MLDS - Rapport TP1

Jeu de données et problématique

Binome :
AZOUAOUI Youcef
BELKESSA Linda

1 Choix du jeu de données

Nom	Tornados
URL	Lien vers le repo Github
Propriétaire	NOAA's National Weather Service Storm Prediction Center
Ordre de grandeur	**Nombre de variables = 27 et Nombre d'observations = 68693**

2 Problématique initiale

Le jeu de données contient un ensemble d'observations indexées par le temps pour des tornades, il indique les caractéristiques géographiques de ces dernières (longitude, latitude, ... etc) ainsi que leurs conséquences en termes de blessés, morts, biens détruits ... etc.

L'aspect temporel nous met dans le cadre d'analyse et prédiction de séries temporelles ce qui nous mène à notre problématique principale qui est de **prédire le comportement future de ces séries** dans l'objectif de :

- **Prédire le degré de fatalité future des tornades en calculant cette dernière à base des variables ['injuries', 'fatalities', 'losses'] (problème supervisé)**
- **Trouver le(s) meilleur(s) attributs qui permettent de donner la meilleure variance inter-clusters de tornades.**

3 La tâche envisagée en supervisé

Nous cherchons dans un premier temps **une formule pour combiner 3 colonnes** qui décrivent les dégâts des tornades de manière à donner plus d'importance aux dommages humains dans le cadre d'études de séries temporelles, **cette nouvelle variable est celle à prédire en utilisant des modèles d'apprentissage supervisé (régression, ARMA, forets aléatoires, XGBoost, ... etc).**

4 La tâche envisagée en non supervisé

Dans cette deuxième partie, nous cherchons à regrouper les tornades selon les caractéristiques qui les distinguent au mieux les unes des autres et les interpréter **pour pouvoir classifier les nouvelles observations en utilisant des algorithmes non supervisés tels que : K-means, DBScan, Classification Ascendante Hiérarchique ... etc.**