Pokémon classification

Nenad Petković SW-37/2018
Isidora Savić SW-72/2018

1. Motivation

Primary, our motivation for this work comes from our love for pokemon. Predicting the main type of pokemon was a great idea we came out with. On the other side, we also wanted to use all the skills we learned in this course.

2. Research questions

In this project, we are predicting the main type of pokemon. We wanted to predict the main type of pokemon based on their other characteristics.
For the dataset, we used Pokémon Stats (Pokémon Go) from Kaggle[1]. This dataset contains stats, attributes, and descriptions of Pokémon in Pokémon Go. This dataset was created so we can find patterns between different primary or secondary types, regions, and categories of Pokémon.

Dataset contains:

- **number** - The number of the Pokémon in the Pokédex
- **pokemon_name** - The name of the Pokémon
- **pic_url** - A URL to a .png file of the Pokémon
- **main_type** - The primary type of the Pokémon
- **secondary_type** - The secondary type of the Pokémon (or NULL only 1 type)
- **region** - Region where Pokémon first appeared in the Pokémon games
- **category** - Category of the Pokémon
- **height** - Height of the Pokémon
- **weight** - Weight of the Pokémon
- **pokemon_family** - Derivative family of the Pokémon (ex. Charmander, Charmeleon, Charizard will all be Charmander Family)
- **attack** - Pokémon base attack stat in Pokémon Go
- **defense** - Pokémon base defense stat in Pokémon Go
- **stamina** - Pokémon base stamina stat in Pokémon Go
- **cp_range** - Pokémon Combat Power (CP) range in Pokémon Go
- **hp_range** - Pokémon Hit Point (CP) range in Pokémon Go
- **capture_rate** - Pokémon capture rate in Pokémon Go ('N/A' values indicate that data was not available from the derived data source)
- **flee_rate** - Pokémon flee rate in Pokémon Go ('N/A' values indicate that data was not available from the derived data source)
- **male_perc** - Percentage of Pokémon found that are male ('N/A' values indicate Pokémon is genderless, only male, or only female)
- **female_perc** - Percentage of Pokémon found that are female ('N/A' values indicate Pokémon is genderless, only male, or only female)

- **resistance** - Dictionary of Pokémon resistances to certain types in the following form: {'percent_resistant': ['type_1', 'type_2']}
- **weakness** - Dictionary of Pokémon weaknesses to certain types in the following form: {'percent_weak': ['type_1', 'type_2']}
- **wild_avail** - Available in the Wild?
- **egg_avail** - Available from the Eggs?
- **raid_avail**- Available as the Raid Boss?
- **research_avail** - Available from Field Research?
- **shiny** - Shiny form available?
- **shadow** - Shadow form available?
- **pkedex_desc** - Pokémon's Pokédex description
- **poss_attacks** - List of possible attacks available to the Pokémon

3. Related work

We did not find any work that is related to this problem. Because of that, we decided that we will use algorithms and methods already created and tested for homework and implement them for this problem.

4. Methodology

Firstly, data from the dataset had to be processed. Some columns, such as Pokemon name or URL to photo, were removed. Columns containing intervals are divided in two: interval beginning and ending. In label encoding, we also implemented a custom categorization for column "*poss_attacks*" where we created an algorithm that made results slightly better.

After the data is processed, the dataset is divided into training and test sets, both containing X and y lists. X contains all columns which are later to be used for predicting Pokemon types, while y contains expected types.

When it comes to classification algorithms, we have tried to implement all of the following: Bagging, Boosting, K-NN, K-means and Naive-Bayes. All of the above, except HistGradientBoostingClassifier, required empty fields in columns to be replaced with actual data. Both replacing and not replacing empty fields has been tested, and we concluded that not replacing any of the missing data gave the best results. Therefore the algorithm used for classification is *HistGradientBoostingClassifier* .

Finally, after the main types have been predicted, accuracy is calculated, and all results are printed in the console.

5. Discussion

The training set is used to fit the model. Only hyperparameter in *HistGradientBoostingClassifier* that had to be learned was *learning rate.* Its value was learned by experimenting with different values around 1, and for that validation dataset was used. The best results were achieved with a learning *rate* of 0.09.

After fitting the model, it is used to predict y values for the test dataset. For model evaluation we use accuracy, and the best one we achieved was 86%.

Results obtained from other models can be found in the rez.md file in the project.

6. References

[1] Pokémon Stats (Pokémon Go) Stats, Attributes and Descriptions of Pokémon in Pokémon Go