Projet Python :

Prédiction du taux de mortalité pour le cancer des poumons, de la tranchée et de la bronche à partir d'indicateurs socio-économiques.

ALBUTIU Ana

DIARRA Assitan

ZIADI Flora

Mastère Spécialisé Data Science

Année 2017-2018

Plan

0. Introduction

L'étude des cancers suscite un intérêt considérable dans la communauté scientifique qui s'inscrit de plus en plus dans une démarche collaborative. En particulier, le site Epidemium met à disposition des données exploitables pour l'étude des cancers.

Pour ce projet, nous avons utilisé deux sources d'informations distinctes : mortality , qui contient les données sur la mortalité dans le monde et worldbank, qui contient des données socio-économiques pour chaque pays. Ainsi à partir de ces deux sources données, nous souhaitons prédire le taux de mortalité par pays en fonction des indicateurs socio-économiques.

Ce projet se décomposera en deux parties : tout d'abord, une première partie qui portera sur l'exploration et le traitement des données, puis une seconde partie qui portera sur la modélisation.

1. Exploration et Analyse de données

1.1. Exploration de la base

Cette première partie nous a permis de visualiser le contenu de la base de données « mortality » pour savoir quelles informations nous pouvons exploiter. On a l'information nombre de morts par code du cancer, tranche d'âge, année, sexe, et pays. Cette exploration nous a permis :

- de savoir qu'il y a plus d'hommes tués par le cancer que de femmes
- d'avoir une intuition sur le choix du cancer à prédire. Le plus fréquent est le cancer des poumons, de la tranchée et de la bronche.
- de connaitre la tranche d'âge la plus touchée par le cancer le plus fréquent : 65-69 ans

A l'issu de cette exploration, nous avons donc décidé de nous restreindre à la prédiction du taux de mortalité d'un cancer pour tous les pays présents dans les bases de données.

Notebook associé : 1.1 exploration des données.ipynb

1.2. Traitement de la base

Dans cette partie, le but est d'avoir la base de données finale qui nous permettra de prédire le taux de mortalité grâce aux indicateurs socio-économiques, il s'agit donc de fusionner les deux bases de données « Worldbank » et « Mortality ». Pour ce faire, il a fallu :

Sur la base Mortality :

- modifier le découpage par tranches d'âge pour harmoniser avec la base « WorldBank »
- modifier les noms des pays pour harmoniser avec la base « WorldBank »
- calculer le taux de mortalité par sexe, par tranche d'âge, par année et par pays
- de se restreindre aux années 2000 à 2015 car on observe une rupture entre 1990 et 2000 qu'on ne sait pas expliquer.

Sur la base WorldBank :

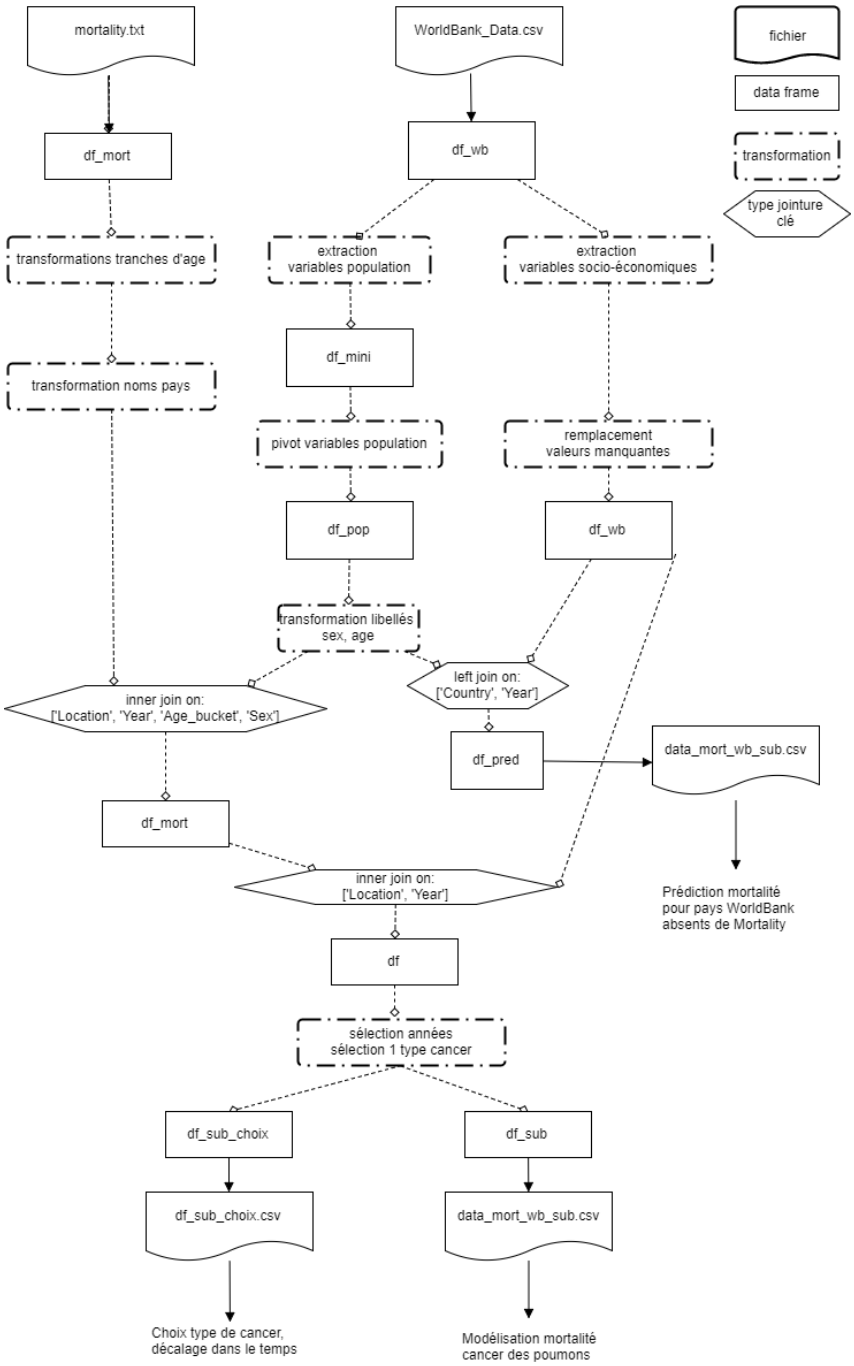
- modifier les codes sexe et tranches d'âges pour harmoniser avec la base « Mortality »
- remplacer des valeurs manquantes
- choisir les indicateurs socio-économiques

Enfin, suite à cette fusion, nous obtenons deux bases de données pour le cancer des poumons, de la tranchée et de la bronche :

- la première correspond à celle avec les 99 pays présents dans les deux bases, elle nous servira pour créer nos modèles.
- la deuxième correspond à celle avec les 199 pays présents dans Wordbank qui nous servira pour prédire le taux de mortalité de nouveaux pays et de l'afficher sur la carte du monde.

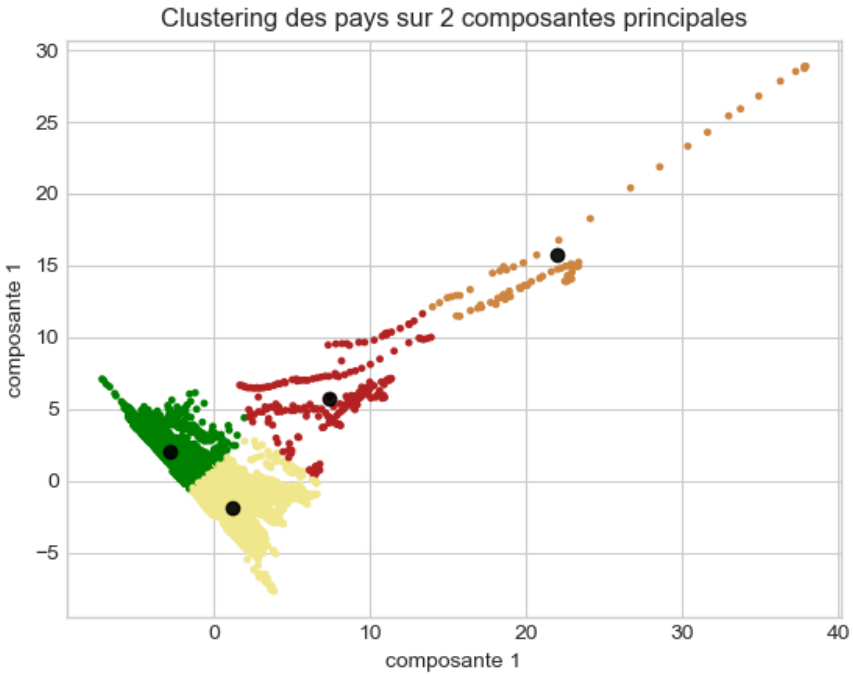
Notebook associé: 1.1 traitement de données.ipynb

Diagramme workflow Traitement de données



1.3. Clustering des pays en fonction des indicateurs sociaux-économiques

Pour se faire une idée sur les informations apportées par les données socio-économiques, on souhaite voir si ces variables pourraient être utilisées pour classifier des pays dans des groupes.



On commence par une analyse en composantes principales, on résume l'espace des variables explicatives, à deux composantes qui ensemble expliquent la moitié de la variance. En projetant les pays sur cet espace, on voit une agglomération des pays qui se ressemblent beaucoup et un nombre plus petit de pays qui de distinguent plus.

Au fil des années, certains pays peuvent voir leur situation économique, démographique ou écologique changer, ce qui se traduit par une modification des groupes - c'est l'exemple de l'Algérie qui passe du groupe 1 au groupe 0.

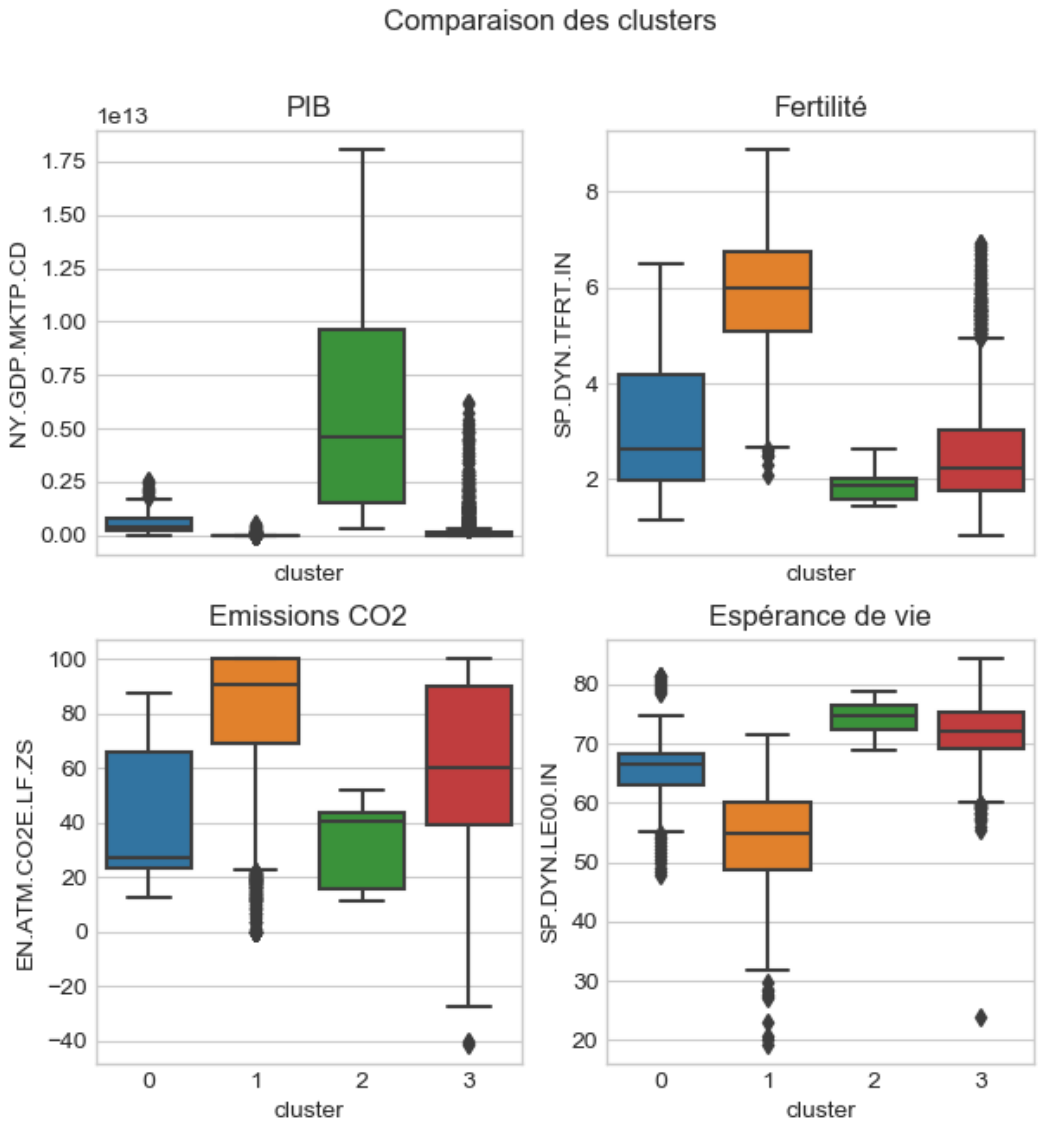
En revanche, un pays plus stable comme la France ne connait pas une évolution similaire - elle reste dans le même groupe pour toutes les années enregistrées.

Pour l'année la plus récente (2015), la composition des clusters est :

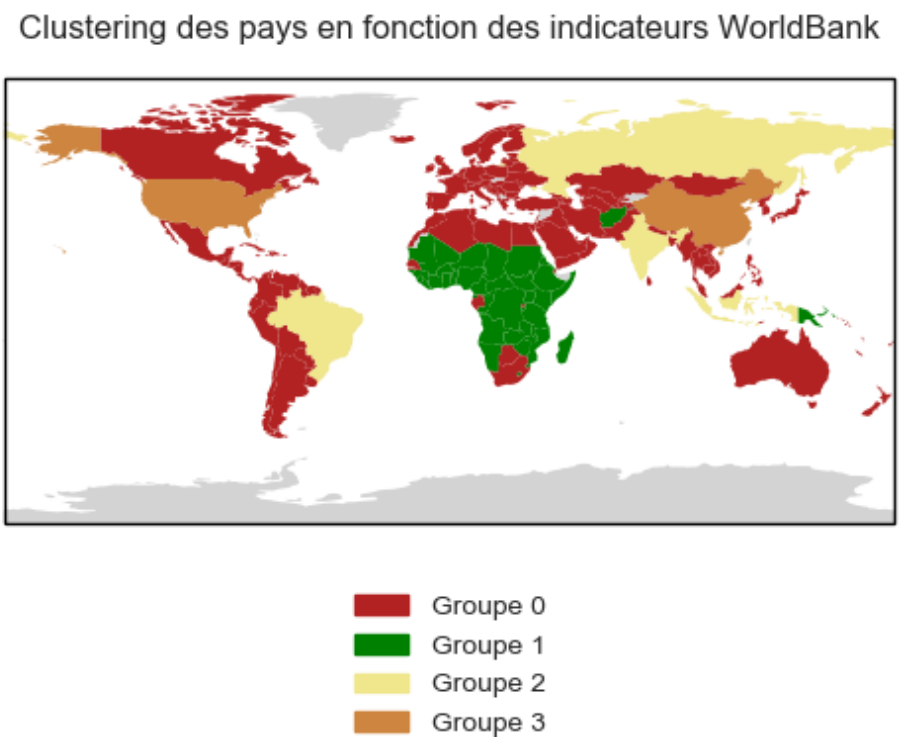
- Groupe 0 : pays de l'Europe, le Mexique, Canada, Amérique du Sud, l'Australie
- Groupe 1 : la Russie, le Brésil, l'Inde, des iles de l'Asie de Sud-Est
- Groupe 2 : Etats-Unis, Chine
- Groupe 3 : pays de l'Afrique

Une analyse de quelques variables plus discriminantes nous permet de caractériser ces groupes :

- Groupe 0 : PIB élevé, fertilité relativement faible, plus petites émissions de CO2, espérance de vie moyenne avec importants points aberrants
- Groupe 1 : PIB faible, fertilité très élevée, fortes émissions de CO2, espérance de vie courte
- Groupe 2 : PIB le plus élevé, fertilité faible, émissions de CO2 relativement réduites et une relativement longue espérance de vie
- Groupe 3 : PIB faible avec variation importante, fertilité moyenne, espérance de vie moyenne



Sur la carte du monde, la représentation des clusters est :

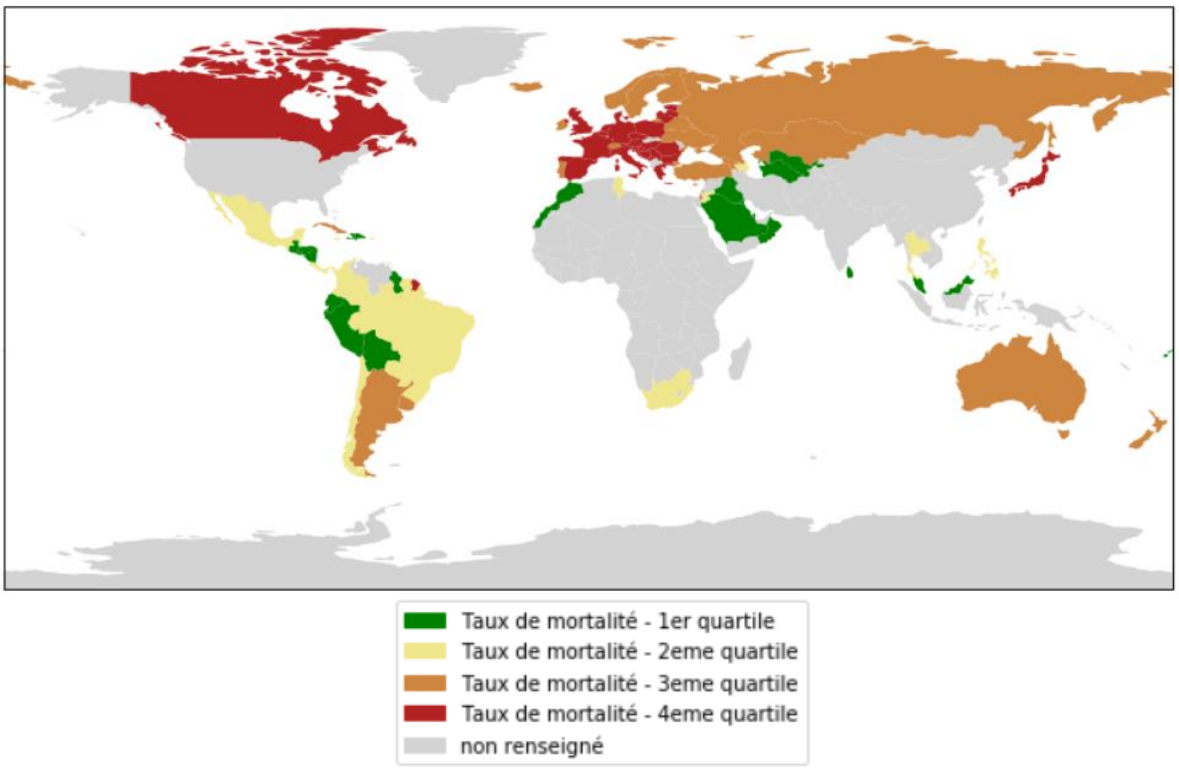


Notebook associé : 1.2 Clustering des pays.ipynb

1.4. Représentation du taux de mortalité par pays

Dans cette partie, nous représentons le taux de mortalité pour le cancer des poumons, de la tranchée et de la bronche par pays de manière descriptive. Nous calculons le taux de mortalité pour chaque pays présent dans la base de modélisation (soit 99 pays), que nous découpons ensuite en 4 tranches égales, enfin nous représentons ces tranches sur une carte du monde.

Taux de mortalité observé par pays



L'objectif sera par la suite de compléter le taux de mortalité pour les pays non renseignés à l'aide de notre modélisation.

Notebook associé : 1.3 Représentation du taux de mortalité par pays.ipynb

2. Présentation des modèles utilisés et comparaison des résultats

2.1. Choix de cancer, décalage de l'impact au niveau temps

Notre but est de voir si, avec la base de données dont on dispose, il existe un **type de cancer** en particulier qui serait mieux expliqué par les indicateurs socio-économiques. On s'attend à ce que certains types de cancers soient héréditaires (ex: cancer des seins), alors que d'autres pourraient être impactés plus par des facteurs liées à l'environnement ou au style de vie (ex: cancer des poumons, mélanome).

Pour cela, on privilégie un modèle simple afin d'itérer rapidement sur les types de cancer et les années de décalage. La régression lasso nous permet en plus d'assigner un poids de 0 aux variables sans pouvoir explicatif, en nous révélant les variables plus pertinentes.

Cancer_code	Nb_features	Score
C14	31	0.3938
C15-C26	33	0.3800
C30-C39	29	0.3737
C71	31	0.3678
C33,C34	45	0.3355
C56	36	0.3297
C64,C65,C68	20	0.3294
C40, C41, C50	26	0.3235
C25	36	0.3169
C40, C41	15	0.3106
C22	35	0.3029
…	…	…
C47	5	0.0117
C58	0	0.0000

Types de cancer mieux prédits par le modèle Lasso

Nous observons que le type de cancer le mieux expliqué par les variables choisies pendant le période observée serait C14 (pharynx), avec 33 variables explicatives significatives et un score de 0.39.

Les cancers les moins influencés par nos variables sont Cancer des nerves périphériques (C47), Cancer du placenta (C58).

Le **cancer des poumons** (C33,C34) a un score très proche et 45 des variables sont significatives dans le modèle correspondant. On décide donc de modéliser ce dernier par la suite.

Dans un deuxième temps, on pourrait croire que les facteurs d'environnement et de style de vie auraient un **impact décalé dans le temps** sur la mortalité, donc on essayera d'expliquer le taux par les facteurs du passé.

A chaque itération, on construit une base de données avec les variables explicatives pour l'année 2000 et une nouvelle variable expliquée décalée de 'delta' nombre d'années(Mortality_shifted).

Base_year	Delta	Nb_features	Nb_obs	R2
2000	14	17	256	0.38454
2000	13	21	608	0.37621
2000	12	33	1008	0.36342
2000	11	43	1504	0.35971
2000	0	41	2704	0.35729
2000	5	38	2048	0.35192
2000	10	41	1792	0.35134
2000	9	39	1888	0.34678
2000	2	42	2432	0.34611
2000	6	43	2048	0.3446
2000	4	41	2176	0.34329
2000	3	44	2336	0.3422
2000	1	42	2512	0.34055
2000	7	43	2016	0.33892
2000	8	43	1888	0.33529

Décalage optimal entre les variables explicatives et la variable expliquée

Le meilleur résultat est obtenu pour un delta = 14, c'est à dire les indicateurs socio-économiques de l'année 2000 expliquent le mieux le taux de mortalité de l'année 2014 pour le cancer des poumons.

Cependant, le nombre d'observations pour ce décalage est très réduit, puisque de moins en moins d'observations sont disponibles pour les années les plus récentes. En plus, seules 17 variables explicatives sont retenues.

Les résultats des modèles avec prise en compte de décalage dans le temps sont assez proches, donc nous décidons de ne pas appliquer un décalage pour les autres modèles qu'on essayera d'optimiser.

Notebook associé : 2.1 Choix cancer, delta années - lasso.ipynb

2.2. Méthodes utilisées pour la prédiction

Pour chaque modèle, nous avons testé de nombreux paramètres, afin d'obtenir les meilleurs résultats.

La majorité de nos variables sont continues et nous les avons standardisées pour optimiser les performances. Les variables catégorielles sont sexe, pays et tranche d'âge. Nous avons utilisé deux approches pour modéliser le taux de mortalité : la classification et la régression. Pour la classification, nous avons découpé le taux de mortalité en 18 classes.

Les résultats de nos modèles sur la **base de test** sont les suivants :

Classification :

Méthode par une Classification	**RMSE**	**R^2**
k plus proches voisins (knn)	2,89	0,43
Arbres de décision	1,9434	0,5895
Forêts aléatoires	1,7148	0,6647
Gradient Boosting (gbm)	1,7907	0,6076

Régression:

Méthode	**RMSE**	**R^2**	**Moyenne**	**Médiane**	**Ecart-type**	**Min**	**Max**
Taux de mortalité observé sur la base de test	x	x	52,13	1,31	135,06	0,0	1308,18
k plus proches voisins (knn)	84,11	0,61	42,61	7,29	135,11	0,0	561,25
Arbres de décision	49,84	0,86	51 ,28	1	1,18	0,21	1002,51
Machine à vecteurs de support (svm)	72.51	0,71	37.64	18.36	57.41	-18.92	597.83
Forêts aléatoires	29,87	0,95	51,87	2,14	130,75	0,0	1294,13

Résultats des modèles

Pour le modèle SVM : le minimum est négatif et le maximum est très éloigné de la vraie valeur. Pour le modèle KNN : le même problème de maximum est observé.

Nous obtenons de meilleurs résultats pour l'arbre de décision et les forets aléatoires en termes de R^2 et de RMSE. De plus, leurs moyennes sont très proches de la moyenne observée. Cependant, le maximum, l'écart type et la médiane pour la forêt aléatoire sont plus proches des valeurs observées. C'est pourquoi, nous retenons la forêt aléatoire comme modèle finale.

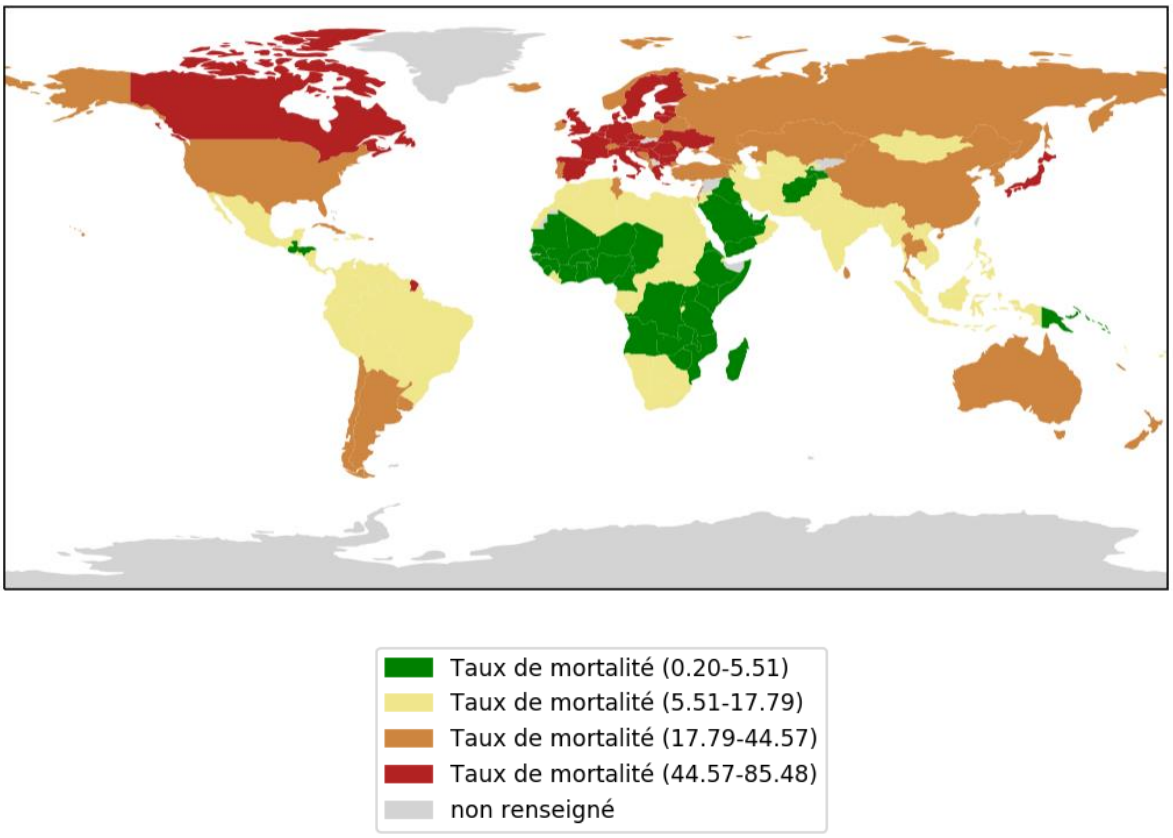
Notebooks associés:

2 .2.a Modèles Knn et Decision Tree.ipynb,
2.2.b Modèle SVM.ipynb et
2.2.c Modèles Ramdom et GBM.ipynb

3. Conclusion

Dans la partie précédente, nous avons sélectionné le modèle pour lequel nous obtenons les meilleurs résultats. L'objectif est de prédire le taux de mortalité pour tous les pays pour lesquels cette information est manquante dans la base de données « mortality ». Pour ce faire, nous appliquons le modèle sur la deuxième base de données (193 pays) obtenue dans la partie traitement de données. Nous représentons ensuite le taux de mortalité prédit sur la carte du monde de la même manière qu'à la partie 1.4. Nous avons gardé également les mêmes bornes pour le découpage par classe que la carte précédente afin de pouvoir comparer les résultats prédits aux résultats observés.

Taux de mortalité prédit par pays



Nos résultats nous semblent cohérents, puisque pour la plupart des pays présents sur les deux cartes, nous obtenons les mêmes couleurs. Ainsi, les taux de mortalité prédits sont assez proches des taux de mortalité observés pour ces pays. Pour exemple des USA dont le taux de mortalité n'est pas renseigné dans nos données, le taux prédit est 42,46 et la valeur indiquée par lung.org est 43,2.

En général, nous observons sur cette carte que le taux de mortalité pour le cancer des poumons est très élevé pour les pays d'Amérique du Nord, l'Europe, l'Asie du Nord et l'Australie. Des taux plus bas sont observés par l'Afrique du Nord et l'Amérique du Sud alors que l'Afrique de l'Ouest et Central ont les taux les plus bas.

4. Extension

Notre modèle ne fournis pas des résultats précis, néanmoins il nous permet d'avoir une idée des ordres de grandeurs des taux de mortalité par pays. Ainsi, notre modèle est exploitable puisqu'il pourrait permettre à une organisation d'obtenir les taux de mortalité pour les pays dont l'information est manquante. Il pourrait être associé à des facteurs socio-économiques complémentaires et mieux renseignés qui affineraient les prédictions.

De plus, il est nécessaire que notre modèle soit rafraichi chaque année en intégrant les nouvelles données. Son rafraichissement nécessite l'intervention humaine, mais il pourrait être automatisé. La mémoire nécessaire supplémentaire pourrait engendrer un cout non négligeable.

L'utilisation de notre modèle par un organisme pour concentrer ses efforts de prévention dans un pays plutôt qu'un autre, signifie qu'en utilisant un taux de mortalité erroné, l'allocation des ressources limitées de cette organisation humanitaire serait efficace.

Un avantage de l'approche Epidemium est la facilité d'obtention des données open source, donc dans la communauté data science est encouragée à s'associer facilement aux spécialistes du domaine médical pour mener ensemble des études plus spécifiques. Par exemple, une meilleure approche pour étudier les causes du cancer serait de suivre une cohorte de patients avec leurs habitudes de vie et leurs informations médicales individuelles, mesurées dans une étude clinique.

Annexes : liste des notebooks

Chaque notebook est associé à une partie de ce rapport. La liste des notebooks est la suivante :

1.1. Exploration des données.ipynb

1.1. Traitement de la base de données.ipynb

1.2. Clustering des pays.ipynb

1.3. Représentation du taux de mortalité par pays.ipynb

2.1. Choix cancer, delta années - lasso.ipynb

2.2.a. Modèles Knn et Decision Tree.ipynb

2.2.b. Modèle SVM.ipynb

2.2.c. Modèles Ramdom et GBM.ipynb

3. Prédiction de la carte finale.ipynb