MLDS - Rapport TP2

Etude descriptive et exploratoire

Binome :
BELKESSA Linda
AZOUAOUI Youcef

1 Introduction

Notre Dataset représente un ensemble d'observations indexées par le temps pour des tornades, il indique les caractéristiques géographiques de ces dernières ainsi que leurs conséquences en termes de blessés, morts et biens détruits. La figure 1 montre la distribution des tornades observées dans les États-Unis depuis 1950.

Figure 1 : Distribution des tornades

Dans ce rapport, nous allons réaliser l'étude descriptive, statistique et aléatoire de nos données.

2 Étude descriptive et statistique

2.1 Distribution et description des données

- Nous avons 27 variables présentes dans notre jeu de données dont 1 booléenne, 21 numériques et cinq de type objet qui représente les dates et le fuseau horaire. Nous avons une seule variable catégorique, qui est "st", qui représente l'État fédéral où s'est produit la tornade.

2.2 Valeurs abberrantes

Nous avons trouvé une instance redondante, ainsi que 756 (1,1 %) valeurs aberrantes dans la colonne magnitude, et 27170 (39,6 %) valeurs aberrantes dans Loss.

Pour remédier à ce problème, nous avons comparé les valeurs statistiques (moyenne, std, min, max) des colonnes "mag" et "loss" avant et aprés l'application des stratégies Backward fill, forward fill, mean, mod et drop NaN. Nous avons opté pour l'utilisation du **drop NaN**, car nous avons constaté que les descripteurs statistiques ne changent pas significativement avec ou sans son utilisation.

2.3 Corrélation des colonnes

Nous avons généré la matrice de corrélation ci-dessous. Nous avons constaté les faits suivants :

- Nous avons trouvé plusieurs fortes corrélations et redondance entre nos attributs, ce qui nous a permis à éliminer certains attributs afin de réduire la taille du Dataset.
- Nous avons aussi observé quelques attributs était déséquilibrés comme "inj" et "fat", où nous avons trouvé que la majorité de leurs valeurs sont à 0, ce qui peut impacter les performances de nos modèles.

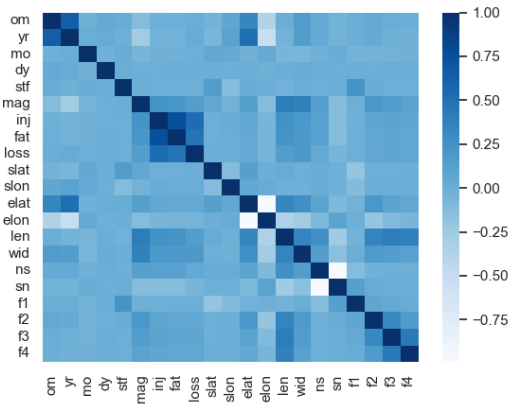

Figure 2 : Matrice de corrélation

Les variables retenues sont : **om, datetime_utc, st, mag, inj, loss, wid, len, ns**.

2.4 Analyse en composantes principales

Nous réalisons une analyse en composantes principales afin de réduire la dimensionnalité des données, facilitant ainsi l'identification des variables qui expliquent la plus grande variance des données.

Nous constatons que la variable qui exprime la plus grande variance des données est "wid".

2.5 Analyse de série temporelle

Nous avons réalisé une série temporelle afin de visualiser l'évolution mensuelle et annuelle des attributs : magnitude des tornades, nombre de blessés et estimations, pertes, comme montré ci-dessous :

- Vu que la série de l'évolution de la magnitude exhibe une dépendance temporelle. Nous avons la possibilité de diriger notre approche supervisée du problème vers la prédiction de la magnitude future, en adoptant un cadre théorique axé sur l'analyse et la prédiction de séries temporelles.
- Nous pouvons opter pour une classification supervisée basée sur les caractéristiques d'une tornade à un instant donné afin de prédire l'intensité des dommages.

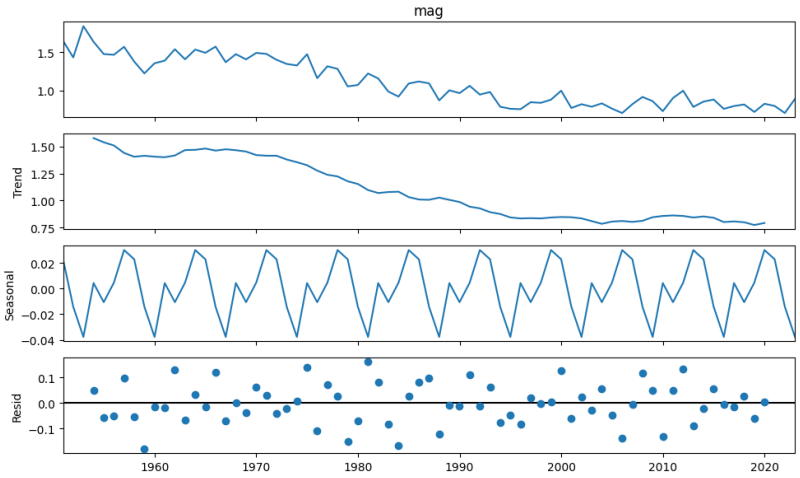

Figure 3 : Décomposition de la série "magnitude"