Évaluation de performance Apache Spark

Capelle Victor - Okumura Ono Lucas

Objectif

-Évaluer le fonctionnement d’une installation distribuée

Installation

- Structure sur Google Cloud Platform, un seul provider

Perspectives

- Performance et Volume
  - 18k lignes - 581k lignes
- Performance et Nombre d’esclaves
  - 0 - 2 - 3
- Performance et Algorithmes
  - Decision Tree et Random Forest Classifiers

Données

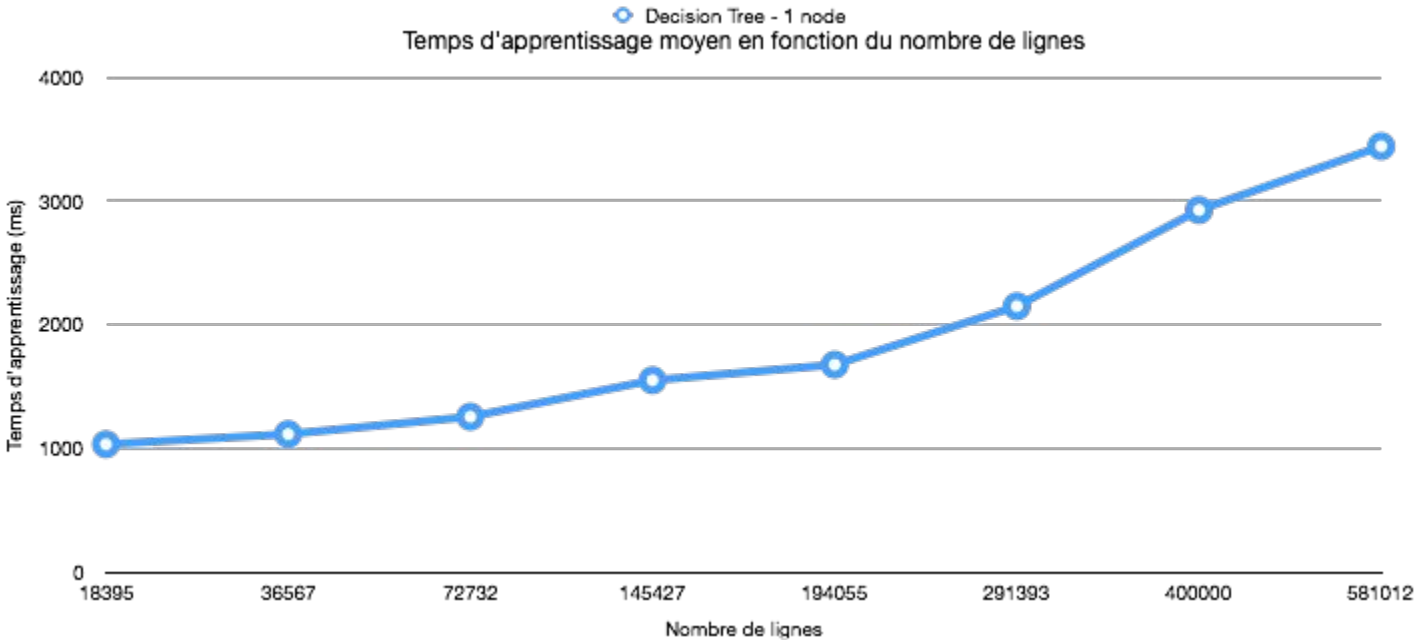
Données sur le type de couverture de terrain en milieu forestier.

Lien : https://archive.ics.uci.edu/ml/datasets/covertype

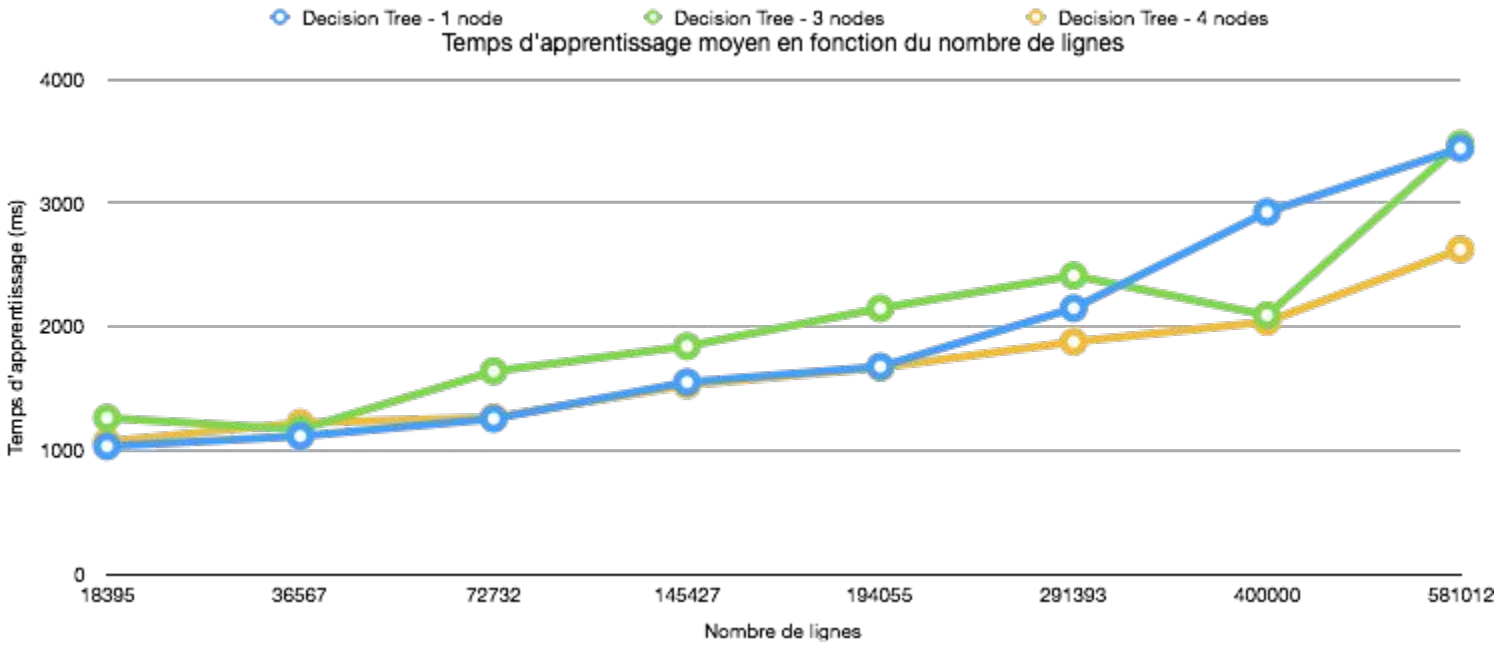
Le but est de prédire le type de couverture de terrain à partir de données cartographiques.

Nous avons fixer la graine de l’aléatoire pour la séparation des données d'entraînement et d’analyse.

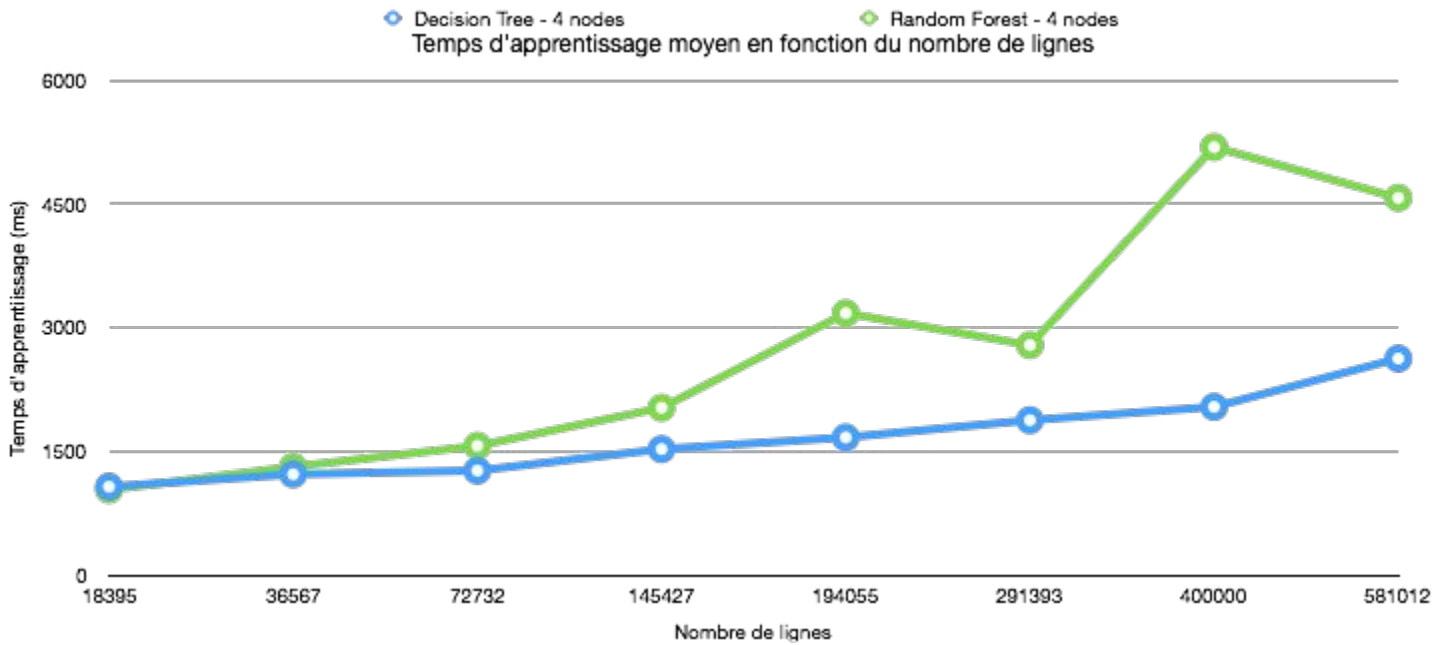
Augmentation du nombre de lignes sur une machine



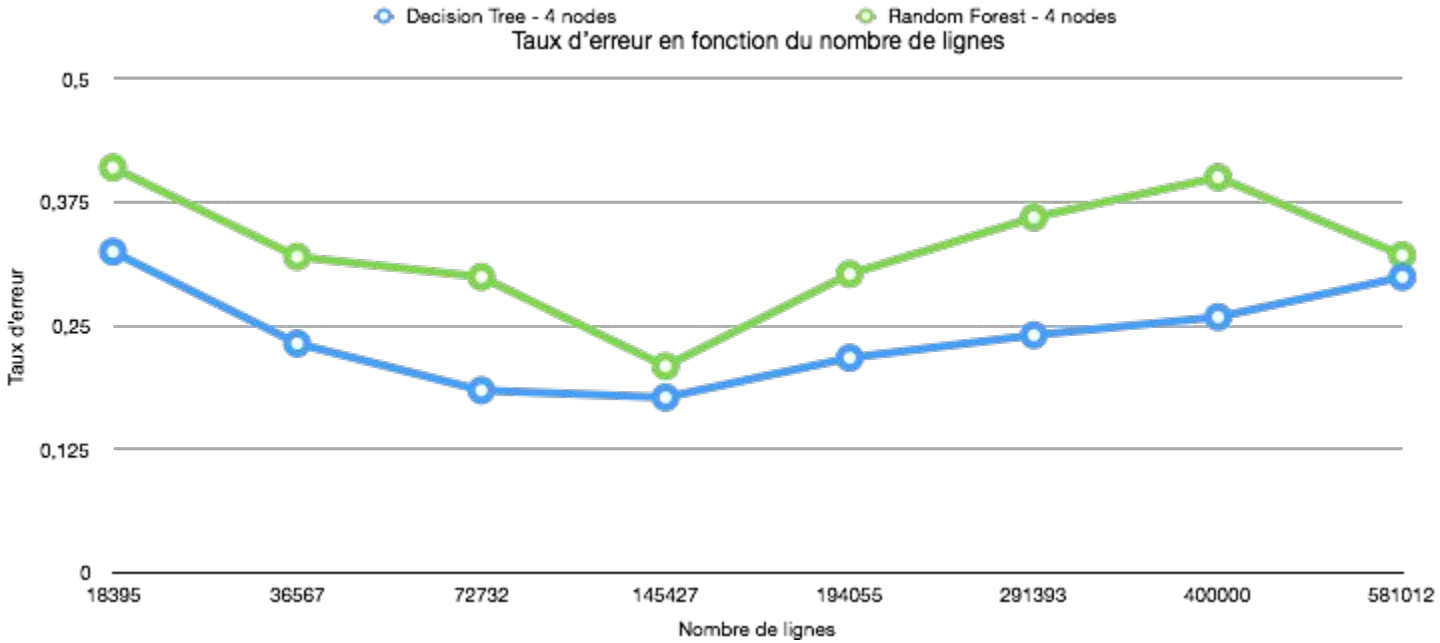
Augmentation du nombre de nodes



Comparaison à un autre algorithme de classification



Comparaison à un autre algorithme de classification



Sources

https://github.com/capellev/SparkMLlib