

# Data Scientist: Klassifikation relevanter Apps für Life Sciences

## Ziel der Aufgabe

Entwickle einen Python-basierten Klassifikator in einem One-Shot- / Low-Resource-Setting, um zu bestimmen, welche Apps für die Branchen Life Science, MedTech und Pharma relevant sind oder welche Apps GxP relevant sein könnten.

## Anforderungen

- Verwende ein Jupyter Notebook oder ein vergleichbares Tool.
- Nutze die im Anhang bereitgestellten Beispieldaten als Grundlage.
- Entwickle mindestens zwei unterschiedliche Klassifikationsansätze (z. B. ML-basiert, Vektorraum-Ansätze).
- Annotiere und verifiziere die Daten
- Visualisiere deine Ergebnisse (z. B. Confusion Matrix, Feature-Importances, Embedding-Plots).
- Identifiziere und erläutere die relevantesten Features für die Klassifikation.
- Dokumentiere deine Strategie, Methodenwahl und Annahmen klar.

## Optionale Tools

- pandas, numpy, scikit-learn
- fasttext, word2vec
- pydantic, weaviate (für strukturierte Daten oder Vektor-Suche)

## Zusätzliche Empfehlungen

- Nutze ggf. Transfer Learning oder Embedding-Techniken, um mit wenigen Beispielen zu arbeiten.
- Begründe deine Modellwahl (z. B. warum ein bestimmter Klassifikator für dieses Setting geeignet ist).
- Achte auf Nachvollziehbarkeit und Reproduzierbarkeit deines Codes.