

## Template 1 (DATAB4)

Vul per groep één keer onderstaande template in en lever deze in op Blackboard

Student informatie	
1. Mohamed Bouallala(2238016)	
2. Rayan Hassan(2208612)	
3. Luuk de Kramer (2213573)	
4. Mohammed Bogatyrev (2182116)	

**1. Theoretisch kader**

**Theoretisch kader deel 1: Wat is data mining? (max. 100 woorden)**

Data mining is het analyseren van grote hoeveelheden data om verborgen patronen en verbanden te ontdekken, vaak met als doel betere beslissingen te nemen of voorspellingen te doen. Dankzij AutoML-tools is dit nu ook mogelijk zonder veel programmeerkennis. Twee veelgebruikte technieken zijn supervised en unsupervised learning. Bij supervised learning leert een model van gelabelde data (bijvoorbeeld e-mails met het label spam/niet-spam). Unsupervised learning zoekt zelfstructuur in ongelabelde data, zoals het groeperen van klanten op basis van koopgedrag. Beide technieken helpen om inzichten uit data te halen zonder handmatige analyse.

Theoretisch kader deel 2a (verplichte onderwerpen):

**Hoe werken classificatie, regressie en clustering? (max. 500 woorden)**

Classificatie is een veelgebruikte techniek binnen data mining waarbij data wordt ingedeeld in vooraf gedefinieerde categorieën. Het algoritme leert dit door een model te trainen met gelabelde voorbeelden. Bijvoorbeeld: als je een dataset hebt met foto's van katten en honden, leert het model de verschillen tussen beide op basis van kenmerken zoals vorm, kleur of grootte. Zodra het model getraind is, kan het nieuwe, onbekende data correct toewijzen aan een categorie (bijv. "kat" of "hond"). Classificatie wordt vaak toegepast in de praktijk, bijvoorbeeld voor het herkennen van spam-mails, het analyseren van medische diagnoses of het herkennen van objecten in beelden. Populaire algoritmes hiervoor zijn beslisbomen, support vectormachines en random Forest. Het doel is altijd om op basis van beschikbare kenmerken van een object of situatie zo nauwkeurig mogelijk te voorspellen tot welke groep het hoort.

Regressie is een techniek binnen het gebied van data mining waarbij je een voorspelling probeert te doen aan de hand van bepaalde gegevens. Het wordt gebruikt om verbanden tussen verschillende variabele binnen een dataset te vinden. Het kan namelijk lastig zijn om de verschillende relaties tussen datapoints te lezen en te begrijpen. Er zijn verschillende soorten regressies. De eerste is lineaire regressie, hierbij wordt geanalyseerd tussen afhankelijke variabele (vaak gekenmerkt met Y) en onafhankelijke variabele (vaak gekenmerkt met X). Dit wordt dan gevisualiseerd met een lineaire lijn. De afhankelijke variabele is dan continue en de onafhankelijke variabele kan continue zijn maar ook onderbrekend. Op deze manier krijg je een inzicht in de relatie tussen de variabele. Dan als tweede heb je non-lineaire regressie. De naam zegt het al het is alles behalve lineair, het kan dus vele vormen aannemen. Hier vallen weer verschillende types onder zoals: logistic regression, ridge regression en Lasso regression.

Clustering is een techniek waarbij je groepen probeert te vinden in de data, zodat er groepen vergeleken kunnen worden. Clustering is bijzonder nuttig bij het analyseren van grote hoeveelheden ongestructureerde of onbekende data, waarbij je vooraf nog geen indeling kent. Het wordt vaak

toegepast op gegevens zoals beroep, leeftijd, inkomen, klantgedrag of geografische locatie. Typische toepassingen zijn: Marktsegmentatie, klanten groeperen op basis van koopgedrag. Beeldverwerking, objecten of regio's groeperen in een afbeelding. Detectie van afwijkingen, clusters gebruiken om abnormale patronen te herkennen.

Theoretisch kader deel 2b (extra onderwerpen): Hoe werken ...? (max. 400 woorden)

### **Onderwerp1:**

#### **Associatieregels**

Naast technieken zoals classificatie, regressie en clustering zijn er binnen data mining ook andere methoden die gericht zijn op het ontdekken van verbanden in data. Eén van die technieken is het gebruik van associatieregels.

**Associatieregels:** Dit wordt ook wel market basket-analyse genoemd. Dit type datamining zoekt naar relaties tussen variabelen. Associatieregels kunnen bijvoorbeeld de verkoopgeschiedenis van een bedrijf bekijken om te zien welke producten het vaakst samen worden gekocht, waardoor bedrijven en onderzoekers weloverwogen beslissingen kunnen nemen op basis van de inzichten die uit de analyse zijn afgeleid. Het concept van associatieregels wordt veel gebruikt op verschillende gebieden, zoals analyse van marktmandjes, mining van webgebruik, inbraakdetectie en bio-informatica, om er maar een paar te noemen.

Een klassiek voorbeeld is dat klanten die luiers kopen ook vaak bier kopen. Deze onverwachte combinatie werd ontdekt via associatieregels. De zogenaamde "bier-en-luiers" case kreeg veel aandacht binnen de wereld van retailanalyse en groeide uit tot een klassiek voorbeeld van hoe data-analyse onverwachte patronen kan onthullen en strategische besluitvorming kan ondersteunen. Het onderstreepte ook het belang van inzicht in klantgedrag en het afstemmen van winkelindelingen om de verkoop te optimaliseren en de winkelervaring te verbeteren.

Hoewel associatieregels krachtige inzichten kunnen opleveren, is het belangrijk om kritisch te blijven.

### **Onderwerp2:**

#### **Sequential Pattern**

Sequential Pattern is een data mining-techniek die bedoeld is om gelijkaardige patronen te ontdekken in datatransacties die zich over een bepaalde periode of in de loop der tijd voordoen. Het primaire doel van deze methode is bedrijfsanalisten te ondersteunen bij het identificeren van verbanden tussen gegevens die zich in een opeenvolgende volgorde ontwikkelen. Daarbij worden patronen opgespoord waarin gebeurtenissen of items in een vaste volgorde voorkomen binnen datasets

Binnen het domein van Web Usage Mining spelen sequential patterns een belangrijke rol bij het analyseren van gebruikersgedrag en navigatiegedrag. De techniek wordt specifiek ingezet om sequentiële sessies te herkennen, waarbij gebruikers pagina's na elkaar bezoeken binnen een bepaalde tijdsvolgorde. Door bijvoorbeeld veelvoorkomende klikroutes of handelingen te herkennen, kun je knelpunten in de UI opsporen. Je kunt dan de website aanpassen zodat gebruikers sneller vinden wat ze zoeken, wat leidt tot hogere tevredenheid.

De kennis die uit deze patronen voortkomt, is bijzonder nuttig voor webbeheerders en marketers. Zij kunnen hiermee trends voorspellen, de structuur van de website optimaliseren of gerichte marketingstrategieën en advertenties ontwikkelen.

## 2. Bronvermelding

Onderwerp 1 + 2a(classificatie):

- Bron 1: Qlik. (z.d.). *What is Data Mining?* Geraadpleegd op 10-5-2025, van <https://www.qlik.com/us/data-analytics/data-mining>
- Delua, J. (z.d.). *Supervised versus unsupervised learning: What's the difference?* IBM. Geraadpleegd op 10 mei 2025, van <https://www.ibm.com/think/topics/supervised-vs-unsupervised-learning>

Onderwerp 2a (regressie):

- Bron 1: Vatsal, S. (2025, 4 februari). *Regression in data mining: A complete overview*. Geraadpleegd op 09-05-2025, van <https://unstop.com/blog/regression-in-data-mining#:~:text=The%20term%20'regression'%20refers%20to,as%20well%20as%20other%20variables>
- Bron 2: Janse, B. (2024, 3 september). *Regressie analyse: de uitleg plus voorbeeld* - Toolshero. Toolshero. Geraadpleegd op 09-05-2025, van <https://www.toolshero.nl/marketing-modellen/regressie-analyse/>

Onderwerp 2a (clustering):

- Bron 1: GeeksforGeeks. (2025, 24 april). *Data mining cluster analysis*. GeeksforGeeks. Geraadpleegd op 09-05-2025, van <https://www.geeksforgeeks.org/data-mining-cluster-analysis/>
- Bron 2: Wikipedia contributors. (2025, 30 april). *Cluster analysis*. Wikipedia. Geraadpleegd op 09-05-2025, van [https://en.wikipedia.org/wiki/Cluster\\_analysis](https://en.wikipedia.org/wiki/Cluster_analysis)

Extra onderwerpen

Onderwerp1 (Associatie):

- Bron 1: *Definitie en uitleg datamining*. (2018, 31 oktober). /. <https://www.kaspersky.nl/resource-center/definitions/data-mining>
- Bron 2: *Datamining associatieregels associatieregels de gouden sleutels van datamining* - FasterCapital. (z.d.). FasterCapital. <https://fastercapital.com/nl/inhoud/Datamining--associatieregels--associatieregels--de-gouden-sleutels-van-datamining.html#Inleiding-tot-associatieregels-bij-datamining>
- Bron 3: Guglani, G. (2023, 5 augustus). *Beer & Diaper-Retail Analytics* - Nerd for Tech - Medium. <https://medium.com/nerd-for-tech/beer-diaper-retail-analytics-e5e7b860dba0>

Onderwerp2 (Sequential Pattern)

- Bron 1: Mughal, J. (2021). *Data Mining: Web Data Mining Techniques, Tools and Algorithms: An Overview*. Geraadpleegd op 10/05/2025 van [https://www.academia.edu/download/72044463/Paper\\_30-Data\\_Mining\\_Web\\_Data\\_Mining\\_Techniques.pdf](https://www.academia.edu/download/72044463/Paper_30-Data_Mining_Web_Data_Mining_Techniques.pdf)
- Bron 2: Olson, D. L., & Delen, D. (2008). *Advanced data mining techniques*. Springer Science & Business Media. Geraadpleegd op: 10/05/2025

### 3. Methode

Beschrijf de onderzoeksmethode bij dit vooronderzoek. Bij deskresearch (waaronder literatuuronderzoek valt): welke databanken heb je gebruikt en welke zoektermen? Wat waren de resultaten en hoe heb je daar een verdere selectie uit gemaakt? (max. 200 woorden)

1. Voor dit vooronderzoek (theoretisch kader 1+ 2a, classificatie) is deskresearch uitgevoerd via online zoekopdrachten in Google. Hierbij zijn zoektermen gebruikt zoals "wat is data mining". Op basis van relevantie en duidelijkheid zijn enkele betrouwbare websites geselecteerd, zoals Qlik en IBM. Ik heb ook gelukkig voorkennis.
2. Voor dit vooronderzoek (theoretisch kader 2a, regressie en clustering) is deskresearch uitgevoerd via online zoekopdrachten in Google. Hierbij zijn zoektermen gebruikt zoals "wat is regressie", "regressie data mining", "wat is clustering" en "clustering data mining". Door deze kernwoorden samen te combineren waren er tal van betrouwbare websites te vinden, in dit geval dus bijvoorbeeld GeeksforGeeks of Wikipedia.
3. Voor dit vooronderzoek (Extra onderwerp, associatie) is deskresearch uitgevoerd via online zoekopdrachten in Google. Hierbij zijn zoektermen gebruikt zoals "wat is associatie" en "association rule beer and diapers ". Enkele betrouwbare websites, zoals Kaspersky, FasterCapital en Medium, zijn geselecteerd op basis van hun relevantie en duidelijkheid.
4. Voor dit vooronderzoek (Extra onderwerp, sequential pattern) is deskresearch uitgevoerd via gerichte zoekopdrachten op Google Scholar. De gebruikte zoekterm was "data mining techniques". Er onder andere gekozen voor recente bronnen en voor boeken die uitgebreid ingaan op dit onderwerp.