Predicting Pokémon Types Using Clustering and Classification

University of Toronto - STA2201 - Winter 2025

Justin Zhang, Isaac Baguisa, Alex Faassen

2025-04-04

Abstract

- Report Summary

Contributions:

- Justin Zhang:
- Isaac Baguisa:
- Alex Faassen:

Introduction

- Research Question
- Motivation
- Context

Over the past two decades, Pokémon has evolved from a niche Gameboy game into a global multimedia franchise encompassing anime, trading cards, e-sports competitions, and mobile applications. For many who grew up in the late 2000s and early 2010s, the world of Pokémon was a formative part of childhood entertainment alongside cultural staples like Mario Bros. and Zelda. As Pokémon continues to captivate audiences across generations, its community-oriented game design and well-structured universe presents an intriguing opportunity for data analysis. In this project, we investigate whether a Pokémon's type—a categorical label used in-game to denote elemental or behavioral characteristics such as Fire, Water, or Grass—can be inferred from its visual appearance and numerical attributes using statistical learning techniques.

Our primary research question is: *Can clustering and classification methods uncover or predict a Pokémon's type based on its image and statstical features?*

The Pokémon type system serves not only as a game mechanic but also as a conceptual grouping based on traits like colour, strength, and theme. We aim to explore whether these groupings have an underlying statistical structure that can be detected through dimensionality reduction, clustering, and classification algorithms. This analysis will provide insight into the extent to which a Pokémon's type is reflected in its appearance and physical characteristics, or whether it is a more arbitrary design choice by game developers.

Data Description

- Data description + overview
- Pre-processing
- Visualizations and summary stats: Stats head; Example image; Mean image

Our analysis is based on two publicly available Kaggle datasets:

1. Pokémon Stats Dataset

URL: The Complete Pokemon Dataset

This structured dataset contains 41 variables for 801 Pokémon. These features include:

- **Physical attributes:** height, weight, base experience
- **Combat statistics:** HP (Health Points), attack, defense, special attack, special defense, speed
- **Categorical indicators:** Legendary status, abilities, and generation
- **Primary and secondary type labels**

All variables are numeric or have been encoded numerically (e.g., legendary status as 0/1). There are no missing values. **Notably**, We only consider primary type labels for this project to maintain a single-label classification structure.

2. Pokémon Image Dataset

URL: Pokemon Image Dataset

The Pokémon Image Dataset consists of 809 PNG files representing unique images of Pokémon from generations 1 through 7. Each file consists of a 3-dimensional array of dimensions 120 by 120 by 4. The first 2 slices represent a 120 by 120 grid of pixels, and the 3rd slice represents the 4 RGBA channels (Red, Blue, Green, Alpha) colour and transparency assignments of each pixel.

Pre-processing

For the images, we begin by flattening each PNG into a (120x120x4) 57600-element vector, each pixel and RGBA value representing a feature. Binding these vectors into a table identified by Pokémon name gives us a usable dataset. To integrate image features with the stats dataset, we filter both tables to common Pokémon requiring several name formatting adjustments to account for unusual characters. Finally, we match the row-order of both datasets, resulting in two easily transferable Pokémon datasets.

Methodology

- What methods/models were chosen and why w.r.t. research question

Image Dimension Reduction

- Image PCA

Pokemon Type Clustering Patterns

- Image PCA results
- t-SNE/UMAP on Pokemon Type

Unsupervised Model

- e.g. Weighted K-means

Supervised Model

- e.g. Log reg, KNN, LDA

Results

- Summarize findings (include tables/plots)
- Interpret results w.r.t. research question

Discussion

- Discuss model performance
- Limitations of methodology
  - Potential sources of bias
- Challenges encountered
- Recommendations for overcoming these + improvements for future work

Notes: - Dual-typing might be causing some difficulties - adding complexity. - Perhaps other approaches, such as deep learning, might do better.

References

- Use `knitcitations`? - manual easier?

(Optional) Appendix

- If we have any "extra" non-essential tables that we want to include, but don't want contributing to the page limit.