



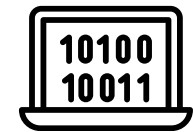
DATAFRYSLÂN
BRON VAN INZICHT

/Project Draagkracht en Draaglast

Nynke Slagter, Anita Nguyen en Carlos de Matos Fernandes



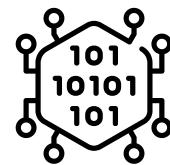
INLEIDING



'HARDE CIJFERS'



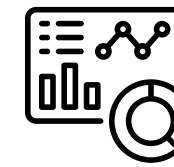
DATA



VOORSPELMODELLEN



ANALYSES



UITKOMSTEN



INLEIDING



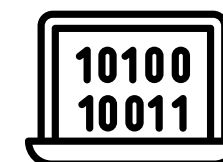
INLEIDING

HET PROJECT DRAAGKRACHT EN DRAAGLAST

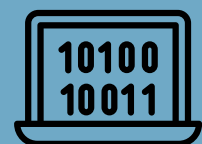
Om beter inzicht te krijgen in de leefbaarheid van buurten, wijken en dorpen is het project **Draagkracht en Draaglast gestart**, een initiatief van Huurdersbelang Fryslân en Planbureau Fryslân.

Het doel is het ontwikkelen van een uniform meetinstrument dat op lokaal niveau de **balans tussen draagkracht** (hulpbronnen, sterktes) **en draaglast** (druk, problematiek) **zichtbaar maakt**. Daarbij wordt niet alleen gekeken naar objectieve registratiedata, maar ook naar de subjectieve beleving van inwoners en de inzichten van professionals.

Het project Draagkracht en Draaglast rust op **drie samenhangende pijlers**. De kracht van het project zit in de samenhang tussen deze drie pijlers. Waar **pijler 1 ‘harde data’ en modellen** levert, zorgt **pijler 2** voor de noodzakelijke **verdieping vanuit de leefwereld van bewoners**. **Pijler 3 vormt de brug naar toepassing en besluitvorming**.



'HARDE CIJFERS'



'HARDE CIJFERS'

DE ROL VAN DATAFRYSLÂN IN HET PROJECT DRAAGKRACHT EN DRAAGLAST

DataFryslân is gevraagd om de uitvoering van Pijler 1 van het project Draagkracht en Draaglast op zich te nemen, met als doel **het ontwikkelen van een kwantitatieve basis voor het meten van ervaren leefbaarheid**. Binnen deze pijler vormt het gebruik van kwantitatieve data uit registraties de kern van de aanpak. De primaire bron van deze gegevens is **CBS Microdata**, waarin persoonsgegevens op verantwoorde wijze beschikbaar worden gesteld voor onderzoek.

In een experimentele setting voert DataFryslân een **modelmatige analyse** uit, gericht op het **inzichtelijk maken** van de bijdrage van afzonderlijke **kenmerken aan de tevredenheid met de woonomgeving**. Daarbij wordt niet alleen het model zelf ontwikkeld, maar ook kritisch gekeken naar de exacte methode en de zeggingskracht van de gebruikte analysetechnieken.

Het **resultaat is een kwantitatief basismodel** dat dient als fundament voor de verdere uitwerking van deze pijler in toekomstige pilotstudies. Hiermee levert DataFryslân een belangrijke bijdrage aan de onderbouwing van beleidsontwikkeling met behulp van robuuste, datagedreven inzichten.



DATA



DATA

WAT IS CBS MICRODATA?

CBS Microdata zijn gedetailleerde, op persoonsniveau verzamelde gegevens die door het Centraal Bureau voor de Statistiek (CBS) beschikbaar worden gesteld voor wetenschappelijk en beleidsgericht onderzoek. Deze data bieden een rijke bron van informatie over uiteenlopende maatschappelijke thema's, zoals werk, inkomen, gezondheid en wonen.

Het gebruik van CBS Microdata is strikt gereguleerd. De gegevens zijn **niet herleidbaar tot individuen** en vallen onder **streng privacyregels**. Alleen erkende **onderzoeksinstellingen** kunnen toegang krijgen, en uitsluitend via een **beveiligde remote-access omgeving** van het CBS.

Onderzoekers moeten vooraf een toets doorlopen om toegang te verkrijgen. Daarnaast wordt alle output die het systeem verlaat, **gecontroleerd door het CBS** om te waarborgen dat er geen vertrouwelijke informatie wordt gedeeld.

Een belangrijk voordeel van CBS Microdata is de mogelijkheid tot het **koppelen van verschillende datasets**, waardoor diepgaande en domeinoverstijgende analyses mogelijk worden. Deze combinatie van detailniveau, veiligheid en flexibiliteit maakt CBS Microdata tot een krachtige en betrouwbare bron voor onderzoek en beleidsontwikkeling.



DATA

DATASETS DIE TEN GRONDSLAG LIGGEN AAN DE ANALYSES

Voor de analyse zijn twee databronnen van uit de **CBS microdata** omgeving de basis: de **Woonbase** en het **Woononderzoek Nederland (WoON)**.

De **Woonbase** is een registergebaseerde databank die informatie bevat over woningen, personen en huishoudens in Nederland. Deze dataset is opgebouwd uit koppelingen tussen verschillende registraties, zoals de Basisregistratie Personen (BRP) en de Basisregistratie Adressen en Gebouwen (BAG). De **Woonbase** biedt:

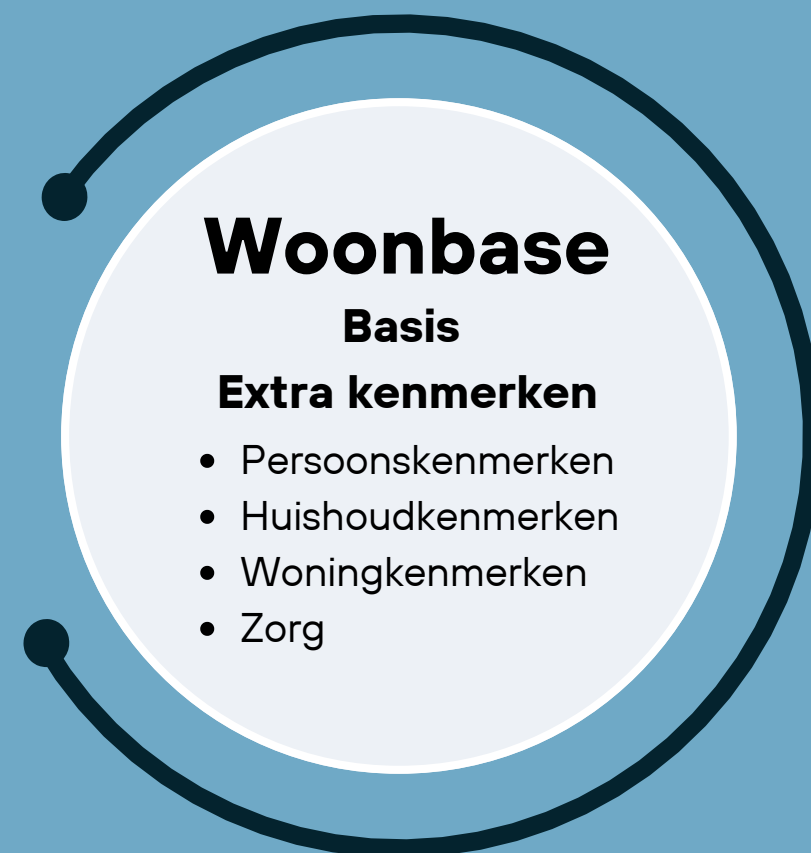
- Volledige populatiegegevens (geen steekproef)
- Informatie over woningkenmerken, huishoudens en verhuisbewegingen
- Mogelijkheid tot koppeling van datasets
- Toegang via een beveiligde remote-access omgeving

Het **WoON** onderzoek is een grootschalig enquêteonderzoek dat periodiek wordt uitgevoerd onder huishoudens in Nederland. Het richt zich op de beleving van de woonomgeving en woonwensen van bewoners. In 2024 bevatte **WoON** 41.295 respondenten.

Belangrijke kenmerken zijn:

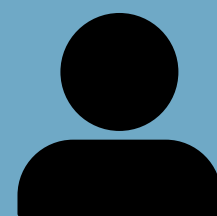
- Representatieve steekproef van huishoudens
- Informatie over woontevredenheid, verhuiscriteria en buurtbeleving
- Subjectieve beleving als aanvulling op objectieve registratiedata

Door deze twee databronnen te combineren ontstaat een rijk en veelzijdig beeld van zowel de feitelijke als de ervaren woonsituatie in Nederland.



DATA

INDICATOREN VAN DE WOONBASE



Persoonskenmerken

- Geslacht
- Burgerlijke staat
- Opleidingsniveau
- Positie ten opzichte van de hoofdkostwinner



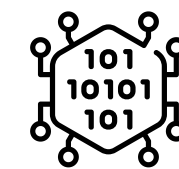
Huishoudkenmerken

- Aantal personen per huishouden
- Type huishouden
- Inkomen ten opzichte van de armoedegrens
- Belastbaar inkomen
- Sociaal economische categorie



Woningkenmerken

- Woningtype
- Bouwjaar (klasse)
- Oppervlakte (klasse)
- WOZ-waarde (klasse)



VOORSPELMODELLEN



VOORSPELMODELLEN

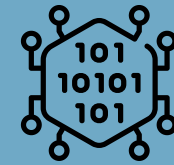
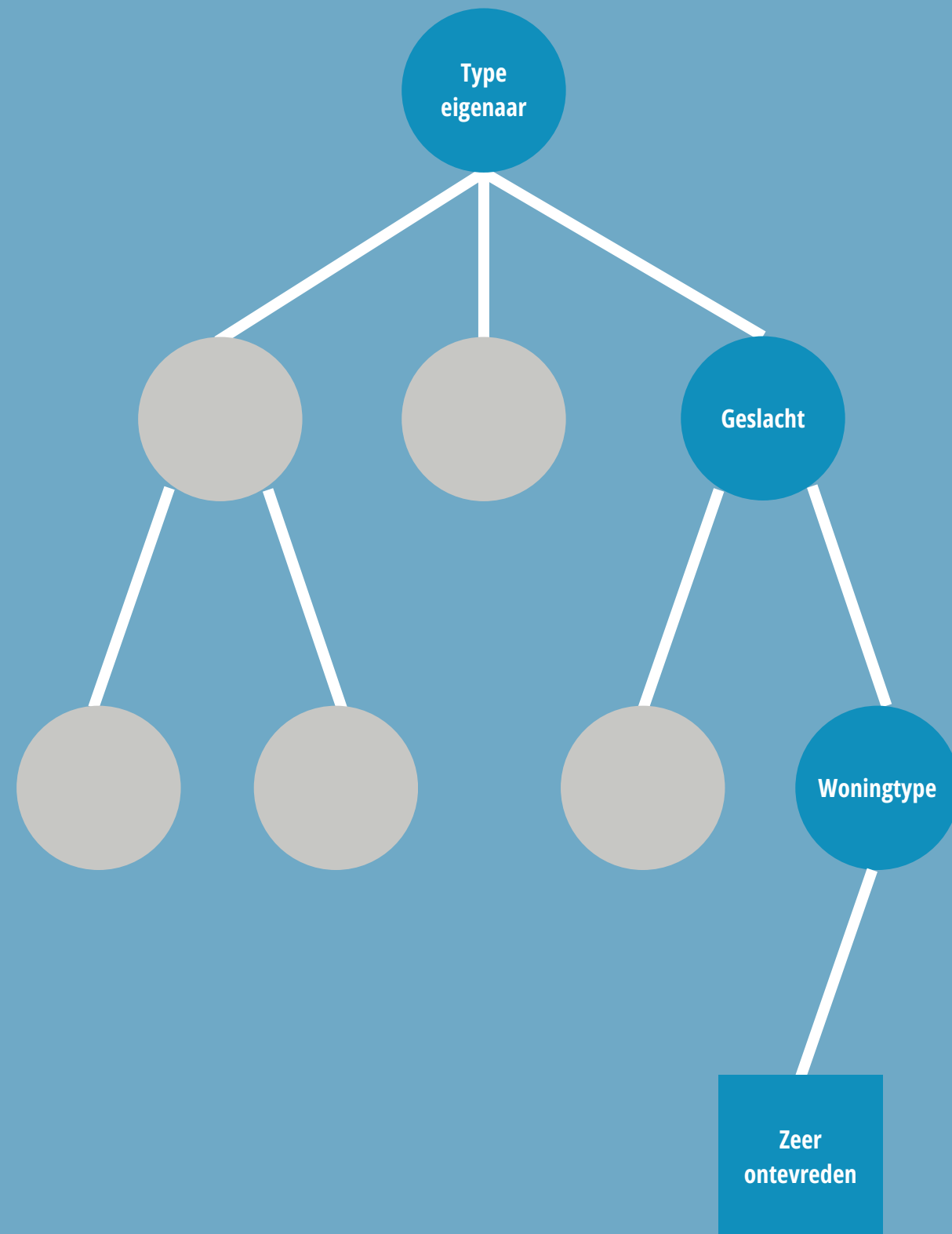
WAAROM GEBRUIKEN WE VOOR DEZE ANALYSES VOORSPELMODELLEN?

Een **voorspelmodel** is een analysemethode die patronen in bestaande data identificeert om een waarde of uitkomst te schatten. Dit type model wordt vaak ingezet om gedrag of behoeften te voorspellen, bijvoorbeeld in het onderwijs, de zorg of mobiliteit.

Voorbeelden van toepassingen zijn het voorspellen van leerachterstanden, zorggebruik, verhuizingen, het weer, verkeersdrukke en de kans op energiearmoede. Deze modellen bieden waardevolle inzichten die kunnen bijdragen aan gericht beleid en efficiëntere dienstverlening.

Een belangrijk voordeel van voorspelmodellen is dat er vooraf geen selectie van variabelen nodig is: **het model bepaalt zelf welke gegevens relevant zijn**. Daarbij wordt alle beschikbare data benut, inclusief onverwachte verbanden die anders mogelijk over het hoofd worden gezien. **Voorspelmodellen kunnen bovendien omgaan met verschillende soorten data, zoals numerieke waarden en categorische gegevens**. Hierdoor is de analyse minder afhankelijk van aannames vooraf.

BESLISBOOM



VOORSPELMODELLEN

HOE WERKEN VOORSPELMODELLEN OP BASIS VAN BESLISBOMEN?

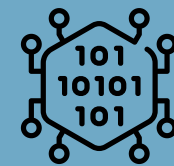
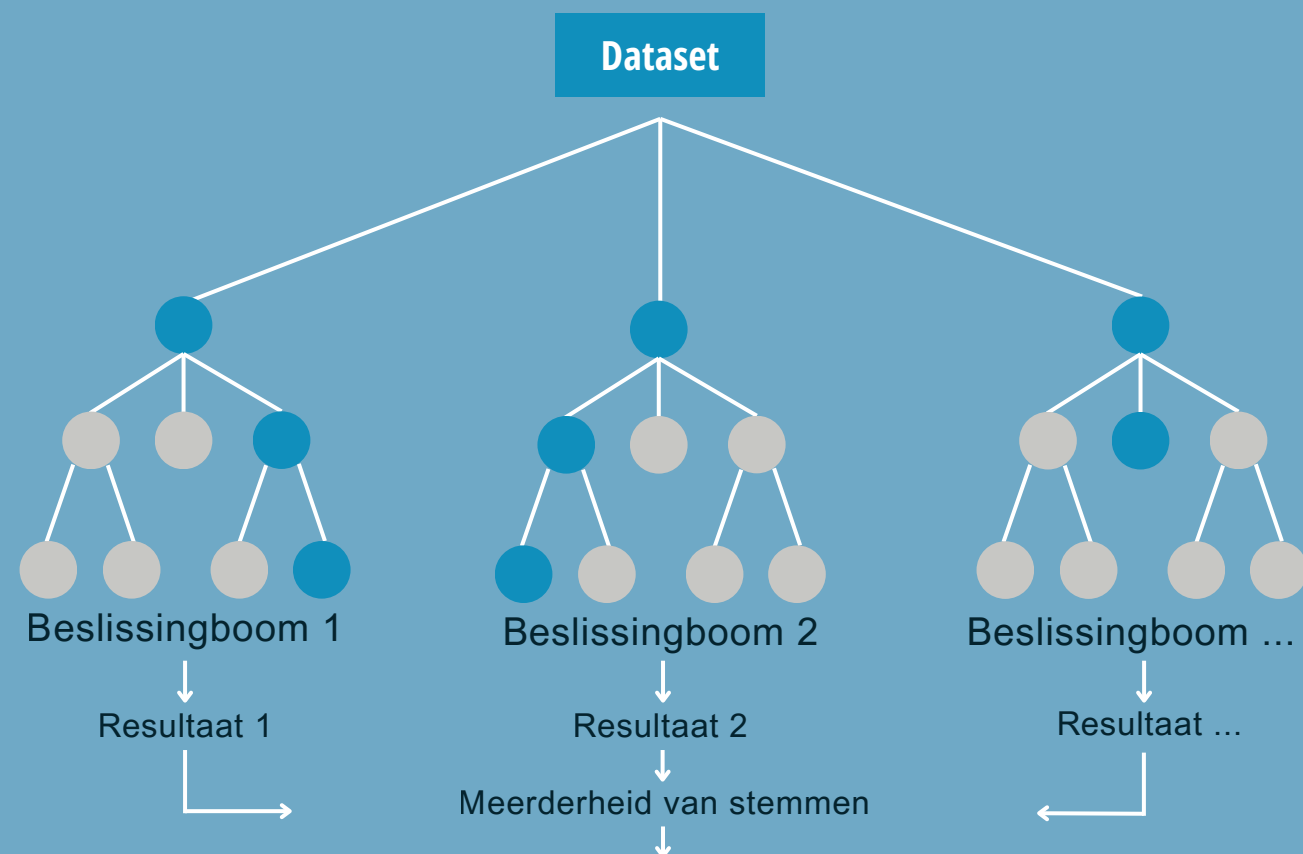
Een beslisboom is een model dat wordt gebruikt om een **doelvariabele** met verschillende categorieën te voorspellen, zoals **tevredenheid met de woonomgeving**. De boom bestaat uit een reeks **knooppunten** en **vertakkingen**. Elk **knooppunt** **vertegenwoordigt een specifiek kenmerk** uit de dataset, zoals leeftijd, inkomen of woningtype. Op basis van de waarde van dat kenmerk splitst de boom zich in vertakkingen naar andere knooppunten.

Aan het einde van elk pad in de boom bevindt zich een **blad**, dat wordt **gelabeld met een categorie van de doelvariabele**, bijvoorbeeld "tevreden", "neutraal" of "ontevreden".

Het model leert deze structuur door patronen in de data te analyseren. Tijdens het leerproces bepaalt het model zelf welke kenmerken het meest informatief zijn en kiest het automatisch de knooppunten die leiden tot de beste voorspellingen.

Beslisbomen vormen de basis voor geavanceerdere machine learning-modellen zoals Random Forest en Gradient Boosting, die meerdere bomen combineren voor hogere nauwkeurigheid en robuustheid.

RANDOM FOREST MODEL



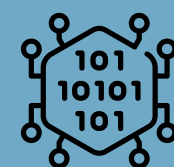
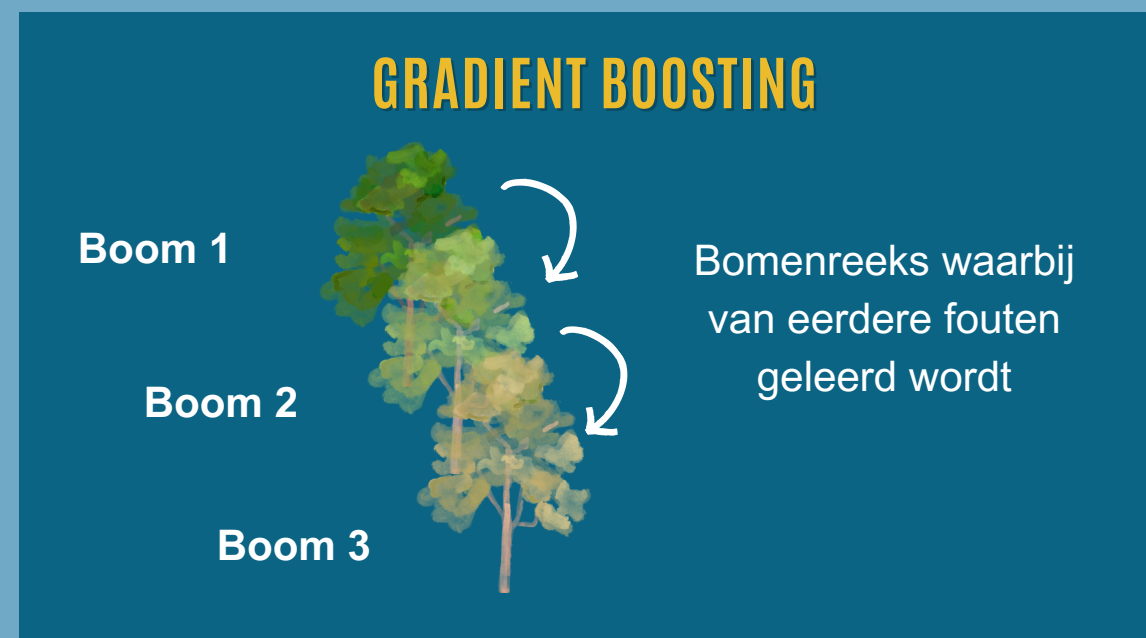
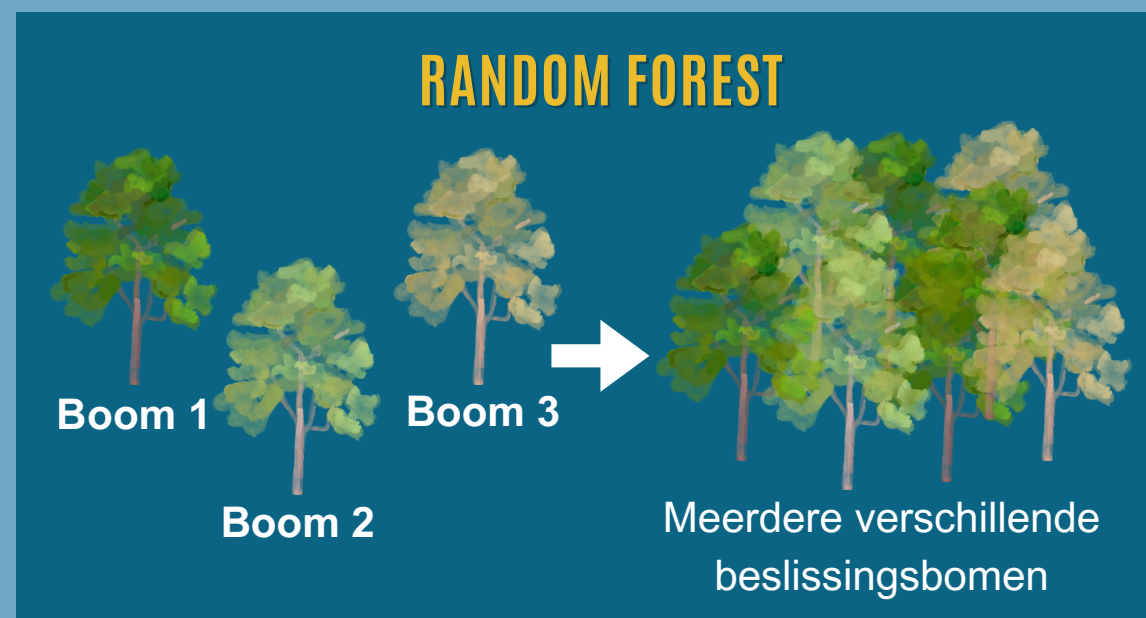
VOORSPELMODELLEN

HOE WERKT EEN RANDOM FOREST MODEL?

Een **Random Forest** is een **machine learning model** dat **voorspellingen doet op basis van meerdere beslissingsbomen**. Elke boom in het model maakt een inschatting door vragen te stellen over de gegevens, zoals bijvoorbeeld leeftijd, inkomen of woonomgeving. In plaats van één enkele boom, gebruikt een random forest **veel bomen tegelijk**, waardoor het model **robuuster en nauwkeuriger** wordt.

Elke boom wordt getraind op een **willekeurige subset van de data**, zodat ze allemaal iets anders leren. Wanneer het model een nieuwe situatie moet inschatten, zoals of iemand waarschijnlijk ontevreden is over de woonomgeving, dan **stemmen alle bomen**. De uiteindelijke voorspelling wordt bepaald door de **meerderheid van de bomen**: als de meeste bomen zeggen dat iemand ontevreden is, dan is dat de voorspelling van het model.

Door deze aanpak kan het Random Forest model goed omgaan met complexe en diverse gegevens, en betrouwbare voorspellingen doen in nieuwe situaties.



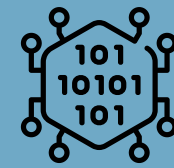
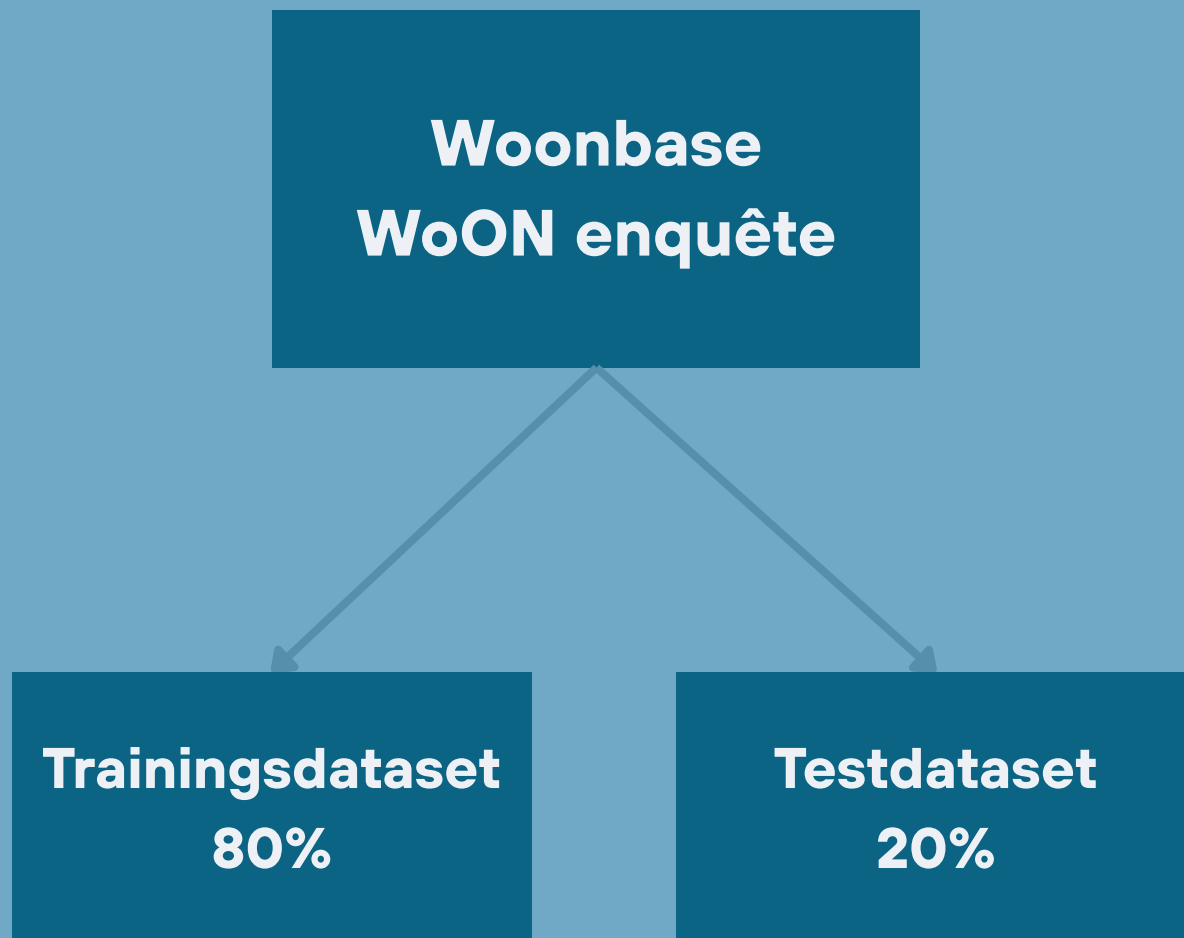
VOORSPELMODELLEN

WAT IS HET VERSCHIL TUSSEN EEN RANDOM FOREST MODEL EN EEN GRADIENT BOOSTING MODEL?

Random Forest en Gradient Boosting zijn beide modellen die gebruikmaken van **beslissingsbomen**, maar ze werken op een andere manier.

Bij **Random Forest** worden **meerdere bomen tegelijk gebouwd**, waarbij elke boom een **willekeurige subset van de data** gebruikt. De uiteindelijke voorspelling wordt bepaald door de **meerderheid van de bomen**.

Bij Gradient Boosting worden de bomen juist **één voor één gebouwd**, waarbij **elke boom leert van de fouten van de vorige**. In dit geval wordt **de volledige dataset** gebruikt, maar met **extra aandacht voor de moeilijk te voorspellen gevallen**. Hierdoor is Gradient Boosting vaak **nauwkeuriger**, maar ook gevoeliger voor **overfitting**, omdat het risico bestaat dat het te veel details uit de trainingsdata gaat volgen.



VOORSPELMODELLEN

DE DATA OM EEN MODEL OP TE TRAINEN

Om een voorspelmodel zoals **Random Forest** of **Gradient Boosting** te trainen, wordt de beschikbare dataset eerst **opgesplitst in twee delen**: een **trainingsdataset** en een **testdataset**.

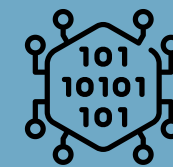
De **trainingsdataset** wordt gebruikt om het model **te leren welke patronen er in de data** zitten, bijvoorbeeld welke kenmerken samenhangen met tevredenheid met de woonomgeving. Op basis van deze gegevens leert het model de relaties tussen de verklarende variabelen en de doelvariabele.

De **testdataset** wordt apart gehouden en pas gebruikt **nadat het model is getraind**. Hiermee wordt beoordeeld **hoe goed het model in staat is om nieuwe, onbekende gevallen te voorspellen**. Door deze aanpak wordt voorkomen dat het model alleen maar goed presteert op de data waarop het is getraind en ontstaat er een realistischer beeld van de kwaliteit van de voorspellingen.

Deze werkwijze is essentieel om **overfitting te voorkomen** en om te zorgen dat het model ook buiten de trainingsdata bruikbaar is, bijvoorbeeld voor het inschatten van leefbaarheid in andere buurten of toekomstige situaties.

CONFUSION MATRIX

	Terecht positief	Onterecht positief
	Onterecht negatief	Terecht negatief



VOORSPELMODELLEN

WAT ZIJN DE UITKOMSTEN VAN VOORSPELMODELLEN?

Voorspelmodellen zoals **Random Forest** en **Gradient Boosting** doen **voorspellingen op basis van patronen** in de data, bijvoorbeeld over (on)tevredenheid over de woonomgeving. De modellen geven inzicht in hoe goed ze daadwerkelijk voorspellen door middel van **validatie op de testdataset**.

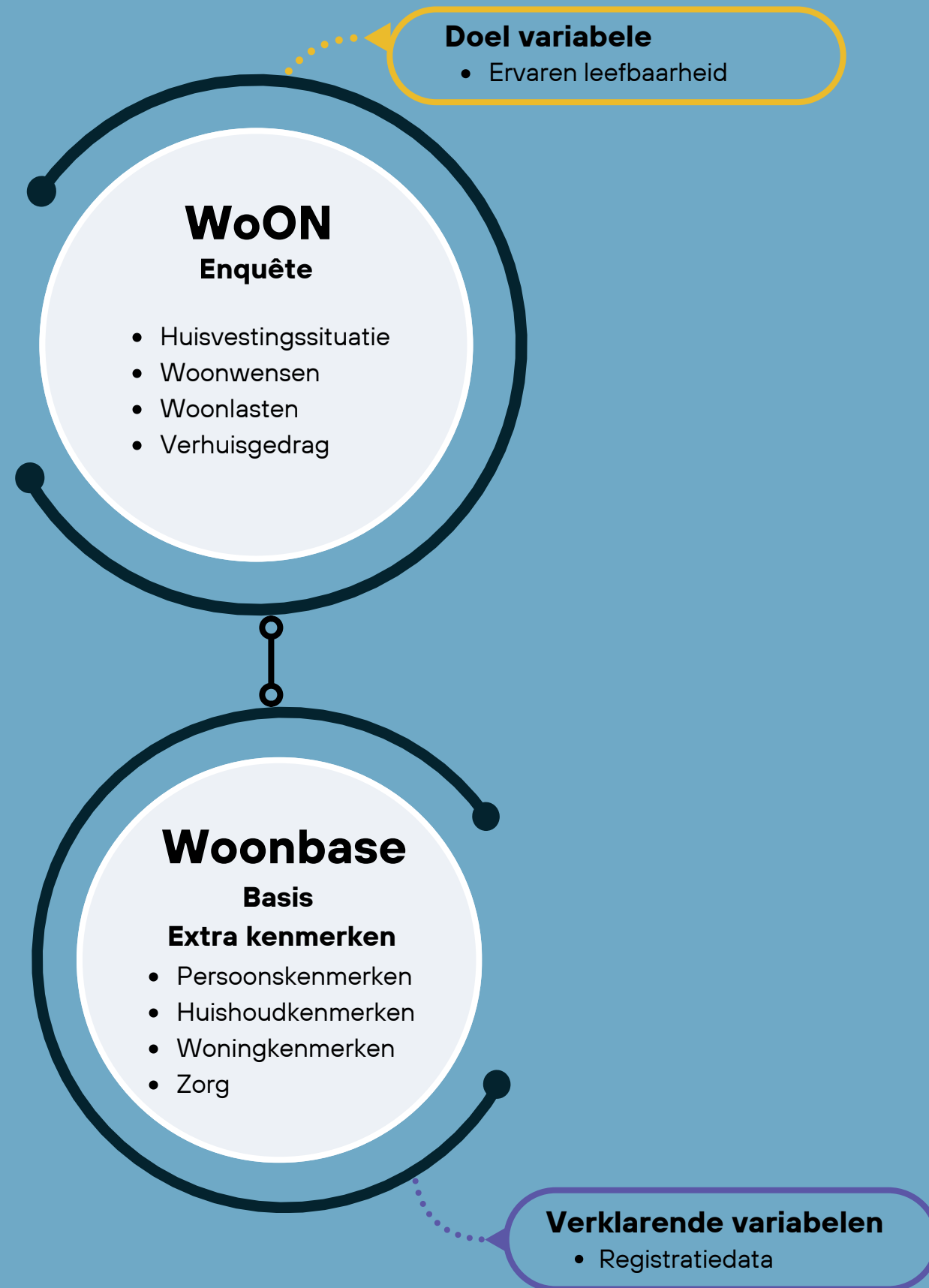
De **nauwkeurigheid** van het model is het **percentage correcte voorspelde uitkomsten**, dit kan worden weergegeven met behulp van een confusion matrix. Een **confusion matrix** geeft in één overzicht weer hoe goed een classificatiemodel presteert door het aantal juiste en foutieve voorspellingen per klasse te tonen.

Naast het geven van een voorspelling, bieden voorspelmodellen ook **inzicht in welke indicatoren het meest bijdragen aan de uitkomst**. Dit gebeurt zonder dat vooraf keuzes hoeven te worden gemaakt over welke variabelen belangrijk zijn. Met behulp van **SHAP-waardes** krijgt elke variabele een kwantitatieve score, die aangeeft **hoeveel invloed die variabele heeft op de voorspelling**, rekening houdend met de impact van andere indicatoren.

Tot slot is de **kwaliteit van de voorspelling sterk afhankelijk van de kwaliteit van de data**. Onvolledige, onnauwkeurige of scheef verdeelde data kunnen leiden tot misleidende uitkomsten. Daarom is een zorgvuldige voorbereiding en evaluatie van de data essentieel bij het gebruik van deze modellen.



ANALYSES



ANALYSES

ANALYSE 1: VOORSPELMODEL

Het doel van analyse 1 is om **tevredenheid met de woonomgeving te voorspellen in Fryslân** op basis van indicatoren uit registratiedata (Woonbase).

De **doelvariabele**, die aangeeft hoe inwoners hun woonomgeving ervaren, is afkomstig uit de **WoON enquête**. Omdat het aantal respondenten beperkt is, is ervoor gekozen om het model te trainen op alle beschikbare respondenten, en niet uitsluitend op Friese respondenten.

De **verklarende variabelen**, zoals woningkenmerken, buurtkenmerken en demografische gegevens, komen uit de **Woonbase**. Deze dataset bevat **registratiedata die beschikbaar is voor alle inwoners van Fryslân**. Voor analyse 1 zijn er maximaal 59 voorspellende variabelen gebruikt.

Wanneer uit de validatie blijkt dat het model **voldoende voorspellende kracht** heeft, kan het worden toegepast op de Woonbase. Daarmee kan het model een inschatting maken van de tevredenheid met de woonomgeving voor **alle inwoners van Fryslân**, ook voor degenen die niet hebben deelgenomen aan de enquête.

Deze aanpak maakt het mogelijk om op basis van objectieve registratiedata een breed beeld te krijgen van de leefbaarheid in de provincie, en biedt waardevolle inzichten voor beleid en gebiedsontwikkeling.



ANALYSES

ANALYSE 2: RELEVANTE FACTOREN

Het doel van analyse 2 is om te onderzoeken **welke factoren relevant zijn voor het voorspellen van tevredenheid met de woonomgeving**.

Hiervoor wordt voortgebouwd op **analyse 1**, waarbij het model wordt uitgebreid met aanvullende verklarende variabelen uit **WoON**. Omdat deze data niet beschikbaar is voor alle inwoners van Fryslân, kunnen deze variabelen **alleen worden gebruikt om het model te trainen**, en niet om voorspellingen te doen voor de gehele populatie.

Naast de enquêtegegevens uit de dataset **WoON** worden ook variabelen uit de **Woonbase** meegenomen. In sommige gevallen zijn variabelen uit de WoON-enquête **geïmputeerd** om ontbrekende waarden aan te vullen. Ook worden er **samengestelde variabelen** gecreëerd om complexere concepten zoals woonkwaliteit of buurtvoorzieningen beter te kunnen modelleren.

Voor analyse 2 zijn maximaal 719 indicatoren meegenomen, waarbij de meeste varianten in de basis gebruikmaakten van maximaal 197 variabelen.

Omdat de data niet beschikbaar is voor de volledige Friese populatie, wordt er **geen voorspelling gedaan van tevredenheid met de woonomgeving in Fryslân**. Wel biedt de analyse inzicht in **welke factoren het meest bijdragen aan (on)tevredenheid**, wat waardevol is voor het verder ontwikkelen van een breed toepasbaar voorspelmodel.



ANALYSES

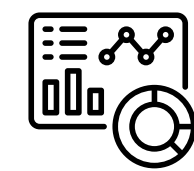
ANALYSE 3: DE INVLOED VAN BUURTINDICATOREN

In **analyse 3** wordt het voorspelmodel verder uitgebreid door **buurtkenmerken toe te voegen** als verklarende variabelen. Deze kenmerken zijn gebaseerd op registratiedata en komen uit de StatLine-tabel **Kerncijfers Wijken en Buurten 2023**. Voorbeelden van mogelijke buurtkenmerken zijn het percentage sociale huurwoningen, de gemiddelde WOZ-waarde van woningen en het gemiddelde inkomen van bewoners in een buurt.

Deze buurtkenmerken worden toegevoegd aan het bestaande model dat in analyse 1 en 2 is opgebouwd. In analyse 1 werd gestart met basisvariabelen uit de **Woonbase** en in analyse 2 zijn aanvullende variabelen uit de dataset **WoON** gebruikt om het model te trainen. De buurtkenmerken uit Statline zijn daarentegen wél beschikbaar voor iedereen in Fryslân, waardoor ze een belangrijke rol kunnen spelen in het verder ontwikkelen van een breed toepasbaar model.

Voor analyse 3 zijn er maximaal 205 indicatoren meegenomen in de analyse.

Het **doel van analyse 3** is om te **onderzoeken in hoeverre deze buurtkenmerken bijdragen aan de voorspelling van tevredenheid met de woonomgeving** en of hun toevoeging leidt tot een verbetering van de modelprestaties.



UITKOMSTEN

VERDELING TEVREDEN MET DE WOONOMGEVING

Categorie	Aantal	Percentage
Zeervredenen	12.653	30,64%
Tevreden	22.041	53,37%
Neutraal	4.367	10,58%
Ontevreden	1.674	4,05%
Zeervredenen	560	1,36%
Totaal	41.295	100,00%



UITKOMSTEN

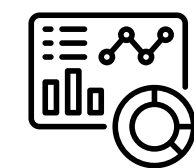
TEVREDEN MET DE WOONOMGEVING

De verdeling van de **doelvariabele tevreden met de woonomgeving** laat zien dat dit een **ongebalanceerde** variabele is. Meer dan de helft van de respondenten is tevreden met de woonomgeving, bijna een derde is zelfs zeer tevreden, maar de percentages van de respondenten die neutraal of (zeer) ontevreden als antwoord hebben ingevuld zijn een stuk lager.

Een ongebalanceerde doelvariabele kan ertoe leiden dat het model vooral leert om de **meerderheidsklasse te voorspellen**, en **minder goed presteert bij het herkennen van de minder vaak voorkomende groep**.

De Random Forest en de Gradient Boosting methodes kunnen omgaan met ongebalanceerde data, maar vereisen wel extra aandacht. Er zijn **verschillende technieken om dit te corrigeren**.

Vooraf bij Gradient Boosting is het belangrijk om te voorkomen dat het model zich te sterk richt op de dominante klasse, omdat het stapsgewijs leert van fouten en daardoor gevoelig is voor scheve verdelingen.



UITKOMSTEN ANALYSE 1



UITKOMSTEN ANALYSE 1

HET PROCES VAN ANALYSE 1

Voor analyse 1 is gebruikgemaakt van **registratiedata uit de Woonbase** om **tevredenheid met de woonomgeving te voorspellen** met behulp van verschillende voorspelmodellen.

Voorafgaand aan de analyse zijn een groot aantal **verklarende variabelen geselecteerd**. Deze variabelen, of subsets daarvan, zijn gebruikt in twee verschillende machine learning modellen, namelijk **het Random Forest en het Gradient Boosting model**. Ter vergelijking is ook een klassiek statistisch model toegepast in de vorm van een **logistische regressie**.

Gedurende het analyseproces is de focus komen te liggen op de Gradient Boosting methode. Binnen deze aanpak zijn **meerdere variabelenselecties getest**, is **geëxperimenteerd met verschillende hyperparameters**, en zijn de **uitkomstklassen samengevoegd** om de prestaties van het model verder te onderzoeken.

CONFUSION MATRIX

Werkelijke waarde	Ze er ontevreden	Ontevreden	Neutraal	Tevreden	Ze er tevreden
	3	307	0	0	0
	3	307	0	0	0
	24	858	0	0	0
	251	4147	0	0	0
	291	2257	0	0	0
Voorspelde waarde					
Ze er tevreden Tevreden Neutraal Ontevreden Ze er ontevreden					



UITKOMSTEN ANALYSE 1

HOE GOED VOORSPELT HET MODEL?

De dataset is opgedeeld in een trainingsdataset en een testdataset. De **trainingsdataset** bevat **80% van de observaties**. Hier worden de modellen op getraind. De **testdataset** bevat **20% van de observaties**. Dat wil zeggen dat het getrainde model wordt getest op $20\% \times 41.295 = 8.259$ observaties. De **testdataset** wordt gebruikt om de **voorspellingen te vergelijken met de echte uitkomsten**.

Hiernaast wordt de confusion matrix weergegeven van een **Gradient Boosting**-model (XGBoost), getraind op een **geselecteerde set variabelen**, waarbij de **hyperparameters zijn geoptimaliseerd** voor maximale modelprestaties.

De **confusion matrix** laat op de **horizontale as** de door het model **voorspelde waarden** zien en op de **verticale as** de **werkelijke waarden** uit de testdataset. In 291 gevallen voorspelt het model de uitkomst **Ze
er
tevreden** correct en in 4.147 gevallen de uitkomst **Tevreden**. De overige uitkomsten worden helemaal niet voorspeld door het model. Dit is het gevolg van de **ongebalanceerde verdeling van de doelvariabele**.

Dit model voorspelt in $(4.147 + 291 + 0 + 0 + 0) / 8.259 \times 100\% = 53,7\%$ de uitkomst correct.

SHAP-WAARDEN



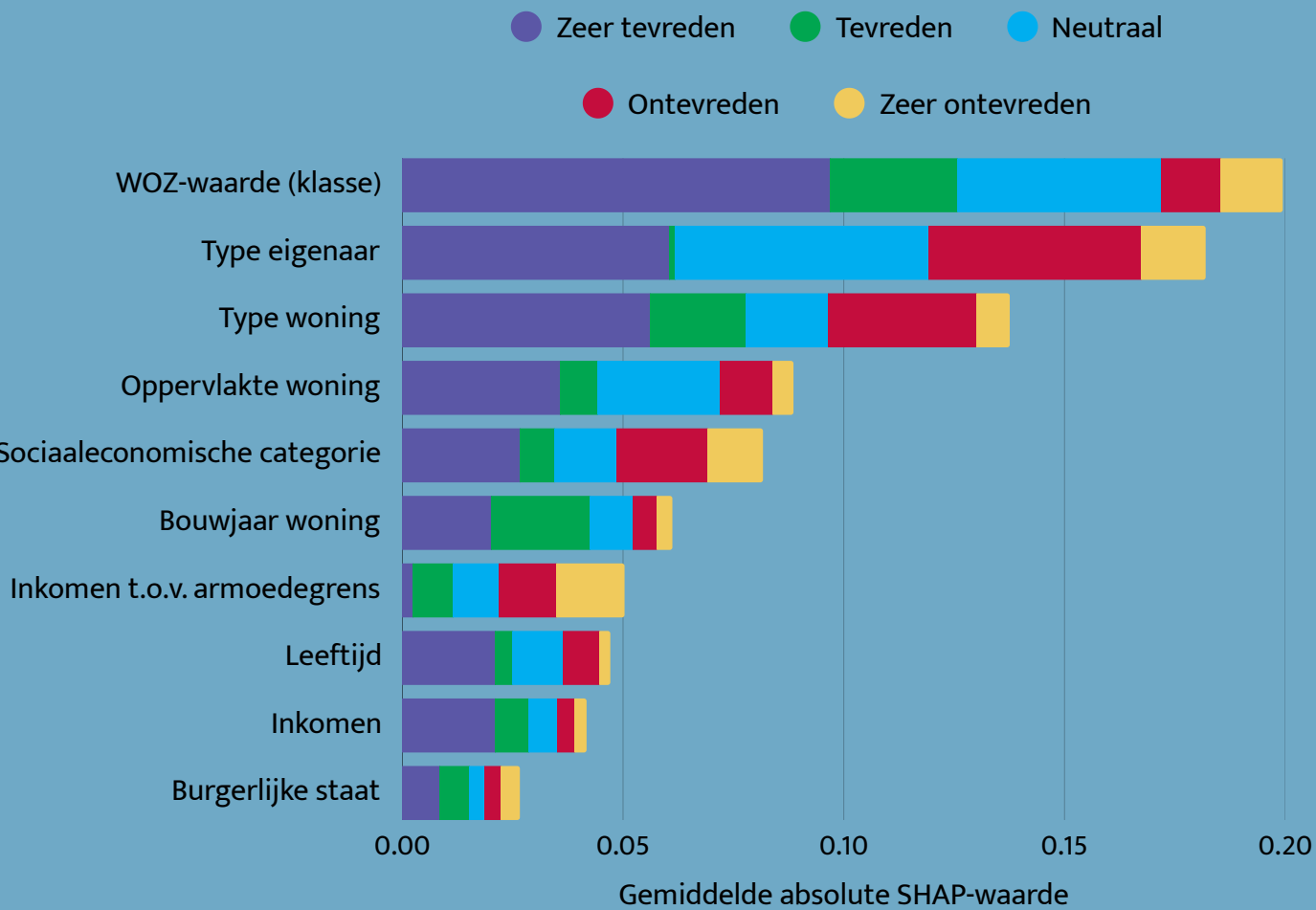
UITKOMSTEN ANALYSE 1

WAT ZIJN SHAP-WAARDEN?

SHAP-waarden geven inzicht in hoe **elk kenmerk (feature)** **bijdraagt aan een individuele voorspelling van een model**. Ze zijn gebaseerd op het idee van Shapley-waarden uit de speltheorie, waarbij de invloed van een kenmerk wordt berekend door te kijken naar alle mogelijke combinaties waarin dat kenmerk kan worden weggelaten. Dit maakt het mogelijk om **interacties tussen variabelen** mee te nemen in de analyse.

Voor elke voorspelling laat SHAP zien hoeveel elk kenmerk bijdraagt aan het verschil tussen de **gemiddelde voorspelling** en de **specifieke uitkomst**. Door het **gemiddelde van alle individuele voorspellingen** te nemen, krijg je een globaal beeld van de belangrijkste kenmerken in het model. Zo kun je niet alleen zien **welke kenmerken het meest invloedrijk zijn**, maar ook hoe ze samenwerken en elkaar beïnvloeden.

GEMIDDELDE IMPACT (SHAP-WAARDE) OP TEVREDENHEID MET DE WOONOMGEVING



UITKOMSTEN ANALYSE 1

WELKE FACTOREN ZIJN BELANGRIJKE VOORSPELLERS?

In het figuur kan je de **10 belangrijkste factoren gebaseerd op de SHAP-waarden** zien. Dit figuur geeft het **gemiddelde (absolute) effect per klasse** weer en laat het **totale effect over alle klassen** zien.

De belangrijkste factor om tevredenheid met de woonomgeving te voorspellen volgens dit model met deze variabelen is de **WOZ-waarde klasse van de woning**. Je ziet dat van deze factor de paarse balk het grootste is. Deze variabele heeft dus vooral een groot effect op het voorspellen van uitkomst "Zeer tevreden", maar de gele balk is relatief klein bij deze factor. De tweede factor, **type eigenaar** (huur of koop), heeft amper invloed op het inschatten van de uitkomst "tevreden", maar wel op de uitkomsten "Zeer tevreden", "Neutraal" en "Ontevreden".

Andere relevante factoren zijn **type woning** (Vrijstaande woning, hoekwoning, enz.), **inkomens variabelen** en **oppervlakte** en **bouwjaar van de woning** en het **geboortejaar van de respondent**.

SAMENVATTEND ANALYSE 1



UITKOMSTEN ANALYSE 1

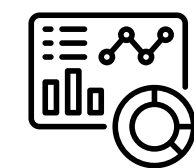
WAT HEBBEN WE GELEERD VAN ANALYSE 1?

Uit deze analyse blijkt dat het voorspellen van tevredenheid met de woonomgeving met de huidige aanpak beperkt effectief is. **De data is ongebalanceerd**, wat betekent dat bepaalde uitkomsten veel vaker voorkomen dan andere. Dit maakt het voor het model **lastig om de minder vaak voorkomende categorieën goed te herkennen**, wat ten koste gaat van de nauwkeurigheid en betrouwbaarheid van de voorspellingen.

Daarnaast laten de resultaten zien dat **verschillende voorspelmodellen vergelijkbare prestaties leveren**. Dit wijst erop dat de keuze van de methode niet doorslaggevend is voor de voorspelkracht. Ook het toepassen van methodes om machine learning modellen geschikt te maken voor ongebalanceerde data, net zoals het tunen van de **hyperparameters** hebben geen significant effect op de voorspelkracht. De beperkende factor lijkt eerder te liggen in de aard en kwaliteit van de beschikbare data.

Het belangrijkste inzicht is dat **de huidige registratiedata onvoldoende zijn om ervaren leefbaarheid goed te voorspellen**. Omdat het overgrote deel van de respondenten tevreden is, lukt het de modellen minder goed om de kleinere groepen met een neutrale of negatieve beleving te herkennen. Objectieve kenmerken zoals woning en persoonlijke kenmerken geven slechts een beperkt beeld van hoe mensen hun leefomgeving daadwerkelijk ervaren. Het **doel** om voor **heel Fryslân te voorspellen hoe tevreden mensen met hun woonomgeving zijn**, kan daarom niet worden uitgevoerd.

Er is behoefte aan aanvullende data die subjectieve beleving, sociale cohesie, veiligheid en andere contextuele factoren beter in kaart brengen. Deze uitbreiding wordt gerealiseerd met analyse 2.



UITKOMSTEN ANALYSE 2



UITKOMSTEN ANALYSE 2

HET PROCES VAN ANALYSE 2: DATA VAN DE WOON ENQUÊTE

Voor analyse 2 is analyse 1 uitgebreid met **verklarende variabelen** uit de **WoON enquête**. Deze enquête bevat vragen over onderwerpen als **woningtype**, **woonlasten**, **woonomgeving**, **verhuiscriteria**, **tevredenheid**, en **sociaaleconomische kenmerken**.

Door het grote aantal variabelen zijn voor de analyse **een groot aantal verklarende variabelen geselecteerd** die mogelijk relevant zijn. Deze variabelen zijn **gegroepeerd in verschillende thema's**:

- algemeen
- leefbaarheid
- verhuizen
- eigendom
- onderhoud
- woongenot
- vorig huis
- hypotheek
- woningkenmerken
- woonlasten
- woonwensen
- zorgbehoefte
- zorggeschiktheid
- energie
- vve

Ondanks het vooraf selecteren van mogelijk relevante verklarende variabelen bevat elke thema veel verschillende variabelen. Net zoals bij de Woonbase is er daarom voor gekozen om voor elke thema ook **een subset van de meer waarschijnlijke relevante variabelen te selecteren**.

Woonbase
WoON enquête

Woonbase

WoON
algemeen

WoON
thema's

ENQUÊTE DATA



UITKOMSTEN ANALYSE 2

HET PROCES VAN ANALYSE 2: WOON ENQUÊTE DATA

Enquêtedata (WoON) verschilt van **registratiedata** (Woonbase), wat specifieke aandachtspunten met zich meebrengt bij de analyse.

De **routing in de enquête** bepaalt welke vragen een respondent te zien krijgt. Hierdoor ontstaan er veel variabelen met **ontbrekende waarden**. Deze ontbrekende waarden zijn echter **niet willekeurig**, maar het directe gevolg van de routinglogica. Dit is een belangrijk aandachtspunt bij het uitvoeren van analyses, omdat het invloed heeft op de interpretatie en verwerking van de data.

Daarnaast is er sprake van **sterke onderlinge samenhang tussen variabelen**, vooral binnen thematische blokken. Dit leidt tot **multicollineariteit**, waarbij verklarende variabelen sterk met elkaar correleren of zelfs direct of indirect van elkaar zijn afgeleid. Dit kan de robuustheid van modellen beïnvloeden en vraagt om zorgvuldige aanpak en interpretatie.

XGBOOST



UITKOMSTEN ANALYSE 2

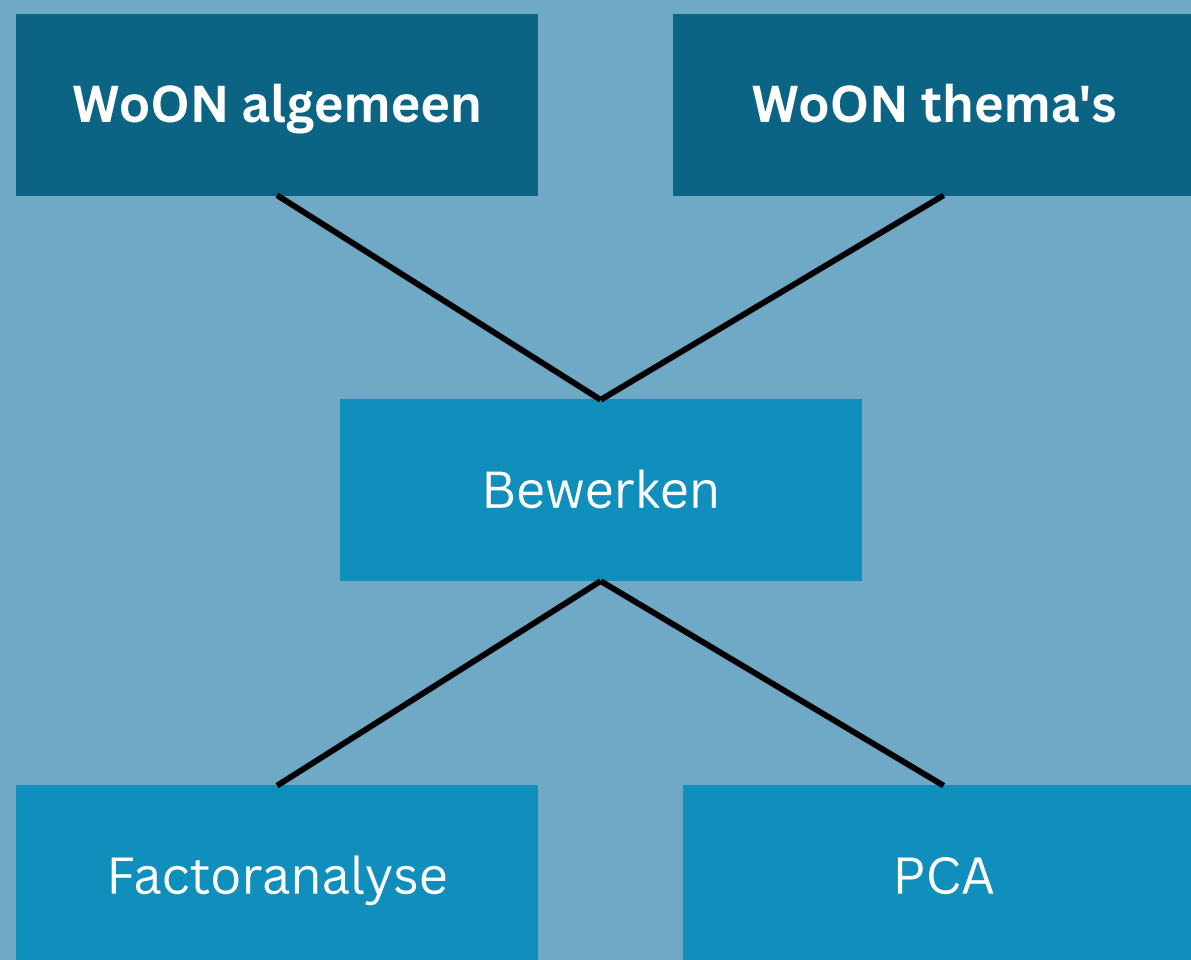
HET PROCES VAN ANALYSE 2: METHODE

Voor analyse 2 is gebruikt gemaakt van een **Gradient Boosting methode (XGBoost)**. Deze **methode kan omgaan met missende waarden** en deze hoeven van te voren niet geïmputeerd (opgevuld) te worden. Tijdens het trainen van de beslisbomen wordt automatisch door het algoritme de optimale splitsing bepaald voor observaties met missende waarden.

Een **sterke onderlinge correlatie van de verklarende variabelen** vormt voor XGBoost een **minder groot probleem** dan bij de klassieke (lineaire) modellen. Tijdens het iteratieve proces wordt telkens gekeken welke splitsing de grootste foutreductie oplevert, waardoor **sterk gecorreleerde variabelen automatisch worden gefilterd**.

In de eerste fase van de analyse zijn alle potentieel relevante variabelen opgenomen in het model om de **algemene voorspelkracht** te evalueren en te identificeren **welke factoren het meest bijdroegen** aan de verklaring van de uitkomstvariabele.

Vervolgens is per thema de **vooraf geselecteerd subset van variabelen** meegenomen in het model om te kijken wat dat voor invloed had op de voorspelkracht van het model.



[Klik hier voor de technische details](#)



UITKOMSTEN ANALYSE 2

HET PROCES VAN ANALYSE 2: VERSCHILLENDE STAPPEN

Voor de volgende fase van de **analyse is de data bewerkt**. De missende waarden zijn ingevuld (**geïmputeerd**) met een geschikte methode, afhankelijk van het type variabele. Daarnaast is voor elke variabele met ontbrekende waarden een aparte **indicatorvariabele** (dummy) aangemaakt, waarmee wordt vastgelegd of een waarde oorspronkelijk ontbrak. Op deze manier blijft de informatie over het ontbreken van data behouden en kan deze ook als verklarende factor worden meegenomen in het model.

De aangepaste variabelen zijn vervolgens gebruikt als input voor het model om te controleren of de aanwezigheid van missende waarden invloed heeft op de prestaties van het model.

In de laatste fase van de analyse zijn op de bewerkte dataset per thema twee **dimensiereductie technieken** toegepast: **factoranalyse** en **Principal Component Analysis (PCA)**. Deze methoden zijn ingezet om de **onderlinge samenhang tussen variabelen binnen elk thema te reduceren** en om multicollineariteit tegen te gaan. Voor elk thema zijn de resulterende factoren of componenten vervolgens opgenomen in het voorspellingsmodel. Op deze manier kon worden onderzocht hoe het model presteert wanneer sterk gecorreleerde variabelen worden samengevat tot een beperkt aantal dimensies, en of dit leidt tot robuustere en beter interpreteerbare resultaten.

VOORSPELKRACHT

Modelomschrijving	Bewerking	Accuracy
Selectie Woonbase (analyse 1)	Geen	52,2%
Selectie Woonbase en algemene variabelen WoON	Geen	53,7%
Selectie Woonbase en selectie algemene variabelen WoON	Geen	52,8%
Selectie Woonbase en algemene en thema variabelen WoON	Geen	69,2%
Selectie Woonbase en selectie algemene en thema variabelen WoON	Geen	69,9%
Selectie Woonbase en selectie algemene en thema variabelen WoON	Thema	69,3%
Selectie Woonbase en selectie algemene en thema variabelen WoON	Thema en algemeen	69,8%
Selectie Woonbase en selectie algemene en thema variabelen WoON	Factoranalyse thema	67,7%
Selectie Woonbase en selectie algemene en thema variabelen WoON	PCA thema	67,5%
Selectie Woonbase en selectie algemene en thema variabelen WoON	PCA totaal	67,7%



UITKOMSTEN ANALYSE 2

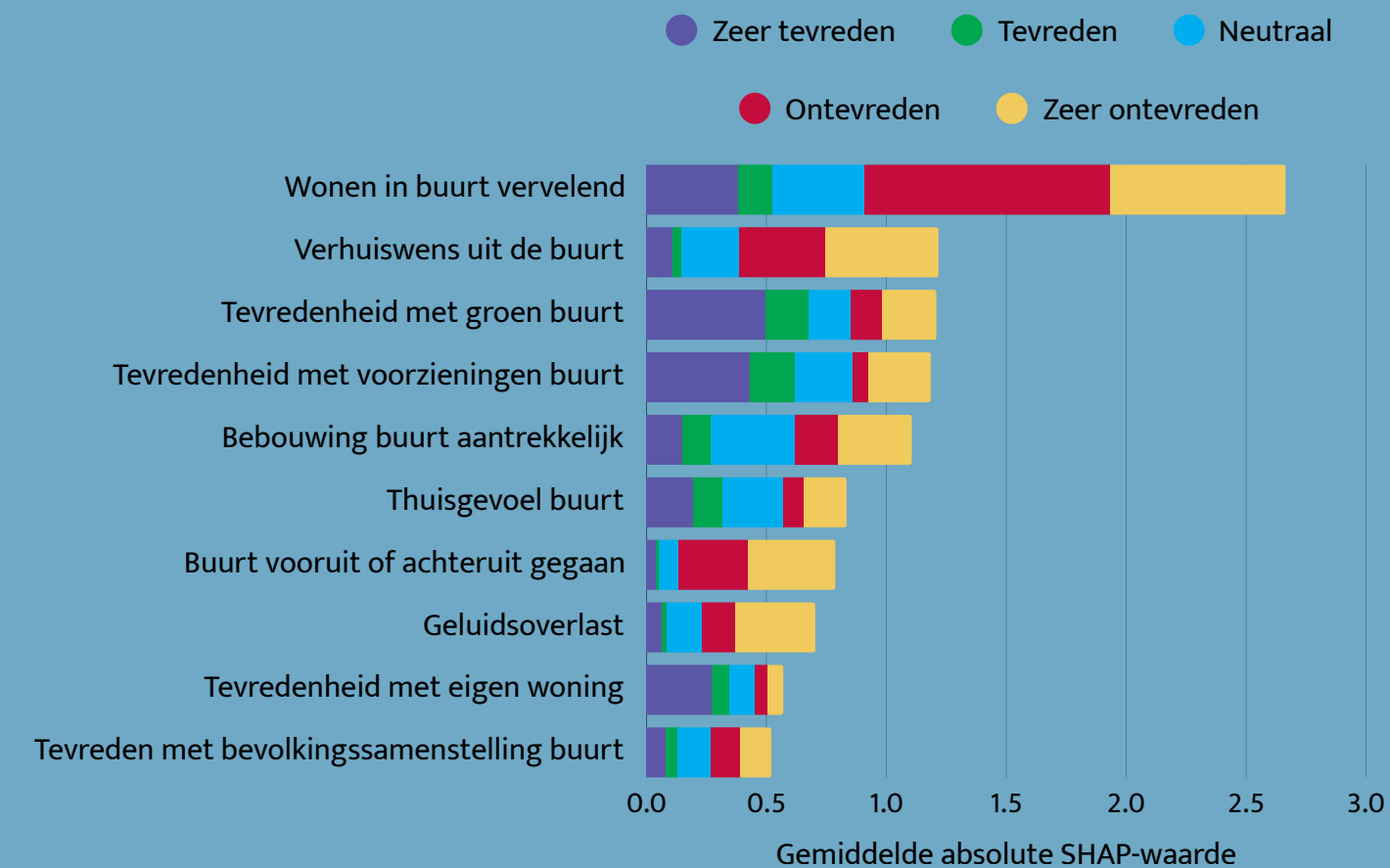
HOE GOED VOORSPELT HET MODEL?

In de tabel is de **voorspelkracht van het model** weergegeven **op basis van verschillende sets van verklarende variabelen**. Als **referentiepunt** is in eerste instantie **de set variabelen uit Analyse 1** gebruikt. Vervolgens zijn **stapsgewijs aanvullende variabelen** toegevoegd. De toevoeging van een selectie algemene variabelen uit de WoON enquête leidde niet tot een duidelijke verbetering in de modelprestaties. Pas **bij toevoeging van de thema-specifieke variabelen nam de voorspelkracht merkbaar toe**.

In de tweede fase is onderzocht of het toevoegen van bewerkte variabelen, waarbij specifiek rekening is gehouden met routingvariabelen met veel missende waarden, leidt tot een verbetering van de modelprestaties. De resultaten laten zien dat **de bewerkte variabelen een vergelijkbare voorspelkracht gaven als de oorspronkelijke, onbewerkte variabelen**.

Tot slot zijn op de bewerkte variabelen, zowel per thema als op het totaal, een factoranalyse en een Principal Component Analysis (PCA) uitgevoerd. Deze analyses hadden als doel om onderliggende structuren in de data te herkennen en de samenhang tussen variabelen te onderzoeken. Daarbij werd ook gekeken naar mogelijke multicollineariteit en naar de mogelijkheid om het aantal variabelen te reduceren zonder dat dit ten koste gaat van de voorspelkracht van het model. Hierdoor nam de voorspelkracht van het model enigszins af. Mogelijk is dit het gevolg van het feit dat de factoranalyse en PCA nog niet volledig waren geoptimaliseerd. De resultaten boden echter geen duidelijke aanleiding om deze optimalisatie door te voeren.

GEMIDDELDE IMPACT (SHAP-WAARDE) OP TEVREDENHEID MET DE WOONOMGEVING



UITKOMSTEN ANALYSE 2

WELKE FACTOREN ZIJN BELANGRIJKE VOORSPELLERS?

Wanneer we inzoomen op **het model met de hoogste accuracy** – het model waarin de verklarende variabelen bestaan uit **een selectie uit de Woonbase en een selectie van algemene en themavariabelen uit WoON**– dan zien we op basis van de gemiddelde absolute SHAP-waarden dat de **belangrijkste factoren voor het voorspellen van tevredenheid met de woonomgeving** juist de variabelen zijn die de **tevredenheid met de buurt** beschrijven. Dit resultaat is logisch en bevestigt de verwachte samenhang.

Echter, ondanks de sterke correlatie tussen deze variabelen, weet het model toch in maar ongeveer 70% van de gevallen de correcte uitkomstcategorie te voorspellen. De prestaties van het model blijven achter bij het classificeren van neutrale en negatieve sentimenten.

VOORSPELKRACHT

Modelomschrijving	Bewerking	Accuracy
basis en algemeen en thema's selectie	geen	69,9%
basis en algemeen en thema's selectie, zonder tevreden	geen	66,6%
basis en algemeen en thema's selectie, zonder leefbaarheid	geen	59,2%
basis en algemeen en thema's selectie, zonder leefbaarheid en tevreden buurt variabelen	geen	58,7%



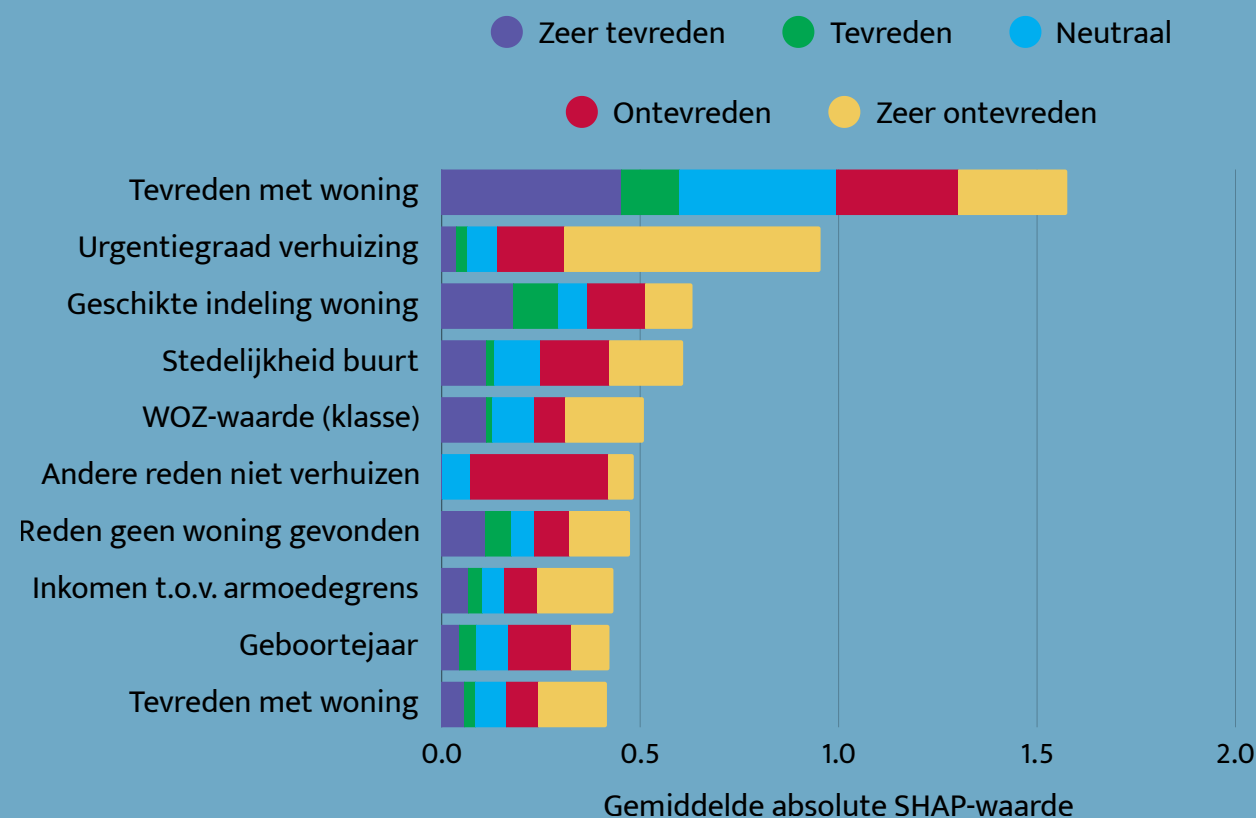
UITKOMSTEN ANALYSE 2

MODELPRESTATIE ZONDER BUURT-GERELATEERDE TEVREDENHEIDSVARIABELEN

Om te onderzoeken in welke mate **de sterke correlatie met buurttevredenheid het model beïnvloedt**, hebben we **stapsgewijs verschillende groepen variabelen weggelaten**. Eerst zijn de variabelen over **tevredenheid met groen en voorzieningen** verwijderd. Vervolgens zijn alle variabelen uit het **thema leefbaarheid** uitgesloten. Tot slot zijn ook andere **sterk gecorreleerde variabelen** met tevredenheid over de buurt verwijderd, die vielen onder het thema verhuizen. Op deze manier kunnen we vaststellen welk effect het elimineren van deze variabelen heeft op de modelprestaties en de voorspelkracht van het model.

Door het weglaten van de variabelen uit het thema leefbaarheid daalt de voorspelkracht van het model van bijna 70% naar bijna 60%. Deze categorie variabelen blijkt daarmee van significante invloed te zijn op de prestaties van het model.

GEMIDDELDE IMPACT (SHAP-WAARDE) OP TEVREDENHEID MET DE WOONOMGEVING



UITKOMSTEN ANALYSE 2

BELANGRIJKSTE VOORSPELLERS NA HET WEGLATEN VAN BUURT- EN LEEFBAARHEIDSVARIABELEN.

Na het verwijderen van de buurt- en leefbaarheidsvariabelen blijven vooral **kenmerken van de woning** zelf als **belangrijke voorspellers** over. Denk hierbij aan factoren zoals de indeling van de woning en de WOZ-waarde. Daarnaast spelen **persoonlijke factoren**, zoals het inkomen en het geboortjaar, een relevante rol in de voorspelling.

Ook blijven **verhuis gerelateerde kenmerken**, zoals de reden om te willen verhuizen, **relevant** voor het **voorspellen van de tevredenheid met de woonomgeving**. Wanneer echter ook deze variabelen worden uitgesloten, daalt de voorspelkracht van het model nog verder.

SAMENVATTEND ANALYSE 2



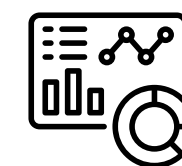
UITKOMSTEN ANALYSE 2

WAT HEBBEN WE GELEERD VAN ANALYSE 2?

Analyse 2 kan worden samengevat als volgt: **het toevoegen van thema-variabelen verbetert de prestaties van het model**, terwijl **het transformeren van de enquêtedata geen significant effect** heeft en **het toepassen van dimensiereductietechnieken ook geen verbetering** oplevert. Het model blijft **moeite** hebben om respondenten te herkennen die **neutraal of niet tevreden zijn over hun woonomgeving**.

Wanneer we inzoomen op de thema-variabelen blijkt dat vooral de sterk gecorreleerde **leefbaarheidsvariabelen van invloed** zijn. Deze variabelen zijn nauw verbonden met de doelvariabelen, maar ondanks deze correlatie bereikt het model slechts een voorspelkracht van circa 70%. Naast de leefbaarheidsvariabelen zijn ook enkele variabelen uit het thema verhuizen belangrijk voor het voorspellen van de tevredenheid met de woonomgeving.

Wanneer deze variabelen worden weggelaten, blijken de belangrijkste voorspellers tevredenheid over het huis, indeling van de woning, WOZ-waarde, stedelijkheid, inkomen en geboortjaar.



UITKOMSTEN ANALYSE 3

Woonbase
WoON enquête

Woonbase

WoON
enquête

Kerncijfers
buurten



UITKOMSTEN ANALYSE 3

HET PROCES VAN ANALYSE 3: BUURTKENMERKEN

Voor deze analyse zijn buurtkenmerken uit de StatLine-tabel **Kerncijfers Wijken en Buurten 2023** gekoppeld aan de datasets van Analyse 1 en 2. Het gaat om variabelen zoals **bevolkingsdichtheid, gemiddelde WOZ-waarde van woningen, percentage koopwoningen, aandeel corporatiewoningen, arbeidsparticipatie, aandeel hoge inkomens en aandeel huishoudens onder of rond het sociaal minimum.**

Analyse 3 is uitgevoerd in drie opeenvolgende fasen. **Fase 1 is gestart vanuit analyse 1**, waarbij individuele kenmerken centraal stonden. **Vervolgens zijn buurtkenmerken toegevoegd** om te onderzoeken of deze extra context de voorspellende kracht van het model versterkt. Als laatste stap zijn de gecorreleerde variabelen uit de Woonbase verwijderd, met als doel het effect van multicollineariteit te onderzoeken.

Fase 2 volgt een vergelijkbare aanpak, maar **is gebaseerd op analyse 2**. Ook hier zijn eerst **de buurtkenmerken toegevoegd**. Daarna zijn niet alleen de gecorreleerde Woonbase-variabelen weggelaten, maar ook de leefbaarheidsvariabelen. In deze fase wordt onderzocht welk effect het toevoegen van buurtkenmerken heeft op de prestaties van het model.

In **fase 3** zijn **uitsluitend buurtkenmerken** meegenomen. Het doel was om te evalueren hoe goed het model presteert op basis van deze kenmerken alleen, en hoe dat zich verhoudt tot modellen waarin individuele kenmerken zijn opgenomen, zoals in analyse 1. Deze fase biedt inzicht in de relatieve voorspellende waarde van buurtkenmerken ten opzichte van persoonlijke data.

VOORSPELKRACHT

Model	Accuracy
Selectie Woonbase (analyse 1)	52,2%
Selectie Woonbase (analyse 1) en buurtkenmerken	52,4%
Selectie Woonbase (analyse 1) en buurtkenmerken, zonder gecorreleerde variabelen	52,2%
Selectie Woonbase en selectie algemene en thema variabelen WoON (analyse 2)	69,9%
Selectie Woonbase en selectie algemene en thema variabelen WoON (analyse 2) en buurtkenmerken	69,1%
Selectie Woonbase en selectie algemene en thema variabelen WoON (analyse 2) en buurtkenmerken, zonder gecorreleerde variabelen	69,7%
Selectie Woonbase en selectie algemene en thema variabelen WoON (analyse 2) en buurtkenmerken, zonder gecorreleerde variabelen en leefbaarheid variabelen	59,0%
Alleen buurtkenmerken	52,3%



UITKOMSTEN ANALYSE 3

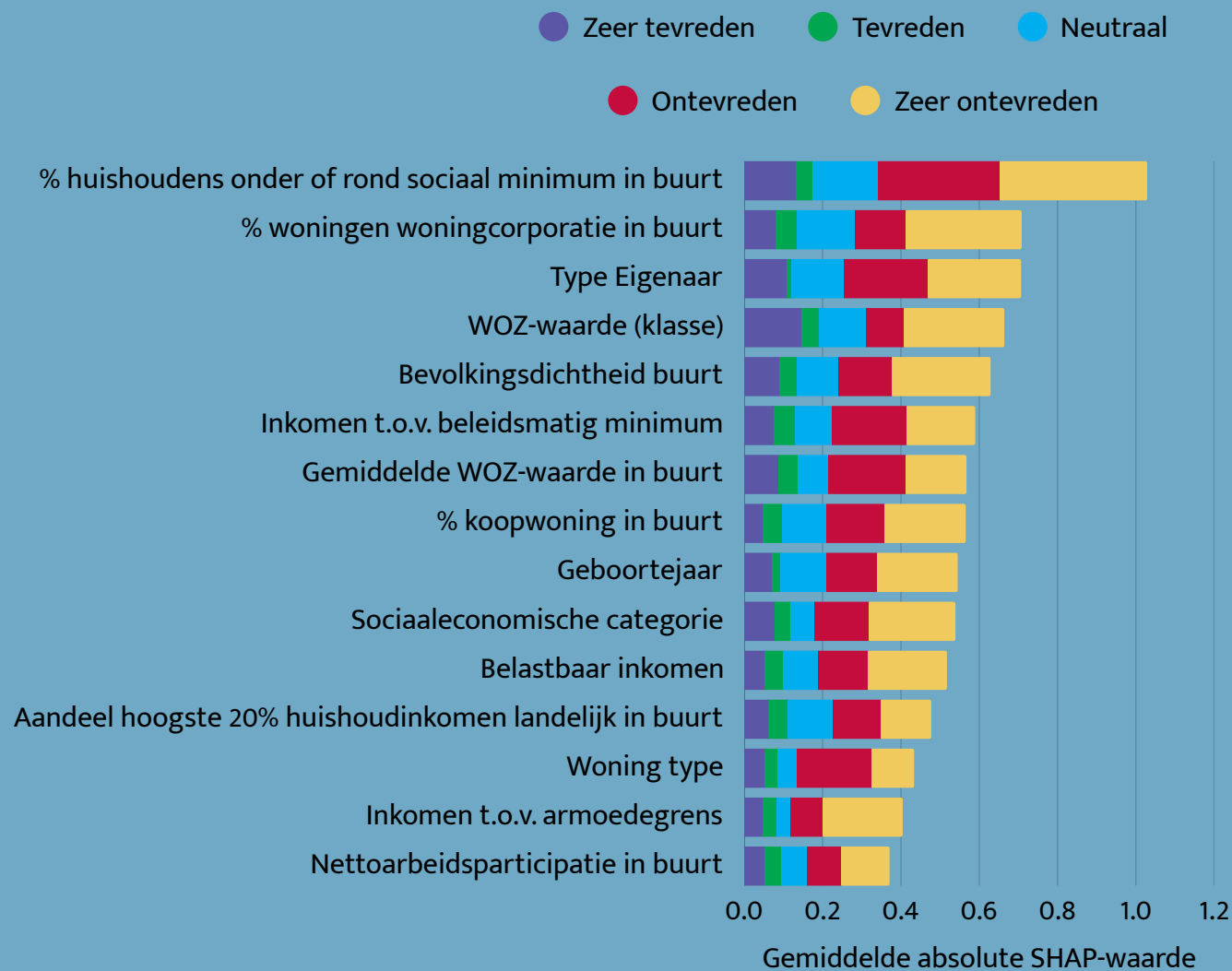
HOE GOED VOORSPELT HET MODEL?

In de tabel is de voorspelkracht weergegeven voor de verschillende fasen van de analyse. **Het toevoegen van buurtkenmerken** aan de registratiedata van **analyse 1** in de eerste fase heeft **geen noemenswaardig effect** op de modelprestaties.

Dit geldt eveneens voor de tweede fase waarbij **de buurtkenmerken zijn toegevoegd aan analyse 2**. Het integreren van variabelen uit de WoON-enquête leidde eerder tot een toename in voorspelkracht in analyse 2, maar het vervolgens **toevoegen van buurtkenmerken heeft nauwelijks invloed**. Ook het weglaten van gecorreleerde variabelen verandert dit beeld niet.

Tot slot zijn **uitsluitend de buurtkenmerken als verklarende variabelen** opgenomen. De modelprestaties blijken in dat geval **vergelijkbaar** te zijn met het model waarin **alleen de registratiedata uit de Woonbase** is gebruikt.

GEMIDDELDE IMPACT (SHAP-WAARDE) OP TEVREDENHEID MET DE WOONOMGEVING



UITKOMSTEN ANALYSE 3

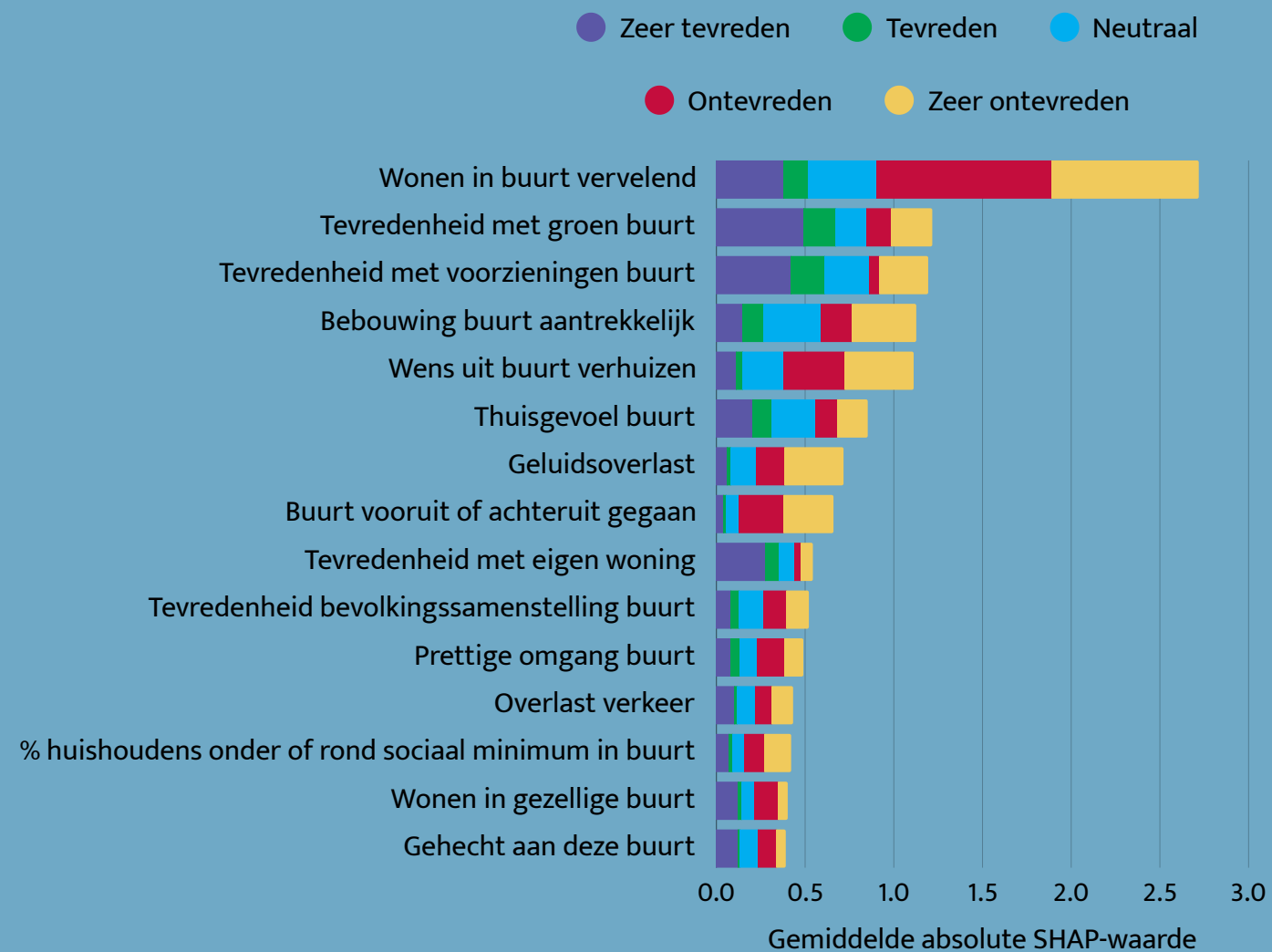
WELKE FACTOREN ZIJN BELANGRIJKE VOORSPELLERS NA HET TOEVOEGEN VAN DE BUURTKENMERKEN AAN ANALYSE 1?

Hoewel de voorspelkracht van het model uit analyse 1 niet toeneemt na het toevoegen van buurtkenmerken, is er wel een verschuiving zichtbaar in de belangrijkste verklarende factoren.

Na het toevoegen van buurtkenmerken aan analyse 1 blijken de belangrijkste factoren voor het voorspellen van tevredenheid met de woonomgeving een combinatie te zijn van buurtkenmerken en individuele kenmerken. Het percentage huishoudens dat onder of rond het sociaal minimum leeft in een buurt, evenals het aandeel woningen van woningcorporaties, zijn de sterkste voorspellers. Deze worden gevolgd door het type eigenaar, de WOZ-waarde en de bevolkingsdichtheid in de buurt. Daarnaast spelen buurtkenmerken gerelateerd aan inkomen, gemiddelde WOZ-waarde en individuele inkomensvariabelen een aanvullende rol.

Samengevat blijken in deze analyse de belangrijkste voorspellers van tevredenheid met de woonomgeving te liggen in buurtkenmerken en individuele kenmerken gerelateerd aan inkomen en woningkenmerken.

GEMIDDELDE IMPACT (SHAP-WAARDE) OP TEVREDENHEID MET DE WOONOMGEVING



UITKOMSTEN ANALYSE 3

WELKE FACTOREN ZIJN BELANGRIJKE VOORSPELLERS NA HET TOEVOEGEN VAN DE BUURTKENMERKEN AAN ANALYSE 2?

Bij het toevoegen van buurtkenmerken aan analyse 2 blijkt dat de **voorspelkracht van het model niet toeneemt**. Wel zijn er kleine **verschuivingen zichtbaar in de belangrijkste verklarende variabelen**. In deze analyse komt **slechts één buurtkenmerk** — het aandeel huishoudens met een inkomen op of onder het sociaal minimum — voor in de top 15 van belangrijkste voorspellers. Daarnaast zijn er kleine verschuivingen zichtbaar tussen de overige variabelen.

Voorals de variabelen die betrekking hebben op de ervaren leefbaarheid blijven relevant binnen het model. Deze factoren behouden hun verklarende kracht, ook wanneer aanvullende buurtkenmerken worden toegevoegd.

SAMENVATTEND ANALYSE 3



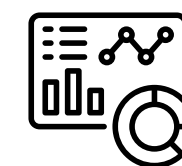
UITKOMSTEN ANALYSE 3

WAT HEBBEN WE GELEERD VAN ANALYSE 3?

Analyse 3 laat zien dat **het toevoegen van buurtkenmerken slechts beperkte invloed heeft op de voorspelkracht van het model**. Hoewel er **nauwelijks verbetering** optreedt in de **modelprestaties**, is er wel sprake van **een verschuiving in de belangrijkste verklarende factoren**.

Met name **het buurtkenmerk over het aandeel huishoudens met een inkomen op of onder het sociaal minimum** komt naar voren als **relevante variabele**. Daarnaast blijven de **ervaren leefbaarheidsvariabelen consistent belangrijk**, wat wijst op de blijvende invloed van subjectieve beleving op woontevredenheid.

Het weglaten van samenhangende variabelen verandert dit beeld nauwelijks, wat duidt op een zekere robuustheid in de samenstelling van het model.



SAMENVATTEND

SAMENVATTEND ANALYSES



UITKOMSTEN ANALYSES

WAT HEBBEN WE GELEERD VAN DE ANALYSES?

Het voorspellen van **tevredenheid met de woonomgeving** op basis van **registratiedata** blijkt slechts een **beperkte voorspelkracht** te hebben. Hoewel deze objectieve gegevens waardevol zijn, bieden ze onvoldoende inzicht in de persoonlijke beleving van bewoners.

Door **subjectieve belevingskenmerken** toe te voegen – zoals ervaren leefbaarheid en het gevoel van veiligheid – **neemt de voorspelkracht van het model aanzienlijk toe**. In circa 70% van de gevallen werd de juiste uitkomst voorspeld. Sommige variabelen correleren sterk met de doelvariabele en dragen bij aan een significante verbetering van de voorspellingen. Toch blijft de **voorspelkracht van het model begrensd**. Omdat het overgrote deel van de respondenten tevreden is, lukt het de modellen minder goed om de kleinere groepen met een neutrale of negatieve beleving te herkennen.

Ook **het toevoegen van buurtkenmerken leidt niet tot een verdere verbetering van het model**. Hoewel dit een lichte verschuiving veroorzaakt in de relevante factoren, blijft de algehele voorspelkracht nagenoeg gelijk.

De **meest bepalende voorspellers van tevredenheid met de woonomgeving** zijn **subjectieve belevingskenmerken**, zoals **ervaren leefbaarheid**, aangevuld met **woningkenmerken en persoonlijke factoren**. Vooral de manier waarop bewoners hun omgeving beleven blijkt logischerwijs sterk samen te hangen met tevredenheid. Daarnaast spelen eigenschappen van de woning, zoals indeling en WOZ-waarde, en persoonlijke kenmerken zoals inkomen en leeftijd een relevante rol.