Zum Einsatz von Maschinellem Lernen in der Umweltverwaltung: Der Simplex4Learning Ansatz

Andreas Abecker[1], Matthias Budde[2], Frank Fuchs-Kittowski[3], Janik Großmann[4], Werner Koch[5], Jonas Lachowitzer[6], Stefan Lossow[7], Erik Rodner[8], Heino Rudolf[9] und Paul Schulze[10]

Abstract: Ziel des im Herbst 2023 gestarteten Forschungsvorhabens Simplex4Learning ist es, die großen und heterogenen Datenbestände der Umweltbehörden für intelligente Analysen mit Methoden des maschinellen Lernens besser zu erschließen und diese Verfahren für Domänenexperten aus dem Umweltbereich ohne vertiefte ML-Kenntnisse praktikabel anwendbar zu machen. Realisiert wird dies (1) durch die Weiterentwicklung der Simplex4Data-Methode zur Datenbereitstellung für ML, ergänzt um (2) AutoML- und MLOps-Funktionalitäten, (3) Funktionalitäten zum Erklären von ML-Ergebnissen, (4) ein ML-Pattern Repository zum Wiederverwenden generalisierter ML-Workflows, all das (5) exemplarisch angebunden an die Datenanalyseplattform Disy Cadenza und das Data Warehouse System Simplex4Data. Der Arbeitsplan des Projekts ist an den konkreten Beispieldaten und Anwendungsfällen von Landesbehörden aus drei Bundesländern orientiert. Der vorliegende Beitrag als „Work-in-Progress"-Bericht skizziert Motivation und Ausgangslage des Vorhabens, den technischen Lösungsansatz und erste Zwischenergebnisse.

Keywords: Maschinelles Lernen, Umweltverwaltung, Umweltmonitoring, Praxiseinsatz, AutoML, MLOps, XAI, Simplex4Data, Cadenza

1 Motivation

Die öffentliche Verwaltung muss im Umweltbereich, bzw. im Kontext von umwelt-, lebens- und geowissenschaftlichen Fragestellungen auf Gebieten wie beispielsweise Naturschutz, Verbraucherschutz, Land- und Forstwirtschaft, Wasserwirtschaft, Infrastruktur, für die verschiedensten Fragestellungen – zu Themen wie zum Beispiel Luftqualität, Wasserquantität und -qualität, Bodengüte, Biodiversität, Tierwohl –

[1] Disy Informationssysteme GmbH, Ludwig-Erhard-Allee 6, 76131 Karlsruhe, andreas.abecker@disy.net
[2] Disy Informationssysteme GmbH, Ludwig-Erhard-Allee 6, 76131 Karlsruhe, matthias.budde@disy.net
[3] Hochschule für Technik und Wirtschaft, Fachbereich 2, Umweltinformatik, Wilhelminenhofstraße 75A, 12459 Berlin, Frank.Fuchs-Kittowski@HTW-Berlin.de
[4] Simplex4Data GmbH, Am Waldschlößchen 4, 01099 Dresden, janik.grossmann@simplex4data.de
[5] Simplex4Data GmbH, Am Waldschlößchen 4, 01099 Dresden, werner.koch@simplex4data.de
[6] Disy Informationssysteme GmbH, Ludwig-Erhard-Allee 6, 76131 Karlsruhe, jonas.lachowitzer@disy.net
[7] Disy Informationssysteme GmbH, Ludwig-Erhard-Allee 6, 76131 Karlsruhe, stefan.lossow@disy.net
[8] Hochschule für Technik und Wirtschaft, Fachbereich 2, Ingenieurwissenschaften, Wilhelminenhofstraße 75A, 12459 Berlin, Erik.Rodner@HTW-Berlin.de
[9] Simplex4Data GmbH, Am Waldschlößchen 4, 01099 Dresden, heino.rudolf@simplex4data.de
[10] Hochschule für Technik und Wirtschaft, Fachbereich 2, Umweltinformatik, Wilhelminenhofstraße 75A, 12459 Berlin, Paul.Schulze@HTW-Berlin.de

doi:10.18420/inf2024_105

vielfältige ökologische Phänomene und ihr komplexes Zusammenspiel mit der Umwelt überwachen und möglichst gut verstehen, um die Politik bei Planungs- und Steuerungsvorhaben zu unterstützen, die Öffentlichkeit zu informieren oder gegebenenfalls in Notfällen auch zu warnen, aber auch der Wirtschaft sowie anderen Verwaltungseinheiten gute Datengrundlagen für Investitionsentscheidungen und ähnliches zu geben. In Zeiten großer gesamtgesellschaftlicher Umbaunotwendigkeiten (wie Energiewende, Mobilitätswende, klimafreundlicherer Industrieumbau, klimaangepasster Städtebau, klimaangepasster Umbau des Waldes), aber auch wachsender Gefährdungen durch Wetterextreme, invasive Arten, sich schnell global ausbreitende Krankheiten und ähnliches, entstehen ganz neue Anforderungen an Montoring von Umweltaspekten, Frühwarnung, Prognosen zu zukünftigen Entwicklungen, Szenarioanalysen für mögliche Entwicklungspfade und ähnliches mehr.

Andererseits gibt es in den öffentlichen Verwaltungen zum Teil bereits umfangreiche und vielfältige Datenbestände und Messnetze, um Aufgaben wie die oben skizzierten datenbasiert anzugehen. Diese Datenbestände wachsen auch, qualitativ und quantitativ, mit zunehmender Geschwindigkeit, befördert durch Technologieentwicklungen wie immer günstigere und leistungsfähigere Fernerkundungstechnologien mit Satelliten und Drohnen, ausgefeilte Online-Messnetze, Citizen Science und anderes mehr.

Moderne Verfahren der KI und des Maschinellen Lernens finden in der Forschung und in der Industrie immer mehr wertvolle Anwendungen für die datenbasierte Entscheidungsunterstützung. In der öffentlichen Verwaltung sind solche Anwendungen noch deutlich seltener. Dies hat sicherlich technische und organisatorische, aber auch fachliche Gründe. Aus technischer Sicht hat die Verwaltung beispielsweise in vielen Prozessen hohe Anforderungen an Nachvollziehbarkeit, Transparenz und Zuverlässigkeit von Ergebnissen. Auch Datenschutz und -sicherheit oder Barrierefreiheit in der Außendarstellung muss man ernster nehmen als dies vielleicht mancher privatwirtschaftliche Akteur tut. Organisatorisch sind häufig sehr verteilte Strukturen und Prozesse (nach Raum, Inhalt und Zuständigkeit/ Kompetenz verteilt), klamme Kassen und langsame Budgetierungsprozesse, Nachwuchsprobleme beim Personal oder strukturell kaum vermeidbare Schwerfälligkeit gegenüber organisatorischen Veränderungen die Hürden bei der Digitalisierung. Aber auch aus fachlicher Sicht bewegen wir uns bei umweltbezogenen Fragestellungen in Bereich höchster Komplexität, wo die betrachteten ökologischen Prozesse und Zusammenhänge mit weiteren komplexen oder gar chaotischen Systemen, wie dem Wetter oder der Wirtschaft, interagieren, so dass das Gesamtverhalten teils kaum vollständig verstanden bzw. modellierbar ist und für eine Modellierung häufig auch sehr vielen Eingangsdaten erforderlich wären, die nicht immer und überall in der benötigten Qualität vorliegen.

Dennoch zeigt die jüngere Vergangenheit einige vielversprechende und interessante Anwendungen des Maschinellen Lernens auf typischen Themenfeldern der (Umwelt-) Verwaltung, die zunehmend auch in Zusammenarbeit mit Behörden entstanden sind oder dort Aufmerksamkeit finden. So verwenden beispielsweise [WLG24] bzw. [Ka23] Ansätze des Deep Learning für die Vorhersage von Grundwasserständen bzw. zum

Verständnis der Verteilung von Nitrat im Grundwasser. [Be24] erzielten sehr gute Ergebnisse bei der ML-basierten Prognose von Pegelständen an Fließgewässern. Auch zur Untersuchung und Prognose der Luftqualität gibt es zahlreiche Arbeiten, die vielfach bereits in kommerzielle Angebote eingeflossen sind und von vielen Kommunen unterstützt bzw. erprobt werden, siehe hierzu z.B. [TRB21] oder [So21]. Viele andere Themen und Technologien werden zunehmend für Umweltverwaltungen interessanter, von Verfahren der Bilderkennung aus Luftbildern etc., über Gefährdungsklassifikationen, zum Beispiel zur Waldbrandgefahr, bis hin zur automatisierten Sensordatenüberwachung für die Anomalieerkennung, Frühwarnung und ähnliches (siehe [Bu24] oder auch [Go22] und [Kl21] für weitere Beispiele).

Um zumindest die technologischen Hürden des ML-Einsatzes zu reduzieren, wurde daher im Oktober 2023 das FuE-Vorhaben Simplex4Learning gestartet. Übergeordnetes Projektziel von Simplex4Learning ist es, (1) die großen, aber komplexen und heterogenen Datenbestände der Umweltbehörden für intelligente Analysen mit Methoden des Maschinellen Lernens besser zu erschließen und (2) die Verfahren des Maschinellen Lernens für Domänenexpert:innen aus dem Umweltbereich praktikabler anwendbar zu machen.

Die Umsetzung von Ziel 1 (Datenbereitstellung) basiert darauf, relevante Datenbestände aus dem Umweltbereich mit ihren vielfältigen zugrundeliegenden Basisdaten (wie Geodaten, Regionalpläne und anderes mehr) und aufgabenspezifisch relevanten Zusatzdaten (wie zum Beispiel Verkehrsströme, Wetter) in einem Data Warehouse harmonisiert zusammenzuführen und dann mit standardisierten Diensten zuzugreifen. Dies nutzt die Methoden und Technologien der Simplex4Data GmbH, auf die wir im vorliegenden Beitrag nicht vertieft eingehen. Wir verweisen hier beispielsweise auf [Ru18], [Gr22], [GR23].

Mit Bezug auf Ziel 2 (Fachanwendertauglichkeit) gibt es zwar heutzutage eine große Breite mächtiger Softwarewerkzeuge, -plattformen und -bibliotheken für Datenintegration, Datenanalyse und Data Science, mit dem Paradigma des automatisierten Machine Learning (AutoML) sogar für eine teilweise Automatisierung des Prozesses, aber all diese Werkzeuge sind zurzeit noch Spezialist:innen vorbehalten, weil ihre zielführende Verwendung außer der Fach- und Domänenexpertise ein großes Maß an Analyse-, Methoden- und Werkzeugwissen erfordert. Wir versuchen daher, diese in einer praktikablen Gesamtkonfiguration zu bündeln und durch die Kopplung an bereits existierende und weitverbreitete Systeme für (Geo-)Dateninfrastrukturen in Behörden besser zugänglich zu machen.

Im vorliegenden Beitrag skizzieren wir im nachfolgenden Kapitel 2 den übergeordneten technologischen Lösungsansatz und einige der verwendeten Basistechnologien. In Kapitel 3 listen wir wichtige Forschungsfragen im Projekt auf und skizzieren erste Ergebnisse. Kapitel 4 schließt mit einer kurzen Zusammenfassung und einem Ausblick.

2 Lösungsansatz und Basistechnologien

Der grobe Lösungsansatz im Projekt wird in der nachfolgenden Abbildung vereinfachend illustriert: Simplex4Data stellt einen harmonisierten Speicher für heterogene Umweltdaten zur Verfügung. Diese Daten werden für die ML-Verfahren mittels standardisierter Diensteschnittstellen von Simplex4Data aus Cadenza heraus zugänglich. Über Cadenza werden Lernaufgaben und -daten von Fachanwender:innen an die angebundene ML-Infrastruktur weitergegeben. Dort sind fachlich beschriebene, vortrainierte ML-Modelle oder ML-Workflows in generalisierter Form in einem ML-Repository abgelegt. Diese können für die aktuelle Aufgabe aufgerufen werden und die Ergebnisse zusammen mit Ergebnissen des XAI wiederum in Cadenza inspiziert werden. Die Befüllung des ML-Repository erfolgt durch ML-Expert:innen, ggf. in Zusammenarbeit mit Domänenfachleuten.

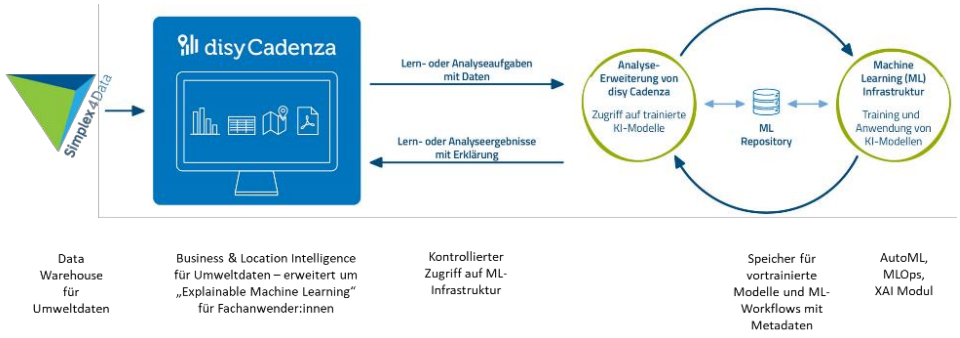

Abb. 1: Vereinfachter allgemeiner Lösungsansatz von Simplex4Learning

Im Vorhaben sollen die folgenden Basistechnologien und Technologiefelder kombiniert werden:

Disy Cadenza ist eine Plattform, die auf vielfältigen (Geo-)Datenbanken, (Geo-) Datendiensten oder (Geo-)Dateien in Form eines virtuellen oder persistenten (Geo-)Data Warehouse aufsetzen kann, auf einem harmonisierten Datenmodell die feinkörnige Definition von Rollen und Rechten ermöglicht und dann einfache Möglichkeiten bereitstellt, um web-basiert Dashboards und Anwendungen zur Visualisierung, Analyse und zum Reporting von Sach- und Geodaten ermöglicht [Di24]. Zusätzliche GIS-Operationen und die Möglichkeit zur tiefen Integration von räumlichen und nichträumlichen Datenverarbeitungsfunktionen machen das Werkzeug zu einer Plattform für „Business & Location Intelligence". Cadenza ist bei zahlreichen deutschen Umweltbehörden im Einsatz und stellt daher das Bindeglied zwischen den etablierten Geodateninfrastrukturen in den Behörden und den Erweiterungen durch das Simplex4Learning Vorhaben dar.

Unter **AutoML** (englisch: Automated Machine Learning [HKV19], [Bi22]) werden Methoden zur Automatisierung der Anwendung von Maschinellem Lernen auf reale Probleme (End-to-End-ML-Prozess) verstanden (siehe auch [Bi20]). Es gibt umfangreiche Veröffentlichungen zur Automatisierung des Maschinellen Lernens, in denen versucht wird, alle Aspekte zu automatisieren. Trotz zahlreicher Erfolge erfordert dieser Prozess immer noch ein hohes Maß an menschlicher Beteiligung. Eine Reihe wichtiger Arbeitsschritte, wie das Schaffen von Domänenverständnis und die Aufgabenformulierung, sind nicht sinnvoll automatisierbar und erfordern üblicherweise eine intensive Kommunikation zwischen Domänenexperten und ML-Experten. Allerdings gibt es viele Systeme zur Automatisierung anderer, mechanischerer Arbeitsschritte, wie zum Beispiel Feature Inference (Merkmalsextraktion aus Datensätzen), Feature Validation (Dimensionsreduktion oder auch Identifizieren eines möglichen Bias bei einem gewissen Merkmal), Model Selection (Auswahl des geeignetsten Modells) oder Hyperparameter Optimization (Tuning der Modellparameter).

Erwartungsgemäß geht es darum, ganze oder partielle ML-Trainings- und Ausführungsprozesse zu dokumentieren, abzuspeichern, wiederzufinden und mit möglichst wenig Aufwand skalierbar anzuwenden, in der Organisation auszuspielen bzw. auch neu zu parametrieren. Mit dieser umfangreichen Zielsetzung bewegen wir uns in den Bereich von **Machine Learning Operations (MLOps)** (siehe [KKH22], [Fe15] oder [Fe21]). Dieses Feld ist noch recht jung und nicht ganz trennscharf definiert, umfasst aber mit dem Management und Deployment von ML-Modellen und ganzen ML-Workflows wesentliche Themen, die auch in Simplex4Learning bearbeitet werden müssen. Abb. 2 unten zeigt eine recht vollständige Sicht auf den MLOps-Kosmos, der insbesondere auch den AutoML-Teil umschließt.

Nach *ml-ops.org* umfasst das Thema MLOps die folgenden Aufgabenbereiche: (a) Data Engineering, (b) Version Control of Data, (c) ML Models and Code, (d) Continuous Integration and Continuous Delivery Pipelines, (e) Automating Deployments and Experiments, (f) Model Performance Assessment sowie (g) Model Monitoring in Production. Henrik Skogström [Sk20] hat eine Template-Darstellung für einen MLOps-Technologiestack gestaltet, der in Abb. 3 beispielhaft instantiiert wird.

Für jeden einzelnen der Schritte gibt es eine Vielzahl potenziell nutzbarer Technologiealternativen, viele davon allerdings proprietärer Code kommerzieller Anbieter, auf die wir hier verzichten wollen. Einige Elemente sind dabei von den Tools der Projektteilnehmer zu besetzen (zum Beispiel Simplex4Data und Disy Cadenza im Bereich Data Engineering bzw. Data Analysis) bzw. werden in eigenen Arbeitspaketen des vorliegenden Projekts gestaltet (zum Beispiel Model Monitoring).

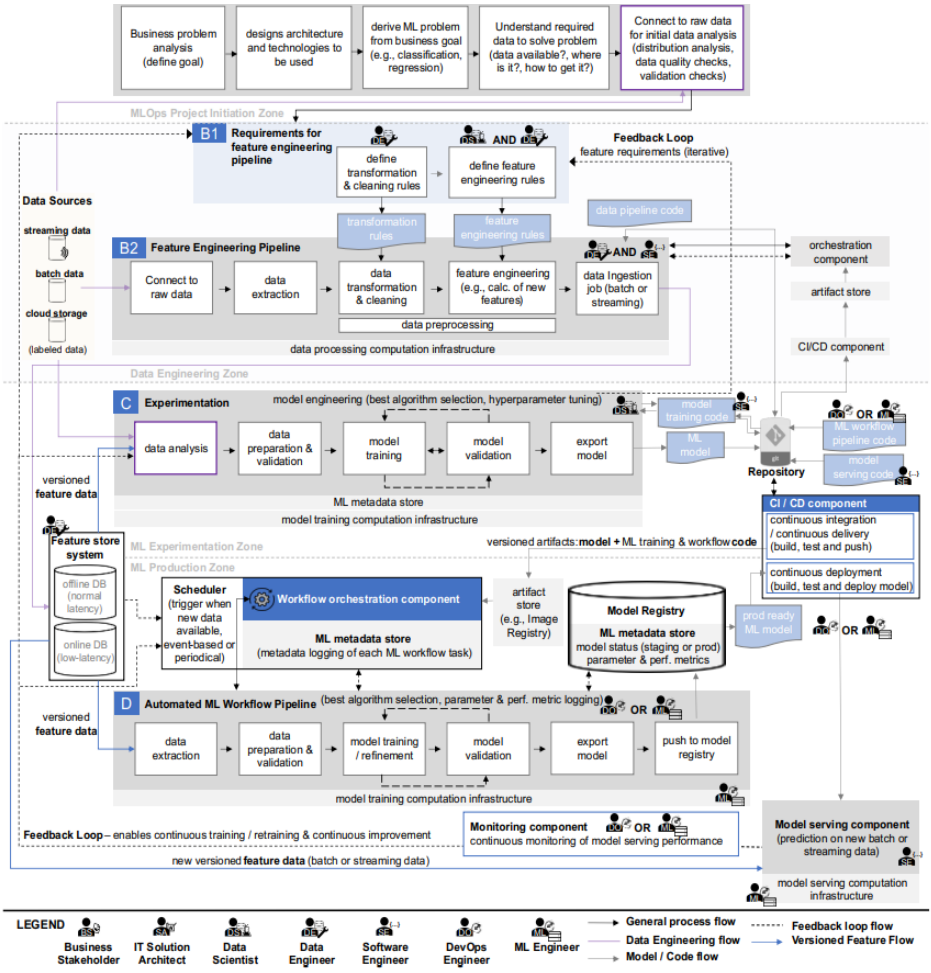

Abb. 2: End-to-end MLOps Architektur und Workflow
mit funktionalen Komponenten und Rollen nach [KKH22]

[He20] gibt eine Übersicht wichtiger aktueller MLOps-Lösungen, wie Algorithmia von DataRobot (Schwerpunkt auf Verwaltung, Ausrollen und skalierender Nutzung von ML-Modellen inklusive Modell-Versionierung, Pipelining und verschlagworteter Modell-Bibliothek; weniger Fokus auf den frühen Phasen der Modellentwicklung), SageMaker von Amazon (vollintegrierte Umgebung für Machine und Deep Learning), Azure Machine Learning für die Microsoft Azure Cloud-Plattform, die Domino Data-Science-Plattform (automatisierte DevOps Funktionen für Data Science Aufgaben, standardmäßig mit AWS verknüpft), HPE Ezmeral ML Ops von Hewlett Packard (für durchgängigen cloud-basierten ML-Lifeycycle Support, von der Experimentierphase bis zum Deployment, *on premise* und auf den gängigen Public Clouds), Metaflow (Open-

Source Python-basiertes Workflow-System für das ML-Lifecycle Management [Me24]), MLflow (Open-source ML-Lifecycle Management Plattform mit Model Repository [Ml24]), Seldon Core (Open-source Plattform mit der Möglichkeit, ML-Modelle auf Kubernetes auszurollen) sowie Seldon Deploy (skalierbares Ausrollen in der Cloud und On-Premise für vielfältige Programmiersprachen und Frameworks Modelle) und Seldon Alibi (Open-Source Python-Bibliothek, für Black-Box Machine Learning, Modell-Inspektionen und -Interpretationen), Google‘s Cloud-AI-Plattform Vertex AI (Funktionen und Workflows für Data Engineering, AutoML und MLOps sowie Verfahren für unstrukturierte Datenquellen), ClearML (Python-basierte Open-Source Plattform für skalierbare und reproduzierbare Data Science Prozesse [Al24]).

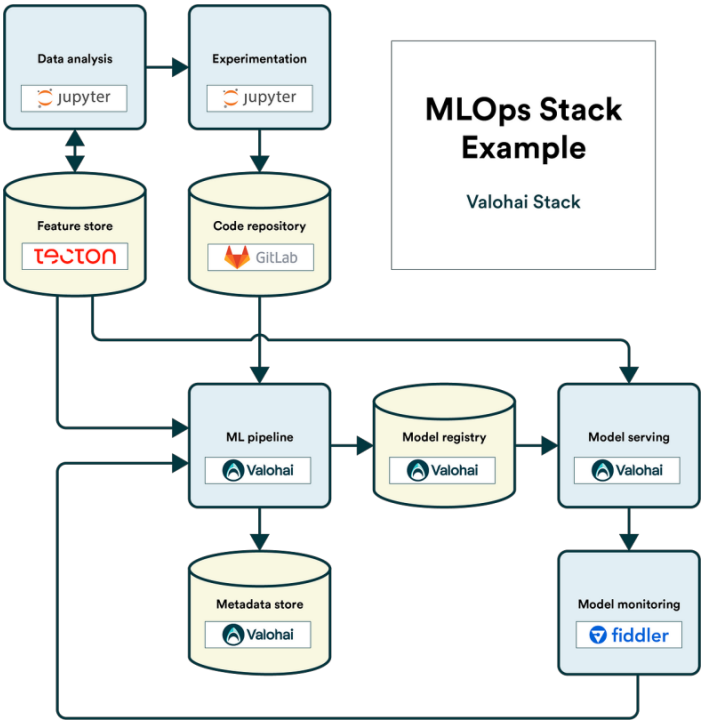

Abb. 3: MLOps Stack von (C) Henrik Skogström,
beispielhaft instantiiert mit Valohai Tools

Unter dem Label der **Explainable AI** finden sich aktuell sehr viele Arbeiten zur Frage, wie man Black-Box ML-Verfahren auf Basis Neuronaler Netze symbolisch verständlich macht (siehe [Mo22], [RSG16] oder [LL17]). Hierbei sind zwei Ansätze sehr dominant. LIME (Local Interpretable Model-Agnostic Explanations) ist eine Technik, die jedes Black-Box-Modell des maschinellen Lernens durch ein lokales, interpretierbares Modell ersetzt, um jede einzelne Vorhersage zu erklären. SHAP (Shapley Additive exPlana-

tions) ist ein spieltheoretischer Ansatz zur Erklärung des Outputs eines ML-Modells. Dieser Ansatz verbindet die optimale Verteilung mit lokalen Erklärungen unter Verwendung der Shapley-Werte aus der Spieltheorie. Die SHAP-Methode wurde in das ExplainerDashboard [Di24b] integriert und steht damit als „Ready-To-Use"-Anwendung zur Verfügung.

Insgesamt soll die vorgeschlagene Simplex4Learning-Lösung Domänenexpert:innen bei den folgenden, aufeinander aufbauenden, Arbeitsschritten unterstützt werden:

1. Auswahl von als potentiell problemrelevant identifizierten Datensätzen für die Analyse,
2. effektive Sichtung der bestehenden Daten, die für ein Modelltraining mit Methoden des Maschinellen Lernens in Frage kommen, sowie Feature Engineering,
3. Auswahl von passenden generischen Modellarchitekturen, Anpassung dieser Modelle an die Fragestellung und Trainieren einer Schar von Modellen,
4. Bewertung der Modellleistung einer trainierten Modellschar, um das passendste Modell zu bestimmen,
5. Ablage und Beschreibung generalisierter Datenanalyse-Pattern in ein ML-Repository, um anderen Fachanwender:innen die spätere Übertragung auf strukturgleiche Fragestellungen mit anderen Daten zu ermöglichen,
6. Suche und Auswahl potentiell geeigneter ML-Pattern für die Nutzung in anderen Kontexten,
7. Instantiierung und Anwendung der ML-Pattern auf neue Daten.

Aufwendige Routineaufgaben wie Datenimport, Datenaufbereitung, Erstellen der Pipelines und Modellexport sollen hierbei zur Entlastung und Fehlervermeidung möglichst vollständig übernommen werden. Dazu wird die Technologie des Automated Machine Learning (AutoML) in Kombination mit Machine Learning Operations (MLOps) Ansätzen auf die mit der envVisio-Methode harmonisierten Umweltdaten angewandt und um Mechanismen der Erklärungsgenerierung und Entscheidungsunterstützung (Explanation-Dashboard) ergänzt. Die envVisio-Methode baut eine einheitliche Fachdatenstruktur im Sinne eines Data Warehouse auf (Simplex4Data Data Warehouse) und stellt flexible Datenschnittstellen für die ML-Operationen zur Verfügung. Dazu sollen im Simplex4Learning-Vorhaben die bestehende envVisio-Methode und die entsprechende Softwareplattform mit ihren Import- und Zugriffsschnittstellen um relevante Datentypen erweitern, insbesondere für das Messdatenmanagement aus der Umweltsensorik.

Zur Vereinfachung der breiten Nutzung ist weiterhin vorgesehen, dass die entwickelten Methoden und Techniken in einen Gesamtdemonstrator auf Basis der Disy Cadenza Plattform für Business & Location Intelligence integriert werden.

Ergänzend zu diesen technischen Entwicklungen muss überlegt werden, welche organisatorischen Rollen (wie zum Beispiel Data Scientist, Umwelt Analyst, …) mit welchen Aufgaben und welcher Werkzeugunterstützung in den zu betrachtenden Behörden realistisch erwartet werden können. In [Bu24] werden bereits als „Personas“ einige solche Rollen vorgestellt, die auf Disy’s Erfahrungen in Umweltverwaltungen beruhen.

Die Technologieentwicklungen im Simplex4Learning-Projekt sollen beispielgetrieben anhand realer Fragestellungen und Datenbestände der assoziierten Projektpartner erfolgen, dem Landesamt für Natur, Umwelt- und Verbraucherschutz Nordrhein-Westfalen (LANUV), der Landesanstalt für Umwelt Baden-Württemberg (LUBW) und dem Landesbetrieb Forst Brandenburg (LFB).

3 Forschungsfragen und erste Ergebnisse

Im Rahmen der Umsetzung des Projekts Simplex4Learning sollen folgende Fragen untersucht werden:

- Was sind konkrete Aufgabenstellungen für die ML-Nutzung in der deutschen Umweltverwaltung?
- Wie ist allgemein der Stand von Wissen, Zielen und Verwendung von ML in der deutschen Umweltverwaltung?
- Wie kann die envVisio-Methode effizient und effektiv für alle relevanten Arten von Umweltdaten genutzt werden, insbesondere Zeitreihendaten?
- Was ist ein praxistaugliches Gesamtszenario für den breiten Einsatz von ML in der Umweltverwaltung, technisch und organisatorisch?
- Was sind praxisrelevante ML-Anwendungen mit gutem Ergebnis- und Nutzenpotential, die im Vorhaben beispielhaft bearbeitet werden sollten?
- Wie kann die Lösung dieser Beispielfragestellungen durch AutoML- und MLOps-Tools vereinfacht und ihre Akzeptanz durch XAI-Methoden erhöht werden?
- Ist die Wiederverwendbarkeit vortrainierter ML-Modelle oder -Workflows in der Umweltpraxis realistisch? Wie ist ein ML-Repository zu gestalten, um Wiederverwendbarkeit und Übertragbarkeit zu ermöglichen?

Zum KIU-Workshop im Herbst 2024 werden einige Arbeitsergebnisse bereits vorliegen:

Ein Überblick zu weitverbreiteten AutoML-Frameworks und eine initiale Bewertung ihrer Eignung für die betrachteten Zwecke wurde bereits vorgenommen. Die AutoML-Tools Autokeras, Auto-Pytorch, H2O, AutoML, NNI, PyCaret, EvalML, MLJAR, FLAML und AutoGluon sowie die MLOps-Tools Mlflow, TensorFlow Extended, ClearML und Kubeflow wurden untersucht. Bewertungs- bzw. Vergleichskriterien

waren Python Support / R Support, Open-Source / kommerziell, On Premise Nutzung, GPU-Support, verfügbare ML-Verfahren, flexible Datenbanken, Kubernetes-Support / Docker-Support, Geodaten-Support (Raster, Vektor), Bedienbarkeit, Erweiterbarkeit, Zugriffsrechte / Keycloak, Monitoring, Dokumentation, Community und die Bewertung durch Dritte (Github). Die Untersuchung wurde überwiegend als Desk Research auf Basis frei verfügbarer Informationen durchgeführt. Einige Frameworks wurden dann auch testweise experimentell eingesetzt. Die einzelnen Bewertungen unterliegen natürlich einer gewissen Subjektivität und sind bei einer sich schnell entwickelnden Technologie auch immer nur eine Momentaufnahme. Außerdem ist die Gewichtung der Bewertungskriterien natürlich individuell unterschiedlich. So wurde für das vorliegende Forschungsprojekt mit dem Ziel einer einfachen späteren Verwertung durch die Simplex4Learning KMU-Partner besonderer Wert auf Open-Source Lösungen mit verwertungsfreundlichen Lizenzen gelegt. Für eine innerbetriebliche Nutzung durch einen Akteur, der gewisse kommerzielle Frameworks bereits im Einsatz hat, könnten sich hier auch ganz andere Präferenz ergeben. Mit den im Simplex4Learning gewählten Bewertungskriterien und -gewichtungen ergab sich zum Untersuchungszeitpunkt, dass (1) MLflow, (2) Kubeflow und (3) ClearML zu präferierende MLOps-Lösungen für das Vorhaben sind und (1) NNI ein hervorragendes AutoML-Framework, sowie (2a) Autokeras bei der Nutzung von Deep Learning Tools bzw. (2b) PyCaret bei der Nutzung symbolischer ML-Algorithmen empfehlenswert sind. Als zu präferierende Gesamtlösung für das Vorhaben bot sich hieraus die Kombination von MLflow mit Autokeras und PyCaret an.

Um die Anforderungen aus der Praxis zu erheben und Ideen für geeignete Beispielanwendungen zu erzeugen, wurden mehrere Workshops mit den assoziierten Partnern durchgeführt, inklusive einer Bestandserhebung verfügbarer Datenbestände. Für die Anwender-Workshops wurde eine themenspezifische Vorgehensweise auf der Basis von Brainwriting entwickelt und realisiert. Diese ist in [Fu24] detailliert beschrieben. Einige erste Beobachtungen aus den ersten Workshops mit Anwender:innen, noch ohne formelle Evaluation, nur subjektive Eindrücke als erste Zusammenfassung der Arbeit mit den Anwendern:

- **Große Sprachmodelle** (Large Language Models) haben inzwischen in der allgemeinen Wahrnehmung bereits solche Bedeutung erlangt, dass Anwender:innen sofort verchiedene Use Cases aus ihrem Arbeitsbereich einfallen, von der automatischen natürlichsprachigen Beschreibung von Grafiken und Kartendarstellungen zwecks Barrierefreiheit über die Erzeugung einfacher Sprache aus Fachtexten bis zu einfachen Zusammenfassungen großer Antwortmengen oder Dokumentbestände.
- Klassische Aufgaben der **Bilderkennung**, also der Identifikation semantischer Entitäten aus Luftbildern, Satelliten- oder Drohnen-Fernerkundungsdaten, könnte ebenfalls an vielen Stellen bestehende Arbeitsabläufe vereinfachen oder beschleunigen und Informationslagen verdichten, sei es bei der Erkennung von Bauwerken oder Infrastrukturelementen, der Verkehrsstromanalyse bis zu sehr speziellen und

auch kleinräumigen Anwendungen wie dem Erkennen einzelner Baumarten in Forstbeständen oder von Mikroorganismen im Wasser in Mikro-Fotografien. Hier ist ein kritischer Punkt häufig die Verfügbarkeit ausreichender Mengen von Trainingsdaten.

- Die intelligente Sensordatenanalyse, namentlich in Form der **Anomalieerkennung auf Zeitreihen**, kann in verschiedensten fachlichen Domänen Anwendung finden, insbesondere bei Luft- oder Wasserqualitätsdaten. Hier sind Ausreißer zunächst zu identifizieren, idealerweise auch vermutliche Ursachen zu finden (akuter Sensordefekt oder langsamer Sensorqualitätsabfall) oder gar Prognosen und/ oder Frühwarnungen zu erzeugen.
- **Ausbreitungsanalysen**, insbesondere für invasive Tier- oder Pflanzenarten, aber auch von Tier- oder Pflanzen-Schädlingen oder -Krankheiten, sind eine Thematik, die zunehmende Bedeutung gewinnt.
- **Korrelations-** und im besten Fall auch **Kausalitätsanalysen** (beschreibende Mustererkennung mit erklärenden Analysen) von Phänomenauftreten und insbesondere von Umweltzuständen aus möglichen Umweltbelastungen stehen natürlich zentral im Aufgabenbereich aller mit Umweltthemen befassten Behörden. Hierzu zählt beispielsweise auch die Bewertung der Wirkung von Maßnahmen oder die Verursacheranalyse bei bestimmten Vorfällen.

Diese Zusammenfassung beschreibt bestimmte Cluster häufig wiederkehrender und mit großer Bedeutung versehener Aufgabenstellungen, die natürlich im Einzelfall noch konkretisiert werden können. Im weiteren Vorhabenverlauf werden einige dieser Frage- bzw. Aufgabenstellungen exemplarisch für konkrete Instantiierungen bearbeitet werden. Aber auch die Sammlung der generischen Themen und passender Lösungsansätze scheint uns einen Mehrwert für die Anwendung zu versprechen.

4 Zusammenfassung und Ausblick

Wir haben die Motivation und die zentrale Zielsetzung sowie den groben Lösungsansatz im Forschungsvorhaben Simplex4Learning beschrieben. Grundsätzlich geht es um die Konzeption, prototypische Entwicklung und intensive Demonstration und Erprobung einer Machine-Learning Plattform für Umweltdaten. Diese entsteht durch aufgabenspezifische Kombination und Zusammenführung von Ansätzen aus der Datenintegration/-harmonisierung, dem automatisierten Maschinellen Lernen und der erklärbaren KI. Die zu entwickelnde Plattform soll das Verständnis bereits existierender Datenbestände verbessern, eine effiziente, flexible und kostengünstige Analyse ermöglichen und somit die Potenziale des ML für die Verwaltung besser ausschöpfen. Die Einbindung in bestehende, in der Verwaltung verbreitete, Werkzeuge soll den Praxiseinsatz fördern. Bei der Gesamtarchitektur sollen auch organisatorische und institutionelle Aspekte berücksichtigt werden, die sich in der Frage widerspiegeln, welche Nutzerrollen mit

welchem Tooling welche Aufgaben bearbeiten können/ sollen. Das Vorhaben ist anspruchsvoll und komplex. Daher sollen anhand konkreter Beispiele aus der Praxis schrittweise praktikable Lösungsansätze gesucht werden.

Zum aktuellen Zeitpunkt werden umfangreiche Anforderungserhebungen mit den drei assoziierten Partnern durchgeführt [Fu24], initiale Tool-Bewertungen und Experimente zu AutoML und MLOps gemacht und eine erste Softwarearchitektur entworfen. Für disy Cadenza soll die neue Analytics-Schnittstelle zur Verwendung externer Advanced-Analytics-Algorithmen genutzt bzw. (weiter-)entwickelt werden. Simplex4Data wird bereits für den effizienten Umgang mit großen Zeitreihendaten ertüchtigt. Eine laufende Umfrage bei verschiedenen weiteren Behörden in Deutschland und Österreich im Rahmen einer studentischen Abschlussarbeit soll die Status-Quo-Erhebung und Anforderungsanalyse vervollständigen.

Danksagungen

Das Verbundprojekt simplex4learning („Intelligente Umweltdatenanalyse durch automatisiertes maschinelles Lernen für Fachanwender“) wird vom Bundesministerium für Bildung und Forschung (BMBF) im Rahmen der Fördermaßnahme „KMU-innovativ IKT“ unter den Förderkennzeichen 01IS23041A bis 01IS23041C finanziell unterstützt. Das Projekt wird gemeinsam durchgeführt von der disy Informationssysteme GmbH in Karlsruhe, der simplex4data GmbH in Dresden und der Hochschule für Technik und Wirtschaft Berlin (HTW Berlin). Das Projekt wird ebenfalls unterstützt von den assoziierten Projektpartnern, dem Landesamt für Natur-, Umwelt- und Verbraucherschutz Nordrhein-Westfalen (LANUV), der Landesanstalt für Umwelt Baden-Württemberg (LUBW) und dem Landesbetrieb Forst Brandenburg (LFB). Die Untersuchung und Bewertung existierender AutoML- und MLOps-Tools wurde von Herrn Gerrit Tombrink (https://geolinked.de/) im Rahmen einer freien Mitarbeit erstellt.

Literaturverzeichnis

[Al24] Allegro AI: Build Better AI At Any Scale, Faster. URL: clear.ml, Stand: 29.07.2024

[Be24] Behrens, G. et al.: Prognose von Pegelständen mit Methoden des Maschinellen Lernens und frei verfügbaren Daten. In Fuchs-Kittowski, F. et al. (Hrsg.) UIS-2024: 31. Workshop Umweltinformationssysteme - Digitalisierung für eine nachhaltige Planetare Zukunft. Springer Vieweg, Wiesbaden, 2024.

[Bi20] Bizety Staff: Open Source AutoML Tools: AutoGluon, TransmogrifAI, Auto-sklearn, and NNI. URL: https://www.bizety.com/2020/06/16/open-source-automl-tools-autogluon-transmogrifai-auto-sklearn-and-nni/, Stand: 29.07.2024.

[Bi22] Bie, T. de et al.: Automating Data Science. Communications of the ACM 65(3), S. 76–87, 2022.

[Bu24] Budde, M. et al.: NiMo 4.0 – Enabling Advanced Data Analytics With AI for Environmental Governance in the Water Domain. at – Automatisierungstechnik, 72(6), S. 564-578, Walter de Gruyter, Berlin, Boston, 2024.

[Di24] Disy Informationssysteme GmbH: Software für Datenanalyse, Business & Location Intelligence. URL: https://www.disy.net/de/produkte/cadenza/datenanalyse-software/, Stand: 29.07.2024

[Di24b] Dijk, O.: Explainer Dashboard. URL: https://github.com/oegedijk/explainerdashboard, Stand: 29.07.2024.

[Fe21] Ferreira, L. et al.: A Comparison of AutoML Tools for Machine Learning, Deep Learning and XGBoost. In 2021 International Joint Conference on Neural Networks (IJCNN), IEEE, 2021.

[Fe15] Feurer, M. et al.: Efficient and Robust Automated Machine Learning. Advances in Neural Information Processing Systems, 28, 2015.

[Fu24] Fuchs-Kittowski, F. et al.: Eine Methode für die Potenzialanalyse zur Identifikation von Anwendungsszenarien für Maschinelles Lernen - Fallstudie in einem Landesumweltamt. In KIU-2024, Gesellschaft für Informatik e.V., 2024.

[Go22] Gotsch, M. et al.: Der Beitrag von Big Data, KI und digitalen Plattformen auf dem Weg zu einer Green Economy - Einsatzbereiche und Transformationspotenziale. UBA Texte 85/2022. Umweltbundesamt, Dessau-Roßlau, 2022.

[Gr22] Großmann, J.: envVisio Service – ein universeller Dienst für Umweltdaten. In: Fuchs-Kittowski, F.; Abecker, A.; Hosenfeld, F. (Hrsg.): Umweltinformationssysteme – Wie trägt die Digitalisierung zur Nachhaltigkeit bei? Springer Vieweg, Wiesbaden, 2022.

[GR23] Großmann, J.; Rudolf, H.: Fachsysteme, Schemaevolution, Datenharmonisierung - Perspektiven für ein neuartiges und effektives Datenmanagement. In: Fuchs-Kittowski, F.; Abecker, A.; Hosenfeld, F. (Hrsg.): Digitalisierung im Zeichen des Klimawandels und der Energiewende. Springer Vieweg, Wiesbaden, 2023.

[He20] Heller, M.: MLops-Plattformen im Überblick. Computerwoche, Blog, 09/2020. URL: https://www.computerwoche.de/a/mlops-plattformen-im-ueberblick,3549851, Stand: 29.07.2024

[HKV19] Hutter, F.; Kotthoff, L.; Vanschoren, J. (Hrsg.): Automated Machine Learning: Methods, Systems, Challenges. The Springer Series on Challenges in Machine Learning. Springer Nature, 2019.

[KSS16] Katz, G.; Shin, E.C.R.; Song, D.: Explorekit: Automatic Feature Generation and Selection. In Proceedings of the 2016 IEEE 16th International Conference on Data Mining (ICDM'16). IEEE, S. 979–984, 2016.

[KMP17] Kaul, A.; Maheshwary, S.; Pudi, V.: AutoLearn - Automated Feature Generation and Selection. In Proceedings of the 2017 IEEE International Conference on Data Mining (ICDM'17). IEEE, S. 217–226, 2017.

[Ka23] Karimanzira, D. et al.: Application of Machine Learning And Deep Neural Networks for Spatial Prediction of Groundwater Nitrate Concentration to Improve Land Use Management Practices. Frontiers in Water, 5-2023, 2023.

[KV15] Kanter, J.M.; Veeramachaneni, K.: Deep Feature Synthesis: Towards Automating Data Science Endeavors. In Proceedings of the IEEE International Conference on Data Science and Advanced Analytics (DSAA'15). IEEE, S. 1–10, 2015.

[Kl21] Klein, S.: Datenbasierte Anwendungsszenarien im grünen Sektor und im Umweltschutz. DigitalAgentur Brandenburg (DABB), Potsdam, 2021.

[Ko17] Kotthoff, L. et al.: Auto-WEKA 2.0: Automatic Model Selection and Hyperparameter Optimization in WEKA. Journal of Machine Learning Research 18(25), S. 1−5, 2017.

[KKH22] Kreuzberger, D.; Kühl, N.; Hirschl, S.: Machine Learning Operations (MLOps): Overview, Definition, and Architecture. ePrint: arXiv:2205.02302, 2022.

[LL17] Lundberg, S.M.; Lee, S.I.: A Unified Approach to Interpreting Model Predictions. In Proceedings of the Advances in Neural Information Processing Systems, S. 4765–4774, 2017.

[Me24] Metaflow: A Framework for Real-life ML, AI, and Data Science. URL: https://metaflow.org/, Stand: 29.07.2024.

[Ml24] Mlflow Project: ML and GenAI made simple. URL: https://mlflow.org/, Stand: 29.07.2024

[Mo22] Molnar, C.: Interpretable Machine Learning: A Guide For Making Black Box Models Explainable. 2nd ed., 2022.

[RSG16] Ribeiro, M.T.; Singh, S.; Guestrin, C.: Why Should I Trust You? Explaining The Predictions of Any Classifier. In Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, S. 1135–1144, 2016.

[Ru18] Rudolf, H.: Umweltdatenmanagement: Eine Geo-Inspiration. Bernhard Harzer Verlag GmbH, Karlsruhe, 2018.

[Ru21] Rudolf, H.: Umweltdaten-Intelligenz - envVisio mit neuen Ansätzen im Umweltdatenmanagement: modelltheoretisch hergeleitet, fachlich ausgearbeitet, praktisch umgesetzt. In: Freitag, U. et al. (Hrsg.) Umweltinformationssysteme – Wie verändert die Digitalisierung unsere Gesellschaft? Springer Vieweg, Wiesbaden, 2021.

[Sk20] Skogström, H.: The MLOps Stack. URL: henrikskogstrom.medium.com/the-mlops-stack-31db61ea1c5b, 2020. Stand: 29.07.2024

[So22] Sokhi, R. et al: Advances in Air Quality Research – Current and Emerging Challenges. Atmospheric Chemistry and Physics 22(7) S. 4615-4703, European Geosciences Union (EGU), 2022.

[TRB21] Tremper, P.; Riedel, T.; Budde, M.: Spatial Interpolation of Air Quality Data with Multidimensional Gaussian Processes. In INFORMATIK 2021, 2. Workshop Künstliche Intelligenz in der Umweltinformatik (KIU-2021), S. 269-286. Gesellschaft für Informatik, Bonn, 2021.

[WLG24] Wunsch, A.; Liesch, T.; Goldscheider, N.: Towards Understanding the Influence of Seasons on Low-Groundwater Periods Based on Explainable Machine Learning. Hydrology and Earth System Sciences, 28, S. 2167-2178, 2024.