# Enquete BiS DNA NOC interpretatie

Voor mijn thesis ontwikkel ik een methode om de schatting van het aantal donoren aan een DNA sample inzichtelijk te maken. Het huidige machine learning model geeft alleen de voorspelling van de NOC aan, maar niet waarop deze voorspelling wordt gebaseerd. Met bepaalde technieken kunnen de voorspellingen van machine learning modellen worden uitgelegd. Met deze enquete probeer ik inzichtelijk te maken waar experts op letten wanneer het aantal donoren wordt ingeschat, zodat de verklaringen van het model op een vergelijkbare manier kunnen worden gepresenteerd. Zo kan een expert bepalen of de voorspelling van het machine learning model aannemelijk is.

1. Wat is uw workflow, en naar welke informatie kijkt u bij het bepalen van het aantal donoren (NOC) in een profiel?

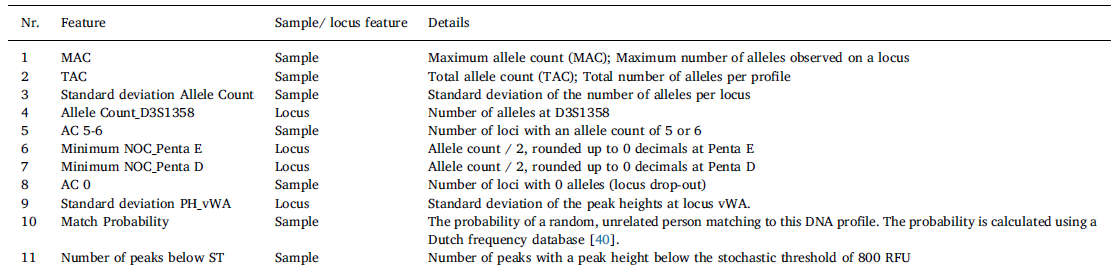
Voorbeeld antwoord*:*

*“Eerst kijk ik naar de MAC. Daarna kijk ik naar het profiel in het algemeen, of er veel pieken zijn, of er veel verschil zit in de piekhoogtes, en de piekhoogte in het algemeen. Meestal is het dan wel duidelijk bij een laag aantal donoren. Als het dan nog niet duidelijk is, bekijk ik nog specifieke loci om daar meer informatie te vinden. Uiteindelijk combineer ik die informatie met de MAC, de TAC, en de informatie over de piekhoogtes om een conclusie te trekken. Ten slotte vergelijk ik mijn antwoord met wat er uit de machine learning NOC tool komt.”*

1. Wat vindt u indicatief van een profiel met:
   1. 1 donor:
   2. 2 donoren:
   3. 3 donoren:
   4. 4 donoren:
   5. 5 donoren:

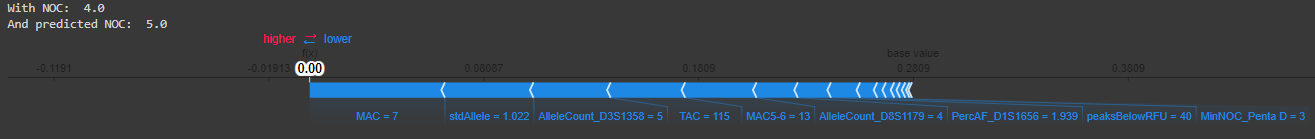
Voorbeeld antwoord:

* 1. *“MAC van 2 in combinatie met een profiel van hoge kwaliteit.*
  2. *…*
  3. *…*
  4. *…*
  5. *Hoge MAC, veel (lagere) pieken, meer noise”*

1. Welke soort uitleg vindt u het meest waardevol wanneer uw verwachting van de NOC niet overeenkomt met de voorspelling van de NOC door het machine learning model? Een combinatie is ook mogelijk. Met features worden bepaalde statistieken van het profiel bedoeld (zie Figuur 1)

Figuur 1: Voorbeeld features die van een profiel kunnen worden berekend.

* 1. Een lijst met features waar het model de beslissing op heeft gebaseerd (zie Figuur 2).



Figuur 2: Verklaring van een voorspelling van het machine learning model. Het model heeft hier 5 donoren voorspelt, terwijl het profiel maar 4 donoren bevat (de onderliggende informatie over het eigenlijke aantal donoren heeft u in een echte situatie natuurlijk niet tot uw beschikking). In deze situatie is het machine learning model dus fout. In de figuur ziet u de features die het sterkst hebben bijgedragen aan de voorspelling van het machine learning model. De sterkste feature is in dit geval dus de MAC van 7, daarna de standaarddeviatie van het aantal allelen per locus, enzovoort.

* 1. Een voorbeeld van welke features er zouden moeten veranderen en met welke waarden zodat het model dit profiel hetzelfde zou interpreteren zoals uw voorspelling (zie onderstaande tekst)

*Het machine learning model had dit profiel ingeschat als een mix van 4 donoren in plaats van 5 als de volgende features werden aangepast naar de volgende waarden:*

* *TAC = 90 en MAC5-6 = 10*
  1. Een voorbeeld-profiel dat erg lijkt op het huidige profiel, maar dat het model hetzelfde zou interpreteren zoals uw voorspelling (zie onderstaande tekst)

*Het machine learning model had dit profiel ingeschat als een mix van 4 donoren in plaats van 5 als het profiel er zo uit had gezien: <voorbeeld profiel met de veranderingen ten opzichte van het originele profiel toegelicht> (bijvoorbeeld hetzelfde profiel, maar met een aantal pieken weggehaald)*

Voorbeeld antwoord:

*a&b*