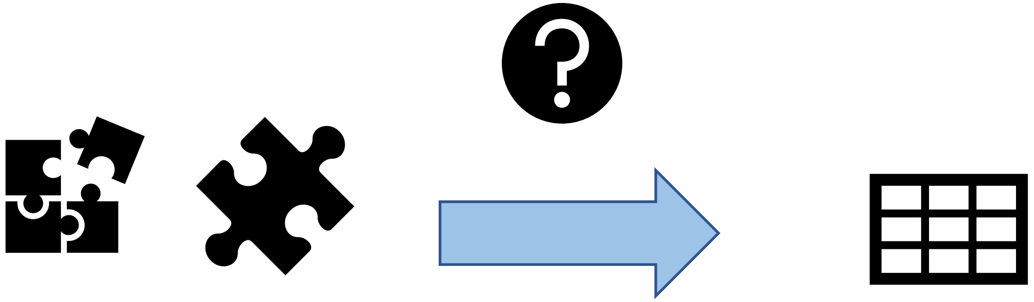
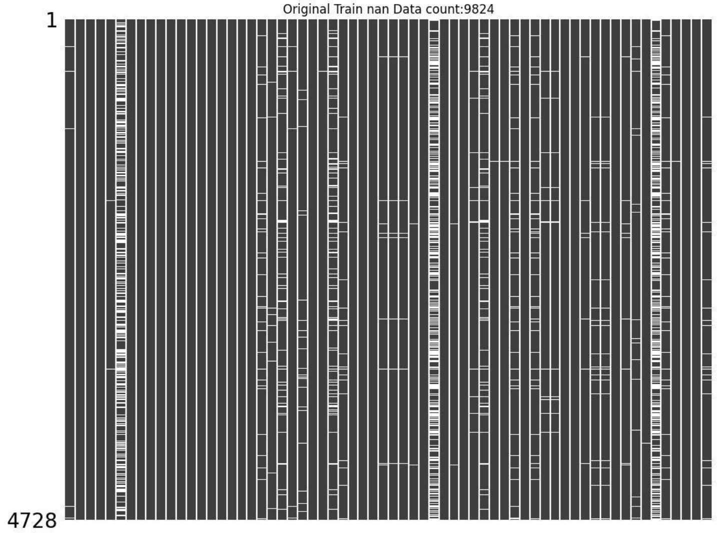
Drexel University Senior Design 2022-2023

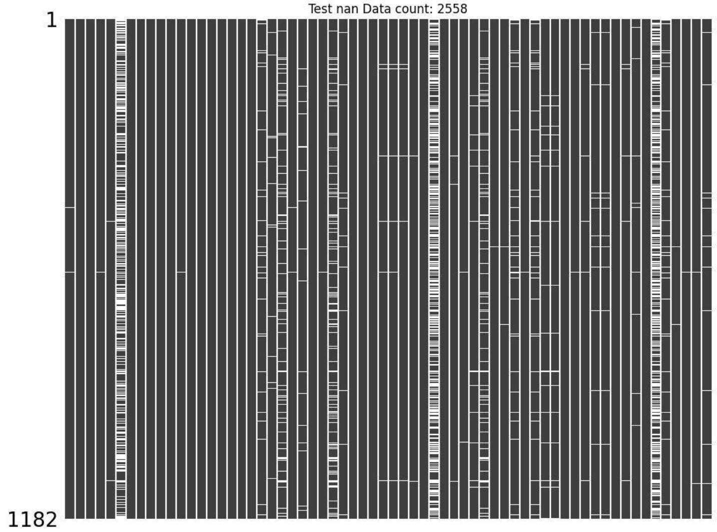
VEX

By Xiyuan Chang
B.S. Data Science.
Advisor: Hegler Tissot

Challenges in standard datasets




Original Train nan Data count:9824


Test nan Data count: 2558

One-hot Encoding

Wine dataset Categorical feature:
Color: [**White**, Pink, Green, Purple, **Red**, Yellow] → Wine's **color** is **Red**/**White**/….

One-hot encoding

Color_White	Color_Pink	Color_Green	Color_Purple	Color_Red	Color_Yellow
0	0	0	0	1	0
1	0	0	0	0	0
0	0	1	0	0	0
0	0	0	1	0	0

Low-Dimensional Embedding Representation

- avoid expanding the datasets

- keep the information in the original dataset as much as possible

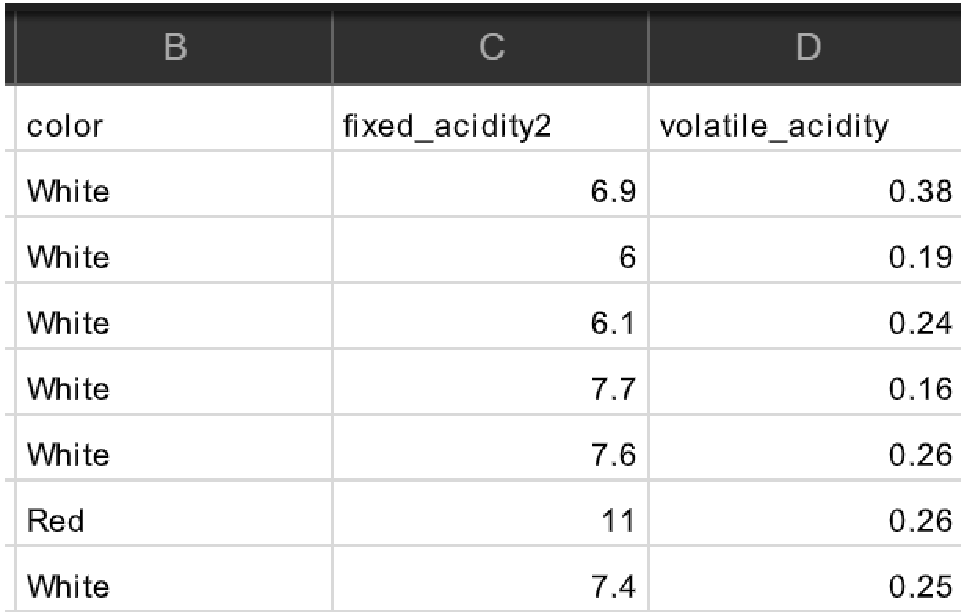
- convey the meanings directly

- capture the relations inside the dataset

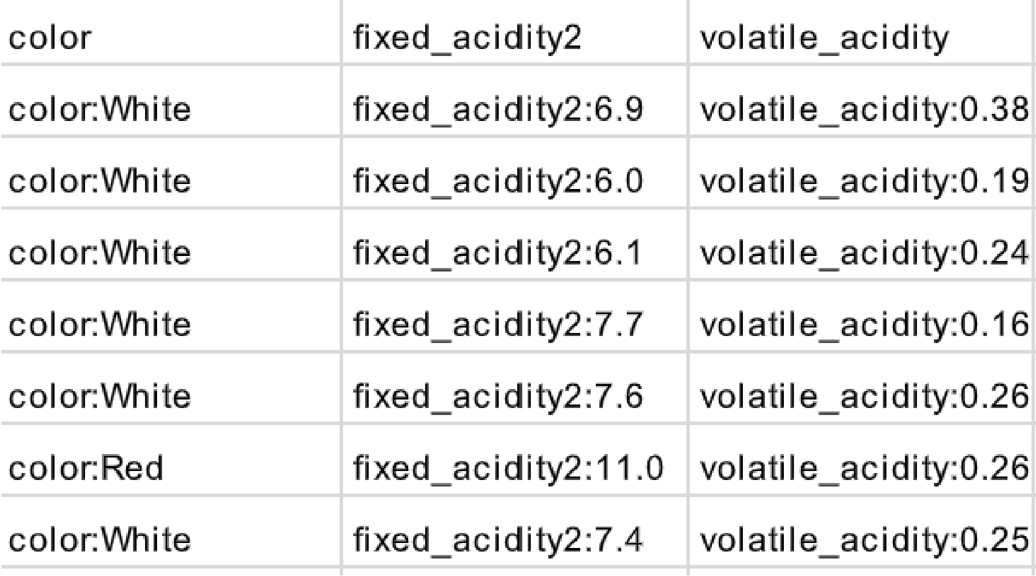
Knowledge Graph Triple

HEXTRATO (c_h:h, r, c_t:t): (Wine:1, color, color: White)

(Wine: 1, fixed_acidity2, fixed_acidity2 : 6.9)



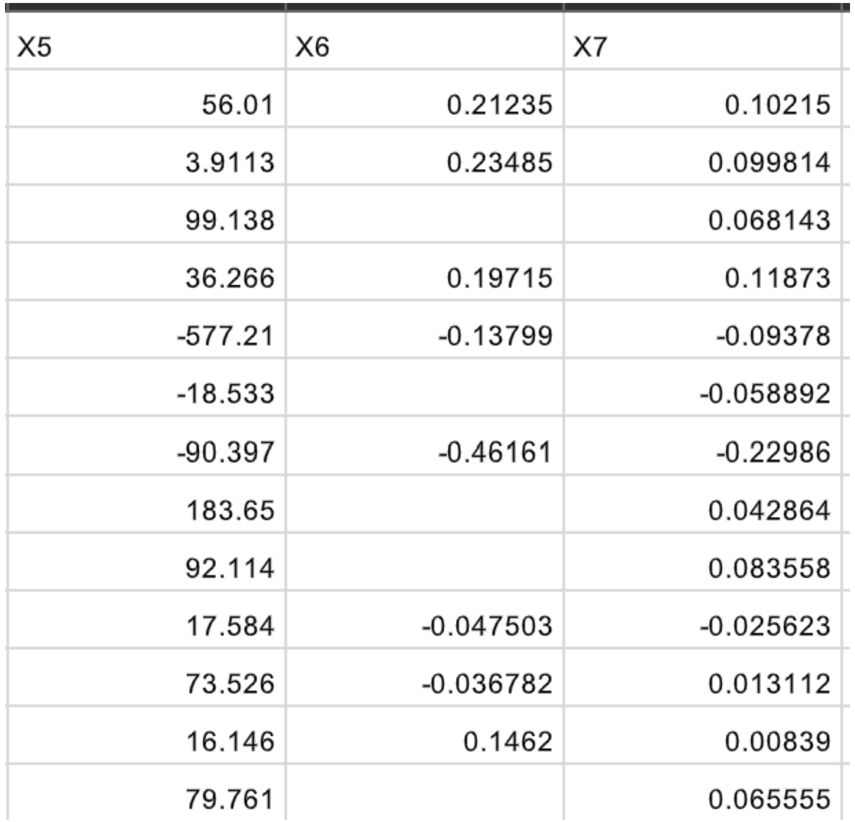
B	C	D
color	fixed_acidity2	volatile_acidity
White	6.9	0.38
White	6	0.19
White	6.1	0.24
White	7.7	0.16
White	7.6	0.26
Red	11	0.26
White	7.4	0.25



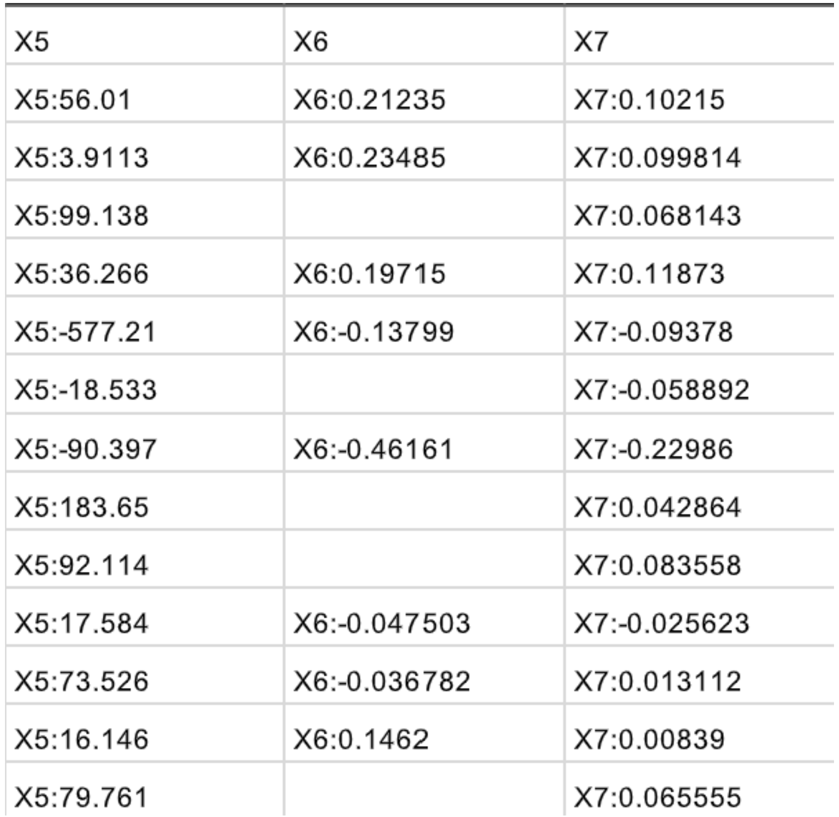
color	fixed_acidity2	volatile_acidity
color:White	fixed_acidity2:6.9	volatile_acidity:0.38
color:White	fixed_acidity2:6.0	volatile_acidity:0.19
color:White	fixed_acidity2:6.1	volatile_acidity:0.24
color:White	fixed_acidity2:7.7	volatile_acidity:0.16
color:White	fixed_acidity2:7.6	volatile_acidity:0.26
color:Red	fixed_acidity2:11.0	volatile_acidity:0.26
color:White	fixed_acidity2:7.4	volatile_acidity:0.25

Knowledge Graph Triple

Missing values



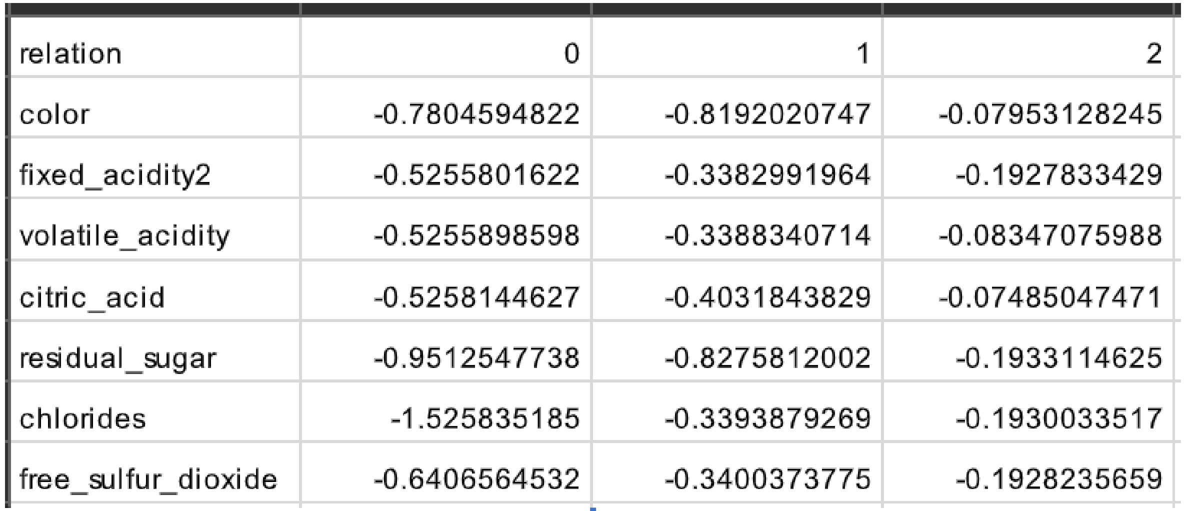
X5	X6	X7
56.01	0.21235	0.10215
3.9113	0.23485	0.099814
99.138		0.068143
36.266	0.19715	0.11873
-577.21	-0.13799	-0.09378
-18.533		-0.058892
-90.397	-0.46161	-0.22986
183.65		0.042864
92.114		0.083558
17.584	-0.047503	-0.025623
73.526	-0.036782	0.013112
16.146	0.1462	0.00839
79.761		0.065555



X5	X6	X7
X5:56.01	X6:0.21235	X7:0.10215
X5:3.9113	X6:0.23485	X7:0.099814
X5:99.138		X7:0.068143
X5:36.266	X6:0.19715	X7:0.11873
X5:-577.21	X6:-0.13799	X7:-0.09378
X5:-18.533		X7:-0.058892
X5:-90.397	X6:-0.46161	X7:-0.22986
X5:183.65		X7:0.042864
X5:92.114		X7:0.083558
X5:17.584	X6:-0.047503	X7:-0.025623
X5:73.526	X6:-0.036782	X7:0.013112
X5:16.146	X6:0.1462	X7:0.00839
X5:79.761		X7:0.065555

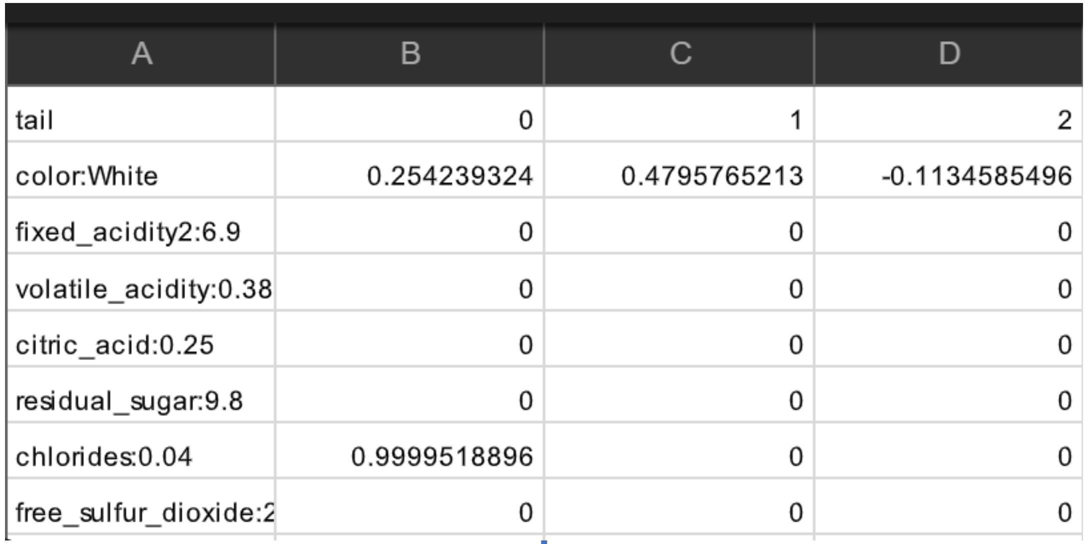
Knowledge Embeddings

Relation embeddings: wine: 32-D



relation	0	1	2
color	-0.7804594822	-0.8192020747	-0.07953128245
fixed_acidity2	-0.5255801622	-0.3382991964	-0.1927833429
volatile_acidity	-0.5255898598	-0.3388340714	-0.08347075988
citric_acid	-0.5258144627	-0.4031843829	-0.07485047471
residual_sugar	-0.9512547738	-0.8275812002	-0.1933114625
chlorides	-1.525835185	-0.3393879269	-0.1930033517
free_sulfur_dioxide	-0.6406564532	-0.3400373775	-0.1928235659

Tail embeddings: wine: 32-D



A	B	C	D
tail	0	1	2
color:White	0.254239324	0.4795765213	-0.1134585496
fixed_acidity2:6.9	0	0	0
volatile_acidity:0.38	0	0	0
citric_acid:0.25	0	0	0
residual_sugar:9.8	0	0	0
chlorides:0.04	0.9999518896	0	0
free_sulfur_dioxide:2	0	0	0

[relation embeddings, tail embeddings]

shape of embedding vector:（10，64）

Model Selection

Dataset	Dataset Info	Classification Task
Polish	Continuous values & Missing values	Binary Classification
Wine	Categorical feature & Continuous Features	Binary Classification
Avila	Continuous Features	Multi-Classification

POLISH	
DATASET	MODEL
TABULAR DATASET WITH ONE-ENCODING	LOGISITIC REGRESSION
TABULAR DATASET WITH ONE-ENCODING	XGBOOST
VARIABLE-LENGTH EMBEDDING VECTORS	BiLSTM-Attention
TAIL EMBEDDING VECTORS	XGBOOST

Model Selection

Wine

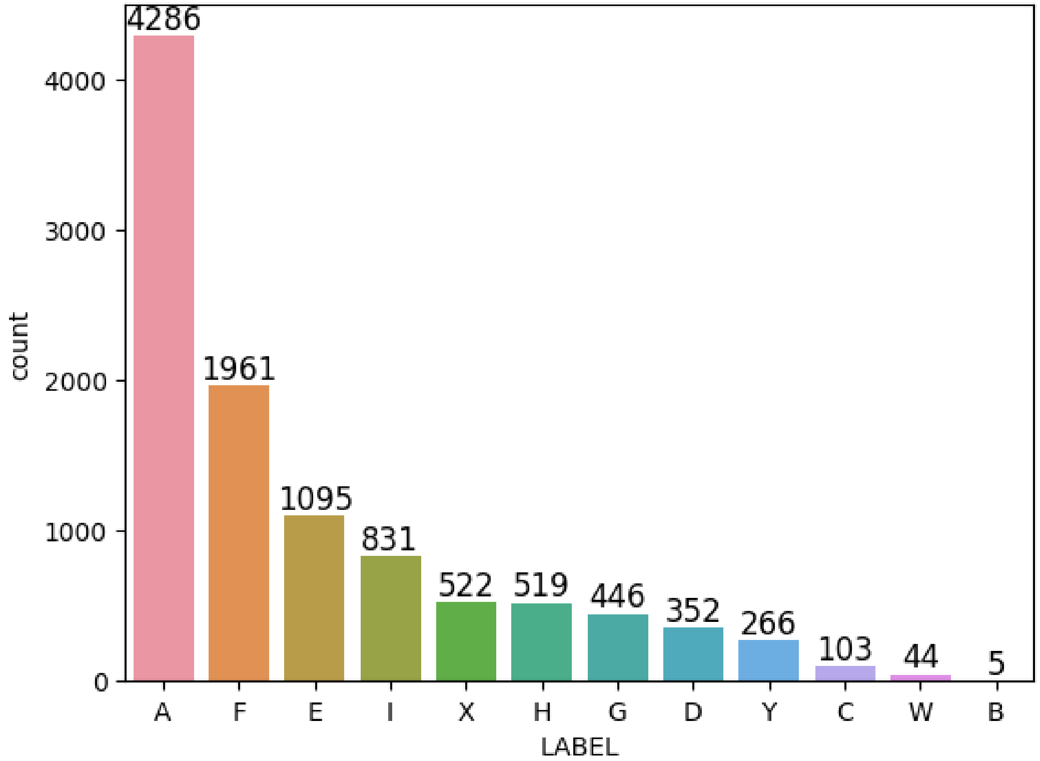
DATASET	MODEL
Tabular dataset with one-hot encoding	Logisitic Rgression
Tabular dataset with one-hot encoding	XGBoost
Variable-Length Embedding vectors	BiLSTM-Attention
Tail Embedding Vectors	XGBoost

Avila

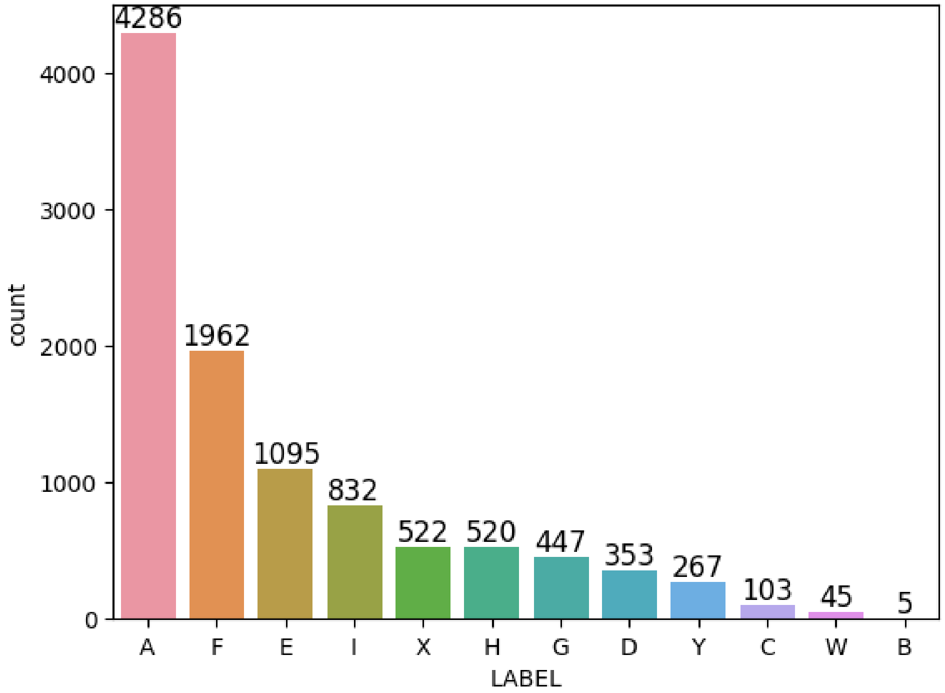
DATASET	MODEL
Tabular dataset with one-hot encoding	Logisitic Rgression
Tabular dataset with one-hot encoding	XGBoost
Variable-Length Embedding vectors	BiLSTM-Attention
Tail Embedding Vectors	XGBoost

Multi-classification

Avila label count



avila y_train label count

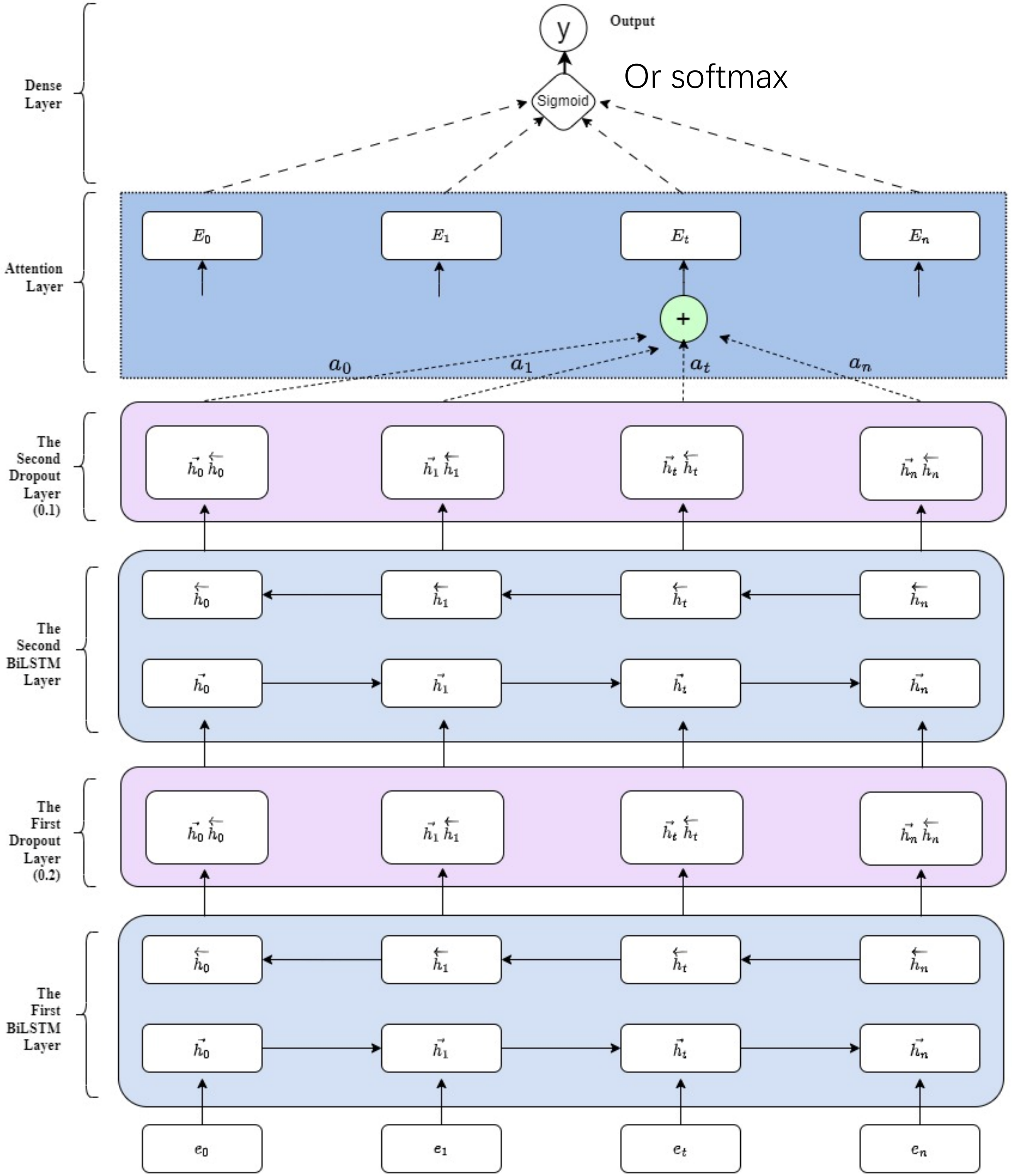


avila y_test label count

```
torch.Size([8, 1])
tensor([[-0.0035],
        [-0.0036],
        [-0.0036],
        [-0.0020],
        [-0.0028],
        [-0.0021],
        [-0.0037],
        [-0.0035]],
```

Tunning & threshold

Input labels:
8 samples with 8 labels
[0,1, 1, 0, 1, 0, 0, 1]



Or softmax

Input batch: [8,1] → per sample has 1 output for deciding label

Input batch: [8, 20]

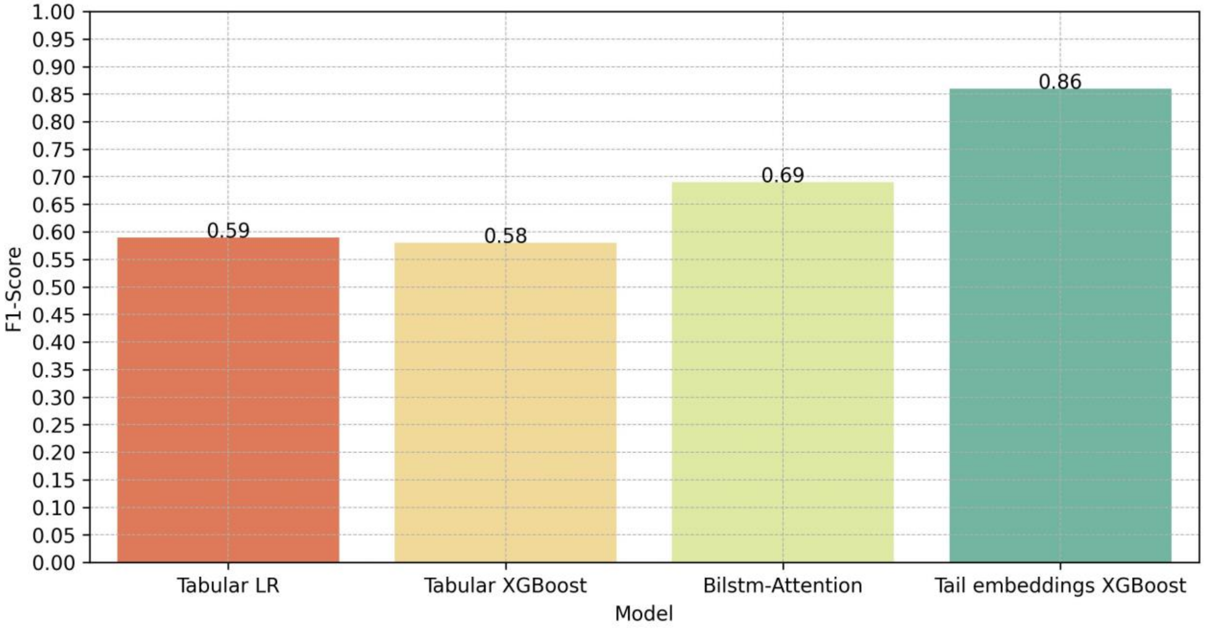
Input batch: [8, 12, 20]

Input batch: [8, 12, 20]

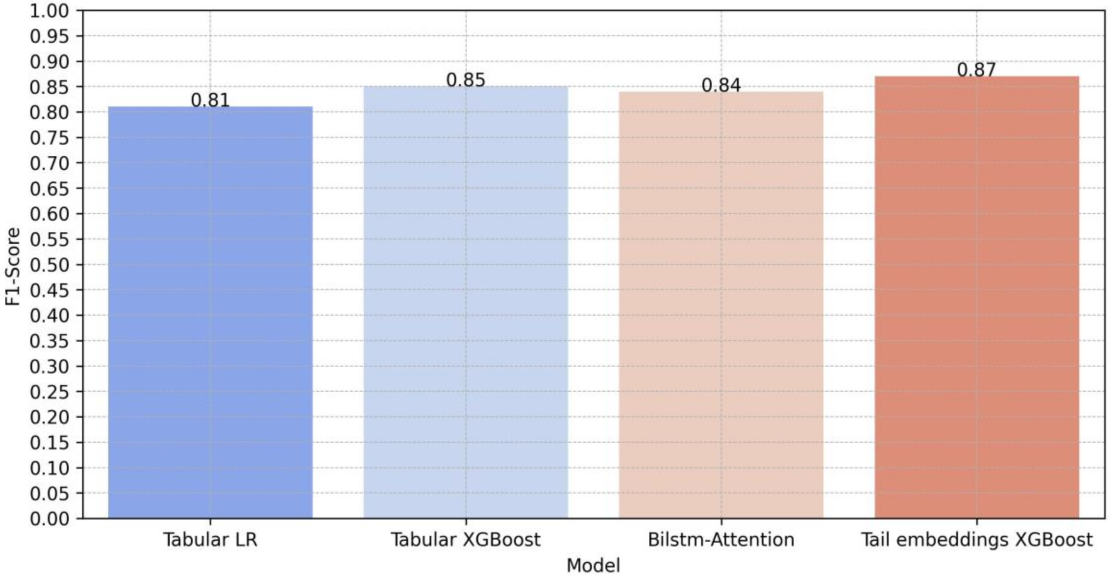
Input batch: [8, 12, 64], hidden units=10

Evaluation F1-score

Binary Classification



Polish model evaluation



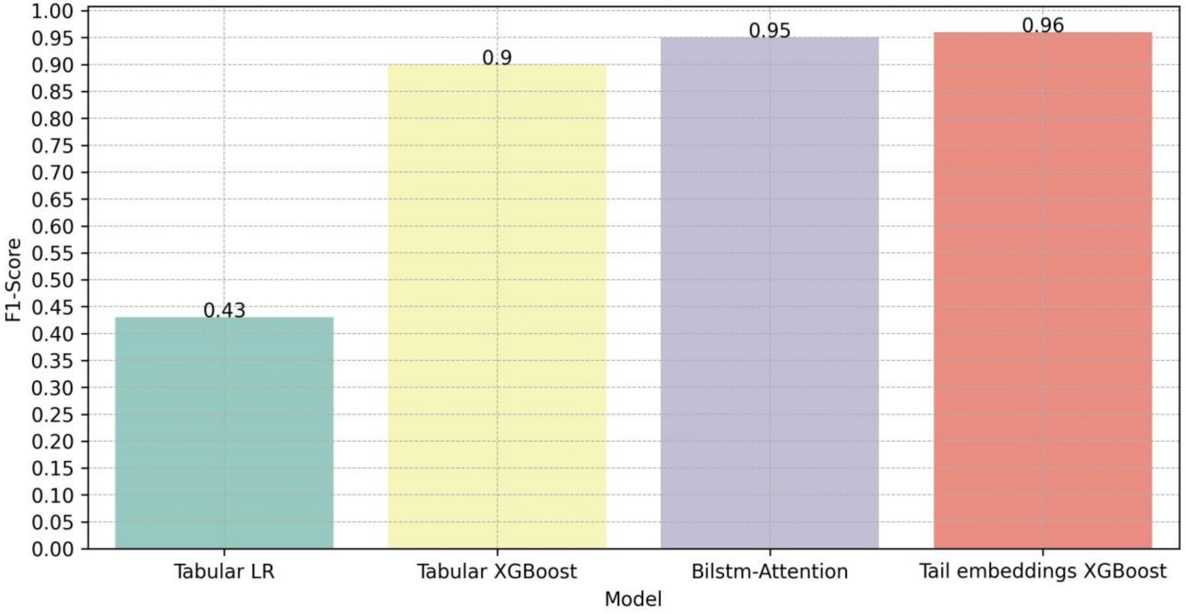
Wine model evaluation

Evaluation F1-score

Multi-Classification

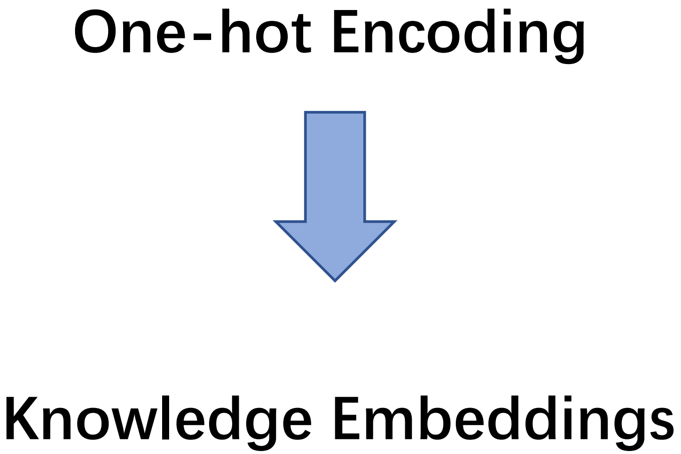
Avila dataset:

Continuous values without
Missing values



Avila model evaluation

Evaluation Model Results



One-hot Encoding

Knowledge Embeddings

Model	Decision
LR using Tabular dataset with one-encoding	✗
XGBoost using Tabular dataset with one-encoding	✗
Neural Network Bilstm-Attention using embeddings	✓
XGBoost using tail embeddings	✓

Contribution

The low-dimensional knowledge embedding representation contributes to:

◆ Handling Missing Values:

Keep the data information as much as possible and do not cause data distortion

◆ Converting Categorical features:

greatly reduce the dimension

◆ Interpretable features:

Categorical features can be more easily because they are represented in terms of real-world entities and their relationships rather than abstract numerical values.

◆ Enhanced machine Learning models:

Embeddings with ontology feature constraints carries more information. Machine learning models have the potential to more effectively capture the relations inherent in the data. This enhancement is frequently manifested in superior performance in tasks such as prediction and classification.

Future Work

Propose A framework that supports low-dimensional vectorial representation for multi-relational data.

Provide an efficient way to integrate temporal information with multi-relational data represented by low-dimensional embeddings.

Construct and Evaluate time-sensitive models on dynamic multi-relational data with temporal information.

Thank you for listening!