**Introduction:**

Soccer is a team sport played between two teams. It is a well known and loved sport, with approximately 3.5 billion fans around the world. But as much as people enjoy sitting in the crowd, at their favorite bars or at home and watching their favorite team plays, there is another aspect to this sport which makes the games even more interesting for their viewers – and that’s betting.

In general, sports betting refers to attempts at predicting sports results, while placing a wager on the outcome. In soccer, there are many aspects of betting – one can bet on simple outcomes in the game, such as the final result of the game (winning team), or the exact result (amount of points per team). However, betting can be much more complicated as one can also try and predict many other outcomes, such as number of “offsides”, number of “corners”, first team to achieve a “corner”, etc.

The possibilities for the gamblers are endless, but the one that gains the most, statistically, is the betting company. There are many soccer betting companies out there, aiming to increase their income. The main way in increasing the company’s income is by setting the betting odds in the optimal way.

The betting odds can be seen as a function of the likelihood of the outcome – each outcome receives a number, which would increase as the likelihood for that outcome decreases. For example, in a simple case, assume the value for team A’s win was set to and a gambler placed a bet of dollars on that outcome. If this outcome happened, the gambler would gain (after “losing” the they bet) dollars. Therefore, if the company set to be very low, they would lose less money to that gambler.

The questions we aim to solve:

So, the question is, how does the company choose the optimal betting odds for a game? Many things can affect a game’s outcome, such as the teams’ levels, the players’ mood, the weather, etc.

In order to understand how to answer the question above, we try to solve the following questions:  **How does the outcome of a match affect the team’s outcome in their following match?  
How does the outcome of a match affect the betting odds in the team’s following match?**

In this project we will use the different methods we learned in the course and calculate the ATE. By using matching and learners, and some more statistical hypotheses testing, we plan to answer this question.

**The Data:**

The data we use in this project is taken from Kaggle. It is a database, composed of multiple tables, but the tables we will be using are:

* “Match” – a table of almost 26,000 records of matches, including the participating teams, dates, outcomes and betting odds given to each team by multiple betting companies.
* “Player Attributes” – A table that describes each player, including some statistics and ratings.
* “Team” – A table with information about different teams.

The full database can be found [here](https://www.kaggle.com/datasets/hugomathien/soccer?select=database.sqlite).

**Exploratory Data Analysis:**

COMPLETE

**Data Processing:**

Now we wish to process the data to get a usable data for our research.

We start by creating two datasets:

**Match dataset**

Using “Match” table in the database, we create a dataset that contains information about each match. The information can be split to two groups:

1. General information: season, date, match id and participating teams ids.
2. Match information:
   1. Amount of goals per team.
   2. Goal difference (“home” team goals minus “away” team goals).
   3. Average betting odds given to each outcome (home team wins, away team wins or draw).
   4. Details about player rating per team (changes between different matches):  
      Minimum, maximum, average, std and median.

**Season dataset**

This dataset contains seasonal information about all teams, which can also be split into two groups:

1. General information: season and team id.
2. Team information:
   1. Amount of wins.
   2. Amount of draws.
   3. Total league points (3 point per win and 1 point per draw).
   4. Amount of goals in the entire season.

We take these two datasets and merge them, so for each match we have the match’s information and the team’s seasonal information.

Now, let’s recall that we want to find if a match’s result affects the next match’s result or the next match’s betting odds. Therefore, for each match, we need to hold information about the following match of the teams.

Therefore we create the final dataset, which holds for each match all the information we discussed earlier, and also holds the same information for the next match of the team.

**The Methods:**

First we want to understand the effect of a match’s outcome on the next match outcome, and therefore we set “prev\_win” variable as the treatment (T) and “next\_win” variable as the outcome (Y). All other variables are the control variables (X).

We use the methods taught in class in order to calculate ATT.

1. IPW

At first we calculate the propensity scores: We train a logistic regression model on X with T as the labels. Thus, for each combination of control variables, we get a probability for getting treatment T.

We use the propensity score to calculate ATT using IPW.

* IPW formula:   
  When is the propensity score of .

We get (CHECK).

1. S-Learner

Here we fit a linear regression model on X (along with T) with Y as labels (predict next win given the control parameters). We fit the model on all the data, and then use it to predict the treated data (teams that lost their previous games) and the same data, this time with treatment 0 (predict the Y value in case they weren’t treated).   
We then calculate the ATT using this formula:

We get .

1. T-Learner

Here we train two linear regression models: One on the treated data () and another one on the untreated data ). We use only the control variables to train the model (as the variable is the same for all data for each model anyway). We use each model to predict the value for the treated group, and calculate the ATT using this formula:

We get .

1. Matching