Documentatie Data pipeline NEC

Gemaakt door: Matthias van Dijk, Victor Duits, Mitchell van Ek, Jochem Kuus en Cindy Tolboom

Table of Contents

Doel/Centrale vraag

Centrale vragen, focus van opdracht.

Resultaten per speler per wedstrijd.
Het aantal zones waarin een speler is geweest per actie.
Het gemiddeld aantal zones waarin spelers zijn geweest in een wedstrijd.
Waar het meeste wordt gespeeld tijdens een wedstrijd per zone.
Het totale rendement per wedstrijd.

Voorbewerking data

Alle bewerkingen op de data die gebruikt zijn om het doel te bereiken.

- Data omzetten naar het juiste formaat, zoals van XML naar CSV bestandsformaat.
- Verwijdering van flags die inlezing verstoren
- Data in dataframes zetten.
- Wedstrijd data koppelen aan spelernamen, voor speler-specifieke data.
- Data opsplitsen naar nieuwe dataframes, zodat de data nog specifieker is.
- Berekeningen met de data uitvoeren.
- Berekende data weer samenvoegen tot één specifiek geheel.
- Bewerkingen uitvoeren waarbij een eventueel verband kan ontstaan tussen de data.
- Aan de hand van berekeningen voorspellingen weergeven met de data.
- Data visualiseren gebruik makend van de juiste grafiek keuze.

Architectuur van data pipeline

Structuur van het Jupyter Notebook gezien vanuit de Data Science Workflow

- Eerst worden alle packages geïmporteerd die nodig zijn om de data te kunnen gaan bewerken.
- Daarna wordt de data omgezet naar dataframes zodat de data gerepresenteerd kan worden in tabellen.
- De data zal daarna gemanipuleerd(data exploratie) worden om bekend te raken met de data dit doen we door te vergelijken/groeperen/filteren op unieke waardes etc.
- Nadat de data is opgeschoond, kunnen er berekeningen gaan plaatsvinden, zoals het tellen van het aantal zones waarin een speler is geweest in een match etc.
- De data is daarna verder opgesplitst, zodat we nog specifieker te werk kunnen gaan. Voorbeeld uit de opdracht: er is een dataframe gemaakt van de zones van het voetbalveld.
- Er kan daarna gekeken worden of er een verhouding is tussen deze data met de andere data, zoals het berekenen van de correlatie en of het normaal verdeeld is.
- Als laatste zal de data gevisualiseerd worden in bijpassende grafieken.

Data collection en munging

De data van NEC was opgeleverd in XML en txt formaat. De data van de XML bestanden hebben we omgezet naar CSV bestanden, zodat de data wat makkelijker in een dataframe omgezet kan worden.

Data storage

Niet gebruikt.

Analysis

1. Voor de betrouwbaarheid van de data, hebben we eerst gekeken naar de data die eventueel verband met elkaar kunnen hebben, zoals spelers en zones. Vanuit daar ga je analyseren welke zones er betreden worden per speler per wedstrijd. Daarna kan je die zones koppelen met bijvoorbeeld een specifieke group zoals het rendement van een wedstrijd. Omdat de data is opgesplitst per speler kan je heel gemakkelijk kijken of dit een verband met elkaar heeft. Het is dus reproduceerbaar en daardoor ook betrouwbaar.
2. Voor de data van NEC hebben we supervised learning toegepast. De data was namelijk al ge-labeled en moest alleen nog maar opgesplitst worden, zodat er geavanceerd onderzoek/manipulatie/berekeningen mee gedaan konden worden.

Visualisation

Voor de visualisatie van de data hebben we Bokeh gebruikt.

Communication

Oplevering:

- Jupyter notebook, staat op GitHub
- Documentatie

Extra databronnen

Geen.

Onderbouwing keuzes

Python packages

- **Pandas**: om de data in dataframes te representeren en om de data op te splitsen in nieuwe dataframes.
- **Numpy**: voor algemene berekeningen.
- **Bokeh**: voor de visualisatie van de data.
- **Matplotlib**: voor de visualisatie van de data.
- **Scipy**: voor de probability plots.
- **Os**: om de bestandsnamen te koppelen aan een wedstrijd.

Missing values

De data wordt bij sommige berekeningen niet meegenomen als het een NaN is.
Bij de meeste dataframes staan de missing values er gewoon nog in omdat het best nuttig kan zijn bij bepaalde berekeningen.

Verdeling dataset

De dataset van NEC hebben we verdeeld aan de hand van de zones. Er zijn dus dataframes gemaakt die gegroepeerd zijn op zone per wedstrijd. Tevens is er een specifieke dataframe gemaakt die per zone bijhoudt hoeveel keer een speler in die zone is geweest, per wedstrijd.