**Executive Summary**

This report covers the methodology and findings of a case study in Machine Learning to see if it is possible to predict whether a Pokémon is Legendary or not. The reason for this study was fueled by personal intrigue. The number of Pokémon currently exceeds one thousand. With such a large number of Pokémon and a popular competitive scene there must exist a sort of balancing for the stats of each Pokémon. If there exists a balancing methodology, then there might be a pattern to discern and used to predict if a Pokémon is Legendary or not.

This case study was possible thanks to Kaggle and the users who have taken the time to collect the data used. While many datasets were experimented with, only one was ultimately chosen to be used due to its completeness and quality. Both supervised and unsupervised learning was used in the case study. Furthermore, several different training algorithms were used in an attempt to get the best results.

**Legends**

The question or problem, if you will, that this assignment sought to solve is, is it possible to predict if a Pokémon is Legendary or not. In order to solve this problem, it was imperative that a quality dataset was used that was as complete as possible.

During the search to procure a dataset that met all the requirements, many were used. Some were even used almost to completion of the assignment, but ultimately ended up falling short. One of the requirements for the dataset was that they had to be complete. For this assignment that means the dataset must include all the Pokémon that exist, their base stats, and whether they are Legendary or not. The reason for the existence of the label in the dataset was because supervised learning was to be used for the assignment.

The data set that was ultimately chosen met all requirements and was also of high quality. This means that the categorical features it had were consistent and there were not a lot, if any, missing values. The reason for these requirements of “quality” is because even though there are over one thousand Pokémon, in terms of datasets, that number is small. The less data I would have to throw out, the better.

Apart from solving the problem of being able to predict if a Pokémon is legendary or not, I also aimed to see if I could glean any other insights. Suffice to say, I sat some interesting relationships and correlations between the different features of a Pokémon.

After the data exploration, pre-processing, and visualization was done, it was time to begin training the models. As previously stated, both supervised learning and unsupervised learning were used along with different training algorithms that we learned in class thus far.

The results of the training were interesting to say the least. I found that unsupervised learning yielded the worst result but did get over 50% one time. In terms of statistics, that’s not too shabby! Supervised learning on the other hand had the best results. Both regression and classification algorithms were used for supervised learning. Unfortunately, the regression algorithms yielded substantially worse results than the classification algorithms. I assume this was because of the binary nature of the label and how some of the actual labels were tightly intertwined with the non-legendary Pokémon. In the end, I was able to achieve a classification accuracy score of over 98% on during one run. Unfortunately, I have not been able to get the same or higher score.

In conclusion, the problem posed was solved and solved rather well. Using supervised learning and classification algorithms, it is possible to predict if a Pokémon is legendary over 97% of the time, consistently. At times, there were scores of a higher value. To iterate on this assignment, I would choose to scrape my own dataset so that I can get the most complete and robust dataset possible that included all features possible.

**Appendix**

The presentation will utilize the notebook itself, as such, there are no slides to attach. However, a copy of the notebook itself and an HTML version will be included along with the report.