# Final Project

maimuna.rahman

May 2023

## 1 Leveraging Synthetic Data for Diabetes Classification: Expanding Disease Prediction Horizons

Introduction:

Advancements in machine learning and artificial intelligence have revolutionized disease prediction and classification in healthcare. This study explores the potential of synthetic data to enhance disease classification, with a focus on diabetes. Traditional approaches to disease classification rely on limited real-world datasets, but synthetic data generation techniques offer opportunities to overcome these limitations and expand disease prediction capabilities.Using a dataset with key diabetes-related features such as Age, Glucose Level, Body Thickness, and Outcome, the project generates an extensive synthetic dataset using advanced data generation algorithms. The objective is to investigate the effectiveness of synthetic data in predicting diabetes, comparing its performance against models trained solely on real-world data. Synthetic data enables a more comprehensive representation of disease-related patterns by introducing diverse scenarios encompassing various patient profiles, demographic factors, and potential risk factors. This approach enhances the robustness and generalizability of predictive models, resulting in improved accuracy in diabetes classification.The study also aims to validate the hypothesis that synthetic data can serve as a viable alternative in healthcare for disease classification. Successful validation could revolutionize the field, providing healthcare practitioners and researchers with a powerful tool to generate large-scale, representative datasets. This would enable more accurate disease predictions and personalized treatments. The efficacy of synthetic data is evaluated by comparing the performance of traditional machine learning algorithms trained on real-world data alone with models incorporating both real and synthetic data. Performance metrics such as accuracy, precision, recall, and F1-score are used to assess classification results and identify potential improvements achieved through synthetic data integration.In conclusion, this project demonstrates the potential of synthetic data in disease classification, particularly for diabetes. By generating an extensive synthetic dataset and comparing its performance against traditional models, the study establishes the feasibility and effectiveness of synthetic data in enhancing disease

prediction accuracy. The findings have significant implications for the healthcare field, paving the way for more accurate and reliable disease classification models, ultimately leading to improved patient care and outcomes.

Dataset: The dataset used in this study originates from the National Institute of Diabetes and Digestive and Kidney Diseases. Its purpose is to make diagnostic predictions regarding the presence of diabetes in patients. This prediction is based on specific diagnostic measurements included in the dataset. Certain criteria were applied when selecting instances from a larger database. Notably, all patients included in this dataset are females, aged at least 21 years, and of Pima Indian heritage.The provided (.csv) file contains multiple variables, some of which are independent variables used as medical predictors, while there is only one target variable that is dependent, referred to as "Outcome."

Link

Literature Review:

Synthetic data generation has great importance in the data science and analytics industry. Before there were challenges faced by data scientists and analysts when working with real-world data, including data privacy, data quality, and data scarcity. To overcome these challenges, the use of synthetic data generation can help in testing proof of concept models without the actual data. There are different approaches used for synthetic data generation, including parametric, non-parametric, and generative adversarial networks (GANs). It also outlines the advantages and disadvantages of each method, which can help data scientists and analysts choose the best approach for their specific use case. One of the strengths is that it provides a step-by-step guide on how to generate synthetic data using Python and the atoti platform. The guide is easy to follow and includes examples of code snippets, making it accessible to both novice and experienced data scientists. Synthetic data generation has significance in the data science and analytics industry. The atoti platform is an excellent tool for generating synthetic data, and the article is a useful resource for data scientists and analysts looking to explore this technology. The article "Machine Learning for Synthetic Data Generation: A Review" published on ResearchGate is a comprehensive and insightful analysis of the role of machine learning in synthetic data generation. The authors provide a detailed review of the state of-the-art methods used in machine learning for generating synthetic data and their application in various fields, including healthcare, finance, and social sciences. Synthetic data generation can help with the challenges of working with real-world data, such as privacy concerns, data bias, and data scarcity. There are different machine learning techniques used for synthetic data generation, including deep learning, GANs, and variational autoencoders. We can use machine learning techniques in synthetic data generation. We can use different machine learning algorithms, including decision trees, support vector machines, and neural networks, for generating synthetic data. The techniques used for data augmentation, data imputation, and data synthesis are commonly used in machine learning-based approaches for synthetic data generation. But there are also limitations of machine learning-based approaches and the challenges associated with evaluating the quality of the generated synthetic data. Synthetic

data can be used as a tool for addressing the challenges associated with using 1 big data in official statistics. We can use synthetic data as a way to address issues related to data privacy and confidentiality, while still allowing for the use of big data in statistical analysis. There are challenges associated with using big data in official statistics, including the need to protect individual privacy and maintain confidentiality. There are potential benefits of using synthetic data as a tool for addressing these challenges, including its ability to protect privacy while still allowing for statistical analysis. There are different methods for generating synthetic data, including the use of statistical models and machine learning algorithms. Synthetic data is used to protect the privacy of individuals in census data and the use of synthetic data to train machine learning algorithms for predictive modeling. There are limitations and challenges associated with the use of synthetic data, including the need to ensure that the synthetic data accurately represents the underlying population and the need for ongoing evaluation and validation of synthetic data methods. Synthetic data is used as a tool for addressing the challenges associated with using big data in official statistics. It is a reliable and effective tool for protecting privacy while still allowing for the use of big data in statistical analysis. Synthetic data is widely used in the UK to create artificial data that can be used in place of real data. Synthetic data is created by using statistical models and algorithms to generate data that has similar statistical properties to the real data but does not contain any personally identifiable information. It is used in research and development as it enables researchers to analyze data without the need to access sensitive or confidential data. This is particularly relevant in fields such as healthcare and social science where access to real data is limited due to privacy concerns. It has different approaches to generating synthetic data, such as generative models and differential privacy, and the challenges associated with each approach. It also provides a review of the current use of synthetic data in different sectors, such as healthcare and finance, and the benefits and limitations of using synthetic data in these sectors. In future the use of synthetic data will be increasing including the need for more collaboration and standardization across different sectors, as well as the need for more research into the validity and reliability of synthetic data. The current use of synthetic data in the UK has importance in research and development, discussing different approaches to generating synthetic data. Synthetic data generation does not contain any identifiable information about individuals. This allows researchers to analyze the data without compromising privacy. There are some challenges associated with real data, such as privacy concerns and the difficulty of obtaining large and diverse datasets. These challenges can limit the potential of research in various fields such as healthcare, social science, and artificial intelligence. Synthetic data has the potential to overcome these challenges by enabling researchers to generate large and diverse datasets that are free from privacy concerns. There is a potential of synthetic data to facilitate collaboration between different organizations and researchers. Synthetic data has been used in research, including studies on healthcare and genetics. Synthetic data can be used in other fields such as autonomous vehicles and cybersecurity. But there are some limitations of synthetic data, such as the

need to validate the accuracy of synthetic data and the potential for bias in the 2 generation of synthetic data

Methodology:

Data Preparation:

In the data preprocessing phase, we start by loading the original diabetes dataset, which contains features such as Age, Glucose, Blood Pressure, Skin Thickness, Insulin, BMI, Diabetes Pedigree Function, and Outcome. Additionally, we use the Faker library to generate synthetic data based on the real dataset.

Using the Faker library, we set generation probabilities for each feature to control the ratio of real and fake data in the synthetic dataset. We iterate over the original dataset and generate fake data for each row, varying the generation probability for each feature. For example, we may generate 90

Data Preprocessing:

With the synthetic dataset prepared, we proceed with data splitting, feature engineering, feature selection, and handling class imbalance. Firstly, we split the data into features (X) and the target variable (y) for both the training and testing datasets.

Next, we perform feature engineering by applying feature scaling using the StandardScaler on the training dataset. This step normalizes the features and ensures they have similar scales, which can improve the performance of machine learning models. We then use the same scaler to transform the testing dataset, maintaining consistency between the two datasets.

Feature selection is carried out using SelectKBest, which employs the $f_classif scoring function to rank the fea$

To handle class imbalance in the training dataset, we employ the Synthetic Minority Over-sampling Technique (SMOTE). This technique oversamples the minority class by generating synthetic samples that are similar to the existing minority class samples. By balancing the class distribution, we prevent bias towards the majority class during model training and improve the model's ability to classify both classes accurately.

Model Training, Selection, and Evaluation:

In this stage, we proceed with model training, selection, evaluation, and computing the confusion matrix. We initialize and train two machine learning models: Random Forest and XGBoost. To find the best combination of hyperparameters for each model, we employ GridSearchCV, which performs an exhaustive search over specified hyperparameter values using cross-validation.

The cross-validation accuracy of the best models obtained from the grid search is calculated. This metric provides an estimate of how well the models generalize to unseen data. We then make predictions on the testing dataset using the trained models and evaluate their performance. Metrics such as accuracy, precision, recall, and F1-score are calculated to assess the models' classification performance. Additionally, we compute the confusion matrix for both the Random Forest and XGBoost models. The confusion matrix provides a detailed breakdown of the model's performance, showing the number of true positives, true negatives, false positives, and false negatives. This information helps evaluate the models' ability to correctly classify instances of diabetes.

By performing these steps, we obtain insights into the models' performance, the effectiveness of synthetic data in diabetes classification, and the impact of feature engineering and selection, as well as class imbalance handling. These evaluations contribute to a comprehensive understanding of the models' capabilities and provide guidance for future improvements in disease classification.

Result: The results obtained from the evaluation of the Random Forest and XGBoost models on the diabetes dataset are presented in this section.

For the Random Forest model, the cross-validation accuracy was found to be 0.7502, indicating a reasonably good performance. The test accuracy achieved on the dataset was 0.7422, which further confirms the effectiveness of the model. The precision and recall scores for class 0 (non-diabetic) were 0.79 and 0.82, respectively, while for class 1 (diabetic), the precision and recall scores were 0.64 and 0.60. The f1-score, which considers both precision and recall, was 0.62 for class 1. Overall, the model demonstrated a weighted average f1-score of 0.74, indicating a satisfactory performance in diabetes classification.

Table 1 :Findings from Random Forest

| Matrix | Value |
|---|---|
| Cross-Validation Accuracy | .752 |
| Test Accuracy | .742 |
| Recall Class(0) | .82 |
| Recall Class (1) | .6 |
| F1-Score(Class 0) | .81 |
| F1 Score(Class 1) | .62 |
| Weighted Average F1 Score | .74 |
| MacroAverage F1 Score | .71 |
| Precision | .64 |
| Weighted Average Accuracy | .74 |

Similarly, for the XGBoost model, the cross-validation accuracy was observed to be 0.717, while the test accuracy was 0.7409. The precision, recall, and f1-score for class 0 were 0.78, 0.83, and 0.81, respectively. For class 1, the precision, recall, and f1-score were 0.65, 0.57, and 0.61. The weighted average f1-score for the XGBoost model was also 0.74, indicating a comparable performance to the Random Forest model.

Table 2 :Findings from XGBoost

| Matrix | Value |
|---|---|
| Cross-Validation Accuracy | .717 |
| Test Accuracy | .741 |

| | |
|---|---|
| Recall Class(0) | .83 |
| Recall Class (1) | .57 |
| F1-Score(Class 0) | .81 |
| F1 Score(Class 1) | .81 |
| Weighted Average F1 Score | .74 |
| MacroAverage F1 Score | .71 |
| Precision | .78 |
| Weighted Average Accuracy | .74 |

Overall, both models achieved reasonably accurate predictions in classifying diabetes. The Random Forest model exhibited slightly higher performance metrics in terms of precision, recall, and f1-score for both classes, while the XGBoost model showed comparable results. These findings suggest that both models have the potential to be effective tools for diabetes classification. However, further analysis and experimentation may be required to identify the optimal model for this specific dataset and to generalize the findings to larger and more diverse populations.

Future Work: In future work, several avenues can be explored to enhance the findings and extend the scope of this project. Firstly, the synthetic data generation process can be further optimized by incorporating more sophisticated techniques, such as generative adversarial networks (GANs), which have shown promising results in generating realistic synthetic data. GANs can potentially capture the underlying data distribution more accurately and produce synthetic samples that closely resemble real-world instances.

The current project focused on diabetes classification, but similar methodologies can be applied to other disease domains as well. Exploring the effectiveness of synthetic data in classifying different diseases would provide valuable insights into the generalizability of the approach and its impact on various medical conditions.Additionally, investigating the transferability of models trained on synthetic data to real-world datasets is an important area for future exploration. It would involve evaluating the performance of models trained on synthetic data when applied to real patient data, assessing the robustness and reliability of the models in practical healthcare settings.Furthermore, conducting a comparative analysis of different synthetic data generation techniques would be beneficial. This would involve evaluating the performance of models trained on synthetic data generated using different algorithms and assessing their effectiveness in disease classification. Understanding the strengths and limitations of various synthetic data generation methods can guide the selection of the most suitable approach for specific healthcare applications.

Overall, future work should focus on refining and expanding the current findings, exploring novel techniques, and validating the applicability of synthetic

data in disease classification across diverse medical domains.

Conclusion: In conclusion, this study demonstrates the potential of synthetic data in enhancing disease classification, specifically focusing on diabetes. By generating a synthetic dataset and comparing it with models trained on real data, we have shown that synthetic data can effectively predict diabetes with comparable performance. The results highlight the value of synthetic data in overcoming limitations of small-scale real-world datasets, enabling a more comprehensive representation of disease patterns. By incorporating synthetic data, predictive models benefit from increased diversity and improved accuracy in diabetes classification. These findings have implications for the healthcare field, providing valuable insights into the use of synthetic data to support accurate disease predictions and personalized treatments. Further research can explore the application of synthetic data in other disease classifications and evaluate its effectiveness across different healthcare domains.

References:

Link 1
Link 2
Link 3
Link 4