Diagnostiseren Parkinson dysphonia

L T Stein

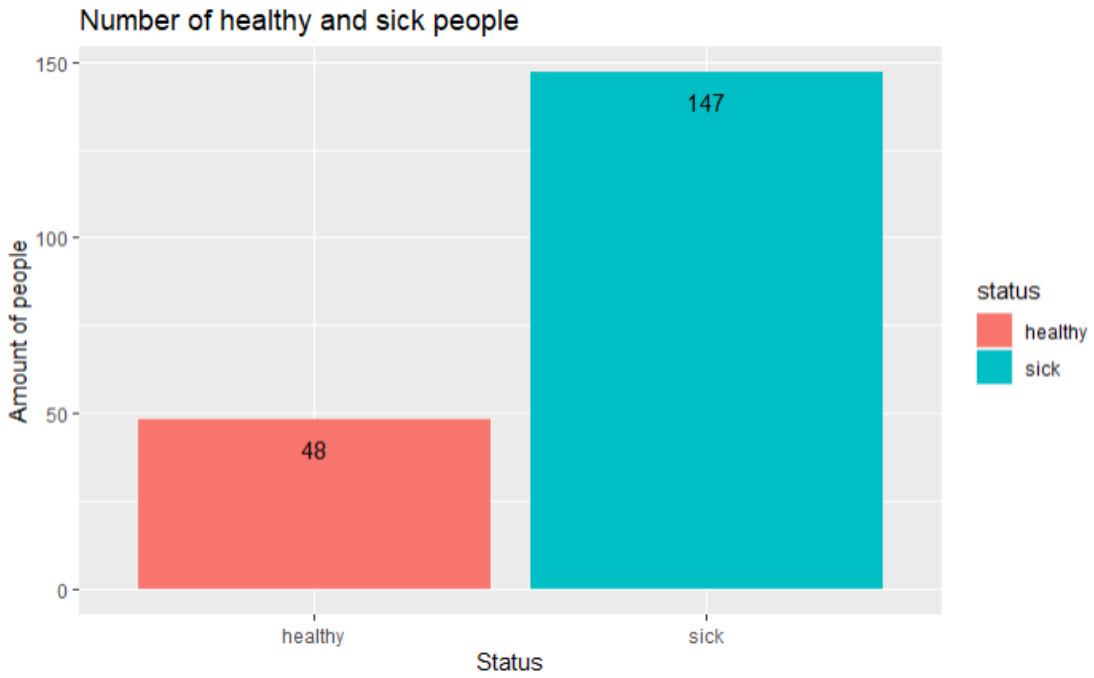
2023-10-05

Resultaten

Volledigheid

Het csv bestand was volledig en er waren geen NA waarden. Daarom hoefden er geen beslissingen genomen worden om waarden te verwijderen. Het bestand bestaat uit 24 kolommen en 195 rijen. Van deze kolommen zijn er 22 numeriek en twee nominaal, namelijk de namen van de patiënten en interessanter de statussen (levels) gezond (geen Parkinson) en ziek (Parkinson patient). Zes kolommen hebben eenheden, namelijk Hertz, procent, milliseconde en decibel; zoals eerder besproken in de inleiding.

De rijen zijn verdeeld in 48 gezonden instanties en 147 PD patiënten instanties. Deze zijn in een barplot hieronder te zien.

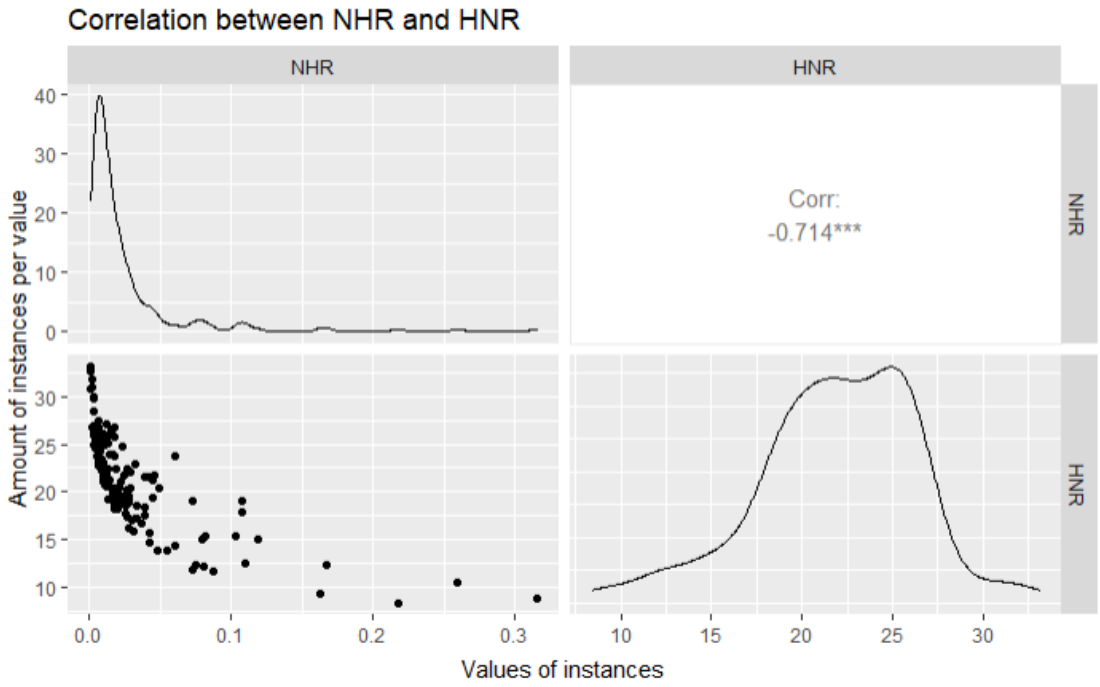


Figuur 1 Barplot dat de aantallen per levels laat zien. Op de x-as zijn de statussen gezond (healthy) en sick(PD patiënt) en op de y-as is de hoeveelheid uitgedrukt in mensen.

Zoals te zien in figuur 1 zijn er drie keer zo veel rijen van zieke mensen dan gezonde mensen in de dataset.

Attributen observeren

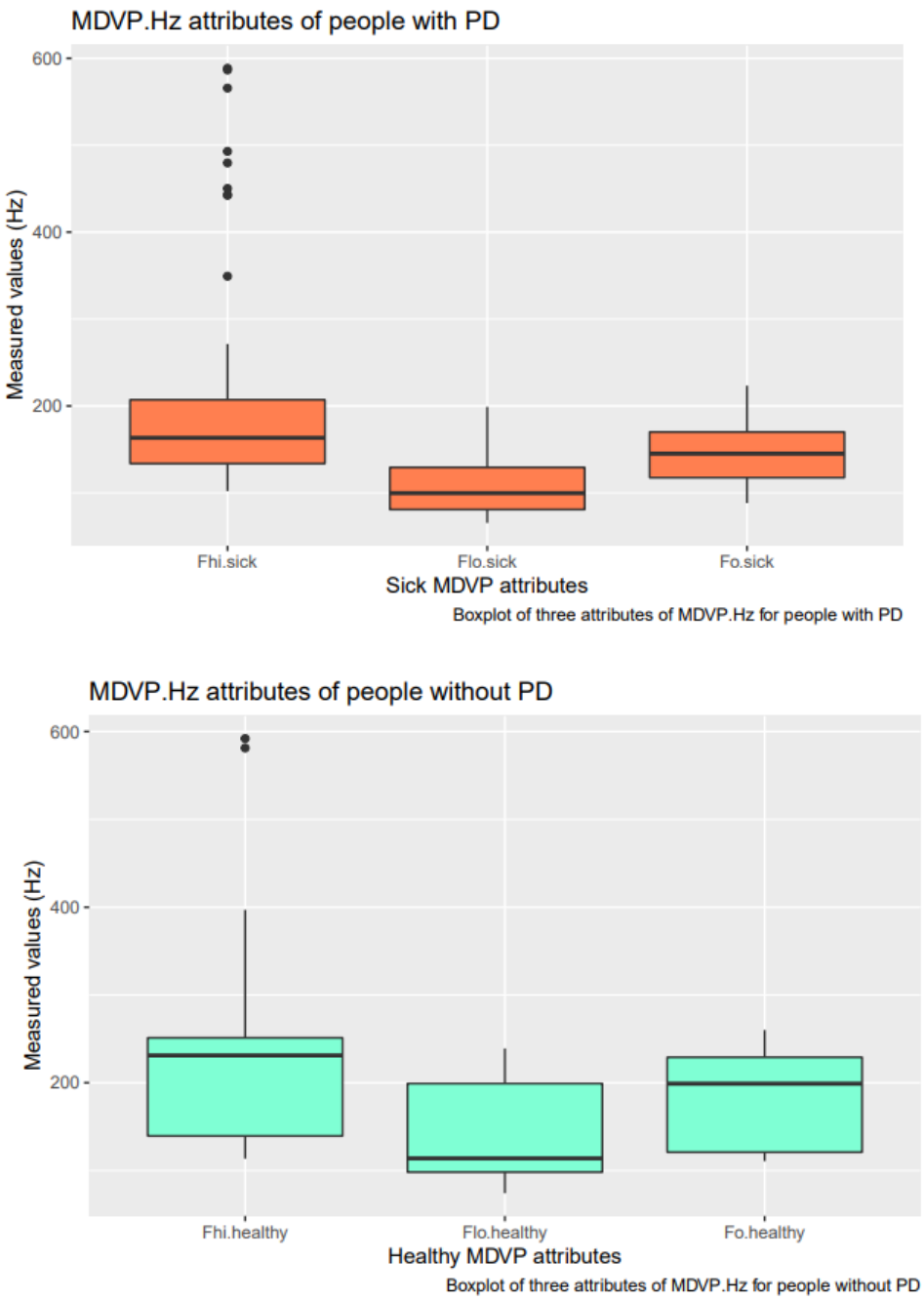
Uit de samenvatting van de data vielen attributen NHR (noise-to-harmonics ratio) en HNR (harmonics-to-noise ratio) op. Deze hebben namelijk waarde die bijna tegenover elkaar lijken te staan, deze zijn te zien in de bijlage bij het samenvattende tabel. Om na te gaan of er inderdaad een verband is tussen de twee wordt er een 'ggpairs' plot gemaakt. Deze plot methode maakt een matrix van de ontvangen attributen en kijkt naar hoe de waarden zijn opgebouwd en vergelijkt deze met elkaar. Dit kan ook wel het gedrag van de gemeten data worden genoemd en de vergelijking wordt uitgedrukt in een correlatie coëfficiënt. Deze ligt tussen 1 (positieve feedback in gedrag), naar nul (weinig tot geen feedback in gedrag) tot -1 (negatieve feedback in gedrag).



Figuur 2 GGpairs plot van attributen NHR en HNR. Op de x-as zijn de groottes van de gemeten waarden en op de y-as de hoeveelheid waarden per grootte. Links boven en rechts onder zijn deze hoeveelheden geplot. Links onder is de correlatie dat in een niet vloeiende boog van links boven naar rechts onder loopt. Rechts boven staat het correlatie coëfficient.

In figuur 2 is te zien dat de boog afloopt, dit geeft aan dat er een negatieve correlatie is. Het feit dat deze niet vloeiend loopt is terug te zien in het correlatie coëfficiënt van -0.714. Dus zijn de gemeten waarden niet exact omgekeerd van elkaar.

Verder zijn er boxplots gemaakt per level van de attributen MDVP.Fo.Hz., MDVP.Fhi.Hz. en MDVP.Flo.Hz. in eenheid Hertz (Hz).



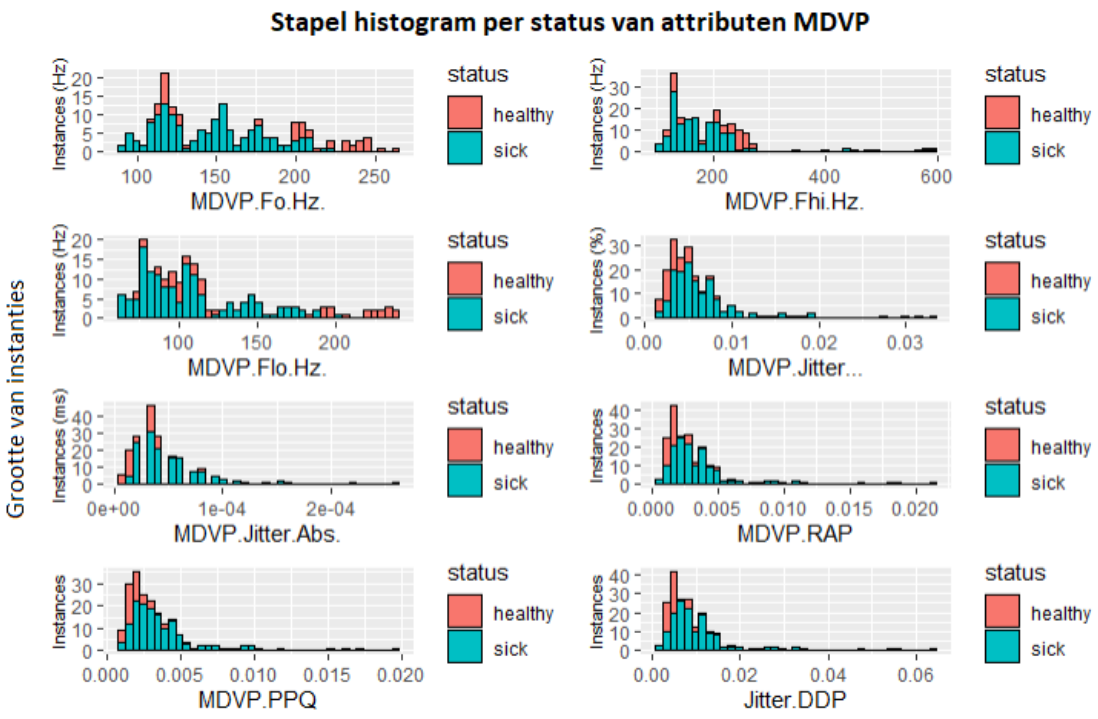
Figuur 3 Metingen van drie MDVP attributen in Hertz per status. Op de x-as staan de attributen en op de y-as de frequenties (Hz).

In figuur 3 hebben gezonde mensen hogere frequenties (F = 150>200) dan mensen met PD (F= 150<200). Uit de EDA werd dit nog sterker bevestigd met ANOVA toetsen die de volgende p-waarden gaven:

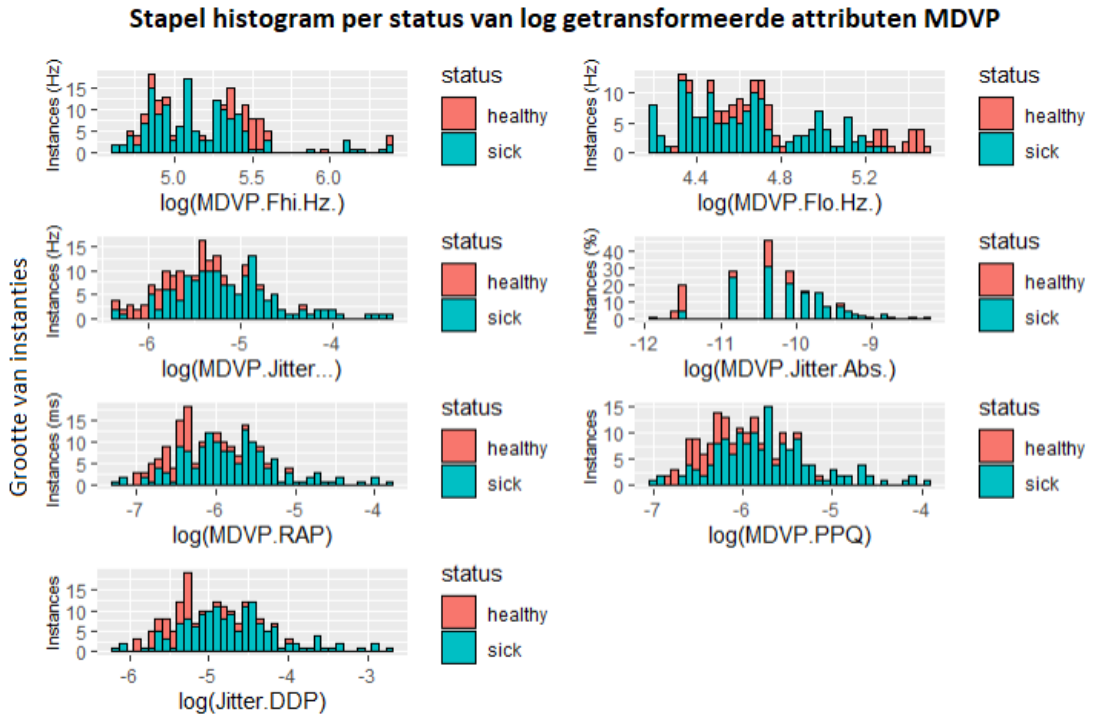
Fo.Hz = 3.121919e-08 Fhi.Hz = 0.02027567 Flo.Hz = 4.197004e-08

Attributes	P-values
Fo.Hz	3.121919e-08
Fhi.Hz	0.02027567
Flo.Hz	4.197004e

Uit het observeren van alle attributen als histogrammen werd het duidelijk dat veertien attributen niet normaal verdeeld waren. Daarom zijn deze met log10 getransformeerd en met succes vertoonden ze daarna wel normaal verdelingen. Als voorbeeld zijn worden de histogrammen van attributen MDVP getoond, omdat deze veel data naar links verdeeld hadden.



Verdeling van data per attribuut

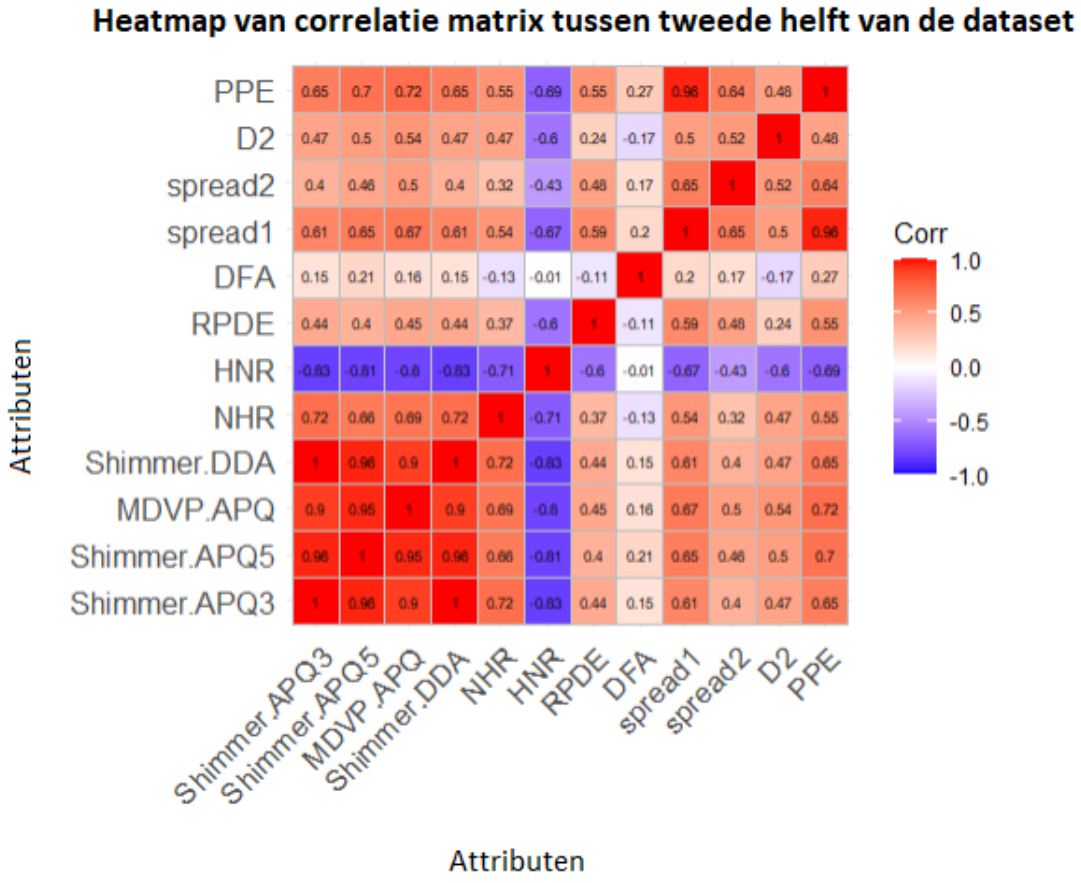


Verdeling van data per attribuut

Figuur 4 Stapel histogrammen van MDVP attributen voor en na log-transformatie. Op de x-as staat de verdeling per attribuut en op de y-as de grootte van de gemeten waarden. Rood is gezond en blauw is ziek.

In figuur 4 is te zien dat na log-tranformatie de data naar het midden verschuift.

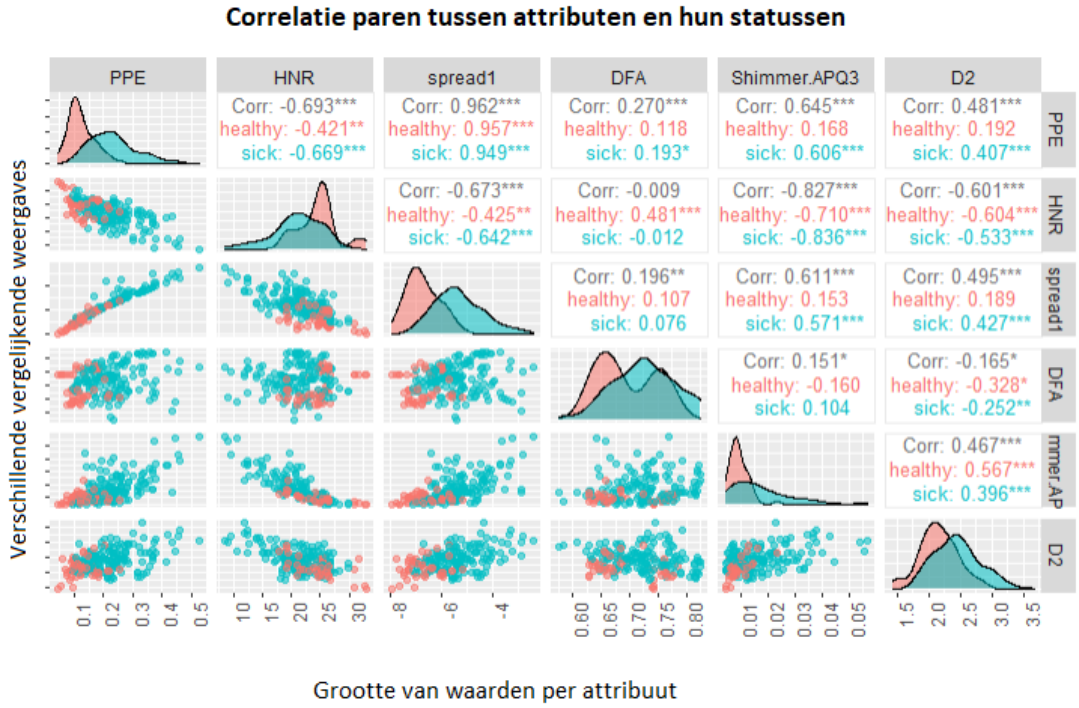
Vervolgens zijn er drie correlatie matrices gemaakt om de data op te delen en te vergelijken. In de onderstaande plot staat de tweede helft van alle numerieke attributen.



Figuur 5 Correlatie heatmap van tweede helft van de dataset. Op zowel de x-as als y-as staan attributen die hun eigen data representeren. De coëfficiënten worden verdeeld van paars (negatief één) naar wit (neutraal nul) tot rood (positief één).

In de heatmap is te zien hoe erg de attributen waarden op elkaar reageren qua gedrag. Links onder in zijn de Shimmer attributen met correlaties van bijna één, dus meetwaarden met erg vergelijkbare oploop. Attributen HNR en NHR zijn paars, deze bevestigen hun negatieve correlatie. Verder zijn er waarden rond nul te zien en heel veel tussen nul en één in.

Om meer beeld te krijgen van het gedrag tussen data dat correlatie aantoont kunnen er grotere GGpairs grafieken gemaakt worden waar de statussen zichtbaar zijn. In de volgende grafiek zijn attributen gekozen die of negatieve correlaties vertoonden of juist erg positief.



Figuur 6 Correlatie paar plot tussen verschillende attributen en de twee statussen. Op de x-as de grootte van hun waarden en op de y-as verschillende weergaves om verschillen of correlaties aan te tonen.

In figuur 6 zijn drie typen weergaven te zien. Links onder in een driehoek met puntgrafieken. Het gedrag van de correlaties is hieruit af te lezen. Bijvoorbeeld de vierde van de eerste kolom PPE gaat van links onder naar rechts boven. Die vertoond een positieve correlatie, hierin is ook een zeker onderscheidt tussen het merendeel van de statussen. Daar rechts naast is een grafiek met meer een wolk spreiding van links boven naar rechts onder. Deze vertoond negatieve correlatie.

Verder wordt links onder van rechts onder gescheiden door berg grafieken die de overlapping tussen statussen per attribuut weergeeft. In andere woorden hoe erg de waarden op elkaar lijken die gemeten zijn. In D2 zit er weinig verschil tussen de meetwaarden van de groepen, maar bij PPE lijkt juist een veel groter verschil.

Overigens maar zeker belangrijk bevat de driehoek van rechts boven in alle correlatie coëfficiënt waarden tussen paren van attributen en hun statussen. Spread1 en PPE heeft een erg positieve correlatie van 0.963 en de groepen verschillen maar met acht. Deze attributen reageren dus erg positief op elkaar. Daartegenover staan Shimmer.APQ3 en DFA met een correlatie vlakbij nul. Er is weinig reactie op elkaar en ook amper verschil tussen

de groepen. Tevens hebben D2 en HNR juist een negatieve correlatie, ze reageren juist van elkaar weg.

Discussie

Vanwege het feit dat de hoeveelheid zieke mensen in figuur 1 drie keer zo veel is als gezonde mensen, moet er rekening gehouden worden dat statistisch gezien meer toevallige data punten voorkomen en de observaties zullen daarom minder als hard bewijs gezien worden. Uit berekening uit de EDA komen er ongeveer tien vals positieven voor, door alpha 0.05 te vermenigvuldigen met het aantal rijen. Al hoewel het voor onderzoek en model training genoeg data is zou als vervolg onderzoek een grotere dataset gebruikt kunnen worden om zekerder te zijn van alle observaties.

In figuur 2 is er sprake van geen volledige negatieve correlatie; een verklaring kan zijn dat de formules gebruikt voor het meten van HNR en NHR niet evenredig omgekeerd zijn. Als voorbeeld 2log(8) = 3 en 2^3 = 8, de formules zijn evenredig omgekeerd verbonden aan elkaar.

In figuur 3 werd het duidelijk dat de eerste MDVP attributen met status “Gezond” significant hogere frequenties vertoonden dan die met status “PD patiënt”. Het feit dat deze opnames op zichzelf al zo een groot onderscheidt maken tussen de statussen, geeft de gedachte dat ze gebruikt kunnen worden als knopen in een model training om de data op te splitsen. OneR zou in dat geval interessant zijn om deze te testen, aangezien het algoritme de attribuut uitkiest die de data het beste splitst vergeleken met de andere attributen.

Figuur 4 laat zien dat log10 transformeren effectief de data normaliseerd. Het is interessant om deze getransformeerde data in een apart bestand te zetten, om te zien of de attributen met outliers (uitschieters) van voor het log-transformeren anders worden beoordeeld door het model dan een dataset met wel getransformeerde data. Een argument hiervoor is dat outliers meer verspreid zijn en daarom zal de splitwaarde tussen de statussen anders zijn, dan wanneer de data dichter bij elkaar is gebracht.

Verder gaf de heatmap in figuur 5 overzichtelijk weer hoe attributen op elkaar reageren. Deze correlaties kunnen positief, neutraal of negatief op elkaar reageren. Dit beeld werd het meest duidelijk met HNR en NHR, zodra de instantie van HNR toenemen worden de waarden van NHR lager. Dit beschrijft de negatieve correlatie die de attributen vertonen in een kruis patroon op de heatmap met elkaar, maar ook met andere attributen.

Alhoewel de heatmap naar verbanden kijkt in attributen, kan er niet uit worden gehaald of de attributen samen daadwerkelijk een onderscheidt kunnen maken tussen de statussen. Er kan alleen worden gespeculeerd dat bij een positieve correlatie en negatieve correlatie de waarden van de statussen zullen verschillen door positieve of negatieve feedback van waarden.

Aansluitend bij figuur 5 geeft figuur 6 extra informatie over het gedrag van data i.v.m. de correlatie coëfficiënten. Het werd duidelijk dat er patronen zijn die correlaties representeren, dus links onder naar rechts boven is positieve correlatie, links boven naar rechts onder is negatieve correlatie en een verspreide wolk geeft zwakkere correlatie aan. Attributen PPE en spread1 vertonen een mooie scheiding tussen de statussen, tegelijkertijd met een hoge correlatie waarde. De vraag is of deze vorm van scheiding een betere grens is tussen de statussen, dan een negatieve correlatie zoals HNR en PPE met een grotere punten wolk en dus zwakkere correlatie.

Omdat er meer zieke waarden zijn, is er tussen beide attribuut paren overlap tussen de clusters. Dit laat zien dat er geen volledige splitsing is. Een manier om tijdens model training meer onderscheidt te maken is door fouten zwaarder mee te tellen voor groep gezond. Door zwaardere kosten te rekenen zal het model voorzichtiger te werk gaan tussen de groepen voor een meer zekere uitkomst hoe goed de attributen samen werken.

Conclusie en reflectie

Door het observeren en testen van de parkinson data is het duidelijk geworden dat er attributen zijn die significante verschillen tussen de groepen aantonen. Dit zijn de eerste drie MDVP attributen, ook uit de correlatie matrix plus de paren matrix-grafiek zijn verschillen tussen de groepen te observeren. Het is duidelijk geworden wat de coëfficiënten betekenen en hoe die zich uiten in een punten grafiek. Ook deze vertonen onderscheidt, alleen is het niet duidelijk hoe sterk dat wordt uitgedrukt in werkelijke data. Ten slotte is het een feit dat er outliers zijn in de data en dat met log10-transformatie deze geminimaliseerd worden. Of deze impact hebben op het onderscheiden van de groepen kan worden onderzocht bij vervolgd onderzoek bij model training.

Om dit onderzoek voort te zetten zou er meer data gemeten moeten worden van mensen met Parkinson en gezonde mensen. Bovendien zou deze dataset ongeveer evenveel instanties per groep moeten hebben om statistische toevalligheden uit te sluiten, die nu meer waarschijnlijk zijn.