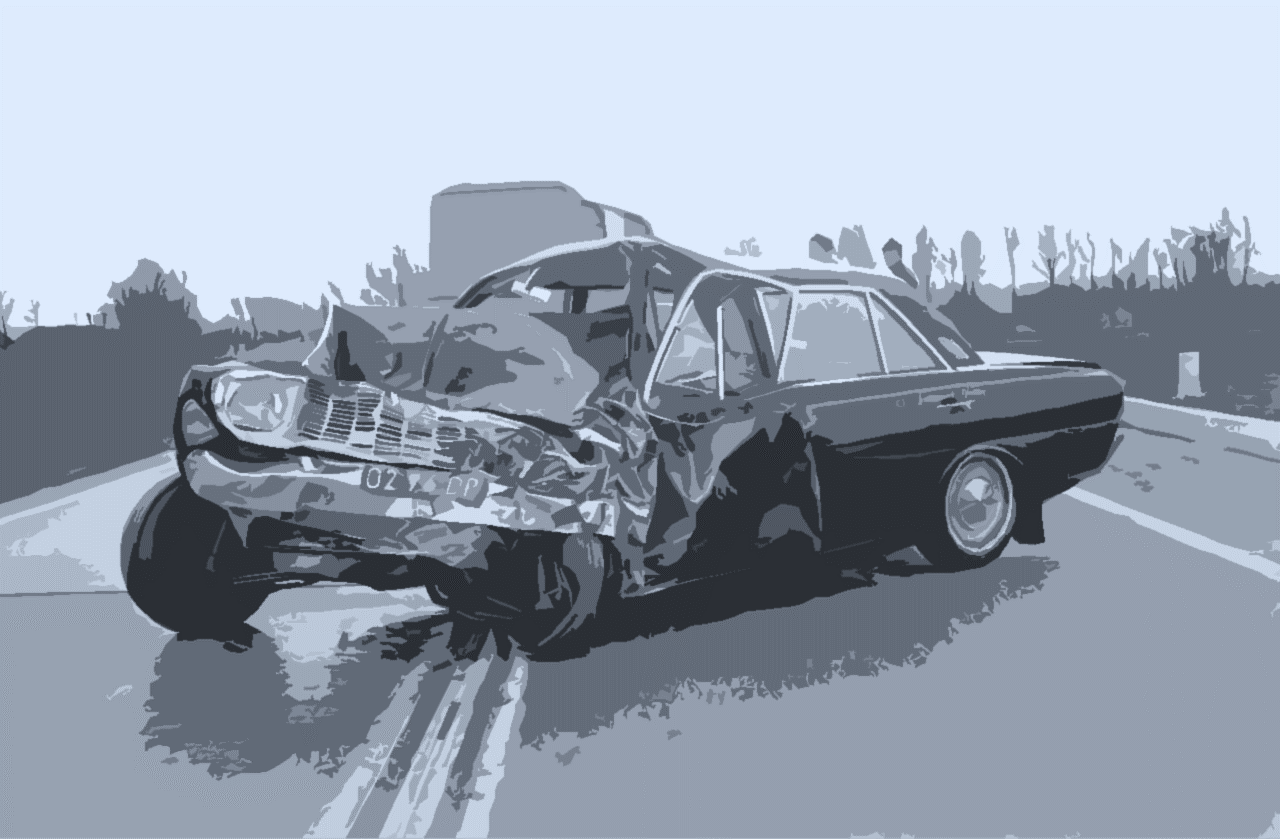
TITEL

Informatiekunde - Universiteit van Amsterdam

Universiteit van Amsterdam  
  
BACHELOR INFORMATIEKUNDE SCRIPTIE

**Stijn Melchior Willem Hering  
12377422**



**Afstudeerbegeleider: Robin Langerak  
Omvang onderzoek: 18EC**

Inhoudsopgave

[1 Introductie 3](#_Toc74254048)

[1.1 Inleiding 3](#_Toc74254049)

[1.2 Het Bestand Geregistreerde Ongevallen in Nederland (BRON) 5](#_Toc74254050)

[2 Theoretisch Kader 6](#_Toc74254051)

[2.1 Verkeersdata 6](#_Toc74254052)

[2.2 Methoden 7](#_Toc74254053)

[2.2.1 Traditionele Methoden Voor Verkeersvraagstukken 7](#_Toc74254054)

[2.2.2 Verkeersincidenten Voorspellen middels Traditionele machine-learning methoden 7](#_Toc74254055)

[2.2.3 Deep Learning 8](#_Toc74254056)

[2.2.4 Verkeersincidenten voorspellen middels Deep-Learning 10](#_Toc74254057)

[2.2.5 Ensemble Learning 10](#_Toc74254058)

[2.2.6 Ensemble Methoden voor Verkeersincidenten 12](#_Toc74254059)

[3 Bibliografie 13](#_Toc74254060)

# 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 gevolgen van verkeersongelukken veroorzaken verkeersongelukken ook aanzienlijk maatschappelijke kosten. Zowel sociale, medische en immateriële kosten als economische kosten. Het voorkomen van verkeersongelukken met verkeersdoden en verkeersgewonden ten gevolgen zijn 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 bij het merendeel van de wereld 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 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 drie kwart van de totale maatschappelijke kosten, ten gevolgen 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 gebracht 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; 500 verkeersdoden en 10.600 ernstige verkeersgewonden (IenW, 2018), zijn zelfs tijdens de Corona pandemie en de restricties die ontstonden 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 omtrent het aantal ernstige verkeersgewonden in 2020 gehaald zal worden als er word gekeken naar het totaal aantal ernstige verkeersgewonden uit 2019. In 2019 ging het om 21.400 ernstige verkeersgewonden, dit is meer dan het dubbelen van wat maximaal word verwacht in 2020. Het verminderen van het aantal verkeersdoden en gewonden is van belang als Nederland zich aan de nationale ambities wilt 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, de investeringsimpuls en de risico gestuurde aanpak vanuit het Strategisch Plan Verkeersveiligheid 2030 meer en effectievere maatregelen nodig (IenW, 2018).

Er is het afgelopen decennia veel onderzoek gedaan naar het gebruik van data wetenschap (*data science*) en kunstmatige intelligentie (*AI*) in het verkeer. In deze onderzoeken wordt er in het algemeen een antwoord gezocht voor drie soorten verkeersvoorspellingsproblemen. Deze zijn als volgt gedefinieerd; 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. Verkeersgeneratie gaat over het generen of modeleren 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 omtrent het voorspellen van verkeerssnelheden, verkeersdrukte, reisbehoeften, reistijden en verkeersongelukken. Veel van deze onderzoeken hebben gezorgd voor nieuwe ontwikkelingen en innovaties omtrent het verkeer en de verkeersveiligheid (Ermagun & Levinson, 2018; Yin et all., 2020; Yuan & Li, 2021). In Nederland zijn dit soort onderzoeken zeer beperkt, terwijl de voordelen van het uitvoeren van dit soort onderzoeken voor de verkeersveiligheid veelvuldig zijn bewezen. De Nederlands staat heeft wel stappen genomen om onderzoek omtrent AI te stimuleren. Door het creëren van het OPENDATA-portaal 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 het maken van een verkeersprognose. Het doel van dit onderzoek is om te toetsen of een voorspellend model kan bij dragen aan de verbetering van de verkeersveiligheid in Nederland. Doormiddel van technieken en methoden uit het domein van de statistiek, data wetenschap (*data science*) en AI toe te passen op het *Bestand Geregistreerde Ongevallen in Nederland* (**BRON**) wordt hier een getracht uitsluitsel over te geven. Het BRON is beschikbaar is gesteld als *OPENDATA* (Rijkswaterstaat, 2020) 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, etc), informatie over het weer, informatie over bestuurders/inzittende, et cetera. 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 het voorspellen van verkeersongelukken aan de hand van historische data?***

*Benoem ik hier ook de sub-vragen?*

* Wat is de maatschappelijk relevantie van het verzamelen van verkeersdata?
* Welke *machine-learning* methoden zijn er voor verkeersgerelateerde vraagstukken?
* Welke machine-learning methoden worden toegepast voor verkeersincidenten voorspelling?
* Wat geven de resultaten aan?
* Wat kan beter?

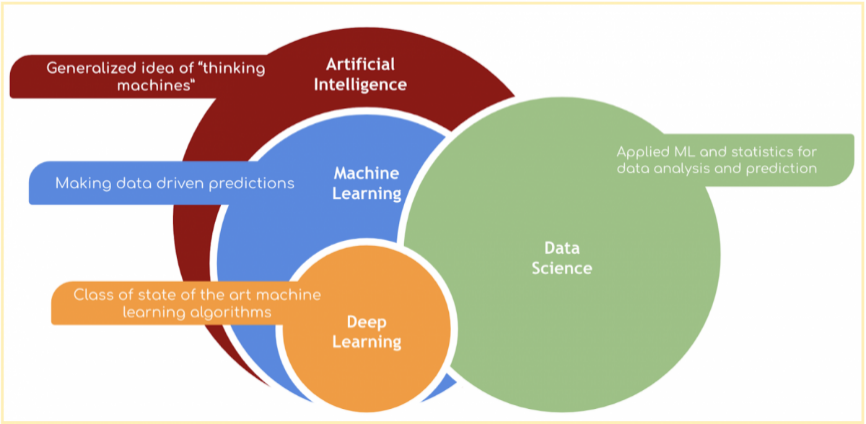
Door middel van kwantitatief onderzoek, via een data-analyse en het maken van een voorspellingsmodel, word kennis en inzicht verkregen over factoren die invloed hebben op verkeersincidenten.Data omtrent verkeersongelukken is verzameld via *Overheid.nl*. De data analyse is uitgevoerd via *JupyterLab* een web gebaseerde interactieve ontwikkelomgeving voor *Jupyter notebooks*, code, en data. HIER KUNNEN LATER NOG DE MODELLEN EN METHODEN KORT WORDEN BESCHREVEN 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 maken van voorspellingen door het voorspellend model. Door het maken van voorspellingen is duidelijk geworden waar, wanneer en met welke factoren de kans op verkeersongelukken het grootst is.

In het volgende segment wordt het belang van verkeersdata toegelicht, vervolgens wordt verduidelijk gegeven welke *data-science* en *machine-learning* toepassingen er al zijn onderzocht en ontwikkeld met betrekking tot het voorspellen van verkeersgerelateerde onderwerpen, deze zijn gecategoriseerd in drie klassen: traditionele methoden, *deep-learning* methoden en ensemble methoden. Per klasse worden voorbeelden van modellen aangehaald die worden toegepast om verkeersincidenten te voorspellen. In het segment dat volgt wordt de methodologie toegelicht en wordt de gebruikte data beschreven. Vervolgens worden de resultaten van de data-analyse en het voorspellend model gedemonstreerd. Het laatste segment behandeld de discussie en conclusie van dit onderzoek.

# Theoretisch Kader

## Maatschappelijke Relevantie Verkeersdata

Het gebruik van data-wetenschap en kunstmatige intelligentie bij het analyseren en voorspellen van verkeersgerelateerde kwesties is de afgelopen jaren enorm toegenomen. Deze ontwikkeling zijn versneld door de opkomst van big-data en het Internet of Things (IoT). Big data en het IoT zijn van significant belang voor het monitoren en analyseren van het verkeer om de gezondheid en veiligheid van mensen te waarborgen (Ge, Bangui Buhnova, 2018).  
Het overgrote deel van de ontwikkelde landen doet aan een vorm van verkeersmonitoring. Door een variërende infrastructuur is er wel een verschil tussen landen in hoe dit wordt aangepakt (Leduc, 2008; Azmat et al, 2019).  
In principe geldt voor elk land hetzelfde, het monitoren van wegen en het verkeer wordt gedaan om verkeersopstoppingen en ander oponthoud in het verkeer tegen te gaan. Verkeersopstoppingen en ander oponthoud hebben namelijk negatieve gevolgen zoals een langere reistijd, een hoger brandstofverbruik en meer luchtvervuiling. Het monitoren van het verkeer en vervoer, in de vorm van data te verzamelen en op te slaan helpt om deze problemen te verminderen en de mobiliteit en stabiliteit in een samenleving te verbeteren (Shi, Abdel-Aty, 2015).



**Figuur 1: Illustratie van AI subcategorieën  
(How AI, Machine-learning, Deep-learning and Data science intersect, z.d.)**

### Verzameling Verkeersdata

Het verzamelen van verkeersdata gebeurt tegenwoordig voornamelijk automatisch door middel van *slimme apparaten* die in een netwerk in verbinding staan met elkaar (IoT). Deze netwerken van *slimme apparaten* worden beheerd door (branche)organisatie of overheidsinstantie die een belang hebben bij het monitoren en managen van de wegen en het verkeer (Barceló, Kuwahara & Miska, 2010). Verkeersdata kan voor verschillende doeleinden gebruikt. Door het uitvoeren van een data-analyse en het toepassen van machine-learning modellen op verkeersdata weten (branche)organisatie of overheidsinstantie welke factoren van belang zijn in het verkeer.

### Toepassingen *machine-learning*

Zo wordt door het toepassen van *machine-learning* modellen op verkeersdata verkeerscongestie verminderd, verkeersopstoppingen automatisch gedetecteerd, wegen en verkeerstromen geoptimaliseerd door een beter gebruik van het huidige wegennet, worden informatiediensten (e.g. verkeersinformatie) verbeterd en efficiënter, word dynamische routegeleiding mogelijk, kunnen verkeersgerelateerde oplossen worden gesimuleerd en kunnen er verkeersvoorspelling worden gedaan (Leduc, 2008). Het maken van accurate verkeersvoorspellingen vormt een grote uitdaging in de commerciële sector en de wetenschap. Hierdoor is er een enorme variatie in methoden en toepassingen die een uitkomst bied voor het verkeervoorspellingsprobleem. De reden hiervoor is af te leiden aan de vorm van verkeersdata. Verkeersdata bestaat voor het grootste deel uit ruimtelijke en temporele (spatio-temporal) ongestructureerd data die in grote hoeveelheid word vergaard door sensoren, camera’s, GPS-signalen en detectie apparatuur (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 veelal automatisch en in *real-time* (Nationale Databank Wegverkeersgegevens (NDW), z.d.)

Verkeersdata transformeert hierdoor voortdurend van locatie en tijd waardoor het complexe en dynamische ruimtelijk-temporele afhankelijkheden vormt. Verder wordt ruimtelijke en temporaire verkeersdata ook beïnvloed door een heleboel externe factoren, zoals weersomstandigheden, gebeurtenissen of wegkenmerken. Hierdoor vertonen verkeersdata een sterke dynamische correlatie in zowel ruimtelijke als temporele dimensies (Yin et all., 2020; Liu, Li, Wu & Li, 2018).

### Relevantie verkeersincidenten voorspelling

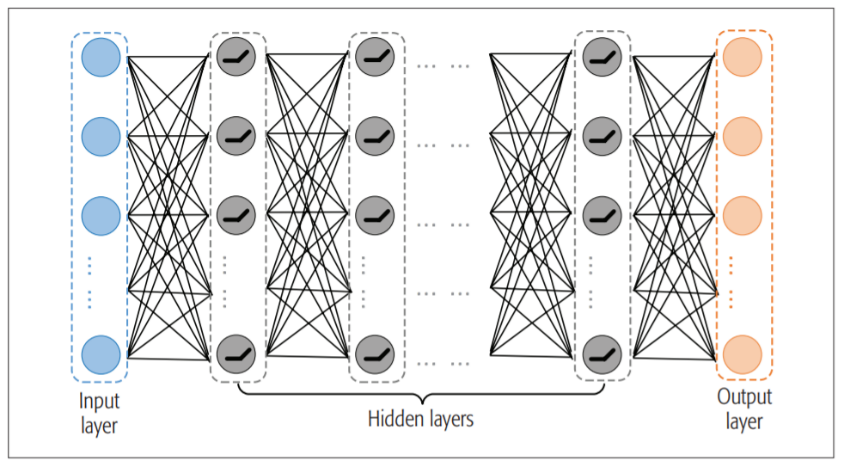
## Machine Learning Methoden

De methodes die het meest voorkomen omtrent het voorspellen van verkeersproblemen zijn onder te verdelen in drie categorieën: *klassieke of traditionele machine-learning*, *deep-learning* en *ensemble* *learning*. In de volgende segmenten worden deze categorieën kort toegelicht. Vervolgens worden per categorie de meest gebruikt methoden voor het voorspellen van verkeersincidenten aangehaald.

### Traditionele Methoden Voor Verkeersvraagstukken

Traditionele methoden voor het oplossen van verkeersvraagstukken omvatten veelal statistische methoden en conventionele *machine-learning* modellen zoals, *decision trees,* regressie en classificatie modellen. Deze methoden vereisen dat de data waar mee wordt gewerkt aan bepaalde veronderstellingen voldoet, hierdoor werken deze modellen niet goed met complexe data. De 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
* *Outliers* in de data



**Figuur 2: Illustratie van een Deep-neural network   
(Liu, Li, Wu & Li, 2018, p. 42)**

* Normaalverdeling
* Homogeniteit tussen groepen

Traditionele methoden slagen er vaak niet in om rekening te houden met sterke niet lineaire temporele verbanden. Verkeersdata is, door de *spatio-temporal*-vorm over het algemeen te complex voor deze klassieke methoden. Hierdoor hebben de klassieke methoden een relatief lage nauwkeurigheid op verkeersdata. Bovendien zijn deze methoden vrijwel alleen goed toepasbaar op relatief kleine datasets (Li & Shahabi, 2018; Yuan, Zhou & Yang, 2018; Yin et all., 2020).

### Verkeersincidenten Voorspellen middels Traditionele machine-learning methoden

Traditionele *machine-learning* methoden, zoals statistische en klassieke regressie en classificatie modellen waren rond de 2000e eeuw erg populair om verkeersincidenten te voorspellen. Statistische modellen werden voornamelijk gebruikt voor factor en scenario analyses en de klassieke regressie en classificatie modellen worden voornamelijk gebruikt om de kansen van verkeersincidenten te voorspellen (bereken). Vanwege de sterke data-assumpties en het beperkte extrapolatievermogen, hing de nauwkeurigheid van deze modellen grotendeels af van de steekproefgrootte en de integriteit (Zheng & Liu, 2009; Chand, Jayesh & Bhasi, 2021). Daarnaast waren, door de ruimtelijke en temporele (spatio-temporal) afhankelijkheden in verkeersdata klassieke methoden niet complex genoeg om een betrouwbare representatie van de verkeersdata te vormen.

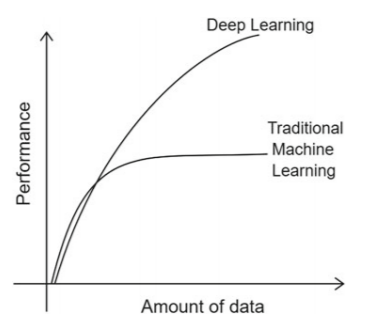
Met de komst van *time-series* methoden werd dit probleem deels verholpen. *Time-series* methoden gebruiken historische tijd gerelateerd data als basis voor het voorspellen van kans waardes. Anders dan bij klassieke statistische en regressiemodellen zijn *time-series* methoden gebaseerd op de correlaties tussen de tijd-data punten. Methodes zoals ARMA, een combinatie van een *autoregressive* (AR) en *moving average* (MA) deel; en ARIMA *autoregressive integrated moving average* konden wel omgaan met tijdgebonden afhankelijkheden in verkeersdata (Zheng & Liu, 2009; Ermagun & Levinson, 2018; Yin et all., 2020).

Voor verkeersincident classificatie word voornamelijk gebruikt gemaakt van *decision trees*, *K-nearest neighbor (KNN), support-vector-machines* (SVM) en *Naive Bayes classifiers*. Deze ondervinden vergelijkbare problemen met klassieke regressiemodellen. De modellen zijn niet complex genoeg voor de verkeersdata of er moet worden voldaan aan bepaalde data-assumpties (Ermagun & Levinson, 2018; Li & Shahabi, 2018; Chand et all., 2021).

### Deep Learning

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



**Figuur 3: (Mathew, Amudha & Sivakumari, 2020, p. 601)**

*Deep-learning* modellen zijn hierdoor uiterst geschikt voor het maken van voorspellingen aan de hand van verkeersdata (Yin et all., 2021). Er zijn 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 optima. *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)

Voordelen en nadelen deep-learning modellen opgesomd:

**Voordelen**

* 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

**Nadelen**

* Heeft veel data nodig om goede resultaten te behalen
* Vergt veel rekenkracht
* Complex om toe te passen
* Slecht interpreteerbaar

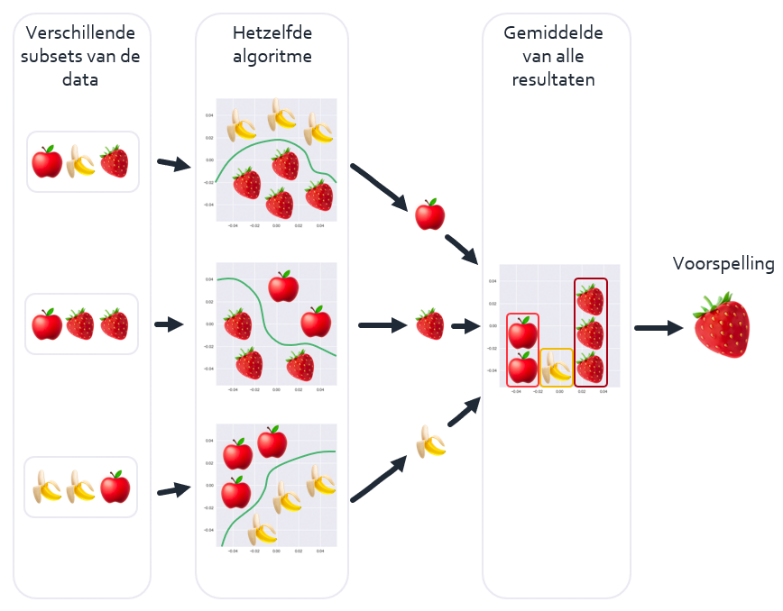
*Deep-Learning* zijn zeer accuraat en goed in het oplossen van bepaalde verkeersgerelateerde vraagstukken, maar door de complexiteit en slechte-interpreteerbaarheid is het belangrijk om na te gaan of het gebruik van simpelere *machine-learning* methodes niet beter aansluiten bij het oplossing van verkeersvraagstukken.

### Verkeersincidenten voorspellen middels Deep-Learning

Verreweg het meeste onderzoek omtrent verkeersincident voorspelling bevat het gebruik van de *Deep-learning* methode. Er zijn tegenwoordig veel variante *Deep-learning* modellen die dit probleem kunnen oplossen maar er zijn twee modellen die veelvuldig voorkomen in verkeersincident voorspellende problemen. Dit zijn het het *Recurrent Neural Network* (RNN) en het *Convolutional Neural Network* (CNN) (Li & Shahabi, 2018).   
Met name de subcategorie van het RNN, de Long Short-Term Memory (LSTM) is populair voor het voorspellen van verkeersincidenten. Het RNN wordt veel gebruikt bij data die in een serie vormt (e.g. time-series), waarbij er rekening gehouden wordt met de invloed van onderlinge data punten. Een LSTM is goed in het modeleren van niet-lineaire temporele correlaties en wordt vaak gecombineerd met andere *deep-learning* modellen zoals een CNN. Het CNN en zijn varianten worden met name gebruikt om ruimtelijke correlatie van de data te modelleren. *Deep-learning* modellen zijn in vergelijking met andere *machine-learning* het meest accuraat als het gaat om het voorspellen van verkeersincidenten (Yin et all., 2020; Akhtar & Moridpour, 2021; Yuan & Li, 2021).

### Ensemble Learning

Bij *Ensemble Learning* wordt een ‘sterk’ model gebouwd met behulp van een verzameling (of "ensemble") van "zwakkere" modellen. Deze ensembles kunnen bestaan uit meerdere *klassieke of traditionele* methodes*,* meerdere *Deep-learning* methodes of een combinatie van de twee. Twee bekende ensemble methodes zijn Extreme Gradient Boosting (XGB) en Random Forrest (RF) (Sagi & Rokach, 2018)



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

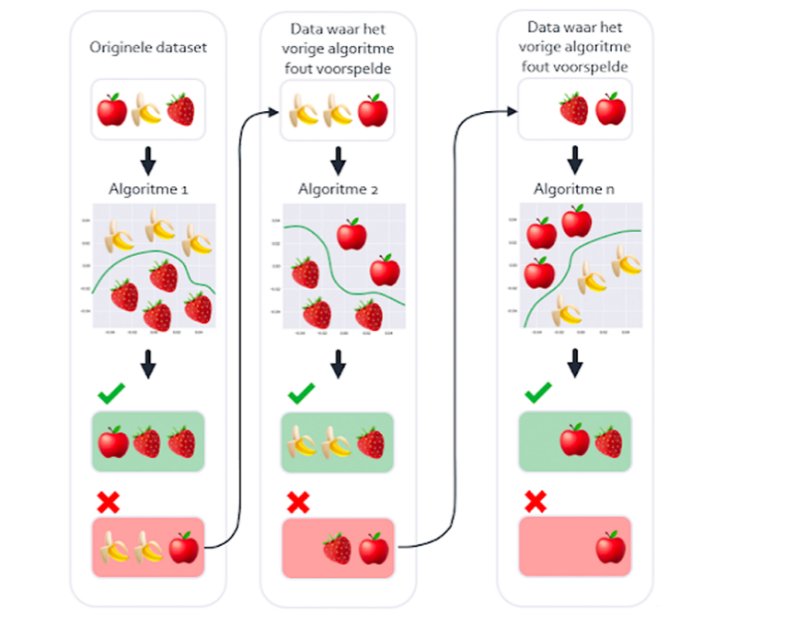
Ensemble methoden hebben een aantal voordelen in tegenstelling tot enkele *machine-learning* en Deep-Learning modellen:

* Vermijd overfitting
* 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. 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.

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 verminderd of vermeid deze bias door gebruik te maken van gerandomiseerde *under-sampling techniques* zoals *bagging, boosting* of *stacking.* Door gebruik te maken van *subsamples* die wel gebalanceerd zijn worden de voorspellende prestaties van de ensemble methoden aanzienlijk verbeteren voor data met een onevenwichtigheid in klasse

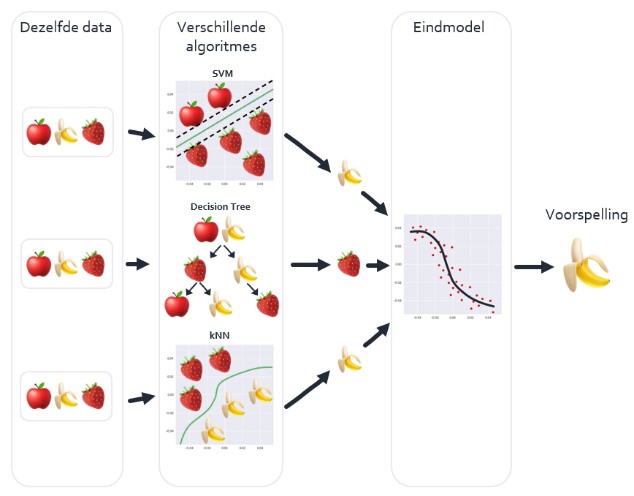
De data die als input dient voor real-time *machine-learning* toepassingen heeft vaak de neiging om te veranderen. Zo kunnen nieuwe features en labels zijn toegevoegd of er ontstaat een verandering in de distributies van deze labels en features. Deze verandering heeft significante gevolgen voor de voorspellende werking van deze toepassing. Door de inherente diversiteit van ensemble modellen en 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 generaliseerd, dit word ook wel de ‘vloek van de dimensionaliteit’ genoemd.



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

Ensemble learning methoden kunnen gebruikt worden om deze dimensionaliteit te verminderen of de meest essentiële dimensies te scheiden van 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).

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).



**Figuur 6: Figuur 5: Illustratie van de Stacking methoden**

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: Your Machine-learning and Data Science Community*, z.d.).

### Ensemble Methoden voor Verkeersincidenten

Ensemble methoden worden al redelijk veel gebruikt bij het voorspellen van verkeersgerelateerde onderwerpen. Het gaat hier voornamelijk om *ensemble decision tree* modellen zoals RF (bagged) en de XGB (boosting) methode (Alajali, Zhou, Wen & Wang, 2018; Yin et all., 2020). Het praktisch aan deze methoden is dat belangrijke features in de verkeersdata direct zichtbaar worden omdat ze bovenin de gecreëerde staan. Hierdoor wordt door het maken van een voorspellend model indirect ook duidelijk welke features de impact hebben op verkeersongelukken. Verder werken *RF* en XGB vanwege hun *decision tree* methode vooral goed categorische data, verkeersincidenten data bestaat voor een groot deel hieruit (Liu & Wu, H, 2017; Alajali et all., 2018; Chand, et all., 2021).

Verder word er ook geëxperimenteerd met ensembles van deep-learning modellen om verkeersincidenten te voorspellen. Er is echter meer onderzoek vereist om deze *deep*-ensembles praktisch relevant te maken en te verbeteren (Dong et all., 2020).

# Methoden Sectie

Het **BRON**

# Bibliografie

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

Alajali, W., Zhou, W., Wen, S., & Wang, Y. (2018). Intersection traffic prediction using decision tree models. Symmetry, 10(9), 386.

Akhtar, M., & Moridpour, S. (2021). A Review of Traffic Congestion Prediction Using Artificial Intelligence. *Journal of Advanced Transportation*, *2021*.

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.

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

Chand, A., Jayesh, S., & Bhasi, A. B. (2021). Road traffic accidents: An overview of data sources, analysis techniques and contributing factors. *Materials Today: Proceedings*.

*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.

*How AI, Machine-learning, Deep Learning and Data science intersect*. (z.d.). [IIlustratie]. Turnitin. <https://www.turnitin.com/blog/artificial-intelligence-and-machine-learning-at-turnitin>

*Kaggle: Your Machine-learning and Data Science Community*. (z.d.). Kaggle. Geraadpleegd op 9 juni 2021, van <https://www.kaggle.com/>

Leduc, G. (2008). Road traffic data: Collection methods and applications. *Working Papers on Energy, Transport and Climate Change, 1*(55), 1-55.

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.

Liu, Y., & Wu, H. (2017, December). Prediction of road traffic congestion based on random forest. In *2017 10th International Symposium on Computational Intelligence and Design (ISCID)* (Vol. 2, pp. 361-364). IEEE.

Mathew, A., Amudha, P., & Sivakumari, S. (2020, February). Deep Learning Techniques: An Overview. In *International Conference on Advanced Machine-learning Technologies and Applications* (pp. 599-608). Springer, Singapore.

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>

Nationale Databank Wegverkeersgegevens. (z.d.). Big data. Nationaal Dataportaal Wegverkeer. Geraadpleegd op 3 mei 2021, van <https://ndw.nu/diensten/big-data>

Tieleman, P. & Data Science Partners. (2020, 23 april). *Stacking, Bagging, Boosting* [Illustratie]. Data Science Partners. <https://pythoncursus.nl/wp-content/uploads/2020/04/ensemble-methods-boosting-768x600.png>

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

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

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

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>

Wijnen, W., Weijermars, W., Schoeters, A., van den Berghe, W., Bauer, R., Carnis, L., ... & Martensen, H. (2019). An analysis of official road crash cost estimates in European countries. Safety science, 113, 318-327.

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

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).

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*.

Zheng, X., & Liu, M. (2009). An overview of accident forecasting methodologies. *Journal of Loss Prevention in the process Industries*, *22*(4), 484-491.