Pilot opleidingsvereisten in vacatureteksten

Een onderzoek met behulp van machine learning

CBS Den Haag
Henri Faasdreef 312
2492 JP Den Haag
Postbus 24500
2490 HA Den Haag
+31 70 337 38 00

www.cbs.nl

projectnummer PR000744

28 april 2022

Inhoudsopgave

1. Inleiding

1.1 Aanleiding

Het ministerie van Onderwijs, Cultuur en Wetenschap (OCW) is verantwoordelijk voor een goede aansluiting tussen het onderwijsaanbod en de arbeidsmarkt. Een van de vragen die hierbij speelt is of bedrijven en instellingen meer behoefte hebben aan breed opgeleide mensen of juist vaker vragen om mensen met een specifieke opleiding. En hoe dit zich heeft ontwikkeld in de loop van de tijd. Zo is bijvoorbeeld de vraag naar iemand met een technische opleiding 'algemeen', terwijl de vraag naar de opleiding werktuigbouwkunde 'specifiek' is. Deze vraag speelt zowel bij de directie hoger onderwijs (ho) als de directie middelbaar beroepsonderwijs (mbo). Kennis hierover kan worden gebruikt voor de inrichting van de curricula van opleidingen om de aansluiting met de arbeidsmarkt te verbeteren. Een manier om te achterhalen wat de opleidingsvereisten zijn in de arbeidsmarkt is te kijken naar vacatureteksten. Het Centraal Bureau voor de Statistiek (CBS) heeft sinds een aantal jaar de beschikking over vacatureteksten afkomstig van de website Werk.nl van het Uitvoeringsinstituut Werknemersverzekeringen (UWV). In overleg tussen het ministerie van OCW en het CBS is toen afgesproken om te onderzoeken of het mogelijk is om met de behulp machine learning en tekstmining technieken een beeld te krijgen van de opleidingsvereisten op basis van deze vacatureteksten.

Dit onderzoek is onderdeel van het OCW, DUO en CBS convenant 2020-2021. Voor dit convenant zijn er maatwerkafspraken gemaakt tussen CBS en OCW waarin het werk is vastgelegd dat CBS in 2020 en 2021 in opdracht van OCW uitvoert. Binnen deze overeenkomst is sprake van doorlopende werkzaamheden en aanvullende projecten. Om nieuwe onderzoeksmethoden te verkennen was de wens dat een van deze aanvullende projecten om een pilot onderzoek met Big Data zou gaan. Dit onderzoek is de uitwerking hiervan.

1.2 Doel

Het doel van het onderzoek is tweeledig:

- Onderzoeken of met behulp van big data technieken een concrete beleidsvraag kan worden beantwoord. In dit geval gaat het om de vraag in hoeverre de op de arbeidsmarkt gevraagde opleidingen in de loop van de tijd veranderen.
- Ervaring op doen met big data-analyse en de mogelijkheden die dat biedt voor statistische vragen op het de beleidsterreinen van OCW. De opgedane inzichten kunnen door OCW en CBS gezamenlijk worden benut voor toekomstige vraagstukken.

1.3 Onderzoeksvragen

Deze doelen zijn vertaald naar de volgende specifieke onderzoeksvragen die we in dit onderzoek proberen te beantwoorden:

- Zijn opleidingsvereisten in vacatureteksten algemeen of specifiek?
- In hoeverre is de verhouding tussen algemene of specifieke opleidingsvereisten veranderd over de tijd?
- Is er verschil tussen het mbo en ho in het aandeel algemene of specifieke opleidingsvereisten?
- Zijn er verschillen tussen beroepsrichtingen in het aandeel algemene of specifieke opleidingsvereisten?

1.4 Pilot onderzoek

Dit is een experimenteel onderzoek dat nog niet eerder op deze manier is uitgevoerd. De uitkomsten van deze pilot zijn dan ook vooral bedoeld om een beeld te krijgen van de mogelijkheden om via tekstmining technieken de opleidingsvereisten voor beroepen in vacatureteksten te identificeren en niet om al daadwerkelijk uitspraken te over de (ontwikkeling van) opleidingsvereisten zelf. Om hier betrouwbare uitspraken over te kunnen doen zijn nog meer vervolganalyses nodig. Binnen de beperkte tijd en capaciteit die was gereserveerd voor deze pilot was geen ruimte voor deze vervolganalyses. Het is belangrijk hier rekening mee te houden bij het interpreteren van de uitkomsten.

2. Data

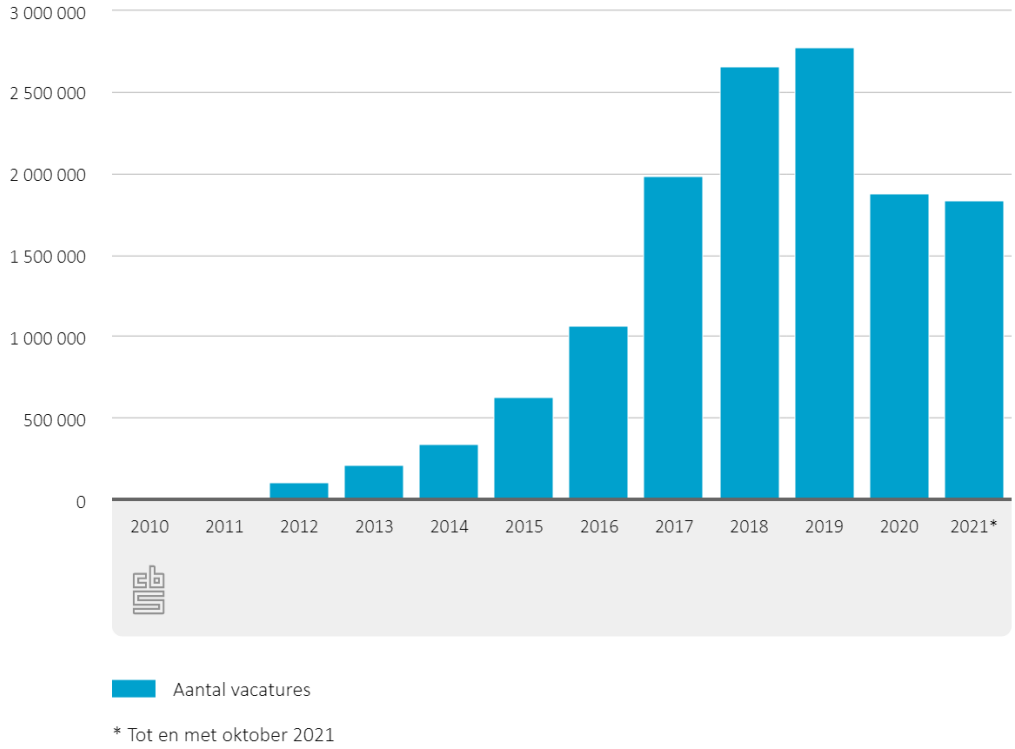
2.1 UWV vacatureteksten

In dit onderzoek is gebruik gemaakt van data over vacatures afkomstig van het Uitvoeringsinstituut Werknemersverzekeringen (UWV). Het UWV heeft deze gegevens aan het CBS geleverd om onderzoek te doen naar de bruikbaarheid voor statistische doeleinden. Dit onderzoek is daar een van de toepassingen van.

Het gaat hier over de vacatures die geplaatst die geplaatst zijn op Werk.nl. Dit is een website van het UWV met informatie voor werkzoekenden, werkgevers en arbeidsmarktinformatie. Het brengt werkzoekenden en werkgevers bij elkaar en heeft ook een uitgebreid vacature aanbod. Naast de originele vacatures van Werk.nl zelf bevat de dataset ook vacatures die het UWV door middel van andere vacaturewebsites heeft gehaald. Deze worden dan ook op Werk.nl gezet.

Het CBS heeft de beschikking over de vacatures die tussen 27 maart 2010 en 30 september 2021 op Werk.nl geplaatst zijn. In totaal gaat dat om 13.665.323 vacatures. Deze vacatures zijn niet dekkend voor alle vacatures in deze periode. Zo is het aantal vacatures in de eerste jaren een stuk kleiner dan in de laatste jaren van de verslagperiode. Figuur 2.1.1 laat het aantal vacatures per verslagjaar zien. Een van de redenen dat het aantal vacatures vooral de laatste jaren een stuk hoger is, is dat het UWV de manier van het verzamelen van de vacatures voor Werk.nl in de loop van de jaren heeft aangepast. Daarnaast is het aantal vacatures ook afhankelijk van het economisch klimaat.

2.1.1 - Aantal vacatures in de UWV dataset, 2010-2021



De UWV dataset bevat 20 variabelen. Voor dit onderzoek waren de volgende variabelen van belang:

- ID: unieke identificatiecode van de vacature. Hiermee konden we de unieke vacatures identificeren en selecteren.
- FUNCTIE_REF: Referentiecode van beroep in het Beroepen en Opleidingenregister (B&O) van het UWV. Hiermee konden we de koppeling maken van de richting van het beroep uit de vacature zoals gebruikt wordt in het B&O register.
- FUNCTIEOMSCHRIJVING: Tekstveld voor de beschrijving van de functie, bedrijf etc.
- WERVENDE_TEKST: Tekstveld voor de wervingstekst.
- OVERIGE EISEN: Tekstveld voor overige eisen voor de functie.

De tekstvelden FUNCTIEOMSCHRIJVING, WERVENDE_TEKST en OVERIGE EISEN zijn samengevoegd en gebruikt als input voor de tekstanalyse.
Een overzicht van alle variabelen uit de dataset is te vinden in bijlage 1.

2.2 Definities

Een belangrijke onderzoeksvraag in dit onderzoek van het ministerie van OCW was of de opleidingsvereisten in vacatures juist specifiek of algemeen zijn, en in hoeverre dat is veranderd over de tijd heen. Om dat in kaart te brengen is een van de eerste stappen geweest om samen met het ministerie een definitie van algemene en specifieke opleidingsvereisten op te stellen. Aan de hand van voorbeelden uit de UWV vacatureteksten is uiteindelijk de definitie opgesteld zoals weergegeven in tabel 2.2.1. Deze definities zijn vervolgens gebruikt voor het samenstellen van het onderzoeksbestand.

2.2.1 Definities algemene en specifieke opleidingsvereisten

Algemeen of specifiek	Omschrijving	Voorbeeld
Algemeen	- Als er alleen een niveau wordt genoemd. - Als er alleen één of meerdere richtingen worden genoemd. - Als er een richting in combinatie met voorbeeldopleidingen worden genoemd.	*"Je beschikt minimaal over een afgeronde relevante hbo-opleiding."* *"Een afgeronde HBO, richting ICT is een pré."* *"Jij hebt een MBO 2 diploma (of hoger) in een technische richting zoals installatietechniek, constructiebankwerker, mechatronica of een ander technisch diploma."*
Specifiek	- Als er één concrete opleiding wordt genoemd. - Als er enkele concrete opleidingen worden genoemd.	*"Je bent in het bezit van een pabo diploma."* *"Jij hebt een MBO 2 diploma (of hoger) in installatietechniek, constructiebankwerker of mechatronica."*

Er is gekozen om naast de dimensie algemeen/specifiek ook de vacatures in te delen aan de hand van het opleidingsniveau. Dit omdat het voor de beleidsdirecties mbo en ho van het ministerie van OCW belangrijk is hier onderscheid in te kunnen maken en het beeld wellicht anders kan zijn per niveau. Hierbij is onderscheid gemaakt tussen mbo (niveau 1-4) en ho (hbo, wo). Vanwege het relatief lage aantal wo vacatures in de data, is dit onderverdeeld bij het ho. Indien er geen opleidingsvereisten in de vacature teksten voorkwamen of het ging om een niveau lager dan mbo werd het als onbekend ingedeeld.

3. Methode

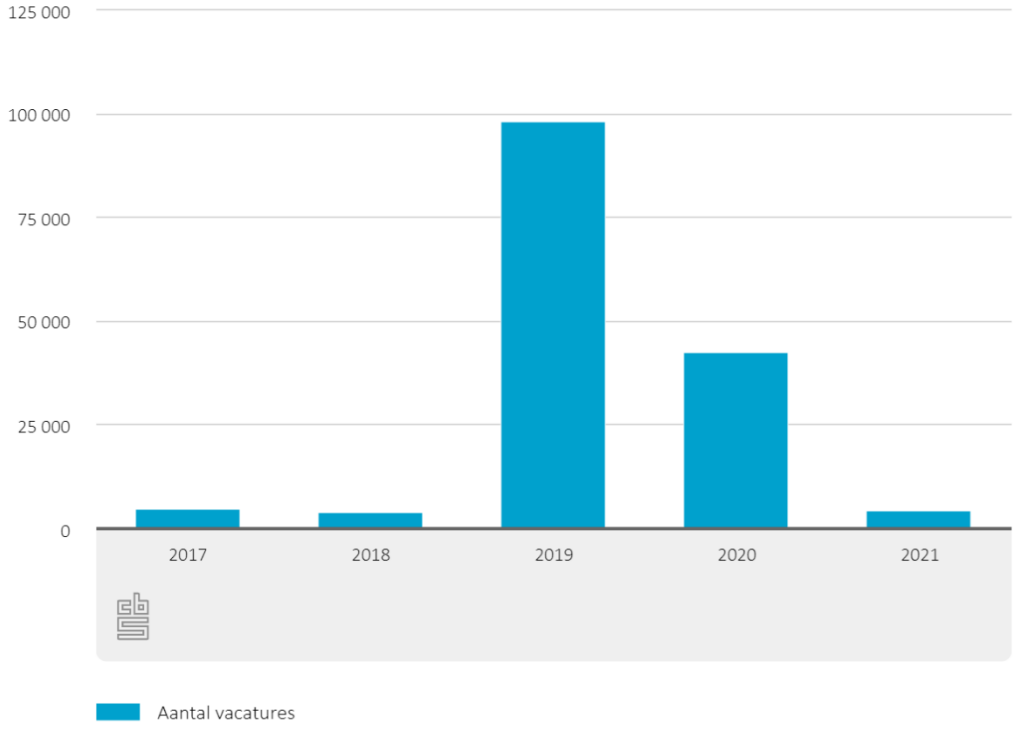
Na het bepalen van de definities voor algemene en specifieke opleidingsvereisten en de indeling voor het opleidingsniveau, zijn de volgende stappen doorlopen in de analyses:

1. Ontdubbelen van de ruwe vacatureteksten;
2. Voorbewerken en opschonen van de tekstuele data;
3. Onderzoeksbestand samenstellen voor het machine learning model;
4. Trainen van het machine learning model en het tunen van hyperparameters;
5. Corrigeren van classificatiefouten.

3.1 Ontdubbelen van vacatures

Voor de verslagjaren 2017 tot en met 2021 kwamen er dubbele vacatures met dezelfde ID voor in de dataset. Om aan de slag te kunnen gaan met de ruwe vacatureteksten, moesten deze daarom eerst ontdubbeld worden. De teksten van de vacatures met dezelfde ID's leken ook inhoudelijk en in woordsamenstelling sterk op elkaar. Van alle dubbele vacatures is er daarom willekeurig één gekozen zodat er voor iedere ID één unieke vacature overbleef. Figuur 3.1.1 laat zien hoeveel vacatures er per verslagjaar verwijderd zijn. Vooral in 2019 waren er relatief wat meer dubbele vacatures ten opzichte van de andere jaren.

3.1.1 Aantal verwijderde vacatures vanwege ontdubbeling, per jaar



3.2 Voorbewerken tekstuele data

Om de ruwe vacatureteksten te kunnen analyseren, zijn er een aantal voorbewerkingsstappen gedaan op de teksten. Hiervoor is o.a. het Nederlandse *spaCy* taalmodel (versie 3.2.0) in Python gebruikt. De volgende bewerkingen zijn gedaan:

- Bij een aantal teksten werd de vacaturetekst afgekapt en werd er verwezen naar de originele website van de vacature. Deze vacatures bevatten allemaal de volgende zin: **** Om de originele vacature te bekijken of hierop te solliciteren, klik op 'solliciteren' en kies vervolgens daaronder voor het "online sollicitatieformulier". ****. Deze zin is overal verwijderd.
- Woorden waarin een hoofdletter wordt vooraf gegaan door een kleine letter zijn verwijderd omdat hier waarschijnlijk een spatie tussen had moeten staan.
- Stopwoorden zijn verwijderd.
- Alleen woorden bestaande uit letters zijn behouden.
- Leestekens zijn verwijderd.
- Alle hoofdletters zijn vervangen door kleine letters.

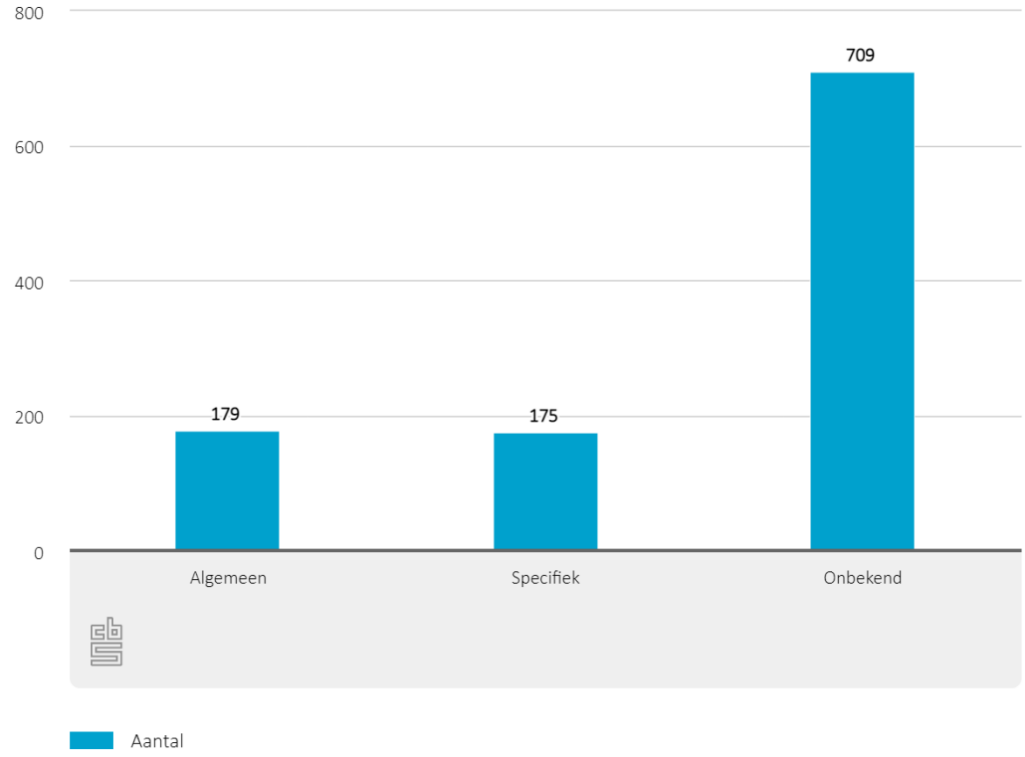
Om de tekst vervolgens te kunnen verwerken in een machine learning model, is het nodig om de tekstuele data om te zetten naar numerieke waarden en de informatie waarde van de teksten mee te geven. Daarvoor gebruiken we de Term Frequency – Inverse Data Frequency (TF-IDF) methode. De Term Frequency (tf) geeft per vacaturetekst de frequentie van woorden aan die voorkomen. Per vacaturetekst wordt dan een ratio berekend van het aantal keren dat een woord voorkomt in die tekst ten opzichte van het totaal aantal woorden in die vacaturetekst. Daarnaast is er de Inverse Data Frequency (idf). Deze maat zegt iets over hoe uniek een woord is en geeft een gewicht mee. Woorden die dus weinig voorkomen krijgen een hoger gewicht mee. Vervolgens kun je de TF-IDF score berekenen per woord in een vacaturetekst.

3.3 Samenstellen van een onderzoeksbestand

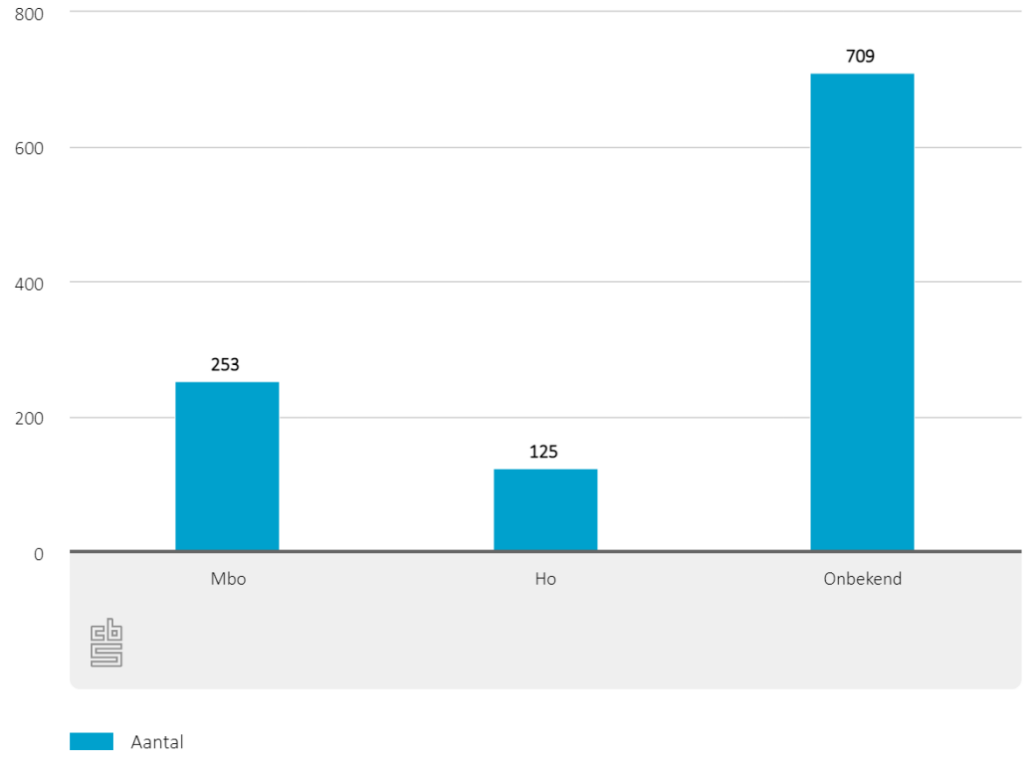
Het doel van het onderzoek was het inzichtelijk maken van de vraag naar algemene of specifieke opleidingsvereisten in vacatureteksten over de tijd. Om een machine learning model te kunnen trainen was het eerst nodig een onderzoeksbestand samen te stellen waarop het model getraind kon worden. Hiervoor is een gestratificeerde steekproef van vacatureteksten getrokken voor de periode 2017 tot en met oktober 2021. Er is gestratificeerd op de beroepsklasse uit de Beroepenindeling ROA-CBS (BRC). De beroepsklasse is het hoogste niveau van de BRC en deelt beroepen in op basis van 13 hoofdrichtingen. zodat we zeker wisten dat de steekproef een representatieve afspiegeling zou zijn van alle verschillende beroepsgroepen die in de data voorkomen. De BRC is samengesteld door het Researchcentrum voor Onderwijs en Arbeidsmarkt (ROA) en CBS om toe te passen bij analyses en statistiek op nationaal niveau.

Daarnaast is het voor het trainen van een machine learning model nodig, om een gelabelde dataset te hebben waarvan we al weten wat de 'echte' waarden zijn voor algemeen/specifiek en het opleidingsniveau. Op deze manier kan het machine learning model leren van een werkelijke dataset, om vervolgens na training schattingen te kunnen maken voor een nieuwe dataset. Deze gelabelde dataset is vervolgens handmatig samengesteld met behulp van de definities voor algemeen/specifiek en het opleidingsniveau. Op basis daarvan zijn de verdelingen van de categorieën naar voren gekomen in de data, zie figuur 3.3.1 en 3.3.2.

3.3.1 Verdeling algemeen, specifiek en onbekend in de gelabelde set



3.3.2 Verdeling mbo, hbo, wo en onbekend in de gelabelde set



Zoals in paragraaf 3.2 beschreven werd, waren er veel vacatures waarbij werd doorverwezen naar de originele website en kon er geen informatie ontleend worden aan de tekst zelf. Deze

vacatures zijn ingedeeld als ‘onbekend’. Ook kwamen relatief veel vacatures voor waarvoor er niet minimaal een opleiding op mbo-niveau werd gevraagd. Ook deze zijn ingedeeld bij de categorie ‘onbekend’. Dit is ook terug te zien in de verdeling van de categorieën waarbij de categorie ‘onbekend’ duidelijk de grootste is.
Omdat ongebalanceerde categorieën invloed kunnen hebben op het machine learning model, hebben we een willekeurige selectie gemaakt uit de onbekende vacatures in het onderzoeksbestand, zodat de verhouding algemeen/specifiek/onbekend gelijker verdeeld was over de categorieën. Hetzelfde is gedaan voor de categorie ‘onbekend’ bij het opleidingsniveau.

3.4 Machine learning

3.4.1 Modellen

Het doel van dit onderzoek is het schatten van de juiste categorieën algemeen/specifiek/onbekend en het opleidingsniveau van vacatureteksten. Om dit te kunnen doen zijn er telkens drie modellen getraind met machine learning op het onderzoeksbestand:

- Algemeen/specifiek/onbekend: Er wordt één model getraind die per vacature aangeeft of het gaat om algemene, specifieke of onbekende opleidingsvereisten.
- Mbo: Er wordt één model getraind die aangeeft of een vacature vraagt om een mbo opleidingsniveau, of geen mbo opleidingsniveau.
- Ho: Er wordt één model getraind die aangeeft of een vacature vraagt om een ho opleidingsniveau, dus hbo of wo, of geen ho opleidingsniveau.

In de volgende stappen komen deze drie modellen telkens terug en wordt ook verder uitgelegd waarom er voor deze modellen is gekozen.

3.4.2 Trainings- en testsets

Om de gelabelde dataset te kunnen gebruiken bij het trainen van een machine learning model is het opgesplitst in een trainings- en testset. Hierbij is de verhouding 70/30 toegepast. Op deze manier kunnen we op 70% van de gelabelde data een model trainen en deze op de overgebleven 30% testen. Zo kan er bepaald worden hoe goed het model kan schatten voor data die nog niet bekend is. Omdat we het model gaan toepassen op de gehele reeks in dit onderzoek, is het dus van belang dat het universeel te gebruiken is op andere vacatureteksten in de onderzoekspopulatie. Voor alle drie de eerder beschreven modellen, wordt er een split gemaakt in een trainings- en testset.

Daarnaast wordt er bij het afleiden van de trainings- en testset ook gestratificeerd getrokken uit alle categorieën van het model, zodat er ongeveer gelijke verhoudingen per categorie zijn in de trainings- en de testset.

3.4.3 Machine learning modellen en kwaliteitsmaten

Tijdens het onderzoek zijn diverse machine learning modellen getest met als doel het model te kiezen met de beste voorspelkwaliteit. De volgende modellen zijn getest: Naïve Bayes, Random Forest, Gradient Boosting, Extreme Gradient Boosting, Logistische regressie en Support Vector Machines. Om het beste model te selecteren is er gekeken naar maten die de kwaliteit van deze schattingen kwantificeren. Reden dat er meerdere maten hiervoor zijn is onder andere dat het vaak van de toepassing van het model afhangt welke maten het meest geschikt zijn.

In dit onderzoek zijn de maten die te maken hebben met classificatie vooral interessant. In tabel 3.4.3 wordt weergegeven waarop deze maten worden gebaseerd, de zogenaamde confusie

matrix. De tabel laat een voorbeeld zien voor een classificatie met twee categorieën: Mbo en overig (ho + onbekend). De confusie matrix bestaat uit de volgende onderdelen:

- TP: het aantal true positives of correct als mbo geclassificeerde vacatures
- FN: het aantal false negatives of incorrect als niet-mbo geclassificeerde vacatures
- FP: het aantal false positives of incorrect als mbo geclassificeerde vacatures
- TN: het aantal true negatives of correct als niet-mbo geclassificeerde vacatures

3.4.3 Voorbeeld confusie matrix

Werkelijk/Geschat	Mbo	Overig
Mbo	True Positives (TP)	False Negatives (FN)
Overig	False Positives (FP)	True Negatives (TN)

Op basis van deze confusie matrix kunnen onder andere de metrieken uit tabel 3.4.4 berekend worden. Deze confusie matrix kan ook uitgebreid worden naar meer dan twee categorieën, zoals algemeen, specifiek én onbekend.

3.4.4 Overzicht metrieken

Metriek	Definitie
Accuracy	$\frac{TP + TN}{TP + TN + FP + FN}$
Recall	$\frac{TP}{TP + FN}$
Precision	$\frac{TP}{TP + FP}$
F1	$\frac{TP}{TP + (\frac{FP + FN}{2})}$

De accuracy kijkt naar het aantal juiste voorspellingen in totaal ten opzichte van het totaal aantal voorspellingen. De recall meet de fractie vacatures met mbo-niveau en die ook als zodanig worden geclassificeerd (een classificatiekans). De precision meet de fractie correct geschatte klassen in die records die geclassificeerd worden als ‘mbo’ (een calibratiekans). De F1-score is het harmonisch gemiddelde van de recall en precision. Door het harmonisch gemiddelde te nemen (en niet het rekenkundig of geometrisch gemiddelde) ligt de F1 dichter bij de laagste van de precision en recall. In de praktijk hangen de precision, recall en F1 af van de fractie vacatures die in werkelijkheid mbo is. Hoe lager deze fractie hoe lager de F1 bijvoorbeeld zal zijn.

Bij de voorselectie van de machine learning modellen is er eerst geselecteerd met de accuracy. Op basis daarvan kwam het Gradient Boosting model als beste naar voren. Later in het onderzoek is gekeken naar het totaal van de metrieken per model.

3.4.4 Gradient Boosting

Het algoritme, Gradient Boosting, dat in dit onderzoek gebruikt wordt combineert beslisbomen om tot gedetailleerde schattingen te komen. Een standaard beslisboom is een eenvoudige machine learning methode. In een beslisboom wordt de data op basis van één kenmerk herhaaldelijk in twee groepen verdeeld. In iedere groep is de geschatte waarde het groepsgemiddelde. De groepen worden zo bepaald dat iedere keer als er subgroepen gemaakt

worden de toename in de modelkwaliteit het grootst is. De methode zal proberen om groepen zo groot mogelijk te maken en te zorgen dat de groepen zoveel mogelijk gelijke records bevat wat betreft de doelvariabele. Dit proces wordt herhaald tot een stopcriterium is bereikt. Het stopcriterium is een combinatie van voldoende grote groepen en voldoende toename van de modelkwaliteit.

Varianten op de beslisboom, zoals Gradient Boosting, zijn ensemble-methoden die worden ontwikkeld om het algoritme stabieler te maken (kleinere variantie), robuuster tegen het modelleren van ruis en tegelijkertijd schaalbaar te houden voor grote datasets. Bij ensemble-methoden wordt niet één beslisboom geschat maar een set beslisbomen. Bij gradient tree boosting worden eenvoudige beslisbomen na elkaar (sequentieel) getraind waarbij elke opeenvolgende boom de schattingen van de voorgaande bomen verfijnd. Bij deze methode ligt de focus dus niet zeer op de kwaliteit van één enkele beslisboom, zoals bij een standaard beslisboom wel wordt gedaan, maar juist op het combineren van eenvoudige beslisbomen en deze sequentieel te verbeteren.

3.4.5 Schatten van het machine-learningmodel

Na het vinden van het Gradient Boosting model als het optimale model voor dit onderzoek, was de volgende stap het instellen van de zogenaamde hyperparameters in het model. Dit proces heet 'tunen'. Hyperparameters zijn dus geen parameters die je kunt trainen met de data, maar die meegegeven worden aan het model. Elke context waarin een machine learning model wordt toegepast is weer anders, waardoor het model ook voor iedere situatie andere instellingen nodig heeft. Er zijn verschillende methoden om de optimale hyperparameters te vinden. In dit onderzoek is er gebruik gemaakt van een Grid Search. Bij een Grid Search wordt er voor verschillende hyperparameters een lijst met mogelijke waarden klaargezet. Vervolgens worden al die verschillende combinaties van hyperparameters getest en wordt de modelkwaliteit beoordeeld met behulp van de recall per label (bijvoorbeeld mbo versus overig).

Er is een aantal hyperparameters waarmee de complexiteit van het model bepaald kan worden. De belangrijkste zijn (in dit onderzoek onderzocht):

- *aantal beslisbomen*: Dit is de belangrijkste parameter. Hoe meer bomen, hoe meer detail de methode kan schatten.
- *boomdiepte*: maximale diepte van de bomen, dus het aantal keer dat de dataset gesplitst wordt. Bij een boomdiepte van 4 kan de dataset in maximaal 24 = 16 groepen verdeeld worden. Hier ook weer: hoe dieper de boom, hoe meer detail.
- *leersnelheid*: hoe snel leert het model. Als deze hoog is dan hebben de eerste geschatte beslisbomen heel veel invloed op de schattingen. Omdat er een zekere mate van toeval zit in deze beslisbomen kan het model hiermee minder goed zijn. Een lagere leersnelheid is in het algemeen beter maar zorgt er wel voor dat het aantal bomen hoger moet zijn en dat het trainen langer duurt.

De Grid Search is vervolgens toegepast op bovenstaande parameters. In deze methode wordt gebruik gemaakt van een kruisvalidatiemethode. Dat betekent dat de trainingsset (70% van de data) wordt opgesplitst in dit geval 5 delen. In elk van deze delen wordt weer een verdeling gemaakt met een trainings- en testset. De testset wordt hier een validatieset genoemd. Vervolgens worden alle combinaties van hyperparameters gebruikt bij het trainen van de trainingsset en wordt de modelkwaliteit getest in de validatieset. Omdat we 5 delen hebben, wordt één specifieke set hyperparameters, dus 5 keer getraind en getest. Uiteindelijk kan een gemiddelde genomen worden over de modelkwaliteit op alle sets samen. De modelkwaliteit is

gemeten d.m.v. de recall, omdat het belangrijk is een zo groot mogelijk aandeel van de algemene en specifieke geclassificeerde vacatures ook echt te vinden. Omdat het model is getraind op 3 categorieën, algemeen/specifiek/onbekend, moeten we voorkomen dat het model vooral goed 'onbekend' kan schatten. Gemiddeld genomen zou een model dat onbekend goed kan schatten dus best goed kunnen zijn, maar in de praktijk de relevante categorieën algemeen en specifiek kunnen missen.

Aan de hand van hier bovengenoemde methoden zijn voor de twee gecodeerde dimensies drie modellen onderzocht. De kruisvalidatiemethode wees uit dat deze dimensies het beste aan de hand van een drietal modellen konden worden voorspeld. Twee van deze modellen hebben betrekking op het opleidingsniveau en voorspellen respectievelijk of de vacature een mbo of een ho opleiding betreft. Het derde model voorspelt tegelijk of een vacature behoort tot de groep vacatures met algemene opleidingsvereisten, specifieke opleidingsvereisten, of dat de opleidingseisen onbekend zijn. De optimale parameters die voor deze drie Gradient Boosting modellen zijn vastgesteld zijn als volgt:

- Mbo vacature: aantal bomen = 400, boomdiepte = 12, leersnelheid = 0,02.
- Ho vacatures: aantal bomen = 100, boomdiepte = 10, leersnelheid = 0,02.
- Algemene, Specifieke, onbekende eisen: aantal bomen = 100, boomdiepte = 12, leersnelheid = 0,01.

3.4.6 Software

De analyses zijn uitgevoerd in Python 3.8. De belangrijkste packages die zijn gebruikt:

- SpaCy 3.2.0 – Nederlandse taalmodel
- Scikit-learn – deze software is gebruikt voor het voorbewerken van de teksten, machine learning modellen en modelkwaliteit.

3.5 Corrigeren van classificatiefouten

Het getrainde algoritme heeft een nauwkeurigheid (op basis van de testset) die lager is dan 100 procent. Daarom is het nodig om de fout die het algoritme maakt tijdens het voorspellen te corrigeren. Aan de hand van de testset kan worden bepaald hoe groot de misclassificatiefouten zijn die het algoritme maakt. Bijvoorbeeld: als een vacature uit de testset door het algoritme is beoordeeld als algemeen, maar in de testset als specifiek, dan dient deze vacature te vallen onder specifiek. Aan de hand van de testset kan bijvoorbeeld bepaald worden hoeveel vacatures door het algoritme zijn beoordeeld als algemeen en in de testset als specifiek. Stel de helft van de vacatures die door het algoritme zijn beoordeeld als algemeen zijn door de testset beoordeeld als specifiek. Dan dient wanneer hetzelfde getrainde algoritme op een nieuwe set wordt losgelaten, de helft van de vacatures die als algemeen zijn beoordeeld als specifiek geteld te worden.

Deze methode[1] zorgt ervoor dat de bias die door het algoritme wordt veroorzaakt, (grotendeels) teniet wordt gedaan. Een bijkomend voordeel van deze methode is dat de onzekerheid (marges) rondom de geschatte percentages bepaald kan worden. De onzekerheden (betrouwbaarheidsintervallen) rondom de schattingen kunnen bepaald worden door het simuleren van testsets. Doordat de testset willekeurig is samengesteld, is er een willekeurigheid rondom bijvoorbeeld het aantal vacatures in de testset die door het algoritme als algemeen zijn beoordeeld en in de testset als specifiek. Deze willekeurigheid heeft invloed

[1] Meertens, Q.A. (2021). *Misclassification bias in statistical learning.*

op de schattingen die op basis van deze aantallen worden gemaakt (deze hangen immers direct af van bovenstaande aantallen). Op basis van het simuleren van testset kan vervolgens een onzekerheidsmarge bepaald worden.

4. Uitkomsten

Het bespreken van de uitkomsten van de machine learning modellen doen we in twee stappen. In paragraaf 4.1 wordt eerst een overzicht gegeven van de kwaliteit van de schattingen door de uitkomsten van de verschillende metrieken voor de drie modellen te bespreken. Dit geeft een beeld van de betrouwbaarheid van de schattingen en daarmee ook de kaders voor de daadwerkelijke schattingen van de verhouding algemeen/specifiek/onbekend en de opleidingsniveaus die worden besproken in paragraaf 4.2. Over het algemeen geldt dat hoe dichter de uitkomsten van de metrieken bij 1 liggen hoe beter de daadwerkelijke schattingen zullen zijn en hoe kleiner de marges rondom de cijfers.

4.1 Kwaliteit van de schattingen

De modelkwaliteit van de schattingen wordt hieronder per geschat model beschreven.

4.1.1 Uitkomsten metrieken model *algemeen/specifiek/onbekend*

Categorie	Precision	Recall	F1-score	Accuracy
Algemeen	0,72	0,67	0,69	
Specifiek	0,72	0,64	0,68	
Onbekend	0,80	0,94	0,86	
Totaal				0,75

De accuracy laat zien dat er 75 procent juiste voorspellingen zijn in totaal ten opzichte van het totaal aantal voorspellingen voor het model algemeen/specifiek/onbekend. Daarnaast laat de recall zien dat er van de werkelijke algemene en specifieke opleidingsvereisten in de vacatures, er 67 procent en 64 procent worden gevonden door het model. Ook laat de precision zien dat van de vacatures die als algemeen en specifiek zijn geclassificeerd door het model, dit voor 72 procent ook juist was voor zowel algemeen als specifiek. De metrieken van de onbekende categorie laten nog betere resultaten zien, maar zijn minder relevant voor het beleid.

4.1.2 Uitkomsten metrieken model *mbo*

Categorie	Precision	Recall	F1-score	Accuracy
Mbo	0,85	0,83	0,84	
Geen mbo	0,88	0,90	0,89	
Totaal				0,87

De accuracy laat zien dat 87 procent juiste voorspellingen zijn in totaal ten opzichte van het totaal aantal voorspellingen voor het model mbo. Daarnaast laat de recall zien dat van de vacatures die werkelijk vragen om een mbo niveau, er 83 procent worden gevonden door het model. Ook laat de precision zien dat van de vacatures die als mbo zijn geclassificeerd, dit voor 85 procent juist was.

4.1.3 Uitkomsten metrieken model *ho*

Categorie	Precision	Recall	F1-score	Accuracy
Ho	0,72	0,55	0,62	
Geen ho	0,87	0,94	0,90	
Totaal				0,85

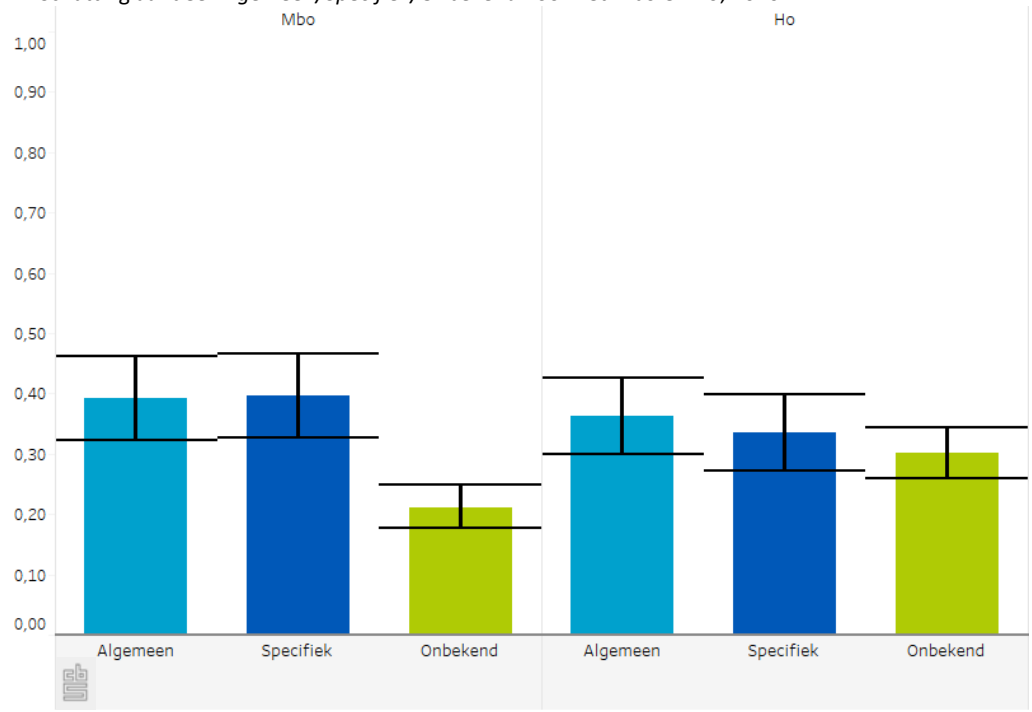
De accuracy laat zien dat 85 procent juiste voorspellingen zijn in totaal ten opzichte van het totaal aantal voorspellingen voor het model ho. Daarnaast laat de recall zien dat van de vacatures die werkelijk vragen om een ho niveau, er 55 procent worden gevonden door het model. Ook laat de precision zien dat van de vacatures die als mbo zijn geclassificeerd, dit voor 72 procent juist was. De resultaten voor ho komen wat minder goed uit dan voor mbo. Dit komt hoogst waarschijnlijk doordat de categorie ho uit zowel hbo als wo vacatures bestaat waardoor er op een ‘ bredere’ categorie getraind moest worden. Daarnaast kwamen er relatief weinig wo vacatures in het onderzoeksbestand voor waardoor die ook lastiger te trainen waren.

4.2 Schattingen

De uitkomsten voor de verhouding algemeen/specifiek/onbekend zijn geschat per opleidingsniveau en worden ook op deze manier gepresenteerd in dit rapport. Daarnaast is het belangrijk om bij de interpretatie van de uitkomsten rekening te houden met de marges rondom de schattingen. Deze marges zijn telkens duidelijk in de figuren weergegeven. Het gaat hier om het 95% betrouwbaarheidsinterval. Dit betekent dat we met 95 procent zekerheid kunnen zeggen dat de schatting tussen deze onder- en bovengrens ligt. Ook betekent dit dat als er twee waarden (schattingen) met elkaar vergeleken worden en deze liggen binnen elkaars betrouwbaarheidsinterval, deze waarden niet significant van elkaar verschillen.

Figuur 4.2.1 geeft een eerste beeld van de uitkomsten voor 2020, het meest recente volledige jaar waarvoor we data beschikbaar hadden. Hierin is te zien dat voor het mbo het aandeel vacatures met algemene opleidingsvereisten met bijna 40 procent vrijwel gelijk is aan het aandeel vacatures met specifieke opleidingsvereisten. Rond deze schattingen zit een marge van ongeveer 14 procent. Het aandeel onbekenden, dus vacatures waarvoor op basis van de vacaturetekst geen indeling kon worden gemaakt, is voor het mbo iets meer dan 20 procent. Voor het ho is het aandeel onbekend in 2020 met ongeveer 30 procent een stuk hoger. De verhouding algemeen/specifiek is hier ook vrijwel gelijk. Algemeen (36 procent) lijkt iets vaker voor te komen dan specifiek (34 procent) maar deze schattingen vallen binnen elkaars marges en zijn dus niet significant verschillend van elkaar.

4.2.1 Schatting aandeel *Algemeen/Specifiek/Onbekend* voor het *Mbo* en *Ho*, 2020



In de volgende paragrafen gaan we dieper op de uitkomsten in. Als eerste komen de ontwikkeling van de verhouding algemeen/specifiek/onbekend aan bod voor het mbo (paragraaf 4.2.1). Vervolgens wordt in paragraaf 4.2.2 hetzelfde gedaan voor het ho. Als laatste worden in paragraaf 4.2.3 ook uitkomsten per beroepsklasse weergegeven.

De volledige uitkomsten van de schattingen zijn terug te vinden in bijlage 2.

4.2.1 Middelbaar beroepsonderwijs

Onderstaande figuur laat zien dat de verhouding algemeen/specifiek/onbekend voor het mbo relatief stabiel is geweest de afgelopen 10 jaar. Zowel het aandeel algemeen als het aandeel specifiek schommelt zo rond de 40 procent. Voor alle jaren vallen de schattingen voor algemeen en specifiek binnen elkaars marges waardoor deze nooit significant van elkaar verschillen.
Het valt verder op dat in 2021, het laatste jaar, relatief wat meer onbekenden voorkomen waardoor ook zowel het aandeel algemeen als specifiek wat naar beneden gaat. Ook in 2011 en 2015 is het aandeel onbekend iets hoger wat in die jaren er vooral voor lijkt te zorgen dat het aandeel algemeen wat lager ligt.

4.2.1.1 Schatting aandeel *Algemeen/Specifiek/Onbekend* voor het *Mbo*, 2010-2021

4.2.2 Hoger onderwijs

Ook voor het ho zijn de ontwikkelingen vrij stabiel. Het aandeel algemeen schommelt tussen de 43 en 32 procent en ligt alle jaren net iets boven het aandeel specifiek. Het aandeel specifiek schommelt tussen 36 en 29 procent. Significant verschillen de aandelen algemeen en specifiek bij het ho echter nooit van elkaar, de schattingen vallen elk jaar binnen elkaars marge. Bij het ho valt de stijging van het aandeel onbekend in 2021 nog meer op dan bij het mbo. En ook hier lijkt bij de overige jaren er een sterker verband te zitten tussen algemeen en onbekend dan tussen specifiek en onbekend. Zo ligt het aandeel algemeen tussen de jaren 2011 en 2016 relatief wat hoger terwijl in dezelfde periode het aandeel onbekend dan wat lager ligt.

4.2.2.1 Schatting aandeel *Algemeen/Specifiek/Onbekend* voor het *Ho*, 2010-2021

4.2.3 Beroepsgroepen

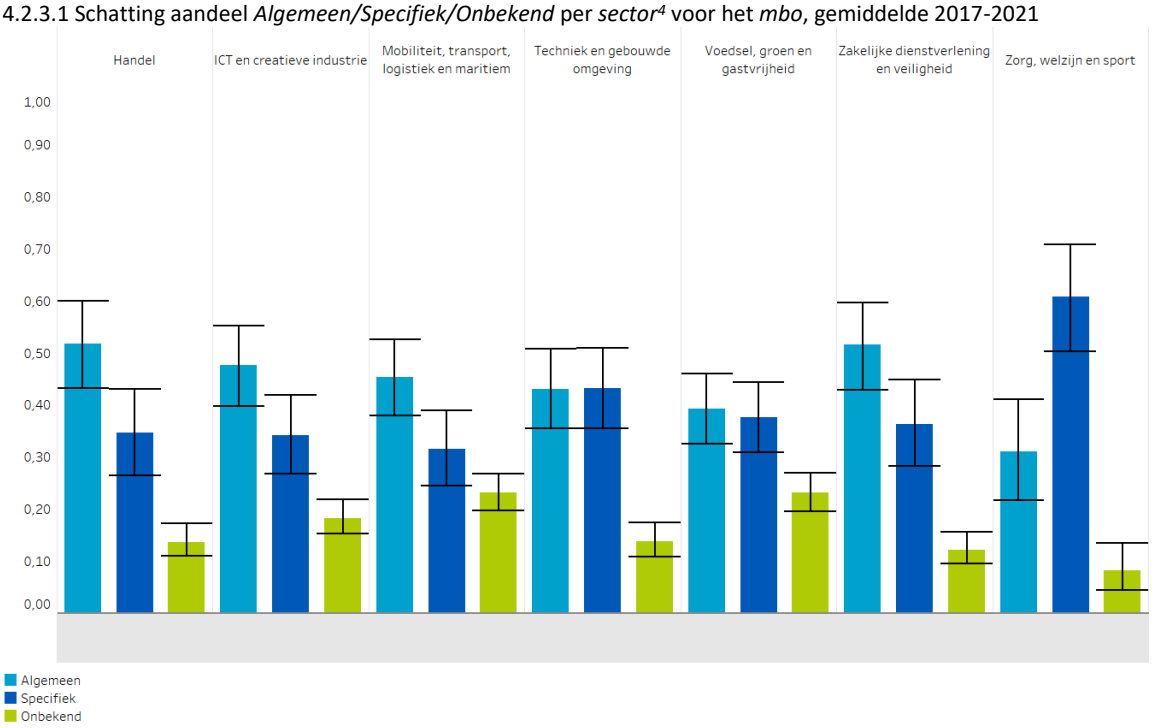
Het ministerie van OCW is ook geïnteresseerd in de vraag of de opleidingsvereisten in vacatures verschillen tussen verschillende beroepsgroepen. Omdat we voor de trekking van de steekproef van het onderzoeksbestand hebben gestratificeerd op BRC beroepsklasse was het ook mogelijk om schattingen te maken van de verdeling algemeen/specifiek/onbekend per beroepsklasse. Deze beroepsklassen zijn vervolgens vertaald naar de mbo sectoren[2] en croho-onderdelen[3] om de indeling beter te laten aansluiten op beleid en de praktijk. In bijlage 3 is deze indeling terug te vinden. De schattingen zijn gemaakt voor een gemiddelde van de jaren 2017 tot en met 2021 om genoeg waarnemingen per beroepsklasse over te houden.

De uitkomsten voor het mbo in de figuur hieronder laten zien dat vooral voor de sector 'Zorg, welzijn en sport' er relatief vaak specifieke opleidingsvereisten in vacatures voorkomen. Dit heeft waarschijnlijk te maken met de vacatures waarin gevraagd wordt naar een specifieke opleiding voor verzorgende of verpleegkundige. Voor de sectoren 'Techniek en gebouwde omgeving' en 'Voedsel, groen en gastvrijheid' was de vraag naar algemene of specifieke opleidingsvereisten ongeveer gelijk aan elkaar. Bij de overige vier sectoren ligt het aandeel algemeen hoger dan het aandeel specifiek. Maar afgezien van de sector 'Handel' lijken het aandeel algemeen en het aandeel specifiek niet significant van elkaar te verschillen.

[2] Voor de indeling in mbo-sectoren is gebruik gemaakt van de sectorkamerindeling van SBB.

[3]

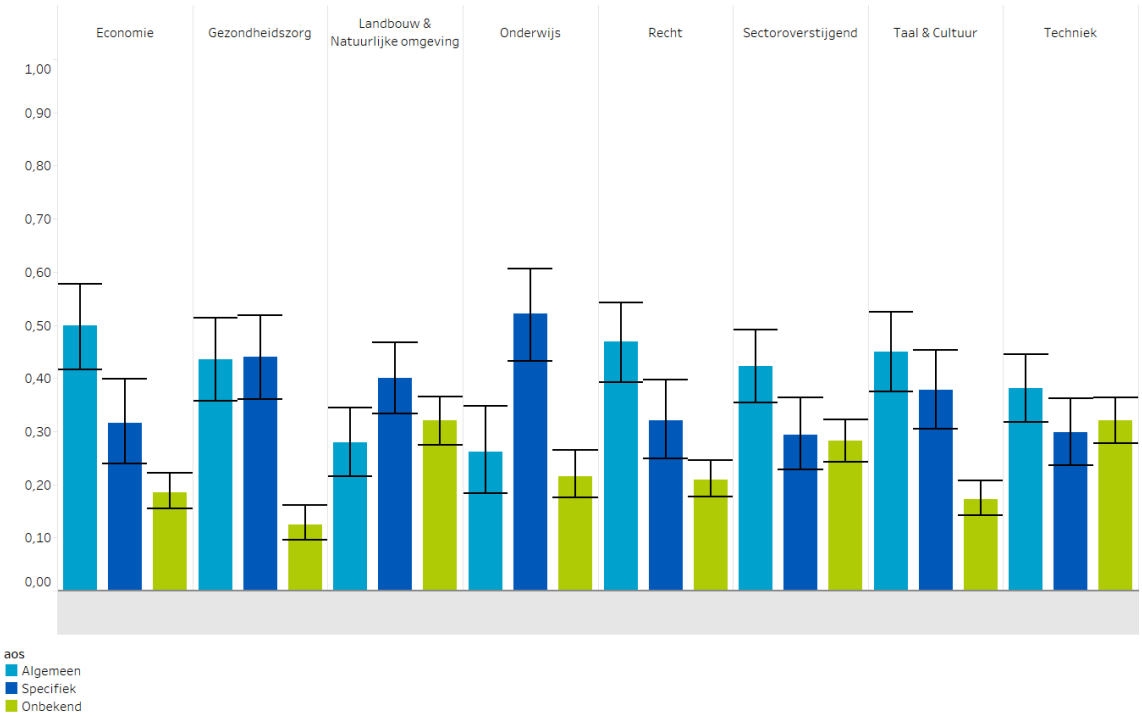
4.2.3.1 Schatting aandeel *Algemeen/Specifiek/Onbekend* per *sector*[4] voor het *mbo*, gemiddelde 2017-2021



Bij het ho valt op dat bij de croho-onderdelen ‘Onderwijs’ en ‘Landbouw & Natuurlijke omgeving’ als enigen de meerderheid van de vacatures specifieke opleidingsvereisten heeft. Wat betreft onderwijs zullen dit overwegend vacatures voor basisschoolleerkracht of docent voor de middelbare school zijn. Bij ‘Gezondheidszorg’ is de verhouding algemeen/specifiek vrijwel gelijk aan elkaar. De croho-onderdelen waar algemene opleidingsvereisten duidelijk de overhand hebben zijn ‘Economie’, ‘Recht’ en ‘Sectoroverstijgend’. Onder ‘Sectoroverstijgend’ vallen de beroepsklassen managers, dienstverlenende beroepen en overig.

[4] De sectoren waarvoor niet genoeg waarnemingen waren worden niet getoond.

4.2.3.2 Schatting aandeel *Algemeen/Specifiek/Onbekend* per *croho-onderdeel*[5] voor het *ho*, gemiddelde 2017-2021



[5] Croho-onderdelen waarvoor niet genoeg waarnemingen waren worden niet getoond.

5. Conclusies en aanbevelingen

Dit onderzoek had de volgende twee hoofddoelen:

- Onderzoeken of met behulp van big data technieken een concrete beleidsvraag kan worden beantwoord. In dit geval gaat het om de vraag in hoeverre de op de arbeidsmarkt gevraagde opleidingen in de loop van de tijd veranderen.
- Ervaring op doen met big data-analyse en de mogelijkheden die dat biedt voor statistische vragen op het de beleidsterreinen van OCW. De opgedane inzichten kunnen door OCW en CBS gezamenlijk worden benut voor toekomstige vraagstukken.

In dit hoofdstuk bespreken we de conclusies met betrekking tot deze doelen en geven we aanbevelingen voor eventueel vervolgonderzoek.

5.1 Conclusies

Er zijn twee soorten conclusies te benoemen: 1) conclusies met betrekking tot de twee hoofddoelen die voor dit onderzoek zijn geformuleerd, en 2) conclusies over de specifieke inhoudelijke onderzoeksvragen. We beginnen hier met het eerste.

Dit onderzoek laat zien dat met een combinatie van machine learning en tekstmining het mogelijk is vacature teksten te classificeren op basis van opleidingsvereisten: de dimensie algemeen/specifiek/onbekend en het opleidingsniveau. Er zat nog wel een ruime marge rondom de schattingen maar er is op deze manier wel een eerste beeld gegeven van de gevraagde opleidingsvereisten over een periode van 10 jaar. De toegevoegde waarde van dit type onderzoek is vooral dat een relatief subjectief construct zoals algemene of specifieke opleidingsvereisten onderzocht kan worden. Dit kan daarom een nuttige aanvulling zijn op de meer reguliere statistieken voor het beantwoorden van beleidsvragen. Er is nog wel aanvullend onderzoek nodig om onder andere de kwaliteit van de schattingen te verbeteren en de invloed van de onbekenden (ontbrekende tekstdata) in beeld te brengen, daarover ook meer in de aanbevelingen.

Wat betreft de inhoudelijke onderzoeksvragen kunnen de volgende conclusies getrokken worden. Voor de vacatures waarvoor de opleidingsvereisten konden worden vastgesteld (dus afgezien van de onbekenden) is ongeveer de helft algemeen en ongeveer de helft specifiek. Dit geldt zowel voor het mbo als ho. Voor het ho lijken er wel net iets vaker algemene opleidingsvereisten voor te komen maar de verschillen zijn minimaal. Ook de ontwikkeling over de tijd is vrij stabiel. Er zijn geen significante verschillen in de verhouding algemeen/specifiek over de periode 2010-2021. Dit geldt zowel voor het mbo als ho. Tussen de sectoren van het mbo en de croho-onderdelen van het ho zagen we wel verschillen in de verhouding algemeen/specifiek. Zo kwamen bijvoorbeeld voor de sector ‘Zorg, welzijn en sport’ er relatief vaak specifieke opleidingsvereisten in vacatures voor. Bij het ho viel op dat bij de croho-onderdelen ‘Onderwijs’ en ‘Landbouw & Natuurlijke omgeving’ de meerderheid van de vacatures specifieke opleidingsvereisten heeft. Deze verschillen tussen de verschillende beroepsgroepen zagen er herkenbaar uit voor het ministerie van OCW.

5.2 Aanbevelingen

Een belangrijk onderdeel van een pilot onderzoek is ervaringen op doen die meegenomen kunnen worden bij eventueel toekomstig onderzoek. In deze paragraaf worden daarom een aantal relevante aanbevelingen besproken die uit dit onderzoek naar voren zijn gekomen.

- De tekstvelden in de UWV vacaturedata waren niet altijd (volledig) gevuld. Dit kwam omdat er veel vacatures in zaten waarin met een link werd verwezen naar de originele vacature die op een andere website stond. Vaak ontbrak daardoor de informatie over de opleidingsvereisten en was de categorie onbekend relatief groot in het onderzoeksbestand. Vooral als het doel van het onderzoek tekstanalyse is, is het belangrijk om vooraf stil te staan bij de mogelijke invloed van deze ontbrekende teksten op de resultaten. En wellicht een manier te bedenken om hier meer om te gaan of voor te corrigeren. In het huidige onderzoek kwamen we er tijdens de analyses pas achter dat de onbekenden een relatief grote categorie waren.
- Er heeft in dit onderzoek geen validatie van de UWV vacaturedata ten opzichte van een secundaire bron plaatsgevonden. Daar was helaas geen ruimte meer voor binnen het onderzoeksbudget. Hierdoor is het lastig conclusies te trekken over de representativiteit van de UWV vacaturedata. Zo zagen we bijvoorbeeld dat er relatief weinig ho vacatures in de dataset voorkwamen, maar weten we niet of dit ook een afspiegeling is van de werkelijkheid. Als er daadwerkelijk statistieken op basis van deze data gemaakt gaan worden is het aan te bevelen deze validatie alsnog uit te voeren.
- Er zat voor de meeste uitkomsten een ruime onzekerheidsmarge rondom de schattingen die uit het machine learning model kwamen. Deze marge is te verkleinen door de kwaliteit van de schattingen te vergroten. Een van de manieren om dit te doen is de testset waarop het model getraind wordt te vergroten. De testset in dit onderzoek bestond uit ruim duizend records. Over het algemeen gaat de kwaliteit van de schattingen met de wortel van de factor waarmee de testset wordt vergroot omhoog. Een voorbeeld: als we de testset 4 keer zo groot maken (dus 4000 records in plaats van 1000 records) dan gaat de kwaliteit van de schattingen met de factor 2 omhoog. Aandachtspunt hierbij is wel dat het samenstellen van de testset in veel gevallen handwerk is en daardoor erg arbeidsintensief.
- In dit onderzoek hebben we drie verschillende modellen getraind: één voor de dimensie algemeen/specifiek/onbekend, één de dimensie mbo/overig en één voor de dimensie ho/overig. Achteraf gezien waren de schattingen waarschijnlijk beter geweest als we één model hadden getraind waarin alle kruisingen van deze drie dimensies voorkwamen. Een belangrijke reden waarom dat in het huidige onderzoek niet is gebeurd is dat er hiervoor ook een grotere testset nodig is.
- Als er een construct gemeten moet worden waar een zekere mate van subjectiviteit in zit is het belangrijk om vooraf goed na te denken over de definitie hiervan. In het geval van dit onderzoek was dat de definitie van algemene of specifieke opleidingsvereisten. Stem met experts vanuit beleid en de praktijk af wat precies wordt bedoeld met het construct zodat bij het samenstellen van de testset hier rekening mee gehouden kan worden. Op deze manier zorg je ervoor dat het machine learning model goede input krijgt wat de kwaliteit van de schattingen vergroot en wat er daardoor ook voor zorgt dat de uitkomsten herkenbaar zijn voor de beleidsmedewerkers.
- In dit onderzoek heeft inhoudelijk vooral de focus gelegen op de onderzoeksvraag of de opleidingsvereisten in de vacatures algemeen of specifiek waren. Er zijn met behulp van de vacaturedata natuurlijk nog veel meer onderzoeksvragen te beantwoorden. Enkele voorbeelden waar wij aan moesten denken zijn:

- Is het taalgebruik in vacatures formeel of juist steeds vaker informeel?
- Hoe vaak worden tijdelijke of vaste contracten aangeboden?
- Hoe belangrijk is het hebben van werkervaring?
- Welke secundaire arbeidsvoorwaarden worden vaak genoemd in vacatures? Bijvoorbeeld op het gebied van hybride werken.

Bijlage 1 UWV dataset

Variabele	Omschrijving
ID	Unieke identificatiecode van de vacature
CREATIE_DATUM	Datum dat de vacature is geplaatst
BRON_CODE	Code die uniek is per leverancier
FUNCTIE_REF	Referentiecode van beroep in B&O-register
EIGEN_FUNCTIE_TITEL	Titel van de functie volgens leverancier
OMSCHRIJVING_L	Omschrijving van titel volgens leverancier
TITEL	Titel van beroep in B&O-register
FUNCTIEOMSCHRIJVING	Tekstveld voor de beschrijving van de functie, bedrijf etc.
WERVENDE_TEKST	Tekstveld voor de wervingstekst
TOELICHTING_ARBEIDSVOORWAARDEN	Tekstveld voor toelichting arbeidsvoorwaarden
OVERIGE_EISEN	Tekstveld voor overige eisen voor de functie
UUR_PER_WEEK_MIN	Minimaal aantal gevraagde uren per week
UUR_PER_WEEK_MAX	Maximaal aantal gevraagd uren per week
CONTRACTVORM	Tijdelijk, uitzicht op vast, vast
INDICATIE_WERKLOCATIE	Vaste locatie, zelfde regio of heel Nederland
WERKGEVER_NAAM_BEDRIJF	Naam bedrijf met de vacature
WERKGEVER_POSTCODE	Postcode (hoofdlocatie) bedrijf met de vacature
PLAATS_WERKLOCATIE	Plaats van de werklocatie
POSTCODE_WERKLOCATIE	Postcode van de werklocatie
IND_DOOR_WERKGEVER_DIRECT	Ingevoerd door tussenpersoon of werkgever

Bijlage 2 Tabellenset

B2.1 Schatting en 95% betrouwbaarheidsinterval voor het aandeel *Algemeen/Specifiek/Onbekend* voor het *mbo en ho*, 2010-2021

		Algemeen			Specifiek			Onbekend		
Niveau	Jaar	Waarde	Ondergrens	Bovengrens	Waarde	Ondergrens	Bovengrens	Waarde	Ondergrens	Bovengrens
Mbo	2010	0,41	0,33	0,48	0,40	0,32	0,47	0,20	0,16	0,24
	2011	0,39	0,32	0,46	0,39	0,32	0,46	0,23	0,19	0,27
	2012	0,41	0,34	0,47	0,38	0,31	0,45	0,21	0,18	0,25
	2013	0,42	0,34	0,49	0,39	0,32	0,46	0,20	0,16	0,23
	2014	0,42	0,35	0,50	0,39	0,32	0,47	0,18	0,15	0,22
	2015	0,42	0,35	0,49	0,41	0,33	0,48	0,17	0,14	0,21
	2016	0,40	0,33	0,47	0,40	0,33	0,47	0,20	0,17	0,24
	2017	0,41	0,34	0,48	0,40	0,32	0,47	0,19	0,16	0,23
	2018	0,40	0,33	0,47	0,41	0,33	0,48	0,19	0,16	0,23
	2019	0,40	0,33	0,47	0,39	0,32	0,46	0,21	0,17	0,24
	2020	0,39	0,32	0,46	0,40	0,33	0,47	0,21	0,18	0,25
	2021	0,37	0,31	0,44	0,36	0,30	0,43	0,26	0,23	0,30
Ho	2010	0,40	0,33	0,47	0,36	0,29	0,43	0,24	0,20	0,29
	2011	0,38	0,31	0,45	0,34	0,28	0,41	0,28	0,23	0,32
	2012	0,41	0,34	0,48	0,34	0,28	0,41	0,25	0,21	0,28
	2013	0,42	0,35	0,49	0,35	0,28	0,42	0,22	0,19	0,26
	2014	0,43	0,36	0,50	0,35	0,29	0,43	0,22	0,19	0,26
	2015	0,42	0,35	0,49	0,36	0,29	0,43	0,22	0,19	0,26
	2016	0,40	0,33	0,46	0,33	0,27	0,40	0,27	0,23	0,31
	2017	0,39	0,33	0,46	0,34	0,27	0,41	0,27	0,23	0,30
	2018	0,38	0,32	0,45	0,33	0,27	0,40	0,29	0,25	0,32
	2019	0,37	0,31	0,44	0,33	0,27	0,40	0,29	0,25	0,33
	2020	0,36	0,30	0,43	0,34	0,27	0,40	0,30	0,26	0,34
	2021	0,32	0,26	0,38	0,29	0,24	0,35	0,39	0,34	0,44

B2.2 Schatting en 95% betrouwbaarheidsinterval aandeel *Algemeen/Specifiek/Onbekend* per *sectorkamer* voor het *mbo*, gemiddelde 2017-2021

	Algemeen			Specifiek			Onbekend		
Sectorkamer	Waarde	Ondergrens	Bovengrens	Waarde	Ondergrens	Bovengrens	Waarde	Ondergrens	Bovengrens
Entree	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00
Handel	0,52	0,43	0,60	0,35	0,27	0,43	0,14	0,11	0,17
ICT en creatieve industrie	0,48	0,40	0,55	0,34	0,27	0,42	0,18	0,15	0,22
Mobiliteit, transport, logistiek en maritiem	0,45	0,38	0,53	0,32	0,25	0,39	0,23	0,20	0,27
Techniek en gebouwde omgeving	0,43	0,36	0,51	0,43	0,36	0,51	0,14	0,11	0,17
Voedsel, groen en gastvrijheid	0,39	0,33	0,46	0,38	0,31	0,44	0,23	0,20	0,27
Zakelijke dienstverlening en veiligheid	0,52	0,43	0,60	0,36	0,28	0,45	0,12	0,10	0,16
Zorg, welzijn en sport	0,31	0,22	0,41	0,61	0,50	0,71	0,08	0,05	0,13

B2.3 Schatting en 95% betrouwbaarheidsinterval aandeel *Algemeen/Specifiek/Onbekend* per *croho-onderdeel* voor het *ho*, gemiddelde 2017-2021

	Algemeen			Specifiek			Onbekend		
Croho-onderdeel	Waarde	Ondergrens	Bovengrens	Waarde	Ondergrens	Bovengrens	Waarde	Ondergrens	Bovengrens
Economie	0,50	0,42	0,58	0,32	0,24	0,40	0,19	0,15	0,22
Gezondheidszorg	0,44	0,36	0,51	0,44	0,36	0,52	0,13	0,10	0,16
Landbouw & Natuurlijke omgeving	0,28	0,22	0,35	0,40	0,33	0,47	0,32	0,28	0,37
Onderwijs	0,26	0,18	0,35	0,52	0,43	0,61	0,22	0,18	0,27
Recht	0,47	0,39	0,54	0,32	0,25	0,40	0,21	0,18	0,25
Sectoroverstijgend	0,42	0,35	0,49	0,29	0,23	0,36	0,28	0,24	0,32
Taal & Cultuur	0,45	0,38	0,52	0,38	0,31	0,45	0,17	0,14	0,21
Techniek	0,38	0,32	0,45	0,30	0,24	0,36	0,32	0,28	0,36

Bijlage 3 Indelingen BRC Beroepsklasse

Sector mbo	BRC Beroepsklasse	BRC Beroepsklasse omschrijving
Voedsel, groen en gastvrijheid	9	Agrarische beroepen
Voedsel, groen en gastvrijheid	11	Dienstverlenende beroepen
ICT en creatieve industrie	8	ICT beroepen
ICT en creatieve industrie	2	Creatieve en taalkundige beroepen
Handel	3	Commerciële beroepen
Handel	5	Managers
Techniek en gebouwde omgeving	7	Technische beroepen
Mobiliteit, transport, logistiek en maritiem	12	Transport en logistiek beroepen
Zakelijke dienstverlening en veiligheid	4	Bedrijfseconomische en administratieve beroepen
Zakelijke dienstverlening en veiligheid	6	Openbaar bestuur, veiligheid en juridische beroepen
Zorg, welzijn en sport	10	Zorg en welzijn beroepen
Zorg, welzijn en sport	1	Pedagogische beroepen
Entree	13	Overig

Croho-onderdeel	BRC Beroepsklasse	BRC Beroepsklasse omschrijving
Gezondheidszorg	10	Zorg en welzijn beroepen
Gedrag & Maatschappij	10	Zorg en welzijn beroepen
Economie	3	Commerciële beroepen
Economie	4	Bedrijfseconomische en administratieve beroepen
Landbouw & Natuurlijke omgeving	9	Agrarische beroepen
Natuur	9	Agrarische beroepen
Onderwijs	1	Pedagogische beroepen
Taal & Cultuur	2	Creatieve en taalkundige beroepen
Recht	6	Openbaar bestuur, veiligheid en juridische beroepen
Techniek	7	Technische beroepen
Techniek	8	ICT beroepen
Techniek	12	Transport en logistiek beroepen
Sectoroverstijgend	13	Overig
Sectoroverstijgend	11	Dienstverlenende beroepen
Sectoroverstijgend	5	Managers