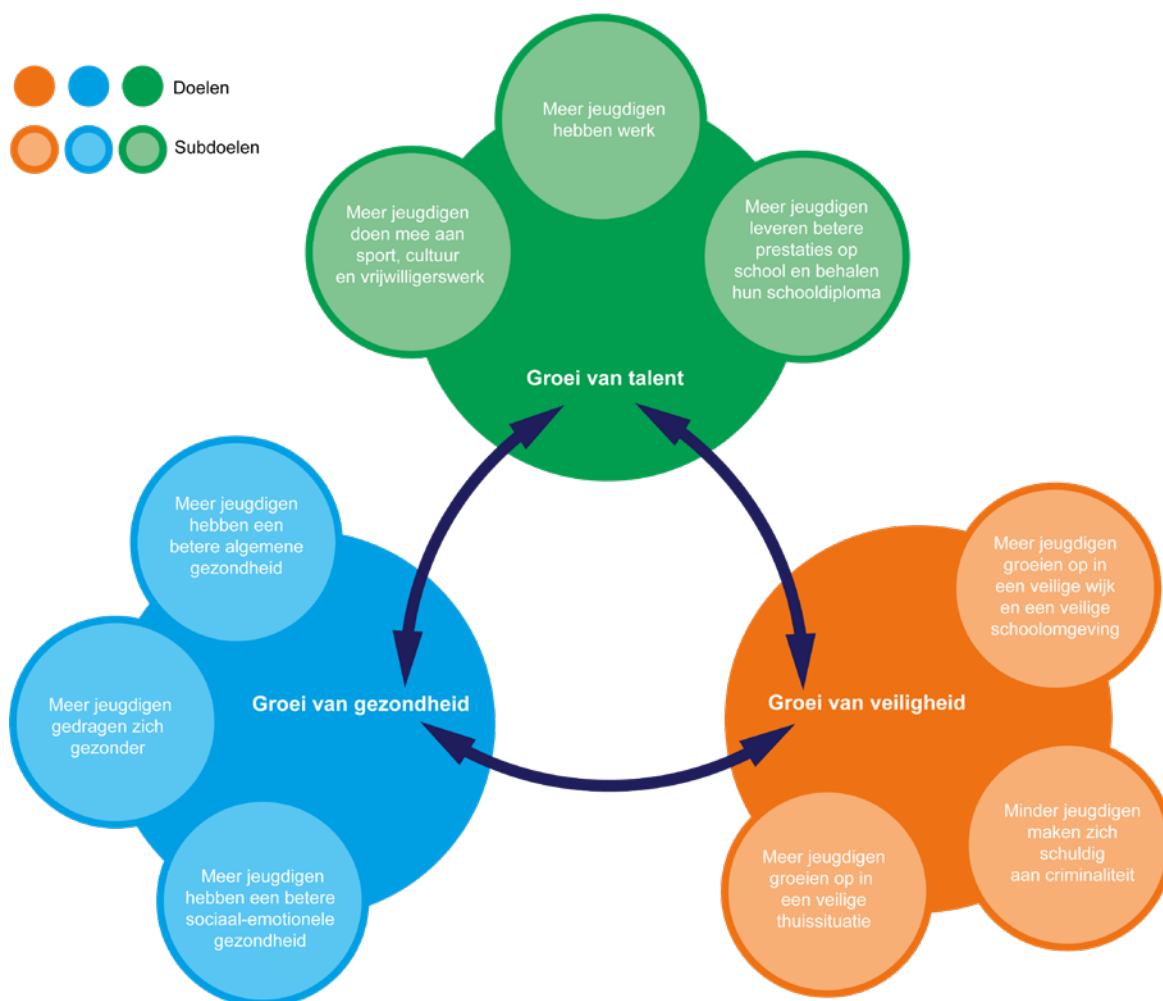


Data science analyse Rotterdam Groeit

1. Inleiding

De potentie van de Rotterdamse jeugd waarmaken: meer kinderen en jongeren groeien kansrijk, veilig en gezond op. Dat is de ambitie van de Gemeente Rotterdam, vertaald in het beleidskader jeugd 'Rotterdam Groeit'. Deze maatschappelijke impact kan bereikt worden door in beleid meer gebruik te maken van wetenschappelijke kennis, actuele informatie en op basis daarvan blijvend te vernieuwen en verbeteren. Data science is één van de manieren om dat te kunnen doen. Rotterdam heeft in samenwerking met het ministerie van Onderwijs Cultuur & Wetenschap (OCW), TNO en het ministerie van Binnenlandse Zaken (BZK) een data science analyse uitgevoerd. Deze factsheet beschrijft het doel en de uitkomsten van dit onderzoek en geeft op basis van deze ervaring een beeld van de kansen voor beleid en praktijk.

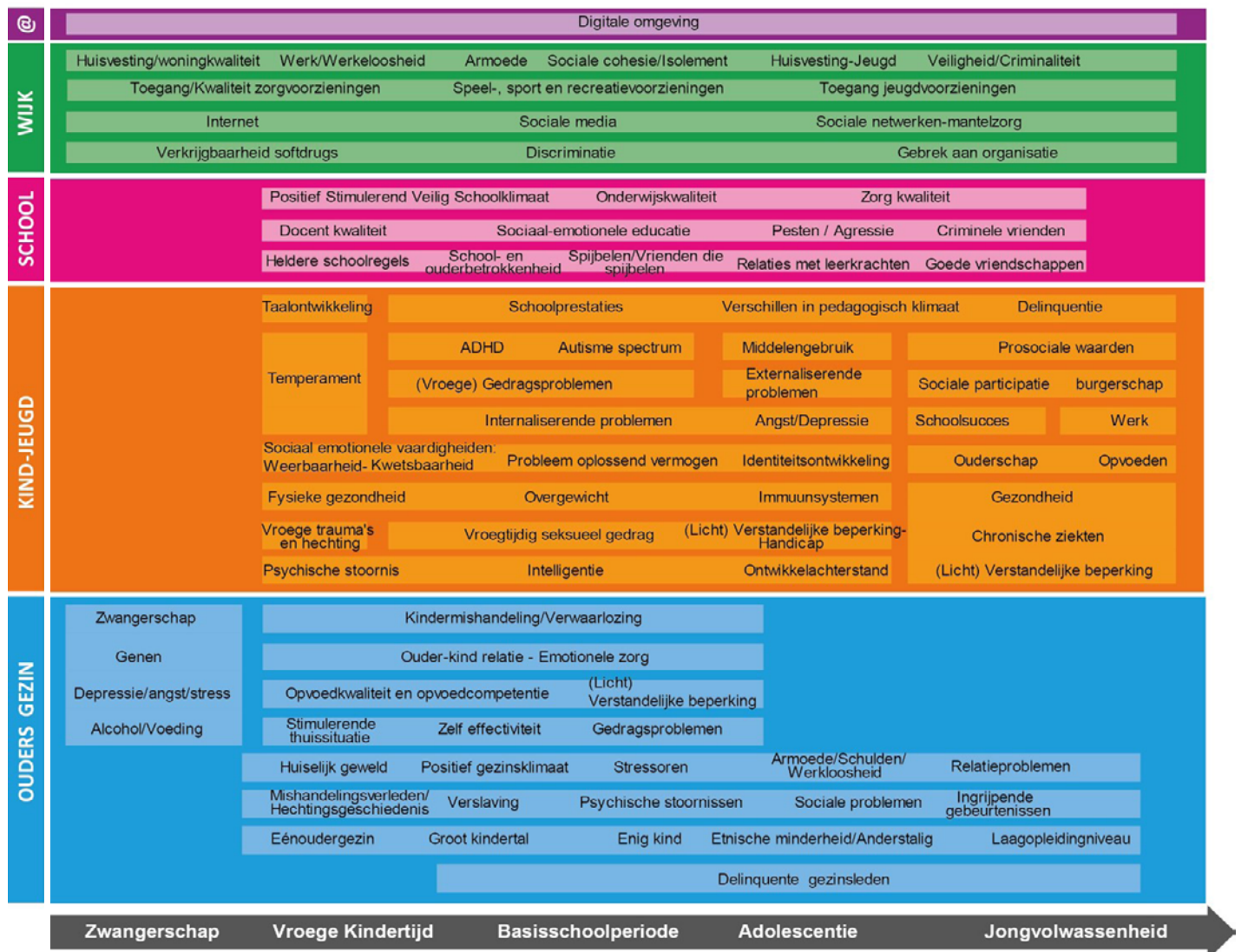
Doelen en subdoelen Rotterdam Groeit



Afb. 1: Doelen en subdoelen Rotterdam Groeit

2. Doel

Het factorenmodel Rotterdam Groeit is het wetenschappelijk fundament voor het Rotterdamse jeugdbeleid. Het maakt duidelijk welke beschermende factoren en risicofactoren van invloed zijn op de ontwikkeling van kinderen en ook welke factoren het meest bepalend zijn. Dit biedt aangrijpingspunten voor beleid. Het factorenmodel is gebaseerd op de meest actuele kennis uit de wetenschappelijke literatuur.



Afb. 2: Factorenmodel Rotterdam Groeit

Het factorenmodel Rotterdam Groeit is dynamisch: nieuwe kennis leidt tot verbetering van het model, bijvoorbeeld tot nu toe onbekende factoren die een rol blijken te spelen in de ontwikkeling van kinderen. Ook data science kan helpen om het factorenmodel verrijken: het kan inzicht geven in de mate van invloed van deze factoren in de ontwikkeling van kinderen en de mate van onderlinge samenhang tussen factoren in de specifieke Rotterdamse situatie. Dat maakt dat wetenschappelijke kennis en actuele data elkaar in een hybride model kunnen versterken.

De behoefte van Rotterdam om met data science nog beter beleid te ontwikkelen sluit aan bij de ambities van het ministerie van OCW, TNO en het ministerie van BZK. Het ministerie van OCW wil via hackathons tot meer en betere big data-analyses komen op het gebied van onderwijs, met gemeenten en regio's in heel Nederland. De ambitie van TNO en het ministerie van BZK is om data science, beleid en praktijk nauwer met elkaar te verbinden en verder te brengen binnen zogenoemde Policy Labs. Deze ambities hebben geleid tot een samenwerking tussen deze partijen in deze Rotterdamse data science analyse.

Het doel van deze specifieke analyse was om met verschillende Machine Learning technieken het factorenmodel Rotterdam Groeit te verrijken en inzichten te verwerven voor potentieel effectief beleid voor het kansrijk, veilig en gezond opgroeien van kinderen in de stad.

De data science analyse levert - naast inhoudelijke uitkomsten - kennis op over:

- Een werkwijze voor het bouwen van een hybride model van generieke wetenschappelijke literatuur en actuele data binnen een lokale context
- Het gebruik van Machine Learning in de sociale context
- Het werken met data van verschillende partners op populatieniveau van de hele stad
- Het ontwikkelen van een dataset met potentie voor kwalitatief hoogwaardige analyses
- De uitlegbaarheid van kunstmatige intelligentie in een publieke beleidscontext

3. Onderwerp data science analyse: kansen voor kinderen

De sociaal emotionele gezondheid van jeugdigen speelt een cruciale rol in hun ontwikkeling, zo blijkt uit de wetenschappelijke literatuur. De mate waarin sociaal emotionele vaardigheden ontwikkeld zijn bij kinderen en jongeren is van invloed op hun kansen op succes op het gebied van school, werk en talentontwikkeling. Daarnaast speelt deze sociaal emotionele gezondheid een rol in de ontwikkeling én het voorkomen van problemen, zoals criminaliteit, psychosociale problemen en schooluitval.

Uit de ranking van factoren op basis van wetenschappelijke literatuur komt de factor sociaal emotionele gezondheid als belangrijkste factor in het factorenmodel Rotterdam Groeit naar voren. Dit betekent dat met beleid gericht op het versterken van deze sociaal emotionele gezondheid bij kinderen grote impact bereikt kan worden als het gaat om het kansrijk, veilig en gezond opgroeien.

In deze data science analyse onderzoeken we:

- oorzaken en (mate van) invloeden van factoren op de sociaal emotionele gezondheid
- welke factoren in welke mate een rol spelen bij het verwezenlijken van kansen voor kinderen, bijvoorbeeld ten aanzien van schoolsucces, maar ook risico's zoals op het gebied van delinquentie

4. Opzet en werkwijze

Een projectteam, bestaande uit strategische beleidsadviseurs, datascientisten en onderzoekers heeft de onderzoeksvragen geformuleerd en indicatoren en bronnen voor relevante factoren uit het factorenmodel Rotterdam Groeit in kaart gebracht. Daarna is de digitale omgeving gereedgemaakt, de privacy impact assessment (PIA) opgesteld en zijn data uit verschillende bronnen verzameld, getoetst en geprepareerd.

Voor de analyse is data samengebracht van verschillende partijen die allemaal in hun eigen context informatie hebben over jeugdigen uit onze stad. Om een integraal beeld te kunnen hebben van het veilig, gezond en kansrijk opgroeien, is samengewerkt met enkele grotere partners die delen van deze informatie in hun data beschikbaar hebben. Door de data van het Centrum voor Jeugd en Gezin, het CBS, de Gemeente Rotterdam en DUO te combineren hebben we nieuwe inzichten kunnen krijgen die voorheen niet mogelijk waren.

In het onderzoek is aan de hand van Machine Learning en statistische methoden gezocht naar de invloed van factoren op het veilig, kansrijk en gezond opgroeien van de Rotterdamse jeugd. De focus ligt bij de analyse op associaties tussen factoren rondom sociaal emotionele gezondheid, schoolsucces en criminaliteit. Machine Learning is een data-gedreven onderzoeksmethode. Op basis van patronen die uit data naar voren komen wordt er een model 'getraind' dat ingezet wordt om voorspellingen te doen.

Gedurende de analyses zijn vakinhoudelijke experts uit wetenschap, beleid en praktijk betrokken om te reflecteren en om lessen uit de data te koppelen aan de werkelijke en de wetenschappelijke wereld. Door de analyse van een grootstedelijk populatieonderzoek met moderne onderzoeksmethoden te koppelen aan inhoudelijke expertise is een unieke combinatie gemaakt die voor een maximaal leereffect zorgt en leidt tot verbeterde inzichten voor beleid, uitvoering en regelgeving.

5. Privacy

In dit project worden analyses uitgevoerd om op populatieniveau relaties te onderzoeken tussen indicatoren van de verschillende factoren. Het project heeft niet als doel de situaties van individuele personen te onderzoeken. Voor de reikwijdte van de analyses is het wel nodig de gegevens te koppelen op individueel niveau. De data van verschillende bronnen zijn door het CBS gepseudonimiseerd en op persoonsniveau gekoppeld, waardoor niet te herleiden is welke factoren bij welke persoon een rol spelen. Doordat de analyses zijn uitgevoerd op populatieniveau, is de privacy van elk individu volledig gewaarborgd. Het onderzoek is uitgevoerd in de beveiligde omgeving van het CBS, door onderzoekers en datascientists van de Gemeente Rotterdam en TNO. De partijen hebben gezamenlijk een privacy impact assessment (PIA) opgesteld om de risico's voor de privacy in kaart te brengen en strategieën om de risico's te minimaliseren vast te leggen.

6. Uitkomsten onderzoek

Voor de analyse zijn Machine Learning modellen gemaakt en getraind met de beschikbare data. Bij het trainen van het model is er gekozen voor decision tree algoritmen, omdat de uitkomsten uit deze analyses uitlegbaar zijn. Daarbij is een statistische analyse uitgevoerd op een beperkte set variabelen. Er zijn analyses uitgevoerd op drie uitkomsten: sociaal-emotionele gezondheid (risico op psychosociale problemen), schoolsucces en delinquentie. Daarbij zijn de analyses voor de verschillende uitkomsten gedaan met zowel alle beschikbare variabelen, als alleen een selectie van variabelen op basis van literatuur (het factorenmodel).

De performance en daarmee de zeggingskracht van het datagedreven model dat is opgesteld op basis van Machine Learning is nog beperkt. Met name de waarden die relatief weinig voorkomen (bijvoorbeeld delinquentie, psychosociale problemen) zijn moeilijk te voorspellen in dit model, omdat de beïnvloedende factoren per (weinig voorkomend) geval verschillen. In alle analyses lijkt het model beter voorspellend als alle variabelen meegenomen worden dan wanneer alleen een selectie van variabelen op basis van de literatuur wordt meegenomen.

Uitkomsten sociaal emotionele gezondheid

In het datagedreven model kwam de WOZ-waarde van het huis waar de jeugdige woont, naar voren als een voorspellende factor voor een risico op psychosociale problemen. Een aantal buurtvariabelen zijn 'best of the rest' en hebben een voorspellende waarde van rond de 4%. Aan dergelijke beperkte percentages kunnen geen conclusies verbonden worden.

In het geval van de uitkomst sociaal emotionele gezondheid (risico op psychosociale problemen) voorspelt een model met variabelen op basis van literatuur beter de risicoscores tegenover een model met alle beschikbare variabelen, maar de positieven (kinderen met risico op psychosociale problemen) worden wel beter voorspeld.

WOZ-waarde als voorspeller?

Een factor die in dit onderzoek een voorspellende waarde lijkt te hebben is de WOZ-waarde van het huis waar een jongere in woont. Hoewel de WOZ-waarde natuurlijk geen directe invloed heeft op de sociaal emotionele gezondheid, kan deze factor gezien worden als een proxy voor het gezinsinkomen. Bij deze data analyse waren er geen gegevens beschikbaar over het precieze gezinsinkomen, enkel of het gezin onder de grens van het sociaal minimumloon zat of niet. Met aanvullende statistische methoden is vervolgens gekeken of er een associatie is tussen de WOZ-waarde van het huis waarin een jongere woont en de sociaal emotionele gezondheid, maar deze associatie bleek niet significant te zijn. Dit wijst erop dat deze relatie niet lineair is en waarschijnlijk complexer dan een directe relatie tussen WOZ-waarde ('inkomen') en sociaal emotionele gezondheid.

Uitkomsten delinquentie

Voor het risico op het plegen van delicten komen in deze analyse de WOZ-waarde, de studierichting VMBO en geslacht (man) naar voren als meest voorspellende variabelen. Ook hier geldt dat de voorspellende waarde van de andere variabelen dan WOZ-waarde slechts rond de 4% ligt en daarmee conclusies niet getrokken kunnen worden.

Uitkomsten schoolsucces

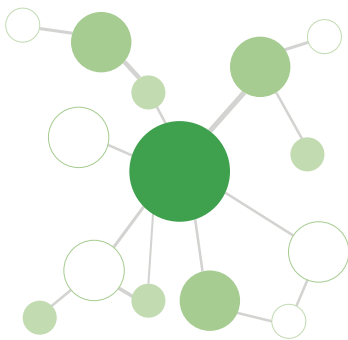
In deze analyse is de kans om te blijven zitten en/of op-/afstroom in kaart gebracht. Voortijdig schoolverlaten kon niet worden voorspeld omdat er een te beperkt aantal positieve uitkomsten in de huidige dataset zit. Ook hier geldt dat de WOZ-waarde als meest voorspellend naar voren komt en 'best of the rest' variabelen een te lage voorspellende waarde hebben om conclusies aan te verbinden.

Potentie

In deze data science analyse is een data gedreven model gemaakt, waarmee het mogelijk is om uitkomsten te voorspellen en daarbij samenhang tussen factoren in beeld te krijgen. De uitkomsten krijgen meer zeggingskracht als we data verder verrijken, in de preparatie de datasets verder analyseren en daarna vervolg analyses doen. Dit verrijken kan door het toevoegen van jaren waarover data beschikbaar zijn vanuit de verschillende bronnen. Verder is nu gebruik gemaakt van nu beschikbare indicatoren per factor. Op een aantal factoren is geen of geen optimale indicator beschikbaar. Dit betekent dat deze factor in de analyse nooit naar voren kan komen als bepalend. Op korte

termijn kunnen indicatoren worden toegevoegd aan de datasets. Daarnaast zijn op een aantal factoren juist meerdere indicatoren beschikbaar. Nadere verkenning van de relaties tussen verschillende indicatoren is nodig om optimaal gebruik te maken van de beschikbare gegevens. Dit draagt ook bij aan meer balans in de datasets als het gaat om het aantal en de waarde van indicatoren per factor. Samenwerking met wetenschappelijke partners en partners waar data verkregen worden leidt op termijn tot meer en steeds betere indicatoren. Met deze ingrediënten werken we verder aan een analyse met een steeds sterkere zeggingskracht.

7. Toekomst



Afb. 3: Ecosysteem van factoren

De uitkomsten uit de data science analyse van de gemeente Rotterdam bieden aanknopingspunten voor de toekomst. De basis van wetenschappelijke kennis die Rotterdam heeft ontsloten in haar factorenmodel kan verder verrijkt worden met actuelere inzichten uit de lokale context. Zo kan meer kennis verworven worden over de samenhang van bepaalde factoren. Zeker als dit soort data science analyses op structurele basis worden herhaald met een nog bredere dataset, legt dit een ecosysteem van factoren bloot. Dit geeft vervolgens meer inzicht over oorsprong van bepaalde problemen of kansen, waardoor helder wordt waar preventievere maatregelen, acties en interventies het meest kansrijk zijn. Daarmee kan beleid zowel effectiever als efficiënter worden. En dat leidt weer tot een sterkere bijdrage aan het kansrijk, veilig en gezond opgroeien van kinderen en jongeren in de stad.

Juli 2019

De factsheet Data science analyse Rotterdam Groeit is een uitgave van de Gemeente Rotterdam.

Auteurs: Rina Stoorvogel, Denis Wiering, Anne Fleur van Veenstra en Maaïke de Boer.

Contact en informatie: www.rotterdam.nl/jeugd, jeugd@rotterdam.nl