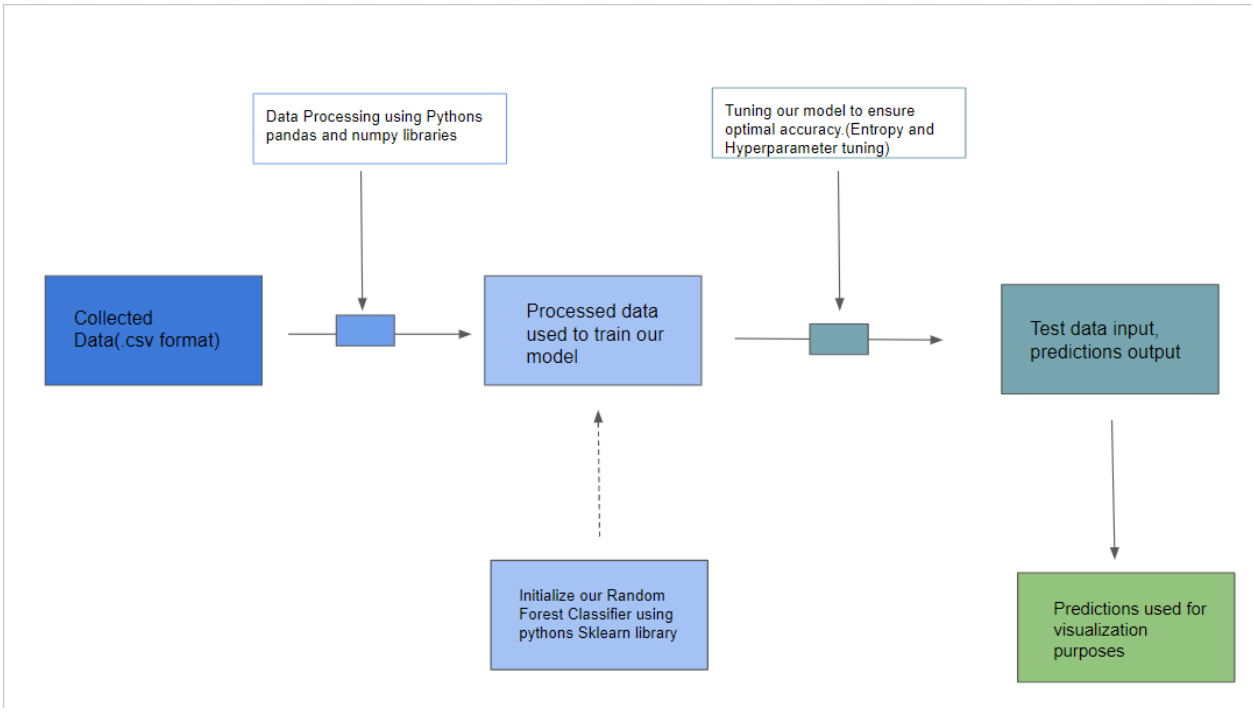
NEWMAC Score Predictor

The general idea for our final project is to use historical NEWMAC soccer data to train a machine learning model to predict the outcomes of an upcoming Clark Men's Soccer Conference run. Specifically, we plan to use a baseline model along with Logistic Regression, Decision Tree, and Random Forest Algorithms for our machine learning models, comparing the accuracies of each individual model.

Our first step towards the implementation of this idea was data collection. While tedious, the process of manually inputting NEWMAC game-to-game data was essential in ensuring that we had a sufficient database to draw conclusions from. Below is the format of our .csv file containing said data(Full scope of data does not fit in screenshot):

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O
1		Home Team	Away Team	H_Prev_Season	A_Prev_Season	FT_Home Goals	FT_Away Goals	H_Team Shots	A_Team Shots	H_Team Shot %	A_Team Shot %	H_Team SOG	A_Team SOG	H_Team SOG %	A_Team S
2	09/18/2010	WPI	Clark	4	5	2	3	9	8	0.222	0.375	5	6	0.556	
3	09/25/2010	Coast Guard	Clark		5	1	1	15	16	0.067	0.063	0	0	0	
4	10/02/2010	Clark	Wheaton	5	2	1	2	7	13	0.143	0.154	3	9	0.429	
5	10/09/2010	Clark	Springfield	5	6	2	2	13	15	0.154	0.133	7	6	0.538	
6	10/23/2010	MIT	Clark	3	5	1	0	19	7	0.053	0	8	3	0.421	
7	10/30/2010	Clark	Babson	5	1	2	3	10	24	0.2	0.125	8	11	0.8	
8	09/22/2012	Clark	Babson	7	1	0	5	12	19	0	0.263	4	9	0.333	
9	09/29/2012	Clark	Coast Guard	7	6	0	1	20	12	0	0.083	9	8	0.45	
10	10/06/2012	Springfield	Clark	2	7	4	1	13	4	0.308	0.25	8	3	0.615	
11	10/13/2012	WPI	Clark	3	7	2	1	12	20	0.167	0.05	5	6	0.417	
12	10/20/2012	Wheaton	Clark	4	7	4	0	17	15	0.235	0	9	5	0.529	
13	10/27/2012	Clark	MIT	7	5	0	3	10	16	0	0.188	4	8	0.4	
14	09/21/2013	Clark	WPI	7	6	1	2	16	7	0.063	0.286	3	3	0.188	
15	09/28/2013	Coast Guard	Clark	2	7	3	0	23	5	0.13	0	14	2	0.609	
16	10/05/2013	Springfield	Clark	4	7	5	0	17	7	0.294	0	12	4	0.706	
17	10/12/2013	Clark	Babson	7	1	1	1	12	30	0.083	0.033	5	13	0.417	
18	10/19/2013	MIT	Clark	3	7	2	0	27	12	0.074	0	10	6	0.37	
19	10/26/2013	Clark	Emerson	7		4	0	18	14	0.222	0	8	7	0.444	
20	11/2/2013	Clark	Wheaton	7	5	1	4	12	14	0.083	0.286	8	8	0.667	
21	09/20/2014	WPI	Clark	6	7	1	0	16	8	0.063	0	8	4	0.5	
22	09/27/2014	Clark	Coast Guard	7	5	2	1	8	18	0.25	0.056	4	4	0.5	
23	10/04/2014	Clark	Springfield	7	4	2	2	19	34	0.105	0.059	10	11	0.526	

From there we will use pythons ***pandas*** and ***numpy*** libraries for data processing, ensuring that we have a data matrix that is compatible with pythons ***sklearn*** library(All values must be numerical and some are unnecessary for prediction). From the ***sklearn*** library, we will get the necessary tools to train the backbone of our Baseline, Logistic Regression, Decision Tree, and Random Forest Models. Of course from here the models will need to be tuned to ensure optimal accuracy. This differs from model to model(also depends on how we process our data), but we will use techniques such as rolling averages, hyperparameter tuning, and entropy to achieve this. As of now we are still working up to that point. Once tuned, our models will be primed and ready to make predictions. We also plan to compare the accuracies and predictions made by each individual model. Our predictions will be multi-class, with baseline return values being 0 for a Clark loss, 1 for a draw, and 2 for a Clark win. We are still finalizing plans for visualization, but our leading idea is to display these predicted results against the actual results from this season. Below is a map that highlights the basic architectural plan for our project:



In our initial report we planned to have our data collected and organized by the midterm deadline. We achieved this goal, with all our desired data imputed into our .csv file. While we did achieve our initial goal, it should be noted that we still have a lot of work ahead of us. With our season ending soon we plan to start putting more focus and energy towards the project and picking up the pace. Our next steps are data processing and initialization of our ML model. We plan to split up and divide these two steps between two groups.

<u>Why did you choose a particular language/framework?</u>

For our project we are planning to use Python as our primary programming language. This is because our project partially falls in the realm of data science, and python has some of the best functionality when it comes to handling and managing datasets. For instance, python's ***numpy*** and ***pandas*** libraries are great tools that we plan to use for processing our data. Along with that, we also plan to incorporate ***sklearn's*** Random Forest Classifier, Decision tree, and Logistic regression frameworks, comparing the accuracies of each model. Sklearn offers great machine learning tools, and for us as a group with little to no machine learning experience the ***sklearn*** library will be imperative in our understanding and implementation of these models. Although we haven't finalized our plans for visualization, we will also most likely incorporate a graph type python library, such as ***matplotlib***.

Did you meet your milestone?

Yes, we did meet our goal set in the initial report, which was to have all of our data collected and organized in our .csv file. However, as mentioned above, this milestone was not the most ambitious, so we still have a lot of work left to do as a group. With that being said we feel that we are still on track for our project, and are now poised to move ahead with the development and fine-tuning of our ML models.

Did you encounter any challenges?

While the data collection process was relatively smooth, we did encounter some challenges in regards to time management. Because we are all in season, we haven't had a ton of time to all focus on the project simultaneously. The process of collecting data was very much a work individually whenever you get the chance kind of approach. While we did achieve our goals, we'll definitely need to focus our energy more as a collective in the future. Additionally, we also see some challenges that could arise in the future. Namely in the implementation and tuning of our ML models. Because we have essentially no experience with machine learning, we feel that the learning curve could be a very steep one. With that being said, however, we feel confident in the research we've done regarding the specifics of the topic as well as the frameworks offered by many of the sources we've collected during the drafting of our initial report.

Do you need to change any milestones?

We don't think we need to change any of our milestones. Our plan is to divide up the goals and work highlighted in our initial report amongst subgroups. This will help maximize efficiency, and allow for a more focused plan to address and solve problems. While we think the order and rigidity of our milestones from our initial report is bound to change, we are still on track to get to the finish line by the highlighted date.