Datamining in een GGZ instelling

Erwin Schuijtvlot

Stageverslag juni 2013

Vrije Universiteit
Faculteit der Exacte Wetenschappen
Studierichting Business Analytics
De Boelelaan 1081
1081 HV Amsterdam

Stagebedrijf:

Stichting Centrum '45
Nienoord 5
1112 XE Diemen

Voorwoord

Deze scriptie is geschreven in het kader van de afstudeerstage die de afronding vormt van de master studie Business Analytics. Dit verslag is het resultaat van 6 maanden onderzoek het landelijke behandel- en expertisecentrum voor psychotrauma, Stichting Centrum ’45.

Bij het uitvoeren van dit onderzoek heb ik naast technische kennis ervaring opgedaan in het voor mij onbekende domein van de psychotrauma. Ik wil hierbij de stichting bedanken voor het aanbieden van de mogelijkheid van deze stage.

Mijn dank gaat in het bijzonder uit naar Lindy van de Westelaken en Niels van de Aa voor de goede begeleiding en interesse die zij hebben getoond in mijn werk. Tevens wil ik alle medewerkers op de afdeling bedanken voor de fijne werksfeer die ik heb mogen ervaren tijdens mijn stage.

Ten slotte gaat mijn dank uit naar Ger Koole en Mark Hoogendoorn voor de begeleiding vanuit de VU.

Management samenvatting

Binnen gezondheidsinstellingen bestaat een toenemende behoefte aan efficiënte methoden om de groeiende hoeveelheid data te analyseren. Stichting Centrum ’45 is hierin geen uitzondering en binnen dit onderzoek zijn een aantal vraagstukken met betrekking tot data behandeld.

Bij dit onderzoek is gebruik gemaakt van drie datasets, de Compass data die resultaten bevat van vragenlijsten afgenomen bij cliënten, de User cliëntgegevens die per cliënt relevante factoren geeft en de User afspraakhistorie die per cliënt de volledige afspraakgegevens vanaf 2005 omvat.

Bij de analyses is gebruik gemaakt van zogeheten machine learning technieken die op intelligente wijze in staat zijn grote hoeveelheden data efficiënt te doorzoeken.

Met betrekking tot de Compass data is de volgende onderzoeksvraag geformuleerd:

-Is de Compass data geschikt als basis voor externe rapportages?

Uit deze hoofdvraag komen de volgende deelvragen:

-Hoe is het gesteld met de data integriteit van de beschikbare dataset?

Bij de integriteit analyse is gebleken dat de Compass data grotendeels correcte gegevens bevat. Er zijn echter statistisch significante verschillen in de verdeling van gegeven antwoorden tussen verschillende jaren gevonden. Dit kan betekenen dat de data behorend bij deze vragen niet correct is. Om dit vast te stellen is verder onderzoek van een domeinexpert nodig.

-Zijn er cliëntgroepen onder- of over-vertegenwoordigd in de data?

Bij de analyse van dit probleem is gekeken naar verschillen tussen cliënten waarvan wel en geen Compass meting bekend was. Hieruit bleek dat cliënten ouder dan 20 jaar die medicatie gebruiken goed binnen de set te zijn vertegenwoordigd. Cliënten die enkel een intakeprocedure hebben gehad bleken slecht in de data te zijn vertegenwoordigd. Cliënten die zijn aangemeld wegens relationele klachten bleken tevens goed te zijn vertegenwoordigd. Cliënten met onbekende aanmeldklacht bleken slecht te zijn vertegenwoordigd.

Om er voor te zorgen dat de Compass data wel representatief is voor de gehele populatie kunnen statistische technieken zoals stratified sampling worden gebruikt om de data filteren. Er kan ook worden gedacht aan beleidsmaatregelen zoals het verplichtstellen van Compassmetingen, het aanbod aan beschikbare talen van de Compass lijsten vergroten en het extra aansporen van ondervertegenwoordigde groepen.

Met betrekking tot de User data is de volgende hoofdvraag geformuleerd: -
Welke inefficiënte patronen zijn er in de data te vinden?

met de volgende sub vragen:

-Welke cliëntgroepen zijn er uit de data te halen?

Met behulp van een automatische procedure om data in groepen in te delen zijn 10 verschillende cliëntgroepen gevonden die voor de stichting herkenbare patronen vertonen.

-Welke groepen hebben een verhoogd risico op een no-show?

Bij het klinische no-show percentage bleek dat oudere vluchtelingen uit het Midden Oosten en onverzekerden therapietrouwer waren dan de overige klinisch behandelde groepen.

Bij het poliklinische no-show percentage bleek de naoorlogse generatie therapietrouw te zijn. Jonge vluchtelingen uit het Midden-Oosten en Afrika bleken relatief hogere percentages te hebben. Ook cliënten met gezinnen bleken hogere no-show percentages te hebben.

Deze informatie kan gebruikt worden om preventieve maatregelen te nemen zoals het versturen van sms alerts naar risicogroepen.

-Welke factoren zijn bepalend voor de behandelintensiteit?

De behandelintensiteit, geformuleerd als het gemiddeld aantal afspraken per 30 dagen bleek vooral afhankelijk te zijn van de behandelsetting. Klinisch en dagklinisch behandelde cliënten hebben gemiddeld meer afspraken dan poliklinisch behandelde cliënten. Het onderzoeken van de behandelintensiteit gebruikmakend van verschillende technieken leverde geen verdere interessante informatie op.

-Welke patronen met betrekking tot no-show en behandelfase zijn in de afspraakhistorie te vinden?

Het sterkste gevonden verband met betrekking tot no-shows bleek het verhoogde risico van het voorkomen van no-shows bij cliënten die reeds een no-show hebben vertoond.

Bij de verschilanalyse tussen behandelfasen was duidelijk te zien dat er in het eerste jaar aanzienlijk meer diagnostische activiteiten plaatsvonden en in latere jaren meer gerichte therapie zoals psychotherapie en medicatieconsulten.

Een opvallend verschijnsel was het voorkomen van intakegesprekken na het vijfde behandeljaar. Dit bleek voor te komen bij 20% van de cliënten die in hun vijfde behandeljaar of verder zaten. Dit is mogelijk te verklaren door een conversie fout. Doordat dit relatief veel cliënten zijn is het sterk aan te raden de fout verder te analyseren en eventueel op te lossen.

Een ander opvallend verschijnsel was de afwezigheid van verschil tussen behandeljaren twee tot en met vier tegenover het vijfde behandeljaar en verder. De verwachting was een toename in begeleiding en een afname in gerichte therapie. Dit was echter niet in de resultaten terug te vinden.

Als suggestie voor verder onderzoek zijn de volgende vragen geformuleerd:

-Wat zijn de verschillen tussen cliënten die bij Compass metingen verbetering laten zien ten opzichte van cliënten die verslechtering tonen?

-Welke factoren zijn bepalend voor de behandelduur?

-Zijn er relaties tussen tijd en no-show in de afspraakhistorie?

-Welke patronen zijn er te vinden in de indirecte uren?

Inhoud

Management samenvatting 1
1 Inleiding 9
1.1 Gerelateerd werk 9
2 Omgevingsanalyse 11
2.1 Organisatie 11
2.2 Cliëntendoorstroom 11
2.3 Gebruik van data 12
3 Probleemanalyse 14
3.1 Probleemstelling 14
3.2 Onderzoeksvragen 14
4 Datamining 15
4.1 Het CRISP model 15
4.2 Machine learning 16
4.2.1 Classification 16
4.2.1.1 Validatie 16
4.2.2 Clustering 17
4.2.2.1 Validatie 17
4.2.3 Sequence mining 17
4.2.3.1 Validatie 18
4.2.3.2 Verschil analyse 20
4.2.4 Gebruikte algoritmes 21
4.2.4.1 ZeroR Classificatie 21
4.2.4.2 OneR Classificatie 21
4.2.4.3 J48 Classificatie 22
4.2.4.4 Expectation Maximization clustering 24
4.2.4.5 Generalized Sequence Patterns 24
5 Exploratieve data analyse 25
5.1 Compass 25
5.1.1 Data Beschrijving 26
5.1.2 Integriteit 27
5.1.2.1 Dubbele records \ kolommen 27
5.1.2.2 Data types 27
5.1.2.3 Antwoordschalen 27
5.1.2.4 Cliëntgegevens 28

5.1.3 Exploratieve analyse 30
5.1.3.1 Ontvangstdatum 30
5.1.3.2 Cliënt nummer 32
5.1.3.3 Dataset opbouw 33
5.1.4 Conclusies 34
5.2 User cliënt data 35
5.2.1 Data Beschrijving 36
5.2.2 Data transformatie 36
5.2.3 Exploratieve analyse 38
5.2.3.1 Doelgroep 38
5.2.3.2 Leeftijd 42
5.2.3.3 Behandelduur 45
5.3 User Afspraakgeschiedenis 46
5.3.1 Data Beschrijving 46
5.3.2 Exploratieve analyse 46
6 Resultaten 48
6.1 Compass data cliënten 48
6.1.1 ZeroR 48
6.1.2 OneR 49
6.1.3 J48 49
6.2 Cliëntgroepen 54
6.2.1 Expectation Maximization clustering 54
6.2.1.1 Cluster 1: Jonge vluchtelingen uit het Midden Oosten 55
6.2.1.2 Cluster 2: Angstige Onbekende, Linners en Veteranen 55
6.2.1.3 Cluster 3 en 5: PDC en intake 55
6.2.1.4 Cluster 4: Oudere vluchtelingen uit het Midden Oosten 56
6.2.1.5 Cluster 6: Naoorlogse generatie met stemmingsklachten 56
6.2.1.6 Cluster 7 en 9: Jonge vluchtelingen uit het Midden Oosten en Afrika 56
6.2.1.7 Cluster 8: Vluchtelingen met gezinnen uit het Midden Oosten 58
6.2.1.8 Cluster 10: Onverzekerde vluchtelingen en onbekende uit een onbekende regio 58
6.2.1.9 Visueel overzicht 59
6.3 Behandelintensiteit 65
6.4 No-show percentages 66
6.4.1 Poliklinisch No-show percentage 66

6.4.2 Klinisch No-show percentage 68
6.5 Afspraakhistorie 69
6.5.1 Directe No-Show patronen 69
6.5.2 Indirecte no-show patronen 70
6.5.3 Behandelfase 71
7 Conclusie 74
7.1 Compass data 74
7.2 User data 76
7.2.1 Cliëntdata 76
7.2.2 Afspraakhistorie 77
7.3 Machine learning 78
7.4 Verder onderzoek 79
Referenties 80
Bijlage 1 Overzicht Compass lijsten 82

1 Inleiding

In dit onderzoek wordt onderzocht welke datamining technieken toegepast kunnen worden binnen een geestelijke gezondheidszorg (GGZ) instelling om diverse vraagstukken die binnen de organisatie leven te kunnen beantwoorden. Het onderzoek is uitgevoerd binnen Stichting Centrum ’45. Deze organisatie houdt zich bezig met het behandelen van cliënten met posttraumatisch stress syndroom (PTSS).

Op basis van een probleemstelling waar de stichting mee kampt, zijn een aantal onderzoeksvragen opgesteld. Om de onderzoeksvragen te beantwoorden wordt geëxperimenteerd met verschillende machine learning technieken. Deze technieken zijn in staat grote hoeveelheden data te doorzoeken op patronen die in een model worden weergegeven.

De modellen die deze technieken opleveren worden vervolgens geëvalueerd in samenwerking met een domeinexpert om zo verder inzicht te verkrijgen en de onderzoeksvragen te kunnen beantwoorden.

Deze scriptie is als volgt opgebouwd. In hoofdstuk 2 wordt een beschrijving van de organisatie gegeven. Er wordt behandeld hoe de stichting is opgebouwd, hoe het cliëntenstroom traject verloopt en hoe er wordt gewerkt met data.

Hoofdstuk 3 behandelt vervolgens de probleemstelling waar de stichting mee kampt en welke onderzoeksvragen hieruit voortvloeien. In hoofdstuk 4 wordt nader ingegaan op het datamining proces en welke technieken er in het onderzoek gebruikt worden en waarom er voor deze technieken is gekozen.

In het 5e hoofdstuk wordt een exploratieve analyse van de gebruikte data in het onderzoek uitgevoerd. In hoofdstuk 6 wordt voorts ingegaan op het beantwoorden van de onderzoeksvragen met behulp van de in hoofdstuk 4 genoemde technieken.

Afsluitend wordt in hoofdstuk 7 een conclusie met aanbevelingen aan de stichting gegeven. In dit laatste hoofdstuk worden ook een aantal suggesties voor vervolg onderzoeken genoemd.

1.1 Gerelateerd werk

Binnen de gezondheidszorg worden datamining toepassingen in toenemende mate populair en essentieel. Door de vaak grote omvang van de data zijn traditionele methoden niet altijd toereikend [1].

Binnen dit domein zijn verschillende toepassingen denkbaar. Een aantal voorbeelden zijn geven in [1] zijn:

- Inzicht verkrijgen in behandeleffectiviteit
- Tijdig risico’s detecteren
- Patiëntenstroom monitoren
- Fraude en misbruik detecteren

Bij het onderzoeken van de behandeleffectiviteit kunnen grote hoeveelheden data met ziekteverloop en toegepaste behandeling worden geanalyseerd. Voor bijvoorbeeld groepen patiënten met eenzelfde aandoening met verschillende behandelingen kan een met datamining verkregen model inzicht bieden in welke behandeling het effectiefst of kostenefficiëntst is.

Het detecteren van risico's kan worden gedaan door gegevens van nieuwe patiënten te toetsen aan een met datamining verkregen model om zo vroegtijdig hoge risico's aan het licht te brengen en preventief te handelen.

Door gegevens te analyseren met betrekking tot patiëntinteracties kunnen patronen aan het licht worden gebracht in het patiëntenstroom traject. Zo kan worden ingesprongen op verwachte voorkeuren en toekomstige activiteiten.

Ten slotte kunnen datamining technieken helpen bij het aan het licht brengen van misbruik en mogelijk frauduleuze handelingen door te zoeken naar ongebruikelijke patronen in bijvoorbeeld data met voorgeschreven medicijnen.

Er wordt vaak gewerkt met classificatie technieken waarbij een model wordt opgesteld om in de toekomst voorspellingen te kunnen doen. Een voorbeeld hiervan is te vinden in [2]. Hier zijn diverse classificatie technieken onderzocht om voorspellingen met betrekking tot diagnoses te kunnen doen.

Binnen dit onderzoek staat het bedrijfsmatige aspect van de instelling centraal. Het doel is om inzicht in de data te verwerven om zo in de toekomst bedrijfsprocessen te kunnen verbeteren.

2 Omgevingsanalyse

Dit hoofdstuk geeft een beschrijving van de organisatie en de huidige werkwijze met betrekking tot data.

2.1 Organisatie

Stichting centrum ’45 biedt landelijk specialistische zorg en diagnostiek voor cliënten die lijden aan posttraumatisch stress syndroom (PTSS). PTSS is de verzamelnaam voor psychotraumaklachten die zich voordoen als gevolg van vervolging, oorlog en geweld. De doelgroep van de stichting bestaat voornamelijk uit eerste en tweede generatie getroffenen in de Tweede Wereldoorlog, vluchtelingen en mensen die beroepsmatig zijn getraumatiseerd zoals politieagenten en militairen.

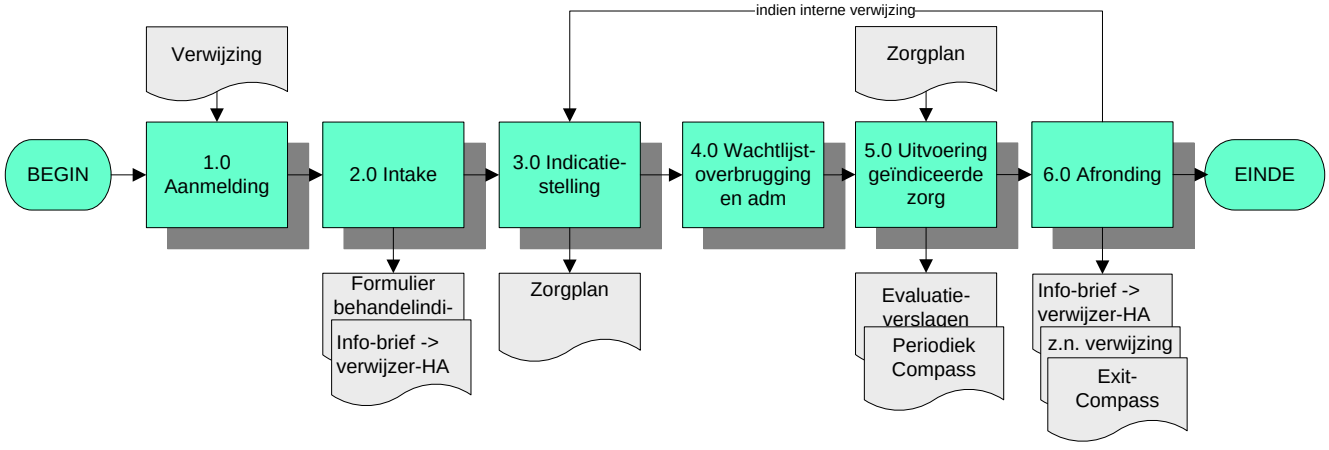
De stichting formuleert haar missie als: *Het bieden van landelijk specialistische psychotraumabehandeling en ontwikkeling van kennis en expertise op het gebied van de psychotraumatologie.*

De stichting maakt deel uit van de overkoepelende organisatie Arq psychotrauma expert groep.

De organisatie bestaat uit twee klinieken, locatie Oegstgeest en Diemen. In Diemen vinden dagbehandelingen en poliklinische behandelingen plaats. In Oegstgeest wordt naast deze behandelvormen klinische behandelingen aangeboden waarbij cliënten meerdere dagen bij de stichting verblijven.

2.2 Cliëntendoorstroom

De stichting heeft het globale behandeltraject van aanmelding tot uitschrijving ingedeeld in 6 stappen. Deze stappen zijn weergegeven in Figuur 1.



Figuur 1 Behandeltraject stichting centrum ‘45

In de eerste stap wordt beoordeeld of een cliënt voldoet aan de door de stichting gestelde aanmeldcriteria. Deze beoordeling wordt uitgevoerd door het bureau aanmelding in samenspraak met een psychiater. Op dit moment wordt er tevens een

zogeheten Diagnose Behandeling Combinatie (DBC) aangemaakt. In deze code staat voor een cliënt de diagnose, behandeling en het aantal behandelminuten. De DBC is sinds 2005 een landelijke standaard voor zorginstellingen [3]. Sinds deze landelijke invoering is de stichting verplicht gedetailleerde afspraakgegevens en diagnose bij te houden.

Wanneer een cliënt aan de criteria voldoet vindt er een intake plaats waarin een diagnose wordt gesteld. Bij deze stap wordt de cliënt tevens uitgenodigd om online een vragenlijst in te vullen. Deze vragenlijst wordt afgenomen in een voor de stichting op maat ontwikkelde webapplicatie, Compass. De cliënt heeft de mogelijkheid om de vragenlijst op de stichting of thuis in te vullen.

De cliënt staat vrij om op dit verzoek in te gaan. Doorgaans wordt een vragenlijst op de stichting ingevuld maar de mogelijkheid bestaat ook om de lijst vanuit huis in te vullen. Voor anderstaligen bestaan vertaalde versies van de lijsten.

Vervolgens wordt in de derde stap een globaal behandelplan opgesteld. In de vierde stap vinden eventuele overbruggingscontacten plaats wanneer er wachtlijsten bestaan voor de geplande zorg. In stap vijf wordt het behandelplan uitgevoerd. Binnen deze fase wordt de cliënt verzocht om periodiek elke 365 dagen een vragenlijst in te vullen in Compass.

De laatste stap vindt plaats wanneer een cliënt voldoende voortgang heeft gemaakt en er geen verdere behandeling binnen de stichting nodig is. Op dit moment wordt de cliënt wederom verzocht een vragenlijst in te vullen binnen Compass. Indien nodig wordt er een verwijzing gedaan voor verdere behandeling buiten de stichting.

2.3 Gebruik van data

Binnen de stichting wordt gewerkt met het software pakket User voor het bijhouden van een elektronisch patiënten dossier. Dit softwarepakket is binnen meerdere geestelijke gezondheidszorg instellingen in Nederland in gebruik.

De stichting deelt de database met twee andere stichtingen, Equator en het Psychotrauma Diagnose Centrum (PDC). Equator specialiseert zich in het behandelen van vluchtelingen en PDC voert diagnostiek uit op het gebied van psychotraumatische klachten naar aanleiding van geweld en geeft onafhankelijk behandeladvies.

In het dossier zijn de gehele afspraakhistorie, cliëntgegevens en voortgang terug te vinden. De data word opgeslagen in een Oracle database. Momenteel wordt deze data binnen de stichting gebruikt voor interne rapportages. Deze worden gegenereerd door het handmatig opstellen en uitvoeren van SQL queries.

Naast deze data beschikt de stichting over de resultaten van de Compass vragenlijsten ingevuld door cliënten. Invullen van de lijsten geschiedt bij de start van een behandeltraject, jaarlijks en wanneer een cliënt uit behandeling gaat.

Compass data is slechts voor een deel van de cliënten populatie beschikbaar. Dit komt doordat het invullen niet verplicht is en cliënten onder de 18 jaar en boven de 75 jaar zijn uitgesloten van deelname. Hiernaast zijn er anderstaligen die geen van de talen spreken waarvoor vertalingen van de lijsten beschikbaar zijn.

Het staat cliënten tevens vrij vragen op de lijst over te slaan waardoor niet alle gegevens voor elke invullende cliënt beschikbaar zijn. Verder zijn er door de jaren

heen diverse vragenlijsten in gebruik geweest waardoor mogelijk inconsistenties in de data aanwezig zijn.

De dataset gebaseerd op de vragenlijsten is op ad-hoc wijze verkregen door de verschillende lijsten samen te voegen in een bestand. De integriteit van de data van deze set is niet gecontroleerd waardoor er momenteel geen goed beeld is omtrent de kwaliteit van de data. Hiernaast is het onbekend in hoeverre de dataset representatief is voor de gehele cliëntenpopulatie van de stichting. De data wordt hierom momenteel enkel intern op experimentele wijze gebruikt.

3 Probleemanalyse

3.1 Probleemstelling

Doordat er in toenemende mate druk staat op het budget dat de stichting en de GGZ in zijn algemeen ter beschikking heeft en er een stabilisatie van de instroom van cliënten wordt verwacht staat de stichting in toenemende mate onder financiële druk.

Daarnaast moet de stichting aan een toenemend aantal kwaliteitseisen voldoen, en moet zij vergaand inzicht bieden in de effecten van de behandeling. Hierdoor is het van steeds groter wordend belang dat de juiste data beschikbaar is zodat er aan externe partijen kan worden gerapporteerd.

Voor behandelinhoudelijke rapportages wenst de stichting gebruik te maken van de vragenlijsten ingevuld in Compass. Deze data kan inzicht bieden in de cliënttevredenheid en effecten van behandeling.

Als gevolg van de toenemende financiële druk is het steeds meer van belang dat er op efficiënte wijze wordt omgesprongen met de beschikbare middelen die de stichting te bieden heeft.

Twee maatstaven van efficiëntie die binnen dit onderzoek worden onderzocht zijn het no-show gedrag en de behandelintensiteit.

3.2 Onderzoeksvragen

Omtrent de externe rapportages is de volgende hoofdvraag geformuleerd:

-Is de Compass data geschikt als basis voor externe rapportages?

De data is geschikt wanneer deze geen incorrecte gegevens bevat en wanneer de data representatief is voor de gehele cliënten populatie van de stichting. Deze hoofdvraag kan hierom verder worden onderverdeeld in de twee subvragen:

-Hoe is het gesteld met de data integriteit van de beschikbare dataset?

-Zijn er cliëntgroepen onder of over vertegenwoordigd in de data?

Uit de behoefte naar een grotere mate van efficiëntie is de volgende hoofdvraag opgesteld:

-Welke inefficiënte patronen zijn er in de data te vinden?

Inefficiënte patronen zijn hierbij patronen die geen direct bijdrage leveren aan de zorg van een cliënt. Om dit te onderzoeken wordt gekeken naar welke cliëntengroepen er in de data te vinden zijn en welke eigenschappen deze groepen bezitten. Tevens wordt de afspraakhistorie geanalyseerd met betrekking tot no-shows en verschillen in behandelfasen.

Deze hoofdvraag is dan onder te verdelen in de volgende subvragen:

-Welke cliëntgroepen zijn er uit de data te halen?

-Welke groepen hebben een verhoogd risico op een no-show?

-Welke factoren zijn bepalend voor de behandelintensiteit?

-Welke patronen met betrekking tot no-show en behandelfase zijn in de afspraakhistorie te vinden?

4 Datamining

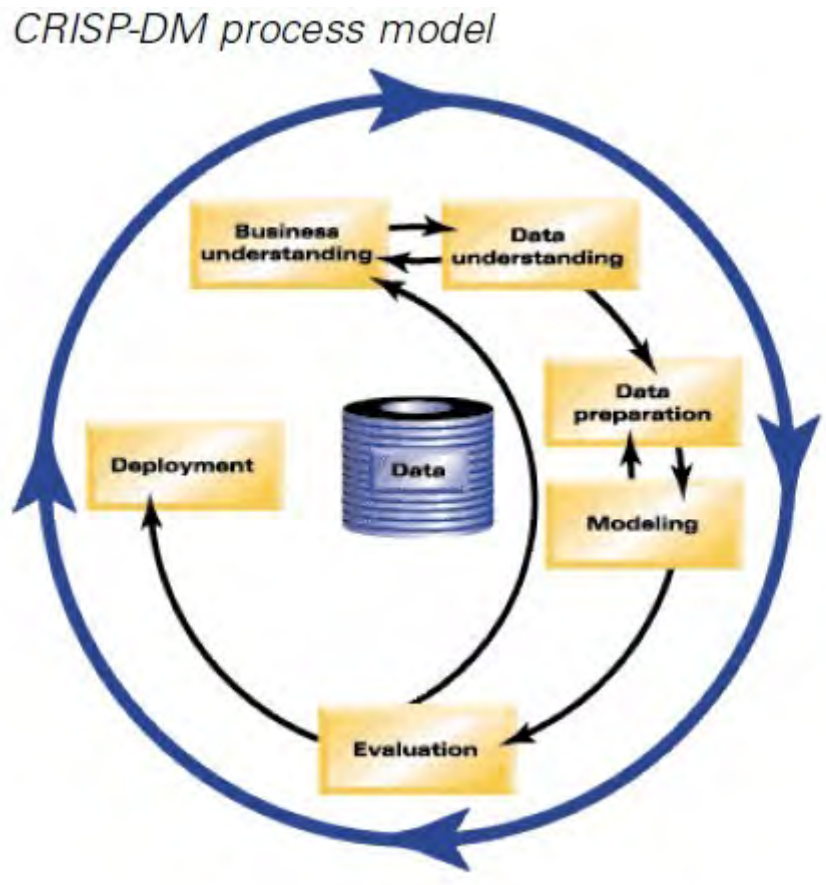
4.1 Het CRISP model

Een datamining project kan worden ingedeeld in een aantal verschillende stappen. Er zijn diverse modellen ontwikkeld waarin deze stappen en de interacties tussen de stappen zijn omschreven.

Twee veel gebruikte modellen zijn het Cross industry standard process for Datamining (CRISP) model [4] en het Knowledge Discovery in Databases (KDD) model [5].

Beide modellen vertonen in grote lijnen dezelfde stappen. Het CRISP model is echter meer toegespitst op een bedrijfsmatige toepassing. Er is daarom voor dit model gekozen als leidraad bij dit masteronderzoek. Er bestaat ook een variant van het CRISP model specifiek voor toepassingen in de gezondheidszorg beschreven in [6]. Voor dit project is deze specialisatie echter niet relevant omdat dit aangepaste model vooral betrekking heeft op het analyseren van diagnostische gegevens.

Het CRISP model beschrijft het datamining proces als een cyclus bestaande uit 6 stappen zoals weergegeven in Figuur 2



Figuur 2 CRISP cyclus

De eerste stap in dit proces bestaat uit een omgevingsanalyse waarbij inzicht wordt verkregen in de voor het project relevante bedrijfsprocessen. In de hierop volgende stap wordt er een exploratieve analyse van de data uitgevoerd om inzicht te verkrijgen in de data. Voorts wordt de data geschikt gemaakt voor de techniek die is gekozen op basis van de uitkomsten van de eerste twee stappen. Vervolgens kan er met de gekozen technieken worden gemodelleerd. Op basis van de uitkomsten hiervan kan het blijken dat er nog verdere aanpassingen van de data nodig zijn. In dit geval wordt de vorige stap opnieuw uitgevoerd.

Wanneer er geschikte resultaten zijn verkregen vindt er een evaluatie plaats en wordt de cyclus eventueel (deels) opnieuw uitgevoerd of worden de resultaten gebruikt.

4.2 Machine learning

Bij de modeleringsstap wordt binnen dit project gebruik gemaakt van zogeheten machine learning technieken. Er wordt bij het beantwoorden van de onderzoeksvragen gebruik gemaakt van classification, clustering en sequence mining technieken. Hiervoor wordt de dataminingtool Weka gebruikt in combinatie met Matlab.

De kracht van machine learning technieken is dat er zonder vooraf vastgestelde hypothese gezocht kan worden naar patronen. Dit in tegenstelling tot een statisticus die een hypothese formuleert en deze gaat bevestigen of ontkrachten door middel van statistische tests.

Er bestaat echter wel een link tussen beide manieren van werken. Veel machine learning technieken maken gebruik van statistische maatstaven voor het valideren van de resultaten.

4.2.1 Classification

Problemen waarbij het de taak is om records te classificeren in een vooraf bekend aantal klasse worden classificatie problemen genoemd [7]. Er wordt hierbij gezocht naar patronen binnen de data die een voorspellende invloed hebben op de te voorspellen klasse variabele ook wel doelvariabele genoemd.

Doordat vooral inzicht verkrijgen in de data van belang en niet zozeer het daadwerkelijk voorspellen is gekozen voor technieken die een model opleveren waarvan de uitvoer eenvoudig is te interpreteren in samenwerking met een domeinexpert.

Het is op dit moment niet de wens van de stichting om data van nieuwe cliënten door een model te laten toetsen om bijvoorbeeld cliënten met een verhoogd no-show risico te detecteren. Modellen met een complexe structuur die niet eenvoudig te interpreteren is zijn daarom voor dit project niet geschikt.

4.2.1.1 Validatie

Er zijn diverse methoden om de kwaliteit van een classificatie algoritme te beoordelen. Voor dit onderzoek is gekozen om gebruik te maken van zogeheten 10-Fold Cross-Validation. Hierbij wordt de dataset verdeeld in 10 verschillende sets. Vervolgens wordt er 10 keer een model opgesteld waarbij 9 sets worden gebruikt

voor het genereren van het model en 1 set voor de validatie. Bij de validatie wordt de voorspelde klasse vergeleken met de daadwerkelijke klasse.

Het eindresultaat is een model en een fout score waarbij de resultaten van de 10 stappen zijn gemiddeld. Deze aanpak is gekozen omdat er relatief weinig data beschikbaar is en zo alle data wordt benut bij het vinden van patronen.

Wanneer verschillende classificatie algoritmes worden gebruikt en er verschillende scores worden verkregen is het van belang om te toetsen of de resultaten statistisch significant verschillen. Een veel gebruikte manier om dit te doen is door gebruik te maken van een zogeheten gepaarde t-test.

Bij deze test wordt een betrouwbaarheidsinterval gemaakt voor het verschil in prestatie van beide algoritmes. Wanneer de waarde 0 zich niet binnen dit interval bevindt betekent dit dat de algoritmes significant verschillen. Voor meer details wordt verwezen naar [8].

4.2.2 Clustering

Bij clustering wordt de data gegroepeerd zonder gebruik te maken van bestaande klassen. Dit wordt gedaan door records die gelijkenis vertonen onder te verdelen in een zelfde groep c.q. cluster. Elk cluster heeft een zogeheten centroïde. Dit centroïde is een verzameling van factoren die het cluster typeren.

Gelijkenis tussen records wordt berekend door afstand tussen records te meten. Afhankelijk van de gebruikte technieken zijn hiervoor verschillende mogelijkheden.

Het aantal clusters kan vooraf worden opgegeven of afhankelijk van de techniek kan worden gezocht naar een optimaal aantal clusters. Voor een uitgebreide beschrijving van diverse clustering technieken wordt verwezen naar [9].

Bij dit onderzoek is clustering gebruikt voor het vinden van cliëntgroepen.

4.2.2.1 Validatie

In tegenstelling tot classification is de kwaliteit van het eindresultaat moeilijker te beoordelen. Er kan worden gekeken naar de gemiddelde afstand van de records tot de centroïdes. Dit getal geeft een indicatie van de spreiding van de data binnen de clusters. Deze afstand kan echter zo klein worden gemaakt als gewenst door het aantal clusters te verhogen. Door net zoveel clusters als records te nemen wordt uiteindelijk een perfecte score van 0 bereikt.

De beoordeling van de clusters wordt daarom gedaan door de resultaten met een domeinexpert door te nemen.

4.2.3 Sequence mining

Bij sequence mining wordt gezocht naar patronen die zich in de tijd voordoen.

De input data wordt bij sequence mining in [10] als volgt geformuleerd:

> Zij $I = \{i_1, i_2, \ldots, i_m\}$ een verzameling van m unieke items, het alfabet van de dataset. Een event is vervolgens een niet lege verzameling van items uit het alfabet. Een sequence is dan gedefinieerd als een geordende verzameling van events.

De input data D bestaat uit sequences met een sequence identifier *sid*. Elke sequence bestaat vervolgens uit een of meerdere events met event identifier *eid*.

Een voorbeeld aankoopgeschiedenis dataset die aan dit stramien voldoet is weergegeven in de onderstaande tabel.

Klantnummer	Volgnummer	Aankoop
1	1	(Portable MP3 speler, Headphone)
1	2	(MP3 speler hoes)
1	3	(Digitale camera)
2	1	(Portable MP3 speler)
2	2	(Digitale camera, Headphone)

Hierbij is het klantnummer de *sid* en het volgnummer het *eid*. Een sequence die in het voorbeeld voor beide klanten in de dataset voorkomt is Portable MP3 speler -> Digitale camera. Aankopen die onder hetzelfde volgnummer vallen en dus op hetzelfde moment plaatsvinden vormen geen sequence. Zo is Portable MP3 speler -> Headphone voor klant 1 geen sequence maar voor klant 2 wel.

Bij het zoeken naar sequences kan verder onderscheid worden gemaakt tussen events die elkaar direct opvolgen zoals Portable MP3 speler -> MP3 speler hoes voor klant 1 en events die elkaar niet direct opvolgen zoals Portable MP3 speler -> Digitale camera voor klant 1. Bij dit onderzoek is met beide manieren gewerkt.

Bij dit onderzoek is sequence mining gebruikt voor het analyseren van de cliënt afspraakhistorie. Hierbij is gezocht naar sequences die leiden tot een no-show. Hiernaast zijn sequences voor cliënten die in een verschillende fase van hun behandeling zitten met elkaar vergeleken om inzicht te verkrijgen in de cliëntendoorstroom.

4.2.3.1 Validatie

Bij de gevonden sequences wordt een zogeheten support gerapporteerd. De support geeft aan voor hoeveel procent van de sequences uit de database de sequence is gevonden. Zo heeft de indirecte sequence Portable MP3 speler -> Digitale camera uit bovenstaand voorbeeld een support van 100%. De indirecte sequence Portable MP3 speler -> MP3 speler hoes heeft daarentegen een support van 50% omdat deze sequence alleen voor klant 1 aanwezig is.

Uit de sequences kunnen vervolgens association rules worden gegenereerd. Dit zijn regels van de vorm “Als A heeft plaatsgevonden dan vindt B plaats”. A is hierbij een sequence bestaande uit een of meerdere events. B bestaat hierbij uit een enkel event.

Bij association rules wordt de support en confidence gerapporteerd. De support van de association rule A -> B is hierbij de support van sequence A, B ook wel genoteerd als $\text{supp}(A \rightarrow B)$.

De confidence is vervolgens gedefinieerd als $conf(A \rightarrow B) = \frac{\text{supp}\,(A \rightarrow B)}{\text{supp}\,(A)}$

De confidence is te interpreteren als de fractie van het aantal keren dat event B volgt na event A en het aantal keren dat even A plaatsvindt.

Het beoordelen van de association rules is geen eenvoudige opgave. De uitvoer kan bestaan uit honderden tot duizenden regels waarvan het grootste deel triviaal is. Om interessante regels uit de uitvoer te halen kan gebruik worden gemaakt van postprocessing technieken zoals het filteren van de uitvoer op basis van de items binnen de events.

4.2.3.2 Verschil analyse

Binnen dit onderzoek worden ook sequences tussen verschillende sets met elkaar vergeleken bij de analyse van de behandelfasen. Het kan interessante informatie zijn dat een bepaald afspraakpatroon voor een bepaalde groep een aanzienlijk hogere support heeft dan voor een andere groep.

Om er voor te zorgen dat de gevonden verschillen statistisch significant zijn is voor dit onderzoek een aanpak bedacht waar wordt gewerkt met een zogeheten binomiaal betrouwbaarheidsinterval voor de support. Een dergelijk interval geeft voor een gegeven betrouwbaarheid een range waarin de werkelijke waarde voor de support hoort te liggen. Wanneer voor een bepaalde sequence de intervallen voor twee verschillende sets niet overlappen is er sprake van een statistisch significant verschil in support. Bij de vergelijking is gebruikt gemaakt van het zogeheten Wilson betrouwbaarheidsinterval gegeven door [11]:

$$\left(p + \frac{Z^2}{2n} \pm Z\sqrt{\frac{p(1-p)}{n} + \frac{Z^2}{4n^2}}\right) \Big/ \left(1 + \frac{Z^2}{n}\right)$$

Waarbij n het aantal observaties is, Z het 1-a percentiel van de standaard normale verdeling is voor gewenste betrouwbaarheid a en p de waargenomen fractie is. Binnen deze context is p de gevonden support van een sequence.

4.2.4 Gebruikte algoritmes

In de onderstaande paragrafen worden de algoritmes die binnen dit onderzoek zijn gebruikt omschreven. Hierbij wordt tevens de keuze van de techniek verantwoord.

Bij het onderzoek is gebruik gemaakt van het softwarepakket WEKA in combinatie met Matlab. Er is voor deze combinatie gekozen omdat in WEKA een brede selectie aan technieken beschikbaar is en omdat in Matlab data eenvoudig kan worden gemanipuleerd en gevisualiseerd.

Naast WEKA is er geëxperimenteerd met het pakket Rapidminer. Beide pakketten bieden soortgelijke functionaliteiten. De ervaring met Rapidminer was echter dat in het programma meer handmatige stappen nodig waren om dezelfde resultaten te bereiken en de rekentijd aanzienlijk langer was.

4.2.4.1 ZeroR Classificatie

ZeroR staat voor “zero rule”. Deze classificatietechniek classificeert simpelweg alle records in een set volgens de meest voorkomende waarde. Wanneer bijvoorbeeld in een dataset met 60% mannen en 40% vrouwen het geslacht moet worden voorspeld worden alle records als man voorspeld. Zo wordt een score van 60% correct voorspelde records verkregen.

Deze techniek wordt binnen dit onderzoek gebruikt als vergelijkingsbasis. Wanneer een techniek geen aanzienlijk betere score oplevert dan ZeroR heeft deze weinig toegevoegde waarde.

4.2.4.2 OneR Classificatie

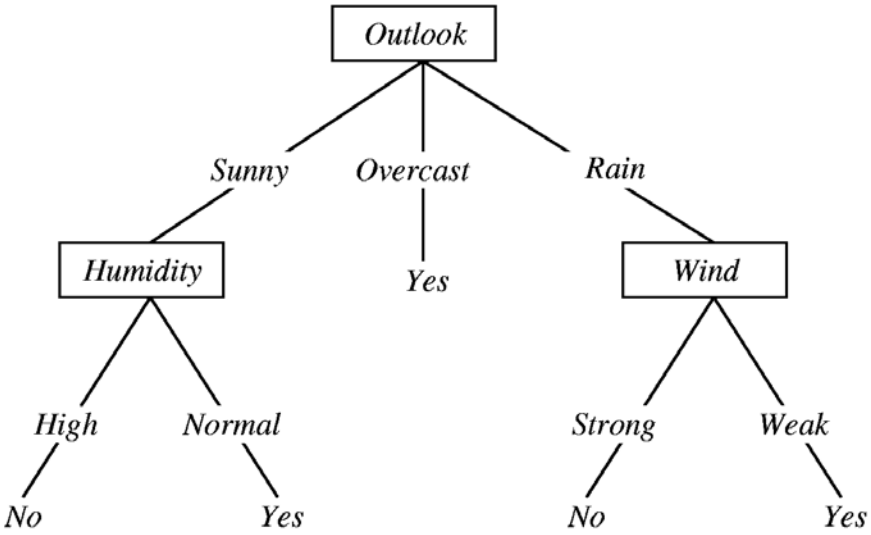
Zoals de naam waarschijnlijk al deed vermoeden staat OneR voor “one rule”. Dit is een classificatietechniek die een stap verder gaat dan het ZeroR algoritme. In plaats van de modale waarde wordt gezocht naar welke geïsoleerde variabele de doelvariabele het best kan beschrijven. Dit wordt gedaan door te tellen welke variabele/waarde combinatie het meest voorkomt in combinatie met de diverse waardes van de doelvariabele. De exacte werking en een evaluatie van dit algoritme wordt besproken in [12].

Deze techniek wordt zowel gebruikt als vergelijkingsbasis en om inzicht te verkrijgen.

4.2.4.3 J48 Classificatie

Het J48 algoritme is een techniek die een zogeheten beslisboom opbouwt.

Het algoritme bouwt zo een boom op waaruit 'als-dan' regels kunnen worden geformuleerd. Een voorbeeld van zo een boom is weergegeven in Figuur 3.



Figuur 3 Voorbeeld beslisboom [7]

De boom in het bovenstaande voorbeeld voorspelt of een dag geschikt is om buiten te gaan sporten aan de hand van vier weersfactoren.

De meest linker tak in de boom kan vertaald worden als de regel "Als het zonnig is en de luchtvochtigheid is hoog is het weer niet geschikt". Een vertakking vanaf de wortel tot en met een eindnode geeft zo steeds een regel weer.

De data op basis waarvan de boom is berekend is te zien in Tabel 1. De tabel geeft voor een aantal dagen vier weersfactoren met hierbij een ja/nee variabele die aangeeft of de dag geschikt is om buiten tennis te spelen. Voor het opbouwen van de boom wordt voor elk van de vier factoren de eerder genoemde information gain berekend. Deze maatstaaf zegt iets over de voorspellende waarde van een factor voor de doelvariabele *PlayTennis*. De boom wordt opgebouwd door de variabele met de meest voorspellende waarde bovenin de boom te plaatsen. In dit geval is dit de *outlook*. In de tabel is direct te zien dat de waarde *Overcast* direct correspondeert de waarde *yes* bij *PlayTennis* waardoor deze variabele een hoge mate van voorspellende waarde heeft.

Het process van steeds de factor selecteren met de hoogst voorspellende waarde wordt nu herhaald totdat alle factoren in de boom zijn gebruikt of totdat er geen verdere vertakkingen meer kunnen worden gemaakt zoals bij de combinatie *Outlook=Overcast* waarbij *PlayTennis* in alle gevallen op *Yes* staat.

Bij de Weka implementatie van dit algoritme is het tevens mogelijk om restricties op te leggen met betrekking tot de grootte van de boom zodat geen onnodig complex model wordt verkregen.

Binnen dit onderzoek is voor deze techniek gekozen omdat het resulterende model direct inzicht verschaft in de data en eenvoudig is te interpreteren.

Day	Outlook	Temp.	Humidity	Wind	Play Tennis
D1	Sunny	Hot	High	Weak	No
D2	Sunny	Hot	High	Strong	No
D3	Overcast	Hot	High	Weak	Yes
D4	Rain	Mild	High	Weak	Yes
D5	Rain	Cool	Normal	Weak	Yes
D6	Rain	Cool	Normal	Strong	No
D7	Overcast	Cool	Normal	Weak	Yes
D8	Sunny	Mild	High	Weak	No
D9	Sunny	Cool	Normal	Weak	Yes
D10	Rain	Mild	Normal	Strong	Yes
D11	Sunny	Mild	Normal	Strong	Yes
D12	Overcast	Mild	High	Strong	Yes
D13	Overcast	Hot	Normal	Weak	Yes
D14	Rain	Mild	High	Strong	No

Tabel 1 Train data voor de beslisboom

4.2.4.4 Expectation Maximization clustering

Het Expectation Maximization (EM) algoritme [13] is een techniek die een gegeven data set in een pseudo optimaal aantal clusters verdeeld. Het algoritme begint met het maken van een willekeurig gekozen cluster. Vervolgens wordt voor elk record de log liklihood berekend. Deze maatstaaf geeft een kans dat het betreffende record tot het cluster behoort. Vervolgens wordt het aantal clusters verhoogd en wordt de berekening opnieuw uitgevoerd. Dit wordt herhaald totdat de gemiddelde log liklihood niet meer verbetert. Het gevonden aantal clusters is nu het eind resultaat.

Bij de analyse is tevens geëxperimenteerd met k-means clustering. Bij deze techniek kiest de gebruiker vooraf het aantal clusters. Het is hierbij echter moeilijker om een goed aantal clusters te bepalen. Voor elk cluster wordt de gemiddelde afstand van de records tot het centroïde van de cluster gemeten. Deze afstand kan willekeurig klein worden gemaakt door het aantal clusters te verhogen. Vanwege dit nadeel is gekozen voor EM clustering.

4.2.4.5 Generalized Sequence Patterns

Het Generalized Sequence Patterns (GSP) algoritme [10] bepaalt iteratief voor een gegeven minimum support en maximale sequence grootte alle frequent sequences in een dataset. Voor dit onderzoek is een Matlab implementatie van het algoritme gemaakt.

Het algoritme begint met het berekenen van de support voor alle enkele items en deze op een lijst te plaatsen. De items waarvan de support onder het opgegeven minimum liggen worden vervolgens van de lijst verwijderd. Vervolgens vindt er een iteratieve stap plaats waarbij er telkens een lijst met kandidaat sequences wordt gemaakt. Deze kandidaat sequences zijn in de eerste iteratie het cartesisch product van alle enkele sequences. De sequences op de lijst zijn dan van grootte twee. Wederom wordt de support voor alle kandidaat sequences bepaald en vallen de sequences af waarvan deze support onder het opgegeven minimum liggen.

In elke volgende iteratie worden de kandidaat sequences steeds een item groter totdat de maximale sequences grootte is bereikt of totdat er geen nieuwe kandidaat sequences meer gemaakt kunnen worden omdat de support van alle kandidaten in een stap onder het opgegeven minimum liggen.

Het GSP algoritme is niet de meest efficiënte techniek maar is wel eenvoudig om snel te implementeren in een omgeving zoals Matlab. Efficiëntie is tevens van minder belang omdat de gebruikte dataset relatief klein is.

5 Exploratieve data analyse

Voorafgaand aan het hoofdonderzoek is er een exploratieve analyse van de data uitgevoerd. In deze analyse is gezocht naar opvallende patronen en eigenschappen van de data. Het doel van deze analyse is om inzicht in de data te verkrijgen en verder bekend te raken met de context door het bespreken van de resultaten met domeinexperts. In deze analyse is tevens een integriteitscheck opgenomen.

Binnen het kader van het Crisp model zoals beschreven in het voorgaande hoofdstuk omvat deze analyse de stappen “Data preperation” en “Data understanding”.

5.1 Compass

Compass is een software systeem om online vragenlijsten bij cliënten af te nemen. Dit systeem is in 2007 door de stichting in gebruik genomen. Op basis van de ingevulde vragen kan voor een cliënt een score worden berekend die vervolgens kan worden vergeleken met de norm-score van de algemene bevolking.

Tijdens het uitvoeren van dit onderzoek is de stichting bezig om een vernieuwde versie van het systeem te ontwikkelen.

Wanneer een cliënt een vragenlijst invult worden de resultaten weggeschreven naar een tijdelijk bestand op de webserver waar de Compass applicatie draait. Er wordt regelmatig een export batch gedraaid waarmee deze bestanden worden weggeschreven naar het archief van de stichting. Deze batch wordt handmatig gestart.

De Compass dataset die gebruikt is voor dit onderzoek is verkregen door de vragenlijsten van alle cliënten in het archief samen te voegen. De volgorde van de kolommen in de individuele lijsten is niet altijd gelijk. Het samenvoegen is hierom gedaan door de kolom namen te vergelijken.

De dataset bestaat uit diverse typen vragenlijsten doordat er door de jaren heen verschillende lijsten zijn gebruikt en doordat voor verschillende doelgroepen verschillende lijsten worden gebruikt.

De vragenlijsten zijn gegroepeerd in drie zogeheten testbatterijen: basis, basis voor vluchtelingen en asielzoekers en aanvullend. De aanvullende batterij bestaat uit een aantal lijsten die speciaal op aanvraag kunnen worden afgenomen. Binnen de stichting wordt hier nauwelijks gebruik van gemaakt. Een afname bestaat uit het invullen van een batterij. Hierdoor vult een cliënt bij afname doorgaans meerdere vragenlijsten in.

5.1.1 Data Beschrijving

Bij de analyses zijn twee verschillende versies van de dataset gebruikt. De eerste versie is een ruwe versie met data zoals die uit de Compass applicatie komt. De tweede versie is een door de stichting opgeschoonde versie. In deze versie is incorrecte data zonder verdere inspectie verwijderd. Hierom is de ruwe versie gebruikt bij de controle op correcte data types en cliënt nummers en de opgeschoonde versie gebruikt bij de overige analyses.

De dataset bestaat uit een samenvoeging van 17 vragenlijsten met elk een verschillend aantal kolommen. Een overzicht van alle lijsten is te vinden in bijlage 1. De ruwe versie van de dataset bestaat uit 1360 kolommen en 3694 records. De opgeschoonde set bevat 631 kolommen en 3686 records. Het verschil in records komt door het verwijderen van cliëntnummers die buiten de toegestane range vallen. Het verschil in kolommen komt doordat data behorende bij een zelfde vraag die op verschillende momenten en waarschijnlijk door verschillende cliënten in de ruwe versie in sommige gevallen onder meerdere kolommen is verdeeld.

De eerste 5 kolommen in de data bevatten de basisgegevens die in principe voor ieder record aanwezig dienen te zijn. Deze gegevens bestaan uit het cliëntnummer, volgnummer, datum van invullen, geboortedatum en geslacht.

5.1.2 Integriteit

Een van de hoofdvragen van het onderzoek is of de Compass data geschikt is als basis voor externe rapportages. De deelvraag die in deze sectie wordt behandeld is:

-Hoe is het gesteld met de data integriteit van de beschikbare dataset?

In overleg met een domeinexpert zijn een aantal aspecten van de data geselecteerd. De resultaten en technieken gebruikt bij deze controle zijn hieronder beschreven.

5.1.2.1 Dubbele records \ kolommen

De gehele dataset is gecontroleerd op dubbele records en kolommen door voor alle records en kolommen een zogeheten MD5 [14] checksum te berekenen. Het MD5 algoritme berekent voor een input string van willekeurige grootte een 128 bits getal. Wanneer voor twee invoer waardes het getal verschillend is kan worden uitgesloten dat de waardes gelijk zijn. Bij gelijke waardes kan worden aangenomen dat de invoer waardes gelijk zijn aan elkaar.

Bij deze controle is aangetoond dat alle kolommen en records in zowel de ruwe als de opgeschoonde versie uniek zijn.

5.1.2.2 Data types

Alle velden in de dataset zijn gecontroleerd op het correcte data type. De kolommen met antwoorden bij de vragenlijsten, het cliëntnummer en het volgnummer bevatten allen numerieke data. De kolom met het geslacht bevat enkel de waardes 1 (man) en 2 (vrouw). De kolom met de ontvangstdatum bevat datums van de vorm dd-mm-yyyy. In de onbewerkte set zijn 29 van de 3959 datums in de geboortedatum kolom van een incorrect formaat, 6 van deze waardes betreffen lege velden. Het cliëntnummer, volgnummer, geslacht en de ontvangstdatum bevatten geen lege waardes. Voor de verdere analyses binnen dit onderzoek is het geboortedatum veld in de Compass data niet van belang. Er zijn daarom geen verdere stappen ondernomen met de incorrecte data.

5.1.2.3 Antwoordschalen

Doordat er door de jaren heen verschillende versies van vragenlijsten zijn gebruikt bestaat er enige twijfel over de inhoudelijke correctheid van de data. Het zou bijvoorbeeld kunnen zijn dat een vraag is geherformuleerd waardoor de antwoordschaal is omgedraaid.

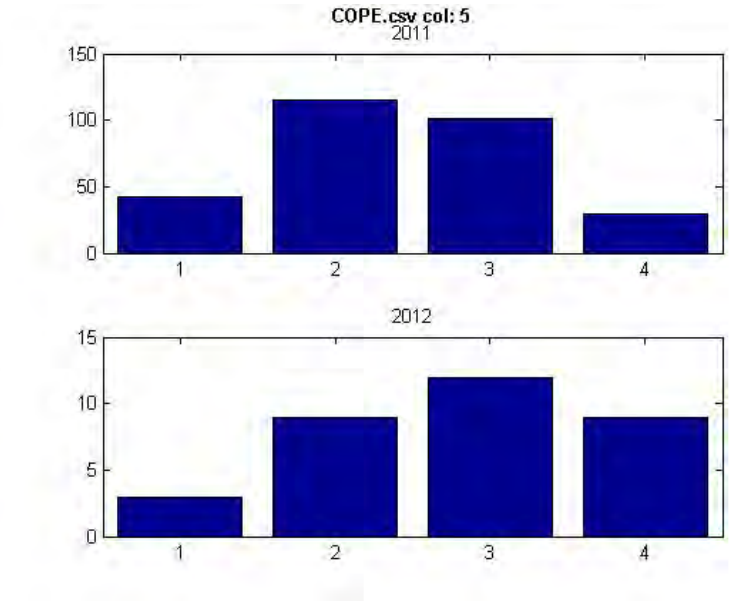
Om potentiële kolommen waarbij dit het geval is op te sporen zijn per jaar per vraag betrouwbaarheidsintervallen berekend voor de gemiddelde waarde bij een kolom. Dit interval is een schatting van het interval waarin het gemiddelde behoord te liggen bij een gegeven betrouwbaarheid. Wanneer twee intervallen voor twee opeenvolgende jaren niet overlappen is er sprake van een significant verschil. Dit interval wordt berekend met [15]

$$\left[\hat{X}_n - \Phi^{-1}(1-\alpha)\frac{S_n}{\sqrt{n}}, \hat{X}_n + \Phi^{-1}(1-\alpha)\frac{S_n}{\sqrt{n}}\right]$$

Waarbij $\hat{X}_n$ staat voor het steekproef gemiddelde, S_n de steekproef variantie. n het aantal observaties en α de betrouwbaarheid. Bij deze analyse is voor α 0,95

genomen zodat een 95% betrouwbaarheidsinterval wordt verkregen. Dit betekent dat wanneer er voor meerdere steekproeven een dergelijk interval wordt berekend het ware gemiddelde zich binnen 95% van deze intervallen bevindt.

Wanneer twee intervallen geen overlap vertonen is dit een indicatie voor een significant verschil. Bij zulke verschillen is vervolgens een histogram geplot om de verdeling van gegeven antwoorden tussen twee jaren visueel te kunnen vergelijken zoals te zien is in Figuur 4. Op deze wijze kunnen mogelijk foutieve kolommen worden geïdentificeerd.



Figuur 4 Verschil in verdeling tussen twee jaren binnen een kolom

De analyse leverde 82 significante verschillen op. Het is aan de domeinexpert om deze resultaten door te nemen en te beoordelen welke gevallen nader dienen te worden onderzocht. Aan de hand van de histogrammen kan dit in een oogopslag worden beoordeeld.

Er moet echter worden opgemerkt dat wanneer er geen significant verschil is gevonden dit niet per definitie betekend dat data voor de betreffende vraag correct is. In bijvoorbeeld het geval van een symmetrisch verdeelde respons blijven de betrouwbaarheidsintervallen gelijk wanneer de schaal wordt omgedraaid.

5.1.2.4 Cliëntgegevens

Van de ruwe data was het bij de stichting reeds bekend dat er enkele records met ongeldige cliëntnummers in de data aanwezig waren. Deze records betroffen 7 cliëntnummers die buiten de toegestane range vielen. In de opgeschoonde set zijn deze records verwijderd.

Voor alle cliëntnummers in de opgeschoonde data is gecontroleerd of deze voorkwamen in de centrale cliënt database van de stichting. Ondanks dat de nummers in de geldige range vielen bleken er 6 cliëntnummers aanwezig te zijn die niet voorkwamen in de database.

Verder bleek voor 53 records de combinatie van cliëntnummer en geboortedatum niet te kloppen met hoe deze in de centrale database stonden opgeslagen. In 11

gevallen was de combinatie cliëntnummer en geslacht incorrect en in 1 geval was zowel het geslacht als de geboortedatum incorrect.

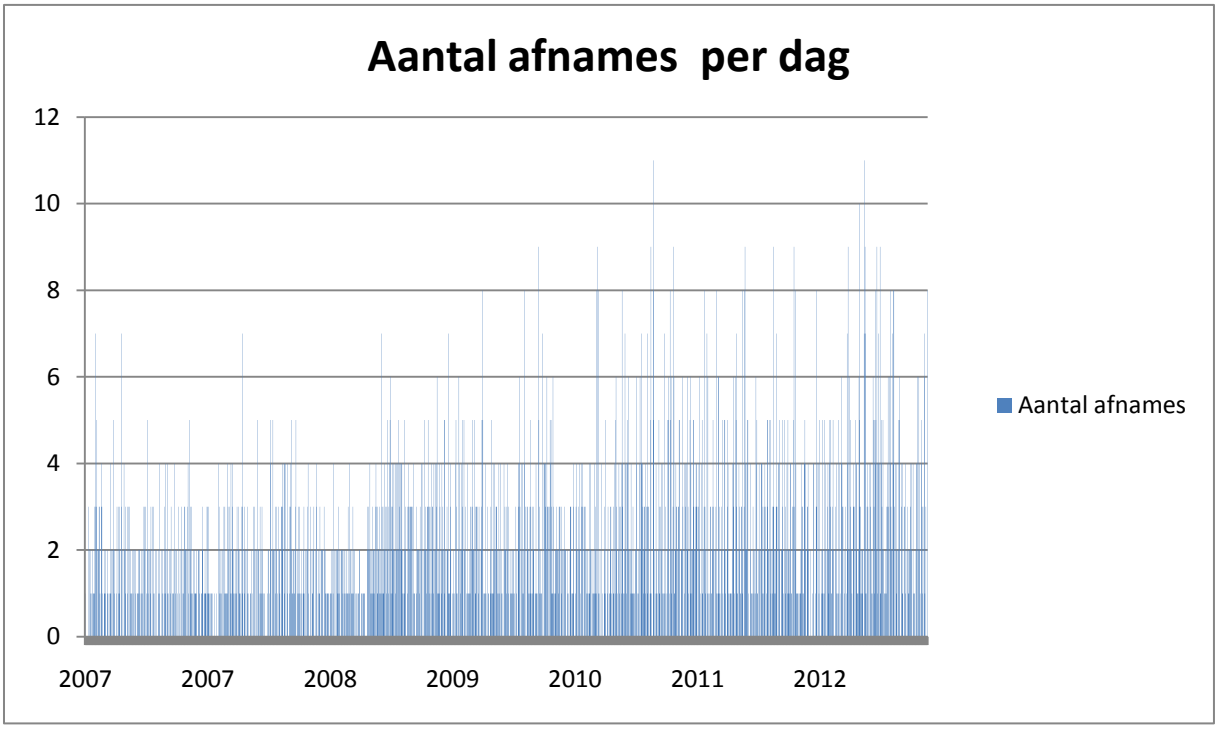
Binnen dit onderzoek zijn alleen de cliëntnummers van belang die voorkomen in de centrale database. De records behorend bij de niet voorkomende cliëntnummers zijn daarom niet relevant.

Bij inspectie bleek dat de foutieve geboortedatums veelal typefouten betroffen. Om deze reden en mede omdat het een relatief klein aantal records betrof is besloten om de betreffende records in de data te laten staan.

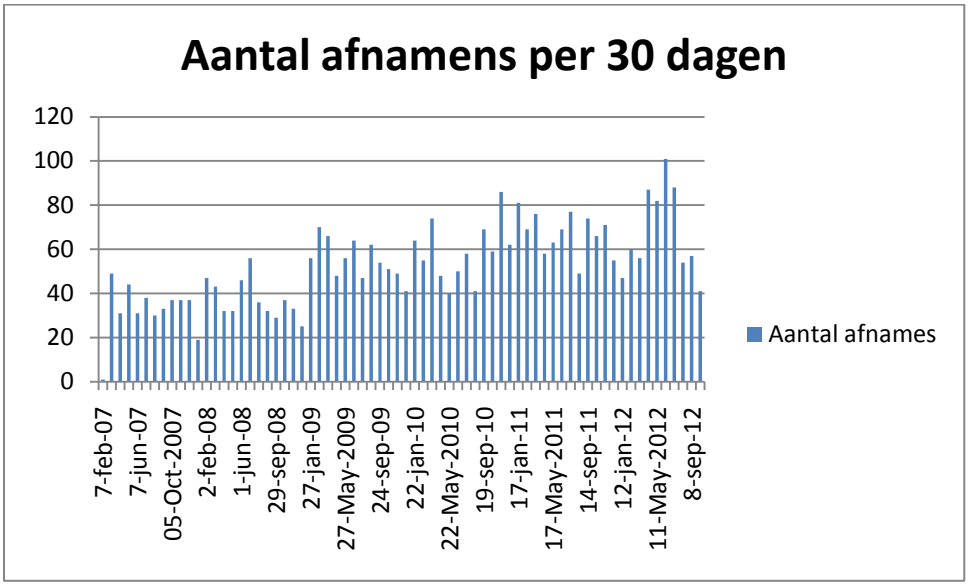
5.1.3 Exploratieve analyse

5.1.3.1 Ontvangstdatum

Aan de hand van de ontvangstdatum kan het gebruik van het systeem door de tijd heen worden geanalyseerd. In Figuur 5 is te zien hoe het aantal records per datum door de tijd heen een lichte stijging lijkt door te maken. In Figuur 6 waarbij het aantal afnames is geaggregeerd per 30 dagen wordt dit beeld bevestigd.

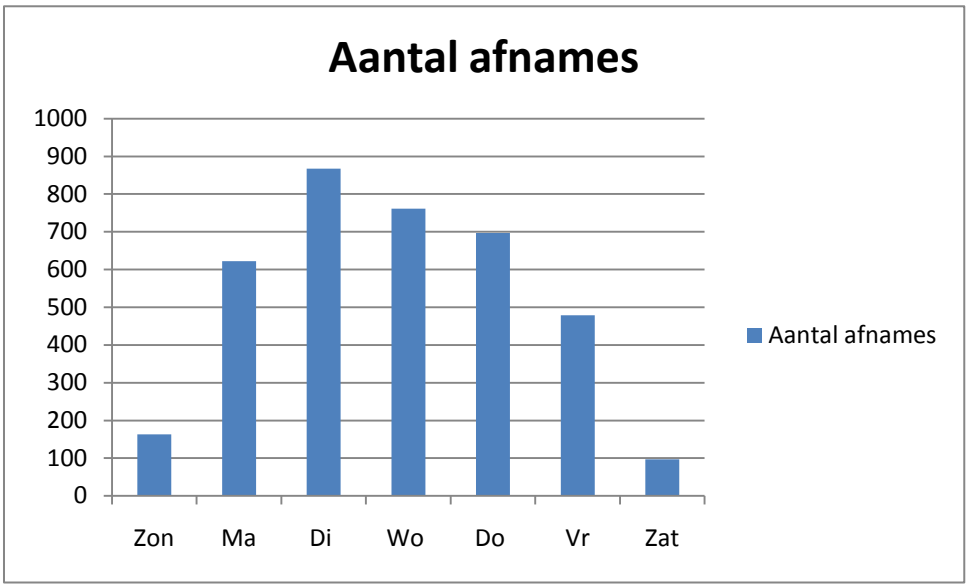


Figuur 5 Aantal Compass afnames door de tijd per dag



Figuur 6 Aantal Compass afnames door de tijd per 30 dagen

Aan de hand van deze kolom kan tevens het aantal afnames per weekdag worden bepaald. Dit is weergegeven in Figuur 7. Te zien is dat de meeste afnames doordeweeks plaatsvinden op dinsdagen. Een kleinere afname in het weekend is te verwachten doordat een aanzienlijk deel van de cliënten de lijst op de stichting invult. Een afname staat hierbij doorgaans voor het invullen van een batterij. Het is in principe mogelijk dat een cliënt het invullen van een batterij over meerdere dagen verdeelt. Hoewel dit uit de data niet is af te lezen gebeurt dit volgens de stichting sporadisch.

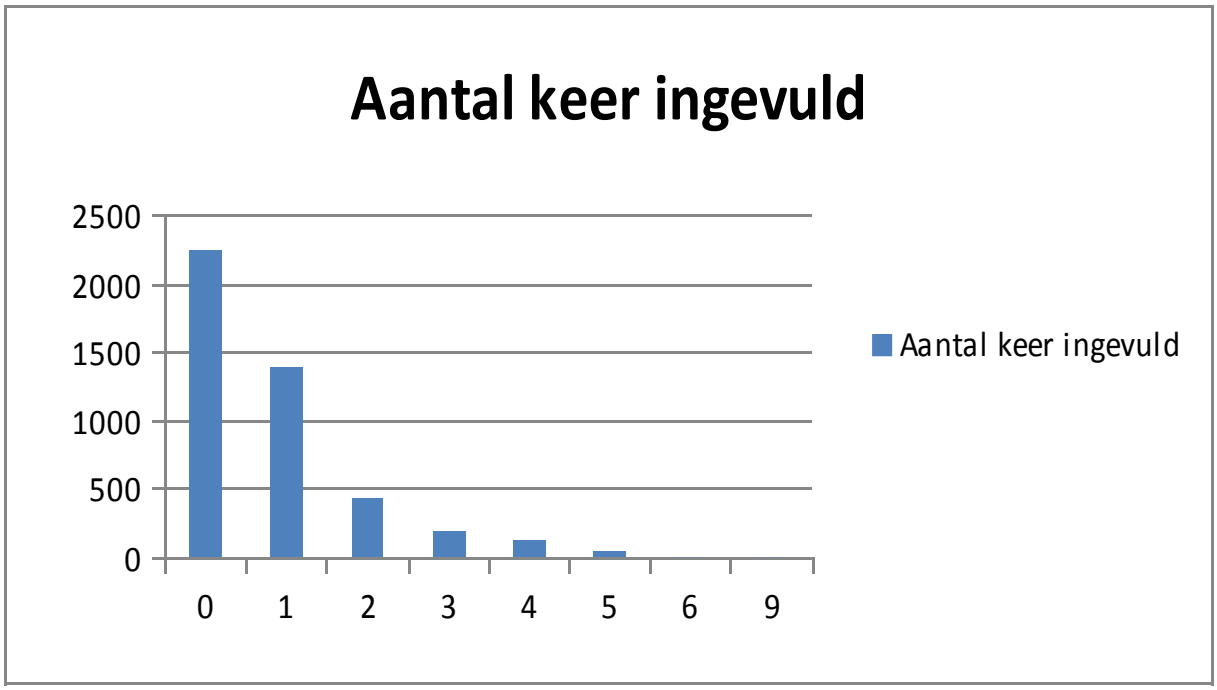


Figuur 7 Aantal Compass afnames per weekdag

5.1.3.2 Cliënt nummer

De 3694 records in de Compass set corresponderen met 2214 cliënten. Hieruit is op te maken dat een groot deel van de cliënten slechts eenmaal een batterij invult.

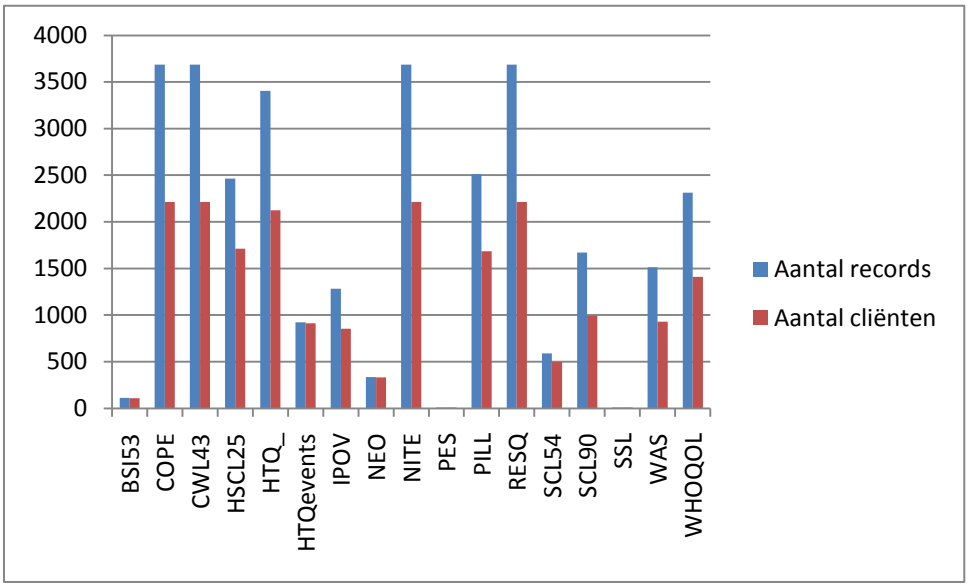
In Figuur 8 is de verdeling weergegeven van het aantal afnames per cliënt. Te zien is dat 1400 cliënten slechts eenmaal een batterij invullen en van slechts een klein deel van de cliënten meerdere afnames beschikbaar zijn. Van de meeste cliënten is echter geen meting bekend. Voor de vergelijking is alleen gekeken naar cliënten die niet bij equator of PDC in behandeling zijn omdat deze groep per definitie is uitgesloten van Compass deelname.



Figuur 8 Histogram van het aantal ingevulde lijsten per cliënt

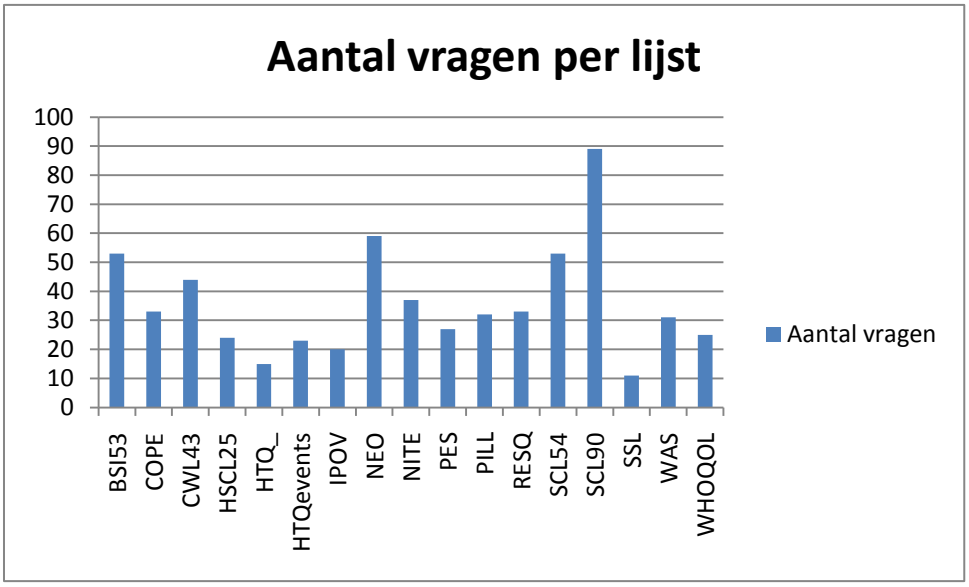
5.1.3.3 Dataset opbouw

De dataset bestaat uit resultaten afkomstig van 17 verschillende vragenlijsten. Voor verschillende cliëntengroepen worden verschillende batterijen gebruikt die bestaan uit een selectie van lijsten. Tussen de batterijen is overlap mogelijk. Figuur 9 geeft de opbouw van de dataset weer door voor iedere lijst binnen de dataset het aantal records en het aantal cliënten per lijst weer te geven.



Figuur 9 Aantal records en cliënten per vragenlijst binnen de dataset

In Figuur 10 is vervolgens per vragenlijst het aantal vragen te zien.



Figuur 10 Aantal kolommen per vragenlijst

5.1.4 Conclusies

De data in de Compass set lijkt voor het grootste deel correct. Er is echter aangetoond dat er statistisch significantie verschillen bestaan tussen de verdelingen in antwoorden gegeven voor gelijke vragen in opeenvolgende jaren. Deze verschillen vereisen nader onderzoek om te bepalen wat hiervan de oorzaak is.

Er was verder te zien dat van slechts een klein deel van de cliënten meerdere Compass metingen beschikbaar zijn.

Voor de verdere analyses binnen dit onderzoek zijn deze punten echter niet relevant omdat de inhoud van de Compass metingen bij het beantwoorden van onderzoeksvragen niet wordt gebruikt.

5.2 User cliënt data

User is een automatiseringsapplicatie voor zorginstellingen ontwikkeld door de softwareontwikkelaar Impulse. Het systeem is bij verschillende geestelijke gezondheidszorg instanties binnen Nederland in gebruik. Sinds eind 2009 wordt het pakket gebruikt bij Stichting Centrum ’45.

Het pakket werkt in combinatie met een Oracle database waar alle gegevens worden opgeslagen. Omdat de stichting niet alle gewenste rapportages rechtstreeks met User kan opstellen is er bij de stichting een medewerker in dienst die de rapportages opstelt door de gegevens door middel van SQL direct uit de database te halen.

Er is echter geen overzicht van de structuur van de database beschikbaar. Tevens is de gehanteerde naamgeving van de kolommen en tabellen niet consistent.

Naast de gegevens die sinds de introductie van User bij de stichting zijn geregistreerd bevat de database geconverteerde gegevens van het oude pakket. Bij deze conversie zijn alleen de cliënt en diagnose gegevens meegenomen. Afspraak en behandeling gegevens zijn slechts na 2005 bij het invoeren van de DBC bekend.

De data bevat diagnostische gegevens waarbij gebruik is gemaakt van de internationale DSM4 classificatie. De classificatie bestaat uit 5 assen. Elke as betreft een bepaald aspect van de diagnose. Op elke as kunnen meerdere diagnoses worden gesteld. Invullen van waardes op de assen is niet verplicht en niet altijd van toepassing.

De eerste as betreft de klinische stoornis die is vastgesteld. De tweede as betreft mogelijke persoonlijkheidsstoornissen. De derde as betreft eventuele lichamelijke klachten. Op de vierde as worden psychosociale en omgevingsfactoren vastgesteld. Ten slotte betreft de vijfde as de GAF score. Deze score geeft het algemeen functioneren van een cliënt aan op een schaal van 0 tot 100. De 0 waarde is een speciale waarde die aangeeft dat er nog geen GAF score is vastgesteld. De 0 waarde is bij verdere analyses beschouwd als leeg veld.

In User kunnen op assen 1, 2 en 4 vijf diagnoses worden ingevuld. In de praktijk zijn dit er echter in de meeste gevallen hoogstens 2. Voor as 1 en 2 zijn hierom de eerste twee diagnoses in de dataset opgenomen. Op as 4 is alleen de eerste diagnose in de dataset opgenomen. Op as 3 kunnen drie diagnoses worden ingevuld. Doorgaans is er op deze as hoogstens 1 diagnose ingevoerd. Van as 3 is daarom alleen de eerste diagnose in de set opgenomen.

5.2.1 Data Beschrijving

Voor dit project zijn alle gegevens van inschrijvingen na 2005 gebruikt. Deze grens is genomen omdat enkel na deze datum afspraakgegevens beschikbaar zijn. Helaas zijn mogelijk interessante variabelen zoals de burgerlijke staat en leefsituatie na deze periode slecht bijgehouden. Voor deze variabelen is voor minder dan een kwart van de cliënten een waarde bekend. De betreffende velden zijn daarom niet in de dataset opgenomen.

De ruwe dataset bestaat uit 6067 records en 30 variabelen.

5.2.2 Data transformatie

Om de dataset geschikt te maken voor verdere analyse zijn er een aantal transformaties op de data uitgevoerd. Zo zijn voor een aantal nominale variabelen bepaalde waardes in een klasse samengevoegd.

Het is mogelijk om de meeste transformaties in SQL uit te voeren. Het nadeel is echter dat SQL hier niet specifiek voor is bedoeld en de nodige code de query onoverzichtelijk maakt. Er is daarom voor gekozen om de ruwe data in Matlab te laden en in deze omgeving de nodige bewerkingen uit te voeren.

Voor de meeste nominale variabelen zijn klassen binnen de variabele samengevoegd. Zo zijn waardes die relatief weinig voorkomen onder een groep ‘overig’ verdeeld. Dit betreft de groepen met 200 cliënten of minder.

Ook zijn waardes die in betekenis weinig verschillen samengevoegd. Dit zijn alle doelgroepen die kinderen betreffen. Deze hercodering is in overleg met een domeinexpert uitgevoerd. De zo verkregen nieuwe variabelen zijn van de suffix _G voorzien. Een uitzondering hierop is het *geboorteland*. Deze variabele is gegroepeerd tot de *geboorteregio*.

Naast het groeperen zijn er een aantal extra variabele aan de set toegevoegd. Deze variabelen zijn de *gaf_globaal*, *inschrijfduur*, *afstand, inwoners* en *in_compass*.

Gaf_globaal is het gemiddelde van de eerste en laatste GAF score van een cliënt. Wanneer slechts een meting beschikbaar is, is deze waarde genomen. Deze waarde zegt iets over het algemene welbevinden van de cliënt gedurende de behandeling.

De inschrijfduur is eenvoudig bepaald aan de hand van de inschrijf- en uitschrijfdatum. De afstand is berekend door de postcode van de cliënt en de postcode van de behandellocatie te gebruiken in combinatie met een externe postcode tabel. Inwoners geeft het aantal inwoners van de woongemeente van de cliënt aan. Deze waarde is bepaald door gegevens van het Centraal bureau voor statistiek aan de dataset toe te voegen. De variabele *in_compass* geeft aan of er een Compass meting voor de cliënt in het Compass data bestand aanwezig is.

Behandelintensiteit is een gedefinieerd als het gemiddeld aantal afspraken per 30 dagen en afgeleid uit *behandeluren* en *inschrijfduur*.

De onderstaande tabel geeft een overzicht van de data na de transformaties. De getransformeerde set bestaat uit 27 variabelen.

Variabele	Omschrijving
LEEFTIJD	De leeftijd bij het moment van inschrijven
GEBOORTE_REGIO	De regio van herkomst van de cliënt. Voor deze variabele is het geboorteland gegroepeerd in: Europa, Afrika, Zuid-Amerika, Amerika, Oceanië, Midden Oosten en Azië.
GESLACHT	Geslacht van de cliënt.
AFSTAND	Reisafstand bepaald aan de hand van de postcode
AANMELDKLACHT	Hoofdklacht van de cliënt tijdens het moment van inschrijven
INSCHRIJFDUUR	Totale inschrijfduur in dagen van de cliënt tot op heden of moment van uitschrijving.
LOCATIE	De locatie van de stichting waarop de cliënt het vaakst is behandeld.
GAF_GLOBAAL	Gemiddelde GAF score van een cliënt.
DIAG_AS1_1_G	Primaire diagnose op as 1. De diagnose op as 1 geeft de primaire psychische ziekte van de cliënt.
DIAG_AS1_2_G	Secundaire diagnose op as 1. Deze variabele is opgenomen in het geval er meerdere ziektes zijn geconstateerd. In het geval van stichting centrum ’45 zijn dit er doorgaans hoogstens twee.
DIAG_AS2_1_G	Primaire diagnose op as 2. Op as 2 worden achterliggende persoonlijkheidsstoornissen geregistreerd.
DIAG_AS2_2_G	Secundaire diagnose op as 2
DIAG_AS3_1_G	Primaire diagnose op as 3. Op as 3 worden eventuele lichamelijke klachten geregistreerd.
DIAG_AS4_G	Diagnose op as 4. Deze diagnose geeft eventuele omgevingsproblematiek aan zoals werkeloosheid.
DBC_REDEN_UIT	Reden voor uitschrijving van een cliënt zoals geregistreerd in de diagnose behandel code.
BEHANDEL_UREN	Het totaal aantal uren behandeling per cliënt.
UITN_COMPASS	Een binaire variabele die aangeeft of een cliënt is uitgenodigd voor een Compass meting.
AANTAL_NEVEN_KL	Het aantal neven cliënten dat bij de stichting in behandeling is. Voornamelijk bij vluchtelingen komt het voor dat er hele gezinnen in behandeling zijn.
MEDICATIE	Geeft aan of een cliënt een consult heeft gehad met betrekking tot medicatie. Cliënten die een dergelijke afspraak hebben gehad gebruiken doorgaans medicatie.
SETTING	Geeft de modale behandelvorm van de cliënt weer. Mogelijke waardes zijn klinisch, poliklinisch en intake.
TOLK	Geeft aan of de cliënt gebruik heeft gemaakt van een tolk bij de behandeling.
VERZ	Geeft aan hoe een cliënt verzekerd is.
DOELGROEP_G	De doelgroep waarbinnen een cliënt valt.
NO_SHOW_KL	Het no-show percentage voor klinische afspraken per cliënt.
NO_SHOW_PL	Het no-show percentage voor poliklinische afspraken per cliënt.
IN_COMPASS	Een binaire variabele die aangeeft of er Compass data beschikbaar is voor de cliënt.
BEHANDEL_INTENSITEIT	Gemiddeld aantal afspraken per 30 dagen

5.2.3 Exploratieve analyse

De exploratieve analyse van de User data is beperkt tot een aantal variabelen. Het doel van de analyse is voornamelijk om inzicht en kennis omtrent de data te verkrijgen alvorens er wordt begonnen aan de hoofdanalyse. In dit hoofdstuk zijn een aantal opvallendheden uitgelicht.

5.2.3.1 Doelgroep

De stichting deelt cliënten in een aantal doelgroepen in. Wegens een systeem fout is de doelgroep variabele echter voor een deel van de cliënten verloren gegaan. Aan de hand van een aantal vuistregels zoals “een onverzekerde cliënt is een vluchteling” is deze variabele deels door een domeinexpert hersteld.

Er wordt oorspronkelijk met 14 doelgroepen gewerkt.

Een aantal doelgroepen zijn samengevoegd met de doelgroep “overig of onbekend” omdat deze groepen relatief weinig cliënten bevatten. De betreffende groepen met het aantal cliënten zijn: *Beroeps gerelateerd* (113), *Eerste generatie WOII* (74), *Kinderen van getraumatiseerden* (74),*Burgeroorlogsgetroffenen (*54), *Partners van getraumatiseerden* (52),*Georganiseerd geweld anderszins* (35).

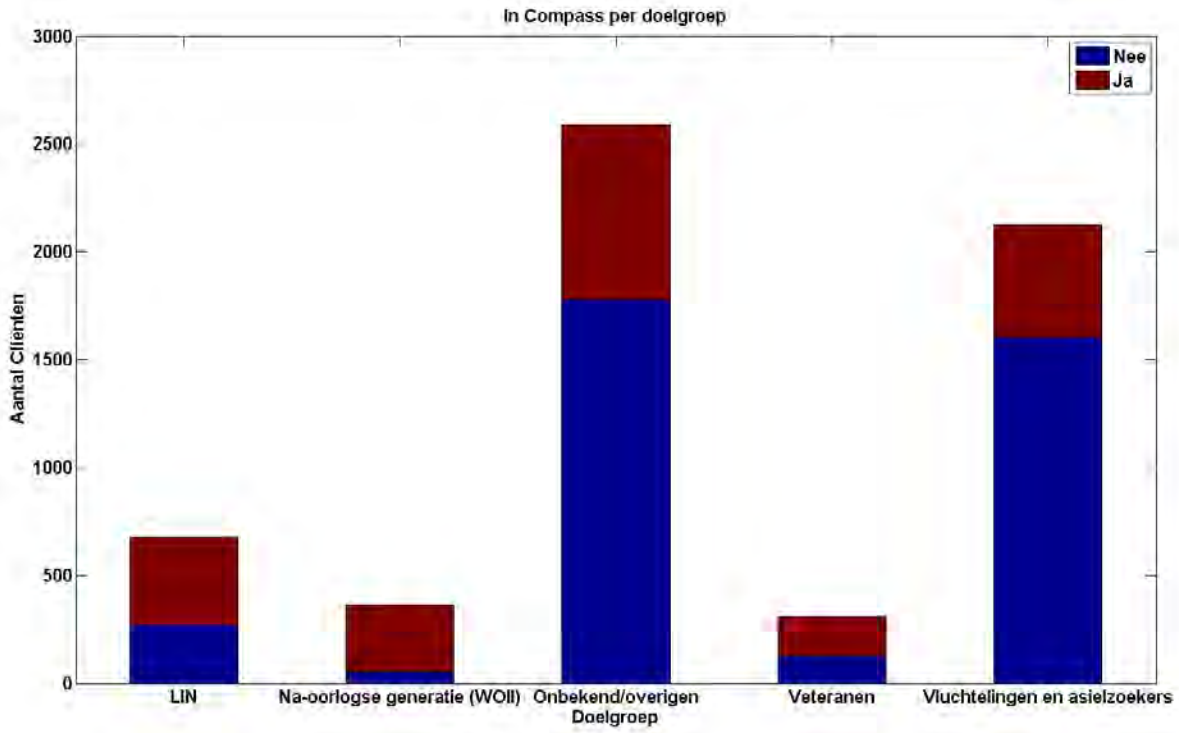
Onderstaande tabel geeft een overzicht van de doelgroepen met hun aantallen zoals gebruikt in dit onderzoek.

Doelgroep	Aantal
Onbekend/Overige	2589
Vluchtelingen en Asielzoekers	2127
Naoorlogse generatie WOII	364
Veteranen	311
Lang in Nederland verblijvende vluchtelingen (LIN)	676

Wanneer de doelgroep wordt geplot in combinatie met *In_Compass,* is te zien welk deel van de cliënten binnen een bepaalde doelgroep in de Compass dataset aanwezig is. Dit is weergegeven in Figuur 11.

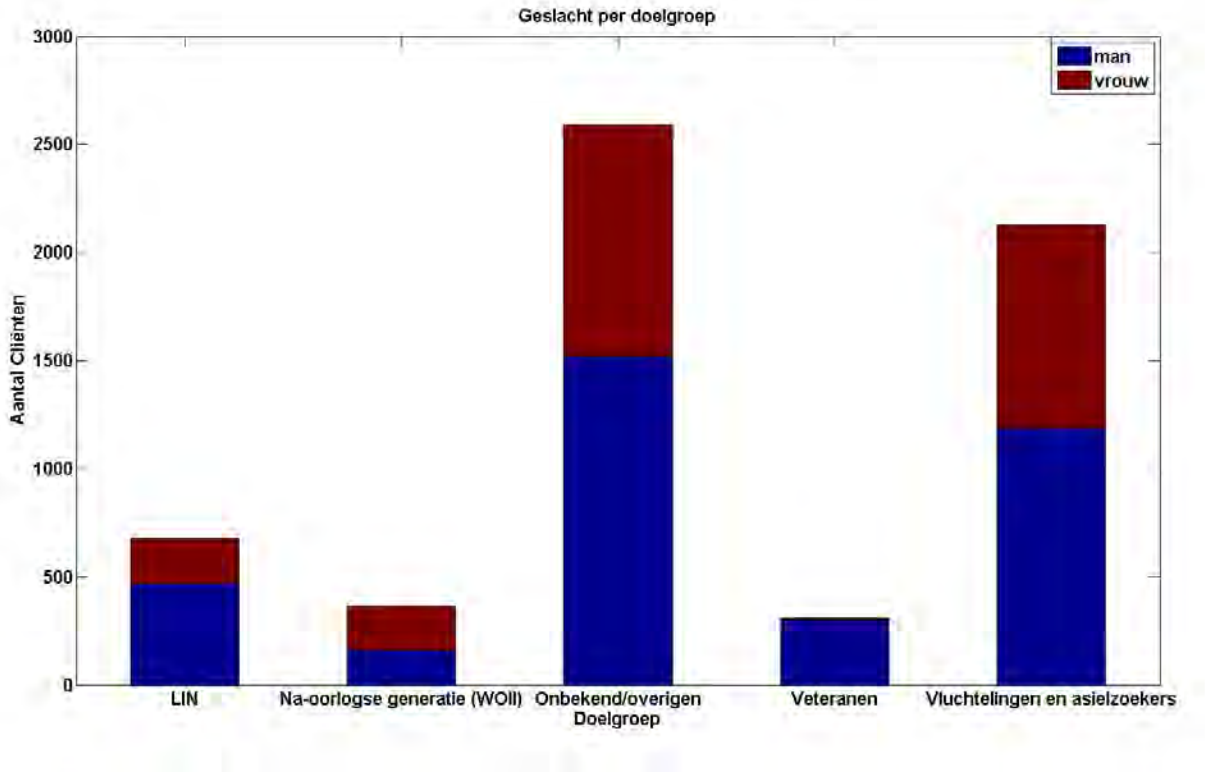
Het rode deel van de staven in dit diagram geeft aan welk deel in de Compass set zit en het blauwe deel van welke cliënten geen Compass data beschikbaar is.

Te zien is dat de kleinere groepen goed vertegenwoordigd zijn binnen de Compass data. De goede vertegenwoordiging van de kleine groepen in de Compass data komt overeen met de verwachting van de stichting.



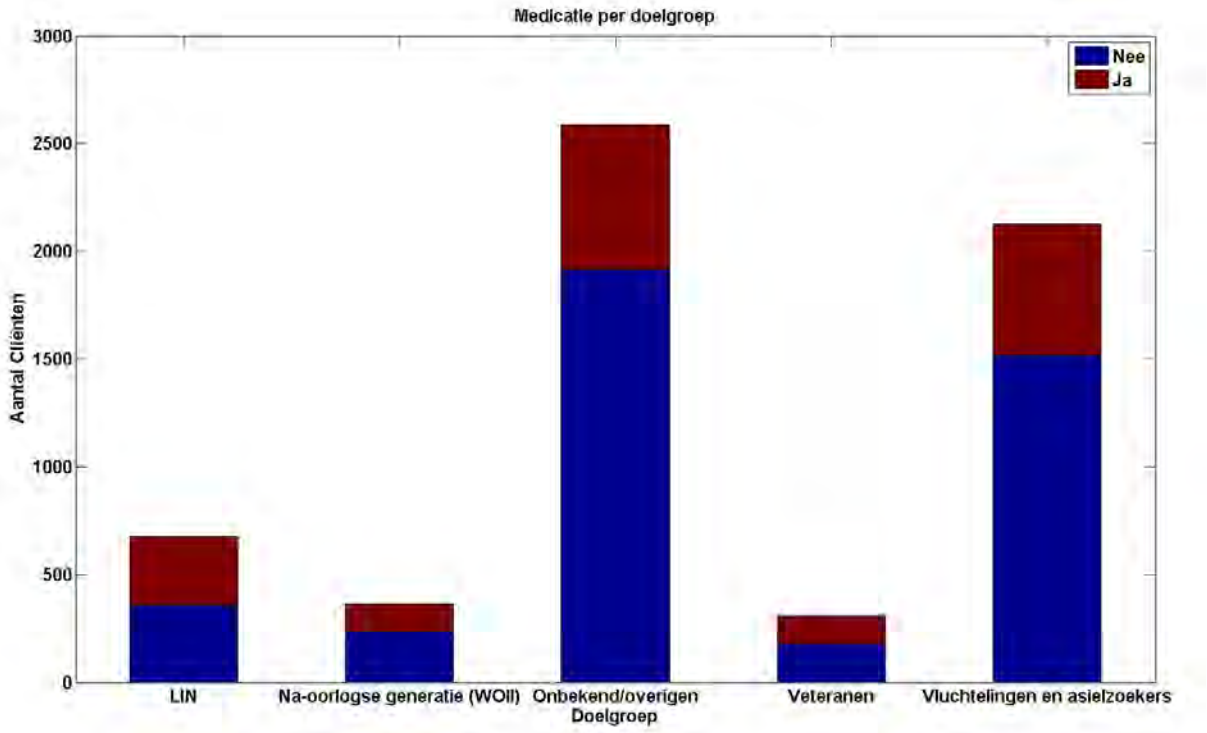
Figuur 11 *Doelgroep* tegenover *In_Compass*. Het rode deel geeft aan welk deel van de cliënten binnen de Compass data aanwezig zijn, het blauwe deel geeft aan van welk deel geen meting beschikbaar is.

Voor het geslacht kan eenzelfde figuur worden gemaakt. Deze is weergegeven in Figuur 12. Er is direct te zien dat de groep *Veteranen* vrijwel exclusief uit mannen bestaat, dit is zoals verwacht omdat dit om beroepsmilitairen gaat. Er is ook te zien dat de groep *Naoorlogse Generatie* voor iets meer dan de helft uit vrouwen bestaat. Bij de *LIN* groep zijn mannen weer oververtegenwoordigd. Bij de groepen *Vluchtelingen en Asielzoekers* en *Onbekend/Overige* is een zelfde verdeling van zo een 2/3 mannen, 1/3 vrouwen te zien.



Figuur 12 *Doelgroep* tegenover *geslacht*. De rode en blauwe delen geven respectievelijk mannelijke en vrouwelijke cliënten weer

Een ander interessant verband is de relatie tussen de doelgroep en het gebruik van medicatie. Dit verband is te zien in Figuur 13 *Medicatie* tegenover *doelgroep*. Wat direct opvalt, is dat bij de *onbekende/overgingen* en *vluchtelingen en asielzoekers* ongeveer een kwart van de cliënten medicatie gebruikt terwijl bij de andere doelgroepen er relatief vaker medicatie wordt gebruikt. Bij de *Naoorlogse generatie* gebruikt een derde van de cliënten medicatie en bij de *LIN* en *veteranen* doelgroepen is dit bijna de helft.



Figuur 13 *Medicatie* tegenover *doelgroep*

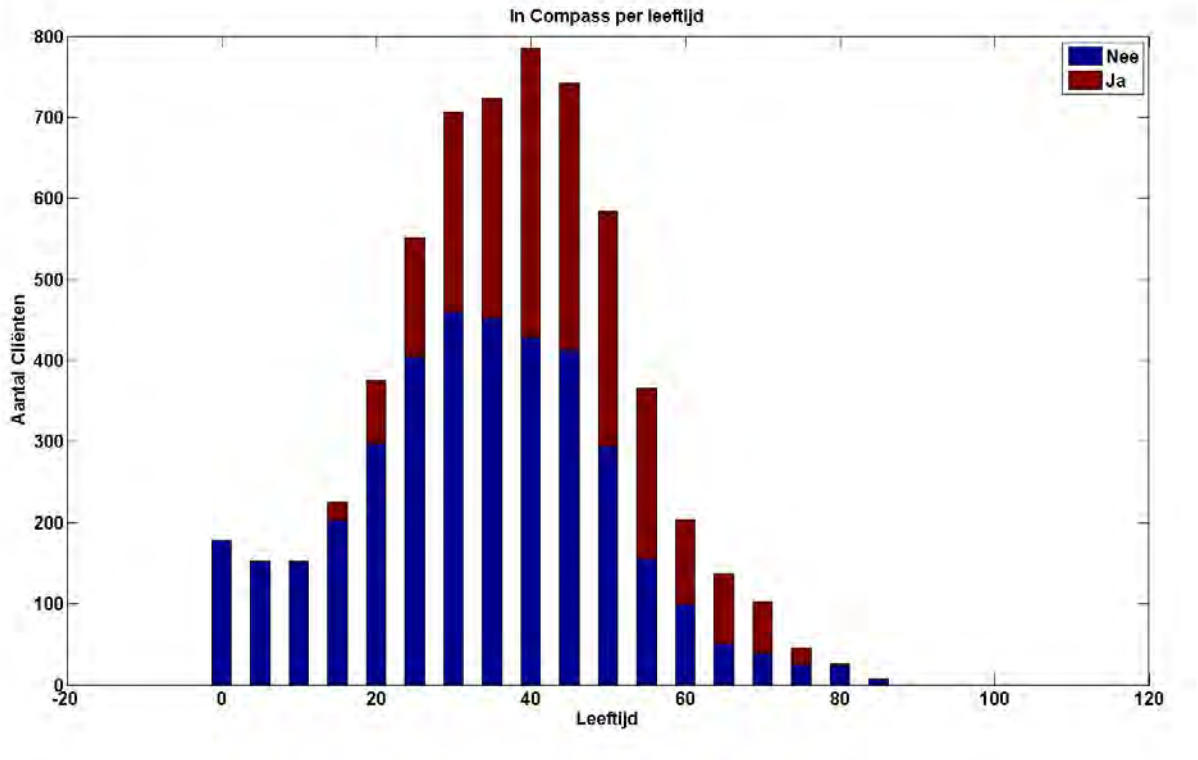
5.2.3.2 Leeftijd

De leeftijd variabele geeft de leeftijd van de cliënt ten tijde van inschrijven. De leeftijden van de cliënten liggen tussen de 0 en 109 jaar met een gemiddelde leeftijd van 38 jaar en een standaard deviatie van 16 jaar. Er zijn 32 cliënten in de dataset aanwezig met een leeftijd van 0 jaar. Bij deze cliënten is geen geboortedatum ingevoerd of in een enkel geval is bij de geboortedatum de inschrijfdatum ingevoerd.

De leeftijd van 109 jaar moet ook om een foutieve geboortedatum gaan omdat het onwaarschijnlijk is dat een hoogbejaarde op deze leeftijd nog aan een psychisch behandel traject begint.

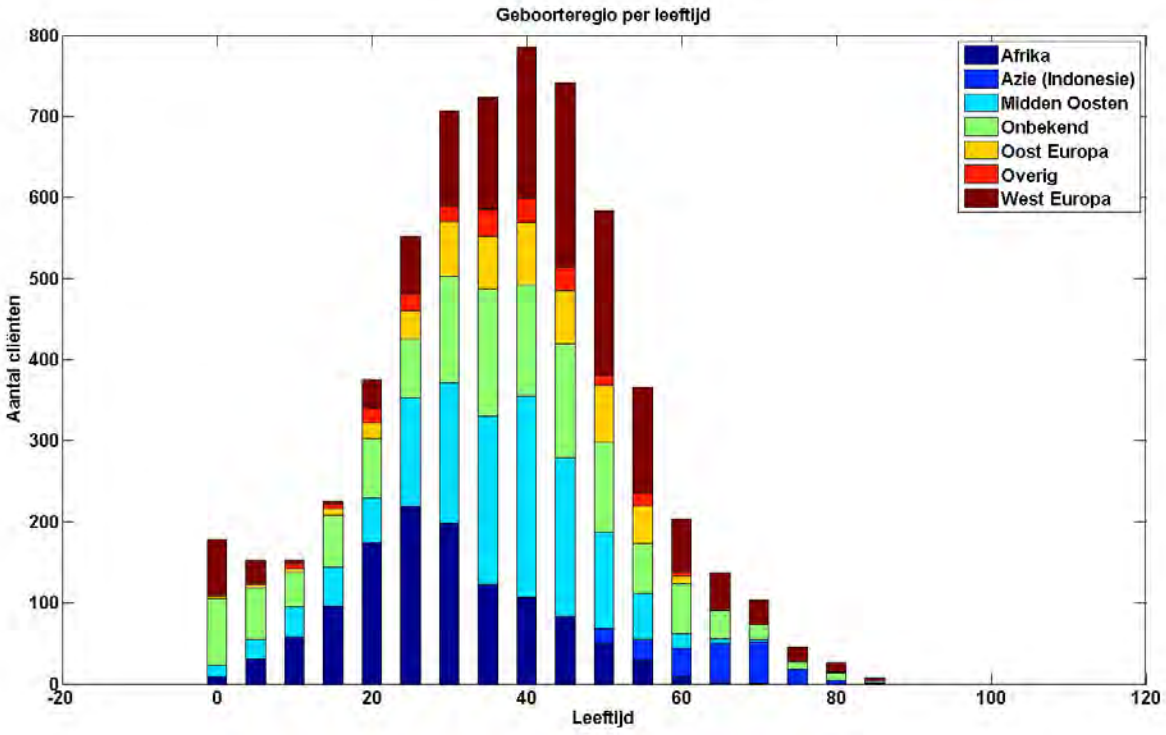
In Figuur 14 is de leeftijdsopbouw van de cliënten van de stichting in combinatie met het al dan niet beschikbaar zijn van Compass data geplot. Er is direct te zien dat van cliënten onder de 20 jaar vaak geen Compass gegevens beschikbaar zijn. Dit is echter te verwachten doordat er voor de Compass onderzoeken een leeftijdsgrens wordt gehanteerd van minimaal 18 jaar. Ook de groep senioren van 80 jaar en ouder vallen buiten de leeftijdsgrens van de Compass onderzoeken.

Binnen de leeftijden die wel in de Compass data aanwezig zijn lijkt het er uit de figuur op dat ouderen beter vertegenwoordigt zijn. Zo is te zien dat meer dan de helft van de cliënten binnen de leeftijd 55-60 in de Compass dataset aanwezig is.



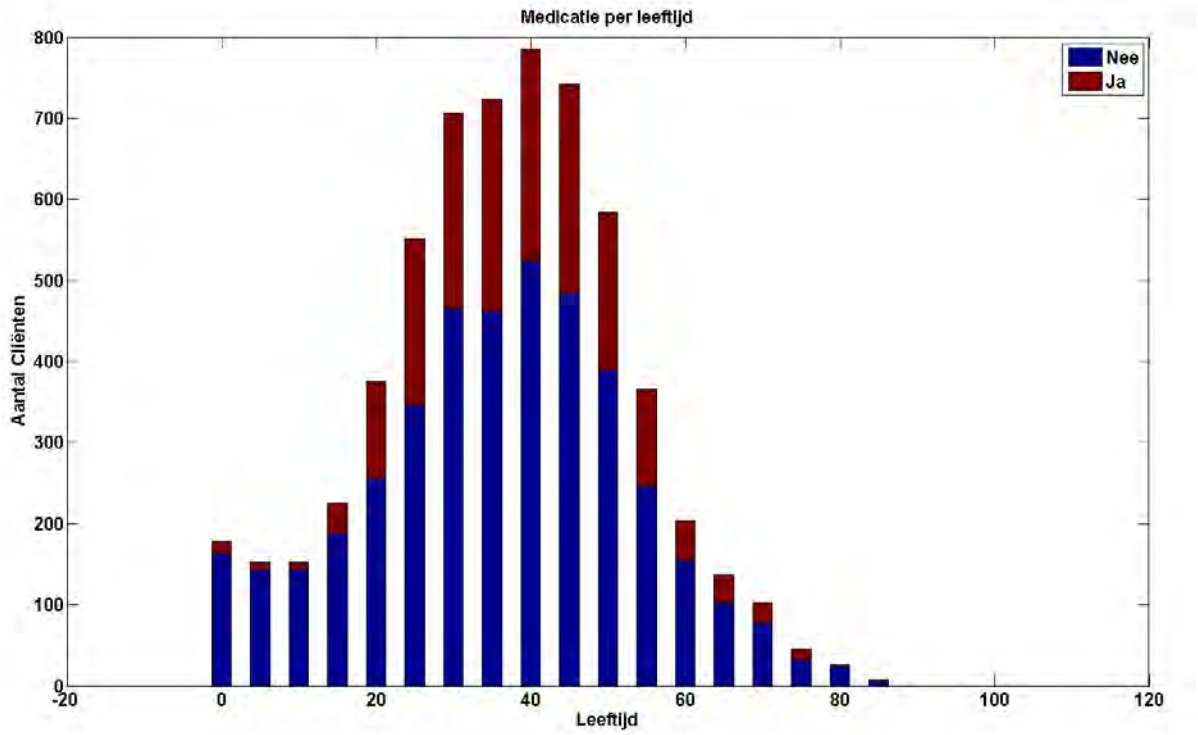
Figuur 14 *Leeftijd* tegenover *In_Compass*. Het rode deel geeft cliënten aan waarvan Compass data beschikbaar is, het blauwe deel geeft cliënten weer waarvoor dit niet het geval is.

Wanneer de *Leeftijd* wordt bekeken in relatie met *Geboorteregio* zijn ook een aantal duidelijk patronen zichtbaar zoals te zien is in Figuur 15. Uit dit figuur is op te maken dat Afrikanen over het algemeen een jongere leeftijd hebben. De Indonesiërs zijn daarentegen vaker ouder, dit is te verwachten doordat deze groep vrijwel exclusief bestaat uit Tweede Wereldoorlog slachtoffers en tweede generatie slachtoffers. Oost-Europeanen zijn weer goed vertegenwoordig tussen de 30 en 55 jaar. Cliënten uit het Midden Oosten zijn doorgaans jonger dan 55 jaar. Cliënten van West Europese afkomst lijken weinig voor te komen tussen de 15 en 20 jaar.



Figuur 15 *Leeftijd* in relatie tot *Geboorteregio*

Wanneer de leeftijd wordt uitgezet tegenover het gebruik van medicatie resulteert dit in Figuur 16. Er valt echter weinig opvallends te zien. Zoals verwacht wordt er weinig medicatie gebruikt door jongeren onder de 20 jaar. Voor de overige leeftijdsklassen lijkt het medicatie gebruik een zelfde patroon te vertonen.

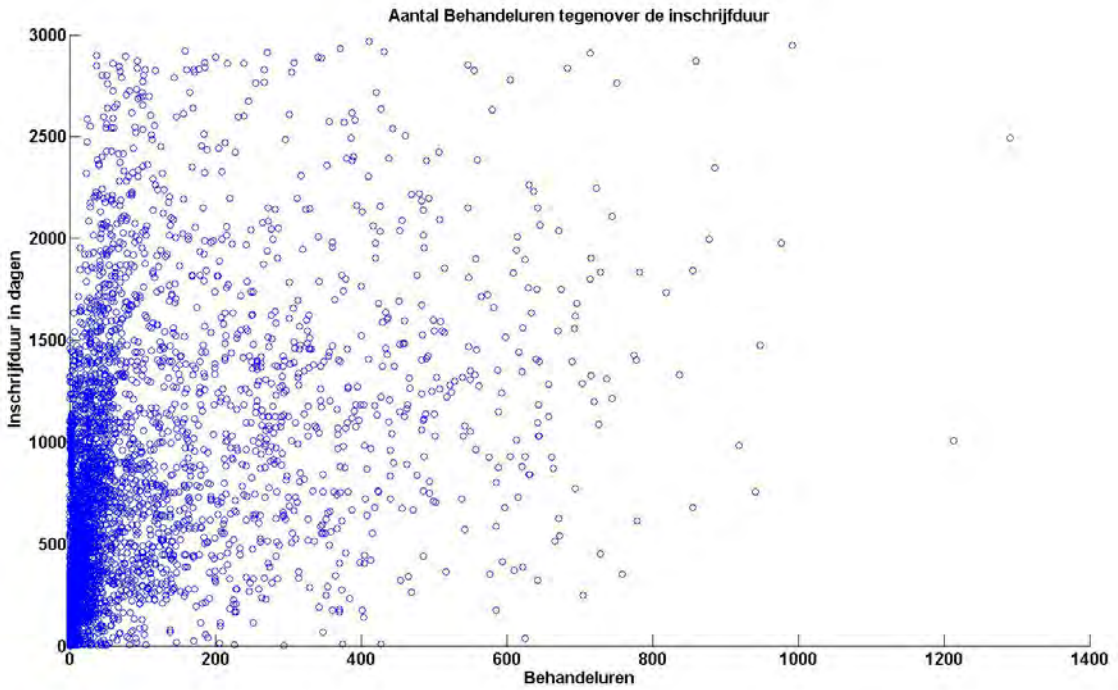


Figuur 16 Medicatie per leeftijd

5.2.3.3 Behandelduur

Bij de probleemanalyse is reeds naar voren gekomen dat de stichting geïnteresseerd is in efficiëntie. Een belangrijke factor hierbij is de behandelduur.

Een vergelijking van de behandelduur met het aantal behandeluren levert de plot in Figuur 17 op. Zoals verwacht correspondeert een korte inschrijfduur sterk met een laag aantal behandeluren. In de afbeelding is dit duidelijk te zien aan de sterke concentratie punten rond de oorsprong van de plot. Naar mate het aantal behandeluren en de inschrijfduur stijgt, wordt de relatie echter steeds minder duidelijk.



Figuur 17 Behandeluren tegenover inschrijfduur in dagen

Een statistische methode om de relatie tussen de twee variabelen te kwantificeren is door de zogeheten Pearson correlatie coëfficiënt te berekenen [16].

De coëfficiënt tussen twee variabelen X en Y, met steekproefgemiddelden $\bar{X}$ en $\bar{Y}$ en standaarddeviaties S(X) en S(Y) wordt berekend met

$$\rho = \frac{\sum_i (X_i - \bar{X})(Y_i - \bar{Y})/(N-1)}{S(X)S(Y)}$$

Deze formule levert een getal tussen de -1 en 1 op waarbij -1 een perfect negatief lineaire relatie weergeeft, 0 duidt op geen lineair verband en 1 duidt op perfect positief lineair verband.

Toegepast op de behandelduren en inschrijfduur geeft deze formule een score van 0.48 met een afgeronde significantie van 0. Deze significantie betekend dat de kans dat de score verkregen wordt terwijl er geen verband is tussen de variabelen. verwaarloosbaar klein is. Geconcludeerd kan worden dat er een significant matig verband bestaat tussen het aantal behandeluren en de inschrijfduur

5.3 User Afspraakgeschiedenis

Naast de set met cliëntgegevens is er ook een set met de volledige afspraakhistorie uit User gehaald. Aan de hand van deze set wordt het cliëntenstroom proces nader geanalyseerd met behulp van de sequence mining techniek die beschreven is in 4.4. De sequence analyse wordt in het volgende hoofdstuk behandeld.

5.3.1 Data Beschrijving

De afspraakgeschiedenis bestaat uit de gehele agenda sinds 2005. De data betreft 890962 afspraken van 7561 cliënten. Binnen de agenda zijn 131 verschillende afspraak types mogelijk.

De opbouw van de data is eenvoudig, in de eerste kolom staat het cliëntnummer gevolgd door het volgnummer van de afspraak en ten slotte het afspraak type.

5.3.2 Exploratieve analyse

Doordat de afspraakgeschiedenis eenvoudig is van opbouw is de exploratieve analyse vrij beperkt. Verdere analyses vereisen als snel complexere technieken die niet bij een exploratieve analyse passen. In overleg met een domein expert is besloten om de data te filteren op zogeheten directe activiteiten. Dit zijn activiteiten waarbij een cliënt aanwezig dient te zijn. Naast directe activiteiten zijn er ook indirecte activiteiten zoals dossierstudies of intern overleg.

Allereerst is er gekeken naar de meest populaire afspraken. De top 10 populairste zijn weergegeven in

0,784	**Intakegesprek**
0,605	Noshow
0,576	Begeleidingscontact
0,443	Compass
0,391	Psychiatrisch onderzoek
0,345	Cliënt gerichte therapie
0,341	Adviesgesprek (nav Intake)
0,329	Medicatieconsult
0,258	Telefonisch beg, contact
0,22	Maatschappelijk werk consult

Tabel 2. De support is hierbij het percentage van de cliënten dat minimaal een afspraak van het betreffende type heeft gehad.

Support	**Activiteit**
0,784	Intakegesprek
0,605	Noshow
0,576	Begeleidingscontact
0,443	Compass
0,391	Psychiatrisch onderzoek
0,345	Cliënt gerichte therapie
0,341	Adviesgesprek (nav Intake)

0,329	Medicatieconsult
0,258	Telefonisch beg, contact
0,22	Maatschappelijk werk consult

Tabel 2 Meest populaire activiteiten

Dat het intakegesprek het meest populaire afspraaktype is, is geen verassing omdat die voor de meeste cliënten de eerste stap vormt van een behandeltraject bij de stichting. Het werkelijke aantal zal echter hoger liggen doordat de afspraakgeschiedenis ook cliënten bevat die voor 2005 bij de stichting waren ingeschreven. De afspraakgeschiedenis voor 2005 is echter helaas niet beschikbaar.

Voor de minst populaire afspraken is hetzelfde gedaan. De support van 0 betekent dat deze kleiner was dan 0.1% waardoor deze in de uitvoer is afgerond. De resultaten hiervan zijn te zien in Tabel 3.

Support	Activiteit
0	Afspraak cliënt
0	Behandelplanbespreking +cliënt
0	Pré intake
0	Regelen tolken
0	z, Beleidsoverleg
0,001	Fysiotherapie
0,001	Huisbezoek
0,001	Int patiëntbespr (MDO)
0,001	Kinderen 0-6
0,001	Ontspannen en bewegen

Tabel 3 Minst populaire activiteiten

6 Resultaten

Dit hoofdstuk is gestructureerd per onderzoeksvraag met uitzondering van de vraag met betrekking tot de data integriteit van de Compass data. Deze vraag is reeds in het voorgaande hoofdstuk behandeld.

Per vraag zijn een of meerdere machine learning technieken toegepast. De resultaten per techniek worden vergeleken en gecombineerd om zo volledig mogelijk antwoord te geven op de onderzoeksvraag.

Binnen het Crisp model uit hoofdstuk 4 valt dit hoofdstuk onder de modelleringstap.

De gehele User dataset zoals beschreven in 5.2.1 is bij het modelleren gebruikt. Naar aanleiding van resultaten bij het modeleren zijn variabelen uit de dataset verwijderd of toegevoegd. Dit is beschreven wanneer dit het geval is, in het andere geval is de volledige set gebruikt zoals beschreven in 5.2.

6.1 Compass data cliënten

De hoofdvraag met betrekking tot de Compass data was in hoofdstuk 3 geformuleerd als:

-Is de Compass data geschikt als basis voor externe rapportages?

Met deelvragen

-Hoe is het gesteld met de data integriteit van de beschikbare dataset?

-Zijn er cliëntgroepen onder of over vertegenwoordigd in de data?

In hoofdstuk 5 werd ingegaan op de eerste deelvraag. De tweede deelvraag wordt in dit hoofdstuk behandeld door met behulp van classificatietechnieken het verschil tussen cliënten in de Compass data en daarbuiten aan het licht te brengen. Bij het analyseren van de cliëntgroepen in 6.2 wordt tevens ingegaan op deze vraag.

Hiervoor wordt de variabele *IN_COMPASS* gebruikt. Bij de classificatie zal dit de doelvariabele zijn.

Voor deze onderzoeksvraag zijn de cliënten die bij PDC en Equator behandeld zijn uit de dataset verwijderd omdat deze groepen bij voorbaat uitgesloten zijn van deelname aan de Compass metingen. Dit betreft de records waar de variabele *LOCATIE_G* op “PDC” of “Equator Diemen” staat. De zo verkregen dataset bestaat uit 4415 records.

6.1.1 ZeroR

Zoals besproken in 4.2.4.1 is ZeroR een classificatietechniek die alle records in een dataset classificeert volgens de modale klasse.

Het ZeroR algoritme classificeert in dit geval alle instanties als 0 omdat de meeste cliënten niet in de Compass dataset aanwezig zijn. Dit resulteert in een score van 51%.

6.1.2 OneR

Zoals behandelt in 4.2.4.2 zoekt het OneR algoritme naar de geïsoleerde variabele die de doelvariabele het best beschrijft.

Deze techniek wees *INSCHRIJFDUUR* als bepalende factor aan. Dit resulteerde in een score van 65%. Uit het model is echter moeilijk af te lezen hoe de inschrijfduur bepalend is doordat de techniek de inschrijfduur in 134 intervallen opdeelde waarvan een deel is te zien in onderstaand fragment van de uitvoer.

INSCHRIJFDUUR:

< 106.5 -> 0
< 109.5 -> 1
< 168.2216 -> 0
< 170.5 -> 1
< 179.7216 -> 0
< 189.2216 -> 1
< 211.2216 -> 0
< 215.7216 -> 1

Het is mogelijk om het minimum aantal records per interval voor numerieke data als setting aan het algoritme mee te geven. Wanneer het aantal naar 500 wordt verhoogd wordt niet de inschrijfduur maar de variabele *UITN_COMPASS* geselecteerd. Dit is uiteraard te verwachten omdat deze variabele aangeeft of een cliënt uitgenodigd is voor een Compass meting. Wanneer deze variabele uit de dataset wordt verwijderd valt de keuze op *BEHANDEL_UREN*. Vreemd genoeg gaat de score in dit geval omhoog naar 70%. Bij verdere inspectie waarbij het algoritme stap voor stap in een debugger is uitgevoerd bleek dat bij cross validatie *INSCHRIJFDUUR* beter scoorde dan *BEHANDEL_UREN*. Ook na het verwijderen van de variabelen was bij de cross validatie dezelfde score te zien. Desondanks werd de hogere totaal score verkregen. Het lijkt er dus op dat er bij de cross validatie een bepaalde bias optreedt.

Gebruikmakend van de T-test zoals genoemd in 4.2.1.1 is gebleken dat de scores significant verschilde van de resultaten verkregen bij ZeroR.

BEHANDEL_UREN:

< 24.5 -> 0
>= 24.5 -> 1

Bij de exploratieve analyse was gebleken dat de inschrijfduur en het aantal behandeluren een correlatie met elkaar vertonen. Dat beide variabele hier terugkomen is dus geen verassing.

6.1.3 J48

Zoals reeds is besproken in 4.2.4.3 levert het J48 algoritme een beslisboom op. Binnen Weka kan de gebruiker opgeven hoe groot een vertakking binnen de boom minimaal moet zijn. Deze instelling kan gebruikt worden om te zorgen dat de boom niet onnodig veel vertakkingen laat zien waardoor het model beter leesbaar en interpreteerbaar is.

Bij het toepassen op het Compass cliënten probleem is het minimum aantal records in een eindnode op 100 gezet.

Wanneer de techniek wordt toegepast op de volledige dataset met *IN_COMPASS* als doelvariabele wordt een score van 80% verkregen. Hierbij is wederom een T-test gebruikt waarbij is uitgewezen dat deze score statistisch significant verschilde van het resultaat van OneR. Het resulterende model is te zien in Figuur 18.

```
J48 pruned tree
------------------

LEEFTIJD <= 20: 0 (712.0/30.0)
LEEFTIJD > 20
|   BEHANDEL_UREN <= 15
|   |   BEHANDEL_UREN <= 2: 0 (367.0/26.0)
|   |   BEHANDEL_UREN > 2
|   |   |   AANMELDKLACHT_G = Klachten nav een traumatische gebeurtenis
|   |   |   |   LEEFTIJD <= 54
|   |   |   |   |   NOSHOW_PL <= 5.882353: 0 (477.0/153.0)
|   |   |   |   |   NOSHOW_PL > 5.882353: 1 (181.0/75.0)
|   |   |   |   LEEFTIJD > 54: 1 (122.0/41.0)
|   |   |   AANMELDKLACHT_G = Klachten mbt relatie partner/gezin/familie: 1 (138.0/40.0)
|   |   |   AANMELDKLACHT_G = onbekend: 0 (291.0/33.0)
|   |   |   AANMELDKLACHT_G = Overig: 1 (29.0/14.0)
|   BEHANDEL_UREN > 15
|   |   BEHANDEL_UREN <= 187
|   |   |   AANMELDKLACHT_G = Klachten nav een traumatische gebeurtenis
|   |   |   |   NOSHOW_PL <= 0.359712
|   |   |   |   |   GAF_GLOBAAL <= 55
|   |   |   |   |   |   LEEFTIJD <= 43: 0 (169.0/64.0)
|   |   |   |   |   |   LEEFTIJD > 43: 1 (143.0/50.0)
|   |   |   |   |   GAF_GLOBAAL > 55: 1 (128.0/30.0)
|   |   |   |   NOSHOW_PL > 0.359712: 1 (574.0/105.0)
|   |   |   AANMELDKLACHT_G = Klachten mbt relatie partner/gezin/familie: 1 (244.0/26.0)
|   |   |   AANMELDKLACHT_G = onbekend
|   |   |   |   BEHANDEL_UREN <= 53: 0 (114.0/40.0)
|   |   |   |   BEHANDEL_UREN > 53: 1 (100.0/29.0)
|   |   |   AANMELDKLACHT_G = Overig: 1 (57.0/10.0)
|   |   BEHANDEL_UREN > 187: 1 (569.0/47.0)
```

Figuur 18 Resulterende beslisboom van het J48 algoritme

De factoren zijn hierbij de knooppunten in de boom en een waarde geeft een vertakking aan. In Figuur 18 is *Leeftijd* het bovenste knooppunt van de boom en de waardes “<=20” en “>20” zijn de eerste twee vertakkingen. De tak “<=20” eindigt gelijk in een eindnode. De 0 geeft aan dat alle cliënten in deze vertakking worden voorspeld als niet aanwezig in de Compass set. De aantallen tussen haakjes geven aan dat er 712 cliënten in deze groep zitten en er van 30 cliënten wel een meting is. Dit resultaat is niet verwonderlijk door de leeftijdsgrens van 18 jaar voor Compass metingen.

Onder de vertakking “>20” zijn twee vertakkingen te vinden totdat er weer een eindnode wordt bereikt. Elke verticale streep geeft steeds aan dat er een verdere vertakking wordt gemaakt. Door alle factoren bij de vertakking tot en met de eindnode te lezen kan de volgende regel worden afgeleid:

“Cliënten ouder dan 20 jaar die 2 behandeluren of minder hebben gehad zijn niet in de Compass data aanwezig ”. Te zien is dat er 367 van deze cliënten zijn waarbij er 26 wel in de Compass data aanwezig zijn.

In zijn geheel laat de boom zien dat met name de leeftijd, het aantal behandeluren, no-show percentage klinisch, de aanmeldklacht en de locatie een bepalende rol spelen. Het no-show percentage kan ook duiden op cliënten die vooral poliklinisch behandeld zijn doordat cliënten die geen poliklinische behandelingen hebben gehad per definitie een no-show percentage van 0 hebben.

Om dergelijke onderlinge verbanden te voorkomen is het algoritme nogmaals uitgevoerd waarbij de no-show percentages uit de dataset zijn verwijderd wegens samenhang met de setting. Ook is de variabele *behandeluren* uit de set verwijderd door samenhang met de inschrijfduur. Naast het verwijderen van deze variabele is stapsgewijs het aantal records onder een eindnode opgevoerd tot 300. Hogere aantallen verslechterde de score aanzienlijk.

Dit resulteert in de compactere boom van Figuur 19.

Het verwijderen van de variabelen en het opvoeren van het minimum aantal records levert een iets slechtere score op van 78%. Ook in dit geval wees een T-test uit dat deze score significant verschilde van het resultaat behaalt met OneR. Deze boom geeft echter een makkelijker te interpreteren beeld en uit de score kan worden opgemaakt dat de no-show percentages weinig invloed uitoefenen.

J48 pruned tree

LEEFTIJD <= 20: 0 (712.0/30.0)

LEEFTIJD > 20

| MEDICATIE <= 0

| | AANMELDKLACHT_G = Klachten nav een traumatische gebeurtenis

| | | SETTING_G = Polikliniek: 1 (374.0/159.0)

| | | SETTING_G = onbekend: 0 (82.0/17.0)

| | | SETTING_G = Dagkliniek: 1 (137.0/42.0)

| | | SETTING_G = Intake: 0 (623.0/226.0)

| | | SETTING_G = Kliniek: 0 (27.0/4.0)

| | AANMELDKLACHT_G = Klachten mbt relatie partner/gezin/familie: 1 (353.0/82.0)

| | AANMELDKLACHT_G = onbekend: 0 (540.0/70.0)

| | AANMELDKLACHT_G = Overig: 1 (60.0/23.0)

| MEDICATIE > 0: 1 (1507.0/308.0)

Figuur 19 Resulterende beslisboom van het J48 algoritme met gereduceerde dataset

Uit de boom zijn de volgende gegevens af te lezen:

-Cliënten jonger dan 20 jaar zijn doorgaans niet in de Compass data te vinden. Van de 712 van deze cliënten is er bij 30 wel een meting afgenomen. Dit is te verwachten door de leeftijdsgrens van 18 jaar die door de stichting wordt gehanteerd voor het afnemen van een Compass meting.

-Cliënten ouder dan 20 die geen medicatie gebruiken en behandeld worden naar aanleiding van een traumatische gebeurtenis zijn afhankelijk van de behandelvorm wel of niet in de Compass data aanwezig.

-Poliklinisch behandelde cliënten zijn voor het grootste deel in de data terug te vinden. 215 van de 347 cliënten in deze groep heeft een meting afgenomen.

-Van de 82 cliënten met onbekende behandelvorm is bij slechts 17 een meting afgenomen. Dit zijn waarschijnlijk cliënten waarvan de afspraakhistorie verloren is gegaan bij de conversie naar User. Compass was op dit moment nog niet in gebruik dus dit is een te verwachten resultaat.

-De dagklinisch behandelde cliënten zijn overwegend wel terug te vinden in de Compass data. Van de 137 cliënten hebben 95 een meting afgenomen.

-De cliënten die enkel een intake hebben gehad zijn slecht in de data vertegenwoordigd. Van slechts 226 van de 623 cliënten in deze groep is een meting beschikbaar.

-Een klein deel van de cliënten in deze groep is klinisch behandeld. Het gaat om 27 cliënten waarbij er van 4 een Compass meting is afgenomen.

-Cliënten ouder dan 20 jaar die geen medicatie gebruiken zijn verder op basis van de aanmeldklacht terug te vinden in de Compass data

-Cliënten die klachten hebben met betrekking tot partner/gezien/familie zijn goed in de data vertegenwoordigd. Van de 353 cliënten is bij 271 een meting afgenomen.

-Cliënten met een onbekende aanmeldklacht zijn slecht in de data vertegenwoordigd. Van de 540 cliënten is bij slechts 70 een meting afgenomen. Waarschijnlijk wegens de eerder genoemde conversie.

-De cliënten met een overige aanmeldklacht heeft iets meer dan de helft een meting afgenomen. Van 37 van de 60 is een meting afgenomen.

-Cliënten die medicatie gebruiken en ouder zijn dan 20 zijn goed in de Compass data vertegenwoordigd. Van de 1507 cliënten hebben er 1199 een vragenlijst ingevuld.

6.2 Cliëntgroepen

De hoofdvraag met betrekking tot de User data is in hoofdstuk 3 geformuleerd als:

-Welke inefficiënte patronen zijn er in de data te vinden?

Met de subvragen:

-Welke cliëntgroepen zijn er uit de data te halen?

-Welke groepen hebben een verhoogd risico op een no-show?

-Welke factoren zijn bepalend voor de behandelduur?

-Welke patronen met betrekking tot no-show en behandelduur zijn in de afspraakhistorie te vinden?

Deze onderzoeksvragen met uitzondering van de laatste, worden met behulp van clustering en classificatie technieken behandeld. De laatste subvraag wordt in 6.5 behandeld. Allereerst wordt clustering toegepast om verschillende cliëntgroepen uit de data te halen. Bij de clustering wordt tevens direct inzicht verkregen in de no-show percentages en andere factoren van de gevonden groepen.

Er moet echter voorzichtig worden gedaan met conclusies trekken op basis van de clustering resultaten. Het kan bijvoorbeeld zijn dat een cluster bestaat uit cliënten met een hoog no-show percentage terwijl soortgelijke cliënten met een laag no-show percentage over andere groepen zijn verdeeld. De cluster analyse is daarom voornamelijk een middel om een beter overzicht over de data te verkrijgen.

6.2.1 Expectation Maximization clustering

Zoals toegelicht in 4.5.4 is bij het onderzoek gebruik gemaakt van Expectation Maximization (EM) clustering. Bij het clusteren zijn alle attributen gebruikt met uitzondering van de *behandelintensiteit* omdat de informatie die deze variabele geeft reeds in de *behandelduur* en *behandeluren* aanwezig is.

In WEKA kan het beginpunt van het zoekproces dat het algoritme uitvoert worden beïnvloed door middel van een zogeheten seed value. Het EM algoritme is met meerdere seed waardes uitgevoerd en de resultaten zijn hierbij met elkaar vergeleken. Het aantal gevonden clusters varieerde hierbij tussen de 6 en 10. Bij elke seed werden echter veelal soortgelijke clusters gevonden.

Zo bleken onder andere de PDC, intake en naoorlogse generatie clusters veelal terug te komen. Deze groepen lijken dus duidelijke eigenschappen te vertonen waardoor ze voor het cluster algoritme telkens te onderscheiden zijn.

Er is gekozen voor de uitvoer met 10 clusters zodat er geen clusters worden gemist. Een aantal clusters is bij de onderstaande beschrijving samen genomen omdat er veel gelijkenis tussen de betreffende clusters aanwezig was.

De kwaliteit van de clusters is beoordeeld door de resultaten door te lopen met een domeinexpert. Hierbij bleek dat de gevonden clusters voor de stichting herkenbare groepen bevatte.

6.2.1.1 Cluster 1: Jonge vluchtelingen uit het Midden Oosten

Deze groep kenmerkt zich door een relatief jonge gemiddelde leeftijd van 28 jaar. Iets minder dan de helft van cliënten in deze cluster komt uit het Midden Oosten en van ongeveer een derde is de geboorteregio onbekend.

Ongeveer de helft van deze cliënten heeft een angststoornis, een derde heeft een overige diagnose en zo een 15% heeft een stemmingsstoornis. Ongeveer de helft van de cliënten heeft een combinatie van twee diagnoses op as 1.

Meer dan de helft heeft problemen binnen de primaire steungroep en zo een 30% heeft overige psychosociale problemen. Op de overige assen is doorgaans geen diagnose aanwezig.

Van deze groep zijn weinig Compass metingen beschikbaar; zo een 15%. De helft van de cliënten gebruikt medicatie.

Deze groep wordt vooral poliklinisch behandeld in Diemen. De gemiddelde inschrijfduur van deze groep ligt rond de 3 jaar met 73 behandeluren.

6.2.1.2 Cluster 2: Angstige Onbekende, Linners en Veteranen.

Dit cluster typeert zich door cliënten waarbij drie kwart van de cliënten een angststoornis heeft en waarbij de doelgroep *onbekend*, *Lin* of *Veteraan* is. De cliënten komen uit diverse regio's en hebben een gemiddelde leeftijd van 45 jaar.

Opvallend aan dit cluster is dat de inschrijfduur rond de 3 jaar ligt net als in cluster 1 terwijl het gemiddeld aantal behandeluren ruim twee keer zo hoog ligt met 198 uur. Deze groep lijkt dus een intensievere behandeling te krijgen. Wat verder opvalt, is dat binnen deze groep 43% van de gevallen een enkelvoudige of complexe diagnose op as3 heeft. De helft van de cliënten heeft problemen binnen de primaire steungroep of overige psychosociale problemen.

Met 80% wordt er binnen deze groep relatief vaak medicatie gebruikt. Poliklinisch of klinische behandelden vormen beide rond de 45% van de groep. De overige 10% is voornamelijk dag klinisch behandeld.

Het overgrote deel van de cliënten wordt in Oegstgeest behandeld.

Verder valt op dat deze groep goed vertegenwoordigd is in de Compass data, van 90% van de cliënten is een meting beschikbaar.

6.2.1.3 Cluster 3 en 5: PDC en intake

Clusters 3 en 5 typeren zich door een overgroot deel van de cliënten dat enkel een intake heeft gehad. Het verschil tussen beide clusters is de locatie. Cluster 3 betreft voornamelijk cliënten van locatie Oegstgeest en cluster 5 cliënten van PDC Diemen. De gemiddelde leeftijd ligt bij beide clusters rond de 43 jaar.

In cluster 3 ligt de inschrijfduur op een jaar met gemiddeld 6 behandeluren. In cluster 5 ligt de inschrijfduur op iets minder dan een jaar met 3 behandeluren.

Van beide clusters zijn van de cliënten vrijwel geen Compass metingen beschikbaar. Van de PDC cliënten is dit te verwachten omdat bij deze groep geen metingen worden afgenomen. Bij cluster 3 is dit enigszins verassend omdat bij de intake procedure regelmatig Compass metingen worden verricht. Het kan echter zijn dat het clustering algoritme voornamelijk intake cliënten waarvan geen meting beschikbaar is

in een cluster heeft geplaatst en de overige intake cliënten waarvan wel metingen zijn verspreid heeft over de overige clusters.

6.2.1.4 Cluster 4: Oudere vluchtelingen uit het Midden Oosten

Dit cluster lijkt op cluster 1 maar verschilt door een hogere leeftijd van 37 jaar en een laag aantal cliënten met onbekende doelgroep of afkomst. Op het gebied van diagnoses komen de clusters grotendeels overeen.

Verdere verschillen de groepen in behandeluren en locatie. Terwijl de inschrijfduur wederom rond de 3 jaar ligt, is het aantal behandeluren meer dan twee keer zo hoog met de 174 behandeluren. In tegenstelling tot cluster 1 worden de cliënten vaker in Oegstgeest behandeld.

Opvallend is ook dat de cliënten in dit cluster wel goed vertegenwoordigd zijn in de Compass data met maar liefst 70% in tegenstelling tot 15% in cluster 1.

6.2.1.5 Cluster 6: Naoorlogse generatie met stemmingsklachten

Dit cluster vormt voor de stichting een duidelijk herkenbare groep. Deze cliënten zijn in hun jeugd getraumatiseerd door de impact die het oorlogsverleden van hun ouders op hen heeft gemaakt.

De doelgroep variabele is voor meer dan de helft van de cliënten in dit cluster *Naoorlogse* generatie. Het overige deel staat op *onbekend*. Meer dan 80% van de cliënten komt uit West Europa en het overige deel uit Indonesië. Vrijwel alle cliënten worden in Oegstgeest behandeld. Iets meer dan de helft dagklinisch en het overige deel poliklinisch. In tegenstelling tot de overige clusters heeft deze groep met iets meer dan de helft van de gevallen vaker een stemmingsstoornis dan een angststoornis. Wat deze groep verder onderscheid is de aanmeldklacht, in plaats van een traumatische ervaring heeft ruim drie kwart van deze cliënten zich aangemeld wegens problemen met gezien en familie. Verder valt ook op dat iets meer dan de helft van de cliënten vrouw is. Dit komt overeen met de exploratieve analyse in 5.2.3.1.

De cliënten hebben een gemiddelde inschrijfduur van 3 jaar met 144 behandeluren. Iets minder dan de helft van de cliënten gebruikt medicatie.

Ook deze groep is goed vertegenwoordigd in de Compass data. Van vrijwel alle cliënten binnen het cluster is een meting beschikbaar.

6.2.1.6 Cluster 7 en 9: Jonge vluchtelingen uit het Midden Oosten en Afrika.

Deze clusters vertonen een grote gelijkenis.

Beide groepen hebben een relatief jonge leeftijd van rond 32 jaar. De cliënten zijn in cluster 7 ongeveer gelijk verdeeld over Oegstgeest, Diemen en Equator. En in cluster 9 zijn meer dan helft van de cliënten in behandeling bij Equator.

In cluster 7 heeft meer dan de helft van de cliënten enkel een intake procedure gevolgd en de rest is poliklinisch behandeld. De korte gemiddelde inschrijfduur van minder dan een jaar in combinatie met gemiddeld 6 behandeluren bevestigen dat de cliënten binnen deze groep vaak niet verder zijn gekomen dan een intake procedure.

In cluster 9 zijn de cliënten echter voornamelijk poliklinisch behandeld. De inschrijfduur ligt binnen dit cluster ook relatief laag met een gemiddelde duur van ongeveer een jaar. Het aantal behandeluren ligt rond de 60.

Verder valt op dat deze cliënten relatief dicht bij de stichting wonen met een gemiddelde afstand van 30 km in cluster 7 en 20 km in cluster 9. Dit in tegenstelling tot de gemiddelde afstand van de gehele populatie van 53 km.

In beide clusters heeft de helft van de cliënten een tolk nodig. Verder verschillen de groepen in het medicatie gebruik. In cluster 7 wordt vrijwel geen medicatie gebruikt terwijl in cluster 9 80% van de cliënten medicatie gebruikt.

Beide groepen zijn slecht vertegenwoordigd in de Compass data, van ongeveer 30% is een meting beschikbaar.

6.2.1.7 Cluster 8: Vluchtelingen met gezinnen uit het Midden Oosten.

Dit cluster bestaat voor de helft uit vluchtelingen en voor het overige deel uit een gelijk aantal Linners en onbekenden.

Een belangrijk onderscheid tussen dit cluster en de overige clusters is het aantal neven cliënten binnen deze groep. Gemiddeld ligt dit aantal op 1.7 binnen dit cluster wat relatief hoog is in contrast met het populatie gemiddelde van 0.18.

Dit betekend dat cliënten binnen dit cluster vaak in behandeling zijn samen met familieleden. In de praktijk zijn dit vaak gezinnen in zijn geheel of partners die samen in behandeling zijn.

Deze groep heeft voornamelijk een angststoornis met in de helft van de gevallen tevens een stemmingsstoornis.

De behandelduur ligt rond de twee jaar met gemiddeld 54 behandeluren. De behandelvorm is gelijk verdeeld over dagbehandeling, poliklinisch of alleen intake.

Wat deze groep verder typeert is een hoog poliklinisch no-show percentage van 32%. Voor de stichting is dit een bekend verschijnsel. Wanneer gezinnen in behandeling zijn komen vaak alle leden niet opdagen wanneer een lid verhinderd is.

Van iets meer dan de helft van de cliënten is een Compass meting afgenomen en iets minder dan de helft heeft een tolk nodig.

6.2.1.8 Cluster 10: Onverzekerde vluchtelingen en onbekende uit een onbekende regio

Dit cluster typeert zich door een relatief hoog aantal onverzekerden. Ongeveer de helft van de cliënten binnen de groep is onverzekerd en 15% heeft een RZA verzekering. Van de gehele populatie is slechts16% onverzekerd of RZA verzekerd.

Van ruim de helft van de cliënten is geen geboorteregio bekend. Een kwart komt uit het Midden Oosten en de rest komt uit de overige regio's.

De gemiddelde leeftijd ligt op 35 jaar. Er zijn vrijwel geen Compass metingen afgenomen. Ruim de helft van deze groep is in behandeling bij locatie Oegstgeest, ruim en kwart in Diemen en de rest is bij overige locaties behandeld.

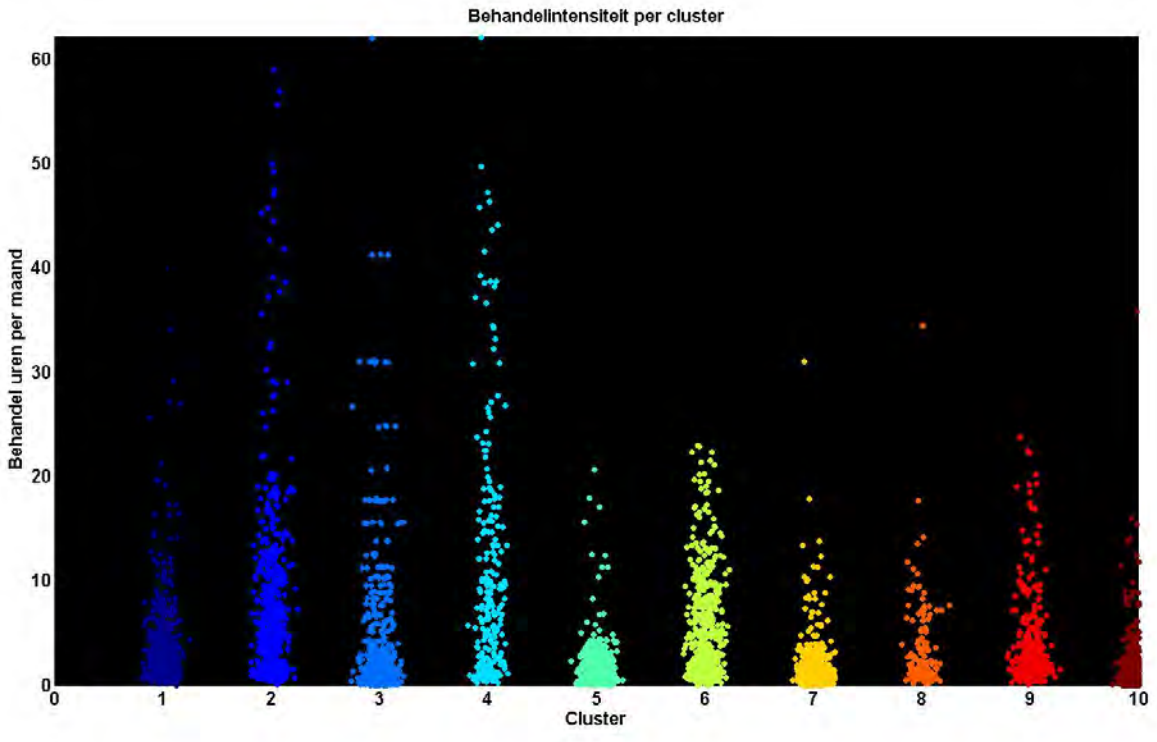
De inschrijfduur ligt bij deze groep rond iets minder dan een jaar met gemiddeld 11 behandeluren.

6.2.1.9 Visueel overzicht

Een manier om verschillen in clusters visueel inzichtelijk te maken is door scatterplots te maken waarin het cluster nummer tegenover een variabele wordt geplot. Deze plots geven inzicht in de mate van spreiding van een variabele binnen een cluster.

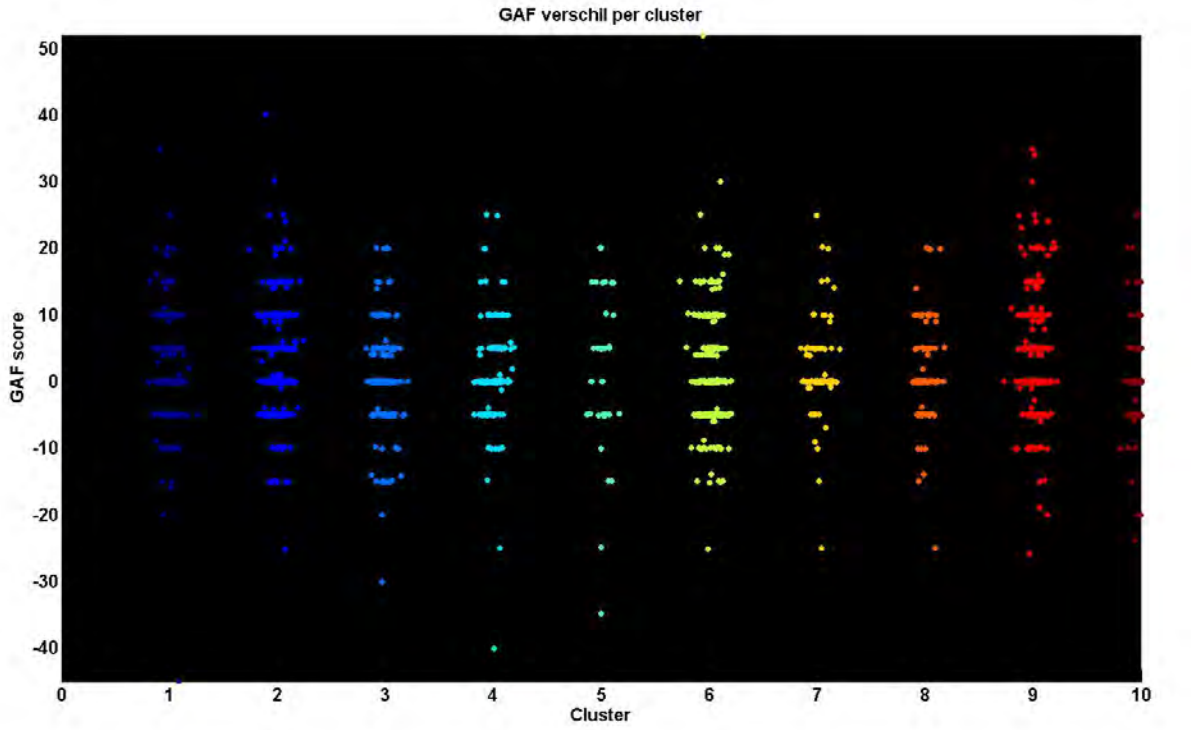
Op de horizontale as van de plots is het cluster nummer te vinden. Om het onderscheid tussen de clusters nog extra te verduidelijken zijn de punten in de plot voor elk cluster verschillend gekleurd. Op de verticale as is vervolgens een variabele te vinden die per cluster wordt vergeleken. In Figuur 20 betreft deze variabele de *behandelintensiteit*. Aan de punten in de plot is normaal verdeelde ruis toegevoegd zodat de punten uit elkaar gaan staan. Hierdoor is de mate van spreiding binnen de clusters inzichtelijk gemaakt.

Uit de cluster uitvoer was reeds op te maken dat er verschillen tussen de clusters bestaan in de verhouding behandeluren en inschrijfduur. Het verband tussen de twee variabele kan worden weergegeven door per cluster de *Behandelintensiteit* te plotten. De verschillen in behandelintensiteit per cluster zijn weergegeven in Figuur 20. Cluster 2 lijkt de meest intensief behandelde cliënten te bevatten, dit cluster bevat de angstige onbekende, Linners en veteranten. Bij cluster 4, de oudere vluchtelingen uit het Midden Oosten is een zelfde soort verdeling te zien. Clusters 5, 7, 8 en 10 vallen weer op door de lagere behandelintensiteit. Met name voor cluster 5 is dit te verwachten doordat dit cluster voornamelijk cliënten van PDC bevat die enkel langskomen voor een onderzoeksdag en advies.



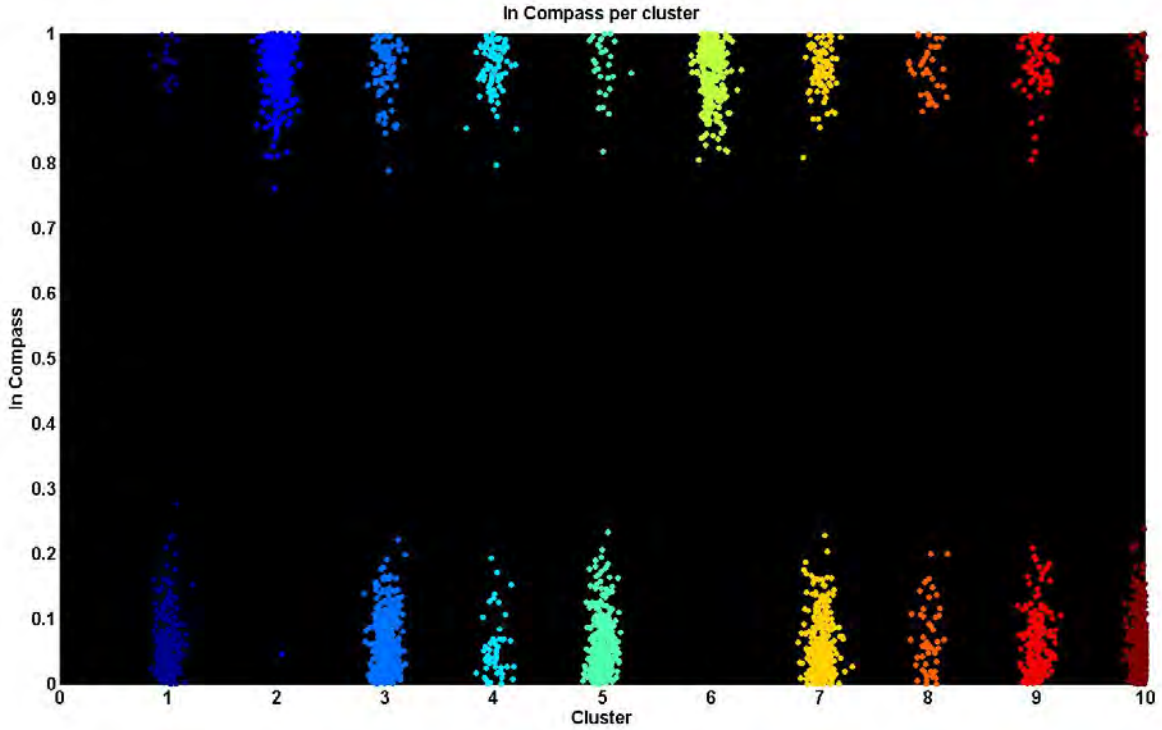
Figuur 20 Behandelintensiteit per cluster

Per cluster kan ook gekeken worden naar de vooruitgang die cliënten boeken door het verschil tussen twee GAF score metingen te bekijken wanneer deze bekend zijn. Voor slechts 2277 zijn twee of meer GAF scores bekend. Voor deze cliënten is het verschil tussen de eerste en laatste meting bepaald. Dit verschil is per cluster weergegeven in Figuur 21. Te zien is dat het grootste deel van de cliënten een positief verschil laat zien en dus vooruitgang boekt. Per cluster lijkt vooral de mate van spreiding te variëren.



Figuur 21 GAF verschil score per cluster

In Figuur 22 is per cluster weergegeven of de cliënten al dan niet in de Compass data aanwezig zijn.



Figuur 22 In Compass per cluster

In de plot zijn per cluster duidelijke verschillen te zien. Te zien is dat de cliënten uit cluster 1 slecht zijn vertegenwoordigd in de Compass data. Dit cluster betrof de jonge vluchtelingen uit het Midden Oosten. Dit is terug te vinden in de resultaten van 6.1.3 waarin te zien was dat jonge cliënten slecht in de data waren vertegenwoordigd. Verder is van iets minder dan de helft van de cliënten in deze groep de aanmeldklacht onbekend. In de eerdere analyse was reeds te zien dat dit samenhing met een slechte vertegenwoordiging in de Compass data.

De cliënten in het tweede cluster zijn duidelijk goed vertegenwoordigd in de Compass data. Dit cluster betrof voornamelijk cliënten met de doelgroep LIN, onbekend of veteraan. 78% van de cliënten binnen dit cluster gebruikt medicatie. In de eerdere analyse was te zien dat cliënten met medicatiegebruik veelal terug waren te vinden in de Compass dataset.

Bij het derde cluster zijn de cliënten overwegend niet in de Compass data aanwezig. Veel van de cliënten binnen dit cluster hebben enkel een intake gehad. Dit is tevens terug te vinden in de eerdere resultaten.

Bij het vierde cluster, vluchtelingen uit het Midden Oosten, is het onderscheid minder duidelijk.

Die cliënten van cluster 5 zijn duidelijk niet in de Compass data aanwezig. Dit is te verwachten doordat dit cluster voornamelijk bestaat uit PDC cliënten waarbij doorgaans geen Compass meting wordt afgenomen.

Bij cluster 6, de naoorlogse generatie lijkt van vrijwel alle cliënten een Compass meting te zijn afgenomen. In de resultaten van 6.1.3 is deze groep terug te vinden als de cliënten met klachten met betrekking tot relatie partner/gezin/familie.

Bij clusters 7 tot en met 9 zijn de verschillen weer minder duidelijk.

Ten slotte zijn de cliënten in cluster 10, de onverzekerden weer slechter in de data vertegenwoordigt. Bij meer dan de helft van de cliënten binnen dit cluster is de aanmeldklacht onbekend. In de eerdere resultaten was te zien dat dit samenhing het niet beschikbaar zijn van een Compassmeting.

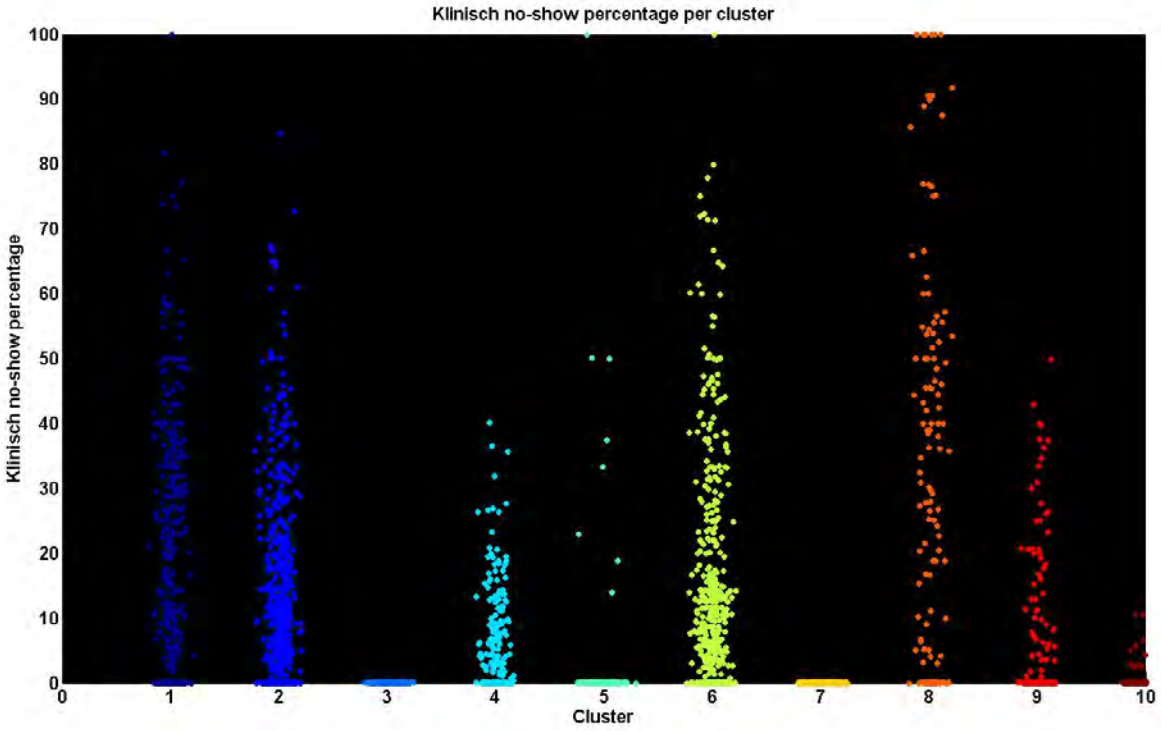
Voor de no-show percentages kunnen soortgelijke plots worden gemaakt. Deze geven een beter beeld van de verdeling van de no-show percentages binnen de groep dan een gemiddelde doordat de punten in de plot de spreiding goed laten zien.

De klinische respectievelijk poliklinische no-show percentages zijn per cluster geplot in

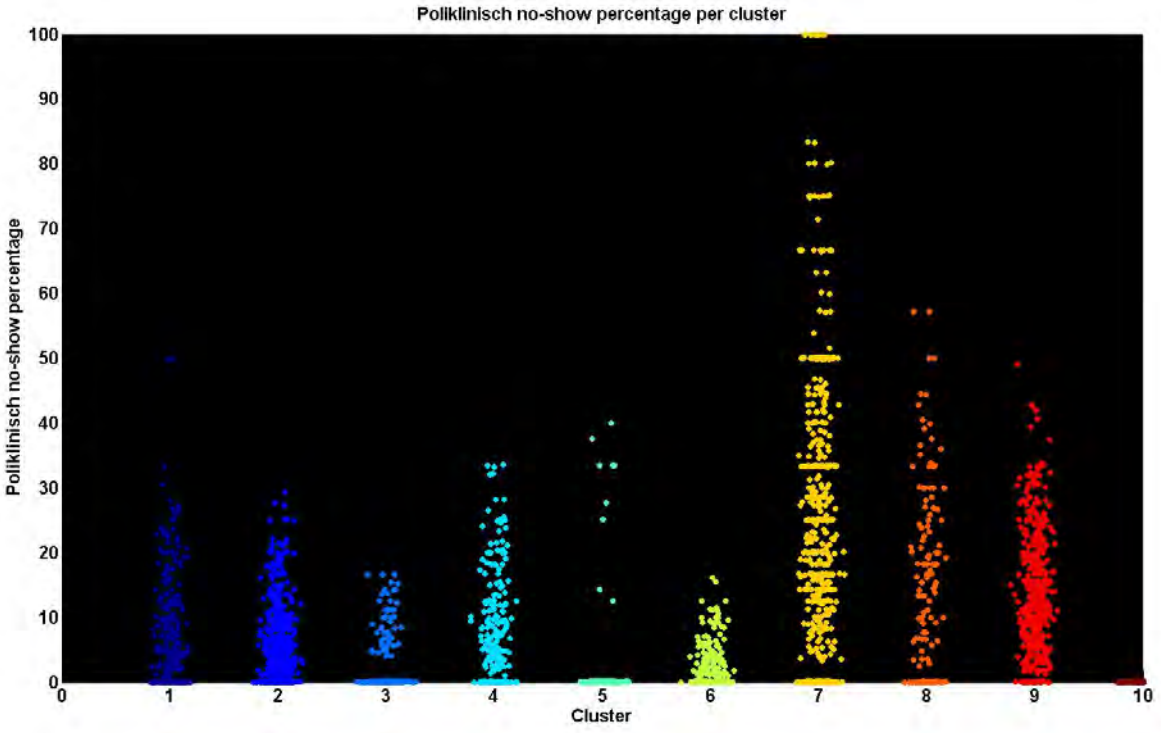
Figuur 23 en Figuur 24. In de eerste figuur valt direct op dat clusters 3, 5, 7 en 10 weinig tot geen cliënten bevatten met een hoog no-show percentage. Dat is niet verwonderlijk omdat deze groepen tevens weinig cliënten bevatten die klinisch worden behandeld met uitzondering van cluster 10, de onverzekerden. Het lijkt er dus op dat deze groep therapietrouw is. In clusters 1, 2 en 6 en 8 lopen de percentages uit een waarbij de meeste cliënten een percentage tussen de 0 en 50% hebben. De cliënten in cluster 4, de oudere vluchtelingen uit het Midden Oosten lijken vaker een laag klinisch no-show percentage te hebben. Deze groep bevat ook vaker cliënten met een hogere behandelintensiteit. Cluster 9, de jonge vluchtelingen uit Afrika en het Midden Oosten, bevat weinig klinische behandelde cliënten. Wat in de figuur te zien is dat de klinische no-show percentages lager liggen dan bij de andere groepen.

In de tweede figuur is te zien dat de poliklinische no-show percentages over het algemeen lager liggen dan de klinische. Opvallend is dat binnen cluster 6, de Naoorlogse generatie, de poliklinische no-show percentages lager liggen dan clusters 1 en 2 terwijl bij de klinische no-show percentages deze groep op elkaar lijken. Het lijkt er dus op dat deze groep extra therapietrouw is wat betreft poliklinische afspraken.

Een groep die opvalt door zijn hoge no-show percentages is cluster 7. Dit cluster bevat vooral jonge vluchtelingen uit Afrika en het Midden Oosten. Cluster 9 dat qua samenstelling vergelijkbaar is met cluster 7 heeft laat tevens een hoger no-show percentage zien. Ook bij cluster 8, de vluchtelingen met gezinnen, is een hoger no-show percentage te zien.



Figuur 23 Klinisch no-show percentage per cluster



Figuur 24 Poliklinisch no-show percentage per cluster

6.3 Behandelintensiteit

Voor de behandelintensiteit is de data verdeeld in twee groepen zodat er in beide groepen evenveel cliënten zitten. Dit leverde een groep op van cliënten die gemiddeld 0 tot 1,3 afspraak per maand hadden en een groep die 1,3 of meer afspraken per maand hadden.

Op deze wijze is er voldoende data aanwezig van beide groepen.

Het ZeroR algoritme geeft hierbij logischerwijs een score van 50%. Het OneR algoritme toonde aan dat de setting de meest bepalende factor was en gaf hierbij een score van 67% correct geclassificeerde cliënten aan. Dit leverde het model van Figuur 25 op. Een T-test wees uit dat de scores statistisch significant verschilde. Zoals verwacht wordt door de stichting zijn de laag intensief behandelde cliënten vooral poliklinisch behandeld of hebben enkel een intake gehad. De hoog intensief behandelde cliënten zijn vooral klinisch en dagklinisch behandeld.

De resultaten komen deels overeen met wat er in de cluster analyse was te zien. De clusters met hoog intensief behandelde cliënten bevatten veel klinische en dagklinische behandelde cliënten. Hoewel cluster 10 veel klinische en dagklinische behandelde cliënt bevat was hierbij geen hogere behandelintensiteit te zien.

```
SETTING_G:
        Polikliniek     -> '(-inf-1.368013]'
        onbekend        -> '(-inf-1.368013]'
        Dagkliniek      -> '(1.368013-inf)'
        Intake  -> '(-inf-1.368013]'
        Kliniek  -> '(1.368013-inf)'
(3410/5055 instances correct)
```

Figuur 25 Behandelintensiteit gemodelleerd met OneR

Andere technieken leverden een soortgelijk model op waarbij de score dicht in de buurt lag van OneR en de resultaten hiervan zijn daarom hier niet verder besproken.

6.4 No-show percentages

Om de no-show percentages nader te onderzoeken zijn twee subsets van de data gemaakt. Om de poliklinische no-show percentages te onderzoeken is een set gemaakt met cliënten die minstens een poliklinische afspraak hebben gehad. Voor het klinische no-show percentage is dit op een analoge wijze gedaan. Dit leverde sets met respectievelijk 4394 en 1413 cliënten op.

6.4.1 Poliklinisch No-show percentage

Voor het poliklinische no-show percentage is de subset opgedeeld in sets met twee ongeveer gelijke groepen. Dit leverde een groep op van cliënten met een no-show percentage van 0 tot 15% en een groep van hoger dan 15%. De groep met het lage no-show percentage bevatte 58% van de data en de tweede groep 42%.

Er is gekozen voor deze aanpak omdat de stichting niet geïnteresseerd is in het voorspellen van een precies no-show percentage. Het is vooral van belang om factoren aan te wijzen die bijdragen aan een hoog of juist laag no-show percentage. De tweedeling van de data zorgt er voor dat van beide groepen voldoende data beschikbaar is. De grenzen van de percentages komen hierbij tevens overeen met wat de stichting als hoog of laag no-show percentage beschouwd.

Het voordeel van het reduceren van de numerieke no-show variabele tot een binaire variabele is dat het resultaat makkelijker te beoordelen is. Een cliënt is in dit geval wel of niet in de juiste klasse geclassificeerd. Er is tevens ook een bredere mogelijkheid aan algoritmes toepasbaar. Bij een numerieke doelvariabele is slechts een beperkt aantal technieken toepasbaar.

Zoals verwacht geeft ZeroR een score van 58%. OneR gaf aan dat de inschrijfduur de meest bepalende factor was met een iets betere score van 61%. Een gepaarde T-test toonde hierbij aan dat het verschil statistisch significant was. Een J48 decision tree score aanzienlijk beter met 73%. Deze boom bestond echter uit een groot aantal vertakking waarbij de eindnodes vaak een klein aantal cliënten bevatte.

Het aantal cliënten in de eindnodes is hierdoor stapsgewijs verhoogd. Bij 100 cliënten per eindnode werd een score van 70% verkregen. Dit aantal kon verhoogd worden tot 400 waarbij een score van 69% werd verkregen en de zeer compacte boom van Figuur 26. Bij het opstellen van de beslisboom zijn *behandeluren* en *inschrijfduur* uit de set verwijderd omdat de informatie van deze variabelen in *behandelintensiteit* zit verwerkt. Een T-test waarbij de J48 uitvoer werd vergeleken met OneR toonde hierbij aan dat het verschil statistisch significant was.

De boom laat twee combinaties van factoren zien die een indicatie vormen voor een hoog no-show percentage.

Allereerst cliënten die geen tolk nodig hebben, medicatie gebruiken en minder dan gemiddeld 5.6 afspraken per maand hebben. 635 van de 1004 cliënten die aan deze factoren voldoen vallen in de groep met een no-show percentage van 15% of hoger. In het licht van de cluster analyse uit het vorige hoofdstuk valt deze groep onder cluster 9. Bij de cluster analyse was ook te zien dat de meeste cliënten binnen dit cluster een poliklinisch no-show percentage hebben van 15% of hoger.

De tweede groep cliënten die in de hoge no-show groep valt zijn de cliënten die een tolk nodig hebben. 476 van de 707 cliënten in deze groep vertoont een hoog poliklinisch no-show percentage. Deze cliënten zijn voornamelijk te vinden in clusters

7, 8 en 9. In Figuur 24 in 6.2.1.9 is duidelijk te zien dat deze groepen veel cliënten bevatten met een hoog no-show percentage.

J48 pruned tree

TOLK <= 0

| MEDICATIE <= 0: '(-inf-0.157729]' (2209.0/567.0)

| MEDICATIE > 0

| | BEHANDEL_INTENSITEIT <= 5.636585: '(0.157729-inf)' (1004.0/369.0)

| | BEHANDEL_INTENSITEIT > 5.636585: '(-inf-0.157729]' (474.0/154.0)

TOLK > 0: '(0.157729-inf)' (707.0/231.0)

Figuur 26 Decision tree verkregen met J48 voor het poliklinisch no-show percentage

6.4.2 Klinisch No-show percentage

Voor het klinisch no-show percentage is de subset wederom opgedeeld in twee gelijke groepen om dezelfde redenen genoemd in de vorige paragraaf. Dit leverde een groep met een percentage tussen de 0 en 11.5% en een groep met een percentage van meer dan 11.5% op. In dit geval bevatte beide groepen 50% van de cliënten.

Logischerwijs laat ZeroR een score van 50% zien. OneR gaf hier de behandelintensiteit aan als meest bepalende factor. Er werd een score van 63% verkregen.

J48 deed het iets beter met een score van 69%. Hierbij is een minimum aantal records per vertakking van 200 gehanteerd. Dit aantal is verkregen door net als bij het poliklinische no-show percentage het aantal stapsgewijs te verhogen. De boom is te zien in Figuur 27.

Bij het uitvoeren van een T-test bleek dat de verschillen van zowel OneR als J48 in vergelijking met ZeroR statistisch significant waren. Het verschil tussen OneR en J48 was echter niet statistisch significant. Aan de beslisboom te zien is dit niet verwonderlijk. De boom wijst net als OneR de behandelintensiteit als meest bepalende factor aan. De verdere vertakking in de setting voegt hier weinig aan toe.

De boom classificeert alle cliënten als cliënten met een hoog no-show met uitzondering van de cliënten met minder dan gemiddeld 8.3 afspraken per maand die voornamelijk klinisch behandeld zijn en cliënten die meer dan gemiddeld 8.3 afspraken per maand hebben. Deze groep vormt met 94 cliënten waarvan er 27 alsnog in de hoge no-show groep vallen een kleine groep.

De relatie tussen de behandelintensiteit en no-show is echter niet duidelijk in de clusters terug te vinden.

J48 pruned tree

BEHANDEL_INTENSITEIT <= 8.311084

| SETTING_G = Dagkliniek: '(11.56982-inf)' (562.0/164.0)

| SETTING_G = Polikliniek: '(11.56982-inf)' (227.0/103.0)

| SETTING_G = Kliniek: '(-inf-11.56982]' (94.0/27.0)

| SETTING_G = Intake: '(11.56982-inf)' (44.0/11.0)

BEHANDEL_INTENSITEIT > 8.311084: '(-inf-11.56982]' (486.0/124.0)

Figuur 27 Decision tree verkregen met J48 voor het klinisch no-show percentage

6.5 Afspraakhistorie

De onderzoeksvraag met betrekking tot de afspraakhistorie is in 3.2 geformuleerd als

> *-Welke patronen met betrekking tot no-show en behandelfase zijn in de afspraakhistorie te vinden?*

Met behulp van sequence mining zoals beschreven in 4.4 wordt deze vraag beantwoord. Bij het onderzoek is er ook nog gezocht naar verschillen in afspraakpatronen tussen cliënten die een verbetering in GAF score laten zien tegenover cliënten die een verslechtering laten zien. Hierbij waren echter geen duidelijke verschillen aangetroffen. Deze vergelijking is daarom verder niet besproken.

6.5.1 Directe No-Show patronen

Bij het analyseren van directe no-show patronen is gezocht naar afspraaktypes die elkaar opvolgen en uiteindelijk leiden tot een no-show. Onderstaande tabel geeft de resultaten hiervan weer.

Regel	Support	Confidence
Groepstherapie -> Noshow	0,12	0,56
Begeleidingscontact -> Noshow	0,31	0,53
Noshow -> Noshow	0,3	0,49
Psychotherapie -> Noshow	0,19	0,43
Telefonisch beg, contact -> Noshow	0,1	0,39
Noshow, Noshow -> Noshow	0,11	0,38
Medicatieconsult -> Noshow	0,12	0,36
Intakegesprek -> Noshow	0,13	0,17
Compass -> Noshow	0,05	0,1

De interpretatie van de bovenste regel is als volgt: groepstherapie gevolgd door een no-show vindt bij 12% van de cliënten plaats. Van de cliënten die groepstherapie hebben vindt de no-show bij 56% plaats.

De support zegt dus iets over de frequentie en de confidence over de sterkte van het verband dat de sequence aangeeft.

Deze percentages zijn voornamelijk nuttig om de patronen onderling te vergelijken. Wanneer bijvoorbeeld gekeken wordt naar de zevende regel is te zien dat wederom 12% van de cliënten een medicatieconsult heeft gehad gevolgd door een no-show. Slechts 36% van de cliënten die medicatieconsults hebben vertonen dit patroon. Hieruit kan vervolgens worden opgemaakt dat het risico op een no-show hoger is na groepstherapie dan na een medicatieconsult.

Uit de tabel wordt ook meteen duidelijk dat het verband intakegesprek gevolgd door een no-show of Compass meting gevolgd door een no-show relatief klein is.

Verder is te zien dat een dubbele no-show bij 30% van de cliënten voorkomt en dat de helft van de cliënten die een no-show vertonen een dubbele no-show vertoont.

Voor de driedubbele no-show nemen deze percentages verder af naar 11% en 38% respectievelijk.

6.5.2 Indirecte no-show patronen

Bij indirecte no-show patronen is gekeken naar patronen waarbij “gaten” binnen de patronen zijn toegestaan. Deze gaten bestaan uit een willekeurig aantal afspraken van een arbitrair type. Onderstaande tabel geeft hiervan de resultaten weer.

Regel	Support	Confidence
Stabiliserende technieken -> Noshow	0,03	0,84
Medicatieconsult -> Noshow	0,27	0,81
Noshow -> Noshow	0,48	0,78
Activerend contact -> Noshow	0,08	0,74
Begeleidingscontact -> Noshow	0,43	0,74
Maatschappelijk werk consult -> Noshow	0,16	0,73
Non-verbaal -> Noshow	0,15	0,73
Groepstherapie -> Noshow	0,15	0,7
Telefonisch beg, contact -> Noshow	0,18	0,69
Psychiatrisch consult -> Noshow	0,08	0,68
Nazorg -> Noshow	0,08	0,66
Psychotherapie -> Noshow	0,3	0,66
Compass -> Noshow	0,31	0,61
Gezinsonderzoek -> Noshow	0,03	0,61
Intakegesprek -> Noshow	0,45	0,58
Psychiatrisch onderzoek -> Noshow	0,2	0,52
Adviesgesprek (nav Intake) -> Noshow	0,17	0,51
Psychodiagnostiek -> Noshow	0,14	0,51
Noshow, Noshow -> Noshow	0,38	0,80

De eerste regel geeft aan dat 3% van de cliënten stabiliserende technieken hebben gevolgd waarna er ooit een keer een no-show heeft plaatsgevonden. De confidence geeft aan dat 84% van de cliënten die stabiliserende technieken volgen ooit een keer een no-show vertonen. In 5.3.2 was te zien dat 60% van de cliënten een keer een no-show vertoont. Bovenstaande resultaten dienen voornamelijk vergeleken te worden met dit percentage. Bij de patronen met een confidence boven de 60% is dus sprake van een verhoogd risico.

De resultaten dienen echter zorgvuldig te worden geïnterpreteerd. Zo vinden stabiliserende technieken vooral plaats bij cliënten die langer in behandeling zijn. Hoe langer iemand in behandeling is, hoe groter de kans dat er ooit een no-show optreedt.

Verder is af te lezen dat 48% van de cliënten twee of meer no-shows vertoont. En 38% van de cliënten vertoont 3 no-shows. 74% van de cliënten waarbij een no-show optreedt, vertoont in de toekomst een of meerdere no-shows. 80% van de cliënten waarbij twee no-shows hebben plaatsgevonden laten in de toekomst een of meerdere no-shows zien. De kans op een no-show neemt dus toe met het aantal al vertoonde no-shows.

6.5.3 Behandelfase

Voor de behandelfase analyse is onderscheid gemaakt tussen 3 fasen:

-Het eerste behandeljaar

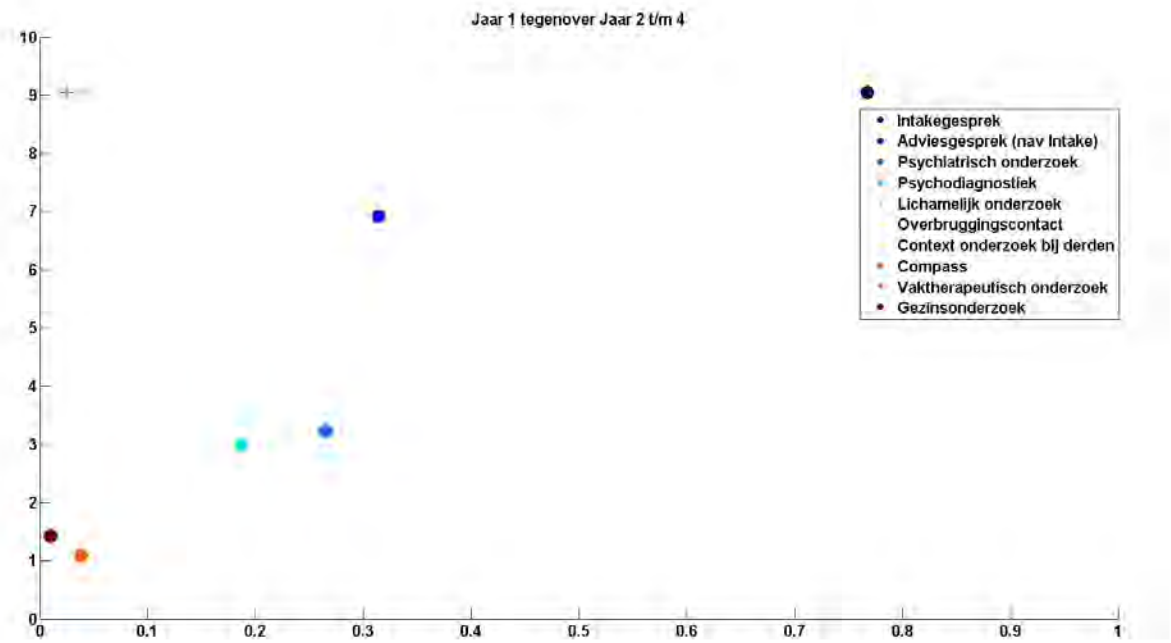
-Het tweede tot en met het vierde behandeljaar

-Het vijfde behandeljaar en verder

Per behandelfase is het verschil in het voorkomen van afspraaktypes geanalyseerd gebruikmakend van de techniek beschreven in 4.5.2. De resultaten zijn vervolgens visueel weergegeven.

In Figuur 28 is een weergave te zien van de verschillen tussen de eerste twee behandelfasen. Elk gekleurd punt staat voor een afspraaktype. De verticale as geeft de ratio tussen de supports in beide sets weer. Het intake gesprek staat ter hoogte van 9, wat betekend dat het intake gesprek 9 keer zo vaak voorkomt in jaar 1 als in jaren 2 tot en met 4. De horizontale as geeft het absolute procentuele verschil tussen de supports in beide sets weer. In jaar 1 in komt het intakegesprek bij 86% van de cliënten voor en in de tweede behandelfase bijna 9% waardoor het intake gesprek op de horizontale as rond de 77% staat in de plot. Voor het adviesgesprek geldt een soortgelijk verband met het verschil dat het verder links staat op de horizontale as wat betekend dat die afspraaktype procentueel voor minder cliënten voorkomt.

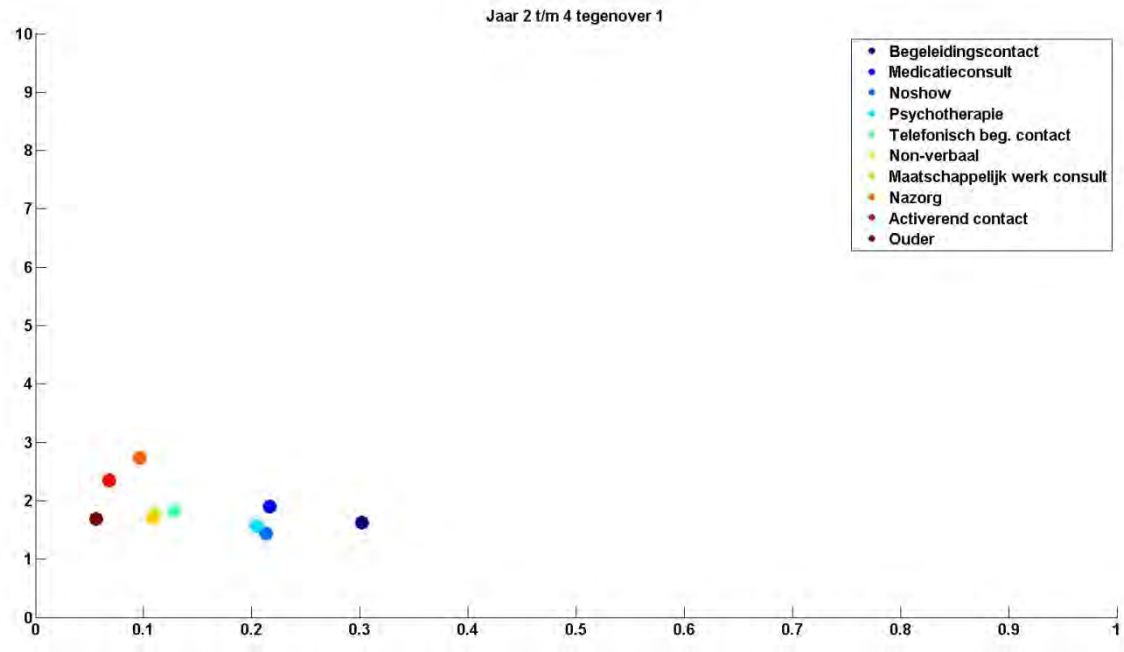
Voor het psychiatrisch onderzoek en psychodiagnostiek is te zien dat deze 3 keer zo vaak voorkomen in de eerste behandelfase als in de tweede. Dit is te verwachten omdat diagnostische activiteiten doorgaans vroeg in het behandeltraject plaatsvinden. Bij de Compassmeting en het gezinsonderzoek is het verschil tussen de twee behandelfase klein. De overige punten weergegeven met een asterisk komen enkel in de eerste behandelfase voor. Echter voor een procentueel klein aantal cliënten. Dit is te zien doordat deze punten op de horizontale as dicht bij de oorsprong liggen. De verticale ligging heeft voor deze punten geen betekenis.



Figuur 28 Verschil in afspraaktypes tussen jaar 1 en jaar 2 t/m 4

In Figuur 29 zijn de verschillen weergegeven tussen de tweede fase en de eerste fase. Er is te zien dat begeleidingscontacten, medicatieconsulten en psychotherapie twee keer zo vaak voorkomen in de tweede behandelfase. Dit is te verwachten omdat na het eerste jaar de diagnostisering gedaan is en er kan worden begonnen met gericht behandelen wat zich in deze afspraaktypes uit. Hetzelfde verschil geldt voor de vorige types die in de plot zijn te vinden, de verschillen zijn echter minder groot omdat het procentueel minder cliënten betreft.

Tevens is te zien dat no-shows bijna twee keer zo veel voorkomen in de tweede behandelfase.

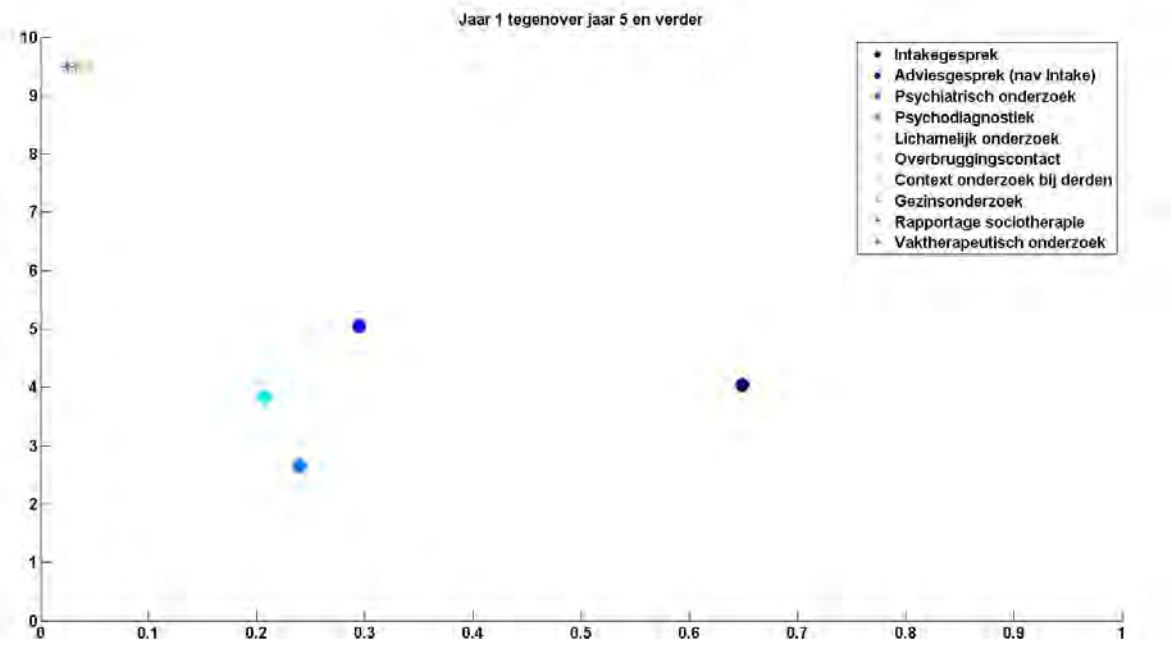


Figuur 29 Verschil in afspraaktypes tussen jaar 2 t/m 4 en jaar 1

In Figuur 30 is de vergelijking tussen de eerste en derde fase te zien. De punten links bovenin zijn alleen terug te vinden in de eerste behandelfase. Deze punten betreffen het lichamelijk onderzoek, overbruggingscontacten, context onderzoek bij derden, gezinsonderzoek, rapportage sociotherapie en vaktherapeutisch onderzoek. De ligging van deze punten op de horizontale as geeft echter aan dat het procentueel weinig cliënten betreft. Minder dan 5% van de cliënten heeft afspraken van dit type in het eerste behandeljaar.

De overige punten betreffen diagnostische en intake activiteiten en deze komen zoals verwacht vaker voor in de eerste behandelfase. Wat echter opvalt, is dat het intakegesprek slechts 4 keer zoveel voorkomt in de eerste behandelfase. Bij de vergelijking tussen de eerste en tweede behandelfase was een verschil van 9 keer zoveel te zien.

Dit wijst er op dat er een aanzienlijke hoeveelheid cliënten is die na hun vierde behandeljaar nog een intakegesprek hebben. Een extra check in de database van de stichting wees uit dat dit inderdaad het geval was. Het blijkt dat maar liefst 21% van de cliënten na hun vierde behandeljaar nog een intakegesprek heeft. De oorzaak hiervan is bij de stichting onbekend.



Figuur 30 Verschil in afspraaktypes 1 en jaar 5 en verder

Bij het vergelijken van behandelfase drie tegenover een was een vergelijkbaar verschil te zien als bij de vergelijking tussen behandelfase twee tegenover een met meer begeleiding en gerichte therapie. Deze vergelijking is daarom hier verder niet besproken.

Bij de vergelijking tussen de tweede en derde behandelfase waren geen grote verschillen te zien met uitzondering van het intakegesprek dat in de derde fase ruim twee keer zo veel voorkomt als in de tweede fase. De afwezigheid van verschillen is voor de stichting een verassend resultaat. De verwachting is dat naar mate iemand langer in behandeling is er minder gerichte therapie plaatsvindt en meer begeleiding. Dit was echter in de resultaten niet te zien.

7 Conclusie

7.1 Compass data

De hoofdvraag behorende bij de Compass dataset was in 3.2 geformuleerd als:

-Is de Compass data geschikt als basis voor externe rapportages?

Met eerste sub-vraag:

-Hoe is het gesteld met de data integriteit van de beschikbare dataset?

Bij de integriteitcontrole uit hoofdstuk 5.1.2 was gebleken dat de Compass data geen dubbele records bevatte. De datatypes van de velden waren op enkele records na correct. Bij een aantal vragen waren verschillen in antwoordschalen gedetecteerd zoals beschreven in 5.1.2.3. Bij een klein deel van de data bleken de cliëntgegevens in de data niet overeen te komen met de gegevens in de centrale database van de stichting.

Gesteld kan worden dat de data voor het grootste deel correct is. De verschillen in antwoordschalen dienen echter nog te worden onderzocht door een domeinexpert. In een aantal gevallen waren de verschillen zo groot dat er mogelijk sprake was van omgedraaide antwoordschalen of veranderde vragen waardoor delen van de data mogelijk inconsistent zijn.

Ondanks dat de cliëntgegevens in slechts een klein aantal gevallen niet correct waren is het sterk aan te raden om in de toekomst een koppeling te realiseren tussen het Compass systeem en de User database zodat foutieve cliëntgegevens niet mogelijk zijn.

De tweede sub-vraag bij de Compass data was:

-Zijn er cliëntgroepen onder of over vertegenwoordigd in de data?

Op deze vraag is ingegaan in 6.1 en 6.2. In 6.1 bleken de leeftijd, aanmeldklacht en medicatiegebruik de meest bepalende factoren te zijn. Cliënten jonger dan 20 jaar, en cliënten met een onbekende aanmeldklacht bleken slecht in de data te zijn vertegenwoordigd. Cliënten die medicatie gebruiken leken goed te zijn vertegenwoordigd binnen de Compass data.

Uit de cluster analyse bleek verder dat er een aantal groepen duidelijk onder of oververtegenwoordigd waren in de data. Zo bleken jonge vluchtelingen uit het Midden Oosten en onverzekerden slecht vertegenwoordigd in de Compass data. Cliënten van de doelgroep naoorlogse generatie, Linners en veteranen bleken echter zeer goed in de data te zijn vertegenwoordigd. Van het overgrote deel van de cliënten in deze groepen is een Compassmeting beschikbaar. De gevonden factoren in 6.1 kwamen overeen met de eigenschappen van de clusters in 6.2.

Er zijn een aantal mogelijkheden om met deze verschillen om te gaan. Er zijn diverse statistische technieken waarmee de data kan worden gefilterd zodat de data representatief is voor de gehele populatie. Een mogelijkheid is om gebruik te maken van stratified sampling [17]. Hierbij wordt de wordt de data verdeeld in groepen ook

wel strata genoemd. Uit elke groep wordt een zodanig aantal datapunten willekeurig geselecteerd zodat de verhouding van de aantallen records tussen groepen gelijk blijft.

In het geval van de Compass data kunnen bijvoorbeeld de gevonden clusters in 6.2 als strata worden gebruikt.

Een andere mogelijkheid is om het beleid met betrekking tot de Compassmetingen in de toekomst aan te passen. De ondervertegenwoordigde groepen zouden bijvoorbeeld extra kunnen worden gemotiveerd om een Compassmeting af te nemen. De meest eenvoudige en effectieve methode is echter om de Compass metingen verplicht onderdeel uit te laten maken van de intakeprocedure.

7.2 User data

7.2.1 Cliëntdata

Met betrekking tot de User data is in 3.2 de volgende hoofdvraag geformuleerd:

-Welke inefficiënte patronen zijn er in de data te vinden?

Met eerste sub-vraag:

-Welke cliëntgroepen zijn er uit de data te halen?

Met behulp van clustering technieken is hier in 6.2 antwoord opgegeven. Na samenvoegen van een aantal clusters waren er 8 verschillende groepen uit de data te halen. Veel van de gevonden typeringen van de gevonden groepen kwamen overeen met de verwachtingen van de domeinexpert.

Aansluitend op de eerste sub-vraag volgt de vraag:

-Welke groepen hebben een verhoogd risico op een no-show?

Door te kijken naar het no-show percentage van de cliënten binnen de groepen bleek dat er duidelijke verschillen tussen de groepen aanwezig waren.

Bij het klinische no-show percentage bleek dat oudere vluchtelingen uit het Midden Oosten en onverzekerden therapietrouwer waren dan de overige klinisch behandelde groepen.

Bij het poliklinische no-show percentage bleek de naoorlogse generatie therapietrouw te zijn. Jonge vluchtelingen uit het Midden-Oosten en Afrika bleken relatief hogere percentages te hebben. Ook cliënten met gezinnen bleken hogere no-show percentages te hebben.

Deze informatie kan in de toekomst worden gebruikt om bij de groepen met een verhoogd risico extra voorzorgsmaatregelen te nemen zoals sms alerts. De informatie kan ook gebruikt worden bij het inplannen van groepsbijeenkomsten zodat een groep niet te veel risicocliënten bevat.

Aanvullend is het no-show percentage onderzocht met behulp van classificatie technieken. Hieruit kwamen factoren die terug te vinden waren in de clusters met verhoogde no-show percentages.

De derde sub-vraag met betrekking tot de User data was geformuleerd als:

-Welke factoren zijn bepalend voor de behandelintensiteit?

De behandelintensiteit, geformuleerd als het gemiddeld aantal afspraken per 30 dagen bleek vooral afhankelijk te zijn van de behandelsetting. Klinisch en dagklinisch behandelde cliënten hebben gemiddeld meer afspraken dan poliklinisch behandeld cliënten. Experimenten met verschillende technieken leverde geen andere factoren op. De behandelintensiteit lijkt daarom een moeilijk te analyseren factor.

7.2.2 Afspraakhistorie

Met betrekking tot de afspraakhistorie is de volgende onderzoeksvraag gesteld:

> -*Welke patronen met betrekking tot no-show en behandelfase zijn in de afspraakhistorie te vinden?*

Uit de no-show analyse kwamen helaas geen onverwachte of sterke verbanden. Er waren wel kleine verschillen tussen patronen te zien. Zo blijkt er een verhoogde kans te zijn op een no-show na een groepscontact dan na psychotherapie.

Verder bleken cliënten die reeds een no-show hebben vertoond een hogere kans te hebben op een no-show in de toekomst. Deze informatie kan gebruikt worden om cliënten te monitoren vanaf het moment dat er een no-show optreedt.

Bij het vergelijken van afspraaktypes per behandelfase zijn duidelijke verschillen gevonden die grotendeels aansluiten bij de verwachting van de stichting. Zo bleken intake en diagnostische afspraken vaker in het eerste behandeljaar voor te komen en gerichte therapie meer in het tweede tot vierde behandeljaar.

Er waren weinig verschillen te zien tussen het tweede tot vierde behandeljaar en vijfde behandeljaar en verder. Dit in tegenstelling tot de verwachting van de stichting waarbij wordt verwacht dat er meer begeleiding en minder gerichte therapie plaatsvindt.

Een verrassende ontdekking bleek de aanwezigheid van intakegesprekken bij cliënten die meer dan 4 jaar in behandeling zijn. 20% van deze cliënten bleek nog een intakegesprek gehad te hebben in deze fase van hun behandeling.

Dit kan bijvoorbeeld betekenen dat er een fout in de database aanwezig is of dat User verkeerd gebruikt is. Het vermoeden is dat dit komt door een conversie fout. Het is aan te raden dit verschijnsel verder te onderzoeken omdat dit om relatief veel cliënten gaat.

Een manier om dit te onderzoeken is door de momenten van de intakegesprekken te vergelijken met het moment van de voorgaande afspraak. Wanneer er een groot gat tussen beide afspraken is kan dit betekenen dat er twee verschillende inschrijvingen waren die bij de conversie zijn samengevoegd.

7.3 Machine learning

Bij het onderzoek is gebruik gemaakt van een aantal machine learning technieken. Er is gewerkt met classificatie, clustering en sequence mining. De classificatietechnieken bleken voor het Compass probleem en de no-show percentages duidelijk te interpreteren en nuttige resultaten te produceren. Bij het toepassen op de behandelintensiteit leverde deze technieken echter minder duidelijk en bruikbare resultaten op.

Clustering van de cliëntgegevens leverde voor de stichting duidelijk herkenbare groepen en patronen op. De bij het clusteren gevonden verbanden met betrekking tot no-show en het Compass probleem kwamen grotendeels overeen met wat er bij het toepassen van classificatietechnieken was gevonden.

Het toepassen van sequence mining met betrekking tot no-shows leverde resultaten op die niet direct duidelijk en intuïtief waren. De overkoepelende verzameling van process mining technieken biedt wellicht geschiktere alternatieven. Process mining is een relatief jong vakgebied waarin veel onderzoek is geweest in de afgelopen jaren [18]. Verschillende tools die bij dit onderzoek zijn ontwikkeld zijn als opensource software beschikbaar gesteld op http://www.processmining.org/.

De vergelijking in sequences tussen verschillende behandelfasen waarbij gebruik werd gemaakt van betrouwbaarheidsintervallen leverde interessante resultaten op. Deze aanpak bleek geschikt om verschillen in afspraakpatronen tussen behandelfasen te vergelijken. De wijze van visualiseren van de resultaten in grafiekvorm bleek een voor de domeinexpert overzichtelijke en duidelijke manier.

7.4 Verder onderzoek

Bij de uitvoer van dit onderzoek zijn verschillende vragen naar boven gekomen die door de tijd en scope beperkingen niet konden worden behandeld. Als suggestie voor verder onderzoek zijn deze vragen hier vermeld.

-Wat zijn de verschillen tussen cliënten die bij Compass metingen verbetering laten zien ten opzichte van cliënten die verslechtering tonen?

Bij de cluster analyse is wel gekeken naar een soortgelijk aspect, de GAF score. Hierbij waren echter geen duidelijke verschillen te zien. Het kan zijn dat er bij de gegevens van de Compass metingen wel duidelijke verschillen zijn. Tevens kan er bij de Compass metingen ingezoomd worden op gerichte aspecten door scores behorend bij specifieke vragen te vergelijken.

-Welke factoren zijn bepalend voor de behandelduur?

Bij dit onderzoek is een verwante factor onderzocht, de behandelintensiteit. Hierbij zijn echter geen sterke of verrassende verbanden naar voren gekomen. Het onderzoeken van de behandelduur leidt mogelijk tot betere resultaten.

Er moet echter wel rekening worden gehouden met de vermoedelijk mogelijk incorrecte inschrijvingen als gevolg van een conversie.

-Zijn er relaties tussen tijd en no-show in de afspraakhistorie?

Het is mogelijk dat no-shows bijvoorbeeld vaker in de ochtend optreden dan in de middag. Dit kan per groep en afspraaktype worden onderzocht. Er kan ook worden gedacht aan het onderzoeken van de tijdspanne tussen het optreden no-shows.

-Welke patronen zijn er te vinden in de indirecte uren?

De analyse van de afspraakhistorie heeft zich binnen dit onderzoek beperkt tot directe uren waarbij een cliënt aanwezig hoort te zijn. Naast het onderzoeken van de indirecte uren in de afspraakhistorie kan ook worden gedacht aan het onderzoeken van de verhouding indirecte/directe uren. Hierbij kunnen mogelijk groepen worden geïdentificeerd waarbij veel indirecte uren worden gemaakt. In het algemeen vormt een groep waarbij de verhouding directe/indirecte uren hoog is een inefficiënt patroon voor de stichting doordat er relatief minder tijd wordt besteed aan daadwerkelijke behandelingen.

Referenties

[1] H. Kob en G. Tan, „Data mining applications in healthcare," vol. 19, nr. 2, 2005.

[2] A. Eapen, *Application of Data mining in Medical Applications,* Waterloo: University of Waterloo, 2004.

[3] „DBC onderhoud factsheet," [Online]. Available: www.dbconderhoud.nl.

[4] P. Chapman, J. Clinton, R. Kerber, T. Khabaza, T. Reinartz, C. Shearer en R. Wirth, „CRISP-DM 1.0 Step-by-step data mining guide," SPSS, 2000.

[5] U. Fayyad, G. Piatetsky-Shapiro en P. Smyth, „The KDD Process for Extracting Useful Knowledge from Volumes of Data," *COMMUNICATIONS OF THE ACM,* vol. 39, nr. 11, 1996.

[6] V. D en A. J, „Process model for data mining In healt care sector," in *Health Ambient Information Systems Workshop*, Colombia, 2011.

[7] T. M. Mitchell, in *Machine learning*, McGraw-Hill, 1997, p. 54.

[8] T. M. Mitchell, in *Machine Learning*, McGraw-Hill, 1997, pp. 145-149.

[9] M. N. Murty, A. K. Jain en P. J. Flynn, „Data Clustering: A Review," vol. 31, nr. 3, 1999.

[10] M. Zaki, *SPADE: an efficient algorithm for Mining Frequent Sequences,* Kluwer Academic Publishers, 2001.

[11] S. Wallis, „Binomial confidence intervals and contingency tests: mathematical fundamentals and the evaluation of alternative methods," 2009. [Online]. Available: http://www.ucl.ac.uk/english-usage/staff/sean/resources/binomialpoisson.pdf.

[12] H. R. C., „Very Simple Classification Rules Perform Well on Most Commonly Used Datasets," *Machine Learning,* nr. 11, pp. 63-91, 1993.

[13] A. Kak, *Expectation-Maximization Algorithm for Clustering Multidimensional Numerical Data,* Purdue University, 2013.

[14] R. Rivest, „RFC 1321," April 1992. [Online]. Available: http://www.ietf.org/rfc/rfc1321.txt.

[15] G. Koole, Optimization of Business Processes, 2010.

[16] J. Stangroom, „Social Sience Statistics," 2013. [Online]. Available: http://www.socscistatistics.com/tests/pearson/Default.aspx. [Geopend 31 05 2013].

[17] StatTrek.com, „Stat Trek," StatTrek.com, 2013. [Online]. Available: http://stattrek.com/survey-research/stratified-sampling.aspx. [Geopend 01 06 2013].

[18] I. T. F. o. P. Mining, „Process Mining Manifesto," *Lecture Notes in Business Information*

Processing, nr. 99, pp. 169-194, 2012.

Bijlage 1 Overzicht Compass lijsten

Tabel 1

Overzicht van meetinstrumenten opgenomen in Compass

Instrument	Omschrijving	Items	Batterij
BSI	Zelfbeoordelingslijst voor het meten van symptomen van psychopathologie. Het instrument omvat 8 dimensies van psychopathologie en een algemene indicator voor psychopathologie.	53	Basis
HSCL	Zelfbeoordelingslijst voor het meten van angst- en depressieve klachten.	25	V&A basis
HTQ – Deel I	Inventarisatie van welke traumatische ervaringen (a) men zelf heeft ondervonden, (b) men getuige is geweest, (c) men heeft horen spreken of (d) men niets mee te maken heeft gehad.	24	Basis (intake); V&A basis (intake)
HTQ – Deel IV	Zelfbeoordelingslijst voor het meten van symptomen van PTSS. De items zijn afgeleid van de DSM-III-R criteria voor PTSS. Het instrument omvat 3 dimensies: herbeleving, vermijding en hyper-arousal.	16	Basis; V&A basis
RESQ	Inventarisatie van een aantal hulpbronnen die mogelijk steun kunnen geven tijdens moeilijke situaties in het algemeen. Het instrument omvat 8 dimensies met de factoren die mogelijk steungeven en een totaalscore.	34	Basis
WHOQOL	Zelfbeoordelingslijst voor het meten van kwaliteit van leven. Het instrument omvat 4 domeinen van kwaliteit van leven en een indicatie voor kwaliteit van leven in het algemeen en tevredenheid met de eigen gezondheid.	26	Basis
PILL – Deel A	Zelfbeoordelingslijst voor inventariseren van somatische en psychische klachten.	33	V&A basis; Aanvullend
NITE	Inventarisatie van slaappatronen, slaap-bevorderende en slaap-verstorende factoren. Het instrument omvat oa de volgende onderdelen: middelen/medicatie, slaappatronen, slaapstoornissen, nachtmerries.	38	Aanvullend
PES	Zelfbeoordelingslijst voor het meten van symptomen van dissociatie. Het instrument omvat een algemene indicator voor dissociatie symptomen en 3 domeinen met verschillende manifestaties van dissociatie symptomen: amnesie, absorptie, depersonalisatie.	28	Aanvullend
COPE-easy	Zelfbeoordelingslijst voor het inventariseren van voorkeuren voor bepaalde copingstijlen. Het instrument omvat 15 subschalen die onderverdeeld worden in 4 copingstijlen: probleem-georiënteerd, emotie-georiënteerd, vermijdend, sociale steun.	34	Aanvullend
NEO-FFI	Zelfbeoordelingslijst voor het meten van de 5 ‘Big Five’ dimensies van persoonlijkheid: neuroticisme, extraversie, openheid, altruïsme, consciëntieus.	60	Aanvullend
WAS	Zelfbeoordelingslijst voor inventariseren van levensopvattingen met betrekking tot mensen, wereld en zelf. Het instrument omvat 8 domeinen van deze levensopvattingen.	32	Aanvullend

SSL	Zelfbeoordelingslijst voor het meten van sociale redzaamheid (eenzaamheid).	12	Aanvullend
IPOV	Zelfbeoordelingslijst voor het inventariseren van probleemoplossend vermogen binnen de partnerrelatie en de tevredenheid met de partner.	21	Aanvullend
LSV	Zelfbeoordelingslijst voor het meten van lichaams- en seksualiteitsbeleving. Het instrument omvat 9 domeinen van lichaams- en seksualiteitsbeleving.	30	Aanvullend