Different Anonymization Methods' Benchmark on Machine Learning Models

Beyza Gündoğan, Burak Yıldırım, Gizem Güneş, Yusuf Erdemirler

1. INTRODUCTION

In today's rapidly expanding and increasingly complex information landscape, protecting sensitive information is crucial for businesses and organizations. Therefore, it is critical for businesses, organizations, etc. to protect sensitive information. Assuring the secrecy of important data is becoming more and more essential because disclosing consumers' sensitive information might create distrust. In addition, information growth is increasing day by day. As a result, maintaining privacy is getting harder. Finding the best algorithm to generalize the information is therefore necessary. We should, however, try to keep the loss of crucial information to a minimum. As a result, expertise in certain areas of data privacy and security is necessary, and GDPR compliance must be met.

In this project, the database that we use contains sensitive information like individual medical costs billed by health insurance. This information is private for the customers and should be best preserved. Selected generalization algorithms are used to generalize the quasi-identifier. A prediction regarding the sensitive information is made using a machine learning model. The results that the machine learning models gathered are used to draw conclusions. Therefore, the utility loss of different anonymization methods, such as Mondrian, Random Generalization, K-means Clustering, and Microaggression methods, are compared by applying machine learning models.

2. DATASET

a. Dataset Features

The dataset consists of 7 features which are **age, sex, bmi, children, smoker, region, and charges** [1]. Age, sex, bmi, children, smoker, and region are Quassi-Identifiers. Charges is a sensitive attribute. Bmi is the body mass index of the individual. Children stands for the number of children that are covered by health insurance. Region corresponds to the living area of the individual. Charges is the medical costs that are costs for medical treatments covered by health insurance. The dataset has 1338 rows. Some samples from raw data can be seen in Figure 1.

age	sex	bmi	children	smoker	region	charges
19	female	27.9	0	yes	southwest	16884.924
18	male	33.77	1	no	southeast	1725.5523
28	male	33	3	no	southeast	4449.462
33	male	22.705	0	no	northwest	21984.47061
32	male	28.88	0	no	northwest	3866.8552
31	female	25.74	0	no	southeast	3756.6216
46	female	33.44	1	no	southeast	8240.5896
37	female	27.74	3	no	northwest	7281.5056
37	male	29.83	2	no	northeast	6406.4107
60	female	25.84	0	no	northwest	28923.13692

Figure 1: Samples of Raw Data

b. Preprocessing

In the preprocessing step, bmi values are rounded to an integer to have a better DGH (domain generalization hierarchy) structure. To turn it into a binary classification problem, charges are converted from float values to labels as greater or equal to 10000 and less than 10000. The dataset is examined by various tools of the pandas library in a Jupyter Notebook. Categorical and continuous variables are identified. Preprocessed data samples can be seen in Figure 2.

age	sex	bmi	children	smoker	region	charges
19	female	28	0	yes	southwest	>=10000
18	male	34	1	no	southeast	<10000
28	male	33	3	no	southeast	<10000
33	male	23	0	no	northwest	>=10000
32	male	29	0	no	northwest	<10000
31	female	26	0	no	southeast	<10000
46	female	33	1	no	southeast	<10000
37	female	28	3	no	northwest	<10000
37	male	30	2	no	northeast	<10000
60	female	26	0	no	northwest	>=10000

Figure 2: Preprocessed Data Samples

c. DGHs

DGHs are constructed considering the possible levels for each quassi identifier. In Figure 3, DGH for bmi and region can be seen respectively. While generalization, numeric values are placed in intervals and categoric values are placed in a broader category.

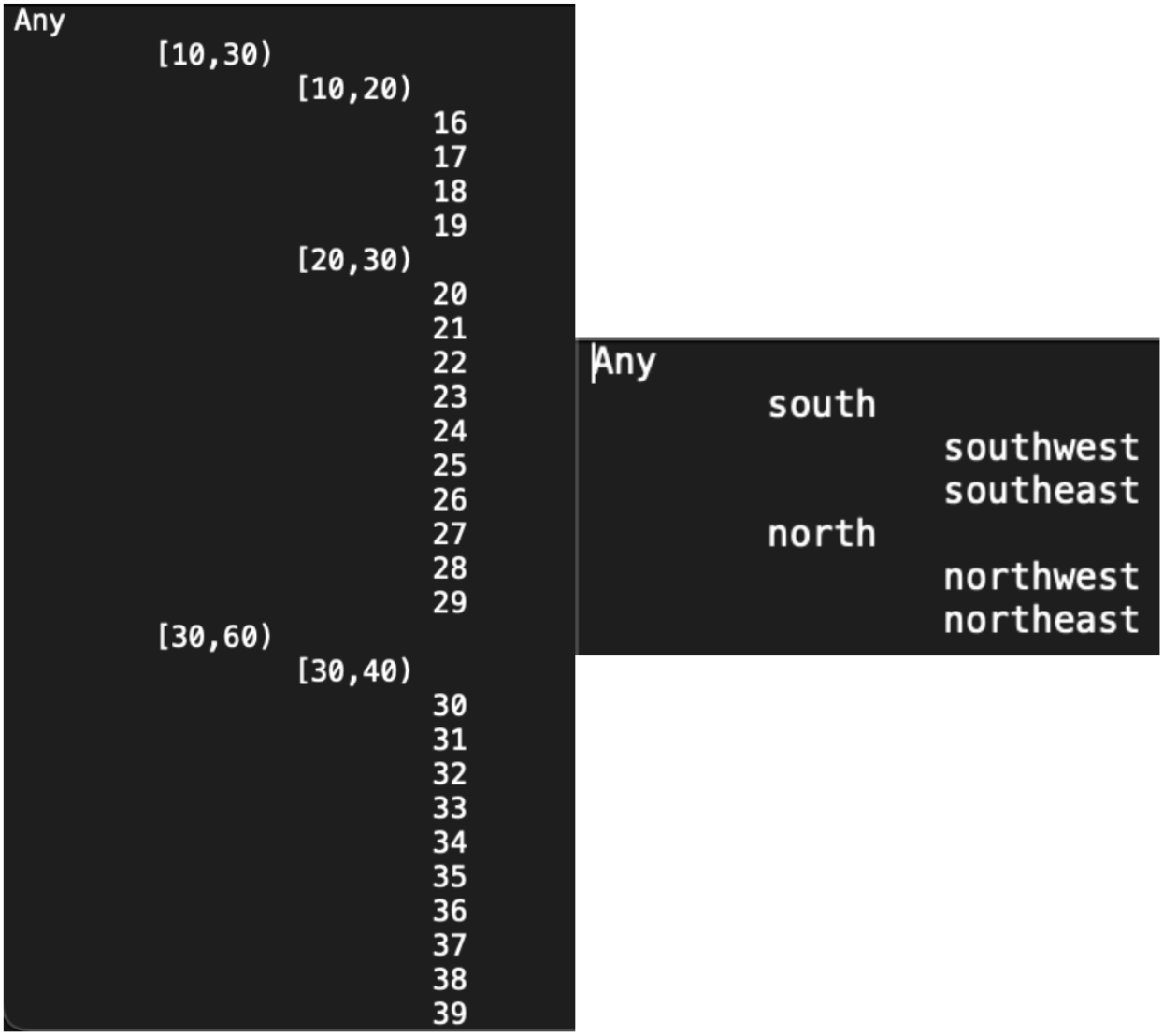

Figure 3: DGH examples

3. ANONYMIZATION METHODS

3 different anonymization methods were used in this project: Random Generalization, K−means Clustering and Microaggregation, and Mondrian. Two of them, Random Generalization, and K−means Clustering and Microaggregation were performed by using the ARX Tool. Partitioning-based anonymization was performed by the Mondrian algorithm. In all of these methods, K-anonymity was achieved with different approaches.

a. Random Generalization

Random Generalization method was performed using ARX Tool. After importing the data to ARX, domain generalization hierarchies were created. Since the dataset has a sensitive attribute which is *charges*, ARX Tool made required to apply ℓ-diversity. Distinct ℓ-diversity was chosen with the parameter ℓ = 2 for all the datasets anonymized with random generalization. On the other hand, to achieve k-anonymous datasets, 4 different k values were used: k = 5, k= 10, k = 20, and k = 50. An overview of prospective solutions can be examined from the exploration perspective after the anonymization procedure on the ARX interface [2]. Optimal default anonymization strategies found by ARX Tool were used to produce anonymized datasets.

b. K−means Clustering and Microaggregation

As in the Random Generalization method, ARX Tool was used to perform K-means Clustering and Microaggregation method. Same domain generalization hierarchies and the same ℓ and k parameters were used to achieve distinct ℓ-diversity and k-anonymization. Again, optimal anonymization strategies found by ARX Tool were used to produce anonymized datasets.

c. Mondrian

Mondrian is known for being an anonymization method that has a partitioning-based top-down greedy approach. With the help of kd-tree, it divides the original dataset into k different groups. Then, it generalizes each group to establish an equivalence class [3]. Basic Mondrian is a variant of Mondrian that uses generalization hierarchies is used for our research [4]. Domain generalization hierarchies are created according to our dataset. Then, basic Mondrian is altered to adapt to our dataset. Several anonymized data are obtained with different K values.

4. MACHINE LEARNING MODELS

3 different machine-learning models were used in this project: Random Forest, XGBoost, and k-Nearest Neighbor. These models are selected due to their popularity and well performance for classification problems by each having unique characteristics.

a. Random Forest

Random forest classifier uses many decision trees to predict the output label. Each decision tree gives an independent prediction but in the end the final output will be the class that is predicted by the majority of the decision trees. In addition, another technique that is used by random forest is feature bagging which enables the classifier to pick random feature subsets used in each tree to increase variety and accuracy.

b. XGBoost

XGBoost means an extreme gradient boosting algorithm. XGBoost classification works using many decision trees. Unlike random forest, each decision tree corrects the mistakes of the prior decision tree to produce an output with less error. It can process tasks simultaneously so that it is fast and efficient compared to other algorithms. It uses regularization to overcome overfitting which is learning the training data too well that it might perform poorly on the test data.

c. k-Nearest Neighbor

kNN (k-Nearest Neighbor) looks for the nearest neighbor of a data point utilizing a distance metric such as euclidean and manhattan distance. In the experiments, k value is chosen as 5 because it lead to a good fit. If it was chosen too low it could create overfitting or if it was chosen too large it could create underfitting problems. Euclidean distance is selected as the distance metric to find the closest distance to other instances. After training the model, in the test data, it looks for the closest 5 neighbors and labels the data as the majority class among these neighbors.

5. EXPERIMENTS

To conduct the experiments, original and anonymized datasets were trained and tested on all the ML models mentioned above. Evaluation metrics are selected as accuracy, precision, and

recall values. They are taken into consideration to compare the performance of ML models on different anonymized data.

The repository of our project can be accessed in Github [5].

6. RESULTS

Accuracy: $\frac{T_p}{T_p + F_p}$

Precision: $\frac{T_p}{T_p + F_n}$

Recall: $\frac{T_p + T_n}{T_p + F_p + F_n + T_n}$

The results indicates the evaluation metrics for 3 different types of machine learning models which are Random Forest, XGBoost, and kNN in order. The models are executed with original data and anonymized data with k-anonymity values of 5,10,20, and 50 respectively. Accuracy, precision, and recall are used as evaluation metrics to compare algorithms.

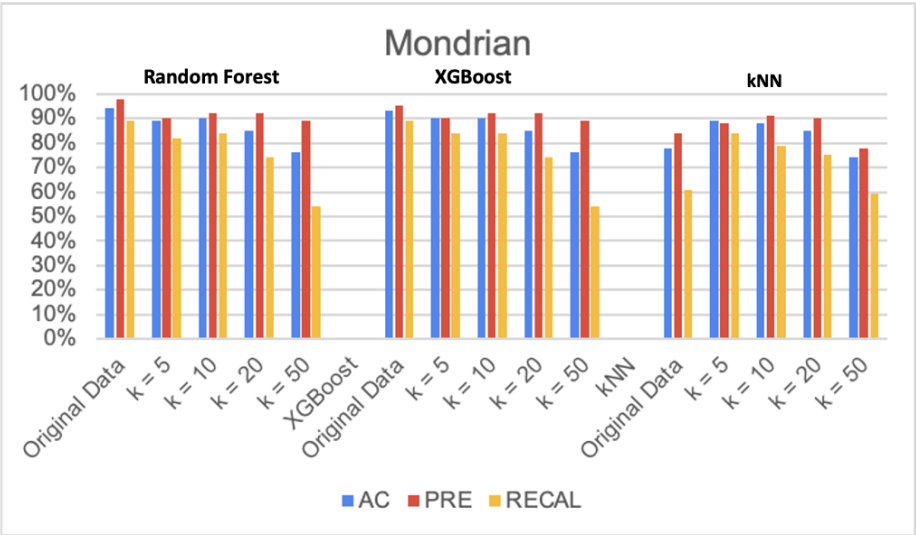

Figure 4: Mondrian Results

In Figure 4, the results of the Mondrian illustrates the accuracy values are decreasing while we are increasing the value of k but as seen in the figure this change is not drastic. It is expected because higher anonymity leads to higher utility loss. Random Forest model and

XGBoost have similar patterns in terms of evaluation metrics but kNN algorithm has a lower performance.

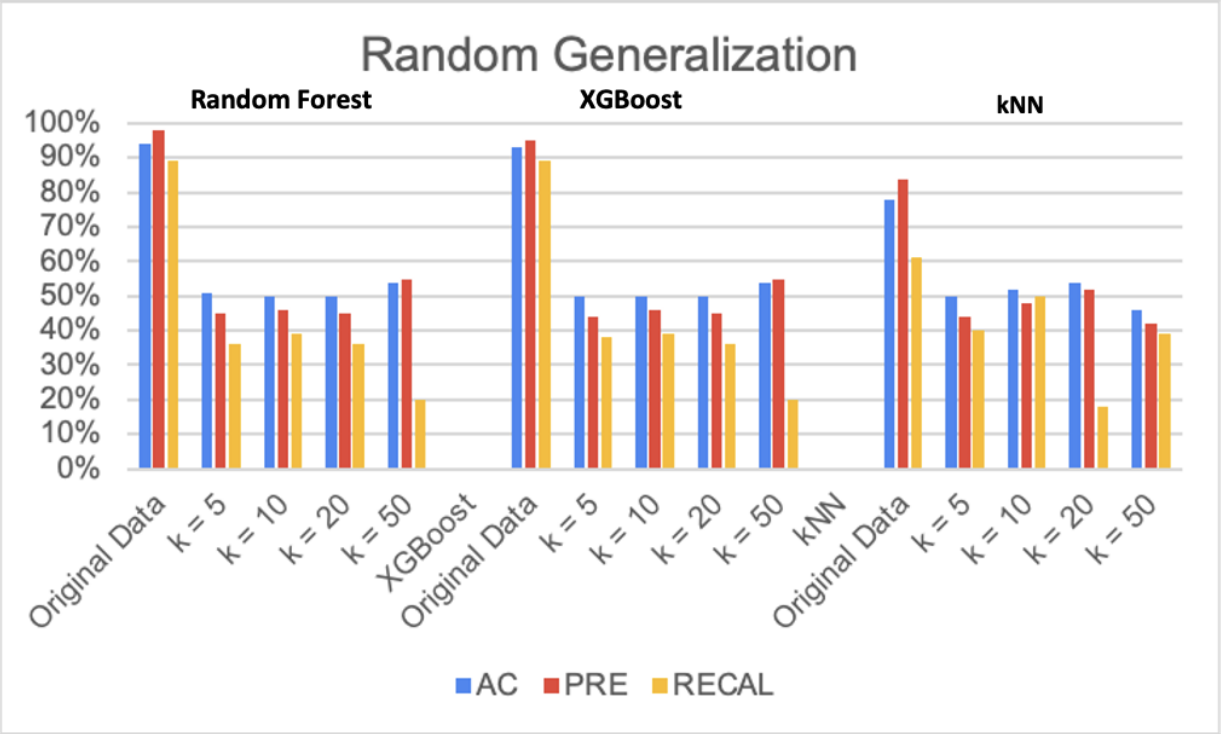

Figure 5: Random Generalization Results

As depicted in Figure 5, random generalization obtained accuracy scores at around 50%. Since it decreased from around 90% to around 50%, utility loss is high. All of the models achieved similar results due to the working principles of random generalization and lack of data.

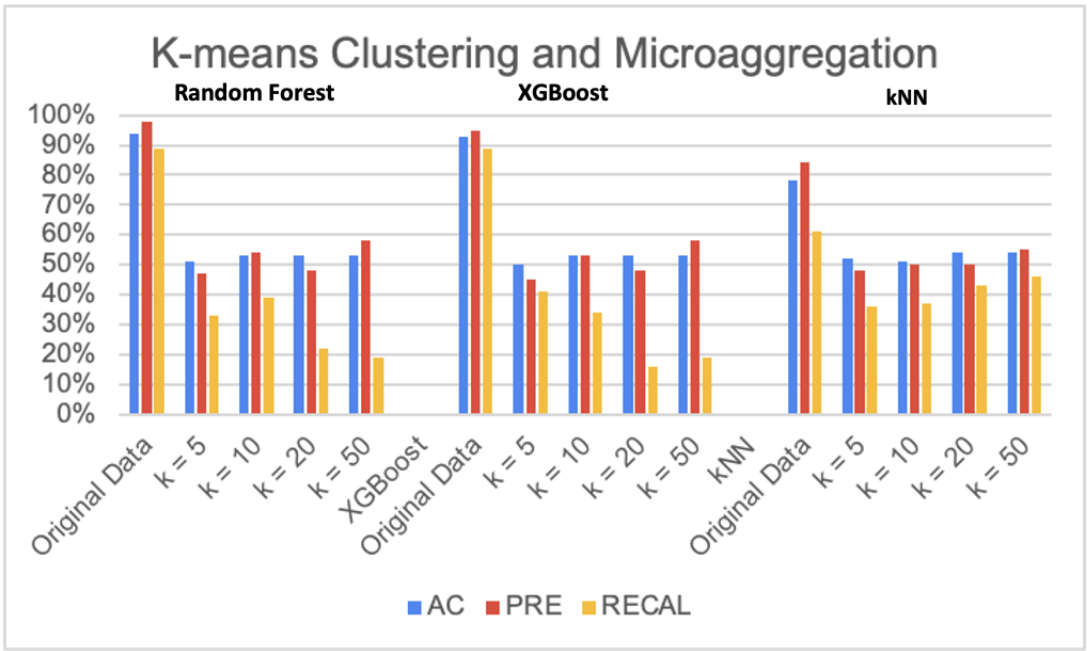

Figure 6: K-means Clustering and Microaggregation Results

The insights that can be observed from Figure 6 are k-means clustering and microaggregation method gave similar accuracy, precision, and recall values for all models. When k values are increased accuracy remained almost the same in all models. There is a similar relationship between k-means clustering and microaggregation and random generalization. This could be a consequence of both of them are being produced by ARX.

7. CONCLUSION

In this project, we compared the utility loss of the Mondrian, Random Generalization, K-means Clustering and Microagression anonymization techniques for machine learning models. The machine learning techniques Random Forest, XGBoost, and k-Nearest Neighbors were used. The models are run on both the raw data and data that has been anonymised. We anonymized the data with different k-values. Such as 5, 10, 20, and 50.

The results show that Mondrian returned more accurate outputs by minimizing the utility loss. Therefore, the use of Mondrian in the industry is of great benefit to the companies that use it. It not only improves the data preservation but also effectively reduces the loss of crucial information. The results also demonstrated that high anonymity led to high utility loss

The dataset we used contains fewer data. Therefore the data may have shown results close to each other. Furthermore, in future research, we can try this project with larger datasets and we can experiment with using many more techniques. Increasing diversity can enhance the ability to detect unusual data. In addition, to effectively preserve privacy, data anonymization is crucial since it obscures any personal information that could be used to identify a specific person or their personal information. This is especially crucial when dealing with sensitive information that could be detrimental to a particular person or a company if it ended up in the wrong hands.

REFERENCES

[1] M. Choi, “Medical Cost Personal Datasets,” *Kaggle*, 21-Feb-2018. [Online]. Available: https://www.kaggle.com/datasets/mirichoi0218/insurance. [Accessed: 26-Jan-2023].

[2] “Anonymization tool,” *ARX*. [Online]. Available: https://arx.deidentifier.org/anonymization-tool/. [Accessed: 26-Jan-2023].

[3] K. LeFevre, D. J. DeWitt, and R. Ramakrishnan, “Mondrian multidimensional K-anonymity,” *22nd International Conference on Data Engineering (ICDE'06)*, 2006.

[4] Qiyuangong, “Qiyuangong/mondrian: Python implementation for Mondrian Multidimensional K-anonymity (Mondrian).,” *GitHub*. [Online]. Available: https://github.com/qiyuangong/Mondrian. [Accessed: 26-Jan-2023].

[5] ggunes18, “GGUNES18/anonymization-on-ML-models: Comp430 term project - fall2022,” *GitHub*. [Online]. Available: https://github.com/ggunes18/anonymization-on-ml-models. [Accessed: 26-Jan-2023].