Geweld(ig): datamining!

Een zoektocht naar het profiel van geweldplegers
met behulp van dataminingtechnieken

Geweld(ig): datamining!

Een zoektocht naar het profiel van geweldplegers
met behulp van dataminingtechnieken

Afstudeerscriptie van Ankie van der Zanden in het kader van de studie Bedrijfswiskunde en Informatica, aan de Vrije Universiteit te Amsterdam.

Den Haag, juni 2005

Begeleiders:

Dr. E. Marchiori	Eerste begeleider, Vrije Universiteit
Dr. ir. G. Jongbloed	Tweede begeleider, Vrije Universiteit
Mw. A. Patty-Hüttmann	Politie Haaglanden
Drs. P. Versteegh	Politie Haaglanden

Voorwoord

Dit is de scriptie van mijn afstudeerstage over datamining, toegepast bij Politie Haaglanden op het onderwerp geweldplegers. De scriptie is geschreven in het kader van mijn zes maanden durende stage in het kader van de opleiding Bedrijfswiskunde en Informatica aan de Vrije Universiteit te Amsterdam. De stage vond plaats op afdeling Analyse & Research, op het hoofdbureau van Politie Haaglanden, te Den Haag.

Dankzij deze interessante stage heb ik veel geleerd over de politie, criminaliteit en criminaliteitsanalyse. Voor deze mooie ervaring ben ik drs. Peter Versteegh dan ook dankbaar. Daarnaast een woord van dank aan dr. Janine Janssen voor het commentaar op mijn scriptie en alle andere collega's voor hun steun en de leuke tijd die ik heb gehad op de afdeling. Een speciaal woord van dank gaat uit naar Astrid Patty-Hüttmann, mijn begeleidster vanuit de politie, voor haar goede begeleiding en enorme enthousiasme. Dankzij haar heb ik veel geleerd over de structuur en de gegevens in het registratiesysteem van de politie en over de strategische criminaliteitsanalyse.

Daarnaast wil ik mijn eerste begeleidster dr. Elena Marchiori bedanken voor haar uitstekende begeleiding. Dankzij haar inbreng en die van de leden van de Weka mailing-list ben ik veel te weten gekomen over de technieken die achter datamining schuil gaan. Daarnaast gaat een woord van dank uit naar mijn tweede begeleider, dr. ir. Geurt Jongbloed.

Ten slotte wil ik mijn vriend en mijn ouders bedanken die al mijn problemen en dilemma's geduldig hebben aangehoord en mij gesteund hebben.

Voor eventuele vragen en opmerkingen naar aanleiding van deze scriptie, kunt u mailen naar: ankievdz@xs4all.nl

Den Haag, juni 2005

Ankie van der Zanden

Samenvatting

Momenteel is er in regio Haaglanden een stijging van het aantal geregistreerde geweldsdelicten. Daarbij komt dat geweld een korpsprioriteit is. Om die redenen wil het korps meer inzicht krijgen in de kenmerken van verdachten van geweld. Bovendien wil men weten wat datamining voor de tactische / strategische analyse kan betekenen. Ook landelijk is er bij de politie vraag naar datamining. In navolging van het onderzoek van het programmabureau Abrio, is er een landelijke expertgroep Datamining samengesteld. Vanuit deze expertgroep is er vraag naar meer onderzoek betreffende datamining.

Om in de toekomst datamining te kunnen toepassen is het noodzakelijk de technieken te onderzoeken en welke technieken geschikt zijn voor de gevraagde analyses. Het doel van dit onderzoek is een overzicht te geven van verschillende dataminingtechnieken met een onderbouwing welke technieken bij welke vraagstellingen passen. Deze technieken zijn vervolgens toegepast op de casus 'geweldplegers in regio Haaglanden van 2004'. Met behulp van datamining wordt inzicht gegeven in de verdachten van geweldsdelicten.

Uit dit onderzoek is gebleken dat dataminingtechnieken nuttig zijn voor de strategische analyses van Politie Haaglanden. Vooral om de aard van geregistreerde criminaliteit vast te stellen biedt datamining een meerwaarde. Het belangrijkste voordeel van datamining ten opzichte van de huidige gebruikte methoden is dat er vooraf geen hypothesen gesteld hoeven te worden. Op die manier kunnen relaties worden ontdekt die niet eerder aan het licht zijn gekomen.

Kort gezegd kan classification gebruikt worden als men geïnteresseerd is in één specifiek kenmerk of variabele. Als er geen doelvariabele bekend is, kan eerst clustering gebruikt worden om groepen te ontdekken waarvan vervolgens met behulp van classification bekeken kan worden wat de kenmerken van die groepen zijn. Als men wil zoeken naar sterke verbanden in een groep verdachten of delicten dan kan de techniek association rules het beste gebruikt worden.

De dataminingtechniek classification is nuttig gebleken in het zoeken van onderscheidende kenmerken tussen twee of meerdere groepen. De kenmerken die onderscheid maken tussen geweldplegers en non-geweldplegers zijn onder andere nationaliteit, harddruggebruik en leeftijd. Er zijn geen persoonskenmerken gevonden op basis waarvan onderscheid te maken is tussen verdachten van verschillende soorten geweld (fysiek geweld, verbaal geweld, seksueel geweld of vermogensgeweld). De delictkenmerken geven echter wel een indicatie van het type geweldpleger maar deze resultaten zijn niet verrassend.

Met de techniek association rules is exploratief gezocht naar sterke verbanden. Dit heeft naast reeds bekende relaties ook nieuwe relaties aan het licht gebracht. Er zijn groepen gevonden die bijna geheel bestaan uit mannelijke verdachten, groepen waarin de verdachten voornamelijk in Den Haag wonen en groepen waarin de Nederlandse etniciteit erg vaak voorkomt. Verder is gebleken dat de techniek association rules minder geschikt is voor datasets met veel variabelen. Er zijn in dat geval te veel gevonden relaties waaruit moeilijk op te maken valt welke bruikbaar zijn.

Clustering heeft inzicht gegeven in samenhangende groepen binnen de dataset. Allereerst zijn er groepen gevonden die vooral samenhangen op basis van hun woonadres. Maar ook andere kenmerken komen daarbij aan het licht zoals verslaving en leeftijd. Bij het clusteren van de verdachten op basis van zowel hun persoonskenmerken als op hun criminele verleden, zijn groepen gevonden die gekenmerkt worden door het aantal antecedenten dat ze op hun naam hebben staan en de soort delicten die ze hebben gepleegd. Er is bijvoorbeeld een groep te onderscheiden die vooral zware delicten pleegt en een groep met verdachten die nooit wapens gebruikt.

Inhoudsopgave

1. Inleiding

"What we call the beginning is often the end.
And to make an end is to make a beginning.
The end is where we start from."
– Thomas Stearns Eliot, Four Quartets

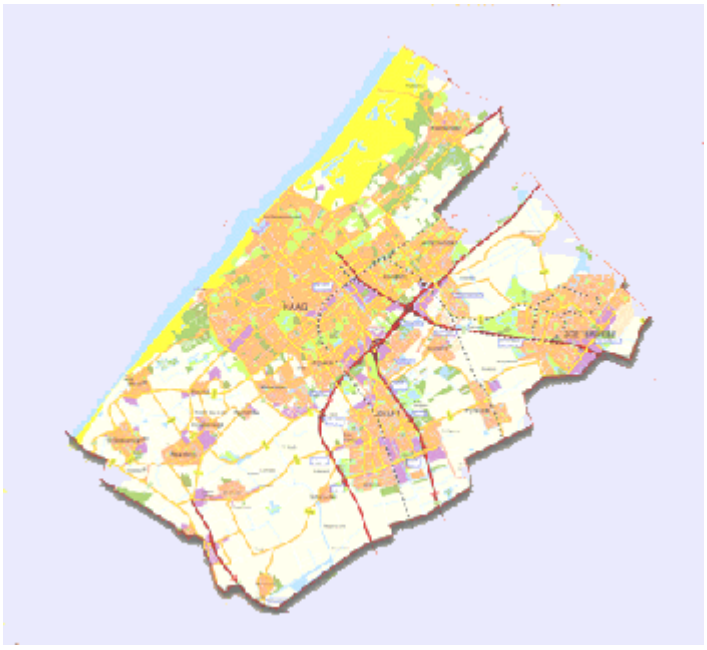
Politie Haaglanden heeft zeer veel gegevens over criminaliteit in de regio. Die gegevens worden onder andere gebruikt voor opsporing (operationele analyse), om een effectieve aanpak van criminaliteit te kunnen realiseren (tactische analyse) en om informatie te verstrekken van de criminaliteit in de regio voor overheidsinstellingen (strategische analyse). Op de afdeling Analyse & Research worden tactische en strategische analyses uitgevoerd. Bij politieregio's door het hele land neemt de interesse naar *datamining* voor zowel het operationele, tactische als strategische vlak toe. Zo ook bij Politie Haaglanden. Zij zijn benieuwd naar wat datamining aan meerwaarde kan bieden boven de meer bekende statistische technieken. Voor dat doel is datamining toegepast op het actuele onderwerp geweld. Zo wordt er gekeken naar de mogelijkheden om een profiel te schetsen van geweldplegers. Daarnaast wordt aangegeven welke dataminingtechnieken het meest bruikbaar zijn voor de strategische analyse.

Het verslag bestaat uit zeven hoofdstukken, te beginnen met een korte beschrijving van Politie Haaglanden en de afdeling Analyse & Research waar de stage is uitgevoerd. Het derde hoofdstuk beschrijft de probleemstelling en de aanpak van het onderzoek. Allereerst is er literatuurstudie gedaan naar de verschillende dataminingmethoden die in hoofdstuk 4 worden beschreven. Hier is tevens te vinden welke methoden voor de afdeling interessant zijn. De theorie die in het hoofdstuk besproken wordt, dient als referentiekader voor hoofdstuk 6. In hoofdstuk 5 worden de gegevens over geweld besproken die voor het dataminen gebruikt zijn. Het werkelijke modelleren is in hoofdstuk 6 beschreven, evenals de resultaten die hieruit voort zijn gekomen. Ten slotte worden de conclusies en aanbevelingen in hoofdstuk 7 gegeven.

2. Politie Haaglanden

Starsky: It's 10 o'clock, you're late; I've been here since 8.
Hutch: 8 o'clock? I didn't even know this place opened that early.
Starsky: It's okay though, because crime called in sick, it's going to get a late start too.
– Starsky and Hutch

De Nederlandse politie bestaat uit 25 regionale korpsen en het Korps landelijke politiediensten (KLPD). Deze korpsen leveren een bijdrage aan veiligheid, leefbaarheid en de bestrijding van criminaliteit in Nederland. Het KLPD organiseert de landelijke politietaken. Politie Haaglanden is één van die korpsen. 365 dagen per jaar, 24 uur per dag actief om de regio veilig en leefbaar te houden. Bij korps Haaglanden werken ongeveer 5000 mensen, met een werkgebied uitstrekkend van het Westland tot en met Wassenaar en van Scheveningen tot en met Zoetermeer. Een aanzienlijk deel van de werknemers is voor de burger zichtbaar als “blauw op straat” maar de voor het publiek minder zichtbare medewerkers, zoals de analisten, zijn minstens zo belangrijk.

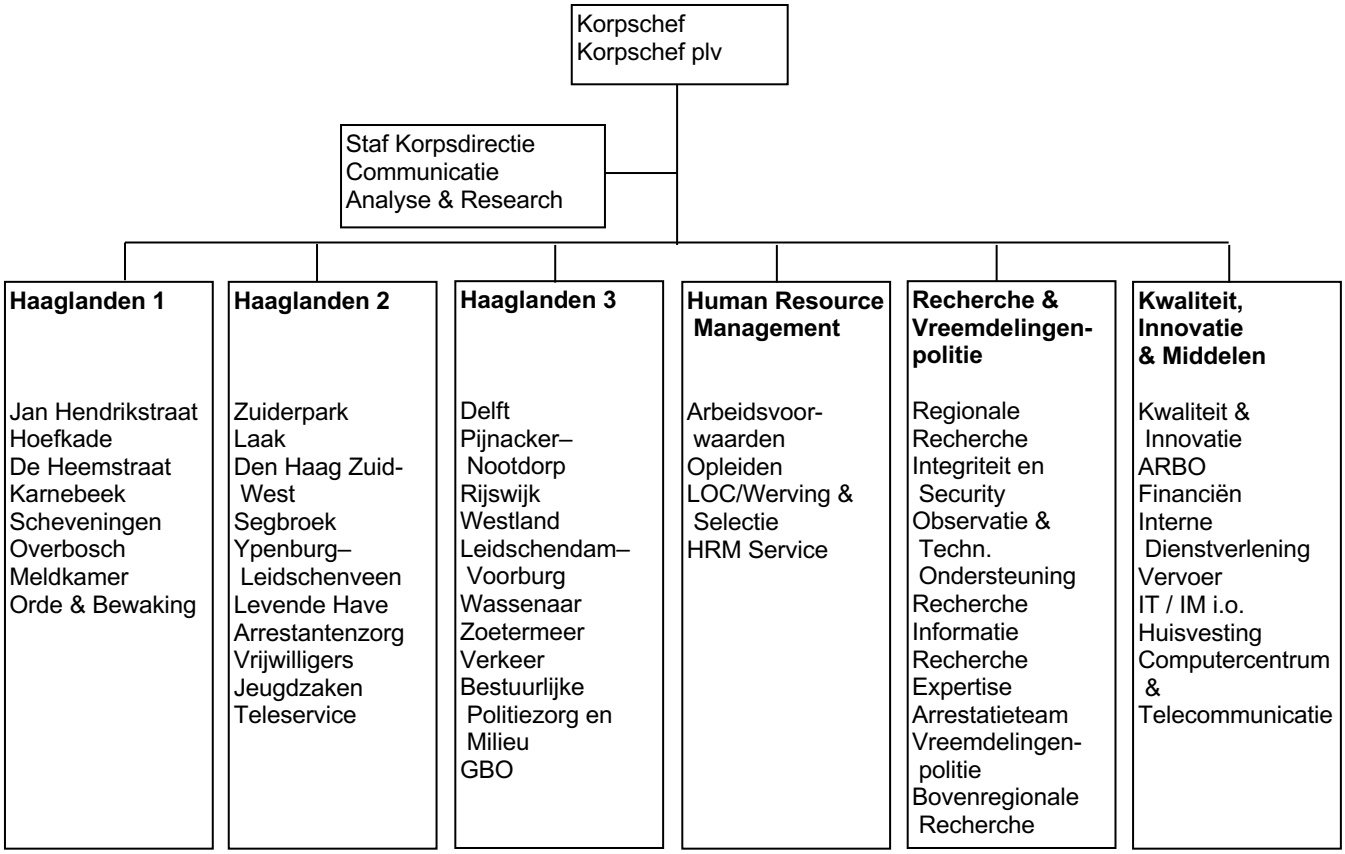


Figuur 2-1: Regio Haaglanden

Het werkterrein van Politie Haaglanden telt negen gemeenten met meer dan één miljoen inwoners. Het gebied is verdeeld in achttien bureaugebieden.

2.1 Analyse & Research

Analyse & Research valt onder Staf Korpsdirectie, zie het organogram in figuur 2-2, op de volgende bladzijde. De afdeling ondersteunt de directie met kennis en informatie over criminaliteit en onveiligheid, waarmee de politiezorg kan worden versterkt. De afdeling doet wetenschappelijk onderzoek naar criminaliteitsproblemen en maakt strategische criminaliteitsanalyses ten behoeve van beleidsvoorbereiding, beleidsvorming en aanpak van criminaliteitsproblemen. Een product dat ieder jaar geleverd wordt is de criminaliteitskaart, een instrument dat de (geregistreerde) criminaliteit en gevoelens van onveiligheid in beeld brengt in een specifiek, geografisch gebied. Ook worden de probleemanalyses op bepaalde thema's jaarlijks uitgevoerd zoals op de korpsprioriteiten veelplegers, jeugd en geweld.



Figuur 2-2: Organogram van Politie Haaglanden

3. Probleemstelling en aanpak

"The real problem is what to do with the problem-solvers after the problems are solved."
– Tell a Friend-Gay Talese

3.1 Probleemstelling

Momenteel is er in regio Haaglanden een stijging van het aantal geregistreerde gewelds*delicten*. Het korps wil mede daarom meer inzicht krijgen in de kenmerken van *verdachten* van geweld. Een andere reden is dat geweld korpsprioriteit is. Dat wil zeggen dat er veel aan dit probleem gedaan wordt. Bovendien wil men weten wat *datamining* voor analyse van geweldplegers kan betekenen. Met de resultaten kunnen zonodig door een andere afdeling strategische en / of tactische besluiten genomen worden.

De data die voor dit onderzoek naar geweld gebruikt worden, zijn afkomstig van het opsporingsregister van de herkenningsdienst (HKS). Het HKS bevat veel gedetailleerde en betrouwbare gegevens van zowel delicten als van verdachten. Als de dader van een delict bekend is dan zijn deze in het systeem gekoppeld. Momenteel gaat het analyseren van de gegevens via het koppelen van tabellen waar vervolgens statistische methoden op worden toegepast. Het is niet goed mogelijk om patronen te herkennen, tenzij er al een vermoeden van een patroon is. Datamining zou hiervoor de oplossing moeten bieden.

Ook landelijk is er bij de politie vraag naar datamining. Het programmabureau Abrio heeft in 2004 een rapportage “Meerwaarde datamining” opgeleverd, mede tot stand gekomen door drie pilots binnen de korpsen Amsterdam-Amstelland, Midden- en West-Brabant en het KLPD. In navolging van dit onderzoek is er een landelijke expertgroep Datamining samengesteld. Vanuit deze expertgroep is er vraag naar meer onderzoek betreffende datamining.

Het is vooralsnog de bedoeling dat analisten in de toekomst met datamining kunnen werken volgens een standaard werkwijze. Deze standaard werkwijze zal onder andere inhouden dat de analist niet verdwaald raakt in de vele bestaande dataminingtechnieken. Het is daarom noodzakelijk te onderzoeken welke technieken geschikt zijn voor de gevraagde analyses. Hiervoor moet uitgezocht worden welke analyses worden verricht, welke vraagstellingen daarbij gesteld worden en welke gegevens daarvoor ter beschikking zijn.

De doelen en vraagstellingen van dit onderzoek kunnen als volgt worden beschreven:

- Een overzicht geven van de verschillende dataminingtechnieken met daarin onderbouwd welke technieken bij welke vraagstelling passen.

  1) Is datamining nuttig voor Politie Haaglanden?

  2) Welke dataminingtechnieken zijn geschikt voor de vraagstellingen binnen Politie Haaglanden?

- Inzicht geven in de verdachten van geweldsdelicten met behulp van datamining. Het gaat hierbij om de verdachten die in regio Haaglanden zijn aangehouden ter zake van een misdrijf in 2004.

  1) Zijn er specifieke kenmerken van geregistreerde verdachten te onderscheiden, als gekeken wordt naar verschillende soorten geweld?
     Onder de soorten geweld wordt verstaan:
     - gewelddadige seksuele delicten,
     - fysiek geweld,
     - verbaal geweld,
     - gewelddadige vermogensdelicten.

     Zo ja, wat zijn die kenmerken?

  2) Zijn er verschillen tussen verdachten van geweldsmisdrijven en andere verdachten?
     Zo ja, wat zijn die kenmerken?

3) Wat zijn de verbanden die exploratief gevonden worden in de dataset?

4) Liggen er groepen verborgen in de gegevens?
   Zo ja, hoe zien die groepen eruit?

3.2 Terminologie

3.2.1 Geweld

De term 'geweld' in dit onderzoek omvat al het geweld dat gericht is tegen personen, of het geweld nu het doel was of het middel. Geweld in dit onderzoek heeft een juridische basis, in tegenstelling tot de manier waarop het begrip in de volksmond gebruikt wordt. De geweldsmisdrijven zijn in te delen in vier categorieën: gewelddadige seksuele delicten, verbaal geweld, fysiek geweld en vermogensdelicten met geweld. Onder gewelddadige seksuele delicten vallen de rubrieken verkrachting en aanranding der eerbaarheid. Onder verbaal geweld valt alleen de rubriek bedreiging. Fysiek geweld bestaat uit mishandeling, poging doodslag, doodslag voltooid en overige misdrijven tegen het leven. Bij vermogensdelicten met geweld gaat het om afpersing en diefstal met geweld. Voor een duidelijker begrip van geweld, wordt verwezen naar paragraaf 5.5 en bijlage A, waarin de wetsartikelen zijn gegeven.

3.2.2 Bron

De gegevens komen uit het opsporingsregister van de herkenningsdienst, het HKS. Hierin bevinden zich gegevens omtrent misdrijven waarvan aangifte is gedaan en gegevens omtrent verdachten met proces verbaal en hun criminele verleden.

De gemaakte extractie bestaat uit alle personen die in de periode 01-01-2004 tot en met 31-12-2004 voorkomen met één of meer *antecedenten*. Van deze personen is hun volledige delictgeschiedenis voor zover deze in het HKS is opgeslagen bekend. Dat wil zeggen dat er ook antecedenten van vóór 01-01-2004 in het bestand voorkomen.

3.3 Aanpak

Om uiteindelijk datamining toe te kunnen passen op het onderwerp geweld, was een aantal dingen noodzakelijk.

3.3.1 Dataset

Allereerst moest er een geschikte dataset gemaakt worden waar de analyse op verricht kan worden. Het voornemen om de technieken op meerdere datasets te evalueren is niet mogelijk gebleken omdat het prepareren van de data erg veel tijd in beslag nam. Bovendien konden de technieken ook op verschillende delen van de dataset toegepast worden om toch conclusies te kunnen trekken over de werking van de technieken.
De gebruikte dataset moest aan een aantal eisen voldoen. Natuurlijk moesten de gegevens valide zijn, maar ook moest de dataset van voldoende grootte zijn. De resultaten van datamining zijn namelijk alleen bruikbaar bij voldoende gegevens. Om de dataset in de geschikte vorm te krijgen voor het dataminen, zijn veel bewerkingen vooraf gegaan.

3.3.2 Software

Software was een tweede vereiste. Een geschikt pakket voor het toepassen van datamining is natuurlijk onmisbaar. Er zijn vele bruikbare applicaties, daarom is er een keuze gemaakt van software die al eerder bij de Nederlandse Politie is gebruikt en software die al bekend was. Voor dit onderzoek is Weka versie 3.4.4 gekozen.
Daarnaast was er een statistisch pakket nodig om de dataset te kunnen beschrijven en te bewerken. Het standaard statistische pakket van de afdeling Analyse & Research, SPSS versie 12, voldeed aan de wensen.

Verder was het handig om een script te schrijven dat diverse technieken van datamining achter elkaar uit kan voeren. Aangezien er op de werkplek geen programmeertaal beschikbaar was, is gekozen om, via de server van de Vrije Universiteit, Unix te gebruiken.

3.3.3 Datamining

Zowel *classification*, *association rules* als *clustering* is toegepast op de dataset. Tussendoor is de dataset zonodig bewerkt tot het de gewenste vorm had voor de te gebruiken techniek. Vooral voor het classificeren zijn vele *algoritmen* getest. De best resulterende algoritmen zijn verder uitgewerkt zoals het variëren van parameters en het toepassen van die *classifier* op de gehele dataset. Voor association rules en clustering is het dataminen tot één algoritme beperkt. De resultaten zijn tenslotte in overleg gecontroleerd op relevantie in de praktijk zodat de resultaten die triviaal waren verwijderd konden worden.

4. Datamining

"In theory, there is no difference between theory and practice.
But, in practice, there is."
– Jan L.A. van de Snepscheut

Het doel van dit hoofdstuk is het begrip datamining te definiëren en uit te leggen wat de verschillende mogelijkheden ervan zijn: classificatie, association rules en clustering. Het is de theoretische achtergrond voor het modelleren in hoofdstuk 6. Om het begrip in een bredere context te plaatsen is in paragraaf 4.2 het verband gelegd tussen datamining en de meer bekende statistische onderzoeksmethode. Voor meer informatie over de werking van de drie dataminingtechnieken wordt verwezen naar het boek van Ian H. Witten en Eibe Frank [*32*].

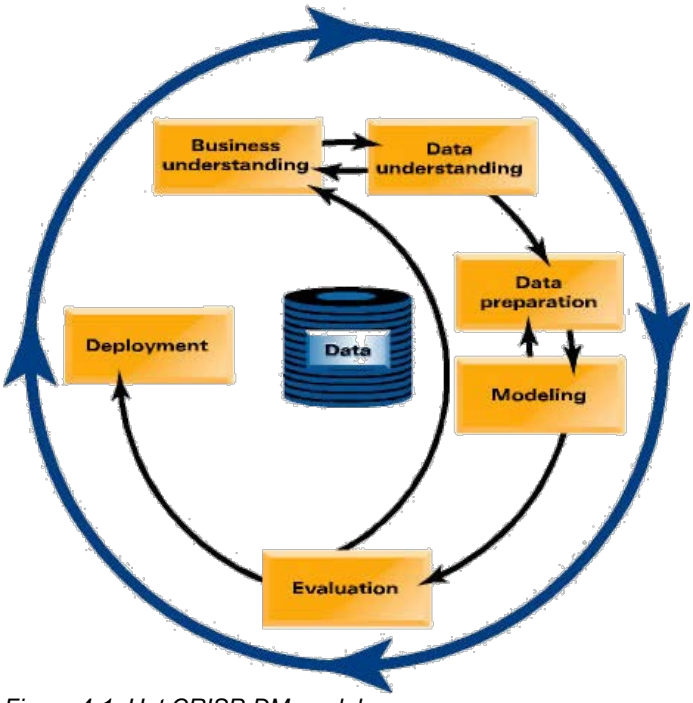
4.1 Wat is datamining?

Datamining is een vorm van kwantitatieve analyse en is bedoeld voor het werken met grote(re) gegevensverzamelingen. Datamining wordt in het rapport van de pilot van Abrio [*1*] omschreven als:

"Het zoeken naar (onbekende) en potentieel nuttige patronen of profielen die verborgen liggen in grote gegevensverzamelingen".

Andere definities uit de literatuur komen hier grotendeels mee overeen en daarom zal deze definitie ook in dit rapport gebruikt worden [*14, 27, 29*].

Een veel gebruikt proces om grootschalige datamining projecten in bedrijven te leiden is CRISP-DM, dat staat voor Cross Industry Standard Process for Data Mining. Het gehele datamining proces bestaat uit zes fases. In figuur 4-1 is het verloop tussen de fasen te zien.



Figuur 4-1: Het CRISP-DM model

Allereerst wordt de doelstelling van het onderzoek en wensen van het bedrijf omgezet naar een probleemomschrijving in termen van datamining. Deze fase wordt bedrijfsbegrip (**business understanding**) genoemd. In de volgende fase verzamelt de dataminer de gegevens en krijgt verstand van de data (**data understanding**) met als doel een eerste indruk te krijgen en problemen te identificeren. De derde fase is het voorbereiden van de data (**data preparation**). Deze omvat alle

activiteiten om de definitieve dataset uit de ruwe gegevens te construeren. Deze activiteiten zijn bijvoorbeeld het selecteren van variabelen, transformeren en het filteren van foutieve of onbetrouwbare gegevens. Als alle voorbereidingen eenmaal gedaan zijn dan kan het modelleren (**modeling**) beginnen. In deze fase worden diverse modelleringtechnieken gekozen en toegepast met parameters die optimaal gekozen zijn. Er zijn diverse technieken om een zelfde probleem op te lossen. Sommige technieken eisen een bepaalde vorm van de gegevens. Daarom is het soms nodig om terug te gaan naar de fase van het voorbereiden van de data. Na het werkelijke dataminen begint de evaluatiefase (**evaluation**). Daarin wordt gekeken of de gestelde doelen behaald kunnen worden met de gekozen modellen en dat er geen kwesties zijn die onvoldoende zijn onderzocht. In deze fase moet een besluit genomen worden over hoe de resultaten van het project moeten worden gebruikt. In de laatste fase wordt tenslotte de plaatsing (**deployment**) van het onderzoek bepaald. Dit kan het schrijven van een rapport inhouden of het reproduceerbaar maken van het model voor de klant.

4.2 Datamining versus statistiek

Wat is het verschil tussen datamining en statistiek?
Eigenlijk moet men niet zoeken naar een scheidingslijn tussen die twee analyse methoden omdat ze in elkaar over lopen. Sommige dataminingtechnieken komen uit vaardigheden voort die in standaard statistiekcursussen worden onderwezen, en anderen liggen dichter bij het soort machine learning dat uit computerwetenschap is voortgekomen. Je zou kunnen zeggen dat statistiek meer gericht is op het testen van hypothesen, terwijl datamining exploratief zoekt naar verbanden. Dit is echter wel erg zwart-wit gesteld. In werkelijkheid houdt statistiek meer in dan alleen het testen van hypothesen en dataminingtechnieken gebruiken niet altijd zoekmethoden. Het blijft echter een feit dat dataminingmethoden gebruik maken van statistische technieken. Veel algoritmen gebruiken bijvoorbeeld statistische toetsen (zoals de Chi-kwadraat toets) om na te gaan of gevonden verbanden significant zijn.

Het is belangrijk altijd in het oog te houden dat het doel van datamining is om verborgen patronen en trends te ontdekken in de gegevens. Analisten die in een dataset zoeken op basis van bekende patronen of ter bevestiging van veronderstellingen, doen geen datamining. Dat is de reden dat in dit onderzoek geen gebruik wordt gemaakt van resultaten uit eerdere onderzoeken. Het betekent uiteraard niet dat eerder gebruikte technieken onbruikbaar zijn, datamining kijkt er alleen op andere manier naar.

4.3 Classification

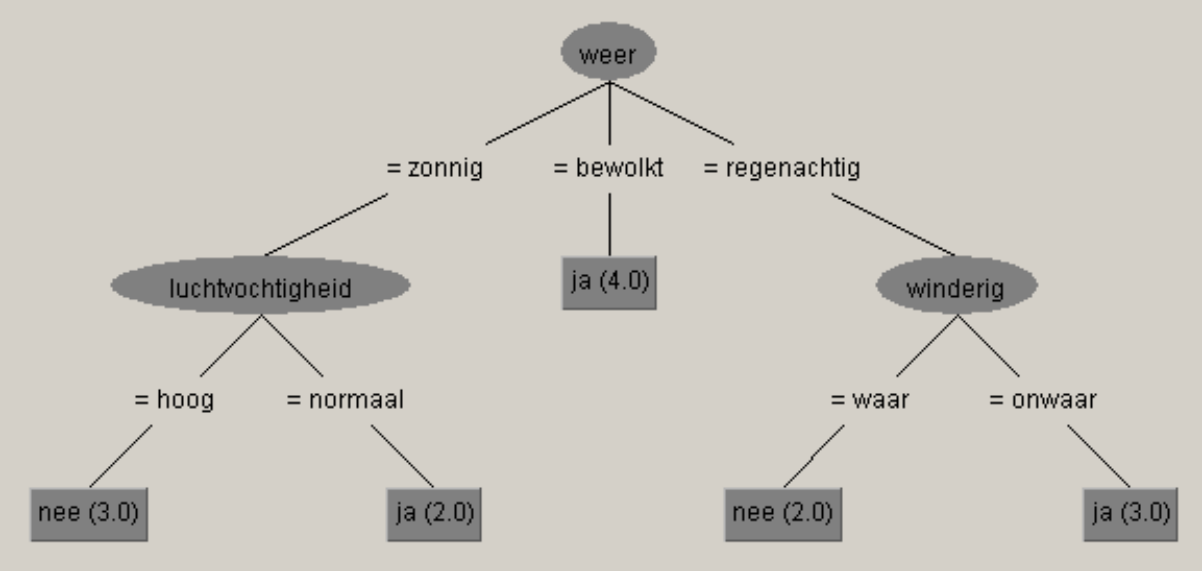
Een van de toepassingen van datamining is classification. Het is mogelijk met classificatie technieken *objecten* op basis van hun kenmerken in te delen in klassen. Met behulp van classification zoekt het systeem patronen in de gegevens die aangeven in welke categorie de objecten behoren. Zo zijn personen in te delen in ‘geweldpleger’ / ‘geen geweldpleger’ of in ‘man’ / ‘vrouw’. Ook kunnen ze bijvoorbeeld ingedeeld worden naar leeftijdsklasse. De variabelen *geweld*, *geslacht* en *leeftijdsklasse* zijn voorbeelden van doelvariabelen.

Doelvariabele De doelvariabele is de variabele waarvan de waarde voorspeld wordt.

Ook kan voor één object voorspeld worden in welke klasse het zal vallen. Als een persoon (het object) voor de eerste keer met de politie in aanraking komt dan is het mogelijk met een zekere mate van waarschijnlijkheid te voorspellen of hij/zij een veelpleger zal worden.
Ook is het mogelijk een numerieke waarde te voorspellen. Een voorbeeld: het voorspellen van het aantal antecedenten dat een persoon in zijn carrière zal oplopen. De doelvariabele is dan het aantal antecedenten. Technieken die numerieke voorspellingen doen worden ook wel regressietechnieken genoemd.

4.3.1 Visualisatie

Een manier om classificering te visualiseren is een beslissingsboom zoals in figuur 4-2. In het voorbeeld wordt bekeken wanneer men wel dan wel niet een spelletje tennis zal spelen. Zo is te zien dat het vier keer voorkomt dat het weer bewolkt was en er wel getennist werd.



Figuur 4-2: Een voorbeeld van een beslissingsboom over tennis.

De structuur van de beslissingsboom is afgeleid uit de gegevens en wordt niet van tevoren opgegeven. De methode werkt door herhaaldelijk de beste variabele te kiezen om de dataset in twee delen te splitsen. Het model kan daarvoor verschillende splitsingscriteria gebruiken die allemaal als doel hebben dat de waarden van de doelvariabele in iedere deelverzameling zo gelijk mogelijk worden.

Ook kunnen uit een beslissingsboom zoals in figuur 4-2, bij voldoende gegevens, regels afgeleid worden. Bijvoorbeeld:

1) Als het weer bewolkt is dan wordt er tennis gespeeld.
2) Als het weer zonnig is en de luchtvochtigheid hoog is dan wordt er geen tennis gespeeld.

4.3.2 Prestatie

Om prestatie te meten wordt een set geclassificeerde objecten genomen waarvan vervolgens een manier wordt aangeleerd waarop een andere ongeclassificeerde dataset in klassen ingedeeld kan worden. Voor dat doel wordt de dataset onderverdeeld in twee sets: een trainingset en een testset. Het woord trainingset geeft de betekenis al aan: de modellen worden erop getraind. Alle gegevens uit de trainingset mogen gebruikt worden om bepaalde patronen te leren ontdekken en voorspellingen te leren doen. Als het model eenmaal heeft geleerd hoe de patronen en verbanden zijn, dan wordt dit toegepast op de testset. Op de testset wordt de prestatie gemeten.

Trainingset — De trainingset is het deel van de originele dataset waarvan de doelvariabele al bekend is en waarvan de classificationmodellen kunnen leren. De modellen worden "getraind".

Testset — De testset is het deel van de originele dataset waarvan de doelvariabele onbekend is en voorspeld wordt door de classification technieken.

Bij het maken van deze verdeling wordt een aanname gemaakt dat zowel de training- als de testset een representatieve deelverzameling van het onderliggende probleem is. Dat wil zeggen dat zowel de trainingset als de testset in verhouding hetzelfde eruit ziet als de originele dataset. Als een deelverzameling aan die aanname voldoet, dan wordt het een stratified sample genoemd.

Stratified — Een stratified sample is een steekproef van de dataset waarin de verdeling van de doelvariabele gelijk is aan de verdeling van de originele dataset.

Het is gebruikelijk de prestatie van een classificatiemodel te meten in termen van een error rate. Het classification model voorspelt de doelvariabele van iedere rij in de testset: als de klasse correct voorspeld is dan wordt het geteld als een succes, zo niet, dan is het een error.

Error rate De error rate is de verhouding van het aantal gemaakte fouten ten opzichte van het aantal objecten in de testset.

4.3.3 Trainen en testen

Er zijn verschillende manieren om trainen en testen toe te passen. Hieronder worden er drie manieren beschreven: een vaste train- en testset, cross validation en leave one out.

Vaste train- en testset
De makkelijkste manier is het maken van een vaste trainingset en testset. De dataset wordt eenvoudigweg gesplitst in twee delen. Als beide deelverzamelingen representatief zijn en er veel data beschikbaar is, geven de resultaten van de testset een goede indicatie van de prestatie. Hoe groter het deel dat voor trainen gebruikt wordt, hoe beter de classification. En hoe groter de testset, hoe nauwkeuriger de resultaten. Een verdeling die vaak gebruikt wordt is tweederde deel van de dataset te gebruiken voor het trainen en éénderde deel voor het testen.

Cross validation
Een andere manier is *k*-fold cross validation. Bij die methode wordt een vast aantal delen (zogenoemde *folds*) van de dataset gekozen, bijvoorbeeld drie. De dataset wordt dan gesplitst in drie ongeveer gelijke delen die om de beurt gebruikt worden voor het testen en de overige delen voor trainen. In het geval dat *k*=3 wordt dus tweederde van de data getraind en éénderde getest. Die procedure wordt drie keer herhaald totdat uiteindelijk ieder deel precies één keer voor het testen gebruikt is. De standaard manier om de error rate te bepalen is stratified 10-fold cross validation. Van de error rates van elk van de tien delen wordt het gemiddelde genomen om een overall error rate te schatten. Dit biedt tevens een mogelijkheid om de variantie te berekenen. Om een nog nauwkeuriger error rate te bepalen kan de cross validation tien keer herhaald worden (dus tien keer 10-fold cross validation).

Leave one out
Leave one out cross validation is feitelijk hetzelfde als n-fold cross validation waarbij *n* het aantal rijen (objecten) in de dataset is. Iedere rij wordt om de beurt eruit gehaald en alle overige rijen worden gebruikt voor trainen. Ook hier wordt het gemiddelde van de error rates op alle *n* delen genomen. Een voordeel van deze methode is dat de hoeveelheid trainingdata zo groot mogelijk is zodat de classification nauwkeurig is. Bij kleine datasets kan deze manier van trainen een uitkomst zijn. Een ander voordeel is dat de resultaten statisch zijn; er wordt geen willekeurige splitsing in de dataset gemaakt en de resultaten zullen bij herhaling hetzelfde zijn. Daar tegenover staat dat de gehele procedure veel tijd kost omdat er *n* keer getraind wordt.

4.4 Association rules

Association rules geven relaties tussen verschillende *attributen* aan. Iedere willekeurige combinatie van variabelen kan een relatie vormen. Er wordt dus niet zoals bij classification gekeken naar één doelvariabele. Association rules geven geen correlatie of afhankelijkheden aan. In plaats daarvan geven ze de meest voorkomende combinaties aan. Het is gemakkelijk te zien dat van een kleine dataset al veel combinaties gemaakt kunnen worden. Zo heeft een dataset van vier variabelen al vijftien mogelijke combinaties. Het is dus belangrijk alleen te kijken naar de meeste betrouwbare en meest voorkomende combinaties. Daarnaast zijn er combinaties die niets zeggend zijn zoals “zaterdag en zondag vallen in het weekend”. Dit verband is erg betrouwbaar maar voegt niets aan kennis toe.

Association rules worden meestal toegepast op categoriale ofwel nominale attributen. Het is echter ook mogelijk numerieke variabelen mee te nemen. Deze numerieke waarden moeten dan eerst ingedeeld worden in klassen.

4.4.1 Visualisatie

Association rules kunnen alleen weergegeven worden in tekst. Een voorbeeld van association rules van de gegevens over het weer is:

```
temperatuur=koud 4 ==> luchtvochtigheid=normaal 4             conf:(1)
weer=zonnig tennis_spelen=nee 3 ==> luchtvochtigheid=hoog 3   conf:(1)
```

De eerste regel betekent dat als de temperatuur koud is, dat de luchtvochtigheid normaal is. De tweede regel geeft aan dat als het zonnig is en er geen tennis gespeeld wordt, de luchtvochtigheid hoog is. In Weka wordt aan iedere kant van de pijl een *frequentie* weergegeven. Achter de regel staat de *betrouwbaarheid* van de gevonden relatie. Op welke manier die waarden ontstaan staat in de volgende paragraaf uitgelegd.

4.4.2 Prestatie

Zoals gezegd is het belangrijk de resultaten te beperken tot de regels met een hoge frequentie en een hoge betrouwbaarheid. Hoe hoger de frequentie en de betrouwbaarheid, hoe beter de prestatie.

Frequentie — De frequentie geeft aan hoe vaak de gevonden relatie voorkomt in de dataset. De frequentie kan ook worden uitgedrukt in percentages.

Betrouwbaarheid — De betrouwbaarheid van een relatie X ==> Y wordt bepaald door het aantal keer dat zowel X als Y in een relatie voorkomt, ten opzichte van het aantal keer dat X voorkomt [*3*].

In het voorbeeld is de betrouwbaarheid van de tweede regel gelijk aan 1. Dat komt doordat de combinatie 'weer = zonnig' en 'tennis spelen = nee' drie keer is gevonden, terwijl in al die gevallen de luchtvochtigheid hoog is, dus de frequentie van de gehele tweede regel is ook gelijk aan drie. Drie delen door drie geeft 1, ofwel een 100% betrouwbare regel. Of het gevonden verband dan ook nuttig is blijft nog de vraag, daar is kennis van de gegevens voor nodig.

4.5 Clustering

Als er geen specifieke doelvariabele is waarop geclassificeerd kan worden, kan clustering worden gebruikt om de gegevens te groeperen. Het groeperen wordt gedaan door overeenkomsten in de dataset te vinden. De groepen worden clusters genoemd. Het aantal clusters is niet van tevoren bekend, dit verschilt per toegepaste methode. Hoe meer gedetailleerd een clustering is, ofwel hoe meer clusters, hoe waarschijnlijker het is dat de clusters zullen veranderen over de tijd.

Clusters worden gemaakt op basis van afstanden. De afstand kan op verschillende manieren berekend worden. Vaak wordt de zogenoemde Euclediaanse afstand genomen. De afstand tussen object *i* en object *j* wordt gegeven door:

$$d_{ij} = \sqrt{\sum_k (w^{(i)}{}_k - w^{(j)}{}_k)^2}\ , \qquad (1)$$

waarbij $w^{(i)}{}_k$ de waarde van object *i* is voor attribuut *k* en $w^{(j)}{}_k$ de waarde van object *j* voor datzelfde attribuut *k*.

Het is belangrijk dat de afstanden van iedere variabele even zwaar mee tellen. Het verschil in leeftijd kan bijvoorbeeld wel honderd zijn terwijl het verschil tussen het aantal kinderen dat iemand heeft niet boven de acht komt. Om te voorkomen dat leeftijd dan een grotere afstand heeft dan aantal kinderen, moeten de variabelen genormaliseerd worden. Dit gebeurt met de formule:

$$w\_norm_k = \frac{w_k - min(w_k)}{max(w_k) - min(w_k)} \qquad (2)$$

Hierin is w_k de werkelijke waarde van attribuut *k* en w_norm_k de genormaliseerde waarde daarvan. Het maximum en minimum worden bepaald over alle waarden van attribuut *i*.

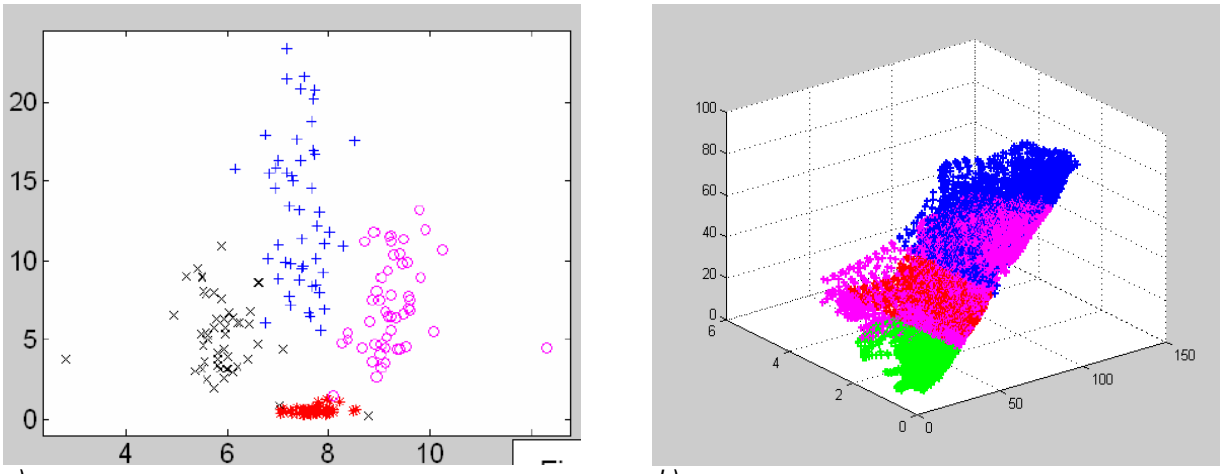
Voorbeeld Een persoon is 75 jaar en heeft drie kinderen. De minimum leeftijd die in de dataset voorkomt is negen en de maximum leeftijd is honderd. Het aantal kinderen loopt van nul tot acht. De persoon krijgt voor leeftijd de genormaliseerde waarde (75-9) / (100-9) = 0,725 en het aantal kinderen wordt (3-0) / (8-0) = 0,375. Zo komen alle waarden van leeftijd tussen nul en één te liggen en tellen alle afstanden even zwaar mee.

Categoriale waarden zorgen voor een ander probleem. Het meten van afstanden tussen categoriale waarden zoals bruin en blauw is niet alledaags. Een manier om hiermee om te gaan is alleen de afstanden 0 en 1 te gebruiken. De afstand is 1 als de waarden verschillen en 0 als ze hetzelfde zijn. De afstand tussen groen en geel wordt op die manier even groot als die tussen geel en zwart.

Missende waarden worden bij clustering op een speciale manier gehanteerd. Categoriale missenden worden als aparte waarde gezien en de afstand tussen een willekeurige waarde en een missende waarde is dus 1. Ook wordt de afstand tussen twee missende waarden op 1 gezet omdat ze verschillend kunnen zijn. Voor twee numerieke missende waarden wordt de afstand ook op 1 gezet. Maar als slechts één van de waarden missend is dan wordt de afstand gelijk gemaakt aan de waarde die wel bekend is of aan 1 min die bekende waarde (de grootste van die twee wordt genomen).

4.5.1 Visualisatie

Clusters visualiseren is over het algemeen lastig. Een uitzondering is het geval dat er slechts twee of drie variabelen worden gebruikt om de gegevens te groeperen. Het is dan mogelijk een twee- of driedimensionale figuur te tonen zoals in figuur 4-3 te zien is.

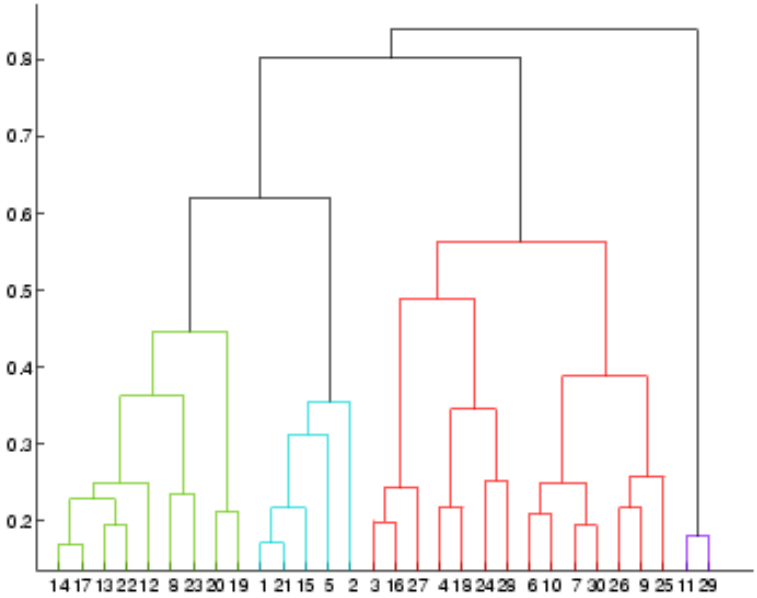


a) *b)*

Figuur 4-3: Links een tweedimensionale en rechts een driedimensionale visualisatie van een clustering

In figuur a zijn twee variabelen tegen elkaar uitgezet waardoor duidelijk vier clusters te zien zijn. Er is als het ware een scheidingslijn tussen te trekken. Maar als er meer dan drie variabelen gebruikt worden is het niet meer mogelijk een dergelijke figuur te maken tenzij een benaderingsmethode gebruikt wordt [*24*].

Er zijn nog andere manieren om een cluster te visualiseren, maar deze worden niet door Weka ondersteund. Een voorbeeld van een andere visualisatie is een *dendrogram* die objecten hiërarchisch met elkaar verbindt (zie figuur 4-4). De verticale as geeft de afstand aan tussen de objecten die met elkaar verbonden zijn [*21*].



Figuur 4-4: Voorbeeld van een dendrogram

Clusters kunnen ook worden beschreven in centrummaten. Het centrum van een cluster kan aangegeven worden met de centroid of de medoid. De centroid geeft de gemiddelde waarde van het cluster aan. In sommige technieken wordt de medoid gebruikt, die het middelste getal aan geeft. Ook kan voor ieder cluster bekeken worden wat de gemiddelde waarde is van ieder numeriek attribuut. Zo kan in het ene cluster de gemiddelde leeftijd 30 zijn en in de andere 55. Voor nominale variabelen kan de modus gebruikt worden als centrummaat. Het is mogelijk dat de centrummaten van twee attributen in verschillende clusters niet (veel) verschillen, dan is de scheiding van de clusters gebaseerd op andere attributen. Om te zien op welke attributen de clusters zijn gescheiden, is het handig om een beslissingsboom te maken (met behulp van classification) met als doelvariabele het clusternummer.

4.5.2 Prestatie

Wanneer een clustering als goed te beschouwen is, valt moeilijk te zeggen. Bij een grafische weergave als in figuur 4-3 is het duidelijk te zien. Hoe meer de verschillende objecten gescheiden zijn, hoe beter de clustering. Maar bij meer dan drie variabelen wordt de afstand tussen de clusters weergegeven in een getal, de log-likelihood genoemd. De formule van de log-likelihood bij twee clusters en *n* objecten is:

$$\log\left[[p_A P(x_1 \mid A) + p_B P(x_1 \mid B)][p_A P(x_2 \mid A) + p_B P(x_2 \mid B)]...[p_A P(x_n \mid A) + p_B P(x_n \mid B)]\right], \qquad (3)$$

waarbij A en B de twee clusters zijn en p_A en p_B de kansen zijn dat een object in cluster A respectievelijk cluster B terecht komt. De objecten zijn gelabeld met x_1, x_2,…,x_n.
Hoe groter de resulterende waarde is, hoe beter de clustering. De ene dataset is echter goed geclusterd als de log-likelihood gelijk is aan 5 terwijl de ander goed is bij de waarde 50076. Er is dus alleen iets over de prestatie te zeggen als het om een vergelijking van clustertechnieken van dezelfde dataset gaat. Het is aan de expert van de data om te zeggen of de gemaakte clusters zinvol zijn.

4.6 Datamining voor strategische analyse

Deze paragraaf beschrijft wanneer dataminingtechnieken goed van pas komen en welke het beste gebruikt kunnen worden. De toepassingen zijn toegespitst op de vraagstellingen van de afdeling Analyse & Research.

De keuze van de te gebruiken technieken hangt grotendeels af van twee componenten:
1) Wat wordt gevraagd, waar is men naar op zoek?
2) Hoe zien de gegevens er uit?

4.6.1 Vraagstellingen

In het vooronderzoek is nagegaan welk soort onderzoeksvragen bij de politie leven om te kunnen inschatten waar datamining in de toekomst voor gebruikt kan worden. Allereerst kunnen de onderzoeksvragen binnen de politie in de volgende categorieën ingedeeld worden [*1*]:

- Beschrijvende vraag: Hoe ziet het probleem eruit en hoe ontwikkelt het zich?
- Verklarende vraag: Wat zijn de oorzaken of gevolgen van het probleem?
- Voorspellende vraag: Hoe zal het probleem er in de toekomst uit zien?
- Evaluerende vraag: Welk effect hebben de maatregelen teweeggebracht en waardoor?
- Prescriptieve vraag: Welke maatregelen moeten we treffen om het probleem aan te pakken?

Afdeling Analyse & Research houdt zich vooral bezig met het beantwoorden van beschrijvende vragen van tactische en strategische aard. Meestal wordt dat uitgedrukt in de vraag:

"Wat is de aard, omvang, spreiding en ontwikkeling van de geregistreerde criminaliteit in regio Haaglanden?"

Deze vraag kan worden toegepast op vele soorten criminaliteit zoals diefstal, mishandeling en moord, maar ook op verschillende geografische niveaus zoals regio, gemeente, wijk en buurt. De vraag kan ook vanuit de kant van de verdachten benaderd worden zoals vragen over geweldplegers, jeugd of veelplegers. Het komt ook voor dat de aard, omvang, spreiding en ontwikkeling van slachtoffers wordt gevraagd.

Onder de aard van criminaliteit vallen vragen als: "Wat zijn de persoonsgebonden- en achtergrondkenmerken van inbrekers in woningen?", "Wat zijn de kenmerken van overvallen in de wijk?". Ook wordt gekeken naar welke criminele carrière de veelplegers in de regio hebben. De omvang betreft het aantal inbraken in de wijk, aantal verdachten per 1000 inwoners en hoe het aantal inbraken zich tot het aantal inbrekers in de wijk verhoudt. Spreiding kijkt vervolgens naar de geografische spreiding van verdachten en delicten. Er wordt onder andere gekeken naar het aantal delicten op een bepaalde plaats. Zo kan men de zogenoemde 'hotspots' analyseren, plaatsen waar veel overlast is. De verandering in tijd wordt verwoord in de ontwikkeling van de geregistreerde criminaliteit. Meestal gaat het om vergelijkingen tussen jaren, kwartalen of maanden.

Naast deze veel omvattende vraagstelling worden er nog enkele andere analyses verricht. Een bekend product is de veelpleger-top500. Dit is een lijst van 500 minderjarige criminelen die het meest actief zijn in regio Haaglanden. Deze personen worden in het bijzonder in de gaten gehouden door de politie.

Veelal worden de analyses weergegeven door middel van frequentietabellen en grafieken. Regelmatig worden ook kruistabellen gebruikt om verbanden weer te geven. Correlatie en toetsen voor significantie worden soms toegepast, regressie wordt zelden gedaan.

4.6.2 Gegevens

Er zijn drie standaard datasets die voor de meeste analyses voldoende zijn. De variabelen die daarvan vaak gebruikt worden voor analyses aan de personenkant zijn: geslacht, leeftijd, leeftijd ten tijde van eerste antecedent, etniciteit, woonplaats, aantal antecedenten, typologie (veelpleger / doorstromer / beginner), geboorteland, alcoholgebruik, harddruggebruik. Door vakkennis en ervaring hebben de analisten deze variabelen gekozen om de standaard analyses mee uit te voeren. Andere variabelen worden soms wel meegenomen voor meer diepgaande analyses.

Sommige dataminingtechnieken zijn alleen geschikt voor numerieke waarden (regressie), sommige juist alleen voor categoriale (association rules). Categoriale variabelen hebben een nominale of ordinale schaal. Een nominale schaal is een uiteenzetting van categorieën zoals man/vrouw, Nederlands/Turks/overig, 0/1. Een ordinale schaal heeft ook categorieën maar met als verschil dat er een oplopende of aflopende volgorde in zit. Voorbeelden zijn 0/1/2/3 en helemaal eens/eens/neutraal/oneens/helemaal oneens.
Daarnaast zijn er numerieke variabelen, die een interval schaal of een ratio schaal hebben. Bij een variabele met interval schaal is de afstand tussen de waarden betekenisvol maar er is geen natuurlijk

nulpunt. Bijvoorbeeld jaartallen. Variabelen met een ratio schaal zijn numerieke waarden waar de afstand betekenisvol is en er een natuurlijk nulpunt is zoals gewicht en lengte.

Op de afdeling worden vooral nominale variabelen gebruikt voor de analyses. Er zijn echter ook numerieke (ratio) variabelen zoals de rubrieken en de jaren die het aantal antecedenten of feiten aangeven per rubriek of per jaar. De dataminingtechnieken moeten dus voor alle soorten gegevens geschikt zijn.

4.6.3 Keuze technieken

Op basis van de meest gebruikte vraagstelling zoals hierboven genoemd, volgt hier een aanbeveling voor de te gebruiken technieken.

Aard → classification, association rules en clustering
Met classification kan een boom geconstrueerd worden om te zien welke persoonskenmerken horen bij een bepaald delict. Er kan een vergelijking mee worden gedaan: wie pleegt wel een woninginbraak en wie niet. Daarnaast is nog keuze tussen het gebruik van een techniek die een boom construeert of regels maakt. Voordeel van een boom is dat er een visueel beeld ontstaat dat gemakkelijk is te interpreteren.
Een andere manier is association rules. Als men wil weten hoe een groep inbrekers eruit ziet dan dient men alleen die inbrekers in de dataset op te nemen. Vervolgens kunnen er exploratief verbanden in worden gezocht. Op die manier komt men gemakkelijk dingen te weten die zonder voorafgaande hypothesen gedaan kunnen worden. Vaak worden hiermee relaties gevonden die niet eerder zijn opgevallen.
Daarnaast kan clustering gebruikt worden om verborgen groepen te vinden in de doelgroep. Zo kan een groep inbrekers, op basis van een aantal persoons- of delictkenmerken, bestaan uit meerdere duidelijk samenhangende groepen. Zo kan men uitvinden wie wat heeft gedaan.

Omvang → frequentietabellen en kruistabellen
Om te meten hoe groot een bepaalde groep is kan het beste een standaard frequentietabel worden gemaakt. Ook kan met kruistabellen de frequentie in subgroepen bekeken worden. Van te voren is bekend van welke groep de grootte gemeten moet worden en daarom is het niet nodig een datamining techniek te gebruiken.

Spreiding → kruistabellen en classification
Om spreiding van bijvoorbeeld veelplegers over bureaugebieden te bepalen is het maken van een kruistabel de gemakkelijkste oplossing. Daarnaast is het mogelijk met classification te bekijken of veelplegers over het algemeen wel of niet actief zijn in een gebied. Op die manier kunnen zogenoemde ‘hotspots’ ontdekt worden.

Ontwikkeling → regressie en classification
Om de ontwikkeling in de tijd te bekijken kunnen er statistische regressietechnieken gebruikt worden. Ook zijn er regressietechnieken in de datamining aanwezig. Om twee groepen van verschillende jaren met elkaar te vergelijken kan classification gebruikt worden om snel verschillen te vinden.

4.7 Software

Er is veel software op de markt voor datamining. De meeste pakketten bevatten echter maar één van de componenten clustering, classification en association rules. Daarnaast is niet alle datamining software even gebruiksvriendelijk. Na onderzoek wat er bij andere regio's van de politie wordt gebruikt, is er een selectie gemaakt. Deze wordt hieronder kort beschreven.

4.7.1 Beschrijving software

SPSS Classification Trees[1]
Leverancier: SPSS
URL: www.spss.com/classification_trees
Functies: Classification
Technieken: Beslissingsbomen (CHAID, Exhaustive CHAID, C&RT en QUEST)
Kosten: Eerste gebruiker € 825,-, vanaf het tweede jaar € 185,-

Classification Trees wordt gebruikt voor het maken van beslissingsbomen. Het is een datamining tool die groepen profileert voor marketing en sales. De CHAID technieken zijn geschikt voor zowel nominale, ordinale als continue variabelen.
Deze module wordt onder andere gebruikt bij het WODC.

SPSS Clementine
Leverancier: SPSS
URL: www.spss.com/clementine
Functies: Classification, clustering, association rules, factor analyse, voorspellen, sequence discovery
Technieken: Apriori, BIRCH, CARMA, Decision trees (C5.0, C&RT), K-means clustering, neural networks (Kohonen, MLP, RBFN), regression (linear, logistic), rule induction (C5.0, GRI)
Kosten: Op aanvraag (circa € 50.000,- serverversie)

Clementine heeft een Windows interface met drag en drop mogelijkheden. De gebruiker hoeft geen ingewikkelde commando's in te tikken maar kan alles met de muis besturen. De complexiteit van de verschillende technieken is verborgen. Ook een gebruiker met minder kennis van de technieken kan daardoor met het pakket werken. Voor deskundigen zijn er steeds 'expert' opties, waarmee de details van de technieken goed te beïnvloeden zijn. Clementine behoort tot de leiders in de markt van datamining software. Het procesmodel CRISP-DM, beschreven in paragraaf 4.1, is ingebakken in Clementine.
Een aantal regio's van de politie hebben Clementine gebruikt of gebruiken het nog steeds.

DataDetective
Leverancier: Sentient
URL: www.sentient.nl/datadetective
Functies: Fuzzy zoeken, kruistabellen, genereren van hotspotkaarten, clustering, relatie-analyse, beslissingsbomen
Kosten: Op aanvraag (circa € 20.000,- per jaar)

DataDetective richt zich vooral op grafische aspecten. Het is gekoppeld aan MapInfo, een applicatie die bij de politie wordt gebruikt om hotspot-kaarten te maken. DataDetective bevat geen classification technieken zoals C4.5. De beslissingsbomen worden gesplitst op variabelen die door de gebruiker opgegeven dienen te worden.
Met fuzzy zoeken wordt gezocht naar de meest overeenkomende incidenten of personen voor opsporingsdoeleinden. Relatie-analyse kan criminele netwerken in kaart brengen.
DataDetective wordt gebruikt bij Politie Amsterdam-Amstelland.

[1] Classification Trees heette voorheen AnswerTree. AnswerTree was een apart product terwijl Classification Trees gemakkelijk als module aan te schaffen is.

Weka

Leverancier:	University of Waikato, New Zealand
URL:	www.cs.waikato.ac.nz/ml/weka
Functies:	Classification, filter (preparatie van data), clustering, association rules, attribute selection
Technieken:	Decision trees, rule learners, model tree generators, support vector machines, locally weighted regression, instance-based learning, bagging, boosting, stacking, EM clustering, k-means, association rules
Kosten:	Gratis

Weka heeft verschillende interfaces waarvan er twee volledig ontwikkeld zijn. De zogehete "Explorer" werkt menu gestuurd en de "simpleCLI" werkt met behulp van commando's. Verder bevat Weka een "Experimenter" waarin verschillende classifications met elkaar vergeleken kunnen worden.
Weka is vooral gericht op classification en *filter* algoritmen. Daarnaast bevat het ook algoritmen om association rules te vinden en clustering waarbij geen doelvariabele opgegeven hoeft te worden. Alle algoritmen zijn geschreven in Java en zijn ook direct aan te roepen vanuit eigen programmacode. Weka wordt beschreven in het boek van Ian H. Witten en Eibe Frank [*31*].
Deze applicatie wordt nog niet gebruikt bij de politie in Nederland. Wel wordt deze veel gebruikt door studenten.

4.7.2 De keuze

Uiteindelijk is Weka gekozen om dit onderzoek te voltooien. Dit is gratis te downloaden software die vele soorten technieken in zich heeft. Het pakket is niet bedoeld voor de beginnende dataminer maar wel zeer geschikt voor onderzoek naar de werking en prestatie van verschillende algoritmen. Een groot voordeel is dat het freeware is en aangezien datamining bij de Politie Haaglanden nog in de kinderschoenen staat is men (nog) niet bereid direct een relatief duur pakket aan te schaffen. Een ander voordeel is dat Weka veel opties heeft en al bekend is bij de stagiair.

Voor de toekomst van datamining bij de politie Haaglanden is Clementine van SPSS aan te raden. Dit product is leider op de markt en behelst vele dataminingtechnieken. Een groot voordeel van Clementine is dat visueel duidelijk is welke stappen genomen worden in het gehele databewerkings-proces. In het prepareren van gegevens is Clementine ook sterk. Bovendien zijn er expertopties zodat altijd inzichtelijk blijft wat er 'achter de schermen' gebeurt. De hoge kosten zijn echter een nadeel. Misschien is er een mogelijkheid tot het sluiten van een raamcontract zodat meerdere korpsen gelijktijdig van de software gebruik kunnen maken.

Aangezien Classification Trees van SPSS alleen geschikt is voor classification en niet voor association rules en clustering, is het beperkt bruikbaar voor datamining. Een voordeel is dat er door enkele medewerkers op het bureau enige ervaring is opgedaan met dit product. Als er alleen vraag is naar classification dan is dit zonder meer een geschikte keuze. Zeker nu het als module in het standaard SPSS pakket wordt aangeboden.

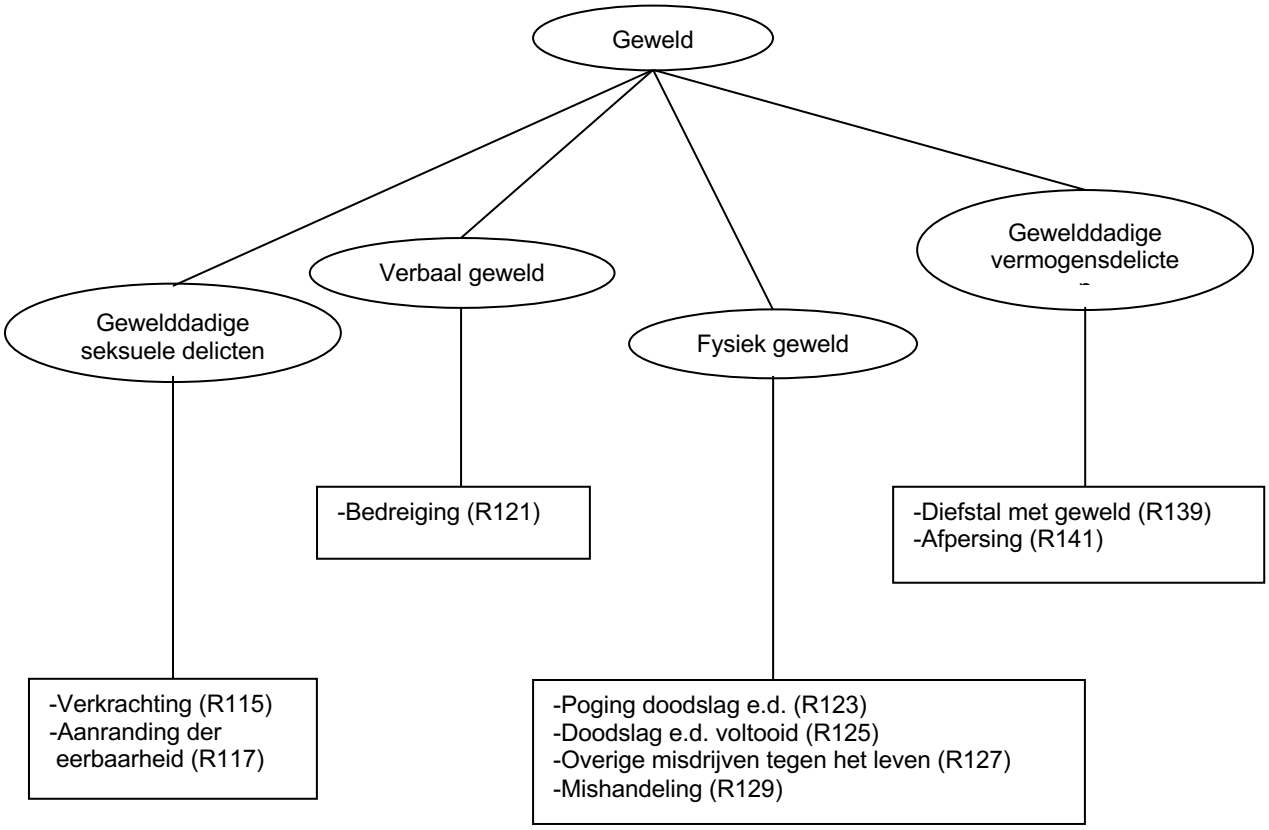
5. De dataset

“There are three kinds of lies: lies, damned lies, and statistics.”
– Mark Twain, Autobiography.

In dit hoofdstuk wordt de gebruikte dataset beschreven. Allereerst wordt de definitie van geweld gegeven zoals die in dit onderzoek gebruikt is en vervolgens worden de bron en de gebruikte variabelen beschreven. In paragraaf 5.4 is een beschrijving van enkele variabelen gegeven.

5.1 Wat is geweld?

De aanpak van geweld is korpsprioriteit, dat wil zeggen dat het korps dit onderwerp als belangrijk ziet en dat er veel aandacht aan besteed wordt. Er zijn echter verschillende definities van geweld. Van Dale geeft de betekenis: "*uitoefening van macht"* en "*kracht die met hevigheid, onstuimigheid wordt uitgeoefend".* Waar echter de grens ligt tussen geweld en bijvoorbeeld een scheldpartij is moeilijk aan te geven. Voor dit onderzoek worden de volgende categorieën onder geweld verstaan: seksuele delicten met geweld, verbaal geweld, fysiek geweld en vermogensdelicten met geweld. Een overzicht van welke rubrieken onder de definitie van geweld vallen is te zien in het volgende schema.



Figuur 5-1: Overzicht van de rubrieken die onder 'geweld' vallen

De wetsartikelen die bij de rubrieken horen zijn gegeven in bijlage A.

5.2 Bron

Voor het onderzoek is gebruik gemaakt van gegevens uit het opsporingsregister van de herkenningsdienst, het zogenaamde HKS. Hierin bevinden zich gegevens omtrent misdrijven die ter kennis zijn gekomen van de politie en gegevens omtrent verdachten en hun criminele verleden. Het HKS wordt binnen regio Haaglanden vaak gebruikt voor analyses vanwege de betrouwbaarheid en goede structuur. Ook wordt het gebruikt voor het verrichten van zoekactiviteiten in het kader van opsporingsonderzoek.

De gemaakte extractie bestaat uit alle personen die in de periode 01-01-2004 tot en met 31-12-2004 voorkomen met een of meer antecedenten. Van deze personen is hun volledige delictgeschiedenis voor zover deze in het HKS is opgeslagen bekend. Dat wil zeggen dat ook de antecedenten van voor 01-01-2004 in het bestand voorkomen.

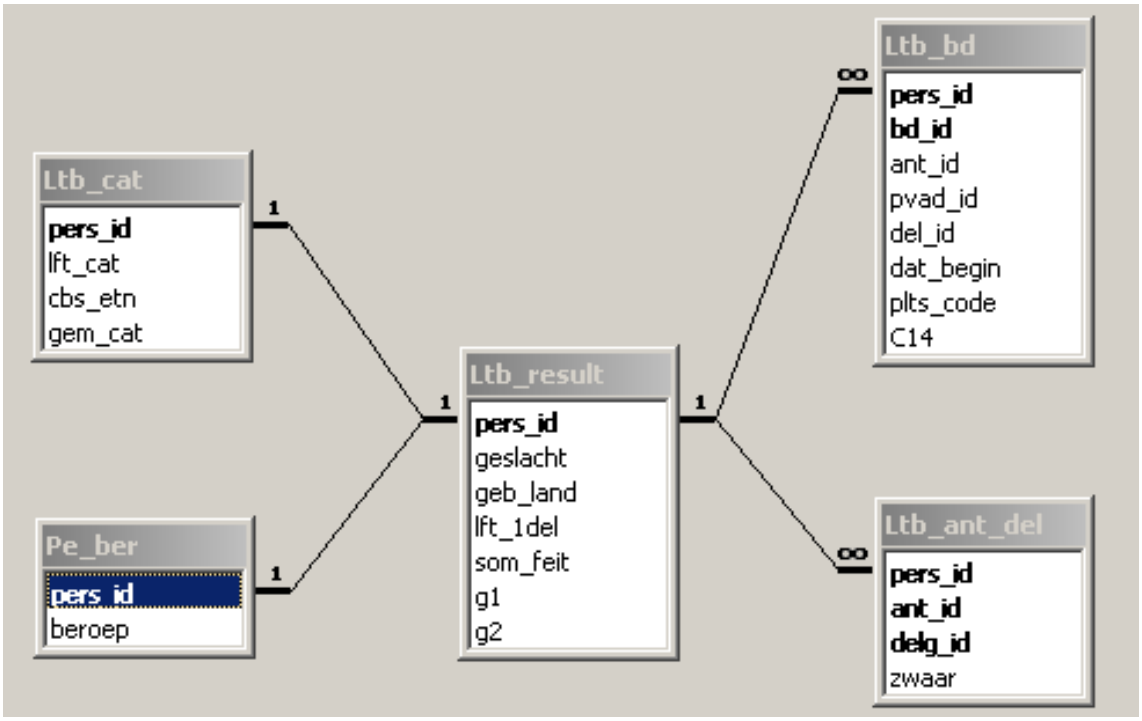
Voor criminaliteitsanalyse kent HKS de Data Extractiemodule (Dex2000), waarmee de analist HKS gegevens kan inlezen in bijvoorbeeld SPSS of Excel. Dex2000 genereert losse tabellen waarvan ltb_result de meest gebruikte is. In die tabel staat informatie over de verdachten zoals woonplaats, geboorteland, etniciteit, leeftijd en geslacht. Er staan geen persoonsgegevens in als naam en adres. Verder staat erin hoeveel antecedenten en feiten de verdachte heeft gepleegd, ook per jaar en per rubriek bekeken. Verder is het in versie 6 van Dex2000 mogelijk een bekende dader tabel te maken die een aantal details van delicten weergeeft zoals plaats en tijd van pleging. Hierin is de persoon niet meer uniek omdat een persoon meerdere feiten op zijn naam kan hebben.

Naast het creëren van bewerkbare datasets voert Dex2000 ook een aantal operaties op de data uit. Zo neemt de module alleen de zwaarst wegende tenlastelegging uit HKS mee in de dataset. Dat wil zeggen dat bij een cq-relatie zoals zware mishandeling cq. mishandeling de zwaarste blijft staan en dus mishandeling weg valt. Verder worden een aantal delicten en/of verdachten verwijderd omdat deze foutief zijn ingevoerd. In foutieve gevallen ontbreekt bijvoorbeeld het gepleegde feit.

In het logbestand van Dex2000 is te vinden welk percentage van de dataset automatisch is gekoppeld, dat wil zeggen dat zowel de persoonsgegevens als de delictgegevens bekend zijn. Bij de extractie die voor deze analyse is gebruikt is dat 85,4% van de personen. Van de overige personen zijn alleen de persoonskenmerken bekend en het wetsartikel dat ze hebben overtreden. Geen verdere details over het gepleegde misdrijf zoals plaats, tijd en modus operandi zijn in dat geval gegeven.

5.3 Gebruikte variabelen

De database van HKS bestaat uit ongeveer 40 afzonderlijke gegevenstabellen die via allerlei sleutels aan elkaar gekoppeld zijn. Niet al deze tabellen zijn voor het onderzoek interessant. Daarom is er een selectie gemaakt van tabellen die voor dit onderzoek interessante informatie bevatten. In onderstaande figuur is weergegeven welke tabellen zijn gebruikt en hoe deze aan elkaar gekoppeld zijn.



Figuur 5-2: Weergave van de koppeling van gebruikte tabellen

De tabel ltb_result is de tabel die de meeste gebruikte variabelen bevat. Deze tabel wordt vaak gebruikt bij analyses op de afdeling. De tabel is opgevraagd voor alle aangehouden verdachten van 2004. De tabel ltb_cat (categorieën) bevat variabelen die zijn afgeleid van andere variabelen. Zo is de variabele *cbs_etn* (etniciteit volgens definitie van het CBS) afgeleid van het geboorteland en de nationaliteit van een persoon en *gem_cat* geeft de grootte van de woonplaats van de persoon aan. Deze tabel is net als ltb_result op persoonsniveau en de sleutel is *pers_id*. Om daarnaast de variabele beroep te kunnen gebruiken is de tabel pe-ber gekoppeld met *pers_id* als sleutel.

Ltb_ant_del (antecedenten en delicten) en ltb_bd (bekende dader) bevatten beide meerdere voorkomens van één persoon. In de tabel met antecedenten en delicten is de sleutel de combinatie van *pers_id*, *ant_id* en *delg_id*. In de bekende dader tabel vormen *pers_id* samen met *bd_id* het unieke veld. De enige gebruikte variabele die uit de antecedenten en delicten tabel is *zwaar*, die aangeeft of het een zwaar delict betrof, op basis van vastgelegde definities van het Openbaar Ministerie. De bekende dader tabel is alleen opvraagbaar met versie 6 van Dex2000. Hier staan interessante variabelen in zoals de modus operandi van ieder delict. Het was niet eerder mogelijk deze tabel met zowel persoonsnummers als delictdetails op te vragen.

In bijlage B is een overzicht van alle gebruikte variabelen per tabel gegeven, met daarbij de betekenissen.

5.4 Voorbereiden van de dataset

Het voorbereiden van de gegevens is een belangrijke en tijdrovende stap in het datamining proces. Het is belangrijk omdat de resultaten van het datamining significant kunnen verbeteren door het transformeren en selecteren van de data.

5.4.1 Transformeren

Omdat er ook interesse is naar delictkenmerken van de verdachten, is er een structuur ontstaan waarin het persoonsnummer niet meer uniek is. Als een persoon 20 feiten heeft gepleegd dan komt die persoon ook 20 keer voor in de dataset. De resultaten van de analyse zullen hierdoor sterk worden beïnvloed. Het aantal delicten kan namelijk oplopen tot wel 300. Het gevolg daarvan is dat de kenmerken van die persoon, zoals leeftijd, ook 300 keer meegerekend zullen worden. Door het veranderen van de structuur is dit probleem aangepakt. Aan de hand van een fictieve variabele *wapen* wordt beschreven hoe dit in zijn werk gaat.

De huidige structuur ziet er uit als in tabel 5-8 te zien is. Iedere rij stelt een gepleegd feit voor.

Pers_id	Bd_id	Wapen
01	01	Steekwapen
01	02	Geen
01	03	Steekwapen
02	01	Vuurwapen
03	01	Vuurwapen
03	02	Steekwapen
…	…	…

Tabel 5-8: Voorbeeld van de huidige structuur van de dataset

Een manier om de personen uniek te maken is om nieuwe variabelen op te nemen, genaamd *steekwapen*, *vuurwapen* en *geen*. De waarden in de kolom *steekwapen* stellen dan het aantal keer voor dat de betreffende persoon een steekwapen heeft gebruikt. De structuur komt er dan uit te zien zoals in tabel 5-9.

Pers_id	Steekwapen	Vuurwapen	Geen
01	2	0	1
02	0	1	0
03	1	1	0
…	…	…	…

Tabel 5-9: De nieuwe structuur van de dataset

Alle variabelen van de bekende dader tabel en de antecedent en delict tabel zijn op deze manier geaggregeerd.

5.4.2 Hoeveelheid variabelen beperken

De dataset bestaat uit 211 variabelen, ook wel attributen genoemd. Die attributen zijn niet allemaal even interessant en er zijn variabelen die bijna dezelfde informatie geven. Als de gegevens die onderzocht worden een groot aantal attributen bevat, kan de runtijd van een datamining algoritme gereduceerd worden door het algoritme op een selectie van de attributen toe te passen. Een andere reden om een selectie te kiezen is dat sommige algoritmen erg gevoelig zijn voor irrelevante variabelen of voor variabelen die van een andere zijn afgeleid. Een algoritme zou slechte resultaten kunnen geven door zulk soort variabelen er in te laten staan.

Aan de andere kant is bij datamining belangrijk dat alle opties open gehouden worden. Het zou zonde zijn een variabele uit te sluiten van onderzoek terwijl er juist een heel onverwachte ontdekking mee gedaan zou kunnen worden. De selectie van variabelen is dus een belangrijk onderdeel van datamining. Het beperken van het aantal variabelen is op twee manieren te realiseren: handmatige selectie en automatische selectie.

Handmatige attribuutselectie

Een voorbeeld van een variabele die van andere variabelen is afgeleid is *versl_cat* die aangeeft in welke categorie verslaving de persoon valt. Deze is afgeleid uit de variabelen *g1* (alcoholist) en *g2* (harddruggebruiker). Zo is ook het jaartal van het eerste delict af te leiden uit de geboortedatum en de leeftijd ten tijde van het eerste delict. Op die manier zijn meerdere variabelen buiten beschouwing gelaten. In bijlage B staat een overzicht van alle geselecteerde variabelen.

In de dataset zijn twee grote groepen variabelen. Allereerst de groep variabelen met namen als *jr_81*. Deze geven het aantal antecedenten aan dat de betreffende verdachte in dat jaar heeft gepleegd. De tweede grote groep bestaat uit rubrieken zoals *R101*. Wetsartikelen vallen onder hoofdrubrieken en sommige hoofdrubrieken zijn weer op te splitsen in subrubrieken. De variabelen geven het aantal feiten aan dat de verdachte in die rubriek heeft gepleegd. De twee groepen variabelen zijn in de dataset opgenomen vanwege het feit dat men nieuwsgierig is naar mogelijke patronen.

Ten slotte is er nog de variabele *typol3*, die de typologie van een verdachte geeft: beginner, doorstromer of veelpleger. De typologie is gebaseerd op het aantal antecedenten en is dus volledig afhankelijk van *som_ant*. Er is echter toch gekozen om deze variabele te behouden omdat de termen beginner, doorstromer en veelpleger veel gebruikte begrippen zijn. Een andere reden is dat de automatische selectie die voor classification plaats zal vinden, niet beide variabelen zal kiezen als dit voor de resultaten van het algoritme nadelig is. De meest bruikbare variabele zal gekozen worden.

Automatische attribuutselectie
De automatische attribuutselectie gebeurt voor het classificeren, één van de dataminingmethoden. Deze techniek is in de meeste datamining software aanwezig. Twee soorten variabelen kunnen de prestaties van de classifiers verslechteren: irrelevante variabelen en afhankelijke variabelen. Een voorbeeld van een afhankelijke variabele is *typol3* en *som_ant* zoals zojuist uitgelegd. Een voorbeeld van een irrelevante variabele is bijvoorbeeld de schoenmaat van een persoon. Deze heeft (naar men vermoedt) geen invloed op het soort delict dat iemand pleegt. De automatische selectie kan op twee manieren uitgevoerd worden: de *filter* methode en de *wrapper* methode.

Bij een filter methode wordt de attribuutverzameling gefilterd om de meest veelbelovende deelverzameling voor het dataminen te produceren. De variabelen die een sterke relatie hebben met de doelvariabele worden geselecteerd. Een wrapper heeft als voordeel dat de te gebruiken classificatietechniek opgegeven kan worden. De wrapper zoekt de beste variabelen uit op basis van die specifieke classifier. Als de te gebruiken classifier bijvoorbeeld niet goed kan omgaan met samenhangende variabelen dan zal de wrapper zulke variabelen niet meenemen in de selectie.

5.4.3 Missende en incorrecte waarden

In de dataset worden missende waarden op verschillende manieren aangegeven. Soms met de waarde 0, soms met '999' en soms met 'ONB'. Bij een aantal variabelen in de dataset komen geen missende waarden voor. Bij andere variabelen komen juist heel veel missende waarden voor zoals bij beroep. Dat komt doordat de verbalisant dit niet altijd vraagt aan de verdachte. De laatste jaren wordt het zelfs nooit meer gevraagd.

De technieken in Weka gaan op verschillende manieren om met missende waarden. Sommigen beschouwen ze als een aparte waarde, sommigen negeren ze. Het is belangrijk de ontbrekende waarden eenduidig weer te geven zodat de technieken ze ook als zodanig zien. Daarom zijn alle onbekende waarden op missing gezet.

Soms is er een verschil tussen missende waarden zoals bij de variabele *gem_cat* die het aantal inwoners van de woonplaats aangeeft. De waarde 0 wil zeggen dat die onbekend is, het getal 8 geeft aan dat de woonplaats in het buitenland ligt en in dat geval is er geen informatie over het aantal inwoners. Het aantal inwoners is in dat geval ook onbekend maar heeft wel meerwaarde ten opzichte van de waarde 0. Er zijn een aantal manieren om hiermee om te gaan. Een manier is het invullen, schatten, van de ontbrekende waarde. In het geval van buitenlandse steden is het moeilijk te schatten. Ze hebben niet dezelfde schaal als Nederlandse steden. Dat is dus geen optie. Bij nominale variabelen zou er een waarde toegevoegd kunnen worden, de waarde 'buitenland'. Maar *gem_cat* is een numerieke variabele. De variabele *land_cd* geeft echter al aan of iemand in het buitenland woont of niet en er is dus geen informatieverlies als de waarde 8 op missend worden gezet, evenals de waarde 0.

Daarnaast dienen incorrecte waarden vervangen te worden door missende waarden of door de correcte waarde. Er is gezocht op uitschieters en daardoor zijn enkele incorrecte waarden aan het licht gekomen. Door het relatief maken van numerieke variabelen (in verhouding tot aantal antecedenten of feiten) zijn er nog eens ongeveer 100 fouten gevonden. Deze personen zijn verwijderd.

5.5 Beschrijving dataset

Voordat het modelleren kan beginnen is het nodig de gegevens te verkennen. Dit houdt het exploratief bekijken van de geprepareerde dataset in om een beter gevoel van de data te krijgen en eventuele fouten te ontdekken.

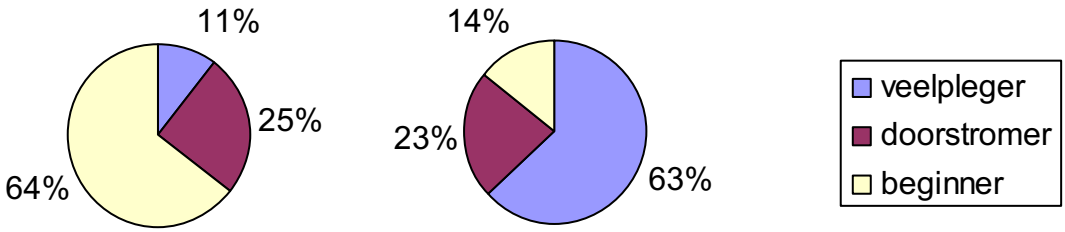
De dataset bestaat uit 18759 rijen die unieke personen voorstellen. Totaal zijn er 211 variabelen waarvan er 14 binair zijn en 186 ratio. De overigen zijn nominaal.

5.5.1 Bureaugebieden

Het werkgebied van Politie Haaglanden bestaat uit 18 bureaugebieden. Daarnaast is er een categorie 'buiten regio'. In Den Haag Zuid-West woont 9,8% van de geregistreerde verdachten. In overige bureaugebieden woont tussen de 1% en 8% van de geregistreerde verdachten. Het grootste deel, namelijk 13,8%, woont echter buiten regio Haaglanden. Van 1027 personen is geen woonplaats geregistreerd, dat komt neer op 5,5%.

5.5.2 Aantal antecedenten

Een verdachte kan ingedeeld worden in drie categorieën op basis van het aantal antecedenten dat hij/zij heeft. Een beginner is iemand die één of twee keer door de politie is aangehouden, een doorstromer drie tot en met tien keer en een veelpleger heeft elf of meer antecedenten. Figuur 5-3 a) laat de verdelingen zien van de typen verdachten ten opzichte van het totaal aantal aangehouden verdachten. Het merendeel van de verdachten is beginner en 'slechts' 11% wordt omschreven als veelpleger. Daar tegenover staat dat in figuur b) te zien is dat die groep veelplegers uit 2004 verantwoordelijk is voor 63% van de antecedenten van 2004 en voorgaande jaren van alle in 2004 aangehouden verdachten.

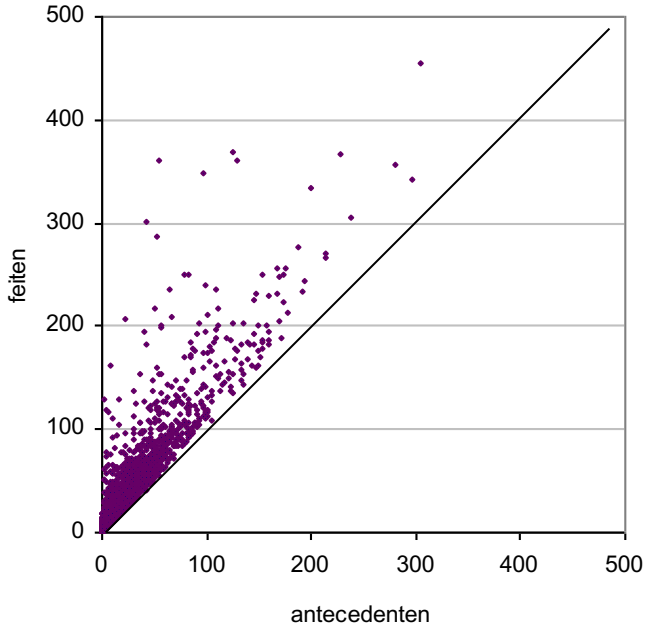


a) Verdachten *b) Hun antecedenten*

Figuur 5-3: Verdeling van de typen verdachten: beginner, doorstromer en veelpleger

5.5.3 Aantal feiten

Het aantal antecedenten staat in relatie tot het aantal feiten dat een verdachte heeft gepleegd. Ze zijn echter niet altijd gelijk. Het is namelijk mogelijk dat een persoon wordt aangehouden op verdenking van inbraak en bij het verhoor nog een andere inbraak bekent. In dat geval heeft die persoon één antecedent (hij is één keer aangehouden) maar twee feiten. In de figuur 5-4 is te zien hoe de twee variabelen samen hangen.

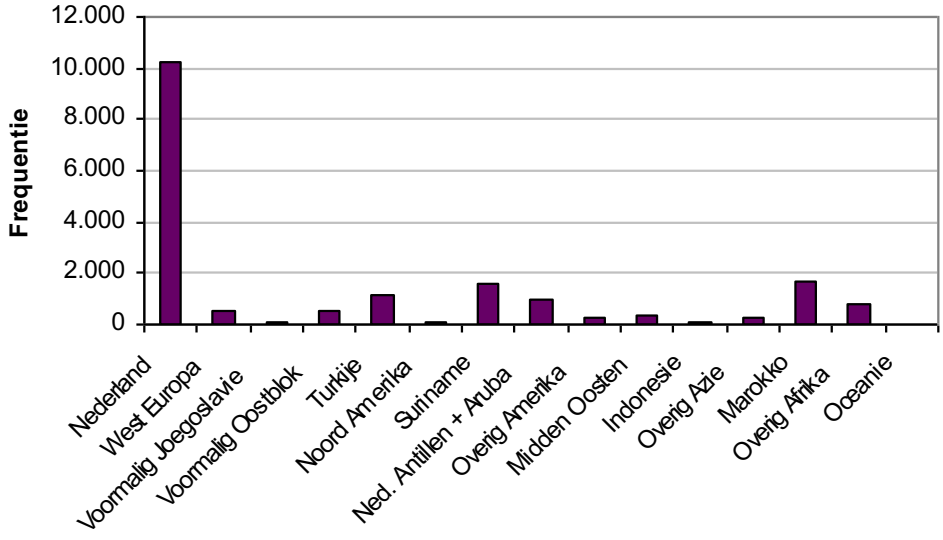


Figuur 5-4: Aantal feiten tegen aantal antecedenten en de lijn y = x.

Zoals verwacht is er een stijgend verband te zien. Het aantal feiten ligt altijd even hoog of hoger dan het aantal antecedenten. De punten die ver afwijken van de lijn $y = x$ hebben relatief veel feiten per antecedent. Het is opvallend dat het aantal feiten een uitschieter heeft, namelijk 455. Deze persoon heeft 304 antecedenten. Een andere extreme waarde is de verdachte die 361 feiten heeft bekend in 'slechts' 54 antecedenten.

5.5.4 Etniciteit

De etniciteit volgens de definitie die het Centraal Bureau voor de Statistiek (CBS) hanteert, wordt bepaald met behulp van de nationaliteit en het geboorteland van de persoon. Buitenlandse etniciteiten hebben hierbij voorrang op Nederlandse. Een voorbeeld: een persoon met de Nederlandse nationaliteit die in Suriname geboren is, heeft de Surinaamse etniciteit. Ook worden verdachten jonger dan 25 jaar bij het bevolkingsregister gecheckt op het geboorteland van de vader en moeder. Dit is echter bij de voor dit onderzoek gebruikte dataset nog niet gebeurd.

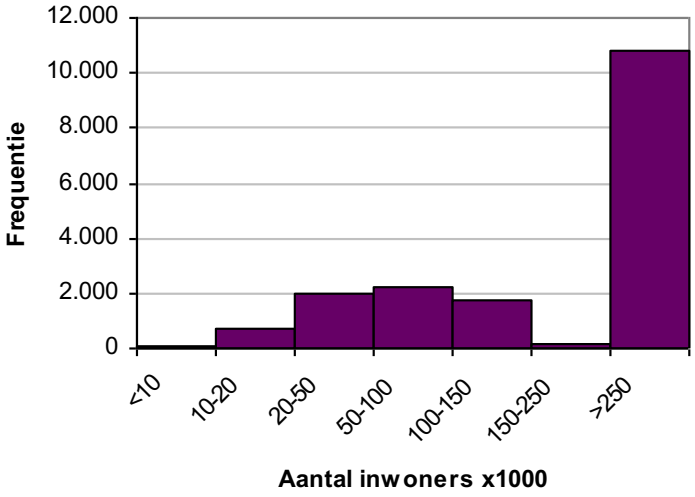


Figuur 5-5: Verdeling van de etniciteiten volgens de definitie van het CBS.

Figuur 5-5 laat zien dat de Nederlandse etniciteit veruit het meest voorkomt, namelijk in 54,6% van de gevallen (10.247). Andere regelmatig voorkomende etniciteiten zijn de Turkse (1.110), Surinaamse (1.604) en Marokkaanse (1.639).

5.5.5 Aantal inwoners woonplaats

De meeste verdachten (57,7%) wonen in een gemeente met meer dan 250.000 inwoners. Dit kan gemeente Den Haag zijn maar ook grote steden buiten de regio. Van de geregistreerde verdachten woonde namelijk 13,8% buiten regio Haaglanden. De verdeling is te zien in onderstaande figuur. Opvallend is dat er weinig verdachten in steden wonen met 150.000 tot 250.000 inwoners. Dat komt doordat er binnen regio Haaglanden geen gemeente is van die orde van grootte. Degene die wel in die klasse behoren, wonen dus buiten de regio.



Figuur 5-6: Verdeling van de verdachten over grootte van gemeenten

5.5.6 Doelvariabele: soort geweld

De doelvariabele geeft aan wat voor type geweldpleger de verdachte is, dan wel dat hij of zij nog nooit geweld heeft gepleegd dat door de politie geregistreerd is. Omdat er in de dataset ook combinaties van soorten geweld voorkomen in de dataset is gekozen voor een binaire notatie, een code bestaande uit enen en nullen. De code is vier tekens lang. Ieder teken staat voor een soort geweld. De code '0000' geeft aan dat de verdachte nooit voor een gewelddadig delict is aangehouden en '1111' geeft aan dat de verdachte ieder soort geweld minstens één keer heeft gepleegd. De betekenissen zijn als volgt:

Teken	Betekenis	Hoofdrubrieken
1e (1000)	Gewelddadige seksuele delicten	R115 (verkrachting) R117 (aanranding der eerbaarheid)
2e (0100)	Verbaal geweld	R121 (bedreiging)
3e (0010)	Fysiek geweld	R123 (poging doodslag e.d.) R125 (doodslag e.d. voltooid) R127 (overige misdrijven tegen het leven) R129 (mishandeling)
4e (0001)	Gewelddadige vermogensdelicten	R139 (diefstal met geweld) R141 (afpersing)

Tabel 5-1: Betekenis van de doelvariabele 'soort geweld'

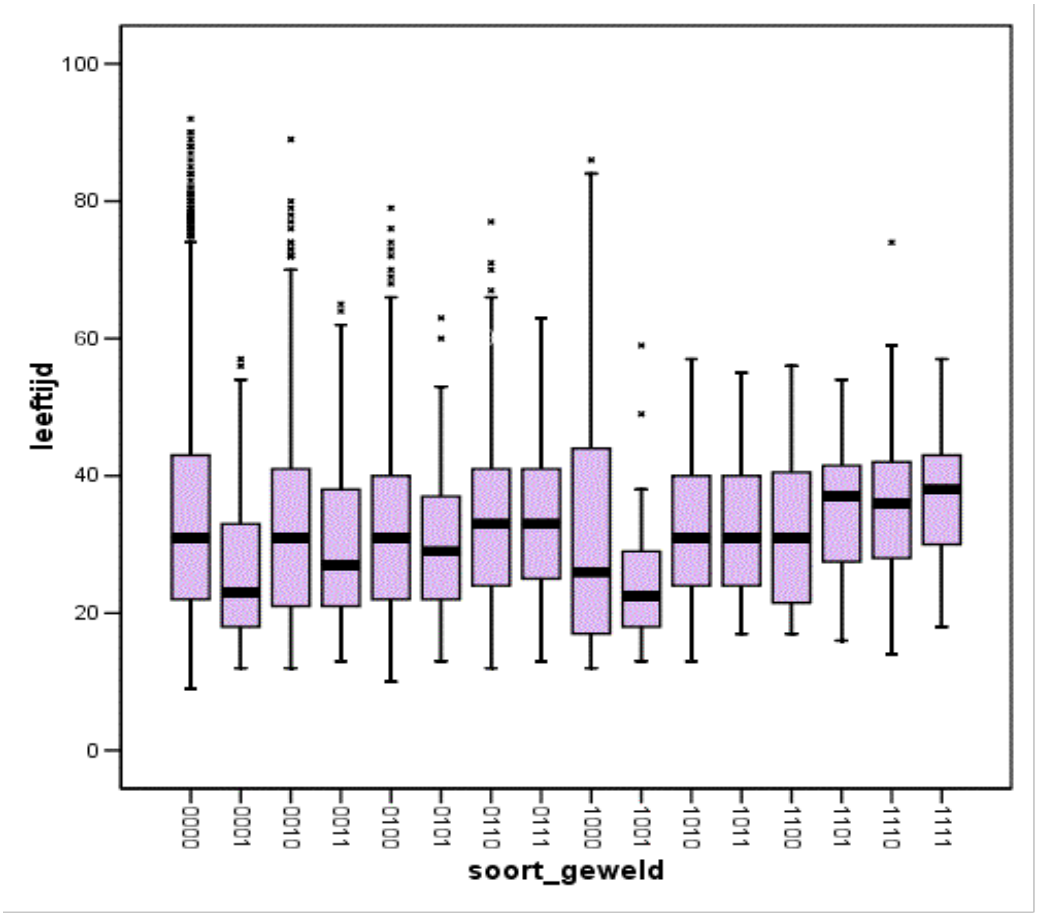
Een combinatie van verbaal en fysiek geweld wordt aangegeven met '0110' en op die manier kunnen ook andere combinaties van geweld weergegeven worden. Totaal zijn er 2^4=16 mogelijke waarden. De frequentietabel staat in tabel 5-2 op volgorde van frequentie.

	Frequentie	Percentage	Cumulatieve percentage
0000	10.952	58,4	58,4
0010	3.205	17,1	75,5
0110	1.144	6,1	81,6
0100	834	4,4	86,0
0111	694	3,7	89,7
0001	682	3,6	93,3
0011	559	3,0	96,3
0101	231	1,2	97,6
1000	117	0,6	98,2
1111	114	0,6	98,8
1010	74	0,4	99,2
1110	72	0,4	99,6
1011	29	0,2	99,7
1001	22	0,1	99,8
1100	19	0,1	99,9
1101	11	0,1	100,0
Totaal	18.759	100,0	

Tabel 5-2: Frequentieverdeling van 'soort geweld'

Opvallend is dat de combinatie fysiek en verbaal geweld (0110) vaker voorkomt dan alleen verbaal geweld (0100). Verder is te zien dat 58% van de verdachten alleen aangehouden is voor non-gewelddadige delicten. De overige 42% heeft dus wel enige vorm van geweld gepleegd. Er zijn 694 van de 18.759 verdachten die zowel vermogensdelicten met geweld, verbaal geweld als fysiek geweld hebben gepleegd.

Boxplots geven de leeftijden per geweldsoort weer in de volgende figuur 5-7. De mediane leeftijden verschillen per geweldtype. Voor non-geweld (0000) ligt de mediaan op 31 jaar (het dikke streepje in het midden van de boks) terwijl de leeftijd van vermogensdelicten met geweld (0001) veel lager ligt, op 23 jaar. Ook de leeftijd van de combinatie gewelddadige seksuele delicten en vermogensdelicten met geweld (1001) is opvallend, deze ligt op 22,5 jaar. Maar hierbij moet wel opgemerkt worden dat slechts 22 verdachten in die categorie vallen. De spreiding van non-geweld (0000) en van gewelddadige seksuele delicten (1000) is relatief groot.



Figuur 5-7: Boxplots van de leeftijd per geweldsoort

De soorten geweld tegenover het aantal antecedenten is in de onderstaande tabel weergegeven.

	soort geweld	Beginner (1-2 ant.)	Doorstromer (3-10 ant.)	Veelpleger (>10 ant.)	Frequentie
geen geweld:	**0000**	82,1%	16,0%	1,9%	10.952
1 soort geweld:	**0001**	36,4%	42,2%	21,4%	231
	0010	57,5%	35,2%	7,3%	3.205
	0100	57,9%	33,6%	8,5%	834
	1000	58,1%	37,6%	4,3%	117
2 soorten geweld:	**0011**	10,4%	45,3%	44,4%	559
	0101	13,0%	37,2%	49,8%	682
	1100	36,8%	47,4%	15,8%	19
	0110	27,0%	49,5%	23,5%	1.144
	1001	22,7%	45,5%	31,8%	22
	1010	29,7%	50,0%	20,3%	74
3 soorten geweld:	**0111**	2,7%	23,2%	74,1%	694
	1011	0,0%	20,7%	79,3%	29
	1101	0,0%	9,1%	90,9%	11
	1110	20,8%	47,2%	31,9%	72
4 soorten geweld:	**1111**	1,8%	14,9%	83,3%	114
	Totaal	**64,5%**	**24,9%**	**10,6%**	**18.759**

Tabel 5-3: Kruistabel van soort geweld tegen typologie van verdachten

De gearceerde waarden in de tabel zijn opvallend. Voor personen die alle soorten geweld hebben gepleegd of een combinatie van drie soorten geweld geldt dat ze veelal veelpleger zijn. Behalve voor de gearceerde combinatie gewelddadige seksuele delicten, verbaal geweld en vermogensdelicten met geweld (1110), die juist weer relatief vaak doorstromer zijn (slechts 32% is veelpleger). De mogelijke combinaties van twee soorten geweld verschillen nogal qua aantal antecedenten. Ten slotte vormen beginners het grootste deel van de non-geweldplegers. Het kan wellicht zo zijn dat verdachten pas na een aantal geweldloze delicten overgaan op delicten met geweld. Dit is echter niet bekend.

5.5.7 Pleegplaats

De plaats van delict kan worden ingedeeld in openbaar, semi-openbaar en privé domein. Van iedere verdachte is bekend hoeveel procent van zijn delicten in die bepaalde domeinen zijn gepleegd. Hieronder is te zien dat een verdachte gemiddeld 15% in privé domein pleegt. De standaarddeviatie is echter redelijk groot, namelijk 30%.

	Openbaar	Semi-openbaar	Privé
Gemiddelde	55,30	28,57	15,15
Standaarddev.	42,39	38,08	29,78

Tabel 5-4: Gemiddelde en standaarddeviatie van het pleegdomein

Daarnaast is geregistreerd in welk bureaugebied een delict is gepleegd. In onderstaande tabel zijn de gemiddelden en standaarddeviaties daarvan gegeven. Opgemerkt moet worden dat de cijfers niet het gemiddeld aantal gepleegde delicten in het bureaugebied aangeven maar juist het gemiddelde percentage dat een verdachte in het bureaugebied pleegt. Zo is te zien dat een verdachte gemiddeld ongeveer 10% van zijn geregistreerde delicten in de Jan Hendrikstraat en Zoetermeer pleegt. De reden voor deze manier van presenteren is dat de dataset uit unieke verdachten bestaat en de delicten geaggregeerd zijn per verdachte.

	Karnebeek	Jan Hendrikstraat	Heemstraat	Hoefkade	Overbosch	Scheve-ningen	Segbroek	Laak	DH ZuidWest
Gemiddelde	2,21	10,37	5,41	4,59	4,42	5,90	5,59	6,22	8,36
Std. dev.	11,83	25,40	18,26	16,66	17,99	20,30	19,17	20,18	23,91

Zuiderpark-laan	Ypenburg Leidschenveen	Zoeter meer	Rijswijk	Westland	Wasse-naar	Voorburg Leidschendam	Delft	Pijnacker Nootdorp	Onbekend DH
5,55	2,22	9,48	4,31	6,00	1,88	6,30	7,02	1,98	0,63
18,90	13,77	27,67	17,40	22,42	12,65	21,63	23,64	13,01	5,88

Tabel 5-5: Gemiddeld aantal delicten gepleegd per bureaugebied en hun standaarddeviatie

5.5.8 Binaire variabelen

In tabel 5-6 is te zien hoe vaak de waarden ja en nee voorkomen bij de binaire variabelen.

Variabele	Nee	Ja	Ja procentueel
Alcoholist[2]	18.588	171	0,9%
Harddrugsgebruiker[2]	17.943	816	4,3%
Medische indicatie	18.720	39	0,2%
Vuurwapengevaarlijk	18.616	143	0,8%
Verzetpleger	18.601	158	0,8%
Vluchtgevaarlijk	18.687	72	0,4%
Zelfmoordneiging	18.724	35	0,2%
Overleden	18.718	41	0,2%
Top 500	18.144	615	3,3%
Bedreiging	16.980	1.779	9,5%
Mishandeling	15.802	2.957	15,8%
Mishandeling gezinslid	18.362	397	2,1%
Stalker	18.642	117	0,6%

Tabel 5-6: Frequentieverdeling van de binaire variabelen

Verder zijn er 15.862 mannelijke verdachten en 2897 vrouwen.

5.5.9 Overige variabelen

Een overzicht van centrummaten en spreidingsmaten van een aantal ratio variabelen is gegeven in tabel. Een overzicht van alle variabelen staat in bijlage B.

Variabele	Minimum	Maximum	Gemiddelde	Std. dev.
lft_1del	12	92	27,82	13,586
lft_ldel	12	92	32,40	13,264
som_feit	1	455	8,30	22,134
som_ant	1	304	5,52	14,110
lft_peil	9	92	32,84	13,265
zwaar_r	0,00	100,00	60,1420	39,20951
poging_r	0,00	100,00	6,4825	18,24871
voor_80_r	0,00	93,33	0,8785	5,79495
jr_03_r	0,00	85,71	5,2815	13,39936
jr_04_r	0,33	100,00	68,1771	36,78342
r101_r	0,00	100,00	4,8642	16,97570
r299_r	0,00	100,00	6,3922	21,16536
gw_sex_r	0,00	100,00	0,6043	6,05232
gw_ovg_r	0,00	100,00	16,9810	30,54600
vm_gw_r	0,00	100,00	2,6394	11,23188
gw_sexpj_r	0,00	100,00	0,3751	5,42862
gw_ovgpj_r	0,00	100,00	12,3250	28,71042

[2] Alcoholgebruik en harddruggebruik worden niet altijd juist geregistreerd. Het vermoeden is dat het werkelijke aantal verslaafden veel hoger ligt.

Variabele	Minimum	Maximum	Gemiddelde	Std. dev.
vm_gwpj_r	0,00	100,00	1,3525	9,60827
januari_r	0,00	100,00	8,7645	22,89013
B02_r	0,00	100,00	21,7247	34,06960
C02_r	0,00	100,00	20,0474	33,03412
D03_r	0,00	100,00	4,2681	16,67260
prive_r	0,00	100,00	15,1482	29,78398
F02_r	0,00	100,00	0,3573	4,72509
H01_r	0,00	100,00	21,1551	37,04814
zondag_r	0,00	100,00	13,8272	28,82755
nacht_r	0,00	100,00	26,1454	37,60780

Tabel 5-7: Centrum- en spreidingsmaten van de meest betekenisvolle ratio variabelen

6. Modelleren

"The purpose of models is not to fit the data but to sharpen the questions."
Samuel Karlin, Royal Society 20, April 1983.

In dit hoofdstuk zullen de vragen betreffende geweld die in paragraaf 3.1 zijn geformuleerd, één voor één worden behandeld. In paragraaf 6.1 wordt een beeld van de geweldplegers geschetst en wordt de techniek uitgelegd die daar het meest geschikt voor is. Er wordt aandacht besteed aan zowel de groepen verdachten die gescheiden worden door het soort geweld als de tweedeling gewelddadig/niet gewelddadig. De kenmerken worden beschreven met behulp van classification. Vervolgens wordt in paragraaf 6.2 gezocht naar exploratieve verbanden in de gegevens van alle verdachten die ooit geweld hebben gepleegd. Hierbij wordt de methode association rules gebruikt. Tot slot wordt met behulp van clustering gezocht naar samenhangende groepen geweldplegers in paragraaf 6.3. Alle datamining is gedaan met Weka 3-4-4.

6.1 Kenmerken van geweldplegers

Het doel is te weten te komen hoe een bepaalde verdachtengroep eruit ziet. Dat kan met behulp van de datamining techniek classification. Het is interessant om te weten wat de kenmerken zijn van een geweldpleger tegenover een non-geweldpleger. Het is echter nog interessanter om te kijken of er kenmerken zijn die verdachten van verschillende soorten geweld van elkaar onderscheiden. Beide varianten zullen besproken worden in deze paragraaf. Daarnaast is er nog de afweging of men kenmerken als aantal antecedenten en andere gegevens over de criminele carrière van de verdachte mee laat wegen, of alleen zijn persoonskenmerken zoals woonplaats, geslacht en etniciteit. Omdat de interesse naar beide opties uit gaat is besloten ze allebei apart uit te voeren. Een overzicht van de toegepaste classificaties:

Kenmerken → Doelvariabele ↓	Persoonskenmerken	Persoons- & delictkenmerken
Soort geweld (4 soorten)	Paragraaf 6.1.2	Paragraaf 6.1.4
Geweld (ja/nee)	Paragraaf 6.1.3	Paragraaf 6.1.5

Tabel 6-1: Overzicht van deze paragraaf

De vier soorten geweld die in dit onderzoek gebruikt worden staan in paragraaf 5.4.6 beschreven en zijn: gewelddadige seksuele delicten, verbaal geweld, fysiek geweld, gewelddadige vermogensdelicten en alle mogelijke combinaties daarvan. Deze doelvariabele heeft een erg scheve verdeling. Geregistreerde verdachten die nooit een gewelddadig delict hebben gepleegd komen erg vaak voor (56%) terwijl verdachten van zedendelicten nog geen 1% van de dataset beslaan. Classificatietechnieken kunnen over het algemeen niet goed met zulke scheve verdelingen omgaan en zullen niet optimaal presteren. Toch is naar deze indeling gekeken om te zien of er misschien discriminerende variabelen te ontdekken zijn, variabelen die kenmerkend zijn voor verdachten met één bepaald soort geweld of voor verdachten met meerdere soorten geweld.

De variabelen die aangeduid worden met "persoonskenmerken" zijn:

gsl_cat	*geb_land*	*beroep_cd*	*g6*
land_cd	*nat_code*	*g1*	*g7*
woon_reg	*cbs_etn*	*g2*	*g8*
gem_cat	*leeftijd*	*g3*	
b_wbs_nw	*ovl*	*g4*	

De delictkenmerken zijn:

lft_1del	*jr_99_r*	*D11_r*	*Vrijdag_r*
lft_ldel	*jr_00_r*	*D12_r*	*Zaterdag_r*
typol3	*jr_01_r*	*D13_r*	*Nacht_r*
top500	*jr_02_r*	*D17_r*	*Ochtend_r*
som_ant	*jr_03_r*	*D19_r*	*Middag_r*
som_feit	*jr_04_r*	*Prive_r*	*Avond_r*
zwaar_r	*januari_r*	*Semi_r*	*Karnebeek_r*
poging_r	*februari_r*	*Openb_r*	*JanHendrikstraat_r*
voor_80_r	*maart_r*	*F02_r*	*Heemstraat_r*
jr_80_r	*april_r*	*F07_r*	*Hoefkade_r*
jr_81_r	*mei_r*	*F08_r*	*Overbosch_r*
jr_82_r	*juni_r*	*F09_r*	*Segbroek_r*
jr_83_r	*juli_r*	*F12_r*	*Laak_r*
jr_84_r	*augustus_r*	*F14_r*	*DHZuidWest_r*
jr_85_r	*september_r*	*H01_r*	*Zuiderparklaan_r*
jr_86_r	*oktober_r*	*H02_r*	*YpenburgLeidschenveen_r*
jr_87_r	*november_r*	*H03_r*	*Zoetermeer_r*
jr_88_r	*december_r*	*H04_r*	*Rijswijk_r*
jr_89_r	*B02_r*	*H05_r*	*Westland_r*
jr_90_r	*C02_r*	*H06_r*	*Wassenaar_r*
jr_91_r	*C06_r*	*H07_r*	*VoorburgLeidschendam_r*
jr_92_r	*C14_r*	*H08_r*	*Delft_r*
jr_93_r	*C15_r*	*H09_r*	*PijnackerNootdorp_r*
jr_94_r	*C20_r*	*Zondag_r*	*onbekendDH_r*
jr_95_r	*C22_r*	*Maandag_r*	*Scheveningen_r*
jr_96_r	*D03_r*	*Dinsdag_r*	
jr_97_r	*D08_r*	*Woensdag_r*	
jr_98_r	*D09_r*	*Donderdag_r*	

De betekenis van de variabelen staat in bijlage B.

6.1.1 Werkwijze

Trainen en testen
Er is gekozen voor 10-fold cross validation omdat bij deze methode tien keer wordt getraind op verschillende stukken van de dataset en het gemiddelde van de resultaten wordt bepaald. Hierdoor wordt meer zekerheid verkregen over de prestatie van de techniek. Uitgebreide tests op verschillende datasets met verschillende technieken heeft aangetoond dat tien ongeveer het juiste aantal folds is om de beste schatting van de fout te krijgen. Er is ook theoretisch bewijs dat dit bevestigt [*31*].
Om de variantie te kunnen bepalen is de zojuist beschreven cross validation tien keer uitgevoerd, ofwel: 10 keer 10-fold cross validation. Om werk te besparen en de resultaten netjes op een rij te krijgen is een awk-script gemaakt die geschikt is voor uitvoering onder het besturingssysteem Unix. Het script is geschreven in EditPlus, een tekstverwerkingsprogramma voor programmeertaal. Unix is op de Vrije Universiteit aanwezig en zodoende is er via een beveiligd programma (Putty) ingelogd om het script te kunnen draaien. De resultaten zijn tenslotte met Microsoft Excel verwerkt om het gemiddelde en de standaarddeviatie te berekenen. Een voorbeeld van een stukje script staat op de volgende pagina.

```
BEGIN
{
  print "J48:" >> "/home/ankie/DM/res_5_verdEnDelGW.txt"
  system("java -Xms512M -Xmx1024M
  weka.filters.supervised.attribute.AttributeSelection
  -i /home/ankie/DM/5_verdEnDelGW.arff -o /home/ankie/DM/temp2GW.arff
  -E 'weka.attributeSelection.WrapperSubsetEval
  -B weka.classifiers.trees.J48'
  -S 'weka.attributeSelection.BestFirst'")

  for (i=1; i<=10; i++)
  {
    system(sprintf("java weka.classifiers.trees.J48 -v -s %i -t
    /home/ankie/DM/temp2GW.arff | grep 'Correctly' >>
    /home/ankie/DM/res_5_verdEnDelGW.txt",i))
  }
}
```

In dit script wordt in een tekstbestand opgeslagen welke classifier gebruikt wordt (in dit voorbeeld J48), vervolgens wordt de attribuutselectie methode wrapper toegepast met als classifier J48. De output hiervan wordt opgeslagen in een arff-file, genaamd temp2GW. Vervolgens wordt tien keer de classifier aangeroepen met verschillende beginwaarde (*seed*) zodat er steeds een andere indeling voor de cross validation wordt genomen. Alleen het percentage goed geclassificeerde objecten wordt opgeslagen in het tekstbestand.

Dataset
Er heeft een bewerking plaats gevonden op de variabelen die het aantal delicten of het aantal antecedenten per unieke verdachte aangeven. Ze zijn relatief gemaakt ten opzichte van het totaal aantal feiten en het totaal aantal antecedenten dat de persoon heeft gepleegd. Als een verdachte namelijk één maal een tasjesroof heeft gepleegd, dan lijkt dit interessant. Maar als daarnaast bekend is dat diegene acht maal een roofoverval heeft gepleegd dan valt die tasjesroof in het niet. De feiten die het meest gepleegd worden zouden zwaarder moeten wegen dan de feiten die minder vaak gepleegd worden. De berekening van bijvoorbeeld jaartal 1981 is als volgt gegaan:

$$jaar81\_relatief = \frac{jaar81}{aantal\ \ antecedenten} \times 100 \qquad (4)$$

Op deze manier liggen alle waarden tussen 0 en 100. Deze berekening is toegepast op alle variabelen waar een ‘X’ bij staat in bijlage B.

Attribuutselectie
Attribuutselectie is uit te voeren met behulp van een filter of een wrapper zoals in paragraaf 5.4.2 is uitgelegd. Over het algemeen werkt een wrapper beter dan een filter. Dit is intuïtief aannemelijk aangezien een filter een algemene selectie maakt terwijl wrapper de te gebruiken classifier als kennis heeft. Er is zoveel mogelijk gebruik gemaakt van de wrapper. Dat is echter een erg tijdrovende bezigheid. De methode kan zelfs langer dan een week in beslag nemen. Daarom is in sommige gevallen gebruik gemaakt van een filter. Het is mogelijk dat technieken iets beter zullen presteren als wrapper gebruikt zou worden of een andere filter. Bovendien zijn er meerdere optimale verzamelingen mogelijk:
“An optimal feature subset need not be unique because it may be possible to achieve the same accuracy using different sets of features (e.g when two features are perfectly correlated, one can be replaced by the other.” [*17*]
Zo zouden bijvoorbeeld nationaliteit en geboorteland vervangen kunnen worden door etniciteit en andersom. Het resultaat van de filter is echter in dit onderzoek niet gewijzigd.

De gebruikte filter is CfsSubsetEval met 10-fold cross validation. CFS (Correlation-based Feature Selection) is een snel algoritme die in het algemeen minder attributen kiest dan andere filters. De filter beoordeelt de waarde van een verzameling attributen door te kijken naar hoe goed ieder attribuut de doelvariabele kan voorspellen in combinatie met de mate van overbodigheid ten opzichte van andere

attributen. Verzamelingen die hoog gecorreleerd zijn met de doelvariabele terwijl ze lage correlatie hebben met andere attributen hebben de voorkeur [*13*].

Steekproeven
Er zijn steekproeven gemaakt van 5%, 10% en 50% van de 18795 verdachten om de prestatie van de classifiers te kunnen meten. Die samples moeten stratified zijn, dat wil zeggen dat de verdeling van de klassen in die samples hetzelfde is als in de hele dataset, zoals in paragraaf 4.3.2 is uitgelegd. Daarvoor is het Weka commando Resample gebruikt. Bij het maken van de steekproeven kan het bij kleine klassen echter voorkomen dat er een waarde van de doelvariabele niet meer in voorkomt. De steekproef is in dat geval niet representatief voor de gehele dataset en de prestatie kan daardoor verslechteren. Om dit te voorkomen is een object dat tot de missende klasse behoort toegevoegd. Een voorbeeld van een sample van 10% staat in tabel 6-2.

Soort geweld	**Hele dataset**		**Sample 10%**		**Stratified sample 10%**	
	Abs. frequentie	Rel. frequentie	Abs. frequentie	Rel. frequentie	Abs. frequentie	Rel. frequentie
0000	10.952	58,40%	1057	56,37%	1057	56,37%
0001	682	3,60%	60	3,20%	60	3,20%
0010	3.205	17,10%	314	16,75%	314	16,75%
0011	559	3,00%	72	3,84%	71	3,79%
0100	834	4,40%	97	5,17%	97	5,17%
0101	231	1,20%	24	1,28%	24	1,28%
0110	1.144	6,10%	129	6,88%	129	6,88%
0111	694	3,70%	64	3,41%	64	3,41%
1000	117	0,60%	15	0,80%	15	0,80%
1001	22	0,10%	1	0,05%	1	0,05%
1010	74	0,40%	13	0,69%	13	0,69%
1011	29	0,20%	2	0,11%	2	0,11%
1100	19	0,10%	0	0,00%	1	0,05%
1101	11	0,1%	1	0,05%	1	0,05%
1110	72	0,40%	10	0,53%	10	0,53%
1111	114	0,60%	16	0,85%	16	0,85%
Totaal:	**18.759**	**100**	**1875**	**100**	**1875**	**100**

Tabel 6-2: Voorbeeld van het maken van een sample

Na enkele tests op steekproeven van de dataset is gebleken dat de prestaties niet veel veranderen naarmate de samples groter worden. Het verschil tussen het aantal goed geclassificeerde verdachten is hooguit 4%. De tijd die het kost om een algoritme uit te voeren verkort wel aanzienlijk naarmate de steekproef kleiner wordt. Er is daarom gekozen om alle classification algoritmen toe te passen op een sample van 5% (937 verdachten). Alleen het best presterende algoritme zal toegepast worden op de gehele dataset om de meest betrouwbare resultaten te verkrijgen.

Keuze technieken
Weka heeft tientallen algoritmen die geschikt zijn voor het classificeren. Maar ze zijn niet allemaal geschikt voor iedere dataset en iedere vraagstelling. Allereerst stellen de algoritmen eisen aan de vorm van de gegevens. De dataset van geregistreerde verdachten bestaat uit zowel nominale als numerieke variabelen. Daarom zijn er alleen technieken gebruikt die met beide soorten variabelen om kunnen gaan. Daarnaast is de vraagstelling belangrijk. Voor de nominale doelvariabele *soort geweld* kunnen alleen technieken gebruikt worden die een doelvariabele met meer dan twee klassen kunnen voorspellen. Voor de binaire doelvariabele *geweld* kunnen ook andere technieken gebruikt worden. Classification technieken voor numerieke voorspellingen (regressie) zijn niet nodig. Daarnaast is besloten geen neurale netwerk technieken te gebruiken. Deze geven over het algemeen een goede classificatie maar geven weinig inzicht in de invloed van de verschillende variabelen. Met zulke technieken is het profiel van een verdachtengroep moeilijk te schetsen.

Met behulp van WekaMetaL (een uitbreiding van Weka) en de Weka Experimenter is bekeken welke technieken waarschijnlijk het beste resultaat op zullen leveren. WekaMetaL gebruikt kennis van resultaten van algoritmen op verschillende standaard datasets uit Weka om te voorspellen hoe ze zullen presteren op de gegeven dataset. De Experimenter vergelijkt meerdere datasets (of samples) en meerdere technieken met elkaar.

De voor dit onderzoek geschikte algoritmen zijn:

Beslissingsbomen	Regels	Andere	Meta
J48	ZeroR	NaiveBayes	AdaboostM1
ADtree	OneR	IB1	Bagging
DecisionStump	DecisionTable	IBk	
NBtree		K-star	
REPtree		SMO	
		SimpleLogistic	

6.1.2 Classification naar soort geweld op basis van persoonskenmerken

Prestatie
Het aantal goed geclassificeerde verdachten op basis van de persoonskenmerken met als doelvariabele *soort geweld* zijn in tabel 6-3 gegeven. De resultaten zijn behaald met behulp van een stratified sample van 5% en de wrapper techniek is toegepast om attributen te selecteren. Van een aantal technieken is het gemiddelde percentage goed geclassificeerde verdachten en hun standaarddeviatie gegeven in onderstaande tabel.

	ZeroR	OneR	J48	NaiveBayes	IB1	AdaBoostM1 met Naïve Bayes
Gemiddelde	55,82%	55,39%	56,31%	**56,36%**	39,28%	**56,59%**
Std. dev.	0	0,21	0,13	0,27	1,05	0,16

Tabel 6-3: Prestatie van enkele algoritmen

Het valt op dat de prestaties van de technieken erg laag zijn. Sowieso is een prestatie van rond de 50% zeer slecht, maar ook in vergelijking met ZeroR leveren de andere technieken geen betere resultaten. ZeroR kan worden gezien als minimale prestatie voor overige algoritmen omdat dit algoritme standaard alle verdachten indeelt in de meest voorkomende klasse.

OneR is een algoritme dat classificeert op basis van slechts één variabele. De ene keer kiest OneR beroep als het belangrijkste attribuut en de andere keer harddruggebruiker (*g2*). Dit hangt af van de gekozen seed voor cross validation. Er is dus geen variabele die significant bepalend is voor de classificatie. NaiveBayes is het algoritme dat het best gepresteerd heeft, op de metatechniek AdaBoostM1 na.

Algoritme: NaiveBayes en AdaBoostM1
Naive Bayes maakt de aanname dat alle variabelen onafhankelijk van elkaar zijn en berekent de kans dat een object in een bepaalde klasse valt. Stel dat van een aangehouden verdachte het volgende is gegeven:

Beroep	**Alcoholist**	**Soort geweld**
Jurist	Ja	?

Met de gegevens van de trainingset wordt voor iedere klasse (*i*) de kans om in die klasse te komen berekend. Dit gebeurt als volgt:

P(i) = P(beroep=jurist | soort geweld=i) x P(alcoholist=ja | soort geweld=i) x P(soort geweld=i)

De verdachte wordt vervolgens ingedeeld in de klasse waarop de kans het hoogst is.

Missende waarden zijn voor Naive Bayes geen probleem. Als in het voorbeeld het beroep missend is, dan telt de voorwaardelijke kans op beroep eenvoudigweg niet mee en wordt de kans op het soort geweld als volgt berekend:

P(i) = P(alcoholist=ja | soort geweld=i) x P(soort geweld=i)

Als in de trainingset een waarde mist, dan is die niet meegerekend bij de frequentie. Ook de kansen zijn gebaseerd op het aantal waarden dat feitelijk voorkomt en niet op het totaal aantal objecten in de trainingset.

Er zijn twee parameters die de resultaten van Naive Bayes kunnen beïnvloeden, namelijk de kernel estimator en supervised discretization. Als de kernel estimator gebruikt wordt dan wordt van iedere variabele een schatting van de verdeling gemaakt in plaats van aan te nemen dat ze normaal verdeeld zijn. Supervised discretization wil zeggen dat de numerieke variabelen omgezet worden in nominale op een manier die het meest voordelig is voor het classificeren, dus rekening houdend met de doelvariabele. De twee opties staan standaard uit in Weka.

Voor meer informatie over Naive Bayes zoals het in Weka geïmplementeerd is, wordt verwezen naar het artikel van George H. John en Pat Langley [*15*].

AdaBoostM1 met als basis Naive Bayes geeft een iets hoger resultaat. AdaBoost is een boosting methode. Boosting werkt door herhaaldelijk een classifier (bijvoorbeeld Naive Bayes) toe te passen op verschillende verdelingen van de trainingset en vervolgens die classifiers te combineren in een enkele samengestelde classifier. Prestaties verbeteren vaak behoorlijk door boosting maar in dit geval maar heel summier. Een uitgebreide beschrijving van AdaBoostM1 is te vinden in het artikel van Y. Freund en R. E. Schapire [*11*].

Resultaten
Na experimenteren met de parameters is gebleken dat het geen verschil maakt voor de prestatie. Er verandert niets of nauwelijks iets aan de classificering van verdachten. Het toepassen van NaiveBayes met de standaard parameters op de gehele dataset resulteert in 59,4% goed geclassificeerde objecten, een verbetering van 3%. De wrapper techniek heeft de variabelen geslacht, harddruggebruiker (*g2*), vuurwapengevaarlijk (*g4*), verzetpleger (*g6*) en vluchtgevaarlijk (*g7*) gekozen om de classificatie van Naive Bayes te optimaliseren. AdaBoostM1 geeft dezelfde prestatie.

Hieronder is de *confusion matrix* te zien van Naive Bayes.

a	b	c	d	e	f	g	h	i	j	k	l	m	n	o	p	<-- classified as
10857	0	2	0	0	0	0	0	1	0	0	0	92	0	0	0	a = 0000
605	0	0	0	0	0	0	0	0	0	0	0	77	0	0	0	b = 0001
3111	0	1	0	0	0	0	0	0	0	0	0	93	0	0	0	c = 0010
804	0	0	0	0	0	0	0	0	0	0	0	30	0	0	0	d = 0100
117	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	e = 1000
18	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	f = 1100
73	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	g = 1010
19	0	0	0	0	0	0	0	0	0	0	0	3	0	0	0	h = 1001
1048	0	0	0	0	0	0	0	0	0	0	0	95	0	0	1	i = 0110
165	0	0	0	0	0	0	0	0	0	0	0	66	0	0	0	j = 0101
464	0	0	0	0	0	0	0	0	0	0	0	95	0	0	0	k = 0011
63	0	0	0	0	0	0	0	0	0	0	0	9	0	0	0	l = 1110
405	0	1	0	0	0	0	0	0	0	0	0	287	0	0	1	m = 0111
16	0	0	0	0	0	0	0	0	0	0	0	13	0	0	0	n = 1011
4	0	0	0	0	0	0	0	0	0	0	0	7	0	0	0	o = 1101
56	0	0	0	0	0	0	0	0	0	0	0	58	0	0	0	p = 1111

Het is te zien dat de meeste verdachten in categorie ‘0000’ (a) zijn ingedeeld. Dat is de meest voorkomende categorie: verdachten die nog nooit voor een gewelddadig delict zijn opgepakt. De getallen die op de diagonaal staan zijn de aantallen goed geclassificeerde verdachten. Er zijn dus 10857 verdachten terecht in categorie ‘0000’ ingedeeld. Opvallend is dat ook veel verdachten in categorie ‘0111’ (m) ingedeeld worden. Dat is de categorie met verdachten die aangehouden zijn voor zowel verbaal geweld, fysiek geweld als vermogenscriminaliteit met geweld maar niet voor seksuele

delicten met geweld. Dit is vooral opmerkelijk omdat in de gehele dataset slechts 4% in deze categorie valt. Een verklaring voor de classificatie is te vinden met behulp van een kruistabel van *soort geweld* en druggebruiker (*g2*).

Soort geweld		Druggebruiker (g2)		
		nee	ja	Totaal
0000	Frequentie	10.847	105	10.952
	% in soort_geweld	99.0%	1.0%	100.0%
0001	Frequentie	601	81	682
	% in soort_geweld	88.1%	11.9%	100.0%
0010	Frequentie	3.129	76	3.205
	% in soort_geweld	97.6%	2.4%	100.0%
0011	Frequentie	468	91	559
	% in soort_geweld	83.7%	16.3%	100.0%
0100	Frequentie	807	27	834
	% in soort_geweld	96.8%	3.2%	100.0%
0101	Frequentie	169	62	231
	% in soort_geweld	73.2%	26.8%	100.0%
0110	Frequentie	1.085	59	1.144
	% in soort_geweld	94.8%	5.2%	100.0%
0111	Frequentie	458	236	694
	% in soort_geweld	66.0%	34.0%	100.0%
1000	Frequentie	117	0	117
	% in soort_geweld	100.0%	0.0%	100.0%
1001	Frequentie	19	3	22
	% in soort_geweld	86.4%	13.6%	100.0%
1010	Frequentie	73	1	74
	% in soort_geweld	98.6%	1.4%	100.0%
1011	Frequentie	17	12	29
	% in soort_geweld	58.6%	41.4%	100.0%
1100	Frequentie	18	1	19
	% in soort_geweld	94.7%	5.3%	100.0%
1101	Frequentie	5	6	11
	% in soort_geweld	45.5%	54.5%	100.0%
1110	Frequentie	65	7	72
	% in soort_geweld	90.3%	9.7%	100.0%
1111	Frequentie	65	49	114
	% in soort_geweld	57.0%	43.0%	100.0%
Totaal	Frequentie	17.943	816	18.759
	% in soort_geweld	95.7%	4.3%	100.0%

Tabel 6-5: Kruistabel van druggebruiker en soort geweld

De kans op harddruggebruik (*g2*) in '0111' is behoorlijk hoog (34%) maar die van '1011', '1101' en '1111' nog hoger. De kans om in de categorie '1101' te vallen is daarentegen zeer klein, aangezien slechts 11 van de 18759 verdachten daarin vallen. Naive Bayes vermenigvuldigt de voorwaardelijke kans zoals P(*g2*=ja | *soort geweld*=0111), met de kans P(*soort geweld*=0111). Omdat die laatste kans relatief hoog is, wordt klasse '0111' aangewezen als juiste klasse als de verdachte druggebruiker is.

Dit gaat ook op voor verdachten die de waarde ‘ja’ hebben voor de variabelen vuurwapengevaarlijk (*g4*), verzetpleger (*g6*) en vluchtgevaarlijk (*g7*).

Conclusie
Door de scheve verdeling in de doelvariabele *soort geweld* kunnen de verdachten niet goed geclassificeerd worden. De beste prestatie op de gehele dataset is 59,4%, behaald met Naive Bayes. Het aantal goed geclassificeerde verdachten in de trainingset is ook laag en er kan dus geconcludeerd worden dat er geen duidelijk beeld te schetsen is van de verschillende typen geweldsplegers, uitsluitend op basis van persoonskenmerken. Wel is gevonden dat druggebruik, vuurwapengevaarlijk, vluchtgevaarlijk en verzetpleging invloed hebben op het soort geweld.

6.1.3 Classification naar wel/geen geweld op basis van persoonskenmerken

Omdat het classificeren op verschillende typen geweldplegers geen goede resultaten oplevert ontstaat de vraag of er misschien wel kenmerken te onderscheiden zijn bij het classificeren naar geweldplegers en non-geweldplegers. Er zijn dan slechts twee verschillende klassen waar een verdachte in kan vallen. De verdeling van deze doelvariabele *geweld* is niet zo scheef: 42% van alle verdachten heeft ooit geweld gepleegd tegenover 58% die nog nooit opgepakt zijn voor een gewelddadig delict.

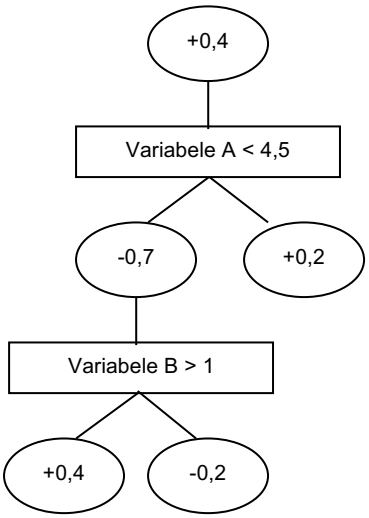
Prestatie
Met de wrapper attribuutselectie worden de volgende resultaten verkregen:

	ZeroR	OneR	J48	NaiveBayes	IB1	ADtree	Bagging met Adtree
Gemiddelde	55,82%	54,22%	59,87%	63,61%	54,61%	**64,86%**	64,19%
Std. Dev	0	0,7639	0,1118	0,4258	1,25	0,3042	0,7024

Tabel 6-6: Prestatie van enkele algoritmen

De Alternating Decisiontree (ADTree) is de techniek die hier het beste heeft gepresteerd. Het gemiddeld aantal goed geclassificeerde objecten is bijna 65% en de standaarddeviatie is laag (0,3). Een boosting algoritme met deze classifier zal geen verschil uitmaken omdat boosting al in ADtree geïmplementeerd is. Bagging verbetert de prestatie niet.

Algoritme: ADtree (Alternating Decisiontree)
Een alternating decisiontree lijkt op een standaard beslissingsboom zoals in figuur 4-2, met als verschil dat bij iedere splitsing een betrouwbaarheid wordt meegegeven. Een voorbeeld van een alternating tree is te zien in figuur 6-1:



Figuur 6-1: voorbeeld van een alternating decisiontree

Het voorbeeld beschrijft of iemand in de klasse ‘ja’ dan wel ‘nee’ valt. De klasse ‘ja’ wordt vertegenwoordigd door een positief teken en ‘nee’ door een negatief teken. Het getal bovenaan de boom (+0,4) geeft aan dat er meer gevallen in klasse ‘ja’ vallen dan in ‘nee’. De classificatie van een object met de waarde A = 0,5 en B = 0,5 gaat via een optelsom, te beginnen bij 0,4. Omdat 0,5 kleiner is dan 4,5 wordt 0,7 van de som afgetrokken en omdat B niet groter is dan 1 wordt er nog eens 0,2 afgetrokken. De totale som wordt dus: +0,4 -0,7 -0,2= -0,5. Het teken van de totale som is negatief en daarom valt dit object in de klasse ‘nee’. De betrouwbaarheid kan niet gezien worden als een kans. De manier waarop de betrouwbaarheid berekend wordt zorgt er namelijk voor dat de uitkomst niet altijd tussen 0 en 1 ligt. Wel geldt: hoe meer de waarde afwijkt van nul, hoe betrouwbaarder de classificatie is [*10*].

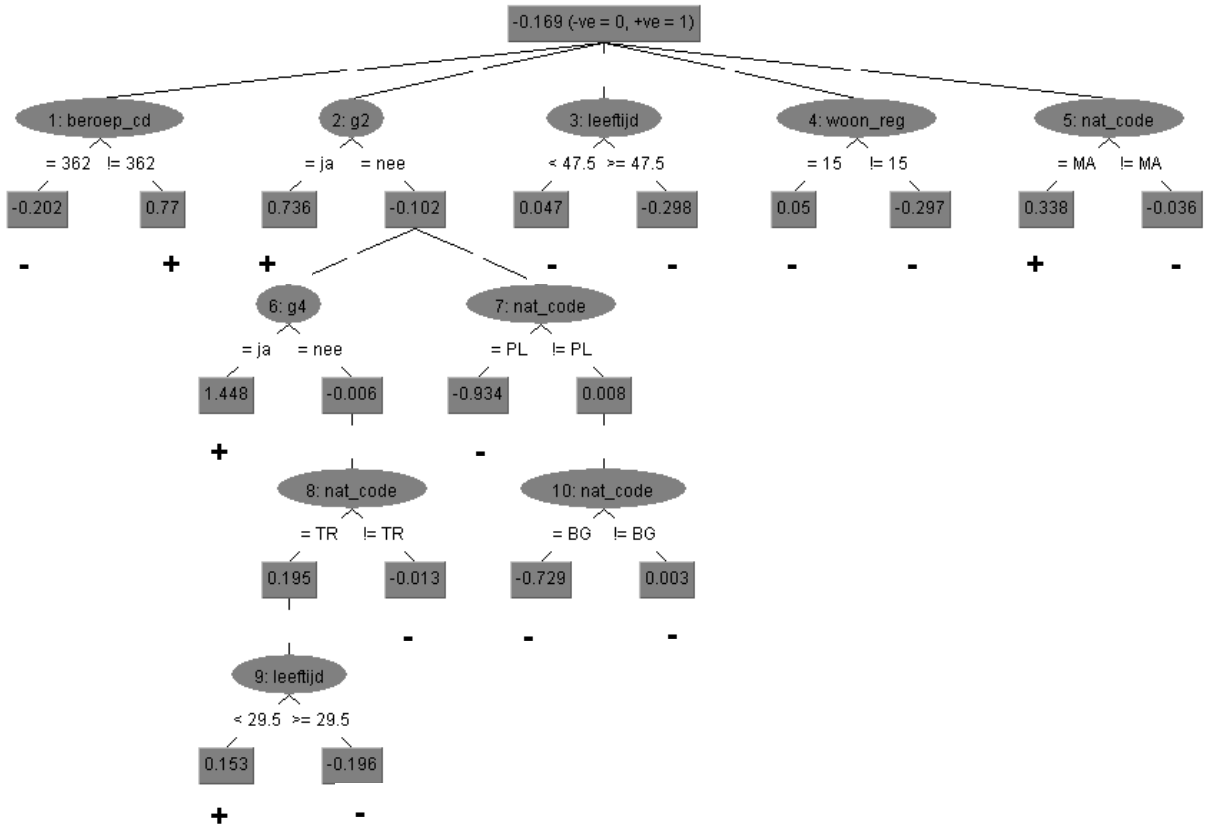
De techniek heeft twee parameters die de resultaten kunnen beïnvloeden. De eerste is het aantal iteraties voor boosting. Meer iteraties resulteert in een grotere, nauwkeurigere beslissingsboom maar maakt het model trager. Iedere iteratie voegt drie niveaus toe aan de boom tenzij er wordt samengevoegd. De tweede parameter is het zoekpad. Deze stelt de zoekmethode in die gebruikt wordt om de boom te bouwen. De standaard instelling ‘expand all paths’ is de beste. Andere zoekmethoden zijn sneller maar geven geen garantie voor het vinden van de optimale oplossing.

Resultaten
Het optimaliseren van het aantal boosting iteraties geeft 30. Het percentage goed geclassificeerd is dan 65,4%. De standaarddeviatie gaat echter ook omhoog en wordt 0,53. De resultaten op de gehele dataset, met de optimale parameters, hebben een gemiddelde van 66,5% en een veel lagere standaarddeviatie van 0,05. De beslissingsboom is echter veel te uitgebreid om een goed inzicht te verkrijgen. Bovendien is het resultaat met 10 boosting iteraties bijna even goed, namelijk 66,3% goed geclassificeerd. Het algoritme heeft de volgende attributen als relevant bestempeld: woonregio, nationaliteit, leeftijd, beroep, druggebruiker en vuurwapengevaarlijk.

In de confusion matrix van ADtree is te zien hoe de verdachten geclassificeerd zijn. Veel verdachten komen onterecht in de klasse non-geweld, namelijk 5115.

```
   0    1  | <-- classified as
-----------+------------------
9750 1202  | 0
5115 2692  | 1
```

De boom van de gehele dataset is te zien op de volgende pagina.



Figuur 6-2: Alternating decisiontree van persoonskenmerken. Een negatief teken is non-geweld, positief teken is geweld. Beroepscode 362 is verplegend personeel / ziekenverzorgende, g2 is harddruggebruik, g4 is vuurwapengevaarlijk, woonregio 15 is regio Haaglanden, nationaliteit MA is Marokko, PL is Polen, TR is Turkije en BG is Bulgarije

10-Fold cross validation neemt steeds 90% van de dataset om een boom van te construeren en vervolgens toe te passen op de overige 10% van de verdachten. Die 10 bomen zullen dus steeds iets verschillen aangezien er steeds een ander deel van de dataset gebruikt wordt. De boom van figuur 6-2 is de boom van de volledige dataset. Zodoende geven de getallen die erin staan een beschrijving van alle geregistreerde verdachten van 2004. De bovenste rij van 5 variabelen zijn opties. Bij classificatie van een verdachte wordt één van de takken gekozen om te classificeren.

Opvallend is dat de variabele *beroep_cd* splitst op beroepcode 362, 'verplegend personeel / ziekenverzorgende'. Als de verdachte dat beroep heeft dan is de kans groter dat diegene geen geweld pleegt en als hij of zij niet dat beroep heeft dan is de kans groter dat hij of zij wel geweld heeft gepleegd. Nadere beschouwing toont echter dat er slechts 13 verdachten het beroep hebben. Er is geen verklaring gevonden voor het feit dat ADtree deze variabele toch als onderscheidend heeft aangemerkt.

Verder is er onderscheid te maken tussen wel of niet plegen van geweld op basis van harddruggebruik (*g2*). In de boom is te zien dat bij de tak 'ja' het getal 0,736 staat. Dit getal wijkt ver af van nul en is dus erg betrouwbaar. Het wil zeggen dat als een verdachte harddrugs gebruikt, de kans op het plegen van geweld veel hoger is dan als de persoon geen harddrugs gebruikt. Als men wel harddrugs gebruikt dan is het percentage gewelddadige verdachten 87%, inderdaad een stuk hoger dan de 40% van niet harddruggebruikers. Het wil echter niet zeggen dat als een verdachte geen harddruggebruiker is, dat hij/zij dan ook een non-geweldpleger is. Dat is te zien aan het getal -0,102 (dat dichtbij nul ligt) en aan de verdere uitsplitsing van de tak. De tak is uitgesplitst naar vuurwapengevaarlijk (*g4*) en nationaliteit, die op hun beurt ook weer uitgesplitst zijn. Verder zijn er nog significante verschillen gevonden tussen geweldplegers en non-geweldplegers in de variabelen woonregio, leeftijd en nationaliteit. Met behulp van de boom en frequentietabellen is de volgende profielschets van een geweldpleger op te maken:

Er bevinden zich relatief veel geweldplegers onder verdachten die…

- .. harddruggebruiker zijn.
  87% van de harddruggebruikers pleegt geweld. Van de verdachten die niet geregistreerd staan als harddruggebruiker pleegt slechts 40% geweld.
- .. de Marokkaanse nationaliteit hebben.
  59% van de verdachten met Marokkaanse etniciteit pleegt geweld. Van verdachten met andere etniciteiten pleegt 40% geweld
- .. vuurwapengevaarlijk zijn.
  96% van deze groep pleegt geweld.
- .. de Turkse nationaliteit hebben en jonger zijn dan 30 jaar.
  56% van deze groep pleegt geweld. Van alle verdachten jonger dan 30 jaar pleegt 42% geweld. Van alle verdachten met Turkse nationaliteit pleegt 50% geweld.

Er bevinden zich relatief weinig geweldplegers onder verdachten die…

- .. de Bulgaarse nationaliteit hebben.
  87% van de verdachten met Bulgaarse nationaliteit pleegt geen geweld tegenover 58% van andere etniciteiten. De Bulgaarse groep telt 68 personen.
- .. de Poolse nationaliteit hebben.
  92% pleegt geen geweld. Er zijn totaal 183 verdachten met Poolse nationaliteit.
- .. buiten de regio wonen.
  73% van buiten de regio wonende verdachten pleegt geen geweld tegenover 56% van binnen de regio wonenden.
- .. ouder zijn dan 47 jaar.
  70% van de verdachten die ouder zijn dan 47 jaar pleegt geen geweld.

De attribuutselectie heeft beroep als relevante variabele aangegeven. Dat is een opvallend gegeven aangezien 90% van de beroepen missend is omdat er de laatste jaren niet meer naar gevraagd wordt bij een aanhouding. Het is interessant om eens te kijken naar een dataset waarin alle beroepen gevuld zijn, om te kijken of deze variabele significante verbeteringen oplevert. Het resultaat van ADtree is 67,88%, een verbetering van slechts 1,4%. Conclusie is dat als beroep goed gevuld zou zijn dat dan ook geen groot onderscheid te maken is tussen geweldplegers en non-geweldplegers op basis van uitsluitend hun persoonskenmerken. Maar wellicht wel iets beter.

Conclusie
Het hoogst behaalde resultaat is 66,5% goed geclassificeerd. Dit is al beter dan de classificatie naar de 16 geweldsoorten. Het resultaat is echter niet naar tevredenheid. De conclusie die hieruit getrokken kan worden is dat er op basis van persoonskenmerken niet met veel zekerheid te zeggen is of een verdachte geweld pleegt of niet. Wel zijn er kenmerken die de kans op geweld plegen vergroten zoals de Marokkaanse nationaliteit, harddruggebruik, Turkse nationaliteit en een leeftijd jonger dan 30 jaar hebbend. Kenmerken die een kleinere kans op geweld tot gevolg hebben, zijn: de Bulgaarse en Poolse nationaliteit, ouder dan 47 jaar en buiten de regio wonen.

6.1.4 Classification naar soort geweld op basis van persoons- en delictkenmerken

Voor deze classificatie zijn meer variabelen gebruikt dan bij voorgaande paragrafen. De verwachting is dat de resultaten daardoor zullen verbeteren aangezien nu ook gegevens bekend zijn over het criminele verleden van de verdachte. Gegevens als pleegplaats, tijdstip en maand worden meegenomen maar ook de modus operandi van de verdachten. Allereerst wordt geclassificeerd naar de soorten geweld.

Prestatie
Verschillende classification technieken op een steekproef van 5% van de verdachten geeft onderstaande resultaten.

	ZeroR	OneR	J48	NaiveBayes	IB1	Bagging met J48
Gemiddelde	55,82%	67,75%	**76,87%**	62,01%	53,71%	75,26%
Std. dev.	0	0,5146	0,3262	0,4388	0,6943	0,5533

Tabel 6-7: Prestatie van enkele algoritmen

J48 geeft een opvallend goed resultaat vergeleken met de andere technieken. Boosting methoden met als basis J48 geven hetzelfde resultaat als enkel J48. Dit komt doordat J48 op zichzelf gezien kan worden als een soort boosting algoritme [*16*]. Bagging geeft geen verbetering en ADTree kan niet toegepast worden op deze data omdat er wordt geclassificeerd naar een multinominale doelvariabele.

OneR kiest de variabele kopstoot/slaan/schoppen (*C14*) als meest relevante variabele. De techniek classificeert als volgt:

```
C14_r:
        < 0.405         -> 0000
        < 7.8450        -> 0011
        < 20.715        -> 0010
        < 25.405        -> 0110
        < 34.845        -> 0010
        < 44.155        -> 0110
        >= 44.155       -> 0010
        ?               -> 0000
646 goed : 937 fout geclassificeerd
```

Waarden zijn missend als het gaat om een handmatig antecedent, dat wil zeggen dat het antecedent niet gekoppeld is aan één of meerdere delicten. De indeling die OneR kiest heeft echter slechts een betrouwbaarheid van 68%.

Algoritme: J48
Het algoritme J48 is meer bekend onder de naam C4.5 / C5.0. Het is een techniek die een beslissingsboom zoals in figuur 4-2 construeert op basis van een maat die informationgain wordt genoemd. Het construeren van een boom gebeurt door het herhaald toepassen van enkele stappen. Eerst wordt er een attribuut geselecteerd die de wortel van de boom voorstelt en er worden takken gemaakt naar iedere waarde die in dat attribuut voor kan komen. Op die manier wordt de dataset gesplitst in delen, één voor elke waarde. Vervolgens wordt dit proces herhaald voor ieder deel. Het algoritme stopt met opsplitsen als alle objecten in een tak in dezelfde klasse vallen. De attributen in de boom worden gekozen met behulp van een maat voor zuiverheid, de information. Hoe kleiner de information, hoe zuiverder het attribuut. Een information van nul wil zeggen dat alle klassen in de tak gelijk zijn en dat er dus niet verder gesplitst hoeft te worden. Het meest zuivere attribuut wordt het hoogst in de boom geplaatst [*22*].

De parameters van J48 zijn:
- Pruned: *Pruned* betekent letterlijk gesnoeid, en dat is ook precies wat deze optie doet. De boom wordt achteraf gesnoeid zodat hij niet zo uitgebreid wordt en daardoor makkelijker te interpreteren. Meestal bevordert pruning de prestatie, daarom wordt het standaard toegepast in Weka.
- Betrouwbaarheid: De betrouwbaarheidsfactor die gebruikt wordt voor pruning. Een kleinere waarde heeft meer pruning tot gevolg. De standaard betrouwbaarheid is 25%.
- Minimum aantal objecten: Het minimum aantal objecten dat in de laatste tak valt. De standaard waarde is twee, dat wil zeggen dat als er drie objecten zijn die in verschillende klassen vallen dat er dan nog een keer gesplitst zal worden totdat de tak 100% zuiver is of totdat er twee of minder objecten zich in de tak bevinden. Een hogere waarde kan gebruikt worden als er veel storing in de data is.

Resultaten
De optimale parameters in dit geval zijn de standaard parameters:

- betrouwbaarheid = 25%,
- pruning,
- minimaal aantal objecten = 2

Deze instellingen toegepast op de classificatie van de gehele dataset geeft een prestatie van 77,58% goed geclassificeerd. De variabelen beroep, typologie, april, bedreigen (*C06*), kopstoot/slaan/schoppen (*C14*), ontrukken (*C15*), schieten (*C20*), steken (*C22*), in verband met politiek (*H04*) worden gekozen door de wrapper.

De confusion matrix is te zien in onderstaande figuur.

a	b	c	d	e	f	g	h	i	j	k	l	m	n	o	p	
10637	6	251	33	0	0	0	0	21	0	1	0	3	0	0	0	a = 0000
198	121	104	129	0	0	0	0	55	15	10	0	50	0	0	0	b = 0001
847	9	2062	30	0	0	0	0	201	2	11	0	43	0	0	0	c = 0010
189	3	10	552	0	0	0	0	61	6	4	0	9	0	0	0	d = 0100
101	1	1	10	0	0	0	0	3	1	0	0	0	0	0	0	e = 1000
6	0	0	11	0	0	0	0	1	0	0	0	1	0	0	0	f = 1100
22	0	33	0	0	0	0	0	13	0	1	0	5	0	0	0	g = 1010
4	4	2	5	0	0	0	0	3	0	1	0	3	0	0	0	h = 1001
86	5	85	84	0	0	0	0	688	4	3	0	189	0	0	0	i = 0110
23	24	6	61	0	0	0	0	30	9	9	0	69	0	0	0	j = 0101
70	26	136	16	0	0	0	0	143	4	38	0	126	0	0	0	k = 0011
5	0	8	6	0	0	0	0	34	0	0	0	19	0	0	0	l = 1110
37	14	25	24	0	0	0	0	124	13	10	0	447	0	0	0	m = 0111
5	1	4	0	0	0	0	0	1	0	2	0	16	0	0	0	n = 1011
0	1	0	5	0	0	0	0	0	0	0	0	5	0	0	0	o = 1101
5	5	1	4	0	0	0	0	12	1	3	0	83	0	0	0	p = 1111

Het is te zien dat verdachten alleen in de volgende categorieën zijn ingedeeld: non-geweld (a), vermogensdelicten met geweld (b), fysiek geweld (c), verbaal geweld (d), fysiek & verbaal geweld (i), verbaal & vermogensdelicten met geweld (j), fysiek & vermogensdelicten met geweld (k), fysiek & verbaal & vermogensdelicten met geweld (m). In de andere categorieën komen geen verdachten voor omdat er geen onderscheidende kenmerken zijn gevonden op basis waarvan gezegd kan worden welk type geweld de verdachte heeft gepleegd.

De beslissingsboom kan helaas niet afgebeeld worden omdat het een te grote figuur is. Wel is het mogelijk regels uit de boom af te leiden. De meest betrouwbare regels die op relatief grote groepen geldig zijn, zijn hieronder gegeven.

Een persoon valt met enige zekerheid in de categorie non-geweld, als:

Betrouwbaarheid = 93% (8890 goed : 621 fout)
bedreigen = 0 EN
kopstoot/slaan/schoppen <= 0,43 EN
typologie = beginner EN
ontrukken <= 10 EN
steken <= 4,46

Betrouwbaarheid = 80% (173 goed : 43 fout)
bedreigen = 0 EN
kopstoot/slaan/schoppen <= 0,43 EN
ontrukken <= 2,78 EN
typologie = doorstromer EN
beroep = onbekend beroep

Een persoon valt met enige zekerheid in de categorie fysiek geweld, als:

Betrouwbaarheid = 79% (1283 goed : 330 fout)
bedreigen = 0 EN
kopstoot/slaan/schoppen > 0,43 EN
ontrukken <= 0,52 EN
typologie = beginner EN
in verband met politiek <= 50

Een persoon valt met enige zekerheid in de categorie vermogensdelicten met geweld, als:

Betrouwbaarheid = 78% (46 goed : 13 fout)
bedreigen = 0 EN
kopstoot/slaan/schoppen <= 0.43 EN
ontrukken > 0.72 EN
typologie = doorstromer

Betrouwbaarheid = 67% (55 goed : 12 fout)
bedreigen = 0 EN
kopstoot/slaan/schoppen <= 0.43 EN
ontrukken > 2.78 EN
typologie = beginner

Een persoon valt met enige zekerheid in de categorie verbaal geweld, als:

Betrouwbaarheid = 70% (45 goed : 19 fout)
kopstoot/slaan/schoppen <= 0.69 EN
bedreigen > 3.8 EN
beroep = onbekend beroep EN
typologie = beginner

Een persoon valt met enige zekerheid in de categorie fysiek & vermogensdelicten met geweld, als:

Betrouwbaarheid = 59% (41 goed : 29 fout)
kopstoot/slaan/schoppen > 0.94 EN
ontrukken > 0 EN
beroep = zonder beroep

Een persoon valt met enige zekerheid in de categorie fysiek & verbaal & vermogensdelicten met geweld, als:

Betrouwbaarheid = 80% (33 goed : 8 fout)
ontrukken > 0 EN <=8.16
typologie = veelpleger EN
beroep = zonder beroep EN
bedreigen > 4.88 EN
schieten = 0 EN
kopstoot/slaan/schoppen > 6.98

Een persoon valt met enige zekerheid in de categorie fysiek & verbaal, als:

Betrouwbaarheid = 71% (146 goed : 59 fout)
kopstoot/slaan/schoppen > 44.44 EN <= 70
ontrukken = 0 EN
typologie = beginner EN
schieten <= 25 EN
bedreigen > 29.03 EN <=58.33
april <= 41.18

Betrouwbaarheid = 70% (176 goed : 76 fout)
kopstoot/slaan/schoppen > 22.5 EN <= 44.44
typologie = beginner EN
ontrukken <= 11.54 EN

Er zijn wel enkele regels gegenereerd die andere soorten geweld aangeven maar die zijn voor minder dan 30 personen geldig en/of hebben een lage betrouwbaarheid. Er zijn geen onderscheidende kenmerken gevonden voor verdachten van seksuele delicten.

Veel regels komen ook voor in combinatie met andere beroepen, met ongeveer dezelfde betrouwbaarheid. Die regels zijn echter van toepassing op veel kleinere groepen en daarom niet weergegeven.

Conclusie
Het algoritme dat in dit geval het beste heeft gewerkt is J48. Er is een prestatie van 77,6% goed geclassificeerde verdachten gehaald. Zelfs op basis van de delictkenmerken kunnen dus niet alle verdachten aan het juiste type geweld toegewezen worden.

De onderscheidende kenmerken zijn niet verrassend. Verdachten van enkel fysiek geweld zijn meestal beginners die met behulp van kopstoot/slaan/schoppen delicten gepleegd hebben en alleen delicten zonder bedreiging. Verdachten van uitsluitend verbaal geweld zijn beginners die juist zonder kopstoot/slaan/schoppen delicten gepleegd hebben, maar met bedreiging. Voor verdachten van gewelddadige vermogensdelicten geldt dat zij beginners of doorstromers zijn die met behulp van ontrukking delicten hebben gepleegd. Van verdachten van gewelddadige seksuele delicten zijn geen onderscheidende kenmerken gevonden.

6.1.5 Classification naar wel/geen geweld op basis van persoons- en delictkenmerken

In deze paragraaf wordt geclassificeerd met dezelfde variabelen als de voorgaande methode. Het verschil is de doelvariabele. De hier gebruikte doelvariabele is geweld. De verdachte kan slechts in twee klassen vallen: geweldpleger of geen geweldpleger.

Prestatie
De prestatie op doelvariabele *geweld* met als dataset een sample van 5% van de verdachten met hun persoons- en delictkenmerken zijn te zien in onderstaande tabel.

	ZeroR	OneR	J48	NaiveBayes	IB1	ADtree	Bagging met J48
Gemiddelde	55,82%	85,27%	**90,15%**	87,87%	83,26%	89,85%	88,63%
Std. Dev	0	0	0,3255	0,2536	0,8038	0,2372	0,8202

Tabel 6-8: Prestatie van enkele algoritmen

Alle algoritmen presteren behoorlijk goed bij deze classificatie. OneR gebruikt attribuut C02 (aanspreken/aanroepen/aanraken) om een relatief goede classificatie te verkrijgen. Volgens die techniek geldt de regel:
Als een verdachte meer dan 1,2% van zijn geregistreerde delicten door middel van aanspreken/aanroepen/aanraken heeft gepleegd, dan is die verdachte ooit aangehouden voor een geweldsdelict.
Die regel heeft een betrouwbaarheid van 85%. Het is een reeds bekend resultaat. Opvallend is dat het resultaat van OneR een variantie van nul heeft. De techniek kiest bij iedere sample dezelfde regel en behaalt daarmee steeds dezelfde prestatie.

De beste techniek is net als bij de vorige classificatie J48 met een resultaat van 90% goed geclassificeerde personen. ADtree presteert bijna net zo goed met een kleinere variantie. Weer geldt dat metatechnieken zoals bagging de prestatie niet verbeteren.

Resultaten
Het veranderen van de parameters beschreven in paragraaf 6.1.4 brengen dit keer wel enige verbetering te weeg. De optimale betrouwbaarheid voor pruning blijft 25% maar het optimale minimum aantal objecten is 4. Dit geeft in de steekproef een verbetering van 0,3% en een halvering van de standaarddeviatie. Het toepassen van J48 met de beste parameters op de gehele dataset geeft 90,84% goed geclassificeerd. Er is dus bijna geen verandering te zien vergeleken met de steekproef. De standaarddeviatie is gedaald naar 0,08.

De volgende variabelen zijn door wrapper gekozen:

som_ant
som_feit
zwaar_r
poging_r
jr_82_r
jr_83_r
jr_99_r
april_r
Segbroek_r
C02_r (aanroepen/aanspreken/aanraken)
C06_r (bedreigen)
C14_r (kopstoot/slaan/schoppen)
C15_r (ontrukken)
C20_r (schieten)
C22_r (steken)
F07_r (jongen/meisje 12-16 jr)
F09_r (man)
F14_r (vrouw)
H04_r (ivm politiek)
H06_r (ivm seksmotieven)

Naast een aantal vanzelfsprekende keuzes zijn er enkele verrassende variabelen gekozen. Zo is zowel het aantal antecedenten als het aantal feiten gekozen terwijl ze sterk van elkaar afhankelijk zijn (de correlatie is 0,93). Voor de techniek NaiveBayes zou dit nadelige gevolgen hebben maar voor J48 is het geen probleem. Daarnaast is opvallend dat er gekozen is voor een aantal jaren, de pleegplaats Segbroek en de pleegtijd april. Ook zijn er slachtofferkenmerken (de F-codes) en motieven (H-codes) bepalend voor het wel of niet plegen van geweld.

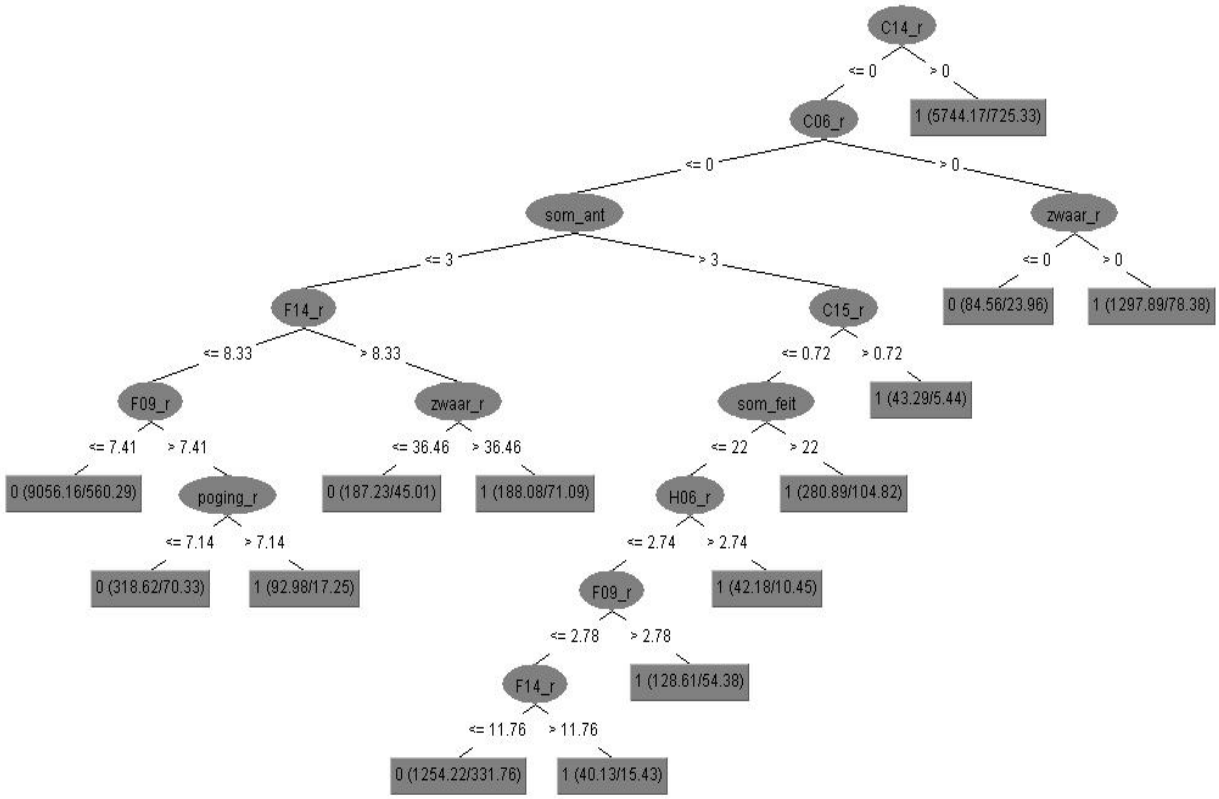
Verder is op te merken dat er geen persoonskenmerken als relevant gezien worden. De boom bestaat geheel uit delictkenmerken.

De confusion matrix van J48 ziet er als volgt uit :

```
    0     1   <-- classified as
-------------
10486   466    0
 1222  6585    1
```

Het is te zien dat er 1222 gewelddadige verdachten geclassificeerd zijn als niet gewelddadig en 466 verdachten die nog nooit opgepakt zijn voor een gewelddadig delict geclassificeerd zijn als gewelddadig. Overige verdachten (17071) zijn goed geclassificeerd.

De beslissingsboom is wederom te groot om weer te geven. Daarom is gekozen voor andere parameters waardoor meer pruning wordt toegepast. De boom in figuur 6-3 is ontstaan uit techniek J48 met een betrouwbaarheid voor pruning van 0,05 en het minimum aantal objecten is 50.



Figuur 6-3: Beslissingsboom van techniek J48. C14= kopstoot/slaan/schoppen, C06 is bedreigen, F14 is vrouwelijk slachtoffer, C15 is ontrukken, H06 is seksmotieven en F09 is mannelijk slachtoffer.

De getallen tussen haakjes in de grijze rechthoeken geven het aantal goed geclassificeerde, respectievelijk het fout aantal geclassificeerde verdachten in die stap. Zo worden door de eerste stap 5744 verdachten goed geclassificeerd tegenover 725 fout geclassificeerd. De decimale getallen hebben te maken met de manier waarop J48 met missende waarden om gaat.

Verder is te zien dat de variabelen *april* en *Segbroek* niet meer gebruikt worden in deze boom. Ze zijn als het ware “gesnoeid”. Alleen de hoofdtakken zijn overgebleven. De belangrijkste variabele is kopstoot/slaan/schoppen (*C14*).

Op basis van de boom in figuur 6-3 zijn de volgende profielen op te stellen:

Er zitten relatief veel geweldplegers in de groepen verdachten die…

- .. minimaal één keer een delict hebben gepleegd met de methode kopstoot/slaan/schoppen.
  95% daarvan pleegt geweld.
- .. minimaal één delict hebben gepleegd met de methode bedreigen.
  98% pleegt geweld van deze groep.
- .. minimaal één delict met bedreigen en minimaal één zwaar delict hebben gepleegd.
  99% pleegt geweld van deze groep. Overigens is 99% van de bedreigingen een zwaar delict.
- .. meer dan drie antecedenten hebben gepleegd.
  76% pleegt geweld tegenover 28% die minder dan 3 delicten heeft gepleegd.
- .. één of meer van hun feiten met de methode ontrukken hebben gepleegd.
  98% daarvan pleegt geweld tegenover 41% geweldplegers die nooit methode ontrukken hebben gebruikt.
- .. bij één of meerdere delicten een seksmotief hadden.
  77% pleegt geweld van deze groep. 42% van de groep zonder seksmotief pleegt geweld.
- .. bij één of meerdere delicten het slachtoffer een vrouw was.
  91% pleegt geweld tegenover 31% als het slachtoffer geen vrouw was.
- .. bij één of meerdere delicten het slachtoffer een man was.
  90% pleegt geweld tegenover 25% als het slachtoffer geen man was.
- .. minimaal één poging hebben gepleegd.
  90% geweld tegenover 32% van de verdachten die nooit een poging hebben gepleegd.

Deze profielen zijn intuïtief logisch en ondersteunen hetgeen al bekend was.

Conclusie
De methode waar geclassificeerd wordt naar *soort geweld* op basis van zowel persoonskenmerken als delictkenmerken werkt het beste. De prestatie is 90,84% goed geclassificeerd met behulp van de techniek J48. Deze techniek werkt precies als het bekende C4.5, waarvan ook een versie in Clementine van SPSS geïmplementeerd is. De belangrijkste onderscheidende variabelen zijn aanspreken/aanroepen/ aanraken (*C02*), bedreigen (*C06*) en *zwaar*, die aangeeft of de verdachte één of meerdere zware delicten heeft gepleegd.

6.2 Exploratieve analyse geweldplegers

In deze paragraaf wordt een exploratieve analyse gedaan van alle geweldplegers die aangehouden zijn in 2004. Geweldplegers zijn gedefinieerd als één of meer keer in 2004 aangehouden verdachten die in hun carrière één of meerdere malen voor een gewelddadig delict zijn aangehouden zoals in paragraaf 5.5 beschreven is. Association rules is de datamining methode die voor exploratieve analyse uitermate geschikt is. De methode zoekt betrouwbare relaties tussen alle variabelen.

6.2.1 Werkwijze

Algoritme: Apriori
Er zijn verschillende technieken in Weka aanwezig die association rules toepassen, namelijk: Apriori, PredictiveApriori en Tertius. Er is gekozen voor één techniek: Apriori. Dit is één van de meest bekende technieken. Een voordeel van Apriori is dat het algoritme snel is, vergeleken met Tertius en PredictiveApriori.

Apriori bestaat uit twee fasen. In de eerste fase worden de vaak voorkomende attribuutverzamelingen bepaald. Dat zijn verzamelingen die tenminste aan de gegeven minimale frequentie voldoen en dus voorkomen in tenminste een bepaald percentage van alle data. In de tweede fase worden association rules gemaakt van de in stap 1 gevonden frequente attribuutverzamelingen [*4*].

De parameters die Apriori heeft, zijn: delta, minimum betrouwbaarheid en minimum en maximum frequentie. Delta is de factor waarmee de frequentie in stapjes verlaagd wordt totdat de minimum frequentie is bereikt of totdat het gevraagd aantal relaties is bereikt. Deze staat ingesteld op 0,05, ofwel 5%. De minimale frequentie staat in Weka standaard op 0,1, dat wil zeggen dat de gehele relatie minstens in 10% van de gevallen moet gelden. Een nadeel daarvan is dat regels die niet vaak

voorkomen maar die wel een hoge betrouwbaarheid hebben, niet getoond worden. De frequentie dient voor dit onderzoek dus laag te zijn en de betrouwbaarheid hoog.

Gebruikte variabelen
Er is een andere dataset gebruikt dan bij classification. De nieuwe dataset bevat uitsluitend in 2004 aangehouden verdachten die in hun criminele loopbaan één of meer gewelddadige delicten hebben gepleegd, terwijl bij classification ook verdachten die nooit voor geweld zijn aangehouden meegenomen werden. De nieuwe dataset bestaat uit 7807 verdachten. Net als bij classification is eerst gekeken naar de persoonskenmerken van de gewelddadige verdachten en daarna ook naar de delictkenmerken. De variabelen die gebruikt worden voor de persoonskenmerken zijn dezelfde als beschreven in paragraaf 1 van dit hoofdstuk. Aan de delictkenmerken zijn alle rubrieken toegevoegd zoals r101, r103, etcetera. Ze zijn weergegeven als binaire variabelen: 1 als de verdachte ooit een gewelddadig delict heeft gepleegd, 0 voor andere delicten.

Daarnaast zijn er een aantal variabelen uit de dataset verwijderd. Het gaat hierbij om: nationaliteit, geboorteland en het aantal antecedenten. Deze laatste is verwijderd omdat de typologie (beginner, doorstromer, veelpleger) hierop gebaseerd is. De nationaliteit en geboorteland maken samen de variabele *cbs_etn* (etniciteit volgens definitie van het CBS). Dit zorgt voor veel triviale regels, als:

```
geb_land=NL cbs_etn=Nederland 10247 ==> nat_code=NL 10247 conf:(1)
```

Om dit te voorkomen zijn nationaliteit en geboorteland verwijderd.

Variabelen aanpassen
Een eis die de techniek Apriori stelt is dat alle variabelen nominaal zijn. Van alle delictkenmerken die relatief waren gemaakt ten opzichte van het aantal antecedenten of feiten (zoals de jaren, rubrieken, zwaar en poging) is een binaire variabele gemaakt. Deze is 1 als één of meer van de gepleegde delicten aan dat kenmerk voldeed en 0 in andere gevallen. Andere numerieke variabelen, zoals leeftijd, zijn omgezet naar nominale variabelen met een klassenindeling. Dit heet discretiseren. Discretiseren is het splitsen van waarden van een continue variabele in een beperkt aantal intervallen. Ieder interval wordt vervolgens behandeld als een discrete waarde (een categorie). Dit kan in Weka gemakkelijk met een discretize filter waarbij je zelf op kunt geven in hoeveel klassen een numerieke variabele ingedeeld moet worden of dat alle klassen dezelfde frequentiebreedte moet hebben. Bij het discretiseren is niet gelet op de verdeling van een doelvariabele omdat bij association rules niet alleen naar de doelvariabele gekeken wordt maar ook naar regels tussen andere variabelen. Weka zoekt zelf naar het optimaal aantal klassen met gelijke breedte.

De dataset voldoet nu aan de eisen die voor association rules gesteld worden. Er worden echter niet zomaar nuttige resultaten verkregen. Apriori creëert duizenden regels waarvan de meeste triviaal zijn zoals:

```
land_cd=NL ==> g7=nee (conf: 0.99)
woonregio=Haaglanden ==> land_cd=NL (conf: 1)
```

De eerste regel klinkt als een nieuw feit. De gevallen dat `g7=ja` (vluchtgevaarlijk) zijn er echter maar zo weinig, dat regels met `g7=nee` op zichzelf al een hogere betrouwbaarheid hebben dan 90%. In combinatie met alle attributen zijn ze dus betrouwbaar en dat soort regels komen erg vaak voor. Voor alle zogenoemde g-codes geldt dat ze vaak ‘nee’ zijn. Samen zorgen die voor duizenden regels die geen kennis toevoegen. Een manier om dit te voorkomen is overal ‘nee’ te vervangen door een missende waarde (‘?’ in Weka). Op die manier kunnen er geen regels gemaakt worden waarbij het antwoord ‘nee’ is maar wel met antwoorden die ‘ja’ zijn. Op die manier is het niet nodig een gehele variabele te verwijderen maar slechts één waarde ervan.
De tweede regel is intuïtief wel logisch. Als een verdachte in Haaglanden woont, dan woont deze in Nederland. Om dit soort regels te vermijden is ook regio Haaglanden op missend gezet. Met eenzelfde redenatie zijn ook de waarden `land_cd=NL` en `ovl=nee` op missend gezet.

Voor de meeste numerieke variabelen geldt dat ze vaak nul zijn. Ook deze waarden zijn op missing gezet om regels als `R114=0 ==> R115=0` te voorkomen. Alleen regels waar de waarde 1 is zijn interessant. Van de variabelen zwaar en poging zijn de nullen niet op missing gezet want die waren niet zo vaak nul en dus is het juist interessant om te weten wanneer die nul zijn.

6.2.2 Association rules van persoonskenmerken

Er is gekozen voor de minimum frequentie = 0,01 en minimum betrouwbaarheid = 0,8. Dat wil zeggen dat het algoritme alleen relaties geeft die van toepassing zijn op meer dan 1% van de data (78 verdachten). Het geval `g8=ja` (zelfmoordneiging) zal daarom niet in de relaties voorkomen omdat er slechts 27 verdachten zijn die met zelfmoordneiging geregistreerd staan en dat is minder dan 1% van de dataset.

Na het aanpassen van de dataset worden er 836 relaties gevonden met Apriori. Daarvan zijn slechts een beperkt aantal relevant. Voor verdachten die in 2004 zijn aangehouden en in het verleden minimaal één keer aangehouden zijn voor een gewelddadig delict gelden de volgende regels:

Regel 1
Van de gewelddadige, van Turkse etniciteit afkomstige verdachten van 18 t/m 25 jaar, woont 86% in steden met meer dan 250.000 inwoners.

Het gaat totaal om een groep van 166 verdachten, waarvan er 143 in een grote stad wonen. In 97% van de gevallen is die grote stad Den Haag. Ter vergelijking: van alle gewelddadige verdachten woont 63% in een gemeente met meer dan 250.000 inwoners en van alle Turkse, gewelddadige verdachten woont 80% in een grote stad (zie regel 4).

Regel 2
Inwoners van Scheveningen die uitsluitend fysiek geweld hebben gepleegd hebben in 84% van de gevallen de Nederlandse etniciteit.

Van alle verdachten die wonen in Scheveningen heeft 78% de Nederlandse etniciteit en van alle gewelddadige Scheveningse verdachten is dat 79%. De verdachten die fysiek geweld plegen hebben dus relatief iets vaker Nederlandse etniciteit dan andere inwoners van Scheveningen. Het gaat om een groep van 159 verdachten waarvan 133 van Nederlandse etniciteit.

Regel 3
Van de druggebruikende, gewelddadige verdachten met Surinaamse etniciteit woont 82% in steden met meer dan 250.000 inwoners.

De groep druggebruikende, gewelddadige, Surinaamse verdachten bestaat uit 147 verdachten. Er wonen er 124 in Den Haag en 4 in een grote stad buiten de regio. Van alle gewelddadige Surinamers woont 76% in zo'n grote stad.

Regel 4
80% van de gewelddadige verdachten met Turkse etniciteit wonen in steden met meer dan 250.000 inwoners.

Er zijn 547 gewelddadige Turkse verdachten waarvan er 440 in een gemeente wonen met meer dan 250.000 inwoners.

Regel 5
Van de Westlanders die uitsluitend fysiek geweld hebben gepleegd heeft 80% de Nederlandse etniciteit.

De groep Westlanders die fysiek geweld hebben gepleegd bestaat uit 151 personen. Terwijl van alle gewelddadige verdachten wonend in het Westland slechts 73% Nederlands is en van alle verdachten in het Westland ook weer 80% Nederlandse etniciteit heeft. De groep Nederlandse fysiek geweldplegers valt dus eigenlijk niet uit de toom terwijl het aantal Nederlanders in de groep geweldplegers wel aan de lage kant is.

Mannen
Van alle gewelddadige verdachten is het grootste deel mannelijk, namelijk 90%. Dit was al bekend. Er zijn echter groepen te vinden die uitzonderlijk veel mannen bevatten. De groepen die een hoger percentage mannen dan 97% bevatten zijn in tabel 6-9 beschreven.

Nr	Groep geweldplegers met..	Aantal verdachten	Aantal mannen	Percentage mannen
1	soort geweld=verbaal & vermogen en woonplaats>250.000 inwoners	165	164	99%
2	woongebied=Zuiderpark en etniciteit=Turkije	84	83	99%
3	etniciteit=Marokko en soort geweld=verbaal & fysiek & vermogen	152	150	99%
4	gemeente>250.000 inwoners en leeftijd=[18-25] en soort geweld=fysiek & vermogen	119	117	98%
5	soort geweld=verbaal & vermogen	231	227	98%
6	woonplaats>250.000 inwoners en etniciteit=Turkije en leeftijd=[26-33]	105	103	98%
7	etniciteit=Marokko en druggebruik=ja	101	99	98%
8	etniciteit=Nederland en soort geweld=verbaal & vermogen	100	98	98%
9	woongebied=Segbroek & leeftijd=[26-33]	100	98	98%
10	woonplaats>250.000 inwoners en leeftijd=[34-42] en soort geweld=verbaal & fysiek & vermogen	144	141	98%
11	woonplaats>250.000 inwoners en leeftijd=[18-25] en soort geweld=verbaal & fysiek & vermogen	93	91	98%
12	etniciteit=Suriname en soort geweld=verbaal & fysiek & vermogen	88	86	98%
13	woongebied=Pijnacker/Nootdorp en etniciteit=Nederland	80	78	98%

Tabel 6-9: Groepen met opvallend veel mannen

6.2.3 Association rules van persoons- en delictkenmerken

Het aantal gevonden relaties op de dataset met 7807 verdachten en 186 variabelen is buitengewoon veel. Zelfs als er alleen gevraagd wordt naar relaties met een hogere betrouwbaarheid dan 90%, zijn er nog steeds meer dan een miljoen relaties. Veel van die gevonden relaties bestaan uit slechts delictkenmerken en hebben de volgende vorm:

```
delictkenmerk ==> delictkenmerk                    (1)
```

Een voorbeeld van zo'n regel is:

```
r133_r=1  Semi_r=1  Openb_r=1  Maandag_r=1  Donderdag_r=1  Avond_r=1  1154  ==>
Middag_r=1 1117    conf:(0.97)
```

Uit zo'n relatie kan niet veel opgemaakt worden. De linkerkant van de pijl zegt namelijk al dat de verdachte meerdere delicten heeft gepleegd omdat er zowel maandag als donderdag in voorkomt en ook semi-openbaar en openbaar. De kans dat die persoon dan ook een delict in de middag heeft gepleegd wordt dan uiteraard al groter. Er kan niet uit opgemaakt worden dat rubriek 133 (eenvoudige diefstal) grotendeels in de middag gepleegd worden. Het kan namelijk ook zo zijn dat er verdachten zijn die ook een delict in rubriek 131 hebben gepleegd en dat die precies in de middag plaats vond. De betrouwbaarheid en de frequentie zijn daarom niet plausibel. Als men meer wil weten over de relaties die tussen delictkenmerken bestaan, dan moet een andere extractie de basis vormen van het datamining proces, namelijk de extractie die uit alle unieke delicten bestaat. De dataset die voor dit onderzoek gebruikt is bevat unieke personen en daardoor zijn de delictkenmerken geaggregeerd.

Relaties die wel interessant zijn, hebben één van de vormen:

```
persoonskenmerk ==> delictkenmerk                  (2)
```

```
delictkenmerk ==> persoonskenmerk                    (3)
```

Ook interessant zijn relaties van de vorm:

```
persoonskenmerk ==> persoonskenmerk                  (4)
```

Maar die zijn al in paragraaf 6.2.2 besproken.

Om alleen relaties van de vorm (2) en (3) te krijgen is er in Visual Basic for Applications in Word een programma geschreven waarmee een filter toegepast wordt die alle regels die niet in de goede vorm staan verwijdert. Van vorm (2) worden er geen nuttige relaties gevonden en van vorm (3) slechts drie relaties. Ze worden hieronder beschreven.

Regel 1
Van de gewelddadige verdachten die minstens één fietsendiefstal hebben gepleegd, één of meerdere delicten hebben gepleegd in de jaren 1987, 1988 en 1989 en één of meerdere delicten in het bureaugebied Overbosch is 91% harddruggebruiker.

Regel 2
Van de gewelddadige verdachten die minstens één fietsendiefstal hebben gepleegd, die één of meerdere delicten hebben gepleegd in de jaren 1988, 1989 en 1990 en één of meerdere delicten in het bureaugebied Overbosch is 91% harddruggebruiker.

Regel 3
Van de gewelddadige verdachten die minstens één fietsendiefstal hebben gepleegd, die één of meerdere delicten hebben gepleegd in de jaren 1987 en 1989 en één of meerdere delicten in het bureaugebied Karnebeek is 91% harddruggebruiker.

Alle drie de regels geven aan dat het aantal harddruggebruikers hoog is (91%) bij groepen verdachten die verdeeld over een aantal jaren feiten hebben gepleegd, in een bepaald bureaugebied een feit hebben gepleegd en minstens één keer voor rubriek 201 (diefstal van fietsen) zijn aangehouden. Deze relaties lijken niets met geweld te maken te hebben maar het gaat hier om de dataset met alleen geweldplegers. Deze verdachten hebben dus naast de fietsendiefstal één of meerdere geweldsdelicten gepleegd.

Het feit dat er meerdere jaren voorkomen in de regels, geeft aan dat de verdachten meer dan één delict hebben gepleegd. Opvallend is dat de relatie tussen fietsendiefstal en druggebruik alleen niet zo sterk is. Van alle verdachten die een fietsendiefstal hebben gepleegd is namelijk maar 36% druggebruiker, nog lang geen 91%. Ook ligt het druggebruik niet enkel en alleen aan de bureaugebieden of aan de jaren. Het is de combinatie van de drie die ervoor zorgt dat er een hoog aantal druggebruikers wordt geconstateerd.

6.3 Verborgen groepen in geweldplegers

Verborgen groepen kunnen opgespoord worden met behulp van clustering. Net als bij de exploratieve analyse worden alleen de verdachten van geweld geanalyseerd. Er wordt dus gezocht naar samenhangende groepen binnen de in 2004 aangehouden verdachten die minimaal één keer in hun loopbaan een gewelddadig delict hebben gepleegd zoals beschreven in paragraaf 5.1.

6.3.1 Werkwijze

In deze paragraaf wordt gezocht naar natuurlijke groepen in de dataset van alle geweldplegers. Als die groepen eenmaal gevonden zijn dan kan er met classification gekeken worden of er bepaalde groepen zijn die veelal één soort geweld plegen of dat die scheiding niet te vinden is. Gelijktijdig kunnen er ongebruikelijke gegevens gevonden worden, de zogenaamde uitschieters. Het zou namelijk kunnen dat deze uitzonderlijke gevallen apart in een cluster terecht komen.

Iedere clustering kan verschillend zijn. Er zijn verschillende indelingen te bedenken. Zo zegt Dr. P. Kruize [*18*]:
"Aan de ene kant kunnen we een verdeling naar betrokkenen maken tussen a) familie en huisgenoten, b) bekenden, c) onbekenden. Anderzijds kunnen we geweldsincidenten indelen naar het domein waarbinnen het voorval plaatsvindt: 1) privaat, 2) semi-openbaar en c) openbaar."
Maar ook kunnen de veelplegers, beginners en doorstromers drie groepen gaan vormen, of juist groepen op basis van leeftijd.

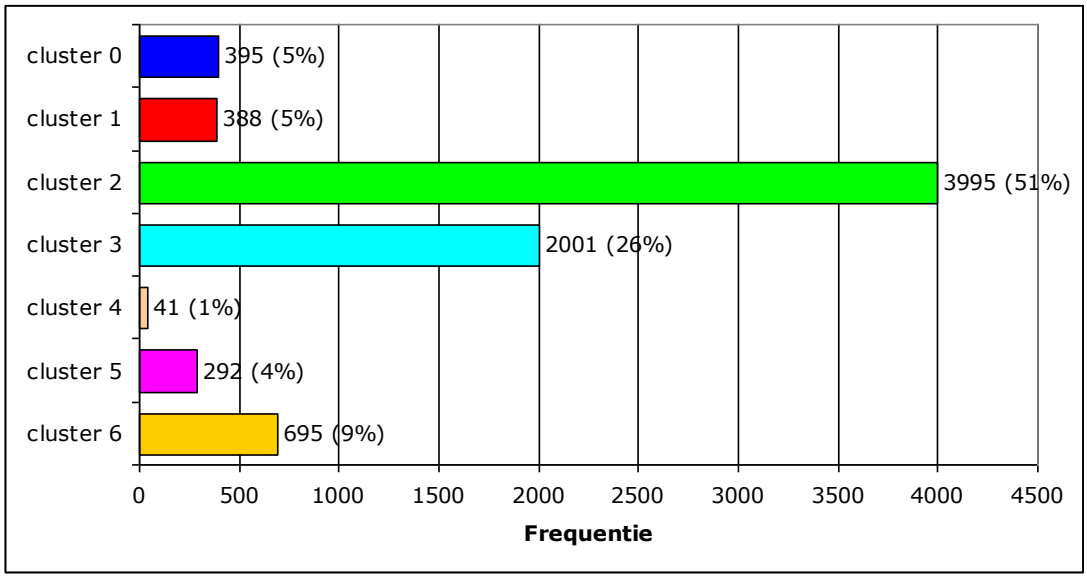
Algoritme: Expected Maximization (EM)
In Weka zijn de volgende clustertechnieken geïmplementeerd: Cobweb, Farthestfirst, MakeDensityBasedClusterer, SimpleKMeans en EM. De laatste twee zijn de meest bekende. K-means is ook geïmplementeerd in het statistische pakket van SPSS. De techniek die voor dit onderzoek is gebruikt, is EM, wat staat voor Expectation Maximization. Dit algoritme is een soort k-means met als belangrijkste verschil dat EM zelf het optimaal aantal clusters bepaalt.

Het EM algoritme begint met het maken van één cluster waar het vervolgens de log-likelihood van berekent. Vervolgens wordt het aantal clusters met één verhoogd en opnieuw de log-likelihood berekend. Dit proces herhaalt zich totdat de log-likelihood niet meer groter wordt. Het beste aantal clusters is dan gevonden [*8*].

Bij deze techniek kunnen onder andere het aantal iteraties opgegeven worden en het aantal clusters. In deze analyse is 50 iteraties gekozen. Het aantal clusters is niet opgegeven omdat het doel van deze analyse is erachter te komen uit hoeveel clusters de dataset bestaat. EM kan het optimaal aantal clusters zelf bepalen.

Dataset
Net als bij association rules en classification is hier weer gekozen voor twee stappen. Eerst is de dataset met alle persoonskenmerken gebruikt. De variabelen die daarbij horen zijn dezelfde als beschreven in paragraaf 6.1. De gevonden clusters worden beschreven in paragraaf 6.3.2. Vervolgens is de clustering op zowel persoons- als delictkenmerken toegepast zoals beschreven in paragraaf 6.1. De resultaten daarvan staan in paragraaf 6.3.3.
Clustering wordt ook wel "unsupervised classification" genoemd. Dat wil zeggen dat er klassen gemaakt worden zonder dat er naar één specifieke variabele (de doelvariabele) wordt gekeken. De clusteringen worden gemaakt op basis van meerdere variabelen. Daarom is de doelvariabele *soort geweld* niet gebruikt in de clustering. Achteraf is van de gevonden clusters wel bekeken of ze een bepaald soort geweld veel of juist weinig bevatten.

6.3.2 Clustering op basis van persoonskenmerken
Met de EM-clustering zijn er zeven clusters gevonden. Het aantal verdachten in ieder cluster is te zien in onderstaande grafiek.



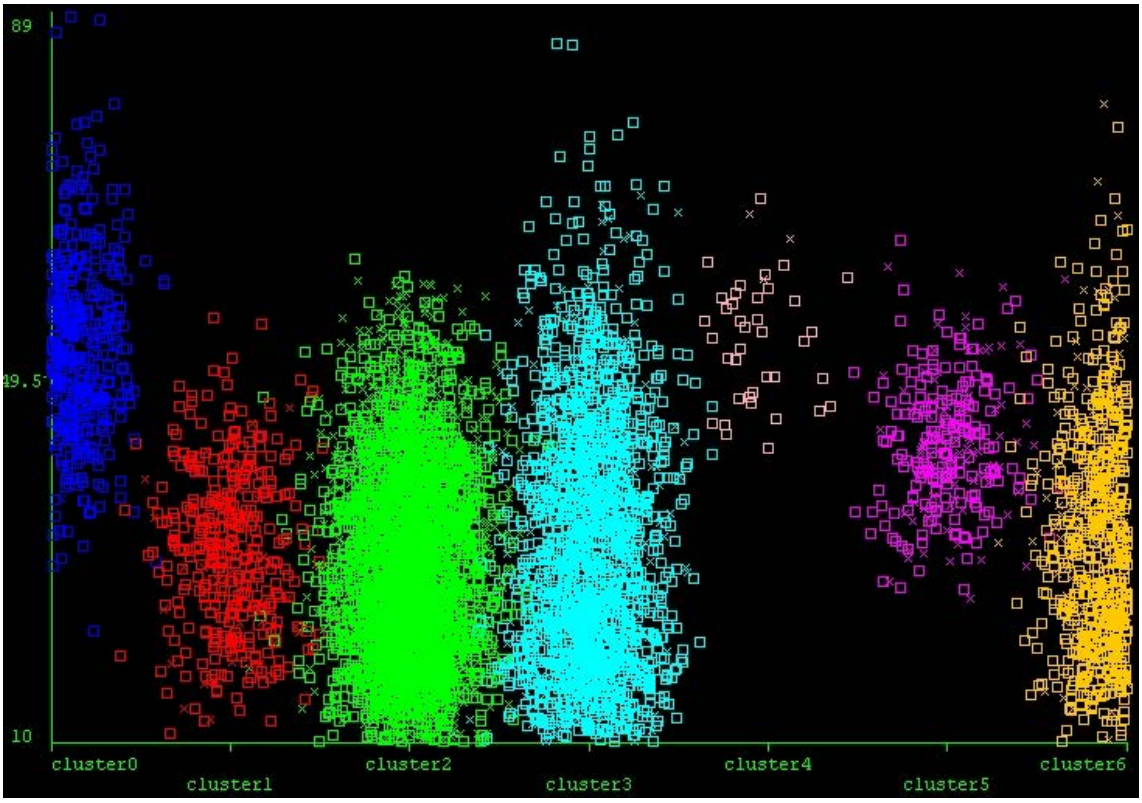
Figuur 6-4: Aantal verdachten per cluster

Het is te zien dat cluster 2 en 3 behoorlijk groot zijn. De anderen zijn relatief klein. De gebruikte kleuren in deze grafiek komen overeen met de kleuren van overige grafieken in deze paragraaf.

De clusters zijn veelal op geografische kenmerken te onderscheiden[3]. Hieronder zijn ze stuk voor stuk beschreven aan de hand van hun meest opvallende variabelen. Bij geen van alle clusters was het beroep een interessante variabele. De verdeling daarvan was altijd min of meer hetzelfde.

Cluster 0: "De oudere Hagenaars"
Bijna alle verdachten in dit cluster wonen in een grote gemeente binnen de regio, aldus Den Haag (94%). De Nederlandse etniciteit komt veel voor, namelijk 82%. Ter vergelijking: 51% van alle gewelddadige verdachten heeft de Nederlandse etniciteit. Verder is opvallend dat er in deze groep geen harddruggebruikers zijn. Er zijn wel zestien alcoholverslaafden.
De gemiddelde leeftijd ligt erg hoog, namelijk 49 jaar (zie figuur 6-5). De leeftijd heeft een standaarddeviatie van 14, redelijk hoog.

Gezien de hoge leeftijd van deze verdachten, reist het vermoeden dat dit de seksueel geweldsverdachten zijn. Nadere analyse levert echter dat maar 7,8% seksueel geweld heeft gepleegd (tegenover 5,9% van alle geweldplegers). Het enige soort geweld dat afwijkt van de verdachten buiten het cluster is vermogenscriminaliteit met geweld. Deze is laag (18,5% heeft ooit vermogenscriminaliteit gepleegd) ten opzichte van andere verdachten (30%).



Figuur 6-5: Op de horizontale as de zeven clusters en op de verticale as de leeftijden

Cluster 1: "De adres onbekend verdachten"
Van alle verdachten in dit cluster is het niet bekend in welke gemeente en welke regio zij wonen. Wel is bekend dat ze in Nederland wonen. Hun gemiddelde leeftijd is 29 jaar met een standaarddeviatie van 7 jaar. Hun etniciteit is zeer uiteenlopend. Zij hebben geen alcoholverslaving maar 13% is wel druggebruiker.

[3] Dat kan komen doordat er zich meerdere afhankelijke variabelen in de dataset bevinden die aangeven waar de verdachte woont. Die variabelen zijn: bureaugebied, woonregio, gemeente categorie en woonland.

Dat van de verdachten geen woonadres geregistreerd is, wil niet per se zeggen dat dit zwervers zijn. Er bevinden zich bovendien geen alcoholverslaafden in dit cluster. Een vermoeden is dat de groep uit veel (ex-) gedetineerden bestaat. Deze mensen hebben geen adres als ze nog in de gevangenis zitten, maar ook als ze net vrij zijn gekomen hebben ze niet gelijk een vast woonadres.

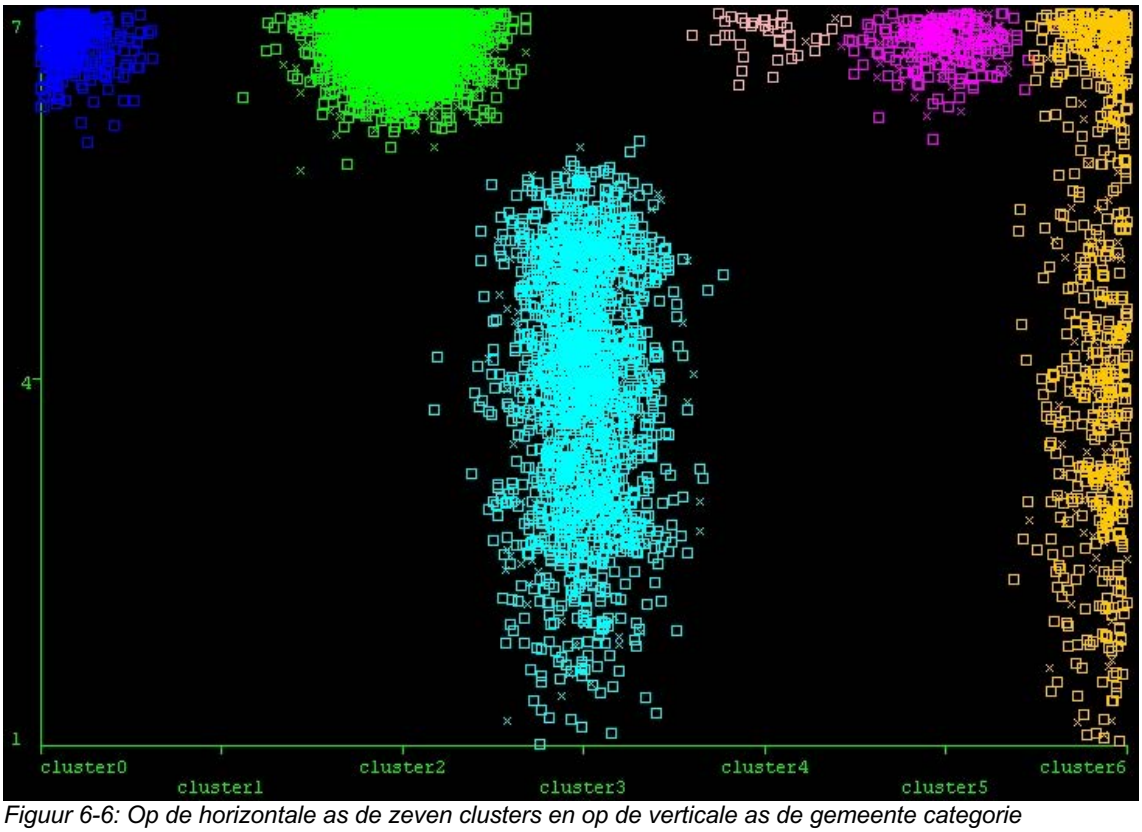
Ze plegen allerlei soorten geweld maar bijna geen seksuele delicten. 16% heeft zowel verbaal, fysiek als vermogensgeweld gepleegd in zijn/haar carrière. In de gehele dataset is dat percentage lager, namelijk 9%. Ook geldt dat vermogensgeweld relatief hoog is: 42% tegenover 30% in de gehele dataset.

Cluster 2 : "De jongere Hagenaars"
De verdachten in dit cluster wonen net als in cluster 0 allemaal in Den Haag, ofwel een grote stad binnen de regio. Hun gemiddelde leeftijd ligt op 30 jaar met een standaarddeviatie van 10 jaar. Er zijn relatief weinig alcoholverslaafden maar het druggebruik is hoger dan normaal (6,5% tegenover 1,9% in de hele dataset).

Deze groep is representatief te noemen voor alle gewelddadige verdachten. De soorten geweld in het cluster hebben namelijk dezelfde verdeling. Er is dus geen soort die opvallend veel of weinig voorkomt.

Cluster 3: "De niet Hagenaars"
Deze in Nederland wonende verdachten wonen in gemeenten met minder dan 250.000 inwoners, binnen regio Haaglanden. Dit zijn de gemeenten Delft, Zoetermeer, Westland, Rijswijk, Wassenaar, Leidschendam/Voorburg en Pijnacker/Nootdorp. In figuur 6-6 is duidelijk te zien dat dit cluster geen verdachten uit grote steden bevat. De gemiddelde leeftijd in de groep is 31 jaar met een standaarddeviatie van 13 jaar. Slechts 1% is alcoholverslaafd en 4% is druggebruiker.

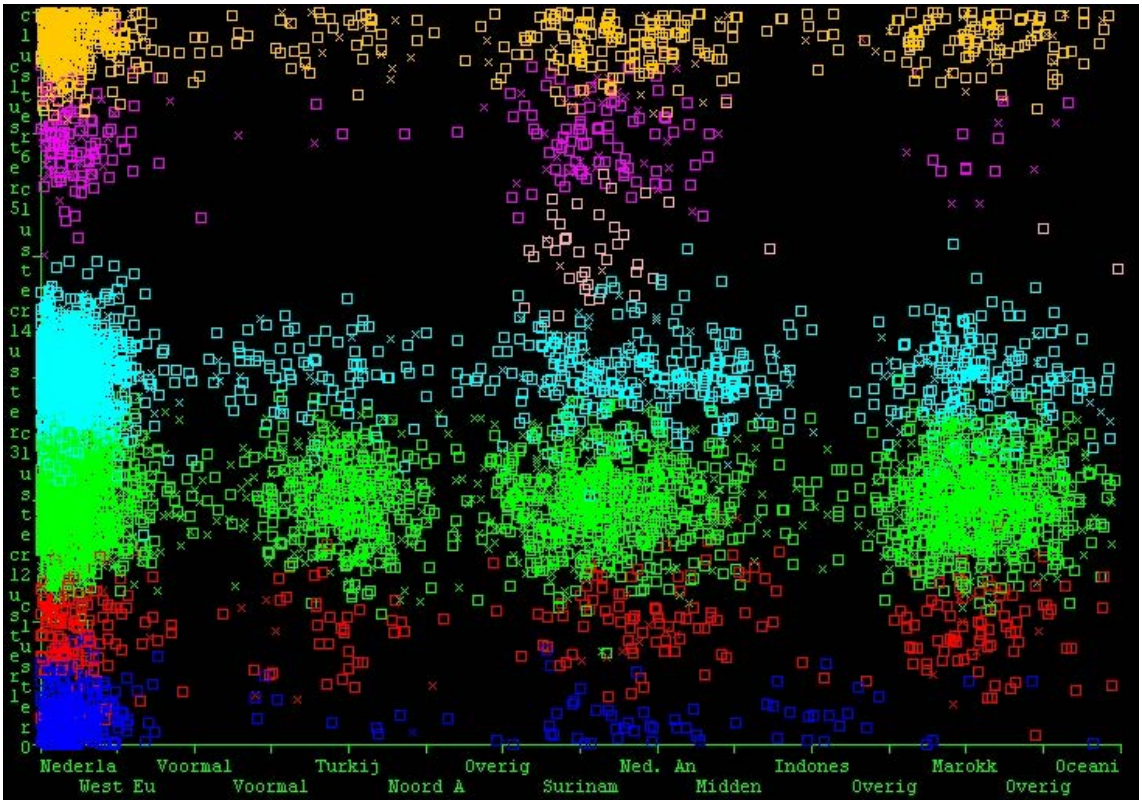
De groep heeft iets minder verdachten die vermogenscriminaliteit met geweld hebben gepleegd (20%) dan in de dataset met alle geweldsverdachten (30%). Verder zijn de aantallen van *soort geweld* niet opvallend.



Figuur 6-6: Op de horizontale as de zeven clusters en op de verticale as de gemeente categorie

Cluster 4: "De oudere, mannelijke Surinamers in Den Haag"
Verdachten in deze groep wonen allemaal in Den Haag, net als cluster 0. Anders is dat er geen vrouwen in aanwezig zijn. Bovendien hebben bijna alle verdachten de Surinaamse etniciteit, zie figuur 6-7. De gemiddelde leeftijd ligt erg hoog, 52 jaar, met een relatief kleine standaarddeviatie van 7 jaar. De verslaving is erg hoog: 20% is alcoholverslaafd en 10% is druggebruiker.

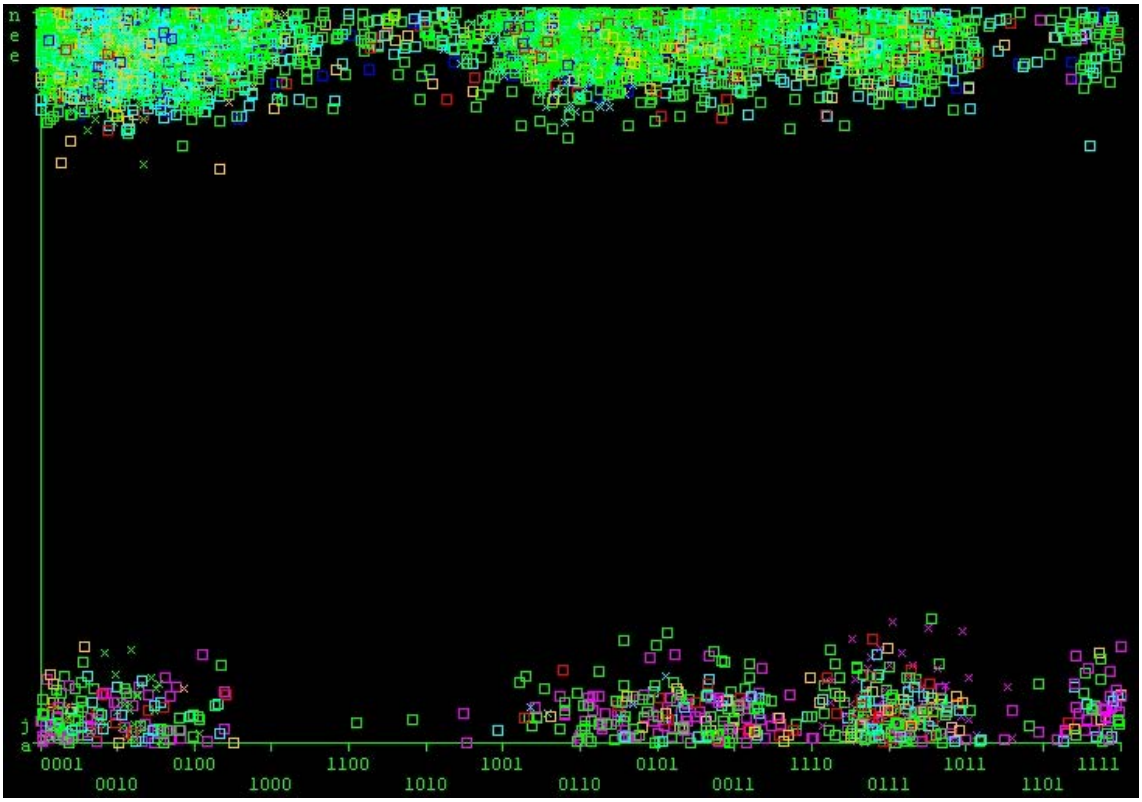
In dit cluster zitten relatief veel verdachten die opgepakt zijn voor een seksueel gewelddadig delict. Van alle gewelddadige verdachten is dit bij 6% het geval. In dit cluster is welliefst 17% ooit opgepakt voor een dergelijk delict. Er moet echter opgemerkt worden dat deze groep bestaat uit slechts 41 mensen.



Figuur 6-7: Op de horizontale as de etniciteiten, op de verticale as cluster 0 tot en met 6

Cluster 5: "De verslaafden"
In dit cluster is 92% van de verdachten druggebruiker en 24% heeft een alcoholverslaving. De groep bevat relatief veel verdachten met Surinaamse etniciteit (28%), maar ook de Nederlandse (52%), Antilliaanse (10%) en Marokkaanse (5%) etniciteit zijn vertegenwoordigd. De gemiddelde leeftijd ligt hier relatief hoog, 44 jaar. Die van de hele groep gewelddadige verdachten ligt namelijk op 32. Bijna alle verdachten (97%) wonen in Den Haag. Van de overige 3% is niet bekend waar zij wonen.

In deze groep bevinden zich veel verdachten die meer dan één soort geweld hebben gepleegd. Zo heeft 11% alle vier de geweldsoorten eens gepleegd (1,5% in de gehele dataset) en 34% heeft zowel verbaal, fysiek en vermogensgeweld gepleegd (9% in gehele dataset). Dit cluster is dus een zeer gewelddadige groep verdachten. In figuur 6-8 is te zien hoe druggebruikers verdeeld zijn over de soorten geweld. Druggebruikers plegen nauwelijks seksuele geweldsdelicten. De paarse punten zijn verdachten van cluster 5. Het is te zien dat deze vaak druggebruiker zijn.



Figuur 6-8: Op de horizontale as soort geweld en op de verticale as druggebruik (g2) in cluster 5.

Cluster 6: "De niet Haaglanders"
Deze verdachten wonen allemaal buiten regio Haaglanden, maar wel in Nederland. Zij wonen in gemeenten van verschillende grootte en hebben uiteenlopende etniciteiten. Hun gemiddelde leeftijd is 33 jaar met een standaarddeviatie van 12 jaar. Zowel het aantal druggebruikers als het aantal alcoholverslaafden is niet uitzonderlijk.

Men vermoedt dat dit de groep uitgaansgeweld is. Mensen uit andere regio's komen naar Den Haag om uit te gaan en plegen daar een vorm van geweld. In dat geval zouden fysiek en verbaal geweld vaak voor moeten komen. Dat is echter niet het geval: 39% heeft verbaal geweld gepleegd en 70% fysiek geweld. In de gehele dataset liggen deze aantallen niet anders: 40% verbaal geweld en 76% fysiek geweld.

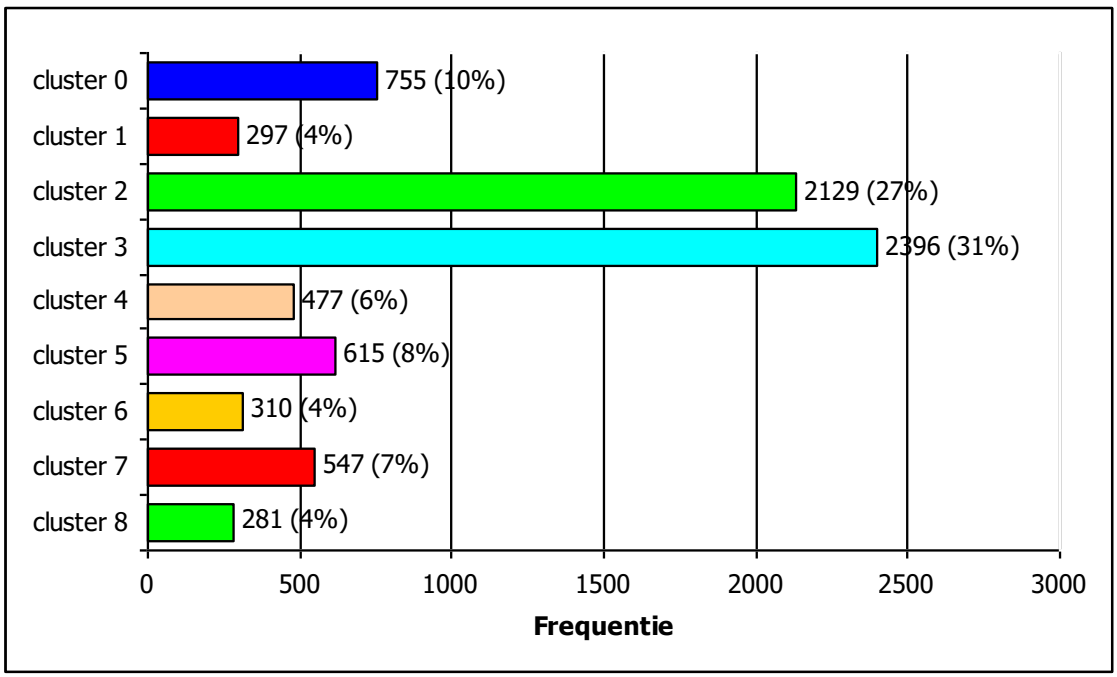
Soort geweld
Samenvattend zijn de soorten geweld als volgt over de clusters verdeeld:

Hele dataset	cluster 0	cluster 1	cluster 2	cluster 3	cluster 4	cluster 5	cluster 6	soort geweld
8,7%	2,5%	10,8%	10,9%	5,2%	0,0%	8,2%	9,5%	**0001**
41,1%	46,8%	30,7%	38,8%	50,6%	31,7%	8,9%	43,2%	**0010**
10,7%	10,1%	10,3%	10,8%	10,0%	9,8%	3,1%	16,0%	**0100**
1,5%	2,5%	0,3%	1,5%	1,6%	7,3%	0,0%	1,3%	**1000**
0,2%	0,5%	0,3%	0,3%	0,1%	0,0%	0,0%	0,4%	**1100**
0,9%	1,3%	1,5%	1,0%	0,9%	0,0%	0,0%	1,0%	**1010**
0,3%	0,3%	0,0%	0,4%	0,2%	0,0%	0,7%	0,1%	**1001**
14,7%	18,5%	14,9%	14,9%	15,6%	9,8%	8,9%	10,6%	**0110**
3,0%	0,5%	4,4%	3,3%	1,6%	2,4%	8,2%	3,0%	**0101**
7,2%	5,8%	8,8%	7,6%	5,6%	14,6%	13,4%	5,9%	**0011**
0,9%	1,8%	0,5%	0,8%	0,8%	7,3%	1,4%	0,9%	**1110**
8,9%	7,8%	16,0%	8,1%	6,3%	14,6%	33,9%	6,8%	**0111**
0,4%	0,0%	0,8%	0,3%	0,3%	0,0%	2,1%	0,4%	**1011**
0,1%	0,3%	0,0%	0,2%	0,0%	0,0%	0,7%	0,1%	**1101**
1,5%	1,3%	0,8%	1,3%	0,9%	2,4%	10,6%	0,7%	**1111**

Tabel 6-10: Frequentieverdeling van soort geweld in de zeven clusters

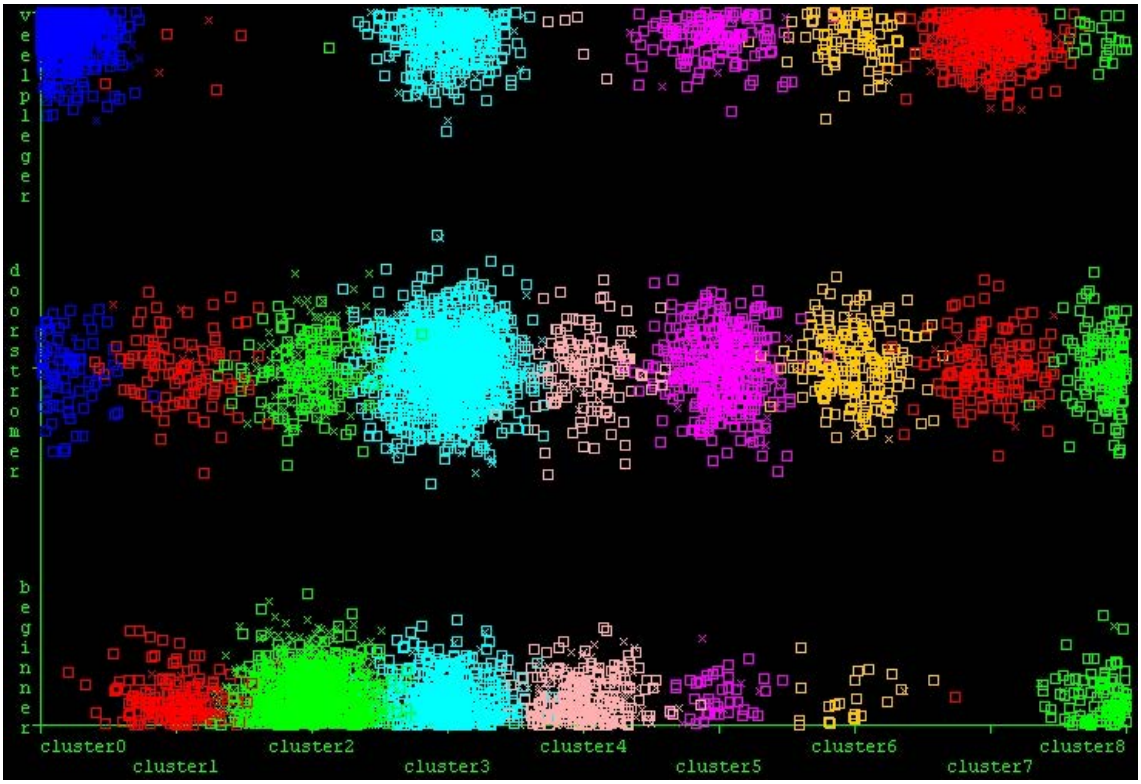
6.3.3 Clustering op basis van persoons- en delictkenmerken

Er zijn op basis van persoonskenmerken en delictkenmerken negen samenhangende groepen gevonden. Figuur 6-9 laat zien hoeveel verdachten in die groepen zitten.



Figuur 6-9: Aantal verdachten per cluster

De clusters worden vooral gekenmerkt door het aantal antecedenten dat de verdachten hebben (zie figuur 6-10) en door het soort delicten dat ze hebben gepleegd.



Figuur 6-10: Op de horizontale as de zeven clusters en op de verticale as de typologie.

Bij de beschrijving van de clusters is het belangrijk in het oog te houden dat iedere verdachte in ieder geval één of meerdere gewelddadige delicten heeft gepleegd.

Cluster 0: "De *residiverende* dieven"
Dit cluster bestaat voor het grootste deel uit veelplegers, namelijk 88%. De rest is doorstromer. Het gemiddeld aantal feiten is 69 per persoon. De verdachten plegen vooral delicten in de rubrieken openbare orde, eenvoudige diefstal, diefstal d.m.v. braak en overige gekwalificeerde diefstal. De subrubrieken winkeldiefstal en diefstal uit voertuigen komen daarbij het meeste voor. De gemiddelde leeftijd van deze personen ligt op 38 jaar. Verder wordt deze groep gekenmerkt door een erg hoog aantal druggebruikers (52%) en een redelijk hoog percentage alcoholverslaafden (8%). Zij plegen geen feiten in Pijnacker/Nootdorp en Ypenburg/Leidschenveen.

Cluster 1: "De seksueel geweldplegers"
In dit cluster bevinden zich vooral beginners en doorstromers. Het gemiddeld aantal feiten is 4. Zij plegen opvallend veel feiten als verkrachting en overige seksuele misdrijven. Het slachtoffer is relatief vaak een prostituee/schandknaap. Ook komen bedreiging en mishandeling voor. Evenals delicten in verband met wraak/minnenijd. Er bevinden zich bijna geen verdachten van (gewelddadige) vermogensdelicten in dit cluster. De verdachten plegen geen feiten in Wassenaar. Er bevinden zich geen vuurwapengevaarlijke personen, verzetplegers en vluchtgevaarlijke personen in dit cluster.

Cluster 2: "De ongewapende geweldplegers"
Het grootste deel van dit cluster is beginner en een iets minder groot deel is doorstromer. Het meest opvallende kenmerk van personen in dit cluster is dat ze geen drugs gebruiken, geen alcoholist zijn en niet vuurwapengevaarlijk zijn. Vooral de rubrieken bedreigingen, mishandeling en discriminatie komen relatief vaak voor. De rubrieken die te maken hebben met seksuele delicten en diefstal komen bijna niet voor. De methode kopstoot/slaan/schoppen heeft 51% van de verdachten in dit cluster wel eens gebruikt en 56% heeft aanroepen/aanspreken/aanraken gebruikt. Het slachtoffer is vaak een man of vrouw.

Cluster 3: "De jongere inbrekers"
De verdachten in dit cluster hebben net als in cluster 0 veel eenvoudige diefstallen en diefstal d.m.v braak op hun geweten. De subrubrieken diefstal uit woning en winkeldiefstal komen het meest voor. Daarnaast is er ook sprake van diefstal met geweld. Ook worden delicten gepleegd met behulp van

ontrukken en het slachtoffer is vaak een jongen/meisje 12-16 jaar. Ook de joyriders bevinden zich in dit cluster. De meeste delicten zijn gepleegd in Den Haag en niet in de andere steden en dorpen in de regio. De residive is hier niet op vallend; beginners, doorstromers en veelplegers komen ongeveer even vaak voor. De lage gemiddelde leeftijd is wel opvallend: 25 jaar met een standaarddeviatie van 7,2. Het gaat hier dus om vrij jonge verdachten. Het aantal verdachten dat met meer personen een delict heeft gepleegd is relatief hoog, namelijk 35%.

Cluster 4: "De fysiek gewelddadigen"
Afpersing, aanranding, mishandeling, doodslag e.d. voltooid, diefstal met geweld, dood/letsel door schuld, overige afpersing, beroving op straat zijn de rubrieken die veel vaker in dit cluster voorkomen dan in andere clusters. Dit alles gebeurt vaker dan gebruikelijk met schieten, aanroepen/aanspreken/aanraken en in verband met politiek. Er bevindt zich niemand in deze groep die een diefstal en dergelijke heeft gepleegd, het zijn allemaal redelijk zware delicten. Het zijn veelal beginners en doorstromers, bijna geen veelplegers. De gemiddelde leeftijd ligt hier ook vrij laag, op 27 jaar, met een standaarddeviatie van 11 jaar. Het aantal druggebruikers en alcoholverslaafden is erg laag. Eén op de drie verdachten heeft wel eens een delict gepleegd met meer personen.

Cluster 5: "De verkeerscriminelen"
In deze groep bevinden zich zowel beginners, doorstromers als veelplegers. Zij hebben gemeenschappelijk dat ze opgepakt zijn voor veelal verkeersdelicten die onder de volgende rubrieken vallen: rijden onder invloed, motorrijtuig besturen tijdens ontzegging, weigeren bloed of ademproef, dood/letsel door schuld en vooral overige misdrijven wegenverkeerswet. Twee van de drie verdachten in deze groep heeft dan ook één of meerdere delicten in openbaar domein gepleegd. De gemiddelde leeftijd is 36 jaar. Mishandeling komt ook regelmatig voor, evenals het gebruik van aanroepen/aanspreken/aanraken en kopstoot/slaan/schoppen.

Cluster 6: "Overtreders bijzondere wetten"
Doorstromers en veelplegers vormen het grootste deel van deze groep. Het gemiddeld aantal gepleegde feiten is 13. Er zijn echter ook een aantal beginners. Zij plegen vooral feiten in de subrubrieken van bijzondere wetten: joyriding, overige milieuwetten, overige wetten economische delicten, verontreiniging oppervlakte water en de arbeidsomstandighedenwet. Delicten in verband met gokken komen vaker voor dan gebruikelijk is. De gemiddelde leeftijd is 39 met een standaarddeviatie van 10 jaar.

Cluster 7: "De oudere verkeerscriminelen"
Er bevinden zich geen beginners in dit cluster. 76% is veelpleger en 24% is doorstromer. Het gemiddeld aantal feiten is 35, met een spreiding van 37. De verdachten in deze groep plegen veel delicten in de rubrieken motorrijtuig besturen tijdens ontzegging, verlaten plaats ongeval en overige misdrijven wegenverkeerswet. Ook dood en letsel door schuld komt vaker voor dan gebruikelijk. Er vinden echter geen delicten plaats in Ypenburg/Leidschenveen. Een verschil met cluster 5 is dat deze groep meer verslaafden heeft. 7% is alcoholist en 18% is druggebruiker. Het percentage vuurwapengevaarlijke verdachten is 8%, tegenover de gebruikelijke 1,7% is dit vrij hoog. Een ander verschil is dat de gemiddelde leeftijd relatief hoog ligt, op 43 jaar, met een spreiding van 8 jaar.

Cluster 8: "De seksueel gewelddadigen"
Dit cluster lijkt veel op cluster 4. Het bevat zowel beginners, doorstromers als veelplegers. De veelplegers zijn echter in de minderheid: 31 verdachten, of wel 9%. Hun delicten vallen binnen veel soorten rubrieken, maar vooral discriminatie, muntsmisdrijven, verkrachting en overige seksuele misdrijven zijn opvallend aanwezig. Er worden meer delicten gepleegd met schieten, een wurgstok/touw en een stukgeslagen glas/fles dan gebruikelijk is. Ze plegen delicten in verband met politiek, seksmotieven en (sport)evenementen. Hun slachtoffer is nooit een bejaarde, maar wel prostituee/schandknaap. Er bevinden zich geen alcoholisten in dit cluster en niet veel druggebruikers. De gemiddelde leeftijd is 29 jaar, met een standaarddeviatie van 10 jaar. Het verschil met cluster 4 is dat deze groep meer seksueel gewelddadig is. Cluster 4 is daarentegen meer fysiek gewelddadig. Een ander verschil is dat er in dit cluster wel delicten zijn in de rubrieken middelenlijst (1 en 2) en overige opiumwet.

7. Conclusies en aanbevelingen

"The opera isn't over till the fat lady sings."
– Dan Cook, The Washington Post, June 13, 1978

7.1 Technische conclusies

Aan de hand van de vragen, gesteld in hoofdstuk 3 kunnen de volgende conclusies getrokken worden.

Is datamining nuttig voor Politie Haaglanden?

Datamining biedt een meerwaarde aan de reeds gebruikte statistische methoden op het gebied van exploratieve analyse. Vooral om de aard van geregistreerde criminaliteit vast te stellen biedt datamining een uitkomst. Het belangrijkste voordeel van datamining ten opzichte van de huidige gebruikte methoden is dat er vooraf geen hypothesen gesteld hoeven te worden. Op die manier kunnen relaties worden ontdekt die niet eerder aan het licht zijn gekomen.

Er zijn echter ook nadelen verbonden aan het dataminen. Ten eerste kost het veel tijd. Het prepareren van de dataset is een belangrijke, maar tijdrovende stap. De resultaten hangen deels af van de preparatie. Echter moet men in het oog houden dat een zeer diepgaande statistische analyse ook veel voorbereiding van de gegevens vereist. Ten tweede kost het uitvoeren van de technieken bij een grote dataset ook veel tijd. Met een snellere computer en wellicht andere software zou dit voor een groot deel opgelost kunnen worden.

Een ander belangrijk punt is dat de dataminer niet alleen veel kennis moet hebben van de gegevens, maar ook van het dataminen. Datamining is een zeer breed vlak, bestaande uit vele algoritmen. Ieder algoritme heeft zijn beperkingen en eigenaardigheden. Een goede training in het dataminen is daarom onmisbaar. Het is gemakkelijker een dataminingspecialist de betekenis van de data te leren dan een dataspecialist datamining te leren.

Welke dataminingtechnieken zijn geschikt voor de vraagstellingen binnen Politie Haaglanden?

De drie technieken die besproken zijn in hoofdstuk 4 zijn allen zeer geschikt om de aard van de criminaliteit te onderzoeken. De ene techniek doet niet onder voor de andere. De keuze van methoden hangt af van wat men wil weten en welke en hoeveel gegevens beschikbaar zijn.

De dataminingtechniek classification is een zeer geschikte techniek om de aard van een groep te bepalen. Er kan met beslissingsbomen of regels inzichtelijk gemaakt worden wat de kenmerken van verdachten of delicten zijn. De techniek die hier het meest geschikt voor is, heet in Weka J48. Dit is een toepassing van de zogenoemde C4.5 techniek die in de meeste datamining software aanwezig is. In Clementine van SPSS wordt deze C5.0 genoemd, een nieuwere versie van C4.5. De techniek levert de beste prestatie en is ook visueel erg duidelijk.

Association rules is een sterk middel om exploratief verbanden te zoeken in gegevens. Deze techniek is minder geschikt voor een dataset met veel variabelen, zeker als er afhankelijkheden zijn. Er worden dan zo veel relaties gevonden dat het niet meer te analyseren is. Het leuke aan deze techniek is dat er regelmatig verbanden gevonden worden die niet eerder zijn ontdekt.

Clustering geeft inzicht in samenhangende groepen binnen de dataset. Met behulp van grafieken kan snel gezien worden welke groepen onderscheidend zijn en welke variabelen voor deze scheiding zorgen. Ook hier geldt dat het een exploratieve vorm is en dus niet gestuurd kan worden.

7.2 Conclusies geweld

Zijn er specifieke kenmerken van geregistreerde verdachten te onderscheiden, als gekeken wordt naar verschillende soorten geweld?
Onder de soorten geweld wordt verstaan:

- *gewelddadige seksuele delicten,*
- *fysiek geweld,*
- *verbaal geweld,*
- *gewelddadige vermogensdelicten.*

Zo ja, wat zijn de kenmerken?

De persoonskenmerken die enigszins onderscheid maken tussen de soorten geweldplegers zijn druggebruik, vuurwapengevaarlijk, verzetpleger en vluchtgevaarlijk. Er bevinden zich geen druggebruikers in de groep die uitsluitend seksueel geweld heeft gepleegd. Het aantal druggebruikers in de categorieën waarbij drie of vier soorten geweld zijn gepleegd is juist relatief hoog.

Ook een aantal delictkenmerken hebben een onderscheidend karakter maar deze resultaten zijn niet verrassend. Zo is een verdachte die een delict heeft gepleegd met de methode kopstoot/slaan/schoppen, zonder bedreiging en met ontrukken in de meeste gevallen een fysiek geweldpleger. Als een beginner nooit een delict met kopstoot/slaan/schoppen, bedreigen, ontrukken en steken heeft gepleegd dan is die verdachte nooit aangehouden voor een gewelddadig delict. Van verdachten van gewelddadige seksuele delicten, verbaal geweld en gewelddadige vermogensdelicten zijn geen onderscheidende kenmerken gevonden.

Zijn er verschillen tussen verdachten van geweldsmisdrijven en andere verdachten?
Zo ja, wat zijn de kenmerken?

Er zijn onderscheidende persoonskenmerken te vinden als gekeken wordt naar wel/geen geweld. De kenmerken die onderscheid maken tussen geweldplegers en non-geweldplegers zijn onder andere nationaliteit, harddruggebruik en leeftijd. Er bevinden zich relatief veel geweldplegers onder verdachten die de Turkse nationaliteit hebben en jonger zijn dan 30 jaar, de harddruggebruikers, verdachten met de Marokkaanse nationaliteit en vuurwapengevaarlijke verdachten. Daar tegenover staat dat er relatief weinig geweldplegers te vinden zijn in de groepen verdachten die de Bulgaarse of Poolse nationaliteit hebben, buiten de regio wonen, of ouder zijn dan 47 jaar.

Verder zijn er ook delictkenmerken die onderscheidend zijn. Evenals bij de classificatie naar *soort geweld* zijn dat de variabelen kopstoot/slaan/schoppen, bedreiging en ontrukken. Daarnaast is ook het aantal antecedenten van invloed, of degene een zwaar delict heeft gepleegd of niet, of de verdachte een seksueel motief had of niet, of er een slachtoffer was en of er een poging gepleegd is. Deze resultaten bevestigen hetgeen al bekend was.

Wat zijn de verbanden die exploratief gevonden worden in de dataset?

De techniek association rules heeft naast reeds bekende relaties ook nieuwe relaties aan het licht gebracht. Er zijn groepen gewelddadige verdachten gevonden die bijna geheel bestaan uit mannelijke verdachten zoals de verdachten van verbaal en vermogensgeweld. Ook een opvallende groep is de verdachten die wonen in bureaugebied Pijnacker/Nootdorp die de Nederlandse etniciteit hebben. In die groep is 98% een man. Verder is gevonden dat veel Surinaamse druggebruikers in Den Haag wonen. Zo zijn er nog meer groepen te vinden die voornamelijk in Den Haag wonen. Ook zijn er groepen waar de Nederlandse etniciteit erg vaak voorkomt.

Op basis van persoons- en delictkenmerken zijn slechts drie relevante relaties gevonden. Die regels geven allemaal aan dat gewelddadige verdachten die minstens één fiets hebben gestolen in bepaalde jaren, in bepaalde buurten, meestal druggebruiker zijn.

Liggen er groepen verborgen in de gegevens?
Zo ja, hoe zien die groepen eruit?

Allereerst zijn er groepen gevonden die vooral samenhangen op basis van hun woonadres. Maar ook andere kenmerken komen daarbij aan het licht zoals verslaving en leeftijd. Er zijn op basis van persoonskenmerken zeven clusters gevonden.

Bij het clusteren van de verdachten op basis van hun criminele verleden zijn vooral de typologiën beginners, doorstromers en veelplegers te onderscheiden. Daarnaast zijn de soort gepleegde feiten per cluster verschillend. Zo is er bijvoorbeeld onderscheid gemaakt tussen verdachten van zware delicten, verkeersdelicten en diefstallen. De clusteringmethode heeft in dit geval negen clusters gevonden.

7.3 Aanbevelingen

Nadat datamining toegepast is op de gegevens over verdachten van geweld kunnen de volgende aanbevelingen gedaan worden:

- Voor een aantal technieken geldt dat de computer erg lange tijd nodig heeft om resultaten te geven. Het clusteren van een dataset met 180 kolommen en 8000 rijen kost bijvoorbeeld een week tijd. De tijd hangt af van welk algoritme gebruikt wordt. Voor intensief gebruik van datamining is het nodig om een computer met voldoende capaciteit te hebben.

- In paragraaf 4.7 zijn een aantal softwarepakketten kort beschreven en is er een keuze gemaakt voor de software voor dit onderzoek. Voor de politie is Weka wellicht minder geschikt omdat er niet veel functies zijn voor het prepareren van de dataset en het programma niet optimaal gebruiksvriendelijk is. Clementine van SPSS is aan te raden maar een groot nadeel daarvan is de prijs. Naast de beschreven opties bestaat er een add-in voor Microsoft Excel, genoemd XLMiner. Het programma kan classification, association rules, clustering en meer. De prijs is relatief laag (ongeveer € 750,- voor twee jaar). Een ander voordeel is dat de werking net zo gaat als in het reeds bekende Excel. Voordat een keuze wordt gemaakt is een dieper onderzoek naar dit programma en andere mogelijkheden aan te raden.

- Het zou erg interessant zijn om te bekijken of het plegen van geweld afhankelijk is van andere componenten dan de variabelen die in dit onderzoek gebruikt zijn. Allereerst is het feit of de verdachte zelf ook een keer slachtoffer is geweest een interessant aspect. Met HKS is het helaas (nog) niet mogelijk om dit te weten te komen. Verder is het ook interessant om te kijken naar kenmerken als opleiding, opleiding ouders, inkomen, burgerlijke staat, etcetera. Helaas is het wettelijk (nog) niet mogelijk deze gegevens bij een aanhouding te registreren.

- Dit onderzoek is gericht op de verdachten van geweld. De delictkenmerken zijn echter ook meegenomen in het onderzoek om na te gaan of er relaties bestaan tussen de persoonskenmerken en delictkenmerken. Het kan echter nuttig zijn om het geweld eens van de andere kant te bekijken, vanuit de unieke delicten. Op die manier kunnen bijvoorbeeld relaties gevonden worden tussen de soort delicten en de plaatsen en tijden waarop deze plaats vinden.

- De clustering op basis van persoonskenmerken is gedaan met veel enigszins afhankelijke geografische variabelen. Om die reden zijn de gevonden clusters ook vooral onderscheidend op de woonplaats van de verdachten. Deze clustering zou nogmaals uitgevoerd kunnen worden met minder variabelen die de woonplaats aangeven.

Definities

Algoritme	Een reeks instructies om een doel te bereiken.
Antecedent	Een aanhouding waarbij een proces verbaal is opgemaakt.
Association rules	Deze dataminingtechniek vindt veel voorkomende relaties in een dataset.
Attribuut	Een variabele, een kolom in de dataset. Bijvoorbeeld: *woon_reg* (woonregio).
Betrouwbaarheid	Bij de dataminingtechniek association rules geldt dat de betrouwbaarheid van een relatie X ==> Y wordt bepaald door het aantal keer dat zowel X als Y in een relatie voorkomt, ten opzichte van het aantal keer dat X voorkomt
Classification	Deze dataminingtechniek bepaalt op basis van reeds bekende classificering in welke klasse ieder nieuw object behoort. Het voorspelt als het ware de klasse.
Classifier	Een algoritme dat classification toepast.
Clustering	Deze dataminingtechniek verdeelt objecten over groepen (clusters) op een manier dat min of meer gelijke objecten in hetzelfde cluster terecht komen. Een cluster is dus een verzameling van objecten die erg op elkaar lijken en ongelijk zijn aan objecten buiten het cluster.
Confusion matrix	Een matrix die aangeeft hoeveel verdachten in elke categorie geclassificeerd zijn. Op de diagonaal staan het aantal juist geclassificeerden.
Cross validation	Een methode om te trainen en testen, waarbij de dataset wordt opgesplitst in delen die om de beurt gebruikt worden als testset en de overige delen als trainingset.
Datamining	Het zoeken naar (onbekende) en potentieel nuttige patronen of profielen die verborgen liggen in grote gegevensverzamelingen.
Delict	Een strafbaar feit, in dit onderzoek altijd zijnde een misdrijf. Er zijn meerdere delicten per antecedent mogelijk.
Dendrogram	Een vorm van visualisatie die objecten hiërarchisch met elkaar verbindt. De verticale as geeft de afstand aan tussen de objecten die met elkaar verbonden zijn.
Doelvariabele	De variabele waarvan de waarde voorspeld wordt. In dit onderzoek zijn er twee doelvariabelen: *geweld* (ja/nee) en *soort geweld* (16 combinaties).
Error rate	De verhouding van het aantal gemaakte fouten ten opzichte van het aantal objecten in de testset.
Filter	Een methode om attributen te selecteren, met als doel de meest veelbelovende deelverzameling voor het dataminen te produceren. De variabelen die een sterke relatie met de doelvariabele hebben worden geselecteerd.
Fold	Bij cross validation wordt de dataset opgesplitst in delen. Die delen worden folds genoemd.
Frequentie	Bij de dataminingtechniek association rules geeft de frequentie aan hoe vaak de gevonden relatie voorkomt in de dataset. De frequentie kan ook worden uitgedrukt in percentages.

Object	Een rij in een dataset. Een object is het doel van het onderzoek, in dit geval de verdachte.
Pruning	Het weglaten van takken in een beslissingsboom omwille van betere prestatie en een duidelijker beeld. Letterlijk: snoeien.
Residiveren	In herhaling treden.
Seed	Beginwaarde voor het genereren van een willekeurige (random) reeks getallen.
Stratified sample	Een steekproef van de dataset waarin de verdeling van de doelvariabele gelijk is aan de verdeling van de originele dataset.
Testset	Het deel van de dataset waarvan de doelvariabele onbekend is en voorspeld wordt door de classification technieken.
Trainingset	Het deel van de dataset waarvan de doelvariabele al bekend is en waarvan de classificationmodellen kunnen leren. De modellen worden "getraind".
Verdachte	In dit onderzoek is de verdachte een persoon die aangehouden is op verdenking van een misdrijf waarvoor proces verbaal is opgemaakt.
Wrapper	Een methode om attributen te selecteren, met als doel de meest veelbelovende deelverzameling voor een specifiek algoritme te produceren.

Literatuurlijst

[1] Programmabureau Abrio (2004), "*Meerwaarde data mining*"

[2] Pieter Adriaans, Dolf Zantinge, "*Data Mining*", 1996, Addison-Wesley

[3] Rakesh Agrawal, Tomasz Imielinski, Arun Swami, *"Mining Association Rules between Sets of Items in Large Databases*" 1993, in Proceedings of the 1993 ACM SIGMOD Conference Washington DC

[4] Rakes Agrawal, Ramakrishnan Srikant, "*Fast Algorithms for Mining Association Rules*", 1994, in Proc. 1994 Ing. Conf. Very Large Data Bases

[5] Catrien Bijleveld, Ronald Meijer, "*Exploratieve analyse van delictcombinaties en verdachtenkenmerken*", in Criminaliteitsanalyse in Nederland, Elsevier

[6] Zhengxin Chen, "*Data Mining and Uncertain Reasoning*", 2001, Wiley-Interscience

[7] DataExpert, *"Criminaliteitsanalyse met datadetective*", 2004

[8] A. Dempster, N. Laird, D. Rubin, "*Maximum likelihood from incomplete data via the EM algorithm*"., 1977, in Journal of the Royal Statistical Society, Series B, 39(1), p. 1–38.

[9] Margaret H. Dunham, "*Data Mining, Introductory and Advanced Topics*", 2003, Pearson Education

[10] Y. Freund, L. Mason, "*The alternating decision tree learning algorithm*", 1999, in Machine Learning: Proceedings of the Sixteenth International Conference, p. 124-133

[11] Y. Freund, R. E. Schapire, "*Experiments with a new boosting algorithm*", 1996, in Proc International Conference on Machine Learning, p. 148-156

[12] Mark A. Hall & Geoffrey Holmes, "*Attribute Selection Techniques for Data Mining*"

[13] Mark A. Hall, Lloyd A. Smith, *"Feature Selection for Machine Learning: Comparing a Correlation-based Filter Approach to the Wrapper*", 1998, in American Association for Artificial Intelligence of Benchmarking

[14] David Hand, Heikki Mannila, Padhraic Smyth "*Principles of Data Mining*", 2001, in Massachusetts Institute of Technology

[15] George H. John, Pat Langley, "*Estimating Continuous Distributions in Bayesian Classifiers*", 1995, in Proceedings of the Eleventh Conference on Uncertainty in Artificial Intelligence, p. 338-345

[16] Michael Kearns, Yishay Mansour, "*On the boosting ability of top-down decision treelearning algorithms*", 1996, in Proceedings of the Twenty-Eighth Annual ACM Symposium on the Theory of Computing

[17] Ron Kohavi, George H. John, "*Wrappers for Feature Subset Selection*", 1996, in AIJ

[18] P. Kruize, "*Geweld in het publieke domein heeft vele gezichten*", 2005, in Het Tijdschrift voor de Politie, jaargang 67

[19] Floor Luykx, Wim Bernasco, "*NSCR Documentatie HKS*", 2004, intern rapport

[20] Anne Mooij, Hans Moerland, "*Visies op de toekomst van de criminaliteitsanalyse*", in Criminaliteitsanalyse in Nederland

[21] Clark F. Olson, "*Parallel algorithm for hierarchical clustering*", 1995, in Parallel Computing 21, p. 1313-1325

[22] Ross Quinlan, "*C4.5: Programs for Machine Learning*", 1993, Morgan Kaufmann Publishers

[23] Ben Rovers, "*Op zoek naar de bekende dader in het HKS*", 2003, intern rapport

[24] John W. Sammon jr., "*A Nonlinear Mapping for Data Structure Analysis*", 1969, in TC, 18(5), p. 401-409

[25] SPSS Inc., "*Clementine 9.0 Algorithms Guide",* 2004

[26] Wouter Stol, "*Informatietechnologie in criminaliteitsanalyse*", in Criminaliteitsanalyse in Nederland, Elsevier

[27] H.J. Thomassen, "*Unix, het standaard operating system*", 1994, Academic Service

[28] Bhavani Thuraisingham, "*Data Mining, Technologies, Technieques, Tools, and Trends*", 1999, CRC Press

[29] Peter Versteegh, "*Geweld in het publieke domein, een probleembeschrijving*", 2001, intern rapport

[30] Christopher Westphal, Teresa Blaxton, "*Data Mining Solutions*", 1998, John Wiley & Sons

[31] Ian H. Witten, Eibe Frank, "*Data Mining: Practical machine learning tools with Java implementations*", 2000, Morgan Kaufmann

Websites:

www.sentient.nl , www.dataexpert.nl
Criminaliteitsanalyse met DataDetective

www.spss.com
SPSS Clementine en Classification Trees

http://www.cs.waikato.ac.nz/~ml/weka/
Weka

Bijlagen

Bijlage A: Wetsartikelen per rubriek

Gewelddadige seksuele delicten

Rubriek 115 – Verkrachting

Art. 242 – Verkrachting
Hij die door geweld of een andere feitelijkheid of bedreiging met geweld of een andere feitelijkheid iemand dwingt tot het ondergaan van handelingen die bestaan uit of mede bestaan uit het seksueel binnendringen van het lichaam, wordt als schuldig aan verkrachting gestraft met gevangenisstraf van ten hoogste twaalf jaren of geldboete van de vijfde categorie.

Rubriek 117 – Aanranding der eerbaarheid

Art. 246 – Feitelijke aanranding der eerbaarheid
Hij die door geweld of een andere feitelijkheid of bedreiging met geweld of een andere feitelijkheid iemand dwingt tot het plegen of dulden van ontuchtige handelingen, wordt, als schuldig aan feitelijke aanranding van de eerbaarheid, gestraft met gevangenisstraf van ten hoogste acht jaren of geldboete van de vijfde categorie.

Verbaal geweld

Rubriek 121 – Bedreiging

Art. 285 – Bedreiging met misdrijf
1. Bedreiging met openlijk geweld met verenigde krachten tegen personen of goederen, met geweld tegen een internationaal beschermd persoon of diens beschermde goederen, met enig misdrijf waardoor de algemene veiligheid van personen of goederen in gevaar wordt gebracht, met verkrachting, met feitelijke aanranding van de eerbaarheid, met enig misdrijf tegen het leven gericht, met gijzeling, met zware mishandeling of met brandstichting, wordt gestraft met gevangenisstraf van ten hoogste twee jaren of geldboete van de vierde categorie.
2. Indien deze bedreiging schriftelijk en onder een bepaalde voorwaarde geschiedt, wordt ze gestraft met gevangenisstraf van ten hoogste vier jaren of geldboete van de vierde categorie.
3. Bedreiging met een terroristisch misdrijf wordt gestraft met gevangenisstraf van ten hoogste zes jaren of geldboete van de vijfde categorie.

285a – Beïnvloeding verklaring
1. Hij die opzettelijk mondeling, door gebaren, bij geschrift of afbeelding zich jegens een persoon uit, kennelijk om diens vrijheid om naar waarheid of geweten ten overstaan van een rechter of ambtenaar een verklaring af te leggen te beïnvloeden, terwijl hij weet of ernstige reden heeft te vermoeden dat die verklaring zal worden afgelegd, wordt gestraft met gevangenisstraf van ten hoogste vier jaren of geldboete van de vierde categorie.
2. Met rechter of ambtenaar wordt gelijkgesteld: een rechter bij onderscheidenlijk een persoon in de openbare dienst van een internationaal gerecht dat zijn rechtsmacht ontleent aan een verdrag waarbij het Koninkrijk partij is.

285b – Belaging
1. Hij, die wederrechtelijk stelselmatig opzettelijk inbreuk maakt op eens anders persoonlijke levenssfeer met het oogmerk die ander te dwingen iets te doen, niet te doen of te dulden dan wel vrees aan te jagen wordt, als schuldig aan belaging, gestraft met een gevangenisstraf van ten hoogste drie jaren of een geldboete van de vierde categorie.
2. Vervolging vindt niet plaats dan op klacht van hem tegen wie het misdrijf is begaan.

Fysiek geweld

Rubriek 123 – Poging doodslag e.d.

Zodra strafrecht artikel 45 (poging) in combinatie met één van de in rubriek 125 genoemde artikelen ten laste is gelegd, dan valt het misdrijf in deze rubriek.

Rubriek 125 – Doodslag e.d. voltooid

Art 287 – Doodslag
Hij die opzettelijk een ander van het leven berooft, wordt, als schuldig aan doodslag, gestraft met gevangenisstraf van ten hoogste vijftien jaren of geldboete van de vijfde categorie.

Art 288 – Doodslag met strafverzwaring
Doodslag gevolgd, vergezeld of voorafgegaan van een strafbaar feit en gepleegd met het oogmerk om de uitvoering van dat feit voor te bereiden of gemakkelijk te maken, of om, bij betrapping op heterdaad, aan zichzelf of andere deelnemers aan dat feit hetzij straffeloosheid hetzij het bezit van het wederrechtelijk verkregene te verzekeren, wordt gestraft met levenslange gevangenisstraf of tijdelijke van ten hoogste twintig jaren of geldboete van de vijfde categorie.

288a
Doodslag, gepleegd met een terroristisch oogmerk, wordt gestraft met levenslange gevangenisstraf of tijdelijke van ten hoogste twintig jaren of geldboete van de vijfde categorie.

Art 289 – Moord
Hij die opzettelijk en met voorbedachten rade een ander van het leven berooft, wordt, als schuldig aan moord, gestraft met levenslange gevangenisstraf of tijdelijke van ten hoogste twintig jaren of geldboete van de vijfde categorie.

289a
1. De samenspanning tot het in artikel 289 omschreven misdrijf, te begaan met een terroristisch oogmerk, alsmede het in artikel 288a omschreven misdrijf, wordt gestraft met gevangenisstraf van ten hoogste tien jaren of geldboete van de vijfde categorie.
2. Artikel 96, tweede lid, is van overeenkomstige toepassing.

Art 290 – Kinderdoodslag
De moeder die, onder de werking van vrees voor de ontdekking van haar bevalling, haar kind bij of kort na de geboorte opzettelijk van het leven berooft, wordt, als schuldig aan kinderdoodslag, gestraft met gevangenisstraf van ten hoogste zes jaren of geldboete van de vierde categorie.

Art 291 – Kindermoord
De moeder die, ter uitvoering van een onder de werking van vrees voor de ontdekking van haar aanstaande bevalling genomen besluit, haar kind bij of kort na de geboorte opzettelijk van het leven berooft, wordt, als schuldig aan kindermoord, gestraft met gevangenisstraf van ten hoogste negen jaren of geldboete van de vijfde categorie.

Art 292 – Deelneming
De in de artikelen 290 en 291 omschreven misdrijven worden ten aanzien van anderen die er aan deelnemen als doodslag of als moord aangemerkt.

Rubriek 127 – Overige misdrijven tegen het leven

Art 293 – Levensberoving op verzoek
1. Hij die opzettelijk het leven van een ander op diens uitdrukkelijk en ernstig verlangen beëindigt, wordt gestraft met een gevangenisstraf van ten hoogste twaalf jaren of geldboete van de vijfde categorie.
2. Het in het eerste lid bedoelde feit is niet strafbaar, indien het is begaan door een arts die daarbij voldoet aan de zorgvuldigheidseisen, bedoeld in artikel 2 van de Wet toetsing levensbeëindiging op

verzoek en hulp bij zelfdoding en hiervan mededeling doet aan de gemeentelijke lijkschouwer overeenkomstig artikel 7, tweede lid, van de Wet op de lijkbezorging.

Art 294 – Helpen bij zelfmoord
1. Hij die opzettelijk een ander tot zelfdoding aanzet, wordt, indien de zelfdoding volgt, gestraft met een gevangenisstraf van ten hoogste drie jaren of geldboete van de vierde categorie.
2. Hij die opzettelijk een ander bij zelfdoding behulpzaam is of hem de middelen daartoe verschaft, wordt, indien de zelfdoding volgt, gestraft met een gevangenisstraf van ten hoogste drie jaren of geldboete van de vierde categorie. Artikel 293, tweede lid, is van overeenkomstige toepassing.

Art 296 – Afbreking van zwangerschap
1. Hij die een vrouw een behandeling geeft, terwijl hij weet of redelijkerwijs moet vermoeden dat daardoor zwangerschap kan worden afgebroken, wordt gestraft met gevangenisstraf van ten hoogste vier jaar en zes maanden of geldboete van de vierde categorie.
2. Indien het feit de dood van de vrouw ten gevolge heeft, wordt gevangenisstraf van ten hoogste zes jaren opgelegd of geldboete van de vierde categorie.
3. Indien het feit is begaan zonder toestemming van de vrouw, wordt gevangenisstraf van ten hoogste twaalf jaren opgelegd of geldboete van de vijfde categorie.
4. Indien het feit is begaan zonder toestemming van de vrouw en tevens haar dood ten gevolge heeft, wordt gevangenisstraf van ten hoogste vijftien jaren opgelegd of geldboete van de vijfde categorie.
5. Het in het eerste lid bedoelde feit is niet strafbaar, indien de behandeling is verricht door een arts in een ziekenhuis of kliniek waarin zodanige behandeling volgens de Wet afbreking zwangerschap mag worden verricht.

R129 – Mishandeling

Art 300 – Mishandeling
1. Mishandeling wordt gestraft met gevangenisstraf van ten hoogste twee jaren of geldboete van de vierde categorie.
2. Indien het feit zwaar lichamelijk letsel ten gevolge heeft, wordt de schuldige gestraft met gevangenisstraf van ten hoogste vier jaren of geldboete van de vierde categorie.
3. Indien het feit de dood ten gevolge heeft, wordt hij gestraft met gevangenisstraf van ten hoogste zes jaren of geldboete van de vierde categorie.
4. Met mishandeling wordt gelijkgesteld opzettelijke benadeling van de gezondheid.
5. Poging tot dit misdrijf is niet strafbaar.

Art 301 – Mishandeling met voorbedachten rade
1. Mishandeling gepleegd met voorbedachten rade wordt gestraft met gevangenisstraf van ten hoogste drie jaren of geldboete van de vierde categorie.
2. Indien het feit zwaar lichamelijk letsel ten gevolge heeft, wordt de schuldige gestraft met gevangenisstraf van ten hoogste zes jaren of geldboete van de vierde categorie.
3. Indien het feit de dood ten gevolge heeft, wordt hij gestraft met gevangenisstraf van ten hoogste negen jaren of een geldboete van de vijfde categorie.

Art 302 – Zware mishandeling
1. Hij die aan een ander opzettelijk zwaar lichamelijk letsel toebrengt, wordt, als schuldig aan zware mishandeling, gestraft met gevangenisstraf van ten hoogste acht jaren of geldboete van de vijfde categorie.
2. Indien het feit de dood ten gevolge heeft, wordt de schuldige gestraft met gevangenisstraf van ten hoogste tien jaren of geldboete van de vijfde categorie.

Art 303 – Zware mishandeling met voorbedachte rade
1. Zware mishandeling gepleegd met voorbedachten rade wordt gestraft met gevangenisstraf van ten hoogste twaalf jaren of geldboete van de vijfde categorie.
2. Indien het feit de dood ten gevolg heeft, wordt de schuldige gestraft met gevangenisstraf van ten hoogste vijftien jaren of geldboete van de vijfde categorie.

Art 304 – Verzwarende omstandigheden
De in de artikelen 300-303 bepaalde gevangenisstraffen kunnen met een derde worden verhoogd:
1°. ten aanzien van de schuldige die het misdrijf begaat tegen zijn moeder, zijn vader tot wie hij in familierechtelijke betrekking staat, zijn echtgenoot of zijn kind;
2°. indien het misdrijf wordt gepleegd tegen een ambtenaar gedurende of ter zake van de rechtmatige uitoefening van zijn bediening;
3°. indien het misdrijf wordt gepleegd door toediening van voor het leven of de gezondheid schadelijke stoffen.

304a
Indien een misdrijf, strafbaar gesteld in artikel 302 of 303, is begaan met een terroristisch oogmerk, wordt de in dat artikel bepaalde tijdelijke gevangenisstraf met de helft verhoogd en wordt, indien op het misdrijf een tijdelijke gevangenisstraf van ten hoogste vijftien jaren is gesteld, levenslange gevangenisstraf of tijdelijke van ten hoogste twintig jaren opgelegd.

304b
1. De samenspanning tot het in artikel 303 omschreven misdrijf, te begaan met een terroristisch oogmerk, wordt gestraft met gevangenisstraf van ten hoogste tien jaren of geldboete van de vijfde categorie.
2. Artikel 96, tweede lid, is van overeenkomstige toepassing.

Art 306 – Deelneming aan aanval of vechterij
Zij die opzettelijk deelnemen aan een aanval of vechterij waarin onderscheiden personen zijn gewikkeld, worden, behoudens ieders verantwoordelijkheid voor de bijzondere door hem bedreven feiten, gestraft:
1°. met gevangenisstraf van ten hoogste twee jaren of geldboete van de vierde categorie, indien de aanval of vechterij alleen zwaar lichamelijk letsel ten gevolge heeft;
2°. met gevangenisstraf van ten hoogste drie jaren of geldboete van de vierde categorie, indien de aanval of vechterij iemands dood ten gevolge heeft.

Vermogensdelicten met geweld

Rubriek 139 – Diefstal met geweld

Art. 312 – Diefstal met geweld of bedreiging
1. Met gevangenisstraf van ten hoogste negen jaren of geldboete van de vijfde categorie wordt gestraft diefstal, voorafgegaan, vergezeld of gevolgd van geweld of bedreiging met geweld tegen personen, gepleegd met het oogmerk om die diefstal voor te bereiden of gemakkelijk te maken, of om, bij betrapping op heter daad, aan zichzelf of andere deelnemers aan het misdrijf hetzij de vlucht mogelijk te maken, hetzij het bezit van het gestolene te verzekeren.
2. Gevangenisstraf van ten hoogste twaalf jaren of geldboete van de vijfde categorie wordt opgelegd:
1°. indien het feit wordt gepleegd hetzij gedurende de voor de nachtrust bestemde tijd in een woning of op een besloten erf waarop een woning staat; hetzij op de openbare weg; hetzij in een spoortrein die in beweging is;
2°. indien het feit wordt gepleegd door twee of meer verenigde personen;
3°. indien de schuldige zich de toegang tot de plaats van het misdrijf heeft verschaft door middel van braak of inklimming, van valse sleutels, van een valse order of een vals kostuum;
4°. indien het feit zwaar lichamelijk letsel ten gevolge heeft.
5°. indien het feit wordt gepleegd met het oogmerk om een terroristisch misdrijf voor te bereiden of gemakkelijk te maken.
3. Gevangenisstraf van ten hoogste vijftien jaren of geldboete van de vijfde categorie wordt opgelegd, indien het feit de dood ten gevolge heeft.

Rubriek 141 – Afpersing

Artikel 317 – Afpersing
1. Hij die, met het oogmerk om zich of een ander wederrechtelijk te bevoordelen, door geweld of bedreiging met geweld iemand dwingt hetzij tot de afgifte van enig goed dat geheel of ten dele aan deze of aan een derde toebehoort, hetzij tot het aangaan van een schuld of het teniet doen van een

inschuld, hetzij tot het ter beschikking stellen van gegevens wordt, als schuldig aan afpersing, gestraft met gevangenisstraf van ten hoogste negen jaren of geldboete van de vijfde categorie.
2. Met dezelfde straf wordt gestraft hij die de dwang, bedoeld in het eerste lid, uitoefent door de bedreiging dat gegevens die door middel van een geautomatiseerd werk zijn opgeslagen, onbruikbaar of ontoegankelijk zullen worden gemaakt of zullen worden gewist.
3. De bepalingen van het tweede en derde lid van artikel 312 zijn op dit misdrijf van toepassing.

Bijlage B: Betekenis variabelen

Deze bijlage geeft alle betekenissen van variabelen die gebruikt zijn bij dit onderzoek. De grijs gedrukte variabelen zijn gebruikt om een nieuwe variabele aan te maken. De variabelen met een 'X' in de laatste kolom zijn bij classification omgezet naar relatieve variabelen zoals beschreven in paragraaf 6.1.1.

Ltb_result

pers_id	uniek persoonsnummer	nominaal	
geb_land	geboorteland	nominaal	
ovl	overleden	nominaal	
nat_code	nationaliteit	nominaal	
geslacht	geslacht	nominaal	
wbs_code	wijk-buurt-subbuurt (indeling politie)	nominaal	
plts_code	woonplaats	nominaal	
b_wbs_nieuw	bureaucode gemaakt van wbs_code en plts_code	nominaal	
land_cd	woonland	nominaal	
woon_reg	woonregio (indeling politie)	nominaal	
lft_1del	leeftijd ten tijde van eerste delict	ratio	
lft_ldel	leeftijd ten tijde van laatste delict	ratio	
som_feit	aantal feiten t/m 2004	ratio	
som_ant	aantal antecedenten t/m 2004	ratio	
g1	alcoholist	nominaal	
g2	harddruggebruiker	nominaal	
g3	medische indicatie	nominaal	
g4	vuurwapengevaarlijk	nominaal	
g6	verzetpleger	nominaal	
g7	vluchtgevaarlijk	nominaal	
g8	zelfmoordneiging	nominaal	
voor_80	aantal antecedenten vóór 1980	ratio	
jr81	aantal antecedenten in 1981	ratio	X
…	…	…	X
jr04	aantal antecedenten in 2004	ratio	X
Hoofdrubrieken:			
R101	openbare orde	ratio	X
R103	discriminatie	ratio	X
R105	gemeengevaarlijke misdrijven	ratio	X
R107	openbaar gezag	ratio	X
R109	muntmisdrijven	ratio	X
R111	overige valsheidmisdrijven	ratio	X
R113	schennis der eerbaarheid	ratio	X
R115	verkrachting	ratio	X
R117	aanranding der eerbaarheid	ratio	X
R119	overige seksuele misdrijven	ratio	X
R121	bedreiging	ratio	X
R123	poging doodslag e.d.	ratio	X
R125	doodslag e.d. voltooid	ratio	X
R127	overige misdrijven tegen leven	ratio	X
R129	mishandeling	ratio	X

R131	dood en letsel door schuld	ratio	X
R133	eenvoudige diefstal	ratio	X
R135	diefstal d.m.v. braak	ratio	X
R137	overige gekwalificeerde diefstal	ratio	X
R139	diefstal met geweld	ratio	X
R141	afpersing	ratio	X
R143	verduistering	ratio	X
R145	bedrog	ratio	X
R147	vernieling auto's	ratio	X
R149	vernieling openbaar vervoer	ratio	X
R151	vernieling openbaar gebouw	ratio	X
R153	overige vernieling	ratio	X
R155	heling/schuldheling	ratio	X
R157	overige misdrijven	ratio	X
Subrubrieken (uitsplitsing van diefstal):			
R201	fietsen	ratio	X
R202	bromfietsen	ratio	X
R203	motoren/scooters	ratio	X
R205	personenauto's+ander motorvoertuig	ratio	X
R206	uit voertuigen	ratio	X
R207	vaartuigen	ratio	X
R208	uit/vanaf vaartuig	ratio	X
R209	gewapende overval	ratio	X
R210	beroving op straat	ratio	X
R211	zakkenrollerij	ratio	X
R212	winkeldiefstal	ratio	X
R213	diefstal uit woning	ratio	X
R214	diefstal uit scholen	ratio	X
R215	diefstal uit bedrijf	ratio	X
R216	diefstal uit/van automaat	ratio	X
R217	diefstal uit sportcomplex	ratio	X
R218	diefstal op kampeerterrein	ratio	X
R290	overige diefstallen sr 311	ratio	X
R291	overige diefstallen sr 312	ratio	X
R292	overige afpersingen sr 317	ratio	X
R299	overige diefstallen	ratio	X
Subrubrieken (bijzondere wetten):			
R317	rijden onder invloed	ratio	X
R321	verlaten plaats ongeval	ratio	X
R323	motorrijtuig besturen tijdens ontzegging	ratio	X
R325	weigeren bloed of ademproef	ratio	X
R327	dood/letsel door schuld	ratio	X
R329	joyriding	ratio	X
R331	overige misdrijven wegenverkeerswet	ratio	X
R335	verontreiniging oppervlakte water	ratio	X
R337	arbeidsomstandighedenwet	ratio	X
R339	wet milieubeheer	ratio	X
R343	afvalstoffenwet	ratio	X
R345	meststoffenwet	ratio	X
R347	overige milieuwetten	ratio	X
R349	overige wetten economische delicten	ratio	X

R351	middelenlijst 1(harddrugs)	ratio	X
R353	middelenlijst 2(softdrugs)	ratio	X
R354	overige opiumwet	ratio	X
R355	wet wapens en munitie	ratio	X
R357	misdrijven andere wetten	ratio	X

Ltb_cat

cbs_etn	afkomst naar cbs_landcode	nominaal	
gem_cat	aantal inwoners van gemeente in klassen	ordinaal	
lft_cat	leeftijd in onderzoeksjaar	ratio	
gw_sex	gewelddadige seksuele delicten	ratio	X
ov_sex	overige seksuele delicten	ratio	X
gw_ovg	gewelddadige delicten overig	ratio	X
vm_gw	vermogensdelicten met geweld	ratio	X
vm_ovg	vermogensdelicten overig	ratio	X
vern_oo	vernieling en openbare orde	ratio	X
verkeer	verkeer misdrijven	ratio	X
opium	drugsdelicten	ratio	X
overig	overige delicten	ratio	X
gw_sexpj	in onderzoeksjaar	ratio	X
ov_sexpj	in onderzoeksjaar	ratio	X
gw_ovgpj	in onderzoeksjaar	ratio	X
vm_gwpj	in onderzoeksjaar	ratio	X
vm_ovgpj	in onderzoeksjaar	ratio	X
vern_oopj	in onderzoeksjaar	ratio	X
verkeerpj	in onderzoeksjaar	ratio	X
opiumpj	in onderzoeksjaar	ratio	X
overigpj	in onderzoeksjaar	ratio	X

Ltb_bd

Alle variabelen in deze tabel zijn getransformeerd zoals beschreven in paragraaf 5.4.2

Plts_code	plaatscode van pleging	nominaal	
Wbs_code	wijk-buurt-subbuurt van pleging	nominaal	
b_wbs_nieuw	bureaucode gemaakt van wbs_code en plts_code	nominaal	X
Poging	wel/geen poging	nominaal	X
Dat_begin	datum begin delict - alleen de maand is gebruikt	interval	X
Gemtijd	gemiddelde tijd - omgezet naar ochtend/middag/avond/nacht	ratio	X
Dagvan	dag van de week	interval	X
B02	Dader - met meer personen	nominaal	X
C02	Hoe gepleegd - aanroepen/aanspreken/aanraken	nominaal	X
C06	Hoe gepleegd - bedreigen	nominaal	X
C14	Hoe gepleegd - kopstoot/slaan/schoppen	nominaal	X
C15	Hoe gepleegd - ontrukken	nominaal	X
C20	Hoe gepleegd - schieten	nominaal	X
C22	Hoe gepleegd - steken	nominaal	X
D03	Met behulp van - chemisch/medisch middel/vergif	nominaal	X

D08	Met behulp van - (schijn)vuurwapen	nominaal	X
D09	Met behulp van - slag-/stootwapen	nominaal	X
D11	Met behulp van - steekwapen	nominaal	X
D12	Met behulp van - steen/katapult	nominaal	X
D13	Met behulp van - stukgeslagen glas/fles	nominaal	X
D17	Met behulp van - vervoersmiddel	nominaal	X
D19	Met behulp van - wurgstok/touw	nominaal	X
E01 t/m 42	Plaats delict - omgezet naar prive/semi/openbaar	nominaal	X
F02	Slachtoffer - bejaarde(n)	nominaal	X
F07	Slachtoffer - jongen/meisje 12-16 jaar	nominaal	X
F08	Slachtoffer - jongen/meisje tot 12 jaar	nominaal	X
F09	Slachtoffer - man	nominaal	X
F12	Slachtoffer - prostituee/schandknaap	nominaal	X
F14	Slachtoffer - vrouw	nominaal	X
H01	In verband met - drankgebruik	nominaal	X
H02	In verband met - drugsgebruik	nominaal	X
H03	In verband met - gokken	nominaal	X
H04	In verband met - politiek	nominaal	X
H05	In verband met - prostitutie/souteneurschap	nominaal	X
H06	In verband met - seksmotieven	nominaal	X
H07	In verband met - (sport)evenement	nominaal	X
H08	In verband met - wraak/minnenijd	nominaal	X
H09	In verband met - verkeerssituatie	nominaal	X

Ltb_ant_del

zwaar	zwaar of licht delict	nominaal	X

Pe_ber

beroepco	bereoepscode	nominaal	

Aangemaakt

typol3	typologie in 3 klassen: veelpleger/doorstromer/beginner	nominaal	
bedreiging 2004	wel/geen bedreiging in 2004 gepleegd	nominaal	
mishandeling 2004	wel/geen mishandeling in 2004 gepleegd	nominaal	
stalking	wel/geen stalking in 2004 gepleegd	nominaal	
huiselijk geweld	wel/geen huiselijk geweld in 2004 gepleegd	nominaal	
geweld	heeft verdachte ooit geweld gepleegd	nominaal	
soort geweld	soort geweld zoals beschreven in paragraaf 5.5.6	nominaal	

Bijlage C: Statistieken van numerieke variabelen

Hieronder staat een lijst van gebruikte variabelen met hun statistieken.

Variabele	Minimum	Maximum	Gemiddelde	Std. Dev.	Scheefheid	Kurtosis
lft_1del	12	92	27,82	13,586	1,099	0,698
lft_ldel	12	92	32,40	13,264	0,711	0,088
som_feit	1	455	8,30	22,134	7,165	72,502
som_ant	1	304	5,52	14,110	7,608	83,976
lft_peil	9	92	32,84	13,265	0,711	0,091
zwaar_r	0,00	100,00	60,1420	39,20951	-0,452	-1,320
somPoging_r	0,00	100,00	6,4825	18,24871	3,808	15,305
voor_80_r	0,00	93,33	0,8785	5,79495	8,296	78,580
jr_80_r	0,00	50,00	0,1788	1,65048	13,135	218,705
jr_81_r	0,00	55,56	0,2327	1,94543	12,466	201,291
jr_82_r	0,00	60,00	0,2640	2,10866	12,406	206,322
jr_83_r	0,00	60,00	0,2916	2,20567	11,438	173,641
jr_84_r	0,00	66,67	0,3368	2,24941	9,841	135,152
jr_85_r	0,00	57,14	0,3428	2,31066	10,136	138,725
jr_86_r	0,00	50,00	0,4296	2,59733	8,078	79,936
jr_87_r	0,00	50,00	0,4672	2,79167	8,357	88,605
jr_88_r	0,00	50,00	0,5402	3,14377	8,053	79,627
jr_89_r	0,00	60,00	0,6426	3,58534	7,871	76,068
jr_90_r	0,00	66,67	0,7579	4,20998	7,821	72,479
jr_91_r	0,00	66,67	0,8652	4,50581	7,285	63,422
jr_92_r	0,00	75,00	0,9640	4,80875	6,911	56,818
jr_93_r	0,00	75,00	1,0781	5,21391	6,697	52,843
jr_94_r	0,00	72,73	1,1058	5,12732	6,424	49,715
jr_95_r	0,00	75,00	1,1974	5,34390	6,113	44,943
jr_96_r	0,00	80,00	1,3320	5,73036	5,843	40,165
jr_97_r	0,00	75,00	1,2868	5,50131	5,775	39,830
jr_98_r	0,00	66,67	1,6452	6,59780	5,223	30,986
jr_99_r	0,00	83,33	1,9538	7,26844	4,870	27,345
jr_00_r	0,00	80,00	2,5494	8,63614	4,147	18,430
jr_01_r	0,00	80,00	3,1181	9,73919	3,748	14,782
jr_02_r	0,00	80,00	4,0828	11,39807	3,196	10,130
jr_03_r	0,00	85,71	5,2815	13,39936	2,783	7,106
jr_04_r	0,33	100,00	68,1771	36,78342	-0,483	-1,463
r101_r	0,00	100,00	4,8642	16,97570	4,455	20,405
r103_r	0,00	100,00	0,1032	2,19700	32,390	1.273,769
r105_r	0,00	100,00	0,7607	6,67005	12,093	162,417
r107_r	0,00	100,00	2,5596	12,13736	6,445	45,044
r109_r	0,00	100,00	0,2456	4,25607	21,163	472,546
r111_r	0,00	100,00	2,5841	12,73266	6,282	41,821
r113_r	0,00	100,00	0,2038	4,06465	22,764	536,528
r115_r	0,00	100,00	0,2323	3,30093	22,183	589,243
r117_r	0,00	100,00	0,3720	5,04538	17,223	318,535
r119_r	0,00	100,00	0,5145	6,17351	14,386	218,178
r121_r	0,00	100,00	4,2501	14,66618	4,785	25,230
r123_r	0,00	100,00	1,4553	8,67913	8,805	87,918

r125_r	0,00	100,00	0,2015	3,54081	23,545	611,105
r127_r	0,00	33,33	0,0018	0,24337	136,963	18.759,000
r129_r	0,00	100,00	11,0397	25,37703	2,620	5,949
r131_r	0,00	100,00	0,0327	1,59829	56,614	3.386,734
r133_r	0,00	100,00	9,3076	23,26616	2,900	7,690
r135_r	0,00	100,00	7,5926	19,34131	3,097	9,771
r137_r	0,00	100,00	6,1653	18,69958	3,877	15,220
r139_r	0,00	100,00	2,3239	10,42423	6,724	52,162
r141_r	0,00	100,00	0,3154	3,86313	19,683	449,761
r143_r	0,00	100,00	1,9689	11,76339	7,309	55,283
r145_r	0,00	100,00	1,1002	8,41630	9,879	104,859
r153_r	0,00	100,00	4,0600	14,53923	4,997	27,265
r155_r	0,00	100,00	2,7770	12,42653	6,271	42,700
r157_r	0,00	100,00	2,8648	12,92914	6,104	40,027
r317_r	0,00	100,00	17,5369	34,10753	1,760	1,442
r321_r	0,00	100,00	3,4895	15,95795	5,341	28,357
r323_r	0,00	100,00	0,6774	5,77740	12,983	197,373
r325_r	0,00	100,00	0,3240	3,71907	19,365	454,333
r327_r	0,00	100,00	0,9462	8,80813	10,494	112,453
r329_r	0,00	100,00	0,1211	2,52268	33,360	1.239,313
r331_r	0,00	100,00	0,0939	1,93153	37,360	1.697,694
r335_r	0,00	100,00	0,0366	1,67001	52,864	2.970,893
r339_r	0,00	50,00	0,0083	0,55214	82,507	7.258,773
r343_r	0,00	100,00	0,0603	1,90314	44,252	2.161,526
r347_r	0,00	100,00	1,6718	12,05941	7,625	57,839
r349_r	0,00	100,00	1,1537	9,76728	9,304	88,084
r351_r	0,00	100,00	2,0474	11,14645	7,233	56,272
r353_r	0,00	100,00	1,4154	9,32187	8,468	78,329
r354_r	0,00	100,00	0,1452	2,40963	31,075	1.168,780
r355_r	0,00	100,00	1,8567	9,95221	7,766	67,371
r357_r	0,00	100,00	0,5189	5,05144	14,399	246,221
r201_r	0,00	100,00	2,3825	13,05357	6,517	43,479
r202_r	0,00	100,00	2,2607	12,20386	6,681	47,012
r203_r	0,00	100,00	0,1765	3,27298	25,779	731,704
r205_r	0,00	100,00	1,8838	10,24291	7,410	60,991
r206_r	0,00	100,00	1,9988	10,55001	6,950	53,155
r207_r	0,00	100,00	0,0301	1,54686	60,273	3.793,126
r209_r	0,00	100,00	0,4886	5,73577	14,911	239,276
r210_r	0,00	100,00	2,0222	11,85317	6,971	50,734
r211_r	0,00	100,00	0,2183	3,93270	22,513	539,334
r212_r	0,00	100,00	11,8861	29,55804	2,370	3,983
r213_r	0,00	100,00	1,6037	10,26378	8,054	69,066
r215_r	0,00	100,00	1,3365	9,07092	8,902	86,205
r290_r	0,00	100,00	5,7795	19,11779	3,858	14,688
r291_r	0,00	100,00	5,5809	18,76178	3,987	15,776
r292_r	0,00	100,00	0,5356	6,05888	14,504	224,236
r299_r	0,00	100,00	6,3922	21,16536	3,705	12,843
gw_sex_r	0,00	100,00	0,6043	6,05232	13,568	203,396
ov_sex_r	0,00	100,00	0,7183	7,42486	12,070	151,522
gw_ovg_r	0,00	100,00	16,9810	30,54600	1,836	2,100
vm_gw_r	0,00	100,00	2,6394	11,23188	6,272	44,932

vm_ovg_r	0,00	100,00	31,7413	39,04955	0,755	-1,057
vern_oo_r	0,00	100,00	12,3477	25,71827	2,402	4,937
verkeer_r	0,00	100,00	23,1891	38,18785	1,314	-0,026
opium_r	0,00	100,00	3,6080	14,68044	5,219	28,728
overig_r	0,00	100,00	8,1710	22,60079	3,278	9,971
gw_sexpj_r	0,00	100,00	0,3751	5,42862	16,700	292,031
ov_sexpj_r	0,00	100,00	0,5444	6,89310	13,612	188,685
gw_ovgpj_r	0,00	100,00	12,3250	28,71042	2,367	4,197
vm_gwpj_r	0,00	100,00	1,3525	9,60827	8,654	79,762
vm_ovgpj_r	0,00	100,00	18,1024	34,49731	1,722	1,291
vern_oopj_r	0,00	100,00	8,0427	23,71787	3,180	8,993
verkeerpj_r	0,00	100,00	18,3952	35,81948	1,685	1,070
opiumpj_r	0,00	100,00	2,5819	13,53073	6,182	39,265
overigpj_r	0,00	100,00	6,1878	21,50354	3,752	12,993
januari_r	0,00	100,00	8,7645	22,89013	3,094	8,867
februari_r	0,00	100,00	7,4954	21,51259	3,430	11,174
maart_r	0,00	100,00	8,1433	22,22538	3,262	9,995
april_r	0,00	100,00	8,4602	23,09105	3,188	9,304
mei_r	0,00	100,00	8,7077	23,22489	3,119	8,902
juni_r	0,00	100,00	8,8520	23,49834	3,076	8,579
juli_r	0,00	100,00	7,3607	21,12107	3,483	11,669
augustus_r	0,00	100,00	8,2298	22,62951	3,226	9,647
september_r	0,00	100,00	8,4258	23,07531	3,197	9,356
oktober_r	0,00	100,00	9,3647	24,24055	2,963	7,804
november_r	0,00	100,00	8,5477	22,87855	3,153	9,199
december_r	0,00	100,00	7,6483	21,53156	3,378	10,880
somB02_r	0,00	100,00	21,7247	34,06960	1,379	0,455
somC02_r	0,00	100,00	20,0474	33,03412	1,538	0,958
somC06_r	0,00	100,00	7,4718	19,61314	3,336	11,423
somC14_r	0,00	100,00	13,5020	27,78057	2,230	3,873
somC15_r	0,00	100,00	0,7347	6,05574	12,180	171,562
somC20_r	0,00	100,00	0,2988	3,98250	19,323	427,244
somC22_r	0,00	100,00	0,7830	6,41627	11,921	162,729
somD03_r	0,00	100,00	4,2681	16,67260	4,690	22,314
somD08_r	0,00	100,00	2,9723	13,13288	5,775	35,948
somD09_r	0,00	100,00	1,0889	7,85981	10,136	114,123
somD11_r	0,00	100,00	2,2905	10,81955	6,774	51,814
somD12_r	0,00	100,00	1,4487	8,69510	8,903	89,172
somD13_r	0,00	100,00	0,2230	3,51365	22,696	585,357
somD17_r	0,00	100,00	31,2277	41,25250	0,860	-1,020
somD19_r	0,00	100,00	0,0181	0,93370	83,264	8.190,416
SomPrive_r	0,00	100,00	15,1482	29,78398	2,009	2,741
SomSemi_r	0,00	100,00	28,5741	38,07878	0,965	-0,667
SomOpenbaar_r	0,00	100,00	55,3024	42,39433	-0,194	-1,655
somF02_r	0,00	100,00	0,3573	4,72509	17,225	326,040
somF07_r	0,00	100,00	3,0949	14,49092	5,551	31,698
somF08_r	0,00	100,00	0,6246	6,91650	13,061	177,853
somF09_r	0,00	100,00	11,9315	26,09604	2,442	5,002
somF12_r	0,00	100,00	0,0715	2,03855	38,286	1.654,903
somF14_r	0,00	100,00	9,1879	24,34968	2,926	7,534
somH01_r	0,00	100,00	21,1551	37,04814	1,446	0,356

somH02_r	0,00	100,00	2,0190	9,57778	7,055	58,888
somH03_r	0,00	100,00	0,1485	3,01257	27,305	835,218
somH04_r	0,00	100,00	0,1441	3,46948	26,630	735,321
somH05_r	0,00	100,00	0,1129	2,90111	32,330	1.092,861
somH06_r	0,00	100,00	1,4006	10,30800	8,512	74,759
somH07_r	0,00	100,00	0,9891	8,05839	10,079	109,805
somH08_r	0,00	100,00	3,8106	15,77856	4,990	25,498
somH09_r	0,00	100,00	6,8197	21,96523	3,546	11,608
SomZondag_r	0,00	100,00	13,8272	28,82755	2,226	3,678
SomMaandag_r	0,00	100,00	10,0307	24,08407	2,868	7,448
SomDinsdag_r	0,00	100,00	10,6739	24,84911	2,736	6,618
SomWoensdag_r	0,00	100,00	11,1990	25,59102	2,640	6,015
SomDonderdag_r	0,00	100,00	12,7039	27,12253	2,392	4,630
SomVrijdag_r	0,00	100,00	14,3345	28,99079	2,186	3,523
SomZaterdag_r	0,00	100,00	15,8214	30,23154	2,002	2,691
SomNacht_r	0,00	100,00	26,1454	37,60780	1,138	-0,322
SomOchtend_r	0,00	100,00	11,5107	26,30743	2,558	5,464
SomMiddag_r	0,00	100,00	26,7212	37,04150	1,087	-0,371
SomAvond_r	0,00	100,00	24,2231	35,47707	1,283	0,155
Karnebeek_r	0,00	100,00	2,2068	11,83198	6,862	50,001
JanHendrikstraat_r	0,00	100,00	10,3663	25,39746	2,690	6,202
Heemstraat_r	0,00	100,00	5,4137	18,26190	4,060	16,577
Hoefkade_r	0,00	100,00	4,5945	16,65836	4,542	21,258
Overbosch_r	0,00	100,00	4,4211	17,98922	4,603	20,593
Segbroek_r	0,00	100,00	5,5918	19,16965	4,019	15,803
Laak_r	0,00	100,00	6,2239	20,18243	3,745	13,493
DHZuidWest_r	0,00	100,00	8,3580	23,90846	3,099	8,473
Zuiderparklaan_r	0,00	100,00	5,5482	18,90226	4,063	16,270
YpenburgLeidschenveen_r	0,00	100,00	2,2230	13,76974	6,576	42,662
Zoetermeer_r	0,00	100,00	9,4762	27,66623	2,793	6,073
Rijswijk_r	0,00	100,00	4,3138	17,39538	4,663	21,447
Westland_r	0,00	100,00	6,0043	22,42137	3,755	12,505
Wassenaar_r	0,00	100,00	1,8847	12,64669	7,220	51,700
VoorburgLeidschendam_r	0,00	100,00	6,3022	21,62846	3,688	12,447
Delft_r	0,00	100,00	7,0172	23,63616	3,409	10,107
PijnackerNootdorp_r	0,00	100,00	1,9829	13,00697	6,996	48,452
SomOnbekendDH	0,00	17,00	0,0491	0,37543	20,381	693,148
onbekendDH_r	0,00	100,00	0,6292	5,87541	13,560	207,798
Scheveningen_r	0,00	100,00	5,9005	20,29840	3,869	14,176