



# Reading Report 1

CleanML: A Study for Evaluating the Impact of Data Cleaning on ML Classification Tasks

Shubham Sanjay Darekar USC ID - 1641138809



### **Overview**

#### Scope:

The paper analyzes 12 real-world datasets, each having 5 common errors among, missing values, outliers, duplicates, inconsistencies and mislabels

### Methodology:

Using various cleaning techniques to these datasets and evaluated their impact on 7 ML algorithms: logistic regression, decision trees, random forests, SVM, KNN, neural networks, and gradient boosting.

#### Approach:

The study employed careful experimental design, using statistical hypothesis testing to control for randomness and the Benjamini-Yekutieli procedure to manage false discovery rates.

### Impact

The study revealed that cleaning doesn't universally improve ML model performance and sometimes negatively impacts the model performance.

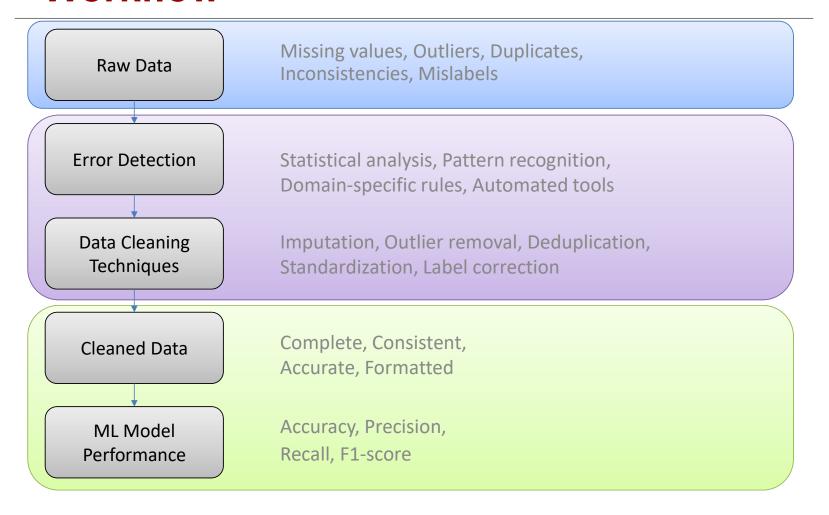


## **Research questions Answered**

- Conduct a first systematic empirical study on the impact of data cleaning on downstream ML classification models, for different error types, cleaning methods, and ML models
- Given their empirical findings, provide a starting point for future research to advance the field of cleaning for ML.



## Workflow





## **Pros of CleanML Research**

#### Real-world Relevance:

The study uses real-world datasets with actual errors, making the findings more applicable to practical scenarios.

### Multiple Cleaning Algorithms:

CleanML incorporates various cleaning algorithms, including both common solutions used in practice and state-of-the-art academic proposals.

#### Holistic Approach:

The study considers the entire ML pipeline, from data cleaning to model training and evaluation, providing a more comprehensive view of the impact of data quality on ML performance.

#### Reproducibility:

The study provides open-source code and experimental results, enabling other researchers to reproduce and build upon their work



## Cons of CleanML Research

#### Potential Bias:

The selection of datasets, error types, and cleaning methods, while extensive, may not cover all possible scenarios, potentially introducing some bias in the results.

#### Complexity:

The multifaceted nature of the study, involving various datasets, error types, and cleaning methods, may make it challenging to draw simple, generalizable conclusions.

#### Limited Consideration of Advanced Techniques:

While the study includes some state-of-the-art cleaning solutions, it may not fully capture the latest advancements in ML-oriented or semi-supervised cleaning methods.



# **Error Types Examined**

- Missing Values
  - Impact varied across datasets; imputation not always beneficial
- Outliers
  - Removal sometimes improved model performance, but not consistently
- Duplicates
  - Elimination generally improved model efficiency, but exceptions existed
- Inconsistencies
  - Correction had mixed effects, depending on the specific dataset and model
- Mislabels
  - Addressing mislabels often led to significant performance improvements



## **Quiz questions**

- What are the five error types studied in the CleanML benchmark?
  - Missing values
  - Outliers
  - Duplicates
  - Inconsistencies
  - Mislabels
- What factors influence ML model performance after cleaning the data?
  - Types of errors present in dataset
  - Cleaning techniques used