Understanding k-Anonymity and l-Diversity

Your Name

December 4, 2024

Introduction

- Data privacy is crucial in today's digital age.
- Re-identification attacks can compromise anonymized datasets.
- **k-Anonymity** and **l-Diversity** are techniques to protect data privacy.

k-Anonymity

- Ensures each record in the dataset is indistinguishable from at least **k-1** others based on quasi-identifiers.
- Achieved through:
  - **Generalization**: Reducing specificity of attributes.
  - **Suppression**: Removing outlier records.

Example

- Quasi-identifiers: Age, ZIP Code, Gender.
- Sensitive attribute: Disease.
- Goal: Ensure each combination of quasi-identifiers appears at least **k** times.

l-Diversity

- Addresses the limitation of k-anonymity (homogeneity attacks).
- Ensures each equivalence class has **at least l distinct sensitive values**.
- Can be achieved by:
  - Adding diversity to sensitive attribute values.
  - Suppressing rows that do not meet the criterion.

Example

- Sensitive attribute: Disease.
- Goal: Each group contains at least **l unique diseases**.

Implementation

- Synthetic dataset generated with attributes: Age, ZIP Code, Gender, Disease.
- Code written in Python to:
  - Check initial k-anonymity and l-diversity.
  - Apply transformations to meet desired k and l values.
- Visualize changes in the dataset.

Conclusion

- **k-Anonymity** ensures groups of indistinguishable records.
- **l-Diversity** adds diversity to sensitive attributes.
- Both methods reduce re-identification risk while balancing data utility.

Takeaways

- Effective data privacy techniques are critical for protecting individuals.
- Practical implementation helps understand trade-offs between privacy and utility.