

Suggesties voor het uitvoeren van Learning Analytics

1 Voorkom overfitting

Bij het uitvoeren van statistische analyses moet voorkomen worden dat beslissingen alleen worden genomen op basis van de data (met name is dit het geval bij kleine steekproeven). Als een uiteindelijk model wordt gepast in een nieuwe steekproef zal het doorgaans minder goed passen. Om dit te voorkomen kan er worden gekruisvalideerd en/of ook inhoudelijke kennis verwerken in de analyse. Bij kruisvalidatie zou bijvoorbeeld gekeken kunnen worden of gevonden patronen in de ene week (bijv. een bepaalde groepsindeling) constant blijven in andere weken.

2 Gebruik gemiddelden of totalen tegen meetfouten

Samenhangend met punt 1, voor herhaalde metingen zoals toetsprestaties of inlogtijden is het vaak handig om te middelen over verschillende momenten zodat de analyses minder last hebben van attenuaties. Voor toetsresultaten is het zinvol om opgaven met gelijkende inhoud te aggregeren.

3 Specifieke keuzes bij mijn eigen analyses

a Pakketten

Ik heb gebruik gemaakt van R voor mijn analyses. Om de betrouwbaarheid en validiteit van de toetsen te bestuderen maak ik gebruik van het pakket psych. Voor exploratieve analyses is het mokken-pakket heel handig omdat daarin mooie visualisaties van itemkwaliteit zitten. Voor confirmatieve analyses raad ik het mirt-pakket aan omdat daarin factoranalyse op dichotome data kan worden gedaan. Voor de clusteranalyse heb ik gebruik gemaakt van het cluster-pakket en heb gekozen voor de clara-functie omdat deze ook data met missing values aankan.

b Aantal groepen

Bij clusteranalyse moet het aantal clusters (dwz voor de toepassing het aantal groepen) worden vastgesteld. Dat kan op twee manieren. Ten eerste vanuit de inhoud van het vak: De onderwijspraktijk geeft het aantal groepen of niveau's dat nodig is voor in de werkgroepen (bijvoorbeeld, er is ruimte voor drie groepen). Hierbij is het ook belangrijk om te bedenken of deze groepen even groot moeten zijn of juist niet. Ten tweede op basis van de analyse zelf. Binnen veel clustersoftware is het mogelijk om goodness of fit indices op te vragen die de geschiktheid van het aantal clusters evalueert. Een uitkomst zou bijvoorbeeld kunnen zijn dat een tweetal clusters het beste bij de data in kwestie past. Als er eenmaal een aantal clusters is gekozen zouden per cluster beschrijvende statistieken kunnen worden opgevraagd (gemiddelden en/of medianen, standaardafwijkingen) om de clustersamenstelling te duiden. Als deze statistieken geen zinvolle beschrijving geven is het wellicht verstandig om een ander clusteraantal te kiezen.