Hello, Welcome to John Le's Data Science Assessment Project for Evil Genius

STEP 1: DATA EXTRACTION

The data from starcraft_player_data.csv is partitioned into training & testing data.
Every 5th entry is considered test data, the rest is training data.

Data is reformated & normalized.

STEP 2: Determine which attributes/features to use

Using all attributes, the training data is used to build 3 different models (KMeans Clustering, BIRCH Clustering, Decision Tree Classification).

The testing data is used to calculate the accuracy of each model. Results can be seen below.
KMeans Accuracy: 0.3387
BIRCH Accuracy: 0.3461
Decision Tree Accuracy: 0.2622

The above process is repeated for each attribute individually
The columns (attributes) with the highest accuracy is seen below.
('kMeans: 8', 0.3446)
('kMeans: 0', 0.3402)
('BIRCH: 6', 0.3299)
('kMeans: 6', 0.3122)
('BIRCH: 0', 0.3034)
('kMeans: 1', 0.299)
('BIRCH: 8', 0.2946)
We can deduce that columns 0, 6, 8 consistently have the highest accuracy.

STEP 3: Create Model & Analyze Results

Models using KMeans Clustering, BIRCH Clustering, & Decision Tree Classification are rebuilt using the selected columns(attributes) only.
The results are shown below.
Kmeans Accuracy Before: 0.3387, Kmeans Accuracy After: 0.3461
BIRCH Accuracy Before: 0.3461, BIRCH Accuracy After: 0.3726
Decision Tree Accuracy Accuracy Before: 0.2622, Decision Tree Accuracy After: 0.268

The highest consistent accuracy is 37.26% using BIRCH Clustering.
This accuracy seems fairly low, lets review the dataset to see if we can understand whats going on.

The pie chart (See Figure1.pdf) shows the distribution of ranks among the training data

From this we can see ranks 1, 7, & 8 make up only 10% of the data.
Lets group ranks 1, 2, & 3 into a group and 6, 7, 8 into another group, and keep ranks 4 & 5 individual.

The accuracy for this 'Approximate Ranking' is 49.34%.
This is still not that great...

Lets reduce the precision and see if the model can predict rankings within +1 or -1 of the actual ranking.
The accuracy for 'Ranking ±1' is 79.68%.

Therefore...

The accuracy of our model to predict the exact ranking is: 37.0%.
The accuracy of our model to predict the ranking within ±1 ranking is: 80.0%.