Differentially Private Analysis of the COMPAS Dataset

Anthony Lewis
College of Engineering and Computer Science
University of Michigan–Dearborn
alewi@umich.edu

Lincoln Badejo
College of Engineering and Computer Science
University of Michigan–Dearborn
lbadejo@umich.edu

Loren Fang
College of Engineering and Computer Science
University of Michigan–Dearborn
fangxy@umich.edu

Margil Funtanilla
College of Engineering and Computer Science
University of Michigan–Dearborn
mfuntani@umich.edu

Tom Tooma
College of Engineering and Computer Science
University of Michigan–Dearborn
ttooma@umich.edu

ABSTRACT

In the modern technology world, data is continuously being shared and collected from users. Few concerns are raised regarding the collection of data as to the ways data is being handled and the privacy apprehensions that may arise when data is utilized in analytics. With data security and privacy, various methods exist as to the privacy and accuracy tradeoff of data being handled through different mechanisms that analysts may consider when studying datasets for information regarding inquiries that occur from the dataset as well as useful visuals for data comprehension. In this project, Gaussian and Laplace mechanisms are used to perform a differentially private analysis of the Correctional Offender Management Profiling for Alternative Sanctions (COMPAS) dataset obtained from the algorithm used for scoring over ten thousand criminal defendants in Broward County, Florida. In this project, results of distinct mechanism utilization are reviewed with benefits for each with the breakdown of data and visuals for effectively comprehending the data being analyzed. Overall, the results highlight that differential privacy can serve as a practical tool for analyzing sensitive criminal justice data, enabling researchers and policy makers to balance privacy protection with the need for reliable insights.

KEYWORDS

Privacy, data, security, defendant, algorithm, mechanism, tradeoff, analyst, gaussian, differential privacy

1 Introduction

Differential privacy has become a leading framework for protecting sensitive information in data analysis and machine learning. The key idea is to introduce randomness to the mechanism that produces the output–for example, a statistic computed from a sample or a model prediction–so that individual data cannot be distinguished. This provides strong, mathematically grounded privacy guarantees while still allowing useful insights to be extracted from the data [1].

Despite its growing adoption, differential privacy remains difficult to apply in practice. Implementers often lack definitive guidance on selecting appropriate privacy budgets (ε), understanding how different variants of the definition affect privacy, or balancing model utility with meaningful protection. As a result, many real-world deployments rely on arbitrary choices of ε, which can substantially diminish the strength of the privacy guarantees. These challenges are especially important in high-stakes domains, where privacy breaches can have real consequences for individuals. One such domain is the criminal justice system. Datasets used in risk assessment tools, such as Correctional Offender Management Profiling for Alternative Sanctions (COMPAS), contain highly sensitive information about individuals' backgrounds, behaviors, and prior interactions with law enforcement. Improper handling of such data risks exposing private details, reinforcing biases, or enabling re-identification. Yet, these datasets are routinely used for research, model evaluation, and policy analysis.

In this project, we use the COMPAS dataset as a case study to demonstrate how differential privacy can be applied to protect individuals in criminal justice data. We explore how adding privacy-preserving noise to key components of the analysis affects both privacy and utility. By applying differential privacy to tasks such as statistical estimation through queries, we illustrate how strong privacy guarantees can coexist with meaningful, interpretable results. Our goal is to provide a clear, practical demonstration of differential privacy in action and to highlight its relevance and importance when working with sensitive, real-world data.

2 Design/Approach

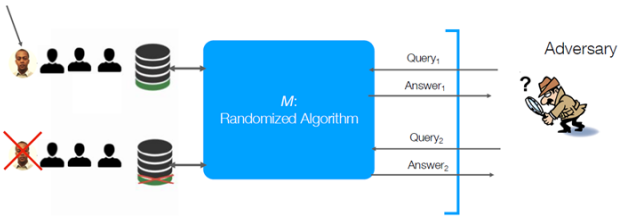
This section reviews the definition of Differential Privacy, explains the Laplace and Gaussian mechanisms, and describes the methodological approach used in our analysis.

2.1 Differential Privacy

The intuition behind differential privacy is illustrated in Figure 1 using two neighboring datasets, *d1* and *d2*, which are identical

except for one record (the example record shown in the class slide featuring Professor Birhanu).

Figure 1: Differential Privacy: Intuition



The goal is to ensure a privacy mechanism *M*, which is an algorithm that takes as input (e.g., a set of queries) and produces an output, is randomized so that an adversary, by observing the output (Answer 1 or Answer 2), cannot determine whether the result was generated from *d1* or *d2*. That is, Answer 1 or Answer 2 is essentially the same regardless of the inclusion of Professor Birhanu's record in *d1* [2].

Formally, a randomized mechanism $M: D \rightarrow R$, with domain D and range R satisfies ε-differential privacy if, for any two neighboring datasets $d1$, $d2 \in D$ and any subset of outputs $S \subseteq Range(M)$ it holds that:

$$Pr[M(d1) \in S] \leq e^{\varepsilon} * Pr[M(d2) \in S] + \delta.$$

Intuitively, differential privacy ensures that the output of *M* does not allow an adversary to determine whether *M* was based on *d1* or *d2* (Figure 1). In other words, the probability of obtaining any valid output is nearly the same whether or not the individual's data is included in the dataset.

The parameter ε is the privacy budget, which controls the strength of the privacy guarantees. The smaller the ε, the harder it is for an adversary to distinguish between outputs based on *d1* and *d2*. However, when ε is too small, it reduces the utility of the data. The parameter δ represents a small probability that quantifies the realistic assumption around the mechanism failing to produce a valid output. When δ=0, we obtain the original definition of ε-DP, i.e., pure differential privacy. It follows that:

$$\ln \frac{Pr[M(d1) \in S]}{Pr[M(d2) \in S]}$$

is the privacy loss. In this project, we implement (ε, δ)-DP by adding noise sampled from Laplace and Gaussian distributions to query results, aiming to add enough noise to improve privacy while minimizing the loss of data utility.

For a function, *f*(*d*), that returns a number, *M*(*d*) satisfies ε-DP:

$$M(d) = f(d) + Lap\left(\frac{s}{\epsilon}\right)$$

where $f(d)$ is the true value of the function and $Lap\left(\frac{s}{e}\right)$ is the act of sampling from Laplace distribution with mean 0 and scale *s*; *s* is the sensitivity of *f* (a query in this analysis). For a function *f* over neighboring datasets *d1* and *d2*, the sensitivity of *M* is the maximum change in the output:

$$\Delta f = max_{d1,d2} \, ||f(d1) - f(d2)||_1.$$

Intuitively, this measures how much one record can affect the output. For counting queries, Δ*f*=1.

Gaussian mechanism is used to add noise to function *f*(*d*) through the following:

$$M(d) = f(d) + \frac{s * \sqrt{2 * log \frac{1.25}{\delta}}}{\epsilon}$$

2.2. Tools and Approach

In this section, we describe the practical implementation of our analysis, including the programming tools and the libraries used, the data preparation process, and the approach taken to apply the Laplace and Gaussian mechanisms for differential privacy on the COMPAS dataset.

In order to perform the comparison analysis of the COMPAS dataset for the project, several tools and libraries are necessary for analysis of the data pertaining to the dataset as well as understanding the privacy tradeoff of mechanisms that are utilized for the project. The programming language, Python, with an associated compiler such as Visual Studio Code is the primary tool that is used for the project in order to program the essential software and mechanisms for the analysis to occur and provide results. Various libraries such as numpy, pandas, and matplotlib are exercised within the project to provide for the required necessities that the analysis will need to effectively depict the data and comparison at hand between Gaussian and Laplace mechanisms. Numpy is often used in analytics for data computation as well as manipulation. In the case of this research, we use numpy to perform many of the analytical computations that are required for the experimentation of the COMPAS dataset breakdown and mechanism comparison. Pandas is a library in Python that is frequently utilized for its data manipulation properties and data cleaning. For the project, we utilize Pandas for effectively performing data cleaning and manipulation, as well as creating necessary DataFrames for the data to perform our analysis of the COMPAS algorithm. Lastly, matplotlib is the main library being utilized to represent visuals of the data. The library is effective at providing means for plot visualization of the dataset.

The approach of the project was to first clean up the data from the COMPAS algorithm so the data is adequately readable in order to perform the privacy tradeoff comparison. Afterwards, several queries were completed to effectively depict the data necessary for displaying key information of the data as well as knowledge regarding the accuracy and privacy tradeoff.

3 Implementation

3.1 Data Cleaning and Preprocessing

The COMPAS dataset contains variables used by the COMPAS algorithm in scoring criminal defendants, along with their outcomes within 2 years of the decision, in Broward County, Florida. COMPAS is an algorithm designed to assess a defendant's likelihood of recidivism–a term used to describe

individuals who re-offend. COMPAS is used by criminal justice agencies to inform decisions regarding the placement, supervision and case management of offenders.

To prepare the COMPAS dataset for analysis, we began with the compas-scores-two-years.csv file provided by ProPublica [3], which has 7,214 records in the raw dataset. We applied the same three preprocessing rules used in the ProPublica methodology to construct a valid two-year recidivism group. First, only data relevant to defendants who have an arrest date within 30 days of their COMPAS screening date was retained to ensure alignment between risk assessment and offense (Rule 1). Next, we excluded rows with indeterminate recidivism labels (is_recid = -1), since those cases do not provide reliable outcome information (Rule 2). Third, we removed data points with ordinary traffic violations (c_charge_degree = 'O'), as these cases usually do not lead to incarceration and fall outside the criteria of our risk analysis (Rule 3). We also improvised by removing any rows without a valid two_year_recid label and casting that field to an integer data type for consistent categorical use (Rule 4). After these filtering rules were applied, the dataset was reduced to 6,172 rows and 53 columns, forming the cleaned two-year group.

Because the COMPAS dataset contains highly sensitive personal information, a privacy-risk audit was conducted, and each field was classified into Personally Identifiable Information (PII), Sensitive Personal Identifiable Information (SPII), or Indirect Identifiers (INDIR). PII fields were full name and numeric ID since they could point directly to an individual [4]. SPII fields included demographic characteristics such as race, sex, date of birth (DOB), and legal case identifiers, which could generate reputational harm on the individual if uncovered [5]. Then INDIR fields were determined that could uniquely identify individuals when combined with other fields (e.g., dates, prior counts) [6]. We constructed separate subsets for each category to confirm coverage and identify fields requiring noise protection during differential privacy. This step verified that 51 out of 53 columns fall into PII, SPII, or INDIR, while time features (start, end) did not clearly fit these classifications. Importantly, no privacy noise was introduced during cleaning. All transformations prepared the data for Laplace and Gaussian mechanism experimentation.

Category	Fields	Count
PII	full name, first, last, id	4
SPII	race, sex, dob, jail/arrest/case details, assessment	13
INDIR	age, days_b_screening_arrest, offense timing, prior counts	34
Neither	start, end	2

3.2 Analytical Workflow

The analysis proceeded in two layers: first, a ground-truth (non-DP) layer; second, a differentially private layer using both Laplace (ε-DP) and Gaussian ((ε, δ)-DP) mechanisms. In the ground-truth layer, we computed histograms of 'DecileScore' by (race, sex), mean scores by two-year recidivism outcome, and summary distributions of scores by race and race/sex. Decile scores are transformed COMPAS scale scores. It is obtained by ranking the scale scores of a group in ascending order and then dividing these scores into ten equal sized groups [7]. We defined a high-risk label as DecileScore ≥ 7 and examined the proportion of individuals classified as high risk in each group. These results were visualized using bar charts to make differences between groups easy to interpret.

In the DP layer, we implemented Laplace and Gaussian mechanisms on various query types: (1) distribution of racial groups, age, and sex in the dataset; (2) average COMPAS score for defendants; (3) recidivism rate; (4) histograms of decile scores by race/sex and (5) mean decile scores by two-year recidivism outcome. Scores were clipped to [1,10] for bounded sensitivity, and we evaluated multiple ε values (0.1, 0.2, 0.5, 1.0, 2.0). For each ε, we reran DP mean queries 50 times and measured the mean absolute error compared to the true means to quantify the privacy–utility trade-off.

4 Experimental Evaluation

A variety of queries were evaluated to support our analysis of the COMPAS dataset. These queries include the use of Laplace and Gaussian mechanisms on the distribution of racial groups in the dataset, age, and gender distributions, COMPAS score distribution, two-year recidivism rates or the proportion of defendants who committed a new crime over the two years, differences in COMPAS score averages, decile score distribution by race, etc. Each query was executed under different privacy budget settings. For clarity, the following sections present results for a selected subset of these scenarios, while the full set of results is available in our implementation code.

4.1 Racial groups distribution

Figure 2 shows the distribution of racial groups in the dataset. Approximately 51.4% of defendants are African-American, 34.1% are Caucasian, 8.2% are Hispanic, 0.5% are Asian, 0.2% are Native American, and almost 6% are categorized as Other.

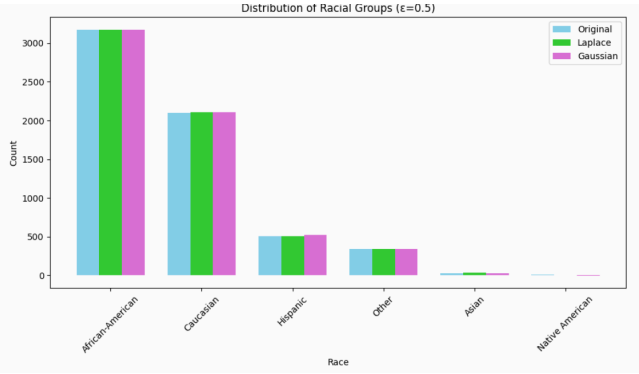


Figure 2: Distribution of Racial Groups (Epsilon = 0.5)

Table 1 reports the differentially private racial distribution produced by the Laplace and Gaussian mechanisms at $\varepsilon = 0.5$. The Laplace mechanism yields proportions that closely track the true distribution, while the Gaussian mechanism introduces slightly larger deviations, including modest increases in the African-American and Hispanic proportions.

	Original		Laplace (ε=0.5)		Gaussian (ε=0.5)	
Race						
African-American	3175	51.4%	3173.4	51.4%	3175.6	51.5%
Caucasian	2103	34.1%	2104.8	34.1%	2103.9	34.1%
Hispanic	509	8.2%	509.4	8.3%	524.9	8.5%
Other	343	5.6%	342.8	5.6%	340.3	5.5%
Asian	31	0.5%	32.9	0.5%	27.2	0.4%
Native American	11	0.2%	6.6	0.1%	-5.5	-0.1%

Table 1: Tabular Distribution of Racial Groups (Epsilon = 0.5)

4.2 Age Distribution

Figure 3 presents the distribution of defendants' ages in the dataset. Approximately 43.4% of the defendants are younger than 30 years old, about 31% are in their 30s, 11.4% are in their 40s, 11.3% are in their 50s, and about 2.9% are 60 or older.. Under the assumption of $\varepsilon = 0.1$, Figure 3 illustrates that adding Laplace noise produces an age distribution that remains relatively close to the true values, while the addition of Gaussian noise introduces greater variability.

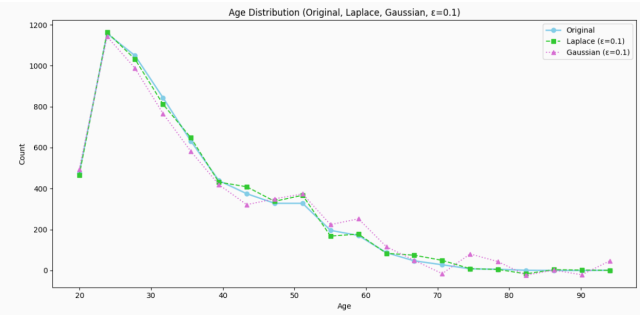


Figure 3: Age Distribution (Epsilon = 0.1)

4.3 Gender Distribution

Figure 4 shows the gender distribution in the dataset, where the vast majority of the defendants are male (81.0%), and only 19.0% are female. Under a privacy budget of $\varepsilon = 0.01$, adding Laplace noise produces a distribution that remains close to the true proportions (81.3% and 18.7%, respectively). More variability appears when Gaussian noise is added, shifting the distribution to 76.6% male defendants and 23.4% female.

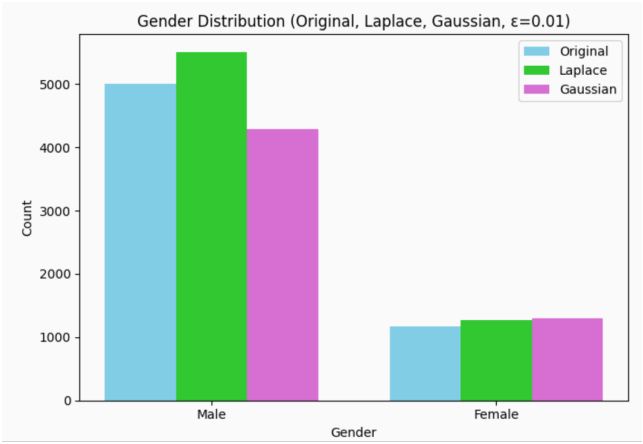


Figure 4: Gender Distribution (Epsilon = 0.01)

4.4 Average COMPAS Score

Figure 5 shows the distribution of COMPAS Decile Score for the defendants. The average Decile Score is indicative of riskiness of the defendant population. The mean score is 4.42, which suggests a generally low risk of recidivism [6]. With $\varepsilon = 0.1$, adding a Laplace noise shows the average is 4.44, and 4.46 with Gaussian noise.

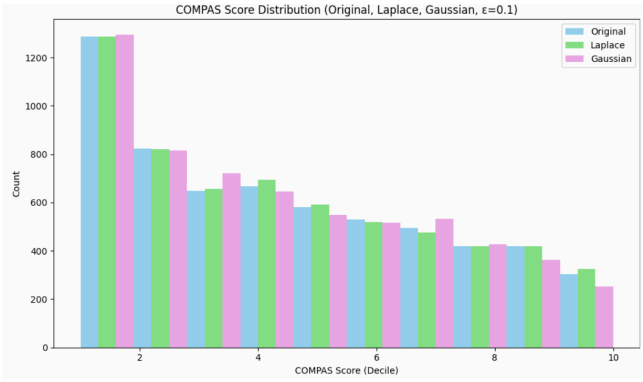


Figure 5: COMPAS Score Distribution (Epsilon = 0.1)

4.5 Recidivism Rate

Figure 6 shows the proportion of defendants who committed a new crime over the next 2 years. The data shows that about 45.5% re-offended in 2 years. With $\varepsilon = 0.01$, the recidivism rate is 46.5% with Laplace noise, while Gaussian noisy recidivism rate is 40.2%.

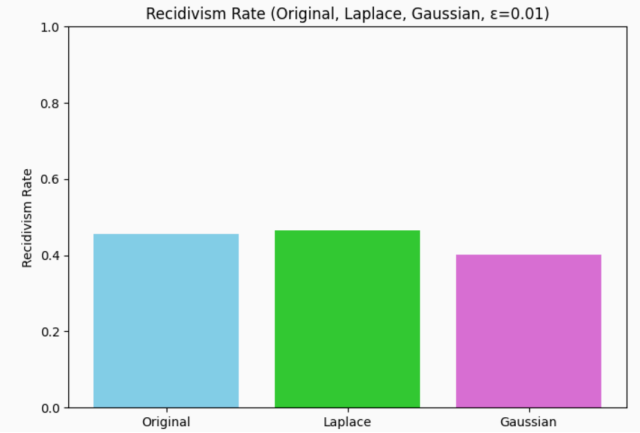


Figure 6: Recidivism Rate (Epsilon = 0.01)

4.6 COMPAS Scores by Race and Gender

From the ground-truth statistics, we deduced two main substantive patterns. First, COMPAS scores are meaningfully correlated with two-year recidivism: individuals who do recidivate within two years have mean decile scores roughly two points higher than those who do not, indicating that the score carries genuine predictive signal at the aggregate level. Second, there are pronounced racial disparities in scoring and high-risk classification. African-American defendants have substantially higher average scores and much higher high-risk rates than Caucasian or Hispanic defendants with Native Americans showing the highest high-risk rates but in very small numbers. Figures 8 and 9 confirm that the entire score distribution for African-American (and especially African-American males) is shifted toward higher deciles, rather than the disparities being driven by a few outliers.

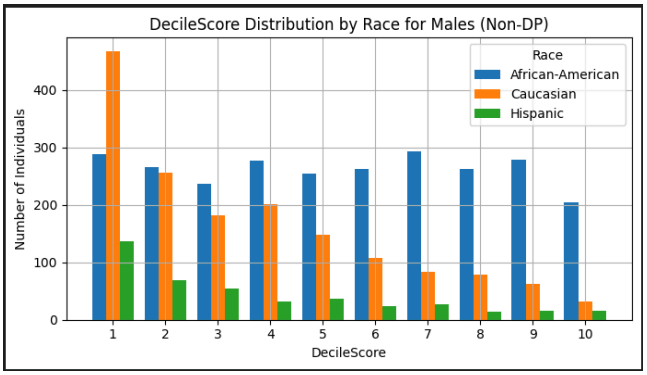


Figure 7: Decile Score Distribution by Race for Males (Non-DP)

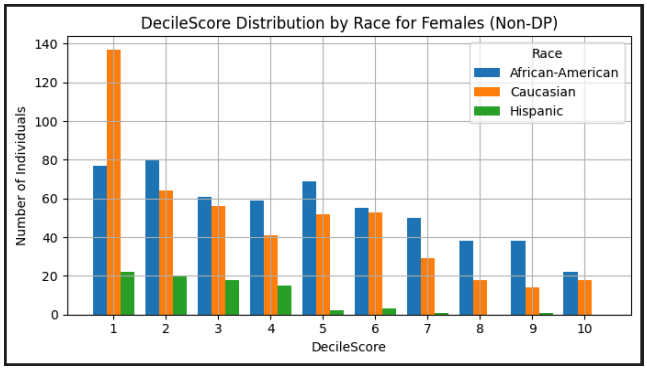


Figure 8: Decile Score Distribution by Race for Females (Non-DP)

The DP layer showed that these key patterns can be preserved while adding formal privacy guarantees. For aggregate mean queries, the Laplace mechanism produced extremely small errors—even at ε as low as 0.1, errors were on the order of only a few hundredths of a decile, and at $\varepsilon \geq 1$ they were essentially negligible. Gaussian DP, with δ = 1e-5, produced slightly larger but still small errors, with mean absolute error decreasing roughly in proportion to $1/\varepsilon$. DP histograms (Figures 10 to 12) for large racial groups retained the overall shape and relative positioning of the non-DP histograms: African-American distributions remained visibly higher-risk than Caucasian distributions under both Laplace and Gaussian mechanisms at $\varepsilon \approx 1$. This leads to two core deductions:

1. It is possible to publish differentially private fairness-relevant statistics (like score distributions and mean scores by group) with minimal loss of interpretability
2. The racial disparities observed in COMPAS scoring are robust enough that they remain visible even after adding DP noise to protect individual-level privacy.

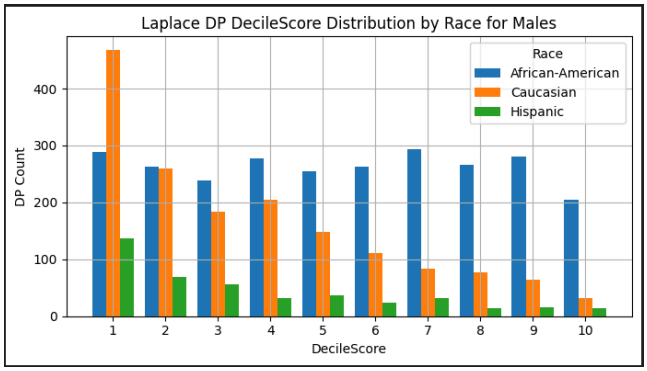


Figure 9: Laplace DP Decile Score Distribution by Race for Males

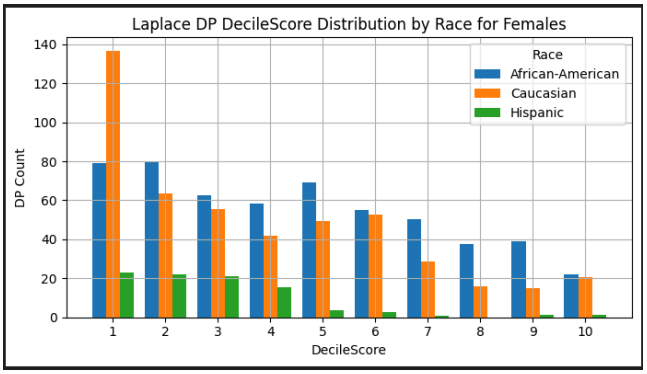


Figure 10: Laplace DP Decile Score Distribution by Race for Females

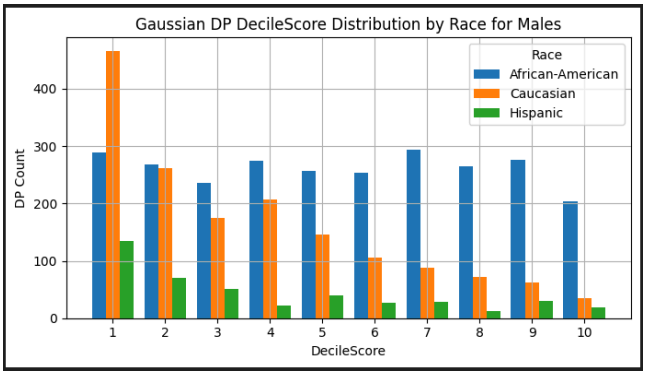


Figure 11: Gaussian DP Decile Score Distribution by Race for Males

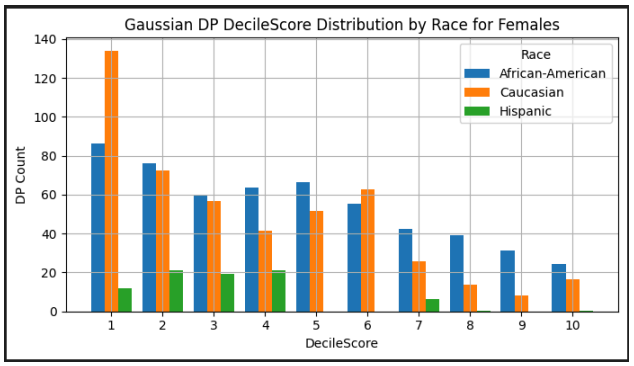


Figure 12: Gaussian DP Decile Score Distribution by Race for Females

Privacy- Utility Trade-off

Figure 13 shows the privacy and data utility trade-off across all tested values of ε. The figure indicates that, overall, the Laplace mechanism outperforms the Gaussian mechanism, yielding lower, mean absolute error for the same privacy budget. In particular, at lower ε values—where the privacy guarantee is strongest—the Laplace mechanism preserves substantially more accuracy. These results suggest that Laplace noise is better suited for releasing aggregate statistics under moderate to strict privacy requirements.

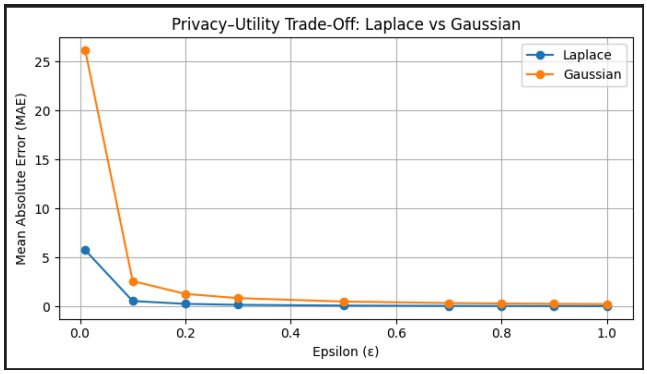


Figure 13: Privacy Utility Trade-Off of Laplace vs Gaussian

5 Conclusion

Our experiments using the COMPAS dataset demonstrate that strong privacy guarantees can be achieved with minimal loss of interpretability in aggregate statistics. In particular, the Laplace mechanism generally produced lower data utility loss compared to Gaussian, especially at moderate and high privacy levels, making it well-suited for releasing summary statistics in sensitive domains.

Despite introducing formal privacy noise, major patterns and disparities observed in the raw data, such as racial gaps in COMPAS scores and recidivism rates, still remained visible after differential privacy was applied. This suggests that differentially private methods can effectively protect individual-level privacy while still enabling meaningful analysis and accountability.

Overall, the results highlight that differential privacy can serve as a practical tool for analyzing sensitive criminal justice data, enabling researchers and policy makers to balance privacy protection with the need for reliable insights.

REFERENCES

[1] Cynthia Dwork, Frank McSherry, Kobbi Nissim, and Adam D. Smith. 2006. Calibrating Noise to Sensitivity in Private Data Analysis. TCC 2006, New York, NY, USA, March 4-7, 2006, Proceedings (Lecture Notes in Computer Science), Shai Halevi and Tal Rabin (Eds.), Vol. 3876. Springer, 265–284.

[2] Cynthia Dwork and Aaron Roth. 2014. The Algorithmic Foundations of Differential Privacy. Found. Trends Theor. Comput. Sci. 9, 3-4 (2014), 211–407.

[3] Jeff Larson and Marjorie Roswell. 2017. COMPAS-analysis. Github. GitHub - propublica/compas-analysis: Data and analysis for 'Machine Bias'

[4] National Institute of Standards and Technology, Guide to Protecting the Confidentiality of Personally Identifiable Information (PII), SP 800-122. Gaithersburg, MD, USA: NIST, 2010. https://csrc.nist.gov/pubs/sp/800/122/final

[5] Protecting Americans From Harmful Data Broker Practices (Regulation V). Federal Register, vol. 89, no. 239, pp. 87621–87727, Dec. 13, 2024. https://www.federalregister.gov/documents/2024/12/13/2024-28690/protecting-americans-from-harmful-data-broker-practices-regulation-v

[6] L. Sweeney. 2000. Simple Demographics Often Identify People Uniquely. Data Privacy Working Paper 3, Carnegie Mellon Univ., Pittsburgh, PA, USA, 2000.

[7] Northpointe. 2015. Practitioner's Guide to COMPAS Core. Practitioner-s-Guide-to-COMPAS-Core.pdf