Final Project Writeup

Problem Statement:

Given a dataset of NHL players, we want to use machine learning to predict if their career number of games played was above or below the mean amount of games played (137), while keeping a privacy guarantee to our data.

Our Solution:

We began by finding a dataset of 12,250 players in the NHL that contained all of the player statistics. The first step was to replace all NaN from the dataset with 0s to increase readability. Then all goalies were removed from the dataset, as they had their own column for the number of games played in the original data set, which was often difficult to navigate around. Following the removal of those goalies, we had to combine some of the positions for each player because there were multiple ways that certain positions were labeled. To allow our data to be usable, we started by label encoding all columns that contained string values using the built-in LabelEncoder function. This allowed the dataset to be compatible with our machine learning model. Next, we created X, which would be our entire dataset, but without the 'games_played' column because that will be our target column also seen as y. Now that all of our data is cleaned we split it into training data, 80% of the data, and testing data, the remaining 20%.

To train our model, we utilized a linear regression approach with our training data sets. We implemented the predict function used in class to return a 'model accuracy' of this approach. To add some noise in order to protect our data, we used both a Renyi gradient descent, and a zero-concentrated gradient descent. We then compared the accuracy of these methods and then compared them while considering their privacy costs.

The Results:

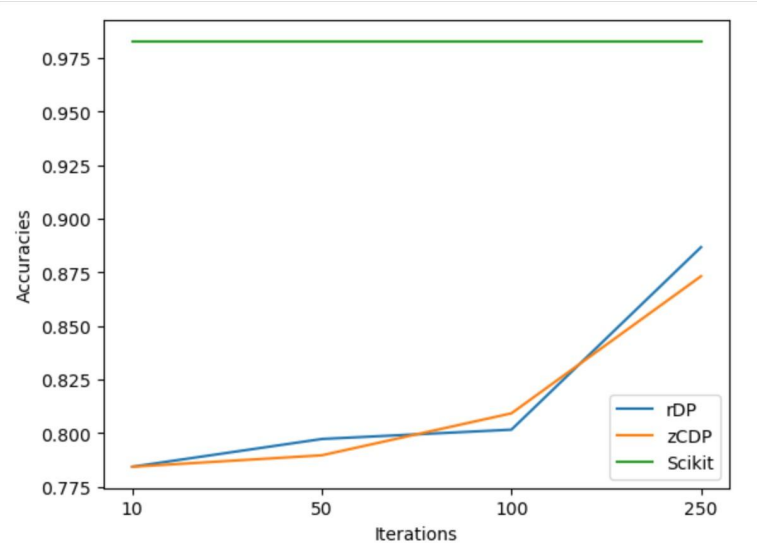

Figure 1: Comparing the accuracies of rDP and zCDP

We decided to graph the accuracy returned by our rDP and zCDP implementations over 4 different iteration counts of 10, 50, 100, 250. Each iteration's count is done 5 times, and the accuracy graphed is the average of those counts to better visualize how the two stacked up against one another. This allowed us to see that these methods are very similar, as they go back and forth for which has the higher accuracy, but when the number of runs increases, the average zCDP accuracy seems to slightly beat the rDP accuracies.