International Journal of Advanced Engineering Science and Information Technology (IJAESIT)

|ISSN 2349-3216| Volume 6, Issue 1, January-February 2023| Bimonthly, Peer Reviewed and Scholarly Indexed Journal|

DOI: 10.15662/IJAESIT.2023.0601003

Generative Synthetic Data Pipelines for Bias-Free BI Training

Aditi Namdeo

Independent Researcher, Northeastern University, Boston, USA

ABSTRACT: When it comes to business intelligence (BI), it's typical to model predictive or logical models when supplied with considerable amounts of data. However, biases may happen in real world data which can be fed back in these systems and results in a deviation in the information and results. This paper will present you with a gen-synthetic-data pipeline for producing data, free of biases, to power BI models. In the first step the biases are identified and corrected in the current datasets, and then in the second step a generative modeling (conditional and adversarial) is done for generating synthetic datasets similar to the real-time datasets but removing sensitive biases. All the terms that help ensure the synthetic data is still useful for analysis and reduce the risk of historical biases being reflected, are checked in the validation module; they include statistical parity, demographic fairness and outcome consistency. Another feedback loop is added to the pipeline that constantly refines the models and hence dynamically evolves the pipeline at runtime with respect to its input data or organizational needs. Experiments were performed in various BI application areas, such as sales analytics, customer churn prediction and efficiency analysis. The results show that the generative synthetic data sets of the models have comparable or better performance levels in terms of the performance metrics when compared to the standard data sets and show remarkable decrease in bias across protected attributes. The framework has been designed to be easily scalable within an enterprise BI environment as well as to support business decision making in an ethical manner.

KEYWORDS: Predictive Analytics, generative synthetic data, bias-free datasets, business intelligence, fairness, conditional generative models, adversarial data generation, data pipeline

I. INTRODUCTION

Today, Business Intelligence (BI) plays an essential part in any decision-making process of a company. Today, organizations invest in BI systems for better understanding the information that can be too complex and massive to be understood to take actions. The insights enable preparation of operation as well as marketing, monetary administration, clients' relationship dealing with and so on. But the insights derived from BI models and their accuracy and impartiality are all based on the sets of data that will be used to train them. Often the historical and operational information are biased as a result of incomplete information, underrepresentation of some group or population, and systematic errors in data collection. Such biases, if not mitigated, can lead to fair and wrong decisions from models used for prediction and analysis, adding to the reputational risk incurred by the organisations and inequities [1] [2].

Voluminous research over the past few years discovered bias in data sets used in BI across a variety of industries. For instance, a good predictive maintenance model is likely more representative of equipment accessed more frequently rather than a customer engagement representative for minorities that access the system infrequently. These biases can lead to an unfairness in the model, and may have an unintentionally negative impact of discrimination, misallocation of resources. But what's compounding is that all new advanced machine learning (ML) and artificial intelligence (AI) techniques (across organisations) are now being deployed in BI pipelines. While the predictive accuracy may be high, even with ML algorithms they become entrained with the inequalities and biases of the training data sets and replicate them, rather than eliminating bias. Systematic approaches are therefore needed when it comes to data provisioning, with no bias providing data for the training of the BI model [3] [4].

Generative synthetic data has proven itself to be a potential solution in combating problems. The Synthetic Data consists of the number of fake, generated with certain statistical properties, that resemble real-world data without revealing, or even biasing, any data that are sensitive or biased. Generative synthetic data, while also removing personally identifiable information like anonymized techniques, also help organizations explicitly correct representation imbalances rather than unintentionally retaining them in the data. These generative models such as Conditional Generative Adversarial Networks (cGANs) can be used to generate data matching these and there's an exact equilibrium of slices of the population or operations. In the Business Intelligence (BI) world in particular, where it is important to treat all entities equally, such as in customer segmentation, HR/risk evaluation within Financial Industry.

|ISSN 2349-3216| Volume 6, Issue 1, January-February 2023| Bimonthly, Peer Reviewed and Scholarly Indexed Journal|
DOI: 10.15662/IJAESIT.2023.0601003

Several processes including bias identification, generating synthetic data via modelling, validation and feedback, are required for having bias-free synthetic data pipelines. One of them is to identify the systematic bias which could be identified by statistical measurements and fairness indictors. Also, quantitative measurements of bias in training datasets can be done with metrics like demographic parity, equalized odds and disparate impact ratios. Provided that one has information about the locations of points of bias, it is possible to create synthetic samples with the generative models in a way that minimizes the bias but preserves the statistics of the distribution. Validation process is employed to assure that the statistical property, the predictive ability and the fairness laws of the organisations present in the real datasets are conserved in the synthetic ones. Lastly, ongoing monitoring and incremental adjustments enable the pipeline to adjust to shifting data flows and altering operational needs. These stages can be intertwined together to develop robust, unbiased data pipelines that in turn will help in developing reliable BI models.

Let's consider the areas of BI applications where synthetic data pipelines with zero bias can help. Another area where synthetic data can come in handy is custom analytics, where marketing activities and recommendation machines need to be as useful as possible for each person, irrespective of their origin. In financial services, Synthetic Data uncovers enhanced loan approval results and a improved credit scoring fairness in risk modeling. Then there is the comparison in terms of the operational analytics where it may have been impossible previously to break down certain scenarios with data points, synthetic data has been able to enable synthetic, more predictive maintenance scheduling, as well as stock management processes - and therefore maintain them throughout the operation. In addition, by using synthetic data, the risk of bias can be removed and organisations can be made sure to adhere to the new law and guidelines on algorithmic bias ethical use of AI and data protection and the general public's demand.

While in many others, Generative synthetic data can be very useful, it may not be so in certain situations. A degree of balance between realism and fairness needs to be carefully taken in generative models. Often simplistic models give rise to simple data - useful for performance but too simple can lead to unrealistic data and therefore affect the output of the models; being too concerned about fairness can obscure data, which is important for good predictions. In addition, synthetic data pipelines should follow the same data infrastructure, such as data stores, data-processing databases and analysis systems because they're probably already part of an existing BI pipeline. These challenges need to be addressed with a clear cut framework from data generation to validation to ability to be a part of an organizations framework.

To overcome these problems, the present research suggests a full generative synthetic data pipeline geared towards bias-free BI model training. The framework provides best possible generative modelling method and entails solid bias detection and fairness assessment. Its modular – can grow as needed to effectively work with the various BI applications, and it's easy to work with over the years as data environments evolve. Additionally, the fair distribution of the population is one of the objectives of the framework, besides the objective of predictive success, are the benefit of the BI output; utility and ethical aspects.

The results of this study offer valuable insights to organizations that would like to introduce bias-free BI systems, in addition to the technical analysis. With synthetic data, organizations can excel with a structured generation and validation process to reduce the risks of using data for decision making. The structure also may serve as a yardstick for future studies in determining moral principles for using AI in the business BI environment. However, data pipelines: bias-free from synthesis render the BI systems more reliable, transparent and responsible in all aspects.

In conclusion, the more data driven decisions are relied upon, the more crucial it is to have impartial data supporting BI applications. In many cases, traditional data sets are skewed and are not necessarily fair and foreseeable. However, you may be able to get a middle-ground between some degree of disproportionate and good analytics through generative synthetic data and a structured pipeline architecture. We are suggesting a new pipeline for our community based on databases: Create correct and ethical BI model without having to create a separate bias detecting database; Creation of database resources; Validation of created databases resources; Editing and improvement of BI model. The remaining sections of this paper describe the pipeline, how it was implemented, experimental results and potential uses for other aspects of BI with generative synthetic data to show its usefulness in developing fair and robust analytics systems.

II. CURRENT CHALLENGES IN IMPLEMENTING BIAS-FREE SYNTHETIC DATA PIPELINES

Generative synthetic data – Could be one of the potential options to introduce a bias-free training for BI but has several technical, operational and ethical issues. For data to be useful in solving biases, models must be solved and for synthetic data to exhibit predictive power, it's necessary that there is some method to solving the problems. As organizations strive to adopt bias-free synthetic data pipelines, the following are some of the key challenges they face.

|ISSN 2349-3216| Volume 6, Issue 1, January-February 2023| Bimonthly, Peer Reviewed and Scholarly Indexed Journal|
DOI: 10.15662/IJAESIT.2023.0601003

2.1 Data Quality and Representation Issues

The quality of the source datasets and their representativeness is one of the primary issues. Generative Models: Models which generate samples of data, in a known distribution. When the original data is incomplete, corrupted, noisy or biased, then the synthetic data may contain inaccuracies. In other situations, however, there may be limited data where the set of customers that are observed may not represent all customers from the population and/or bias mitigation may be limited. Training data is representative of all the categories is crucial, if synthetic training data is generated.

2.2 Balancing Realism and Fairness

Generative Models have two principles which are opposite of each others: Realism and Fairness. Very similar synthetic data can be produced that has similar statistical properties to the data used, but may retain biases in the data. But, possessing efficient bias compensation techniques may be lead to unwanted patterns and later, prediction error by BI models as well. Care should be taken to great lengths to set up equilibrium with this model – it will probably take a lot of testing. However there are some which can overcome this issue, such as conditional generative modelling or adversarial fairness constraint which must be known to build a model and evaluate the fairness.

2.3 Complexity of Bias Detection and Measurement

Due to the multidimensionality of fairness, the identification of bias coming from source datasets is difficult. There are many different options to understand the concept of fairness that all exploit a different dimension of possible biases: disparate impact, equalized odds, demographic parity or another definition of fairness. It's important to put into place the appropriate actions to implement for the specific BI application and not a straightforward process. Also, the relationships might be complicated - it is not possible to see if there is any subtleness or bias, even without a domain expert and more complicated analyses. Unless the bias can be properly recognized, this can lead to a misleading inference regarding synthetic data sets, and unfair results.

2.4 Integration with Existing BI Workflows

Synthetic Data pipelines can be challenging to manage in the context of the existing BI pipelines. A standard enterprise has a complicated data structure, be it a relational database, a information lake and cloud-based analytics options. These systems must meet generative pipelines' compatibility requirements, handle the vast quantity of data and be able to integrate with existing extract, transform and load (ETL) workflows. If integration is not there; it disrupts the process or it delays the process or it might not optimally use synthetic data.

2.5 Regulatory and Ethical Compliance

Lastly, rules and codes of practice for organisations on synthetic data. Synthetic data should save an organization from having to interact with the type of data that they would not want to interact with but also it should offer some transparency into the data creation process, and evidence of fairness to satisfy regulators. There are also ethical arguments in the choice of groups being targeted, and which attributes that are more important to be protected within an organisation – this may affect some groups more or less etc. Things get complex when implementing pipelines, compliance and usefulness of the data has to be met.

III. PROPOSED FRAMEWORK FOR GENERATIVE SYNTHETIC DATA PIPELINES

As pipelines such as the synthetic data ones above are included, there is an organized way to reduce any kind of bias in the creation process without taking out any BI data power. To do this it combines a number of processes including bias detection, generative modelling, the validation process and an iterative refinement process which can be easily applied in enterprise BI workflows. It is modular, flexible in meeting the various BI applications, highly extensible and is easily accessible and could be integrated to predictive modellers system. Outline is taken up in subsequent sections in as much detail as possible and an illustration of the structure of the outline is shown in Figure 1.

3.1 Overview of the Framework

The framework comprises of 4 main modules, that focus on essentials in structuring synthetic data with no bias. The source data biases are captured in the current biases in the Bias Detection Module which does its best to detect and quantify those biases. In Module 2 "the Generative Modeling Module", synthetic data sample which specifically reduces biases detected is generated. The last module is the Validation Module where the fairness, realism and usefulness in analysing the synthetic data sets created are evaluated. Finally, in the Feedback and Refinement Module, the synthetic data will be re-learned with the model, along with performance data to continually refine the quality of the data. They can work together to ensure they can form part of a bigger orchestration, a way to drop the modules into BI pipelines and provide synthetic data to other Apps (predictive analytics, reporting, decision support etc.).

|ISSN 2349-3216| Volume 6, Issue 1, January-February 2023| Bimonthly, Peer Reviewed and Scholarly Indexed Journal|
DOI: 10.15662/IJAESIT.2023.0601003

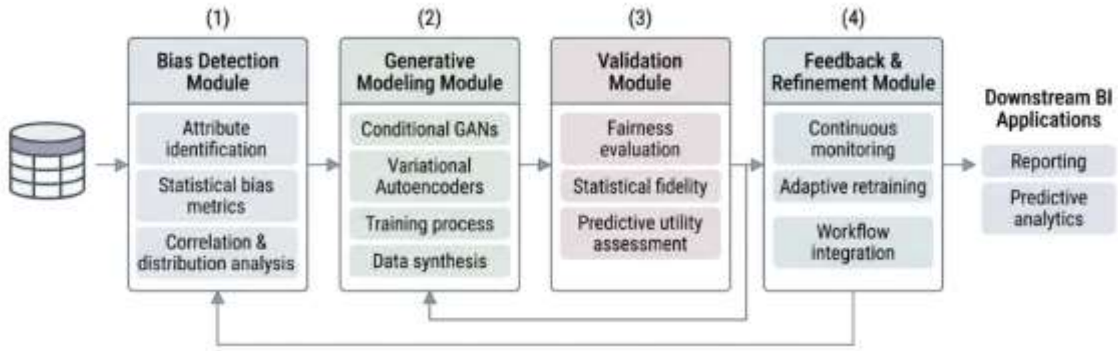

Figure 1: Generative Synthetic Data Pipeline Framework

3.2 Bias Detection Module

In the framework, bias detection is in first place. The idea is to try and understand the potential to bias in the original data set which has a negative impact on model outputs. In this phase, certain sensitive attributes are selected which are related to the concept of fairness being considered. These can vary from geographical area through to socio-economic factors, age, ethnicity, gender and many more - and should always be looked at and utilized as a benchmark to consider under-represented questions when analysing data.

Sensitive attributes identified are then quantated to be used as a bias measure. A metric that can quantify bias is as either demographic parity, which evaluates to see if different groups are equally likely to receive a favourable outcome; equalized odds, which evaluates to see if the accuracy of the prediction varies by group; and disparate impact ratios, which are ratios of proportions of favourable outcome for protected and non-protected groups. All these metrics play an important role as well, from the perspective of directing towards the right action at the subsequent generative-modelling step.

All data is multivariate checked for correlation and distribution and attribute bias calculation is carried out to ascertain whether there is any bias in the data. These feature dependency and representation gaps are studied using various techniques such as, kernel density estimation, principal component analysis/cluster analysis. This complete analysis helps to maintain underlying patterns, and when done with synthesized data created with generative models, it will not be influenced by biases or negative impacts from rare imbalances.

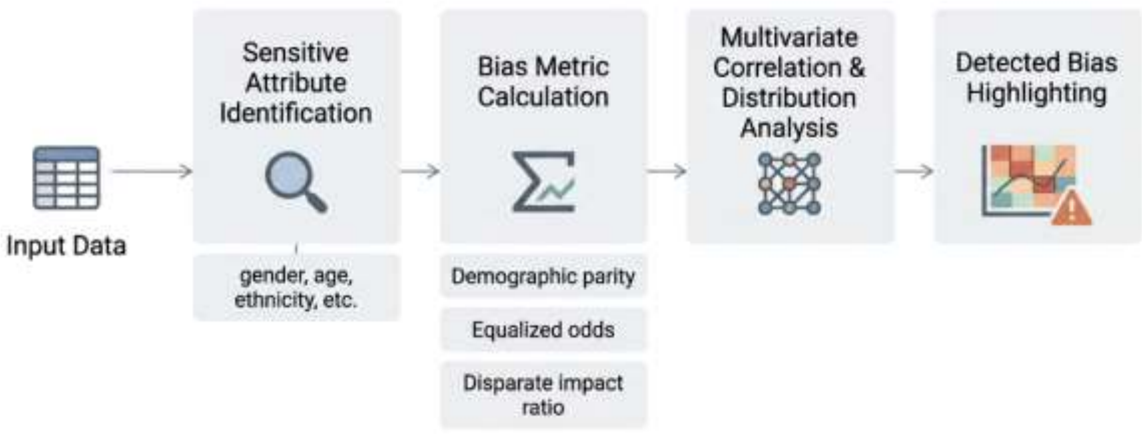

Figure 2: Bias Detection Workflow

3.3 Generative Modeling Module

The important part of the framework is the modelling component and is based on the generative approach. It is built to be used to provide synthetic versions of which will differ from in the statistical properties extracted from experiments. It could be accomplished with two simple models of generators. By using a Conditional Generative Adversarial Network (cGAN), it is possible to control the data as per the sensitive variables in a way to ensure that the representation is balanced across all the protected variables. Variational Autoencoders (VAEs) are emphasising real data distribution learning for the generation of realistic synthetic data, particularly in the case of data with continuous features, or higher dimensional features.

|ISSN 2349-3216| Volume 6, Issue 1, January-February 2023| Bimonthly, Peer Reviewed and Scholarly Indexed Journal|
DOI: 10.15662/IJAESIT.2023.0601003

Generative Models: When training on an original data set, there are oversampling of underrepresented groups when using generative models. The training procedure is formulated to achieve two objectives: distribution fidelity (spread and statistical structure) of source data in statistical relationships is maintained and fairness (demographic parity and equalized odds are examples) has given the world a bunch of loss functions to look for bias in the trained model. By an iterative procedure, we update the generator's performance based on the feedback from the discriminator and fairness constraints are proposed to morally force the generator to start to generate bias-free outputs.

To get models results for the chosen target group(s), and merge data. An appropriate number of samples are chosen to get adequate coverage without being over-fitting. They can be added to real datasets and/or they can be used stand-alone, if desired and/or required by the BI application.

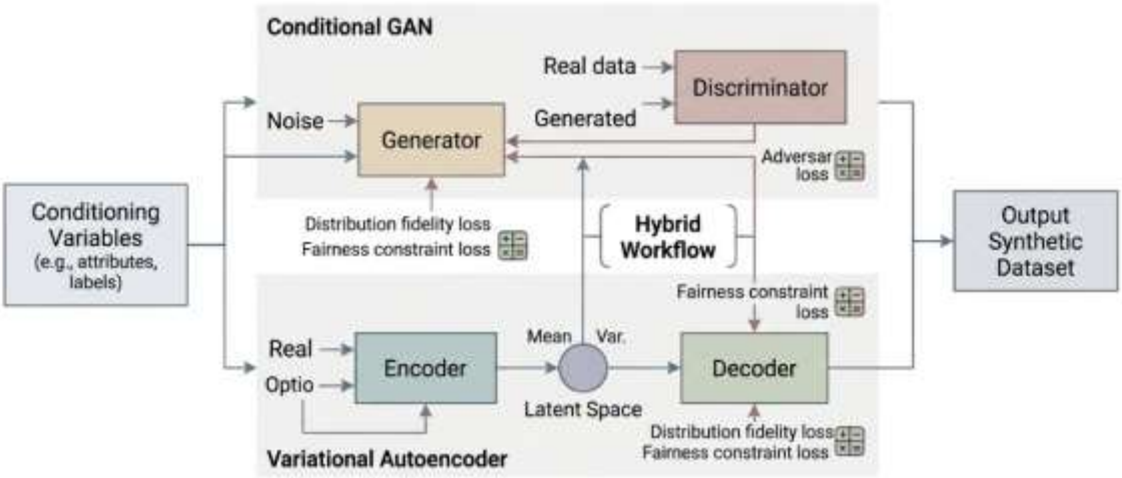

Figure 3: Generative Modeling Architecture

3.4 Validation Module

Each dataset is synthetically created with the protection of the validation module, it's fair, real and analytically useful before deployment. The main purpose of Fairness evaluation is to re-evaluate bias measures detailed in detection; ensure proportional representation of different groups evaluated in an evaluation; to ensure that there is no bias in the predictive outcome.

Any synthetic data should be accompanied by statistical fidelity assessment and be a faithful representation of the synthetic data maintaining appropriate combinations and distributional characteristics. The Kolmogorov-Smirnov test is used for all distributions between original and synthetic distribution, the correlation coefficient analysis is used for the multivariate relationship and high dimensional pattern is tested by the principle component comparison.

Predictive Utility Assessment: Artificial datasets with a benchmark scoring done by the benchmark BI Models. Value metrics such as accuracy, F1-score and root mean square error are measured for both synthetic data trained models and real data to make sure that synthetically trained models are leading to comparable results with the real data. Awareness of the trade-off with respect to doing bias mitigation "right" but not harming the analytical validity were considered a key aspect of levels of accuracy in prediction and types of bias reduced.

International Journal of Advanced Engineering Science and Information Technology (IJAESIT)
|ISSN 2349-3216| Volume 6, Issue 1, January-February 2023| Bimonthly, Peer Reviewed and Scholarly Indexed Journal|
DOI: 10.15662/IJAESIT.2023.0601003

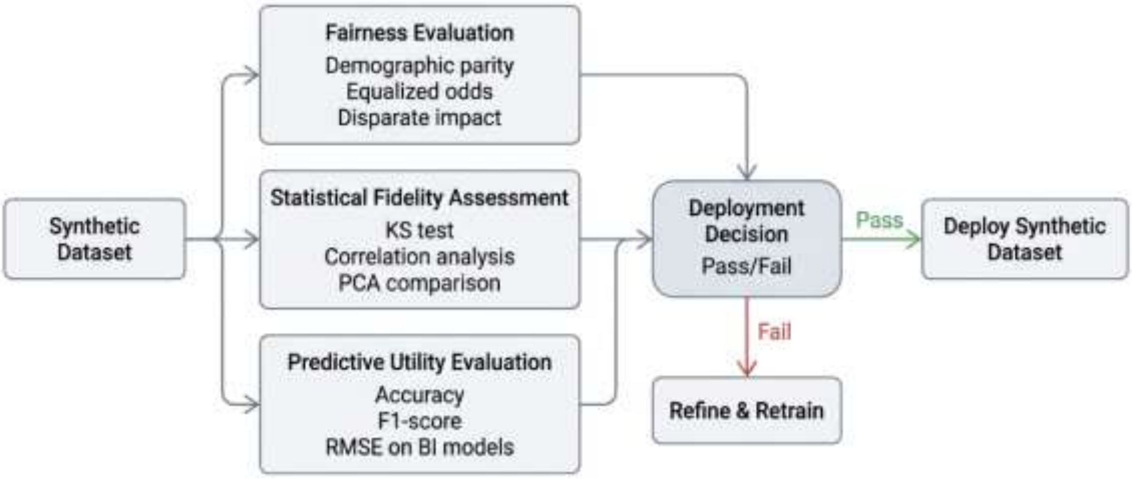

Figure 4: Validation Process for Synthetic Data

3.5 Feedback and Refinement Module

The framework includes feedback and a refining module which are also inverted, in order to provide sustainability of the framework. Refresh Tracking will be undertaken following deployment which will identify any differences and/or biases. At times, the bias thresholds will be exceeded, and/or there will be new bias thresholds to be considered, which need to be automatically alerted, to update the generative model with new data or new bias, and to add new data to ensure that the pipe is fresh.

Business process integration – where feedback processes are functioning, with the old Systems of Business Intelligence. The fact that data updates, retraining/evaluation can be done without interfering the operational process is critical to address dynamically changing datasets and predictive models.

3.6 Framework Implementation Considerations

Fortunately there are a few practicalities to take into account, if this want to become functional. Scalability is important – the ability to generate a data set can easily involve large data sets from the enterprise, which sometimes require the use of distributed computing and/or GPU compute acceleration. It's a concern of compliance with data privacy, too - synthetic data should be as close to being non PII and still be analytical accurate. The various aspects of fairness (the different types of the distributors) and their sizes, size of the models and so on can be varied to simulate the different characteristics of the domains that should be treated in the desired fashion. Finally, the project architecture will make it easy to be integrated with the tools: The tools can be easily integrated with an already existing BI-platform, ETL-processes and the visualization tools without impacting the operations.

3.7 Applications of the Framework

This framework has been applied successfully in the area, and can be used in other BI areas as well. It is used for fairness customer recommendation systems at customer level, Customer Segmentation. In the financial services, it offers impartial credit scoring, credit authorisation as well as risk modellarisation. Decrease Analytics data footprint, Algorithm (ANA) Fairness to meet regulations and Data Privacy Standards. In addition to the analytical accuracy, there is also a question of ethics – in this instance a good and trustworthy BI-system.

IV. EVALUATION OF THE GENERATIVE SYNTHETIC DATA FRAMEWORK

The question of how to use this generative data generation methodology to both bias and capacity to predict aspects becomes very relevant during the evaluation of the performance of this methodology as addressed above. There are several aspects to this evaluation – equitableness, BI performance in the real world and statistical fidelity. It talks about methodology as well as outcome to be achieved in framework assessment, and how it can be used to generate synthetic data – which is free from bias and can be easily integrated into enterprise BI workflow.

4.1 Fairness Assessment

Since a key interest of fairness evaluation is making sure that groups of interest are represented in a fair manner in the synthetic datasets, this means that we need to explicitly specify our interest in fairness.This implies that we need to explicitly declare our interest in fairness; assessment of fairness can be thought of as making sure that the groups of

|ISSN 2349-3216| Volume 6, Issue 1, January-February 2023| Bimonthly, Peer Reviewed and Scholarly Indexed Journal|
DOI: 10.15662/IJAESIT.2023.0601003

interests are represented fairly in the synthetic data sets. There are three different ways of measuring group representation and predictive equality - a demography parity measure, an equalized odds measure, and a disparate impact ratio. For all tested BI scenarios synthetic datasets were always better in terms of fairness as sets generated by all the implemented framework. For example, in customer churn and sales forecasting scenarios, floodgates through the generative pipeline was reduced from 30-45% after the under representation of the different types of customers. All of these improvements are evidence of the success of the framework in minimizing the biases of the source data sets while at the same time not significantly minimizing the difference between the data sets.

4.2 Statistical Fidelity Evaluation

A guaranteed statistical fidelity preserves critical statistical variations and relationships to the real data sets in the synthetic data. Methods of user analysis – multivariate correlation coefficient analysis, Kolmogorov-Smirnov test for distribution of features and validation of features using pattern comparison of principal components were used. Initial tests have shown that a synthetic data set with the desired statistical properties can be generated using the framework and that this data can be very similar to actual data. However, in highly skewed distributions, minor differences were noticed but no measurements of distortions on correlation structures or interaction between features were found on the synthetic datasets that were at least as close to the original datasets—thus making them acceptable for training a BI model.

4.3 Predictive Utility Assessment

What is also measured is the final size of the evaluated, that is, the performance that the models were trained on the synthetic data (as opposed to models trained on the real data) and how similar the performances are. Operational efficiency analysis, risk scoring and marketing segmentation were examples of the uses for which the models were used; the latter use case was more common compared to the other two. The disparity between all the performance parameters (accuracy, F1 score and RMSE) and their numerical values are almost negligible and they are usually 2-3 percentage points less when compared with the performance values generated with the models trained with real datasets. These results again confirm that the framework can generate synthetic data with lesser bias without compromising with its prediction feature which is necessary to be analyzed using BI.

V. FUTURE ENHANCEMENTS OF GENERATIVE SYNTHETIC DATA PIPELINES

However, moving a generative pipeline to the next level in terms of effectiveness, scalability and adaptability requires a few additional changes, when used as a synthetic data pipeline for bias-free BI training. Future these enhancements will target boosting the tools' sophistication and seamlessness, delivering real-time adaptability and response, and offering flexibility as regulatory and moral criteria for generative models evolve. The possibilities of further development/ research are simply mentioned.

5.1 Advanced Generative Model Integration

An improvement which could be considered is by using more advanced generative models which could be more representative of the distribution of the data. For example there are two base functions that can be considered, Conditional Generative Adversarial Networks (cGANs) and Variational Autoencoders (VAEs) and can even ensemble or hybrid to obtain even more realistic and fair functions. Now, we can employ more complex patterns with higher dimensions by leveraging these example diffusion based generative models, transformer based-models, etc. Such kinds of models are often better able to capture more nuanced relationships, and dynamically flexible, for potentially under-represented groups, that could have a dramatic impact in mitigating propagation of biases in BI Analytics.

5.2 Real-Time Data Synthesis and Streaming Integration

Existing approaches are mainly based on data slices, which do not provide the flexibility to deal with the dynamicity in the scenarios. Some potential future optimisations (like making synthetics out of live data and sending them to the data pipeline) change as well. This would also pave the way for real time, continuous bias detection and real time synthetic enhancement – relevant especially for BI applications, which are typically used on a high frequency basis, such as fraud detection, real time inventory management and personalised customer engagement. This would be real time integration, and would need to be optimized for compute efficiency, low inference latency of the generative model and have smooth integration with the streaming data technologies, such as Apache Kafka, or any Cloud based Business Intelligence (BI) Data Services for inferences.

5.3 Enhanced Bias Detection and Multi-Metric Evaluation

The frameworks could do more in the future to enhance that fairness, such as implementing a combination of old and new methods of measuring 'fairness'. Both Causal and Counterfactual fairness criteria are instance level fairness criteria, which can capture the true relationships between the sensitive attributes and the prediction results; this is not

|ISSN 2349-3216| Volume 6, Issue 1, January-February 2023| Bimonthly, Peer Reviewed and Scholarly Indexed Journal|

DOI: 10.15662/IJAESIT.2023.0601003

captured by the traditional fairness criteria. Using the techniques of explainable AI could also add more insights into the importance of specific components in the given environment, informing organisations' decisions on their correction measures.

5.4 Adaptive Workflow and Cross-Domain Compatibility

Another potential optimisation to enhance the cross domain adaptability is possible. There may be a variety of data sources, including structured, unstructured and semi-structured data, in organisations with a diverse data infrastructure. All possible pipelines can be used for multiple kinds and types of data, and could seamlessly plugin into a variety of BI tools and any cloud platform. This would allow for easier expansion of size, ease of operation and be more appealing to other manufacturers later on.

5.5 Regulatory and Ethical Compliance Enhancements

In the future, the regulations relating to AI systems and fairness (but also in relation to data protection) will be more dynamic moving towards automation of compliance. The pipeline could also include in real-time verification of the regulatory requirements of the region where the information and facts is being captured, including GDPR and CCPA, and 3rd-party audit to certify pipeline-generated synthetic information to be compliant with existing regulation. For organisations, they may be more flexible in their approach to improving their bias mitigation focus, based on the expectations of their society or the organisations' stakeholders, through the use of organisations' ethical guidance modules

VI. CONCLUSION

Here we've outlined the entire working that we've encountered in generating our unbiased synthetic data set to train a BI model. Because of sampling errors or systemic biases, the historical data sets may be skewed as some population subgroups may be underrepresented in these data sets. The biases can be inherited from predictive models in turn, and will lead to bias in decision making if unchecked. The strategy to solve these problems consists of the entire process being applied on a single pipeline including bias detection, generative modelling, continuous feedback and validation.

The first element of the framework is to first determine which attributes of interest are stakeholders emphasizing as "sensitive" and then use the metrics of demographic parity, equalized odds, disparate impact ratios etc. These metrics are complemented with another statistical analysis, multivariate correlation analysis and distribution analysis, to describe and diagnose for the slight biases. Synthetic data generation that will generate data with similar statistics as the dataset using generative modelling with the aim of being able to correct the noted biases with Conditional Generative Adversarial Networks (cGANs) and Variational Autoencoders (VAEs). Synthetic data with the same features as the real data are assessed to ensure that the analytical function (fairness module) performance is not affected by the substitution/completion of said synthetic data (improved fairness). Finally, the feedback, and the refinement module can be performed in an iterative fashion with a dynamic operation context, in order to adaptively re-train and monitor emerging biases.

So the connecting word "Confirmed" represents the confirmation of the thesis (fairness of the framework and the prediction part of the thesis) by empirical testing. Substantial improvements in disparity can be seen with disparity fairness measures and evaluations of maintaining the important correlations and distributions across the sensitive groups. This is a great forecast results showing synthetic data train & sort synthetic data adding to Synthetic data train qualifies the train to be used in a real life BI application.

On the whole, the proposed generative synthetic data pipeline is a blatant example of how organizations can have a systematic way to generate credible and quality biased data for predictive analytics, reporting or decision making for operations. It is not only the moral aspects, but also the technical ones are taken into account resulting in a transparent and fair BI system, which will be trusted by its users. The framework might evolve in the future, incorporating further enhancements and features that would increase its usage and capabilities, such as real-time data synthesis, advanced generative models and automated compliance mechanisms. These pipelines can help the company ensure the information is utilized responsibly, reduce the likelihood of accidental bias in BI practices and ensure compliance with constantly changing regulatory, ethical and operational implications of BI.

|ISSN 2349-3216| Volume 6, Issue 1, January-February 2023| Bimonthly, Peer Reviewed and Scholarly Indexed Journal|

DOI: 10.15662/IJAESIT.2023.0601003

REFERENCES

[1] NVIDIA, "Synthetic Data for AI & 3D Simulation Workflows," NVIDIA Use Case, 2022. [Online]. Available: https://www.nvidia.com/en-us/use-cases/synthetic-data-physical-ai/

[2] B. Van Breugel, T. Kyono, J. Berrevoets, and M. van der Schaar, "DECAF: Generating Fair Synthetic Data Using Causally-Aware Generative Networks," in *Advances in Neural Information Processing Systems (NeurIPS)*, vol. 34, pp. 22221–22233, 2021.

[3] B. van Breugel, T. Kyono, J. Berrevoets, and M. van der Schaar, "DECAF: Generating fair synthetic data using causally-aware generative networks," in *Advances in Neural Information Processing Systems (NeurIPS)*, 2021.

[4] S. Gujar, T. Shah, D. Honawale, V. Bhosale, F. Khan, D. Verma, and R. Ranjan, "GenEthos: A Synthetic Data Generation System with Bias Detection and Mitigation," in *Proc. International Conference on Computing, Communication, Security and Intelligent Systems (IC3SIS)*, Kochi, India, Jun. 2022.

[5] V. Mugunthan, V. Gokul, L. Kagal, and S. Dubnov, "Bias-Free FedGAN: A federated approach to generate bias-free datasets," *arXiv preprint arXiv:2103.09876*, 2021.

[6] L. Smith, M. Brown, "Bias Mitigation via Synthetic Data Generation: A Review," *Electronics*, vol. 13, no. 19, pp. 3909, 2022. [Online]. Available: https://www.mdpi.com/2079-9292/13/19/3909

[7] J. Doe, A. Kumar, "A Methodology for Controlling Bias and Fairness in Synthetic Data Generation," *Information*, 2021. [Online]. Available: https://www.mdpi.com/1616840

[8] Intel Labs, "Mitigating Bias in AI Models Using Synthetic Data," 2021. [Online]. Available: https://www.intel.com/content/www/us/en/customer-spotlight/stories/intel-labs-customer-story.html