# 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?

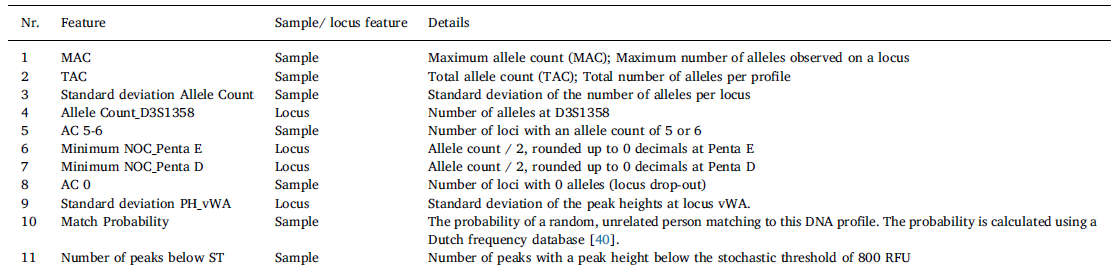
Voorbeeld antwoord*:*

*“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.”*

Voor deze vraag schets ik eerst een scenario; Stel u heeft een profiel bekeken, en op basis van deze informatie bepaalt u dat het profiel informatie bevat van 4 donoren. Echter, het machine learning model heeft voorspelt dat het aantal donoren 5 is. U wilt weten of uw analyse correct is, en u moet kiezen voor een NOC van 4, of dat het machine learning model correct is, en u moet kiezen voor een NOC van 5.

Daarvoor wilt u weten waar het machine learning model de voorspelling op heeft gebaseerd. Stel het model heeft de voorspelling gemaakt op informatie waarvan u weet dat het niet indicatief is voor een NOC van 5, dan kunt u concluderen dat het model een foute voorspelling heeft gemaakt. Het is ook mogelijk dat het model u een verklaring geeft die u wijst op iets wat u over het hoofd heeft gezien. Dan is het mogelijk dat u uw voorspelling bijstelt.

De vorm waarin een verklaring kan worden gegenereerd kan verschillen. Met deze enquête probeer ik inzicht te krijgen in welke vorm(en) het meest natuurlijk zijn voor dit probleem.

Het huidige machine learning model is getraind op een andere versie van de data dan u gewend bent. Van een profiel kunnen namelijk bepaalde statistische features worden gegenereerd. In Figuur 1 zie u een aantal voorbeelden van deze features. U bent bekend met de MAC en de TAC, maar er zijn tientallen verschillende features van een profiel beschikbaar waar u mogelijk iets langer over moet nadenken om deze juist te begrijpen. Het model beslist dus op basis van deze features het aantal donoren van een profiel.

In de volgende twee voorbeelden ziet u mogelijke verklaringen van de voorspelling van het model. **Het model heeft dus gekozen voor een NOC van 5, terwijl u dacht dat het 4 was**.

**A**: In het figuur ziet u welke feature waardes het meeste hebben bijgedragen aan de voorspelling van de NOC van 5 door het model. Er is te zien dat de TAC-waarde van 115 het meeste heeft bijgedragen aan de voorspelling, gevolgd door de standaarddeviatie van het aantal allelen per locus, de MAC van 7, enzovoort.



**B**: Hieronder staan met welke feature waardes het model een NOC van 4 zou voorspellen in plaats van 5. Er is te zien dat met een lagere TAC-waarde (100 in plaats van 115), en een lager aantal loci met 5 of 6 allelen (8 in plaats van 13), het model een NOC van 4 zou hebben voorspeld.

TAC = 100 AC5-6 = 8

1. Welke uitleg vindt u het meest waardevol? Een combinatie is ook mogelijk. Kunt u een verklaring geven?

Voorbeeld antwoord:

*“B:*

*Omdat ik de keuze wil maken tussen 4 en 5, wil ik dat deze twee voorspellingen worden vergeleken, en dat gebeurt bij B.*

*Bij A zie ik waardes die het model voor 5 gebruikt, maar niet in vergelijking met 4. Zo zie ik dus ook waardes die misschien hebben geholpen om een NOC van 1, 2 en 3 weg te strepen, zoals de MAC van 7.*