Une image contenant texte

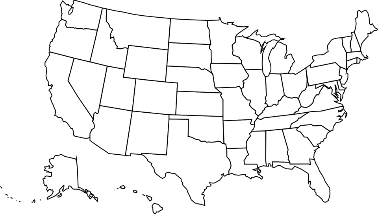
Description générée automatiquement

Masse de données et fouille de données

-

**Etude du Covid-19 aux Etats-Unis**

-

CHEVRIER Jean-Christophe

LUC Tristan

NOIROT Quentin

Sommaire

[1. Contexte 4](#_Toc98700133)

[2. Analyse exploratoire des données 4](#_Toc98700134)

[2.1. Volumétrie des données 4](#_Toc98700135)

[2.2. Structure des données 4](#_Toc98700136)

[2.3. Analyse exploratoire introductive 4](#_Toc98700137)

[2.5. Analyse exploratoire de la colonne « date » 4](#_Toc98700138)

[2.6. Analyse exploratoire de la colonne « county » 4](#_Toc98700139)

[2.7. Analyse exploratoire de la colonne « state » 4](#_Toc98700140)

[2.8. Analyse exploratoire de la colonne « fips » 4](#_Toc98700141)

[2.9. Analyse exploratoire de la colonne « cases » 4](#_Toc98700142)

[2.10. Analyse exploratoire de la colonne « deaths » 4](#_Toc98700143)

[2.11. Vérification de l’unicité du couple (date, comté) 4](#_Toc98700144)

[3. Nettoyage des données 4](#_Toc98700145)

[3.1. Suppression des lignes avec des valeurs manquantes 4](#_Toc98700146)

[3.2. Conversion au même format pour les dates 4](#_Toc98700147)

[3.3. Suppression des "-" dans les colonnes quantitatives 4](#_Toc98700148)

[3.4. Transformation des colonnes quantitatives en entiers 4](#_Toc98700149)

[4. Fouille des données 4](#_Toc98700150)

[4.1. Complétion des données 4](#_Toc98700151)

[4.2. Sélection des données 4](#_Toc98700152)

[4.3. Première normalisation des données 4](#_Toc98700153)

[4.4. Mélange des données 4](#_Toc98700154)

[4.5. Conception d’un modèle de prédiction du nombre de morts par régression 4](#_Toc98700155)

[4.5.1. Division en deux ensembles : données des prédictions, et du modèle 4](#_Toc98700156)

[4.5.2. Division en deux ensembles : variables explicatives, et variable résultat 4](#_Toc98700157)

[4.5.3. Division en deux ensembles : données des entrainements, et des tests 4](#_Toc98700158)

[4.5.4. Seconde normalisation des données 4](#_Toc98700159)

[4.5.5. Première configuration de modèle 4](#_Toc98700160)

[4.5.6. Deuxième configuration de modèle : plus de neurones 5](#_Toc98700161)

[4.5.7. Troisième configuration de modèle : plus de paquets 5](#_Toc98700162)

[4.5.8. Quatrième configuration de modèle : taux d’apprentissage de 5% 5](#_Toc98700163)

[4.5.9. Cinquième configuration de modèle : taux d’apprentissage de 0.1% 5](#_Toc98700164)

[4.5.10. Sixième configuration de modèle : taux d’apprentissage de 0.01% 5](#_Toc98700165)

[4.6. Conception d’un modèle de prédiction du nombre de morts par classification 5](#_Toc98700166)

[4.6.1. Classification des observations par classes de nombre de morts 5](#_Toc98700167)

[4.6.2. Création de l’architecture du modèle 5](#_Toc98700168)

[4.6.3. Test manuel de la justesse de prédiction du modèle 5](#_Toc98700169)

[4.6.4. Etude du modèle à partir de graphique 5](#_Toc98700170)

[5. Conclusion 5](#_Toc98700171)

# 1. Contexte

Dans le cadre de notre formation de Master 2 MIAGE SID, et pour le module de masse de données et de fouille de données, nous avions reçu la mission d’implémenter une solution informatique devant répondre à un certain cahier des charges.

# 2. Analyse exploratoire des données

## 2.1. Volumétrie des données

## 2.2. Structure des données

## 2.3. Analyse exploratoire introductive

## 2.5. Analyse exploratoire de la colonne « date »

## 2.6. Analyse exploratoire de la colonne « county »

## 2.7. Analyse exploratoire de la colonne « state »

## 2.8. Analyse exploratoire de la colonne « fips »

## 2.9. Analyse exploratoire de la colonne « cases »

## 2.10. Analyse exploratoire de la colonne « deaths »

## 2.11. Vérification de l’unicité du couple (date, comté)

# 3. Nettoyage des données

## 3.1. Suppression des lignes avec des valeurs manquantes

## 3.2. Conversion au même format pour les dates

## 3.3. Suppression des "-" dans les colonnes quantitatives

## 3.4. Transformation des colonnes quantitatives en entiers

## 4. Fouille des données

## 4.1. Complétion des données

## 4.2. Sélection des données

## 4.3. Première normalisation des données

## 4.4. Mélange des données

## 4.5. Conception d’un modèle de prédiction du nombre de morts par régression

### 4.5.1. Division en deux ensembles : données des prédictions, et du modèle

### 4.5.2. Division en deux ensembles : variables explicatives, et variable résultat

### 4.5.3. Division en deux ensembles : données des entrainements, et des tests

### 4.5.4. Seconde normalisation des données

### 4.5.5. Première configuration de modèle

#### 4.5.5.1. Création de l’architecture du modèle 4.5.5.2. Test manuel de la justesse de prédiction du modèle

#### 4.5.5.3. Etude du modèle à partir de graphique

### 4.5.6. Deuxième configuration de modèle : plus de neurones

### 4.5.7. Troisième configuration de modèle : plus de paquets

### 4.5.8. Quatrième configuration de modèle : taux d’apprentissage de 5%

### 4.5.9. Cinquième configuration de modèle : taux d’apprentissage de 0.1%

### 4.5.10. Sixième configuration de modèle : taux d’apprentissage de 0.01%

## 4.6. Conception d’un modèle de prédiction du nombre de morts par classification

### 4.6.1. Classification des observations par classes de nombre de morts

### 4.6.2. Création de l’architecture du modèle

### 4.6.3. Test manuel de la justesse de prédiction du modèle

### 4.6.4. Etude du modèle à partir de graphique

# 5. Conclusion