Inspectie van het Onderwijs
Ministerie van Onderwijs, Cultuur en Wetenschap

ALGORITMISCHE SIGNALERING RISICOSCHOLEN: TECHNISCH RAPPORT

Utrecht, juni 2020

Voorwoord

Dit rapport beschrijft een verkennend onderzoek naar de toepasbaarheid van voorspellingsmodellen voor risicobeoordeling in het onderwijstoezicht. Het onderzoek voerden we uit samen met onderzoekers van de Vrije Universiteit in onze academische werkplaats. Voor deze verkenning hebben we data gebruikt over het regulier basisonderwijs uit de periode 2011 – 2018.
Dit is een methodische verkenning en dus geen rapport waarin we tot nieuwe conclusies komen over risico's in het basisonderwijs. Evenmin gebruiken we de resultaten nu voor het toezicht.

Leerlingen, studenten en ouders moeten erop kunnen vertrouwen dat het onderwijs op scholen en instellingen goed is. De onderwijsbesturen zijn hiervoor verantwoordelijk. De Inspectie van het Onderwijs ziet toe op de basiskwaliteit. Eens in de vier jaar doen we een uitgebreid onderzoek bij elk bestuur en zijn scholen. Daarnaast analyseren we jaarlijks alle scholen op basis van de gegevens die we hebben. Als we daarbij risico's zien, kan dit leiden tot extra gesprekken of onderzoeken. Op deze manier belasten we besturen, scholen en instellingen alleen indien nodig en zetten we onze beschikbare capaciteit zo goed mogelijk in. Daarnaast willen we ook binnen onze cyclus van vierjaarlijkse onderzoeken de gespreksagenda voeden met informatie over risico's.

Voor het maken van risicoschattingen analyseren we al sinds lange tijd een veelheid van gegevens, zoals leerresultaten, leerlingprognoses en financiële data. De toename van data leidt tot behoefte aan nieuwe analysetechnieken. Eén van die technieken betreft het gebruik van voorspellingmodellen (zelflerende algoritmen) voor het voorspellen van (toekomstige) risico's.
Met deze rapportage laten we zien dat het gebruik van voorspellingsmodellen binnen de inspectie kan helpen bij het prioriteren van scholen. Dit blijkt met name iets toe te voegen bij kwaliteitsgebieden die minder 'telbaar' zijn. Kwaliteit van het onderwijsproces, kwaliteitszorg & ambitie en schoolklimaat zijn voorbeelden. Op die gebieden kunnen voorspellingmodellen menselijke beoordelaars ondersteunen bij de taak om veelzijdige en soms indirecte informatie goed te wegen.

Naast veel enthousiasme leven er binnen de samenleving natuurlijk ook zorgen over toepassing van algoritmen en big data. Deze zorgen gaan bijvoorbeeld over de omgang met gevoelige gegevens en over mogelijke vooringenomenheid (bias) van voorspellingsmodellen. Wat dan helpt, is zo transparant mogelijk zijn en het gesprek aangaan over de kansen en eventuele zorgen. Met de publicatie van onze bevindingen hopen we hiertoe een aanzet te geven. We zien vragen of reacties graag tegemoet.

dr (A.) Bert Bulder
Directeur directie Kennis, Inspectie van het Onderwijs

juni 2020

INHOUD

Samenvatting

Dit rapport geeft een uitgebreide, technische, beschrijving van het verkennende onderzoek dat binnen de Inspectie van het Onderwijs (IvhO) en in samenwerking met de Vrije Universiteit (VU) is uitgevoerd naar het gebruik van voorspellingsmodellen in het risicogericht toezichtsproces. De uitkomsten van dit rapport zijn bovendien samengevat in een (beknopte) hoofdrapportage. Daarnaast wordt op dit moment gewerkt aan een wetenschappelijk artikel waarin de mogelijkheid van het gebruik van voorspellingsmodellen bij algoritmisch bepaalde schoolonderzoeken verder wordt geëvalueerd.
In dit project is onderzocht of voorspellingsmodellen van toegevoegde waarde kunnen zijn voor de prioritering van risicoscholen voor nader onderzoek door inspecteurs of analisten. Voorspellingsmodellen kunnen op basis van historische voorbeelden patronen leren herkennen in grote hoeveelheden gegevens, die daarmee ook gebruikt kunnen worden om voorspellingen te doen over de toekomst. Binnen dit project zijn hiervoor gegevens over scholen in het reguliere basisonderwijs verzameld van binnen en buiten de inspectie. Deze gegevens zijn gekoppeld aan beoordelingen die gegeven zijn door inspecteurs gedurende de jaren 2011-2018. In het onderzoek zijn een aantal belangrijke vragen aan bod gekomen: 1) Welke fases in het risicogericht toezichtsproces van de IvhO lenen zich voor het gebruik van voorspellingsmodellen?; 2) Is de huidige kwaliteit en hoeveelheid van de beschikbare gegevens voldoende om zinvolle risicoschattingen te maken over scholen?; 3) Welke vormen van databewerkingen en welke modelvormen zijn het meest geschikt voor risicoschattingen in het toezicht op scholen?; 4) Wat is de te verwachten accuraatheid van voorspellingen bij implementatie als risicoproduct?; 5) Kunnen voorspellingsmodellen betere risicoschattingen genereren dan huidige risicoproducten? 6) Op welke manier kan zogenaamde 'bias' in voorspellingsmodellen inzichtelijk gemaakt worden?
Uit de verkenning is gebleken dat voor het trainen van voorspellingsmodellen op dit moment het beste gebruik gemaakt kan worden van de historische beoordelingen die zijn gegeven naar aanleiding van inspectiebezoeken aan scholen. In de verkenning is gebruik gemaakt van beoordelingen op het niveau van individuele standaarden (de verschillende kwaliteitsdomeinen), alsook op het niveau van de gegeven eindoordelen over een school als geheel. Vooral de beoordeling op het niveau van standaarden biedt de mogelijkheid om de risico's bij scholen te schatten voor specifieke kwaliteitsaspecten.
Een ander niveau van beoordeling dat ook is overwogen binnen dit project behelst de zogenaamde expertanalyse (deskresearch door analisten en inspecteurs, wat voorafgaat aan het besluit tot een schoolbezoek). Vergeleken met oordelen door inspecteurs tijdens een schoolbezoek sluit het voorspellen van de uitkomst van de expertanalyse nog directer aan bij het belangrijkste doel van risicoproducten (de prioritering van scholen voor de expertanalyse). Bovendien vergt deskresearch minder capaciteit dan schoolbezoeken, wat het aantal beschikbare beoordelingen zou kunnen verhogen. Voor het gebruik van de beoordelingen naar aanleiding van alleen deskresearch waren echter onvoldoende gestructureerde gegevens beschikbaar. In de toekomst is het daarom wenselijk om ook de documentatie van het deskresearch verder te standaardiseren. Dit zou het mogelijk maken om voorspellingsmodellen ook toe te passen bij het voorspellen van de uitkomst van deze vorm van risicoschatting.
Verder is gebleken dat de beschikbare gegevens die gebruikt konden worden als voorspeller (features) doorgaans van voldoende kwaliteit zijn, wat een belangrijke voorwaarde is voor het gebruik van algoritmen voor risicoschattingen. Deze methoden zijn namelijk afhankelijk van het gebruik van grote hoeveelheden

gegevens. Ook hierbij geldt echter dat er in de toekomst actief gezocht moet worden naar nieuwe databronnen en betere indicatoren. Bovendien zijn er verschillende vormen van datavoorbewerking en verschillende modelvormen onderzocht. Een succesvolle strategie behelst uitgebreide vormen van voorbewerking zoals het meenemen van informatie over meerdere jaren en het construeren van voorspellers die aansluiten op specifieke probleemscenario's zoals herkend door inspecteurs.
De voorspelkracht van de algoritmen kon vergeleken worden met die van de kennisanalyse, (het risicoproduct dat tot voor kort in gebruik was voor de prioritering van scholen), omdat de risicocategorieën van de kennisanalyse voor meerdere jaren in de dataset beschikbaar zijn. Voor de prestatiemonitor (het huidige risicoproduct) bleek deze vergelijking nog niet goed mogelijk, omdat het pas sinds recent in gebruik is. De vergelijking tussen het getrainde algoritme en de kennisanalyse heeft ten eerste laten zien dat ook de kennisanalyse al aanzienlijke voorspellende waarde heeft, met name waar het de beoordeling van de leerresultaten betreft (standaard OR1; domein leerresultaten). Deze beoordelingen zijn echter historisch sterk gebaseerd geweest op vrij harde gegevens zoals eindtoetsscores. Bovendien is de kennisanalyse niet alleen gebruikt als model van risicoschatting maar ook in belangrijke mate als model voor oordeelsvorming (een hoge risicoscore ging meestal gepaard met een onvoldoende oordeel). Dit maakt een objectieve vergelijking op dit kwaliteitsdomein lastig.
Voor de meer contextuele standaarden blijken in zijn algemeenheid minder harde gegevens beschikbaar (in dit project zijn beoordelingen op de domeinen Onderwijsproces [OP1, OP2, OP3]; Kwaliteitszorg en Ambitie [KA1, KA2, KA3] en Schoolklimaat [SK1] meegenomen, hierna 'zachte standaarden' genoemd). Hierdoor zijn risico's op deze kwaliteitsdomeinen in het geheel moeilijker te voorspellen dan voor een 'harde standaard' als OR1. Door het gebrek aan dergelijke harde gegevens bleek echter de toegevoegde waarde van voorspellingsmodellen *juist* op deze kwaliteitsdomeinen het sterkst. Voorspellingsmodellen lieten qua voorspelkracht juist voor deze standaarden een sterke verbetering zien ten opzichte van de kennisanalyse. Vanuit de gedachte dat goede of slechte onderwijsresultaten uiteindelijk slechts een gevolg moeten zijn van het presteren van scholen op kwaliteitsdomeinen zoals het onderwijsproces, de kwaliteitszorg en zaken zoals sociale veiligheid wordt in recente jaren het belang dat binnen de IvhO aan deze kwaliteitsdomeinen gegeven wordt in de beoordeling sterker. De toegevoegde waarde van voorspellingsmodellen op dit vlak kan daarom een belangrijke rol spelen bij de verdere ontwikkeling van datagedreven risicogericht toezicht. Een concreet resultaat dat dit project op dit aspect al heeft opgeleverd in 2019 is dat de prestatiemonitor PO is aangepast door onder andere het percentage ziekteverzuim onder leraren als indicator toe te voegen (deze kwam uit de analyses als sterk voorspellend naar voren).Voor de verdere ontwikkeling van risicogericht toezicht binnen de IvhO kunnen er op basis van deze verkenning een aantal aanbevelingen gedaan worden. Het lijkt wenselijk om de resultaten van dit onderzoek op korte termijn te betrekken bij de verdere ontwikkeling van de prestatiemonitor (in eerste instantie voor de sector Primair Onderwijs), en om ook te onderzoeken hoe het gebruik van deze technieken kan worden opgeschaald naar andere sectoren, en naar het toezicht op besturen in het algemeen.
Daarnaast zal het belangrijk zijn om rekening te houden met het feit dat er op dit moment een verschuiving plaatsvindt naar meer bestuursgericht toezicht. In deze vorm zal het aantal individuele schoolbezoeken waar een beoordeling uit voortkomt afnemen. Dat betekent dat er ook gezocht moet worden naar nieuwe alternatieven voor de officiële beoordelingen om modellen mee te trainen en te evalueren. Dit is belangrijk omdat de IvhO voor het vervullen van haar waarborgfunctie bij een bestuursgerichte aanpak in toenemende mate zal moeten leunen op adequate en tijdige risicomodellen. Een van de stappen daartoe zou kunnen zijn om de uitkomsten van het deskresearch (de expertanalyse) zodanig te standaardiseren dat

deze ook gebruikt kunnen worden om voorspellingsmodellen te trainen en evalueren, en om daarnaast op zoek te gaan naar nieuwe informatiebronnen (indicatoren) om risicoschattingen verder te verbeteren. Daarnaast zal het voor de ontwikkeling en evaluatie van risicoproducten in de toekomst ook belangrijk blijven om niet *alleen* scholen te onderzoeken met hoge risico's. Hoewel dit uitgangspunt op gespannen voet kan lijken te staan met de wens tot verdere verbetering van doelmatigheid op de korte termijn (een zo beperkt aantal scholen onderzoeken/bezoeken) zal een te sterke focus op bekende risico's leiden tot een vorm van tunnelvisie waarbij nieuwe of onbekende risico's op de lange termijn een tijd onzichtbaar kunnen blijven. Voor het evalueren van risicomodellen zijn immers voorbeelden van zowel goede als minder goede scholen nodig.
Als laatste is het belangrijk te benoemen dat dit project zich voornamelijk gericht heeft op de technische mogelijkheden van voorspellingsmodellen binnen het risicogericht toezicht van de IvhO. Het gebruik van voorspellingsmodellen roept echter ook belangrijke bestuurlijke vraagstukken op. In dit stuk wordt kort ingegaan op mogelijke vormen van bias (vooringenomenheid) van voorspellingsmodellen. Het is echter raadzaam om bij het toekomstige gebruik van voorspellingsmodellen verder in te gaan op thema's zoals vooringenomenheid, maatschappelijke draagvlak, en de juridische basis voor het gebruik van algoritmen.

1 Inleiding

Dit is de technische rapportage voor het verkennende onderzoek naar het gebruik van voorspellingsmodellen voor de prioritering van risicoscholen t.b.v. nader onderzoek door analisten of inspecteurs. In dit rapport staat een uitgebreide beschrijving van de gekozen methodieken, de resultaten en de daaruit voortvloeiende adviezen voor het verdere gebruik van voorspellingsmodellen voor risicosignalering binnen de IvhO.

1.1 Achtergrond en opdracht

Tot de kerntaken van de Inspectie van het Onderwijs (IvhO) behoort het bewaken van de kwaliteit van het Nederlands onderwijs. De beperkte middelen en het diffuse onderwijslandschap in Nederland maken dit tot een uitdagende taak. Met toezicht dat met name gericht is op risicovolle scholen kunnen de beschikbare middelen mogelijk efficiënter worden benut en daarmee een uitkomst bieden. Bovendien bestaat de wens om onnodige toezichtlast voor scholen en besturen te vermijden door de proportionaliteit van het toezicht te vergroten. Het toezicht van de IvhO is om die reden al ruim 10 jaar risicogericht. De huidige risicoschattingen zijn echter gebaseerd op een beperkte hoeveelheid indicatoren (zoals scores op de eindtoetsen) en vooraf vastgestelde grenswaarden, waarbij een beperkte statistische evaluatie van het gebruikte model is toegepast. Bovendien kan dergelijk deterministisch risicogericht toezicht leiden tot blinde vlekken. Vooral nieuwe of voorheen onbekende risico's zullen mogelijk niet —of te laat— herkend worden wanneer inspecteurs zich er niet van bewust zijn. Dit pleit voor het onderzoeken van methoden die goed in staat zijn om op objectieve wijze relaties te vinden tussen beschikbare gegevens en toezichtsbevindingen.
Het project Algoritmische Signalering Risicoscholen is geïnitieerd als verkenning om te bepalen of voorspellingsmodellen kunnen helpen om tot betere risicoschattingen te komen. Binnen dit project is daarom de toepasbaarheid van voorspellingsmodellen onderzocht. Bovendien is gekeken of deze modellen naar verwachting betere risicoschattingen kunnen maken dan de recent gebruikte risicoproducten van de IvhO zoals de kennisanalyse en de prestatiemonitor. Ten slotte heeft dit onderzoek enkele belangrijke vraagstukken rond algoritmische risicosignalering binnen de IvhO proberen te duiden.
Er is in een vroeg stadium besloten om de verkenning uit te voeren voor het regulier basisonderwijs. Dit is vooral besloten omdat deze onderwijssector een grote en relatief homogene groep objecten van toezicht behelst, wat het gebruik van voorspellingsmodellen sterk vergemakkelijkt. Het gebruik van voorspellingsmodellen is namelijk sterk afhankelijk van de beschikbaarheid van grote hoeveelheden data (historische voorbeelden van 'goede' en 'slechte' scholen). Om dezelfde reden zijn in deze verkenning het speciaal basisonderwijs en speciaal onderwijs buiten beschouwing gelaten.

1.2 Projectorganisatie

Dit project behelst een samenwerking tussen de Vrije Universiteit (VU) en de IvhO. Daarnaast hebben we voor dit project verschillende interviews gehouden en zijn er bijeenkomsten georganiseerd met experts binnen en buiten de IvhO.
Begin 2019 hebben er interviews plaatsgevonden met inspecteurs en analisten van de sectors Primair onderwijs (PO); Voortgezet Onderwijs (VO); Middelbaar Beroeps Onderwijs (MBO); Hoger Onderwijs (HO) en Speciaal Onderwijs (SO). Daarnaast is er een gezamenlijke bijeenkomst gehouden met de betrokken experts. Deze bijeenkomsten hadden als doel om inzicht te verkrijgen in de huidige werkwijze rond

risicogericht toezicht binnen de verschillende sectoren; om suggesties van experts te verzamelen voor vernieuwing in het datagedreven risicogericht toezicht; om ideeën op te doen voor mogelijke risicoindicatoren; en om de verschillende sectoren te informeren over het project.
Daarnaast hebben er gedurende 2019 verschillende bijeenkomsten plaatsgevonden met datascientists en onderzoekers werkzaam bij andere overheidsorganisaties in Nederland zoals het ministerie van Sociale Zaken en Werkgelegenheid; de Inspectie van Gezondheidszorg en Jeugd; de Inspectie van Leefomgeving en Transport; de Dienst Uitvoering Onderwijs (DUO); en enkele anderen. Omdat deze organisaties in verschillende mate ook bezig zijn met het onderzoeken en ontwikkelen van voorspellingsmodellen voor het toezicht waren deze bijeenkomsten vooral gericht op het uitwisselen van technische en theoretische kennis over het gebruik van algoritmen. Ten slotte zijn er ook meerdere gesprekken gevoerd met Ofsted, de Engelse onderwijsinspectie. Ofsted gebruikt al sinds enkele jaren een risicodetectiemodel gebaseerd op het gebruik van voorspellingsmodellen. Uit deze contacten is verder duidelijk geworden dat het onderzoek naar voorspellingsmodellen binnen overheden een vlucht neemt. Dit lijkt daarmee een goed moment om te onderzoeken of de IvhO bij deze ontwikkeling aan moet sluiten.
Gedurende het project zijn er om de drie maanden bijeenkomsten gehouden met de zogenaamde adviesgroep van het project, waaraan onder andere de leden van de projectgroep, een afdelingshoofd van de directie kennis, een strategisch inspecteur, en onderzoekers van de VU deelnamen.

1.3 Leeswijzer

De objecten van toezicht in het primair basisonderwijs zijn zogenaamde clusters. Wanneer in dit document gesproken wordt over scholen worden clusters bedoeld.
Hoofdstuk 2 van het document schetst in het kort het gebruik van algoritmen voor risicosignalering bij andere overheidsorganen.
Hoofdstuk 3 beschrijft de belangrijkste karakteristieken van de dataset. Daarbij gaat het vooral over de fase in het toezicht-proces waarin we risico's willen kunnen voorspellen en de features (indicatoren) die gebruikt kunnen worden om voorspellingen te doen. Hierbij wordt vooral ingegaan op de aspecten die de dataset uitdagend maken bij het gebruik voor risicomodellen.
Hoofdstuk 4 beschrijft de fase van het modelleren en optimaliseren van de voorspelkracht van de modellen
Hoofdstuk 5 beschrijft de resultaten door te kijken naar de verdelingen van toegekende risicoscores en maakt een vergelijking tussen voorspellingen op basis van algoritmen en voorspellingen op basis van de kennisanalyse en, voor zover mogelijk, de prestatiemonitor.
Hoofdstuk 6 geeft een korte verhandeling over het begrip bias –modelmatige vooringenomenheid- in de context van de gebruikte dataset.
Ten slotte wordt in Hoofdstuk 7 afgesloten met conclusies en de belangrijkste adviezen op basis van het verkennende onderzoek.

2 Algoritmen voor risicosignalering

2.1 Voorspellingsmodellen

Het werk van de inspectie wordt op veel verschillende manieren ondersteund door het gebruik van geautomatiseerde processen. Bij de meeste geautomatiseerde processen voert een programma de instructies uit die door een persoon expliciet zijn ingevoerd. Hierbij is het belangrijk dat die persoon bekend is met de verschillende mogelijke invoerwaarden en de bijbehorende gewenste uitkomst (denk bijvoorbeeld aan een waarschuwing wanneer de resultaten van een school een grenswaarde overschrijden). Soms komt het echter voor dat die relaties niet bekend zijn, of dat ze te complex zijn om handmatig te definiëren. Vanuit de Kunstmatige Intelligentie[1] zijn er verschillende methoden ontwikkeld om met dergelijke vraagstukken om te gaan. Een belangrijke methode daarvan bestaat uit zogenaamd 'gecontroleerd machinaal leren' (supervised machine learning; ook bekend als 'zelflerende algoritmen'). In dit stuk gebruiken we de term *voorspellingsmodellen* voor de algoritmen die uit deze methode voortkomen.

Voorspellingsmodellen worden getraind op basis van historische gegevens en zoeken op geautomatiseerde wijze naar statistische relaties tussen de invoergegevens (denk in dit project aan voorspellers zoals eindtoetsresultaten of aantal schorsingen van leerlingen) en de bijbehorende uitkomsten (voorbeelden van 'voldoende' en 'onvoldoende' beoordelingen). Deze relaties worden als parameters vastgelegd in het model. Vervolgens kan dit model in combinatie met nieuwe invoergegevens ook *voorspellingen* doen voor tot dan toe nog ongeziene uitkomsten. Voor de IvhO zouden dat voorspellingen kunnen zijn voor scholen die nog niet zijn bezocht. De belangrijkste bijdrage van voorspellingsmodellen aan risicogericht toezicht is dus dat het een methode biedt om op statistisch gefundeerde wijze tot een prioritering van risicoscholen te komen die zo optimaal mogelijk aansluit bij de manier van beoordelen door inspecteurs in het verleden.

Het zoeken naar statistische relaties kan op verschillende manieren plaatsvinden. Hierbij kan men denken aan methoden zoals lineaire regressie, beslisbomen en zeer uiteenlopende vormen van neurale netwerken. Hoewel deze methoden zelflerend zijn in de zin dat ze zelf parameters vaststellen zijn ze volledig afhankelijk van de gegevens en de probleemstelling zoals ze door mensen worden aangeboden. Het is belangrijk om te benadrukken dat deze methoden voorspellingen doen op basis van historische beoordelingen van inspecteurs en/of analisten. Daarmee zullen deze algoritmen dus in principe ook historische *manieren* van beoordelen toepassen in de voorspellingen voor de toekomst. Hoewel dit de kracht van voorspellingsmodellen is, kan het ook een probleem vormen wanneer er de wens bestaat om de manier van beoordelen juist te veranderen. Een voorbeeld van een dergelijke verandering is de recente wens binnen de IvhO om in de beoordeling naar een breder palet aan kwaliteitsdomeinen te gaan kijken dan alleen onderwijsresultaten. Een model dat echter getraind is op data uit een periode waarin eindoordelen voornamelijk gebaseerd werden op eindtoetsresultaten zal deze wijze van beoordeling ook op de toekomst projecteren. Om veranderingen in werkwijze in te passen is het dus belangrijk om op gerichte wijze van de voorspellingen af te wijken.

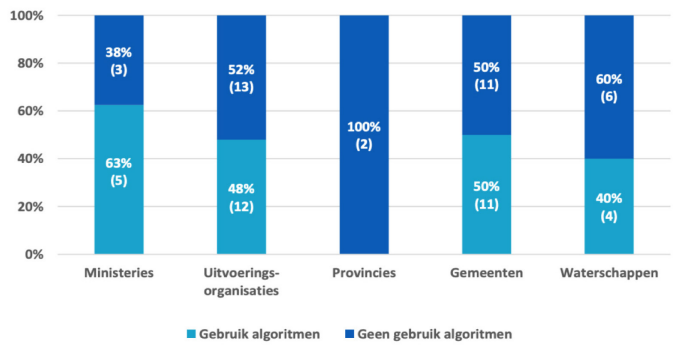
2.2 Gebruik van algoritmen bij overheden

In het afgelopen decennium zijn steeds meer bedrijven en overheden gebruik gaan maken van voorspellingsmodellen. Deze aanpak heeft tot grote ontwikkelingen

1 Kunstmatige Intelligentie behelst een breed vakgebied dat zich o.a. bezighoudt met de ontwikkeling van complexe computersystemen die kunnen leren; een relatief groot probleemoplossend vermogen hebben; en die cognitieve eigenschappen vertonen/simuleren die voorheen vooral geacht werden aan mensen voorbehouden te zijn.

geleid in het gebruik van technieken zoals als automatische beeld- en spraakherkenning binnen die organisaties, en wordt op dit moment ook al toegepast voor risicodetectie binnen overheidsorganen zoals de politie[2] en de belastingdienst[3]. Een recent verkennend onderzoek van het Centraal Bureau voor de Statistiek (CBS)[4] laat zien dat ongeveer de helft van de responderende organisaties bewust gebruik maakt van algoritmen in enige vorm (Figuur 2.1). Hierbij is het overigens belangrijk om onderscheid te maken tussen zogenaamde rule-based (beslisregel) algoritmen en zogenaamde case-based algoritmen (het type voorspellingmodellen zoals ook in dit project onderzocht). Beide zijn namelijk meegenomen in het desbetreffende onderzoek. Nieuwe beschikbare technieken gaan vooral over de tweede categorie (voorspellingsmodellen). Binnen de IvhO worden expliciet vastgelegde beslisregels immers ook al in verschillende vormen gebruikt, bijvoorbeeld in de kennisanalyse.

Figuur 2.1



Gebruik van algoritmen onder responderende overheidsorganisaties naar type organisatie. Figuur overgenomen uit het rapport "Verkennend onderzoek naar het gebruik van algoritmen binnen overheidsorganisaties", bron: CBS[4].

Van de respondenten in het onderzoek van het CBS die aangaven algoritmen te gebruiken, gebruikte 16% alleen beslisregel algoritmen, 37% alleen case-based algoritmen en 47% beide. Het gebruik van voorspellingsmodellen blijkt daarmee inmiddels een veelvoorkomende praktijk binnen overheden. Dit lijkt daarmee het belang te bevestigen voor de IvhO om te onderzoeken of er ook binnen toezicht op het onderwijs toegevoegde waarde bestaat voor het gebruik van deze methodiek. Verder heeft de onderwijsinspectie in Engeland (Ofsted) de afgelopen jaren Voorspellingsmodellen toegepast voor risicodetectie bij scholen[5]. Uit dit project is gebleken dat het doelmatig gebruik van voorspellingsmodellen ook sterk af hangt van het draagvlak, zowel binnen de inspectie als ook bij schoolbesturen en binnen de samenleving als geheel.

---

[2] 'Criminaliteits Anticipatie Systeem verder uitgerold bij Nationale Politie', via: https://www.politie.nl/nieuws/2017/mei/15/05-cas.html (laatst geraadpleegd 15 januari 2019).

[3] WRR, p53. via https://www.wrr.nl/publicaties/rapporten/2016/04/28/big-data-in-een-vrije-en-veilige-samenleving (laatst geraadpleegd 15 januari 2019).

[4] Verkennend onderzoek naar het gebruik van algoritmen binnen overheidsorganisaties: https://www.cbs.nl/nl-nl/maatwerk/2018/48/gebruik-van-algoritmen-door-overheidsorganisaties

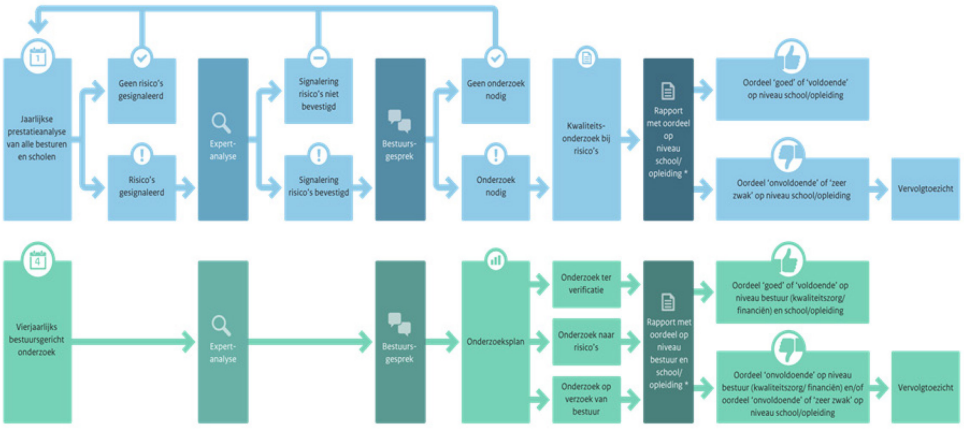
[5] https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/737583/Methodology_note_risk_assessment_of_good_and_outstanding_maintained_schools_and_academies_030918.pdf

3 Databronnen

3.1 Toezicht op scholen

Het toezicht op scholen in het Nederlandse basisonderwijs verloopt in belangrijke mate volgens twee lijnen. De eerste lijn behelst de zogenaamde jaarlijkse monitoring en richt zich op het identificeren van scholen die beneden de ondergrens van de vereiste kwaliteit zijn geraakt of dreigen te raken (de bovenste –blauwe- lijn in Figuur 3.1). Binnen deze lijn bestaat het toezicht uit verschillende fases. De eerste fase behelst de jaarlijkse prestatieanalyse. Hierbij wordt aan alle scholen in het basisonderwijs een risicoscore toegekend waarmee scholen geprioriteerd worden voor verder onderzoek. Het is in deze fase van het risicogerichte toezicht dat het gebruik van voorspellingsmodellen een belangrijke invloed kan hebben (d.m.v. een betere prioritering van scholen). Scholen die als relatief risicovol uit deze eerste analyse komen, worden (na intern overleg) onderworpen aan de zogenaamde expertanalyse. De expertanalyse behelst vooral deskresearch, uitgevoerd door analisten in samenwerking met inspecteurs. Als deze expertanalyse de risico's bevestigt, kan besloten worden tot een bestuursgesprek. Als uit dit bestuursgesprek blijkt dat een onderzoek op de school nodig is volgt een kwaliteitsonderzoek, waarbij inspecteurs tijdens een bezoek aan de school beoordelingen geven met betrekking tot een of meer kwaliteitsaspecten uit het onderzoekskader. Hierbij is het belangrijk om te vermelden dat inspecteurs een school zelden op álle kwaliteitsaspecten beoordelen, maar vaak slechts op een select (vooraf besloten) aantal kwaliteitsdomeinen. Deze kwaliteitsdomeinen en de bijbehorende kwaliteitseisen zijn beschreven in de zogenaamde standaarden. Bij risico-onderzoeken worden de beoordelingen van de standaarden samengevat in een eindoordeel (waarbij een school als geheel bijvoorbeeld 'zwak' bevonden kan worden).

Figuur 3.1



Stroomdiagram van het toezicht. Het toezicht kent twee lijnen: 1) in blauw het risicogericht toezicht, op basis van een risico-instrument (nu de prestatiemonitor), 2) in groen de 4-jaarlijkse bestuurs-onderzoeken met bijbehorende verificatie onderzoeken.

De tweede belangrijke lijn waarbinnen onderzoek wordt uitgevoerd zijn de 4-jaarlijkse bestuursonderzoeken. De inspectie voert minstens iedere vier jaar een inspectie uit bij alle schoolbesturen in Nederland. Als onderdeel van deze onderzoeken worden ook enkele van de scholen onder het bestuur bezocht en worden enkele standaarden beoordeeld. Uit deze bezoeken volgt normaliter geen

samenvattend eindoordeel over de school (bij onderzoek naar risico's en bij onderzoek op verzoek van het bestuur gebeurt dit normaliter wel).
Naast deze twee schoolspecifieke lijnen voert de IvhO ook zogenaamde stelsel- en themaonderzoeken uit, waarbij gekeken wordt naar aspecten van het onderwijsstelsel als geheel zoals leesvaardigheid, rekenvaardigheid of bijvoorbeeld onderwijs gericht op het Fries. Het doel van deze onderzoeken is om inzicht te krijgen in het onderwijsbestel als geheel. Individuele scholen (of schoolbezoeken) vormen onderdeel van een steekproef met als doel een representatief beeld te vormen over kwaliteitsaspecten van scholen in Nederland. Deze vorm van toezicht is dus niet primair gericht op de individuele scholen. Binnen deze onderzoeken worden vaak beoordelingen gegeven over standaarden die in relatie staan tot het onderwerp van het specifieke stelsel- of themaonderzoek. Deze onderzoeken resulteren echter nooit in een eindoordeel over de school. Mocht een inspecteur tijdens een dergelijk onderzoek echter ernstige tekortkomingen tegenkomen dan kan een dergelijk onderzoek vanzelfsprekend alsnog worden omgezet naar een risicogericht kwaliteitsonderzoek.

3.2 Evaluatie/selectie van de labels

Een zeer belangrijke keuze in dit project betrof de selectie van de labels, ofwel de afhankelijke variabele. Dit betreft namelijk de vraag: *wat beogen we te voorspellen?* Het definiëren van risicoscholen kan namelijk op velerlei manieren. Een voor de hand liggende keuze zou zijn om de gegeven eindoordelen te voorspellen. Een belangrijk nadeel van deze beoordelingen is echter dat er jaarlijks maar weinig scholen een onvoldoende krijgen als eindoordeel. Dit maakt het lastiger voor voorspellingsmodellen om te 'leren' hoe deze beoordelingen voorspeld kunnen worden. Een ander nadeel van eindoordelen is dat deze in voorgaande jaren voor een belangrijk deel op behaalde eindtoetsscores gebaseerd zijn geweest. In recente jaren stuurt de IvhO meer op de beoordeling van een breder palet aan kwaliteitsaspecten. In dit project zijn daarom verschillende opties overwogen voor de definitie van risico's. Deze sluiten aan op de verschillende fasen van het onderzoek in het risicogericht toezicht (zoals beschreven in Figuur 3.1). Deze opties hebben echter ieder belangrijke voor- en nadelen.

3.2.1 Uitkomst van de expertanalyse

Voordeel: Net als de risicoproducten is de uitkomst van de expertanalyse (deskresearch) grotendeels gebaseerd op de gegevens (data) die de IvhO tot haar beschikking heeft. Daarmee sluit de expertanalyse nauw aan bij de informatie die ter beschikking kan staan voor risicoproducten. Dit betekend dat er relatief weinig 'extra informatie' tussen de twee soorten risicoschattingen zullen zitten.

Nadeel: Er heeft slechts zeer beperkte historische verslaglegging van deskresearch plaatsgevonden. Bovendien blijkt deze verslaglegging niet goed gestandaardiseerd. Dit maakt deze fase van het risicogericht toezicht op dit moment een slechte kandidaat.

3.2.2 Wel of geen risico-onderzoek

Voordeel: Zoals aangegeven behelst een van de belangrijke uitdagingen in dit project het vinden van een label met een groot aantal targets (risico-objecten). Wanneer het wel of niet uitvoeren van een risico-onderzoek gebruikt wordt als afhankelijk variabele, worden ook scholen die tijdens een schoolbezoek als nét voldoende beoordeeld zijn aangewezen als voorbeeld van een mogelijke onvoldoende school. Dit zorgt voor een groter aantal voorbeelden van onvoldoende scholen, wat de voorspelkracht naar verwachting ten goede zou komen.

Nadeel: Het al dan niet uitvoeren van een risico-onderzoek geeft weinig inzicht in de onderliggende redenen voor dit besluit. Daarnaast is dit besluit in de voorgaande jaren voornamelijk gebaseerd op behaalde eindtoetsscores. Het aanwijzen als targets van scholen die aan een risico-onderzoek zijn onderworpen zou daarmee waarschijnlijk een te eenzijdig beeld van kwaliteit kunnen schetsen. Dit maakt het wel of niet uitvoeren van een risico-onderzoek tot een slechte kandidaat.

3.2.3 Standaarden

Voordeel: Hoewel inspecteurs conservatief zijn bij het geven van onvoldoende eindoordelen zijn ze doorgaans minder conservatief bij het geven van onvoldoendes op de onderliggende standaarden. Dus wanneer een school in het geheel (net) voldoende presteert, kan de inspecteur een duidelijke aanwijzing tot verbetering geven door een onvoldoende te geven op een specifieke standaard. Dit zorgt voor een groter aantal targets (voorbeelden van onvoldoende scholen op een specifiek aspect) en zou daarmee de voorspelkracht ten goede kunnen komen. Een bijkomend voordeel is dat inzicht in risicoscores op standaarden inzicht kan geven in waar eventuele risico's bij een school zitten. Deze informatie zou analisten en inspecteurs kunnen helpen bij het voorbereiden van deskresearch of een schoolbezoek.

Nadeel: De IvhO werkt met waarderingskaders die van tijd tot tijd worden aangepast om aan te sluiten bij nieuwe regelgeving en inzichten. Ook recent heeft er een belangrijke transitie plaatsgevonden van het zogenaamde 2012-kader naar het 2017-kader. Deze transitie was zo ingrijpend dat de categorieën van kwaliteitsaspecten op veel punten sterk zijn gewijzigd. In de 2012-kaders werden beoordelingen bijvoorbeeld gegeven op zogenaamde indicatoren (een groot aantal, zeers specifieke kwaliteitsaspecten) en vanaf de 2017-kaders op standaarden (een beperktere set breed geformuleerde kwaliteitsaspecten). Ook is er bij de overgang gewisseld tussen verschillende schalen (4 en 5 puntschaal van kwaliteitsniveaus). Daarnaast kunnen inspecteurs tijdens schoolbezoeken observaties doen die niet beschreven worden in de beschikbare data. Dit zou een nadelige invloed kunnen hebben op de voorspelkracht van voorspellingsmodellen.

3.2.4 Eindoordelen

Voordeel: Eindoordelen sluiten het meest aan op de intuïtieve definitie van risicoscholen voor een breder publiek. Het behelst immers een totaal-oordeel van een school.

Nadeel: Zoals aan het begin van deze sectie is aangegeven krijgen jaarlijks relatief weinig scholen een onvoldoende eindoordeel. Bovendien geven ze vanwege de historische focus op onderwijsresultaten een eenzijdig beeld van kwaliteitsaspecten. Wederom geldt ook hier dat inspecteurs tijdens schoolbezoeken observaties doen die niet beschreven worden in de beschikbare data. Dit zou een nadelige invloed kunnen hebben op de voorspelkracht van voorspellingsmodellen.

3.2.5 Standaarden en eindoordelen: beschrijving

Op basis van de hierboven beschreven overwegingen is besloten om zowel de eindoordelen als de standaarden te gebruiken als labels. De eindoordelen vanwege de meest directe relatie tot risicoscholen en de standaarden vanwege de grotere aantallen geobserveerde onvoldoendes en het mogelijk betere inzicht in de onderliggende kwaliteitsaspecten.

De manier waarop de IvhO het toezicht op scholen invult wordt beschreven in het zogenaamde onderzoekskader[6]. Het onderzoekskader omvat de werkwijze van de inspectie en het zogenaamde waarderingskader waarin de normen voor de verschillende kwaliteitsgebieden beschreven staan. In de afgelopen jaren zijn er wisselingen geweest in de gebruikte onderzoeks- en waarderingskaders. Om te trainen op basis van historische beoordelingen was het van belang om de verschillende soorten beoordelingen om te coderen naar een uniforme manier van beoordeling. Voor voormalige waarderingskaders zijn daarom, in overleg met enkele inspecteurs, de beoordelingen gehercodeerd van de toenmalig gebruikte kwaliteitsaspecten (deze werden indicatoren genoemd) naar scores op standaarden zoals in het huidige waarderingskader. Tabel 8.1 (bijlage) beschrijft de structuur van omcoderen. Bij het omcoderen zijn de meest zwaarwegende standaarden meegenomen (waaronder de zogenaamde 'kernstandaarden' OP2, OP3, SK1, OR1, zie onder). Om de overwegingen bij het omcoderen te illustreren: indicator 1.1 (Opbrengsten) uit het 2012 kader geeft de volgende beschrijving: *"De resultaten van de leerlingen aan het eind van de basisschool liggen ten minste op het niveau dat op grond van de kenmerken van de leerlingenpopulatie mag worden verwacht"*. Deze indicator is gehercodeerd naar standaard OR1 (Onderwijsresultaten), omdat deze het best aansluit bij de bijbehorende beschrijving: *"De school behaalt met haar leerlingen leerresultaten die ten minste in overeenstemming zijn met de gestelde norm"*. Ten slotte zijn beoordelingen op standaarden gedichotomiseerd naar onvoldoende ("slecht" en "onvoldoende") en voldoende ("voldoende" en "goed") beoordelingen omdat dit het modelleerproces sterk versimpelt. Binnen de IvhO wordt het onderscheid tussen 'voldoende' en 'goed' aangeduid met de term waardering (i.p.v. beoordeling). Omdat dit onderscheid in het huidige project expliciet niet gemaakt wordt, gebruiken we voor de consistentie simpelweg de term 'beoordeling' voor het onderscheid tussen voldoende en onvoldoende scholen.

Beoordelingen op de volgende kernstandaarden zijn meegenomen:

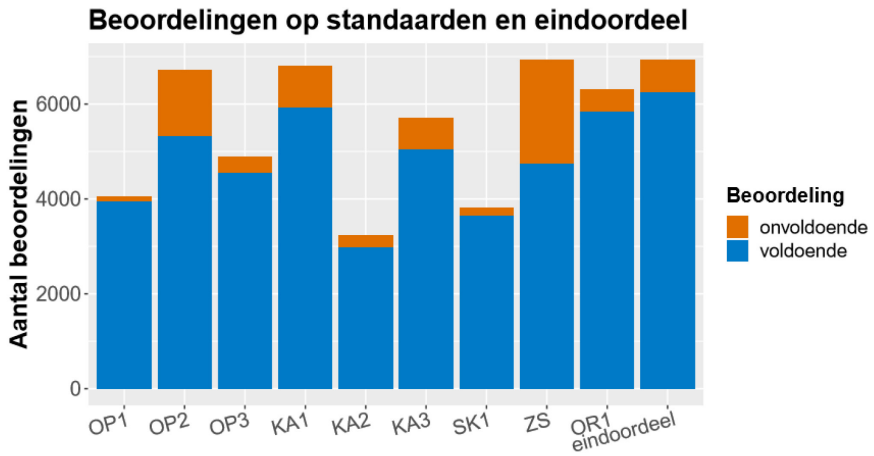
1) OP1 (Aanbod)
2) OP2 (Zicht op ontwikkeling)
3) OP3 (Didactisch handelen)
4) SK1 (Veiligheid)
5) OR1 (Resultaten)
6) KA1 (Kwaliteitszorg)
7) KA2 (Kwaliteitscultuur)
8) KA3 (Verantwoording en dialoog)

Daarnaast is een samengevoegde standaard berekend, die aangeeft of één van de zogenaamde 'zachte standaarden' (OP1, OP2, OP3, SK1, KA1, KA2, KA3) onvoldoende was (deze gecombineerde standaards wordt aangeduid met ZS). We definiëren deze standaarden hier als 'zacht' om het contrast aan te geven met OR1 welke op vrij 'harde' data (zoals eindtoetsresultaten) gebaseerd kan worden. Daarnaast is per onderzoek het gegeven eindoordeel meegenomen. Bij onderzoeken waarbij wel op standaarden is gescoord -maar geen eindoordeel is geregistreerd- is het eindoordeel 'voldoende' geïmputeerd om daarmee de dekking van dit oordeel te verhogen. Wanneer er op een school sterke aanwijzingen zouden zijn geweest voor problemen zou een onderzoek namelijk zijn omgezet in een risico-onderzoek en zou ook een bijbehorend eindoordeel (al dan niet onvoldoende) zijn geregistreerd.

---

6 https://www.onderwijsinspectie.nl/binaries/onderwijsinspectie/documenten/rapporten/2019/06/06/onderzoekskader-2017-po-en-vve/Onderzoekskader+po+versie+aug19.pdf

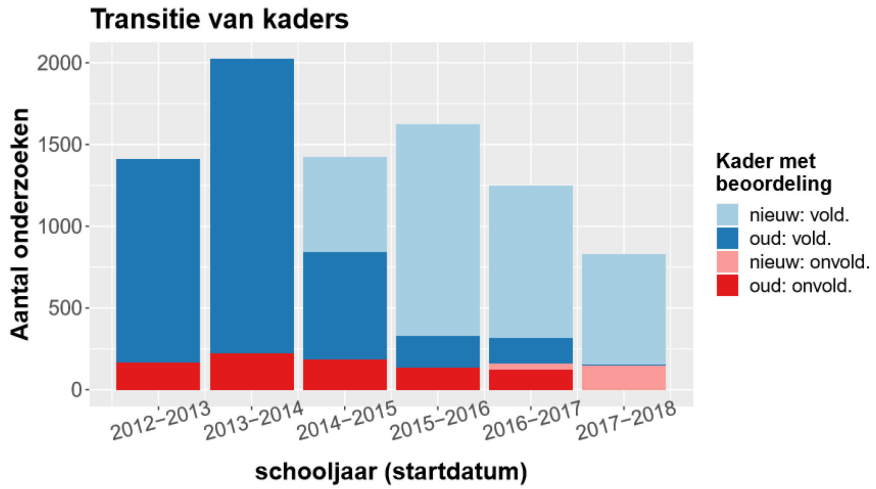
Figuur 3.2



Historische beschrijving van beoordelingen op 8 standaarden (deels gehercodeerd vanaf oude kaders); een standaard die de individuele zogenaamde "zachte standaarden" combineert (ZS); en de eindoordelen. Alle beoordelingen zijn gedichotomiseerd tot voldoende en onvoldoende. Inspecteurs geven aanzienlijk vaker voldoendes dan onvoldoendes. Zie figuur 8.1 voor de aantallen onderzochte standaarden over de jaren.

In Figuur 3.2 wordt een overzicht gegeven van de beoordelingen die volgens onze wijze van coderen op de verschillende standaarden (en het eindoordeel) zijn gegeven. Er bestaat een sterke disbalans in de beoordelingen. Voor de meeste labels wordt slechts in ongeveer 10-15% van de onderzochte scholen een onvoldoende gegeven. Reguliere voorspellingsmodellen zijn gevoelig voor een 'majority class', waardoor de signalering van de 'minority class' (de risicoscholen) suboptimaal is. In de gecombineerde standaard ZS is dit probleem enigszins verlicht met een percentage onvoldoendes van ongeveer 30%, maar nog steeds aanwezig. Hoewel een uitdaging, vormt dit een veelvoorkomend probleem, vooral in de context van toezichthouders, waarbij objecten van toezicht die onvoldoende zijn, doorgaans uitzonderingen zijn. Een belangrijk gevolg van dit aspect is dat de kwaliteit van modellen niet geëvalueerd moet worden op basis van het percentage correct voorspelde labels. Een model dat simpelweg aan alle scholen een voldoende toekent komt dan immers al gauw op een hoge 'accuratesse' uit. Een betere maat voor deze beoordeling vormt de Area Under the Curve (AUC). Daarover meer in het hoofdstuk "Modelleren".
Zoals aangegeven vormen de verschillende gebruikte waarderingskaders een andere belangrijke uitdaging in de data over beoordelingen. Figuur 3.3 laat zien welke onderzoeken gestart zijn onder de twee verschillende kaders. Daarnaast is ook weergegeven welke eindoordelen onder desbetreffende kaders gegeven zijn. Hieruit is op te maken dat er vanaf schooljaar 2014-2015 begonnen is met het scoren op het nieuwe kader (het zogenaamde 2017- kader). Dit kader werd echter alleen gebruikt bij schoolbezoeken die uitmondden in voldoende eindoordelen. Voor risico-onderzoeken, waarbij er grote problemen geconstateerd (konden) worden, werd teruggegrepen op het oude kader; het op dat moment wettelijke kader. Pas vanaf 2017-2018 werd ook voor hoog-risico schoolbezoeken overwegend het nieuwe kader gehanteerd. Vanwege het belang van de onvoldoendes (de targets) voor het trainen van modellen, kan dit patroon een verstoring opleveren in de jaar-op-jaar voorspellingen. Hier wordt op teruggekomen bij de beschrijving van de resultaten. Verder is ook te zien dat het totaal aantal schoolbezoeken waarbij een oordeel is gegeven in recente jaren is afgenomen (terwijl het totaal aantal onvoldoendes redelijk gelijk gebleven is).

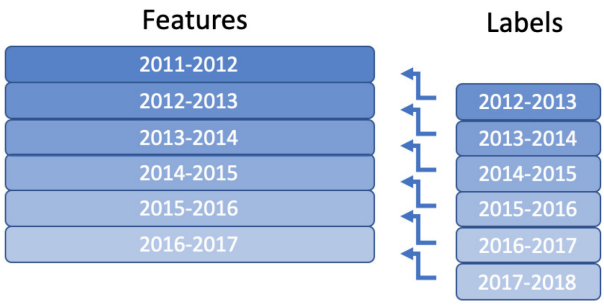
Figuur 3.3



De historische dataset omvat een transitie van waarderingskaders. Waarbij binnen oude kaders gescoord werd op indicatoren en nieuwe kaders op standaarden. Opvallend daarbij is dat het nieuwe kader vanaf schooljaar 2014-2015 al gebruikt werd bij een groot deel van de bezoeken die uitmondden in een voldoende eindoordeel (laag-risico bezoeken). Pas vanaf schooljaar 2017-2018 is het nieuwe kader ook gebruikt bij onderzoeken die uitmondden in onvoldoende eindoordelen (risico-onderzoeken).

Het doel van dit project behelst het *voorspellen* van risicoscholen. Dat wil zeggen: gegeven de data in jaar X, wat zal de beoordeling zijn van een inspecteur in jaar X + 1? Bij het koppelen van de labels aan de features zijn daarom de labels met een jaar teruggeschoven. Het gevolg daarvan is dat het optimaliseren van modellen al direct gebaseerd is op statistische relaties die voorspelkracht hebben m.b.t. de beoordelingen die een jaar later gegeven zouden worden bij een schoolbezoek. Bovendien sluit deze benadering ook beter aan bij de praktijk: wanneer een inspecteur bijvoorbeeld in maart 2020 een inspectiebezoek brengt aan een specifieke school, dan zal de inspecteur de beoordeling op bijvoorbeeld OR1 baseren op de eindtoetsscores zoals behaald in april 2019 (dus schooljaar 2018-2019). Deze benadering is weergegeven in Figuur 3.4.

Figuur 3.4



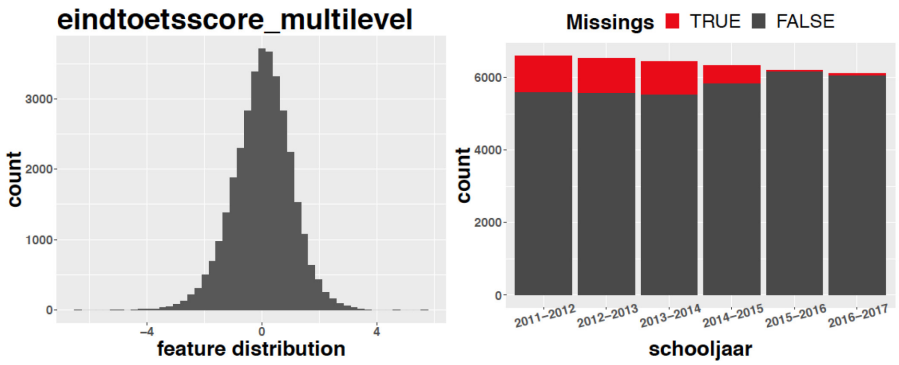
Bij het koppelen van de datasets zijn de labels (oordelen) een jaar teruggeschoven, t.b.v. het voorspellende karakter van dit project.

3.3 Features: selectie en beschrijving

Voor de selectie van features (voorspellers) is uitgegaan van een zo breed mogelijk scala van gegevens die binnen afzienbare tijd verzameld konden worden. Dit zijn gegevens waarvan vooraf geschat werd dat ze een mogelijke correlatie met

onderwijskwaliteit zouden kunnen hebben. De keuze daarin is vooral ingegeven door gesprekken met experts en op basis van de inschattingen van de teamleden van het project. Deze dataset staat beschreven in de Data Documentatie Algoritmische Selectie Risicoscholen (het "Codebook"; Edocsnummer 5471020). De dataset behelst gegevens zoals: verschillende behaalde leerresultaten van leerlingen; demografische karakteristieken van de leerling populatie; school-aspecten zoals denominatie; gegevens over het bestuur en het personeel; financiële gegevens over het bestuur; demografische gegevens over de buurt van de school; en geografische informatie (zoals provincie). Deze dataset behelst ongeveer 160 verschillende features. Om een beeld te geven van de inhoud van de datadocumentatie en de structuur van de gegevens worden hieronder bij wijze van voorbeeld van enkele van deze features de verdeling en het percentage missende waarden per jaar weergegeven.
Een voorspeller die historisch zeer belangrijk is geweest voor het beoordelen van onderwijskwaliteit betreft de behaalde eindtoetsscores. Figuur 3.5 laat de kernstatistieken van deze features zien. Het paneel links laat zien dat deze variabele numeriek en normaal verdeeld is. Het paneel rechts laat zien dat het aantal missende waarden afneemt voor recentere jaren.

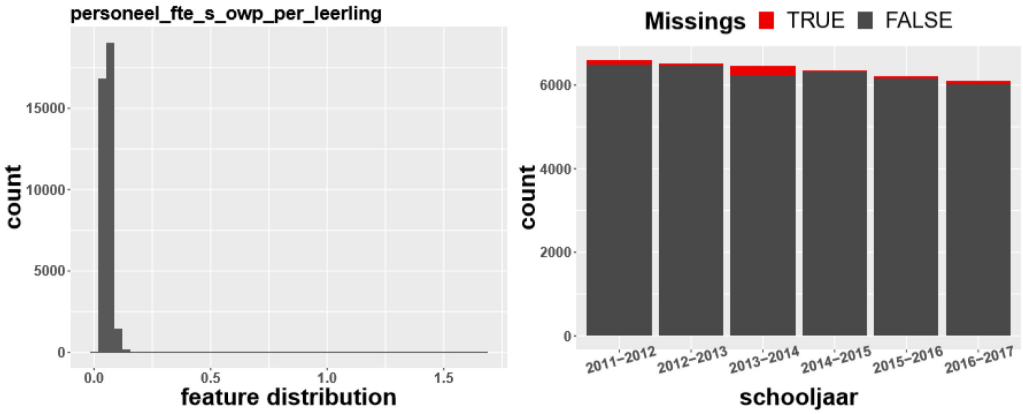
Figuur 3.5



Beschrijving van het feature eindtoetsscore_multilevel. Deze feature omvat gestandaardiseerde eindtoetsscores, gecorrigeerd voor % gewichtenleerlingen. Paneel 1 (links) beschrijft de verdeling van waarden d.m.v. een histogram. Paneel 2 (rechts) beschrijft het percentage missende waarden (NA) per schooljaar.

Een andere voorspeller die meegenomen is in dit onderzoek betreft het gemiddelde aantal fte's onderwijzend personeel per leerling (Figuur 3.6). Het figuur laat zien dat de verdeling zeer scheef is. Dit onderstreept de noodzaak voor een aantal verdere data-voorbewerkingsstappen. Het paneel rechts laat zien dat deze feature weinig missende waarden heeft.

Figuur 3.6



beschrijving van het feature aantal fte onderwijzend personeel per leerling. Zie Figuur 3.5 voor verdere beschrijving van de panelen.

3.4 Uitdagingen in de dataset

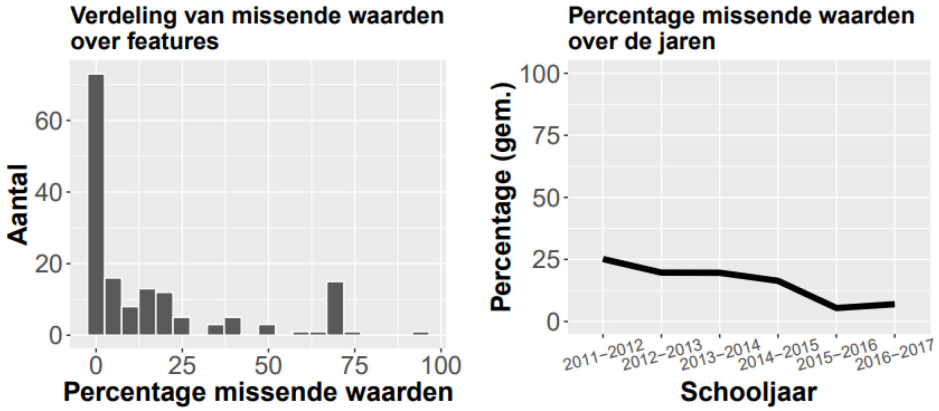
De dataset bevat een aantal specifieke uitdagingen. Hoewel zeker niet uniek voor deze data, maken deze uitdagingen het belangrijk om weloverwogen keuzes te maken over hoe er mee om te gaan. Naast de disbalans in beoordelingen (zoals besproken in de sectie over standaarden en eindoordelen), gaat dit over het aanzienlijke aantal missende waarden; de sterke covariantie tussen verschillende features; en bijvoorbeeld de hiërarchische relaties in de data. Hieronder bespreken we deze uitdagingen kort. Een belangrijk deel van het optimaliseren van modellen behelst het toepassen van vormen van voorbewerking om met deze uitdagingen om te gaan. Daarbij kan gedacht worden aan imputatie (het vervangen van missende

waarden door bijvoorbeeld de mediaan of het gemiddelde) en het toepassen van technieken zoals Principale Componenten Analyse om covariantie tegen te gaan. In de sectie feature engineering worden de effecten van enkele oplossingen voor deze uitdagingen besproken.

3.4.1 Missende waarden

Een aanzienlijk deel van de features is niet compleet gevuld voor de dataset (Figuur 3.7). Verder heeft slechts 61% van de features minder dan 1% missende waarden. Wel is te zien dat het aantal missende waarden aanzienlijk afneemt voor de meer recente schooljaren, wat de voorspelkracht van modellen waarschijnlijk ten goede zal komen.

Figuur 3.7

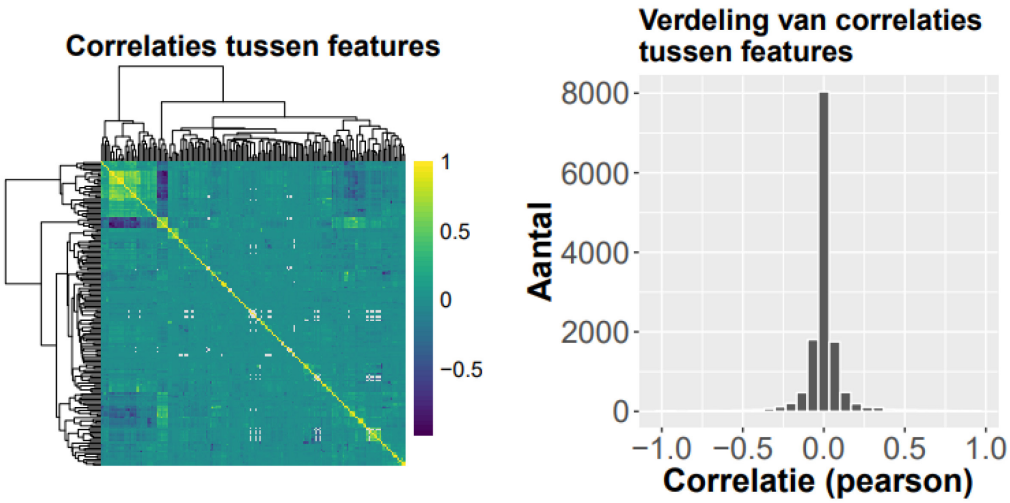


Verdeling van missende waarden (links) en het gemiddelde percentage missende waarden per schooljaar (rechts) over de features in de dataset. Veel features bevatten missende waarden. Het aantal missende waarden neemt af in de recentere jaren.

3.4.2 Samenhang tussen features

Een flink aantal features in de dataset vertonen onderlinge samenhang. Zo bestaat er bijvoorbeeld vanzelfsprekend samenhang tussen het percentage leerlingen dat referentieniveau 1F behaald voor leesvaardigheid en het percentage leerlingen dat niveau 2F heeft behaald. Om de onderlinge samenhang tussen de features inzichtelijk te maken worden eerst alle categorische variabelen d.m.v. dummy codering omgezet naar een numerieke codering. Vervolgens zijn de zogenaamde ‘pairwise complete’ correlaties berekend tussen alle features.

Figuur 3.8



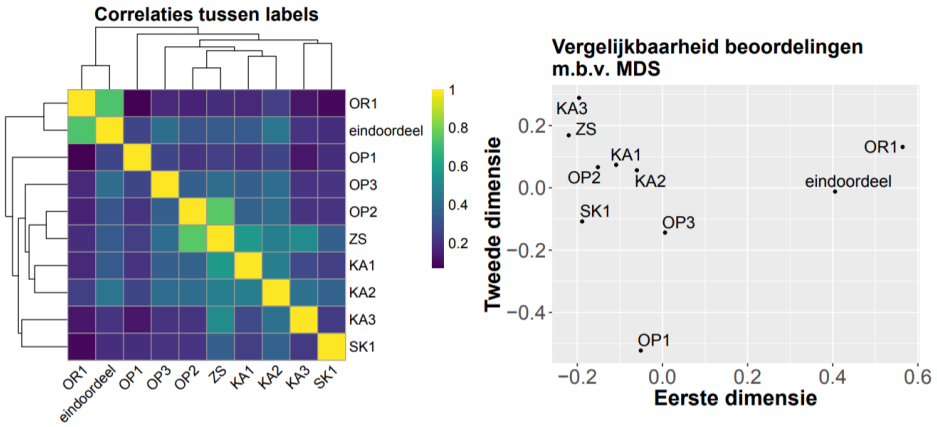
Gegroepeerde correlaties tussen features (links) en de verdeling van de correlaties tussen alle features (rechts). Er zijn groepen features die onderling sterk correleren. Voor de leesbaarheid van het figuur zijn de rij en kolomnamen van het linker paneel weggelaten.

Het linker paneel van Figuur 3.8 beschrijft deze correlaties d.m.v. een correlatiematrix, waarbij getracht is om features met sterke samenhang te groeperen (o.b.v. hiërarchische clustering). Hieruit wordt zichtbaar dat er inderdaad groepen features met sterke samenhang in de dataset zitten. De correlaties die linksboven rond de diagonaal gegroepeerd staan, beschrijven voornamelijk demografische karakteristieken van de leerlingenpopulatie en de buurt van de school (de features uit de leefbaarometer, apcg-scores en wijkgegevens). Dit zijn features zoals het percentage niet-westerse leerlingen en de apcg-score die beschrijft welk percentage van de leerlingen van een school uit een gezin komt waarvan een of beide ouders een uitkering krijgen. Het kleine cluster rechtsonder langs de diagonaal (met sterk positieve correlaties) beschrijft de samenhang tussen een aantal features die leerresultaten beschrijven (zoals behaalde referentieniveaus).

3.4.3 Samenhang tussen labels

Er bestaat ook aanzienlijke samenhang tussen beoordelingen op de verschillende standaarden onderling en met de eindoordelen. Figuur 3.9 visualiseert de relaties tussen de beoordelingen op basis van de correlaties (correlaties tussen complete observatie paren).

Figuur 3.9



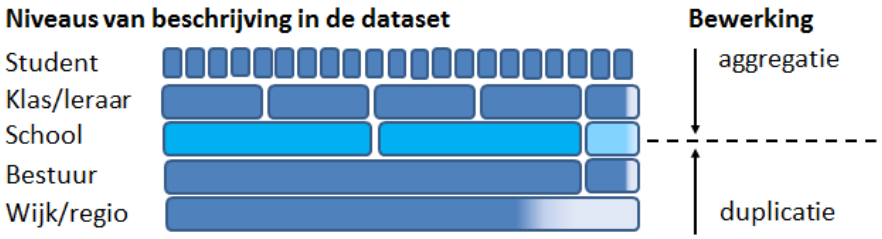
Samenhang tussen beoordelingen op de verschillende labels. Het linker paneel geeft de correlatiematrix weer. Labels waarvan de beoordelingen sterke samenhang vertonen zijn lichtgeel gekleurd (een hoge correlatie), en staan gegroepeerd op basis van hiërarchische clustering. Het rechterpaneel beschrijft dezelfde data maar dan met behulp van Multidimensional Scaling (MDS). Labels met sterke samenhang staan in deze beschrijving wederom dicht bij elkaar, maar dan verdeeld over twee dimensies.

Het linker paneel geeft een correlatie matrix weer waarbij getracht is om features met sterke samenhang te groeperen (o.b.v. hiërarchische clustering). Het rechter paneel beschrijft dezelfde data maar dan op basis van Multidimensional Scaling. In beide visualisaties staan labels met sterke samenhang dicht bij elkaar gegroepeerd. Uit deze visualisaties wordt zichtbaar dat beoordelingen op OR1 en de eindoordelen historisch gezien sterke samenhang vertonen. Verder vertonen de zachte standaarden onderling ook sterke samenhang met uitzondering van OP1 (deze standaard is historisch gezien overigens relatief weinig beoordeeld, zie Figuur 8.1). De gecombineerde zachte standaard (ZS) vertoont (zoals verwacht) ook goede samenhang met de meeste andere zachte standaarden. Binnen de individuele zachte standaarden bestaat er vooral sterke samenhang tussen beoordelingen op OP2 (zicht op ontwikkeling); OP3 (Didactisch handelen); KA1 (Kwaliteitszorg); en KA2 (Kwaliteitscultuur).

3.4.4 Hiërarchische relaties

Scholen zijn onderdeel van een hiërarchische structuur die bestaat uit leerlingen, klassen, scholen, besturen en regio's. Het vormt een uitdaging om op een elegante manier met deze vorm van nesting om te gaan (zie Figuur 3.10). In deze fase van het project is er voor gekozen om gegevens over leerlingen en klassen te aggregeren naar schoolniveau. Verder is informatie over besturen en regio's toegepast op alle onderliggende scholen. Dat betekent dat verschillende scholen onder een bestuur bijvoorbeeld dezelfde financiële feature-waarden zullen bevatten omdat deze alleen beschikbaar zijn op het niveau van een bestuur. Meer informatie over deze relaties per feature is te vinden in het Codebook. In deze fase is dus besloten om op deze relatief simpele manier om te gaan met de hiërarchische relaties. In de toekomst zou onderzocht kunnen worden of aan voorspelkracht gewonnen kan worden door gebruik te maken van bijvoorbeeld vormen van linear mixed effects regression omdat deze methode goed om kan gaan met data in een hiërarchische structuur. Dergelijke methoden zouden bijvoorbeeld informatie op zowel leerling- als bestuursniveau als zodanig kunnen meenemen. Gezien het aantal leerlingen in het regulier basisonderwijs zullen deze modellen naar verwachting echter zeer veel rekenkracht (en tijd) vereisen.

Figuur 3.10



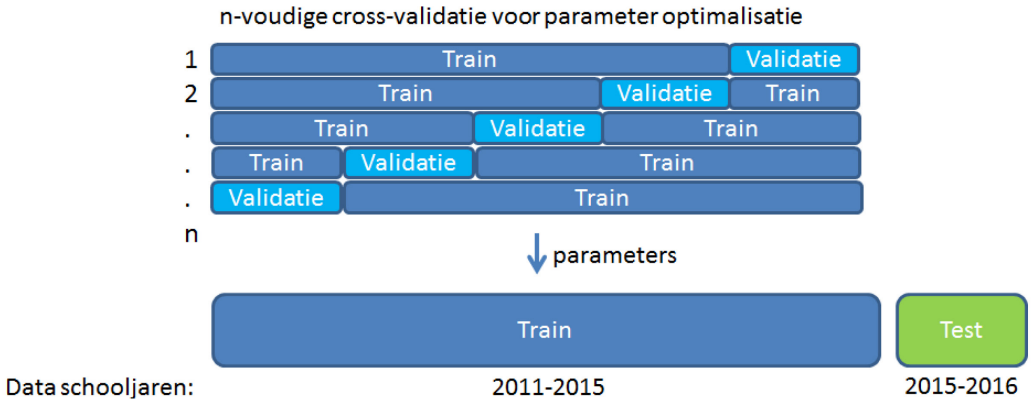
De dataset bevat informatie in hiërarchische relaties (nesting). Omdat het project voorspellingen doet op schoolniveau is informatie over kleinere eenheden geaggregeerd en dat van grotere eenheden gedupliceerd naar schoolniveau.

3.5 Opdeling in training-, validatie- en testsets

Het trainen van voorspellingsmodellen behelst het geautomatiseerd aanpassen van een (soms groot) aantal parameters totdat het model de data zo goed mogelijk benadert. Dat wil zeggen, tot het model gegeven de features een zo goed mogelijke voorspelling kan doen over de bijbehorende labels (oordelen). Een belangrijke en bekende tekortkoming van voorspellingsmodellen is dat deze algoritmen kunnen resulteren in zogenaamde overfitting. Het gevaar daarvan is dat het model zeer goed wordt in het beschrijven van de gebruikte trainingsset maar daarbij moet inboeten op voorspelkracht voor nieuwe – ongeziene - data. Gezien het juist voorspellende karakter van de toepassing van voorspellingsmodellen bestaan er een aantal noodzakelijke stappen om deze overfitting te voorkomen.
Een belangrijke benadering in het gebruik van voorspellingsmodellen voor het tegengaan van overfitting bestaat uit het opdelen van datasets in zogenaamde train, validatie en testsets: de onderzoeker traint modellen op basis van een trainingsset en verschillende modellen (met verschillende parameters) worden met elkaar vergeleken aan de hand van de validatieset. De voorspelkracht van het uiteindelijk gekozen (optimale) model wordt getoetst aan de hand van de testset. Voor het project signalering risicoscholen bestonden minimaal twee potentiele vormen van opdeling. Een (veelgebruikte) methode is dat train-, validatie- en testsets bestaan uit willekeurig gekozen subsets uit de gehele dataset, vaak in een verhouding rond de 60% train; 20% validatie; 20% testset. Een benadering die in ons geval echter gepaster leek definieert de train en validatieset, en de testset op basis van verschillende schooljaren.
De eerste benadering (willekeurige toebedeling) brengt namelijk enkele problemen met zich mee voor onze toepassing: wanneer scholen willekeurig worden geselecteerd uit de gehele populatie dan kunnen er binnen een schooljaar scholen uit hetzelfde bestuur in de verschillende sets terechtkomen (bijvoorbeeld in de train en testset). Dit is onwenselijk omdat een deel van de features (zoals financiële gegevens) hetzelfde zijn voor deze scholen. Dergelijke samenhang tussen scholen in de verschillende sets zou vervolgens een té rooskleurig beeld kunnen geven van de model voorspelkracht. Maar zelfs wanneer we scholen onder eenzelfde bestuur zouden samenvoegen in dezelfde sets, dan nog bestaan er verschillende vormen van samenhang binnen schooljaren in de dataset zoals groepering op basis van buurtkenmerken en andere regionale invloeden die gedeeld zullen zijn over scholen. Een ander nadeel is dat deze benadering niet direct aansluit op de manier van werken van de inspectie: we proberen geen voorspellingen te doen over de kwaliteit van willekeurige (ongeïnspecteerde) scholen in het verleden. De inspectie wil juist voorspellingen doen over de toekomst.
Een logischere benadering is daarom de keuze om de opdeling tussen train(+validatie) en testset te maken op basis van schooljaar. In deze benadering

roteren de train + validatieset bijvoorbeeld over verschillende subsets van data tussen 2011 en 2015 en bestaat de testset uit gegevens over schooljaar 2015-2016 (Dit behelst dus de beoordelingen over schooljaar 2016-2017; zie Figuur 3.4). Dit principe wordt gevisualiseerd in Figuur 3.11. In de gebruikte methode zijn de train- en validatiesets overigens niet opgedeeld per schooljaar, omdat deze opdeling lastiger aan te passen is (dit gebeurt "onder de motorkap" van veelgebruikte functies voor het trainen van modellen; dit zou een mogelijke toekomstige verbeterstap kunnen zijn).
Dit project kende in het modelleren twee belangrijke fasen. De eerste modelleerfase behelsde de hackathon. Voor deze fase zijn de labels van schooljaar 2016-2017 (gekoppeld aan de features van schooljaar 2015-2016) als testset aangewezen. De tweede modelleerfase betrof de periode van optimalisatie ná de hackathon. Voor deze periode zijn ook de labels uit schooljaar 2015-2016 betrokken bij de train+validatiesets en zijn de labels uit schooljaar 2017-2018 (gekoppeld aan features uit 2016-2017) achtergehouden als testset.

Figuur 3.11



Voorbeeld van gekozen indeling van de data in een train-, validatie- en testset voor de risicoscholen dataset voor de hackathon. Modellen worden geoptimaliseerd op basis van train- en validatiesets (data uit schooljaren 2011-2015). Deze indeling vindt meerdere keren plaats. Het model met de optimale parameters (of de optimale modelvorm) heeft een hoge voorspelkracht voor de verschillende validatiesets. De testset (compleet ongeziene labels) dient slechts ter uiteindelijke evaluatie van de voorspelkracht van het gekozen model. Na de hackathon werd ook schooljaar 2015-2016 aan de train en validatieset toegevoegd, en bestond de testset uit features van schooljaar 2016-2017 (labels uit schooljaar 2017-2018).

4 Modelleren

4.1 De afweging tussen borging van kwaliteit en doelmatigheid

Wanneer inspecteurs besluiten welke scholen bezocht gaan worden zal er altijd een afweging bestaan tussen twee belangrijke aspecten.

1) Vanuit het oogpunt van de borging van kwaliteit van het onderwijssysteem is het belangrijk om geen enkele onvoldoende school te missen.
2) Vanuit het oogpunt van de doelmatigheid is het belangrijk om geen scholen als risicovol aan te merken (en te bezoeken) als vervolgens blijkt dat deze scholen toch voldoende kwaliteit bieden.

Deze twee overwegingen zijn niet specifiek voor de onderwijscontext maar komen naar voren in vrijwel elk classificatieprobleem. Denk daarbij bijvoorbeeld aan het detecteren van ziektes (we willen zoveel mogelijk dragers van een ziekte als zodanig identificeren, maar toch ook zo weinig mogelijk mensen onnodig aan medische tests onderwerpen), of het controleren van zeecontainers op mogelijke aanwezigheid van drugs etc.
Deze afweging kan inzichtelijk gemaakt worden door een kruistabel waarbij in de rijen de verschillende soorten scholen staan (onvoldoende vs. voldoende scholen) en in de kolommen de voorspelde soorten risicocategorieën (voorspeld risico vs. geen voorspeld risico). In Tabel 4.1 komen de verschillende combinaties naar voren. De rode, cursief gedrukte cellen, beschrijven de onwenselijke gevallen: voldoende scholen die toch bezocht zijn (zogenaamde Vals-Positieven; ook bekend als Type I fouten) en de onvoldoende scholen die toch niet bezocht zijn (zogenaamde Vals-Negatieven; ook bekend als Type II fouten).

Tabel 4.1

	Voorspeld risico	Geen voorspeld risico
Onvoldoende scholen	**Onvoldoende school, terecht bezocht**	***Onvoldoende school, toch niet bezocht***
Voldoende scholen	***Voldoende school, toch bezocht***	**Voldoende school, terecht niet bezocht**

Een ogenschijnlijk intuïtieve manier om de voorspellingen van risico-instrumenten voor scholen te vergelijken is om te kijken welk percentage van alle beoordelingen correct voorspeld zijn. Omdat er echter voor de verschillende beoordelingen een sterke disbalans bestaat (meer voldoendes dan onvoldoendes) geeft deze metriek een sterk vertekend beeld. Wanneer er slechts 10% onvoldoende scholen in de dataset zijn dan zal een simpel model, welke aan *iedere* school een voldoende toekent, al een percentage correct behalen van 90%. Toch zou dit model in het toezicht geen enkele waarde hebben. We missen immers *alle* onvoldoende scholen. Een veelgebruikte manier om naar classificatieproblemen te kijken behelst daarom een afweging tussen zogenaamde sensitiviteit en specificiteit. In de context van risicoscholen beschrijft sensitiviteit hoeveel van de onvoldoende scholen ook daadwerkelijk als risicovol worden aangemerkt. Specificiteit kijkt echter naar de set van voldoende scholen en geeft weer hoeveel van de voldoende scholen ook daadwerkelijk een laag voorspeld risico hebben. Idealiter zijn zowel de specificiteit als de sensitiviteit hoog. In werkelijkheid is dit echter zelden haalbaar en zal een

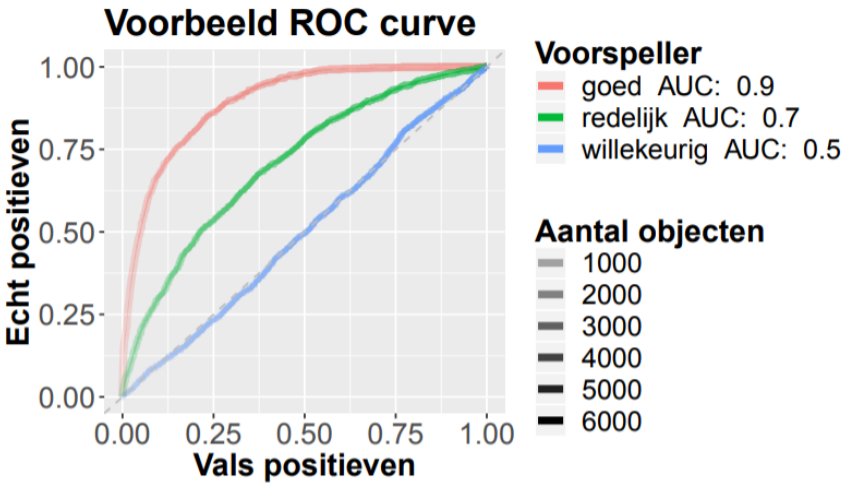
afweging gemaakt moeten worden. Om een bepaald percentage van alle onvoldoende scholen te bezoeken zal men in de praktijk toch met enige regelmaat ook scholen moeten bezoeken die achteraf voldoende bleken te zijn. En vice versa; om niet alle scholen te hoeven bezoeken, moeten we vaak accepteren dat we ook enkele onvoldoende scholen zullen missen. Maar hoe kunnen we tot een geïnformeerd besluit komen over hoeveel scholen dan bezocht moeten worden? En als we de risicogrens moeilijk te bepalen vinden, hoe kunnen we verschillende modellen dan met elkaar vergelijken?

4.2 Het vergelijken van risicomodellen: AUC en precision at k

Een veelgebruikte methode om sensitiviteit en specificiteit te gebruiken voor het vergelijken van voorspellingsmodellen is door deze te combineren in een zogenaamde Receiver Operating Curve (zie Figuur 4.1). De ROC curve beschrijft de relatie tussen sensitiviteit en specificiteit bij een reeks aan grenswaarden: wanneer een risicomodel erg conservatief is (of conservatief wordt toegepast), dan zal de sensitiviteit laag zijn, maar de specificiteit erg hoog (punten linksonder in de grafiek). Een liberaal model heeft een hoge sensitiviteit maar een zeer lage specificiteit (punten rechtsboven in de grafiek). In dit figuur beschrijft elke lijn een hypothetische modelvorm. Om het verschil tussen deze soorten modellen te kwantificeren wordt de zogenaamde Area Under the Curve (AUC) gebruikt. Dat wil zeggen, de AUC behelst het totale oppervlak onder de lijnen. Modellen die erg goed voorspellen, lopen langs de linker bovenhoek en geven een AUC van 1 (perfecte classificatie: detectie van alle Echt Positieven maar geen Vals Negatieven). De rechte diagonaal beschrijft de voorspelkracht van een compleet willekeurig model (het voorspelt niet beter dan kans, ongeacht de grenswaarde) en geeft een AUC van 0.5. Een AUC van 0.7, wat realistischer is in de context van het voorspellen van risicoscholen, beschrijft daarmee een redelijk scorend model. AUC is gebruikt als de belangrijkste maat om modellen met elkaar te vergelijken in dit project.
Naast de AUC zijn er nog verschillende andere manieren om naar voorspelkracht te kijken in de context van onderzoek bij risicoscholen, afhankelijk van het doel van classificatie. Een aspect dat in de context van onderwijs relevant is, is dat de capaciteit van inspecteurs voor risico-onderzoeken zeer beperkt is. In 2018 zijn er in het primair onderwijs bijvoorbeeld 1774 onderzoeken uitgevoerd op scholen. Daarvan vielen er echter slechts 132 in de categorie Risico-Onderzoek op een School en 104 in de categorie Herstelonderzoek[7]. De overige onderzoeken zijn dus in eerste instantie niet uitgevoerd in het kader van een vermoeden van ernstig verhoogde risico's (verificatieonderzoeken, stelsel- en themaonderzoeken, etc.). Gegeven de beperkte capaciteit is het dus vooral belangrijk om modelvormen met elkaar te vergelijken op redelijk conservatieve grenswaarden. Een mogelijke metriek is de zogenaamde precision at k. Dat wil zeggen, het aantal correct voorspelde labels wanneer we uitgaan van een grenswaarde die slechts een beperkt aantal scholen als risico-school aanwijst. Dit aantal is aangeduid als k. In het beschrijven van de resultaten zullen we in een aantal gevallen ook de precision at k geven voor verschillende modellen. Als k gebruiken we hier 600 (i.p.v. 132). Een belangrijke reden daarvoor is dat risicoproducten binnen de IvhO in eerste instantie vooral gebruikt worden om de inzet van expertanalyses mee te bepalen. Het is aannemelijk dat er jaarlijks meer expertanalyses dan daadwerkelijke schoolbezoeken uitgevoerd kunnen worden.

[7] Jaarverslag 2018 Inspectie van het Onderwijs: https://www.onderwijsinspectie.nl/binaries/onderwijsinspectie/documenten/jaarverslagen/2019/06/11/jaarverslag-2018/Jaarverslag+Inspectie+van+het+Onderwijs+2018.pdf

Figuur 4.1

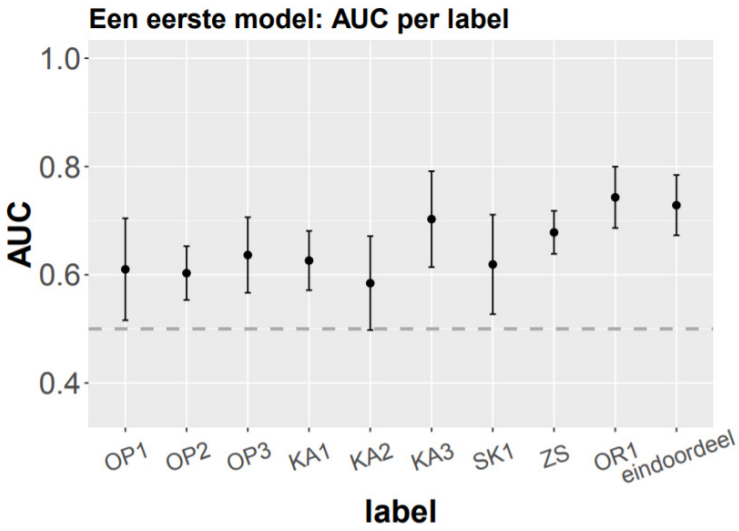


Voorbeeld Receiver Operating Curves (ROC) voor drie soorten hypothetische voorspellers met de bijbehorende AUC waarden. Een goede voorspeller zal zorgen voor een efficiënte sortering. Als gevolg zal het aantal correct aangewezen targets (Echt Positieven) sneller toenemen dan het aantal Vals Positieven wanneer objecten een-voor-een onderzocht worden, beginnend bij het object met de hoogste risicoscore. De transparantieschaal geeft het aantal onderzochte objecten weer over een continu verschuivend criterium.

4.3 Een eerste model: logistische regressie

Ter introductie bespreken we hier een eerste model en de bijbehorende resultaten. In dit geval vinden er geen vormen van voorbewerking plaats (behalve imputatie met de mediaan per feature voor missende waarden). Alle features zoals beschreven in het Codebook worden in deze analyse meegenomen. We trainen een standaard logistisch regressiemodel voor alle 10 de type beoordelingen (beoordelingen als afhankelijke variabelen) en voorspellen vervolgens op basis van het model de beoordelingen voor het jaar daarna. Op basis van deze voorspellingen en de daadwerkelijke beoordelingen kan een AUC-score berekend worden. Figuur 4.2 visualiseert de behaalde AUC-scores voor dit model over de verschillende labels.

Figuur 4.2



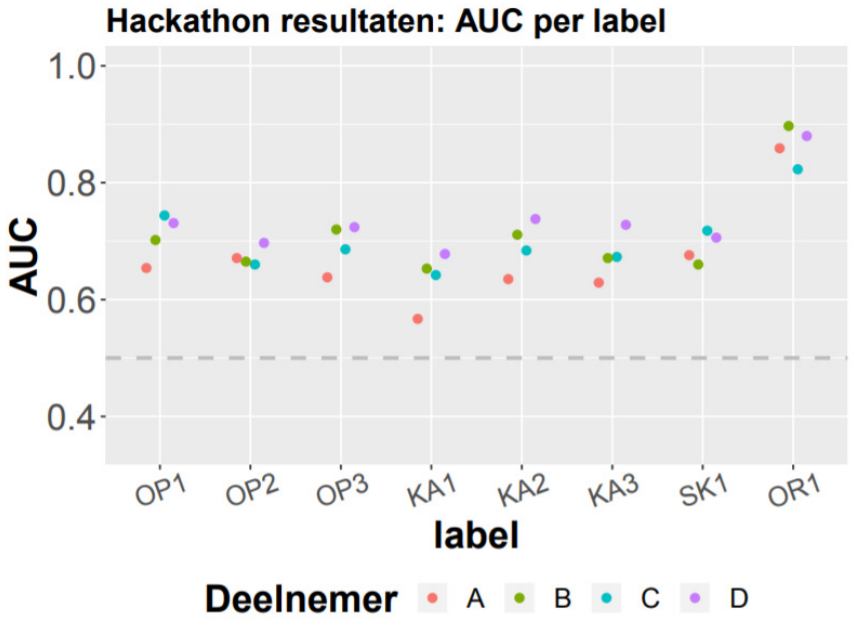
Voorspelkracht van een standaard logistisch regressiemodel voor de verschillende labels. Het model is getraind op basis van gegevens over 2014-2015 met de bijbehorende labels uit 2015-2016. Voorspellingen zijn gegenereerd voor schooljaar 2016-2017 op basis van gegevens over 2015-2016. De resulterende risicoscores zijn vergeleken met daadwerkelijke beoordelingen. Voorspelling op kans niveau geeft een AUC van 0.5 (grijze stippellijn). Error-bars reflecteren 95% betrouwbaarheidsintervallen op basis van bootstrapping van de AUC.

Deze eerste analyse laat zien dat er verschillen bestaan tussen de mate waarin beoordelingen op de verschillende labels te voorspellen zijn. De meeste zachte standaarden behalen een AUC rond de 0.6. OR1 (Resultaten), KA3 (Verantwoording en dialoog), de gecombineerde zachte-standaarden en eindoordelen zijn iets beter te voorspellen. Deze laten een AUC rond de 0.7 zien.

4.4 Hackathon

In mei 2019 heeft de projectgroep een hackathon georganiseerd. Meerdere deelnemers hebben geprobeerd de beoordelingen op de 8 individuele standaarden te voorspellen (de eindoordelen en het gecombineerde label waren geen onderdeel van de Hackathon). Daarbij konden ze gebruik maken van modelvormen naar keuze. Vooral wanneer een dataset, zoals in dit project, veel verschillende features bevat, zijn er verschillende technieken die potentieel betere resultaten kunnen behalen dan standaard logistische regressie. Dit komt vooral omdat deze modelvormen minder geneigd zijn tot overfitting en beter in staat zijn om zich te beperken op de invloed van slechts enkele - sterk voorspellende - variabelen.
Vier deelnemers wisten binnen de termijn van drie dagen voorspellingen in te dienen voor schooljaar 2015-2016. Figuur 4.3 visualiseert de behaalde AUC-scores. Het gebruik van meer geavanceerde modelleertechnieken leidt tot zichtbaar betere scores dan het simpele logistische regressie model waarvan de resultaten zijn weergegeven in het vorige figuur. Zo valt te zien dat de maximale AUC-scores van vrijwel alle zachte standaarden een maximaal behaalde AUC waarde heeft van boven de 0.7. De maximale waarde voor OR1 komt zelfs tot vlak onder de 0.9.

Figuur 4.3



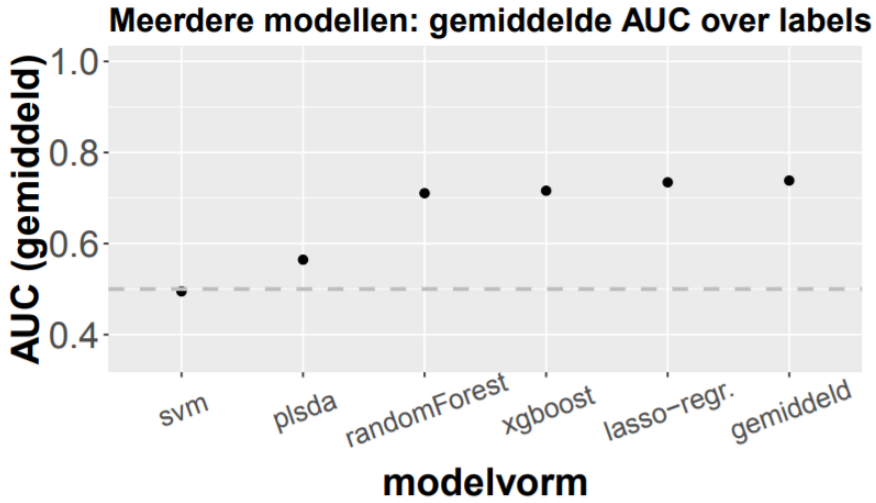
Voorspelkracht van de verschillende deelnemers op de verschillende labels van voorspellingen voor schooljaar 2016-2017. Deelnemer A gebruikte een regressiemodel met stepwise feature selectie; Deelnemer B gebruikte lasso-regressie; Deelnemer C gebruikte random forests; Deelnemer D selecteerde verschillende modellen voor de verschillende labels. Deelnemer D behaalde de gemiddeld hoogste AUC-waarden. De labels voor de gecombineerde zachte standaarden en de eindoordelen waren geen onderdeel van de Hackathon.

De deelnemers aan de hackathon hebben verschillende benaderingen gekozen wat betreft vormen van voorbewerking en modelleertechnieken. In de analyses na de hackathon zijn deze verschillende benaderingen verder uitgewerkt en waar nodig verder ontwikkeld. Hieronder worden de belangrijkste voorbewerkingsstappen en modelleervormen die hieruit voortgekomen zijn beschreven.

4.5 De voorspelkracht van verschillende modelvormen

Het succes van deelnemer D was voor een belangrijk deel te danken aan een strategie waarbij er per label een modelvorm werd gekozen die de beste prestaties leverde op subsets van de trainingsdata (in plaats van dezelfde modelvorm voor alle features, zoals gebruikt door de andere deelnemers). Het lijkt dus zinvol om deze methodiek in het vervolg toe te passen.
Bovendien is bekend dat verschillende modelvormen vaak goed zijn in het modelleren van specifieke aspecten van de data. Een veelgebruikte methode is daarom om de voorspellingen van verschillende modelvormen te combineren (een vorm van zogenaamde 'ensemble averaging'). Deze strategie is hier daarom ook onderzocht.

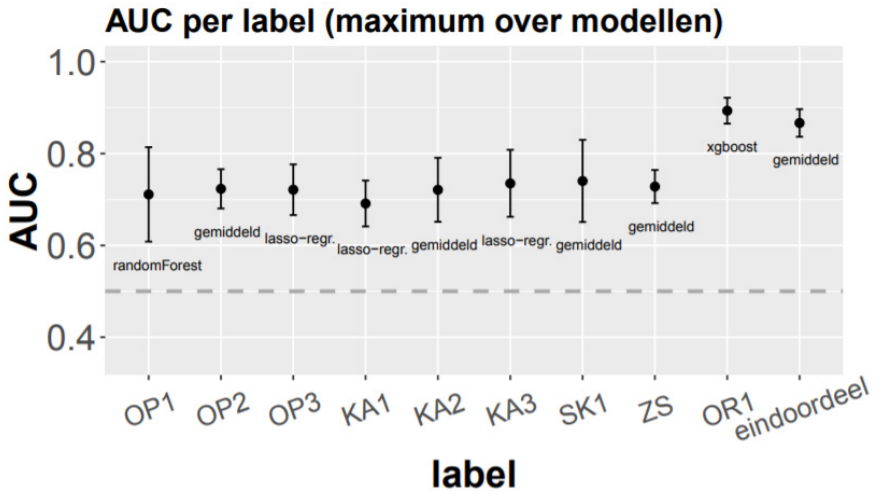
Figuur 4.4



Voorspelkracht van een reeks modelvormen. De behaalde AUC-waarden zijn voor dit figuur gemiddeld over de verschillende labels. De modelvormen zijn gesorteerd op gemiddelde behaalde AUC. De modellen zijn getraind op basis van data over 2014-2015. Voorspellingen zijn gegenereerd op basis van gegevens uit 2015-2016 en de beoordelingen van het daaropvolgende schooljaar. De voorspellingen zijn vergeleken met daadwerkelijke beoordelingen in dat schooljaar. Er worden geen error bars weergegeven omdat deze waarden gemiddelden betreffen en de betrouwbaarheidsintervallen zijn berekend over de onderliggende AUC-waarden per label.

Voor een eerste verkenning van de modelvormen starten we met de dataset zonder uitgebreide vormen van voorbewerking (wederom alleen imputatie met de globale mediaan). Een reeks modelvormen zijn getraind om de optimale relatie vast te stellen tussen data van schooljaar 2014-2015 en beoordelingen van schooljaar 2015-2016. De resulterende modellen zijn gebruikt om voorspellingen te doen op basis van gegevens over schooljaar 2015-2016. Vijf veelgebruikte modelvormen zijn toegepast: svm (Support Vector Machine); plsda (Partial Least Squares Discriminant Analysis); random forest (gecombineerde beslisbomen); xgboost (gradient boosting); en lasso-regressie. Daarnaast zijn de voorspellingen van deze 5 modelvormen gecombineerd door de voorspelling van ieder model te transformeren naar een rangorde (met gehele waarden van 1 tot het aantal scholen) en per school de gemiddelde rangorde te berekenen (wederom afgerond naar gehele getallen). De resultaten van de verschillende modelvormen zijn zichtbaar in Figuur 4.4 (zie Figuur 8.2 in de Bijlagen voor een uitsplitsing op het niveau van labels). Uit het figuur blijkt dat, met de onderzochte parameters, vooral svm en plsda geen goede resultaten behalen met deze dataset. Verder is duidelijk dat lasso-regressie vrij goede voorspelkracht biedt. Echter, de voorspelling gebaseerd op de gemiddelde rank-sortering over modellen bereikt in deze fase de beste voorspelkracht.
Het is natuurlijk ook belangrijk om te kijken naar de voorspelkracht op de verschillende labels. Figuur 4.5 laat per label de maximaal behaalde AUC-waarde zien, wederom voor voorspellingen die gedaan zijn op basis van gegevens uit schooljaar 2015-2016. Ook hier behalen vooral de eindoordelen en de standaard OR1 hoge AUC-waarden en bevinden de zachte standaarden zich rond de 0.7. Bij ieder datapunt is weergegeven welke modelvorm het maximum behaalde. Hier is wederom zichtbaar dat (met de onderzochte parameters) svm en plsda nooit tot het beste model leidden. Deze modelvormen worden daarom in het vervolg niet meegenomen en dus ook niet langer in de 'gemiddelde' voorspeller.

Figuur 4.5



Maximaal behaalde AUC-waarden per label op basis van een dataset met minimale voorbewerking. Onder ieder datapunt is weergegeven welk model deze waarde bereikte. Error-bars reflecteren 95% betrouwbaarheidsintervallen op basis van bootstrapping van de AUC.

4.6 Feature engineering

Naast het testen van verschillende modelvormen bestaat een belangrijk onderdeel van het modelleren uit het verkennen van verschillende vormen van voorbewerking en het construeren van nieuwe features op basis van inzichten (feature engineering). De belangrijkste vormen van voorbewerking die zijn verkend, hebben betrekking tot het omgaan met missende waarden; het omgaan met de correlaties tussen de features d.m.v. PCA; het construeren van nieuwe features op basis van gesprekken met experts en eigen overwegingen; het betrekken van features over meerdere jaren bij een voorspelling; en het aftoppen van het bereik van feature-waarden en z-transformatie van de features. Hieronder worden de uitgevoerde bewerkingen beschreven. Vervolgens vergelijken we het effect van de bewerkingen op de voorspelkracht van de modellen.

4.6.1 Imputatie

De dataset bevat een groot aantal features met missende waarden. De meeste van de gebruikte modelvormen kunnen hier niet goed mee omgaan. Een veelgebruikte methode om met dit probleem om te gaan, is imputatie. Hier zijn twee vormen van imputatie getest. De eerste vorm behelst het vervangen van alle missende waarden met de globale mediaan voor numerieke features en met de modus voor categorale variabelen. Deze vorm van imputatie is ook toegepast in de hierboven beschreven verkenningen. De tweede vorm bestaat uit een getrapte imputatie waarbij voor missende waarden eerst is gekeken of er gegevens van hetzelfde bestuur binnen hetzelfde schooljaar beschikbaar zijn voor imputatie; vervolgens is gekeken of er voor de overgebleven missende waarden binnen het bestuur over alle jaren gegevens beschikbaar zijn voor imputatie en uiteindelijk zijn de overgebleven missende waarden vervangen door de globale mediaan of modus van de desbetreffende feature.

4.6.2 Principale Componenten Analyse

Zoals eerder beschreven bestaan er aanzienlijke correlaties tussen verschillende features in de dataset. Veel van de hogere correlaties kunnen worden samengenomen in drie aspecten. Een eerste aspect behelst de correlaties tussen

demografische features. Deze lijken veelal samen te hangen met aspecten rond Sociaal Economische Status (SES) van de populatie leerlingen op een school en de buurt waarin een school gelegen is. Een tweede aspect staat vooral in relatie tot de behaalde referentieniveaus (lees, taal en rekenvaardigheid). De laatste staan in relatie tot een aantal features die aspecten rond vervolgsucces beschrijven. Voor ieder van deze drie aspecten zijn (apart) de groepen van features die sterke samenhang vertonen, samengenomen en is getracht door middel van Principale Componenten Analyse (PCA) de voornaamste onderliggende componenten te beschrijven met de vier sterkst verklarende componenten. Het is hierbij belangrijk te benoemen dat hiermee weliswaar een deel van de correlaties weggenomen zijn maar zeker niet alle. In een aantal gevallen lijkt het immers belangrijk om vanwege de interpreteerbaarheid de oorspronkelijke features in de dataset te behouden ondanks onderlinge correlaties (bijvoorbeeld eindtoetsresultaten in het afgelopen jaar en de eindtoetsresultaten gemiddeld over de drie voorgaande jaren).

4.6.3 Meerjaren features

Voor een reeks features zijn de waarden van voorgaande twee jaren aangeplakt aan de data van een gegeven jaar. Dit is gedaan voor de features met betrekking tot: het aantal leerlingen; de genormeerde eindtoetsscores; percentage niet-westerse migranten leerlingen; het percentage gewichtenleerlingen. Daarnaast zijn voor een aantal features de gegevens over meerdere jaren samengevoegd tot een gemiddelde waarde. Dit is gedaan voor de features met betrekking tot: het aantal leerlingen; de genormeerde eindtoetsscores; percentage niet-westerse migranten leerlingen; het percentage gewichtenleerlingen; en een schatting van het al dan niet aan- of aftreden van directieleden.

4.6.4 Expert features

Tijdens de gesprekken met inspecteurs werd onder andere een specifiek probleemscenario voor scholen geschetst. Dit behelst een situatie waarin een school te maken krijgt met een afname in leerlingaantallen en tegelijkertijd een relatieve toename in percentage niet-westerse migranten leerlingen. Bovendien zou dit vooral een probleem kunnen zijn voor kleinere besturen. Om deze dynamiek te beschrijven is een feature geconstrueerd die een signaal bevat (waarde 1) wanneer een school een relatieve toename in niet-westerse migranten leerlingen vertoont en tegelijkertijd een afname in het totaal aantal leerlingen (in beide gevallen is jaar x vergeleken met het gemiddelde over jaar x-1 en x-2).
Een andere suggestie die gedaan is, is dat vooral de KA-standaarden een relatie vertonen met sterke fluctuaties over de jaren in de financiële kengetallen zoals liquiditeit, solvabiliteit en rentabiliteit. Om deze dynamiek te beschrijven is voor alle financiële indicatoren een feature geconstrueerd die de standaarddeviatie over de drie voorgaande jaren weergeeft. Daarnaast is deze stap ook uitgevoerd voor: het aantal leerlingen; het aantal niet-westerse migranten leerlingen; het aantal gewichtenleerlingen; en de genormeerde eindtoetsscores.
Het is belangrijk op te merken dat deze bewerkingen makkelijker uitgevoerd konden worden nadat de meerjaren-features berekend waren. De effecten van deze stappen worden daarom gecombineerd beschreven.
Ten slotte is bij de expert features ook de risicoscore van de kennisanalyse toegevoegd als voorspeller. Dit is een score die bepaald is op basis van de genormeerde eindtoetsscores over de afgelopen drie jaar. Een school die drie jaar op rij onder de genormeerde score is gevallen krijgt daarbij een hoog risico.

4.6.5 Transformaties

Voor alle features zijn de extreme waarden teruggebracht tot 3 standaarddeviaties boven of onder het gemiddelde. Vervolgens is op iedere feature een z-transformatie toegepast

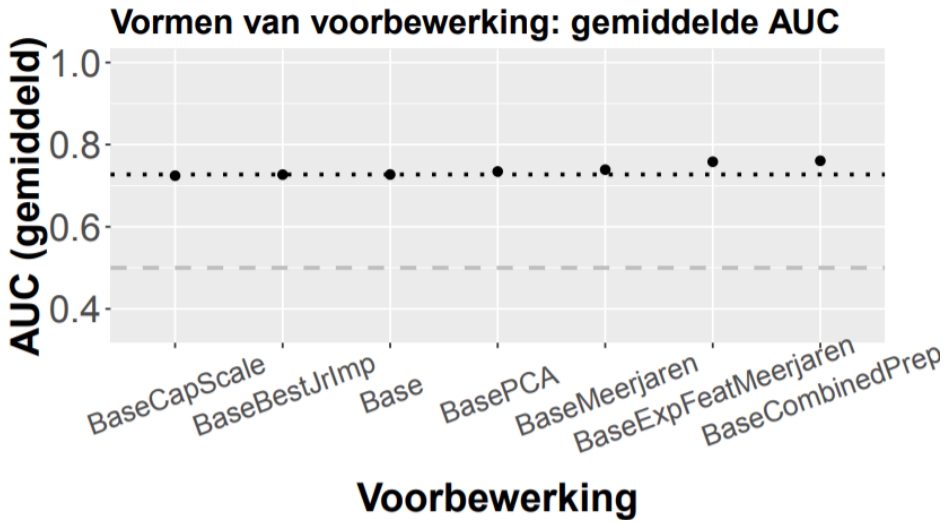
4.6.6 De effecten van feature engineering-stappen op voorspelkracht model

Om de effecten van de voorbewerking stappen inzichtelijk te maken is begonnen met een dataset met minimale voorbewerking. Vervolgens zijn voor ieder label de drie overgebleven modellen getraind en is bovendien wederom een gemiddelde modelvorm berekend. De volgende voorbewerkingen worden onderscheiden:

1) Base: slechts imputatie met de globale mediaan
2) BaseCapScale: het aftoppen en schalen van de features
3) BaseBestJrImp: getrapte imputatie
4) BasePCA: vervangen van sterk correlerende features door de 4 eerste PC's
5) BaseMeerjaren: de meerjaren features
6) BaseExpFeatMeerjaren: de expert features
7) BaseCombinedPrep: combinatie van bovenstaande bewerkingen

Figuur 4.6 beschrijft de gemiddelde AUC-waarden voor de verschillende voorbewerkingsstappen. Het aftoppen en herschalen van de features en de getrapte imputatie lijken weinig effect te hebben op de voorspelkracht vergeleken met de basisdataset. Ook het effect van PCA, de meerjaren-features en de expert-features lijken per stuk gering. Toch geeft de gecombineerde voorbewerking de hoogste AUC-waarde. In het vervolg zal daarom van deze voorbewerking worden uitgegaan.

Figuur 4.6

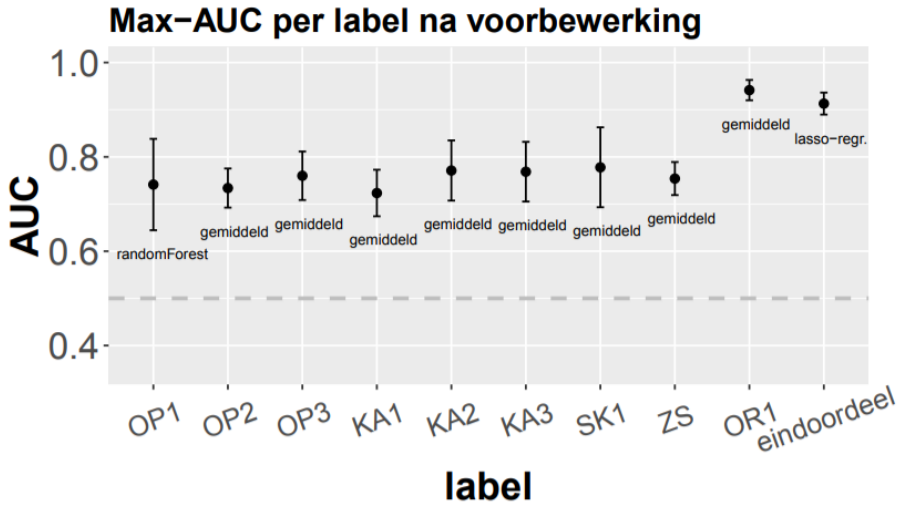


Voorspelkracht (op basis van gegevens uit 2015-2016) na verschillende voorbewerking stappen. De behaalde AUC-waarden zijn voor dit figuur gemiddeld over de verschillende labels. De modelvormen zijn gesorteerd op gemiddelde behaalde AUC. De grijze gestreepte lijn geeft kans niveau weer. De zwarte stippellijn geeft de gemiddelde voorspelkracht weer voor de basis set met minimale voorbewerking.

Figuur 4.7 visualiseert de AUC-waarden per label na de toegepaste (gecombineerde) voorbewerking. Het is hierbij overigens van belang dat bij het trainen van modellen er optimalisatie plaatsvindt door middel van het uitproberen van (deels) willekeurig gekozen parameter waarden. Dit betekent dat Figuur 4.7 er na verschillende trainingsronden net anders uit kan komen te zien (het beeld dat hier zichtbaar wordt is echter representatief voor verschillende uitgevoerde trainingsronden). Daarbij is het dus ook belangrijk om naar de betrouwbaarheidsintervallen te kijken.
Hoewel de winst uit Figuur 4.6 van de voorbewerkingen beperkt leek te zijn maakt een vergelijking tussen Figuur 4.5 en Figuur 4.7 duidelijk dat de voorspelkracht voor vrijwel alle individuele labels er op vooruit is gegaan (zie Figuur 8.3 voor een

vergelijking tussen de verschillende fases in het project). Verder is ook duidelijk dat nu nog vaker het gemiddelde model als sterkst voorspellende model naar voren komt. Dit lijkt dus voor veel van de features een goede modelvorm te zijn. Alleen OP1 en het eindoordeel komen op een andere optimale modelvorm uit. Nadere analyse heeft laten zien dat het verschil tussen lasso-regressie en het gemiddelde model voor het eindoordeel verwaarloosbaar is. Voor OP1 is dit echter niet het geval. Voor dit label komt het gemiddelde model uit op een AUC van 0.67 (0.59-0.76) terwijl het random forest een AUC behaald van 0.76 (0.67 - 0.85). Het lijkt in dit geval dus niet wenselijk om bij voorbaat voor alle labels het gemiddelde model toe te passen.

Figuur 4.7



Maximaal behaalde AUC-waarden per label op basis van een dataset met uitgebreide (gecombineerde) voorbewerking (op basis van gegevens uit 2015-2016). Onder ieder datapunt is weergegeven welk model deze waarde bereikte. Error-bars reflecteren 95% betrouwbaarheidsintervallen op basis van bootstrapping van de AUC.

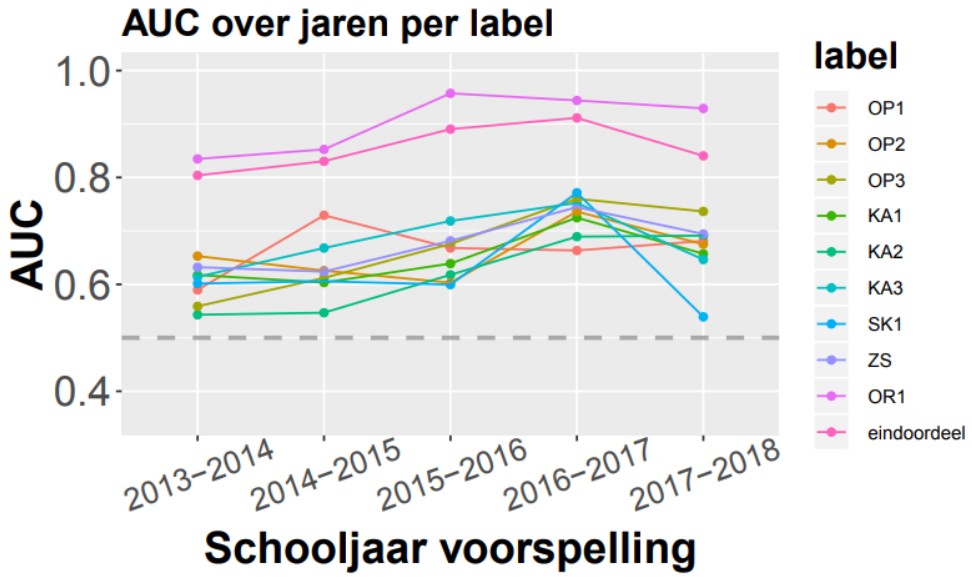
4.7 Voorspellingen voor verschillende tijdsperiodes

Tot zover zijn modellen getraind om de optimale relatie vast te stellen tussen gegevens over schooljaar 2014-2015 en de beoordelingen die in 2015-2016 zijn gegeven. Deze zijn vervolgens gebruikt om voorspellingen te genereren op basis van gegevens over 2015-2016 voor beoordelingen in schooljaar 2016-2017. Vooral schooljaar 2015-2016 heeft een goede dekking van gegevens met relatief weinig missende waarden (Figuur 3.7). Bovendien was dat ook het jaar op basis waarvan de deelnemers van de hackathon hun voorspellingen voor beoordelingen in 2016-2017 aanleverden. Het is echter ook belangrijk om inzicht te krijgen in eventuele veranderingen over de jaren van de voorspelkracht. Om dit inzichtelijk te maken zijn individuele modellen getraind voor de verschillende jaren in de dataset, waarbij de AUC-waarden steeds bepaald zijn op basis van de beoordelingen in het daaropvolgende jaar. Hierbij zijn de gecombineerde voorbewerkingsstappen uit de vorige sectie gebruikt. Modellen zijn getraind per schooljaar op basis van gegevens over de schooljaren 2011-2016.

Vooral de voorspellingen voor schooljaar 2017-2018 zijn van belang. Omdat de labels voor dit schooljaar pas na de hackathon aan de dataset gekoppeld zijn (de testset). Bovendien zijn er op basis van dit schooljaar geen exploratieve analyses uitgevoerd in het kader van het optimaliseren van voorbewerkingen en modellen. Daarmee zou deze data een relatief helder beeld moeten kunnen geven van de te verwachten voorspelkracht van getrainde modellen (kleiner risico op overfitting). Zoals eerder benoemd is er echter een belangrijk nadeel aan de data van dit

schooljaar. Het is namelijk het eerste schooljaar waarin ook de onvoldoende beoordelingen zijn toegekend op basis van het 2017-kader (zie Figuur 3.3). In dit geval moeten modellen dus over de transitie van kaders heen kunnen voorspellen. Zoals hieronder zichtbaar zal worden is het mogelijk dat dit een negatieve invloed heeft gehad op de voorspelkracht van beoordelingen voor dit schooljaar.
De meest succesvolle strategie tijdens de hackathon bestond er uit dat per label het beste model geselecteerd werd. In het vervolg wordt hier dezelfde strategie toegepast, waarbij de voorspellingen in een gegeven schooljaar gedaan worden op basis van het model dat in het jaar ervoor het meest succesvol was. Bijvoorbeeld: om voorspellingen te doen voor beoordelingen in schooljaar 2017-2018 wordt het model geselecteerd dat de beste AUC-waarden behaalde voor de voorspellingen van het jaar 2016-2017. Dat betekend dat, over de jaren, de voorspellingen van een gegeven label gebaseerd kunnen zijn op verschillende modelvormen.
Figuur 4.8 toont de behaalde AUC-waarden voor de jaar-op-jaar voorspellingen per label. Hieruit valt op te maken dat vooral in de eerdere jaren (2011-2013) de getrainde modellen geen goede voorspelkracht hebben voor beoordelingen die in de daaropvolgende jaren zijn gegeven. Dit heeft mogelijk te maken met het relatief grote aantal missende (en dus geïmputeerde) feature-waarden in die jaren (zie wederom Figuur 3.7). Bovendien kunnen voor deze data de meerjaren-features niet berekend worden en dus niet aan de voorspellingen bijdragen. De AUC-waarden zijn het hoogst voor modellen die voorspellingen doen voor schooljaar 2016-2017. Belangrijk is ook dat de voorspelkracht voor verschillende labels aanzienlijk daalt voor de voorspellingen van labels in schooljaar 2017-2018 (het laatste jaar in de reeks).

Figuur 4.8



Voorspelkracht voor de voorspellingen van de gegeven beoordelingen voor de labels over meerdere jaren. De behaalde AUC-waarden zijn gebaseerd op de voorspelling van het model dat het jaar ervoor optimaal presteerde (zie tekst). Voorspellingen zijn gegenereerd op basis van gegevens uit de voorgaande jaren en vergeleken met daadwerkelijke beoordelingen in de schooljaren op de x-as. Er worden geen error bars weergegeven in het belang van de leesbaarheid van het figuur.

Een mogelijke reden voor de achteruitgang in voorspelkracht voor het meest recente schooljaar is de transitie in kaders en de bijbehorende manier van scoren. Een andere belangrijke mogelijkheid is echter dat het evalueren van de

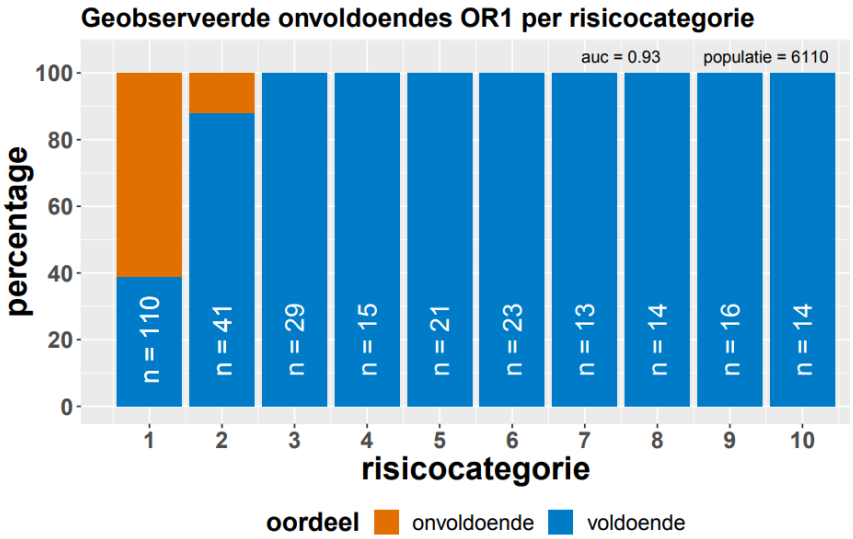
voorbewerkingsstappen wellicht tot overfitting heeft kunnen leiden op dit specifieke jaar. Dit zou betekenen dat de AUC-scores voor het laatste jaar wel degelijk een realistisch beeld schetsen van de te verwachten voorspelkracht in risicoselectie. Het is hierbij belangrijk te benoemen dat de zorg rond overfitting niet in grote mate van toepassing is voor de selectie van modelvormen. Het 'beste' model voor een gegeven label is namelijk steeds gebaseerd op de voorspelkracht van de verschillende modellen in het jaar ervoor. De verwachting is daarom dat het gevolg van overfitting beperkt zou moeten zijn. Als laatste is het ook mogelijk dat de data voor dit schooljaar van overwegend slechtere kwaliteit is. Uit Figuur 3.3 bleek in ieder geval dat er aanzienlijk minder scholen beoordeeld zijn in schooljaar 2017-2018.
In de huidige fase van het project is expliciet besloten om de beoordelingen van het jaar 2018-2019 op geen enkele manier te onderzoeken in het kader van dit project. Dit vanuit het belang om voor deze data het risico op overfitting te minimaliseren. Begin 2020 zal ook deze data beschikbaar gemaakt worden om de voorspelkracht van de modellen zo onbevooroordeeld mogelijk te kunnen inschatten. Omdat het bij deze data mogelijk zal zijn om zowel modellen te trainen en testen op data die exclusief onder het 2017 kader is verzameld zal dit vooral duidelijkheid verschaffen over de mate waarin toch overfitting heeft plaatsgevonden. Als het effect van overfitting beperkt is geweest, dan zal de voorspelkracht voor deze jaren naar verwachting vergelijkbaar zijn met die van de voorspellingen 2016-2017.

5 Resultaten

5.1 De sortering van scholen naar risico's en een indeling in risicocategorieën

Het belangrijkste doel van risicoschattingen binnen de IvhO is om een sortering mogelijk te maken over scholen, zodat de scholen met de hoogste risico's geselecteerd kunnen worden voor nader onderzoek. Het is daarom belangrijk om inzichtelijk te maken hoe de voorspellingen van voorspellingsmodellen deze sortering beinvloeden. Voor Figuur 5.1 zijn alle scholen in het regulier basisonderwijs in het schooljaar 2017-2018 ingedeeld in 10 risicocategorieen op basis van de voorspelde risicos op standaard OR1 (Resultaten). De categorieën lopen af van hoog risico (categorie 1) tot laag risico (categorie 10). De keuze voor het aantal categorieen is arbitrair, maar zodanig gekozen dat inzichteljik kan worden gemaakt of er visueel inderdaad (betekenisvolle) sortering plaatsvindt met het oog op risicoselectie.

Figuur 5.1

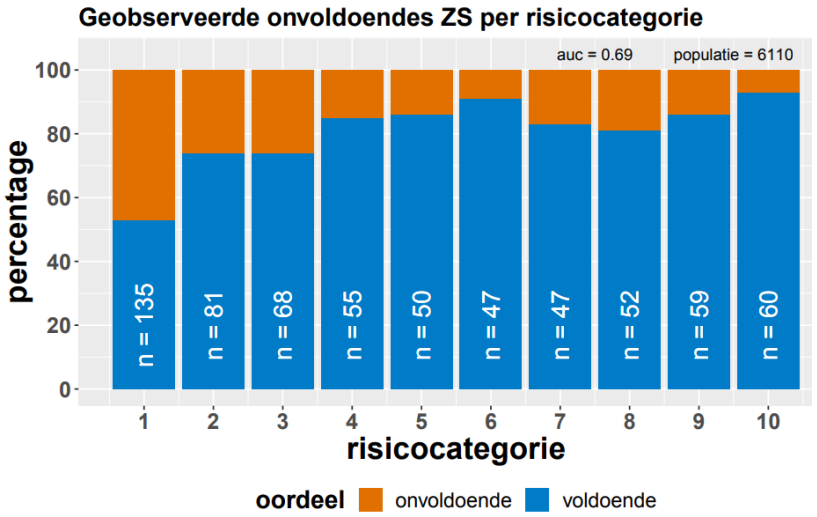


Oordelen van inspecteurs op standaard OR1, afgezet tegen voorspelde risicocategorieën (semi arbitrair opgedeeld in 10 categorieën). De sortering van scholen in risicocategorieën zijn gedaan op basis van het 'beste' model in het voorgaande jaar (in dit geval het gradient-boosting model). De onderliggende modellen zijn getraind op basis van gegevens over 2015-2016 met de beoordelingen over 2016-2017. Op basis van deze modellen en de gegevens over 2016-2017 zijn voorspellingen gemaakt voor beoordelingen in schooljaar 2017-2018 (de uiteindelijke testset). De totale populatie scholen in 2017-2018 in de dataset bestaat uit 6110 scholen (zie rechtsboven in het figuur). Per staaf is aangegeven hoeveel van de desbetreffende scholen daadwerkelijk zijn beoordeeld op OR1 door een inspecteur (witte tekst). De verdeling tussen de kleuren geeft per staaf aan welk percentage van de geïnspecteerde scholen een voldoende of onvoldoende beoordeling heeft gekregen op OR1.

Figuur 5.1 laat zien dat het voor schooljaar 2017-2018 goed mogelijk is om op basis van de modelgebaseerd voorspellingen scholen in te delen in risicocategorieën. Van de 110 onderzochte scholen in risicocategorie 1 bleek minstens 60% een onvoldoende beoordeling op OR1 te krijgen. Zoals eerder aangegeven is dit niet verassend, omdat de beoordelingen op OR1 voor een belangrijk deel gebaseerd zijn op eindtoetsresultaten (waar de IvhO gedetailleerde informatie over heeft). Deze

uitkomsten suggereren overigens ook dat de andere (niet-onderzochte) scholen in deze risicocategorieën een vergelijkbare kans op een onvoldoende op OR1 hebben. Het is verder ook zichtbaar dat de scholen in de hoogste risicocategorie in 2017-2018 ook daadwerkelijk vaker bezocht zijn waarbij een beoordeling op OR1 is gegeven (de witte getallen in de staven geven aan hoeveel scholen daadwerkelijk beoordeeld zijn op de betreffende standaard). Dit bevestigt het beeld dat het risicogericht toezicht in 2017-2018 ook sterk georiënteerd was op eindtoetsresultaten.
Vooral de beoordelingen op zachte standaarden zijn moeilijker te voorspellen. Figuur 5.2 laat de modelgebaseerde risicocategorieën voor zachte standaarden zien, afgezet tegen beoordelingen die gegeven zijn door inspecteurs in schooljaar 2017-2018. In vergelijking met het vorige figuur valt te zien dat de voorspellingen over het algemeen minder accuraat zijn. Er zitten bijvoorbeeld relatief meer scholen die een onvoldoende kregen op een van de zachte standaarden in de laag-risicocategorieën vergeleken met OR1. Desalniettemin blijkt er wel degelijk in belangrijke mate voorspellende waarde uit te gaan van de modellen voor de zachte standaarden. Van de 135 onderzochte scholen in risicocategorie 1 kreeg ongeveer 45 procent een onvoldoende oordeel. Van de 60 onderzochte scholen in de laagst-risicocategorie (categorie 10) kreeg slechts 10 procent een onvoldoende. In de appendix worden de risicocategorieën tegen beoordelingen afgezet voor alle labels voor voorspellingen voor schooljaar 2017-2018 (Figuur 8.5).
De eerdere discussie rond overfitting in acht nemend is het relevant om te benoemen dat de voorspelkracht voor de zachte standaarden in veel gevallen aanzienlijk beter is voor schooljaar 2016-2017. De verdeling in risicocategorieën voor de verschillende labels voor schooljaar 2016-2017 zijn ook toegevoegd in de appendix (Figuur 8.4). Uit de desbetreffende figuren wordt eveneens de relatie tussen de AUC score en de kwaliteit van de sortering goed zichtbaar.

Figuur 5.2



Oordelen van inspecteurs op de gecombineerde zachte standaarden (ZS), afgezet tegen voorspelde risicocategorieën voor schooljaar 2017-2018. Voor verdere details zie Figuur 5.1. Ook voor dit label was het 'beste' model uit het voorgaande jaar het gemiddelde model.

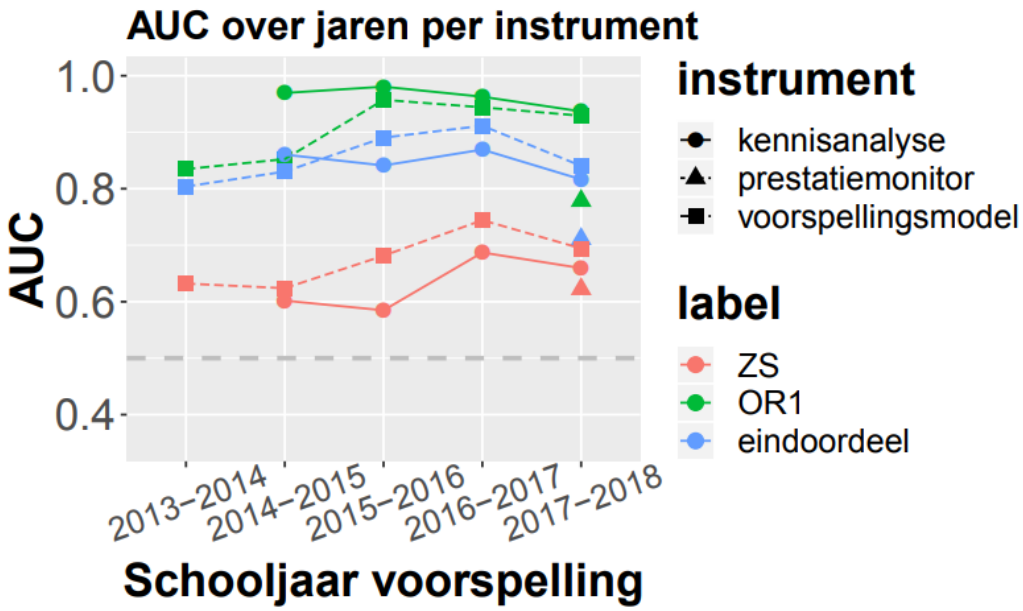
Een belangrijk bijkomend inzicht dat uit deze analyses naar voren komt is dat het voor de meeste kwaliteitsaspecten (standaarden) onmogelijk zal blijken om alle onvoldoende scholen te bezoeken, zonder ook alle scholen in het reguliere basisonderwijs te bezoeken. Gegeven de beperkte capaciteit voor risico-onderzoeken is het aannemelijk dat niet meer dan ongeveer 600 scholen op een of

andere manier onderzocht of bezocht zullen worden als gevolg van vermoedens van problemen. Dit komt ongeveer overeen met alle scholen in de eerste staaf van de figuren hierboven. Bij het beoordelen van scholen in categorie 1 zal dus naar verwachting 50% ook daadwerkelijk een onvoldoende krijgen op een of meer zachte standaarden.

5.2 Vergelijking met prestatiemonitor & kennisanalyse

Een doel van dit project was om te onderzoeken of voorspellingsmodellen tot betere risicoschattingen kunnen komen dan de huidige informatieproducten die gebruikt worden om risico's te schatten. Tot het schooljaar 2017-2018 was de zogenaamde kennisanalyse het voornaamste informatieproduct dat in het regulier basisonderwijs gebruikt werd om risicoscholen te identificeren en verder te onderzoeken. De kennisanalyse is grotendeels gebaseerd op een beslisregel. Wanneer de eindtoetsresultaten van een school drie jaar op rij onder de verwachte normscores lagen (deze scores zijn gecorrigeerd voor leerling populatie op basis van de gewichtenregeling), dan kreeg de school het label 'risico' en was het aannemelijk dat een school dat jaar bezocht zou gaan worden. Bovendien speelde deze regel ook een belangrijke rol bij de daadwerkelijke beoordeling. Met andere woorden, wanneer een school drie jaar onder de normscores zat, dan was een bezoek -en een beoordeling onvoldoende- erg waarschijnlijk (in ieder geval op OR1 en meestal ook als eindoordeel). Deze vorm van evalueren van onderwijskwaliteit was daarmee sterk gericht op onderwijsresultaten. Binnen het onderwijsveld bestond er onvrede over deze sterke focus op resultaten en klonk dan ook de roep om een manier van beoordelen die een breder palet aan kwaliteitsaspecten mee zou nemen.
In het kader van deze ontwikkeling is sinds 2017 gewerkt aan de implementatie van de opvolger van de kennisanalyse; de zogenaamde prestatiemonitor. Het doel van de prestatiemonitor is net als bij de kennisanalyse om risicoscores te genereren per school, op basis waarvan verder gericht onderzoek kan worden uitgevoerd. De risicoscore uit de prestatiemonitor is echter gebaseerd op een veel breder palet aan indicatoren, zoals veranderingen in schoolgrootte, onverwacht hoge of lage advisering (gegeven de eindtoetsen), de eindtoetsresultaten en een aantal andere aspecten. Deze indicatoren zijn grotendeels geselecteerd op basis van de expertise van inspecteurs die bij de ontwikkeling van de prestatiemonitor betrokken waren en de ontwikkelaars van het risico-instrument. De selectie was echter niet gebaseerd op uitgebreide statistische evaluatie.
Omdat ook de kennisanalyse en de prestatiemonitor risicoscores genereren kan een AUC-waarde berekend worden om de voorspelkracht van deze risico-instrumenten te vergelijken. Figuur 5.3 laat de behaalde AUC-waarden van de verschillende risicoproducten zien voor beoordelingen op de drie belangrijkste labels: OR1; de gecombineerde zachte standaarden (ZS); en de eindoordelen voor jaar-op-jaar voorspellingen. Het is hierbij echter wel belangrijk om op te merken dat de kennisanalyse en de prestatiemonitor slechts één risicoscore per school genereren. De behaalde AUC waarden voor de verschillende labels zijn voor de kennisanalyse en de prestatiemonitor dus op een en dezelfde sortering (risicoscore) gebaseerd. In dit project zijn modellen echter getraind om risicoscores voor alle onderzochte labels aparts te optimaliseren. Een directe vergelijking is daarmee ingewikkeld omdat de kennisanalyse nooit ontwikkeld is om bijvoorbeeld alleen SK1 te voorspellen. Desalniettemin maakt deze vergelijking wel inzichtelijk op welke kwaliteitsdomeinen de voorspellingsmodellen de meeste winst in voorspelkracht kunnen realiseren.

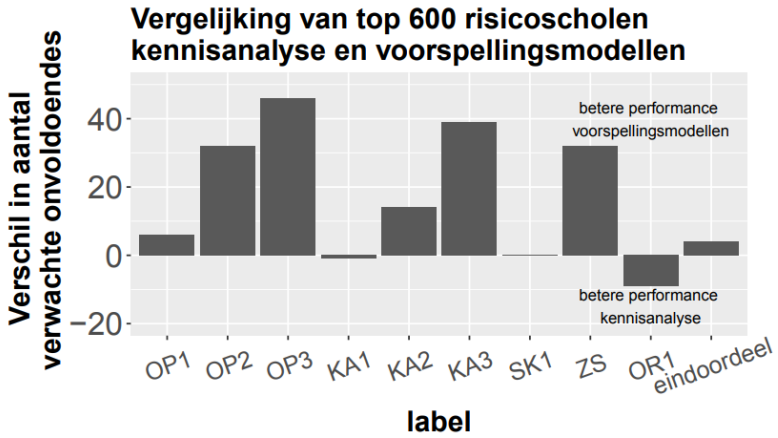
Figuur 5.3



Voorspelkracht voor de voorspellingen van de kennisanalyse, de prestatiemonitor, en voorspellingsmodellen, voor drie soorten labels (zachte standaarden [ZS], OR1 en eindoordelen). De risicoscores van de kennisanalyse en de prestatiemonitor zijn niet voor alle jaren beschikbaar (zie tekst). Wederom is het voorspellingsmodel getraind op basis van gegevens van het voorgaande jaar.

Wat opvalt, is dat de kennisanalyse (doorgetrokken lijn) voor OR1 (groene lijnen) een betere voorspelkracht behaalt dan het voorspellingsmodel (gebroken gestreepte lijn) voor de hele reeks aan beschikbare jaren. Zoals hierboven beschreven is de kennisanalyse in deze periode gebruikt als beslismodel en vooral gericht op onderwijsresultaten. Dit maakt het niet verassend dat de voorspelkracht dermate goed is. Dit maakt duidelijk dat de toegevoegde waarde van algoritmen vooral gezocht moet worden in de verklarende kracht op zachte standaarden. Uit Figuur 5.3 blijkt inderdaad dat voor de gecombineerde zachte standaarden (rode lijnen) het beeld omgekeerd is. Hier behaalt het algoritme een betere voorspelkracht dan de kennisanalyse. Ook voor de eindoordelen behaalt het voorspellingsmodel in de drie meest recente jaren een betere voorspelkracht.
Voor de prestatiemonitor zijn slechts voor het laatste jaar voorspelkracht-waarden zichtbaar. Deze scores zijn lager dan voor de kennisanalyse en het algoritme voor beide standaarden. In schooljaar 2017-2018 was de prestatiemonitor nog in ontwikkeling en het is aannemelijk dat deze waarde geen goed beeld geeft van de voorspelkracht van de huidige versie van de prestatiemonitor.
Naast het vergelijken van voorspelkracht op basis van AUC is het ook belangrijk om inzicht te verkrijgen in het effect van het gebruik van voorspellingsmodellen op de *aantallen* scholen die in de hoogste risicocategorieën vallen. Zoals eerder aangegeven kan hierbij bijvoorbeeld gekeken worden naar het aantal scholen in bijvoorbeeld de top 600 die daadwerkelijk een onvoldoende krijgen bij het gebruik van verschillende risicoproducten.

Figuur 5.4



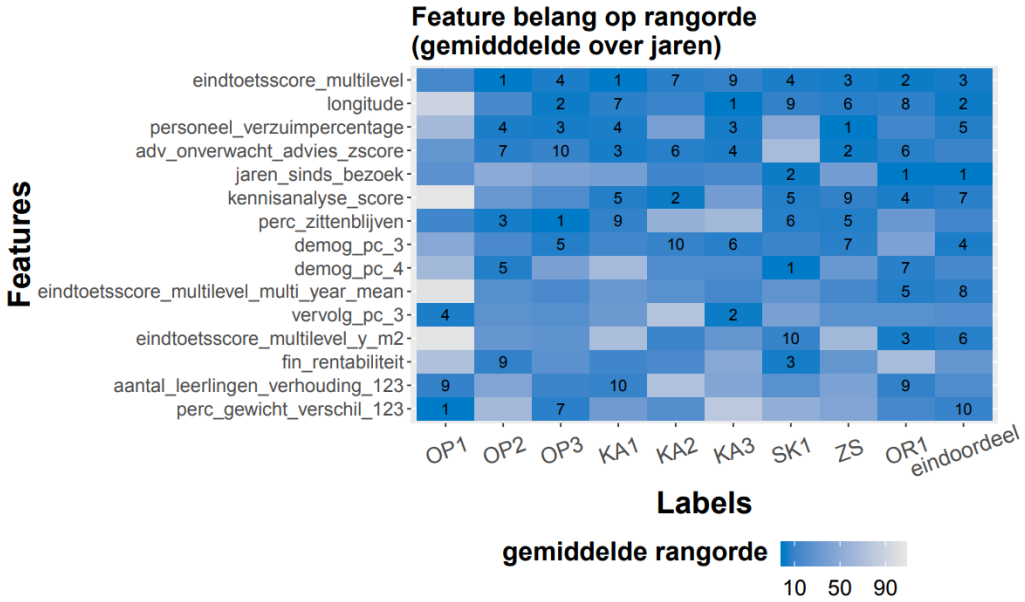
Verschil in het aantal scholen met een onvoldoende beoordeling dat in de top 600 van meest risicovolle scholen valt n.a.v. een voorspellingsmodel of de kennisanalyse (voor schooljaar 2017-2018).

Figuur 5.4 laat per standaard zien hoeveel *extra* onvoldoende scholen er in de top 600 gevallen zouden zijn bij gebruik van een voorspellingmodel in plaats van de kennisanalyse voor schooljaar 2017-2018. Het gaat daarbij om een verwachting omdat in dat schooljaar niet voor alle scholen in de top 600 daadwerkelijk alle standaarden beoordeeld zijn (een flink aantal is zelfs überhaupt niet bezocht). Ook hier is zichtbaar dat de te behalen winst bij gebruik van voorspellingsmodellen vooral te verwachten is in de schatting van risico's van de zachte standaarden.

5.3 Belangrijke voorspellers

Het gebruik van voorspellingsmodellen kan ook inzicht geven in de relatieve bijdrage van verschillende features aan de voorspelkracht. Figuur 5.5 geeft voor een select aantal voorspellers (de gemiddeld sterkst verklarende voorspellers) het relatieve belang weer in de voorspelkracht van de modellen voor de verschillende labels. Wanneer een feature in de top 10 voorspellers valt van een label, dan is de rangorde ook als getal weergegeven. Wanneer we bijvoorbeeld naar de kolom van OR1 kijken (tweede kolom van rechts) dan valt te zien dat de top 5 van voorspellers bestaat uit: 1) het aantal jaren sinds het vorige bezoek; 2) de gewogen eindtoetsscores van het voorgaande jaar; 3) de gewogen eindtoetsscores van twee jaar ervoor; 4) de risicoscore van de kennisanalyse; 5) het gemiddelde van de gewogen eindtoetsscores van de afgelopen drie jaar. Deze lijst geeft weer dat voorspellingen voor OR1 zwaar leunen op onderwijsresultaten of aspecten die daaruit voortkomen. Zo is het aantal jaren sinds het voorgaande bezoek sterk afhankelijk van de eindtoetsscores in de voorgaande jaren (bij lage scores volgde immers zeer waarschijnlijk een bezoek).

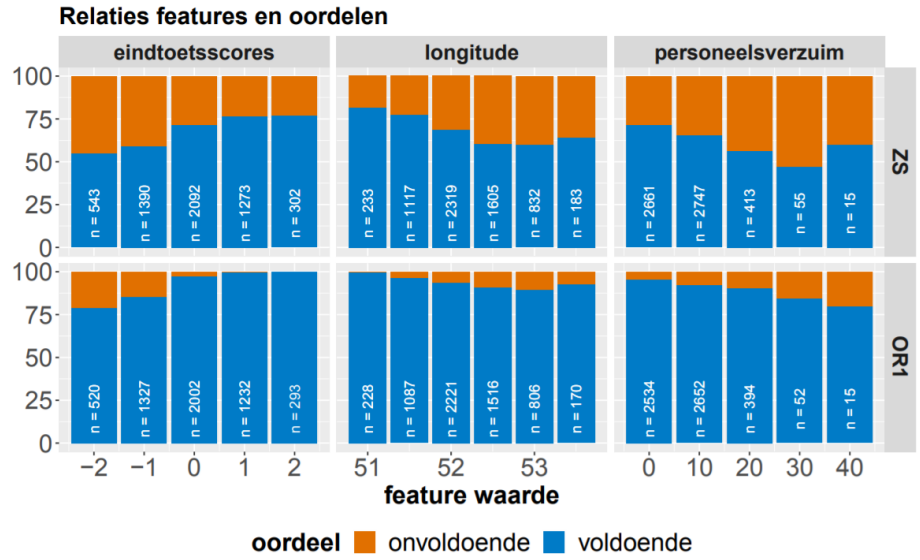
Figuur 5.5



Overzicht van het belang van de 15 sterkst verklarende features voor voorspellingsmodellen per label. Sterk verklarende features eindigen hoog in de rangorde en hebben dus een waarde richting 1. Zwak-verklarende features hebben een waarde richting 160. Rangorden per cel zijn gebaseerd op het gemiddeld belang over meerdere jaren en over drie modelvormen (lasso-regressie, random forests en gradient-boosting). Wanneer een notering in de top 10 eindigt is deze waarde weergegeven in de cel. De volgorde van de rijen geeft het gemiddelde belang (gemiddeld over de standaarden) weer, waarbij eindtoetsscore_multilevel gemiddeld de hoogste notering behaald. Het is bij de interpretatie belangrijk om in ogenschouw te nemen dat er correlaties bestaan tussen features.

Voor de gecombineerde zachte standaarden (ZS) valt te zien dat de top 5 van voorspeller bestaat uit: 1) het verzuimpercentage van leraren; 2) het geven van onverwachte vervolgadviezen door leraren (bijvoorbeeld onverwacht hoge advisering); 3) de gewogen eindtoetsscores van het voorgaande jaar; 4) de gemiddelde leeftijd van het onderwijzend personeel (niet getoond in het figuur); 5) het percentage leerlingen dat blijft zitten. Het is belangrijk om voor deze voorspellers ook te bekijken wat dan precies de relatie is tot de beoordelingen. Figuur 5.6 beschrijft de relatie tussen 3 features die gemiddeld genomen (over alle standaarden) het hoogst in de sorteringen terugkwamen: de genormeerde eindtoetsscores; longitude (noord-zuid verdeling); en ziekteverzuim onder leraren. Uit het figuur blijkt dat deze drie features inderdaad een duidelijke samenhang vertonen met de beoordelingen door inspecteurs: lage eindtoetsscores (gestandaardiseerd) gaan gepaard met relatief veel onvoldoendes; hoge longitude (meer noordelijke scholen) gaan gepaard met relatief meer onvoldoendes; en een hoog verzuimpercentage gaat gepaard met relatief veel onvoldoendes.
Verder geeft Figuur 5.5 ook voor de andere labels weer welke voorspellers relatief sterk bijdragen aan de voorspelkracht. Zo blijkt ook het percentage zittenblijvers een belangrijk rol te spelen. Daarnaast lijken ook demografische factoren een belangrijke rol te spelen (demog_pc3 & 4: demografische aspecten samengevat d.m.v. Principale Component Analyse; zie Sectie 4; Modelleren).

Figuur 5.6



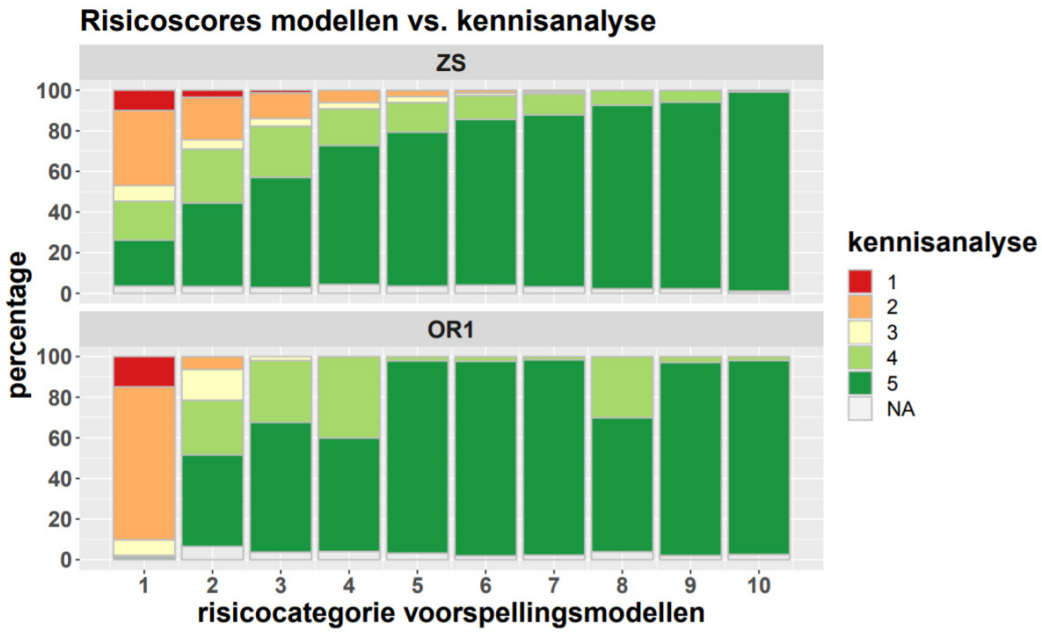
Overzicht van de relatie tussen drie belangrijke features en de beoordelingen op ZS (gecombineerde zachte standaarden; bovenste rij) en OR1 (Resultaten; onderste rij) door inspecteurs over de schooljaren 2011-2017. Ten opzichte van Figuur 5.5 is de naamgeving versimpeld (eindtoetsscores = eindtoetsscore_multilevel; personeelsverzuim = personeel_verzuimpercentage).

Hoewel deze analyse inzicht verschaft in het belang van features bestaan er twee aspecten die belangrijk zijn om te onderkennen. Ten eerste bestaan er sterke correlaties tussen verschillen de features. Zo bestaat er bijvoorbeeld logischerwijs een correlatie tussen de gewogen eindtoetsscores van het voorgaande jaar en de gemiddelde gewogen eindtoetsscores over de voorgaande drie jaar (en beide eindigden in de top 10 voor OR1). Dit betekent dat individuele features onderdeel kunnen zijn van een cluster aan samenhangende features (zie Figuur 8.6 in de bijlage voor een overzicht van de correlaties tussen de features uit Figuur 5.5). Dit is vooral het geval voor de features rond leerresultaten. De bijdragen van de verschillende features zijn dus niet per definitie onafhankelijk. Hierdoor kan in principe de individuele bijdrage van features worden onder- of overschat.
Ten tweede is het belangrijk om te onderkennen dat de voorspellende waarde van een feature gebaseerd is op de correlatie tussen een feature en de beoordelingen maar dat dit geenszins betekend dat er een causale relatie bestaat. Een sprekend voorbeeld is bijvoorbeeld de observatie dat het aantal ijsjes dat per maand verkocht wordt sterk samenhangt met het aantal doden door verdrinking. Tussen deze variabeen bestaat vanzelfsprekend geen direct *oorzakelijk* verband. Beide hangen echter samen met de gemiddelde temperatuur (een interveniërende variabele). Ook bij de relaties zoals die naar voren komen in Figuur 5.5 is het goed mogelijk dat er geen sprake is van een oorzakelijke relatie.
Desalniettemin laat het verschil in de soorten feature rangordes tussen de modellen voor OR1 en de gecombineerde zachte standaarden zien dat er wel degelijk betekenisvolle verschillen lijken te bestaan in de voorspelkracht van features tussen verschillende kwaliteitsdomeinen. Waar voorspellingen voor OR1 (zoals verwacht) vooral gedreven worden door features m.b.t leerresultaten, worden voorspellingen voor zachte standaarden vooral gedreven door meer contextuele factoren zoals ziekteverzuim en de leeftijd van leraren. Zolang deze features vooral bezien worden op basis van hun *voorspelkracht* hoeft het geen probleem te zijn dat de relatie niet oorzakelijk is.

5.4 Verschillende risicoprofielen in de kennisanalyse en voorspellingsmodellen

Uit de analyse in sectie 5.2 blijkt dat er belangrijke verschillen bestaan tussen risicoschattingen op basis van de kennisanalyse en de voorspellingsmodellen. Dit geldt vooral waar het de zachte standaarden betreft. Om verder inzicht te verkrijgen in deze verschillen laat Figuur 5.7 de relatie zien tussen risicoscores uit de kennisanalyse en risicoscores o.b.v. het voorspellingsmodel voor de gecombineerde zachte standaarden en voor OR1. Zoals zichtbaar wordt uit het figuur bestaat er voor beide labels een sterke relatie tussen de risicoscores van de twee instrumenten. Voor OR1 (het onderste paneel) is de relatie zeer sterk: vrijwel alle scholen die in de hoogste categorie van het voorspellingsmodel vallen kregen ook een risicoscore van 1 of 2 uit de kennisanalyse. Interessanter is de relatie voor de gecombineerde zachte standaarden (bovenste paneel). Van de scholen die in de hoogste risicocategorie vallen op basis van het voorspellingsmodel (hoogste 10%; de meest linker staaf) krijgt een aanzienlijk deel ook een hoge risicoscore op basis van de kennisanalyse (~50% krijgt een score 3 of lager op de kennisanalyse). Er is echter ook een aanzienlijk deel dat een lage risicoscore krijgt uit de kennisanalyse (de groene delen in de linker staaf; ~20% krijgt een risicoscore 5). Het is interessant om te onderzoeken waar de verschillen mogelijk vandaan komen.

Figuur 5.7

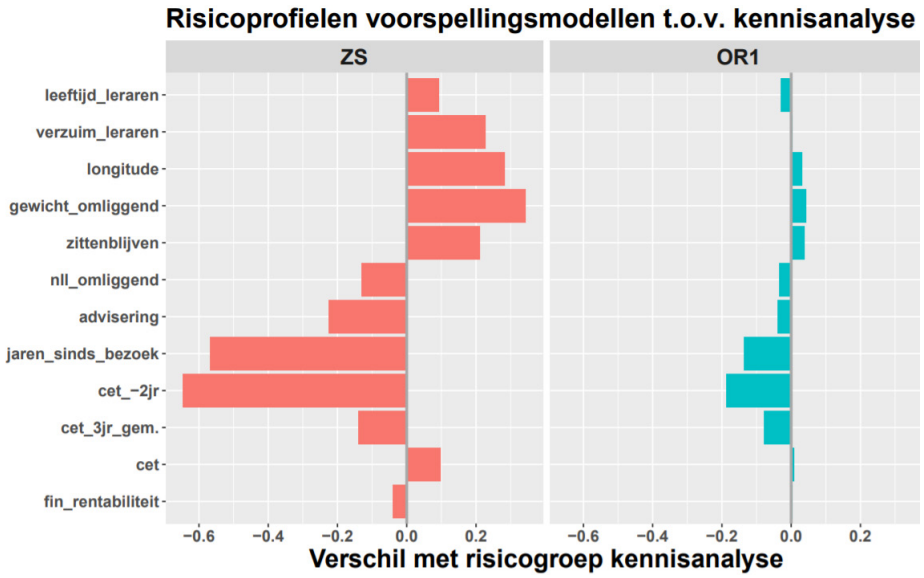


De relatie tussen risicoscores op basis van de voorspellingsmodellen en die van de kennisanalyse op basis van gegevens uit schooljaar 2016-2017. Voor beide schalen geldt een score "1" als hoog risico. Risicoscores op basis van voorspellingsmodellen zijn wederom opgedeeld in 10 groepen van gelijke grootte. De relatie is weergegeven voor risicoschattingen op de gecombineerde zachte standaarden (ZS; bovenste paneel) en OR1 (onderste paneel).

Het is aannemelijk dat de verschillen in risicoscores voor individuele scholen hun grondslag hebben in de waarden van de features die relatief belangrijk zijn in de modellen voor de gecombineerde zachte standaarden (Figuur 5.5). Dat wil zeggen, de verschillende risicomodellen herkennen waarschijnlijk verschillende "risicoprofielen" in de data. Om dit verder te onderzoeken zijn de twee risicogroepen zoals gedefinieerd uit de kennisanalyse en de voorspellingsmodellen met elkaar vergeleken door naar de verschillen in de feature distributies van de belangrijkste features te kijken.

Figuur 5.8 toont het *verschil* tussen de gemiddelde feature waarden van de risicogroep uit de kennisanalyse en de risicogroep uit de voorspellingsmodellen. De risicogroep zoals herkend door het voorspellingmodel voor zachte standaarden vertoont een aantal belangrijke verschillen met de risicogroep zoals herkend door de kennisanalyse. Voor de zachte standaarden (van boven naar onder), zijn de volgende patronen zichtbaar: Deze groep bestaat uit scholen met relatief hoge verzuimwaarden; een hoge longitude (noordelijke scholen); het gemiddelde leerlinggewicht is doorgaans relatief hoger dan dat van omliggende scholen; en de scholen hebben relatief veel zittenblijvers. Daarnaast hebbend deze scholen vaak lagere leerlingaantallen dan omliggende scholen; relatief lage advisering (onderadvisering); is er relatief kort geleden al een inspectiebezoek geweest; en is de eindtoetsscore van 2 jaar eerder relatief vaak laag. Er bestaan geen grote verschillen in de risicopopulaties wat betreft de meest recente eindtoetsscores; het 3-jaars gemiddelde van de eindtoetsen, en de financiële rentabiliteit.
Wat verder opvalt, is dat er voor de risicoschattingen voor de standaard OR1 veel kleinere verschillen bestaan tussen de risicogroepen zoals geïdentificeerd door de kennisanalyse en de voorspellingsmodellen. Het belangrijkste verschil lijkt te zijn dat in de laatstgenoemde groep de risicoscholen relatief vaak ook 2 jaar eerder een lage eindtoetsscore hadden, en dat er relatief kort geleden al een inspectiebezoek heeft plaatgevonden. Dit laat zien dat voorspellingsmodellen kunnen helpen om verschillen in risicoprofielen te identificeren, en zoals in dat geval ook met name om het verschil in het risicoprofiel t.o.v. de kennisanalyse. Figuur 8.7 (bijlage) laat de risicoprofielen voor de kennisanalyse en de voorspellingsmodellen zien naast de verdeling in de gehele populatie.

Figuur 5.8



Verschil in gemiddelde feature waarden (gestandaardiseerd o.b.v. totale populatie) tussen de risicogroep zoals geïdentificeerd door de kennisanalyse (risicoscores "1" of "2"; n = 590) en de risicogroep zoals geïdentificeerd door de voorspellingsmodellen (hoogste 10% risicoscores; n = 611) voor 12 belangrijke features (zie Figuur 5.5). De features denom_pc en vervolg_pc zijn weggelaten omdat deze lastig direct te interpreteren zijn. De kennisanalyse feature is weggelaten omdat in dit figuur juist een vergelijking met de kennisanalyse wordt weergegeven. Waarden zijn gebaseerd op gegevens uit schooljaar 2016-2017. Verschilwaarden worden weergegeven voor risicogroepen m.b.t. de gecombineerde zachte standaarden (ZS; links) en OR1 (rechts).

6 Bias en vooringenomenheid in risicomodellen

Een van de veelgehoorde zorgen over het gebruik van algoritmen in het publieke domein gaat over het bestaan van bias in algoritmen. In die context wordt bias vaak beschreven als een onwenselijke (of onrechtmatige) disbalans in het toewijzen van een label zoals 'risicovol' aan bijvoorbeeld minderheidsgroepen. Zulke bias in modellen wordt ook wel 'vooringenomenheid' genoemd (of discriminatie). De term bias wordt echter ook gebruikt binnen de statistiek, waar het eerder gezien wordt als een kenmerk van een dataset (vaak een steekproef) die een vertekend beeld geeft van de werkelijkheid (de gehele populatie): bepaalde groepen mensen kunnen bijvoorbeeld structureel over het hoofd gezien worden in steekproeven. Hoewel deze concepten sterk gerelateerd zijn, beschrijven ze verschillende aspecten. Ze zijn sterk gerelateerd omdat een statistische bias in het selecteren van de voorbeelden (in ons geval voldoende en onvoldoende scholen) ook kan leiden tot bias (vooringenomenheid) in een getraind voorspellingsmodel. Vooringenomenheid kan echter ook op andere manieren in een model terecht komen. In dit hoofdstuk zal vooral gericht worden op bias in de zin van algoritmische vooringenomenheid. Dit hoofdstuk beoogt vooral om duidelijk te maken op welke manier onderzocht kan worden of modellen mogelijk bias vertonen.

Bias is met name relevant voor kenmerken waarover er maatschappelijke consensus bestaat dat ze beschermd zouden moeten worden om discriminatie te voorkomen, bijvoorbeeld op basis van geslacht of etniciteit. Er zijn vele voorbeelden van voorspellingmodellen die gebruikt zijn/worden door publieke organen waarvan is aangetoond dat ze een vorm van bias laten zien[8,9]. Een bekend voorbeeld gaat over een systeem (COMPAS) dat in verschillende staten van de VS gebruikt wordt door rechters om risicoschattingen te maken over een verdachte (bijvoorbeeld het risico op recidive). Uit onderzoek is gebleken dat dit systeem zwarte verdachten lijkt te benadelen ten opzichte van witte verdachten[10]. Op dit moment bestaat er wetenschappelijke discussie over de beste manieren om bias in modellen inzichtelijk te maken en te corrigeren[11].

Vooralsnog bestaat er geen universele methode om bias in risicomodellen tegen te gaan. Het is daarom vooral belangrijk om *inzichtelijk* te maken in hoeverre een risicomodel verschillende groepen indeelt in verschillende risicogroepen. En als dat in de context van onderwijstoezicht zo is, in welke mate dat dan verklaard kan worden op basis van relatief objectieve criteria zoals eindtoetsscores of aantallen binnengekomen signalen. In dit hoofdstuk zullen we twee aspecten bespreken die in dit licht interessant zijn om te bekijken: 1) Mogelijke bias op basis van de migratieachtergrond van leerlingen, en 2) mogelijke bias op basis van geografie. Bij het onderzoeken van bias is het belangrijk om opnieuw te benoemen dat dit in zelflerende modellen vaak een gevolg kan zijn van het bestaan van een vorm van bias in de beschikbare features en vooral in labels waar het model op getraind is. In die zin kunnen voorspellingsmodellen ook een middel zijn om bestaande vormen van bias in een (menselijke) werkwijze aan het licht te brengen.

8 Obermeyer, Z., Powers, B., Vogeli, C., & Mullainathan, S. (2019). Dissecting racial bias in an algorithm used to manage the health of populations. Science, 366(6464), 447-453

9 Chouldechova, A. (2017). Fair prediction with disparate impact: A study of bias in recidivism prediction instruments. Big data, 5(2), 153-163

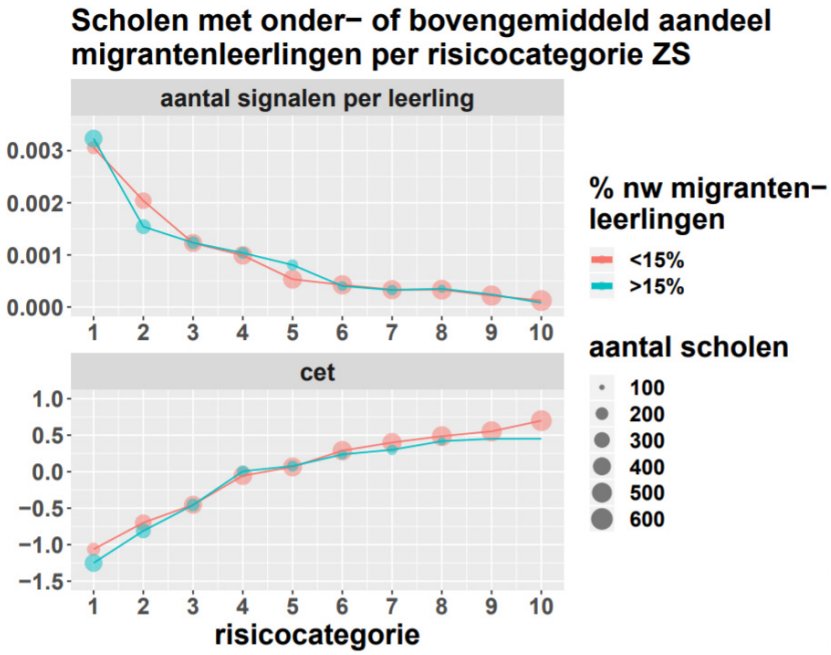
10 Angwin J et al., 'Machine bias: There's software used across the country to predict future criminals. And it's biased against blacks'. ProPublica, May 23, 2016 https://www.ProPublica.org/article/machine-bias-riskassessments-in-criminal-sentencing accessed 19 September 2018.

11 Corbett-Davies, S., & Goel, S. (2018). The measure and mismeasure of fairness: A critical review of fair machine learning. arXiv preprint arXiv:1808.00023.

6.1 Percentages niet-westerse migrantenleerlingen

In het Nederlandse basisonderwijs bestaan er grote verschillen in het percentage kinderen met een migratieachtergrond per school. We weten ook dat kinderen met een niet-westerse migratieachtergrond vaak lagere eindtoetsscores behalen, en dat scholen met relatief veel migrantenleerlingen vaak moeite hebben om goede leerkrachten aan te trekken. Desalniettemin is de feature die sterk samenhangt met het percentage niet-westerse migrantenleerlingen (de demografische pc1) niet als sterk verklarend naar voren gekomen uit de voorspellingsmodellen. Mogelijk komt dit doordat deze factor ook sterke samenhang vertoonde met de behaalde eindtoetsscores (die, zoals weergegeven in Figuur 5.5, structureel wel als sterk verklarend naar voren komt). Het kan zo zijn dat scholen met veel niet-westerse migrantenleerlingen vaker hogere risicoscores krijgen. Als dit inderdaad het geval is dan roept dat echter de vraag op of scholen met hoge percentages migrantenleerlingen *onevenredig* oververtegenwoordigd worden in de hogere risicocategorieën.

Figuur 6.1



De relatie tussen risicocategorieën uit het voorspellingsmodel voor de gecombineerde zachte standaarden, het percentage niet-westerse migrantenleerlingen per school, en objectieve criteria zoals aantallen signalen per leerling (bovenste paneel) en genormeerde eindtoetsscores (onderste paneel). Zie tekst voor verdere uitleg.

Figuur 6.1 geeft de verdeling over de risicocategorieën weer voor het voorspellingsmodel van de gecombineerde zachte standaarden. In dit figuur zijn de scholen met boven- (blauwe lijnen, n = 1763) of onder-gemiddeld aantal niet-westerse migranten leerlingen (rode lijnen; n = 4334) apart weergegeven. Het aantal scholen dat in een risicocategorie valt is weergegeven aan de hand van de grootte van de stippen. In het bovenste paneel zijn deze verdelingen afgezet tegen het gemiddelde aantal signalen per leerling in de subgroepen, en in het onderste paneel is de verdeling afgezet tegen de genormeerde eindtoetsscores. Ter illustratie: Uit het figuur valt op te maken dat op scholen in de hoogste risicocategorie

(categorie 1) het gemiddelde aantal signalen per leerling ongeveer 0.003 bedraagt (3 signalen per jaar per 1000 leerlingen; zie bovenste paneel). Hoewel scholen met een bovengemiddeld aantal niet-westerse migrantenleerlingen zijn oververtegenwoordigd in de hoogste risicocategorie (de blauwe bolletjes zijn groter dan de rode) blijkt dat er binnen deze categorie voor beide groepen gemiddeld evenveel signalen per leerling binnenkomen.
Voor de gehele verdeling in Figuur 6.1 kunnen de volgende aspecten worden waargenomen: 1) Er blijkt een sterke relatie te bestaan tussen de risicoscores en de aantallen signalen en de genormeerde eindtoetsscores. Dit is niet verrassend, aangezien deze features ook beschikbaar waren voor het model om risicoschattingen te maken; 2) Het blijkt dat scholen met bovengemiddeld (>15%) veel leerlingen met een niet-westerse migratieachtergrond relatief vaak in de hoogste risico categorieën van het voorspellingsmodel terecht komen: De blauwe bolletjes zijn doorgaans groter dan de rode richting risicocategorie 1, maar juist niet richting risicocategorie 10; 3) Zowel de analyse van de bijbehorende aantallen binnenkomende signalen en de bijbehorende (genormeerde) eindtoetsscores suggereren echter dat scholen met bovengemiddeld veel niet-westerse migrantenleerlingen niet *onevenredig* vaak in de hoogste risico categorieën terecht lijken te komen: d.w.z. de toegekende risicoscores lijken in overeenstemming met de doorgaans lagere (genormeerde) eindtoetsresultaten die de leerlingen op deze scholen behalen, en de doorgaans hogere aantallen signalen per leerling. Relatief lage eindtoetsscores en relatief hoge aantallen signalen per leerling zouden immers aanleiding moeten zijn voor verder onderzoek door analisten of inspecteurs (en het sturen van de inzet van dergelijk onderzoek is het doel van het risicoproduct).
Figuur 8.8 (bijlage) geeft een vergelijkbare analyse weer voor de risicoscores op OR1. Dat figuur schetst een deels vergelijkbaar beeld, met als uitzondering dat -gegeven de risicoscore voor OR1- de scholen met bovengemiddeld veel niet-westerse migrantenleerlingen relatief iets meer signalen binnenkrijgen, en relatief iets lagere genormeerde eindtoetsscores behalen. Het model voor OR1 lijkt daarmee een lichte bias te behelzen om scholen met bovengemiddeld veel niet-westerse migrantenleerlingen *lagere* risicoscores te geven (gegeven de eindtoetsscores en aantallen signalen).
Uit de hier beschreven analyse lijkt naar voren te komen dat -gegeven de eindtoetsscores en aantallen signalen- scholen met bovengemiddeld veel niet-westerse migrantenleerlingen niet onevenredig oververtegenwoordigd worden in de hoogste risico categorieën.

6.2 Longitude (noord-zuid verdeling)

Een ander interessant aspect in het licht van mogelijke bias betreft de verdeling over longitude (de geografische noord-zuid verdeling). Anders dan het percentage niet-westerse migranten leerlingen (vooral gevat in pc1) komt de feature longitude namelijk in de modellen *wel* vaak naar voren als een van de belangrijke indicatoren (zie bijvoorbeeld Figuur 5.5 en Figuur 5.6). Een analyse van de verdeling over risicocategorieën van scholen die in het zuiden, het midden, of het noorden van Nederland gelegen zijn (deze zijn voor de analyse opgedeeld in drie even grote categorieën) laat zien dat vooral de scholen in het zuiden relatief vaker in de lagere risicocategorieën terecht komen waar het de voorspellingen over de zachte standaarden betreft. Een analyse van de percentages binnenkomende signalen laat desalniettemin een evenredige verdeling zien over regio's binnen de risicocategorieën. Een analyse van de bijbehorende (genormeerde) eindtoetsscores suggereert echter dat de lagere risicoscores op de zachte standaarden niet in overeenstemming zijn met de gerealiseerde eindtoetsresultaten. Dat wil zeggen, gegeven de onderwijsresultaten lijkt het dat scholen in het zuiden van Nederland door het model minder streng beoordeeld worden waar het de zachte standaarden betreft. Een analyse van de risicoscores op OR1 laat *wel* een evenwichtiger verdeling

van regio's over de verschillende risicocategorieën zien (evenveel scholen uit de verschillende regio's in de verschillende risicocategorieën). Bovendien zijn de risicoscores voor de verschillende regio's in overeenstemming met de bijbehorende aantallen signalen en eindtoetsresultaten. Deze resultaten tezamen suggereren dat in het zuiden van Nederland met name de relatie tussen eindtoetsresultaten en scores op de zachte standaarden anders is dan in de rest van Nederland.
De overwegend lagere risicoscores op de zachte standaarden die door het getrainde model worden toegekend aan zuidelijke scholen zijn een gevolg van overeenkomstige verschillen in de historische dataset waar de modellen op zijn getraind (over de periode 2011-2017 zijn er voor zuidelijke scholen relatief minder onvoldoendes gegeven op zachte standaarden). De verdeling over longitude hangt binnen de inspectie nauw samen met de inspectiekantoren die betrokken zijn bij het coördineren van het toezicht op scholen: zuidelijke scholen worden overwegend overzien vanuit kantoor Tilburg; scholen in het midden van Nederland overwegend vanuit kantoor Utrecht; en scholen in het noorden van Nederland overwegend vanuit kantoor Zwolle. Het is daarom mogelijk dat de hier waargenomen verschillen voort komen uit historische verschillen in de werkwijze tussen verschillende regiokantoren. Het strekt daarom tot aanbeveling om deze patronen verder te onderzoeken, al dan niet in de context van het gebruik van voorspellingsmodellen in het toezicht.

6.3 Evaluatie van bias in risicoproducten

Bij implementatie van een voorspellingsmodel (maar ook bij de risicoproducten die op dit moment in gebruik zijn) zal het belangrijk zijn continue te monitoren of er sprake kan zijn van onevenredige verdeling van verschillende groepen over risicocategorieën, en in welke mate zulke verdelingen (on)wenselijk zijn. Dit zou ten minste een analyse behelzen vergelijkbaar met de hierboven beschreven methode. De vraag welke aspecten gecontroleerd moeten worden (moet dit bijvoorbeeld ook gaan over het percentage gewichtenleerlingen, of nog andere karakteristieken?); en de vraag of –en op welke manier- er gecorrigeerd zou moeten worden voor een disbalans, moet in overleg met domeinexperts (inspecteurs en analisten) plaatsvinden.

7 Conclusies en advies

7.1 Conclusies

In dit rapport is een beschrijving gegeven van verkennend onderzoek naar voorspellingsmodellen dat in 2019 binnen de IvhO is uitgevoerd in samenwerking met de VU. Het voornaamste doel van het project was om te bepalen of voorspellingsmodellen kunnen helpen om tot betere prioritering van risicoscholen te komen voor nader onderzoek door analisten of inspecteurs. Daartoe is onderzoek gedaan naar de toepasbaarheid (lenen de beschikbare data en modeltechnieken zich voor dit doel) van voorspellingsmodellen binnen de IvhO en is gekeken of deze modellen betere risicoschattingen kunnen maken dan de recent gangbare risicoproducten van de IvhO zoals de kennisanalyse en de prestatiemonitor.
Een belangrijke vraag die ook aan bod is gekomen in dit project is in welke fase van het risicogericht toezicht we voorspellingen willen en kunnen maken. Met andere woorden: welke vorm van risico's willen we kunnen voorspellen? Het bleek niet goed mogelijk om de uitkomsten van de expertanalyse voor dit doel te gebruiken omdat de historische verslaglegging daarvan niet consistent genoeg blijkt. Het bleek wel goed mogelijk om de eindoordelen en de beoordelingen op standaarden n.a.v. schoolbezoeken te gebruiken als afhankelijke variabelen (de labels). Op korte termijn kunnen de beoordelingen op de standaarden en de eindoordelen goed gebruikt worden om risicomodellen te trainen en evalueren. De daaruit voortvloeiende risicoscores kunnen gebruikt worden om de prioritering van scholen ten behoeve van de expertanalyses te bepalen.
Uit het onderzoek is verder gebleken dat de beschikbare features/indicatoren zich ten dele goed lenen voor het gebruik van voorspellingsmodellen. Voor dit onderzoek zijn diverse databronnen verzameld van binnen en buiten de IvhO om te gebruiken als features. Hoewel er van deze data een aantal belangrijke complicerende factoren duidelijk zijn geworden (zoals missende gegevens, gegevens op verschillende niveaus van beschrijving, etc.), is het goed mogelijk gebleken om deze gegevens zodanig voor te bewerken dat ze gebruikt konden worden in de context van dit project. Wel is het duidelijk dat er in de toekomst op dit vlak winst te behalen is door betere indicatoren te identificeren en gebruiken voor risicoschattingen. Uit het onderzoek is ook gebleken dat de faciliteiten, zoals de beschikbare methoden voor de implementatie van voorspellingsmodellen die binnen de IvhO ter beschikking stonden (er is vooral gemodelleerd binnen de open source programmeertaal R), in ieder geval voldoende zijn om risicomodellen te trainen.
Uit de resultaten komt naar voren dat het voorspellen van beoordelingen op de standaard OR1 (Resultaten) en de eindoordelen zeer succesvol is. Dit komt niet als verassing, aangezien de standaard OR1 en de eindoordelen historisch in zeer sterke mate gebaseerd zijn geweest op de eindtoetsresultaten. De grootste uitdagingen in het bereiken van succesvolle voorspellingen blijken te zitten in het voorspellen van beoordelingen op de standaarden die niet op dergelijke harde gegevens berusten. Dit betreft de overige standaarden die hier zijn onderzocht (OP1, OP2, OP3, KA1, KA2, KA3 en SK1). In dit rapport zijn deze standaarden aangeduid als de 'zachte standaarden'. Ook voor deze standaarden was het mogelijk om ruim boven kansniveau de beoordelingen te voorspellen. Het is ook vooral op deze beoordelingen waar voorspellingsmodellen naar verwachting een belangrijke rol kunnen spelen bij toekomstige verbetering in het risicogerichte toezicht.
Voor de meeste schooljaren waarvoor data beschikbaar was in dit project waren ook de risicoscores van de kennisanalyse beschikbaar. Het was daarom goed mogelijk om de kwaliteit van voorspellingen van voorspellingsmodellen en de kennisanalyse met elkaar te vergelijken. Al bleek het wel een complicerende factor te zijn dat de

kennisanalyse (het historisch gebruikte risicoproduct) niet alleen als risicoschatting is gebruikt, maar ook in belangrijke mate als beoordelingsinstrument (hetgeen het vanzelfsprekend ook een historisch goede "voorspeller" maakt). Vooral waar het de beoordelingen op de zachte standaarden betreft blijken voorspellingsmodellen echter betere risicovoorspellingen te kunnen doen dan de kennisanalyse. Een vergelijking met de prestatiemonitor bleek nog niet goed mogelijk omdat dit risicoproduct nog niet als standaard risico-instrument in gebruik was in de periode die is onderzocht. Afgelopen jaar (2018-2019) is er echter wel met dit product gewerkt en de gegeven beoordelingen kunnen binnenkort in vervolgonderzoek meegenomen worden.
Naast de beschreven verkenning bestaat een concreet resultaat uit het feit dat in 2019 het rekenmodel van de prestatiemonitor is aangepast naar aanleiding van dit project. Er is een indicator aan het model toegevoegd (verzuimpercentage onder leraren) omdat deze als sterk verklarend naar voren kwam, en een andere indicator is verwijderd (jaren sinds bezoek, omdat deze een effect had in een richting *tegenovergesteld* aan de verwachting). Daarnaast is er nog een aantal andere indicatoren geïdentificeerd die in de prestatiemonitor opgenomen zouden kunnen worden. Omdat er samenhang tussen indicatoren kan bestaan zal het echter belangrijk zijn om bij de verdere ontwikkeling van de prestatiemonitor de impact van het toevoegen/verwijderen van individuele indicatoren op de voorspelkracht te evalueren.
Ten slotte is het belangrijk om te benoemen dat het toezicht binnen de IvhO sinds enkele jaren een belangrijke transitie ondergaat waarmee het belang van goede risicoschattingen zal toenemen. Het toezicht verschuift namelijk van toezicht dat vooral gericht is geweest op individuele scholen, naar een vorm van toezicht dat zich vooral op besturen richt. In de huidige dataset bleek dan ook dat het totaal aantal schoolbezoeken waarbij oordelen worden gegeven al gestaag afneemt sinds 2014. Deze transitie is ingezet vanuit de wens om besturen nog sterker verantwoordelijkheid te laten nemen voor de kwaliteit van onderwijs van hun scholen dan voorheen. Enerzijds betekent dit dat het aantal scholen dat een officiële beoordeling krijgt in de komende jaren zal afnemen. Anderzijds blijft de IvhO vanuit haar wettelijke waarborgfunctie onverminderd verantwoordelijk op het bewaken van de ondergrens van kwaliteit, ook op individuele scholen. In deze context zal daarom het belang van adequate en tijdige risicoproducten de komende jaren enkel maar groter worden. Om dergelijke risicomodellen te creëren en evalueren zal het nodig zijn om nieuwe kwaliteitscriteria te ontwikkelen.

7.2 Adviezen

Op basis van de verkenning geven wij de volgende adviezen. Deze zijn thematisch opgedeeld maar in een aantal gevallen onderling afhankelijk.

7.2.1.1 Implementatie in de prestatiemonitor PO

Het onderzoek heeft aangetoond dat de beschikbare gegevens goed gebruikt kunnen worden om op basis van voorspellingsmodellen voorspellingen te doen over risico's op scholen. De toegevoegde waarde van het gebruik van voorspellingsmodellen ten opzichte van bestaande risicoproducten geldt vooral voor risico's op kwaliteitsgebieden waar een gebrek aan harde gegevens voor is. Een van de belangrijkste aspecten van de werkwijze rond voorspellingsmodellen is dat de voorspelkracht van verschillende potentiele risicomodellen met elkaar vergeleken wordt op basis van historische data. Het strekt tot aanbeveling om op korte termijn de resultaten van dit onderzoek te betrekken bij de verdere ontwikkeling van de prestatiemonitor PO, ten minste waar het de manier van evalueren van mogelijke risicomodellen betreft. Voor het implementeren van de resultaten van dit onderzoek bestaan grofweg twee mogelijkheden:

1) De meest invloedrijke (belangrijkste) voorspellers worden –in overleg met domeinexperts- als indicatoren toegevoegd aan de prestatiemonitor (en onbelangrijke voorspellers verwijderd). Daarnaast kan bijvoorbeeld de relatieve weging van indicatoren worden aangepast. Elke potentiele aanpassing dient geëvalueerd te worden op basis van veranderingen in voorspelkracht voor historische gegevens.
2) Een tweede mogelijkheid is om expliciet een risicovoorspelling van het algoritme mee te nemen als individuele indicator in de prestatiemonitor (een zogenaamde 'black-box' voorspeller, die overigens ook op basis van historische gegevens geëvalueerd is). Tijdens de ontwikkeling van de prestatiemonitor kan dan door de inspecteurs besloten worden in welke mate deze voorspeller meegewogen moet worden.

In eerste instantie heeft de eerste optie de voorkeur omdat deze vorm meer inzicht geeft in de indicatoren die bijdragen aan risicoschattingen. Bovendien kunnen specifieke indicatoren (zoals ziekteverzuim) vervolgens ook relatief simpel gebruikt worden voor implementatie in andere sectoren. Wanneer echter blijkt dat volgens de tweede methode structureel betere voorspellingen gedaan kunnen worden dan kan dat reden zijn om toch voor deze meer 'black box' variant te kiezen (die daarmee ook meer sector specifiek zal zijn).
Op basis van deze bevindingen is het zinvol om in samenwerking met de directie PO op korte termijn de prestatiemonitor zodanig verder te ontwikkelen dat de hier opgedane kennis gebruikt kan worden voor risicoschattingen. Voor dit advies is samenwerking nodig van de directie PO en de directie Kennis.

7.2.1.2 Verder betrekken bij, en scholen van, inspecteurs en analisten in de ontwikkeling van risicoproducten

Bij de introductie van risicoschattingen die (deels) gebaseerd zijn op voorspellingsmodellen zou het goed zijn om een brede groep van analisten en inspecteurs te betrekken bij vragen die centraal staan bij de implementatie, zoals de vorm waarin algoritmen meegenomen worden en de wenselijkheid van het gebruiken van specifieke features (zoals demografische karakteristieken). Dit kan bijvoorbeeld door trainingen te organiseren waarin er onder begeleiding interactief met verschillende varianten van de prestatiemonitor geëxperimenteerd kan worden om de invloed op de daaruit volgende voorspelkracht te laten zien. Daarbij kan gedacht worden aan varianten die van elkaar verschillen in de meegenomen indicatoren, wegingen en risicogrenzen. Dit zou er voor zorgen dat inspecteurs en analisten een beter begrip ontwikkelen voor het 'gedrag' van risicoproducten en de invloed van verschillende indicatoren op de voorspelkracht.
Op dit moment wordt er onderzocht of er binnen het inwerkprogramma van de Academie een module verzorgd kan worden met betrekking tot risicoschattingen waarmee aan dit advies vormgegeven kan worden. Voor het opzetten van deze module is in eerste instantie samenwerking nodig van de directie PO en de directie Kennis.

7.2.2 Documentatie expertanalyse

Zoals beschreven in het rapport is de historische documentatie van het deskresearch (de expertanalyse) onvoldoende gestructureerd vastgelegd om te gebruiken als labels voor de voorspellingsmodellen. Omdat het deskresearch dichter aansluit bij het gebruik van risicoproducten binnen de IvhO is voor de toekomst sterk aan te raden om deze verslaglegging meer gestructureerd en volledig te maken. Daarbij zouden analisten naar aanleiding van het deskresearch bijvoorbeeld officieuze risicoscores kunnen toewijzen aan scholen voor de verschillende kwaliteitsdomeinen (standaarden). Deze zouden dan vervolgens gebruikt kunnen worden als labels voor

de verdere evaluatie en ontwikkeling van risicoproducten. Dergelijke documentatie zou ook een rol kunnen spelen bij het mitigeren van potentieel verlies in zelflerend vermogen als gevolg van de afname in beoordelingen op individuele scholen in het bestuursgericht toezicht.
Het advies is daarmee om op korte termijn te onderzoeken hoe de werkwijze rond de expertanalyse kan worden aangepast om te komen tot meer gestructureerde documentatie van de resultaten. Voor dit advies is in eerste instantie samenwerking nodig van de sector directies SO, PO, MBO, VO, de directie Kennis.

7.2.3 Duurzaam ruimte maken voor verbetering van risicogericht toezicht.

Risico's en de beoordeling van risico's, veranderen continue. Wanneer dit bewuste veranderingen in oordeelsvorming betreft worden deze meestal vastgelegd in kaders, zoals de wisselende waarderingskaders die gebruikt zijn in het toezicht in de afgelopen jaren. Maar veranderingen in risico's kunnen ook onzichtbaar plaatsvinden. Als gevolg van grote demografische veranderingen kunnen bijvoorbeeld nieuwe risico-scenario's ontstaan, zoals het geleidelijk opkomen van de problematiek van het lerarentekort en de specifieke risico's die daaraan verbonden zijn (overbelasting van personeel?). Deze veranderingen, en vooral de manier waarop ze de kwaliteit van onderwijs beïnvloeden, zijn niet per definitie vooraf te voorspellen.
Dit continu veranderende risicolandschap vraagt om een duurzame benadering van risicoschattingen die optimaal in staat is om nieuwe ontwikkelingen in kaart te blijven brengen. Wanneer elke school ieder jaar aan een uitgebreide expertanalyse onderworpen zou worden - of zelfs bezocht-, dan zou dit probleem beperkt kunnen blijven. De IvhO gebruikt echter risicogericht toezicht, waardoor individuele scholen voor lange tijd niet uitgebreid geanalyseerd (deskresearch) of bezocht kunnen worden. Hier zijn zwaarwegende redenen voor zoals de noodzaak tot doelmatigheid en proportionaliteit van toezicht gegeven de beperkte capaciteit en de toezichtslast die scholen ondervinden. Dit betekent echter wel dat de IvhO een methodiek moet ontwikkelen die gericht is op het vinden van mogelijke 'blinde vlekken' in het risicolandschap.
Het ontdekken van nieuwe risico's (en het bijstellen van bestaande) vraagt om een afweging tussen verificatie van verwachte risico's en de exploratie van nieuwe risico's. Deze afweging is niet uniek voor de context van het toezicht bij de IvhO, maar doet zich voor in verschillende situaties waarbij men met beperkte capaciteit op zoek is naar manieren om algoritmen te verbeteren. Binnen het veld van voorspellingsmodellen bestaat de zogenaamde 'active-learning' methode[1 2]. Hierbij worden op basis van voorspellingen specifieke ongelabelde items aangedragen door het model om door een expert te laten beoordelen. Afhankelijk van de vorm van Active Learning worden deze items bijvoorbeeld geselecteerd omdat het model er het meest *onzeker* over is (en dus niet omdat het item de grootste kans heeft op een risico). Deze methode is in staat om relatief snel de risicogrenzen aan te passen en te optimaliseren. Deze methodiek zou potentieel een belangrijke bijdrage kunnen leveren aan het duurzaam in kaart blijven brengen van risico's in het onderwijs. Omdat deze methode niet gericht is op het beoordelen van risico-items per sé, past kan het geen vervanging zijn van de risicogerichte onderzoeken (het is immers onwenselijk dat de IvhO scholen verzuimt te onderzoeken of bezoeken, omdat het model zeer *zeker* is van een onvoldoende). Het zou echter wel mogelijk zijn om deze vorm van onderzoek naast de reguliere risico-onderzoeken uit te voeren. De Active Learning methodiek zou bijvoorbeeld toegepast kunnen worden in een speciaal daarvoor opgezet (jaarlijks) thema-onderzoek. In dit onderzoek zouden analisten en of inspecteurs dus gevraagd worden om scholen uitgebreid te analyseren of te

[1 2] Settles, Burr (2010). "Active Learning Literature Survey". Computer Sciences Technical Report 1648. University of Wisconsin–Madison: http://burrsettles.com/pub/settles.activelearning.pdf

bezoeken omdat we op basis van de beschikbare gegevens *onzeker* zijn over de risico's op specifieke kwaliteitsdomeinen.
Om deze manier van werken vorm te geven stellen wij voor om in eerste instantie een verkenning uit te voeren binnen de sector PO.

7.2.4 Verbreding van beschikbare indicatoren

De belangrijkste bron voor mogelijke verbeteringen in voorspelkracht van risicomodellen zit in het vinden van betere risico-indicatoren. Vooral in de sector PO speelt het probleem dat robuuste gegevens over leerresultaten pas in groep 8 verkregen worden bij het maken van de eindtoetsen. Het lijkt belangrijk om te onderzoeken of hier al eerder gegevens over beschikbaar zouden kunnen komen.
Een ander aspect aan risico's bij scholen dat veel door inspecteurs genoemd is behelst de 'kwaliteit van bestuurders'. Veel inspecteurs gaven aan dat een goede bestuurder een relatief zwakke school er weer bovenop kan helpen. Op dit moment heeft de IvhO echter geen gegevens die inzicht verschaffen in de ervaring en of kwaliteit van schoolbestuurders. Daarnaast zouden ook gegevens over maatschappelijke waardering van scholen (bijvoorbeeld op basis van social media) een rol kunnen spelen.
Het zal daarom in de toekomst belangrijk zijn om samen met inspecteurs, analisten en andere onderwijsexpert zoals schoolbesturen actief op zoek te gaan naar mogelijke nieuwe informatiebronnen. Voor dit advies is in eerste instantie samenwerking nodig van de sector directies SO, PO, MBO en VO en de directie Kennis.

7.2.5 Verkenning van voorspellingsmodellen in andere sectoren

Uit dit project is gebleken dat voorspellingsmodellen kunnen bijdragen aan het schatten van risico's bij PO-scholen. Gegeven deze uitkomst is het zinvol om te verkennen of voorspellingsmodellen ook bruikbaar kunnen zijn voor andere onderwijssectoren waar de IvhO toezicht op houdt. Een voor de hand liggende sector zou de sector Voortgezet Onderwijs (VO) zijn, vanwege de omvang en het datagedreven karakter van het toezicht op die sector. Daarnaast zou echter ook prioriteit gegeven kunnen worden aan risicoschattingen op bestuursniveau om daarmee snel aan te sluiten op de transitie naar bestuursgericht toezicht binnen de IvhO.
Voor dit advies is samenwerking nodig van de directie VO en de directie Kennis.

7.2.6 Ethische kaders voor het gebruik van algoritmen in het toezicht

Er bestaat een toenemende maatschappelijke zorg over het gebruik van "algoritmen" door de overheid[13]. Recent berichtte de NOS bijvoorbeeld dat de overheid op grote schaal voorspellende algoritmen gebruikt, zonder dat dat altijd even duidelijk voor burgers is, terwijl deze algoritmen wel negatieve effecten kunnen hebben. Naar aanleiding van deze berichtgeving spraken politieke partijen (o.a., D66, CDA) de wens uit om strengere regels op te stellen voor het gebruik van algoritmen door de overheid. Ook werd aangegeven dat er een richtlijn moet komen die bepaalt in welke gevallen gebruik van algoritmen gerechtvaardigd is en dat er een waakhond zou moeten worden aangewezen die daar toezicht op houdt[14].
Het betrekken van inspecteurs, analisten en mogelijk andere belanghebbenden in het onderwijsveld bij het besluiten over de manier waarop algoritmen gebruikt kunnen worden in risicosignalering vormt een zeer belangrijke eerste stap om met deze zorgen om te gaan. Daarnaast is het ook belangrijk om transparant te communiceren over deze ontwikkelingen en om actief samenwerking op te zoeken

13 bijv., https://nos.nl/artikel/2286848-overheid-gebruikt-op-grote-schaal-voorspellende-algoritmes-risico-op-discriminatie.html
14 https://nos.nl/artikel/2289495-d66-en-cda-willen-richtlijn-en-toezichthouder-voor-overheidsalgoritmes.html

met organisaties die zich bezig houden met het verantwoord gebruik van algoritmen bij overheden.
Voor dit advies is in eerste instantie inspanning nodig van de directie Kennis.

7.2.7 Beknopte adviezen:

1a) Gebruik de resultaten van deze verkenning voor de verdere ontwikkeling van de prestatiemonitor PO, vooral waar het risico's op de zachte standaarden betreft.

1b) Organiseer een discussie met inspecteurs en analisten over de effecten van (het gebruik van) verschillende indicatoren in de prestatiemonitor op de voorspelkracht als risicomodel.

2) Verbeter de documentatie van de expertanalyse binnen de verschillende sector directies van de IvhO.

3) Onderzoek het inzetten van specifieke jaarlijkse themaonderzoeken ('Active Learning') met als doel het continue verbeteren van risicoschattingen.

4) Investeer in het vinden van nieuwe indicatoren door het beschikbaar maken van nieuwe databronnen. Het identificeren van mogelijke indicatoren kan bijvoorbeeld door gerichte gesprekken met verschillende onderwijsexperts.

5) Onderzoek het gebruik van voorspellingsmodellen voor risicoschattingen binnen andere onderwijs sectoren, bijvoorbeeld VO, of risicoschattingen voor besturen.

6) Zorg voor transparantie over het gebruik van algoritmen binnen de IvhO en zoek samenwerking met organisaties die gespecialiseerd zijn in het verantwoord gebruik ervan.

8 Bijlagen

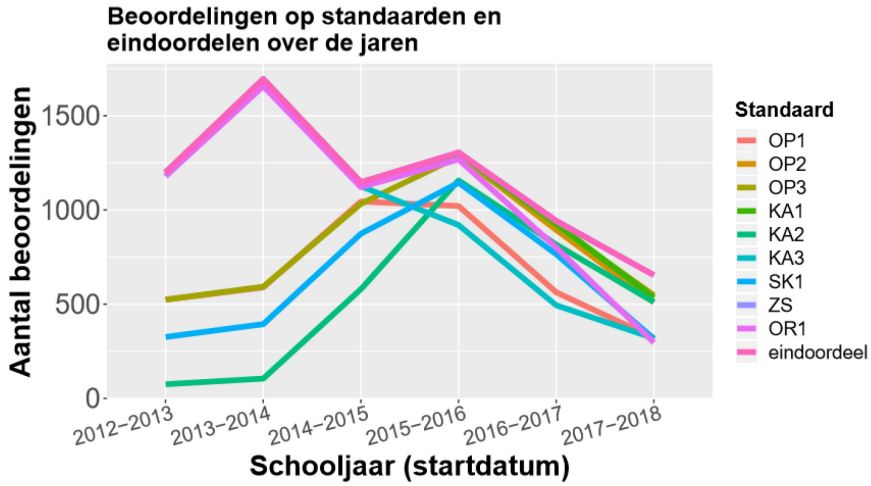
Tabel 8.1

Structuur van het omcoderen van de onderzoekskaders

	waard. Kad. Pil. 2020	Voorl. Pil BST PO (ovt's)	oude kaders	
standaard	indicatorcodes	indicatorcodes	indicatorcodes	Norm-indicatorcodes
OP1	2.1	2.1	2.7, 2.5, 2.6, 2.8, 2.3,	2.1, 2.2, 2.4
OP2	2.2	2.2	7.2, 8.1, 7.6, 8.2	7.1, 8.3
OP3	2.3	2.3	5.4, 3.2, 3.1, 6.1, 6.2, 6.3, 6.4, 5.8	5.2, 5.3, 5.1
KA1	4.1	5.1	8.4	9.4, 9.5, 9.1, 9.2, 9.3
KA2	4.2	5.2	10.1, 10.3, 10.2, 4.11, 4.9	
KA3	4.3	5.3	10.4, 8.6	9.6
SK1	3.2	4.1	4.2, 4.4, 4.5, 4.6	
OR1	1.1	1.1	1.4	1.1

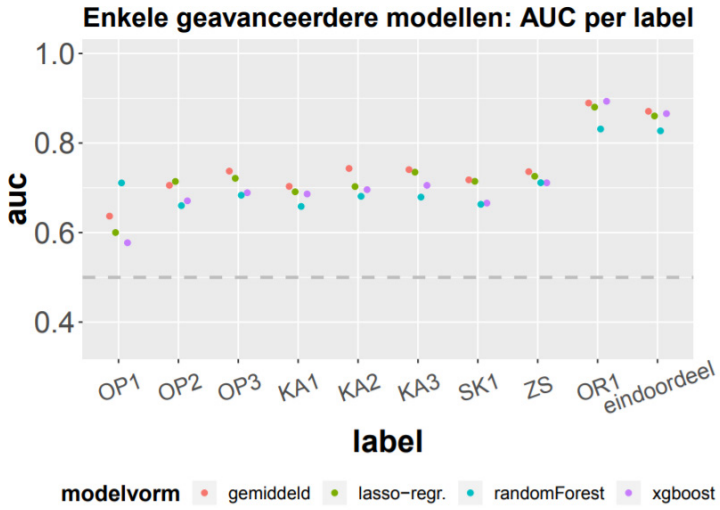
Het doel van het omscoren was om oude oordelen te beschrijven volgens de indeling in standaarden zoals toegepast in de huidige kaders (PO-kader BGT Onderzoekskader 2017; en PO Waarderingskader 2017). Voor de set van meer recente kaders (de Pilots) is slechts de code aangepast om aan te sluiten op de standaarden zoals gebruikt in het 2017 kader. Dit behelst de kaders: Waarderingskader pilots tz 2020; en Voorlopig wk pilots bestuurstoezicht PO [ovt's](2e en 3e kolom). Het omscoren van de oude kaders was complexer. Dit behelst de kaders: Basisset regulier PO 2012; Basisset 4JB 2012; Basisset Stelselonderzoek 2012-2013; Overige kernindicatoren vierjaarlijks bezoek po; Overige kernindicatoren stelselonderzoek 2012-2013; Uitbreidingsset 2012. Voor deze omcodering zijn scores op verschillende indicatoren omgezet naar een numerieke code (0: Onvoldoende; 1 Voldoende/Goed) en vervolgens zijn de scores samengevoegd d.m.v. middeling en afgerond tot oordelen op standaarden. Bij het omscoren van deze kaders hebben zogenaamde "normindicatoren" een dubbele weging gekregen (zie 4e en 5e kolom van de tabel).

Figuur 8.1



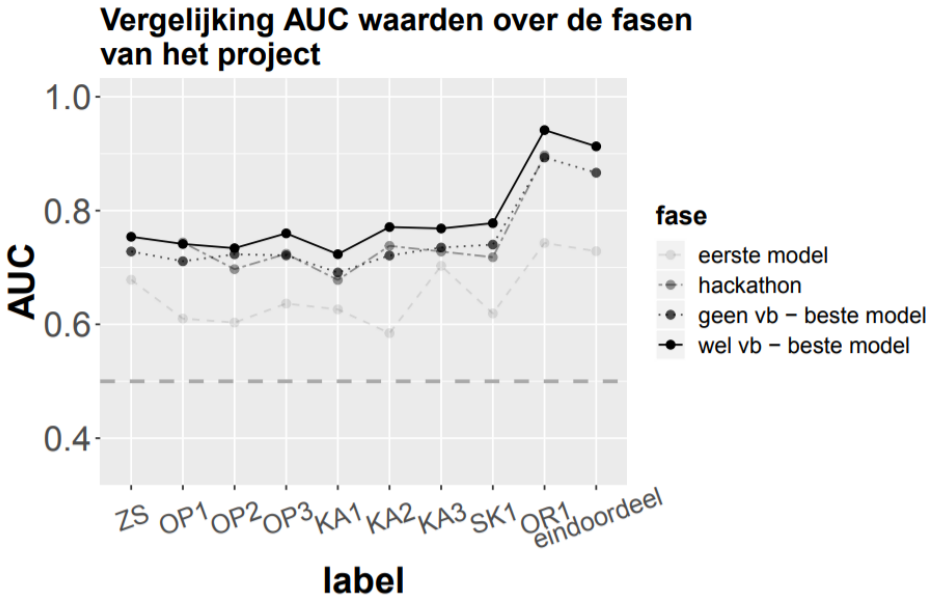
Aantallen beoordelingen per standaard over de onderzochte schooljaren. Wat opvalt, is dat vooral OP1; OP3; KA2; en SK1 in de jaren tussen 2012-2014 (het oude kader) in verhouding weinig beoordeeld zijn. Dit is waarschijnlijk een gevolg van de manier van hercoderen van de kaders. In de meer recente jaren (2016-2018) zijn OP1; KA3; SK1; en OR1 in verhouding weinig beoordeeld.

Figuur 8.2



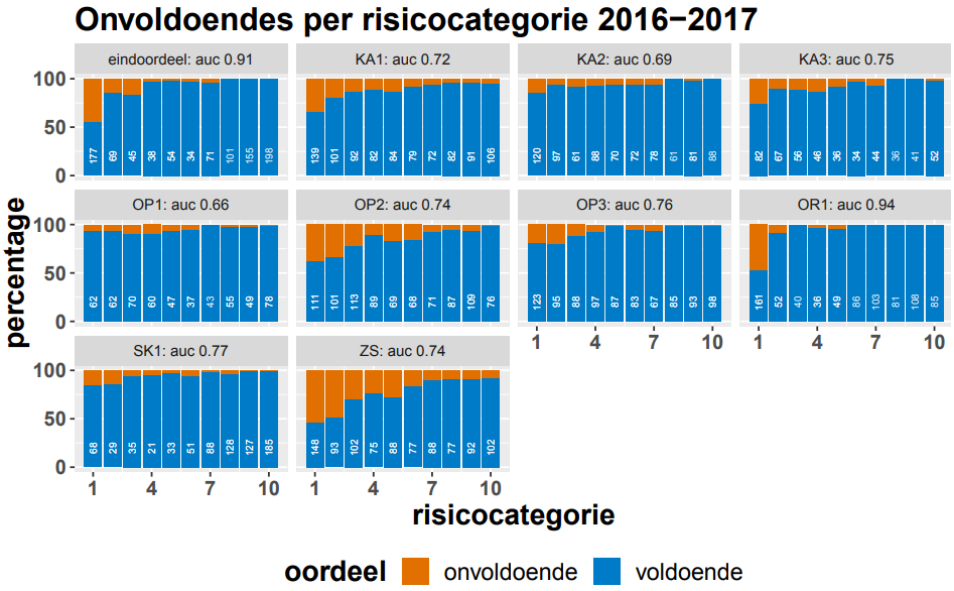
Voorspelkracht van een reeks modelvormen per label. De modellen zijn getraind op basis van data over 2014-2015. Deze dataset heeft geringe voorbewerking ondergaan (imputatie met de globale mediaan). Voorspellingen zijn gegenereerd voor 2015-2016 en vergeleken met daadwerkelijke beoordelingen in dat schooljaar. Voorspelling op kans niveau geeft een AUC van 0.5.

Figuur 8.3



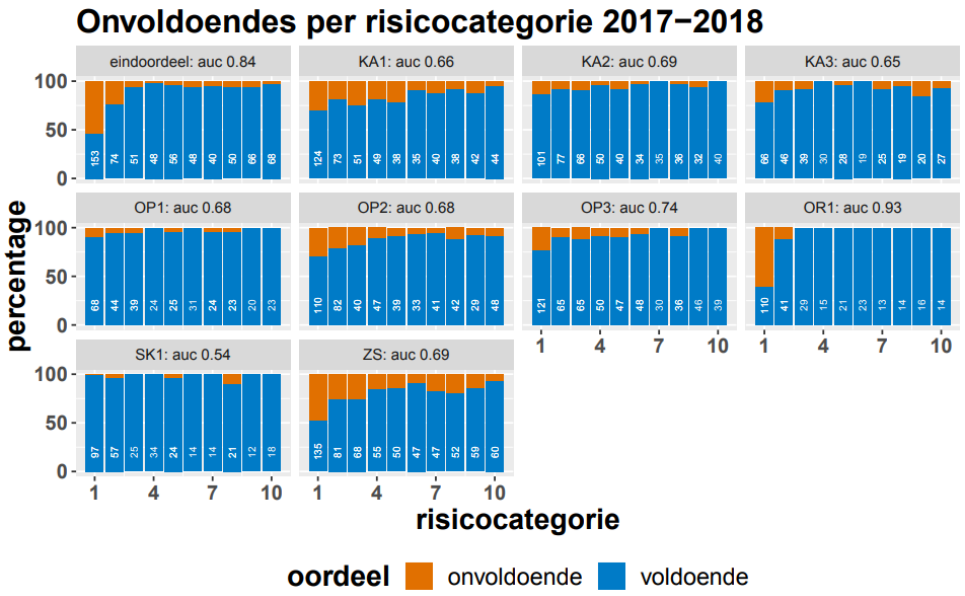
Voorspelkracht over de fases van het project. Voor de resultaten van de hackathon is per label de maximale voorspelkracht gekozen. Vergeleken met een simpel (lineaire regressie) model neemt de voorspelkracht aanzienlijk toe bij het gebruik van complexere modelvormen die vormen van feature selectie toepassen. Daarnaast heeft de uitgebreidere voorbewerking tot een verdere toename in voorspelkracht geleid.

Figuur 8.4



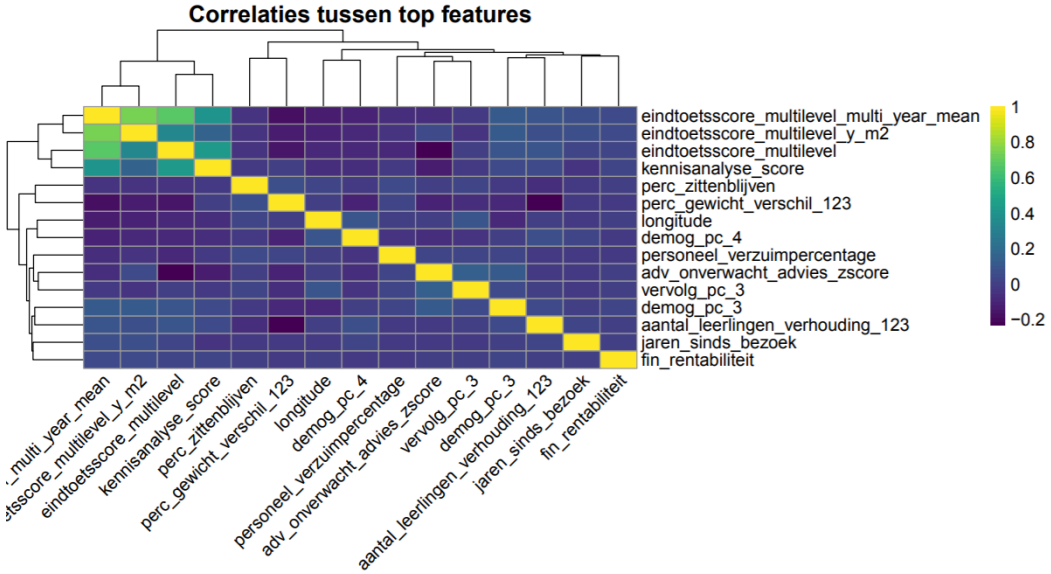
Oordelen van inspecteurs op de verschillende labels, afgezet tegen voorspelde risicocategorieën voor schooljaar 2016-2017. Voor verdere details zie Figuur 5.1.

Figuur 8.5



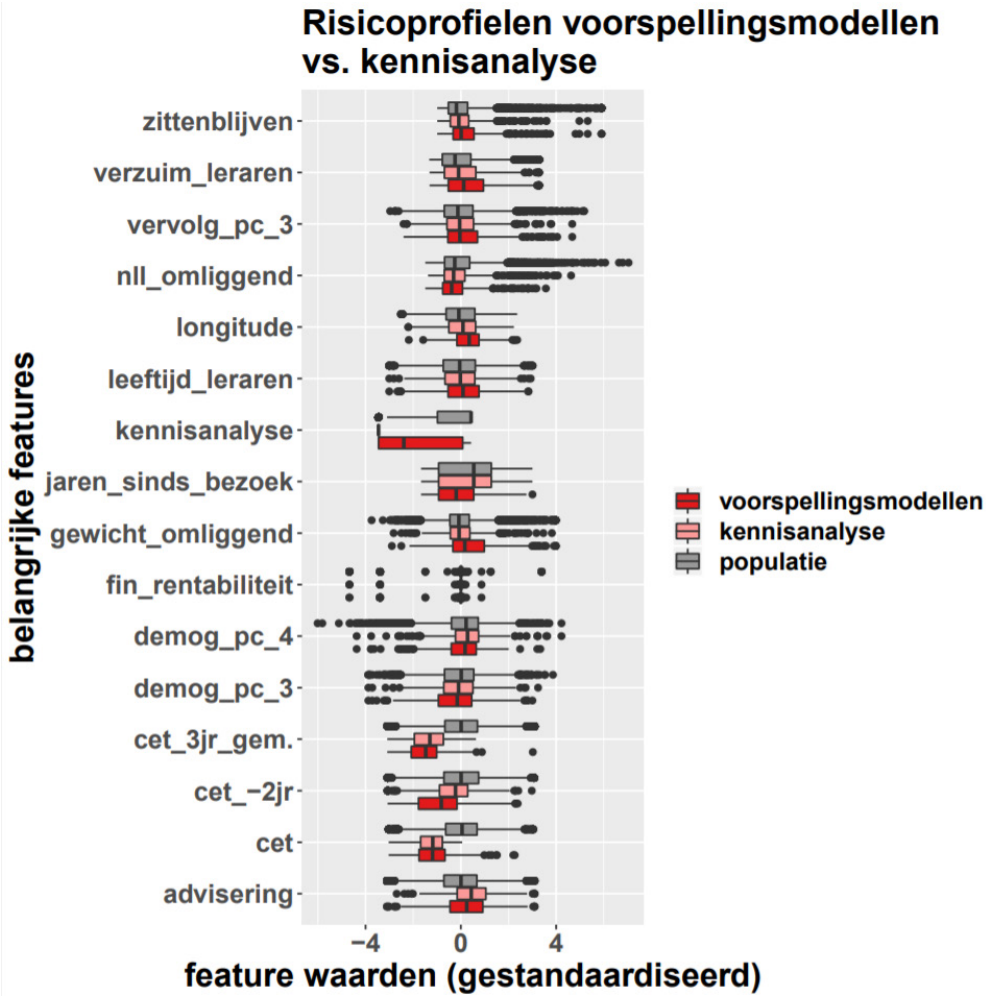
Oordelen van inspecteurs op de verschillende labels, afgezet tegen voorspelde risicocategorieën voor schooljaar 2017-2018. Voor verdere details zie Figuur 5.1.

Figuur 8.6



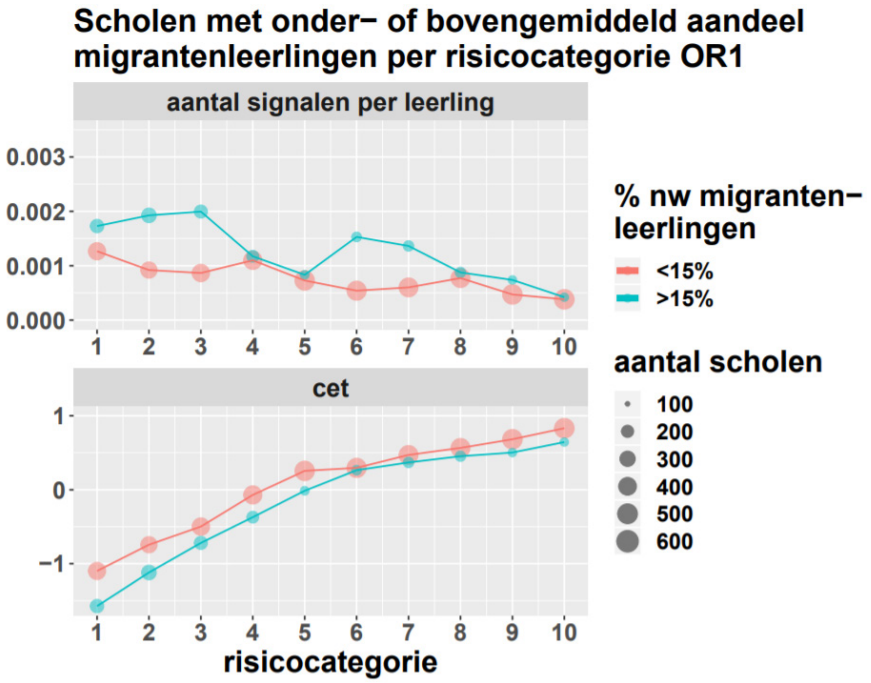
Gegroepeerde correlaties tussen de 15 belangrijkste features

Figuur 8.7



Risicoprofielen op basis van de kennisanalyse en de voorspellingsmodellen voor de gecombineerde zachte standaarden. Feature waarden voor de 15 belangrijkste features (zie Figuur 5.5) + de gemiddelde leeftijd van leraren (een belangrijke voorspeller voor ZS). Feature waarden worden apart weergegeven voor scholen in de dataset die in 2016-2017 een score 1 of 2 behaalden op de kennisanalyse (n = 590); de scholen die in de 10% hoogste risicocategorie vallen op basis van het voorspellingsmodel (n = 611); en de populatie als geheel (n = 6110).

Figuur 8.8



De relatie tussen risicocategorieën uit het voorspellingsmodel voor OR1, het percentage niet-westerse migrantenleerlingen per school, en objectieve criteria zoals aantallen signalen per leerling (bovenste paneel) en genormeerde eindtoetsscores (onderste paneel). Zie tekst voor verdere uitleg. Voor dit figuur is gebruik gemaakt van het 'gemiddelde' model.

Colofon

Inspectie van het Onderwijs
Postbus 2730 | 3500 GS Utrecht
www.onderwijsinspectie.nl

Een exemplaar van deze publicatie is te downloaden vanaf de website van de Inspectie van het Onderwijs: www.onderwijsinspectie.nl.