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.

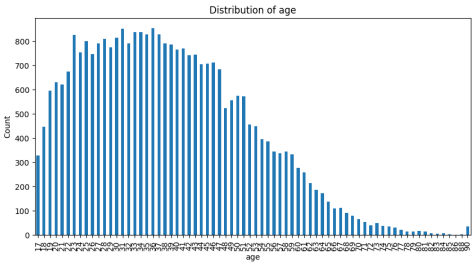

Figure 1: Age Distribution

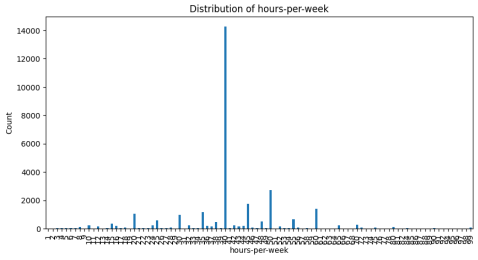

Figure 2: Hours Per Week Distribution

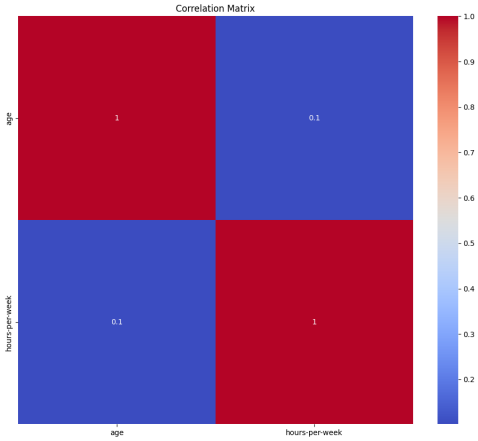

Figure 3: Correlation Matrix of Numerical Features

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
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