A closer look at Machine Learning Pipelines in Python and SQL

Christian Fuchs

14 June 2023

Outline

- Preprocessing Pipelines
- Between python commands and SQL commands
- Biases
- Mlinespect Framework
- Datasets
- Runtime comparison

Prepossessing Pipeline

Prepossessing Pipelines:
Data sources gets modified. Use of wide range of different operations, we look at:

- table
- join
- aggregate project
- filter
- split

Prepared for use in machine learning

Operations

Translation of operations often line by line:

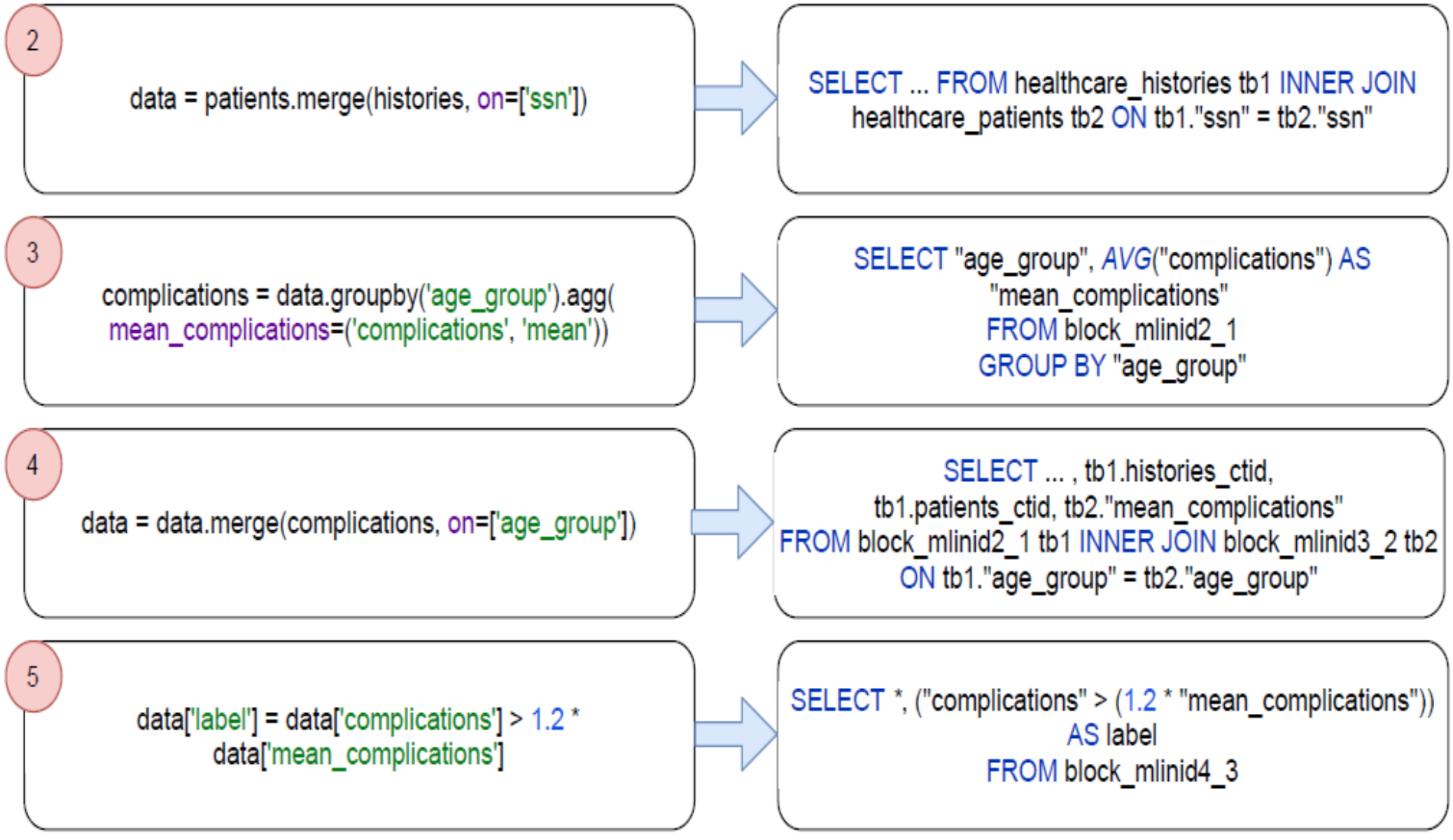

Figure 1: Translations

Why translation?

Advantages:

- Performance benefits ?
- Easy readable code
- Enables in-memory performance
- Eliminates the overhead of function calls
- Need to materialise subresults

Disadvantages:

- Need to materialise subresults.
- Existing libraries can't be used

Operations

Focus on Basic Panda Function begins with:

- Read from CSV
- Merge/Join
- Selection and Projection
- Arithmetic/Boolean Operations
- Group-By and Aggregation
- Drop Null Values
- Replace
- Row-Wise Operations

Operations

Typical Scikit-Learn Functions:

- Simple Imputer
- One-Hot-Encoder
  - Standard Scaler:

5.2.3 Standard Scaler. The standard score[6] z of a sample $l \in X$ is calculated as $z(l) = \frac{l - mean(x \in X)}{stddev(x \in X)}$. The mean and standard deviation is calculated in the fitting step (Listing 17) and reused for any other transformation.

```
1 SELECT (("label" - (SELECT AVG("label") FROM origin)))
2        / (SELECT STDDEV_POP("label") FROM origin) AS "label"
3 FROM origin
```

Figure 2: Runtime Analysis

  - KBins Discretizer
    - Binarize

BIAS

Bias is as a systematic error leading to irrational preferences or aversions. Can be reproduce and amplified when using machine learning. Difference between technical and introduced biases.

Technical Bias:

Appears in SQL and python scripts. Often after a preprocessing pipeline missclassifies data.

Introduced Bias:

Systematic error, whether through dataset with already existing problems or appearing in process. Materialising View/CTE hepls detecting them. It's possible to use Mlinespect to detect them.

mlinespt framework

Provides two checks:

- NoIllegalFeature verifies that none of the used features in provided dataset are contained in blacklist of illegal feature names
- NoBiasIntroducedFor targets pre-existing and technical biases

Mlinspect looks at data pipelines created with pandas and scikit-learn.
Mlinspect is designed to understand the semantics of preprocessing operations of popular Python frameworks from the data science space like scikit-learn and pandas.

Datasets

Uses two different datasets for training and testing. We look at four different pipelines which differ in tupel size and used operations.

- healthcare
- compas
- adult simple
- adult complex

All use pandas and scikit-learn operations. The instrumentation based on captured function calls described so far is independent of the specific library.

Graphs

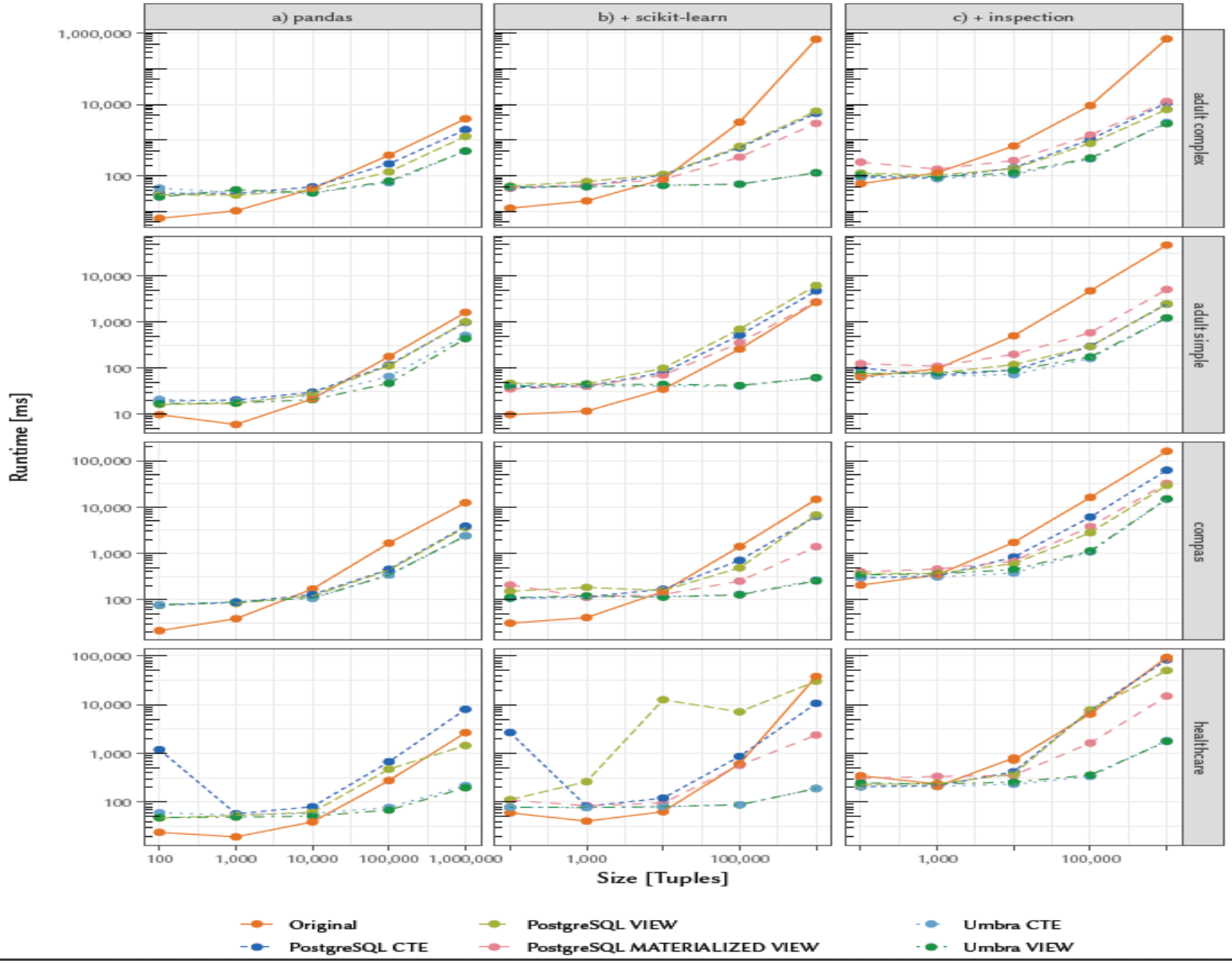

Figure 3: Runtime Analysis

Graphs

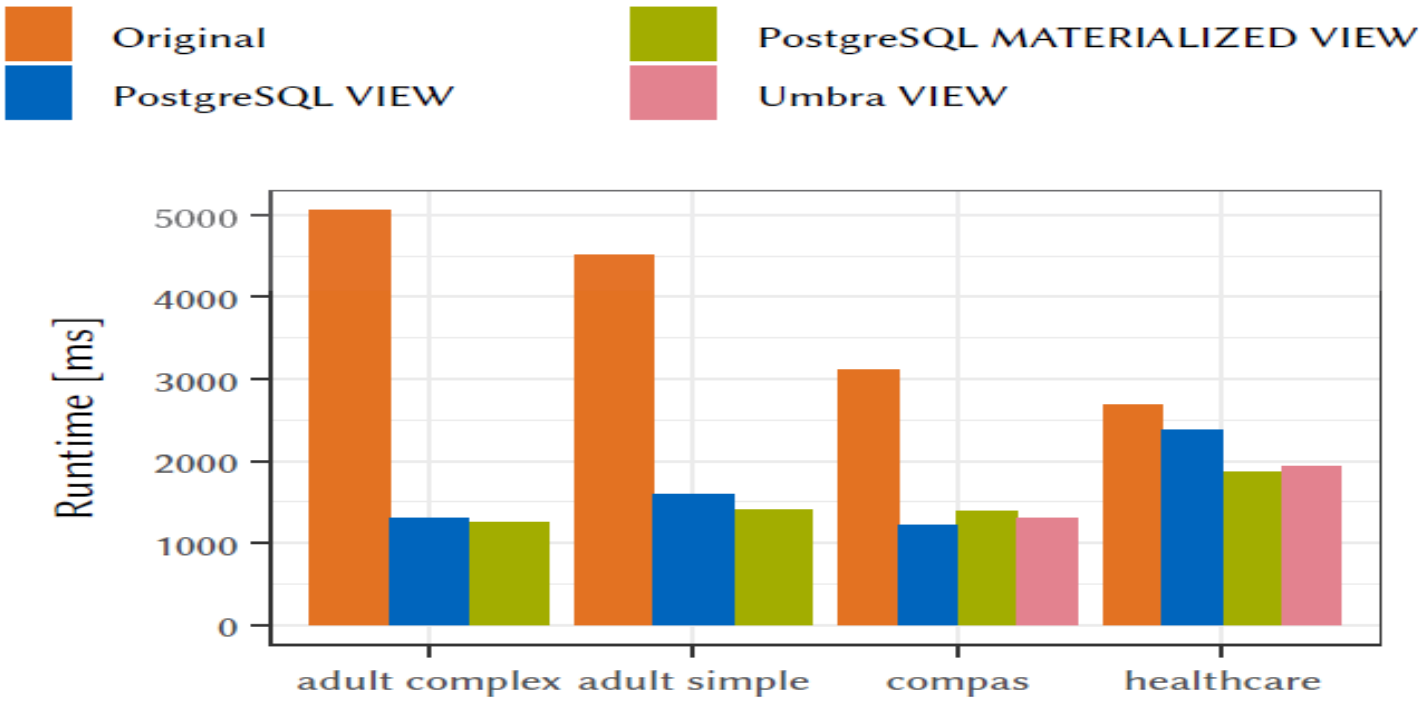

Figure 4: Runtime Analysis

Sources

https://www.cidrdb.org/cidr2021/papers/cidr2021_paper27.pdf
https://ubc-cs.github.io/cpsc330/lectures/05_preprocessing-pipelines.html#
https://openproceedings.org/2023/conf/edbt/paper-168.pdf