**Rapport DDB**

**V2A, Groep 6**

| **Alexander van Schaick Zillesen** | **1789777** |
| --- | --- |
| **Wytze Ketel** | **1797080** |
| **Timo Hak** | **1771225** |



**Inhoudsopgave**

[**Inleiding**](#_b5ztcsq33k3q) **3**

[Probleem:](#_x5athhb0emm7) 3

[Doelstelling:](#_olbqyyh778fb) 3

[**Samenvatting**](#_u3urnmi2k293) **4**

[**Business Understanding**](#_txtvdzs3g7sa) **5**

[Stakeholders/doelstelling:](#_e8mhr8mra4hk) 5

[Huidige situatie:](#_4xsaouefrod5) 5

[Gewenste situatie:](#_hvwziyf4l11s) 6

[Target Variabele:](#_qub2wfozwquq) 6

[**Data Understanding**](#_xixelyazlly1) **7**

[Selectie:](#_bx73r6c2lfis) 7

[Correlatie:](#_yvmx3g1e09k4) 7

[**Data preparation**](#_cyk2ko9msf0q) **8**

[Data opschonen:](#_rja1qt1vwyy6) 8

[**Modeling**](#_by3rqhpur6hc) **9**

[Technieken:](#_ho4y75pfoimt) 9

[Modellen:](#_lyb7yiv2z4z5) 9

[Decision tree:](#_psvmw5knpqgc) 9

[KNN:](#_uyb1coppzsss) 9

[Regressie:](#_70228krwds9p) 9

[Random Forest:](#_vpoy6ca70eiv) 9

[**Evaluation**](#_pu56qo205xwf) **10**

[Criteria:](#_1f9anr7zd4yo) 10

[Review:](#_xok9zh7xeu53) 10

[**Deployment**](#_9mqgn20riq) **11**

[Handleiding:](#_blde6c6a7o0k) 11

[Applicatie:](#_6h8yqc6y84ff) 11

[**Conclusie**](#_pqib5sbjnp62) **12**

[**Bijlagen**](#_v844naj32rdg) **13**

# Inleiding

#### 

ProRail is verantwoordelijk voor het spoorwegnet van Nederland. Samen met de vervoerders zoals de NS doen zij hun best om reizigers en goederen veilig en op tijd naar hun bestemming te brengen.

#### Probleem:

Wanneer er een probleem verschijnt die het treinverkeer hindert, krijgt de NS een initiële indicatie van hoe lang de verwachting is voor hoe lang het oplossen van het probleem gaat duren. Hiermee kunnen zij de reizigers inlichten en rekening houden met eventuele voorbereidingen. Maar ze beginnen pas echt met de voorbereidingen wanneer de monteur ter plekke is en een verwachte reparatietijd heeft gegeven. Het probleem is echter dat de reparaties vaak eerder klaar zijn dan de voorbereidingen van de planners, en hierdoor ligt het treinverkeer langer stil dan nodig.

#### Doelstelling:

In de praktijk worden voorspellingen van monteurs vaak te hoog ingeschat, waardoor het treinverkeer langer stilstaat dan nodig. De doelstelling van het project is daarom ook om een betere voorspelling te maken waardoor het treinverkeer sneller hervat kan worden. Ook is het de bedoeling dat de voorspellingen in een dashboard getoond kan worden met aanvullende informatie zoals betrouwbaarheid erbij.

# Samenvatting

# 

We hebben simpelweg de stappen gevolgd uit het crisp-dm model en hebben ook op de eerste dag van elke sprint de planning gemaakt voor die week.

Te beginnen met de Business understanding, we hebben goed gekeken naar hoe de huidige situatie is, en wat de stakeholders zijn. Daarna hebben we gekeken hoe we de huidige situatie kunnen verbeteren, zodat de belangrijke stakeholders daarvan profiteren. Ook hebben we aan de hand van de eisen van de opdrachtgever, de target variabele bepaald.

Bij de data understanding zijn we dieper in de data gedoken om het beter te kunnen begrijpen. Met de kennis die we daarbij opgedaan hebben, hebben wij uiteindelijk een selectie gemaakt van de data waar wij mee wilde gaan werken. Daarnaast hebben wij nog verder onderzoek gedaan naar bijvoorbeeld de correlaties.

De data was heel erg rommelig, en we hebben uit de selectie van de data die wij gebruiken vrijwel alles schoongemaakt, al hebben we bij sommige velden hele specifieke methodes moeten gebruiken om de data bruikbaar te maken.

Wij hebben verschillende technieken geprobeerd om tot een goed model te komen, en hebben uiteindelijk 2 technieken over gehouden. Dit zijn de knn en polyreg. Deze 2 vonden wij uiteindelijk het beste om mee te werken.

Voor de evaluatie hebben wij nog even goed naar de criteria gekeken en hoe het eindproduct er nou precies uit moet komen te zien. En hebben wij hebben een validatie van de oplossing aan de hand van een validatieset met relevante visualisaties erbij.

# 

# 

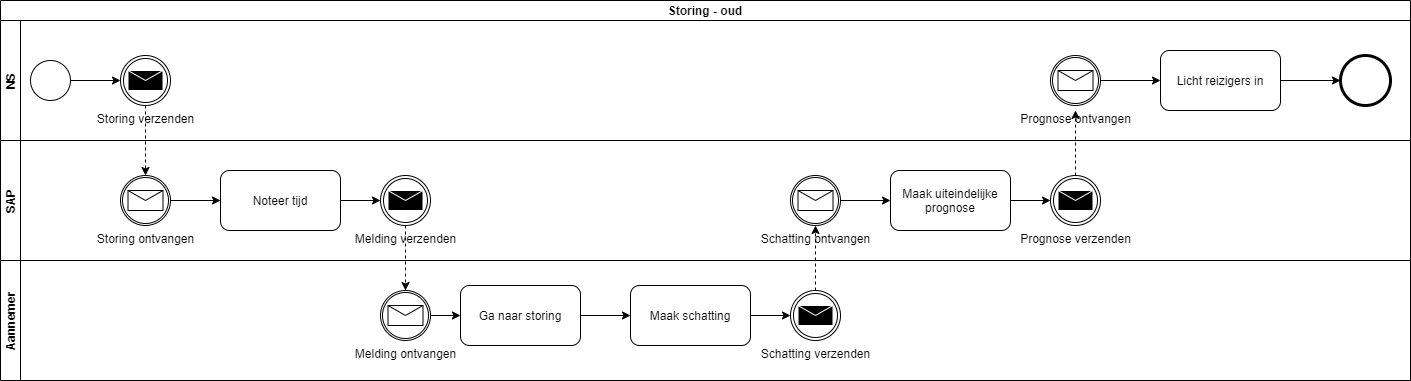
# 

# Business Understanding

#### Stakeholders/doelstelling:

* Planners NS
* Prorail
* De doelstelling is om een betere voorspelling te geven dan de monteurs, zodat de planners op tijd met hun voorbereidingen kunnen beginnen waardoor het treinverkeer minder lang stil staat/gehinderd wordt.

#### Huidige situatie:



De belangrijkste stakeholder zijn de planners van de NS, zij zijn afhankelijk van de informatie van prorail en monteurs die de problemen op lossen. Monteurs schatten vaak niet goed in hoe lang het gaat duren om een probleem op te lossen, en dit leidt tot het onnodig lang stilstaan/hinderen van het treinverkeer. De planners bij de NS zijn vaak nog niet klaar met hun voorbereidingen terwijl de monteurs het probleem al opgelost hebben. Voor de planners is het belangrijk om te weten wanneer een probleem opgelost zal zijn, zodat zij op tijd kunnen beginnen om alle treinen/personeel weer op de juiste plek te krijgen.

#### 

#### Gewenste situatie:

De gewenste situatie zal zijn dat de planners een goede schatting krijgen van de monteurs en de benodigde informatie van prorail zodat zij op tijd met de voorbereidingen kunnen beginnen. Hierdoor zal er weinig tijd zitten tussen het oplossen van het probleem, en het moment dat alle treinen/personeel zich weer op de juiste plek bevinden. Dit zorgt voor minder vertragingen, dus minder overlast voor de reizigers.

###### **Target Variabele:**

De target variabele die wij gebruiken is ‘stm\_fh\_duur’, dit is de totale duur van de functieherstel in minuten.

# Data Understanding

#### Selectie:

We hadden eerst een selectie gemaakt aan de hand van de omschrijvingen van de data. We kozen natuurlijk de data waarvan wij dachten dat het enige invloed zou kunnen hebben op ons model. Hiermee gingen we verder onderzoek doen.

Selectie:

'stm\_oorz\_groep'

'stm\_oorz\_code'

'stm\_geo\_gst'

'stm\_fh\_ddt'

'stm\_sap\_storeind\_ddt'

'stm\_sap\_melddatum'

'stm\_sap\_meldtijd'

'stm\_aanntpl\_tijd'

'stm\_fh\_dd'

'stm\_fh\_tijd'

'stm\_fh\_duur'

'stm\_sap\_storeinddatum'

'stm\_sap\_storeindtijd'

'stm\_prioriteit'

‘stm\_progfh\_in\_duur’

Met deze selectie keken we eerst naar de numerieke variabelen, zodat we kunnen zien welke variabelen nog omgezet moesten worden tot een voor een model bruikbaar getal, we kwamen erachter dat er nog een aantal variabelen waren die verwerkt moesten worden.

Daarna keken we naar het aantal lege rijen in de database, en kwamen tot de conclusie dat er, percentueel gezien, weinig waardes ontbraken en ze dus zonder al te veel invloed konden laten vallen.

Tot slot keken we of alle data types klopte en zagen dat er een aantal nog moesten worden omgezet naar numerieke waardes.

#### Correlatie:

In het begin hadden wij gevonden dat er vrijwel geen correlatie was tussen de verschillende tabellen, hiermee konden wij dus weinig doen. Na het opschonen van de data was er wel een grote sprong in correlatie tussen onze target variabele en ‘stm\_prioriteit’, hier hebben we natuurlijk geprobeerd om mee te werken.

# Data preparation

#### Data opschonen:

De data was heel erg rommelig, en we hebben dus goed ons best moeten doen om het op te schonen. We hebben vrijwel alles opgeschoont, dit hebben wij onder andere gedaan om hopelijk een goede correlatie te vinden na het opschonen van de data.

Voor alle data hebben wij het datatype aangepast waar nodig, ook hebben wij de outliers verwijderd, maar voor een aantal tabellen hebben wij specifieke aanpassingen moeten doen:

* Lege velden in ‘stm\_prioriteit’ gevuld met 9 (de laagste prioriteit)
* Bij ‘stm\_ooz\_codes’ hebben wij een nieuwe code 0 aangemaakt voor onbekende oorzaak codes.
* Onze target variabele ‘stm\_fh\_duur’ hebben wij alles verwijderd waar de duur 0 is, en ook de sterke outliers verwijderd.
* Alle data in ‘stm\_geo\_gst’ die niet numeriek is, hebben wij verwijderd.
* Alle tabellen met tijden en data worden omgezet naar pure getallen, zodat ons model daarmee kan werken.

# 

# Modeling

#### Technieken:

Wij hebben uiteindelijk meerdere technieken gebruikt voor het maken van een model, de decision tree, knn, lineaire regressie, polynomial en random forest.

#### Modellen:

We hebben verschillende modellen geprobeerd, met name om te kijken welk model accurater is, maar ook om te kijken of andere modellen niet over fit zijn.

###### **Decision tree:**

We komen uit op de tabellen ‘stm\_sap\_meldtijd’ en de ‘stm\_aanntpl\_tijd’ in dit model, na verschillende combinaties van tabellen te hebben geprobeerd.

Volgens het model heeft het een accuracy van 30%

###### **KNN:**

We komen uit op de tabellen ‘stm\_sap\_meldtijd’ en de ‘stm\_aanntpl\_tijd’ in dit model, na verschillende combinaties van tabellen te hebben geprobeerd.

Volgens het model heeft het een accuracy van 30%

Het KNN model wordt by default gebruikt om een voorspelling te maken, omdat we dan eenvoudiger een accuratie kunnen geven van de voorspelling zelf.

###### **Regressie:**

We gebruiken de tabellen ‘prioriteitscode’,‘stm\_sap\_meldtijd’, ‘stm\_aanntpl\_tijd’,’monteur voorspeltijd’. Het is een iets complexer model dan bij KNN of Decision tree, maar claimd ook een hogere accuracy te hebben van ongeveer 40%

###### **Random Forest:**

Maakt ook gebruik van weer ‘stm\_sap\_meldtijd’ en de ‘stm\_aanntpl\_tijd’, het model is echter langzaam in het gebruik en claimd slechts een 31% zekerheid, we hebben het dus maar achterwege gelaten.

# 

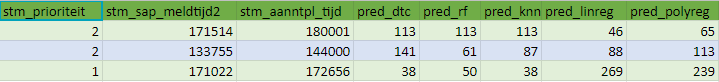
# Evaluation

#### Criteria:

* Voorspellingen moeten getoond worden in een dashboard
* Relevante informatie over de voorspelling, zoals de betrouwbaarheid, moet getoond worden

#### Review:

Validatie aan de hand van een validatieset:



# Deployment

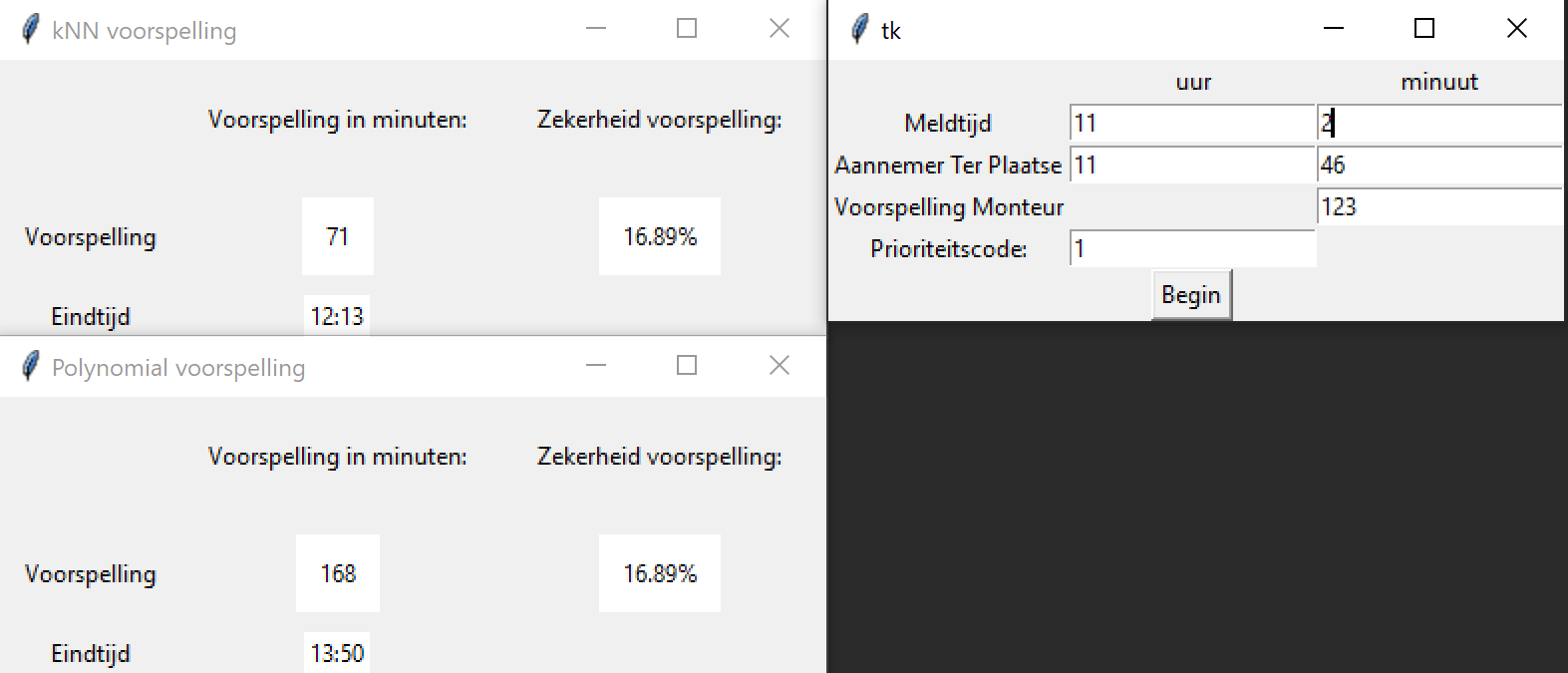
#### Handleiding:

We hebben het programma zo simpel mogelijk gehouden, er zijn maar een paar stappen te volgen om het programma goed te gebruiken:

1. Vul de tijd in wanneer het probleem is gemeld, er zijn losse input velden voor uren en minuten.
2. Vul de tijd in wanneer de Aannemer ter plaatse is, hier zijn ook losse input velden voor uren en minuten.
3. Vul de voorspelling van de monteur in (in minuten).
4. Vul de prioriteitscode in.
5. Druk op ‘Begin’ en het programma zal heel snel met een antwoord komen.

#### Applicatie:

De applicatie is vrij simpel, met een aantal input velden en een knop om de berekening te starten:

****

# Conclusie

We hebben veel verschillende dingen geprobeerd om tot een goed model te komen, zo hebben we bijvoorbeeld veel verschillende combinaties van tabellen getest voor het model.

Na het testen van verschillende combinaties in het model zijn we tot de conclusie gekomen dat knn en polyreg modellen het beste zouden werken voor ons.

Het dashboard is heel simpel te gebruiken, en is grotendeels foolproof gemaakt, ook wordt er weergegeven hoe zeker het model is dat hij de voorspelling goed heeft.

# Bijlagen

<https://github.com/Alexandervsz/Storingen>