Visual Analytics Project Proposal

A.Y. 2024-2025

Davide Fortunato 1936575

Davide Mazzatenta 2140428

Dataset

The dataset is about a football game called FIFA. It has data of all the players from different versions of the game from 2015 to 2023. In our case the AS index is roughly 161584 * 110 = 17.774.240 (more on dataset treatment in the “Visual Analytics Cycle“ part). We found it on Kaggle at the following link https://www.kaggle.com/datasets/joebeachcapital/fifa-players and its name is *male_players (legacy)_23.csv*. Every tuple is a football player with attributes concerning physical/technical performances and team/personal information.

General Idea

Analytics Part: The project will use dimensionality reduction techniques to condense high-dimensional football players features into a 2D space and clustering algorithm like k-means++ in order to divide data into groups based on similar attributes.
Visual Part: The project will feature 6 coordinated and interactive visualizations:

1. A football field representing different roles of the players;
2. A bar chart showing single player features compared to aggregated feature values of its cluster;
3. A scatter plot representing aggregated data in which each point is a player and it is coloured based on clusters;
4. A line chart to represent temporal information and series like wage over years;
5. A radar chart that shows the principal characteristics of the player:
6. Player image that will pop up when a single player is selected through the scatter plot.

Intended user

This project aims to provide a useful tool for a broad range of football-engaged people: from **coaches** and **football analysts** interested in tracking performances of a specific player across seasons or technical planning, to the **scouts** who are looking for players with specific characteristics and that can be similar to more famous players, besides the **football fans** and **FIFA game players** who crave for in-depth players statistics and comparisons.

Used analytics

PCA or t-SNE will be used to reduce the dataset's numerical features to two dimensions in the scatter plot.
Football players will be grouped based on their most important features and aggregated into clusters with the k-means++ algorithm.
When a user selects a role in the football field the scatter plot with all the players in that role will update and through it the user can select a player that will appear in the image and in the other charts with the top 5 or 10 similar players.

Relation with the visual analytics cycle

1. **Data Preparation**: The dataset has lots of useless attributes for our analysis so we will proceed to clean the dataset, keeping information related to numerical attributes like speed, strength, shoot_power and so on. We will also be using encoding algorithms, like one-hot encoding or label encoding, for those categorical features that we consider useful for the analysis.
2. **Dimensionality Reduction**: Refer to the “Used analytics” part above.
3. **Interactive Exploration**: Users can interact with visualizations, such as the scatterplot or the football field, to explore player clusters or specific player features.
4. **Feedback Loop**: Insights from user interactions help refine further analysis based on specific attributes like role, player wage and so on, besides comparing players with similar play styles.
5. **Knowledge Generation**: Coordinated visualizations reveal trends, relationships, and outliers in player data, assisting users in making informed evaluations and discoveries.

Mockup of the user interface