**2022 SOA student case challenge analysis – team Scout Cookies🍪**

An imaginary country Rarita wants to qualify for the FSA League within the next five years. Rarita: East Rarita, Central Rarita and West Rarita. The motivation is behind this proposal is to benefit economically. With money awarded for winners, they may invest in local sports arenas and sports leagues. The winning country also gains a global visibility, providing opportunities for more investment, tourism, and political influence to achieve final economic and growth.

The key objectives of the project are as follows:

• Construct a competitive team using the defined constraints

• Analyze the impact of building a Football “brand” on Rarita’s economy over the next 10 years using key economic indices, probabilities, and a cost/benefit analysis.

# Team selection

## Background

A soccer team usually consists of 18-21 players, of which 11 are allowed to play in the field at the same time and the others are substitutes. There are 4 positions in a team: forward, midfielder, defender, and goalkeeper. Generally speaking, the first three positions are transferable and share common skills. Therefore, we select players for these positions from the same mixed pool of forward-midfielder-defender candidates. The goalkeeper's role is very different from the other players. So the training data and model analysis of the goalkeeper is separated from the other positions.

## Goals

1. Use information of *player* *data,* select around 20 players from around 5000 the candidates. (completed)
2. Give probability ranges of the “success” of being competitive. (ongoing…)
3. The amount of spending required on assembling the team over the next 10 years, including discussions of the sources of non-governmental funding. (completed)
4. Direct team revenues generated from having a “competitive” team. (completed)

## Method

K-nearest-neighbors model (KNN model)

## Implementation

#### Data preparations

Inner join 3 tournament tables (tournament shooting, tournament passing and tournament defense) and select useful attributes for analysis. Create a joined table, *Tournament.Joined*. The same operations are applied to 3 league tables (league shooting, league passing and league defense). The joined table is *League.Joined*.

Match tournament team rankings (1-24) to players, based on the nation team they belong to. This is used as the first criteria to select candidates. A play with small ranking of nation team is desirable. We use KNN model to achieve the selection.

A similar data preparation process is applied to goalkeeper as well.

#### Decision tree method attempt (not used)

After doing some research and reading some academic papers related to similar problems, we find that the decision tree model is one popular option for sports player selection.

The decision tree model is a supervised machine learning algorithm that can potentially solve the problem of selecting adequate soccer players for Team Rarita. Similar to any other machine learning algorithm, we first split the data into 2 parts, the training set and the testing set. By using the R package ‘rpart’, we build a partitioned regression model that predicts the rank of a player given some of its attributes. With rpart, we obtain a trained decision tree model fit to the training dataset.

After generating the predicted player ranks using the trained model, we test the accuracy of the model. Unfortunately, in this case, decision tree yields a low accuracy rate of 16.8%. One potential problem of this method is overfitting as it fails to capture the important features. Another potential problem is poor attribute choice, the attributes we chose may not be adequate for this particular model.

As a result, we decided to apply other machine learning approaches to build our model.

#### KNN Model training, testing and application

We first normalized all the players’ data. We then divide non-goalkeeper tournament data (in *Tournament.Joined*) into two parts, where 90% of the data are for training and the other 10% are for testing. In the testing set, we compare the predicted team rankings for players with the actual rankings. With an error tolerance of 3, using cross validating, we find that the optimal k value that maximizes the accuracy rate is 13 where the accuracy rate reaches more than 50%. The use tournament data as training data and league data as testing data with k = 3, we the predicted team ranks for the players.

A similar modelling process is applied to goalkeeper data as well. (optimal k=1, testing accuracy rate=83.3%)

#### Team assembling

For non-goalkeeper players, with the predicted rankings, we initially selected 319 players out of around 1500 candidates in league data. We consider *goals per 90 minutes, pass completion %* and *tackles wining the possession of the ball %* as key attributes which make a good non-goalkeeper player. By filtering on these key attributes, 18 top performers are finally selected as non-goalkeeper players for the team.

For goalkeepers, with the predicted rankings, we initially selected 22 players out of around 425 candidates in goalkeeper league data. We consider *goals against per 90 minutes* and *saves per shot on target %* as key attributes which make a good goalkeeper player. By filtering on these key attributes, 2 top performers are finally selected as non-goalkeeper players for the team.

#### Cost & revenue analysis

Cost is the next 10 years’ total salary of selected players. Revenue is the total loaning fees of the Raritan-league players of the next 10 years. Inflation rate and salary increasing are considered based on judgements. Calculations are based on the following formula.

Where is the 2021 salary for player i, is the constant inflation rate, and is the salary increasing rate for player .

Sensitivity testing may be implemented for cost and revenue analysis in the later process.

# What to do next?

1. Construct implementation plan.
2. Perform economic impact analysis.
3. Perform risk analysis and risk mitigation considerations.

# Works cited

Scikit-learn: *Machine Learning in Python*, Pedregosa et al., JMLR 12, pp. 2825-2830, 2011.

Venkataraman, A. (1999). *Issues in decision tree learning*. From: http://pandamatak.com/people/anand/771/html/node29.html

Patil, N., Sequeira, B. H., Gonsalves, N. N., & Singh, A. A. (2020). *Cricket team prediction using Machine Learning Techniques*. SSRN Electronic Journal. From: https://doi.org/10.2139/ssrn.3572740

Singh, B. D. (2020, August 14). *Buying a soccer team: A machine learning approach*. From: https://towardsdatascience.com/buying-a-soccer-team-a-machine-learning-approach

Sirohi, K. (2018, December 30). *K-nearest neighbors algorithm with examples in R (simply explained knn).* From: https://towardsdatascience.com/k-nearest-neighbors-algorithm-with-examples-in-r-simply-explained-knn