Generator syntetycznych danych tabelarycznych z użyciem przestrzeni osadzeń: studium użycia w medycynie

DR INŻ. JAROSŁAW DRAPAŁA

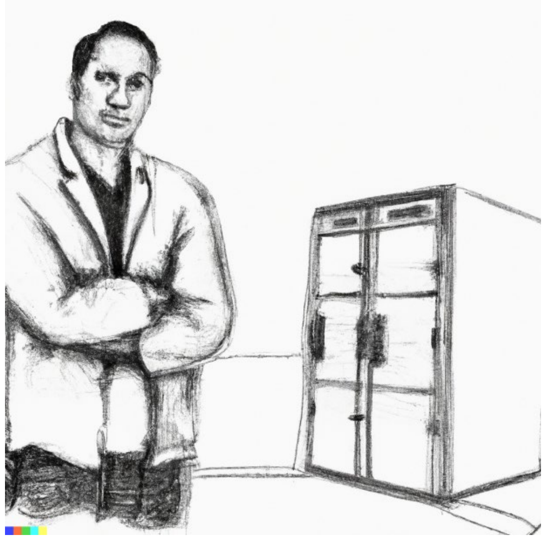
Plan prezentacji

- Opis **problemu** – geneza tematu
- Propozycja rozwiązania **spoza dziedziny uczenia głębokiego**
- Studium przypadku – rezultaty obliczeń
- **Porównanie** ze standardową metodą uczenia głębokiego
- Przykłady **innych zastosowań** przedstawionych koncepcji

Problem

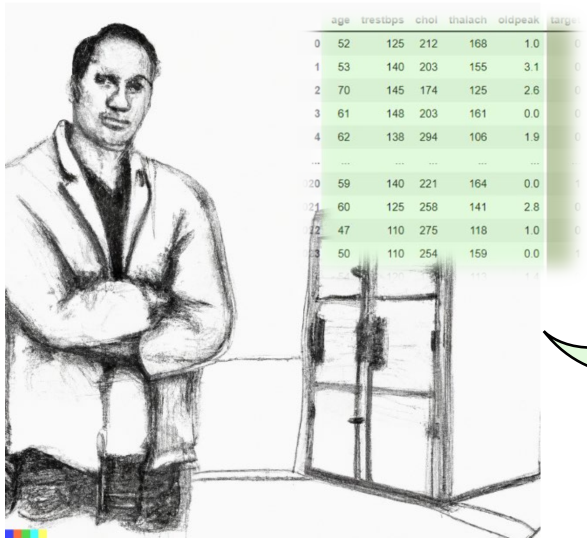
	Height	Weight	LDL cholesterol	HDL cholesterol	Total cholesterol	CRP ultrasensitive	Age	Gender	Hypertension	Diabetes mellitus	Healthy	BMI
653	177	94	76	25	124	18.04	83	Male	Yes	Yes	No	30.0
594	169	71	81	64	169	2.72	76	Male	Yes	No	No	24.9
218	172	72	193	37	248	16.31	71	Male	No	No	Yes	24.3
155	158	80	64	28	127	1.91	82	Female	Yes	Yes	No	32.0

This is a database that stores the data of my patients, but you cannot access them.



I can design and develop an ML solution for you, but I need your data to train models.

Credible fake dataset

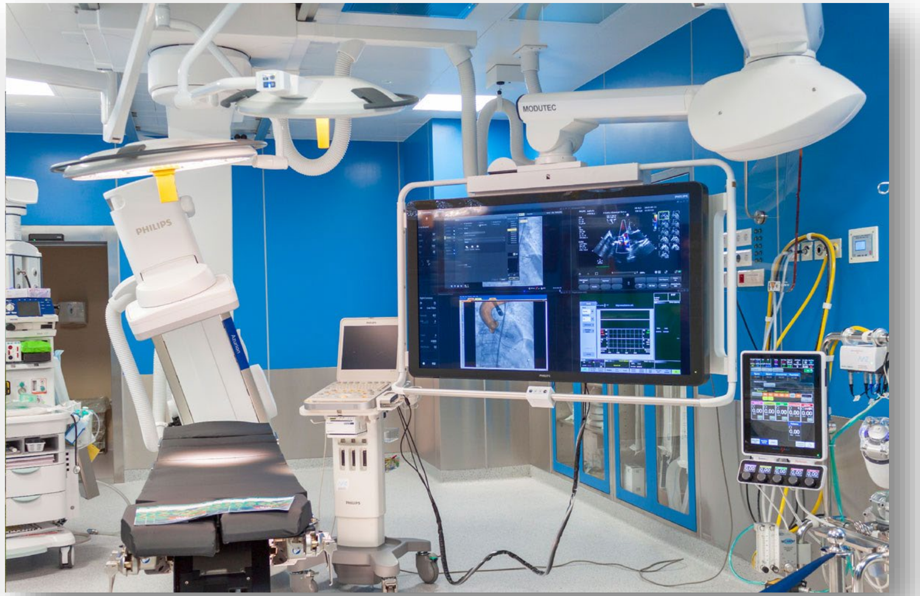


THE SYNTHETIC DATA VAULT

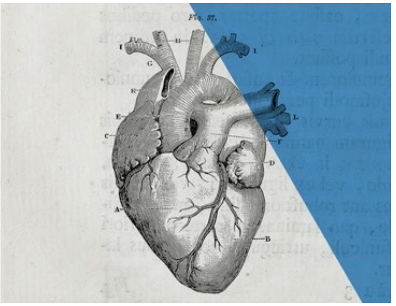
CTGAN

	age	trestbps	chol	thalach	oldpeak	target
0	72	164	248	122	1.1	1
1	55	181	234	161	0.7	0
2	60	163	341	136	3.1	0
3	63	143	207	138	2.1	0
4	68	148	212	168	0.8	1
...	...	...	...	...	...	...
99995	66	122	235	180	0.1	1
99996	51	105	206	132	2.3	1
99997	57	126	196	163	0.0	0
99998	46	114	218	105	1.8	0
99999	54	126	253	179	0.1	1

Case study



INSTYTUT CHORÓB SERCA
UNIWERSYTETU MEDYCZNEGO WE WROCŁAWIU



Centre for Heart Diseases

The Center for Heart Diseases at the University Hospital in Wroclaw – a leading center integrating the work of cardiologists and cardiac surgeons, offering a full profile of cardiovascular therapy for adults around the clock.

Dataset

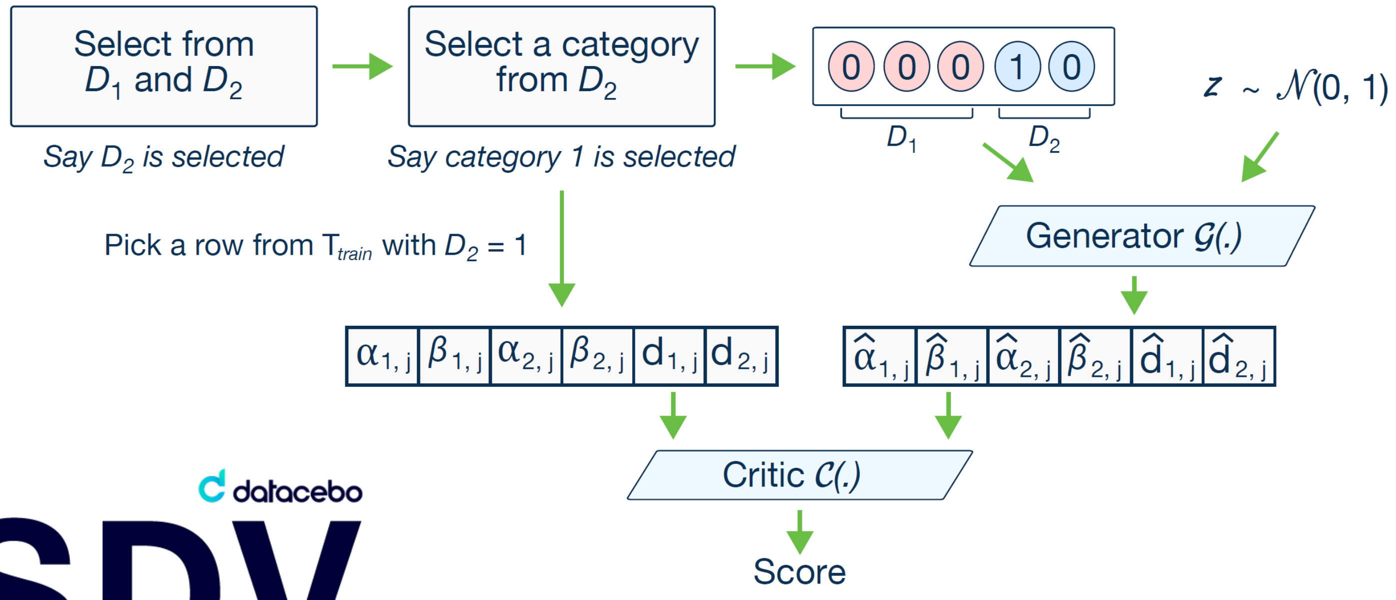
	Height	Weight	LDL cholesterol	HDL cholesterol	Total cholesterol	CRP ultrasensitive	Age	Gender	Hypertension	Diabetes mellitus	Healthy	BMI
653	177	94	76	25	124	18.04	83	Male	Yes	Yes	No	30.0
594	169	71	81	64	169	2.72	76	Male	Yes	No	No	24.9
218	172	72	193	37	248	16.31	71	Male	No	No	Yes	24.3
155	158	80	64	28	127	1.91	82	Female	Yes	Yes	No	32.0
448	164	110	120	37	172	10.98	77	Female	Yes	No	No	40.9
394	160	68	125	42	194	17.36	69	Female	Yes	No	No	26.6
244	158	72	76	38	141	3.67	78	Female	No	No	Yes	28.8
443	175	70	70	43	126	16.47	64	Male	Yes	No	No	22.9
439	175	68	90	18	128	4.39	65	Male	No	No	Yes	22.2
601	170	87	41	24	80	20.43	71	Male	Yes	Yes	No	30.1
203	178	100	100	29	148	149.17	69	Male	Yes	Yes	No	31.6
35	180	87	178	48	273	5.71	48	Male	No	Yes	No	26.9
503	167	62	86	44	149	2.15	69	Male	Yes	No	No	22.2
94	178	90	202	87	315	0.84	62	Male	Yes	Yes	No	28.4

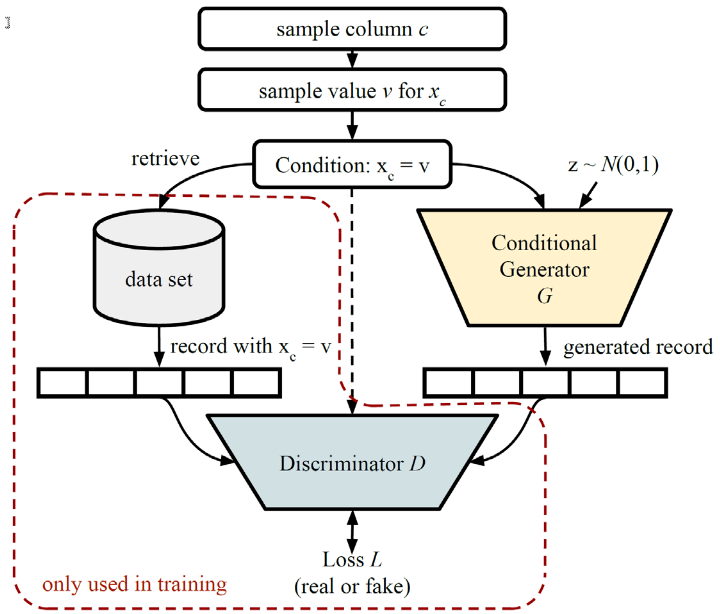
Dataset

710 patients out of 1068
10 variables out of 39
2 variables are dummy

Feature name	type	range
Height	*numerical*	$\langle 142, 200 \rangle$
Weight	*numerical*	$\langle 36.4, 198 \rangle$
LDL cholesterol	*numerical*	$\langle 8, 226 \rangle$
HDL cholesterol	*numerical*	$\langle 8, 121 \rangle$
Total cholesterol	*numerical*	$\langle 51, 368 \rangle$
CRP ultrasensitive	*numerical*	$\langle 0.08, 263.77 \rangle$
Age	*numerical*	$\langle 24, 97 \rangle$
Gender	*categorical*	{*Female*: 33%, *Male*: 67%}
Hypertension	*categorical*	{*Yes*: 74%, *No*: 26%}
Diabetes mellitus	*categorical*	{*Yes*: 44%, *No*: 56%}
BMI	*numerical*, dummy	$\langle 11.95, 70.15 \rangle$
Healthy	*categorical*, dummy	{*Yes*: 18%, *No*: 82%}

CTGAN
*C*onditional *T*abular *G*enerative *A*dversarial *N*etworks





datacebo
SDV
The Synthetic Data Vault

Borisov, V., Leemann, T., Seßler, K., Haug, J., Pawelczyk, M., & Kasneci, G. (2022). Deep neural networks and tabular data: A survey. *IEEE Transactions on Neural Networks and Learning Systems*.

Xu, L., Skoularidou, M., Cuesta-Infante, A., & Veeramachaneni, K. (2019). Modeling tabular data using conditional gan. *Advances in neural information processing systems*, *32*.

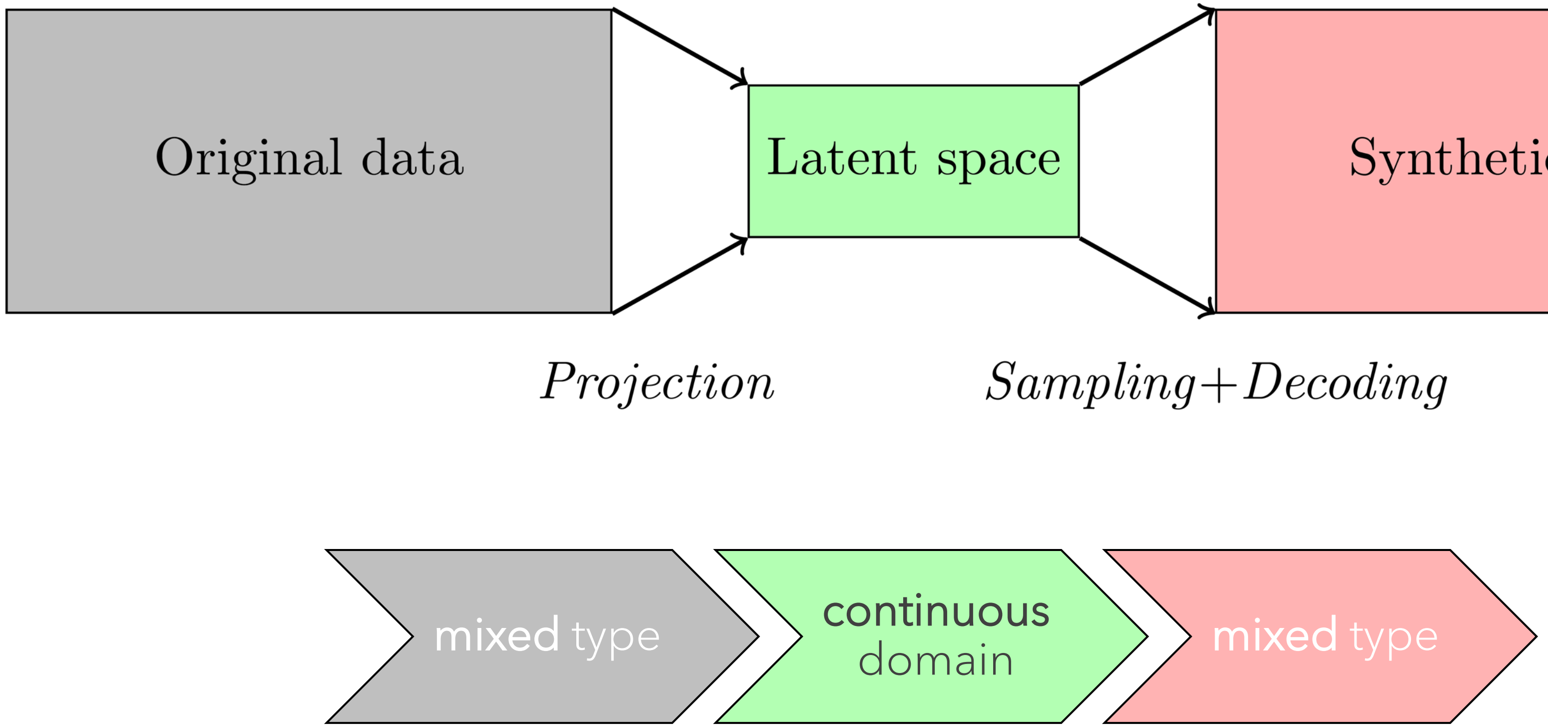
PROPOSED SOLUTION

Composite SDG

Ingredients

- [ ] Multidimensional Scaling
- [ ] Kernel Density Estimator
- [ ] Support Vector Machines
- [ ] Random Forests

The role of a latent space

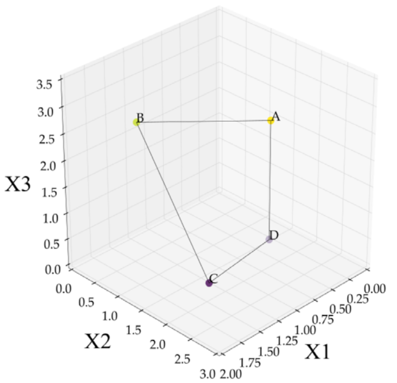


Multidimensional scaling – MDS

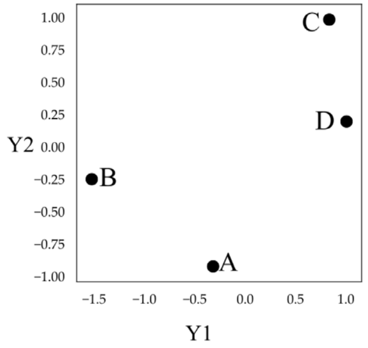
$$E\left(\widehat{X}\right) = \sum_{m=1}^{N}\sum_{n=1}^{N}\left(D_{mn} - \widehat{D}_{mn}\right)^2 A_{mn} \qquad \widehat{D}_{mn} = \left[d(\widehat{\mathbf{x}}_m, \widehat{\mathbf{x}}_n)\right]$$

	Height	Weight	LDL cholesterol	HDL cholesterol	Total cholesterol	CRP ultrasensitive	Age	Gender	Hypertension	Diabetes mellitus	Healthy	BMI
653	177	94	76	25	124	18.04	83	Male	Yes	Yes	No	30.0
594	169	71	81	64	169	2.72	76	Male	Yes	No	No	24.9
218	172	72	193	37	248	16.31	71	Male	No	No	Yes	24.3

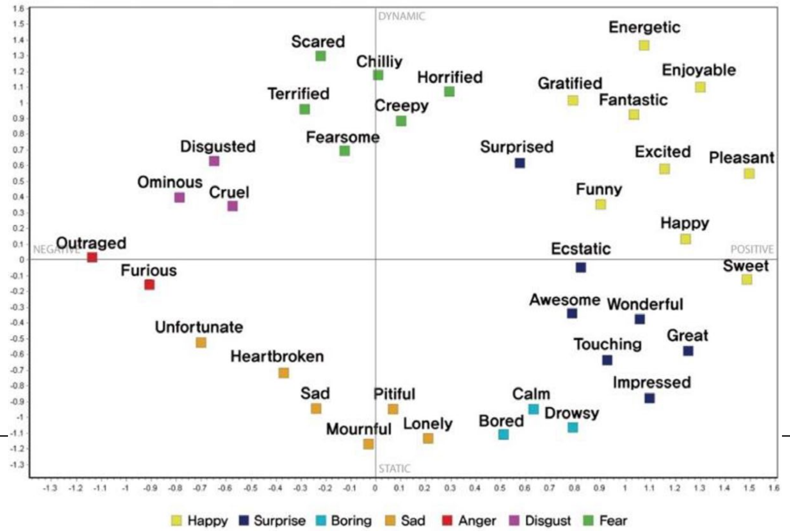
$$\begin{bmatrix} 1.41 \\ 0.91 \end{bmatrix} \quad \begin{bmatrix} 0.87 \\ -0.81 \end{bmatrix} \quad \begin{bmatrix} 0.37 \\ 0.22 \end{bmatrix}$$



Euclidean Distance in Original Space (3-dimensions)				
A	B	C	D	Entity
	1.69	2.53	2.20	A
		2.66	2.61	B
			0.82	C
				D



Euclidean Distance in Lower Dimension (2-dimensions)				
A	B	C	D	Entity
	1.71	2.53	2.20	A
		2.67	2.59	B
			0.82	C
				D



Probabilistic model operating in a latent space

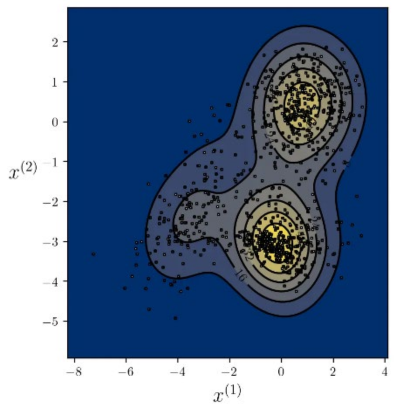
Original data → MDS → Data encoded in latent space → KDE based generator

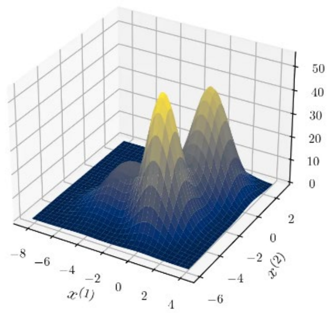
Projection

Estimation of PDF

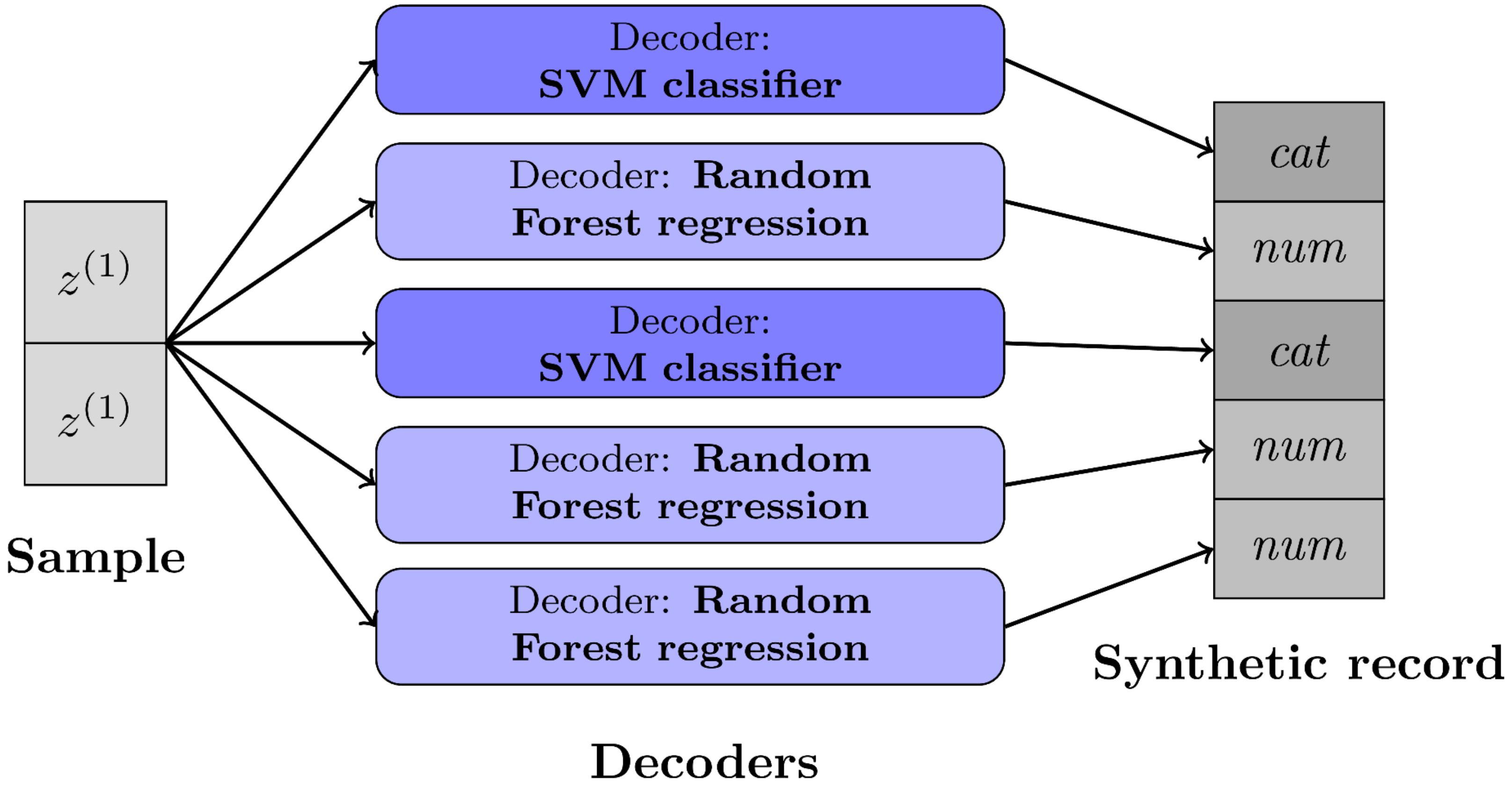
$$f(\mathbf{x}) = \frac{1}{n} \sum_{i=1}^{n} K\left(\mathbf{x} - \mathbf{x}_i\right)$$

$$K(\mathbf{x}) = (2\pi)^{-\frac{d}{2}} \det(\mathbf{H})^{-\frac{1}{2}} \; e^{-\frac{1}{2}\mathbf{x}^{\mathrm{T}}\mathbf{H}^{-1}\mathbf{x}}$$

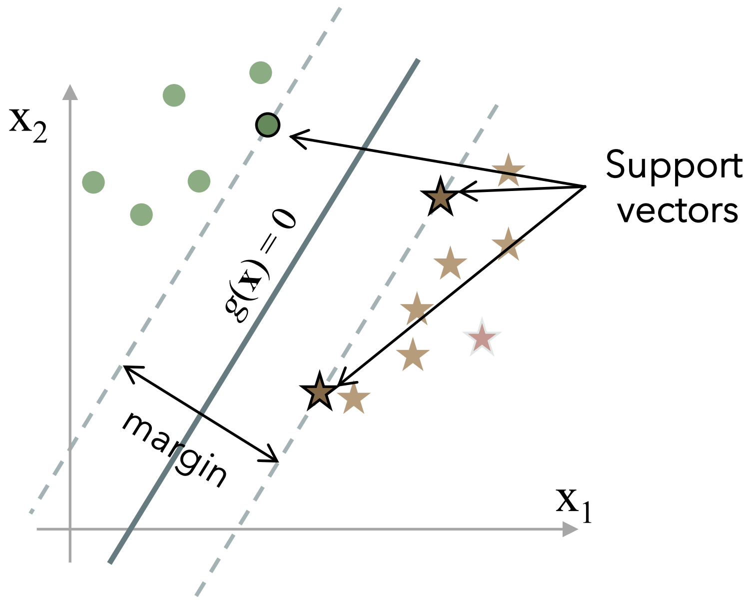




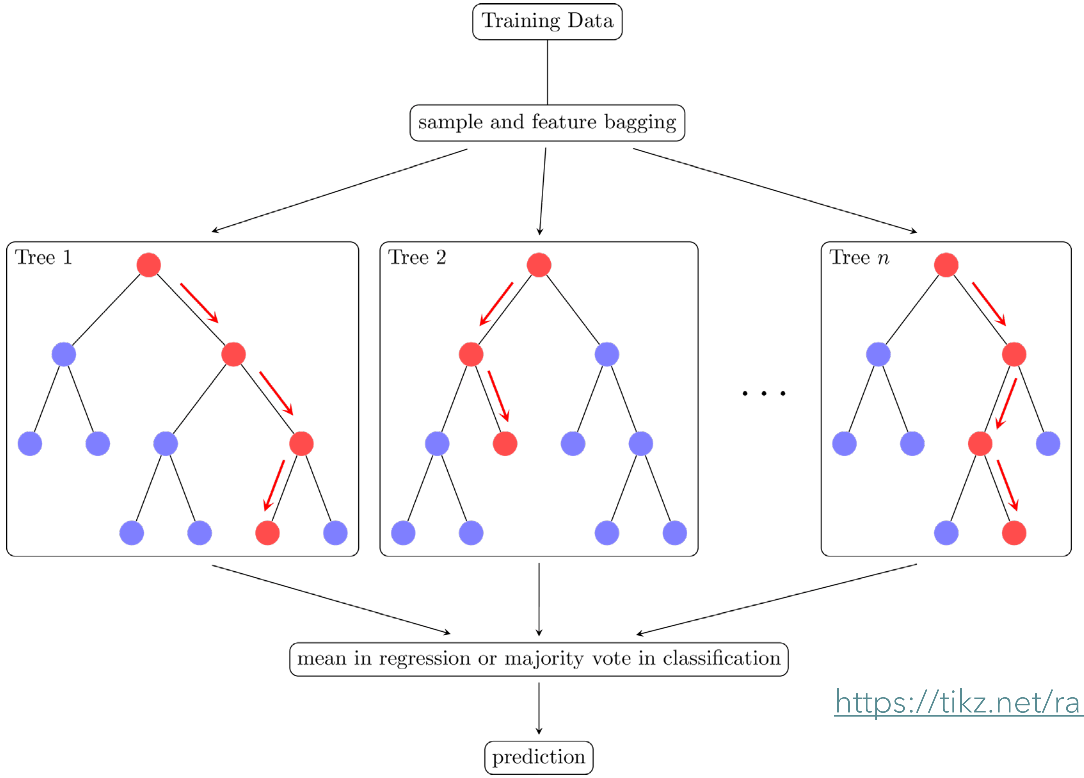
Decoding a latent space sample to its full form



Decoders



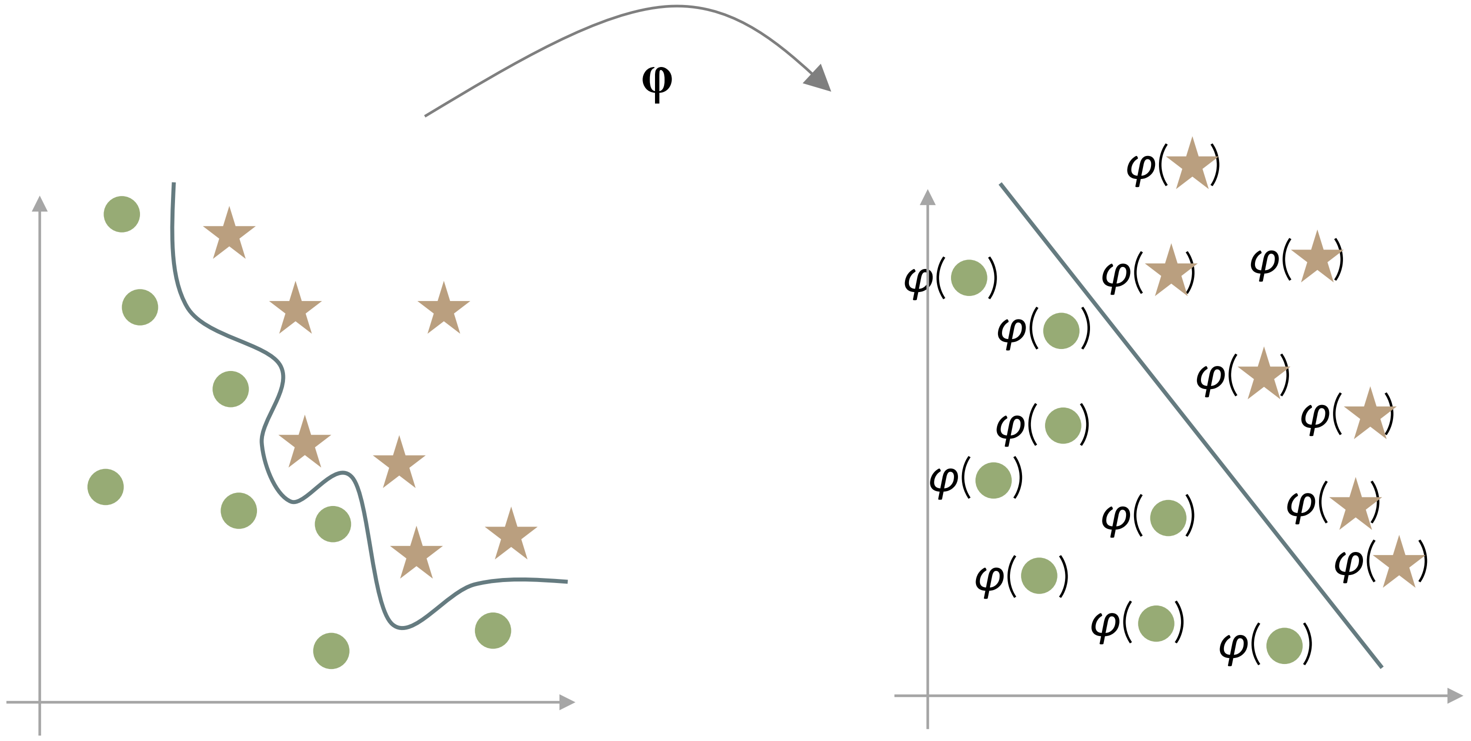
Support Vector Machine
for Classification



https://tikz.net/random-forest/

Random Forest Regression

Decoders

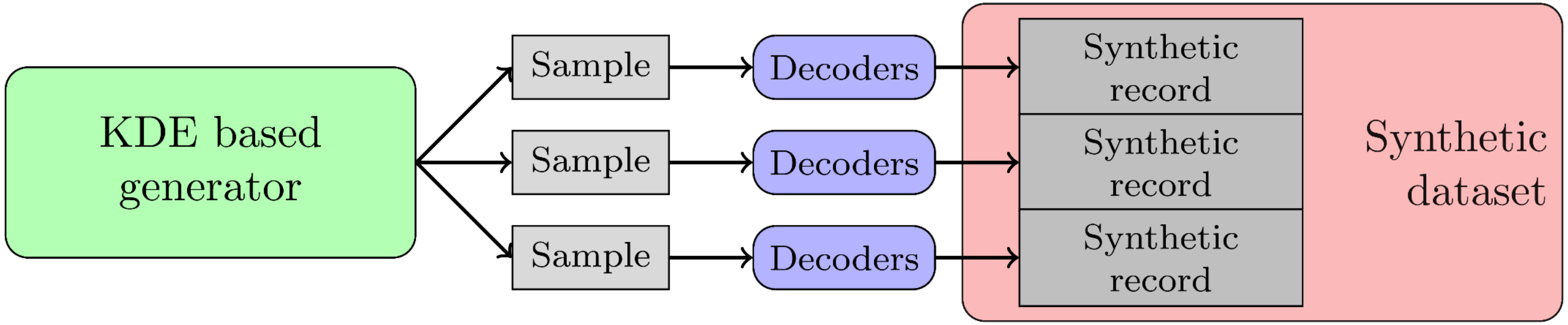


Kernel SVM

$$g(\mathbf{x}) = \sum \lambda_n y_n \varphi^T(\mathbf{x}_n) \varphi(\mathbf{x}) + w_0$$

$$K(\mathbf{x}_n, \mathbf{x}_m) = \varphi^T(\mathbf{x}_n) \varphi(\mathbf{x}_m)$$

Generation of synthetic records

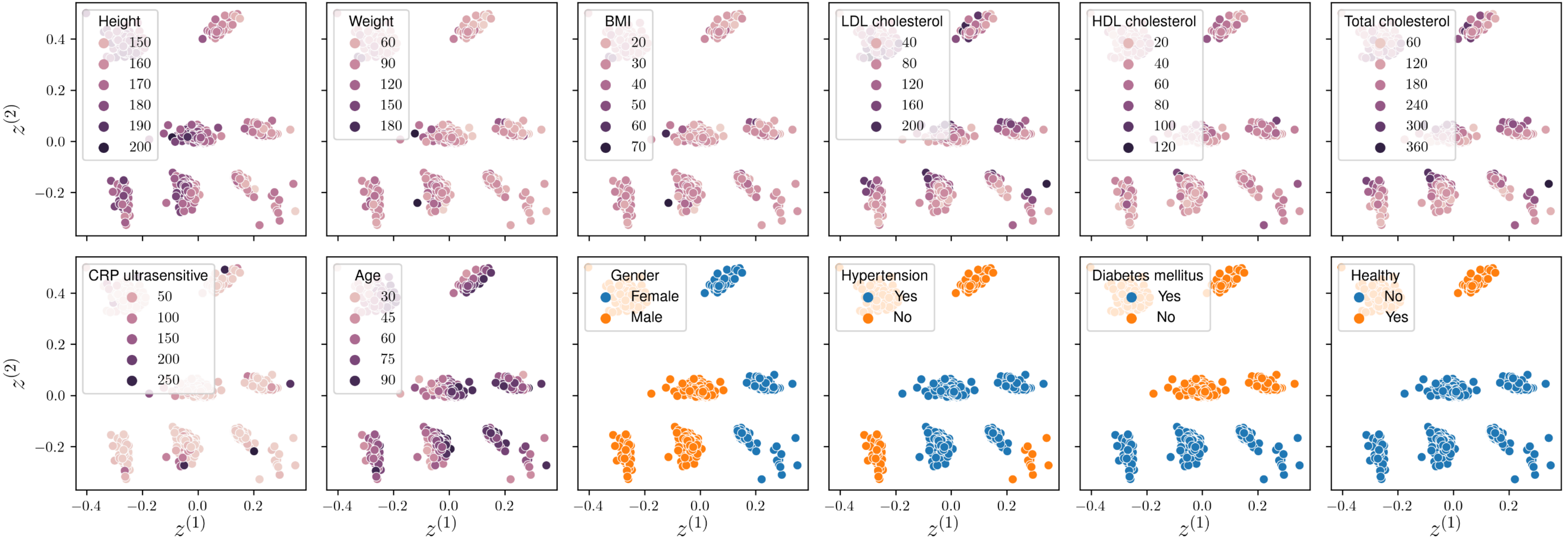


https://github.com/jdrapala/CompositeSDG

Results obtained for the cardiological dataset

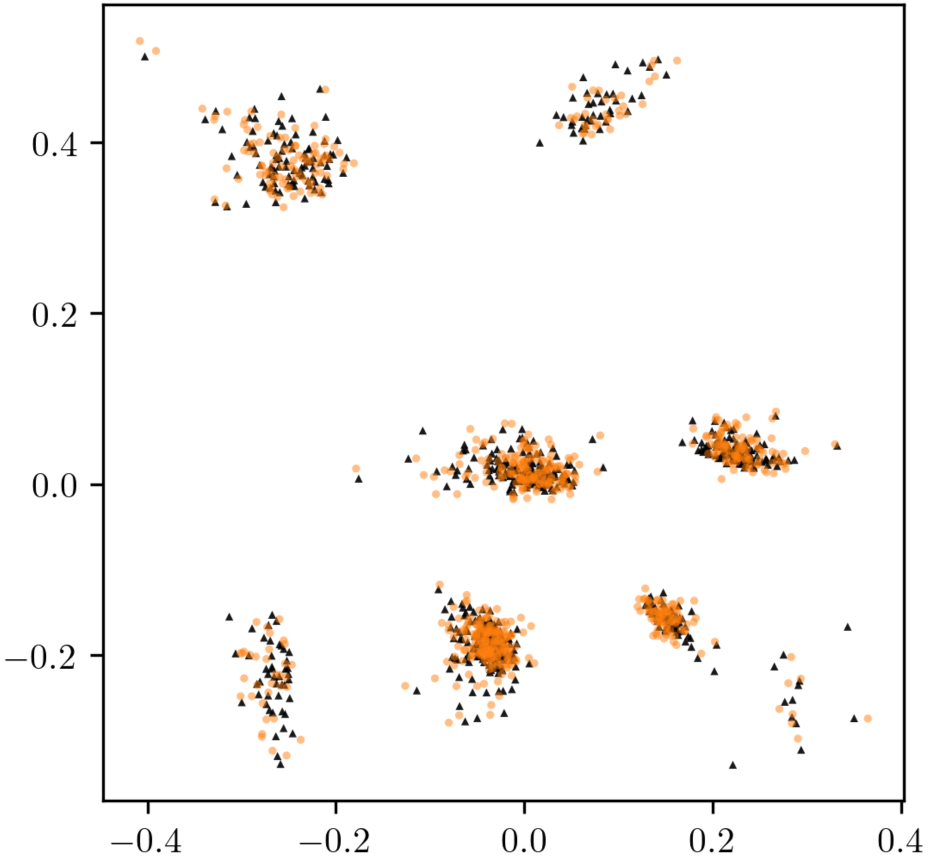
https://kacperswirkula.pythonanywhere.com/
login: zpi / hasło: zpi

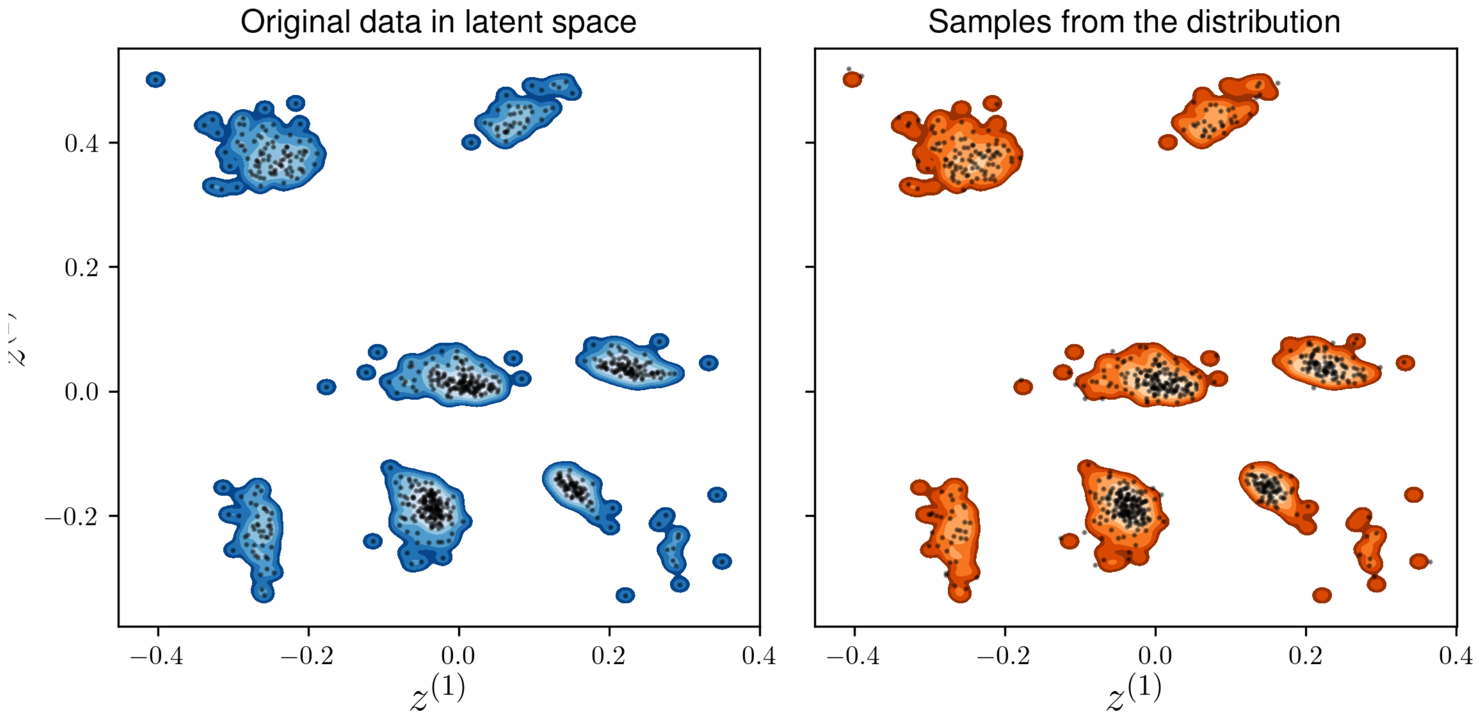
Latent space representation of dataset




```
Dist_matrix = pairwise_distances(df_dataset_scaled, metric='cosine')
projected_data = MDS(n_components=2, dissimilarity='precomputed', normalized_stress='auto').fit_transform(Dist_matrix)
```

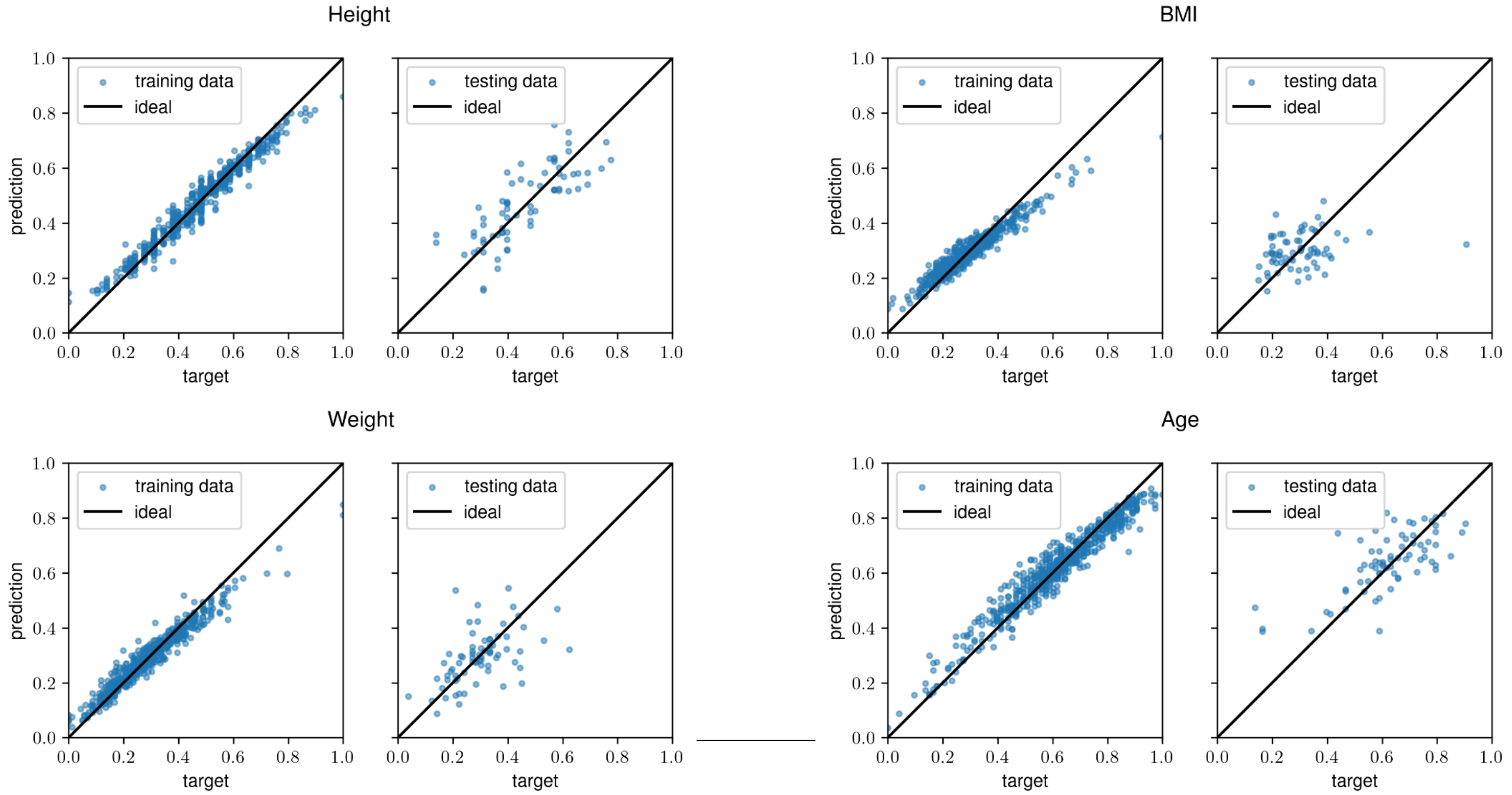
Kernel Density Estimator – KDE





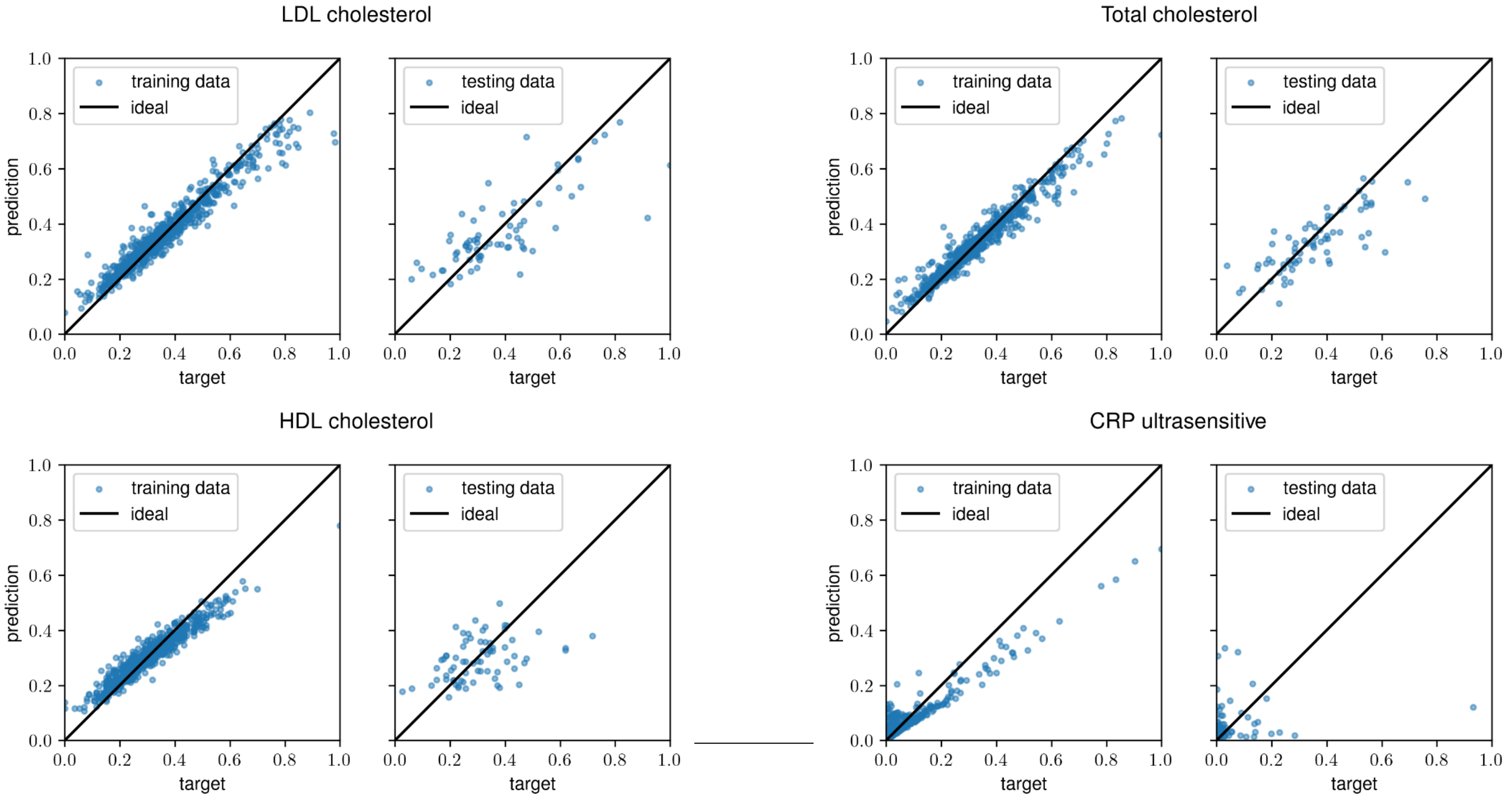

```
kde = KernelDensity(kernel='gaussian', bandwidth=0.008).fit(df_dataset_latent)
```

Performance of decoders – numerical



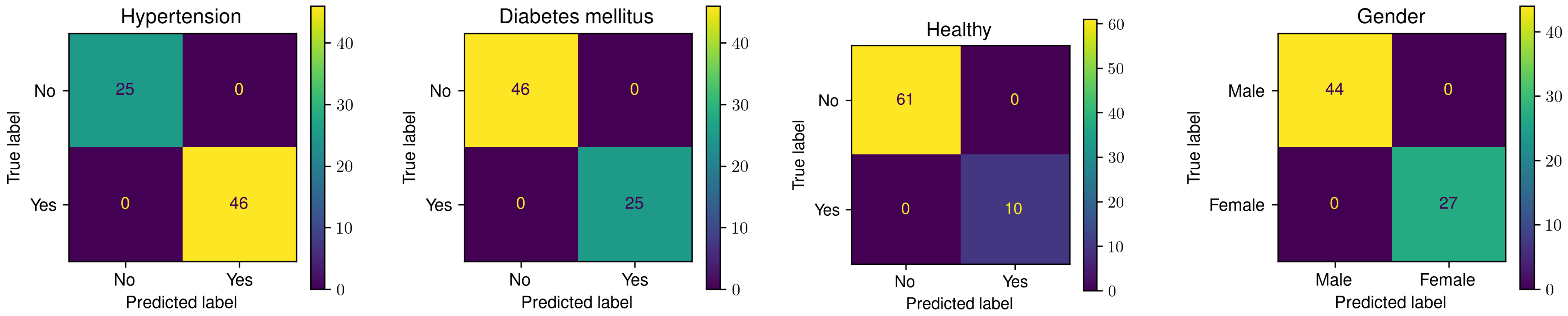

```
model=RandomForestRegressor()
```

Performance of decoders – numerical




```
model=RandomForestRegressor()
```

Performance of decoders – categorical



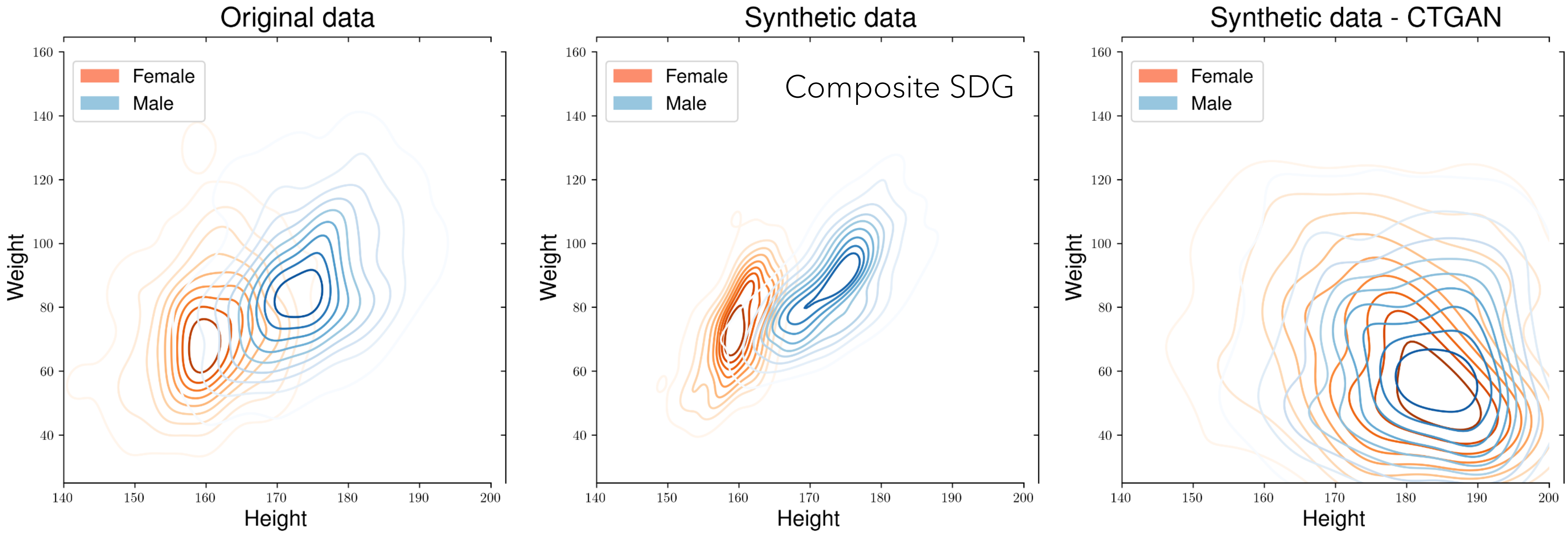

```
model=SVC()
```

Composite SDG vs CTGAN

Joint distribution

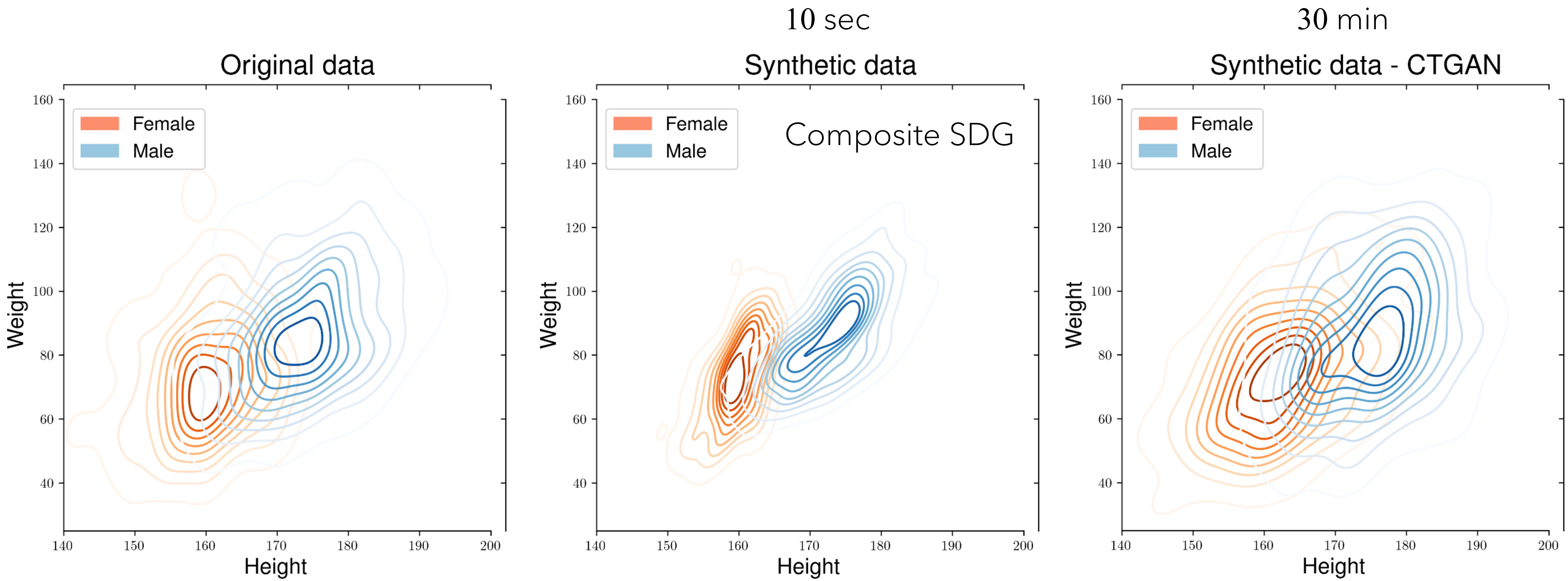
10 sec

50 sec




```
ct = ctgan.CTGAN(epochs=20000)
```

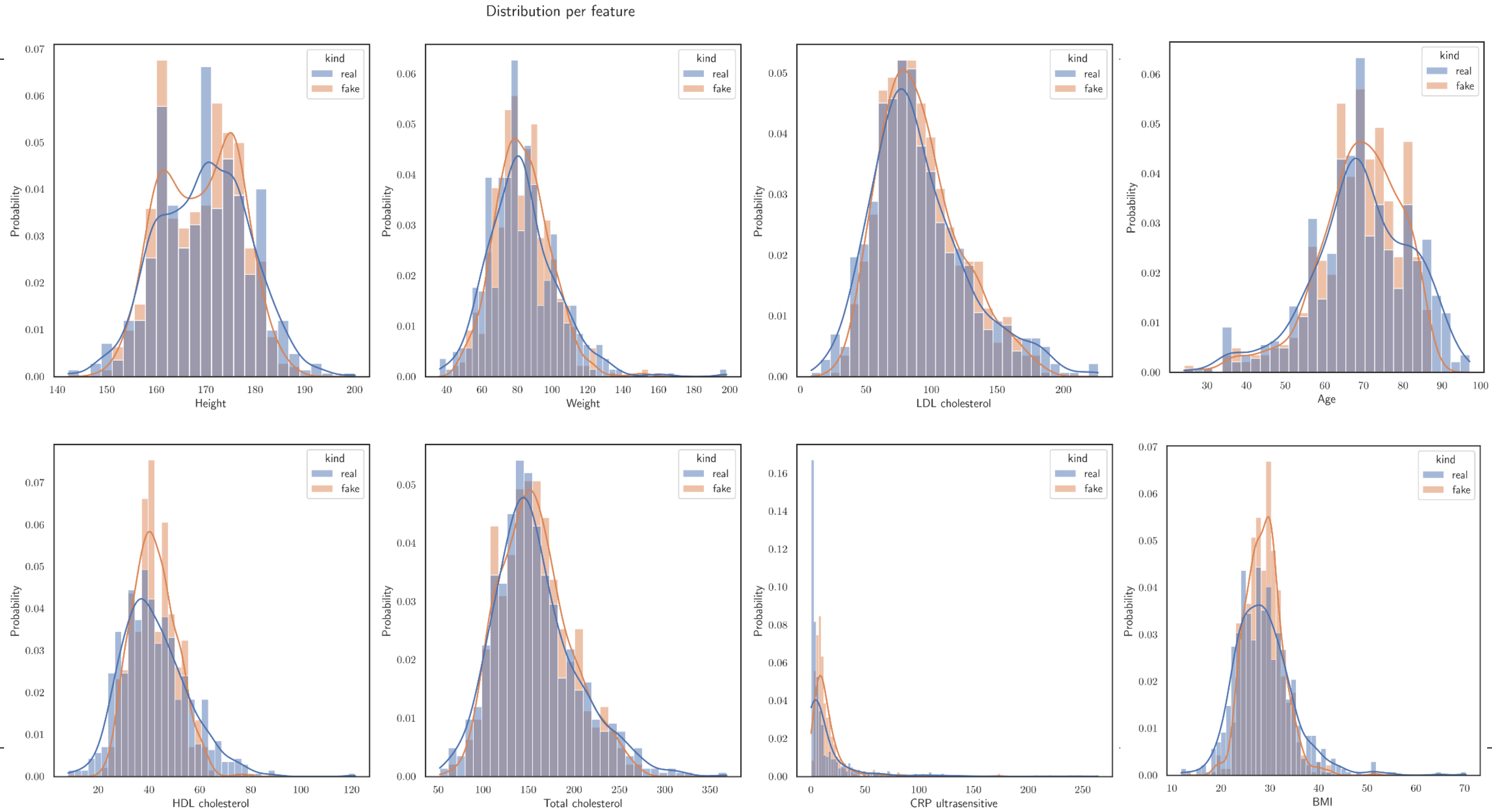
Joint distribution




```
ct = ctgan.CTGAN(epochs=20000)
```

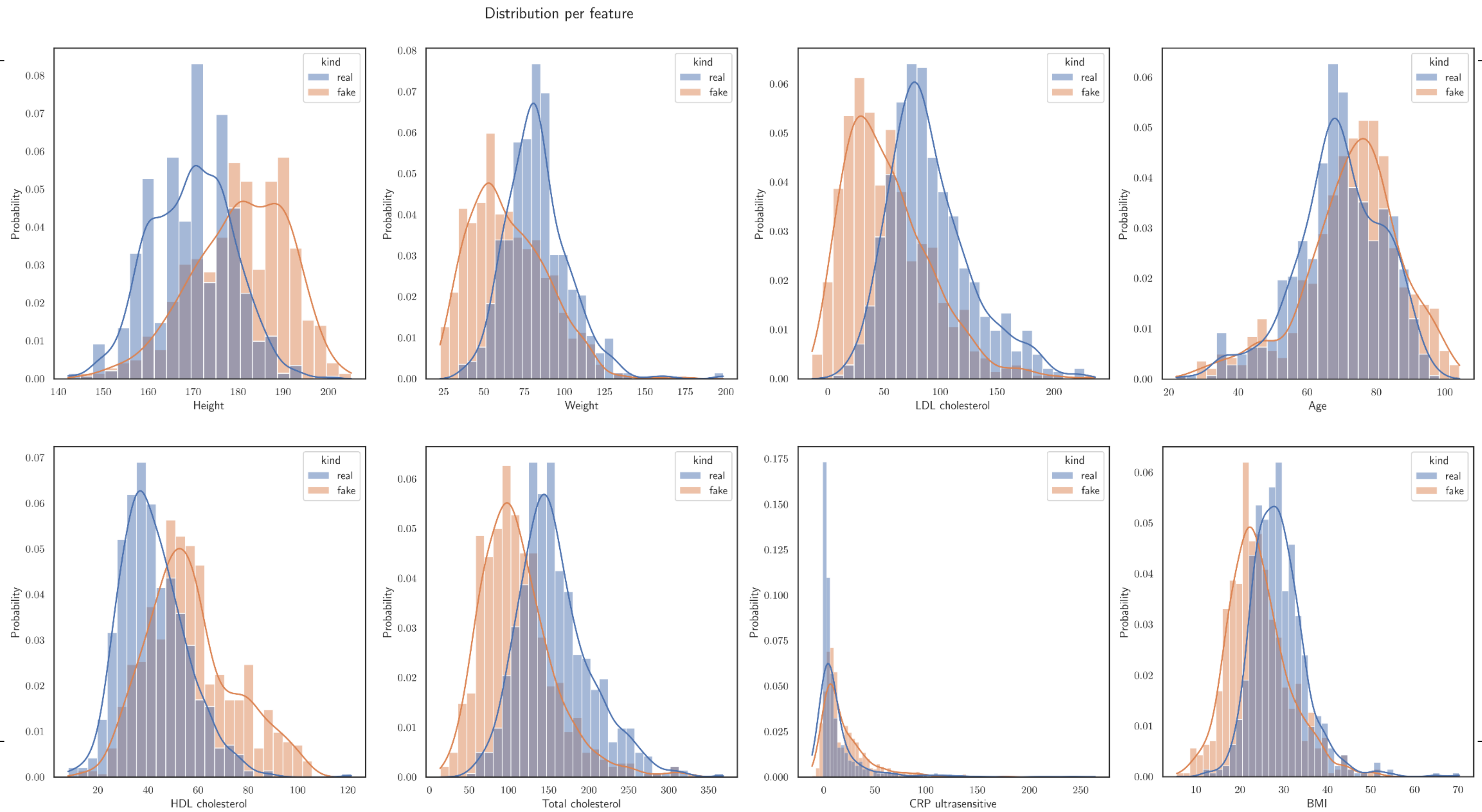
Distributions

Composite SDG



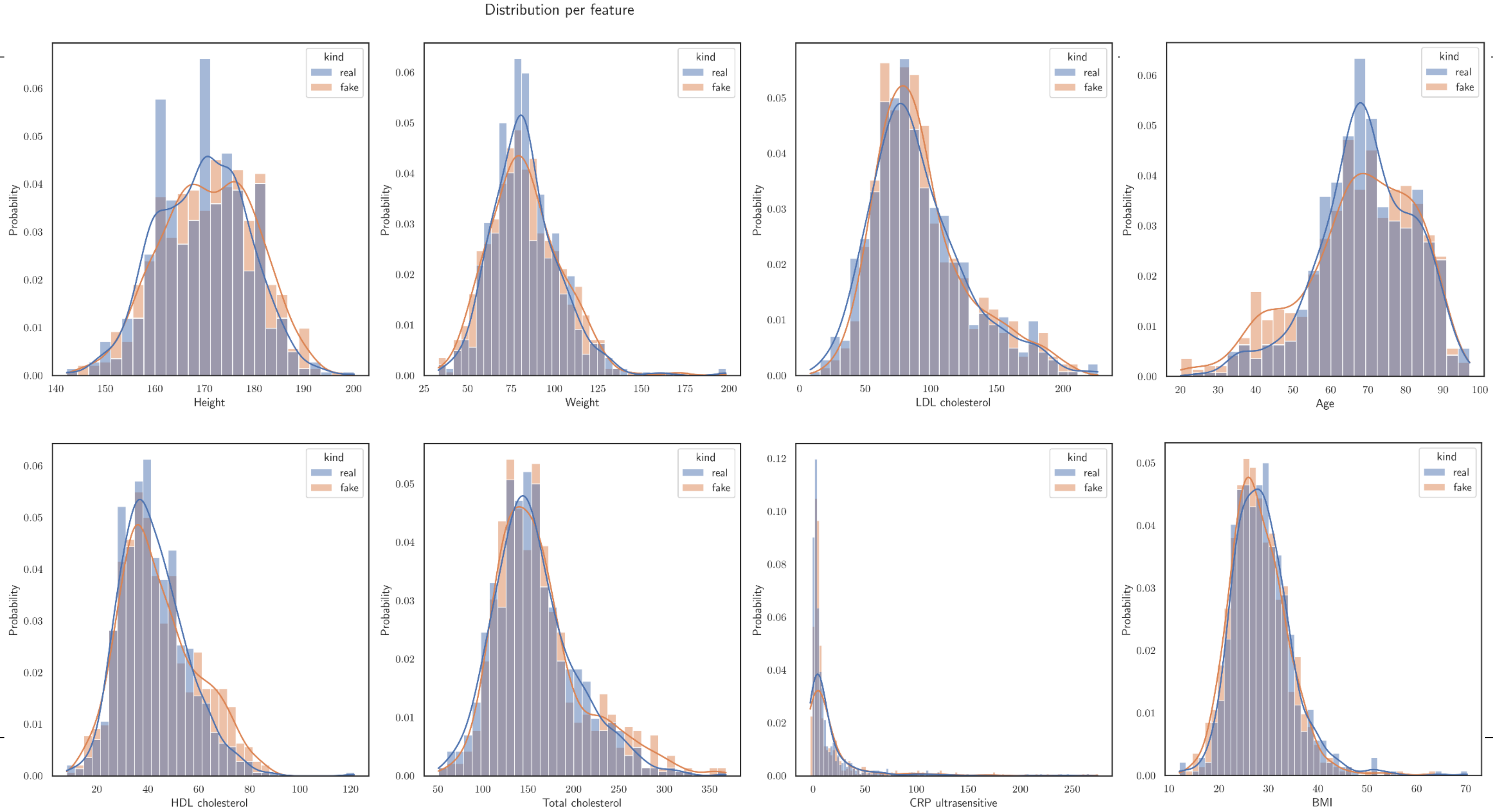
Distributions

CTGAN 50 sec



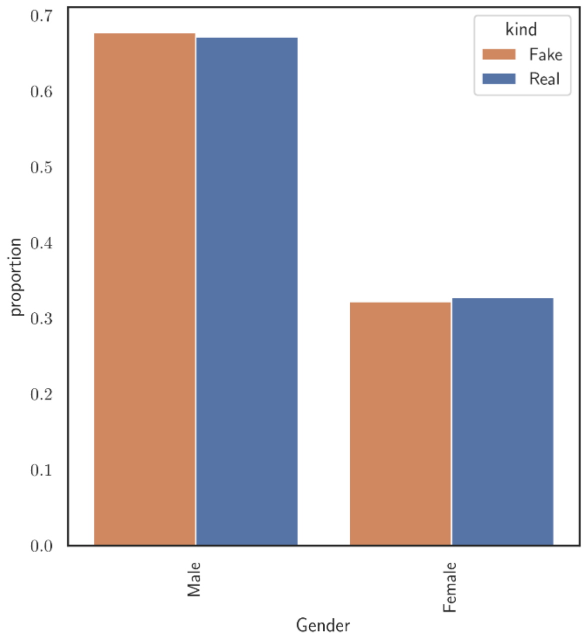
Distributions

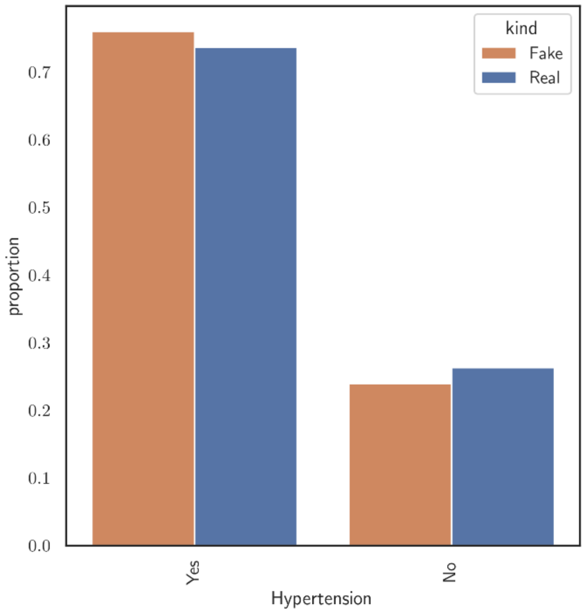
CTGAN 30 min

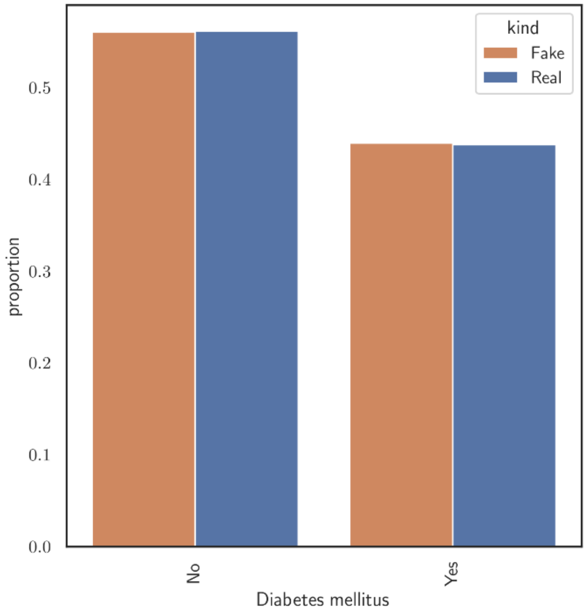


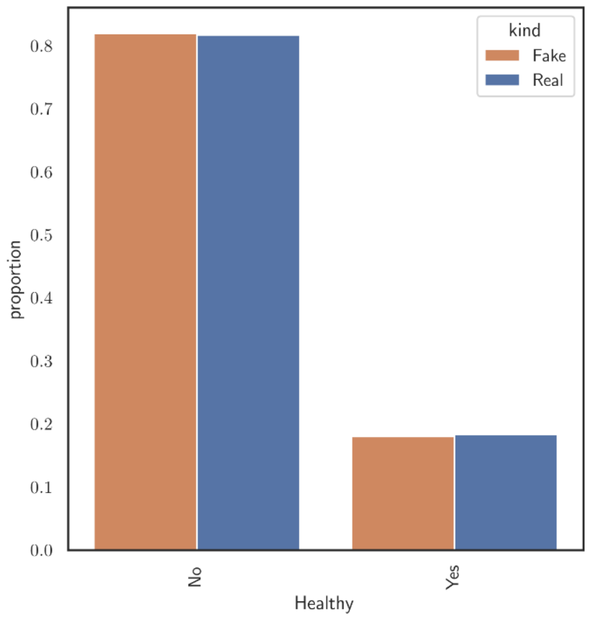
Group counts

Composite SDG



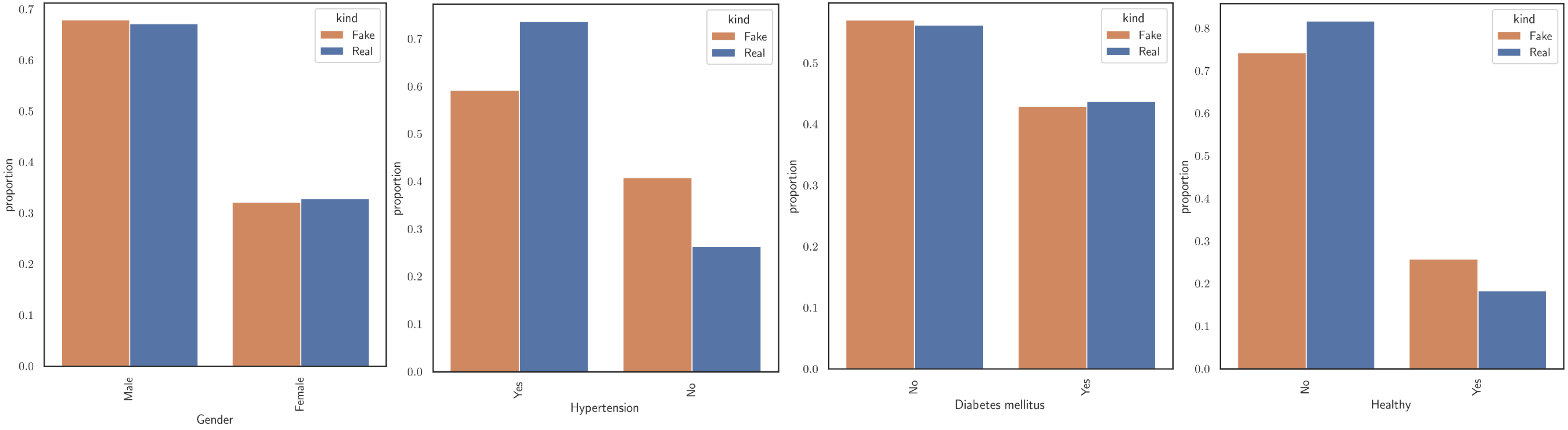






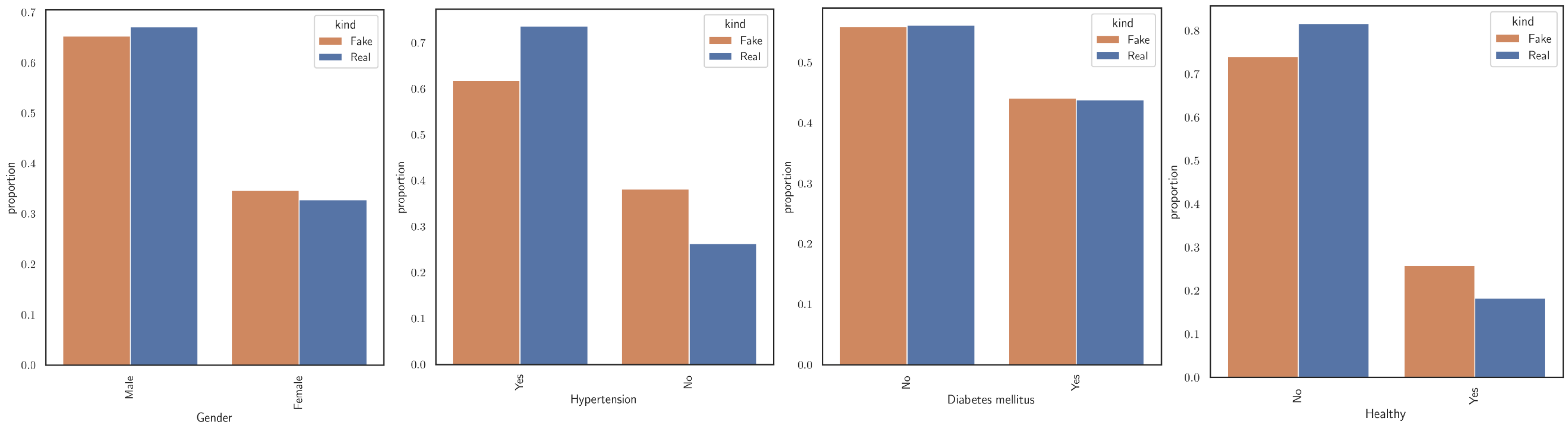
Group counts

CTGAN 50 sec



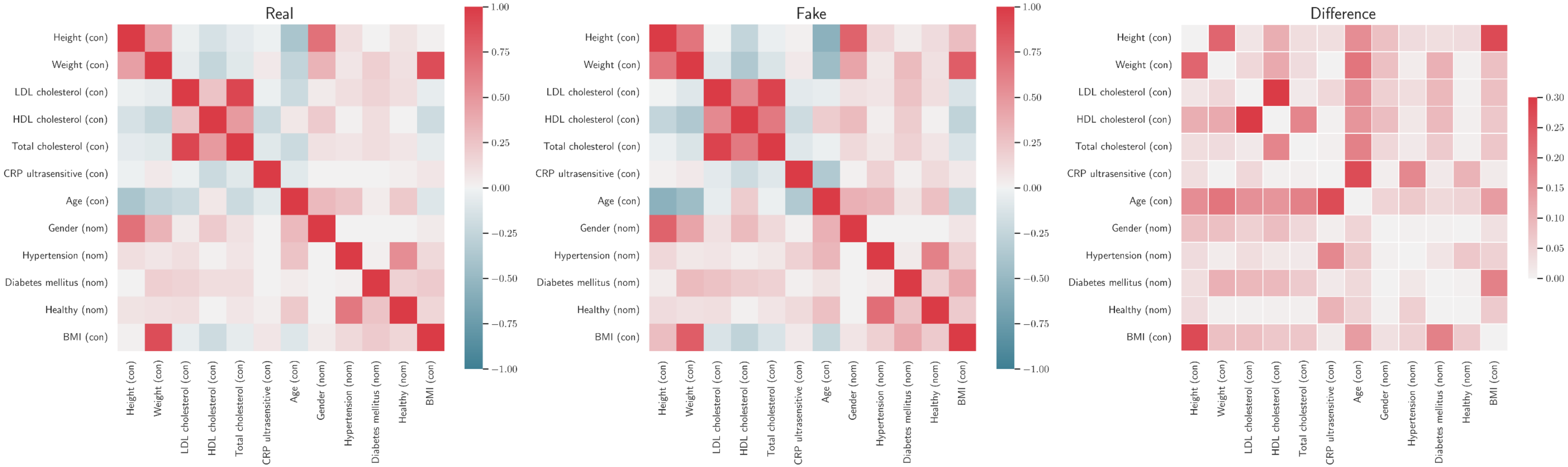
Group counts

CTGAN 30 min



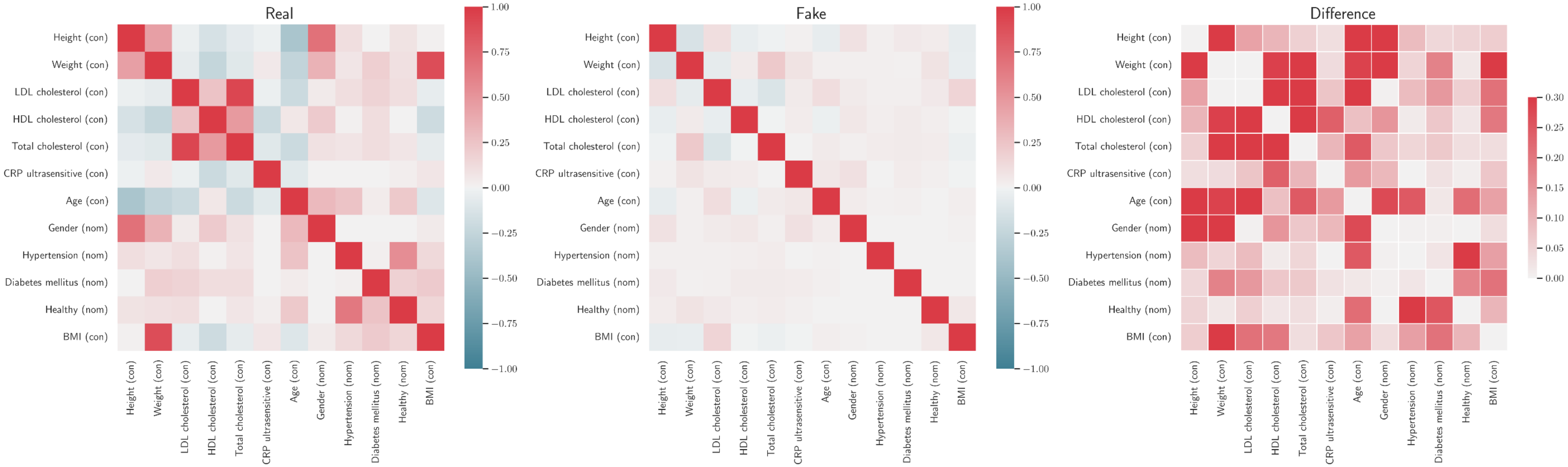
Correlations

Composite SDG



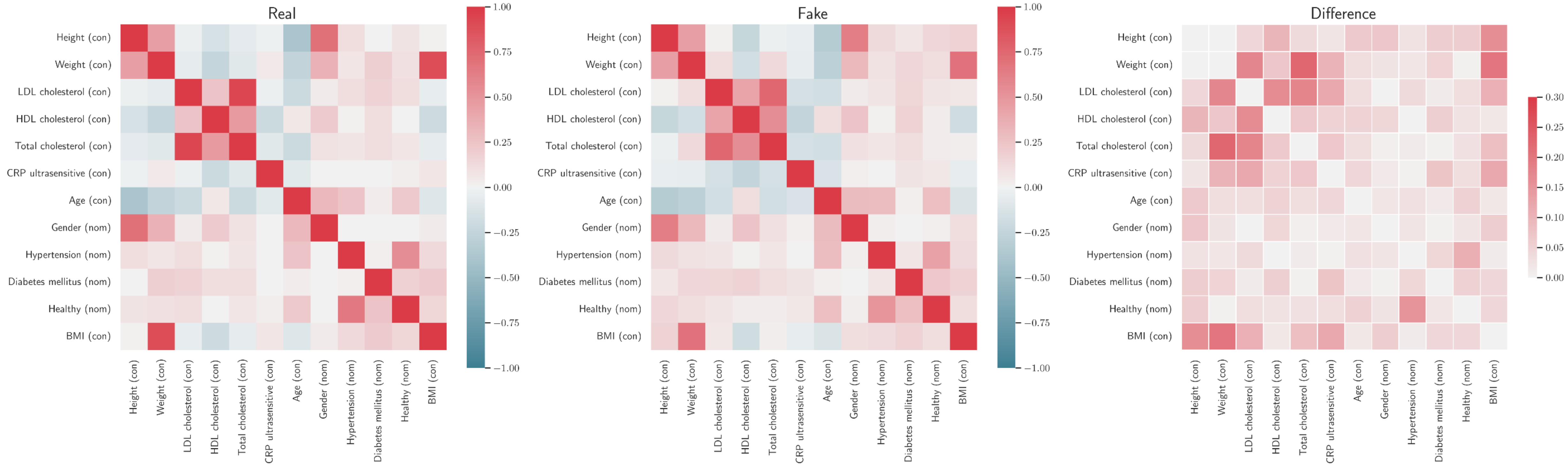
Correlations

CTGAN 50 sec

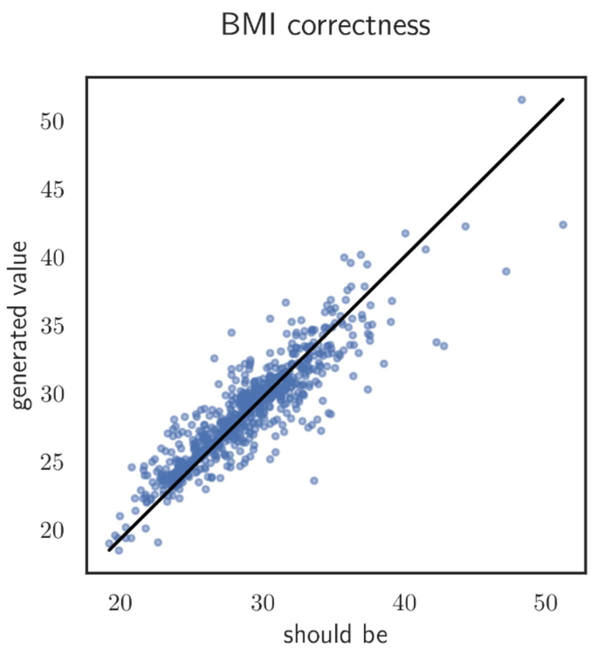


Correlations

CTGAN 30 min

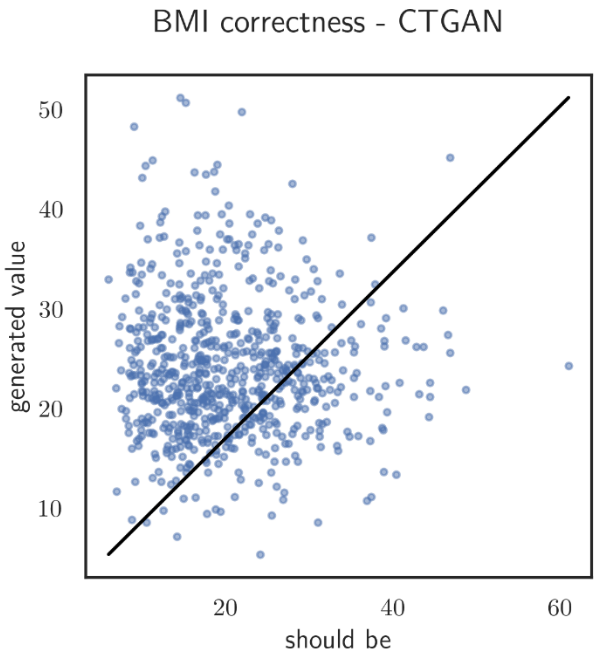


Reconstruction of dummy variables



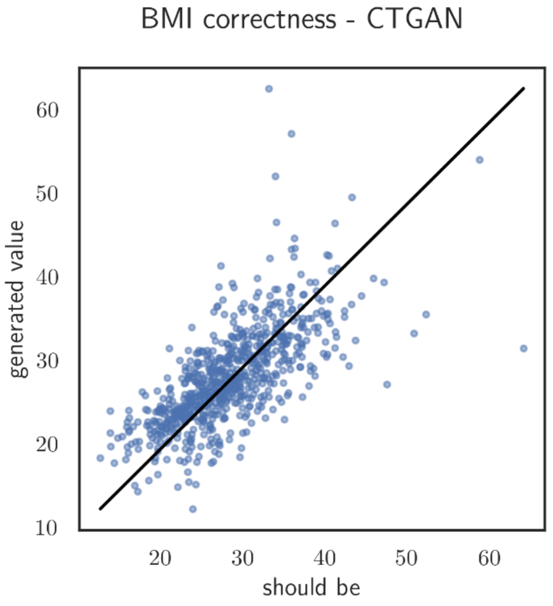
Healthy: 100 %

Composite SDG



Healthy: 61 %

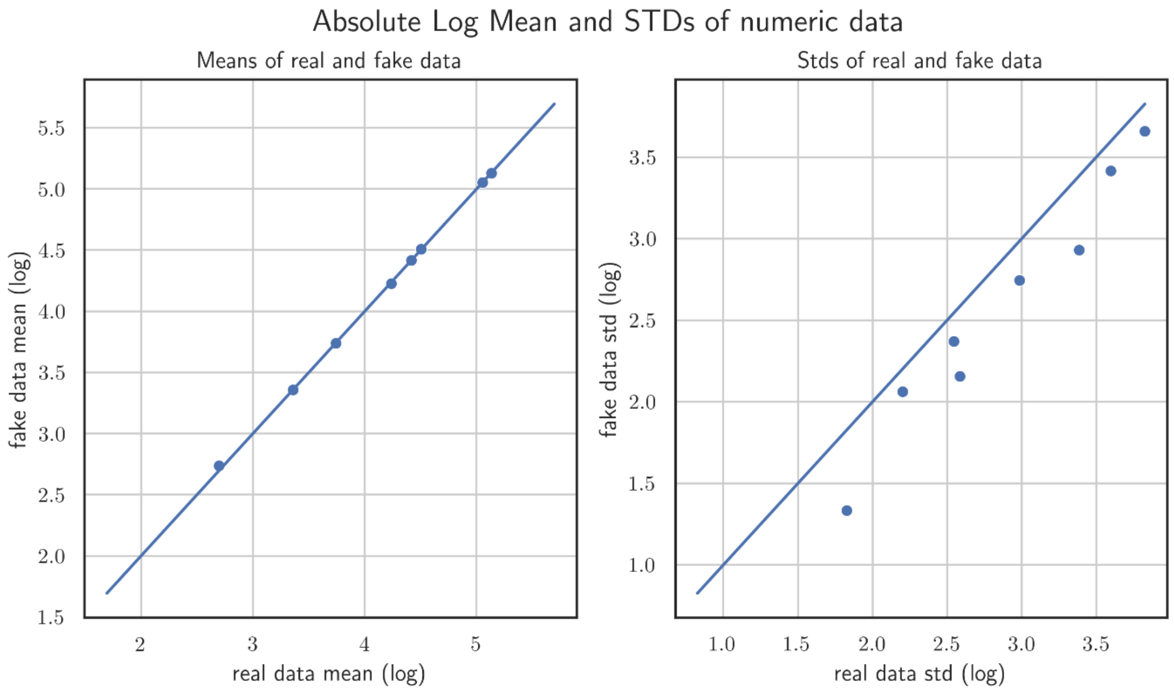
CTGAN 50 sec



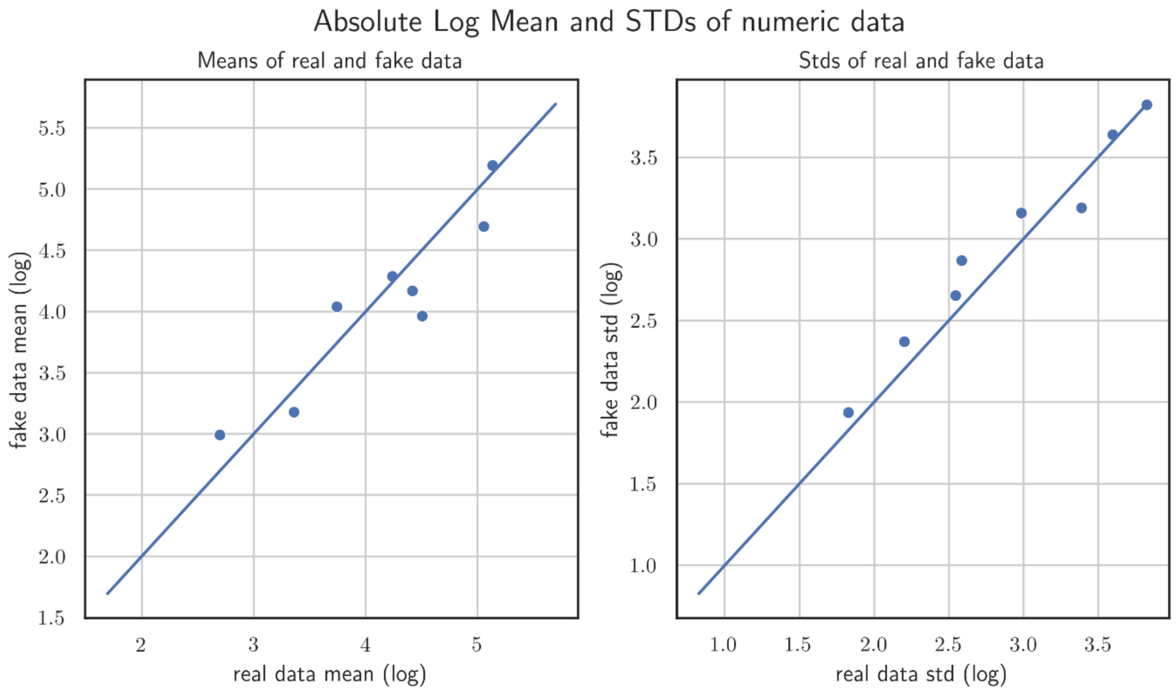
Healthy: 94 %

CTGAN 30 min

Bias and variance

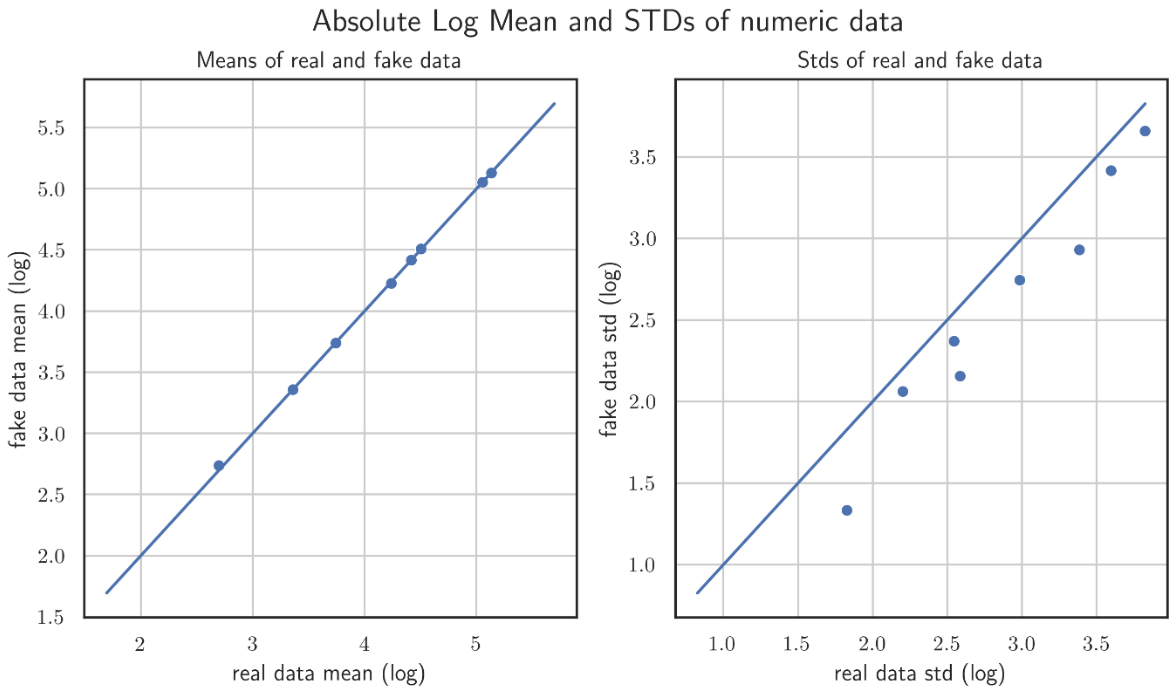


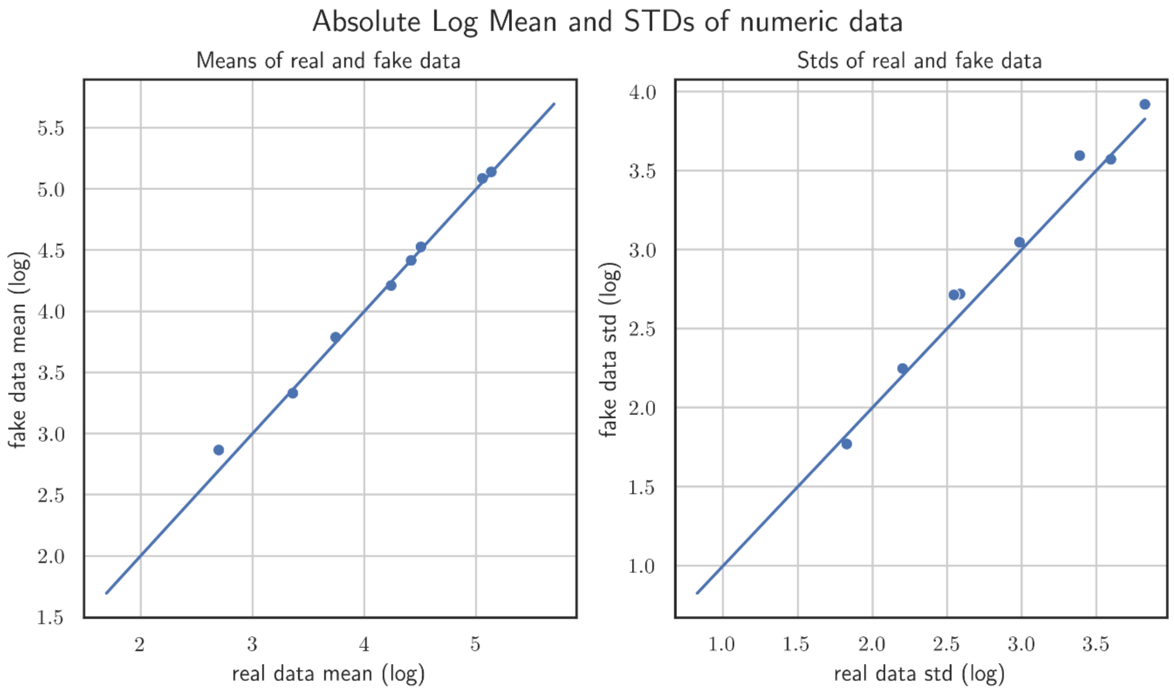
Composite SDG



CTGAN 50 sec

Bias and variance

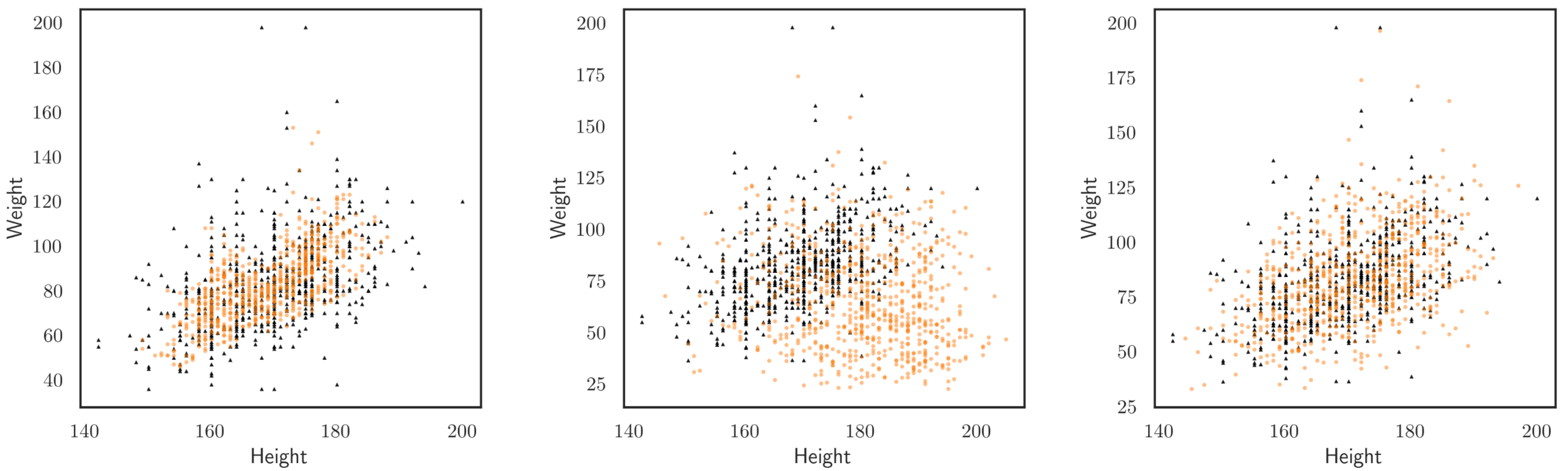




Composite SDG

CTGAN 30 min

Mode collapse?



Composite SDG

CTGAN 50 sec

CTGAN 30 min

Highlights

- A new generative model was developed using **only classical machine learning** methods
- It is **easier to customize** than solutions based on deep learning
- It better captures the statistical relationships between variables of mixed types
- Note: The case analysis presented **should be extended to verify** the above statements

REFERENCES

1. Alaa A, Van Breugel B, Saveliev ES, van der Schaar M. *How faithful is your synthetic data? Sample-level metrics for evaluating and auditing generative models*. In: International Conference on Machine Learning, PMLR. 2022. p. 290–306.
2. Borg I, Groenen PJ. *Modern multidimensional scaling: Theory and applications*. Springer Science & Business Media; 2005.
3. Borup D, Christensen BJ, Mühlbach NS, Nielsen MS. *Targeting predictors in random forest regression*. International Journal of Forecasting. 2023;39:841–68.
4. Brenninkmeijer B, de Vries A, Marchiori E, Hille Y. *On the generation and evaluation of tabular data using GANs* [dissertation]. 2019.
5. Cervantes J, Garcia-Lamont F, Rodríguez-Mazahua L, Lopez A. *A comprehensive survey on support vector machine classification: Applications, challenges, and trends*. Neurocomputing. 2020;408:189–215.
6. Dankar FK, Ibrahim MK, Ismail L. *A multi-dimensional evaluation of synthetic data generators*. IEEE Access. 2022;10:11147–58.
7. Deisenroth MP, Faisal AA, Ong CS. *Mathematics for machine learning*. Cambridge University Press; 2020.
8. Drapała J, Szczepanowski R, Świątek J, Uchmanowicz I, Czapla M, Biegus J, et al. *Two-stage approach to cluster categorical medical data*. In: International Conference On Systems Engineering. Springer; 2022. p. 178–86.
9. Drapała J. Composite SDG [Internet]. GitHub. Available from: https://github.com/jdrapala/CompositeSDG.git
10. El Emam K. *Seven ways to evaluate the utility of synthetic data*. IEEE Security & Privacy. 2020;18:56–9.

REFERENCES

11. Esteban C, Hyland SL, Rätsch G. *Real-valued (medical) time series generation with recurrent conditional GANs*. arXiv preprint arXiv:1706.02633. 2017.
12. Fonseca J, Bacao F. *Tabular and latent space synthetic data generation: A literature review*. Journal of Big Data. 2023;10:115.
13. Goncalves A, Ray P, Soper B, Stevens J, Coyle L, Sales AP. *Generation and evaluation of synthetic patient data*. BMC medical research methodology. 2020;20:1–40.
14. Hernandez M, Epelde G, Alberdi A, Cilla R, Rankin D. *Synthetic data generation for tabular health records: A systematic review*. Neurocomputing. 2022;493:28–45.
15. Kristan M, Leonardis A, Skočaj D. *Multivariate online kernel density estimation with Gaussian kernels*. Pattern recognition. 2011;44:2630–42.
16. Murtaza H, Ahmed M, Khan NF, Murtaza G, Zafar S, Bano A. *Synthetic data generation: State of the art in health care domain*. Computer Science Review. 2023;48:100546.
17. Park N, Mohammadi M, Gorde K, Jajodia S, Park H, Kim Y. *Data synthesis based on generative adversarial networks*. arXiv preprint arXiv:1806.03384. 2018.
18. Patki N, Wedge R, Veeramachaneni K. *The synthetic data vault*. In: 2016 IEEE international conference on data science and advanced analytics (DSAA). IEEE; 2016. p. 399–410.

REFERENCES

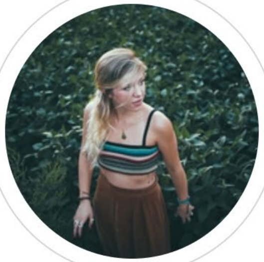
19. Romero-Corral A, Somers VK, Sierra-Johnson J, Korenfeld Y, Boarin S, Korinek J, et al. *Normal weight obesity: A risk factor for cardiometabolic dysregulation and cardiovascular mortality*. European heart journal. 2010;31:737–46.
20. Węglarczyk S. *Kernel density estimation and its application*. In: ITM Web of Conferences. EDP Sciences; 2018. p. 00037.
21. Xu L, Skoularidou M, Cuesta-Infante A, Veeramachaneni K. *Modeling tabular data using conditional GAN*. Advances in neural information processing systems. 2019;32.
22. Zhang Y, Zaidi N, Zhou J, Li G. *Interpretable tabular data generation*. Knowledge and Information Systems. 2023;65:2935–63.
23. Zhang Y, Zaidi NA, Zhou J, Li G. GANBLR: *A tabular data generation model*. In: 2021 IEEE International Conference on Data Mining (ICDM). IEEE; 2021. p. 181–90.

Other use cases

Fake Person Generator

Custom Generate

Gender: Random | Age: Random | State: Random | City: Random | Generate



Constance D Bowers

Gender: **female**

Race: **White**

Birthday: **12/19/1975** (**48** years old)

Street: **4559 Jerome Avenue**

City, State, Zip: **Edinburg, Texas(TX), 78539**

Telephone: **956-292-9208**

Mobile: **956-305-7370**

BASIC INFORMATION

Temporary Gmail(real)	i.nt.re.p.idnmw@gmail.com *This is a real Gmail. Click here to receive emails.*
Email(fake)	sedrick19710@gmail.com
Height	5' 7" (170 centimeters)
Weight	135.1 pounds (61.28 kilograms)
Hair Color	Brown
Blood Type	A+

ONLINE PROFILE

Login Times	95 times
On-line Time	21755 seconds
Points	295 (0-10,000 points)
Level	2 (1-10)
Number of Comments	32 comments posted
Posted Articles	30 articles posted
Friends	24 friends