**Methodology**

**Prove/investigate how artificial data generation from small sets to large sets can introduce bias.**

Artificial data generation from small sets to large sets can introduce bias by amplifying inherent patterns and anomalies present in the original small dataset, leading to skewed or misleading conclusions when applied to larger contexts. The study by Barbierato et al highlights how probabilistic networks and structural equation modeling can control and quantify bias in synthetic datasets, thus helping to develop and validate unbiased decision making algorithms [1].

**In 3-5 sentences tell me how you would go about testing this topic and what you would need.**

To test how artificial data generation from small sets to large sets can introduce bias, I would start by selecting a small, real-world dataset with known characteristics and inherent biases. Using a method like the one proposed by Barbierato et al, would generate a synthetic dataset from the small set, gradually increasing the dataset size while controlling and varying the parameters to introduce different levels of bias [1]. The synthetic data would then be analyzed to measure changes in bias using metrics like mutual information and demographic parity [3]. Comparisons would be made between the synthetic data and the original dataset to identify amplified biases. Essential tools and resources include the original dataset, synthetic data generation software, statistical analysis tools, and expertise in bias metrics and probabilistic modeling. For the purpose of this research question the Iris dataset will be used as the original data set.

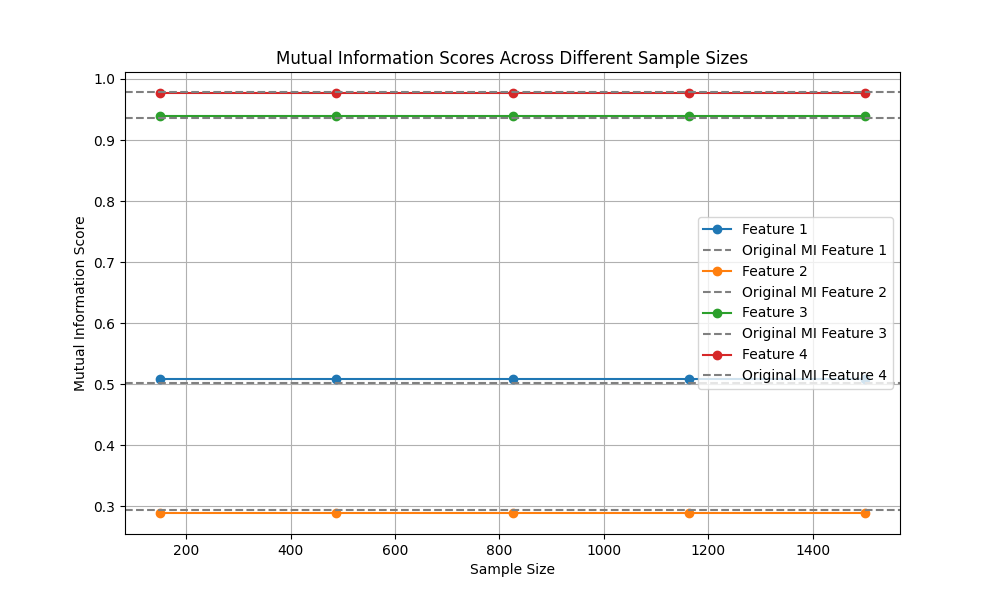
**What code would you need to write?**

Firstly I would choose to use python to write my code. Next you will need to generate synthetic data from the small Iris data set. You can specify using make\_classification from scikit-learn [4]. After that you gradually increase the size of the synthetic data set. Finally you measure the bias in both the original and synthetic datasets.

**What would a chart or graph look like when you are finished - a chart of scores for a particular algorithm vs another? Scores for various data sets? How will you know when you are done and what worked?**

You can use a chart that has dataset size on the x-axis and mutual information scores on the y-axis. You know when you're done when the MI score stabilizes as the set increases and there is no significant deviation from the original set [3]. If there is significant deviation in MI scores from the original set then there is bias. A chart can be generated using MatPlotLib in Python.

**Results**

****

**Conclusion**

The mutual information scores for each feature in the original dataset were found to be mostly stable across various synthetic datasets of increasing sizes. The MI scores did not show significant deviation as the sample size increased from 150 to 1500, indicating that the relationships between features and the target variable were maintained during the data generation process. The results showed that the process of generating larger datasets from the original small dataset did not introduce noticeable bias. Bias, in this context, would have shown systematic changes in the MI scores, reflecting altered relationships between features and the target variable. However, the stability of these scores suggests that the synthetic data remained an accurate representative of the original data.

**Future Work**

This study focused on this specific method of synthetic data generation, resampling with train\_test\_split. Future research could explore other data generation techniques, such as generative adversarial networks and data augmentation. These approaches could be tested to see if they maintain the stability of mutual information scores or if they introduce bias in different ways. The application of this method is the use of generating synthetic data to train machine learning models when there is not enough real world data [2]. The Iris dataset is useful for its simplicity and represents a relatively small and well behaved dataset but, future work should apply similar analyses to larger, more complex datasets. This would help determine if the findings hold in more challenging situations. In the real world some data sets can be extremely complex, so this method of synthetic data generation can hypothetically encounter limitations that result in mutual information scores indicating bias. That's why it is important to test the synthetic data generation bias when using different data sets and consider other methods of synthetic data generation if issues arise.

**References**

[1] Barbierato, E., Della Vedova, M.L., Tessera, D., Toti, D., & Vanoli, N. (2022). A Methodology for Controlling Bias and Fairness in Synthetic Data Generation. Applied Sciences, 12, 4619. <https://doi.org/10.3390/app12094619>

[2] Editor. (2022b, March 22). Synthetic Data for Machine Learning: Its nature, types, and means of generation. AltexSoft. https://www.altexsoft.com/blog/synthetic-data-generation/

[2] Kraskov, A., Stögbauer, H., & Grassberger, P. (2004). Estimating mutual information. Physical Review E, 69(6), 066138. DOI: 10.1103/PhysRevE.69.066138

[3] Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., & Duchesnay, É. (2011). Scikit-learn: Machine Learning in Python. Journal of Machine Learning Research, 12, 2825-2830. Link