# Section # 01: Introduction

## What is Machine Learning?

Machine learning is a technique that performs statistical inference and pattern analysis to work out mathematical models to define various systems such as predictive analysis, classification, clustering, etc. It is an amalgamation of several different fields such as data mining, Optimization, Statistics and Mathematics. The variety of models described in these fields combine to form the powerful tool we now know as machine learning. In this task we were asked to perform three different machine learning tasks. While performing this task, we also come across one of the most important properties of machine learning which is generalization ability. We can see that how these powerful techniques can learn from very few observations and then predict the output for a similar but different valued set of observations.

## What is PySpark?

Another major component of our project was PySpark. Since we were required to code the entire assignment in PySpark it is important that we set the ground rules for what PySpark is. The major functionality and selling point of PySpark is scalability and pipeline building. PySpark helps with scaling the machine learning based pipelines where we can perform several different tasks for various data and then deploy them at a distributed level. Furthermore, it is possible to setup the clusters and then deploy the PySpark based applications. In this task, we use the PySpark library to perform the various tasks.

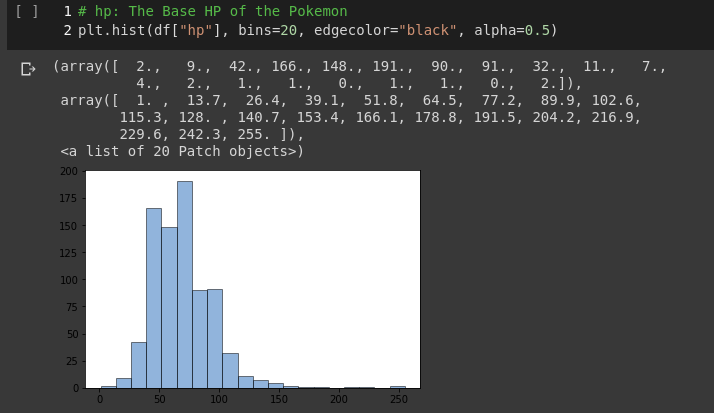
## Dataset:

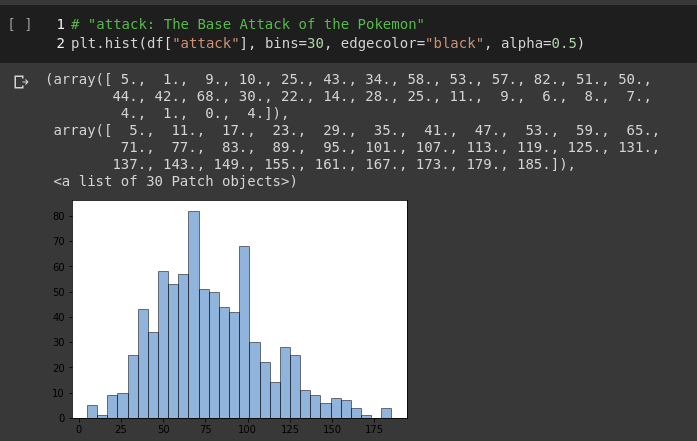
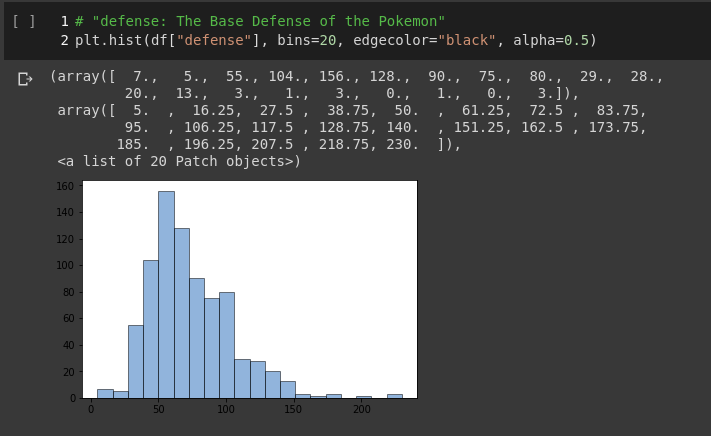
https://www.kaggle.com/ranja7/movieratingsbyusers

https://www.kaggle.com/rounakbanik/pokemon

## Introduction

In this assignment we first did exploratory data analysis on the dataset we had to choose from the Kaggle. For this assignment, I used the Pokemon dataset from Kaggle which can be found at this link: <https://www.kaggle.com/rounakbanik/pokemon>. According to the dataset provider, “This dataset contains information on all 802 Pokemon from all Seven Generations of Pokemon. The information contained in this dataset include Base Stats, Performance against Other Types, Height, Weight, Classification, Egg Steps, Experience Points, Abilities, etc. The information was scraped from <http://serebii.net/>”.

At first we performed various data set analyses to find out what features are necessary for the task. I plotted the histograms for various feature points of the Pokemon dataset. The graphs:



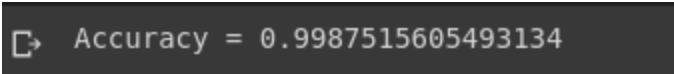
These graphs illustrate the necessary features that are required for the task that we are going to perform. I perform three tasks which are collaborative filtering using the Alternative Least Squares (ALS) method which is a PySpark speciality, Logisitc Regression with a regularization i.e. L2 and finally I perform the K-Means clustering on the dataset. We discuss the strategy and results of each technique we perform under their specific heading.

# Section # 02: Techniques

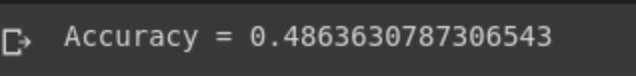
## Collaborative Filtering with ALS

Filtering is a mechanism which is used to filter out the most important information which is of use to the users from any given dataset. This technique was famously used with the Netflix Movie recommendation challenge where they asked data scientists to come up with a technique to solve the recommendation problem they had faced and gave cash prizes to the winner. Collaborative filtering with ALS employs matrix factorization in a parallel fashion. This is the speciality of PySpark which makes almost all of the famous machine learning techniques scalable in this case due to the massive parallelization that is performed before performing the factorization. The factorization process helps churn the large matrices into bite sized pieces which can be then processed to figure out the necessary things. I also used mean squared error which is error calculation based on the squared mean of the filtered out elements. The task is to minimize the mean error so that the best possible recommendation is acquired. This is a good error measure since once we have minimum mean we can safely say that unnecessary recommendations are not there anymore i.e. variation from mean is also minimum. In this specific dataset, our task was to recommend a Pokemon based on several features of there such as speed, agility, power, shape, size, etc. Without the factorization, the mean of each is very large and after application of technique it is reduced to around about 3, which is very less and we can be rest assured that we will be recommended the best Pokemon to choose from. Only if Ash new this.

## Logistic Regression

In the second task, I performed logistic regression to check if we can classify the Pokemons based on their properties such as speed, power, shape, size, color, skills, etc. The logistic regression with LBFGS employs a regularization technique which is a Euclidean regularization i.e. L2 which helps with making sure all the values are normalized and we do not see much of the deviation from mean. We get favorable accuracy figures on the dataset which we split into training and testing before performing te necessary regression tasks. We get about 99.8% accuracy figures on the dataset which speaks for the powerful technique this. It is important to note that this technique is highly dependent on the dataset i.e. a smaller dataset can be learned and we get good figures. It gets increasingly difficult as the dataset size increases.

## K-Means Clustering

For the last task, we were asked to perform K-Means clustering to cluster the Pokemons based on their properties. This helps to analyze which Pokemon is most similar to which Pokemon and then we can select based on their skills which cluster to choose from. This can also help normalize the hierarchy of Pokemons. For example, one might need Pokemons with high speed and water specialty. With this clustering techniques, we can perform such a task. This technique is also focused on reducing the distance between points and the cluster centers and optimizes it. A simple Euclidean distance can be used to perform the task. However, we can also use a more specialized distance accordingly. We also get very good accuracy figures with this technique.

# Section # 03: Conclusion

In this assignment, we worked on three main tasks of machine learning and complemented with exploratory data analysis which helps with working what features to choose. I used PySpark as recommended to perform the different tasks such as collaborative filtering, logistic regression and K-Means clustering.

# Section # 04: References

1. <https://en.wikipedia.org/wiki/K-means_clustering>
2. <https://en.wikipedia.org/wiki/Machine_learning>
3. <https://towardsdatascience.com/a-brief-introduction-to-pyspark-ff4284701873>
4. <https://www.kaggle.com/rounakbanik/pokemon>
5. <https://stackoverflow.com/questions/49932310/what-algorithm-does-ml-classification-logisticregression-use-in-spark>
6. <https://stats.stackexchange.com/questions/17436/logistic-regression-with-lbfgs-solver>