**Template – Use Case identificeren**

**Mijlpaal 1: Wat is het doel van de use case?**

|  |
| --- |
| *Beschrijf wat de use case is en welk doel je wilt bereiken* |

**Bouwstenen** (Deze zijn alleen ter inspiratie, je antwoorden moeten in het lege vak hierboven staan)

|  |
| --- |
| Hoe ga je je doel meten?   * Procentuele omzetstijging * Verbeterde klantwaardering * Verhoogde efficiëntie (vermindering van tijd en/of kosten, verbeterde kwaliteit) * Concurrentievoordeel realiseren (verhoogd marktaandeel, verbeterde marktpositie) * Bedrijfsrisico's verminderen |

|  |
| --- |
| Wat is de reden om AI in dit geval te overwegen?   * Wat is de huidige situatie en het belangrijkste probleem? * Wat is de gewenste situatie en waarom zou dit een oplossing zijn? |

|  |
| --- |
| Op wie zal dit project een impact hebben en op welke manier?   * Wie zijn de belangrijkste belanghebbenden? (Intern en extern aan uw bedrijf.) * Wat zijn de behoeften van het beoogde publiek? |

|  |
| --- |
| Wat zijn de voordelen en risico's van het benaderen van deze zaak met AI?   * Wat zijn voordelen? * Wat zijn risico's? |

**Mijlpaal 2: De vraag definiëren die je wilt beantwoorden**

|  |
| --- |
| *Formuleer de vraag die de AI-toepassing moet beantwoorden:* |

**Bouwstenen**

|  |
| --- |
| Wat is de zakelijke vraag die je wilt beantwoorden? We willen...   * Beschrijven... * Ontdekken... * Voorspellen... * Voorschrijven...   Is er maar één vraag of valt de use case uiteen in meerdere vragen?   * Denk na over de volgorde van vragen en nesting (fractale benadering) |

|  |
| --- |
| Wat zijn de grenzen van de vraag?   * Waar is het model wel/niet van toepassing?   + Voorbeeld: Europa ja, Azië nee. Medische industrie: ja, maakindustrie: nee. * Kenmerken binnen de dataset die grenzen vertegenwoordigen?   + Voorbeeld: Geslacht, leeftijd, land * Product of dienst: Voor welk product of dienst is het model? |

|  |
| --- |
| Wat zijn de beperkingen van de vraag?   * Tijdigheid: hoe snel of frequent kun je het model trainen en evalueren? * Juridisch: welke ethische of juridische dilemma's kun je verwachten? * Interpretabiliteit: Is het belangrijk dat mensen de afwegingen begrijpen (White box) of is uitleg minder belangrijk en meer resultaatgericht? (Black box) * Technisch: gegevenscomplexiteit, gegevensgrootte, gegevensopslag of -snelheid. |

**Mijlpaal 3: Wat is het probleemtype?**

|  |
| --- |
| *Overweeg en identificeer de aard van het probleem.* |

**Bouwstenen**

|  |
| --- |
| **Supervised Learning**  Een benadering om een algoritme te laten leren hoe gegevensinvoer gerelateerd is aan (vooraf bepaalde) uitvoer. Het algoritme wordt getraind op basis van voorbeelden waarin de gewenste uitvoer bekend is, vandaar de term 'supervised'. Er zijn veel verschillende algoritmen binnen deze categorie, maar op een algemeen niveau worden ze gebruikt om ofwel [1] gegevens in te delen in een bepaald aantal klassen ('classificatie') of [2] een (numerieke) waarde te voorspellen, meestal aangeduid met de term 'regressie'. In beide gevallen is er gelabelde uitvoer om het algoritme mee te trainen, maar het verschil is het type uitvoer dat het algoritme moet modelleren (klassen of waarden).   1. Classificatie: Gegevens indelen in een bepaald aantal klassen (voor voorspelling of interpretatie). 2. Regressie:    1. Voorspelt een doelwaarde op basis van gerelateerde gegevenskenmerken.    2. Schat het effect van een factor(en) op de doelstelling. |

|  |
| --- |
| **Unsupervised learning**  Bottom-up benadering voor het vinden van patronen in de gegevens. In tegenstelling tot supervised learning (classificatie en regressie hierboven) wordt de gewenste output niet gespecificeerd door de gebruiker, maar wordt de doorvoer gespecificeerd c.q. hoe de computer de gegevens kan verwerken om bepaalde patronen en output te vinden. Soorten unsupervised learning die relatief vaak worden gebruikt:   1. Clusteren: wordt gebruikt om afzonderlijke groepen gegevens te vinden. Dit kan worden toegepast op elk type gegevens (getallen, tekst, afbeeldingen, ...) 2. Dimensionaliteitsreductie: wordt gebruikt om het aantal dimensies te verminderen. Dit kan bijvoorbeeld worden toegepast om een multidimensionale vector in 2D of 3D te visualiseren, of om veel vragen in een enquête terug te brengen tot een paar onderliggende concepten. 3. Process mining: wordt gebruikt om procesgegevens te beschrijven op basis van de feitelijke gegevens. 4. Collaborative filtering: gebruikt om gegevens op individueel niveau te filteren op basis van collectieve patronen. |

|  |
| --- |
| **Reinforcement Learning**  Gebruikt om prestaties te evalueren op basis van een beoogd resultaat en dienovereenkomstig te belonen/ bestraffen. Dit probleemtype is belangrijk voor algoritmen die taken uitvoeren die menselijk gedrag nabootsen. |

**Mijlpaal 4: Kies een oplossingsmethode voor het probleemtype**

**4.1** **Classificatie**

|  |
| --- |
|  |

**Bouwstenen**

|  |  |  |
| --- | --- | --- |
| **Algoritme** | **Uitleg** | **Voorbeeld Use Case** |
| **Logistic Regression** | Schat de relatie tussen een binair doel (bijvoorbeeld waar of onwaar) en een of meer kenmerken door de waarschijnlijkheid van een van beide opties tussen 0 en 1 te geven. | Credit scoring: Riskant of niet  Hotelmanagement: Gast zal annuleren of niet |
| **k-Nearest Neighbours** | Gebruikt de nabijheid van gegevenspunten om klassen te maken met de laagste afstand tot punten in dezelfde klasse en de maximale afstand tot een andere klasse | Geneeskunde: Heeft kanker, vereist meer aandacht en geen kanker (Meer dan 2 klassen)  Productaanbod: groepeer vergelijkbare producten in productgroepen |
| **Decision Tree or Forest** | Deel waarnemingen herhaaldelijk in op basis van meerdere verstrekte kenmerken en de waarschijnlijkheid dat dat verschil nauwkeurig is | Marktonderzoek: Consumenten categoriseren op basis van een doel/doel  Management: Helpt bij het nemen van beslissingen in uiteenlopende contexten |
| **Neural Network** | Een reeks algoritmen die een onderliggende relatie in gegevens probeert te ontdekken via een structuur die menselijke neuronen nabootst | Beweging: Een computer die klassen van acties detecteert  Zicht: Augmented reality die specifieke objecten detecteert  Robotica: Roomba kiest de meest efficiënte richting (links, rechts, terug, vooruit) |
| **Support Vector Machine** | Een supportvectormachine probeert het optimale hypervlak (grens tussen gegevensgroepen) te vinden dat wordt gebruikt om gelabelde gegevens te scheiden en te classificeren. | Detectiesystemen: Spamdetectie per e-mail, fraudedetectie  Gezichtsdetectie: Kan worden gebruikt om onderscheid te maken tussen gezichten en niet gezichten, door vierkantjes rond gezichten te plaatsen. |
| **Naïve Bayes** | Ervan uitgaande dat kenmerken onafhankelijk van elkaar zijn, waarbij elk kenmerk wordt gebruikt als een individuele classificeerder voor een classificatietaak | Tekstanalyse: Kan tekstinvoer gebruiken om onderwerpen of onderliggende concepten te voorspellen  Real-time voorspellingen: Aangezien dit algoritme efficiënt is, kan het in-realtime en effectief worden gebruikt. |
| **LLMs/GPT Models** | Getraind op grote hoeveelheden tekst, waardoor ze op menselijke wijze tekst kunnen genereren op basis van de patronen die ze hebben geleerd. Door de context van invoerprompts te verwerken en te begrijpen, kunnen LLM's samenhangende reacties produceren, vragen beantwoorden en helpen bij een verscheidenheid aan tekstgebaseerde taken. | Sentimentclassificatie: Sentiment in klantbeoordelingen detecteren, zoals negatief/positief of boos/twijfelachtig/gelukkig, enz. Inhoudsclassificatie: Het genre van een boek of verhaal bepalen door de inhoud te analyseren en het in te delen in categorieën zoals romantiek, mysterie of sciencefiction.  Opmerking: beide voorbeelden gaan ervan uit dat er vooraf gedefinieerde labels voor categorieën zijn. |

**4.2** **Regressie**

|  |
| --- |
|  |

**Bouwstenen**

|  |  |  |
| --- | --- | --- |
| **Algoritme** | **Uitleg** | **Voorbeeld Use Case** |
| **Linear Regression** | Schat de relatie tussen een doel en meerdere kenmerken met behulp van een rechte lijn | Financiële prestaties: Winstgevendheid of promotionele effectiviteit voorspellen  Marktonderzoek: Interpretatie van consumentengedrag aan de hand van geregistreerde gegevens over hun acties |
| **Time Series Forecasting** | Een voorspellingsmethode die toekomstige uitkomsten probeert te voorspellen op basis van historische uitkomsten en de consistente overdracht van dynamiek van relevante voorspellers | Financiële prestaties: Historische prestaties die worden gebruikt om toekomstige prestaties te voorspellen, zodat trends kunnen worden onderzocht. |
| **Stochastic Gradient Descent (Optimization)** | Algoritme benadert iteratief de gradiënt in een enkele stap bij het bepalen van classifiers tot het de laagst mogelijke fout bereikt | Kostenreductie: kosten minimaliseren gegeven gegevenskenmerken.  Inkomsten verbeteren: Winstmaximalisatie met behulp van factoren die informatie geven over producten, vraag, aanbod, enz. |
| **Neural Network** | Een reeks algoritmen die een onderliggende relatie in gegevens probeert te ontdekken door middel van een structuur die menselijke neuronen nabootst | Management: Verkoop-, winst-, omzet- en kostencijfers voorspellen  Visie: Augmented reality meting van objecten detecteren  Robotica: Roomba kiest de meest efficiënte richting (360 graden) |
| **Decision Tree or Forest** | Deel waarnemingen herhaaldelijk in op basis van meerdere verstrekte kenmerken en de waarschijnlijkheid dat dat verschil nauwkeurig is | Marktonderzoek: Consumenten indelen in categorieën op basis van leeftijd, inkomen of andere numerieke kenmerken |

**4.3** **Unsupervised learning**

|  |
| --- |
|  |

**Bouwstenen**

|  |  |  |
| --- | --- | --- |
| **Algoritme** | **Uitleg** | **Voorbeeld Use Case** |
| **k-Means Clustering** | Dit algoritme probeert clusters te vinden op basis van een vooraf gedefinieerd aantal k clusters, waarbij gebruik wordt gemaakt van de gegevenspunten die het dichtst bij elkaar liggen. | Marktonderzoek: klantsegmentatie (Customer metrics)  Supply chain management: Groepeer leveranciers op basis van relatiestatus of risicoclassificatie |
| **F-P Growth (Pattern or Association Algorithm)** | Dit algoritme ontwikkelt een frequentieboom of -patroon die de meest frequente associatiepatronen tussen kenmerken in een dataset markeert. | Marktonderzoek: Ontdek merk- of productassociaties, gebaseerd op hoe het algoritme ze verbindt |
| **Process discovery** | Dit algoritme vertaalt loggegevens naar een procesmodel | Marktonderzoek: ontdek hoe klanten interageren met de organisatie, welke stappen worden genomen in de klantenzorg, welke 'journeys' mensen afleggen. |
| **k-Modes Clustering** | Dit algoritme is een variatie op k-means die specifiek omgaat met niet-numerieke gegevens. In plaats van waarnemingen te clusteren op basis van dichtstbijzijnde gemiddelden, worden waarnemingen gegroepeerd op basis van de modus (hoe vaak een waarneming voorkomt) van die categorieën. | Marktonderzoek: Maakt clustering op basis van categorieën mogelijk, wat kan worden gebruikt om privacy en gegevens te beschermen door specifieke metingen niet te tonen |
| **Gaussian Mixture Model** | GMM gaat ervan uit dat er een mengsel is van k componenten, elk met een kansverdeling en k mengselgewichten, die vervolgens worden gebruikt om waarnemingen te groeperen | Marktonderzoek: Klantsegmentatie als subpopulaties niet kunnen of zijn geïdentificeerd |
| **Principal Component Analysis (PCA)** | Dit is een gegevenstransformatietechniek die de gegevens wijzigt naar een nieuw coördinatensysteem zodat correlaties tussen gelijksoortige dimensies kunnen worden gegroepeerd of verwijderd. | Marktonderzoek: controleer of de juiste concepten zijn gemeten  Klantreis: Informatieverzameling stroomlijnen door het aantal gestelde vragen of gebruikte enquêtes te verminderen  Algemeen: visualiseer gegevens in 2D of 3D als de gegevens die je wilt visualiseren multidimensionaal zijn (bijv. 40-dimensionaal) |
| **Latent Dirichlet Analysis (LDA)** | Een techniek voor de verwerking van natuurlijke taal die de aanwezigheid van woorden telt die betrekking hebben op vooraf gedefinieerde onderwerpen in een set documenten. | Natuurlijke taalverwerking: Ontdek belangrijke onderwerpen en hun frequentie in tekstblokken  Natuurlijke taalverwerking: Tekstblokken samenvoegen die door een groep mensen worden gebruikt om in te schatten of hun mening positief of negatief is in die context |
| **Isolation Forest (Anomaly Detection)** | Het algoritme gebruikt beslissingsbossen als basis, maar splitst gegevens vervolgens op om ervoor te zorgen dat slechts 1 gegevenspunt wordt uitgefilterd bij elke splitsing. | Anomaliedetectie: Kan worden gebruikt om unieke problemen of obstakels in gegevens te detecteren, zoals specifieke gevallen van verkeerde invoer of gebeurtenissen. |
| **Collaborative filtering** | Dit algoritme gebruikt gegevens op groepsniveau om gegevens voor een individuele entiteit te filteren. | Dit algoritme wordt veel gebruikt in aanbevelingssystemen waarbij een individu een aanbeveling krijgt op basis van de 'consumptie' van relatief vergelijkbare klanten. |
| **LLMs/GPT Models** | Getraind op grote hoeveelheden tekst, waardoor ze op menselijke wijze tekst kunnen genereren op basis van de patronen die ze hebben geleerd. Door de context van invoerprompts te verwerken en te begrijpen, kunnen LLM's samenhangende reacties produceren, vragen beantwoorden en helpen bij een verscheidenheid aan tekstgebaseerde taken. | Tekstblokken categoriseren: Nieuwsartikelen indelen in onderwerpen zoals politiek, sport of entertainment op basis van hun inhoud.  Anomalieën identificeren: Het identificeren van spam e-mails door de LLM e-mails te laten classificeren als spam of legitiem op basis van patronen in hun tekst.  Algemeen/basis gebruik: LLM's/GPT-modellen worden veelvuldig gebruikt om tekst te genereren wanneer er een vraag wordt gegeven. Dit is in wezen "unsupervised learning", waarbij het model een invoer krijgt en de uitvoer moet zelf bepalen. |

**Mijlpaal 5: Gegevens begrijpen**

|  |
| --- |
| Beschrijf de gegevens die nodig zijn voor de AI-oplossing |

**Bouwstenen**

|  |
| --- |
| Welke gegevens zijn nodig om de bedrijfsvraag te beantwoorden?   * Welke gegevens zijn beschikbaar, hebben jullie hier toegang toe? * Welke gegevens zijn niet beschikbaar en hoe komen we hieraan? |

|  |
| --- |
| Wat zijn de vereiste kenmerken van de gegevens?   * Wat is er nodig om deze kenmerken te identificeren? * Gestructureerd / Semi- / Ongestructureerd |

|  |
| --- |
| Hoe moeten deze gegevens worden geïnterpreteerd?   * Syntaxis en semantiek? Etikettering? * Weerspiegelen de gegevens de essentie van de omgeving die ze vertegenwoordigen? |

|  |
| --- |
| Wie is de eigenaar van de gegevens?   * Toegang en gebruik van de gegevens, zijn er beperkingen rond de gegevens? Voorbeeld: GDPR |

|  |
| --- |
| Hoe worden vooroordelen in de gegevens geïdentificeerd? |

**Mijlpaal 6: Voorbereiding en validatie van gegevens**

*Hoe bereid je je gegevens voor?*

* Schone gegevens (gegevensverdelingen visualiseren)
* Nieuwe attributen genereren die nodig zijn voor analyse (ook wel feature engineering genoemd)
* Gegevens integreren (samenvoegen en aggregeren)

*Hoe valideer je je gegevens?*

* Is de omvang van de gegevens voldoende in vergelijking met de populatie?
* Is er een steekproefbias?
* Controleer of de gegevens een correlatie hebben met de doelstelling? (Correlatie betekent geen causaliteit, maar kan een indicator zijn?
* Controleren of de gegevens een betrouwbaar kenmerk bevatten om te voorspellen en te verifiëren?
* Hebben de gegevens transformatie nodig om te verwerken? Welke coderingstechnieken passen hierbij?
* Andere...

**Mijlpaal 7: Statistische evaluatie**

**Belangrijkste statistische gegevens**

|  |
| --- |
| *Kies minimaal twee metrieken (indien van toepassing, beschrijf anders subjectieve factoren om resultaten zonder supervisie te evalueren).* |

**Building blocks**

|  |  |  |
| --- | --- | --- |
| **Classificatie Metrisch** | **Berekening** | **Uitleg** |
| Confusion matrix | Werkelijk vs Voorspeld | Diagram showing: True Positives, True Negatives, False Negatives, False Positives. |
| Accuracy | (True Positives + True Negatives) / Total events | Aantal correcte voorspellingen vergeleken met het totaal aantal voorspellingen. |
| Precision | True Positives / (True Positives + False Positives) | Aantal correcte positieve voorspellingen (verhouding waar versus voorspeld positief) |
| Recall | True Positives / (True Positives + False Negatives) | Aantal correcte voorspellingen voor de positieven (rantsoen ware versus werkelijke positieven) |

|  |  |  |
| --- | --- | --- |
| **Regression Metric** | **Berekening** | **Uitleg** |
| Mean Absolute Error (MAE) | For each prediction P – A = Diff | Som van het absolute verschil tussen Werkelijk en Voorspeld. |
| Mean Square Error (MSE) | For each prediction P – A = Diff2 | Som van het absolute verschil in het kwadraat tussen Werkelijk en Voorspeld. |
| R Squared (R2) | MSE / Constant Baseline = R2 | Hoeveel beter dan constante basislijn? 0 = niet beter , 1 = perfect |

|  |  |  |
| --- | --- | --- |
| **Cluster Metric** | **Berekening** | **Uitleg** |
| Within Cluster Sum of Squares Error (WSS) | *WSS*=∑∑∣∣*x*−*μi*​∣∣2 | WSS meet de compactheid van de clusteringoplossing, waarbij een lagere WSS aangeeft dat datapunten dichter bij de middelpunten van hun respectieve clusters liggen, wat wijst op een potentieel betere clustering. |
| Between Cluster Sum of Squares (BSS) | *BSS*=∑*i*=1*k*​∣*Ci*​∣⋅∣∣*μi*​−*μ*∣∣2 | BSS meet de scheiding tussen clusters, waarbij een hogere BSS aangeeft dat clusters verder uit elkaar liggen, wat duidt op een beter onderscheid en mogelijk een geschiktere clustering. |
| F-Ratio | BSS / WSS | De intuïtie achter de F-verhouding is dat een goede clusteringoplossing een hoge interclustervariantie heeft (clusters zijn goed gescheiden) en een lage interclustervariantie (gegevenspunten binnen een cluster liggen dicht bij elkaar). Daarom wijst een grotere F-verhouding op een potentieel betere clusteringoplossing. |
| Elbow Plot  Door het aantal clusters uit te zetten tegen een clustercriterium (meestal de Within Cluster Sum of Squares of WSS), ontstaat er een "elleboogpunt" waar de snelheid van afname sterk verandert. Deze "elleboog" vertegenwoordigt een buigpunt waar het toevoegen van meer clusters leidt tot afnemende opbrengsten in termen van model fit. Idealiter is het optimale aantal clusters het punt waar de curve begint af te vlakken, vergelijkbaar met de hoek in een elleboog, vandaar de naam. | | |

|  |
| --- |
| Dimensional Reduction bevat geen metrieken, omdat deze probleemtypes geen doel bevatten. Deze probleemtypes worden gebruikt om nieuwe inzichten uit gegevens te genereren. |

**Mijlpaal 8: Bedrijfsevaluatie**

**Hoe zou je evalueren of de oplossing geschikt is voor het doel?**

|  |
| --- |
| *Wat zijn de belangrijkste criteria om de oplossing te beoordelen? Beantwoordt de oplossing de vraag?* |

**Hoe zou je evalueren of het ethisch en legaal is?**

|  |
| --- |
| *Overweeg het EU-kader voor betrouwbaarheid en de komende wettelijke vereisten* |

|  |
| --- |
| *Zou het gebruik van een AI-oplossing de omgeving negatief beïnvloeden?*  *(Denk bijvoorbeeld aan het effect van een vals negatief tegenover een vals positief in de omgeving)* |

