Polytech Sorbonne

Projet Apprentissage Statistique : Forest Cover Type Prediction

Boussad Merhane, Romaric Kanyamibwa

MAIN5

Année 2018 - 2019

Table des matières

1 Introduction

Le projet dont nous allons discuter dans la suite de ce rapport a pour objectif la classification de différents types de forêts à partir d'observations faites sur le terrain. Il s'agit d'un problème de classification multi-classe avec 7 classes. C'est un problème qui avec l'avènement du Machine Learning et le Deep Learning a été beaucoup étudier depuis les années 2000.

L'intérêt de ce projet repose dans son application potentielle.En effet la gestion des ressources etant une problématique actuelle, ce projet avec différents modelés predictive pourrait potentiellement aider les gestionnaires des ressources naturelles d'élaborer des stratégies des gestions des différents écosystèmes sur leur responsabilité .

2 Données

L'ensemble des données est composé de 581 012 instances, où chaque observation (instance) correspond à une cellule de 30 x 30 mètres. Pour chaque observation, 12 mesures (attributs) sont données. Ainsi, l'ensemble de données final est composé de 10 variables quantitatives, de 4 **wilderness areas** binaires et de 40 variables binaires de **soil type** , pour un total de 54 colonnes de données. Ci-dessous, quelques descriptions à propos de ces est donné :

- Elevation : Altitude en mètres.
- Aspect : Orientation.
- Slope : Pente en degrés.
- Horizontal_Distance_To_Hydrology : Distance horizontale par rapport aux éléments aquatiques de surface les plus proches.
- Vertical_Distance_To_Hydrology : Distance verticale par rapport aux éléments aquatiques de surface les plus proches.
- Horizontal_Distance_To_Roadways : Distance horizontale de la route la plus proche.
- Hillshade_9am : Indice d'ombrage à 9h, solstice d'été (0 à 255).
- Hillshade_Noon : Indice d'ombrage à midi, soltice d'été (0 à 255).
- Hillshade_3pm : Indice d'ombrage à 15h, solstice d'été (0 à 255).
- Horizontal_Distance_To_Fire_Points : Distance horizontale par rapport aux points d'inflammation d'incendie de forêt les plus proches.
- Wilderness_Area : Désignation d'une réserve intégrale. 4 colonnes binaires, 0 (absence) ou 1 (présence).
- Soil_Type : Désignation du type de sol. 40 colonnes binaires, 0 (absence) ou 1 (présence).
- Cover_Type : Désignation du type de couverture forestière.C'est l'attribut à prédire (1 à 7).

En ce qui concerne l'attribut Covert_Type, la distribution suivante est observée dans l'ensemble du jeu de données :

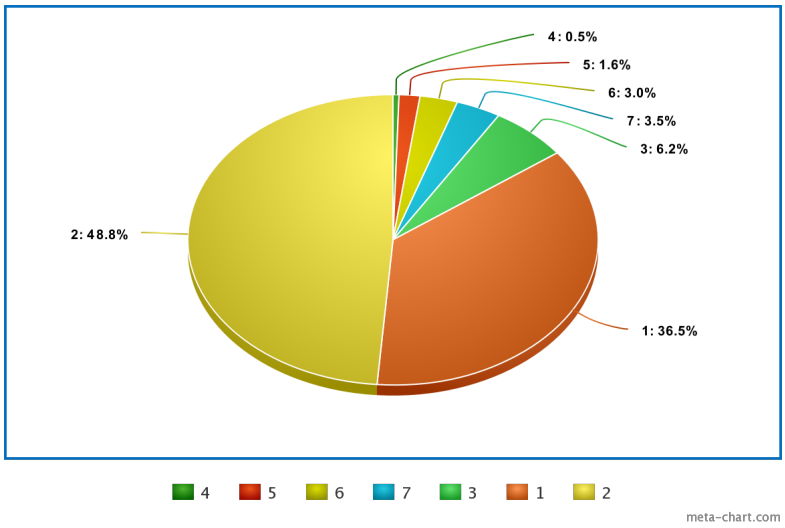

Figure 1 – *Distribution des Cover_Type:Piechart*

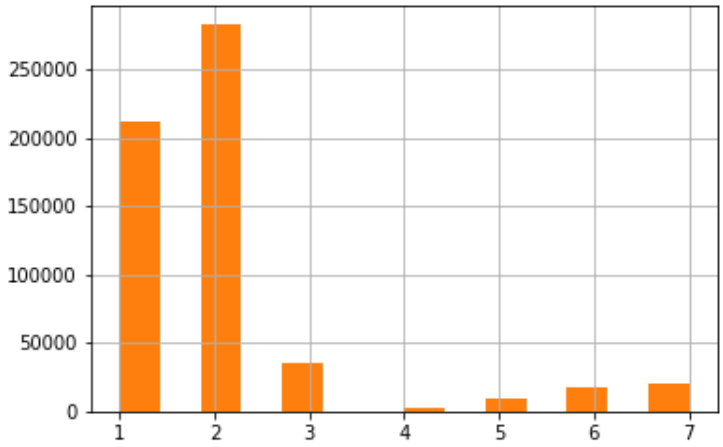

Figure 2 – *Distribution des Cover_Type*

Nous observons que environ **85%** de nos données appartiennent aux classes 1 ou 2 et que **0.5%** de données font partie de la classe 4,ces caractéristiques particulières de notre jeu données vont nous poser des problèmes quand nous allons commencer à implémenter des modelés de prédictions.

3 Analyse des données

L'analyse des données est une étape primordiale dans toute modélisation statistique.Elle nous permet de décrire et explorer les données avant d'en tirer de quelconques lois ou modèles prédictifs de plus elle nous permet d'extraire tous les informations pertinentes contenu dans nos données.

En outre une analyse de données exhaustif nous permet d'e tirer conclusion et pouvoir implémenter des modelés predictives bien adapté et avec les bons features.Il faut savoir aussi que une bonne analyse de données peut améliorer considérablement nos résultats.

Par conséquence pour faire face aux problèmes d'équilibrage vu précédemment nous allons utiliser différentes techniques d'analyse de données.

3.1 Explorations de données

Comme dit précédemment les données à analyser possède 54 features[1] différents plus ou moins significatives.Dans un premier temps pour extraire des informations nous avons voir à travers de graphique comment ces différents features influence le Cover_Type de données.

[1] caractéristique

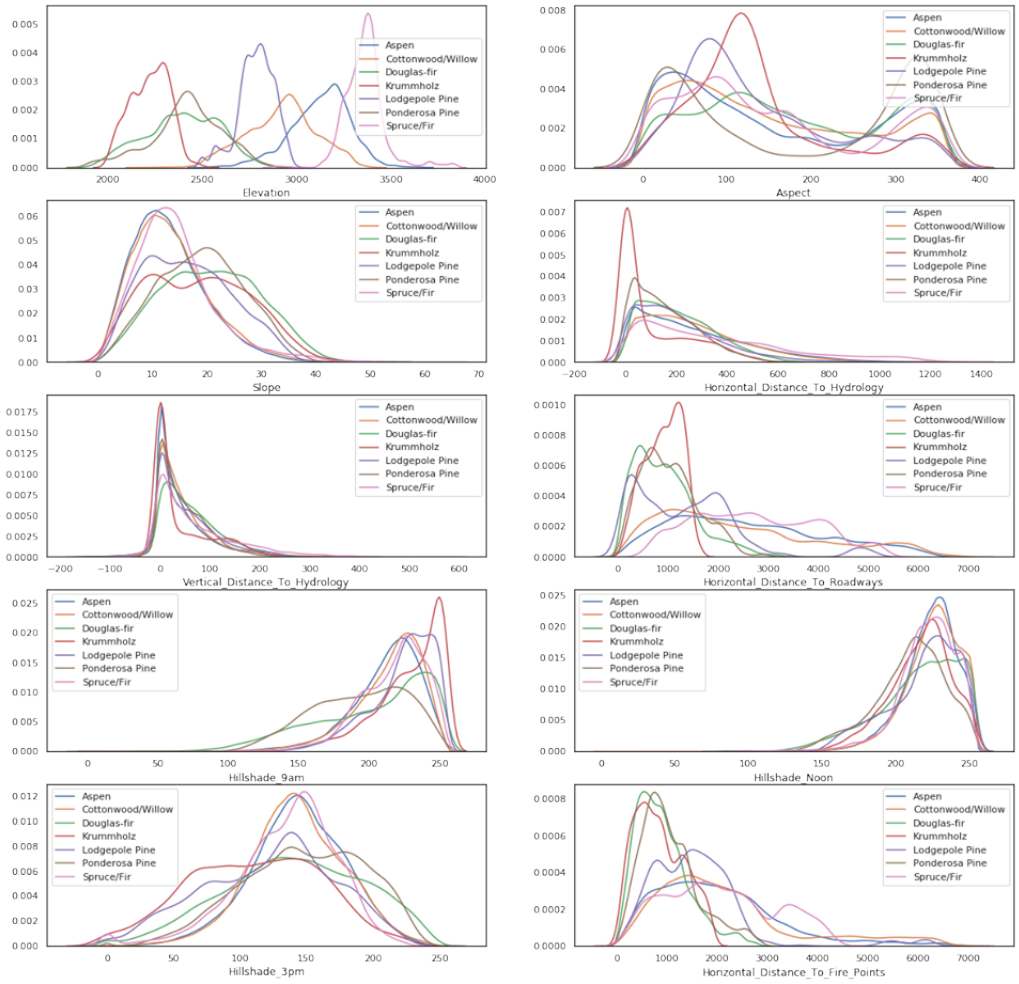

Figure 3 – *Graphes de densité pour les differents features*

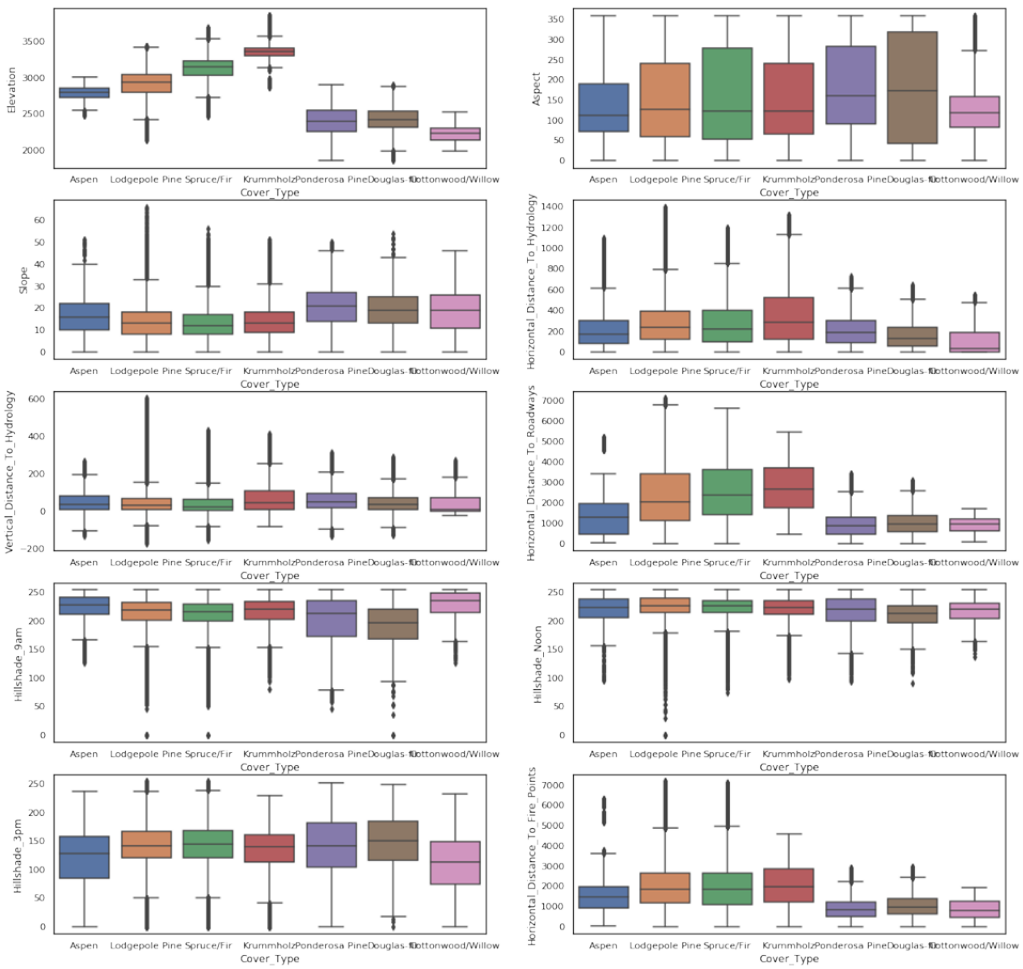

Figure 4 – *Boxplotles differents features*

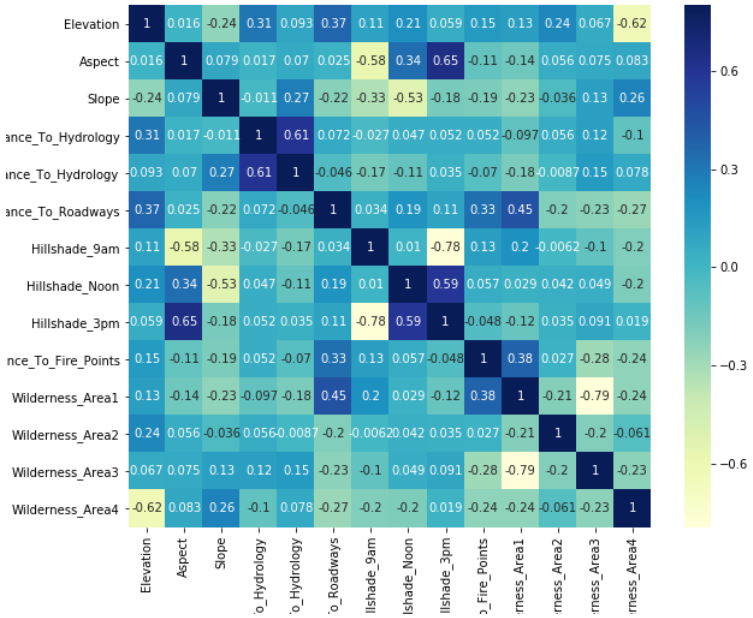

Figure 5 – *Matrice de correlation (heatmap)*

3.2 Sélection des Variables

3.2.1 ACP

4 Classification