0
5.33	0.972	1.248	0.291	0.599	0.104	0.292	0
7.53	1.926	1.62	0.559	0.738	0.25	0.187	2
5.818	1.51	1.249	0.468	0.666	0.115	0.145	0
6.125	1.454	1.25	0.387	0.558	0.131	0.137	0
4.198	0.763	0.637	0.106	0.441	0.121	0.059	1
3.495	0.637	0.479	0.189	0.49	0.139	0.129	3
4.876	1.465	1.102	0.411	0.281	0.229	0.13	1
4.137	0.77	0.642	0.161	0.367	0.149	0.136	1
3.204	0.758	0.881	0.069	0.363	0.112	0.117	3
6.65	1.815	1.539	0.532	0.707	0.144	0.113	2
6.961	1.881	1.484	0.541	0.656	0.218	0.364	2
4.614	1.714	1.148	0.467	0.125	0.095	0.096	1
5.722	1.537	1.385	0.424	0.563	0.17	0.061	0
6.463	1.861	1.37	0.351	0.682	0.093	0.17	2
5.684	1.24	1.187	0.329	0.648	0.103	0.06	0
4.501	0.561	0.628	0.137	0.54	0.154	0.14	1
6.436	1.798	1.491	0.567	0.533	0.101	0.157	2
5.36	0.979	1.027	0.281	0.567	0.215	0.104	0
4.432	0.785	1.144	0.201	0.425	0.197	0.051	1
5.559	1.343	1.173	0.476	0.56	0.079	0.069	0
6.26	1.767	1.474	0.477	0.511	0.12	0.139	0
3.138	0.67	0.54	0.092	0.371	0.193	0.051	3

+ Bộ test:

Happy sco	GDP	Social sup	Healthy li	Freedom	Generosit	Perceptio	Cluster
5.275	1.417	1.428	0.149	0.464	0.09	0.019	0
5.267	0.921	0.665	0.145	0.464	0.134	0.136	0
6.15	1.287	1.188	0.31	0.631	0.106	0.066	0
4.442	1.422	1.224	0.426	0.539	0.12	0.086	1
6.661	1.856	1.433	0.566	0.582	0.083	0.27	2
6.587	2.168	1.354	0.607	0.66	0.17	0.561	2
7.395	1.921	1.51	0.562	0.754	0.225	0.52	2
7.586	1.949	1.548	0.537	0.734	0.208	0.525	2
5.526	1.39	1.153	0.499	0.549	0.073	0.027	0
3.545	0.914	0.327	0.215	0.117	0.129	0.145	3
7.228	2.2	1.357	0.549	0.71	0.149	0.418	2
7.804	1.888	1.585	0.535	0.772	0.126	0.535	2
5.035	1.438	1.021	0.183	0.346	0.036	0.102	1
4.393	1.025	1.024	0.283	0.768	0.176	0.051	1

- Bộ dữ liệu train và test đảm bảo đủ các nhãn đã được gán
- Nhóm khởi tạo mô hình KNN với giá trị k = 1 (Đang huấn luyện mô hình với k =1), và sau đó thử test lại với bộ dữ liệu test. Nhóm thu được kết quả như sau:

  + Chúng ta sẽ tập trung chính vào chỉ số accuracy ở mỗi lần test.

```
KNN Model Evaluation with k = 1 :
[[6 3 0 0]
 [0 9 0 0]
 [0 0 9 0]
 [0 0 0 1]]
              precision    recall  f1-score   support

           0       1.00      0.67      0.80         9
           1       0.75      1.00      0.86         9
           2       1.00      1.00      1.00         9
           3       1.00      1.00      1.00         1

    accuracy                           0.89        28
   macro avg       0.94      0.92      0.91        28
weighted avg       0.92      0.89      0.89        28
```

  => k = 1 thì accuracy là 0.86

+ Với k = 3:

```
KNN Model Evaluation with k = 3 :
[[7 2 0 0]
 [0 9 0 0]
 [0 0 9 0]
 [0 0 0 1]]
              precision    recall  f1-score   support

           0       1.00      0.78      0.88         9
           1       0.82      1.00      0.90         9
           2       1.00      1.00      1.00         9
           3       1.00      1.00      1.00         1

    accuracy                           0.93        28
   macro avg       0.95      0.94      0.94        28
weighted avg       0.94      0.93      0.93        28
```

=> k = 3 thì accuracy là 0.93

+ Với k = 5:

```
KNN Model Evaluation with k = 5 :
[[7 2 0 0]
 [0 9 0 0]
 [0 0 9 0]
 [0 0 0 1]]
              precision    recall  f1-score   support

           0       1.00      0.78      0.88         9
           1       0.82      1.00      0.90         9
           2       1.00      1.00      1.00         9
           3       1.00      1.00      1.00         1

    accuracy                           0.93        28
   macro avg       0.95      0.94      0.94        28
weighted avg       0.94      0.93      0.93        28
```

=> k = 5 thì accuracy là 0.93

- Đồ thị sai số, bảng test error, train error:

	K	Train Error	Test Error
0	3	0.009259	0.071429
1	5	0.037037	0.071429
2	7	0.027778	0.071429
3	9	0.018519	0.107143
4	11	0.037037	0.142857

- Để mô hình trên có tính hiệu quả cao thì thông qua những chỉ số k , accuracy, cũng như là đồ thị sai số ta có thể thấy được là k = 3, k = 5, k = 7 thì sẽ đưa ra test error ở mức thấp nhất cũng như là chỉ số accuracy ở mức cao. Nhóm quyết định chọn k = 3 hoặc k = 5 hay k = 7 để mô hình KNN trở nên hiệu quả.
- Kết quả gắn nhãn dựa trên dự đoán của thuật toán:

```
KNN Classification Results on New Data:
[0 1 0 1 2 2 2 2 0 3 2 2 1 1 0 1 0 0 1 1 1 2 0 1 1 1 2 0]
```

- Kết quả gán nhãn sau khi kiểm tra (dựa trên bộ dữ liệu test để kiểm tra lại):

84	0
85	0
42	0
111	1
20	2
24	2
5	2
1	2
74	0
129	3
8	2
0	2
93	1
114	1
82	0
123	1
67	0
83	0
87	0
107	1
125	1
3	2
32	2
115	1
92	1
106	1
14	2
81	0

- Như chúng ta có thể thấy với giá trị k nằm trong khoảng 3 đến 7 ta có thể huấn luyện mô hình KNN hiệu quả vì sự sai khác giữa dự đoán cho nhãn gán (predict) và kiểm tra nhãn dựa trên bộ dữ liệu test là không quá nhiều (chỉ sai 1 vài chỗ).

Kết luận: Vậy có thể thấy thông qua thuật toán phân lớp KNN, ta có thể thấy được rằng các nhãn được gán cho các dòng dữ liệu khi gom cụm theo K Mean là chính xác. Thuật toán KNN này cũng giúp ta hình dung rằng GDP tác động vào Điểm hạnh phúc là rất nhiều.

4.4. Naive bayes

Naive Bayes là một phương pháp trong thống kê và Machine learning được sử dụng cho việc phân loại và dự đoán. Phương pháp này dựa trên Định lý Bayes và giả định "ngây thơ" (naive) rằng các đặc trưng đầu vào là độc lập với nhau khi đã biết lớp. Trong mô hình naive bayes được sử dụng đối với dataset Điểm hạnh phúc thì chúng ta đã biết được trước các lớp được phân lớp thông qua thuật toán KNN đã làm trước đó.

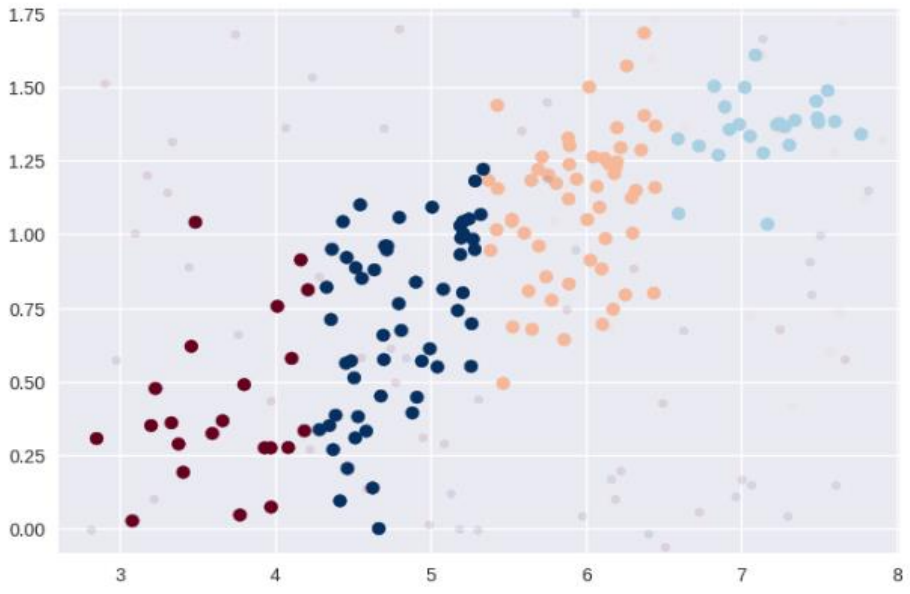
- **Đối với dataset 2019**

```
Naive-Bayes Model Evaluation on Test Data:
[[ 1  0  0  2]
 [ 0 11  0  0]
 [ 0  0  5  0]
 [ 0  0  0 13]]
              precision    recall  f1-score   support

           0       1.00      0.33      0.50         3
           1       1.00      1.00      1.00        11
           2       1.00      1.00      1.00         5
           3       0.87      1.00      0.93        13

    accuracy                           0.94        32
   macro avg       0.97      0.83      0.86        32
weighted avg       0.95      0.94      0.92        32
```

- Chúng ta thấy rằng thuật toán Naive bayes đều cho ra một giá trị chính xác (accuracy = 0.94) là một giá trị khá cao nhưng độ chính xác vẫn không cao bằng mô hình phân lớp KNN với k =9 (accuracy = 0.97).
- Trực quan hóa các lớp của mô hình Naive bayes:



- Có thể nhận thấy rằng mô hình Naive bayes này khi trực quan hóa lên cho ra kết quả gần giống như mô hình KNN vì accuracy của Naive bayes là 0.94 cũng là một tỷ lệ khá cao.
- Kết luận : đối với bộ dataset 2019 thì ta mô hình KNN với k =9 sẽ là tốt nhất cho việc phân lớp.

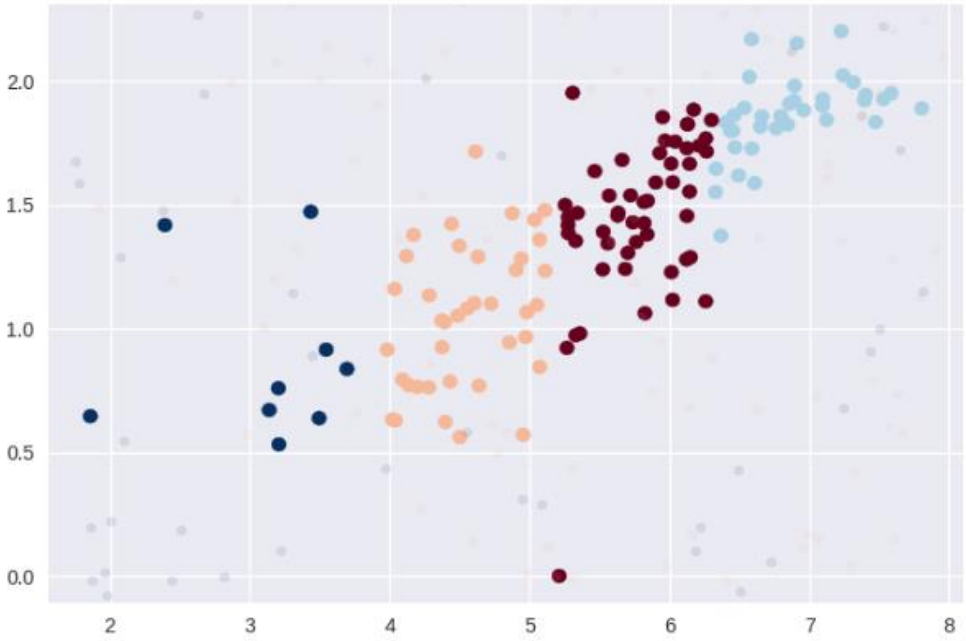
- **Đối với dataset 2023**

```
Naive-Bayes Model Evaluation on Test Data:
[[6 3 0 0]
 [0 8 0 1]
 [1 0 8 0]
 [0 0 0 1]]
              precision    recall  f1-score   support

           0       0.86      0.67      0.75         9
           1       0.73      0.89      0.80         9
           2       1.00      0.89      0.94         9
           3       0.50      1.00      0.67         1

    accuracy                           0.82        28
   macro avg       0.77      0.86      0.79        28
weighted avg       0.85      0.82      0.82        28
```

- Chúng ta thấy rằng thuật toán Naive bayes đều cho ra một giá trị chính xác (accuracy = 0.84) là một giá trị tạm chấp nhận được nhưng vẫn không cao bằng mô hình phân lớp KNN với k nằm trong khoảng 3 đến 7 (accuracy = 0.93).
- Trực quan hóa các lớp của mô hình Naive bayes:



- Có thể nhận thấy rằng mô hình Naive bayes này khi trực quan hóa lên cho ra kết quả cũng gần giống như mô hình KNN, điều này vẫn chưa đảm bảo là kết quả trực quan này có tỉ lệ chính xác cao vì accuracy của Naive bayes đối với dataset 2023 là không cao.
- Kết luận : đối với bộ dataset 2023 thì ta mô hình KNN với k nằm trong khoảng từ 3 đến 7 sẽ là tốt nhất cho việc phân lớp.

4.5. Luật kết hợp

- Mục tiêu của luật kết hợp là tìm ra những "kết hợp có ý nghĩa, đáng chú ý, quan trọng".
- Một trong các thuật toán để tìm ra luật kết hợp là Apriori. Ý tưởng thực hiện thuật toán này là ban đầu phải tìm tập hợp 1 phổ biến, rồi kết hợp các tập hợp 1 đã chọn được để tạo thành các tập hợp 2, và cứ lặp lại quá trình để tìm các tập hợp lớn hơn cho đến khi không thể tạo thêm tập hợp kết hợp mới có tần suất xuất hiện lớn hơn min_support. Cuối cùng kết hợp tất cả các tập họp lại để tạo ra tập luật kết hợp.

- **Tìm luật kết hợp theo thuật toán Apriori**

- Nhóm sẽ tìm kiếm luật kết hợp với 2 bộ dữ liệu của năm 2019 và năm 2023.
- Để chuẩn hóa lại bộ dữ liệu cho phù hợp trước khi tìm kiếm luật kết hợp bằng thuật toán Apriori, nhóm đã sử dụng quy luật sau: với mỗi cột dữ liệu, chọn ra một giá trị ngưỡng n sao cho giá trị n chia cho giá trị lớn nhất của cột đó ra gần bằng với 0.7 nhất. Ở từng cột, nếu giá trị của dòng dữ liệu nào lớn hơn giá trị ngưỡng n của cột đang xét thì chuyển giá trị đó thành 1, nếu không thì chuyển thành 0.
- Vì số lượng luật kiếm được ở mỗi bộ dữ liệu là khá nhiều, nên nhóm chỉ sẽ chọn ra một số luật tiêu biểu để giải thích ở từng bộ dữ liệu. Với mỗi luật nhóm sẽ so sánh kết quả của luật kết hợp với những kết luận ở phần sự tương quan dữ liệu ở trên để chứng minh là những kết luận đó đúng.

- **Đối với dataset 2019:**

Sau khi sử dụng thuật toán Apriori lên bộ dữ liệu năm 2019, nhóm đã rút ra được 70 luật:

- Một số luật:

	antecedents	consequents	antecedent support	consequent support	support	confidence	lift	leverage	conviction	zhangs_metric
0	frozenset({'GDP'})	frozenset({'Happy score'})	0,3076923077	0,4743589744	0,2820512821	0,9166666667	1,932432432	0,1360946746	6,307692308	0,696969697
1	frozenset({'Happy score'})	frozenset({'Social support'})	0,4743589744	0,6794871795	0,4615384615	0,972972973	1,431922489	0,13921762	11,85897436	0,5738482385
2	frozenset({'Happy score'})	frozenset({'Healthy life expectancy'})	0,4743589744	0,4807692308	0,3717948718	0,7837837838	1,63027027	0,1437376726	2,401442308	0,7354920101
3	frozenset({'Healthy life expectancy'})	frozenset({'Happy score'})	0,4807692308	0,4743589744	0,3717948718	0,7733333333	1,63027027	0,1437376726	2,319004525	0,7445721584
4	frozenset({'Freedom to make life choices'})	frozenset({'Happy score'})	0,4230769231	0,4743589744	0,3076923077	0,7272727273	1,533169533	0,1070019724	1,927350427	0,6027777778
5	frozenset({'GDP'})	frozenset({'Social support'})	0,3076923077	0,6794871795	0,3076923077	1	1,471698113	0,09861932939	inf	0,462962963
6	frozenset({'GDP'})	frozenset({'Healthy life expectancy'})	0,3076923077	0,4807692308	0,2884615385	0,9375	1,95	0,1405325444	8,307692308	0,7037037037
7	frozenset({'Healthy life expectancy'})	frozenset({'Social support'})	0,4807692308	0,6794871795	0,4551282051	0,9466666667	1,393207547	0,1284516765	6,009615385	0,5435576421
8	frozenset({'Freedom to make life choices'})	frozenset({'Social support'})	0,4230769231	0,6794871795	0,3653846154	0,8636363636	1,271012007	0,07790927022	2,35042735	0,3695906433
9	frozenset({'GDP', 'Happy score'})	frozenset({'Social support'})	0,2820512821	0,6794871795	0,2820512821	1	1,471698113	0,09040105194	inf	0,4464285714
10	frozenset({'GDP', 'Social support'})	frozenset({'Happy score'})	0,3076923077	0,4743589744	0,2820512821	0,9166666667	1,932432432	0,1360946746	6,307692308	0,696969697
11	frozenset({'GDP'})	frozenset({'Happy score', 'Social support'})	0,3076923077	0,4615384615	0,2820512821	0,9166666667	1,986111111	0,1400394477	6,461538462	0,7171717172
12	frozenset({'GDP', 'Happy score'})	frozenset({'Healthy life expectancy'})	0,2820512821	0,4807692308	0,2628205128	0,9318181818	1,938181818	0,1272189349	7,615384615	0,6742160279
13	frozenset({'GDP', 'Healthy life expectancy'})	frozenset({'Happy score'})	0,2884615385	0,4743589744	0,2628205128	0,9111111111	1,920720721	0,1259861933	5,913461538	0,6736980883
14	frozenset({'Happy score', 'Healthy life expe	frozenset({'GDP'})	0,3717948718	0,3076923077	0,2628205128	0,7068965517	2,297413793	0,1484220907	2,36199095	0,8989547038
15	frozenset({'GDP'})	frozenset({'Happy score', 'Healthy life expecta	0,3076923077	0,3717948718	0,2628205128	0,8541666667	2,297413793	0,1484220907	4,307692308	0,8157181572
16	frozenset({'GDP', 'Freedom to make life ch	frozenset({'Happy score'})	0,1987179487	0,4743589744	0,1923076923	0,9677419355	2,040104621	0,09804404997	16,29487179	0,6362666667
17	frozenset({'Happy score', 'Healthy life expe	frozenset({'Social support'})	0,3717948718	0,6794871795	0,3717948718	1	1,471698113	0,119165023	inf	0,5102040816
18	frozenset({'Happy score', 'Social support'})	frozenset({'Healthy life expectancy'})	0,4615384615	0,4807692308	0,3717948718	0,8055555556	1,675555556	0,1499013807	2,67032967	0,7487684729
19	frozenset({'Healthy life expectancy', 'Social	frozenset({'Happy score'})	0,4551282051	0,4743589744	0,3717948718	0,8169014085	1,722116483	0,1559007232	2,870808679	0,7695740365
20	frozenset({'Happy score'})	frozenset({'Healthy life expectancy', 'Social su	0,4743589744	0,4551282051	0,3717948718	0,7837837838	1,722116483	0,1559007232	2,520032051	0,7977291842
21	frozenset({'Healthy life expectancy'})	frozenset({'Happy score', 'Social support'})	0,4807692308	0,4615384615	0,3717948718	0,7733333333	1,675555556	0,1499013807	2,375565611	0,7765006386

- Nhận xét một số luật đã rút ra được:

+ [GDP] → [Happy score] với Confidence: 0.92:

⇒ GDP đầu người cao sẽ dẫn đến điểm hạnh phúc cao, tương tự với kết luận của sự tương quan dữ liệu ở trên.

+ [Healthy life expectancy] → [Happy score] với Confidence = 0.77:

⇒ Tuổi thọ khỏe mạnh cao sẽ dẫn đến điểm hạnh phúc cao, tương tự với kết luận của sự tương quan dữ liệu ở trên.

+ [GDP, Healthy life expectancy] → [Happy score] với Confidence = 0.91:

⇒ GDP đầu người cao và tuổi thọ khỏe mạnh kỳ vọng cao sẽ dẫn đến điểm hạnh phúc cao, tương tự với kết luận của sự tương quan dữ liệu ở trên.

+ [GDP, Social support] → [Happy score] với Confidence = 0.92:

⇒ GDP đầu người cao và hỗ trợ xã hội cao sẽ dẫn đến điểm hạnh phúc cao, tương tự với kết luận của sự tương quan dữ liệu ở trên.

+ [Healthy life expectancy, Freedom to make life choices, Social support] → [Happy score] với Confidence = 0.93:

⇒ Tuổi thọ khỏe mạnh kỳ vọng cao, tự do khi lựa chọn trong cuộc sống cao và hỗ trợ xã hội cao sẽ dẫn đến điểm hạnh phúc cao, tương tự với kết luận của sự tương quan dữ liệu ở trên.

Ngoài ra thì ta cũng rút ra được một số luật giữa 6 thuộc tính chính:

+ [GDP] → [Social support, Healthy life expectancy] với Confidence = 0.94:

⇒ GDP đầu người cao sẽ dẫn đến hỗ trợ xã hội cao và tuổi thọ khỏe mạnh kỳ vọng cao.

+ [GDP, Healthy life expectancy] → [Social support] với Confidence = 1:

⇒ GDP đầu người cao và tuổi thọ khỏe mạnh kỳ vọng cao sẽ dẫn đến hỗ trợ xã hội cao.

+ [GDP, Freedom to make life choices] → [Social support] với Confidence = 1:

⇒ GDP đầu người cao và tự do khi lựa chọn trong cuộc sống cao sẽ dẫn đến hỗ trợ xã hội cao.

+ [GDP, Freedom to make life choices] → [Healthy life expectancy] với Confidence = 0.97:

⇒ GDP đầu người cao và tự do khi lựa chọn trong cuộc sống cao sẽ dẫn đến tuổi thọ khỏe mạnh kỳ vọng cao.

=> Ta có thể thấy được những luật đã tìm ra được là tương đồng với các kết luận ở phần tương quan dữ liệu

- **Đối với dataset 2023:**

Sau khi sử dụng thuật toán Apriori lên bộ dữ liệu năm 2023, nhóm em đã rút ra được 97 luật:

- Một số luật:

	antecedents	consequents	antecedent support	consequent support	support	confidence	lift	leverage	conviction	zhangs_metric
0	frozenset({'GDP'})	frozenset({'Happy score'})	0,4117647059	0,5588235294	0,3897058824	0,9464285714	1,693609023	0,1596020761	8,235294118	0,6962264151
1	frozenset({'Happy score'})	frozenset({'Social support'})	0,5588235294	0,625	0,5367647059	0,9605263158	1,536842105	0,1875	9,5	0,7917808219
2	frozenset({'Social support'})	frozenset({'Happy score'})	0,625	0,5588235294	0,5367647059	0,8588235294	1,536842105	0,1875	3,125	0,9315068493
3	frozenset({'Healthy life expectancy'})	frozenset({'Happy score'})	0,2426470588	0,5588235294	0,2352941176	0,9696969697	1,735247209	0,09969723183	14,55882353	0,5594660194
4	frozenset({'Happy score'})	frozenset({'Freedom to make life choices'})	0,5588235294	0,5661764706	0,4632352941	0,8289473684	1,464114833	0,1468425606	2,536199095	0,7185185185
5	frozenset({'Freedom to make life choices'})	frozenset({'Happy score'})	0,5661764706	0,5588235294	0,4632352941	0,8181818182	1,464114833	0,1468425606	2,426470588	0,7306967985
6	frozenset({'GDP'})	frozenset({'Social support'})	0,4117647059	0,625	0,4117647059	1	1,6	0,1544117647	inf	0,6375
7	frozenset({'Healthy life expectancy'})	frozenset({'GDP'})	0,2426470588	0,4117647059	0,2352941176	0,9696969697	2,354978355	0,1353806228	19,41176471	0,7597087379
8	frozenset({'GDP'})	frozenset({'Freedom to make life choices'})	0,4117647059	0,5661764706	0,3161764706	0,7678571429	1,356215213	0,0830449827	1,868778281	0,4465116279
9	frozenset({'Healthy life expectancy'})	frozenset({'Social support'})	0,2426470588	0,625	0,2426470588	1	1,6	0,09099264706	inf	0,4951456311
10	frozenset({'Freedom to make life choices'})	frozenset({'Social support'})	0,5661764706	0,625	0,4705882353	0,8311688312	1,32987013	0,1167279412	2,221153846	0,5717690678
11	frozenset({'Social support'})	frozenset({'Freedom to make life choices'})	0,625	0,5661764706	0,4705882353	0,7529411765	1,32987013	0,1167279412	1,755952381	0,6614583333
12	frozenset({'Healthy life expectancy'})	frozenset({'Freedom to make life choices'})	0,2426470588	0,5661764706	0,1985294118	0,8181818182	1,445100354	0,0611483564	2,386029412	0,4066882416
13	frozenset({'GDP', 'Happy score'})	frozenset({'Social support'})	0,3897058824	0,625	0,3897058824	1	1,6	0,1461397059	inf	0,6144578313
14	frozenset({'GDP', 'Social support'})	frozenset({'Happy score'})	0,4117647059	0,5588235294	0,3897058824	0,9464285714	1,693609023	0,1596020761	8,235294118	0,6962264151
15	frozenset({'Happy score', 'Social support'})	frozenset({'GDP'})	0,5367647059	0,4117647059	0,3897058824	0,7260273973	1,763209393	0,1686851211	2,147058824	0,9344115004
16	frozenset({'GDP'})	frozenset({'Happy score', 'Social support'})	0,4117647059	0,5367647059	0,3897058824	0,9464285714	1,763209393	0,1686851211	8,647058824	0,7358490566
17	frozenset({'GDP', 'Healthy life expectancy'})	frozenset({'Happy score'})	0,2352941176	0,5588235294	0,2279411765	0,96875	1,733552632	0,0964532872	14,11764706	0,5533498759
18	frozenset({'Happy score', 'Healthy life expectancy	frozenset({'GDP'})	0,2352941176	0,4117647059	0,2279411765	0,96875	2,352678571	0,1310553633	18,82352941	0,7518610422
19	frozenset({'Healthy life expectancy'})	frozenset({'GDP', 'Happy score'})	0,2426470588	0,3897058824	0,2279411765	0,9393939394	2,410520297	0,1333801903	10,06985294	0,7726276229
20	frozenset({'GDP', 'Happy score'})	frozenset({'Freedom to make life choices'})	0,3897058824	0,5661764706	0,3088235294	0,7924528302	1,399656947	0,08818122837	2,090240642	0,4678714859
21	frozenset({'GDP', 'Freedom to make life choices'})	frozenset({'Happy score'})	0,3161764706	0,5588235294	0,3088235294	0,976744186	1,747858017	0,1321366782	18,97058824	0,6257040451

- Nhận xét một số luật đã rút ra được:

+ [GDP] → [Happy score] với Confidence: 0.95:

⇒ GDP đầu người cao sẽ dẫn đến điểm hạnh phúc cao, tương tự với kết luận của sự tương quan dữ liệu ở trên.

+ [Healthy life expectancy] → [Happy score] với Confidence = 0.97:

⇒ Tuổi thọ khỏe mạnh cao sẽ dẫn đến điểm hạnh phúc cao, tương tự với kết luận của sự tương quan dữ liệu ở trên.

+ [GDP, Healthy life expectancy] → [Happy score] với Confidence = 0.97:

⇒ GDP đầu người cao và tuổi thọ khỏe mạnh kỳ vọng cao sẽ dẫn đến điểm hạnh phúc cao, tương tự với kết luận của sự tương quan dữ liệu ở trên.

+ [GDP, Social support] → [Happy score] với Confidence = 0.95:

⇒ GDP đầu người cao và hỗ trợ xã hội cao sẽ dẫn đến điểm hạnh phúc cao, tương tự với kết luận của sự tương quan dữ liệu ở trên.

+ [GDP, Healthy life expectancy, Social support] → [Happy score] với Confidence = 0.97:

⇒ GDP đầu người cao, tuổi thọ khỏe mạnh kỳ vọng cao và hỗ trợ xã hội cao sẽ dẫn đến điểm hạnh phúc cao, tương tự với kết luận của sự tương quan dữ liệu ở trên.

Ngoài ra thì ta cũng rút ra được một số luật giữa 6 thuộc tính chính:

+ [GDP] → [Social support] với Confidence = 1:

⇒ GDP đầu người cao sẽ dẫn đến hỗ trợ xã hội cao.

+ [GDP, Healthy life expectancy] → [Social support] với Confidence = 1:

⇒ GDP đầu người cao và tuổi thọ khỏe mạnh kỳ vọng cao sẽ dẫn đến hỗ trợ xã hội cao.

+ [Healthy life expectancy] → [Freedom to make life choices, Social support] với Confidence = 0.82:

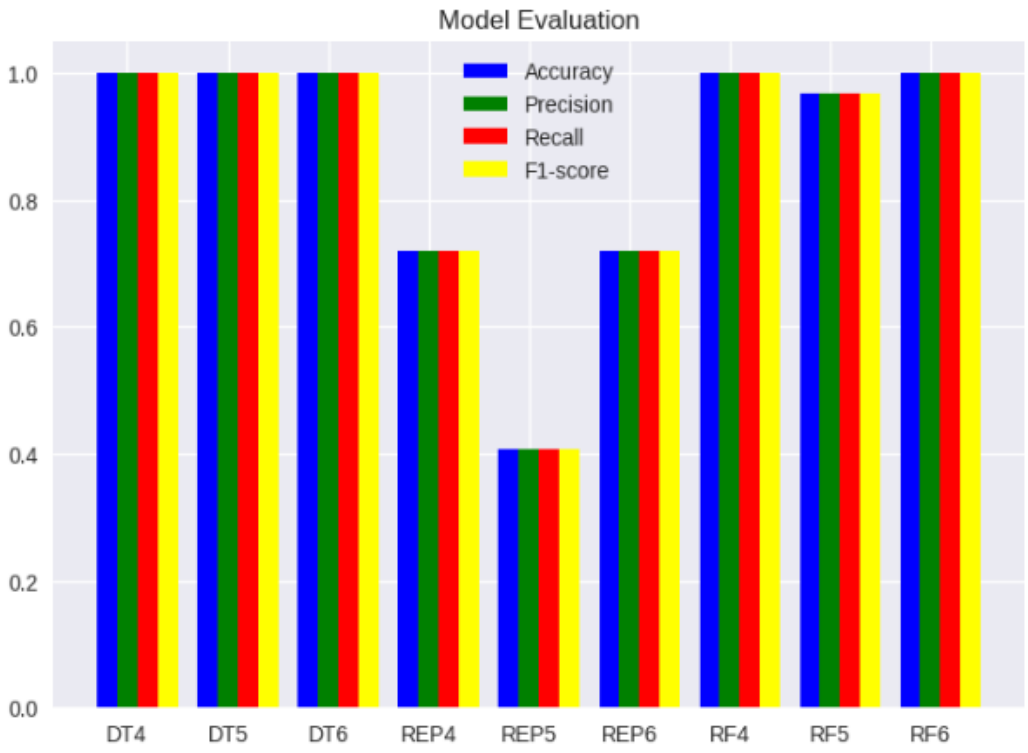
⇒ Tuổi thọ khỏe mạnh kỳ vọng cao sẽ dẫn đến tự do khi lựa chọn trong cuộc sống cao và hỗ trợ xã hội cao.

+ [GDP, Freedom to make life choices] → [Social support] với Confidence = 1:

⇒ GDP đầu người cao và tự do khi lựa chọn trong cuộc sống cao sẽ dẫn đến hỗ trợ xã hội cao.

=> Ta có thể thấy được những luật đã tìm ra được là tương đồng với các kết luận ở phần tương quan dữ liệu

4.6. Đánh giá Decision Tree, REP tree, Random forest



- DT: decision tree
- REP: reduced error pruning tree
- RF: random forest

Nhận xét:

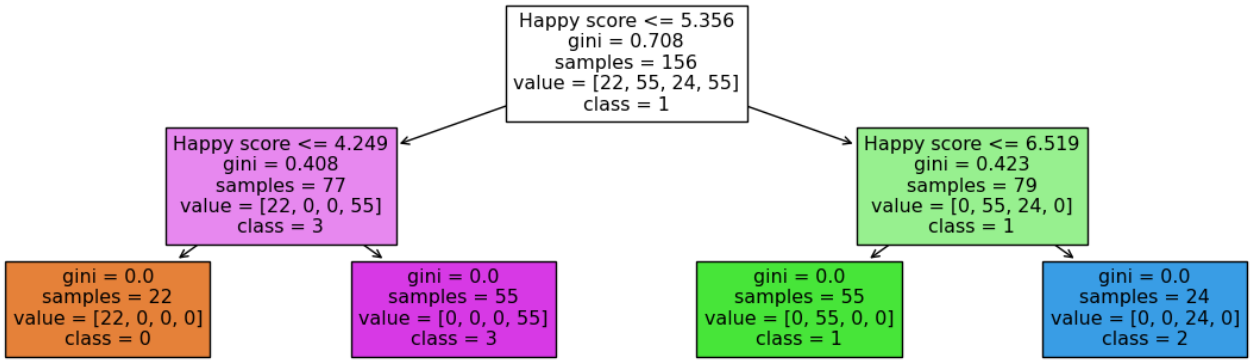
- Các chỉ số DT4, DT5, DT6, RF4, RF6 cao nhất. Chỉ có Decision tree là thuộc dạng cao nhất nên nhóm dùng làm Cây quyết định.
- Các chỉ sốt REP 4, REP 5, REP 6 thấp nhất. Vì vậy nhóm không sử dụng Reduced error pruning tree và Random forest cho Cây quyết định.
- Nhóm đã trực quan các chỉ số về accuracy(độ chính xác), precision, recall, F1-score của các mô hình cây lên thành một biểu đồ cột. Thông qua biểu đồ trên có thể thấy decision tree với max-depth là 4, 5 , 6 đều cho ra những thông số tốt nhất nên nhóm tiến hành chọn decision tree để làm.

4.7. Cây quyết định

- Nhóm sẽ sử dụng lại kết quả của gom cụm K-Means để làm các lớp cho bộ dữ liệu.
- Ở mỗi node trong cây, nhóm sẽ phân tích những thuộc tính sau:

  + Tên_feature </>/= giá_trị: Đây là điều kiện kiểm tra xem Tên_feature có đạt một ngưỡng giá trị nào đó hay không..
  + gini = giá_trị: Chỉ số gini dùng để đo lường sự không đồng nhất trong một tập dữ liệu. Giá trị chỉ số gini càng thấp thì sự phân chia dữ liệu càng tốt..
  + samples = giá_trị: Tổng số mẫu đã được quan sát ở node này
  + value = [danh sách số mẫu ở các lớp]: Phân bố của các lớp ở node này.
  + class = tên_lớp: Lớp được dự đoán ở node này dựa trên số lượng lớp có số lượng mẫu nhiều nhất ở lớp

- **Cây quyết định năm 2019**



- **Node gốc:**

  + Happy score ≤ 5.356: Đây là điều kiện kiểm tra feature Happy score có giá trị ≤ 5.356 hay không.
  + gini = 0.708: Chỉ số gini ở node này là 0.708.
  + samples = 156: Tổng số mẫu đã được quan sát ở node này là 156.
  + value = [22, 55, 24, 55]: Phân bố của các lớp ở node này. Có 22 mẫu thuộc lớp "0", 55 mẫu thuộc lớp "1", 24 mẫu thuộc lớp "2" và 55 mẫu thuộc lớp "3".
  + class = 1: Lớp được dự đoán ở node này là "1" vì đây là class có số lượng mẫu nhiều nhất ở node này. Số lượng mẫu của lớp 1 và 3 là như nhau, và khi gặp trường hợp này, một trong những cách đơn giản nhất mà thuật toán có thể lựa chọn để giải quyết là chọn lớp có index nhỏ hơn.

⇒ Như vậy, ở node gốc này sẽ dựa vào đặc điểm "Happy score ≤ 5.356" để phân chia tập dữ liệu. Node này có 156 mẫu với phân bố lớp như trên. Lớp được dự đoán ở đây là lớp "1". Nếu như điều kiện ở node gốc là True thì sẽ xét tiếp đến node con ở bên trái, nếu là False thì sẽ xét tiếp đến node con ở bên phải.

- **Node con bên trái:**
  + Happy score ≤ 4.249: Đây là điều kiện kiểm tra feature Happy score có giá trị ≤ 4.249 hay không.
  + gini = 0.408: Chỉ số gini ở node này là 0.408.
  + samples = 77: Tổng số mẫu đã được quan sát ở node này là 77.
  + value = [22, 0, 0, 55]: Phân bố của các lớp ở node này. Có 22 mẫu thuộc lớp "0", 0 mẫu thuộc lớp "1", 0 mẫu thuộc lớp "2" và 55 mẫu thuộc lớp "3".
  + class = 3: Lớp được dự đoán ở node này là "3" vì đây là lớp có số lượng mẫu nhiều nhất ở node này.

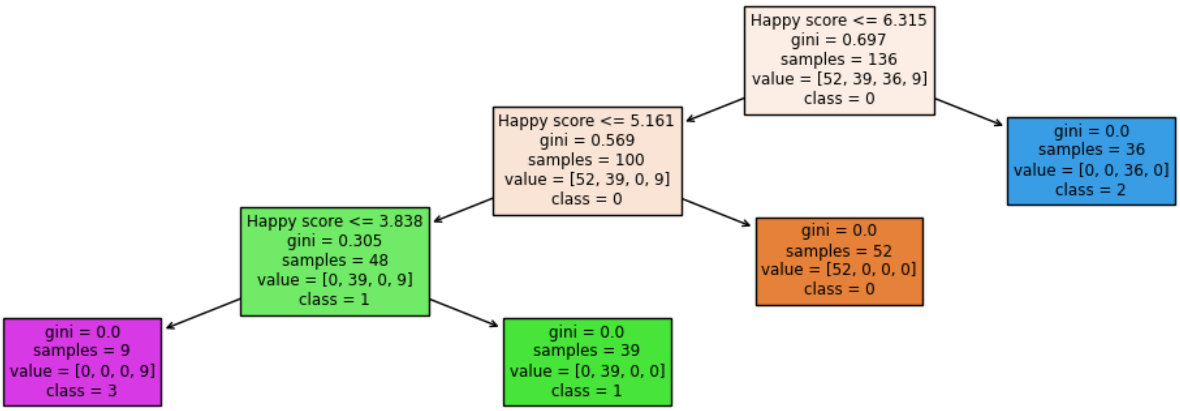
⇒ Như vậy, ở node này sẽ dựa vào đặc điểm "Happy score ≤ 4.249" để phân chia tập dữ liệu. Node này có 77 mẫu với phân bố lớp như trên. Lớp được dự đoán ở đây là lớp "3". Nếu như điều kiện ở node gốc là True thì sẽ kết luận lớp là lớp "0", nếu là False thì sẽ kết luận lớp là lớp "3"

- **Node con bên phải:**
  + Happy score ≤ 6.519: Đây là điều kiện kiểm tra feature Happy score có giá trị ≤ 6.519 hay không.
  + gini = 0.423: Chỉ số gini ở node này là 0.423.
  + samples = 79: Tổng số mẫu đã được quan sát ở node này là 79.
  + value = [0, 55, 24, 0]: Phân bố của các lớp ở node này. Có 0 mẫu thuộc lớp "0", 55 mẫu thuộc lớp "1", 24 mẫu thuộc lớp "2" và 0 mẫu thuộc lớp "3".
  + class = 1: Lớp được dự đoán ở node này là "1" vì đây là lớp có số lượng mẫu nhiều nhất ở node này.

⇒ Như vậy, ở node này sẽ dựa vào đặc điểm "Happy score ≤ 6.519" để phân chia tập dữ liệu. Node này có 79 mẫu với phân bố lớp như trên. Lớp được dự đoán ở đây là lớp "1". Nếu như điều kiện ở node gốc là True thì sẽ kết luận lớp là lớp "1", nếu là False thì sẽ kết luận lớp là lớp "3"

=> Dựa vào thuộc tính điểm hạnh phúc, ta đã tạo ra được cây quyết định với độ sâu là 1.

- **Cây quyết định năm 2023**



- **Node gốc:**

  + Happy score ≤ 6.315: Đây là điều kiện kiểm tra feature Happy score có giá trị ≤ 6.315 hay không.
  + gini = 0.697: Chỉ số gini ở node này là 0.697.
  + samples = 136: Tổng số mẫu đã được quan sát ở node này là 136.
  + value = [52, 39, 36, 9]: Phân bố của các lớp ở node này. Có 52 mẫu thuộc lớp "0", 39 mẫu thuộc lớp "1", 36 mẫu thuộc lớp "2" và 9 mẫu thuộc lớp "3".
  + class = 0: Lớp được dự đoán ở node này là "0" vì đây là class có số lượng mẫu nhiều nhất ở node này.

⇒ Như vậy, ở node gốc này sẽ dựa vào đặc điểm "Happy score ≤ 6.315" để phân chia tập dữ liệu. Node này có 136 mẫu với phân bố lớp như trên. Lớp được dự đoán ở đây là lớp "0". Nếu như điều kiện ở node gốc là True thì sẽ xét tiếp đến node con ở bên trái, nếu là False thì sẽ kết luận lớp là lớp "2".

- **Node con bên trái mức 1:**

  + Happy score ≤ 5.161: Đây là điều kiện kiểm tra feature Happy score có giá trị ≤ 5.161 hay không.
  + gini = 0.569: Chỉ số gini ở node này là 0.569.
  + samples = 100: Tổng số mẫu đã được quan sát ở node này là 100.
  + value = [52, 39, 0, 9]: Phân bố của các lớp ở node này. Có 52 mẫu thuộc lớp "0", 39 mẫu thuộc lớp "1", 0 mẫu thuộc lớp "2" và 9 mẫu thuộc lớp "3".
  + class = 0: Lớp được dự đoán ở node này là "0" vì đây là lớp có số lượng mẫu nhiều nhất ở node này.

⇒ Như vậy, ở node này sẽ dựa vào đặc điểm "Happy score ≤ 5.161" để phân chia tập dữ liệu. Node này có 100 mẫu với phân bố lớp như trên. Lớp được dự đoán ở đây là lớp "0". Nếu như điều kiện ở node gốc là True thì sẽ xét tiếp đến node con ở bên trái, nếu là False thì sẽ kết luận lớp là lớp "0".

- **Node con bên trái mức 2:**

+ Happy score ≤ 3.838: Đây là điều kiện kiểm tra feature Happy score có giá trị ≤ 3.838 hay không.
+ gini = 0.305: Chỉ số gini ở node này là 0.305.
+ samples = 48: Tổng số mẫu đã được quan sát ở node này là 48.
+ value = [0, 39, 0, 9]: Phân bố của các lớp ở node này. Có 0 mẫu thuộc lớp "0", 39 mẫu thuộc lớp "1", 0 mẫu thuộc lớp "2" và 9 mẫu thuộc lớp "3".
+ class = 1: Lớp được dự đoán ở node này là "1" vì đây là lớp có số lượng mẫu nhiều nhất ở node này.

⇒ Như vậy, ở node này sẽ dựa vào đặc điểm "Happy score ≤ 3.838" để phân chia tập dữ liệu. Node này có 48 mẫu với phân bố lớp như trên. Lớp được dự đoán ở đây là lớp "1". Nếu như điều kiện ở node gốc là True thì sẽ kết luận lớp là lớp "3", nếu là False thì sẽ kết luận lớp là lớp "1".

=> Dựa vào thuộc tính điểm hạnh phúc, ta đã tạo ra được cây quyết định với độ sâu là 2.

5. Tổng kết

Trong bài báo cáo này nhóm đã sử dụng các thuật toán đã được học trong môn Khai phá dữ liệu này để khai phá 5 bộ dữ liệu Báo cáo hạnh phúc thế giới năm 2019 đến 2023. Các thuật toán đã được sử dụng bao gồm K Mean, KNN, Naive Bayes, Apriori, DecisionTreeClassifier,…
Sử dụng các thuật toán nhóm đã xác định những thuộc tính ảnh hưởng nhiều tới sự hạnh phúc của một quốc gia, sự phân bố theo khu vực của các quốc gia theo Điểm hạnh phúc, và cuối cùng là tìm ra những quy luật để vận dụng vào việc khai phá các bộ dữ liệu tương tự sau này.

6. Tài liệu tham khảo

[1] World Happiness Report, "FAQ", 20/03/2023. [Online]. Available: https://worldhappiness.report/faq/ [Accessed 18/01/2024]

[2] Helliwell, J. F., Layard, R., Sachs, J. D., De Neve, J.-E., Aknin, L. B., & Wang, S. (Eds.). (2023). *World Happiness Report 2023*. New York: Sustainable Development Solutions Network.

[3] World Happiness Report, “Cities and Happiness: A Global Ranking and Analysis”, 20/03/2020. [Online]. Available: https://worldhappiness.report/ed/2020/cities-and-happiness-a-global-ranking-and-analysis/ [Accessed 18/01/2024]

[4] Gallup, “Understanding How Gallup Uses the Cantril Scale”, 24/08/2009. [Online]. Available: https://news.gallup.com/poll/122453/understanding-gallup-uses-cantril-scale.aspx [Accessed 18/01/2024]

[5] F. Marquez-Padilla and J. Alvarez, "Grading happiness: what grading systems tell us about cross-country wellbeing comparisons," *Economics Bulletin*, Vol. 38, No. 2, pp.1138–1155, 2018.