**Methods**

RQ1 - What type of factors most heavily influence player injuries in the National Football League?

Data Organization

There are many factors that can lead to an injury in the National Football League. We are attempting to decide the leading factors that lead to these injuries in games. We are using a combination of two datasets to get a play-by-play description of each play, who was injured, and what position they play. To prepare our data we first had to determine all the columns that would most likely be useful for our research. The data set had 372 different columns in which we found 47 to be potentially useful for our research. These 47 columns were separated as potentially useful but are not necessarily used throughout the analysis. Using the ‘play\_by\_play’ dataset, in the ‘desc’ column, we are able to extract any time a player was injured from 2019-2023. Then merging that dataset with the ‘player\_posistion19to23’ we can get all the information on the player. After removing the unnecessary columns and adding the player information, our dataset is ready for use.

There are several factors that can be looked at for the probability of a player being injured on a given play. However, the key factors being looked at are weather, stadium conditions, turf, player position, and player usage. These categories were most represented in the various literature reviews that were conducted. The set of nearly 247,000 plays was dissected find only the plays that resulted in a player injury. K- means clustering was used to cluster injuries based on position vs play type. The plays were represented by 5 clusters and included; No-plays, passes, runs, punts, and kickoffs. This allowed for the breakdown of injuries for each position based on the play that was taking place. The no play cluster is determined to be a play that happened in real time but was called as a ‘no-play’ due to stoppage, penalties or other situations.

Modeling Techniques:

* Logistic Regression
  + - Coefficients are calculated for each predictor variable to determine how each variable changes the log-odds of an injury during a play.
* Decision Tree
  + Figure out the tree ‘branch’ that can determine a predicted class. The decision tree attempts to find variable values d for the data that best split the predicted classes. This determines the most accurate class predictors for each class.

RQ2 – Can we predict the probability that an injury will occur under certain conditions?

Maximum Likelihood Estimation

Our predictions for which positions are most likely to suffer an injury given compounding factors during a game relies on multiple variables that together, contextualize a specific play where the injury happened. These variables give information on when the injury happened during a drive. Key variables give information on things like which down and how many net yards were gained at that time; what the play type was, run or pass; and what the player usage was by using their total snap count at that point in the game as a proxy.

We’ll divide up just the injury observations into an 80:20 train/test split to ensure proper cross-validation. We’ll run the model on the train data and test it on the test data. Once we have our results, we’ll plug them into a confusion matrix and find the accuracy of the model. The confusion matrix will provide us with the sensitivity of the model, which tells how exact the model is at correctly predicting a player position. We’ll have eight different sensitivities, one for each position group.

Modeling Techniques:

* Naïve Bayes
  + Takes conditional probabilities for each variable in the target category, conditioned on all the variables we provide the model to train on. Using these conditional probabilities, it gives maximum likelihood estimates to make predictions.

RQ3 – What types of Injuries can be predicted in future games?

Based on all the factors that were looked at it is difficult to determine what factors overall feed into injuries. It was hypothesized that turf and weather were going to be major factors in whether an injury occurred or not. However, after our EDA it shows that there is slight difference in these categories. We realized that there may be other factors that maybe are overlooked that contribute to higher levels of injuries. Using a random tree to figure out if we can accurately predict injuries based on factors using random noise. We were able correctly predict non-injury plays with a 99% accuracy while only getting an accuracy score of 67% on injury predictions. The accuracy score was 99% indicating a high rate of correctly predicted outcomes across both classes.

Gradient boosting was used to predict if weather was a major factor in influencing injuries. We evaluated the model's performance using classification metrics such as precision, recall, and F1-score. Although the outcome was a high accuracy for predicting plays that an injury did not occur, it was terrible at predicting plays where an injury would occur. This is likely due to the dataset being so imbalanced. We tried combating this by adjusting class weights, but still no luck. This highlighted one of the many challenges of trying to predict such a relatively rare event in such a largely imbalanced dataset.

* Random Forest
  + Creates an ensemble of decision trees in which each tree is trained with specific random noise. Combines the output of multiple decision trees to form a single result for predicting injuries.
* Gradient Boosting:
  + We chose Gradient Boosting to predict injuries based on weather conditions due to its effectiveness in handling complex datasets.