Lab 4: Data-Centric AI vs Model-Centric AI

This lab explores two contrasting approaches to improving machine learning models: **Data-Centric AI** and **Model-Centric AI**. The focus is on demonstrating the power of improving data quality to achieve superior results, even with standard models.

Student Name	ID	Section
Lujain Bukassim Almarri	S20106753	1

Objective

The lab aims to:

1. **Understand the Problem**:
    - Train a classifier to predict sentiment (good or bad) for product reviews in the magazine category.
    - Highlight limitations of focusing solely on model performance.
2. **Explore Data-Centric AI**:
    - Identify and address issues in the dataset.
    - Demonstrate how cleaning data impacts performance.
3. **Compare Approaches**:
    - Use model-centric techniques like hyperparameter tuning and advanced classifiers.

- Evaluate the results of data cleaning versus model tuning.

Steps in the Lab

1. Baseline Model Training

- **Dataset**: Reviews labeled as "good" or "bad."
- **Model**: Support Vector Machine (SVM) using **TF-IDF** for text representation.
- **Outcome**: Achieved an initial accuracy of **76.5%**, highlighting room for improvement.

2. Model-Centric Approach

- Tested different classifiers:
  - **Naive Bayes Classifier**: Improved accuracy to **85.3%**.
  - **Random Forest**: Poor performance with **49.8%** accuracy.
- Tried hyperparameter tuning and ensemble methods but faced diminishing returns.

3. Identifying Data Issues

- Examined sample data:
  - Found mislabeled examples (e.g., a negative review labeled "good").
  - Detected noisy data (e.g., HTML tags in reviews).

- Observed that data quality significantly affected model performance.

4. Data-Centric Approach

- Implemented a heuristic to identify noisy data (e.g., reviews with HTML tags).
- Removed problematic reviews to create a **cleaned dataset**.
- Retrained the baseline model using the cleaned dataset.

Results

- Accuracy with the cleaned dataset improved dramatically to **97%**, far exceeding improvements achieved through model-centric approaches alone.

Key Takeaways

1. **Data Quality Matters**:
    - Cleaning and preprocessing the data can yield more significant performance gains than model tuning.
    - Identifying and fixing mislabeled or noisy data is critical.
2. **Model-Centric Limitations**:
    - Optimizing models or trying complex algorithms has diminishing returns when working with low-quality data.

3. **Practical Insights**:

   - Focus on data-centric methods for real-world projects where data quality issues are common.

   - Combine data and model-centric approaches for optimal results.