The goal was to use previous season and current season data to predict the playoff outcome for the 2024-2025 NFL Playoffs.

Initially, I started with the current 2024-2025 season data. Although accuracy appeared to be high, I included the previous two regular seasons (2022-2023, 2023-2024) data in order to provide additional data to help train the model.

I began by loading and creating the CSV’s, once loaded I created a DataFrame from the three year stats. The data was cleaned, converting the results of home and away games into win and loss totals. Using provided data, calculations for win percentage, weather averages, and per quarter score statistics were added to help train the model.

The model was trained using Random Forest Classifier with a 92.68% accuracy. The model used 20 features and 816 rows of data. One limitation to this data is lack of player performance. It also does not track injuries, illnesses or other reasons players are missing games. Recent game performance and coaching changes are also not being considered. The focus of this model was on weather impacts, scores of games, quarter scores, and home and away records.

Feature Importances show the highest consideration for the total points at the end of the game either home or away (away\_cumulative\_full & home\_cumulative\_full) at 15.97% and 15.72%. When running the model for one season, the 2nd quarter point differential (quarter\_2\_point\_diff) was more important than the other quarter differentials at almost 5%, with the three seasons of data the 4th quarter point differential was higher. I would like to look at this further given the right data set.

A Correlation Matrix was made using the weather temperature, humidity, and home win probability data. I found humidity has a weak relationship with the win probability. There is a slight positive correlation between weather temp and home win probability. Due to the weak correlation, these features could be dropped when running future models.

When making the predictions for playoff winners, both models made the same predictions, this could be due to the high accuracy rating of both models using Random Forest (over 90%).