ABSTRACT:

Machine learning (ML) is a branch of Artificial Intelligence which enables computers to analyse the data and learn without any explicit programming. This task requires some amount of training data, however acquiring precise data for training the model is one the most expensive and difficult parts and accuracy of the model largely depends upon the type, quality and quantity of it. In this report, we have tried to study how much training data is enough to train the models effectively by applying some well-known algorithms on two multivariate datasets having continuous and binary dependent output variables, respectively. The result shows that increasing the size of training data will increase the overall accuracy but for some algorithms there is no any significant improvement could be achieved.

Introduction:

Machine Learning is a process by which computer can make prediction through analysing the input data and it is either curve fitting or classification tasks. [Domingos] In last few years, the use of machine learning has been increased tremendously due to increase of computational power. Nowadays, it is widely used in Web search, object detection, recommender systems, drug design and many other applications. A report published by McKinsey Global Institute claims that ML will revolutionize the future innovation [15].

In ML, data plays an important role and based on that the input features, algorithm, and accuracy metric are selected. In order to train the algorithm, the data is divided into training and testing datasets. Therefore, the first question that arise while implementing an algorithm is that how much data is required to train the model effectively. As per our knowledge, there is no definite answer to this, but in most scenario, it depends on various factors like complexity of the algorithm, input features, correlation between data etc. For example, non-linear algorithms will need more training data compared to linear models.

Related Work:

In [Claudia], it was proposed that training size should be defined by specifying confidence interval widths for classification algorithm in bio spectroscopy field. As mentioned in [ESL], increasing the training dataset will overfit the model. Hence, the model will adjust closely with training data and it will not generalize well. It was found in [Hajian-Tilaki] that how the performance of models vary with the training dataset size in biomedical applications. The investigation in [Cho] describe about how much training data is required to have an accurate model in medical image deep learning systems. In [Sun], it was found that it is possible to have a better accuracy in machine translation systems even with large training datasets.

**4. Methodology**

During the research, following steps were followed: Identifying relevant datasets & their target features, data preprocessing, splitting the dataset into test and train, splitting training dataset into chunks of different length, build models upon these chunks, evaluate these models with the test data to find accuracy, plot the results to compare the accuracy vs train data size.

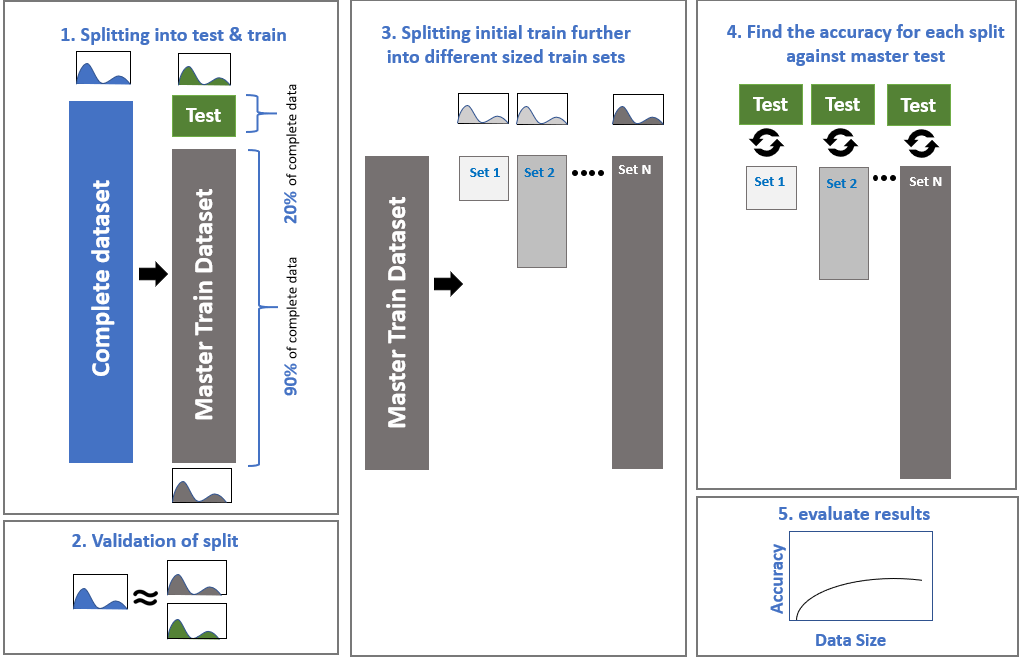
**4.1 Data Sets**

Two multivariate datasets from the UCI machine learning library were used for the research, “Bike Sharing” and “Bank Marketing“ dataset. Bike sharing dataset has 17,379 observations and 16 features recorded for 2 years (2011, 2012) at day-hour level. The target variable is the number of bikes rented & features are environmental conditions at the hour. The “Bank Marketing Dataset” is a bank customer level data which has 45,212 observations and 17 features. Target variable is if the contacted customer subscribed to the bank term deposit or not. The features are bank client information.

**4.2 Data Pre-processing**

In both the datasets, null values were treated and EDA was performed to understand the data. Label encoding was performed on categorical variable. In “Bike Sharing” dataset, new features like Sunday flag (day is Sunday or not) & day period (“noon” ,”evening” etc.) were created.

**4.3 Data Splitting**

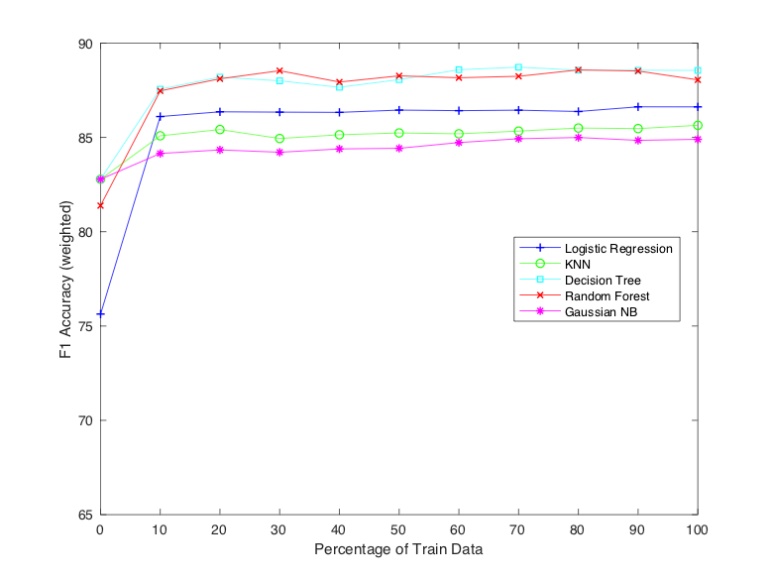


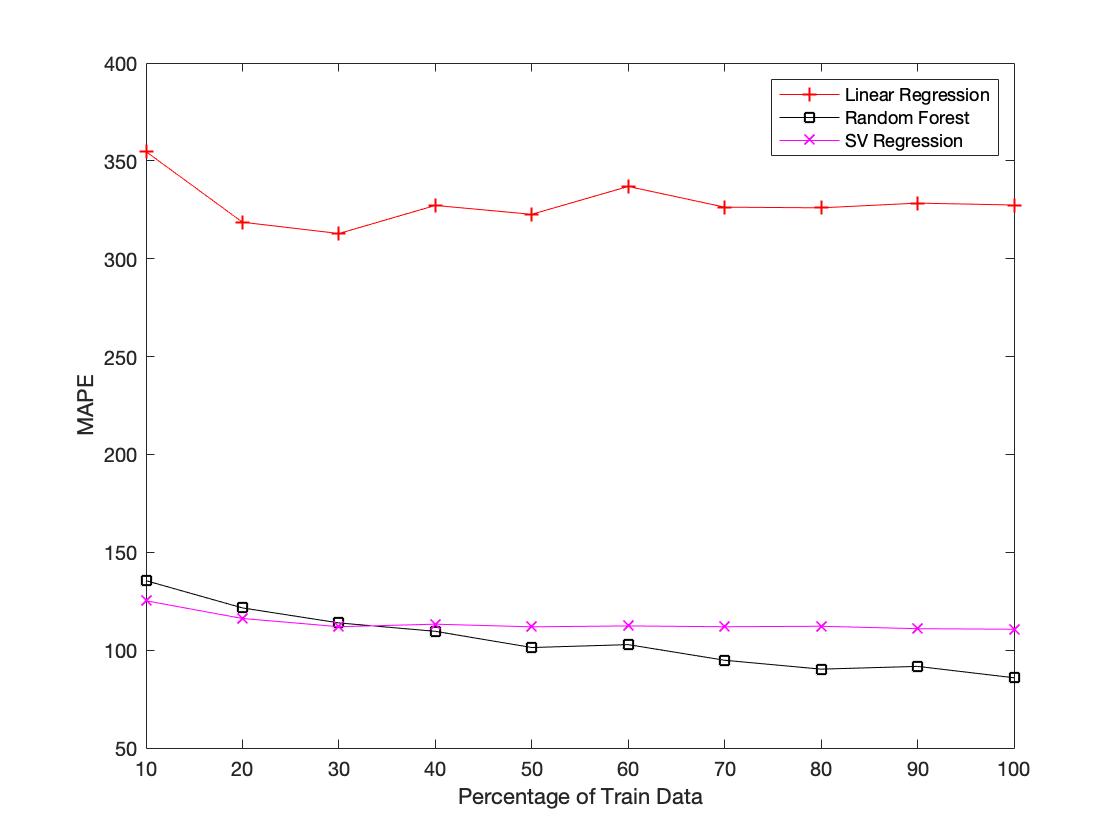
To analyze the impact of data amount on the performance of ML algorithms, the selected datasets were split into test and train in ratio of 20:80. 10 training datasets were containing increment number of observations (10% of the initial train dataset) were then generated from the initial train dataset [above figure]. Distributions of all test & train datasets were compared with that of the complete dataset to identify any biased split.

**4.4 Model building**

As dependent variable in the “Bike sharing” dataset is continuous, linear regression, support vector regression and random forest algorithms were used. In “Bank Marketing” dataset, as dependent variable is binary true/false, logistic regression, K nearest neignbour (KNN), decision tree, random forest & Gaussian Naive Bayes algorithms were implemented. All the models were implemented on the individual training sets generated from the initial train dataset.

**Results:**

****



As shown in Fig. 1, in Bank Marketing dataset, the impact of changing training dataset size is related to the selected classifier algorithm. In this figure, the first point on x-axis is 0.02% and the accuracy is low because of the underfitting phenomenon but by increasing the train dataset size, the accuracy starts to improve. However, after increasing the size of dataset beyond 10%, there is no significant enhancement in the performance of the models. It can be seen that more complex models like Decision Tree and Random Forest have higher performance compared to the simple models like Logistic Regression, Gaussian Naïve Bayes and KNN. In addition, it is clear that after increasing the size of training dataset to 10%, all the applied classification methods in this setup behave in a same fashion without any obvious change in the overall accuracy.

For the Bike Sharing dataset (Fig. 2), all the three algorithms behave differently when size of training dataset is increased. For linear regression model, mean absolute percentage error (MAPE) reduces up until 30% of train data size and then the error increase eventually becoming nearly constant after 80% . In case of support vector regression model, error decreases till the train size is 30% of overall training data but beyond this, error becomes almost constant. The random forest model’s error decrease constantly when size of training data increase. Random forest also has least amount of error (MAPE 85.9 ) when compared with the three models.

The size of dataset will have a significant impact of machine learning models up to a certain level. Based on simplicity of datasets, better accuracy is obtained with low amount of data. Complex models will have better accuracy compared to linear/simple models. However, this rate of improvement depends on the complexity of datasets.

**Limitations and outlook:**

In order to further investigate about the impact of training data set size on accuracy we can have optimized feature engineering and find a better way to label the features. Also, it is necessary to find a better correlation between various features. In addition, the performance of models could be checked with different accuracy metrics based on balanced/imbalanced datasets. Next plan would be to implement algorithms on Census Income Data Set and a few more datasets related to different applications.

**References**

Beleites, C., Neugebauer, U., Bocklitz, T., Krafft, C., and Popp, J. (2013). Sample size planning for classification models. *Analytica chimica acta*, 760:25–33.

Cho, J., Lee, K., Shin, E., Choy, G., and Do, S. (2015). How much data is needed to train a medical image deep learning system to achieve necessary high accuracy? *ArXiv e-prints*.

Domingos, P. (2000a). A unified bias-variance decomposition. In *Proceedings of 17th International Conference on Machine Learning*, pages 231–238.

Domingos, P. (2000b). A unified bias-variance decomposition for zero-one and squared loss. *AAAI/IAAI*, 2000:564–569.  
Figueroa, R. L., Zeng-Treitler, Q., Kandula, S., and Ngo, L. H. (2012). Predicting sample size required for classification performance.

*BMC medical informatics and decision making*, 12(1):8.

Geman, S., Bienenstock, E., and Doursat, R. (1992). Neural networks and the bias/variance dilemma. *Neural Computation*, 4:1–58.

Hajian-Tilaki, K. (2014). Sample size estimation in diagnostic test studies of biomedical informatics. *Journal of biomedical informatics*, 48:193–204.

Mukherjee, S., Tamayo, P., Rogers, S., Rifkin, R., Engle, A., Campbell, C., Golub, T. R., and Mesirov, J. P. (2003). Estimating dataset size requirements for classifying DNA microarray data. *Journal of computational biology*, 10(2):119–142.

Sun, C., Shrivastava, A., Singh, S., and Gupta, A. (2017). Revisiting unreasonable effectiveness of data in deep learning era. *arXiv preprint arXiv:1707.02968*.

P. Domingos, “A few useful things to know about machine learning,” Commun. ACM, vol. 55, no. 10, pp. 78–87, 2012.

J. Manyika, M. Chui, B. Brown, J. Bughin, R. Dobbs, C. Roxburgh, and A. Byers. Big data: The next frontier for innovation, competition, and productivity. Technical report, McKinsey Global Institute, 2011.