Synthetic Data for Alzheimer's Diagnosis

Kristiana Karshelieva, Hunter Merrill

Oct 28, 2025

There are many ways to go about generating synthetic data. We have chosen to go top-to-bottom through the methods discussed in Mendes, et. al in order to fully explore the suite of options available for use in the medical field. As a reminder, the data that we reference from here onward is a classification set of Alzheimer's patients. Our goal is to address the imbalance of ethnicities represented in the data set. Currently, the dataset features 4 enumerated ethnicity values (0 to 3) in the following distribution: 0 - 1278, 1 - 454, 3 - 211, 2 - 206. Ethnicity 0 is the predominant class, with 3 to 6 times more entries in the dataset.

Our baseline for comparison is the MLPClassifier from scikit-learn, trained on the original data for 5000 iterations. Throughout our exploration, we will retain the same classifier and will train it on the various synthetic data we generate to compare performance and fairness. The results from our baseline can be seen in Table 1. Interestingly, despite the skewed dataset, accuracy across all groups is fairly consistent, albeit not very good. There are some notable issues with low precision and recall.

For this initial milestone, we explored the first of Mendes' methods: the rule-based approach. Rule-based approaches apply pre-defined rules to extrapolate from existing data. In our case, we strictly set the number of samples for each ethnicity. We accomplished this by means of the imbalanced-learn modules for under- and oversampling.

'Undersampling' simply means leaving out observations from overrepresented groups in order to manipulate the observation ratios. However, it drastically reduces the amount of data the model can work with (which, in terms of efficacy, is nigh always a bad thing). We undersampled ethnicities 0, 1, and 3 to match the count of ethnicity 2 (206). Observe in Table 2 that undersampling caused a significant negative swing in almost every performance metric, particularly with respect to positive patients—whose small share of the original data was undersampled to a near-insignificant n.

Oversampling, on the other hand, involves repeating rows with under-represented attributes until the total count matches that of the over-represented one(s). While more data is conducive to more learning, bootstrapping it in this way raises the risk of overfitting. In our case, we oversampled ethnicities 1, 2, and 3 to the count of ethnicity 0 (1278). This rubber-banded the positive scores of every category, but confusingly tanked the negatives, as seen in 3 (note especially the recalls). We will have to investigate further as to why this phenomenon occurred.

Finally, in an effort to counteract the small sample size of the original data set, we oversampled *all* of the ethnicities to 10× the count of ethnicity 0 (12780). This approach revived the negative f1-scores to levels well above that of the original data's model, although the positive sits proportionally lower. With significantly improved averages across the board, this approach has potential (Table 4).

Another curious synthetic data generation method we employed is asking GPT-5 to create synthetic data for us. To simulate use by an individual with little expertise in data and ML, we let the model "choose" how to generate the data rather than giving it a specific approach. We wanted to assess the outcomes of trusting an LLM that has not been fine-tuned for the task. GPT-5 successfully generated new data without altering the original dataset. Other than that, we followed the same approach for training as in the baseline. The performance can be seen in Table 5 - it is better than the baseline in terms of accuracy in all categories, while precision and recall are comparable. However, such an approach should not be trusted lightly since the methodology of data generation is not transparent.

One of our main difficulties is that our synthetic generation methods have failed to produce any significant improvements over the baseline despite improved balance. We have not been able to identify why the baseline is performing so well with skewed data, but possible reasons are overfitting of the baseline model, small relevance of the ethnicity attribute for diagnosis and poor or inappropriate for the tasks synthetic generation methods. Additionally, we rely on classic performance metrics provided by scikit-learn to analyze our results. These metrics are a good start, but we want to integrate fairness specific metrics as seen in our literature review. However, we struggled to identify exactly how to tailor these for assessing bias with respect to several ethnicity groups.

Our goals for the next milestone are to figure out how to tailor the various fairness metrics we discovered in our literature review as well as any knew ones we may find to our specific task of balancing our dataset by ethnicity. Additionally, we want to develop the other 2 approaches for generation outlined in Mendes, et. al, with a particular focus on ML generation due to its current popularity. Finally, we may look into exploring several models for the classification itself rather than using just the standard MLPClassifier from scikit-learn and see what interactions emerge with the synthetic data.

ethnicity		precision	recall	f1-score	support
0	0	0.90	0.60	0.72	222
	1	0.55	0.88	0.68	124
	accu			0.70	346
	m. avg	0.72	0.74	0.70	346
	w. avg	0.77	0.70	0.70	346
1	0	0.89	0.63	0.74	63
	1	0.54	0.84	0.66	32
	accu			0.71	95
	m. avg	0.71	0.74	0.70	95
	w. avg	0.77	0.71	0.71	95
2	0	0.78	0.61	0.68	23
	1	0.64	0.80	0.71	20
	accu			0.70	43
	m. avg	0.71	0.70	0.70	43
	w. avg	0.71	0.70	0.70	43
3	0	0.90	0.65	0.75	40
	1	0.44	0.79	0.56	14
	accu			0.69	54
	m. avg	0.67	0.72	0.66	54
	w. avg	0.78	0.69	0.70	54

Table 1: Classification metrics by ethnicity group (original data)

ethnicity		precision	recall	f1-score	support
0	0	0.73	1.00	0.85	33
	1	0.00	0.00	0.00	12
	accu			0.73	45
	m. avg	0.37	0.50	0.42	45
	w. avg	0.54	0.73	0.62	45
1	0	0.69	1.00	0.82	34
	1	0.00	0.00	0.00	15
	accu			0.69	49
	m. avg	0.35	0.50	0.41	49
	w. avg	0.48	0.69	0.57	49
2	0	0.64	1.00	0.78	32
	1	1.00	0.22	0.36	23
	accu			0.67	55
	m. avg	0.82	0.61	0.57	55
	w. avg	0.79	0.67	0.60	55
3	0	0.68	1.00	0.81	36
	1	1.00	0.19	0.32	21
	accu			0.70	57
	m. avg	0.84	0.60	0.56	57
	w. avg	0.80	0.70	0.63	57

Table 2: Classification metrics by ethnicity group (undersampled to lowest count)

ethnicity		precision	recall	f1-score	support
0	0	1.00	0.04	0.08	210
	1	0.36	1.00	0.52	111
	accu			0.37	321
	m. avg	0.68	0.52	0.30	321
	w. avg	0.78	0.37	0.24	321
1	0	1.00	0.08	0.15	246
	1	0.21	1.00	0.35	60
	accu			0.26	306
	m. avg	0.60	0.54	0.25	306
	w. avg	0.85	0.26	0.19	306
2	0	1.00	0.09	0.16	235
	1	0.27	1.00	0.43	80
	accu			0.32	315
	m. avg	0.64	0.54	0.29	315
	w. avg	0.81	0.32	0.23	315
3	0	1.00	0.14	0.24	295
	1	0.14	1.00	0.24	41
	accu			0.24	336
	m. avg	0.57	0.57	0.24	336
	w. avg	0.89	0.24	0.24	336

Table 3: Classification metrics by ethnicity group (oversampled to max count)

ethnicity		precision	recall	f1-score	support
0	0	0.87	0.97	0.92	2717
	1	0.60	0.23	0.34	512
	accu			0.85	3229
	m. avg	0.74	0.60	0.63	3229
	w. avg	0.83	0.83	0.83	3229
1	0	0.91	0.97	0.94	2820
	1	0.56	0.27	0.36	365
	accu			0.89	3185
	m. avg	0.74	0.62	0.65	3185
	w. avg	0.87	0.89	0.87	3185
2	0	0.84	0.97	0.90	2580
	1	0.66	0.21	0.32	612
	accu			0.83	3192
	m. avg	0.75	0.59	0.61	3192
	w. avg	0.80	0.83	0.79	3192
3	0	0.91	0.99	0.95	2847
	1	0.69	0.18	0.29	327
	accu			0.91	3174
	m. avg	0.80	0.59	0.62	3174
	w. avg	0.89	0.91	0.88	3174

Table 4: Classification metrics by ethnicity group (oversampled to 10x max count)

ethnicity		precision	recall	f1-score	support
0	0	0.85	0.83	0.84	210
	1	0.70	0.72	0.71	112
	accu			0.80	322
	m. avg	0.77	0.78	0.78	322
	w. avg	0.80	0.80	0.80	322
1	0	0.93	0.86	0.89	248
	1	0.55	0.71	0.62	59
	accu			0.83	307
	m. avg	0.74	0.79	0.75	307
	w. avg	0.85	0.83	0.84	307
2	0	0.81	0.94	0.87	233
	1	0.67	0.38	0.49	81
	accu			0.79	314
	m. avg	0.74	0.66	0.68	314
	w. avg	0.78	0.79	0.77	314
3	0	0.94	0.97	0.95	295
	1	0.70	0.53	0.60	40
	accu			0.92	335
	m. avg	0.82	0.75	0.78	335
	w. avg	0.91	0.92	0.91	335

Table 5: Classification metrics by ethnicity group (ChatGPT generated)

Mendes, Jorge M., et al. “Synthetic Data Generation: A Privacy-Preserving Approach to Accelerate Rare Disease Research.” Frontiers in Digital Health, vol. 7, Mar. 2025, p. 1563991. PubMed Central, https://doi.org/10.3389/fdgth.2025.1563991.