Few-shot Learning with LLM Generated Features

UNIVERSITY OF MANNHEIM
School of Business Informatics and Mathematics

Team Project HWS 2026

Idea

UNIVERSITY OF MANNHEIM
School of Business Informatics and Mathematics

- Implement a scikit learn classifier / preprocessor which improves classification/regression tasks based on LLMs

- Example Dataset
  - Predicting the Fuel Consumption of Cars

Auto MPG Data Set

Download: Data Folder, Data Set Description

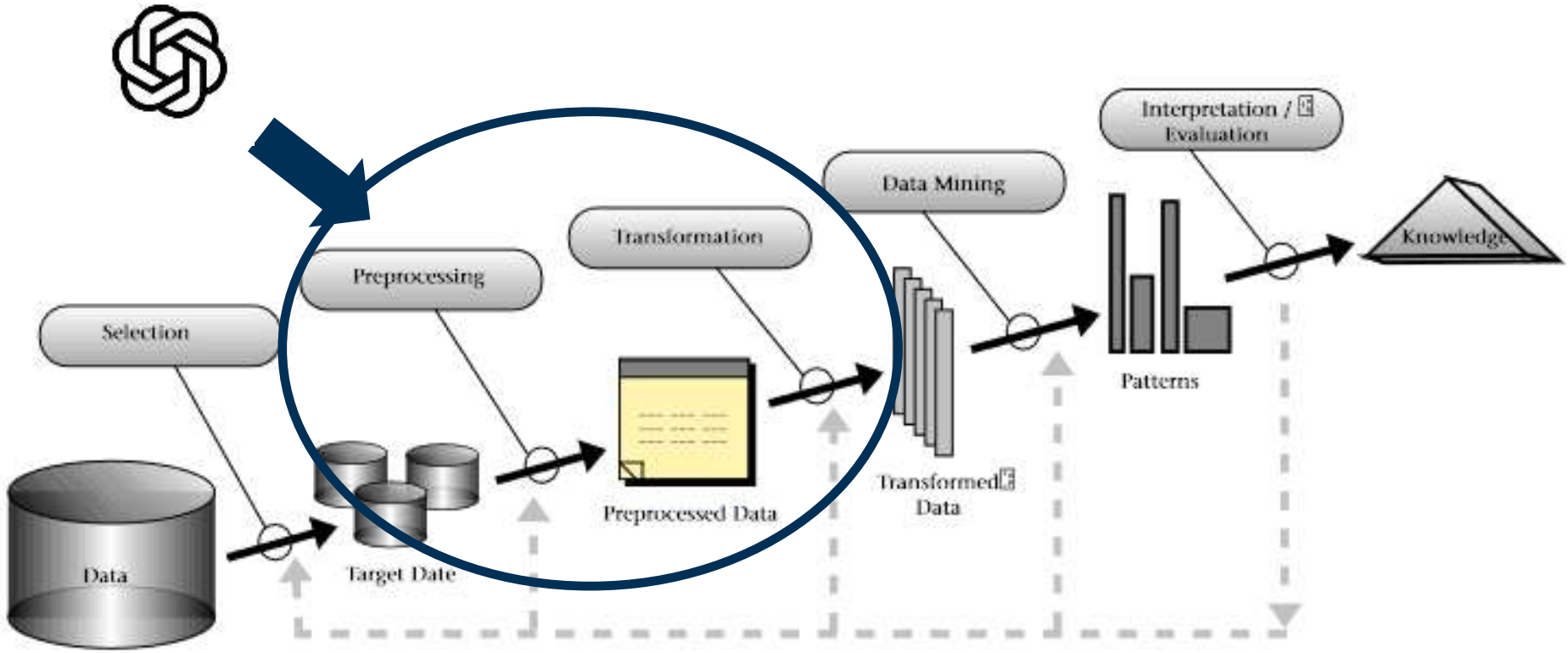
Abstract: Revised from CMU StatLib library, data concerns city-cycle fuel consumption

Data Set Characteristics:	Multivariate	**Number of Instances:**	398	**Area:**	N/A
Attribute Characteristics:	Categorical, Real	**Number of Attributes:**	8	**Date Donated**	1993-07-07
Associated Tasks:	Regression	**Missing Values?**	Yes	**Number of Web Hits:**	509360

http://archive.ics.uci.edu/ml/datasets/Auto+MPG

Idea

- Idea A)
  - Preprocessor which add features based on background knowledge



Idea

- Idea A)
  - Preprocessor which add features based on background knowledge

Model	cylinders	PS	weight	accel
VW Golf	4	125	1200	9.1

way to less attributes to come up with a good model

LLM Background Information

Model	cylinders	PS	weight	accel	brand	length	width	assembly	production start year
VW Golf	4	125	1200	9.1	VW	4255.0	1799	Indonesia	2012-01-01

Idea

UNIVERSITY OF MANNHEIM
School of Business Informatics and Mathematics

- Idea B)
  - Preprocessor which add more examples

Model	cylinders	PS	weight	accel
VW Golf	4	125	1200	9.1

LLM Background Information

Model	cylinders	PS	weight	accel
VW Golf	4	125	1200	9.1
chevrolet chevelle malibu	8	130	3504	12
ford torino	8	302	3449	10.5

Idea

- Idea C)
  - For datasets including multimodal data like image/video /audio:
    - use a coding harness for the LLM and ask it to implement functions that extract **leightweight** features
    - E.g. skin cancer prediction (HAM10000)

image	age	sex	location	diagnosis
	45	male	back	melanoma

Let the LLM do the feature engineering for you

image	age	sex	location	Output of Python Function: radius	diagnosis
	45	male	back	0.3	melanoma

Idea

- Idea D)
  - For datasets including multimodal data like image/video /audio:
    - Give e.g. the image directly to an LLM to getn the feature representation directly (costly because for <u>each example during inference</u> a LLM needs to be executed)

image	age	sex	location	<u>diagnosis</u>
	45	male	back	melanoma

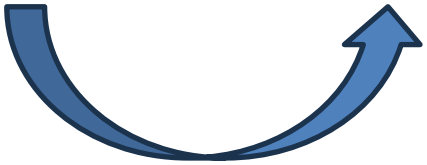
Let the LLM do the feature engineering for you

image	age	sex	location	Output of LLM: skin type	<u>diagnosis</u>
	45	male	back	Skin Type: IV	melanoma

Idea

- Idea E)
  - Let the LLM do the classification itself
    (either with or without additional features or additional examples)

image	age	sex	location	diagnosis
	45	male	back	melanoma



LLM

- Which prompts ?
- How much background information?
- Confidence extraction?

Organisational information

UNIVERSITY OF MANNHEIM
School of Business Informatics and Mathematics

- The project runs for one semester (six months)
- 2-4 students
- There will be regular meetings
  - Some in person, most can be online
- Computing infrastructure is provided by the DWS group
- Required skills of the team
  - Good programming skills in Python
  - Attended (or attending) Data Mining

Supervisor

Dr. Sven Hertling
Substitute Professor for Data Science

E-Mail: sven.hertling@uni-mannheim.de

Links

UNIVERSITY OF MANNHEIM
School of Business Informatics and Mathematics

- Idea A
  - Previous team project which uses background knowledge from knowledge graphs
    - https://madoc.bib.uni-mannheim.de/60076/
    - https://openreview.net/pdf?id=2FQgZhV-LDL
    - kgextension
      - https://github.com/om-hb/kgextension

- Idea E (without additional features)
  - Scikit-LLM
    - https://github.com/fnnx-ai/scikit-llm
    - https://skllm.beastbyte.ai/

Questions