ISSN(Online): 2319-8753
ISSN (Print): 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

*(A High Impact Factor, Monthly, Peer Reviewed Journal)*

*Visit: www.ijirset.com*

**Vol. 9, Issue 3, March 2020**

# Data Anonymization of Bank Data Using Suppression and Randomization

Daphne Sneha S [1], Princy KS [2], Sathia Priya R [3]

U.G. Student, Department of Computer Science and Engineering, Loyola-ICAM College of Engineering and Technology, Chennai, Tamil Nadu, India[1]

U.G. Student, Department of Computer Science and Engineering, Loyola-ICAM College of Engineering and Technology, Chennai, Tamil Nadu, India [2]

Assistant Professor, Department of Computer Science and Engineering, Loyola-ICAM College of Engineering and Technology, Chennai, Tamil Nadu, India[3]

**ABSTRACT**: This paper aims to provide an approach for data anonymization using a bank dataset. In order to perform the process of anonymizing data, techniques such as suppression and randomization. The main objective of this proposal is to provide privacy of Personally Identifiable Information (PII). The process of data anonymization can be used to solve this problem. We have designed an approach which will make the process effective. Our proposal will enable the use of a bank dataset which contains critical Personally Identifiable information (PII). The data anonymization is performed in such a way that it complies with a regulation called the GDPR (General Data Protection Regulation). The proposal makes use of various techniques to anonymize all the attributes of the dataset with a unique approach. The approaches such as suppressing quasi-identifiers, separating non-quasi-identifiers based on the uniqueness of the attribute as categorical and continuous attributes, label encoding, generating histogram of an attribute, generation of probability distribution function, identifying the best distribution for a continuous attribute etc., These techniques are applied to a bank dataset to generate a dataframe that is completely normalized and usable in the data analysis

**KEYWORDS**: Data Anonymization, Suppression, Randomization, Personally Identifiable Information (PII), Bank Dataset

## I. INTRODUCTION

The main objective of data anonymization is to protect private or sensitive data by deleting or encrypting the Personally Identifiable Information (PII) from a database. The purpose of data anonymization is to protect an individual's or company's private activities while maintaining the integrity of the data gathered and shared. Other terms for data anonymization are "data obfuscation," "data masking," or "data de-identification". Data anonymization is whereby sanitization and masking of classical information should be done in such a way that if a breach occurs, the acquired data is useless to the culprits. The need to protect data should be followed as the highest priority practices in every company, as classified information that falls into the wrong hands can be misused, intentionally or unintentionally. Lack of sensitivity when handling sensitive client information can come at a great cost to businesses due to regulatory authorities cracking down on gross negligence. The General Data Protection Regulation(GDPR) states seven principles for maintaining the privacy of personal data such as collection, organization, structuring, storage, alteration, consultation, use, communication, combination, restriction, erasure or destruction of personal data. The authorities sign

ISSN(Online): 2319-8753
ISSN (Print): 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(A High Impact Factor, Monthly, Peer Reviewed Journal)

Visit: www.ijirset.com

Vol. 9, Issue 3, March 2020

a license agreement on the GDPR with the stakeholders (or) research persons who are in need of data to perform analysis. Fig. 1 shows the process of data anonymization, where the data publisher and the data recipient are involved.

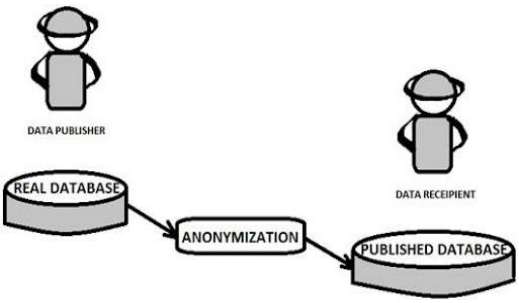



Fig. 1. Data Anonymization Process

There are certain cases where the GDPR is violated and data theft cases are recorded. Our proposal addresses this issue. In order to solve this problem, we make use of the Safe Harbor law, which is a part of the HIPAA (Health Insurance Portability and Accountability Act). The Safe Harbor law of HIPAA deals with the protection of data in the healthcare sector. In our proposal, we make use of the law and apply it in a bank dataset. We follow the generic data anonymization approaches to make sure the PII is completely removed.

## II. RELATED WORKS

Disha Dubli and D.K Yadav (2017) [1] have proposed a system that is used to perform data anonymization for privacy preservation. Here, secure techniques such as generalization, bucketization and slicing are used for performing data anonymization.

Graham Cormode and Divesh Srivastava (2017) [2] have presented a unified framework of data anonymization techniques. Different working models are generated which guarantees privacy offered by k-anonymization and i-diversity. Existing data anonymization techniques can be classified in several dimensions on the basis of nature of data, anonymization approaches and anonymization techniques.

Paul Francis, Sebastian Probst Eide and Reinhard Munz (2018) [3] designed Diffix : High Utility Database Anonymization, which is a general-purpose, easy to use and data quality preserving framework. Diffix adds a minimal amount of noise to answers—Gaussian with a standard deviation of only two for counting queries—and places no limit on the number of queries an analyst may make.

Nirzari Patel and Mehul P Barot (2018) [4] presented an observation on anonymization based privacy preserving data publishing where they concentrate about the data anonymization in a globally-network society. No-explicit identifiers such as name and phone number, birth date and ZIP code are anonymized.

The above references gave us a clear idea about the process of data anonymization, the distinction to choose the correct techniques which apply to the anonymization process, the choice of appropriate dataset for implementation, and the approaches to different attributes of the dataset.

ISSN(Online): 2319-8753
ISSN (Print): 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

*(A High Impact Factor, Monthly, Peer Reviewed Journal)*

*Visit: www.ijirset.com*

**Vol. 9, Issue 3, March 2020**

## III. PROPOSED SYSTEM

The IDE used is Google Colaboratory, which is supported by Google Compute Engine Server as backend, with RAM and ROM services provided for free by Google. The code for anonymization will be implemented at the server. The process of anonymization is implemented in modules, in which the degree of anonymization increases at the end of each stage. The important modules of implementation are suppression of quasi-identifiers, randomization of categorical attributes, and anonymization of continuous attributes to derive the anonymized dataset. Fig. 2. shows the basic system architecture used to build the proposal.

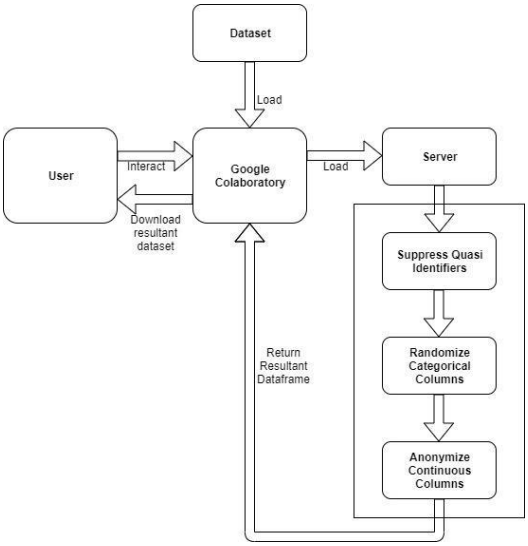



Fig. 2. Proposed System Architecture

## IV. ALGORITHM

*A) Suppression*

Suppression is a disclosure limitation method which involves removing data (e.g., from a cell or a row in a table) to prevent the identification of individuals in small groups or those with unique characteristics. This method may often result in very little data being produced for small populations, and it usually requires additional suppression of non-sensitive data to ensure adequate protection of PII (e.g., complementary suppression of one or more non-sensitive cells in a table so that the values of the suppressed cells may not be calculated by subtracting the reported values from the row and column totals). Correct application of this technique generally ensures low risk of disclosure; however, it can be difficult to perform properly because of the necessary calculations (especially for large multi-dimensional tables). Further, if additional data are available elsewhere (e.g., total student counts are reported), the suppressed data may be re-calculated.

*B) Randomization:*

Randomization is the process of generating random values in the place of original values. The objective of randomization is to preserve the privacy associated with the data by modifying the data in such a way that the analysis property and the frequency of distribution of data are minimally affected (or) unaffected. The attributes of the dataset that are subjected to randomization cannot be eradicated as it might be needed for analysis. Hence, the values need to

ISSN(Online): 2319-8753
ISSN (Print): 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(A High Impact Factor, Monthly, Peer Reviewed Journal)

Visit: www.ijirset.com

Vol. 9, Issue 3, March 2020

be modified in such a way that the privacy is not compromised and the property of the dataset is not modified as well. The randomization technique can be applied to select attributes of a dataset depending upon its frequency of unique values. The attributes for randomization are selected and might not contain personal data, but still contain a considerable amount of importance that its privacy is essential.

## V. IMPLEMENTATION

*A) Preparation of Dataset:*

The dataset has been obtained from "PKDD'99 Discovery Challenge" based on "Guide to the Financial Data Set". The dataset has been prepared by Petr Berka and Marta Sochorova.The dataset contains 8 relational tables. We are using "transaction's" to work on our project**.** The transaction table contains 10 attributes and 1056320 rows. Fig. 3. Shows a glimpse of the dataset.

| id | trans_id | account_ | date | type | operatior | amount | balance | k_symbo | bank | account |
|---|---|---|---|---|---|---|---|---|---|---|
| 0 | 695247 | 2378 | 930101 | PRIJEM | VKLAD | 700 | 700 | | | |
| 1 | 171812 | 576 | 930101 | PRIJEM | VKLAD | 900 | 900 | | | |
| 2 | 207264 | 704 | 930101 | PRIJEM | VKLAD | 1000 | 1000 | | | |
| 3 | 1E+06 | 3818 | 930101 | PRIJEM | VKLAD | 600 | 600 | | | |
| 4 | 579373 | 1972 | 930102 | PRIJEM | VKLAD | 400 | 400 | | | |
| 5 | 771035 | 2632 | 930102 | PRIJEM | VKLAD | 1100 | 1100 | | | |
| 6 | 452728 | 1539 | 930103 | PRIJEM | VKLAD | 600 | 600 | | | |
| 7 | 725751 | 2484 | 930103 | PRIJEM | VKLAD | 1100 | 1100 | | | |
| 8 | 497211 | 1695 | 930103 | PRIJEM | VKLAD | 200 | 200 | | | |
| 9 | 232960 | 793 | 930103 | PRIJEM | VKLAD | 800 | 800 | | | |
| 10 | 505240 | 1726 | 930103 | PRIJEM | VKLAD | 1000 | 1000 | | | |
| 11 | 144541 | 485 | 930104 | PRIJEM | VKLAD | 300 | 300 | | | |
| 12 | 637741 | 2177 | 930104 | PRIJEM | VKLAD | 800 | 800 | | | |
| 13 | 689827 | 2357 | 930104 | PRIJEM | VKLAD | 800 | 800 | | | |
| 14 | 846006 | 2881 | 930104 | PRIJEM | VKLAD | 700 | 700 | | | |
| 15 | 637742 | 2177 | 930105 | PRIJEM | PREVOD | 5123 | 5923 | DUCHOD | YZ | 6E+07 |

Fig. 3. Dataset

*B) Suppression of Quasi-identifiers:*

The quasi-identifiers are identified by evaluating the uniqueness of the column in the dataset. A column is a quasi-identifier if it is unique (or) almost unique for every record in the dataset.Quasi-identifiers are pieces of information that are not of themselves unique identifiers, but are sufficiently well correlated with an entity that they can be combined with other quasi-identifiers to create a unique identifier. Quasi-identifiers can thus, when combined, become personally identifying information. Fig. 4. Shows the status of the attributes of the dataset after suppression.

```
(230465, 7)
```

| | account_id | date | account | type | operation | k_symbol | bank |
|---|---|---|---|---|---|---|---|
| **15** | 2177 | 930105 | 62457513.0 | 0 | 1 | 1 | 12 |
| **24** | 1972 | 930107 | 14132887.0 | 0 | 1 | 1 | 10 |
| **46** | 3592 | 930110 | 73166322.0 | 0 | 1 | 1 | 6 |
| **49** | 576 | 930111 | 30300313.0 | 0 | 1 | 1 | 12 |
| **53** | 2357 | 930112 | 34144538.0 | 0 | 1 | 1 | 7 |

Fig. 4. Shape and head of the dataset after suppression

Null values that are associated with the dataset are removed as they pose no meaning during the analysis performed using the dataset. Removing null values also cleans the dataset.

*C) Separating Non - Quasi Identifiers*

In order to anonymize the attributes of the dataset that are not very unique and contain very less of unique values, we are randomizing them by replacing each unique value by a common pattern of labels.This makes the data secure and does not affect the property of the data in the dataset.The threshold value for separating the non-quasi attributes into

ISSN(Online): 2319-8753
ISSN (Print): 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

*(A High Impact Factor, Monthly, Peer Reviewed Journal)*

*Visit: www.ijirset.com*

**Vol. 9, Issue 3, March 2020**

categorical and continuous attributes is an important step for giving different treatment to the attributes of the dataset based on the number of unique values. The threshold value is set to "20 Unique Values". The number of unique values of the attributes is directly proportional to the degree of privacy (importance) they exhibit.

```
['type', 'operation', 'k_symbol', 'bank']
['account_id', 'date', 'account']
```

Fig. 5. Separated Non Quasi-Identifiers

*D) Randomizing categorical non-quasi-identifiers*

Once we identify the categorical columns, we can randomize them. The process of randomization is performed by generating random values of each categorical attribute. If the value of the attributes are of the type "text" or "string", there are possibilities that the property of the attribute can be leaked. In order to avoid this situation, the "text" or "string" data can be replaced with "integer labels" which hides the property of the attribute. Also, the analysis property of the attribute remains unaffected. Fig. 6. Shows a sample label encoding.

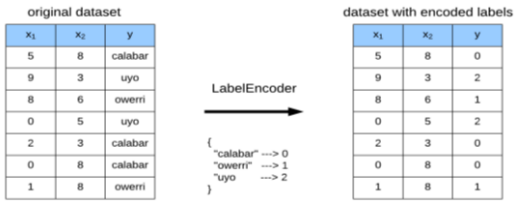



Fig. 6. Label Encoding

This is done after the completion of the label encoding process for the attributes if needed. For each categorical attribute, the number of unique values are identified. A list consisting of the unique values of the attribute is sent to the randomizing function along with the probability of occurrence of each of those unique values.

*E) Anonymizing continuous non-quasi-identifiers*

The process of anonymizing the continuous attributes involves the generation of histogram and manipulating the parameters of the histogram to generate the probability distribution function(PDF) for various distributions to choose the best distribution for each continuous attribute. A histogram is a graphical display of data using bars of different heights. In a histogram, each bar groups numbers into ranges. Taller bars show that more data falls in that range. In other words, it provides a visual interpretation of numerical data by showing the number of data points that fall within a specified range of values (called "bins"). It is similar to a vertical bar graph without gaps between the bars. Fig. 7. Shows a sample histogram. Each blue bar refers to a bin. The size of the blue bar (refer X - axis) is the bin size of the histogram. The average values of the bin can be inferred from the Y – axis.

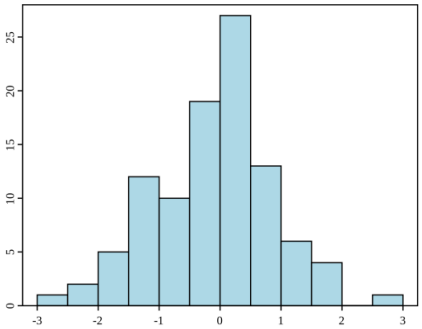



Fig. 7. Histogram

ISSN(Online): 2319-8753
ISSN (Print): 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

*(A High Impact Factor, Monthly, Peer Reviewed Journal)*

*Visit: www.ijirset.com*

**Vol. 9, Issue 3, March 2020**

The parameters of the histogram are input data, bins, density and bin edges.
Input Data - This refers to the input data. The histogram is computed over the flattened array.
Bins - In a histogram, the total range of the data set (i.e from minimum value to maximum value) is divided into a number of equal parts. These equal parts are known as bins or class intervals. No overlap is allowed between the bins. Any observation can occupy one and only one bin.
Density - If False, the result will contain the number of samples in each bin. If True, the result is the value of the probability density function at the bin, normalized such that the integral over the range is 1.
Bin Edges - The bin edges refer to the start and end values for each of the bins present in the generated histogram.
The statistical functions module contains a large number of probability distributions as well as a growing library of statistical functions. Each univariate distribution is an instance of a subclass of rv_continuous (rv_discrete for discrete distributions). The Sum of Squared Estimates of Error (SSE) measures the total deviation of the response values from the fit to the response values. It is also called the summed square of residuals and is usually labeled as SSE.

$$SSE = Sum_{(i=1 \text{ to } n)}\{w_i\,(y_i - f_i)^2\}$$

Here yi is the observed data value and fi is the predicted value from the fit. wi is the weighting applied to each data point, usually wi = 1. A value closer to 0 indicates that the model has a smaller random error component, and that the fit will be more useful for prediction.
The histogram that is generated using the data of the continuous attribute and the parameters of the histogram are obtained. The best distribution is chosen by comparing the best parameters with the current parameters. At the end of the comparison, the best distribution for the current continuous attribute is chosen. The same process is repeated for all the continuous attributes and the respective best distributions are obtained. Fig. 8-10. Shows the best distributions and its parameters of all the continuous attributes of the dataset.

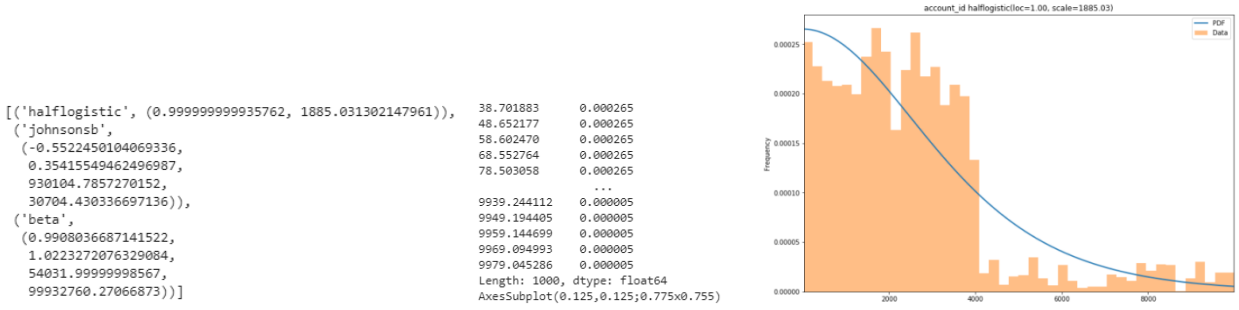



Fig. 8. - 10. List of best distributions and its parameters, values of histogram, histogram and PDF of a continuous attribute

*F) Integration of modules*
The parameters of the best distribution of each continuous attribute are sent to generate the Probability Density Function(PDF) of the corresponding continuous attribute of the dataset. The PDF is generated using PPF(Percent Point Function), which is the inverse of the CDF(Cumulative Distribution Function). The generated PDF can be used to generate random values in accordance with the PDF for the chosen continuous attribute of the dataset. The result is generated by merging the process of anonymizing categorical and continuous columns and applying it on the original dataframe. The resulting data frame is converted to CSV format and is downloaded. The size of the dataset is set to the size of the original dataset after removing the null values.

ISSN(Online): 2319-8753
ISSN (Print): 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(A High Impact Factor, Monthly, Peer Reviewed Journal)

Visit: www.ijirset.com

Vol. 9, Issue 3, March 2020

(230465, 7)

| | type | operation | k_symbol | bank | account_id | date | account |
|---|---|---|---|---|---|---|---|
| 0 | 1 | 1 | 3 | 11 | 452.881306 | 958436.330178 | 2.888777e+07 |
| 1 | 1 | 0 | 3 | 8 | 3091.462799 | 956447.277738 | 3.768307e+05 |
| 2 | 1 | 0 | 3 | 9 | 6440.705771 | 951810.265892 | 4.384225e+06 |
| 3 | 1 | 0 | 0 | 0 | 16.801648 | 960286.966722 | 9.625548e+07 |
| 4 | 1 | 0 | 4 | 7 | 838.150426 | 944710.782464 | 3.490908e+07 |

Fig. 11. Shape and head of the dataset after anonymization

## VI. RESULTS AND ANALYSIS

The best distribution of each continuous attribute is found by the analysis of the parameters of the histogram. The histogram parameters such as arg, loc, scale etc., are calculated by fitting the data of each continuous attribute into the histogram which is separated by the bins. The Probability Density Function (PDF) of the best distribution of the continuous attribute is generated by making use of the Percent Point Function (PPF) over a range of values. It can be found that the PDF of a continuous distribution closely resembles the distribution pattern of the histogram. The matching in pattern is heavily influenced by the closely related parameter of the histogram, which is the Estimated Sum of Squared Errors (SSE).

```
930308.378096    0.000047          9.906644e+05    1.057724e-08
930338.900397    0.000046          1.088643e+06    1.056734e-08
930369.422697    0.000045          1.186622e+06    1.055831e-08
930399.944998    0.000044          1.284601e+06    1.055003e-08
930430.467299    0.000043          1.382580e+06    1.054236e-08
                    ...                               ...
960678.067314    0.000418          9.847969e+07    9.230018e-09
960708.589615    0.000476          9.857767e+07    9.216091e-09
960739.111916    0.000561          9.867565e+07    9.201190e-09
960769.634216    0.000703          9.877363e+07    9.185164e-09
960800.156517    0.001042          9.887161e+07    9.167827e-09
Length: 1000, dtype: float64      Length: 1000, dtype: float64
AxesSubplot(0.125,0.125;0.775x0.755)  AxesSubplot(0.125,0.125;0.775x0.755)
```

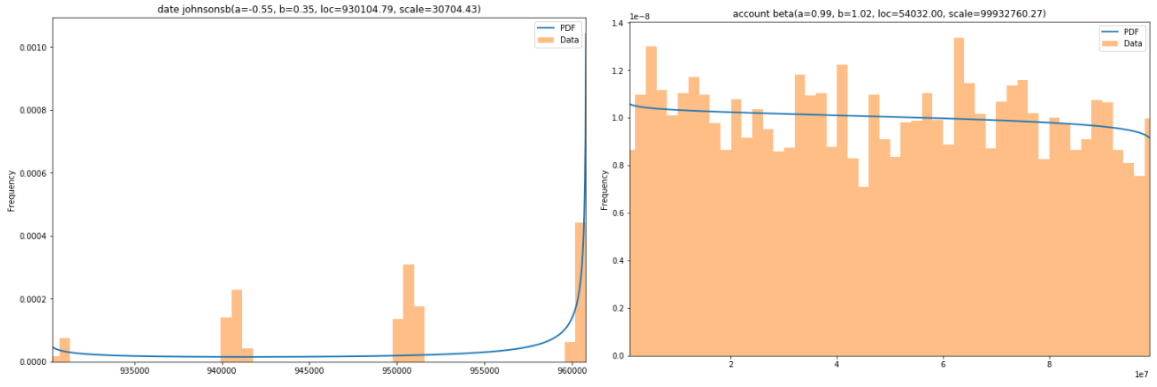



Fig. 12. - 16. Values of histogram, histogram and PDF for the continuous attributes “date” and “account”

The generation of PDF with reasonable accuracy with the fitted data helps to anonymize the continuous attribute without the property being affected. ie., The frequency and probability of occurrence of the unique values of the continuous attributes. The anonymized dataset still has the columns of the dataset named which helps the data recipient with the analysis. It is expected that the analysis performed using the anonymized dataset will give similar results when compared with the analysis performed using the original dataset.

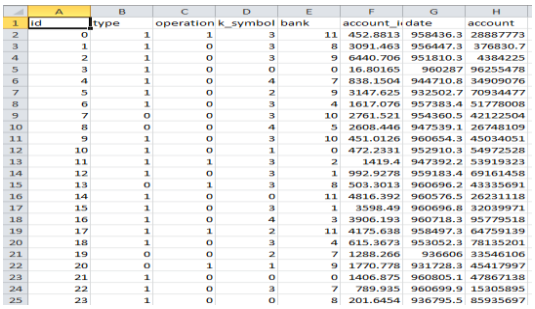
ISSN(Online): 2319-8753
ISSN (Print): 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

*(A High Impact Factor, Monthly, Peer Reviewed Journal)*

*Visit: www.ijirset.com*

**Vol. 9, Issue 3, March 2020**



| | A | B | C | D | E | F | G | H |
|---|---|---|---|---|---|---|---|---|
| 1 | id | type | operation | k_symbol | bank | account_i | date | account |
| 2 | 0 | 1 | 1 | 3 | 11 | 452.8813 | 958436.3 | 28887773 |
| 3 | 1 | 1 | 0 | 3 | 8 | 3091.463 | 956447.3 | 376830.7 |
| 4 | 2 | 1 | 0 | 3 | 9 | 6440.706 | 951810.3 | 4384225 |
| 5 | 3 | 1 | 0 | 0 | 0 | 16.80165 | 960287 | 96255478 |
| 6 | 4 | 1 | 0 | 4 | 7 | 838.1504 | 944710.8 | 34909076 |
| 7 | 5 | 1 | 0 | 2 | 9 | 3147.625 | 932502.7 | 70934477 |
| 8 | 6 | 1 | 0 | 3 | 4 | 1617.076 | 957383.4 | 51778008 |
| 9 | 7 | 0 | 0 | 3 | 10 | 2761.521 | 954360.5 | 42122504 |
| 10 | 8 | 0 | 0 | 4 | 5 | 2608.446 | 947539.1 | 26748109 |
| 11 | 9 | 1 | 0 | 3 | 10 | 451.0126 | 960654.3 | 45034051 |
| 12 | 10 | 1 | 0 | 1 | 0 | 472.2331 | 952910.3 | 54972528 |
| 13 | 11 | 1 | 1 | 3 | 2 | 1419.4 | 947392.2 | 53919323 |
| 14 | 12 | 1 | 0 | 3 | 1 | 992.9278 | 959183.4 | 69161458 |
| 15 | 13 | 0 | 1 | 3 | 8 | 503.3013 | 960696.2 | 43335691 |
| 16 | 14 | 1 | 0 | 0 | 11 | 4816.392 | 960576.5 | 26231118 |
| 17 | 15 | 1 | 0 | 3 | 1 | 3598.49 | 960696.8 | 32039971 |
| 18 | 16 | 1 | 0 | 4 | 3 | 3906.193 | 960718.3 | 95779518 |
| 19 | 17 | 1 | 1 | 2 | 11 | 4175.638 | 958497.3 | 64759139 |
| 20 | 18 | 1 | 0 | 3 | 4 | 615.3673 | 953052.3 | 78135201 |
| 21 | 19 | 0 | 0 | 2 | 7 | 1288.266 | 936606 | 33546106 |
| 22 | 20 | 0 | 1 | 1 | 9 | 1770.778 | 931728.3 | 45417997 |
| 23 | 21 | 1 | 0 | 0 | 0 | 1406.875 | 960805.1 | 47867138 |
| 24 | 22 | 1 | 0 | 3 | 7 | 789.935 | 960699.9 | 15305895 |
| 25 | 23 | 1 | 0 | 0 | 8 | 201.6454 | 936795.5 | 85935697 |

Fig. 17. Resultant Anonymized Dataset

## VII. CONCLUSION

This idea proposes an automated method for performing data anonymization using suppression and randomization techniques. Here, we make use of HIPAA Safe Harbour technique by adopting its principles and merging it with the GDPR guidelines in such a way that the inconsistencies and loopholes in the law can be handled and the failures are minimized. Our proposal makes use of a bank dataset which consists of Personally Identifiable Information (PII). The threshold for separating the sensitive attributes of the dataset is maintained and the attributes of the dataset are anonymized with different approaches that are related to the sensitivity of the unique values (or) degree of privacy of the attribute. Techniques such as label encoding, histograms, probability distributions are used along with the algorithms such as suppression and randomization. The data anonymization process is done in such a way that the dataset supports machine learning. The attribute names of the dataset are not masked in order to help the data recipients handle the anonymized data during analysis.

## ACKNOWLEDGEMENT

We thank faculties from the Department of Computer Science and Engineering, Loyola-ICAM College of Engineering and Technology for their valuable support and guidance.

## REFERENCES

[1] Disha Dubli and D.K Yadav, Department of Computer Applications, National Institute of Technology Jamshedpur, India. “Secure Techniques of Data Anonymization for Privacy Preservation”, *International Journal of Advanced Research in Computer Science*, May-June 2017.
[2] Graham Cormode, Divesh Srivastava. “Anonymized Data: Generation, Models, Usage” *AT&T Labs-Research,*. IEEE, 2017.
[3] TransCelerate BioPharma Inc. “Model Approach: De-identification and Anonymization of Privacy Information in Data .” IEEE, 2016.
[4] Nirzari Patel, Mehul P Barot. “Observations on Anonymization Based Privacy Preserving Data Publishing.” *International Journal of Recent Technology and Engineering (IJRTE).* IJRTE, 2018.
[5] Tiancheng Li, Ninghui Li. “Injector: Mining Background Knowledge for Data Anonymization.” *Department of Computer Science, Purdue University.* IEEE, 2016.
[6] Paul Francis, Sebastian Probst Eide and Reinhard Munz. “Diffix: High-Utility Database Anonymization.” IEEE, 2018.
[7] Dhamodran.P, Priyadharsini.P, Kavitha.M.S. “A Survey on Data Anonymization Techniques for large data sets.” *International Journal of Computer Science and Mobile Computing.* IJCSMC, Nov 2014.
[8] Bin Zhou, Jian Pei, Wo-Shun Luk. “A Brief Survey on Anonymization Techniques for Privacy Preserving Publishing of Social Network Data.” IEEE, 2017.