**Documentatie specialisatie opdracht**

Ahmed dhiblawe, Mohammed Aarab en Lucas Bos

**Requirements**

Jupyter notebook, python (3), de gps en xml bestanden (zonder ‘flag’ tags, deze staan in github) van NEC

**Data munging**

Voorbewerking van data:

In de xml bestanden die we van school hebben gekregen staat altijd wanneer de tekst bespreking is een tag ‘flag’ maar deze wordt niet gesloten. Omdat dit problemen opleverde met het uitlezen van het xml bestand hebben we overal deze ‘flag’ tags weggehaald.

Omdat we niks aan de rijen hebben zonder GPStijd hebben wij deze verwijderd. (Zie 1 in jupyter notebook)

Aan de gps-data hebben we nog een kolom toegevoegd met de timestamp van de gps-tijd zodat we makkelijker tijden kunnen vergelijken met elkaar. (Zie 2)

We hebben het xml bestand van alle wedstrijden uitgelezen. We hebben nog een kolom met de wedstrijd eraan toegevoegd zodat het duidelijk is om welke wedstrijd het gaat. De wedstrijd hebben we uit de naam van het bestand gepakt. (Zie 3)

We hebben de vakken ingedeeld in een lengte en een breedte vak omdat we alleen maar met nummers kunnen voorspellen en geen tekst. (Zie 4)

Aan de tabel met vakken hebben we nog een kolom ‘goal’ toegevoegd. Hier staat in of er een goal is gemaakt of niet. Dit hebben we gedaan omdat we dan later makkelijk kunnen voorspellen of je wel of niet scored vanuit een vak en dan kun je later makkelijk alle rijen ophalen waarin in goal is gemaakt. (Zie 5)

**Statistische analyse**

In de gps-data stond hoe hard een speler heeft bewogen in meter per seconde. Wij zijn ervan uitgegaan dat dit de gemiddelde snelheid van de speler was tijdens het interval sinds de vorige gps-meting. Aan de hand van deze snelheid hebben we de totale gelopen afstand van de speler berekend door het te vermenigvuldiger met het tijdsverschil tussen de huidige meting en de vorige. (Zie 6)

Ook hebben we gekeken wanneer de pauze is. Dit hebben we gedaan door te kijken waar het tijdsinterval tussen een meting en die daarvoor groter dan vijf minuten is. (Zie 7)

**Machine learning**

We hebben een voorspelling gemaakt waarmee je een vak kan aangeven en dan wordt er voorspeld of je wel (1) of niet (0) scored vanuit dit vak. Hiervoor hebben we k nearest neighbours gebruikt van scikit-learn met als x het lengte en breedte vak en als y of er een doelpunt is gemaakt of niet. Hierdoor kun je als je lengte en breedte vak geeft zien of je een goal zal maken of niet. (Zie 8)

De uitkomst van de voorspelling is echter altijd 0 (geen doel), omdat de bal heel vaak in alle vakken komt en er maar vier keer is gescored.

Daarom hebben wij zelf een berekening gedaan waaruit blijkt vanuit welk vak je het meeste hebt gescored. Hoe vaker er wordt gescored hoe betrouwbaarder deze percentages zullen zijn. Als je ziet dat er vanuit een vak het meeste doelpunten zijn gemaakt heb je meer kans om vanuit dit vak nog eens te scoren. (Zie 9)

Ook hebben we voorspeld hoeveel meter een speler heeft gelopen om een bepaalde tijd. Hiervoor hebben we lineare regressie gebruikt, omdat er een zeer hoge correlatie is tussen de tijd en gelopen afstand. (Zie 10)

**Visualisatie**

We hebben het percentage van de goals vanuit een vak geplot in bar chart, omdat dit overzichtelijk is. (Zie 11)

Ter verduidelijking hebben we de gelopen afstand in een grafiek gezet samen met onze voorspelling. (Zie 12)