**Portfolio Machine Learning**

**Student: Esmee Kraan Studentnummer: s1173469**

**Versie: 0.7 Datum:** 01-11-2024

Inhoudsopgave

[1 Mijn beroepstaken en -producten (logboek) 2](#_Toc181094625)

[2 Showcases 5](#_Toc181094626)

[2.1 Mijn showcases voor het data mining proces 6](#_Toc181094627)

[2.1.1 Showcase 1: Het Datamining Proces met CRISP-DM 6](#_Toc181094628)

[2.1.2 Showcase 2: <Titel> 8](#_Toc181094629)

[2.2 Mijn showcases voor ethiek 9](#_Toc181094630)

[2.2.1 Showcase 1: Opdracht F 9](#_Toc181094631)

[2.2.2 Showcase 2: <Titel> 9](#_Toc181094632)

[2.3 Mijn showcases voor business understanding 10](#_Toc181094633)

[2.3.1 Showcase 1: Huiswerkopgave - Boekclassificatie 10](#_Toc181094634)

[2.3.2 Showcase 2: <Titel> 12](#_Toc181094635)

[2.4 Mijn showcases voor data understanding 13](#_Toc181094636)

[2.4.1 Showcase 1: Huiswerkopgave - NL-FR Vertaling 13](#_Toc181094637)

[2.4.2 Showcase 2: <Titel> 15](#_Toc181094638)

[2.5 Mijn showcases voor data preparation 16](#_Toc181094639)

[2.5.1 Showcase 1: Themaopdracht 2 - AWL 16](#_Toc181094640)

[2.5.2 Showcase 2: <Titel> 18](#_Toc181094641)

[2.6 Mijn showcases voor modelling 19](#_Toc181094642)

[2.6.1 Showcase 1: Huiswerkopgave – Boekclassificatie 19](#_Toc181094643)

[2.6.2 Showcase 2: <Titel> 21](#_Toc181094644)

[2.7 Mijn showcases voor evaluation 22](#_Toc181094645)

[2.7.1 Showcase 1: Huiswerk opgave – Old or Not 22](#_Toc181094646)

[2.7.2 Showcase 2: <Titel> 24](#_Toc181094647)

[2.8 Mijn showcases voor deployment 25](#_Toc181094648)

[2.8.1 Showcase 1: Huiswerkopgave – Boekclassificatie 25](#_Toc181094649)

[2.8.2 Showcase 2: <Titel> 27](#_Toc181094650)

[3 Reflectie op professionele ontwikkeling tijdens het traject 28](#_Toc181094651)

[Bijlages 29](#_Toc181094652)

# Mijn beroepstaken en -producten (logboek)

**Instructie:**

In dit hoofdstuk geef je een overzicht van datgene wat jij hebt uitgevoerd in het kader van het semester machine learning*.*

Dit logboek kun je gebruiken om in het volgende hoofdstuk de showcases te schrijven waarmee je gaat onderbouwen waarom jij de leeruitkomsten hebt behaald.

|  |  |  |  |
| --- | --- | --- | --- |
| Thema-opdracht | Dit heb ik deze sprint gedaan | Tevreden (1-10) | Bruikbaar om deze leeruikomsten aan te tonen |
| 1 | Tijdens themaopdracht 1 hield ik me bezig met de **data understanding**. Ik onderzocht de datasets door te kijken naar datatypen, non-null waarden. Terwijl mijn groepsgenoten de data preparation uitvoerden, richtte ik me verder op het **filteren** van de data. Ik controleerde of het donker was door te kijken naar globale stralingswaarden (*Q*), of de verwarming uit stond via de kolom *'temp\_set\_\_degC'*, en vergeleek temperatuurverschillen tussen binnen en buiten. Uitschieters filterde ik er ook uit.  Daarna probeerde ik een **loss-functie** op te zetten met MSE, wat ik nog uitdagend vond omdat ik niet alles ervan begreep. Ik zette ook een functie op om per dag de Tau-waarde per huis te berekenen, wat complexer bleek dan verwacht en uiteindelijk ook niet werkte.  Een belangrijk punt dat ik meenam naar de volgende opdrachten is om stap voor stap te werken en tussentijds de resultaten te evalueren. Door tijdsdruk werkten we snel naar een resultaat toe, wat de logica en netheid van de uitkomst beïnvloedde. In het vervolg wil ik nauwkeuriger werken om het proces beter te begrijpen. | 7 | Data mining proces, ethiek, business understanding, data understanding, data preparation, modeling, evaluation, deployment |
| 2 | Voor de tweede themaopdracht schreef ik een script om **data in te lezen en om te zetten** naar een pandas DataFrame. Samen met mijn groep verkende we eerst het .yaml-bestand om de variabelen en hun betekenissen te begrijpen. Dit hielp ons om gerichte keuzes te maken in welke variabelen we wilden gebruiken bij het inlezen.  We werkten gezamenlijk aan **data understanding**, waarbij we weer keken naar de datatypen en mogelijke mankementen in kaart brachten. Voor de **data preparation** kon ik grotendeels gebruik maken van eerdere huiswerkopgaven, wat het proces versnelde.  Vervolgens maakte ik het multi-model voor deze opdracht.  Bij deze opdracht namen we meer tijd voor elke stap, wat leidde tot een beter begrip van de data en de preparation. Toch ervaarde ik uitdagingen bij het multi-model: hoewel ik een tip had gekregen om deze techniek te gebruiken, begreep ik niet volledig hoe deze werkte. Voor toekomstige projecten wil ik beter begrijpen wat elke stap precies doet, vooral bij modelling.  Ik werkte bij deze opdracht soms te snel, waardoor de groep het gevoel kreeg minder te kunnen bijdragen. Dit hebben we besproken, en we spraken af beter te communiceren en dat iedereen meer initiatief neemt, bijvoorbeeld door te experimenteren met meerdere modellen zodra de data prep klaar is. | 8 | Data mining proces, ethiek, business understanding, data understanding, data preparation, modeling, evaluation, deployment |
| 3 | In de laatste themaopdracht werkte ik in eerste instantie samen met de groep, totdat we merkten dat ons .txt-bestand, gebaseerd op een voorbeeld van Mischa, niet goed functioneerde. Ik ging op zoek naar een andere aanpak om het .docx-bestand te parseren en kwam uit bij **LlamaParse**, een tool die dit automatisch kan doen. We besloten toen over te stappen op een .json-formaat vanwege de kant-en-klare opmaak van zo’n bestandstype.  Daarna werkte ik met Iris en Raymond aan de app, waarbij we veel hebben getest om de juiste opzet te vinden en het **GEITJE**-model te installeren. We kwamen uiteindelijk tot een versie die algemene vragen redelijk kon beantwoorden en ook vragen over het OER enigszins accuraat beantwoorde. Iris en Raymond verbeterden dit verder zodat het model alleen reageerde op vragen over het OER.  Sanne en ik hebben toen een groot deel van de documentatie geschreven, waaronder business understanding, data understanding, data preparation, modelling, evaluation, en deployment. Raymond en Iris keken deze later na en vulden aan. Over het geheel genomen verliep dit project beter, mede doordat we duidelijk inzicht hadden in de data en de werking van het model.  Hoewel er opnieuw wat taakverdelingsproblemen waren, hielden we na de presentatie een retrospective waarin iedereen feedback kon geven aan elkaar. Dit leverde nuttige conclusies op, en we besloten deze manier van evalueren voortaan aan te houden voor de eindopdracht. | 8 | Data mining proces, ethiek, business understanding, data understanding, data preparation, modeling, evaluation, deployment |

|  |  |  |  |
| --- | --- | --- | --- |
| Eindopdracht Sprint | Dit heb ik deze sprint gedaan | Tevreden (1-10) | Bruikbaar om deze leeruikomsten aan te tonen |
| 1 |  |  | Data mining proces, ethiek, business understanding, data understanding, data preparation, modeling, evaluation, deployment |
| 2 |  |  | Data mining proces, ethiek, business understanding, data understanding, data preparation, modeling, evaluation, deployment |
| 3 |  |  | Data mining proces, ethiek, business understanding, data understanding, data preparation, modeling, evaluation, deployment |
| 4 |  |  | Data mining proces, ethiek, business understanding, data understanding, data preparation, modeling, evaluation, deployment |
| 5 |  |  | Data mining proces, ethiek, business understanding, data understanding, data preparation, modeling, evaluation, deployment |

# Showcases

**Instructie**

Jouw showcases zijn die leeruitkomsten, die jij in het semester machine learning hebt uitgevoerd en die het beste laten zien dat jij de leeruitkomsten

* Data mining process
* Ethiek
* Business understanding
* Data understanding
* Data preparation
* Modeling
* Evaluation
* Deployment

op het juiste niveau binnen een machine learning context beheerst.

De showcases die je beschrijft in dit portfolio zijn de basis voor de beoordeling. Voor elk van de acht hierboven genoemde leeruitkomsten moet je minimaal 2 showcases opvoeren. Doelgroep voor deze beschrijving: mensen met kennis van machine learning, die jou en de opdracht (nog) niet kennen.

In het beoordelingsformulier staat bij elke leeruitkomst waaruit de complexiteit blijkt. Goede showcases laten zien dat jij dit beheerst.

Geef per showcase in minimaal 1 tot maximaal 2 pagina’s een korte inleiding, leg gebruikte methoden/technieken kort uit, geef relevante en concrete voorbeelden. Voeg als bewijsmateriaal alle gerealiseerde beroepsproducten als bijlage aan het portfolio toe en verwijs hier in de showcase naar. Beschrijf in de showcase dus vooral het proces dat jij bij het uitvoeren van de leeruitkomst hebt doorlopen, dus hoe je dingen hebt aangepakt, welke keuzes je hebt gemaakt en waarom. Tijden de formatieve toetsmomenten krijg je feedback, neem deze mee in de beschrijving van je showcase.

## Mijn showcases voor het data mining proces

### Showcase 1: Het Datamining Proces met CRISP-DM

In deze showcase reflecteer ik op mijn aanpak binnen het data mining proces volgens de CRISP-DM-methodologie. Door eerdere ervaringen in de minor Data Science heb ik mijn kijk op het proces vernieuwd. In deze showcase bespreek ik hoe ik CRISP-DM nu gebruik in een machine learning context en mijn verbeterpunten voor een effectiever, gestructureerder proces.

**Introductie en Probleem Domein**

Het notebook begint met een introductie waarin ik het project en de te behandelen onderwerpen schets. Deze introductie geeft de lezer een duidelijke structuur en verwachtingen, wat helpt om de inhoud en het doel van het notebook direct helder te maken.

In de sectie “Probleem Domein” beschrijf ik de relevantie van het project en geef ik context aan het onderwerp met visuals. Hiermee ontstaat er een beter beeld van de casus, en de lezer begrijpt waarom dit project uitgevoerd wordt en welke vragen het beantwoordt.

**Business Understanding**

In deze fase werk ik het doel van de klant uit. Hier identificeer ik de problemen die het project oplost en definieer ik belangrijke succesfactoren (KSF’s) en KPI’s om het eindresultaat te kunnen beoordelen. Ik stel mezelf vragen zoals:

* "Weet ik welke vraag ik ga oplossen?"
* "Zijn de KSF’s en KPI’s van het project duidelijk?"
* "Begrijp ik de context en het doel van de vraagstelling?"

Deze vragen helpen me om mijn aanpak in de Business Understanding-fase te verdiepen. Voorheen behandelde ik deze fase als een snelle inventarisatie, maar nu neem ik meer tijd om de opdracht goed te begrijpen voordat ik verderga. Dit voorkomt mogelijke fouten en helpt me om vanaf het begin gericht te blijven werken.

**Vereiste Libraries**

In dit deel beschrijf ik de gebruikte libraries en geef ik aan hoe deze geïnstalleerd kunnen worden. Door packages te benoemen en toe te lichten, bied ik documentatie en instructies voor het reproduceren van het project. Ik vermeld hoe Anaconda kan worden geïnstalleerd, voeg een lijst toe van libraries met hun functies en geef installatiestappen voor de lezer. Dit bevordert de reproduceerbaarheid en toegankelijkheid van het project voor anderen.

**Data Understanding**

In de Data Understanding-fase voer ik een stapsgewijze analyse uit van de dataset. Ik begin met een verkenning van de data om inzicht te krijgen in de structuur en mogelijke mankementen. In plaats van alleen outputs te genereren, leg ik bij elke stap uit waarom ik de analyses uitvoer. Ik beschrijf bijvoorbeeld de resultaten van *df.head()* en *df.describe()* om ontbrekende waarden of foutieve gegevens in kaart te brengen. Deze aanpak maakt het proces inzichtelijker, en een verbeterpunt voor mezelf is om deze analyse af te sluiten met een overzicht van bevindingen.

**Data Preparation**

De Data Preparation-fase bestaat uit het opschonen van de data om deze klaar te maken voor modellering. Hier documenteer ik de stappen die ik doorloop om de data te prepareren en geef ik toelichting bij elke stap. Dit helpt anderen om mijn werkwijze te begrijpen en het proces nauwkeurig te volgen. Ook hier ben ik begonnen met een template voor meer consistentie, zodat ik de data preparation-fase in volgende projecten efficiënter en betrouwbaarder kan doorlopen.

**Modelling**

In deze sectie beschrijf ik de gebruikte modellen en licht ik de keuzes toe die ik maak. Ik bespreek welke modellen ik heb geselecteerd, waarom ik voor specifieke instellingen heb gekozen en hoe deze keuzes aansluiten op de data en doelstellingen. Documentatie is belangrijk in deze fase om de reproduceerbaarheid en begrijpelijkheid van de resultaten te waarborgen. Een verbeterpunt is om uitgebreider te verantwoorden waarom ik voor bepaalde parameters kies en wat ze doen.

**Evaluation**

In de Evaluation-fase beoordeel ik de prestaties van de modellen aan de hand van visualisaties. Hier reflecteer ik op wat wel en niet heeft gewerkt en geef ik een overzicht van mogelijke vervolgstappen. Door deze evaluatie in de context van de projectdoelen te plaatsen, schets ik een duidelijk beeld van de effectiviteit van mijn aanpak. Daarnaast wil ik in de toekomst mijn beslissingen en mijn werkwijze in deze fase beter toelichten, zodat ik mijn keuzes consistent kan onderbouwen.

**Deployment**

Door een gestructureerde en gedocumenteerde opzet van het notebook, maak ik de overdracht van mijn werk al eenvoudiger. Dit helpt bij de toekomstbestendigheid van het project. In eerdere projecten werd deployment slechts als advies opgenomen, maar in dit project besteed ik ook aandacht aan onderhoudbaarheid, beveiliging en de technische vereisten voor het model in een operationele setting. Showcase 1 biedt hierin meer detail.

Een belangrijke toevoeging voor reproduceerbaarheid is het creëren van een requirements.txt-bestand met versies van de gebruikte packages en een samenvatting van de code in één codeblokc. Dit helpt anderen om mijn notebook probleemloos te hergebruiken en snel te runnen.

**Conclusie**

Door CRISP-DM stapsgewijs en zorgvuldig toe te passen, heb ik me beter in elke fase van het data mining proces kunnen verdiepen. Door kritischer na te denken over de context en reproduceerbaarheid, bied ik nu een vollediger en flexibelere werkwijze voor toekomstige projecten. Met dit verbeterde proces zorg ik ervoor dat de eindresultaten niet alleen voldoen aan de huidige eisen, maar ook bijdragen aan een toekomstbestendige en herbruikbare oplossing.

**Bewijsmateriaal:** Zie *Voorbeeldnotebook.ipynb* voor de template van het proces wat ik boven heb benoemd.

### Showcase 2: <Titel>

## Mijn showcases voor ethiek

### Showcase 1: Opdracht F

### Showcase 2: <Titel>

## Mijn showcases voor business understanding

### Showcase 1: Huiswerkopgave - Boekclassificatie

In deze showcase reflecteer ik op de business understanding die ik formuleerde voor de boekclassificatieopdracht en hoe ik dit kan verbeteren voor toekomstige projecten. Het doel van de boekclassificatieopdracht was om een neuraal netwerk te ontwikkelen dat het genre van een boek voorspelt op basis van tekstuele samples uit Project Gutenberg.

Afbeelding met tekst, schermopname, diagram, Lettertype

Automatisch gegenereerde beschrijving**Het Probleem Domein**

De context van het probleem heb ik duidelijk vastgesteld: het classificeren van boeken in verschillende genres was van groot belang voor bibliotheken en boekwinkels. De keuze voor tekstuele samples van Project Gutenberg bood een rijke dataset, maar ik erkende dat de ongelijke verdeling van genres (zie Figuur 2.3.1.1) en variatie in schrijfstijlen uitdagingen met zich meebrachten. Dit waren twee belangrijke punten waar over nagedacht moest worden voor het maken van het model.

**Business Understanding**

Het doel van het project, het ontwikkelen van een neuraal netwerk voor genrevoorspelling, was helder. Echter ik kon dieper ingaan op de specificaties van de businessvraag. Wie is de doelgroep? In welke context gaat het model gebruik worden? Etc.

**Wat ging ik oplossen?**

De centrale vraag was duidelijk gedefinieerd. Toch zou een bredere context of een meer gedetailleerde uitleg over de specifieke behoeften van de eindgebruikers waardevol zijn. Wat waren hun verwachtingen? Hoe konden ze de voorspellingen gebruiken in hun dagelijkse praktijk? Eigenlijk hetzelfde verbeterpunt wat ik bij de business understanding opnoem, net wat meer diepgang.

Figuur 2.3.1.1: Genre verdeling in 1 v/d 3 dataframes

**Kritische Succesfactoren (KSF's) en Key Performance Indicators (KPI's**

De focus op de nauwkeurigheid van het model was een relevante KSF, maar ik kon deze verder uitwerken. Waarom was een nauwkeurigheid boven de 50% gekozen als doel? Een betere onderbouwing van deze keuze, bijvoorbeeld door te kijken naar benchmarks in vergelijkbare projecten, zou de validiteit van deze KSF versterken.De KPI's waren benoemd, maar een beter beeld op hoe deze indicatoren zouden kunnen worden gemeten was handig geweest.

**Toepasbaarheid van de Huidige Data**

Het was goed dat ik de datakwaliteit en variabiliteit erkende, maar er moest meer aandacht zijn voor specifieke datamanagementstrategieën. Hoe zou ik omgaan met de ongelijke verdeling van genres in de dataset? Waren er technieken zoals oversampling of undersampling die ik kon overwegen om deze uitdaging te mitigeren? Ging ik werken met de hele dataset, ging ik een subsetmaken, ging ik met batchverwerking werken. Dit zijn belangrijke dingen om van te voren enigzins duidelijk te hebben.

**Verbeterpunten**

Hoewel deze business understanding met een goed is beoordeeld heb ik toch nagedacht over welke dingen ik beter wil doen voor toekomstige projecten. Een aantal punten die naarboven kwamen waren:

1. **Uitbreiden van de Context**

Ik had de behoeften en verwachtingen van de eindgebruikers die baat hadden bij het model verder moeten verkennen. Dit zou hebben geholpen bij het verder concretiseren van de businessvraag.

1. **Diepere Analyse van KSF's**

Ik moet meer context en onderbouwing geven aan de gekozen KSF's. Dit kon door benchmarks of best practices in de sector te onderzoeken en deze te relateren aan de projectdoelen.

1. **Specificeren van KPI's**

Ik had de methodologie voor het meten van KPI's moeten definiëren en een baseline vaststellen voordat het model werd getraind. Dit zou een referentiepunt voor evaluatie bieden.

1. **Strategieën voor Datakwaliteit**

Ik had een plan moeten implementeren voor het omgaan met datakwaliteit, inclusief technieken voor het balanceren van de dataset en het verwerken van ontbrekende waarden.

Door deze reflectie op de business understanding te maken, kon ik mijn aanpak verder verfijnen en ervoor zorgen dat de machine learning oplossing niet alleen technisch robuust was, maar ook goed aansloot bij de behoeften van de organisatie en haar gebruikers. Hoewel ik technisch gezien alle stappen van de rubric netjes ben langsgegaan mist er wat diepgang, dat wil ik beter doen bij het groepsproject.

****Bewijsmateriaal****: Zie bijlage boekclassificatie\_esmee\_kraan.pdf voor de uitgewerkte business understanding.

### Showcase 2: <Titel>

## Mijn showcases voor data understanding

### Showcase 1: Huiswerkopgave - NL-FR Vertaling

Voor mijn eerste showcase in Data Understanding heb ik gekozen om mijn werk aan het Nederlands-naar-Fries vertaalmodel te beschrijven. Bij deze opdracht heb ik feedback op eerdere projecten verwerkt om de dataset nauwkeuriger te analyseren en vroegtijdig mogelijke problemen in kaart te brengen. In deze showcase maak ik inzichtelijk hoe ik de kwaliteit en betekenis van de data onderzocht en afwijkingen en mankementen verklaarde en corrigeerde.

**Proces en Verbeterpunten**

Uit eerdere feedback bleek dat ik de data understanding-fase vaak te beperkt hield, en dat belangrijke analyses pas in de data preparation-fase werden uitgevoerd. Bij deze opdracht besloot ik dat te verbeteren door de dataset grondig te verkennen en mogelijke mankementen al vroeg te analyseren. Hierdoor kon ik de kwaliteit van de data op meerdere niveaus evalueren, wat een stevige basis legde voor het project en ons einddoel.

**Verkenning en Controle van Data**

Ik begon met een verkenning van de dataset om inzicht te krijgen in de structuur, datatypen en variabelen. Dit gaf me de kans om de betekenis van de data beter te begrijpen en maakte het makkelijker om afwijkingen gericht op te sporen. Vervolgens analyseerde ik de verdeling van de vertaalde zinnen om te controleren of er een gelijk aantal Nederlandse en Friese zinnen was, wat essentieel is om het model gebalanceerd te trainen (zie figuur 2.4.1.1).

Afbeelding met tekst, schermopname, diagram, Rechthoek

Automatisch gegenereerde beschrijvingToen ik zag dat deze verdeling al 50/50 was, kon ik verder met andere verdelingen, zoals de minimale en maximale lengte van zinnen in karakters en woorden. Deze aanpak bood mij een bredere kijk op de dataset en mogelijke problemen die ik anders wellicht over het hoofd zou hebben gezien.

Afbeelding met tekst, schermopname, Lettertype, menu

Automatisch gegenereerde beschrijving

Tijdens deze analyse ontdekte ik dat sommige Nederlandse zinnen slechts drie karakters hadden, terwijl sommige Friese zinnen maar één karakter bevatten. Door dit soort afwijkingen vroeg op te sporen, kon ik de kwaliteit van de dataset proactief bewaken. Ik vond deze variaties opmerkelijk en besloot deze rijen nader te onderzoeken (zie figuur 2.4.1.2). Ik printte alle rijen met drie karakters in het Nederlands en één karakter in het Fries.

Figuur 2.4.1.2: Rijen met Min. karakters

Figuur 2.4.1.1: Verdeling FY en NL

De meeste Nederlandse zinnen klopten, maar de Friese variant was incorrect, *“in de wandeling”* werd vertaal naar *“e”*. Ook controleerde ik rijen met twee karakters (Zie Figuur 2.4.1.3) en nam deze door met mijn Friese grootouders om de inhoud nauwkeuriger te kunnen beoordelen. Met hun hulp filterde ik nog meer verkeerde rijen uit de dataset.

Afbeelding met tekst, schermopname, menu

Automatisch gegenereerde beschrijving *“op die wijze”* zou google translate naar *“op dy wize”* vertalen maar *“sa”* is dus wel een logische vertaling omdat het eigenlijk “zo” betekent, wat we in de Nederlandse taal ook wel gebruiken om iets aan te duiden, voor bijvoorbeeld op die manier, zo. Dit soort rijen had ik zelf wellicht kunnen uitvogelen met mijn minimale fries, maar ik wou toch iemand betrekken die zelf echt de taal spreekt.

Bij *“per kilo”*, *“per ons”* en *“per pond”* ging het dus vaak fout dat het woord kilo, ons en pond niet na “de” of “it” stond. Ook zoiets wat me zelf zou zijn ontschoten. Het feit dat de vertaling dus een heel woord mistte was dan ook de reden dat ik die rijen heb verwijderd.

Figuur 2.4.1.3: Rijen met 2 FY karakters

Afbeelding met tekst, schermopname, Lettertype

Automatisch gegenereerde beschrijvingVoor de analyse van de woordlengtes onderzocht ik afwijkingen in zowel lange als korte zinnen. Sommige korte zinnen met richtingaanwijzingen leken soms niet helemaal correct (zie Figuur 2.4.1.4), maar ik besloot ze te behouden omdat ze in bepaalde contexten wel logisch konden zijn. In de korte Friese zinnen bleek alles inhoudelijk kloppend. Deze stappen toonden mijn bewuste keuzes voor het wel of niet behouden van contextuele data.

**Verdere Inconsistenties**

Afbeelding met tekst, schermopname, Lettertype

Automatisch gegenereerde beschrijvingNaast de lengteanalyse heb ik ook non-null waarden, unieke waarden en de meest voorkomende zinnen onderzocht. Door deze analyse uit te voeren, kon ik nauwkeurig afwijkingen en herhalingen ontdekken die de betekenis en kwaliteit van de data sterk kunnen beïnvloeden met hun frequentie. In de Nederlandse kolom kwam bijvoorbeeld de zin "Eigenlijk moeten we bruinbrood eten..." 192 keer voor. De vertalingen leken echter volgens mijn grootouders in grote lijnen te kloppen, daarom heb ik deze zin er niet uitgefilterd.

Figuur 2.4.1.4: Rijen met richtingaanwijzingen

Ook vond ik de Friese zin *"trije by fiif trije by fiif meter"* 32 keer, die vaak verkeerd werd vertaald (Zie figuur 2.4.1.5). Omdat dit soort herhalingen en getalafwijkingen het model konden misleiden, heb ik deze rijen uit de dataset verwijderd. Hiermee zorgde ik ervoor dat het model zich niet op herhaalde fouten baseert en zo een vertekend beeld opbouwt.

Hoewel ik in theorie had kunnen vertrouwen op de door docenten aangeleverde dataset, bracht mijn diepergaande analyse toch belangrijke problemen aan het licht. Door een uitgebreide data understanding-analyse en kritische blik op afwijkingen en mankementen kon ik diverse fouten en onnodige rijen verwijderen, wat de kwaliteit van de dataset en daarmee de modelprestaties ten goede zou komen. Dit proces laat zien dat ik in staat ben om een dataset op waarde te schatten en potentiële problemen vroegtijdig en gericht aan te pakken.

Figuur 2.1.4.5: Aantal vertaalde rijen van de Friese zin

**Bewijsmateriaal:** Zie bijlage *NL\_FY\_esmee\_kraan.pdf* voor de uitgewerkte analyses, methodes en bijbehorende visualisaties van deze data understanding.

### Showcase 2: <Titel>

## Mijn showcases voor data preparation

### Showcase 1: Themaopdracht 2 - AWL

Voor de eerste showcase Data Preparation heb ik gekozen voor themaopdracht 2 van AWL, een groot gedeelte van de data preparation heb ik op me genomen. Dit project had als doel om een model te ontwikkelen dat in staat is om pakketten automatisch te segmenteren en te classificeren op basis van hun materiaal en vorm. Met een dataset van ongeveer 1000 afbeeldingen van verschillende verpakkingen, heb ik een gestructureerde aanpak gehanteerd om de data voor te bereiden.

Afbeelding met tekst, schermopname, diagram, Perceel

Automatisch gegenereerde beschrijving**Data Extractie en Organisatie**

Bij de Data Understandings fase hebben Annefien en ik een aantal foutieve rijen geïdentificeert waarbij we vermoedde dat of de vorm of het materiaal niet klopte. In figuur 2.5.1.1 Laten we alle combinaties zien van de vorm en materiaal. De combinaties van materialen en verpakkingen waren over het algemeen logisch, alleen hadden we onze twijfels bij een papieren doos en een plastic doos. Ook waren er 5 undefined rijen. Bij nader onderzoek bleken de papieren- en plastic doos enigszins te kloppen. Daarom hebben we die er niet uitgefilterd, maar het laat wel zien dat we in staat zijn om foutieve data te vinden en eventueel te filteren.

Bij de data preparation heb ik allereerst de vijf rijen met `*undefined`* waarden gefilterd, waardoor het aantal rijen terugging van 4936 naar 4931. Daarna heb ik bepaalde kolommen, zoals *mask*, *min\_box*, *max\_box*, *roi*, en *roi\_center*, niet noodzakelijk waren voor het trainen van het model. Deze gegevens waren overbodig en heb ik weggelaten om de dataset te vereenvoudigen en te optimaliseren. Door me te concentreren op de essentiële informatie in de kolommen *shape*, *material*, *path*, en *crop\_part*, heb ik ervoor gezorgd dat de dataset geschikt en geoptimaliseerd was voor de Modelling fase.

Figuur 2.5.1.1: Plot met combinaties van vorm en materiaal

Om de kwaliteit van de data te waarborgen, heb ik het pad van de afbeeldingen geüpdatet naar het volledige pad, zodat alle afbeeldingen correct konden worden geladen. Dit was een belangrijke stap, aangezien foutieve paden zouden leiden tot missende waarden bij het inladen van de data. En omdat ik deze stap ook had gedaan bij de huiswerkopgave, waar een deel van de code vanaf was geleid. Vervolgens heb ik de dataset gesplitst in een train- en testset.

Afbeelding met tekst, schermopname, menu, ontwerp

Automatisch gegenereerde beschrijving**Verwerking en Voorbereiding**

Bij het verwerken van de afbeeldingen heb ik de functie van het huiswerk als voorbeeld gebruik en herschreven, deze was verantwoordelijk is voor het uitsnijden, normaliseren, en desgewenst verhogen van de helderheid van de afbeeldingen. Door afbeeldingen te normaliseren naar waarden tussen 0 en 1, heb ik ervoor gezorgd dat de data in een geschikt formaat is voor het model.

Vervolgens heb ik de afbeeldingen georganiseerd in een gestructureerde mappenstructuur (zie Figuur 2.5.1.2), die de labels voor vorm en materiaal weerspiegelt. Hierdoor zijn de afbeeldingen eenvoudig toegankelijk voor verdere verwerking en training van het model. Dit zorgde ervoor dat zowel de enhanced als unenhanced datasets goed georganiseerd waren en klaar voor gebruik in de machine learning pipeline.

**Kwaliteitscontrole en Evaluatie**

Afbeelding met tekst, Lettertype, schermopname, zwart-wit

Automatisch gegenereerde beschrijvingOm de kwaliteit van de verwerkte data te evalueren, heb ik de shape van de geladen afbeeldingen en labels gecontroleerd (Zie Figuur 2.5.1.3). De resultaten bevestigden dat zowel de training- als testsets correct waren ingeladen, met aantallen die overeenkwamen met de verwachtingen: 3994 afbeeldingen voor training en 987 voor testen. Deze controle toont aan dat de data-extractie en verwerking succesvol waren, zonder missende of foutieve waarden.

Afbeelding met schermopname, collage

Automatisch gegenereerde beschrijvingDaarnaast heb ik willekeurige afbeeldingen geplot met hun respectieve labels en materialen om visueel te bevestigen dat de gegevens correct waren opgeslagen en verwerkt (Zie Figuur 2.5.1.4). Dit helpt bij het waarborgen van de nauwkeurigheid van de labels en de relevantie van de data voor het trainingsproces.

Figuur 2.5.1.2: Mappenstructuur

Figuur 2.5.1.3: Shape Images en Labels

**Conclusie**

In dit data preparation proces heb ik aangetoond dat ik in staat ben om data effectief te extraheren uit een databron, deze te organiseren in een bruikbaar formaat voor machine learning, en rekening te houden met de kwaliteit van de data. Door zorgvuldige filtering van onbruikbare gegevens, het waarborgen van correcte paden, en het verwerken en organiseren van afbeeldingen op basis van labels, heb ik een solide basis gelegd voor de verdere ontwikkeling van een Neuraal Netwerk Model om de paketten te classificeren.

Figuur 2.5.1.4: Plot met willekeurige afbeeldingen en labels

**Bewijsmateriaal:** Zie bijlage *awl\_groep5.pdf* voor de uitgewerkte analyses, methodes en bijbehorende visualisaties van deze data preparation.

### Showcase 2: <Titel>

## Mijn showcases voor modelling

### Showcase 1: Huiswerkopgave – Boekclassificatie

In de modeling fase van dit project heb ik verschillende modellen voor tekstclassificatie ontwikkeld en geëvalueerd. Het doel was om een dataset van teksten in zeven verschillende genres te categoriseren. Door gebruik te maken van verschillende architecturen en technieken, zoals SimpleRNN, LSTM, GRU, en convolutionele netwerken, heb ik inzicht gekregen in hun prestaties en beperkingen. Deze showcase beschrijft de aanpak die ik heb gevolgd, de keuzes die ik heb gemaakt, en de resultaten van elk model.

Afbeelding met tekst, schermopname, diagram, Lettertype

Automatisch gegenereerde beschrijving****Data Preparation****

Voordat ik met de modeling fase begon, heb ik de dataset geanalyseerd en geprepareerd. De dataset bestond uit een verzameling teksten, elk gelabeld met een genre. Ik heb de dataset verkend door gebruik te maken van statistieken en visualisaties om de distributie van genres, de lengte van de teksten en andere relevante kenmerken te begrijpen (Zie Figuur 2.6.1.1).

Vervolgens heb ik Ongewenste tekens, dubbele waarden en ontbrekende gegevens zijn verwijderd. Ook zijn teksten genormaliseerd door ze om te zetten naar kleine letters. Daarna heb ik de woorden omgezet naar numerieke waarden door middel van tokenization, en vervolgens heb ik een embedding-laag geïmplementeerd voor de vectorrepresentatie van de woorden.

****Modeling****

Ik heb verschillende modellen gebouwd om de prestaties van verschillende architecturen te vergelijken en om te experimenteren.

Figuur 2.6.1.1: Verdeling genres in training dataframe

*Tabel 1: Gerealiseerde modellen boekclassificatie*

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Model** | **Architectuur** | **Training Accuracy** | **Validatie Accuracy** | **Opmerkingen** |
| Model 0.0: Eenvoudig Model | Embedding-laag, Twee Dropout-lagen, Flattening, Dense output-laag | 65,14% | 63,54% | Stabiel. |
| Model 0: SimpleRNN | Embedding-laag, SimpleRNN-laag, Dense output-laag | 80.8% | 17.89% | Overfitting zichtbaar. |
| Model 1: 1 LSTM Layer | Embedding-laag, LSTM-laag, Dense output-laag | 86.15% | ~50% | Overfitting zichtbaar. |
| Model 2: 2 LSTM Layers | |  | | --- | |  |  |  | | --- | | Embedding-laag, 2 LSTM-lagen, Dense output-laag | | 98.48% | ~55.75% | Duidelijke overfitting; val\_loss stijgt. |
| Model 3: GRU met Dropout | Embedding-laag, GRU-laag met dropout, Dense output-laag | 99.19% | ~54-55% | Overfitting zichtbaar, maar dropout helpt. |
| Model 4: Conv1D Model | Embedding-laag, Conv1D-lagen, MaxPooling, Dense output-laag | 99.48% | ~55-57% | Stijging val\_loss; moet beter generaliseren. |
| Model 5: Verbeterde LSTM | Bi-directionele LSTM met dropout, Dense output-laag | 83.20% | ~55% | Bi-directionele netwerken waren uitdagend. |
| Model 6: Verbeterd Conv Model | Convolutioneel model met Dropout, Dense output-laag | ~99.XX% | ~56-58% | Succesvol maar overfitting blijft een probleem. |

**Evaluatie en Confusion Matrix**

Na het trainen van de modellen heb ik hun prestaties geëvalueerd door een confusion matrix te genereren. Deze matrix hielp bij het identificeren van welke genres vaak verkeerd werden gecategoriseerd. Het was belangrijk om de prestaties van elk model op deze manier te visualiseren, aangezien het hielp bij het begrijpen van de sterke en zwakke punten van elk model.

****Conclusie en Reflectie****

Deze modeling fase gaf me de kans om verschillende neurale netwerkarchitecturen te onderzoeken en hun effectiviteit in tekstclassificatie te vergelijken. Ik heb geleerd dat complexere modellen zoals LSTM en GRU veelbelovende resultaten kunnen opleveren, maar ook dat ze vatbaar zijn voor overfitting. Het integreren van technieken zoals dropout en early stopping was cruciaal om de generalisatie van de modellen te verbeteren.

In de toekomst zal ik meer nadruk leggen op het gebruik van validatie tijdens de training en op het optimaliseren van hyperparameters. Dit project heeft me ook doen inzien dat het tijdmanagement bij het trainen van modellen belangrijk is, vooral bij het werken met grotere datasets.

****Bewijsmateriaal****: Zie bijlage boekclassificatie\_esmee\_kraan.pdf voor de uitgewerkte analyses, methodes en bijbehorende visualisaties van de modellings fase.

### Showcase 2: <Titel>

## Mijn showcases voor evaluation

### Showcase 1: Huiswerk opgave – Old or Not

Voor deze showcase beschrijf ik het evaluatieproces van een leeftijdsclassificatie-model dat ik heb gemaakt voor de Old or Not opdracht. De bedoeling was om te zien of een machine learning-model de leeftijd van gezichten kon inschatten, specifiek om te bepalen of iemand ouder dan 42 jaar was. De dataset die ik gebruikte bevatte gezichten van beroemdheden (21 jaar en ouder), en ik heb twee verschillende CNN-modellen (convolutionele neurale netwerken) geprobeerd om te zien welk beter presteerde.

**Methode en Technieken**

Voor de evaluatie gebruikte ik de volgende technieken:

* **Test Accuracy:** Voor een snelle kijk op hoe goed het model werkte op de trainings- en validatie dataset.
* **Confusion Matrix (CM):** Om te zien hoeveel fouten het model maakte bij het labelen van “oud” of “niet oud.”
* Afbeelding met tekst, schermopname, diagram, Rechthoek

  Automatisch gegenereerde beschrijving**Individuele Voorspellingen:** Ik probeerde het model uit met enkele persoonlijke foto’s om te zien hoe het reageerde op nieuwe data.

Figuur 2.7.1.4: CM Model 2

Figuur 2.7.1.3: Train- en Val Plot Model 2

Figuur 2.7.1.1: Train- en Val Plot Model 1

Figuur 2.7.1.2: CM Model 1

Afbeelding met tekst, Perceel, lijn, diagram

Automatisch gegenereerde beschrijvingAfbeelding met tekst, schermopname, diagram, Rechthoek

Automatisch gegenereerde beschrijvingAfbeelding met tekst, Perceel, lijn, diagram

Automatisch gegenereerde beschrijving

**Evaluatieproces**

Het evaluatieproces heb ik in stappen aangepakt, beginnend bij het eerste model die ik heb gemaakt met behulp van de stof besproken in de werkcolleges.

Het eerste model had een simpele architectuur en bereikte een test accuracy van 65.58% (Zie figuur 2.7.1.1). Dit gaf me een basisniveau, maar ik merkte dat het model vooral moeite had met foutieve classificaties, oftewel veel false positives en false negatives (Zie Figuur 2.7.1.2).

Daarna heb ik onderzoek gedaan naar andere studies die een soort gelijk CNN-Model hebben gemaakt. Uit de studie *“Gender Prediction and Age Estimation using CNN from Face Images”* heb ik het tweede model gebasseerd.Het tweede model was iets complexer en verbeterde de test accuracy lichtjes naar 66.85% (Zie figuur 2.7.1.3). De aanpassing leidde tot minder fouten, maar ze waren nog steeds hoog genoeg om de betrouwbaarheid in twijfel te trekken.

Met de confusion matrixen ontdekte ik dat het tweede model minder fouten maakte dan het eerste, maar er was nog steeds een hoog aantal false positives en false negatives (Zie figuur 2.7.1.4). Dit liet zien dat beide modellen nog niet ideaal waren om in de praktijk te gebruiken.

Model 1 had wel bijvoorbeeld veel last van overfitting. Het model heeft specifieke patronen in de trainingsdata sterk opgepikt, maar kan die kennis niet goed toepassen op nieuwe data. Hierdoor presteert het minder goed op de validatieset. Bij Model 2 liggen de training en validation accuracy juist dichter bij elkaar, model 2 is dus minder vatbaar voor overfitting en stabieler dan Model 1.

Tot slot testte ik het model met een paar persoonlijke foto’s (Zie Figuur 2.7.1.5) en merkte dat beide modellen vaak foutief “oud” voorspelde. Dit kon betekenen dat de modellen misschien gevoelig waren voor de soort belichting die ik gebruikte en accessoires zoals een helm en zonnebril.

Afbeelding met tekst, kleding, zonnebril, Menselijk gezicht

Automatisch gegenereerde beschrijving

**Resultaten en Aanbevelingen**

Na het evalueren van beide modellen, bleek dat beide versies een redelijke basis legden maar nog veel ruimte voor verbetering hadden.

Voor toekomstige stappen adviseerde ik om de modellen verder te trainen en meer te experimenteren met bijvoorbeeld de tuning, en om meer data te gebruiken voor meer representatie.

Figuur 2.7.1.5: De uitkomsten van beide modellen op 3 selfie’s van mijzelf.

**Conclusie**

Het evaluatieproces liet me zien wat goed werkte en wat niet, en gaf me praktische ideeën voor verbeteringen. Dit proces toonde mijn vaardigheid om machine learning-oplossingen te beoordelen, de waarde ervan in te schatten en te onderzoeken en experimenteren om tot betere resultaten te komen.

**Bewijsmateriaal:** Zie bijlage *Old\_or\_Not\_esmee\_kraan.pdf* voor de uitgewerkte analyses, methodes en bijbehorende visualisaties van de Evaluation

### Showcase 2: <Titel>

## Mijn showcases voor deployment

### Showcase 1: Huiswerkopgave – Boekclassificatie

Het doel van de boekclassificatie-opdracht was een neuraal netwerk te ontwikkelen dat automatisch het genre van een boek kan voorspellen op basis van tekstsamples uit Project Gutenberg. Deze showcase behandelt het advies dat ik opgesteld heb voor de deployment van dit model, met aandacht voor onderhoudbaarheid en overdraagbaarheid. Deze deployment was met een voldoende beoordeeld, ik heb hem gekozen voor deze showcase om ook in te gaan op de stappen die ik moet ondernemen om dit naar een goed te halen.

**Samenvatting van de Modellen en Keuzes**

Tijdens de modellering heb ik verschillende modellen getest, waaronder Eenvoudige modellen, RNN’s, LSTM’s, GRU’s en convolutionele netwerken (zie Hoofdstuk 2.6.1 en tabel 1). Model 0.0 bleek het best presterende model, met een val\_acc van 64%, zonder de risico’s op overfitting die ik bij complexere modellen zoals LSTM en GRU had. Deze modellen bleken weliswaar beter tijdens training, maar hadden veel last van overfitting en een hogere belasting op geheugen en CPU, wat ze minder geschikt maakt voor een praktische deployment.

**Deploymentaanpak en Verbeterpunten**

Bij de het deployment gedeelte van deze opdracht gaf ik een algemene conclusie waar ik de resultaten van elk model evalueerde en gaf ik aanbevelingen voor een aantal vervolgstappen. De vervolgstappen gingen in op schaalbaarheid, batchverwerking en enigszins over onderhoudbaarheid van het model in de praktijk. Ook lichtte ik toe waarom sommige resultaten en experimentaties gelimiteerd waren. Bij de onderstaande kopjes ga ik per punt wat meer in op wat ik heb gedaan voor de opdracht en waar de verbeterpunten liggen voor het groepsproject.

**Schaalbaarheid en Batchverwerking**

Voor de deployment raadde ik aan om met een batch-verwerkingssysteem te werken om grote hoeveelheden data in delen te kunnen verwerken, wat verwerkingskracht en tijd bespaarde. Deze keuze zou het model al redelijk schaalbaar in productie maken. Zelf begon ik met een dataset van 100K rijen zonder batch-verwerkingssysteem wat al gauw een probleem bleek bij de experimentatie vanwege mijn laptops gebrek aan kracht.

In toekomstige projecten wil ik beginnen met een kleine subset van de dataset, waardoor ik sneller kan experimenteren en modelkeuzes kan finetunen zonder de hoge belasting van een grote dataset. Dit stelt me in staat om eerder bepaalde hyperparameters en architecturen te bepalen, wat leidt tot een meer experimentatie en dus een beter advies. En als ik tevreden ben over een bepaald model kan ik een batch-verwerkingssysteem toepassen om het model te trainen op een grotere dataset.

**Onderhoudbaarheid en Modelbeheer**

Model 0.0 is relatief onderhoudbaar vanwege de eenvoud van het model (zie Tabel 1). Het model was ook snel te trainen, waardoor het geschikt is voor bijwerking en tuning zonder zware aanpassingen. Omdat ik minder met documentatie werkte, was er echter minder duidelijkheid in hoe het model efficiënt geüpdatet en gefinetuned kan worden.

Voor een soepelere overdracht en onderhoudbaarheid zal ik in toekomstige projecten mijn aanpak en keuzes beter documenteren, inclusief alle preprocessing-stappen en de modelparameters. Met deze documentatie, kan een ander teamlid of externe gebruiker het model eenvoudiger recreëren, beheren en updaten.

**Documentatie en Overdraagbaarheid**

Voor deze opdracht heb ik beperkte documentatie opgezet, waardoor het model overdraagbaar is op basisniveau, maar voor een ervaren teamlid mogelijk nog extra uitleg zou vergen.

In toekomstige projecten zal ik volledigere documentatie aanleveren, inclusief een README met uitleg over de architectuur en configuratie-instructies voor de benodigde software en hardware om het model te draaien en te testen. Dit helpt andere teamleden en een opdrachtgever om het model probleemloos over te nemen en te begrijpen zonder mijn hulp.

**Conclusie**

**Voor de toekomst ga ik de deployment zeker anders aanpakken, momenteel keek ik er meer naar als een kort advies over hoe iemand verder zou kunnen gaan met een model en de notebook zonder enig detail. Maar het gaat veel dieper dan dat, het is een heel advies over niet alleen het model zelf. Maar ook over de onderhoudbaarheid, security en de daadwerkelijke deployment van het model. Hoe zou je dit model bijvoorbeeld toepassen in een praktische situatie, voor deze opdracht heb ik daar niet over nagedacht. Maar het is wel relevant voor de opdrachtgever.**

**Bewijsmateriaal**: Zie bijlage *boekclassificatie\_esmee\_kraan.pdf* voor de uitgewerkte conclusie en deployment advies.

### Showcase 2: <Titel>

# Reflectie op professionele ontwikkeling tijdens het traject

**Instructie**

Om de beroepstaken succesvol uit te kunnen voeren, heb je bepaalde professionele vaardigheden nodig. In dit hoofdstuk reflecteer je op je professionele ontwikkeling van dit proof of concept project en beschrijf je aan de hand van de uitgevoerde beroepstaken in minimaal 1 tot maximaal 2 pagina’s hoe en waarom jouw professionele vaardigheden hierin naar voren komen. Daarbij richt je je op de volgende professionele vaardigheden:

1. Persoonlijk leiderschap: aanpak en initiatief.

Vragen die je jezelf kunt stellen: heb ik mij altijd goed voorbereid? Zo ja, hoe dan en zo nee, waarom niet? Welke doelen heb ik mezelf bij het uitvoeren van de beroepstaken gesteld? Hoe heb ik ervoor gezorgd dat ik deze doelen ook heb behaald? Zijn er situaties geweest waarin het anders liep dan ik vantevoren bedacht had en hoe ben ik daar mee omgegaan? Heb ik hulp gevraagd als ik dat nodig had?

1. Lerende en onderzoekende houding: keuzes maken.

Vragen die je jezelf kunt stellen: heb ik bij het uitvoeren van de taken nog alternatieven overwogen of onderzocht (aanpak / methode / techniek)? Welke keuzes heb ik gemaakt en op basis waarvan? Heb ik misschien in eerste instantie verkeerde keuzes gemaakt? Hoe heb ik deze hersteld?

1. Communicatie: effectiviteit

Vragen die je jezelf kunt stellen: hoe heb ik met (gast)docenten en medestudenten gecommuniceerd en was dat altijd de beste manier? Waarom wel/niet? Zijn er situaties geweest waarin ik feedback heb gekregen of heb ik zelf in een bepaalde situatie feedback gegeven? Hoe ging dat en welk effect heeft dat gehad? Hoe staat het met mijn schriftelijke taalvaardigheid?

1. Zelfreflectie: inzicht in effecten van eigen handelen.

Vragen die je jezelf kunt stellen: wat heb ik in de afgelopen periode over mezelf geleerd? Wat ging goed en wat ging minder goed? Wat ga ik de volgende keer anders doen, waarom en hoe? Wat blijf ik vooral hetzelfde doen en waarom?

Je professionele ontwikkeling wordt formatief getoetst. Dit betekent dat je hier wel op beoordeeld wordt, maar dat de beoordeling niet wordt meegenomen in het cijfer voor je portfolio.

Bijlages

Uitwerkingen huiswerkopdrachten

Uitwerkingen themaopdrachten

Uitwerkingen ethiekopdrachten

Uitwerking Proof of Concept inclusief:

* Projectplan
* Ontwerpdocumentatie
* Code
* Vrijgaveadvies
* Voorgangsrapportages
* ….