Verslag Distributed Databases

Dylan Cluyse, Laura Renders, Liam Goethals

14 november 2022

Inleiding

Tijdens het opleidingsonderdeel Distributed Databases maakten wij kennis met Apache Spark. Spark biedt data-analyse voor grootschalige gegevensverwerking. Deze technologie wordt aangeboden voor Java, Python, R en Scala. Voor dit opleidingsonderdeel werd Java gekozen als programmeertaal.

In dit verslag nemen wij u mee in de wereld van machine learning gedreven door Spark.

1. Calculating Air Polution with Spark Regression

Aanpak

Wij wilden verschillende regressiemodellen uitproberen.
    * Lineaire regressie * Random Forest Regressie

Evaluatie

2. Tweet detection with NLP

Aanpak

We hebben hier twee klassen: correcte tweets en fraudulente tweets. We moeten een aanpak vinden voor een binaire classificatiemodel.

Evaluatie

Bij het evalueren van het classificatiemodel ging onze voorkeur uit naar de confusionmatrix. Zo hebben wij een zicht op hoe goed ons model de klassen kan voorspellen. Bij de logistische regressie merken wij een goed evenwicht op tussen valse positieven en valse negatieven. De classificatiemodellen met Random Forest en Decision Trees voorspellen daarentegen weinig valse negatieven, maar wel uitbundig veel valse positieven. Als extra hebben wij ook de nauwkeurigheid van het model berekent.

3. Predicting the burning area of Spanish forest fires.

Aanpak

Gebruikte regressiemodellen: * Lineaire regressie * Random forest regressie

De resultaten van de metrieken waren zwak. Onze presumptie lag bij de volatiele waarden. Sommige waarden waren tussen het bereik van 0 - 10. De ... feature daarentegen had een range van 0 - 1000. Dit hebben we opgelost door een MinMaxScaler toe te voegen aan de pipeline.

Evaluatie

Conclusie

Uit document: