Project Title: A Short, Descriptive Title

First Author
University / Organization
first.author@email.com

Second Author
University / Organization
second.author@email.com

ABSTRACT

This report details the implementation and evaluation of a scalable local recoding anonymization approach using locality sensitive hashing (LSH), primarily based on the research paper "Scalable Local-Recoding Anonymization using Locality Sensitive Hashing for Big Data Privacy Preservation" [2]. This approach proposed to enhance the scalability and efficiency of k-anonymity in big data environments. Using the UCI Adult dataset [1] the method integrates a new provenance-set-based semantic distance metric with LSH(specifically MinHash) for efficient data partition followed by a parallelized, recursive agglomerative k-member clustering algorithm(Beta-AC) [2].

1. EXPLORATORY DATA ANALYSIS (EDA)

1.1 Data Inspection

The dataset used is a preprocessed version of the Adult dataset from the UCI Machine Learning Repository [1]. It contains over 30000 records across 10 relevant attributes including 8 quasi-identifiers and 2 numerical features implicitly discretised by the methodology.
The analysis began with loading the Adult dataset consisting of 15 columns initially.

- Initial Data Count: The raw dataset had 32561 entries.
- column dropping: Several columns were considered non-quasi-identifying or as non-sensitive like capital-gain, capital-loss, fnlwgt, education-num, and income; in order to focus on the key quasi-identifier(QI) attributes.
- Final Quasi identifiers set: The resulting dataset contained 10 columns consisting of 8 categorical attributes(workclass, education, marital-status, occupation, relationship, race, sex, native-country), and two numerical attributes(age, hours-per-week) that would be discretized by the method. Work Class is used as the sensitive attribute.
- Missing Data: The initial inspection revealed missing data represented by a '?' string. These records were removed, reducing the total to 30162 clean records.

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

Include key plots to explore distributions or relationships:

- Histogram of key numerical features
- Boxplot to show outliers
- Pairplot / Correlation heatmap for relationships

1.3 Insights

Summarize interesting patterns:

- Which variables correlate strongly?
- Any skewed distributions or outliers?
- Early hypotheses about clusters or classes

2. DATA PREPROCESSING

2.1 Handling Missing Data

Explain how you dealt with missing data (imputation, deletion, etc.) and justify your choice.

2.2 Feature Engineering

List any new variables or transformations you applied (e.g., encoding, log transforms, ratios).

2.3 Standardisation / Normalisation

Discuss any scaling applied (e.g., z-score, min–max) and why it was necessary.

3. DATA MINING METHODS AND ANALYSIS

3.1 Methods

Describe which algorithms or analytical methods were applied:

- Clustering (e.g., K-Means, DBSCAN)
- Dimensionality Reduction (e.g., PCA, t-SNE)
- Classification/Regression (if applicable)

3.2 Results

Summarize the main results. Use figures/tables for clarity:

- Cluster quality metrics (e.g., silhouette score)
- Feature importances
- Visualizations of clusters or decision boundaries

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

[1] R. K. Barry Becker. Adult, 1996.

[2] X. Zhang, C. Leckie, W. Dou, J. Chen, R. Kotagiri, and Z. Salcic. Scalable Local-Recoding Anonymization using Locality Sensitive Hashing for Big Data Privacy Preservation. In *Proceedings of the 25th ACM International on Conference on Information and Knowledge Management*, CIKM '16, pages 1793–1802, New York, NY, USA, Oct. 2016. Association for Computing Machinery.