Scalable Local-Recoding Anonymization using Locality Sensitive Hashing for Big Data Privacy Preservation

Khotso Bore
University Of Pretoria
u19180642@tuks.co.za

Innocentia Ledimo
University Of Pretoria
u22928678@tuks.co.za

Busisiwe Vemba
University Of Pretoria
u22928678@tuks.co.za

Siphesihle Khumalo
University Of Pretoria
u25759257@tuks.co.za

ABSTRACT

A brief summary of your project. Mention the dataset, goal (e.g., clustering/classification/EDA), and key findings in 4–5 sentences.

1. EXPLORATORY DATA ANALYSIS (EDA)

1.1 Data Inspection

Describe your dataset: number of rows, columns, types of variables, missing values, and summary statistics. Include a short table or paragraph summarizing key properties.

Table 1: Dataset Overview

Feature	Type	Missing (%)
Age	Numerical	2.3
Gender	Categorical	0.0
Income	Numerical	5.1

1.2 Visualisations

Here are some key visualizations from the EDA phase. We explore the distributions of the numwerical features age and hours worked, and their correlations.

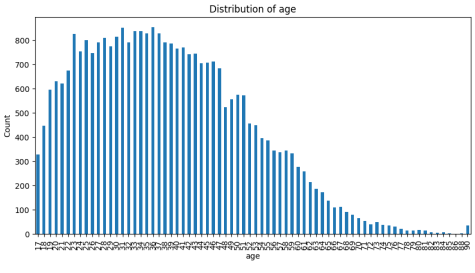


Figure 1: Age Distribution

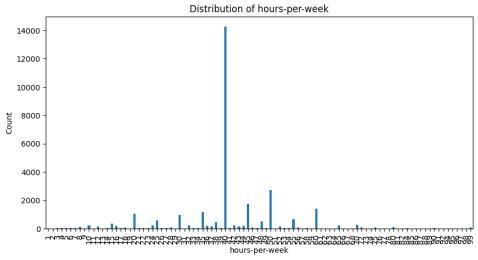


Figure 2: Hours Per Week Distribution

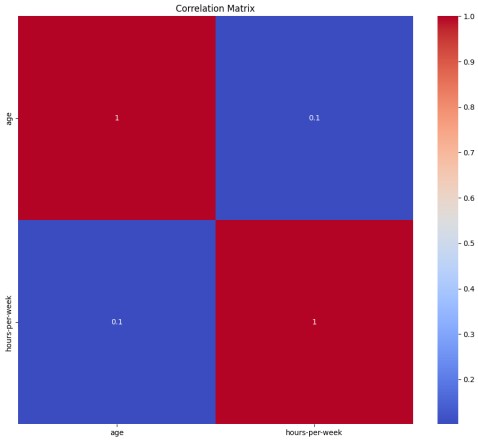


Figure 3: Correlation Matrix of Numerical Features

1.3 Insights

The correlation between age and hours worked suggests that while there may be slight tendencies—such as younger or older employees working somewhat more or fewer hours—the relationship is not strong or consistent across the group.

A large portion of employees work standard full-time hours (around 40 hours per week), with fewer employees working significantly more or less than this amount. Very likely because most records are from the United States, with a smaller representation from other countries.

2. DATA PREPROCESSING

2.1 Handling Missing Data

The dataset contained both standard and non-standard forms of missing values, which were handled in two stages:

2.1.1 Initial Removal of Missing Entries:

After loading the data, all records containing standard missing values were removed. This created a clean baseline for subsequent processing.

2.1.2 Handling Non-Standard Missing Indicators:

Some attributes, such as `workclass`, used the symbol "?" to represent missing values instead of a recognized missing data marker, such as `NaN`. These entries were first identified and converted into proper missing values, after which they were removed from the dataset.

This two-step process ensured that all incomplete records were removed. Maintaining a complete dataset is crucial for the LSH-based anonymization pipeline, since missing values can distort similarity measurements and clustering outcomes.

2.2 Feature Engineering

Feature engineering focused on selecting only the attributes relevant for anonymization and preparing them for efficient processing.

2.2.1 Feature Selection:

Several columns were removed from the dataset because they were either:

- Not quasi-identifiers that could link to external data sources,
- Redundant representations of existing information, or
- Sensitive or irrelevant for the anonymization process.

This reduced the dataset from fifteen to ten columns, retaining only the quasi-identifiers such as *age*, *workclass*, *education*, *marital status*, *occupation*, *relationship*, *race*, *sex*, *native country*, and *hours per week*. These attributes were used as the basis for the local-recoding anonymization.

2.2.2 Data Processing Throughout the LSH-Based Anonymization Pipeline:

Additional transformations were applied throughout the pipeline to ensure that data representation was suitable for the various algorithms at every stage of the pipeline:

- Binary vector conversion was used to transform categorical values into formats suitable for MinHash operations.
- Distance calculation functions were defined to measure semantic similarity between records.
- String standardization was applied to ensure consistent formatting across all categorical values.

These preprocessing components established the foundation for the LSH-based anonymization pipeline, which relies on accurate and consistent data representation. Further details on these transformations are discussed in Section 3.

2.3 Standardisation and Normalisation

Unlike numerical datasets that require techniques such as z-score standardisation or min–max scaling, this dataset consists mainly of categorical variables. Therefore, normalisation was incorporated directly into the anonymisation pipeline through specialised distance-based approaches:

- **Taxonomy-based distance metrics** that automatically bound all categorical distances within [0, 1] based on tree height and path length.
- **Provenance set-based semantic distances** that ensure comparability across different attributes and taxonomy structures.
- **String trimming** that removes whitespace to ensure consistent categorical value matching.

These methods maintain the original categorical form of the data while ensuring that all similarity and distance calculations are standardised. The details of these normalisation techniques are presented in Section 3.

3. DATA MINING METHODS AND ANALYSIS

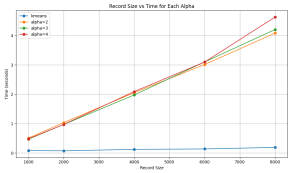
3.1 Methods

The anonymization process treats privacy preservation as a k-member clustering problem under k-anonymity. The proposed approach integrates these components:

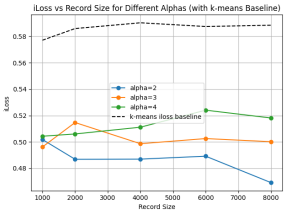
- **Provenance-Set Semantic Distance:** Defines similarity between records using Jaccard distance of their provenance sets, enabling LSH to approximate semantic closeness.
- **MinHash-Based LSH Partitioning:** Transforms each record's provenance vector into MinHash signatures. Similar records are hashed into the same β-clusters (coarse groups).
- **Recursive k-Member Clustering (LSH-RC):** Within each β-cluster, a recursive partitioning and agglomerative clustering algorithm is applied. Combines smaller clusters until each group size $\geq$ k.
- **Local recording:** Within each cluster, categorical attributes are generalized according to attribute taxonomies, ensuring that each cluster meets the k-anonymity requirement while minimizing information loss. The implementation constructs inverted taxonomies to compute necessary generalizations efficiently.

3.2 Results

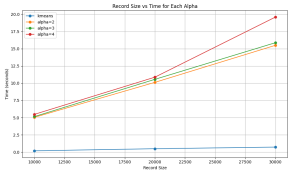
The performance of the proposed LSH-RC method was evaluated against a baseline k-means clustering approach across varying dataset sizes. The key metrics assessed were execution time and information loss (ILoss) as functions of the number of records. Overall our results indicate that LSH-RC scales much less efficiently than k-means while maintaining comparable ILoss levels.



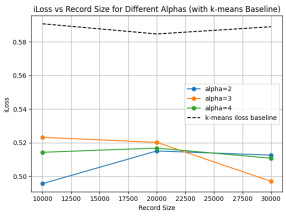
(a) Execution Time A



(b) ILoss A



(c) Execution Time B



(d) ILoss B

Figure 4: Execution time (a) & (c), and ILoss (b) & (d) w.r.t. #(records).

3.3 Discussion

Interpret the results:

- What patterns or groups emerged?
- Were the methods appropriate?
- Any limitations or anomalies?

4. CONCLUSION AND REFLECTION

Summarize what you found and learned:

- Key insights from data mining
- Challenges faced and how you overcame them
- Potential future work or improvements

Acknowledgements

(Optional) Acknowledge any data sources, collaborators, or funding.

5. ADDITIONAL AUTHORS

6. REFERENCES