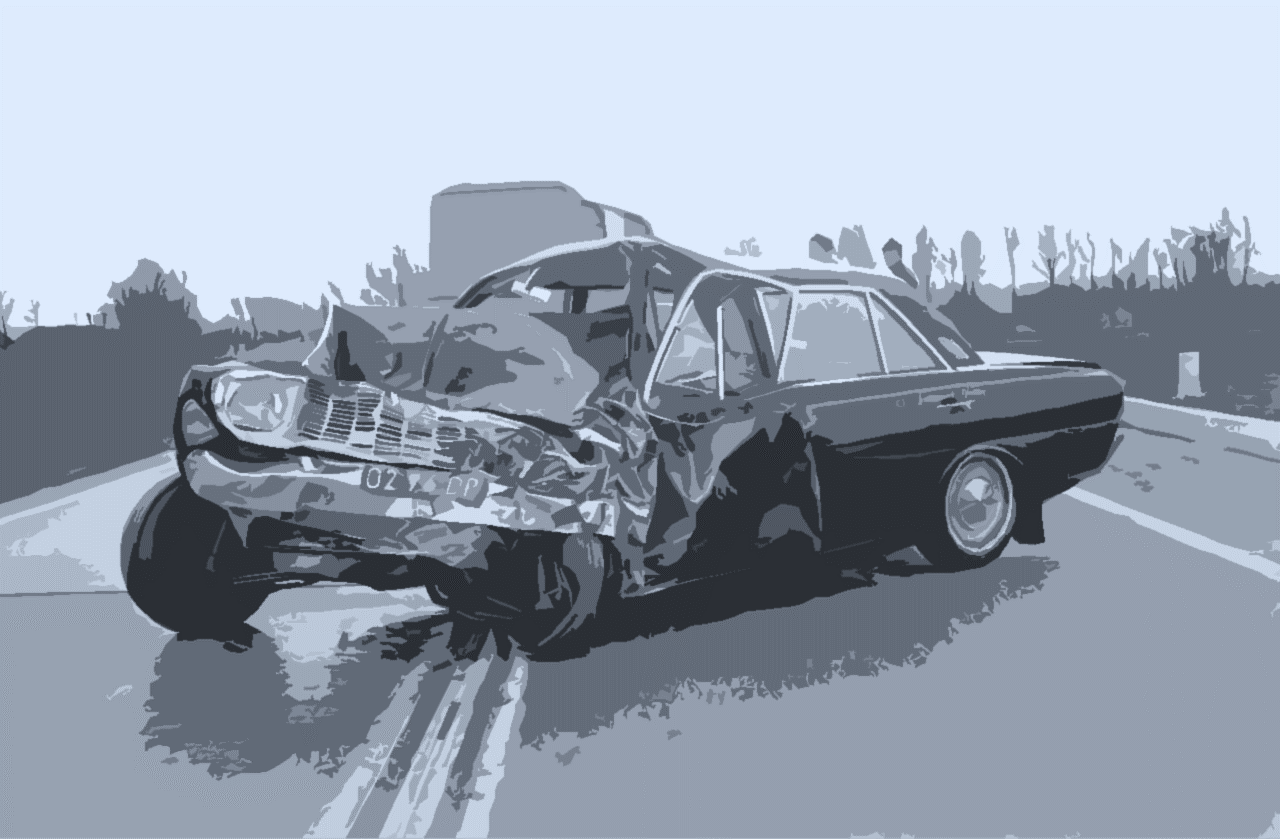
VERKEERSVEILIGHEID IN NEDERLAND:   
HET GEVAAR ZIT IN DE FACTOREN

Informatiekunde - Universiteit van Amsterdam

Universiteit van Amsterdam  
  
BACHELOR INFORMATIEKUNDE SCRIPTIE

**Stijn Melchior Willem Hering  
12377422**



**Afstudeerbegeleider: dr. Robin Langerak  
Tweede beoordelaar: prof. dr. Paul Groth  
Omvang onderzoek: 18EC**

# Abstract

Volgens cijfers van de ***World Health Organization*** sterven er jaarlijks ongeveer 1,35 miljoen mensen door verkeersongelukken. Nog eens 20 tot 50 miljoen mensen lopen niet-dodelijk letsel op. In Nederland is deze trend ook aanwezig. Ondanks de verplichte verminderde mobiliteit en het noodzakelijke thuiswerken in 2020 en begin 2021, zijn er volgens het**Ministerie van Infrastructuur en Waterstaat** 110 méér verkeersdoden geregistreerd dan de doelstelling van maximaal 500. Verder lijkt het onwaarschijnlijk dat de doelstelling van het aantal ernstige verkeersgewonden in 2020. Er is de afgelopen decennia veel onderzoek gedaan naar het gebruik van *datascience* en *artificial intelligence* (*AI*) in het verkeer. In Nederland zijn dit soort onderzoeken beperkt, terwijl de voordelen van het uitvoeren van dit soort onderzoeken voor de verkeersveiligheid veelvuldig zijn bewezen. In dit onderzoek worden technieken en methoden uit het domein van de statistiek, *datascience* en AI toegepast op het *Bestand Geregistreerde Ongevallen in Nederland* (**BRON**). Het doel van dit onderzoek is om te toetsen of een classificatiemodel kan bijdragen aan de verbetering van de verkeersveiligheid in Nederland om zo het aantal doden en gewonden die vallen per jaar te reduceren. In dit onderzoek is er gebruikt gemaakt van twee machine learning modellen die de drie afloop klassen, uitsluitende materiele schade, letsel en dodelijk classificeren aan de hand van categorische variabelen. Een model dient als baseline en de ander als experimenteel model. Het baseline model is een ‘simpele’ *logistieke regressie* model. Het bestaat uit een *Catboost* *classifier*, het gaat hier specifiek om een *gradient boosting trees* model dat ontworpen is voor categorische data. Het *logistieke regressie* model scoort een gewogen *accuracy* van 0,64, de *Catboost* *classifier* scoort een gewogen *accuracy* van 0.85. Beiden modellen kunnen helpen de verkeersveiligheid in Nederland te verbeteren. Door het analyseren van de *feature importance* en het berekenen van waarschijnlijkheden kunnen verkeersscenario’s met verschillende factoren en locaties met elkaar worden vergeleken. Hierdoor moet duidelijk worden welke factoren het meest bijdragen aan een veilige en onveilige verkeersituatie. Door deze kennis kunnen beleidsmakers meer informatieve keuzes maken om de verkeersveiligheid in Nederland te verbeteren.

Inhoudsopgave

[1 Abstract 2](#_Toc75873953)

[3 Introductie 4](#_Toc75873954)

[3.1 Inleiding 4](#_Toc75873955)

[3.2 Het Bestand Geregistreerde Ongevallen in Nederland (BRON) 6](#_Toc75873956)

[4 Theoretisch Kader 8](#_Toc75873957)

[4.1 Praktische relevantie van het verzamelen verkeersdata 8](#_Toc75873958)

[4.1.1 Verkeersincidenten data 9](#_Toc75873959)

[4.2 Gebruikte Toepassingen 9](#_Toc75873960)

[4.2.1 Traditionele methoden 9](#_Toc75873961)

[4.2.2 Neurale Netwerken 10](#_Toc75873962)

[4.2.3 Deep - Learning 11](#_Toc75873963)

[4.2.4 Ensemble Methoden 12](#_Toc75873964)

[4.3 Gerelateerd onderzoek 15](#_Toc75873965)

[5 Methodologie 16](#_Toc75873966)

[5.1 Ongevallen en Partijen 16](#_Toc75873967)

[5.2 Pre-Processen 17](#_Toc75873968)

[5.3 Feature Engineering 17](#_Toc75873969)

[5.4 Data-Analyse 18](#_Toc75873970)

[5.5 AI-modellen 18](#_Toc75873971)

[5.6 CatBoost Classifier 19](#_Toc75873972)

[6 Resultaten 21](#_Toc75873973)

[6.1 Data-analyse 21](#_Toc75873974)

[6.2 Chi-Kwadraattest 21](#_Toc75873975)

[6.3 Feature Importance Catboost 21](#_Toc75873976)

[6.4 Logistic Regression 21](#_Toc75873977)

[6.5 Catboost Classifier 22](#_Toc75873978)

[7 Discussie 23](#_Toc75873979)

[8 Conclusie 24](#_Toc75873980)

[9 Bibliografie 26](#_Toc75873981)

[10 Appendix 29](#_Toc75873982)

# Introductie

## Inleiding

Volgens cijfers van de *World Health Organization* (**WHO**) sterven er jaarlijks ongeveer 1,35 miljoen mensen door verkeersongelukken. Nog eens 20 tot 50 miljoen mensen lopen niet-dodelijk letsel op. Velen van hen raken gehandicapt als gevolg van dit letsel (WHO, 2020).   
Naast het leed dat mensen ondervinden ten gevolge van verkeersongelukken veroorzaken verkeersongelukken ook aanzienlijke maatschappelijke kosten. Het gaat om zowel sociale, medische en immateriële als economische kosten. Het voorkomen van verkeersongelukken met verkeersdoden en verkeersgewonden tot gevolg is hierdoor een volksgezondheidsprioriteit geworden waarvan de sociale en economische implicaties zich tot ver buiten de vervoerssector uitstrekken (Fumagalli et al., 2017).   
Volgens de **WHO** (2020) zijn de totale kosten die gemaakt worden door verkeersongelukken wereldwijd goed voor drie procent van het bruto binnenlands product (bbp). Zo ook in Nederland. Volgens de meest recente cijfers van de Stichting Wetenschappelijk Onderzoek Verkeersveiligheid (SWOV) uit 2018 worden de totale maatschappelijke kosten van verkeersongevallen geschat op gemiddeld € 17 miljard (€ 15,8 tot € 18,6 miljard); ruim 2% van het bbp. Dit is aanzienlijk hoger dan andere maatschappelijke kosten die door het verkeer in Nederland worden gemaakt. Zo zijn de verkeerskosten met betrekking tot congestie ongeveer 3,3 tot 4,3 miljard euro en schade aan het milieu goed voor 7 miljard euro (SVOW, 2020a).

Volgens cijfers van het *Ministerie van Infrastructuur en Waterstaat* (**IenW**)(2019) is meer dan driekwart van de totale maatschappelijke kosten, ten gevolge van verkeersongelukken toe te rekenen aan verkeersdoden en verkeersgewonden. Het overige deel bestaat uit materiële schade. Gemiddeld bedragen de kosten per verkeersdode circa 2,8 miljoen en per ernstig verkeersgewonde ruim 300.000 . In Nederland bestaan de kosten zowel uit ‘harde’ economische kosten – zoals medische kosten en schade aan voertuigen – als uit immateriële kosten. Bij immateriële kosten gaat het om verlies van levensjaren en kwaliteit van leven. Bijna de helft van de totale kosten zijn immateriële kosten (IenW,2019).

Sinds 1999 was er een dalende trend zichtbaar in het aantal verkeersdoden in Nederland. Van 1999 tot 2010 halveerde het aantal verkeersdoden. De afgelopen tien jaar stagneerde dit cijfer. Al een decennia lang is er geen significante verandering te zien in het aantal verkeersgewonden en verkeersdoden in Nederland, blijkt uit de cijfers van het **IenW**(2021). Met 19.700 ernstige verkeersgewonden in 2011 en 21.400 in 2019 (Rijkswaterstaat, 2020a) is er in de afgelopen tien jaar zelfs een positieve trend te zien. De doelstelling die waren gezet in Nederland voor het jaar 2020, namelijk minder dan 500 verkeersdoden en 10.600 ernstige verkeersgewonden (IenW, 2018), zijn omdanks de Corona pandemie en de restricties die bestonden door nieuwe regels en noodwetten niet gehaald.   
Ondanks de verplichte verminderde mobiliteit en het noodzakelijke thuiswerken in 2020 en begin 2021, zijn er volgens het**IenW** (2021) 110 méér verkeersdoden geregistreerd dan de doelstelling van maximaal 500. Verder lijkt het onwaarschijnlijk dat de doelstelling van het aantal ernstige verkeersgewonden in 2020 gehaald zal worden als er wordt gekeken naar het totaal aantal ernstige verkeersgewonden uit 2019. In 2019 ging het om 21.400 ernstige verkeersgewonden Dit is meer dan het dubbele van wat maximaal wordt verwacht in 2020. Het verminderen van het aantal verkeersdoden en gewonden is van belang als Nederland zich aan de nationale ambitie wil houden: de halvering van doden en ernstig verkeersgewonden in 2030 (SVOW, 2020b). Verder draagt een vermindering van het aantal verkeersdoden en gewonden bij aan een verbeterde verkeersveiligheid in Nederland en worden de maatschappelijke lasten gereduceerd. Om dit doel te behalen is het essentieel dat er wordt geëxperimenteerd met meerdere innovatieve ideeën om het aantal verkeersincidenten te verminderen. Volgens de SWOV (2020b) zijn er ondanks de maatregelen uit het regeerakkoord van het kabinet Rutte 3, de investeringsimpuls en de risico-gestuurde aanpak vanuit het Strategisch Plan Verkeersveiligheid 2030 meer en effectievere maatregelen nodig (IenW, 2018).

Er is de afgelopen decennia veel onderzoek gedaan naar het gebruik van *datascience* en *artificial intelligence* (*AI*) in het verkeer. In deze onderzoeken wordt er in het algemeen een antwoord gezocht voor drie soorten problemen, de verkeersclassificatie, verkeersgeneratie en verkeersprognose-of voorspelling. Verkeersclassificatie gaat over het classificeren van verkeersdata, zoals het classificeren van de drukte op de weg in lichte congestie, middelmatige congestie en hevige congestie of het classificeren van de afloop van een ongeluk. Verkeersgeneratie gaat over het generen of modelleren van representatieve verkeersdata. Door het generen van verkeersdata kunnen onderzoekers of verkeerinstanties gegeneerde data gebruiken om verkeersstromen te simuleren of nieuwe wetgeving of innovaties te testen zonder ‘echte’ verkeersdeelnemers in gevaar te brengen. Het verkeersprognose probleem heeft betrekking op het maken van verkeersvoorspellingen of verkeersprognoses. Denk hierbij aan onderzoeken naar het voorspellen van verkeerssnelheden, verkeersdrukte, reisbehoeften, reistijden en verkeersongelukken. Veel van deze onderzoeken hebben gezorgd voor nieuwe ontwikkelingen en innovaties voor het verkeer en de verkeersveiligheid (Ermagun & Levinson, 2018; Yin et all., 2020; Yuan & Li, 2021). In Nederland zijn dit soort onderzoeken beperkt, terwijl de voordelen van het uitvoeren van dit soort onderzoeken voor de verkeersveiligheid veelvuldig zijn bewezen. De Nederlandse overheid heeft wel stappen genomen om onderzoek omtrent AI te stimuleren. Door het creëren van het Open Data-portaal (Nederlandse Overheid, 2019) kunnen burgers en professionals zoeken in regelgeving, bekendmakingen en andere datacollecties van de overheid (Dataregister van de Nederlandse Overheid | Data overheid, z.d.).

## Het Bestand Geregistreerde Ongevallen in Nederland (BRON)

In dit onderzoek ligt de focus voornamelijk op classificeren van de afloop van een verkeersongeluk aan de hand van categorische waarden. Het doel van dit onderzoek is om te toetsen of een classificatiemodel kan bijdragen aan de verbetering van de verkeersveiligheid in Nederland. Door middel van technieken en methoden uit het domein van de statistiek, *datascience* en AI toe te passen op het *Bestand Geregistreerde Ongevallen in Nederland* (**BRON**) wordt getracht uitsluitsel te geven of dit mogelijk is. Het BRON is beschikbaar is gesteld als *OPENDATA* (Rijkswaterstaat, 2020b) door het Ministerie van Economische Zaken en Klimaat (EZK) via de Digitale Delta (Ministerie van Economische Zaken en Klimaat, 2019) en wordt vanaf 2010 bijgehouden door de Nederlandse politie en/of weginspecteurs van Rijkswaterstaat. Het BRON bevat veel informatie over verkeersongevallen zoals locatie, staat van het wegdek (nat, droog, ijzel etc), informatie over het soort wegdek, informatie over het voertuig. Een deel van het BRON is in 2018 afgeschermd om aan deAlgemene verordening gegevensbescherming (AVG) te voldoen. Hierdoor is data uit het BRON niet terug te leiden naar eventuele slachtoffers of betrokkenen. Met behulp van het BRON wordt er antwoord geven op de volgende onderzoekvraag: ***Is het mogelijk om de verkeersveiligheid in Nederland te verbeteren door de afloop van ongelukken te voorspellen aan de hand van historische data?***  
Om een antwoord te formuleren voor deze onderzoeksvraag moet er eerst een antwoord worden gegeven op de volgende subvragen.

* Wat is de maatschappelijk relevantie van het verzamelen van verkeersdata?
* Welke *machine-learning* methoden worden gebruik voor verkeersgerelateerde vraagstukken?
* Welke machine-learning methoden worden toegepast voor het classificeren van de afloop van ongelukken?
* Hoe zien de data eruit?
* Welke modellen zijn toegepast in dit onderzoek?
* Wat kan er verbeterd worden aan het model?
* Wat geven de resultaten aan en waarom kunnen deze resultaten helpen bij het verbeteren van de verkeersveiligheid in Nederland?

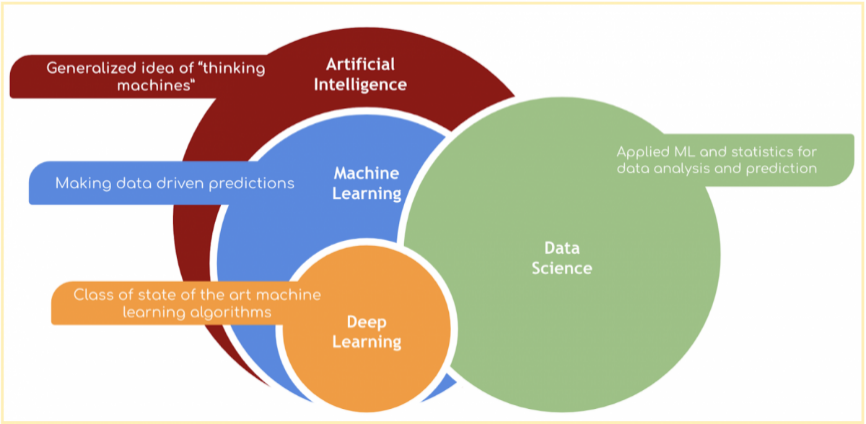
Door middel van kwantitatief onderzoek, via data-analyse en het maken van een classificatiemodel, wordt kennis en inzicht verkregen over factoren die invloed hebben op verkeersincidenten.Data over verkeersongelukken is verzameld via *Overheid.nl*. De data analyse is uitgevoerd via *JupyterLab* een web gebaseerde interactieve ontwikkelomgeving voor *Jupyter notebooks*, code, en data. In de *JupyterLab* environment is vervolgens ook een voorspellend model gebouwd. Uit de data-analyse is duidelijk geworden welke factoren en omstandigheden het meeste invloed hebben op de kansen op verkeersongelukken. Deze factoren worden gebruikt bij het classificeren van verkeersongelukken in de klassen *uitsluitend materiele schade, verkeersgewonden* en *verkeersdoden*. Door het classificeren wordt duidelijk waar en met welke factoren de kans op ‘*ernstige’* verkeersongelukken het grootst is.

In het volgende segment wordt het belang van verkeersdata toegelicht, vervolgens wordt verduidelijk gemaakt welke *data-science* en *AI* toepassingen er al zijn onderzocht en toegepast met betrekking tot het classificeren van de afloop van verkeersongelukken. In het segment dat volgt wordt de methodologie toegelicht en worden de gebruikte dataset beschreven. Vervolgens worden de resultaten van de data-analyse en het voorspellend model gedemonstreerd. Het laatste segment behandelt de discussie en conclusie van dit onderzoek.

# Theoretisch Kader

## Praktische relevantie van het verzamelen verkeersdata

Het gebruik van *data-science* en *AI* in het verkeer is de afgelopen jaren enorm toegenomen. Deze ontwikkeling zijn versneld door de intrede van big-data en het Internet of Things (IoT). Big data en het IoT zijn van significant belang voor het managen en analyseren van het verkeer om de gezondheid en veiligheid van mensen in het verkeer te waarborgen (Ge, Bangui Buhnova, 2018).



**Figuur 1: Illustratie van AI subcategorieën  
(Wang, 2020)**

Het overgrote deel van de ontwikkelde landen verzamelt verkeersdata. Door een variërende infrastructuur is er echter een verschil tussen landen in hoe dit wordt aangepakt. Het verzamelen van verkeersdata gaat voornamelijk door middel van *smart devices* die in een netwerk in verbinding staan met elkaar (IoT). Deze netwerken van smart devices worden beheerd door organisaties of overheidsinstanties die een belang hebben bij de wegen en het verkeer. Deze smart devices zijn er in verschillende vormen. Verder is het type data dat ze (automatisch) verzamelen ook erg divers. Zo wordt er data verzameld via sensoren, camera’s, voertuigen en detectie apparatuur. Verder wordt er ook nog verkeersdata handmatig verzameld door toezichthouders van overheidsinstanties of organisaties. Het gaat hier wel om een kleiner gedeelte van dataverzameling omdat handmatige dataverzameling vaak wordt gedaan voor specifieke doeleinden zoals bij verkeersincidenten(Barceló, Kuwahara & Miska, 2010).   
Alleen in Nederland wordt er volgens het Nationaal Dataportaal Wegverkeer (NDW), een nationaal loket voor wegverkeersgegevens in Nederland, zowel van rijkswegen en provinciale wegen als van gemeentelijke hoofdwegen; per dag meerdere terabytes aan verkeersdata verzameld. Het verzamelen van deze data gebeurt voor het grootste gedeelte automatisch en in *real-time* (NDW, 2020)

Aangezien het verkeer en vervoer een belangrijke basis vormen in een samenleving, is het belangrijk dat wegen vrij en toegankelijk blijven. Verkeerdata heeft hier een significante rol in. Zo worden verkeersdata gebruikt voor: routeplanning, verkeerstoezicht, calamiteitenbeheer, verkeersregeling, anomaliedetectie, situatieherkenning en verkeersvoorspelling. Bovendien kunnen verkeersdata het risico van vervuiling en verkeersincidenten die de gezondheid en veiligheid van de burgers kunnen schaden, tot een minimum beperken (Ge, Bangui & Buhnova, 2018). In dit onderzoek zullen we ons focussen op verkeersdata met betrekking tot verkeersincidenten.

### Verkeersincidenten data

In het verleden werd alle data over verkeersincidenten verzameld door de politie. Deze data werden vervolgens doorgestuurd naar desbetreffende ministeries of branche organisaties die alle gegevens verzamelden en er vervolgens statische analyses op uitvoerden. Deze data kunnen worden onderverdeeld in drie categorieën.

* informatie met betrekking tot de plaats van het ongeval: zoals, datum en tijd van ongeval, locatie in de vorm van coördinaten, naam van gemeente waar het ongeval plaatsvond, informatie omtrent de weg(weg type, maximum snelheid, conditie van de weg, etc), weersomstandigheden, rijbaanverlichting, etc.
* informatie met betrekking tot voertuigen: type van betrokken voertuigen, betrokkenheid van aanhangwagens, doorrijden na ongeval, nummerplaat/landcode, etc.
* informatie met betrekking tot het personen: soort betrokkenheid (bestuurder, passagier, voetganger), geboortejaar, geslacht, afloop van het ongeluk, etc.

De intrede van elektronische gegevensverwerking; in de vorm van big data en IoT heeft ervoor gezorgd dat de politie deze gegevens accurater kan verzamelen en eerder kan analyseren. Daarnaast werd het ook mogelijk om betrouwbaardere conclusies te trekken omdat de verzamelde data een hogere mate van validiteit bezitten (Schlögl & Stütz , 2019).

## Gebruikte Toepassingen

Er zijn verschillende onderzoeken gedaan voor het classificeren/voorspellen van de afloop van ongelukken. De methoden die worden gebruikt in deze onderzoeken zijn onder te verdelen in drie categorieën:

* Traditionele methoden
* Neurale Netwerken of Deep-Learning Methoden
* Ensemble Methoden

Voor elk van deze methoden zal een korte toelichting worden gegeven. Vervolgens wordt een aantal onderzoeken aangehaald die gebruik hebben gemaakt van de eerder genoemde methoden.

### Traditionele methoden

Traditionele methoden omvatten veelal statistische methoden en conventionele *machine-learning* modellen zoals, *decision trees,* regressie en classificatie modellen. Traditionele methoden vereisen dat de data waar mee wordt gewerkt aan bepaalde veronderstellingen voldoet, hierdoor werken deze modellen niet goed met complexe data. Veel traditionele methoden werken aan de hand van sterke stationaire data assumpties. Er moet eerst aan deze assumpties worden voldaan voordat een model op een juiste manier kan worden toegepast en uiteindelijk accurate voorspellen kan doen. Voorbeelden van enkele data-assumpties zijn:

* Afhankelijke of onafhankelijke variabele
* Lineaire relaties
* Geen *outliers* in de data
* Normaalverdeling
* Homogeniteit tussen groepen

Traditionele methoden slagen er vaak niet in om rekening te houden met sterke, niet-lineaire verbanden. Dit is terug te zien in onderzoek omtrent verkeersdata. Door de vaak *spatio-temporal*-vorm is verkeersdata over het algemeen te complex voor traditionele methoden. Het gevolg hiervan is dat (enkele) traditionele methoden relatief weinig gekozen worden als oplossing voor verkeersgerelateerde vraagstukken. Bovendien zijn deze methoden vrijwel alleen goed toepasbaar op kleinere datasets (Li & Shahabi, 2018; Yuan, Zhou & Yang, 2018; Yin et all., 2020).

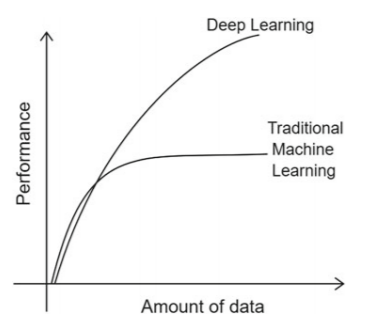
### Neurale Netwerken

Rond de eeuwwisseling, leidden de grotere beschikbaarheid van data en de toenemende rekenkracht tot grotere successen van neurale netwerken (NN). NN modellen zijn een type *machine-learning* algoritme die zijn geïnspireerd op de structuur en functie van de hersenen, vandaar worden ze neurale netwerken genoemd . Een NN maakt gebruikt van meerdere lagen van *nodes* die geleidelijk kenmerken of features van een hoger niveau extraheren uit een ruwe input (Mathew, Amudha & Sivakumari, 2020).   
NN zijn tegenwoordig immens populair en zijn in principe voor elk AI probleem toepasbaar. Door de populariteit van NN zijn er vele soorten en subcategorieën ontstaan. Daarnaast zijn de NN steeds groter en complexer geworden, dit heeft er uiteindelijk voor gezorgd dat deze netwerken een nieuwe naam kregen; *Deep-Learning* modellen (Aggarwal, 2018).

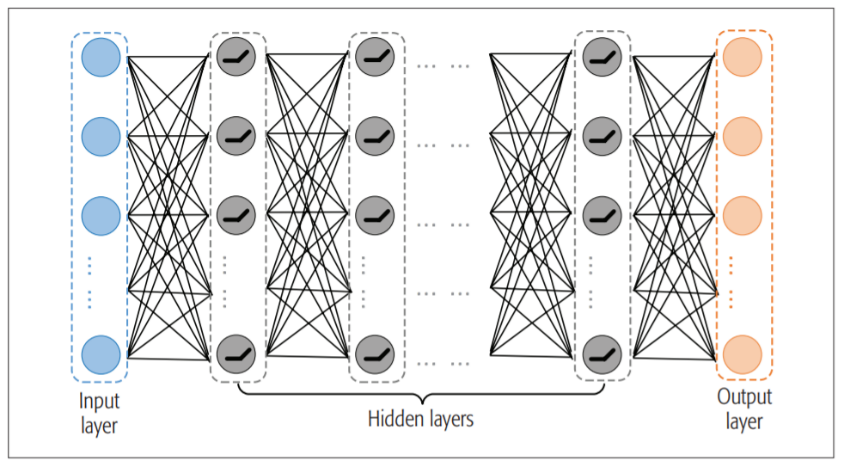
### Deep - Learning

*Deep-learning*-modellen zijn, in tegenstelling tot de traditionele modellen gebaat bij het gebruik van grote en complexere data architecturen. *Deep-learning* modellen worden, in tegenstelling tot de traditionele modellen beter naarmate er meer data gebruikt worden waarbij de prestatie afvlakt als het de maximale hoeveelheid trainings-data bereikt (Mathew, Amudha & Sivakmari, 2020).

*Deep-learning* modellen zijn hierdoor uiterst geschikt voor het maken van voorspellingen aan de hand van verkeersdata (Yin et all., 2021). Er is echter wel een aantal nadelen bij het gebruik van *Deep-learning* modellen. Ten eerste is er enorm veel data nodig om deze modellen te trainen en testen. Dit heeft als gevolg dat er ook veel rekenkracht nodig is om deze modellen te gebruiken. Het grootste nadeel van *Deep-learning* modellen is inherent aan de techniek die de basis vormt van *Deep-learning* modellen. Door iteratief te optimaliseren door het gebruik te maken van *gradient descent* bereikt een *Deep-neural network* uiteindelijk een lokaal optimum. *Deze Deep-neural network* modellen zijn door deze techniek slecht interpreteerbaar en erg complex. Het is vaak niet duidelijk wat een *Neural Network* daadwerkelijk heeft *‘geleerd’*. Om deze reden worden NN of *Deep-learning* modellen ook wel *black* *boxes* genoemd (Castelvecchi, 2016; Yin et all., 2021; Yuan & Li, 2021)



**Figuur 2: (Mathew, Amudha & Sivakumari, 2020, p. 601)**



**Figuur 3: Illustratie van een Deep-neural network   
(Liu, Li, Wu & Li, 2018, p. 42)**

*Deep-Learning* modellen zijn zeer accuraat en goed in het oplossen van bepaalde verkeersgerelateerde vraagstukken, maar door de complexiteit en slechte-interpreteerbaarheid is het belangrijk om na te gaan of het gebruik van simpelere *machine-learning* methodes niet beter aansluiten bij het oplossen van verkeersvraagstukken. De voordelen en nadelen van deep-learning modellen opgesomd:

|  |  |
| --- | --- |
| Deep-Learning | |
| Voordelen | **Nadelen** |
| * Correlaties en afhankelijken in de data worden automatisch verwerkt * Kunnen vrijwel voor elk probleem worden toegepast * Werkt zeer goed met complexe data * Behaald zeer goede prestaties | * Heeft veel data nodig om goede resultaten te behalen * Vergt veel rekenkracht * Complex om toe te passen * Slecht interpreteerbaar |

### Ensemble Methoden

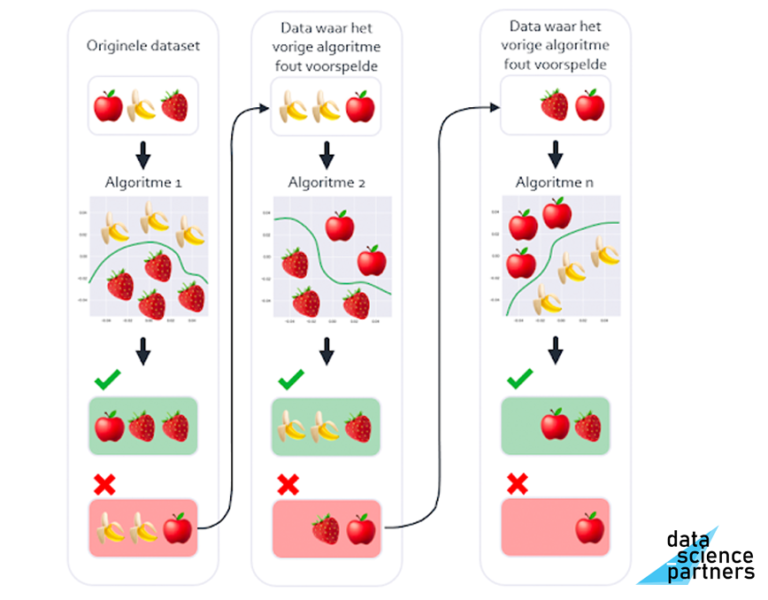
Bij *Ensemble Learning* wordt een ‘sterk’ model iteratief gebouwd met behulp van een verzameling (of "ensemble") van "zwakkere" modellen. Er zijn verschillende technieken die worden gebruikt bij ensemble methoden, deze zijn onder te verdelen in twee *frameworks:*

|  |  |
| --- | --- |
| Ensemble Learning | |
| Dependent framework | **Independent framework** |
| * Boosting | * Bagging * Stacking |

Bij ensemble methoden die gebruik maken van het *dependent framework* beïnvloedt de output van elk model eerder in de serie van modellen de constructie van het volgende model. De kennis die eerdere modellen hebben aangeleerd wordt overgedragen naar volgende modellen, deze techniek heet *Boosting*. Bij ensemble learning methoden die gebruik maken van het *independent framework* zijn alle modellen onafhankelijk van elkaar. Bij methoden uit het *independent framework* wordt er voornamelijk gebruik gemaakt van *Bagging* en *Stacking*. Bij Bagging wordt elk model onafhankelijk getraind met een *sub-sample* van de originele data. Vervolgens wordt er bij nieuw data door middel van een ‘stemming’ tussen de modellen beslist welke voorspelling het meest waarschijnlijk is. Bij *Stacking* worden alle onafhankelijk modellen gecombineerd. Eerst worden de modellen getraind met de beschikbare data, vervolgens worden deze modellen gecombineerd tot een enkel model. Deze wordt uiteindelijk gebruikt om voorspelling mee te maken op nieuwe data (Sagi & Rokach, 2018, Dong et all., 2020).   
Twee bekende en veelgebruikte ensemble methodes zijn Extreme Gradient Boosting (XGB) en Random Forrest (RF) (Sagi & Rokach, 2018).

Ensemble methoden hebben een aantal voordelen in tegenstelling tot enkele *machine-learning* en deep-learning modellen:

* Kans of *overfitting* en *underfitting* is klein.
* Computationeel voordeel
* Representatief
* Goed te interpreteren
* Functioneert zelfs met onevenwichtigheid in klassen
* Flexibel
* Werkt met een hoge dimensionaliteit

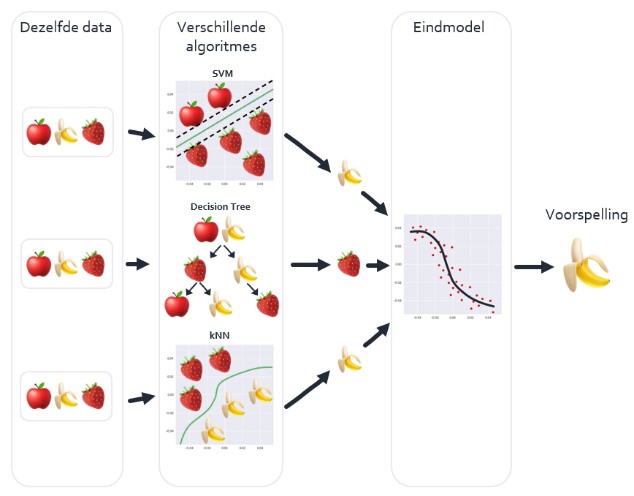
Ensemble learning methoden *overfitten* vrijwel nooit. Door het ‘gemiddelde’ te nemen van verschillende hypothesen vermindert het risico dat een onjuiste hypothese wordt gekozen en verbetert derhalve de algemene voorspellende prestatie. Daarnaast kunnen ensemble learning methoden door een groot aanbod aan parameters zeer complex worden. Door te doen aan *hyperparameter-tuning* en de juiste parameters te instantiëren is hierdoor de kans op *underfitting* ook zeer klein. Een individueel *machine-learning* model eindigt vaak tijdens het optimalisatie proces in een lokaal optimum. Door het combineren van verschillende *machine-learning* modellen, verminderen ensemble-methoden het risico op het bereiken van een lokaal minimum en verhoogt de kans op het vinden van het globale optimum, hierdoor behalen ze een computationeel voordeel. De optimale oplossing voor een probleem kan net buiten de scope liggen van een enkel model. Door het combineren van verschillende modellen wordt de scope vergroot en sluit het ensemble van modellen beter aan op de beschikbare data. Het ensemble model is hierdoor goed representatief voor de data.

**Figuur 4: Illustratie van de Boosting methoden  
(Tieleman & Data Science Partners, 2020)**

De ensemble methode werkt in tegenstelling tot veel andere *machine-learning* methoden goed op data met ongebalanceerde klassen. Bij data met ongelijke verdeling in klassen wordt bij het trainen van een *machine-learning* model vaak een bias aangeleerd voor een enkele klasse. De ensemble methode vermindert of vermijdt deze bias door gebruik te maken van gerandomiseerde *under-sampling techniques (*Rokach, 2005*).* Door gebruik te maken van *subsamples* die wel gebalanceerd zijn, worden de voorspellende prestaties van de ensemble methoden aanzienlijk verbeterd voor data met een onevenwichtigheid in klasse

**Figuur 5: Illustratie van de Bagging methode  
(Tieleman & Data Science Partners, 2020)**

De data die als input dient voor real-time *machine-learning* toepassingen heeft vaak de neiging om te veranderen. Zo kunnen nieuwe features en labels zijn toegevoegd of er ontstaat een verandering in de distributies van deze labels en features. Deze verandering heeft significante gevolgen voor de voorspellende werking van deze toepassing. Door de inherente diversiteit van ensemble modellen en door gebruik te maken van *dynamic weighted majority voting* is de ensemble methode erg flexibel en kan het gemakkelijk omgaan met deze verandering. Data met een hoge dimensionaliteit vormt voor de meeste traditionele *machine-learning* modellen een probleem. Een hoge dimensionaliteit resulteert vaak in een model dat slecht generaliseert, dit wordt ook wel de ‘vloek van de dimensionaliteit’ genoemd.

Ensemble learning methoden kunnen gebruikt worden om deze dimensionaliteit te verminderen of de meest essentiële dimensies te scheiden van de rest. Dit doet het door gebruik te maken van *attribute bagging* (AB). Dit is een techniek die de accuraatheid van verschillend featuresubsets met elkaar vergelijkt en hieruit de beste featuresubset selecteert(Sagi & Rokach, 2018; Dong et all., 2020).

**Figuur 6: Illustratie van de Stacking methoden.  
(Tieleman & Data Science Partners, 2020)**

De belangstelling voor ensemble methoden en het gebruik van ensemble methoden in onderzoeken is de afgelopen paar jaar significant toegenomen volgens een literatuur analyse uitgevoerd door Sagi & Rokach (2018). Ensemble learning methoden zijn flexibel, snel, betrouwbaar, werken met *ruwe* en complexe data, zijn in tegenstelling tot NN en met name *Deep-learning* modellen goed interpreteerbaar en produceren zeer nauwkeurige voorspellen (Liu, Gegov & Cocea, 2016; Sagi & Rokach, 2018; Dong et all., 2020).

Ensemble learning methoden en met name *XGB* zijn hierdoor erg populair op *Kaggle*; een data science platform waar *machine-learning* competities worden gehouden en waar dit model vaak de hoofdprijs pakt (Kaggle Inc., 2021).

## Gerelateerd onderzoek

In een vergelijkbaar onderzoek hebben Cuenca, Aliane, Puertas en Andres (2018) aan de hand van historische data over verkeersongelukken, betrokken voertuigen en mensen de ernst van verkeersongelukken getracht te voorspellen. Het betreft hier een binair voorspellingsprobleem waarbij voorspellingen zijn gemaakt voor de categorieën, geen ernstig ongeval en ernstig ongeval. Deze voorspellingen zijn gedaan aan de hand van categorische en numerieke data. Door een klassenongelijkheid in de gebruikte data is er gebruikt gemaakt van een *subsampling* techniek. In dit onderzoek zijn drie *machine learning* modellen met elkaar vergeleken. Een *Naive Bayes* (traditioneel), *Gradient Boosting Trees* (ensemble) en een *Deep-Learning* model. De statistieken accuraatheid (accuracy), precisie (precision) en *F1-measure* zijn gebruikt om de modellen te evalueren. Het *deep-learning* presteerde het beste met een accuracy van 0,87, precision van 0,84 en een F1-score van 0,87. Het *gradient boosting trees* model had vergelijkbare scores van 0,87, 0,85 en 0,87. Het naive bayes model presteerde minder goed met een score van 0,76, 0,74 en 0,77.

In een ander vergelijkbaar onderzoek hebben AlMamlook, Kwayu, Alkasisbeh en Frefer (2019) de ernst van verkeersongelukken voorspeld met behulp van vier modellen. AdaBoost (ensemble), Logistic Regression (traditioneel), Naive Bayes (traditioneel) en een Random Forrest(ensemble). In dit onderzoek is ook gebruik gemaakt van historische data omtrent verkeersongelukken. Deze data zijn zowel categorisch als numerieke van aard. In dit onderzoek zijn de modellen vergeleken aan de hand van de *accuracy, precision, recall, F1-measure* en *Area under the Receiver Operating Characteristic Curv*e (AUC). Alle modellen behaalden vergelijkbare resultaten met betrekking tot de *accuracy, precision, recall, F1-measure*. Vervolgens is gekeken naar de AUC om het best model te selecteren. Het Random Forrest model behaalde de hoogste AUC score van 75.5%.

Alkheder, S., Taamneh, M., & Taamneh, S. (2017) hebben onderzocht of de ernst van ongelukken kan worden voorspeld aan de hand van enkel categorische data. In dit onderzoek is een Neural Netwerk gebruikt om de ernst van ongelukken te classificeren in vier klassen; dodelijk, ernstig, matig en licht. Het betreft hier een meervoudig classificatie probleem. Ook in dit onderzoek was er sprake van een sterke klassenongelijkheid. Het Neurale Network is geëvalueerd aan de hand van de *accuracy* en AUC. Het model behaalde voor de klassen dodelijk ernstig, matig, lichtelijk en *accuracy* van 0%, 0%, 78.4% en 82% respectievelijk, de AUC score was slechts 0.697

# Methodologie

## Ongevallen en Partijen

Het BRON is een gestructureerde heterogene dataset en bestaat uit twee hoofbestanddelen en een aantal koppelbestandsdelen. De hoofdbestandsdelen zijn het bestand ongevallen en het bestand partijen. De attributen met beschrijvingen van deze bestanden zijn te vinden in de a ppendix, zie Tabel 1 en Tabel 2. De overige bestanden uit het BRON dienen als referentiemateriaal voor de attribuutwaardes in de twee hoofdbestanden. Alle bestanden in het BRON kunnen, aan de hand van sleutelattributen met elkaar worden verbonden. Het relatieschema van het BRON is te vinden in Figuur 8 (in de appendix).

Het bestand ongevallen bevat alle informatie omtrent de ongelukken zelf, zoals locatie, aantal betrokkenen, afloop van het ongeluk etc. Het bestand partijen bevat informatie over de partijen die betrokken waren bij de ongevallen. In dit onderzoek wordt er gebruik gemaakt van de openbare variant van het BRON. Een groot gedeelte van de openbare variant is afgeschermd wegens de AVG. Het BRON bevat attributen die verplicht moeten worden ingevuld en attributen die optioneel zijn. Daarnaast is er ook door de jaren heen een aantal attributen uit de bestanden ongevallen en partijen verwijderd en toegevoegd. Dit heeft erin geresulteerd dat beide bestanden niet consistent zijn ingevuld, waardoor er veel waardes ontbreken. De openbare variant van het BRON bestaat vrijwel compleet uit attributen die categorische van aard zijn, zowel nominaal als ordinaal. Het gaat hier om attributen van voertuigsoort (fiets, personenauto, vrachtauto, etc.)- tot aan zichtafstand(onbeperkt (200m >), beperkt ( 200m), ernstig beperkt ( 50m)). De numeriek attributen in het BRON zijn, op het attribuut ANTL\_PTJ na, allemaal afgeschermd door de AVG of bevatten geen wezenlijke informatie. De rest van de attributen bestaan uit categorische variabelen.

## Pre-Processen

Na het samenvoegen van de bestanden ongevallen en partijen ontstaat er een bestand met **1224079** rijen en **36** kolommen, zie Tabel 3 in de appendix. Vervolgens is er een selectie van de data gemaakt. Omdat vanaf 2017 de meeste attributen in de data consistent zijn ingevuld door de Rijkswaterstaat en politie, worden alleen de verkeersongelukken van 2017 tot heden gebruikt. Verder zijn alle lege kolommen verwijderd. Dit om zoveel mogelijk bruikbare en relevante data te behouden. Verder zijn alle attributen met naam *anders* (-AN) uit het bestand gehaald. Deze attributen omschrijven een verkeersongelukken attribuut aan de hand van een aantal woorden wanneer nog geen standaard attribuut code voor bestaat. Omdat deze *anders* attributen vrij weinig informatie bevatten zijn deze niet meegenomen in het onderzoek. Het gaat hier om een paar honderd tot paar duizend items.

Vervolgens zijn er aan de hand van het attribuut FK\_VELD5, items uit het bestand puntlocaties gekoppeld aan het hoofdbestand. Deze puntlocaties maken het mogelijk om buurten en wijknamen van het geografisch forum Publieke Dienstverlening Op de Kaart (PDOK) (2020) toe te voegen aan het hoofdbestand.   
Door het toevoegen van de buurt en wijknamen zijn de geografische coördinaten in het BRON gekoppeld aan een meer expliciete categorie dan de veelomvattende geografische attributen gemeente en provincie. Door de koppeling van puntlocaties aan buurten en wijken zijn een aantal geografische attributen (FK\_VELD5, X\_COORD\_RD, Y\_COORD\_RD en GEOMETRY) niet meer van essentieel belang. Deze attributen zijn vervolgens uit het bestand verwijderd. Door het verwijderen van attributen en het toevoegen van data is het bestand getransformeerd naar 566460 rijen en 21 kolommen   
(Tabel 4 in de appendix). De overgebleven geografische attributen, LONG en LAT zijn enkel gebruikt voor het visualiseren van verkeersongelukken op de kaart van Nederland. Vervolgens zijn deze attributen ook verwijderd uit het bestand.

## Feature Engineering

Het enige numeriek attribuut ANTL\_PTJ is getransformeerd naar een ordinaal categorische attribuut door gebruik te maken van een kwantiel discretisatiefunctie. Hierdoor zijn de verschillende numerieke variabelen van dit attribuut onderverdeeld in de categorieën Laag, Gemiddeld en Hoog. Zoals te zien in Tabel 4 is er nog een aantal missende waarden te vinden in de data. Door gebruik te maken van een *imputer* (Scikit-learn Developers, 2020) zijn alle missende waarden in de ordinaal categorische attributen en nominaal categorische attributen vervangen door het woord ‘onbekend’. Vervolgens is de data door een *encoder* getransformeerd. Hierdoor zijn alle nominale categorische variabelen via een *one-hot-encoder* omgezet naar een binair schema. Vervolgens zijn alle features met het woord ‘onbekend’ weer uit de data gefilterd, hierdoor worden alleen relevante features behouden. De ordinale categorische variabelen zijn omgezet via een ordinale-encoder waarbij de variabelen worden omgezet in een ordinale schaal [0, N-klassen]. Door het *encoderen* van alle variabelen in de data is het mogelijk om een data-analyse uit te voeren en een classificatie model toe te passen op de data. Na het *encoderen* van de data ontstaat een *sparse*-matrix van 566460 rijen en 15270 kolommen.

## Data-Analyse

In dit onderzoek zijn twee methoden gebruikt om categorische attribuut gewichten te berekenen. Dit zijn de Chi-kwadraattest en de *feature importance* methoden van *gradient boosting trees* (ensemble learning) model.  
De chi-kwadraattest is een niet-parametrische test die wordt gebruikt bij categorische variabelen Een chi-kwadraattoets wordt in de statistiek gebruikt om de onafhankelijkheid van twee verschijnselen te testen. In dit onderzoek wordt de chi-kwadraattest gebruikt om te meten welke categorische waarden er verantwoordelijk zijn voor de verschillen in klassen (McHugh, 2013).   
De andere methoden betreffen een ingebouwde functie binnen het model waar voorspellingen mee worden gedaan.   
Deze kan de *feature importance* score berekenen aan de hand van twee technieken:

* Loss Function Change
* Prediction Values Change

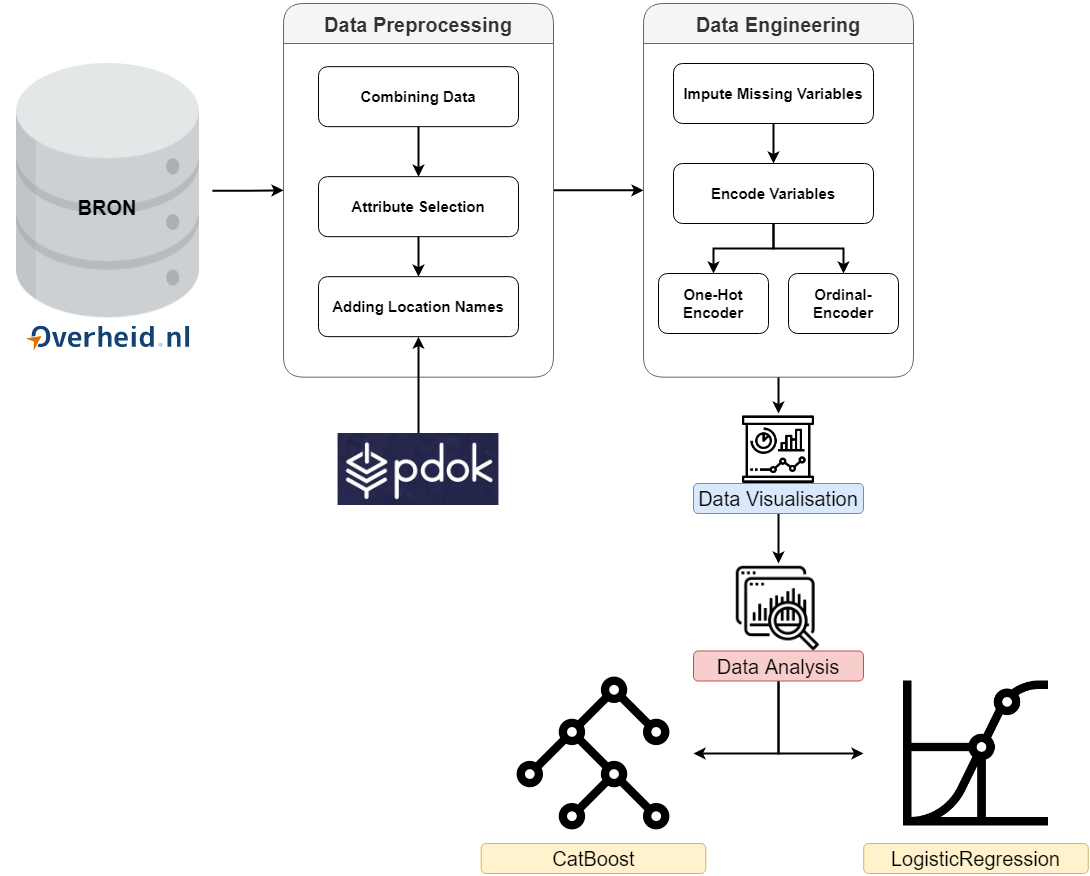
Bij *Loss Function Change* wordt het gewicht van een attribuut berekend op basis van het verschil tussen de *loss function* score van het model mét dit attribuut en zonder dit attribuut. De *Loss Function Change* kan alleen toegepast worden bij binaire classificatie problemen. De *Prediction Values Change* techniek geeft aan hoeveel de voorspelling gemiddeld verandert als de waarde van een attribuut verandert. Hoe groter de verandering in de voorspelling, hoe belangrijker het attribuut. In dit onderzoek wordt een CatboostClassifier (CC) toegepast op een multi-classificatie probleem. Om de *feature importance* te berekenen wordt hierdoor gebruik gemaakt van de *Prediction Values Change.* Door het uitvoeren van deze data-analyse wordt duidelijk welke attributen (factoren) de meest invloed hebben op het maken van verkeersongelukken én welke attributen het de meeste invloed hebben op de afloop van een verkeersongeluk.

## AI-modellen

In dit onderzoek is er gebruikt gemaakt van twee machine learning modellen. Een model dient als baseline en de ander als experimenteel model. Het baseline model is een ‘simpele’ logistieke regressie model. Dit model is van origine een binaire *classifier* maar kan met behulp van een one-vs-rest schema toch classificaties produceren voor meerdere klassen. Het experimenteel model maakt gebruik van de ensemble methode, het gaat hier specifiek om een *gradient boosting trees* model dat ontworpen is voor categorische data. Het model is valt onder de software CatBoost, een open-source softwarebibliotheek (CatBoost, 2017). Er is specifiek niet gekozen voor een Deep-Learning model, dit om de interpreteerbaarheid van het model en de betrouwbaarheid van het onderzoek te waarborgen. Beiden modellen berekenen waarschijnlijkheden van de drie afloop categorieën, waarbij de afloop met de hoogste waarschijnlijkheid zorgt voor de uiteindelijk classificering. De drie waarschijnlijkheden samen staan gelijk aan de totale waarschijnlijkheid van het maken van een verkeersongeluk. Beide modellen zijn geëvalueerd aan de hand van de accuracy, recall, precision en F1-score.

## CatBoost Classifier

Om de klasse ongelijkheid (Figuur 12) in de target data tegen te gaan wordt er gebruikt gemaakt van de klasse balans parameter van CC. Deze functie berekent de gewichten van de klasse op basis van de ratio’s van de afloop-klassen of het totaal aantal mogelijkheden items voor elke klasse. Deze gewichten worden gebruikt als vermenigvuldigingsfactor voor de features. De CC heeft een groot scala aan parameters die, als ze goed zijn afgesteld een zeer nauwkeurig model kunnen produceren. Om het model te optimaliseren is er gebruik gemaakt van hyperparameter tuning en cross-validatie. Hierdoor is het model geoptimaliseerd voor dit probleem. De CC construeert bij het trainen een *decision tree* waarbij de belangrijkste features bovenaan de boom komen te staan. Hierdoor wordt overzichtelijk welke categorische variabelen de meest impact hebben op de uiteindelijke waarschijnlijkheden van de drie afloop klassen uitsluitend materiele schade, letsel en dodelijk.



Figuur 7: Illustratie van de Methode

# Resultaten

## Data-analyse

Door het visualiseren van de puntlocaties in de data, via een density heatmap zijn er verschillende verkeersongelukken-hotspots (concentraties) zichtbaar in Nederland. Deze density heatmaps zijn onderverdeeld in de verkeersongeluk afloop categorieën:

* Uitsluitend Materiele Schade (Figuur 9)
* Letsel (Figuur 10)
* Dodelijk (Figuur 11)

Deze density heatmaps zijn te vinden in de appendix. Deze concentraties van verkeersongelukken zijn voor alle drie de categorieën voornamelijk te zien in en rond de Randstad. Verkeersongelukken waar letselschade de afloop was zijn voornamelijk te vinden in de steden; Amsterdam, Rotterdam, Den Haag en Utrecht. Gevaarlijke locaties waar doden zijn gevallen bevinden zich door heel Nederland, met hoge concentraties in de provincie Brabant en Zuid-Holland en Friesland.

## Chi-Kwadraattest

De resultaten van de Chi-Kwadraattest zijn te vinden in Figuur 13 en Tabel 5.  
De top vijf belangrijkste factoren bestaan uit WGD\_CODE\_1, OTE\_ID\_Boom, AOL\_ID\_Geparkeerd\_voertuig, OTE\_ID\_Scootmobiel en PVE\_NAAM\_Zuid-Holland met respectievelijke chi-kwadraat waarden 28891, 21058, 15836, 10841 en 10022 (p-waarden 0).

## Feature Importance Catboost

De feature gewichten die berekend zijn door het CatBoost-model zijn te zien in Figuur 14. De features AOL\_ID en BUURTNAAM zijn beide voor meer dan 50% verantwoordelijk voor de output van het model. De geconstrueerde *decision tree* is te vinden in de appendix Figuur 15

## Logistic Regression

Het Logistic Regression (LR) model met een *one-vs-rest* instelling behaalde een *accuracy* van 0.64. Voor de afloop dodelijk is een *precision* gehaald van 0.76, een *recall* van 0.65 en een F1-score van 0.70. Voor de categorie letsel zijn de respectievelijke scores van 0.58, 0.59 en 0.59 gehaald. Voor de categorie uitsluitend materiële schade gaat het om 0.60, 0,69, 0,64. Het classificatie rapport is te vinden in tabel 6.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| *Tabel 6: Classificatie Report: LogisticRegression* | | | | |
| LogisticRegression | | | | |
|  | **Precision** | **Recall** | **F1-Score** | **Support** |
| Dodelijk | 0,76 | 0,65 | 0,7 | 56646 |
| Letsel | 0,58 | 0,59 | 0,59 | 56646 |
| Uitsluitend materiele schade | 0,60 | 0,69 | 0,64 | 56646 |
| Accuracy | 0,64 | 0,64 | 0,64 |  |
| Macro Avg | 0,65 | 0,64 | 0,64 | 169938 |
| Weighted Avg | 0,65 | 0,64 | 0,64 | 169938 |

## Catboost Classifier

De CC met de multiclass instelling behaalde een accuracy van 0.82, voor de afloop ***dodelijk*** is een *precision* gehaald van 0.99, een *recall van 0.75 en een F1-score* van0.85*.* Voor de klasse ***letsel*** zijn de respectievelijke scores van 0.76, 0.86 en 0.81 gehaald. Voor de klasse ***uitsluitend materiële schade*** gaat het om 0.78, 0,87, 0,84. Alle evaluatie-scores zijn te vinden in tabel 7.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| *Tabel 7: Classificatie Report: Catboost Classifier* | | | | |
| Catboost Classifier | | | | |
|  | **Precision** | **Recall** | **F1-Score** | **Support** |
| Dodelijk | 0,99 | 0,75 | 0,85 | 56646 |
| Letsel | 0,76 | 0,86 | 0,81 | 56646 |
| Uitsluitend materiele schade | 0,78 | 0,87 | 0,84 | 56646 |
| Accuracy | 0,82 | 0,82 | 0,82 |  |
| Macro Avg | 0,84 | 0,82 | 0,83 | 169938 |
| Weighted Avg | 0,84 | 0,82 | 0,83 | 169938 |

# Discussie

In dit onderzoek is een data-analyse uitgevoerd en zijn doormiddel van een LR en CC model classificaties gedaan op historische verkeersongelukken data (BRON). Hier zijn enkel categorische variabelen voor gebruikt. Bij de data-analyse is er gebruikt gemaakt van een aantal technieken. Eerst zijn de verkeersongelukken gevisualiseerd op een kaart van Nederland. Vervolgens is de data onderworpen aan een data-analyse. Uit de data-analyse is gebleken dat een groot aantal factoren significant bijdragen aan het veroorzaken van verkeersongelukken. Vervolgens zijn het LR en CC model toegepast om waarschijnlijkheden van drie afloop klassen te berekenen. Het LR model diende als baseline en het CC als experiment. Uit de resultaten blijkt dat het LR model minder goed presteert dan het CC model. Dit was te verwachten omdat het LR model in vergelijking met het CC model minder complex is. Omdat het CC model complexer is dan het LR model heeft het de input data beter kunnen omvatten. Het is hierdoor ook niet opmerkelijk dat het CC betere resultaten heeft geproduceerd. De onderzoeken van Cuenca, Aliane, Puertas en Andres (2018) en AlMamlook, Kwayu, Alkasisbeh en Frefer (2019) corresponderen met deze resultaten. Het CC model maakt bij het trainen van het model een *decision tree* waarbij de verschillende variabelen en waarschijnlijkheden de vertakkingen vormen. Door de data-analyse en creëren van een classificatie model kunnen de risico’s van afloop categorieën worden berekend. Verder kan de totale kans van verkeersongelukken aan de hand van categorische variabelen worden berekend door de som van de afloop waarschijnlijkheden te nemen.

Dit huidige onderzoek is een aanvulling op bestaande onderzoeken over het classificeren van de afloop van verkeersongelukken. In Nederland zijn dit soort (publieke) onderzoeken nog redelijk beperkt, terwijl het belang van dit soort onderzoeken in de wetenschappelijk literatuur veelvuldig zijn bewezen. Er is in dit onderzoek uitsluitend gebruikt gemaakt van een select aantal categorische variabelen die beschikbaar zijn gesteld in het BRON. Het BRON bevat echter veel meer informatie over de ongelukken en betrokken partijen zelf. Deze informatie is door de invoering van de AVG voor een groot deel afgeschermd of compleet verwijderd uit het bestand. Het advies voor een vervolg onderzoek is dan ook om deze attributen te gebruiken voor een *machine learning* model. Attributen zoals tijd en datum, merkvoertuig en leeftijd voertuig en informatie over de betrokken partijen, zoals leeftijd bestuurder, rijbewijs categorieën, nationaliteit en medicijngebruik zouden kunnen helpen bij het verder specificeren, optimaliseren en accurater maken van het classificerend model.   
Wanneer tijd en datum variabelen worden toegevoegd aan het bestand wordt het ook mogelijk om regressie modellen toe te passen op de data. Hierdoor kan er geanalyseerd worden op trends en kunnen time-series modellen worden toegepast om voorspellingen mee te doen.

Een andere insteek is om data van andere bronnen toe te voegen aan het bestand. Denk hierbij aan data over het weer of statistieken die afhankelijk verbonden zijn aan een locatie. Dit alles om de *machine learning* modellen nog accurater en betrouwbaarder te maken zodat deze beter kunnen bijdragen aan de verkeersveiligheid van Nederland.

# Conclusie

Het doel van dit onderzoek was om te onderzoeken welke factoren invloed hebben op ongelukken in het algemeen en specifiek welke factoren de meeste invloed hebben op de ernst van ongelukken. Er is getracht een antwoord te vinden op de vraag: *Is het mogelijk om de verkeersveiligheid in Nederland te verbeteren door de afloop van ongelukken te voorspellen aan de hand van historische data?* Om een antwoord te formuleren is er kwantitatief onderzoek uitgevoerd op een verkeersongelukken dataset (*dimensies,* 566460 x 15270).

Er is een data-analyse toegepast op de dataset en er zijn vervolgens twee voorspellende modellen toegepast. Een controle model bestaand uit een LR model en een experimenteel model bestaande uit een CC. De CC valt onder de ensemble *gradient boosting trees* modellen en het LR model onder de traditionele classificatie modellen. Beide modellen berekenen waarschijnlijkheden voor de drie verschillende afloop categorieën, Uitsluitend Materiële Schade, Letsel en Dodelijk. De som van deze drie klassen samen staat gelijk aan het totale risico om een verkeersongeluk te maken. De waarschijnlijkheden zijn per afloop categorie berekend aan de hand van categorische variabelen en variabelen die gebonden zijn aan een puntlocaties.

Uit de data-analyse is duidelijk dat het grootste aantal ongelukken wordt veroorzaakt in en rond de grote steden. Wanneer er specifiek per afloop categorie wordt gekeken geldt ditzelfde patroon voor *uitsluitend materiële schade* en *letsel* gevallen. Voor een verkeersongeluk met dodelijke afloop is dit niet het geval. Hier zijn in de provincies Brabant en Friesland een aantal opmerkelijke concentratiepunten te vinden. Door het visualiseren van verkeersongelukken op de kaart van Nederland krijgen gemeenten en Rijkswaterstaat een inzicht waar de gevarenzones liggen van het veroorzaken van verkeersongelukken.

De resultaten van de *density heatmap* en de chi-kwadraattest geven een indicatie van de factoren en locaties die invloed hebben op het veroorzaken van verkeersongelukken. Dit is zowel op algemeen niveau als per afloop. Deze kennis stelt overheidsinstanties, gemeenten en branche organisaties in staat om nieuwe regels op te stellen en/of infrastructuur in het algemeen en per locatie beter en veiliger in te delen om zo de verkeersveiligheid op locatie te verhogen. Door gebruik te maken van een classificatie model kunnen deze nieuwe regels en/of infrastructuur getest worden. Door de waarschijnlijkheden te berekenen van de nieuwe situatie en oude situatie per locatie kunnen deze met elkaar vergeleken worden.

Verder kunnen de classificatie modellen, mits deze betrouwbaar zijn helpen bij het analyseren van de verkeersveiligheid van nieuwe verkeerssituaties. Door het berekenen van de afloop waarschijnlijkheden van nieuwe verkeersituaties, kan de verkeersveiligheid worden gemeten. De verkeersveiligheid van bijvoorbeeld nieuwe wegen of kruispunten kan per categorie of over het algemeen worden gemeten. Hierdoor weten beleidsmakers en wegbeheerders waar ze op moeten letten bij het construeren van verkeerswegen.

Concluderend valt te stellen dat het in zeker mate mogelijk is om de verkeersveiligheid te verbeteren in Nederland door de afloop van ongelukken te voorspellen aan de hand van historische data. Door het analyseren van de *feature importance* en het berekenen van de waarschijnlijkheden van een *machine learning* model kunnen verkeersscenario’s met verschillende factoren en locaties met elkaar worden vergeleken. Hierdoor moet duidelijk worden welke factoren het meest bijdragen aan een veilige en onveilige verkeersituatie. Door deze kennis kunnen beleidsmakers meer informatieve keuzes maken om de verkeersveiligheid in Nederland te verbeteren.

# Bibliografie

Aggarwal, C. C. (2018). Neural networks and deep learning. *Springer*, *10*, 978-3.

Alkheder, S., Taamneh, M., & Taamneh, S. (2017). Severity prediction of traffic accident using an artificial neural network. *Journal of Forecasting*, *36*(1), 100-108.

AlMamlook, R. E., Kwayu, K. M., Alkasisbeh, M. R., & Frefer, A. A. (2019, April). Comparison of machine learning algorithms for predicting traffic accident severity. In *2019 IEEE Jordan International Joint Conference on Electrical Engineering and Information Technology (JEEIT)* (pp. 272-276). IEEE.

Barceló, J., Kuwahara, M., & Miska, M. (2010). Traffic data collection and its standardization. *In Traffic Data Collection and its Standardization (pp. 1-10)*. Springer, New York, NY.

*CatBoost* (0.26). (2017). [Open-source software library]. Yandex. https://catboost.ai/

Castelvecchi, D. (2016). Can we open the black box of AI?*. Nature News, 538*(7623), 20*.*

Cuenca, L. G., Puertas, E., Aliane, N., & Andres, J. F. (2018, September). Traffic accidents classification and injury severity prediction. In *2018 3rd IEEE International Conference on Intelligent Transportation Engineering (ICITE)* (pp. 52-57). IEEE.

*Dataregister van de Nederlandse Overheid | Data overheid.* (z.d.). Overheid.nl. Geraadpleegd op 1 juni 2021, van <https://data.overheid.nl/>

Dong, X., Yu, Z., Cao, W., Shi, Y., & Ma, Q. (2020). A survey on ensemble learning. *Frontiers of Computer Science*, *14*(2), 241-258.

Ermagun, A., & Levinson, D. (2018). Spatiotemporal traffic forecasting: review and proposed directions. *Transport Reviews*, *38*(6), 786-814.

Fumagalli, E., Bose, D., Marquez, P., Rocco, L., Mirelman, A., Suhrcke, M., & Irvin, A. (2017). The high toll of traffic injuries: unacceptable and preventable. World Bank.

Ge, M., Bangui, H., & Buhnova, B. (2018). Big data for internet of things: a survey. *Future generation computer systems*, *87*, 601-614.

Kaggle Inc. (2021). *Kaggle: Your Machine Learning and Data Science Community.* Kaggle. <https://www.kaggle.com/>

Li, Y., & Shahabi, C. (2018). A brief overview of machine learning methods for short-term traffic forecasting and future directions. *SIGSPATIAL Special*, *10*(1), 3-9.

Liu, Z., Li, Z., Wu, K., & Li, M. (2018). Urban traffic prediction from mobility data using deep learning. *IEEE Network, 32*(4), 40-46.

Liu, H., Gegov, A., & Cocea, M. (2016). Ensemble learning approaches*. In Rule Based Systems for Big Data* (pp. 63-73). Springer, Cham.



McHugh, M. L. (2013). The chi-square test of independence. *Biochemia medica*, *23*(2), 143-149.

Ministerie van Infrastructuur en Waterstaat. (2018, december). Het strategisch Plan Verkeersveiligheid 2030: Veilig van deur tot deur. Rijksoverheid. <https://www.rijksoverheid.nl/documenten/rapporten/2018/12/05/bijlage-1-het-strategisch-plan-verkeersveiligheid-2030-veilig-van-deur-tot-deur>

Ministerie van Infrastructuur en Waterstaat. (2021, 20 april). Actuele verkeersongevallencijfers. Rijkswaterstaat. <https://www.rijkswaterstaat.nl/wegen/wegbeheer/onderzoek/verkeersveiligheid-en-ongevallencijfers/actuele-verkeersongevallencijfers>

Ministerie van Infrastructuur en Waterstaat (2019). Kennisinstituut voor Mobiliteitsbeleid KiM, 2019, 203 p., KiM-19-A12 - ISBN 978-90-8902-216-5

Ministerie van Economische Zaken en Klimaat. (2019, 31 mei)., Dutch Digital Delta., *Informatie- en communicatietechnologie (ICT)*., Rijksoverheid.nl., <https://www.rijksoverheid.nl/onderwerpen/ict/ict-en-economie/dutch-digital-delta>

Nationaal Dataportaal Wegverkeer (NDW). (2020, 21 juni). *NDW open data*. Nationaal Dataportaal Wegverkeer. <http://opendata.ndw.nu/>

Nederlandse Overheid. (2019). Dataregister van de Nederlandse Overheid | Data overheid. Data Overheid. <https://data.overheid.nl/>

Publieke Dienstverlening Op de Kaart. (2020). *Home - PDOK*. PDOK. Geraadpleegd op 16 juni 2021, van <https://www.pdok.nl/>

Rijkswaterstaat. (2020a, 7 mei). Actuele verkeersongevallencijfers. <https://www.rijkswaterstaat.nl/wegen/wegbeheer/onderzoek/verkeersveiligheid-en-ongevallencijfers/actuele-verkeersongevallencijfers>

Rijkswaterstaat. (2020b, 6 november). *Verkeersongevallen - Bestand geRegistreerde Ongevallen Nederland | data overheid* [Dataset]. Overheid.nl. <https://data.overheid.nl/dataset/9841-verkeersongevallen---bestand-geregistreerde-ongevallen-nederland>

Rijkswaterstaat - CIV. (2020, 18 augustus). *Handleiding product: Bestand geRegistreerde Ongevallen Nederland (BRON)* [Illustratie]. Rijkswaterstaat. <https://www.rijkswaterstaat.nl/apps/geoservices/geodata/dmc/bron/Documentatie/Handleiding%20product%20Bestand%20geRegistreerde%20Ongevallen%20Nederland.pdf>

Rokach, L. (2005). Ensemble methods for classifiers. In *Data mining and knowledge discovery handbook* (pp. 957-980). Springer, Boston, MA.

Sagi, O., & Rokach, L. (2018). Ensemble learning: A survey. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, *8*(4), e1249.

Schlögl, M., & Stütz, R. (2019). Methodological considerations with data uncertainty in road safety analysis. *Accident Analysis & Prevention*, *130*, 136-150.

Scikit-learn Developers. (2020). SimpleImputer. Scikit-Learn. <https://scikit-learn.org/stable/modules/generated/sklearn.impute.SimpleImputer.html>

Stichting Wetenschappelijk Onderzoek Verkeersveiligheid (SWOV). (2020a). Kosten van verkeersongevallen. SWOV-factsheet, maart 2020, SWOV Den Haag.

Stichting Wetenschappelijk Onderzoek Verkeersveiligheid (SWOV). (2020b). *De Staat van de Verkeersveiligheid 2020 Doelstellingen 2020 worden niet gehaald*. Geraadpleegd van: <https://www.swov.nl/publicatie/de-staat-van-de-verkeersveiligheid-2020>.

Wang, E. (2020, 21 april). *How AI, Machine Learning, Deep Learning and Data science intersect.* [Illustratie]. Turn it in. https://www.turnitin.com/blog/artificial-intelligence-and-machine-learning-at-turnitin

World Health Organization. (2020, 7 februari). Road traffic injuries. World Health Organization (WHO). <https://www.who.int/news-room/fact-sheets/detail/road-traffic-injuries>

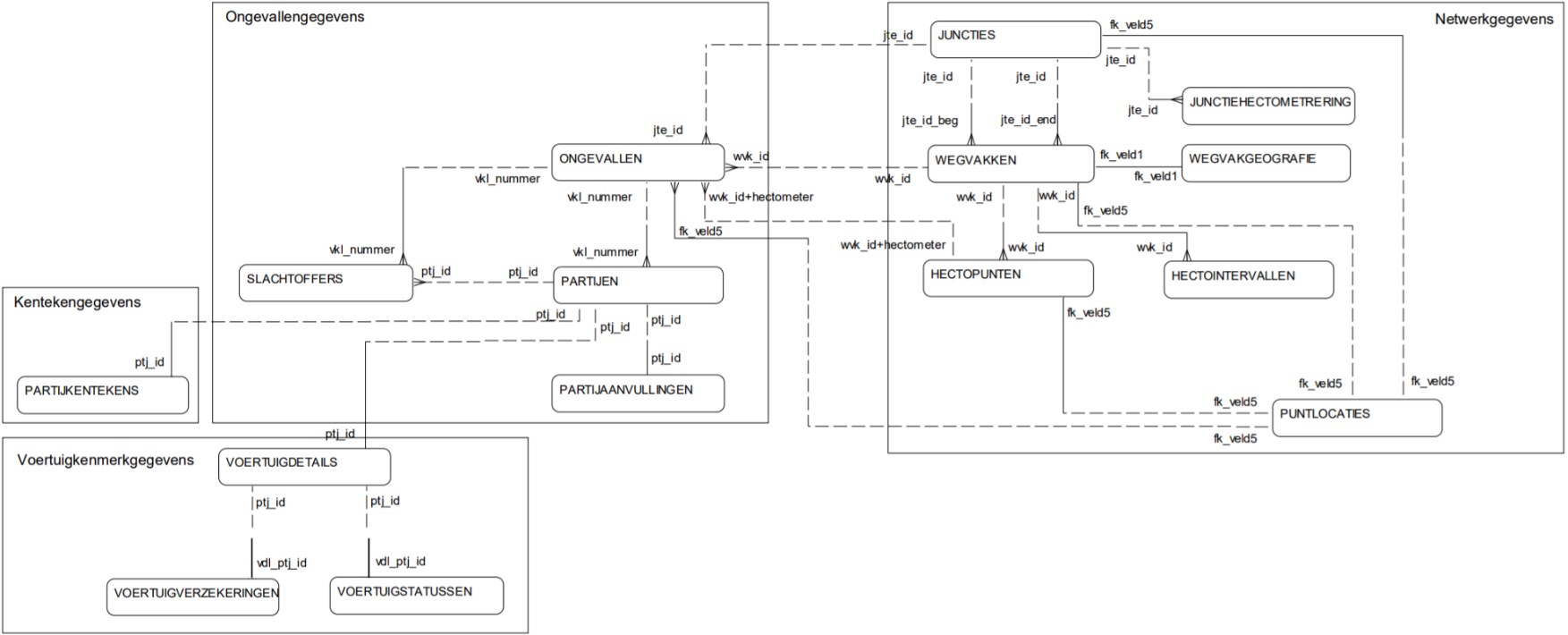
Yin, X., Wu, G., Wei, J., Shen, Y., Qi, H., & Yin, B. (2020). A comprehensive survey on traffic prediction. *arXiv preprint arXiv:2004.08555*.

Yin, X., Wu, G., Wei, J., Shen, Y., Qi, H., & Yin, B. (2021). Deep Learning on Traffic Prediction: Methods, Analysis and Future Directions. *IEEE Transactions on Intelligent Transportation Systems*.

Yuan, H., & Li, G. (2021). A Survey of Traffic Prediction: from Spatio-Temporal Data to Intelligent Transportation. Data Science and Engineering, 6(1), 63-85.

Yuan, Z., Zhou, X., & Yang, T. (2018, July). Hetero-convlstm: A deep learning approach to traffic accident prediction on heterogeneous spatio-temporal data. In *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining* (pp. 984-992).

# Appendix

****

Figuur 8: Relatieschema BRON (Rijkswaterstaat – CIV, 2020)

Tabel 1: Bestand Ongevallen

|  |  |  |  |
| --- | --- | --- | --- |
| **ONGEVALLEN** | | | |
| Naam | Formaat | **Verplicht** | Definitie |
| **Vkl\_nummer** | *Num(12)* | **Verplicht** | Unieke identificatie van het verkeersongeval |
| **Regnummer** | *Tekst(30)* | Optioneel | Registratienummer; intern nummer of code van een ver­keers­ongeval, dat dient als communicatiemiddel tussen politie en CIV |
| **Pvopgem** | *Tekst(1)* | Optioneel | Aanduiding of er een procesverbaal voor het verkeers­ongeval is opgemaakt:  J = Ja |
| **Datum\_vkl** | *Datum* | Optioneel | Datum waarop het verkeersongeval heeft plaatsgevonden  (In verband met de AVG is dit attribuut leeg, zie 2.2.5) |
| **Dag\_code** | *Tekst(2)* | Optioneel | Unieke identificatie van de dag van de week waarop het verkeersongeval heeft plaatsgevonden; verwijzing naar referentiebestand Dagen  (In verband met de AVG is dit attribuut nu leeg, zie 2.2.5) |
| **Mnd\_nummer** | *Num(2)* | Optioneel | Unieke identificatie van de maand waarin het verkeers­ongeval heeft plaatsgevonden; verwijzing naar referentie­bestand Maanden |
| **Jaar\_vkl** | *Num(4)* | **Verplicht** | Jaar waarin het verkeersongeval heeft plaatsgevonden |
| **Tijdstip** | *Tekst(4)* | Optioneel | Tijdstip, uitgedrukt in uren en minuten, waarop het verkeers­ongeval heeft plaatsgevonden  (In verband met de AVG is dit attribuut leeg, zie 2.2.5) |
| **Uur** | *Tekst(2)* | Optioneel | Uur waarbinnen het verkeersongeval heeft plaatsgevonden |
| **Ddl\_id** | *Num(1)* | Optioneel | Unieke identificatie van het dagdeel waarbinnen het verkeers­ongeval heeft plaatsgevonden; verwijzing naar referentie­bestand Dagdelen  (In verband met de AVG is dit attribuut leeg, zie 2.2.5) |
| **Ap3\_code** | *Tekst(3)* | **Verplicht** | Unieke identificatie van de afloop van het verkeersongeval, onderverdeeld in 3 categorieën; verwijzing naar referentie­bestand Aflopen3 |
| **Ap4\_code** | *Tekst(3)* | **Verplicht** | Unieke identificatie van de afloop van het verkeersongeval, onderverdeeld in 4 categorieën; verwijzing naar referentiebestand Aflopen4  (In verband met de AVG is dit attribuut leeg, zie 2.2.5) |
| **Ap5\_code** | *Tekst(3)* | **Verplicht** | Unieke identificatie van de afloop van het verkeersongeval, onderverdeeld in 5 categorieën; verwijzing naar referentiebestand Aflopen5  (In verband met de AVG is dit attribuut leeg, zie 2.2.5) |
| **Antl\_sla** | *Num(3)* | **Verplicht** | Het totaal aantal slachtoffers bij het verkeersongeval  (In verband met de AVG is dit attribuut leeg, zie 2.2.5) |
| **Antl\_dod** | *Num(3)* | **Verplicht** | Het aantal dodelijke slachtoffers bij het verkeersongeval  (In verband met de AVG is dit attribuut leeg, zie 2.2.5) |
| **Antl\_gzh** | *Num(3)* | **Verplicht** | Het aantal gewonde slachtoffers bij het verkeersongeval, die in het ziekenhuis opgenomen zijn  (In verband met de AVG is dit attribuut leeg, zie 2.2.5) |
| **Antl\_seh** | *Num(3)* | **Verplicht** | Het aantal slachtoffers bij het verkeersongeval, die spoedeisende hulp toegediend kregen  (In verband met de AVG is dit attribuut leeg, zie 2.2.5) |
| **Antl\_gov** | *Num(3)* | **Verplicht** | Het aantal overige gewonde slachtoffers bij het verkeersongeval  (In verband met de AVG is dit attribuut leeg, zie 2.2.5) |
| **Antl\_ptj** | *Num(3)* | **Verplicht** | Het aantal partijen betrokken bij het verkeersongeval |
| **Antl\_tdt** | *Num(2)* | **Verplicht** | Het aantal toedrachten vastgelegd bij de vermoedelijke veroorzaker van het verkeersongeval  (In verband met de AVG is dit attribuut leeg, zie 2.2.5) |
| **Mne\_code** | *Tekst(3)* | **Verplicht** | Unieke identificatie van de manoeuvre toegekend aan het verkeersongeval; verwijzing naar referentiebestand Manoeuvres  (In verband met de AVG is dit attribuut leeg, zie 2.2.5) |
| **Aol\_id** | *Num(2)* | **Verplicht** | Unieke identificatie van de aard ongeval toegekend aan het verkeersongeval; verwijzing naar referentiebestand AardOngevallen |
| **Niveaukop** | *Tekst(1)* | **Verplicht** | Aanduiding op welk niveau het verkeersongeval gekoppeld is aan het BN:  E = Ongeval exact gekoppeld aan BN  K = Ongeval gekoppeld op kruispuntniveau  S = Ongeval gekoppeld op straat niveau  G = Ongeval gekoppeld op gemeente niveau |
| **Wse\_id** | *Num(2)* | Optioneel | Unieke identificatie van de wegsituatie op de plaats van het verkeersongeval; verwijzing naar referentiebestand Wegsituaties |
| **Wse\_an** | *Tekst(30)* | Optioneel | Wegsituatie anders; een wegsituatie op de plaats van het verkeersongeval, die niet als keuze vermeld staat |
| **Bebkom** | *Tekst(2)* | Optioneel | Aanduiding of het verkeersongeval binnen of buiten de bebouwde kom plaatsvond:  BI = Binnen  BU = Buiten |
| **Maxsnelhd** | *Num(3)* | Optioneel | De aangegeven maximum snelheid op de plaats van het verkeersongeval:  15 = 15 km/u (stapvoets)  30 = 30 km/u  50 = 50 km/u  60 = 60 km/u  70 = 70 km/u  80 = 80 km/u  90 = 90 km/u  100 = 100 km/u  120 = 120 km/u  130 = 130 km/u  Voor ongevallen die hebben plaatsgevonden op wegen met beheerder ‘Rijk’ en waar de Politie geen maximum snelheid heeft ingevuld’, zijn de maximum snelheden aangevuld met snelheden uit WEGGEG. |
| **Wvl\_id** | *Num(2)* | Optioneel | Unieke identificatie van de wegverlichting op de plaats en ten tijde van het verkeersongeval; verwijzing naar referentie­bestand Wegverlichtingen |
| **Wvg\_id** | *Num(2)* | Optioneel | Unieke identificatie van de wegverharding op de plaats van het verkeersongeval; verwijzing naar referentiebestand Wegverhardingen |
| **Wvg\_an** | *Tekst(30)* | Optioneel | Wegverharding anders; een wegverharding op de plaats van het verkeersongeval, die niet als keuze vermeld staat |
| **Wdk\_id** | *Num(2)* | Optioneel | Unieke identificatie van de toestand van het wegdek op de plaats en ten tijde van het verkeersongeval; verwijzing naar referentiebestand Wegdekken |
| **Wdk\_an** | *Tekst(30)* | Optioneel | Wegdek anders; een toestand van het wegdek op de plaats en ten tijde van het verkeersongeval, die niet als keuze vermeld staat |
| **Lgd\_id** | *Num(1)* | Optioneel | Unieke identificatie van de lichtgesteldheid op de plaats en ten tijde van het verkeersongeval; verwijzing naar referentiebestand Lichtgesteldheden |
| **Zad\_id** | *Num(2)* | Optioneel | Unieke identificatie van de zichtafstand op de plaats en ten tijde van het verkeersongeval; verwijzing naar referentie­bestand Zichtafstanden |
| **Wgd\_code\_1** | *Tekst(1)* | Optioneel | De eerste code van de weersgesteldheid die van toepassing is op de plaats en ten tijde van het verkeersongeval, volgens de prioritering  D = Droog  R = Regen  M = Mist  S = Sneeuw/Hagel  H = Harde windstoten  O = Onbekend |
| **Wgd\_code\_2** | *Tekst(1)* | Optioneel | De tweede code van de weersgesteldheid die van toepassing is op de plaats en ten tijde van het verkeersongeval, volgens de prioritering  D = Droog  R = Regen  M = Mist  S = Sneeuw/Hagel  H = Harde windstoten  O = Onbekend |
| **Bzd\_id\_vm1** | *Num(3)* | Optioneel | Unieke identificatie van de bijzonderheid ter plaatse van het verkeersongeval, van het type verkeersmaatregel, met het laagste id; verwijzing naar referentiebestand Bijzonder­heden |
| **Bzd\_id\_vm2** | *Num(3)* | Optioneel | Unieke identificatie van de bijzonderheid ter plaatse van het verkeersongeval, van het type verkeersmaatregel, met het op 1 na laagste id; verwijzing naar referentiebestand Bijzonderheden |
| **Bzd\_id\_vm3** | *Num(3)* | Optioneel | Unieke identificatie van de bijzonderheid ter plaatse van het verkeersongeval, van het type verkeersmaatregel, met het op 2 na laagste id; verwijzing naar referentie­bestand Bijzonderheden |
| **Bzd\_vm\_an** | *Tekst(30)* | Optioneel | Bijzonderheid verkeersmaatregel anders; een bijzonderheid inzake verkeersmaatregelen ter plaatse van het verkeers­ongeval, die niet als keuze vermeld staat |
| **Bzd\_id\_if1** | *Num(3)* | Optioneel | Unieke identificatie van de bijzonderheid ter plaatse van het verkeersongeval, van het type infrastructuur, met het laagste id; verwijzing naar referentiebestand Bijzonder­heden |
| **Bzd\_id\_if2** | *Num(3)* | Optioneel | Unieke identificatie van de bijzonderheid ter plaatse van het verkeersongeval, van het type infrastructuur, met het op 1 na laagste id; verwijzing naar referentiebestand Bijzonder­heden |
| **Bzd\_id\_if3** | *Num(3)* | Optioneel | Unieke identificatie van de bijzonderheid ter plaatse van het verkeersongeval, van het type infrastructuur, met het op 2 na laagste id; verwijzing naar referentiebestand Bijzonder­heden |
| **Bzd\_if\_an** | *Tekst(30)* | Optioneel | Bijzonderheid infrastructuur anders; een bijzonderheid inzake infrastructuur ter plaatse van het verkeersongeval, die niet als keuze vermeld staat |
| **Bzd\_id\_ta1** | *Num(3)* | Optioneel | Unieke identificatie van de bijzonderheid ter plaatse van het verkeersongeval, van het type tijdelijke aard, met het laagste id; verwijzing naar referentiebestand Bijzonder­heden |
| **Bzd\_id\_ta2** | *Num(3)* | Optioneel | Unieke identificatie van de bijzonderheid ter plaatse van het verkeersongeval, van het type tijdelijke aard, met het op 1 na laagste id; verwijzing naar referentiebestand Bijzonder­heden |
| **Bzd\_id\_ta3** | *Num(3)* | Optioneel | Unieke identificatie van de bijzonderheid ter plaatse van het verkeersongeval, van het type tijdelijke aard, met het op 2 na laagste id; verwijzing naar referentiebestand Bijzonder­heden |
| **Bzd\_ta\_an** | *Tekst(30)* | Optioneel | Bijzonderheid tijdelijke aard anders; een bijzonderheid van tijdelijke aard, ter plaatse van het verkeersongeval, die niet als keuze vermeld staat |
| **Jte\_id** | *Num(10)* | Optioneel | Unieke identificatie van de junctie, waarop het verkeers­ongeval heeft plaatsgevonden; is alleen ingevuld bij kruispunt-ongevallen |
| **Wvk\_id** | *Num(10)* | Optioneel | Unieke identificatie van het wegvak, waarop het verkeers­ongeval heeft plaatsgevonden; is alleen ingevuld bij wegvak-ongevallen |
| **Hectometer** | *Num(4)* | Optioneel | Het getal dat op het hectometerbord staat, waarbij het verkeersongeval heeft plaatsgevonden; is alleen ingevuld bij wegvak-ongevallen op gehectometreerde wegen |
| **FK\_veld5** | *Tekst(17)* | **Verplicht** | Relatieveld voor koppeling met bestand Puntlocaties:  afhankelijk van of alleen junctie-id, resp. alleen wegvak\_id, dan wel wegvak\_id en hectometrering ingevuld is:  ‘JTE’+jte\_id, of  ‘WVK’+wvk\_id, of  ‘HTT’+wvk\_id+hectometer |
| **Huisnummer** | *Tekst(10)* | Optioneel | Het huisnummer ter hoogte van de locatie en ten tijde van het verkeers­ongeval, zoals opgegeven door de politie; is alleen ingevuld bij wegvak-ongevallen, waarvoor geen hecto­metrering is vastgelegd |
| **Gme\_id** | *Num(4)* | **Verplicht** | Unieke identificatie van de gemeente waarin conform de in het BN actuele gemeentegrenzen het verkeersongeval heeft plaatsgevonden |
| **Gme\_naam** | *Tekst(24)* | **Verplicht** | Naam van de gemeente waarin conform de in het BN actuele gemeentegrenzen het verkeersongeval heeft plaatsgevonden |
| **Pve\_code** | *Tekst(2)* | **Verplicht** | Unieke identificatie van de provincie waarin conform de in het BN actuele gemeente-indeling en -grenzen het verkeersongeval heeft plaatsgevonden |
| **Pve\_naam** | *Tekst(24)* | **Verplicht** | Naam van de provincie waarin conform de in het BN actuele gemeente-indeling en -grenzen het verkeersongeval heeft plaatsgevonden |
| **Kdd\_naam** | *Tekst(40)* | Optioneel | Naam van het kaderwetgebied (zone met zonesoort 04 en zonecode beginnend met KW) waarin conform de in het BN actuele gemeente-indeling en -grenzen het verkeersongeval heeft plaatsgevonden |
| **Plt\_naam** | *Tekst(40)* | Optioneel | Naam van het politiedistrict (zone met zonesoort 17 of 28) waarin conform de in het BN actuele polygoongrenzen het verkeers­ongeval heeft plaatsgevonden; is alleen ingevuld indien een zone met zonesoort 07 of 08 als selectie­parameter is ingegeven |
| **Dienstcode** | *Tekst(40)* | Optioneel | Code van de regio binnen de regio waarin conform de in het BN actuele polygoongrenzen het verkeersongeval heeft plaatsgevonden |
| **Dienstnaam** | *Tekst(40)* | Optioneel | Naam van regio waarin conform de in het BN actuele polygoongrenzen het verkeers­ongeval heeft plaatsgevonden |
| **Distrcode** | *Tekst(40)* | Optioneel | Code van het disctict binnen de regio waarin conform de in het BN actuele polygoongrenzen het verkeersongeval heeft plaatsgevonden |
| **Distrnaam** | *Tekst(40)* | Optioneel | Naam van het district waarin conform de in het BN actuele polygoongrenzen het verkeersongeval heeft plaatsgevonden |
| **Dagtype** | *Tekst(5)* | Optioneel | Geeft aan in welk deel van de week het verkeersongeval heeft plaatsgevonden  MA-VR = Maandag t/m Vrijdag  ZA-ZO = Zaterdag t/m Zondag |
| **Ind\_alc** | *Tekst(1)* | Optioneel | Indicator die aangeeft of er binnen het ongeval sprake is geweest van alcoholgebruik.  J = Ja  N = Nee |
| **Weeknr** | *Num(2)* | **Verplicht** | Het weeknummer waarbinnen het ongeval in dat jaar heeft plaatsgevonden |

Tabel 2: Bestand Partijen

|  |  |  |  |
| --- | --- | --- | --- |
| **PARTIJEN** | | | |
| Naam | Formaat | **Verplicht** | Definitie |
| **Ptj\_id** | *Num(12)* | **Verplicht** | Unieke identificatie van de partij |
| **Vkl\_nummer** | *Num(12)* | **Verplicht** | Unieke identificatie van het verkeersongeval, waarbij de partij betrokken is |
| **Nummer** | *Num(3)* | **Verplicht** | Een aaneensluitend volgnummer voor partijen binnen een verkeersongeval |
| **Doorrijder** | *Tekst(1)* | Optioneel | Aanduiding of het voertuig doorgereden is na het verkeers­ongeval:  J = Ja |
| **Ote\_id** | *Num(2)* | Optioneel | Unieke identificatie van het objecttype; verwijzing naar referentiebestand Objecttypes |
| **Ote\_an** | *Tekst(40)* | Optioneel | Objecttype anders; een objecttype, dat niet als keuze vermeld staat |
| **Ntt\_code\_v** | *Tekst(3)* | Optioneel | Unieke identificatie van de nationaliteit van het voertuig; verwijzing naar referentiebestand Nationaliteiten  (In verband met de AVG is dit attribuut leeg, zie 2.2.5) |
| **Vtgverz** | *Tekst(1)* | Optioneel | Aanduiding of het voertuig verzekerd was ten tijde van het verkeers­ongeval:  J = Ja  N = Nee |
| **Schade** | *Tekst(1)* | Optioneel | Aanduiding of er schade aan het voertuig is:  J = Ja  N = Nee |
| **Getraanh** | *Tekst(1)* | Optioneel | Aanduiding of het voertuig een aanhangwagen trok:  J = Ja |
| **Gevstof** | *Tekst(1)* | Optioneel | Aanduiding of het voertuig een niet afgedekte gevaarlijke stoffen plaat voerde:  J = Ja |
| **Vtgverl** | *Tekst(2)* | **Verplicht** | Aanduiding of bij het voertuig verlichting aanwezig was en zo ja, of deze wel of niet brandde:  NB = Niet brandend  BR = Brandend  NA = Niet aanwezig/nvt  XX = Onbekend |
| **Antl\_pas** | *Num(3)* | Optioneel | Aantal passagiers; het aantal personen, die tijdens het verkeersongeval in het betrokken voertuig aanwezig waren |
| **Gebdat** | *Datum* | Optioneel | De kalenderdatum waarop de bestuurder of voetganger is geboren  (In verband met de AVG is dit attribuut leeg, zie 2.2.5) |
| **Gebjaar** | *Num(4)* | Optioneel | Het kalenderjaar waarin de bestuurder of voetganger is geboren, in het formaat eejj  (In verband met de AVG is dit attribuut leeg, zie 2.2.5) |
| **Leeftijd** | *Num(3)* | Optioneel | De leeftijd van de bestuurder of voetganger op de dag van het verkeersongeval  (In verband met de AVG is dit attribuut leeg, zie 2.2.5) |
| **Lke\_id** | *Num(2)* | Optioneel | Unieke identificatie van de leeftijdsklasse waartoe de leeftijd van de bestuurder (of voetganger) behoort; verwijzing naar referentiebestand Leeftijdsklassen |
| **Ntt\_code\_b** | *Tekst(3)* | Optioneel | Unieke identificatie van de nationaliteit van de bestuurder (of voetganger); verwijzing naar referentiebestand Nationaliteiten  (In verband met de AVG is dit attribuut leeg, zie 2.2.5) |
| **Geslacht** | *Tekst(1)* | Optioneel | Het geslacht van de bestuurder of voetganger:  M = Mannelijk  V = Vrouwelijk |
| **Blaastest** | *Tekst(1)* | Optioneel | Aanduiding of er een alcohol blaastest is afgenomen ten tijde van het verkeersongeval:  J = Ja, afgenomen  N = Nee, niet afgenomen  (In verband met de AVG is dit attribuut leeg, zie 2.2.5) |
| **Art8** | *Tekst(1)* | Optioneel | Aanduiding of er artikel 8 is geconstateerd:  J = Ja, geconstateerd  W = Artikel 8 niet geconstateerd, wel alcohol  G = Geen alcohol  (In verband met de AVG is dit attribuut leeg, zie 2.2.5) |
| **Medicgebr** | *Tekst(1)* | Optioneel | Aanduiding of sprake is van drugs- en/of medicijngebruik:  J = Ja  (In verband met de AVG is dit attribuut leeg, zie 2.2.5) |
| **Rijbewgel** | *Tekst(1)* | Optioneel | Aanduiding of de bestuurder in het bezit is van een geldig rijbewijs:  J = Ja |
| **Rijbewcat** | *Tekst(20)* | Optioneel | Een opsomming in alfabetische volgorde van de codes van de categorieën van rijbevoegdheden, welke in het bezit zijn van de voertuigbestuurder; codes worden onderling gescheiden door een komma (,); de volgende codes worden gehanteerd:  A1 = Rijbewijscategorie A1  A2 = Rijbewijscategorie A2  B = Rijbewijscategorie B  C = Rijbewijscategorie C  D = Rijbewijscategorie D  E = Rijbewijscategorie E  F = Militair Rijbewijs  X = Geen, n.v.t. |
| **Rijbewbeg** | *Tekst(1)* | Optioneel | Aanduiding of de bestuurder in het bezit is van een beginnersrijbewijs:  J = Ja  N = Nee |
| **Bromfcert** | *Tekst(1)* | Optioneel | Aanduiding of de bestuurder in het bezit is van een bromfietscertificaat:  J = Ja  N = Nee |
| **Uitgpos1** | *Num(3)* | Optioneel | De 1e vastgelegde plaats waar het voertuig zich bevond vlak voor het verkeersongeval:  1 = Rijbaan  2 = Fietspad/fietsstrook  3 = Trottoir/berm  4 = Vluchtheuvel/middenberm  5 = Inrit/uitrit  6 = Vluchtstrook  7 = Parkeervoorziening  8 = Tram-/busbaan |
| **Uitgpos2** | *Num(3)* | Optioneel | De 2e vastgelegde plaats waar het voertuig zich bevond vlak voor het verkeersongeval:  1 = Rijbaan  2 = Fietspad/fietsstrook  3 = Trottoir/berm  4 = Vluchtheuvel/middenberm  5 = Inrit/uitrit  6 = Vluchtstrook  7 = Parkeervoorziening  8 = Tram-/busbaan |
| **Uitgpos\_an** | *Tekst(30)* | Optioneel | Uitgangspositie anders; een uitgangspositie, die niet als keuze vermeld staat |
| **Voorgbew** | *Num(3)* | Optioneel | De voorgenomen beweging van de bestuurder voor het verkeersongeval:  1 = Oversteken  2 = Vooruit  3 = Links rijstrook wisselen  4 = Stilstand  5 = Rechts rijstrook wisselen  6 = Linksaf  7 = Links omkeren  8 = Achteruit  9 = Rechts omkeren  10 = Rechtsaf  11 = Parkeerstand |
| **Agt\_type** | *Tekst(1)* | Optioneel | Aanduiding waar het aangrijppunt betrekking op heeft:  V = Voertuig  A = Aanhangwagen |
| **Agt\_id\_1** | *Num(3)* | Optioneel | Unieke identificatie van het aangrijppunt met het laagste volgordenummer, ter aanduiding van de plaats(en) waar de partij tijdens het verkeersongeval het eerst werd geraakt; verwijzing naar referentiebestand Aangrijppunten |
| **Agt\_id\_2** | *Num(3)* | Optioneel | Unieke identificatie van het aangrijppunt met het op 1 na laagste volgordenummer, ter aanduiding van de plaats(en) waar de partij tijdens het verkeersongeval het eerst werd geraakt; verwijzing naar referentiebestand Aangrijppunten |
| **Bwg\_id\_1** | *Num(2)* | Optioneel | Unieke identificatie van de beweging met het laagste id, ter aanduiding van de beweging(en) van de partij, ten gevolge van het verkeersongeval en direct na het verkeersongeval; verwijzing naar referentiebestand Bewegingen |
| **Bwg\_id\_2** | *Num(2)* | Optioneel | Unieke identificatie van de beweging met het op 1 na laag­ste id, ter aanduiding van de beweging(en) van de partij, ten gevolge van het verkeersongeval en direct na het verkeers­ongeval; verwijzing naar referentiebestand Bewegingen |
| **Bwg\_an** | *Tekst(30)* | Optioneel | Beweging anders; een beweging na ongeval, die niet als keuze vermeld staat |
| **Tdt\_id\_1** | *Num(3)* | Optioneel | Unieke identificatie van de toedracht met het laagste id, ter aanduiding van de omstandigheden van de vermoedelijke veroorzaker, welke hebben geleid tot het verkeers­ongeval; verwijzing naar referentiebestand Toedrachten  (In verband met de AVG is dit attribuut leeg, zie 2.2.5) |
| **Tdt\_id\_2** | *Num(3)* | Optioneel | Unieke identificatie van de toedracht met het op 1 na laagste id, ter aanduiding van de omstandigheden van de vermoedelijke veroorzaker, welke hebben geleid tot het verkeers­ongeval; verwijzing naar referentiebestand Toedrachten  (In verband met de AVG is dit attribuut leeg, zie 2.2.5) |
| **Tdt\_id\_3** | *Num(3)* | Optioneel | Unieke identificatie van de toedracht met het op 2 na laagste id, ter aanduiding van de omstandigheden van de vermoedelijke veroorzaker, welke hebben geleid tot het verkeers­ongeval; verwijzing naar referentiebestand Toedrachten  (In verband met de AVG is dit attribuut leeg, zie 2.2.5) |
| Tdt\_an | *Tekst(30)* | Optioneel | Toedracht anders; een toedracht, die niet als keuze vermeld staat |

Tabel 3: Partijen, Ongevallen samengevoegd

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Indicatie van de data** | | | | |
| **#** | **Columns** | **Non-Null** |  | **Dtype** |
| **0** | PTJ\_ID (key) | 1224079 | non-null | int64 |
| **1** | VKL\_NUMMER (key) | 1224079 | non-null | int64 |
| **2** | JAAR\_VKL | 1224079 | non-null | int64 |
| **3** | AP3\_CODE | 1224079 | non-null | object |
| **4** | ANTL\_PTJ | 1224079 | non-null | int64 |
| **5** | AOL\_ID | 1224079 | non-null | int64 |
| **6** | NIVEAUKOP | 1224079 | non-null | object |
| **7** | WSE\_ID | 1088378 | non-null | float64 |
| **8** | WSE\_AN | 300 | non-null | object |
| **9** | BEBKOM | 927439 | non-null | object |
| **10** | MAXSNELHD | 1224079 | non-null | object |
| **11** | WVL\_ID | 1074245 | non-null | float64 |
| **12** | WVG\_ID | 1063267 | non-null | float64 |
| **13** | WVG\_AN | 5207 | non-null | object |
| **14** | WDK\_ID | 1064466 | non-null | float64 |
| **15** | WDK\_AN | 2892 | non-null | object |
| **16** | LGD\_ID | 815109 | non-null | float64 |
| **17** | ZAD\_ID | 29 | non-null | float64 |
| **18** | WGD\_CODE\_1 | 1051296 | non-null | object |
| **19** | WGD\_CODE\_2 | 793 | non-null | object |
| **20** | BZD\_ID\_VM1 | 115087 | non-null | float64 |
| **21** | BZD\_ID\_VM2 | 8253 | non-null | float64 |
| **22** | BZD\_ID\_VM3 | 481 | non-null | float64 |
| **23** | BZD\_VM\_AN | 4106 | non-null | object |
| **24** | BZD\_ID\_IF1 | 54296 | non-null | float64 |
| **25** | BZD\_ID\_IF2 | 25 | non-null | float64 |
| **26** | BZD\_IF\_AN | 77172 | non-null | object |
| **27** | BZD\_ID\_TA1 | 22080 | non-null | float64 |
| **28** | BZD\_ID\_TA2 | 334 | non-null | float64 |
| **29** | BZD\_TA\_AN | 415 | non-null | object |
| **30** | IND\_ALC | 109947 | non-null | object |
| **31** | GME\_NAAM | 1224079 | non-null | object |
| **32** | PVE\_NAAM | 1224079 | non-null | object |
| **33** | FK\_VELD5 | 1224079 | non-null | object |
| **34** | NUMMER | 1224079 | non-null | int64 |
| **35** | OTE\_ID | 1163309 | non-null | float64 |
| *Dimensies:* | 1224079 x 36 |  |  |  |

Tabel 4: Data na pre-processing

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Indicatie van de Data** | | | | |  |
| **#** | **Column** | **Non-Null** |  | | **Dtype** |
| 0 | PTJ\_ID (key) | 566460 | | non-null | int64 |
| 1 | VKL\_NUMMER (key) | 566460 | | non-null | int64 |
| 2 | ANTL\_PTJ | 566460 | | non-null | int64 |
| 3 | AOL\_ID | 566460 | | non-null | category |
| 4 | AP3\_CODE | 566460 | | non-null | category |
| 5 | BEBKOM | 516099 | | non-null | category |
| 6 | OTE\_ID | 549269 | | non-null | category |
| 7 | IND\_ALC | 15399 | | non-null | category |
| 8 | MAXSNELHD | 566460 | | non-null | category |
| 9 | WDK\_ID | 513020 | | non-null | category |
| 10 | WGD\_CODE\_1 | 506282 | | non-null | category |
| 11 | WSE\_ID | 509428 | | non-null | category |
| 12 | WVG\_ID | 504143 | | non-null | category |
| 13 | WVL\_ID | 507922 | | non-null | category |
| 14 | LGD\_ID | 566452 | | non-null | category |
| 15 | LONG | 566460 | | non-null | float64 |
| 16 | LAT | 566460 | | non-null | float64 |
| 17 | BUURTNAAM | 566442 | | non-null | category |
| 18 | WIJKNAAM | 566442 | | non-null | category |
| 19 | GME\_NAAM | 566460 | | non-null | category |
| 20 | PVE\_NAAM | 566460 | | non-null | category |
| *Dimensies:* | 566460 x 19 |  | |  |  |

Afbeelding met kaart

Automatisch gegenereerde beschrijving

Figuur 9: Density Heatmap: Uitsluitend Materiele Schade

Afbeelding met kaart

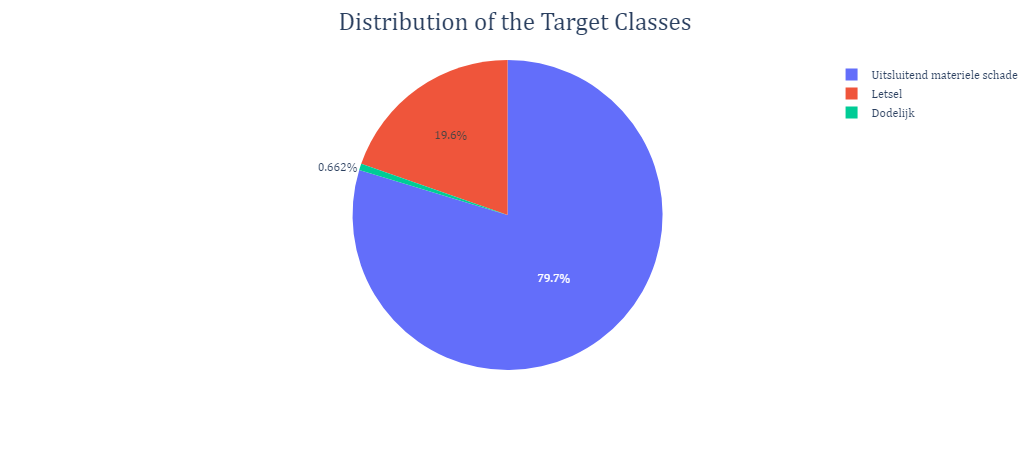
Automatisch gegenereerde beschrijving

Figuur 10: Density Heatmap: Letsel

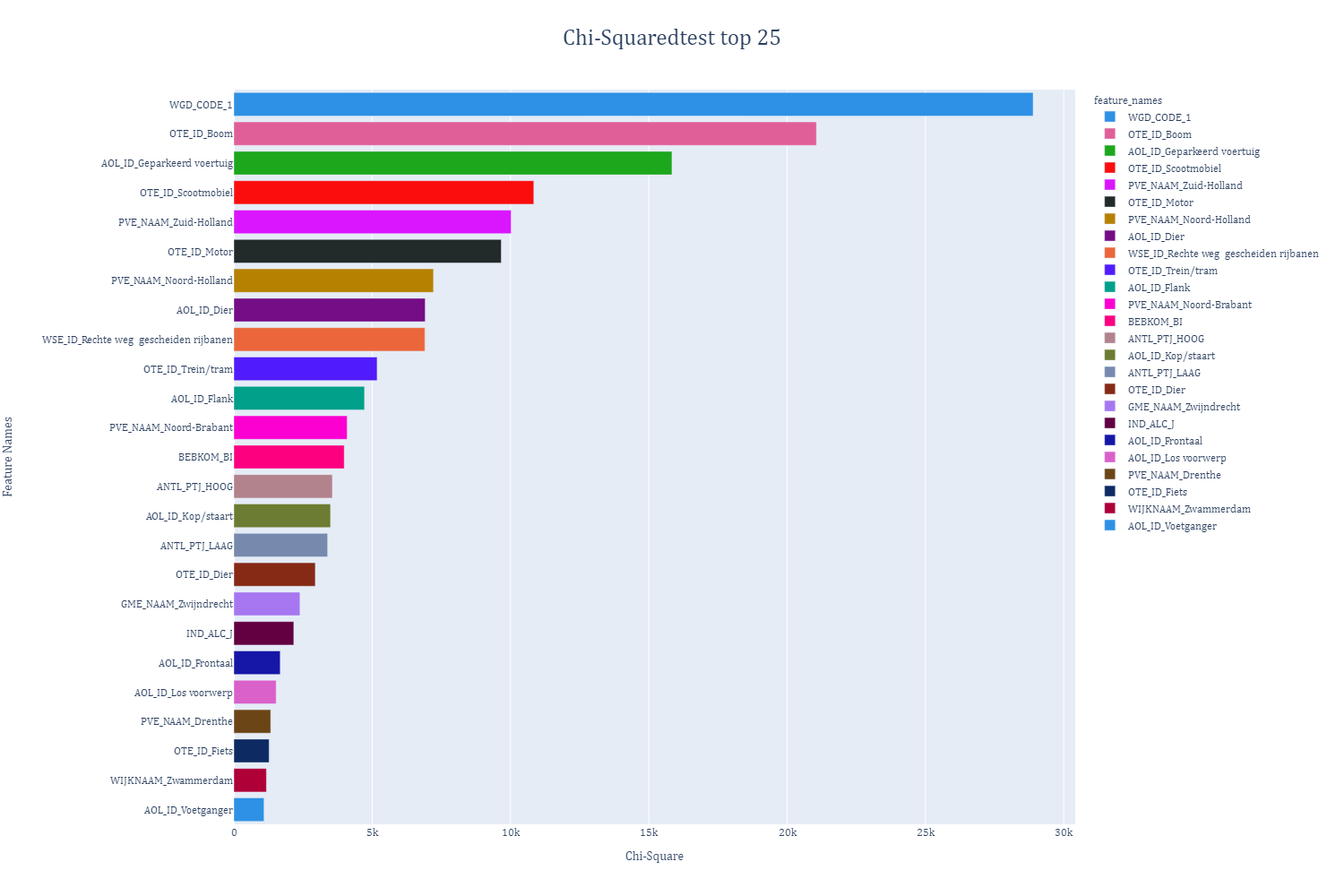
Afbeelding met kaart

Automatisch gegenereerde beschrijving

Figuur 11: Density Heatmap: Dodelijk



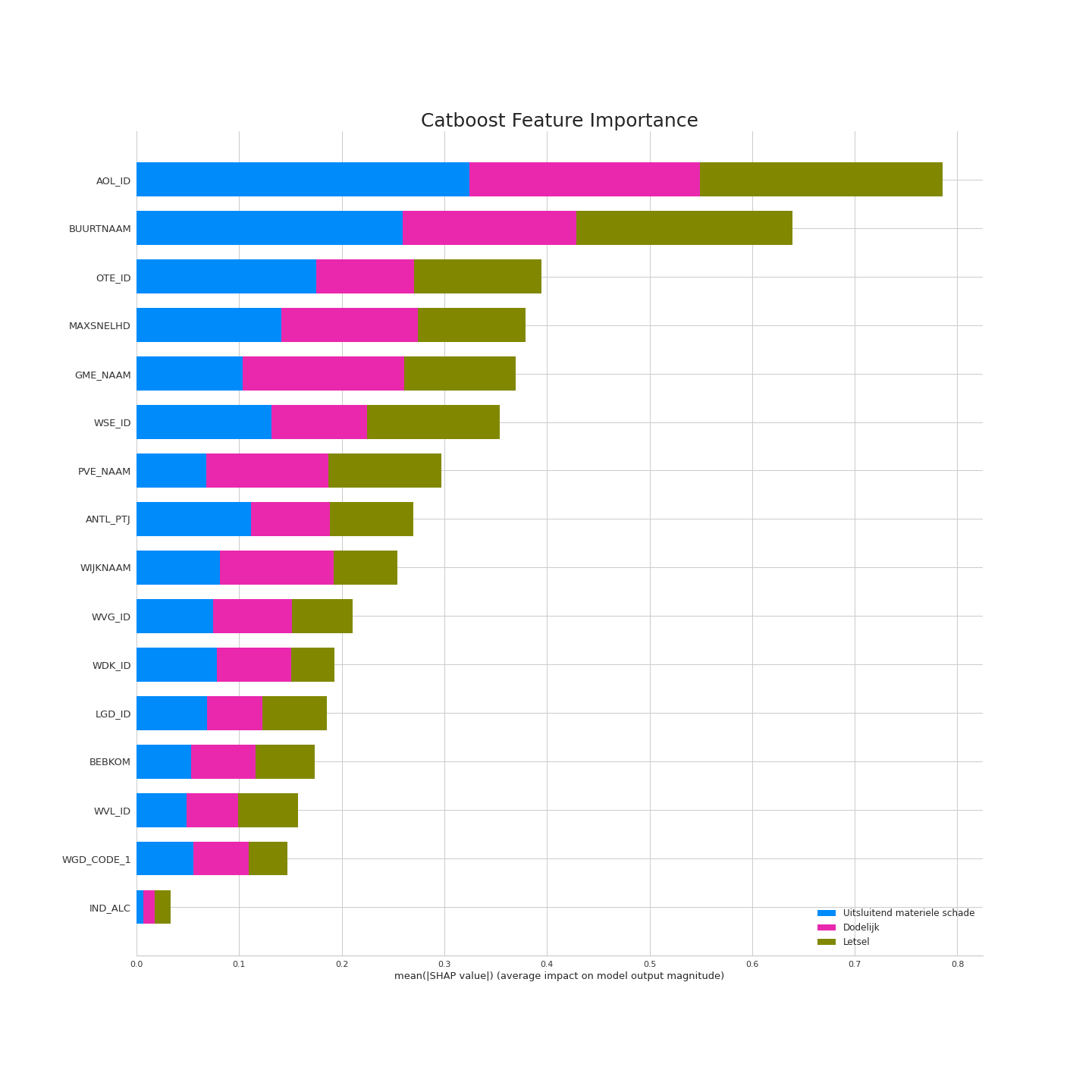
Figuur 12: Visualisatie van klassenongelijkheid



Figuur 13: Chi-Kwadraat resultaten

Tabel 5: Chi-Kwadraat resultaten van de top 50 features

|  |  |  |
| --- | --- | --- |
| **Indicatie van de Data** | | |
| ***Feature Names*** | ***χ2*** | ***p-value*** |
| WGD\_CODE\_1 | 28891,75 | 0 |
| OTE\_ID\_Boom | 21058,12 | 0 |
| AOL\_ID\_Geparkeerd voertuig | 15836,04 | 0 |
| OTE\_ID\_Scootmobiel | 10841,81 | 0 |
| PVE\_NAAM\_Zuid-Holland | 10022,19 | 0 |
| OTE\_ID\_Motor | 9665,55 | 0 |
| PVE\_NAAM\_Noord-Holland | 7216,071 | 0 |
| AOL\_ID\_Dier | 6917,232 | 0 |
| WSE\_ID\_Rechte weg gescheiden rijbanen | 6905,945 | 0 |
| OTE\_ID\_Trein/tram | 5179,079 | 0 |
| AOL\_ID\_Flank | 4723,512 | 0 |
| PVE\_NAAM\_Noord-Brabant | 4095,308 | 0 |
| BEBKOM\_BI | 3989,252 | 0 |
| ANTL\_PTJ\_HOOG | 3561,252 | 0 |
| AOL\_ID\_Kop/staart | 3494,114 | 0 |
| ANTL\_PTJ\_LAAG | 3388,99 | 0 |
| OTE\_ID\_Dier | 2943,951 | 0 |
| GME\_NAAM\_Zwijndrecht | 2389,433 | 0 |
| IND\_ALC\_J | 2166,416 | 0 |
| AOL\_ID\_Frontaal | 1676,285 | 0 |
| AOL\_ID\_Los voorwerp | 1528,498 | 0 |
| PVE\_NAAM\_Drenthe | 1331,512 | 7,3E-290 |
| OTE\_ID\_Fiets | 1274,034 | 2,2E-277 |
| WIJKNAAM\_Zwammerdam | 1176,574 | 3,2E-256 |
| AOL\_ID\_Voetganger | 1090,36 | 1,7E-237 |
| PVE\_NAAM\_Gelderland | 998,6031 | 1,4E-217 |
| OTE\_ID\_e-bike | 955,0534 | 4,1E-208 |
| WSE\_ID\_Rechte weg | 941,1751 | 4,2E-205 |
| AOL\_ID\_Vast voorwerp | 923,0249 | 3,7E-201 |
| BUURTNAAM\_Verspreide huizen Nummer Eén en Slijkplaat | 922,623 | 4,5E-201 |
| GME\_NAAM\_Landsmeer | 866,6388 | 6,5E-189 |
| WVG\_ID | 844,975 | 3,3E-184 |
| GME\_NAAM\_Lansingerland | 796,8794 | 9,1E-174 |
| GME\_NAAM\_Grave | 779,7835 | 4,7E-170 |
| GME\_NAAM\_Hulst | 749,388 | 1,9E-163 |
| OTE\_ID\_Trekker | 735,5248 | 1,9E-160 |
| ANTL\_PTJ\_GEMIDDELD | 674,8146 | 2,9E-147 |
| BUURTNAAM\_Verspreide huizen Millingen | 663,9593 | 6,7E-145 |
| WIJKNAAM\_Wijk 02 Markelo | 644,0011 | 1,4E-140 |
| OTE\_ID\_Snorfiets | 643,6661 | 1,7E-140 |
| PVE\_NAAM\_Flevoland | 635,6651 | 9,3E-139 |
| GME\_NAAM\_Neder-Betuwe | 575,0686 | 1,3E-125 |
| GME\_NAAM\_Eijsden-Margraten | 568,7387 | 3,2E-124 |
| PVE\_NAAM\_Groningen | 562,6316 | 6,7E-123 |
| BUURTNAAM\_Verspreide huizen Nulde | 557,6234 | 8,2E-122 |
| BUURTNAAM\_Gorecht-West | 557,6234 | 8,2E-122 |
| WIJKNAAM\_Wijk 14 Kamperpoort-Veerallee | 549,1819 | 5,6E-120 |
| PVE\_NAAM\_Zeeland | 534,9309 | 6,9E-117 |
| OTE\_ID\_Los voorwerp | 521,125 | 6,9E-114 |
| BUURTNAAM\_Buurt Ermelo ten dele bij de kern | 467,455 | 3,1E-102 |



Figuur 14: Catboost Feature Gewicht

**Figuur 15: Decision Tree (Klik op het figuur)**

