Data Set Reduction to Improve Computing Efficiency and Energy Consumption in Healthcare Domain

Priyadarshan Dhabe, Param Mirani, Rahul Chugwani and Sadanand Gandewar

[1] *Department of Information Technology*
Vishwakarma Institute of Technology
Bibwewadi, Pune-411037, Maharashtra, India
{priyadarshan.dhabe, param.mirani17, rahul.chugwani17, sadanand.gandewar17}@vit.edu

Abstract – Many data sets used in the healthcare domain are huge and thus, need a considerable amount of computations, memory, electrical energy, and processing time for training and testing of Machine Learning (ML) algorithms. In this chapter, we proposed a new dataset reduction technique to reduce the size of datasets without affecting much of classification and recognition accuracy. This can be done by identifying and removing the redundant data samples from the training data set, which leads to the creation of reduced data sets, which are termed as "*Bonsai Data sets*". As per our experimentation with the Skin segmentation data set, we reduced 39,993 samples out of 1,83,793 samples, which is around 21% samples, without sacrificing much on percentage classification and recognition. We used Logistic Regression (LR) [1],[2], K-nearest Neighbor (KNN) [3],[4] and Support Vector Machine (SVM) [5],[6] for testing classification and recognition accuracy using original as well as *Bonsai data sets.*

Keywords: dataset, data points, bonsai, recognition, distance factor

1 Introduction

Nowadays, computer-assisted medical diagnosis is pervasive due to better accuracy and lesser efforts. Various ML algorithms are used to train models that can be used for reasoning about the diagnosis. We need a data set to develop a model using ML algorithm. The more the samples are there in the data set better will be the developed model and hence the diagnosis. But a very large number of unnecessary samples in a given data set may lead to the following problems.

1. Increased computing time: - Many ML algorithms need multiple passes through the data set for model learning.
2. Increased Number of computations
3. Increased electrical power consumption.
4. Increased memory space

Dataset plays an important role in developing an ML model. It is a collection of n-dimensional data points. Each such data point is defined with $n-D$ vector containing n-values, one for each attribute. These data points are grouped into classes based on the similarity of their features. An input to the ML algorithm is such a data set and

output is the developed model and is used for prediction. These data sets are created and maintained by organizations and/or repositories without taking much care of redundant samples within them. It is not much of their fault since data collection is their prime goal. Thus, we are proposing a new approach to identify and remove redundant data samples for preparation of "*Bonsai data sets*" and their obvious advantages like energy efficiency, reduction in computations, improved execution time and memory space, without affecting much on the accuracy.

While reducing the dataset one must have to consider that, the reduction should not degrade substantially recognition and classification accuracies of the trained model. To accomplish this, we identify and remove redundant data points. A sample is called as redundant if it does not contribute to training since their peculiarities are already represented by some another pattern/sample. Thus, their removal will definitely reduce the size of data set but not have an adverse effect on the recognition and classification accuracy

Our proposed approach is very simple, we remove all the data samples of a pattern class that are falling close to the mean of that sample population with a user-defined distance parameter d , which needs to be learned (tunable) or decided for each data set. Since mean consists of all the properties of those removed data points it will contribute in the same way to the learning as compared to the removed redundant data points. There are many classifiers available for testing/verification of the classification and recognition accuracy. But, due to their popularity/wide applicability, we used LR [1],[2], KNN [3],[4] and SVM [5],[6] for verification of classification and recognition accuracy, before and after removal of redundant patterns from the training set. Independence of the proposed technique from the classifier is also another feature and provides the freedom to the user to choose any classifier.

1.1 Related Work

Reducing the data set has two-fold meaning. One interpretation is to reduce the dimensionality of the data set and the other is to reduce the number of data samples. The former approach is called as *dimensionality reduction* [6] and the latter is known as *Sampling* [7] [8]. Reducing data-sets for optimization is a common practice while building the ML model. Each of the approaches has its specific method and process of reducing the dataset. There are seven methods reported in [6] for dimensionality reduction. Our proposed method, the *Bonsai data set*, aims to remove data points that do not contribute to the learning.

Data columns with little changes in the data carry little information. Missing Value Ratio [6] is a technique in which the data columns having too many missing values are removed. Low Variance Filter [6] is another technique in which all data columns with variance lower than a given threshold are removed. Since variance depends on range, we have to normalize the dataset before the application of this technique. High Correlation filter [6] is another method in which we eliminate based on the correlation of two attributes. This is because if two attributes have a high correlation then only one of them will contribute to learning. Hence one of the two columns is removed.

Decision Tree Ensembles [10], also referred to as random forests, are useful for feature selection in addition to being effective classifiers. One approach to dimensionality reduction is to generate a large and carefully constructed set of trees against a target attribute and then use each attribute's usage statistics to find the most informative subset of features.

Principal Component Analysis (PCA) [1] is also a statistical procedure that orthogonally transforms the original n coordinates of a data set into a new set of n coordinates called principal components and then eliminates un-necessary attributes. Backward Feature Elimination [6] is a method where attributes are removed, one at a time, and then the model is re-built. Then the error rate of the new model and previous model is compared. This process continues until the error rate of a new model is more than the older model. The older-model then becomes the best performing model. Forward Feature Construction [6] is the inverse process to the Backward Feature Elimination. We start with one feature only, progressively adding one feature at a time, i.e. the feature that produces the highest increase in performance.

Patterns recognition algorithms such as LR [1][2], KNN [3],[4], SVM [5],[6] commonly suggest that their computations increase with the increase in the number of data-points in the data set. These methods are also used for experimentation purposes for computing classification and recognition accuracy of *the Bonsai data set*.

There are some sampling algorithms given in [7] [8] [9] for the selection of a subset of samples using various criteria from bigger data sets. Our proposed approach of *Bonsai Data set preparation* is somehow more similar to *sampling* than *dimensionality reduction* (*feature selection)* [6], but different than sampling too. In sampling, one sample at a time is randomly/systematically chosen and a decision is made to add it to the reduced set or not. We can consider the proposed approach of *Bonsai Data set preparation* as a *batch sampling* (we are coining this term and refers to), where a batch of patterns i.e subset of patterns within a given class are declared redundant and removed from that pattern class.

While developing these algorithms, everyone proposed/improved these algorithms without considering any limitations on electrical energy consumption. Due to this, a lot of electrical energy is consumed in building and using these algorithms. Based on [11], almost 28-35% of the total energy is spent in moving data across memory hierarchy (e.g. register to RAM, RAM to Disk, etc). Algorithms such as SVM consume more power [12] than other learning algorithms [13]. However, this consumption of energy is not even constant over its time of training. The following papers [12], [13] were also helpful in understanding the consumption of power for training and using ML algorithms.

The main research objective of this paper is to propose a method that detects and removes redundant data samples from a data set, which do not (very little) affect the reasoning accuracy, for optimizing computations, time, space, and electrical energy also. Cross-checking the accuracy of using redundant and reduced Bonsai data set using classifiers is another research question that is discussed in this paper.

2 Proposed Method of Preparation of Bonsai data set

In this section, we are presenting the proposed method of preparation of *the Bonsai data set*, which is the smallest subset of non-redundant data samples from the training set, such that, it must provide comparable classification and recognition accuracy. The proposed method is explained as follows.

Let S be the K sets of training pairs (r_h, d_h), where $r_h = (r_{h1}, r_{h2}, ..., r_{hn})$ be an n-dimensional input pattern for $n > 0$ and d_h is the class label. Each r_{hi} is one of the numeric features of the pattern r_h $i = 1, 2, ..., n$. An input pattern r_h has a class label $d_h \in \{d_{h1}, d_{h2}, ..., d_{hC}\}$ i.e set S has training patterns of C classes $C > 0$.

While deciding the redundant patterns we consider all the patterns belonging to a given class from training set, say $m \le C$, and computed the mean pattern $R_m = \{r_{m1}, r_{m2,...,r_{mn}}\}$ as an arithmetic mean of all the patterns of m^{th} class. We define a maximum distance $d_{\max}$ as the maximum distance of any pattern of m^{th} class from its mean R_m. We use a user-defined, tunable, fractional parameter $0 < \alpha \le 1$ for deciding the redundant patterns. A pattern of a m^{th} class is redundant if its distance d from mean R_m is satisfying condition $d \le (\alpha * d_{\max})$. The parameters $d_{\max}$, $(\alpha * d_{\max})$ and R_m used for computation of *the Bonsai data set* are pictorially illustrated in Fig. 1.

In our experimentation we start with a smaller value of α, identify and remove the redundant patterns from each class. Then, we check the classification and recognition accuracy. By gradually increasing α, we decide the final value of α, where we are getting maximum classification and recognition accuracy nearly the same as that of exhibited by the original training data set. We settle at this value of α and measure the total number of redundant patterns from the training data set. We recommend all the researchers not to use these patterns for their work to save computations, time, memory space, and electrical energy. Thus, the proposed approach provides benefits in all possible ways and hence strongly recommended.

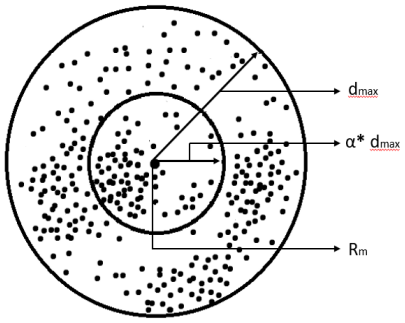

Fig. 1. Pictorial description of parameters used for computation of the Bonsai data set.

3 Experimental Results

We experiment with the Fisher Iris data set [14] for generating *proof-of-concept* (*POC*) of our approach. Then, we worked with the Skin Segmentation data set [15]. For both the data sets, we used 75% (roughly) randomly selected patterns as a training set and the remaining 25% as a testing set. The training set is used for classification accuracy and testing set for knowing recognition performance.

As per our study, there is no consensus on how much must be the split of training and testing set. A lot of debate is found on StackOverflow [16], between ration 75:25%, 80:20% and even on 50:50%. We learned that this ratio needs to be decided based on the size of the data set. If the data set is very large then even 50% samples in the training set and the remaining 50% samples in the test set will work fine. Due to the large data set, 50% of samples are also enough to fit the training data. But, if the data set is small to medium 80% to 20% and 75% to 25 % is common practice. The 80% training and 20% testing set is also supported by the "*Pareto principle*" [17]. But we used [18] and [19] as a logical base of 75% samples in the training set and 25% in the testing set, this is also a default setting in *sklearn* library in Python [19].

3.1 Fisher Iris Data Set

In this sub-section, we compare the recognition and classification accuracy of the Fisher Iris data set, as described in Table 1, using both the bonsai set and originally provided dataset.

We start experimentation with a value of α from 0 to 0.5 and the percentage reduction of patterns, obtained percentage classification, and recognition are shown in Table 2. In Table 2, %Rd represents percentage reduction, %C indicates percentage classification, and %R represents recognition.

Table 1. Description of Iris Data Set [14].

Number of Datapoints	150
Number of Attributes	4
Number of Classes	3
Associated Task	Classification

The value α is used to identify and eliminate redundant data samples from the training set. For $\alpha = 0$, i. e without any reduction (using all 115 samples from the training set), we obtained 96.52% classification and 97.37% recognition accuracy, respectively. This result is the first highlighted row of Table 2. Then by increasing α up to 0.25, we obtained the same % recognition accuracy of 97.37 as far $\alpha = 0$ and there is a small drop in % classification from 96.52% to 95.35%. This small drop in classification accuracy should not be treated as a serious handicap of the proposed method since most of the time we need recognition than the classification. Thus, by reduction of 25.22% of the training set (86 samples out of 115), we can provide

exactly the same recognition accuracy and comparable classification accuracy. This is shown in the 5th highlighted row of Table 2.

Table 2. Experimental results of the Fisher Iris data set.

Distance Factor (α)	Number of samples used in training	%Rd	%C	% R
0	115	0	96.52	97.37
0.1	112	2.61	96.43	97.37
0.15	108	6.09	96.30	97.37
0.2	97	15.65	95.88	97.37
0.25	86	25.22	95.35	97.37
0.3	75	34.78	94.67	94.74
0.35	60	47.83	93.33	94.74
0.4	48	58.26	93.75	97.37
0.45	42	58.26	92.86	92.11

Such, a *Bonsai data set* is recommended for training various ML systems rather than using the original data set.

If we increase α further from 0.25, we observed a decrease in both classification and recognition accuracy and thus not recommended to use these values. The plot of α versus classification and recognition accuracy is as shown in Fig. 2. The plot of percentage reduction in samples and accuracy is explained by Fig. 3. We suggest choosing the maximum value of α where we obtained classification and recognition accuracy close to the original training set.

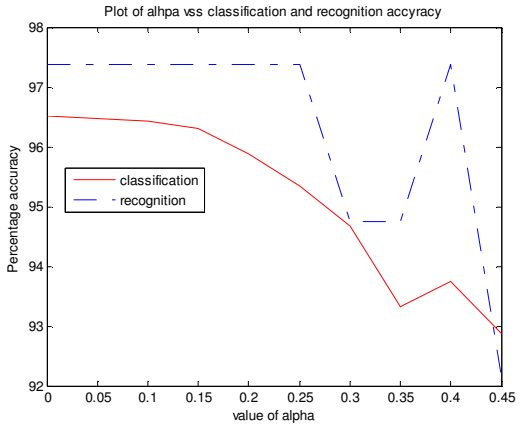

Fig. 2. Plot of alpha versus classification and recognition accuracy.

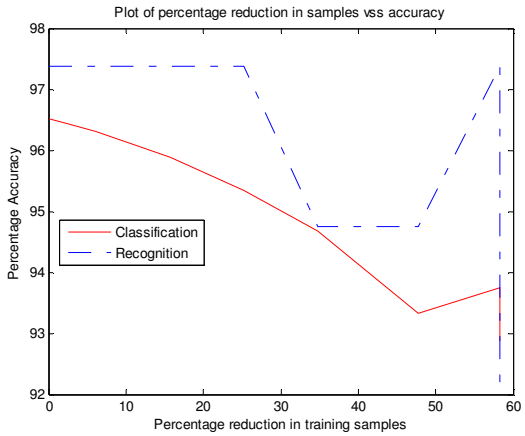

Fig. 3. Plot of percentage reduction in samples versus classification and recognition accuracy.

Let's summarize the different results which were found after training the bonsai datasets for different distance factors α. At the initial state distance factor is $\alpha = 0$ which means that 0% reduction has been done i.e. we are using the complete training dataset and the result obtained is shown as the first row of Table 2. As we go down in the table you will observe that distance factor α is increasing thereby, increasing the percentage reduction. Till the distance factor $\alpha = 0.25$ results are quite good and surprisingly close to the initial results, which is a great thing because the percentage reduction is nearly 25.22% which means only 75% of the original training set is sufficient to provide around the same classification and recognition accuracy.

After the distance factor $\alpha = 0.25\,5$ we observe the drop in recognition percentage so for this dataset we suggest that you should not go beyond a 25.22% reduction to get the best possible results. The 25% reduction means that out of 115 in Iris data set, we are training from only 86 patterns i.e. we can remove 29 patterns without much compromising on the accuracy.

We also measured the performance of the *Bonsai data set* using various ML algorithms and the performance is as shown in Table 3.

As shown in the above Table 3, we applied different training algorithms on bonsai datasets having different distance factor. As you can see all the algorithms give a really good recognition percentage till the distance factor is 0.25 when you go beyond that percentage recognition drops for every algorithm.

The motive behind applying different algorithms on bonsai datasets was to show that independent of any specific classification algorithm *bonsai datasets* performs well. From the above experimentation, one can conclude that without losing much on the performance we can reduce the dataset and it works with any ML algorithm.

Table 3. Accuracy of Bonsai datasets using different algorithms.

Algorithm	Distance Factor (α)							
	0.1		0.25		0.4		0.5	
	C%	R%	C%	R%	C%	R%	C%	R%
KNN(3)	96.43	97.37	95.35	97.37	93.75	97.37	97.14	86.84
KNN(5)	96.43	97.37	95.35	97.37	93.75	97.37	97.14	86.84
LR	98.21	97.37	97.67	97.37	95.83	94.74	97.14	86.84
SVM	98.21	97.37	97.67	97.37	97.92	94.74	94.29	86.84

3.2 Skin Segmentation Data Set [15]

Features of this dataset collected from the UCI repository [15] are mentioned below in Table 4. It is one of the famous data set from the healthcare domain.

Table 4. Features of the Skin Segmentation data set.

Number of Datapoints	245057
Number of Attributes	4
Number of Classes	2
Associated Task	Classification

To summarize the results mentioned in Table 5, highlighted 1st row indicates 0% reduction for $\alpha = 0$, we obtained 99.97% classification and 99.96% recognition accuracy. When we increase the distance factor α, percentage reduction will also increase and percentage recognition goes down slowly. The drop observed in performance is in decimal points $\alpha = 0.2$. For this particular dataset we suggest ideal reduction percentage would be 21% for $\alpha = 0.2$, the reason behind this is that after this point there is more drop in recognition accuracy. This means that only 79% of patterns of the existing training set is sufficient for comparable classification and recognition accuracy. Thus, remaining 21.76% patterns i.e around 39,993 patterns out of 183793 are declared redundant and can be removed. Removal of this 21%, huge number, of redundant patterns leads to a reduction in 21% number of computations, 21% computational time, and 21% amount of memory space and 21% less energy consumption. Fig. 4 shows the plot of α versus accuracy for *Skin segmentation data set*.

Table 5. Results of the Skin Segmentation Dataset.

Distance Factor (α)	Number of Samples used in training	%Rd	%C	% R
0	**183793**	**0**	**99.97**	**99.96**
0.1	176769	3.82	99.97	99.96
0.15	161330	12.22	99.96	99.82
0.175	156249	14.99	99.96	99.42
0.2	**143793**	**21.76**	**99.97**	**99.96**
0.25	133794	27.20	99.97	96.08
0.3	118091	35.75	99.97	92.25
0.5	57261	68.84	99.98	78.06

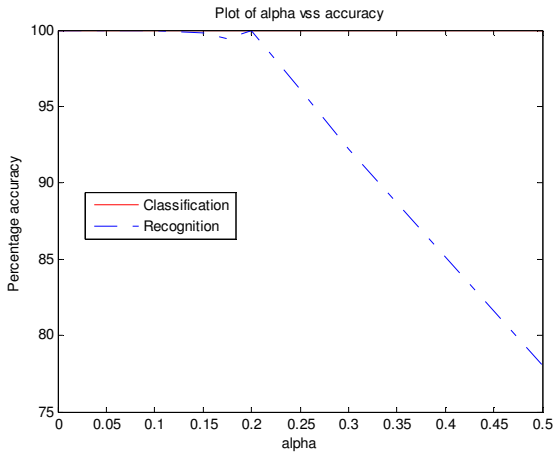

Fig. 4. Plot of alpha versus percentage accuracy.

From the experimentation using both the data sets, we conclude that there can be considerable redundant samples in the training data sets used for ML applications. We proposed a new sampling [7] [8] [9] type method of preparation of *the Bonsai data set* to detect and eliminate the redundant samples that do not contribute much in ML training. In the proposed approach we need to decide the best value of the parameter α experimentally, such that we can get around comparable classification and recognition accuracy using minimum samples from the training set. Such reduction will be helpful for a reduction in the number of computations, time, memory space, and consumption of electrical energy for training ML algorithms.

Reduced data set (Bonsai Data Set) will cut down the training time and recall time of the system used for medical diagnostic purposes and thus, keeps their users happy without compromising the quality of output. It also supports green computing [20] by reducing the number of computations to be done, since more computations need more consumption of electrical energy. More energy consumption can have a cascading effect of heat generation in the system and may need cooling.

4. Discussion of the results

As per our experimentation with both Iris data [14] and Skin Segmentation data [15], the proposed method of Bonsai data set, suggest the redundant patterns within each pattern class that can be removed so that ML algorithms can be trained quickly, with less memory space and thus, need less consumption electrical energy. A tunable parameter $0 < \alpha \leq 1$ needs to be decided for each data set for determining the redundant training samples. More the value of α, more will be the redundant training samples, like *sampling technique* [7][8][9]. By experimenting with α, we need to decide the value of α, where we get comparable recognition and classification accuracy as compared to the original training data set. Results show that the proposed approach allows us to reduce 21% computations, 21% computational time, 21% of memory space, and a similar amount of electrical energy (since 21% fewer computations are performed), for the skin segmentation data set [15]. Similar results can be expected from other ML data sets too and this proves the importance of the proposed approach of the *Bonsai data set*.

We suggest to all the websites like *UCI machine learning repository* [21] and kaggle [22] hosting huge data sets for ML purpose to also host reduced Bonsai data sets, such that many researchers can use the Bonsai data set and can work quickly with it by saving computations, time, memory space and electrical energy.

5 Conclusion

It is concluded from the proposed sampling method of *the Bonsai data set* that, we can remove the redundant patterns from ML data sets and can prepare the reduced bonsai datasets. We obtained bonsai data set with a 21% reduction in training samples for Skin Segmentation Data Set [15] and hence the same percentage reduction in computations, time, memory space, and electrical energy. Then, we can train the ML models on these Bonsai datasets without much compromising on classification and recognition accuracy. It is observed that using bonsai datasets for model training, we can achieve reduced computations, memory space, and reduced electrical power consumption as compared to datasets containing redundant patterns. Because of these advantages *bonsai sets* help users to train ML models efficiently, thus, they are strongly recommended as opposed to original data sets, in the health care domain.

References

1. Abdi. H. & Williams, L.J. (2010). "Principal component analysis". *Wiley Interdisciplinary Reviews: Computational Statistics*. **2** (4): 433–4598. Logistic Regression - Austin, J. T., Yaffee, R. A., & Hinkle, D. E. (1992). Logistic regression for research in higher education. Higher Education: Handbook of Theory and Research, 8, 379–410
2. Austin, J. T., Yaffee, R. A., & Hinkle, D. E. (1992). Logistic regression for research in higher education. Higher Education: Handbook of Theory and Research, 8, 379-410.
3. Thomas M. Cover and Peter E. Hart, "Nearest neighbor pattern classification," IEEE Transactions on Information Theory, (1967), Vol. 13 (1) pp. 21-27.
4. Boser, B. E., I. Guyon, and V. Vapnik (1992). A training algorithm for optimal margin classifiers . In Proceedings of the Fifth Annual Workshop on Computational Learning Theory, pages. 144 -152. ACM Press 1992.
5. D.K. Srivastava, L. Bhambhu (2009). Data classification using support vector machines. JATIT journal. Available online http://www.jatit.org/volumes/research-papers/Vol12No1/1Vol12No1.pdf
6. Andrea Manero-Bastin, "Seven Techniques for Data Dimensionality Reduction", https://www.analyticsvidhya.com/blog/2015/07/dimension-reduction-methods/
7. Yıldırım, Ahmet & Ozdogan, Cem & Watson, Dan. (2014). Parallel Data Reduction Techniques for Big Datasets. 10.4018/978-1-4666-4699-5.ch004.
8. Pang-Ning Tan; Michael Steinbach; Anuj Karpatne; Vipin Kumar, "Introduction to data mining", Second Ed., Pearson, 2019
9. Harsh Darji, "Sampling: An An approach to solve the bird counting problem", 2019, available [online] https://towardsdatascience.com/sampling-79075e9176cb
10. Leo Breiman, Jerome H. Friedman, Richard A. Olshen, and Charles J. Stone (1984). Classification and Regression Trees. Wadsworth International Group, Belmont, California.
11. G. Kestor, R. Gioiosa, D. J. Kerbyson, Hoisie (2013). Quantifying the energy cost of data movement in scientific applications. IEEE International Symposium on Workload Characterization (IISWC).
12. E. G. Martin, C.F. Rodrigues, G. Riley, H. Graham (2019). Estimation of energy consumption in machine learning. Journal of Parallel and Distributed Computing.
13. E. G. Martin, N. Lavison, H. Graham, Casalicchio, V. Boeva (2019). How to Measure Energy Consumption in Machine Learning Algorithms. Joint European Conference on Machine Learning and Knowledge Discovery in Databases.
14. *Fisher Iris* Data Set – Available online https://archive.ics.uci.edu/ml/datasets/iris
15. Skin Segmentation Data Set – Available online https://archive.ics.uci.edu/ml/datasets/skin+segmentation
16. https://stackoverflow.com/questions/13610074/is-there-a-rule-of-thumb-for-how-to-divide-a-dataset-into-training-and-validatio
17. Pareto principle-https://en.wikipedia.org/wiki/Pareto_principle
18. I. Guyon, A scaling law for the validation-set training-set size ratio, AT & T Bell Laboratories, Berkeley, Calif, USA, 1997.
19. sklearn library- https://scikit-learn.org/stable/
20. Kurp, Patrick. (2008). Green Computing. Commun. ACM. 51. 11-13. 2008, 10.1145/1400181.1400186.
21. https://archive.ics.uci.edu/ml/index.php
22. https://www.kaggle.com/datasets