

WORLD HAPPINESS REPORT

2023

Cherine BENAMAR , Franck COUSIN , Kervin DHOLAH , Bastien SALAÜN



SOMMAIRE

Introduction	p.2
I) Analyse exploratoire des données	
a) Comprendre le contexte de l'étude pour une analyse approfondie.	p.3
b) Définir clairement les objectifs et le cadre de l'étude pour orienter la recherche.	p.4
c) Sélectionner les variables et les combiner de manière cohérente pour une analyse complète et précise.	p.6
II) Visualisation des données	
a) Analyse des moyennes et des évolutions des données.	p.14
b) Exploration de la distribution des données et leur classement.	p.16
c) Étude de la corrélation entre les différentes variables.	p.19
III) Pré-processing des données	
a) Traitement des valeurs manquantes pour une analyse complète et fiable.	p.21
b) Exploration statistique approfondie des données pour en tirer des insights pertinents.	p.21
c) Représentation visuelle des corrélations entre les variables pour identifier les relations significatives.	p.23
IV) Identification de tendances et de modèles significatifs	
a) Utilisation des techniques de Machine Learning pour analyser les données.	p.26
b) Sélection de l'algorithme approprié pour atteindre les objectifs de l'étude.	p.36
c) Ajustement des hyperparamètres pour améliorer les performances du modèle.	p.37
Conclusion	p.50
Annexes	p.53

Introduction

Le World Happiness Report est une initiative importante qui vise à mesurer et à comprendre le bonheur dans le monde entier. Les données collectées dans ce rapport permettent aux chercheurs de mieux comprendre les facteurs qui contribuent au bonheur et de proposer des solutions pour améliorer la qualité de vie des individus et des sociétés.

Le bonheur est une notion complexe qui a toujours suscité l'intérêt des philosophes, des psychologues, des sociologues et de nombreux autres chercheurs. Bien que le bonheur soit souvent considéré comme un état d'esprit subjectif et personnel, il existe des facteurs universels qui semblent influencer le bien-être des individus.

Le concept de bonheur a évolué au fil du temps, passant d'une définition basée sur la satisfaction des besoins physiques et matériels à une définition plus holistique qui prend en compte les dimensions émotionnelles, sociales et spirituelles. Aujourd'hui, le bonheur est souvent considéré comme un état de satisfaction générale de la vie, qui comprend des éléments tels que la joie, l'épanouissement, la réalisation de soi et la connexion avec les autres.

Le World Happiness Report est une enquête annuelle menée par l'Organisation des Nations Unies pour évaluer le niveau de bonheur dans les pays du monde entier. Le rapport est basé sur des enquêtes auprès des citoyens des pays participants, ainsi que sur des données économiques et sociales. Le rapport fournit des informations sur la situation du bonheur dans le monde, ainsi que des données sur les facteurs qui influencent le bonheur, tels que la richesse, le soutien social et la santé. Le rapport utilise principalement les données du Gallup World Poll. Chaque rapport annuel peut être téléchargé par le public sur le site Web du World Happiness Report.

Les données sont collectées auprès de personnes dans plus de 150 pays. Chaque variable mesurée révèle un score moyen pondéré sur une échelle allant de 0 à 10 qui est suivi au fil du temps et comparé à d'autres pays. Ces variables comprennent actuellement :

- PIB réel par habitant
- Soutien social
- Espérance de vie en bonne santé
- Liberté de faire des choix de vie
- Générosité
- Perceptions de la corruption

BIAIS POSSIBLES

Biais culturel : Les mesures de bonheur, de richesse et de soutien social peuvent varier en fonction des cultures et des valeurs culturelles. Par exemple, certains pays peuvent valoriser davantage l'individualisme, tandis que d'autres peuvent valoriser davantage le collectivisme, ce qui peut influencer les résultats.

Biais de sélection : Le dataset ne comprend que des données sur les pays qui ont participé à l'enquête, ce qui peut ne pas être représentatif de tous les pays du monde. De plus, les données ne

sont collectées que parmi les personnes qui ont répondu aux enquêtes, ce qui peut introduire un biais de sélection.

Biais lié à la méthode de collecte de données : Les données ont été collectées en utilisant des enquêtes auto-déclarées. Les personnes peuvent ne pas être honnêtes ou exactes dans leurs réponses, ce qui peut influencer les résultats.

Biais temporel : Les données ont été collectées à un moment spécifique dans le temps et peuvent ne pas être représentatives des conditions actuelles. Les mesures de bonheur peuvent varier en fonction des événements mondiaux, de l'économie et de l'environnement politique, entre autres facteurs.

I) Analyse exploratoire des données

a) Comprendre le contexte de l'étude pour une analyse approfondie

Contexte

- **Contexte d'insertion du projet dans le métier de Data Analyste.**

En tant que Data Analyste, mon travail consiste à extraire des informations significatives à partir des données et à les utiliser pour aider à prendre des décisions éclairées dans divers domaines d'activités. Le projet qui consiste à analyser le World Happiness Report s'inscrit parfaitement dans mon métier, car il me permet de travailler sur un sujet important pour la société tout en mettant en pratique mes compétences en analyse de données.

En effet, l'analyse du bonheur à l'échelle mondiale permet de mieux comprendre les facteurs qui influencent la qualité de vie des individus, ce qui peut aider les gouvernements et les organisations à concevoir des politiques et des programmes visant à améliorer le bien-être des citoyens. En tant que Data Analyste, je suis en mesure de travailler avec des données complexes et de les transformer en informations utiles qui peuvent aider à informer les décisions des décideurs et des responsables politiques.

Ce projet me permet également de développer mes compétences en visualisation de données et en communication, car il est important de pouvoir présenter les résultats de manière claire et compréhensible pour les parties prenantes. En résumé, ce projet me donne l'opportunité d'utiliser mes compétences en analyse de données pour contribuer à une cause importante tout en développant mes compétences professionnelles.

- **Du point de vue technique.**

D'un point de vue technique, l'analyse du World Happiness Report implique plusieurs étapes importantes dans le traitement des données.

Tout d'abord, il faut recueillir les données, puis prendre le temps de bien comprendre les informations à disposition afin d'effectuer une analyse exploratoire du jeu de données. Il faudra

utiliser différentes techniques de visualisation de données pour identifier les tendances et les relations entre les variables. Cette analyse nous permettra de faire ressortir des tendances concernant les facteurs qui ont le plus d'influence sur le score du bonheur, notamment grâce aux méthodes statistiques, telles que la régression linéaire ou la corrélation.

Ensuite, il sera important de nettoyer et de transformer les données pour éliminer les valeurs aberrantes, les doublons et les données manquantes. Ce Pré-processing nous permettra d'avoir un jeu de données prêt à être utilisé pour les algorithmes de Machine Learning.

b) Définir clairement les objectifs et le cadre de l'étude pour orienter la recherche

Objectifs

Notre objectif est de proposer, à l'aide de données fiables et librement accessibles, une analyse du bien être sur Terre, en répondant à la problématique suivante :

- **Quels sont les facteurs qui influencent le score du bonheur ?**

L'objectif de ce projet est de déterminer les combinaisons de facteurs permettant d'expliquer pourquoi certains pays sont mieux classés que les autres et de mettre en place un algorithme de prédiction du score du bonheur.

Cadre

Nous avons à notre disposition deux jeux de données accessibles librement sur Kaggle.

- [world-happiness-report.csv](#) concernant les années allant de 2005 à 2021
- [world-happiness-report-2021.csv](#) ne concernant que l'année 2021

df_report_21.head()

	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
0	Finland	Western Europe	7.842	0.032	7.904	7.780	10.775	0.954	72.0	0.949	-0.098	0.186	2.43	1.446
1	Denmark	Western Europe	7.620	0.035	7.687	7.552	10.933	0.954	72.7	0.946	0.030	0.179	2.43	1.502
2	Switzerland	Western Europe	7.571	0.036	7.643	7.500	11.117	0.942	74.4	0.919	0.025	0.292	2.43	1.566
3	Iceland	Western Europe	7.554	0.059	7.670	7.438	10.878	0.983	73.0	0.955	0.180	0.673	2.43	1.482
4	Netherlands	Western Europe	7.484	0.027	7.518	7.410	10.932	0.942	72.4	0.913	0.175	0.338	2.43	1.501

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
1.106	0.741	0.691	0.124	0.481	3.253
1.108	0.763	0.686	0.208	0.485	2.868
1.079	0.816	0.653	0.204	0.413	2.839
1.172	0.772	0.698	0.293	0.170	2.967
1.079	0.753	0.647	0.302	0.384	2.798

df_report.head()

	Country name	year	Life Ladder	Log GDP per capita	Social support	Healthy life expectancy at birth	Freedom to make life choices	Generosity	Perceptions of corruption	Positive affect	Negative affect
0	Afghanistan	2008	3.724	7.370	0.451	50.80	0.718	0.168	0.882	0.518	0.258
1	Afghanistan	2009	4.402	7.540	0.552	51.20	0.679	0.190	0.850	0.584	0.237
2	Afghanistan	2010	4.758	7.647	0.539	51.60	0.600	0.121	0.707	0.618	0.275
3	Afghanistan	2011	3.832	7.620	0.521	51.92	0.496	0.162	0.731	0.611	0.267
4	Afghanistan	2012	3.783	7.705	0.521	52.24	0.531	0.236	0.776	0.710	0.268

Volumétrie

Le premier fichier world-happiness-report.csv est stocké dans le dataframe : **df_report**, il comporte 1949 lignes, 11 colonnes ainsi que 373 valeurs manquantes.

Le deuxième fichier world-happiness-report-2021.csv quant à lui est stocké dans le dataframe : **df_report_21**, il comporte 149 lignes, 20 colonnes et aucune valeur manquante.

c) Sélectionner les variables et les combiner de manière cohérente pour une analyse complète et précise

Choix des variables et concaténation

CHOIX DES VARIABLES

Lorsque nous travaillons avec des jeux de données complexes, il est courant de constater qu'ils contiennent un grand nombre de variables potentiellement utiles pour notre analyse. Cependant, inclure toutes ces variables peut rendre l'analyse difficile et peut également nuire à la performance du modèle. **De plus, les rapports officiels du calcul du bonheur sont édités par l'ONU uniquement sur la base des 6 critères énoncés précédemment.** C'est pourquoi, dans notre cas, il est nécessaire d'exclure certaines variables de l'analyse. **Nous devons donc ne garder que les variables ci-dessous :**

« Ladder Score » ou « Life Ladder »

Score de bonheur. Il s'agit de la réponse moyenne nationale à la question sur l'évaluation de la vie. La question est formulée comme suit : « Veuillez imaginer une échelle, avec des marches numérotées de 0 en bas à 10 en haut. Le sommet de l'échelle présente la meilleure vie possible pour vous et le bas de l'échelle représente la pire vie possible pour vous. À quelle marche de l'échelle diriez-vous que vous vous trouvez personnellement en ce moment ? »

« Country name » : nom du pays

« Regional indicator »

Région du monde auquel appartient le pays

« year » : année de sondage

« Logged GDP per capita » ou « Log GDP per capita »

PIB par habitant en parité de pouvoir d'achat

« Social support »

Mesure du soutien social perçu, mesurée sur une échelle de 0 à 1

« Healthy life expectancy at birth » ou « Healthy life expectancy »

L'espérance de vie en bonne santé mesurée en années

« Freedom to make life choices »

Mesure de la liberté perçue pour faire des choix de vie, mesurée sur une échelle de 0 à 1

« Generosity »

Mesure de la générosité perçue, mesurée sur une échelle de 0 à 1

« Perceptions of corruption »

Mesure de la perception de la corruption dans le pays, mesurée sur une échelle de 0 à 1

CONCATÉNATION DES 2 JEUX DE DONNÉES

Suite à cette fusion, nous nous retrouvons avec un dataset de 2098 lignes et 10 colonnes.

Remplacement des valeurs manquantes créées lors de la concaténation du fait que la variable « Regional indicator » n'est pas présente dans **df_report**, et que la variable « year » n'est pas présente dans **df_report_21**

À ce stade nous avons traité uniquement les valeurs manquantes de ces deux variables car il s'agit d'une gestion simple et logique. Les autres valeurs manquantes seront traitées après avoir effectué l'analyse du jeu de données.

	Regional_indicator	Country_name	year	Life_Ladder	Log_GDP_per_capita	Social_support	Healthy_life_expectancy_at_birth	Freedom_to_make_life_choices	Generosity
0	South Asia	Afghanistan	2008	3.724	7.370	0.451	50.80	0.718	0.168
1	South Asia	Afghanistan	2009	4.402	7.540	0.552	51.20	0.679	0.190
2	South Asia	Afghanistan	2010	4.758	7.647	0.539	51.60	0.600	0.121
3	South Asia	Afghanistan	2011	3.832	7.620	0.521	51.92	0.496	0.162
4	South Asia	Afghanistan	2012	3.783	7.705	0.521	52.24	0.531	0.236

Perceptions_of_corruption

0.882
0.850
0.707
0.731
0.776

Analyse exploratoire

Dans toute analyse statistique, il est important de déterminer quelle variable représente l'objectif ou la variable à prédire. Dans ce contexte, la variable cible est celle qui sera utilisée pour évaluer la performance du modèle et pour prendre des décisions basées sur ses prévisions. Dans ce cas précis, la variable ici est « **Life_Ladder** ».

Les variables explicatives **retenues** quant à elle sont :

« **Log_GDP_per_capita** », « **Social_support** », « **Healthy_life_expectancy_at_birth** », « **Freedom_to_make_life_choices** », « **Generosity** », « **Perceptions_of_corruption** »

La variable cible et les variables explicatives sont de type quantitative. Elles sont utilisées dans tous les rapports du World Happiness Report depuis 2005.

Le dataset contient les scores du bonheur de 2005 à 2021. La répartition géographique est sur 10 régions avec 166 pays au total.

Liste des régions :

- 'South Asia'
- 'Central and Eastern Europe'
- 'Middle East and North Africa'
- 'Sub-Saharan Africa'
- 'Latin America and Caribbean'
- 'Commonwealth of Independent States'
- 'North America and ANZ'
- 'Western Europe'
- 'Southeast Asia'
- 'East Asia'

Les 166 pays ne sont pas présents sur toutes les années. Concernant les régions, 7 régions sont représentées en 2005 puis à partir de 2006 nous retrouvons les 10 régions.

Analyse descriptive du jeu de données

	Life_Ladder	Log_GDP_per_capita	Social_support	Healthy_life_expectancy_at_birth	Freedom_to_make_life_choices	Generosity	Perceptions_of_corruption
count	2098.000000	2062.000000	2085.000000	2043.000000	2066.000000	2009.000000	1988.000000
mean	5.471402	9.373060	0.812709	63.478503	0.746094	-0.001027	0.745650
std	1.112676	1.154247	0.118203	7.468780	0.140766	0.161400	0.186261
min	2.375000	6.635000	0.290000	32.300000	0.258000	-0.335000	0.035000
25%	4.652250	8.470500	0.750000	58.704500	0.652000	-0.115000	0.688750
50%	5.392000	9.462000	0.835000	65.280000	0.767000	-0.027000	0.801000
75%	6.282500	10.360750	0.905000	68.660000	0.859000	0.089000	0.869000
max	8.019000	11.648000	0.987000	77.100000	0.985000	0.698000	0.983000

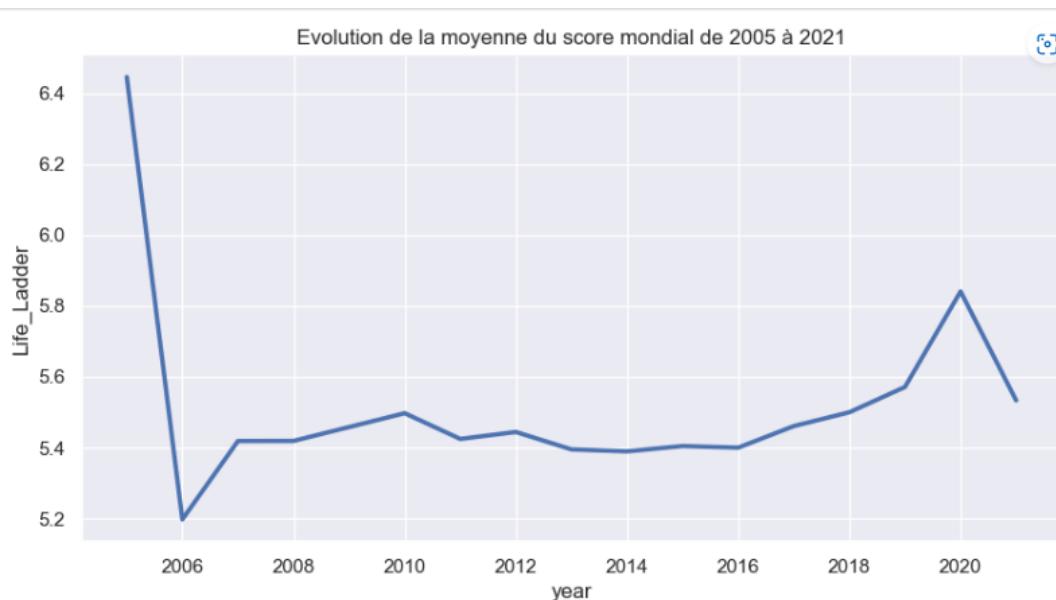
La note minimale du score du bonheur : "Life_Ladder" est de **2.375**, la note maximale est de **8.019**.

La moyenne du score du bonheur dans le monde de 2005 à 2021 : **5.47**

Globalement, les moyennes et les médianes sont proches, cela signifie généralement que la distribution des données est symétrique ou approximativement symétrique. Cela peut également indiquer que les valeurs aberrantes sont rares ou absentes dans les données.

Cela suggère également que les valeurs extrêmes ont peu d'impact sur la tendance centrale des données, ce qui peut être interprété comme un signe de stabilité ou de cohérence des données.

Évolution du score du bonheur dans le monde de 2005 à 2021



Nous pouvons remarquer que le meilleur score du bonheur (moyenne mondiale) est en 2005 et que l'année suivante le score chute de 6.45 à 5.20.

- **Comment peut-on expliquer cette baisse ?**

En explorant le dataset nous pouvons remarquer que le nombre de pays est passé de 27 en 2005 à 89 en 2006. Une augmentation du dataset de 62 pays peut expliquer une variation du score du bonheur.

- **Quels pays sont entrés dans le jeu de données en 2006 ?**

Trois nouvelles régions sont entrées dans le World Happiness report en 2006 : « **Common wealth of Independent States** », « **Sub-Saharan Africa** » et « **South Asia** ».

Ces 3 régions sont arrivées dans le dataset avec un score du bonheur faible :

- Commonwealth of Independent States : **4.83**
- Sub-Saharan Africa : **4.07**
- South Asia : **5.23**

Liste des pays de ces 3 régions :

- Armenia, Azerbaijan, Belarus, Georgia, Kazakhstan, Kyrgyzstan, Moldova, Russia, Tajikistan, Ukraine, Uzbekistan
- Benin, Botswana, Burkina Faso, Cameroon, Chad, Ghana, Kenya, Madagascar, Malawi, Mali, Mozambique, Niger, Nigeria, Rwanda, Senegal, Sierra Leone, South Africa, Tanzania, Togo, Uganda, Zambia, Zimbabwe
- Bangladesh, India, Nepal, Sri Lanka

Ces pays à faibles scores du bonheur sont également des pays à faible PIB par habitant. Cette relation entre le PIB par habitant et le score du bonheur sera à analyser par la suite !

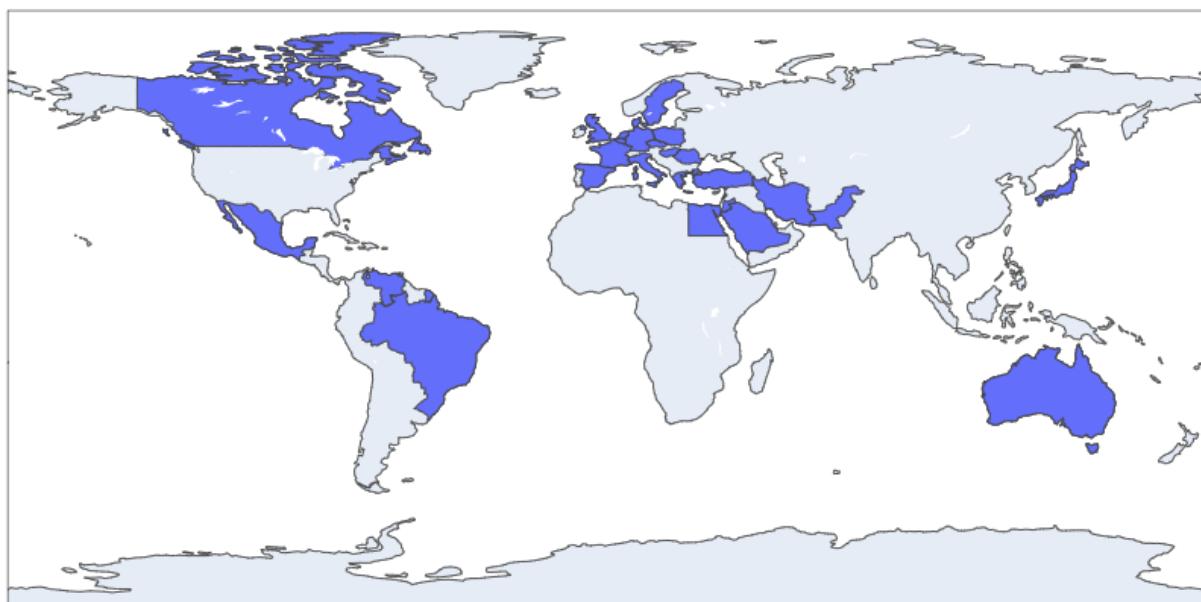
Scores du bonheur et PIB par habitant par régions en 2006

Regional_indicator	year	Life_Ladder	Log_GDP_per_capita
Central and Eastern Europe	2006	5.422200	10.153200
Commonwealth of Independent States	2006	4.834727	8.976273
East Asia	2006	5.398000	10.091250
Latin America and Caribbean	2006	5.581000	9.163316
Middle East and North Africa	2006	5.870400	10.153000
North America and ANZ	2006	7.243500	10.725000
South Asia	2006	4.644750	8.114000
Southeast Asia	2006	5.239500	9.026500
Sub-Saharan Africa	2006	4.074182	7.673455
Western Europe	2006	6.881625	10.778125

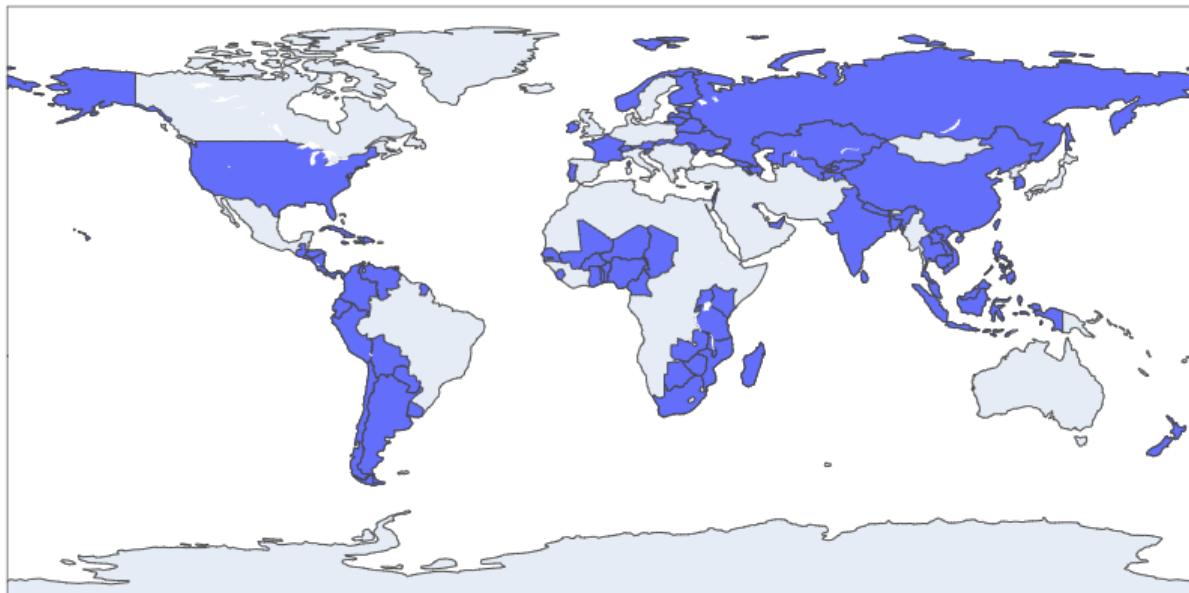
Les 3 nouvelles régions (« Commonwealth of Independent States », « Sub-Saharan Africa » et « South Asia ») ont un PIB par habitant très nettement en dessous des autres régions déjà présentes en 2005. Leurs scores du bonheur sont également plus faibles que les autres régions.

Il semble qu'il y ait une corrélation entre le PIB par habitants et le score du bonheur. La suite de l'analyse pourra nous le confirmer.

Carte des pays présents en 2005



Carte des pays présents en 2006



Nous pouvons également remarquer une importante variation du score du bonheur entre 2019 et 2021 (période COVID). Le dataset passe de 144 pays en 2019 à 95 pays en 2020 puis remonte à 149 pays en 2021.

Les pays qui ne sont plus présents en 2020 :

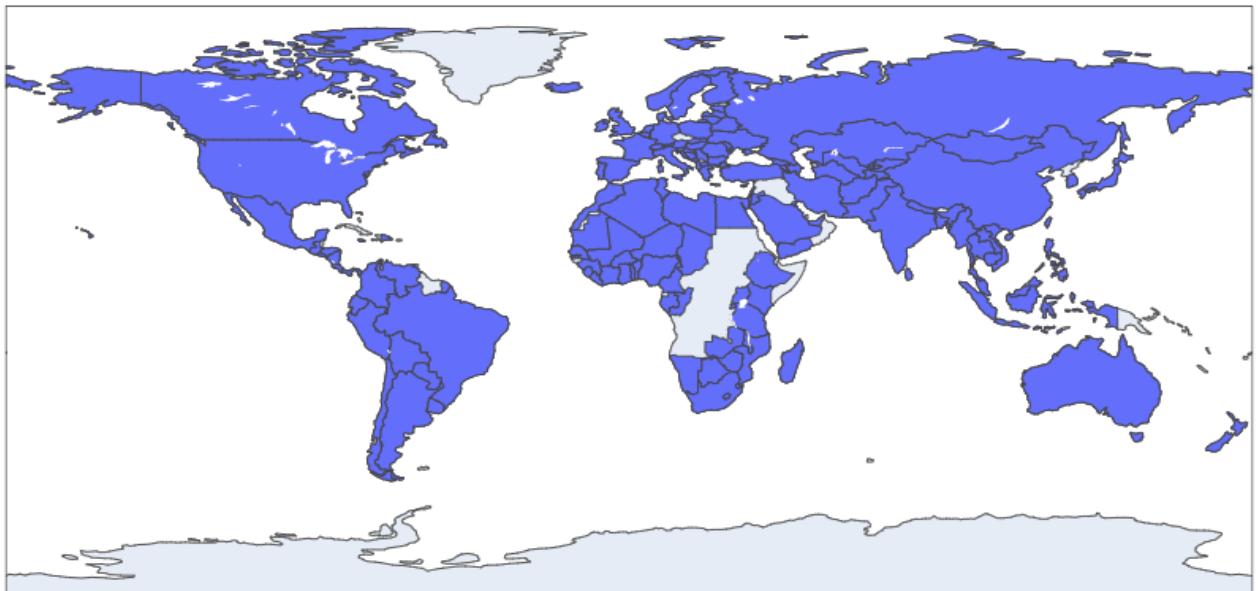
'Afghanistan', 'Algeria', 'Armenia', 'Azerbaijan', 'Belarus', 'Botswana', 'Burkina Faso', 'Chad', 'Comoros', 'Congo (Brazzaville)', 'Costa Rica', 'Gabon', 'Gambia', 'Guatemala', 'Guinea', 'Honduras', 'Indonesia', 'Jamaica', 'Kuwait', 'Lebanon', 'Lesotho', 'Liberia', 'Libya', 'Luxembourg', 'Madagascar', 'Malawi', 'Malaysia', 'Mali', 'Mauritania', 'Mozambique', 'Nepal', 'Nicaragua', 'Niger', 'North Cyprus', 'Pakistan', 'Palestinian Territories', 'Panama', 'Paraguay', 'Peru', 'Romania', 'Rwanda', 'Senegal', 'Sierra Leone', 'Singapore', 'Sri Lanka', 'Swaziland', 'Togo', 'Turkmenistan', 'Uzbekistan', 'Vietnam', 'Yemen'.

Les pays qui se rajoutent en 2021 :

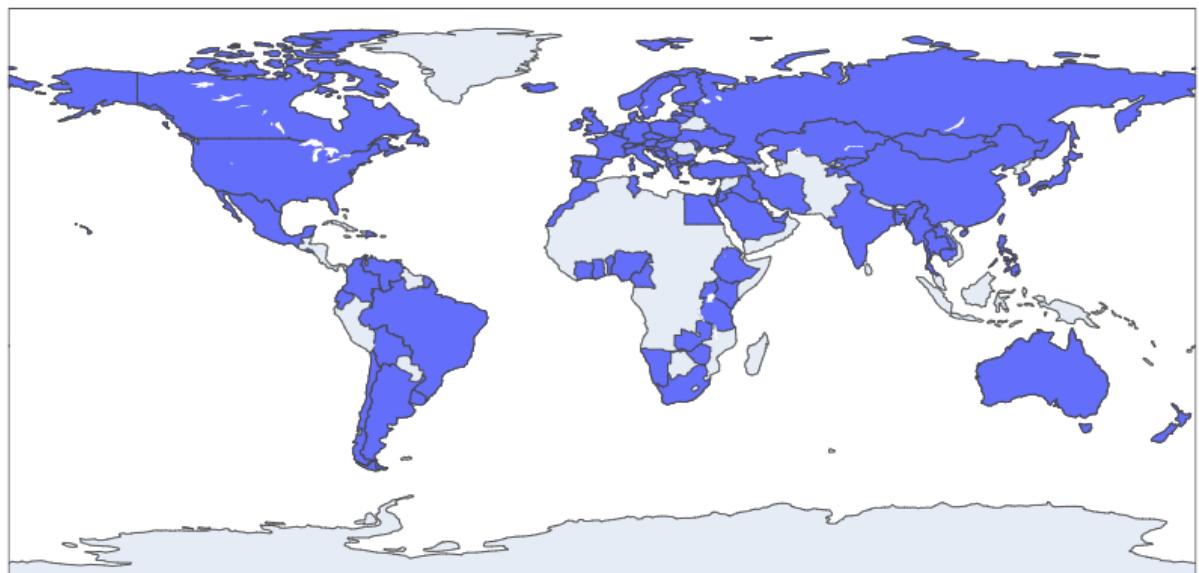
'Luxembourg', 'Costa Rica', 'Guatemala', 'Singapore', 'Jamaica', 'Panama', 'Uzbekistan', 'Romania', 'Kuwait', 'Nicaragua', 'Honduras', 'Peru', 'Paraguay', 'North Cyprus', 'Belarus', 'Vietnam', 'Libya', 'Malaysia', 'Indonesia', 'Congo (Brazzaville)', 'Armenia', 'Nepal', 'Maldives', 'Azerbaijan', 'Senegal', 'Niger', 'Turkmenistan', 'Gambia', 'Guinea', 'Pakistan', 'Algeria', 'Gabon', 'Burkina Faso', 'Mozambique', 'Mali', 'Liberia', 'Lebanon', 'Palestinian Territories', 'Chad', 'Sri Lanka', 'Swaziland', 'Comoros', 'Mauritania', 'Madagascar', 'Togo', 'Sierra Leone', 'Burundi', 'Yemen', 'Haiti', 'Malawi', 'Lesotho', 'Botswana', 'Rwanda', 'Afghanistan'

Nous pouvons constater qu'en majorité, les pays qui disparaissent du dataset en 2020 et qui réapparaissent en 2021 sont des pays à moyen ou faible PIB par habitants. **Encore une fois nous pouvons apercevoir un lien entre le score du bonheur et le PIB par habitant.**

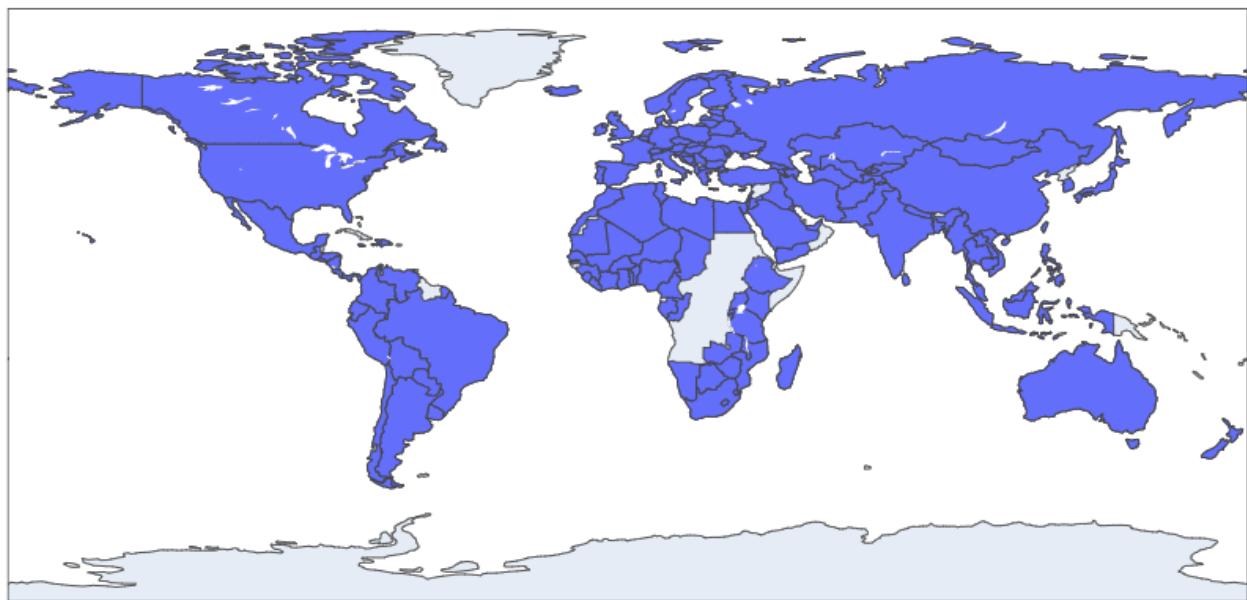
Carte des pays présents en 2019



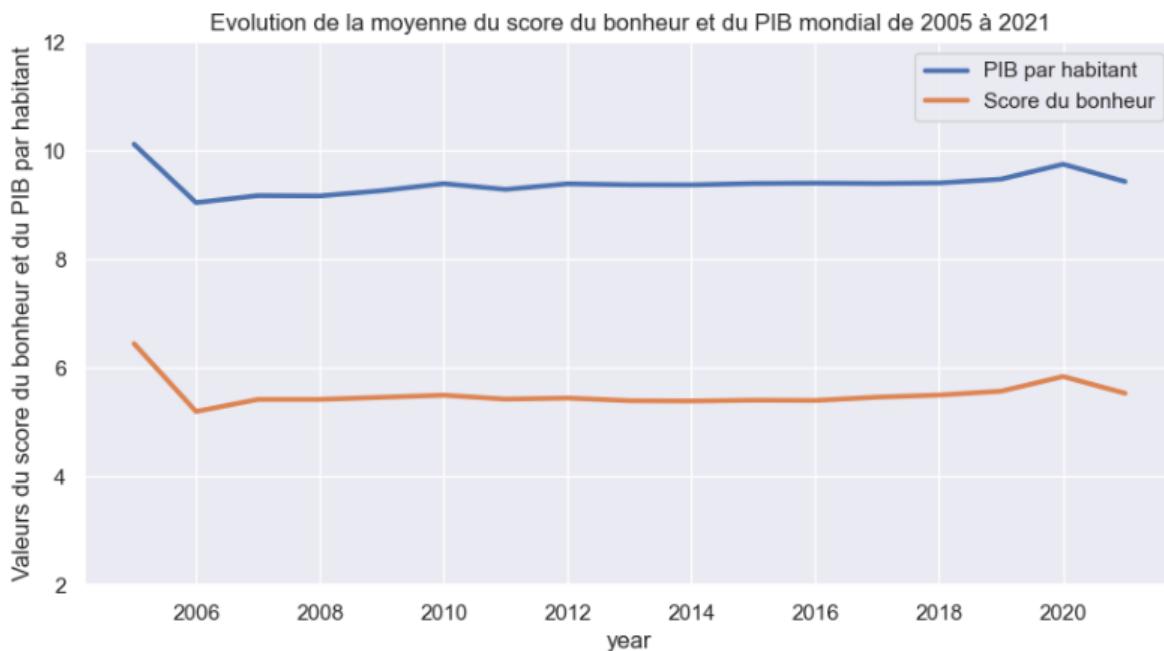
Carte des pays présents en 2020



Carte des pays présents en 2021



Evolution de la moyenne du score du bonheur et du PIB par habitants dans le monde de 2005 à 2021

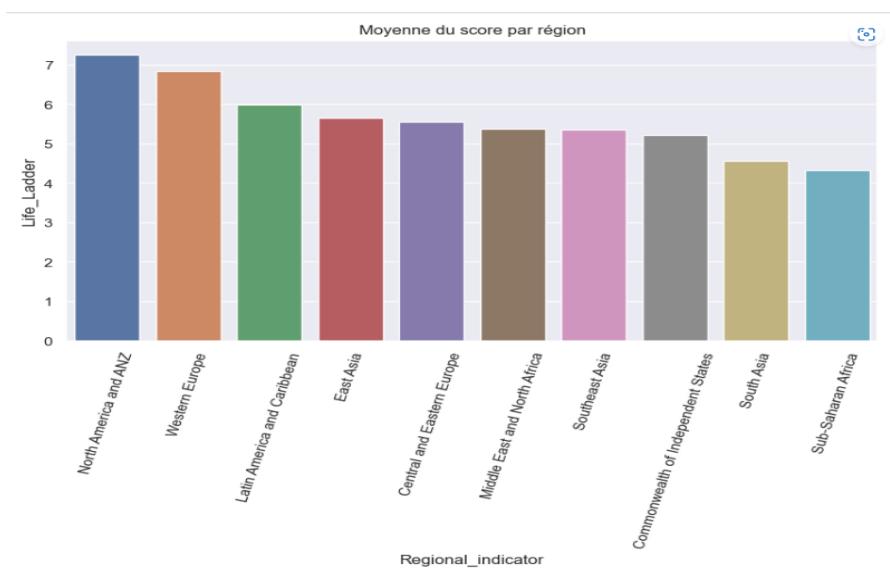


Les deux courbes se suivent : La corrélation entre le PIB par habitants et le score du bonheur devra être confirmée à l'aide de méthodes statistiques.

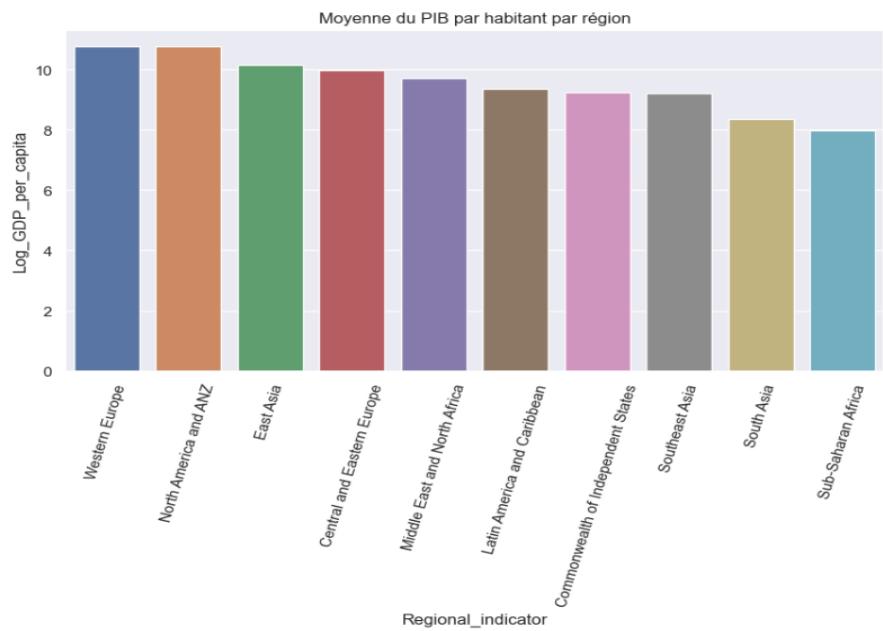
II) Data visualisation

a) Analyse des moyennes et des évolutions des données

Moyenne du score du bonheur par région



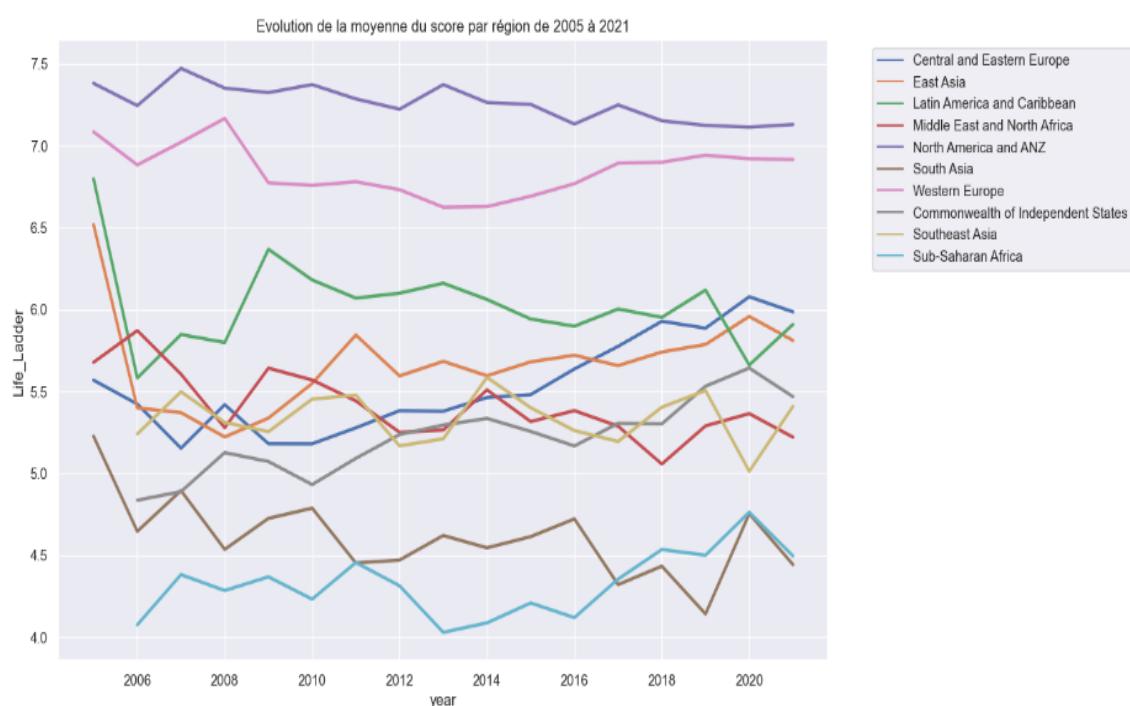
Moyenne du PIB par habitants par région



Le classement des régions par PIB par habitant correspond au classement des régions par score du bonheur.

Regional_indicator	Life_Ladder	Log_GDP_per_capita
North America and ANZ	7.254919	10.766823
Western Europe	6.833096	10.777580
Latin America and Caribbean	5.992900	9.349141
East Asia	5.645511	10.160571
Central and Eastern Europe	5.557930	9.977241
Middle East and North Africa	5.375423	9.728855
Southeast Asia	5.345960	9.220936
Commonwealth of Independent States	5.212835	9.251269
South Asia	4.558404	8.360521
Sub-Saharan Africa	4.318261	7.967635

Évolution de la moyenne du score par région de 2005 à 2021

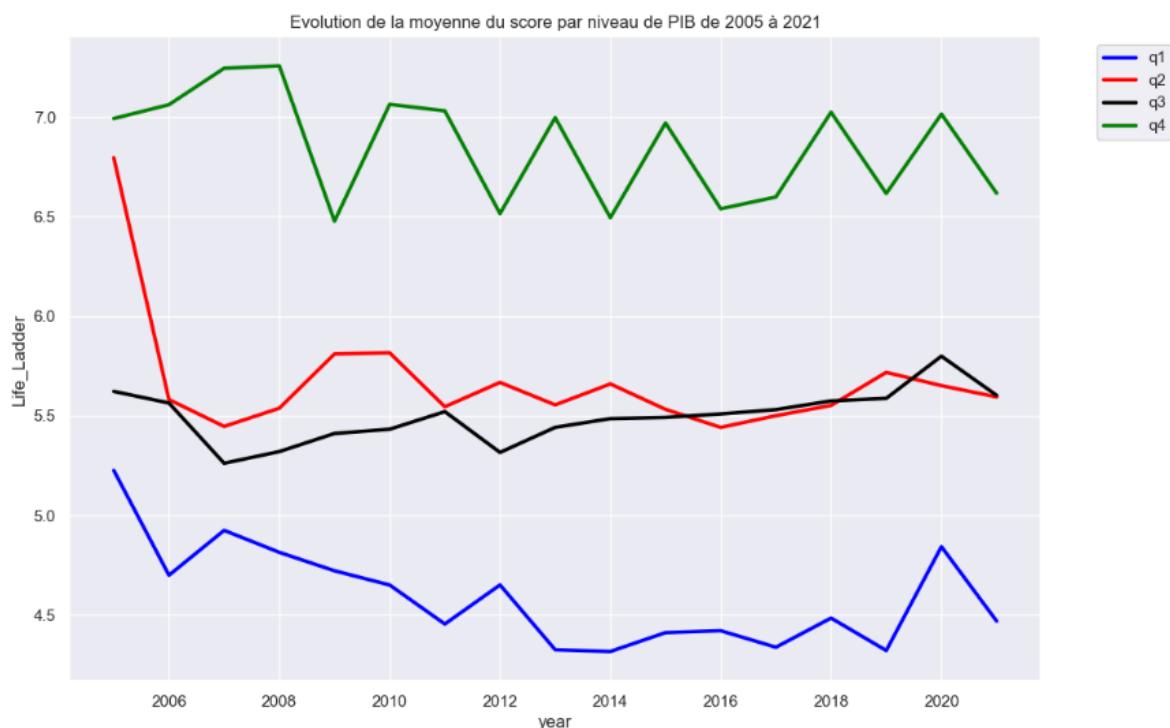


Ce graphique est difficilement lisible surtout pour les 6 courbes situées au centre.
Nous pouvons quand même constater très peu de variations du score du bonheur pour les deux régions avec les scores les plus élevés (« North America and ANZ » et « Western Europe »).

Dans la continuité de notre analyse et pour suivre la tendance que nous avons observé qui est la relation entre le PIB par habitants et le score du bonheur, il semble intéressant de suivre l'évolution du score du bonheur selon le PIB par habitant.

Évolution de la moyenne du score du bonheur selon le niveau de PIB par habitant de 2005 à 2021

Nous avons divisé la variable « Log_GDP_per_capita » (PIB par habitant) selon les quartiles afin d'obtenir 4 groupes de niveau de PIB.

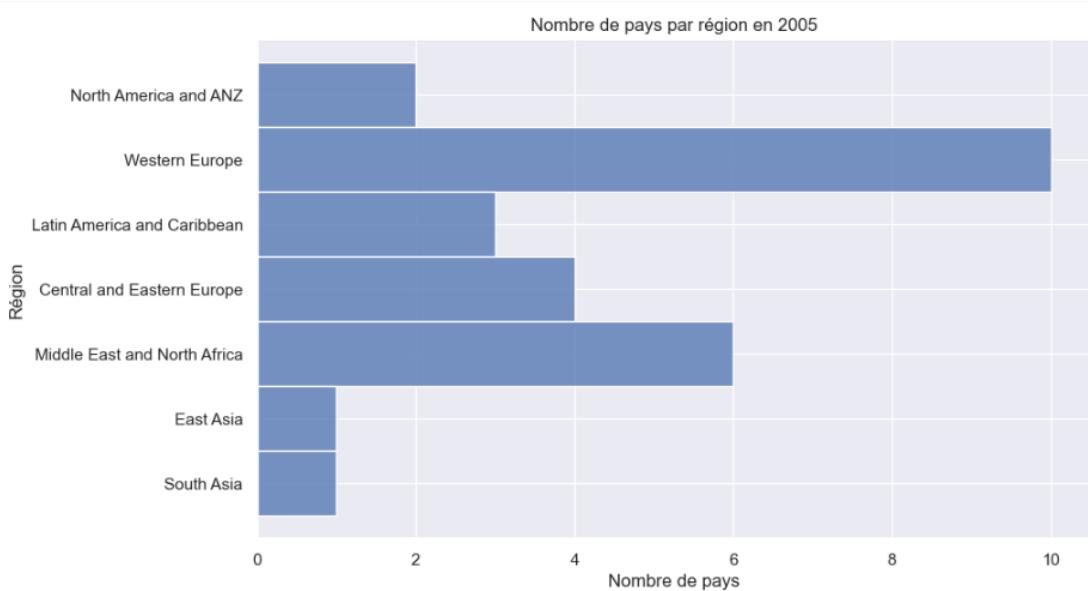


Les scores de bonheur varient légèrement, mais ils sont généralement associés à une plage de scores correspondant au niveau de PIB par habitant.

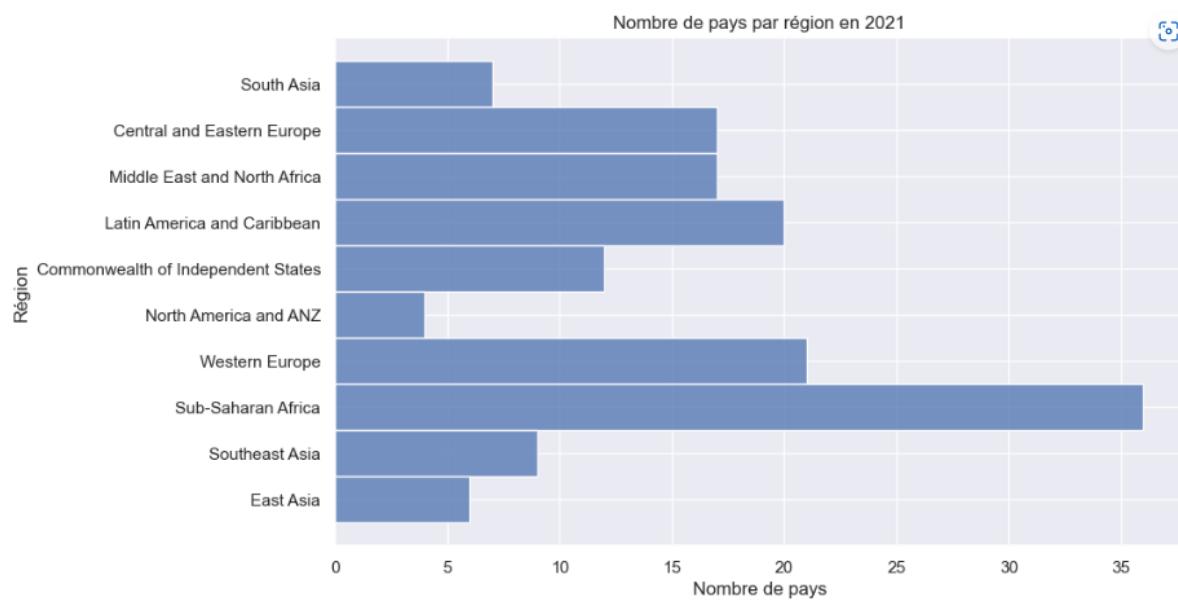
b) Exploration de la distribution des données et leur classement

Distribution par région

Nombre de pays par région en 2005



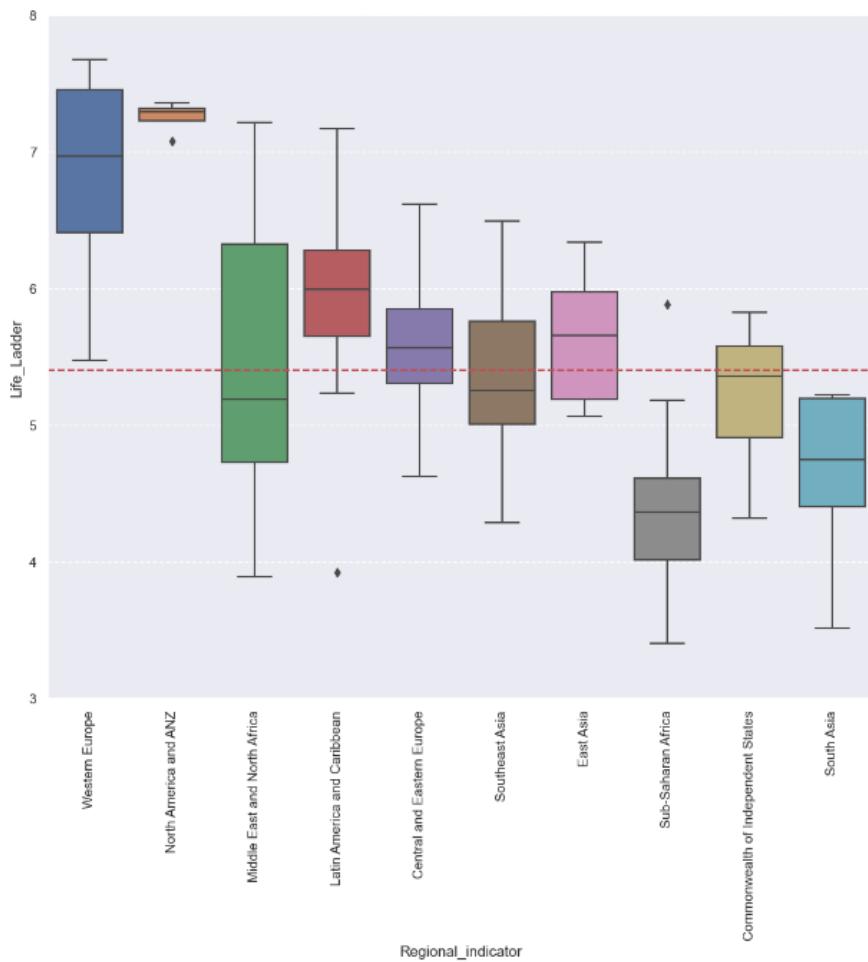
Nombre de pays par région en 2021



Entre 2005 et 2021 la répartition du nombre de pays passe à plus de 150 avec une surreprésentation de l'Afrique sub-saharienne.

DIAGRAMME EN BOÎTE (boxplot)

Score du bonheur (« Life_Ladder ») par région



Sur ce diagramme en boîte on peut s'apercevoir que deux régions sont largement en tête dans le classement du bonheur « Western Europe » et « North America and ANZ ». Pour ces deux régions, tous les pays ont un score du bonheur supérieur à la moyenne mondiale.

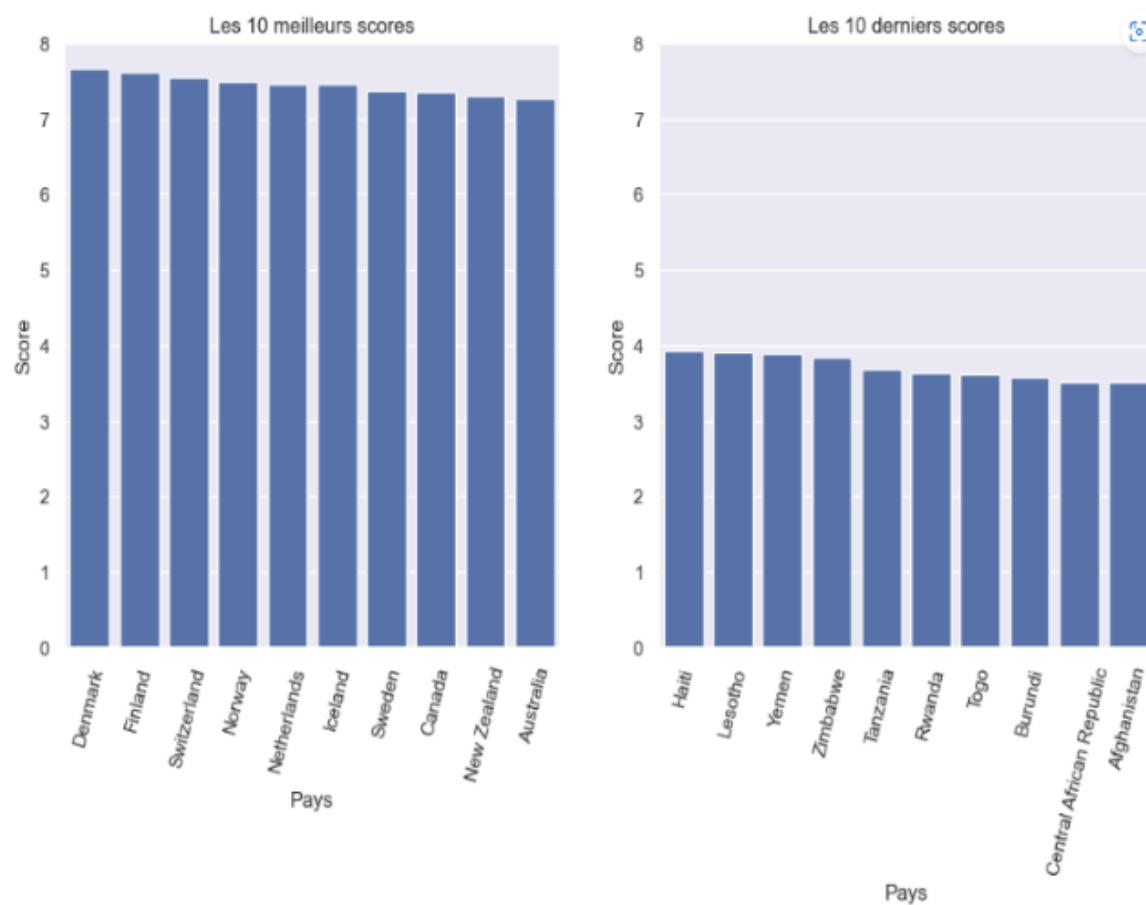
Tous les pays de la région « North America and ANZ » et 50% des pays de la région « Western Europe » ont un Life_Ladder supérieur ou égal à 7, c'est-à-dire supérieur à tous les autres pays du monde (à l'exception des top scores de « Middle East and North Africa » et « Latin America and Caribbean »).

Les scores les plus bas se trouvent dans les régions « Sub-Saharan Africa » et « South Asia » où tous les pays ont un Life_Ladder inférieur à la moyenne mondiale.

Quasiment 75% des pays de « Sub-Saharan Africa » sont en-dessous de 4.5

Les scores les plus hétérogènes se situent dans la région « Middle East and North Africa ».

Classement Top 10 et Bottom 10 des pays selon le score du bonheur



Tous les pays du top 10 sont soit des pays de « Wester Europe » soit des pays de « North America and ANZ ». Le dernier pays du classement est de « South Asia » et 6 pays du bottom 10 sont des pays de « Sub-Saharan Africa ».

c) Étude de la corrélation entre les différentes variables

Corrélations entre les variables

Pour tester l'indépendance de variables lorsque les variables sont quantitatives, le test de corrélation de **Pearson** s'impose. Pour mesurer la corrélation entre les deux variables, on s'appuiera sur le **coefficient de corrélation de Pearson**.

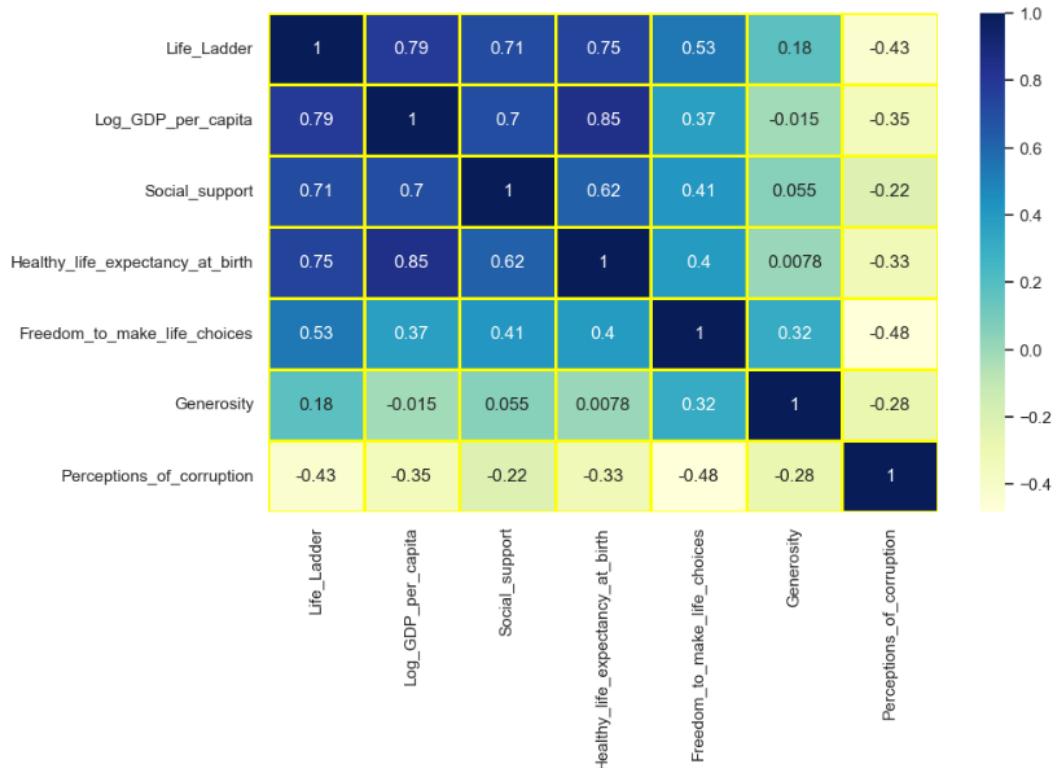
Le coefficient de corrélation de Pearson est une formule qui permet de quantifier la relation linéaire entre deux variables : le coefficient est un réel entre -1 et 1 avec :

- **1 les variables sont corrélées**

- **0 les variables sont décorrélées**
- **-1 les variables sont corrélées négativement**

La matrice de corrélation déjà implémentée dans la bibliothèque Pandas renvoie un dataframe contenant les coefficients de corrélation de pearson entre chaque variable quantitative et les autres. Elle est facilement calculable à l'aide de la méthode `corr`.

Nous affichons une HeatMap (fonction graphique de Seaborn permettant d'afficher un tableau, colorisé en fonction des résultats du test de corrélation de Pearson ainsi que le coefficient de corrélation).



Les corrélations les plus importantes avec le score du bonheur :

- **Log_GDP_per_capita** (PIB par habitant)
- **Social_support** (le soutien social)
- **Healthy_life_expectancy_at_birth** (l'espérance de vie en bonne santé)

On peut également noter une très forte relation entre le pouvoir d'achat et l'espérance de vie en bonne santé.

Pour confirmer les corrélations identifiées entre les variables explicatives (pouvoir d'achat, soutien social, espérance de vie en bonne santé) et la variable cible, nous allons calculer la p-value.

La **p-value** est la probabilité, sous H_0 , d'obtenir une statistique aussi extrême (pour ne pas dire aussi grande) que la valeur observée sur l'échantillon.

Elle représente la probabilité de rejeter l'hypothèse nulle quand celle-ci est vraie. Plus la **p-value** est petite, plus la probabilité de faire une erreur en rejetant l'hypothèse nulle est faible.

Avant de calculer la p-value nous devons gérer en amont les valeurs manquantes.

III) Pré-processing des données

a) Traitement des valeurs manquantes pour une analyse complète et fiable

Gestion des valeurs manquantes

df.isnull().sum()	
Regional_indicator	0
Country_name	0
year	0
Life_Ladder	0
Log_GDP_per_capita	36
Social_support	13
Healthy_life_expectancy_at_birth	55
Freedom_to_make_life_choices	32
Generosity	89
Perceptions_of_corruption	110

Dans les cas où une valeur est manquante pour une variable explicative et que le pays concerné est présent que sur une année, nous avons supprimé le pays du dataset car nous n'avons pas le moyen de remplacer la valeur manquante. Autrement nous avons remplacé les valeurs manquantes par la moyenne du pays pour la variable concernée.

Suite au nettoyage des données, nous nous retrouvons avec un dataset de 2024 lignes.

b) Exploration statistique approfondie des données pour en tirer des insights pertinents

Analyses statistiques

```

Variable explicative : Log_GDP_per_capita
Corrélation de Pearson : 0.7950042114259807
P-value : 0.0

Variable explicative : Social_support
Corrélation de Pearson : 0.7145383539876242
P-value : 4.7465691e-316

Variable explicative : Healthy_life_expectancy_at_birth
Corrélation de Pearson : 0.7503498544045164
P-value : 0.0

Variable explicative : Freedom_to_make_life_choices
Corrélation de Pearson : 0.5312107506137111
P-value : 8.952343695564951e-148

Variable explicative : Generosity
Corrélation de Pearson : 0.17686409470812284
P-value : 1.0989636108273123e-15

Variable explicative : Perceptions_of_corruption
Corrélation de Pearson : -0.45418575645922193
P-value : 1.419070944385589e-103

```

Variable explicative	Corrélation de Pearson	P-Value
Log_GDP_per_capita	0.7950	0.0
Social_support	0.7145	Proche de 0
Healthy_life_expectancy_at_birth	0.7503	0.0
Freedom_to_make__life_choice	0.5312	Proche de 0
Generosity	0.1768	Proche de 0
Perception_of_corruption	-0.4542	Proche de 0

Log_GDP_per_capita : Il y a une corrélation positive significative (corrélation de Pearson = 0.7950042114259807) entre le PIB par habitant et le score du bonheur (variable cible), avec une p-value très faible (p-value = 0.0) qui indique que la corrélation observée entre la variable "Log_GDP_per_capita" et la variable "Life_Ladder" est statistiquement significative à un niveau de signification de 0,05 ou moins. Cela suggère que la relation entre ces deux variables est peu

probablement due au hasard. Il existe une relation linéaire forte entre ces deux variables : plus le PIB par habitant est élevé, plus le score du bonheur a tendance à être élevé.

Social_support : Il y a une corrélation positive significative (corrélation de Pearson = 0.7145383539876242) entre le soutien social et le score du bonheur, avec une p-value faible (p-value = 4.7465691e-316), ce qui suggère une relation linéaire importante entre ces deux variables. Cela indique que plus le soutien social est élevé, plus le score du bonheur a tendance à être élevé.

Healthy_life_expectancy_at_birth : Il y a une corrélation positive significative (corrélation de Pearson = 0.7503498544045164) entre l'espérance de vie en bonne santé à la naissance et le score du bonheur, avec une p-value très faible (p-value = 0.0), ce qui suggère une relation linéaire forte entre ces deux variables. Cela indique que plus l'espérance de vie en bonne santé à la naissance est élevée, plus le score du bonheur a tendance à être élevé.

Freedom_to_make_life_choices : Il y a une corrélation modérée (corrélation de Pearson = 0.5312107506137111) entre la liberté de faire des choix dans la vie et le score du bonheur, avec une p-value très faible (p-value = 8.952343695564951e-148), ce qui suggère une relation linéaire modérée entre ces deux variables.

Generosity : Il y a une corrélation positive faible mais significative (corrélation de Pearson = 0.17686409470812284) entre la générosité et le score du bonheur, avec une p-value très faible (p-value = 1.0989636108273123e-15), ce qui suggère une relation linéaire faible mais significative entre ces deux variables. Cette relation est moins forte que pour les autres variables explicatives.

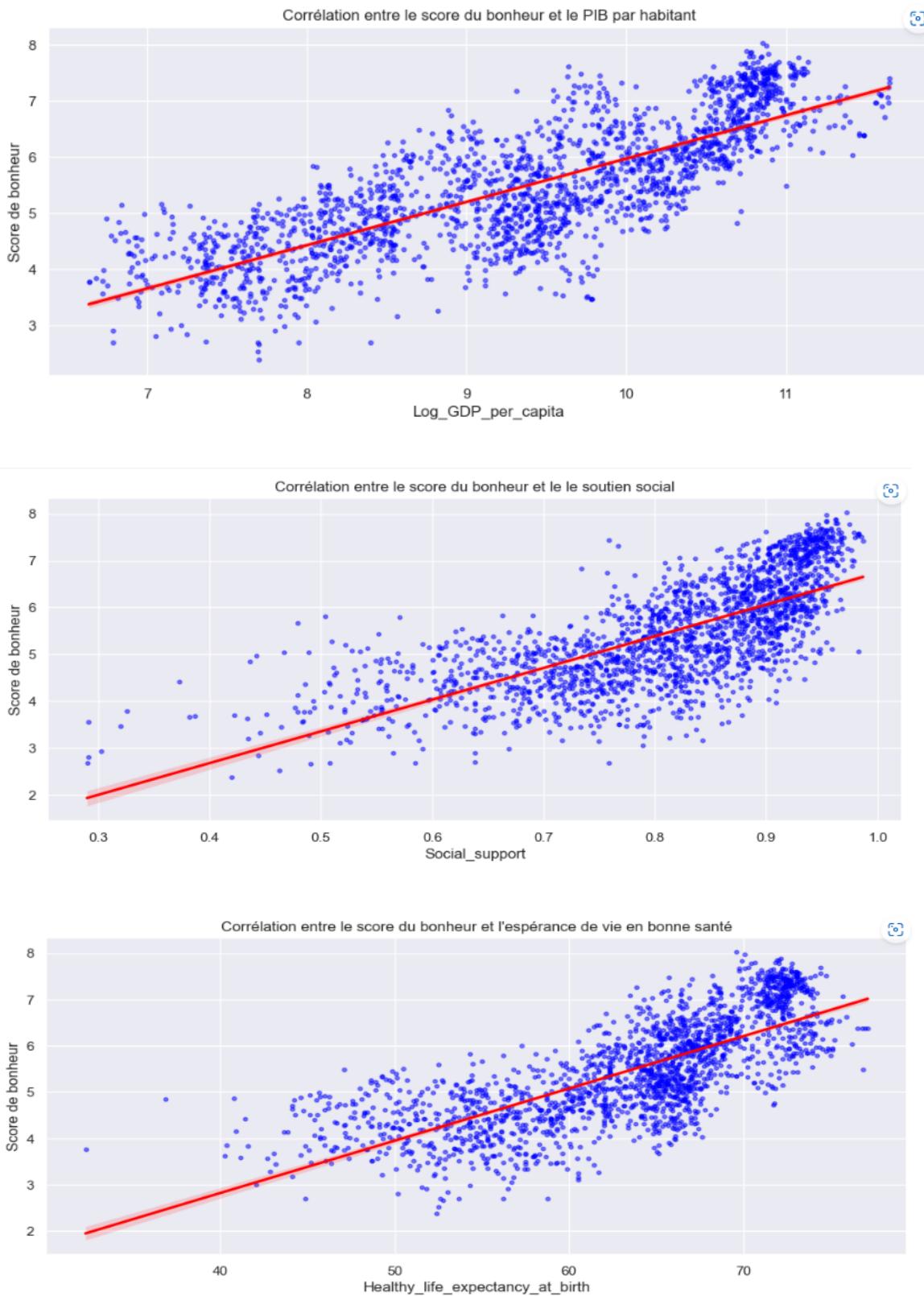
Perceptions_of_corruption : Il y a une corrélation négative modérée (corrélation de Pearson = -0.45418575645922193) entre les perceptions de la corruption et le score du bonheur, avec une p-value très faible (p-value = 1.419070944385589e-103), ce qui suggère une relation linéaire modérée entre ces deux variables.

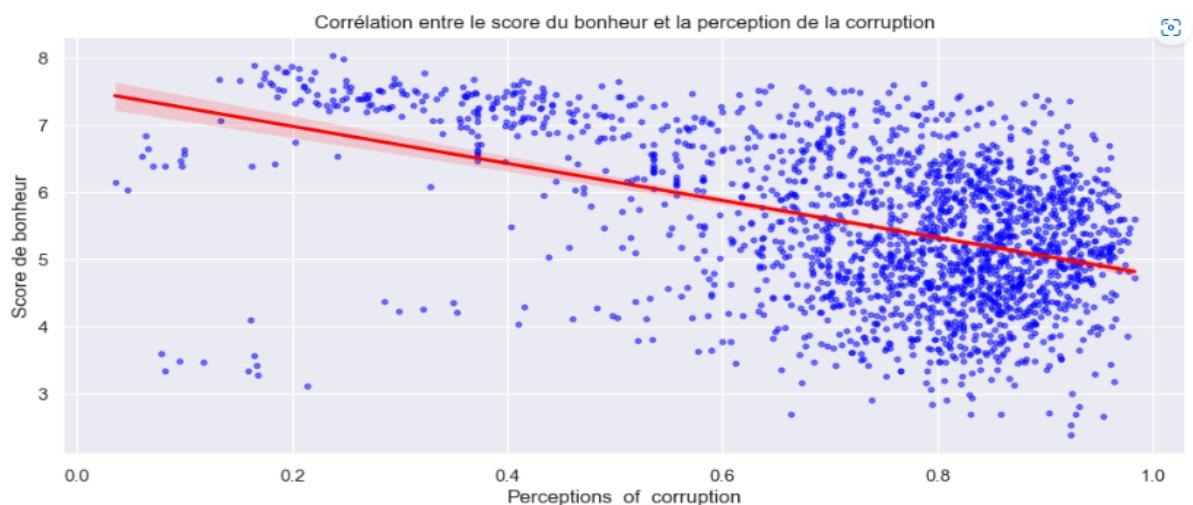
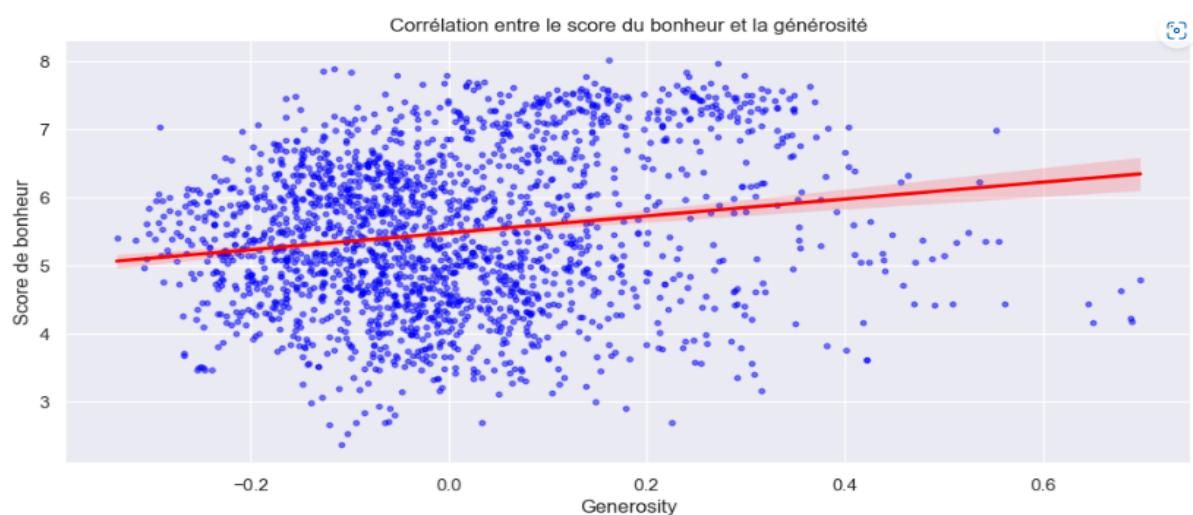
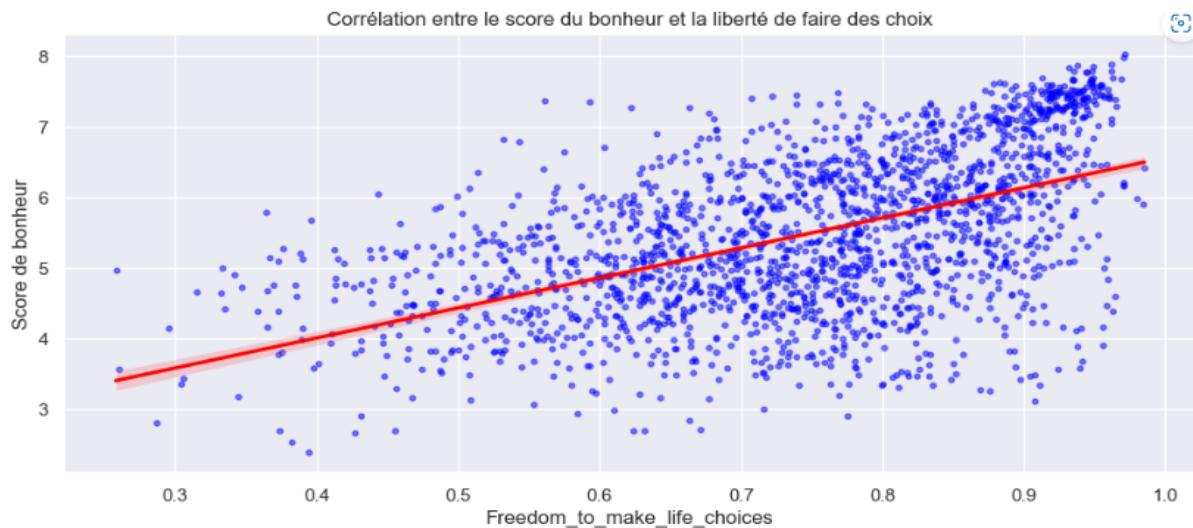
En conclusion, les variables **Log_GDP_per_capita**, **Social_support**, et **Healthy_life_expectancy_at_birth** montrent des corrélations positives fortes avec le score du bonheur, indiquant que ces facteurs ont une influence positive sur le bonheur perçu. La liberté de faire des choix dans la vie montre une corrélation modérée avec le score du bonheur, tandis que la générosité montre une corrélation faible mais significative. Les perceptions de la corruption montrent une corrélation négative modérée avec le score du bonheur, suggérant que moins il y a de corruption perçue, plus le bonheur est élevé.

Cependant, il est important de noter que la corrélation ne signifie pas forcément une relation de causalité, et d'autres facteurs non mesurés peuvent également influencer le bonheur des individus.

c) Représentation visuelle des corrélations entre les variables pour identifier les relations significatives.

Visualisation des corrélations entre les variables explicatives et la variable cible





Nous pouvons constater visuellement qu'effectivement les corrélations sont importantes entre la variable cible et les variables « Log_GDP_per_capita », « Social_support » et « Healthy_life_expectancy_at_birth ». Ces relations sont linéaires.

Nous optons donc pour l'étape de modélisation pour un algorithme d'apprentissage supervisé et pour un modèle de Régression Linéaire.

Feature scaling

Toutes nos valeurs sont quantitatives mais elles ne sont pas à la même échelle. C'est pourquoi nous devons normaliser les données avant d'entraîner nos algorithmes de Machine Learning.

La normalisation permet de ramener les variables à une même échelle, ce qui permet à toutes les variables d'avoir une influence équilibrée sur l'algorithme. La normalisation des données est une étape essentielle dans la préparation des données pour la mise en place d'un algorithme de machine learning. Elle permet de rendre les variables comparables, d'améliorer la convergence de l'algorithme, d'éviter les erreurs numériques et de faciliter l'interprétation des résultats. Elle est importante pour garantir des résultats précis et fiables.

Normalisation ou Standardisation ?

La Normalisation consiste à borner toutes les valeurs entre 0 et 1.

La Standardisation consiste à retrancher à chaque valeur la valeur moyenne de la variable et de la diviser par son écart type.

Avec la Normalisation, les prédictions ne pourront pas être inférieures à la valeur minimum actuelle et supérieures à la valeur maximale actuelle. C'est pourquoi nous devons choisir d'utiliser la STANDARDISATION (Z-Score). Nous utiliserons la méthode **StandardScaler**.

Cette étape sera effectuée après la séparation de notre jeu de données en jeu d'entraînement et jeu de test, pour éviter les fuites d'informations.

Concernant l'étape de Machine Learning, nous devrons utiliser plusieurs modèles, les évaluer, les améliorer et étudier les résultats en utilisant les métriques les plus appropriées pour choisir le modèle le plus performant.

IV) Identification de tendances et de modèles significatifs

a) Utilisation des techniques de Machine Learning pour analyser les données

MACHINE LEARNING

Nous avons décidé de tester différents modèles de Machine Learning de deux manières différentes. La première consiste à utiliser la méthode classique en divisant l'ensemble des données en train test split. La deuxième méthode consiste à entraîner le modèle sur toutes les années à l'exception de 2021, puis à le tester et l'évaluer spécifiquement sur les données de cette année-là.

LES MÉTRIQUES

Dans une prédition :

- L'**Erreur** = valeur réelle - valeur prédite, donc celle-ci peut être positive ou négative.
- L'**Erreur quadratique** = la valeur réelle - valeur prédite au carré, (elle est donc forcément positive).
- **MSE** (Mean squared error / Erreur quadratique moyenne) = la moyenne de toutes les Erreurs quadratiques.
- **RMSE** (Root mean squared error / racine carrée de l'Erreur quadratique moyenne) = `np.sqrt` de la MSE. Cette métrique permet de remettre à l'échelle les erreurs calculées par la MSE
- L'**Erreur absolue** est la valeur absolue de l'Erreur (elle du coup forcément positive). **MAE** (Mean absolute error / Erreur absolue moyenne) = la moyenne des erreurs absolues.

La **MSE** est utilisée lorsqu'on accorde une importance exponentielle aux erreurs. C'est-à-dire que la MSE pénalise beaucoup plus les grandes erreurs. Exemple : pour un estimateur de distance de freinage : une erreur de 10 mètres n'est pas 10 fois plus grave qu'une erreur de 1 mètre mais peut être 100 fois plus grave !

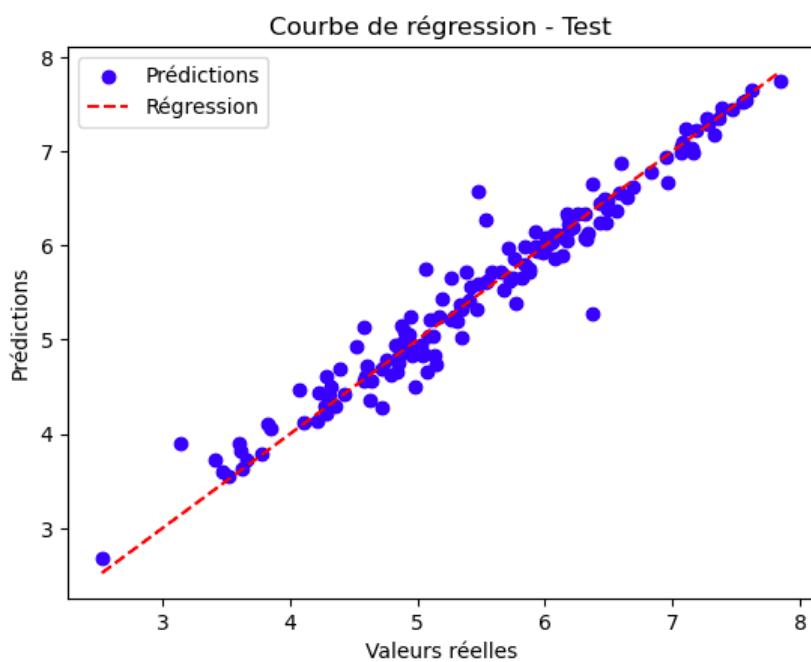
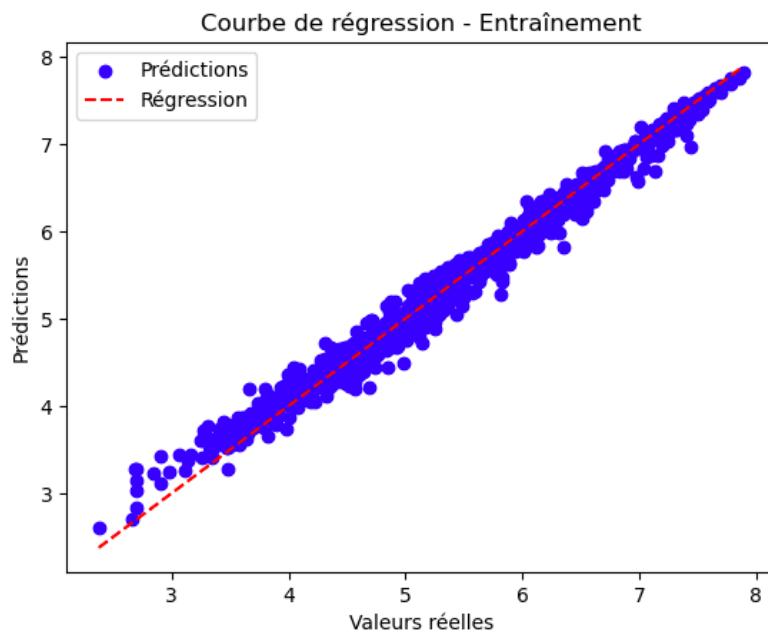
La **MAE** quant à elle est utilisée si l'importance d'une erreur est linéaire et si les plus grandes erreurs sont surtout dues aux outliers. Dans ce cas, la MAE donnera une meilleure représentation de la performance du modèle. La **MEDIAN ABSOLUTE ERROR** = la médiane de toutes les Erreurs : elle est peu sensible aux outliers.

Le **Coefficient de détermination (R2)** évalue la performance du modèle par rapport au niveau de variation présent dans les données. $= 1 - \text{Somme des erreurs quadratiques} / \text{la Variance}$. Le R2 permet de savoir à quel niveau le modèle décrit les variations de la variable cible. Exemple : le modèle décrit 90% des variations du score du bonheur !

Comparaison de divers algorithmes de Machine Learning sur l'ensemble de test de l'année 2021 (séparation de l'année 2021) :

Courbe de régression d'apprentissage et de test pour RandomForest Regressor

Le RandomForest Regressor est un algorithme d'apprentissage automatique qui utilise un ensemble d'arbres de décision pour effectuer des prédictions précises et robustes dans des problèmes de régression.

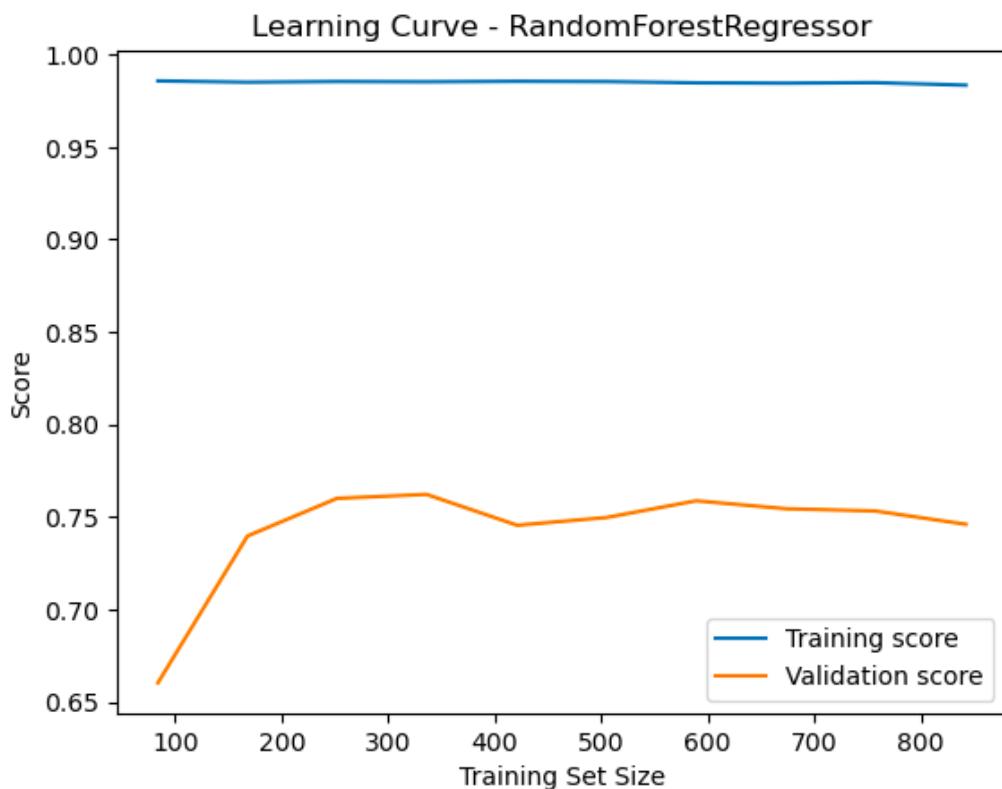


Les courbes de régressions montrent une bonne adéquation entre les prédictions et les valeurs réelles du modèle. En effet les prédictions semblent suivre étroitement les valeurs réelles indiquant une bonne performance du modèle.

LEARNING CURVE

La fonction de la courbe d'apprentissage (learning curve) est d'évaluer les performances d'un modèle d'apprentissage automatique en fonction de la quantité de données d'entraînement disponibles. Elle permet de visualiser comment les performances du modèle évoluent à mesure que la taille de l'ensemble d'entraînement augmente.

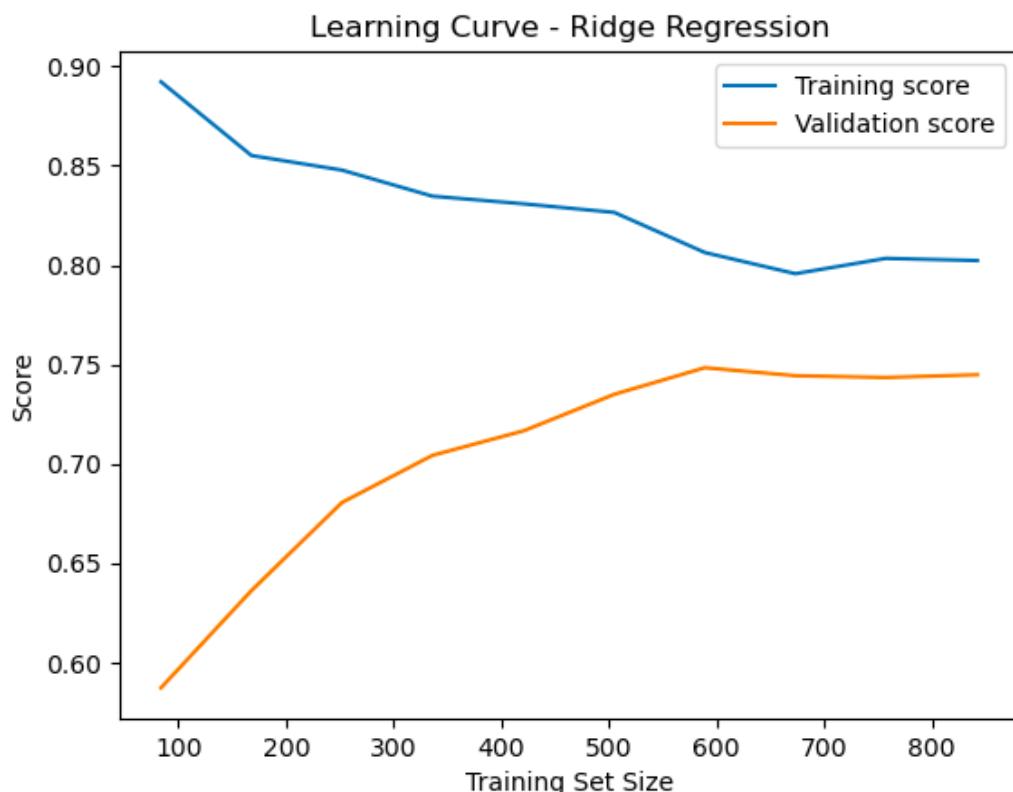
La courbe d'apprentissage trace le score de performance du modèle (par exemple, précision, erreur quadratique moyenne, etc.) en fonction du nombre d'exemples d'entraînement utilisés. Elle permet de détecter des problèmes tels que le surapprentissage (overfitting) ou le sous-apprentissage (underfitting).

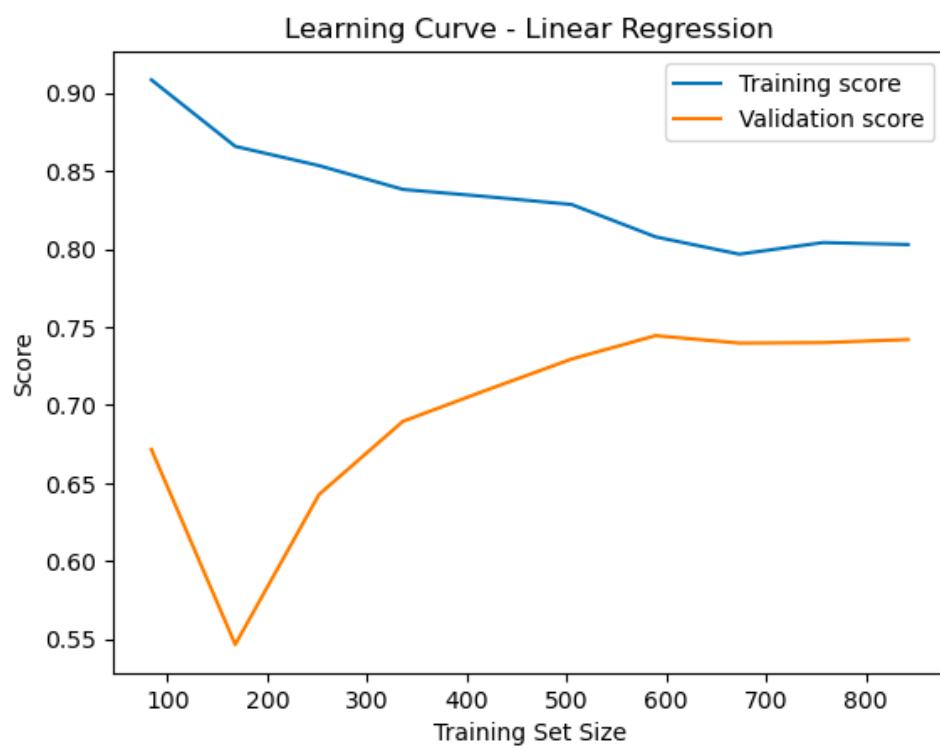


Si la courbe d'apprentissage pour l'ensemble d'entraînement reste horizontale à un score de 1, cela pourrait signifier que le modèle est capable de parfaitement s'adapter aux données d'entraînement. Cela peut se produire lorsque le modèle est surajusté (overfitting) aux données d'entraînement.

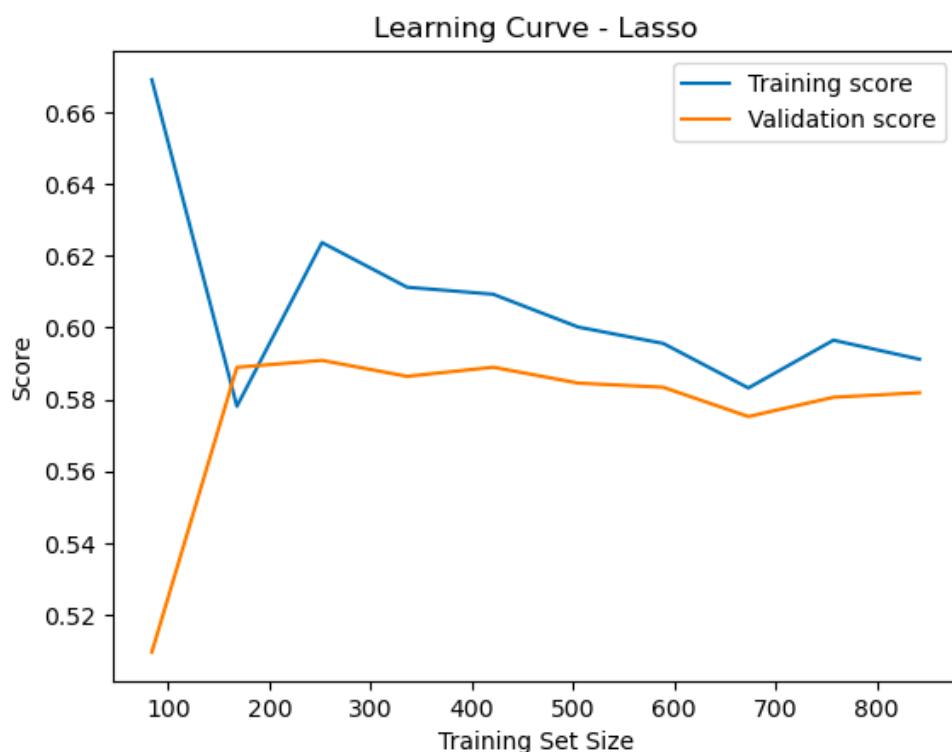
Il y a quelques raisons possibles pour lesquelles cela se produit :

- **Sur-ajustement (overfitting)** : Le modèle mémorise les données d'entraînement au lieu de généraliser.
- **Données d'entraînement limitées** : Si la taille de l'ensemble d'entraînement est petite, il est possible que le modèle puisse mémoriser les données sans avoir à généraliser. L'ajout de plus de données d'entraînement peut aider à améliorer la performance du modèle.

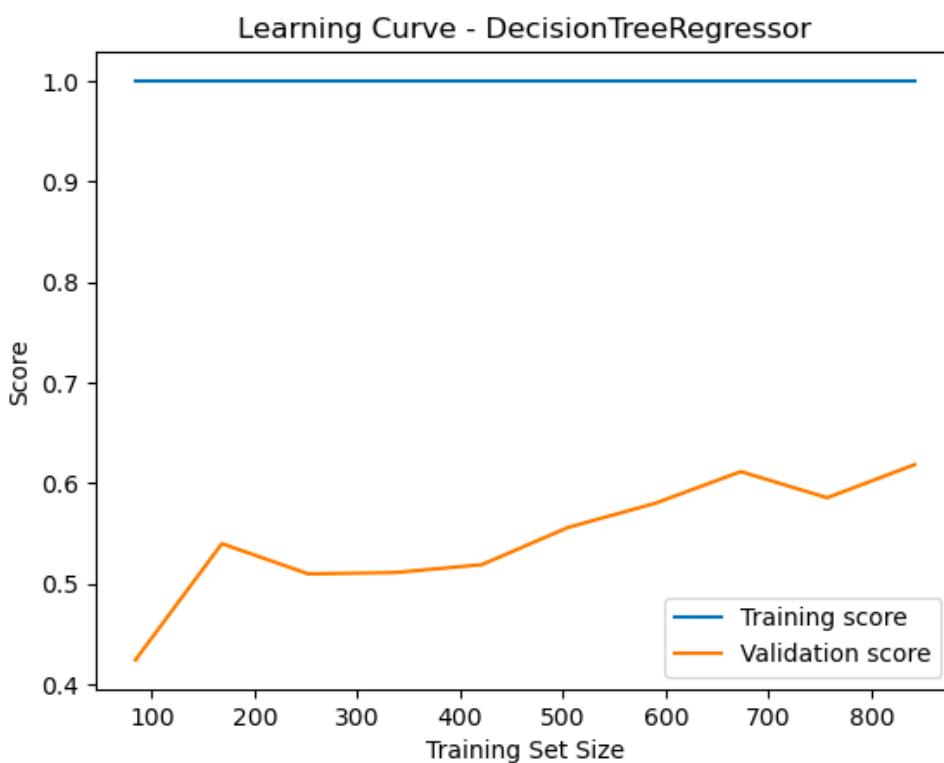




Avec un trainsetsize de 600 valeurs, le modèle commence à présenter une cohérence notable, où les scores de test et d'entraînement se rapprochent progressivement.



À partir d'un trainsetsize de 750/800, le modèle montre une certaine cohérence, mais ses performances restent modérées.

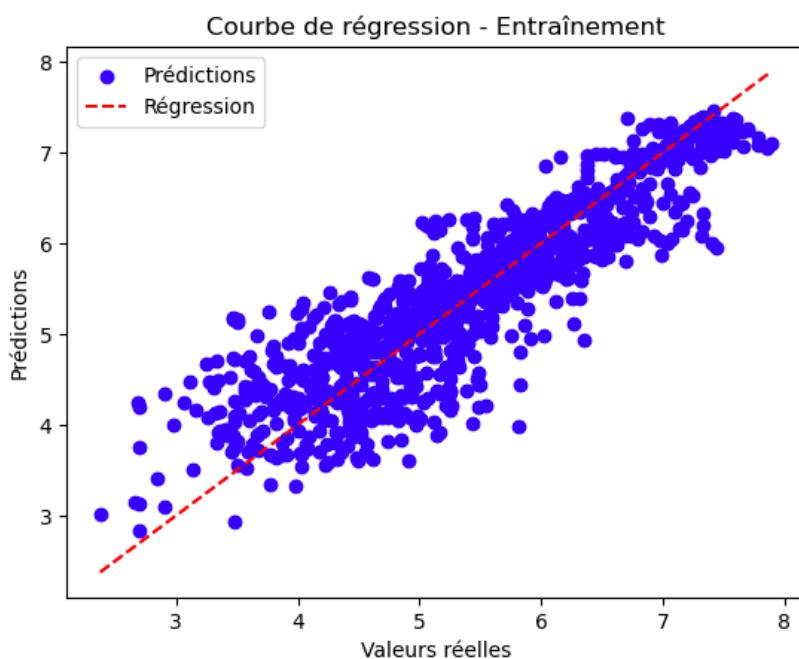


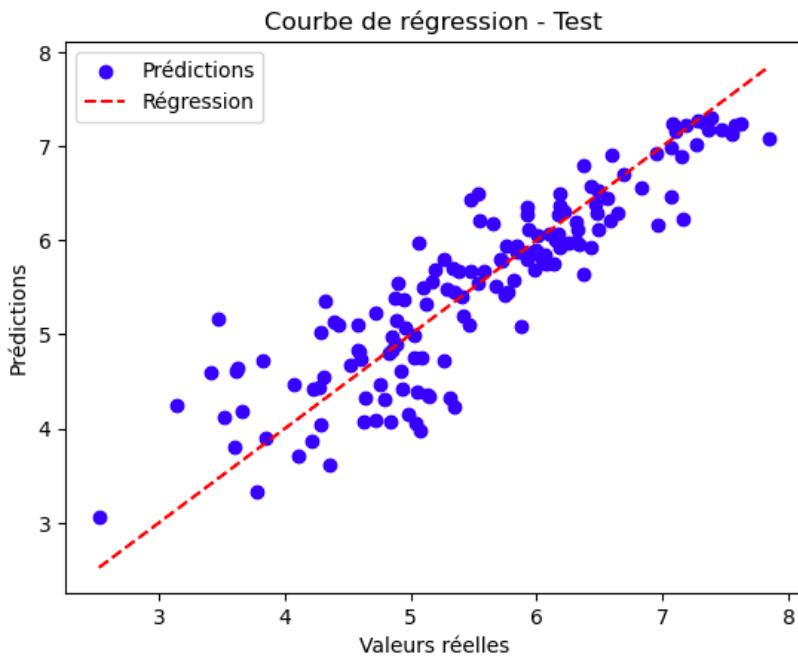
Lorsque la courbe d'apprentissage de l'ensemble d'entraînement reste horizontale (stagnante) avec une valeur de 1, cela indique généralement que le modèle est en sur-apprentissage. Il y a un écart non cohérent entre la courbe d'apprentissage et la courbe de validation de score.

```
(      model      mse      rmse      r2      mae
0  LinearRegression  0.241572  0.491500  0.789125  0.377233
1          Lasso  1.147823  1.071365 -0.001966  0.877528
2          Ridge  0.241518  0.491445  0.789173  0.377214
3 DecisionTreeRegressor  0.099188  0.314942  0.913416  0.202664
4 RandomForestRegressor  0.056624  0.237958  0.950571  0.164801,
      model      mse      rmse      r2      mae
0  LinearRegression  0.259414  0.509327  0.798572  0.389878
1          Lasso  1.287871  1.134844  0.000000  0.945533
2          Ridge  0.259414  0.509327  0.798571  0.389878
3 DecisionTreeRegressor  0.000000  0.000000  1.000000  0.000000
4 RandomForestRegressor  0.021650  0.147138  0.983190  0.108021)
```

Courbe de régression d'apprentissage et de test pour RIDGE

La fonction de régularisation Ridge, également connue sous le nom de régression Ridge, est utilisée dans les problèmes de régression pour réduire le sur apprentissage (overfitting) et améliorer la généralisation du modèle. Elle est basée sur la pénalisation des coefficients du modèle pour les maintenir à des valeurs plus petites.



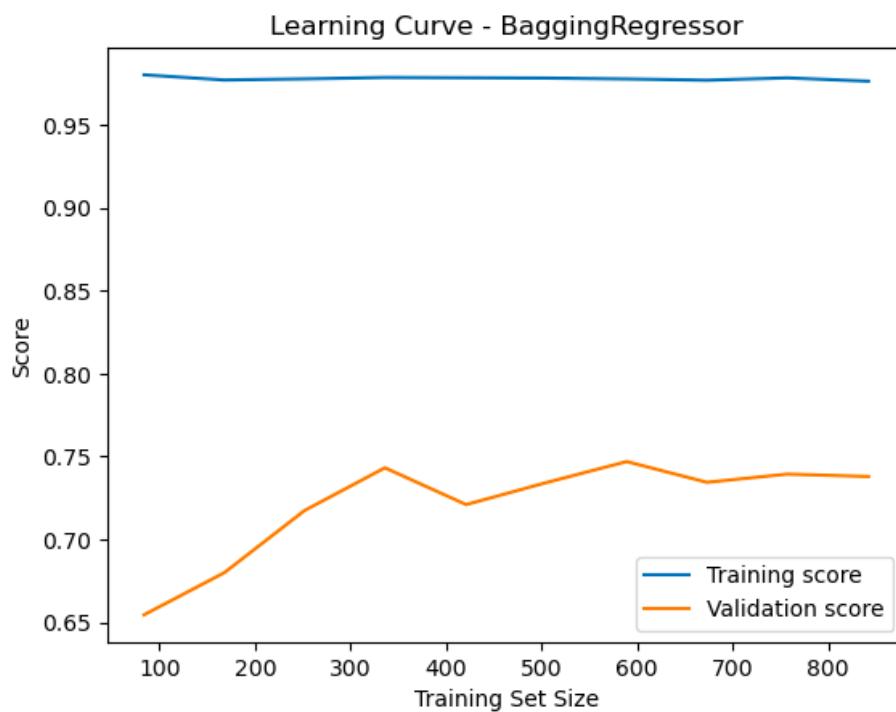


Pour ces deux courbes de régression, on remarque que plus les scores du bonheur sont élevés, plus nos prédictions sont précises.

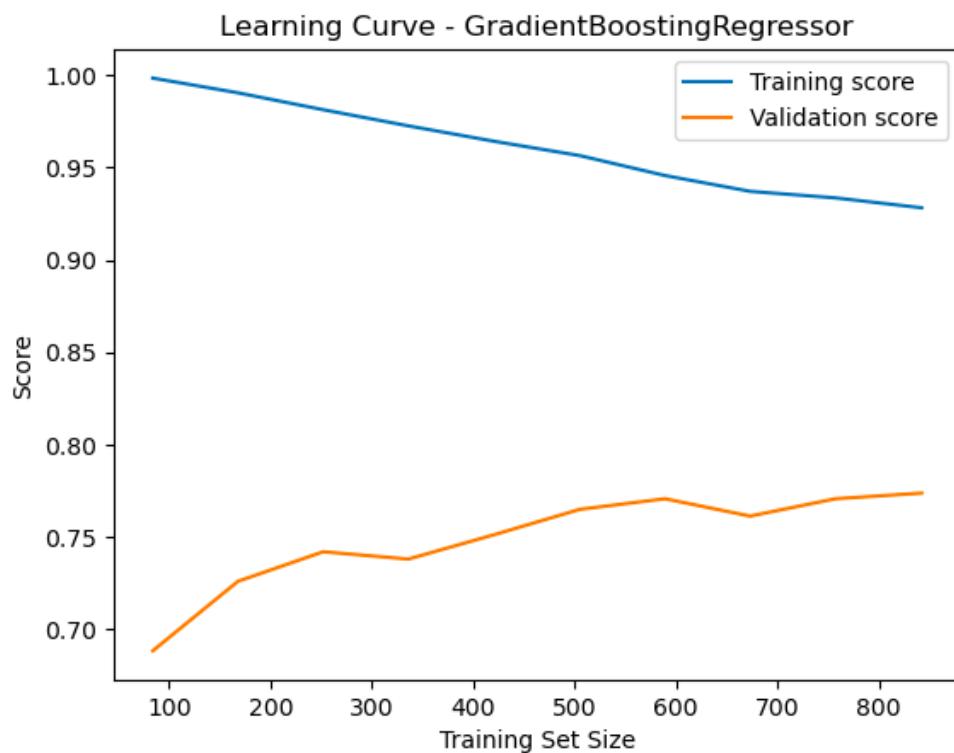
BOOSTING ET BAGGING

Le boosting et le bagging sont deux techniques d'ensemble utilisées pour améliorer les performances des modèles prédictifs en combinant les prédictions de plusieurs modèles individuels. Le boosting construit des modèles successifs en mettant l'accent sur les erreurs de prédiction des modèles précédents, tandis que le bagging entraîne des modèles indépendants sur des sous-ensembles aléatoires des données d'entraînement.

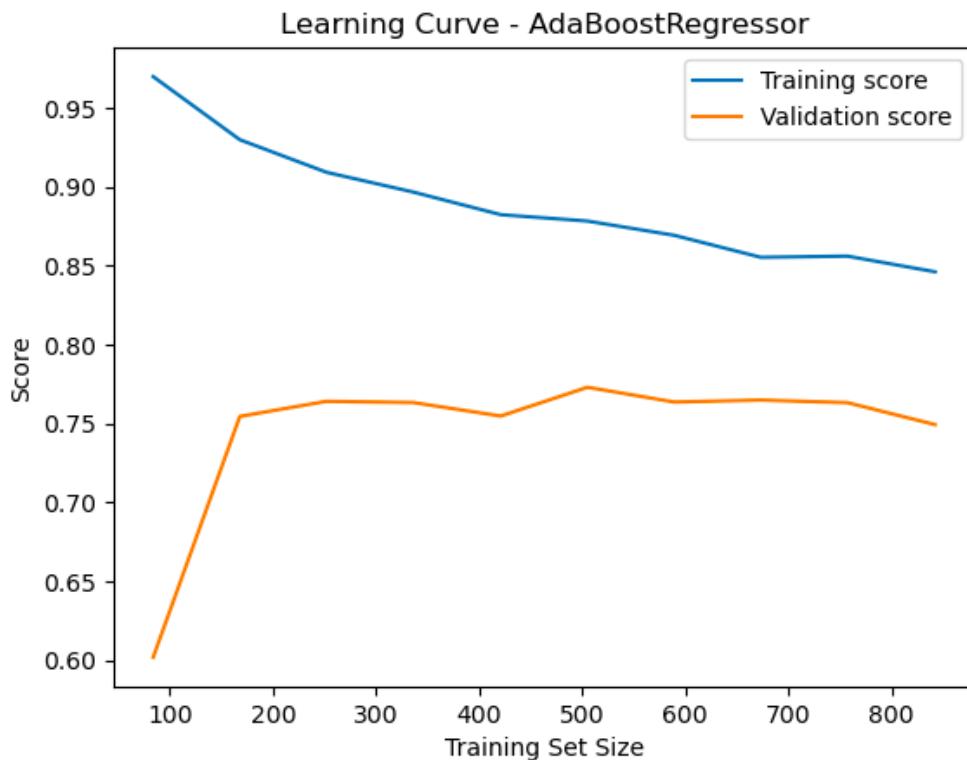
		model	mse	rmse	r2	mae
0		BaggingRegressor	0.074783	0.273465	0.934720	0.174058
1		GradientBoostingRegressor	0.120530	0.347174	0.894786	0.268643
2		AdaBoostRegressor	0.211777	0.460193	0.815134	0.377700,
		model	mse	rmse	r2	mae
0		BaggingRegressor	0.029994	0.173189	0.976710	0.121196
1		GradientBoostingRegressor	0.109331	0.330653	0.915107	0.255767
2		AdaBoostRegressor	0.204933	0.452696	0.840874	0.370706)



Nous pouvons remarquer le même problème que le DecisionTreeRegressor plus haut.



Un écart significatif entre le score d'entraînement et le score de validation, ce qui suggère que le modèle nécessite un grand volume de données pour atteindre de bonnes performances. En d'autres termes, le modèle semble avoir besoin d'un ensemble de données volumineux pour obtenir des résultats satisfaisants.



À mesure que le volume de données augmente, la performance du modèle diminue progressivement.

b) Sélection de l'algorithme approprié pour atteindre les objectifs de l'étude

CHOIX DE L'ALGORITHME POUR LA PRÉDICTION DU SCORE DU BONHEUR

Le choix se porte sur l'algorithme de la régression Ridge en se basant sur les métriques R2 et MAE et les courbes d'apprentissage (learning curves) où on peut observer la stabilité et la cohérence entre le score d'entraînement et le score de validation.

Le coefficient de détermination R2 de 0.789173 indique que le modèle explique environ 79% de la variance totale de la variable cible. Cela signifie que les variables explicatives incluses dans le modèle expliquent une grande partie de la variation observée dans les valeurs réelles de la variable cible. Une valeur de R2 élevée suggère une bonne adéquation du modèle aux données.

L'erreur absolue moyenne (MAE) est une mesure de la performance du modèle qui quantifie l'erreur moyenne entre les prédictions du modèle et les valeurs réelles. Une MAE de 0.377214 signifie qu'en moyenne, les prédictions du modèle diffèrent de 0.377214 unités de la valeur réelle.

Dans notre cas, cela signifie que les prédictions sont en moyenne proches des valeurs réelles.
La régression Ridge est une méthode de régularisation qui permet d'éviter le surajustement.

c) Ajustement des hyperparamètres pour améliorer les performances du modèle

AJUSTEMENT DES HYPERPARAMÈTRES

L'introduction au processus d'ajustement des hyperparamètres pour la régression Ridge implique plusieurs étapes clés. Dans cette approche, nous cherchons à trouver les meilleurs paramètres pour notre modèle en utilisant une technique appelée recherche par grille (GridSearch) :

1 - Définir une grille de valeurs pour les hyper paramètres à ajuster. Dans le cas de la régression Ridge, le principal hyperparamètre à régler est l'alpha, qui contrôle le degré de régularisation.

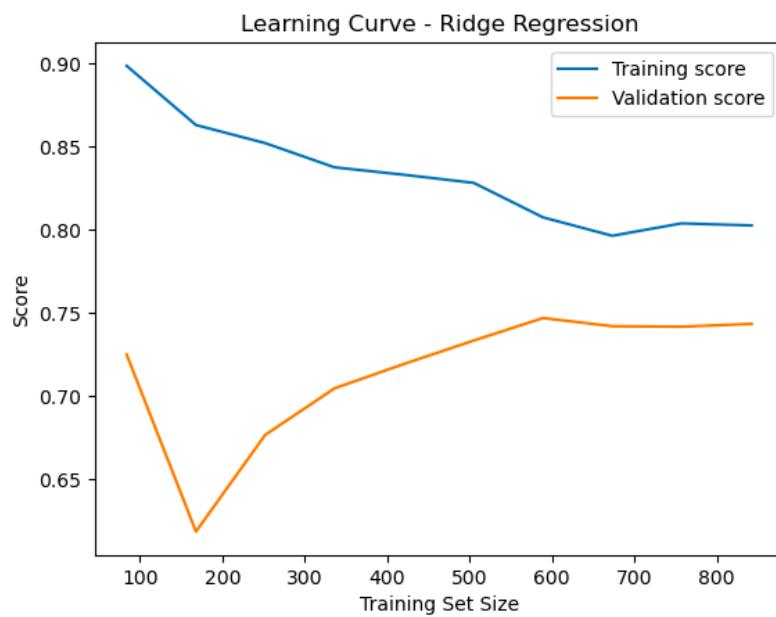
2 - Utiliser la fonction GridSearchCV pour effectuer la recherche des meilleurs hyperparamètres en utilisant une validation croisée.

3 - Appliquer la recherche des hyperparamètres sur le jeu de données d'entraînement pour ajuster les modèles Ridge correspondant à chaque combinaison d'hyperparamètres et évaluer leur performance à l'aide de la validation croisée.

4 - Obtenir les meilleurs hyperparamètres en utilisant l'attribut best_params_ de l'objet GridSearchCV.

5 - Réentraîner le modèle Ridge en utilisant les meilleurs hyperparamètres sur l'ensemble de données d'entraînement complet.

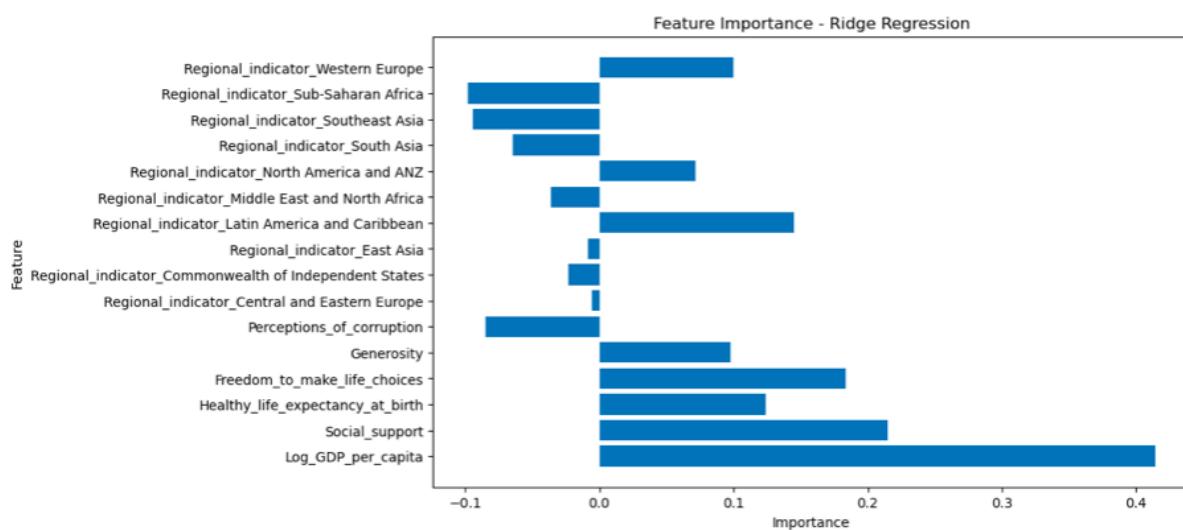
```
MSE:  0.24110034491084523
RMSE:  0.4910196991067112
R2:  0.7895370566874194
MAE:  0.37706380938803846
```

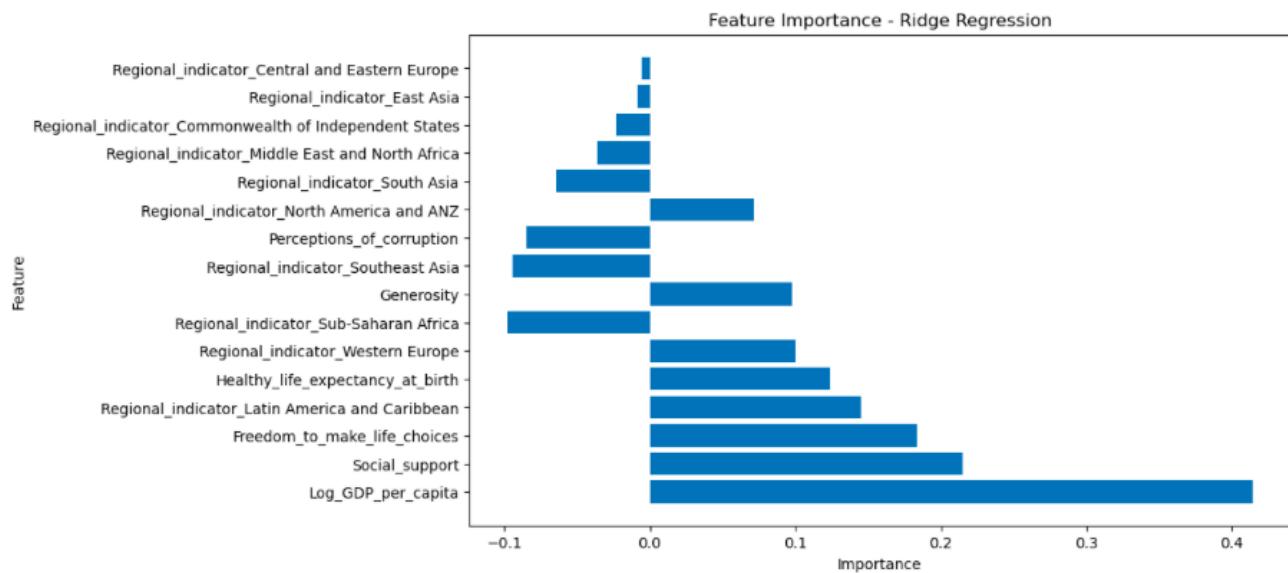


Malgré une légère amélioration, l'ajustement des hyperparamètres n'a pas entraîné d'augmentation significative des performances du modèle.

FEATURE IMPORTANCE

La feature importance est une mesure utilisée en apprentissage automatique pour évaluer l'influence relative de chaque caractéristique (ou variable) sur les prédictions d'un modèle. Elle permet de déterminer quelles caractéristiques sont les plus informatives et contribuent le plus à la capacité prédictive du modèle.





```

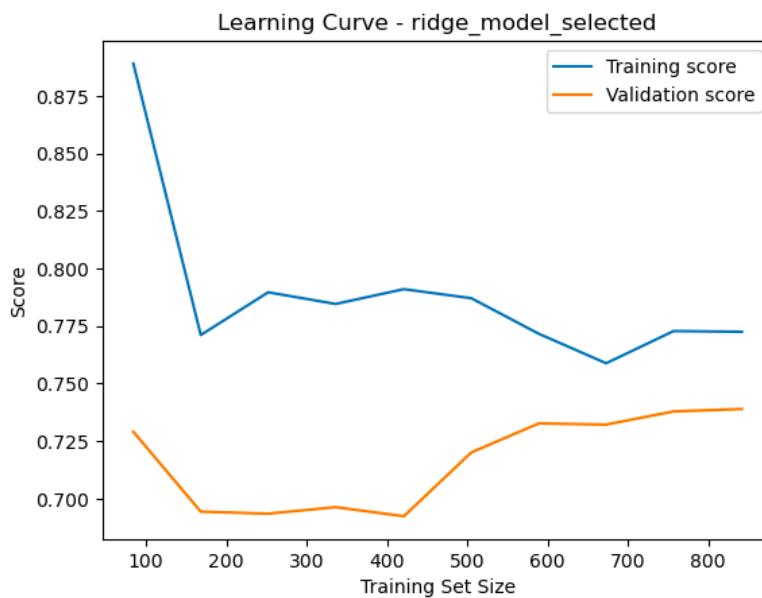
Nombre optimal de variables sélectionnées: 8
Variables sélectionnées: Index(['Log_GDP_per_capita', 'Social_support', 'Freedom_to_make_life_choices',
   'Generosity', 'Perceptions_of_corruption',
   'Regional_indicator_Latin America and Caribbean',
   'Regional_indicator_North America and ANZ',
   'Regional_indicator_Western Europe'],
  dtype='object')

```

```

MSE (Mean Squared Error) : 0.2757635257031133
RMSE (Root Mean Squared Error) : 0.5251319126687248
R2 Score : 0.759278638530388
MAE (Mean Absolute Error) : 0.40867884769527896

```



Finalement le modèle est plus performant lorsqu'il utilise toutes les variables explicatives.

MODIFICATION DE LA MÉTHODE DE MACHINE LEARNING SELON L'ÉTUDE DES RÉSULTATS OBTENUS

Nous devons à présent adapter notre méthodologie car nous avons observé deux problèmes :

1 - Les régions en tant que variables explicatives introduisent des biais et des incohérences. En effet, nous avons pu observer dans la Feature Importance que certaines régions étaient plus importantes que des variables explicatives directement liées aux scores du bonheur. De plus, l'importance d'une région va être également liée aux nombres de valeurs qu'elle contient. C'est-à-dire que la distribution du jeu de données selon les régions va être déterminante.

Nous décidons donc de ne pas retenir les régions en tant que variables explicatives pour notre modèle de Machine Learning.

2 - L'utilisation de l'année 2021 en tant que jeu de données de test ne permet pas d'avoir suffisamment de données car nous nous retrouvons avec un test size d'à peine 15%. Les résultats sur une ensemble de test réduit peuvent être plus sensibles aux fluctuations aléatoires et peuvent ne pas être représentatifs de la performance réelle du modèle sur de nouvelles données.

De plus, l'année 2021 étant la plus récente, peut comporter des caractéristiques uniques et des événements spécifiques qui ne se reproduisent pas régulièrement. En utilisant cette année comme ensemble de test, il existe un risque de biais temporel, car notre modèle pourrait être trop ajusté à ces événements spécifiques et ne pas généraliser de manière optimale sur de nouvelles données.

Nous décidons donc de ne pas séparer l'année 2021 de l'ensemble du dataset afin de permettre à notre modèle de prédire de manière plus robuste les scores du bonheur sur de nouvelles données.

Évaluation de plusieurs algorithmes en excluant les régions non incluses, tout en utilisant l'année 2021 comme ensemble de données :

```
In [3]: df = pd.read_csv(fp)
df.head()
```

	Regional_indicator	Country_name	year	Life_Ladder	Log_GDP_per_capita	Social_support	Healthy_life_expectancy_at_birth
0	South Asia	Afghanistan	2008	3.724	7.370	0.451	50.80
1	South Asia	Afghanistan	2009	4.402	7.540	0.552	51.20
2	South Asia	Afghanistan	2010	4.758	7.647	0.539	51.60
3	South Asia	Afghanistan	2011	3.832	7.620	0.521	51.92
4	South Asia	Afghanistan	2012	3.783	7.705	0.521	52.24

In [4]:

```
# Séparation des variables explicatives de la variable cible
X = df.drop(['Life_Ladder', 'year', 'Country_name', 'Regional_indicator'], axis=1)
y = df['Life_Ladder']
```

In [6]:

```
# Diviser les données en ensemble d'entraînement et ensemble de test
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.2, random_state=42)
```

In [7]:

```
display(X_train.shape)
display(X_test.shape)
```

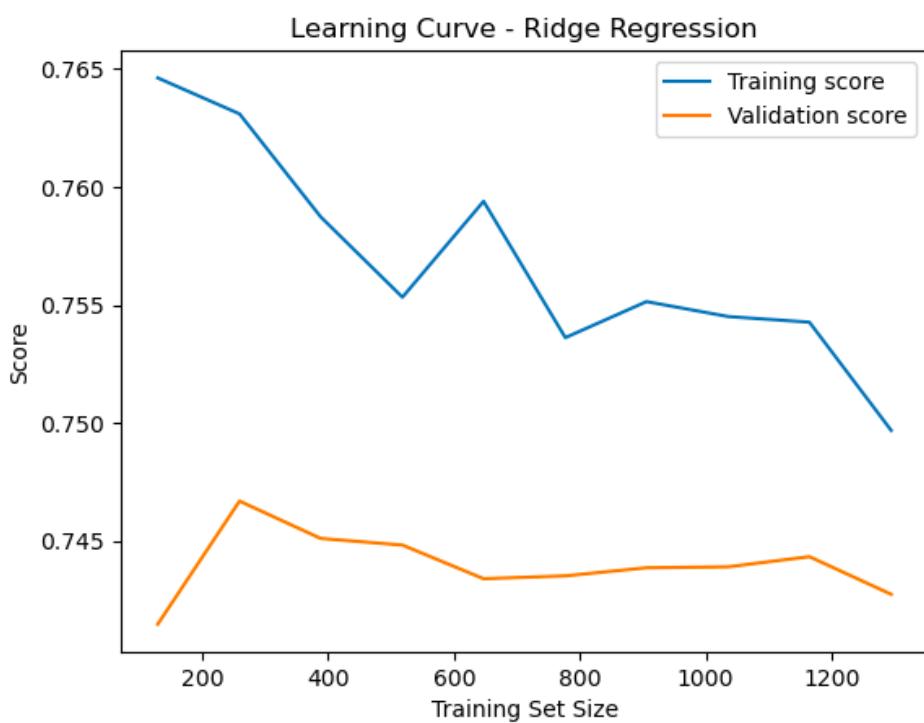
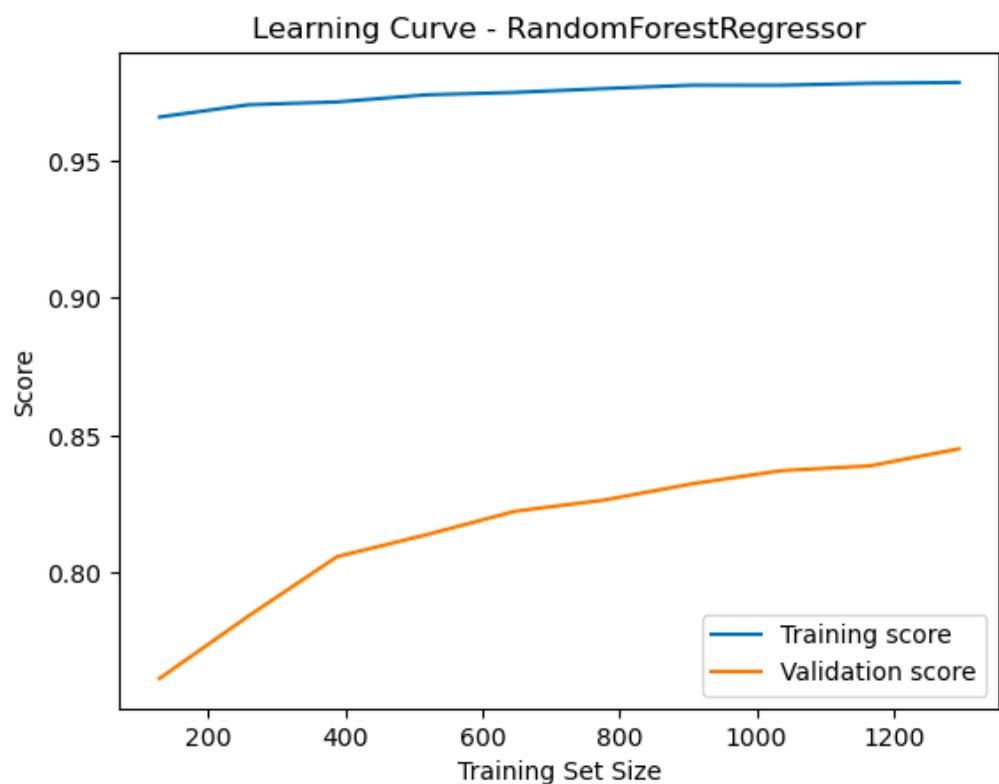
```
(1619, 6)
(405, 6)
```

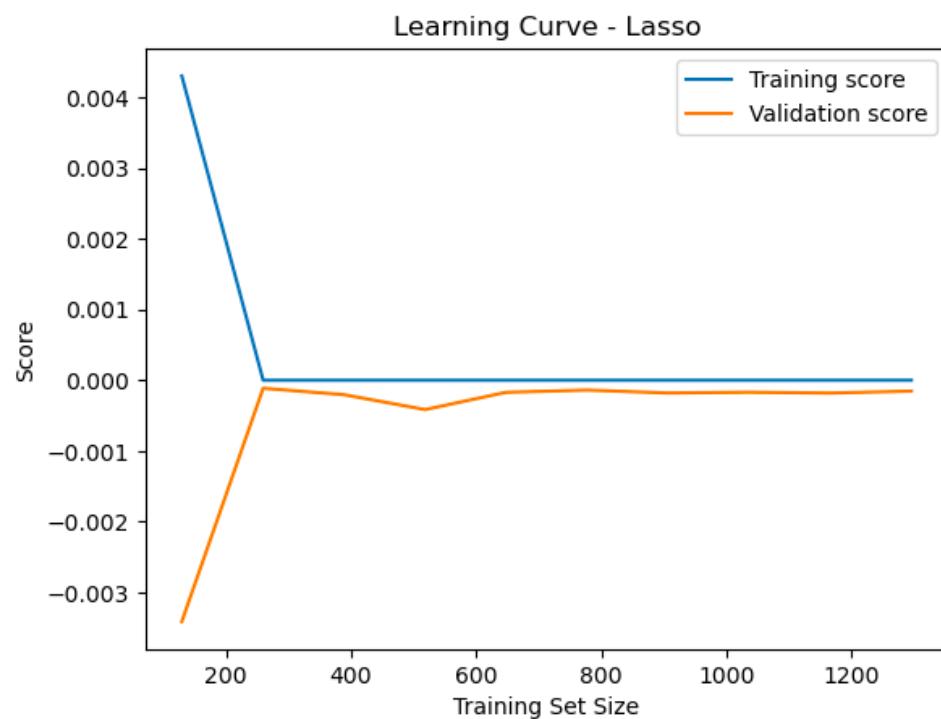
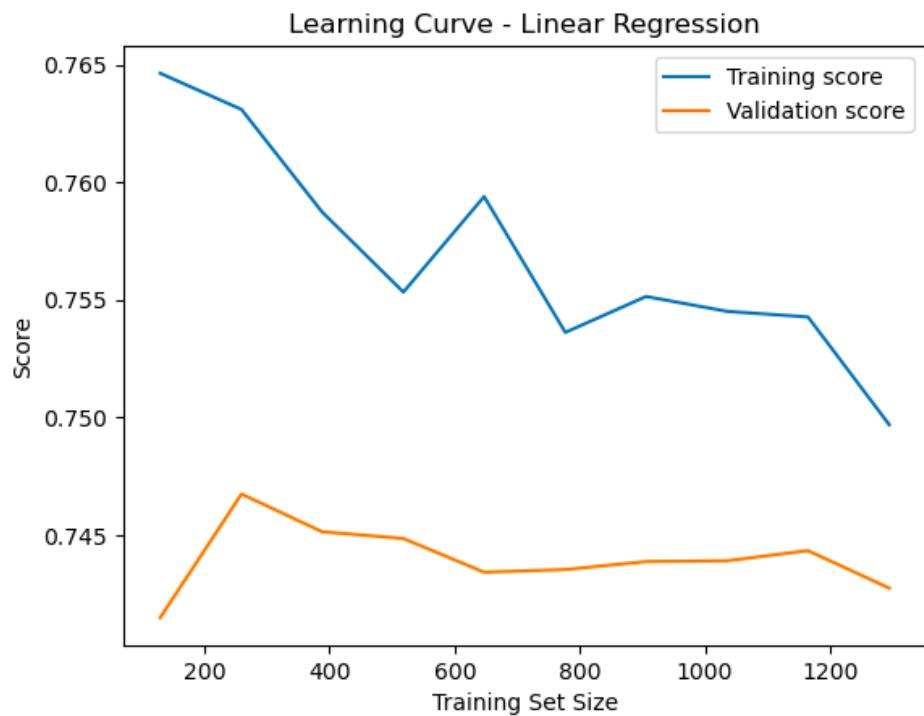
Out[8]:

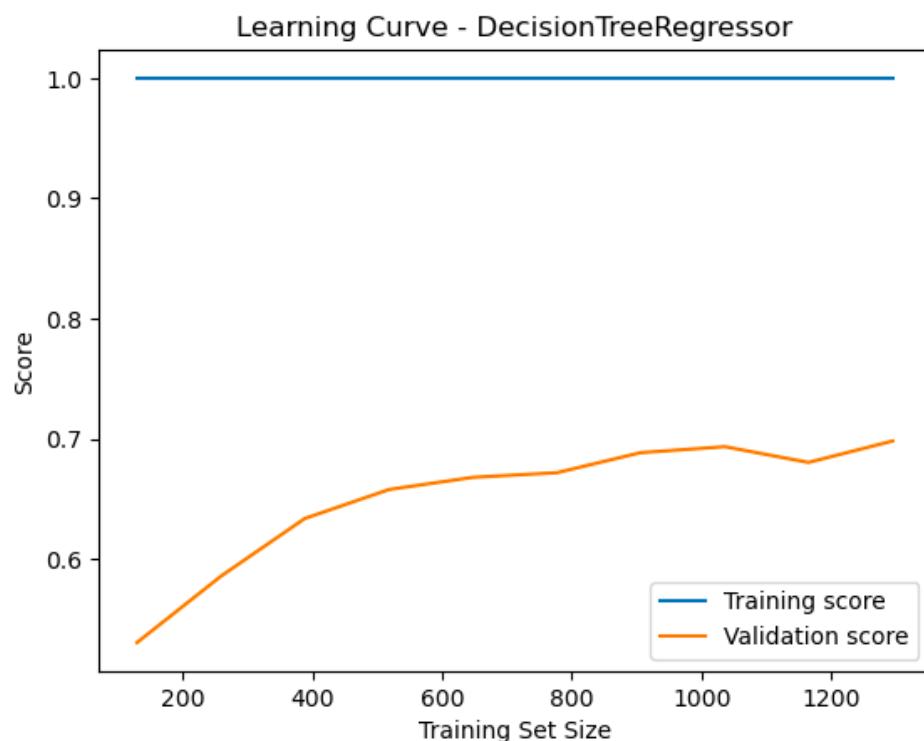
	model	mse	rmse	r2	mae
0	LinearRegression	0.293944	0.542166	0.766619	0.415058
1	Lasso	1.259534	1.122290	-0.000024	0.933234
2	Ridge	0.293952	0.542173	0.766613	0.415062
3	DecisionTreeRegressor	0.337905	0.581296	0.731716	0.418741
4	RandomForestRegressor	0.162418	0.403011	0.871046	0.303299,
	model	mse	rmse	r2	mae
0	LinearRegression	0.317312	0.563305	0.748979	0.436069
1	Lasso	1.264085	1.124315	0.000000	0.937425
2	Ridge	0.317312	0.563305	0.748979	0.436085
3	DecisionTreeRegressor	0.000000	0.000000	1.000000	0.000000
4	RandomForestRegressor	0.026854	0.163871	0.978756	0.120252)

Résultats de la fonction qui teste plusieurs algorithmes.

LEARNING CURVE







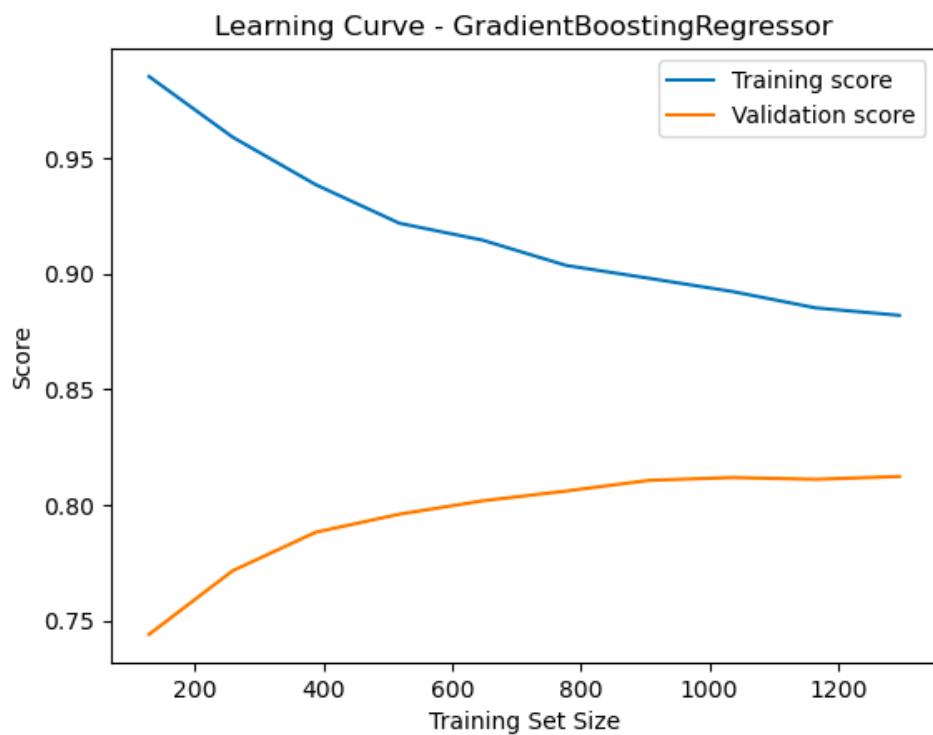
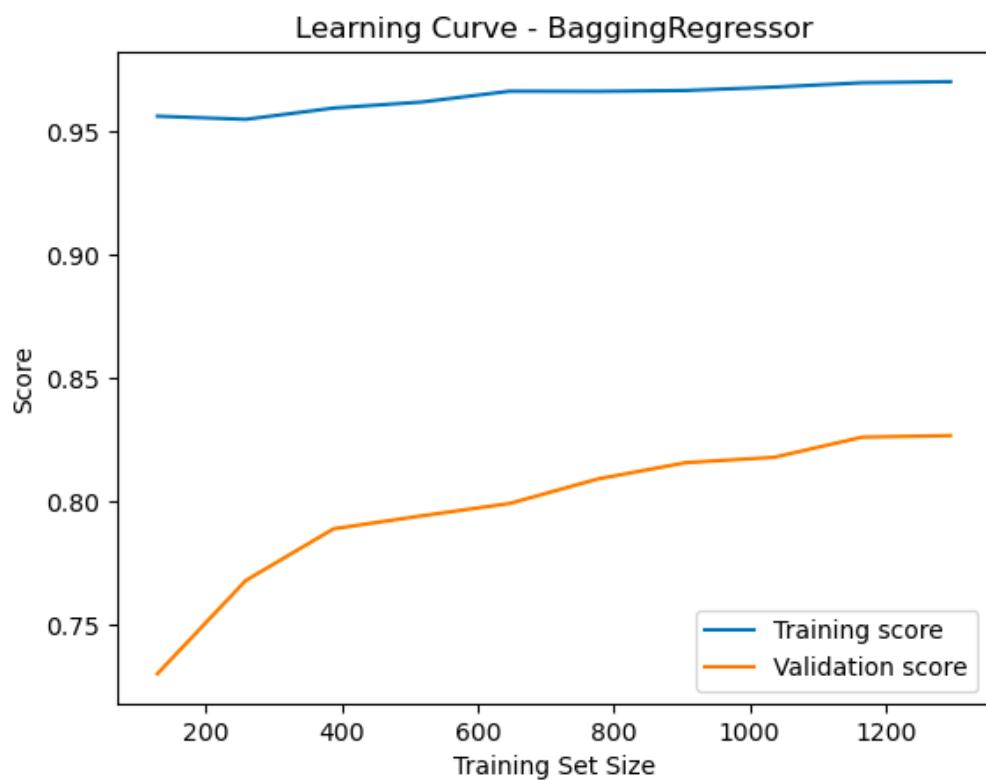
Dans notre cas, les algorithmes testés sont soit en situation d'overfitting soit en manque de stabilité (pour le Ridge regression et le Linear regression, les performances chutes à partir d'un training size de 1200)

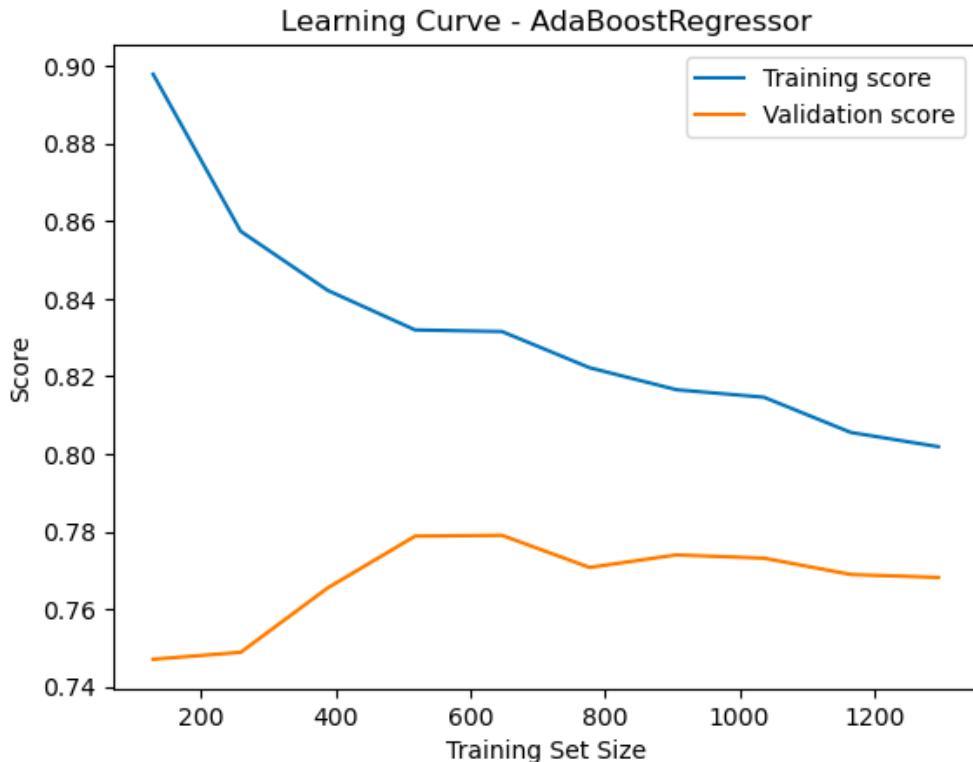
- Pour palier à ces soucis, nous allons tester des algorithmes d'ensemble :

Les algorithmes d'ensemble permettent souvent d'obtenir de meilleures performances prédictives par rapport à un seul modèle. En combinant les prédictions de plusieurs modèles, les algorithmes d'ensemble peuvent réduire le biais et la variance, ce qui conduit généralement à des résultats plus précis.

L'un des principaux problèmes en apprentissage automatique est le surapprentissage, c'est-à-dire lorsque le modèle est trop spécialisé dans les données d'entraînement et ne se généralise pas bien aux nouvelles données. Les algorithmes d'ensemble, comme le bagging et le boosting, permettent de réduire le surapprentissage en agrégant les prédictions de plusieurs modèles différents, ce qui favorise la généralisation.

BOOSTING ET BAGGING





CHOIX DE L'ALGORITHME POUR LA PRÉDICTION DU SCORE DU BONHEUR

Le choix se porte sur l'algorithme **GradientBoostingRegressor** en se basant sur les métriques R2 et MAE et les courbes d'apprentissage (learning curves) où on peut observer la stabilité et la cohérence entre le score d'entraînement et le score de validation.

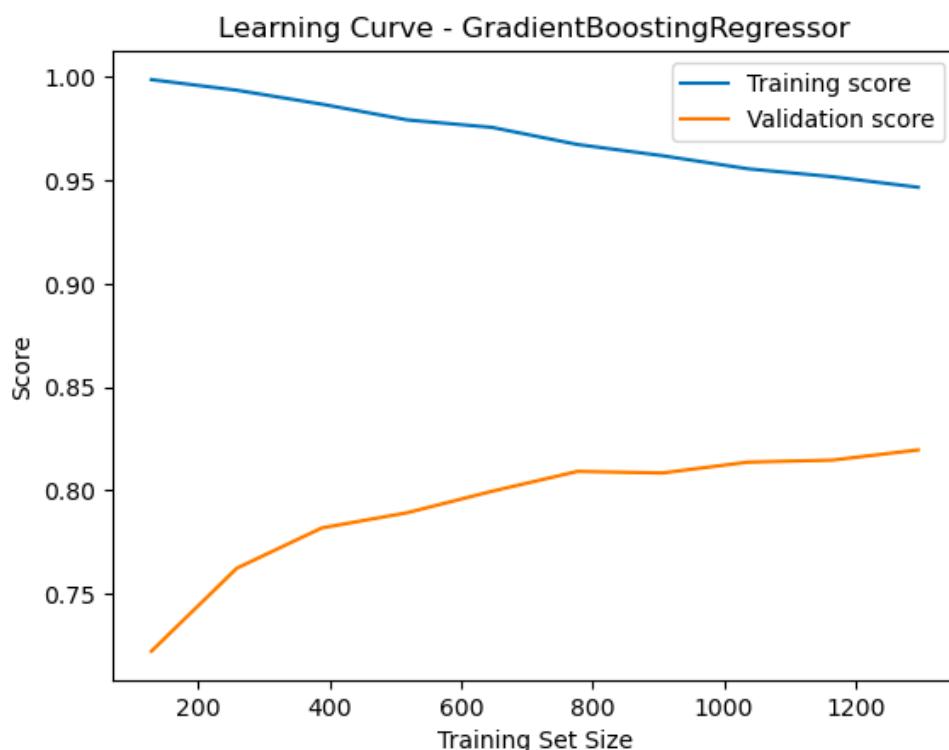
- Le coefficient de détermination R2 de 0.82 indique que le modèle explique environ 85% de la variance totale de la variable cible.
- L'erreur absolue moyenne (MAE) de 0.36 signifie que, en moyenne, les prédictions du modèle diffèrent de 0.329074 unités de la valeur réelle.

L'algorithme GradientBoostingRegressor est une méthode d'apprentissage automatique utilisée pour la régression. Il s'agit d'une approche basée sur l'ensemble, où plusieurs modèles de régression faibles sont combinés pour former un modèle plus fort. Il est également robuste aux valeurs aberrantes grâce à l'utilisation des résidus dans les itérations successives.

AJUSTEMENT DES HYPERPARAMÈTRES

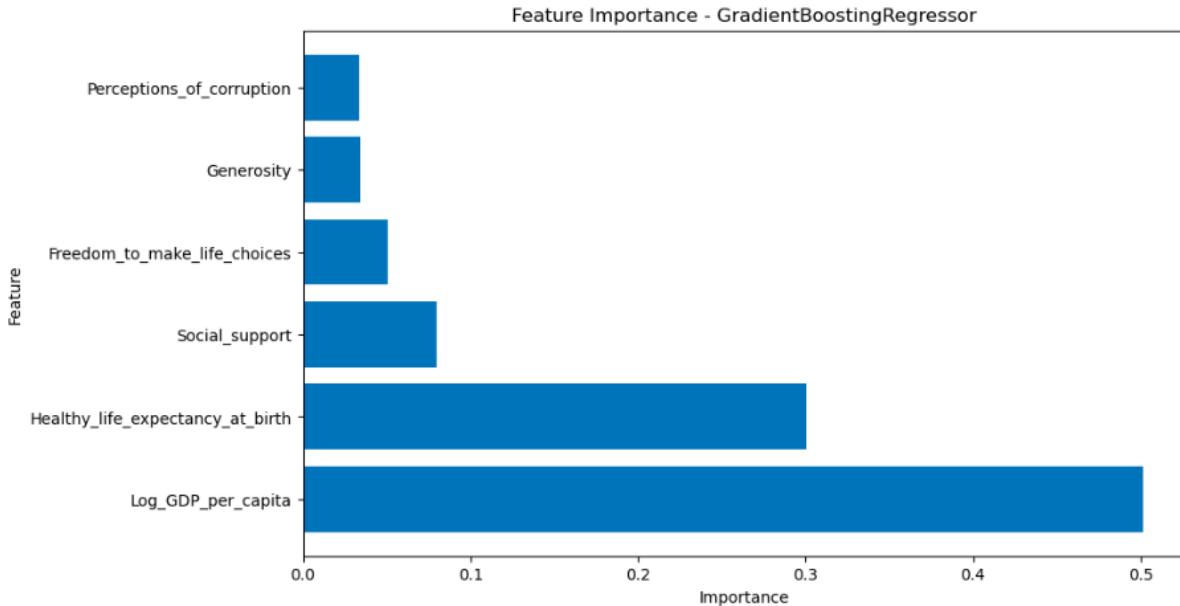
```
MSE:  0.19538430968712117
RMSE:  0.4420229741621143
R2:  0.84487204026263
MAE:  0.32998530280798166
```

L'ajustement des hyperparamètres nous permet d'améliorer les performances de l'algorithme. Nous passons d'un **R2 de 0.82 à 0.84** et d'une **MAE de 0.361853 à 0.33**



FEATURE IMPORTANCE

	Feature	Importance
0	Log_GDP_per_capita	0.501275
2	Healthy_life_expectancy_at_birth	0.300320
1	Social_support	0.079749
3	Freedom_to_make_life_choices	0.050847
4	Generosity	0.034484
5	Perceptions_of_corruption	0.033325



Le nombre optimal de variables est finalement toutes les variables, c'est-à-dire les 6.

SAUVEGARDE DU MODÈLE

```
In [61]: # Chargement du modèle sauvegardé
loaded_model = load('gradientboostregressor.joblib')

# Utilisation du modèle chargé pour faire des prédictions
loaded_model.predict(X_test_scaled)

Out[61]: array([5.32042028, 7.36359444, 4.99551849, 7.05564195, 7.29365756,
   4.45205665, 4.83115252, 4.23892593, 5.70399885, 7.13896873,
   4.32402626, 5.51490257, 3.57968853, 6.01364304, 6.00387382,
   4.29324264, 4.22447145, 7.34690755, 4.81675976, 4.20157018,
   6.34256939, 3.72898028, 5.562297 , 4.71550291, 6.53745326,
   5.83418889, 5.3494631 , 5.01712815, 5.66711048, 6.36339032,
   3.94894697, 4.97798222, 7.01066967, 3.77992879, 4.74797973,
   7.08650093, 4.94019849, 6.21057451, 4.27766002, 4.67125518,
   4.32759422, 5.32639707, 4.05257401, 4.88479818, 5.46303931,
   6.67857802, 3.94251436, 4.74609802, 5.16616777, 4.2670047 ,
   6.41613035, 5.15393868, 4.74598568, 5.88749639, 5.62826143,
   4.17729135, 5.74373503, 4.41756609, 6.30877137, 5.35324176,
   4.35925579, 3.907814 , 6.52524002, 6.49889627, 7.41296162,
   6.51235741, 5.13330556, 5.77837598, 5.1031721 , 4.20520594,
   6.51655178, 4.16135194, 5.04566216, 6.04801381, 6.20970568,
   5.02899151, 6.29621096, 4.23021037, 6.65279497, 6.02380056,
   4.0809669 , 6.25493041, 5.4644389 , 4.2558578 , 5.69485725,
   4.64061183, 6.17795546, 7.28568086, 6.13919558, 7.41879757,
   5.47903019, 5.53767527, 5.86981304, 6.85214195, 5.06870791,
   5.51695276, 4.53760152, 4.3526917 , 5.39587858, 6.9107685 ,
   4.6080293 , 4.08993361, 5.49614201, 7.3514231 , 4.84180232,
   5.46182997, 5.31470909, 5.12967656, 3.60374063, 5.74779036,
   4.75883027, 7.00833083, 7.58782893, 6.16119299, 6.32479817,
   4.4617671 , 4.02161241, 5.30074976, 5.79452711, 2.82705088,
   5.19213926, 4.3837921 , 5.13250199, 5.64758434, 4.3995491 ,
   5.54628798, 4.41153181, 6.68447257, 4.67747879, 7.2553343 ,
   4.25648858, 5.86405507, 4.83610898, 7.60964713, 6.28467256,
   6.46247511, 5.88073591, 4.09154753, 6.21033199, 4.30292775,
   6.47425939, 7.16671112, 7.21689444, 5.34792481, 5.33654048,
   4.58631289, 4.64783939, 7.13173427, 3.98232012, 6.80190834,
   6.12888806, 4.74265473, 4.9233026 , 6.99223028, 5.75427642,
   7.51652756, 5.49590892, 5.58280814, 4.72783468, 5.19141874,
   3.82720629, 5.46534742, 5.62536163, 4.5727515 , 7.46102137,
   5.58230913, 5.52085676, 7.01691553, 4.31633764, 4.56509992,
```

5.58230913, 5.52085676, 7.01691553, 4.31633764, 4.56509992,
5.40669204, 4.72065886, 4.94426361, 5.2344581 , 6.32229034,
4.38939203, 5.65249682, 7.02392936, 5.10136902, 5.0097143 ,
6.39117518, 3.95841674, 7.11621123, 4.95190272, 4.34179142,
5.71540586, 5.34115071, 7.29003548, 5.90027656, 5.57155652,
3.74456552, 5.00358164, 4.36777313, 6.48484993, 4.2276093 ,
5.85408579, 5.15597462, 4.87091673, 3.91588728, 4.00200995,
5.64099342, 6.69222959, 6.45458425, 6.5798965 , 5.15138147,
6.22646342, 6.00328126, 6.07468516, 6.22980711, 5.85715168,
4.16691204, 4.61595601, 6.13224777, 4.85333081, 5.66028808,
6.59459101, 5.55813056, 4.69650452, 4.99724395, 6.46660631,
4.21697697, 5.77841454, 6.2037808 , 6.74141492, 6.08931499,
5.06485644, 4.24974166, 4.62351917, 7.38846473, 5.75180513,
4.6698338 , 6.02151269, 4.33491314, 4.16189956, 7.47470781,
6.55962522, 4.26972506, 4.6366229 , 5.8781033 , 4.57628851,
5.36107583, 6.86797874, 3.66058849, 5.18005398, 5.79510967,
6.33846799, 5.67471958, 5.75435037, 5.61230821, 6.38277801,
4.68640743, 7.11938641, 5.3123632 , 4.22876947, 4.65800995,
4.70579617, 5.29723141, 6.12017246, 5.20268246, 4.61024469,
5.32688944, 6.62239805, 4.52451353, 4.3616835 , 6.4896905 ,
4.52417011, 4.16718417, 4.71650518, 7.24921879, 5.35749437,
5.03832838, 7.15376538, 6.49324935, 5.26295887, 4.35564093,
3.72277336, 5.39712838, 4.09485854, 5.83673062, 4.42792891,
5.85348445, 6.40230293, 4.61784256, 5.64425497, 6.36347671,
5.61737672, 5.24862772, 5.68716413, 5.55375686, 6.57020582,
5.29659229, 4.84609213, 5.51387352, 4.39832917, 7.09356721,
6.92747175, 7.22758063, 5.41102561, 4.16374754, 7.20533577,
4.72072001, 7.02719574, 4.58233002, 4.07726288, 6.35335827,
6.63549844, 6.02273349, 4.96731702, 6.5532303 , 7.388051 ,
5.70389172, 5.98244111, 7.07253813, 4.61636024, 5.17889591,
4.37168598, 5.93817094, 4.788565 , 5.20993315, 5.96191306,
5.33558584, 5.71614132, 3.99773364, 5.01721926, 4.1898603 ,
4.55698645, 5.81523989, 4.32089899, 7.5368252 , 6.34190526,
4.31269452, 3.88645488, 7.28953113, 4.91684718, 6.00941613,
6.36067409, 5.82908125, 4.14319786, 7.24627511, 3.79253064,
4.57210381, 6.20084339, 4.76539019, 5.9699741 , 6.14598309,
7.29528994, 5.58423651, 5.60223625, 5.71174504, 4.48273906,
4.03436241, 5.20060428, 7.27177445, 6.53928886, 5.19318988,
5.07525083, 4.61587571, 4.43939821, 5.31405989, 5.51595058,
4.08168755, 5.03343419, 5.01146726, 7.45740866, 5.39724561,
4.38794582, 6.5459047 , 5.9056606 , 4.73067712, 5.94138904,
4.0297593 , 5.60299244, 6.24529008, 5.81572571, 4.6549787 ,
7.53585798, 4.76953604, 6.62165489, 3.88157166, 5.76996227,

```

6.59459101, 5.55813056, 4.69650452, 4.99724395, 6.46660631,
4.21697697, 5.77841454, 6.2037808 , 6.74141492, 6.08931499,
5.06485644, 4.24974166, 4.62351917, 7.38846473, 5.75180513,
4.6698338 , 6.02151269, 4.33491314, 4.16189956, 7.47470781,
6.55962522, 4.26972506, 4.6366229 , 5.8781033 , 4.57628851,
5.36107583, 6.86797874, 3.66058849, 5.18005398, 5.79510967,
6.33846799, 5.67471958, 5.75435037, 5.61230821, 6.38277801,
4.68640743, 7.11938641, 5.3123632 , 4.22876947, 4.65800995,
4.70579617, 5.29723141, 6.12017246, 5.20268246, 4.61024469,
5.32688944, 6.62239805, 4.52451353, 4.3616835 , 6.4896905 ,
4.52417011, 4.16718417, 4.71650518, 7.24921879, 5.35749437,
5.03832838, 7.15376538, 6.49324935, 5.26295887, 4.35564093,
3.72277336, 5.39712838, 4.09485854, 5.83673062, 4.42792891,
5.85348445, 6.40230293, 4.61784256, 5.64425497, 6.36347671,
5.61737672, 5.24862772, 5.68716413, 5.55375686, 6.57020582,
5.29659229, 4.84609213, 5.51387352, 4.39832917, 7.09356721,
6.92747175, 7.22758063, 5.41102561, 4.16374754, 7.20533577,
4.72072001, 7.02719574, 4.58233002, 4.07726288, 6.35335827,
6.63549844, 6.02273349, 4.96731702, 6.5532303 , 7.388051 ,
5.70389172, 5.98244111, 7.07253813, 4.61636024, 5.17889591,
4.37168598, 5.93817094, 4.788565 , 5.20993315, 5.96191306,
5.33558584, 5.71614132, 3.99773364, 5.01721926, 4.1898603 ,
4.55698645, 5.81523989, 4.32089899, 7.5368252 , 6.34190526,
4.31269452, 3.88645488, 7.28953113, 4.91684718, 6.00941613,
6.36067409, 5.82908125, 4.14319786, 7.24627511, 3.79253064,
4.57210381, 6.20084339, 4.76539019, 5.9699741 , 6.14598309,
7.29528994, 5.58423651, 5.60223625, 5.71174504, 4.48273906,
4.03436241, 5.20060428, 7.27177445, 6.53928886, 5.19318988,
5.07525083, 4.61587571, 4.43939821, 5.31405989, 5.51595058,
4.08168755, 5.03343419, 5.01146726, 7.45740866, 5.39724561,
4.38794582, 6.5459047 , 5.9056606 , 4.73067712, 5.94138904,
4.0297593 , 5.60299244, 6.24529008, 5.81572571, 4.6549787 ,
7.53585798, 4.76953604, 6.62165489, 3.88157166, 5.76996227,
4.00749906, 6.0429575 , 4.81748165, 7.09312896, 7.54729276,
7.45833528, 5.81278741, 7.52900598, 5.04493698, 4.74203529,
6.39137026, 5.72705893, 6.57643333, 7.28209997, 7.00244337,
5.96068872, 4.58018902, 7.1679019 , 7.26726156, 5.69725013,
5.27642156, 6.26760415, 6.47570967, 4.90761954, 5.47292575])

```

CONCLUSION

Lors de notre analyse exploratoire, nous avons observé une tendance : le PIB par habitant semblait être une variable explicative importante dans le score du bonheur. Les courbes de l'évolution de la moyenne du score du bonheur dans le monde et de la moyenne du PIB par habitant dans le monde se suivaient. La matrice de corrélation et l'analyse statistique nous ont confirmé cette tendance.

Cependant, il est important de souligner que le PIB par habitant seul ne suffit pas à expliquer complètement le score du bonheur. Au-delà du facteur économique, d'autres variables ont également montré leur importance dans le calcul du score du bonheur. Deux de ces variables sont l'espérance de vie en bonne santé et le soutien social. D'ailleurs dans la mise en place de notre algorithme de

prédition, notre modèle était le plus précis et le plus robuste lorsqu'il prenait en compte l'ensemble des variables explicatives.

En résumé, notre analyse a mis en évidence l'importance du PIB par habitant dans le score du bonheur (les pays les plus riches sont les plus heureux). Cependant elle souligne également l'importance d'adopter une approche holistique et de considérer plusieurs dimensions pour comprendre le bonheur dans une société.

Difficultés rencontrées lors du projet :

Concaténation de deux jeux de données qui a engendré beaucoup de valeurs manquantes. Cela étant dû au fait que le world-happiness-report.csv ne contenait pas la feature "Regional indicator" et que le world-happiness-report-2021.csv ne contenait pas la feature "year".

Le défi majeur a été d'évaluer la pertinence de notre approche, de notre modèle et des données utilisées. Il était essentiel de garantir que notre modèle était capable de capturer les facteurs pertinents et significatifs pour prédire le score du bonheur. Cela nécessitait une analyse approfondie des variables et des relations potentielles, ainsi qu'une réflexion critique sur notre méthodologie.

Bilan :

En ce qui concerne les objectifs du projet, nous avons réussi à les atteindre. Nos analyses ont permis de déterminer les facteurs qui influencent le score du bonheur et notre modèle a été capable de le prédire.

En fonction du contexte spécifique et des besoins des utilisateurs, le modèle peut être adapté et intégré dans différents processus métiers. Il pourrait être utilisé dans le secteur du marketing pour comprendre les facteurs qui influencent la satisfaction des clients, pour prédire le niveau de satisfaction des clients en fonction de différentes variables (âge, sexe, niveau de revenu, préférences d'achat...)

Pistes d'amélioration pour améliorer la performance du modèle :

- Exploration de nouvelles variables : enrichir le modèle avec par exemple des données sur le niveau d'éducation, la qualité de l'environnement ou d'autres indicateurs socio-économiques pour améliorer la précision des prédictions.
- Exploration de modèles plus avancés : par exemple des modèles tels que les réseaux de neurones ou des modèles basés sur des techniques de deep learning pourraient être explorés pour voir s'ils offrent des améliorations significatives.

OUVERTURE

Une piste d'ouverture intéressante serait d'approfondir l'analyse en se concentrant sur des pays spécifiques. En examinant les facteurs influençant le bonheur dans des contextes nationaux particuliers, nous pourrions découvrir des tendances, des différences culturelles et des spécificités régionales qui éclairent davantage notre compréhension du bonheur.

Il serait également intéressant d'étendre l'analyse à d'autres domaines. Le modèle pourrait être à d'autres domaines, tels que l'éducation, le travail, l'environnement. En adaptant notre modèle et en incorporant des variables spécifiques à chaque domaine, nous pourrions obtenir des informations précieuses sur le bonheur dans ces contextes spécifiques.

Pour apporter des perspectives supplémentaires sur le bonheur, on pourrait envisager l'intégration de données qualitatives telles que des enquêtes de satisfaction, des interviews ou des témoignages. En combinant des données quantitatives et qualitatives, nous pourrions obtenir une compréhension plus riche.

Une piste d'ouverture intéressante pour approfondir ce projet consiste à explorer si l'importance des features change en fonction du niveau de PIB par habitant. Nous pourrions nous demander si, à partir d'un certain niveau de PIB, celui-ci cesse d'être un facteur déterminant du bonheur et quels autres aspects deviennent alors plus importants pour prédire le score du bonheur. Nous pourrions alors identifier des seuils où l'importance relative du PIB change.

Cette exploration nous conduirait à une compréhension plus fine et contextualisée des facteurs qui contribuent au bonheur, en tenant compte des spécificités liées au développement économique.

Annexes

I) Analyse exploratoire des données

https://github.com/DataScientest-Studio/JAN23_DA_Bien-Etre/blob/kervin/notebooks/Analyse%20-%2001%20-%20concat%C3%A9nation.ipynb

II) Visualisation des données

https://github.com/DataScientest-Studio/JAN23_DA_Bien-Etre/blob/kervin/notebooks/Analyse%20-%2001%20-%20macro%20vers%20micro.ipynb

III) Pré-processing des données

https://github.com/DataScientest-Studio/JAN23_DA_Bien-Etre/blob/kervin/notebooks/Analyse%20-%20Notebook%20Rendu%20%201%20-%20avril%202023%20%2B%20enregistrement%20dataset%20pour%20ML.ipynb

IV) Identification de tendances et de modèles significatifs

https://github.com/DataScientest-Studio/JAN23_DA_Bien-Etre/blob/kervin/notebooks/ML%20PROJET%20-%20AVEC%20s%C3%A9paration%20ann%C3%A9e%202021%20-%20RIDGE%20sauvegarde.ipynb

https://github.com/DataScientest-Studio/JAN23_DA_Bien-Etre/blob/kervin/notebooks/ML%20PROJET%20-%20AVEC%20s%C3%A9paration%20ann%C3%A9e%202021.ipynb

https://github.com/DataScientest-Studio/JAN23_DA_Bien-Etre/blob/kervin/notebooks/ML%20PROJET%20-%20SANS%20s%C3%A9paration%20ann%C3%A9e%202021%20-%20SANS%20les%20regions%20-%20GradientBoostingRegressor.ipynb

https://github.com/DataScientest-Studio/JAN23_DA_Bien-Etre/blob/kervin/notebooks/ML%20PROJET%20-%20sans%20s%C3%A9paration%20ann%C3%A9e%202021.ipynb

https://github.com/DataScientest-Studio/JAN23_DA_Bien-Etre/blob/kervin/notebooks/exploratory_data_analysis.ipynb

https://github.com/DataScientest-Studio/JAN23_DA_Bien-Etre/blob/kervin/notebooks/projet_wor_id_happiness_pour_machine_learning.csv

https://github.com/DataScientest-Studio/JAN23_DA_Bien-Etre/blob/kervin/notebooks/ML%20PROJET%20-%20SANS%20s%C3%A9paration%20ann%C3%A9e%202021%20-%20SANS%20les%20regions%20-%20GradientBoostingRegressor.ipynb