**Incisive Excel Anonymization Tool**

# **Introduction**

This tool is created as part of the Incisive project. It is a cli tool that the purpose of it is to anonymize the excel data of patients with breast, lung colorectal and prostate cancer. It works specifically for the Incisive excel data and their needs. Mostly it applies differential privacy on numerical data, leveling up medical history and medicines data that follow “who.int” and “who.cc” terminology respectively, grouping ethnicities and changing Ids of patients and data Providers. It also applies some curations to the data and after execution calculates anonymization metrics.

# **Requirements**

This tool requires Python 3.x for proper execution. Please ensure that Python 3.x is installed on your system before running the program.

To install Python 3.x, please visit the official Python [website](https://www.python.org/) and follow the installation instructions specific to your operating system.

Once Python 3.x is installed, you can follow the below installation process.

# **Installation**

For the installation of the tool, you must follow the below steps.

First, you need to clone in your machine this [repository](https://github.com/DimitrisTrakos/Anonymazation_Excel_Data) from github.

|  |
| --- |
| git clone https://github.com/DimitrisTrakos/Anonymazation\_Excel\_Data.git |

After cloning the repository, it is preferred to create a venv environment.

|  |
| --- |
| python3 -m venv venv |

Activate the environment according to the operating system used by the machine that is the tool .

Macos/Linux

|  |
| --- |
| source venv/bin/activate |

Windows

|  |
| --- |
| source venv/bin/activate |

Install Requirements

|  |
| --- |
| pip install -r requirements.txt |

The tool is installed and ready for use.

# **Usage**

First before running the tool, you must have a folder with this structure name:

<cancer\_type/data\_provider/data>.

|  |
| --- |
| An example of a folder path is this: |

|  |
| --- |
| incisive2/breast/auth/data |

It also needs a json file that will have the mapping of the ids of the patients.

It must follow this structure name:

|  |
| --- |
| id\_mapping\_<cancer\_type>.json |

An example of a name of the json file is this:

|  |
| --- |
| id\_mapping\_prostate.json |

An example of a json file content is this:

{

"034438": "959256",

"0043237": "56535"

}

You can run the tool with is command:

|  |
| --- |
| python anonymize\_excel.py <folder\_path> |

# **Configurations**

For every cancer type and their sheets there is a folder with a text file dedicated for modifying the columns that differential privacy it will apply and setting the parameters of the algorithm. These parameters are epsilon, maximum change of values, minimum accepted value in the column and round the values.

The folder that has the differential privacy configurations is located inside the root folder of the tool with the name “data”.

Differential privacy parameters must follow these types:

|  |
| --- |
| Column Name: str Epsilon: float Max change: float Min Value: float Round Values: boolean |

Example of the format of content of the text file

|  |
| --- |
| <column\_name>,<epsilon>,<max\_change>,<min\_value>,<round\_values> |

An example of a content of a text file is this:

|  |
| --- |
| Age at diagnosis,1,1,0,True Delivery Time,1,1,0,True |

# 

# **Calculated Metrics**

The tools after the anonymization is calculated and prints as a report the below metrics.

### Total Unique Values Original

This metric indicates the total number of unique values present in the original dataset column before anonymization. It provides insight into the diversity of data types or categories present in the original dataset, which informs decisions about how to generalize or transform the data during anonymization.

### Total Unique Values Anonymized

Like the Total Unique Values Original metric, this measures the total number of unique values in the anonymized dataset column after the anonymization process. It reflects the level of data reduction or generalization applied during anonymization to protect privacy while retaining data utility.

### Common Values Count

This metric counts the number of values that remain consistent between the original and anonymized datasets. It assesses the degree to which the anonymization process preserves key information or characteristics from the original dataset while obfuscating sensitive details.

### Exclusive Values in Original

This metric identifies any unique values present only in the original dataset column. It highlights data elements that may require special handling or transformation during anonymization to ensure comprehensive privacy protection without compromising data integrity.

### Exclusive Values in Anonymized

Similarly, this metric measures any unique values present only in the anonymized dataset column. It helps evaluate the effectiveness of anonymization techniques in preventing reidentification by ensuring that no new identifying information is inadvertently introduced.

### Percentage of Original Data Retained

### This metric is calculated by dividing the number of unique values present in both the original and the anonymized/transformed dataset by the total number of unique values in the original dataset, and then multiplying the result by 100. It provides an insight into how much of the original information is kept intact after the data processing steps.



Percentage Overlap:

This metric is calculated by dividing the number of unique values that are present in both datasets by the total number of unique values in one of the datasets (usually the modified or smaller dataset), and then multiplying the result by 100. It is especially useful in assessing how much of the data in a derived dataset (such as an anonymized or transformed dataset) is representative of the original dataset or another dataset it is being compared with.



### Percentage of New Values in Anonymized

### This metric calculates the percentage of new, previously unseen values introduced in the anonymized dataset. It helps assess the potential impact of anonymization on data quality and completeness by identifying any discrepancies or additions in the anonymized data compared to the original dataset.

### Jaccard similarity coefficient

Measures the similarity between two sets by calculating the intersection divided by the union of the sets. It offers a standardized way to quantify the degree of similarity or overlap between the original and anonymized datasets, facilitating comparison and evaluation of anonymization techniques.