Criminal Soccer – a Big Data Perspective

Research Question:

Does the crime rate of a players nation, impact his behavior in regards to committed fouls.

Characteristics

This section of the document is dedicated to discussing the 5 v's of big data characteristics and their relevance to this project.

Variety:

In order to answer the research question, we needed to combine multiple data sources that differ in structure and format.

- **Foul Statistics for players:**
  We found a dataset that contains foul statistics across multiple csv files.
- **Player Nationalities:**
  As the csv dataset did not contain player nationalities, we needed to gather that information from another source. We built a web scraper that returned html data.
- **National Crime Statistics:**
  In order introduce national crime rates we created a web scraper that fetched data from Wikipedia. This program gathered html documents.

As this project collects data from multiple sources in various formats, it classifies as Big Data from the perspective of Variety.

Value:

Answering this research question would positively impact talent development leagues and children's teams as it could provide insights on where to increase the focus on fair play. That would lead to less injuries and increased companionship amongst young players.

Therefor this project classifies as Big Data from the perspective of Value.

Veracity:

This project needs to handle inconsistencies, when combining data sources. As the csv data did not provide information regarding player's nationalities, this information needed to be collected from the web. Within this process, the research team needed to handle the following cases:

Case 1: No information for given search term:

For some players the Webpage did not deliver search results for their name. The project team decided to handle these cases by setting nationality to None.

Case 2: Multiple search results for given search term:

When entering a player name, the search Table would show multiple results. The project team decided to handle these cases by selecting the top ranked search result.

Due to the inconsistencies between data sources this project classifies as Big Data from the perspective of Veracity.

Velocity:

In order to answer the given research question real time analytics were not required.

Therefor this project does not classify as Big Data from the perspective of Velocity

Volume:

This project works with information that can be stored on a single and does not need clustering. Therefor this project does not classify as Big Data from the perspective of Volume.

This project can be classified as Big Data as it fits 3 out of 5 criteria points.

4 Levels of Data Handling

This section discusses the decisions made along the data processing pipeline

Data Source Layer:

The Data for this project was sources from csv files and html files. In order to collect the base data the project team has downloaded a Kaggle dataset in the “comma separated value” format. Further the team created 2 web scrapers do download html files from the web.

Data Storage Layer:

The data was initially stored using the ETL (Extract, Transform, Load) approach.

Extract: Data extracted as further described in Data Source Layer.

Transform: The extracted data was then turned into python dictionaries. Furthermore typecasting was performed for predefined contents of interest . Some data sources where combined to form relevant documents.

Load: After transformation, the single dictionaries where loaded to MongoDB collections. MongoDB is a NoSQL document database designed to store large amounts of data with flexible schema.

Data Analysis Layer:

As this project did not require any machine learning or AI processing steps, Data analysis was primary performed by using the python framework pandas.

Additionally the team implemented a custom made map reduce process that analyses the payer appearances per country. That information is then stored in a MySQL database.

Data Output Layer:

The results are presented in a Jupiter notebook. Charts where primarily created using matplotlib, as it is one of the standard frameworks for plotting and visualizing Information.