Zoe Want
DS 325
Professor Roth
May 2, 2025

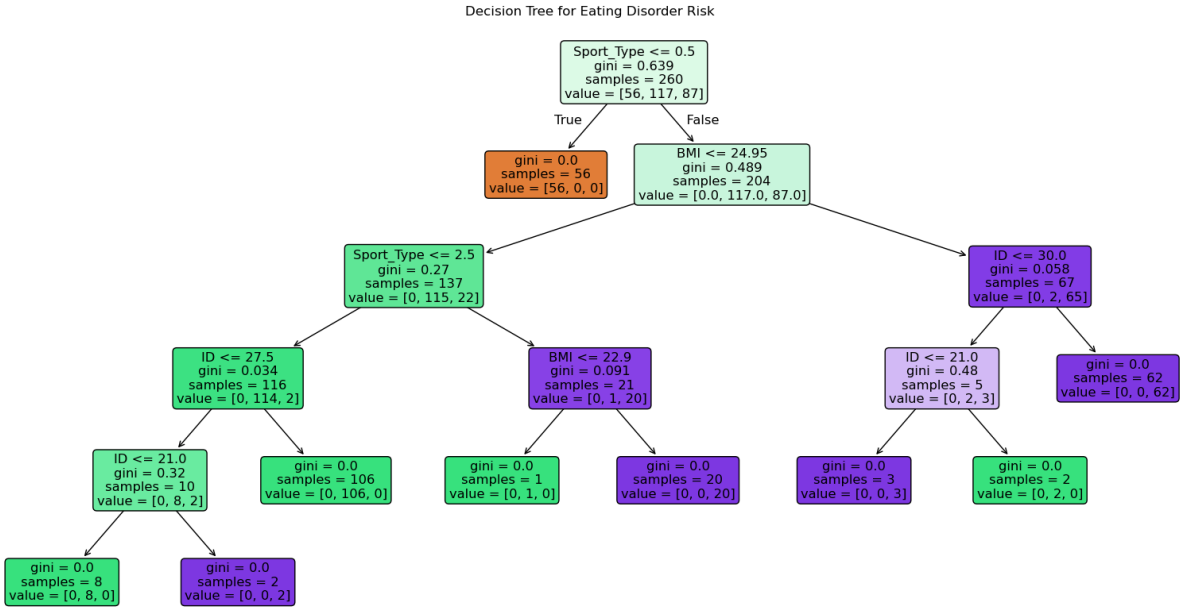
Predicting Eating Disorder Risk Using a Decision Tree Classification Model

In recent years, NCAA athletes have spoken out about the pressure to maintain specific body standards, often leading to disordered eating. While many previous studies have examined prevalence rates, few have quantitatively identified what factors put athletes most at risk. This project aims to identify the factors that contribute to an NCAA athlete, whether that be gender, the sport played, time spent exercising, or other factors that will be discussed later on. I predict that BMI and time spent exercising will be the factors that contribute most to the athletes identified as at risk for an eating disorder. I curated a data set with AI, using data from Bratland-Sanda and Sundgot-Borgen's academic study on *Eating disorders in athletes: Overview of prevalence, risk factors and recommendations for prevention and treatment* combined with raw data sets from The Art of Statistics. Using the generated data set, I applied a Classification model, specifically a decision tree, to determine if a collegiate athlete is at risk for developing an eating disorder. From there, I also tested the importance of each feature to identify what factor most closely correlates to eating disorder risk. The results of the test were that sport type, BMI, and ID held the most importance in predicting eating disorder risk.

As briefly mentioned in the introduction, the model used multiple datasets. First data, from Bratland-Sanda and Sundgot-Borgen's academic study on *Eating disorders in athletes: Overview of prevalence, risk factors and recommendations for prevention and treatment.* From this academic study, I used the Collegiate female athletes data set by Greenleaf et al. (2009), as well as the Collegiate male athletes data set by Petrie et al. (2008). I synthesized these data sets, with the raw data sets on College Athletes, Eating Disorders, Anorexia, Exercise Hours, and College Female Athletes from The Art of Statistics. Using ChatGPT, I was able to synthesize all of these data sets to create a final data set. The final data set included 325 athletes, and the included features consisted of: gender, sport type, eating disorder risk, BMI, age, division/region, years played, hours of exercise per week, and dieting history. In respect to cleaning and preprocessing, I cleaned the data prior to importing it using a text editor. In order to determine if a collegiate athlete is at risk for developing an eating disorder, a classification model was used. More specifically, a decision tree along with a confusion matrix. I selected this model because the target variable, eating disorder risk, is discrete not continuous. The results are either: none, low, or high. As I am classifying athletes to one of these three groups, classification is the preferred method over a regression. I compared this model to a k-nearest neighbors model, as I thought it would be beneficial to see which performed better. However, the accuracy was not as high in the KNN model, and therefore I stuck with the decision tree. After selecting the

appropriate model, I then went through the following steps to get my final decision tree and confusion matrix:

1. Imported the data set
    a. I cleaned the data in a text editor prior to importing it for efficiency
2. Printed the data set to ensure there were no missing variables
3. Encoded the data, as not all of it was numerical data
4. Used the ordinal encoder for my ordinal variables
    a. Ordinal variables: eating disorder risk
5. Used the one hot encoder for my categorical variables
    a. Categorical variables: gender, sport type, and division/region
6. Refit the new encoded data
7. Defined the target variable as eating disorder risk
8. Trained and tested the data
9. Fit the data to a decision tree classifier, which can be seen in Figure 1 below:



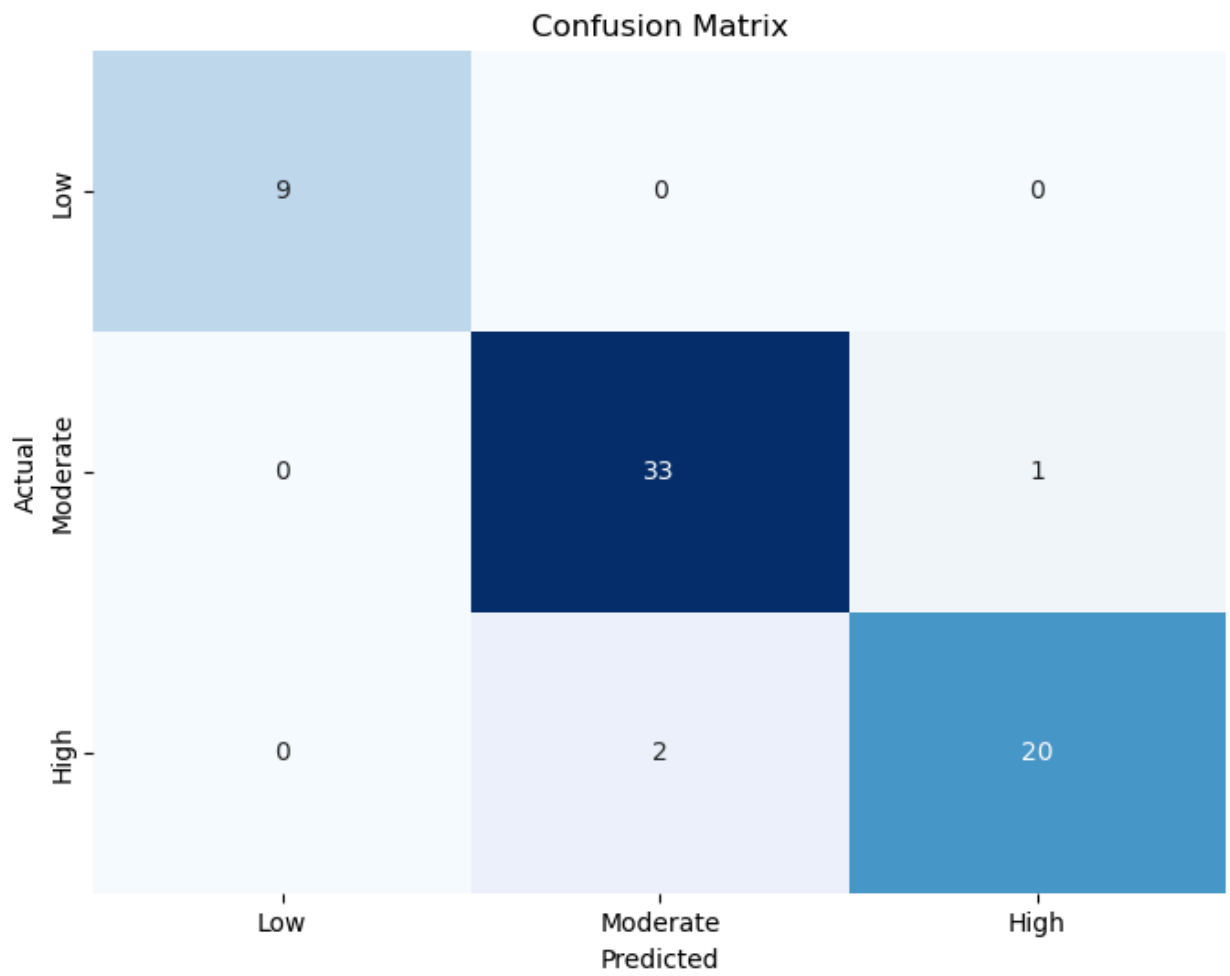
10. Printed the tree rules in order to better conceptualize the decision tree, and can be seen below in Figure 2:

```
|--- Sport_Type <= 0.50
|   |--- class: 0
|--- Sport_Type >  0.50
|   |--- BMI <= 24.95
|   |   |--- Sport_Type <= 2.50
|   |   |   |--- ID <= 27.50
|   |   |   |   |--- ID <= 21.00
|   |   |   |   |   |--- class: 1
|   |   |   |   |--- ID >  21.00
|   |   |   |   |   |--- class: 2
|   |   |   |--- ID >  27.50
|   |   |   |   |--- class: 1
|   |   |--- Sport_Type >  2.50
|   |   |   |--- BMI <= 22.90
|   |   |   |   |--- class: 1
|   |   |   |--- BMI >  22.90
|   |   |   |   |--- class: 2
|   |--- BMI >  24.95
|   |   |--- ID <= 30.00
|   |   |   |--- ID <= 21.00
|   |   |   |   |--- class: 2
|   |   |   |--- ID >  21.00
|   |   |   |   |--- class: 1
|   |   |--- ID >  30.00
|   |   |   |--- class: 2
```

a.

11. Accuracy, precision, and recall scores were calculated
12. Feature importance was printed, to determine which helped identify eating disorder risk best
13. A confusion matrix was created to visualize how many eating disorder risk athletes were correctly identified, which can be seen in Figure 3 later on

In order to assess the decision tree, as previously stated, I calculated precision, recall, accuracy, and also curated a confusion matrix. The decision tree had a 0.95 score for precision, 0.95 for recall, and 0.95 for accuracy. For precision, this means when the model predicted at risk for an eating disorder, it was correct 95% of the time. It is also correlated with a low false positive rate. For recall, approximately 95% of predicted true positives, predicted at risk for eating disorder, when they were actually at risk, were captured. This is correlated to a low false negative rate. Lastly, for accuracy, the 0.95 correlates to 95% of the total predictions being correct, both predicted at risk and not at risk for an eating disorder. Lastly, to visualize and interpret the true positives, true negatives, false positives, and false negatives more easily I constructed a confusion matrix. That can be pictured below in Figure 3.



As you can see, although there were 325 athletes only 65 of the participants were included. This can be due to multiple things, potentially missing variables or the model excluded athletes for another reason. More importantly, it is important to note that all of the low at risk for an eating disorder were identified correctly, of 35 of the moderate at risk only one was misclassified as high at risk, and of the high at risk 20 were correctly predicted, while two were misclassified as moderately at risk.

As mentioned in the previous section, it is safe to conclude that trainers and coaches in the NCAA can look at factors to determine if athletes are at risk for developing an eating disorder. Superiors can advise weekly or monthly check ins to manage how much athletes are training, their BMI, etc. By doing this it will help mitigate the number of athletes dealing with eating issues. During my project, I ran into an issue with data. There were not an abundant number of raw data sets on my research question, however using ChatGPT I was able to curate a data set using a combination of multiple raw data sets. One thing I did not expect in my research was for many factors to be omitted. If I could re-run my project I would try to correct this or get to the root of why this was happening, as I do find each of these as strong indicators of a potential eating disorder. Concluding, I hope this project is a stepping stone to helping schools

ensure their athletes do not develop eating disorders. Additionally, I think it sparks additional research to what factors best help reverse these disorders and help athletes recover.

Citations

Art of Stat. (n.d.). *General Social Survey Data*. https://artofstat.com/datasets

Bratland-Sanda, S., & Sundgot-Borgen, J. (2012, November 13). Eating disorders in athletes: Overview of prevalence, risk factors and recommendations for prevention and treatment - bratland‐sanda - 2013 - european journal of sport science - wiley online library. https://onlinelibrary.wiley.com/doi/10.1080/17461391.2012.740504

ChatGPT. (2025). *Dataset compiled on youth physical activity and sport development trends based on academic literature*[Unpublished dataset]. Created using data from Vaeyens et al. (2013), Keating et al. (2005), and Weiss & Wiese-Bjornstal (2009).

Petrie, T. A., Greenleaf, C., Reel, J., & Carter, J. (2008). Prevalence of eating disorders and disordered eating behaviors among male collegiate athletes. Psychology of Men & Masculinity, 9(4), 267–277. https://doi.org/10.1037/a0013178