**Title: Dataset Citation Classifier for Scientific Literature**

**1. Introduction**

Machine Learning (ML) is a subset of artificial intelligence that enables computers to learn patterns and make decisions without explicit programming. Deep Learning (DL), a specialized branch of ML, uses artificial neural networks with multiple layers to model high-level abstractions in data. In natural language processing (NLP), deep learning has enabled models to understand and classify human language in tasks such as sentiment analysis, named entity recognition, and question answering.

This project focuses on dataset citation detection and classification in scientific papers, based on the “Make Data Count - Finding Data References” Kaggle competition. Scientific data citations are typically buried in unstructured text and vary in format, making them hard to detect using rule-based methods. By applying machine learning techniques, we aim to identify references to research datasets and classify them as either “Primary” or “Secondary” citations.

**2. Literature Review**

Data citation plays a crucial role in enabling reproducible research and recognizing data as a first-class research output. However, research has shown that up to 86% of scientific data remains uncited (Peters et al., 2016), mainly due to the lack of standardized citation practices and automated tools.

The Make Data Count initiative promotes open data citation standards and developed the MDC Data Citation Corpus. Previous attempts to automate citation recognition used rule-based heuristics or manually curated reference patterns. However, these methods fail to generalize across diverse scientific disciplines and formats.

Modern NLP methods using machine learning can automatically extract dataset mentions from full-text papers. Techniques such as TF-IDF vectorization, Logistic Regression, and ensemble models have proven effective in text classification. Our project builds upon this by developing a supervised classifier trained on labeled citation examples from the MDC dataset.

**3. Exploratory Data Analysis (EDA)**

The competition dataset includes:

* train\_labels.csv: Contains article IDs, dataset identifiers (DOI or Accession ID), and citation types (Primary/Secondary)
* train/: Contains XML and PDF full-text versions of scientific articles
* test/: Contains test articles for submission evaluation

Each dataset citation is either:

* **Primary**: Data generated and used in the study
* **Secondary**: Previously published data reused in the study

Sample XML and DOI mention:

<article>  
 <abstract>The data used in this publication can be accessed from Dryad at doi:10.5061/dryad.6m3n9</abstract>  
</article>

After parsing XMLs, we extracted paragraphs containing dataset mentions. Citations with missing or invalid dataset IDs were filtered out. The dataset was imbalanced, with more Secondary citations than Primary.

**4. System Architecture**

The system consists of the following pipeline:

1. **Data Preprocessing**:
   * Parse XML to extract article text
   * Search for known dataset identifiers in context
   * Normalize DOI/Accession ID format
2. **Feature Extraction**:
   * Clean text (lowercase, remove punctuation)
   * Convert context to TF-IDF vectors
3. **Model Training**:
   * Use TfidfVectorizer + LogisticRegression in a scikit-learn pipeline
   * Train on labeled dataset from train\_labels.csv
4. **Model Evaluation**:
   * Evaluate using F1-score on validation set
   * Handle class imbalance via stratified sampling
5. **App Deployment**:
   * Build a Streamlit app
   * Load model + vectorizer with joblib
   * User inputs a paragraph → app detects datasets → predicts citation type

**5. Model Evaluation and Implementation**

We trained a Logistic Regression classifier with TF-IDF features. After filtering valid citations, the model was evaluated on a holdout set:

precision recall f1-score support  
 Primary 0.44 1.00 0.61 7  
 Secondary 1.00 0.74 0.85 34  
 Accuracy 0.78 41

This showed good generalization, especially on the dominant Secondary class. The final model was submitted to Kaggle and achieved a **public leaderboard score of 0.021**, ranking **#9** at the time of submission.

### Streamlit App

The classifier was integrated into a live web app using Streamlit:

* Hosted via Streamlit Community Cloud
* Repo: <https://github.com/kristy999/Keggle-Competition>
* Demo URL: https://kristy999-keggle-competition.streamlit.app

Users can paste a paragraph from a scientific paper. The app scans for dataset mentions (DOIs or IDs like GSE12345) and classifies each as Primary or Secondary.

**6. Conclusion**

This project demonstrates the effectiveness of machine learning in solving real-world text classification problems. We developed a model to identify and classify dataset references in scientific papers, contributing to open science infrastructure. By deploying the model in an interactive app, we enable researchers to explore citation types and improve dataset discoverability.

Future improvements could include:

* Using transformer models (e.g., BERT) for context-aware predictions
* Expanding to multilingual scientific texts
* Connecting to live APIs (e.g., DataCite) for validation

Overall, this project highlights how AI can support data citation tracking, reproducibility, and research impact analysis in scientific communication.

**References**

* Peters, I., Kraker, P., Lex, E., Gumpenberger, C., & Gorraiz, J. (2016). Research data explored: an extended analysis of citations and altmetrics. *Scientometrics*, 107(2), 723–744. <https://doi.org/10.1007/s11192-016-1887-4>
* Make Data Count. (2025). Finding Data References. Kaggle. <https://www.kaggle.com/competitions/make-data-count-finding-data-references>
* Scikit-learn: Machine Learning in Python. <https://scikit-learn.org>
* Streamlit Docs. <https://docs.streamlit.io>