**FIFA World Cup 2022 Analysis & Predictions**

**Goal:** To create a FIFA prediction model to predict the 2022 Qatar world cup winners.

**Importance of the project:**

* Prediction models in Sports industry are generally built on skillsets. But when players are playing on bigger stage It takes more than skillsets to win a world Cup.
* Mentalities of these players makes or breaks the team. It is this mentality that sets a great team apart from the good team it is this mentality that helps a team to win a tournament like World Cup.
* So, in this project by Using mentalities along with skillset we are building a prediction model which can help in predicting the world cup winner (as accurate possible by cross verifying with real time progress).

**How was it made Possible:**

* **Data gathering and Transformation**
  + **Web scraping**
    - We scrape the data from official FIFA html pages using html tags and obtain final data in a data frame and export it as csv to obtain FIFA ratings df.
  + **Obtaining mentalities**
    - From the Kaggle we obtain so and so and filter by nationality name.
    - Select best 50 players of each country by their overall rating
    - We obtain average of these 50 player mentalities by country to obtain national mentalities df.
  + **Obtaining 'y' for predictive modeling**
    - consider the goal difference b/w home\_score and away\_score d = (home\_score - away\_score).
    - convert it into binary class 1 or 2 based on which team wins d>0 or d<0.
* **Data Augmentation**
  + For data augmentation we join national mentalities dataset with the official FIFA country ratings.
  + To obtain the final merged df we join matches df with the national mentalities and FIFA ratings based on home\_team and away\_team columns.
* **Data imputation**
  + We impute the missing attributes by using 25 percentile imputation.
* **Data generation for Simulations**
  + For generating data to predict match winners, we first construct the input dataframe for the model, given 2 countries.
  + Next, we add randomness to the input dataframe to run the simulation N number of times to obtain real results as much as possible.