Group 7 Writeup by: Kodee Bonalewicz, Daniel Howell, Benjamin Howell, Zephan Keach

Data Pipeline: The data was scraped from ufcstats.com. The dataset includes over a dozen variables that record different events that happen in the match. The events include: shots landed and shots attempted, which are separated into head shots, body shots, leg shots as well as standing shots, ground shots and clinch shots, takedown attempts and takedown successes, reversals in ground control, total control time and knockdowns. It also includes the weight class of the fighter, the date the fight happened, and whether they are a male or female. We are using the data for k-means clustering as well as for skill estimates that will be used to determine transition probabilities on our Markov chain.

Baseline Model Running:
We have a logistic regression model that predicts a win or loss for fighter i against fighter j by their differences in cumulative significant strikes landed per cumulative minutes fought (CSLpCM). We use this as a baseline to see whether our k-means clustering, markov chain and skill estimates offer better predictive power than a simple logistic regression.

Preliminary Results:
In the screenshot of our K-Means clustering results below we can see our data is grouped into three separate clusters. These clusters are using data and organized by parameters we selected that distinguish fighter types and make outcomes more predictable. We can see in the K-Means data here that cluster 2 has far greater ground shots landed and attempts indicating a fighting style that revolves more on the ground which we can cluster as wrestlers in our data set. We can say that cluster 1 is a more aggressive fighter without much time on the ground because of the higher proportion of head and body shots landed/absorbed. For cluster 0, we can see this cluster as more defensive in nature because of the overall fewer body head and body shots landed/absorbed meaning the fighters in cluster 0 are much more careful about when they attack.

K-Means:

Cluster Type 1	Head Shots Landed	Head Shots Absorbed	Body Shots Landed
0	19.120412	19.002699	7.352031
1	41.142183	36.752857	14.697895
2	36.661355	12.080952	6.756227

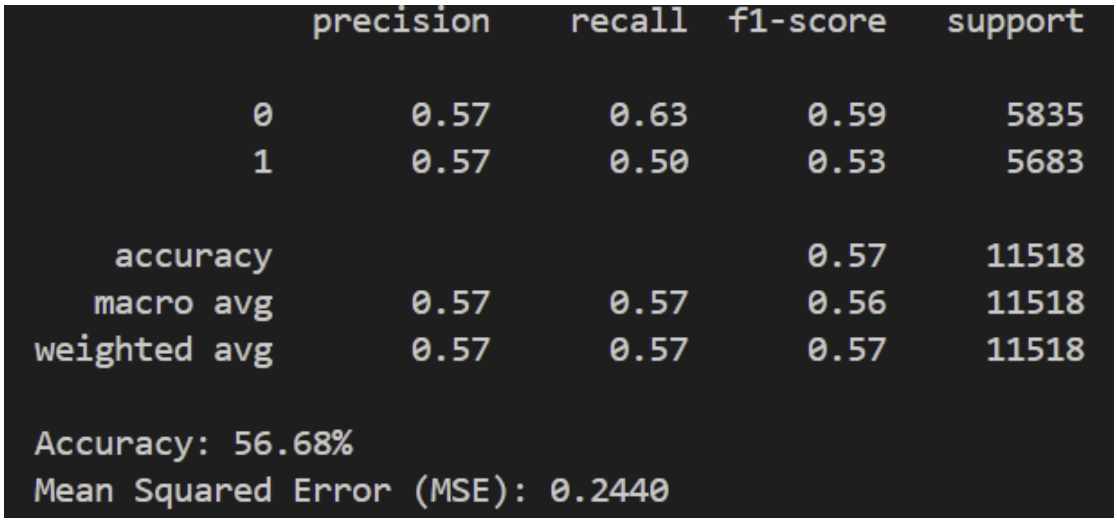
Cluster Type 1	Body Shots Absorbed	Takedowns	Takedown Attempts	\
0	7.157853	0.966645	2.512194	
1	12.907257	0.764938	2.259959	
2	2.756593	2.619597	5.230769	

Cluster Type 1	Takedowns Absorbed	Takedowns Attempted Against	\
0	0.926056	2.470272	
1	1.124917	3.551214	
2	0.164103	0.913553	

Cluster Type 1	Ground Shots Landed	Ground Shots Attempted	\
0	3.614541	5.057515	
1	5.077771	7.216610	
2	25.059158	35.379670	

Cluster Type 1	Ground Shots Absorbed	Ground Shots Attempted Against
0	4.701668	6.625629
1	3.770766	5.330545
2	2.271429	3.283883

We also have skill estimates for the transition probabilities on the markov chain, and their excel files are in the github. Lastly, we have results for the baseline logistic regression, shown below.



	precision	recall	f1-score	support
0	0.57	0.63	0.59	5835
1	0.57	0.50	0.53	5683
accuracy			0.57	11518
macro avg	0.57	0.57	0.56	11518
weighted avg	0.57	0.57	0.57	11518

Accuracy: 56.68%
Mean Squared Error (MSE): 0.2440

The baseline model achieves a 56.68% accuracy, outperforming naive guessing (~49-50%) by a meaningful margin. This suggests that CSLpCM is a valuable predictor of fight outcomes. We will use this model as a benchmark for evaluating the performance of our Markov chain simulations.

Next Steps: Currently, our K-Means uses 3 clusters, representing fighting, wrestling, and mixed styles. We can review further if creating more clusters will more effectively isolate matchups where our markov chains work better. Also, we will simulate the markov chain and predict fights in the second half of 2023, and use the k-means clustering to see if our model predicts better for certain style matchups. We are also going to compare our results against the baseline model.

Contribution Breakdown: Kodee worked on K-means clustering and analysis, Ben worked on skill estimates and the markov chain, Dan worked on data collection, the baseline model and skill estimates, Zephan reviewed accuracy of the model and helped with the baseline.