

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

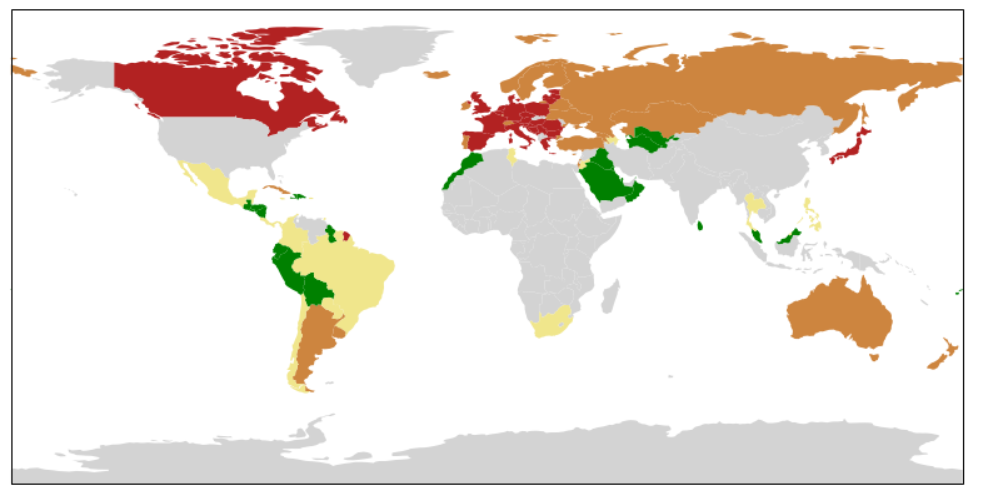
**Année 2017-2018**

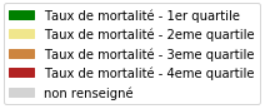
1. **Introduction :**

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
   2. Retraitement de la base
   3. Analyse des pays en fonction des données socio économiques; représenter sur une carte les groupes de pays (clustering)
   4. 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
      1. k plus proches voisins (knn)
      2. Arbres de décision
      3. Machine à vecteurs de support (svm)
      4. Forêts aléatoires
      5. Gradient Boosting (gbm)

Les résultats de nos modèles sont les suivants :

|  |  |  |
| --- | --- | --- |
| **Méthode** | **RMSE sur la base de test** | **R2 sur la base de test** |
| k plus proches voisins (knn) |  |  |
| Arbres de décision |  |  |
| Machine à vecteurs de support (svm) |  |  |
| Forêts aléatoires Régression | 0,0282 | 0,9901 |
| Forêts aléatoires Classification | 1,7782 | 0,9801 |
| Gradient Boosting (gbm) | 1,7907 | 0,6076 |

**Voir avec les filles si un tableau leur convient ou si elles veulent qu’on détaille plus..**

1. **Conclusion**

Flora : à venir

1. **Annexes : listes des notebooks**

Flora : à venir