Accuracy of Gaussian Naive Bayes and Logistic Regression Models with Differential Privacy

1st Minh Dang Truong
Department of Computer Science
Denison University
Granville, Ohio
truong_m1@denison.edu

2nd Andrew Pham
Department of Computer Science
Denison University
Granville, Ohio
pham_l2@denison.edu

3rd Minh Nguyen
Department of Computer Science
Denison University
Granville, Ohio
nguyen_v2@denison.edu

***Abstract*—In this paper, we explore the impact of differential privacy on the accuracy of two classification models: Gaussian Naive Bayes and Logistic Regression. Through comprehensive experimentation, we draw two key conclusions. Firstly, we find that the choice of datasets significantly influences the accuracy of both models under differential privacy. Secondly, we observe distinct behavior in the models' responsiveness to privacy constraints. Specifically, the Gaussian Naive Bayes model demonstrates a higher level of robustness to differential privacy compared to Logistic Regression. Notably, even with a well-constructed dataset, the Gaussian Naive Bayes model consistently achieves high accuracy, while the accuracy of Logistic Regression exhibits noticeable fluctuations.**

***Index Terms*—Gaussian Naive Bayes, Logistic Regression, differential privacy, accuracy**

I. Introduction

In recent years, the proliferation of data-driven technologies has led to unprecedented advancements in various domains, ranging from healthcare to finance and beyond. Machine learning models, powered by algorithms capable of learning patterns from vast datasets, have become integral tools for decision-making, automation, and prediction. Classification algorithms, in particular, have proven instrumental in various tasks, such as spam filtering, image recognition, and medical diagnosis.

As the prevalence of machine learning applications grows, there is an increase in concerns about the potential misuse or compromise of sensitive data. The very success of these algorithms relies on access to comprehensive datasets, often containing personally identifiable information. Consequently, there is a pressing need to reconcile the benefits of machine learning with the imperative to protect individual privacy.

Differential privacy [1] has emerged as a foundational concept in the privacy-preserving machine learning field. It offers a rigorous mathematical framework to quantify and control the disclosure of individual information, even in the presence of powerful adversaries. By introducing controlled noise to the training data or model parameters, differential privacy ensures that the inclusion or exclusion of any single data point does not unduly influence the outcomes.

This paper focuses on assessing the impact of differential privacy on the accuracy of two foundational classification models: Gaussian Naive Bayes and Logistic Regression. Both models serve as cornerstones in the machine learning toolkit, with Gaussian Naive Bayes excelling in simplicity and Logistic Regression being a powerful tool for binary classification. Understanding how these models perform under the constraints of differential privacy is crucial for building privacy-preserving machine learning applications.

The objectives of this study include:

- Evaluating the accuracy of differentially private Gaussian Naive Bayes and Logistic Regression models across varying levels of privacy parameters.
- Investigating how datasets can affect the accuracy of differentially private models.

The remainder of this paper is organized as follows: Section 2 provides background knowledge on Gaussian Naive Bayes, Logistic Regression, and differential privacy. Section 3 details the methodology employed in our study, including dataset selection, model training, and evaluation metrics. Section 4 presents and discusses the results of our experiments. Finally, Section 5 summarizes our findings, discusses their implications, and outlines potential avenues for future research.

By undertaking this project, we hope to contribute to the ongoing discourse on privacy-preserving machine learning and offer practical guidance for researchers, practitioners, and policymakers grappling with the intersection of data science and data privacy.

II. Background Information

A. Gaussian Naive Bayes: A Probabilistic Framework

Gaussian Naive Bayes (GNB) [2], deeply rooted in the principles of Bayes' Theorem, works by computing the posterior probability of each class based on input features under the assumption of conditional independence among these features. This assumption, though an oversimplification at times, significantly enhances computational efficiency, especially in high-dimensional data contexts. GNB, diverging from many machine learning algorithms, does not engage in iterative optimization techniques like gradient descent due to its probabilistic nature. This property allows direct calculation of probabilities from the training dataset.

Another notable aspect of GNB is the incorporation of prior probabilities, which reflects pre-existing knowledge about class distributions. This is particularly advantageous in situations with inherent likelihood differences among classes or imbalanced datasets. The algorithm presumes continuous features associated with each class follow a Gaussian distribution. While this assumption facilitates computational simplicity, it may not always align with the distributions of real-world data, thus posing potential limitations.

B. Logistic Regression: An Iterative Optimization Approach

Logistic Regression (LR) [3] is a widely used machine learning algorithm primarily designed for binary classification tasks, where the outcome variable consists of two classes. At the core of LR is the sigmoid (logistic) function [4], which transforms real-valued input into a range between 0 and 1. This function is pivotal for mapping the linear combination of input features to a probability score.

The algorithm assumes a linear relationship between the input features and the log-odds (logit) [5] of the probability of the positive class. The output of the sigmoid function represents the probability that a given input belongs to the positive class. If this probability exceeds a specified threshold (typically 0.5), the instance is classified as the positive class; otherwise, it is classified as the negative class.

LR models the log-odds of the probability of the positive class and employs maximum likelihood estimation [6] during training. The goal is to adjust the model parameters to maximize the likelihood of the observed outcomes given the input features. The performance of LR is often evaluated using the log loss [7] (cross-entropy) as the cost function, measuring the difference between predicted probabilities and actual class labels.

LR models are interpretable, as the coefficients associated with each feature indicate the strength and direction of their influence on the predicted probability. This interpretability makes Logistic Regression a valuable tool for binary classification tasks. The algorithm finds applications across diverse fields, including medicine for predicting disease occurrence, marketing for customer churn prediction, and finance for credit scoring.

C. Differential Privacy

Differential privacy is a sophisticated framework rooted in rigorous mathematical principles designed to address the increasing concerns about protecting sensitive information during data analysis [8]. At its core, differential privacy introduces a formal and quantifiable guarantee that the inclusion or exclusion of any individual's data does not impact the outcome of a computation. This assurance is particularly valuable in scenarios where privacy is of paramount importance.

One of the primary mechanisms employed in differential privacy is the intentional introduction of noise into the computation or analysis of data [9]. This noise serves as a protective measure, preventing precise inferences about individual data points from the output. The main goal is to ensure that the privacy of each individual is preserved, regardless of whether their data is part of the analysis. This protection extends to both local and global levels of data aggregation: local differential privacy [10] perturbs individual data points before aggregation, while global differential privacy adds noise to the final aggregated result.

An inherent trade-off exists between privacy and utility in differential privacy. While perfect privacy could be achieved by adding excessive noise, the output of such computation will be of little utility. Balancing between privacy and utility is crucial to enable meaningful analysis while protecting sensitive information. Differential privacy also introduces the concept of a privacy budget, representing the cumulative amount of privacy loss permitted over multiple analyses. Once this budget is exhausted, further analyses may compromise individual privacy.

III. Evaluation Methodology

We first describe the datasets that we use for evaluation, followed by the description of our experimental setup. We then present the accuracy of the two models when trained on five different datasets and study in detail how the dataset's characteristics affect the accuracy of each model across a wide range of privacy budgets.

A. Datasets

All the datasets we use for our experiments are from the UC Irvine Machine Learning Repository. We purposefully selected datasets that only contain numeric values (either discrete or continuous) since the Gaussian Naive Bayes and Logistic Regression models only work on numeric datasets.

Breast cancer Wisconsin (diagnostic) [11]. This dataset contains features computed from a digitized image of a fine needle aspirate of a breast mass. Some typical features are radius, texture, perimeter, area, smoothness, and symmetry. A binary diagnosis (malignant or benign) is then labeled for each record of feature values. The dataset is 570 rows long and 30 columns/features wide. Within those around 570 rows, 62% of them are training examples for the label "benign", and the remaining rows (37%) are for "malignant".

Wine quality [12]. Originally, there were two datasets related to red and white variants of the Portuguese "Vinho Verde" wine, but due to privacy and logistic issues, only physicochemical and sensory variables are available; for instance, there is no data about grape types, wine brand, wine selling price, etc. For our experiments, we only use the red wine dataset. Given those variables of each example, we have a 0-10 rating for its quality. The dataset is roughly $1,600$ rows long and 12 columns/features wide. The label distribution is as follows: 42% for rating 5, 39% for rating 6, 12% for rating 7, 3% for rating 4, 1% for rating 8, 0.6% for rating 3, and 0% for the remaining ratings.

Rice (Cammeo and Osmancik) [13]. This dataset focuses on two certified rice varieties cultivated in Turkey: the Osmancik species, extensively planted since 1997, and the Cammeo species introduced in 2014. A total of $3,810$ rice grain images were taken and processed for the two species, and feature inferences were made from these images. Specifically, seven morphological features were obtained for each grain of rice, including area, perimeter, major axis length, minor axis length, eccentricity, convex area, and extent. Within those $3,810$ rows, 57.3% are training examples for the Cammeo species, and the rest (42.7%) are for the Osmancik one.

Letter recognition [14]. The dataset categorizes a large number of black-and-white rectangular pixel displays, assigning each to one of the 26 capital letters in the English alphabet (A-Z). These letter images were generated from 20 different fonts, and each letter, across these diverse fonts, underwent random distortions, resulting in a dataset of $20,000$ distinct stimuli. Each stimulus was transformed into 16 basic numerical attributes, encompassing statistical moments and edge counts, which were subsequently normalized to fit within an integer range from 0 to 15. Some notable numerical attributes are x-box, y-box, x-bar, y-bar, width, and height. Regarding label distribution, each letter has around 3-4$\%$ of the total dataset as its training examples.

MAGIC gamma telescope [15]. This dataset is computer-generated using Monte Carlo simulation to replicate the registration process of high-energy gamma particles in a ground-based atmospheric Cherenkov gamma telescope that employs imaging techniques. This type of telescope observes high-energy gamma rays by detecting the Cherenkov radiation emitted during electromagnetic showers initiated by the gamma rays in the atmosphere. The Cherenkov radiation, occurring in visible to UV wavelengths, penetrates the atmosphere and is captured by the detector. This captured information enables the reconstruction of the shower parameters. The available information consists of pulses left by the incoming Cherenkov photons on the photomultiplier tubes, arranged in a plane, the camera. Depending on the primary gamma energy, a varying number of Cherenkov photons (ranging from a few hundred to around $10,000$) are collected in patterns forming the shower image. This allows for statistical classification between images caused by primary gammas (signal) and those resulting from hadronic showers initiated by cosmic rays in the upper atmosphere (background). The dataset is approximately $19,000$ rows long and 10 columns wide. Within those $19,000$ rows, 64.8% are training examples for the label "gamma" (signal) and the remaining 35.2% are for "hadron" (background).

The datasets are diverse in dimensionality, number of records and labels, and label distribution/skewness. For example, we used datasets with a small number of records (wine quality dataset) and ones with a larger number of records (letter recognition and MAGIC gamma telescope datasets). The dataset can also be simple with a few features (rice and MAGIC gamma telescope datasets) or complicated with many features (breast cancer and letter recognition datasets). The number of labels in our training dataset also determines the complexity of our classification model. Therefore, we used both simple datasets with binary labels (breast cancer, rice, and MAGIC game telescope datasets) and more complicated ones with a much larger number of labels (wine quality and letter recognition datasets). In addition, the overall accuracy of our classification model also depends on how balanced/skewed a dataset is. Hence, we decided to include both highly skewed (wine quality dataset) and more balanced ones (rice and letter recognition datasets). The motivation for our dataset selection is that we want to evaluate our models in the most comprehensive way by using a diverse group of datasets because we believe the resulting model will be heavily affected by the dataset it is trained on.

B. Experimental Setup

In this project, we attempted to comprehensively evaluate the accuracy of differentially private GNB and LR classification models across various datasets. The experimental setup involved an 80/20 split of the datasets, with 80% of the data dedicated to training the models and the remaining 20% reserved for rigorous testing. The decision to employ an 80/20 split for the datasets was motivated by the need to balance the training effectiveness and robust model evaluation.

Allocating 80% of the data to the training set was aimed at providing the machine learning models with substantial information to capture the underlying patterns inherent in datasets. This larger training set contributes to the efficient training of the models, especially when dealing with datasets of small size (wine quality dataset). At the same time, reserving 20% of the data for testing plays a crucial role in assessing the generalization capabilities of the models. The testing set consists of examples the models have not encountered during training, offering a fair evaluation of how well they can extend their learned knowledge to new instances. This emphasis on generalization aligns with a fundamental goal in machine learning — building models that can perform well on real-world data beyond the training set.

We used the GNB implementation from the IBM differential privacy library. In addition to tuning the epsilon (privacy budget) parameter for the model, specifying the bounds of the data is crucial in preventing potential privacy leaks. Differentially private models provide privacy guarantees by introducing controlled noise to the computations to protect individual data points from being uniquely identified. The bounds of data are essential for ensuring that the noise added to the results of these computations is appropriately calibrated to the data. Failure to provide these bounds can result in privacy leak warnings from the IBM library, indicating that the privacy guarantees of the differentially private mechanism may be compromised. We achieved this by observing values in our datasets first and then choosing a common bound for all the datasets, not an individual bound for each dataset.

We also used the differentially private LR model from the IBM differential privacy library. To prevent potential privacy leaks, we need to give the model the maximum l2 norm of any rows of the datasets. If not specified, the maximum norm is taken from the dataset the first time we provide the data to the model. However, this practice will result in a privacy leak warning from the library as it reveals information about the data.

More importantly, choosing the appropriate number of iterations is crucial to the model's accuracy. LR is trained using iterative optimization algorithms like gradient descent, where the model parameters are updated to minimize the cost function. The number of iterations directly impacts whether the model converges to an optimal solution or not. If the maximum number of iterations is set too low, the algorithm might terminate prematurely, resulting in suboptimal parameter values and an underfit model. In addition, inadequate iterations can lead to underfitting, where the model fails to capture the underlying patterns in the data, thus giving low prediction accuracy. Therefore, we manually tuned the maximum number of iterations until the model converges for each dataset. Specifically, the model will have different numbers of maximum iterations for different training datasets.

IV. Result discussion

In this section, we conduct a detailed analysis of the graphical representations derived from our experimental data. The aim is to meticulously interpret the patterns and anomalies evident in the graphs, thereby providing a comprehensive understanding of the two models in the context of differential privacy.

A. Accuracy of differentially private Gaussian Naive Bayes model

In the non-private setting, the GNB model shows high accuracy (above 75%) when it uses the “breast cancer”, “rice”, and “gamma telescope” training datasets, which is illustrated in figure 1a, 1c, and 1e, respectively. The reason for this high accuracy is that the model, in these three cases, is a simple binary classifier, and the label distribution in each training dataset is quite balanced, meaning we have enough training examples for every label. However, with differential privacy, the accuracy of the model that uses the “breast cancer” training dataset fluctuates wildly (between 0.4 and 0.6). The expected pattern of “the bigger the privacy budget, the less privacy the model preserves, the more accurate the model becomes” does not show here. While the model is highly accurate in a non-private context, its performance becomes less predictable when noise is applied. This unpredictability can be explained by the tiny training dataset size (569 records) since, in a small dataset, noise can easily overwhelm real values. However, for the private model that uses the “rice” and “magic gamma” training datasets, we see the result that we expected - an overall upward trend in accuracy. The accuracy gets closer to the baseline of the non-private setting when we increase our epsilon value from 10^{-2} to 10^{2}. The bigger sizes of these two training datasets ($3,810$ records and $19,000$, respectively), combined with their smaller number of features (7 and 10, respectively), might have contributed to the emergence of that pattern.

For the two remaining training datasets: “wine quality” and “letter recognition”, the GNB model’s accuracy is illustrated in figure 1b and 1d, respectively. The resulting model already shows modest accuracy (0.55 and 0.65, respectively) without differential privacy. This average accuracy is probably because the classification task becomes more complex as more labels need to be classified. When differential privacy is applied, the model trained on the “wine quality” dataset shows lower but consistent accuracy (0.4) as long as the privacy budget is not too stringent. This case is interesting because our accuracy cannot be higher than 0.4, although the privacy budget has become more generous. For the private model trained on the “letter recognition” dataset, the result is very surprising to us because the accuracy drops to below 0.1, meaning the model classifies wrong on almost every instance. We believe this is because the task of letter recognition is inherently complex (i.e., distinguish between “i” and “l” or “b” and “d”), so the model needs a much higher-quality dataset to be trained on to gain better accuracy in a private setting.

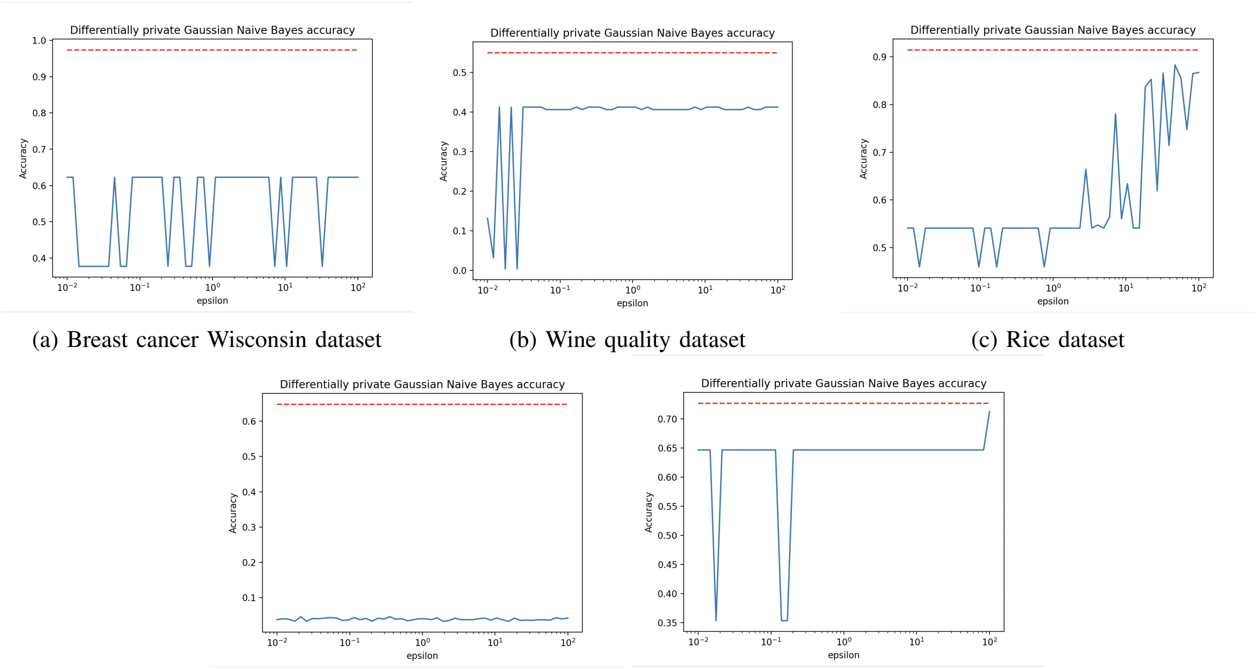
B. Accuracy of differentially private Logistic Regression model

Similar to the GNB model, in the non-private setting, the LR model shows high accuracy (above 79%) when it uses the “breast cancer”, “rice”, and “magic gamma” training datasets, which is illustrated in figure 2a, 2c, and 2e, respectively. Again, we believe the simple binary classification task and the balanced label distribution in these three datasets contribute to this high accuracy. However, with the introduction of differential privacy, what we observed in the GNB approach no longer shows here. Although the “rice” dataset is better than the “breast cancer” dataset and the “magic gamma” dataset is much better than both (in the sense that there are more records and fewer features), the model’s accuracy, in all three cases, fluctuates significantly and lacks predictability. The reason for this behavior may lie in the LR model itself: it is not robust to differential privacy. We believe that as soon as differential privacy is introduced to our LR model, accuracy will drop and become unpredictable no matter how good our dataset is and how generous our privacy budget is.

For the two remaining training datasets “wine quality” and “letter recognition”, the LR model’s accuracy is illustrated in figure 2b and 2d, respectively. In a non-private setting, the resulting model shows higher accuracy (78% versus 58%) if trained on the “letter recognition” dataset than the “wine quality” dataset. This difference may be because, with similar dimensionality, there are significantly more records in the “letter recognition” dataset (around $20,000$ versus $1,600$ records), and the training examples for each label are more evenly distributed in the “letter recognition” dataset. When differential privacy is applied, though the model’s accuracy is lower, the supposed pattern of “the bigger the privacy budget, the less privacy the model preserves, the more accurate the model becomes” shows up again. Specifically, we can still see an overall upward trend in accuracy when we increase our privacy budget from 10^{-2} to 10^{2}. Compared to our GNB model when trained on the “letter recognition” dataset, our LR model seems to be better suited.

V. Conclusion

In conclusion, our investigation into the impact of differential privacy on the accuracy of Gaussian Naive Bayes



(a) Breast cancer Wisconsin dataset

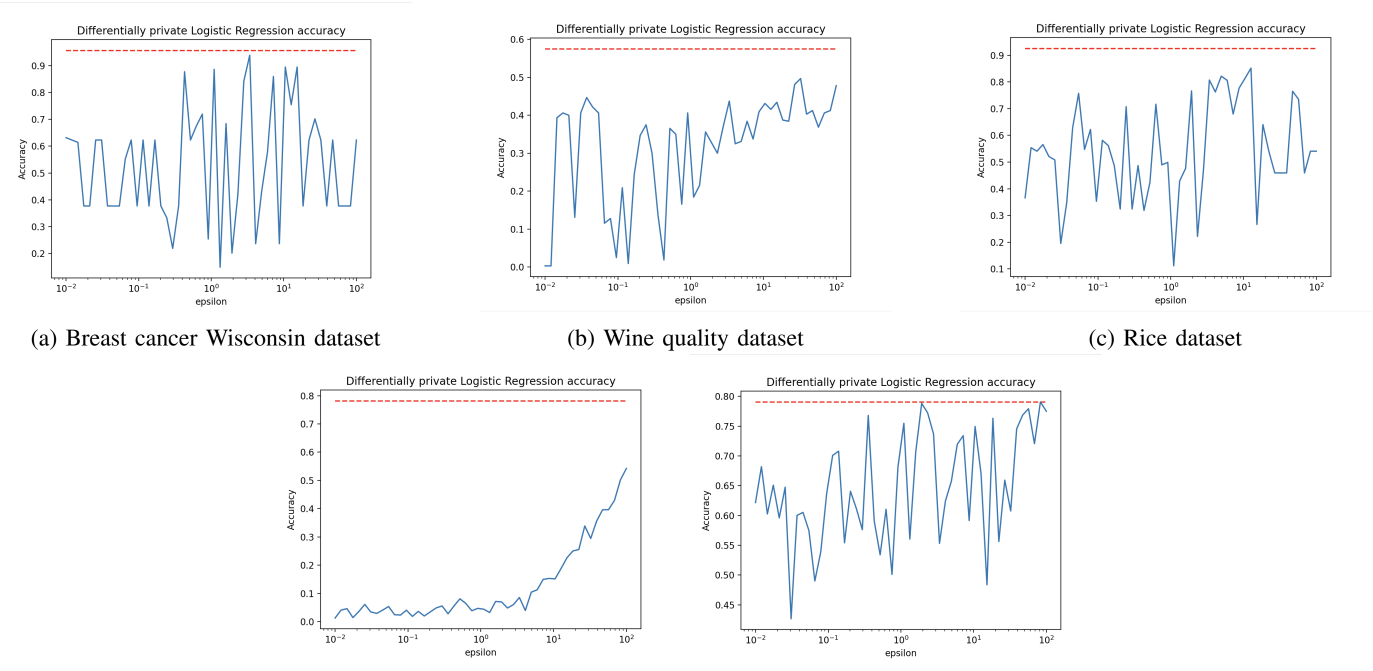
(b) Wine quality dataset

(c) Rice dataset

(d) Letter recognition dataset

(e) MAGIC gamma telescope dataset

Fig. 1: Accuracy of differentially private Gaussian Naive Bayes classifier under five different datasets



(a) Breast cancer Wisconsin dataset

(b) Wine quality dataset

(c) Rice dataset

(d) Letter recognition dataset

(e) MAGIC gamma telescope dataset

Fig. 2: Accuracy of the differentially private Logistic Regression classifier under five different datasets

and Logistic Regression has provided valuable insights into the interplay between privacy constraints and classification model performance. Our first key finding underscores the significant influence of dataset selection on the accuracy of both models under differential privacy. The sensitivity of model accuracy to varying datasets highlights the need for careful consideration when implementing privacy-preserving measures. The second conclusion reveals nuanced differences in the models' responses to privacy constraints. Gaussian Naive Bayes emerges as the more robust model under differential privacy, consistently maintaining high accuracy levels, even with a well-constructed dataset. In contrast, Logistic Regression exhibits noticeable accuracy fluctuations, suggesting a higher susceptibility to privacy constraints. Our study adds depth to the ongoing discourse on the practical implications of differential privacy on classification models. The project also underscores the importance of thoughtful model selection and parameter tuning in privacy-conscious data analytics.

Moving forward, additional research efforts can delve deeper into optimizing Logistic Regression under the constraints of differential privacy. One avenue for exploration involves developing novel regularization techniques tailored to enhance LR's robustness while ensuring privacy preservation.

References

[1] C. Dwork, "Differential privacy," in *Automata, Languages and Programming*, M. Bugliesi, B. Preneel, V. Sassone, and I. Wegener, Eds. Berlin, Heidelberg: Springer Berlin Heidelberg, 2006, pp. 1–12.

[2] G. H. John and P. Langley, "Estimating continuous distributions in bayesian classifiers," in *Proceedings of the Eleventh Conference on Uncertainty in Artificial Intelligence*, ser. UAI'95. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc., 1995, p. 338–345.

[3] J. Tolles and W. J. Meurer, "Logistic Regression: Relating Patient Characteristics to Outcomes," *JAMA*, vol. 316, no. 5, pp. 533–534, 08 2016. [Online]. Available: https://doi.org/10.1001/jama.2016.7653

[4] S. R. Dubey, S. K. Singh, and B. B. Chaudhuri, "A comprehensive survey and performance analysis of activation functions in deep learning," *CoRR*, vol. abs/2109.14545, 2021. [Online]. Available: https://arxiv.org/abs/2109.14545

[5] Wikipedia, "Logit." [Online]. Available: https://en.wikipedia.org/wiki/Logit

[6] C. Gourieroux and A. Monfort, "Asymptotic properties of the maximum likelihood estimator in dichotomous logit models," *Journal of Econometrics*, vol. 17, no. 1, pp. 83–97, 1981. [Online]. Available: https://www.sciencedirect.com/science/article/pii/0304407681900609

[7] G. Cybenko, D. P. O'Leary, and J. Rissanen, *The Mathematics of Information Coding, Extraction and Distribution*, 1999.

[8] C. Dwork, "A firm foundation for private data analysis," *Commun. ACM*, vol. 54, no. 1, p. 86–95, jan 2011. [Online]. Available: https://doi.org/10.1145/1866739.1866758

[9] C. Dwork, F. McSherry, K. Nissim, and A. Smith, "Calibrating noise to sensitivity in private data analysis," in *Theory of Cryptography*, S. Halevi and T. Rabin, Eds. Berlin, Heidelberg: Springer Berlin Heidelberg, 2006, pp. 265–284.

[10] U. Erlingsson, V. Pihur, and A. Korolova, "Rappor: Randomized aggregatable privacy-preserving ordinal response," in *Proceedings of the 2014 ACM SIGSAC Conference on Computer and Communications Security*, ser. CCS'14. ACM, Nov. 2014. [Online]. Available: http://dx.doi.org/10.1145/2660267.2660348

[11] W. Wolberg, O. Mangasarian, N. Street, and W. Street, "Breast Cancer Wisconsin (Diagnostic)," UCI Machine Learning Repository, 1995, DOI: https://doi.org/10.24432/C5DW2B.

[12] P. Cortez, A. Cerdeira, F. Almeida, T. Matos, and J. Reis, "Wine Quality," UCI Machine Learning Repository, 2009, DOI: https://doi.org/10.24432/C56S3T.

[13] "Rice (Cammeo and Osmancik)," UCI Machine Learning Repository, 2019, DOI: https://doi.org/10.24432/C5MW4Z.

[14] D. Slate, "Letter Recognition," UCI Machine Learning Repository, 1991, DOI: https://doi.org/10.24432/C5ZP40.

[15] R. Bock, "MAGIC Gamma Telescope," UCI Machine Learning Repository, 2007, DOI: https://doi.org/10.24432/C52C8B.