Impact of Normal, Uniform, and Laplace Noise Addition Techniques on Central Tendencies of Data in Privacy-Preserving Data Mining

Saurab Gupta[1], S Harshitha Devi[2], Shashidhar V[3]
[1,2]Student, [3]Assistant Professor
[1,2,3]Department of Computer Science & Engineering,
[1,2,3]RV Institute of Technology and Management, Bengaluru, India

Abstract - Through this paper, we deal with the problem of privacy preservation in data mining using various techniques of noise addition such as normal, uniform, and Laplace. In this study, we examined the impact of noise addition techniques on a dataset containing student scores of three different subjects where mean, median, and standard deviation were used as three different central tendencies. On performing multiple iterations on datasets, the results obtained conveyed that the uniform addition performed better than normal and Laplace in preserving the mean and median on datasets. Whereas the normal addition performed better than the uniform and Laplace addition in preserving the standard deviation on original datasets. Our work is been motivated by the need to safeguard sensitive information such that we can enable its use for various purposes.

Index Terms - Normal Noise, Uniform Noise, Laplace Noise, Noise Addition, Central Tendencies, Privacy-Preserving Data Mining.

I. INTRODUCTION

Data Mining is a growing discipline that involves the procedure of identifying patterns, relationships, and insights from massive databases. To find different patterns, correlations, and trends in data, a variety of techniques from machine learning, data analysis, and database management systems are used. Data mining is used in various industries such as finance, health care, marketing, and e-commerce[1] to discover outcomes and improve decision-making.
Educational institutions often collect data of students, parents and faculty members that includes vast amount of personal information such as academic records, health records, thereby it is essential to secure the sensitive information from unauthorized user[2,3]. If the institution's data is damaged in any manner, there may be severe penalties for the institution. To build a sense of trust among the members of institute it is highly essential to maintain the privacy of the data.

Confidentiality issues in data mining:
Datasets often involve confidential information which without proper protective measures can lead to a violation of privacy. In some cases, organizations share data with third-party analytics vendors by depersonalizing the customer's data to analyze the data patterns. This often involves identifying information that can be tracked back to individuals. Such situations can put the organization at risk. To overcome such risks several techniques such as clustering and classification are used in to implement privacy[4].

Privacy-preserving data mining techniques:
Noise addition involves adding random noise to the data in order to provide privacy in data mining[5]. The different noise distribution techniques used are normal distribution, uniform distribution, and Laplace distribution throughout the paper.

II. LITERATURE SURVEY

An overview of privacy protection in DL, where the datasets are maintained by noise addition in deep learning, is provided by Ahmed EI Ouadrhiri et al[6]. In simple terms, the datasets are protected against model inversion attacks, which pose a threat to those datasets that contain sensitive data when the attackers do not have direct access over the data. To determine whether a person was a member of the training set or not, an attacker can use a different attack strategy known as a membership inference attack[7]. The author has addressed a number of approaches for dealing with deep learning privacy issues. 1) The three methods are K-anonymity, L-diversity, and T-closeness. These methods build a new dataset that safeguards user information, preventing the extraction of sensitive data even if attackers gain access to the entire dataset.2) A method that safeguards user privacy while multiple parties are being trained and each party's dataset is maintained private.3) A technique based on differential privacy, in which the user's privacy is attained through three stages of training, which are:1) Generating a privacy-preserving dataset prior to training.2) In the process of training, safeguard the data sent from the client to the server.3) On the basis of training process, models that hinder model inversion and membership are generated Nisha Chaudhary et al[8] summarize the use of differential privacy in medical records which will evaluate patients' records and identify patients' illnesses from the provided set of data. This proposed model includes very little human participation. The differential privacy technique uses the method of adding noise to statistical data on the basis of Laplace distribution and Gaussian distribution to classify various parameters in medical records.

A concept to stop the disclosure of private user information on apps like Facebook and Instagram has been put forward by Aghasian et al[9]. The Bernstein polynomial theorem is the recommended approach for ensuring privacy of users on social media.
Throughout the paper of Hillol Kargupta, the effectiveness of the random-value distortion approach[10] for maintaining privacy is raised. It next proposes a random matrix-based spectral-filtering approach. The suggested approach compares the spectrum produced from the observed data with that of random matrices to see how well it performs. The detailed experimental findings presented in this

study combine to show that, in many situations, random data distortion retains very little data privacy. This paper's analytical methodology identifies a number of potential directions for the creation of unique privacy-preserving data mining methods. Examples include algorithms that use multiplicative and colored noise to preserve privacy in data mining applications, and algorithms that clearly assure against privacy breaches using linear transformations.

III. DEFINITIONS

(1)Mean

Mean is the mathematical measure of central tendency. By dividing the total of the given data with number of values the mean can be computed[11].

$$\bar{x} = \frac{1}{n}\left(\sum_{i=1}^{n} x_i\right)$$

(2)Median

A median value is one that is in the center[12]. A large number of data points may be represented by a single point using median. Median is an easiest mathematical statistic that can be calculated. After the data is arranged in ascending order for computing the median, the middle data point reflects the median of the data.

$$\text{If } n \text{ is odd, } \text{median}(x) = x_{(n+1)/2}$$

$$\text{If } n \text{ is even, } \text{median}(x) = \frac{x_{(n/2)} + x_{((n/2)+1)}}{2}$$

(3)Standard Deviation

The distribution of data around the mean value is measured by standard deviation. It is employed in comparisons of the consistency of various data sets.

$$\sigma = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(x_i - \mu)^2}, \text{ where } \mu = \frac{1}{N}\sum_{i=1}^{N} x_i$$

(4)Normal Distribution

The Normal Distribution[13], which is known by the name Gaussian Distribution, is a continuous probability distribution which is symmetric about its mean. The mean, median and mode of a normal distribution is same. The shape of the distribution is a bell curve, which conveys that the major values are clustered around the mean.The mean and standard deviation are two parameters used to determine the distribution. The center of distribution is determined by the mean, while its dispersion is determined by the standard deviation.

The distribution is defined by the following probability distribution function:

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}$$

(5)Uniform Distribution

In Uniform Distribution[14] all the values have equal probability of occurrence within a range. The probability density function is constant over a specified interval and zero elsewhere in uniform distribution. The shape of the distribution is rectangular. The two parameters, the upper limit, and lower limit, describe the range of values that can occur in the distribution.

The distribution is defined by the following probability distribution function:

$$f(x) = \begin{cases} \frac{1}{b-a}, & a \le x \le b \\ 0, & x < a \text{ or } x > b \end{cases}$$

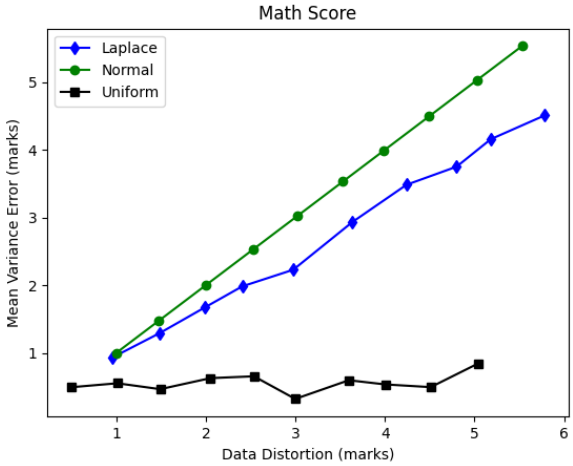
(6)Laplace Distribution

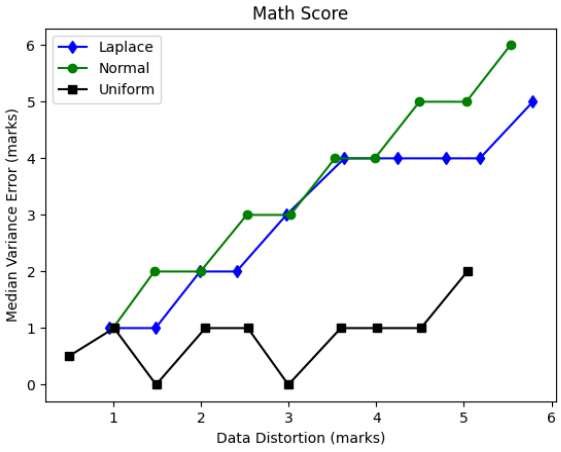
The Laplace Distribution[15],which is known by the name Double Exponential Distribution, is a continuous probability distribution which has a shape similar to bell curve. The distribution is similar to normal distribution but has heavier tails resulting in a higher probability of extreme values when compared to normal distribution. The location parameter and the scale parameter are the two parameters that control the distribution. The distribution's center is represented by the location parameter, while the distribution's spread is determined by the scale parameter.

The distribution is defined by the following probability distribution function:

$$f(x) = \frac{1}{2b}\exp\left(-\frac{|x-\mu|}{b}\right)$$

IV. GRAPHS





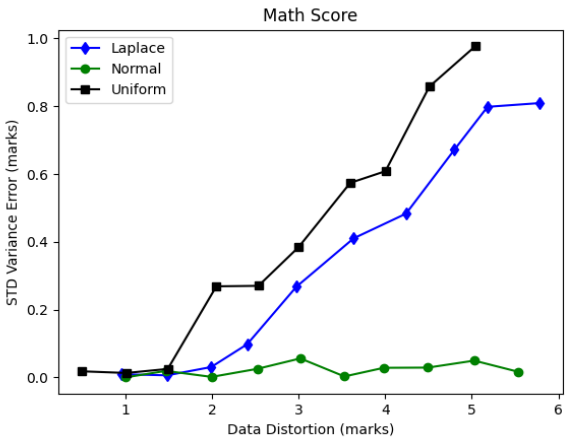
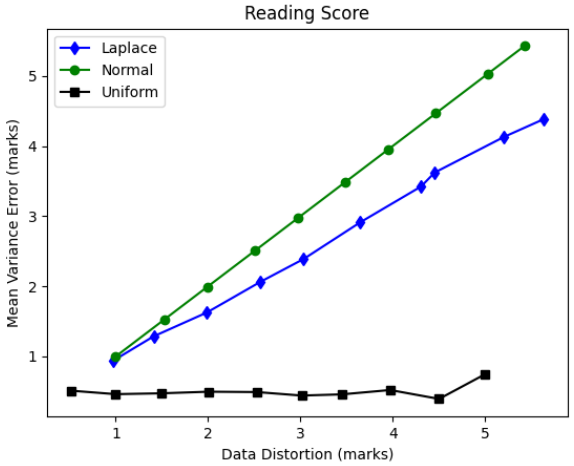
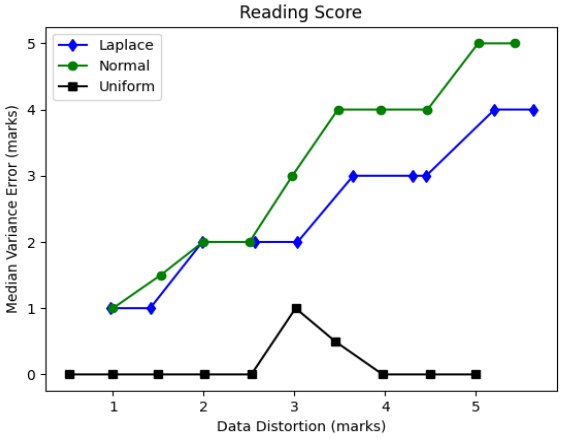


Fig 1: Graphs of Math Score





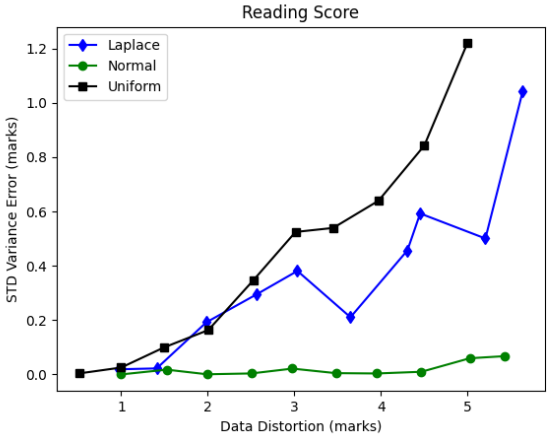
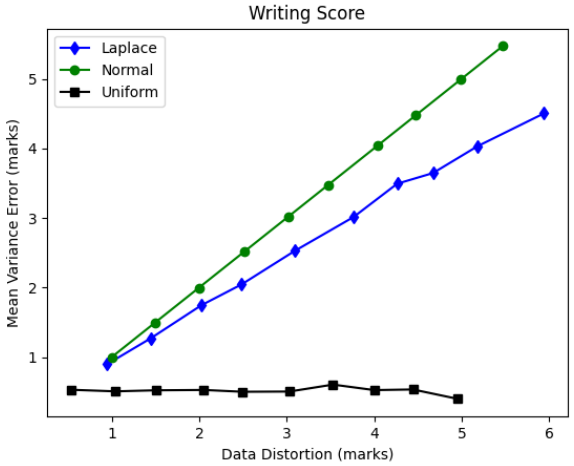
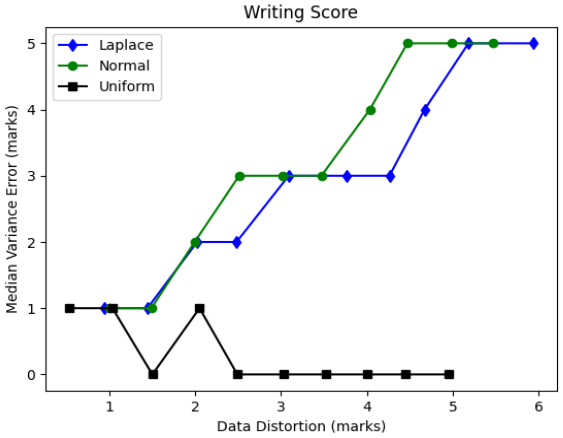


Fig 2: Graphs of reading score





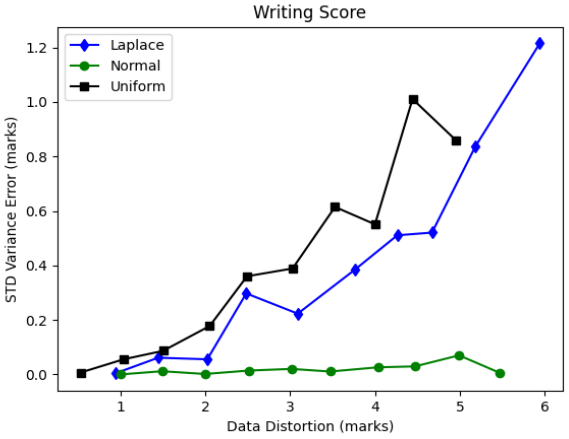


Fig 3: Graphs of writing score

V. RESULTS

In this study, we examined the impact of uniform, normal, and Laplace noise addition techniques on the central tendencies of a dataset containing students' scores from three different subjects. As measures of central tendency, we used mean, median, and standard deviation. After each noise addition, we calculated the variance error for each metric. We performed ten iterations for each noise technique with different parameters and calculated the data distortion i.e. mean absolute difference between the new values and the original values for each. Then the results were plotted on a graph for comparison with the data distortion on the x axis and the variance error for each metric on the y axis. The results show that uniform noise performed better than normal and Laplace noise in preserving the mean of original dataset, as indicated by the mean variance error graph. The above result is attributed to the uniform distribution property for generating random values uniformly across the range, without introducing larger values that may affect the mean of dataset. Furthermore, the graph for median variance error showed that uniform noise had a value closer to 0 compared to other techniques, indicating that it preserved the median of the original dataset better. This can be attributed to the uniform distribution's property of generating values uniformly across a range, which slightly affects the relative ordering of the dataset. For preserving the standard deviation of original dataset, normal noise performed better than uniform and Laplace noise, especially for higher data distortion, as indicated by standard deviation variance error graph. This result for normal distribution can be attributed to the bell shaped curve obtained in the distribution graph, where the majority of the generated values are relatively close to the distribution's mean, causing only a slight change in the dataset's standard deviation after adding normal noise.

VI. CONCLUSIONS

Our experimentation aimed to study effect of normal, uniform, and Laplace noise on the central tendencies of a dataset, and on analyzing the graphs generated, we conclude that adding uniform noise is the preferred technique for preserving the mean and median of the original dataset. The property of uniform distribution to add random values uniformly across a range without introducing larger values helps in avoiding a significant shift in the mean and median of the dataset. The bell shaped curve of normal noise, produces values that are near to the mean making it preferable to preserve the original dataset's standard deviation. As a result, introducing normal noise exhibits just a small modification in the standard deviation of datasets, making it better for preserving the standard deviation.

VII. REFERENCES

[1] Chamikara, M. A. P., Bertok, P., Liu, D., Camtepe, S., & Khalil, I. (2018). Efficient data perturbation for privacy preserving and accurate data stream mining. Pervasive and Mobile Computing.
[2] Salloum, S.A., Al-Emran, M., Monem, A.A., Shaalan, K.: Using text mining techniques for extracting information from research articles. In: Studies in Computational Intelligence, vol. 740. Springer (2018).
[3] Salloum, S.A., Al-Emran, M., Abdallah, S., Shaalan, K.: Analyzing the Arab gulf newspapers using text mining techniques. In: International Conference on Advanced Intelligent Systems and Informatics, pp. 396–405 (2017).
[4] Maheyzah Md Siraj, Nurul Adibah Rahmat, and Mazura Mat Din. 2019. A Survey on Privacy Preserving Data Mining Approaches and Techniques. In Proceedings of the 2019 8th International Conference on Software and Computer Applications (ICSCA '19). Association for Computing Machinery, New York, NY, USA, 65–69.
[5] Elnaz Lashgari, Dehua Liang, Uri Maoz,Data augmentation for deep-learning-based electroencephalography,Journal of Neuroscience Methods,Volume 346,2020.
[6] A. E. Ouadrhiri and A. Abdelhadi, "Differential Privacy for Deep and Federated Learning: A Survey," in IEEE Access, vol. 10, pp. 22359-22380, 2022, doi: 10.1109/ACCESS.2022.3151670.
[7] H. Ren, J. Deng and X. Xie, "GRNN: Generative regression neural network—A data leakage attack for federated learning", 2021.
[8] N. Chaudhary, V. Gupta, K. Sandhir, R. Gupta, S. Chhabra and A. K. Singh, "Privacy Preserving Ensemble Learning Classification Model for Mental Healthcare," 2022 Seventh International Conference on Parallel, Distributed and Grid Computing (PDGC), Solan, Himachal Pradesh, India, 2022, pp. 513-518, doi: 10.1109/PDGC56933.2022.10053268.
[9] Aghasian, E., Garg, S. and Montgomery, J., 2018. A privacy-enhanced friending approach for users on multiple online social networks. Computers, 7(3), p.42.
[10] J. Shan, Y. Lin and X. Zhu, "A New Range Noise Perturbation Method based on Privacy Preserving Data Mining," 2020 IEEE International Conference on Artificial Intelligence and Information Systems (ICAIIS), Dalian, China, 2020, pp. 131-136, doi: 10.1109/ICAIIS49377.2020.9194850.
[11]ShashidharVirupaksha& D. Venkatesulu (2022) Subspacebased aggregation for enhancing utility, information measures, and cluster identification in privacy preserved data mining on high-dimensional continuous data, International Journal of Computers and Applications
[12] M.A.P. Chamikara, P. Bertok, I. Khalil, D. Liu, S. Camtepe,Privacy preserving distributed machine learning with federated learning,Computer Communications,Volume 171,2021.
[13]M.A.P. Chamikara, P. Bertok, D. Liu, S. Camtepe, I. Khalil,Efficient privacy preservation of big data for accurate data mining,Information Sciences,Volume 527,2020.
[14] Virupaksha, S., Dondeti, V. Anonymized noise addition in subspaces for privacy preserved data mining in high dimensional continuous data. Peer-to-Peer Netw. Appl. 14, 1608–1628 (2021).
[15] Weibei Fan, Jing He, Mengjiao Guo, Peng Li, Zhijie Han, Ruchuan Wang,Privacy preserving classification on local differential privacy in data centers,Journal of Parallel and Distributed Computing,Volume 135,2020