ML Algorithms with Anonymized Data

İrem Yurdakurban 72314
Kabil Mert Küçükerdem 71972
Barkın Kılıçkını 71918

INTRODUCTION

In this study, our primary goal is to explore how the choice of k value in anonymization affects the accuracy of machine learning algorithms when applied to a dataset. We have selected the well-known adults dataset, which contains sensitive information, and our focus is on safeguarding individual privacy through anonymization using three distinct k values. The anonymization process is pivotal to ensure the concealment of personal details while preserving the dataset's utility.

Our approach involves employing three machine learning algorithms—k-nearest neighbors, gradient boosting, and random forest—on both the original and anonymized datasets. By assessing the impact of anonymization on model accuracy across different versions of the dataset, we aim to understand the trade-off between data privacy and machine learning performance. This analysis provides valuable insights into balancing the need for accurate predictions with the imperative of protecting individual privacy. The anticipated outcomes include a comprehensive understanding of how varying levels of anonymization influence model accuracy, offering essential considerations for decision-making in situations where both data privacy and predictive accuracy are critical.

In the context of data-driven decision-making, the delicate balance between data privacy and the effectiveness of machine learning models is a significant concern. Our project specifically addresses the optimization of k values in anonymization to uphold individual privacy while sustaining the accuracy of machine learning algorithms. As organizations increasingly utilize predictive models on datasets containing sensitive information, the trade-off between data privacy and machine learning accuracy becomes more crucial. Through the examination of different k values on key algorithms—k-nearest neighbors, gradient boosting, and random forest—our study aims to provide nuanced insights into the trade-offs between anonymization and model accuracy.

TECHNICAL APPROACH

1. **Dataset:**
    - We used the Adult dataset we got from: https://archive.ics.uci.edu/dataset/2/adult
    - We identified ‘salary’ as sensitive attributes. And according to that, we will k-anonymize our dataset using 3 anonymization techniques.
    - We used the Employee database we got from: https://www.kaggle.com/datasets/tawfikelmetwally/employee-dataset
    - We identified ‘LeaveOrNot’ as sensitive attributes. And according to that, we will k-anonymize our dataset using 3 anonymization techniques.

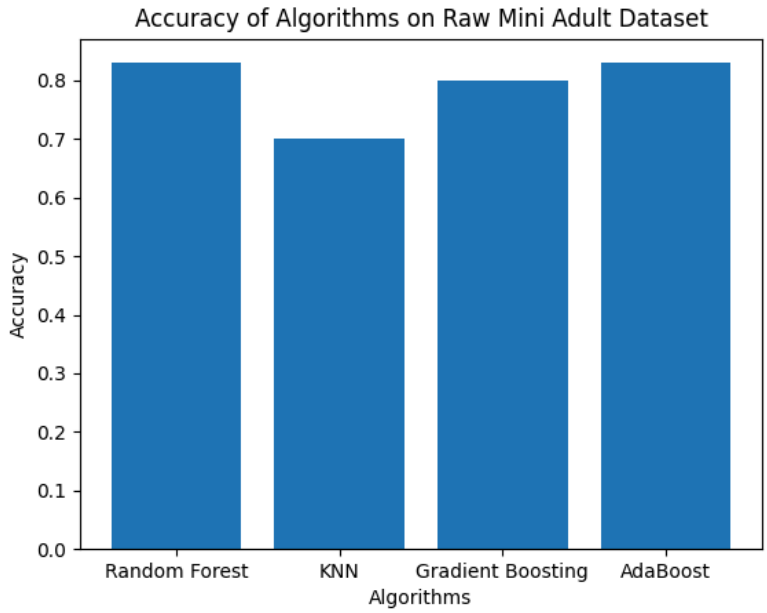
2. **Algorithms:**
    - We will implement four machine learning algorithms, K-nearest Neighbors, Gradient Boosting, Random Forest and Adaptive Boosting. We will apply all of them for both raw and anonymized datasets.

3. **Implementation:**
    - All of our processes will be handled in Python and Jupyter Notebook will be used as the platform. In the data pre-processing step, we will use ‘pandas’ and ‘numpy’ libraries. For the machine algorithms, we will specifically get use of the ‘scikit-learn’ library since it is one of the most used and easy to implement libraries in machine learning projects.

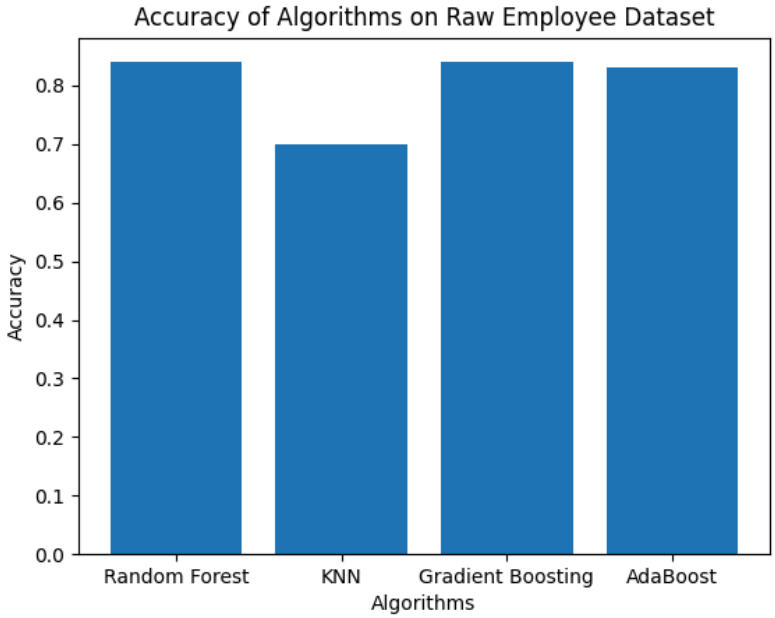
FIGURES

Raw Datasets

Mini Adult



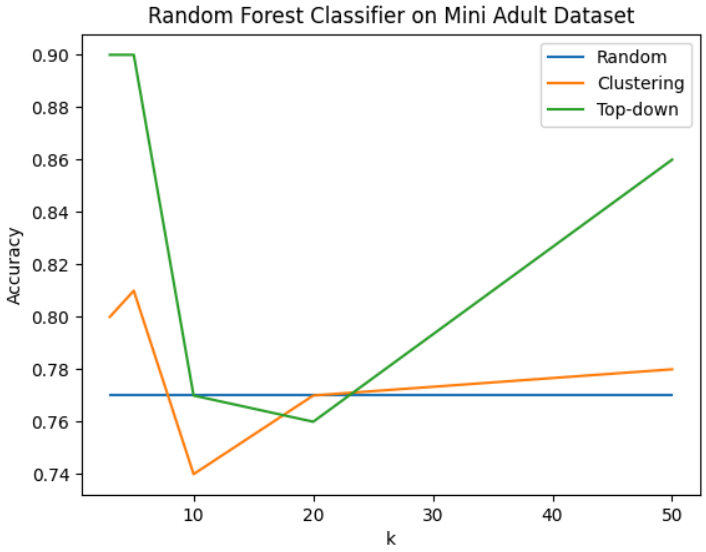
Employee



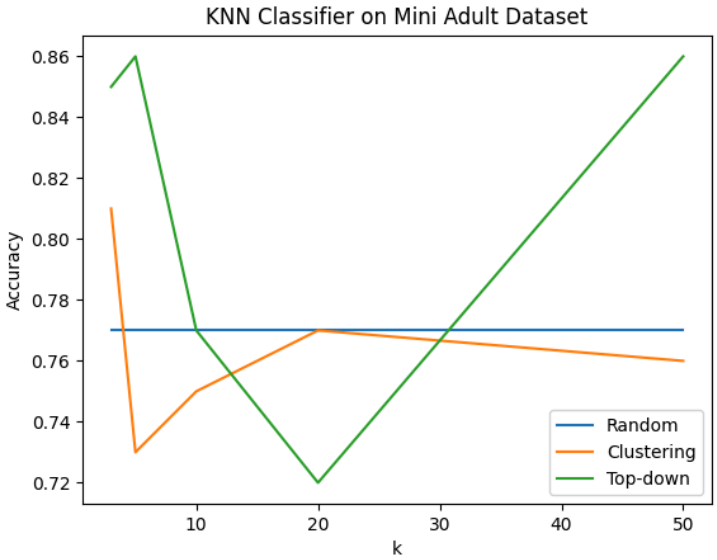
The following figures display the accuracies achieved by various machine learning algorithms on three distinct anonymized versions of the Adult dataset, utilizing k values ranging from 0 to 50.

Mini Adult Dataset:

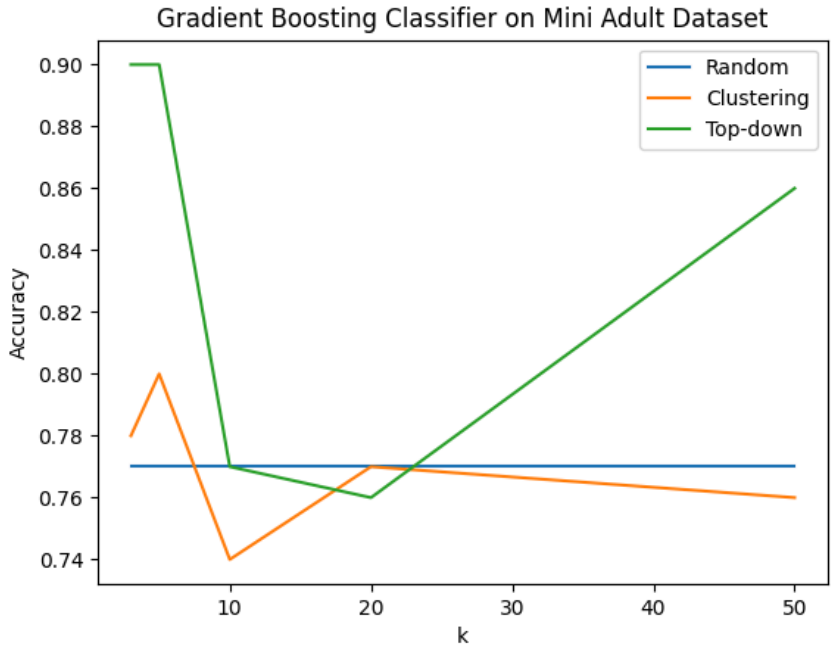
- **Random Forest Algorithm**



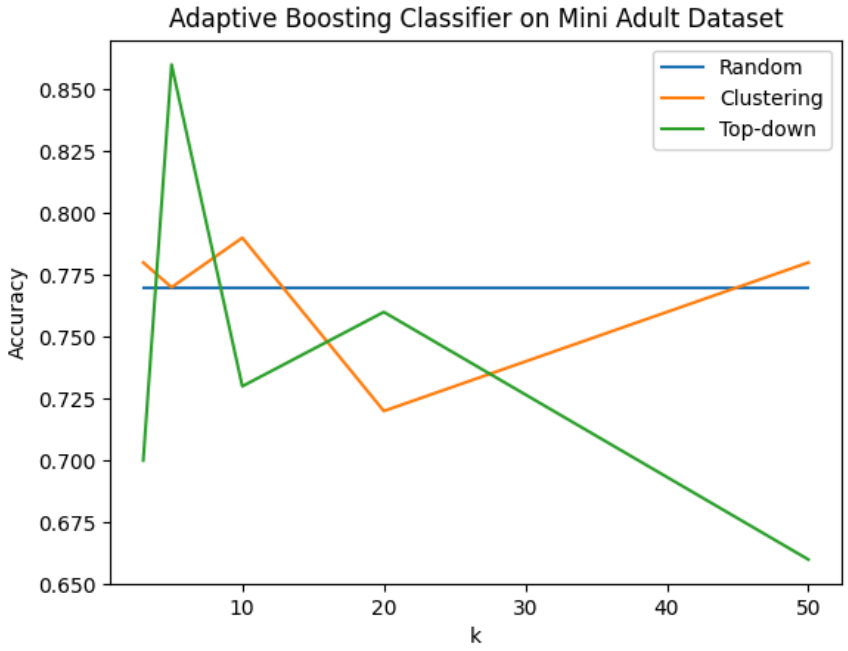
- **K-nearest Neighbors**



- **Gradient Boosting**



- **Adaptive Boosting**

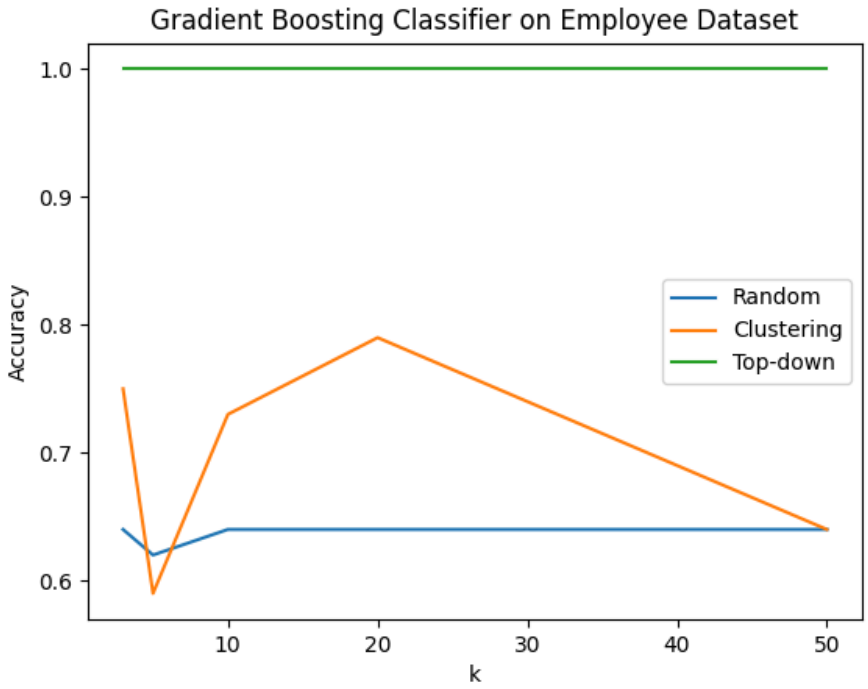


Employee Dataset:

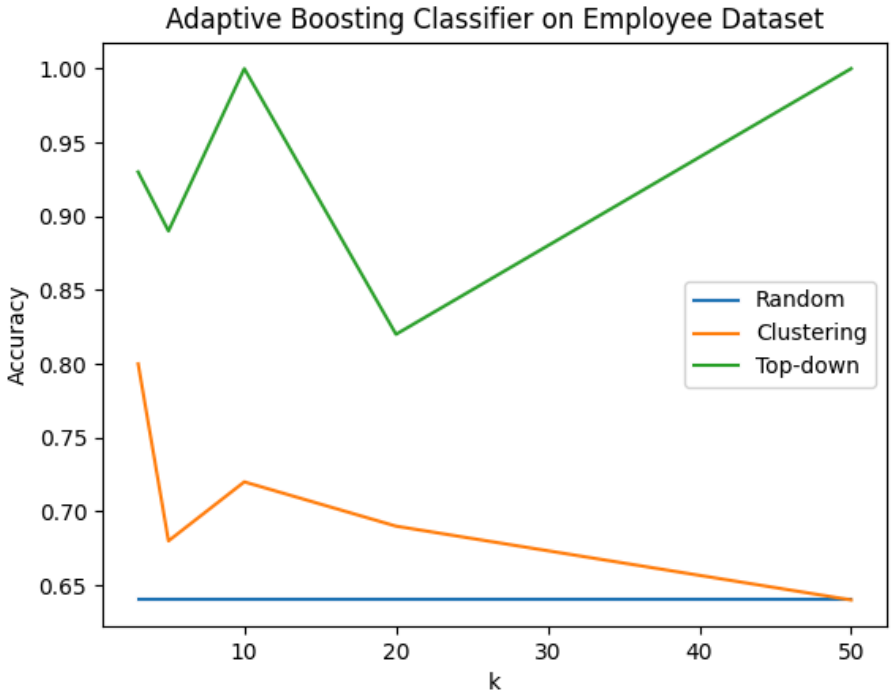
- **Random Forest Algorithm**

- **K-nearest Neighbors**

- **Gradient Boosting**



- **Adaptive Boosting**

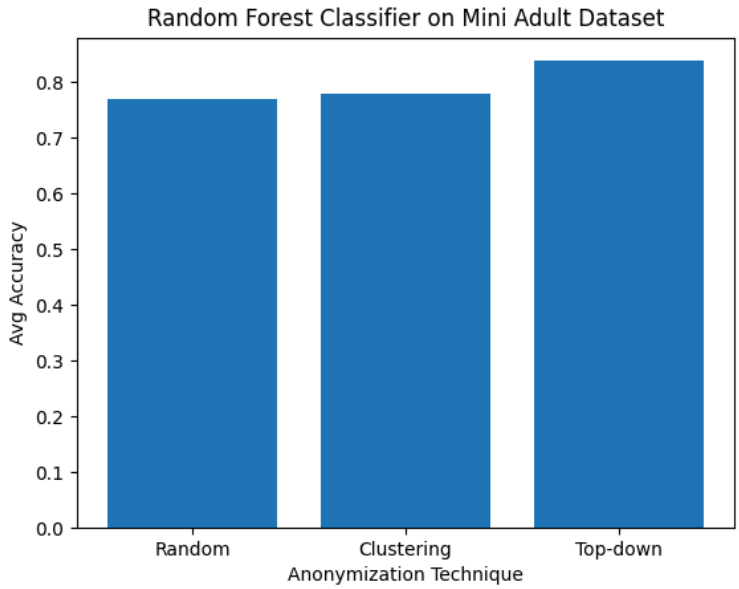


The figures below illustrate the average accuracies achieved by individual machine learning algorithms, contingent on the specific anonymization technique applied to the dataset.

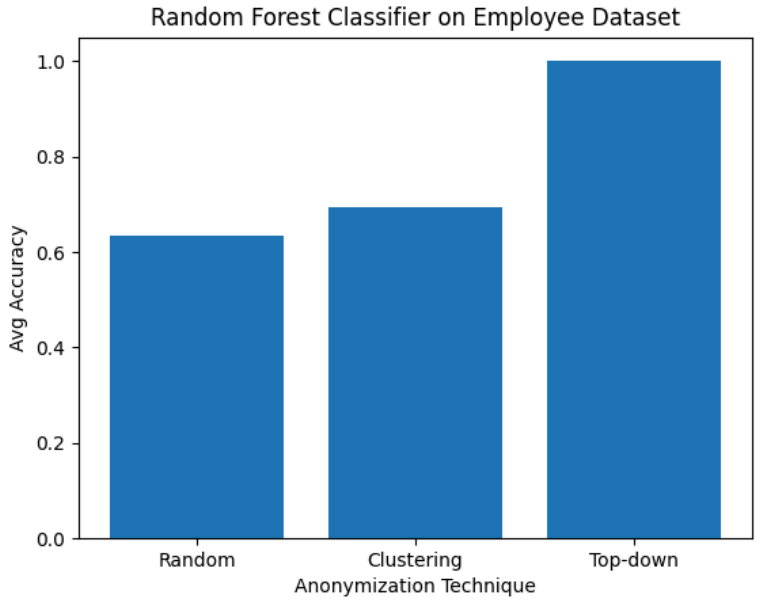
Anonymization techniques

- **Random Forest Algorithm**

Mini adult dataset:

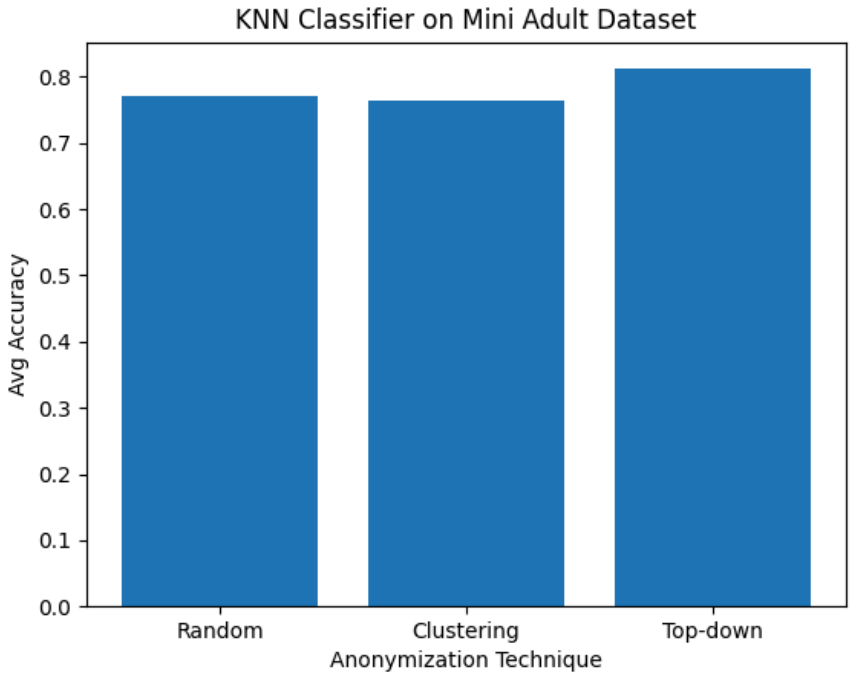


Employee dataset:

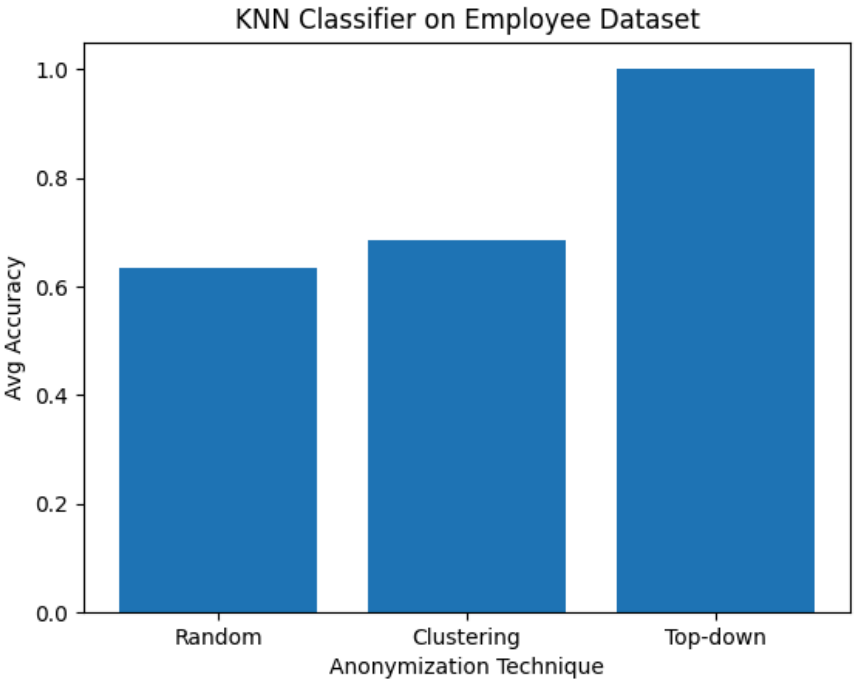


- **K-nearest Neighbors**

Mini adult dataset:

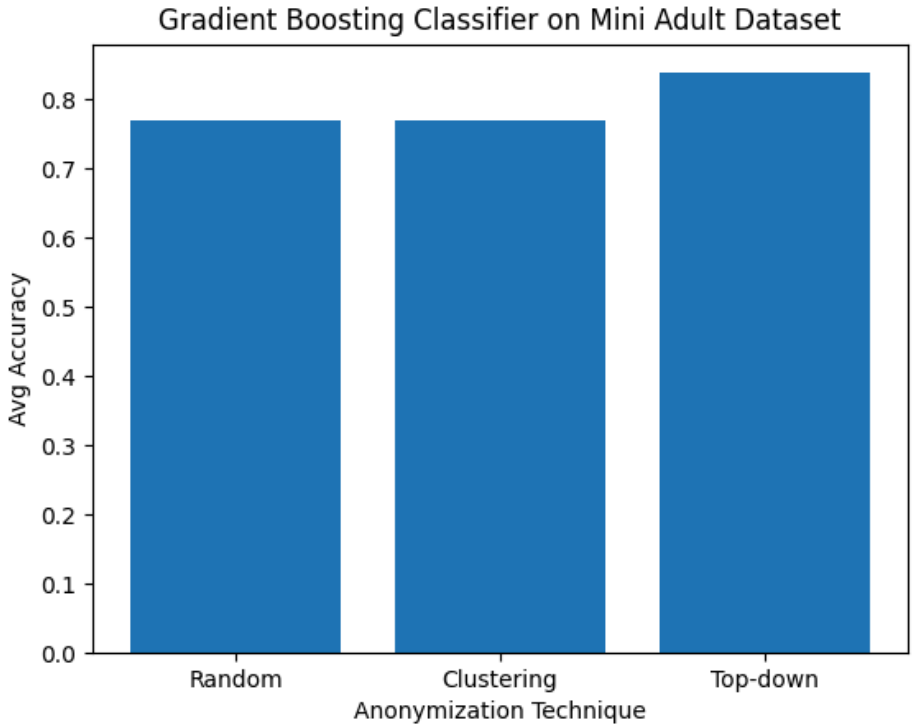


Employee dataset:

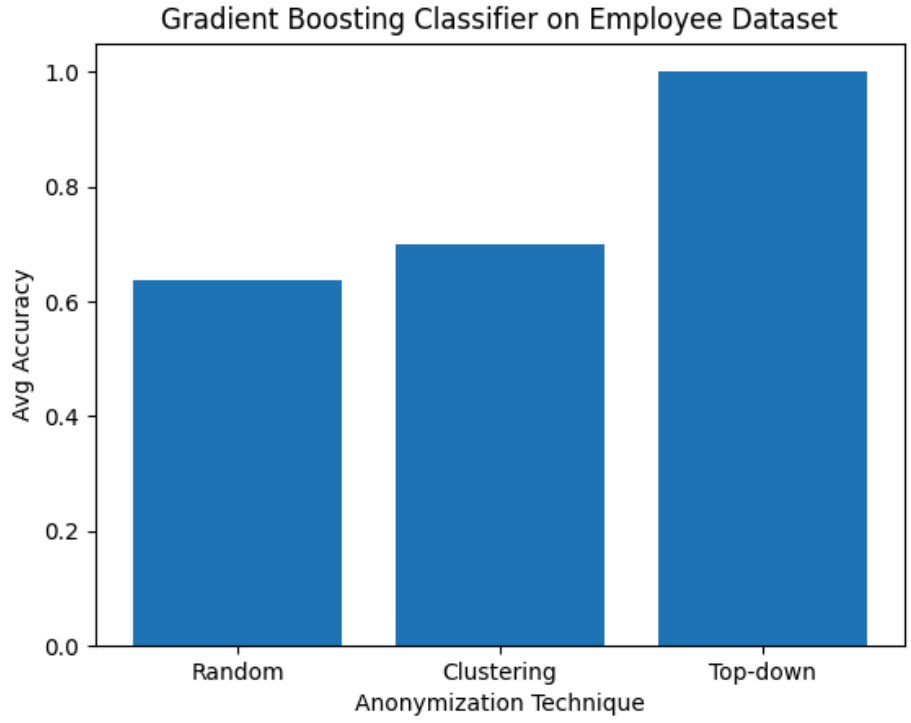


- **Gradient Boosting**

Mini adult dataset:

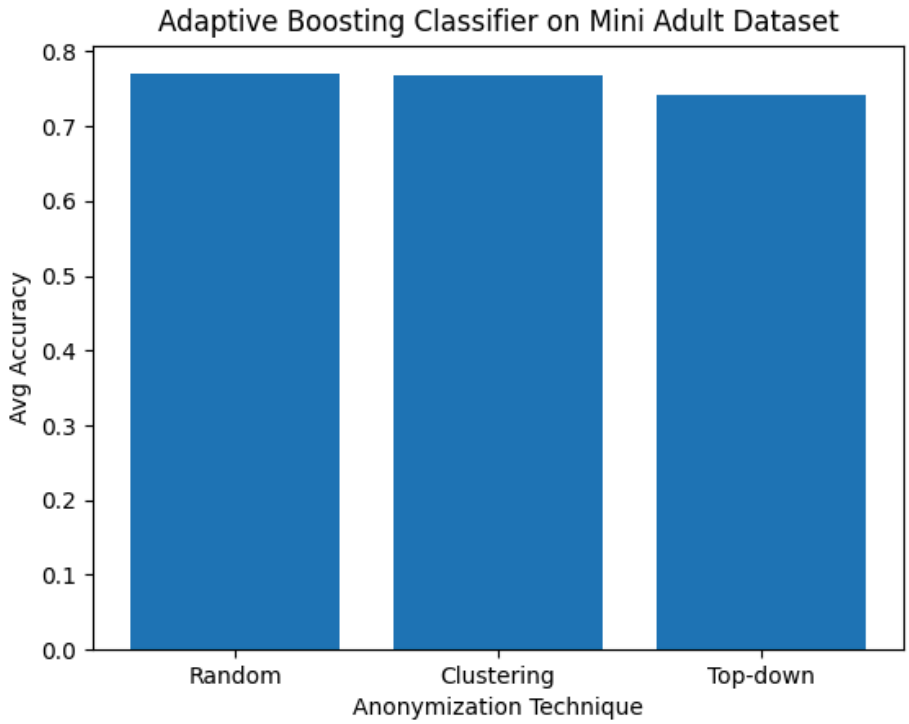


Employee dataset:

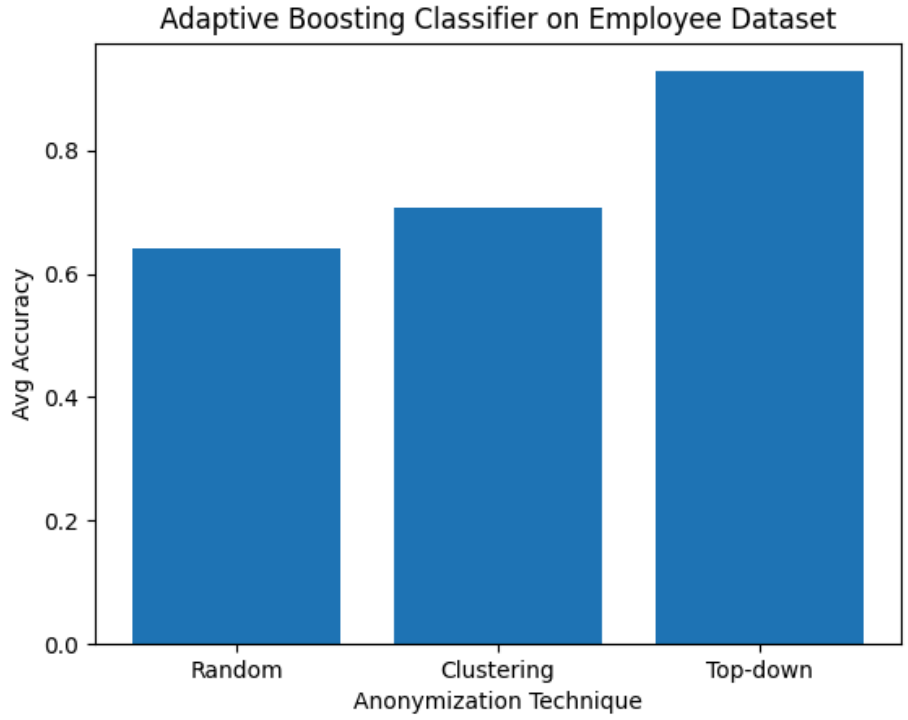


- **Adaptive Boosting**

Mini adult dataset:



Employee dataset:

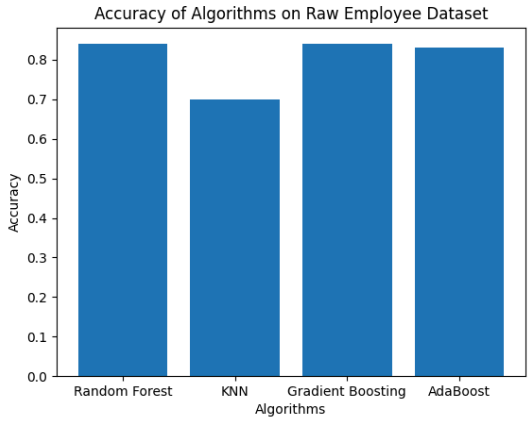


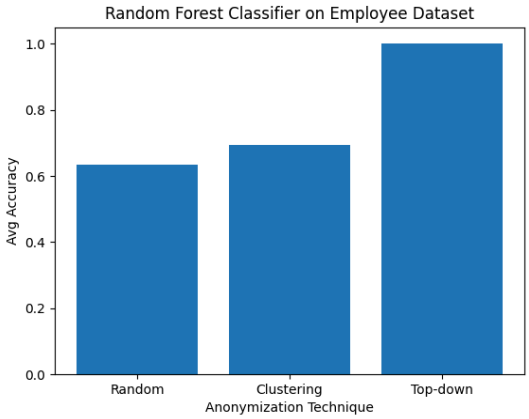
RESULTS

Raw vs Anonymized Dataset

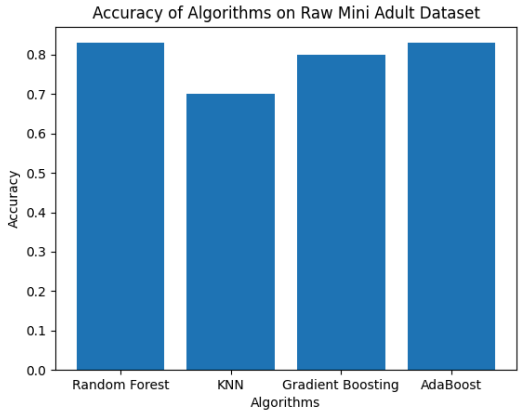
Prior to initiating the project, we held certain expectations regarding the outcomes. Initially, we anticipated that machine learning algorithms would demonstrate higher accuracy on raw datasets, given their precision with more detailed information. However, our observations revealed an unexpected result as we anticipated a direct correlation between anonymization and accuracy. Contrary to our expectations, we found no clear-cut relationship between accuracy and anonymization. Surprisingly, there were instances where the anonymized dataset exhibited greater accuracy.

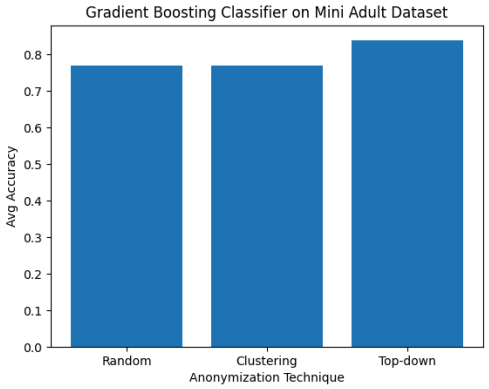
For instance, when examining the Random Forest algorithm on the Employee dataset, we noted an increase in accuracy with top-down anonymization. This unexpected finding underscores the nuanced relationship between anonymization techniques and algorithm performance.





In the mini adult dataset, we observed a notable improvement in accuracy for the Gradient Boosting Algorithm, particularly with top-down anonymization. This additional observation further emphasizes the nuanced impact of anonymization techniques on the performance of specific machine learning algorithms.





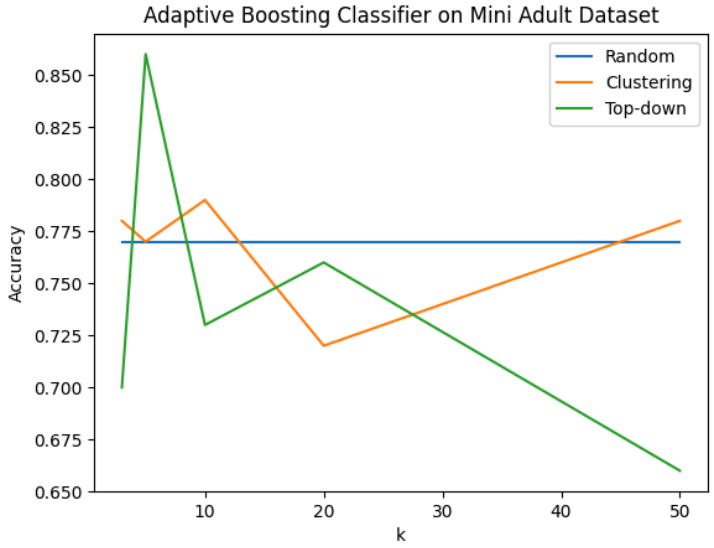
Impact of k-value

We initially anticipated a direct correlation between the k value and the accuracy of machine learning algorithms. With the expectation that the k value would linearly influence accuracy, we assumed a straightforward increase or decrease pattern. However, our findings contradicted this assumption.

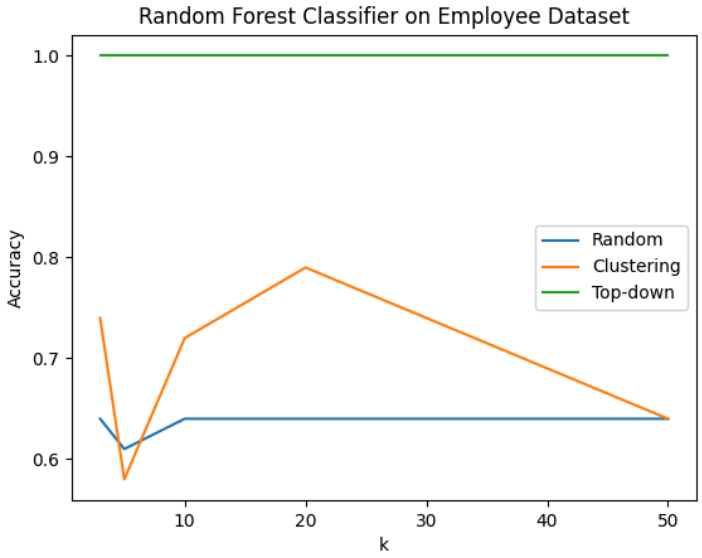
In certain instances, we observed that the k value had no discernible effect on accuracy, such as with the K-nearest Neighbors algorithm on the anonymized versions of Employee dataset.

Additionally, there were cases where accuracy exhibited a non-linear response to changes in the k value. For example, the Adaptive Boosting algorithm on the Clustering Anonymized Mini

Adult dataset showcased an initial decrease in accuracy with an increase in the k value, followed by a subsequent increase.



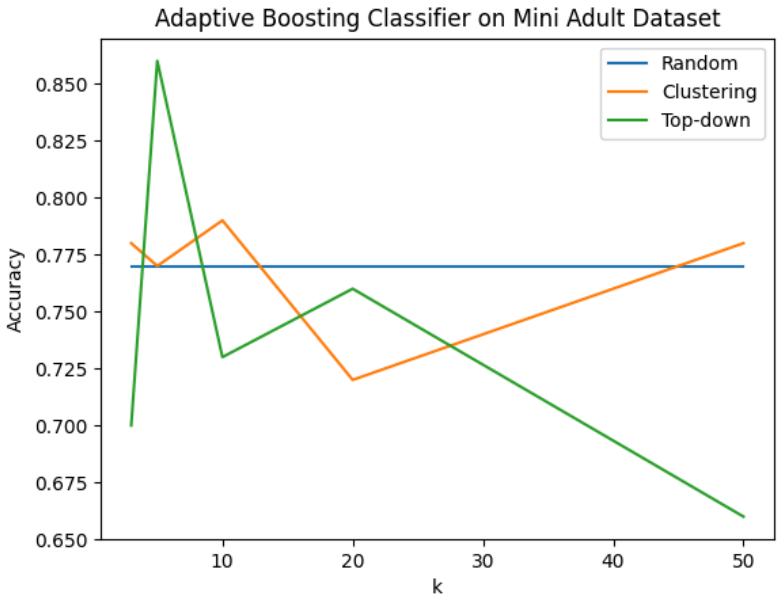
Furthermore, we identified scenarios in which the k value initially impacted accuracy, but beyond a certain point, the accuracy plateaued. One such case was observed with the Random Forest algorithm on the Random Anonymized Employee dataset. These nuanced patterns underscore the intricate relationship between the choice of k value and the performance of machine learning algorithms under different anonymization scenarios.



Upon comprehensive analysis of all the results, it becomes apparent that alterations in the k-value do not exhibit a correlated change in accuracy, regardless of the combination of machine learning algorithm, dataset, or anonymization method.

Anonymization Methods

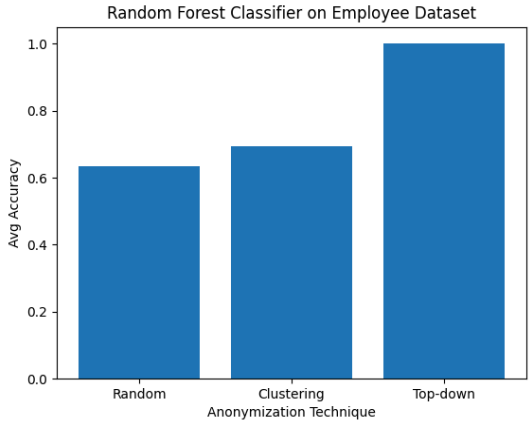
Before commencing the project, our assumption was that the influence of anonymization methods on machine learning algorithms would be minimal. Contrary to this expectation, we observed significant variations in accuracy based on the anonymization method, particularly when the machine learning algorithm, dataset, and k value remained constant. For instance, the impact of anonymization methods on the Adaptive Boosting algorithm in the Mini Adult dataset was distinct: when we look at the beginning, top-down methods led to an increase in accuracy, clustering resulted in a decrease, and random methods maintained a consistent accuracy level. This unexpected variability underscores the substantial role that anonymization methods play in shaping the performance of machine learning algorithms under specific conditions.

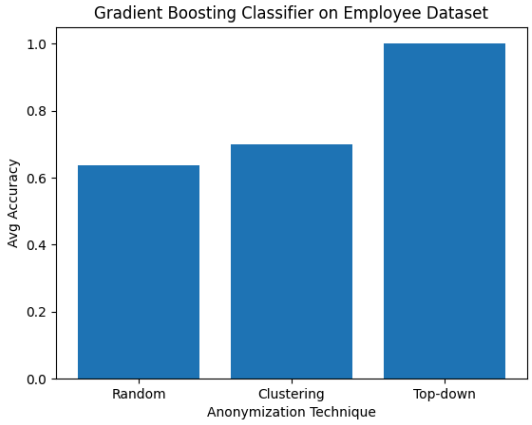


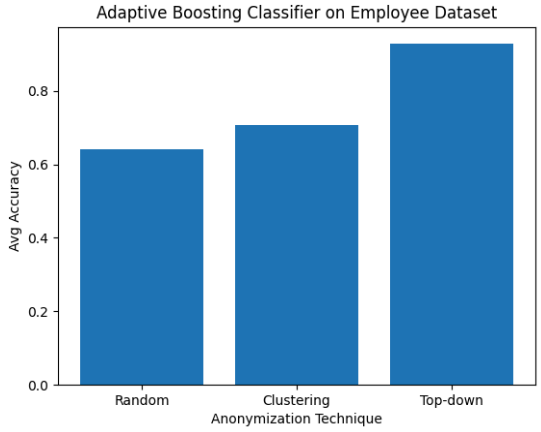
Upon analyzing all our results comprehensively, it becomes evident that the top-down anonymization method consistently outperforms other variations. The highest accuracies across various scenarios were consistently achieved using the top-down model. This observation underscores the effectiveness and reliability of the top-down anonymization method in maintaining or even enhancing accuracy in machine learning algorithms within the scope of our study.

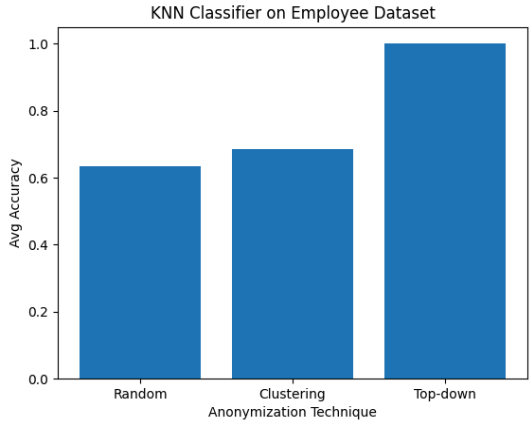
Machine Learning Algorithms

Upon comparing the average accuracies of different machine learning algorithms, a notable trend emerges: the variations in accuracy are more significantly influenced by the choice of anonymization models rather than the specific machine learning algorithm employed. This is evident in the similar average accuracies observed among different machine learning algorithms when applied to the same anonymized dataset. However, it's worth noting that, on average, the K-nearest Neighbor algorithm consistently exhibits lower accuracy compared to the other machine learning models in our study. This observation underscores the impact of anonymization methods in shaping the overall performance of machine learning algorithms.









CONCLUSION

In our exploration of the interplay between data privacy, security, and machine learning accuracy, several key insights have emerged. Contrary to our initial expectations, the impact of anonymization on accuracy did not follow a straightforward correlation. The nuances of this relationship were particularly evident in unexpected scenarios where anonymized datasets exhibited greater accuracy, challenging the assumption that raw datasets inherently provide superior results.

The influence of the k-value on accuracy further highlighted the complexity of the data privacy and machine learning balance. Contradicting our initial assumptions, the k-value's impact on accuracy did not follow a linear pattern. Instead, we identified instances where the k-value showed no significant effect, cases of non-linear response, and scenarios where initial impacts plateaued. These findings underscore the importance of a nuanced approach to k-value selection in the context of maintaining both data privacy and accurate machine learning models.

Anonymization methods played a pivotal role, revealing that top-down anonymization consistently outperformed other variations across various scenarios. This emphasizes the crucial role that anonymization methods play in shaping the performance of machine learning algorithms. Additionally, when comparing machine learning algorithms, we observed that variations in accuracy were more prominently influenced by the choice of anonymization methods than by the specific algorithm employed. Notably, the K-nearest Neighbor algorithm consistently exhibited lower average accuracy compared to other models, highlighting the relevance of anonymization in shaping the overall performance of machine learning algorithms. In conclusion, our study emphasizes the intricate and dynamic nature of the relationship between data privacy, anonymization methods, and machine learning accuracy, offering valuable insights for navigating this delicate balance in real-world applications.