The (un?) reasonable effectiveness of data

This goal of this task is twofold. 1) Get familiar with scikit-learn and some of its machine-learning algorithms; 2) Find out how the effectiveness of machine learning differs for simple and complex tasks on different datasets and dataset sizes. More precisely, your goal is to answer the following questions:

1. To what extent does the effectiveness of machine-learning algorithms depend on the size (i.e. number of instances) and complexity (i.e. number of features) of the data? [200-300 words]
2. Looking only at the performance of your best performing machine-learning 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 would have performed. [100 words max]

Details:

1. Use scikit-learn.
2. Choose two regression and two classification algorithms
    a. Regression
        i. Linear Regression
        ii. One algorithm of your choice
    b. Classification
        i. Logistic Regression
        ii. One algorithm of your choice
3. Choose four datasets, ~~each with at least 100,000 instances,~~ one with at least 10,000 instances, and three with at least 100,000 instances (ideally, all four have 100,000+ instances).
    a. One dataset of your choice with a low complexity (maximum of 5 features – you can also use a dataset with more features, but only pick 5 features or less)
    b. The SUM dataset (without noise)
    c. The SUM dataset (with noise)
    d. One dataset of your choice with a medium complexity (around 30+ features)
4. Divide your datasets into chunks of 100; 500; 1,000; 5,000; 10,000; 50,000; 100,000; 500,000; 1,000,000; 5,000,000; 10,000,000; 50,000,000; 100,000,000 instances. Treat each chunk as a separate dataset on which you do your training and testing.
    a. If a dataset has e.g. 9.x million entries, feel free to consider this equivalent to 10 million
    b. If your dataset has only e.g. 200.000 instances, well, then you can only use chunks of 100…100,000 instances. In that case, fill the results in the CSV file for the 500,000 to 100,000,000 chunks with a "non available" value.
    c. Always use the first x instances from the dataset to create the chunks, i.e. do not pick randomly e.g. 100k instances from the whole dataset but take the first 100k instances.
5. Apply the machine learning algorithms with each chunk (consider each chunk as separate dataset on which you train and test your algorithms). Keep in mind that even a dataset for regression usually can be used for classification (introduce your own classes if necessary, or reduce the number of classes to apply e.g. logistic regression), and vice versa.

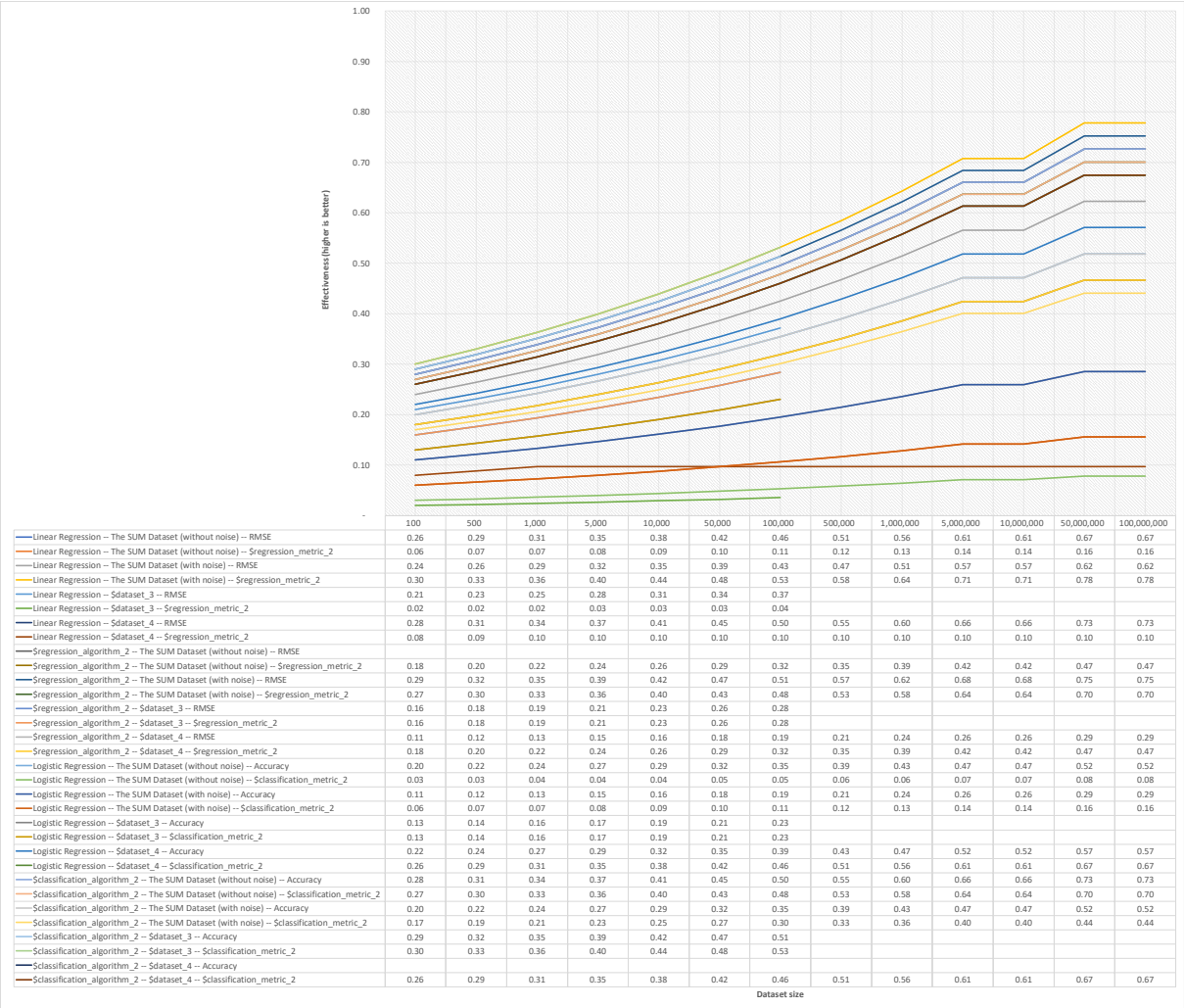
a. If you discover that it would take too long to train an algorithm on one of the large chunks (or the model wouldn't fit into your memory), feel free to not use that algorithm on that chunk.
b. Use a 10-fold cross validation. For the very large chunks: If the 10-fold cross method causes the training and evaluation to take too long, use a 70/30 split instead (and mention it in your report)
c. Use RMSE (if possible) and another metric of your choice for the evaluation of the regression algorithms
d. Use accuracy (if possible) and another metric of your choice for the classification algorithms.

The expected deliverables are

1. A CSV file with your results.
    a. The content of the CSV file must follow exactly (!) as specified in the example file.

	100	500	1000	5000	10000	50000	100000	500000	1000000	5000000	10000000	50000000	100000000
Linear Regression -- The SUM Dataset (without noise) -- RMSE	0.110	0.121	0.133	0.146	0.161	0.177	0.195	0.214	0.236	0.259	0.259	0.285	0.285
Linear Regression -- The SUM Dataset (without noise) -- $regression_metric_2	0.190	0.209	0.230	0.253	0.278	0.306	0.337	0.370	0.407	0.448	0.448	0.493	0.493
Linear Regression -- The SUM Dataset (with noise) -- RMSE	0.150	0.165	0.182	0.200	0.220	0.242	0.266	0.292	0.322	0.354	0.354	0.389	0.389
Linear Regression -- The SUM Dataset (with noise) -- $regression_metric_2	0.160	0.176	0.194	0.213	0.234	0.258	0.283	0.312	0.343	0.377	0.377	0.415	0.415
Linear Regression -- $dataset_3 -- RMSE	0.040	0.044	0.048	0.053	0.059	0.064	0.071	#N/A	#N/A	#N/A	#N/A	#N/A	#N/A
Linear Regression -- $dataset_3 -- $regression_metric_2	0.160	0.176	0.194	0.213	0.234	0.258	0.283	#N/A	#N/A	#N/A	#N/A	#N/A	#N/A
Linear Regression -- $dataset_4 -- RMSE	0.040	0.044	0.048	0.053	0.059	0.064	0.071	0.078	0.086	0.094	0.094	0.104	0.104
Linear Regression -- $dataset_4 -- $regression_metric_2	0.040	0.044	0.048	0.048	0.048	0.048	0.048	0.048	0.048	0.048	0.048	0.048	0.048
$regression_algorithm_2 -- The SUM Dataset (without noise) -- RMSE	#N/A	#N/A	#N/A	#N/A	#N/A	#N/A	#N/A	#N/A	#N/A	#N/A	#N/A	#N/A	#N/A
$regression_algorithm_2 -- The SUM Dataset (without noise) -- $regression_metric_2	0.090	0.099	0.109	0.120	0.132	0.145	0.159	0.175	0.193	0.212	0.212	0.233	0.233
$regression_algorithm_2 -- The SUM Dataset (with noise) -- RMSE	0.120	0.132	0.145	0.160	0.176	0.193	0.213	0.234	0.257	0.283	0.283	0.311	0.311
$regression_algorithm_2 -- The SUM Dataset (with noise) -- $regression_metric_2	0.140	0.154	0.169	0.186	0.205	0.225	0.248	0.273	0.300	0.330	0.330	0.363	0.363
$regression_algorithm_2 -- $dataset_3 -- RMSE	0.080	0.088	0.097	0.106	0.117	0.129	0.142	#N/A	#N/A	#N/A	#N/A	#N/A	#N/A
$regression_algorithm_2 -- $dataset_3 -- $regression_metric_2	0.080	0.088	0.097	0.106	0.117	0.129	0.142	#N/A	#N/A	#N/A	#N/A	#N/A	#N/A
$regression_algorithm_2 -- $dataset_4 -- RMSE	0.230	0.253	0.278	0.306	0.337	0.370	0.407	0.448	0.493	0.542	0.542	0.597	0.597
$regression_algorithm_2 -- $dataset_4 -- $regression_metric_2	0.260	0.286	0.315	0.346	0.381	0.419	0.461	0.507	0.557	0.613	0.613	0.674	0.674
Logistic Regression -- The SUM Dataset (without noise) -- Accuracy	0.230	0.253	0.278	0.306	0.337	0.370	0.407	0.448	0.493	0.542	0.542	0.597	0.597
Logistic Regression -- The SUM Dataset (without noise) -- $classification_metric_2	0.090	0.099	0.109	0.120	0.132	0.145	0.159	0.175	0.193	0.212	0.212	0.233	0.233
Logistic Regression -- The SUM Dataset (with noise) -- Accuracy	0.140	0.154	0.169	0.186	0.205	0.225	0.248	0.273	0.300	0.330	0.330	0.363	0.363
Logistic Regression -- The SUM Dataset (with noise) -- $classification_metric_2	0.250	0.275	0.303	0.333	0.366	0.403	0.443	0.487	0.536	0.589	0.589	0.648	0.648
Logistic Regression -- $dataset_3 -- Accuracy	0.260	0.286	0.315	0.346	0.381	0.419	0.461	#N/A	#N/A	#N/A	#N/A	#N/A	#N/A
Logistic Regression -- $dataset_3 -- $classification_metric_2	0.270	0.297	0.327	0.359	0.395	0.435	0.478	#N/A	#N/A	#N/A	#N/A	#N/A	#N/A
Logistic Regression -- $dataset_4 -- Accuracy	0.290	0.319	0.351	0.386	0.425	0.467	0.514	0.565	0.622	0.684	0.684	0.752	0.752
Logistic Regression -- $dataset_4 -- $classification_metric_2	0.230	0.253	0.278	0.306	0.337	0.370	0.407	0.448	0.493	0.542	0.542	0.597	0.597
$classification_algorithm_2 -- The SUM Dataset (without noise) -- Accuracy	0.080	0.088	0.097	0.106	0.117	0.129	0.142	0.156	0.171	0.189	0.189	0.207	0.207
$classification_algorithm_2 -- The SUM Dataset (without noise) -- $classification_metric_2	0.290	0.319	0.351	0.386	0.425	0.467	0.514	0.565	0.622	0.684	0.684	0.752	0.752
$classification_algorithm_2 -- The SUM Dataset (with noise) -- Accuracy	0.190	0.209	0.230	0.253	0.278	0.306	0.337	0.370	0.407	0.448	0.448	0.493	0.493
$classification_algorithm_2 -- The SUM Dataset (with noise) -- $classification_metric_2	0.220	0.242	0.266	0.293	0.322	0.354	0.390	0.429	0.472	0.519	0.519	0.571	0.571
$classification_algorithm_2 -- $dataset_3 -- Accuracy	0.290	0.319	0.351	0.386	0.425	0.467	0.514	#N/A	#N/A	#N/A	#N/A	#N/A	#N/A
$classification_algorithm_2 -- $dataset_3 -- $classification_metric_2	0.280	0.308	0.339	0.373	0.410	0.451	0.496	#N/A	#N/A	#N/A	#N/A	#N/A	#N/A
$classification_algorithm_2 -- $dataset_4 -- Accuracy	#N/A	#N/A	#N/A	#N/A	#N/A	#N/A	#N/A	#N/A	#N/A	#N/A	#N/A	#N/A	#N/A
$classification_algorithm_2 -- $dataset_4 -- $classification_metric_2	0.230	0.253	0.278	0.306	0.337	0.370	0.407	0.448	0.493	0.542	0.542	0.597	0.597

    b. The file name must follow exactly (!) the pattern "task1, $team_id, performance by dataset size, data.csv", e.g. "task1, team_01, performance by dataset size, data.csv"
2. A PNG image illustrating your results, and named "task1, $team_id, performance by dataset size, chart.png", e.g. "task1, team_55, performance by dataset size, chart.png". The image might look like:



	100	500	1,000	5,000	10,000	50,000	100,000	500,000	1,000,000	5,000,000	10,000,000	50,000,000	100,000,000
Linear Regression -- The SUM Dataset (without noise) -- RMSE	0.26	0.29	0.31	0.35	0.38	0.42	0.46	0.51	0.56	0.61	0.61	0.67	0.67
Linear Regression -- The SUM Dataset (without noise) -- $regression_metric_2	0.06	0.07	0.07	0.08	0.09	0.10	0.11	0.12	0.13	0.14	0.14	0.16	0.16
Linear Regression -- The SUM Dataset (with noise) -- RMSE	0.24	0.26	0.29	0.32	0.35	0.39	0.43	0.47	0.51	0.57	0.57	0.62	0.62
Linear Regression -- The SUM Dataset (with noise) -- $regression_metric_2	0.30	0.33	0.36	0.40	0.44	0.48	0.53	0.58	0.64	0.71	0.71	0.78	0.78
Linear Regression -- $dataset_3 -- RMSE	0.21	0.23	0.25	0.28	0.31	0.34	0.37						
Linear Regression -- $dataset_3 -- $regression_metric_2	0.02	0.02	0.02	0.03	0.03	0.03	0.04						
Linear Regression -- $dataset_4 -- RMSE	0.28	0.31	0.34	0.37	0.41	0.45	0.50	0.55	0.60	0.66	0.66	0.73	0.73
Linear Regression -- $dataset_4 -- $regression_metric_2	0.08	0.09	0.10	0.10	0.10	0.10	0.10	0.10	0.10	0.10	0.10	0.10	0.10
$regression_algorithm_2 -- The SUM Dataset (without noise) -- RMSE													
$regression_algorithm_2 -- The SUM Dataset (without noise) -- $regression_metric_2	0.18	0.20	0.22	0.24	0.26	0.29	0.32	0.35	0.39	0.42	0.42	0.47	0.47
$regression_algorithm_2 -- The SUM Dataset (with noise) -- RMSE	0.29	0.32	0.35	0.39	0.42	0.47	0.51	0.57	0.62	0.68	0.68	0.75	0.75
$regression_algorithm_2 -- The SUM Dataset (with noise) -- $regression_metric_2	0.27	0.30	0.33	0.36	0.40	0.43	0.48	0.53	0.58	0.64	0.64	0.70	0.70
$regression_algorithm_2 -- $dataset_3 -- RMSE	0.16	0.18	0.19	0.21	0.23	0.26	0.28						
$regression_algorithm_2 -- $dataset_3 -- $regression_metric_2	0.16	0.18	0.19	0.21	0.23	0.26	0.28						
$regression_algorithm_2 -- $dataset_4 -- RMSE	0.11	0.12	0.13	0.15	0.16	0.18	0.19	0.21	0.24	0.26	0.26	0.29	0.29
$regression_algorithm_2 -- $dataset_4 -- $regression_metric_2	0.18	0.20	0.22	0.24	0.26	0.29	0.32	0.35	0.39	0.42	0.42	0.47	0.47
Logistic Regression -- The SUM Dataset (without noise) -- Accuracy	0.20	0.22	0.24	0.27	0.29	0.32	0.35	0.39	0.43	0.47	0.47	0.52	0.52
Logistic Regression -- The SUM Dataset (without noise) -- $classification_metric_2	0.03	0.03	0.04	0.04	0.04	0.05	0.05	0.06	0.06	0.07	0.07	0.08	0.08
Logistic Regression -- The SUM Dataset (with noise) -- Accuracy	0.11	0.12	0.13	0.15	0.16	0.18	0.19	0.21	0.24	0.26	0.26	0.29	0.29
Logistic Regression -- The SUM Dataset (with noise) -- $classification_metric_2	0.06	0.07	0.07	0.08	0.09	0.10	0.11	0.12	0.13	0.14	0.14	0.16	0.16
Logistic Regression -- $dataset_3 -- Accuracy	0.13	0.14	0.16	0.17	0.19	0.21	0.23						
Logistic Regression -- $dataset_3 -- $classification_metric_2	0.13	0.14	0.16	0.17	0.19	0.21	0.23						
Logistic Regression -- $dataset_4 -- Accuracy	0.22	0.24	0.27	0.29	0.32	0.35	0.39	0.43	0.47	0.52	0.52	0.57	0.57
Logistic Regression -- $dataset_4 -- $classification_metric_2	0.26	0.29	0.31	0.35	0.38	0.42	0.46	0.51	0.56	0.61	0.61	0.67	0.67
$classification_algorithm_2 -- The SUM Dataset (without noise) -- Accuracy	0.28	0.31	0.34	0.37	0.41	0.45	0.50	0.55	0.60	0.66	0.66	0.73	0.73
$classification_algorithm_2 -- The SUM Dataset (without noise) -- $classification_metric_2	0.27	0.30	0.33	0.36	0.40	0.43	0.48	0.53	0.58	0.64	0.64	0.70	0.70
$classification_algorithm_2 -- The SUM Dataset (with noise) -- Accuracy	0.20	0.22	0.24	0.27	0.29	0.32	0.35	0.39	0.43	0.47	0.47	0.52	0.52
$classification_algorithm_2 -- The SUM Dataset (with noise) -- $classification_metric_2	0.17	0.19	0.21	0.23	0.25	0.27	0.30	0.33	0.36	0.40	0.40	0.44	0.44
$classification_algorithm_2 -- $dataset_3 -- Accuracy	0.29	0.32	0.35	0.39	0.42	0.47	0.51						
$classification_algorithm_2 -- $dataset_3 -- $classification_metric_2	0.30	0.33	0.36	0.40	0.44	0.48	0.53						
$classification_algorithm_2 -- $dataset_4 -- Accuracy													
$classification_algorithm_2 -- $dataset_4 -- $classification_metric_2	0.26	0.29	0.31	0.35	0.38	0.42	0.46	0.51	0.56	0.61	0.61	0.67	0.67

You can create the image either directly with scikit-learn (it should look roughly like the example); or after creating the CSV file, you just copy and paste the results to the Excel template, which automatically creates an image (you only need to store it as PNG then).

3. A PDF report that
    a. Contains the team id, individual student IDs, and the total time that was spent to complete the task
    b. Answers the above-mentioned questions
    c. Contains the created image (please ensure that the quality of the image is good. It doesn't matter if we must zoom in e.g. 400%, but the numbers etc. must be readable at some zoom level).
    d. Provides a brief overview (table) of the used datasets, algorithms, etc.
    e. Lists the contributions of all team members
    f. Contains additional information if required
    g. Is in about the same format as the report-template and named "task1, $team_id, performance by dataset size, report.pdf"
4. Your source code (no compiled/binary files) in the sub folder /task1, $team_id, performance by dataset size, source code/. Ensure you follow the "developer guidelines".