Class Exercise 2

NFL data

- For avid football fans and aspiring analysts, a wealth of statistical information is available online to dissect every play, player, and team performance.
- Whether you're building complex predictive models or simply want to settle a debate with friends, these websites offer a treasure trove of NFL data.
- **Pro-Football-Reference:** A favorite among sports researchers and analysts, this site is a veritable encyclopedia of NFL history.
- What if we can predict whether a team will have a winning season based on their regular season statistics.
- **Goal:** To predict if an NFL team will have a winning season.
- We'll use Pro-Football-Reference, as its website structure uses clean HTML <table> elements, which is perfect for the "pandas.read_html() "function. We will scrape the main standings page from the recently completed 2023 season, which contains both team records and key statistics.

Exercise 2 – Data Integration with read_html()

Build a model that can predict a winning team in NFL

1. https://www.pro-football-reference.com/years/2023/ - this website has some tables which you can checkout.

2. There are two tables of the conferences that we will use - The first table is AFC, the second is NFC

3. Tasks to do:

   - Use the given url to scrape the tables data

   - Combine and clean the two conference tables and convert stat columns to numeric types for modeling ('PF', 'PA', 'PD', 'SoS') and these are described as "points for, points against, points differential, strength of schedule"

   - Remove 'W-L% ' column. The main reason is to prevent a problem called data leakage. In machine learning, data leakage happens when information from your target variable (the thing you're trying to predict) accidentally "leaks" into your feature variables (the data you're using to make the prediction).

   - Create a label column suitable for classification (1s and 0s) – think of a logical way of doing this as it is not readily given.

   - Create an SQL database called 'NFL' and store the cleaned data in a table called 'stats'. Use the pandas 'to_sql' function.

   - Stats columns will be your features, so select them and drop the rest.

   - Use a classification model to predict win or lose in a season

Exercise 2 – contd …

4. Predict on new data of a hypothetical team ‘Atlanta Vipers’ using the model. Add this data and the prediction as a new row in the database table you have created.

   - 'PF': [465],
   - 'PA': [380],
   - 'PD': [85],
   - 'SoS': [1.5]

- Finally, deploy it using streamlit.