ML That Minds Its Business (and Yours): Fair, Accurate, and Private

Michelle Buraczyk, Holly Lopez, Leo Mercado, Dr. Anaanta Kotal
NSF RET Site: Cybersecurity Research Experience For Educators Through Data Science (CREEDS)
The University of Texas at El Paso

Abstract

In today's data driven world, machine learning models are increasingly responsible for making decisions that affect people's lives – from loan approval and job recommendations to medical diagnoses and personalized services. With this growing influence comes a critical responsibility; ensuring that these models are not only accurate, but also fair and privacy preserving.

Introduction

Our research explores how machine learning can make accurate predictions while ensuring fairness and privacy. We focused on two datasets with different ethical challenges.

Dataset	Prediction Target	Key Features
1. Diabetes Dataset	Likelihood of developing diabetes	Sex, Age, Diet, Lifestyle
2. Recidivism Dataset	Likelihood of recommitting a crime	Age, Prior Offenses, Race
3. Income Dataset	Whether someone makes over $50K per year	Age, Education, Occupation, Hours per Week, etc.
4. Student Performance Dataset	Academic performance of students	Study Time, Absences, Family Support, Test Scores

These cases allowed us to examine how machine learning models can remain accurate while minimizing **bias** and protecting **sensitive information**, especially in critical areas like **healthcare** and **criminal justice**.

Methodology

We began by preprocessing the datasets—removing missing values, applying one-hot encoding, and scaling numerical features—followed by a train-test split. We trained and evaluated models including Logistic Regression, Random Forest, KNN, Linear Regression, and Naive Bayes to identify the most accurate baseline.

1. Accuracy First

We train models to minimize prediction error:

$$\mathcal{L}(f_\theta) = \frac{1}{n}\sum_{i=1}^{n} \ell(f_\theta(x_i), y_i)$$

This gives us strong baseline performance.

Next, we integrated differential privacy using IBM's *diffprivlib*, varying **Epsilon (ϵ)** to explore the trade-off between accuracy and privacy. Smaller ε values ensured stronger privacy at the cost of performance.

2. Then Privacy

We use **DP-SGD** to protect individual data. Gradients are clipped and noised: $\tilde{g} = \frac{1}{n}\sum_{i=1}^{n} \text{clip}(g_i) + \mathcal{N}(0, \sigma^2 I)$

This ensures **(ε,δ)-differential privacy**.

Finally, we applied fairness interventions with the *fairlearn* library, testing bias mitigation techniques to produce a model that balanced accuracy, privacy, and fairness.

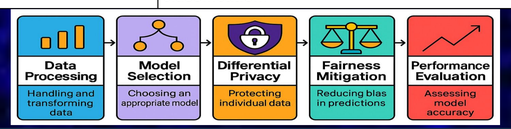
3. Fairness, Finally

We add fairness constraints, like demographic parity, during training using: $\min_\theta \mathcal{L}(f_\theta) + \lambda \cdot \text{FairnessPenalty}(f_\theta)$

Balancing performance, privacy, and fairness—together.

Overview of Common Machine Learning Models

Model	Purpose	Detailed Description
Logistic Regression	Predicts categories	Estimates the probability that a data point belongs to a specific class using the **logistic (sigmoid) function.** Often used in binary and multiclass classification tasks.
Random Forest	Improves prediction reliability	An **ensemble of decision trees** that reduces overfitting by averaging multiple trees' outputs. Increases model stability and accuracy.
K – Nearest Neighbors (KNN)	Compares similarity	Classifies based on the **majority class among K closest neighbors** using a distance metric. No training phase makes it simple and intuitive.
Linear Regression	Predicts continuous values	Fits a **straight line** through the data to model the relationship between input features and a numerical target, ideal for trend prediction.
Naive Bayes	Predicts categories using probabilities	A **probabilistic classifier** based on Bayes' Theorem. Assumes feature independence, making it fast and effective for **text and spam classification**.



RESULTS: UCI Diabetes

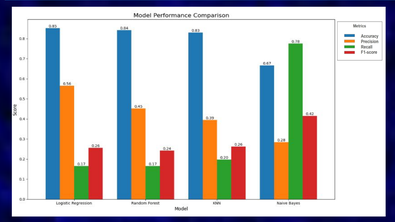


Figure 1.1 – ***Performance comparison of various machine learning models based on accuracy, precision, recall, and F1-score. Logistic Regression outperformed the other models across most metrics and was selected as the final model for the project.***

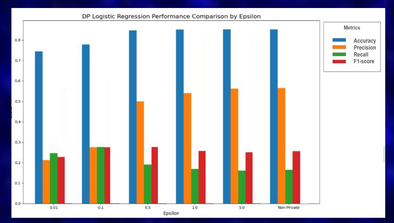


Figure 1.2 – ***Impact of differential privacy on model performance across five epsilon values. Varying levels of noise were introduced to assess the trade-off between privacy and accuracy. An epsilon value of 0.5 was ultimately selected, achieving an accuracy of 0.8470 while ensuring a reasonable level of data privacy.***

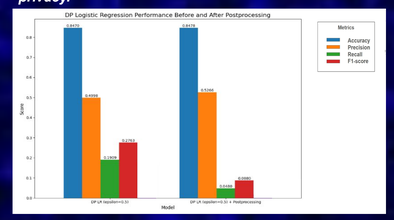


Figure 1.3 – Fairness was introduced to the differentially private model using a post-processing mitigation technique. This approach balanced fairness and privacy while maintaining strong performance, improving the model's accuracy from 0.8470 (private only) to 0.8478 (private and fair).

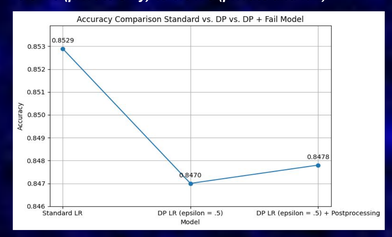


Figure 1.4 – Comparison of accuracy across three models: the Standard Logistic Regression model, the Differentially Private model, and the Private and Fair model. This highlights the trade-offs and performance impact of adding privacy and fairness constraints.

RESULTS: . COMPAS Recidivism Racial Bias

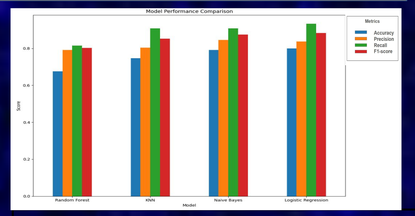


Figure 2.1 – Performance comparison of various machine learning models based on accuracy, precision, recall, and F1-score. Logistic Regression outperformed the other models across most metrics and was selected as the final model for the project.

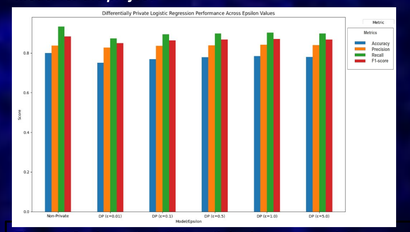


Figure 2.2 – Impact of differential privacy on model performance across five epsilon values. As epsilon increases (from 0.01 to 5.0), the performance metrics generally tend to increase, moving closer to the performance of the non-private model. A smaller epsilon (e.g., 0.01) provides stronger privacy but results in lower performance metrics. A larger epsilon (e.g., 5.0) provides weaker privacy but results in higher performance metrics.

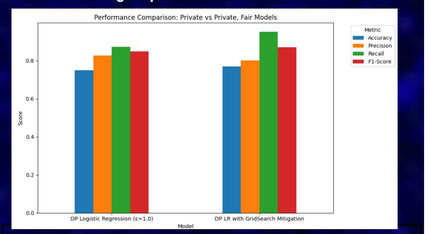


Figure 2.3 - The bar chart comparing the accuracy of the Standard Logistic Regression model, the Differentially Private Logistic Regression model (epsilon=1.0), and a Standard Logistic Regression model with post-processing.

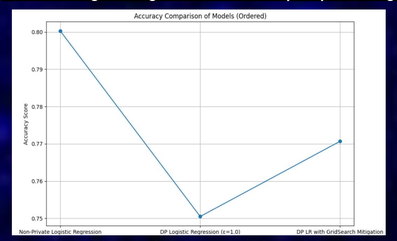


Figure 2.4 – The bar chart is a comparison of the model accuracy for the standard, differentially private and fair data sets using a Differential Privacy Logistic Regression epsilon value of 5 and the Random Forest Classifier Machine Learning Model.

Conclusion

Achieving a machine learning model that is fully accurate, private, and fair is inherently challenging due to the trade-offs among these goals. Our objective was to fine-tune the privacy parameter (epsilon) and apply fairness mitigation techniques to develop a model that balances all three as effectively as possible.

UCI Diabetes Dataset	COMPAS Recidivism Dataset
Private & fair model accuracy: **0.8478**	Private & fair model accuracy: **0.9428**
Standard model accuracy: **0.8529**	Standard model accuracy: **0.9999**
Marginal difference: **0.0051**	Known for exposing racial bias in predictive algorithms

These results demonstrate that while some accuracy may be sacrificed, it is possible to design models that are meaningfully private and fair without significantly compromising predictive performance.

Potential Lesson Plan Elements

Learning Objectives

By the end of this lesson or unit, students will be able to:

- Open and run a notebook in **Google Colab**.
- Use **Pandas** to:
  - Load and inspect datasets.
  - Clean and filter data to improve model accuracy
  - Perform basic statistical analysis (mean, median, mode, etc.).
    - Use **Matplotlib** to:
  - Create visualizations such as histograms, scatter plots, and box plots.
  - Interpret distributions, outliers, and trends in context to include differentially private models and differentially private/fair models

AP Statistics Tie-ins

- **Center & Spread** – Use `.mean()`, `.median()`, `.std()` in Pandas
- **Data Distribution** – Visualize using histograms, box plots
- **Outliers** – Identify from boxplots or using IQR method
- **Two-variable analysis** – Create scatter plots to show relationships (e.g., Glucose vs. Age)

Assessment Ideas

- **Mini Project**: Students choose a column in the dataset, calculate stats, and create visualizations.
- **Exit Ticket**: "What does the histogram of 'Age' tell you about the distribution? Is it skewed?"
- **Notebook Check**: Review importance of creating private and fair machine learning models.

References

1. Dua, D., & Graff, C. (2019). *UCI Adult Dataset*. UCI Machine Learning Repository. Retrieved from https://archive.ics.uci.edu/dataset/2/adult
2. Dua, D., & Graff, C. (2019). *CDC Diabetes Health Indicators Dataset*. UCI Machine Learning Repository. Retrieved from https://archive.ics.uci.edu/dataset/891/cdc+diabetes+health+indicators
3. Cortez, P., & Silva, A. M. G. (2008). *Student Performance Dataset*. UCI Machine Learning Repository. Retrieved from https://archive.ics.uci.edu/dataset/320/student+performance
4. Ofarrell, D. (2016). *COMPAS Recidivism Risk Score Data and Analysis*. Kaggle. Retrieved from https://www.kaggle.com/datasets/danofer/compass

Acknowledgements

This work is supported by National Science Foundation Award # 2206982
Thank you Dr. Kotal, Dr. Piplai, Dr. Ceberio, Dr. Tosh for all your hard work.