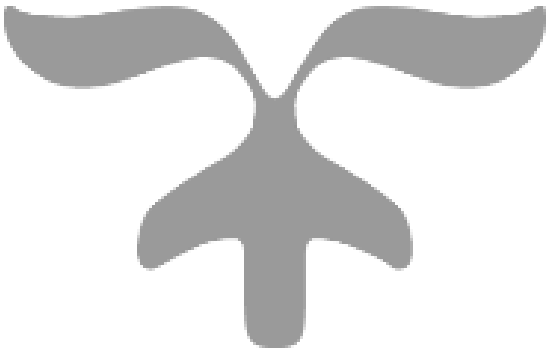
Justification Document

Jinne Haan en Daan Derks



Contents

List of Tables

List of Figures

0.1 Excersize 1

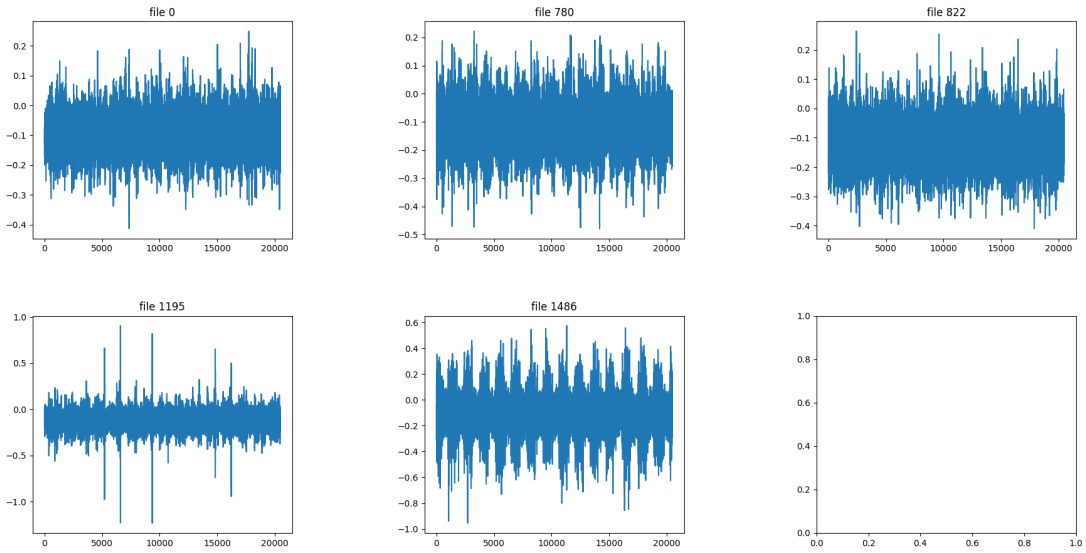


Figure 1: Figuur x in verschillende stadiums

1.

2. **Toegevoegde features**

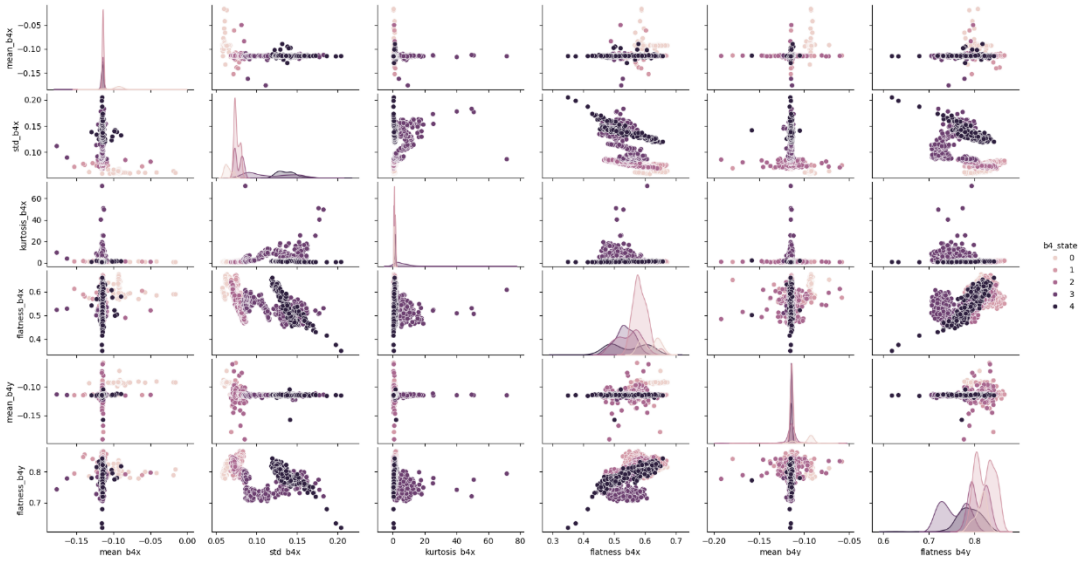
   Spectral Flatness: Deze maat vergelijkt het geometrisch gemiddelde met het rekenkundig gemiddelde van het vermogensspectrum van een signaal. Een hoge spectral flatness duidt op een vlak spectrum, wat geassocieerd wordt met storingen of schade aan lagers, doordat defecten vaak leiden tot bredere frequentiecomponenten.

   Kurtosis: Kurtosis meet de spitsheid van de verdeling van het signaal. Signalen met defecten bevatten vaak plotselinge pieken, wat resulteert in een hogere kurtosis. Dit kan een sterk onderscheidend kenmerk zijn tussen gezonde en beschadigde lagers.

   Spectral Flatness: Peeters, B., & De Roeck, G. (2001). Stochastic system identification for operational modal analysis: A review. [IEEE]

   Kurtosis in fault detection: Randall, R. B. (2011). Vibration-based Condition Monitoring. Wiley.

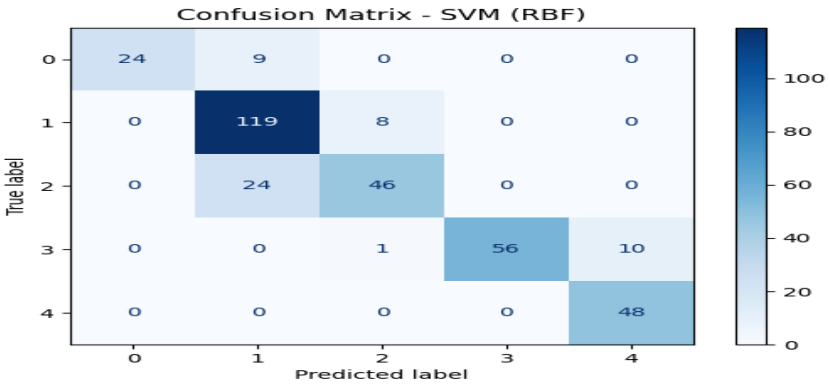
1. Graphics Type 'SHAPE' is not supported yet. Please insert it as image.



We hebben Random Forest en SVM (RBF) gekozen omdat we een sns.pairplot hebben gemaakt waarin te zien is dat samples met stage 1 visueel goed te scheiden zijn van de andere stadia, terwijl stages 2, 3 en 4 veel overlap vertonen. Dit patroon werd bevestigd door een aanvullende PCA-analyse: de eerste twee hoofdcomponenten scheiden stage 1 duidelijk van de rest, maar tonen sterke overlap tussen de andere drie stadia.

Op basis van deze observaties verwachten we dat classificatie van stage 1 versus de andere stadia relatief eenvoudig is, maar dat het onderscheiden van alle vier de degradatiestadia meer complexiteit vereist. Daarom kiezen we voor modellen die goed kunnen omgaan met niet-lineaire scheidingen en complexe interacties tussen features.

Random Forest is geschikt vanwege zijn robuustheid tegen overfitting, het omgaan met overlappende klassen en het leveren van feature importance. SVM met RBF-kernel is gekozen vanwege zijn vermogen om niet-lineaire beslissingsgrenzen te modelleren, wat waarschijnlijk nodig is gezien de overlappende clusters.



1. Graphics

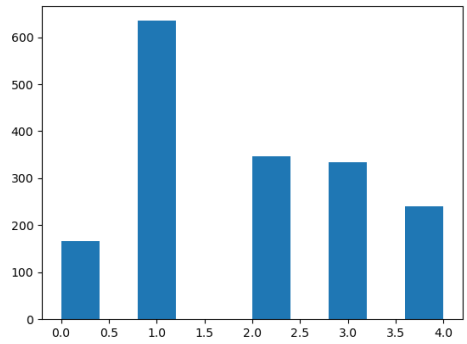
   Type ‘SHAPE’ is not supported yet. Please insert it as image.

   De standaarddeviatie en de RMS (root mean square) zijn sterk gecorreleerd omdat ze beide de spreiding van de data meten. Het gemiddelde is redelijk constant en dat is het enige verschil in de twee formules.

2. Random forest is betere op alle 3 de evaluaties dus kiezen we dit model

1. We hebben de cutoff van de multicolineariteit op absolute waarde 0.8 gezet om minder multicolineariteit te krijgen in het model.

2. Graphics Type ‘SHAPE’ is not supported yet. Please insert it as image.



1. Accuracy is niet een goed model criterea omdat de verdeling van de groepen niet gelijk is, er zit veel meer in groep 1 dan in de andere groepen. We hebben meer gelet op f1 score.

2. We hebben train set en een test set gebruikt om overfitting te voorkomen.

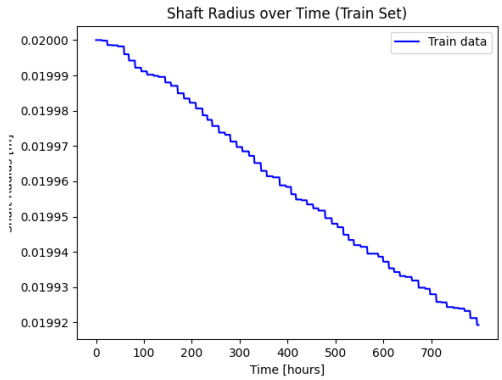
3. Als je voor bearing 4 wilt classificeren dan wel, maar wanneer je andere bearings wilt classificeren dan niet want we kunnen er niks over zeggen.

4. Een nog uitgebreidere grid-search met meer parameter opties

5. In conclusie is het een goed model om bearing 4 te classificeren maar in de praktijk heb je waarschijnlijk meer dan bearing 4 nodig.

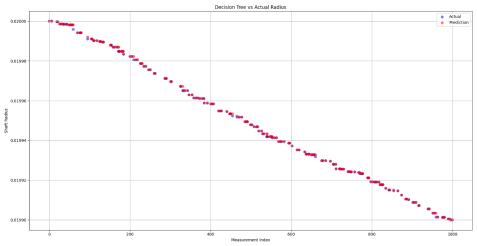
0.2 Excersize 2

1. Graphics Type ‘SHAPE’ is not supported yet. Please insert it as image.



We hebben gekozen voor decision tree regression model omdat de radius in stapjes omlaag gaan en decision tree is hier precies voor bedoeld

Graphics Type ‘SHAPE’ is not supported yet. Please insert it as image.



1. We hebben gekozen voor R2 en MSE. MSE vertelde ons weinig omdat de fouten heel klein waren maar relatief houd MSE geen rekening mee. Dus vooral gelet op R2
2. Dit model is reliable omdat de R2 score erg hoog is.
3. In de grid-search op meerdere parameters letten.
4. We hebben een model gemaakt dat uitstekend werkt. Dat is te zien aan een R2 score van 0.9998 !

0.3 Excersize 3

1. We hebben gekozen voor mean, std, max_std, rms, kurtosis, flatness, crest.

   We hebben dezelfde features van excersize 1 omdat ze daar al nuttig bleken en daarnaast hebben we twee nieuwe features crest en max_std toegevoegd. Crest is de verhouding tussen de piek-waarde en de rms waarde van een signaal deze waarde geeft inzicht in hoe 'impulsief' een trillingsignaal is. Max_std is de maximale standaard deviatie over de tijd dit leek ons een goed idee omdat we de std over de tijd hebben geplot en we zagen dat std in sprongen omhoog ging.

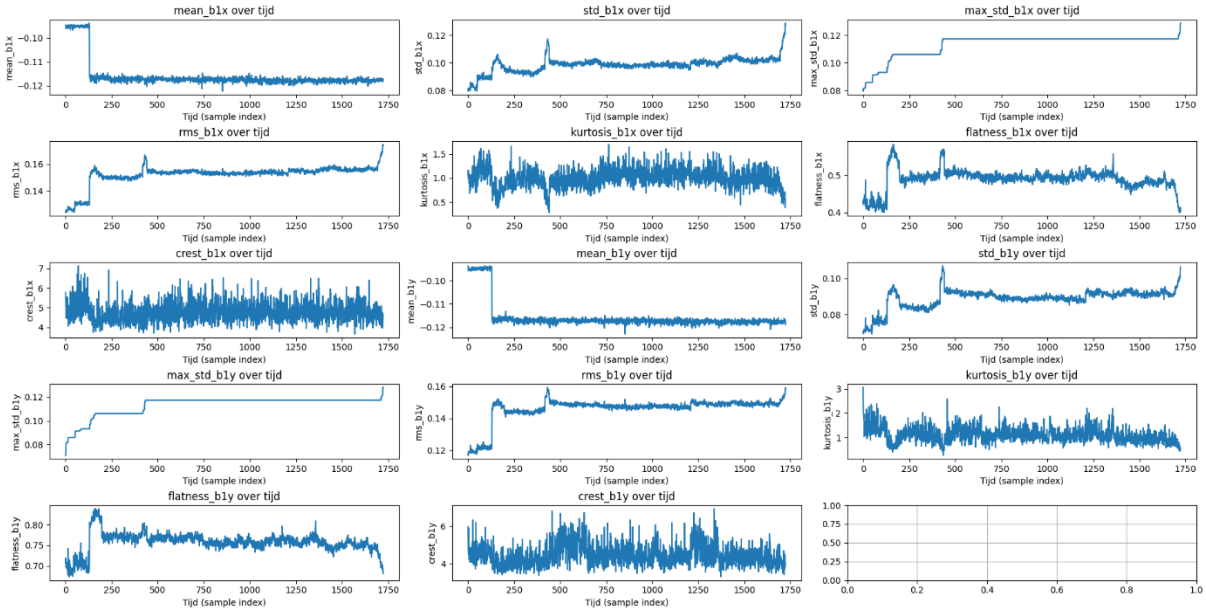


Figure 2: Figuur sns pairplot van de gekozen features

1. We hebben gekozen voor K-means clustering en gaussian Mixture omdat we wisten dat er 5 groepen waren. Een nadeel aan K-means clustering is dat deze slecht omgaat met groepen van verschillende grootte, groep 1 is een stuk groter dan de rest bij bearing 4 dus dit zou ook kunnen bij bearing 1. K-means en gaussian mixture werkt allebei goed op niet-lineare data. We hebben PCA dimensionality reduction geprobeerd maar dit maakte de resultaten slechter. We hebben de

resultaten van verschillende methodes gecritiseerd op een plot van index waarbij de verschillende groepen unieke kleuren hebben, bij een perfecte clustering zou een groep één aaneengesloten kleur vormen zonder overlap met andere kleuren op de index-as. Hoe meer het door de kleuren in de war zijn hoe slechter de clustering is gelukt.

1. We hebben van de groep de gemiddelde index gebruikt, vervolgens is de volgorde van 0 tot 4 met het laagste gemiddelde als stadium 0 en hoogste gemiddelde als stadium 4.

0.4 AI gebruik

We hebben voor verschillende onderdelen AI gebruikt.
Syntax: Het programmeren doen we zelf, we bedenken wat we willen doen en hoe de flow ongeveer zal gaan. Wanneer we bezig zijn met programmeren en de syntax lukt niet helemaal, dan vragen we ai hoe dat moet en om voorbeelden. Een voorbeeld: we hebben gevraagd om voor te doen hoe je vanuit de fetch_ucirepo datasets kunt inladen in python en transformeren in een Pandas dataframe. We wisten dat dit kon, maar niet precies hoe, want het is de eerste keer dat we met die package werken.
Test-scores: nadat we hebben gekeken naar welke test-score we gaan gebruiken hebben we nog even aan ai nagevraagd of dat die het eens is met de keuze of, of dat hij toch een andere test-score zou aanraden. Hieruit bleek dat we goed bezig waren.
Errors: soms krijgen we errors en is het moeilijk te begrijpen waar het mis gaat. Dan vragen we ai om inzicht te bieden in wat er mis gaan en een mogelijk oplossing suggestie te krijgen.

0.5 Wie heeft wat gedaan?

We hebben samen aan alle code gewerkt. Dit hebben we gedaan door naast elkaar te zitten en te programmeren, meestal hebben we dan allebei kleine taken die de hele tijd samen vallen. We hebben naast elkaar samengewerkt via de 'live share' extensie van visual studio code gemaakt door Microsoft. Op deze manier kunnen we een beetje als een online word document tegelijk in een document typen.
Alle keuzes die gemaakt zijn tijdens deze opdracht zijn dan ook samen genomen en overlegd.
Het justification document is gemaakt door:
Keuze van de datasets: Daan
Gemaakte keuzes: beide
AI gebruik: Jinne
Wie heeft wat gedaan?: beide

Sources:
OpenAI. (2023). *ChatGPT* (Mar 14 version) [Large language model]. `https://chat.openai.com/chat`
Course modules: FEB25: - Data Science 8. (n.d.). `https://canvas.fontys.nl/courses/26895/modules`