ПРАВИТЕЛЬСТВО РОССИЙСКОЙ ФЕДЕРАЦИИ

ФГАОУ ВО НАЦИОНАЛЬНЫЙ ИССЛЕДОВАТЕЛЬСКИЙ УНИВЕРСИТЕТ

«ВЫСШАЯ ШКОЛА ЭКОНОМИКИ»

Факультет компьютерных наук

Образовательная программа «Прикладная математика и информатика»

УДК ХХХХХ

**Отчет об исследовательском проекте на тему:**

**Статистическое исследование по футбольным лигам Англии и Аргентины**

|  |  |  |  |
| --- | --- | --- | --- |
| **Выполнил**: |  |  | **25.05.2023** |
| студент группы БПМИ218  *Ковинский Руслан* |  |
|  | *(подпись)* |  | *(дата)* |
|  |  |  |  |
| **Принял руководитель проекта:** | |  |  |
| Приглашенный преподаватель, Аспирант Факультета компьютерных наук  / Департамента больших данных и информационного поиска  *Горшков Сергей Сергеевич* | | | |
| A close up of a signature  Description automatically generated with medium confidence |  |  | **25.05.2023** |
| *(подпись)* |  |  | *(дата)* |

**Москва 2023**

**Содержание**

[Аннотация 3](#_Toc489549988)

[1 Актуальность и постановка задачи 4](#_Toc1872581299)

[2 Обзор литературы 4](#_Toc1532304473)

[2.1 Изучение тем анализа данных и машинного обучения с нуля 4](#_Toc1803666303)

[2.2 Исследование “Who Will Score? A Machine Learning Approach to Supporting Football Team Building and Transfers”[1] 5](#_Toc1184495348)

[3 Предварительная обработка данных 5](#_Toc980933674)

[4 Анализ индивидуальных игроков 8](#_Toc1547985850)

[5 Анализ команд 11](#_Toc1492089283)

[6 Заключение 13](#_Toc1878622316)

[7 Список Литературы 14](#_Toc1238960333)

Аннотация

Целью данного курсового проекта является проведение статистического исследования, основанного на анализе данных о футбольных лигах Англии и Аргентины. Используя набор данных, содержащий информацию о командах, игроках и статистике матчей, мы исследуем различные аспекты футбольной игры и выявим интересные зависимости.

Мы начинаем с анализа индивидуальных игроков и их статистики. При помощи методов статистического анализа и машинного обучения, находим интересные закономерности и предсказываем различные параметры игроков, включая количество забитых голов и передач, а также оцениваем их влияние на результат команды.

Затем мы переходим к изучению команд в целом. Исследуем среднее количество голов, забитых и пропущенных каждой командой, и рассматриваем их распределение внутри лиги.

В итоге, данное исследование позволяет получить полноценное представление о футбольных лигах Англии и Аргентины. Оно раскрывает статистические тенденции, связанные с игроками и командами, а также предлагает интересные зависимости, которые могут быть полезны для понимания футбольной игры в этих регионах.

# 1 Актуальность и постановка задачи

Футбол является одной из самых популярных и любимых спортивных игр в мире. Множество футбольных лиг существуют в разных странах, предлагая нам захватывающие матчи и впечатляющие выступления. Понимание и анализ футбольных данных может предоставить нам ценную информацию о том, как команды и игроки достигают успеха, а также о факторах, влияющих на результаты матчей. Анализ данных о футбольных лигах может помочь нам лучше понять, какие факторы влияют на успех команд. Мы можем исследовать индивидуальные характеристики игроков и их влияние на результаты матчей. Также можем исследовать команды в целом, анализируя среднее количество голов, забитых и пропущенных каждой командой, и определяя их позиции в таблице.

В мире футбола уже давно участвуют не только игроки или традиционный персонал клуба, но также очень активно принимают участие спортивные аналитики, благодаря которым получается создавать все более удачные стратегии тренировок, проработок ошибок и даже питания, все это не говоря о создании спортивных приложений. Ожидаемые результаты этого исследования могут предоставить ценную информацию тренерам, скаутам и аналитикам, которые занимаются футболом.

2 Обзор литературы

2.1 Изучение тем анализа данных и машинного обучения с нуля  
 Для решения поставленной задачи были использованы материалы доступные на курсах [“Машинное обучение”](https://stepik.org/course/8057/promo#toc) и [“Анализ данных в спорте”](https://plucky-growth-d29.notion.site/9853f4e4506747d7ac874408c0ab499a), благодаря которым изучены следующие темы:

* + Основные понятия статистики, таких как среднее, дисперсия, корреляция, распределение и вероятность.
  + Классификация: Понимание методов классификации, таких как логистическая регрессия, метод опорных векторов (SVM) и решающие деревья.
  + Обзор методов машинного обучения: Понимание различных типов алгоритмов машинного обучения, таких как линейная регрессия, случайный лес, градиентный бустинг и методы классификации.
  + Оценка моделей: Понимание метрик оценки моделей, таких как среднеквадратичная ошибка (MSE) и точность (accuracy).

## 2.2 Исследование “Who Will Score? A Machine Learning Approach to Supporting Football Team Building and Transfers”[1]

В этой статье Bartosz Cwiklinski, Agata Giełczyk и Michał Choras представляют возможный подход к созданию профессиональной команды с помощью инструмента, основанного на машинном обучении. Предлагаемое решение дает многообещающие результаты и может помочь скауту или менеджеру команды в процессе планирования трансфера. Тем не менее, существуют некоторые угрозы достоверности, которые могут повлиять на полученные результаты и весь подход. Кроме того, спорт непредсказуем. Следовательно, указания, исходящие от ML, следует рассматривать исключительно как советы. Примечательно и то, что они учитывают не только технические и физические параметры профессионального футболиста, но и его психологическое состояние.

[1] [“Who Will Score? A Machine Learning Approach to Supporting Football Team Building and Transfers”](https://www.researchgate.net/publication/348384792_Who_Will_Score_A_Machine_Learning_Approach_to_Supporting_Football_Team_Building_and_Transfers)

# 3 Предварительная обработка данных

Процесс предварительной обработки данных включает несколько этапов, включая удаление пропущенных значений, преобразование категориальных переменных, масштабирование признаков и другие манипуляции для подготовки данных к анализу.

Данные были предоставлены в формате двух Excel документов, один из которых создан, используя исключительно испанский язык, который был переведен на английский.



Рис. 3.1: Пример столбцов таблицы.

Следующим шагом было преобразование типов данных в DataFrame. Преобразование типов данных в DataFrame — это процесс изменения типов данных столбцов DataFrame на соответствующие типы, которые лучше соответствуют хранящимся в них значениям. В случае, когда числовые данные имеют тип object после чтения из файла Excel, необходимо выполнить следующие шаги для преобразования типов данных:

* + Идентификация столбцов с числовыми данными: Сначала необходимо идентифицировать столбцы в DataFrame, которые должны иметь числовой тип данных, но вместо этого имеют тип object.ё

print(df.dtypes)

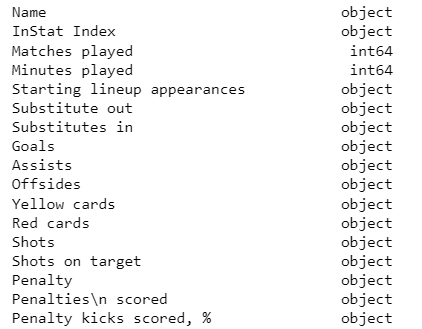


Рис. 3.2: Типы данных элементов каждого стобца.

* + Удаление неподходящих символов и пробелов: Перед преобразованием типов данных, потребуется удалить неподходящие символы, такие как символы валюты или разделители тысяч, а также пробелы из значений в столбцах, в нашем случае - это знаки « - » на местах, в которых должны стоять нули.

df = df.replace('-', '0')

* + Преобразование типов данных: Для преобразования типов данных в числовой формат можно использовать методы из библиотеки pandas. Например, можно использовать метод *astype(float)* для процентных значений и *astype(int)* для числовых для преобразования столбцов. Преобразование может быть выполнено следующим образом:

df = df.astype({name:'int' **for** name **in** \

list(df.columns[**1**:]) **if** name[-**1**] != '%'})  
 for each **in** df:  
 **if** each[-**1**] == '%':  
 df[each] = df[each].str.replace('%', \ '').astype(float) / **100**

Проверка результатов преобразования:

**print**(df.dtypes)

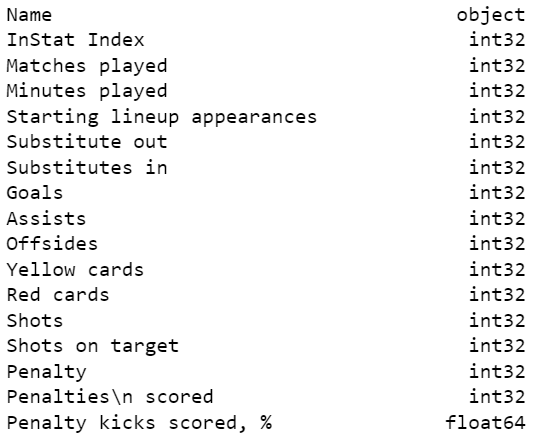


Рис. 3.3: Типы данных элементов каждого столбца после изменения.

Как можно заметить из рисунка, все числовые значения теперь имеют тип данных **int32**, а процентные **float64.** Предварительная обработка данных закончена.

# 4 Анализ индивидуальных показателей игроков

В этой части проведен анализ индивидуальных показателей игроков, их статистики и влияния на результаты команд. Использованы различные модели машинного обучения, такие как линейная регрессия, случайный лес и градиентный бустинг, чтобы предсказать различные параметры игроков. В данном отчете показана только основная часть статистического анализа, остальное находится по [ссылке](https://github.com/KovinskiyRuslan/Statistical_study_of_football_leagues).

Благодаря библиотеке pandas, мы можем очень быстро находить всевозможную информацию о игроках, например:

Найдем игрока, который провел на поле больше всего времени на поле:

most\_played\_player = df[df['Minutes played'] == \

df['Minutes played'].max()]['Name']

Найдем игрока, который забил больше всего голов:

top\_scorer = df[df['Goals'] == df['Goals'].max()]['Name']

Найдем игрока, который сделал больше всего пасов:

most\_passes\_player = df[df['Passes'] == df['Passes'].max()]['Name']

Благодаря библиотеке seaborn и matplotlib возможно визуализировать данные.

correlation\_matrix = df[["Goals", "Assists", "Shots on target"]].corr()   
  
sns.heatmap(correlation\_matrix, annot=True, cmap='coolwarm')   
plt.title("Correlation matrix of Goals, Assists and Shots on target")   
plt.show()

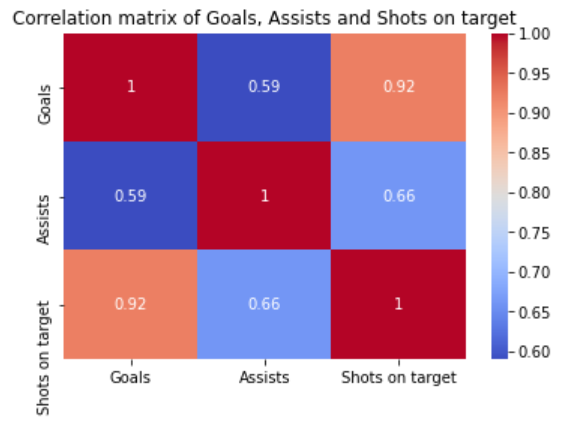


Рис. 3.3: Корреляционная матрица голов, пасов и ударов по воротам.

Гистограмма распределения игроков по количеству сыгранных матчей:

plt.figure(figsize=(10, 6))  
sns.histplot(df["Matches played"], kde=True, bins=20, color='skyblue')  
plt.title("Distribution of Players by Matches Played")  
plt.xlabel("Matches Played")  
plt.ylabel("Number of Players")  
plt.show()

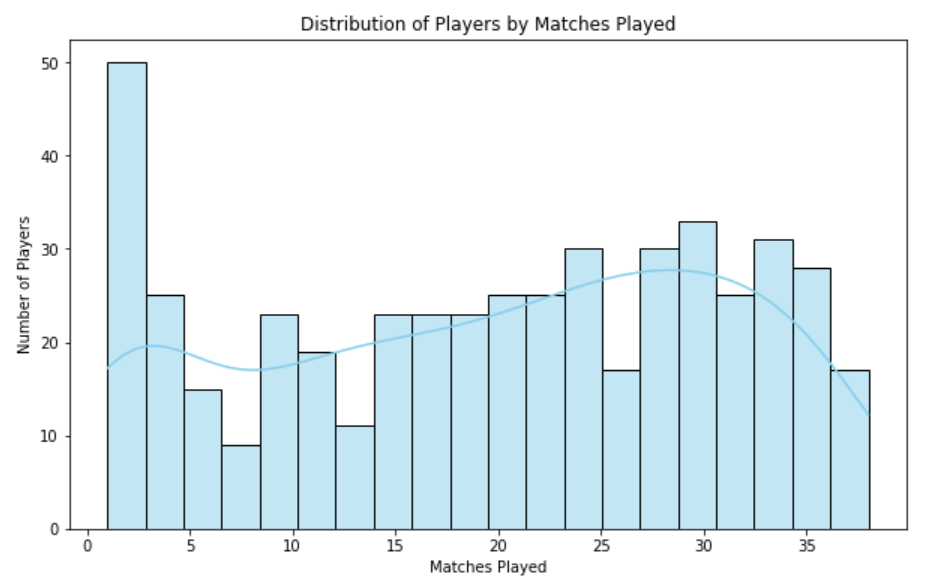


Рис. 4.2: Гистограмма распределения игроков по количеству сыгранных матчей.

Следующим шагом будет использование библиотеки sklearn. Sklearn - это популярный пакет Python, который широко применяется в области Data Science и Machine Learning. Он предоставляет набор функций и алгоритмов для решения задач машинного обучения, таких как классификация, прогнозирование и кластеризация данных.

Произведем сравнение двух моделей машинного обучения.

Предскажем количество забитых голов игрока на основе других переменных, таких как 'Assists', 'Shots on target', 'Matches played' и 'Minutes played' с помощью линейной регрессии.

from sklearn.model\_selection import train\_test\_split  
from sklearn.linear\_model import LinearRegression  
from sklearn.metrics import mean\_squared\_error, r2\_score  
  
Извлечение нужных переменных:  
X = df[['Assists', 'Shots on target', 'Matches played', 'Minutes played']]  
y = df['Goals']  
  
Разделение данных на тренировочный и тестовый наборы:  
X\_train, X\_test, y\_train, y\_test = train\_test\_split(X, y, test\_size=0.2, random\_state=42)  
  
Обучение модели линейной регрессии:  
model = LinearRegression()  
model.fit(X\_train, y\_train)

Предсказание на тестовых данных:  
y\_pred = model.predict(X\_test)  
  
Оценка модели:  
mse = mean\_squared\_error(y\_test, y\_pred)  
r2 = r2\_score(y\_test, y\_pred)

**Mean Squared Error: 1.5561937760741906**  
**R^2 Score: 0.8373375040094866**

В этот раз используем модель случайного леса (Random Forest), которая является более сложной моделью машинного обучения. Она состоит из ансамбля решающих деревьев и может предоставить более точные прогнозы, чем линейная регрессия.

from sklearn.ensemble import RandomForestRegressor  
from sklearn.preprocessing import StandardScaler  
  
Масштабирование данных:  
scaler = StandardScaler()  
X\_train\_scaled = scaler.fit\_transform(X\_train)  
X\_test\_scaled = scaler.transform(X\_test)  
  
Обучение модели случайного леса:  
rf\_model = RandomForestRegressor(n\_estimators=100, random\_state=42)  
rf\_model.fit(X\_train\_scaled, y\_train)  
  
Предсказание на тестовых данных:  
rf\_y\_pred = rf\_model.predict(X\_test\_scaled)  
  
Оценка модели:  
rf\_mse = mean\_squared\_error(y\_test, rf\_y\_pred)  
rf\_r2 = r2\_score(y\_test, rf\_y\_pred)

**Random Forest Mean Squared Error: 1.4892855670103093**  
**Random Forest R^2 Score: 0.8443311422413793**

Как мы можем видеть, модель случайного леса оказалась удачнее.

5 Анализ команд

Проведем анализ команд в целом, их средней статистики и распределения результатов. Исследовуем среднее количество забитых и пропущенных голов, определим самые результативные команды или команды с лучшей защитой. Можно также классифицировать команды на атакующие, защитные или сбалансированные, используя методы классификации.

Расчет среднего количества голов по командам:  
avg\_goals\_by\_team = df.groupby('Team')['Goals'].mean().sort\_values(ascending=False)

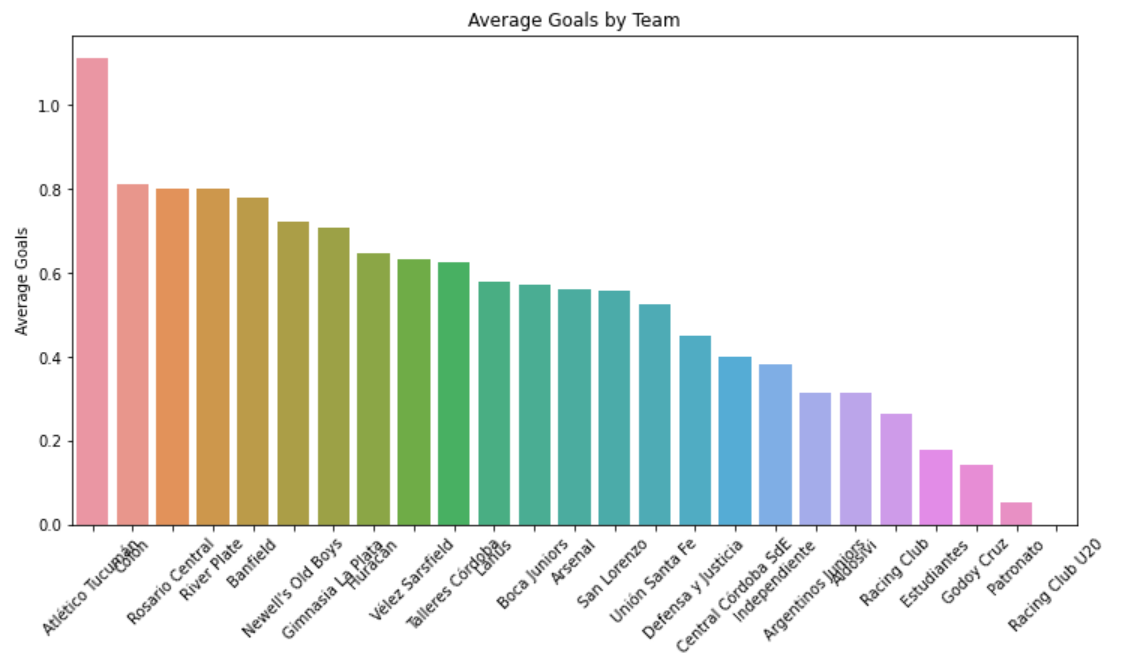


Рис. 5.1: Гистограмма распределения среднего количества голов на команду.

Для анализа эффективности команд в нападении можно взять во внимание следующие показатели:

* "Goals/90": Среднее количество забитых мячей за 90 минут игры. Это основной показатель эффективности в атаке.
* "MeanSucAtt/90": Среднее количество успешных атакующих действий за 90 минут. Это помогает понять, насколько команда активна в атаке.
* "MeanShots /90": Среднее количество ударов по воротам соперника за 90 минут. Это показывает, насколько команда часто пытается нанести удар по воротам соперника.
* "Assists /90": Среднее количество голевых передач за 90 минут. Этот показатель помогает понять, насколько команда способна создавать шансы для забивания голов.
* "xG/90": Среднее ожидаемое количество голов за 90 минут. Этот статистический показатель позволяет оценить качество создаваемых командой атак.

Группируем данные по командам и вычисляем среднее значение для интересующих нас столбцов:  
team\_stats = df.groupby('Team')[['Goals/90', 'MeanSucAtt/90', 'MeanShots/90', 'Assists/90', 'xG/90']].mean()  
  
Теперь у нас есть средние показатели для каждой команды. Мы можем сортировать команды по любому из этих показателей. Например, отсортируем команды по среднему количеству голов за 90 минут в убывающем порядке:  
best\_attack\_teams = team\_stats.sort\_values('Goals/90', ascending=False)  
  
Выводим первые 5 команд с наибольшим средним количеством голов за 90 минут:  
print(best\_attack\_teams.head())

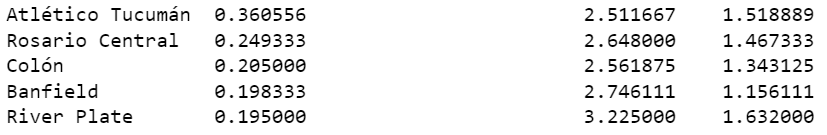


Рис. 5.2: 5 команд с наибольшим средним количеством голов за 90 минут.

Предположим, что мы хотим предсказать количество голов на основе следующих показателей:  
features = ['GamesPlayed', 'MeanSucAtt/90', 'MeanShots/90', 'Assists/90', 'xG/90']  
  
Создаем переменные X и y:  
X = df[features]  
y = df['Goals']  
  
Разбиваем данные на обучающую и тестовую выборки:  
X\_train, X\_test, y\_train, y\_test = train\_test\_split(X, y, test\_size=0.2, random\_state=42)  
  
Создаем и обучаем модель:  
model = LinearRegression()  
model.fit(X\_train, y\_train)  
  
Делаем прогнозы на тестовой выборке:  
predictions = model.predict(X\_test)  
  
Вычисляем среднеквадратическую ошибку:  
mse = mean\_squared\_error(y\_test, predictions)

**Mean Squared Error: 0.4930805746477233**

Можно разработать модель классификации, которая предсказывает, принадлежит ли игрок к определенной команде, основываясь на его статистических показателях. Это может быть полезно для выявления ключевых характеристик, которые делают игроков подходящими для определенной команды.

Помимо этого, модель может помочь в управлении командой, позволяя вам предсказать, какие игроки могут быть подходящими для перехода в другую команду, основываясь на их статистических показателях.

Используем градиентный бустинг с помощью библиотеки XGBoost:

import xgboost as xgb  
  
Создаем и обучаем модель:  
model = xgb.XGBClassifier(use\_label\_encoder=False, eval\_metric='mlogloss')  
model.fit(X\_train, y\_train)  
  
Делаем прогнозы на тестовой выборке:  
predictions = model.predict(X\_test)  
Accuracy: 0.7209302325581395

# 6 Заключение

В рамках данного курсового проекта было проведено статистическое исследование футбольных лиг Англии и Аргентины с использованием анализа данных и методов машинного обучения. Целью исследования было получить глубокое понимание футбольных лиг, их команд и игроков, а также выявить интересные зависимости и закономерности.

В результате исследования были получены ценные знания о футбольных лигах Англии и Аргентины. Анализ данных позволил нам лучше понять факторы, влияющие на успех команд и игроков, а также выявить интересные зависимости и закономерности. Эти результаты могут быть полезными для тренеров, скаутов и аналитиков, работающих в футбольной индустрии. Также качество предсказаний в лиге Аргентины оказались лучше, чем в лиге Англии, потому что в ней было намного больше различных параметров. Также в процессе исследования выяснилось, что команды лиги Аргентины в среднем забивают больше голов, чем в лиге Англии из-за большей ориентированности команд на атаке, нежели на защите.

В заключение, данное статистическое исследование предоставляет ценную информацию о футбольных лигах Англии и Аргентины. Оно подтверждает важность анализа данных и использование методов машинного обучения в футбольной аналитике. Результаты исследования могут быть использованы для принятия решений в футбольных клубах, для разработки стратегий и тактик, а также для предсказания успеха команд и игроков.

7 Список Литературы

* 1. [“Машинное обучение” курсна Stepik от](https://stepik.org/course/8057/promo#toc) [Артёма Шевлякова](https://stepik.org/users/41665007)
  2. [“Анализ данных в спорте”](https://plucky-growth-d29.notion.site/9853f4e4506747d7ac874408c0ab499a)
  3. [“Who Will Score? A Machine Learning Approach to Supporting Football Team Building and Transfers”](https://www.researchgate.net/publication/348384792_Who_Will_Score_A_Machine_Learning_Approach_to_Supporting_Football_Team_Building_and_Transfers)
  4. [5 Important Uses Of Data Analytics In Football Author : Aniruddh S Kumar](https://madaboutsports.in/5-important-facts-data-analytics-in-football/#1_Performance_Analysis)