Evaluating Differential Privacy Mechanisms in Ensemble Models for Adult Income Prediction

No Author Given

No Institute Given

Abstract. Machine learning models often face a trade-off between predictive performance and privacy when trained on sensitive data. Differential Privacy (DP) offers formal privacy guarantees by injecting noise during data analysis, but this typically comes at the cost of reduced model accuracy. In this paper, we present a comprehensive empirical study of the privacy-utility trade-off for ensemble classification models on a real-world dataset. We focus on the UCI *Adult* income dataset and three popular ensemble methods (Random Forest, LightGBM, and XGBoost). We apply two common DP mechanisms, Laplace and Gaussian noise, to the training data at varying privacy budgets ϵ, and evaluate the impact on model performance. Our results show that model accuracy and F1-score degrade as privacy becomes stricter (lower ϵ), with particularly severe drops in recall due to models favoring majority class predictions under heavy noise. Among the mechanisms, Laplace noise generally yields better performance than Gaussian noise at moderate privacy levels (e.g. $\epsilon \geq 1$), while both mechanisms converge to poor performance in the extreme privacy regime. We also find that XGBoost is more robust to DP noise compared to Random Forest and LightGBM, maintaining higher accuracy and F1 across the ϵ range. This work provides insights into how different DP noise mechanisms and ensemble models balance privacy and utility, informing practitioners about effective choices for privacy-preserving machine learning.

Keywords: Differential Privacy · Ensemble Learning · Random Forest · XGBoost · LightGBM · Income Prediction

1 Introduction

In recent years, privacy-preserving machine learning has gained prominence as models are frequently trained on sensitive user data. One formal approach to protecting privacy is *Differential Privacy* (DP), which ensures that the inclusion or exclusion of any single data point minimally affects the output of an algorithm [5]. DP is typically achieved by adding calibrated noise to data or model computations, providing a quantifiable privacy guarantee at the expense of some loss in accuracy.

While DP has been successfully applied to simple statistical analyses and deep learning [1], its impact on classical machine learning models, such as ensemble methods, warrants thorough investigation. Ensemble models like *Random*

Forests [3, 6] and gradient boosting machines (e.g., *XGBoost* [4] and *LightGBM* [11]) are widely used due to their high accuracy [8, 12]. However, training data may contain sensitive information (for example, the UCI *Adult* income dataset used in this study contains individuals' demographic and financial data). Incorporating DP into such models could mitigate privacy risks but may degrade performance. The key question is how large that degradation is, and whether some models or DP mechanisms offer a better privacy-utility balance than others.

This paper empirically investigates the impact of DP on ensemble classifiers. We compare two noise injection mechanisms—Laplace and Gaussian—applied during training, and evaluate three representative methods (Random Forest, LightGBM, XGBoost) on the *Adult* dataset. By varying the privacy budget ϵ, we analyze trade-offs between stronger privacy (lower ϵ) and performance across Accuracy, Precision, Recall, F1-score, and Specificity.

Our contributions are as follows: (1) We provide a systematic comparison of Laplace vs. Gaussian noise for different levels of privacy on ensemble models, highlighting their impact on various performance metrics. (2) We analyze the relative robustness of different ensemble learning algorithms under privacy constraints, finding that boosted trees (XGBoost, LightGBM) can sustain higher utility than bagged trees (Random Forest) when noise is added. (3) We discuss practical insights, such as which DP mechanism might be preferable and how much accuracy one might trade for a given privacy level, to guide practitioners building privacy-preserving classifiers.

The remainder of the paper is organized as follows. Section 2 reviews related work on differentially private machine learning and prior approaches to privatizing tree-based models. Section 3 describes the dataset, models, DP mechanisms and experimental setup. Section 4 presents the results of our experiments, including visualizations of performance trends. We discuss the implications of these results in the context of privacy-utility trade-offs. Finally, Section 5 concludes the paper with a summary and suggestions for future work.

2 Related Work

Differential Privacy has been applied to various machine learning techniques. Early research on DP for machine learning focused on simple models and query answering. For example, Friedman and Schuster[9] presented one of the first differentially private decision tree algorithms by adding noise to the splitting criteria counts. Similarly, efforts like Abadi et al. [1] brought DP to deep neural networks through a method known as DP-SGD, which adds noise to gradient updates during training.

Specific to ensemble methods, several studies have explored how to build or adapt these models under DP constraints. Hou et al. [10] proposed a differentially private random forest (DPRF) algorithm that introduces noise in the tree-building process and allocates privacy budget across the forest to improve utility. Their approach improved on earlier methods by optimizing how noise is

distributed at different tree depths. Another line of work (e.g., [9]) has examined random decision trees and other tree ensemble techniques with noise injection to preserve privacy.

Our approach in this paper is closest to an *input perturbation* strategy: instead of modifying the learning algorithm itself (as in [10] who adjust tree node splits, or [1] who modify the training process for neural networks), we simply add noise to the training data features according to DP mechanisms. This approach has the advantage of being straightforward and model-agnostic: any standard classifier can be trained on the noised dataset without alteration. However, the challenge is that input noise can significantly affect model accuracy, especially for complex models. By comparing Laplace and Gaussian noise, our work also relates to prior studies on the efficacy of different DP mechanisms. The Laplace mechanism is the canonical approach for DP with pure ϵ-privacy, while the Gaussian mechanism achieves (ϵ, δ)-privacy and may add less noise when a small probability of failure δ is allowed [5]. Understanding which mechanism yields better performance for a given task is important for practitioners.

Overall, while prior works have established theoretical frameworks and advanced algorithms for differentially private learning, our contribution is an empirical evaluation focusing on ensemble classifiers. We aim to quantify how much accuracy and other metrics deteriorate under various privacy levels and identify which methods offer more favorable trade-offs in practice.

3 Methodology

3.1 Dataset and Preprocessing

We use the UCI Adult dataset [2]. The goal is to predict whether an individual's income exceeds \$50K/year based on census features. The dataset contains 48,842 instances (after removing examples with missing values) with 14 input features and a binary target label (income $>$50K or $\leq$50K). Input features include both numeric attributes (e.g., *age*, *hours-per-week*, *education-num*) and categorical attributes (e.g., *education level*, *occupation*, *marital-status*, *race*, *sex*, etc.).

Before training models, we perform data cleaning and preprocessing. We discard records with missing values (a standard practice for this dataset, resulting in a slightly smaller dataset than the original 48K). We then split the data into a training set and a test set using the provided division from the UCI repository (approximately 32,561 training and 16,281 test instances). For numeric features, we apply standardization (zero mean, unit variance scaling) to ensure they are on comparable scales; for categorical features, we apply one-hot encoding to convert categories into binary indicator variables. The preprocessing pipeline produces a numeric feature matrix for training and testing.

3.2 Differential Privacy Mechanisms for Input Perturbation

Let X_{train} and X_{test} denote the processed training and test feature matrices, and $y_{\text{train}}, y_{\text{test}}$ the corresponding labels. To enforce differential privacy during

model training, we add random noise to the training feature matrix X_{train} using two different mechanisms:

- **Laplace Mechanism**: We add noise drawn from a Laplace distribution to each entry of X_{train}. The Laplace mechanism for achieving ϵ-DP uses noise scaled to the L_1-sensitivity of the query. In our case, if we assume each feature is bounded (or has been scaled) to a known sensitivity Δ (difference in feature value caused by one individual's data), then we add noise Laplace$(0, \Delta/\epsilon)$ independently to each feature value. We set $\Delta = 1$ for all features by virtue of our preprocessing (numeric features are standardized and most categorical features are one-hot encoded to 0/1, so a single individual's change can alter a feature by at most 1 unit).
- **Gaussian Mechanism**: We also consider the Gaussian mechanism, which provides (ϵ, δ)-DP. The Gaussian mechanism adds noise from a normal distribution $\mathcal{N}(0, \sigma^2)$, where σ is chosen based on the L_2-sensitivity of the data and the desired (ϵ, δ) parameters. We fix $\delta = 10^{-5}$ (a common choice for DP with large datasets) and compute σ using the standard formula $\sigma = \frac{\sqrt{2\ln(1.25/\delta)}\,\Delta}{\epsilon}$ [5]. In our implementation, we set $\Delta = 1$ analogous to the Laplace case, so $\sigma \approx \frac{\sqrt{2\ln(1.25/\ 10^{-5})}}{\epsilon} \approx \frac{4.91}{\epsilon}$. This means for a given privacy budget ϵ, the Gaussian noise standard deviation is about $4.91/\epsilon$. We add independent Gaussian noise $N(0, \sigma^2)$ to each entry of X_{train}.

Intuitively, smaller values of ϵ (approaching 0) enforce stronger privacy by adding more noise, whereas larger ϵ allow data to be more accurate but less private. An ϵ of ∞ would correspond to no noise (no privacy protection). We experiment with a range of ϵ values $\{0.1, 0.5, 1.0, 2.0, 5.0\}$ to observe privacy-utility trade-offs. Additionally, we include a baseline model trained on the original data with *no DP noise* (equivalent to $\epsilon = \infty$) for comparison.

It is worth noting that adding noise to training features is a straightforward way to achieve DP, sometimes referred to as input perturbation. However, this approach can be conservative in terms of the amount of noise needed because it perturbs all features rather than a sufficient statistic or model parameters. We therefore expect notable performance degradation, especially at low ϵ. More advanced DP training methods (such as perturbing tree node splits or gradient perturbation in neural networks) might achieve better accuracy for the same privacy, but our aim here is to compare Laplace vs Gaussian noise in a controlled, simple setup across different ensemble models.

3.3 Models and Training Procedure

We evaluate three ensemble classification models:

1. **Random Forest (RF)** [3, 6]: an ensemble of decision trees trained via bootstrap aggregation (bagging). We use a RandomForestClassifier from scikit-learn with default hyperparameters (100 trees, Gini impurity for splits, etc.).

2. **LightGBM (LGBM)** [11]: a gradient boosting framework that builds trees sequentially. We use the LightGBM library's classifier (LGBMClassifier) with default settings.
3. **XGBoost (XGB)** [4]: an optimized gradient boosting library. We use XGBClassifier from the XGBoost library with default settings (and with its default behavior of using histogram-based splits and regularization).

Each model is trained on several versions of the training data: - One without any added noise (baseline, no DP). - One for each combination of mechanism $\in \{\text{Laplace}, \text{Gaussian}\}$ and $\epsilon \in \{0.1, 0.5, 1, 2, 5\}$.

In total, for each model, we train 1 (baseline) + 2 (mechanisms) $\times$ 5 (epsilons) = 11 models. For fairness, all other conditions are kept the same across these runs.

During training on noisy data, the target labels y_{train} remain unchanged (we only noise the features). We assume that label values are not sensitive or that label privacy is out of scope; one could also consider flipping labels for privacy, but that is beyond our current focus.

3.4 Evaluation Metrics

We assess the performance of each model on the test set using a set of standard classification metrics that collectively capture both overall correctness and class-specific behavior. These metrics are particularly important given the imbalanced nature of the dataset, where the majority class (income $\leq$50K) significantly outweighs the minority class (income >\$50K). The evaluation metrics used are as follows:

- **Accuracy**: The proportion of total correct predictions [7]. It is defined as:

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN}$$

- **Precision**: The proportion of predicted positives that are actually positive (income >\$50K). It is defined as:

$$\text{Precision} = \frac{TP}{TP + FP}$$

- **Recall**: Also known as sensitivity or true positive rate, it measures the ability to detect actual positives. It is defined as:

$$\text{Recall} = \frac{TP}{TP + FN}$$

- **F1-Score**: The harmonic mean of precision and recall, which provides a balanced evaluation metric for imbalanced datasets. It is defined as:

$$\text{F1} = 2 \times \frac{\text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}}$$

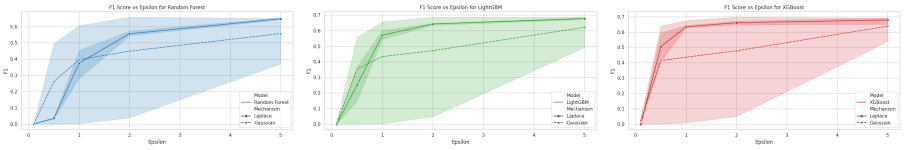
– **Specificity**: The proportion of correctly identified negative instances (income ≤50K). It is defined as:

$$\text{Specificity} = \frac{TN}{TN + FP}$$

The Adult dataset is somewhat imbalanced (24% positive, 76% negative), so focusing only on accuracy can be misleading. F1, precision, and recall provide insight into performance on the minority class. In particular, under extreme noise we anticipate models might predict almost everyone as earning ≤50K to maximize accuracy, which would yield high specificity but near-zero recall for the high-income class. Tracking these metrics helps us characterize such behavior.

4 Results and Analysis

We present the experimental results on the test set, analyzing how different privacy levels and mechanisms affect each model's performance. Figure 1 summarizes the impact of privacy budget ϵ on F1-score for each model and DP mechanism. Figure 2 provides a closer comparison of Laplace vs. Gaussian mechanisms at a representative privacy level ($\epsilon = 1.0$). Finally, Figure 3 shows a comprehensive heatmap of all performance metrics across the different conditions.



(a) Random Forest (b) LightGBM (c) XGBoost

Fig. 1: F1 Score vs. privacy budget ϵ for each model under Laplace vs. Gaussian noise. Each subplot corresponds to one ensemble model. Curves show how the F1-score increases as the privacy budget ϵ becomes larger (less noise). $\epsilon = \infty$ corresponds to the no-noise baseline point on each curve.

4.1 Overall Privacy-Utility Trade-off

As expected, all models suffer performance degradation as the privacy budget ϵ decreases (i.e., noise increases). In the most extreme privacy setting we tested ($\epsilon = 0.1$), the models effectively lost almost all predictive power: F1-scores dropped near zero in most cases (Figure 1). This is because at $\epsilon = 0.1$, the added noise is very large relative to the scale of the features (recall $\sigma \approx 49$ for Gaussian, and Laplace scale = 10 for each feature), completely obscuring any informative signal. In fact, for $\epsilon = 0.1$, the classifiers tended to predict the

majority class for every instance, achieving only about 75–76% accuracy (which equals the proportion of negatives in the data) with virtually zero recall for the positive class.

As privacy constraints are relaxed, utility improves. By $\epsilon = 1.0$, all models regained a substantial portion of their accuracy and F1. For instance, as shown in Figure 1, at $\epsilon = 1$ LightGBM reached an F1 around 0.5–0.6 (depending on mechanism) compared to its no-noise F1 of about 0.71. XGBoost similarly achieved around 0.6–0.63 F1 at $\epsilon = 1$ (versus 0.71 baseline). Random Forest was the most affected, with F1 about 0.38–0.40 at $\epsilon = 1$ (versus 0.67 baseline). By $\epsilon = 5.0$ (the highest finite privacy budget we tested), performance was much closer to the no-DP baseline for all models: e.g., XGBoost had F1 ≈ 0.68 vs. 0.71 baseline, and accuracy within 3 percentage points of the baseline. This indicates that a relatively modest privacy budget (like $\epsilon = 5$, which is still considered a fairly weak privacy guarantee in DP terms) can almost preserve full utility, whereas very strong privacy ($\epsilon \leq 0.5$) comes with a heavy cost in accuracy.

4.2 Comparison of DP Mechanisms: Laplace vs. Gaussian

One of our key questions was whether the choice of noise distribution (Laplace or Gaussian) makes a difference in model performance for a given privacy level. Our findings show that the mechanism does matter, and its impact varies with the privacy budget:

- For more lenient privacy requirements (ϵ around 1 and above), the Laplace mechanism consistently outperformed the Gaussian mechanism in terms of model utility. This trend is evident in Figure 1: the Laplace curve for each model is above the Gaussian curve for $\epsilon \geq 1$. For example, at $\epsilon = 1$, Laplace noise yielded an F1 of 0.57 for LightGBM compared to 0.43 with Gaussian, and for XGBoost 0.63 vs 0.44. At $\epsilon = 2$, the gap persists (LightGBM F1: 0.64 Laplace vs 0.47 Gaussian; XGBoost: 0.66 vs 0.48; Random Forest: 0.56 vs 0.45).
- At very tight privacy levels ($\epsilon < 1$), both mechanisms severely degrade performance, but Gaussian noise has a slight edge for the most affected models at the lowest ϵ. In particular, Random Forest under Laplace noise was almost completely broken at $\epsilon = 0.5$ (F1 ≈ 0.04), whereas Gaussian noise at $\epsilon = 0.5$ gave RF a somewhat better F1 of 0.26. LightGBM also saw Gaussian yield a higher F1 than Laplace at $\epsilon = 0.5$ (0.35 vs 0.25). The intuition here is that for extremely high noise, the heavier-tailed Laplace distribution occasionally produces very large perturbations that can flip a feature value beyond recognition, whereas Gaussian noise (with the same DP guarantee, which in practice allowed slightly less variance due to the δ parameter) might be a bit "softer" in those cases. However, for XGBoost, even at $\epsilon = 0.5$, Laplace noise performed better (F1 0.51 vs 0.41 with Gaussian), indicating that mechanism superiority at low ϵ may also depend on the model's inherent robustness.
- At $\epsilon = 0.1$, performance was so low for both mechanisms that neither could be said to meaningfully outperform the other; all models essentially defaulted to majority-class predictions as noted.

Figure 2 provides a visual snapshot at $\epsilon = \mathbf{1.0}$, comparing Laplace vs. Gaussian F1 scores for each model. We see that Laplace bars are higher for all three models at this privacy level. This suggests that if one is targeting a moderate privacy budget (which still provides a decent balance of privacy and utility), Laplace mechanism might be the preferable choice for injecting noise into the training data of ensemble models.

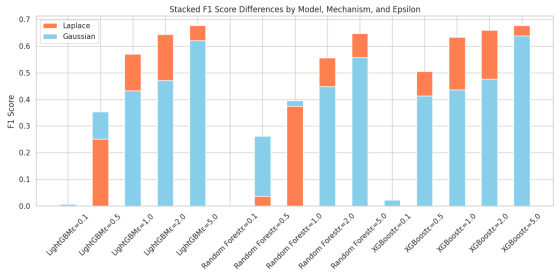


Fig. 2: F1-scores of ensemble models (Random Forest, LightGBM, XGBoost) under Laplace vs. Gaussian noise at a representative privacy level $\epsilon = 1.0$. Laplace noise yields higher F1 than Gaussian for all three models in this setting. Error bars (if visible) indicate variability across any randomness in training; here the variation is minimal since results are deterministic given a fixed random seed.

4.3 Model-wise Performance Differences

Our results reveal that XGBoost is the most resilient model under differential privacy noise, followed by LightGBM, and then Random Forest:

- **XGBoost** not only starts with a high baseline performance (Accuracy 87%, F1 0.71 without noise, matching LightGBM and slightly above Random Forest) but also retains better performance at lower ϵ. For instance, at $\epsilon = 1$ with Laplace noise, XGBoost achieved F1 about 0.63, whereas LightGBM was 0.57 and Random Forest 0.38. Even with Gaussian noise, XGBoost at $\epsilon = 1$ had F1 0.44 vs LightGBM 0.43 and RF 0.40. This indicates XGBoost's boosted tree approach copes with noisy features relatively well.
- **LightGBM** performed similarly to XGBoost under many conditions but was slightly more affected by noise. At very low ϵ, LightGBM's F1 dropped to nearly 0 (especially with Laplace at $\epsilon = 0.1$ it completely failed to identify any positives), while XGBoost still managed a few correct positive predictions (reflected by a non-zero but small F1).
- **Random Forest** was the most vulnerable. With Laplace noise in particular, RF struggled: e.g., as noted, at $\epsilon = 0.5$ Laplace, RF's F1 was essentially zero (the model predicted almost all instances as negative class to maximize accuracy). Random Forest's baseline (no-noise) F1 was also slightly lower

(0.67) than the boosted models (0.71), implying it inherently had a bit less capacity on this problem, which noise only exacerbated.

One reason for XGBoost's robustness might be its combination of robust loss functions and regularization. The boosting process could also be better at correcting mistakes caused by noise in earlier iterations. Random Forest, by contrast, relies on averaging many fully grown trees; if most trees are rendered inaccurate by noise at their splits, the ensemble average still fails. LightGBM, being a boosting method like XGBoost, also handled noise well but perhaps its default settings (like leaf-wise growth) made it slightly more sensitive to heavily noisy features compared to XGBoost's level-wise growth with regularization.

4.4 Precision, Recall, and Specificity Behavior

The inclusion of multiple metrics allows us to diagnose how the models' prediction behavior shifts under DP noise. The heatmap in Figure 3 shows Accuracy, Precision, Recall, F1, and Specificity for each combination of model, mechanism, and ϵ (including the no-noise baseline as "None, $\epsilon = \infty$").

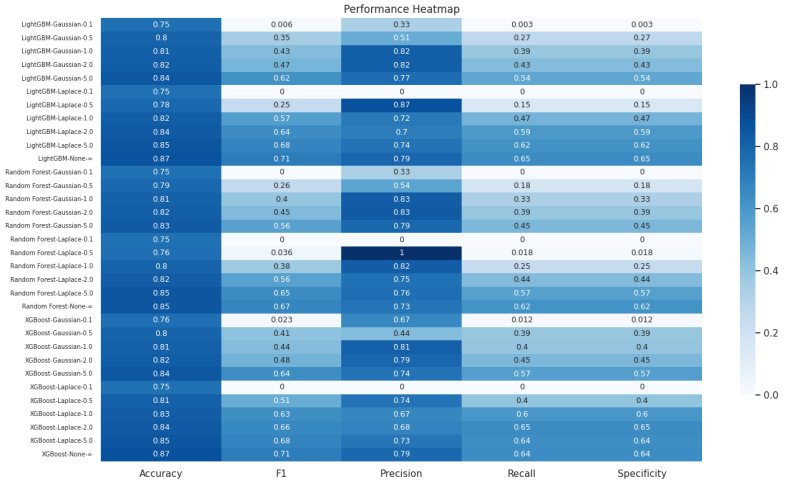


Fig. 3: Heatmap of performance metrics for all models under each DP mechanism and privacy level. Each row corresponds to a model-mechanism-ϵ combination (including a baseline "None, $\epsilon = \infty$" with no DP). Each column is a metric (Accuracy, Precision, Recall, F1, Specificity). Values are color-coded from low (blue) to high (red). This visualization highlights the drastic drop in Recall (and F1) at low ϵ, and comparatively smaller changes in Accuracy and Specificity.

From Figure 3, we observe the following:

- At strict privacy (e.g., $\epsilon = 0.1$ or 0.5), Recall (for the positive class) plunges towards 0, whereas Specificity remains high and even improves in some cases. This confirms that models under heavy noise default to predicting the negative class (since the dataset has more negatives, this yields decent accuracy and very high specificity, but zero or near-zero recall for positives). For example, Random Forest with Laplace at $\epsilon = 0.5$ had Specificity 1.0 but Recall 0.0, meaning it predicted almost no one as high-income. Similarly, LightGBM at $\epsilon = 0.1$ (either mechanism) has Recall 0.
- Precision for the positive class can sometimes be paradoxically high when recall is extremely low. This happens in cases where the model predicts only a handful of instances as positive (perhaps out of noise or random chance) and those happen to align with actual positives. For instance, in the heatmap, some entries show moderately high Precision even when Recall is near zero. However, these are situations with almost trivial models and should not be misconstrued as good performance. The F1 metric accounts for both and remains near zero in those cases. Generally, as ϵ increases and models start predicting positives again, precision might dip slightly when recall rises (a typical precision-recall tradeoff), but F1 increases overall with ϵ.
- Accuracy remains somewhat high even for low ϵ relative to recall. For example, at $\epsilon = 0.1$, accuracy is around 75-80% for all models, which is just the baseline accuracy of always guessing the majority class. This underscores a crucial point: accuracy alone is not a sufficient indicator of model usefulness under DP noise. One could achieve 75% accuracy on this dataset by a trivial non-private model that predicts all instances as "$\leq$50K". DP noise pushes models towards this trivial solution at high privacy, hence accuracy degrades only a little (from 85% to 75%), while recall and F1 plummet. For meaningful utility, we need privacy budgets that allow models to significantly beat that baseline (which we see happens around $\epsilon = 1$ and above in our experiments).
- The heatmap also emphasizes again the relative performance of mechanisms and models. For instance, comparing Laplace vs Gaussian rows for a given model and ϵ, the Laplace row often has more "warm" colors (higher values) in the F1 column. And comparing down the rows, XGBoost rows are generally warmer in F1 and Recall than the corresponding LightGBM and RF rows at the same ϵ.

4.5 Summary of Findings

In summary, our experimental analysis shows:

- There is a sharp privacy-utility trade-off for ensemble models on the Adult dataset. Privacy budgets $\epsilon \leq 0.5$ virtually cripple the models, while $\epsilon \approx 1$ to 2 allows moderate performance, and $\epsilon \geq 5$ recovers most of the model's original accuracy.
- The Laplace mechanism tends to preserve model accuracy and F1 better than the Gaussian mechanism at moderate and high ϵ (within the (ϵ, δ) regime

we tested), though Gaussian can have a minor advantage in the extremely noisy regime for some models.
- Among the ensemble models tested, XGBoost is the most privacy-robust, followed by LightGBM, and then Random Forest which struggles under heavy noise.
- Differential privacy noise heavily skews model predictions toward the majority class, evidenced by high specificity and low recall under strong privacy. This suggests that practitioners should be cautious: a DP-protected model might appear to have reasonable accuracy while essentially ignoring the minority class.

These results provide empirical guidance on choosing ϵ and mechanisms. For example, if one requires $\epsilon = 1$ (often considered a reasonably strong privacy guarantee), using Laplace noise and a model like XGBoost can yield an F1 around 0.6 on Adult, whereas Random Forest with Gaussian noise might only get 0.4. If an even stronger privacy of $\epsilon = 0.5$ is needed, one should expect significant drops in utility (F1 below 0.3 for the best case in our tests).

5 Conclusion and Future Work

In this paper, we examined the impact of differential privacy on ensemble classification models through a detailed case study on the UCI Adult dataset. We compared Laplace and Gaussian noise mechanisms for achieving DP, and evaluated three popular ensemble methods: Random Forest, LightGBM, and XGBoost. Our results show that injecting DP noise into training data can significantly degrade model performance, especially for strict privacy settings, but the extent of degradation varies by mechanism and model type. Laplace noise generally provided better accuracy/F1 than Gaussian noise at comparable privacy levels (for ϵ around 1 and above), and XGBoost emerged as the most resilient model under noise, maintaining higher F1 scores than LightGBM and Random Forest across the board.

We also highlighted how DP primarily impacts recall of the positive class in this imbalanced classification task, as models under heavy noise resort to predicting the majority class. This insight warns that high accuracy under DP can be misleading if not considered alongside recall and precision.

For practitioners, our study suggests a few actionable points: If using tree-based ensembles on sensitive data, consider using boosting algorithms (XGBoost/LightGBM) with DP input perturbation, as they handle noise better than bagged trees in our experiments. If the choice of DP mechanism is flexible, Laplace noise may yield slightly better model utility than Gaussian for moderate privacy budgets, though both perform similarly poorly at very low ϵ. Finally, always evaluate a range of metrics under DP to ensure the model is still accomplishing the intended task (e.g., detecting the positive class) and not just achieving okay-looking accuracy by defaulting to safe predictions.

In future work, there are several avenues to explore. One direction is to investigate *output perturbation* or *objective perturbation* methods for these models,

which might add noise in a more targeted way (for example, adding noise to the trained model parameters or to the tree splitting criterion, rather than to all input features) and potentially achieve better privacy-utility trade-offs. The approach by Hou et al. [10] to distribute privacy budget within random forests is one such example that could be extended to boosting. Another direction is to test on other datasets and tasks (e.g., highly imbalanced or multiclass problems) to see if the patterns observed here generalize. Additionally, hyperparameter tuning under DP is an open challenge: all our models used default settings, but tuning could possibly regain some performance, albeit with care since tuning itself would have to be done in a privacy-aware manner. Finally, exploring theoretical analysis to complement these empirical findings would be valuable—for instance, explaining why Laplace vs. Gaussian noise yields different outcomes for tree ensemble learning, or deriving bounds on the performance degradation as a function of ϵ for these classifiers.

We hope that this work contributes to a better understanding of practical differential privacy for ensemble models and guides both researchers and practitioners in building machine learning systems that responsibly balance privacy with predictive performance.

References

1. Abadi, M., Chu, A., Goodfellow, I., McMahan, H.B., Mironov, I., Talwar, K., Zhang, L.: Deep learning with differential privacy. In: Proceedings of the 2016 ACM SIGSAC conference on computer and communications security. pp. 308–318 (2016)
2. Becker, B., Kohavi, R.: Adult. UCI Machine Learning Repository (1996), DOI: https://doi.org/10.24432/C5XW20
3. Breiman, L.: Random forests. Machine learning **45**(1), 5–32 (2001)
4. Chen, T., Guestrin, C.: Xgboost: A scalable tree boosting system. In: Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining. pp. 785–794 (2016)
5. Dwork, C., Roth, A., et al.: The algorithmic foundations of differential privacy. Foundations and trends® in theoretical computer science **9**(3–4), 211–407 (2014)
6. Fauzi, M.A.: Random forest approach fo sentiment analysis in indonesian. Indones. J. Electr. Eng. Comput. Sci **12**, 46–50 (2018)
7. Fauzi, M.A., Arifin, A.Z., Yuniarti, A.: Arabic book retrieval using class and book index based term weighting. International Journal of Electrical and Computer Engineering **7**(6), 3705 (2017)
8. Fauzi, M.A., Yang, B.: Continuous stress detection of hospital staff using smartwatch sensors and classifier ensemble. In: pHealth 2021, pp. 245–250. IOS Press (2021)
9. Friedman, A., Schuster, A.: Data mining with differential privacy. In: Proceedings of the 16th ACM SIGKDD international conference on Knowledge discovery and data mining. pp. 493–502 (2010)
10. Hou, J., Li, Q., Meng, S., Ni, Z., Chen, Y., Liu, Y.: Dprf: a differential privacy protection random forest. Ieee Access **7**, 130707–130720 (2019)

11. Ke, G., Meng, Q., Finley, T., Wang, T., Chen, W., Ma, W., Ye, Q., Liu, T.Y.: Lightgbm: A highly efficient gradient boosting decision tree. Advances in neural information processing systems **30** (2017)
12. Zheng, Q., Yu, C., Cao, J., Xu, Y., Xing, Q., Jin, Y.: Advanced payment security system: xgboost, lightgbm and smote integrated. In: 2024 IEEE International Conference on Metaverse Computing, Networking, and Applications (MetaCom). pp. 336–342. IEEE (2024)