Title: A Smart Pokedex: Harnessing the power of Machine Learning

Team: Tyler Walsh, Christian Cipolletta

Kaggle dataset:
Pokémon with stats and image
(https://www.kaggle.com/datasets/christofferms/pokemon-with-stats-and-image/data)

Part 1: Implementation contributions:
We plan to implement an array of different models in this project. We plan to implement a Convolutional Neural Network (CNN) which will guess the main type of the pokemon in an image and compare its performance to a K-Nearest Neighbor (KNN) model. For the CNN we plan to use YOLO architecture through an Ultralytics package based on PyTorch (https://github.com/ultralytics/ultralytics). For the KNN we will implement our own using python packages like PyTorch and Scikit. The other experiment we are going to run is testing different models on their ability to guess the "Total Score" of a pokemon based on 6 other parameters. For this we plan to compare the accuracy of linear regression models with and without regularization. Again we will most likely create our own using the homeworks from this class and python packages.

If we have time, we would like to explore other model types like transformers and reinforcement learning, but do not want to be too ambitious at this time.

Part 2: Evaluation contributions:
We will compare the CNN and the KNN by recording the Test Accuracy on a held-out test set. We plan to use k-fold to ensure there is not a lucky run by one of our models. We will also run a determined hyperparameter sweep to judge the effect they have on accuracy. For the CNN we will change the learning rate, weight decay, and number of epochs. For the KNN, we will be changing the value of K. This will allow us to not only understand which architecture is more effective, but the effect of hyperparameters as well. We plan to record the time taken for training and testing in addition, so that we can take that into account.

For the linear regression model we will also record the Test Accuracy on a held-out test set and use k-fold verification. We will determine the accuracy by using the Mean Squared Error (MSE) of the model. The hyperparameter we will be tuning for the linear regression models is lambda.

Prior work:

1. https://itnext.io/how-to-use-machine-learning-to-build-a-Pokémon-search-engine-pt-1-2db1f8551426
2. https://docs.ultralytics.com/
3. https://www.kaggle.com/code/devraai/analyzing-pokmon-stats-and-predictive-modeling

Which parts of the curriculum from this class do you expect to apply?:
We plan to apply KNN, CNN, and linear regression lessons we learned from the curriculum. We have a dataset with both image and quantitative data which will allow us to implement a plethora of different models. We expect that reference 1 will give us a good starting point for the KNN, reference 2 will help us set up the CNN, and reference 3 will help us set up the linear regression.

Expected challenges and risk mitigation:
One challenge will have to deal with is training and inference time. In this day and age, everyone expects lightning fast speeds. The KNN and linear regression models are expected to train and predict fast, but the CNN might be a bigger burden. We will make sure to record the time taken to train and predict to compare this for the future. Furthermore, we plan to test the use of fine tuning a pretrained model which, in theory, should take a shorter time to train and be more accurate.

Another one of our expected challengers is data manipulation. The dataset is impressive as it gives us the images, names, types, and statistics of the pokemon. However, if we wanted to do any more analysis, for example some Pokémon are legendary status, we would have to manipulate the data ourselves. However, this does give an interesting idea, maybe a reinforcement learning model could be used to guess which Pokémon are legendaries.

Ethical considerations and broader social impact:
The major ethical concern behind our project is that Pokémon is technically not in the public domain. While we are not selling anything nor profiting in any way, the publishers of Pokémon (Nintendo) are very hostile against people using their work. Therefore if the project becomes big, it could easily be taken down by the publishers, Nintendo or Game Freak. Another impact is that machine learning models oftentimes use much more processing power (and therefore more electricity) than conventional methods. There is a chance that there will be more electricity consumption globally because of implementing these methods.