Exploring High School Student Academic Metrics with Differential Privacy Techniques

Kiera Clarke, Noah Sullivan

Problem Statement:

The aim of this project is to investigate the correlation between high school students' academic performance in mathematics and sex while ensuring data privacy. The challenge lies in analyzing sensitive academic metrics without compromising individual student privacy.

Technical Description:

The solution involves implementing differential privacy mechanisms, particularly Laplace and Gaussian mechanisms, to preserve data privacy while analyzing academic metrics. Initially, the dataset, sourced from Kaggle, was preprocessed by selecting relevant columns such as age, sex, and academic grades ('grade_1', 'grade_2', 'final_grade').

Unique aspects of this implementation include:

Differential Privacy Mechanisms: Both Laplace and Gaussian mechanisms were utilized to add noise to aggregate statistics such as mean grades. This process involved defining mechanisms to compute noisy means for male and female students separately, preserving individual privacy while maintaining data utility.

Comparative Analysis: The project undertook a comparative analysis between Laplace and Gaussian mechanisms, evaluating their effectiveness in preserving privacy while accurately analyzing gender-based academic performance trends.

Synthetic Data Generation: As an integral part of ensuring differential privacy, synthetic data generation was implemented to maintain the statistical characteristics of the original dataset while preserving individual privacy. Specifically, the project utilized the two-marginal differential privacy mechanism to generate synthetic data based on marginal distributions of columns ('age', 'sex', 'grade_1', 'grade_2', 'final_grade'). The ‘dp_synthetic_data_two_marginal’ function was employed, which constructs the synthetic data based on four marginal distributions, ensuring differential privacy with an epsilon value of 1.0. The generated synthetic data exhibits a

similar structure to the original dataset, safeguarding individual privacy through the inclusion of noise while preserving statistical fidelity.

Synthetic Data Generation (Five-Marginals):
To further enhance the privacy and utility of the dataset, a five-marginal differential privacy mechanism was employed for the generation of synthetic data. Leveraging the dp_synthetic_data_five_marginal function, the synthetic dataset was created based on the joint distribution of five columns ('age', 'sex', 'grade_1', 'grade_2', 'final_grade') while maintaining differential privacy with an epsilon value of 1.0.

Results Description:

The Laplace and Gaussian mechanisms were applied to compute the mean final grades for male and female students. The actual mean final grades were compared with the noisy means generated by both mechanisms. The average percentage error was calculated to assess the accuracy of the mechanisms.

The synthetic data generation process demonstrated the application of differential privacy to create synthetic datasets. The generated two-marginal and five-marginal synthetic datasets aimed to maintain the statistical properties of the original dataset while ensuring individual privacy.

Conclusion:

The project successfully explored the relationship between high school students' academic performance and gender while preserving individual privacy through differential privacy mechanisms. The comparative analysis highlighted the trade-offs between privacy and accuracy in analyzing sensitive educational data.

The results demonstrate the efficacy of differential privacy techniques in protecting individual privacy while enabling valuable statistical analysis on sensitive datasets.