**To Be or Not to Be: Legendary**

As a long time Pokémon fan, I have always been captivated by how Gamefreak, the developer, is able to manage and maintain such a large set of Pokémon while continuing to add new ones. Most importantly, I have always been intrigued with figuring out if they have a formula for how they decide to assign attributes to the Pokémon and what makes a legendary Pokémon stand above the rest. With the discovered dataset covering all Pokémon thus far, my goal is to develop a model that, if possible, can predict whether a Pokémon is legendary or not.

The dataset in question is composed of 1032 rows and 44 columns. Each row relates to one and only one Pokémon while the columns relate to that Pokémon’s attributes. There are both numerical and categorical columns present in the dataset. Additionally, some of the values in the dataset are NULL, however, this is to be expected because the column in which this occurs represents an optional attribute that a Pokémon may have. Seeing as there are over 40 columns in this dataset, I will describe them in groups. The first group, characteristics, will include columns such as name, type 1, type 2, height, weight, [is] legendary, etc. Type 1 and type 2 indicate the ‘typing’ of a Pokémon, that is, is the Pokémon of the type grass, fire, water, etc. The second group will be the ‘stats’ group. The stats group includes columns such as att, def, spa, etc. The third group will be referred to as the ‘effectiveness’ group. These columns indicate how effective a Pokémon is against the type specified in the column name.

Due to the fact that the dataset contains the information as to whether or not a Pokémon is legendary, this will be a supervised model utilizing classification methods because we are trying to learn what makes a Pokémon legendary or not. However, I am curious to see if taking an unsupervised approach would yield similar results. This can be done by removing the label, the [is] legendary column, and seeing if any models can classify the Pokémon accurately. While my main approach is that of supervised, I would like to compare results between both supervised and unsupervised models.

My biggest concern is that there may be one column that is a near perfect predictor of what makes a Pokémon legendary. This would mainly be concerning to me since it could essentially make the other columns unneeded and hide any other possible patterns present in the dataset. If that were to occur, I will probably end up training two models to see if there can be greater insight learned without the near perfect predictor column.