Privacy Preserving Data Mining

Minor project II

Submitted by:  
Dharmesh Malav (19803005)  
Nikhil Paleti (19803024)  
Divyanshu Tiwari (19803025)

Under the supervision of:  
Mrs. Niyati Aggrawal



Department of CSE/IT  
Jaypee Institute of Information Technology University, Noida

May 2022

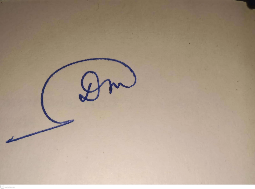
ACKNOWLEDGEMENT

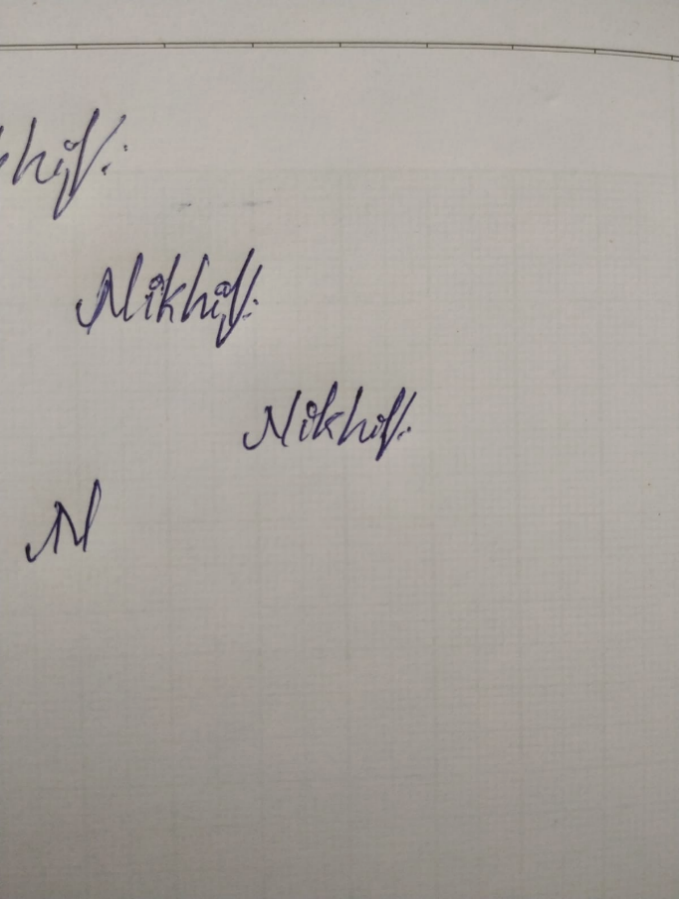
I would like to place on record my deep sense of gratitude to Mrs. Niyati Aggrawal , Designation, Jaypee Institute of Information Technology, India for his/her generous guidance, help and useful suggestions.

I express my sincere gratitude to Mrs. Niyati Aggrawal , Dept. of CSE, India, for his/her stimulating guidance, continuous encouragement and supervision throughout the course of present work.

I also wish to extend my thanks to \_\_\_\_\_\_\_\_\_\_\_\_\_ and other classmates for their insightful comments and constructive suggestions to improve the quality of this project work.

Signature(s) of Students

Dharmesh Malav (19803005) 

Nikhil Paleti (19803024)

Divyanshu Tiwari (19803025) 

DECLARATION

We hereby declare that this submission is our own work and that, to the best of our knowledge and beliefs, it contains no material previously published or written by another person nor material which has been accepted for the award of any other degree or diploma from a university or other institute of higher learning, except where due acknowledgment has been made in the text.

Place: Jaypee Institute of information technology ,Noida

Date: 16 May, 2022

Name: Dharmesh Malav

Enrolment No.: 19803005

Name: Nikhil Paleti

Enrolment No.: 19803024

Name: Divyanshu Tiwari

Enrolment No.:19803025

CERTIFICATE

This is to certify that the work titled “Privacy Preserving Data Mining” submitted by Nikhil Paleti , Dharmesh Malav and Divyanshu Tiwari of B.Tech of Jaypee Institute of Information Technology, Noida has been carried out under my supervision. This work has not been submitted partially or wholly to any other University or Institute for the award of any other degree or diploma.

Digital Signature of Supervisor

Name of Supervisor : Mrs. Niyati Aggrawal

Designation :

Date :

ABSTRACT

The objective of the proposed project is to look into and investigate about Privacy Preserving Techniques that are existing, and employed in the real world. The proposed project would also evaluate the privacy preserving techniques as deployed, on a real-life dataset.

TABLE OF CONTENTS

Page No.

*Abstract* *i*

*Acknowledgement ii*

*List of Tables* *iv*

*List of Figures* *v*

*List of Abbreviations vi*

*List of Nomenclature vii*

Chapter 1: INTRODUCTION

1.1GENERAL 2

1.2 . . . . . . . . . . . . . 2

1.2.1 General 3

1.2.2 . . . . . . . . . . . 3

1.2.3 . . . . . . . . . . . . 4

1.3 . . . . . . . . . . .. . . . . . . 41.4 . . . . . . . . . . . . . . . . . . 5

Chapter 2: LITERATURE REVIEW

2.1 GENERAL 7 2.2 ……………. 8:

:

:

References

Appendices- 1, 2, 3 (If any) 15

Power Point Presentation 16

List of Tables

Table Title Page

3.1 Values Assigned to Standard *k-*ε Turbulence Model Coefficients ................ xx

3.2 Values Assigned to RNG *k-*ε Turbulence Model Coefficients ...................... xx

4.1 Engine Specifications. .................................................................................... xx

4.2 Geometrical Details of the Injector ................................................................ xx

4.3 Boundary and Initial Conditions .................................................................... xx

4.4 Grid Independence Study ............................................................................... xx

List of Figures

Figure Title Page

3.1 Lagrangian Droplet Motion ........................................................................... xx

4.1 Vertical Manifold. ........................................................................................ xx

4.2 20O Bend Manifold......................................................................................... xx

4.3 90O Bend Manifold....................................................................................... xx

4.4 Spiral Manifold .............................................................................................. xx

4.5 Spiral Manifold Configuration (θ = 225o) ..................................................... xx

4.6 Spiral Manifold with Different Flow Entry Angles (20O, 32.5O and 45O) ........ xx

Nomenclature

English Symbols

A Pre-exponential constant

Cp Specific heat,J/kg-K

c Reaction progress variable

*Dd* Instantaneous droplet diameter, m

*Dm* Instantaneous droplet diameter

INTRODUCTION

Privacy is the point of conversation of many people lately, because of the complexity, scale and importance of data, more specifically – Personal Data, in the current era of computing, which relies heavily on such data to train Soft Computing models and personalize our experiences.

This data, however, is often stored as plain text, often unsecured, on databases, to be processed as quickly as required.

This calls for research and work to be done on techniques to protect the privacy of the users who have their data spread on the internet.

There are 3 major privacy preserving techniques that have been coined and researched to an extent.

K-anonymity: K-anonymity is a property of a dataset that indicates the re-identifiability

of its records. A dataset is k-anonymous if quasi-identifiers for each person in the dataset

are identical to at least k – 1 other people also in the dataset.

k-anonymity has 2 major techniques that can be deployed:

a. Suppression: In this method, certain attributes, or entire columns are replaced by any

generic character, like hash ‘#’ or arrow ‘^’. All or some values of a column may be replaced.

b. Generalization: In this method, individual values of attributes are replaced with a broader

category, i.e, a range of values.

L-diversity: “A q⋆-block is l-diverse if contains at least l “well-represented” values for the

sensitive attribute S. A table is l-diverse if every q⋆-block is l-diverse.”

In l-diversity, we aim to segregate the dataset into “l” diverse sections, where the

attributes falling under one of the sections, shares the same combinations of the key attributes.

l-diversity is most prone to skewness or attribute disclosure attacks.

T-closeness : “An equivalence class is said to have *t*-closeness if the distance between the

distribution of a sensitive attribute in this class and the distribution of the attribute in the

whole table is no more than a threshold *t*. A table is said to have *t*-closeness if all

equivalence classes have *t*-closeness .”

HITS Algorithm –

Hyperlink-Induced Topic Search (also known as hubs and authorities) is a link analysis

algorithm that rates Web pages.

Despite being originally designed for webpages, this algorithm can be applied in modified form,

on any directed graph, including our dataset

The HITS Algorithm is based on a fundamental model of the internet, which was basically a

directional graph. The HITS Algorithm is an iterative algorithm, which works by rating each

webpage (Node) according to two metrics:

Authority Value, which estimates the value of the content/node itself (in-degree).

Hub Value, which estimates the value of its links to other nodes (out-degree).

However, the disappointing fact is that Privacy Preservation techniques like these are almost

completely unused in the real world, in the field of data mining, even by major organizations

like Apple, Microsoft or Facebook, who collect user data on a large scale for purposes like

training their AT/Soft Computing Models, or more, simply because of the time involved in

these, and the lack of perceived monetary benefit.

This project aims to throw more light, and more “HITS” on the topic of privacy, and the (fairly)

easy to utilize techniques that can go a long way in protecting consumer privacy, in case of data

breaches

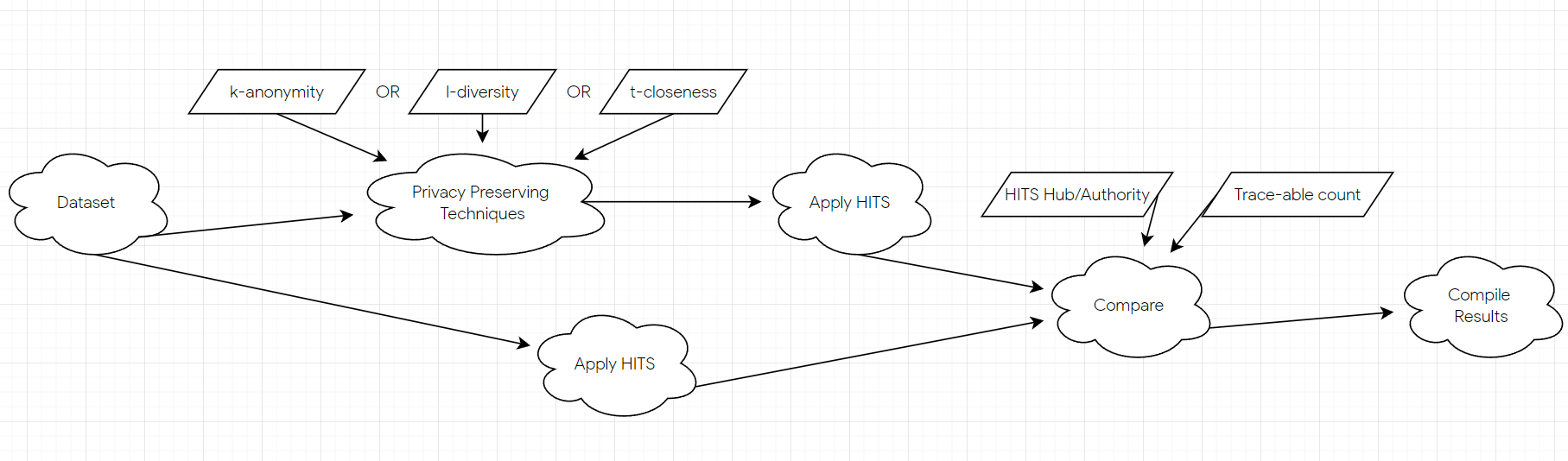
Requirement Analysis

1. Co-Authorship Network Analysis Dataset  
   <https://www.kaggle.com/code/bkoseoglu/co-authorship-network-analysis/data>This dataset is a freely available dataset, of research papers and scholarly articles published about coronaviruses like CoVID-19 or CoV-2.
2. It contains over 40,000 data points about authors, papers published, and collaborations, if any.  
     
   It seems to be an ideal candidate for our research, since the dataset contains certain private information about authors, which is important data that needs to be privatized.

Detailed design

The process of the project can be broken down into the following steps:

1. Data Ingestion and Pre-processing.
2. Apply Privacy Preserving Techniques on the dataset.
3. Apply Performance Measures on Dataset
4. Evaluate scores “before” and “after” the Privacy Preserving Techniques.



As the Project flow suggests, the work-flow begins from procuring a dataset, in our case, the

“Co-Authorship Network Analysis Dataset” from Kaggle.

Then we apply the pre-processing as necessary and store the dataset in a variable, as a numpy

array. A graph data structure will also be implemented.

A copy of the dataset variable will be made and then the copy of the dataset will be used to apply

privacy preserving metrics as mentioned above.

Both of the datasets, before and after privacy-preserving techniques will be run through the HITS

Algorithm, to procure Hub/Authority Scores. We also measure count of re-traceable unique data

members for the same.

Then the data from both the variables is compared, and results are compiled accordingly

Implementation

Data Preprocessing and Visualization

1. The data is ingested into a numpy array, from the original 3 CSV files of the dataset.
2. A “graph” data structure is implemented using the author\_papers.csv file
3. Pandas is used to handle the dataset, and any incomplete data points are removed.
4. The outliers will be visualized (scatter plots, or Tensorboard), and removed as necessary.

Performance Measures

1. HITS Scores  
   We will run the HITS Algorithm on the Dataset before and after applying Privacy Preserving Techniques. The delta of the Hub and Authority Scores from HITS would be noted.  
   This measure will be an indicator of the actual usability of the dataset even when the data has been privatized. The delta needs to be minimized to indicate that the effect of the Privacy Preservation is less, or negligible.
2. Trace-ability.  
   We will be making count of the number of unique values for authors/papers that we can trace back to, for specific attributes, before and after applying Privacy Preserving Techniques.  
   This count must be reduced after applying privacy measuring techniques since this indicates that the “traceability”, as the name suggests, of the users, is reduced, thus reducing privacy concern

Experimental Results and Analysis

Conclusion of the Report and Future Scope

REFERENCES

1. Duan, Y., Wang, J., Kam, M., & Canny, J. (2005). Privacy preserving link analysis on dynamic weighted graph. Computational & Mathematical Organization Theory, 11(2), 141-159.

2. Kerschbaum, F., & Schaad, A. (2008, October). Privacy-preserving social network analysis for criminal investigations. In Proceedings of the 7th ACM workshop on Privacy in the electronic society (pp. 9-14).

3. LeFevre, K., DeWitt, D. J., & Ramakrishnan, R. (2006, April). Mondrian multidimensional k-anonymity. In 22nd International conference on data engineering (ICDE'06) (pp. 25-25). IEEE.

4. Sweeney, L. (2002). k-anonymity: A model for protecting privacy. International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems, 10(05), 557-570.

5. Liu, X., Lin, H., & Zhang, C. (2012). An Improved HITS Algorithm Based on Page-query Similarity and Page Popularity. J. Comput., 7(1), 130-134.

6. http://pi.math.cornell.edu/~mec/Winter2009/RalucaRemus/Lecture4/lecture4.html