

|  |
| --- |
| **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**

1. **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.

1. **Exploration et Analyse de données :**
   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 en tirer. Les colonnes de la base de données sont : code du cancer, tranche d’âge, l’année, sexe, pays et taux de mortalité. Cette exploration nous a permis de voir que la répartition des sexes est égale, le cancer le plus fréquent et la tranche d’ache la plus touchée. Nous avons donc décidé de prédire le taux de mortalité pour un unique cancer qui est le cancer des poumons, de la tranchée, de la bronche.

* 1. **Retraitement de la base**

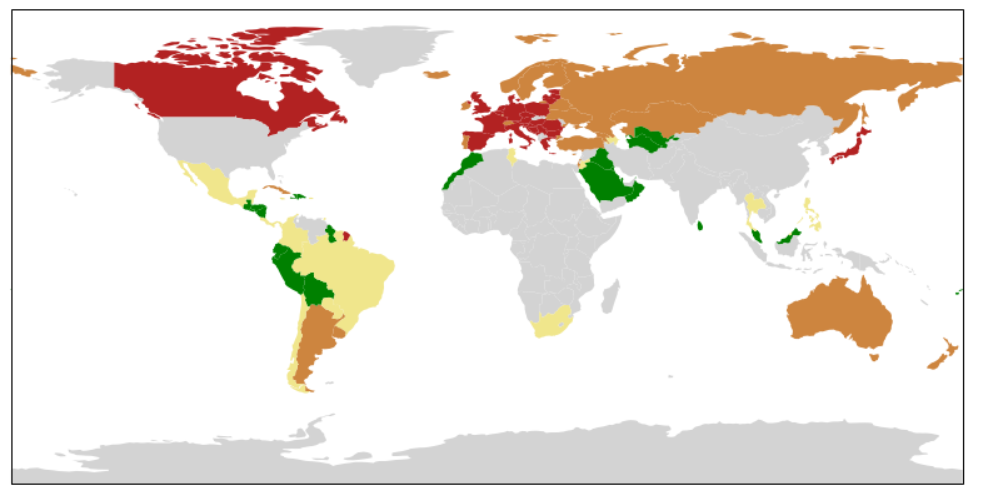
Dans cette partie, le but est d’avoir la base de données finale qui nous permettra de prédire le taux de mortalité grâce aux indicateurs sociodémographiques, il s’agit donc de fusionner les deux bases de données worldbank et mortality.

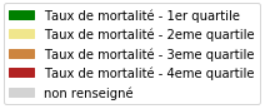
Il s’agissait tout d’abord de faire correspondre les tranches d’âge et les noms de pays car l’une des bases contenait plus de tranche d’âge que l’autre et certains pays présents dans les deux bases n’avaient pas le même nom. Puis nous avons calculé le taux de mortalité par sexe, par tranche d’âge, par année et par pays grâce aux indicateurs tels que la population totale de femmes et d’hommes. Puis, nous avons fait la fusion des deux bases de données par année et pays. Enfin nous avons nettoyé la base de données car elle contenait des valeurs non renseignées.

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

* 1. **Analyse des pays en fonction des données socio économiques; représenter sur une carte les groupes de pays (clustering)**
  2. **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.

1. **Présentation des modèles utilisés et comparaison des résultats**
   1. **Boucle pour le choix du cancer et le décalage de l'impact au niveau temps (lasso)**
   2. **Méthodes utilisées pour la prédiction**

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 |

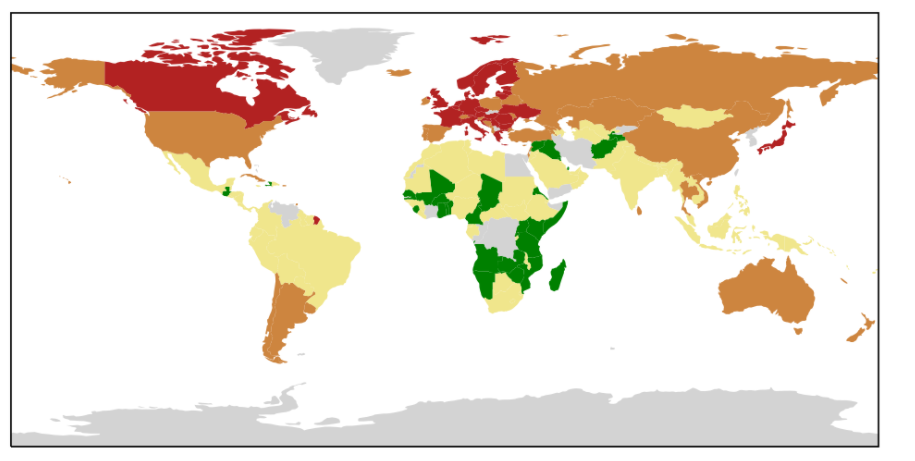
**Régression:**

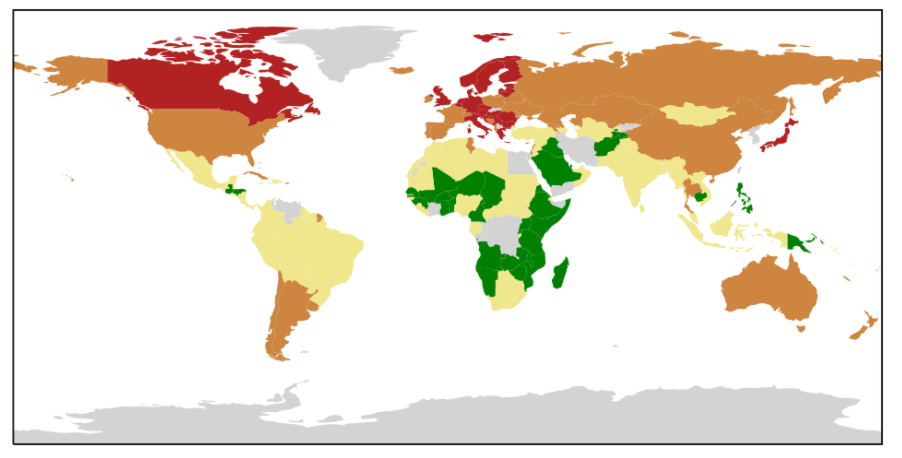
|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| **Méthode** | **RMSE** | **R2** | **Moyenne** | **Médiane** | **Ecart-type** | **Min** | **Max** |
| Taux de mortalité observé sur la base de test | x | x | 52,13 | 1,31 | 135,06 | 0,0 | 1308,18 |
| k plus proches voisins (knn) | 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 |
| Forêts aléatoires  automatique | 29,8759 | 0,9503 | 51,87 | 2,14 | 130,75 | 0,0 | 1294,13 |

Nous obtenons de meilleurs résultats avec une forêt aléatoire pour une régession.

1. **Conclusion**

**Carte obtenue par la méthode manuelle**

 **Carte obtenue par la recherche automatique**



**Les filles quels résultats penser vous que je dois mettre ???**

**Les stat sont un peu mieux avec la méthode automatique, mais la carte pour l’europe est un peu moins bonne.**