Leveraging Knowledge Graphs for Enhancing Machine Learning-based Heart Disease Prediction

WIRTSCHAFTS
UNIVERSITÄT
WIEN VIENNA
UNIVERSITY OF
ECONOMICS
AND BUSINESS

Majlinda Llugiqi, Fajar J. Ekaputra, Marta Sabou

WU Vienna

EQUIS ACCREDITED AACSB ACCREDITED ASSOCIATION OF MBAS AMBA ACCREDITED

27.03.2024

Outline

WIRTSCHAFTS UNIVERSITÄT WIEN VIENNA UNIVERSITY OF ECONOMICS AND BUSINESS

- Introduction and motivation
- Knowledge graph construction
- Infusing KG into ML pipeline
- Results
- Conclusion and future work

Introduction & Motivation

- ML algorithms widely used across various domains for predictive tasks.
- However the effectiveness of ML models is constrained by the scarcity of annotated datasets needed for training accurate models.
- The challenge of limited annotated datasets for training ML models is particularly critical in domains where accuracy is crucial, such as medical diagnosis.

Knowledge Graphs

- Use of KGs to enrich the data - enhance the performance.
- KGs provide a structured representation of domain-specific knowledge.
- By utilizing existing ontologies and KGs, we can infuse our datasets with rich, contextual information that goes beyond raw data.

Preprocessing Knowledge Graphs

- Heart disease domain:

  age, chest pain type, resting blood pressure, heart rate…

- Ontologies that describe the dataset’s features

- Three different ontologies in the heart disease domain:
  - *Small* ontology - existing ontology from Trepan Reloaded [1]
  - *Extended* ontology - extended HFO ontology [2]
  - *SNOMED* ontology - extracted from SNOMED [3]

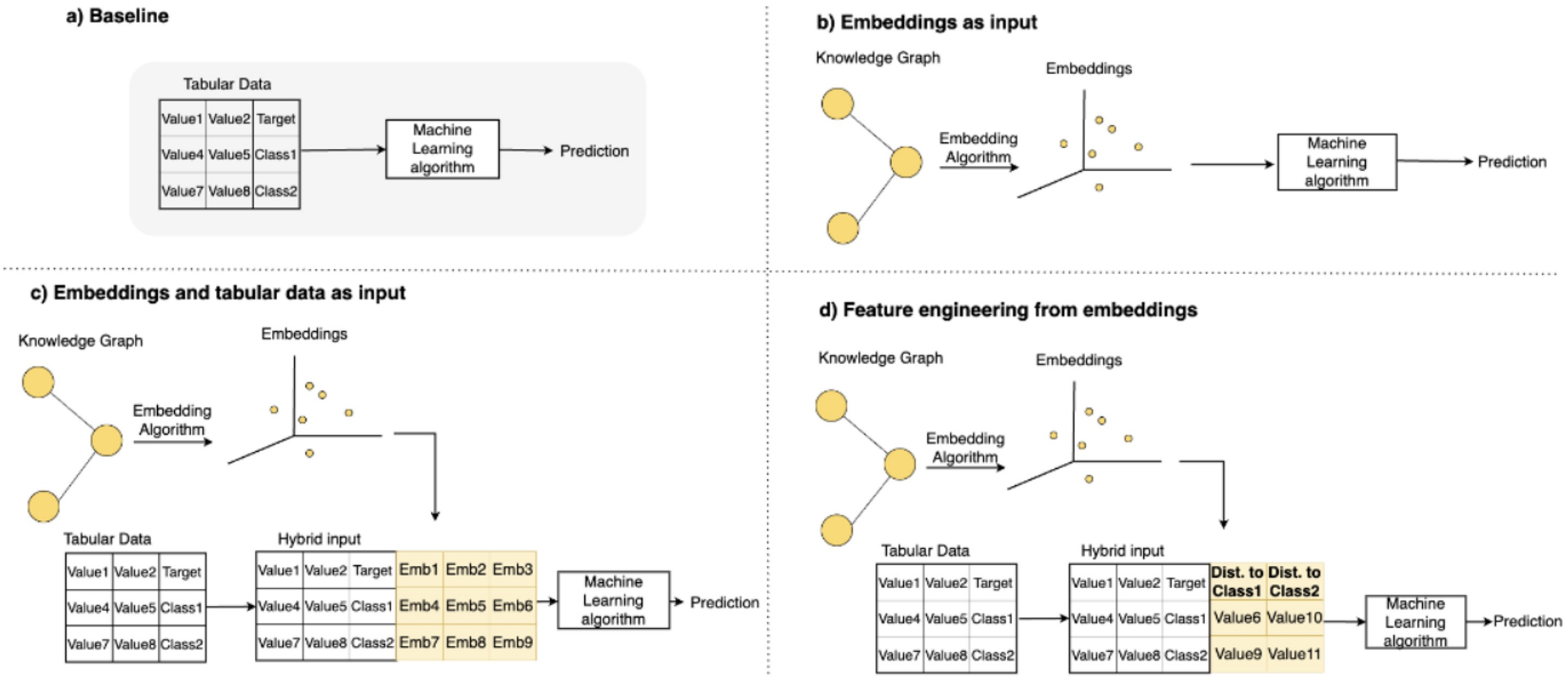
Preprocessing Knowledge Graphs

- Mapped the features of the dataset to the concepts/relations
- KGs construction - population of the ontologies with dataset instances

Table 1: Details of the KGs for heart disease domain.

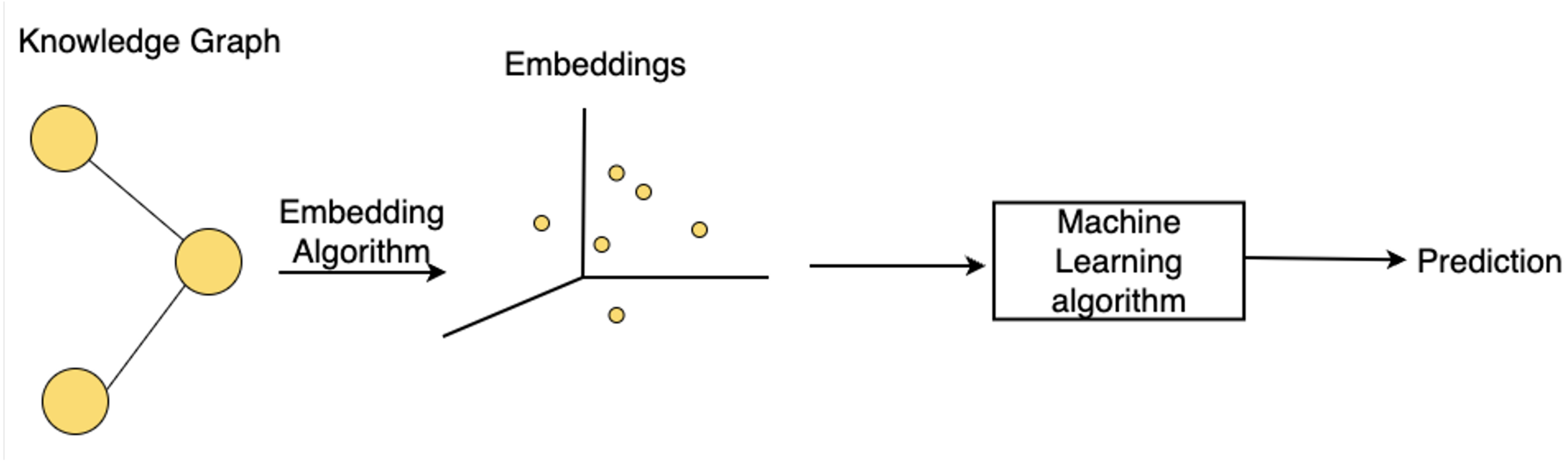
KG	Logical Axioms	Classes	Object prop.	Data prop.
Small	4637	29	6	10
Extended	6682	1664	6	10
Snomed	1963	80	24	10

Knowledge Graphs infusion into ML pipeline



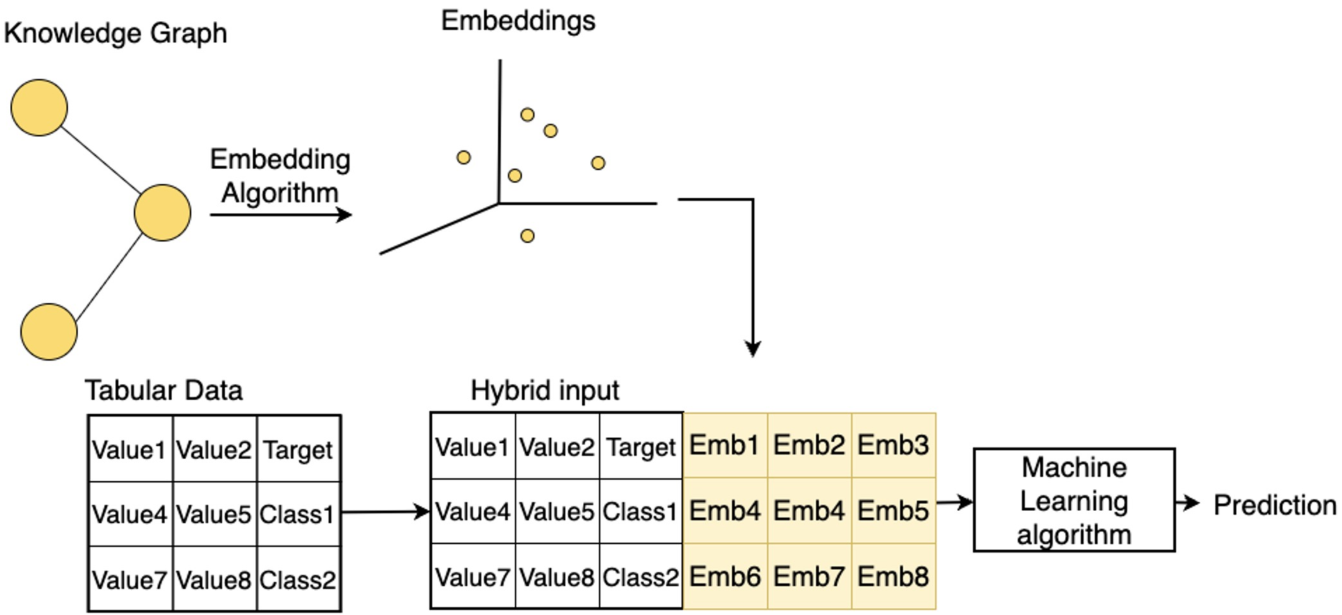
Embedding as input

- Training ML learning with embeddings from KG only.
- Embeddings represent patients in vector space



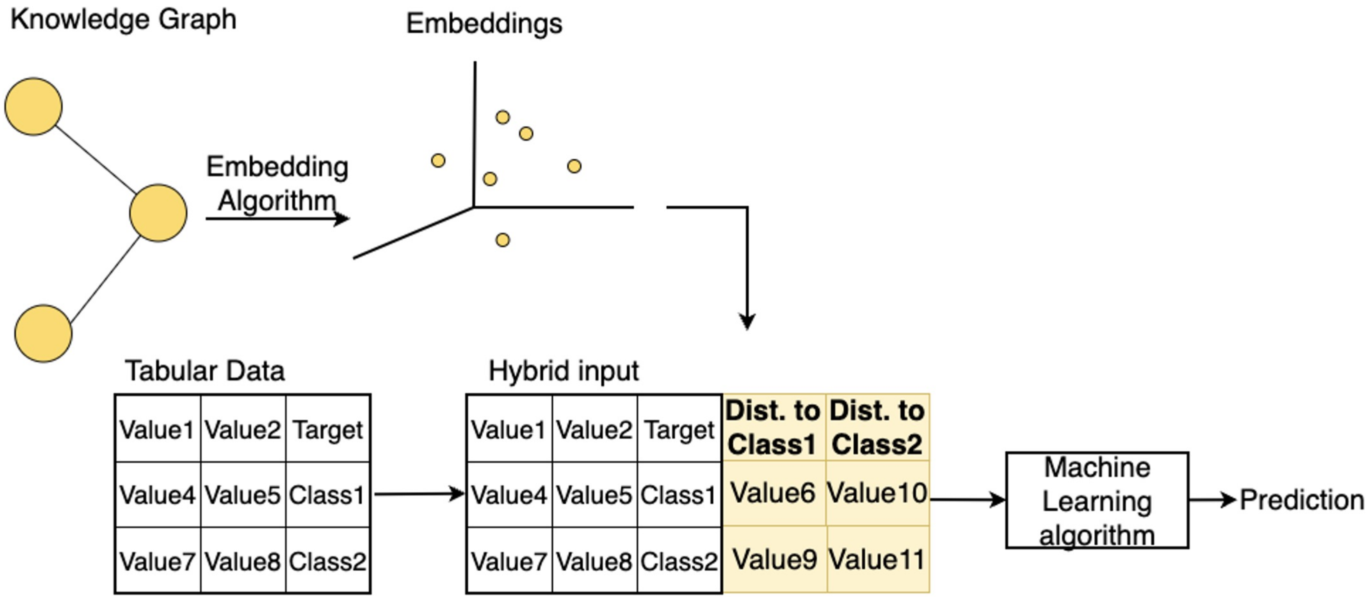
Enriching tabular data with embeddings

- For each patient, embedding vectors are added as extra columns.



Feature Engineering from embeddings

- For each patient in the embedding space, their Euclidean distance to ‘disease’ and ‘no disease’ class is added.



Experiment Setup

- Dataset: Heart disease prediction from Kaggle (303 patients)
- ML models:
  - KNN, SVM, XGBoost, FFNN
- Metrics used :
  - Accuracy, F2 Score
- Embedding algorithms:
  - RDF2Vec, Node2vec

age	sex	cp	trestbps	chol	fbs	restecg	thalach	exang	oldpeak	slope	ca	thal	label
63.0	Male	TypicalAngina	145.0	233.0	Yes	LeftVentricularHypertrophy	150.0	No	2.3	Downsloping	Zero	FixedEffect	0
67.0	Male	Asymptomatic	160.0	286.0	No	LeftVentricularHypertrophy	108.0	Yes	1.5	Flat	Three	Normal	1
67.0	Male	Asymptomatic	120.0	229.0	No	LeftVentricularHypertrophy	129.0	Yes	2.6	Flat	Two	ReversableEffect	1
37.0	Male	NonAnginalPain	130.0	250.0	No	Normal	187.0	No	3.5	Downsloping	Zero	Normal	0
41.0	Female	AtypicalAngina	130.0	204.0	No	LeftVentricularHypertrophy	172.0	No	1.4	Upsloping	Zero	Normal	0
56.0	Male	AtypicalAngina	120.0	236.0	No	Normal	178.0	No	0.8	Upsloping	Zero	Normal	0
62.0	Female	Asymptomatic	140.0	268.0	No	LeftVentricularHypertrophy	160.0	No	3.6	Downsloping	Two	Normal	1

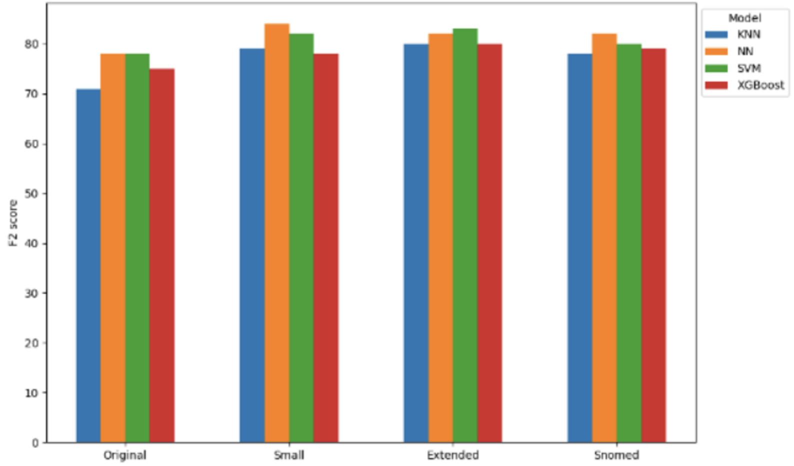
Results

Table 3: Comparison of Accuracy and F2 Scores Across Models using various KG Inputs.

Model	Original Acc.	Original F2	Rdf2Vec Acc.	Rdf2Vec F2	Node2Vec Acc.	Node2Vec F2	Comb-R2V Acc.	Comb-R2V F2	Comb-N2V Acc.	Comb-N2V F2	FE-R2V Acc.	FE-R2V F2	FE-N2V Acc.	FE-N2V F2
Small KG														
KNN	0.81	0.71	0.51	0.34	0.72	0.59	0.81	0.71	0.81	0.71	0.65	0.53	**0.83**	**0.79**
NN	0.82	0.78	0.53	0.04	0.81	0.79	0.82	0.78	0.82	0.77	0.73	0.69	**0.85**	**0.84**
SVM	0.82	0.78	0.54	0.00	0.81	0.78	0.82	0.78	0.83	0.81	0.74	0.65	**0.84**	**0.82**
XGB	0.79	0.75	0.50	0.40	0.73	0.67	0.80	0.75	0.81	0.75	0.65	0.57	**0.80**	**0.78**
Snomed KG														
KNN	0.81	0.71	0.54	0.34	0.76	0.66	0.81	0.71	0.81	0.72	0.65	0.53	**0.83**	**0.78**
NN	0.82	0.78	0.57	0.29	0.79	0.75	0.82	0.78	0.82	0.77	0.70	0.65	**0.84**	**0.82**
SVM	0.82	0.78	0.54	0.00	0.80	0.75	0.82	0.78	0.82	**0.80**	0.69	0.62	**0.83**	**0.80**
XGB	0.79	0.75	0.58	0.48	0.76	0.70	**0.82**	0.77	**0.81**	0.77	0.64	0.58	**0.81**	**0.79**
Extended KG														
KNN	0.81	0.71	0.53	0.37	0.80	0.72	0.81	0.71	0.81	0.72	0.52	0.24	**0.84**	**0.80**
NN	0.82	0.78	0.54	0.05	0.80	0.78	0.82	0.79	0.83	0.80	0.55	0.26	**0.84**	**0.82**
SVM	0.82	0.78	0.54	0.00	0.79	0.76	0.82	0.78	0.83	0.80	0.56	0.22	**0.84**	**0.83**
XGB	0.79	0.75	0.53	0.43	0.77	0.73	0.82	0.77	0.81	0.77	0.54	0.41	**0.82**	**0.80**

Results - Impact of KG Size and Structure

- Different algorithms favor different KG characteristics
- NN best performance - Small KG
- KNN and XGB best performance - Extended KG



Conclusion

WIRTSCHAFTS UNIVERSITÄT WIEN VIENNA UNIVERSITY OF ECONOMICS AND BUSINESS

- Using RDF2Vec and Node2Vec embeddings from KGs improves the accuracy and F2 scores - especially when distances to the classes are added as additional features
- The performance of ML algorithms is affected by the size and structure of KGs, with different algorithms favoring different KG characteristics.
- Adding KG information to ML algorithms enhances performance across all models without altering their inherent performance hierarchy.

Future work

WIRTSCHAFTS UNIVERSITÄT WIEN VIENNA UNIVERSITY OF ECONOMICS AND BUSINESS

- Apply this approach in various domains beyond heart disease.
- Investigate different embedding algorithms and use different ML algorithms.
- Measure data-dependency of ML algorithms and compare this with the complementary contributions from KGs.

Questions?

Majlinda Llugiqi
WU Vienna
majlinda.llugiqi@wu.ac.at