MINI-PROJET MACHINE LEARNING

Statistical analysis on factors influencing life expectancy

Data-set Kaggle
https://www.kaggle.com/kumarajarshi/life-expectancy-who

Elisa DESMETZ
M1 IOT – H3 Hitema

Sommaire

Introduction....1

1. Nettoyage des données....1

1.1. Description des variables....2

1.2. Valeurs aberrantes, manquantes et ouliers....2

2. Exploration des données....3

3. Régression linéaire multiple....4

Introduction

Le data-set d'intérêt utilisé pour ce projet est composé de données récoltées par le Global Health Observatory sous la direction de l'OMS. Les données sont récoltées dans l'objectif de suivre le niveau de la santé mondiale ainsi que les facteurs liés pour tous les pays.

Le data-set lié à l'espérance de vie est composé de facteurs santé recensés pour 193 pays obtenus sur le site de l'OMS, les données économiques sont récoltées sur le site des Nations Unies.

Ce data-set a l'avantage de proposer plusieurs angles d'approche. Dans mon cas je vais m'intéresser à deux problématiques :

- Quels sont les effets de la couverture vaccinale sur l'espérance de vie ?
- Comment l'évolution du PIB influence-t-elle la part du PIB allouée à la santé ?

Afin de répondre à ces problématiques je passerais par une étape de data cleaning, avant d'explorer les données, puis j'effectuerais des analyses.

1. Nettoyage des données

Pour pouvoir nettoyer le data-set il faut comprendre les variables avec lesquelles on travaille. Les critères à prendre en compte pour les variables sont :

- Qu'est-ce que la variable représente et de quel type est-elle ?
- Est-ce que la variable a des valeurs manquantes ? Si oui comment les traiter ?
- Est-ce que la variable a des outliers ? Si oui comment les traiter ?

Dans mon cas, le data-set étudié est une agrégation de données réalisée par l'OMS, il regroupe de nombreux indicateurs estimés pour un pays et pour une année.

Dans un premier temps il a fallu effectuer un passage sur les noms des variables, ceux-ci n'étant pas homogènes. J'ai choisi pour cela d'uniformiser les noms en passant tout en minuscule et en transformant les espaces en underscores.

```
cols_old = list(data_dirty.columns)
cols_new = []
for col in cols_old:
    cols_new.append(col.strip().replace('  ', ' ').replace(' ', '_').lower())
data_dirty.columns = cols_new
```

De plus, l'une des variables avait un nom qui ne correspondait pas à ce qu'elle contenait, j'ai donc corrigé cela également.

```
data_clean=data_dirty.rename(columns = {'thinness_1-19_years':'thinness_10-19_years'})
```

1.1. Description des variables

Le data-set est composé de 2938 observations sur 22 variables de suivi.

Les les variables d'intérêt pour l'étude sont :

country : pays d'origine des variables de suivi, variable qualitative nominale
year : année de mesure des variables de suivi, variable quantitative de type date
percentage_expenditure: dépense du pays sur la santé en pourcentage du PIB, variable quantitative de ratio
hepatitis_b: couverture vaccinale à l'hépatite B chez les moins de 1 an, variable quantitative de ratio
measles: nombre de cas reportés de rougeole pour 1000, variable quantitative de ratio
polio: couverture vaccinale à la polio chez les moins de 1 an, variable quantitative de ratio
total_expenditure : dépense du gouvernement sur la santé en pourcentage des dépenses totales du gouvernement, variable quantitative de ratio
gpd: Produit Intérieur Brut per capita, variable quantitative de ratio

REPONSE
life_expectancy: espérance de vie des individus en années pour un individu donné dans un pays donné à une année donnée, variable quantitative

Les variables country et years ont des variables de regroupement, elles ne seront pas utilisées dans les analyses

Les variables prédictives peuvent être divisées en plusieurs catégories : facteurs liés à l'immunité, facteurs liés à la mortalité, facteurs économiques, facteurs sociaux.

1.2. Valeurs aberrantes, manquantes et ouliers

	adult_mortality	infant_deaths	alcohol	percentage_expenditure	hepatitis_b	measles
count	2928.000000	2938.000000	2744.000000	2938.000000	2385.000000	2938.000000
mean	164.796448	30.303948	4.602861	738.251295	80.940461	2419.592240
std	124.292079	117.926501	4.052413	1987.914858	25.070016	11467.272489
min	1.000000	0.000000	0.010000	0.000000	1.000000	0.000000
25%	74.000000	0.000000	0.877500	4.685343	77.000000	0.000000
50%	144.000000	3.000000	3.755000	64.912906	92.000000	17.000000
75%	228.000000	22.000000	7.702500	441.534144	97.000000	360.250000
max	723.000000	1800.000000	17.870000	19479.911610	99.000000	212183.000000

Avec un simple tableau descriptif des variables on peut remarques des incohérences :

- Un taux de mortalité de 1 est très probablement une erreur.
- Un taux de mortalité infantile de 0 est très probablement également une erreur.
- Un taux de mortalité chez les moins de 5 ans de 0 est très probablement également une erreur.
- Des IMC de 1 et de 87.3 sont impossibles car absolument pas réalistes en fonction de ce que la variable représente.
- Une population de 34 individus est improbable.

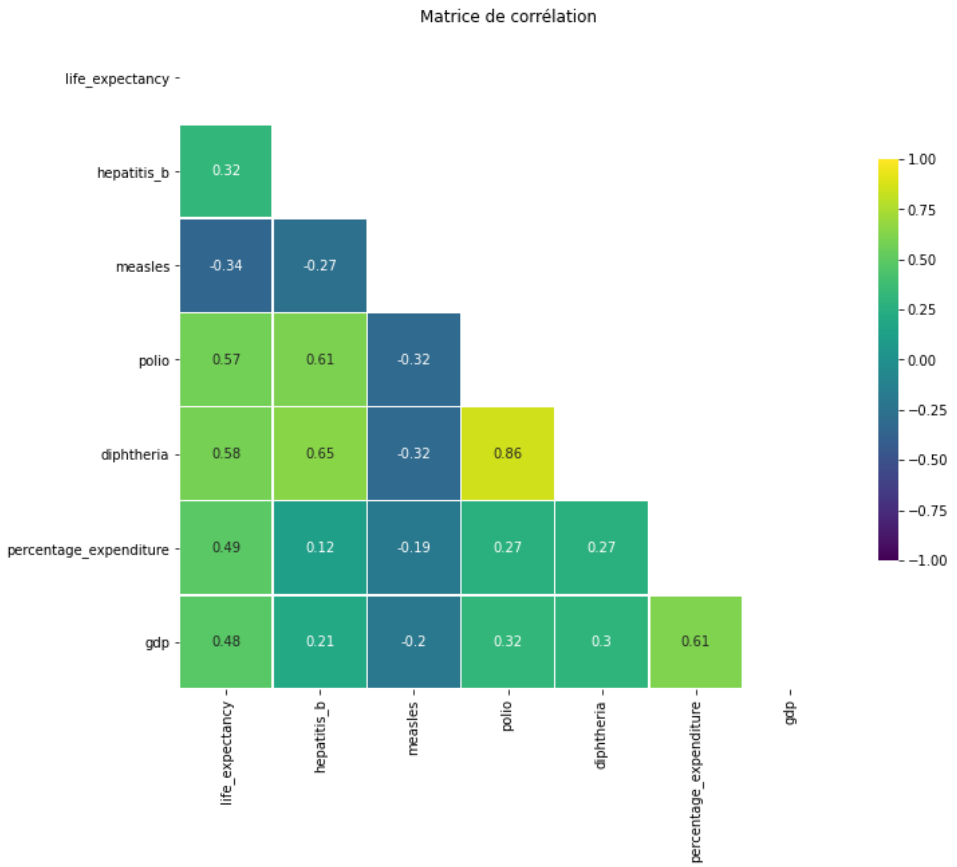
Pour certaines erreurs, on va considérer que ce sont des valeurs aberrantes et donc on va les transformer en NaN pour qu'elles ne polluent pas l'information.

- La mortalité infantile ne doit pas être nulle
- La mortalité des individus inférieurs à 5 ans ne doit pas être nulle
- L'IMC ne doit pas être inférieur à 10 et supérieur à 50

Afin de corriger la présence de valeurs manquantes, j'impute la valeur de l'observation manquante par la moyenne de celle de l'année correspondante.

Le data-set comporte un grand nombre d'outliers, pour régler ce souci, et au vu de la disparité du nombre d'outliers, j'effectue une winsorisation.

2. Exploration des données



Une fois les données winsorisées, on peut s'intéresser à la matrice de corrélation.

L'espérance de vie est corrélée avec :

- Le PIB (faible positif)
- Le pourcentage du PIB alloué à la santé (faible positif)
- La couverture vaccinale à la diphtérie et à la polio (faible positif)

La couverture vaccinale à l'hépatite B est corrélée avec :

- La couverture vaccinale à la diphtérie et à la polio (faible positif)

La couverture vaccinale à la polio est corrélée avec :

- La couverture vaccinale à la diphtérie (fort positif)

Le pourcentage du PIB alloué à la santé est corrélé avec :

- Le PIB (fort positif)

On constate que les couvertures vaccinales sont souvent corrélées entre elles et qu'elles sont corrélées avec l'espérance de vie. On constate également que le pourcentage du PIB alloué est corrélé positivement avec le PIB, plus celui-ci est élevé, plus la part allouée à la santé est élevée.

3. Régression linéaire multiple

Afin de conclure sur l'effet de la couverture vaccinale sur l'espérance de vie, j'effectue une régression linéaire multiple prenant comme variables prédictives les couvertures vaccinales de la polio, de la diphtérie et de l'hépatite B.

```
Intercept:
 51.4333644749866
Coefficients:
 [ 0.13133464  0.10859986 -0.02458323]
Predicted Life Expectancy:
 [69.14133521]
```

OLS Regression Results

Dep. Variable:	life_expectancy	R-squared:	0.266
Model:	OLS	Adj. R-squared:	0.265
Method:	Least Squares	F-statistic:	353.5
Date:	Mon, 16 Nov 2020	Prob (F-statistic):	5.60e-196
Time:	23:38:43	Log-Likelihood:	-10332.
No. Observations:	2938	AIC:	2.067e+04
Df Residuals:	2934	BIC:	2.070e+04
Df Model:	3		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	51.4334	0.658	78.131	0.000	50.143	52.724
diphtheria	0.1313	0.009	14.313	0.000	0.113	0.149
polio	0.1086	0.009	12.371	0.000	0.091	0.126
hepatitis_b	-0.0246	0.008	-3.149	0.002	-0.040	-0.009

Omnibus:	111.569	Durbin-Watson:	1.874
Prob(Omnibus):	0.000	Jarque-Bera (JB):	137.049
Skew:	-0.425	Prob(JB):	1.74e-30
Kurtosis:	3.631	Cond. No.	637.

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.

Avec un R^2 de 0.27, on peut conclure que le modèle n'est pas précis.