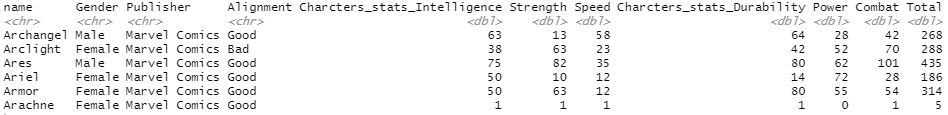
Across many planets and galaxies alike there will continue to be the debate of which is better: Marvel or DC. I decided to take this age-old question and put a spin on it to determine if we can differentiate between the two. To do so, I am going to conduct a series of experiments using data pertaining to various superheroes/villains of both Marvel and DC. Specifically, I am utilizing two datasets from Kaggle both which contain the name of the character, the publisher, their alignment, and gender. One of the datasets, the Score set has mostly numeric variables and the other, Power set, has a majority of categorical ones.

The first dataset found [here](https://www.kaggle.com/dannielr/marvel-superheroes?select=charcters_stats.csv) contains a score from 1-100 for traits such as speed, durability, strength and combat.

Score Set:

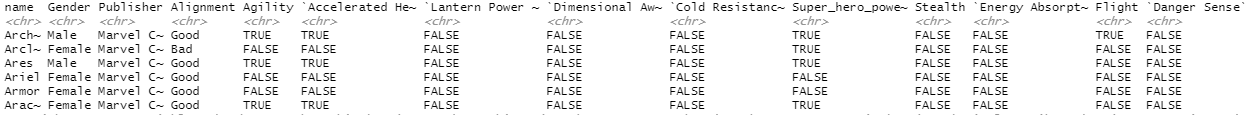
* 418 instances, 11 variables
* Features: 9 (2 categorical, 7 numeric) (No missing values)
* Target: categorical (Publisher / Alignment)
* Metas: string (Name)



The other dataset found [here](https://www.kaggle.com/dannielr/marvel-superheroes?select=superheroes_power_matrix.csv) contains upwards of 169 Boolean values for various powers such as Agility, Flight, Stealth and Durability.

Power Set:

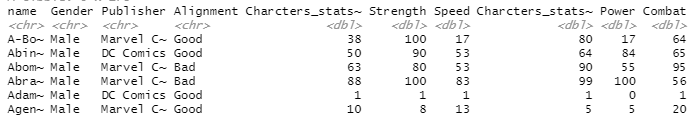
* 418 instances, 171 variables
* Features: 169 categorical (No missing values)
* Target: categorical (Publisher / Alignment)
* Metas: string (Name)



I then utilized Microsoft Access to join both the Score and the Power dataset to create a Master dataset which will be utilized in my experiments.

Master Set:

* 418 instances, 178 variables
* Features: 176 (169 categorical, 7 numeric) (No missing values)
* Target: categorical (Publisher / Alignment)
* Metas: string (Name)



Consistencies: Gender (categorical) Name (string) , Publisher & Alignment (one is target, other is just feature, depending on the experiment, both categorical)

To conduct each experiment, I imported all of the datasets into Orange where I then set up classification algorithms including SVM, kNN, Random Forest and Naïve Bayes to determine which model and dataset provides the greatest classification accuracy. The set up is as follows:

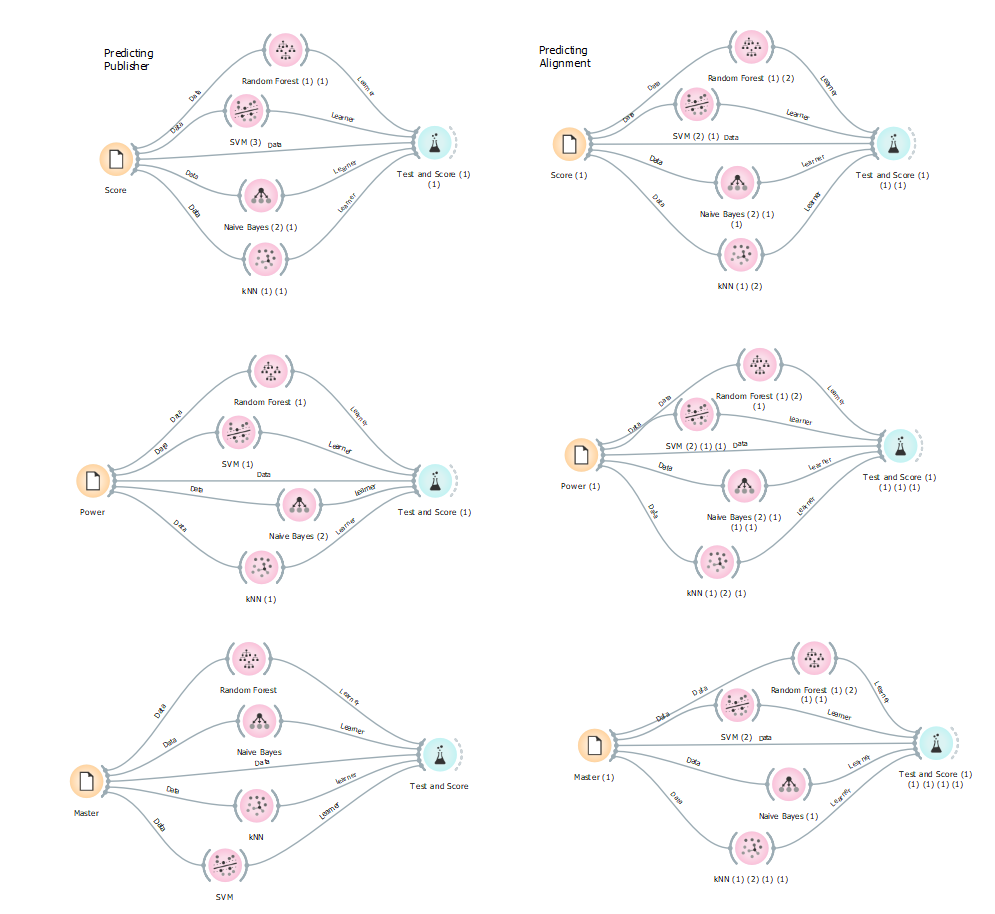
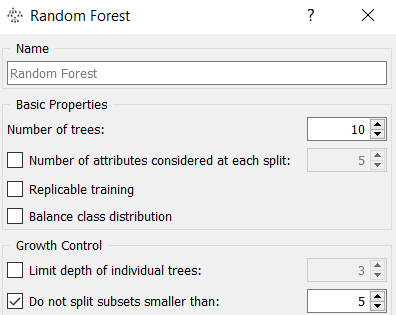
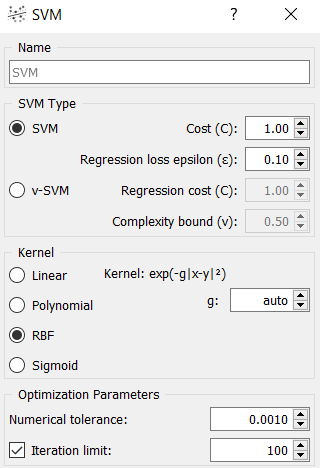
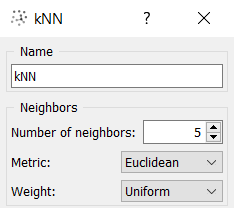


Figure 1

To maintain consistency across all experiments, each of the algorithms are set to the following parameters:

For SVM, I have C at 1 and Regression loss epsilon at .1 since this is a 2D dataset. Random Forest has been set for 10 trees and subsets no smaller than 5. For kNN, I have it k set to 5 with a Euclidean distance. Naïve Bayes does not require tuning for these experiments.





Next, I evaluated the various models on a 3-fold cross validation to determine how they can predict publisher and alignment for the power dataset, the rating statistics dataset, and the combined dataset. This is done by the test and score buttons which can be seen in each experiments design in Figure 1. So, in total, my experiment includes 6 different parts which we will now determine which is the best set of data and algorithm to predict both publisher and alignment based on our classification accuracy results:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Experiment** | **Random Forest CA** | **Naïve Bayes CA** | **kNN CA** | **SVM CA** |
| Experiment 1: Predict Publisher with Score Dataset | 62.4% | 58.4% | 64.8% | 62.9% |
| Experiment 2: Predict Alignment with Score Dataset | 62.2% | 32.3% | 61% | 65.1% |
| Experiment 3: Predict Publisher with Power Dataset | 67.2% | 65.6% | 66.3% | 68.4% |
| Experiment 4: Predict Alignment with Power Dataset | 61% | 14.8% | 62.4% | 66.7% |
| Experiment 5: Predict Publisher with Master Dataset | 67.5% | 65.6% | 68.4% | 69.1% |
| Experiment 6: Predict Alignment with Master Dataset | 64.6% | .5% | 63.9% | 68.2% |

When analyzing our results, it is clear that the SVM algorithm yielded the greatest percent of classification accuracy. In fact, the only experiment where the greatest percentage of classification accuracy wasn’t from the SVM algorithm was in Experiment 1, which has a kNN classification accuracy of 64.8%. I can speculate that the reason this test was more effective for Experiment 1 was due to the fact that the kNN algorithm weights all of its variables the same, and therefore weighted Publisher and Alignment the same. Also, it is possible that the algorithm was able to associate the various scores with a specific publisher as this variable could be more distinct compared to alignment, which may be tougher to decipher between as there are 3 possible answers, good, bad and neutral. With that said, the overall results of the Power dataset are greater than the Score, as it is possible the scores may lack consistency as they were derived based on the creator of the datasets valuation of each character. The power dataset, being Boolean, ensures us of a concrete answer for each character (true or false) based on whether the character exhibits the trait in the slightest bit.

In comparing the various classification algorithms and their accuracy in predicting both Publisher and Alignment, Naïve Bayes has proven to be least effective in working with all three datasets. This algorithm is the fastest and works best on smaller datasets. It is also intended on being useful when working with numeric values. This could explain why the classification accuracy for Experiment 6 was the lowest of all. Furthermore, for a classification problem, Random Forest gives you the probability of belonging to a class, while SVM gives you the distance to the boundary. This is seen based on the classification accuracy for Random, Forest, as it is greater for experiments predicting publisher than those for alignment. SVM, on the other hand, appears to be effective in establishing the boundaries of good, bad and neutral when classifying alignment although the target variable is assumed to be binary in an SVM test.

We can conclude that the Master dataset is most effective, as Experiments 5 and 6 have the greatest classification accuracy percentage for predicting Publisher and Alignment respectfully.

|  |  |
| --- | --- |
| Experiment | SVM CA |
| Experiment 5: Predict Publisher with Master Dataset | 69.1% |
| Experiment 6: Predict Alignment with Master Dataset | 68.2% |

In summary, The Random Forest and kNN algorithms had the results most comparable to SVM, which ranged from 2-5% less than the classification accuracy of the SVM algorithms per the majority of experiments. The Random Forest algorithm is intended for classification, similar to the Decision Tree, but it builds multiple trees to determine the best way to classify the data, which is why it is significantly greater. As for how it compares to kNN and SVM, Random Forest is better for comparing both numeric and categorical data, where the others are better for numeric specific data, like the set we are working with.

With that said our classification accuracy tells us that the SVM algorithm is most effective when predicting the publisher, Marvel or DC, using both the Power and Score data. I took some time to adjust the parameters of the SVM test understanding that it may be possible to obtain a greater classification accuracy if I adjust the kernel, however, I only managed to obtain lower accuracies. Based on the kNN results one can assume that our classes are quite separable, yet altering the number of k does not provide better results. It is important to note that SVM is better for distinguishing between two objects (i.e. Publisher), rather than Alignment which has three objects. Overall, these are just some assumptions as to why the SVM algorithm yielded the highest classification of all the algorithms used across the experiments.

For one final investigation, I decided to see if we can determine whether the 16 ‘neutrally’ aligned characters would be classified as good or bad based on the Master set which yielded the greatest accuracy. To do so, I split the dataset and set up the prediction for the 4 classification algorithms. The results are as follows:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **name** | **SVM** | **Random Forest** | **Naive Bayes** | **kNN** |
| Bizarro | Good | Bad | Bad | Good |
| Blackwulf | Good | Good | Bad | Good |
| Deadpool | Good | Good | Bad | Good |
| Deathstroke | Good | Good | Bad | Bad |
| Galactus | Good | Bad | Bad | Good |
| Juggernaut | Good | Good | Bad | Good |
| Lobo | Bad | Bad | Bad | Bad |
| Raven | Good | Good | Good | Good |
| Red Hood | Good | Good | Good | Good |
| Red Hulk | Good | Good | Bad | Good |
| Sandman | Good | Good | Bad | Good |
| Sentry | Good | Good | Bad | Good |
| Sinestro | Bad | Good | Bad | Good |
| The Comedian | Good | Good | Good | Good |
| Toad | Bad | Good | Bad | Bad |
| Trickster | Good | Good | Good | Good |

Of the 16 characters, 6 of them have 3 of the 4 tests with the same alignment. An additional 5 characters have a conclusive alignment based on the unanimous results of all the algorithms. This tells us that although their alignment is considered neutral, the majority of these characters are leaning in one direction of good or bad. It is important to note that there is a percentage of probability associated with each of these alignments for each algorithm. I have included this breakdown below as it is vital in understanding the confidence of each character’s alignment.

