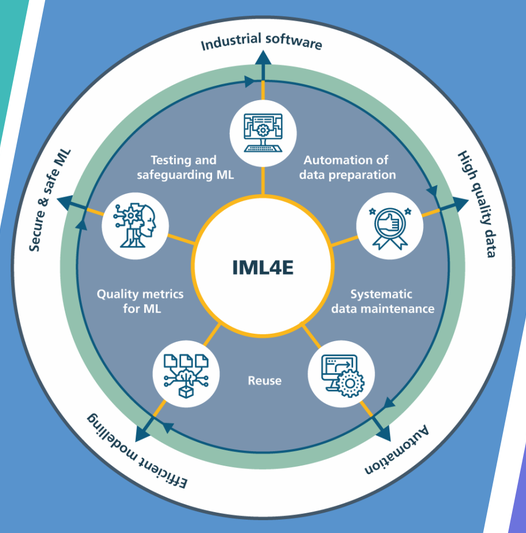
Benchmarking von Bewertungstools für tabellarische Daten in ML-Pipelines

Projektmotivation

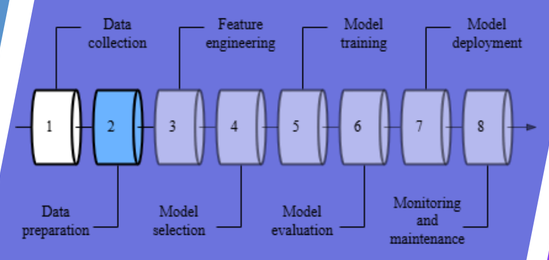
Problem

- Daten aus der echten Welt sind häufig Fehlerbehaftet
- Fehler entstehen bei der Datensammlung oder Datenverarbeitung

Forschungsprojekt



Anwendungsbereich



- Datenvorbereitung ist ein grundlegender Schritt einer ML-Pipeline

Grundlage

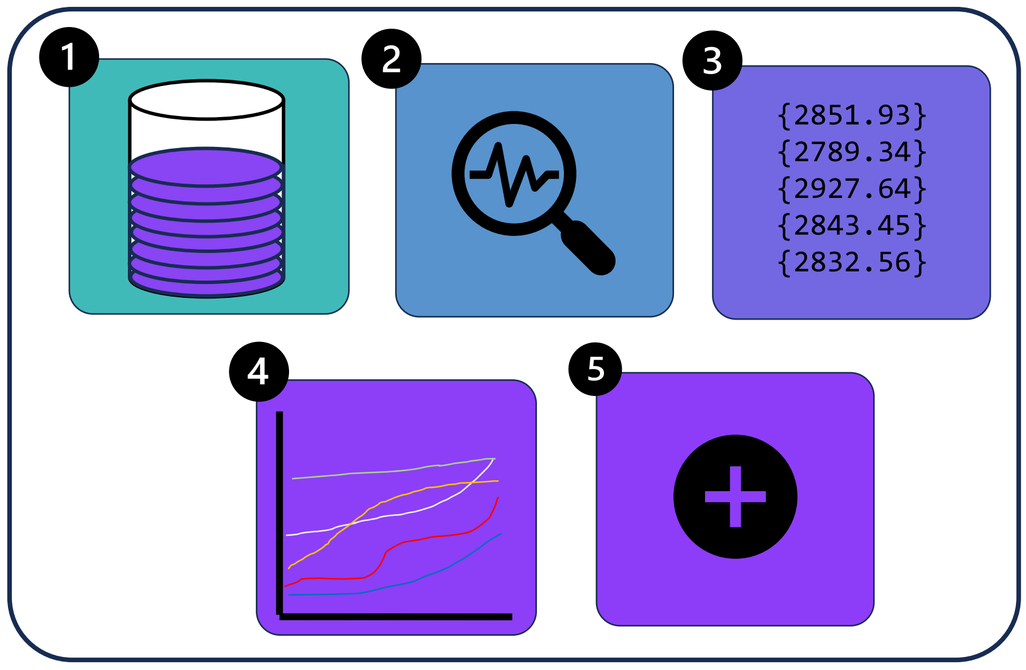
Attr. 1	Attr. 2	Wert
a	x	0.34
b	y	0.65
c	z	0.74

- Datenbewertung als Grundlage für Anwendungsfälle
- Anwendungsfälle wie z.B. Data summarization

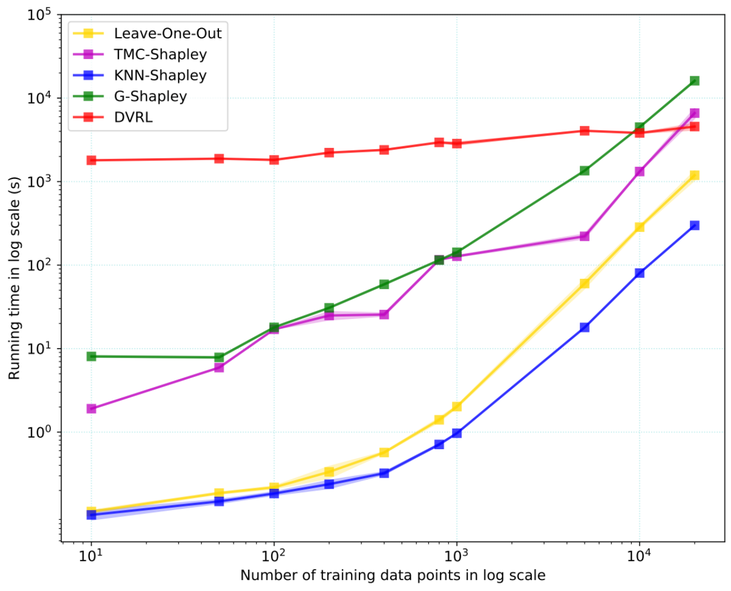
Projektablauf

1	Datensatz und Datenmenge wählen
2	Datenbewertungstool wählen
3	Ergebnisse in eine CSV-Datei schreiben
4	Ergebnisse grafisch darstellen
5	Hinzufügen neuer Datenbewertungstools

- Die Basis des Benchmarks wurde bereits entwickelt
- Der Workflow des Benchmarks wurde wie in 1,2,3,4 beschrieben automatisiert
- Hinzufügen neuer Datenbewertungstools
- Verwendung eines Jupyter Notebooks



Projektergebnisse



- Vergleichen verschiedener Datenbewertungstools
- Mehrmalige Wiederholung der Tests, um auch die Standardabweichung darstellen zu können
- Untersuchung der Laufzeit bei verschieden großen Datensätzen
- Benchmarking von unterschiedlichen Datensätzen
- Automatisierung des Benchmark-Ablaufs
- Updaten der bereits bestehenden Codebasis
- Anpassen der bereits bestehenden Codebasis, um das Hinzufügen weiterer Bewertungstools zu vereinfachen
- Datenbewertung als Grundlage für Anwendungsfälle

Attr. 1	Attr. 2	Wert
a	a	0.34
b	b	0.65
c	c	0.74
d	d	0.83

Attr. 1	Attr. 2	Wert
b	b	0.65
c	c	0.74
d	d	0.83

Fazit und Ausblick

- Die gewonnenen Ergebnisse liefern wertvolle Erkenntnisse für Forscher und Unternehmen
- Daten aus der echten Welt sind häufig fehlerbehaftet, weswegen eine Datenvorbereitung unabdinglich ist
- Beim Durchführen des Benchmarks sind Probleme beim Hinzufügen neuer Datenbewertungstools aufgrund von Versionskonflikten aufgetreten, welche behoben werden mussten
- Die zukünftige Erweiterung des Benchmarks mit weiteren Datensätzen, Datenbewertungstools und Bewertungsparametern ermöglicht es für jeden Anwendungsfall die besten Datenbewertungstools herauszufinden

h_da hochschule darmstadt informatik
member of EUT+ EUROPEAN UNIVERSITY OF TECHNOLOGY
software AG A SOFTWARE GMBH BRAND

Praxisprojekt 2
Informatik Dual
2025

Gruppe B