IPRI INTERNATIONAL PERFORMANCE RESEARCH INSTITUTE

IPH Institut für Integrierte Produktion Hannover

LEITFADEN ZUR AUFBEREITUNG VON DATEN FÜR DIE ANWENDUNG VON MACHINE LEARNING

Bundesministerium für Wirtschaft und Klimaschutz

AiF

Im Rahmen des Forschungsprojekts MLready (https://ipri-institute.com/forschungsprojekte/mlready/) wurde dieser Leitfaden entwickelt.

Er soll kleine und mittelständische Unternehmen im produzierenden Gewerbe dazu befähigen, eigenständig Datenquellen zu identifizieren und Daten für die Verwendung von Machine Learning vorzubereiten.

Im folgenden wird eine Checklist der Schritte definiert, unten finden sich zusätzliche Informationen zu den Schritten.

Checkliste

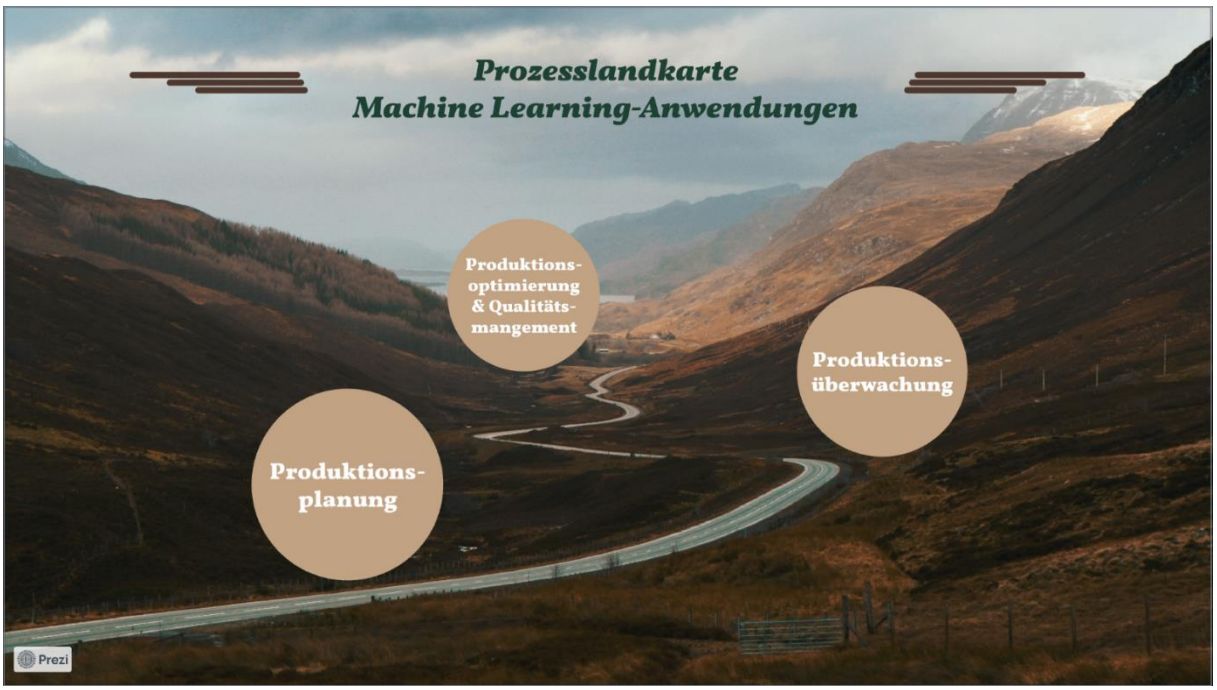
- ☐ Einen Anwendungsfall definieren
- ☐ Daten auswählen
- ☐ Datenqualität bemessen
- ☐ (Bei Bedarf) Datenqualität verbessern

Fragen, Anmerkungen und Verbesserungsvorschläge können jederzeit an Herrn Garlef Hupfer am IPRI (ghupfer@ipri-institute.com) gesendet werden. Wir helfen Ihnen gerne bei der Anwendung dieses Leitfadens!

1. Auswahl eines Machine Learning Anwendungsfalls

Daten müssen spezifisch für den geplanten Anwendungsfall ausgewählt werden. Daher muss zunächst ein Anwendungsfall definiert werden.

Im Projekt wurde die Prozesslandkarte der Machine Learning-Anwendungsfälle entwickelt. Unternehmen können sich hier (https://prezi.com/view/6074HVoARB98SFgSAxdA/) einen Anwendungsfall auswählen.



2. Auswahl der Datengrundlagen

In der Prozesslandkarte sind für jeden Anwendungsfall Datengrundlagen definiert, die für den jeweiligen Anwendungsfall benötigt werden bzw. unterstützend wirken können.

Optional: Zusätzlich können weitere Datenquellen mit Hilfe des morphologischen Kastens identifiziert werden (s. nächste Seite).

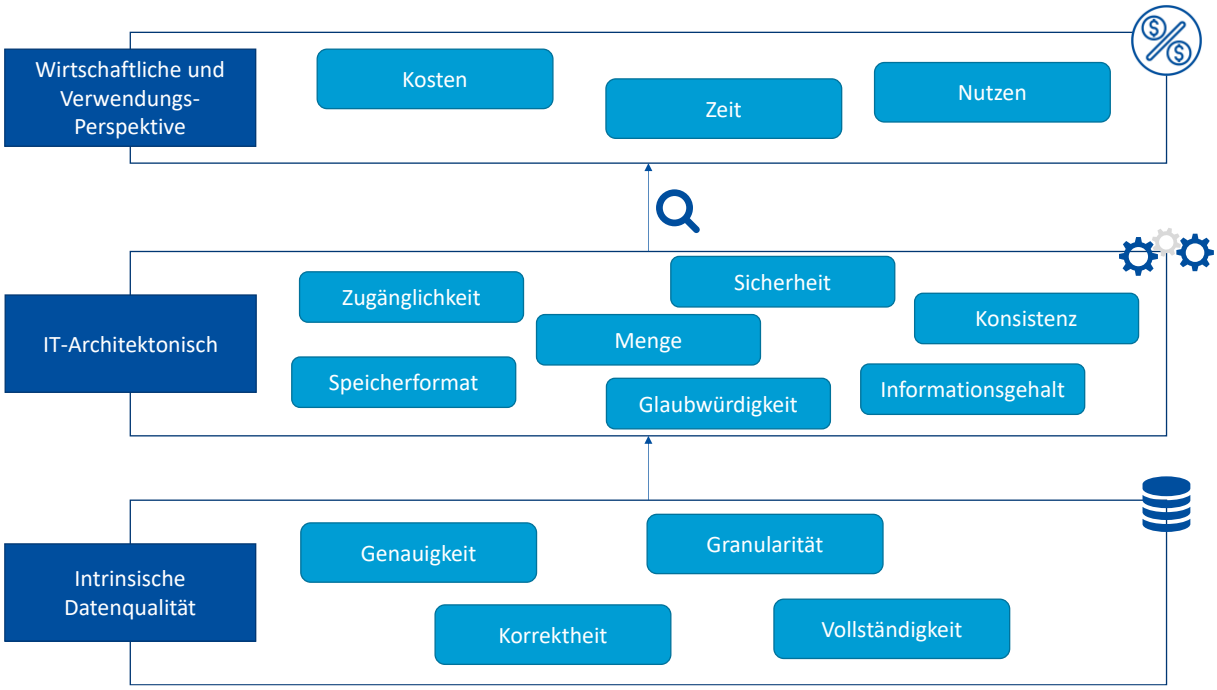
Durch die Kombination von Attributen von Datensätzen sollen Ideen für weitere Datenquellen geschaffen werden.

Merkmal	Ausprägungen						
Herkunft Werden Daten intern erhoben oder von extern zugezogen?	Intern				Extern		
Strukturiertheit Liegen die Daten maschinell verwertbar vor?	Umstruktuiert			Semi-strukturiert			Strukturiert
Datenformat In welchem Format liegen die Daten vor?	handschriftlich	Textdateien	Bilddateien	json, xml etc.	csv, xls oder Datenbank-formate	Audio-dateien	Videodateien
Veränderlichkeit Wie oft werden Daten erhoben?	Sekündlich		Minütlich		Stündlich	Täglich	(>) wöchtenlich
Zugänglichkeit Kann direkt auf die Daten zugegriffen werden?	Im eigenen System		Mit Zugriffsbeschränkungen		Frei im Internet zugänglich	Kostenpflichtig	Unzugänglich
Volumen Wie viele Daten sind verfügbar?	Einzelfälle			Regelmäßige Dokumentation			Hochfrequente Dokumentation über längere Zeit
Abhängigkeiten Können Daten unabhängig verwendet werden?	nur in Verbindung in komplexen Systemen aussagekräftig			in Verbindung mit wenigen anderen Datensätzen aussagekräftig			kann eigenständig verwendet werden

3. Analyse der Datenqualität

Garbage in – garbage out. Für die Anwendung von Machine Learning muss die Datenqualität *ausreichend* gut sein.

Um abzuschätzen, ob die Datengrundlage für die Anwendung ausreicht, wurde ein Datenqualitätsmodell entwickelt, nachdem die Qualität von Daten bemessen werden kann. Dieses Modell sollte für jede der ausgewählten Datengrundlagen angewendet werden.



Intrinsische Datenqualität

Bezeichnet Faktoren, die bemessen, ob und wie die erhobenen Daten die Realität abbilden.

- Die **Genauigkeit** der Daten ist dann gegeben, wenn jede erhobene Information jeweils genau einem Wert in den Daten zugeordnet ist.
- Daten sind **korrekt**, wenn der jeweilige Wert korrekt gemessen und demenstsprechend in die Datenbank eingetragen wurde.
- Daten sind **vollständig**, wenn alle für den jeweiligen Nutzen erforderlichen Informationen in die Datenbank aufgenommen wurden und fehlende Werte minimiert werden.
- Die **Granularität** bezeichnet die Kleinteiligkeit der Daten, ob also Informationen auf kleinteiliger Ebene enthalten sind.

IT-Architektonisch

Bezeichnet Faktoren hinsichtlich der Verwendung und Verwendbarkeit der Daten (ob die Daten für die jeweilige Verwendung geeignet sind).

- Daten sind **zugänglich**, wenn sie für Nutzer möglichst klar und schnell abrufbar sind.
- Das **Speicherformat** bezeichnet die informationstechnische Speicherung der Daten und beinhaltet unter Effizienz im Speichervolumen und die maschinelle Interpretierbarkeit.
- Daten sind erst dann von hoher Qualität, wenn bei ihrer Erhebung, Speicherung und Weiterverarbeitung **Sicherheitsstandards** eingehalten werden. Zum Beispiel müssen die Daten vor Manipulationssversuchen sicher sein und die Informationsweitergabe darf nur an Befugte erfolgen.
- Daten sing **glaubwürdig**, wenn bei der Informationserhebung bestimmte Qualitätsstandards eingehalten werden, nach denen die Nutzer der Korrektheit der Daten vertrauen können.
- **Konsistenz** bedeutet, dass Daten immer in gleicher Form sowie regelmäßig und zeitlich gleichmäßig erhoben und gespeichert werden.
- Das richtige **Datenvolumen** bzw. die geeignete **Menge** der Daten ist dann gegeben, wenn die Quantität der Daten für die jeweilige Nutzung der Daten ausreicht.
- Die Qualitätsanforderungen des **Informationsgehalts** beziehen sich darauf, ob die erhobenen Informationen einen tatsächlichen Nutzen haben, für andere Anwendungen verwertet werden und somit ob sie nützlich sind.

Verwendung

Bezeichnet die Sinnhaftigkeit der Verwendung der Daten (ob der Aufwand der Erhebung, Verarbeitung und Verwendung der Daten den Nutzen rechtfertigt).

- Die Datenspeicherung und Datenpflege kann für das Unternehmen mit hohen finanziellen **Kosten** einhergehen. Bei der Verwendung der Daten muss der Aufwand dem Nutzen angemessen sein.
- Für eine hohe Datenqualität müssen die Daten für den jeweiligen Nutzen **aktuell** sein, das Alter der Daten muss für den jeweiligen Nutzen sinnvoll sein.
- Der **Nutzen** von Daten ergibt sich aus dem umgesetzten Potenzial der Anwendung nach Erhebung, Instandhaltung und Verarbeitung der Daten.

4. Verbesserung der Datenqualität

Durch die Anwendung des Datenmodells werden mögliche Defizite der Datengrundlage aufgedeckt. Zur Behebung sollte ein langfristig nachhaltiges Datenqualitätsmanagement aufgebaut werden, die diese Defizite bereinigen. Im Folgenden werden einige generelle Richtlinien angeschnitten.

Geschäftsprozessoptimierung: Die Abläufe in einem Unternehmen werden durch die Geschäftsprozesse festgelegt. Diese Abläufe sind grundlegend und werden aufgrund ihrer Komplexität nur ungern verändert. Dennoch kann es notwendig sein, in die Hauptabläufe des Unternehmens einzugreifen, um die Qualität der Daten sicherzustellen. So kann eine Adaption bestehender Prozesse notwendig sein, um langfristig die Fehler in der Datenqualität zu beheben.

Systemoptimierung: Um die Datenqualität zu steigern können zudem die verwendeten Informationssysteme optimiert werden. Dies umfasst bspw. die Anpassungen des Datenmodells, die Einführung einheitlicher Datenbezeichnungen, die Einführung von verschiedenen Constraints sowie die Überarbeitung von Geschäftsprozessapplikationen. Dadurch werden Fehler bei der Datenerfassung, -speicherung und -nutzung vermieden.

Datenbereinigung: Die Verbesserung der Datenqualität beinhaltet die Bereinigung von identifizierten Fehlern, entweder automatisiert oder manuell durch Mitarbeitende. Um eine permanente Datenqualität zu wahren ist die Integration von Qualitätsmessungen in die Geschäftsprozesse erforderlich.

Schulung von Mitarbeitenden: Die Schulung der Mitarbeitenden ist ein weiterer Aspekt, welcher zur Verbesserung der Datenqualität beitragen kann. Neben technischen Maßnahmen ist es wichtig, sie durch klare Arbeitsanweisungen zu unterstützen. Schulungen informieren die Mitarbeitende darüber, wie sie mit den Daten umgehen sollen und mit ihrer Arbeit zur unternehmensweiten Datenqualität beitragen können.

Das IGF-Vorhaben 22312 N der Forschungsvereinigung Institut für Energie- und Umwelttechnik e.V. – IUTA wird über die AiF im Rahmen des Programms zur Förderung der industriellen Gemeinschaftsforschung (IGF) vom Bundesministerium für Wirtschaft und Klimaschutz (BMWK) aufgrund eines Beschlusses des Deutschen Bundestages gefördert.