School of Computing Electronic
Engineering and Mechanical & Manufacturing Engineering

Data Warehousing & Data Mining
Continuous Assessment

Name	Course	Student ID
Cormac Duggan	CASE	17100348
Gergely Gellert	CASE	17379616

Plagiarism Statement:

I understand that the University regards breaches of academic integrity and plagiarism as grave and serious.
I have read and understood the DCU Academic Integrity and Plagiarism Policy. I accept the penalties that may be imposed should I engage in practice or practices that breach this policy.
I have identified and included the source of all facts, ideas, opinions, viewpoints of others in the assignment references. Direct quotations from books, journal articles, internet sources, module text, or any other source whatsoever are acknowledged and the sources cited are identified in the assignment references.
I declare that those sections, which I now submit for assessment, that I have been required to write individually, are entirely my own work and have not been taken from the work of others save and to the extent that such work has been cited and acknowledged within the text of my work.
I declare that those sections, which I now submit for assessment, that I have been required to write as part of a group, are entirely the work of my group and have not been taken from the work of others save and to the extent that such work has been cited and acknowledged within the text.
I have read and understood the DCU library referencing guidelines (available at: http://www.library.dcu.ie/LibraryGuides/Citing&ReferencingGuide/player.html) and/or recommended in the assignment guidelines and/or programme documentation.
By submitting this material online I confirm that this assignment, or any part of it, has not been previously submitted by me or any other person for assessment on this or any other course of study.
By submitting material for assessment online I confirm that I have read and understood the DCU Academic Integrity and Plagiarism Policy (available at: https://www4.dcu.ie/sites/default/files/registry/docs/IntegrityPlagiarism.pdf)

Section 1

For our data classification project we have decided to take the stats and historical placement of Pokemon in different tiers of the competitive metagame. In the Pokemon video games there are over 800 assorted monsters to choose from, each of which has different stats, types, abilities, and movesets. With the correct combination of all of those a single Pokemon can go from competitively useless to AG tier (AG meaning it is so good it is banned from competitive usage).

We acquired our first dataset from kaggle and it had the majority of what we needed formatted incorrectly for how we wanted to use it. The original dataset provided 51 columns which we cut down to 14 as the rest were negligible when determining competitive viability. We originally wanted to include movesets as they are a large determining factor in tier placement but then realized it would be about 800 columns worth of data to include movesets and an 800x800ish dataset seemed a little more than unrealistic for the scope of this particular project. To make the data set more usable for ourselves when trying to make placement decisions we redetermined the numbers provided in each of the stat columns we had and replaced them with Ubers, OU, UU, RU, NU, PU, and LC as more general groupings.

The kaggle dataset also did not include one of our most important columns which was competitive tier. A website called smogon keeps track of all the competitive usage rates and tiers of each Pokemon since 2014 or earlier. As the kaggle dataset we found was from 2017 we took the competitive set from that year to determine the tier of each Pokemon. After ripping all the necessary data from the smogon datasets we added usage rates and tier to our dataset. A few anomalies came up because sometimes people play with unpredictable strategies, but for at least 99% of the Pokemon, the smogon dataset placed them into appropriate tiers. With the tiers and usage acquired we added those two columns to the rest of the data giving us the set we would be working with.

Name	Type 1	Type 2	Stat Total	Hit Points Stat	Attack Stat	Defence Stat	Special Atk Stat	Special Def Stat	Speed Stat	Resistances	Weaknesses	Usage in Tier	Tier

For our classification algorithm we will be taking into account the above categories which are black to determine the appropriate competitive tier the Pokemon would likely be placed in if they were to be added to the game. The name being a unique identifier won't be taken into account as it should not affect the outcome. It is there merely to keep a complete dataset and for testing purposes. The completed classification will place Pokemon in one of the following tiers. Ordered from best to worst the tiers are; Ubers, OU, UU, RU, NU, PU, LC.

Section 2

One of the first things we did when we got our hands on our data was see how various stats for each pokemon are distributed. To our surprise we found that all of the pokemon's stats almost perfectly followed a positively skewed bell curve distribution. This surprised us because each pokemon has a seemingly randomly allocated stats for each of their "base stats". Despite this all the stats seem to take on a positively skewed bell curve. This makes sense because of the amount of pokemon available in game. In the game there's a limited amount of "specialised" pokemon. For example if you have a high Attack or Special Attack it would be beneficial to also have a high speed stat so you can hit the opposing pokemon hard, potentially knocking it out before it has a chance to retaliate and knock you out in return. In these cases stats such as Defense or Health Points (HP) aren't as important as this certain Pokemon if used successfully can simply just knock out it's opponent in one hit before it can take any damage itself as retaliation.

This caused us to value some combination of typings more when it came to deciding what competitive tier a pokemon should land in. For example for a pokemon that has a fire typing, having either a high Attack or Special Attack stat, paired with a high Speed stat would be beneficial since Fire has a good matchup against a higher than average number of other typings, hitting them either for normal damage or increased damage when used. Along with this, a pokemon that uses a Fire type attack while also having a Fire typing gains a 1.5 times damage boost to that attack, hence why it's good to have this typing. To incorporate this into our tier predictions we decided to give any types that are deemed to have an "offensive" typing a boost to its Attack or Special Attack weighting. We didn't give the pokemon's speed weighting a boost because there are some good pokemon that have offensive typing but have a low speed stat but usually it's recommended to have a high speed stat.

In a similar vein if a pokemon has typing that has been deemed to be a good defensive typing we have given its Health Points weighting a boost and either it's Defence or Special Defence weighting depending on which is better. We gave these weightings a softer boost compared to Attack or Special Attack weighting as we're boosting 2 stats weighting instead of just 1.

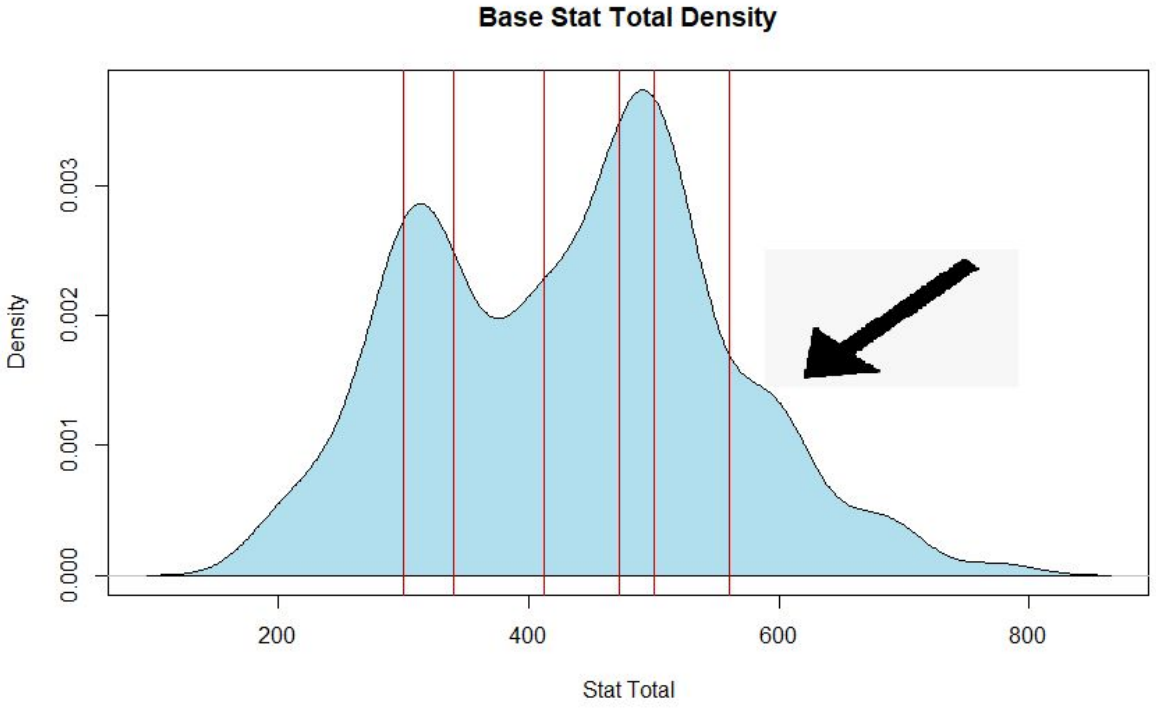
On the contrary we weren't really surprised about how the graph turned out for the base stat totals. Base stat is when you add all of the stats such as Health Points, Attack, Defence etc. together as a singular number. In pokemon, most pokemon can "evolve" to become more powerful. This, in terms of numbers and statistics though, means the numbers behind each of it's stats get modified, usually boosted to a higher number. Sometimes the pokemon can gain a new typing to which factors into our weightings explained above. Usually these "evolved" pokemon have a base stat number they reach for example 400, that is then used to determine how much Health that pokemon has or how hard it can hit with an attack.

The first peak of the data on the left side is usually what unevolved pokemon, also known as basic pokemon base stats add up to. This should make up most of

our “lower tiered” pokemon as they won’t have outstanding stats. Although due to their typing and some specialised stat distribution they could potentially be placed in a higher tier.

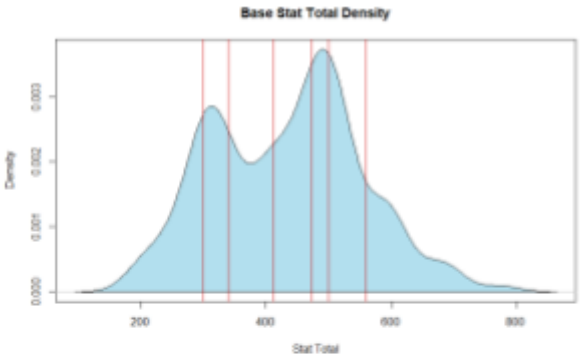
The second peak of the data is pokemon that have either evolved once or pokemon that don’t evolve at all leading them to have high natural base stats. This tier is the highest since these types of pokemon are the most common as both standalone pokemon and intermediate pokemon that go on to evolve once more reside in this base total.

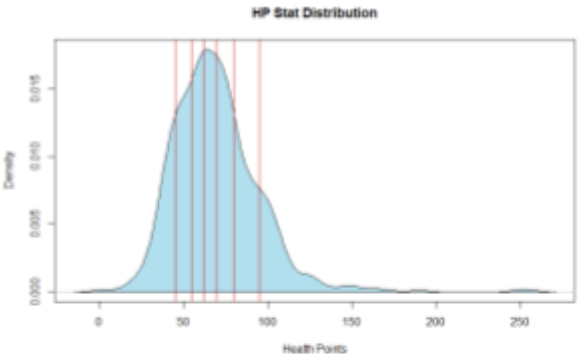
The third notable feature we managed to identify in the graph is this bump. This bump as shown below represents the pokemon that have evolved for the second time and a class of pokemon known as legendary pokemon. Legendary

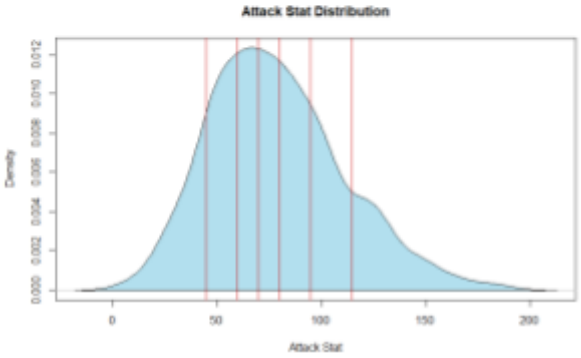


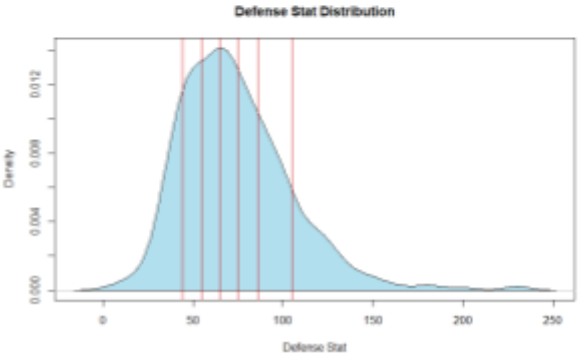
Pokemon are a special subsection of the total list of Pokemon. Usually one or two legendary Pokemon are the main focus of a particular pokemon game and obtaining them is one of the main objectives of the game. To honour this legendary status they’re given higher than normal stats, often at times have the highest base stats in their respective games.

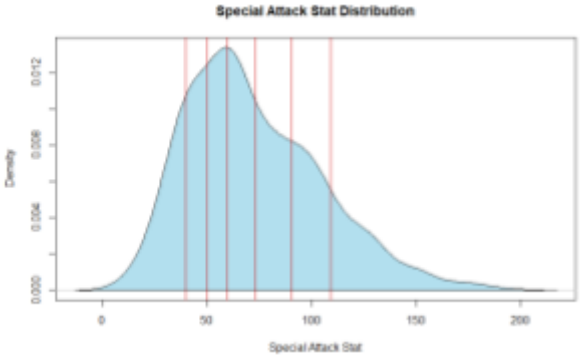
The red lines present in all of the diagrams denote the upper threshold for the average level of base stats to be in the allocated tiers. Because of these tier divisions we concluded that we should use percentile range partitioning when we’re predicting what tier pokemon will end up in. We choose this because the data shows that the higher your stats are, along with having good typing to complement your stats the higher ranked tier the pokemon will end up in.

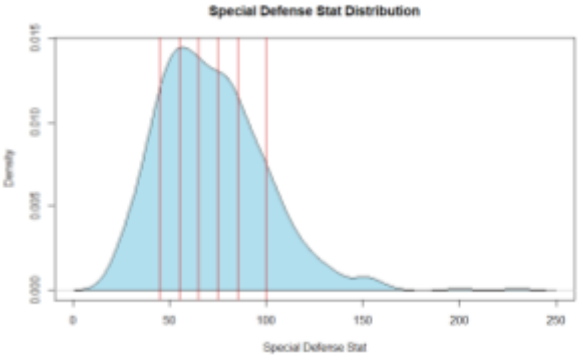

Base Stat Total Density
Density
0.000
0.001
0.002
0.003
200
400
600
800
Stat Total

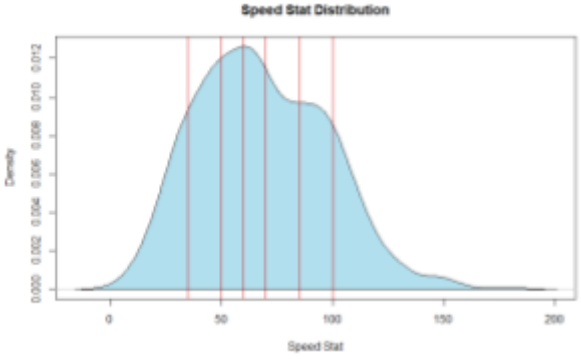

HP Stat Distribution
Density
0.000
0.005
0.010
0.015
0
50
100
150
200
250
Health Points


Attack Stat Distribution
Density
0.000
0.002
0.004
0.006
0.008
0.010
0.012
0
50
100
150
200
Attack Stat


Defense Stat Distribution
Density
0.000
0.004
0.008
0.012
0
50
100
150
200
250
Defense Stat


Special Attack Stat Distribution
Density
0.000
0.004
0.008
0.012
0
50
100
150
200
Special Attack Stat


Special Defense Stat Distribution
Density
0.000
0.005
0.010
0.015
0
50
100
150
200
250
Special Defense Stat

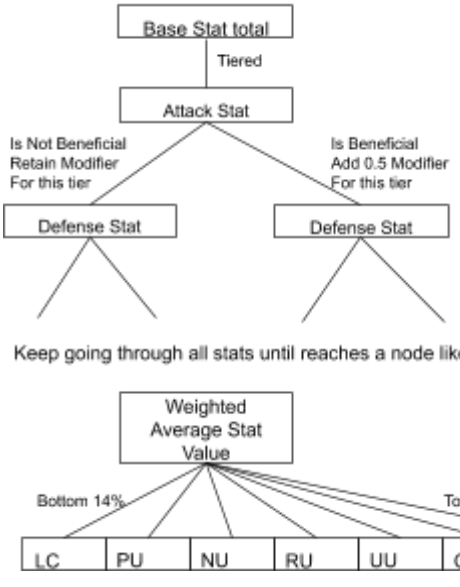

Speed Stat Distribution
Density
0.000
0.002
0.004
0.006
0.008
0.010
0.012
0
50
100
150
200
Speed Stat

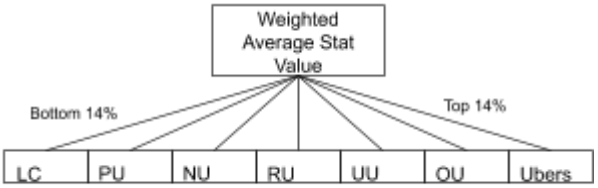
Section 3

After finishing the initial analysis of the data and formatting the dataset appropriately for our usage we created a decision tree that would use a few different partitions to attempt to determine the appropriate competitive tier for each Pokemon. Using the dataset in which we replaced the numerical stat values with tier names we partitioned the stats based on whether or not the highest stats an individual Pokemon had were beneficial to their typing or not. This decision was made by taking into account the most common move types for the given typing of the Pokemon (each move is either Special or Physical relating to the two attack stats each pokemon has). For example: fire type moves typically are either special attack based or attack based moves. As a Pokemon gains an advantage when attacking with the proper move type and having the first turn it was determined that speed and either special or physical attack stats would be most beneficial for fire type Pokemon.

We went on to determine which stats were most beneficial for each typing and gave modifiers to the stats based on how useful they were for the type. If a Pokemon had a beneficial stat, the value weight for that stat would gain an additional 0.5 point modifier. Otherwise each stat was valued based on tier; LC being worth 1 point, PU at 2 point, NU at 3, RU at 4, UU at 5, OU at 6, and Ubers at 7. A weighted average was taken to determine the overall stat viability of each individual Pokemon using the modifier and the points awarded for each stat. As a simple explanation: Pokemon X has a fire typing with 4 stats that are tiered in Ubers, 3 that are in OU, and 2 in NU with 1 of the 4 Ubers stats being valuable to its typing. This would result in (4.5(7) + 3(6) + 2(3))/9 as the stats weighted average.

After getting the stat weighted average for each Pokemon we partitioned our decision tree to categorize them into tiers. As our sample set was 798 and we have 7 tiers each tier would consist of approximately 114 Pokemon. The Uber tier would consist of the top 114 stat weighted averages while OU would contain the next 114 and so on down the list. The extended decision tree would come to look something like the first example while the code and major partitionings of the tree would suggest it looked more like the second.





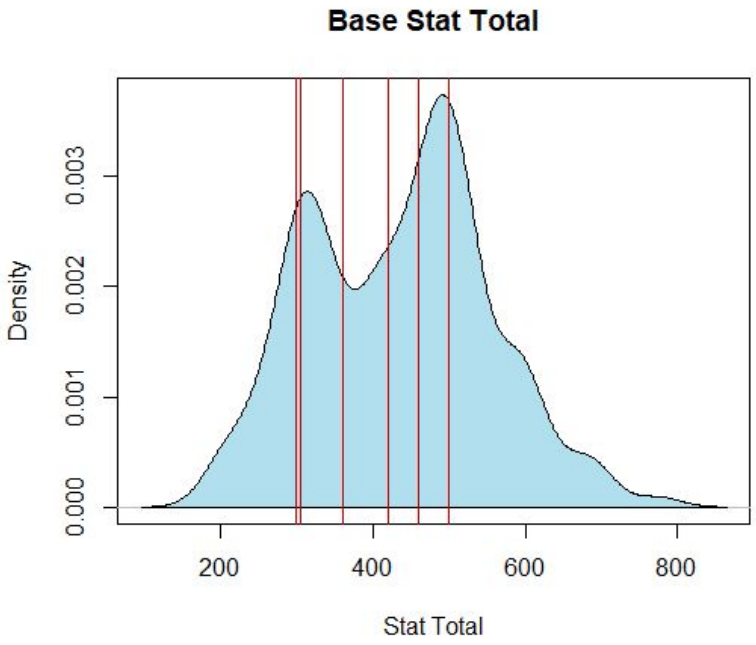
Section 4

Initially when looking at the result of our data we were rather disappointed as looking solely at percentage accuracy it was pretty inaccurate. However, upon second looks and more in depth analysis we became more content with the percentage we received.

To determine the accuracy of our data at the end we ran every data entry we had through the algorithm then compared the output classification with our previously stored classifications that we had obtained from the smogon website. The classification results had an accuracy of 21% which initially looks really low. Based on our disappointment we were determined to find out why we were so inaccurate. Coming back to the fact that we had 7 different tiers we determined that each Pokemon had a better chance of being placed in the proper category than if we were just to roll a 7 sided die. This was our first glimmer of hope that our algorithm might not be as terrible as it looks.

Following this we wanted to determine how accurate the classification was in terms of wrong classifications. Using a numbering system we calculated that the average standard deviation for an incorrect classification was + or - 1 tier. This information was significantly more encouraging for the both of us. This meant that on average for the 79% that came back incorrect they were typically only incorrect by 1 tier out of 7. This would be extra encouraging because as long as it is a positive rating (meaning an OU classified as Ubers) it is still usable. If it were a negative classification the Pokemon wouldn't be allowed into the tier it was classified as.

When trying to determine the issues with our classification algorithm and why the accuracy was low the first thing that came to mind was our distribution for typings. As most types had two stats that could be beneficial and get boosts but fire and grass were anomalies we looked at them first. The fire and grass type were considered "neutral attackers" so they could have either special or physical attack as well as speed and receive a boost to their score over all. It seems because of this leniency on those two typings the accuracy for them was greatly decreased to a 15.8%. Additionally the average deviation for inaccurate answers was +2 tiers meaning that almost 85% of Pokemon that had the fire or grass typing would be placed about 2 tiers higher than they were supposed to be. As is evident in the following graph the base stat expectation for these types are far lower than average when compared to the original distribution in section 2.



The tiers from the above graph show the significant benefit of being a fire or grass type when being placed into a tier. While the cutoff for LC remains in about the same position for Base Stat Total, every other tier is moved back. The least viable Pokemon remain not viable because their stats are just too low but following that the PU tier is almost nonexistent and every following tier cutoff is 1 tier behind where it was when taking into account all Pokemon.

From this information we can assume that other typings have somewhere between a 23 and 26% accuracy to compensate for these two typings which is better than we had expected. Apart from these two types, however, our biggest issues for accuracy probably came from the bonus Pokemon received for beneficial typings. While it is a necessity it can perfectly portray usage or how viable Pokemon are. Simply put it is a vast oversimplification of the mechanics of battling in the game. For example, while we decided that psychic types were special attackers and flying types were physical attackers, neither of these can be completely accurate. If we take Lugia, a legendary Pokemon introduced in 2000, it has the psychic and flying types, but it's stat distribution makes it a very good defensive Pokemon. Our method of analysis can't determine this anomaly for the typing as it is a simplified classification. While Lugia has very high stats and was placed in the Uber tier anyway, other type anomaly Pokemon would likely not be so lucky to be placed properly.

Apart from what was previously mentioned there is also the issue of how we actually partitioned our tiers. In the actual competitive meta there is no basis that defines tier based on percentile ranges. The actual tier list is more of a distribution of usages than percentiles. One key example being the Uber tier which actually only contains 47 Pokemon and not 114. This discrepancy can be seen across all tiers as the actual size of each tier is not going to be exactly 114 and depends on where

each Pokemon gets utilized most. For the sake of partitioning we didn't place tiers by how many they actually contain as of right now and instead by percentile. This is almost definitely a cause of the low percentage for accuracy.

On top of the other issues that our algorithm faced it also had the challenge of predicting human interaction with the game and strategy. A large amount of Pokemon end up getting placed in higher tiers than the stats would predict simply because people like the way they look or have found unique strategies to use them in. A perfect example of this would be what is called the F.E.A.R. strategy. This takes one of the lowest stat total Pokemon in the game and utilizes their lack of health and an attack called endeavor to bring any opponent that doesn't have the ghost typing down to 1 health point immediately. To even think about incorporating usages like these into our algorithm would take far more time and computing power than is realistically available to us for this project.

The last issue would be all the other things that determine a Pokemon's tier including items, abilities, and unusual stat distributions. While all attacking moves will have either a special or physical basis and by taking the majority available to each typing we determined the bonuses. However there are also set up and defensive moves that Pokemon can get access to which make them more viable competitively. To implement this into our algorithm would make our dataset 800x800 instead of 13x800. Moves like 'protect' which make you invulnerable to the enemy for a turn giving you a turn scouting cant be determined as a benefit or not when only taking into account the stats of an individual Pokemon. Items that give you a boost to speed at the cost of only being able to choose 1 move can not be algorithmically determined for value. Abilities that make you gain stat boosts in certain weather, give you health upon taking damage from a certain type of move, or give you the ability to survive any move that would otherwise have one hit knocked you out cant have their viability determined by an algorithm. The only way to determine how each of the compounding variables plays into the viability of any given Pokemon is simply to play around with them and see what works.

In the end it can be said that placing Pokemon based on their typing, stats, and resistances and weaknesses does not provide enough detail into the complexities of the game to be conclusive or even decently accurate. We did our best to work with the most numerical portions of what make Pokemon competitively viable, but in the end it came up short and our algorithm was only 21% accurate in deducing 1 of the 7 tiers for each. Some minor tweaking to the values and exact partitioning methodology may have given us an accuracy rate of closer to 30% but it is doubtful that we could have gotten it much higher than that. The overall complexity of the game, especially when played competitively and against people who spend hours analysing the best abilities, movesets, and items thus min maxing every aspect of the game just leaves too much up to the player. Algorithmically it is unreasonable to assume a Pokemon can accurately be placed in a tier that would relate to what the community deams it to be.