Jackson Boyle

5/2/2025

DS 325 Prof. Roth

Final Project – College Basketball ACL Injuries

Introduction:

Injuries are a major obstacle in the success of both athletes and teams across all sports, including college basketball. As a sports fan and athlete myself, one of the most devastating injuries to see a player endure is an ACL or Anterior Cruciate Ligament injury. These injuries can completely derail a career and account for 8% of all game injuries in NCAAW basketball (Kitman Labs, 2016). So, being able to determine what factors influence ACL injuries among college basketball players is crucial for the financial and physical health of players as well as their teams. I think that the amount of games played would be the biggest determinant in ACL injuries. To do this, I ran two correlation tests between actual ACL injury statistics and factors like rest between games and an already calculated fatigue score. After finding the most correlated features, I ran a linear regression to determine if a few of these features could predict whether or not a player actually had an ACL injury.

Methods:

The dataset used in this analysis is called "Athlete Injury and Performance Dataset" and was designed to analyze how scheduling affected ACL injury rates on college basketball. It included several features such as position, age, and height which I excluded from the analysis as these are factors which a coach or manager could not influence or change. I wanted to focus on two types of features in the dataset, raw scheduling values and pre-calculated player scores to figure out what factors influence injury or if it even can be predicted. This does assume that all data was reported accurately and that subjective features like fatigue score were determined relatively evenly across athletes.

The key steps to run this analysis included:

1. Filtering out unwanted features and splitting the features into raw values and calculated scores
2. Running a correlation test for the raw values and injury indicator as well as the calculated scores and injury indicator
3. Choosing the features from both that had the highest correlation to the injury indicator (Recovery_Days_Per_Week, Fatigue_Score, and Load_Balance_Score)
4. Running a linear regression model on these features to predict the injury indicator.
5. Running a decision tree model on these features to predict the injury indicator. I chose to run two different models as they both have different ways of making predictions and could have interesting distinctions.

Results:

After running the tests on the correlations between features and the injury indicator, I found that many features had no or very little correlation. In the raw values correlation test, I found that recovery days per week (-.26) was the only feature with a correlation greater than .2 (Figure X). In the calculated scores correlation test, I found that fatigue score, a subjective fatigue level on a scale of 1 (low) to 10 (high), (.29) and load balance score, a calculated score (0-100) where a higher value represents a better balance between training load and recovery (-.49) were the only two features with a correlation greater than .2 (Figure Y). Both the decision tree and logistic regression had high accuracy scores in predicting injury indicator. For the decision tree, 155 of the 160 training data were accurately predicted with only 5 injuries bring predicted as non-injuries (Figure 1). The accuracy score was 0.97 while the macro average recall was .77.

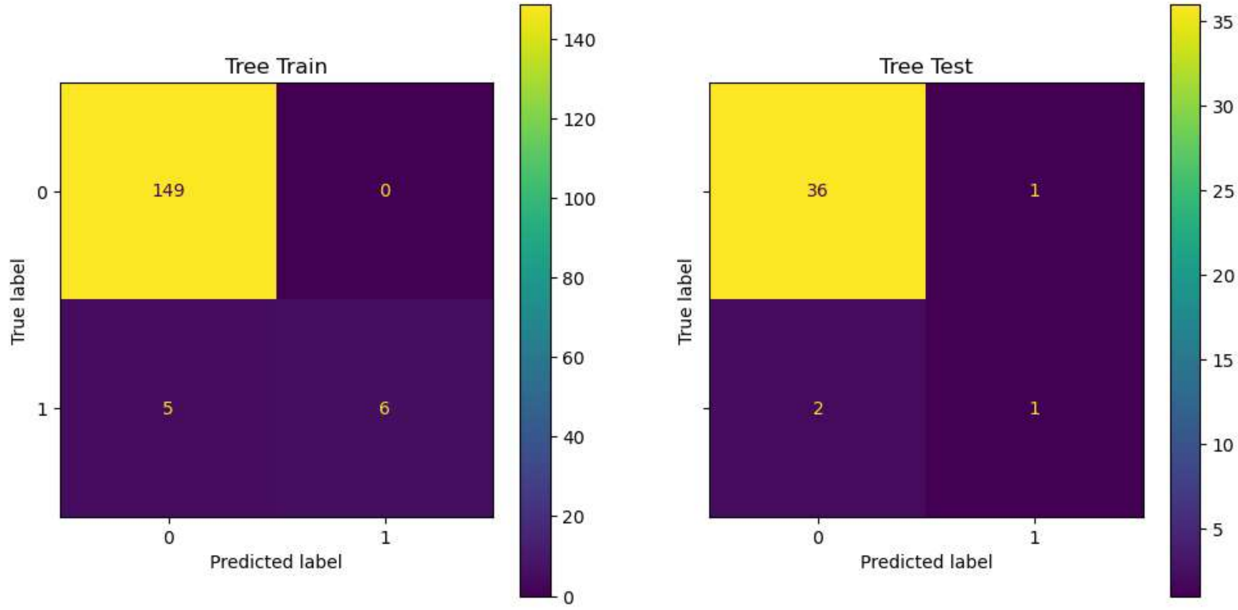


Figure 1. A confusion matrix showing the results of the decision tree model for predicting injury indicator. 0 represents non-injured athletes while 1 represents those with an ACL injury.

For the logistic regression, 153 of the 160 training data were accurately predicted with 5 injuries being predicted as non-injuries and 2 non-injuries bring predicted as injuries (Figure 2). The accuracy was .96 while the macro average recall was .77. So, the decision tree was slightly more accurate than the logistic regression.

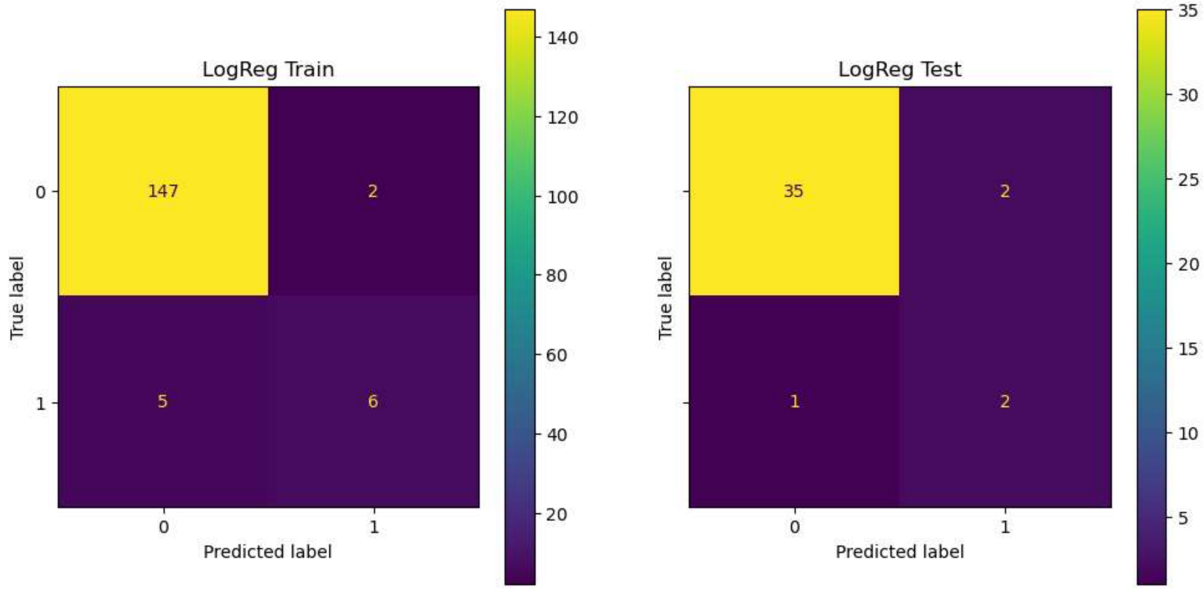


Figure 2. A confusion matrix showing the results of the logistic regression model for predicting injury indicator. 0 represents non-injured athletes while 1 represents those with an ACL injury.

Discussion:

There are several interesting insights in these results that could influence how coaches manage their players' schedules and injuries. First, the fact that recovery days per week was the only really significant raw value predictor of ACL injury, and that it was negatively correlated, shows that by incorporating more recovery days, potentially even with longer days could be more effective in reducing injury. It also seems as though reducing player fatigue and improving the balance between recovery and training load could be effective at reducing ACL injuries. This aligns with what past medical studies have found which say to avoid playing when you are overly fatigued (NYU Langone Health, 2021) and that, for basketball specifically, rest is a crucial part of the prevention of and recovery from ACL injuries (Crystal, 2024).

While both of the models had very high accuracy scores, it is hard to say whether either is necessarily a perfect tool for predicting ACL injuries. The fact that there were so few athletes in the dataset which actually had experienced an ACL injury means that my models did not get a ton of chances to predict non-injuries. I don't think this invalidates my results or makes them any less significant but does highlight how future research could be interesting. Adding more athletes into the dataset would allow for a more complete and accurate analysis, making it more broadly applicable. Another interesting avenue of future research would be to compare to professional NBA or WNBA players or college athletes from other sports to see how the results are similar and different.

Overall, I feel as though finding the features with the greatest correlation was very successful and has interesting implications for coaches and players. While the models are also interesting, I think they have more limited applicability and could benefit from further research or greater dataset size.

Citations:

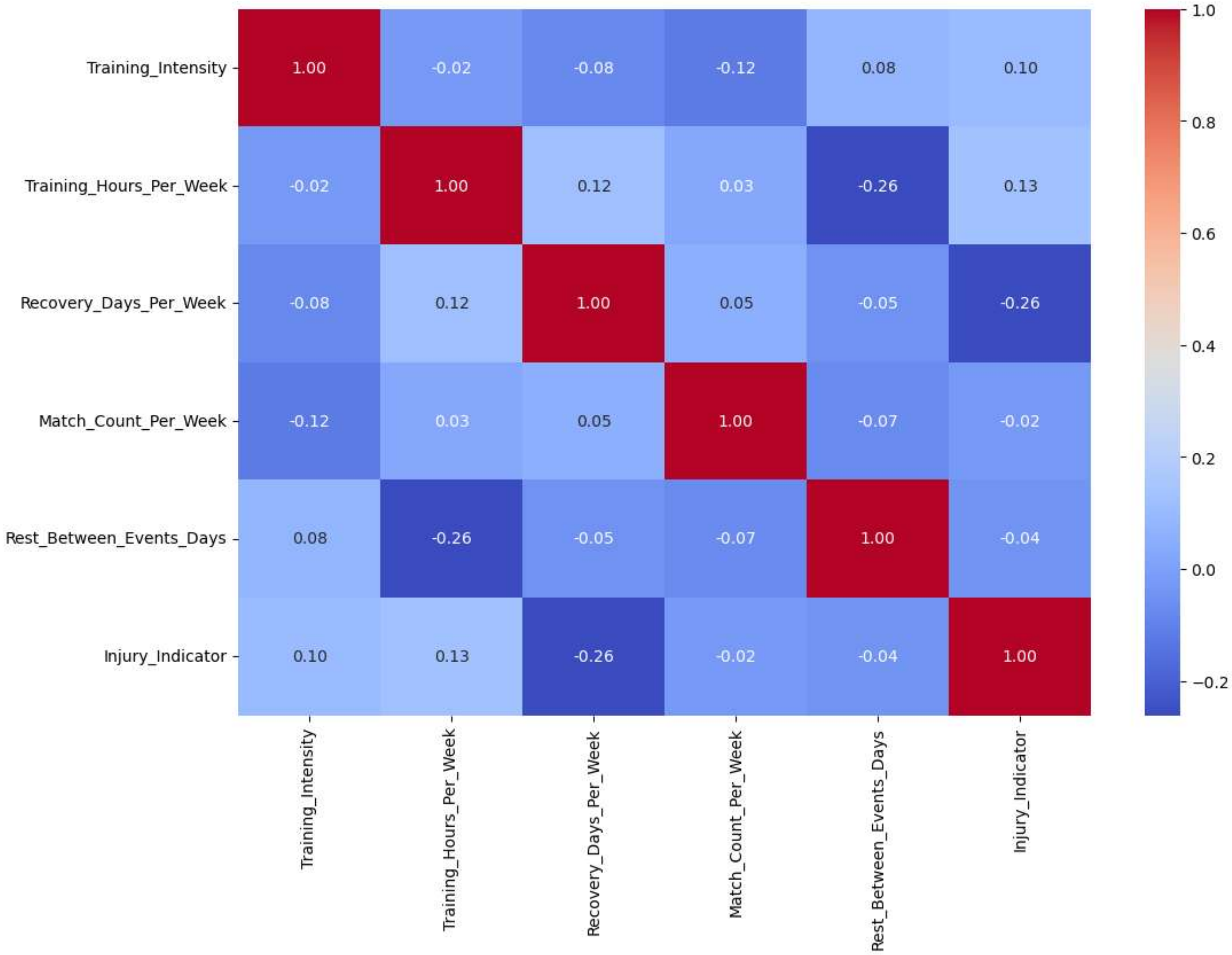
Crystal, Y. (2024, March 13). *ACL Injuries In Basketball: Causes, Treatments & Preventions*. Genourob Knee Laxity Arthrometers. https://arthrometer.com/acl-injuries-in-basketball/#elementor-toc__heading-anchor-9

Kitman Labs. (2016, April 20). *Examining NCAA basketball and it's most common injuries*. Kitman Labs. https://www.kitmanlabs.com/blog/examining-ncaa-basketball-and-its-most-common-injuries/

NYU Langone Health. (2021, March 26). *Five Ways to Prevent Anterior Cruciate Ligament Injuries*. NYU Langone News. https://nyulangone.org/news/five-ways-prevent-anterior-cruciate-ligament-injuries

Ziya. (2024). *Athlete Injury and Performance Dataset*. Kaggle.com. https://www.kaggle.com/datasets/ziya07/athlete-injury-and-performance-dataset?resource=download

Figures:



	Training_Intensity	Training_Hours_Per_Week	Recovery_Days_Per_Week	Match_Count_Per_Week	Rest_Between_Events_Days	Injury_Indicator
Training_Intensity	1.00	-0.02	-0.08	-0.12	0.08	0.10
Training_Hours_Per_Week	-0.02	1.00	0.12	0.03	-0.26	0.13
Recovery_Days_Per_Week	-0.08	0.12	1.00	0.05	-0.05	-0.26
Match_Count_Per_Week	-0.12	0.03	0.05	1.00	-0.07	-0.02
Rest_Between_Events_Days	0.08	-0.26	-0.05	-0.07	1.00	-0.04
Injury_Indicator	0.10	0.13	-0.26	-0.02	-0.04	1.00

Figure 3. Correlation heatmap showing the correlation between each of the raw value features on each axis with a focus on those correlated with injury indicator

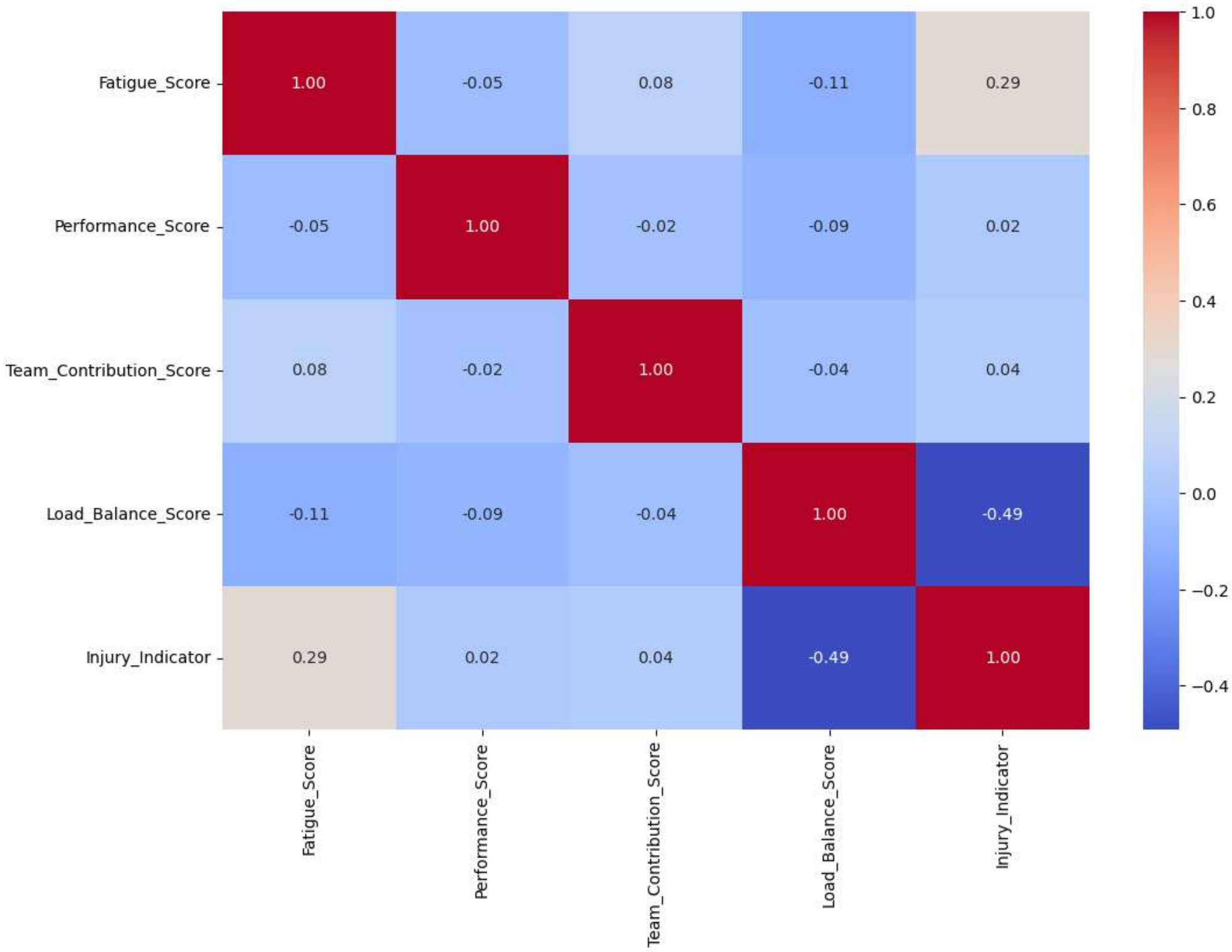


Figure 4. Correlation heatmap showing the correlation between each of the calculated score features on each axis with a focus on those correlated with injury indicator