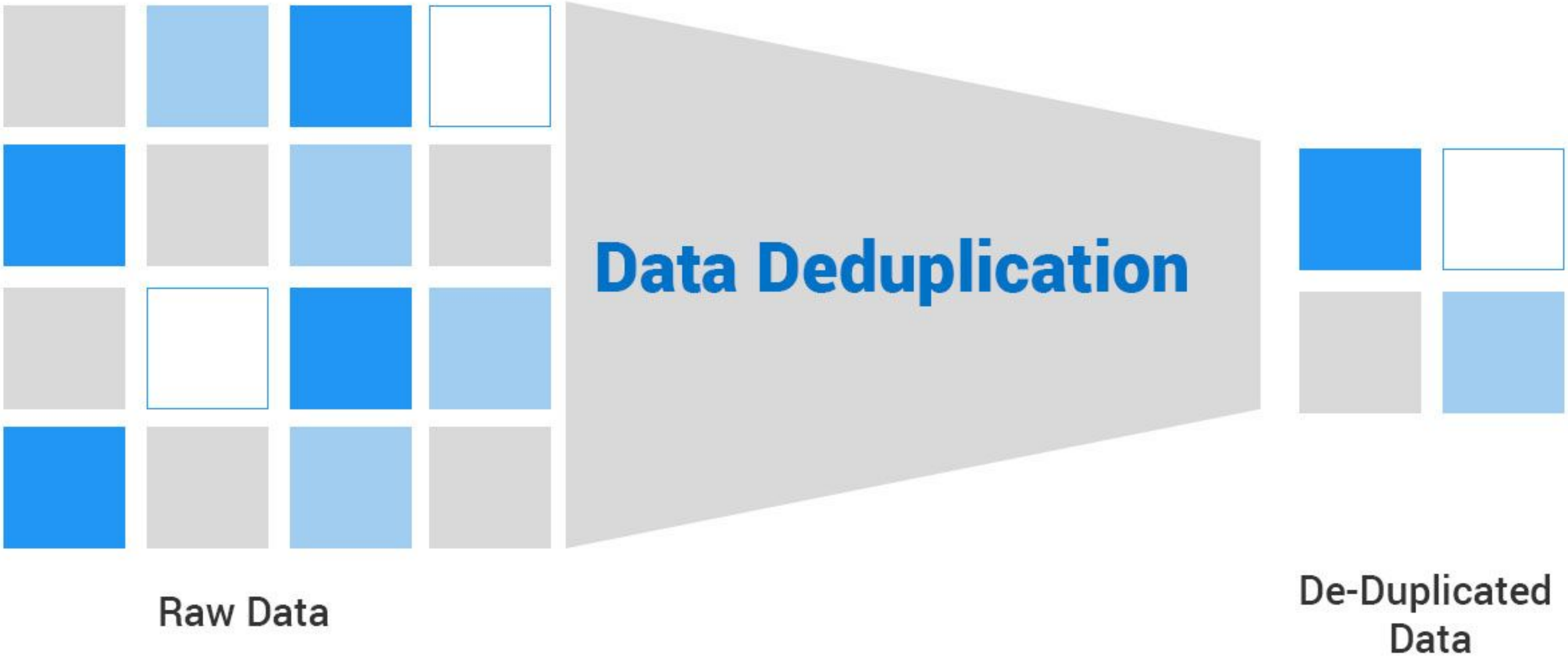
Data duplication removal using machine learning

NAME : Vatshayan



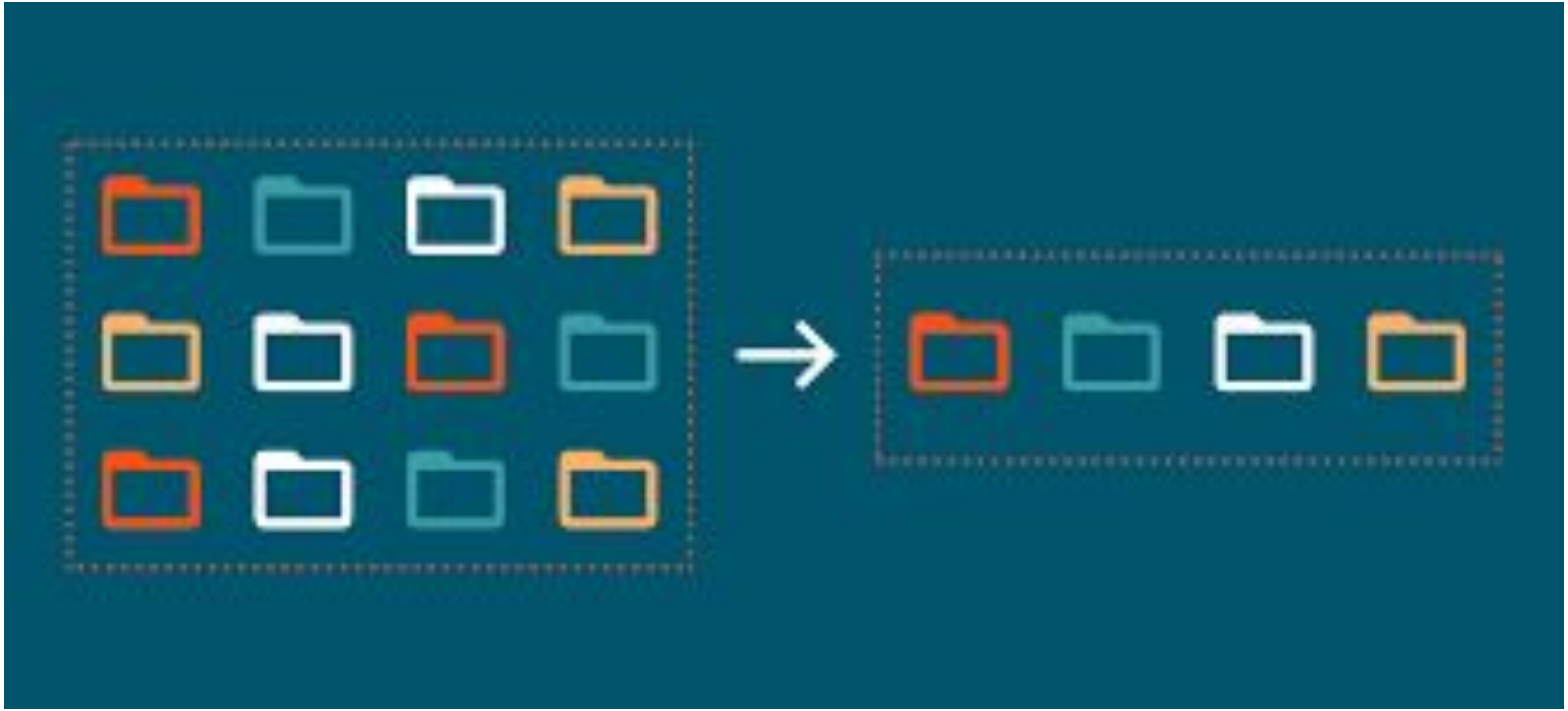
Data Deduplication

Raw Data

De-Duplicated Data

Abstract

Data duplication removal through machine learning is crucial for enhancing data quality. This process utilizes various algorithms, such as clustering and similarity matching, to automatically identify and eliminate duplicate records within large datasets. By considering diverse data attributes, such as text, numbers, and categories, machine learning models assess record similarity. Successful deduplication results in improved data quality, reduced storage costs, and enhanced performance in downstream applications. It plays a vital role in data management across various domains, constantly advancing to provide more accurate and automated duplicate detection and removal solutions.



INTRODUCTION

In computing, data deduplication is a technique for eliminating duplicate copies of repeating data. A related and somewhat synonymous term is single-instance (data) storage. This technique is used to improve storage utilization and can also be applied to network data transfers to reduce the number of bytes that must be sent. In the de-duplication process, unique chunks of data, or byte patterns, are identified and stored during a process of analysis. As the analysis continues, other chunks are compared to the stored copy and whenever a match occurs, the redundant chunk is replaced with a small reference that points to the stored chunk. Given that the same byte pattern may occur dozens, hundreds, or even thousands of times (the match frequency is dependent on the chunk size), the amount of data that must be stored or transferred can be greatly reduced. One of the most common forms of data de-duplication implementations works by comparing chunks of data to detect duplicates.

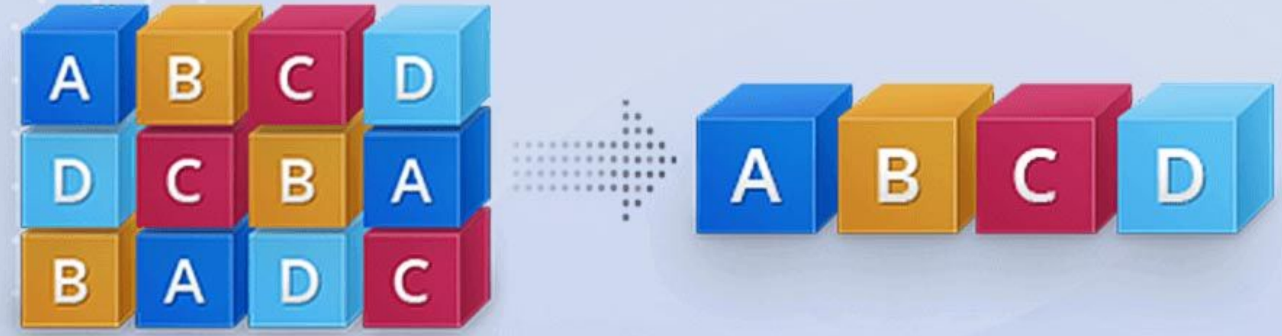
Conti.

For that to happen, each chunk of data is assigned an identification, calculated by the software, typically using cryptographic hash functions. In many implementations, the assumption is made that if the identification is identical, the data is identical, even though this cannot be true in all cases due to the pigeonhole principle; other implementations do not assume that two blocks of data with the same identifier are identical, but actually verify that data with the same identification is identical. If the software either assumes that a given identification already exists in the de-duplication namespace or actually verifies the identity of the two blocks of data, depending on the implementation, then it will replace that duplicate chunk with a link.

DATA DEDUPLICATION

Data Deduplication, often called Dedup for short, is a feature that can help reduce the impact of redundant data on storage costs. When enabled, Data Deduplication optimizes free space on a volume by examining the data on the volume by looking for duplicated portions on the volume. Duplicated portions of the volume's dataset are stored once and are (optionally) compressed for additional savings. Data Deduplication optimizes redundancies without compromising data fidelity or integrity. Data de-duplication is a technique for eliminating duplicate copies of repeating data.

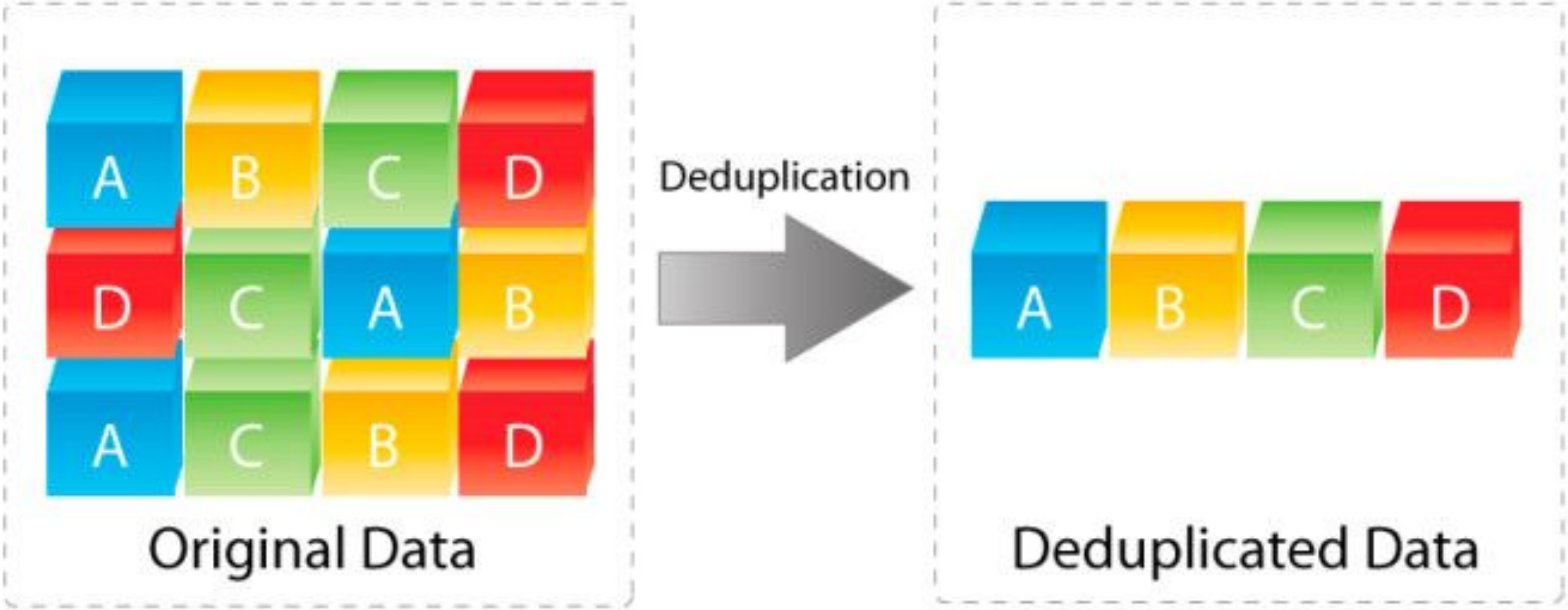
Data deduplication



OBJECTIVES

The following are the objectives of this project:

- Detecting and removing duplicates using Machine Learning by calculating the digest of files which takes less time than other pre-implemented methods.
- The project proposes an efficient method for detecting and removing duplicates using machine learning algorithms.
- Storage optimization by de-duplication.



Original Data

Deduplication

Deduplicated Data

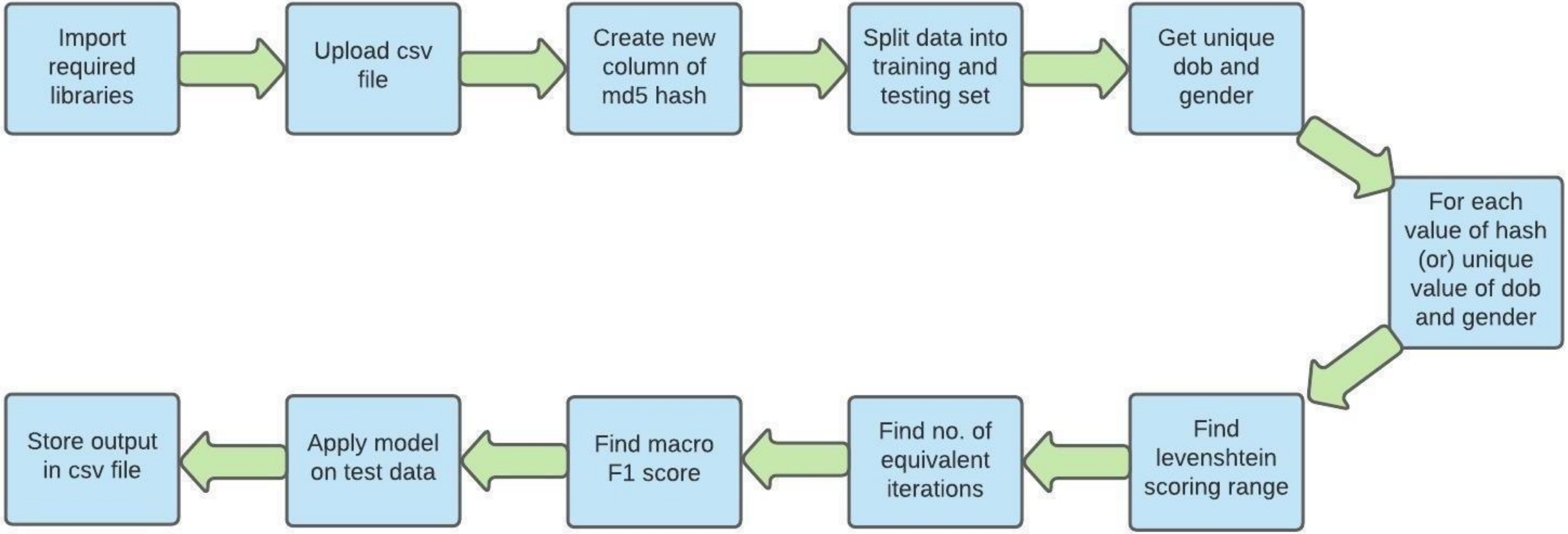
LITERATURE SURVEYS

1. Data de-duplication is a technology of detecting data redundancy, and is often used to reduce the storage space and network bandwidth. Now it is one of the hottest research topics in the backup storage area. In this paper, five representative chunking algorithms of data de-duplication are introduced and their performance on real data set is compared. The experiment result shows that the performance of these methods is improved obviously from the whole-file chunking to the TTTD chunking. According to the analysis of their features, we can provide some references for backup storage systems to choose the best chunking algorithm for eliminating data redundancy.
2. One of the best techniques is the deduplication data. In this paper a study on previous researches will do. Focus on the gaps and the problems that they could not solve or introduce them as future work. Finally, we propose a new method depending on the literature to fill the gaps. This method based on the checking the data, whether it is in the cloud storage before storing it

METHODOLOGY

We perform and analyse data deduplication in two methods. Firstly, we implement a pure ML based algorithm which is used to find the duplicate entries. Secondly, we will formulate an algorithm using both checksum and Machine learning approaches. We will find the checksum of each entry in the file, and then use Machine learning to predict if a given entry is unique.

FLOW



Import required libraries

Upload csv file

Create new column of md5 hash

Split data into training and testing set

Get unique dob and gender

For each value of hash (or) unique value of dob and gender

Find levenshtein scoring range

Find no. of equivalent iterations

Find macro F1 score

Apply model on test data

Store output in csv file

Conclusion

Our project successfully recognised the original values from the csv filled with duplicate values in both ML technique and the ML & hash technique .The deduplicated data frames were then saved in separate csv files.The ML model was the better model when compared to the hash model. Even though it was slow, it gave accurate results. With the emerging need for Automation, AI and various ML based approaches come in handy for realising our needs. Our approach combines the usage of both previously used methods like hashing and the upcoming field of machine learning.This helps us build more effective models with reduced error rates.Making it beneficial for the long run.

References

1. Title: "A survey of data deduplication techniques"
   Authors: Sotiris Ioannidis, Manolis Sifalakis, and Vasilis Samoladas
   Published in: ACM Computing Surveys, 2011
2. Title: "Deep Deduplication: A deep neural network approach to record deduplication"
   Authors: Thomas Tuytelaars and Cees G. M. Snoeyink
   Published in: Proceedings of the 25th ACM International on Conference on Information and Knowledge Management, 2016
3. Title: "An Efficient Block-Sorting-Based Deduplication Scheme for Large Scale Backup Storage"
   Authors: Fang Liu, Hong Jiang, and Dan Feng
   Published in: IEEE Transactions on Computers, 2011
4. Title: "Robust Hybrid Record Deduplication"
   Authors: Divesh Srivastava, Muralikrishna Vepakomma, and Sunita Sarawagi
   Published in: Proceedings of the VLDB Endowment, 2007