# VAN DATA NAAR INZICHT: ONDERZOEK NAAR WONINGBRANDEN EN BUURTKENMERKEN IN VRMWB

Hoofdonderzoeker: Charley Bosman datascientist GGD West-Brabant, in opdracht van Veiligheidsregio Midden- en West-Brabant.

Projectteamleden: Rob van Bussel, Metha de Heer, Ike Kroesbergen, Floor van Lintel, Dirk Suchy, Steven Troost en Leonard Vanbrabant.

Opdrachtgever: Louise Schneider, programmamanager risicogerichtheid van de Veiligheidsregio Midden- en West-Brabant, i.s.m. Jolande van Balen, programmamanager informatiepositie.

Breda, juni 2023

# INHOUD

# INLEIDING

In oktober 2021 startte de projectgroep woningbranden vanuit de speerpunten van de Veiligheidsregio Midden- en West-Brabant (VRMWB) 'risicogerichtheid' en 'informatiepositie'. Feit is dat in de VRMWB 70% van de gebouwbranden een woningbrand is. Bij deze woningbranden vallen de meeste slachtoffers en de impact op de inwoners in materiële en immateriële schade, letsel, risico te overlijden is groot, terwijl de regelgeving rondom woningbranden slechts beperkt is. Het versterken van de risicogerichtheid en de informatiepositie zijn beide speerpunten in het beleid van de VRMWB. Daardoor is er behoefte aan meer inzicht over de risico's in de buurten en buurtkenmerken. Bijvoorbeeld voor de veilige en gezonde aanpak in buurten en wijken. Het inzicht in buurten en buurtkenmerken in relatie tot woningbranden ontbrak, wat heeft geleid tot de volgende onderzoeksvragen:

- Kunnen buurten in de VRMWB geclusterd worden tot een aantal buurttypen op basis van buurtkenmerken die verder gaan dan het aantal inwoners in de buurt?
    - En verschillen deze buurttypen in het aantal woningbranden gerelateerd aan het aantal huishoudens?
- Welke buurtkenmerken hangen in ieder van deze buurttypen het meest samen met een woningbrand?
- Verschilt de omvangschade en het letsel tussen de buurttypen?

De achtergrond van de onderzoeksvragen en de eerste bevindingen met de buurtindeling van 2016 zijn vastgelegd in een interne rapportage (Max Geelen, juni 2022). Tevens is de onderzoeksmethode beschreven en wetenschappelijk getoetst in een artikel in het Tijdschrift voor Veiligheid (Februari, 2023).

Deze huidige rapportage beschrijft de doorontwikkeling vanaf juli 2022 tot juni 2023 met de buurtindeling 2020 en verrijkte data. In de Appendix zijn de analysemethoden en resultaten in detail vastgelegd. De ontwikkelde algoritmen zijn opgeslagen in het register van de VRMWB, zodat het hergebruikt kan worden. Een managementsamenvatting is apart beschikbaar.

# DATA

Een *woningbrand* is gedefinieerd als een melding bij de Gemeenschappelijke Meldkamer vastgelegd in het GMS van de brandweer, met een inzet van in ieder geval de brandweer. In dit onderzoek beoordelen we een woningbrand als: 1) incident geclassificeerd als woningbrand of 2) gebouwbrand in een gebouw met woonfunctie. Bij deze woningbranden kan letsel of schade zijn ontstaan; in de dataset van het GMS zijn geen gegevens bekend over eventueel letsel en schade aan eigendommen. Het betreft unieke incidenten (dubbele meldingen zijn samengevoegd), waarvan de status is afgesloten. Het gebruikte databestand is gefilterd voor oefen-, test- of onterechte meldingen en/of incidenten waarbij de brandweer niet is uitgerukt. Alle incidenten die aan de definitie voldoen vanaf 1 januari 2013 tot en met 31 december 2022 zijn meegenomen in de analyses. In totaal gaat het in de VRMWB in de genoemde periode om 3749 woningbranden. De incidenten zijn geaggregeerd op CBS-buurtniveau. De gebruikte buurtindeling is afkomstig van het CBS (2020). Verder gebruiken we de buurtkenmerken uit een openbare dataset van het CBS met kerncijfers over wijken en buurten (CBS, 2020). Kenmerken over leefstijl en gezondheid van de buurtbewoners komen uit de Gezondheidsmonitor 2020 (GGD'en, RIVM & CBS, 2020). Dit betreft een vierjaarlijkse vragenlijst die is afgenomen bij een steekproef van inwoners vanaf 18 jaar. De uitkomsten hiervan zijn geschat op CBS-buurtniveau en zijn openbaar beschikbaar.

In ons onderzoek naar woningbranden in Midden- en West-Brabant hebben we drie bronnen gebruikt om informatie te verkrijgen. Ten eerste hebben we het GMS-systeem gebruikt om de woningbranden in te delen in CBS-buurten. Daarnaast hebben we openbare CBS-wijk- en buurtgegevens gebruikt, die statistieken bevatten over specifieke buurten, wijken en gemeenten in een bepaald jaar. Deze gegevens bieden inzicht in demografische informatie, migratieachtergrond, huishoudtypen, energiegegevens, woningtypen en meer. De gegevens variëren van aantallen tot percentages en gemiddelden. Tot slot hebben we gegevens van de GGD over de gezondheid per buurt gebruikt, die elke vier jaar gepubliceerd worden en inzicht geven in verschillende gezondheidscijfers zoals aantallen alcoholdrinkers, rokers, inwoners met langdurige aandoeningen en dergelijke. We hebben besloten om gegevens uit het jaar 2020 te gebruiken omdat de GGD-Gezondheidsmonitor (een vragenlijst over gezondheid, gedrag en leefstijl) om de vier jaar wordt uitgevoerd en de meest recente toekomstige gegevens pas in 2024 beschikbaar zullen zijn. Zo konden we kerncijfers van het CBS op buurtniveau combineren met de gezondheidsgegevens die de GGD bij haar inwoners heeft verzameld.

# PRE-PROCESSING

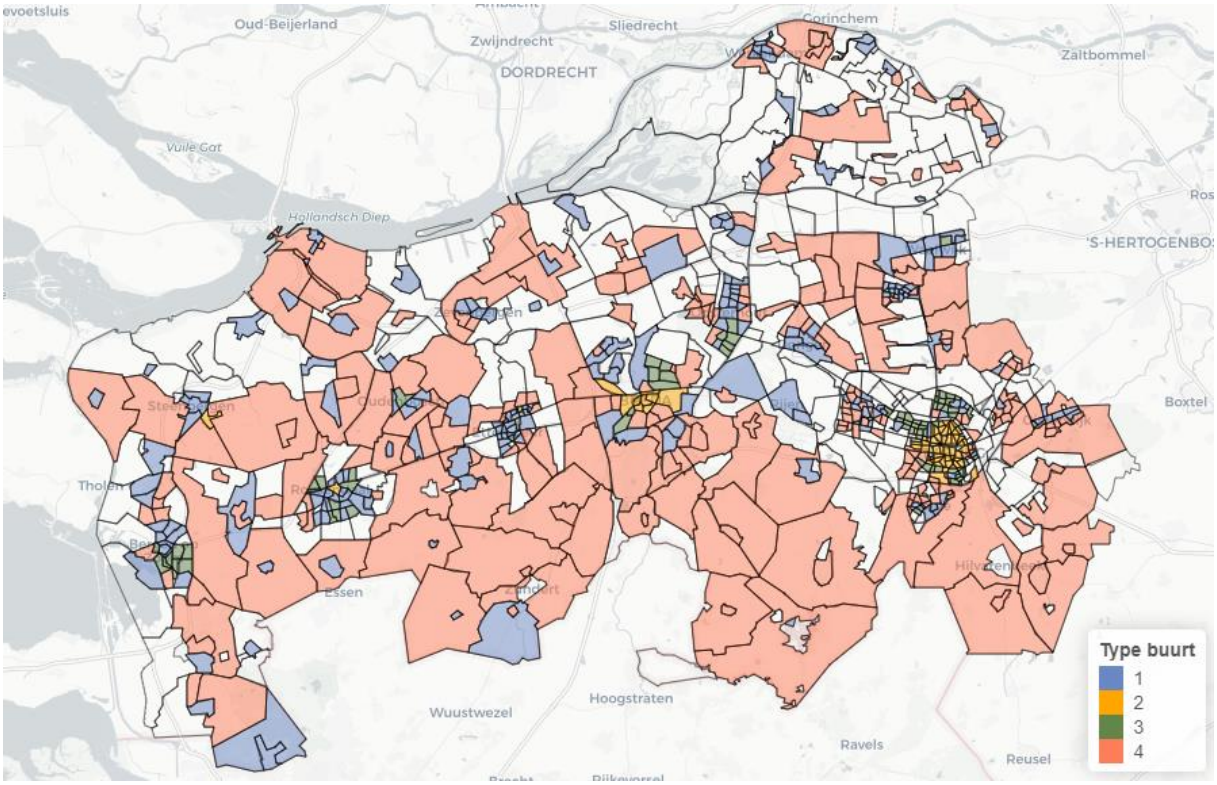
De dataset van het Centraal Bureau voor de Statistiek (CBS) bevat veel verschillende variabelen, maar niet alle variabelen zijn elk jaar openbaar beschikbaar voor elke buurt. Dit betekent dat de beschikbaarheid van gegevens per jaar kan verschillen voor een specifieke buurt, wat het lastig kan maken om de analyse opnieuw uit te voeren na verloop van tijd. Voor dit project hebben we in eerste instantie gegevens uit 2016 gebruikt omdat er op dat moment geen recentere gegevens beschikbaar waren. Wanneer we bijvoorbeeld de analyse op data uit het jaar 2020 wilden uitvoeren, kwamen de kolommen niet meer overeen met die van de data uit het jaar 2016. Om ons onderzoek reproduceerbaar te maken, hebben we besloten om alleen die variabelen te selecteren die waarschijnlijk nog steeds beschikbaar zullen zijn in de CBS-dataset van het volgende jaar. Om te bepalen welke variabelen dit zijn, hebben we gekeken naar welke variabelen sinds 2015 in de CBS-dataset voorkwamen. Door deze benadering te hanteren, is het waarschijnlijker dat we onze analyses kunnen herhalen en vergelijken met latere jaren. Zie Appendix A om meer details te verkrijgen over het opschonen van de dataset, ook staat hier een overzicht van de variabelen die zijn gebruikt voor dit onderzoek.

Naast het selecteren van de relevante variabelen, hebben we ook zorgvuldig gekeken naar welke buurten we wilden opnemen in ons onderzoek. We hebben ervoor gekozen om buurten uit te sluiten waar minder dan 100 huishoudens wonen. Dit komt omdat deze buurten vaak bedrijventerreinen zijn en er weinig woningen te vinden zijn, waardoor deze buurten minder relevant zijn voor ons onderzoek. Dit heeft geresulteerd in 609 buurten die meegenomen zijn in het onderzoek. Daarnaast kijken we naar percentages en niet naar aantallen, en zo kunnen de percentages van de kleinste buurten snel hoog oplopen, omdat er simpelweg ook minder mensen wonen.

# CLUSTERING VAN BUURTEN

De buurten van Midden- en West-Brabant hebben we vergeleken op het gebied van demografie, gezondheid, type gezinnen en woningen. Om deze buurten te groeperen en inzicht te krijgen in hun overeenkomsten en verschillen, hebben we een clusteranalyse uitgevoerd. Dit houdt in dat we de buurten hebben verdeeld in verschillende clusters op basis van de overeenkomsten tussen de variabelen die we hebben geanalyseerd. Er zijn vier clusters van buurten geïdentificeerd uit de analyse, die elk unieke eigenschappen hebben die hen onderscheiden van andere buurten. Zie Appendix B voor meer uitleg hoe deze clusterindeling is gemaakt.

Figuur 1 geeft de clusterindeling weer voor de regio Midden- en West-Brabant. De niet-ingekleurde vlakken zijn buurten met minder dan 100 huishoudens die niet zijn opgenomen in de clusteranalyse. Deze buurten worden daarom niet weergegeven met een kleur op de afbeelding.

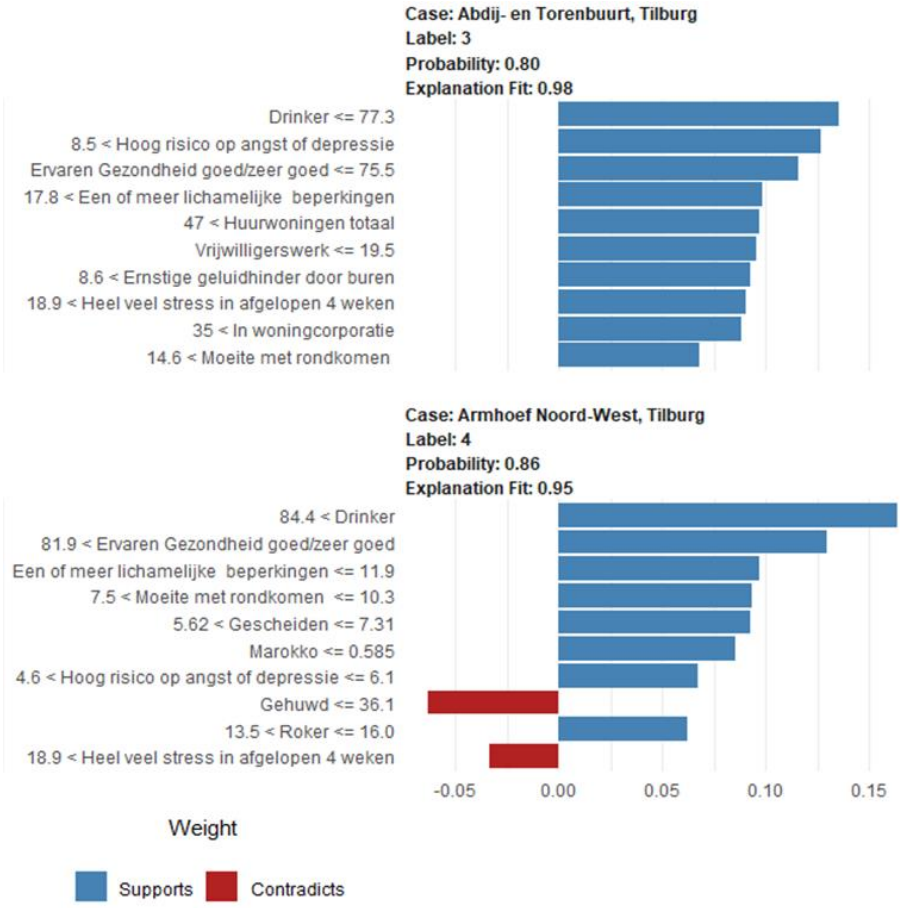



**FIGUUR 1**

### *WAT TYPEERT EEN CLUSTER?*

We zijn gestart met het visualiseren van de belangrijkste kenmerken die de verschillende clusters van elkaar onderscheiden. Voor elk onderwerp (zoals woningtypes, leeftijdscategorieën, lichamelijke gezondheid, mentale gezondheid, etc.) hebben we de relevante gegevens geschaald en weergegeven in grafische vorm. Op deze manier konden we opvallende patronen en inzichten snel identificeren.

Daarnaast maken we ook gebruik van een methode die voor elke buurt afzonderlijk uitlegt waarom het model een specifieke voorspelling heeft gedaan. Dit wordt uitgelegd in Appendix C. Deze methode identificeert belangrijke variabelen die kenmerkend zijn voor een bepaald buurtcluster en laat zien of deze variabelen een positieve of negatieve invloed hebben gehad op de indeling van dat cluster. Hierdoor krijgen we niet alleen inzicht in de belangrijkste variabelen per cluster, maar kunnen we ook meer inzoomen op specifieke buurten en hun karakteristieken. In Figuur 2 zie je een voorbeeld van de uitkomsten voor twee buurten.




**FIGUUR 2**

In dit geval zijn de buurten Abdij- en Torenbuurt en Armhoef Noord-West in Tilburg geanalyseerd. De grafiek laat zien welke variabelen een hoge invloed hebben gehad op de voorspelling van het cluster waarin de buurt is ingedeeld, en hoe goed deze variabelen de voorspelling kunnen verklaren. Deze grafieken zijn geïntegreerd in de buurtcluster kaart van de regio. Dit is een interactieve kaart waar alle buurten te zien zijn met hun specifieke buurtcluster en hun kleuren (zie Figuur 1), en als je op een buurt klikt krijg je de bijbehorende grafiek te zien. In het najaar van 2023 wordt hier een aparte sessie voor ingepland om de kaart goed over te dragen aan degene die hem zullen gebruiken. Voor meer technische details over deze methode zie Appendix C en Appendix D.

Het is van belang om te benadrukken dat de vermelde variabelen niet de enige factoren zijn die hebben bijgedragen aan de indeling van de buurten in clusters. Het gaat om de combinatie van meerdere factoren die hebben geleid tot de voorspelling van de clusters. Het is tevens mogelijk dat een bepaalde variabele een negatieve invloed heeft gehad op de voorspelling.

Om de betrouwbaarheid van de cluster indeling te waarborgen, zijn deze bevindingen voorgelegd aan deskundigen in de brandweersector in de regio, en hun input werd gebruikt om de resultaten te valideren.

Kort samenvat kunnen we de clusters als volgt omschrijven:

Cluster 1: In dit buurtcluster vinden we weinig opvallende punten maar meer een gemiddeld cluster. Er is wel een iets hoger percentage mensen van 65 jaar en ouder, en er is relatief gezien weinig migratieachtergrond in vergelijking met andere clusters. De woningen in dit cluster variëren tussen huur- en koopwoningen.

Cluster 2: In dit buurtcluster zien we een piek in de leeftijdscategorieën 16-25 en 25-45 jaar. Er is een opvallend hoog percentage personen dat niet getrouwd is, en het aantal meergezinswoningen (huur) is relatief hoog in

vergelijking met andere clusters. Daarnaast zien we een relatief hoger percentage (westerse) migratieachtergrond, hogere percentages rokers, alcoholgebruikers, mensen met stress, angst en depressie ten opzichte van andere clusters.

Cluster 3: Personen in de buurten in dit cluster vallen iets meer in de leeftijdscategorie van 25 tot 45 jaar en wonen in stedelijke gebieden. We zien hier een hoog percentage meergezinswoningen van woningcorporaties en we zien ook een hoger percentage migratieachtergrond, zoals Marokko, Turkije, de Antillen en Aruba, Suriname, of andere niet-westerse landen. Vergeleken met andere clusters hebben de buurten een hoger percentage mensen met lichamelijke beperkingen, rokers en mensen met stress, angst en depressie, sociale/emotionele eenzaamheid, en overgewicht.

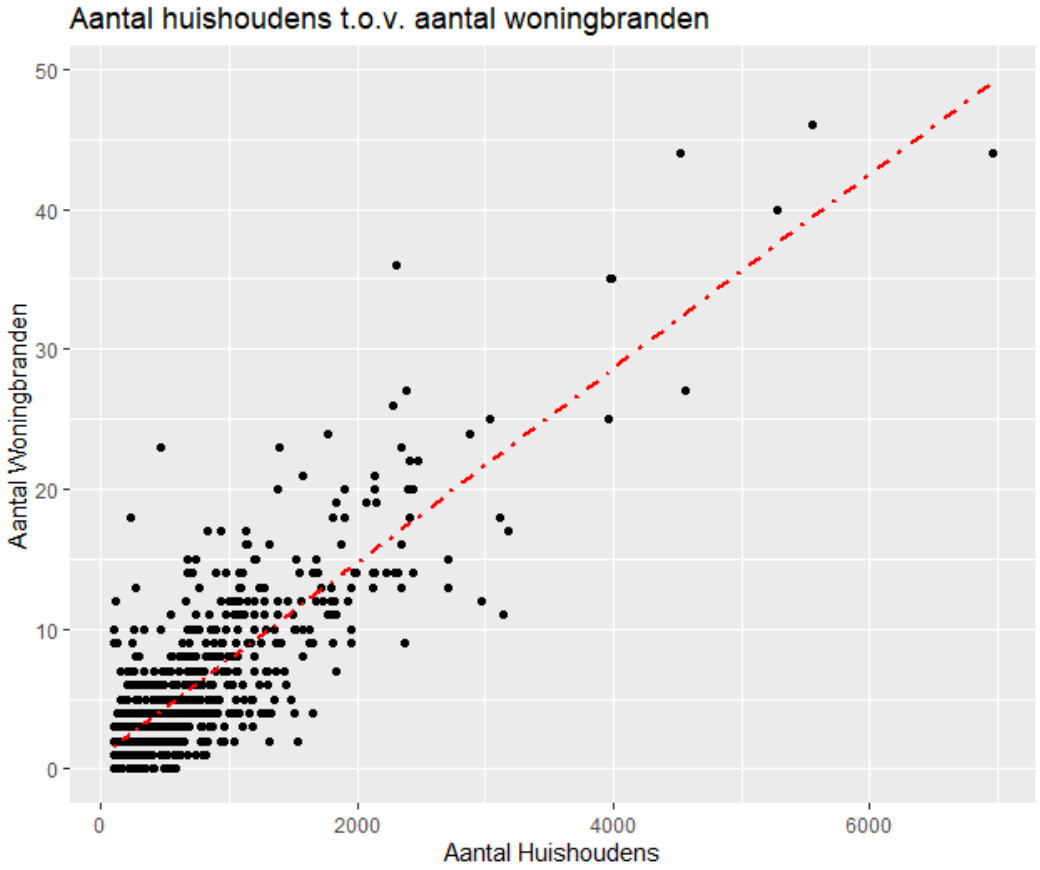
Cluster 4: Dit buurtcluster omvat relatief meer inwoners in de leeftijdscategorie van 45 tot 65 jaar die in landelijke gebieden wonen. Ze wonen voornamelijk in eengezinswoningen (koop) en vaak gaat het om huishoudens met kinderen. Over het algemeen hebben zij een goede (mentale) gezondheid en is er weinig tot geen migratieachtergrond in vergelijking met andere buurtclusters.

# RESULTATEN ANALYSE WONINGBRANDEN HELE REGIO EN PER CLUSTER

Dit hoofdstuk presenteert de resultaten van de analyse van woningbranden in de hele regio en per cluster van buurten. We kijken naar het aantal woningbranden in elk cluster om patronen en trends te identificeren en eventuele verschillen tussen de clusters te onderzoeken. We beginnen met een overzicht van het totale aantal woningbranden in de regio en hoe dit zich verhoudt tot het aantal huishoudens en vervolgens presenteren we de resultaten per cluster.

## HELE REGIO

Figuur 3 toont de relatie tussen het aantal huishoudens en het aantal woningbranden. De rode lijn geeft de geschatte waarden van het aantal woningbranden weer op basis van een eenvoudig lineair regressiemodel. De lijn die is getekend is de best passende lijn tussen de punten. De lijn wordt getekend op zo'n manier dat de afstanden tussen de punten en de lijn het kleinst is. De punten in de grafiek vertegenwoordigen de buurten en hun locatie geeft de verhouding weer tussen het aantal huishoudens in die buurt en het feitelijk aantal woningbranden dat in die buurt heeft plaatsgevonden. Met andere woorden, de punten laten zien hoeveel woningbranden er hebben plaatsgevonden per huishouden in een bepaalde buurt. Zie Appendix E voor meer details over deze regressieanalyse. In de grafiek kan je zien dat er een positief sterk verband blijkt te zijn tussen het aantal huishoudens en het aantal woningbranden voor de hele regio Midden- en West-Brabant.

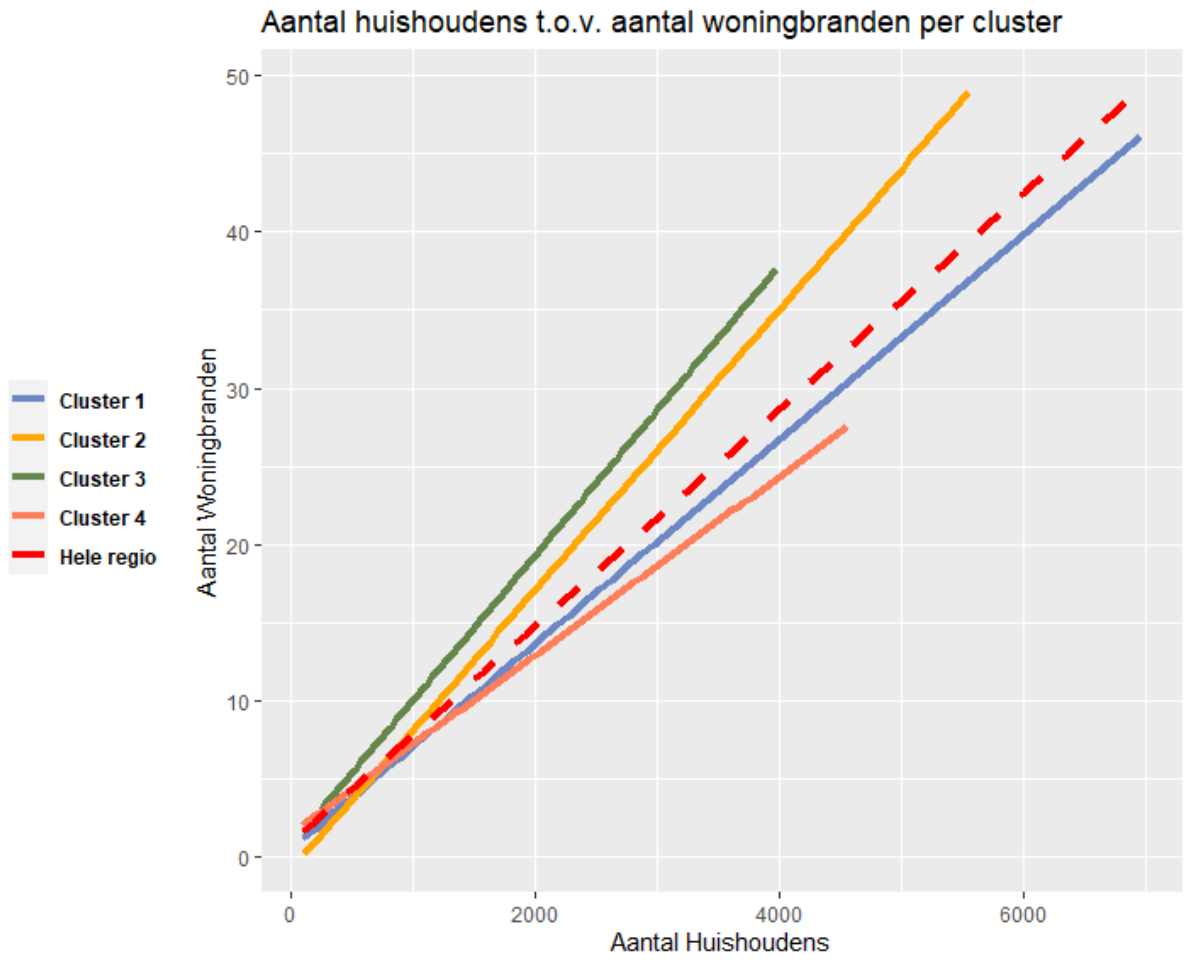



**FIGUUR 3**

## PER CLUSTER

De grafiek in Figuur 4 laat het verband zien tussen het aantal huishoudens en het aantal woningbranden in de verschillende clusters van buurten, in vergelijking met het verband tussen het aantal woningbranden en het

aantal huishoudens voor de hele regio, zoals te zien in Figuur 3. Er zijn in totaal vijf lijnen in de grafiek van Figuur 4, één voor elk buurtcluster en één voor de hele regio. Door de lijnen te vergelijken, kunnen we zien of er verschillen zijn in de relatie tussen het aantal huishoudens en het aantal woningbranden tussen de verschillende clusters en de hele regio.



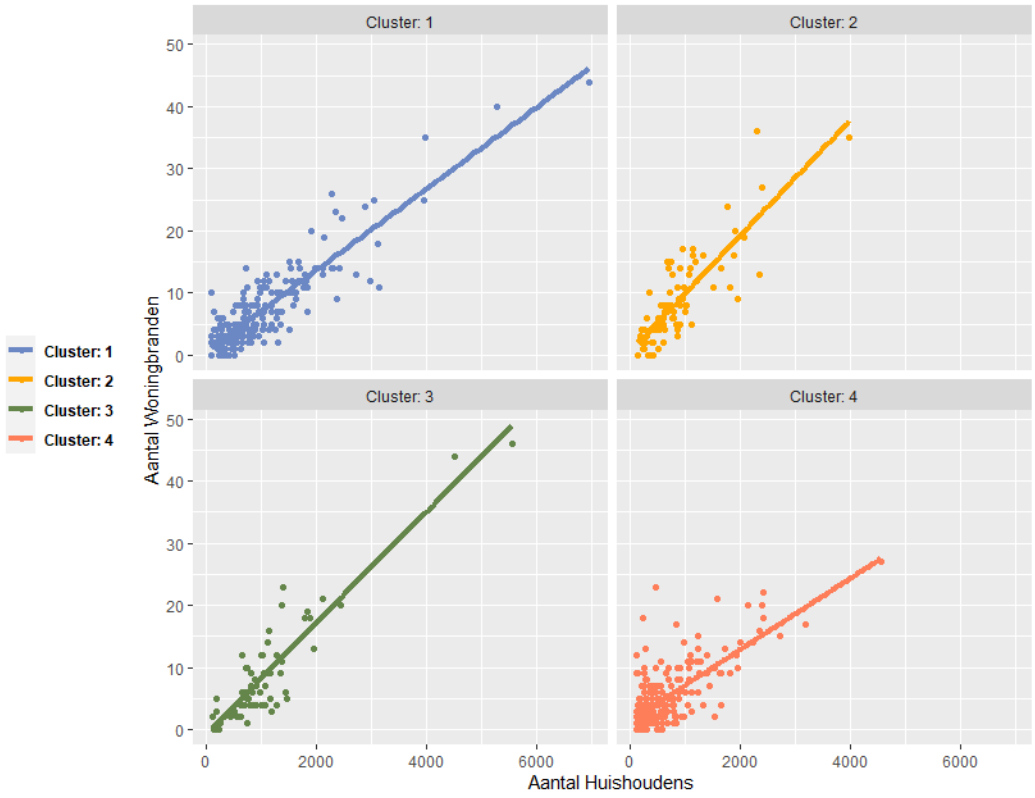

**FIGUUR 4**

De lijnen van Cluster 2 en Cluster 3 stijgen steiler dan die van Cluster 1, Cluster 4 en de hele regio in Figuur 4. Dit geeft aan dat in Cluster 2 en Cluster 3 het aantal woningbranden sneller toeneemt naarmate het aantal huishoudens groter wordt. Aangezien de grafiek de verhouding tussen het aantal woningbranden en het aantal huishoudens in verschillende clusters toont, zijn de individuele buurten (de punten) niet weergegeven om de leesbaarheid van de grafiek te behouden. ***Kort gezegd, neemt het aantal woningbranden meer toe in Cluster 2 en Cluster 3 naarmate het aantal huishoudens stijgt in een buurt.***

Figuur 5 toont vier subplots, één voor elk cluster. Elke subplot bevat punten die de relatie tussen het aantal huishoudens en het aantal woningbranden in elke buurt binnen het cluster weergeven. De lijn is een lineaire regressielijn die het verband tussen het aantal huishoudens en het aantal woningbranden in dat cluster weergeeft.

Door elk cluster afzonderlijk te bekijken, kunnen patronen en trends specifiek voor dat cluster worden geïdentificeerd. Sommige clusters hebben bijvoorbeeld een meer lineair verband tussen het aantal huishoudens en het aantal woningbranden, terwijl andere clusters een minder lineair verband laten zien, vooral bij Cluster 4. Om te bepalen of er een lineair verband is tussen de variabelen, kun je naar de vorm van de puntenwolk kijken. Als de punten min of meer langs een rechte lijn liggen, dan is er sprake van een lineair verband. Als de punten

juist meer verspreid zijn en niet langs een rechte lijn lijken te liggen, dan is er waarschijnlijk geen sprake van een lineair verband. Als je kijkt naar de punten van Cluster 4, dan lijkt het alsof er sprake is van een cluster van punten die meer verspreid zijn dan de punten in de andere clusters. Dit kan verschillende oorzaken hebben. Er kunnen bijvoorbeeld andere factoren zijn dan het aantal huishoudens die het aantal woningbranden beïnvloeden, die specifiek gelden voor dit cluster. Hierdoor ontstaat er meer spreiding in de data.




**FIGUUR 5**

De resultaten van de lineaire regressies per cluster zijn te vinden in Appendix E. Er is ook gekeken naar de interactie van het aantal huishoudens en het type cluster van een buurt. Bij het onderzoeken van het effect van het aantal huishoudens op het aantal woningbranden per cluster, kijken we naar de manier waarop het aantal woningbranden afhankelijk is van het aantal huishoudens en het type cluster van de buurt. De uitkomsten hiervan zijn dat het effect van het aantal huishoudens verschilt tussen de verschillende clusters. De coëfficiënten van deze interactie-effecten zijn significant en positief voor Cluster 2 en Cluster 3, wat betekent dat het effect van het aantal huishoudens op het aantal woningbranden groter is in die clusters dan in Cluster 1 (het referentiecluster). Voor Cluster 4 is het interactie-effect negatief en niet significant. In Cluster 2 en Cluster 3 is dus het effect van het aantal huishoudens op het aantal woningbranden groter dan in Cluster 1, wat betekent dat de kans op woningbrand toeneemt naarmate het aantal huishoudens in de buurt toeneemt, vooral in deze clusters. In Cluster 4 is er echter geen significant effect van het aantal huishoudens op het aantal woningbranden, wat suggereert dat andere factoren belangrijker kunnen zijn bij het beïnvloeden van de kans op woningbrand in dit cluster.

### AANTAL WONINGBRANDEN PER 1000 HUISHOUDENS

Het absolute aantal woningbranden kan op zichzelf een nuttige maatstaf zijn om de veiligheid van een gebied te meten, maar het kan misleidend zijn als je alleen dit cijfer gebruikt om verschillende gebieden te vergelijken. Als je rekening houdt met de grootte van elke buurt in aantal huishoudens en het aantal woningbranden relateert aan het aantal huishoudens, kun je een betere vergelijking maken tussen de buurten. Daarom is het nuttig om het aantal woningbranden per 1000 huishoudens te berekenen. Door deze ratio te berekenen, kunnen we een

meer accurate vergelijking maken tussen de verschillende buurten. Dit kan helpen bij het identificeren van buurten waar de verhouding tussen het aantal woningbranden en het aantal huishoudens hoog is en die mogelijk extra aandacht nodig hebben op het gebied van brandveiligheid. Uit de resultaten blijkt dat het aantal woningbranden per 1000 huishoudens verschilt tussen de buurtclusters (zie Appendix E voor de statistische toets). Specifiek is gebleken dat in Cluster 3 de mediaan van het aantal woningbranden per 1000 huishoudens hoger ligt dan in de andere clusters.

### BELANGRIJKSTE INDICATOREN BUITEN HET AANTAL HUISHOUDENS

We hebben een analyse uitgevoerd om de belangrijkste demografische en gezondheidsfactoren te identificeren die van invloed zijn op de voorspelling van het aantal incidenten (woningbranden) in de verschillende clusters. Hierbij hebben we eerst inclusief het aantal huishoudens gekeken. Later hebben we ook naar indicatoren buiten het aantal huishoudens om gekeken, om de invloed van deze factor te isoleren en ons te concentreren op andere mogelijke variabelen. Dit geeft ons een beter inzicht in welke factoren het meest relevant zijn voor het aantal woningbranden in verschillende clusters. Meer uitleg over de analyse vind je in Appendix F.

In alle clusters spelen meerdere factoren een rol bij de kans op woningbranden per buurt. Per cluster komen verschillende kenmerken naar voren die in combinatie met elkaar belangrijk zijn in de voorspelling van een woningbrand. Een paar van die factoren lichten we hieronder toe:

In Cluster 1 spelen meerdere factoren een rol bij de kans op woningbranden per buurt. Ten eerste gaat een hoger percentage woningen gebouwd vóór 2000 gepaard met een verhoogde kans op woningbranden. Met andere woorden, buurten met een groter aandeel oudere woningen lopen een hoger risico op woningbranden. Daarnaast heeft het percentage inwoners dat ernstig eenzaam is ook een positieve invloed op het aantal woningbranden. Aan de andere kant zijn er bepaalde factoren zoals het percentage eengezinswoningen, woningen in bezit door overige verhuurders en het percentage huishoudens zonder kinderen die de kans op woningbranden juist verkleinen.

In Cluster 2 is er een verhoogde kans op woningbranden in buurten waar het percentage inwoners met een Marokkaanse achtergrond toeneemt, evenals het percentage gescheiden inwoners. Aan de andere kant heeft het percentage vrijwilligerswerk een omgekeerd effect, waarbij een hoger percentage vrijwilligerswerk juist de kans op woningbranden vermindert.

Binnen Cluster 3 blijkt dat een hoger percentage inwoners met een Marokkaanse achtergrond en een hoger percentage inwoners tussen de 25 en 45 jaar oud samengaat met een toename van het aantal woningbranden in een buurt. Aan de andere kant is er een verband waarbij een groter aantal huishoudens zonder kinderen geassocieerd wordt met een afname van het aantal woningbranden in een buurt.

Binnen Cluster 4 blijkt dat een hoger percentage koopwoningen in een buurt samengaat met een lagere kans op woningbrand. Daarnaast wordt een lagere kans op woningbrand gemeten met een hoger percentage inwoners dat aangeeft hun gezondheid als goed te ervaren. Aan de andere kant laat een hoger percentage huurwoningen van woningcorporaties in een buurt een verhoogde kans op woningbrand zien, evenals een hoger percentage woningen met een bouwjaar voor 2000 en een hoger percentage verweduwde mensen.

### VOORSPELLING MET HET AANTAL HUISHOUDENS

Daarnaast voeren we een analyse uit waarmee we het aantal woningbranden per buurt kunnen voorspellen. We hebben ontdekt dat het aantal huishoudens een belangrijke voorspeller is voor het aantal woningbranden. Om de voorspellingen zo accuraat mogelijk te maken, nemen we als eerste stap het aantal huishoudens wel als één van de voorspellende variabelen mee.

Nadat we de voorspellingen hebben gedaan, passen we een explainer methode toe, op dezelfde manier als in het Hoofdstuk Clustering van buurten. Deze methode stelt ons in staat om te begrijpen waarom een bepaalde

voorspelling is gedaan, naast het aantal huishoudens. Het geeft ons inzicht in de specifieke factoren en kenmerken van een buurt die hebben bijgedragen aan de voorspelling.

In Appendix F wordt uitgebreider uitgelegd hoe deze analyse is uitgevoerd en hoe het explainer model is toegepast. Hier vind je meer gedetailleerde informatie over de methodologie en de stappen die zijn genomen om een beter begrip te krijgen van de voorspellende factoren en de redenering achter de voorspellingen voor elke individuele buurt. Ook bij deze methode zie je weer verschillen in welke variabelen naar voren komen om een goede voorspelling te maken van het aantal woningbranden, per cluster en per buurt.

## DATASET VERRIJKEN MET BRANDOORZAKEN

Om meer inzicht te krijgen in woningbrandincidenten, hebben we onze dataset uitgebreid met informatie over de oorzaken en locaties van woningbranden uit de TBO en AG5 dataset. We hebben interessante inzichten verkregen, met name dat brandoorzaken en -locaties kunnen verschillen in de buurtclusters. De TBO en AG5 dataset bevat gegevens zoals incidentnummers, evenals kenmerken van de brand, zoals de plaats in de woning waar de brand is ontstaan en de vermoedelijke oorzaak ervan. De AG5 dataset bevat kladbokregels die gegevens bevatten over het type brand. Door het incidentnummer te gebruiken, kunnen we de TBO en AG5 data koppelen aan de dataset uit het meldkamersysteem, en daarmee bepalen in welk buurtcluster de brand plaats vond. De TBO en AG5 dataset bevat gegevens over brandincidenten uit de jaren 2019, 2020 en 2021. De dataset afkomstig van het meldkamersysteem ging over een bredere periode namelijk 2013 tot en met 2022. Het is verder belangrijk op te merken dat de jaren 2020 en 2021 werden beïnvloed door de coronapandemie, wat mogelijk van invloed is geweest op het soort brandincidenten dat heeft plaatsgevonden. De TBO en AG5 dataset bevat geen aanvullende informatie over het type gebouw, letsel of geschatte schade als gevolg van een brand. Daarom is het essentieel om nog andere bronnen te raadplegen waarmee we onze GMS-dataset verder kunnen verrijken.

Om ons onderzoek naar woningbranden in verschillende clusters te versterken, hebben we contact opgenomen met Stichting Salvage om kennis en gegevens uit te wisselen. Stichting Salvage verleent eerste hulp aan slachtoffers van brandschade (en ook bliksem-, explosie-, water-, storm- en aanrijdingschade) namens de gezamenlijke brandverzekeraars in Nederland. Bij elk incident waar ze ter plaatse komen, worden specifieke kenmerken van de brand genoteerd. We hebben gevraagd data te leveren op het niveau van CBS-buurten, zodat we ze kunnen koppelen aan de CBS-buurtindeling en de bijbehorende clusterinformatie. Door nauw samen te werken met Stichting Salvage kunnen we waardevolle gegevens verkrijgen die ons een diepgaander inzicht bieden in de woningbranden en hun kenmerken op het niveau van buurtclusters.

In Tabel 1 wordt een overzicht gegeven van het aantal branden per cluster op basis van gegevens uit het meldkamersysteem, evenals het aantal branden per cluster in de AG5- en TBO-datasets en het aantal branden in de Salvage-dataset. Om het aantal branden in perspectief te plaatsen, is ook het aantal buurten per cluster vermeld. Het is belangrijk op te merken dat er in Cluster 1 en 4 aanzienlijk meer buurten zijn opgenomen volgens het clusteralgoritme, waardoor er automatisch meer branden worden waargenomen in die clusters. De getoonde aantallen branden zijn cumulatief opgeteld.

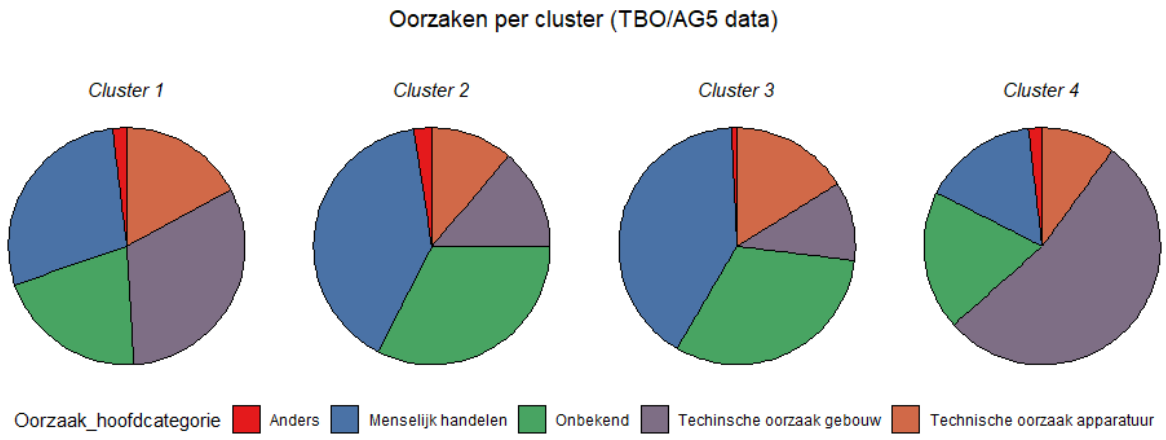
Het is interessant om op te merken dat zowel Stichting Salvage als het team Brandonderzoek betrokken zijn geweest bij verschillende brandincidenten in de verschillende buurtclusters. Het lijkt erop dat Stichting Salvage vaker aanwezig was bij woningbranden in Cluster 2 en Cluster 3, terwijl het team Brandonderzoek juist meer branden heeft onderzocht in Cluster 4. Allebei de bronnen kunnen belangrijk zijn in het verder verkennen van de woningbranden in de verschillende clusters.

**TABEL 1.** Per cluster het aantal buurten, aantal branden uit GMS dataset, aantal branden uit Salvage dataset, en aantal branden uit AG5/TBO dataset.

| **Cluster** | **Aantal buurten** | **Aantal branden GMS (meldkamersysteem)** | **Aantal branden Salvage** | **Aantal branden AG5/TBO** |
|---|---|---|---|---|
| 1 | 209 | 1447 | 476 | 429 |
| 2 | 65 | 557 | 213 | 108 |
| 3 | 86 | 736 | 336 | 159 |
| 4 | 249 | 1178 | 228 | 379 |

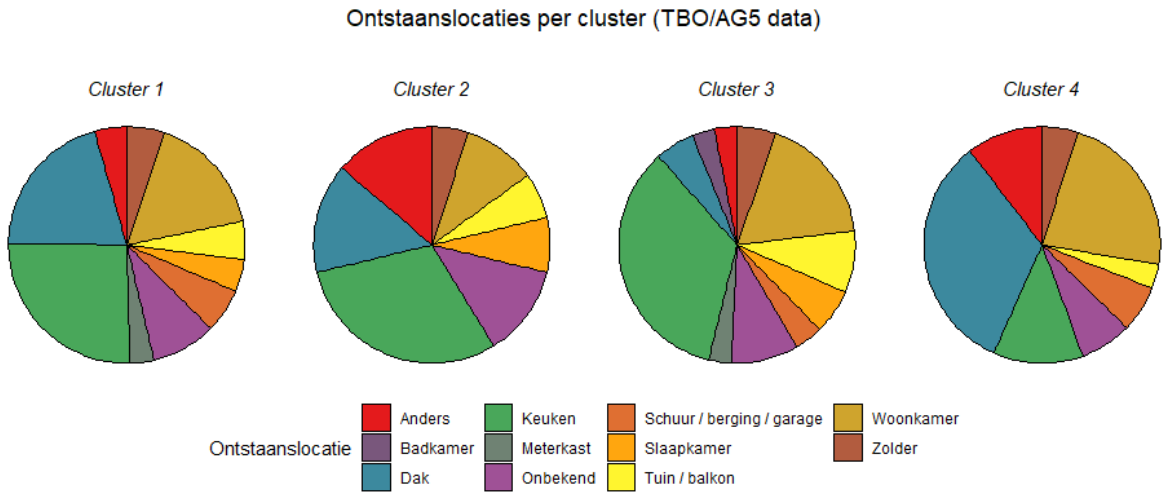
### *RESULTATEN UIT AG5/TBO DATASET*

In Figuur 6 worden de oorzaken per cluster weergegeven op basis van de AG5/TBO-dataset. In Cluster 4 valt op dat technische oorzaak gebouw een groot aandeel heeft gehad in woningbranden, terwijl in Cluster 2 en Cluster 3 menselijk handelen vaker voorkomt. Het is echter belangrijk op te merken dat menselijk handelen ook indirect verband kan houden met andere brandoorzaken.




**FIGUUR 6**

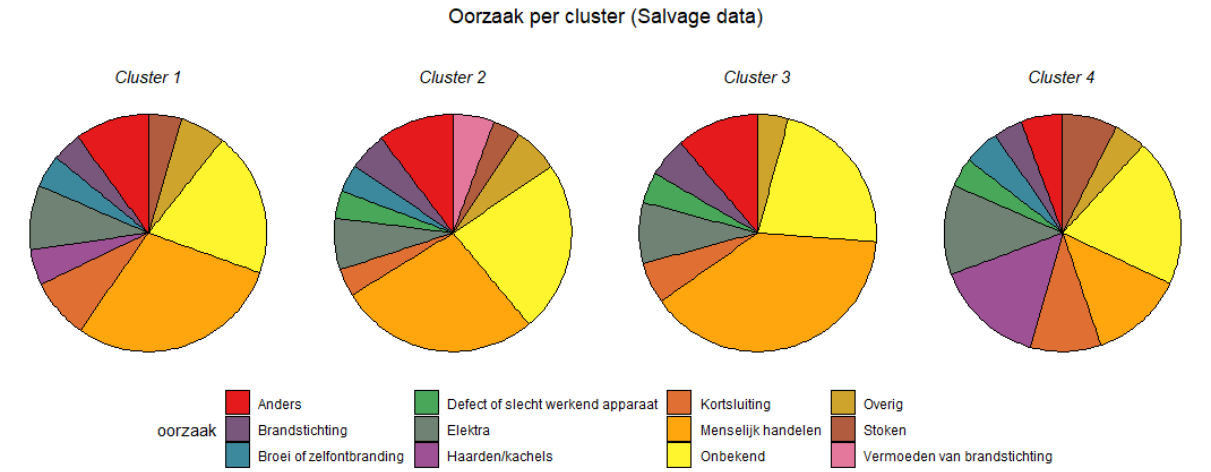
In Figuur 7 wordt de verdeling van de ontstaanslocaties getoond uit de AG5/TBO dataset. Binnen Cluster 4 is er een duidelijke concentratie van dak branden en in Cluster 3 van keuken branden.




**FIGUUR 7**

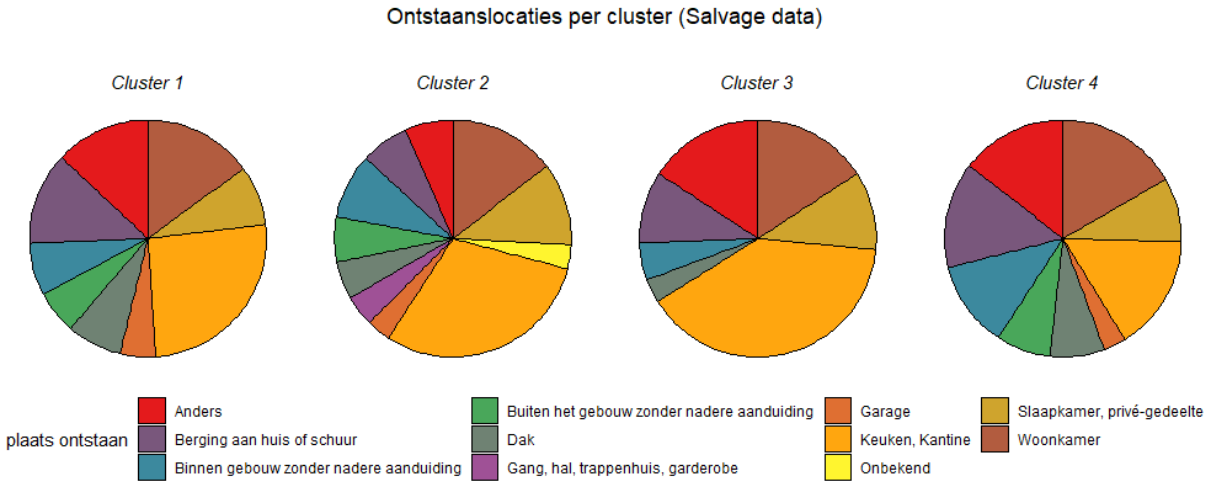
*RESULTATEN UIT SALVAGE DATASET*

In Figuur 8 wordt de verdeling van de verschillende oorzaken per cluster weergegeven uit de Salvage data. Wat opvalt is dat vooral in Cluster 3 een groot aandeel menselijk handelen is en in Cluster 4 valt het aandeel haarden/kachels op.




**FIGUUR 8**

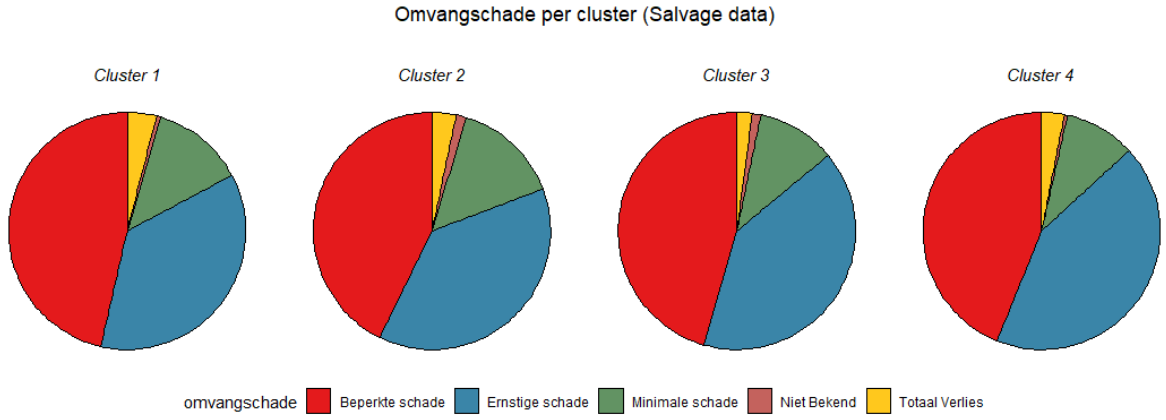
In Figuur 9 wordt de verdeling van de ontstaanslocaties uit de Salvage data weergegeven per cluster. Ook hier zie je weer verschillen tussen de clusters. Wat opvalt is dat Cluster 3 voornamelijk keukenbranden heeft, terwijl dit in Cluster 4 aanzienlijk minder voorkomt. Cluster 4 heeft juist vaker branden die ontstaan zijn in de berging aan huis of schuur of binnen gebouw zonder nadere aanduiding.



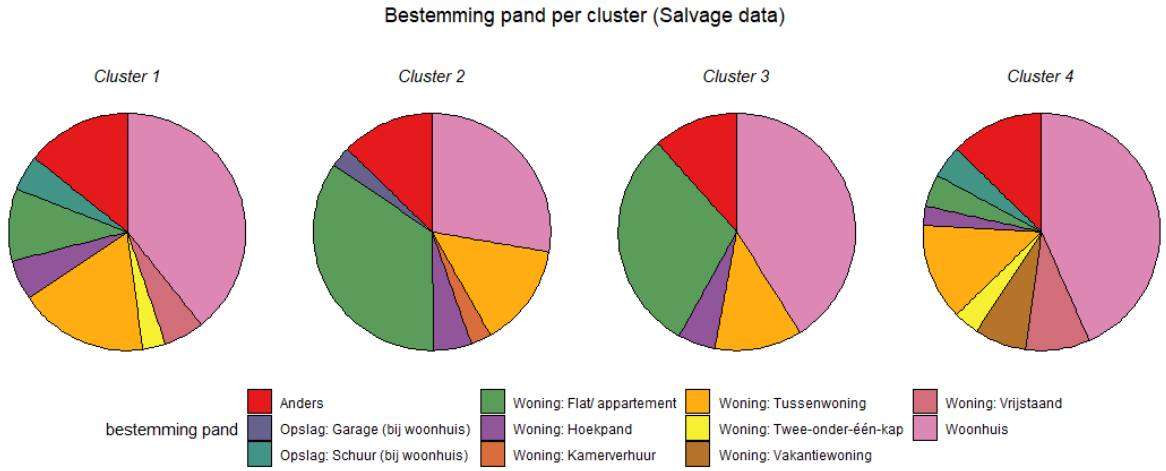

**FIGUUR 9**

De dataset van Salvage biedt daarnaast ook informatie over omvangschade van de woningbrand (Figuur 10), type gebouw (Figuur 11), letsel (Figuur 12) en of er sprake is van hennep (Figuur 13).

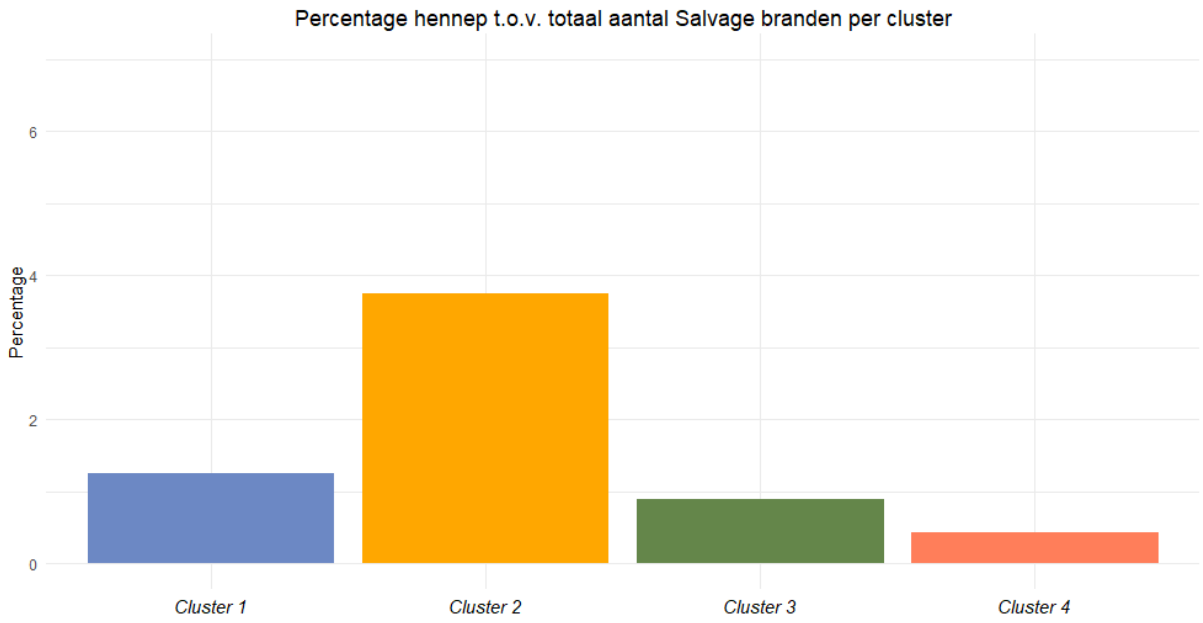
Figuur 10 laat ook iets opvallends zien, namelijk dat de geschatte omvangschade per cluster vrijwel identiek is. Dus hoewel er absoluut gezien in Cluster 3 per 1000 huishoudens meer woningbranden zijn, is de gemiddelde procentuele omvangschade van de branden niet anders dan in de andere clusters. Verder zie je in Figuur 11 dat er in Cluster 2 en Cluster 3 voornamelijk branden zijn in appartementen/flats.



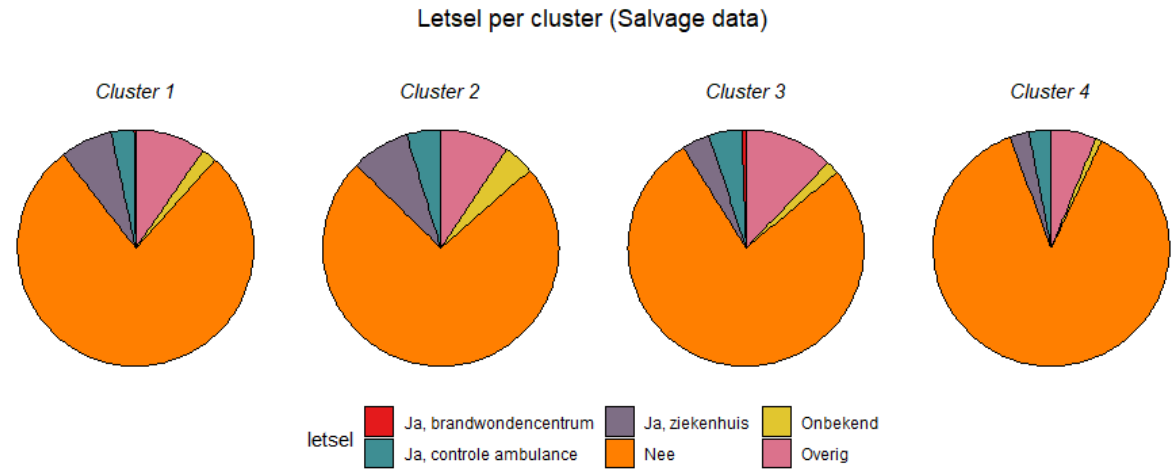

**FIGUUR 10**




**FIGUUR 11**




**FIGUUR 12**




**FIGUUR 13**

Naast het analyseren van de visualisaties is het ook van belang om te onderzoeken of de verschillen in oorzaken, ontstaanslocaties, type woning en letsel significant verschillen tussen de clusters. In Appendix G worden verschillende tabellen gepresenteerd met de actuele aantallen per cluster, evenals de bijbehorende proporties en informatie over de significantie van deze verschillen.

# SAMENWERKINGEN

## SAMENWERKING MET VEILIGHEIDSREGIO GELDERLAND-MIDDEN

Het project dat is uitgevoerd, is gericht op de Veiligheidsregio Midden- en West-Brabant en de analyses zijn specifiek voor die regio uitgevoerd. Echter, binnen de projectgroep werken we samen met een data analist van de Veiligheidsregio Gelderland-Midden. Deze samenwerking biedt vele voordelen, omdat we op deze manier kennis en expertise kunnen bundelen gedurende het hele project. Door intensief contact te onderhouden, kunnen we snel schakelen bij experimenten en profiteren van elkaars ervaringen. Een bijkomend voordeel van deze samenwerking is dat we ook relatief snel de analyses kunnen uitvoeren voor de regio Gelderland-Midden. Zo kunnen we controleren of de resultaten vergelijkbaar zijn met die van de regio Midden- en West-Brabant. Dit is van groot belang, omdat het ons inzicht geeft in de betrouwbaarheid van de analyses en de mogelijke generaliseerbaarheid van de resultaten naar andere regio's. De eerste resultaten zijn positief gebleken, aangezien de analyses die voor de regio Midden- en West-Brabant zijn uitgevoerd ook succesvol herhaald konden worden voor de regio Gelderland-Midden.

## UITKOMSTEN VAN DE SAMENWERKING MET NIPV

Ons algoritme is gedeeld met het Nederlands Instituut Publieke Veiligheid (NIPV) om de reproduceerbaarheid van onze analyses en de generaliseerbaarheid van onze resultaten voor heel Nederland te onderzoeken. Het eerste aspect dat is onderzocht, is of het aantal buurtclusters en de kenmerken van die buurtclusters in onze studie overeenkomen met de buurtclusters voor heel Nederland. We hebben vastgesteld dat dit inderdaad het geval is, wat een positief resultaat is. Dit impliceert dat de regio Midden- en West-Brabant een representatieve weerspiegeling biedt van diverse buurttypes, meer en minder verstedelijkt, die in heel Nederland voorkomen.

Daarnaast hebben we gekeken of er verschillen zijn in het aantal woningbranden tussen de buurtclusters in heel Nederland, vergelijkbaar met de verschillen die we hebben waargenomen in de buurtclusters van de regio Midden- en West-Brabant. We hebben vastgesteld dat er specifieke buurtclusters zijn waarin het aantal woningbranden relatief sneller stijgt in verhouding tot het aantal huishoudens. Deze patronen komen zowel voor in de buurtclusters van de regio Midden- en West-Brabant als in de bredere context van heel Nederland.

Deze uitgebreide analyse benadrukt de relevantie van de buurtclusters in de regio Midden- en West-Brabant als representatieve eenheden voor heel Nederland. Bovendien onderstreept het de aanwezigheid van specifieke clusters waarin het aantal woningbranden een verhoogde trend vertoont, wat wijst op een mogelijk verhoogd risico in die gebieden. Dit biedt waardevolle inzichten voor het ontwikkelen van gerichte brandpreventiemaatregelen en interventiestrategieën, niet alleen in de regio Midden- en West-Brabant, maar ook in andere veiligheidsregio's in Nederland.

# VERVOLGACTIES EN TOEPASSINGSMOGELIJKHEDEN

Dit onderzoek zou over ongeveer 4 jaar herhaald moeten worden, omdat buurten, buurtkenmerken en risico's kunnen veranderen. Daarom hebben we het gebruikte algoritme opgeslagen in het register van de VRMWB.

De vervolgacties in najaar 2023 zijn gericht op overdracht van de resultaten zodat deze toegepast kunnen gaan worden in:

- Dekkings- en spreidingsplan om bij verspreiding van materieel en bij vraagstukken over herplaatsing van kazernes het kaartmateriaal te benutten, en de kans op woningbrand in de buurtclusters mee te wegen.
- Mobiliteitsplannen van gemeenten om het kaartmateriaal te benutten zodat de invalshoek kans op woningbrand wordt betrokken bij mobiliteits- en bereikbaarheidsvraagstukken in gemeenten.
- Veilige en gezonde wijkenaanpak in gemeenten om de buurtkenmerken en kans op woningbrand te benutten in de dialoog met veiligheidspartners over een veilige en gezonde wijkaanpak.
- Risicomonitoring waaronder brandrisicoprofielen omdat afgezet tegen het aantal huishoudens de buurten in Cluster 2 en 3 een steilere samenhang hebben met woningbrand. De buurten in Cluster 3 hebben per 1000 huishoudens meer woningbrand. De buurtclusters verschillen onderling in brandoorzaak en ontstaanslocatie.
- Interventies brandveilig leven waaronder risicocommunicatie: er kan gericht in activiteiten Brandveilig Leven aandacht worden besteed aan bepaalde brandoorzaken en hoe deze te voorkómen in een buurtcluster.

# GELEERDE LESSEN

## BUURTINDELINGEN (2016 & 2020) VERANDEREN

Dit project is gestart met het gebruik van gegevens van de GGD en het CBS uit 2016. Elk jaar levert het CBS nieuwe gegevens, terwijl de GGD-Gezondheidsmonitor eens in de vier jaar wordt uitgevoerd. Aan het begin van dit project was de meest recente GGD-dataset afkomstig uit 2016, aangezien de gegevens van 2020 nog niet beschikbaar waren. Tot en met juli 2022 zijn alle analyses gebaseerd op de gegevens van 2016. Toen de GGD-dataset voor 2020 beschikbaar kwam, werd overwogen om dezelfde clustering toe te passen op de nieuwere dataset. Een uitdaging bij het opnieuw uitvoeren van de analyses was dat de buurtindelingen van de CBS-gegevens ook in de loop der jaren zijn veranderd. Gemeentes bepalen zelf wanneer een buurt wordt opgesplitst of samengevoegd, en deze veranderingen kunnen elk jaar plaatsvinden. Dit betekende dat er een zorgvuldige afweging moest worden gemaakt bij het opnieuw uitvoeren van de analyses, zodat de gegevens van 2020 konden worden vergeleken met die van 2016. Het was noodzakelijk om te zorgen voor een consistente buurtindeling voor beide datasets (die uit 2016 en die uit 2020), anders zouden de resultaten niet vergelijkbaar zijn. Dit vormde dus een belangrijke uitdaging bij het updaten van de analyses voor de gegevens van 2020. Er zijn methoden gebruikt om de CBS- en GGD-data van 2020 op te splitsen of samen te voegen naar de buurtindeling van 2016. Vervolgens is er een nieuwe clusterindeling gemaakt, maar vanwege de verschillen tussen de variabelen konden de clusters niet goed worden vergeleken. De variabelen die door het CBS worden geleverd, kunnen van jaar tot jaar verschillen, dus sommige gegevens die wel beschikbaar waren in de dataset van 2016, waren niet meer beschikbaar in de dataset van 2020. Het clustering-algoritme gebruikt andere variabelen als invoer, waardoor het andere keuzes maakt over welke buurt in welke cluster moet worden geplaatst. Uiteindelijk is besloten om de buurtclustering van 2016 los te laten en opnieuw te beginnen met de dataset van 2020, maar alleen met variabelen die hoogstwaarschijnlijk elk jaar beschikbaar zullen zijn in de CBS-dataset. Om te bepalen welke variabelen vrijwel altijd terugkeren in de CBS-dataset, zijn alleen kolommen behouden die van 2015 tot 2022 consistent aanwezig waren. Wanneer er in 2024 weer een nieuwe dataset beschikbaar komt van de GGD, kan er opnieuw een analyse worden uitgevoerd.

## CONNECTIE MET DATABASE VOOR INCIDENTEN

Het huidige script laadt de open data van het CBS en de GGD in via een CBS API in R, maar het aantal woningbranden moet nog handmatig worden toegevoegd. Hiervoor wordt momenteel een csv-bestand gebruikt dat gefilterd wordt op alleen woningbranden tussen 2013 en 2023. Omdat de incident data in twee delen is geleverd, namelijk vanaf 2013 tot 2021 en voor de branden van 2022 apart, moeten deze eerst worden samengevoegd voordat het aantal woningbranden per buurt kan worden berekend.

In de toekomst raden wij aan om een connectie te maken met de database waar de incident data is opgeslagen, zodat altijd de meest actuele data beschikbaar is en er niet onnodig csv-bestanden met branddata opgeslagen hoeven te worden.

## GEBOUWEN ANALYSE

We hebben geprobeerd om de GMS data van het meldkamer systeem te koppelen aan de KRO aanzien dataset. De KRO aanzien dataset geeft informatie over het type gebouwen in de regio. Door deze datasets te koppelen via de geometrie van de gebouwen, konden we informatie verzamelen over in welk type gebouw de woningbrand plaats had. We hoopten hiermee te kunnen achterhalen welk type woningen het meest brandgevoelig was, zodat we betere brandpreventie-adviezen konden geven.

Helaas bleek bij het koppelen van de datasets dat we van 677 woningbranden het gebouwtype niet konden achterhalen. Bovendien bleek de registratie van de gebouwtypen erg vervuild te zijn, waardoor het ons geen extra informatie opleverde. We hebben geprobeerd om gebruiksklasse en bouwbestemmingen van de panden te gebruiken als aanvullende informatie, maar deze gaven ons ook geen bruikbare gegevens voor ons onderzoek. Soms kwamen we woningen tegen die waren bedoeld voor een specifieke doelgroep, maar we hadden geen

duidelijk begrip van wie deze doelgroep was. Bovendien werd deze informatie niet altijd geregistreerd omdat dit handmatig moest worden ingevoerd.

### BUURTINDELING FUNCTIE NIET ALTIJD ROBUUST

Aan het begin van ons project hebben we gebruik gemaakt van het nationaal geo-register om via R de accurate buurtindeling op te halen van een specifiek jaar in een specifieke regio. Dit verliep lange tijd zonder problemen, totdat we op een gegeven moment merkten dat de URL niet meer werkte in onze functie. Na onderzoek bleek dat er intern bij het nationaal geo-register wijzigingen werden doorgevoerd, waardoor onze functie niet meer goed werkte. Om dit probleem op te lossen, hebben we de functie aangepast zodat deze de buurtindeling kon ophalen via PDOK. Hierdoor konden we de functie blijven gebruiken en ons onderzoek voortzetten. Het is echter belangrijk om te beseffen dat externe websites en bronnen kunnen veranderen, waardoor bepaalde functies mogelijk niet meer werken. Dit kan van invloed zijn op de betrouwbaarheid en reproduceerbaarheid van ons onderzoek. Het is daarom belangrijk om hier rekening mee te houden en eventuele wijzigingen in externe bronnen op te merken en hierop te anticiperen.

# APPENDIX

Het R-script voor onderstaande analyses is opgeslagen in het algoritmeregister van de Veiligheidsregio Midden- en West-Brabant. De tooling die is gebruikt voor dit project is R en RStudio, inclusief de benodigde packages.

## APPENDIX A: PREPROCESSING

### *MISSENDE WAARDES*

Tijdens het onderzoek hebben we ontdekt dat er in sommige kolommen missende waardes voorkwamen. Om de kwaliteit van onze dataset te waarborgen, hebben we besloten om alle kolommen met meer dan 5% missende waardes te verwijderen. Voor de kolommen waar minder dan 5% missende waardes waren, hebben we de missende waardes opgevuld met behulp van de 'mice' functie uit het *mice* package van R. De 'mice' functie is een imputatiemethode die ontworpen is om missende waardes in datasets op te vullen. Het package gebruikt meervoudige imputatie om de ontbrekende waardes te schatten op basis van de overige waardes in de dataset. Dit wordt gedaan door modellen te maken voor elke variabele met missende waardes en deze modellen te gebruiken om de missende waardes op te vullen. De 'mice' functie gebruikt de Predictive Mean Matching (PMM) methode om de ontbrekende waarden te schatten op basis van andere waarden in de dataset. De functie produceert in ons geval 5 verschillende geïmputeerde datasets. Ten slotte worden de geïmputeerde waarden samengevoegd met de oorspronkelijke dataset. De uiteindelijke dataset bevat dan geen ontbrekende waarden meer in de geselecteerde numerieke kolommen en kan worden gebruikt voor verdere analyses.

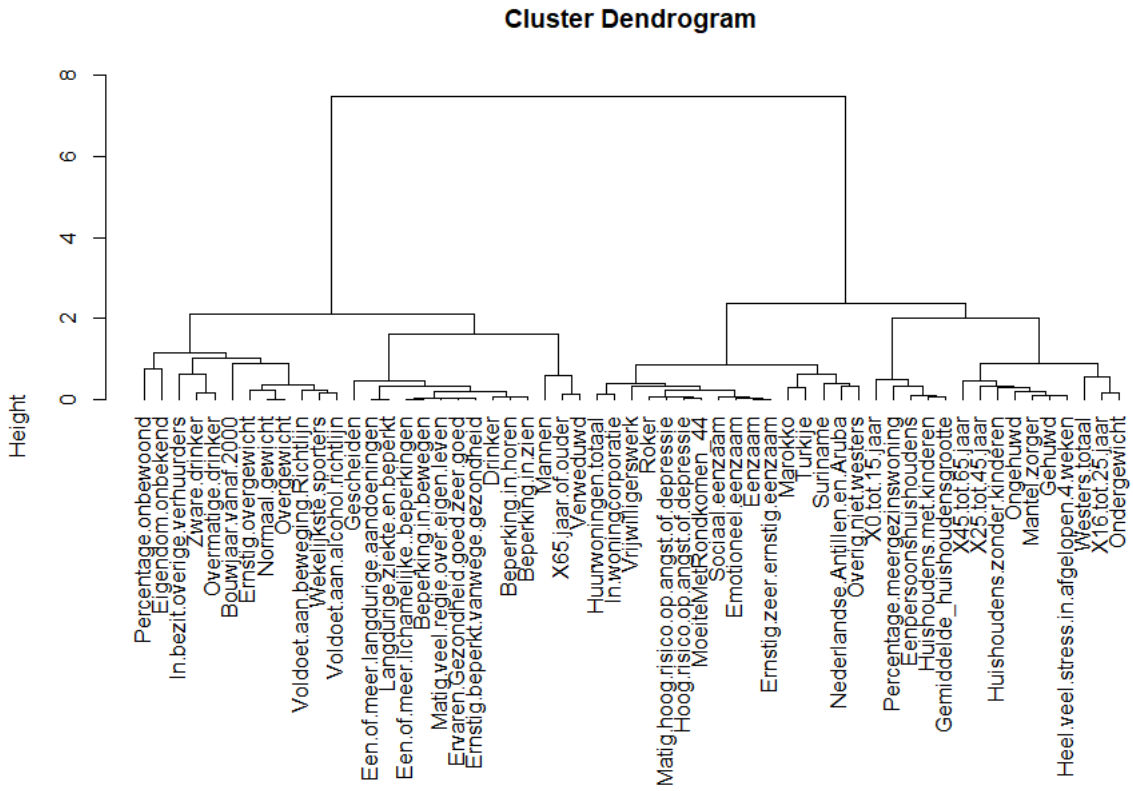
### *FACTOR ANALYSE*

We hebben ook een factoranalyse uitgevoerd om te begrijpen hoe de verschillende variabelen gerelateerd zijn aan elkaar en om ze te groeperen op basis van deze relaties. Bij variabelen die sterk met elkaar correleren, hebben we ervoor gekozen om slechts één variabele te behouden die het meest bijdraagt aan de variatie in de dataset. Hiermee voorkomen we dat we variabelen behouden die in feite hetzelfde aangeven en behouden we alleen de belangrijkste variabelen die de meeste informatie bevatten. Dit helpt ons om de dataset te vereenvoudigen en de nadruk te leggen op de belangrijkste variabelen.

De Kaiser-Meyer-Olkin (KMO) test is een maatstaf voor de geschiktheid van de gegevens voor factoranalyse. Het meet de mate van gemeenschappelijke variantie tussen de variabelen en geeft aan in hoeverre deze gegevens geschikt zijn voor factoranalyse. De data is getransformeerd om aan de aannames te voldoen. De KMO-waarde op onze getransformeerde dataset geeft een waarde van 0,92 en dit geeft aan dat de gemeenschappelijke variantie tussen de variabelen hoog is en dat de gegevens goed geschikt zijn voor factoranalyse.

Het kan voorkomen dat sommige kolommen een zeer hoge correlatie vertonen, wat betekent dat ze bijna perfect met elkaar overeenkomen. In dit geval kan het verwijderen van een van deze kolommen voordelig zijn. Hoog correlerende variabelen bevatten vaak dezelfde informatie en dragen daardoor niet bij aan het voorspellende vermogen van een model. Ook kan het makkelijker worden om het model te interpreteren.

Met behulp van hiërarchische clustering kunnen we groepen van variabelen identificeren die vergelijkbare eigenschappen hebben en daarom dicht bij elkaar liggen. Dit kan ons helpen om patronen en relaties tussen de variabelen te begrijpen en te visualiseren. In Figuur 14 zie je een overzicht van welke variabelen bij elkaar zouden worden gegroepeerd.




**FIGUUR 14**

Een dendrogram is een grafische weergave waarbij de onderlinge afstand tussen de variabelen wordt weergegeven in de vorm van een boomstructuur. Het kan gebruikt worden om inzicht te krijgen in de onderliggende structuur van de data en om groepen van variabelen te identificeren die vergelijkbaar zijn. Elk blad aan de onderkant van het dendrogram vertegenwoordigt een variabele in de dataset. De variabelen worden vervolgens samengevoegd in factoren, die worden weergegeven door de takken van het dendrogram. Door te kijken naar de lengte van de lijnen in het dendrogram kan je een idee krijgen van de mate van gelijkenis tussen de variabelen. Hoe langer de lijn, hoe minder gelijk de variabelen zijn. Variabelen die zeer vergelijkbaar zijn, worden vaak samengevoegd tot één groep, terwijl variabelen die minder op elkaar lijken, in afzonderlijke groepen blijven. Als er twee variabelen zijn die heel dicht bij elkaar staan in het dendrogram en dus zeer vergelijkbaar zijn, kan dit suggereren dat deze variabelen sterk gecorreleerd zijn en dat één van de twee kan worden verwijderd uit de dataset zonder veel informatie te verliezen.

Daarnaast hebben we gekeken bij de factor analyse en bij het clusteren van de variabelen welke variabelen het minste bijdragen aan de synthetische factor. Bij de factoranalyse kan je kijken naar de communaliteit van de variabelen, wat aangeeft hoeveel van de variantie in de oorspronkelijke variabele wordt verklaard door de factoren. Variabelen met een lage communaliteit dragen dus weinig bij aan de synthetische factor en kunnen eventueel worden verwijderd om de dataset te vereenvoudigen. Bij de analyse kan je kijken naar de squared loading correlation, die aangeeft hoeveel van de variantie in de oorspronkelijke variabele wordt verklaard door de factor. Variabelen met een lage squared loading correlation dragen dus weinig bij en kunnen eventueel worden verwijderd; deze variabelen halen we weg om de dataset te vereenvoudigen.

Een overzicht van de variabelen die wij hebben verwijderd uit de dataset zijn:

| Variabele (in %) | Reden |
|---|---|
| Een of meer langdurige aandoeningen | Overlapt met andere variabele |
| Ernstig beperkt vanwege gezondheid | Overlapt met andere variabele |
| Beperking in bewegen | Overlapt met andere variabele |
| Beperking in zien | Overlapt met andere variabele |
| Beperking in horen | Overlapt met andere variabele |
| Langdurige ziekte en beperkt | Overlapt met andere variabele |
| Eenzaam | Overlapt met andere variabele |
| Sociaal eenzaam | Overlapt met andere variabele |
| Emotioneel eenzaam | Overlapt met andere variabele |
| Matig hoog risico op angst of depressie | Overlapt met andere variabele |
| Voldoet aan beweging Richtlijn | Overlapt met andere variabele |
| Overmatige drinker | Overlapt met andere variabele |
| Voldoet aan alcohol richtlijn | Overlapt met andere variabele |
| Matig veel regie over eigen leven | Overlapt met andere variabele |
| Mantelzorger | Overlapt met andere variabele |
| Normaal gewicht | Overlapt met andere variabele |
| Ernstig/zeer ernstig eenzaam | Overlapt met andere variabele |
| Eigendom onbekend | Weinig toevoeging |
| Onbewoond | Weinig toevoeging |
| Mannen | Weinig toevoeging |
| Eengezinswoning | Overlapt met andere variabele |
| Bewoond | Overlapt met andere variabele |
| Koopwoningen totaal | Overlapt met andere variabele |
| Bouwjaar voor 2000 | Overlapt met andere variabele |
| Beperkt vanwege gezondheid | Overlapt met andere variabele |

#### *VAN AANTALLEN NAAR RELATIEF*

We hebben alle variabelen omgezet naar percentages in plaats van aantallen om de data te normaliseren, omdat buurten sterk kunnen verschillen in grootte en aantal huishoudens. Hierdoor wordt de data relatief en zijn de buurten beter te vergelijken.

#### *RESULTAAT NA HET OPSCHONEN*

Na het opschonen zijn de volgende variabelen overgebleven die wij meenemen in dit onderzoek:

#### VARIABELEN

| |
|---|
| Buurtnaam |
| Aantal inwoners |
| Aantal huishoudens |
| % 0 tot 15 jaar |
| % 16 tot 25 jaar |
| % 25 tot 45 jaar |
| % 45 tot 65 jaar |
| % 65 jaar of ouder |
| % Een of meer lichamelijke beperkingen |
| % Ernstig overgewicht |
| % Ernstige geluidhinder door buren |
| % Ervaren Gezondheid goed/zeer goed |
| % Heel veel stress in afgelopen 4 weken |

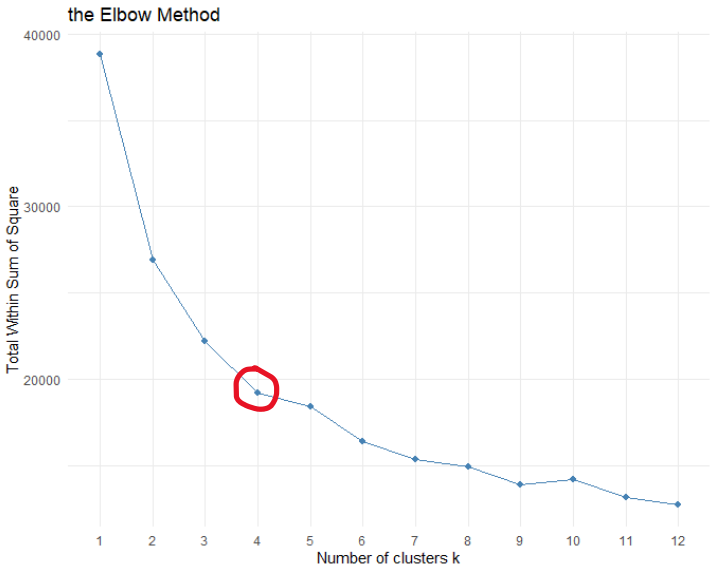
| |
|---|
| % Hoog risico op angst of depressie |
| % Huishoudens met kinderen |
| % Huishoudens zonder kinderen |
| % Huurwoningen totaal |
| % In woningcorporatie |
| % Moeite met rondkomen |
| % Nederlandse Antillen en Aruba |
| % Overig niet westers |
| % Percentage meergezinswoning |
| % Wekelijkse sporters |
| % Westers totaal |
| % Zware drinker |
| % Drinker |
| % Eenpersoonshuishoudens |
| % Gehuwd |
| % Bouwjaar na 2000 |
| % In bezit overige verhuurders |
| % Gescheiden |
| % Marokko |
| % Ondergewicht |
| % Ongehuwd |
| % Overgewicht |
| % Roker |
| % Suriname |
| % Turkije |
| % Verweduwd |
| % Vrijwilligerswerk |

## APPENDIX B: CLUSTERING

De variabelen aantal huishoudens en aantal inwoners worden niet meegenomen in de clustering functie omdat we puur geïnteresseerd zijn in de clustering van de relatieve kenmerken van een buurt (bijv. percentage rokers) en niet in of een buurt een groot of klein aantal inwoners heeft.

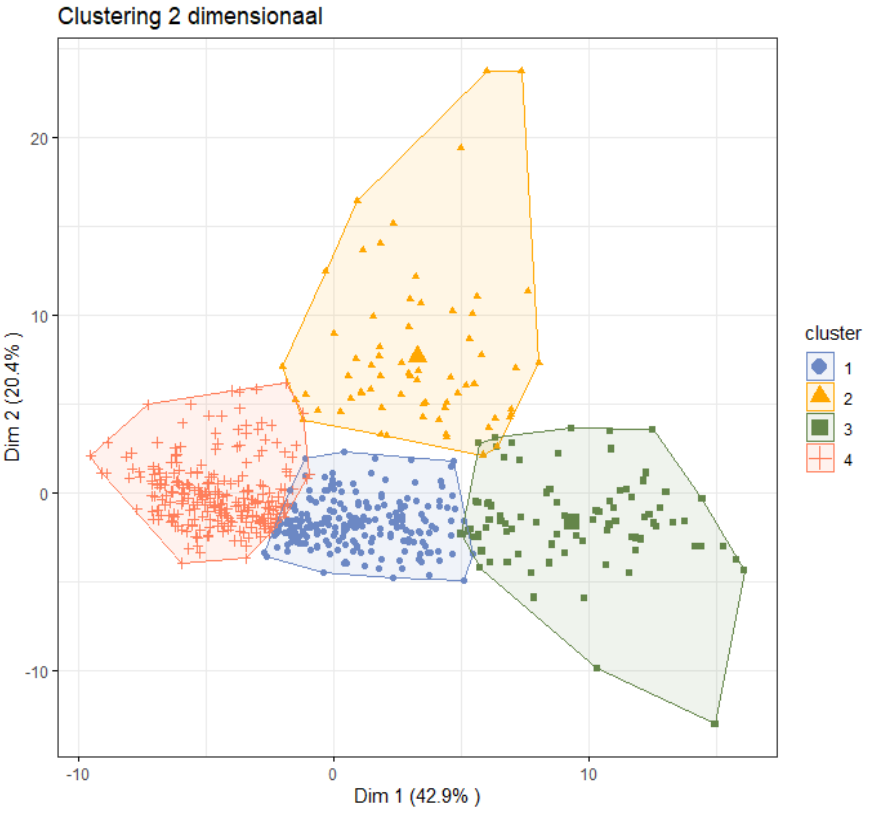
We hebben besloten om het K-means cluster algoritme te gebruiken omdat dit een veelgebruikte en robuuste methode is om groepen of clusters te vormen op basis van gegevens. K-means is een *unsupervised learning* algoritme dat iteratief datapunten groepeert in *k* clusters, waarbij elk datapunt wordt toegewezen aan het dichtstbijzijnde centrum van het cluster.

Om te bepalen hoeveel clusters we zouden gebruiken, hebben we een techniek gebruikt die bekend staat als de 'elbow methode'. Hierbij hebben we het aantal clusters gevarieerd en gekeken naar de verandering van de som van de kwadraten van de afstanden tussen de datapunten en hun toegewezen clustercentra. Het punt waar de verandering niet beduidend verder daalt, wordt de 'elleboog' genoemd en geeft aan dat verdere toename van het aantal clusters weinig verschil zal maken in de groepering. Zie Figuur 15. We hebben deze methode toegepast en hieruit bleek dat *k=4* het beste aantal clusters was voor onze analyse.



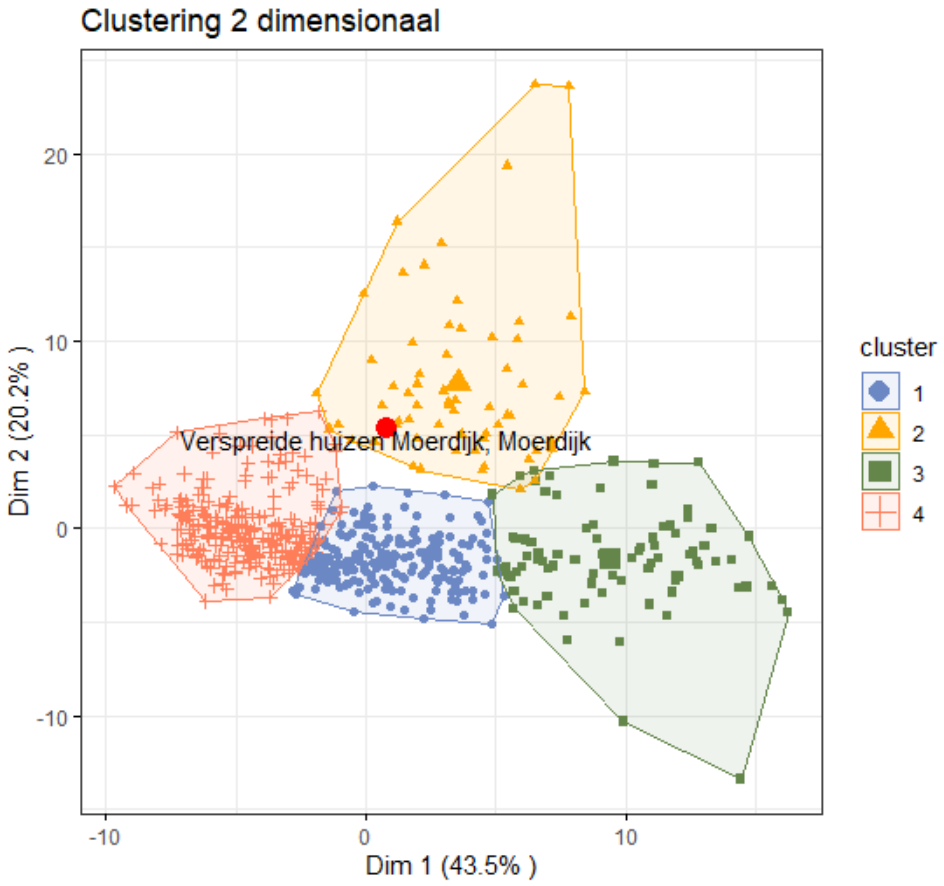

**FIGUUR 15**

De plot die in Figuur 16 wordt weergegeven is een visualisatie van de clustering resultaten van een K-means algoritme. De dataset is geschaald en verminderd tot 2 dimensies met behulp van de Principal Component Analysis (PCA) methode. De X-as toont de waarde van de Principal Component (Dim 1) en de Y-as toont de waarde van de tweede Principal Component (Dim 2). Elke buurt in de dataset wordt weergegeven als een punt in de plot en gekleurd volgens het cluster waartoe het behoort. Hierin zie je welke buurten dichter bij elkaar liggen en welke buurten 'outliers' zijn, oftewel welke buurten zich verder van de andere buurten in hun cluster bevinden en daardoor een afwijkend profiel hebben.




**FIGUUR 16**

Na overleg met brandweerexperts uit de regio hebben we de buurt 'Verspreide huizen Moerdijk, Moerdijk' handmatig toegewezen aan Cluster 4. Uit de analyse bleek dat de buurt uit Moerdijk beter past in Cluster 4, aangezien dit cluster vooral de buurten omvat die zich meer in het buitengebied bevinden. Zie Figuur 17 voor verheldering waar die buurt eerst in de cluster indeling lag.




**FIGUUR 17**

## APPENDIX C: CLASSIFICATIE VAN DE BUURTEN OP DE CLUSTER INDELING MET FEATURE IMPORTANCES

### *BELANGRIJKHEID VAN DE KENMERKEN OM DE CLUSTERS TE VOORSPELLEN MET EEN CLASSIFICATIE MODEL*

Om nog dieper inzicht te krijgen in de indeling van de buurten in de clusters, hebben we ook een classificatiemodel getraind op onze data, waarbij we alle CBS- en GGD-kenmerken gebruiken per buurt om te voorspellen in welk cluster deze buurt zou moeten vallen. Dit hebben we gedaan om beter te begrijpen waarom de buurten in bepaalde clusters zijn ingedeeld. We hebben gebruik gemaakt van methoden om te bepalen welke variabelen het belangrijkst waren voor het model om een bepaalde buurt in een specifiek cluster in te delen. Je ziet in Tabel 2 belangrijkheid van de variabelen een overzicht met de variabelen en hoe belangrijk ze waren voor het model. Zoals verwacht zijn er verschillen tussen de clusters. Als voorbeeld: voor Cluster 2 geldt dat 'Percentage eenpersoonshuishoudens' en 'Percentage inwoners 16 tot 25 jaar' relatief belangrijker zijn dan voor andere clusters. Uit onze analyse blijkt ook dat de variabele 'Percentage inwoners ervaren gezondheid goed/zeer goed' van groot belang is voor zowel Cluster 1 als Cluster 4. Daarnaast hebben we vastgesteld dat het 'Percentage inwoners met een Marokkaanse achtergrond' van belang is voor de indeling van buurten in Cluster 3. Het is echter belangrijk op te merken dat deze methode niet aangeeft of een variabele een positieve of negatieve invloed heeft op de voorspelling van het model. We kunnen alleen vaststellen hoe belangrijk een variabele is

geweest voor het indelen van een buurt in een bepaald cluster. Nu hebben we tekstueel één variabele toegelicht, maar houd er rekening mee dat er meerdere factoren in onderlinge samenhang hebben meegespeeld bij de clustering van buurten.

**TABEL 2** Belangrijkheid variabelen per cluster voor classificatie clusterindeling.

| | Cluster1 | Cluster2 | Cluster3 | Cluster4 |
|---|---|---|---|---|
| % 0 tot 15 jaar | 28,1 | 59,75 | 31,01 | 44,76 |
| % 16 tot 25 jaar | 60,64 | 55,12 | 21,57 | 17,64 |
| % 25 tot 45 jaar | 42,39 | 47,36 | 19,77 | 30,31 |
| % 45 tot 65 jaar | 38,46 | 51,59 | 35,78 | 43,26 |
| % 65 jaar of ouder | 46,83 | 15,24 | 28,53 | 39,93 |
| % Ongehuwd | 66,13 | 62,25 | 28,04 | 33,84 |
| % Gehuwd | 62,98 | 77,01 | 34,86 | 51,76 |
| % Gescheiden | 44,62 | 0,28 | 52,8 | 69,2 |
| % Verweduwd | 60,26 | 12,54 | 27,57 | 58,54 |
| % Westers totaal | 0 | 51,99 | 38,36 | 34,14 |
| % Marokko | 5,33 | 22,54 | 61,25 | 43,37 |
| % Nederlandse Antillen en Aruba | 26,03 | 30,27 | 25,34 | 39,55 |
| % Suriname | 30,9 | 30,63 | 23,86 | 36,61 |
| % Turkije | 7,71 | 21,79 | 63,03 | 49,6 |
| % Overig niet westers | 17,81 | 26,94 | 65,1 | 49,12 |
| % Eenpersoonshuishoudens | 58,92 | 77,45 | 70,32 | 71,99 |
| % Huishoudens zonder kinderen | 54,82 | 45,06 | 41,29 | 31,81 |
| % Huishoudens met kinderen | 38,99 | 66,06 | 28,37 | 63,35 |
| Gemiddelde huishoudensgrootte | 44,62 | 73,13 | 40,23 | 73,64 |
| % Meergezinswoning | 42,53 | 43,01 | 39,66 | 51,9 |
| % Onbewoond | 25,21 | 26,11 | 19,15 | 13,44 |
| % Huurwoningen totaal | 56,63 | 55,75 | 73,94 | 77,92 |
| % In woningcorporatie | 43,24 | 10,52 | 66,85 | 73,24 |
| % In bezit overige verhuurders | 32,14 | 42,66 | 13,9 | 34,79 |
| % Eigendom onbekend | 6,96 | 14,94 | 27,2 | 6,68 |
| % Bouwjaar vanaf 2000 | 33,97 | 13,87 | 34,58 | 39,6 |
| % Ervaren Gezondheid goed/zeer goed | 97,19 | 28,64 | 88,21 | 97,88 |
| % Wekelijkste sporters | 70,47 | 65,67 | 84,85 | 59,79 |
| % Ondergewicht | 56,96 | 73,22 | 28,71 | 35,87 |
| % Overgewicht | 72,16 | 69,07 | 63,91 | 63,68 |
| % Ernstig overgewicht | 79,94 | 55,2 | 81,13 | 72,46 |
| % Roker | 53,59 | 69,26 | 64,11 | 67,61 |
| % Drinker | 83,19 | 65,13 | 93,06 | 78,85 |
| % Zware drinker | 58,95 | 77,95 | 49,07 | 42,35 |
| % Een of meer lichamelijke beperkingen | 100 | 36,09 | 91,3 | 96,17 |
| % Hoog risico op angst of depressie | 71,9 | 66,55 | 86,43 | 74,06 |
| % Heel veel stress in afgelopen 4 weken | 66,05 | 74,38 | 58,68 | 56,27 |
| % Vrijwilligerswerk | 30,9 | 37,34 | 75,04 | 63,54 |
| % Moeite met rondkomen | 68,49 | 64 | 81,22 | 79,55 |

We hebben het random forest classificatiealgoritme gebruikt als machine learning model om buurten in te delen in clusters op basis van verschillende variabelen. Het random forest algoritme werkt door het creëren van een verzameling van beslissingsbomen. Elke boom wordt getraind op een subset van de gegevens en geeft een voorspelling af. Het uiteindelijke resultaat is de gemiddelde voorspelling van alle bomen in de verzameling. Eerst wordt er een training set (80% van de data) en een test set (20% van de data) gecreëerd op basis van de cluster variabele van de dataset. Dit gebeurt door middel van de createDataPartition functie uit de caret package, die de indices van de trainings- en testdata splitst op basis van de cluster variabele. Hierbij wordt 80% van de data gebruikt voor training en 20% voor testing. Er wordt een random forest model gecreëerd met de trainingsdata. Dit gebeurt door middel van de caret package en de train functie, waarbij "rf" wordt gebruikt als de methode voor het trainen van het model. Het model werd beoordeeld door middel van cross-validatie, waarbij de gegevens in 10 delen werden verdeeld en het model 10 keer opnieuw werd getraind en gevalideerd, waarbij elk deel eenmaal werd uitgesloten als test set. Dit proces werd 5 keer herhaald om een robuuste schatting te maken van de prestaties van het model. Het model voorspelt de cluster variabele op basis van alle andere variabelen in de dataset. De set.seed functie wordt gebruikt om de resultaten reproduceerbaar te maken en de importance = TRUE parameter zorgt ervoor dat de variabelen die het meest bijdragen aan de voorspelling van het cluster worden gerapporteerd.

De resultaten tonen dat het model een hoge nauwkeurigheid heeft, aangezien de gemiddelde nauwkeurigheid over alle modellen en iteraties 0.958 is, met een Kappa-score van 0.938. Dit betekent dat het model bijna 96% van de tijd correct voorspelt tot welk cluster een buurt behoort.

Daarnaast zijn de resultaten van het afstemmen van de mtry-parameter weergegeven. Het model presteerde het beste met mtry = 2, wat betekent dat bij het bouwen van elke boom in het bos slechts twee variabelen willekeurig werden geselecteerd om de boom te bouwen. Dit toont aan dat het model niet alle variabelen nodig heeft om zeer nauwkeurige voorspellingen te maken.

Om te begrijpen welke variabelen belangrijk zijn geweest voor de classificatie, hebben we de feature importances van het random forest algoritme gebruikt. De variabele importances worden berekend met behulp van de functie varImp() uit de caret-package. Hierbij wordt de scale parameter op TRUE gezet om de schaal van de variabelen gelijk te maken en de vergelijking van de belangrijkheid van de variabelen te vergemakkelijken. Dit stelt ons in staat om te zien welke variabelen een grote invloed hebben gehad op de indeling van de buurten in clusters.

## APPENDIX D: CLASSIFICATIE VAN DE BUURTEN OP DE CLUSTER INDELING MET MACHINE LEARNING EXPLAINER

LIME (Local Interpretable Model-Agnostic Explanations) is een techniek voor het uitleggen van de uitkomst van een machine learning-model door te kijken naar welke kenmerken van de invoer de grootste invloed hebben op de voorspelling die het model doet. Het idee is om een lokaal interpreteerbaar model te maken rondom een specifieke voorspelling, waarbij het model alleen kijkt naar de kenmerken van de invoer die relevant zijn voor die voorspelling. Dit lokale model kan vervolgens worden gebruikt om uit te leggen waarom het model tot die voorspelling is gekomen.

In ons geval gebruiken we LIME om uit te leggen waarom een bepaalde buurt is ingedeeld in een specifiek cluster. We kijken naar welke kenmerken van die buurt de grootste invloed hebben op de clustering. Door deze uitleg kunnen we beter begrijpen waarom buurten in specifieke clusters zijn ingedeeld en kunnen we dit per buurt analyseren in plaats van alleen cluster breed.

Het belangrijkste voordeel van LIME is dat het een model-agnostische methode is. Dit betekent dat het kan worden toegepast op elk type machine learning model, ongeacht de complexiteit ervan. Bovendien is LIME ontworpen om te werken met zowel gestructureerde als ongestructureerde gegevens, wat het een zeer flexibele methode maakt voor model interpretatie.

De functie “explain” wordt gebruikt om de voorspelling van het model te verklaren.

In Figuur 2 staan twee voorbeelden van een uitleg. Het model heeft de buurten ingedeeld in respectievelijk Cluster (=Label) 3 en Cluster 4 met een waarschijnlijkheid van 80% en 86%. De grafiek toont welke variabelen hebben bijgedragen aan deze indeling. In de Abdij- en Torenbuurt is het percentage alcoholdrinkers kleiner of gelijk aan 77,3%, en het percentage inwoners met een hoog risico op angst of depressie is hoger dan 8,5%. Deze twee factoren hebben een relatief grote invloed gehad op de voorspelling van de buurt in Cluster 3. Aan de andere kant heeft de Armhoef Noord-West buurt een hoger percentage drinkers dan 84,4%, en meer dan 81,9% van de inwoners beschouwen hun gezondheid als goed tot zeer goed. Deze factoren hebben een relatief grote invloed gehad op de indeling van deze buurt in Cluster 4.

## APPENDIX E: RESULTATEN ANALYSE WONINGBRANDEN PER REGIO EN PER CLUSTER

### *RESULTATEN LINEAIRE REGRESSIE*

**NOTE: HIERONDER LICHTEN WE DE RESULTATEN VAN EEN LINEAIRE REGRESSIE TOE. DAARNAAST ZIJN OOK EEN POISSON-REGRESSIE EN EEN ROBUUSTE LINEAIRE REGRESSIE ONDERZOCHT, DIE MEER WEERSTAND BIEDT TEGEN UITSCHIETERS. ER WERD ECHTER SLECHTS EEN MINIMAAL VERSCHIL WAARGENOMEN TUSSEN DE VERSCHILLENDE REGRESSIEMODELLEN. VANWEGE DE BETERE INTERPRETATIE VAN EEN LINEAIRE RELATIE, IS ERVOOR GEKOZEN OM ALLEEN DE LINEAIRE REGRESSIE IN DIT RAPPORT TOE TE LICHTEN. NIETTEMIN ZIJN IN HET SCRIPT DE CODES OPGENOMEN OM ALLE REGRESSIEMODELLEN UIT TE VOEREN.**

*lm(formula = aantal meldingen ~ aantal huishoudens, data = data_incidenten)*

| | Estimate | Std. Error | t-value | p-value |
|---|---|---|---|---|
| (Intercept) | 0.864795 | 0.209972 | 4.119 | 4.34e-05 |
| Aantal huishoudens | 0.006943 | 0.000189 | 36.729 | < 2e-16 |

Residual standard error: 3.585 on 607 degrees of freedom

Multiple R-squared: 0.6897, Adjusted R-squared: 0.6892

F-statistic: 1349 on 1 and 607 DF, p-value: < 2.2e-16

De lineaire regressieanalyse die is uitgevoerd, heeft een model opgeleverd om de relatie tussen het aantal huishoudens en het aantal woningbranden te modelleren voor de hele regio, dus niet per cluster. De uitkomst van deze analyse kan worden samengevat in de volgende punten:

De analyse toont aan dat het intercept van het model 0.865 is, wat betekent dat het verwachte aantal woningbranden gelijk is aan 0.865 wanneer er 0 huishoudens zijn. Dit heeft echter geen praktische betekenis, aangezien er in werkelijkheid altijd huishoudens zijn in een buurt. Verder blijkt uit de analyse dat de helling van het model 0.007 is. Dit betekent dat voor elke toename van één huishouden, het aantal woningbranden met 0.007 zal toenemen.

Het R-kwadraat van het model is 0.6897. Dit betekent dat ongeveer 68,97% van de variatie in het aantal woningbranden wordt verklaard door de variatie in het aantal huishoudens. Dit suggereert dat het model een redelijke mate van nauwkeurigheid heeft bij het voorspellen van het aantal woningbranden op basis van het aantal huishoudens.

Tot slot is de p-waarde voor de helling kleiner dan 0.001, wat betekent dat er een significante lineaire relatie is tussen het aantal huishoudens en het aantal woningbranden.

*RESULTATEN LINEAIRE REGRESSIE PER CLUSTER*

*lm(formula = aantal meldingen ~ aantal huishoudens + cluster, data = data_incidenten)*

| | Estimate | Std. Error | t-value | p-value |
|---|---|---|---|---|
| (Intercept) | 0.0917138 | 0.3060045 | 0.300 | 0.7645 |
| Aantal huishoudens | 0.0069693 | 0.0001911 | 36.478 | < 2e-16 *** |
| cluster2 | 2.5514495 | 0.4489047 | 5.684 | 2.05e-08 *** |
| cluster3 | 1.1265319 | 0.4970584 | 2.266 | 0.0238 * |
| cluster4 | 0.6634999 | 0.3374136 | 1.966 | 0.0497 * |

Residual standard error: 3.499 on 604 degrees of freedom

Multiple R-squared: 0.7059, Adjusted R-squared: 0.7039

F-statistic: 362.4 on 4 and 604 DF, p-value: < 2.2e-16

De resultaten van het model laten zien dat de intercept, oftewel het aantal woningbranden in een buurt zonder huishoudens, 0.092 is. Dit heeft echter geen praktische betekenis, aangezien er in werkelijkheid altijd huishoudens zijn in een buurt. De helling van 0.007 geeft aan dat voor elke toename van één huishouden in een buurt, het aantal woningbranden gemiddeld met ongeveer 0.007 zal toenemen, op basis van de andere variabelen in het model.

De hellingen van de clusters geven de relatie weer tussen het aantal woningbranden en het cluster waarin de buurt is ingedeeld, waarbij rekening is gehouden met het aantal huishoudens. In dit specifieke voorbeeld zijn de hellingen van de clusters als volgt: Cluster 2: 2.551, Cluster 3: 1.127, en Cluster 4: 0.663. Dit betekent dat als we de invloed van het aantal huishoudens uitsluiten, de verwachte toename van het aantal woningbranden in Cluster 2 ten opzichte van Cluster 1 2.551 is. Voor Cluster 3 is deze toename 1.127 en voor Cluster 4 0.663. De p-waarde geeft aan dat de schattingen van deze hellingen statistisch significant zijn, met een p-waarde kleiner dan 0.05. Dit betekent dat we kunnen concluderen dat er een significant verschil is in het aantal woningbranden tussen de verschillende clusters van buurten.

Uit de resultaten blijkt dat er significant meer woningbranden zijn in Clusters 2, 3 en 4, waarbij Cluster 2 de hoogste coëfficiënt heeft. Het model lijkt een goede pasvorm te hebben, aangezien de R-kwadraat (R-squared) van 0.706 aangeeft dat ongeveer 70,6% van de variantie in het aantal woningbranden kan worden verklaard door de variabelen in het model. De p-waarde van het F-statistiek is laag (minder dan 2.2e-16), wat betekent dat het model significant is.

*RESULTATEN LINEAIRE REGRESSIE MET INTERACTIE EFFECT*

*lm(formula = aantal meldingen ~ aantal huishoudens * cluster, data = data_incidenten)*

| | Estimate | Std. Error | t value | Pr(>\|t\|) |
|---|---|---|---|---|
| (Intercept) | 0.4941217 | 0.3483410 | 1.419 | 0.1566 |
| Aantal huishoudens | 0.0065588 | 0.0002637 | 24.869 | < 2e-16 *** |
| cluster2 | 0.1738269 | 0.7047914 | 0.247 | 0.8053 |
| cluster3 | -1.3978385 | 0.7446063 | -1.877 | 0.0610 . |
| cluster4 | 0.9708751 | 0.4570798 | 2.124 | 0.0341 * |
| Aantal huishoudens:cluster2 | 0.0027378 | 0.0006378 | 4.292 | 2.06e-05 *** |
| Aantal huishoudens:cluster3 | 0.0024223 | 0.0005489 | 4.413 | 1.21e-05 *** |
| Aantal huishoudens:cluster4 | -0.0008337 | 0.0004451 | -1.873 | 0.0615 . |

Residual standard error: 3.375 on 601 degrees of freedom

Multiple R-squared: 0.7276, Adjusted R-squared: 0.7245

F-statistic: 229.4 on 7 and 601 DF,  p-value: < 2.2e-16

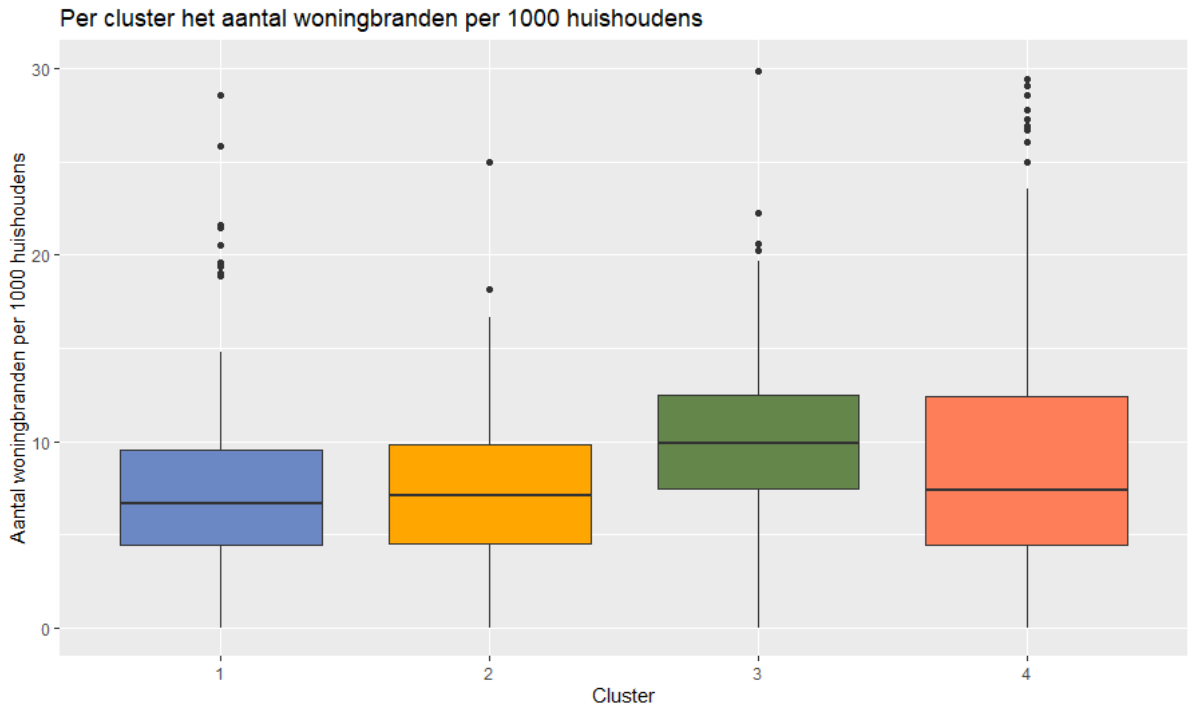
Het lijkt erop dat het effect van het aantal huishoudens verschilt tussen de verschillende clusters. Dit wordt aangegeven door de interactie-effecten (Aantal meldingen ~ Aantal huishoudens:clusterX). De coëfficiënten van deze interactie-effecten zijn significant en positief voor Cluster 2 en Cluster 3, wat betekent dat het effect van aantal huishoudens op het aantal incidentmeldingen groter is in die clusters dan in Cluster 1 (het referentiecluster). Voor Cluster 4 is het interactie-effect negatief en niet significant.

Het model heeft een goede pasvorm omdat het aangepaste R-kwadraat 0,7245 is. Dit betekent dat het model ongeveer 72% van de variantie in het aantal incidentmeldingen kan verklaren. Het toevoegen van een interactie effect voegt maar 2% extra verklaarde variantie toe in vergelijking met het model zonder interactie effect.

### *Het aantal woningbranden per 1000 huishoudens*

In Figuur 18 zie je vier boxplots, elk voor een verschillend cluster. De boxplot toont informatie over het aantal woningbranden per 1000 huishoudens in elk cluster. De gekleurde 'box' van de boxplot laat zien waar de meeste data zich bevindt, terwijl de middelste lijn het midden van de data laat zien. De lijntjes die uit de boxplot steken, worden whiskers genoemd. Zij laten zien hoe ver de data zich uitstrekt.

Uit de boxplot is duidelijk af te lezen dat de verschillende clusters ongelijk zijn verdeeld qua aantal woningbranden per 1000 huishoudens. Cluster 3 valt op omdat hier relatief meer woningbranden plaatsvinden per 1000 huishoudens in vergelijking met de andere clusters. We kijken voornamelijk naar de mediaan, en niet naar het gemiddelde als centrum maat omdat het aantal woningbranden per buurt niet normaal verdeeld is maar erg negatief 'geskewed' is. Verder valt op dat in Cluster 4 de spreiding van het aantal woningbranden wat groter is. Dit kan verklaard worden door het feit dat dit cluster voornamelijk bestaat uit buitengebieden met weinig inwoners. Hierdoor kan zelfs een laag aantal woningbranden per buurt al snel leiden tot een hoge ratio van woningbranden per 1000 inwoners.




**FIGUUR 18.**

We hebben daarnaast het aantal woningbranden per 1000 huishoudens (= "ratio_incidenten") statistisch getoetst met een ANOVA toets om te kijken of er een significant verschil zit tussen de verschillende clusters.

De boxplot (Figuur 18) laat dus voornamelijk de mediaan zien, terwijl de ANOVA gemiddeldes vergelijkt.

*aov(ratio_incidenten ~ cluster, data = data_incidenten)*

| | Degrees of freedom | Sum of squares | Mean squares | F-value | Pr(>\|t\|) |
|---|---|---|---|---|---|
| Cluster | 3 | 1677 | 559 | 5.838 | 0.000619 *** |
| Residuals | 605 | 57931 | | | |

In dit geval laat de ANOVA-tabel zien dat de verklaarde variantie door de "cluster" factor significant is (F-waarde = 5.838, p-waarde = 0.000619). Dit betekent dat er een significant verschil is tussen de clusters met betrekking tot de gemiddelde waarden van "ratio_incidenten" (het aantal woningbranden per 1000 huishoudens).

## APPENDIX F: BELANGRIJKSTE INDICATOREN OM WONINGBRANDEN TE VOORSPELLEN

### *LASSO/ELASTIC-NET/RIDGE REGRESSIE*

In situaties waarin er een grote hoeveelheid onderling gecorreleerde variabelen aanwezig is, kunnen traditionele lineaire regressieanalyses problemen ondervinden bij het identificeren van de belangrijkste indicatoren voor het model. Om dit probleem aan te pakken, maakt lasso-regressie gebruik van een "penalty term" die aan de kostfunctie wordt toegevoegd. Deze term verkleint bepaalde gewichten van variabelen in het model of brengt ze zelfs tot nul. Hierdoor kunnen we een meer beknopt model verkrijgen met minder variabelen, waarbij de coëfficiënten van irrelevante variabelen tot nul worden gereduceerd. Dit stelt ons in staat om de belangrijkste indicatoren te identificeren en minder belangrijke variabelen uit te sluiten.

Naast lasso-regressie zijn er ook andere benaderingen onderzocht, zoals ridge-regressie en elastic net-regressie. Bij ridge-regressie worden de overige coëfficiënten sterk verkleind, maar niet tot nul gebracht. Dit is vooral geschikt wanneer alle variabelen belangrijk zijn en er geen weggelaten mogen worden. Elastic net-regressie is eigenlijk een combinatie van lasso- en ridge-regressie. Bij het vergelijken van de prestaties van de regressiemodellen tussen elastic-net en lasso (gemeten met RSME en MAE) was het verschil zo klein dat er gekozen is voor lasso-regressie omdat de interpreteerbaarheid van dit model beter is. Het behouden van minder indicatoren in het model maakt de interpretatie ervan gemakkelijker.

In het R-script wordt een trainControl-object gedefinieerd dat de parameters bevat voor het trainen van het model, in ons geval een 10-voudige cross-validatie. Vervolgens wordt een lassoGrid gedefinieerd, dat de waarden van alpha en lambda bevat die zullen worden gebruikt in de lasso regressie. In dit geval is alpha gelijk aan 1, wat betekent dat er alleen lasso-regressie wordt gebruikt en geen elastic net-regressie. Lasso-regressie maakt gebruik van L1-norm regularisatie, wat betekent dat het probeert om de coëfficiënten van de onafhankelijke variabelen zo veel mogelijk naar nul te brengen, terwijl het de fout in de afhankelijke variabele zo laag mogelijk houdt. Dit resulteert in schattingen van de coëfficiënten die "schaars" zijn, wat betekent dat veel coëfficiënten precies nul zullen zijn. Het model wordt hierdoor beter interpreteerbaar.

Er wordt een 'loop' uitgevoerd die de lasso-regressie voor elk cluster afzonderlijk uitvoert. Binnen de 'loop' wordt eerst een set voorspellers en de respons variabele gedefinieerd op basis van de gegevens van het huidige cluster. Vervolgens wordt de lasso-regressie getraind met behulp van de train functie van de caret package, waarbij de voorspellers, responsvariabele en andere parameters zoals de trainControl en tuneGrid worden meegegeven. Ten slotte worden de coëfficiënten van de lasso-regressie voor elk cluster opgeslagen.

**TABEL 3.** Lasso coëfficiënten per cluster – zonder het aantal huishoudens als predictor.

| attribute | cluster1 | cluster2 | cluster3 | cluster4 |
|---|---|---|---|---|
| (Intercept) | 6,99 | 8,62 | 8,6 | 4,83 |
| % 16 tot 25 jaar | 0,01 | | | |
| % 25 tot 45 jaar | | | 0,38 | |
| % Gescheiden | | 0,24 | | |
| % Verweduwd | | | | 0,58 |

| | | | | |
|---|---|---|---|---|
| % Westers totaal | | -0,22 | | |
| % Marokko | -0,23 | 2 | 1,16 | |
| % Nederlandse Antillen en Aruba | -0,01 | -0,11 | | |
| % Suriname | -0,3 | | | |
| % Turkije | -0,09 | | | |
| % Overig niet westers | | | | 0,05 |
| % Eenpersoonshuishoudens | 0,01 | | | 0,02 |
| % Huishoudens zonder kinderen | -0,71 | | -0,33 | |
| % Eengezinswoning | -0,86 | | | |
| % Bewoond | | | -0,68 | |
| % Koopwoningen totaal | | | | -0,72 |
| % In woningcorporatie | 0,03 | | | 0,31 |
| % In bezit overige verhuurders | -0,21 | | | |
| % Eigendom onbekend | -0,29 | 0,02 | | |
| % Bouwjaar voor 2000 | 0,5 | | | 0,27 |
| % Ervaren Gezondheid goed/zeer goed | | | | -0,2 |
| % Voldoet aan beweging Richtlijn | 0,1 | | | |
| % Ondergewicht | -0,62 | | | |
| % Normaal gewicht | 0,38 | | | |
| % Overgewicht | | | | -0,49 |
| % Ernstig overgewicht | -0,63 | | | -0,22 |
| % Voldoet aan alcohol richtlijn | | | | -0,09 |
| % Zware drinker | -0,14 | | | |
| % Overmatige drinker | -0,65 | | | 0,17 |
| % Langdurige ziekte en beperkt | -0,65 | | | |
| % Beperking in horen | 0,83 | | | |
| % Beperking in bewegen | | | | 0,34 |
| % Matig veel regie over eigen leven | | | 0,49 | -0,45 |
| % Ernstig/zeer ernstig eenzaam | 0,55 | | | |
| % Mantelzorger | -0,07 | | | |
| % Vrijwilligerswerk | | -0,17 | | |

**TABEL 4**. Lasso coëfficiënten per cluster – met het aantal huishoudens en inwoners als predictor.

| attribute | cluster1 | cluster2 | cluster3 | cluster4 |
|---|---|---|---|---|
| (Intercept) | 6,99 | 8,62 | 8,6 | 4,83 |
| Aantal inwoners | | | 4,09 | |
| % 16 tot 25 jaar | 0,29 | | | |
| % 45 tot 65 jaar | -0,01 | | | |
| % Gescheiden | | 0,01 | | |
| % Suriname | | 0,04 | | |
| Sterfte totaal | 0,29 | | | 0,4 |
| Huishoudens totaal | 5,49 | 7,03 | 1,18 | 3,04 |
| % Eenpersoonshuishoudens | 0,87 | | | |
| % Huishoudens zonder kinderen | | | -0,53 | |

| | | | | |
|---|---|---|---|---|
| % Bewoond | | | | -0,61 |
| % Koopwoningen totaal | 0,31 | | | |
| % In woningcorporatie | -0,05 | | | |
| % Normaal gewicht | 0,25 | | | |
| % Roker | 0,39 | | | |
| % Drinker | -0,38 | | | |
| % Zware drinker | | | | 0,13 |
| % Overmatige drinker | | | | 0,21 |
| % Beperking in zien | | 0,27 | | |
| % Heel veel stress in afgelopen 4 weken | | | | -0,42 |
| % Matig veel regie over eigen leven | | | | -0,03 |
| % Vrijwilligerswerk | 0,26 | -0,05 | | 0,13 |

In Tabel 3 en 4 wordt voor elke cluster en elk attribuut weergegeven wat de coëfficiënt is van de variabele in de lasso regressie van dat cluster. Een positief cijfer geeft aan dat er een positieve relatie is tussen het attribuut en het aantal meldingen woningbrand in dat cluster. Met andere woorden, hoe hoger het cijfer, hoe groter de invloed van dat attribuut op de aantal woningbranden van dat cluster.

Aan de andere kant geeft een negatief cijfer aan dat er een negatieve relatie is tussen het attribuut en het aantal meldingen in dat cluster. Dit betekent dat hoe lager het cijfer, hoe groter de invloed van dat attribuut op het aantal woningbranden van dat cluster, maar dan in negatieve zin.

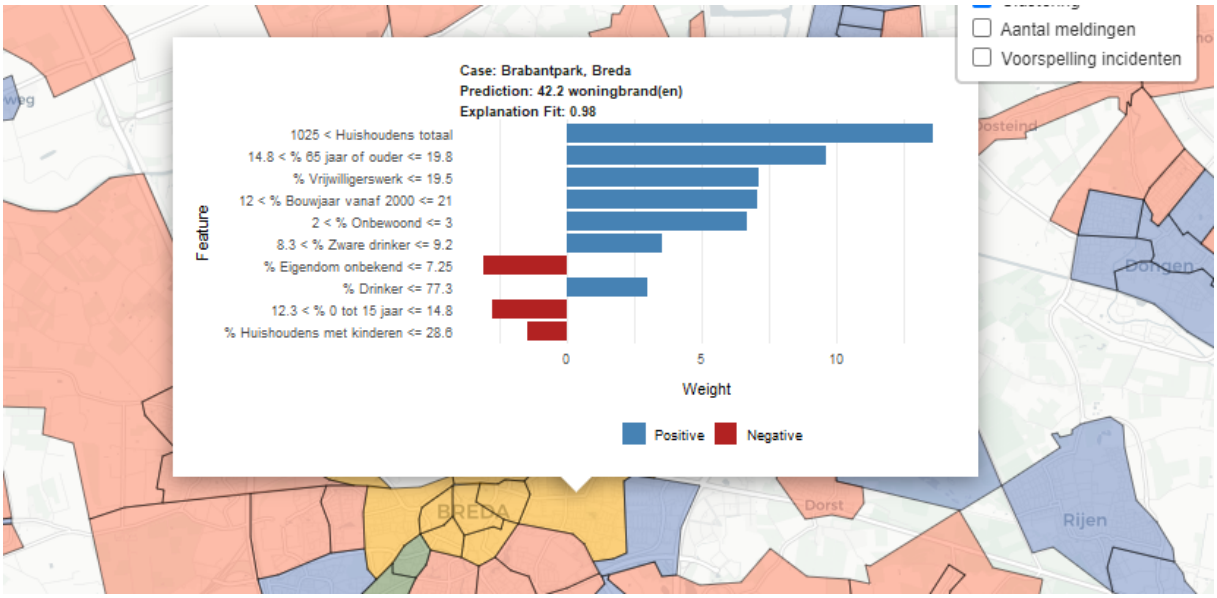
Wanneer de coëfficiënt van een attribuut gelijk is aan nul, betekent dit dat de variabele geen bijdrage levert aan het model. Met andere woorden, het attribuut is niet veelbetekenend voor het verklaren van de variatie in de aantal meldingen woningbrand van dat specifieke cluster. In de tabel worden deze waarden als "leeg" weergegeven.

### *Aantal meldingen voorspellen per buurt*

Naast het analyseren van de clusters hebben we ook individuele voorspellingen per buurt gedaan en gekeken naar de belangrijkste variabelen die daarbij een rol spelen. We kunnen elk willekeurig model trainen om het aantal woningbranden in een buurt te voorspellen en de belangrijkste variabelen per buurt identificeren die bijdragen aan deze voorspelling. Bovendien kunnen we LIME-resultaten gebruiken om de redenen achter een specifieke voorspelling in een buurt mogelijk te verklaren.

Het Random Forest-model wordt getraind met behulp van de train-functie uit het caret package. Hier worden verschillende parameters ingesteld, zoals de methode ("rf"), de belangrijkheid van variabelen, en trainingscontroleparameters voor cross-validatie. Het getrainde model wordt gebruikt om voorspellingen te doen voor de testset (test). De beste resultaten werden behaald bij een 'mtry'-waarde van 42. Het gekozen model had een gemiddelde RMSE van 3.322724, een gemiddelde Rsquared van 0.6861333, en een gemiddelde MAE van 2.487866. Bij het selecteren van het optimale model werd de RMSE gebruikt en het model met de kleinste waarde van RMSE werd gekozen. Het uiteindelijk gekozen model had een 'mtry'-waarde van 42.

Een 'lime'-explainer wordt gecreëerd met behulp van de trainingsgegevens (X_train) en het getrainde Random Forest-model (model_rf). Figuur 19 laat een voorbeeld van een voorspelling van het aantal woningbranden in een buurt inclusief de uitleg.




**FIGUUR 19**

## APPENDIX G: VERSCHILLEN IN OORZAKEN/ONTSTAANSLOCATIES/LETSEL/SCHADE PER CLUSTER

Naast het analyseren van de visualisaties is het ook van belang om te onderzoeken of de verschillen in oorzaken, ontstaanslocaties, type woning en letsel significant verschillen tussen de clusters. Hiervoor maken we gebruik van de Fisher's Exact-test. Deze test is geschikt in situaties waarin de aantallen incidenten per categorie klein zijn. Door de Fisher's Exact-test uit te voeren voor elke cluster, kunnen we bepalen of er een significant verschil is in de aantallen tussen de clusters, rekening houdend met de totale woningbranden per cluster. Het is belangrijk op te merken dat er in Cluster 1 en 4 relatief meer buurten zijn opgenomen volgens het clusteralgoritme, waardoor er automatisch meer branden worden waargenomen in die clusters. De getoonde aantallen branden zijn cumulatief opgeteld. In onderstaande tabellen, geeft het label 'Totaal aantal branden (Salvage)' aan hoeveel branden er zijn geweest waarbij Salvage betrokken was in dat specifieke cluster.

Hieronder staan per onderwerp een aantal resultaten uitgelicht:

### *OORZAKEN*

**Oorzaak: Kortsluiting**

| Cluster | Totaal aantal branden (Salvage) | Aantal oorzaak kortsluiting | Proportie |
|---|---|---|---|
| 1 | 476 | 39 | 8.1% |
| 2 | 213 | 8 | 3.8% |
| 3 | 336 | 19 | 5.7% |
| 4 | 228 | 22 | 9.6% |

*Fisher's Exact Test for Count Data*
*p-value = 0.0453*
*alternative hypothesis: two.sided*

**Oorzaak: Haarden/kachels**

| Cluster | Totaal aantal branden (Salvage) | Aantal oorzaak haarden/kachels | Proportie |
|---|---|---|---|
| 1 | 476 | 23 | 4.8% |
| 2 | 213 | 7 | 3.3% |
| 3 | 336 | 9 | 2.7% |
| 4 | 228 | 34 | 14.9% |

*Fisher's Exact Test for Count Data*
*p-value = 6.801e-08*
*alternative hypothesis: two.sided*

**Oorzaak: Menselijk handelen**

| Cluster | Totaal aantal branden (Salvage) | Aantal oorzaak menselijk handelen | Proportie |
|---|---|---|---|
| 1 | 476 | 123 | 29.2% |
| 2 | 213 | 58 | 27.2% |
| 3 | 336 | 131 | 39% |
| 4 | 228 | 29 | 12.7% |

*Fisher's Exact Test for Count Data*
*p-value = 9.799e-11*
*alternative hypothesis: two.sided*

**Oorzaak: Vermoeden van brandstichting**

| Cluster | Totaal aantal branden (Salvage) | Aantal oorzaak vermoeden van brandstichting | Proportie |
|---|---|---|---|
| 1 | 476 | 11 | 2.3% |
| 2 | 213 | 12 | 5.7% |
| 3 | 336 | 10 | 3% |
| 4 | 228 | 2 | 0.9% |

*Fisher's Exact Test for Count Data*
*p-value = 0.02426*
*alternative hypothesis: two.sided*

**Oorzaak: Stoken**

| Cluster | Totaal aantal branden (Salvage) | Aantal oorzaak stoken | Proportie |
|---|---|---|---|
| 1 | 476 | 22 | 4.6% |
| 2 | 213 | 8 | 3.8% |
| 3 | 336 | 5 | 1.5% |
| 4 | 228 | 17 | 7.5% |

*Fisher's Exact Test for Count Data*
*p-value = 0.0042*

*alternative hypothesis: two.sided*

*ONTSTAANSLOCATIES*

**Ontstaanslocatie: Keuken, kantine**

| Cluster | Totaal aantal branden (Salvage) | Aantal ontstaanslocatie keuken/kantine | Proportie |
|---|---|---|---|
| 1 | 476 | 123 | 25.8% |
| 2 | 213 | 64 | 30% |
| 3 | 336 | 133 | 39.6% |
| 4 | 228 | 36 | 15.8% |

*Fisher's Exact Test for Count Data*
*p-value = 6.377e-09*
*alternative hypothesis: two.sided*

*LETSEL*

**Letsel: Geen letsel**

| Cluster | Totaal aantal branden (Salvage) | Aantal geen letsel | Proportie |
|---|---|---|---|
| 1 | 476 | 370 | 77.7% |
| 2 | 213 | 157 | 73.7% |
| 3 | 336 | 259 | 77.1% |
| 4 | 228 | 199 | 87.3% |

*Fisher's Exact Test for Count Data*
*p-value = 0.0019*
*alternative hypothesis: two.sided*

**Letsel: Ja, ziekenhuis**

| Cluster | Totaal aantal branden (Salvage) | Aantal letsel ziekenhuis | Proportie |
|---|---|---|---|
| 1 | 476 | 34 | 7.1% |
| 2 | 213 | 17 | 8% |
| 3 | 336 | 13 | 3.9% |
| 4 | 228 | 6 | 2.6% |

*Fisher's Exact Test for Count Data*
*p-value = 0.0147*
*alternative hypothesis: two.sided*

*BESTEMMING PAND*

**Bestemming pand: Flat/appartement**

| Cluster | Totaal aantal branden (Salvage) | Aantal bestemming pand flat/appartement | Proportie |
|---|---|---|---|
| 1 | 476 | 34 | 9.9% |

| | | | |
|---|---|---|---|
| 2 | 213 | 17 | 34.7% |
| 3 | 336 | 13 | 30.4% |
| 4 | 228 | 6 | 4.4% |

*Fisher's Exact Test for Count Data*
*p-value < 2.2e-16*
*alternative hypothesis: two.sided*

**Bestemming pand: Woonhuis**

| Cluster | Totaal aantal branden (Salvage) | Aantal bestemming pand woonhuis | Proportie |
|---|---|---|---|
| 1 | 476 | 187 | 39.3% |
| 2 | 213 | 59 | 27.7% |
| 3 | 336 | 138 | 41.1% |
| 4 | 228 | 99 | 43.4% |

*Fisher's Exact Test for Count Data*
*p-value = 0.00257*
*alternative hypothesis: two.sided*

**Bestemming pand: Kamerverhuur**

| Cluster | Totaal aantal branden (Salvage) | Aantal bestemming pand kamerverhuur | Proportie |
|---|---|---|---|
| 1 | 476 | 1 | 0.2% |
| 2 | 213 | 6 | 2.8% |
| 3 | 336 | 5 | 1.5% |
| 4 | 228 | 1 | 0.4% |

*Fisher's Exact Test for Count Data*
*p-value = 0.008434*
*alternative hypothesis: two.sided*

**Bestemming pand: Schuur (bij woonhuis)**

| Cluster | Totaal aantal branden (Salvage) | Aantal bestemming pand schuur (bij woonhuis) | Proportie |
|---|---|---|---|
| 1 | 476 | 23 | 4.8% |
| 2 | 213 | 2 | 0.9% |
| 3 | 336 | 6 | 1.8% |
| 4 | 228 | 10 | 4.4% |

*Fisher's Exact Test for Count Data*
*p-value = 0.01085*
*alternative hypothesis: two.sided*

**Bestemming pand: Vrijstaande woning**

| Cluster | Totaal aantal branden (Salvage) | Aantal bestemming pand vrijstaande woning | Proportie |
|---|---|---|---|
| 1 | 476 | 26 | 5.5% |
| 2 | 213 | 3 | 1.4% |
| 3 | 336 | 2 | 0.6% |
| 4 | 228 | 20 | 8.8% |

*Fisher's Exact Test for Count Data*
*p-value = 3.823e-07*
*alternative hypothesis: two.sided*

*HENNEP*

**Hennep volgens brandweer: Ja**

| Cluster | Totaal aantal branden (Salvage) | Aantal hennep volgens brandweer: Ja | Proportie |
|---|---|---|---|
| 1 | 476 | 6 | 1.3% |
| 2 | 213 | 8 | 3.8% |
| 3 | 336 | 3 | 0.9% |
| 4 | 228 | 1 | 0.4% |

*Fisher's Exact Test for Count Data*
*p-value = 0.03096*
*alternative hypothesis: two.sided*