

# Examen Machinaal Leren OPLOSSINGEN

Eerste zittijd  
19 januari 2022

## Deel B

Voornaam: \_\_\_\_\_


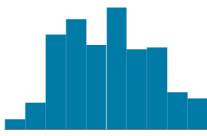
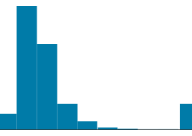
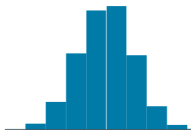

Familienaam: \_\_\_\_\_

**DRAAI DEZE PAGINA PAS OM NA HET SEIN VAN DE BEGELEIDERS**

- Dit examen is gesloten boek. Er zijn geen hulpmiddelen toegelaten.
- Haal deze bundel niet uit elkaar.
- Noteer je antwoorden in deze bundel. Bij sommige vragen moet je een uitdrukking of waarde invullen in de daartoe voorziene ruimte én je antwoord motiveren. Vergeet deze motivatie niet. Zonder motivatie krijg je geen punten, zelfs al is je antwoord correct.
- Schrijf leesbaar.
- Achteraan deze bundel vind je een aantal lege bladen. Duid op elk blad goed aan of het over een kladblad of antwoordblad gaat.

## 1. Data preprocessing [7 punten]

Onderstaande tabel vat een dataset voor “predictive maintenance” samen. Het doel is om aan de hand van machine learning te voorspellen wanneer een elektromotor zal falen aan de hand van observaties zoals de temperatuur, de snelheid en het koppel (torque). We beschouwen dit probleem als een classificatieprobleem.

Machine type	Temperature	Speed	Torque	Failed
				
A: 5% B: 80% Other: 15%	min: 10 max: 150 mean: 60	min: 0 max: 2000 mean: 100	min: -100 max: 100 mean: 10	No failure: 95% Failure: 5%

- (a) Geef 3 mogelijke problemen die je zou verwachten bij deze dataset. Verklaar kort waarom dit een probleem is en wat je zou doen om dit op te lossen.

**Antwoord:**

**Oplossing:** Enkele mogelijke problemen zijn (er werden er slechts 3 gevraagd):

- (a) Ongebalanceerde dataset:

- **Probleem:** 95% van de datapunten behoort tot één klasse. Dit kan problemen veroorzaken tijdens het trainen omdat het model dan kan terugvallen op een triviale oplossing waarbij het steeds de klasse met het meeste datapunten zal voorspellen.
- **Oplossing:** Hoger gewicht toekennen aan de datapunten uit de kleinste klasse. Zodat beide klassen voor ongeveer even veel meetellen in de loss functie.
- **Oplossing:** Subsampling van de grootste klasse of over sampling van de kleinste klasse.
- **Oplossing:** Data augmentatie om zo extra trainingsdata voor de kleinste klasse te genereren.

- (b) Sterk verschillende feature range:

- **Probleem:** Bepaalde features hebben een grotere range dan andere features. Afhankelijk van het type metriek (bv bij Euclidean distance) zal het resultaat voornamelijk bepaald worden door de feature met de grootste range.
- **Oplossing:** Normalisatie of standaardisatie om de feature waarden naar dezelfde range terug te brengen.

- (c) Categorische variabele:

- **Probleem:** De variabele “Machine Type” heeft geen numerieke waarde maar een categorische, deze kunnen niet rechtstreeks als input voor het model gebruikt worden omdat het model de numerieke waarde zal gebruiken.
- **Oplossing:** One-hot encoding.

- (d) Outliers:

- **Probleem:** De variabele “Speed” heeft een aantal outliers met heel hoge waarden. Dit kan het moeilijker maken voor het model om een goede oplossing te vinden.
- **Oplossing:** Analyse van het probleem om te achterhalen of deze waarden nuttig zijn. Indien nodig outlier removal toepassen of de waarde herschalen (bv door een logaritme te berekenen).

- (b) Welke metriek zou je gebruiken om je model te evalueren ? Verantwoord kort je keuze.

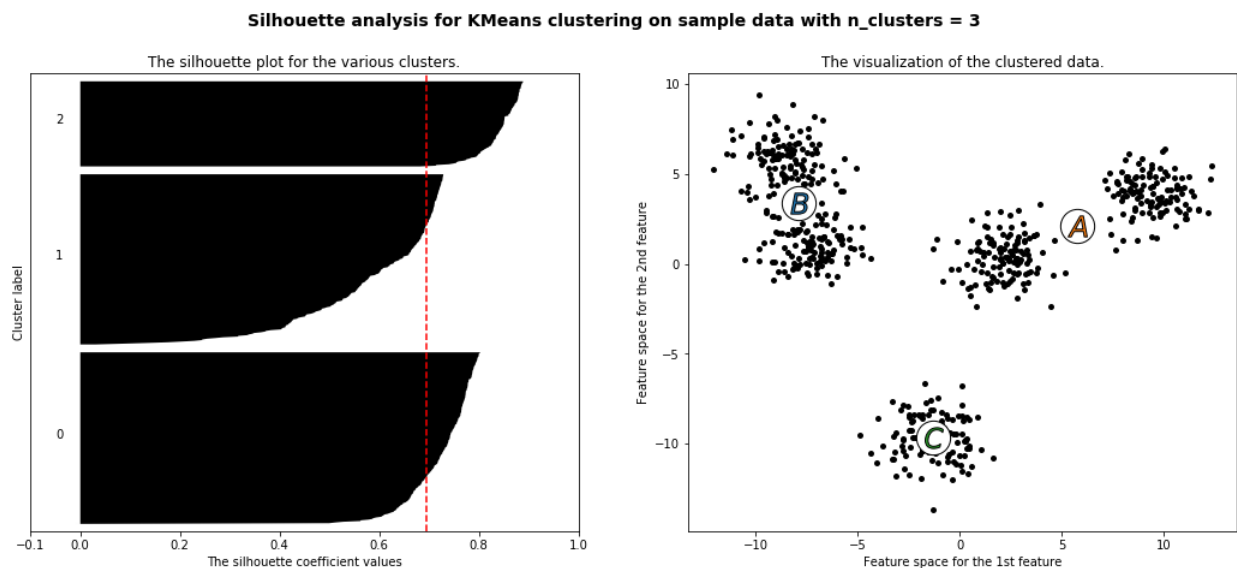
**Antwoord:**

**Oplossing:** Omwille van de onbalans in de data is accuracy niet de meest geschikte metriek. Als het model altijd “no failure” zou voorspellen zou het een accuracy van 95% halen. Een goede oplossing zou zijn om een ROC curve te plotten die het gedrag van de classifier toont naarmate we de threshold voor binaire classificatie aanpassen. De ROC curve toont de relatie tussen de True positive rate (TPR) en de False positive rate (FPR). De true positive rate is het aantal true positives gedeeld door het werkelijke aantal positives (hoeveel van de failures kunnen we detecteren ?). De False positive rate is het aantal false positives gedeeld door het totaal aantal negatives (hoeveel “no failures” detecteren we als “failure” ? ). Naarmate we de threshold hoger zetten en het model dus zekerder moet zijn vooraleer we een failure voorspellen, zullen beide metrieken dalen. De trade-off tussen beiden definieert dan de ROC curve. De ROC curve kan samengevat worden in één getal: de area under the curve: AUC.

Een andere oplossing is om de precision en recall te berekenen. Precision geeft aan hoeveel van de teruggegeven resultaten werkelijk failures zijn. Recall geeft aan hoeveel van de failures er werkelijk gevonden werden. Deze twee getallen kunnen samengevat worden in één getal aan de hand van de F1-score.

## 2. Clustering met K-means [4 punten]

Onderstaande figuur toont rechts een dataset met twee features. De letters A,B en C duiden de drie gevonden clusters aan na het uitvoeren van het K-means algoritme met  $K = 3$ . In de linker figuur wordt de overeenkomstige silhouette plot getoond.



- (a) Geef voor elk van de cluster labels uit de silhouette plot (0,1 en 2) aan met welke cluster (A, B of C) ze overeenkomen.

**Oplossing:** C = 2 , B = 0, A = 1

- (b) Hoe wordt deze silhouette plot opgebouwd? Wat wordt er juist voorgesteld ?

**Oplossing:**

Na de clustering wordt er voor elk punt een metriek berekend die de afstand van dat punt tot zijn cluster center vergelijkt met de afstand tot andere clusters. Deze waarde ligt tussen -1 en 1. Een waarde van 1 geeft aan dat het punt dicht bij het center ligt en ver van andere clusters. Een waarde van nul geeft aan dat het op de grens ligt. Een sterk negatieve waarde geeft aan dat het punt dicht bij een ander cluster center dan bij de cluster waaraan het is toegekend (dit kan niet voorkomen bij k-means). Eens deze metriek is berekend, wordt er voor elke cluster een cumulatief histogram getoond van de waarden.

### 3. SVM [4 punten]

De (vereenvoudigde) signatuur van de Scikit-learn SVM classifier (SVC) is als volgt:

```
class sklearn.svm.SVC(C=1.0, kernel='poly ')
```

- (a) Waarvoor dient de C parameter ? Hoe kies je een waarde voor deze parameter ?

**Oplossing:** Bij een soft-margin SVM bepaalt de C parameter de trade-off tussen twee termen in de loss functie: een term die de marge zo groot mogelijk probeert te nemen en een term die de (zachte) classificatiefout zo klein mogelijk maakt.

Hoe groter de waarde van C, hoe kleiner de margin omdat het model sterker afgestraft wordt voor classificatie fouten.

C is een hyperparameter die kan bepaald worden aan de hand van een grid-search/ random search, eventueel in combinatie met cross-validatie.

- (b) Waarvoor dient de “kernel” parameter ? Geef een andere mogelijke waarde voor deze parameter.

**Oplossing:** De kernel zorgt ervoor dat de SVM een niet-lineaire classificatie kan maken. Data wordt getransformeerd naar een ruimte met hogere dimensie door de kernelfunctie, en er wordt een hyperplane (lineaire classifier) in die ruimte gezocht.

De kernel trick laat ons toe om de classificatie in de hoger dimensionale ruimte te doen zonder de features expliciet te transformeren. Enkele andere opties voor deze parameter zijn “linear”, “RBF” of “sigmoid”.

#### 4. Gaussian Mixture models [5 punten]

Wat is een Gaussian Mixture model ? Zorg dat je in je uitleg de termen “Gaussian” en “Mixture” duidelijk verklaart. Geef ook aan waarvoor ze kunnen gebruikt worden en wat hun parameters zijn.

**Oplossing:** De volgende elementen moesten aanwezig zijn in het antwoord, eventueel via een figuur.

- GMM is een distributiefamilie, bestaande uit de gewogen som (=mixture) van unimodale Gaussiaanse distributies (normaalverdelingen).
- Parameters: gemiddelde (vector) en covariantiematrix van de individuele normaalverdelingen + mixture coëfficiënten die het relatieve gewicht van elke unimodal distributie aangeven.
- Gebruikt voor generatieve modellen, bv. om nieuwe samples uit een geleerde distributie te maken; voor clustering (toewijzen van een nieuwe element aan een van de distributies); voor anomaliedetectie (detecteren van samples met zeer lage probabilliteit volgens de geleerde distributie).