**Startup Dataproject voorspellen van verzuim**

**Introductie**

In het kader van de data & analytics Python cursus die ik volg bij Novalinq ga ik een portfolio opdracht maken om de cursus mee af te sluiten. Vanuit Robidus heb ik vernomen dat eer een wens is om door middel van data analyse meer inzicht te krijgen in de factoren die invloed hebben op het verzuim van een werknemer en de mogelijkheid om op basis van data te kunnen voorspellen wat de kans is dat iemand langdurig verzuimt

**Bussinessvragen:**

1. Welke factoren hebben invloed op de kans dat iemand langdurig verzuimt?
2. Kunnen we voorspellen hoeveel verzuim iemand krijgt in het komende jaar (in uren en percentage van totale verloonde uren)?
3. Kunnen we voorspellen welke verzuimstatus iemand krijgt in het komende jaar (statussen: arbeidsgeschikt, gedeeltelijk arbeidsongeschikt, niet duurzaam arbeidsongeschikt, duurzaam arbeidsongeschikt)?

**Stakeholders**

* Wia (Wet werk en inkomen naar arbeidsvermogen) casemanagers en consultants die hierdoor beter advies kunnen geven aan klanten.
* Klanten van Robidus die hierdoor preventief in kunnen spelen op risico’s binnen hun huidige en toekomstige werknemersbestand.
* Werknemers omdat zij vroegtijdig geholpen kunnen worden wanneer er een verhoogd risico is op verzuim en zo minder lang thuis hoeven te zitten (indirect voordeel).

**Data**

Data vanuit het datawarehouse van Robidus (eventueel geanonimiseerd) wanneer mogelijk.

Beschikbare data die ik wil gebruiken:

* Individuele factoren
  + Aantal gewerkte uren
  + Sekse
  + Inkomen
  + Leeftijd
  + Type arbeidsrelatie (Flexibel, vast, uitzend)
  + Woonplaats (postcode4)
  + Verhuizing (verandering in postcode4)
  + Verzuimgeschiedenis (Status en duur in voorgaande jaar of jaren)
  + Afstand tot werk (bereken door middel van afstand tussen postcode4 van werknemer en adres werkgever)
* Bedrijfsfactoren
  + Grote van werkgever
  + Ziekteverzuim bij werkgever (percentage)
  + Sector van werkgever

Data die ik zou willen gebruiken maar waarvan ik niet weet of die beschikbaar is:

* + Aantal gewerkte overuren
  + (verandering in) Burgerlijke staat
  + Hiërarchie binnen het bedrijf

Anders datasets van Kaggle.com over absentism mocht de data van Robidus niet gebruikt kunnen worden. Hoewel de uitkomst van deze datasets niet gelijk bruikbaar is voor Robidus is dit wel een proof of concept voor het voorspelen van (langdurig) ziekteverzuim en zorgt het voor meer expertise van voorspelmodellen bij mij en (na uitwisseling van de uitkomsten) het team/de organisatie.

**Doelen**

* De juiste data beschikbaar stellen binnen twee weken na de start van het project.
* Het presenteren van de huidige stand van zaken en de oorzaken van langdurig verzuim in de dataset binnen 4 weken na de start van het project.
* Het door predictive modeling voorspellen van verzuim in het komende jaar en dit binnen 8 weken na de start van het project kunnen presenteren aan (de stakeholders) binnen Robidus.