Project Overview

This project combines web scraping and machine learning to predict the outcomes of English Premier League (EPL) football matches. We'll start by scraping historical match data, clean and preprocess it, and then build machine learning models to predict the outcome of future matches.

Part 1: Web Scraping Match Data

The first part of the project focuses on scraping football match statistics from the EPL. Using Python libraries like requests and BeautifulSoup, we scrape data such as match scores, shots, and other relevant match statistics. This data is cleaned and organized using pandas to prepare it for machine learning.

Part 2: Predicting Match Winners Using Machine Learning

In the second part, we use the cleaned and structured match data to train a Random Forest machine learning model to predict the winner of future matches. We evaluate the model's performance and make improvements to enhance prediction accuracy.

Features:

- Web Scraping:
  - Collect match data, including scores, shots on target, and other stats, from multiple EPL seasons.
  - Use Python libraries such as requests and BeautifulSoup to scrape and parse data.

- Data Preprocessing:
  - Clean and organize raw data with pandas to ensure it's ready for analysis.

- Machine Learning:
  - Train a Random Forest model using the cleaned data to predict the winners of future matches.
  - Assess model accuracy and apply improvements for better predictions.

- Scalability:
  - The methodology can be applied to other leagues or sports by adjusting the scraping and model-building process.

Technologies Used:

- Web Scraping: requests, BeautifulSoup
- Data Manipulation: pandas
- Machine Learning: scikit-learn (Random Forest model)

Data Sources

The dataset used for this project is scraped from the English Premier League website, containing match statistics for seasons starting from 2020 to 2022. The data includes match scores, shots, shots on target, and other performance metrics for each match.

How It Works

Part 1: Web Scraping Match Data

- **Data Downloading:**
  We use the requests library to download match data from the EPL website.

- **Data Parsing:**
  BeautifulSoup is used to parse the raw HTML and extract the match statistics such as scores, shots, shots on target, and other relevant metrics.

- **Data Cleaning:**
  The data is then loaded into a pandas DataFrame for cleaning, removing any irrelevant or redundant information, and structuring it for machine learning.

Part 2: Predicting Match Winners with Machine Learning

- **Model Training:**
  The cleaned match data is used to train a Random Forest model, predicting the outcome of each match.

- **Improvement and Evaluation:**
  After training the initial model, we evaluate its performance and improve it by integrating key features like rolling averages of stats (e.g., shots on target, goals scored).

- **Prediction:**
  The model can then predict the winner of future EPL matches by analyzing current statistics and historical trends.