

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

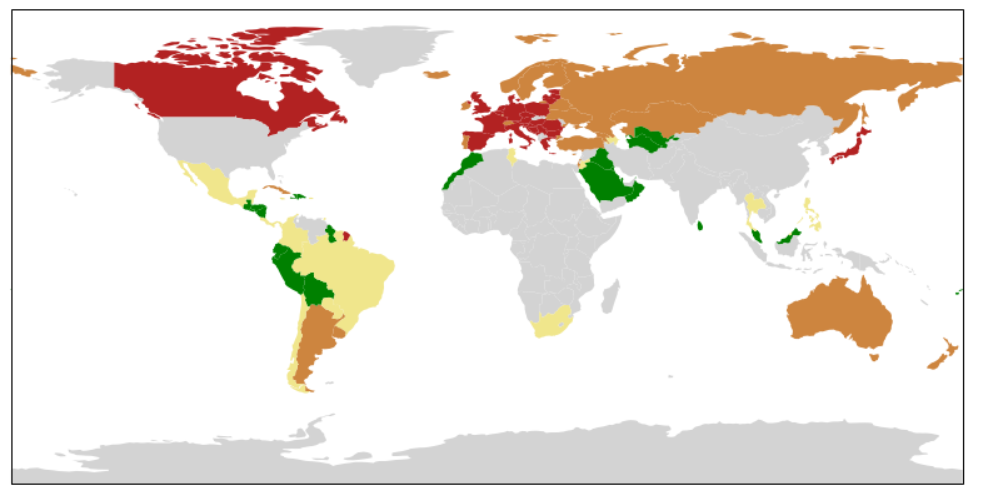
Dans cette partie, le but est d’avoir la base de données finale qui nous permettra de prédire le taux de mortalité grâce aux indicateurs sociodémographiques, il s’agit donc de fusionner les deux bases de données worldbank et mortality.

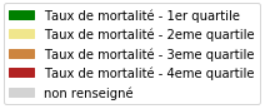
Il s’agissait tout d’abord de faire correspondre les tranches d’âge et les noms de pays car l’une des bases contenait plus de tranche d’âge que l’autre et certains pays présents dans les deux bases n’avaient pas le même nom. Puis nous avons calculé le taux de mortalité par sexe, par tranche d’âge, par 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**

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 |

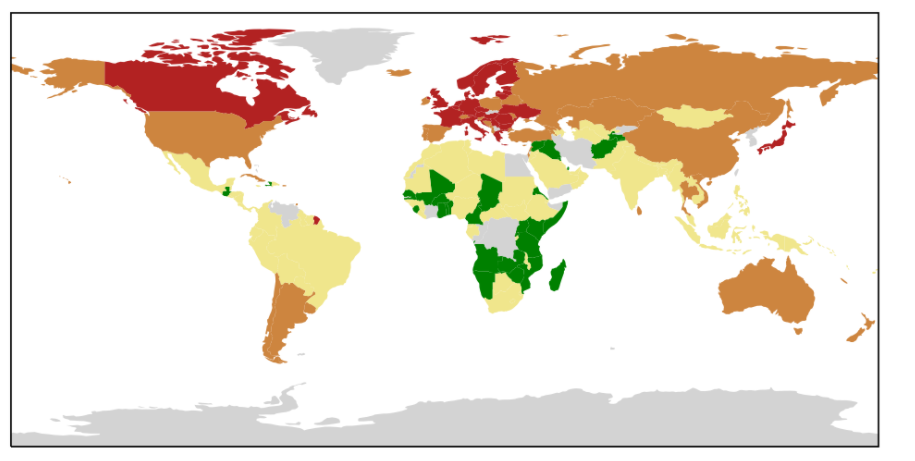
**Régression:**

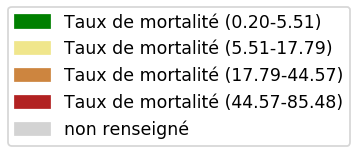
|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| **Méthode** | **RMSE** | **R2** | **Moyenne** | **Médiane** | **Ecart-type** | **Min** | **Max** |
| Taux de mortalité observé sur la base de test | x | x | 52,13 | 1,31 | 135,06 | 0,0 | 1308,18 |
| k plus proches voisins (knn) | 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.

1. **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