1) Team name
Mention your team name.

AACK

(0 points)

2) Member names
Mention the names of your team members.

(0 points)

Catherine Erickson, Karrine Denisova, Ally Bardas, Annabel Skubisz

3) Link to the GitHub repository
Share the link of the team's project repository on GitHub.

Also, put the link of your project's GitHub repository here.

We believe there is no harm in having other teams view your GitHub repository. However, if you don't want anyone to see your team's work, you may make the repository *Private* and add your instructor and graduate TA as *Collaborators* in it.

(0 points)

https://github.com/kdnu24/STAT303-2--project.git

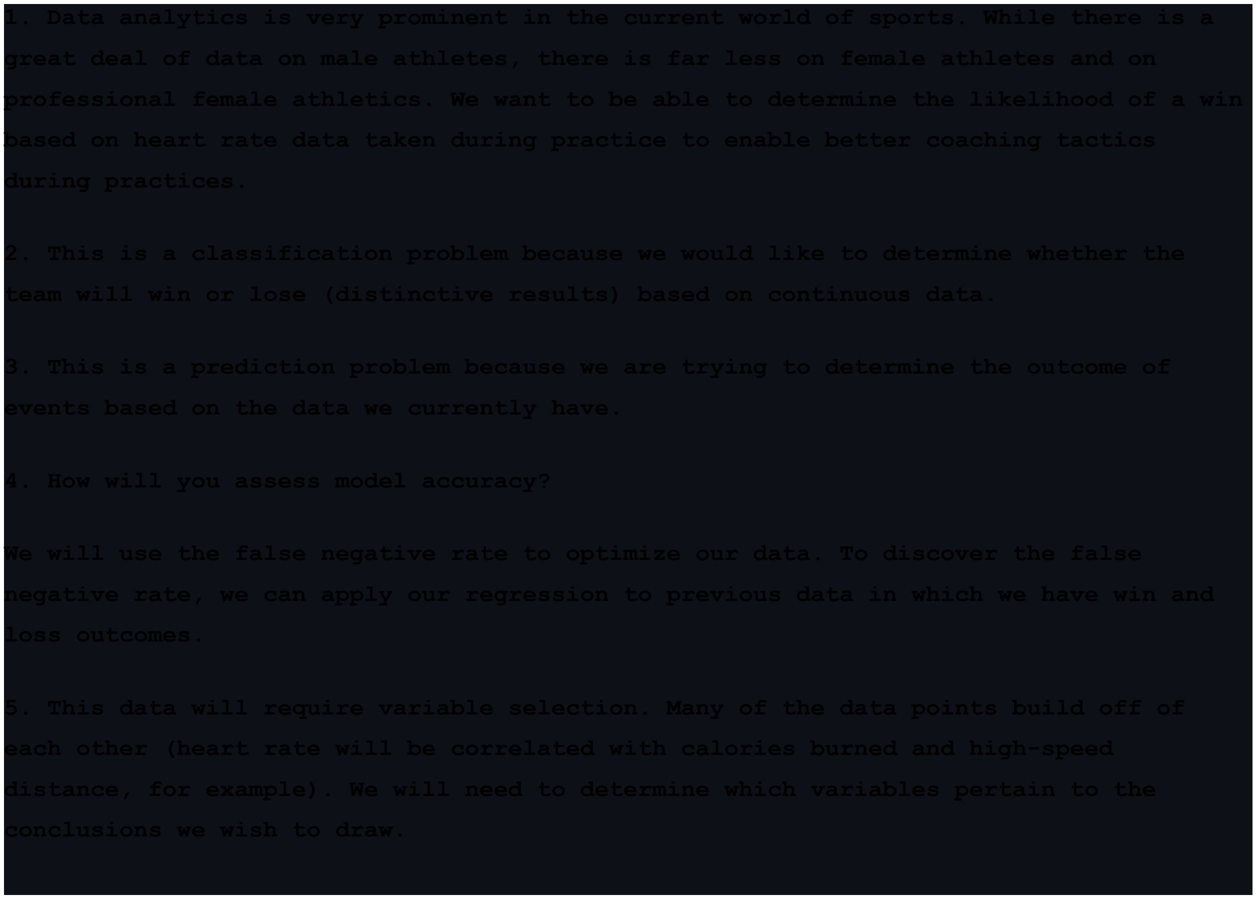
4) Topic
Mention the topic of your course project.

We are going to analyze the heart rate data from one of the Varsity teams at Northwestern. We are going to look at if the data can predict game wins and losses.
(0.25 points)

5) Problem statement
Explain the problem statement. The problem statement must include:

1. **The problem.** Data analytics is very prominent in the current world of sports. While there is a great deal of data on male athletes, there is far less on female athletes and there is far less on professional female athletics. We want to be able to determine the likelihood of a win based on heart rate data taken during practice to enable better coaching tactics during practices.

2. **Is it a regression or classification problem or a combination of both?** This is a classification problem because we would like to determine whether the team will win or lose (distinctive results) based on continuous data.
3. **Is it an inference or prediction problem or a combination of both?** This is a prediction problem because we are trying to determine the outcome of events based on the data we currently have.
4. **How will you assess model accuracy?**
    - **If it is a classification problem, then which measure(s) will you optimize for your model – precision, recall, false negative rate (FNR), accuracy, ROC-AUC etc., and why?** We will use the false negative rate to optimize our data. To discover the false negative rate, we can apply our regression to previous data in which we have win and loss outcomes.
    - **If it is a regression problem, then which measure(s) will you optimize for your model – RMSE (Root mean squared error), MAE (mean absolute error), maximum absolute error etc., and why?**



1. Data analytics is very prominent in the current world of sports. While there is a great deal of data on male athletes, there is far less on female athletes and on professional female athletics. We want to be able to determine the likelihood of a win based on heart rate data taken during practice to enable better coaching tactics during practices.

2. This is a classification problem because we would like to determine whether the team will win or lose (distinctive results) based on continuous data.

3. This is a prediction problem because we are trying to determine the outcome of events based on the data we currently have.

4. How will you assess model accuracy?

We will use the false negative rate to optimize our data. To discover the false negative rate, we can apply our regression to previous data in which we have win and loss outcomes.

5. This data will require variable selection. Many of the data points build off of each other (heart rate will be correlated with calories burned and high-speed distance, for example). We will need to determine which variables pertain to the conclusions we wish to draw.

5. **What techniques do you think you may need to use to improve your model? If you have too many variables, some of which are correlated or**

collinear, you may need to do variable selection *(techniques for variable selection that you will learn later in the course - stepwise regression, lasso, ridge regression)*. If the variables do not have a linear relationship with the response, or if some of the modeling assumptions are not satisfied, you may need to transform the predictors and/or the response (variable transformation) to obtain a better fit. This data will certainly require variable selection. Many of the data points build off of each other (heart rate will clearly be correlated with calories burned and high speed distance).

(5 points)

6) Data sources
What data sources will you use, and how will the data help answer the questions? Explain. The strength and conditioning coach of the Northwestern Field Hockey team provided it. This data has all of the variables including heart rate, wins and losses, length of training, calories burned, distance traveled, away or home game, and many more variables.

If the data is open source, share the link of the data.

(1 point)

Not open source - from the athletic trainer

7) Stakeholders
Who are the stakeholders, and how will your project benefit them? Explain.

(1 point)

Our stakeholders are the Northwestern field hockey coaches. This project will help them determine the optimal conditions for practice in order to win games. By finding which variables have the highest correlation with wins, the coaches will be able to target these heart rates or speed or distance run etc during practice.

Additionally, sports sponsors and field hockey sponsors can benefit from the results of this data as it will lead to more games won with the deeper understanding of the predictors that contribute to the game.

Athletes themselves are stakeholders, because understanding training load and predictors to win allow athletes to peak at right time and not overwork in areas not necessary, reducing injury and overload.