Semantic-Driven Data Augmentation for Improved Machine Learning Predictions (Extended Abstract)

Majlinda Llugiqi[1,*], Fajar J. Ekaputra[1,2] and Marta Sabou[1]

[1] Vienna University of Economics and Business, Vienna, Austria
[2] TU Wien, Vienna, Austria

Abstract

Machine learning (ML) models frequently struggle in domains where labeled data is limited or sensitive. To address this challenge, we explore a semantic-driven data augmentation approach that incorporates external knowledge into tabular datasets. Our method leverages neuro-symbolic techniques to enrich training data with structured context derived from knowledge graphs (KGs), aiming to enhance the predictive capabilities of standard ML algorithms. We evaluate multiple approaches for integrating KG information into ML pipelines and examine their impact on model performance across binary classification tasks involving medical datasets such as heart disease and chronic kidney disease. The experimental setup includes four ML models and four distinct KG embedding algorithms, with performance evaluated using accuracy and F2 score. Results show that augmenting tabular features with semantic distance metrics from KG embeddings yields notable improvements. For instance, XGBoost achieves a significant F2 score increase from 75.19% to 90.85% in heart disease prediction. These results suggest that semantic augmentation of tabular datasets has the potential to enhance ML prediction[1].

Introduction Machine learning (ML) has achieved remarkable success across domains such as computer vision [2, 3] and language processing [4, 5], largely driven by the availability of large-scale datasets. However, in domains such as healthcare, where data is often scarce or protected by privacy regulations, ML models often encounter performance limitations [6, 7].

To address these issues, our recent work [1] proposed integrating structured domain knowledge into ML pipelines through the use of knowledge graphs (KGs). This approach combines structured knowledge with data-driven learning by embedding KGs into vector representations to enrich tabular datasets. We evaluate various techniques for incorporating KG embeddings into binary classification tasks for heart and chronic kidney disease prediction. We investigate how different embedding strategies and ML models interact, focusing on accuracy and F2 scores. Our findings demonstrate that semantic augmentation, particularly through distance-based features derived from KG embeddings, can significantly enhance predictive performance, highlighting the potential of knowledge-infused learning in data-scarce environments.

[1] This work is based on a full article accepted at Neurosymbolic AI [1]

The Second Workshop on Knowledge Graphs and Neurosymbolic AI (KG-NeSy), co-located with SEMANTiCS'25: International Conference on Semantic Systems, September 3–5, 2025, Vienna, Austria

[*] Corresponding author.

majlinda.llugiqi@wu.ac.at (M. Llugiqi); fajar.ekaputra@wu.ac.at (F. J. Ekaputra); marta.sabou@wu.ac.at (M. Sabou)
0000-0002-5008-6856 (M. Llugiqi); 0000-0003-4569-2496 (F. J. Ekaputra); 0000-0001-9301-8418 (M. Sabou)

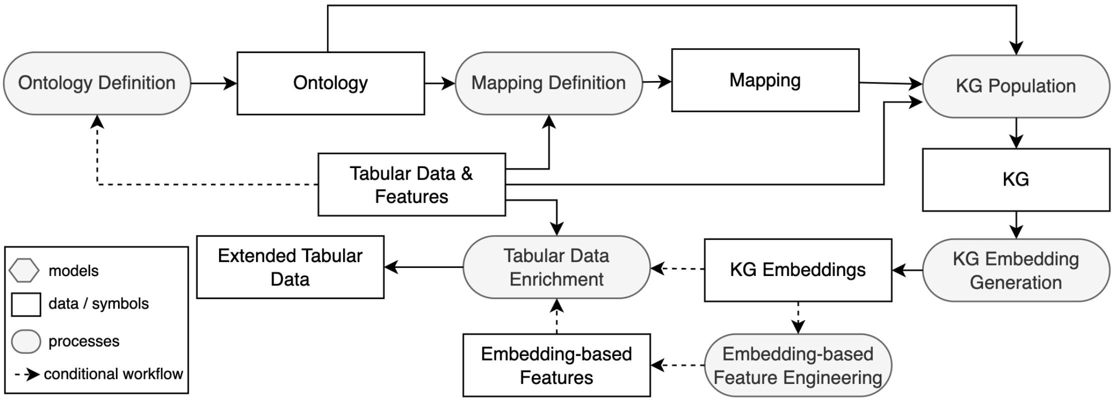

Figure 1: Overview of the proposed methodology for enriching tabular data with KGE. (adapted from [1] following the boxology notation [8]).

Methodology To enrich tabular data with KG information, we propose a pipeline that integrates KG embeddings into ML models (see Figure 1). The process begins by defining domain ontologies to represent tabular dataset's features and constructing KGs using mappings from tabular features to ontology concepts and relations. These KGs are then embedded into vector spaces using different KG embedding methods. We explore several methods for computing information in the embedding space which is then used to augment tabular data with KG-derived features to enhance ML performance, as illustrated in Figure 2. These strategies differ in how they leverage the semantic information encoded in KG embeddings. One method, EmbedOnly, uses only the embeddings to assess whether the learned semantic structure can replace raw features. In EmbedAugTab, embeddings are combined with the original tabular features to enrich the data with latent relational information. The DistAugTab approach introduces distance-based features: for each instance, we calculate the distance from its embedding to class centroids, capturing proximity-based semantics. This is extended in EmbedDistTabAug, which integrates both the raw embeddings and their distance-based metrics into the dataset. To capture higher-order semantics, ClusterAugTab applies clustering in the embedding space, assigning each instance a cluster membership to reflect latent groupings. EmbedClusterAugTab builds on this by combining cluster memberships with the full embeddings for even more expressive augmentation. Finally, we investigate feature interaction strategies. In InteraAugTab and EmbedInteraAugTab, we compute element-wise interactions between original features and embeddings, enabling the model to learn complex joint effects between raw clinical measurements and semantic knowledge from the KG.

Our experiments apply these strategies to binary classification tasks in the healthcare domain (heart disease and chronic kidney disease), using four ML models (KNN, SVM, XGBoost and a feedforward neural network (NN)). Performance is evaluated using accuracy and F2 score.

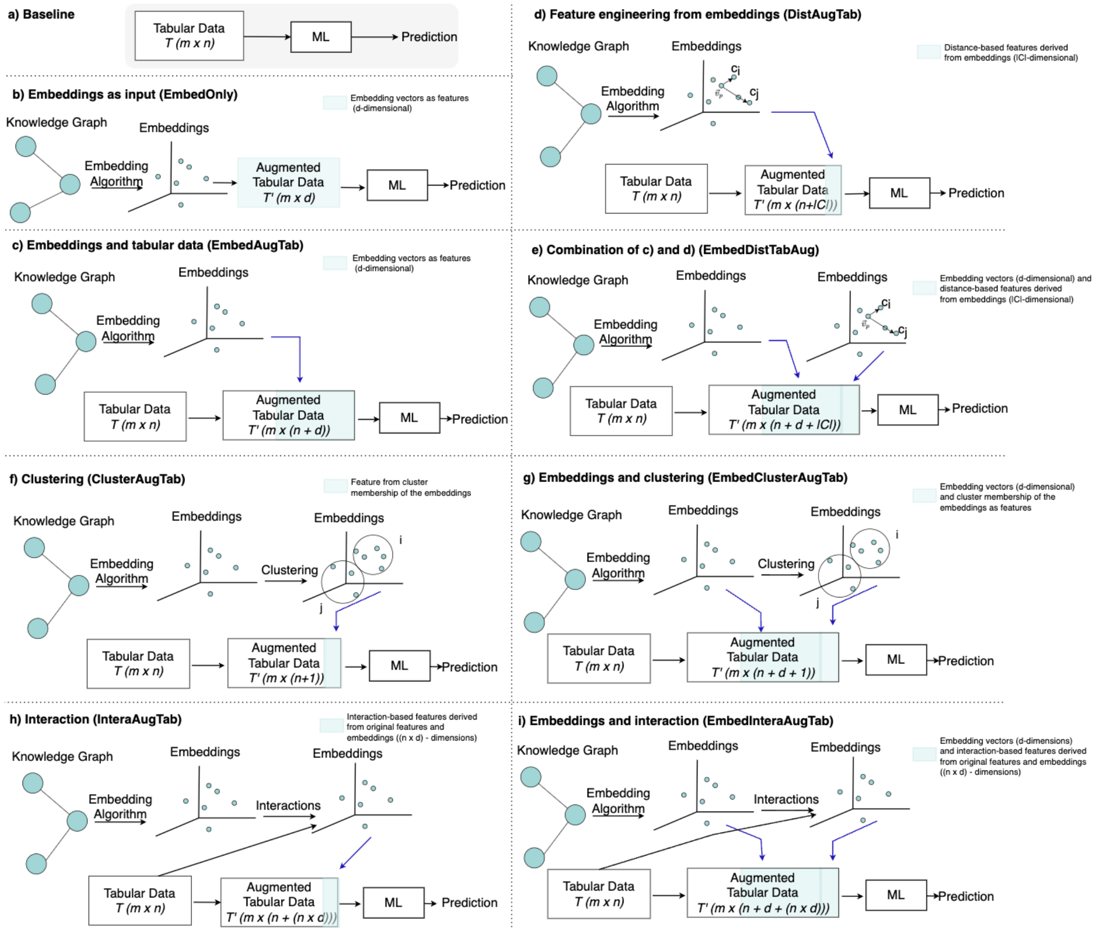

Figure 2: Overview of semantic augmentation approaches to enrich tabular data with KG embeddings for improved ML performance (adapted from [1]).

Results Our evaluation covered eight KG-based data augmentation strategies applied to binary classification tasks in heart disease and chronic kidney disease prediction. In Table 1,[1] we present the average accuracy and F2 scores across different embedding methods, classifiers, and augmentation strategies for heart disease prediction. Similar trends were observed in the kidney disease experiments, where integrating KG-based information also led to improvements in predictive performance. Among the evaluated strategies, those that incorporated distance-based features from the embedding space (e.g., DistAugTab) outperformed other approaches.

Among the embedding methods, RDF2Vec emerged as the most effective overall, yielding stable and strong results across both tasks and models, likely due to its ability to capture semantic paths in the KG. Node2Vec also performed well in scenarios where local graph structure was more informative, e.g., improving KNN in kidney disease and XGBoost in heart disease prediction. Notably, XGBoost, despite achieving the highest F2 score in some configurations, showed

[1] Due to space limitations, we report results only for heart disease prediction in this extended abstract. Full results, including those for chronic kidney disease and additional scenarios, are available in [1].

Table 1
Average accuracy and F2 scores, accross various approaches and models for heart disease prediction.

	KNN		NN		SVM		XGBoost	
Methods	Acc.	F2	Acc.	F2	Acc.	F2	Acc.	F2
Baseline	81.02	71.33	81.77	77.44	79.75	77.18	79.32	75.19
Node2Vec								
DistAugTab	81.17	71.54	**82.17**	**78.78**	**81.81**	**78.36**	**92.51**	**90.85**
EmbedDistAugTab	**81.43**	**71.70**	77.71	76.10	**81.67**	**78.57**	**91.82**	**89.27**
EmbedDistAugTabRed	80.66	70.76	80.06	75.74	79.46	77.71	79.32	75.15
EmbedClustAugTab	81.17	70.97	72.43	72.96	76.10	75.01	55.32	57.39
ClustAugTab	81.21	71.16	78.01	76.03	77.58	76.02	62.93	65.05
RDF2Vec								
DistAugTab	81.02	71.33	81.96	78.57	79.75	77.18	**84.38**	**81.62**
EmbedDistAugTab	81.02	71.33	81.85	78.46	79.75	77.18	80.60	77.20
EmbedDistAugTabRed	79.95	69.59	80.49	76.45	79.32	77.63	78.77	75.25
EmbedClustAugTab	**81.18**	71.12	81.44	77.48	**80.16**	**77.36**	78.64	75.34
ClustAugTab	**81.18**	71.12	81.81	78.38	**80.16**	**77.36**	79.10	75.22
DistMult								
DistAugTab	80.88	71.04	**82.18**	**78.90**	**80.27**	77.39	53.42	54.84
EmbedDistAugTab	80.94	71.14	80.57	78.55	80.11	77.56	50.49	61.31
EmbedDistAugTabRed	80.16	70.02	81.30	77.69	79.34	77.71	78.16	73.92
EmbedClustAugTab	**81.39**	71.32	72.17	72.09	75.31	73.72	50.12	55.90
ClustAugTab	**81.43**	**71.45**	76.78	75.76	76.17	74.26	59.35	62.11
TransH								
DistAugTab	80.98	71.27	**81.99**	**78.55**	**80.10**	77.30	75.34	73.77
EmbedDistAugTab	80.95	71.19	80.89	77.97	**80.11**	77.50	55.48	57.76
EmbedDistAugTabRed	80.08	69.95	80.72	76.61	79.38	**77.78**	78.20	74.17
EmbedClustAugTab	80.76	70.42	76.68	74.95	78.32	74.42	48.52	50.38
ClustAugTab	80.82	70.55	80.24	75.94	78.52	74.52	60.35	56.58

variability across embedding strategies, highlighting the importance of matching embedding methods with suitable models.

These findings support the hypothesis that semantic enrichment, particularly through embedding-derived features, can enhance ML performance on tabular data, especially in domains with limited or sensitive datasets.

Conclusion This work shows that semantically enriching tabular data using KG embeddings can enhance the predictive performance of ML models, particularly in low-data medical scenarios. Integrating structured domain knowledge through augmentation strategies, especially distance-based features, led to notable F2 score gains. While promising for context-aware predictions in sensitive domains, challenges remain in generalizability. Future work will explore the effectiveness of KGs across diverse, data-scarce domains to address ML models' data dependency.

Acknowledgments This work was supported by the FFG SENSE (894802) and FAIR-AI (904624) projects, as well as by the Austrian Science Fund (FWF) BILAI 10.55776/COE12 and HOnEst (V 745-N) projects. For open access purposes, the author has applied a CC BY public copyright license to any author accepted manuscript version arising from this submission.

Declaration on Generative AI During the preparation of this work, the author(s) used GPT-4o in order to brainstorm ideas about the title. After using these tool, the authors reviewed and edited the content as needed and take full responsibility for the publication's content.

References

[1] M. Llugiqi, F. J. Ekaputra, M. Sabou, Semantic-based data augmentation for machine learning prediction enhancement, Neurosymbolic Artificial Intelligence 1 (2025) 29498732251340160.

[2] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, L. Fei-Fei, Imagenet: A large-scale hierarchical image database, in: 2009 IEEE conference on computer vision and pattern recognition, Ieee, 2009, pp. 248–255.

[3] M. Wortsman, G. Ilharco, S. Y. Gadre, R. Roelofs, R. Gontijo-Lopes, A. S. Morcos, H. Namkoong, A. Farhadi, Y. Carmon, S. Kornblith, et al., Model soups: averaging weights of multiple fine-tuned models improves accuracy without increasing inference time, in: International conference on machine learning, PMLR, 2022, pp. 23965–23998.

[4] J. D. M.-W. C. Kenton, L. K. Toutanova, Bert: Pre-training of deep bidirectional transformers for language understanding, in: Proceedings of naacL-HLT, volume 1, Minneapolis, Minnesota, 2019, p. 2.

[5] A. Ramesh, P. Dhariwal, A. Nichol, C. Chu, M. Chen, Hierarchical text-conditional image generation with clip latents, arXiv preprint arXiv:2204.06125 1 (2022) 3.

[6] K. Poulinakis, D. Drikakis, I. W. Kokkinakis, S. M. Spottswood, Machine-learning methods on noisy and sparse data, Mathematics 11 (2023) 236.

[7] D. Jarrett, E. Stride, K. Vallis, M. J. Gooding, Applications and limitations of machine learning in radiation oncology, The British journal of radiology 92 (2019) 20190001.

[8] M. van Bekkum, M. de Boer, F. van Harmelen, A. Meyer-Vitali, A. ten Teije, Modular design patterns for hybrid learning and reasoning systems, Appl. Intell. 51 (2021) 6528–6546. doi:10.1007/S10489-021-02394-3.