**Handreiking**

|  |  |  |
| --- | --- | --- |
| **Onderwerp** | **:** | Methode gezamenlijke data classificatie |
| **Van** | **:** | Lex Slaghuis, VNG Data en monitoring |
| **Aan** | **:** | Klik hier als u tekst wilt invoeren. |
| **Datum** | **:** | 1 april 2021 |

## Samenvatting

Vanuit de pilot MOR is onderzocht welke meerwaarde uit een landelijke dataset van open MOR data behaald kunnen worden. Een van de onderdelen hiervna was een test met een algoritme waarbij de meldingen op basis van Machine Learning naar BOR MELD worden gecategoriseerd. BORMELD is een categoriseringssysteem van het CROW dat ontworpen is voor het beheer van de openbare ruimte. Voor vragen over BOR MELD kunt u contact opnemen met het CROW ([www.crow.nl](http://www.crow.nl)). Voor algemene vragen kunt u contact opnemen met Lex Slaghuis ([lex.slaghuis@vng.nl](mailto:lex.slaghuis@vng.nl) of 070 373 8008).

Bijlagen:

ontwikkel\_classificatie\_meldingen.ipnb

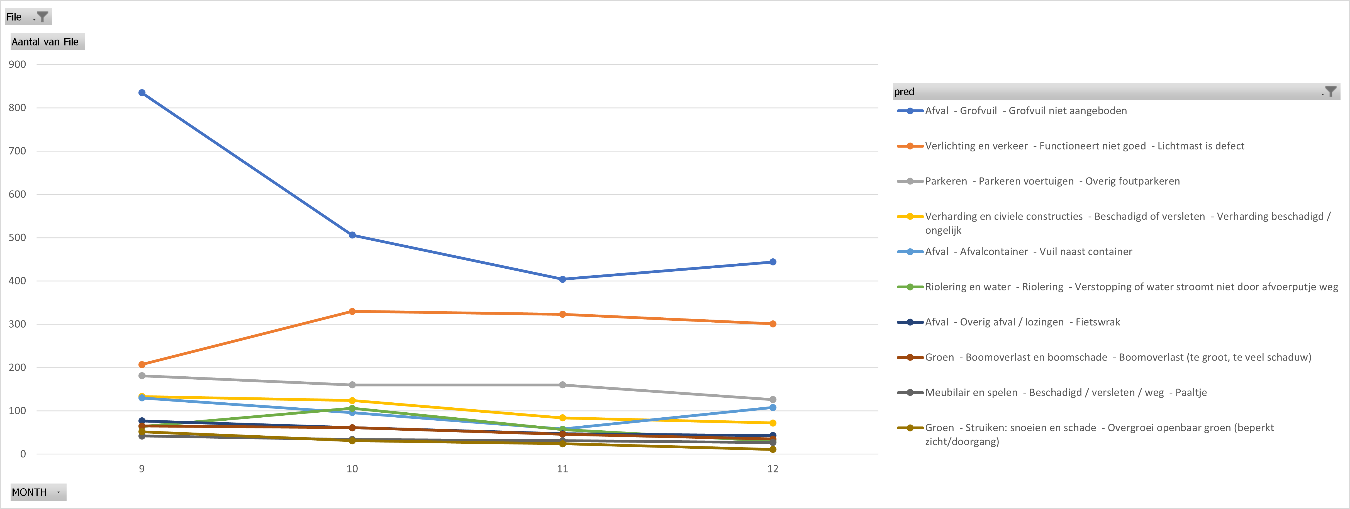
pas\_toe\_classificatie\_meldingen.ipnb

## Data analyse en classificatie

Op basis van algoritmen kunnen meldingen geclassificeerd worden. Door de gemeente Amsterdam is hiervoor een generiek algoritme ontwikkeld. De gebruikte broncode hiervoor is beschikbaar via: <https://gitlab.com/commondatafactory/experiments/signalen-classificatie>. Door deze met de juiste trainingsdata te voeden kan dit algoritme in werking gebracht worden. Indien het algoritme getraind wordt met uw eigen classificatie systeem dan zal deze, indien de datakwaliteit goed is, in hoge mate de juiste classificatie kunnen voorspellen van niet geclassificeerde meldingen. Zo voegt de gemeente Amsterdam zelf op <https://meldingen.amsterdam.nl/incident/beschrijf> de meest waarschijnlijke classificatie toe, in plaats van dit bij gebruikers uit te vragen.

Het is ook mogelijk om BOR Meld classificatie systematiek toe te passen. Hiervoor is een trainingsdataset nodig, waarbij meldingen volgens de BOR MELD standaard geclassificeerd zijn. VNG realisatie beschikt over een gesloten trainings-set op basis van data van meerdere gemeenten hiervoor, en stelt het resultaat een getraind model met een F1 (accuraatheid) van ongeveer 75%, vrij beschikbaar. Het model is beschikbaar in de bijlage ‘model-db-bor-meld.pkl’ bij deze handreiking.

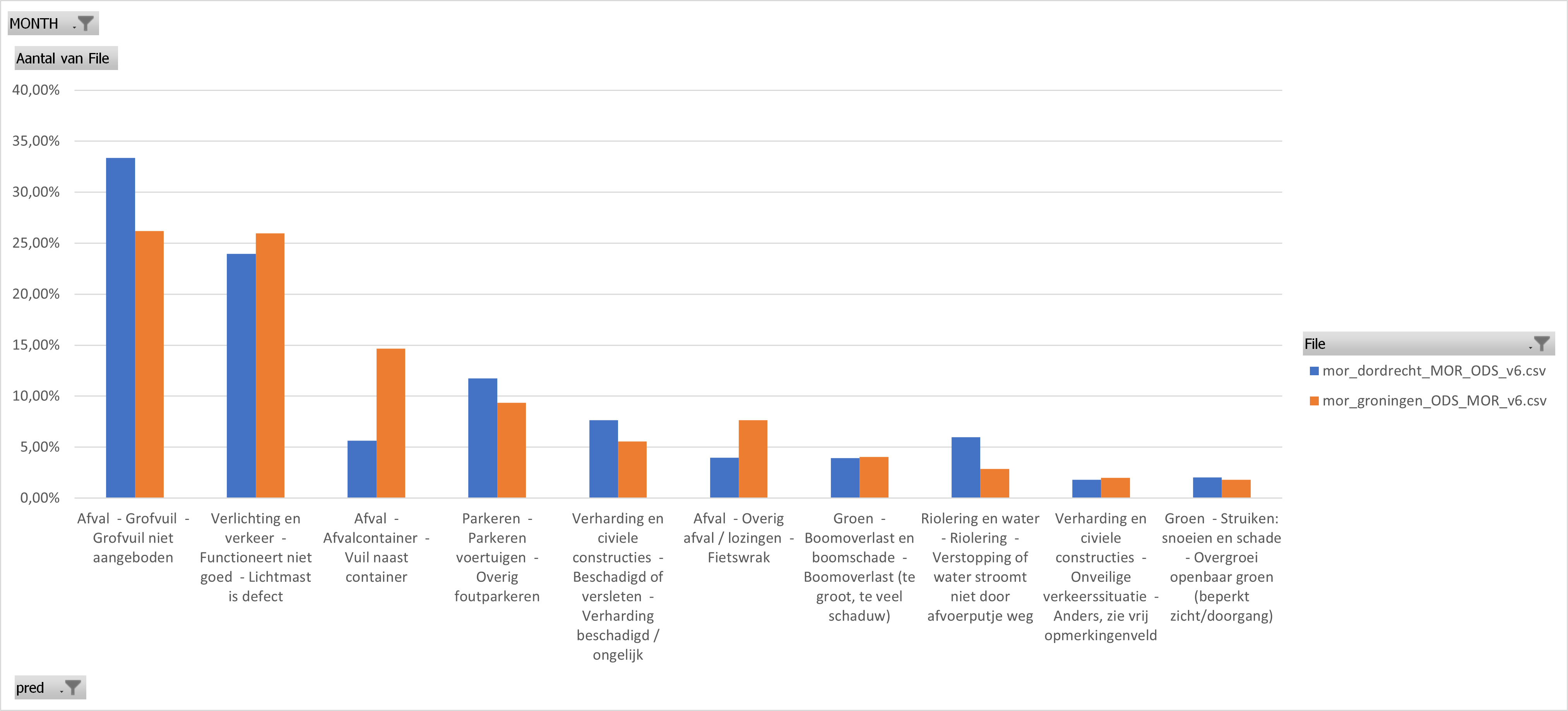
Hiermee kunt u vervolgens het algoritme toepassen op uw eigen data (bijlage pas\_toe\_classificatie\_meldingen.pynb). Om vervolgens het algoritme zelf door te trainen zodat in de toekomst zelf uw eigen meldingen automatisch meldingen kan classificeren. Zodra er voldoende (open ) data beschikbaar kunnen VNG Realisatie, leveranciers en gemeenten gezamenlijk een (of meerdere ) trainings-sets en modellen onderhouden.

Voor sturingsinformatie is het benchmarken van gemeenten door meldingen te classificeren mogelijk.

Figuur Ontwikkeling maanden 9-12

**Nu niet goed genoeg**

Echter de betrouwbaarheid van het model is hiervoor op dit moment niet goed genoeg om hier conclusies uit te trekken. Bij beide gemeenten is ‘grofvuil niet aangeboden’ de meest voorkomende melding. Als gekeken wordt naar het aandeel in het totaal aantal meldingen dan is bij machine classificatie dit 25% EN resp. 22 % voor de gemeenten. Oorspronkelijk zijn er met de hand 17% en 23% meldingen geclassificeerd. Bij de ene gemeente wordt een hoop meer geclassificeerd; bij de andere juist minder. Met substantiële procentuele afwijkingen. Deze afwijkingen zorgen ervoor dat de benchmark op dit moment niet accuraat genoeg is om uitspraken te doen over trends en verdelingen.



Figuur Verdeling van meldingen uitgesplitst tussen gemeenten

## Helaas blijkt in de praktijk dus ook dat de accuraatheid (F1 rond de 75%) niet toereikend voor benchmarking. Alhoewel een algoritme nooit perfect is, zitten er in de gebruikte BOR meld standaard zitten veel ambiguïteiten, die een eenduidige classificatie in de weg staan.

## Een mooi voorbeeld hiervan is zwerfvuil dat bij BOR meld op ‘straat’ , in het ‘water ‘ of in het ‘groen ‘ gemeld kan worden. Bewoners voegen tekstueel deze informatie niet toe, en als dit wel gebeurt is het snel ingewikkeld om te detecteren: ‘ Er is zwerfvuil bij het watertje in het park.(….)’

## Een ander type ambiguïteit is rond ondergrondse vuilcontainers en bijvoorbeeld fietsen (of zwerfvuil, ….). Zinnen als ‘ naast de ondergrondse container staat een zwerfvuil en een roestende fiets ‘ raken de discussie of er nou sprake is van zwerfvuil, een zwerffiets, of een ontoegankelijke ondergrondse container’. Misschien wel alle drie, en dat roept de vraag op of er een classificatieboom (waarbij altijd het een of het ander is) wel zo nuttig is in deze context. Wellicht dat ‘tags’ waarbij er meerdere tags met een indicatie van kans een nuttigere classificatie (en benchmark!) kan opleveren. Immers, het benchmark inzicht in de lijn van…. ‘ zwerfvuil betreft 13% van de meldingen , of er zijn 123 zwerffietsen per 1000 inwoners in wijk X.’