

|  |
| --- |
| **Projet Python :**  **Prédiction du taux de mortalité pour le cancer des poumons à partir d’indicateurs socio-économiques.** |

**ALBUTIU Ana**

**DIARRA Assitan**

**ZIADI Flora**

**Mastère Spécialisé Data Science**

**Année 2017-2018**

**Plan**

[0. Introduction : 3](#_Toc503086202)

[1. Exploration et Analyse de données : 3](#_Toc503086203)

[1.1. Exploration de la base 3](#_Toc503086204)

[1.2. Retraitement de la base 3](#_Toc503086205)

[1.3. Analyse des pays en fonction des données socio économiques; représenter sur une carte les groupes de pays (clustering) 3](#_Toc503086206)

[2. Présentation des modèles utilisés et comparaison des résultats 7](#_Toc503086207)

[2.1. Boucle pour le choix du cancer et le décalage de l'impact au niveau temps (lasso) 7](#_Toc503086208)

[2.2. Méthodes utilisées pour la prédiction 7](#_Toc503086209)

[3. Conclusion 8](#_Toc503086210)

# Introduction :

L’objet de ce projet est de prédire le taux de mortalité pour le cancer des poumons. Pour cela, nous avons utilisé les données du site [Epidemium](http://www.epidemium.cc/data.html). Nous avons utilisé deux bases de données distinctes : [mortality](http://qa.epidemium.cc/data/mortality_dataset/) , qui contient les données sur la mortalité dans le monde et [world bank](http://qa.epidemium.cc/data/epidemiology_dataset/world_bank_data/), 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 l’analyse des données, puis une seconde partie qui portera sur la modélisation.

# Exploration et Analyse de données :

## 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 en tirer. Les colonnes de la base de données étant : code du cancer, tranche d’âge, année, sexe, pays et nombre de mort, cette exploration nous a permis :

* de savoir qu’il y a plus d’hommes atteints que de femmes
* de se restreindre aux années les plus renseignées c’est-à-dire entre 2000 et 20015
* d’avoir une intuition sur le choix du cancer à prédire c’est à dire le plus fréquent : 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 en particulier dans tous les pays présents dans les bases de données et en se basant sur les données de 2000 à 2015.

## Retraitement 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 sociodémographiques, il s’agit donc de fusionner les deux bases de données Worldbank et Mortality. Pour se faire, il a fallu :

**Sur la base Mortality :**

* modifier sur le découpage par tranches d’âge pour harmoniser avec la base WorldBank
* modifier sur 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

**Sur la base WorldBank :**

* modifier sur les codes sexe et tranches d'ages pour harmoniser avec la base Mortality
* remplacer des valeurs manquantes
* choisir les indicateurs socio-économiques

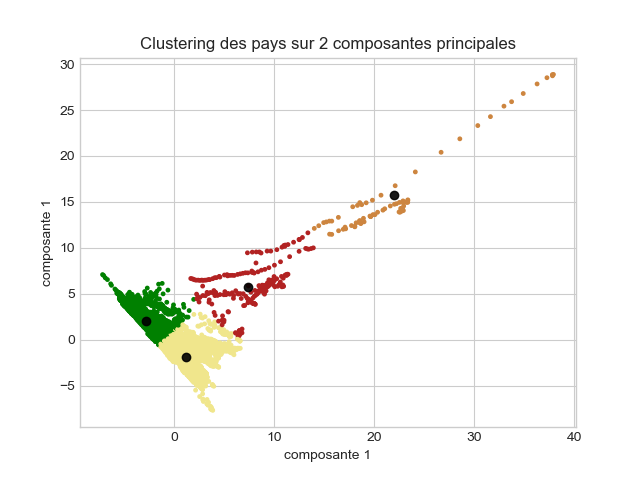
Enfin, nous avons fait la fusion des bases de données :

* 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.

**!!! Ajouter le graphe d’Ana + quelques stat descriptives de la base finale (nombre de ligne, description du taux de mortalité (moyenne, ..)**

## Analyse des pays en fonction des données socio économiques; représenter sur une carte les groupes de pays (clustering)

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 features 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 un changement de group auquel ils appartiennent - c'est l'exemple de l'Algérie.

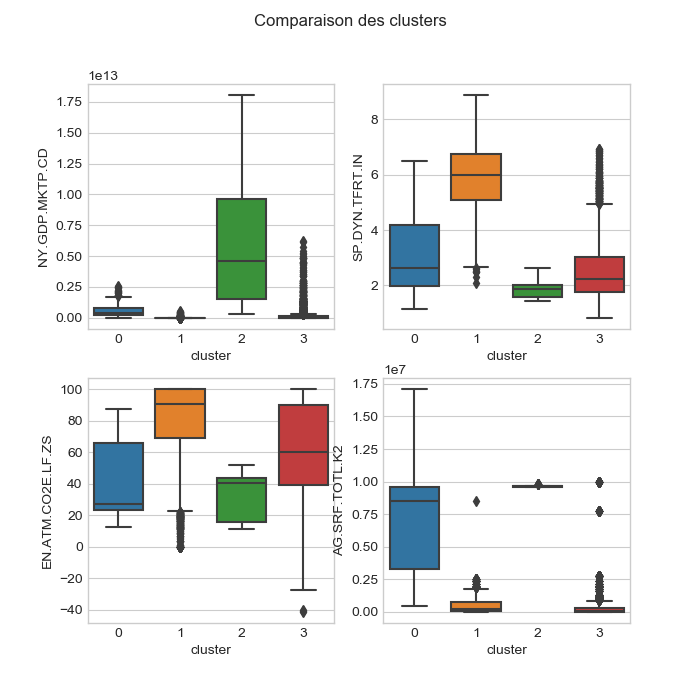
En revanche, un pays plus stable comme la France ne connait pas une évolution similaire - elle reste dans le meme groupe pour toutes les années enregistrées.

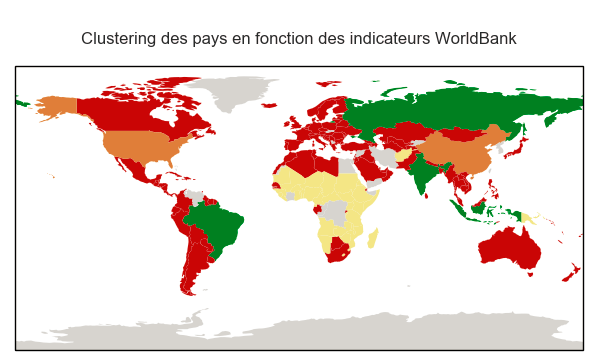
Pour l’année la plus récente, la composition des clusters est :

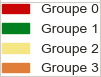
* Groupe 0 : pays de l’Europe, le Mexique, Canada, Amérique du Sud, l’Australie
* Groupe 1 : la Russie, le Brésil, l’Inde, des iles de l’Asie de Sud-Est
* Groupe 2 : Etats-Unis, Chine
* Groupe 3 : pays de l’Afrique

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

* Groupe 0 : PIB élevé, fertilité relativement faible, plus petites émissions de CO2, grande surface, mais variabilité dans le cluster
* Groupe 1 : PIB faible, fertilité très élevée, fortes émissions de CO2, petite surface
* Groupe 2 : PIB le plus élevé, fertilité faible, émissions de CO2 relativement réduites et une grande surface.
* Groupe 3 : PIB faible avec variation importante, fertilité moyenne, émissions de CO2 relativement élevées et petite surface

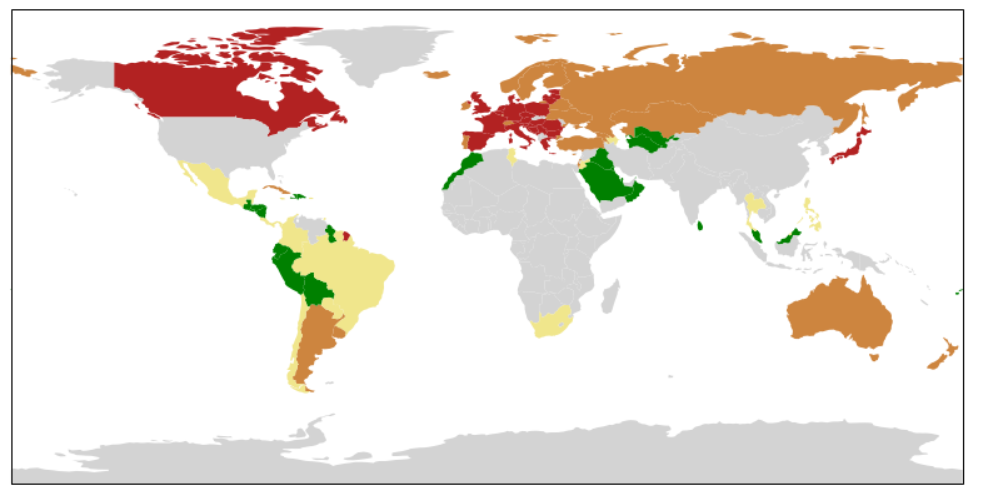


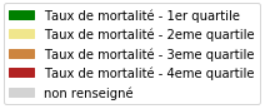
Sur la carte du monde, la représentation des clusters est : 



* 1. **Représentation du taux de mortalité par pays**

Dans cette partie, nous représentons le taux de mortalité par pays de manière descriptive. Nous calculons le taux de mortalité pour chaque pays présent dans la base data\_2000 (soit 99 pays), que nous découpons ensuite en 4 tranches égales, enfin nous représentons ces tranches sur une carte du monde.





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

# Présentation des modèles utilisés et comparaison des résultats

## 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 |

Il paraît que le type de cancer le mieux expliqué par les variables choisies pendant le période observée serait C14 (pharynx), avec 33 features 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.

Par la suite, 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 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 |

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 features 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.

## Méthodes utilisées pour la prédiction

Pour chaque modèle, nous avons testé de nombreux paramètres, afin d’obtenir les meilleurs résultats. Les résultats de nos modèles sur la **base de test** sont les suivants :

**Classification :**

|  |  |  |
| --- | --- | --- |
| **Méthode par une Classification** | **RMSE** | **R2** |
| k plus proches voisins (knn) | 2,7576 | 0.4141 |
| Arbres de décision | 1,9434 | 0,5895 |
| Forêts aléatoires | 1,7148 | 0,6647 |
| Gradient Boosting (gbm) | 1,7907 | 0,6076 |

**Régression:**

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| **Méthode** | **RMSE** | **R2** | **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) | 62,9766 | 0,7652 |  |  |  |  |  |
| Arbres de décision | 39,0767 | 0,9096 |  |  |  |  |  |
| Machine à vecteurs de support (svm) | 60,06 | 0,80 | 40,91 | 2,76 | 85,53 | -19,63 | 663,36 |
| Forêts aléatoires  Méthode manuelle | 29.5590 | 0,9513 | 51,47 | 3,01 | 129,69 | 0,0 | 1281,61 |

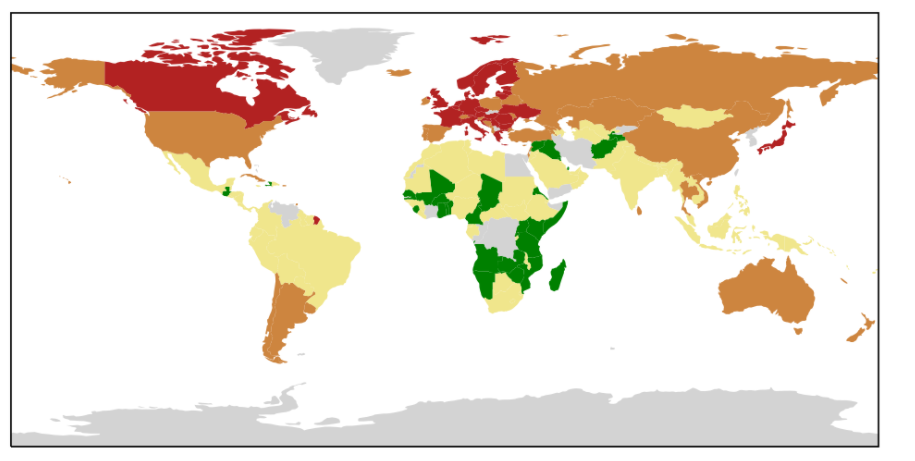
Nous obtenons de meilleurs résultats avec une forêt aléatoire pour une régression. Nous retenons ainsi ce modèle pour la suite.

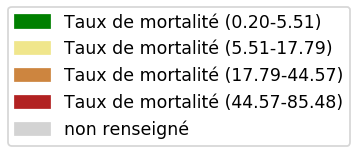
# Extension

Notre modèle pourrait être exploitable car il donne une approximation du taux de mortalité dans les pays pour lesquels cette information est manquante, dans le cadre notamment d’une étude menée par une ONG afin de créer un modèle qui prédirait le taux de mortalité dans le temps. En cas de mauvaises prédictions, le résultat dans le temps serait dans ce cas biaisée et donc inexploitable. Il pourrait être associé à d’autres facteurs socio-économiques qui affineraient les prédictions car l’erreur du modèle actuel est non négligeable. De plus, sa durée de vie est infinie à condition qu’il soit rafraichi chaque année grâce aux nouvelles données. Son rafraichissement nécessite l’intervention humaine mais il pourrait être automatisé et cela nécessiterait des ressources en mémoire considérables car le volume des données sera amené à grandir.

# 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 nous ne disposons pas du taux de mortalité et ainsi compléter la carte du monde présente à la partie 1.4. Nous allons ainsi fitter le modèle retenu sur une nouvelle base de données « data\_pred\_wb.csv », qui est issue de la base « world\_bank ». Nous représentons ensuite le taux de mortalité prédit sur la carte du monde de la manière qu’à la partie 1.4. Nous gardons également les mêmes tanches que la carte précédente afin de pouvoir comparer équitablement les résultats prédit aux résultats observés.





Nous sommes assez satisfaits de nos résultats, 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. De plus, nous avons comparé les taux de mortalité prédits au (demander à la source à ana) pour quelques pays qui ne sont pas présent dans la table mortalité. Nos résultats sont en accord, c’est notamment le cas pour les Etats-Unis (demander les chiffres à Ana).

Nous observons 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. Nous pouvons nous demander si c’est réellement le cas où cela vient plutôt du fait que la cause de mort par le cancer des poumons n’est pas toujours détectée dans des pays plus pauvres.

1. **Annexes : liste des notebooks**

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

* 1. Exploration.ipynb
  2. Traitement base de données.ipynb
  3. Classification pays.ipynb
  4. Taux de mortalité par pays.ipynb

2.1. Choix cancer, delta années - lasso.ipynb

2.2.a. Knn et decision tree.ipynb

2.2.b. Modèle SVM - régression et choix paramètres.ipynb

2.2.c. Forêts aléatoires et GBM.ipynb

3. Prédiction de la carte finale.ipynb

**Demain, Harmonisation les noms des notebooks avec les filles.**

**Par exemple,**

* 1. Exploration des données.ipynb
  2. Traitement de la base de données.ipynb
  3. Classification des pays.ipynb
  4. 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