

卒業論文

# ランダムフォレストによるサッカー W 杯の優勝国予想

2014011 石田亜斗武

指導教員 六井 淳 教授  
副査 渡邊貴之 教授

2024 年 1 月

静岡県立大学経営情報学部



## 概要

日本は 2022 年 FIFA ワールドカップでスペインやドイツと同じグループに入り、予想外の好成績を収めた。このように予測困難なスポーツの試合結果を予測するために、統計的手法や機械学習が注目されている。そこで本研究では自身のサッカー経験と AI や機械学習の知識を活用して、ワールドカップの優勝国を予測するモデルの提案に着目した。

本研究は、過去の FIFA ワールドカップのデータを使用して、優勝国を予測するための予測モデルを構築することを目的とする。スポーツの試合に関するデータの活用が注目されており、予測結果は FIFA ワールドカップに関わる人々に有益な情報を提供することが期待できる。本研究では、ランダムフォレストを用いてモデルを作成する。多数の決定木を組み合わせることで過学習を防ぎながら、重要な特徴量を評価し複雑なデータ構造を学習することができる、スポーツの勝敗予測において他の手法に比べ予測精度が高いなどの理由からランダムフォレストを採用する。モデルの評価では精度や再現率、F1 スコアなどの評価指標を使用する。

## Abstract

Japan unexpectedly performed well in the 2022 FIFA World Cup, placing in the same group as Spain and Germany. Statistical methods and machine learning have been attracting attention for predicting the outcome of such difficult-to-predict sports matches. Therefore, this study focused on proposing a model to predict the winning country of the World Cup by utilizing his own soccer experience and knowledge of AI and machine learning.

This study aims to build a prediction model to predict the winning country using data from past FIFA World Cups. The use of data on sports matches is attracting attention, and the prediction results are expected to provide useful information to people involved in the FIFA World Cup. In this study, the model is created using a random forest. Random forests are employed because they can evaluate important features and learn complex data structures while preventing overlearning by combining a large number of decision trees, and they have higher prediction accuracy than other methods in sports win/loss prediction. Evaluation of the model will use metrics such as accuracy, reproducibility, and F1 score.

# 目次

第 1 章	はじめに	1
1.1	研究背景 . . . . .	1
1.2	研究目的 . . . . .	1
1.3	論文の構成 . . . . .	2
第 2 章	関連技術	3
2.1	ランダムフォレスト . . . . .	3
2.1.1	特徴選択 . . . . .	3
2.2	データの前処理 . . . . .	3
2.3	ハイパーパラメータチューニング . . . . .	3
第 3 章	検証実験	4
3.1	データ詳細 . . . . .	4
3.2	評価指標 . . . . .	4
3.2.1	混合行列 . . . . .	4
3.2.2	正解率 . . . . .	4
3.2.3	適合率 . . . . .	4
3.2.4	再現率 . . . . .	4
3.2.5	F 値 . . . . .	4
3.3	検証内容 . . . . .	4
3.4	検証結果 . . . . .	4
3.5	考察 . . . . .	4
第 4 章	まとめと今後の課題	5
4.1	まとめ . . . . .	5
4.2	今後の課題 . . . . .	5
	謝辞	6
	参考文献	7



# 第 1 章

## はじめに

### 1.1 研究背景

2022 年 FIFA ワールドカップでは、日本は死のグループと呼ばれる、スペインやドイツといったサッカーの強豪国と同じグループ E に属することになった。しかし、結果はスペインとドイツの両国に勝利し、グループステージを突破し、敗戦濃厚とされていた強豪国クロアチアに PK 戦であと一步のところまで負けてしまったが、世界中に大きな驚きを与えた [1]。このように、近年では予測が困難なスポーツの試合結果を科学的に分析するアナリティクスの分野が注目され、予測の精度を高めるために統計的手法や機械学習が利用されるようになっている [2]。そこで本研究では、私自身の長年のサッカーの経験と、人工知能や機械学習の知識を使い、ワールドカップの優勝国を予測できるモデルの提案に着目した。

### 1.2 研究目的

本研究の目的は、過去の FIFA ワールドカップのデータをもとに、優勝国を予測する高精度な予測モデルを構築することである。近年ではスタッツデータと呼ばれるスポーツの試合に関するデータの活用の注目が高まり、リーグやチームの選手強化や戦術立案、スポーツベッティングで利用されている [3] ことから、本研究の予測結果は FIFA ワールドカップに関わる人々にとって有益な情報を提供することが期待できる。

本研究でランダムフォレストを採用した理由は、その柔軟性と予測精度の高さにある。スポーツの勝敗予測に対して有用な特徴変数を確定することが困難である [4] という問題に対して、ランダムフォレストは多数の決定木を組み合わせることで過学習 [5] を防ぎながらも、重要な特徴量を効果的に評価し、複雑なデータ構造を学習することができる [6]。また、先行研究 [7][8] から他手法に比べて予測精度が高いことも分かる。モデルの評価では、精度や再現率、F1 スコアなどの評価指標 [8] を用いてモデルを評価する。

## 2 第1章 はじめに

### 1.3 論文の構成

第2章では、本研究の関連技術と用語について述べる。

第3章では、検証実験について述べる。

第4章では、まとめと今後について述べる。



## 第 2 章

# 関連技術

### 2.1 ランダムフォレスト

#### 2.1.1 特徴選択

### 2.2 データの前処理

### 2.3 ハイパーパラメータチューニング

## 第 3 章

# 検証実験

### 3.1 データ詳細

### 3.2 評価指標

#### 3.2.1 混合行列

#### 3.2.2 正解率

#### 3.2.3 適合率

#### 3.2.4 再現率

#### 3.2.5 F 値

### 3.3 検証内容

### 3.4 検証結果

### 3.5 考察

## 第 4 章

# まとめと今後の課題

### 4.1 まとめ

### 4.2 今後の課題

## 謝辭

## 参考文献

- [1] 公益財団法人日本サッカー協会, FIFA ワールドカップカタール 2022, JFA.jp, [https://www.jfa.jp/samuraiblu/worldcup\\_2022/schedule\\_result/](https://www.jfa.jp/samuraiblu/worldcup_2022/schedule_result/)
- [2] 谷岡広樹, ”スポーツアナリティクスにおけるデータと AI 活用”, P1, 2020
- [3] 経済産業省, スポーツ庁, ”スポーツ DX レポート”, 2022
- [4] Milad Keshtkar Langaroudi, Mohammad Reza Yamaghani, “Sports Result Prediction Based on Machine Learning and Computational Intelligence Approaches A Survey” , P1, 2019
- [5] TRYETING”機械学習における過学習とは何か？原因・回避方法をくわしく解説”, <https://www.tryeting.jp/column/6846/>
- [6] Matthias Schonlau, Rosie Yuyan Zou, “The random forest algorithm for statistical learning” , P4, 2020
- [7] Rahul Baboota, Harleen Kaur, “Predictive analysis and modelling football results using machine learning approach for English Premier League” P15 2018
- [8] Teno Gonz’alez, Dos Santos, Chunyan Wang Niklas, Carlsson Patrick Lambrix, Linköping University, Sweden. “Predicting Season Outcomes for the NBA” , P11, 2021
- [9] TRYETING”機械学習における評価指標とは？種類から学習方法まで徹底解説”, <https://www.tryeting.jp/column/6042/>

## 付録 A