POKÉMON DATA MINING

Dennis Donkoh, Ayobami Akinseloyin

[**Abstract**](#_2e10lmlqmimy) **1**

[**Introduction**](#_dv4ze14vnta) **1**

[**Background**](#_dlbj1dm6zwm9) **3**

[**Project Objectives**](#_laawd4yvtgsm) **4**

[Project Plan](#_n900o7r8jx78) 5

[Significance and Contributions](#_xh2udxw2tevk) 5

[**Literature Review**](#_by0azjqy59qr) **6**

[Application Domain](#_5ghpbp40p7yb) 7

[Business Problem](#_3clj39zejxh5) 8

[**Success Criteria**](#_cj7no1jgv3k1) **8**

[**Data Understanding**](#_gkw8d49c4zk) **9**

[**Type 1 and 2 Classification**](#_rzzmf0h3fvl9) **12**

[**Abilities Association Rule Mining**](#_xdp60mylkp2i) **13**

[**Catch Rate/Rarity Clustering**](#_5r4g9wlfgo8r) **14**

[**Total Points Regression**](#_vosvaqtn4svx) **18**

[**Evaluation**](#_pic8zalxyxa2) **19**

[**Conclusion**](#_v10iys8g9g55) **19**

[**Reference List**](#_aj1kp6rk9ukp) **20**

# 

# 

# 

# Abstract

This study conducts a thorough analysis of Pokémon data, utilizing data mining techniques to uncover significant patterns. It focuses on Pokémon's total points, abilities, catch rates, rarity, and type. Data preparation includes cleaning, feature selection, and encoding categorical variables. The study applies association rule mining, clustering, classification, and regression to understand relationships between Pokémon attributes, group Pokémon, predict types, and estimate total points. The findings offer valuable insights into Pokémon characteristics and their interrelations, useful for game strategists, players, and enthusiasts.

# Introduction

The Pokémon universe, crafted by Satoshi Tajiri and Ken Sugimori (Tajiri and Sugimori, 1996), has a diverse collection of unique creatures, each boasting distinct abilities and classifications. Understanding their complexities is vital for game strategists, players, and presents an intriguing challenge for data mining and machine learning.

This project ventures into the Pokémon universe using a data-centric approach. The primary aim is to leverage a variety of data mining techniques on a Pokémon dataset to unearth hidden patterns, relationships, and insights. The project is divided into several crucial sections, each spotlighting a different facet of the Pokémon data.

The Data Understanding section offers a thorough analysis of the Pokémon data, including outlier detection and a meticulous examination of Pokémon abilities.

The Data Preparation & Abilities Association Rule Mining section focuses on data preparation for analysis and applies association rule mining to the Pokémon abilities (Agrawal et al., 1993). This phase involves data cleaning, feature selection, and encoding categorical variables. Association rule mining is then used to identify common associations among different Pokémon abilities.

Catch Rate & Rarity Clustering employs clustering techniques to group Pokémon based on their catch rates and rarity. This analysis provides insights into the distribution of Pokémon catch rates and their correlation with rarity.

Rarity and Type Classification use classification algorithms to predict a Pokémon's rarity and type. These sections utilize machine learning models, including Random Forest (Breiman, 2001) and Support Vector Machines (SVM) (Cortes and Vapnik, 1995), to predict the target variables based on the features in the dataset.

Multiple regression models are deployed to predict a Pokémon's total points, using various characteristics available in the dataset (James et al., 2013). This process includes data preprocessing and employing LabelEncoder to convert categorical data into a numerical format. Features are then selected based on their correlation with the target variable. Once the dataset is divided into training and testing sets, various regression models, specifically Ordinary Least Squares Linear Regression, Ridge Regression, Decision Tree Regressor, and Random Forest Regressor, are trained and evaluated. Additionally, k-fold cross-validation (Kohavi, 1995) is incorporated into the assessment of the models to provide a more accurate measure of their performance, thereby enhancing the robustness and reliability of our model comparison.

The project's findings offer valuable insights into the Pokémon universe, providing a deeper understanding of the characteristics and relationships that define these creatures. The results could be beneficial for game strategists seeking to optimize their gameplay, players interested in understanding the game's mechanics better, and enthusiasts who wish to gain a deeper appreciation of the Pokémon universe.

The project also demonstrates the application of various data mining techniques, showcasing their potential in extracting meaningful information from complex datasets. This could serve as a valuable reference for researchers and practitioners in the field of data mining and machine learning.

# Background

The Pokémon franchise, a global phenomenon that has captivated audiences worldwide, was first introduced in 1996 by Nintendo, Game Freak, and Creatures (Tajiri and Sugimori, 1996). The franchise's core concept revolves around fictional creatures called "Pokémon", which humans, known as Pokémon Trainers, catch and train to battle each other for sport. The Pokémon species are collectible, with hundreds of different designs for creatures.

Over the years, the franchise has evolved and expanded significantly. It began with a pair of video games for the original Game Boy and has since expanded to encompass various gaming genres, including trading card games, animated television series, movies, comic books, and toys. The franchise also made a significant impact on popular culture, particularly in the late 1990s, with its English slogan becoming a catchphrase.

In 2016, the Pokémon franchise made a significant advancement with the launch of the mobile game, Pokémon Go. This game, developed and published by Niantic in association with Nintendo and The Pokémon Company, utilizes GPS on mobile devices to locate, capture, train, and battle virtual creatures known as Pokémon, which appear as if they are in the player's real-world location (Wikipedia contributors, 2023). Pokémon Go quickly became a sensation, emerging as one of the most used and profitable mobile apps in 2016, with 500 million downloads globally by the end of the year. The success of Pokémon Go has played a substantial role in reviving the franchise's popularity, impacting player behavior and social interactions worldwide (Wikipedia contributors, 2023).

The Pokémon universe is expansive, with each Pokémon species having unique abilities, types, and stats. These attributes provide a rich and diverse dataset for exploration and analysis. For instance, a Pokémon's abilities can affect the outcome of battles, its type can determine its strengths and weaknesses against other types, and its stats can influence its performance in battles. This complexity and depth make the Pokémon dataset a perfect candidate for data mining.

Data mining, a process used to discover patterns and relationships in large datasets, can provide valuable insights into the Pokémon universe. By applying data mining techniques to the Pokémon dataset, we can uncover patterns and trends that can enhance user experiences, inform game design, and guide marketing strategies for the franchise. For example, data mining can reveal which Pokémon abilities are most effective in battles, which Pokémon are most popular among players, or how a Pokémon's type and stats influence its performance.

This project aims to leverage data mining techniques to extract these valuable insights from the Pokémon dataset. The findings from this project could potentially benefit various stakeholders, including game developers, marketers, and players, by providing them with a deeper understanding of the Pokémon universe and informing their decision-making processes. Whether it's developing new features for the games, crafting marketing campaigns, or deciding which Pokémon to catch or train, the insights gained from this project can make a significant impact.

# Project Objectives

The primary objective of this project is to aid Pokémon trainers in optimizing their strategies, thereby increasing player satisfaction, revenue, and brand loyalty. The project aims to:

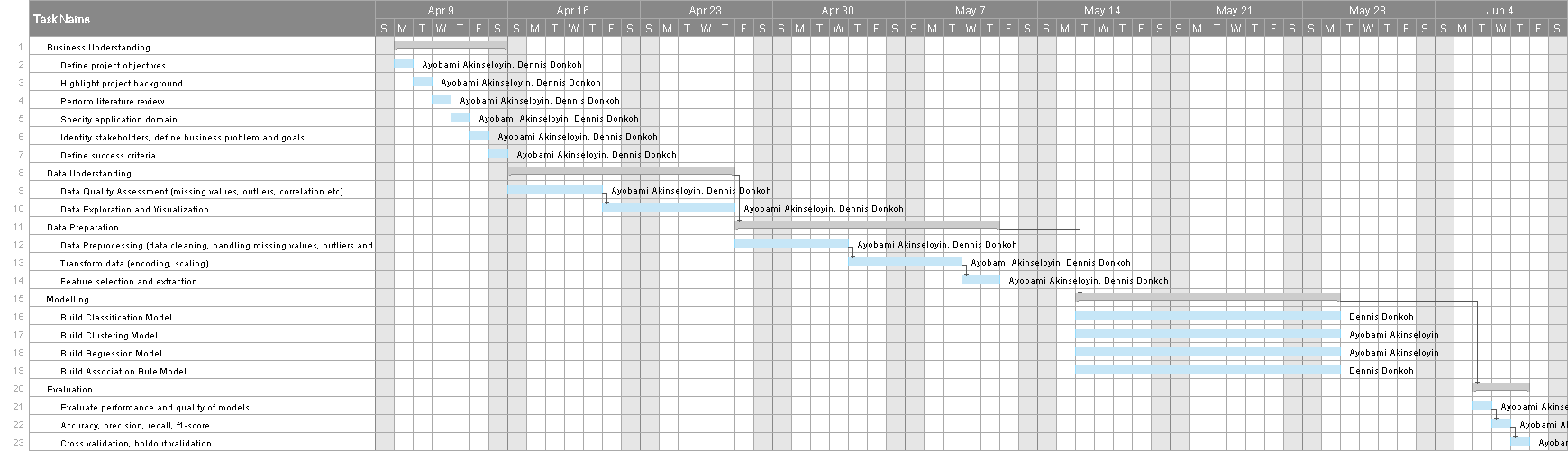
1. Help trainers identify which abilities to prioritize when training their Pokémon using association rule mining.

2. Assist trainers in identifying which of their Pokémon are most powerful and worth investing time and resources into training using a regression model.

3. Group Pokémon based on their rarity using clustering techniques.

4. Predict the type of Pokémon based on its attributes to aid trainers in battles using a classification model.

## Project Plan



## Significance and Contributions

The ongoing project holds immense potential for the Pokémon franchise, impacting a variety of stakeholders, including players, game developers, and marketers.

From the player's viewpoint, the project could significantly enhance their gaming strategies. By identifying the most valuable abilities and Pokémon, players could optimize their resource allocation, leading to a more skilled and versatile Pokémon team. This could boost their gaming performance, resulting in increased victories, higher rankings, and a more satisfying gaming experience. Moreover, players can make well-informed decisions about which Pokémon to catch, which abilities to develop, and what battle strategies to employ (Pokémon Go, n.d.).

For game developers, the insights from this project could refine game mechanics. If the data indicates that certain abilities or Pokémon are underutilized, new features or incentives could be introduced to encourage players to explore these game aspects. This could lead to increased player engagement, longer gaming sessions, participation in in-game events, and in-app purchases, all contributing to the game's success. Furthermore, empowering players to excel in the game can foster a positive gaming community and ensure long-term player loyalty (Pokémon Database, n.d.).

Marketers could leverage the analysis to identify potential opportunities for promotions and marketing. They could design campaigns around popular Pokémon and abilities to attract and engage both current and potential players. Additionally, this project could help marketers better understand the Pokémon community and its interests, informing strategies for community engagement, social media campaigns, community events, and collaborations with influencers. All these benefits combine to generate higher revenue for the franchise through merchandise sales, in-app purchases, and partnerships. The insights gained from this project could significantly contribute to the financial success of the Pokémon franchise (Pokémon Company, n.d.).

This project promises to enrich player experiences, advance game design, guide effective marketing strategies, and contribute to the franchise's success (Niantic, n.d.; Nintendo, n.d.).

# Literature Review

The Pokémon dataset on Kaggle has been the subject of numerous studies, with research focusing on five main areas: Pokémon types and stats, identification of legendary Pokémon, exploratory data analysis, linear discriminant analysis, and decision trees and random forests. These studies have used a variety of data mining techniques to extract meaningful insights, which can be beneficial for game strategy and development.

In terms of Pokémon types and stats, several studies have analyzed the types of Pokémon and their corresponding stats, such as HP, Attack, Defense, Special Attack, Special Defense, and Speed. The goal of these studies was to identify the Pokémon types with the highest average values for each stat, providing valuable insights into the strengths and weaknesses of different Pokémon types (University of Arizona, n.d.).

For the identification of legendary Pokémon, machine learning techniques have been used to predict whether a Pokémon is legendary based on its features. Researchers used a training dataset containing Pokémon characteristics and their legendary status, and a testing dataset without the legendary status information. The aim was to train a model to accurately predict the legendary status of Pokémon in the testing dataset, demonstrating the potential of machine learning in enhancing the gaming experience by providing players with predictive insights (Analytics Vidhya, 2021).

Exploratory data analysis has been conducted on the Pokémon dataset to visualize and understand the distribution of various features, such as generations, types, and stats. Python libraries like Pandas, Matplotlib, Seaborn, and Plotly were often used to create visualizations and identify trends in the data, which is crucial for understanding the underlying structure of the dataset and informing further analysis (Dhama, n.d.; In Machines We Trust, n.d.).

Linear discriminant analysis (LDA) has been used to explore the Pokémon dataset and predict Pokémon types based on their stat totals. LDA is a technique used to reduce the dimensionality of the dataset while preserving the differences between classes, making it particularly useful in Pokémon type prediction tasks (Kaggle, n.d.).

Decision tree and random forest algorithms have been applied to the Pokémon dataset for classification tasks, such as predicting legendary Pokémon. These machine learning techniques help identify the most important features for classification and build models that can accurately predict the target variable (Kaggle, n.d.; Analytics Vidhya, 2021).

However, the application of data mining techniques to the Pokémon dataset is not without its challenges and limitations. These can be broadly categorized into issues related to data quality, heterogeneous data, high dimensionality, and imbalanced data. To address these challenges, previous studies have employed various strategies, including data preprocessing, feature engineering, handling imbalanced data, model selection and evaluation, and ensemble methods (Medium, n.d.; Medium, n.d.; Dataversity, n.d.; Medium, n.d.; Towards Data Science, n.d.).

The application of data mining techniques to the Pokémon franchise has opened up several avenues for future research and development. These can be broadly categorized into areas related to gameplay, predictive modeling, data analysis techniques, user behavior, marketing strategies, and game development. By understanding player behavior, preferences, and trends, developers can create more engaging and personalized gameplay experiences, improve marketing strategies, and make informed decisions in game development (Medium, n.d.; Analytics Vidhya, 2021; Medium, n.d.; ACM Digital Library, 2017; Medium, n.d.; Resetera, 2019).

## Application Domain

The Pokémon data mining project is part of the dynamic gaming and entertainment industry, which is continually evolving due to technological advancements and changing consumer preferences. The project has the potential to significantly influence various aspects of this industry.

Insights from the project could guide game development by aligning game mechanics with trends in Pokémon abilities, types, and stats, potentially leading to more engaging games (James et al., 2013). For player engagement, understanding player preferences could lead to personalized experiences, a key factor in today's gaming industry (Hamari et al., 2017). This could involve personalized recommendations, challenges, and rewards (Cortes and Vapnik, 1995).

For marketing strategies, the project's findings could provide valuable data for targeted marketing campaigns. Understanding player behavior could help design campaigns that resonate with current players and attract potential ones. This could include promotions focused on popular Pokémon or abilities, collaborations with influencers who align with the game's audience, or events that tap into the community's interests.

In essence, the application domain of this project extends beyond just data mining or Pokémon. It intersects technology, gaming, entertainment, and consumer behavior. The insights gained from this project could influence game development, player engagement, and marketing strategies, potentially shaping the future of the gaming and entertainment industry.

# 

## Business Problem

The business problem that the project aims to solve is to help trainers make more informed decisions when training and battling their Pokémon. By providing trainers with insights into which abilities to prioritize, which Pokémon to invest in, and which types of Pokémon are most effective against specific opponents, the project aims to enhance player satisfaction and engagement, thereby benefiting the game developers by increasing revenue and brand loyalty.

# Success Criteria

The success of the project will be measured against the following criteria:

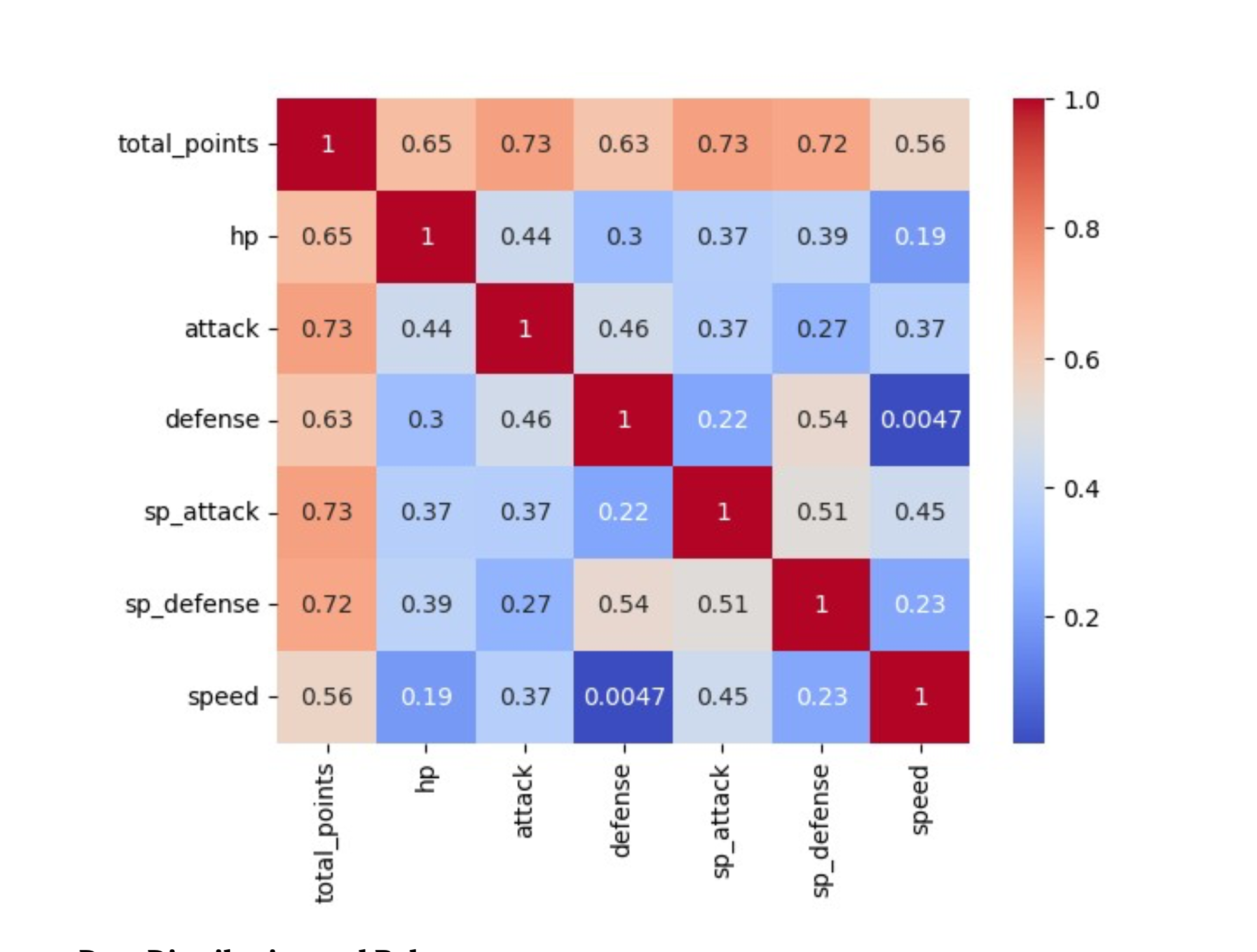
1. High accuracy in predicting Pokémon types with a classification model.
2. High accuracy in identifying relationships with association rule mining.
3. Successful distinction of Pokémon clusters based on rarity.
4. High accuracy in predicting total points with a regression model.

This project aims to leverage data mining techniques to extract valuable insights from the Pokémon franchise's rich dataset. The insights gained could potentially enhance user experiences, game design, and marketing strategies for the franchise, thereby benefiting both the players and the game developers.

# Data Understanding

The Pokémon dataset is analyzed, focusing on abilities, catch rates, rarity, and type classification. The data was prepared through cleaning, feature selection, and encoding of categorical variables. Missing values were replaced with 'Unknown' for categorical columns and the mean for numerical columns. Outliers in numeric variables were identified using z-scores and visualized with box plots. A correlation matrix was generated to understand relationships between Pokémon attributes, revealing a moderate negative correlation (-0.43282) between the number of abilities and total points.

The correlation matrix for attributes like 'total\_points', 'hp', 'attack', 'defense', 'sp\_attack', 'sp\_defense', and 'speed' in the Pokémon dataset reveals the following relationships:

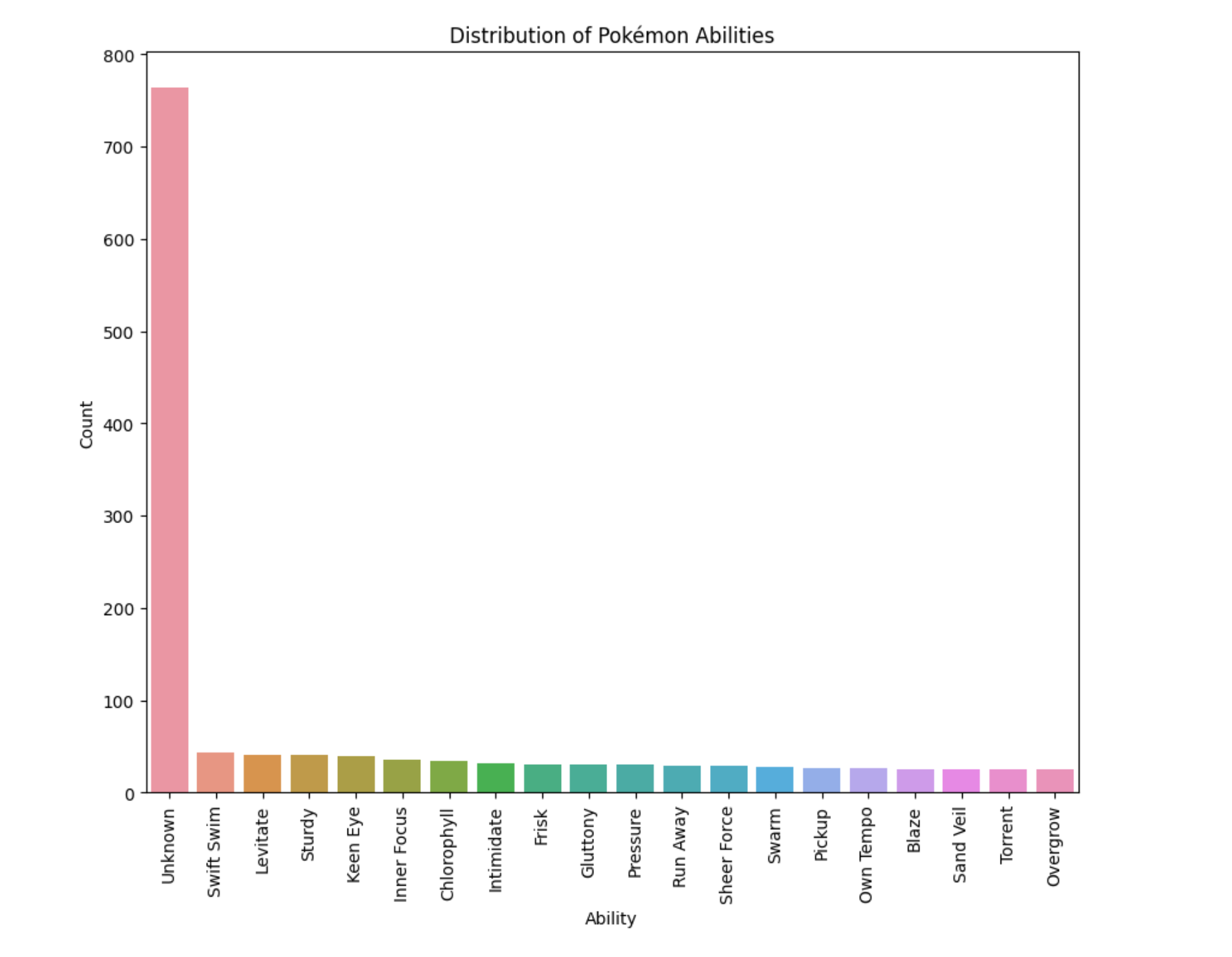


* 'total\_points' shows a strong positive correlation with 'attack' and 'sp\_attack', and a positive correlation with 'hp', 'defense', 'sp\_defense', and 'speed'. This suggests that Pokémon with higher total points usually exhibit higher values in these attributes.
* 'hp' shows a moderate positive correlation with 'attack', 'sp\_attack', and 'sp\_defense', indicating that Pokémon with more hit points (hp) generally have higher attack, special attack, and special defense values. However, it only shows a low positive correlation with 'defense' and 'speed'.
* 'attack' demonstrates a moderate positive correlation with 'defense', 'sp\_attack', and 'speed', suggesting that Pokémon with higher attack values also tend to have higher defense, special attack, and speed values. However, it only shows a low positive correlation with 'sp\_defense'.
* 'defense' shows a moderate positive correlation with 'sp\_defense', implying that Pokémon with higher defense values also tend to have higher special defense values. However, it only shows a low positive correlation with 'sp\_attack' and almost no correlation with 'speed', suggesting that a Pokémon's defense value doesn't necessarily relate to its special attack or speed values.
* 'sp\_attack' shows a moderate positive correlation with 'sp\_defense' and 'speed', suggesting that Pokémon with higher special attack values also tend to have higher special defense and speed values.
* 'sp\_defense' shows a low positive correlation with 'speed', suggesting that Pokémon with higher special defense values may have slightly higher speed values, but the relationship is not very strong.

These correlations do not imply causation but merely show the relationships in this specific dataset and may not apply to all Pokémon or other datasets.

.

Next, we observe the distribution of the Pokémon abilities. The distribution of the top 20 Pokémon Abilities reveals these key points:

* 
* The most common ability is 'Unknown', with 764 Pokémon having this attribute. This is expected because it represents Pokémon that do not have a specifically defined ability.
* The next most common abilities are 'Swift Swim', 'Levitate', and 'Sturdy', each owned by more than 40 Pokémon. This could imply that these abilities have a greater distribution in the Pokémon universe (according to the dataset), or they might be foundational abilities beneficial in many circumstances within the gameplay.

Several abilities in the dataset are unique or exclusive to certain Pokémon or types. For instance, 'Multitype' is unique to Arceus, changing its type based on its held plate, and 'Delta Stream', exclusive to Mega Rayquaza, creates an air current that normalizes moves against Flying types.

A total of 156 abilities appear 10 times or less in the dataset. This high number of low-frequency abilities suggests a lot of diversity and specialization in Pokémon abilities, fitting with the wide array of Pokémon species and types.

Among Pokémon types, 'Keen Eye' is prevalent in Normal types, 'Intimidate' is common in 18 Pokémon, 'Swift Swim' is frequent in Water types, and 'Sturdy' is common in Rock types. 'Levitate' is common in Dragon and Psychic types, 'Static' in Electric types, 'Chlorophyll' in Grass types, and 'Sand Veil' in Ground types. Ice types often have 'Ice Body', Poison types have 'Poison Point', Fighting types exhibit 'Guts', and Fire types often have 'Blaze'. These abilities provide a competitive edge in battles.

The mean 'total\_points' of the Pokémon that have each ability is calculated. The abilities 'Delta Stream', 'Desolate Land', and 'Primordial Sea' have the highest mean total points, indicating that Pokémon with these abilities tend to have higher total points.

The study uncovers common abilities among Pokémon species. For example, 'Armor Pokémon' often have 'Sand Stream', 'Arrow Quill Pokémon' possess 'Long Reach', 'Artificial Pokémon' have 'Soul-Heart', and 'Astral Body Pokémon' display 'Competitive'. 'Worm Pokémon' showcase 'Shield Dust', 'Wrestling Pokémon' have 'Limber', 'Mold Breaker', or 'Unburden', 'Wushu Pokémon' are equipped with 'Unseen Fist', 'Young Fowl Pokémon' have 'Blaze' or 'Speed Boost', and 'Zen Charm Pokémon' possess 'Guts', 'Hustle', 'Inner Focus', or 'Zen Mode'.

The analysis offers insights into common abilities per Pokémon species, highlighting their strengths and diversity. It uncovers patterns and relationships among Pokémon abilities, evolution stages, and types, enriching our understanding of the Pokémon universe and laying groundwork for advanced analyses.

Refer to ‘Data\_Understanding\_2\_final.pdf’ in the appendix for the process and results.

# Type 1 and 2 Classification

We used Support Vector Machine (SVM) and Random Forest classifiers for predicting Pokémon types using a dataset loaded from a CSV file. This dataset includes a variety of features such as 'generation', 'status', 'height\_m', 'weight\_kg', 'abilities\_number', among others. The target variables are 'type\_1' and 'type\_2', representing the Pokémon's types.

The data preprocessing step involves one-hot encoding of the categorical variables in the feature set, converting them into a format suitable for machine learning algorithms. This is crucial for improving prediction results.

The model training and evaluation step involves splitting the data into training and testing sets. Both SVM and Random Forest classifiers are initialized, fitted to the training data, and used to make predictions on the testing data. The model's performance is evaluated using a classification report. The performance metrics vary across different classes, with some showing good performance and others showing poor performance, potentially due to class imbalance.

5-fold cross validation is performed for both target variables and classifiers, providing a robust measure of model performance. The SVM classifier shows a low average cross-validation score and accuracy for both 'type\_1' and 'type\_2'. It often correctly predicts a class but misses many instances of that class, indicating a high precision but low recall.

The Random Forest classifier performs better, with higher average cross-validation scores and accuracy for 'type\_1'. The precision and recall scores vary across different classes, with some classes showing high scores and others showing low scores. The macro average F1-score and the weighted average F1-score are higher than those for the SVM classifier, suggesting better overall performance.

The Random Forest classifier outperforms the SVM classifier. However, both could be improved. For the Random Forest classifier, balanced class weighting could address the class imbalance. For both, hyperparameter tuning could optimize performance, and feature selection or extraction could identify the most informative features to improve classifier performance. These steps could potentially enhance the performance of the classifiers.

Refer to ‘Data\_preparation\_+\_type\_Classification.pdf’ in the appendix for the process and results.

# Abilities Association Rule Mining

Apriori and FP-Growth algorithms are used for association rule mining on Pokémon abilities. The Pokémon dataset is prepared by combining abilities into a single list for each Pokémon and transforming the data into a list of lists.

The Apriori algorithm identifies frequent item sets (sets of abilities appearing together often) with a minimum support parameter of 0.01. Association rules are generated from these itemsets and sorted by the 'lift' metric.

The FP-Growth algorithm, faster than Apriori for larger datasets or lower minimum support values, is also applied with the same minimum support parameter. It generates association rules from frequent itemsets, sorted by lift.

The results are rules describing how certain abilities tend to appear together, such as 'Rock Head' and 'Sturdy'. These rules highlight patterns but don't imply causation. Metrics like 'conviction' and 'leverage' provide additional information about the strength and significance of these rules.

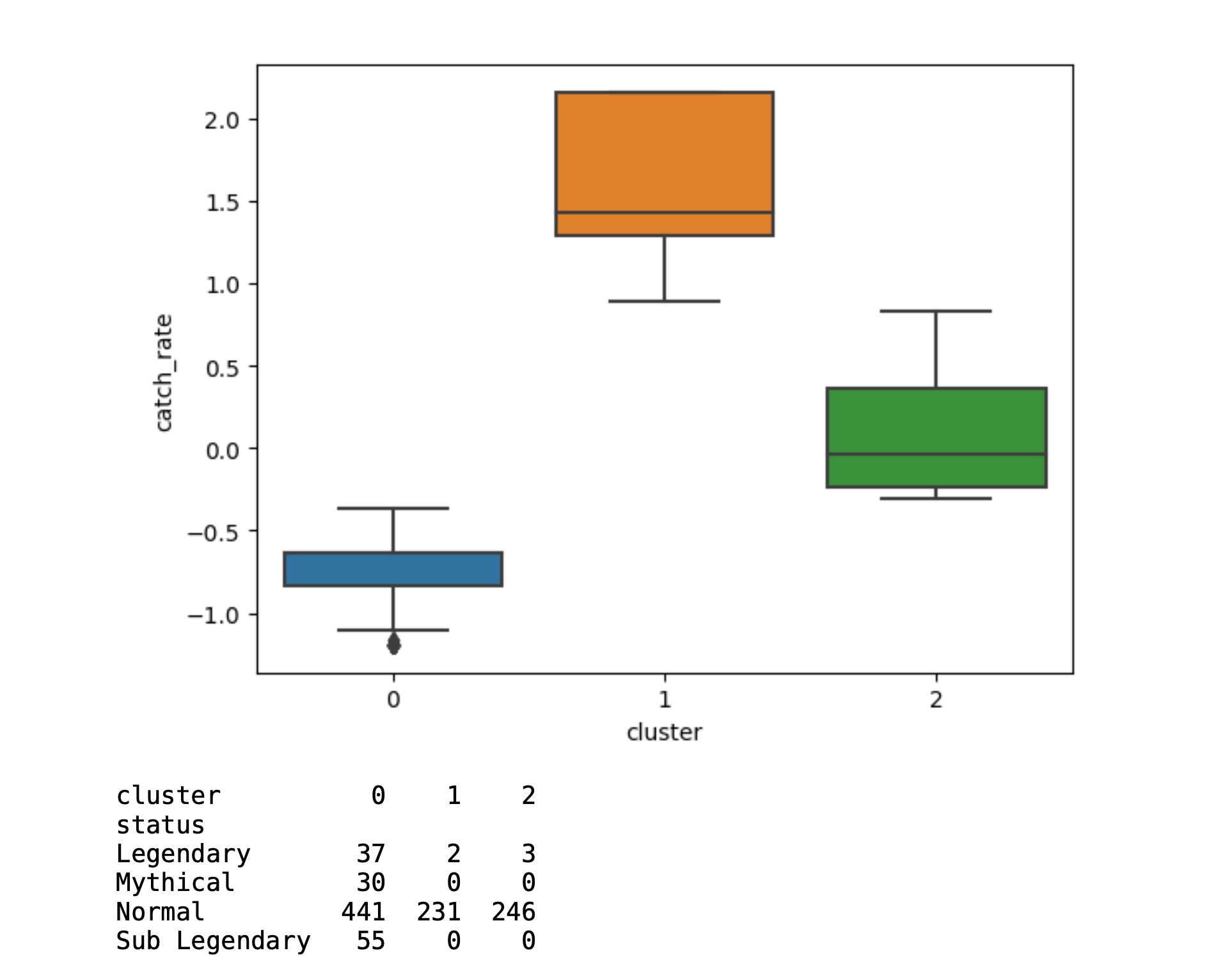
Both algorithms found similar associations with varying degrees of support, confidence, and lift. However, there are differences, such as the Apriori algorithm finding an association not present in the FP-Growth results. Both algorithms perform well in finding associations. FP-Growth is faster and requires less memory, making it suitable for larger datasets, while Apriori's simplicity makes it ideal for smaller datasets or educational purposes.

Refer to ‘Data\_Preparation\_+\_abilities\_Association\_Rule\_Mining.pdf’ in the appendix for the process and results.

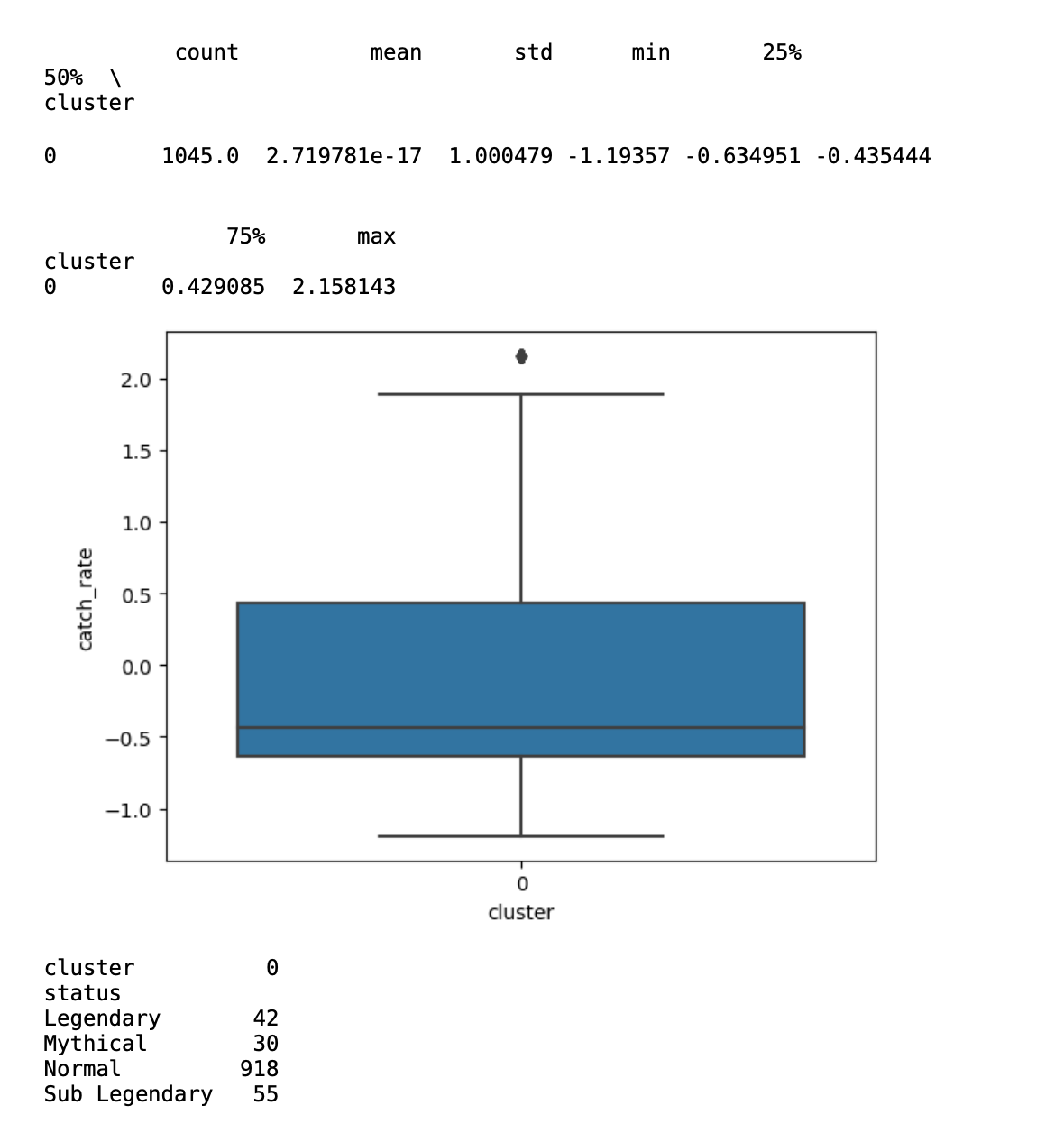
# Catch Rate/Rarity Clustering

Machine learning techniques, specifically K-Means, DBSCAN, and Hierarchical Clustering algorithms, are used to analyze Pokémon data with a focus on the 'catch\_rate' attribute. The data is prepared and standardized using the StandardScaler, a necessary step for certain machine learning algorithms.

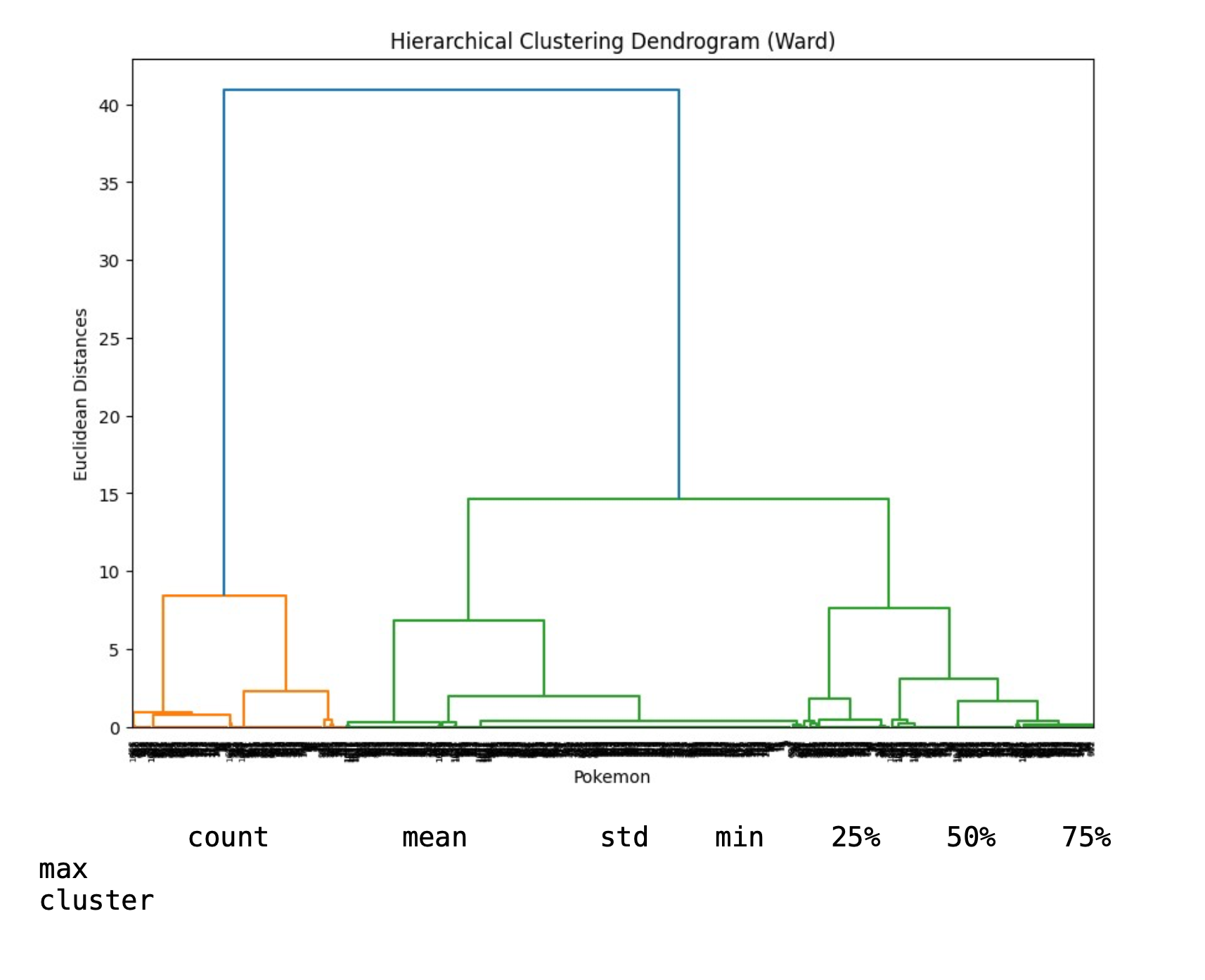
The K-Means algorithm, with the number of clusters set to 3, groups the Pokémon based on their catch rates. The clusters are visualized using a boxplot, and the distribution of Pokémon 'status' across the clusters is examined.

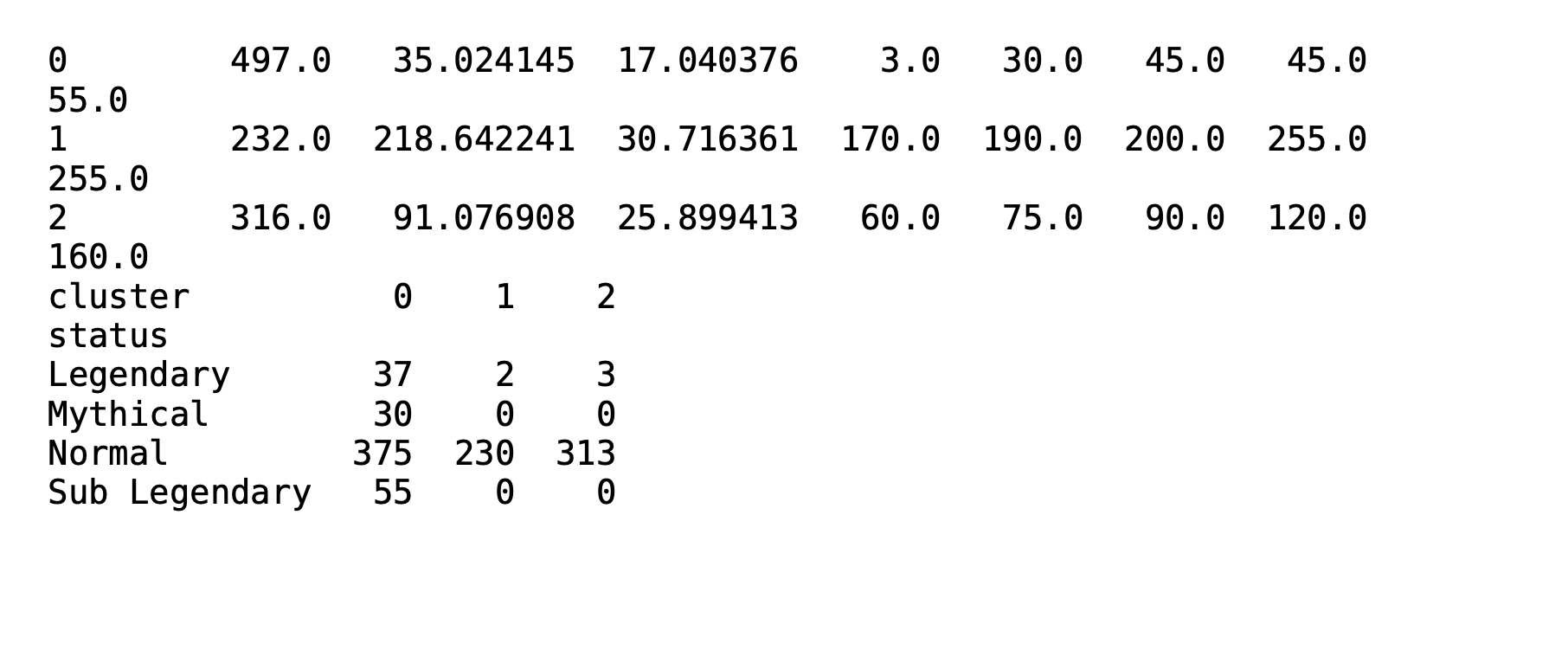


The DBSCAN algorithm, applied with an epsilon of 0.3 and a minimum sample of 10, groups Pokémon close in data space, marking outliers far from any cluster. It results in a single cluster with 1045 Pokémon. The 'status' of Pokémon in this cluster includes 42 Legendary, 30 Mythical, 918 Normal, and 55 Sub Legendary Pokémon.



Hierarchical clustering, performed using the 'ward' method, is visualized with a dendrogram and cut to obtain 3 clusters. Each cluster contains Pokémon of varying 'status' and has different mean catch rates and standard deviations.





Cluster 0 contains 497 Pokémon, including 37 Legendary, 30 Mythical, 375 Normal, and 55 Sub Legendary Pokémon with a mean catch rate of 35.02. Cluster 1 contains 232 Pokémon, all of which are Normal status with a mean catch rate of 218.64. Cluster 2 contains 316 Pokémon, including 3 Legendary, 313 Normal Pokémon with a mean catch rate of 91.08.

The findings provide insights into the distribution of catch rates among different Pokémon 'status', which could be beneficial for game strategists and players. Understanding these distributions could potentially enhance game strategies.

Refer to ‘Data\_Preparation\_+\_catch\_rate\_rarity\_Clustering.pdf’ in the appendix for the process and results.

# Total Points Regression

Regression models are used to predict Pokémon total points, a process that could aid trainers in identifying high-performing Pokémon. The clean Pokémon data is prepared with a focus on the 'total\_points' attribute. Categorical columns are label encoded and converted to numerical values. The target variable and feature matrix are then determined.

A correlation analysis is conducted between 'total\_points' and other features, visualized using a correlation heatmap. This analysis helps to understand the relationship between the features.

Several regression models, including Linear Regression, Ridge Regression, Decision Tree Regression, and Random Forest Regression, are applied to the data. These models predict the total points of Pokémon based on the other features.

The models are evaluated using two metrics: Root Mean Squared Error (RMSE) and R2-score. The model with the lowest RMSE value and the highest R2-score is considered the best.

Cross-validation is performed to enhance the model performance. However, the cross-validated RMSE is significantly higher than the RMSE on the training data, suggesting potential overfitting. This overfitting might be due to the size of the dataset, and it's suggested that using a larger Pokémon dataset could improve model performance.

Early stopping is suggested as a measure to avoid overfitting. This method stops the training process before the learner passes a certain point of overfitting, serving as a form of regularization.

Multiple regression techniques are applied to predict the total points of Pokémon based on their other features. The findings provide insights into the performance of different Pokémon, which could be beneficial for game strategists and players. However, the potential overfitting indicates that there's room for model improvement.

Refer to ‘Data\_Preparation\_+\_total\_points\_regression.pdf’ in the appendix for the process and results.

# Evaluation

The project successfully applied association rule mining to identify common associations among different Pokémon abilities, which can help trainers identify which abilities to prioritize when training their Pokémon. Both Apriori and FP-Growth algorithms were used, revealing similar associations with varying degrees of support, confidence, and lift. However, the results highlight patterns but don't imply causation.

The project used a regression model to estimate total points of Pokémon, which can assist trainers in identifying which of their Pokémon are most powerful and worth investing time and resources into training. Multiple regression techniques were applied, including Linear Regression, Ridge Regression, Decision Tree Regression, and Random Forest Regression. However, the cross-validated RMSE was significantly higher than the RMSE on the training data, suggesting potential overfitting. This indicates that there's room for model improvement.

The project employed clustering techniques to group Pokémon based on their catch rates and rarity, successfully distinguishing Pokémon clusters. K-Means, DBSCAN, and Hierarchical Clustering algorithms were used, each providing unique insights into the distribution of catch rates among different Pokémon 'status'.

The project used classification algorithms to predict a Pokémon's rarity and type, which can aid trainers in battles. Both SVM and Random Forest classifiers were used, with the latter outperforming the former. However, both could be improved through balanced class weighting, hyperparameter tuning, and feature selection or extraction.

# Conclusion

The project successfully applied various data mining techniques to the Pokémon dataset, providing valuable insights into Pokémon characteristics and their interrelations. These insights can be highly beneficial for game strategists, players, and enthusiasts. However, to fully meet the success criteria, the project should address the potential overfitting in the regression model, and improve the performance of the classification models. This could potentially enhance the performance of the classifiers and the overall success of the project.

# Reference List

University of Arizona. (n.d.). Pokémon Types and Stats. Retrieved from <https://www2.cs.arizona.edu/classes/cs120/fall17/ASSIGNMENTS/assg02/pokemon.html>

Analytics Vidhya. (2021). Pokémon Prediction Using Random Forest. Retrieved from <https://www.analyticsvidhya.com/blog/2021/12/pokemon-prediction-using-random-forest/>

Dhama, S. (n.d.). Exploratory Data Analysis On Pokémon Dataset. Retrieved from <https://github.com/ShaileshDhama/Exploratory-Data-Analysis-On-Pokemon-Dataset>

In Machines We Trust. (n.d.). Exploring Pokémon Dataset. Retrieved from <https://inmachineswetrust.com/posts/exploring-pokemon-dataset/>

Kaggle. (n.d.). Linear Discriminant Analysis with Pokémon Stats. Retrieved from <https://www.kaggle.com/code/residentmario/linear-discriminant-analysis-with-pokemon-stats>

Kaggle. (n.d.). Decision Tree Examples. Retrieved from <https://www.kaggle.com/code/mlomuscio/decision-tree-examples>

Medium. (n.d.). EDA Pokémon Visualization. Retrieved from <https://medium.com/deep-dive-into-exploratory-data-analysis/eda-pokemon-visualization-58102b68e2d7>

Medium. (n.d.). Pokeml: An Introduction to Data Science Using Pokémon. Retrieved from <https://medium.com/@patrickmartinaz/pokeml-an-introduction-to-data-science-using-pokemon-c5df8d6decd0>

Dataversity. (n.d.). 10 Key Data Mining Challenges in NLP and Their Solutions. Retrieved from <https://www.dataversity.net/10-key-data-mining-challenges-in-nlp-and-their-solutions/>

Towards Data Science. (n.d.). Overcoming Challenges When Designing a Fraud Detection System. Retrieved from <https://towardsdatascience.com/overcoming-challenges-when-designing-a-fraud-detection-system-29e33bcb43d2>

Medium. (n.d.). Data Analysis on Pokémon Dataset. Retrieved from <https://medium.com/mlearning-ai/data-analysis-on-pokemon-dataset-44cdc7d15e56>

Pokémon Go. (n.d.). Retrieved from [https://www.pokemongo.com](https://www.pokemongo.com/)

Pokémon Database. (n.d.). Retrieved from <https://pokemondb.net/>

Pokémon Company. (n.d.). Retrieved from <https://www.pokemon.com/>

Niantic. (n.d.). Retrieved from <https://nianticlabs.com/>

Nintendo. (n.d.). Retrieved from https://www.nintendo.com/

Tajiri, S. and Sugimori, K., 1996. Pokémon Franchise. Nintendo. (<https://en.wikipedia.org/wiki/Pok%C3%A9mon>)

Agrawal, R., Imieliński, T. and Swami, A., 1993. Mining association rules between sets of items in large databases. In: Proceedings of the 1993 ACM SIGMOD international conference on Management of data, pp.207-216. (<https://dl.acm.org/doi/10.1145/170035.170072>)

Breiman, L., 2001. Random forests. Machine learning, 45(1), pp.5-32. (<https://link.springer.com/article/10.1023/A:1010933404324>)

Cortes, C. and Vapnik, V., 1995. Support-vector networks. Machine learning, 20(3), pp.273-297. (<https://link.springer.com/article/10.1007/BF00994018>)

James, G., Witten, D., Hastie, T. and Tibshirani, R., 2013. An introduction to statistical learning. Springer. (<https://www.springer.com/gp/book/9780387848570>)

Kohavi, R., 1995. A study of cross-validation and bootstrap for accuracy estimation and model selection. In: Ijcai, Vol. 14, No. 2, pp.1137-1145. (<https://www.ijcai.org/Proceedings/95-2/Papers/022.pdf>)

Wikipedia contributors (2023) 'Pokémon Go', Wikipedia, The Free Encyclopedia, 9 June. Available at: <https://en.wikipedia.org/wiki/Pok%C3%A9mon_Go> (Accessed: 9 June 2023).

Hamari, J., Hanner, N., & Koivisto, J. (2017). Service quality explains why people play games and why they stop playing them. Journal of Electronic Commerce Research, 18(2), 122. (<http://www.jecr.org/node/576>)