**Relational and NoSQL Databases**

**Advanced Databases/Databases Technologies**

2024/2025 Project – 1st Checkpoint

1. **Group Details:**

Group number: 8

Members: Afonso Baptista 58213; Miguel Borges 58187; Miguel Dinis 58198; Rafael Correia 58256

Every member of the group contributed equally to the development of this project.

* Afonso Baptista:
  + .
  + .
  + .
* Miguel Borges:
  + .~
  + .
  + .
* Miguel Dinis:
  + ..
  + ..
* Rafael Correia:
  + .
  + .

1. **Project Description**

2.1. Phase 1: Data Modelling and Querying

Dataset Description

The dataset we chose is a Kaggle dataset (link to the dataset is available at the end of the report) on international football match results, goal scorers and shootouts. It includes 47917 results of international football matches from 1872 up to 2024. The matches are strictly men’s full internationals, and the data doesn’t include Olympic Games or matches where at least one of the teams was the B team, U-23 or a league select team. This dataset is divided into three files:

* *results.csv*: has the list of match results. Includes the following columns:
* *date*: date of the match.
* *home\_team*: the name of the home team.
* *away\_team*: the name of the away team.
* *home\_score*: full-time home team score including extra time, not including penalty-shootouts.
* *away\_score*: full-time away team score including extra time, not including penalty-shootouts.
* *tournament*: the name of the tournament.
* *city*: the name of the city/town/administrative unit where the match was played.
* *country*: the name of the country where the match was played.
* *neutral*: whether the match was played at a neutral venue.
* *goalscorers.csv*: has the list of scorers in the games. Includes the following columns:
* *date*: date of the match.
* *home\_team*: the name of the home team.
* *away\_team*: the name of the away team.
* *team*: the name of the team that scored the goal.
* *minute*: the minute the goal was scored.
* *scorer*: name of the player that scored the goal.
* *own\_goal*: whether the goal was an own goal.
* *penalty*: whether the goal was a penalty.
* *shootouts.csv*: has the list of winner teams of penalty-shootouts. Includes the following columns:
* *date*: date of the match.
* *home\_team*: the name of the home team.
* *away\_team*: the name of the away team.
* *winner*: winner of the penalty-shootout.
* *first\_shooter*: the team that went first in the shootout.

Observations

* In *goalscorers*, some entries lacked the *minute* field, leading to their removal.
* The *first\_shooter* in the table *shootouts* had roughly 64% missing values, requiring imputation.
* *data* fields were stored as objects instead of datetime formats.

Data Cleaning Process

* Removing missing values: rows missing the critical *minute* value in *goalscorers* were dropped as they could not reliably represent goals.
* Imputation: missing *first\_shooter* values in the *shootouts* table were filled with a random selection between the *home\_team* and the *away\_team*. These values are not critical, and their imputation ensures that the data is complete without biased results.
* Data type conversion: *date* values were converted into actual datetime objects for consistency and to enable time-based filtering.

2.1.1. Relational Schema

The data was modelled into three core entities:

* Results: Represents matches with attributes such as date, teams, scores
* ..
* ..

2.1.2. MongoDB Collections

2.2. Phase 2: Indexing and Optimization

1. **Conclusion**
2. **References**

Original dataset’s link: <https://www.kaggle.com/datasets/martj42/international-football-results-from-1872-to-2017>