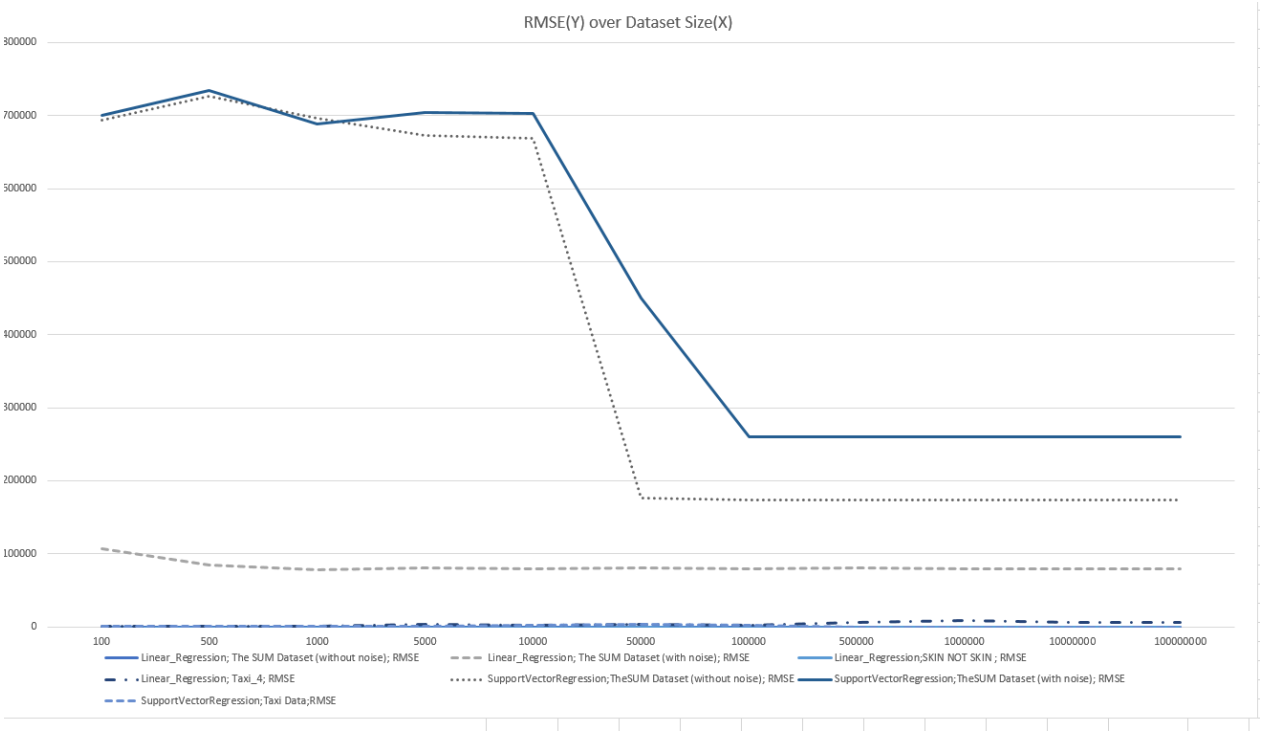
The (Un?) Reasonable Effectiveness of Data: Report

Team: team 13
Student IDs: 14311128, 14314613,14311771
Total Time Required (in hours): 15 hours

Findings/Answer

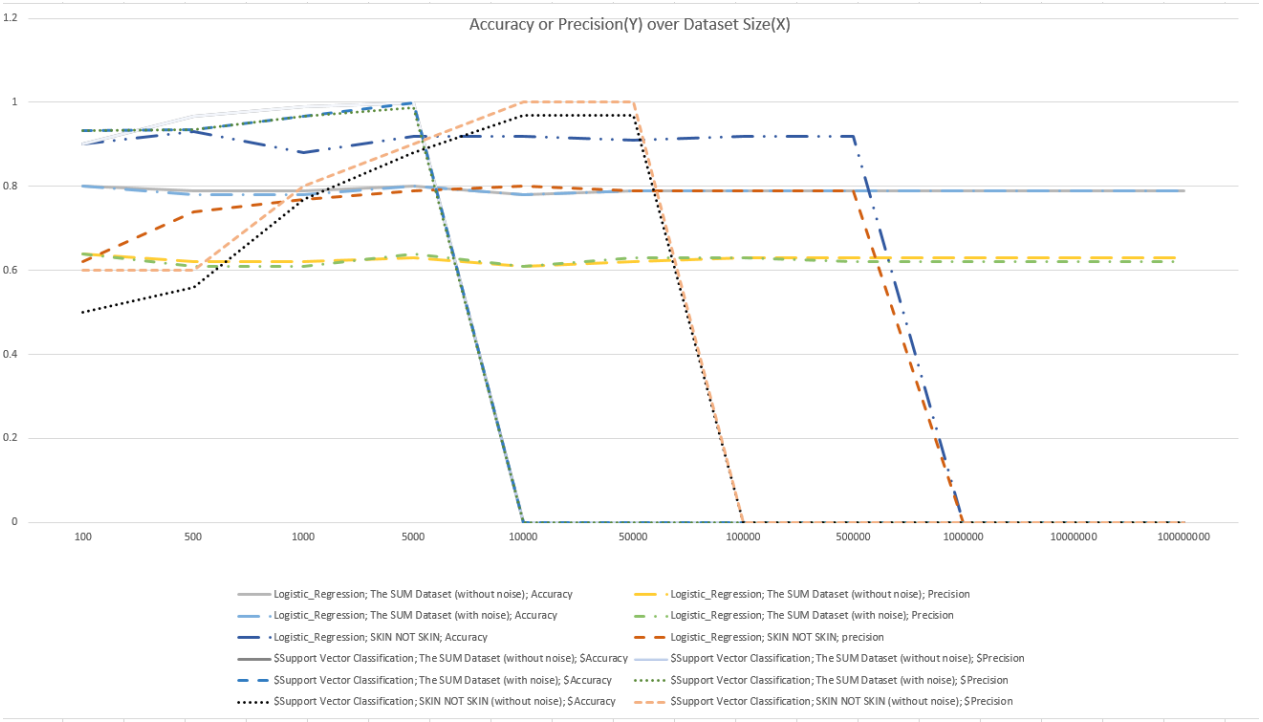
Question 1: To what extent does the effectiveness of machine-learning algorithms depend on the size and complexity of the data? [200-300 words]

"We don't have better algorithms. We just have more data." This was Peter Norvig's claim in the article "The Unreasonable Effectiveness of Data" published in the IEEE. It lays out the assertion that different algorithms perform virtually the same with the addition of more data. However, this claim that that the solution to the problem is to throw more data at a problem and home the algorithm finds a solution. Sometime the amount of data is not the problem but the quality, and at some point, there is a diminishing marginal utility of data where the processing time it takes to run the algorithm with larger and larger number of samples doesn't result in any significant changes in improved performance of the algorithm.



This is the performance of our Regression algorithms measured by Root Mean Squared Error. The first thing to note is that we copied over the results of our last suitable example for datasets larger than available. The data shows decreases in the overall RMSE as we increase the number of samples but depending on the implementation and the data the rate of decrease is larger. We also had our Linear

Regression algorithm return a score of 0 for RMSE at every available dataset size (this we feel is an error on our part).



This is the accuracy and precision of our classification algorithms over dataset size, we should note the where the results drop to 0, this is because the algorithm did not finish running and timed out for these dataset sizes. Overall precision and accuracy were closely linked and often improved (to levels nearing perfect scores of 1) as we increased to the larger datasets.

Overall, we can conclude that algorithm performance can be improved somewhat by increasing the amount of data used, but after a certain point the improvement in performance reduces to a point of irrelevance.

Question 2: Looking only at the performance of your best performing algorithm on "The SUM dataset (without noise)": how well was machine-learning suitable to solve the task of predicting a) the target value and b) the target class? Consider in your assessment, how well a simple rule-based algorithm could have performed. [100 words max]Data, Algorithms, etc.

The best performing Classification algorithm was Support Vector Classification. This is based on support vector machines. The algorithm achieved near perfect results on the dataset up to 5000 (where the compute time became too much for my computer. It was helped by the fact that very large and larger numbers were more frequent and so it was able to guess it incorrectly for some samples with out having a major impact on its accuracy or precision. A rule based algorithm may have struggled but could have operated under the rules if feature(1) is > than x and feature(10) is greater than x than it can be classed as y. Where x is a number in the range 1 – 10000 and y is the label applied. As features 1 and 10 have a large impact on the overall size of the computed number these are singled out in the rule based algorithm.

The best performing prediction (Regression) algorithm was linear regression, it reported a perfect score on every dataset size. This is probably an error, but this would perform better than any conceivable rule based system.

Algorithm 1	Linear Regression
Algorithm 2	Support Vector Regression
Algorithm 3	Logistic Regression
Algorithm 4	Support Vector Classification
Dataset 1	The SUM Dataset (without noise)
Dataset 2	The SUM Dataset (with noise)
Dataset 3	NY Taxis Data
Dataset 4	Skin Not Skin
Metric 1	RMSE
Metric 2	R2
Metric 3	Accuracy
Metric 4	Precision

Contributions (max. 200 words)

Gordon Hind (14311128): implemented Support Vector Classification and Support Vector Regression algorithms helped write the report.

Brian De Buiteach (14314613): implemented Logistic Regression and managed CSV files, including converting the skin not skin dataset from a text file to a usable csv format in Java.

Aodan O'Laoghaire (14311771):implemented linear regression, helped write report

Additional Information

If you feel that any additional information is needed to understand your work, please provide it here.

Gordon Hind (14311128) The amount of time it took to run many of the algorithms with the higher sample amounts or data complexity was too long for my laptop and would result in it shutting off the process. I believe that the code is technically correct but is not efficient or runnable on slow machines.