Teammates Emails:

Daniel Bueno: dbueno3@buffalo.edu

Christopher Deck: cmdeck2@buffalo.edu

CSE 469 Data Mining Final Project:

Predicting if a Pokemon is Legendary Based on Their Stats

Introduction:

In the game Pokemon, it is rare and hard to catch a legendary Pokemon because it is hard to come across. You have to do a lot of work to be able to catch the legendary Pokemon. While we know this for a fact we wanted to explore a Pokemon dataset that involves each Pokemon's stats and compare them to the Legendary Pokemon who are also in the dataset which helped us better understand how a Legendary Pokemon's stats differ from a non-legendary Pokemon. Therefore, Our data mining problem revolves around exploring the process of predicting a Pokemon is a legendary type based on their stats in Pokemon. The impact that our process of predicting a Pokemon is a legendary type based on the Pokemon stats will be bringing a classifier algorithm to be used in the future Pokemon and see if that Pokemon is Legendary or Non-legendary based on their stats.

Formulation:

Our data mining task can formulated into a classification that can be expanded to predict whether a Pokemon is Legendary or not, based on its states. The reason is that classification is suitable because we have a distinct category that each Pokemon needs to be assigned based on their attributes. With the input being the Pokemon stats we can have an expected output of each

Pokemon in the dataset be classified as either Legendary or Non-Legendary and get a graph of how both Pokemon differ from each other and why Legendary Pokemon are different from Non-Legendary Pokemon.

Dataset:

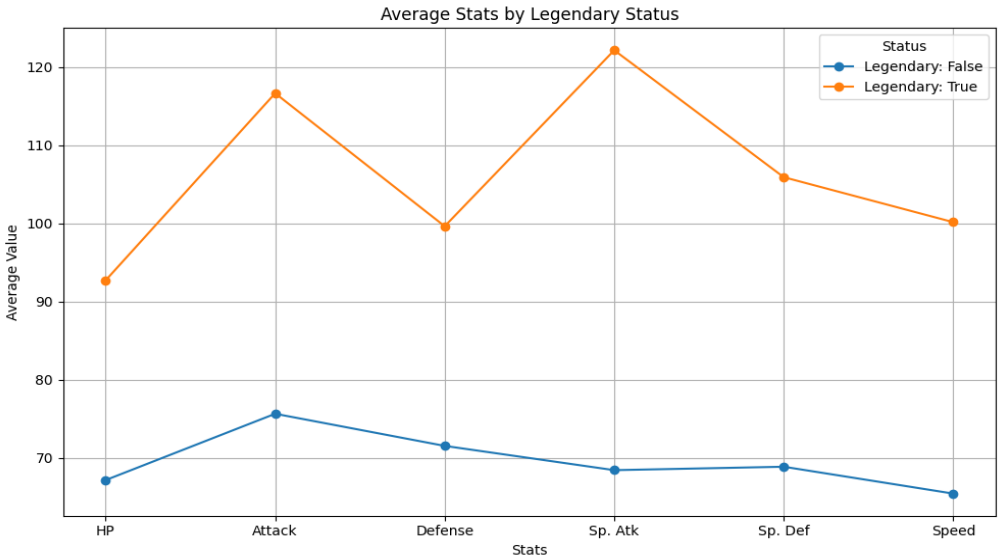
As for our dataset, we took from Kaggle which is our primary dataset and source. While looking at the dataset we noticed that the data did not need extra work to clean the data or preprocess the dataset to be in a format that our data mining algorithm could be applied on the dataset. So there is a plus to getting our data from Kaggle and the developers that use that same data and clean up the dataset for us to get straight into applying our algorithm to the dataset and see how the dataset and algorithm behave with each other and analyze the output we get back.

Algorithm:

We wrote a classification algorithm and a clustering algorithm to be used on our Pokemon dataset. As for our classification algorithm, we ended up creating a Random Forest Classifier which is similar to the Decision tree algorithm but we combined those trees to reach a single result and plotted those on a graph to see the difference and see where each Pokemon expands on the graph. With this algorithm, we explored the average of Legendary and Non-Legendary Pokemon's stats and see if there were any dramatic differences (Line Graph). We also explored the average of the attack and defense stats against special attack and special defense and see if there were any dramatic differences and if there were any outliers in the graph (Scatter Plot). Our algorithms both display a Simple Linear classifier with the two classes being Legendary type and Non-Legendary type Pokemon.

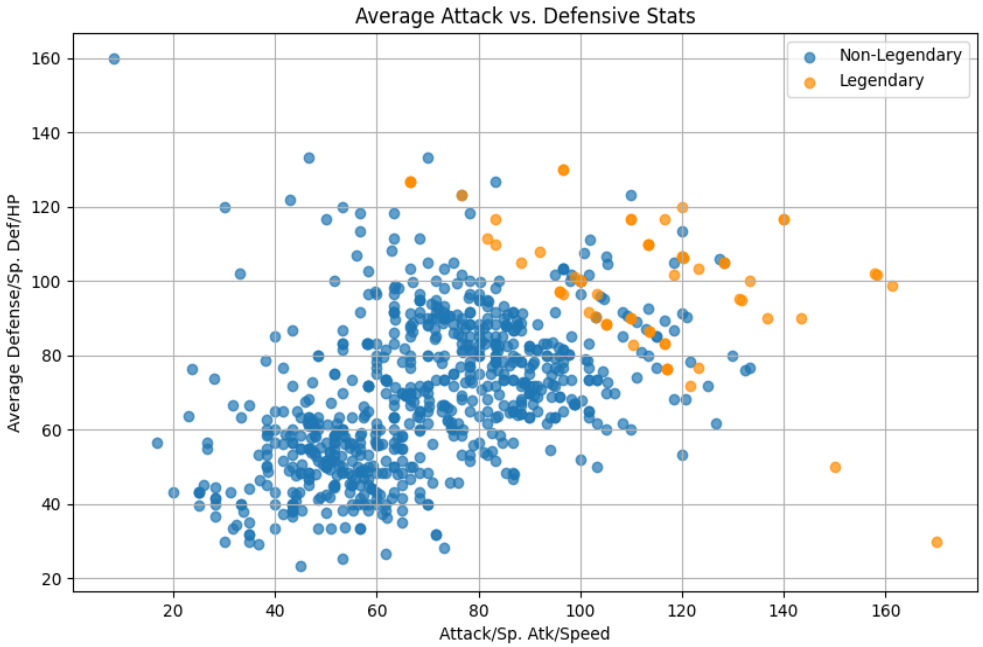
Experiments:

Average Stats by Legendary Stats:



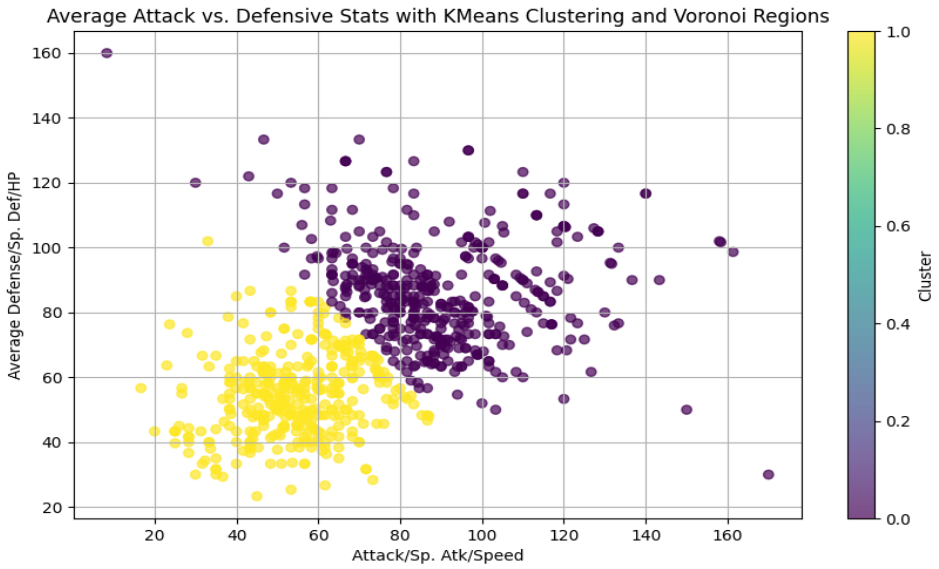
This line graph compares the average stats for the Legendary being labeled as true and Non-Legendary as False Across the 6 different metrics (HP/Attack/Defense/Special Attack/Special Defense/Speed). With that being said there are noticeable differences in the average states between Legendary and Non-Legendary Pokemons. Legendary Pokemons have a higher average in all 6 metrics than Non-Legendary Pokemons.

Average Attack and Defense stats:



This Scatter Plot shows the distribution of Pokemon based on their combined Attack, Special Attack, and Speed against their combined defensive stats including Defense, Special defense, and Health Points. The Legendary Pokemons who are labeled as orange dots tend to have higher values in both attack and defense stats compared to Non-Legendary Pokemons who are shown in blue. There is a clear distinction between the two groups, with the Legendary Pokemon being clustered towards the top right corner of the graph which indicates that they have a higher average in both metrics.

K-Means clustering Average Attack vs. Defensive Stats:



This Scatter Plot shows the application of K-Means clustering on the dataset, with each point representing a Pokemon characterized by its average Attack and Defense stats. The color represents different clusters identified by K-Means. The color legend on the right side of the graph represents the Pokemons membership with the probability scale of a metric of distance from the cluster centroid. For the cluster insights, we notice the separation in the cluster it shows that the Pokemon dataset which shows that there are Legendary and Non-legendary Pokemons can both be considered not an outlier. Also, the majority of Pokemon who have below average on attack and defense are considered outliers which is interesting to think about.

References:

https://www.kaggle.com/datasets/abcsds/pokemon

https://en.wikipedia.org/wiki/Random_forest