iJRASET
International Journal For Research in Applied Science and Engineering Technology

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 10 **Issue:** IV **Month of publication:** April 2022

DOI: https://doi.org/10.22214/ijraset.2022.41743

www.ijraset.com

Call: 08813907089 | E-mail ID: ijraset@gmail.com

International Journal for Research in Applied Science & Engineering Technology (IJRASET)
ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.538
Volume 10 Issue IV Apr 2022- Available at www.ijraset.com

Digital Data Forgetting: A Machine Learning Approach

Aneri Vasani[1], Sahana Shettigar[2], Snehlata Rana[3], Prof. Shirish K Sabnis[4]
[1, 2, 3, 4]*Manjara Charitable Trust's Rajiv Gandhi Institute of Technology, Mumbai, India*

Abstract: ***Our lives have become increasingly reliant on data as the globe undergoes a rapid digital change. People nowadays collect a vast amount of data in order to obtain useful information, as the outcomes obtained from them are critical for planning. It's becoming more difficult to acquire and use data on computers as it grows in size. As a result, we may encounter data storage issues in the future. As a result, we propose a machine learning approach called "Digital Data Forgetting" to address this issue. These algorithms will not only delete non-valuable data, but they will also lower the size of the dataset. This is referred to as 'Big Cleaning.' For experimental purposes, we use Principal Component Analysis, K-nearest Neighbours, and Logistic Regression.***
Keywords: Digital Data Forgetting, Deep Autoencoders, PCA, Cleaning, Machine Learning, Logistic Regression.

I. INTRODUCTION

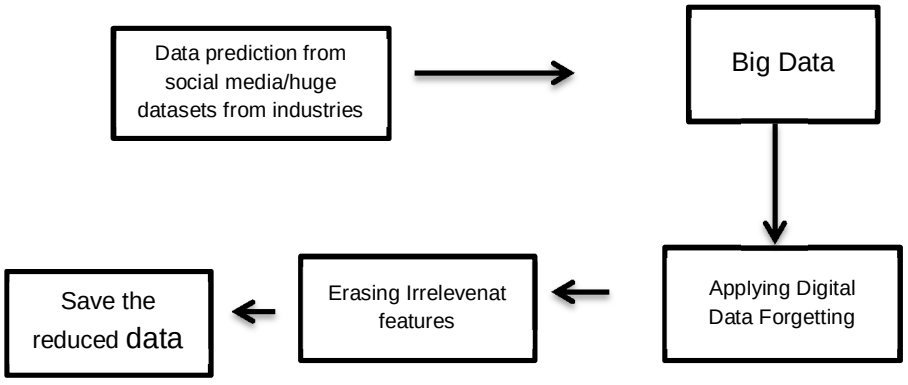
Data is a representation of the digital world's evolution. To advance to the next stage of digital existence, data storage, exchange, manufacturing, and processing are required. The amount of data available in the past was restricted, thus it was saved on paper. However, sharing was tough at the time. Digital change has advanced in recent decades, bringing with it the ability to produce precise data. Data is now producing capacity generated in a variety of ways. Social media, platforms such as Facebook, YouTube, and Instagram, and electronic trade on measuring values of equipment such as weather forecasts, gas pressure, and other gadgets are all examples of this.

Data has become an integral aspect of everyone's life today. Data is used in a variety of sectors, including medical, and by a variety of organisations, like Amazon, Flipkart, and others, for purposes such as analysis and prediction. Because storing and analysing such enormous amounts of data has become problematic, we now attempt to save all of the data we generate. Although the storage capacity is large, it is limited, and we shall hit that limit in the not-too-distant future. These massive amounts of data, termed "Big Data," contain information based on a variety of standards.As a result, it's critical to limit the amount of data while keeping the most valuable aspects, such as information for analysis.

As a result, we want to present the 'Digital Data Forgetting' strategy, which minimises data size by wiping the data that is least important.

The rest of paper is organised as follows:

1) Chapter 2: Construction of main framework of area is done with some questions.
2) Chapter 3: Some machine learning methods are explained.
3) Chapter 4: Some experiments are done.
4) Chapter 5: A conclusion is presented.



II. RELATED WORK AND RESEARCH QUESTIONS

Despite the fact that the rising cost of keeping data has resulted in a loss of time in machine learning models, there is no data cleansing method.

Similar to this difficulty, the community has come up with solutions to similar data issues.

Experts choose the most essential real-life data and issues. In a library, for example, deciding which books should be maintained on the shelves and which should be archived. It is handled by a librarian who is an expert in this sector, who develops the criteria based on her previous experience and chooses amongst them. The shelf layout is then completed. Because this type of research hasn't been done before in the literature, we've developed a method that supports real-world issues, such as in a library. We offer our strategy by producing replies to various inquiries in order to identify a novel approach for solving a problem that influences the success of a Machine Learning model arising from a data set.

1) Is there any data in the dataset that is inconsistent, despite the fact that the class is specified?
2) Is there a structure to the data set that causes overfitting?
3) Which component of the data set provides us with the most useful attributes?

We conclude that the strategy used to answer the aforementioned study questions will not only increase the performance of Machine Learning models, but will also save businesses money on storage costs.

III. MACHINE LEARNING FOR DIGITAL FORGETTING

In this research, we implement four strategies that are widely used in the field of machine learning. For experimental and analysis purposes, K-Nearest Neighbours, PCA, Deep autoencoders, and Linear Regression are used. It is vital to understand the concept behind how we can forget data utilising these techniques.

A. KNN

One of the most widely used machine learning algorithms is the k-NN algorithm. It can be used for regression analysis as well as categorization. The location of data from a test set whose class is already known is used to build a k-NN model. It basically categorises itself into different categories, eg: 3 various types of candies in a box.

B. PCA

PCA is another machine learning algorithm that is used mostly. Using the PCA algorithm, we may determine the most important components that reflect the data. The eigenvector-based approach is PCA.

3 simple steps:

1) Standardization..
2) Calculation of covariance matrices
3) IDENTIFY THE PRINCIPAL COMPONENTS BY COMPUTING THE EIGENVECTORS AND EIGENVALUES OF THE COVARIANCE MATRIX.

C. Deep Autoencoders

We also use a deep auto encoder, which is a type of neural network, in our research. This procedure is mostly used to reduce dimension. This method is quite good at detecting irrelevant data. It is an unsupervised learning technique with two components: an encoder and a decoder. The data is fed into the encoder, which creates a model/code to encode or convert it.

The converted data is referred to as a better qualified data set than the input, such as a noise-free data set.

D. Logistic Regression

Logistic regression is a "supervised machine learning" approach that can be used to model the likelihood of a specific class or occurrence. When the data is linearly separable and the result is binary or dichotomous, this method is applied. As a result, for Binary classification tasks, Logistic regression is commonly utilised.

PCA and Deep Autoencoders are thus utilised to reduce the dataset's dimensionality, while KNN and Logistic Regression are employed for classification and regression.

The goal of this research is to locate and eliminate less valuable material. We employ PCA and Deep Autoencoders for this.

The use of KNN Algorithm and Logistic Regression is to check the improved accuracy after removing irrelevant data.

ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.538
Volume 10 Issue IV Apr 2022- Available at www.ijraset.com

IV. IMPLEMENTATION

Three datasets are used in this analysis. The first dataset is the iris dataset, which has 50 samples of each kind, each of which has been classified into four additional attributes. The wine prediction dataset is the second dataset. It has 13 characteristics ranging from alcohol to proline k, as well as one dependent variable, customer segment.
The third dataset is the cancer dataset, which includes 30 features that can help clinicians discriminate between benign and malignant tumours.
We aim to extract the least relevant characteristic while erasing data from data storage because it will not be able to store all data in digital storage in the future.

The 4 techniques for experimental analysis are:

A. Application of PCA on Digital Data Forgetting

To determine the most essential feature in the dataset, the PCA algorithm is utilised. We delete the n-variety of data that aren't really necessary. The goal is to find and remove the least important features:

1) Calculate the average of the data
2) Use the mean of the data to scale the data.
3) Get the diagonals of the covariance matrix to find the data variances.
4) Delete the first n records with the smallest variance.

B. Application of KNN

We apply the KNN algorithm to determine the class of each record. We remove the record if it falls within the cluster's middle threshold.
The steps are as follows:
To begin, we use the iris dataset to train a k-NN model and determine the accuracy of k-NN prediction.
Then we use a deep autoencoder technique with a digital data forgetting strategy. The input layer of our deep autoencoder model contains 150 nodes, the middle layer (final layer of encoder) has 1 node, and the output layer has 150 nodes. We notice that after a 10% forgetting period, the prediction accuracy improves.

C. Application of Deep Auto Encoder

We utilize a deep autoencoder for data compression and discover the farthest records to compress the data in this manner. We eliminate the records that are the farthest away.
Algorithm steps are as follows:

1) Normalize the input
2) Design a deep autoencoder
3) The last layer of the encoding section must have one node
4) Train the deep autoencoder
5) Predict with the encoding layer We will have compressed data after this step.
6) Find the distance between compressed data and normalised data. Sort the distances discovered.

The record/feature with the largest distance is the least relevant. Finally, get rid of the records/features that aren't critical.

D. Application of Logistic Regression

Logistic regression is a technique for estimating the likelihood of a discrete outcome based on a single variable. The most common logistic regression models produce a binary result, such as true or false, yes or no, and so on.
The algorithm's steps are as follows:

1) Data pre-processing.
2) The Training set is subjected to Logistic Regression.
3) Predicting how an exam will turn out.
4) Visualizing the test set's results to ensure accuracy (Creation of Confusion matrix).

As a result, machine learning techniques can be used to change data. We can save the most valuable characteristics after we've identified them, and the remainder can be deleted.

V. RESULTS

To begin, we use classification and regression methods without any sort of forgetting. The dataset is then subjected to digital data loss. The reduced dataset is saved. Then we repeat the forgetting procedure to see if the accuracy has improved.

For the purposes of this study, we used three datasets. The iris dataset, the wine prediction dataset, and the cover dataset are all available.

Sr. no	Dataset	Algorithm used for prediction	Result before DigitalData Forgetting	Result after DigitalData Forgetting
1.	Iris dataset	K-NN	96.67%	100%
2.	Wine prediction data	Logistic Regression	94.44%	97.92%
3.	Cancer dataset	Logistic Regression	96.5%	97.08%

VI. CONCLUSION

We used three datasets: the iris dataset, the Wisconsin Breast Cancer dataset, and the wine prediction dataset. We did not apply any digital data forgetting mechanism to the Iris dataset at initially, instead training the KNN model on it. After that, we use PCA to forget digital data, and then we train the KNN model on the dataset after removing irrelevant features. We noticed a 3.33 percent boost in accuracy when K= 5 was used as the best value.

On datasets with larger dimensions, we can use the digital data forgetting process. The accuracy of the breast cancer dataset and the wine prediction dataset improved by 0.58 percent and 3.48 percent, respectively.

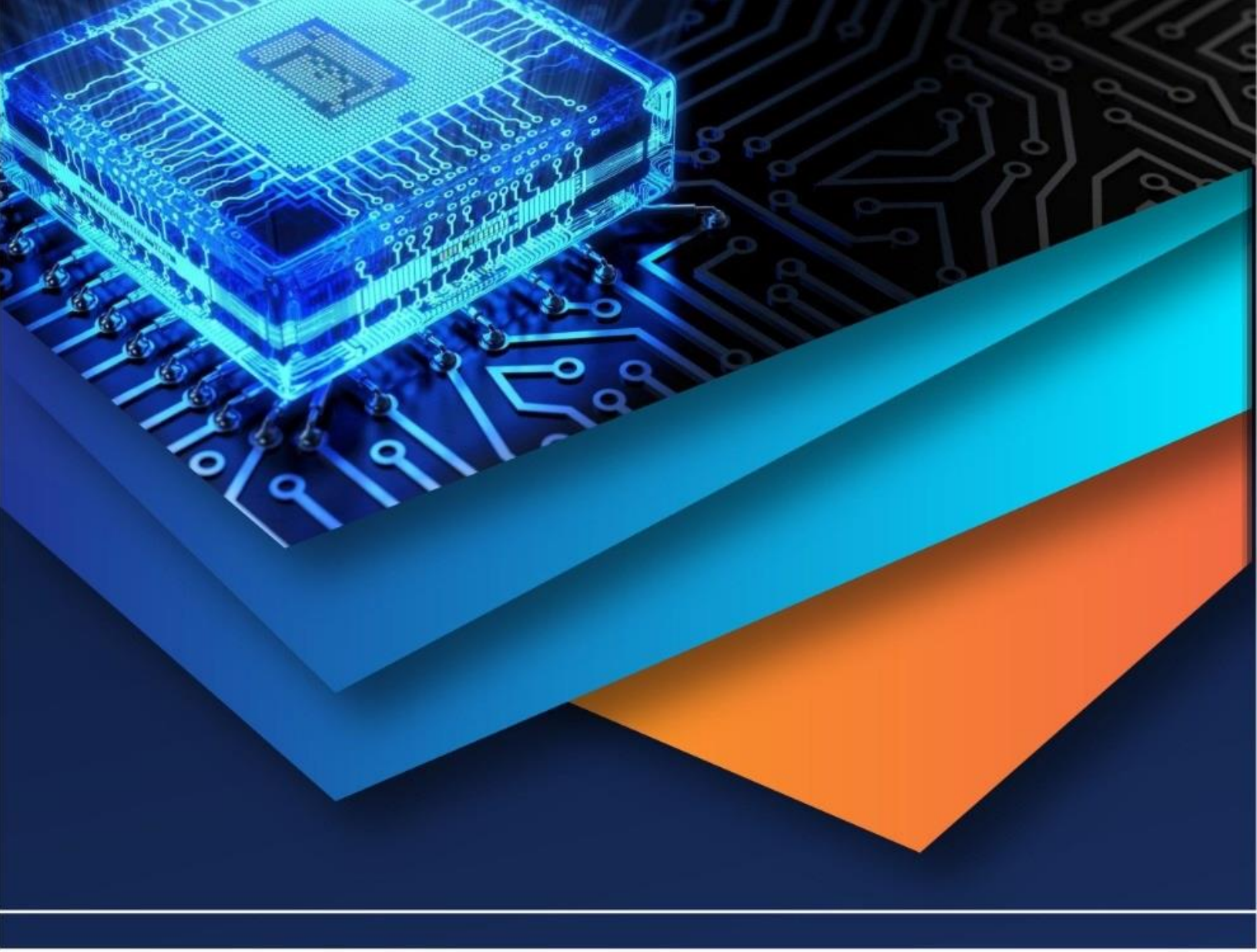
We found that reducing irrelevant data not only shrinks the dataset and makes it easier to work with, but it also improves the accuracy of the machine learning algorithm.

We provide a digital data forgetting approach for huge data called 'Big Cleaning' in this paper. Storage devices will be able to store more data as a result of this method.

Furthermore, it improves accuracy, which will benefit a variety of organisations and can be used in medical disciplines, among other things.

REFERENCES

[1] Melike GÜNAY, Digital Data Forgetting: A Machine Learning Approach. IEEE Journal
[2] Digital Data Forgetting: A Machine Learning Approach Mir Mohammad Yousuf1 , Shravan Kumar2 ,Mamoon Rashid3 ,Bisma Majid pp. December 2019
[3] Visualization Study of High-dimensional Data Classification Based on PCA-SVM, ZHAO Zhongwen pp. Second international conference 2017
[4] Big Data Dimension Reduction using PCA Tonglin Zhang Department of Statistics, Interenational Conference of smart cloud 2016
[5] Dimension Reduction Big Data using recognition of data, data features based on Copula Function and Principal Component Analysis, pp. 2021
[6] I. Jolliffe, “Principal component analysis. In: International encyclopedia of statistical science”, Springer, Berlin, Heidelberg, 2011. p. 1094-1096
[7] Lei J H, Yang J H, Zhong J C. High-dimensional Data Visualization Based on Principal Component Analysis and Parallel Coordinate[J]. Computer Engineering, 2011, 37(1):48-50



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY