**CSYE7200 FINAL PROJECT REPORT**

**Shangqing Hu 1374342**

**Junlong qiao 2784609**

**Yuxuan cheng 2773368**

**Abstract**

This project aims to analyze data on players in FIFA23 using the Scala programming language. The project involves choosing a Scala-based data processing framework ── Apache Spark and mining data using FIFA23's API and loading the data into the framework. The data is cleaned and pre-processed to remove inconsistencies and missing values. Exploratory data analysis is conducted using Scala-based data visualization libraries. Predictive models are developed using Scala-based machine learning libraries such as MLlib, Smile, or Deeplearning4j. Use clustering algorithms to discover players with potential in lower leagues, discover players with the same technical characteristics, and show the technical style of players in each league. The project has practical applications in sports management and scouting, allowing teams to identify and recruit players that meet their specific needs and goals using the power and flexibility of Scala.

**User cases**

1. The gamer could use this system to find their target player with the attribute they want.
2. The gamer could use the recommend system to find the players similar to the target player.
3. The gamer could use the predictive model to find high potential players below the premier league.

**Methodology**

In soccer games, we often encounter situations when players are injured and live in poor form. If you don ‘t wants to change the whole tactical system of your team because of the absence of a player, it is important to find a replacement for that player with the same technical characteristics. Therefore, the system clusters players according to the position they play and their usual foot, speed of holding the ball, dribbling skills, etc. as feature values to find players with similar tactical styles. We use the k-means algorithm from the machine learning library in Spark to help us achieve the clustering of players at each position.

The High Potential Player Search System is a powerful tool for uncovering a player's potential based on their performance.

This project seeks to go beyond basic scouting by analyzing various features such as overall score, potential score, value, age, height, and weight to determine whether a player has the potential to enter a top league, such as La Liga, Bundesliga, etc.

**Data Source**

* <https://www.kaggle.com/datasets/stefanoleone992/fifa-23-complete-player-dataset>
* ~1000000 row
* 110 attributes

**Acceptance criteria**

1. Search for relevant players can be completed in less than 8 seconds (approximately 7.84s).
2. The recommend system can give at least 10 similar players.
3. Our accuracy of our predictive model is higher than 0.9.

**System environment**

1. Java-version:8
2. Scala:2.13
3. Tomcat 9
4. Apache.spark-sql:2.13-3.3.0
5. Apache.spark-mllib:2.13-3.3.0
6. Apache.sparktest: 2.13- 3.2.15

**Reference**

**[1]. Wikipedia(Online), Available: <https://en.wikipedia.org/wiki/Confusion_matrix>**