The name of our project is the NEWMAC Score Predictor. Using historical NEWMAC soccer data (1980-2022), we will train a machine learning model to predict the results of future NEWMAC soccer games. The statistics used are the following but not limited to: goals, home/away, record in the season, rank in the conference, saves, and shots. We plan on implementing a Random Forest and/or Decision Tree for our specific machine learning model. A big part of our motivation for choosing this project is that we are all part of the Men's Soccer Team here at Clark, and have noticed first hand how statistics are having more of an impact on the beautiful game than ever before. To start we plan to implement a Clark centric approach, with the model focusing on results from the perspective of Clark. We believe that our place on the team gives us good insight into the importance of certain statistics and details. Ultimately the grand goal is to be able to implement this approach for other sports or schools, but to start we will focus on Clark Men's Soccer. Potentially we could pitch our model to our coach or even the NEWMAC. The two biggest challenges that we face come in data collection and implementation. One problem we've noticed is that a lot of the data is not uniform and not organized in an easily accessible way. In solving this we plan to input data into our .csv by manually handpicking what data we need from different locations and potentially even creating data points of our own, such as a historical program rating for all the teams in the NEWMAC. The implementation problem is more simple, but we may have difficulty implementing the different algorithms we plan to use in a timely and effective manner. We haven't entirely finalized our ideas for visualization and displaying results, but right now we are leading with an idea to display the scores over a season using a graph. For this graph we plan to include predictions from all of our data, predictions from data just from the last ten years, and actual results. The planned timeline for our project is below:

We've broken our plan into 4 parts - Data Collection/Processing, theory/planning implementation framework, Coding/Implementation, and Expansion.

10/13 - Finalized Initial Report

Data Collection/Processing

10/20 - All data collected and organized, ready for manual .csv inputting

10/27 - Finalized .csv with all data points, Midterm report done

Theory/planning implementation framework

11/3 - Have our approach finalized and planned out accounting for timing and difficulty. Also how we plan to divide the work amongst our group members

Coding/Implementation

11/10 - Coding, taking the framework from some of our sources and working to adjust it for our desired functionality.

11/17 - Coding, finalizing model, debugging, and tying up loose ends.

Expansion

11/27 – 12/1- Final Presentations. Anything to add to the project or polish with the model. Work on and finish presentation(Depending on timeline for the implementation we may not get to this part)

Below is the link to our sources. Many of the sources offer good insight into the theory behind our machine learning models, as well as a framework that we can use in our implementation.

https://docs.google.com/document/d/1lvqGVor_ilm5t5-u1ZKjeigb3y03XSAEEWxeQJd7rqw/edit