

Voetbalanalyse

met betrekking tot gegevens uit de Eredivisie en Jupiler League

Author(s)

Piersma, Nanda

Publication date

2016

[Link to publication](#)

Citation for published version (APA):

Piersma, N. (2016). *Voetbalanalyse: met betrekking tot gegevens uit de Eredivisie en Jupiler League*.

General rights

It is not permitted to download or to forward/distribute the text or part of it without the consent of the author(s) and/or copyright holder(s), other than for strictly personal, individual use, unless the work is under an open content license (like Creative Commons).

Disclaimer/Complaints regulations

If you believe that digital publication of certain material infringes any of your rights or (privacy) interests, please let the Library know, stating your reasons. In case of a legitimate complaint, the Library will make the material inaccessible and/or remove it from the website. Please contact the library: <https://www.amsterdamuas.com/library/contact>, or send a letter to: University Library (Library of the University of Amsterdam and Amsterdam University of Applied Sciences), Secretariat, Singel 425, 1012 WP Amsterdam, The Netherlands. You will be contacted as soon as possible.

Voetbalanalyse

met betrekking tot gegevens uit de Eredivisie en Jupiler League



Datum

10 januari 2016

in opdracht van *Johan Sports*

Nanda Piersma

Voetbalanalisten

Dave Baas

Jeroen Kits

Mitchell Schut

Bas Smissaert

Inhoudsopgave

1. Inleiding	3
2. Positievoorspelling	4
2.1. Visualisatie.....	4
2.2. Gemiddelden	6
2.3. Weka	8
3. Eventvoorspelling	14
3.1. Visualisatie.....	14
3.2. Gemiddelden	15
3.3. Weka	15
4. Niveauvoorspelling.....	17
4.1 Visualisatie	17
4.2 Gemiddelden	17
4.3 Weka	19
5. Alternatieve Aanpak	20
6. Evaluatie	23
Literatuurlijst.....	Error! Bookmark not defined.
Bijlagen	24

1. Inleiding

Deze data-analyse heeft betrekking op diverse loop- en sprintgegevens van profvoetballers uit de Eredivisie (niveau 1) en Jupiler League (niveau 2). Aan de hand hiervan zijn er hopelijk voorspellingen op het gebied van spelersposities, events (wedstrijden of trainingen) en niveau te doen. Johan Sports voert dit onderzoek uit omdat wordt geconstateerd dat sportprestaties kunnen worden verbeterd aan de hand van bepaalde trackers. De verkregen dataset kan gebruikt worden om diverse individuele spelersprestaties in kaart te brengen. Aan de hand hiervan kunnen bepaalde spelers met een hogere potentie worden ontdekt of kunnen er voor andere spelers specifieke trainingsprogramma's worden opgesteld om bepaalde eigenschappen te verbeteren. Gedurende dit onderzoek zal op zoek worden gegaan naar gelijkenissen waarbij het van belang is welke spelers qua sportprestaties met elkaar vergeleken kunnen worden.

De opdrachtgever Johan Sports (2016) heeft ons de vraag gesteld of er een clustering te maken is uit de gegeven dataset waarin gegevens van 27-07-2015 tot en met 1-12-2015 zijn gemeten. Er zal met name worden ingegaan op het onderscheid wat gemaakt wordt tussen wedstrijden en trainingen, niveau en spelersposities. De onderzoeksvraag luidt in eerste instantie:

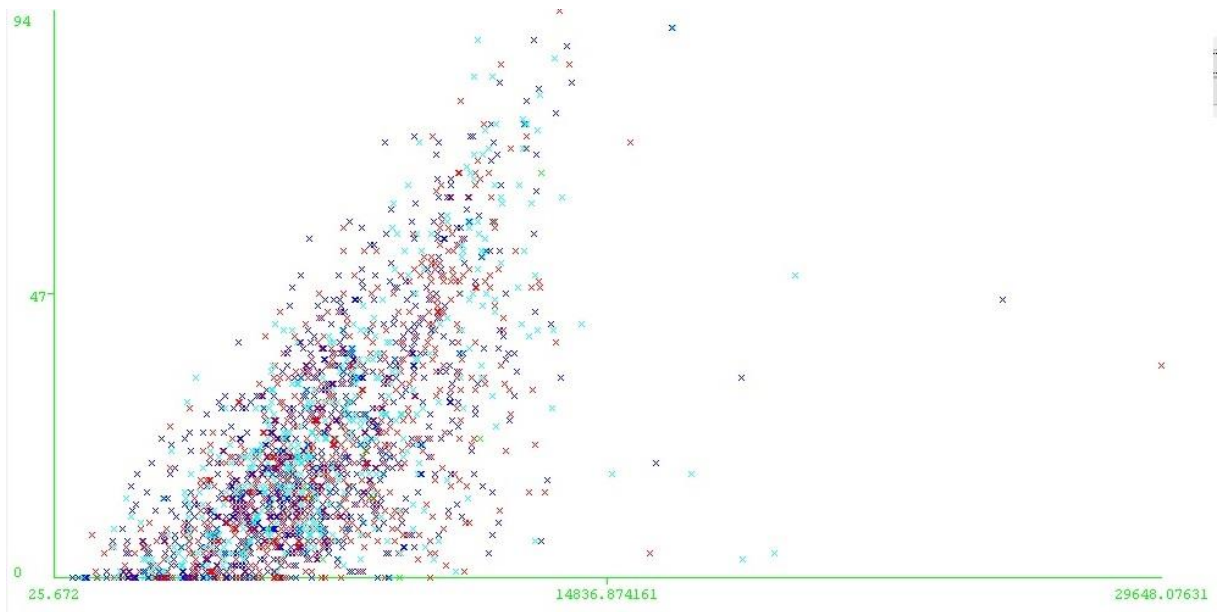
Welke clustering en welke classificaties zijn te vinden in de trackerdata van profvoetballers van twee niveaus (Eredivisie en Jupiler League), in de periode 28-07-2015 t/m 1-12-2015, waarbij gebruik wordt gemaakt van het softwarepakket Weka? (Johan Sports, 2016)

Met behulp van het data-analyseprogramma Weka is het onderzoeksteam op zoek gegaan naar eventuele clusters en classificaties. Logischerwijs is de dataset op voorhand gefilterd in Excel, zodat er geen onzuivere of onbekende datapunten meegenomen zijn in de data-analyse. Door ons wiskundig inzicht te gebruiken wordt er een basis gelegd voor een bepaald deelonderzoek, waarna gericht kan worden gezocht naar eventuele resultaten met betrekking tot wedstrijden en trainingen, niveau en spelersposities.

2. Positievoorspelling

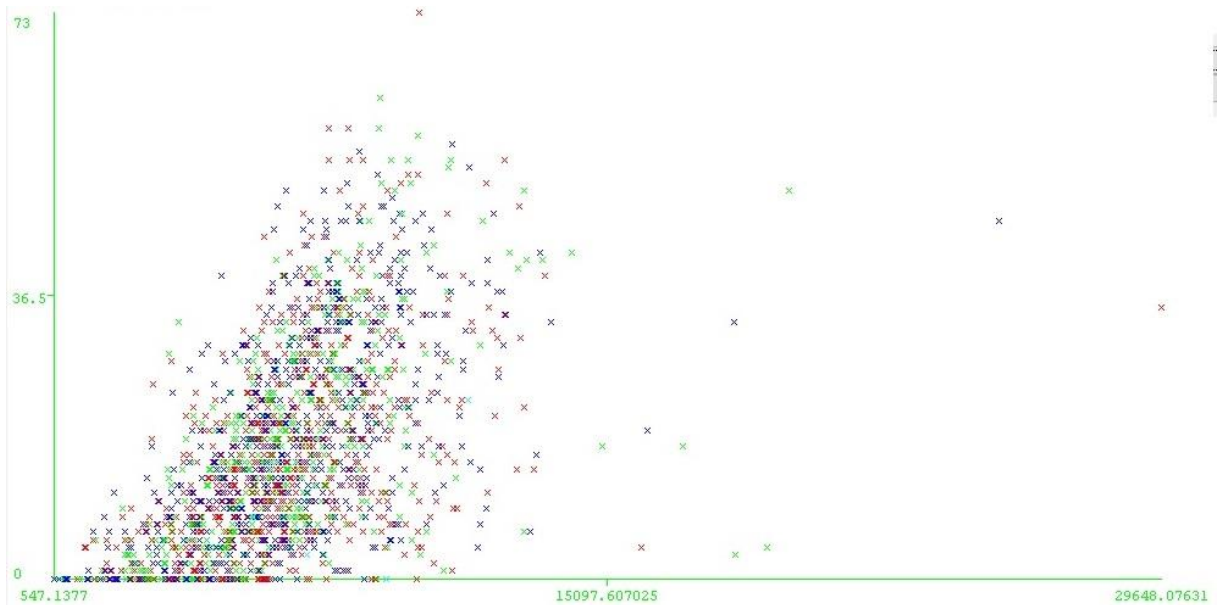
2.1. Visualisatie

Aanvankelijk was het de bedoeling om aan de hand van de verkregen dataset, waarin data van trainingen en wedstrijden beschikbaar is, een voorspelling te doen over de posities van voetballers uit de Jupiler League en Eredivisie. Om hierover uitspraken te kunnen doen, zijn allereerst diverse plots gemaakt waarmee in beeld kon worden gebracht hoe bepaalde variabelen zich tegenover elkaar verhouden. Al snel bleek dat variabelen met betrekking tot het afleggen van (sprint)afstanden weinig invloed hebben op de positie. De posities hebben een verschillende kleur gekregen. Er is te zien dat er niet bepaalde kleur ergens uitschieten. In figuur 1 is hier een voorbeeld van te zien, waarbij de totaal afgelegde afstand staat weergegeven tegenover het aantal sprints.



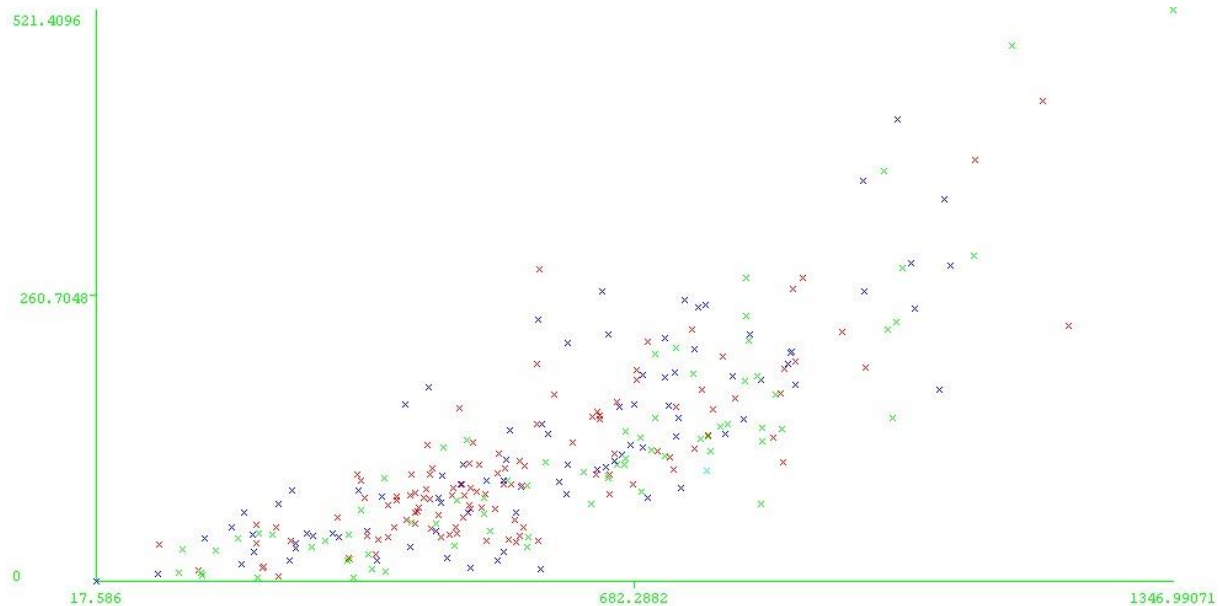
Figuur 1: totaal afgelegde afstand versus aantal sprints, met gekleurde posities (gehele dataset)

Nadat de volledige dataset geanalyseerd is en daar geen resultaat uit voort kwam, is dezelfde analyse nogmaals uitgevoerd met enkel de trainingen. Wellicht zou er aan de hand van de afgelegde afstanden op trainingen wel een voorspelling van de posities kunnen worden gedaan. Ook ditmaal zijn er geen opvallende elementen zichtbaar wanneer er wordt gekeken naar de verschillende posities, zoals in figuur 2 te zien is. In deze figuur staan wederom de totaal afgelegde afstand en het aantal sprints tegenover elkaar.



Figuur 2: totaal afgelegde afstand versus aantal sprints, met gekleurde posities (dataset trainingen)

Tevens is er een mogelijkheid om de trainingen buiten beschouwing te laten en enkel focus te leggen op de wedstrijden. Wanneer dit in de praktijk wordt gebracht, blijkt dat er wederom geen opvallende patronen ontstaan in de uitgevoerde plots. In figuur 3 is hier een voorbeeld van te zien, waarin de sprintafstand en de hoge intensiteitssprintafstand tegenover elkaar in kaart zijn gebracht.



Figuur 3: sprintafstand versus hoge intensiteitssprintafstand, met gekleurde posities (dataset wedstrijden)

Doordat de bovengenoemde plots geen zichtbare aanleidingen geven tot vervolgonderzoek in Weka, zal eerst worden gekeken naar de gemiddelden. Wellicht zorgt dit voor een nieuwe ingeving, waarop kan worden voortgeborduurd.

2.2. Gemiddelden

In de onderstaande tabel, tabel 1, staan diverse gemiddelden die berekend zijn aan de hand van de dataset. Deze gemiddelden hebben betrekking op de verschillende posities van de spelers.

	Positie	Event	Competitie	Tijdsduur (s)	Afgelegde Afstand (m)	Spelersbelasting (-)	Sprintafstand (m)	Aantal Sprints (-)	Afgeleverde Afstand Hoge Intensiteitssprints (m)	Aantal Hoge Intensiteitssprints (-)
Deel 1	A	W	1+2	4885.623	8355.452	319.2569	494.3657	44.232	116.4825	10.81651
	M	W	1+2	5131.808	9137.519	354.398	508.9402	45.32632	98.49328	9.388889
	V	W	1+2	5190.49	9066.893	354.7372	501.5115	43.96992	105.6901	9.826446
Deel 2	A	W+T	1	5736.653	6895.222	239.1761	299.7725	24.63793	77.08696	4.982456
	A	W+T	2	6715.793	7252.416	280.442	261.4508	22.40747	73.47698	4.528302
	M	W+T	1	5884.084	7184.411	254.1535	293.7227	24.08939	73.22865	4.603352
	M	W+T	2	6535.74	7073.873	277.6851	240.4069	20.89394	59.57076	3.910761
	V	W+T	1	5781.611	7252.685	256.1025	304.8668	24.10048	83.9118	5.134615
	V	W+T	2	6813.82	7184.29	278.2141	235.0218	19.83837	66.63453	4.050562
Deel 3	A	W	1	4464.259	7742.683	270.4145	474.3898	40.84	128.2839	10.63636
	A	W	2	4996.078	8516.08	332.0602	499.3596	45.08	113.4983	10.86207
	A	T	1	5967.221	6741.656	233.5154	270.4743	21.91946	69.52769	4.147651
	A	T	2	7099.193	6970.69	268.934	209.9554	17.5	65.61727	3.284424
Deel 4	M	W	1	4918.417	9479.31	335.3423	549.6068	43.70833	117.3564	9.833333
	M	W	2	5200.093	9028.147	360.4959	495.1938	45.87324	91.63396	9.227273
	M	T	1	6033.606	6829.072	241.5824	254.1019	21.05161	66.39596	3.793548
	M	T	2	6843.966	6622.886	258.5749	184.7458	15.43692	52.85275	2.796825
Deel 5	V	W	1	5192.737	9415.942	359.1094	509.226	44.97222	104.5367	9.628571
	V	W	2	5189.676	8940.57	353.1548	498.6484	43.59794	106.1595	9.906977
	V	T	1	5910.96	6777.518	233.4767	262.3412	19.75723	79.73913	4.225434
	V	T	2	7179.775	6788.562	261.3283	180.1468	14.8927	59.04715	2.926339

Tabel 1: gemiddelden per positie, weergegeven per verschillende events en niveaus.

Zoals in de bovenstaande tabel te zien is, is deze onderverdeeld in vijf verschillende delen. Binnen ieder deel zijn er een aantal opvallendheden die hieronder worden besproken. De opvallende getallen zijn weergegeven in een donkeroranje kleur, zodat het onderscheid duidelijk is.

Gemiddelden van onderlinge posities in wedstrijden in Eredivisie en Jupiler League (Deel 1 in tabel)

Wanneer er wordt gekeken naar de bovenstaande tabel, valt op dat aanvallers in het algemeen minder afstand afleggen in vergelijking met middenvelders en verdedigers. Echter, zij nemen over het algemeen minder lang deel aan de wedstrijd. De spelersbelasting bij aanvallers ligt dan ook gemiddeld iets lager dan bij middenvelders en verdedigers. De sprintafstand en aantal sprints is nagenoeg gelijk, terwijl aanvallers en verdedigers gemiddeld meestal wel meer hoge intensiteitssprints afleggen. De totale afstand binnen deze intensiteitssprints is bij alle posities ongeveer gelijk.

Gemiddelden van wedstrijden en trainingen gezamenlijk (Deel 2 in tabel)

Zodra de gegevens van de wedstrijden en trainingen worden samengenomen, ontstaan er vertekende beelden in de totale gemiddelden. Zo valt het bijvoorbeeld op dat de gemiddelden in vergelijking met enkel de wedstrijden aanzienlijk lager zijn wanneer wordt gekeken de variabelen. Om concrete uitspraken te kunnen doen, wordt er bij voorkeur gekeken naar aangepaste datasets waarin enkel wedstrijden of enkel trainingen staan weergegeven.

Aanvallers, onderverdeeld op event en niveau (Deel 3 in tabel)

Als er wordt gekeken naar de spelersbelasting, valt het op dat aanvallers in de Jupiler League over een relatief hoge spelersbelasting beschikken in vergelijking met de Eredivisie. Hierbij moet worden opgemerkt dat aanvallers in de Jupiler League gemiddeld langer deelnemen aan de wedstrijd dan aanvallers uit de Eredivisie. Ondanks de lagere spelersbelasting van Eredivisie spelers, leggen zij wel meer afstand tijdens hoge intensiteitssprints. Hoewel er in de Jupiler League gemiddeld veel langer getraind wordt, leggen de aanvallers in de Eredivisie gemiddeld ruim 60 meter meer af in de sprints op de trainingen.

Middenvelders, onderverdeeld op event en niveau (Deel 4 in tabel)

Middenvelders in de Eredivisie leggen niet alleen algemeen meer afstand af gedurende de wedstrijden en trainingen, ook is hun afgelegde (hoge intensiteits)sprintafstand hoger. Voor zowel de wedstrijden als de trainingen geldt dat men in de Jupiler League dit binnen een langer tijdsbestek gaat.

Verdedigers, onderverdeeld op event en niveau (Deel 5 in tabel)

Gemiddeld leggen verdedigers in de Eredivisie tijdens wedstrijden een halve kilometer meer af dan spelers in de Jupiler League. Ondanks dat er in de Jupiler League vele malen langer wordt getraind, is de sprintafstand op trainingen in de Eredivisie gemiddeld ruim 80 meter hoger. Ook de afstand van het aantal afgelegde hoge intensiteitssprints is ongeveer 20 meter hoger in vergelijking met de trainingen in de Jupiler League.

2.3. Weka

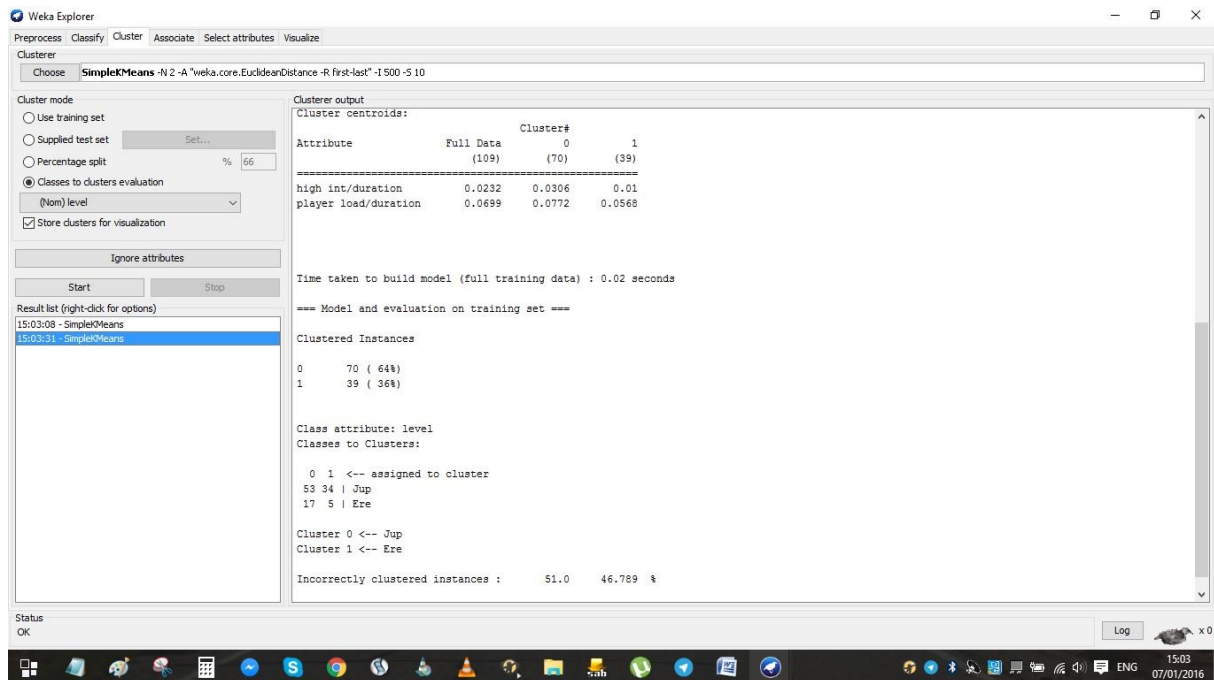
Uit de gegevens van tabel 1 blijkt dat er enige opvallende verschillen zijn te vinden tussen het niveau waarop spelers spelen en hun positie. Deze waarden zijn in de tabel allemaal in het donkeroranje gemarkeerd. De variabelen behorend bij deze gekleurde waarden zijn in Weka geanalyseerd om te kijken of er geclassificeerd en geclusterd kan worden. In de tabel bij deel 3 is bijvoorbeeld te zien dat er een verschil is op te merken tussen spelersbelasting en afgelegde afstand in hoge intensiteitssprints voor de verschillende wedstrijd niveaus, namelijk Jupiler league en Eredivisie. Zo zijn er ook bij deel 4 en deel 5 van de tabel verschillende terug te vinden in bepaalde type metingen voor middenvelders en verdedigers in de Jupiler League, vergeleken met de Eredivisie. Ook al deze data is in Weka geanalyseerd om te proberen ze te classificeren en te clusteren en deze clusters visueel uit te beelden. Omdat dit een hoop verschillende schermen met resultaten opleverde zijn de meeste hiervan toegevoegd aan de bijlagen.

Onze aanpak was als volgt. Allereerst is er een aangepaste dataset gemaakt waarin alle metingen werden gedeeld door tijdsduur om zo een wat evenwichtiger beeld te geven per speler. Immers niet elke speler maakt evenveel minuten tijdens wedstrijden, omdat deze kunnen worden gewisseld of later pas worden ingezet. Op deze manier is de dataset nog net wat zuiverder gemaakt en kan hij worden geanalyseerd met behulp van Weka. Hierin werd gebruik gemaakt van SimpleKMeans, om te kijken of er clusters te vinden waren. Dit deden we eerst met alle variabelen, dus ook diegenen waarin eigenlijk geen groot verschil zat. Hieruit bleek dat er weinig te concluderen viel en er eigenlijk voor alle drie de posities aanvallers, middenvelders en verdedigers geen clusters te ontdekken zijn.

Vervolgens werd er wederom met behulp van de SimpleKMeans gefilterd op alleen de oranje gemarkeerde attributen van spelers, die eventueel wel een significant verschil zouden moeten kunnen weergeven aan de hand van classificaties en clusters. Hieronder volgen per positie de variabelen die er per positie opvielen, de resultaten die van de SimpleKMeans en nog een visualisatie om aan te tonen of er wel of geen clusters te vormen waren.

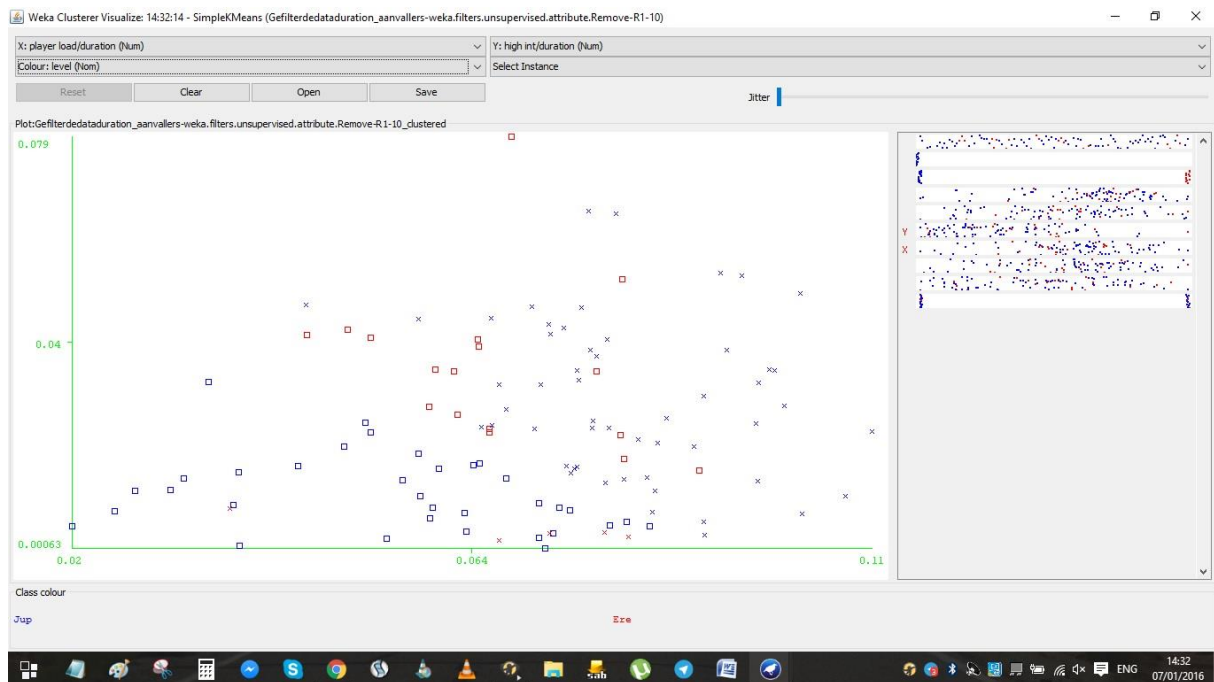
Aanvallers

Bij aanvallers viel het op dat er een groot verschil was tussen spelersintensiteit en hoge intensiteitssprints uit de Eredivisie en Jupiler League. In figuur 4 is de SimpleKMeans zichtbaar, behorend bij deze situatie.



Figuur 4: spelersintensiteit en afgelegde afstand van hoge intensiteitssprints met betrekking tot aanvallers

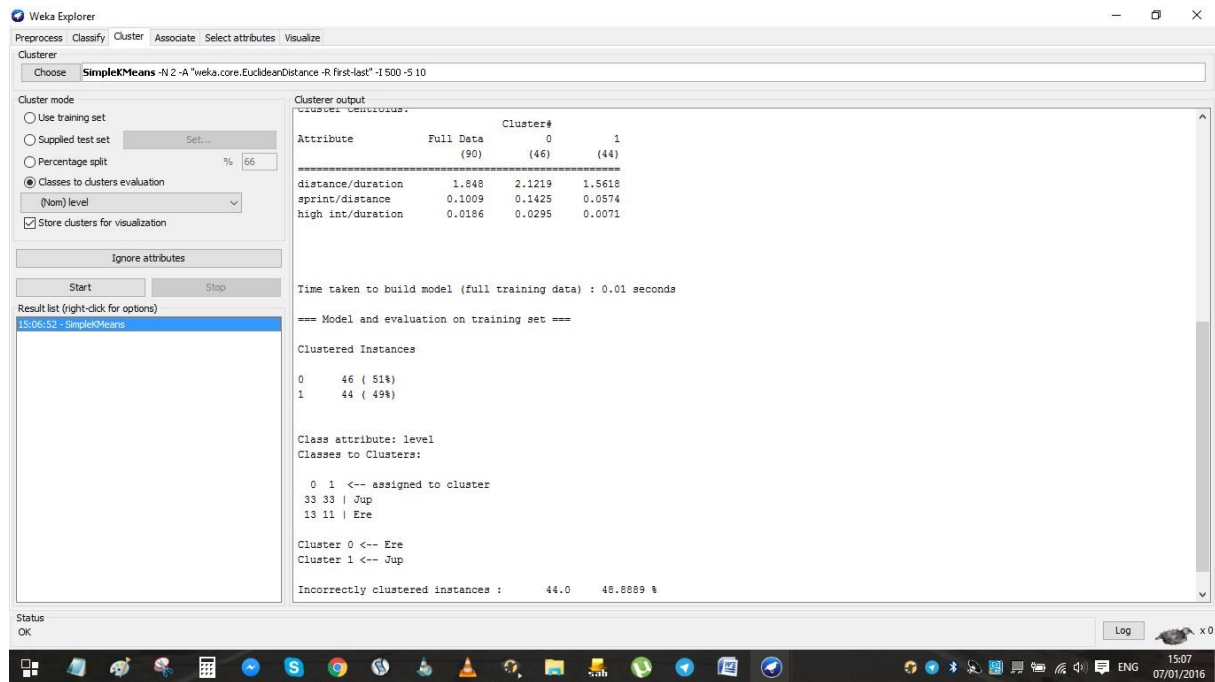
Uit deze data valt te interpreteren dat er geen clusters op te maken zijn tussen de verschillende niveaus van de aanvallers. Beide niveaus vallen ongeveer evenveel in zowel het nulde als het eerste cluster en tevens is het percentage incorrect geclusterde data 46%. Dit is veel te hoog om te concluderen dat er daadwerkelijk een cluster is. Dit werd vervolgens ondersteund door de data te visualiseren, deze visualisatie is te zien in de figuur 5. Hierbij zijn de rode punten representatief voor de Eredivisie, terwijl de blauwe punten de gegevens uit de Jupiler League weergeven.



Figuur 5: visualisatie spelersintensiteit en afgelegde afstand van hoge intensiteitssprints met betrekking tot aanvallers

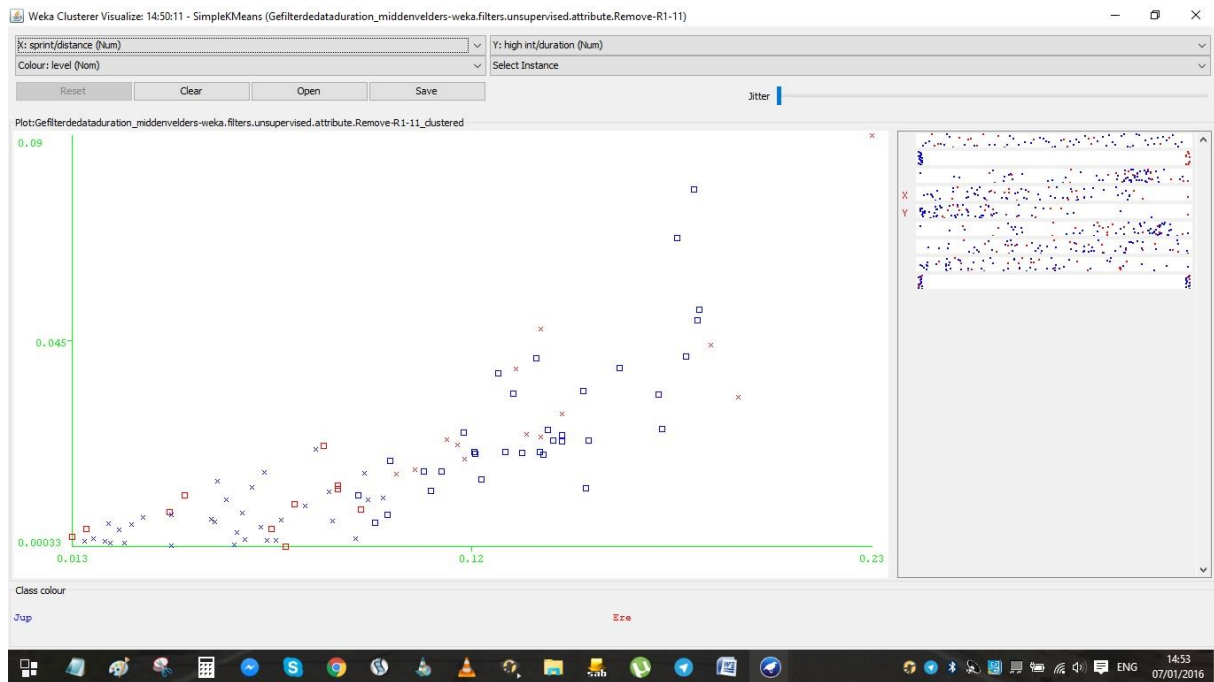
Middenvelders

Vervolgens is dezelfde methode toegepast wanneer er werd gefocust op de middenvelders, waarbij wederom op zoek is gegaan naar verschillen tussen bepaalde variabelen wat betreft het niveau. Dezelfde stappen zijn in Weka uitgevoerd. Eerst werden alle variabelen meegenomen waaruit niks op te maken was, vervolgens zijn alleen de interessante variabelen meegenomen in de analyse. Dit leverde de resultaten op die weergegeven zijn in figuur 6.



Figuur 6: afgelegde afstand en afgelegde afstand van hoge intensiteitssprints met betrekking tot middenvelders

Wederom valt er af te lezen dat beide niveau's gelijk zijn verdeeld over de clusters en is de error zelfs bijna 49%. Ook uit de visualisatie blijkt dat er duidelijk geen clusters te vinden zijn. Er waren drie variabelen dus deze hebben we alle drie tegenover elkaar gezet. In figuur 7 is een voorbeeld weergegeven, terwijl de overige situaties zijn weergegeven in de bijlagen.



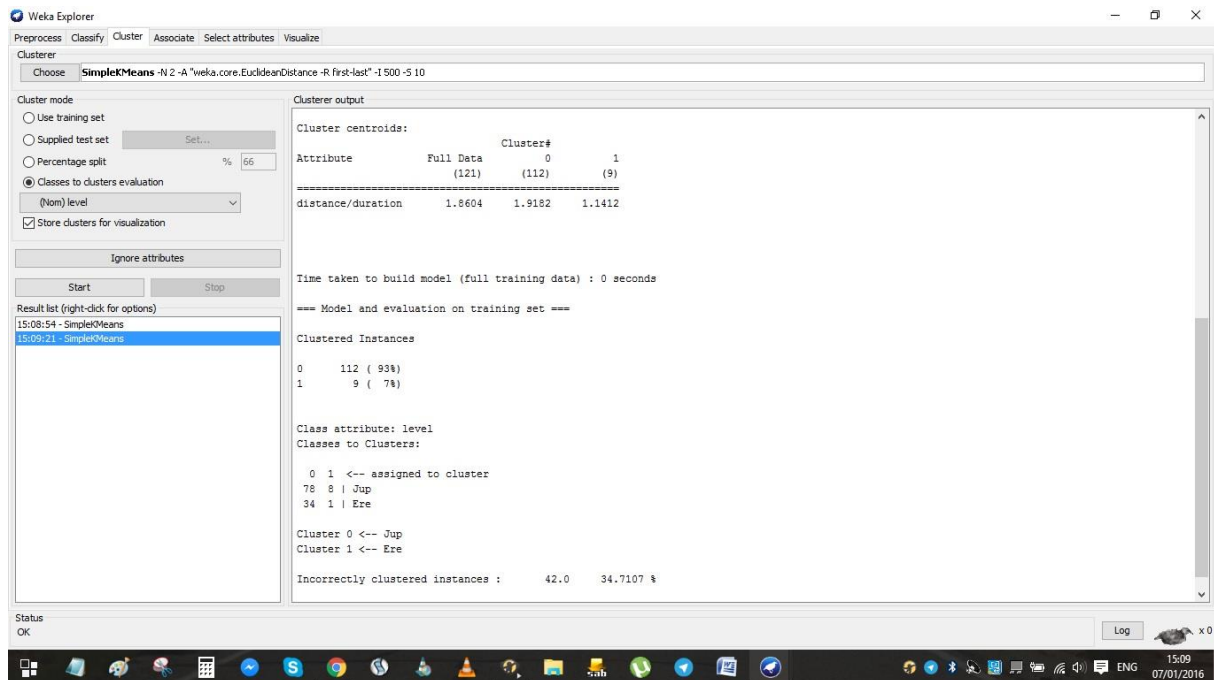
Figuur 7: visualisatie afgelegde afstand en afgelegde afstand van hoge intensiteitssprints met betrekking tot middenvelders

Hierbij is de afgelegde sprintafstand tegen de afstand van de hoge intensiteitssprints weergegeven, die beide van elkaar verschilden per niveau bij middenvelders. Dit is af te lezen in tabel 1.

Verdedigers

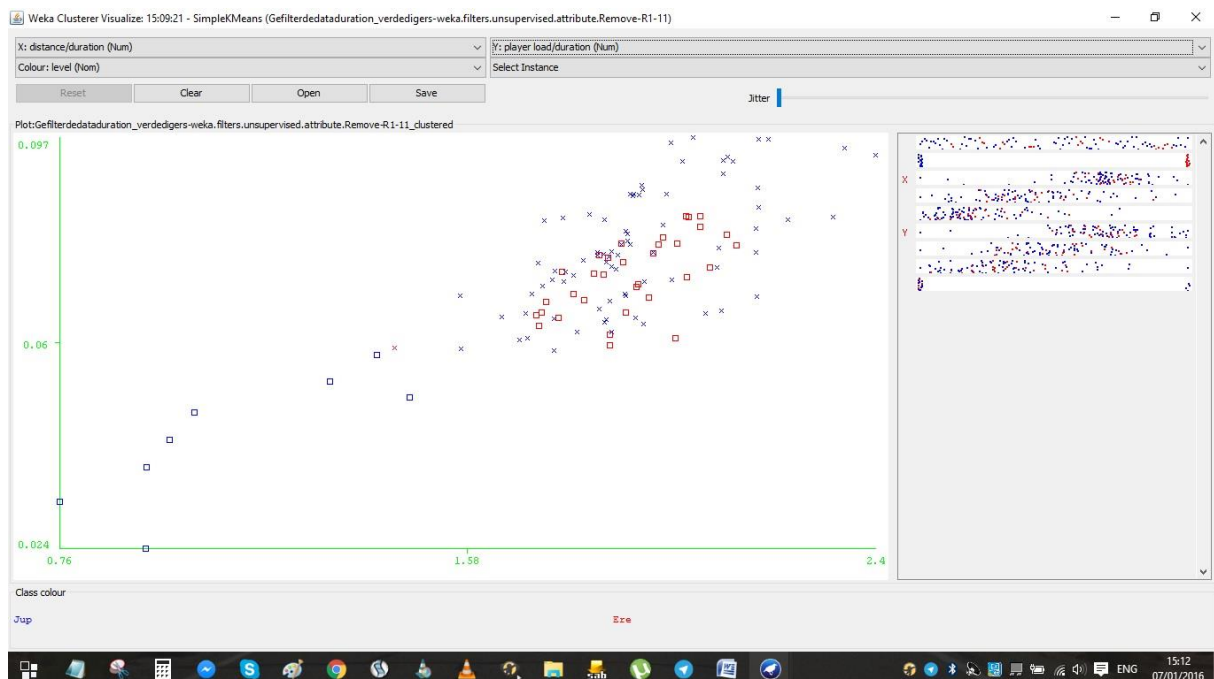
Tot slot is er gekeken naar opvallende verschillen in variabelen van verdedigers in de Jupiler League, vergeleken met Eredivisie. In tabel 1 valt af te lezen dat er slecht één variabele aanzienlijk verschilt, namelijk de totale afgelegde afstand per wedstrijd van de spelers.

Wederom is de dataset allereerst aangepast, waarna de analyse is uitgevoerd middels SimpleKMeans. Hieruit kwam weinig verrassends, er waren geen clusters op te maken. Daarna werden alleen de totale afgelegde afstand per spelers per niveau meegenomen in de SimpleKMeans. Hieruit volgde het resultaat in figuur 8.



Figuur 8: afgelegde afstand met betrekking tot verdedigers

Wederom valt er weinig te clusteren omdat het overgrote deel van zowel Jupiler league als Eredivisie in het eerste cluster vallen. Tevens is er een error van 34%, wat nog steeds veel te hoog is om clusters te kunnen constateren. Dit wordt ondersteund door een visualisatie van de SimpleKMeans voor verdedigers, die in figuur 9 te zien is.

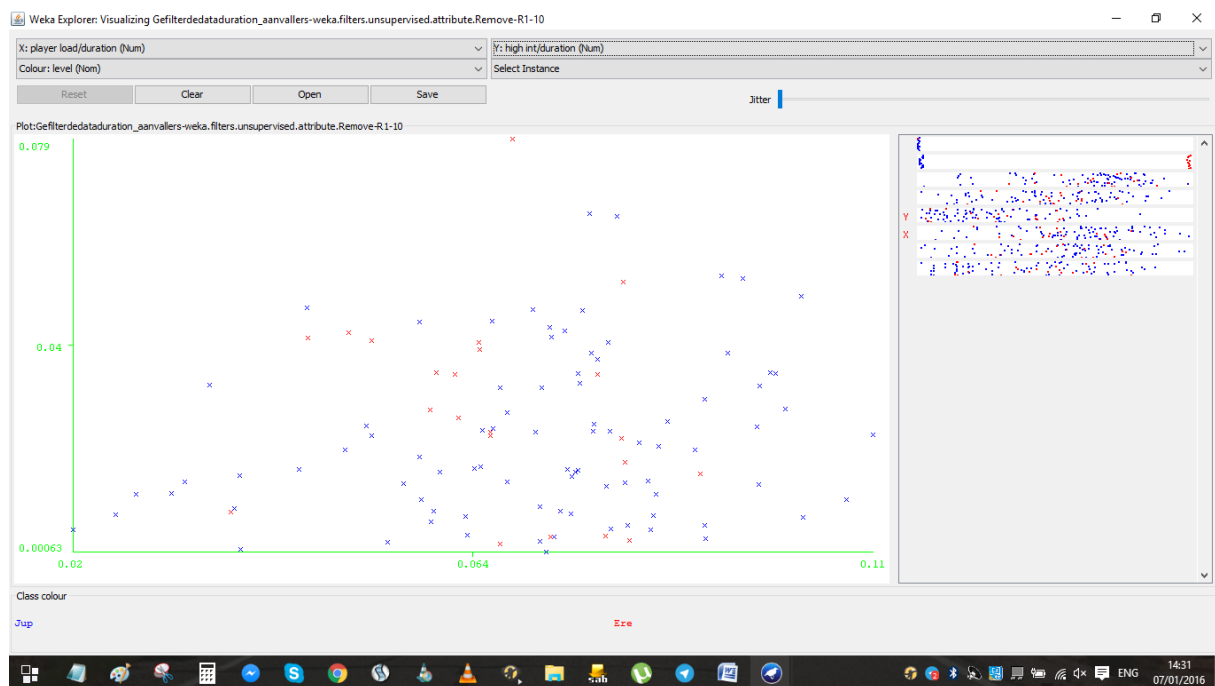


Figuur 9: visualisatie afgelegde afstand en spelersintensiteit met betrekking tot verdedigers

Omdat er maar één variabele was die enigszins verschilde, namelijk de totale afgelegde afstand, is deze afgezet tegen variabelen als het aantal sprints en de spelersintensiteit. De laatstgenoemde is in figuur 9 weergegeven. Wederom is er geen clustering op te merken en zijn de blauwe en rode datapunten door elkaar verspreid, met hier en daar wat outliers voor de spelers uit de Jupiler League.

Al met al vielen de uitgevoerde analyses aan de hand van vooraf geconstateerde verschillen bij spelers per positie in de Jupiler League vergeleken met Eredivisie dus tegen. Waar op voorhand werd gedacht wat opvallende verschillen te kunnen ontdekken, bleken deze na analyse in Weka toch niet significant genoeg te zijn om daadwerkelijk clusters te kunnen vormen en apart te kunnen classificeren.

Daarnaast is er per dataset van spelerspositie gevisualiseerd in Weka, waarmee gezocht is naar clusters door de niveaus verschillend te kleuren en vervolgens in de rijtjes van opvallende variabelen te speuren naar eventuele clusters tussen twee variabelen. Ook dit leverde weinig op. In figuur 10 is een voorbeeld te zien voor de aanvallers waar de spelersintensiteit is afgezet tegenover de afgelegde afstand in hoge intensiteitssprints. Hierbij zijn de spelers uit de Eredivisie in het rood weergegeven en de spelers uit de Jupiler League blauw. Voor de overige posities is dit terug te vinden in de bijlagen.



Figuur 10: spelersintensiteit en afgelegde afstand van hoge intensiteitssprints met betrekking tot aanvallers

Ook hieruit vielen geen clusters op te maken. Omdat er nog steeds weinig resultaat geboekt is, is vervolgens besloten om simpelweg alle wedstrijden van beide niveaus te gaan vergelijken met alle trainingen van beide niveaus. Op voorhand wordt hier een verschil verwacht, wat ook het geval bleek. Dit wordt duidelijk in het volgende hoofdstuk.

Aan de hand van de plots valt direct op dat er een duidelijk verschil is tussen trainingen en wedstrijden. De trainingen zijn in het blauw weergegeven, terwijl de rode datapunten de wedstrijden in kaart brengen. Over het algemeen lijken de rode datapunten zich meer aan de rechterkant van de plots te bevinden. Dit geeft aan dat er meer (hoge intensiteits)sprints- en afstanden zijn afgelegd in wedstrijden in vergelijking met de trainingen. Ons eerdergenoemde vermoeden lijkt dus te kloppen, blijkbaar is de intensiteit tijdens wedstrijden hoger dan tijdens trainingen. Na de afbeeldingen grondiger te hebben bestudeerd, is er allereerst gekeken naar de bijbehorende gemiddelden.

3.2. Gemiddelden

In de onderstaande tabel, tabel 2, staan diverse gemiddelden die berekend zijn aan de hand van de dataset. Deze gemiddelden hebben betrekking op de verschillende events.

Event	Competitie	Tijdsduur (s)	Afgelegde Afstand (m)	Spelersbelasting (-)	Sprintafstand (m)	Aantal Sprints (-)	Afgeleverde Afstand Hoge Intensiteitssprints (m)	Aantal Hoge Intensiteitssprints (-)
W	1	4897.764	8925.412	325.793	510.382	43.4	114.785	9.9630
W	2	5118.870	8796.685	346.988	499.016	44.836	104.803	10.079
T	1	5968.388	6783.068	236.123	262.204	20.853	72.214	4.061
T	2	7065.886	6822.220	263.847	191.363	15.989	59.364	3.005

Tabel 2: gemiddelden per event, weergegeven per verschillende niveaus

Zoals al eerder is geconstateerd, valt het op dat de gemiddelden van de trainingen in vergelijking met de wedstrijden stukke lager uitvallen. Daarnaast blijkt er in de Jupiler League aanzienlijk langer te worden getraind dan in de Eredivisie. Tussen de Eredivisie en Jupiler League onderling zijn verschillen zichtbaar als het gaat om de trainingen, met name wanneer er wordt gefocust op de sprints. Nu zowel de plots als de gemiddelden dezelfde signalen afgeven, zal er worden gekeken naar een cluster.

3.3. Weka

Met behulp van de clusterfunctie SimpleKMeans zijn er vervolgens twee clusters gevonden, zoals in figuur 12 te zien is. Allereerst is er een cluster gevonden met veel trainingen en een cluster met veel wedstrijden. In het nulde cluster zijn er 218 trainingen geclusterd en 285 wedstrijden geclusterd, terwijl er in het eerste cluster 1407 trainingen geclusterd in combinatie met 49 wedstrijden geclusterd zijn. Hierbij kan dus worden geconstateerd dat er 13,62% van de wedstrijden en trainingen verkeerd geclusterd is. Er blijkt sprake te zijn van een heel groot cluster, waarin over het algemeen alleen maar trainingen in terug te vinden zijn. Daarnaast is er nog een cluster gevonden met een merendeel aan wedstrijden. Aan de hand hiervan kan worden geconcludeerd dat wedstrijden met een hogere intensiteit worden gespeeld dan trainingen. Vrijwel alle variabelen vallen op het gebied van de wedstrijden hoger uit in vergelijking met de trainingen.

Clusterer

Choose
SimpleKMeans -N 2 -A "weka.core.EuclideanDistance -R first-last" -I 500 -S 10

Cluster mode

☐ Use training set
☐ Supplied test set
☐ Percentage split
☒ Classes to clusters evaluation

Set...
% 66
(Nom) event_type
Store clusters for visualization

Ignore attributes
Start
Stop

Result list (right-click for options)

12:58:47 - XMeans
12:59:00 - SimpleKMeans
12:59:34 - SimpleKMeans
13:00:33 - SimpleKMeans
13:00:49 - SimpleKMeans
13:01:45 - SimpleKMeans

Clusterer output

	(1959)	(503)	(1456)
distance/duration	1.2527	1.7986	1.0641
player load/duration	0.0479	0.0686	0.0408
sprint distance/duration	0.0479	0.1097	0.0266
number sprints/duration	0.0041	0.0086	0.0026
high int distance/duration	0.0121	0.0282	0.0065
high int number/duration	0.0008	0.002	0.0004

Time taken to build model (full training data) : 0.07 seconds

=== Model and evaluation on training set ===

Clustered Instances

0 503 (26%)
1 1456 (74%)

Class attribute: event_type
Classes to Clusters:

0 1 <-- assigned to cluster
218 1407 | TR
285 49 | MA

Cluster 0 <-- MA
Cluster 1 <-- TR

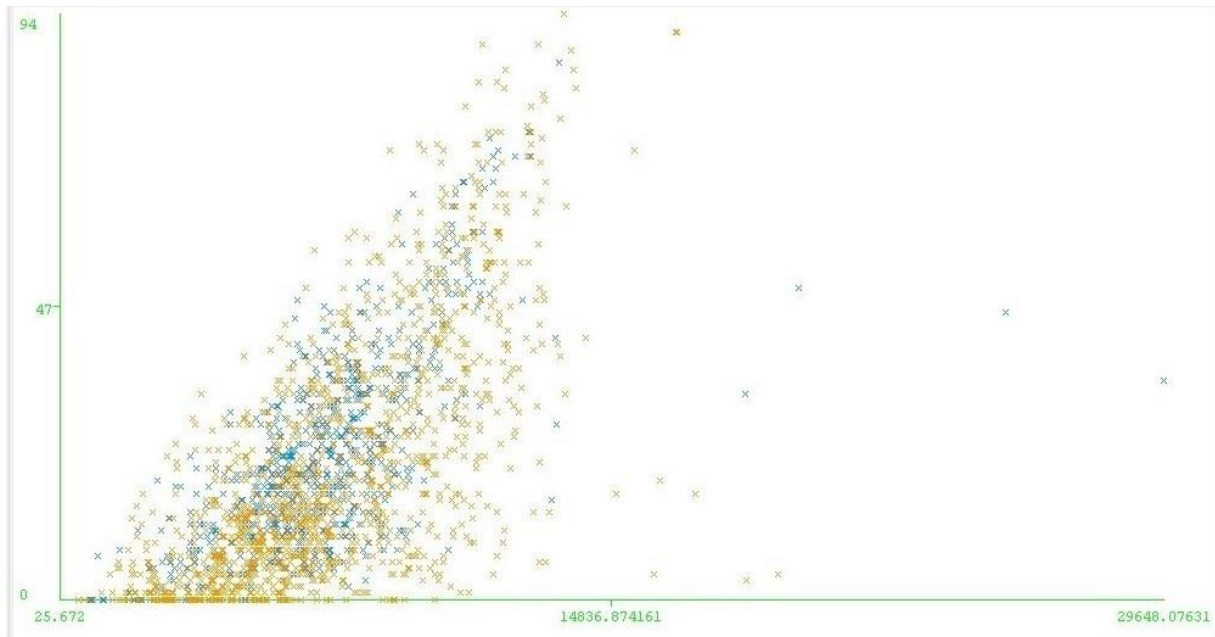
Incorrectly clustered instances : 267.0 13.6294 %

Figuur 12: SimpleKMeans met betrekking tot wedstrijden en trainingen (gehele dataset)

4. Niveauvoorspelling

4.1 Visualisatie

Ook wanneer wordt gekeken naar de voorspelling van het niveau, zal er allereerst aan de hand van wat plots een indicatie worden verkregen. Het was lastig om hieruit constatering te doen. In figuur 13 hieronder staat het aantal sprints weergegeven tegenover de totaal afgelegde afstand, waarbij onderscheid gemaakt is tussen Eredivisie (blauw) en Jupiler League (oranje). Hieruit wordt zichtbaar dat er weinig verschil zit in de onderlinge verhoudingen. Ditzelfde geldt voor de overige variabelen.



Figuur 13: aantal sprints tegenover totaal afgelegde afstand (gehele dataset)

4.2 Gemiddelden

In de onderstaande tabel, tabel 3, staan diverse gemiddelden die berekend zijn aan de hand van de dataset. Deze gemiddelden zijn gebaseerd op de gehele dataset, waarin wedstrijden gecombineerd zijn met trainingen.

Competitie	Tijdsduur (s)	Afgelegde Afstand (m)	Spelersbelasting (-)	Sprintafstand (m)	Aantal Sprints (-)	Afgeleverde Afstand Hoge Intensiteitssprints (m)	Aantal Hoge Intensiteitssprints (-)
1	5800.039	7119.939	250.223	299.740	24.263	78.393	4.918
2	6707.886	7185.267	279.134	245.312	21.048	66.844	4.169

Tabel 3: totale gemiddelden per niveau

Aan de hand van de bovenstaande tabel is het lastig om uitspraken te doen. Enerzijds geeft de tabel een vertekend beeld, aangezien de trainingen en de wedstrijden zijn gecombineerd. Al eerder is geconcludeerd dat de intensiteit in de wedstrijden aannemelijk hoger is dan op de trainingen. Anderzijds geeft de tabel wel een indicatie van de totale belasting voor spelers in het algemeen. Over het algemeen lijkt deze ongeveer gelijk te zijn, hoewel de sessies in de Jupiler League wel van een langere duur zijn. Daarnaast ligt de spelersbelasting in de Jupiler League gemiddeld hoger dan in de Eredivisie, wat betekent dat de som van de acceleraties hoger is in de Jupiler League in vergelijking met de Eredivisie.

4.3 Weka

Weka is op zoek gegaan naar clusters en vond er in deze situatie twee, zoals in figuur 14 te zien is. Uit deze gegevens kan worden afgelezen dat Weka in totaal 1584 elementen heeft gevonden uit de Jupiler League. Daarbij heeft Weka 564 elementen gevonden uit de Eredivisie. Van de 1584 elementen die in de Jupiler League zijn geplaatst zijn er 442 elementen verkeerd geplaatst. Daarbij komen ook nog de misplaatste waarnemingen uit de Eredivisie. Van de in totaal 564 geplaatste waarnemingen zijn er 162 verkeerd geplaatst. Wanneer je de fout ingedeelde waarnemingen bij elkaar optelt en deelt door het totaal, vind je het fout ingeclusterde percentage. Dit komt op een percentage van ruim 39%. Dit percentage is helaas te hoog. Er wordt namelijk vanuit gegaan dat een foutpercentage van 10% acceptabel zou zijn. Door dit te hoge percentage kan worden geconcludeerd dat Weka geen betrouwbare clusters kan vinden.

The screenshot shows the Weka Explorer interface with the SimpleKMeans classifier selected. The 'Cluster mode' section on the left has 'Classes to clusters evaluation' selected. The 'Clusterer output' section on the right displays the following data:

Clusterer output	1.2343	1.028	1.8134
total distance (m)	0.0471	0.0388	0.0704
player load (-)	0.9003	0.7612	1.2907
sprint distance (m)	#DIV/0!	#DIV/0!	0.0
number of sprints (-)	0.0	0.0	#DIV/0!
high intensity sprint distance (m)	#DIV/0!	#DIV/0!	0.0
high intensity number of sprints (-)			

Time taken to build model (full training data) : 0.5 seconds

=== Model and evaluation on training set ===

Clustered Instances

Cluster	Instances	Percentage
0	1584	(74%)
1	564	(26%)

Class attribute: Level

Classes to Clusters:

Cluster	Instances	Percentage	Assigned to cluster
0	1584	(74%)	442 162 Ere
1	564	(26%)	1142 402 Jup

Cluster 0 <-- Jup
Cluster 1 <-- Ere

Incorrectly clustered instances : 844.0 39.2924 %

Figuur 14: SimpleKMeans met betrekking tot Eredivisie en Jupiler League (gehele dataset)

5. Alternatieve Aanpak

Op voorhand is de gehele dataset verdeeld in verschillende delen, waarin er onderscheid gemaakt is tussen posities, events en niveau. Zoals eerder genoemd zijn er geen concrete resultaten verkregen door op deze manier de data te analyseren. Zodoende moesten we op zoek gaan naar een alternatieve aanpak. Mevrouw Piersma gaf het advies om verschillende klassen samen te voegen. Wellicht zou dit mogelijkheden bieden tot het vinden van clusters in Weka. Weka is in staat om meerdere punten in dezelfde klasse te plaatsen, waardoor er wellicht wel clusters te vinden zijn.

In figuur 15 is te zien welke uitkomst de bovengenoemde methode biedt, wanneer gebruik wordt gemaakt van de SimpleKMeans. Hieruit blijkt dat er vier clusters zijn gevonden op de zelfgemaakte klasse *Event_Type_Level*. Echter geeft deze een foutpercentage van ruim 49%. De clusters die gevonden zijn, blijken dus niet goed genoeg voor dit onderzoek. Dit percentage ligt te hoog.

The screenshot displays the Weka SimpleKMeans interface. The 'Clusterer' dropdown is set to 'SimpleKMeans'. The 'Cluster mode' section has 'Classes to clusters evaluation' selected. The 'Class attribute' is 'Event_Type_Level'. The 'Ignore attributes' section is empty. The 'Start' button is visible. The 'Result list' shows three entries for SimpleKMeans, with the last one selected. The 'Clusterer output' pane shows the following results:

```
Time taken to build model (full training data) : 0.25 seconds

=== Model and evaluation on training set ===

Clustered Instances

0      1128 ( 53%)
1       517 ( 24%)
2       231 ( 11%)
3       272 ( 13%)

Class attribute: Event_Type_Level
Classes to Clusters:

  0   1   2   3  <-- assigned to cluster
295 148  50  13 | TR Ere
 12  28   1  57 | MA Ere
759 295 170  32 | TR Jup
 62  46  10 170 | MA Jup

Cluster 0 <-- TR Jup
Cluster 1 <-- TR Ere
Cluster 2 <-- MA Ere
Cluster 3 <-- MA Jup

Incorrectly clustered instances :      1070.0  49.8138 %
```

Figuur 15: SimpleKMeans met betrekking tot alternatieve aanpak (gehele dataset)

Nadat de bovenstaande clustering geen gewenste resultaten gaf, is er geprobeerd om handmatig clusters te zoeken. Dit kan worden gerealiseerd door plots in Weka weer te geven, zoals dit ook al eerder is gedaan. Ook dit leverde helaas geen clusters op.

Aangezien ook dit geen uitkomst bood, is ervoor gekozen om nog een stapje verder te kijken wat betreft het samenvoegen van klassen. Ditmaal wordt er nadrukkelijk gelet gefocust op de posities van de spelers. Ditmaal gaat Weka op zoek naar clusters die de zelfgemaakte klasse *Event, Type, Level & Player Type* herkent. Bovendien is de dataset op sommige vlakken aangepast, waardoor het zoeken naar clusters gemakkelijker wordt. De numerieke data is gedeeld door de tijdsduur, zodat de data onderling vergelijkbaar is. Het aantal sprints wat speler X in een bepaalde wedstrijd gemiddeld verricht kan bijvoorbeeld meer zijn dan het aantal sprints van speler Y.

Hieronder in figuur 16 is te zien wat Weka als uitkomst geeft wanneer binnen het programma wordt gezocht op vier clusters, als gebruik wordt gemaakt van de SimpleKMeans-functie. Weka vindt weldegelijk clusters. Echter, wederom met een onnacceptabel foutpercentage. Dit geeft weinig hoop op het daadwerkelijk vinden van clusters.

Clusterer

Choose **SimpleKMeans** -N 2 -A "weka.core.EuclideanDistance -R first-last" -I 500 -S 10

Cluster mode

☐ Use training set

☐ Supplied test set Set...

☐ Percentage split % 66

☒ Classes to clusters evaluation

(Nom) Level

☒ Store clusters for visualization

Ignore attributes

Start Stop

Result list (right-click for options)

- 13:55:29 - SimpleKMeans
- 13:56:26 - SimpleKMeans
- 14:11:29 - SimpleKMeans
- 14:39:49 - SimpleKMeans
- 14:43:47 - SimpleKMeans**
- 14:44:38 - SimpleKMeans
- 14:44:48 - SimpleKMeans
- 14:45:50 - SimpleKMeans

Clusterer output

Clustered instances

Cluster	Count	Percentage
0	1128	(53%)
1	517	(24%)
2	231	(11%)
3	272	(13%)

Class attribute: Event_type_Level_Player_type

Classes to Clusters:

Cluster	0	1	2	3	Assigned to cluster
94	42	18	6	11	TR Ere ATT
7	11	0	11	11	MA Ere ATT
265	112	73	15	15	TR Jup ATT
22	20	5	58	11	MA Jup ATT
109	55	16	3	11	TR Ere DEF
3	11	0	29	11	MA Ere DEF
291	99	62	14	14	TR Jup DEF
22	19	2	63	11	MA Jup DEF
92	51	16	4	11	TR Ere MID
2	6	1	17	11	MA Ere MID
203	84	35	3	11	TR Jup MID
18	7	3	49	11	MA Jup MID

Cluster 0 <-- TR Jup DEF

Cluster 1 <-- TR Jup MID

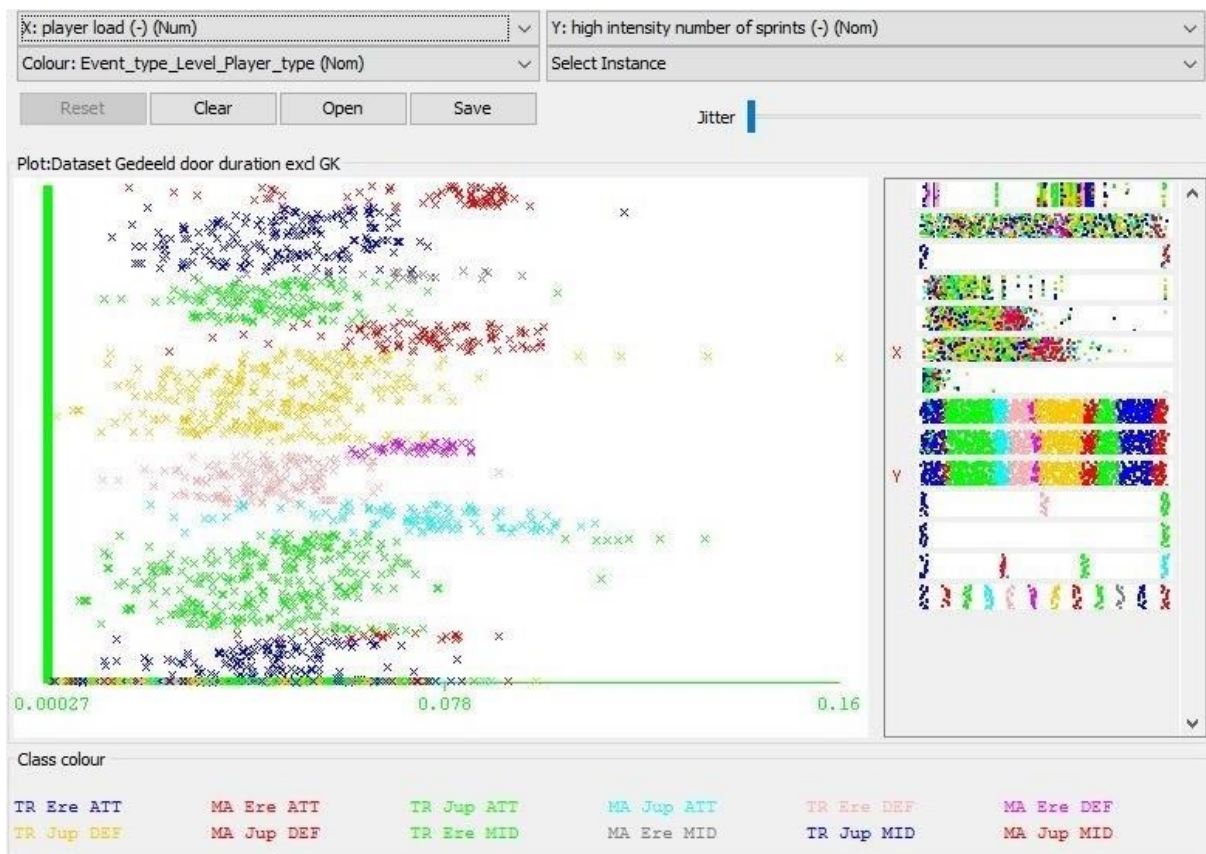
Cluster 2 <-- TR Jup ATT

Cluster 3 <-- MA Jup DEF

Incorrectly clustered instances : 1637.0 76.2104 %

Figuur 16: SimpleKMeans met betrekking tot alternatieve aanpak (gehele dataset, gelet op posities)

Nadat Weka zelfstandig naar clusters heeft gezocht, blijft wederom de mogelijkheid over om zelf visueel op zoek te gaan naar clusters. In dit geval blijken er weldegelijk clusters te vinden te zijn. In figuur 17 is te zien hoe deze clusters er visueel uitzien. Doordat er onderscheid gemaakt is middels verschillende kleuren, worden de clusters duidelijk zichtbaar. Weka herkent zelfstandig geen clusters wanneer de spelersbelasting en het aantal hoge intensiteitsprints tegenover elkaar staan, terwijl deze visueel wel zichtbaar zijn.



Figuur 17: plot met betrekking tot alternatieve aanpak (gehele dataset, gelet op posities)

Nu deze clusters alsnog gevonden zijn, kunnen er een aantal dingen voorspeld worden. Hierbij moet worden gericht op de spelersbelasting en het aantal sprints dat een speler getrokken heeft. Aan de hand hiervan is te zien of een speler een wedstrijd of een training heeft afgewerkt. Daarbij is duidelijk zichtbaar op welk niveau de desbetreffende wedstrijd of training afgewerkt is. De positie waar een speler speelt in het veld, is daarbij blijkbaar ook van invloed.

6. Evaluatie

De onderzoeksvraag

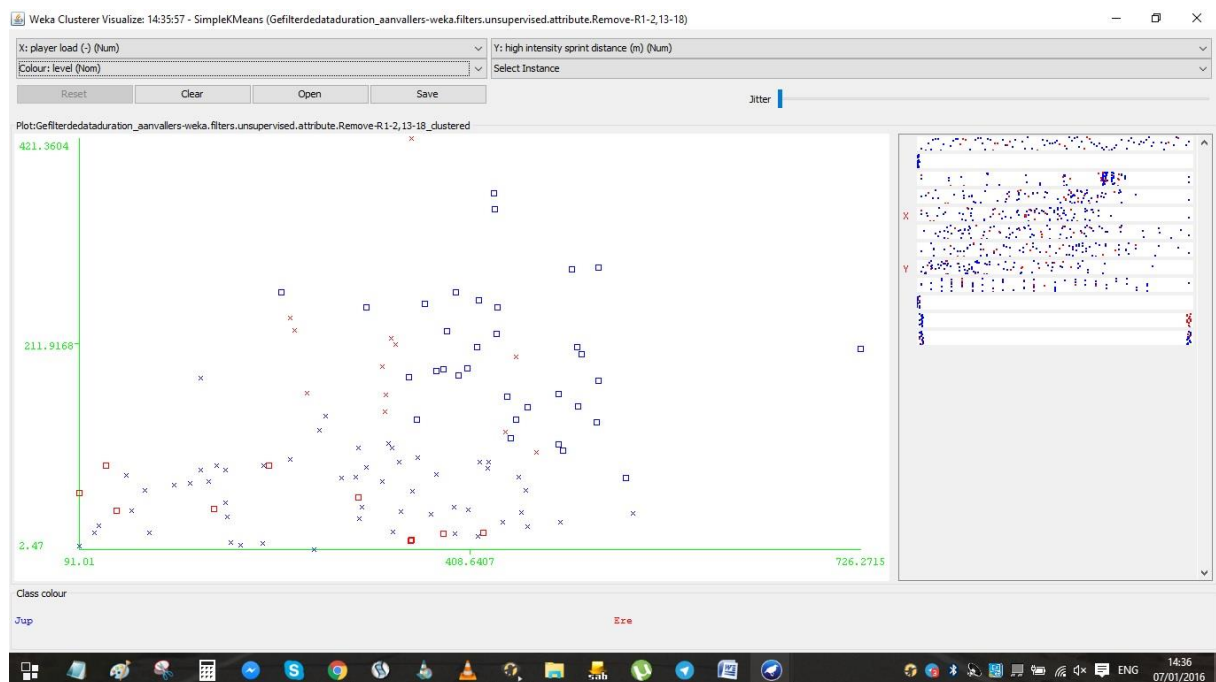
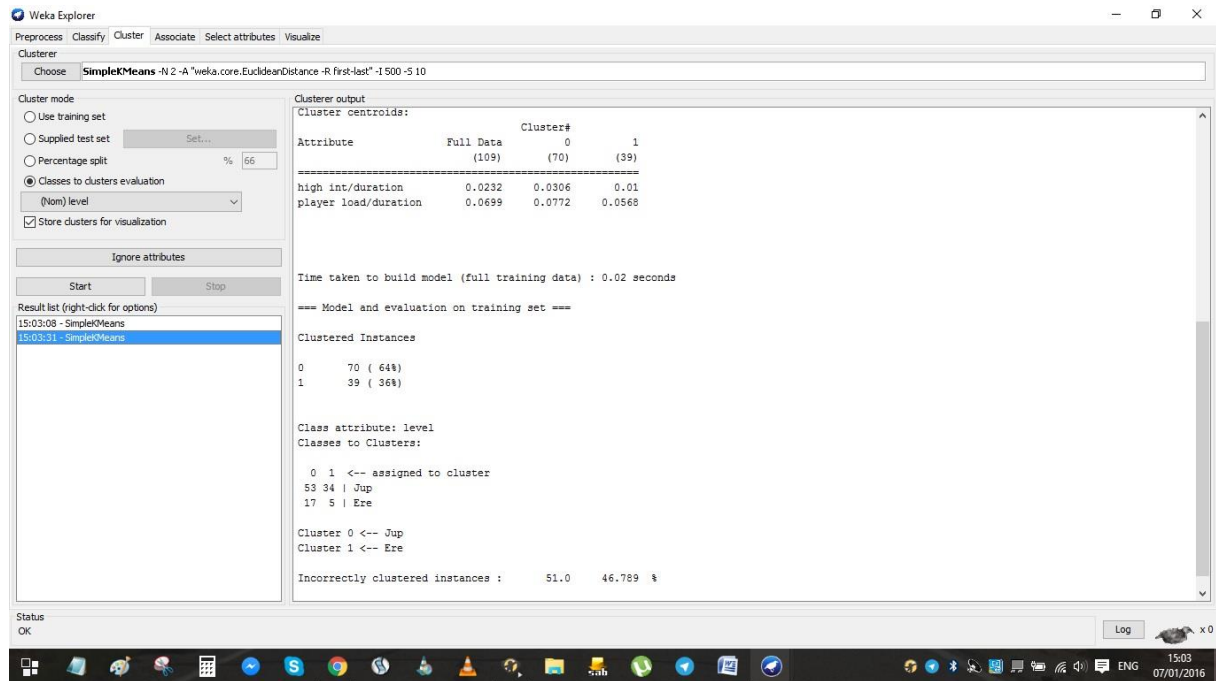
Welke clustering en welke classificaties zijn te vinden in de trackerdata van profvoetballers van twee niveaus (Eredivisie en Jupiler League), in de periode 28-07-2015 t/m 1-12-2015, waarbij gebruik wordt gemaakt van het softwarepakket Weka?

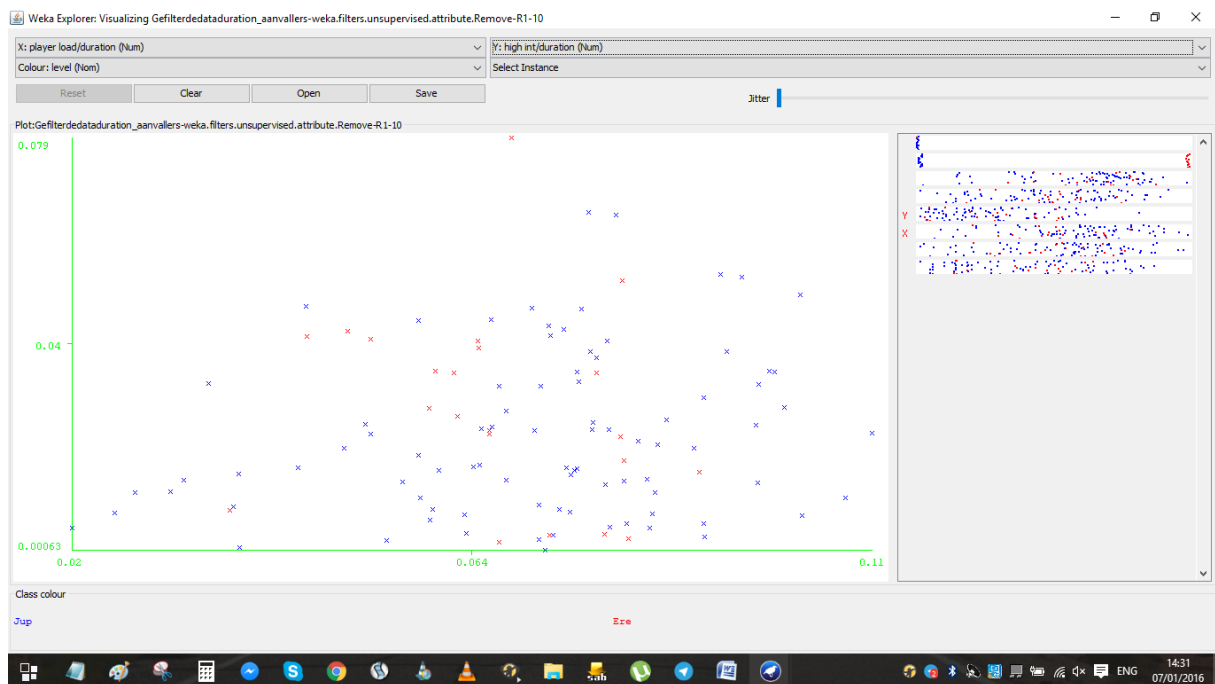
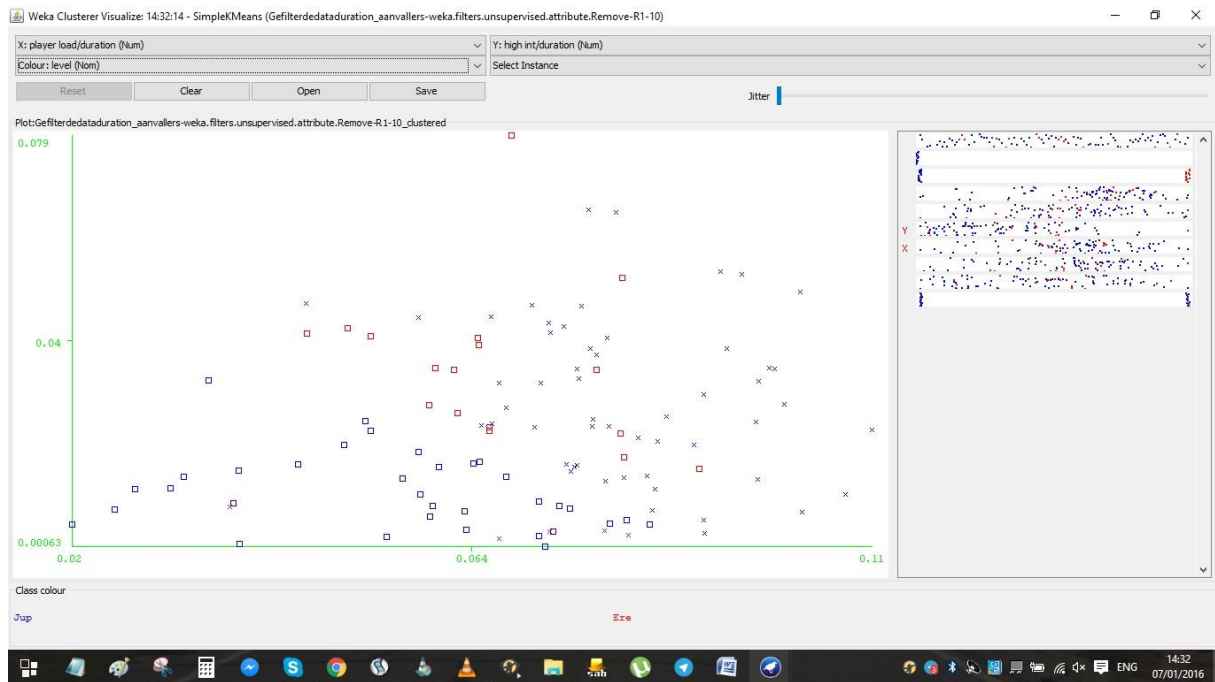
Door de verkregen dataset in verschillende stukken op te delen, hebben we wel kleine opvallendheden uit de data kunnen halen. Er kan worden geconcludeerd dat het mogelijk is om posities, niveau en trainingstype te kunnen voorspellen wanneer er verschillende voetbalklassen worden samengevoegd. Wel kan geconcludeerd worden dat de data onvoldoende rijk is om duidelijke clusters en classificaties te maken van verschillende soorten voetballers.

Bijlagen

Hoofdstuk 2

Aanvallers, diverse screenshots uit Weka





Middenvelders, diverse screenshots uit Weka

Weka Explorer

Preprocess Classify Cluster Associate Select attributes Visualize

Clusterer

Choose **SimpleKMeans** -N 2 -A "weka.core.EuclideanDistance" -R first-last -I 500 -S 10

Cluster mode

☐ Use training set

☐ Supplied test set Set...

☐ Percentage split % 66

☒ Classes to clusters evaluation

(Nom) level

☒ Store clusters for visualization

Ignore attributes

Start Stop

Result list (right-click for options)

15:06:52 - SimpleKMeans

Cluster output

Cluster Statistics:

Attribute	Full Data (90)	Cluster# 0 (46)	Cluster# 1 (44)
distance/duration	1.848	2.1219	1.5618
sprint/distance	0.1009	0.1425	0.0574
high int/duration	0.0186	0.0295	0.0071

Time taken to build model (full training data) : 0.01 seconds

=== Model and evaluation on training set ===

Clustered Instances

0 46 (51%)

1 44 (49%)

Class attribute: level

Classes to Clusters:

0 1 <-- assigned to cluster

33 33 | Jup

13 11 | Ere

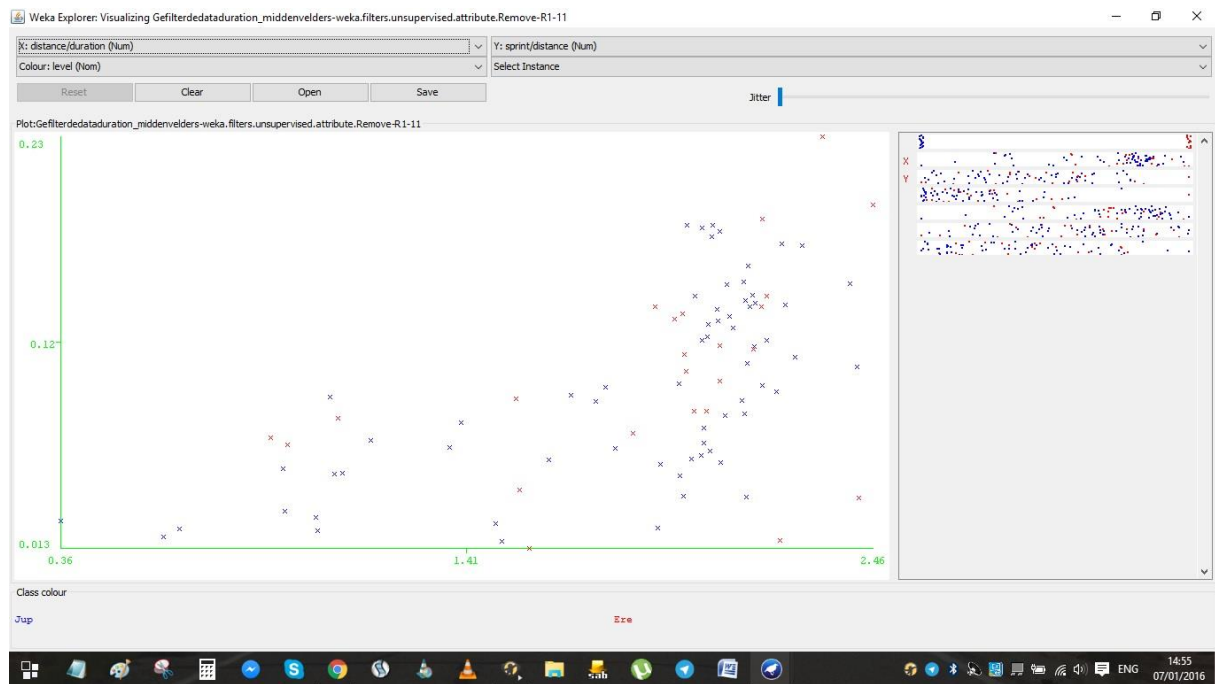
Cluster 0 <-- Ere

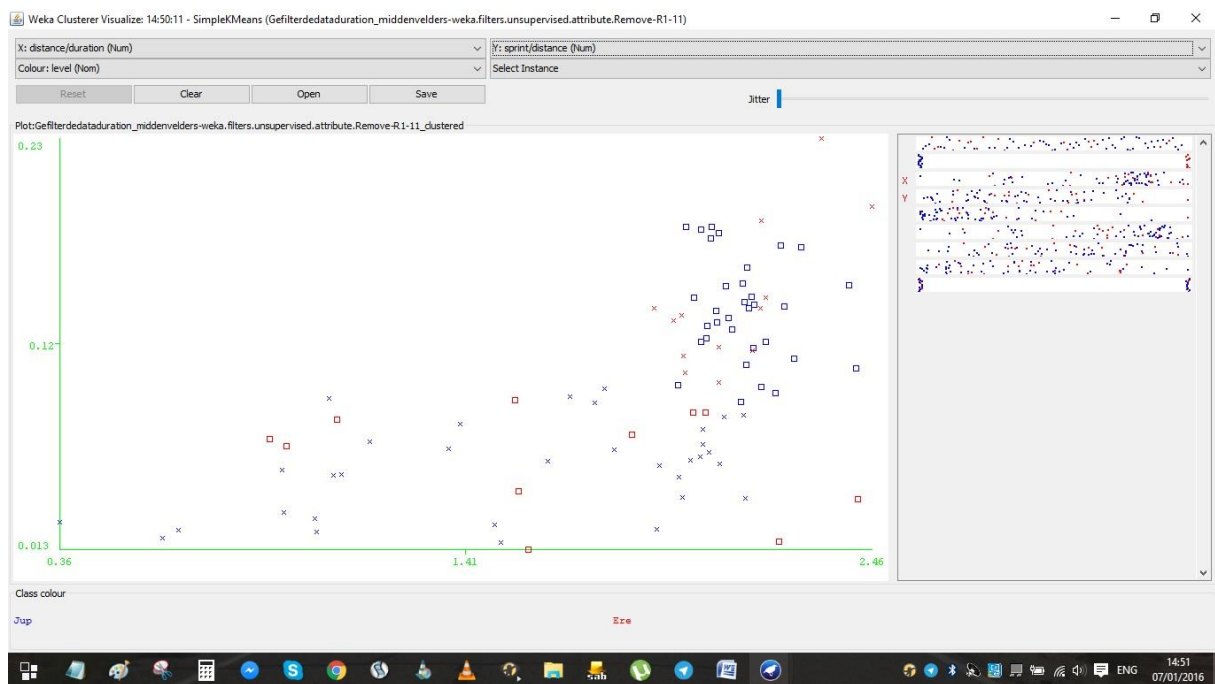
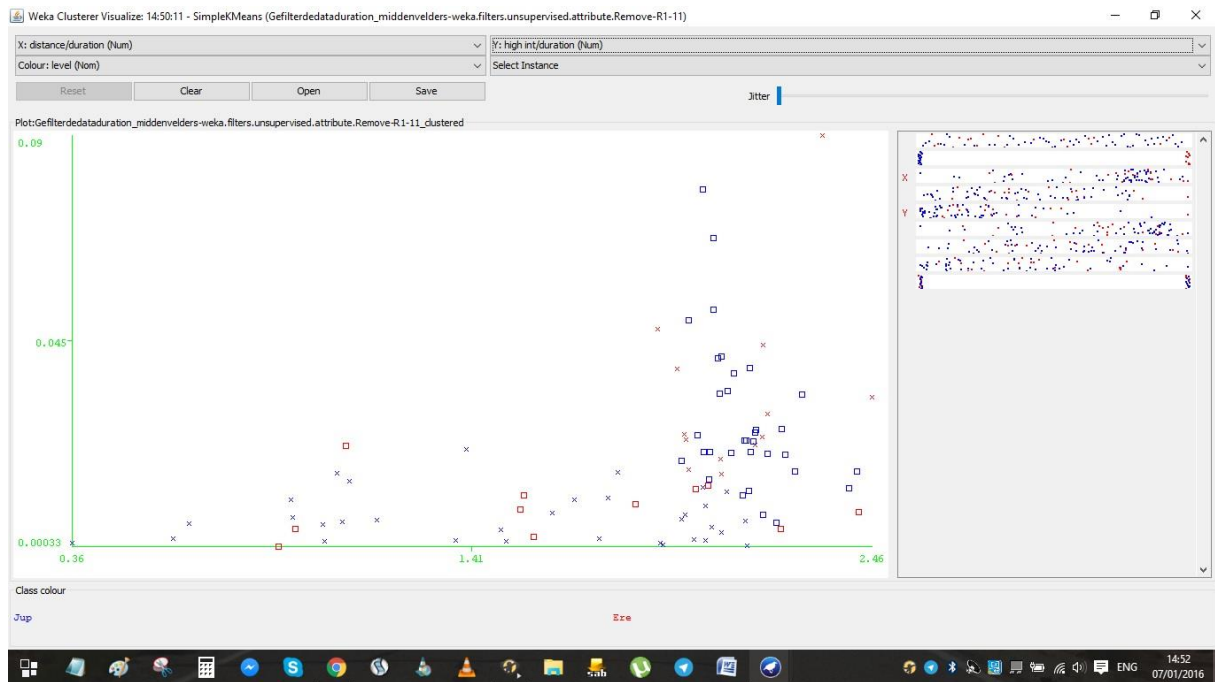
Cluster 1 <-- Jup

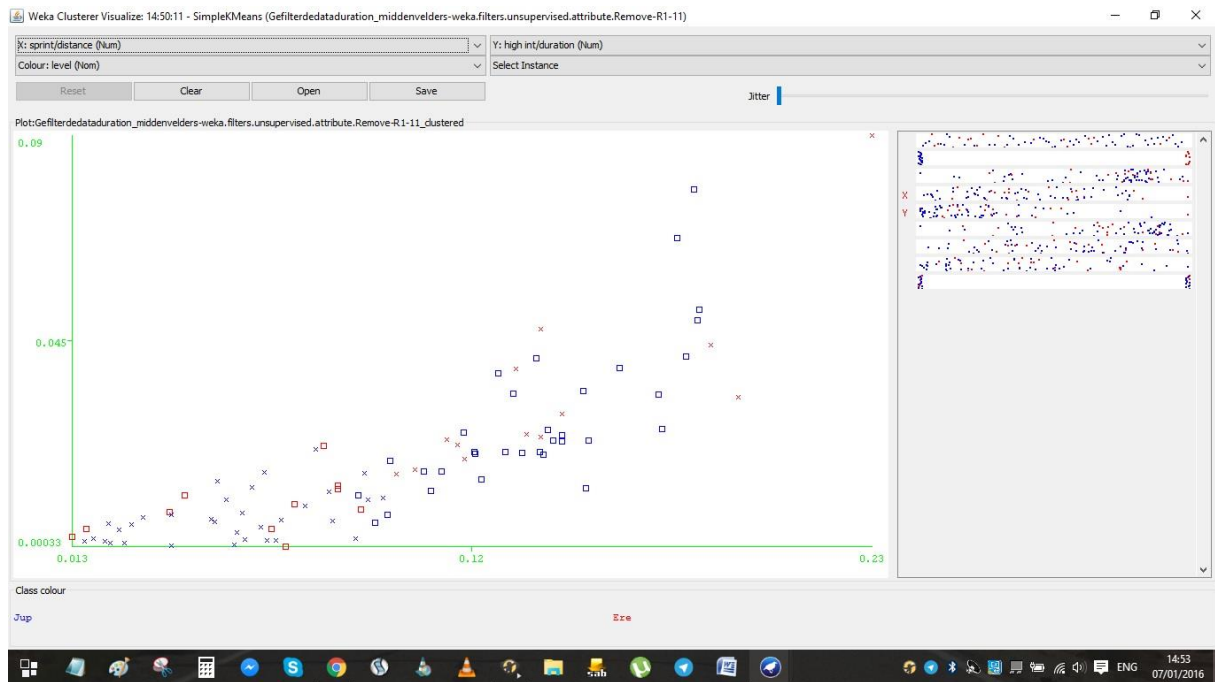
Incorrectly clustered instances : 44.0 48.8889 %

Status OK

Log







Verdedigers, diverse screenshots uit Weka

Weka Explorer

Preprocess Classify Cluster Associate Select attributes Visualize

Clusterer
 Choose **SimpleKMeans -N 2 -A "weka.core.EuclideanDistance -R first-last" -I 500 -S 10**

Cluster mode
☐ Use training set
☐ Supplied test set Set...
☐ Percentage split % 66
☒ Classes to clusters evaluation
 (Nom) level
☒ Store clusters for visualization

Ignore attributes

Start Stop

Result list (right-click for options)
 15:08:54 - SimpleKMeans
 15:09:21 - SimpleKMeans

Cluster output

Cluster centroids:

Attribute	Full Data (121)	Cluster# 0 (112)	Cluster# 1 (9)
distance/duration	1.8604	1.9182	1.1412

Time taken to build model (full training data) : 0 seconds

=== Model and evaluation on training set ===

Clustered Instances

Cluster	Instances	Percentage
0	112	(93%)
1	9	(7%)

Class attribute: level
 Classes to Clusters:

Cluster	Assigned to cluster
0	1
1	0

Cluster 0 <-- Jup
 Cluster 1 <-- Ere

Incorrectly clustered instances : 42.0 34.7107 %

Status OK Log

