Algorithmic and Theoretical Aspects of Differential Privacy – Assignment

In this assignment, you will have a chance to try applying a machine learning algorithm on the data that is protected under the differential privacy notion. There are five tasks in this assignment. Your grade will be i if you submit correct solutions for i tasks.

Before you start:
1) Download the dataset at https://www.kaggle.com/majidarif17/weight-and-heightcsv, unzip it, and upload the csv file to your Google Drive.
2) Download the Python notebook "ComparingDifferentialPrivacyAlgorithms.ipynb" in ITC-LMS. Then, upload the file to your Google Colab (https://colab.research.google.com).
3) Run the first code cell of the notebook to mount your Google Drive to the Python notebook.
4) Run the second code cell of the notebook to check if you can correctly have the data in your notebook.

About the dataset:
Each tuple represents gender, height, and weight of our users. **Let us suppose in this assignment that the table size is public information, gender and height are quasi-identifiers, and weight is sensitive information**. We want to do linear regression to find a relationship between height and weight in male and female.

Task 1: Linear regression of two variables
1) Write a function to calculate linear regression result using the formulation given in the following website:
   https://www.statisticshowto.com/probability-and-statistics/regression-analysis/find-a-linear-regression-equation/
   The input of the function must be $\sum x_i$, $\sum y_i$, $\sum x_i^2$, $\sum x_i y_i$, and n.
2) Use your function in 1) to calculate the relationship between height and weight in male and female. Please assume that x_i is height of i and y_i is weight of i.

Task 2: Laplacian mechanism and Composition Theorem
We want to be sure that the publication of the relationship between height and weight is 0.1-differentially private. We will use the Laplacian mechanism to achieve that. We will add the Laplacian noise to the input of the functions ($\sum x_i$, $\sum y_i$, $\sum x_i^2$, $\sum x_i y_i$, and n).
1) Discuss why we need *not* to add noise to $\sum x_i$, $\sum x_i^2$, n in this setting.
2) Let $f(T) = \sum y_i = \sum weight_i$. Calculate $GS(f)$. Please give an assumption on weight and height ranges by yourself.
3) Let $g(T) = \sum x_i y_i = \sum height_i weight_i$. Calculate $GS(g)$. Please give an assumption on weight and height ranges by yourself.
4) We will add the Laplacian noise to $\sum y_i$ and $\sum x_i y_i$. What should be the noise parameters? Discuss why by the noise parameters we will have 0.1-differential privacy when we publish the linear regression result.

Task 3:
Observe the differences before and after adding noise. Do you think that the noise makes the linear regression results significantly worse?

Task 4: SmallDB Algorithm

Consider the task by Students 1-3. In this task, we will aim provide a smaller database which can give precise inputs for the function that calculates the linear regression results ($\sum y_i$ and $\sum x_i y_i$). From next question, let us assume that $\alpha = 0.1$ and $\epsilon = 0.1$.

1) What is the size of the smaller database we should have given to the data scientist?
2) Let assume that we will *select* records of the smaller database from the original database. What is the number of possible publish database in the exponential mechanism.
3) Recall the following inequality:
$$Pr\,Pr\left[E \leq OPT - \frac{2\Delta Utility}{\epsilon}(ln\,ln\,(\#possible\ outputs) + t\,)\right] \leq e^{-t}.$$
What the inequality say when $t = 100$?
4) What can we say from your answer in 3)? Can we say that SmallDB algorithm will give us a good result here? Give your reason why or why not? What could be reasons behind good/bad results of the SmallDB algorithm?

Task 5:

Compare the results that we have in Task 3 with provisional results from the SmallDB algorithm. Which of the implementation should give a better result? Discuss the reasons.