**Explication de la base**

Il s’agit d’une tâche de régression difficile, dont l’objectif est de prédire la superficie brûlée par les incendies de forêt, dans la région nord-est du Portugal, en utilisant des données météorologiques et d’autres données.

Les incendies de forêt sont un problème environnemental majeur, créant des dommages économiques et écologiques tout en mettant en danger des vies humaines. La détection rapide est un élément clé pour contrôler un tel phénomène. Pour y parvenir, une alternative consiste à utiliser des outils automatiques basés sur des capteurs locaux, tels que ceux fournis par les stations météorologiques. En effet, on sait que les conditions météorologiques (par exemple, la température, le vent) influencent les incendies de forêt et plusieurs indices d’incendie, tels que l’indice météorologique des incendies de forêt (IFM), utilisent de telles données. Dans ce travail, nous explorons une approche d’exploration de données (DM) pour prédire la superficie brûlée des feux de forêt. Cinq techniques de DM différentes, par exemple les machines à vecteurs de support (SVM) et les forêts aléatoires, ainsi que quatre configurations distinctes de sélection de caractéristiques (utilisant des composants spatiaux, temporels, FWI et des attributs météorologiques), ont été testées sur des données réelles récentes collectées dans la région nord-est du Portugal. La meilleure configuration utilise un SVM et quatre données météorologiques (c’est-à-dire la température, l’humidité relative, la pluie et le vent) et elle est capable de prédire la surface brûlée des petits incendies, qui sont plus fréquents. Ces connaissances sont particulièrement utiles pour améliorer la gestion des ressources de lutte contre les incendies (par exemple, en hiérarchisant les cibles pour les avions-citernes et les équipes au sol).

La table de la base de données est constituée de 13 colonnes (variables) qui permettent d’analyse de la situation.

1. **Coordonnées spatiales de l’axe X** : X dans la carte du parc de Montesinho : 1 à 9

2. **Coordonnées spatiales Y**: axe Y dans la carte du parc de Montesinho : 2 à 9

3. **Mois** : mois de l’année : 'jan' à 'dec'

4. **Jour** : jour de la semaine : 'mon' à 'dim'

5. **FFMC**: Indice FFMC du système IFM : 18,7 à 96,20

6. **DMC** : Indice DMC du système FWI : 1,1 à 291,3

7. **DC** : Indice DC du système FWI : 7,9 à 860,6

8. **ISI** : Indice ISI du système FWI : 0,0 à 56,10

9. **Temp**: température en degrés Celsius : 2,2 à 33,30

10. **HR** : humidité relative en % : 15,0 à 100

11. **Vent** : Vitesse du vent en km/h : 0,40 à 9,40

12. **Pluie :** pluie extérieure en mm/m2 : 0,0 à 6,4

13. **Superficie** : la superficie brûlée de la forêt (en ha) : 0,00 à 1090,84 (cette variable de sortie est très biaisée vers 0.0, elle peut donc faire sens pour modéliser avec la transformée logarithmique).

**Piste d’analyse**

Dans [Cortez et Morais, 2007], l’aire de sortie a d’abord été transformée avec une fonction ln(x+1). Ensuite, plusieurs méthodes de Data Mining ont été appliquées. Après l’ajustement des modèles, les sorties ont été post-traité avec l’inverse de la transformation LN(x+1). Quatre configurations d’entrée différentes ont été utilisé. Les expériences ont été menées à l’aide d’une validation croisée de 10 fois x 30. Deux des mesures de régression ont été mesurées : MAD et RMSE. Une machine à vecteurs de support gaussienne (SVM) alimentée avec seulement 4 conditions météorologiques directes (température, HR, vent et pluie) ont obtenu la meilleure valeur MAD : 12,71 +- 0,01 (moyenne et intervalle de confiance à 95 % près en utilisant une distribution de l’étudiant t). Le meilleur RMSE a été atteint par le prédicteur moyen naïf. Analyse de la courbe d’erreur de régression (REC) montre que le modèle SVM prédit plus d’exemples à l’intérieur d’une erreur admise plus faible. En effet, le modèle SVM prédit de meilleurs petits incendies, qui sont majoritaires.