**Jigyas Sharma**

**Dr. Han Wang**

**EECS 765**

**14 September 2023**

**Homework 1: Report**

**Environment:** Google Colab, Github

**Libraries:** Pandas

**Language:** Python

**Input Dataset:** adult.data

**Achieving K-Anonymity using Generalization and Suppression**

*K-Value = 10*

First, after setting up the environment for your dataset, we look at the data and look at the unique values of the each of the provided quasi identifiers.

A screenshot of a computer program

Description automatically generated

Based on this information above, we divide each of the attribute to different hierarchies. Since, age was continuous data, there was no point in extracting unique values for that quasi identifier. The hierarchy for age was divided with 10 year interval from ages 0-50, after that we use 2 intervals, 50-64 as 64 is the retirement age and then everything is defined into one group. The hierarchy for education was then setup to being in four different buckets. The first bucket maps anyone lower than a high school education to “Elementary”, the second bucket includes everyone who has some level of high school education, the third bucket maps people who did either vocational training or professional school to “Trade School” and then everyone with a higher education is mapped to “University Education”. The hierarchy for Marital Status was devised in a similar manner where the first bucket was everybody who was Single, the second bucket includes everyone who was married and is currently with their spouse, the last bucket includes all the people who were married but are not with their spouse anymore. The hierarchy for Race was mapped in a way that everybody who had a race assigned was classified as either as a major race or other.

After defining the hierarchies for generalization for all the quasi identifiers, we create an algorithm to generalize the data into the hierarchies for each attribute as defined above. The algorithm is defined below:

A screenshot of a computer program

Description automatically generated

After concluding the generalization over the dataset, we check if the data is k-anonymous, taking the k-value = 10. The generalization over the dataset did not ensure k-anonymity, therefore, we apply another method to achieve k-anonymity. The data anonymization technique I chose to go with was “Suppression”. The implementation for suppression was done by grouping the data in the sensitive columns, which for us were Age, Education, Marital-Status and Race. Then we add all the non-anonymized data to a bucket and drop the data that has less than k-value of indistinguishable tuples. At this point we would receive a k-anonymous dataset regardless, however, we need to measure the data loss over dropped values. If the data loss, is significant, we have to go with another data anonymization technique. Upon calculation, we only lose about 1% of our data, therefore, suppression for our dataset would work and the dataset we output to is k-anonymized, which is also checked in the code.

A screenshot of a computer program

Description automatically generated

**Output Dataset:** k\_anonymous\_dataset.csv