Comparing Accuracy Among Various ML Models on Synthetic Data for Preserving Privacy

Md. Ashraf Uddin
University of Manitoba
Winnipeg, Canada
uddinma1@myumanitoba.ca

ABSTRACT

Data is a very important thing in today's world. It is very confidential information everywhere such as in hospitals, banks, government sectors, etc. Nowadays, data is very sensitive in every sector, therefore data privacy is becoming a popular field to research. A lot of techniques have been used for the last two decades for ensuring data privacy. Today, synthetic data generation is one of the most popular ways of data protection. All organizations like the medical sector, banking sector, and government sectors are very careful about data privacy. As a result, researchers have developed synthetic data generation techniques, that will create synthetic data depending on the real dataset. In our work, we will show the accuracy of synthetic data compared to the real dataset using some basic machine learning models (Logistic Regression, Random Forest Classifier, Decision Tree, and Support Vector Machine). Here, we will use three real datasets of cancer patients, train and test that dataset with the four machine learning models, and find accuracy, precession, and recall values for the real data. At the same time, we will generate synthetic data using the real dataset and train it and test it with the real dataset, and find accuracy, precession, and recall values for the synthetic dataset. Finally, we will show the comparison between synthetic data and real data for accuracy, precession, and recall values by applying four machine learning models and decide which model gives the best result for which dataset.

KEYWORDS

Synthetic data, machine learning models, privacy, cancer dataset, and Gretel.

ACM Reference Format:
Md. Ashraf Uddin. 2018. Comparing Accuracy Among Various ML Models on Synthetic Data for Preserving Privacy. In *Proceedings of Make sure to enter the correct conference title from your rights confirmation emai (Conference acronym 'XX)*. ACM, New York, NY, USA, 9 pages. https://doi.org/XXXXXXX.XXXXXXX

Conference acronym 'XX, June 03–05, 2018, Woodstock, NY

ACM ISBN 978-1-4503-XXXX-X/18/06…$15.00
https://doi.org/XXXXXXX.XXXXXXX

1 INTRODUCTION

Hospital is a very sensitive data source because in that dataset there is a lot of confidential information like name, age, gender, location, disease, etc. Therefore, the hospital authority does not want to disclose the dataset to the public, even for research purposes [1]. As a result, it is very difficult for researchers to work with the real dataset for proposing some mechanism to ensure data privacy. Moreover, some issues have to face while working with the dataset like the limited availability of or access to the healthcare dataset. For solving this situation, researchers have developed a technique that makes a new dataset depending on the features and attributes of the real dataset, and this technique is called synthetic data generation.

In our research work, for ensuring data privacy we have used the synthetic data generation technique. Here, we have selected three medical datasets [2,3] for our task, and all of the datasets are cancer-related. In our experiment, Gretel [4] is used for synthetic data generation purposes. Initially, we train four machine learning models (Logistic Regression, Random Forest Classifier, Decision Tree, and Support Vector Machine) with the real dataset and test them with the real data. Here, we have taken 80% of the dataset for the training purpose and 20% of the dataset for the testing purpose and we note accuracy, precession, and recall values for this work.

Next, we generated synthetic data using Gretel for our experiment and tested it with the real dataset. In Gretel, we have modified the tokenizer section for getting more accurate tokens to get perfect results. In our experiment, we have used 20% of the real data for testing synthetic datasets. In every case, we have used three cancer datasets for our experimental purpose and we calculated accuracy, precession, and recall values for both synthetic data and real dataset. In our study, we find that all machine learning models perform almost the same both for the real and synthetic datasets. Among all of the models, three models (logistic regression, random forest, and support vector classifier) work accurately, but the decision tree gives low results in comparison with other models. Moreover, all models perform properly in the breast cancer-1 and cancer datasets, but a little bit of low performance is detected from the breast cancer-2 dataset.

All of our experiments are described below sections. In the second section, background and related works are described. The data selection section is used for describing three dataset and their records in the third section. In the fourth section, we have described our detailed methodology and synthetic data generation procedures. Finally, the fifth section is used for experimental result analysis. The sixth section describes the discussion part of our whole experiment and the last section is used for the conclusion and future works.

2 BACKGROUND AND RELATED WORKS

Synthetic data generation is a machine learning-based algorithmic technique that is essential for creating an artificial dataset from any sample of real-world events. It is generated programmatically and used for testing datasets of original or production data. Moreover, it is used to validate mathematical models and train various machine learning models. This synthetic dataset is created from the real data depending on the same statistical characteristics, features, and attributes of the real dataset. It represents the original data and does not corrupt the vital attributes of the real dataset. There are three types of synthetic datasets: fully synthetic, partially synthetic, and hybrid synthetic data and they can be used in many industries like financial services, automotive and robotics, healthcare, manufacturing, security, social media, etc. In the case of the medical datasets, it does not publish the real dataset to the public because of many confidential issues. Hence, synthetic data plays an important role in data privacy.

Since 2003, the US Census Bureau has been investigating the validity and disclosure risk of synthetic data, which is created by combining sensitive data from the Census Bureau's Survey of Income and Program Participation, the Internal Revenue Service's individual lifetime earnings data, and the Social Security Administration's individual benefit data to create public-use data [19, 20]. The goal was to make it feasible to share synthesis person-level records from private datasets that included personal and financial information while protecting privacy. Synthetic data files for public use have been published as a result of the favorable outcomes.

The public release of this data has benefitted both the academic community and the general public by allowing parties that previously did not have access to critical data to do more in-depth economic policy analyses. Data from an annual economic census of US firms make up the Synthetic Longitudinal Business Database [21]. Synthetic data has also been examined in the United Kingdom as a means of allowing public access to rich data from UK longitudinal studies [22-23] that contain highly sensitive data linking national census data to administrative data for persons and their families. These datasets allow researchers to explore data and develop and test code and models outside of the secure environment where real data is stored without restrictions, while data owners provide a mechanism for researchers' results, code, and models to be validated on real data within the secure environment and feedback to be provided. This method boosts research output while also assuring the construction of reliable and accurate models [24].

A huge number of works have been done by many researchers in this area. Various data perturbation techniques are used for data masking, data swapping and adding noise to modify the dataset before publishing the dataset to the public for ensuring data privacy. However, those methods do not reduce data disclosure risk and those techniques also lose data utility. Rubin D.B [5], and Little R. [6] proposed the synthetic dataset for the first time. Further, Raghunathan et al. [7] implemented it for the first time and extended it. Later, many researchers worked on that topic [8-13]. A nonparametric tree-based technique that uses classification and regression trees (CART) was used for data synthesizing [14]. A recent technique is proposed using the Bayesian network approach for synthetic data generation [15].

Synthetic data is seen as a secure means of disseminating sensitive information to the general public. It creates data that is outside the scope of traditional de-identification techniques. A fictional dataset that contains no genuine data but can be traced back to the original data source, keeping the original data's accurate statistical properties. As a result, reverse engineering, or the identification of an actual individual is thought to be improbable [16]. This study [17] builds on a previous study to see if completely synthetic data can maintain the hidden complex patterns that supervised machine learning can uncover from real data, and thus if it can be used as a legitimate substitute for real data when developing eHealth apps and health-care policy solutions. In this study [18], the authors created synthetic data and used five machine learning models to compare the accuracy of the synthetic data to the accuracy of the real dataset.

3 DATA SELECTION

For our experiment, we have selected 3 datasets that are cancer-related. We have collected our dataset from the University of California Irvine Machine Learning Repository [3], and Kaggle [2]. Missing values have been removed from the datasets by eliminating features having a large number of missing values. Table 1 summarizes the experimental datasets and their attributes. These datasets are chosen to allow for an examination of synthetic data performance when we have applied them to datasets of various sizes and sorts.

4 METHODS

4.1 Methodology

In figure 1, we have shown our proposed methodology. Our main target is to generate synthetic data and finding accuracy of the generated data compared with the accuracy of the real dataset. In our proposed design, we have taken cancer-related three real datasets. After getting raw medical data we have to process it. In this step, we have distinguished categorical and numerical features. Then we ensure the quality by removing different anomalies like mixed data values, special characters, and mismateched data types. Data cleaning step is a very important step that is the process of adding missing data and correcting, repairing, or removing incorrect or irrelevant data from a data set. In this step, we ignore tuples, sometimes manually fill the missing data, etc. After completing data preprocessing step, we use it for two purposes: one is synthetic data generation and the other is to train machine learning models. In our experiment, we have used four common machine learning models (Logistic Regression, Random Forest Classifier, Decision Tree, and Support Vector Machine) for our result analysis purposes. We have split our dataset into two parts: 80% dataset is used for the training purposes and 20% is used for the testing purposes. When all of the models are trained with the real dataset we have tested it with the 20% real dataset, and we have calculated accuracy, precession, and recall values for our experimental necessity. Moreover, the original dataset is also used for synthetic data generation. Here, we have generated more than 1,10,000 records from three datasets.

In our experiment, we have used Gretel for synthetic data generation. When data is generated, we have trained our models with the synthetic data. When training is accomplished we have tested our trained data with the real dataset. Here, we have also used a

Table 1: Dataset selection

Dataset name	Attributes	Categorical attributes	Numerical Attributes	Number of records	Synthetic records	Data sources
Breast cancer-1	31	1	30	570	30,000	UCI ML repository
Breast cancer-2	10	0	10	117	50,000	Kaggle
Cancer	10	0	10	859	30,000	Kaggle

20% real dataset for testing purposes. When testing is done, we get accuracy, precession, and recall values for 4 machine learning models. At last, we compare both results (real dataset result and synthetic dataset result) and find which model performs best in which dataset.

4.2 Synthetic Data Generation

Generally, some basic steps are required to generate synthetic data. In figure 2, a general structure is shown for synthetic data generation. Raw data is collected from any data source and data could be in .csv format or .txt format. After getting raw data we have to go to the preprocessing step where we clean data, create features, and normalize it for the generator. Then we have used the processed dataset for training purposes. Lastly, after training the generator synthetic dataset will be generated and it can be used for various purposes.

In our experiment, we have used Gretel for synthetic data generation. In Gretel, there are four main components of synthetic data generation. Those are configurations, tokenizers, training, and generation.

Configurations are classes that are unique to the underlying machine learning engine that is used to train and create data. For example, using TensorFlowConfig to build all of the necessary parameters to train a model based on TF. For backward compatibility with previous versions of the library, LocalConfig is aliased to TensorFlowConfig. A model is saved to a specified directory, which may be archived and used later if desired.

Tokenizers convert input text into tokens that are used by the ML engine and these tokens are used as training input. In this step of synthetic data generation, we have modified the existing token generation technique and applied Gensim for getting more accurate results. Here, CSV file is used as an input for the token generator.

Training a model combines the configuration and tokenizer to create a model that can be used to produce new data and is saved in the appropriate directory.

Any number of new lines or records can be generated once a model has been trained and finally the synthetic data will be produced.

Figures 3 and 4 are used for breast cancer synthetic data generation. 30,000 synthetic data are generated from the real breast cancer dataset. Figure 3 shows the correlation between training and synthetic data and the right plot shows the correlation difference. This correlation figure shows synthetic data generation is good enough with compare to the training dataset and it is shown that the correlation difference is not very much. Component analysis shows that generated synthetic dataset is also similar to the training set.

Figures 5 and 6 are used for the breast cancer second dataset. Here, 50,000 data are generated from the real dataset. Figure 5 and figure 6 show correlation and principal component analysis respectively. Moreover, figures 7 and 8 show the correlation graph and component analysis graph for the cancer dataset and here, 30,000 synthetic datasets are generated from the real dataset.

5 EXPERIMENTAL RESULT ANALYSIS

In our experiment, initially, we work with the real dataset and then we analyze the synthetic dataset. In figure 9, it is shown the correlation matrix both for real and synthetic datasets for breast cancer. A correlation matrix is a tool for summarizing a huge dataset, identifying trends, and making decisions based on them. In this figure real dataset correlation matrix is shown on the left side whereas the synthetic dataset correlation matrix is shown on the right graph. From the matrix, we can see that the synthetic data generated matrix is almost similar to the real dataset correlation matrix. In the next two figures 10 and 11, we show the same experimental output for the rest two datasets. In every case, we can see that the synthetic data generated correlation matrix is close to the real dataset correlation matrix.

In table 2, we show accuracy, precession, and recall values both for the real and synthetic datasets. In our experiment, we have used 4 different machine learning models. From the table, we can see that accuracy, precession, and recall values for logistic regression both for the real and synthetic datasets are almost the same. The decision tree does not perform well for both datasets whereas the rest other models give more accurate results.

In table 3, we can see that all models can not perform well for the real dataset. In this dataset, accuracy, precession, and recall values are not satisfactory according to the previous dataset. Among all four models, random forest performs worse in the real dataset, but for the synthetic dataset, it works best. Support vector classifier works better compare to the rest other models for both real and synthetic datasets. Almost all models perform best for the third dataset, in table 4. Comparatively decision tree does not perform well in this dataset. Otherwise, logistic regression, random forest, and support vector classifier work well.

Figure 12, represents the accuracy between real and synthetic datasets for breast cancer-1. It is shown that all models perform well for both datasets and the accuracy is around 90%. Here, logistic regression, random forest, and support vector classifier give the accuracy of more than 90% and the decision tree is close to 90%. From figure 13, the graph shows different values depending on 4 machine learning models. For the breast cancer-2, dataset only logistic regression works well for the real dataset, whereas the rest of other models' accuracy is below 80% and random forest accuracy for the real dataset is below 70%, but all models perform properly

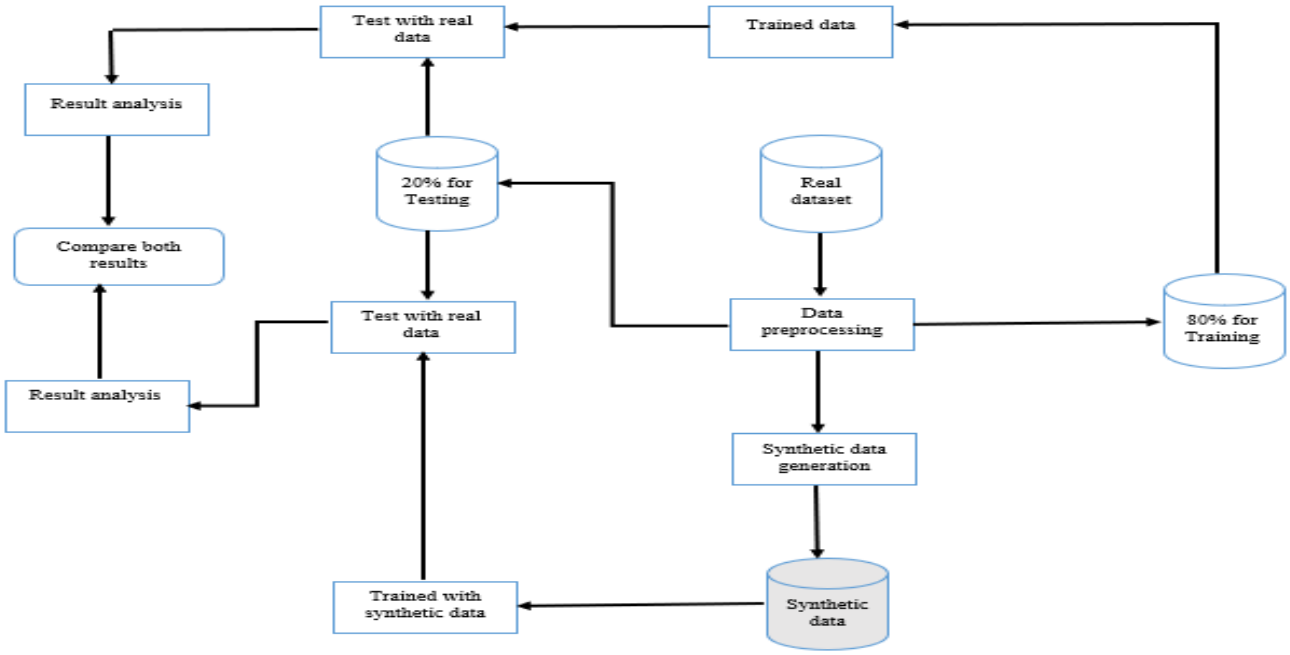

Figure 1: Proposed methodology

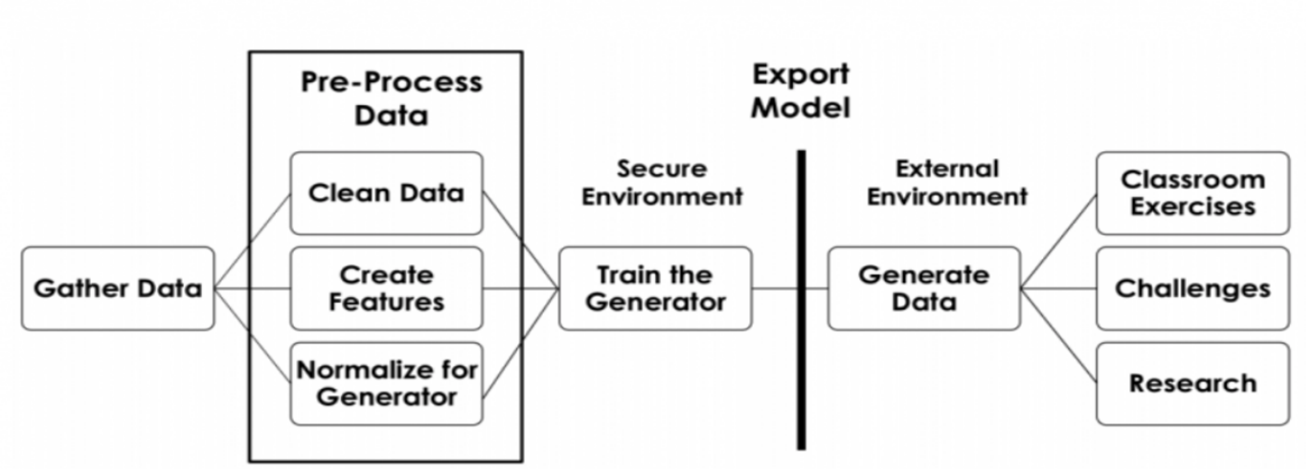

Figure 2: Synthetic data generation

Table 2: Comparing accuracy, precession, and recall values for both real and synthetic dataset (breast cancer-1)

Model name	Accuracy (Real)	Accuracy (Synthetic)	Precession (Real)	Precession (Synthetic)	Recall (Real)	Recall (Synthetic)
Logistic Regression	0.92	0.93	0.93	0.93	0.92	0.93
Random Forest Classifier	0.92	0.93	0.92	0.93	0.92	0.93
Decision Tree	0.89	0.91	0.89	0.91	0.88	0.91
Support Vector Classifier	0.92	0.93	0.93	0.93	0.92	0.93

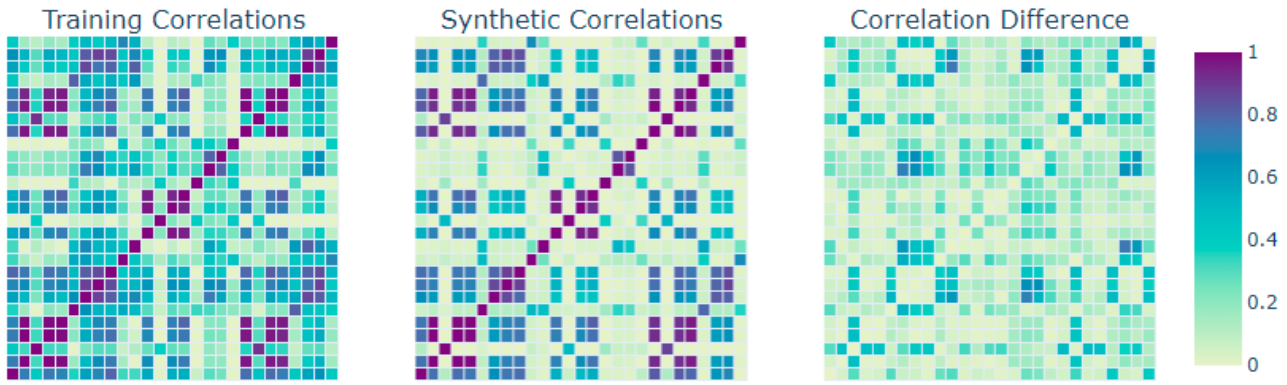

Figure 3: Training and synthetic data correlation (breast cancer dataset-1)

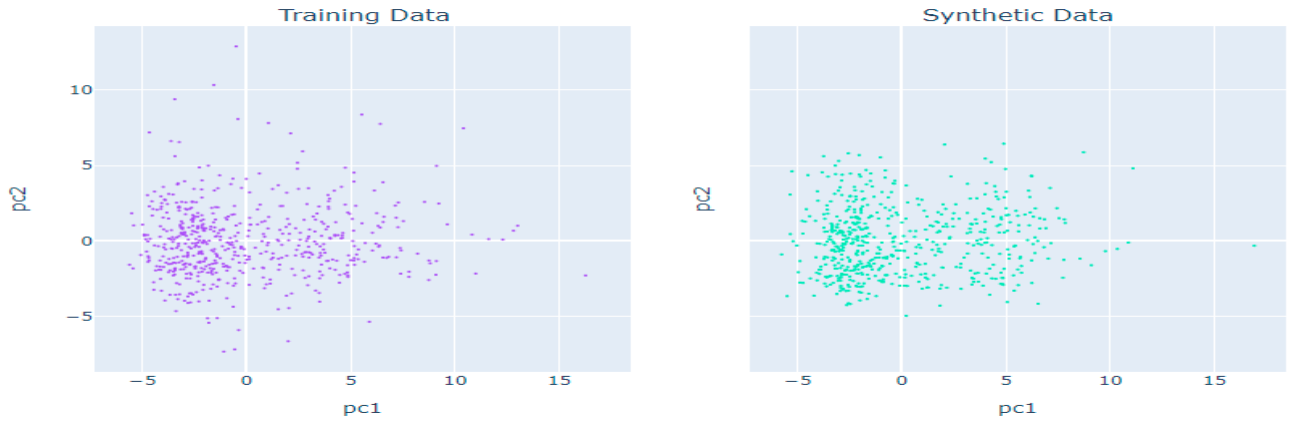

Figure 4: Principal component analysis (breast cancer dataset-1)

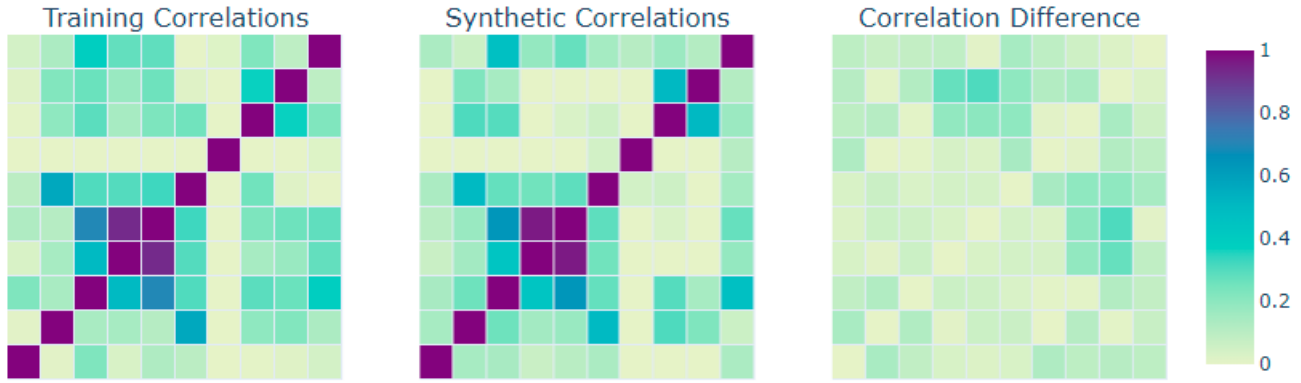

Figure 5: Training and synthetic data correlation (breast cancer dataset-2)

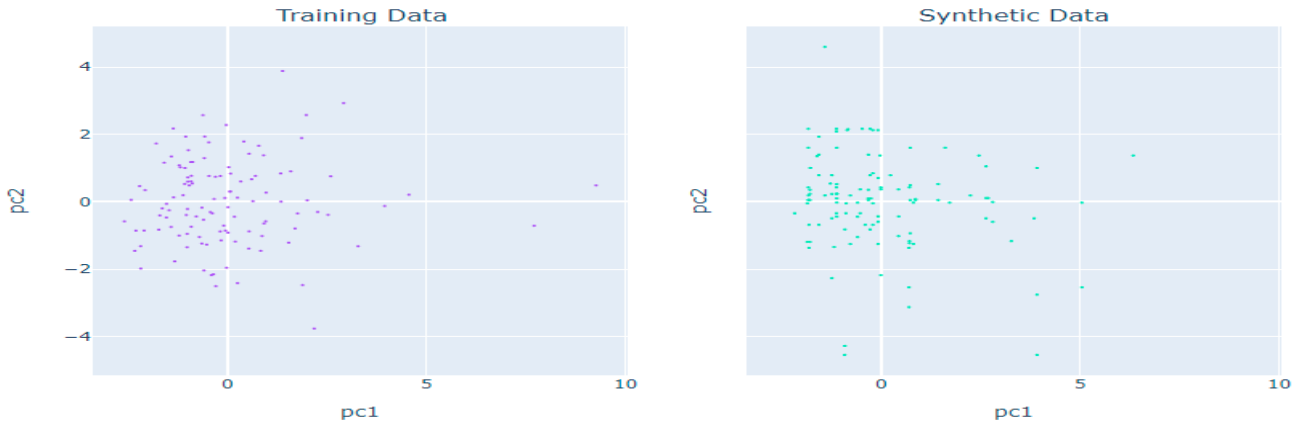

Figure 6: Principal component analysis (breast cancer dataset-2)

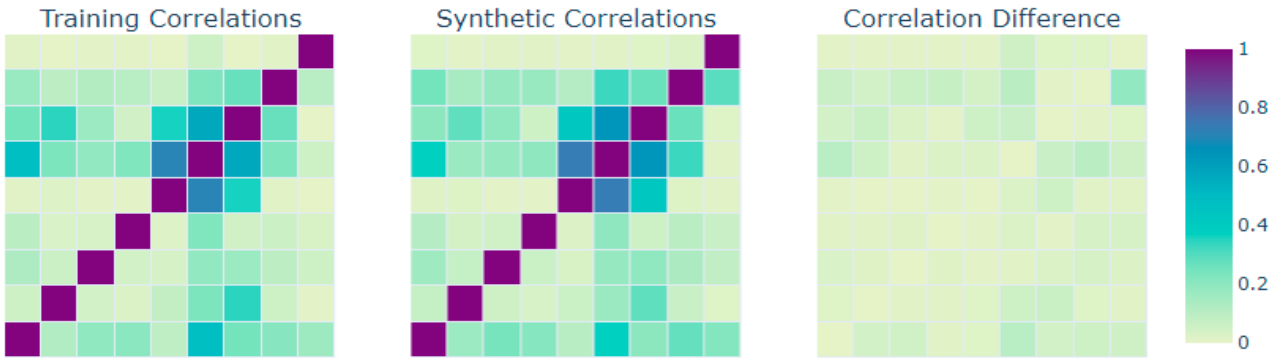

Figure 7: Training and synthetic data correlation (cancer dataset)

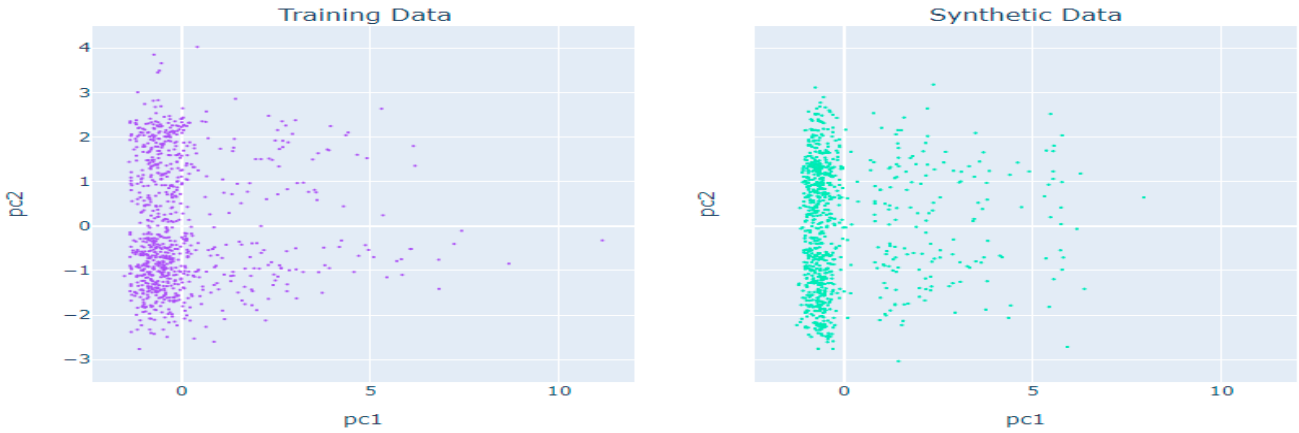

Figure 8: Principal component analysis (cancer dataset)

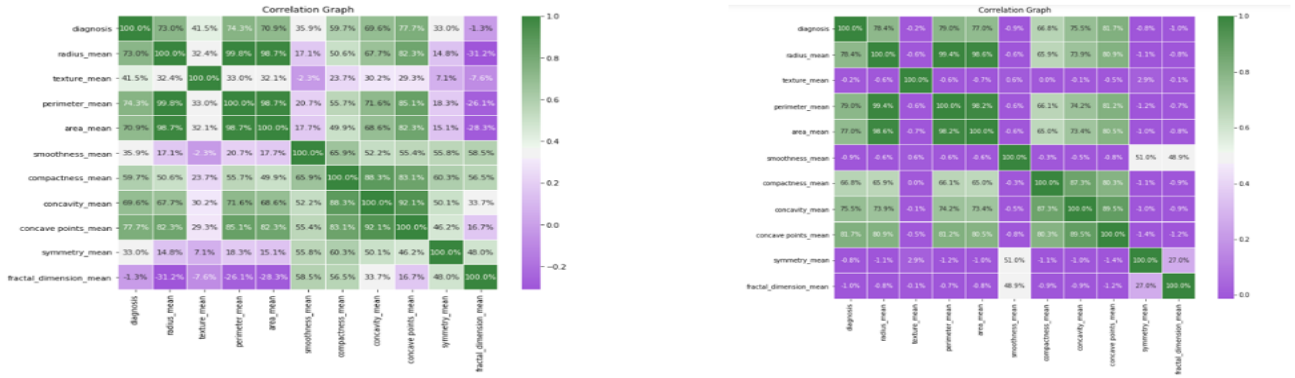

Figure 9: Correlation matrix for real dataset (breast cancer-1 left) and synthetic dataset (breast cancer-1 right)

Figure 10: Correlation matrix for real dataset (breast cancer-2 left) and synthetic dataset (breast cancer-1 right)

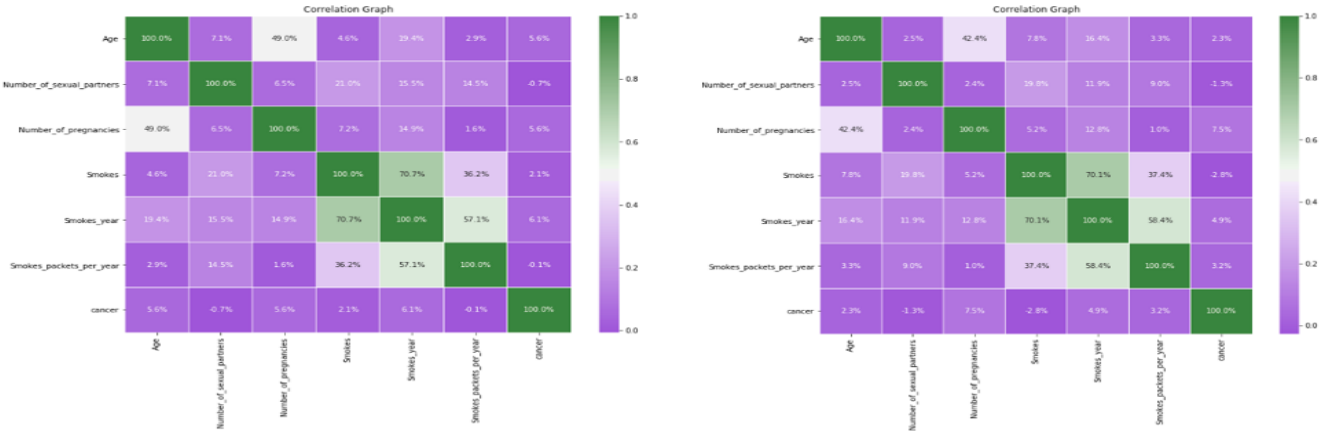

Figure 11: Correlation matrix for real dataset (cancer left) and synthetic dataset (cancer right)

Table 3: Comparing accuracy, precession, and recall values for both real and synthetic dataset (breast cancer-2)

Model name	Accuracy (Real)	Accuracy (Synthetic)	Precession (Real)	Precession (Synthetic)	Recall (Real)	Recall (Synthetic)
Logistic Regression	0.88	0.79	0.85	0.84	0.92	0.76
Random Forest Classifier	0.67	0.97	0.67	1.00	0.67	0.94
Decision Tree	0.79	0.92	0.77	0.98	0.83	0.88
Support Vector Classifier	0.79	1.00	0.77	1.00	0.83	1.00

Table 4: Comparing accuracy, precession, and recall values for both real and synthetic dataset (cancer)

Model name	Accuracy (Real)	Accuracy (Synthetic)	Precession (Real)	Precession (Synthetic)	Recall (Real)	Recall (Synthetic)
Logistic Regression	0.95	0.93	0.90	0.86	0.95	0.93
Random Forest Classifier	0.92	0.95	0.91	0.71	0.92	0.63
Decision Tree	0.85	0.95	0.91	0.67	0.85	0.59
Support Vector Classifier	0.95	0.93	0.90	0.83	0.95	0.93

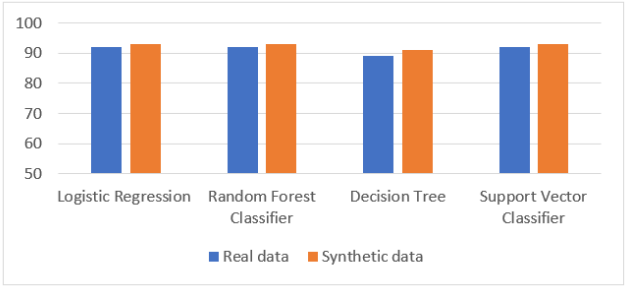

Figure 12: Comparing accuracy between real and synthetic dataset for breast cancer-1

Figure 13: Comparing accuracy between real and synthetic dataset for breast cancer-2

with the synthetic dataset because of 50,000 records are generated synthetically. Figure 14, shows the result for the third dataset (cancer). Here, 30,000 records are synthetically generated and all models perform smoothly both for real and synthetic datasets. We can find that logistic regression, random forest, and support vector classifier give the accuracy of more than 90% both for real and synthetic datasets. Here, the decision tree shows the accuracy of around 85% for the real dataset and 95% for the synthetic dataset.

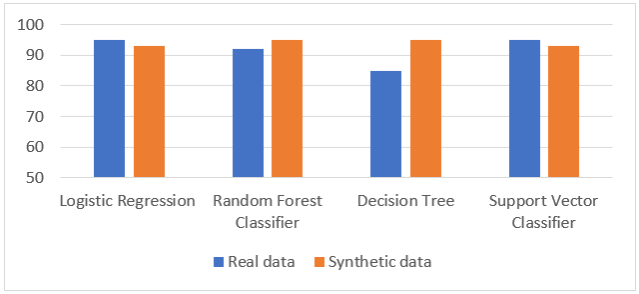

Figure 14: Comparing accuracy between real and synthetic dataset for cancer

6 DISCUSSION

As privacy protection methods progressively fail to secure current data, the necessity for synthetic data, particularly in the health care industry, is gaining traction. Real health care data is sometimes difficult or impossible to distribute due to legitimate privacy concerns, obstructing crucial machine learning research that may use this data to enhance patient outcomes and health policy decision-making. Synthetic data has the potential to alleviate data scarcity by serving as a viable substitute for actual data. In the literature, we have used Gretel for synthetic data generation and modified the tokenizer part of this Gretel. Here, we have used Gensim for token generation and we found a slightly more accurate value. From this experiment, it is shown that all machine learning models have almost the same accuracy when trained on synthetic data and tested on real data. In our observation, accuracy goes slightly lower for the second dataset (breast cancer-2) because of the maximum amount of synthetic data generation.

7 CONCLUSION

In this study, we have shown the effectiveness of synthetic data for training different machine learning models for use of the health care-related cancer dataset. The results are promising, the accuracy of synthetic data after testing with the real dataset is almost similar

to the real data testing with the real dataset. All machine learning models perform properly with two datasets; breast cancer-1 and cancer datasets, whereas a model shows a slightly low performance for the breast cancer-2 dataset. Logistic regression, random forest, and support vector classifier work perfectly, whereas decision tree shows a little bit lower-performing accuracy. In our experiment, we have generated 1,10,000 synthetic records from 1546 real records gathered from 3 datasets. In the future, we want to work on more datasets with new synthetic data generation techniques.

REFERENCES

[1] Rumbold, J.M., Pierscionek, B.K. (2018). Contextual Anonymization for Secondary Use of Big Data in Biomedical Research: Proposal for an Anonymization Matrix. JMIR Medical Informatics, 6.

[2] https://www.kaggle.com/datasets

[3] https://archive.ics.uci.edu/ml/datasets.php

[4] https://gretel.ai/

[5] Rubin D. Statistical disclosure limitation. J Off Stat 1993;9(2):461-468. https://doi.org/10.1002/9781118445112.stat00072

[6] Little R. Statistical analysis of masked data. J Off Stat 1993; 9:407-426.

[7] Raghunathan, T.E., Reiter, J.P., Rubin, D.B. (2003). Multiple Imputation for Statistical Disclosure Limitation. Journal of Official Statistics, 19, 1.

[8] Reiter, J.P. Simultaneous Use of Multiple Imputation for Missing Data and Disclosure Limitation.

[9] Reiter, J.P. (2005). Releasing multiply imputed, synthetic public use microdata: an illustration and empirical study. Journal of the Royal Statistical Society: Series A (Statistics in Society), 168.

[10] Reiter, J.P. (2005). Significance tests for multi-component estimands from multiply imputed, synthetic microdata. Journal of Statistical Planning and Inference, 131, 365-377.

[11] Reiter, J.P. (2009). Using Multiple Imputation to Integrate and Disseminate Confidential Microdata. International Statistical Review, 77.

[12] Reiter, J.P., Raghunathan, T.E. (2007). The Multiple Adaptations of Multiple Imputation. Journal of the American Statistical Association, 102, 1462 - 1471.

[13] Reiter, J.P., Drechsler, J. (2007). Releasing multiply-imputed synthetic data generated in two stages to protect confidentiality. Statistica Sinica, 20, 405-421.

[14] Reiter, J.P. (2005). Using CART to generate partially synthetic public use microdata. Journal of Official Statistics, 21, 441-462.

[15] Ping, H., Stoyanovich, J., Howe, B. (2017). DataSynthesizer: Privacy-Preserving Synthetic Datasets. Proceedings of the 29th International Conference on Scientific and Statistical Database Management.

[16] Nowok, B., Raab, G.M., Dibben, C. (2016). synthpop: Bespoke Creation of Synthetic Data in R. Journal of Statistical Software, 74, 1-26.

[17] Heyburn, R., Bond, R.R., Black, M.M., Mulvenna, M.D., Wallace, J.G., Rankin, D., Cleland, B. (2018). Machine learning using synthetic and real data: Similarity of evaluation metrics for different healthcare datasets and for different algorithms. Data Science and Knowledge Engineering for Sensing Decision Support.

[18] Rankin, D., Black, M.M., Bond, R.R., Wallace, J.G., Mulvenna, M.D., Epelde, G. (2020). Reliability of Supervised Machine Learning Using Synthetic Data in Health Care: Model to Preserve Privacy for Data Sharing. JMIR Medical Informatics, 8.

[19] Abowd, J.M., Stinson, M.H., Benedetto, G. (2006). Final Report to the Social Security Administration on the SIPP/SSA/IRS Public Use File Project..

[20] Benedetto, G., Stanley, J.C., Totty, E. (2018). The Creation and Use of the SIPP Synthetic Beta v7.0.

[21] Kinney, S.K., Reiter, J.P., Reznek, A.P., Miranda, J., Jarmin, R.S., Abowd, J.M. (2011). Towards Unrestricted Public Use Business Microdata: The Synthetic Longitudinal Business Database. U.S. Census Bureau Center for Economic Studies Research Paper Series.

[22] Boyle, P., Feijten, P., Feng, Z., Hattersley, L., Huang, Z., Nolan, J., Raab, G.M. (2009). Cohort Profile: the Scottish Longitudinal Study (SLS). International journal of epidemiology, 38 2, 385-92.

[23] O'Reilly, D., Rosato, M., Catney, G., Johnston, F., Brolly, M. (2012). Cohort description: the Northern Ireland Longitudinal Study (NILS). International journal of epidemiology, 41 3, 634-41.

[24] Miranda, J., Vilhuber, L. (2013). Looking back on three years of Synthetic LBD Beta.