Data Analytics

Element 1 - Group 2

Team Members

BIBI KURIAN (w9457273)

CORCORAN, JAMES (j9049758)

LUI, YU SHING (a0201746)

NG, CHUN YU (w9190570)

ODUNOLA, BOLA (a0141109)

PYBUS, DANIEL (p4261837)

Dataset

- Dataset:
  - Complete Pokemon Dataset
- List of Pokedex:
  - Generation One to Eight
- Data Content:
  - 38 columns
  - 1027 rows

- Data Source:
  - https://www.kaggle.com/mariotormo/complete-pokemon-dataset-updated-090420?select=pokedex_%28Update_05.20%29.csv

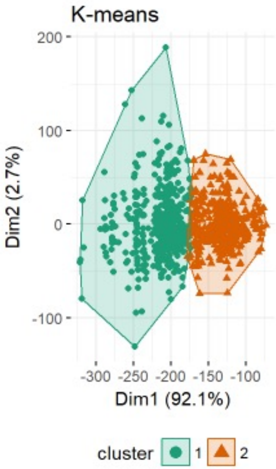
Case study’s scope and objectives

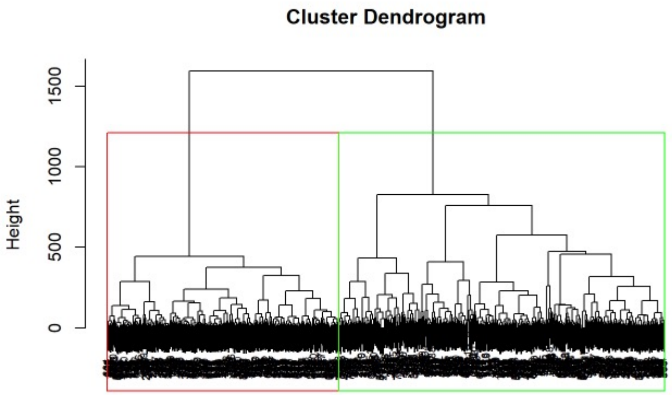
- Can we use clustering to identify different groups of Pokemons?
- Can legendary Pokémon be identified through the use of classification methods?
- What species are the strongest and weakest in Pokemon?

Can we use clustering to identify different groups of Pokemons?

- Objective
  - Use clustering to find groups of Pokemons
  - Attributes: HP, Attack, Defense, SP Attack, SP Defense, Speed, Height, Weight
  - Identify the strong points of different Pokemons

- Clustering methods
  - Hierarchical clustering
  - K-means





Chun Yu, Ng
Yu Shing, Lui

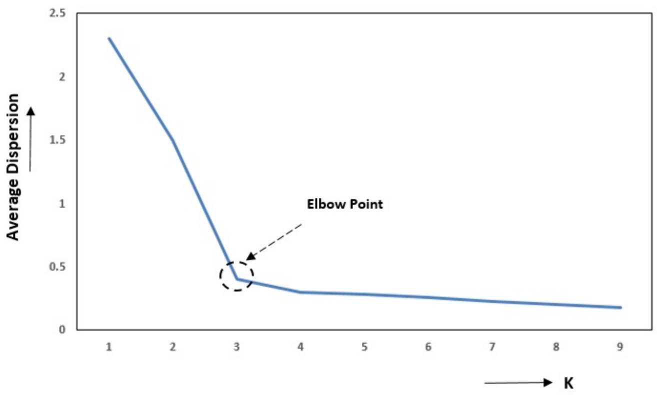
Can we use clustering to identify different groups of Pokemons?

- Compare results within and between the two methods
  - Between methods
  - Within methods
- Compare with Pokemon types

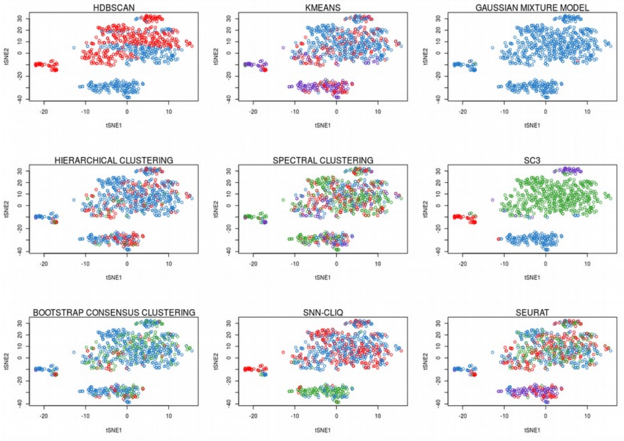
- Validation
  - Internal validation indexes and stats
  - Relative measures
  - Visual exploration

(Brock et al., 2008; Halkidi, Batistakis, & Vazirgiannis, 2001)

Elbow Method for selection of optimal "K" clusters



Bartoschek et al., Clustering on Raw Expression Matrix, N = 716, optPC = 30, optPerp = 27, N_clust = 4

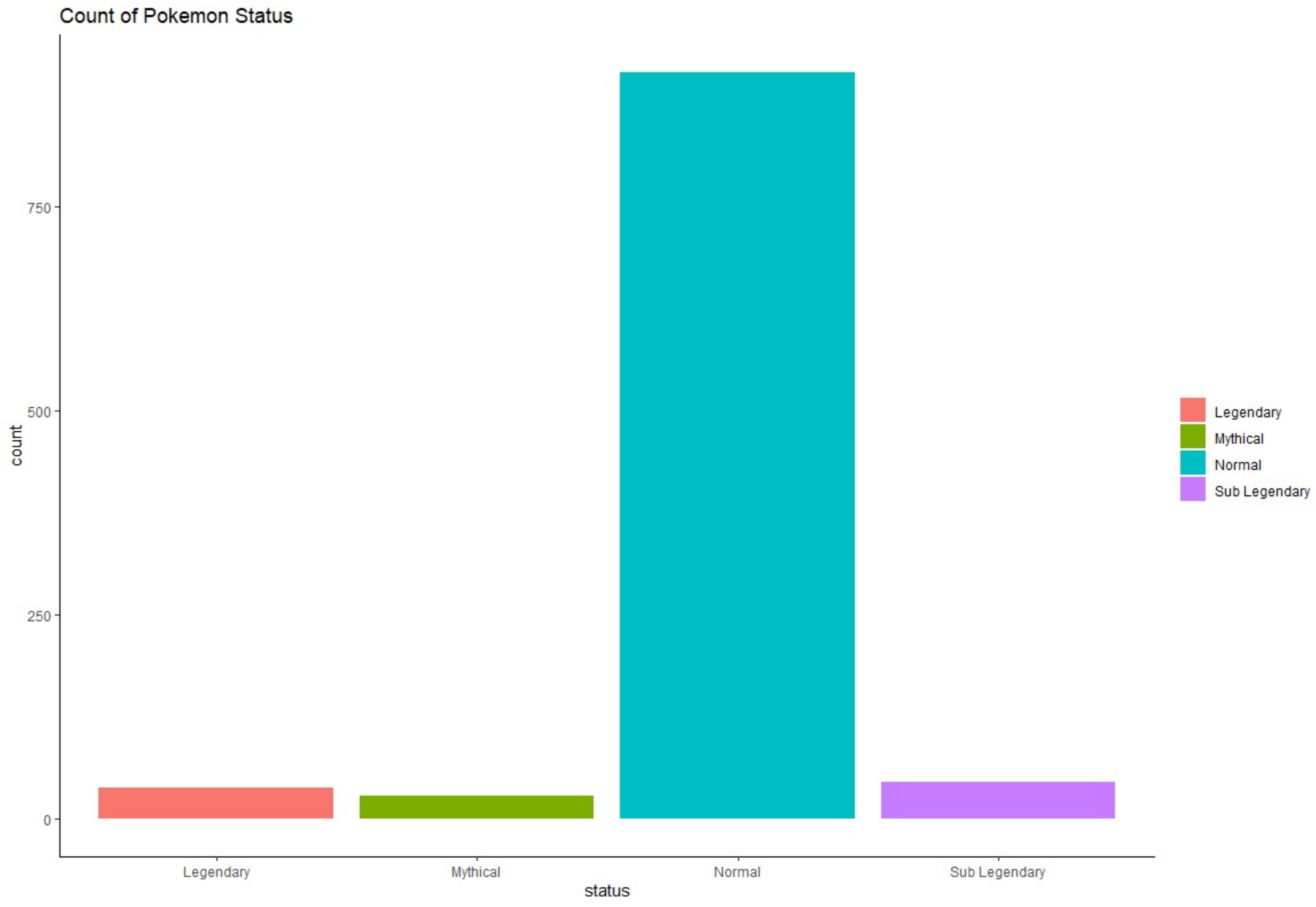


Daniel Pybus
James Corcoran

Can legendary Pokémon be identified through the use of classification methods?

Teesside University

Data Exploration

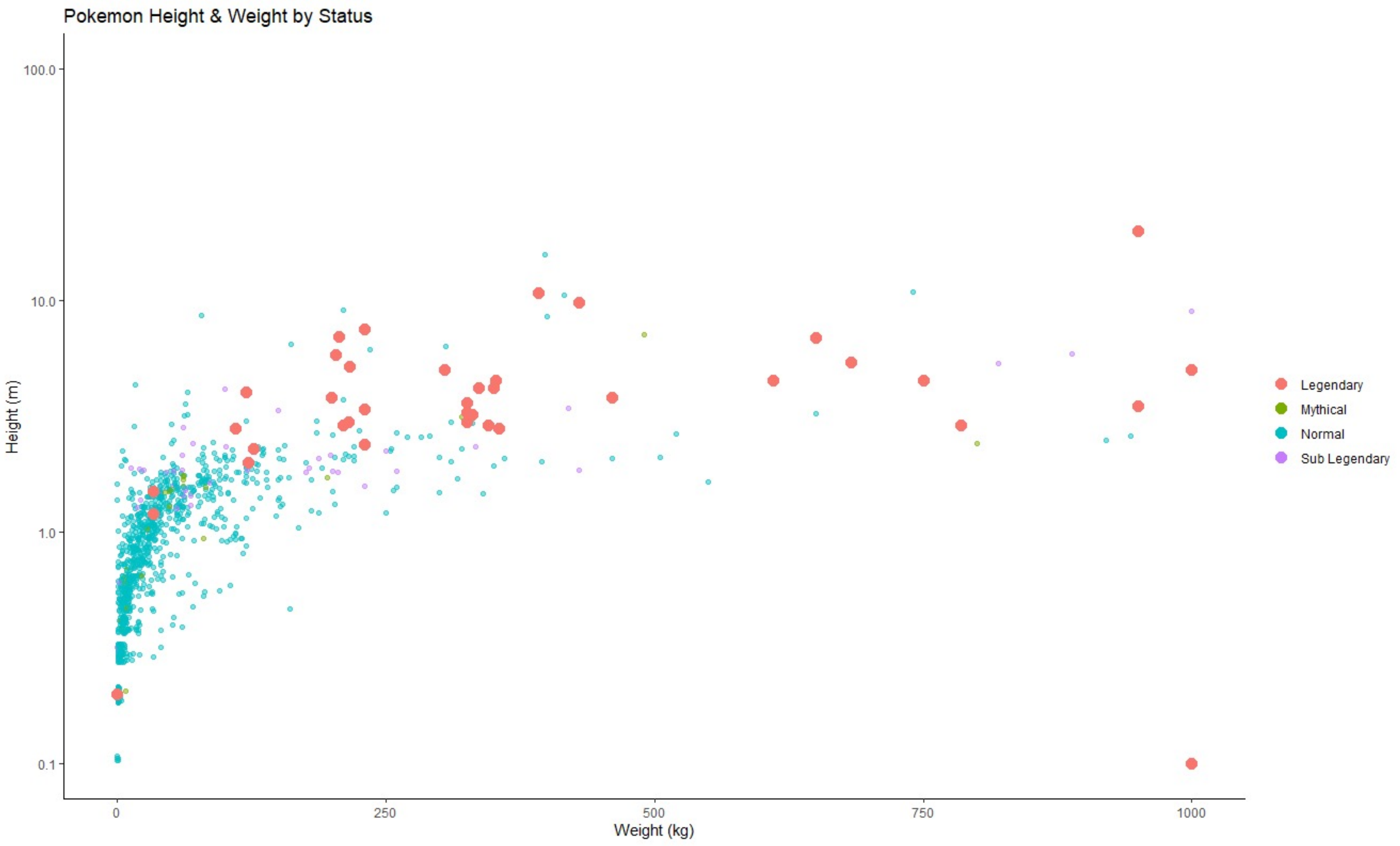


Daniel Pybus
James Corcoran

Can legendary Pokémon be identified through the use of classification methods?

Teesside University

Data Exploration

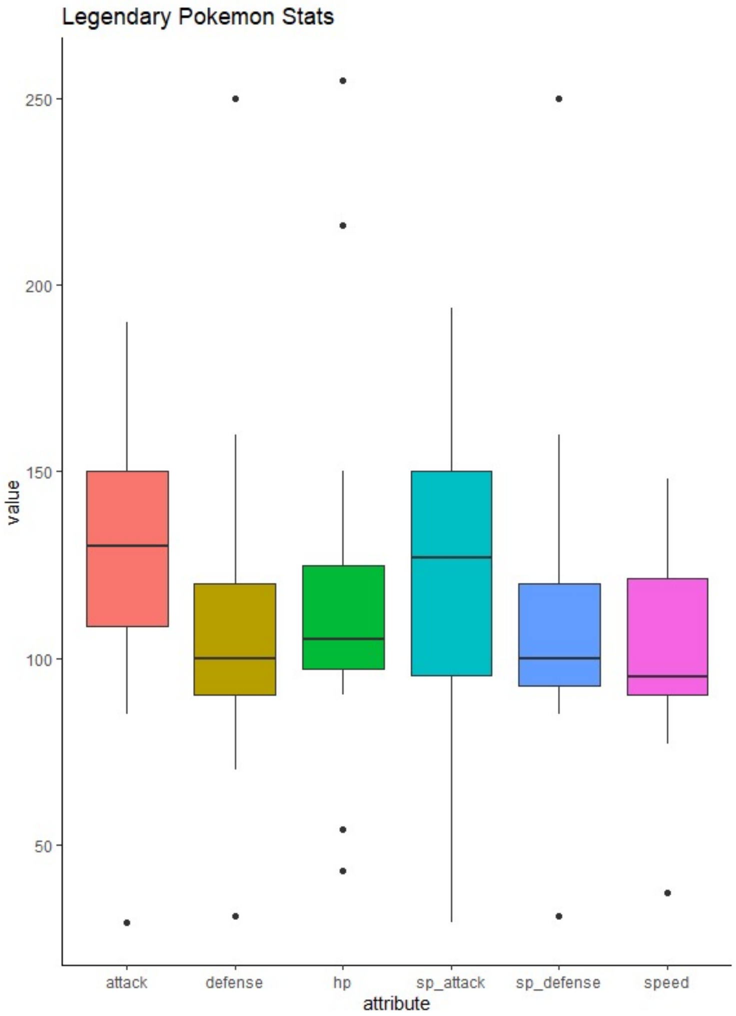


Daniel Pybus
James Corcoran

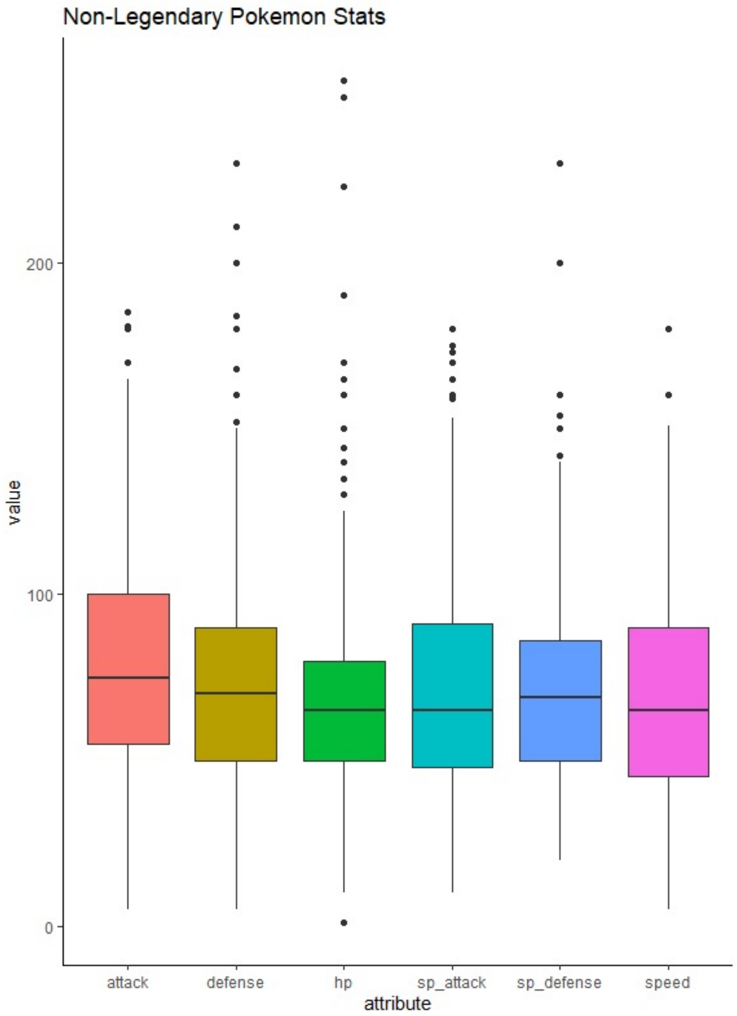
Can legendary Pokémon be identified through the use of classification methods?

Teesside University

Data Exploration



Legendary Pokemon Stats



Non-Legendary Pokemon Stats

Daniel Pybus
James Corcoran

Can legendary Pokémon be identified through the use of classification methods?

Decision Trees

Method

CART (Classification & Regression Trees) methodology: Each region of the tree is continuously divided into smaller sub-groups formed by asking yes/no questions in relation to features. (*Breiman, 2017*)

Measure

Gini Index: Each split is an attempt to minimise node impurity i.e. consisting mostly of observations from a single class. (*Boehmke et al, 2020*)

Evaluation

Early Stoppage: Restricting the depth or growth of the tree.

Pruning: Finding optimal tree depth through the use of cross-validation and complexity parameters. (*Boehmke et al, 2020*)

Decision Trees

Process Summary

Import data into R

Further data exploration

Tidy and process data based on findings

Train/test split

Build model

Comparison of evaluation methods to determine best practice

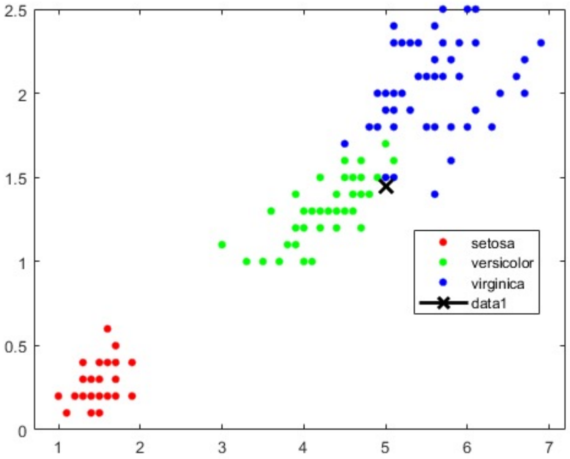
Identify importance of features i.e. VIP variables

Further Investigation

Random Forests

K Nearest Neighbour

- Simple algorithm that stores all available cases, and classifies new data based on similarity measures (Subramanian, 2019).



www.mathworks.com

- ‘K’ is a parameter that refers to the number of nearest neighbours to include in the majority of the classification process.
- Small K = Noisy Large K = Increased Bias
- Generally, K = sqrt(total number of data points)

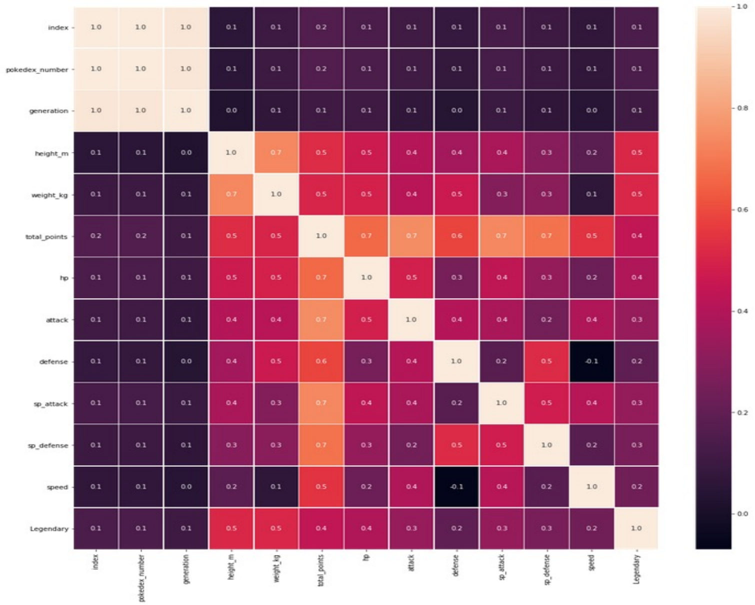
Daniel Pybus
James Corcoran

Can legendary Pokémon be identified through the use of classification methods?

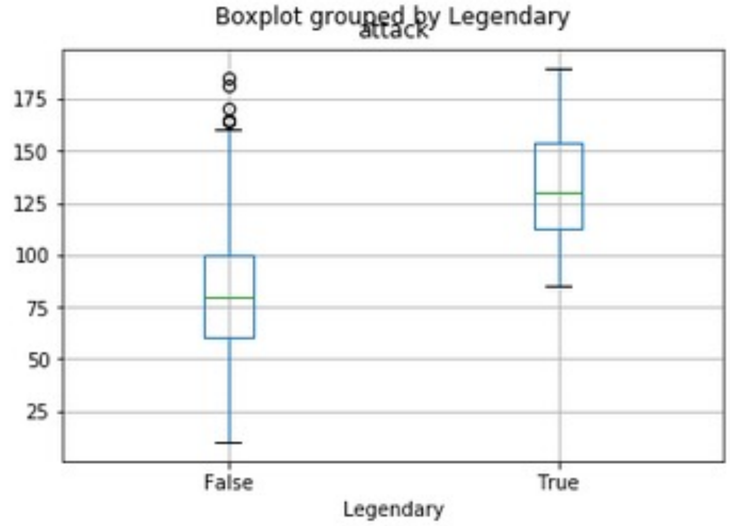
Teesside University

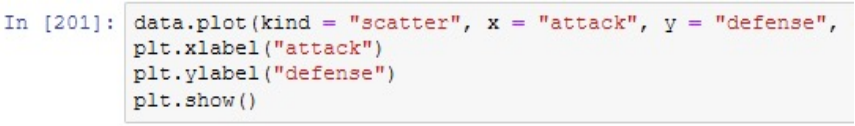
Proposed Method:

- Read csv, clean data
- Test-Train data, 0.2 test size (20%)
- Standardise columns – StandardScaler
- Determine K value
- Predict data – classifier.predict
- Evaluate model to check accuracy – confusion matrix
- Check f1 score & Accuracy score
- Plot graph – Is Legendary – TRUE/FALSE

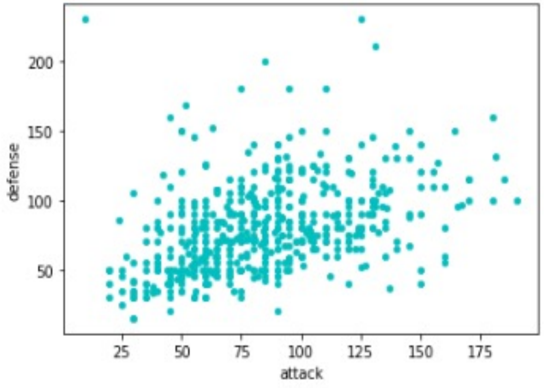


	index	pokedex_number	generation	height_m	weight_kg	total_points	hp	attack	defense	sp_attack	sp_defense	speed	Legendary
index	1.0	1.0	1.0	0.1	0.1	0.2	0.1	0.1	0.1	0.1	0.1	0.1	0.1
pokedex_number	1.0	1.0	1.0	0.1	0.1	0.2	0.1	0.1	0.1	0.1	0.1	0.1	0.1
generation	1.0	1.0	1.0	0.0	0.1	0.1	0.1	0.1	0.0	0.1	0.1	0.0	0.1
height_m	0.1	0.1	0.0	1.0	0.7	0.5	0.5	0.4	0.4	0.4	0.3	0.2	0.5
weight_kg	0.1	0.1	0.1	0.7	1.0	0.5	0.5	0.4	0.5	0.3	0.3	0.1	0.5
total_points	0.2	0.2	0.1	0.5	0.5	1.0	0.7	0.7	0.6	0.7	0.7	0.5	0.4
hp	0.1	0.1	0.1	0.5	0.5	0.7	1.0	0.5	0.3	0.4	0.3	0.2	0.4
attack	0.1	0.1	0.1	0.4	0.4	0.7	0.5	1.0	0.4	0.4	0.2	0.4	0.3
defense	0.1	0.1	0.0	0.4	0.5	0.6	0.3	0.4	1.0	0.2	0.5	-0.1	0.2
sp_attack	0.1	0.1	0.1	0.4	0.3	0.7	0.4	0.4	0.2	1.0	0.5	0.4	0.3
sp_defense	0.1	0.1	0.1	0.3	0.3	0.7	0.3	0.2	0.5	0.5	1.0	0.2	0.3
speed	0.1	0.1	0.0	0.2	0.1	0.5	0.2	0.4	-0.1	0.4	0.2	1.0	0.2
Legendary	0.1	0.1	0.1	0.5	0.5	0.4	0.4	0.3	0.2	0.3	0.3	0.2	1.0





```
In [201]: data.plot(kind = "scatter", x = "attack", y = "defense",
          plt.xlabel("attack")
          plt.ylabel("defense")
          plt.show()
```



References

- Mario Tormo Romero, M. 2020. Kaggle. [Online]. [12 November 2020]. Available from: https://www.kaggle.com/mariotormo/complete-pokemon-dataset-updated-090420?select=pokedex_%28Update_05.20%29.csv
- The Pokemon company. 2020. Pokemon. [Online]. [12 November 2020]. Available from: https://www.pokemon.com/uk/pokedex/
- Boehmke, B.C. and Greenwell, B. (2020) *Hands-on machine learning with R.* 1st edn.
- Breiman, L. (2017) *Classification and regression trees.*
- Brock, G., Pihur, V., Datta, Susmita., and Datta, Somnath. (2008) “ClValid: An R Package for Cluster Validation.” *Journal of Statistical Software* 25 (4): 1–22.

References

- Mathworks.com. 2020. *Classification Using Nearest Neighbors- MATLAB & Simulink*. [online] Available at: <https://www.mathworks.com/help/stats/classification-using-nearest-neighbors.html> [Accessed 4 November 2020].
- Subramamanian, D., 2020. *A Simple Introduction To K-Nearest Neighbors Algorithm*. [online] Medium. Available at: <https://towardsdatascience.com/a-simple-introduction-to-k-nearest-neighbors-algorithm-b3519ed98e> [Accessed 6 November 2020].
- Mitchell, T., 2017. *Machine Learning*. New York: McGraw Hill.
- Halkidi, M., Batistakis, Y., & Vazirgiannis, M. (2001). On clustering validation techniques. *Journal of intelligent information systems*, *17*(2-3), 107-145.

Q & A